Action Potential

JOHN A. WHITE
Boston University

GLOSSARY

activation The time-dependent growth of a membrane conductance in response to membrane depolarization.

all-or-nothing A term that refers to the property that action potentials, if they occur, have a stereotyped shape that is largely independent of the size and form of the suprathreshold stimulus.

current clamp An experimental protocol in which transmembrane current is controlled, usually at a series of constant values, and resulting transmembrane potentials are measured.

deactivation The time-dependent reversal of activation in response to membrane hyperpolarization; leads to a decrease in membrane conductance.

deinactivation The time-dependent reversal of inactivation, triggered by hyperpolarization; leads to an increase in membrane conductance.

depolarization Making membrane potential less negative.

hyperpolarization Making membrane potential more negative.

inactivation The time-dependent decline of a conductance (e.g., the Na^+ conductance), which follows after its activation; triggered by depolarization.

membrane potential The voltage difference across the neural membrane, determined by the balance of ionic fluxes across the plasma membrane.

refractory period The period immediately after an action potential, in which it is difficult or impossible to induce a second action potential.

space clamp The condition in which membrane potential is the same throughout the spatial extent of the cell.

threshold The value of membrane current or membrane potential necessary to induce an action potential.

voltage clamp An experimental protocol in which membrane potential is controlled, usually in a stepwise fashion, and resulting transmembrane currents are measured.

voltage-gated ion channels Transmembrane proteins that open in response to changes in membrane potential, allowing a particular ionic species to cross the membrane.

The action potential is the all-or-nothing electrical impulse used to communicate information between neurons and from neurons to muscle fibers. The energy used to generate action potentials is in the form of electrochemical gradients of ions (in particular, sodium and potassium) that are established by ion pumps. The rising phase of action potentials is caused by the autocatalytic activation of many Na^+-selective ion channels in response to sufficiently large increases in membrane potential. The falling phase of the action potential is caused by two factors that develop more slowly but dominate the electrical response after a few milliseconds: the inactivation of sodium channels and the activation of potassium channels, both of which occur in response to depolarization. Understanding the diverse mechanisms underlying electrical excitability in neurons remains a rich field of experimental and theoretical study, with wide-ranging implications for human health.

I. BASIC PROPERTIES OF THE ACTION POTENTIAL

The basic properties of the action potential can be studied using a microelectrode constructed from a glass capillary tube with a fine tip and containing artificial intracellular solution. This microelectrode,

inserted into the cell body or axon of a neuron (Fig. 1a, inset), measures the value of membrane potential relative to the extracellular space. At rest, typical values of membrane potential range from −40 to −90 mV. Passing positive electrical current into the cell depolarizes it (i.e., makes membrane potential less negative). In response to small depolarizing stimuli, the neuron's response is small as well (Fig. 1a, bottom). In response to larger stimuli, above a threshold value, the response is fundamentally different; the membrane potential quickly rises to a value well above 0 mV and then falls over the course of 1–5 msec to its resting value (Fig. 1a, middle). Often, the falling phase of the action potential undershoots resting potential temporarily. The action potential is said to be all-or-nothing because it occurs only for sufficiently large depolarizing stimuli, and because its form is largely independent of the stimulus for suprathreshold stimuli. In some neurons, a single action potential can be induced by the offset of a hyperpolarizing stimulus

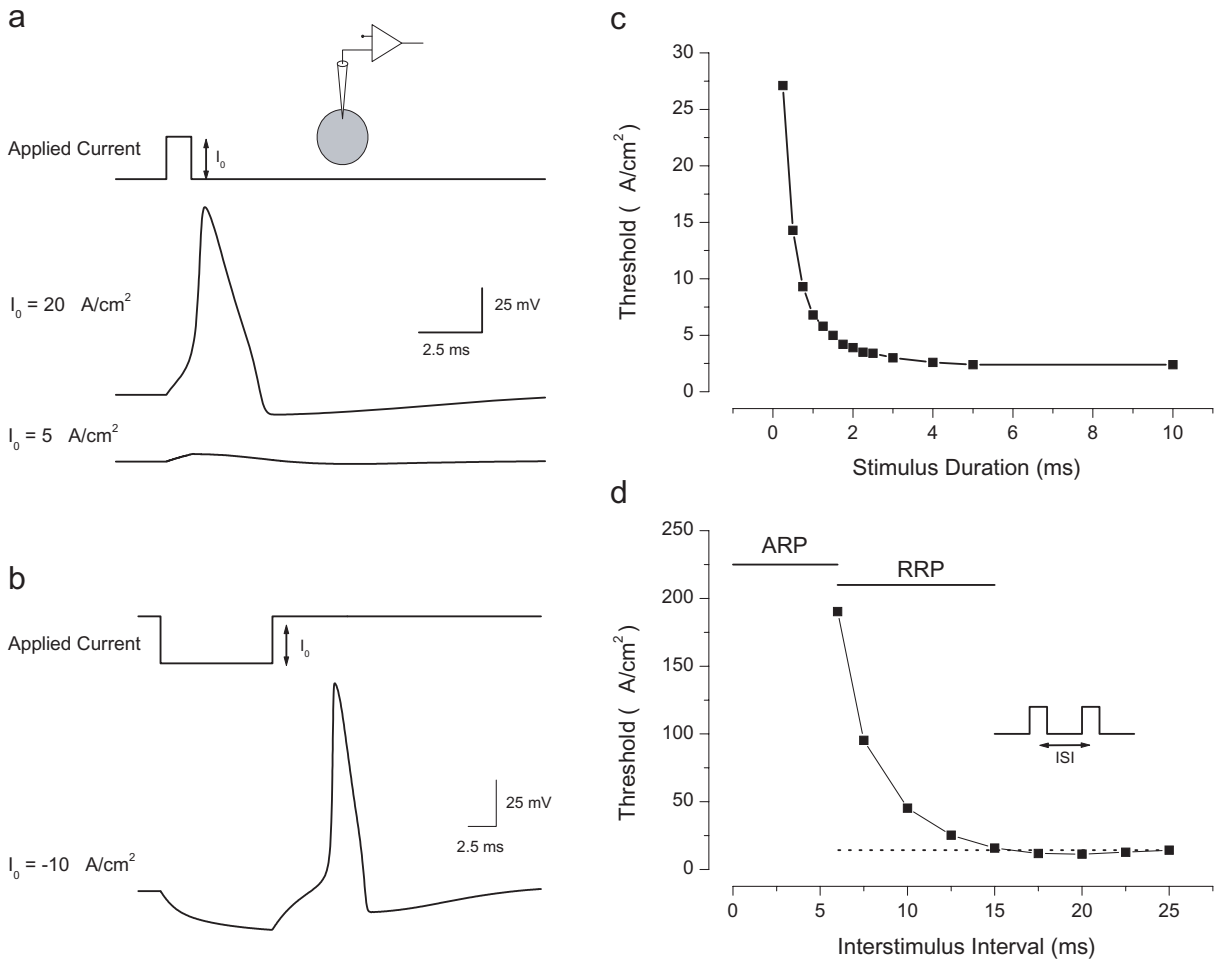


Figure 1 Basic properties of the action potential. (a) Traces show responses of a simulated space-clamped squid axon ($T = 6.3°C$) to intracellularly injected current pulses of duration 0.5 msec (top trace). The simulated recording configuration is shown in the inset. Sufficiently large inputs evoke all-or-nothing action potentials (middle trace). The response is minimal to subthreshold stimuli (bottom trace). The inset shows the basic recording configuration. (b) A simulation demonstrating anode-break excitation in response to the offset of a hyperpolarizing current pulse (duration = 10 msec). (c) Current threshold (the minimal amplitude of a current step necessary to evoke an action potential) plotted vs stimulus duration. (d) Simulation results demonstrating refractoriness. Two current pulses (duration = 0.5 msec each) were delivered to the model, with interstimulus interval (ISI) varied systematically. The first pulse had magnitude twice the threshold for evoking an action potential. The *y*-axis shows the magnitude of the second pulse necessary to evoke a spike. For ISI < 15 msec, threshold is above its normal value (dashed line). During the relative refractory period (RRP), threshold is elevated; during the absolute refractory period (ARP), it is not possible to evoke a second action potential.

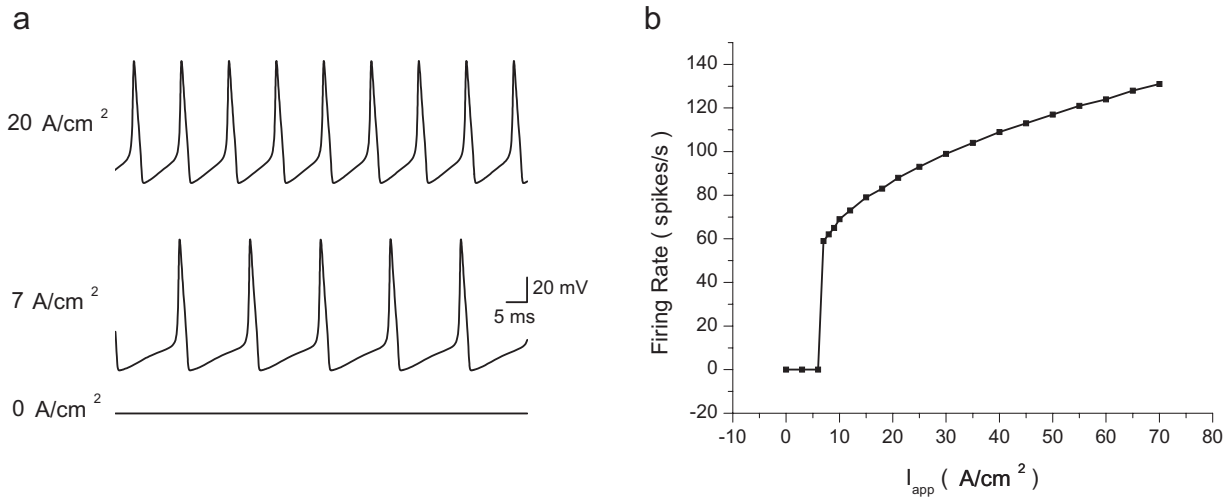


Figure 2 Spike rate depends on the magnitude of applied current. (a) Simulated traces of space-clamped squid giant axon ($T = 6.3°C$) to constant applied current. (b) Firing rate increases with increasing applied current. Note that the minimal firing rate is well above zero spikes/sec.

(Fig. 1b). This phenomenon is called anodal break excitation or rebound spiking.

The value of threshold depends on the duration of the stimulus (Fig. 1c); brief stimuli are required to be larger to evoke an action potential. Threshold also depends on more subtle features of the stimulus, such as its speed of onset. For a short time after an action potential has occurred, it is impossible to evoke a second one (Fig. 1d). This period is referred to as the absolute refractory period (ARP). After the ARP comes the relative refractory period (RRP), in which an action potential can be evoked, but only by a larger stimulus than was required to evoke the first action potential. Stimulation by an ongoing suprathreshold stimulus leads to repetitive firing at a rate that is constant once any transients have settled out (Fig. 2a). The rate of repetitive firing increases with increasing depolarization (Fig. 2b), eventually approaching the limit imposed by the ARP.

Once initiated, the action potential propagates down the axon at an approximately constant velocity. The leading edge of the action potential depolarizes adjacent unexcited portions of the axon, eventually bringing them to threshold. In the wake of the action potential, the membrane is refractory, preventing reexcitation of previously active portions of the cell. In unmyelinated axons, the action potential travels smoothly, with constant shape and at constant velocity. In myelinated axons, conduction is saltatory: The action potential ''jumps'' nearly instantaneously from one node of Ranvier to the next, greatly increasing the speed of propagation.

II. CLASSICAL DESCRIPTIONS OF THE ACTION POTENTIAL

A. Electrochemical Potentials and Voltage-Dependent Membrane Conductances

Changes in electrical potential in excitable cells are driven by movement of ions through ion-specific membrane conductances. For a perfectly specific conductance G, the current across the membrane $I = G \times (V_m - V_n)$, where V_m is the electrical potential across the membrane and V_n is the equilibrium potential for the ion, given by the Nernst equation:

$$V_n = \frac{RT}{z_n F} \ln \left(\frac{[X_n]^{\mathrm{o}}}{[X_n]^{\mathrm{i}}} \right)$$

where $R = 8.314$ J/(mol K) is the gas constant, T is absolute temperature, z_n is the valence of ion n, $F = 9.648 \times 10^4$ C/mol and is Faraday's constant, $[X_n]^{\mathrm{o}}$ is the outer concentration of ion n, and $[X_n]^{\mathrm{i}}$ is the inner concentration of ion n. Intuitively, the equilibrium potential is the value of membrane potential at which ionic fluxes due to concentration gradients and voltage gradients cancel one another, leading to zero net flux of the ion. Note that ionic

current, as defined, is positive for outward flux of a positive ion. For neurons, $[Na^+]^o > [Na^+]^i$; consequently, V_{Na} typically ranges from 40 to 50 mV and $I_{Na} < 0$. In contrast, $[K^+]^i > [K^+]^o$; V_K ranges from -70 to -100 mV and $I_K > 0$. The fact that resting membrane potential is far from V_{Na} and close but typically not equal to V_K implies that these ions are out of equilibrium, and thus that there is a constant trickle of each, even at rest. This flux is opposed by energy-expending ionic pumps, most notably the Na^+/K^+ ATPase, which serve to maintain Na^+ and K^+ concentration gradients and, consequently, equilibrium potentials.

Among the first quantitative clues regarding the mechanisms underlying the action potential came from ionic substitution experiments demonstrating that Na^+ and K^+ are the primary ions responsible for the phenomenon. Other early experiments demonstrated that membrane conductance, but not membrane capacitance, changes during the course of the action potential. Together, these results suggested the hypothesis that fluxes of Na^+ and K^+, driven by changes in ion-specific conductances, are responsible for the action potential. It is important to note that changes in membrane potential are not induced by changes in intracellular concentrations of Na^+ and K^+: Ionic fluxes during individual action potentials are small enough that concentrations remain essentially unperturbed. Instead, ionic fluxes alter V_m by changing the distribution of charge very near the membrane.

B. The Hodgkin–Huxley Model of the Space-Clamped Action Potential

Researchers in the middle of the 20th century hypothesized that the Na^+ and K^+ conductances underlying the action potential are "gated" (i.e., turned on and off) by changes in membrane potential. To test this hypothesis, they devised methods to measure ionic fluxes while controlling membrane potential V_m at a fixed value throughout the length of the axon. The process of controlling V_m, called voltage clamping, simplifies the behavior of the hypothesized voltage-dependent "gates." The process of making V_m the same throughout the axon, called space clamping, prevents complex spatial spread of excitation. Under these conditions, Na^+ and K^+ fluxes can be isolated either by manipulation of ionic concentrations (and thus equilibrium potentials) or by using specific blockers of particular conductances. (Tetrodotoxin is the classic blocker of Na^+ conductances; tetraethyl ammonium blocks many K^+ conductances.) Isolated Na^+ and K^+ fluxes from simulations are shown in Fig. 3 for many values of membrane potential. As hypothesized, these fluxes are voltage dependent. Na^+ and

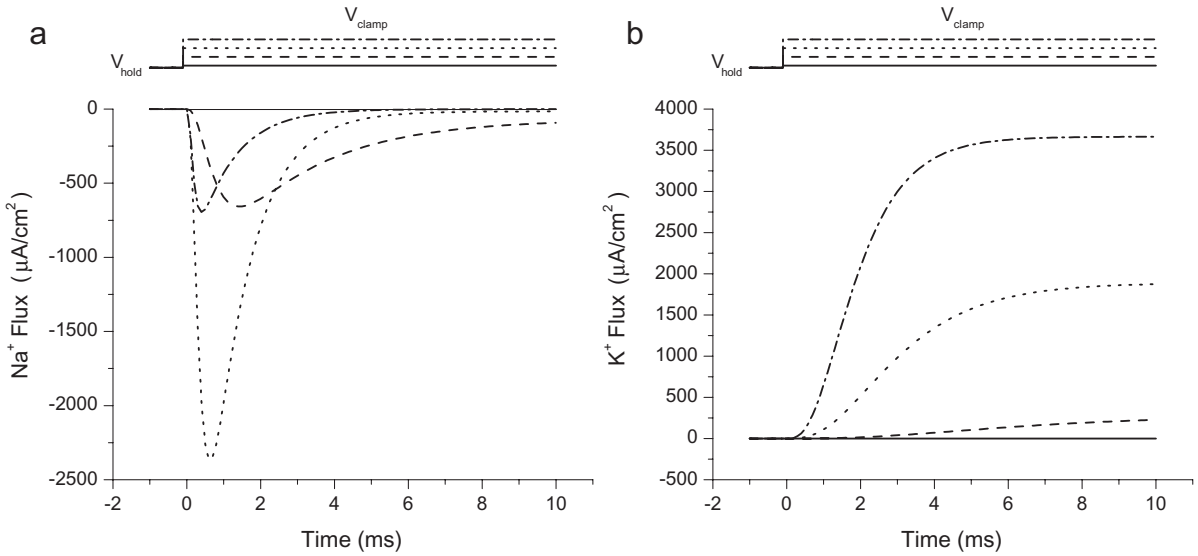


Figure 3 Simulated responses to voltage-clamp stimuli. Simulated Na^+ (a) and K^+ (b) in response to voltage-clamp steps from a holding potential of -90 mV to clamp potentials of -80 (solid lines), -40 (dashed lines), 0 (dotted llines), and 40 mV (dashed and dotted lines). The Na^+ flux is inward (negative) and is characterized by rapid activation and slower inactivation. The K^+ flux is outward (positive) and activates significantly more slowly than the Na^+ flux.

K^+ fluxes differ in several important aspects. First, they are opposite in sign for most values of membrane potential: The Na^+ flux depolarizes the neuron, whereas the K^+ flux hyperpolarizes the cell. Second, the Na^+ flux turns on ("activates") much more quickly than the K^+ flux. Third, the Na^+ flux turns off ("inactivates") after a brief period of depolarization. In contrast, the K^+ conductance remains activated in response to a prolonged depolarizing stimulus.

A watershed event in the history of neuroscience was the development by Hodgkin and Huxley of a relatively simple mathematical model, derived from voltage-clamp studies of the giant axon of the squid, that accounts for the generation and propagation of the action potential. The Hodgkin–Huxley model describes the membrane as an electrical circuit (Fig. 4a) that includes descriptions of membrane capacitance C_m; voltage-gated, Na^+- and K^+-selective conductances (G_{Na} and G_K, respectively), each in series with a battery representing the appropriate equilibrium potential; and a constant "leak" conductance that passes more than one ion. Mathematically, the Hodgkin–Huxley model includes four differential equations, which describe how the derivatives of membrane potential and three "gating variables" (variables that range between 0 and 1 and determine the values of voltage-gated conductances) behave. Two time- and voltage-dependent gating variables determine the size of the Na^+ conductance: The m gate captures the rapid activation of the Na^+ conductance after a step depolarization, whereas the h gate describes the slower inactivation process by which the

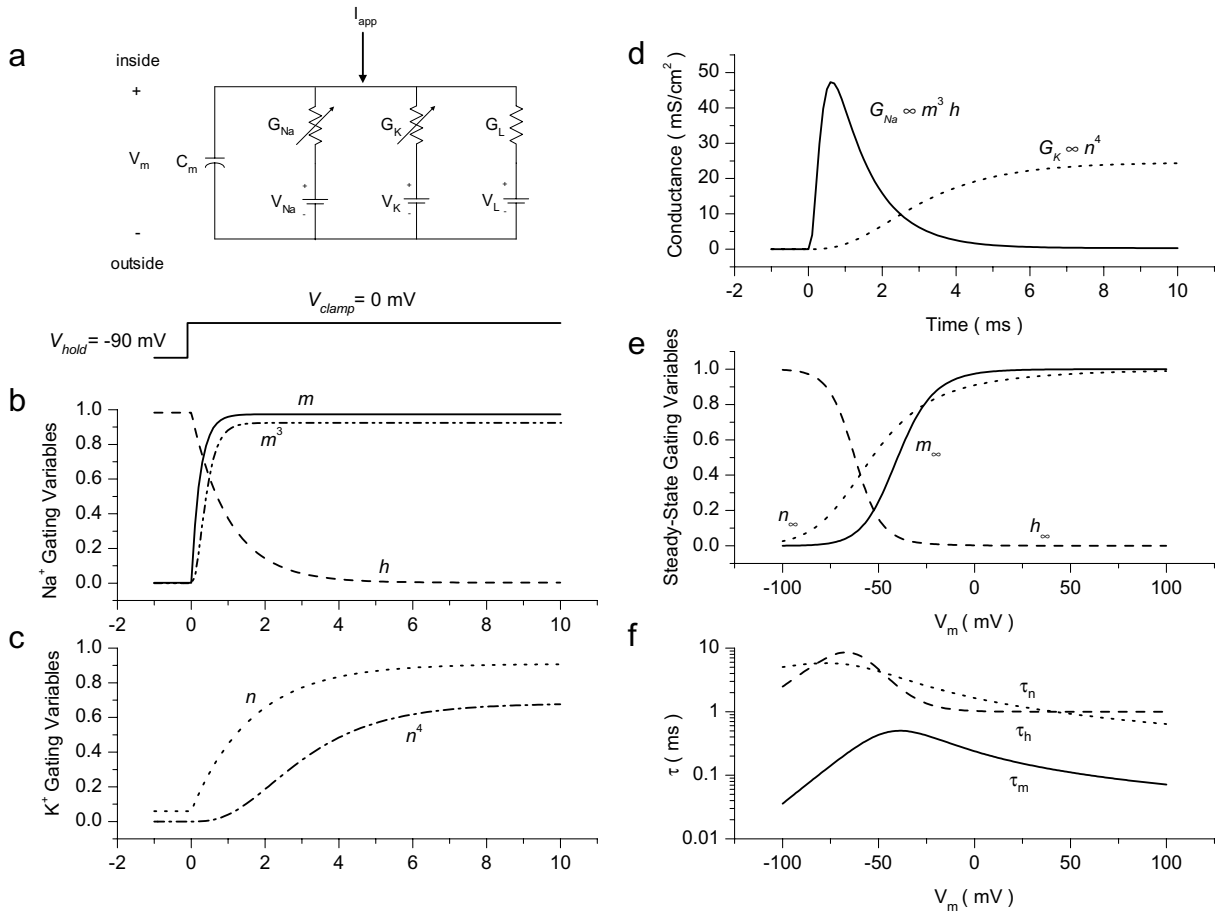


Figure 4 The Hodgkin–Huxley model of voltage-gated Na^+ and K^+ conductances. (a) Electrical circuit representation of the Hodgkin–Huxley model of squid giant axon under space-clamped conditions. (b–d) Responses of the Hodgkin–Huxley gating variables to a voltage-clamp step from $V_{hold} = -90\,mV$ to $V_{clamp} = 0\,mV$. The m and h gates determine the value of $G_{Na} \propto m^3h$. The n gate determines the value of $G_K \propto n^4$. (e) Steady-state values of m, h, and n for the entire physiologically relevant range of membrane potential V_m. (f)Time constants describing how quickly the gating variables m, h, and n reach their steady-state values, plotted vs V_m. Note the log scale on the y-axis.

Table I
Definitions and Units of Mathematical Symbols

C_m	Membrane capacitance per unit area ($\mu F/cm^2$)
$\bar{G}_{Na}, \bar{G}_K$	Maximal values of voltage-gated Na^+ and K^+ conductances per unit area (mS/cm^2)
G_{Na}, G_K	Voltage-gated Na^+ and K^+ conductances per unit area (mS/cm^2)
G_L	Leak conductance per unit area (mS/cm^2)
m, h, n	Voltage-dependent gating variables that determine magnitudes of voltage-gated conductances (dimensionless)
$m_\infty, h_\infty, n_\infty$	Voltage-dependent steady-state values of gating variables (dimensionless)
τ_m, τ_h, τ_n	Voltage-dependent time constants associated with gating variables (msec)
$\frac{dm}{dt}, \frac{dh}{dt}, \frac{dn}{dt}$	Time derivatives of gating variables (1/msec)
V_{Na}, V_K	Equilibrium potentials for Na^+ and K^+ (mV)
V_L	Reversal potential for the multi-ion leak conductance (mV)
I_{app}	Applied current flux ($\mu A/cm^2$)
V_m	Membrane potential (mV)
$\frac{dV_m}{dt}, \dot{V}_m(t)$	Time derivative of membrane potential (mV/msec)

Na^+ conductance turns off in response to prolonged depolarization. One gating variable, the n gate, describes voltage-dependent activation of the K^+ conductance G_K. An additional conductance (the leak conductance G_L) is voltage independent and small. In mathematical terms, the Hodgkin–Huxley equation is written as follows, with symbols defined in Table I:

$$C_m \frac{dV_m}{dt} = -[G_{Na}(V_m - V_{Na})$$

$$+G_K(V_m - V_K) + G_L(V_m - V_L)] + I_{app}$$

$$G_{Na} = \bar{G}_{Na} m^3 h \qquad G_K = \bar{G}_K n^4$$

$$\frac{dm}{dt} = \frac{m_\infty(V_m) - m}{\tau_m(V_m)} \qquad \frac{dh}{dt} = \frac{h_\infty(V_m) - h}{\tau_h(V_m)}$$

$$\frac{dn}{dt} = \frac{n_\infty(V_m) - n}{\tau_n(V_m)}$$

In response to a step change from an initial value of membrane potential (often referred to as the holding potential, V_{hold}) to the clamp potential, V_{clamp}, each of the Hodgkin–Huxley gating variables (m, h, and n) changes from an initial value to a steady-state value with an exponential time course (Figs. 4b and 4c). The steady-state values (m_∞, h_∞, and n_∞) and exponential time constants (τ_m, τ_h, and τ_n) are determined solely by the current value of V_m, which equals V_{clamp} for voltage-clamp experiments (Figs. 4e and 4f). The initial values of the gating variables are determined by the holding potential. G_{Na} is proportional to $m^3 \times h$; G_K is proportional to n^4 (Figs. 4b–4d). The powers to which the gating variables m and n are raised were used by Hodgkin and Huxley to induce small delays in activation of the conductance in order to better match experimental data.

C. The Hodgkin–Huxley Model Accounts for the Basic Properties of the Action Potential

Although the Hodgkin–Huxley model is incorrect in some details (see Section III.A), it can account for many of the basic properties of the neuronal action potential. For example, two properties of the m gate account for the phenomenon of the all-or-nothing action potential with a distinct threshold in response to a particular type of stimulus:

1. Because the m gate activates (opens) with depolarization, and its activation leads to further depolarization, this gate is prone to autocatalytic "positive feedback" that can magnify a small depolarization into a full-blown action potential. The current threshold for firing an action potential is the amount of current required to engage this cycle of positive feedback (see Section III.D for a discussion of the threshold value of membrane potential). In contrast with the m gate, the h and n gates react to oppose depolarization (h because it becomes smaller with depolarization, and n because its activation increases the size of an outward K^+ current).
2. The m gate is much faster than the h and n gates (Fig. 4e). The speed of the m gate means that the rapid rising phase of the action potential can occur before the stabilizing influences of the h and n gates can engage to bring V_m back to near resting potential. The speed of inactivation of h (i.e., its decrease with depolarization) and activation of n determine the width of the action potential.

Traces of each of the gating variables during the course of a depolarization-induced action potential are shown in Fig. 5a. Like spike threshold and spike duration, the phenomena of absolute and relative refractory periods can be accounted for by tracking gating variables. The ARP is associated with elevated values of n and, most important, greatly reduced

values of *h* after a spike (Fig. 5a). These factors make it impossible for a second spike to be elicited soon after the first. The RRP lasts as long as it takes for the *h* and *n* gates to return to their baseline values (about 15 msec in squid giant axon).

At resting potential, the voltage-gated Na^+ conductance is partially inactivated in that the *h* gate is partly closed and the *n* gate is partly open (Figs. 4 and 5). Hyperpolarizing the neuron below resting potential increases the value of *h*, in a process called deinactivation, and decreases the value of *n*, in a process called deactivation. Deinactivation of the Na^+ conductance and deactivation of the K^+ conductance can leave the neuron more excitable after hyperpolarization, and thus can account for anodal break excitation (Fig. 5b), also known as rebound spiking. Neurons typically fire only one rebound spike because after that spike the Na^+ and K^+ conductances return to their baseline states and the cell returnes to its resting level of excitability.

The first major success of the Hodgkin–Huxley model was that this relatively simple model, derived from voltage-clamp experiments, accounted successfully for many aspects of the action potential in the current-clamped and space-clamped axon. Even more impressive, and strikingly demonstrative of the level of understanding this model represents, was its ability to accurately account for the shape and velocity of the propagating action potential. The quantitative arguments involved are complex and will not be discussed here, but Fig. 6 shows qualitatively how action potentials propagate in unmyelinated axons.

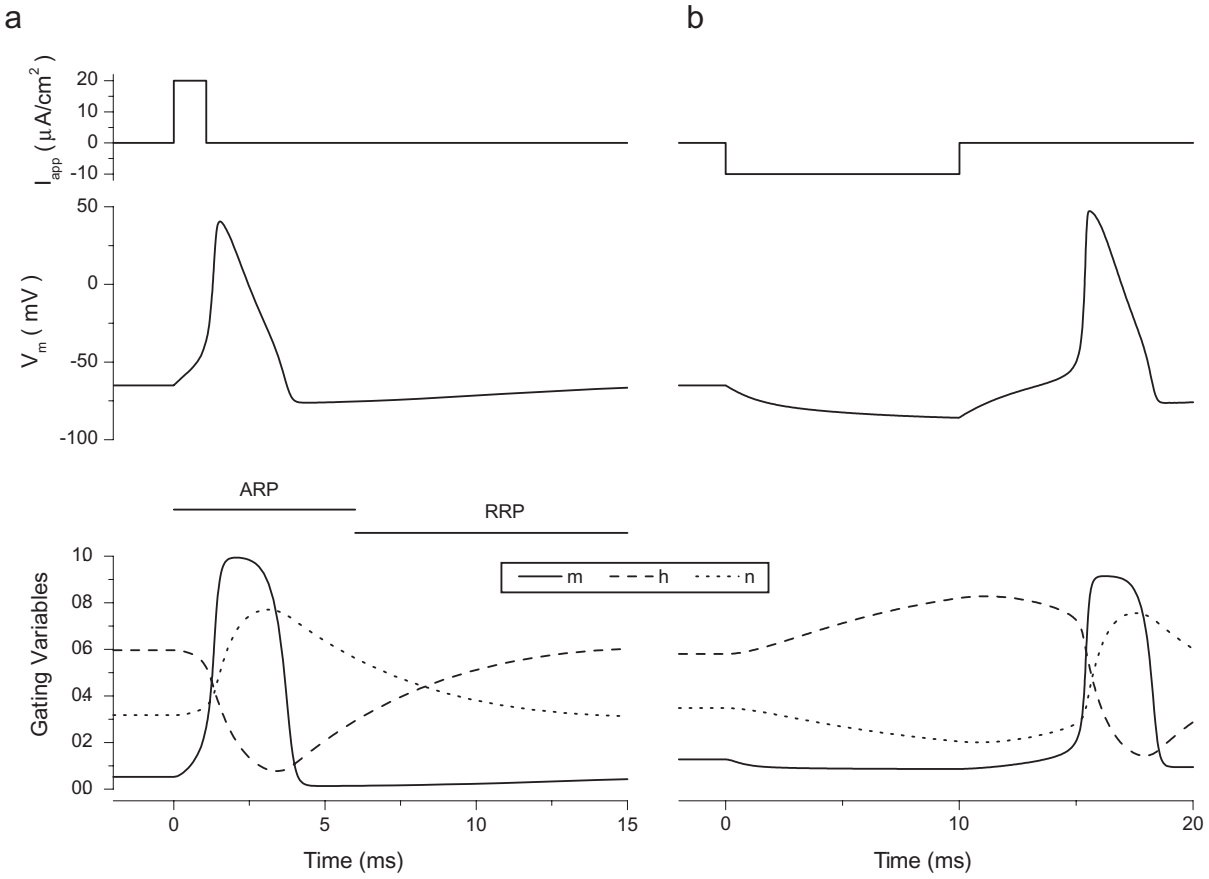


Figure 5 The Hodgkin–Huxley model accounts for axonal excitability. (a) From top to bottom, applied current I_{app}, membrane potential V_m, and each of the gating variables are plotted vs time. Curves were derived from the Hodgkin–Huxley model at 6°C. The rapid activation of the *m* gate underlies the action potential in response to this brief current pulse. The slower inactivation of the *h* gate and activation of the *n* gate repolarize the membrane a few milliseconds later. The duration of the absolute and relative refractory periods (ARP and RRP, respectively) is controlled by the duration of *h* gate inactivation and *n* gate activation. (b) After a hyperpolarizing input, the Hodgkin–Huxley model (6°C, with enhanced density of the Na^+ conductance) can produce a rebound spike (also known as anode break excitation). Data are plotted in the same order as in (a). Deactivation of the *n* gate and deinactivation of the *h* gate underlie this phenomenon.

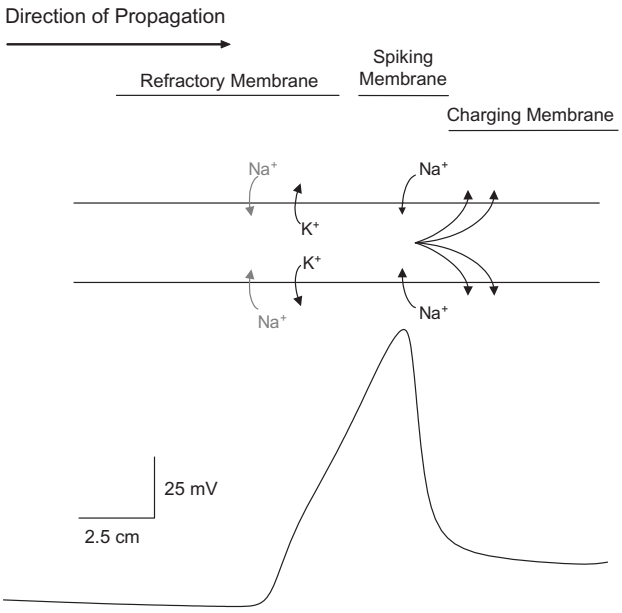


Figure 6 Propagation of the action potential in an unmyelinated axon. (Top) Schematic of the unmyelinated axon showing the sequence of events as an action potential propagates from left to right. The point of maximal Na^+ flux characterizes the locus where V_m is greatest. Positive charge from this point spreads to the right, gradually depolarizing the membrane on the leading edge of the action potential until threshold is reached. At the trailing edge of the action potential, to the left, the membrane is refractory. (Bottom) "Snapshot" of V_m plotted vs axial distance for the propagating action potential, with an assumed conduction velocity of 20 m/sec. Note that the form of the action potential is reversed when plotted vs distance rather than time.

Membrane potential reaches its peak value near the position of maximal Na^+ flux. Adjacent, unexcited membrane (Fig. 6, right) is depolarized by positive current from the site of Na^+ flux. Eventually, this depolarization is large enough that the membrane at the leading edge is excited as well. The membrane in the wake of the action potential is refractory (i.e., dominated by small G_{Na} and large G_K, and thus unable to spike) and thus unlikely to be reexcited. Elaborations of this model can account for saltatory conduction in myelinated axons.

III. CURRENT TOPICS RELATED TO THE ACTION POTENTIAL

A. Voltage-Gated Ion Channels Underlie Voltage-Dependent Membrane Conductances

In the past two decades, improvements in recording methodologies, pioneered by the group of Sakmann and Neher, as well as an explosion of knowledge in the field of molecular biology have demonstrated directly that membrane protein assemblies called ion channels underlie voltage-gated conductances in neurons and other excitable cells. The main (α) subunit of voltage-gated Na^+ channels (Fig. 7a) is a membrane protein approximately 2000 amino acids in length and has a molecular weight $>200{,}000$. This protein includes four domains, each of which consists of six putative transmembrane segments. A smaller β subunit is not necessary to form the channel but is crucial for regulating aspects of channel function such as the speed of inactivation. The main subunit of the K^+ channel has a very similar structure, except that the each protein codes for only one domain. Thus, four α subunits are necessary to form a K^+ channel.

Recordings from single channels (Fig. 7b) typically reveal two conductance states: "open" and "closed." Switching between open and closed states appears random, but the probability that the channel is in the open state varies with membrane potential. These probabilities correspond approximately to the values of gating variables in the Hodgkin–Huxley formulation. Sums of repeated recordings from an individual ion channel show behavior that is very similar to the macroscopic behavior of the channel population (Fig. 7b). Sophisticated analyses of single-channel recordings yield probabilistic models that can account for both microscopic and macroscopic behavior. These

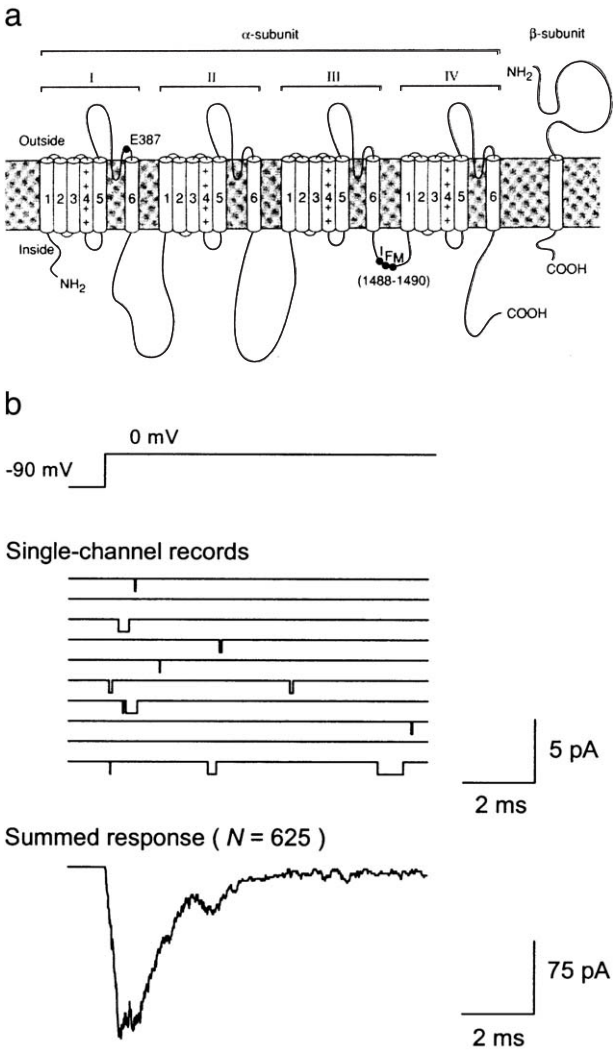


Figure 7 Na^+ channels underlie the voltage-gated Na^+ conductance. (a) Putative structure of the α and β subunits of the rat brain Na^+ channel (IIA). Roman numerals indicate the domains of the α subunit, each of which includes six putative transmembrane segments. Indicated residues are implicated in binding the channel blocker tetrodotoxin (E387) and in forming the inactivation gate (IFM 1488–1490) (adapted from Ashcroft (2000)). (b) The 10 middle traces show simulated single-channel recordings from a Na^+ channel under voltage clamp. The top trace shows the voltage-clamp command. The bottom trace shows the sum of 625 single-channel records.

models are similar in structure to Hodgkin–Huxley-type models but different in some details. For example, careful analysis of recordings from single Na^+ channels shows that their inactivation state is not controlled by an activation-independent process like the h gate but, rather, that Na^+ channels must activate before they can inactivate.

Experiments with channels that have been subjected to site-directed mutations have revealed close connections between particular loci on the protein and specific aspects of channel behavior. Among the properties that have been tied to specific loci on the α subunit are those of pore formation, ionic selectivity, voltage dependence, inactivation, and blockage by specific toxins (Fig. 7a).

B. Diversity of Mechanisms Contributing to Neuronal Excitability

In the approximately 50 years since the development of the Hodgkin–Huxley model, researchers have discovered many ion channel-based mechanisms for generating and controlling electrical activity in neurons. Often, these mechanisms rely on Na^+, Ca^{2+}, and K^+ channels. Sodium channels are relatively consistent in

their properties from case to case but do show some variety. In particular, some Na^+ channels do not exhibit fast inactivation; it is not clear whether these noninactivating Na^+ channels are a separate population from the typical, fast-inactivating Na^+ channels or a subset of the fast-inactivating Na^+ channels that have slipped into a different "mode" of gating. Calcium and potassium channels show more diversity than Na^+ channels. In particular, different classes of Ca^{2+} and K^+ channels show widely diverse properties with regard to the presence and speed of inactivation, pharmacological properties, and voltage ranges of activation. Some K^+ channels are sensitive to both membrane potential and the local intracellular concentration of Ca^{2+}, giving rise to interesting interactions between these two systems.

C. Modulation of Voltage-Gated Ion Channels

The properties of voltage-gated ion channels are not static. Many neuromodulators have been shown to alter ion channel function and the properties of action potentials by phosphorylating or dephosphorylating one or more sites on the channel protein. A host of neuronal firing properties, including spike width, average firing rate, and refractory period, are subject to metabolic control.

Ion channel expression profiles and neuromodulatory states can have profound consequences for neuronal function. A striking example of this point derives from thalamic relay neurons (Fig. 8). Depending on the level of depolarization or neuromodulatory state, these cells have two distinct firing modes. When relay neurons are hyperpolarized, they exhibit rhythmic bursting (Fig. 8a). The long interburst interval is determined by interaction of the low-threshold Ca^{2+} current and slowly activating, inwardly rectifying cation current. Superimposed on each slow Ca^{2+} spike are many fast action potentials, mediated by Na^+ and K^+ channels. When relay neurons are depolarized, either by electrical current or any of a number of neuromodulators, the low-threshold Ca^{2+} channels and inwardly rectifying cation channels are unimportant and the neurons fire in a tonic pattern much more reminiscent of the Hodgkin–Huxley model (Fig. 8b).

D. Mathematical Analyses of Neuronal Excitability

Although much about neuronal excitability can be learned by simply tracking changes in gating variables of the Hodgkin–Huxley-style model, computational neuroscientists and applied mathematicians have used mathematically based techniques to gain more general (and thus deeper) insights. Mathematical approaches

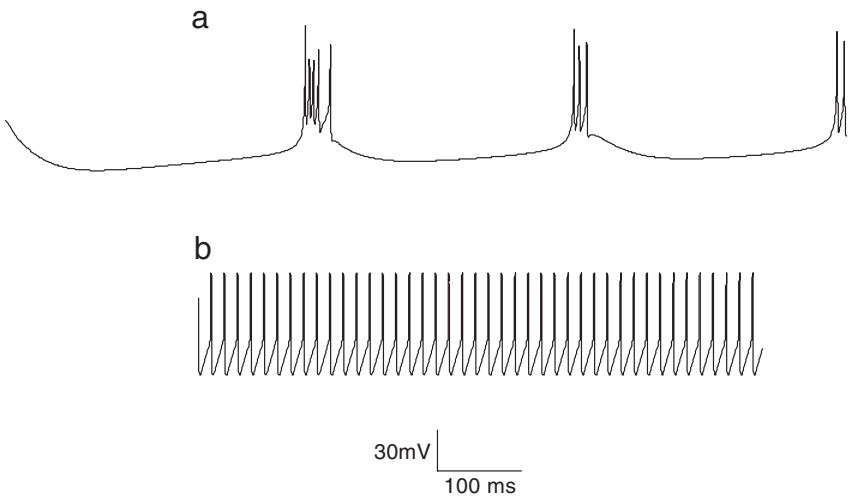


Figure 8 Thalamic relay neurons show two distinct firing modes. Shown are simulated responses of a thalamic relay neuron [Figure generated using the computational model from D. A. McCormick and J. R. Huguenard (1992). A model of the electrophysiological properties of thalamocortical relay neurons. *J. Neurophysiol.* **68,** 1384–1400]. (a) Under hyperpolarized conditions, relay neurons fire in an oscillatory manner. Slow oscillations are generated by interactions of the transient Ca^{2+} current I_t and the slow, hyperpolarization-activated cation current I_h. Fast action potentials, mediated by Na^+ and K^+, occur at the peaks of the Ca^{2+} spikes. (b) Depolarization by any of a number of means (e.g., current injection or neuromodulation) puts the neuron in a "tonic firing" mode. In this mode, I_t is inactivated and I_h is deactivated. Consequently, the cell's behavior is dominated by the Na^+ and K^+ currents exclusively.

based on the techniques of nonlinear dynamics have been successful in identifying the particular features that determine a host of important features of neuronal excitability, including threshold, the relationship between sustained firing rates and applied current, and particular patterns of bursting.

Figure 9 shows how nonlinear dynamics can be applied to understand neuronal threshold. Figures 9a and 9b show results from simulations of the Hodgkin–Huxley equations with zero applied current but two different initial conditions in membrane potential. As Figs. 9a and 9b demonstrate, excitable cells can be exquisitely sensitive to initial conditions. Examining the time derivative of membrane potential $\dot{V}_m(t)$ shortly after the perturbation in initial conditions gives insight into this phenomenon. Figure 9c shows results from hundreds of simulations conducted over a large range of initial values V_0. $\dot{V}_m(t)$, evaluated at $t = 0.5$ msec, is plotted vs V_0 (solid line; the value $t = 0.5$ msec was chosen because this is long enough for the m gate to react significantly to the perturbation away from resting potential but short enough that the h or n gates remain relatively near their resting values). Also plotted are the contributions to $\dot{V}_m(0.5)$ from Na^+ conductance (dashed line) as well as K^+ and leak conductances (dotted line). These contributions can be obtained easily from the Hodgkin–Huxley equation describing $\dot{V}_m(t)$. The effect of the Na^+ conductance is to elevate $\dot{V}_m(t)$; the effect of the K^+ conductance is to reduce $\dot{V}_m(t)$.

Figure 9d shows a magnified view of $\dot{V}_m(0.5)$ in the region near spike threshold. The plot shows two zero-crossings, at $V_0 = -65$ and -59 mV. These zero crossings, often called fixed points, are especially important because where $\dot{V}_m(t) = 0$, membrane potential V_m is by definition not changing, meaning that $\dot{V}_m$ has the potential to remain fixed at that point indefinitely (for a system with constant parameters). The slope of the curve at each zero crossing tells us much about the stability of that fixed point in response to small fluctuations (e.g., due to noise). For example, consider the special case of no perturbation. In this case, V_0 equals resting potential (-65 mV), which we expect to be a fixed point. For V_m slightly less than -65 mV, $\dot{V}_m(0.5) > 0$, implying that V_m will return to its resting value. For V_m slightly higher than -65 mV, $\dot{V}_m(0.5) < 0$, implying again that V_m will return to

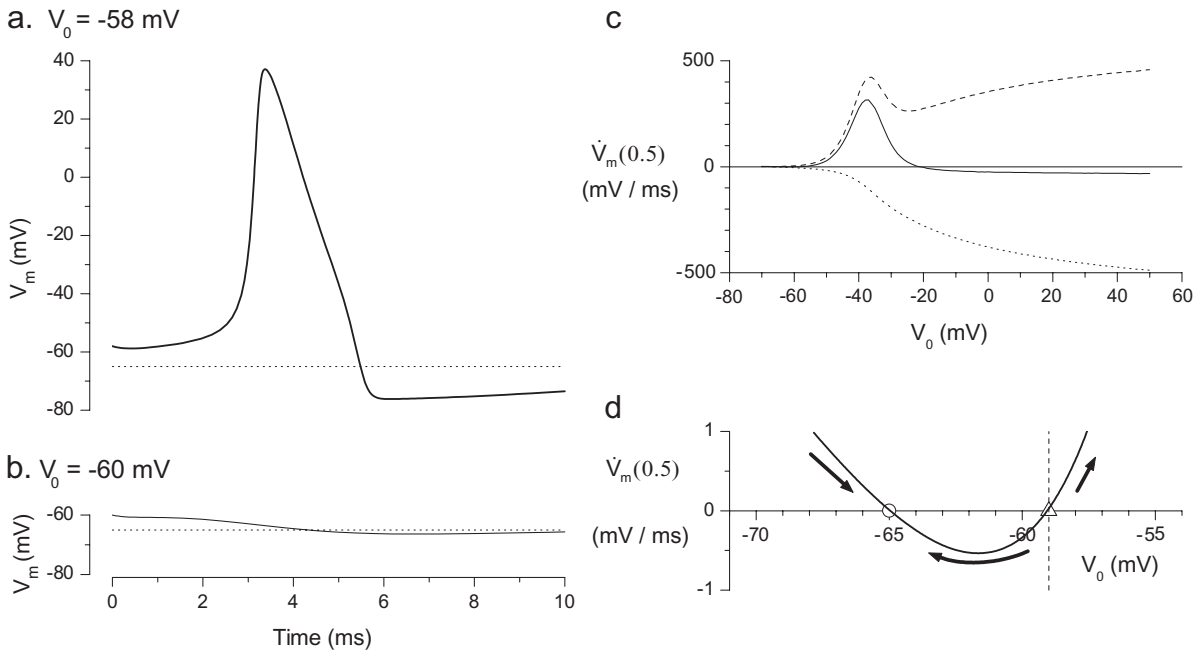


Figure 9 Threshold of the Hodgkin–Huxley model can be explained by examining the stability of "fixed points." (a, b) Spiking behavior in the Hodgkin–Huxley model is very sensitive to V_0, the initial value of voltage. Resting potential (-65 mV) is depicted by the dotted lines. (c) $\dot{V}(0.5)$, the time derivative of membrane potential at $t = 0.5$ msec, plotted vs V_0. Also plotted are the contributions of G_{Na} (dashed line), as well as G_K and G_L (dotted line), to $\dot{V}(0.5)$.The contribution of G_L is minimal because this conductance is very small. (d) A magnified plot of $\dot{V}(0.5)$ vs V_0. Two fixed points [points where $\dot{V}(0.5) = 0$] are shown. As indicated by the arrows, the sign of $\dot{V}(0.5)$ makes the solution flow toward the open circle at (-65, 0), indicating stability; the solution flows away from the open triangle at (-59, 0), indicating instability. For $V_0 > -59$ mV, an action potential will be generated.

–65 mV. The value $V_m = -65$ mV is said to be a stable fixed point because after small perturbations above or below that value, V_m will return to it.

Next, consider the fixed point at $V_0 = -59$ mV. In this case, $\dot{V}_m(0.5) < 0$ for $V_m < V_0$, and $\dot{V}_m(0.5) > 0$ for $V_m > V_0$. This result implies that this fixed point is unstable: After small perturbations above or below V_0, V_m will move away from the fixed point. The implication of this fact is that the point $V_0 = -59$ mV serves as a threshold. For $V_0 < -59$ mV, the model returns to resting potential (Fig. 9a); for $V_0 > -59$ mV, V_m rapidly increases as the action potential begins [Fig. 9b; in these cases, $\dot{V}_m(t)$ continues to evolve, eventually bringing the cell to rest]. For these simulations, the fixed point at $V_0 = -59$ mV corresponds exactly with the threshold for generation of an action potential (vertical dashed line in Fig. 9d).

E. Backpropagation of the Action Potential

In the traditional view, information flow in the mammalian neuron is in one direction only: The dendrite receives input from the presynaptic population, the soma integrates this information, the decision regarding whether or not to fire an action potential is made at or near the axon hillock, and the action potential is propagated to other neurons by the axon. This view has been amended in recent years as scientists have developed techniques for recording simultaneously from multiple locations within the neuron (e.g., the soma and a primary dendrite). In pyramidal neurons of layer V of neocortex, for example, suprathreshold synaptic input to the apical dendrites leads to initiation of an action potential near the soma. This action potential can then travel back from the soma toward the distal end of the apical dendrite. The reliability of backpropagation depends on recent patterns of input and spike generation. Given the crucial role of dendritic depolarization for synaptic plasticity, backpropagating action potentials may be important for experience-dependent alterations of neuronal circuitry in learning and memory.

F. Diseases Related to Mutations in Ion Channels

Given the central role that electrical excitability plays in nervous system function, it is not surprising that mutations of voltage-gated ion channels alter neuronal function. Although work in this area is just beginning, a host of maladies have been associated with nonlethal mutations of neuronal voltage-gated channels, including the following:

- Generalized epilepsy with febrile seizures, so-named because patients have fever-induced seizures that develop later in life into seizures without a clear trigger, is associated in some cases with a rare mutation of the β_1 subunit of the Na^+ channel. This mutation may promote epilepsy by slowing the inactivation process in neuronal Na^+ channels, leaving the brain hyperexcitable.
- Benign familial neonatal epilepsy is associated with mutations that lead to reduced expression of slow, voltage-gated K^+ channels of the KCNQ family, thereby leaving some neurons hyperexcitable.
- Some forms of episodic ataxia, a condition of triggered events of imbalance and uncoordinated movements, has been associated with many missense mutations of the $K_v1.1$ channel, which gives rise to an inactivating K^+ conductance. Ataxia-associated mutations of $K_v1.1$ have been shown to lead to pathologically rapid deactivation, enhanced inactivation, and increases in the threshold of activation. These disparate changes all have the effect of broadening the neuronal action potential, but it is not known how the broadened spike may lead to ataxia.

See Also the Following Articles

ELECTRICAL POTENTIALS • ELECTROENCEPHALOGRAPHY (EEG) • EVENT-RELATED ELECTROMAGNETIC RESPONSES • ION CHANNELS • NEURON • NEURAL NETWORKS

Suggested Reading

Aidley, D. J., and Stanfield, P. R. (1996). *Ion Channels: Molecules in Action*. Cambridge Univ. Press, Cambridge, UK.

Ashcroft, F. M. (2000). *Ion Channels and Disease*. Academic Press, San Diego.

Häusser, M., Spruston, N., and Stuart, G. (2000). Diversity and dynamics of dendritic signaling. *Science* **290,** 739–744.

Hille, B. (2001). *Ionic Channels of Excitable Membranes*, 3rd ed. Sinauer, Sunderland, MA.

Johnston, D., and Wu, S. M.-S. (1995). *Foundations of Cellular Neurophysiology*. MIT Press, Cambridge, MA.

Koch, C. (1999). *Biophysics of Computation*. Oxford Univ. Press, New York.

Koch, C., and Segev, I. (Eds.) (1998). *Methods in Neuronal Modeling*, 2nd ed. MIT Press, Cambridge, MA.

Nicholls, J. G., Martin, A. R., Wallace, B. G., and Fuchs, P. A. (2000). *From Neuron to Brain*, 4th ed. Sinauer, Sunderland, MA.

Sakmann, B., and Neher, E. (Eds.) (1995). *Single-Channel Recording*, 2nd ed. Plenum, New York.

Weiss, T. F. (1996). *Cellular Biophysics*. MIT Press, Cambridge, MA.

White, J. A., Rubinstein, J. T., and Kay, A. R. (2000). Channel noise in neurons. *Trends Neurosci.* **23,** 131–137.

Adolescent Brain Maturation

JAY N. GIEDD, ELIZABETH A. MOLLOY, and JONATHAN BLUMENTHAL
National Institute of Mental Health

GLOSSARY

arborization From the Latin word for tree, the process of brain cells growing extra "branches, twigs, and roots."

corpus callosum A collection of myelinated neuronal axons connecting similar areas of the left and right cerebral hemispheres.

magnetic resonance imaging A technique that combines a powerful magnet, radio waves, and computer technology to acquire images of the human brain without the use of harmful radiation.

pruning The process by which axonal or dendritic branches are cut back or eliminated.

use-it-or-lose-it principle The hypothesis that brain cells and connections that are used will survive and flourish, whereas those that are not used will wither and perish.

Any parent of a teen can attest that the brain of a 13-year-old is very different from the brain of a 9-year-old. However, actually defining these differences in a scientific way has been elusive, because nature has gone through a great deal of trouble to protect the brain. It is wrapped in a tough, leathery membrane surrounded by a protective moat of fluid and completely encased in bone. This has shielded the brain from falls or attacks from predators but it has also shielded the brain from scientists. However, recent advances in imaging technology now allow us to examine the living, growing human brain as never before. Magnetic resonance imaging (MRI) provides exquisitely accurate pictures of brain anatomy. It does so without the use of ionizing radiation, permitting not only the scanning of healthy children and adolescents but also repeat scans of the same individuals throughout their development. In this article, we discuss the anatomical changes that occur in the adolescent brain as detected by MRI.

I. BRAIN DEVELOPMENT

Human brain development, like central nervous system development in all vertebrates, takes place by an overproduction and then selective elimination of cells. Most of the dynamic activity of brain development takes place *in utero*. However, as indicated in Fig. 1, competitive elimination, myelination, and dendritic and axonal arborization continue throughout childhood and adolescence.

Brain cells are of two general types—neurons and glial cells. Although neuronal number peaks during gestation, neuronal size changes with age as axonal thickness, dendritic number, and the number of synaptic connections undergo cyclic changes throughout development. Environmental factors influence which synaptic connections and neurons thrive and remain viable. For instance, children with cataracts who do not receive treatment prior to age 1 suffer irreversible cortical blindness.

The other main class of central nervous system cells is the glial cells, which unlike neurons continue to

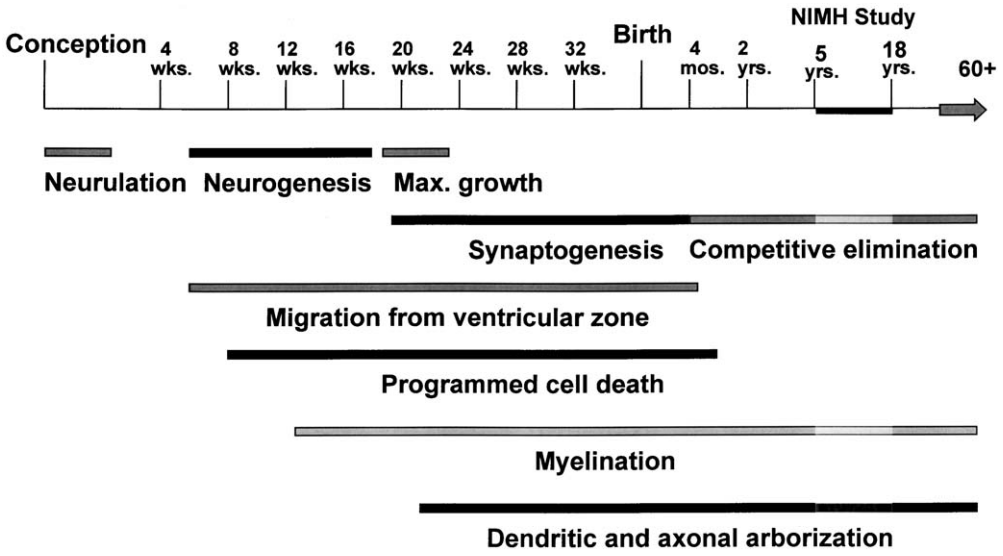


Figure 1 Time course of critical events in the determination of human brain morphometry (reproduced with permission from J. N. Giedd, 1997, Normal development. *Neuroimaging* **6,** 265–282).

actively proliferate and die postnatally. Glial cells outnumber neurons by ratios ranging from 1.7 to 10. Myelination by a subclass of glial cells, oligodendrocytes, is an important determinant of increases in structure size during childhood and adolescence. Ultimate structure size is determined by this dynamic interplay between glial cells and decreasing numbers but increasing size of neurons. Synaptic pruning is an important aspect in the functional development of the brain but may have little impact on overall structure size. Estimates from research on the primary visual cortex of the macaque monkey indicate that a total loss of all boutons would result in only a 1–2% decrease in volume. However, synaptic pruning may have an effect on the thickness of the parent axon or dendritic branches. Another parameter to consider in structure size is packing density, which is influenced by degree of vascularity, extracellular volume, and hydration.

Genetics, hormones, growth factors, nutrients in the developing nervous system, diet, infections, toxins, trauma, stress, and degree of enriched environment all play a role in determining structure size, and the complexity of these factors and their interactions should be considered in any interpretation of the significance of gross structural volume.

II. TOTAL CEREBRAL VOLUME

The brain consists of two cerebral hemispheres, the ventricles, the cerebellum, and the brain stem. The total size of the cerebral hemispheres changes little from childhood to adolescence, reaching 95% of its adult size by the age of 5 years. This is perhaps surprising to anyone who has watched an adult's hat falling down over the eyes of a child. The seeming discrepancy is due to the fact that head circumference does indeed increase from ages 4 to 18 (approximately 2.0 in. in boys and 1.9 in. in girls), but the increase is accounted for by an increase in skull thickness and less so by an increase in ventricular volume. Many factors, including intelligence, handedness, psychiatric illness, body size, and gender, have been related to total brain size in teens as well as adults.

A. Intelligence

Recent studies have found small but statistically significant relationships between brain size and intelligence. Although in the most robust of these findings IQ accounts for only 17% of the variance, this parameter should be considered in any group comparison. Education and socioeconomic status have been reported to influence brain size as well, although interdependence with factors such as nutrition, prenatal care, and IQ is not clear.

B. Psychiatric History

Abnormalities of brain structure have been observed in many pediatric neuropsychiatric illnesses, including

autism, attention deficit/hyperactivity disorder (ADHD), childhood-onset schizophrenia, dyslexia, eating disorders, fetal alcohol syndrome, obsessive–compulsive disorder, Sydenham's chorea, and Tourette's syndrome. It is evident that a normative sample must be carefully screened to rule out these conditions. Likewise, affective disorders and substance abuse have been associated with structural anomalies in adults and should be considered as potential confounds in pediatric samples as well.

C. Handedness

Beginning with Geschwind, several investigators have noted a relationship between handedness and structural symmetry measures of the brain. Handedness should not be viewed as strictly left or right but as a continuum, and it should be quantified as such. Patient and control groups must be matched for handedness since symmetry differences are often key features in discriminating groups such as ADHD, dyslexia, or Tourette's disorder.

D. Body Size

The relationship between brain size and body size in humans is surprisingly weak. In contrast to the relative stability of brain weight after childhood, body weight varies widely among individuals and can vary substantially within individuals from time to time. Height is also a poor indicator of brain size, as can be implied by contrasting the notable increases in height from ages 4 to 18 years with the lack of corresponding increase in brain size. This general trend for the young to have disproportionately large head-to-height ratios compared to adults is widely observed throughout the mammalian species.

E. Gender

As indicated by autopsy and imaging studies, the male brain is approximately 10% larger than the female brain across all ages. Of course, gross structural size may not be sensitive to sexually dimorphic differences in connectivity between different neurons, known differences in receptor density, or more subtle differences in the size or connectivity of various nuclei. Given the multiple parameters determining brain size, a larger size should not be interpreted as imparting functional advantage or disadvantage.

A plot of total cerebral volume versus age for 146 healthy boys and girls is presented in Fig. 2. As can be seen, brain sizes are highly variable. This large variability of brain sizes means that a larger number of subjects, or following the same subjects over time, is necessary to discern how brain anatomy changes during adolescence. The relative stability of total brain size throughout childhood and adolescence belies the dynamic activity of the various subcomponents of the brain. In the following sections, the effects of age and gender on these different parts of the brain are examined.

III. GRAY MATTER

Brain tissue can be broadly categorized as gray matter, white matter, or cerebrovascular fluid. These different tissue types are discriminated by MRI and define the anatomical boundaries of most brain structures. Gray matter generally consists of cell bodies and white matter of myelinated axon fibers. The gray matter changes during adolescence are presented for different regions of the cerebral cortex as well as for the subcortical basal ganglia, amygdala, and hippocampus.

A. Cortical Subdivisions

The brain is often divided into four lobes with functionally distinct properties: the frontal, temporal, parietal, and occipital lobes.

1. Frontal Gray Matter

The functions of the frontal lobe include planning, organizing, strategizing, as well as initiating, shifting, and sustaining attention. It is sometimes referred to as the "executive" of the brain. As seen in Fig. 2, frontal gray matter increases throughout childhood, peaks at 11 years in girls and 12 years in boys, and then declines throughout adolescence. The peaks correspond to the onset of puberty, although a direct relationship between hormonal changes of puberty and this process has yet to be established. The thickening of cortical gray matter is thought not to reflect an increase in the number of neurons but to be caused by an increase in the number and thickness of branches and connections on the dendrites and axons of existing neurons, a process called arborization. Following this peak of arborization is a pruning process whereby the number of branches and connections are selectively cut back.

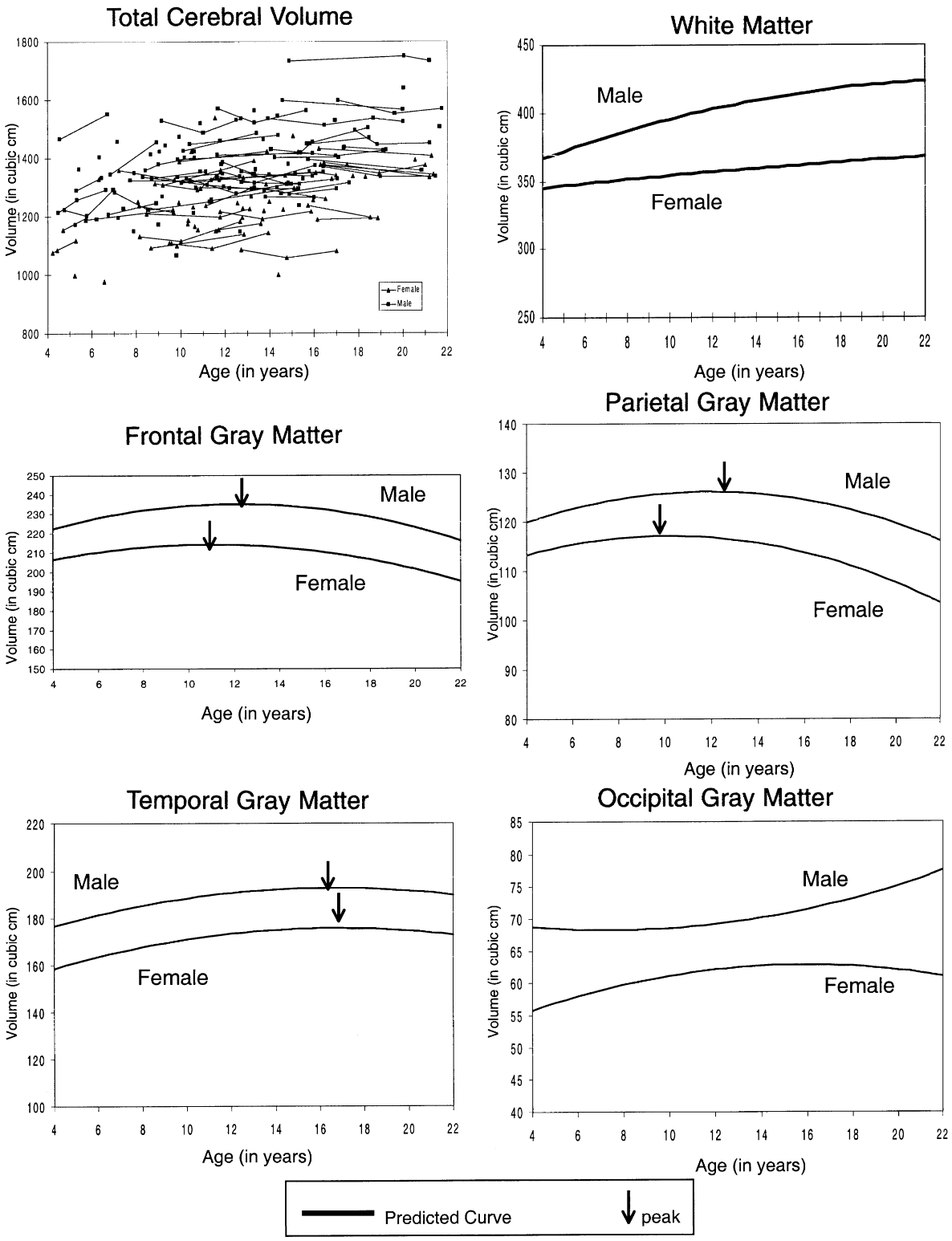


Figure 2 Predicted size with 95% confidence intervals for cortical gray matter in frontal, parietal, temporal, and occipital lobes for 243 scans from 89 males and 56 females, ages 4–22 years. The arrows indicate peaks of the curves (reproduced with permission from J. N. Giedd, *et al.*, 1999, Brain development during childhood adolescence. *Nat. Neurosci.* **2,** 861–863).

2. Parietal Gray Matter

The parietal lobes are primarily responsible for receiving and processing sensory input such as touch, pressure, heat, cold, and pain. The parietal lobes are also involved in the perception of body awareness and the construction of a spatial coordinate system (mental map) to represent the world around us. Individuals with damage to the right parietal lobe often show striking abnormalities in body image and spatial relations, such as failing to attend to part of the body or space (contralateral neglect). Patients with damage to the left parietal lobe often experience difficulty with writing (agraphia), an inability to recognize familiar objects (agnosia), and language disorders (aphasia). Bilateral damage can lead to problems with visual attention and motor abilities. As with the frontal gray matter, parietal gray matter increases during childhood and decreases during adolescence, peaking at 10.2 years in girls and 11.8 years in boys (Fig. 2).

3. Temporal Gray Matter

The temporal lobes subserve functions of language, emotion, and memory. As opposed to the frontal and parietal gray matter, the temporal gray matter does not peak until age 16.7 years in girls and 16.5 years in boys (Fig. 2). Electroencephalographic studies of adolescents and young adults also indicate ongoing maturation of the temporal lobes throughout the second decade of life.

4. Occipital Gray Matter

The occipital lobes are involved in visual information processing and object recognition. The gray matter in the occipital lobes continues to increase during childhood and adolescence (Fig. 2).

B. Subcortical Divisions

1. Basal Ganglia

The basal ganglia are composed of the caudate nucleus, putamen, globus pallidus, subthalamic nucleus, and substantia nigra. These structures are well-known to influence movement and muscle tone as indicated by their dysfunction in Parkinson's and Huntington's disease, but they are also integral components of circuits mediating higher cognitive functions, attention, and affective states. Of the basal ganglia components, only the caudate, putamen, and globus pallidus are large enough to be readily quantifiable by MRI. Like frontal and parietal cortical gray matter, the basal ganglia generally decrease in volume during the teen years.

2. Amygdala/Hippocampus

The amygdala and the hippocampus are adjacent gray matter structures in the medial temporal lobe. The amygdala is an almond-shaped structure that plays a key role in the brain's integration of emotional meaning with perception and experience. It coordinates the actions of the autonomic and endocrine systems and prompts release of adrenaline and other excitatory hormones into the bloodstream. The amygdala is involved in producing and responding to nonverbal signs of anger, avoidance, defensiveness, and fear. The amygdala has been implicated in emotional dysregulation, aggressive behavior, and psychiatric illnesses such as depression. It has also been shown to play an important role in the formation of emotional memory and in temporal lobe epilepsy.

The hippocampus is a horseshoe-shaped region that is involved in short-term memory storage and retrieval. Human capacity for these functions undergoes marked changes from ages 4 to 18 years. However, the relationships between changes in these abilities and changes in brain morphology are not well understood. New memories are kept in the hippocampus before being transferred to the cerebral cortex for permanent storage. This may explain why people with brain damage to their hippocampal region retain previous memories of faces and places, which are stored in the cortex, but have difficulty forming new short-term memories. The hippocampus is also implicated in the learning and remembering of space (spatial orientation). Animal studies show that damage to the hippocampus results in the inability to navigate through familiar areas. In humans, hippocampal damage results in a failure to remember spatial layouts or landmarks. Following stroke damage to the parahippocampus, patients lose the ability to learn new routes or to travel familiar routes.

During adolescence the size of the amygdala increases sharply for males and less sharply for females. In contrast to the amygdala, the hippocampus increases more robustly in adolescent females. These sex-specific maturational patterns are consistent with nonhuman primate studies that have found a predominance of androgen receptors in the amygdala and a predominance of estrogenic receptors in the hippocampus.

Other lines of evidence also support the influence of estrogen on the hippocampus. Female rats that have had their ovaries removed have a lower density of dendritic spines and decreased fiber outgrowth in the hippocampus, which can be alleviated with hormone replacement. In humans, smaller hippocampi have been reported in women with gonadal hypoplasia, and a recent MRI study of 20 young adults found relatively larger hippocampal volumes in females.

In addition to receptors for gonadal steroids, the hippocampus and amygdala are rich in receptors for adrenal steroids, thyroid hormone, and nerve growth factor. In addition to direct effects on hippocampal development, estrogen may influence development by blocking neurodegenerative effects of glucocorticoids.

The intricacy of various neurochemical systems and the diversity of afferent and efferent connections to the many distinct nuclei of most brain structures make straightforward relationships between volumes of a single structure and performance on a particular cognitive task uncommon. One of the rare exceptions to this rule is the relationship between hippocampal size and memory function. Birds that store food need better memory than their non-food-storing counterparts and correspondingly have larger hippocampi. Likewise, male polygamous prairie voles travel far and wide in search of mates and have significantly larger hippocampi and perform better on laboratory measures of spatial ability than their female counterparts. In the monogamous vole species, which do not show male-female differences in spatial ability, no sexual dimorphism of hippocampal size is seen. Correlations between memory for stories and left hippocampal volume in humans have also been noted.

Anomalies of temporal lobe and medial temporal lobe structures have been reported for a variety of psychiatric disorders, including affective disorders, autism, and, most consistently, schizophrenia, which is increasingly understood as a neurodevelopmental disorder. These disorders have marked sex differences in age of onset, symptomatology, and risk factors. The sex-specific maturational differences may have relevance to the expression of these disorders.

IV. WHITE MATTER

Unlike the nonlinear regionally specific changes in gray matter during adolescence, white matter volume increases linearly with age in all lobes. The net increase across ages 4–22 is approximately 12% (Fig. 2). In addition to volumetric changes in white matter, the density of white matter increases particularly in language-relevant left frontotemporal pathways.

V. CORPUS CALLOSUM

The linear increases in lobar white matter parallel increases for the major white matter tract in the brain, the corpus callosum. The corpus callosum is a bundle of approximately 200 million nerve fibers connecting the left and right hemispheres of the brain. Most of the fibers are myelinated and most connect homologous areas of the cortex. The corpus callosum integrates the activities of the left and right cerebral hemispheres. It combines information from bilateral sensory fields, facilitates memory storage and retrieval, allocates attention and arousal, and enhances language and auditory functions. Efficiency of interhemispheric integration has been linked to creativity and intelligence and becomes more crucial as task difficulty increases. Capacities for these functions improve during childhood and adolescence, making morphologic changes in the corpus callosum during these ages intriguing.

The myelinated fibers of the corpus callosum make it particularly easy to see on midsagittal MR images, which along with its clinical interest have made it a common target of investigations. The corpus callosum is arranged in an approximately topographic manner, with anterior segments containing fibers from the anterior cortical regions, middle segments containing fibers from the middle cortical regions, and so on. Progression of corpus callosum development continues throughout adolescence, and corpus callosum anomalies have been reported in several childhood disorders. Sexual dimorphism of the corpus callosum remains a controversial topic, with some reports indicating sex differences and many others no differences.

VI. VENTRICLES

The ventricles are the cerebrovascular fluid-filled cavities of the brain. Lateral ventricular volume increases about 30% between the ages of 4 and 20, with the greatest increase occurring before adolescence. There is not a significant gender difference in the shapes of the developmental curves.

VII. CEREBELLUM

Although the cerebellum has historically been viewed as a brain region primarily involved in the coordination of movement, an increasing number of neuropsychological, neurophysiological, and neuroimaging studies show that the cerebellum plays a role in higher cognitive function, language, and emotional control. The development of the cerebellum in childhood and adolescence demonstrates two unique characteristics compared to other structures quantified by MRI. First, it is the most sexually dimorphic, being robustly larger in males (Fig. 3). Second, cerebellar size is the least heritable. That is, its size is similar in monozygotic versus dizygotic twins. Also, it is relatively late maturing. These features imply that the cerebellum is particularly sensitive to environmental influences during critical stages of development.

VIII. CONCLUSIONS

Adolescence is a tumultuous time in the development of the brain. Although scientists are now able to observe and quantify anatomic changes, much work remains to be done in interpreting the changes and understanding the forces that shape and guide them. One hypothesis for what influences the adolescence gray matter pruning is the "use-it-or-lose-it" principle. According to this principle, those cells and connections that are used will survive and flourish, whereas those cells and connections that are not used will wither and die. If this theory is correct, then the activities of the teen would have a powerful influence on the ultimate "hardwiring" of the adult brain.

The sexual dimorphism of the developing brain is of particular interest in child psychiatry, in which nearly all of the disorders have different ages of onset, symptomatology, and prevalences between boys and girls. The extent to which the sex differences in healthy brain development interact with other genetic and environmental effects to account for some of these clinical differences between boys and girls is an active focus of study.

Future directions of adolescent brain imaging are likely to include studies of identical and nonidentical twins, longitudinal studies, and the use of new tools such as functional MRI and diffusion tensor imaging.

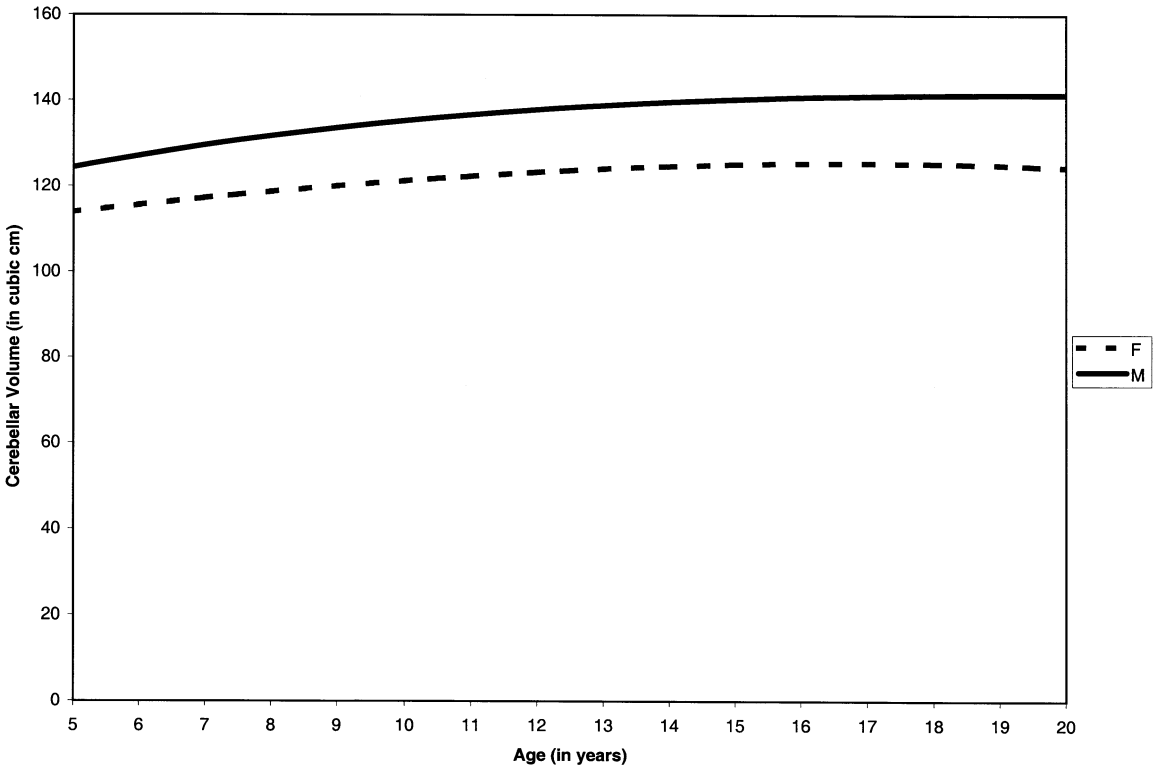


Figure 3 Cerebellar growth curves based on 145 individuals (243 scans).

Understanding the nature and influences of brain changes during childhood and adolescence may help parents, teachers, society, and teens take steps to optimize the development of the brain.

See Also the Following Articles

AGING BRAIN • BRAIN DEVELOPMENT • CORPUS CALLOSUM • NEUROPSYCHOLOGICAL ASSESSMENT, PEDIATRIC • READING DISORDERS, DEVELOPMENTAL • VISUAL SYSTEM DEVELOPMENT AND NEURAL ACTIVITY

Suggested Reading

Dawson, G., and Fischer, K. W. (Eds.) (1994). *Human Behavior and the Developing Brain*. The Guilford, New York.

Jacobson, M. (1991). *Developmental Neurobiology*, 3rd ed. Plenum, New York.

Johnson, M. (Ed.) (1993). *Brain Development and Cognition: A Reader*. Blackwell, Oxford.

Mitchel, G. F., and Moore, C. L. (1995). *Developmental Psychobiology: An Interdisciplinary Science*. MIT Press, Cambridge, MA.

Ratey, J. J. (2001). *A User's Guide to the Brain*. Pantheon, New York.

Rutter, M., and Rutter, M. (1993). *Developing Minds*. Basic Books, New York.

Aggression

JEFFREY L. SAVER
University of California, Los Angeles

GLOSSARY

amygdala Almond-shaped neuronal structure located in the inferomesial temporal lobe, anterior to the hippocampus, a component of the limbic system.

hypothalamus Neuronal structure constituting the inferiormost component of the diencephalon, connected with the pituitary gland via the infundibular stalk, and intimately concerned with visceral, autonomic, and endocrine function.

impulsivity Prone to act rashly; unable to resist urges.

prefrontal cortex Cortical sectors in the frontal lobe anterior to the motor strip.

sham rage A state in which an animal deprived of all cortical, or bilateral amygdaloid, input reacts to stimuli with behavioral displays of extreme anger.

violence The forceful infliction of physical injury on another individual or object.

Aggression is physical or verbal behavior that is threatening, destructive, or intended to harm. Aggression is a fundamental component of social behavior, essential to drive fulfillment and survival. Aggressive behavior in humans may be broadly divided into at least two phenomenological forms, an irritable–impulsive–affective subtype and a controlled-instrumental predatory-instrumental subtype. Each subtype is likely controlled by distinct, albeit overlapping, neuroanatomic and neurochemical substrates.

I. NATURE AND NURTURE IN AGGRESSIVE BEHAVIOR

Attempts to understand the brain basis of human aggression have often been thwarted by the outmoded dichotomies of nature versus nurture. Cultural and social influences are of critical importance in the genesis of human aggressive behavior. Violence may be motivated by the highest political and religious ideals and promoted and reinforced by cultural and governmental authorities. However, every violent behavior, whether springing from the most elevated or the basest of impulses, requires a neurobiological substrate to orchestrate the complex array of perceptual, motor, and autonomic components of acts that constitute aggressive conduct.

Recent developmental studies demonstrate that nature and nurture are closely intertwined in the genesis of aggressive behaviors. Early life experiences strongly shape an adult's regulation of aggressive

conduct. Large-scale cohort studies demonstrate that key environmental variables promoting the development of repeatedly violent individuals include rearing in disordered households, physical abuse in childhood, sexual abuse in childhood, and social deprivation. However, these environmental influences interact with an individual's biological vulnerabilities. Twin and adoption studies suggest that genetic factors control half the variation in propensity to oppositional temperament (angry, vindictive, and resistant to control), lack of empathy (callousness), and impulsivity, and each of these traits increases the risk of chronically violent behavior. Longitudinal cohort studies may be summarized as demonstrating that most violent children do not become violent adults, but that most violent adults were violent children. These observations indicate that although biological vulnerabilities are critical to the genesis of adult aggressive behavior, social interventions early in life can often avert the development of chronically violent behavior in adulthood. Recent studies tie nature and nurture together even more directly, demonstrating in nonhuman primates and in humans that early life social deprivation and adverse events can alter brain seroternergic systems, reshaping the neural systems for regulation of hostile behaviors.

Patients with acquired brain lesions represent a distinctive, and informative, population. When focal injuries disrupt the neural networks that regulate aggression, hostile behaviors may appear that have no relevant developmental or environmental precipitant or only minimal social provocation. More often, damage to neuronal systems controlling assertive behavior leads not to random acts of overt aggression but to alterations in temperament and inappropriate choices of targets and settings for aggressive behavior. Studies correlating loci of focal injury and patterns of altered aggressivity afford unique insights into the neural architecture responsible for implementing and inhibiting hostile behavior.

II. AN INTEGRATIVE NEUROSCIENCE APPROACH TO AGGRESSION

An overly simplified concept of "organicity" has been an additional obstacle to past attempts to understand the brain substrates of human aggression. Explicitly or implicitly, some investigators have suggested that neurologic insults produce a unitary "organic aggression syndrome," posessing stable, invariant behavioral features independent of lesion location or lesion type. In most formulations, this stereotypic organic aggression syndrome is postulated to lower a general threshold for aggression or to result in episodic dyscontrol. This view is imprecise, obscuring fundamental evolutionary, neurochemical, neurophysiologic, and behavioral distinctions among discrete aggression-related neural circuits in the brain stem, diencephalon, limbic system, and neocortex. This article instead advances an integrative neuroscience perspective, drawing on converging sources of evidence from evolutionary studies, ethology, neurophysiology, pharmacology, and clinical neuroscience to delineate a multiregional, hierarchical model of the neural regulation of aggression.

III. EPIDEMIOLOGY

Aggression is common among neurologically normal individuals and in diverse populations of brain-injured individuals. Homicide is the second most common cause of death among young adults and the 12th leading cause of death at any age in the United States. One population-based survey suggested that 3% of individuals commit violent acts annually, with a lifetime prevalence of 24%. However, other studies suggest that 94% of parents in the United States physically punish their toddlers, 11–17% of husbands physically hit their wives, and 11% of wives physically hit their husbands. Up to 40% of individuals admitted to psychiatric wards of general hospitals are violent immediately before admission, and up to 37% assault staff or patients during their hospitalization. Relatives of individuals with traumatic brain injury identify temper and irritability as major behavioral difficulties in 70% of patients. Imprisoned criminals evidence schizophrenia, major depression, and manic-depressive disorder at twice the rate of the general population.

IV. EVOLUTIONARY PERSPECTIVES

Aggression serves vital, evolutionarily adaptive functions. Along with the fundamental drives of fear, hunger, sexual desire, and social affiliation, adaptive aggression is present throughout the order Mammalia, triggered by environmentally appropriate and highly specific stimuli. Agonistic behavior to obtain food,

defend a territory, win a mate, or protect offspring is essential for the survival of the individual and for propagation of its genetic material.

Recent formulations of evolutionary theory suggest that competitive selection fosters the development of closely regulated and intertwined aggressive and affiliative behaviors. Unregulated, wanton aggression would rapidly reduce support among an organism's conspecifics and impair reproductive success. Conversely, uniformly submissive and avoidant behavior would prevent an organism from gaining access to critcial resources. Selection pressures on social behavior consequently tend to favor evolutionarily stable strategies with which an organism may variably express either aggressive or affiliative behaviors, depending on the state of several variables at the moment of a particular encounter with conspecifics, including its past interactions with that individual, position in dominance hierarchies, age, strength, and the general availability of environmental resources. In social primates the need for precise neural regulation of aggression is particularly advanced compared with that of species that lead a more solitary existence.

One domain in which the influence of natural selection may be clearly discerned is in the principles of outward display of aggressive and submissive signaling among conspecifics that recur in invertebrates, dogs, cats, primates, and humans. Emotional displays are highly stereotyped and opposing emotions are frequently conveyed by antagonistic postures, promoting accurate communication of an organism's emotional state. Darwin was the first to suggest that human facial expressions originated from postures associated with adaptive actions among mammalian forebears. He proposed that the sneer, an expression of anger, evolved from the baring of the canine teeth prior to a biting attack. Over time, fighting postures expressed in the body and facial musculature come to signal the threat, rather than solely the enactment, of attack. Inhibition of the neuronal assembly triggering an aggressive posture and activation of neurons enabling opposing postures conveys the opposite emotion of friendliness or submissiveness. The substantial role of subcortical neuromodulators, such as norepinephrine, acetylcholine, and serotonin, in modifying aggressive propensities in humans likely developed from their phyologenetically ancient function in promoting relevant peripheral skeletal and autonomic processes relevant to conspecific displays. From these origins a complex affective communication system developed, hardwired at its lowest neuronal level of implementation, regulated by higher centers that moderate the timing and intensity of display.

The complexity of neural regulation of aggression in mammals is driven in part by the existence of several distinct subtypes of aggressive behavior. For each subtype, aggressive acts are triggered, targeted, promptly terminated, and specifically inhibited by distinct classes of environmental stimuli. Ethologists have advanced several typologies of aggressive behavior, with each class demonstrating a specific outward display and set of determining stimuli. Moyer's widely recognized classification scheme divides hostile behavior into predatory, territorial, intermale, maternal, defensive, fear-induced, irritable, and instrumental subtypes (Table I). In various animal models, neuronal recording and lesion studies have identified important loci participating in neural networks controlling these discrete assertive behaviors (Table II). Evidence for broadly similar clustering of aggressive behaviors into an impulsive–reactive–hostile affective subtype and a controlled–proactive–instrumental–predatory subtype has been reported among violent children, psychiatric patients, and perpetrators of murder.

Table I
Behavioral Classification of Aggression

Type	Eliciting stimulus	Form
Predatory	Natural prey	Efficient, little affective display
Territorial	Boundary crossing	—
Intermale	Conspecific male	Ritualized responses
Fear induced	Threat	Autonomic reactions, defensive behaviors
Maternal	Distress calls, threat to offspring	Initial attempts to avoid conflict
Irritable	Frustration, deprivation, pain	Hyperactivity, affective display
Instrumental	—	—

Table II
Neuroanatomic Correlates of Aggressive Behavior Subtypes in Experimental Animals[a]

Triggers	Suppressors
Predatory offensive aggression	
Anterior hypothalamus	Prefrontal cortex
Lateral hypothalamus	Ventromedial hypothalamus
Lateral preoptic nuclei	Basolateral amygdala
Ventral midbrain tegmentum	Mammillary bodies
Ventral midbrain	
Ventromedial periaqueductal gray matter	
Intermale (competitive) aggression	
Laterobasal septal nuclei	Dorsolateral frontal lobe
Centromedial amygdala	Olfactory bulbs
Ventrolateral posterior thalamus	Dorsomedial septal nuclei
Stria terminalis	Head of caudate
Fear-induced aggression	
Centromedia amygdala	Ventromedial hypothalamus
Fimbria fornix	Septal nuclei
Stria terminalis	Basolateral amygdala
Ventrobasal thalamus	Ventral hippocampus
Maternal–protective aggression	
Hypothalamus	Septal nuclei
Ventral hippocampus	Basolateral amygdala
Anterior hypothalamus	Frontal lobes
Ventromedial hypothalamus	Prefrontal cortex
Dorsomedial hypothalamus	Medial prepiriform cortex
Posterior hypothalamus	Ventromedial hypothalamus
Anterior cingulate gyrus	Head of caudate
Thalamic center median	Dorsomedian nucleus of thalamus
Ventrobasal thalamus	Stria terminalis
Ventral hippocampus	Dorsal hippocampus
Ventral midbrain tegmentum	Posterior cingulate gyrus
Ventromedial periaqueductal gray matter	Periamygdaloid cortex
Cerebellar fastigium	
Sex-related aggression	
Medial hypothalamus	Septal nuclei
Fimbria fornix (male)	Fimbria fornix (female)
Ventral hippocampus	Cingulate gyrus
	Dorsolateral amygdala

[a]Reproduced with permission from D. M. Treiman (1991), Psychobiology of ictal aggression. *Adv. Neurol.* **55,** 343.

Whereas in simple organisms neurohormonal mediation of aggressive–submissive postures may suffice for regulating hostile behavior, in social mammals and especially primates, the need for flexible and precise control of aggressive and other emotional behaviors has propelled the evolution of hierarchic levels of intermediate and higher neural circuitry. In general, nervous system evolution, like somatic evolution, proceeds not by replacing existing structures but by modification and addition of subtler levels of control over older structures—evolution by tinkering. The human brain has developed through a progressive elaboration of neural elements surrounding the brain stem core found in less complex organisms. In the regulation of emotion, more recently evolved limbic, paralimbic, and neocortical components of the nervous system have established anatomic and physiologic controls over brain stem and primary cortical structures that implement autonomic, endocrine, and motoric somatic states.

Discrete structures controlling drives exist at each level of the neuraxis, mediating between sensory and motor systems. Relatively simple protoreptilian nervous system regulatory mechanisms persist in the human brain at the level of the brain stem and hypothalamus. Limbic structures provide a second critical level of control in the mammalian brain, each projecting directly to the hypothalamus. The frontal neocortex constitutes a third level, greatly expanded in higher primates, that modulates both limbic and hypothalamic output. These levels are functionally distinctive, with characteristic inputs (afferent circuitry), outputs (efferent circuitry), and functional organization, providing varying representations of the internal milieu and external environment (Fig. 1).

Neurons controlling such basic drives as feeding and reproduction are closely associated at each anatomical site. Circuitry regulating aggression, often an instrumental response in the service of these drives, is localized in adjacent regions. Because of this anatomical juxtaposition, dysregulation of aggression caused by brain injury is frequently accompanied by simultaneous abnormalities in feeding and sexual behavior.

V. BRAIN STEM REGULATION OF AGGRESSION

Pontine and mesencephalic nuclei coordinate full-fledged aggressive behaviors in rodents but only fragments of aggressive behavior in primates. Inputs

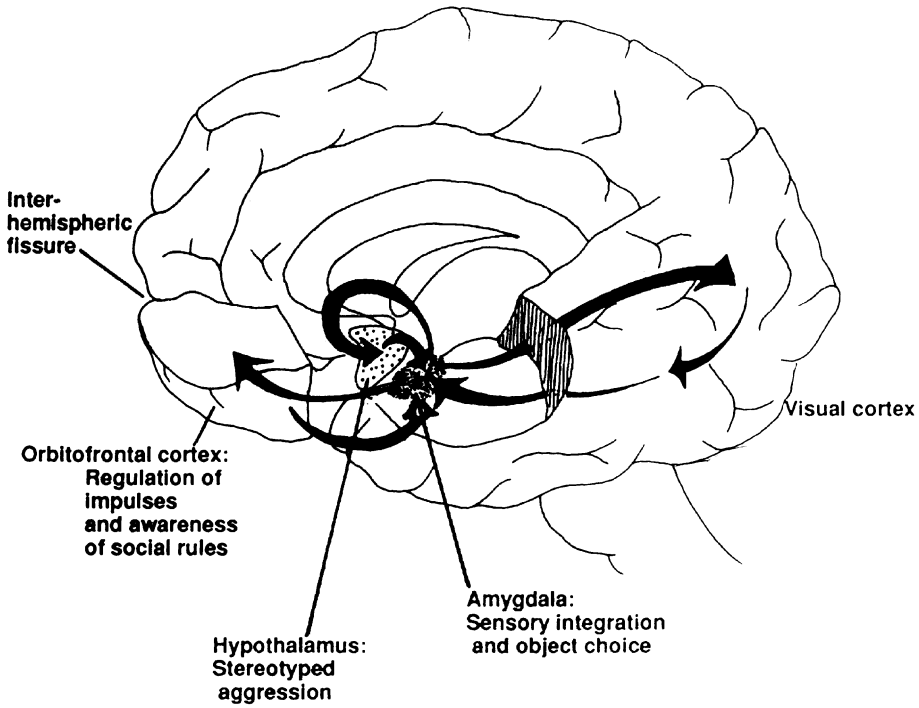


Figure 1 Critical sites in the hierarchical, multiregional neural system regulating aggression. Portions of the left temporal and orbitofrontal regions are cut away in this midsagittal view of the human brain to optimize visualization. [Reproduced with permission from Saver *et al.*, 1996, Neuropsychiatry of aggression. In: *Neuropsychiatry: A Comprehensive Textbook* (Fogel, B. S., Schiffer, R. B., Rao, S. M., Eds.) p. 532. Williams & Wilkins, Baltimore].

to the system include spinoreticular proprioceptive and nociceptive sensory circuits. Outputs incorporate pontine (facial) and descending motor centers, leading to stereotypic movements. Electrical stimulation of upper brain stem nuclei in subhuman primates produces different fragments of aggressive facial expressions and vocalizations, dissociated from offensive or defensive behavior. Other brain stem circuits contribute to somatomotor patterns for gestures and approach and avoidance behaviors. Medullary sympathetic and parasympathetic nuclei exert direct autonomic effects on cardiovascular, respiratory, and gastrointestinal peripheral organ systems. In humans, however, full-fledged aggression-related behavior patterns are not produced at the brain stem level. Response coordination and decision making are carried out at higher processing stations. As a result, brain stem lesions, although sometimes disturbing fragments of aggressive behavioral output, generally do not produce major syndromes of altered aggressivity.

However, an important clinical aggression syndrome does arise from disruption of brain stem regulation of sleep–wake states—the rapid eye movement (REM) sleep behavior disorder. In normal individuals, neurons in the vicinity of the locus ceruleus actively inhibit spinal motorneurons during REM sleep, preventing the expression of motor programs being actively generated in the motor cortex. In the cat, bilateral pontine tegmental lesions compromise REM sleep muscle atonia, permitting the enactment of oneiric behaviors, including frequent biting and other attack conduct. In humans, focal or degenerative lesions in the pontine tegmentum similarly produce the REM sleep behavior disorder, in which affected individuals physically perform their dream movements. Although normal dreaming subjects report an extremely diverse array of movements and activities fictively performed in dreams, violent actions comprise the overwhelming preponderance of behaviors actually enacted by individuals with REM sleep behavior disorder. Most commonly, middle-aged men experience a violent dream in which they are attacked by animals or unfamiliar people. In response, the dreamer engages in vigorous, coordinated motor acts that are often violent in nature. Individuals may

jump off the bed, smash furnishings, or attack their bed partner, with frequent resulting injury to both patient and spouse. It is likely that aggressive responses generated in the forebrain during REM dreaming produce more powerful descending motor outputs than feeding, sexual, or other drive-related activities and disproportionately override the partial residual muscle atonia in these patients.

VI. HYPOTHALAMIC REGULATION OF AGGRESSION

The human hypothalamus receives inputs conveying information regarding the internal state of the organism through chemoceptors, osmoceptors, and viscerosensory cranial nerves. In contrast to these rich sources of interoceptive data, the hypothalamus does not directly receive sensory input regarding the external world from primary or association sensory cortices. Important outputs of the hypothalamus are to the pituitary gland through releasing factors and directly transported peptides, to the autonomic nervous system, for which it serves as "head ganglion," and to midbrain and spinal motor centers that coordinate stereotypic movements.

In controlling biological drives, the hypothalamus employs a functional decision-making strategy of hardwired antagonism of excitatory and inhibitory nuclei. Algebraic, neurophysiologic comparison of biochemically coded inputs leads to either graded and homeostatic or threshold, stereotypic all-or-none responses. These characteristics are illustrated by the reciprocally active lateral and medial hypothalamic centers controlling hunger and satiety. In rodents, stimulation of the lateral hypothalmic area initiates feeding, and ablation may lead to starvation; stimulation of the ventromedial hypothalamus terminates eating, and lesions produce obesity. These hardwired systems generate predictable responses independent of the animal's experience.

Numerous studies in animals suggest that the priniciple of threshold elicitation of stereotypic response similarly characterizes hypothalamic control of aggression. In cats, when neural structures rostral to the hypothalamus are destroyed, the decorticate animals periodically produce rage displays with little or no provocation, exhibiting hissing, piloerection, pupil dilation, and extension of claws. Direct electrical stimulation of the posterior lateral hypothalamus reliably elicits this "sham rage" in animals that have undergone cortical ablation. In the intact feline and rodent brain, stimulation of the posterior lateral hypothalamus shortens the latency for species-specific predatory attack. Instillation of acetycholine or cholinomimitic drugs in the lateral hypothalamus similarly facilitates attack behavior, promoting biting attacks of a cat on a rodent, or a rat on a mouse or frog, even by previously docile animals. Injection of cholinergic antagonists will eliminate biting attacks, even among usually aggressive felines and rodents. Conversely, ventromedial hypothalamic stimulation inhibits rather than facilitates aggressive behaviors and may promote assumption of a defensive posture. Bilateral ablation of the ventromedial nucleus increases the overall level of aggression in operated cats, including unprovoked attacks by previously friendly animals on their caretakers.

Clinical observations in humans suggest a broadly similar role for the hypothalamus in human aggression. Neoplasms that destroy the ventromedial hypothalamic area bilaterally are associated with attacks on caregivers reminiscent of animal aggression following ventromedial lesions. In the classic report of Reeves and Plum, a 20-year-old woman developed bulimia and obesity, amenorrhea, diabetes insipidus, and profound behavioral change. Over a 2-year period, she displayed outbursts of aggression characterized by indiscriminately scratching, hitting, or biting examiners who approached. She denied experiencing angry or vindictive internal feelings toward these individuals and expressed surprise and regret regarding her attacks. The outbursts tended to occur more frequently when she had not eaten for several hours, suggesting the emergence of predatory-like aggression. Postmortem examination revealed a hamartoma destroying the ventromedial hypothalamus. In another case report, a patient with bilateral hypothalamic lesions exhibited aggressive outbursts that appeared to be influenced by seasonal light levels, erupting more often in dark, winter months. These and other clinical cases suggest that in the human brain the hypothalamus plays an important role in the setting of a threshold for aggressive responses.

VII. AMYGDALA AND TEMPOROLIMBIC CORTEX REGULATION OF AGGRESSION

In contrast to the hypothalamus, the amygdaloid complex is reciprocally connected with multiple

cortical sensory systems capable of conveying highly processed information regarding the external world. Rich connections are established with a variety of both unimodal and polymodal sensory regions, such as the perirhinal cortex and the superior temporal sulcus, allowing convergence of information from visual, auditory, tactile, and gustatory cortices. The basolateral amygdala receives extensive projections from unimodal visual cortices in the inferotemporal cortex (such as area TE in primates) that are specialized for recognizing objects such as faces in central vision. Extensive intrinsic connections within the amygdala promote further coordination of sensory information.

Important outputs from the amygdala in primates are to the hypothalamus, through the stria terminalis and ventral amygdalofugal pathway; to brain stem centers controlling heart rate and respiration through the central nucleus projection pathway; and to the extrapyramidal motor system, especially the ventral striatum, also through the stria terminalis and ventral amygdalofugal pathway.

The amygdala appears to provide a critical link between sensory input processed in the cortical mantle to produce a model of external reality and hypothalamic and somatomotor centers evoking pain, fear, and other basic drive–related emotions. Many observations in animals and humans suggest that a fundamental function performed by the amygdaloid complex and related temporolimbic structures is linking perceived objects with appropriate emotional valences. The result is a qualitative steering of behavior rather than quantitative regulation of threshold. On the basis of prior experience, sensory–emotional associations direct consummatory behavior to appropriate targets in the external world.

The importance of the amygdaloid complex in the recall of the affective signficance of stimuli is demonstrated by the drive–object dysregulation of the Kluver–Bucy syndrome observed in animals when the amygdala (and often overlying temporal neocortex) are removed bilaterally. Monkeys with such lesions engage in a continuous olfactory and oral exploration of their environment in which each object evokes the same response of tasting and sniffing as though the monkeys had never encountered it. The animals fail to distinguish food from inedible objects and eat metal bolts and feces as readily as normal dietary items. Animals have difficulty distinguishing appropriate from inappropriate sexual partners; similarly, lesioned cats will attempt copulation with chickens or other animals. These results suggest that lesioned animals cannot identify particular objects as being appropriate or inappropriate to satisfy hypothalamic drives.

The effects of bilateral amygdalectomy on aggressive behavior are consistent with such a hypothesis. Amygdala removal results in taming and placidity in most animals. Objects that previously evoked signs of fear or provoked attack appear to lose their past associations. Monkeys no longer behave aggressively toward experimenters, becoming docile and easy to handle. Unilateral amygdalectomy with lesions of all commissural pathways produces taming when stimuli are presented to the operated hemisphere but appropriate hostile responses when the stimuli are displayed to the unoperated hemisphere. However, amygdalectomy in submissive monkeys has led to a maintained or increased level of aggression, consonant with the view that the fundamental effect of amygdalectomy on aggression is not a change in aggressive threshold but a modification of previously acquired patterns of linking stimuli with aggressive responses. Fundamentally appetitive drives, such as feeding and reproduction, are released onto inappropriate targets. An instrumental drive such as aggression is no longer elicited or suppressed according to past, learned responses of the animal.

In humans, extensive bilateral temporolimbic damage produces behavior that is similar to that of lesioned monkeys, frequently accompanied by amnesia, aphasia, and visual agnosia (Fig. 2). Patients may engage in indiscriminate oral and tactile exploration of their environment (hyperorality and hypermetamorphosis) and change their sexual preferences. Affected individuals exhibit a flattened affect and report diminished subjective emotional responses to stimuli. Aggressive behaviors become uncommon, and apathy with lack of either strongly positive or negative responses becomes the rule. In one series of 12 patients with acquired Kluver–Bucy syndrome, placidity was noted in all. Functional imaging and lesion studies in humans suggest that the amygdala plays a critical role in processing of perceived threat stimuli. Fearful responses to threatening faces and objects are diminished in individuals with lesions largely confined to the amygdala bilaterally. A group performing bilateral amygdalotomies in aggressive patients reported that among 481 cases, approximately 70% showed a reduction in either restlessness or destructiveness, and one-half of these remained placid even when purposefully provoked.

The first reported case of the Kluver–Bucy syndrome in humans illustrates the characteristic clinical picture. A 19-year-old man sequentially underwent left

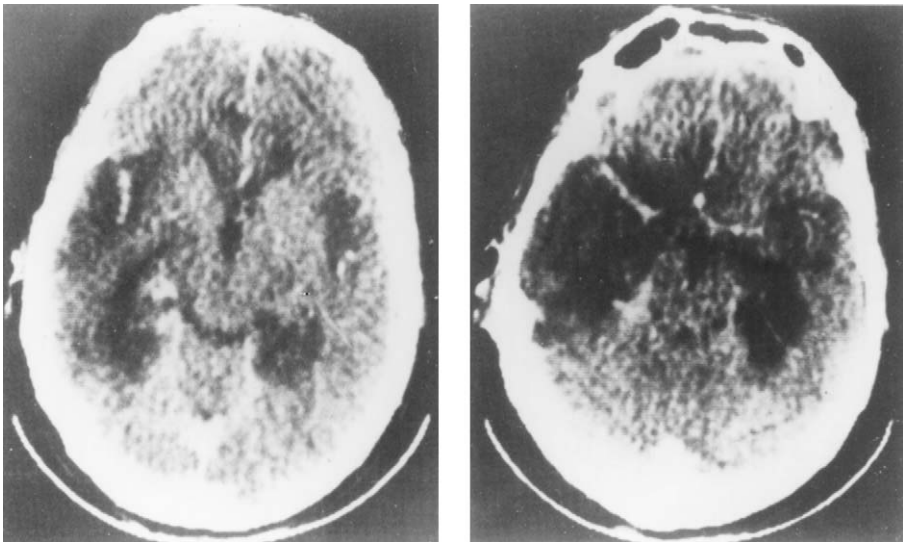


Figure 2 Bilateral temporolimbic lesions from herpes simplex encephalitis producing Kluver–Bucy syndrome including hypoaggression (passivation). This 42-year-old man was apathetic, indifferent, and impassive. In addition, he constantly manipulated and frequently mouthed objects, made sexual propositions to staff, especially men (prior to his illness he had been heterosexual), and exhibited severe anterograde amnesia and visual agnosia. CT axial images demonstrate large right temporal and smaller left mesial temporal hypodense lesions (reproduced with permission from S. Bakchine, F. Chain, and F. Lhermitte, 1986, *Rev. Neurol.* **142,** 126–132. © Masson Editeur).

and then right temporal lobectomies for treatment of a refractory seizure disorder accompanied by frequent outbursts of violent behavior. Following the second operation, he demonstrated dramatic behavioral changes, including compulsive manual manipulation of objects in the environment, insatiable appetite, sexual exhibitionism with frequent masturbation, severe retrograde and anterograde amnesia, and prosopagnosia. The reporting physicians were particularly surprised by a new placidity and the resolution of his previously aggressive behavior:

> *He no longer manifested the slightest rage reactions toward the nurses and doctors, upon whom, before the second operation, he used to rush as soon as they came into sight. The patient, on the contrary, now assumed an extremely childish and meek behavior with everyone and was absolutely resistant to any attempt to arouse aggressiveness and violent reactions in him.*

Related, modality-specific alterations in aggressive responding appear when bilateral lesions spare the amygdaloid complex but selectively interrupt pathways linking unimodal cortical sensory processing areas with the temporolimbic region. Stimuli presented solely within the sensory modality disconnected from the amygdala then fail to evoke learned associations, but stimuli that may be processed through other sensory channels elicit normal responses. One reported case of modality-specific limbic disconnection concerned a 39-year-old college graduate who suffered severe brain injury in a motorcycle accident. Computed tomography (CT) scans demonstrated bilateral cerebral hemorrhages in the inferior occipitotemporal region, interrupting visual input to polar and mesial temporolimbic structures. In addition to right hemiparesis, left hemidystonia, and prosopagnosia, he exhibited visual hypoemotionality—a diminished ability to react affectively to visual stimuli. A former assistant city planner, he was no longer moved by aesthetic differences between buildings. He ceased hiking because he now found natural scenery dull. He complained of total loss of emotional reaction to seeing attractive women in everyday encounters and to erotic visual stimuli. However, he maintained a strong interest in music, to which he listened almost constantly. He could be sexually aroused by verbal–auditory stimuli and derived pleasure from touching and being touched. This modality-specific limbic disconnection extended to fear and aggressive responses. In laboratory testing, when exposed to a series of slides, he rated as neutral and unemotional threatening images such as a gun and a snake, which normal controls scored as negative and highly arousing.

An intriguing contrast to the behavioral alterations that result from removal of the temporal lobes is provided by a far more common clinical condition, temporal lobe epilepsy, in which abnormal neuronal excitability develops within temporolimbic cell populations. Within the temporal lobe, the amygdaloid complex is particularly sensitive to the phenomenon of kindling, in which repeated stimulation of neurons leads to a progressive lowering of the threshold for discharge. Because many processing pathways converge on the amygdala, activity of epileptic foci throughout and beyond the temporal lobe can affect amygdalar excitability. The resulting enhancement of amygdaloid activity may, in a general sense, be the converse of the decreased activity underlying Kluver–Bucy syndrome.

In normal animals, individual amygdaloid neurons respond selectively to biologically significant food and social stimuli. Kindling may lead to long–term changes in limbic physiology that alter and enhance aggressive and other emotional responses to both drive–related and neutral stimuli. Rather than losing previously acquired associations between sensory stimuli and drives, some temporal lobe epilepsy patients appear to forge new, fortuitous associations. Rather than a lack of emotional response to stimuli, they exhibit deepened and generalized affective associations.

Interictal behavioral changes consistent with this model have been observed in a subset of patients with temporal lobe epilepsy. These individuals exhibit a cluster of interictal behaviors that have been labeled the Geschwind–Gastaut syndrome, encompassing deepened emotions, a sensitivity to moral issues, often with religious and philosophical preoccupations, and hypergraphia—a tendency to write about these subjects at great length. As a consequence of strongly felt emotions, these individuals may become highly sensitive to slights or violations of principle and experience intense anger. These patients' strong moral and philosophical beliefs often preclude violent acts. However, if they do act aggressively, their behavior typically is performed in clear consciousness and often followed by sincere regret.

In an illustrative case, a 40-year-old man developed complex partial seizures in his 20s, characterized by fear followed by flushing, tachycardia, and loss of consciousness. He had suffered febrile seizures in childhood. Electroencephalography (EEG) showed bilateral temporal discharges, and pneumoencephalography demonstrated a dilated temporal horn of the left lateral ventricle. His interictal behavior was remarkable for extreme seriousness with virtually no sense of humor and a sensitivity to infractions of minor military procedures. When fellow servicemen lightheartedly violated minor rules, he would attempt to reason with them. However, he became incensed by their failure to appreciate his concerns, and brawls often ensued. The patient was enraged when sentenced to a military stockade for 1 week, especially because his elaborate ethical justification for his actions was not taken seriously. To indicate his anger, he destroyed plumbing fixtures in his cell and subsequently threatened to kill the magistrate whom he believed had treated him unfairly.

Following release and neuropsychiatric treatment, his temper became better controlled as he developed strong religious and philosophical convictions that prohibited violence. Nonetheless, several years after overt violent behavior had ceased, he told an examiner, "I have more of a problem with anger than anybody I have ever met in my life." He described a constant internal tension between feelings of being treated unjustly and a sincere desire not to harm another individual. Other aspects of his behavior consistent with the interictal behavior syndrome included evangelical religiosity, extensive and detailed writing, and inappropriately prolonged encounters with fellow patients and caretakers (enhanced social cohesion/viscosity).

VIII. PREFRONTAL CORTICAL REGULATION OF AGGRESSION

The dorsolateral prefrontal cortex receives extensive afferents from multiple posterior neocortical association areas, including dense connections with the inferior parietal lobule, a region responsible for surveying extrapersonal space for relevant stimuli. The orbitofrontal cortex is reciprocally connected to the rest of the neocortex principally via the dorsolateral convexity of the frontal lobe. Projections from the hypothalamus through the dorsal medial nucleus of the thalamus and from the rostral temporal lobe through the uncinate fasciculus inform the frontal lobes of both internal (hypothalamus) and external (neocortical association to temporal lobe) stimuli of affective significance.

The prefrontal cortex has direct outputs to the pyramidal motor system, the neostriatum, temporal neocortex, and the hypothalamus. Schematically, prefrontal cortices appear to integrate a current account of the outside world, the state of the internal

milieu, and the recognition of drive-relevant objects with knowledge of learned social rules and previous experiences relating to reward and punishment. The prefrontal cortex may play a particularly important role in both working memory and social modeling, maintaining an abstract representation of the world that allows anticipation of the effects of one's actions on other individuals and the likely consequences of such actions. The prefrontal cortices construct a behavioral plan that is consistent with experience and especially the rules of socialization in order to optimize the satisfaction of biological drives.

The simplest summary of these complex functions in humans is judgment, which should not be simply equated with purely rational cost–benefit calculations that may be quite time-consuming and biologically uneconomical. Rather, it has been proposed that, in selecting among alternative response options, prefrontal cortices are guided by internal, somatic state markers—physiological cues that allow rapid choice of previously rewarded, effective options. Damage to the dorsal convexity in humans results in a diminution of long-term planning and a state of apathy or indifference. On formal neuropsychological testing, shifting of response set and the ability to apply strategy to problem solving are impaired. In contrast, damage to the orbital undersurface of the frontal lobe has classically been described as resulting in superficial, reflexive emotional responses to stimuli in the immediate environment. Patients are impulsive, acting without foresight or consideration of the remote consequences of their actions. Orbital frontal lesions thus lead to episodes of transient irritability. Often, a patient strikes out quickly after a trivial provocation, with little consideration of social prohibitions limiting aggressive behavior or untoward future consequences. The targets of aggression are categorically appropriate, but patients are unable to apply abstract rules that would override the immediate environmental provocation.

Numerous case reports illustrate the tendency of patients with orbitomedial frontal injuries to act impulsively, without regard to long-term consequences or sustained courses of action. Most well-known is the paradigmatic case of Phineas Gage, a railroad worker who suffered an injury primarily to the left orbitomedial frontal lobe when an explosion projected a tamping rod through his skull. Subsequently, in the words of his physician, "he was no longer Gage." Previously a temperate, hardworking individual, he became "disrespectful," "irreverent," and "profane." He rejected all attempts to restrain him from satisfying desires of the moment but rapidly abandoned plans he made to achieve these desires.

In addition to detailed single case studies, case series of murderers have demonstrated a high incidence of frontal structural abnormalities on CT and magnetic resonance imaging (MRI), frontal hypofunction on position emission tomography (PET), and abnormal neuropsychologic performance on frontal systems tasks. Frontal ventromedial lesion location independently predicted aggression and violence among 271 America veterans who suffered penetrating head injuries in the Vietnam war.

Convergent evidence for a critical role of the orbitofrontal cortex in evoking internal somatic markers to regulating aggression is supported by studies of neurologically normal individuals with antisocial personality disorder and high psychopathy scale scores, who demonstrate diminished autonomic reactions to negative stimuli, decreased orbitofrontal activity in some functional imaging studies, and reduced prefrontal gray matter volumes on volumetric MRI studies.

IX. COMBINED LESIONS OF TEMPORAL AND FRONTAL LOBES

Some pathological processes tend to damage simultaneously multiple levels of neural circuitry critical to the regulation of aggression. The orbitofrontal surface and rostral temporal poles are particularly susceptible in closed head injuries, and conjoint lesions in the same patient are not uncommon. Temporolimbic epilepsy is a frequent sequel of both closed and open head injury.

Aggressive behavioral syndromes may result that have features associated with dysfunction of several brain regions. For example, as a result of brain trauma, a patient may, develop a temporolimbic epileptic focus as well as a contusion of the orbital frontal cortex. Such a patient can display the deepened emotions and anger associated with the interictal behavior syndrome as well as a failure to inhibit or modulate hostile responses typical of a frontal lesion. In one reported case, a young man suffered severe brain injury in a motor vehicle accident. Imaging studies demonstrated enlargement of the frontal and temporal horns of both lateral ventricles and EEG abnormalities were recorded over the right frontotemporal area. The patient displayed intermittent apathy suggestive of frontal lobe damage but also developed personality changes characteristic of the interictal behavior syndrome of

temporal lobe epilepsy. He had outbursts of extremely violent behavior and eventually attempted to murder his parents and former girlfriend. When questioned regarding his aggression, the patient failed to appreciate that his behavior might be distressing to others. PET studies provide further evidence for frontal and temporolimbic dysfunction in psychiatric patients and in violent criminals. In one study of eight repeatedly violent psychiatric patients, low metabolic rates were noted in prefrontal and medial temporal cortices. Similarly, in a study of 41 murderers, reduced glucose metabolism was noted in the prefrontal cortex, superior parietal gyrus, left angular gyrus, and corpus callosum, and reduced left hemisphere activity compared with right hemisphere activity was noted in the amygdala, thalamus, and medial temporal lobe. These differences were particularly pronounced among individuals classified as having performed affective, impulsive violence compared with planned, predatory violence.

Figure 3 Association of left hemisphere lesions with aggression. Template mapping of CT scan lesions in 18 consecutive stroke patients with violent aggressive outbursts demonstrates a preponderance of left hemisphere lesions, especially left frontal (reproduced with permission from S. Paradiso, R. G. Robinson, and S. Arndt, 1996, *J. Nerv. Mental Dis.* **184,** 750).

X. HEMISPHERIC ASYMMETRIES IN THE REGULATION OF AGGRESSION

Several lines of neuropsychological research suggest differences in left and right hemisphere specialization for the processing of emotion, including anger and aggression. The left hemisphere plays a greater role in decoding linguistically conveyed emotional information, and the right hemisphere is more important in processing nonverbal emotional cues, such as prosody and facial expression of emotion. Moreover, the right hemisphere may be more highly specialized for mediating emotional responses in general and negative emotional responses such as fear and anger in particular. These conclusions are supported by studies of functional asymmetry for verbal versus nonverbal expression of affect, verbal versus nonverbal decoding of affects expressed by others, and asymmetric facial expression of affect in patients with unilateral stroke, other asymmetric neurological injuries, and transient hemisphere inactivation during Wada testing and also in normal experimental subjects.

Studies in focal lesion patients suggest an important role of hemispheric specialization in the genesis of hostile behaviors. A case control study in focal stroke patients identified lesion location in the anterior, left hemisphere as a predictor of aggressive behavior, independent of the presence of depression (Fig. 3). Among 50 patients with temporal lobe epilepsy, those exhibiting intermittent explosive disorder were more likely to have left amygdala and left periamygdala lesions due to encephalitis or other structural insults.

Neuropsychological and psychophysiologic studies in unselected populations of violent criminal offenders also show frequent evidence of left hemispheric dysfunction. When neuropsychological deficits are observed in studies of violent groups, they tend to involve not only frontal-executive functions but also verbal comprehension, expressive speech, and other left hemisphere language functions. These findings are consonant with those of many studies in conduct-disordered and delinquent juveniles. In addition to lower average overall IQ, these individuals frequently have a disproportionately lowered verbal IQ (language, left hemisphere) compared to performance IQ (visuospatial, right hemisphere). Psychophysiologic studies employing computerized EEG spectral analysis indicate that persistently violent behavior among psychiatric inpatients is linked to increased 8-band

slow-wave activity in left frontotemporal derivations. Divided visual field, dichotic listening, skin conductance asymmetry, and lateral preference studies additionally suggest subtle abnormalities of left hemispheric function in sociopathic individuals without overt neurological lesions. In one small series of subjects examined by PET, decreased blood flow and metabolism were observed in the left temporal lobe of all four institutionalized, sexually violent offenders studied, and left frontal hypometabolism was observed in two of the four. In a SPECT study, left anterior temporal and bilateral dorsofrontal hypoperfusion were among the abnormalities distinguishing dementia patients with aggression from those without.

Based on the results of these and other laterality studies in violent individuals, different theorists have proposed that either left hemispheric or right hemispheric networks play a predominant functional role in the regulation of aggression, and that subtle developmental hemispheric abnormalities underlie repeatedly aggressive behavior in a proportion of "functional" psychopaths. A broader view is that each hemisphere performs complementary processing related to hostile behavior, and functional abnormalities of either hemisphere may produce disturbed aggressive responding through distinctive mechanisms. Left hemisphere dysfunction, implicated in a preponderance of studies, may lead to overt expression of negative affects mediated by the right hemisphere, diminished linguistic regulation over behavior, and adverse social encounters due to impaired verbal communication. Right hemisphere dysfunction may lead to improper intrahemispheric decoding and encoding of prosody, facial expressions, and other nonverbal emotional responses, overreliance on semantic processing, impaired self-awareness of physiological arousal, and a distinctive pattern of altered aggressivity.

Right hemisphere parietofrontal lesions may impair an individual's ability to interpret and produce emotional gestures, prosody, and affect and to monitor internal somatic states. This may produce an acquired "sociopathy," even in the absence of bifrontal lesions. Although the most frequently noted pattern of altered behavior after right parietal cortical lesions is one of inappropriate cheerfulness and denial of illness, irritability and aggressive outbursts may also ensue. A well-known case is that of former Supreme Court Justice William Douglas, who exhibited disturbed social judgment following a right hemisphere stroke. Although he retained sufficient linguistic and abstract reasoning ability to return to the bench, inappropriate social behaviors soon forced his retirement.

XI. NEUROCHEMISTRY OF AGGRESSION

During the past two decades, there has been an explosion of knowledge regarding neurotransmitter–receptor systems that modulate aggressive behavior in animals and humans. However, many studies have focused exclusively on the effects of a specific neurotransmitter or receptor subtype on one or more aspects of violent or aggressive behavior. Integration of such fine-scale neurochemical data with the large-scale neurocognitive network data acquired in neuroanatomical and neurophysiological studies has been limited. Most experimental and clinical reports on the "neurochemistry of aggression" likely describe the effects of neuromodulators at peripheral, brain stem, hypothalamic, and diffusely projecting hemispheric sites, which can reduce or raise the individual's overall predisposition to aggression. Still lacking are studies that evaluate the multisynaptic integration of parallel-processed streams of complex sensory and limbic information that link amygdala, orbitofrontal, and other higher cortical centers. It is quite likely that these networks are subserved by diverse messenger systems and not exclusively controlled by a single neurotransmitter. We also have insufficient understanding of the complex interactive effects that neurotransmitters and neurohormones exert on one another or the specific receptor subtypes mediating a particular response. However, because neurochemical studies provide the basis for pharmacological interventions in aggressive patients, data from such studies are clinically invaluable.

The hypothalamus, amygdala, and frontal lobe are richly innervated by monoaminergic neurotransmitters, acetylcholine, and neuropeptides. Neurotransmitter systems strongly linked to mediation of aggressive behaviors in animal and human studies include serotonin, acetylcholine, norepinephrine, dopamine, γ–aminobutyric acid (GABA), and testosterone and other androgens as well as nitric oxide, opioids, and glutamate. Serotonergic systems appear to be particularly important and have been the subject of intense experimental and clinical investigation.

A. Serotonin

Diverse animal and human studies suggest that serotonin is a critical modulator of aggressive behavior. Experimental work in animal models of aggressive conduct supports a critical role of serotoninergic

systems in hostile behavior. Eltoprazine, a 5-HT_1 agonist, reduces attack behavior in several species. In the rat brain, eltoprazine binding is greatest in the dorsal subiculum, substantia nigra, ventral pallidum, and globus pallidus. Other 5-HT_{1A} agonists, including buspirone and gepirone, reduce isolation-induced aggression in mice without causing sedation or incoordination. It has been proposed that serotonin increases the ability of an organism to arrange for and tolerate delay. Decreased serotonin leads to an increase of behaviors that are usually suppressed. Studies of isolation-induced aggression, shock-induced fighting, and muricidal and filicidal behavior have demonstrated an inverse relationship between serotonin activity and aggression in rats and mice. Recent studies have begun to delineate more complex species-specific and receptor-specific effects of serotonin on aggression. One recent rat study found that agonists at 5-HT_{1A}, 5-HT_{1B}, and 5-HT_2 receptors all reduced offensive aggression, but only 5-HT_2 agonists reduced defensive aggression.

Findings of studies of naturally behaving animal populations are consonant with the findings from experimentally induced aggression paradigms. Domesticated silver foxes, who easily tolerated human contact, had a higher level of midbrain and hypothalamic serotonin than wild silver foxes bred in captivity. The domesticated foxes also had a reduced density of 5-HT_{1A} receptor binding in the hypothalamus. Rhesus monkeys with the highest blood levels of serotonin were socially dominant, whereas animals with decreased whole blood serotonin tended to be ostracized. Aggressive behavior was also associated with high levels of cortisol, suggesting greater stress. Adolescent male rhesus macaque monkeys show an inverse correlation between cerebrospinal fluid (CSF) levels of the serotonin metabolite 5-hydroxyindoleacetic acid (5-HIAA) and risk taking and aggression in the wild.

Serotonin interacts with other neurotransmitter and neurohumoral systems in modulating impulsivity and aggression. For example, one group of investigators examined the effects of testosterone and serotonin administration on dominance and aggression in rats. Male rats given testosterone became dominant. Quipazine, a serotonin agonist, blocked aggression in both naturally dominant and testosterone–induced dominant rats. Nonspecific serotonin antagonists blocked aggression only in testosterone–induced dominant males. This study demonstrates pharmacoselectivity among different forms of aggression, a property that may be desirable for the development of pharmacologic treatments.

Data from human clinical studies are consistent with data from the experimental and observational animal literature, suggesting a critical role for serotonergic systems in human aggression. One major domain of investigation has focused on serotonergic markers in patients who have attempted or committed violent suicide. Caution must be exercised when interpreting these studies. The psychologic and biologic substrates of aggression against the self, manifest in suicide, are likely to differ from the underpinnings of aggression directed against others. Nonetheless, there are important behavioral affiliations between self-directed and outwardly directed violence and the two actions often cosegregate. Lifetime externally directed aggression is more frequent in suicide attempters than in others. Accordingly, findings in suicide patients offer important insight into general neurobiologic substrates of violent behavior.

Many postmortem studies have found decreased levels of serotonin and of presynaptic serotonin binding sites, such as the serotonin transporter site, in subcortical and neocortical brain regions in patients completing violent suicide. Recent studies suggest these abnormalities are more prominent in the orbital prefrontal cortex than in the dorsolateral prefrontal cortex. Similarly, postsynaptic serotonin receptors, such as 5-HT_{1A} and 5-HT_{2A} receptors, are generally elevated in the frontal cortex of violent suicides, especially in orbitofrontal sectors. These findings are consistent with a regulatory increase in postsynaptic serotonin receptors in response to decreased presynaptic serotonergic activity.

Cerebrospinal fluid levels of 5-HIAA are decreased in patients attempting suicide and in violent criminal offenders. A stronger correlation has been noted between CSF 5-HIAA and suicidal behavior than suicidal ideation alone, suggesting that low CSF 5-HIAA levels are a marker not simply of depression and suicidal risk but also of a tendency to aggressive and impulsive behavior. Several studies in criminal and interpersonally violent psychiatric populations support this view. Lowered CSF 5-HIAA levels were found in samples of impulsive arsonists and impulsive murderers. In a group of soldiers with behavior problems, CSF 5-HIAA was negatively correlated with aggressive behavior, and CSF 5-HIAA was reduced in a group of borderline patients with aggressive and suicidal behavior.

Neuroendocrine challenge studies have been utilized to probe serotonergic function in aggression. Serotonin administration in normals causes a release of prolactin. Suicidal depressed patients and patients

with personality disorders who exhibit impulsive and aggressive behavior have a blunted prolactin response to fenfluramine, a releaser of presynaptic serotonin stores, and to *m*-chlorophenylpiperazine, a 5-HT_2 agonist. This work suggests that suicidal and impulsive/aggressive patients have serotonergic hypoactivity. On the other hand, patients with depression and suicidality demonstrated an increased release of cortisol to 5-hydroxytryptamine, suggesting hypersensitivity of other postsynaptic serotonin receptors in this group. Other workers have not found a significant correlation between neuroendocrine challenge abnormalities and suicidality.

Additional support for the influence of central serotonergic activity on aggressive propensities comes from studies of normal volunteers administered selective serotonin reuptake inhibitors. The resulting increase in central serotonergic activity correlated with decreases in hostility and in negative affect generally and an increase in affiliative behaviors.

At least 14 different receptors for serotonin exist in the human brain. Recent investigations of the relationship between serotonin and aggression have begun to more precisely dissect serotonergic systems by employing molecular probes of specific serotonergic receptor subtypes. Buspirone, a 5-HT_{1A} agonist, produced a normal prolactin release when given intravenously to healthy male volunteers. This effect was blocked by the nonselective 5-HT receptor antagonist metergoline and by pindolol, a β-adrenergic and 5-HT_1 antagonist, in a dose-related fashion. Prolactin response to buspirone was inversely correlated with levels of "irritability" in patients with personality disorders, suggesting that decreased sensitivity of the 5-HT_{1A} receptor may be responsible for components of impulsive–aggressive behavior in patients with personality disorders.

Recently, genetic studies have provided further evidence of an important role of serotonin in aggression regulation. A polymorphism in the gene for tryptophan hydroxylase, the rate-limiting enzyme in the biosynthesis of serotonin, has been associated with suicidal behavior in impulsive alcoholic criminals and in individuals with major depression. Amino acid substitutions in the 5-HT_7 receptor gene have been reported among alcoholics with violent behavior.

B. Acetylcholine

Some of the earliest work on the neurochemistry of aggression focused on acetylcholine. Electrical stimulation of the lateral hypothalamus in rats leads to predatory attack on mice in animals that previously tolerated mice in their cage without attacking them. The attack terminates as soon as the electrical stimulation is discontinued. Applying carbachol, a cholinergic agonist, to the lateral hypothalamus provokes the sterotypic aggressive response, which can be blocked by atropine and facilitated by acetylcholinesterases. This cholinergic-induced predatory response is target specific—directed only at the animal's usual prey—and without affective display. Electrical stimulation of the lateral or dorsal amygdala facilitates predatory attack through its connections to the lateral hypothalamus. Applying carbachol to the amygdala also induces a predatory response. Aggressive behavior following human exposure to cholinesterase inhibitors has been observed in several clinical case reports. Despite well-documented early animal experimentation, cholinergic mediation of aggression and its clinical implications have been understudied in recent years. For example, the muscarinic receptor subtypes mediating hypothalamic aggression and cholinergic regulation of aggression in the frontal cortex have not been characterized in detail.

C. Norepinephrine and Dopamine

Catecholamine systems are associated with aggressive behavior in several animal models and clinical populations. Peripherally administered norepinephrine (NE) enhances shock induced fighting in rats. α_2 receptor agonists increase rat aggressive behavior, whereas clonidine decreases rodent aggressive behavior acutely. β-Adrenergic blocking decreases aggressive behavior in laboratory animals. Several human studies have found a correlation between increased CSF or frontal cortex NE or its metabolite 3-methoxy-4-hydroxyphenylglycol and aggressive behavior. It has been proposed that central noradrenergic tracts originating in the locus coeruleus innervate a behavioral inhibitory system that projects widely to the hippocampus, septal region, and frontal lobes. Modulatory disturbances in central norepinephrine would then lead to impulsivity and episodic violence. Long-term β-adrenergic blockade with agents such as propranolol is a well-established, effective therapy to reduce aggressive responding in diverse neuropsychiatric patient groups with violent behaviors.

L-Dopa can induce aggressive behavior in rodents and humans. Apomorphine, a potent dopamine agonist, can induce fighting in rats. Dopamine antagonists

tend to reduce aggression but usually at doses that also slow motor and cognitive performance. A few studies have shown reduced levels of a dopamine metabolite, homovanillic acid, in suicidal patients.

Recent genetic studies support an important role of catecholamine systems in human aggression. The genetic loci for MAO and catechol-*O*-methyltransferase (COMT), two enzymes critical in the catabolism of catecholamines, are located on the X chromosome. In a large human kindred, impulsive violent behavior and mental retardation among several males co-segregated with a point mutation in the MAO type A gene that produced enzyme deficiency and presumably increased central catecholaminergic activity. Male knockout mice lacking the MAO-A gene also show aggressive behavior. COMT gene alleles include a common polymorphism that produces three- or four-fold variation in enzyme activity. The allele coding for the less active form of the enzyme (and resulting increased central catecholaminergic activity) has been associated with violent behavior in two studies of schizophrenic and schizoaffective patients and in male knockout mice.

D. γ-Aminobutyric Acid

Several lines of evidence suggest that GABA inhibits aggression in animals and humans. GABA injected into the olfactory bulbs in rats inhibits mouse killing, whereas GABA antagonists can induce muricidal behavior. Benzodiazepines and other agents that facilitate GABA can decrease isolation-induced fighting in mice and attenuate aggression caused by limbic lesions. In humans, despite their tranquilizing and antiaggressive effect in the vast majority of patients, benzodiazepines can rarely lead to a transient increase in aggressive behavior (''paradoxical rage'').

E. Testosterone and Other Androgens

Testosterone is an important mediator of aggressive responding in diverse mammalian species. In rats, dominant males have higher levels of testosterone than submissive males. Cortisol increases in both groups, but cortisol is higher in the submissive group, suggesting a greater level of stress. In vervet monkeys, increases in serum and salivary testosterone levels correlated with the number of aggressive encounters. Moyer suggested that androgens increased intermale and irritable, but not predatory, sexual, fear-induced, and maternal forms of aggression. An interaction between androgens and other neuromodulators such as the monoamine neurotransmitters appears to govern aggressive responding. Testosterone-induced dominance in rats is reduced after treatment with 5-HT1A, -1B, and -2A/2C receptor agonists.

The association of androgens with aggression suggested by the simple observation that males enact aggressive behaviors more frequently than females in most mammalian species, including humans, is supported by observations of convergent hormonal–behavioral relationships in female spotted hyenas. The spotted hyena is one of the most aggressive animals in the wild. Spotted hyenas also have a very organized and highly nurturant clan society. Male and female hyenas are approximatley equal in size, and female genitalia are masculinized. The colonies are dominated by the females in a tightly ranked hierarchy. Females are more aggressive than males, and adult males are usually not able to feed from a kill while the dominant females are eating. The females' large body habitus, androgenous genitalia, and aggressive behavior are related to the high circulating levels of androstenedione. The role of androgens in mediating aggression thus transcends sexual lines in this species.

In humans, numerous studies support an important link between circulating androgens and aggressive behavior. Elevated testosterone levels in adolescent boys correlate with low frustration tolerance and impatience. Boys with increased testosterone are more likely to respond aggressively when provoked. Increased levels of free testosterone have been measured in the saliva of incarcerated violent criminals. Victorious collegiate wrestlers show a greater increase in their serum testosterone than do their defeated counterparts. Violent behaviors have been reported in individuals taking anabolic steroids for body-building programs. Male alcoholics who abused other people had higher levels of testosterone and lower levels of cortisol than those who did not. In a treatment study, inhibiting gonadal function with a GnRH antagonist reduced outward-directed aggression. A meta-analysis of reported studies demonstrated a strong positive correlation between testosterone levels and observer-rated aggressiveness.

It is important to note some common weaknesses in the data currently available on the neurochemistry of aggression. Most studies on neurotransmitter and neurohormonal effects on aggression have been conducted on male animals and men. Endocrine and neurochemical factors influencing aggression in

females have not been fully evaluated. Caution must be exercised when generalizing findings across species, particularly when comparing responses between humans, other primates, and other mammalian orders. Many aggression-related neurotransmitters are conserved across species, but there are likely important variations in receptor subsystems. Precision in defining and measuring aggression, impulsivity, and irritability in many animal models and humans is difficult to achieve. Also, many studies fail to recognize and fully clarify state–trait distinctions. Neuroendocrine challenge studies are subject to wide variability, depending on agent dosage, route of administration, and outcome measure.

Nonetheless, progress in basic investigations of the neurochemical and neuroendocrine mediators of aggression sets the stage for advances in the pharmacotherapeutics of violent disorders. Moreover, once a better understanding of the individual neurochemical and neuroendocrine factors contributing to aggression is attained, interactions among the multiple systems operating convergently and divergently at hierarchial sites in the neuraxis that regulate hostile behavior may be more fully explored.

XII. MAJOR CLINICAL SYNDROMES OF AGGRESSION

A. General Considerations

The recognition that specific neurological lesions may lead to violent behavior in human beings, and that abnormalities at different levels of the neuraxis result in distinctive types of aggressive behavior, provides a guiding schema for the evaluation and treatment of inappropriately aggressive (and inappropriately hypoaggressive) individuals (Tables III–V). In addition to emphasizing the need for careful neuropsychiatric evaluation of every violent patient, the hierarchical model for the regulation of aggression suggests important parameters that may help to characterize any aggressive act. Integrating information regarding the clinical manifestations of aggressive behavior, additional aspects of the history (particularly related to other drive-related behaviors), the neurological and psychiatric examinations, and structural and functional laboratory studies allows classification of individual patients among major syndromes of impulse dysregulation and aggression.

1. Hypothalamic/Brain Stem Syndromes

In patients with brain stem-mediated sleep–related disorders of violence, aggressive actions are generally nocturnal, associated with incomplete maintenance of REM or non-REM sleep states. The REM sleep behavior disorder predominantly occurs in middle-aged men. Violence most commonly occurs during a vivid and frightening dream, is often directed at a bed partner mistaken for a dream figure, and is unplanned, without the use of weapons. Affected patients are difficult to arouse from their dreaming state. Afterwards, they generally recall the dream material that provoked aggression but report believing they were attacking animal or human oneiric figures rather than their furniture or spouse. They exhibit remorse about their actions and, before a diagnosis is made, may self-treat their disorder by tying themselves in restraining devices at night.

Table III
Distinguishing Features of Focal Lesion Syndromes of Aggression in Humans

Syndrome	Provocation	Eliciting stimulus	Outbursts	Complex plans	Amnesia for acts	Remorse
Hypothalamic	Basic drive (e.g. hunger); unprovoked	Individuals who happen to be present	Yes	No	No	Yes
Ictal	None	Any near individual or inanimate object	Yes	No	Yes	Yes
Postictal	Attempts to restrain/protect patient	Caretakers	Yes	No	Yes	Yes
Interictal	Perception of moral injustice; threat, including misinterpretation of trivial stimulus	Individuals	Occasional	Yes	No	Yes, may be intense
Orbitofrontal	Minor provocation	Individuals	Yes	No	No	No

The REM sleep disorder must be distinguished from other parasomnias that may be injurious to patient and spouse, such as somnambulism, sleep drunkenness, and sleep terrors, which arise out of non-REM sleep. Nocturnal seizures must also be excluded. The evaluation in suspected cases includes a thorough history of sleep complaints from patient and bed partner, neurological and psychiatric examination, overnight polysomnographic study, and MRI. REM sleep behavior disorder has been associated with a variety of neurological conditions, including Parkinsonism, dementia, stroke, multiple sclerosis, and alcohol withdrawal. However, more than 50% of cases are idiopathic. Pontine tegmental lesions, which might be expected from animal studies, are rare, possibly because pontine injury frequently produces devastating motor and arousal deficits that preclude expression of the disorder. Approximately 90% of patients exhibit sustained improvement when treated with clonazopam.

Table IV
Selected Causes of Syndromes of Hyperaggressivity and Hypoaggressivity (Passivity)

Site/syndrome	Common etiologies
Hyperaggressivity	
Hypothalamic	Hamartoma
	Craniopharyngioma, astrocytoma
Temporal lobe epilepsy	Mesial temporal sclerosis
	Vascular malformation
	Glioma
	Traumatic brain injury
Orbitofrontal systems	Traumatic brain injury
	Anterior communicating artery aneurysm
	Orbital meningioma
	Huntington's disease
	Frontotemporal dementias
	Herpes simplex encephalitis
Combined frontotemporal	Traumatic brain injury
	Frontotemporal dementias
	Herpes simplex encephalitis
Multifocal or poorly localized	Attention deficit disorder
	Toxic–metabolic encephalopathies
	Vitamin B12 deficiency
	Alcohol, cocaine
	Multiple sclerosis
	Vascular dementia
Delusional cognition	Paranoid schizophrenia
	Late-life paraphrenia
	Endogenous depression
	Mania
	Alzheimer's disease
	Vascular dementia
Hypoaggressivity	
Bilateral amygalotemporal	Herpes simplex encephalitis
	Frontotemporal dementias
	Posterior cerebral artery infarctions
	Traumatic brain injury
	Urbach–Wiethe disease
	Temporal lobectomies
Dorsolateral frontal systems	Subdural hematomas
	Glioma
	Progressive supranuclear palsy
	Anterior cerebral artery infarctions

In patients with hypothalamic lesions, outbursts of violent behavior in the awake period may be precipitated by internal or visceroceptive states, such as hunger, fatigue, light deprivation, or hormonal stimulation. Alternatively, patients may exhibit a heightened general level of aggressivity. Frequently, attacks are on individuals who happen to be near the patient, without the formation of complex plans. Patients often have diminished insight into the reasons for their actions, although they recall and may demonstrate remorse for their behaviors. Subjects with hypothalamic lesions may demonstrate altered patterns of sleeping or eating, polydipsia, and polyuria or deficient regulation of sex hormones, thyroid, or adrenocortical function. Heteronymous visual field impairments may be evident if lesions extend to involve the optic nerves, chiasm, or tracts. The workup of patients with suspected hypothalamic lesions should include MRI or other structural imaging of the region of the third ventricle, endocrine studies, and formal visual fields. The differential diagnosis includes benign and malignant tumors such as craniopharyngiomas and astrocytomas, which present with subacute alterations in behavior. Another etiology is hypothalamic hamartoma, which usually presents with a distinctive clinical profile of childhood onset of gelastic epilepsy (ictal laughter), sometimes accompanied by precocious puberty, along with interictal bouts of uncontrolled rage (Fig. 4).

2. Temporolimbic Epilepsy Syndromes

Few topics in behavioral neuroscience are more controversial than the relationship of aggression to

Table V
Strategies for Treating Aggressive Behavior

Setting of aggression	Initial therapeutic approaches
Impulsive aggressive acts in setting of congenital or acquired intellectual impairment or hypothalamic injury	Control of appetite, sleep, diurnal cues
	β-Adrenergic blocker
	Selective serotonin reuptake inhibitors
	Cholinergic (muscarinic) antagonists (?)
	Valproic acid
	Avoid barbituates, benzodiazepines, sedatives
Aggression related to deepened affect or ideation in the interctal syndrome of temporolimbic epilepsy (moralistic conviction)	Antiepileptic medications
	Selective serotonin reuptake inhibitors
	Reality-oriented psychotherapy
	Avoid lithium carbonate (may worsen seizures)
Disinhibited aggression in response to transient environmental stimuli, $\pm$ frontal systems dysexecutive signs	Explicit, concrete social structure
	Selective serotonin reuptake inhibitors
	β-Adrenergic blocker
	Avoid barbituates, benzodiazepines, sedatives
Aggressions precipitated by delusions, hallucinations	Atypical antipsychotics
	Neuroleptics
Irritability related to manic or hypomanic states	Mood stabilizers
Acute agitation	Neuroleptics
	Benzodiazepines

epilepsy. Failure to adequately distinguish between aggressive actions in the ictal, postictal, and interictal periods has contributed greatly to the confusion regarding the relationship between temporolimbic epileptic foci and violent behavior.

Ictal aggression does occur, but with extreme rarity. An international consensus panel found that only 7 of 5400 patients studied on video EEG monitoring exhibited aggressive behavior during recorded seizures. Hostile behaviors ranged from physical violence directed toward inanimate objects to mild verbal or physical aggression directed toward a person. When aggressive acts during complex seizures occur, they may appear without provocation or in response to an environmental stimulus and are characterized by spontaneous, undirected aggression toward nearby objects or individuals. The patient is amnestic for actions and often expresses remorse.

A much more common form of aggressive behavior in epilepsy is resistive violence during the postictal period. Following a complex partial seizure or, more frequently, a generalized convulsion, patients may be disoriented and confused. During this epoch, well-intended attempts at physical restraint can provoke aggression, which almost always ceases when restraint is withdrawn. The attacks generally involve striking out without the use of a weapon or sometimes with objects that happen to be close at hand. Patients have no memory for their actions upon clearing of consciousness and will express dismay if they have injured others.

Aggression in the subacute period following the end of a seizure activity can also occur. Often, the aggression appears in the context of postictal psychosis and mania, especially in patients with paranoid delusions and threatening hallucinations. However, recently a syndrome of subacute postictal aggression was described in four patients, occurring hours to days after a seizure, without postictal psychosis or mania. Subacute postictal aggression appears to be a rare phenomenon. The attack behaviors are intentionally directed after minor provocations, and patients retain full recall of the episodes.

Overt aggression related to the interictal behavior syndrome of temporolimbic epilepsy is unusual because the heightened moral and religious values that

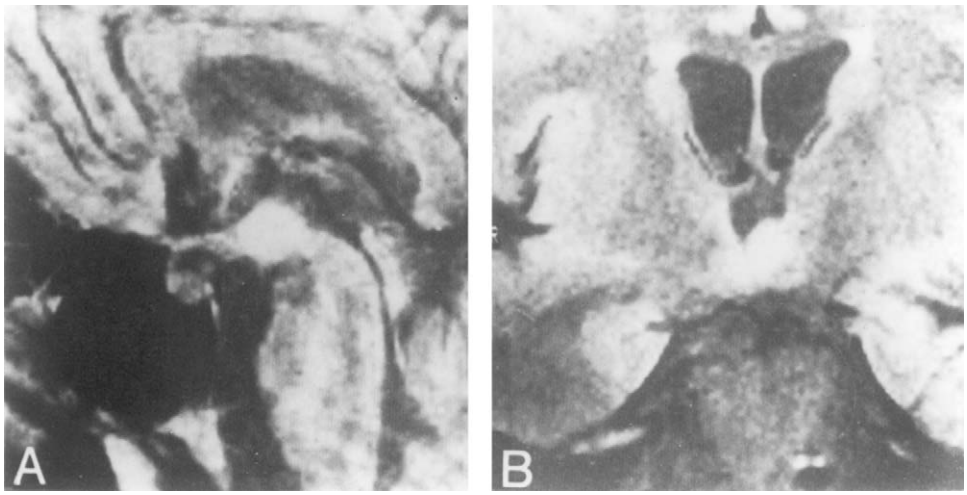


Figure 4 Aggression due to hypothalamic lesion. A 19-year-old man presented with several years of aggressive behavior, poor social adjustment, and seizures of multiple types, including ictal laughter. Coronal proton density-weighted MRI scan demonstrates a hypothalamic hamartoma, evident as a 1-cm-high signal intensity mass in the substance of the hypothalamus (reproduced with permission from S. F. Berkovic, F. Andermann, and D. Melanson, *et al*., 1988, *Ann. Neurol*. **435**).

are features of the syndrome preclude violent actions. However, in rare circumstances, intense emotional reactions to perceived injustice or threat can lead subjects to formulate and carry out complex plans of violent response. Attacks may be directed against a specific individual and could involve the use of a weapon. Not all hostile actions by these patients involve long-term planning—rarely, the intensity of feelings evoked in a particular situation might lead to an immediate response. Patients fully recall their actions and often exhibit extreme remorse. Some individuals continue to believe their acts had ample moral justification.

In individuals with epilepsy, sedative antiepileptic medications such as barbiturates may contribute to hostile behaviors by impairing impulse control. Irritability, a common complaint among patients with poorly controlled partial and primary generalized seizures, may result from environmental factors, medications, or in relation to the underlying cerebral pathology or epileptogenic process.

The laboratory evaluation for epilepsy in violent individuals includes routine scalp EEG and more sensitive sleep recordings and the use of nasopharyogeal or spheroidal leads. In select cases, ambulatory EEG, long-term inpatient video EEG monitoring with scalp and sphenoidal electrodes, or invasive subdural grids or depth electrodes may be necessary to establish the seizure focus.

CT and especially MRI are utilized to exclude slowly growing gliomas and other mass lesions. Volumetric MRI or careful visual analysis of the hippocampus and amygdala (seen best on T_1-weighted coronal cuts) may aid in the diagnosis of mesial temporal sclerosis by demonstrating unilateral or bilateral atrophy. Metabolic imaging with PET or SPECT may show increased blood flow and hypermetabolism in mesial temporal structures during ictal discharges and decreased blood flow and hypometabolism interictally. PET is the more sensitive technique interictally; SPECT is more practical for capturing ictal events. Common etiologies of temporolimbic epilepsy include mesial temporal sclerosis, hamartomas, dysplasia, low-grade astrocytomas, oligodendrogliomas, vascular malformations, and traumatic brain injury.

3. Orbitofrontal Systems Syndromes

Patients with lesions in orbitofrontal cortices or associated subcortical structures such as the caudate nucleus may engage in directed acts of aggression. However, they are often incapable of planning or executing a complex response that requires an extended sequence of actions. Failure to consider long-term and especially social consequences of violent outbursts is a salient feature. Frequently, the patient engages in impulsive, unreflective responses to identifiable but trivial environmental provocations. In some patients, extended but inefficient rational analysis of social situations without guidance from somatic–emotional systems is observed. Patients remember their actions but often lack remorse, and they fail to link aggressive

actions with punishment or other long-term adverse outcomes, contributing to repeated offenses.

The neurological exam may reveal anosmia due to damage to the olfactory nerves or tracts on the undersurface of the frontal lobes and release phenomena such as the grasp reflex. If the lesion is confined strictly to orbitofrontal cortices, subjects may show few deficits on conventional IQ tests or even on neuropsychological tests explicitly designed to probe frontal-executive function. When lesions trespass upon dorsolateral frontal territories, deficits in go-no-go testing, verbal and nonverbal fluency, and set shifting may be evident. CT and especially MRI studies are helpful in screening for structural lesions. Common etiologies indude traumatic brain injury, anterior communicating artery aneurysm rupture, anterior cerebral artery infarction, orbital meningioma, the frontotemporal dementias (Fig. 5), and Huntington's disease.

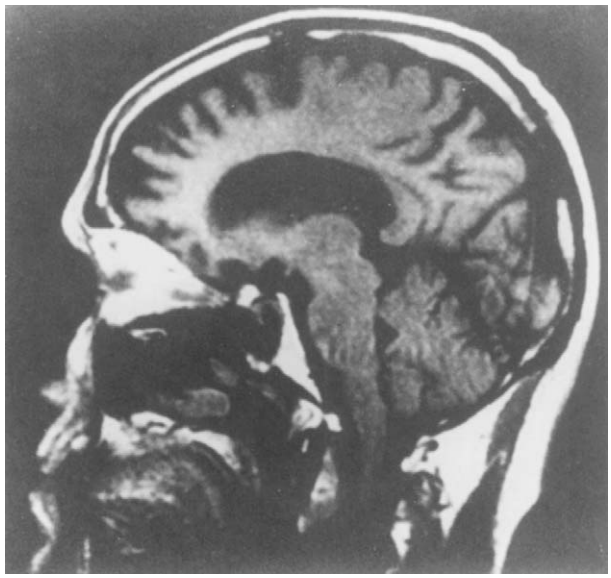


Figure 5 Aggression with bifrontal lesions. Over 3 years, a 56-year-old man developed frequent violent outbursts and persistent foul language, as well as ritualistic behaviors, disinhibition, and jocularity. MRI demonstrates focal frontal lobar atrophy, consistent with Pick's disease. [Reproduced with permission from Saver *et al.*, 1996, Neuropsychiatry of aggression. In *Neuropsychiatry: A Comprehensive Textbook*. (Fogel, B. S., Schiffer, R. B., Rao, S. M., eds.) p. 540. Williams & Wilkins, Baltimore].

4. Syndromes of Diffuse or Multifocal Brain Injury

Several medical conditions that produce aggression have effects on the brain that are diffuse or multifocal. A large body of investigative literature has demonstrated an increased frequency of "minimal brain dysfunction" and poorly localized neurological "soft signs" in violent patients. In one study, minor and irregularly distributed perceptual, motor, sensory, reflex, and cognitive defects were noted in 119 of 286 patients with a history of recurrent attacks of uncontrollable rage. Minor neurologic findings are common in violent individuals in juvenile reform school and on death row. Attention deficit disorder was significantly correlated with criminal and violent offenses in a prospective study. Not surprisingly, developmental or acute medical conditions producing scattered minor neurological impairments place an individual at higher risk of expressing aggressive impulses. Diffuse or multifocal brain insults are likely to affect one or several circuits within the multiregional, hierarchical aggression regulatory system. In addition, a history of being abused or reared in an unstable household is likely to interact synergistically with multifocal brain injuries. An individual who has learned a model of acting on impulse and possesses a limited repertoire of other response options is all the more likely to demonstrate diminished flexibility and inhibition of aggression after diffuse, especially frontal, injuries.

The clinical manifestations of aggression in these patients are heterogeneous. History of school difficulty, learning disability, hyperactivity, or head trauma associated with brief loss of consciousness or amnesia are often elicited. School records are helpful because patients may deny or minimize past academic or disciplinary problems. A clouded sensorium or subtle sensorimotor or visual impairments may be found on neurological examination. Etiologies of diffuse brain dysfunction with episodic violence include, in addition to attention deficit disorder, toxic–metabolic encephalopathies (such as hyper- and hypoglycemia, B_{12} deficiency, and thiamine deficiency), multiple schlerosis, and subcortical vascular dementia.

5. Delusional Syndromes

Delusional individuals are prone to violent behavior. In these patients, aggression-related neural systems may, like the intellect, be placed in service of the psychosis. In a variety of neuropsychiatric disorders, including schizophrenia, endogenous depression, mania, Alzheimer's disease (AD), and other dementias, the presence of thought disorder and hallucinations,

especially of the persecutory type, increases the risk of violent outbursts. In a series of 181 subjects with probable AD, physical aggression was observed in 30% of patients; delusions and misidentifications frequently preceded violent outbursts. Dementia patients with paranoia and aggressive behavior have an increased rate of early institutionalization. Delusions are more likely to lead to aggression if frontal systems are also impaired.

The clinical approach to these patients is focused on the diagnosis and treatment of the underlying psychotic disorder. Etiologies of delusional disorders include paranoid schizophrenia, affective illness, late-life paraphrenia, AD, multiinfarct dementia, and subcortical dementias.

XIII. TREATMENT

Consideration of multiregional neural processing may guide selection of appropriate environmental and biological interventions to control aggression.

Regular satisfaction of feeding drives and sleep–awake cycles may minimize hostile outbursts in patients with hypothalamic aggression. Much of our current pharmacological armamentarium for the treatment of violence is focused on neuromodulators with cell bodies in the brain stem and hypothalamus that project widely to hemispheric structures, relying on a pharmacologically induced bias against all types of aggressive responses. For example, brain stem nuclei are the likely sites of action for drugs that block β–adrenergic receptors, activate 5-HT receptors, and enhance GABA activity. By raising the threshold for aggressive responding, such agents may have an ameliorative effect on aggression resulting from lesions at all levels of the neuraxis. Agents modulating muscarinic cholinergic receptors in the lateral hypothalamus are worthy of more systematic investigation for antiaggressive effects based on observations that cholinomimetics directly instilled in the lateral hypothalamus elicit aggression. In one patient with lesions thought to disconnect hypothalamic circuitry from higher control, oral administration of a centrally active cholinergic antagonist dramatically suppressed aggressive, biting behavior. Intervention at the temporolimbic level of control introduces additional considerations. Patients in whom a focal seizure disorder is diagnosed should be treated with agents effective for complex partial seizures, especially those such as carbamazepine, valproate, and gabapentin which have mood-stabilizing effects. Neuroleptics, which lower the seizure threshold, should be employed with caution. Avoiding restraint and providing gentle reassurance and reorientation in the postseizure period reduces postictal resistive violence. Patients whose clinical seizures remain refractory to antiepileptic therapy and who have well-defined unilateral foci are candidates for surgical resection. Temporal lobectomies or amygdala hippocampectomies produce excellent seizure control in the preponderance of such patients and may have a beneficial effect on aggression in patients whose epilepsy is associated with violent behaviors. Reduced aggression has been reported in more than 35% of violent epilepsy patients after temporal lobectomy.

The principles of pharmacological therapy for interictal aggression differ from those directed at reducing aggression levels in the hypothalamus. At the amygdalar level, the problem of timing is critical. To form appropriate sensory–emotional associations, the normal amygdala must be active in a brief time window to bind a specific stimulus or event with the hypothalamic signal of hunger or anger. Prolonged responses evoked by temporolimbic epileptic foci lead to an inappropriate broadening of the range of associated stimuli. Antiepileptic agents such as carbamazepine or valproic acid, which reduce prolonged, rapid discharges of neurons, may favor more selective, adaptive associations.

Recent studies have suggested that excitatory amino acid receptors, such as the class activated by *N*-methyl-D-aspartate (NMDA), may be critically involved in the associative process of long-term potentiation within the hippocampus and amygdala. Agents that modulate this response, such as selective NMDA–receptor blockers or nitric oxide synthetase inhibitors, merit investigation for antiaggressive and serenic properties in temporolimbic epileptic patients.

Processing within the orbitofrontal cortices represents an advanced stage of synaptic interaction occurring at the convergence of multiple streams of prior sensory evaluations. It is unlikely that a single neurotransmitter could be modulated to duplicate or restore prefrontal functions. Conversely, drugs that nonselectively inhibit neuronal function through inhibition of chloride channels or other mechanisms, such as ethanol, benzodiazopines, and barbiturates, exert a disproportionate effect on prefrontal functioning, which is particularly dependent on polysynaptic inputs. A paradox in the current psychopharmacology of aggression may be illuminated by considering the prefrontal effects of agents that simultaneously act at

other levels of the neuraxis. For example, benzodiazepines have antiaggressive properties in many species and tranquilizing effects in humans, likely mediated by their potentiating effects at GABAergic inhibitory receptors. However, these compounds are known to precipitate paradoxical rage, the release of previously inhibited hostile responses. A possible explanation is that benzodiazepines impair prefrontal processing through nonselective neuronal inhibition in a manner similar to ethanol intoxication. In some patients, the resulting loss of social insight and judgment more than offsets the general tranquilizing effects of these agents.

Serotonergic agents may also exert substantial effects on prefrontal function. Serotonergic efferents from prefrontal cortices appear to serve an important role in the inhibition of impulsive responses. Lowered levels of 5-hydroxyindoleacetic acid and blunted responses to fenfluramine in impulsive aggressive individuals are consonant with this formulation, suggesting that serotonergic agents, especially those acting at 5-HT_1 receptors, may be beneficial to patients with orbitofrontal dysfunction. α_2 adrenergic agonists, such as clonidine and guafacine, have been shown to ameliorate frontal focused attention deficits and may thus be helpful to individuals with orbitofrontal derangements.

Supportive psychotherapy is helpful to many patients with temporolimbic and orbitofrontal aggression syndromes. Insight-oriented therapy is unlikely to benefit the patient with prefrontal injury who may verbally comprehend a behavioral problem and propose solutions but cannot reliably call on this knowledge to control behavior. Some patients with interictal behavior changes associated with temporal lobe epilepsy are amenable to insight psychotherapy, but many accept neither criticism nor advice. However, by alerting patients to their intensified emotional responses, and capitalizing on their heightened moral and religious sensitivities, a therapist may reduce their likelihood of aggressive actions.

XIV. CONCLUSIONS

Many aggressive behaviors are neither maladaptive nor the result of neurological disease. Even when violent behavior and neurological lesions coexist, they may not be causally related. A violent lifestyle may lead to head trauma and neurological abnormalities that are the consequence, rather than the cause, of aggression. While fully accepting these qualifications, behavioral neuroscience is enriched by recognizing both that a diverse array of neurological lesions may contribute to violent behavior in human beings and that abnormalities at different levels of the neuraxis produce distinctive subtypes of aggression. Basic and clinical studies that consolidate and extend our understanding of the multiregional, hierarchical neural networks regulating aggression are urgently needed to refine diagnostic and therapeutic approaches to violent individuals.

See Also the Following Articles

ANGER • BEHAVIORAL NEUROGENETICS • BEHAVIORAL PHARMACOLOGY • COGNITIVE PSYCHOLOGY, OVERVIEW • HYPOTHALAMUS • NEUROPSYCHOLOGICAL ASSESSMENT • PREFRONTAL CORTEX • PSYCHONEUROENDOCRINOLOGY • SEXUAL BEHAVIOR • VIOLENCE AND THE BRAIN

Suggested Reading

Benjamin, S. (1999). A neuropsychiatric approach to aggressive behavior. In *Neuropsychiatry and Mental Health Services* (F. Ovsiew, Ed.), pp. 149–196. American Psychiatric Press, Washington, DC.

Davidson, R. J., Putnam, K. M., and Larson, C. L. (2000). Dysfunction in the neural circuitry of emotion regulation—A possible prelude to violence. *Science* **289,** 591–594.

De Waal, F. B. M. (2000). Primates—A natural history of conflict resolution. *Science* **289,** 586–590.

Mann, J. J. (1998). The neurobiology of suicide. *Nature Med.* **4,** 25–30.

Raine, A., Lencz, T., Bihrle, S., LaCasse, L., and Colletti, P. (2000). Reduced prefrontal gray matter volume and reduced autonomic activity in antisocial personality disorder. *Arch. Gen. Psychiatr.* **57,** 119–127.

Saver, J. L., Salloway, S., Devinsky, O., and Bear, D. M. (1996). The neuropsychiatry of aggression. In *Neuropsychiatry: A Comprehensive Textbook* (B. S. Fogel, R. B. Schiffer, and S. M. Rao, Eds.), pp. 523–548. Williams & Wilkins, Baltimore.

Volavka, J. (1995). *Neurobiology of violence.* American Psychiatric Press, Washington, DC.

Volavka, J. (1999). The neurobiology of violence: An update. *J. Neuropsych. Clin. Neurosci.* **11,** 307–314.

Aging Brain

LORI L. BEASON-HELD and BARRY HORWITZ
National Institute of Aging

GLOSSARY

Alzheimer's disease A progressive neurodegenerative disease causing dementia that is common in the elderly, whose symptoms include memory dysfunction, deficient cognitive function, and an inability to deal with the activities of daily living. Neuropathologically, it is characterized by the presence of senile plaques, neurofibrillary tangles, and neuronal loss in the hippocampal area, association areas of the cerebral cortex, and a number of subcortical nuclei, including the nucleus basalis of Meynert.

nucleus basalis of Meynert A group of neurons in the basal forebrain that project to the cerebral cortex, constituting the cholinergic input to the cerebral cortex.

typical and successful aging Typical aging refers to individuals who are cognitively intact, even though they may have some nondementing illness that could affect brain structure and/or function (e.g., hypertension); successful aging refers to individuals who are free of any illness that could affect brain structure or function.

white matter hyperintensities Abnormal signals in the white matter observed with magnetic resonance imaging. Two kinds are denoted: periventricular white matter hyperintensities and deep white matter hyperintensities.

This article reviews some of the changes that occur in the human brain with advancing age. After some introductory comments, we give a brief overview of the age-related alterations in cognition and sensorimotor performance, discuss senescent changes in brain structure at the macroscopic and cellular levels, and examine age-associated changes in brain function as assessed by functional neuroimaging.

I. INVESTIGATING THE AGING BRAIN

The changes in the brain associated with aging are seemingly both evident and elusive. Obviously, aging continually takes place from conception to death, but in this article, we address only that aspect of aging associated with senescence. What is evident is that there is apparently a clear decline in numerous brain-mediated sensory, motor, and cognitive processes with advancing age, although what exactly declines is not fully understood. Likewise, the gross morphology of aged brains looks different than that of young brains; generally, they have larger sulcal widths and correspondingly smaller gyri. Neurons also should be affected by age; after all, these cells are postmitotic, and once they die most will not be replaced. However, documenting neurobiological alterations due to aging in humans (and nonhuman primates) has been remarkably difficult and the results controversial.

Four fundamental problems make the investigation of the senescent human brain complex. First, how should one study aging? Comparing a group of young versus a group of old subjects (i.e., a cross-sectional design) has numerous disadvantages, including that one is not studying the aging of any single subject, and the results can be compromised by secular effects. For example, the young subjects of today grew up in a world of better health care than did the old subjects. They may be physically larger, may have engaged in

different kinds of physical activity, and may have different diets. On the other hand, elderly subjects have the advantage that they have survived to be old subjects. Also, comparable levels of formal education in each group may not mean the same thing. The other option is to study a single group of subjects as they grow old (i.e., a longitudinal design). This is quite difficult for many reasons: Some age-related changes can begin as early as the fifth or sixth decade, if not earlier, but following subjects over several decades is clearly expensive and logistically complex. Moreover, a longitudinal design often forces one to continue using a technique that can become outmoded [e.g., employing X-ray computed tomography (CT) to examine the structure of the brain when magnetic resonance imaging (MRI) scans give much better information]. Also, as subjects age, they may develop a brain-related disorder; therefore, the question arises as to how much of their previously acquired data should be used. Although longitudinal designs probably are the better choice, the majority of aging studies employ a cross-sectional design.

The second fundamental problem in studying the aged brain has to do with brain disease. Presumably, we are interested in those changes due to the aging process. However, the incidence of many neurological and psychiatric disorders increases with advancing age. Some neurodegenerative diseases, especially Alzheimer's disease, are extremely difficult to differentiate in their early stages from normal aging. Often, the diagnosis for some of these diseases can only be made postmortem by searching for specific neuropathological markers in autopsied tissue; thus, if a healthy aged subject shows a behavioral change, is it because there is subclinical disease present or does the change actually represent something that is age related? What makes this especially troubling for the scientific investigator is that the pathology of some disorders such as Alzheimer's disease may start to occur one or two decades prior to the onset of clinical symptoms. Until reliable *in vivo* markers for these diseases are available, the problem of knowing what constitutes an aging change from an alteration due to disease will continue to affect this field of research. A similar problem occurs for MRI white matter signal abnormalities, which are also present in many disorders, such as hypertension, vascular disease, and diabetes. Moreover, what exactly is meant by the term "healthy aging"? For example, it is still an unanswered question as to whether everyone will become demented if they live long enough; that is, some of the pathological features and some of the cognitive symptoms of Alzheimer's disease (or a related dementia) are found in old nondemented people, but to a lesser degree.

One relatively recent set of findings illustrates this problem particularly well. Several studies have shown that the female sex hormone estrogen may affect neurons so that higher hormonal levels improve learning and memory. Moreover, some research suggests that estrogen may protect against Alzheimer's disease. Therefore, when women go through menopause, they could show a decline in cognitive function due to reduced estrogen levels, and thus, there will be a sex-based aging difference. For those women at risk for Alzheimer's disease, there may be a further insult on neuronal function. Estrogen replacement therapy, more common now than in previous years, may ameliorate these effects. Whether or not these specific findings hold up as research continues, this illustration shows the complex interactions involving aging changes, sex, cognitive alterations, and disease risk factors that can confound experimental studies of aging and that can affect our ability to define what healthy aging even means.

These considerations have led investigators to try to distinguish between typical (or usual or normal) and successful aging. Cognitively intact individuals in the first category may have nondementing illnesses that increase in prevalence with advancing age. For example, they may have vascular risk factors such as hypertension that are associated with impaired cognition in the old. The second category, successful aging, refers to individuals who are free of these illnesses. Because various research groups differ as to whether they study typical or successful aging, their results often are at odds with one another.

The third fundamental problem for the investigator of brain senescence has to do with neuroplasticity, a term we use in its most general sense. Changes due to neuroplasticity can include such regionally local processes as axonal and dendritic sprouting as well as long-range processes best expressed in terms of altered patterns of task- and behavior-related neural activity in functional systems distributed across the brain. The brain's inherent plasticity allows it to compensate to a degree for the degradative processes associated with aging. This means that behavioral performance may appear unchanged even in the presence of structural and functional neurobiological age-related alterations. It means that it can be difficult to correlate any specific neurobiologic change with a corresponding cognitive deficit. It also means that some of the changes seen with aging may represent the effects of plasticity and thus are related to aging only indirectly. Finally,

neuroplasticity can make it difficult to distinguish subjects who have disease-related pathology from the healthy aged. On the other hand, neuroplasticity also means that the brain has some capacity to adjust to both aging and disease in such a way that many individuals can maintain a normal and productive mental life even into advanced old age.

The fourth problem, which can be particularly acute for those who work at the cellular and subcellular levels of analysis, is that the brain is incredibly heterogeneous in terms of structure and function. Different parts of the brain may show differential aging effects, as might different neuronal populations even within the same brain region. Because the brain operates by means of distributed networks, it can be difficult to relate the age-associated anatomical changes seen in one brain area to functional changes in behavior.

These points are worth keeping in mind as we review what has been learned about the neurobiological alterations with advancing age. These issues provide a framework for appreciating the difficulty investigators have found in arriving at a detailed consensus about what features of the brain decline during old age.

Before we begin our review, however, several other items need to be pointed out. First, a central finding in almost all studies is the increased variability in whatever one is measuring in the old compared to the young; this often means that values for the quantity of interest of many old subjects are well within the range of values found in the young subjects. Second, the age-related differences we discuss are quite small compared to the differences found between normal subjects and those with brain disease. Finally, unless stated otherwise, when we talk about the old or elderly, we are usually referring to subjects between the ages of 65 and 90 years. The "oldest old" individuals appear to constitute a special group, perhaps due to genetic factors.

II. COGNITIVE CHANGES WITH AGE

There is a rich literature detailing the age-related changes in cognitive and sensorimotor function. Cognitive ability is generally tested using a battery of neuropsychological measures that includes tests of intelligence, executive function, language, visuospatial ability, attention, and memory. From such batteries it has been concluded that healthy aging is associated with a decline in several domains of cognitive function, some of which can begin as early as the sixth decade of life.

In evaluating these tests, it is important to distinguish aspects of performance that engage peripheral processes from central processes. For example, as we age, the lens of the eye changes so that we become farsighted. Obviously, to assess visual processing, an investigator must take account of this peripheral age-related effect. The declines in cognitive performance with age discussed next are thought to be mediated primarily by the central nervous system and not by peripheral neural processes.

A. General Intellect

Whereas general knowledge is preserved with aging, studies show a differential decline on some intelligence measures, leading aging researchers to propose that there are two distinct types of intelligence. Crystallized intelligence refers to knowledge accumulated over the life span and includes vocabulary, general information knowledge, comprehension, and arithmetic. Performance on tests of crystallized intelligence is generally preserved with aging. The ability to evaluate and respond to novel events is referred to as fluid intelligence. Tests of fluid intelligence evaluate deductive reasoning, problem solving, memory spans, and figural rotation. These measures decline with age. Because many of these tests must be completed in a given time period, poorer performance in the elderly on tests of this nature may be related to slower information processing speed and slower motor response times, which also show age-related decrements.

B. Executive Function

Executive function includes processes associated with high-level cognitive abilities. Tests of executive function involve mental flexibility, abstraction, and planning; many are thought to be related to frontal lobe function since patients with frontal lobe lesions perform badly on such tests. Mental flexibility can be assessed with tasks requiring sorting and set shifting, such as a card sorting task in which the rules governing the sorting process change at intervals during the task. Performance on tests of this nature is impaired with age. These impairments are worse if the tasks involve a memory component. Abstraction can be assessed using tests involving concept formation, reasoning,

and categorization. Tests which involve the ability to detect similarities between stimuli and the rules governing the sequencing of a string of letters or numbers are used to measure abstraction capabilities. Although these abilities decline with age, some impairments appear to be education dependent. For example, on tasks of concept formation, the higher the education of the subject, the slower the decline in performance ability with age.

C. Motor Processes

Slowing of motor processes has long been associated with aging. For example, performance on tests of walking speed and finger tapping rates declines with age. Elderly subjects are also slower than young subjects when moving a hand toward a target, sorting a stack of cards, and writing. Many studies that require a motor response, such as pushing a button, also show increased reaction times in the elderly, with or without a decrease in response accuracy. It is of interest that tests that require a verbal response do not show significant changes with age, suggesting that speed may be an important factor on those tasks requiring hand or finger movements, and especially on tasks that must be completed in a given time period. Some skills acquired early in the life span which require motor abilities, such as typing and playing musical instruments, are preserved with age.

Areas of the brain that have been shown to be important in the control of movement include motor and premotor cortex in the frontal lobe, the basal ganglia, and the cerebellum.

D. Language

Most components of language are preserved with age and those that change do so late in the life span. Four aspects of linguistic ability that have been measured are phonological, lexical, syntactic, and semantic knowledge. Phonologic knowledge, or the ability to use different sounds and understand the rules of their combination, is preserved with age. However, because there is a decline in hearing sensitivity with age, the elderly often report difficulties in understanding speech. Nonetheless, under normal conditions, when hearing sensitivity is controlled for, the young and old seem to understand speech equally well. Lexical knowledge, or understanding the actual name of an object or action, is preserved with age. The elderly can also discriminate as well as young subjects between words and nonsense words. The ability to meaningfully combine words into sentences is referred to as syntactic knowledge. This ability is also preserved with age unless a memory component is involved. Semantic knowledge, which is one type of long-term memory, refers to word meaning. Naming and verbal fluency assess semantic ability and decline with age, although impairments usually do not become significant until age 70.

For right-handed individuals, the regions of the brain considered essential for language include several areas bordering the Sylvian fissure in the left cerebral hemisphere (e.g., Wernicke's area in the posterior superior temporal gyrus for comprehension and Broca's area in the inferior frontal lobe for production).

E. Visuospatial Processes

Visuospatial tasks assess the ability to recognize and reproduce geometric drawings, perform construction tasks, and carry out object recognition tasks that require the subject to identify specific object features. Although elderly subjects can perform some types of visuospatial tasks with the same accuracy as young subjects, albeit more slowly, many visuospatial abilities do decline with age. Elderly individuals are impaired on tasks of construction with blocks, object assembly, maze learning, and performance of puzzles with a spatial component. Impaired ability to draw and to judge the accuracy of drawings also occurs with age. A decline in the ability to detect a target position in the visual field and to detect the direction of motion is additionally observed with aging. Some of these tasks require the integrity of regions in the occipital and parietal lobes, especially in the right hemisphere.

F. Attention

Attention can be defined in a number of ways, but for this discussion, perhaps the best definition is the ability to concentrate or focus mental powers on the task at hand. Studies have examined different types of attentional capacities. One type, sustained attention, involves the ability to focus on a simple task without losing track. No decrease in the ability to perform tasks involving simple vigilance or sustained attention is seen with aging. However, if the task is more complex and requires the subject to remember previous stimuli,

for example, an impairment can be observed. Another type of attention, selective attention, requires the subject to ignore extraneous information and to focus only on the relevant stimuli. Early studies suggested that older subjects were impaired on this type of task because they were less efficient at ignoring irrelevant stimuli. Recent studies involving the ability to search an array of stimuli for specific targets indicate that there is no change with age, although there does seem to be an impairment in controlling the focus of attention during visual search. Another type of test used to assess concentration ability involves divided attention in which subjects must perform at least two tasks simultaneously. This testing paradigm requires not only that the subject focus attention but also that the subject switch his or her attention from one task to the other. On tests involving divided attention, older subjects appear to be impaired on all tasks except those involving simple perception. Examples of more complex tasks used to study divided attention include remembering strings of letters and numbers simultaneously and responding when a specific number of dots appear in the field of view while performing a driving simulation. Some researchers attribute impairments on measures such as these to task difficulty and slower response times in the elderly.

G. Memory

Memory problems are the most common complaint of the elderly. Difficulties with memory can begin as early as the 50s. In general, the ability to access remote information and the capacity to attend to new information are not affected by age, although there does seem to be an age-related increase in difficulties with word finding and naming, as mentioned previously. Instead, aging deficits are seen in the ability to learn and retain new information. It is clear that there are several types of memory as well as several processes involved in the ability to remember. Sensory or immediate memory, which involves the ability to remember items instantly, is preserved with aging. Short-term (or working) memory involves memory for items held over seconds or minutes. This is a limited capacity store where the material must be actively rehearsed to be maintained in memory. Tests of short-term memory include remembering lists of words or letters. This ability is relatively preserved with aging, although there are some suggestions of a reduced short-term storage capacity in the elderly. The modern conception of working memory also entails a component that permits the information in the short-term store to be manipulated. Evidence suggests that this component becomes impaired with advancing age. The dorsolateral frontal cortex is thought to play a central role in working memory.

Items stored in long-term memory may be accessed over hours, days, or years. Tests of recall or recognition of stimuli are often used to assess long-term memory. Both types of tests are impaired in old subjects, with recall appearing to be worse than recognition. Three processes are involved in long-term memory: encoding, storage, and retrieval of information. Long-term storage abilities appear intact in the elderly because the results of studies examining rates of forgetting are similar for old and young subjects. It is theorized, instead, that older subjects have difficulties with encoding and retrieval. Studies of encoding suggest that the elderly are less efficient at using proper strategies when committing information to memory. Because tests of recall are impaired with age, it is also thought that the retrieval process is disrupted in elderly subjects.

There is also a distinction between the types of material to be remembered. Episodic memory includes memory for events, whereas semantic memory involves memory for facts, ideas, and concepts. Episodic memory is impaired in aging. Early studies found that semantic memory was relatively preserved with age, but recent studies suggest that there is also a slight decline in the ability to remember semantic information. Another distinction involving types of material includes implicit and explicit memory. Implicit memory refers to memory for information known to the subject. This can be tested using word fragment completion and degraded picture tasks. These types of tasks are performed relatively well by elderly subjects. Explicit memory involves conscious recollection of new information. This type of memory is impaired with age.

Based on lesion studies in nonhuman primates and rodents, and on research with amnestic patients, structures in the medial temporal lobe, particularly the hippocampus and entorhinal cortex, are thought to play a central role in encoding information into long-term memory.

H. Summary: Cognitive Changes with Age

Aging results in declines in a variety of cognitive domains, but some abilities appear to be relatively preserved. General intellectual knowledge and

crystallized intelligence measures such as vocabulary and comprehension remain largely intact with age, as do attention processes that allow one to remain vigilant or to selectively attend to situations. Language abilities related to phonologic, lexical, and syntactic knowledge are also relatively stable, as are motor skills that are learned early in the life span and repeatedly used. Although the most common complaint among the elderly involves memory problems, some aspects of memory are also preserved with age, including processes involved in immediate and implicit memory, and some aspects of short-term memory.

Processing speed needed for most types of cognitive operations slows with age, and there are declines in the ability to reason and solve problems—areas of fluid intelligence. Older subjects also have difficulty dividing their attention between two tasks when these tasks are performed concurrently. Executive function domains such as mental flexibility, abstraction, and concept formation are impaired, as are visuospatial skills required for drawing, construction, and maze learning. Declines are also observed in language skills involving semantic knowledge needed for naming and verbal fluency. Motor skills that require speed are impaired. Also, as early as age 50, studies show that there are age-related declines in the ability to learn and retain new information for long-term access. These memory impairments are thought to be related to deficits in encoding and retrieving information to be remembered.

III. ANATOMICAL CHANGES WITH AGE

There are two ways to assess changes in brain structure in the senescent brain: *in vitro* studies of postmortem tissue and *in vivo* studies of the living brain using techniques such as X-ray CT or MRI. Until about 1990, most *in vivo* studies used X-ray CT, whereas most current investigations generally employ MRI because of its better spatial resolution and contrast. Studies of structural changes at the cellular and subcellular level require the use of either light or electron microscopy and generally are performed on postmortem tissue that has been chemically fixed.

A. Gross Anatomical Changes

1. Postmortem Tissue Studies

Brain weight is known to decrease with age. Changes begin in the third and fourth decades and show a progressive decline throughout the life span. At age 20, the average weight of the male brain is approximately 1400 g, and by the age of 65 brain weight is approximately 1300 g. Brain weight for females follows a similar trend, although the total weight is 100–150 g less than that of males. Most changes occur after age 55, with a total loss of up to 15% of the peak brain weight by age 90.

Shrinkage of brain tissue or atrophy also occurs with age. Atrophy is clearly seen along the surface of the cerebral hemispheres where the gyri or cortical ridges become progressively more narrow and the spaces between the gyri, referred to as sulci, widen. Mild to moderate cerebral atrophy occurs in a heterogeneous fashion, with frontal, parasagittal, and temporal regions affected more than other areas of the cortex. The cerebral ventricles also dilate with age, typically becoming apparent in the 60s. Severe atrophy is generally associated with disease processes. Changes in brain volume are also associated with aging and begin after age 50, with a 2 or 3% progressive decrease per decade. Until age 50, a decrease in volume is predominantly observed in the gray matter of the cerebral cortex. After age 50, decreases in white matter are greater. Autopsy data suggest that the volume of white matter is decreased in old (75–85 years) compared to young subjects by approximately 11%.

2. *In Vivo* Structural Changes

Both cross-sectional and longitudinal aging studies of changes in brain structure have been performed on healthy subjects using either X-ray CT or MRI. Our focus is mainly on the results of the MRI studies because of the better spatial resolution and contrast obtainable with this method. One can manipulate the scanning parameters of a MRI device in several ways, each of which emphasizes signals corresponding to somewhat different features of brain tissue. In this overview, we discuss senescent changes in volumetric measures of gray matter, white matter, cerebrospinal fluid (CSF) space, and a few fairly well-defined brain regions, such as the hippocampus and the basal ganglia. The absence of clearly and easily identifiable landmarks makes it difficult to measure the volumes of specific regions of the neocortex in a rigorous way. We also review some of the findings about age-related alterations in white matter hyperintensities.

a. Some Fundamentals Concerning Structural Brain Imaging Before discussing the specific structural brain imaging findings, it is worthwhile to provide some information about how these types of data are

acquired. Both CT and MRI generate images of the brain in the form of slices parallel to one another. With MRI, these slices can be parallel to any plane that one chooses. The slices have a certain thickness and may or may not be contiguous. Recent imaging studies have tended to use quite thin slices (about 1 mm thick).

There are essentially two distinct methods that have been used to evaluate brain volumetrics on CT or MRI images. The first involves determining the volumes of specific brain regions by manually tracing their areas on individual slices. The second method uses some type of computer algorithm to segment automatically each image into specific tissue types (e.g., gray matter, white matter, and CSF). Each method has limitations. The trace method is more subjective, although it enables the tracer to make use of his or her knowledge of neuroanatomy. The most precise results are obtained with thin slices, but the analysis can be quite time-consuming. The segmentation method has the advantage of objectivity, but it can be especially susceptible to the partial volume problem: Even a single pixel (the smallest element in an image) may contain a mixture of brain tissues, and thus it becomes difficult to categorize every element in the image as belonging to one kind of tissue versus a second. In an image in which the blackest pixels correspond to one tissue type and the whitest to a second, gray pixels may correspond either to a third tissue type or may represent the partial volume averaging of the first two types. As investigations of this sort have continued, improvements in both the trace and segmentation methods have been devised, but the types of problems indicated previously persist.

One last technical point needs to be mentioned. Because the size of an individual's brain and its components is related to the size of the subject (and thus, for example, men have on average larger brains than women), almost all studies of brain volumetrics use normalized volumes; for instance, a commonly used volumetric measure is the percentage of the intracranial volume of the subject.

b. Age-Related Volumetric Changes There seems to be essentially uniform agreement that the total amount of brain tissue decreases with advancing age. Also, all investigations have reported an increase in the volume of CSF in the brain of elderly compared to young subjects. When these changes begin and whether the loss of brain tissue corresponds primarily to gray matter, white matter, or both have been contentious issues. Several investigations have indicated that gray matter shows a gradually accelerating decline with age, but after approximately age 50 the white matter of the brain shows a more pronounced age-related decrease and seems to be the main contributor to the age-related loss of brain volume. Also reasonably clear is that these changes are different for men and women. Whereas the increase with age in ventricular CSF is the same for the two sexes, men show a greater increase in peripheral CSF (i.e., subarachnoid CSF) than do women. Peripheral CSF is considered to be a marker of cortical atrophy.

The frontal lobes have been shown to decline in volume with advancing age, with some research groups reporting that the loss of tissue is greater in men than in women. The parietal lobes have also been shown to be reduced in aging, but to a greater extent in women compared to men. Some groups have found no age-related change in the volume of the temporal lobes, whereas others have; the different findings could be attributable to whether or not the subjects under study represented successful (no change in temporal lobe volume) versus typical aging. Some studies have also examined the size of the corpus callosum (the large bundle of nerve fibers connecting the left and right cerebral hemispheres) as a function of age. Several of these have found that the anterior portion of the corpus callosum shows age-related atrophy, which is consistent with the age-related reduction in frontal lobe volume.

Most brain structures that are measurable by manual tracing with MRI have been found to show age-related decreases in volume. These include basal ganglia structures, such as the caudate and lenticular nuclei, and the anterior portion of the thalamus. Many studies have examined the hippocampus and other components of the medial temporal lobe. Most, although not all, investigations have reported a decrease in the size of the hippocampus with advancing age. In one longitudinal study of individuals ranging in age from 70 to 89 years, it was found that the hippocampus decreased in volume by approximately 1.5% per year.

In summary, although there is still much debate in the scientific literature, it seems that most brain structures show a measurable decrease in size with advancing age. However, compared to neurodegenerative diseases such as Alzheimer's disease, these age-related decreases are small. For example, the study that reported the 1.5% annual rate of hippocampal atrophy also found that this rate was approximately 4% in patients with Alzheimer's disease. Moreover, these changes generally do not begin until the sixth decade. The most notable feature of these data is the

increase in variance with advancing age. That is, there are elderly individuals with values within the young normal range, even though some older subjects show significant atrophy. This aspect of aging is not restricted to brain volumetrics, however; as mentioned previously, it essentially typifies most quantitative studies of aging. Finally, it should be noted that the relationship of these measures of increased atrophy to cognitive decline is unclear. Very few investigations have examined the correlation between cognitive decline and brain volumetrics, especially in a longitudinal design.

c. White Matter Hyperintensities With the advent of MRI, abnormal signals were observed in the white matter in a number of neurological diseases known to affect the white matter (e.g., multiple sclerosis and Binswanger's disease), in various dementias, as well as in the elderly. Because these MRI signals are often best observed using scanning parameters that result in their appearing as bright lucencies against a black (low-signal) white matter, they have been termed white matter hyperintensities (WMHs). Two kinds have been distinguished: (i) Periventricular white matter hyperintensities (PWMHs) appear either as frontal or occipital caps of the cerebral ventricles or as a thin lining surrounding the ventricles, and (ii) deep white matter hyperintensitites (DWMHs) are seen as subcortical punctate foci, although larger confluences of foci form with increased severity and often merge with the PWMHs. The neuropathological substrate for these signals can vary depending on the disease. In aging, the PWMH most likely results from the breakdown of the ventricular ependyma, which leads to an increase in the water content of the nearby myelin, demyelination, and reactive gliosis. DWMHs likely correspond to gliosis, demyelination, and atrophy and shrinkage of axons and myelin around blood vessels. Because DWMHs are found in patients with cerebrovascular disease, vascular dementia, and hypertension, their appearance in the elderly may have an ischemic origin.

The general finding of MRI studies is that WMHs are rare in healthy individuals less than 50 years of age, but their presence increases with advancing age. WMHs are present in a majority of elderly subjects but are generally mild in those individuals who have no cerebrovascular disease; the PWMHs appear as a capping or pencil-thin lining of the lateral ventricles, whereas DWMHs are seen only as diffuse, focal, punctate foci. One meta-analysis of 156 studies showed that the prevalence of WMHs increased from 25% at age 30 to 75% at age 80. This study also found that age and hypertension were the major predictors of the presence of WMHs.

The relation between severity of WMHs and deficient cognition is less clear. As indicated previously, aging leads to measurable deficits in certain cognitive functions. Individuals at all ages with essential hypertension have also been shown to have impaired cognition in several domains compared to normotensive control subjects, even if the hypertension has been well controlled by drugs. Elderly hypertensive subjects have a greater amount of brain atrophy than do normotensive age-matched controls. Also, as stated previously, the extent of WMHs is greater in hypertensives than in controls. However, even in healthy subjects free of cerebrovascular risk factors, it has been reported that increased volume of WMHs is associated with increased ventricular volume and deficient cognitive function, especially on neuropsychological tests sensitive to frontal lobe dysfunction. Interestingly, there seems to be a correlation between elevated systolic blood pressure, even in the normal age-related range, and WMH burden.

B. Microscopic Changes in Neuronal Structure and Neurotransmitter System Integrity

1. Cellular Changes

There are a number of cellular changes that occur in the gray matter of the brain with aging. Although these changes are more dramatic in disease states, there are also degenerative processes associated with healthy aging. One of the most fundamental and currently unresolved issues is whether or not there is a loss of neurons in the cerebral cortex of the aged brain. Early studies suggested that cell loss did occur with age, but recent studies have shown a decrease in the number of large cells and an increase in the number of small cells, suggesting that shrinkage of the cell body, as opposed to a decrease in number, is associated with senescence.

Some areas of the brain do show a decrease in neuronal number, including the hippocampus, thalamus, putamen, cerebellum, and subcortical nuclei such as the substantia nigra, locus coeruleus, nucleus basalis of Meynert (NBM), and inferior olive. The hippocampal region is of interest because of its purported role in encoding new memories; the earliest neuropathological changes in Alzheimer's disease occur in parts of this structure and in surrounding tissue. Parts of the basal ganglia (putamen, globus pallidus, and portions of the

thalamus), along with the cerebellum and substantia nigra, are key components of the neural system involved with regulating movement. Neuronal death of the dopaminergic cells in the substantia nigra causes Parkinson's disease; several symptoms of this degenerative disease increase their frequency in the aged. Some studies have reported neuronal loss in the NBM, the source of the cholinergic projection to the cerebral cortex. This nucleus shows a significant loss of neurons in Alzheimer's disease.

There are also degenerative changes that occur within the nerve cell body, including accumulation of lipid products, vacuoles, inclusions, and abnormal protein within the cytoplasm. Pigment accumulation or lipofuscin occurs at different rates within different areas of the brain. The large neurons of the precentral gyrus are particularly predisposed to lipofuscin deposits, which are composed of lipids, proteins, and carbohydrates. Neuromelanin, which results from peroxidation of the lipofuscin granules, also occurs in the neurons of the substantia nigra and locus coeruleus. Granulovascular degeneration, which results in the accumulation of cytoplasmic vacuoles, is common in aging. Lewy bodies, commonly found in Parkinson's disease, occur in a small number of healthy aged subjects, whereas colloid inclusions or fine granular material in the cisterns of rough endoplasmic reticulum are often found in the senescent brain. Marinesco bodies, composed of fine granules and filaments in a lattice-type network, also commonly occur with advancing age. Neurofibrillary tangles (abnormal fibrous protein accumulation in the cytoplasm), which are one of the key pathological markers of Alzheimer's disease, occur within parts of the hippocampus and amygdala, several subcortical nuclei, and some regions of the cerebral cortex (primarily in the entorhinal cortex in the temporal lobe) but with a density far less than that found in patients diagnosed with Alzheimer's disease.

A few studies have found senescent changes in the dendritic system of neurons, although this kind of analysis is sensitive to the type of fixation used to preserve the brain. These changes include a decrease in dendritic number, which usually begins with dendrites farthest away from the cell body. A decrease in the number of synaptic terminals is also seen. A proliferation of existing dendrites can be observed and is thought to be a compensatory mechanism for those that are lost. These changes can be seen after age 60. Neuroaxonal dystrophy, or enlargement of the distal ends of axons, occurs predominantly in some nuclei in the medulla and increases in frequency with age.

Neuropathologic changes that occur in the neuropil, or the area surrounding cells, include the development of neuropil threads, senile plaques, and Hirano bodies. Neuropil threads consist of the abnormal protein seen in neurofibrillary tangles, but these proteins are located in the neuronal process surrounding amyloid plaque cores and are rare in aging. Senile plaques are composed of amyloid protein, degenerating neuronal processes, and reactive glial cells. There are different types of plaques and these are classified by the organization of amyloid and the presence or absence of degenerating cell processes. In normal aging, senile plaques may exhibit dystrophic neurites, but these neurites usually lack the paired helical filaments seen in Alzheimer's disease. These plaques are commonly found in the frontal and temporal cortex and in the hippocampus. Hirano bodies are spindle or rod shaped structures found in the neuropil surrounding neurons and are commonly seen in senescent brains.

2. Neurotransmitter Changes

There are numerous substances in the brain which play a critical role in neural transmission. The class of neurotransmitters that we discuss are called neuromodulators; their effect on neurotransmission has a longer time course than classical transmitters such as glutamate and GABA, and they seem to affect the responsiveness of target neurons to other inputs. These substances, which originate from midbrain and brain stem nuclei, include acetylcholine, dopamine, serotonin, and noradrenaline. They play a central role in many of the neurological and psychiatric illnesses that are common in the elderly, including Alzheimer's disease (acetylcholine), Parkinson's disease (dopamine), and depression (serotonin).

a. Cholinergic System The NBM in the basal forebrain provides most of the cholinergic innervation of the cerebral cortex. One study found that cells within the NBM increase in size until age 60 and then begin to atrophy, particularly in posterior regions of the nucleus. Acetylcholine (ACh) is the primary neurotransmitter produced by cells within the NBM. There is little change in acetylcholinesterase content, an enzyme responsible for the breakdown of ACh, in elderly subjects. There are also minimal changes in high-affinity choline uptake, which is the rate-limiting step in ACh production. However, age-related changes in ACh receptors are observed. There are two principal cholinergic receptor types: muscarinic and nicotinic. A 10–30% reduction in muscarinic receptor density is

seen in the cerebral cortex and striatum. The cortex and hippocampus exhibit a decrease in nicotinic receptors, whereas the thalamus shows a decrease in nicotinic and an increase in muscarinic receptor density.

b. Dopaminergic System Dopamine innervation of the cortex, limbic system, and basal ganglia originates from the ventral tegmentum and the substantia nigra. There are substantial changes to neurons of the substantia nigra with age. After age 65, there is a progressive decline in cell number within this region. The remaining cells exhibit decreased nucleolar volume and mild accumulation of neurofibrillary tangles, Lewy bodies, and neuromelanin. Studies that have examined the effects of aging on dopamine have shown that the density of presynaptic D_1 receptors decreases, whereas the density of postsynaptic D_1 receptors increases in the striatum with age. The striatum also shows a decrease in pre- and postsynaptic D_2 receptor density.

c. Noradrenergic System Noradrenaline is produced by locus coeruleus neurons found in the brain stem. There is a progressive loss of noradrenergic neurons from the brain stem beginning from age 30–40. Noradrenergic neurites can be found in senile plaque formations in aging. A decrease in tyrosine hydroxylase, which is needed for the production of dopamine and noradrenaline, is also observed in elderly subjects. The cerebral cortex contains both α- and β-adrenergic receptors. Adrenergic α_2 receptors significantly decrease with age. Loss of β-adrenergic receptors occurs in a more heterogeneous fashion based on cortical region. Receptors in the frontal lobe show no decrease in number, whereas those in the precentral, temporal, occipitotemporal, and cingulate cortical regions exhibit a linear decline with age.

d. Serotonergic System The raphe nuclei in the midbrain supply the serotonergic innervation of the brain. The primary metabolite resulting from the breakdown of serotonin, 5-hydroxyindoleacetic acid, does not decline with age. However, there is a decrease in imiprimine binding with aging; the imiprimine binding site is a presynaptic marker for the reuptake of serotonin. Two types of serotonergic receptors (S_1 and S_2) have been found to decline with age. S_1 receptors demonstrate up to 70% reduction in number; S_2 receptors density is decreased by 20–50% in elderly subjects. Functional imaging studies using markers for the S_2 receptor have found decreases in receptor density in occipital, parietal, temporal, and frontal lobes of the brain. Of these regions, the most significant decreases in density are observed in frontal and temporal cortices.

C. Summary: Anatomical Changes with Age

There is a general decrease in brain weight with advancing age, and macroscopic atrophy in the form of increased sulcal and ventricular size becomes clearly evident. The volumes of a number of brain structures are reduced in old versus young brains. The most prominent of these include the hippocampal formation, the frontal lobes, and a number of subcortical structures in the basal ganglia and the nuclei that are the source of the neuromodulatory transmitters (e.g., NBM and the substantia nigra). White matter changes, which are associated with vascular problems such as hypertension, are also more common in the elderly.

At the cellular level, the aging brain exhibits many neuropathologic changes, but these changes are diffuse and relatively modest in relation to those associated with neurodegenerative diseases. Decreased overall size and volume of gray and white matter occur relatively early in the life span. Degenerative processes that affect the gray matter of the brain appear to begin later and include diffuse accumulation of abnormal products within and around neurons and changes in the ultrastructure of neuronal processes. Neuromodulatory neurotransmitter systems show declines in concentration and/or receptor densities in aging brains, but these declines are generally mild to moderate in degree.

IV. FUNCTIONAL NEUROIMAGING STUDIES OF AGING

A. Fundamentals of Functional Neuroimaging

Although other articles in this volume treat functional neuroimaging in detail, it is important to mention a few methodological issues here as they pertain to aging. The two basic types of functional neuroimaging methods are those that measure the electric or magnetic fields generated by neural activity [electroencephalography (EEG) and magnetoencephalography (MEG)] and those that measure the hemodynamic or metabolic consequences of neural activity [positron emission tomography (PET), single photon emission

computed tomography (SPECT), and functional MRI (fMRI)]. These two types differ in many ways, but they differ primarily in terms of spatial and temporal resolution and in the extent of the brain about which information is provided: PET (and SPECT; most research studies on brain aging have used PET; we will not specifically distinguish between SPECT findings and those obtained using PET) and fMRI can localize activity for most of the brain to a resolution of a few millimeters but with temporal resolution on the order of several seconds at best. EEG and MEG have a temporal resolution in the millisecond range but with poor localizability for activity of many brain regions simultaneously. PET can be used to ascertain markers for a number of neurophysiological and neurochemical processes, including glucose metabolism (most commonly using ^{18}F-fluoro-deoxyglucose as the radioligand), cerebral blood flow (often using [^{15}O]water as the radiotracer), and neuroreceptor concentration, whereas most fMRI studies provide information about the amount of blood oxygenation (oxygenated and deoxygenated blood have different magnetic susceptibilities). Blood oxygenation, cerebral blood flow, and glucose metabolism are mostly used as indirect indicators of regional neural activity.

In the next section, we review a number of studies of brain cognitive function in aging that have used either PET or fMRI. Two kinds of PET studies are discussed: (i) resting state studies, in which a subject lies in the scanner "at rest," having no specific task requirement except to remain awake, and (ii) activation studies, in which a subject in the scanner is asked to perform a set of specific tasks while the neural activity in his or her brain is measured. Although resting studies are easier to perform, two advantages of activation studies are that specific cognitive systems can be probed, and subject performance during the scanning session can be obtained and correlated with the functional neuroimaging data.

Two different data analysis strategies are used to analyze PET and fMRI data so that inferences can be made about the brain processes involved in cognition. The first is called the subtraction paradigm, which is concerned with the functional specialization of different brain areas. In functional neuroimaging studies using PET/fMRI, this paradigm is implemented in its simplest form by comparing the functional signals between two scans, each representing a different experimental condition. The locations of the major differences in signal between the two presumably delineate the brain regions differentially involved in the two conditions. For example, if the two conditions differed by the presence of an additional cognitive operation in one compared to the other, then the brain areas "activated" would, it is assumed, be associated with that cognitive operation. For aging studies, one would be interested in those brain regions that show a task-by-age interaction. In resting brain studies, signals obtained from young and old subjects are compared directly.

The second data analysis method is called the covariance paradigm. This approach aims at determining how different brain regions interact with one another during the performance of specific cognitive tasks. It does this by examining the interregional covariances in brain activity (often called the interregional functional connectivity). Functional neuroimaging methods obtain data simultaneously from multiple brain regions and therefore are ideal for use with the covariance paradigm. These two paradigms for analyzing functional neuroimaging data complement one another; both are necessary to get a clear picture of how the brain works.

The possibly confounding issue of partial volume artifacts needs to be mentioned before discussing the functional changes observed in aged brains. PET and fMRI provide images of function, not structure. As such, these functional images must be interpreted in the context of an underlying anatomical substrate. When comparing young and old subjects, the anatomical region that is the source of the functional neuroimaging signal may be reduced in size due to atrophy in old compared to young subjects. The question then arises as to whether a reduced functional signal indicates a functional deficit or simply reflects atrophy. Although attempts to correct PET and fMRI images for atrophy have been made in a few studies, it is not commonly performed due to inherent technical difficulties. Thus, most results presented in the next few sections have not addressed this issue.

B. Resting Studies

Resting studies of cerebral function generally have been performed using PET and have measured either regional cerebral glucose metabolism or regional cerebral blood flow (rCBF). Global or overall resting metabolism shows a decline of approximately 12–20% over the adult age range of 20–80 years. One study found a 12% reduction in resting metabolism after age 60. Regional assessment of metabolism shows a differential decline in glucose utilization with age.

The frontal, temporal, and parietal lobes all exhibit decreased resting metabolism, with the frontal lobes, particularly the dorsolateral cortex, showing the largest decline with age. In the temporal cortex, both lateral and medial temporal areas have been found to have resting metabolism that declines with advancing age. The largest age-related difference within the temporal lobe is in anterior temporal cortex.

Measurements of resting cerebral blood flow show a linear decline in global gray matter values beginning after age 55. An overall decline in the range of 18–28% of global flow is observed in the elderly compared to young subjects. Regional analysis reveals that association areas are affected more than primary cortical areas with age. Decreased blood flow is observed in the frontal, temporal, parietal, and occipital lobes of the brain as well as in the cingulate cortex. In the frontal cortex, a linear decrease in blood flow of 30% from ages 22 to 82 has been reported. In the temporal lobe, superior temporal cortical regions appear to be affected most.

Whether the modest declines observed in resting cerebral metabolism and blood flow represent a true functional decrease or occur because of atrophy-related partial volume effects is not known. Also, almost all the measurements discussed previously have the feature mentioned in other sections of this article—their variance increases with advancing age.

Interregional correlations of resting glucose metabolic rates have also been evaluated in young and old subjects to assess changes in brain interactions during aging. Older subjects have significantly fewer large correlations between frontal and parietal regions than do young subjects, suggesting an age-related decrease in integrated function between anterior and posterior cortex.

A number of PET studies have examined differences in neurotransmitter function between young and old subjects, with most focusing on the integrity of the dopaminergic system. Several groups have found an age-related decline in D_2 receptor (postsynaptic) availability in the striatum of approximately 8% per decade. A similar percentage decline in the amount of dopamine transporter (a presynaptic marker) has been reported. Loss of striatal presynaptic dopamine function, evaluated by measuring the uptake of a positron-labeled version of L-dopa, likewise has been found during aging. The age-related loss of these makers has been shown to correlate with motor function and also with performance on some tests sensitive to frontal lobe function. These studies thus point to a decline in nigrostriatal function with age.

C. Cognitive Activation Studies

Studies using brain activation paradigms can be divided into two domains: those examining the effects of sensory stimulation or motor function and those investigating higher level cognitive functions. Interestingly, few activation studies have been performed that examine motor performance in aging. With regard to the sensory domain, it has been shown that vibratory stimulation of either hand in healthy subjects results in increased rCBF in the contralateral primary sensorimotor and supplementary motor areas of the brain. No correlation is observed with age, suggesting that this form of response does not change in the elderly. Conversely, fMRI has been used to demonstrate that there are age-related changes in basic visual function. Using goggles that generate alternating checkerboard stimuli to examine photic stimulation in the elderly, researchers found that the primary visual cortex in aged subjects demonstrated a significantly lower response level than that seen in young subjects. Another study using PET to assess visual function in the cerebral cortex also found age-related differences. When subjects were shown visual textures made up of black-and-white check patterns, the brains of older subjects demonstrated decreased activation of visual extrastriate and temporal regions and increased activation of frontal lobe areas relative to young subjects. Together, these findings suggest that there are cortical changes in the way visual material is processed with age.

Studies have also been performed that examined differences in visual perception. When subjects were asked to identify the two identical human faces out of a choice of three faces in a display, the occipitotemporal object visual pathway of the brain was activated. Older subjects, however, exhibited slightly different regional activation of the visual cortices than did younger subjects. Similar findings were observed when subjects were required to perform a task involving location or identification of similar spatial configurations. Here, activation of the occipitoparietal spatial visual pathway was observed, with old subjects again demonstrating a different regional activation of the visual cortices, along with exhibiting additional activation of frontal lobe regions not seen in the young subjects. These results suggest that recruitment of cortical regions different from those observed in young subjects occurs with age during visual perception. Reaction time data showed that older subjects were slower in responding than young subjects on both types of tasks, although accuracy was well above chance.

An analysis of interregional functional connectivity performed on these data adds some insight into the nature of cortical recruitment. For the face-matching task, it was found in young subjects that the functional connection between ventral frontal cortex and posterior visual cortex was fairly weak. In the healthy aged, however, there was a strong functional linkage between these two brain areas, perhaps indicating greater use of the frontal cortex for strategic processing or monitoring of behavior in the old compared to the young subjects. These results suggest that changes in functional connectivity in the aged brain, a form of neuroplasticity, may be one way by which behavioral performance can be maintained.

In terms of other cognitive functions, studies have been performed that evaluate attention, executive function, and memory for both verbal and visual material. Using a visual search paradigm to assess attention, age-related differences have been demonstrated with PET in both brain activation and behavioral performance. When subjects were asked to selectively attend to a central target position in a grid of letters, no age differences were observed. However, differences were seen when subjects were required to divide their attention between different portions of the grid in search of a specific target. Although both young and old subjects exhibited activation of occipitotemporal, occipitoparietal, and frontal regions, the young group demonstrated more activation of posterior brain regions (occipitotemporal), whereas the old group showed more anterior (frontal) activation patterns. The old subjects were also slower and less accurate in choosing the correct target during the divided-attention task.

The ability to sort objects into specific categories is a measure of executive function. Brain activation patterns obtained using PET that result from the ability to sort cards based on number, color, or shape showed involvement of frontal, parietal, and occipital regions in both young and old subjects. Although similar regions were activated in both age groups, the magnitude of the increased rCBF was reduced in the old group, suggesting that there is a decrease in cortical efficiency associated with aging. The elderly again performed more slowly and made more errors than the young subjects on this task.

Several functional imaging studies have examined processes related to the ability to remember specific stimuli, such as faces or words. With regard to memory for faces, it has been shown that young subjects activate frontal cortex in the left hemisphere as well as temporal lobe regions, including the hippocampus, while encoding a series of faces. Old subjects, however, show no significant activation of these regions during this condition. When retrieving or recognizing the faces, young subjects demonstrate involvement of frontal cortex in the right hemisphere as well as parietal regions, whereas old subjects only activate frontal regions. These findings led to the conclusion that memory dysfunction seen in the elderly results from impaired encoding of material to be remembered. Another study investigated brain function during a working memory task involving face stimuli. Here, the subjects were required to remember specific faces over varying delay intervals. During this task both young and old subjects exhibited activation of the frontal cortex, which increased in magnitude in the left hemisphere with increasing delay, and of extrastriate visual cortex, which showed decreasing activation with increasing delay. However, in old subjects there was less activation overall, and they had increased activation of occipitoparietal cortex and different frontal regions than those seen in the young. Although there were subtle differences in response time and accuracy in task performance, it was suggested that the differences in activation may represent different strategies employed by the elderly in maintaining a short-term memory for faces.

Verbal material has also been used to examine memory function in the elderly. Several studies have been performed using retrieval paradigms of words from cues given in the form of three-letter word stems. It was found that explicit retrieval of words results in frontal activation in young and old subjects, but the regions differ between the two groups. One study also found medial temporal activation only in the old, whereas another found it in both groups when the task became more difficult. These findings suggest that there are functional differences in retrieving verbal information with age, especially in the frontal lobes of the brain. Decreased accuracy in the old group also supports the conclusion that retrieval processes are impaired in the elderly.

Other studies have investigated encoding and retrieval of single words or word pairs. Although the results of these studies vary to a certain degree, there are similarities among the brain activation patterns. During encoding of words or the letters that make up words, young subjects demonstrated activation of the frontal cortex in the left hemisphere in addition to occipital and temporal regions. Age-related differences in activation were observed in frontal regions of both cerebral hemispheres. During retrieval, young subjects showed involvement of the right frontal

cortex. Occipitotemporal activation was also observed in this condition. Old subjects demonstrated bilateral activation of the frontal cortex and both increased and decreased activation of occipitotemporal regions relative to the young groups. A study examining the functional connectivity of the brain during these tasks revealed that young subjects used the frontal cortices of the right and left hemispheres differentially during encoding and retrieval, whereas the old tended to use networks in both cerebral hemispheres for both encoding and retrieval. Together, these results again suggest that there is a functional reorganization of brain regions mediating memory processes with age.

D. Summary: Functional Neuroimaging Studies

Studies of resting brain function indicate that there are small reductions in both glucose metabolism and cerebral blood flow with age and an age-related reduction of nigrostriatal dopamine function. The declines in global measurements occur within the average range of 10–20% after age 60. Regional metabolism and blood flow also decrease with age. Changes have been observed most commonly in frontal, temporal, and parietal lobes of the brain. Of these regions, the frontal lobes appear to exhibit the greatest decline when assessed with either glucose or blood flow techniques.

Brain function during activation also demonstrates age-related changes. Studies of visual function show activation differences in occipital, temporal, and frontal regions. Generally, these differences are seen as decreases in activation of visual areas within the occipital and temporal lobes and increases in activation of areas in the frontal cortex. Although individual studies of memory function yield slightly different results, the similarities suggest that there is a reorganization of brain activation patterns in the elderly during memory tasks. This reorganization can been seen during tasks involving both encoding and retrieval of visual and verbal information. Some differences are observed in occipital and parietal lobes of the brain, but the most prominent age-associated differences are seen in frontal lobe activation, especially involving prefrontal cortical regions. These frontal lobe differences include increased activation as well as involvement of regions unlike that observed in young subjects. Often, in tasks in which frontal activation is predominant in one hemisphere in young subjects, the activation becomes bilateral in the elderly. Together, the activation data suggest that decreases in cortical function and subsequent recruitment of novel regions during task performance may represent some form of functional compensation in the aging brain, although perhaps not sufficient to maintain performance at the same levels of accuracy and speed as in young subjects.

V. CONCLUSIONS

Several conclusions can be drawn from our review of selective changes in neural structural and function with aging. First, as we have stressed, compared to the types of brain-related changes observed in neurological and psychiatric disorders, the alterations seen in the brain with advancing age are modest. Almost all the measures of cognition, brain structure, and brain function that we examined showed essentially the same behavior: If change was found, it became worse after approximately 60–70 years of age. However, the variability in the old increased significantly, and often there were old subjects whose values on these measures were well within the range of those of young subjects.

Although much controversy exists in the research literature, some findings seem fairly consistent. The two areas of cognition that show clear declines with advancing age are some types of memory and measures of mental flexibility (i.e., fluid intelligence). Speed of cognitive and sensorimotor processing also seems to decline with age. Areas of the brain that demonstrate structural alterations include medial temporal lobe regions such as the hippocampus, the frontal lobes, regions within the basal ganglia, and a number of subcortical nuclei, such as the nucleus basalis of Meynert and the locus coeruleus, which are the sources of neuromodulatory neurotransmitters in the cortex. These transmitters play a key role in numerous cognitive processes, including memory and attention. There also seems to be an increase in the presence of white matter abnormalities with advancing age, although how this relates to alterations in cognitive performance is unclear.

Although we did not discuss in detail brain abnormalities associated with age-related neurodegenerative disorders, the distinction between aging changes and disease-associated alterations was difficult for us to maintain. Long ago, it was thought that senescence was part of a continuum with dementia and perhaps also with movement disorder of the Parkinsonian type. About two decades ago, the view changed to the notion that healthy brain aging could be distinguished from brain diseases, particularly those that lead to dementia. That is, there are changes that

occur in the brain that are part of the aging process, but these are distinct from the changes due to specific brain disorders such as Parkinson's disease, cerebrovascular disease, and especially Alzheimer's disease. In our view, this distinction has become blurred. There are interactions between genetic risk factors for various brain disorders and environmental factors, such as diet, exercise, and education, that can modulate the neural changes associated with aging and brain disease. The estrogen example we presented earlier illustrates this point. Successful aging seems to be based on a sufficiently robust neuroplasticity that can keep dysfunction due to the aging and subclinical disease-related processes in check.

See Also the Following Articles

ALZHEIMER'S DISEASE, NEUROPSYCHOLOGY OF • BRAIN DAMAGE, RECOVERY FROM • COGNITIVE AGING • COGNITIVE REHABILITATION • DEMENTIA • SHORT-TERM MEMORY

Acknowledgments

We thank Drs. Stanley I. Rapoport, Cheryl Grady, and Helen Creasey for many useful discussions over the years about the aging brain. We also thank Drs. Catherine Connolly and Giampiero Giovacchini for critically reading the manuscript.

Suggested Reading

Albert, M. S. (1988). Cognitive function. In *Geriatric Neuropsychology* (M. S. Albert and M. B. Moss, Eds.), pp. 33–56. Guilford, New York.

Creasey, H., and Rapoport, S. I. (1985). The aging human brain. *Ann. Neurol.* **17,** 2–10.

de Groot, J. C., de Leeuw, F. E., and Breteler, M. M. B. (1998). Cognitive correlates of cerebral white matter changes. *J. Neural. Transm.* **53**(Suppl.), 41–67.

Drachman, D. A. (1997). Aging and the brain: A new frontier. *Ann. Neurol.* **42,** 819–828.

Giannakopoulos, P., Hof, P. R., Michel, J.-P., Guimon, J., and Bouras, C. (1997). Cerebral cortex pathology in aging and Alzheimer's disease: A quantitative survey of large hospital-based geriatric and psychiatric cohorts. *Brain Res. Rev.* **25,** 217–245.

Grady, C. L. (1998). Brain imaging and age-related changes in cognition. *Exp. Gerontol.* **33,** 661–673.

Horwitz, B. (1998). Using functional brain imaging to understand human cognition. *Complexity* **3,** 39–52.

Kausler, D. H. (1991). *Experimental Psychology, Cognition, and Human Aging*. Springer-Verlag, New York.

Peters, A., Morrison, J. H., Rosene, D. L., and Hyman, B. T. (1998). Are neurons lost from the primate cerebral cortex during normal aging? *Cereb. Cortex* **8,** 295–300.

Pietrini, P., and Rapoport, S. I. (1994). Functional neuroimaging: Positron-emission tomography in the study of cerebral blood flow and glucose utilization in human subjects at different ages. In *The American Psychiatric Press Textbook of Geriatric Neuropsychiatry* (C. E. Coffey and J. L. Cummings, Eds.), pp. 195–213. American Psychiatric Press, Wastington, DC.

Powers, R. E. (1994). Neurobiology of aging. In *The American Psychiatric Press Textbook of Geriatric Neuropsychiatry* (C. E. Coffey and J. L. Cummings, Eds.), pp. 35–69. American Psychiatric Press, Washington, DC.

Schochet, S. S. (1988). Neuropathology of aging. *Neurol. Clin. North Am.* **16,** 569–580.

Smith, C. D. (1996). Quantitative computed tomography and magnetic resonance imaging in aging and Alzheimer's disease. *J. Neuroimaging* **6,** 44–53.

Agnosia

JONATHAN J. MAROTTA and MARLENE BEHRMANN
Carnegie Mellon University

GLOSSARY

alexia An acquired condition, usually as a result of brain damage (such as follows strokes in adults), marked by an impairment in reading, in which reasonable vision, intelligence, and most language functions other than reading remain intact.

apperceptive agnosia A form of visual agnosia in which a person cannot reliably name, match, or discriminate visually presented objects, despite adequate elementary visual function (visual fields, acuity, and color vision).

associative agnosia A form of visual agnosia in which a person cannot use the derived perceptual representation to access stored knowledge of the object's functions and associations but is able to copy and match the drawing even though unable to identify it.

Balint's syndrome Agnosic syndrome that results from large bilateral parietal lesions and is composed of three deficits: (i) paralysis of eye fixation with inability to look voluntarily into the peripheral visual field, (ii) optic ataxia, and (iii) disturbance of visual attention such that there is neglect of the peripheral field.

dorsal simultanagnosia An inability to detect more than one object at a time, with difficulty shifting attention from one object to another.

dorsal stream The stream of cortical visual projections from primary visual cortex to posterior parietal cortex, concerned primarily with the visual control of action and representation of spatial information.

inferotemporal cortex Inferior surface of the temporal lobe that is particularly important for object recognition.

Klüver–Bucy syndrome A group of impairments, including visual agnosia, resulting from bilateral damage to the temporal lobes.

optic aphasia A condition in which a person cannot name a visually presented object, despite being able to indicate the identity of the object through gesture and to sort the visual stimuli into categories.

prosopagnosia A form of visual agnosia in which a person cannot recognize faces, despite adequate elementary visual function (visual fields, acuity, and color vision).

ventral simultanagnosia A reduction in the ability to rapidly recognize multiple visual stimuli, such that recognition proceeds in a part-by-part fashion.

ventral stream The stream of cortical visual projections from primary visual cortex to the inferotemporal cortex, concerned primarily with representing the identity of stimuli by such characteristics as shape and color.

Visual agnosia is a disorder of recognition confined to the visual realm, in which a person cannot arrive at the meaning of some or all categories of previously known visual stimuli despite normal or near-normal visual perception and intact alertness, intelligence, and language. This article takes a multidisciplinary approach in discussing this impairment and considers clinical and neurological studies in humans as well as neurophysiological data in nonhuman primates.

I. CASE STUDIES

JW is a relatively young man, in his early forties, who, despite many preserved cognitive abilities, fails to recognize many common objects. In August 1992, JW suffered a severe cardiac event while exercising and was subsequently anoxic. A computed tomography (CT) scan revealed multiple hypodensities in both occipital lobes with minor hypodensities in his right parietal lobe. Although JW has normal visual acuity as well as intact color and motion perception, Behrmann and colleagues have shown that he recognizes approximately 20% of black-and-white line drawings and a slightly higher percentage of color pictures. He is almost totally unable to recognize photographs of famous people. He is poor at copying simple line drawings presented to him (Fig. 1), at matching rectangles and squares of various dimensions, at simple shape detection (e.g., deciding that an "X" is present among a background of visual noise), and even at detecting symmetry in a visual image. Despite these impairments, he is able to recognize objects well from tactile/haptic input and from definitions that are read to him. These findings suggest that his long-term knowledge of objects is preserved. This is further confirmed by his ability to generate visual images in his "mind's eye" and to describe those in detail. Needless to say, this impairment significantly limits his ability to interact with objects and his world. Whereas JW was the owner of a hardware computer company (and had a master's degree in computer science), currently he works as a volunteer and provides instruction on computer use to people who are blind.

CK, like JW, is impaired at recognizing objects and has been studied extensively by Behrmann, Moscovitch, and Winocur. CK sustained brain damage in a motor vehicle accident in 1988; he was struck on the head by the side mirror of a truck while he was jogging. Except for a hint of bilateral thinning in the occipitotemporal region, no obvious circumscribed lesion is revealed on magnetic resonance imaging (MRI) or CT scan. This may not be surprising given that his lesion was sustained via a closed head injury which often results in shearing of axons or more microscopic neuronal damage. Despite his deficits, CK functions well in his life; he has a responsible managerial job and makes use of sophisticated technology that allows him to translate written text into auditory output.

When asked to identify line drawings, CK misrecognized a candle as a salt shaker, a tennis racquet as a fencer's mask, and a dart as a feather duster, presumably because of the feathers on the end (Fig. 2). As

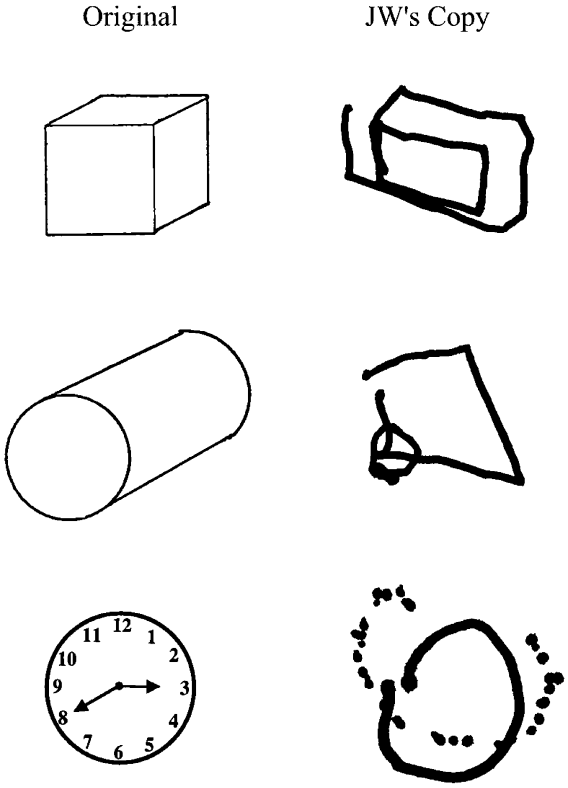


Figure 1 Patient JW's copies of simple line drawings.

illustrated by these examples, CK, like JW, is clearly not blind. However, despite his intact visual acuity, he fails to recognize even familiar and common visually presented objects. This deficit holds irrespective of whether the objects are drawn in black-and-white on a piece of paper or whether they are shown in slides or even as real three-dimensional objects, although the addition of information such as color and surface texture does assist recognition to some extent.

CK, like JW, can also use tactile/haptic information to recognize objects; he was perfectly able to recognize a padlock and a paper clip by touch alone. CK can also provide detailed definitions for an object whose name is presented to him verbally; for example, he defined a pipe as "a long cylindrical hollow object to convey liquid or gas" and a card of matches as "a cardboard container containing long sticks or matches which are struck against cordite." These definitions clearly demonstrate that his deficit is not attributable to a failure to name objects nor a loss of semantic knowledge.

CK is unable to read and, although he writes flawlessly, he cannot read his own writing presented to him at a later point in time. CK's hobbies have also

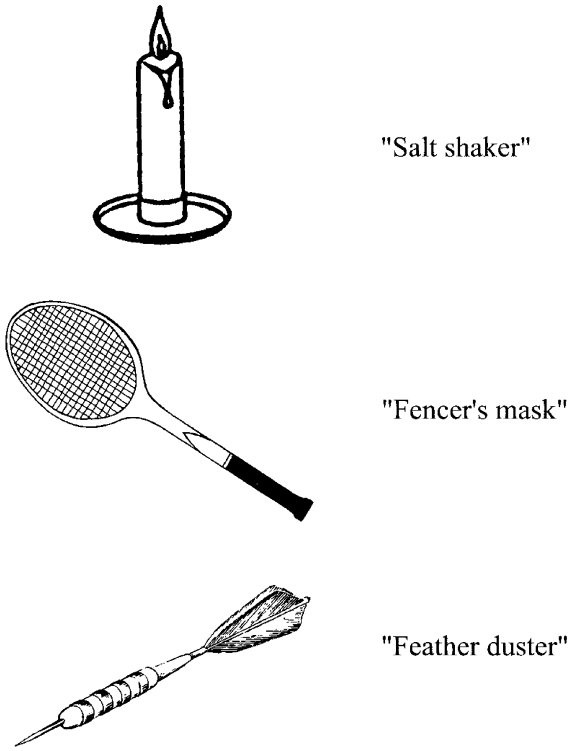


Figure 2 Line drawings misnamed by patient CK.

been affected; he is no longer able to design complex configurations of his large plastic soldier collection or visually differentiate airplanes, a domain in which he had rather extensive knowledge premorbidly.

II. BACKGROUND

Despite the behavioral differences between JW and CK, they have a dramatic deficit: They are unable to recognize even common, familiar objects—a disorder termed "agnosia" by Sigmund Freud (coined from the Greek "without knowledge"). Visual agnosia is a disorder of recognition, in which a person cannot arrive at the meaning of some or all categories of previously known visual stimuli, despite normal or near-normal visual perception and intact alertness, intelligence, and language. Despite the visual recognition problems associated with agnosia, there is normal recognition of objects through modalities other than vision (touch, auditory, and verbal definitions or description of their function), which suggests that the deficit is not simply a difficulty in retrieving names or in accessing the necessary semantic information. Visual recognition has been more extensively studied than recognition in other modalities, although similar deficits have been observed in patients with auditory (auditory agnosia) or tactile (tactile agnosia) deficits.

The traditional view of agnosia as a specific disorder of recognition has undergone considerable challenge in the past, with critics contending that all visual agnosias can be explained by a subtle alteration in perceptual functions likely accompanied by a generalized intellectual deterioration. Despite this early skepticism, there is now widespread acceptance of this disorder as a legitimate entity and detailed case studies have been concerned with characterizing both the underlying mechanisms that give rise to this disorder and the overt behaviors.

Lissauer was the first to classify visual object agnosia into two broad categories: apperceptive "mindblindness" and associative mindblindness. These impairments were evaluated by requiring patients to (i) describe the formal features of a pattern, (ii) reproduce it by drawing, and (iii) recognize it among similar alternatives. Using Lissauer's classifications, a person with apperceptive agnosia is assumed to be impaired at constructing a perceptual representation from vision and subsequently is unable to copy, match, or identify a drawing. In contrast, a person with associative agnosia is one who cannot use the derived perceptual representation to access stored knowledge of the object's functions and associations but is able to copy and match the drawing even though he or she is unable to identify it.

Recent neuropsychological accounts by Humphreys and Riddoch as well as by Warrington and colleagues and computational accounts such as that of Marr and colleagues have sought to extend Lissauer's dichotomy for two reasons. The first reason is the growing understanding that visual object recognition comprises many distinct steps not captured by the simple dichotomy. For example, it has been suggested that apperceptive processes include encoding the primitive dimensions of shape and segmentation of figure from ground. Associative processes may also be subdivided to include access to stored visual knowledge of objects, followed by access to stored associative and functional (semantic) knowledge from the description derived from the image. The second reason for further differentiation of the underlying processes, and the lesion types, derives more fine-grained neuropsychological analysis. One such example is of patients who show impaired access to knowledge of associative and functional properties of the object but have well-preserved understanding of the object's shape, as

reflected in a high-complexity object decision task (differentiating real objects from novel objects that are composed of parts of real objects). Other patients perform relatively poorly at object decision but are still able to carry out many high-level perceptual tasks, such as matching objects across different viewpoints and sorting pictures into basic categories. These developments have forced a further refinement of our understanding of visual processing and the types of breakdown that are possible. Despite the simplicity of Lissauer's dichotomy and its clear inadequacy, it provides a coarse framework that has proved useful in describing agnosia, as illustrated in the book *Visual Agnosia* by Farah. Following Farah, we adopt this dichotomy as a starting point and describe these two forms of agnosia, we also provide a detailed discussion of the patients described previously and the implications of such disorders for our further understanding of visual object recognition. Before we continue our exploration of these types of agnosia, however, we first identify the underlying neuromechanisms responsible for this visual perceptual processing.

III. NEUROANATOMY

Milner and Goodale proposed that the two prominent cortical visual pathways that have been identified in the primate brain (after Mishkin, Ungerleider, and Macko) are each involved in two very different processes. The underlying mechanisms in the ventral stream, which projects from primary visual cortex to the inferotemporal cortex (via many routes involving areas V2, the ventral portion of V3, V4, and TEO) are thought to be involved in visual perception, whereas the dorsal stream, which projects from primary visual cortex (and the superior colliculus via the pulvinar) to the posterior parietal cortex is thought to be involved in the visual control of action (Fig. 3). Both streams are thought to process information about object features and their spatial locations, but each stream uses this visual information in different ways. The transformations carried out by the dorsal stream deal with moment-to-moment information about the location and orientation of objects and thereby mediate the visual control of skilled actions, such as manual prehension, directed at those objects. In contrast, visual information is transformed in the ventral stream to deliver the enduring characteristics of objects and their relations, permitting the formation of long-term perceptual representations of the world. Such representations play an essential role in the recognition and identification of objects and enable us to classify objects and events, attach meaning and significance to them, and establish their causal relations. Such operations are essential for accumulating a knowledge base about the world.

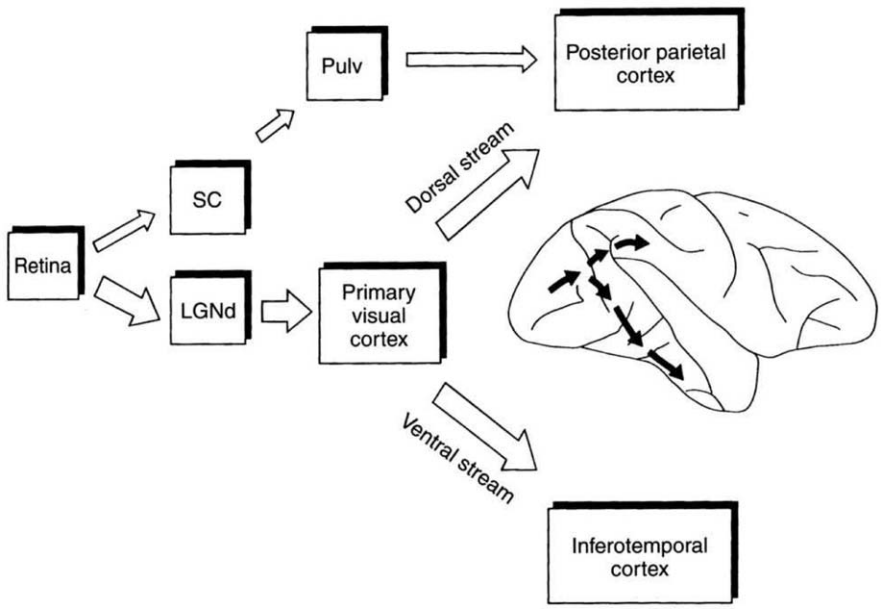


Figure 3 Diagram of the major routes leading from the retina into the dorsal and ventral streams. LGNd, lateral geniculate nucleus, pars dorsalis; Pulv, pulvinar; SC, superior colliculus. Reprinted from *Current Biology* **4**(7), Goodale, M. A., Meenan, J. P., Bulthoff, H. H., Nicolle, D. A., Murphy, K. J., and Racicot, C. L., Separate neural pathways for the visual analysis of object shape in perception and prehension, pp. 604–610, copyright 1994, with permission from Elsevier Science.

Many of the cells in inferotemporal cortex, the terminus of the ventral stream, respond best to complex visual stimuli, such as hands and faces; in particular, the more anterior parts of the inferotemporal cortex are remarkably selective in their responses to object attributes. The receptive field of virtually every cell in the inferotemporal (IT) cortex, a complex of areas lying ventrally below the superior temporal sulcus, including various subdivisions of area TE along with area TEO, includes the foveal region, where fine discriminations are made. These cells also have large receptive fields that allow for generalization across the entire visual field and for coding the intrinsic features of an object independent of its location. The critical features that activate cells in the anterior IT cortex are moderately complex and can be thought of as partial features common to images of several different natural objects. There are also neurons in the IT cortex that demonstrate properties consistent with object constancy in that they remain selectively responsive to a visual stimulus despite changes in stimulus viewpoint, retinal image size, or even color. Thus, the ventral stream is uniquely set up to process visual information into perceptual representations to which meaning and significance can be attached and stored. Damage to the ventral stream is believed to cause the disturbances of object recognition that are characteristic of visual agnosia.

Evidence for this derives from nonhuman primate work, in which large bilateral resections of the temporal lobe result in a form of visual agnosia, Klüver–Bucy syndrome. Lesions of the IT cortex impaired the monkey's ability to identify objects when the discriminations required use of color, pattern, or shape. These monkeys had difficulty using vision to learn associations with objects and could no longer recognize objects or distinguish between objects on the basis of their visual dimensions. They were unable to distinguish food from nonfood objects using vision alone and were unable to learn new visual discriminations between patterns for food reward. Although they incessantly examined all objects in sight, these animals recognized very little and often picked up the same item repeatedly. Klüver–Bucy syndrome can also be achieved with just the removal of IT but, like human visual agnosia, the IT monkey's recognition deficits cannot be explained by "low-level" sensory impairments since large bilateral lesions of IT have been found to have no residual effect on visual acuity.

Recent functional neuroimaging studies of regional blood flow in normal human subjects have revealed many different visual areas beyond primary visual cortex that appear to correspond to those in the ventral stream of the monkey brain that are specialized for the processing of color, texture, and form differences of objects. These studies have shown that face-matching tasks involve the occipitotemporal regions, detection of shape activates regions along the superior temporal sulcus, and the ventral region of the temporal lobe, and the perception of color is associated with activation of the lingual gyrus (V4).

IV. APPERCEPTIVE AGNOSIA

Individuals with apperceptive agnosia, such as patient JW, have profound difficulty recognizing and naming line drawings; their ability to recognize, copy, or match simple shapes as well as more complex objects is severely impaired. However, their elementary visual functions, such as acuity, brightness discrimination, and color vision, are relatively preserved, along with reasonable sensory and semantic memory functioning in the visual domain. These patients have normal visual fields and can maintain fixation on a visual target. The fundamental deficit involves an inability to process features, such that they are not fully available for developing a percept of the overall structure of an object.

One of the classical cases of apperceptive agnosia, described by Benson and Greenberg, was thought to be blind for several months following carbon monoxide-induced anoxia until he was seen successfully negotiating his wheelchair down a passage. Testing revealed that his fields were full to a 3-mm stimulus, that he could reach accurately for fine threads placed on a piece of paper and detect small changes in size, luminance, and wavelength, and that he was aware of small movements. Despite these fundamental abilities, he was unable to recognize objects, letters, or numbers and was unable to discriminate between any visual stimuli that differed only in shape.

Even though recognition of real objects is also impaired in these individuals, it is often better than recognition of line drawings; identifications of objects are typically inferences, made by piecing together color, size, texture, and reflectance clues. These individuals can often make accurate guesses about the nature of objects from such cues, such as the shininess of the glass and metal on a salt shaker or the color of an apple. A striking feature of this disorder is that many patients spontaneously use quite laborious and time-consuming tracing strategies of the hand or

head to aid in the recognition of visual objects. These strategies, although helpful, may not always produce an accurate result because one needs a reasonably good visual image in the first place for the purposes of tracing.

Apperceptive agnosia corresponds to the breakdown at the stage at which the sensory features of the stimulus are processed and its structural description is achieved—a relatively early stage of the visual recognition networks in the human equivalent of the ventral stream. The deficit appears to be at the level of shape or form discrimination. Some apperceptive agnosic patients are more impaired at perceiving curved than straight lines. JW, for example, is poor at deciding whether two line features are the same or different unless their orientations are very different. He also does not show "popout" of a target that differs from the background distractors if the difference is one of curvature or orientation (unless the differences are very great). Some patients may also fail to achieve perceptual constancy, interpreting a circle as an ellipse.

Interestingly, at least one apperceptive agnosic patient, DF, reported in the literature by Milner and Goodale appears to have implicit knowledge of object attributes that is not available for explicit report. As a result of carbon monoxide-induced anoxia, DF sustained damage to her occipital lobes bilaterally that extends into ventral occipital and dorsal occipitoparietal regions, while largely sparing primary visual cortex. Even though DF's "low-level" visual abilities are reasonably intact, she can no longer recognize common objects on the basis of their form or even the simplest of geometric shapes. Nevertheless, despite her profound inability to perceive the size, shape, and orientation of visual objects, DF can direct accurate and well-formed grasping movements, indistinguishable from those shown by normal subjects, toward the very same objects she cannot identify or discriminate. It has been argued that this intact visuomotor function is mediated by relatively intact parietofrontal cortical mechanisms (accessed via the dorsal stream) in DF, which continue to operate efficiently despite severely damaged occipitotemporal (ventral stream) structures. At this point, it is worth noting that DF also appears to have implicit knowledge of visual attributes even though she appears not to have this information available to her when tested directly. The critical evidence comes from studies that show that DF is influenced by the McCullough effect. This is a color aftereffect that is contingent on the orientation of grating patterns. When shown white-and-black line gratings in horizontal or vertical orientations, DF is poor at reporting orientation explicitly. However, after she was adapted to a green-and-black vertical grating alternating with a red-and-black horizontal grating, she reported seeing color on white-and-black gratings with the horizontal subcomponent appearing greenish and the vertical component appearing pinkish. As in control subjects, the effect was strongly dependent on the congruence of the angles in the testing and adaptation phase. In a follow-up, DF revealed a preserved McCullough effect with oblique gratings, indicating more fine orientation discrimination ability than simply vertical and horizontal.

The neurological damage in apperceptive agnosia tends to be diffuse and widespread and can involve damage to the posterior regions of the cerebral hemispheres, involving occipital, parietal, or posterior temporal regions bilaterally. This damage is often the result of cerebral anoxia, where a lack of oxygen to the brain produces neuronal death in "watershed" regions or regions lying in the border areas between territories of different arterial systems. Carbon monoxide-induced anoxia not only produces multifocal disseminated lesions but also affects the interlaminar connections between neurons. Mercury poisoning, which is also known to give rise to apperceptive agnosia, affects the white matter, thereby compromising connections between neurons rather than the neurons themselves.

V. SIMULTANAGNOSIA

A person with simultanagnosia can perceive the basic shape of an object but is unable to perceive more than one object, or part of an object, at a time. Thus, these patients appear to have limited ability to process visual information in parallel, although they are relatively good at identifying single objects. Farah distinguished between two forms of simultanagnosia according to whether the patients had lesions affecting the dorsal or ventral visual stream. Each is discussed in turn here.

Although a person with dorsal simultanagnosia is able to recognize most objects, he or she generally cannot process more than one at a time, even when the objects occupy the same region of space. These individuals often have counting deficits and their descriptions of complex scenes are slow and fragmentary. The underlying impairment in dorsal simultanagnosia appears to be a disorder of visual attention so severe that these individuals cannot explicitly report perceiving the unattended objects. Dorsal simultanagnosia is often observed in the context of Baliant's

syndrome, is accompanied by oculomotor deficits and optic ataxia, and results from a bilateral parietooccipital lesion.

Individuals with apperceptive agnosia and those with dorsal simultanagnosia share many characteristics. In some cases, they may act effectively blind, being unable to negotiate visual environments of any complexity, and their perception appears to be piecemeal and confined to a local part or region of the visual field. The piecemeal nature of their perception, however, differs in significant ways. In apperceptive agnosia, only very local contour is perceived, whereas in dorsal simultanagnosia whole shapes are perceived, but only one at a time. Individuals with apperceptive agnosia use color, size, and texture to guess at objects but cannot use shape information. In contrast, individuals with dorsal simultanagnosia have intact shape perception. In contrast to apperceptive agnosia, the deficit in dorsal simultanagnosia appears to be attention related rather than shape related.

A person with ventral simultanagnosia usually has a lesion to the left inferior temporooccipital region. Although such a patient is generally able to recognize a single object, he or she does poorly with multiple objects and with single complex objects, irrespective of their size. Although they cannot recognize multiple objects, they differ from individuals with dorsal simultanagnosia in that they can perceive multiple objects. These individuals can count scattered dots and, if given sufficient time, can also recognize multiple objects. They respond slowly and often describe explicitly individual elements of the picture without appreciating the whole scene. This is also true in reading, and these patients are classified as letter-by-letter readers because they only recognize one letter of a word at a time, and hence show a linear relationship between reading speed and the number of letters in a word.

A recent reconceptualization of simultanagnosia by Humphreys and colleagues suggests that the two different forms are well characterized in terms of impairments in constructing different forms of spatial representations. Although those with ventral lesions are limited in the number of parts of an object they can represent, those with dorsal lesions are limited in the number of separate objects they can represent.

VI. HIGHER ORDER APPERCEPTIVE DEFICITS

Many patients have been identified who, although still classified as having apperceptive agnosia, appear to have some residual visual processing ability and can copy and match objects to some degree. These patients have object recognition difficulty under challenging conditions and do somewhat better in more optimal conditions. For example, they are impaired at recognizing objects under poor lighting conditions when shadows are cast, creating misleading contours. An additional manipulation that proves difficult for these patients is recognition of foreshortened or degraded objects. They are also poor at recognizing objects from unusual viewpoints or unconventional angles relative to more standard viewpoints. The classification of these patients is unclear because there is ongoing debate regarding the mechanisms that give rise to such deficits and whether, indeed, these deficits arise from a common source. Warrington and colleagues argued that the failure to identify objects from unusual but not conventional views is consistent with a deficit that occurs once sensory information has been processed and is attributable to a problem in categorizing two instances of the same stimulus as identical. Some patients appear to be impaired at deriving viewpoint-independent representations, despite the fact that they are able to construct a viewpoint-dependent representation. This distinction between viewpoint-dependent and -independent representation parallels the distinction made by Marr in his well-known theory of vision.

VII. ASSOCIATIVE AGNOSIA

Unlike apperceptive agnosia, a person with associative agnosia can make recognizable copies of a stimulus that he or she may not recognize subsequently and can also successfully perform matching tasks. Teuber elegantly referred to this deficit as "perception stripped of meaning." As in the case of apperceptive agnosia, recognition is influenced by the quality of the stimulus and performance on three-dimensional objects is better than on photographs, and performance on photographs is better than on line drawings. The recognition deficit appears to result from defective activation of information pertinent to a given stimulus. There is a failure of the structured perception to activate the network of stored knowledge about the functional, contextual, and categorical properties of objects that permit their identification. In effect, this is a deficit in memory that affects not only past knowledge about the object but also the acquisition of new knowledge. Unlike individuals with apperceptive agnosia, who guess at object identity based on color and

texture cues, people with associative agnosia can make use of shape information. When these individuals make mistakes in object identification, it is often by naming an object that is similar in shape to the stimulus. For example, FZ, a patient of Levine, misidentified a drawing of a baseball bat several times. Interestingly, his answer differed on each occasion, referring to it as a paddle, a knife, or a thermometer.

Although these patients can copy drawings well, the drawings are not necessarily normal; the end product might be a fairly good rendition of the target but the drawing process is slow and slavish and they can lose their place if they take their pen off the paper since they do not grasp component shapes. As can be seen in Fig. 4, a copy of a geometric configuration by patient CK, an individual with associative agnosia, was reasonably good, although the process by which he copied indicates a failure to bind the contours into meaningful wholes.

One of the important claims of associative agnosia is that perception is intact and it is meaning that is inaccessible. Much effort has been directed at evaluating this claim and the general finding is that even patients with associative agnosia have some form of visual impairment. For example, LH, a well-known and thoroughly documented agnosic patient studied by Levine and Calvanio, was moderately impaired on several tests of perception. He was considerably slower than normal subjects in making rapid comparisons between shapes or in searching for a prespecified target figure. His performance was also poor on tasks that required him to identify letters that were fragmented or degraded by visual noise, relative to control subjects. Based on the findings from LH and other associative agnosic patients, it is clear that their perception is not normal. Although it may be considerably better than that of apperceptive agnosic patients, it is still impaired to some extent.

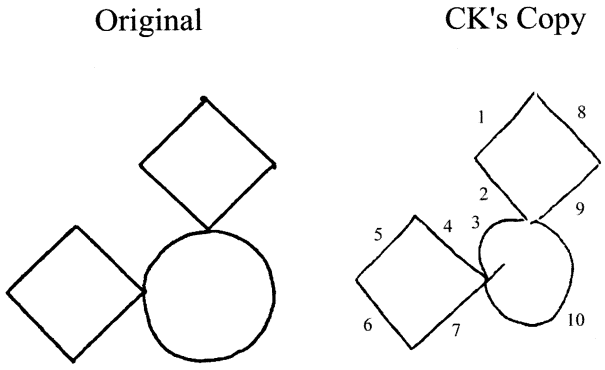


Figure 4 Patient CK's copy of a geometric configuration. The numbers assigned to his copy indicate the order in which the lines were drawn and show that he copies in a very literal fashion, failing to integrate lines 1, 2, 8, and 9 into a single shape.

The brain damage in associative agnosia is more localized than in apperceptive agnosia. Some cases appear to involve only unilateral damage to the occipital lobe and bordering posterior temporal or parietal lobe. The lesions are often more circumscribed, sometimes involving the left inferior longitudinal fasciculus, which connects fusiform gyrus to temporal structures, or the bilateral posterior hemispheric areas in posterior cranial artery territory.

VIII. INTEGRATIVE AGNOSIA

Lissauer's dichotomy makes provision for two main stages of processing, apperception and association. Object recognition, however, involves more than matching stimuli coded in terms of primitive features such as line orientation to stored knowledge. Instead, the spatial relations between the lines and features need to be coded, the object needs to be segregated from its ground, and parts of an object need to be related and integrated. These processes are typically thought of as serving intermediate-level vision. There have been several recent detailed reports of patients with deficits due to poor perceptual integration of form information ("integrative agnosia"). Patient HJA, studied by Riddoch and Humphreys, appears to oversegment identify objects piecemeal. For example, when presented with a paintbrush, HJA responded that "it appears to have two things close together or else you would have told me." CK's descriptions of errors reveal a similar pattern; he is able to perceive and report some part of the image but not the whole. Indeed, CK appears to oversegment the image, as illustrated in his copying performance. Whether or not integrative agnosia might previously have been considered an apperceptive or associative form of agnosia is difficult to determine. On the one hand, these patients appear not to be able to exploit Gestalt grouping principles, in which case the problem is closer to that of apperceptive agnosia. On the other hand, these patients perform well on standardized testing of perceptual processes, suggesting that they are more akin to associative agnosic patients. For example, these patients can discriminate between Efron shapes (squares and rectangles that have the same surface area but vary in aspect ratio) and can make orientation and

size matching judgments at a normal level. Given the uncertainty of the classification, a separate category is obviously warranted.

The hallmark of integrative agnosia is an impairment in integrating elements, which are well detected, into a perceptual whole. For example, these patients do more poorly at recognizing overlapping objects than when the same objects presented nonoverlapping, presumably because of the difficulty in segregating and assigning boundaries in the former case. Counterintuitively, these patients perform better when the stimuli are silhouettes rather than line drawings. This finding suggests that the lines that usually enhance perception and add information to the display disrupted the patients' ability to recognize objects. This is likely a consequence of their inability to group elements into a whole and this process is less taxing in the case of silhouettes. These patients also do more poorly when the exposure duration of stimuli is limited. Finally, the performance of the patients improves under limited exposure duration. This is not surprising because it has been suggested that the segmental visual processing is done in a serial fashion and limiting the exposure of the items likely affects this adversely.

Because these patients are impaired at grouping, they appear to be oversensitive to segmentation cues and to parse stimuli inappropriately into parts. In fact, grouping and segmentation are likely two sides of the same coin, with the grouping impairment leading to piecemeal and fragmented image processing. The integrative agnosia deficit appears to affect a stage of visual coding intermediate between basic shape coding and visual access to memory representations, concerned with parallel perceptual grouping and the integration of parts into wholes. It is revealed most strikingly under conditions when segmentation or grouping are stressed.

IX. OPTIC APHASIA

Like agnosic patients, patients with optic aphasia have a modality-specific (visual) recognition deficit and can recognize objects from both auditory and tactile presentation. The critical distinction between agnosia and optic aphasia, however, is that optic aphasic patients can recognize objects. This is evidenced by their nonverbal identifications (e.g., through gesture) and their ability to sort a visual stimulus with other stimuli of the same category. Additionally, these patients are also not particularly sensitive to the visual quality of the stimulus and performance is approximately equivalent for object, photographs, and line drawings, unlike the case for associative agnosic patients. Another important distinction concerns the error types. The errors made by patients with optic aphasia are usually semantic in nature and hardly ever visual, whereas in associative agnosia the errors are primarily visual. One of the best studied patients by Lhermitte and Beauvois, Jules F., when shown a picture of a boot correctly gestured pulling on a boot but called it a hat. Optic aphasia patients also appear to make many perseverative errors, reproducing the error response from previous trials and sometimes producing "conduites d'approache" or progressive approximation to the target.

X. CATEGORY-SPECIFIC VISUAL AGNOSIA

For some patients with associative agnosia, recognition is not equally impaired for all categories of objects. Category-specific visual agnosia (CSVA) is a deficit in which the boundary between impaired and intact recognition can be approximately defined along the semantic criterion of biological vs nonbiological objects. In other words, these patients can show severely impaired visual recognition of objects from biological categories while recognition of most other categories is largely spared. For example, a patient may be able to recognize all manners of tools or artifacts but show marked difficulties in recognizing even the most common fruits or vegetables. CSVA is believed to be a semantic disorder, in which patients have problems associating the view of an object, in a specific category, with stored knowledge of its identity. The mechanisms underlying visual perception do not appear to have access to the semantic knowledge of certain categories of objects.

It has also been proposed that this dissociation may be the result of the recognition of living things depending on some specialized neural mechanisms that are not needed for the recognition of nonliving things. Evidence for this derives from the findings that CSVA for biological objects usually follows inferior-temporal damage. Moreover, recent studies have found that defective recognition of persons was associated with damage to right temporal regions, defective recognition of animals was associated with damage to the right mesial occipital/ventral temporal region and left mesial occipital region, and defective

recognition of tools was associated with damage in the occipital–temporal–parietal junction of the left hemisphere. As we discuss later, these studies have helped to reveal the extent to which there is modular organization in the visual system.

XI. RELATIONSHIP BETWEEN VISUAL OBJECT AGNOSIA AND WORD AND FACE RECOGNITION

One of the interesting recent developments in our investigations of object agnosia concerns different forms of category specificity, but here the category refers to different forms of visual stimulus recognition, such as face and word recognition. The critical issue is whether agnosia can be restricted to object recognition or whether it is reflects a broader form of visual impairment. In an extensive review of the literature, Farah suggested that the latter is more correct and that because visual recognition procedures for objects, words, and faces are not neurally separated, not all pure forms of visual deficit are possible. This argument was based on the fact that some but not all patterns of dissociation have been observed between these three classes of stimuli. In particular, Farah argues that there have been no convincing reports of patients with visual object agnosia without alexia or prosopagnosia or with prosopagnosia and alexia without visual object agnosia. The failure to find these two patterns of deficit has been taken to suggest that, instead of there being independent and separate mechanisms subserving objects, words, and faces, there may be a continuum of recognition processes. At one end of this continuum is a more holistic or gestalt form of processing that is optimized for processing nondecomposable perceptual wholes, such as faces, and at the other end is a process that is optimized for processing multiple perceptual parts, such as letters of words. Object recognition may be mediated by either process depending on the nature of the stimulus and its perceptual characteristics. By this account, patients may be selectively impaired as a consequence of damage to one of these two processes. Thus, in its pure form, damage to the more holistic process will result in prosopagnosia in isolation, whereas damage to the more part-based processes will result in alexia in isolation.

An obvious claim of this account is that it should not be possible to observe a patient for whom the recognition of objects is impaired, in isolation, given that object recognition is subserved by one of the two other processes. Despite this interesting hypothesis, there have been several recent case studies that challenge it. Thus, for example, there have been several detailed studies of patients who have a selective deficit in object recognition with retained face and word recognition. The presence of such a pattern undermines the two-process account of visual recognition and is more consistent with a view in which there is neural differentiation between all types of visual stimuli. Whether this differentiation refers to the fact that different mechanisms are involved in encoding the three stimulus types or in accessing their stored knowledge remains unclear. The proposal that the three types of visual stimuli are differentiated to some extent is generally (although not perfectly) consistent with recent functional neuroimaging data that shows that different brain areas are activated for the different stimulus types. Thus, for example, word recognition is associated with an increase in cortical activation in the left medial extrastriate region, whereas face recognition is associated with increased activation in the right fusiform gyrus. Object recognition is a little more problematic. Although enhanced activity is observed in a host of regions in the left hemisphere and some in the right hemisphere, some of these activations appear to overlap those associated with face recognition and the extent to which there is some sharing of mechanisms for faces and objects remains controversial.

XII. AGNOSIA AND ACTION

Previously we discussed patient DF, who developed a severe form of apperceptive agnosia following carbon monoxide-induced anoxia. Although DF's visual system is unable to use shape information to make perceptual judgments and discriminations about an object, she is able to use the same information to accurately guide her prehensile movements to those same targets. For example, even though DF is unable to discriminate solid blocks of differing dimensions, she accurately scales her grasp when picking up the blocks, opening her hand wider for larger blocks than she does for smaller ones, just as people with normal vision do. DF also rotates her hand and wrist appropriately when reaching out to objects in different orientations, despite being unable to describe or distinguish the size, orientation, and shape of the different objects. It appears that although the perceptual mechanisms in DF's damaged ventral stream can no longer deliver any perceptual information about the

size, orientation, and shape of objects she is viewing, the visuomotor mechanisms in her intact dorsal stream that control the programming and execution of visually guided movements remain sensitive to these same object features.

Although the discussion of the dorsal and ventral streams in this article has emphasized their separate roles, there are many connections between the two streams and thus the opportunity for "cross-talk." Recent investigations have shed light on the role that the communication between these two streams plays in object recognition. In some cases of associative agnosia, it has been reported that the ability to identify actions and to recall gestures appropriate to objects could play a significant role in preserving recognition of certain objects. Sirigu suggested that sensorimotor experiences may have a critical role in processing information about certain objects. It has been reported that the object categories that individuals with associative agnosia have difficulty reporting are those that they could not recall their action. The objects that they do not recognize would thus appear to be those that they do not associate with their sensorimotor experiences. The objects that they do recognize may be those whose action plays a critical part. This could help explain the "living" versus "nonliving" dissociation seen in CSVA. Action is certainly an important element for knowing tools, kitchen utensils, and clothes. In contrast, most animals do not evoke any gestures, and the only action linked with most fruits and vegetables is a simple gripping. It also appears that the recognition of action is well preserved in these individuals. The impairments in recognizing static objects perceived visually in associative agnosia sharply contrast with the relatively better ability to recognize objects from gestures illustrating their use and to recognize actions shown in line drawings.

It appears that the dorsal stream not only provides action-relevant information about the structural characteristics and orientation of objects but also is involved in the recognition of actions and in the recognition of objects when sensorimotor experience is evoked. This suggests that the dorsal pathway is involved in conscious visual perception and in the interpretation of goal-oriented action, even when shown in a static way. It is possible that when ventral stream damage in agnosia prevents direct access to representations of an object for perception, sensorimotor information from the dorsal stream may provide a limited mechanism for recognition. In other words, semantic information about objects may be accessed by the dorsal stream and passed onto the ventral stream for recognition. The preservation of how to manipulate an object may play a crucial part in assisting object recognition in patients with associative agnosia.

XIII. WHAT AGNOSIA TELLS US ABOUT NORMAL VISION

A major obstacle to understanding object recognition is that we perform it so rapidly and efficiently that the outcome belies the underlying complexity. One approach to discovering the processes that mediate object recognition is to study the performance of individuals who have an impairment. This breakdown approach has proven extremely illuminating and has provided important insights into the mechanisms involved in normal object recognition. The breakdown approach as reflected in the study of neuropsychological patients with agnosia is related to other approaches that also examine the system in its nonoptimal state. These approaches include the study of visual illusions in which the perception of normal subjects is distorted through some stimulus manipulation and the study of perception when cues are reduced, such as in monocular versus binocular vision.

Neuropsychological studies of agnosia have not only identified a major distinction between "early" and "late" stages of object recognition, as well as differentiated more discrete impairments within each of these stages, but also uncovered deficits associated with "intermediate"-level vision. Additionally, investigations with patients have allowed us to address issues such as category specificity both within the domain of objects and across visual domains, relating faces and words to objects. Finally, how perception might be related to action has been a focus of neuropsychological research and important observations have been gleaned from the detailed and thorough study of these patients with agnosia.

Studies of patients with agnosia have also shed light on the extent to which there is modular organization in the visual system. Although we have been concerned only with deficits of object recognition following brain damage, there are also patients with selective deficits of depth, motion, and color processing. One interpretation of these selective deficits is that there are independent regions of the brain that are specialized for certain functions. An even more extreme view, but one that has been tempered recently, is that these independent regions are exclusively dedicated for particular visual functions. At a higher

level, whether there are truly independent areas for recognition of different categories of visual objects (living/nonliving) or for different types of stimuli (faces, words, or objects) remains a matter of ongoing investigation.

Perhaps most important is that these studies of patients with visual object agnosia have constrained our theories of object recognition and, in turn, these theories have guided our investigation of these interesting and illuminating deficits.

See Also the Following Articles

ALEXIA • AUDITORY AGNOSIA • CATEGORIZATION • COLOR VISION • EYE MOVEMENTS • SENSORY DEPRIVATION • SPATIAL VISION • VISION: BRAIN MECHANISMS • VISUAL DISORDERS • VISUAL SYSTEM DEVELOPMENT AND NEURAL ACTIVITY

Acknowledgments

The writing of this article was supported by grants from the McDonnell–Pew Program in Cognitive Neuroscience and NSERC to JJM and by NIH to MB.

Suggested Reading

De Renzi, E. (1999). Agnosia. In *Handbook of Clinical and Experimental Neuropsychology* (G. Denes and L. Pizzamiglio, Eds.), pp. 371–407. Psychology Press/Erlbaum/Taylor & Francis, Hove, UK.

Farah, M. J. (1990). *Visual Agnosia: Disorders of Object Recognition and What They Tell Us about Normal Vision.* MIT Press, Cambridge, MA.

Humphreys, G. W., and Riddoch, M. J. (1987). The fractionation of visual agnosia. In *Visual Object Processing: A Cognitive Neuropsychological Approach* (G. W. Humphreys and M. J. Riddoch, Eds.), pp. 281–306. Erlbaum, London.

Milner, A. D., and Goodale, M. A. (1995). *The Visual Brain in Action.* Oxford University Press, Oxford, UK.

Mishkin, M., Ungerleider, L., and Macko, K. A. (1983). Object vision and spatial vision. *Trends in Neurosciences* **6,** 414–417.

Parkin, A. J. (1996). *Explorations in Cognitive Neuropsychology.* Blackwell, Oxford, UK.

Rumiati, R. I., Humphreys, G. W., Riddoch, M. J., and Bateman, A. (1994). Visual object agnosia without prosopagnosia or alexia: Evidence for hierarchical theories of visual recognition. *Visual Cognition* **1**(2/3), 181–225.

Warrington, E. K. (1985). Agnosia: The impairment of object recognition. In *Handbook of Clinical Neurology: Clinical Neuropsychology* (J. A. M. Frederiks, Ed.), Vol. 1, pp. 333–349. Elsevier, Amsterdam.

Agraphia

STEVEN Z. RAPCSAK*,† and PELAGIE M. BEESON†

*VA Medical Center * and University of Arizona, Tucson †*

GLOSSARY

afferent control systems Neural systems involved in monitoring visual and kinesthetic feedback during the execution of handwriting movements.

agraphia Acquired disorders of spelling and writing caused by neurological damage.

allographic conversion The process by which abstract graphemic representations are converted into the appropriate physical letter shapes.

allographs Different physical forms of a letter (i.e., upper- vs lowercase or print vs script).

grapheme Letter or cluster of letters that corresponds to a single phoneme in the language.

graphemic buffer Working memory system that temporarily stores abstract orthographic representations while they are being converted into codes appropriate for various output modalities (i.e., writing, oral spelling, typing, or spelling with anagram letters).

graphemic output lexicon The memory store of learned spellings. Representations in the graphemic output lexicon contain information about the orthographic structure of familiar words.

graphic innervatory patterns Motor commands to specific muscle effector systems involved in the production of handwriting. Graphic innervatory patterns determine the correct sequence of muscle activations and set the appropriate kinematic parameters for the writing task, including absolute stroke size, duration, and force.

graphic motor programs Abstract spatiotemporal codes for writing movements. Graphic motor programs contain information about the sequence, position, direction, and relative size of the strokes necessary to create different letters.

phoneme–grapheme conversion The mental process by which individual units of sound are translated into the corresponding letters.

The term agraphia refers to disorders of spelling and writing caused by neurological damage. In this article, we review the neuropsychological characteristics and anatomical correlates of the various agraphia syndromes encountered in clinical practice. Patterns of abnormal performance are interpreted within the framework of a cognitive information processing model of normal spelling and writing. Converging evidence from patients with focal brain lesions and functional neuroimaging studies (position emission tomography and functional magnetic resonance imaging) in normal subjects is used to identify the neural systems involved in writing.

I. INTRODUCTION

Writing is a system of communication in which conventional graphic signs are used to represent various elements of spoken language (i.e., words, syllables, or phonemes). In its fully developed form, writing is probably no more than 6000 years old—a relatively recent achievement in human evolution. The neuropsychological study of writing has an even shorter history that begins with the work of nineteenth-century European neurologists. These early investigators noted that although spoken and written expression were both frequently impaired following damage to the language-dominant hemisphere, disorders of writing, or agraphia, occasionally occurred

in isolation. This important clinical observation led to the proposal that speech and writing may have distinct neuroanatomic substrates. The potential independence of writing from speech was cogently argued by William Ogle in a landmark paper published in 1867 which also included the first classification system of agraphia. Specifically, Ogle described a type of linguistic agraphia in which patients produced well-formed letters but spelling was inaccurate or one word was substituted for another. He also noted that in other patients, agraphia was characterized by defective motor execution resulting in poorly formed letters that were frequently unrecognizable. Although Ogle's views about the relationship between speech and writing drew criticism from many of his contemporaries, the proposed clinical distinction between linguistic and motor forms of agraphia was generally accepted.

Ogle's pioneering observations were soon followed by attempts to localize the brain areas involved in writing and to clarify their anatomical and functional interactions with the cortical speech centers identified by Broca and Wernicke (Fig. 1). These efforts, spearheaded by the so-called "diagram makers," culminated in neuroanatomical models that postulated two distinct cortical centers for writing. Following Dejerine's suggestion, orthographic information relevant to the correct spelling of words was believed to be stored in the dominant angular gyrus in the form of "optic word images." In contrast, the "motor graphic images" responsible for controlling the execution of writing movements were localized to a premotor cortical area at the foot of the second frontal convolution, sometimes referred to as Exner's writing center. The act of writing presumably required the coordinated activity of the putative parietal and frontal cortical centers.

In addition to these anatomical considerations, the diagram makers also engaged in lively debates about the nature of the linguistic codes or representations used in writing. A central and often controversial topic concerned the role of phonology. The fact that writing is acquired relatively late in linguistic development, after oral language functions are firmly established, suggested to many that written expression was parasitic upon speech and that it involved obligatory phonological mediation. Consistent with this view, simple introspection reveals that writing is normally

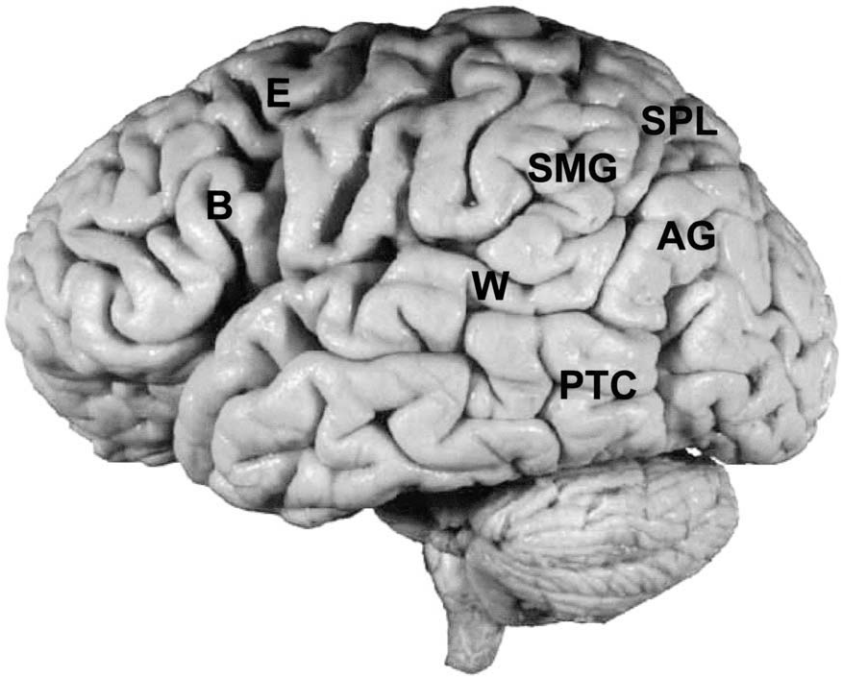


Figure 1 Cortical areas involved in speech and writing. B, Broca's area; W, Wernicke's area; E, Exner's writing center; AG, angular gyrus; PTC, posterior temporal cortex; SMG, supramarginal gyrus; SPL, superior parietal lobule [reproduced with permission from Nolte, J. (1998). *The Human Brain: An Introduction to Its Functional Anatomy*. Mosby, St. Louis, Missouri].

accompanied by "inner speech." Phonological theories of writing took two distinct forms. Some investigators explicitly denied the existence of word-specific orthographic representations and proposed that writing entailed segmenting spoken words into their constituent sounds, followed by the conversion of each sound into the appropriate letter (i.e., phoneme–grapheme conversion). As we shall see, this hypothesis is falsified by patients who lose the ability to perform phoneme–grapheme conversion but who can nonetheless spell familiar words accurately. Another version of the phonological theory of writing allowed for the possibility of stored orthographic representations for familiar words but maintained that these representations could only be activated indirectly via the spoken form of the word. The problem with this proposal is the clinical observation that in some aphasic patients written expression is superior to or is qualitatively different from speech production. These findings demonstrate that access to orthography is possible even when the corresponding phonological representation of the word is unavailable and further suggest that lexical representations for written and spoken words are neuroanatomically distinct. Note, however, that although the neuropsychological evidence is clearly at odds with the view that writing involves obligatory phonological mediation, it need not imply that phonology plays no role in writing under normal circumstances.

Although the fundamental questions raised by nineteenth-century investigators about the neural substrates and linguistic mechanisms of writing are still relevant today, contemporary work on agraphia has also been strongly influenced by cognitive models of language processing. The cognitive method of analysis relies on an information processing approach and seeks to understand complex language skills by decomposing them into several potentially independent processing components with distinct functional roles. In order to interpret the performance of neurological patients within this type of a theoretical framework, it is necessary to make the additional assumption that the proposed processing modules are also neuroanatomically distinct and can therefore be selectively impaired by brain damage. Following this line of reasoning, we begin our discussion of agraphia by presenting a cognitive model of normal spelling and writing. Next, we describe the clinical characteristics and neuroanatomical correlates of various agraphia syndromes and attempt an explanation of abnormal writing performance in terms of damage to different functional components of the model.

II. COGNITIVE MODEL OF WRITING

According to the model presented in Fig. 2, writing requires the coordinated activity of "central" and "peripheral" processing modules. Central processing components are linguistic in nature and are involved in generating spellings for familiar or unfamiliar words, whereas the peripheral components are responsible for converting orthographic information into motor commands for writing movements.

A. Central Components

The central components of the model correspond to three potentially independent linguistic spelling routes. Two of these, the lexical–semantic route and the lexical–nonsemantic route, are used for spelling familiar words. In contrast, plausible spellings for unfamiliar words or pronounceable nonwords (e.g., sprunt) are assembled by the nonlexical route. An additional central component that receives and temporarily stores the abstract orthographic representations computed by the three spelling routes is referred to as the graphemic buffer.

1. Lexical–Semantic Route

Spelling by the lexical–semantic route relies on interactions between the semantic system and the graphemic output lexicon (Fig. 2). The semantic system represents conceptual knowledge of word meanings independent of word forms. The graphemic output lexicon contains information about the orthographic structure of familiar words and thus functions as the memory store of learned spellings. As depicted in Fig. 2, semantic input to orthography may be direct (pathway A) or indirect via the phonological representation of the word that is also used in speech production (pathways B and C). It has been proposed that representations in the graphemic output lexicon are normally activated by combined input from both the direct and the indirect routes. Such dual coding may safeguard against errors that might occur if one or the other route was used exclusively. For instance, relying on the indirect route via the phonological output lexicon may result in homophone confusions (e.g., "stair"–"stare") since these words have the exact same sound pattern even though they are spelled differently. On the other hand, spelling by the direct route might be susceptible to semantic errors (e.g.,

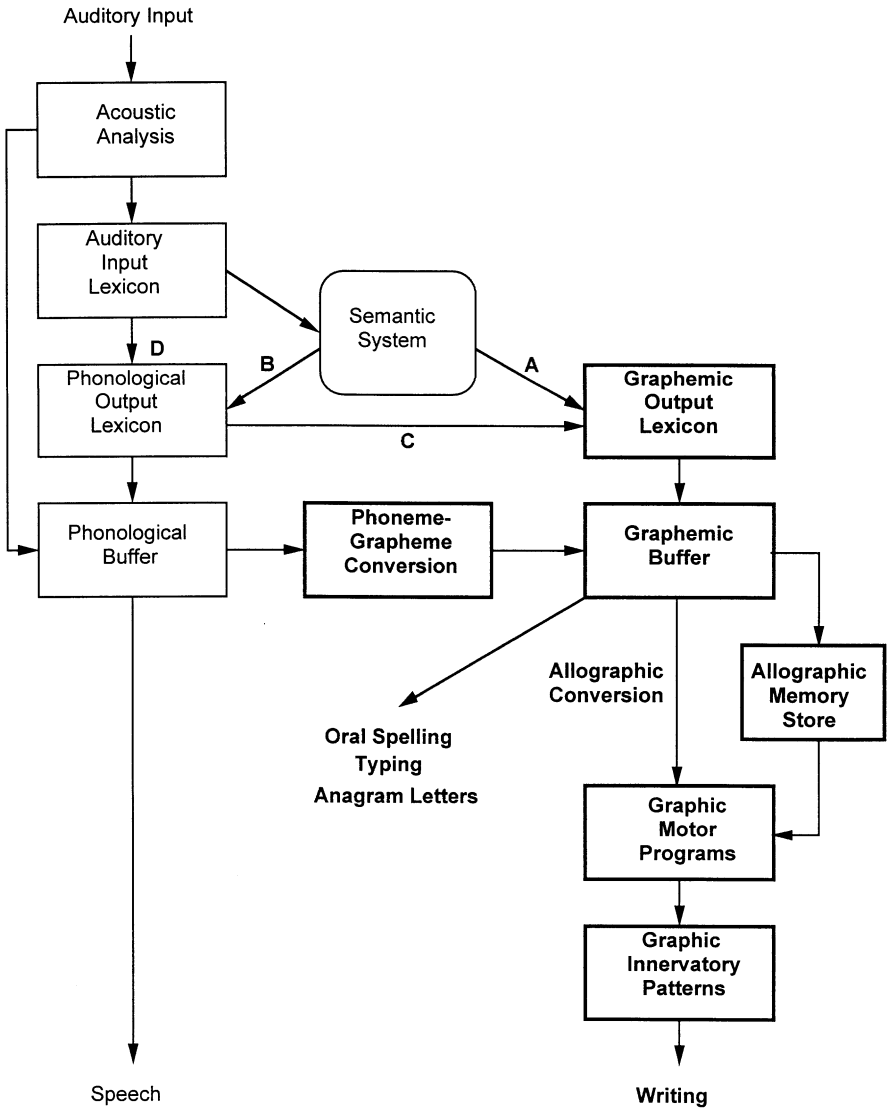


Figure 2 Cognitive information processing model of spelling and writing.

"day"–"week") unless the response was constrained by simultaneous phonological input to orthography via the indirect route. Postulating dual access to the graphemic output lexicon may also provide an explanation for the clinical observation that some aphasic patients with damage to the phonological output lexicon can write words they are no longer able to say (by relying on the direct route to the graphemic output lexicon via pathway A).

As can be seen in Fig. 2, the lexical–semantic route provides the only mechanism for incorporating meaning into writing. Therefore, this spelling route plays a critical role in conceptually mediated writing tasks such as written composition and written naming. It is likely that writing familiar words to dictation is normally performed via the lexical–semantic route as well, although the lexical–nonsemantic route can also potentially be used for this purpose.

2. Lexical–Nonsemantic Route

Spelling to dictation by the lexical–nonsemantic route relies on connections between the spoken and written forms of the same word that bypass the semantic system. One possibility is that the dictated word first activates its representation in the auditory input lexicon, followed by the retrieval of the corresponding entry from the phonological output lexicon, which in

turn activates the appropriate orthographic word form in the graphemic output lexicon (Fig. 2, pathways D and C). An alternative mechanism might involve direct connections between corresponding representations in the auditory input and the graphemic output lexicons. As it will become apparent later, the evidence for lexical–nonsemantic spelling comes mostly from patients with agraphia, and it is not entirely clear what function this spelling route might serve under normal circumstances.

3. Nonlexical Route

Normal individuals can produce plausible spellings for unfamiliar words or nonwords without significant difficulty. Since these novel items are not represented in the graphemic output lexicon, spelling cannot simply rely on the activation of stored orthographic patterns. According to our model, the spelling of unfamiliar words and nonwords is accomplished via the nonlexical spelling route. Unlike the lexical spelling routes that rely on a whole-word retrieval process, spelling by the nonlexical route involves a subword-level algorithmic procedure based on phoneme–grapheme conversion rules (Fig. 2). In this process, the novel auditory stimulus is first broken down into its component sounds, following which each constituent phoneme is converted into the corresponding grapheme. It is possible, however, that the nonlexical route can also perform phonological-to-orthographic translations based on units larger than individual phonemes and graphemes (e.g., syllables). Finally, it has been suggested that spelling nonwords may not be entirely nonlexical and may instead be based on lexical analogy with familiar words.

Phoneme–grapheme conversion plays an important role in learning to spell, but this rule-based nonlexical procedure is resorted to much less frequently once the normal adult spelling vocabulary is established. However, the nonlexical route can serve as a backup strategy when word-specific orthographic information is temporarily unavailable or is incomplete. This might happen when normal individuals attempt to spell words they do not use very often (i.e., low-frequency words). It should be noted that in orthographically opaque languages such as English, in which sound-to-spelling relationships are notoriously inconsistent, the nonlexical procedure can only succeed with unambiguous or regular words that have highly predictable phoneme–grapheme correspondences (e.g., ''mint''). Ambiguous words, in which the same phonology can be realized by more than one combination of letters (e.g., ''drain'' and ''drane''), and irregular words that contain exceptional phoneme–grapheme mappings (e.g., ''choir'') cannot be spelled correctly by the nonlexical route since the straightforward application of phoneme–grapheme conversion rules for such words will result in phonologically plausible errors (e.g., ''yot'' for ''yacht''). The spelling of ambiguous and irregular words, therefore, depends critically on access to precise word-specific orthographic knowledge.

4. Graphemic Buffer

Central spelling routes compute abstract orthographic representations that can be externalized in writing, oral spelling, typing, or as an arrangement of anagram letters. The graphemic buffer is a working memory system that temporarily stores these abstract orthographic representations while they are being converted into codes appropriate for the various output modalities (e.g., letter shapes for written spelling or letter names for oral spelling). As shown in Fig. 2, this processing module receives the output of all three spelling routes and therefore occupies a strategic position between the central and peripheral components of the writing process. The maintenance of information within the graphemic buffer is influenced by stimulus length (i.e., the number of graphemes that make up the spelling of a word or nonword) since longer items require more storage capacity than shorter ones. Dysfunction of the buffer is associated with the loss of information relevant to the identity and serial ordering of stored graphemes, leading to letter substitutions, additions, deletions, and transpositions.

B. Peripheral Components

The peripheral processing components of writing are responsible for converting abstract orthographic information into movements of the pen. This complex sequence, which begins with letter shape selection and ends with neuromuscular execution, is carried out by a set of hierarchically organized processing modules, the operational characteristics of which are discussed in this section. Peripheral conversion mechanisms that subserve other output modalities (i.e., oral spelling, typing, and spelling with anagram letters) will not be discussed here, except to note that these systems most likely diverge from writing at the level of the graphemic buffer (Fig. 2).

1. Allographic Conversion

The first step in producing writing is the allographic conversion process, which involves the selection of the appropriate letter shapes for the string of graphemes held in the graphemic buffer. In handwriting, letters can be realized in different case (upper vs lower) or style (print vs script). The various physical forms each letter of the alphabet can take are referred to as allographs (e.g., B, b, *b*, *b*). The choice of allographs is influenced by convention (e.g., capitalizing letters in sentence-initial position), contextual factors (e.g., filling out a form vs writing a note or a letter), and individual style.

The exact nature of the representations involved in allographic conversion remains poorly understood. Some investigators have proposed that allographs are stored in long-term memory as abstract visuospatial descriptions of letter shape (the allographic memory store in Fig. 2). Others have suggested that allographs correspond to letter-specific graphic motor programs in which shape information is specified in terms of the sequence of strokes necessary to produce the desired letter. Although it is currently not possible to adjudicate between these competing proposals, the latter interpretation certainly has the appeal of parsimony. In particular, it is not entirely clear why one would need to retrieve abstract visuospatial information about letter shapes once the characteristic stroke patterns of different letters are firmly established in procedural memory and writing becomes an automatic motor task.

2. Graphic Motor Programs

Handwriting is a highly specialized motor activity that takes years to master. Similar to other complex motor skills, the neural control of writing movements is organized in a hierarchical fashion, with the general outline of the movement represented at the highest level and lower levels regulating increasingly specific details of neuromuscular execution. At the highest level, writing movements are controlled by graphic motor programs. It is assumed that these programs are stored in long-term memory rather than being assembled *de novo* every time a particular letter is written. Graphic motor programs contain information about abstract spatiotemporal movement attributes, including the sequence, position, direction, and relative size of the strokes necessary to produce different letters. However, graphic motor programs do not specify concrete kinematic parameters such as absolute stroke size, duration, and force. Although graphic motor programs are letter specific, they are effector independent in the sense that they do not determine which particular muscle groups are to be recruited for movement execution. An interesting aspect of writing is that it can be performed by using different muscle–joint combinations of the same limb (e.g., writing with a pen is accomplished with the distal muscles of the hand and wrist, whereas writing on a blackboard mostly involves the proximal muscles of the shoulder and elbow) or by using different limbs altogether (e.g., dominant vs nondominant hand or foot). The finding that writing produced by different effector systems displays striking similarities with respect to overall letter shape suggests a high degree of motor equivalence, consistent with the notion of effector-independent graphic motor programs.

3. Graphic Innervatory Patterns

The final stage in handwriting involves the translation of the information encoded in graphic motor programs into graphic innervatory patterns containing sequences of motor commands to specific effector systems. It is at this lower stage in the motor hierarchy that the appropriate combinations of agonist and antagonist muscles are selected and concrete movement parameters specifying absolute stroke size, duration, and force are inserted into the program. Since the biophysical context of the actual writing task may vary from one occasion to another (e.g., different writing instruments or surfaces), the motor system must have the flexibility to compute the appropriate kinematic parameters "on-line." Therefore, parameter estimation is viewed as a more variable and dynamic process than the retrieval of stored graphic motor programs specifying relatively invariant movement attributes. Once the kinematic parameters for the given writing context have been selected, the motor system executes the strokes required to produce the desired letters as a rapid sequence of ballistic movements (between 100 and 200 msec/stroke).

4. Afferent Control Systems

Central motor programs for skilled movements can be executed in an "open-loop" fashion (i.e., relatively independent of sensory feedback). Consistent with this general principle of motor physiology, it is possible to write letters in the absence of vision or when the writing hand is deafferented as a result of severe peripheral or central sensory loss. It is also clear, however, that

normal handwriting requires afferent input for maximum speed and accuracy. This fact can be readily demonstrated by depriving normal individuals of visual feedback (either by having them write with their eyes closed or by using delayed visual feedback). Characteristic production errors under these conditions include the tendency to duplicate or omit letters or strokes, especially when writing sequences of similar items (e.g., words with double letters or letters containing repeated stroke cycles such as "m" or "w"). These errors can also be observed when subjects are asked to write while performing another task simultaneously (e.g., counting aloud or tapping with the other hand). In the dual-task situation, visual and kinesthetic feedback are available, but sensory information is not being used efficiently because attention is diverted from it by the secondary task. It has been proposed that the accurate monitoring of afferent feedback plays an important role in updating graphic motor programs as to which letters or strokes have already been executed. This "place keeping" function becomes especially critical when similar or identical stroke patterns have to be produced or when the complex sequence of muscle activations required for handwriting needs to be sustained over longer periods of time. Sensory feedback is also required to maintain the correct spacing between letters and words and to keep the line of writing properly oriented on the page.

III. NEUROPSYCHOLOGICAL DISORDERS OF WRITING

Having identified the major functional components of the normal writing process, we now turn our attention to clinical disorders of writing in patients with neurological damage. If we are correct in assuming that the various processing components of our cognitive model are subserved by dedicated neural systems, then damage to specific brain regions should be associated with distinct types of agraphia. Selective damage to a single processing component should result in a "pure" agraphia syndrome with a characteristic and predictable combination of impaired and preserved writing abilities. Specifically, the particular functions assigned to the damaged module should be disrupted, whereas those mediated by other modules should be relatively spared. Although pure cases fulfilling these criteria are encountered occasionally, in clinical practice writing disorders often display features consistent with simultaneous damage to several processing modules, resulting in mixed or multicomponent agraphia syndromes.

From a neuropsychological perspective, agraphia syndromes can be subdivided into central and peripheral types. In essence, this classification system is similar to the distinction between linguistic versus motor forms of agraphia introduced by previous investigators. Central agraphias reflect damage to the proposed linguistic spelling routes or the graphemic buffer. These syndromes are characterized by qualitatively similar spelling deficits across all possible modalities of output (i.e., writing, oral spelling, typing, and spelling with anagram letters). In contrast, in the peripheral agraphias the damage involves processing components located distal to the graphemic buffer (Fig. 2), and the impairment primarily affects the selection and production of letters in handwriting.

A. Central Agraphias

Central agraphia syndromes include lexical or surface agraphia, phonological agraphia, deep agraphia, semantic agraphia, and graphemic buffer agraphia. The characteristic linguistic features and neuroanatomical correlates of these syndromes are discussed in this section.

1. Lexical Agraphia: The Selective Impairment of Word-Specific Orthographic Knowledge

Lexical agraphia is characterized by an impairment of vocabulary-based spelling. Patients have difficulty spelling familiar words, especially words that contain ambiguous or irregular phoneme–grapheme mappings. In contrast, spelling of regular words is relatively preserved, as is the ability to spell unfamiliar words or nonwords. Errors in spelling ambiguous or irregular words are usually phonologically plausible (e.g., "oshen" for "ocean"). In addition to the strong effect of orthographic regularity, spelling is influenced by word frequency, with an advantage of high-frequency words over low-frequency ones. However, spelling performance is typically unaffected by other lexical–semantic variables such as imageability (i.e., high imageability or concrete words such as "apple" vs low imageability or abstract words such as "pride") and grammatical word class (i.e., content words including nouns, verbs, and adjectives vs functors such as prepositions, pronouns, articles, and auxiliaries).

The central linguistic features of lexical agraphia can be accounted for by postulating an impairment at the level of the graphemic output lexicon (Fig. 2). The loss

or unavailability of stored word-specific orthographic information forces patients to generate spellings by relying on the preserved nonlexical route. As indicated earlier, this route is primarily used to compute plausible spellings for unfamiliar words and nonwords, but it can also handle regular words that strictly obey phoneme–grapheme conversion rules. However, attempts to spell ambiguous or irregular words by the nonlexical route result in phonologically plausible errors.

In most reported cases of lexical agraphia the responsible lesion involved the left temporo–parieto–occipital junction. The lesion sites overlap in the region of the angular gyrus (Brodmann area 39) and posterior middle and inferior temporal gyrus (Brodmann area 37), typically sparing the perisylvian language cortex that includes Wernicke's area (Brodmann area 22), the supramarginal gyrus (Brodmann area 40), and Broca's area (Brodmann area 44). Lexical agraphia has also been described in patients with Alzheimer's disease (AD) and semantic dementia. Similar to patients with focal brain lesions, lexical agraphia in these neurodegenerative disorders may reflect the frequent involvement of left hemisphere temporal and parietal cortical association areas by the disease process and the relative sparing of the perisylvian language zone. Consistent with this hypothesis, one study found that lexical agraphia in AD correlated with reduced metabolic activity in the left angular gyrus, as measured by positron emission tomography (PET). In summary, the neuroanatomical findings in patients with lexical agraphia suggest that information relevant to the spelling of familiar words is stored in the left angular gyrus and/or ventrally adjacent posterior temporal cortex, making this extrasylvian cortical area a likely neural substrate of the graphemic output lexicon. The critical role of posterior temporal cortex in orthographic processing is also supported by PET and functional magnetic resonance imaging studies in normal subjects which demonstrated activation of left Brodmann area 37 during various writing tasks.

2. Phonological Agraphia: Dysfunction of the Nonlexical Spelling Route

In phonological agraphia, spelling of unfamiliar words and nonwords is significantly impaired, whereas the spelling of familiar words is relatively spared. The contrast between phonological agraphia and lexical agraphia constitutes an important double dissociation and provides strong support for the proposed distinction between the linguistic procedures involved in spelling unfamiliar vs familiar words. The existence of phonological agraphia also effectively disproves the claim that spelling familiar words must normally rely on phoneme–grapheme conversion.

Phonological agraphia reflects the selective disruption of the nonlexical spelling route. The spelling impairment may be attributable to at least two qualitatively different types of processing deficits. Some patients seem to have lost their knowledge of phoneme–grapheme correspondence rules, as suggested by their inability to write even single letters correctly when given their characteristic sounds. Others can translate single phonemes into the appropriate graphemes, and in these cases the nonlexical spelling deficit may reflect an inability to segment novel auditory stimuli into their constituent sounds. However, it is important to keep in mind that defective phoneme–grapheme transcoding may also be caused by imperception of the stimuli (i.e., phoneme discrimination deficit) or reduced phonological short-term memory.

In pure cases of phonological agraphia, spelling of familiar words (including ambiguous and irregular words) can be performed at a fairly high level, consistent with preserved access to representations in the graphemic output lexicon via the lexical–semantic route. The fact that spelling in phonological agraphia relies on a lexical–semantic strategy is also suggested by the observation that some patients cannot write words to dictation unless they have access to their meaning. Furthermore, spelling performance may be influenced by lexical–semantic variables such as imageability, grammatical word class, and frequency. Specifically, some patients spell high-imageability words better than low-imageability words, and content words may have an advantage over functors. In addition, high-frequency words may be spelled more accurately than low-frequency words. Orthographic regularity, however, does not have a significant effect on spelling. Spelling errors in phonological agraphia are usually phonologically implausible, although they may have visual similarity to the target. Morphological errors (e.g., ''works''–''working'') and functor substitutions (e.g., ''over''–''here'') are also observed occasionally.

The most common neuroanatomical correlate of phonological agraphia is damage to the perisylvian language zone, including Wernicke's area, the supramarginal gyrus, and, in some cases, Broca's area. In contrast, the extrasylvian temporoparietal cortical areas implicated in lexical agraphia are typically spared. It has been suggested that the critical neural

substrate of phonological agraphia is damage to the supramarginal gyrus. Therefore, this cortical area may play an important role in subword-level phonological-to-orthographic transcoding. In general, the neuroanatomical observations in patients with phonological agraphia are consistent with the results of functional neuroimaging studies in normal subjects that demonstrated prominent activation of perisylvian cortical areas during language tasks requiring phonological processing. However, the precise role of the different perisylvian cortical regions in lexical vs nonlexical phonology, subvocal rehearsal, and phonological short-term memory remains to be elucidated.

3. Deep Agraphia: Semantic Errors in Writing

The hallmark of deep agraphia is the presence of frequent semantic errors in writing. Semantic errors have no orthographic or phonological similarity to the target; the only relationship is based on word meaning (e.g., "apple"–"banana"). Apart from these pathognomonic errors, deep agraphia has several linguistic characteristics in common with phonological agraphia. These include the severe difficulty in spelling unfamiliar words and nonwords; the effect of imageability, grammatical word class, and frequency in spelling familiar words; and the presence of morphological errors, functor substitutions, and visually similar misspellings. The substantial overlap in terms of both linguistic features and anatomical lesion location suggests that a strict taxonomic separation between phonological and deep agraphia may be artificial.

Semantic errors in writing indicate dysfunction of the lexical–semantic spelling route and may reflect several different processing impairments. Possible mechanisms include damage to the semantic system or faulty transmission of information between the semantic system and the graphemic output lexicon (Fig. 2). In writing to dictation, semantic errors may also result from damaged connections between the auditory input lexicon and the semantic system. However, dysfunction of the lexical–semantic route alone is not sufficient to explain semantic errors in writing to dictation since the correct spelling of the target word could potentially be generated by spelling routes that normally bypass the semantic system and are therefore not susceptible to errors based on word meaning. Specifically, potential semantic errors could be blocked by phonological-to-orthographic transcoding via either the lexical–nonsemantic or the nonlexical spelling routes. For instance, although the damaged lexical–semantic route may activate the entry "banana" in the graphemic output lexicon in response to the dictated word "apple," the overt semantic error could be avoided if the phonological code /æpəl/ was simultaneously released from the phonological output lexicon and provided a second source of input to orthography via the lexical–nonsemantic spelling route (Fig. 2, pathways D and C) or if the correct spelling was computed by phoneme–grapheme conversion via the nonlexical route. Therefore, the appropriate conditions for semantic errors arise when the output of the dysfunctional lexical–semantic spelling route remains completely unconstrained by orthographic information generated by the lexical–nonsemantic and nonlexical routes. In conclusion, an adequate linguistic explanation of deep agraphia requires that we postulate multiple processing impairments affecting all three central spelling routes.

Deep agraphia is typically associated with large left hemisphere lesions that produce extensive damage to the perisylvian language zone. The near-complete destruction of left hemisphere language areas documented in some cases of deep agraphia raises important questions about the neural systems that mediate the residual writing abilities of these patients. One possibility is that writing is generated by spared left hemisphere cortical/subcortical structures. Alternatively, deep agraphia may arise when writing is produced by the right hemisphere. According to the latter hypothesis, the syndrome reflects the intrinsic functional limitations of the intact right hemisphere language system. Several features of right hemisphere language are consistent with this view. For instance, neuropsychological studies of auditory and written word comprehension in split-brain patients suggest that the semantic system of the right hemisphere may not be as precisely organized as its left hemisphere counterpart, making it susceptible to semantic errors. In addition, the right hemisphere has reduced phonological competence and may completely lack the ability to perform nonlexical or subword-level phonological operations. Finally, the right hemisphere lexicon is thought to be biased toward high-imageability words and may have only limited representation of low-imageability words, functors, and bound morphemes (i.e., prefixes and suffixes). Therefore, right hemisphere writing remains a viable explanation of deep agraphia in patients with massive left hemisphere lesions. The isolated right hemisphere's capacity to generate elementary written expression has also been documented in split-brain patients and in patients with left hemispherectomy.

4. Semantic Agraphia: Writing without Meaning

As previously discussed, knowledge of word meanings is incorporated into writing via the lexical–semantic spelling route (Fig. 2). Semantic influence on writing is essential for written composition and for written naming. Damage to the semantic system or a disconnection of the semantic system from the graphemic output lexicon may interfere with the ability to use meaning to guide the selection of the appropriate orthographic word forms, resulting in a writing disorder known as semantic agraphia. In semantic agraphia, spontaneous writing and written naming may be significantly compromised, but writing to dictation is relatively preserved since this task can be accomplished without semantic mediation by relying on the lexical–nonsemantic or nonlexical spelling routes (Fig. 2). Using spelling routes that bypass the semantic system, patients with semantic agraphia can write familiar words to dictation even when their meaning is not understood (either because of damage to the semantic system or because of a disconnection of the semantic system from the auditory input lexicon). Accurate spelling of ambiguous and irregular words without comprehension of their meaning in semantic agraphia provides the best available evidence for the existence of a lexical–nonsemantic spelling route (i.e., direct transcoding between phonological and orthographic lexical representations via pathways D and C in Fig. 2), since these words could not be spelled correctly by relying on a nonlexical strategy.

Although writing to dictation in semantic agraphia may be reasonably accurate, homophones create major difficulties for these patients. Since homophones have identical sound patterns (e.g., ''blue''–''blew''), the selection of the correct orthographic form depends on the semantic context in which the word is used (e.g., ''she has *blue* eyes'' vs ''the wind *blew* hard''). In semantic agraphia, meaning has limited influence on spelling. Therefore, patients are unable to use semantic context to disambiguate dictated homophones and they frequently produce the semantically inappropriate orthographic word form.

Semantic agraphia has been associated with both frontal and temporoparietal left hemisphere lesion sites that typically spare the perisylvian language zone. The widespread anatomical distribution of lesions in semantic agraphia is consistent with functional neuroimaging studies in normal subjects that documented activation of multiple frontal and temporoparietal extrasylvian cortical association areas during language tasks requiring semantic processing. Homophone confusions, consistent with impaired semantic influence on spelling, have also been observed in patients with lexical agraphia, demonstrating the close anatomical proximity of semantic and orthographic processing modules within the left extrasylvian temporoparietal cortex. Semantic agraphia in patients with AD or semantic dementia may also reflect the pathological involvement of these posterior left hemisphere regions by the cortical degenerative process.

5. Graphemic Buffer Agraphia: The Interaction of Spelling, Memory, and Attention

As can be seen in Fig. 2, the output of all three central spelling routes converges on the graphemic buffer. Therefore, damage to this working memory system results in similar errors across all writing tasks (i.e., written composition, written naming, and writing to dictation). Furthermore, dysfunction of the buffer is expected to have a detrimental effect on the processing of all stored graphemic representations, regardless of lexical status (i.e., words vs nonwords), lexical–semantic features (i.e., frequency, imageability, and grammatical word class), or orthographic regularity. In contrast, spelling accuracy is strongly influenced by word length since each additional grapheme introduces a potential error by increasing the demand on limited storage capacity. Because the graphemic buffer is a central processing component that supports all possible modalities of output, qualitatively similar errors are observed in writing, oral spelling, typing, and spelling with anagram letters.

Dysfunction of the buffer interferes with the short-term retention of information concerning the identity and serial ordering of stored graphemes. Spelling errors take the form of letter substitutions, deletions, additions, and transpositions, and their frequency increases as a function of word length. The distribution of errors may also be influenced by letter position. In some cases, more errors were produced on letters at the end of the word, consistent with the hypothesized serial left-to-right readout process from the buffer (i.e., items retrieved last are more vulnerable to degradation because they have to be retained in memory longer). Errors on terminal letters may decrease in frequency if patients are asked to spell backward, suggesting that it is the order of retrieval rather than the spatial position of the letter within the word that is the critical factor. In other patients, spelling errors primarily involved letters in the middle of the word. The usual explanation offered for this phenomenon is that letters in internal

positions are more susceptible to interference by neighboring items than are letters located at either end.

In addition to memory storage capacity, the retrieval of information from the graphemic buffer is also influenced by attentional factors. For instance, it has been demonstrated that some patients with neglect produce more spelling errors on the side of the word opposite their lesion, regardless of stimulus length or modality of output (i.e., oral vs written spelling) and independent of whether they spell in conventional left-to-right order or in reverse. These observations are consistent with a unilateral defect in the distribution of attention over an internal graphemic representation in which the order of graphemes is spatially coded within a word-centered coordinate system (i.e., a spatial coordinate frame whose center corresponds to the midpoint of the word).

Lesion sites in patients with agraphia due to dysfunction of the graphemic buffer have been variable. Damage to the left posterior parietal cortex is a common finding, although some patients have evidence of frontal lobe involvement. This provisional localization of the graphemic buffer is consistent with the prominent role of frontoparietal cortical networks in verbal working memory and spatial attention, as revealed by recent functional neuroimaging studies in normal subjects.

B. Peripheral Agraphias

Clinically, the peripheral agraphias are characterized by defective selection or production of letters in handwriting. In pure cases, the central or linguistic components of writing are intact, and the preservation of orthographic knowledge can be demonstrated via spared output modalities that include oral spelling, typing, and spelling with anagram letters. The major subtypes of peripheral agraphia include allographic disorders, apraxic agraphia, nonapraxic disorders of motor execution, and afferent dysgraphia.

1. Allographic Disorders: Impairments of Letter Shape Assignment

In allographic writing disorders the main difficulty involves the activation or selection of letter shapes appropriate for the abstract graphemic representations held in the graphemic buffer. As noted earlier, it is not clear whether the underlying deficit reflects a failure to access visuospatial letter shape information from an independent allographic memory store or whether it is caused by a functional impairment of the procedures necessary for activating the correct letter-specific graphic motor programs (Fig. 2).

The breakdown of the allographic conversion process can take different clinical forms. In some cases it may result in partial or complete inability to write the intended letters, apparently because of a failure to remember how to create the appropriate physical letter shapes. The letter production deficit may be specific to certain allographic forms. For instance, patients have been described who could write uppercase letters significantly better than lowercase letters, whereas others showed the opposite dissociation. These findings imply independent neural representations for upper- versus lowercase letters that can be selectively disrupted by brain damage. Allographs for different lowercase writing styles may also be organized separately, as suggested by the reported dissociation between printing lowercase letters versus writing the same letters in a cursive style.

Other patients with allographic disorders have no difficulty in producing individual letters but they seem to have trouble specifying the contextually appropriate allographic repertoire, resulting in an uncontrollable mixing of upper- and lowercase letters in handwriting. In still other patients, case and style specification procedures are intact, but written spelling is characterized by numerous well-formed letter substitution errors. Substitution errors typically involve the production of letter shapes physically similar to the intended target, and the occurrence of these errors may be influenced by letter frequency (i.e., more errors on letters used less frequently in the English language). Whether the documented physical similarity effects in letter substitution errors are based on shared visuospatial features or whether they reflect the fact that graphic motor programs containing similar stroke sequences are more likely to be confused has not been established conclusively.

In the majority of reported cases, allographic writing disorders were associated with left temporo-parieto-occipital damage. Therefore, this posterior cortical region may play an important role in activating and selecting the appropriate letter shapes for written output.

2. Apraxic Agraphia: Defective Motor Programming of Writing Movements

Apraxic agraphia is characterized by poor motor execution of the stroke patterns necessary to produce letters. In order to classify the writing disorder as

"apraxic," it is important to demonstrate that the letter production deficit is not caused by more elementary sensorimotor (i.e., weakness or deafferentation), basal ganglia (i.e., tremor or rigidity), or cerebellar (i.e., dysmetria or ataxia) dysfunction affecting the writing limb.

Although patients with apraxic agraphia may be able to grasp the pen correctly, the spatiotemporal attributes of handwriting are severely disturbed and the facile strokes normally used to produce the required spatial trajectory are replaced by slow, effortful, and imprecise movements. In severe cases, all attempts at writing may result in an illegible scrawl. When the writing disorder is less severe, it may be possible to distinguish between various writing styles and between upper- and lowercase forms, although individual letters may be difficult to recognize. Typical errors of letter morphology include spatial distortions, stroke omissions, and the insertion of anomalous strokes resulting in nonletters (Fig. 3A). When letters are more legible, it may be possible to demonstrate that written spelling is actually preserved. The writing difficulty may be specific to letter formation since in some cases numbers could be produced correctly. Patients with apraxic agraphia can sometimes copy letters better than they can write them spontaneously or to dictation. However, copying is slow and fragmented, and it is usually accomplished in a "stroke-by-stroke" fashion relying heavily on visual feedback. These production features are characteristic of unskilled graphomotor performance and are reminiscent of the way children first learn to write.

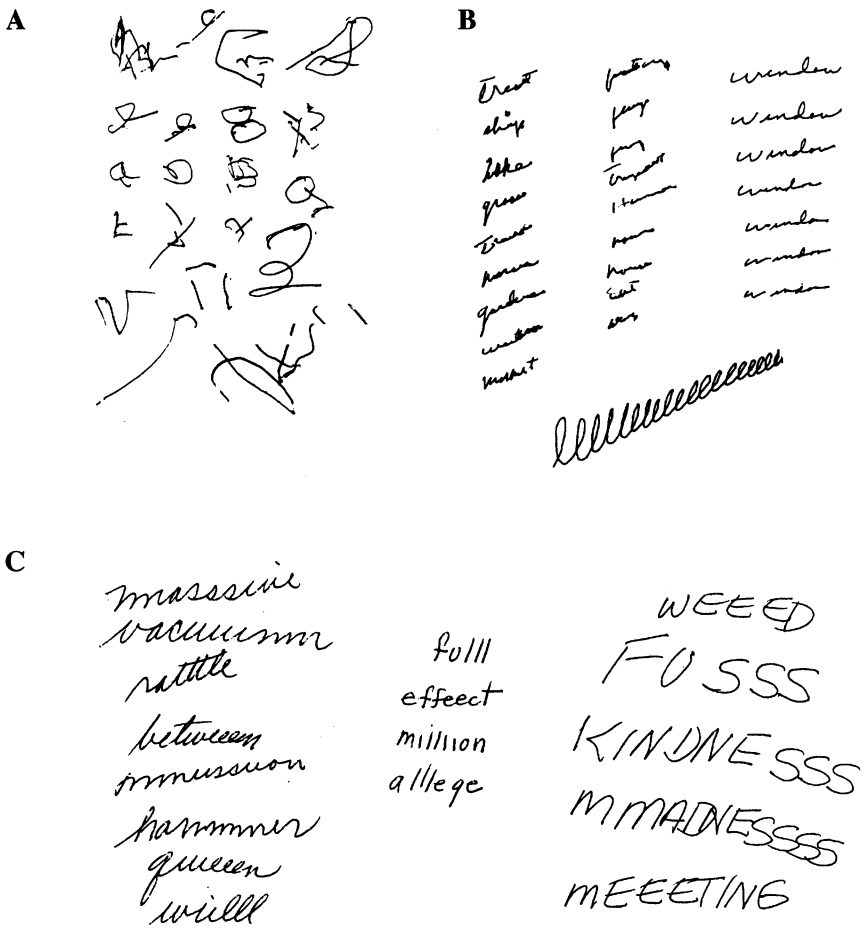


Figure 3 (A) Errors of letter morphology in the writing produced by a patient with apraxic agraphia following left parietal lobe damage. (B) Micrographia in Parkinson's disease. Note overall reduction of letter size. Progressive reduction of writing amplitude is seen with repeated attempts to write the same word or letter. (C) Afferent dysgraphia in a patient with right parietal lobe damage. Duplications occur mostly in words with double letters and when writing letters that contain repeated stroke cycles.

Within the framework of our model, apraxic agraphia can be explained by postulating damage to processing components involved in the programming of handwriting movements. Possible neuropsychological mechanisms include the destruction or disconnection of graphic motor programs or damage to systems involved in translating the information contained in graphic motor programs into graphic innervatory patterns (Fig. 2). Apraxic agraphia is dissociable from limb apraxia, suggesting that motor programs for writing are distinct from programs for other types of purposeful skilled movements.

Apraxic agraphia has specific clinicoanatomical correlations. In right-handers, the responsible lesions typically involve the left posterior–superior parietal region. In particular, damage to cortical areas surrounding the intraparietal sulcus (i.e., superior parietal lobule and the superior portions of the angular and supramarginal gyri) is a common finding. In other cases, the lesions involve dorsolateral frontal cortex, including the premotor area at the foot of the second frontal convolution known as Exner's writing center. Finally, apraxic agraphia has been described following damage to the supplementary motor area (SMA). Taken together, these neuroanatomical observations suggest that the motor programming of handwriting is controlled by a distributed neural system that includes parietal and frontal cortical components with distinct functional roles. Specifically, posterior–superior parietal cortex may contain abstract spatiotemporal codes for writing movements (i.e., graphic motor programs), whereas the frontal components of the network (dorsolateral premotor cortex and SMA) may be responsible for generating the appropriate motor commands to specific muscle effector systems. Parietal lesions may cause apraxic agraphia by damaging or destroying graphic motor programs, whereas frontal premotor lesions may interfere with the process of translating these programs into graphic innervatory patterns specifying the proper sequence of muscle activations necessary for producing the appropriate stroke patterns. White matter lesions located deep to these cortical areas may cause apraxic agraphia by disconnecting the parietal and frontal components of the network.

The existence of a distributed frontoparietal cortical system responsible for controlling handwriting movements receives additional support from functional neuroimaging studies in normal subjects (Fig. 4). These studies have consistently demonstrated activation in posterior–superior parietal cortex, dorsolateral premotor cortex, and the SMA during the performance of various writing tasks. Similar brain regions are activated during imagined writing movements, suggesting that mentally executed and real graphomotor gestures are subserved by partially overlapping neural systems. Frontoparietal cortical networks are also implicated in other types of skilled hand movements typically performed under visual guidance, including reaching, grasping, and object manipulation. These observations suggest that distinct frontoparietal cortical systems may form the neural substrate of specific object-oriented motor behaviors.

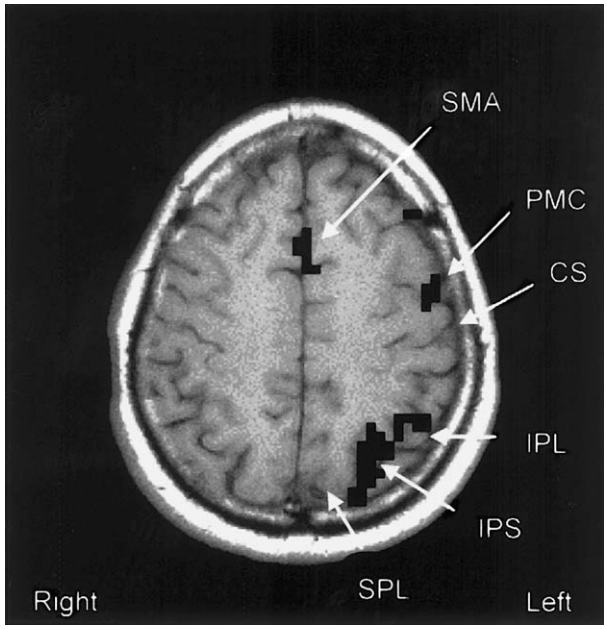


Figure 4 Functional MRI scan in a normal right-handed subject demonstrating the cortical network involved in the production of handwriting movements. Regions of activation correlate with writing to dictation minus a control task of drawing circles. SMA, supplementary motor area; PMC, premotor cortex (including Exner's area); IPL, inferior parietal lobule (angular and supramarginal gyri); IPS, intra parietal sulcus; SPL, superior parietal lobule; CS, central sulcus.

In most right-handed persons the left hemisphere is dominant for writing. Consequently, writing with the left hand in these individuals must involve transfer of linguistic and motor information from the left to the right hemisphere across the corpus callosum. Consistent with this hypothesis, damage to the corpus callosum in right-handers produces unilateral agraphia of the left hand. Neuroanatomical observations in patients with callosal agraphia suggest that information critical for programming the skilled movements of writing is transferred through the body of the corpus callosum, whereas linguistic information is transferred

more posteriorly through the fibers of the splenium. Furthermore, the fact that unilateral apraxic agraphia and ideomotor apraxia are occasionally dissociable following callosal damage implies that motor programs for writing and programs for other types of skilled limb movements are transferred through anatomically distinct callosal pathways.

3. Impaired Selection and Control of Movement Parameters: Nonapraxic Disorders of Writing Force, Speed, and Amplitude

In addition to determining the correct sequence of muscle activations, the neural systems responsible for generating graphic innervatory patterns must also select the appropriate kinematic parameters for the writing task. The breakdown of these operations may result in the insertion of incorrect movement parameters into otherwise intact graphic motor programs, leading to defective control of writing force, speed, and amplitude.

A typical example of this type of motor production deficit is the micrographia of patients with Parkinson's disease. As its name implies, micrographia is characterized by a striking reduction in handwriting size (Fig. 3B). Letters may get progressively smaller during the writing process. Even though letter size is diminished, writing remains legible in milder cases and there are no stroke-level errors of the type seen in apraxic agraphia. These findings indicate that the control of writing movements at the level of graphic motor programs is preserved. Although patients with micrographia can activate the correct muscles in the appropriate sequence, they cannot generate the forces necessary to maintain proper letter size. Writing speed is also significantly reduced. These production features are consistent with the general reduction of movement amplitude and speed in Parkinson's disease.

Micrographia in Parkinson's disease reflects basal ganglia dysfunction caused by the loss of striatal dopamine. Dopaminergic projections to the striatum originate in the substantia nigra of the midbrain. It has been demonstrated that focal lesions of the substantia nigra or the striatum can produce micrographia of the contralateral hand. Basal ganglia structures exert their influence on motor behavior through reciprocal connections to cortical motor areas, including dorsolateral premotor cortex and SMA. Operating as a functional unit, the cortical and subcortical components of the basal ganglia–thalamocortical motor loop play an important role in the control of movement force, speed, and amplitude. Consistent with this hypothesis, micrographia has been observed not only in patients with basal ganglia lesions but also following damage to the SMA.

Poor penmanship is also typical of the writing produced by patients with cerebellar dysfunction. Cerebellar lesions interfere with the smooth and automatic execution of the rapid alternating movement sequences that characterize normal handwriting. Writing movements are slow, effortful, and disjointed. Precise control over movement direction, force, speed, and amplitude is no longer possible. As a result, the writing trajectory becomes irregular and may be subject to unpredictable perturbations that the patient is unable to correct. Letters with curved shapes tend to be decomposed into straight lines, reflecting abrupt transitions in movement direction.

Similar to the basal ganglia, the cerebellum is connected to frontal premotor areas via re-entrant neuronal circuitry. Clinical observations in patients with cerebellar lesions suggest that the corticocerebellar motor loops are involved in the selection and control of kinematic parameters for skilled movements. The cerebellum also plays an important role in monitoring motor performance by comparing premotor commands for the intended movement with sensory feedback about the actual movement taking place. This comparator function is critical for error detection and for adjusting the evolving movement to changing contextual requirements.

In summary, basal ganglia–thalamocortical and cerebellocortical motor networks are possible neural substrates of the system responsible for generating the graphic innervatory patterns that guide the execution of handwriting movements. This conclusion is supported by functional neuroimaging studies that demonstrated activation of premotor cortex, basal ganglia, and cerebellum during the performance of various writing tasks.

4. Afferent Dysgraphia: The Role of Sensory Feedback in Handwriting

Patients with afferent dysgraphia have difficulty using sensory feedback to monitor and control the execution of handwriting movements. The writing errors produced by these patients are similar to those observed in normal subjects under experimental conditions that interfere with the efficient use of visual or kinesthetic feedback. Typical findings include duplications and omissions that are especially common when writing sequences of similar letters or strokes (e.g., "shampoo" written as "shampo" or "shampooo," or the addition

of extra loops in writing letters containing repeated stroke cycles) (Fig. 3C). Writing errors are usually not detected by the patients and therefore remain uncorrected. Cursive handwriting tends to be more severely affected than writing in uppercase print. In addition to letter/stroke duplications and omissions, patients may have difficulty keeping the correct spacing between letters and words and may not be able to write in a straight, horizontal line.

Afferent dysgraphia is typically seen following right parietal lobe damage, although features of the syndrome can occasionally be observed in patients with left parietal lesions. These findings suggest that the parietal lobes in general, and the right parietal lobe in particular, play an important role in monitoring visual and kinesthetic feedback for the proper control of writing movements. Writing produced by patients with right parietal damage also frequently shows evidence of left-sided spatial neglect. However, neglect-related errors (e.g., the tendency to write on the right side of the page or the failure to cross t's and dot i's at the beginning of the word) and feedback-related errors (e.g., letter/stroke repetitions and omissions) are dissociable. Furthermore, the type of neglect that affects the peripheral aspects of writing is distinct from the type of neglect that affects the internal graphemic representations of words and leads to unilateral spelling errors during the readout process from the graphemic buffer. These observations suggest that attention interacts with the writing process at several different levels. Thus, we must distinguish between attentional resources directed toward spatially coded internal graphemic representations computed by central spelling routes and attentional resources directed toward stimuli located in external visual space, including the letters and words produced on the page during writing.

In addition to being mediated by neural systems distinct from those involved in spatial attention, afferent control of stroke and letter production may also be separate from control over the more general (i.e., nonlateralized) visuospatial aspects of writing, such as maintaining the proper horizontal orientation and keeping the correct spacing between letters and words. According to this hypothesis, the core deficit in afferent dysgraphia is the inability to use sensory feedback to keep track of the number of letters and strokes produced in handwriting, whereas the other features of the syndrome may reflect simultaneous damage to functionally independent visuoperceptual and spatial attention modules located in close anatomical proximity within parietal cortex.

Letter/stroke duplications have also been observed in the writing produced by patients with right frontal lobe lesions, and it has been proposed that these errors may represent a form of motor perseveration. In addition, features of afferent dysgraphia have been documented following cerebellar damage. This finding is consistent with the view that the cerebellum is normally involved in the monitoring of sensory feedback during the execution of handwriting movements.

IV. CONCLUSIONS

Writing is a complex cognitive activity that requires the interaction of linguistic, motor, spatial, and perceptual systems for normal performance. For these reasons, and perhaps also because it is the least frequently used language modality, writing remains a fragile skill that is highly vulnerable to disruption by brain damage.

In this article, we provided evidence that the various agraphia syndromes observed in neurological patients are interpretable within the framework of a cognitive model of normal spelling and writing. It is anticipated that the functional architecture of the model will undergo further modifications and refinements not only as a result of neuropsychological case studies but also on the basis of relevant observations in normal subjects and insights gained from computational models of spelling and writing.

Most of our knowledge concerning the neural substrates of writing comes from clinicoanatomical observations in patients with focal brain damage. Currently, there are only a few neuroimaging studies of writing in normal subjects, but it is clear that this methodology holds tremendous promise for functional localization. By combining patient lesion data with neuroimaging studies in normal individuals, it will be possible to map the cognitive operations involved in writing onto specific brain regions with much greater precision than either technique alone would permit. We are confident that future research in the related fields of cognitive neuropsychology, neurolingustics, and functional neuroimaging will lead to exciting new insights into the neural mechanisms underlying this unique cognitive skill, the development of which had the revolutionary impact of allowing human beings to communicate their ideas across space and time.

See Also the Following Articles

ALEXIA • APHASIA • BROCA'S AREA • DYSLEXIA • LANGUAGE AND LEXICAL PROCESSING •

LANGUAGE DISORDERS • READING DISORDERS, DEVELOPMENTAL • WERNICKE'S AREA

Acknowledgments

This work was supported by the Cummings Endowment Fund to the Department of Neurology at the University of Arizona and by National Multipurpose Research and Training Center Grant DC-01409 from the National Institute on Deafness and Other Communication Disorders. The authors acknowledge the contributions of Elena Plante, Lee Ryan, Theodore Trouard, and Amy Ramage to the functional neuroimaging studies reported here.

Suggested Reading

Beauvois, M.-F., and Dérousné, J. (1981). Lexical or orthographic agraphia. *Brain* **104,** 21–49.

Bub, D., and Chertkow, H. (1988). Agraphia. In *Handbook of Neuropsychology* (F. Boller and J. Grafman, Eds.). Elsevier, New York.

Caramazza, A., Miceli, G., Villa, G., and Romani, C. (1987). The role of the graphemic buffer in spelling: Evidence from a case of acquired dysgraphia. *Cognition* **26,** 59–85.

Ellis, A. W. (1982). Spelling and writing (and reading and speaking). In *Normality and Pathology in Cognitive Functions* (A. W. Ellis, Ed.). Academic Press, London.

Ellis, A. W., and Young, A. W. (1988). *Human Cognitive Neuropsychology*. Erlbaum, Hove, UK.

Hillis, A. E., and Caramazza, A. (1995). Spatially specific deficits in processing graphemic representations in reading and writing. *Brain Language* **48,** 263–308.

Margolin, D. I., and Goodman-Schulman, R. (1992). Oral and written spelling impairments. In *Cognitive Neuropsychology in Clinical Practice* (D. I. Margolin, Ed.). Oxford Univ. Press, New York.

McCarthy, R. A., and Warrington, E. K. (1990). *Cognitive Neuropsychology: A Clinical Introduction*. Academic Press, San Diego.

Rapcsak, S. Z. (1997). Disorders of writing. In *The Neuropsychology of Action* (L. J. G. Rothi and K. M. Heilman, Eds.). Apraxia: Psychology Press, Hove, UK.

Rapcsak, S. Z., Beeson, P. M., and Rubens, A. B. (1991). Writing with the right hemisphere. *Brain Language* **41,** 510–530.

Rapp, B., and Caramazza, A. (1997). From graphemes to abstract letter shapes: Levels of representation in written spelling. *J. Exp. Psychol. Hum. Perception Performance* **23,** 1130–1152.

Roeltgen, D. P. (1993). Agraphia. In *Clinical Neuropsychology* (K. M. Heilman and E. Valenstein, Eds.). Oxford Univ. Press, New York.

Shallice, T. (1981). Phonological agraphia and the lexical route in writing. *Brain* **104,** 413–429.

Shallice, T. (1988). *From Neuropsychology to Mental Structure*. Cambridge Univ. Press, Cambridge, UK.

Van Galen, G. P. (1991). Handwriting: Issues for a psychomotor theory. *Hum. Movement Sci.* **10,** 165–192.

Alcohol Damage to the Brain

JOHN W. OLNEY
Washington University, St. Louis, School of Medicine

GLOSSARY

alcoholic hallucinosis Auditory hallucinations and accompanying delusions, often of a paranoid type, occurring in a patient who is not withdrawing from ethanol but rather is engaged in ongoing drinking.

blackout A phenomenon in which an individual who is intoxicated with ethanol behaves as if he or she is in command of his or her faculties and is aware of the acts he or she is performing but upon sobering up has no recollection of having performed these acts.

delerium tremens A toxic psychosis featuring vivid and often frightening visual hallucinations and tremulousness that typically develops in chronic alcoholics 24–96 hr after cessation of ethanol intake.

fetal alcohol effects A syndrome observed in children who were exposed to ethanol *in utero* consisting of some but not all of the components of the fetal alcohol syndrome. A term currently recommended for partial syndromes primarily featuring central nervous system disturbances is alcohol-related neurodevelopmental disorder.

fetal alcohol syndrome A syndrome observed in children who were exposed to ethanol *in utero* consisting of growth deficiency, a pattern of dysmorphic facial features, and neurobehavioral disturbances.

Korsakoff syndrome A syndrome, first described by Korsakoff in 1887, characterized by peripheral neuropathy, confusion, disorientation, memory loss, and a tendency to confabulate.

Wernicke's encephalopathy A syndrome, first described by Wernicke in 1881, in which neurodegenerative changes occur in the thalamus, hypothalamus, periaqueductal region, and floor of the fourth ventricle. The clinical symptoms consist of ocular abnormalities, ataxia, and alterations in the state of consciousness varying from mild confusion to coma.

In modern society, we live with an alcohol conundrum: Alcohol (ethanol) is, and has been for much of recorded time, the euphoriant of choice for human adults who perceive ethanol as user-friendly. However, it has damaged the brains of more human fetuses than any other agent in the human environment, and today it continues to damage the brains of human fetuses. In addition, either directly or indirectly, it has caused neurodegenerative changes in the brains of countless human adults, causing them to have cognitive disturbances ranging from mild memory loss to profound dementia. Medical science has been slow in deciphering the mechanisms underlying ethanol's deleterious effects on the brain, but recent findings, particularly pertaining to fetal alcohol syndrome, are beginning to shed new light on how ethanol damages the human brain.

I. CLINICAL SYNDROMES ASSOCIATED WITH ETHANOL CONSUMPTION

A. Acute Ethanol Intoxication

Ethanol is a simple molecule that freely enters the brain and interacts with many cellular or subcellular systems. Ethanol's well-known ability to cause acute intoxication qualifies it as a central nervous system (CNS) intoxicant, a word that implies a toxic action on the CNS. However, toxic actions can be transient, reversible, and relatively harmless or long-lasting,

irreversible, and very harmful. On the basis of all available evidence, it appears that acute exposure of human adult to ethanol, even in large doses, produces an intoxication syndrome that is transient, reversible, and not damaging to the brain.

The mechanism(s) by which ethanol transiently renders an individual intoxicated is not fully understood. However, accumulating evidence suggests that ethanol interacts with, and alters the status of, several transmitter systems in the brain, and this provides the most promising clues to its intoxicating effects. There is evidence that ethanol, by either a direct or indirect action, alters the status of the acetylcholine, dopamine, serotonin, glutamate, and GABA transmitter systems. Of these systems, the two that are most likely to play a primary role in ethanol's intoxicating effects are the glutamate and GABA systems, which are responsible for most of the excitatory (glutamate) and inhibitory (GABA) transmitter activity in the brain. The basis for this interpretation is that ethanol's intoxicating effects are akin to the effects that general anesthetics have on the brain, and it is becoming increasingly well recognized that all agents currently used as general anesthetics act by one of two mechanisms. Either they block excitatory neurotransmission through NMDA glutamate receptors or they promote inhibitory neurotransmission through $GABA_A$ receptors.

In recent years, researchers have shown repeatedly that ethanol blocks the excitatory functions mediated by NMDA glutamate receptors and promotes the inhibitory functions mediated by $GABA_A$ receptors. Both of these actions (reduced excitation and increased inhibition) translate into a suppressant effect on neural circuits in the brain. Although ethanol has an immediate activating effect on human behavior, its overall net effect is one of CNS depression. For example, at excessive doses it produces a comatose state similar to that induced by general anesthetic agents. The immediate activating effects may be due to the NMDA blocking action of ethanol in that other NMDA blocking drugs are known to have a disinhibition effect on certain neural circuits. These are neural circuits that are ordinarily held under restraint by a mechanism whereby glutamate, acting at NMDA receptors on GABAergic inhibitory neurons, tonically activates these neurons causing them to maintain a high level of inhibitory tone in the circuit. When ethanol blocks the NMDA receptors, glutamate ceases activating the inhibitory neurons and these neurons cease inhibiting the circuit. In the absence of inhibition the circuit becomes hyperactive, which would explain ethanol's activating effect on human behavior. It should be noted that in these particular circuits glutamate is performing an unusual function: It is serving as a regulator of inhibitory tone and the net effect of glutamate stimulation of NMDA receptors in these circuits is inhibitory. The net effect of glutamate stimulation of NMDA receptors in many other neural circuits is excitatory so that ethanol blockade of NMDA receptors in these other circuits would have a depressant effect on behavior. Similarly, ethanol's promotion of GABAergic inhibitory transmission would have a depressant effect. Thus, the simultaneous action of ethanol at both $GABA_A$ and NMDA glutamate receptors in neural circuits that are sometimes wired in a paradoxical manner can explain why it has an immediate activating effect on human behavior that is transient and soon gives way to a more prolonged depressant effect.

B. Blackouts

It has been observed that individuals who imbibe large quantities of ethanol may appear to be in conscious control of their behavior and may behave as if they have an intact memory, but on the following day after they have sobered up they have a blank memory for important events that they participated in during the inebriation episode. This phenomenon is referred to as a "blackout." It is most likely due to the important role that NMDA glutamate receptors play in memory. It has been shown that various drugs that block NMDA glutamate receptors interfere with the acquisition of memory. While under the influence of an NMDA receptor blocking drug, laboratory animals lose their ability to acquire new information and commit it to memory. However, they are still able to recall and act on information previously learned and committed to memory. Moreover, in the future when they are not under the influence of the drug they are able to acquire new information and commit it to memory. It has also been shown that NMDA receptor blocking drugs interfere with the phenomenon known as long-term potentiation (LTP), which is a synaptic mechanism that is thought to mediate memory functions. LTP has been studied extensively in the hippocampus, a brain region that is prominently involved in memory functions. It is well established that when NMDA glutamate receptors at synapses in the hippocampus are stimulated at a high frequency they become conditioned so that subsequent application of glutamate to the receptor elicits a much more robust (potentiated) response than if the receptor had not received the

high-frequency stimulation. This conditioned status is referred to as LTP. Drugs that block NMDA receptors, including ethanol, interfere with the induction of LTP at hippocampal synapses. Thus, there is evidence at both a synaptic level and a behavioral level that ethanol, by its NMDA receptor blocking action, can interfere with memory functions in a way that would explain blackouts associated with heavy ethanol consumption. It is not known with certainty whether chronic heavy drinking with repetitive blackouts can result in permanent damage to memory mechanisms at a synaptic or molecular level or whether it entails only reversible effects on neural pathways that mediate memory function.

C. Alcoholic Hallucinosis

Alcoholic hallucinosis refers to a condition in which a chronic heavy user of ethanol displays persistent auditory hallucinations and related delusional beliefs but remains oriented, with intact memory, and does not manifest other psychotic behaviors. This is a relatively rare complication of chronic alcohol use, and the underlying mechanisms are poorly understood. However, it is not unlikely that the NMDA antagonist properties of ethanol play a role because drugs that block NMDA receptors are notorious for inducing psychotomimetic symptoms. As discussed later, drugs that have NMDA antagonist properties have the potential to cause psychotic symptoms, including auditory hallucinations and delusions, and also have the potential to cause physical injury and even death of cerebrocortical neurons. Phencyclidine (PCP; angel dust) is a well-known example of a drug that has powerful psychotomimetic actions that are solely attributable to its blocking action at a recognition site within the NMDA receptor ion channel. Another well-known NMDA antagonist that has psychotomimetic activity is ketamine, a drug that is used in human medicine as a general anesthetic. Because of its psychotomimetic properties, ketamine has recently become a popular drug of abuse (street name, special K). The psychotic reactions associated with ketamine anesthesia have been referred to as "emergence" reactions and it was learned many years ago that these reactions could be substantially ameliorated by treatment with drugs that activate $GABA_A$ receptors, sometimes called GABAmimetic agents. Recently, it has been found that PCP, ketamine, and many other drugs that have NMDA antagonist properties cause pathomorphological changes in cerebrocortical neurons of the adult rat brain, and it has been clearly demonstrated that these changes can be completely prevented if a GABAmimetic drug is administered together with the NMDA antagonist. In light of this information, the fact that ethanol has strong GABAmimetic activity as well as NMDA antagonist activity should cause it to behave as if it did not have psychotomimetic and neurotoxic properties because ethanol's potential to cause these adverse effects due to its blocking action at NMDA receptors would be cancelled out by its counterbalancing action at $GABA_A$ receptors. It is interesting to note that in the ketamine anesthesia situation the psychotic reactions were often not completely prevented because it was the practice to administer the GABAmimetic drug toward the end of the anesthesia period or in the recovery room after the brain changes conducive to psychotic behaviors had already occurred. The door was being closed after the horse was already out of the barn. In contrast, in animal experiments, the GABAmimetic drug has been administered in advance or at the same time as the NMDA antagonist, and this results in complete prevention of the neurotoxic effects of the NMDA antagonist. Thus, since ethanol has the GABAmimetic property built-in, its user is always protected against the expression of its psychotomimetic or neurotoxic potential.

The psychosis produced by NMDA antagonist drugs closely resembles a schizophrenic psychosis and this has given rise to a currently popular hypothesis that hypofunction of the NMDA glutamate receptor system may contribute to the pathophysiology of schizophrenia. Previously, the emphasis in schizophrenia research was on the dopamine system, the hypothesis being that hyperactivity of this system may be responsible for schizophrenic symptomatology. These two hypotheses are not mutually exclusive. There is evidence that dopamine is an inhibitory regulator of the release of glutamate at NMDA glutamate receptors. Thus, if the dopamine system were hyperactive, this would result in hyperinhibition of glutamate release at NMDA receptors which would produce an NMDA receptor hypofunctional state that could explain the symptoms of schizophrenia. Ethanol hallucinosis occurs in people who are not withdrawing from alcohol but rather are chronically ingesting alcohol on a steady basis. Under this condition, the NMDA transmitter system is being blocked and hence is hypofunctional, and the dopamine system is believed to be in a hyperactive state. Thus, it is reasonable to propose that dopamine hyperactivity, NMDA receptor hypoactivity, or both may be responsible for the

symptoms of alcoholic hallucinosis. Because alcoholic hallucinosis is a relatively rare condition, it seems likely that it occurs only in individuals who harbor a genetic predisposition to manifest psychotic symptoms, perhaps due to dysfunction of one or more of the transmitter systems with which ethanol interacts. The genetic predisposition may not be strong enough by itself to trigger psychotic symptoms, but ethanol's interaction with the genetically defective transmitter system(s) may be enough to tip the balance and cause the genetic predisposition to be expressed as a clinical illness. However, in the absence of a genetic predisposition, the GABAmimetic properties of ethanol may be strong enough to prevent propsychotic mechanisms from breaking through.

D. Tolerance, Dependence, and Addiction

Mechanisms underlying tolerance, physical dependence, and addiction are exceedingly complex and are not very well understood. These topics will be addressed only briefly because they are only indirectly related to the topic of alcohol damage to the brain. This is not to say that they are unimportant in relation to the brain damage problem. On the contrary, they are exceedingly important. If ethanol did not have properties that cause an individual to become psychologically and physically dependent on it, no one would drink it chronically in large enough amounts to cause brain damage either directly or indirectly. If ethanol did not have these properties, large segments of society would not habitually imbibe ethanol, and if large segments of society were not imbibing ethanol, many fewer fetuses would be exposed during critical periods when the developing brain is particularly vulnerable to damage from ethanol. It is likely that ethanol induces long-term (but not necessarily permanent) changes in transmitter systems, and such changes may be largely responsible for physical and psychological dependence and addiction. However, it has not been demonstrated that such changes represent a permanent alteration in the brain that could be considered irreversible damage.

E. Withdrawal Phenomena

1. Delirium Tremens

The withdrawal syndrome experienced by an individual who is physically dependent on ethanol is called delirium tremens because its most prominent manifestations are tremors of the extremities and a delirious mental state. The picture is one of a toxic psychosis consisting of agitation, distractability, sleeplessness, visual hallucinations, disorientation, and confusion. Although, the mechanisms are not completely understood, the symptoms reflect CNS hyperactivation, and it is very likely that an imbalance between excitatory and inhibitory transmitter systems in various brain regions is responsible. Transmitter systems that have been primarily implicated are the NMDA glutamate excitatory system, the $GABA_A$ inhibitory system, and the dopamine inhibitory system. There is good agreement among investigators that chronic heavy alcohol intake causes an upregulation of the NMDA glutamate excitatory system, a condition that develops presumably as a compensatory effort of the NMDA receptor system to overcome the persistent blockade of NMDA receptors by ethanol. The upregulation consists of an increased amount of NMDA receptor protein, an increased expression of mRNA for synthesis of NMDA receptor subunits, and an enhanced capacity of this receptor system to mediate excitatory activity. In addition, chronic ethanol intake is associated with increased extracellular concentrations of glutamate presumably reflecting increased release of glutamate at both NMDA and other glutamate receptors. In the withdrawal state, the absence of ethanol leaves the upregulated and overly responsive NMDA receptor system unblocked and free to respond in an exaggerated way to glutamate that is present in excessive concentrations in the synaptic cleft. Chronic exposure to ethanol also produces changes in the $GABA_A$ transmitter system but there is less agreement among investigators regarding the nature of these changes, one problem being that the changes are apparently different in different brain regions. The consensus appears to be that the $GABA_A$ system becomes downregulated and less capable of performing its inhibitory functions. In the withdrawal state, the dopamine system shows a marked decrease in functional activity, which apparently is fostered by hyperactivity of the NMDA glutamate system. It seems likely that the tremors associated with alcohol withdrawal are a reflection of the combined disturbances in the dopamine and NMDA receptor systems, and the psychotic manifestations, including vivid visual hallucinations, may also relate to disturbances in one or both of these systems.

2. Convulsions

Patients who are undergoing withdrawal from chronic heavy alcohol use are prone to epileptiform seizures. The seizure activity may be quite severe and can be

injurious to the brain or even life threatening. Years ago, the tendency of chronic alcoholics to have seizures was recognized but it was assumed that it was the excessive, continuous exposure to ethanol that triggered the seizures. Thus, the condition was referred to as rum fits. Today, it is recognized that chronic heavy ethanol intake *per se* does not typically trigger seizure activity; rather, it causes changes in the brain that set the stage for seizures to occur whenever there is an abrupt decrease in ethanol intake. In ordinary circumstances, brain circuits are not subject to epileptiform activity because a stable balance is maintained between glutamatergic excitatory activity and GABAergic inhibitory activity. However, any substantial decrease in inhibitory activity or increase in excitatory activity alters the balance in favor of excitation and increases the chances that seizures (runaway excitatory activity) can occur. As described previously, chronic heavy ethanol intake causes chronic blockade of NMDA glutamate receptors and a compensatory upregulation of the NMDA glutamate transmitter system. At the same time, it causes chronic overactivation of the $GABA_A$ receptor system, which produces a compensatory downregulation of the GABAergic system. Upregulation of glutamatergic excitation coupled with downregulation of GABAergic inhibition sets the stage for seizure activity, but seizures do not occur as long as steady ethanol intake continues because ethanol prevents runaway excitation by blocking NMDA excitatory receptors and activating $GABA_A$ inhibitory receptors. In the interval from 12 to 48 hr after chronic alcohol intake has been abruptly discontinued, brain circuits are in a seizure-prone hyperirritable state in which stimuli that would ordinarily be innocuous can trigger seizure activity that may be difficult to control with a standard anticonvulsant drug such as dilantin. If seizure activity in limbic circuits is not promptly brought under control it feeds on itself in a reverberating manner and can cause neurons that have NMDA receptors on their surface to undergo excitotoxic degeneration, which entails a cascade of events initiated by excessive influx of sodium and calcium through NMDA receptor ion channels. Some anticonvulsant drugs, such as barbiturates and benzodiazepines, are more effective than others in the clinical management of rum fits which is consistent with the postulated role of $GABA_A$ receptor down regulation as a causative mechanism, in that these drugs act at $GABA_A$ receptors to restore GABAergic inhibition to more normal levels.

In addition to seizures associated with ethanol withdrawal, there are individuals who have an epileptic diathesis, which implies that a state of excitatory/inhibitory imbalance already exists. In this case, even isolated episodes of acute alcohol intake may upset the balance further and trigger seizure activity.

F. Wernicke's Encephalopathy

Neuropathologists have recognized for many years that chronic heavy ethanol intake can result in a syndrome first described by Wernicke in 1881, in which neurodegenerative changes occur in the thalamus, periaqueductal region, and floor of the fourth ventricle. Individuals with Wernicke's encephalopathy typically manifest a triad of clinical symptoms consisting of ocular abnormalities, ataxia, and a disturbance in consciousness varying from mild confusion to a profound comatose state. Whether a direct action of ethanol in the brain contributes to this neuropathological syndrome is unclear. A very similar syndrome is known to occur in individuals who have severe nutritional deficiencies, especially thiamine deficiency, and the Wernicke pattern and type of brain damage can be induced in experimental animals by rendering them severely thiamine deficient. Chronic alcoholics are typically deficient in thiamine because they tend to substitute ethanol for other more nutritional sources of calories, and because ethanol intereferes with the gastrointestinal absorption of thiamine. Also strongly implicating thiamine deficiency in Wernicke's encephalopathy is evidence that treatment of patients manifesting early signs of a Wernicke syndrome with massive doses of intramuscular thiamine prevents the Wernicke pattern of brain damage and markedly diminishes its symptomatic manifestations. In fact, there has been a striking reduction in the incidence of Wernicke's encephalopathy since clinicians have adopted the practice of treating chronic alcoholics with large doses of thiamine as a prophylactic measure. Thus, it is generally believed that nutritional deficiency is primarily responsible for Wernicke's encephalopathy, and that ethanol contributes only indirectly by promoting thiamine and other nutritional deficiencies.

G. Korsakoff Syndrome

In 1887, Korsakoff described a clinical syndrome associated with chronic ethanol intake featuring amnesia, confabulation, disorientation in time and place, and peripheral neuropathy. The tendency of chronic alcoholics to suffer from nutritional deficiency

and to display a mixed clinical picture, including symptoms of both the Wernicke and Korsakoff type, has led some investigators to categorize the two syndromes together under the heading Wernicke–Korsakoff syndrome and to ascribe the overall syndrome to a multivitamin-deficiency state. Reinforcing this tendency is the observation that both syndromes tend to occur initially in an acute alcohol withdrawal context as part of a delirum tremens episode. However, the Korsakoff syndrome reportedly can occur in individuals who have no history of delirium tremens and who lack the typical symptoms of Wernicke's encephalopathy. Moreover, the pattern of brain damage in the Korsakoff syndrome, although sometimes overlapping with the Wernicke pattern, tends to involve other brain regions, including the cerebral cortex, not typically affected in the Wernicke syndrome. Therefore, it must be assumed that the Wernicke and Korsakoff syndromes are two separate conditions, but the role of ethanol vs vitamin deficiencies or other contributing causes in producing the Korsakoff syndrome remains to be clarified.

H. Alcoholic Deterioration

Progressive ataxia, degeneration of cerebellar Purkinje neurons, cerebral atrophy, and mild to severe dementia have been described by many authors as findings typically associated with chronic heavy alcohol use. In one study, loss of Purkinje neurons from the cerebellar vermis correlated with long-term daily ingestion of moderate doses of ethanol, but in other studies a definite correlation could not be established between either cerebrocortical or cerebellar atrophy and the long-term pattern or amount of ethanol intake. Dementia syndromes of mild to moderate degree have been described in chronic alcoholics, with stabilization or improvement occurring if ethanol ingestion is discontinued. Thus, there is suggestive evidence that ethanol can have deleterious effects on the adult brain which may be separate and distinct from the Wernicke and Korsakoff syndromes, but it remains to be determined whether or how ethanol directly contributes to these brain damage syndromes.

I. Fetal Alcohol Syndrome

For many centuries, perhaps several millennia, ethanol has served as the euphoriant of choice by human adults throughout the world. Undoubtedly, over that period of time countless human fetuses have been exposed *in utero* to ethanol, but it was less than 30 years ago that P. Lemoine and colleagues in France and K.L. Jones and D.W. Smith in the United States reported the first evidence that ethanol can have deleterious effects on the human fetus. For several years after these authors described the fetal alcohol syndrome there was a tendency to disbelieve that ethanol could cause such dramatic changes in the human fetus. Their description of the syndrome focused on craniofacial malformations and severe neurobehavioral disturbances, including frank mental retardation, as the primary clinical manifestations. In subsequent research it was determined that all fetuses affected by ethanol do not show all features of the syndrome. For example, some may show either craniofacial malformations or CNS effects, but not both. Moreover, sometimes the CNS effects were limited to relatively mild degrees of learning impairment and/or hyperactivity/attention deficit disorder. By convention, the term fetal alcohol syndrome (FAS) has been used to refer to the full syndrome featuring both dysmorphogenic and neurobehavioral aspects, and an alternate term, fetal alcohol effects (FAE), has sometimes been used to refer to partial or less severe syndromes that tend to feature primarily neurobehavioral aspects. A Committee of the Institute of Medicine, National Academy of Sciences proposed in 1996 that partial syndromes affecting primarily the CNS be termed alcohol-related neurodevelopmental disorder (ARND). Regardless of terminology, it is generally recognized that the CNS effects are the most debilitating component of the syndrome. It is not known what percentage of births currently or in the past qualify for a diagnosis of ARND, but there is little doubt that ethanol has contributed to more cases of developmental neuropsychiatric impairment than any other substance in the human environment and perhaps more than all other substances combined.

FAS research initially focused on efforts to identify the critical period when the developing fetus is vulnerable to the toxic effects of ethanol. Initially, there was a tendency to assume that there was only one window of vulnerability during which all aspects of the syndrome were caused. However, when it was realized that some fetuses may have predominantly craniofacial malformations and others subtle neurobehavioral disturbances, it became evident that ethanol may act by more than one mechanism and each mechanism may have its own critical period of vulnerability. Because the progenitor cells that form the skull and

bony skeleton differentiate much earlier than those that give rise to neurons in the brain, it was logical to assume that the vulnerability time window for producing the craniofacial malformations must be early, within the first trimester, whereas the vulnerable period for causing neurobehavioral disturbances would have to be later, perhaps in the second or third trimester, when neuronal progenitor cells are differentiating, migrating, and undergoing maturation. While mechanisms giving rise in the first trimester to the craniofacial malformations remain largely unknown, insights described below have been gained into third trimester mechanisms responsible for the deleterious effects of ethanol on the developing brain.

II. MECHANISTIC CONSIDERATIONS

A. Mechanisms Operative in the Adult Brain

It is generally agreed that chronic heavy use of ethanol by human adults results in a substantially higher incidence of neurodegenerative disorders than is encountered in a comparable population of individuals who are not chronic heavy users of ethanol. However, in a high percentage of cases, nutritional deficiency, uncontrolled seizure activity, head trauma, or factors other than direct impingement of ethanol on CNS neurons can account for the damage to the brain. The question that remains unanswered is whether ethanol, by a more direct toxic action on neurons, can trigger neurodegeneration in the adult brain. To address this question it may be useful to examine the known mechanisms by which ethanol interacts with neurons and consider whether ethanol, acting through any of these mechanisms, might have an injurious effect on neurons.

Of the known mechanism by which ethanol interacts with CNS neurons, there is only one mechanism that is known to have neurotoxic potential—the NMDA receptor blocking action of ethanol. In several studies, ethanol has been shown to have NMDA antagonist properties when applied to neurons in concentrations comparable to those attained in the ethanol-intoxicated human brain. It is well established that systemic administration of drugs that block NMDA glutamate receptors causes acute pathomorphological changes in cerebrocortical neurons of the adult rat brain. Low doses of NMDA antagonist drugs produce reversible pathomorphological changes in restricted cerebrocortical brain regions, whereas higher doses induce irreversible neuronal degeneration, not only in these restricted regions but also in many other corticolimbic brain regions and in the cerebellum. Most experiments demonstrating the neurotoxic properties of NMDA antagonist drugs have involved acute administration. However, it has also been shown that continuous infusion of low to moderate doses of an NMDA antagonist drug to adult rats over a 4-day period results in widespread irreversible degeneration of corticolimbic neurons. Therefore, because of its NMDA antagonist properties, ethanol would be expected, following either acute high-dose intake or chronic lower-dose intake, to be damaging to the brain, and the logical question to be addressed is how adult humans have been able to use ethanol in high doses either acutely or chronically without routinely incurring damage to the brain.

It might be argued that the neurotoxic action of NMDA antagonist drugs is a species-specific effect to which rats are vulnerable but humans are not. However, it is well-known that human adults are vulnerable to another type of toxic effect caused by NMDA antagonist drugs, namely, a psychotoxic effect. Phencyclidine is an NMDA antagonist that was introduced into human medicine as a general anesthetic in the 1950s and was soon withdrawn because of the severe psychotic reactions it produced. Subsequently, it has become notorious as a psychotomimetic drug of abuse. Ketamine, an NMDA antagonist that is less potent than PCP, has been used in human medicine as an anesthetic for several decades and is well-known to produce psychotomimetic reactions that are termed emergence reactions. In recent years, many newly developed NMDA antagonist drugs have undergone clinical trials for certain neurotherapeutic indications, and all such trials have been prematurely stopped because of the psychotomimetic side effects of these drugs. Given this clear-cut evidence that NMDA antagonist drugs cause psychotic reactions in adult humans, how is it possible for adult humans to use ethanol in high doses either acutely or chronically without its NMDA antagonist action triggering psychotic symptoms?

The most likely answer is that ethanol has a built-in mechanism that serves as an antidote to counteract and cancel out the neurotoxic and psychotoxic effects that would otherwise occur due to its NMDA antagonist properties. In support of this interpretation, ethanol has been shown to have GABAmimetic activity and it is well established that GABAmimetic drugs prevent either the neurotoxic action of NMDA antagonist drugs in adult rats or the psychotomimetic actions of these drugs in adult humans. Therefore, a

logical explanation for the observation that in human experience ethanol triggers neither acute psychotoxic nor neurotoxic reactions is that in the same circuits through which the NMDA antagonist action of ethanol would be expected to trigger psychoneurotoxic effects it exerts an action at $GABA_A$ receptors that nullifies these effects. In other circuits, ethanol may exert both NMDA antagonist and GABAmimetic actions, which combine in an additive manner to produce relaxation and a sense of euphoria. This would explain why ethanol has been perceived by human adults for many centuries as a user-friendly and relatively innocuous nectar from the gods.

From this information, an interesting question arises: Is the postulated GABA protective mechanism a totally fail-safe mechanism? A working hypothesis currently being entertained is that the GABAmimetic action of ethanol may effectively counterbalance and nullify its toxic potential in acute situations and in the context of intermittent or long-term moderate use, but under conditions of heavy chronic intake the GABA protective mechanism may gradually weaken (e.g., through receptor downregulation) and lose its ability to prevent the NMDA antagonist psychotomimetic or neurotoxic actions from being expressed. This might explain why alcoholic hallucinosis occurs in some chronic alcoholics and why a pattern of corticolimbic and cerebellar neurodegeneration occurs in others. It is noteworthy that the Wernicke pattern of neurodegeneration that can be attributed to thiamine deficiency is a pattern involving the thalamus, hypothalamus, midbrain, and brain stem, whereas another pattern of damage frequently seen in alcoholics which is not explained by thiamine deficiency involves corticolimbic and cerebellar neurons. Consistent with this hypothesis are recent reports that a single high dose of ethanol does not induce brain damage, but a steady infusion of ethanol at a dose that maintains a state of constant inebriation for 4 days results in degeneration of corticolimbic neurons in certain regions of the adult rat brain in which the NMDA antagonist mechanism would be expected to be operative.

Although this proposal is speculative, it is built on a foundation of evidence pertaining to the known properties of ethanol and is the most promising hypothesis that has been generated to date to explain how ethanol can appear to be relatively harmless for human adults but be associated with deterioration of neurological or cognitive function in more cases than can be easily explained by obvious mechanisms such as nutritional deficiency, head trauma, or uncontrolled seizure activity.

B. Mechanisms Operative in the Developing Brain

Many efforts during the past two decades to develop a suitable animal model for studying FAS met with only limited success. Microencephaly (reduced brain mass) is a characteristic finding in FAE/FAS victims, and it was demonstrated that exposure of immature rodents to ethanol in late gestation or during the first two postnatal weeks caused a reduction in brain mass. However, numerous additional animal studies failed to provide an explanation for the reduced brain mass. A modest loss of neurons from the cerebellum was described, but this cannot explain an overall reduction in brain mass, nor can it explain the types of neurobehavioral disturbances observed in FAE/FAS victims. The fact that treatment during the neonatal period (the first 2 weeks after birth) caused cerebellar neuronal loss and a reduced brain mass helped to narrow the period of peak vulnerability to the early neonatal period, which in the rodent is a time of rapid brain growth, sometimes called the brain growth spurt period. These findings have potential relevance to the human FAS because in humans the comparable developmental period when the brain growth spurt occurs is in the third trimester of gestation.

In the year 2000, three decades after FAS was first described, Ikonomidou and colleagues demonstrated that administration of ethanol to infant rodents during the brain growth spurt period triggers widespread apoptotic degeneration of neurons in many regions of the developing brain. The degeneration pattern they observed (Fig. 1) was so extensive that it could readily explain the reduced brain mass and also the myriad neurobehavioral disturbances associated with FAE/FAS. Numerous prior studies had failed to detect this dramatic pattern of neurodegeneration because optimal methods were not applied at the optimal time for detecting this type of damage, the acute period when the degenerating neuronal profiles are conspicuously evident. Examining the brains of ethanol-treated animals after a long delay interval is not optimal because at late intervals there is nothing to detect except a scattered pattern of missing neurons. If the pattern of neuronal dropout is diffusely distributed over many brain regions, tens of millions of neurons can be deleted without gross or conspicuous alteration in any given brain region. In addition, when neuronal dropout occurs during development by an apoptotic mechanism, it does not elicit a scarring response which the neuropathologists often rely on for detecting a site of brain injury long after the injury has occurred.

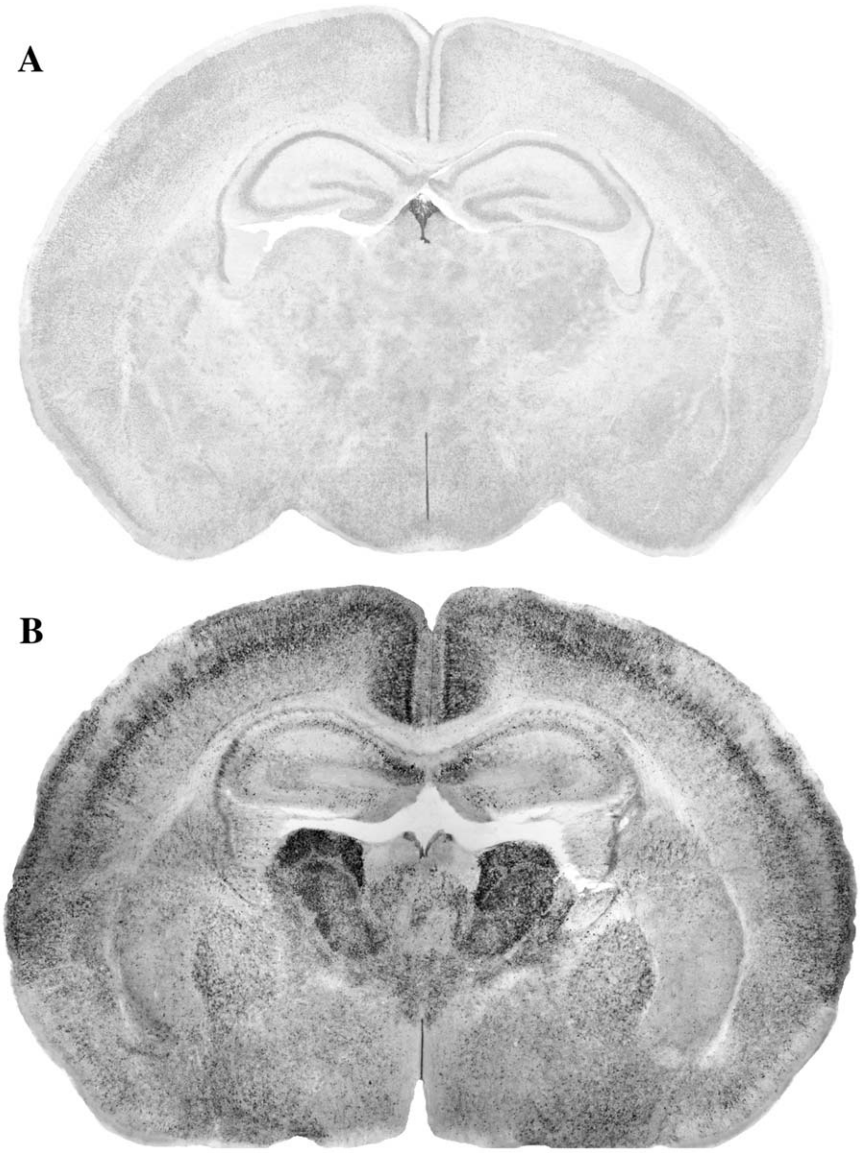


Figure 1 Ethanol-induced apoptotic neurodegeneration in the C57BL/6 mouse brain. (A) Photomicrograph of a silver-stained brain section from an 8-day-old mouse 24 hr after saline treatment. In this normal control brain, as in any normal brain during development, a few neurons in scattered distribution are undergoing degeneration. However, because the concentration of degenerating neurons in any given region is so low, the degenerating profiles are barely visible at this low magnification. (B) Photomicrograph of a silver-stained brain section from an 8-day-old mouse 24 hr after ethanol administration. Degenerating neurons (small black dots) are so abundant that they make various brain regions containing a high density of vulnerable neurons stand out in relief. Regions that are heavily affected at this brain level include the caudate nucleus, globus pallidus, hippocampus, hypothalamus, cingulate and parietal cortices, and various anterior thalamic nuclei. [From Olney, J. W., *et al.* (2000). Ethanol-induced apoptotic neurodegeneration in the developing brain. *Apoptosis* **5,** 515–521.]

In a recent neuroimaging study of the brains of living FAE/FAS subjects, there was evidence for a generalized reduction in brain mass, especially in the thalamus and basal ganglia, but no sign of a gross or conspicuous defect in any given brain region. It was concluded that the deleterious effects of ethanol on the developing brain must occur by a mechanism that reduces brain mass at a cellular or molecular level in an evenly distributed manner. The findings of Ikonomidou *et al.* in infant rodents treated with ethanol fit this description very well. The pattern of neuronal deletion was very diffuse but with the most dense degeneration

occurring in specific structures such as the thalamus and basal ganglia.

Ikonomidou *et al*. demonstrated that the cell death process induced by ethanol is an apoptotic process in which the developing neurons commit suicide. They concluded that ethanol drives neurons to commit suicide by a dual mechanism—blockade of NMDA receptors and excessive activation of $GABA_A$ receptors. They determined that the time window of vulnerability to this brain damage mechanism coincides with the brain growth spurt period. The brain growth spurt period is a period when synaptogenesis is occurring at a rapid rate, and neurons are expanding their dendritic arbors extensively to provide additional surface area for receiving newly formed synaptic connections. During this period neurons depend on a balanced level of excitatory and inhibitory input through NMDA glutamate and $GABA_A$ receptors, respectively. Either blockade of NMDA receptors or hyperactivation of $GABA_A$ receptors abnormally suppresses neuronal activity. By mechanisms that remain to be deciphered, suppressed activity during synaptogenesis translates into a message for the neuron to commit suicide. The role of the NMDA and $GABA_A$ receptor systems in this neurodegenerative syndrome was established by treating infant rodents with various agents that block NMDA receptors or agents that promote $GABA_A$ neurotransmission and showing that all such agents trigger massive apoptotic neurodegeneration during synaptogenesis. Treating infant rodents with agents that interact as either agonists or antagonists of other transmitter receptor systems did not elicit a neurodegenerative response.

An important feature of these new findings is that only a transient exposure to ethanol—a single episode of ethanol intoxication—was required to trigger extensive apoptotic neurodegeneration. In terms of blood ethanol concentrations, it required elevations in the range of 180 mg/dl, lasting for approximately 2–4 hr, to produce a robust neurodegenerative response. If blood ethanol concentrations remained at this elevated level for longer than 4 hr, the severity of degeneration escalated rapidly. Extrapolating to the human situation, it seems unlikely that maternal ingestion of a single glass of wine with dinner during the third trimester would cause neurons to degenerate in the fetal brain, but if on a single occasion several alcoholic beverages are imbibed within a period of a few hours, this might approach or exceed the threshold for causing neurons in the fetal brain to commit suicide. A major problem in assessing risk is that there is no way to know precisely how to extrapolate from rodents to humans, but prudence dictates that the extrapolation be made conservatively because of unknown variables that might cause the human fetus to be substantially more sensitive than the rodent fetus to this brain damage mechanism.

Another important point is that within the brain growth spurt period different neuronal populations were found to have different temporal patterns for responding to the apoptosis-inducing effects of ethanol. Thus, depending on the timing of exposure, different combinations of neuronal groups were deleted, which signifies that this is a neurodevelopmental mechanism that can contribute to a wide spectrum of neuropsychiatric disturbances. Consistent with this observation are recent findings by Famy and colleagues pertaining to FAE/FAS subjects who were studied in adulthood. In addition to a history of childhood hyperactivity/attention deficit disorder and varying degrees of learning impairment, a high percentage of these individuals were found to have adult-onset neuropsychiatric disturbances, including a 44% incidence of major depressive disorder and 40% incidence of psychosis. This is an important study that used a longitudinal research design to assess for the first time the full range of neuropsychiatric disturbances that human fetal exposure to ethanol can cause. Because ethanol effects on the fetus were not even suspected until 27 years ago, a prospective longitudinal study could not be completed until a cohort of individuals bearing the FAE/FAS diagnosis had grown to adulthood and begun manifesting adult-onset disturbances.

It is interesting to consider how mechanisms by which ethanol damages the immature brain compare with mechanisms by which ethanol might be damaging to the adult brain. The first important consideration is that ethanol has NMDA antagonist properties and it is known that NMDA antagonist drugs typically cause a specific type of neurodegenerative reaction in the adult brain to which the immature brain is not sensitive. The adult neurodegenerative reaction occurs by an excitotoxic mechanism and is detectable in the neuronal cytoplasm as a vacuolization reaction within 2–4 hr after drug treatment, and it has been shown that immature animals are totally insensitive to this neurotoxic mechanism. However, if immature animals during the synaptogenesis period are treated with an NMDA antagonist drug, it causes a different type of neurodegenerative response—a response that is apoptotic rather than excitotoxic and causes neurodegeneration that becomes detectable at 16–24 hr and

distributes in a pattern different from the adult pattern of degeneration. Thus, it is clear that drugs with NMDA antagonist properties have two different mechanisms by which they are damaging to the brain; one mechanism is operative only during a certain period in development, whereas the other is operative only in adulthood.

Given that ethanol has NMDA antagonist properties, it would be expected to cause one type of degeneration in the immature brain and another type of degeneration in the adult brain. However, empirical findings in animal studies indicate that ethanol produces a robust neurodegenerative reaction in the immature brain but does not, even at very high doses given acutely, produce the expected neurodegenerative reaction in the adult brain. To understand this paradox, it is necessary to analyze another major property of ethanol—its GABAmimetic property. During the period of synaptogenesis, GABAmimetic drugs are very toxic to the developing brain. They trigger massive apoptotic neurodegeneration in many regions of the developing brain. However, GABAmimetic drugs are not toxic to the adult brain. Instead, they are neuroprotective. They do not produce any neurotoxic effects when administered by themselves; when administered together with an NMDA antagonist drug, they protect against the mechanism by which NMDA antagonists damage the adult brain. Since ethanol has both NMDA antagonist and GABAmimetic properties, the logical conclusion, based on all available evidence, is that the NMDA antagonist and GABAmimetic properties of ethanol act in concert to produce a ''double-whammy'' type of damage in the developing brain, whereas in the adult brain these two properties are subtractive—one property is neurotoxic and the other is neuroprotective—with the end result being that the neurotoxic potential is there but is not expressed, except perhaps under conditions of heavy chronic ethanol abuse.

III. SUMMARY AND CONCLUSIONS

Ethanol is a simple molecule that freely enters the brain and interacts with many cellular or subcellular systems. Recent research suggests that its acute intoxicating effects and its ability to induce more permanent deleterious effects in the brain may be due to its interaction with ion channels through which many of the brain's normal functions are mediated. Of particular importance are ion channels associated with neurotransmitter function, especially channels that mediate the excitatory actions of glutamate and those that mediate the inhibitory actions of GABA. Recent evidence indicates that ethanol, by a dual mechanism (blockade of NMDA glutamate receptor ion channels and excessive activation of $GABA_A$ channels), can trigger widespread apoptotic neurodegeneration in many regions of the developing brain. This is a mechanism that can explain the reduced brain mass and lifelong neurobehavioral disturbances associated with the human fetal alcohol syndrome.

Mechanisms operative in the adult brain are different from those operative in the developing brain Wernicke's encepalopathy, a well-recognized ethanol-related adult syndrome involving neurodegenerative changes in the thalamus, midbrain, and brain stem, is believed to be due to thiamine deficiency that typically accompanies heavy chronic ethanol intake. Several lines of evidence suggest that ethanol can cause neurodegenerative changes in other regions of the adult brain by other mechanisms, but the nature of the other mechanisms remains largely a matter of speculation. Promising for its explanatory potential is a hypothesis involving the same transmitter systems (NMDA and GABA) that have been implicated in FAS. It is known that drugs that block NMDA receptors trigger a disinhibition syndrome in the adult brain that can injure or kill neurons by an excitotoxic mechanism (as opposed to the apoptotic mechanism operative during development). It is also known that drugs that activate $GABA_A$ receptors reverse this disinhibition syndrome and prevent it from injuring neurons in the adult brain. Thus, in the adult brain ethanol has neurotoxic potential because of its NMDA blocking property, but it is prevented from expressing neurotoxicity by its $GABA_A$ activating property. Recent evidence supports this concept and suggests that the protective mechanism is more than adequate to prevent expression of ethanol's neurotoxic potential in circumstances of acute heavy intake or intermittent chronic intake, but the protective mechanism may weaken and be overcome under conditions of heavy, steady chronic intake.

This proposed concept is appealing because it can readily explain the historical reality that ethanol has been used by human adults throughout the world for many centuries without apparent harm to most users, but with a subpopulation of chronic heavy abusers developing alcoholic hallucinosis and/or an irreversible neurodegenerative dementia syndrome. The other notable aspect of this historical reality is that because adult society has tended to perceive ethanol as user-friendly, countless human fetuses have been exposed

in utero by mothers who were unaware that ethanol can act in the fetal brain by a dual neurotoxic mechanism (and no counteractive protective mechanism) to drive millions of neurons to commit suicide—neurons that would otherwise have developed normally and made a positive contribution to the brain's intellectual potential. Deletion of these neurons from the developing brain results in a variety of neurobehavioral disturbances, ranging from hyperactivity/attention deficit and learning disorders in childhood to major depressive and psychotic disturbances in adulthood.

See Also the Following Articles

BRAIN DAMAGE, RECOVERY FROM • BRAIN DEVELOPMENT • DEMENTIA • DEPRESSION • GABA • NEURODEGENERATIVE DISORDERS • NEUROPSYCHOLOGICAL ASSESSMENT, PEDIATRIC • SHORT-TERM MEMORY • WERNICKE'S AREA

Suggested Reading

Bonthius, D. J., and West, J. R. (1990). Alcohol-induced neuronal loss in developing rats; Increased brain damage with binge exposure. *Alcohol Clin. Exp. Res.* **14,** 107.

Corso, T. D., Sesma, M. A., Tenkova, T. I., Der, T. C., Wozniak, D. F., Farber, N. B., and Olney, J. W. (1997). Multifocal brain damage induced by phencyclidine is augmented by pilocarpine. *Brain Res.* **752**, 1.

Famy, C., Streissguth, A. P., and Unis, A. S. (1998). Mental illness in adults with fetal alcohol syndrome or fetal alcohol effects. *Am. J. Psychiatr.* **155,** 552.

Harris, R. A., Proctor, W. R., McQuilkin, S. J., Klein, S. J., Mascia, M. P., Whatley, V., Whiting, P. J., and Dunwiddie, T. V. (1995). Ethanol increases $GABA_A$ responses in cells stably transfected with receptor subunits. *Alcohol Clin. Exp. Res.* **19,** 226.

Ikonomidou, C., Bosch, F., Miksa, M., Vockler, J., Bittigau, P., Pohl, D., Dikranian, K., Tenkova, T., Turski, L., and Olney, J. W. (1999). Blockade of glutamate receptors triggers apoptotic neurodegeneration in the developing brain. *Science* **283,** 70.

Ikonomidou, C., Ishimaru, M. J., Woznial, D. F., Price, M. T., Tenkova, T., Dikranian, K., and Olney, J. W. (2000). Ethanol-induced apoptotic neurodegeneration and fetal alcohol syndrome. *Science* **287,** 1056.

Jevtovic-Todorovic, V., Kirby, C. O., and Olney, J. W. (1997). Isoflurane and propofol block neurotoxicity caused by MK-801 in the rat posterior cingulate/retrosplenial cortex. *J. Cereb. Blood Flow Metab.* **17,** 168.

Lovinger, D. M., White, G., and Weight, F. F. (1989). Ethanol inhibits NMDA-activated ion current in hippocampal neurons. *Science* **243,** 1721.

Olney, J. W., and Farber, N. B. (1995). Glutamate receptor dysfunction and schizophrenia. *Arch. Gen. Psychiatry* **52,** 998–1007.

Olney, J. W., Labruyere, J., and Price, M. T. (1989). Pathological changes induced in cerebrocortical neurons by phencyclidine and related drugs. *Science* **244,** 1360–1362.

Olney, J. W., Labruyere, J., Wang, G., Wozniak, D. F., Price, M. T., and Sesma, M. A. (1991). NMDA antagonist neurotoxicity: Mechanism and prevention. *Science* **254,** 1515.

Pierce, D. R., and West, J. R. (1986). Alcohol-induced microencephaly during the third trimester equivalent: Relationship to dose and blood alcohol concentration. *Alcohol* **3,** 185.

Reich, D. L., and Silvay, G. (1989). Ketamine: An update on the first twenty-five years of clinical experience. *Can. J. Anaesth.* **36,** 186.

Stratton, K. R., Howe, C. J., and Battaglia, F. C. (Eds.) (1996). *Fetal Alcohol Syndrome: Diagnosis. Epidemiology, Prevention, and Treatment*. National Academy Press, Washington, DC.

Victor, M., Adams, R. D, and Collins, G. H. (1989) *The Wernicke–Korsakoff Syndrome,* 2nd ed. Davis, Philadelphia.

Alertness

VERITY J. BROWN and ERIC M. BOWMAN
University of St. Andrews

GLOSSARY

ascending reticular activating system The collection of forebrain projections, arising in the brain stem reticular formation, that were traditionally regarded as a single arousal system.

attention A variety of cognitive processes that enable selectivity of information processing.

readiness A state of premotor activation in preparation for action.

reticular formation Tissue in the central brain stem, from the spinal cord to the diencephalon, having a net-like appearance in histological section due to the arrangement of fiber bundles.

vigilance A form of attention involving a state of expectation or preparedness for the receipt and processing of specific information.

Alertness refers to a state of enhanced readiness to receive and process information and to respond. When a subject is alert, stimuli are processed more efficiently (indicated by improved response accuracy) and responses are initiated more rapidly (indicated by shorter response latencies). Alertness carries connotations of arousal, vigilance, and readiness, although it may be distinguished from all of these.

I. FUNCTIONAL CONSIDERATIONS

A. Arousal and Alertness

Alertness differs from arousal in that alertness refers specifically to a cognitive state, whereas arousal is a more general term that includes peripheral physiological states as well as forebrain activation and cognitive state. Arousal and alertness are related in that a state of cognitive alertness often accompanies a state of high physiological arousal. Nevertheless, physiological arousal is associated with activity of the sympathetic nervous system, which is manifest in various measures (including increased heart rate and pupil dilation) that do not necessarily correlate with subjective mental alertness. The converse is also the case: Alertness indicated by forebrain cortical activation, as when directed to the processing of task relevant stimuli, is not necessarily associated with an increase in arousal. Thus, a state of high alertness is an attentive state, whereas a state of high physiological arousal may or may not be associated with a particular attentive state. Factors that increase alertness may also increase arousal and vice versa, but the two processes are psychologically, physiologically, and neuroanatomically distinct.

In 1909, Yerkes and Dodson proposed that an inverted U-shaped function described the relationship between arousal and performance. The Yerkes–Dodson

law states that (i) increasing arousal is associated with improved task performance to a certain (task-specific) optimum and thereafter further increases in arousal impair performance and (ii) the optimum level of arousal for a given task is inversely related to task difficulty. Although this idea has been hugely influential and seems to apply to many situations, there are problems with it, not the least being the assumption of a unitary process of arousal. However, by regarding arousal and alertness as distinct processes, the Yerkes–Dodson inverted-U function can be seen to be a product of both of these processes. Optimal performance of any task is likely to be achieved in a state of high alertness rather than high arousal *per se*. The rising portion of the Yerkes–Dodson function can be accounted for to the extent to which arousal and alertness correlate; increasing arousal improves alertness, which improves performance. However, a state of high physiological arousal can result in physiological reactions that interfere with cognitive processes, which could account for the falling portion of the Yerkes–Dodson function.

The second part of the Yerkes–Dodson law states that the optimum level of arousal decreases with increasing task difficulty. Tasks that are cognitively more challenging would almost certainly be improved by higher levels of alertness. One way to achieve higher alertness is by increasing arousal. Nevertheless, arousal also increases the probability of physiological effects that might be distracting, such as a racing heart rate, sweating, and an increase in respiratory rate. There might also be distracting cognitive factors, such as anxiety if arousal is induced by aversive stimuli. It is possible that tasks that are more cognitively demanding might also be more susceptible to disruption by such factors. In these circumstances, arousal and alertness conflict: Arousal can increase cognitive alertness but, as a side effect, cognitive performance suffers from sympathetic nervous system activation. Thus, the relationship between task difficulty and arousal might depend on the extent to which the side effects of arousal compromise cognitive function.

B. Vigilance and Alertness

Just as alertness can be distinguished from arousal, it can also be distinguished from vigilance. Vigilance has also been described as the focused readiness to detect and respond to a given stimulus. When a warning is given prior to a stimulus, the response to the stimulus is improved—the warning is said to "alert" the subject to a forthcoming event. This "alerting" effect of cues actually refers to an increase in the state of vigilance. Vigilance, therefore, refers to directed alertness—one can be vigilant for a specific event, whereas alertness is a state of more generalized readiness.

Vigilance also carries connotations of effort, in contrast with alertness, which does not carry such connotations. Thomas Jefferson did not say the price of freedom is eternal alertness but rather, eternal vigilance. With vigilance, one is alert to particular possibilities, threats or dangers, and signs or signals. Eternal vigilance is possible to imagine, whereas eternal alertness is impossible. However, eternal vigilance is not easy because vigilance requires effort. The cost of vigilance is clearly seen in the laboratory as the "vigilance decrement"—a decline in performance over time. In a classic vigilance task, the target stimuli to which the subject must respond are temporally unpredictable. They may also be spatially unpredictable or of low intensity. When a subject performs such a task, over time the detection rate of targets decreases and the reaction time to targets increases. This is the vigilance decrement and it is greatest under conditions in which the targets are most difficult to detect and the requirement for vigilance is greatest. Not only is adequate performance of a target detection task difficult to maintain over time but also the effort involved is tiring.

Motor readiness is also called phasic alertness. This refers to the short duration (typically measured within a second rather than over minutes or even hours) performance enhancement, which is related to temporal expectation of targets. Reaction times to targets are faster as the temporal probability of the target increases. Phasic alertness is related to vigilance and may indeed reflect operation of similar processes.

C. The Orienting Response

Orienting refers to the movement of the head and eyes to foveate a target of interest. Bringing the most sensitive region of the retina to a target of interest improves the visual processing of that target. Covert orienting refers to attentional orienting (the movement of the mind's eye), which may occur in the absence of an overt orienting response. When a subject is alert, orienting responses are more rapid. Orienting is so fundamental to information processing that it might be regarded as a manifestation of alertness, with no

possibility of making an operational distinction between the state of alertness and the orienting response. Alternatively, the orienting response might be improved by increased alertness in the same way that vigilance performance is improved by increased alertness, although neither one is synonymous with alertness. This is perhaps most obvious when one considers disorders of consciousness in which there may be evidence of visual tracking and smooth pursuit eye movements in a subject who is otherwise unresponsive. The patient is unconscious but is regarded as alert solely because of the presence of such orienting responses.

Thus, alertness is defined here as a general state of cognitive readiness, reflected in cortical arousal. It is distinct from physiological arousal and from the demanding and costly processes of vigilance and phasic motor readiness. The processes are related in that high arousal may be accompanied by increases in alertness and increased overall alertness will benefit specific cognitive functions, including vigilance performance.

II. SUBJECTIVE ALERTNESS

The most common method of measuring alertness is to ask subjects how they feel, often employing a checklist of adjectives or paired adjectives (e.g., sleepy–awake, anxious–relaxed, and lethargic–energetic). The problem with this approach is that subjective measures do not always correspond to performance measures, such as measures of reaction time or problem-solving tasks. This may be because the relationship between alertness and performance is not necessarily linear (see Section I.A). Alternatively, it may be because the performance measures commonly used do not measure alertness directly but another function, such as vigilance. Finally, it could be because responses to such adjective checklists are contaminated by arousal level and therefore they are not good indicators of alertness. Whatever the difficulty, it is clear that resorting to adjective checklists is less than satisfactory.

III. CORTICAL ELECTROENCEPHALOGRAM

An alternative method to assess alertness is to measure neural activity using an electroencephalogram (EEG). The EEG is the summed electrical activity (postsynaptic potentials) of many thousands of neurons. It is recorded noninvasively using scalp electrodes. Because the site of recording is on the scalp, much of the recorded signal emanates from cortical neurons. When neurons fire in synchrony, the EEG has a pattern of high amplitude and low frequency. Desynchrony, when the neurons fire independently, is reflected as low amplitude and high frequency in the EEG. The amplitude and frequency of the neural activity correlate with sleeping and waking states. Four different patterns of neural activity may be identified: alpha, beta, theta, and delta activity. Theta waves are high-amplitude, low-frequency (4–7 cycles per second) waves that characterize the period of drowsiness before the delta waves (fewer than 3 cycles per second) of deep, so-called "slow-wave" sleep are seen. When subjects are awake, two patterns of neural activity are recorded in the EEG—alpha waves and beta waves. In quiet resting, particularly when the eyes are closed, alpha waves are recorded with a frequency of about 10 cycles per second. The higher amplitude, lower frequency pattern that defines alpha waves reflects neural synchrony. When subjects are alert, low-amplitude, high-frequency beta waves (15–30 cycles per second) dominate the EEG: This is the state of desynchrony. Desynchrony in the EEG can be elicited by stimulating the subject with a loud noise or by engaging the subject in a mental activity such as problem solving. These different patterns of neural activity recorded in the EEG represent a continuum of neural states rather than absolute qualitative differences. Thus, in the EEG trace, it is possible to identify periods of alpha activity or synchrony and beta activity or desynchrony, but it is not necessarily possible to determine when one type of activity gives way to another. The presence and proportion of beta activity in the EEG are correlated with a subjective sense and behavioral evidence of alertness in the subject.

The EEG also has features that may be related to vigilance and motor readiness; therefore, this technique might be a useful method to distinguish alertness from vigilance and motor readiness. The noise in the EEG signal is such that it is not easy to identify changes in the raw electrical signal evoked by stimuli. However, if the signals are averaged over many occurrences of an event, it is possible to extract an event-related potential from the background noise. In vigilance tasks, sustained negativity in the EEG reliably follows a warning signal that precedes a stimulus event. This particular scalp potential, with increasing negativity in the period preceding the expected target, is referred to as the N1. It is also called the expectancy wave or contingent negative variation (CNV) because it is elicited by

neither the warning signal nor the stimulus event alone but by the expectancy of the target following the warning signal due to the contingency between them. It is not possible to conclude with certainty that the CNV does indeed reflect a single psychological process (by implication, the warning-target contingency) rather than the totality of multiple processes, which might include orientation to the warning signal and expectancy of the target stimulus. Nevertheless, the CNV appears to be a correlate of vigilance and the issue of whether the CNV reflects a single process informs the debate about whether vigilance is a single process. Another negativity in the EEG, the N2, is recorded over secondary motor cortex, preceding the onset of movement. This is called the readiness (or Bereitschafts) potential. This negative potential reflects movement preparation or motor readiness. The ''vigilance decrement'' is reflected in the overall amplitude of the CNV and the motor readiness potentials; therefore, there is support for the contention that these negative brain potentials reflect the operation of effortful attentional vigilance.

IV. MODULATION OF ALERTNESS

Knowing the determinants of alertness has implications for the ability to control levels of alertness, which in turn has economic implications. We live in a 24-hr-day society, in which many industries and services operate in shifts. Shift workers want to feel alert and are expected to maintain optimal performance even when working during the night. Nevertheless, mistakes of all kinds are more likely during the early morning hours, particularly when body temperature is at its minimum, indicating the low point in the circadian cycle. Therefore, there is considerable interest in methods to improve alertness, with implications for health and safety.

Alertness shows diurnal variation. There are individual differences in cycles of alertness, but generally after a normal night of sleep people are most alert during the morning and experience a ''low point'' in alertness in the mid-afternoon, which rises again in the early evening. It is most difficult to maintain alertness if one is awake during the early morning hours, between 2 and 6 AM. Alertness is also modulated by time since sleep; that is, with an increasing sleep deficit, there is decreasing alertness.

Alertness is altered by ingested substances. Apart from drugs, food can modulate alertness. Some food items have psychoactive effects. For example, certain foods, such as warm milk, turkey, or bananas, facilitate sleep. Turkey and bananas have a known direct physiological effect: They are good dietary sources of tryptophan, which is a precursor chemical for serotonin. Milk also contains tryptophan, but the quantities are lower: The value of milk, particularly, warm milk, could well have more to do with stimulus-induced expectations (the warm milk might be a conditioned stimulus, associated with a bedtime routine). Nevertheless, although a banana at bedtime might improve sleep quality (which might perhaps impact alertness the following day), it is not obvious that a banana at breakfast would make a person less alert. Indeed, there is evidence that it might improve alertness. Dietary supplements of tryptophan are reported to be useful to reduce the symptoms of depression, including psychomotor retardation. Glucose is an important substance for neural functioning. Because there are no stores of glucose in the brain, availability in the blood is critical. Hunger, or, more specifically, low blood glucose, is associated with feelings of mental and physical fatigue. Conversely, reaction times are faster (both during a baseline period and after a glucose drink) if the blood glucose values are high or increasing. Epinephrine does not cross the blood–brain barrier, but when administered peripherally it improves alertness and enhances performance on cognitive tests. One of its actions is to cause the release of stores of glucose from the liver and thus increase blood glucose levels. Interestingly, with increasing age, glucose regulation appears to become less efficient and this might account for some of the decrement in mental acuity that accompanies aging.

Alertness varies according to the demands of current activity, with internal cognitive factors (such as feedback following performance errors) or a change in the task demands (such as an increase in the density of traffic on the road for a motorist) modulating levels. An increase in activity or mild exercise are known to improve alertness. A modern office is often warm, the chairs are comfortable, and computers generate a constant soporific hum—the ideal conditions for sleep. Alertness can be improved by reducing the temperature and making the inhabitants slightly uncomfortable. The occurrence of an unpredictable event or a change in posture and activity are arousing, improve cardiovascular function, and also improve alertness.

Alertness is improved by napping. A prophylactic nap, before a long period without sleep, is beneficial. Cat-napping is also beneficial, particularly a nap lasting up to 20 min, which is sufficient to be refreshing

but not disorienting on awakening. If a sleeper is awakened from rapid eye movement sleep, a period of which would be inevitable in a longer nap, he or she will be less alert and more disoriented than if he or she had less overall sleep.

Night-shift workers benefit from bright lighting in their working environment. The beneficial effect of this is likely to be due to the influence of the circadian clock, which is reset by environmental light.

V. ANATOMICAL BASIS OF ALERTNESS

The anatomical basis of alertness is likely to be common to mammals; therefore, in addition to the information derived from clinical studies of patients with brain damage, the anatomy of arousal systems can be investigated in laboratory animals. In the following sections, brain regions commonly agreed to be involved in cortical arousal and attention are discussed with regard to how these functions relate to alertness.

A. The Mesencephalic Reticular Formation

Work in the 1930s and 1940s established the importance of the brain stem reticular formation and its ascending projections in the sleep–wake cycle and arousal. In experiments in the cat, Frédéric Bremer had shown that knife cuts at different levels of the neuraxis had different effects on the patterns of sleep and wakefulness. A knife cut made between the inferior and superior colliculi (which Bremer called the "cerveau isolé" preparation) resulted in a persistent sleep-like state from which the animal did not recover. The pupils were constricted and the EEG showed only synchronized activity. If a posterior cut was made severing the spinal cord from the brain (which Bremer called the "encephalé isolé" preparation), the cat, although paralyzed, showed sleep–wake cycles with normal EEG signs of alternating synchrony and desynchrony.

In the 1940s, the research of Guiseppe Moruzzi and Horace Magoun firmly established the functional significance of these ascending projections using electrophysiological techniques: recording of single neurons and electrical stimulation. In their classic 1949 paper, they reported that stimulation of the reticular core of cats (particularly of the rostral mesencephalon) results in cortical EEG desynchrony and a change in behavior to alert wakefulness.

The reticular formation-to-forebrain projections, severed by Bremer's knife cuts and stimulated by Moruzzi and Magoun, became known collectively as the ascending reticular activating system, reflecting their supposed importance in arousal and activation. The ascending reticular projections are to thalamic nuclei, the hypothalamus, and basal forebrain. Integrity of the ascending reticular activating system is essential for an animal to be alert and awake. If these pathways are severed, the animal is in a state of coma; if stimulated, the animal shows EEG desynchronization and behavioral activation.

The reticular formation continues to be regarded as an important structure in alerting processes. However, it is now accepted that the reticular formation is not merely a single "wakefulness center," the loss of which results in sleep and the stimulation of which results in arousal. Rather, the reticular formation is composed of multiple areas, with distinct functional and neurochemical contributions, that collectively are able to gate sensory information as well as modulate cortical excitability and motor output. In particular, it was ascertained that a midpontine transection (between Bremer's cerveau isolé and encephalé isolé preparations) resulted in a large *reduction* in sleep. Bremer interpreted the effects of the midcollicular section as due to a loss of forebrain activation and arousal because sensory information was no longer passed to the cortex. The midpontine section also severed sensory input to the forebrain but resulted in a cat that showed very little sleep. This suggested that the coma resulting from Bremer's midcollicular section was due to damage to an arousal mechanism, but that there were also complementary sleep-producing mechanisms, which were isolated by the more posterior knife cut. Following the midpontine section, there are cycles of synchrony and desynchrony in the EEG. During the periods of desynchrony, pupil dilation and visual tracking (with pursuit eye movements, vergence, and accommodation) are seen in response to visual stimuli, suggesting normal arousal mechanisms.

Just as the concepts of arousal, activation, alertness, and attention have been distinguished and dissociated, so to have the functions of the reticular formation. In particular, the variety of neurochemical systems arising in the mesencephalon necessitates that the traditional view of the reticular formation be replaced by consideration of the function of a multiplicity of brain stem neurochemical systems.

B. Thalamus

The thalamus, located in the diencephalon, is functionally interposed in the sensory and motor pathways from the brain stem to the basal ganglia and cortex. The thalamus is subdivided according to functional criterion, with the principal component being sensory and motor relay nuclei and the intralaminar nuclei.

1. Thalamic Relay Nuclei

Each of the relay nuclei may be distinguished and defined by their cortical connectivity. For example, the medial and lateral geniculate nuclei are auditory and visual, respectively. The anterior group is connected to cingulate cortex and has "limbic" functions, the medial group is connected to prefrontal cortex, presumably mediating cognitive, or so-called executive, functions, whereas the lateral group (which includes the pulvinar) is associated with visual and somatosensory association cortices and premotor cortex and mediates higher order sensorimotor integrative functions. The relay nuclei are not connected to each other but, rather, might be regarded as independent channels of information.

The oscillation of thalamic cells is the origin of the synchronous activity measured in the cortical EEG. Release of acetylcholine in the thalamus, which precedes EEG desynchrony and wakefulness, returns the thalamic cells to the mode in which information can be relayed to cortex. Release of acetylcholine in the thalamic relay nuclei depolarizes the relay cells, increasing their excitability. There is a simultaneous hyperpolarization of the local inhibitory interneurons: Because this releases the relay cells from local circuit inhibition, the likelihood of the relay cell firing is further increased. Thus, acetylcholine release in thalamic relay nuclei influences the speed, likelihood, and veracity of information transmission to cortex.

2. Intralaminar Nuclei

The intralaminar nuclei comprise cell groups located in the paramedian thalamus, within the white matter called the internal medullary lamina. The mesencephalic reticular formation provides noradrenergic and cholinergic input to the intralaminar nuclei. There are also inputs to the intralaminar nuclei from cortex, globus pallidus, cerebellum, and spinal cord. Output is to the dorsal (caudate nucleus and putamen) and ventral (nucleus accumbens) striatum of the basal ganglia and to widespread areas of cortex. The anterior intralaminar nuclei project to prefrontal cortex, parietal cortex, and primary sensory areas. The posterior nuclei project to the parietal cortex and premotor (including the frontal eye fields) cortex. These intralaminar–cortical projections are likely to provide the route by which thalamus directly influences cortical arousal. Whereas the cortical projections of the relay nuclei transmit specific sensory and motor information to the relevant areas of cortex, the cortical projections of the intralaminar nuclei do not follow the boundaries of cortical areas. Electrical stimulation of the intralaminar nuclei results in eye movements. When neurons in the intralaminar nuclei are driven by input from the mesencephalic reticular formation, EEG desynchrony results. Thus, the intralaminar nuclei can mediate the effects of mesencephalic influence on the activity of frontal cortical–basal ganglia–thalamic circuits carrying complex sensory and motor information as well as directly influencing cortical tone.

3. The Thalamic Reticular Nucleus

Reflecting the organization of the relay nuclei, the thalamus is often referred to as the "gateway" to cortex. Surrounding the thalamus on its anterior, ventral, and lateral surfaces is the thalamic reticular nucleus, which is a sheet of GABAergic cells. Corticothalamic and thalamocortical fibers pass through the thalamic reticular nucleus and fibers passing in both directions have axon collaterals that synapse in the reticular nucleus. Within the thalamic reticular nucleus, signals are segregated by sensory modality, maintaining the topography accorded by their thalamic and cortical origins. The cells of the thalamic reticular nucleus do not project to cortex but rather to the thalamus (including to nuclei other than those from which they received signals) and to other sectors of the thalamic reticular nucleus. These connections imply that the reticular formation is privy to information passing between thalamus and cortex and can moderate that flow of information. Activity in the thalamic reticular nucleus may gate the flow of information from thalamus to cortex by sharpening receptive fields and response times of thalamic neurons and by modulating cortical arousal.

The thalamic reticular nucleus also receives cholinergic input from the reticular formation, specifically from the cholinergic cells of the laterodorsal and pedunculopontine tegmental nuclei, as well as from the

basal forebrain nucleus basalis of Meynert. The reticular nucleus also plays a role in the origin of EEG desynchrony by hyperpolarizing the thalamic relay cells, increasing the probability of burst firing and oscillatory activity. Because of the topography of cortical connections, the thalamic reticular nucleus is able to control the functional state (synchronous oscillatory activity as opposed to desynchronous activity compatible with information transmission) of individual specific channels to cortex.

In a series of studies in the 1970s, Skinner and Yingling examined the role of the system comprising the reticular formation, the thalamic reticular nucleus, and the corticothalamic circuits. Yingling and Skinner proposed that the circuits linking cortex and thalamus control the implementation of intention. The input from the mesencephalic reticular formation to the thalamic reticular nucleus is able to interrupt this pathway by controlling the inhibition by thalamic reticular nucleus of specific thalamic relay nuclei. Position emission tomography and functional magnetic resonance imaging and functional magnetic resonance imaging scans in human subjects have confirmed the role of thalamus and the mesencephalic reticular formation when subjects are mentally engaged and performing a task compared to resting scans.

VI. NEUROCHEMICAL BASIS OF ALERTNESS

A. Norephinephrine

The principal noradrenergic cell groups are found in the locus coeruleus, the lateral tegmentum, and the dorsal medulla. The most important of these for the processes discussed here, being implicated in processes of arousal and attention, are the neurons of the locus coeruleus. These neurons project widely to the forebrain, including thalamus and hypothalamus, cerebellum, basal forebrain, hippocampus, and throughout the neocortex. They also have descending projections to the sensory nuclei of the brain stem and to the spinal cord. Inputs to the locus coeruleus contrast with these widely distributed outputs because they are restricted to two nuclei of the rostral medulla—the nucleus paragigantocellularis and the nucleus hypoglossi prepositus. The locus coeruleus does not receive direct sensory input, or indeed any direct input from the forebrain, but rather excitatory sensory input is relayed through these medullary nuclei.

The activity of the locus coeruleus varies with the sleep-wake cycle. The neurons are most active during alert waking and least active during rapid eye movement sleep, even though the cortical EEG shows desynchrony under both of these conditions. Pharmacological stimulation of the locus coeruleus in anesthetized cats results in desynchrony in the EEG.

Neurons of the locus coeruleus respond to sensory input, particularly salient–novel or noxious–environmental stimuli. On the other hand, these neurons are quiet when the subject is awake and behaviorally alert but merely engaged in activities such as grooming or eating. This suggests that the neurons are responding to external stimuli but not to internal stimuli. However, although the neurons respond to salient, unexpected events, the response of the locus coeruleus is not dependent simply on the sensory properties of stimuli but also on their relevance. For example, in vigilance tasks the neurons are responsive to targets but not to nontargets. Furthermore, these neurons show a vigilance decrement, with a decrease in their responsivity over time in the task that corresponds to the behavioral decline in performance. This evidence suggests that locus coeruleus norepinephrine may play a role in vigilance.

B. Acetylcholine

Acetylcholine is metabolized from choline and acetyl coenzyme A by choline acetyltransferase. Choline is available in food such as egg yolks, or by breakdown of phosphatidylcholine, more commonly known as lecithin, which is used as an emulsifier in foods such as chocolate. Choline is the rate-limiting step in the production of acetylcholine, such that insufficiency of choline in the diet can lead to deficits in acetylcholine, whereas dietary supplements can increase the production of acetylcholine. Acetylcholine is broken down in the synaptic cleft by acetylcholinesterase; thus, acetylcholinesterase inhibitors, such as physostigmine or tetrahydroaminoaridine, have the effect of increasing the availability of acetylcholine in the synapse. There are two types of cholinergic receptors—nicotinic (found in striated muscle and in the central nervous system) and muscarinic (found in smooth muscles and in the central nervous system).

The two brain stem cholinergic groups (at the junction of the pons and mesencephalon) are the pedunculopontine tegmental nucleus, giving rise to the dorsal tegmental pathway, which innervates the

thalamus, and the laterodorsal tegmental nucleus, giving rise to the ventral tegmental cholinergic pathway, which innervates the thalamus and hypothalamus. Brain stem cholinergic neurons are active when the EEG shows desynchrony (i.e., during alert waking and during rapid eye movement sleep). Furthermore, an increase in activity of these neurons precedes the onset on desynchrony, suggesting a causal relationship.

In the forebrain, the septal cholinergic cell groups give rise to projections to the hippocampus, whereas the basal forebrain nucleus basalis of Meynert projects to the cortex and the thalamic reticular nucleus. These basal forebrain cholinergic projections are intimately related to the brain stem cholinergic system and represent a functional extension of it. The forebrain cholinergic cells receive input from the ascending reticular cholinergic cells. Like brain stem cholinergic cells, the activity of basal forebrain cholinergic neurons correlates with EEG desynchrony and shows the same circadian fluctuations related to the sleep–wake cycle. Cortical cholinergic innervation from the basal forebrain has been implicated in states of attention, particularly sustained and selective attention. In the case of well-practiced performance, with predicable and expected stimuli, cortical cholinergic input from the basal forebrain maintains vigilance performance. As such, this system is involved in attention and does not determine levels of alertness, although the basal forebrain is likely to be the substrate in which attentional performance is modified by changes in alertness.

It is not possible to be certain where systemic drugs exert their effects. In the case of cholinergic drugs, it could be at the level of the relay nuclei of the thalamus, the thalamic reticular nucleus, or cortex, all of which are believed to be important in alertness and attentional processes but in different ways. Cholinomimetics (e.g., nicotine) are alerting and improve cognitive performance of normal humans and patients with Alzheimer's disease. Conversely, anticholinergics (e.g., scopolamine) impair the cognitive function of normals and have sedative effects. The cognitive deficits of scopolamine most likely reflect specifically the role of acetylcholine cognitive functions rather than being due to the sedative effects *per se*: Amphetamines increase arousal and alertness but exacerbate the cognitive deficits following scopolamine. The effects of cholinomimetics in normal humans and patients with Alzheimer's disease (who have measurable loss of cortical acetylcholine) are more likely due to facilitation of attention and the efficient allocation of limited cognitive resources rather than enhancement of memory.

C. Dopamine

Dopamine-containing cells of the mesencephalon are located in the substantia nigra (A9 cell group), the ventral tegmental area (A10 cell group), and the retrorubral nucleus (A8 cell group). The cell projections are used to define two systems: The mesostriatal (from the substantia nigra and ventral tegmental area to the dorsal striatum, as well as other nuclei of the basal ganglia) and the mesolimbic and mesocortical (from the ventral tegmental area to the ventral striatum and frontal cortex).

Given the specificity of these projections, it is reasonable to assume that the functions of these pathways should be easily defined. The mesostriatal pathway is known to degenerate in Parkinson's disease, which results in imbalances of activity in circuits of the basal ganglia of the forebrain and the attendant motor impairments of the disease. The function of dopamine in this pathway can be described as motor activation: Akinetic patients with Parkinson's disease (with degeneration of the mesostriatal dopamine pathway) have been reported to declare that they desire to move but are unable to initiate the action. Conversely, the drug amphetamine, one of the actions of which is to stimulate the release of dopamine in this pathway, is dubbed "speed" for its motor activation effects. The mesolimbic–mesocortical pathway can also be stimulated by drugs, and the ability of a drug to stimulate dopamine in *this* system is predictive of a drug's potential for abuse in humans and self-administration in animals.

The spontaneous activity of dopaminergic neurons does not correlate with the sleep–wake cycle. Dopaminergic neurons respond to environmental stimuli but not in the same manner that would be expected of a sensory system. In particular, the neurons are likely to respond initially to the presentation of a primary reinforcer such as food when the reinforcer is unexpected. Over repeated presentations, when the primary reinforcer is no longer unexpected, the neuronal response is seen to diminish, while preceding predictive stimuli come to elicit the neuronal response. Thus, it is thought that the mesostriatal and the mesolimbic–mesocortical dopamine systems play a role in motivation and behavioral activation in the pursuit of a goal.

D. Serotonin

The raphe nuclei, running rostro-caudally along the midline of the pons and medulla, contain the majority of the serotonergic neurons of the brain. The posterior groups send descending projections to the cerebellum and to sensory and motor neurons and also modulate neuronal excitability. The ascending forebrain projections arise from the anterior groups. The dorsal raphe projects to frontal cortex and striatum and the median raphe projects to the hippocampus and the septum.

Neurons of the raphe nuclei fire tonically during active alert wakefulness when there is EEG desynchrony, but they fire at very low rates during sleep and are silent during rapid eye movement sleep when the EEG is also desynchronized. However, lesions of the raphe nuclei result in reduction of forebrain serotonin and a correlated reduction in time spent sleeping. Furthermore, drugs that acutely reduce forebrain serotonin also lead to a reduction in sleep. With chronic inhibition of serotonin synthesis (maintaining levels of serotonin below 10% of normal), there is nevertheless a partial recovery of sleep. This evidence suggests that serotonin is necessary for the expression of normal sleep–wake cycles, although serotonin does not appear to be necessary for initiation of sleep.

A role for serotonin in cognitive processes is suggested by the specificity of the neuronal responses during alert wakefulness. There is evidence that serotonin might be involved in processes of behavioral inhibition, particularly the ability to withhold punished responses. Such deficits would be manifest as impulsivity.

E. Combined Effects

The serotonergic, noradrenergic, and cholinergic transmitter systems show changes in their activity as a function of the sleep–wake cycle. These transmitter systems are implicated in setting levels of cortical excitability, which presumably forms the basis for variations in alertness. Nevertheless, it is still the case that very little is understood about how these systems interact to modulate or mediate alertness and varieties of attention in the waking state. The role of glutamatergic neurons of the mesencephalon, which project to the thalamus and have similar patterns of activation as those of the brain stem cholinergic projections across the sleep–wake cycle, also remains relatively uninvestigated.

VII. DRUGS THAT CHANGE ALERTNESS

A. Stimulants

Psychomotor stimulants, which include amphetamine and related compounds, caffeine, and nicotine, increase alertness, whereas sedatives decrease alertness. The term ''psychomotor stimulant'' is of interest because it draws attention to the cognitive effects of the drugs and emphasizes that they do not merely stimulate the motor system, although motor stimulation is a feature of many of these drugs.

The methylxanthine derivatives (caffeine, theophylline, and theobromine found in coffee, tea, and chocolate, respectively) are the most popular and commonly used stimulants. They increase the release of dopamine, serotonin, and norepinephrine. Fatigue is reduced and there is an increase in alertness and improvement in vigilance performance. However, caffeine consumption can impair complex motor tasks. At doses exceeding approximately 250 mg (the quantity in three fresh-brewed cups), tremor, nervousness, and anxiety can result. Contrary to popular belief, caffeine does not counteract the effects of alcohol: Caffeine (or another stimulant) will arouse a drunk but will not make him or her sober.

Amphetamines are stimulant drugs that cause euphoria, hyperactivity, increased confidence and subjective alertness, and reductions in fatigue and appetite (hence their use for weight control). Response latency is decreased in a variety of tasks, with increases in errors generally being due to anticipations (i.e., not waiting to process a target signal). Amphetamines also cause sympathetic arousal (increasing heart rate and blood pressure). The principal pharmacological effect of amphetamines is to increase the release and decrease the reuptake of the neurotransmitters dopamine and norepinephrine throughout the brain.

The stimulants adrafinil and modafinil are agonists at the noradrenergic postsynaptic (α_1) receptors. At doses that have alerting effects, these drugs are reported to have few side effects. Specifically, unlike other stimulant drugs that inhibit sleep, reports suggest that these drugs do not impair the ability to fall asleep and do not alter sleep patterns. Furthermore, they are less likely than amphetamines to result in rebound hypersomnolence, and the motor effects (tremors and hyperactivity) are much less pronounced. In addition, although modafinil increases dopamine turnover in the striatum, the motor stimulatory effects of modafinil are not antagonized by drugs that block dopamine receptors, suggesting that the mechanism of

the stimulant properties of modafinil is entirely independent of the motor stimulatory effects of amphetamines. Reboxetine, an antidepressant noradrenegic reuptake inhibitor, is also a stimulant.

Nicotine is an agonist at cholinergic nicotinic receptors, but by binding at presynaptic cholinergic receptors it has the indirect effect of causing the release of dopamine. Nicotine improves performance on a variety of tests of attention and cognitive function.

B. Sedatives

The narcotic analgesics (or narcotics) belong to the class of opioid drugs. Opiates (e.g., opium, morphine, and codeine) are derived from the opium poppy, whereas the synthetic derivatives (e.g., heroin) and purely synthetic drugs (e.g., Demerol and methadone) are called opioids. Heroin crosses the blood–brain barrier faster than morphine, giving a larger and faster "high," but it is converted to morphine once in the brain. These drugs are sedative analgesics, because they have pain-relieving properties as well as being central nervous system depressants. All of the drugs also have abuse potential because they result in euphoria. The sedative effects of opioids result from opiate receptors in the mesencephalic reticular formation, whereas the reinforcing effects are due to stimulation of the mesolimbic dopamine system. The psychoactive effects of opium have been known for centuries. After the opiate receptor was identified, the search began for the endogenous chemical that bound to this receptor. The endorphines (the name is derived from "endogenous morphine") were identified. Endorphines are released with stress and exercise; they appear to function as natural painkillers. It is the release of endorphines that is responsible for "runner's high," the feeling of euphoria and sedation that follows strenuous exercise.

Central nervous system depressants are similar to opiates but without the analgesic effect. They include barbiturates (e.g., pentobarbital, secobarbital, amobarbital, and phenobarbital), sedative hypnotics [e.g., scopolamine and methaqualone (quaaludes)], and anxiolytics (e.g., diazepam). Barbiturates are used as sleeping pills and there is cross tolerance and cross dependence with alcohol, which has similar depressant effects on cognitive function. Alcohol is a central nervous system depressant, which has effects on membrane excitability.

VIII. DISORDERS OF ALERTNESS

Fatigue is a deficit in alertness, which is a normal response following physical or mental exertion and instructs the body to rest and repair. Fatigue is a symptom of many illnesses, which reflects in part the importance of rest in recuperation from illness. However, there are occasions, such as after resting, when fatigue is inappropriate and, when alertness is required, debilitating. If such unexplained fatigue is persistent and is accompanied by other symptoms (such as sleeping problems, depression, concentration or memory problems, headache, sore throat, swelling of the lymph nodes, and muscle or joint pain) it is likely to be diagnosed as chronic fatigue syndrome.

Alertness is generally considered a good thing, but it is worth considering whether there are conditions in which a patient might complain of hyperalertness. Insomnia is the most obvious candidate; however, this is better thought of as a condition of inability to sleep rather than inability to be "not alert." Typically, the insomniac will complain that he or she is drowsy and fatigued and, in fact, suffers from a lack of alertness when awake because of an inability to sleep. The best example of hyperalertness is perhaps mania.

A. Mania

Mania describes a state of high, perhaps frenzied, psychic and physical energy in which the patient shows hyperactivity, impulsivity, and disproportionate emotional reactions (often mirth, but also irritability). Mania, which may be regarded as the converse of depression, is rarely seen alone without the patient also experiencing depressive periods, alternating with the periods of mania. The cycle of mania and depression in bipolar disorder is many months, with more than four cycles in a year described as "rapid cycling."

The neurochemical basis of bipolar disorder is not known, although serotonin and norepinephrine are implicated, as they are in unipolar depression. Clues to the nature of the neurochemical dysfunction, and therefore clues to the nature of normal neurochemical function, are often found in the effective treatment regimes. The most effective treatment for bipolar disorder is chronic administration of the drug lithium carbonate. Lithium has a rapid and beneficial effect on calming mania and once the mania is treated, it seems that the depression does not follow. The neurochemical effects of lithium are on the noradrenergic and

serotonergic systems. The biological action of lithium is complex, with effects seen on membrane excitability, neurotransmitter receptor density, and modulation of second messenger systems. The functional consequences of such actions are that reuptake of norepinephrine and serotonin is enhanced and the release of serotonin is enhanced. One of the current hypotheses regarding bipolar disorder is that there is a synergy between these neurotransmitters in the determination of mood, such that low norepinephrine "sets the scene" for affective disorder while the levels of serotonin determine the form the illness will take. It is unlikely that the explanation will be as simple as the relative overall levels of serotonin and norepinephrine, but understanding the complex action of lithium might provide important insight into the normal functional harmony of these brain transmitter systems.

B. Depression

The converse of mania is depression, and psychomotor retardation is one of the cardinal symptoms of depressive disorder. Psychomotor retardation is manifest as poverty of movements, generalized lethargy, and inertia. There appears to be a loss of motivation, with a consequent decline in productivity, which might be described as mental and physical fatigue.

Diminished ability to think or concentrate or indecisiveness nearly every day are depressive symptoms of considerable diagnostic and research interest. In the *DSM-IV* manual, they are included in the diagnostic criterion for major depressive episode.

Atypical depression is diagnosed when there is reactive mood ("mood swings") plus at least two of four additional symptoms: a lifetime pathological sensitivity to rejection, hyperphagia (or overeating), hypersomnia, and leaden paralysis. The latter two are of most relevance to the discussion of disorders of alertness. Hypersomnia describes the tendency to sleep excessively: Some patients may sleep as much as 20 hr a day. Leaden paralysis describes a state in which the patient suffers extreme psychomotor retardation. All physical activity ceases and the patient will spend many hours in apparent stupor.

Depression is treated with drugs, which increase the availability of serotonin and/or norepinephrine either by increasing their production or by blocking their reuptake, thereby making more of the neurochemical available in the synapse to stimulate receptors.

C. Disorders of Consciousness

At the extreme of disorders of alertness are the so-called disorders of consciousness, including coma, persistent vegetative states, and hyperkinetic or akinetic mutism. In the case of coma, the patient cannot be aroused and is unresponsive to internal and external stimuli. In contrast, the persistent vegetative state is differentiated from coma by the presence of cyclic arousal (resembling sleep–wake cycles) but without cognitive alertness or responsivity. Both coma and persistent vegetative state are associated with damage in the brain stem and both conditions are described as deficits of alertness.

In akinetic and hyperkinetic mutism, the deficit is less obviously one of alertness, although patients with either condition are unresponsive and apparently unconscious. Akinetic mutism resembles the persistent vegetative state in that the patient is unresponsive but shows spontaneous eye opening and may show smooth pursuit tracking of visual stimuli. This state has been described as "hypervigilant": The patient appears to be alert and attentive although he or she remains unresponsive. This disorder results from damage to frontal cortical and subcortical areas, such as that which occurs following rupture of the anterior communicating artery. The converse condition appears to be hyperkinetic mutism, in which the patient displays coordinated motor activity but without attention and apparent awareness. Hyperkinetic mutism is seen following damage to posterior attentional circuits of the parietal and occipital lobe, leaving frontal cortex intact. In both cases, the brain stem activating systems are intact: The forebrain executive systems are compromised. Thus, these conditions might represent the dissociation of consciousness (which is impaired) and alertness (which is present).

IX. IS ALERTNESS A USEFUL CONCEPT?

It is both possible and useful to make a distinction between alertness and processes such as arousal, attention, and vigilance. Traditionally, arousal was the term used to encompass all forms of activation, be it the cognitively and spatially specific arousal achieved by covert orienting of attention or the general and nonspecific sympathetic arousal resulting from the perception of imminent danger. Identifying the specific functions of the multiple transmitter

systems, arising in the brain stem to innervate the forebrain, requires a fractionation of the concept of arousal into its components. It is not possible to determine a single neurochemical system that is responsible for alertness: It is likely that alertness is the product of the synergistic activation of multiple transmitter systems. Nevertheless, alertness is a useful concept because it represents another dimension along which cognitive processing varies and further fractionates and refines the traditional concept of arousal.

See Also the Following Articles

AROUSAL • ATTENTION • DEPRESSION • DOPAMINE • ELECTROENCEPHALOGRAPHY (EEG) • INFORMATION PROCESSING • NOREPINEPHRINE • THALAMUS AND THALAMIC DAMAGE • VIGILANCE

Suggested Reading

Eysenck, M. W. (1982). *Attention and Arousal. Cognition and Performance*. Springer-Verlag, Berlin.

Moruzzi, G., and Magoun, H. W. (1949). Brainstem reticular formation and activation of the EEG. *Electroencephalogr. Clin. Neurophysiol*. **1,** 455–473.

Parasuraman, R., Warm, J. S., and See, J. E. (1998). Brain systems of vigilance. In *The Attentive Brain* (R. Parasuraman, Ed.), pp. 221–256. MIT Press, Cambridge, MA.

Passani, M. B., Bacciottini, L., Mannaioni, P. F., and Blandina, P. (2000). Central histaminergic system and cognition. *Neurosci. Biobehav. Rev.* **24,** 107–113.

Robbins, T. W. (1997). Arousal systems and attentional processes. *Biol. Psychol.* **45,** 57–71.

Sarter, M., and Bruno, J. P. (2000). Cortical cholinergic inputs mediating arousal, attentional processing and dreaming: Differential afferent regulation of the basal forebrain by telencephalic and brainstem afferents. *Neurosci.* **95,** 933–952.

Steriade, M. (1993) Central core modulation of spontaneous oscillations and sensory transmission in thalamocortical systems. *Curr. Opin. Neurobiol.* **3,** 619–625.

Alexia

RHONDA B. FRIEDMAN
Georgetown University Medical Center

GLOSSARY

affixes Grammatical word endings (e.g., *ly, ing, ed*).

derivational paralexia A paralexia in which the root morpheme, but not the part of speech, is retained.

functors Grammatical words, including pronouns, prepositions, articles, conjunctions, and auxiliary verbs (e.g., *him*, *the*, *with*, and *are*).

inflectional paralexia A paralexia in which the root morpheme and the part of speech are retained, but the form of the word is incorrect.

orthographic Pertaining to the letters of which a word is composed and the order in which the letters occur.

orthographic paralexia A paralexia in which the response shares at least 50% of its letters with the target word.

paralexia The incorrect production of a word in oral reading.

part of speech Syntactic classification of a word (e.g., noun, verb, adjective, adverb, and functor).

phonologic Pertaining to the pronunciation of the word.

pseudowords Pronounceable nonwords (e.g., *zilt* and *rog*).

semantic Pertaining to the word's meaning.

semantic paralexia A paralexia that consists of a real word that is related in meaning to the target word.

In this article, *alexia* is the term used to refer to acquired disorders of reading, subsequent to brain injury, in persons who had been literate prior to the injury. This is distinguished from dyslexia, which refers to developmental disorders of reading. In other texts, the term *acquired dyslexia* is used synonymously with alexia and is distinguished from *developmental dyslexia*.

I. INTRODUCTION

Alexias are common in patients who have sustained damage to the left hemisphere of the brain and are present in most patients with aphasia. In recent years, several different varieties of alexia have been identified. This is not surprising: Reading is a skill consisting of many cognitive components. Although there are many different models of reading, most agree that certain basic processes must occur: The written symbols are processed perceptually; letters are segregated and identified; orthographic units are matched to stored orthographic representations (knowledge of the letters that comprise a word and the order in which they occur); and the activation of orthographic representations, in turn, leads to activation of phonologic representations (the word's pronunciation) and semantic representations (the word's meaning). Disturbances in these various subcomponents of reading might very well be expected to disturb the process of reading in different ways. The identification of different types of alexia has led to new, more specific, approaches to treatment as well as to more detailed assessments of acquired reading disorders.

II. SYMPTOMS OF ALEXIA

The different types of alexia are distinguished from one another on the basis of two basic features: The

properties of the words that the patient has difficulty reading and the types of *paralexias*, or reading errors, that are produced.

A. Paralexias

The term *paralexia* is applied to an error produced in response to a task requiring the oral reading of a written word. Ordinarily, this term is used only for single-word responses and does not include multiword phrases.

Orthographic paralexias are reading errors in which the response shares at least 50% of its letters with the target word, in approximately the same order. Some examples are shown in Table I. These paralexias have sometimes been called "visual paralexias." This is a misnomer: The overall visual similarity between target word and response is often minimal with regard to overall shape and length (e.g., *political* → "police"). Another reason not to use the term visual paralexia is that in most cases, orthographic paralexias are not caused by visual problems. Orthographic paralexias are produced by patients with all forms of alexia, but the reasons these errors are produced differ between alexias. Therefore, these paralexias are not particularly useful in the classification of an alexia.

Semantic paralexias are reading errors in which the meaning of the response word is related to the meaning of the target word. Semantic paralexias are of many types. The response word may be a synonym, an antonym, a subordinate, a superordinate, a coordinate, or an associate of the target. Examples of semantic paralexias are provided in Table I. By definition, all alexic patients who produce semantic paralexias as more than 5% of their total reading errors are classified as having "deep alexia," which will be discussed later.

Table I
Examples of Paralexias

Type	Example
Orthographic	winter → "water"
	badge → "bandage"
Semantic	river → "stream"
	play → "game"
Derivational	instructive → "instructor"
	strength → "strengthen"
Inflectional	wished → "wishing"
	wants → "want"
Orthographic-then-semantic	sympathy → "orchestra"
	favor → "taste"
Functor substitutions	before → "into"
	his → "our"
Regularization errors	bread → "breed"
	shoe → "show"

Inflectional and derivational paralexias are reading errors in which the root morpheme is retained but the incorrect form of the word is produced. Inflectional paralexias refer to errors in which the correct part of speech is retained (e.g., *happiest* → "happier" and *buy* → "bought"). Derivational paralexias refer to errors in which the part of speech has been changed (e.g., *applaud* → "applause"). Inflectional and derivational errors are always seen in patients with deep alexia. They are often produced by patients with phonologic alexia as well. They are not characteristic of pure alexia, surface alexia, or attentional alexia. Prefix errors may occur in left neglect alexia, and suffix errors may occur in right neglect alexia.

Function word substitutions are errors in which a function word such as a preposition, conjunction, pronoun, or auxiliary verb is incorrectly read as another, seemingly unrelated, function word. Many of these paralexias are easily classified as function word substitutions; there is simply no other way in which the response word and the target word are connected (e.g., *her* → "which"). On other occasions, function word substitutions may be related orthographically (e.g., *her* → "here") or semantically (e.g., *her* → "she"). Although these latter paralexias may indeed represent instances of orthographic or semantic paralexias, they are typically all categorized as function word substitutions. Function word substitutions are produced by all patients with deep alexia, and by some patients with phonologic alexia.

Regularization errors are paralexias that are produced when an alexic patient reads a word as it *would* be pronounced if it were being read via some sort of spelling-to-sound correspondence rules (e.g., the word *come* is read as "comb"). These errors are not always easy to classify because the "rules" that were employed are not always transparent. Regularization errors may occur, sometimes with great frequency, in patients with surface alexia.

Orthographic-then-semantic paralexias are the result of two processing errors. The target word is first altered orthographically, then the altered word is misread semantically. These paralexias may at first appear to be random, unrelated responses until the mediating word is deduced (e.g., pivot → [pilot] → "airplane").

B. Word Properties Affecting Reading Accuracy

Written words differ along many dimensions. Some differences among written words, such as the word's initial letter or physical characteristics such as color, height, or type font, have little relevance to alexia. Other characteristics of words are extremely important in diagnosing and understanding alexia. These include part of speech, concreteness, length, regularity, and familiarity.

1. Part of Speech

For some alexic patients, the probability of reading a word correctly is dependent on the word's syntactic class. What is remarkable is that when a patient shows such a part-of-speech effect, the order of difficulty of the word classes is usually predictable. Nouns and adjectives are typically read best; verbs are read with greater difficulty; and functor words, including prepositions, conjunctions, pronouns, articles, and auxiliary verbs (e.g., *have* and *was*), are read most poorly. Part-of-speech effects are always seen in deep alexia and often in phonological alexia.

2. Concreteness

Another dimension along which words can be divided is the degree to which their referents are concrete or accessible to the senses. This is highly correlated with imageability, the ease with which a word's referent can be imaged. For some alexic patients, words that are highly concrete or imageable (e.g., *chair*) are more likely to be read correctly than words low in imageability or concreteness (e.g., *truth*). Patients with deep alexia always display a concreteness effect, and patients with phonological alexia may show this effect as well.

3. Length

The length of a written word will affect its likelihood of being read correctly for patients with certain types of alexia but not for others. For some patients, particularly those with pure alexia, words containing more letters will be more difficult. For other patients, especially those with phonologic/deep alexia, the number of syllables (not letters) might affect the difficulty of reading the word. For many patients, however, neither of these measures of length will be significantly correlated with reading success.

4. Regularity

This refers to the degree to which a word's pronunciation can be determined by its spelling—that is, whether it can be "sounded out" on the basis of spelling-to-sound correspondences. For some patients, particularly those with surface alexia, words that are highly regular such as *pin* and *tub* are more likely to be read correctly than irregular (also called "exception") words such as *pint* and *touch*.

5. Familiarity

Familiarity, whether or not a word is known to the reader as a real word, affects the reading of some alexic patients. An unimpaired adult reader can read both familiar real words and unfamiliar pronounceable nonwords (pseudowords), such as "rithy" or "Mr. Jamport." The reading of patients with phonologic/deep alexia is sensitive to this variable; real words such as *rot* may be read better than pseudowords such as *bot*.

III. TYPES OF CENTRAL ALEXIA

A. Pure Alexia

The syndrome of pure alexia was first described in the 19th century. The most striking feature of this reading disorder is that the patient retains the ability to write and spell; thus, pure alexia is also known as alexia without agraphia. Patients who have alexia without agraphia cannot read that which they have just written.

A characteristic feature of pure alexia is that patients with this form of alexia retain the ability to recognize words that are spelled aloud to them. That is, although patients with pure alexia have difficulty recognizing written words, they do not have difficulty identifying those same words upon hearing the names of the letters in the word in serial order. In fact, many of these patients discover that they can "read" written words if they name the letters of the words. The use of this compensatory reading strategy has been termed letter-by-letter reading. The ability to identify letters may be impaired early in the course of pure alexia. However, letter-naming ability often recovers over time, or it can be successfully retrained in most cases. The error most likely to be produced is the orthographic paralexia. These errors may be the result of incorrect letter naming or failure to hold on to all letter names in a word while the word is being identified (Table II).

Table II
Central Alexias and Their Characteristic Paralexias

Alexia type	Paralexias
Pure alexia	Orthographic
Surface alexia	Orthographic
	Regularization
Phonological alexia	Orthographic
	Inflectional and derivational
	Function word substitutions
Deep alexia	Orthographic
	Inflectional and derivational
	Function word substitutions
	Semantic
Phonological text alexia	*Errors occur only in text reading*
	Orthographic
	Inflectional and derivational
	Function word substitutions

Table III
Central Alexias and Word Properties Affecting Reading

Alexia	Word properties affecting reading
Pure alexia	Letter length
Surface alexia	Regularity
Phonological alexia	Part of speech
	Concreteness
	Syllable length
	Familiarity
Deep alexia	Part of speech
	Concreteness
	Syllable length
	Familiarity
Phonological text alexia	*Errors occur only in text reading, except PW reading errors*
	Part of speech
	Concreteness
	Syllable length
	Familiarity

Pure alexia is also characterized by a length effect, in which words with more letters are read more slowly and are less likely to be read correctly than words with fewer letters (Table III). The length effect is likely a consequence of the explicit letter-by-letter reading strategy: The more letters that must be identified and named, the longer it must take to do so. There are no effects of concreteness, part-of-speech, regularity, or familiarity either when these patients are attempting to read words or when words are spelled aloud to them.

Patients with pure alexia usually have intact language, although there may be some degree of anomia (difficulty retrieving words), which may be particularly pronounced for colors (color anomia). They frequently have a visual field cut called a right homonymous hemianopia, in which the right side of visual space cannot be seen.

Both the retained ability to spell and write and the retained ability to recognize orally spelled words suggest that orthographic information about words remains intact in patients with pure alexia. Thus, it has been suggested that the disorder reflects a disconnection of visual information from intact language processing areas of the brain.

The anatomy of pure alexia is consistent with the notion of a disconnection between visual and language processing centers of the brain. Pure alexia typically results from a stroke within the distribution of the left posterior cerebral artery or from a tumor located in the posterior left hemisphere of the brain. In most cases, the left occipital lobe is damaged such that the primary visual cortex is destroyed. This accounts for the right homonymous hemianopia. All visual input, then, must be processed initially by the right visual cortex. The results of this processing must then be transferred to the left angular gyrus for orthographic processing. However, this relay of the information is not possible because the lesion also damages the splenium of the corpus callosum, which connects the hemispheres in the posterior part of the brain, or the lesion damages the fibers adjacent to the angular gyrus within the left hemisphere. In either case, the visual information cannot reach the left angular gyrus, and reading fails (Table IV).

B. Surface Alexia

The cardinal feature of surface alexia is the presence of a measurable regularity effect in reading. When presented with a list of words with regular spelling-to-sound correspondences and a second list of words matched to the first list in letter length and frequency but consisting of words with irregular spelling-to-sound correspondences, patients with surface alexia have considerably more difficulty reading the second list. The regularity effect is likely to be even more pronounced when the words are of low frequency.

Some patients with surface alexia produce regularization errors; that is, an irregular word is pronounced (incorrectly) as it would be according to the

Table IV
Alexias and Site of Lesion

Alexia	Lesion site
Pure alexia	Left occipital lobe and splenium of the corpus callosum
Surface alexia	Parietal or temporoparietal, multifocal cortical degeneration, or closed head injury
Phonological alexia	Variable sites, but usually in the distribution of the left middle cerebal artery
Deep alexia	Left frontal extending into the parietal and temporal lobes
Phonological text alexia	Variable sites, but usually in the distribution of the left middle cerebral artery
Attentional alexia	Left posterior tumor or left parietal infarct
Neglect alexia	Typically right parietal lesion

spelling-to-sound correspondence rules of the language. However, not all surface alexic patients produce regularization errors in reading. Indeed, some of these patients produce a large number of errors that appear to be the result of a *misuse* of correspondence rules. Vowels, which typically have more than one pronunciation, are often mispronounced. One common error is the production of the short vowel rather than the long vowel in syllables consisting of vowel–consonant–e (e.g., reading *hate* as "hat"). Consonants with multiple pronunciations may be misread as well (e.g., *get* → "jet").

The patient with surface alexia appears to be unable to access the meanings of written words without first accessing their pronunciations. Comprehension of written words appears to depend on the pronunciation given to the word. Thus, if a written word is read incorrectly, the meaning attributed to that word will correspond to the pronunciation given. For example, if the word *come* is read according to spelling-to-sound correspondence rules, and is thus pronounced so as to rhyme with "home," then the surface alexic patient may interpret the word to mean an implement used for fixing one's hair.

As might be predicted, homophones are the source of a great deal of confusion for surface alexic patients. Intact readers make use of the orthography (spelling) of words such as *for* and *four* to determine that the former is a preposition and the latter a number. The surface alexic patient, upon pronouncing them both identically, may not know which for/four is which.

Indeed, the surface alexic patient frequently relies on the pronunciation of a written word not just to determine its meaning but also to determine whether or not it is a real word. If the patient reads the word *pint* so as to rhyme with *mint*, then the word will be judged to be a nonword, because there is no English word with that pronunciation. The patient is unable to rely on the familiarity of the sequence of letters in a previously known word to determine that it is a real word.

Similarly, pseudowords are not immediately recognized as being nonwords and must be pronounced before such a judgment can be made. Pseudowords that are homophonic with real words (e.g., *hoam*) may be accepted as real words by surface alexic patients.

The reading of patients with surface alexia usually does not show an effect of familiarity. [If anything, there is a reverse familiarity effect, in that pseudowords are read better than many (irregular) real words.] Likewise, effects of concreteness, part of speech, and length are not typically seen in surface alexia. In most reported cases of surface alexia, spelling is impaired in a manner analogous to the reading deficit. That is, words with irregular or ambiguous spellings are likely to be spelled incorrectly more often than words with predictable spellings (lexical or surface agraphia).

Most patients with surface alexia have lesions in the parietal or temporoparietal region of the left hemisphere. Surface alexia has also been described in cases of multifocal cortical degeneration and is frequently seen following closed head injury.

C. Phonological Alexia

The defining feature of phonological alexia is a strong familiarity effect (i.e., a marked deficit in reading pseudowords) in the face of a relatively intact ability to read real words. Although this may seem to have little relevance for reading in the real world (we are rarely called on to read pseudowords), in fact patients with phonological alexia do complain of difficulty reading, although their complaints are often nonspecific.

Some patients with phonological alexia also have difficulty reading functor words (prepositions, conjunctions, etc.) and may have a tendency to delete, add, or substitute affixes (e.g., read *faded* as "fade" or "fading").

An explanation of phonological alexia that accounts for both the primary difficulty reading pseudowords and the secondary difficulty reading functors is that it represents a disturbance in connections between written words (orthography) and their pronunciation (phonology), forcing reading to proceed via direct connection between orthography and meaning (semantics). Most functors and affixes serve primarily a syntactic role and have weak representations within the semantic network. Pseudowords, of course, have no semantic value and thus cannot be read via meaning.

The errors that these patients produce when attempting to read pseudowords often seem to be derived from the target word in some way. Commonly, the initial phoneme (sound) is correct. Often, a word that is orthographically similar to the target word is produced. Some very short pseudowords may be read entirely correctly; long pseudowords are rarely read correctly.

The reading of patients with phonological alexia does not exhibit a regularity effect. If a length effect is seen, it is dependent on the number of syllables or phonologic complexity, not the number of letters. A part-of-speech effect may be seen, particularly for functors, as noted previously. A concreteness effect is occasionally seen in phonological alexia. Patients with phonological alexia do not always display the analogous deficit in writing (phonological agraphia).

The lesion causing phonological alexia is quite variable but is normally located within the distribution of the left middle cerebral artery.

D. Deep Alexia

The reading of patients with deep alexia displays all the alexia symptoms of phonological alexia, with the addition of the defining feature of the disorder, the production of semantic paralexias. When semantic paralexias are produced at a rate that is greater than chance, the entire symptom complex is almost guaranteed to be seen: poor pseudoword reading, a part-of-speech effect in which verbs are read more poorly than nouns, and functors are read more poorly than verbs, a concreteness effect, and the production of derivational errors and functor word substitutions. The pseudoword reading deficit in deep alexia tends to be more severe than that seen in phonological alexia. Responses may be completely dissimilar to the target pseudoword, or no response at all is produced.

It has been suggested that deep alexia represents the (most impaired) end point of a continuum that includes the various manifestations of phonological alexia. As in phonological alexia, the pronunciation of written words cannot be accessed directly; reading proceeds semantically. However, in deep alexia there is an impairment within the semantic reading route that results in the semantic paralexias. This impairment may be within the semantic processing system or in the ability of that system to access the correct phonological code.

The typical lesion that produces deep alexia is a large one, affecting much of the left frontal lobe and extending into the parietal and temporal lobes as well. It has been posited that the symptom complex of deep alexia is actually a manifestation of right hemipshere reading in the presence of a left hemisphere that is greatly damaged. Support for this notion has come from studies of reading in split-brain patients. It has been demonstrated that the right hemisphere of some of these patients can read concrete words but not abstract words or functors and cannot determine the pronunciation of nonwords (ascertained with a rhyme task). This pattern is similar to the reading pattern of patients with deep alexia.

E. Phonologic Text Alexia

Some patients complain of difficulty reading following a stroke or head injury, but the examiner may be unable to find any class of words—or indeed any words at all—that the patient cannot read correctly. However, the accurate reading of words presented singly is not replicated when the words are presented within the context of text. When reading text, these patients tend to produce functor word substitutions, and they make errors on affixed words, as is frequently observed when phonological/deep alexic patients are asked to read single words. Also, like phonological alexic patients, these patients have difficulty reading pseudowords. It therefore appears that the reading problems of these patients are related to the problems seen in patients with phonological alexia. Hence, this reading disorder has been labeled phonologic text alexia.

Patients with phonologic text alexia typically have auditory comprehension deficits as well as reading

comprehension deficits. They also have impaired short-term phonologic memory, as demonstrated by decreased span for recall of digits, words, and pseudowords. The combination of all symptoms associated with phonologic text alexia leads to the proposition that the reading disorder seen in these patients is in some way the result of a deficit in phonologic processing and/or retention of phonologic information.

The paralexias produced by patients with phonologic text alexia in text reading (few errors are produced in single-word reading) are predominantly derivational paralexias and orthographic paralexias; semantic paralexias are not part of this syndrome, nor are regularization errors. A length effect may be observed, but this will be dependent on the number of syllables, not the number of letters, reflecting the phonologic processing deficit.

As with phonological alexia, phonological text alexia is normally the result of a lesion in the distribution of the left middle cerebral artery.

IV. TYPES OF VISUAL ALEXIA

Some acquired reading disorders are not the result of deficits within the language/reading processing systems per se, but rather reflect difficulties at a more peripheral stage of visual processing. Attentional alexia and neglect alexia are the most common of these visual alexias.

A. Attentional Alexia

Attentional alexia, like phonologic text alexia, is seen only when the patient is attempting to read text; the reading of words in isolation is basically intact. However, when viewing a multiword display (typically normal text), the patient appears to have difficulty maintaining the separation between words. Letters from adjacent words somehow infiltrate the word being read, resulting in errors that appear to be similar to those of orthographic paralexias. Attentional alexia does not appear to be a problem specific to reading. Rather, it is part of a more general problem with selective attention. The reported etiologies of attentional alexia include left posterior tumor and left parietal lobe infarct.

B. Neglect Alexia

The term neglect alexia refers to a pattern of reading in which one side of each word and/or one side of a page of text is not read or is misread. Often, although not always, this is seen within the context of a neglect syndrome that is not specific to reading. That is, the patient tends to ignore one side of space, even when not engaged in reading. Typically, neglect is produced by a lesion in the right parietal lobe, resulting in neglect of the left side of space. However, there have been reported cases of right-sided neglect alexia following left hemisphere lesions and even some cases of left-sided neglect alexia following left hemisphere lesions.

See Also the Following Articles

AGRAPHIA • ANOMIA • DYSLEXIA • LANGUAGE AND LEXICAL PROCESSING • LANGUAGE DISORDERS • READING DISORDERS, DEVELOPMENTAL

Suggested Reading

Coltheart, M. (Ed.) (1996). Phonological dyslexia. Special edition of *Cognitive Neuropsychol.* **13**(6).

Coltheart, M., Patterson, K., and Marshall, J. C. (Eds.) (1980). *Deep Dyslexia.* Routledge and Kegan Paul, London.

Friedman, R. B. (1996). Phonological text alexia: Poor pseudoword reading plus difficulty reading functors and affixes in text. *Cognitive Neuropsychol.* **13,** 869–885.

Friedman, R. B. (1996). Recovery from deep alexia to phonological alexia: Points on a continuum. *Brain Language* **52,** 114–128.

Friedman, R. B. (2002). *Clinical diagnosis and treatment of reading disorders.* In *Handbook of Adult Language Disorders: Integrating Cognitive Neuropsychology, Neurology, and Rehabilitation* (A. E. Hillis, Ed.) Psychology Press, Philadelphia.

Friedman, R. B., and Alexander, M. P. (1984). Pictures, images, and pure alexia: A case study. *Cognitive Neuropsychol.* **1**(1), 9–23.

Friedman, R. B., Ween, J. E., and Albert, M. L. (1993). *Alexia.* In *Clinical Neuropsychology, Third Ed.* (K. Heilman and E. Valenstein, Eds.), pp. 37–62. Oxford University Press, New York.

Marshall, J. C., and Newcombe, F. (1973). Patterns of paralexia: A psycholinguistic approach. *J. Psycholing. Res.* **2,** 175–199.

Patterson, K., and Kay, J. (1982). Letter-by-letter reading: Psychological descriptions of a neurological syndrome. *Quart. J. Exp. Psychol.* **34A,** 411–441.

Patterson, K., Marshall, J. C., and Coltheart, M. (Eds.) (1985). *Surface Dyslexia: Neuropsychological and Cognitive Studies of Phonological Reading.* Erlbaum, Hillsdale, NJ.

Riddoch, M. J. (Guest Ed.) (1990). Neglect and the peripheral dyslexias. *Cognitive Neuropsychol.* **7**(5/6) (Special issue).

Alzheimer's Disease, Neuropsychology of

ROBIN G. MORRIS and CLAIRE L. WORSLEY
Institute of Psychiatry, London, United Kingdom

GLOSSARY

divided attention Involves two or more simultaneous trains of cognitive activity in which the attentional focus has to be split between these activities either by simultaneously maintaining attention or rapidly switching between them.

episodic memory A type of memory involving recollection of specific experiences or episodes or where the context of material is also remembered. It can be contrasted with semantic memory, which is knowledge of the world, independent of the context in which it is acquired.

mental control Processes involved in the control and sequencing of cognitive activity. The term can be used in a similar fashion as executive functioning, which in turn encompasses the control processes relating to cognition.

perseveration The tendency to continue (inappropriately) a particular pattern of behavior or response. This can either occur continuously, such as repeating the same motor action, or after a delay, such as the unintentional repetition of a previously emitted response or idea.

Alzheimer's disease (AD) is the most common dementia, accounting for approximately 65% of cases and resulting in widespread neurodegenerative brain changes and accompanying progressive cognitive decline. The brains of people with AD are atrophic, with the most pronounced loss of brain tissue in the parietal and temporal lobes and with relative sparing of the occipetal, primary motor, and somatosensory cortices. The main histological changes include the presence of neuritic plaques, neurofibrillary tangles, granulovacuolar degeneration, gliosis, and neuronal cells loss. Accompanying these changes are depleted neurotransmitter functioning, affecting the main systems, including cholinergic, noradrenergic, and serotinergic function. A corresponding global reduction in cognitive function is the main feature of AD, but this has distinct patterns and can vary between patients. Cognitive neuropsychological investigations have helped establish the pattern of cognitive decline, mainly through studying groups of AD patients.

I. THE PRODROMAL PHASE

The insidious onset of Alzheimer's disease (AD) means that there is a significant prodromal phase of varying length, with research indicating a length of up to 20

years. During this period, there is an accumulation of neuropathological alterations accompanied by subtle cognitive changes, but not yet detected as AD. The nature of these changes has been investigated by following up a random sample of individuals without dementia in a prospective fashion and determining the characteristics of those who develop AD. Such studies indicate that poor performance on tests of episodic memory are highly predictive of AD. There are also individuals who develop isolated substantial memory impairment in the absence of a diagnosis of dementia who then develop AD. This prodromal phase may last many years. Studies suggest that subsequent AD in individuals with isolated memory impairment can be predicted on the basis of short batteries of neuropsychological tests of episodic memory and mental control. Allied to this is the finding that atrophy of the hippocampus is predictive of subsequent conversion to AD. So too are decreased cerebral perfusion rates in the posterior and anterior cingulate, the hippocampal–amygdala complex, and the anterior thalamus, which are structures comprising the network that is thought to support episodic memory function. The prodromal phase has been studied in individuals at greater risk for developing AD, specifically those who are at risk for autosomal-dominance familial AD. This research indicates that the onset of AD is preceded by measurable decline approximately 2 or 3 years before symptoms are observable and 4 or 5 years before the individual fits the criteria for AD. Again, memory dysfunction is the best predictor of subsequent AD.

II. COGNITIVE DECLINE IN AD

AD is perhaps the most well-known example of a progressive dementia. The trajectory of decline, however, can differ among people. Nevertheless, there is evidence that the decline is not linear, with the rate of decline early on being slower than in the middle phase. It is not clear whether the rate of decline slows again in the later stages, partly because there is a lack of measuring instruments that can measure change across the range of severity of cognitive impairment. Many factors have been investigated regarding whether they influence the rate of decline overall. High levels of education seem to be associated with a lower prevalence of AD. Thus, education has been interpreted as acting as a mitigating factor in relation to the neuropathological changes that occur in AD. Conversely, a lower level of education may act as a risk factor in terms of determining the effects of decreased cerebral functioning. Other factors, such as age of onset and whether or not the person is in an institution or receiving community care, do not appear to have any systematic relationship with severity of dementia.

III. ATTENTION

Attention has been investigated in AD in four main areas: phasic, sustained, selective, and divided attention. Phasic attention involves maintaining a state of readiness to respond to a stimulus for very short periods. This appears to be normal in AD, as shown by the normal facilitatory effect of a warning signal prior to having to respond to a stimulus. For example, when a tone precedes having to press one of two keys, according to the position of a "square" on a computer screen, the speed advantage provided by the tone is normal in AD. Sustained attention or "vigilance" is the ability to attend to a particular sensory input or set of sensory inputs for long time periods. Preservation of vigilance in AD has been shown on various tasks, including that requiring a person to listen to trains of short tones.

Selective attention is the ability to shift attention at the perceptual stage. Within the visual domain this has been studied in relation to spatial attention, or moving the focus of attention between specific locations. Here, the processes of engaging or focusing attention on a location, disengaging, and shifting to another location can be studied. In AD, there is evidence that a deficit exists in disengaging and switching attention. Performance on the Posner selective attention task supports this conclusion. AD has been investigated by employing a task in which a particular response key has to be pressed according to whether a vowel or consonant appears to the left or the right of a central fixation stimulus. An arrow is shown just before the stimulus, pointing to either the left or right side of the screen, predicting the stimulus (valid cue) or pointing in the wrong direction (invalid). Most of the time the cue points in the right direction. In AD, there is a normal advantage when the valid cue is shown, indicating that engaging attention is normal. For the invalid cue, there is a greater increase in response time for AD, indicating problems with disengaging or switching attention away from the invalid cue. Impairments in selective attention have also been shown in relation to stimulus attributes. This has been investigated by showing digits

constructed from arrays of smaller digits and then requiring attentional shifts between the larger digits (global) and the smaller ones (local). This requirement leads to much larger response times in AD. Auditory selective attention has been studied using the dichotic listening task, in which simultaneous presentation of two strings of digits is provided. The normal right ear advantage when recalling the two-digit strings is lost in AD. Additionally, when instructed to recall the left ear first, there is an inability to benefit from a left ear advantage. This is interpreted as indicating that attention cannot be switched at will in AD.

People with AD perform very poorly on tasks that involve divided attention. This has formally been studied by requiring patients to combine tracking a moving node on a visual display unit using a light pen and other distracter tasks, such as pressing a foot pad when a tone is heard or repeating a string of digits. This type of task results in very high levels of interference in AD patients when compared to controls, even if the difficulty level of the tasks is adjusted to ensure that overall the difficulty of each individual task matches the ability of the participant.

IV. EXECUTIVE FUNCTION

The term "executive function" is used to cover a broad category of neuropsychological functions concerned with the sequencing and coordination of mental activity and behavior. These aspects tend to show impairment in the early stages of AD, with considerable impairment in the middle phase. This includes impairment on a range of standard tests, such as the Wisconsin Card Sorting Test, the Stroop Test, Random Generation of Digits, and Verbal Fluency. Although some of the deficits on these types of tasks are due to impairments in other domains of functions (e.g., language in relation to verbal fluency), there does appear to be a core deficit in relation to such functions as mental flexibility, initiation, and response inhibition. It is possible to detect subgroups of AD in which executive dysfunction appears early in a more isolated form, although the impairment tends to be become more general as the dementia progresses. It should be noted, however, that progressive dysexecutive impairment as an early presenting feature is likely to occur in people with frontotemporal dementia rather than AD.

Behavioral signs of executive dysfunction at a more basic level include perseveration and utilization behavior. Various forms of perseveration are seen in AD. These can be elicited using simple tasks, such as performing movement to command or getting the patient to draw alternating squares. The unchecked repetition of movement, sometimes termed continuous perseveration, is thought to be due to a disturbance in motor output in which there is postfacilitation of motor impulses. Operating at a different level are "stuck in set" perseverations, which involve difficulty in switching from one activity to another. A third type, which has been termed recurrent perseveration, is the most common type reported in AD and involves unintentionally repeating a response after a delay, cued by a new stimulus. For example, when a patient is asked to define a series of words, such as in the Wechsler Adult Intelligence Scale Vocabulary test, he or she may produce definition material from earlier items in response to a new item in an inappropriate fashion. It has been found that 88% of patients with AD have produced at least one recurrent perseveration within a standard test battery, whereas in normal older adults this type of error is rarely seen. In utilization behavior, responses are cued by objects in the environment in an exaggerated or inappropriate fashion —for example, being impelled to reach out and grasp proximal objects. Although this has not been studied extensively in AD, it can be observed in patients in the middle or later stages.

V. SHORT-TERM AND WORKING MEMORY

Short-term memory encompasses memory for material or events up to a period of approximately 30 sec. A basic method for assessing this is the memory span procedure, which involves the repetition of sequences of items, for example, words or tapping out a sequence of moves on an array of blocks. This type of memory is only slightly impaired in early AD. A more substantial impairment is observed if a delay is introduced between presenting material and recall and when the delay is filled by a distracter task.

This pattern of short-term memory dysfunction has been best characterized by research that attempted to identify the integrity of the different cognitive subcomponents of short-term memory. In relation to the verbal domain, the functioning of the articulatory loop system is a major component. This acts as a phonological store for verbal material, with a verbal rehearsal mechanism. There is evidence that this system is relatively spared in early AD. This is based on the finding that the phonological similarity effect is normal in AD. The effect relates to the tendency of

phonologically similar letters (e.g., PTCVBG) to be recalled worse than dissimilar letters (e.g., KWYRZQ). The size of the effect reflects the strength of the phonological store, and this effect has been found to be normal in AD, despite an overall reduction in memory span. A related finding is the normal word-length effect, in which longer words are more difficult to remember than shorter ones, reflecting the greater amount of time that the longer words take to be recycled in the articulatory rehearsal mechanism.

A major contributor to short-term or working memory impairment in AD is the divided attention of a person. This is seen in experimental paradigms, such as combining tracking with remembering digits, and also on tasks such as the Brown–Peterson task, in which three verbal items (e.g., letters) have to be remembered over intervals of up to approximately 30 sec whilst the person is distracted by a subsidiary task. When combined with tasks such as counting backwards by three's, there can be a severe impairment, even in the early stages of AD. Also, even simple tasks, such as a simple tapping movement, are sufficient to cause impairment in AD.

VI. LONG-TERM MEMORY

A. Episodic Memory

To many clinicians, an impairment in episodic memory, the inability to recall information or events within spatiotemporal contexts in the period of minutes, hours, or days, is the hallmark of AD. However, it should not be considered as a single distinguishing feature. Episodic memory impairment accounts for the loss of spatial or temporal orientation that can occur early on in AD. Formal testing of this type of memory almost always reveals substantial impairment across a range of tasks, for example, recalling lists of words, sentences, and stories or recognition memory for words, faces, and pictures. From a behavioral standpoint, memory loss has a distinct pattern as the dementia progresses. The following are the stages in the breakdown of memory function in AD.

Early AD: Mild memory lapses occur but cause only a few problems for the person. They are often falsely attributed to other factors, such as the effects of normal aging, stress, or depression. Examples are forgetting errands, failing to forward messages, and becoming disorientated in unfamiliar surroundings. Memory of episodes in the near distant past is poor, including memory of conversations with other people. These types of memory problems do not necessarily indicate a progressive neuropsychological impairment or dementing illness.

Moderate AD: The memory impairment starts to have a very significant effect on daily living activities. Memory errors include becoming disorientated even in familiar surroundings, forgetting familiar people or friends, and confusing the time of day or day of week. The person becomes increasingly unable to keep track of daily events.

Severe AD: Memory errors become more severe and can present safety problems for the person, such as those associated with wandering or forgetting to turn off the gas stove. Close relatives may be forgotten. Marked positive signs become apparent, such as confabulation and paramnesia.

The substantive impairment in episodic memory has been related to the pattern of neurodegeneration in AD. Studies of pure amnesia have implicated the mesial temporal lobe structures in memory functioning, including the hippocampus, the parahippocampal gyrus, and the perirhinal and entorhinal cortices. There is evidence that these structures are more heavily damaged early on in the time course of AD, including neurodegeneration of the hippocampus in which severe changes in the CA1 field are observed. Additionally, neurofibrillary tangles (NFTs) are found in large quantities in the entorhinal cortex (affecting layers II and IV). This structure receives projections from the perirhinal and parahippocampal cortices, which in turn project to association cortex including regions of the frontal, parietal, and temporal lobes. A hierarchical vulnerability of individual cytoarchitectural fields has been established through neuropathological studies in which NFTs appear in highest densities first in the entorhinal cortex and then in the CA1 fields and the association and sensory cortices (Fig. 1). Structural magnetic resonance imaging (MRI) has also shown that loss of volume of the hippocampus is related to episodic memory loss rather than other aspects of cognition (e.g., language and constructional praxis). In addition to mesial temporal lobe involvement, there is evidence that damage to the diencephalon contributes to memory disorder. First, a feature of episodic memory impairment in AD is a lack of sensitivity to the context in which a memory is formed, similar to that found in amnesic patients with diencephalic damage, such as occurs in Wernicke–Korsakoff's syndrome. Second, structural MRI has revealed an association between episodic memory loss and shrinkage of the diencephalic (thalamic) region.

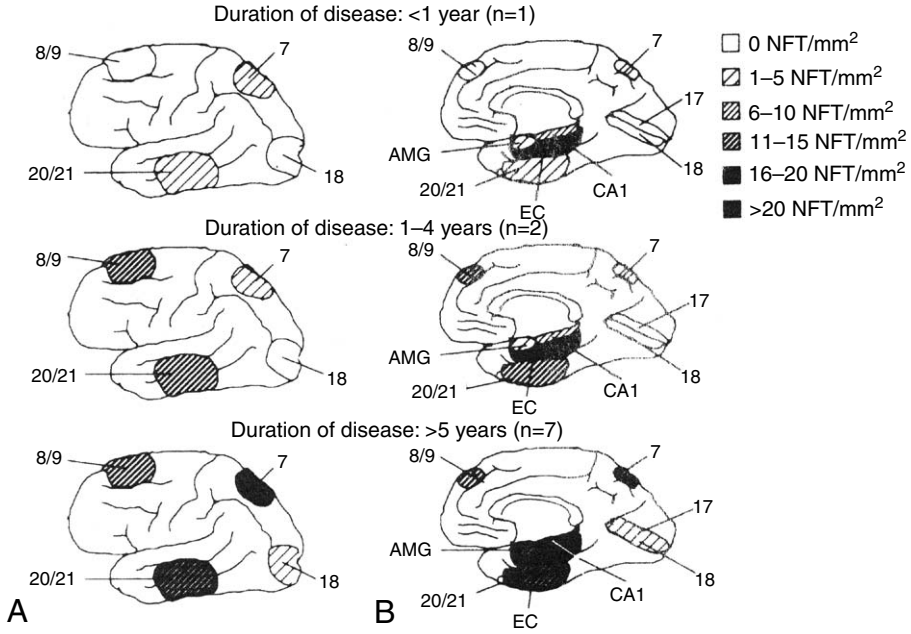


Figure 1 A schematic representation of the relationship between density and distribution of neurofibrillary tangles (NFTs) and duration of AD. (A) lateral view; (B) lateral medial view. The legend gives the density in number of NFTs per millimeter^2 represented by different shadings. EC, entorhinal cortex; AMG, amydala: Numbers refer to Brodman mapping (reproduced with permission from Arriagada *et al*., 1992).

The episodic memory impairment in AD can be seen as an amnestic syndrome overlaid by various processing deficits that further compromise memory functioning. This is supported by the inability of people with AD to make use of semantic processing to aid episodic memory. Normally, memory for material can be improved substantially if it can be better structured during the encoding process or if associative links are made with other material, as in the case of the "depth of processing" procedure in which a semantic orientating instruction is given, such as preceding a word with the question "Is it a type of bird?" In AD there is a failure to take advantage of these types of processes, with no improvement in memory. Nevertheless, when orientating instructions are given when the material is presented and the same orientating instructions are used to cue retrieval, there are memory improvements in early AD. Thus, this fact can be useful in rehabilitation programs relating to memory function.

B. Semantic Memory

The term semantic memory refers to the long-term store of facts, rules, concepts, and associations. The neuronal basis for this type of memory is thought to be diffuse within the neocortex but involving the inferolateral regions of the temporal lobes, with bilateral representation. The vulnerability of this region in AD underpins the loss of semantic memory functioning even at very early stages. Loss of semantic memory should also be distinguished from another disorder, semantic dementia, in which there is an early and marked pure semantic memory impairment. Evidence for the impairment in AD derives from a variety of tests. One is an impairment in verbal fluency, the ability to produce series of words that begin with a specific letter or that belong to a certain category (e.g., fruits). Fewer items are produced within a fixed time period. There is also a problem with object naming, in which the tendency is to produce superordinate or within-category errors [e.g., naming a trumpet a "musical instrument" (superordinate) or a "violin" (within-category)]. These impairments appear to relate to loss of material rather than to a problem accessing it, and they cannot be explained solely on the basis of language impairment. The main supporting evidence for this is that within a particular individual with AD, errors occur relating to the same specific material with different tests. For example, if there is a problem with naming an item, there is more likely to also be a

problem matching the name of the object to a picture of it. This can be true across a range of tests, including naming, matching words and pictures, sorting into categories, and providing definitions in response to being given the name of an item.

In other types of patients with brain damage, sometimes there is a dissociation between the types of semantic memories that are impaired. Specifically, those with focal damage to temporal structures can have difficulties naming living things, whereas those with frontoparietal damage can be impaired with regard to nonliving things. This distinction is mirrored in AD, for which some studies show particular problems with naming living things and some only with nonliving things, whereas others show no difference between the two. It has been suggested that such category-specific impairment depends in part on the stage of the dementia, with the early stages associated with more difficulty with nonliving items, and the pattern reversing as the dementia progresses. However, this is not proven and an alternative idea is that the type of category-specific deficit depends on the relative involvement of either temporal or frontoparietal regions.

The breakdown of semantic memory may explain some of the difficulties that people with AD have in responding appropriately in social or practical situations. There is evidence that impaired semantic memory extends to loss of schema or script knowledge. This refers to memories for typical sequences of events common to the shared cultural experience. For example, this might include "going to the theatre" or "a meal at a restaurant." Script knowledge has been examined in AD, and it has been shown that only the central and most frequent components of a script tend to be recalled, with difficulties in assigning particular activities to certain scripts.

The overall breakdown of semantic memory in AD may explain the clinically observed "emptiness" of thought and vague linguistic utterances, observed more acutely as the dementia progresses into the moderate form.

C. Remote Memory

When interviewing a person with AD, there is often the impression that distant events (e.g., early adult or childhood memories) are remembered better than recent ones. This type of "Remote Memory" has been explored in AD using various techniques and has provided insight into the circumstances in which the very remote memories are stronger. One technique is to present photographs of the faces of people who were famous in different past time periods. On the whole, such studies support the clinical impression, with famous people from the distant past more likely to be remembered, but there is an overall impairment. People with AD tend to do better if asked to say whether a face represents a famous person rather than to name the person or provide the identity; this finding has been taken to indicate that the primary problem is a loss of stored semantic information about the person who is represented rather than difficulties in accessing the name. A second technique consists of to showing pictorially famous scenes and asking the person to recall information about the event. Again, pictures of events in the distant past tend to produce fewer memories in AD.

Another approach is to investigate autobiographical memory, in which personal memories are cued using single words (e.g., "car") or people are asked specific questions about their past. Single-word cueing tends to produce more vague and generic responses. Specific questions, which may be about personal semantic information (e.g., schoolteachers) or episodic (e.g., recollections of the first day at school), result in a higher than normal number of memories about the distant past, similar to the pattern with famous faces. A further variant on this approach is to ask the person to talk about events that have been important in his or her life. In normal older adults, there is a tendency to produce most stories relating to early life, very few stories regarding the middle years, and then a modest increase in the amount of recent stories. In AD this pattern is repeated, but with a downward shift in performance. The result is that the past stories remain prominent, whereas middle-life stories become extremely sparse and there are few recent ones. In other words, people with AD may be left with memories of only salient personal events in early life, which may explain why one of the characteristics of temporal disorientation is that they think they are living in an earlier era.

VII. IMPLICIT MEMORY

A. Conceptual Priming

Conceptual priming is a form of repetition priming in which the subsequent reprocessing of the stimulus

takes place at the meaning or content level of the target stimulus. A main example is word-stem completion, in which the first few letters of a word are represented (e.g., the word "motel" with the stem "mot"). The task is to generate a word beginning with these letters; the fact that the particular word was presented previously increases the chances that it will be produced and not some other word. There is evidence for impairment of this type of task in AD, with the prior presentation of words failing to bias the subsequent production of the same words. This contrasts with "pure" amnesia or Huntington's disease, in which normal priming may occur.

However, there are circumstances in which there is normal conceptual priming, including the use of the homophone spelling bias test. Here, the manner in which a homophone is spelled by a subject can be biased in one direction by presenting semantically related items beforehand that are related to a particular spelling For example, if the word "son" (or "sun") is presented aurally, then it is more likely to be spelled "sun" if it is preceded by the words "moon" and "stars." The semantic context primes the production of the related spelling. This phenomenon has been shown to be as strong in AD patients as in normal subjects. There is also evidence that providing greater support for the original semantic processing of a word can reinstitute the priming effect in AD patients. Specifically, this has been achieved by requiring the subject to complete a sentence that ends in the word used for the priming (e.g., "He hit the nail with a … ." must be completed, where the priming word is "hammer"). A similar procedure that results in normal priming is to require the person to provide the meaning of each of the words that are used as primes. The presence of normal priming effects in certain instances has been interpreted as indicating that whether or not conceptual priming occurs normally in AD depends on the demands on semantic memory. If the demands are high, such as in the lexical priming task, then a partial degradation of semantic memory will cause an impairment.

B. Perceptual Priming

Perceptual priming occurs when processing the perceptual features of a stimulus (e.g., the visual or auditory form) increases the subsequent response time to the same material. Several studies have shown normal "repetition priming" in AD when previously presented material increases the rate at which a geometrically transformed script is read. Similarly, facilitation of the identification of briefly presented words following previous presentation of the same items is normal in AD. This type of learning mechanism may be located within the brain regions responsible for lower level perceptual processes rather than in the temporal–parietal regions, hence the degree of preservation in AD.

C. Perceptual Bias

A related type of learning is the perceptual biasing that can occur within sensorimotor processing in response to repeated activity. For example, lifting a heavy weight for a certain amount of time will bias the judgment of the weight of another item in the direction of thinking it is lighter. AD patients, despite having difficulties explicitly remembering the initially biasing experience, show the normal perceptual illusion. This also applies to the perceptual adaptation test, in which distorting prisms are used to shift the perceived locations of targets, which then have to be pointed to, with feedback given regarding the accuracy of response. In AD, the rate of learning is normal.

VIII. SKILL LEARNING

The ability to learn skilled motor behavior is not impaired in AD patients, at least in the early stages. This is shown by normal rates of learning on a variety of different tests. One is the pursuit rotor test, in which a handheld stylus has to be held so that it maintains contact with a rotating metal disc. Over a series of trials, accuracy increases and the rate of learning is taken as the measure of learning ability. In AD patients, the rate of learning is the same as that in controls, and this is true when the initial error rates on the task are carefully matched across groups by varying the speed of the disc to suit each person. Another task is mirror-reversed reading of text, which in normal circumstances improves with practice. Again, the rate of improvement appears to be normal in AD. In order to test more complex visuomotor skill learning, a serial reaction time test has been used. Here, a typical task consists of four lights being presented and each time a light comes on a key in front of it must be pressed as quickly as possibly. A random sequence is presented, but embedded in this sequence is an order

that is repeated several times. Visuomotor skill learning is shown by the increase in response time following repetition of the sequence, contrasting with a relative decline in response time with novel sequences. In AD, the normal pattern of learning is found, in the context of slower overall responses.

In all of the skill learning tasks described previously, the preserved learning ability suggests that AD patients rely on a different neuranatomical substrate than, for example, episodic memory. Comparison studies have been performed using other neurological conditions. These show a consistent learning deficit in Huntington's chorea, indicating that the neostriatal system is involved at the level of modification of central motor programs following sensorimotor feedback.

IX. LANGUAGE

A disturbance of language function is an intrinsic part of early AD and shows a steady progression in the time course of the dementia. At the early stages, the main feature in everyday language is anomia, accompanied by impoverished and circumlocutious language. As the dementia progresses, problems with comprehension appear to be more pronounced and the content of language becomes more vague, with syntactical and paraphasic errors. At the end stages, language is reduced to a few phases or even muteness. The following are the stages in the breakdown of language functioning:

- Early AD
  - Word-finding difficulties
  - Naming impairment
  - Circumlocutory discourse
  - Some problems with semantics
- Moderate AD
  - Impaired comprehension, particularly with complex material
  - Simplified syntax
  - Content vague and sometime meaningless
  - Paraphasias
  - Verbal perseveration
- Severe dementia
  - Meaningless repetition of words
  - Repetition of nonsense words
  - Mutism

Systematic studies of language function in AD have focused on the different components and suggest a general pattern of disorder. In relation to phonetic or phonemic processing, the processing and production of the sound characteristics, there is relative preservation. Until the advanced stages, phonetic articulation of word and sentence systems is largely intact. To a certain extent this is true for syntactic processing, in the sense that simpler syntactic constructions appear normal. Nevertheless, syntactic simplification in written language has been noted as an early indication of AD. Phonology and syntax contrast markedly with word retrieval impairment, which is manifest in everyday language as an impoverishment of vocabulary associated with difficulties in finding the right words. Nouns tend to be more affected than adjectives, with verbs the least affected. Formal analysis of this impairment has focused on confrontation naming for objects and word fluency. The naming impairment may reflect in part a perceptual difficulty since it is exacerbated if the stimulus is a photograph or line drawing rather than an actual object. Analysis of the errors in naming suggests that they are also semantic in nature, with a breakdown in semantic memory in part underlying the language impairment. Word fluency impairment likewise may reflect a combination of word retrieval, semantic memory, and executive function impairment. The nature of the semantic impairment was discussed previously. Allied to this is a deficit in comprehension of both written and aurally presented material.

The language impairment in AD fits none of the main categories of aphasia, but comparison studies indicate that the closest similarity is with forms of aphasia with more posterior lesions within the neuronal language system (e.g., transcortical sensory dysphasia). This is consistent with the neuropathology of AD, in which the main region of early cortical damage resides in the parietal and temporal lobes.

X. READING

Reading impairment in AD can be better understood in relation to cognitive neuropsychological models of reading processes. A major feature of these models is the identification of three main routes for reading. First, there is a sublexical route that connects the visual analysis of the word via grapheme to phoneme conversion and is used in learning to read. Second, a lexical route involves accessing the word lexicon, but with direct access to the language output system, bypassing analysis of the semantic system. This allows for the identification of words and pronunciation, but

without accessing meaning. Finally, there is input via lexical analysis to the semantic system and then language output.

In AD, the impairment of semantic processing contributes heavily to problems with reading and there is evidence that semantic memory deteriorates before lexical access. For example, in AD, there is more impairment in defining words than in reading comprehension, followed by oral reading of single words. Regular words can be read successfully, but problems occur when attempting to match words to pictures. In AD, there is preservation of reading irregular words, and this is exemplified in the National Adult Reading Test (NART), which is used to estimate the premorbid intelligence of a person with AD. Here, the irregular spellings mean that the grapheme-to-phoneme route cannot be used successfully, with reliance on either the semantic or the lexical route. NART performance is much less affected in AD than performance on other tests that would normally measure intellectual function; thus, it can be used as a premorbid predictor. Nevertheless, with moderate dementia, NART performance does decline and the errors made tend to reflect resorting to using the grapheme-to-phoneme conversion rules to arrive at an incorrect pronunciation. There is also evidence that the ability of AD patients to read NART words can be enhanced by placing them in the context of sentences, as in the Cambridge Contextual Reading Test. This is interpreted as the sentence context either facilitating the reader to identify the word as familiar, which then improves access to the lexical entry, or facilitating the use of a semantic route, resulting in a greater likelihood of correct reading. The reading of nonwords is impaired in AD, unless reading is done in circumstances in which it is possible to draw on orthographically similar words to construct pronunciation. A decline in nonword reading also follows the severity of dementia, indicating that the sublexical route is compromised in line with the dementing process.

Taken together, the pattern of reading ability in AD indicates that the lexical system shows less marked deterioration, with more reliance on the lexical rather than semantic or sublexical route.

XI. WRITING AND SPELLING

Writing becomes impaired early on in AD and is correlated with the severity of overall cognitive decline. This is in turn related to difficulties in accessing semantic information and is reflected in impairment in such tasks as written naming, narrative writing, and the ability to write the correct form of a homophone according to the given meaning. Reduced accuracy can also be attributed to impairments in nonlinguistic factors such as executive control and praxis rather than mainly to a disturbance within language-specific processing. Despite obviously impaired spelling performance, AD does not appear to produce any significant change in the effect of lexical, orthographic, and phonological stimulus variables in writing real words and pseudowords to dictation. For example, there is a normal advantage for spelling words with regular sound-to-spelling correspondence. However, there is a tendency for AD patients to show a slightly increased effect of spelling regularity and to have relatively more difficulty spelling pseudowords than normal controls.

XII. CALCULATION

Impairment in calculation is seen early on in AD. Although the pattern tends to be heterogeneous and inconsistent, there is greater impairment in relation to execution of calculation procedures rather than in answering arithmetic facts (rote-learned calculation outcomes). Within the latter, there is more impairment with multiplication rather than addition and subtraction facts. The deficit in execution has been related in part to the higher executive demand, particularly when the task emphasizes divided attention, such as writing and memorizing. Errors may also be due to problems in organizing and monitoring the sequence of cognitive operations.

XIII. VISUOSPATIAL FUNCTION

Visuospatial impairment in AD may affect either constructional or visuoperceptive ability, and these may reflect the involvement of the dorsal or ventral streams of neuronal processing, respectively. Constructional impairment is unlikely to be a presenting symptom, but it can be detected through formal assessment and may relate to parietal lobe dysfunction, particularly in the right hemisphere. Assessment procedures include the block design subtest from the Wechsler Adult Intelligence Scale, which usually reveals impairment in moderate dementia but can detect difficulties in the mild form. A simpler alternative is to require the patient to draw a series of line drawings, with drawing impairment frequently

observed in early AD. There are some indications that drawing to command (e.g., the patient is requested to "draw a clock") is more sensitive than copying drawings. Additionally, copying three-dimensional figures such as a cube tends to reveal impairment more readily than copying unidimensional figures. Many of these procedures, however, may confound constructional dysfunction with problems of praxis or planning. Visuoperceptual impairment may in part underlie impairment associated with naming objects. This is supported by the fact that naming impairment is increased when photographs of objects or line drawings are used rather than the actual objects.

XIV. EMOTIONAL PROCESSING

The ability to process emotion has been assessed in AD using a variety of tasks, including those relating to the auditory and visual domains. These tasks include recognizing the emotion of audiotaped voices, facial expressions of emotion, gestures, body movements, and videotaped vignettes. There is impairment in AD on these tasks, but this has been found to relate to the other components of the tasks used (e.g., visual processing or abstraction). The lack of primary deficit in processing emotion may explain why people with mild to moderate AD can demonstrate correct emotional responses in social settings, appropriately detecting the emotions of other people.

XV. MOTOR FUNCTIONING

The following are features of motor impairment in AD:

- Pyramidal signs
  - Hyperreflexia
  - Plantar responses
- Extrapyramidal signs
  - Tremor
  - Rigidity
  - Bradykinesia
  - Gait disturbances
- Primitive reflexes
- Apraxia
  - Ideational
  - Ideomotor

Not all people with AD exhibit the full spectrum of deficits. A general feature of AD is that the primary motor cortex is relatively preserved. However, pyramidal signs may be seen in AD; these include hyperreflexia, extensor plantar responses, hyperactive jaw jerk reflex, and ankle clonus. The mechanism for this is unclear but may involve either diffuse white matter or associated vascular dementia. Extrapyramidal signs exist, and the most common are bradykinesia (slowness of movement) and rigidity (increased resistance to passive movement). Tremor is infrequent. Gait abnormalities consist of slowness in walking, with decreased step length, paucity arm swing, and a droop posture, all of which may become more apparent as the dementia progresses. Myoclonus can occur early on, but motor seizures are regarded as happening later. Primitive reflexes of two types, nocioceptive and prehensile, are also seen in AD. The nocioceptive type, which includes the snout reflex, the glabellar blink reflex, and the palmomental reflex, occurs in approximately 20–50% of cases and can be seen early on. Of these, the snout and palmomental reflexes are the most common. The prehensile type, which includes grasping, sucking, and rooting, is less frequent (10–20% of cases) and is associated with late-stage dementia.

The two common types of apraxia are seen in AD, according to the strict definition of this term. Ideational apraxia, which involves impairment of the ability to perform actions appropriate to real objects, is frequently observed in AD, as is ideomotor apraxia. The latter refers to selection of elements that constitute movement, such as in copying gestures and miming usage. It has been estimated that the presence of both ideational and ideomotor apraxia occurs in 35% of patients with mild, 58% with moderate, and 98% with severe AD. Tool action knowledge has also been characterized as conceptual apraxia and found to be dissociable in AD from semantic language impairment.

In the moderate or severe range, another motor feature that can be observed is motor impersistence. This is an impairment in the ability to sustain a voluntary movement (e.g., exerting a steady hand grip or keeping the eyes closed). This is distinguished from apraxia because the movement can be performed and maintained with instruction. This type of motor impairment is strongly related to frontal lobe involvement.

XVI. HETEROGENEITY

The previous description of AD is mostly based on group studies with the aim of establishing general

features of neuropsychological dysfunction. Nevertheless, considerable heterogeneity exists and there is growing evidence for different subforms of AD, defined in terms of their initial presentation. Although a typical pattern is early memory disorder followed by the development of attentional, executive, and semantic memory and praxis and visuospatial impairment, other patterns have been observed. These include early aphasic disorder of fluent and nonfluent type and also early visuospatial or perceptual impairment. These tend to reflect the pattern of neuropathological changes, with the aphasic cases having temporal lobe involvement, those with visuospatial disorders having greater occipetoparietal alterations, and visual cortex involvement in the visual variant cases. It has been estimated that these atypical presentations comprise 14% of patients with AD.

XVII. CONCLUSION

This article shows that AD can be characteractized as an interlocking set of neuropsychological impairments that combine to provide a distinctive profile, considering AD as a single entity. At least in the early stages, areas of relative preservation of function can be identified, and these in turn can be used to optimize functioning. In addition, no person with AD is the same; individuals deviate substantially from the general profile, such that in clinical practice an individual analysis of the presenting deficits and their resultant disabilities should always be considered.

See Also the Following Articles

AGING BRAIN • ALEXIA • ANOMIA • ATTENTION • COGNITIVE AGING • DEMENTIA • INTELLIGENCE • MEMORY, EXPLICIT AND IMPLICIT • SEMANTIC MEMORY • SHORT-TERM MEMORY • WORKING MEMORY

Suggested Reading

Arriagada, P. V., Growdon, J. H., Hedley-Whyte, T., and Human, B. T. (1992). Neurofibrillary tangles but not senile plaques parallel duration and severity of Alzheimer's disease. *Neurology* **42,** 631–639.

Fleischman, D., and Gabrielli, J. (1999). Long-term memory in Alzheimer's disease. *Curr. Opin. Neurobiol.* **9,** 240–244.

Hodges, J. R., and Patterson, K. (1995). Is semantic memory consistently impaired early in the course of Alzheimer's disease? Neuroanatomical and diagnostic implications. *Neuropsychologia* **33,** 441–459.

Morris, R. G. (1996). *The Cognitive Neuropsychology of Alzheimer-Type Dementia*. Oxford Univ. Press, Oxford.

Morris, R. G. (1999). *The neuropsychology of Alzheimer's disease and related dementias*. In *Psychological Problems of Ageing: Assessment, Treatment and Care* (R. T. Woods Ed.). Wiley, Chichester, UK.

Parasuraman, R., and Haxby, J. V. (1993). Attention and brain function in Alzheimer's disease: A review. *Neuropsychology* **7,** 242–272.

Anger

CHRISTOPHER R. LONG and JAMES R. AVERILL
University of Massachusetts, Amherst

GLOSSARY

aggression In humans, the intentional infliction of harm on another, against their wishes and not for their own good; in other animals, the infliction of harm per se.

executive functioning A superordinate level of cognitive functioning that coordinates goal-directed activities and mediates conscious experience.

localization of function The attribution of a specific psychological process or function to a particular neural structure or brain area.

phrenology A pseudoscience of character analysis popular in the early 19th century based on the mistaken belief that patterns of indentations and protrusions in the skull indicate brain areas that mediate psychological functions.

Type A behavior Behavior characterized by feelings of time pressure, competitiveness, and hostility or chronic anger. Type A behavior has been linked to cardiovascular disease.

Anger is an emotion that involves both an attribution of blame for some perceived wrong and an impulse to correct the wrong or prevent its recurrence. The angry response may or may not involve aggressive behavior.

I. INTRODUCTION

Anger is a complex phenomenon deeply rooted in both our social and biological history. In the fourth century BC, Aristotle observed that

> *Anyone can get angry—that is easy; … but to do this to the right person, to the right extent, at the right time, with the right motive, and in the right way, that is not for everyone nor is it easy; wherefore goodness is both rare and laudable and noble.*
>
> *Nichomachean Ethics*, 1109a25

Elsewhere, writing from the perspective of a "physicist," Aristotle defined anger as a "boiling of the blood or warm substance surrounding the heart" (*De Anima*, 403a30). These two approaches to anger—the social and the physical—are not unrelated. Aristotle was not able to articulate that relation, however, except in abstract, logical terms. His knowledge of the physiological mechanisms was woefully inadequate. Today, advances in neurophysiology allow us to do considerably better.

At first, our knowledge of the relation of anger to the brain might seem a straightforward empirical issue, dependent only on technological advances. However, such an interpretation is only partially correct. As the above passage by Aristotle suggests, anger is inextricably linked to a network of concepts, an implicit folk theory that encompasses notions of action and passion, right and wrong, and retribution and conciliation. Neural mechanisms may be necessary conditions for anger, as they are for all behavior, but they are not sufficient. In this article, we focus

on those necessary conditions, but we do so in a way that also respects the irreducibly social aspects of anger.

II. TWO MEANINGS OF ANGER

The term "anger" is used in two distinct ways, both in everyday discourse and in scientific writings. First, in a generic sense, anger refers to a family of closely related emotions, including annoyance, rage, hostility, contempt, envy, fury, frustration, and jealousy. Second, in a specific sense, anger refers to one emotion among others in the same general class, such as when anger is contrasted with annoyance or rage. It is necessary to distinguish between the generic and specific senses of anger because what is true at one level of generality need not be true at another level.

In its generic sense, anger is often used to indicate almost any aggressive emotional response. Aggression is one characteristic that some episodes of anger share with other emotions in its class (annoyance, rage, hostility, contempt, etc.). However, as a specific emotion, anger is only occasionally accompanied by physical aggression, and even verbal aggression is relatively uncommon. For example, think of the last time you became angry, probably in the past day or two. What did you do? If you are like the majority of people, you likely talked the incident over with the instigator of your anger or with a neutral third party, or else you engaged in a variety of calming activities to "let off steam."

If anger is typically as benign as the above observations suggest, why should we be concerned with it and its underlying neural mechanisms, other than for academic reasons? The answer is threefold.

First, anger can erupt into aggression. When it does, the consequences can be serious both for the angry person and the target. In recent years, major advances have been made in our understanding of the evolutionary and physiological bases of aggression. As are new breakthroughs in the understanding of the neurophysiology of emotional behavior in general, advances in the study of aggression are reviewed in detail elsewhere in this encyclopedia and will only be mentioned briefly here.

Second, the attribution of anger is often used to excuse behavior after the fact. Anger shifts the blame from the aggressor ("I couldn't help it; I was overcome") onto the target ("Besides, he deserved it"). The existence of such a ready-made excuse may facilitate the occurrence of subsequent aggression against the same or other target, as in the case of child and spouse abuse.

Third, anger can cause problems even when no aggression is involved. A person who is unable to express anger in an effective manner may continue to suffer affronts (e.g., harassment in the workplace), leading to chronic stress and associated psychophysiological disorders (e.g., hypertension). In addition, a propensity toward anger and hostility seems to be the component of the Type A behavior pattern that is associated with increased risk for coronary heart disease.

In short, there are good reasons for studying anger and its relation to the brain. It is important to emphasize at the outset, however, that this article focuses on anger as a phenomenon in its own right, distinct from both aggression and emotion in general. This discussion is intended to complement, not replicate or supercede, information provided in the articles in this encyclopedia specifically focusing on aggression and emotion.

III. THE LOCALIZATION OF FUNCTION

To what degree can a complex psychological phenomenon such as anger be localized in the brain? The answer to this question depends in part on the level of organization under consideration. The more elementary the response (e.g., a reflex-like reaction), the more feasible its localization. At higher levels of organization (e.g., an instrumental goal-directed act), the nature of the response depends less on a specific neural structure than on the interaction among different parts of the brain and on the integration of information from a variety of sources, both internal and external.

Historically, there has been a tendency to replace the questions of what is happening and how it is happening with the seemingly more tractable question of where it is happening. In the early 19th century, for example, phrenologists claimed to have identified locations in the brain for a multitude of psychological functions, including everything from verbal ability to love of the picturesque, based on the perceived correlation of relevant behaviors with individual differences in patterns of cranial bulges and indentations. Although the empirical bases on which phrenology presumably rested were soon disproved (e.g., the shape of the skull bears little relation to the shape of the brain),

phrenology remained popular among laypersons well into the 20th century.

Modern researchers have grown much more sophisticated in their ability to assess relationships between the brain and behavior. For instance, by mapping event-related potentials, as well as using imaging techniques such as functional magnetic resonance imaging and positron emission tomography, researchers have recently demonstrated a surprising degree of modularity in brain function. For example, there appears to be specialized neural circuitry for the ability to recognize melody formed by variations in pitch as well as specialized neural circuitry for the ability to recognize the number of objects in an array. In a similar vein, research has demonstrated associations between angry cognitions and certain areas of the brain, often (but not always) the prefrontal cortex, the orbitofrontal cortex, and the bilateral temporal poles.

Before reviewing in detail possible associations between anger and specific areas of the brain, a caveat is worth emphasizing. Recall the example of the recognition of melody. Specialized circuitry for recognizing melodic changes in pitch does not suggest that Beethoven's *Ninth Symphony*, for instance, is somehow hardwired into the brain. Both the composition and the experience of a symphony are dependent on a particular social context as well as any biological predispositions, and the neurological correlates of a symphony are not only established through experience but also variably and widely distributed. Something similar can be said about anger. Specialized circuits likely exist that help mediate aspects of anger. However, as the quotation from Aristotle that opened this article implies, anger also belongs to the realm of moral discourse. A person cannot simply be angry; rather, he or she must be angry at someone for some reason (e.g., a perceived wrong). If anger is to be fully localized, its "place" is as much in the social system (i.e., in the normatively structured interaction between individuals) as it is in the nervous system.

Stated differently, the functions of the nervous system must be identified on their own terms, which include the processing, storage, and retrieval of information and the organization of behavior. The question then becomes: Are some kinds of processes unique to anger (as opposed, for example, to aggression and emotion in general)? No simple answer can be given to a question such as this. The human brain is an exceedingly complex organ that has evolved over millions of years. Past adaptations are seldom discarded but instead are maintained and incorporated into newer systems. Thus, some specialization of function relative to anger undoubtedly exists. There is, however, no reason to believe that a one-to-one relationship exists between a particular process and angry behavior, on the one hand, and another process and nonangry behavior, on the other hand. Rather, depending on the circumstances, the same processes may help mediate angry behavior on one occasion and nonangry behavior on another occasion. Although this complicates analysis, it also yields a major advantage; it means that insights gained in one area of study (e.g., the neural mechanisms involved in memory) can be extrapolated *mutatis mutandis* to other areas, including anger.

IV. THE ORGANIZATION OF ANGER EPISODES

Anger is a hierarchically organized pattern of behavior. In the following discussion, we distinguish five levels of organization, three of which can be observed in infrahuman animals, whereas the other two are specific to humans (Fig. 1). In discussing the infrahuman levels, we focus on aggression rather than anger per se for reasons that will become evident. However, as emphasized earlier, human anger need not—and typically does not—involve aggressive behavior.

Using radio-controlled electrodes implanted in the brains of *Macaca mulatta* monkeys, José Delgado observed the effects of brain stimulation on behavior as the animals roamed freely in their colonies. Stimulation of some brain sites elicited elementary (level I) responses, such as baring the teeth and vocalizations. Stimulation of other sites elicited responses that, although still fragmentary, were more complex and better organized: walking around, circling, climbing, and the like. Such level II responses were sensitive to the environment, but they remained divorced from the remainder of the animal's behavior.

At a still more complex level of organization (level III), one that implicates executive functioning, stimulation could elicit well-coordinated attacks against other members of the colony. However, the occurrence of an attack was not a straightforward response to brain stimulation; rather, it was mediated by past encounters between the stimulated animal and the other members of the colony as indicated, for example, by their relative positions in the troop's dominance hierarchy. When an animal was dominant, stimulation might elicit aggression against subordinate monkeys in the colony. However, if the subordinate monkeys were replaced by more dominant ones, stimulation of the

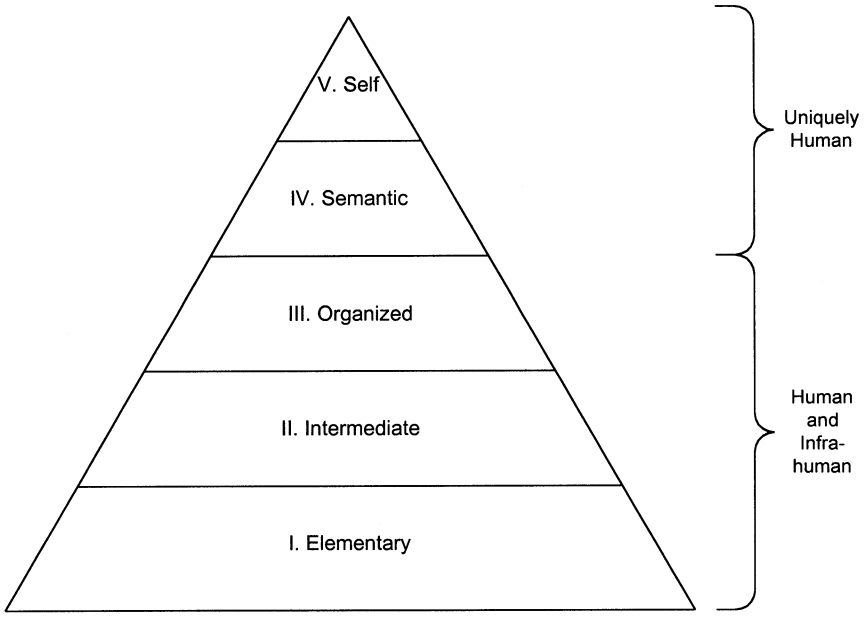


Figure 1 The hierarchical organization of anger episodes.

same animal (and at the same site in the brain) might cause the former aggressor to become the target of aggression by others.

In the particular monkey whose position in the troop's hierarchy was manipulated, the electrode stimulated her right pedunculus cerebellaris medius close to the lateral lemniscus. However, the important detail is not the precise location of the stimulation but that different responses could be elicited depending on the animal's circumstances at the time of stimulation. This exemplifies the point made earlier that a specific mechanism may help mediate angry behavior on one occasion and nonangry behavior on another occasion.

We do not wish to imply that the monkeys studied by Delgado experienced anger, except in a metaphorical sense. As discussed earlier, the concept of anger implies an attribution of blame; and an attribution of blame in turn presumes a network of concepts, for example, of intentionality, right and wrong, and justice and retribution. To the extent that the experience of anger is informed by the concept of anger, a monkey cannot be angry: It can be frustrated and aggressive but not angry.

To differentiate human anger from infrahuman aggression, two additional levels of organization must be considered; namely, the behavior must conform to anger as conceived of by the individual (level IV), and it must be related to the self (level V). To an even greater extent than level III, these distinctly human levels of organization involve executive functions.

V. EXECUTIVE FUNCTIONS AND THE EXPERIENCE OF ANGER

Executive functions coordinate goal-directed behavior and mediate conscious experience. Little direct evidence is available on the neurological mediators of these functions, at least with regard to anger. However, the work of Endel Tulving and colleagues on memory and self-awareness provides some insight. Some of the same capacities required for memory are also required for the full experience and expression of anger. For example, recognizing one's place in a dominance hierarchy, as in the case of Delgado's monkeys described earlier, presumes a memory of past encounters based on associative learning. Knowing the meaning of anger and related concepts, which is a kind of factual knowledge, involves semantic memory. However, objective knowledge of the meaning of anger is not sufficient. In anger, the appraised wrong is experienced personally (subjectively) as an affront to the self. The capacity for self-awareness, which Tulving relates to episodic memory, is thus necessary for the full experience of anger.

The three levels of executive functioning discussed by Tulving and colleagues correspond approximately to the top three levels of organization (III–V) depicted in Fig. 1. All three levels of executive functioning (and their corresponding levels of organization) appear to be localized in the prefrontal cortex. The first level, which can be found in infrahuman animals as well as in humans, interacts directly with posterior cortical and subcortical processes. It is charged with integrating diverse responses into a meaningful sequence, such as when Delgado's monkey's recognition of her subordinate status mitigated her inclination to attack. The second level of executive functioning appears to be localized in a somewhat more anterior portion of the frontal lobes. In addition to mediating semantic memory and factual knowledge, as discussed previously, this level also helps regulate behavior in situations that require novel solutions. Frustration, which is a common occasion for aggression, is one such situation. The third and "highest" level of executive functioning is closely allied with the second. Its distinguishing feature is its ability to relate events to one's own life, not simply in a factual, objective manner but as part of the temporal sequence of events, extending from the past and into the possible future, that comprise a person's sense of self. There would be no anger in the fullest sense—only momentary flaring to the immediate stimulus—without involvement of the self. Tulving and colleagues speculate that this capacity for self-awareness is mediated by processes localized in the most recently evolved anterior portions of the frontal lobes, an area of the brain whose relationship to emotional functioning is currently under investigation, with new findings appearing almost monthly.

Anger is often depicted as a biologically primitive response mediated by subcortical and paleocortical regions of the brain (e.g., the limbic system). It might therefore seem anomalous to suggest that anger is, in an important way, a function of the most recently evolved parts of the brain (anterior frontal lobes). However, this anomaly is not as great as it might at first appear. In the first complete work devoted exclusively to anger, the 4th-century theologian Lactantius used the wrath of God, not animal aggression, as the starting point of his analysis. Also, in traditional ethical teachings, the failure to become angry at perceived wrong was treated as a failure in humanity, not as a sign of superior virtue. In other words, the theoretical conception of anger as a highly evolved psychological function is not incompatible with many long-standing understandings of anger.

VI. RECOGNIZING ANOTHER'S ANGER

In recent years, considerable research has been devoted to specifying the brain mechanisms that help mediate a target's perception of another person's anger. This reflects, in part, methodological considerations: It is relatively easy to ask persons (e.g., brain-damaged patients) to evaluate pictures of angry facial expressions or audio recordings of angry speech; it is difficult to elicit anger in the context of a research study. However, the recognition of anger is an important issue in its own right. As discussed earlier, anger is an interpersonal emotion; it presumes a target as well as an angry person. An inability to recognize another's anger may thus be as disruptive of social relationships as is the inability to express anger appropriately.

The results of research on anger recognition have been variable, allowing few generalizations. The amygdala, the anterior cingulate cortex, and the frontal lobes are areas of the brain frequently implicated in studies of the visual recognition of angry expressions, although these areas are often implicated in the recognition of other emotional expressions as well. Damage to the amygdala has also been associated with difficulty in auditory recognition of anger cues, although several recent studies have called into question the importance of the amygdala in recognizing vocal expressions of emotion.

When interpreting results of research on anger recognition, several considerations must be kept in mind. First, the expressions of anger used in these studies are typically responses (e.g., facial displays) that occur at the organizational level III or below, as depicted in Fig. 1. That is, they are more indicative of aggressive intent than of anger per se. Second, the recognition of an emotional expression need not elicit the same emotion in the perceiver as in the sender; that is, when recognition of an angry expression does occur, the response of the perceiver may be fear or remorse, not anger in return. Third, even when the same emotion (anger) is elicited in the perceiver as in the sender, the brain mechanisms involved in recognition need not be the same as those involved in expression.

VII. IMPLICATIONS FOR REHABILITATION FROM BRAIN INJURY

Life is often unfair, or so it seems. Not surprisingly, people frequently become angry when confronted with life's inevitable misfortunes, including injury due to

accident or disease. For example, a brain-injured patient suffering paralysis or aphasia following a stroke may, like any other patient, experience anger at the seeming unfairness of events. When the anger is misdirected at health care providers, not to mention friends and family, treatment may be disrupted and recovery prolonged. Nevertheless, the anger is understandable, even if unjustified, as the patient comes to terms with a painful event that makes little rational sense.

Such "normal" if misdirected anger must be distinguished from that which results more or less directly from injury to parts of the brain related to aggression. For instance, an injury that results in stimulation of the amygdala may induce an aggressive response that is interpreted post hoc by the patient as anger. As humans, we like to think that our actions are meaningful. When, for extraneous reasons (such as brain injury), we respond in ways that make no sense, we nevertheless impose meaning on the response. Interpreting a brain injury-induced aggressive response as anger is one such meaning-making device. (Recall the earlier discussion of the generic and specific senses of anger. As noted, anger is often used generically to refer to almost any aggressive response.)

One distinguishing feature between normal (albeit misguided) anger following injury and brain injury-induced anger/aggression is that the former is typically manifested soon after the injury's occurrence, and it abates as the patient adjusts to life following the injury. However, differential diagnosis is difficult. Once an aggressive response is interpreted as anger, it is in a sense "normalized"; that is, it is made to conform to the beliefs and rules that help guide normal anger. The underlying condition that produced the aggression may thus be masked.

In short, the assessment and management of anger and aggression in neurological rehabilitation requires careful exploration of the patient's entire repertoire of behavior and the instigating factors. It is easy to be misled by focusing uncritically on a patient's claim that he or she was simply acting out of anger.

The most efficacious treatment of anger and aggression following brain injury requires an individualized rehabilitation program that incorporates an array of neurological, behavioral, and social therapies. In the same way that anger cannot be localized to any specific neural structure, the most effective anger management program will not rely exclusively on any one form of treatment.

See Also the Following Articles

AGGRESSION • COGNITIVE REHABILITATION • EMOTION • STROKE • VIOLENCE AND THE BRAIN

Suggested Reading

Averill, J. R. (1982). *Anger and Aggression: An Essay on Emotion*. Springer-Verlag, New York.

Berkowitz, L. (1993). *Aggression: Its Cause, Consequences, and Control*. McGraw-Hill, New York.

Davidson, R. J., Putnam, K. M., and Larson, C. L. (2000). Dysfunction in the neural circuitry of emotion regulation: A possible prelude to violence. *Science* **289,** 591–594.

Delgado, J. M. R., and Mir, D. (1969). Fragmental organization of emotional behavior in the monkey brain. *Ann. N.Y. Acad. Sci.* **159,** 731–751.

Panksepp, J. (1998). *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford Univ. Press, New York.

Patrick, P. D., and Hebda, D. W. (1994). Management of aggression. In *Neuropsychological Rehabilitation: Fundamentals, Innovations, and Directions* (J. León-Carrión, Ed.), pp. 431–451. GR/St. Lucie Press, Delray Beach, FL.

Wheeler, M. A., Stuss, D. T., and Tulving, E. (1997). Toward a theory of episodic memory: The frontal lobes and autonoetic consciousness. *Psychol. Bull.* **121,** 331–354.

Anomia

LISA TABOR CONNOR*,‡ and LORAINE K. OBLER†,‡

*Washington University School of Medicine, †City University of New York Graduate School and University Center, and ‡Boston University School of Medicine

GLOSSARY

aphasia A disorder of language produced by damage to the language zone, usually in the left hemisphere of the brain. Aphasia is not due to the inability to articulate language but is caused by damage to language representations or mechanisms.

circumlocution When a target word cannot be retrieved, a multiword response describing the characteristics of the object it names. For example, "you write with it" for "pen."

confrontation naming A task used to evaluate a person's ability to spontaneously name an object or an action; either a picture or an actual object is presented to the person for him or her to name, or a picture, pantomime, or video enactment of an action is presented to the person to elicit its name.

literal (or phonemic) paraphasia Sound substitutions or additions that result in either an unintended real-word utterance or a nonsense utterance. For example, with the word target *flea*, the utterances *free*, *fleep*, and *flur* would all be classified as literal paraphasias.

perseveration The unintended repetition of an idea, verbal utterance, or motor output.

verbal (semantic) paraphasia Substitution of a real word for a verbal target; if the substituted word is semantically related to a verbal target, such as *leaf* for the intended target *tree*, this would be considered a semantic paraphasia.

Anomia is a profound difficulty in coming up with words in the course of discourse and/or on a naming task, particularly those words that are heavily meaning-laden. This definition presumes both that a word being sought was originally known by the speaker and that the difficulty in retrieval is not due to failed articulation. Anomia is a hallmark of aphasic syndromes and occurs in aphasic persons with fluent, running speech and in aphasic persons with nonfluent, halting speech. That is, all aphasic individuals have word-finding difficulties to some degree. Such difficulties retrieving words are also characteristic of Alzheimer's dementia and, to some extent, of normal aging. This review focuses on the three populations in which anomia, sometimes called dysnomia, is prevalent: persons with aphasia, persons with Alzheimer's disease, and healthy older adults. A review follows of category-, modality-, and grammatical-class-specific anomias and of how these fractionations of word-finding into subtypes inform our understanding of the neuroanatomy of naming. Rehabilitation for anomia and recovery from it are discussed. Finally, a cognitive framework for naming is proposed to provide a means for discussing features of word-finding impairment.

I. ANOMIA IN APHASIA

As mentioned in the article on aphasia (this volume), it is a set of disorders of language. In adults, the aphasias are typically caused by stroke in the left hemisphere of the brain in the perisylvian region. The perisylvian region is bounded by the third convolution of the frontal lobe and the angular gyrus in the parietal lobe,

by the supramarginal gyrus, and by the superior border of the inferior temporal gyrus. Nearly all language deficits are produced by lesions inside the perisylvian region, and lesions outside of this region rarely produce language deficits, at least in right-handed individuals. Both areas traditionally associated with the classic aphasia syndromes, Broca's area in the frontal lobe and Wernicke's area in the temporal lobe, lie inside the perisylvian region.

Aphasic individuals with lesions inside this "language zone" exhibit different characteristics depending upon the location and extent of the lesion. Language deficits may be characterized by severe limitations of verbal and/or written production with relatively well-preserved comprehension (as in Broca's aphasia) or by severe limitations of auditory comprehension with relatively preserved verbal production (as in Wernicke's aphasia). Many other patterns of impairment and preservation of language faculties are possible, and particular combinations are characteristic of other aphasia syndromes. Regardless of whether the person with aphasia exhibits nonfluent speech or fluent speech, however, a common deficit is evident, that is, a difficulty accessing words, anomia. In fact, anomia was recognized in the late nineteenth and early twentieth centuries as a crucial component of aphasia by early aphasiologists such as Broca, Wernicke, Freud, Pitres, Head, and Goldstein. Anomia may be either a central feature of the aphasic picture or more peripheral to other features of language dysfunction. Because of its crucial importance to aphasic syndromes, however, most aphasia assessments begin with examination of the individual's ability to name in conversational speech and to name an object or drawing of an object to confrontation.

Naming is first assessed by eliciting a sample of speech. The speech sample may be obtained by asking the patient open-ended questions like, "What brought you here today?" or by showing a picture of a visual scene and asking the patient for a verbal description of it. Advantages of eliciting speech through a picture description task are that specific targets for naming are present, providing the patient with more structure, and that there may be objective standards for quantifying the patient's impairment in naming. One picture-description task commonly used in the assessment of aphasia is the Cookie Theft Picture Description from the *Boston Diagnostic Aphasia Examination*.

Naming to confrontation is the next step in assessing aphasia. The manner by which confrontation naming ability is tested is quite straightforward. Either an object or a picture or drawing of an object is presented and the patient is asked to say its name. Likewise, an action may be depicted, the examiner can pantomime an action, or a video clip may be shown portraying an action and the patient is asked to say what is happening. A commonly used test of confrontation naming of objects is the *Boston Naming Test*. To examine the nature of retrieval failures in confrontation naming, examiners probe the patient for more information. For example, they may ask "Do you know what this is? Is there another word for that? What is the woman doing?" If the patient is still unable to respond, cues may be given that are related to the meaning of the pictured object. A meaning-based cue, called a semantic cue, is often given if it is unclear that the patient perceives the picture accurately. For example, if shown a picture of a beaver, a cue may be either general in nature, such as "It is an animal," or specific in nature, such as "It builds a dam." If the patient is still unable to respond correctly or if it is clear from the patient's comments that the picture is perceived correctly, a sound-based cue, called a phonemic cue, may be given to aid the patient in producing the correct target. Typically, a phonemic cue consists of a small fragment of the target, such as the initial sound and initial vowel. For instance, if the target is beaver, the cue "bee" would be spoken. In some instances, latencies to produce a picture name may be measured to detect subtle difficulties in initiation or naming difficulties with particular categories of words. The significance of responding to cueing and category-specific deficits in naming will be discussed later in this article.

The characteristics of anomia differ among aphasia classifications. As illustrative examples, consider the features of anomia in Broca's aphasia, Wernicke's aphasia, and anomic aphasia. Anomia plays a dramatic role in Broca's aphasia. Patients with Broca's aphasia have difficulty initiating speech, have nonfluent, agrammatic speech with poor prosody and phrase length, and have particular difficulty producing the "small" function words of the language such as articles and prepositions, as compared to substantive words. Their speech is hesitant and labored, with only critical items related to the meaning of the message being produced, such as nouns and some verbs. Their total output is severely reduced. During confrontation naming, the naming performance of Broca's aphasics may be severely or mildly impaired, but some degree of difficulty will be encountered. Further, the Broca's aphasic may say the name of a previously produced picture even when a new one is presented (a phenomenon called perseveration) or may produce a word that

is semantically related to the target, such as *leaf* for the target *tree*. This latter type of substitution is called a verbal or semantic paraphasia.

Patients with Wernicke's aphasia, by contrast, have "fluent," not labored, fairly grammatical, speech with generally preserved prosody and phrase length. They often have difficulty with confrontation naming tasks, producing verbal paraphasias and word-sound substitutions called literal or phonemic paraphasias. It is often quite difficult to point to specific instances of anomia in the speech of a Wernicke's aphasic because there are no pauses to search for words and the intended meaning of the utterances in free conversation is often unclear. Wernicke's aphasics with severe impairment may produce neologisms, that is, nonsense words that have been hypothesized to reflect word-finding difficulties. A common feature of running speech in milder Wernicke's aphasics, like that of other fluent aphasics such as conduction, anomic, or transcortical sensory aphasics, is circumlocution, the verbal description of a target that the patient is unable to retrieve. In addition, Wernicke's aphasics fill their discourse with words "empty" of content (e.g., *thing*, *something*, *do*), again, probably reflecting difficulty in accessing the substantive nouns and verbs they intend. Severe Wernicke's aphasics are unaware that their speech does not make sense to the listener and may be frustrated that the message is not comprehended. Wernicke's aphasics also tend to perseveratively repeat ideas, particular phrases, or individual words or nonsense words.

Anomic aphasics have a severe word-finding impairment with little evidence of other language difficulty; that is, comprehension and repetition are both well-preserved. The anomia in anomic aphasia is characterized by circumlocution and a keen awareness of retrieval failure. The person with anomic aphasia may be quite frustrated at the inability to produce the desired target. The speech output of an anomic aphasic is still classified as "fluent" even though word-finding difficulties may be quite frequent, because phrase length is in the normal range with circumlocutions consisting of many words in a run and prosody is unimpaired.

II. CATEGORY-SPECIFIC, GRAMMATICAL-CLASS-SPECIFIC, AND MODALITY-SPECIFIC ANOMIAS

In addition to general difficulties retrieving lexical targets in aphasia, considerable attention has been devoted to concomitant category-specific dissociations in naming in aphasia. Several dissociations have been reported, including disproportionate deficits in naming natural objects compared to manmade objects as well as the converse, in retrieving proper names versus common nouns, in naming to visual confrontation as compared to auditory or tactile confrontation, and in naming actions versus objects and the converse. Evidence for semantic category-, grammatical-class-, and modality-specific anomias will be reviewed in turn.

A. Semantic-Category-Specific Anomia

Though reports of both category-specific sparing and deficits arose earlier than the seminal work of Warrington in the 1970s, Warrington and her colleagues brought attention to the increasing number of reports that words from different semantic categories may be unequally affected in aphasia. They began a series of systematic investigations into the selective preservation and impairment of semantic categories, seeking the best way to understand the islands of impairment that their patients exhibited. In early work they focused on the abstract versus concrete dimension and then later focused on the natural versus manmade object dimension. Many studies of semantic-category-specific deficits followed. Although at first it appeared that these more broadly defined categories captured the nature of the impairments in these patients, it became apparent through studies reported by other investigators that many exceptions arose that could not be readily incorporated into such broad categories. For instance, reports surfaced of patients with selective impairments in finding the names of fruits and vegetables but no other living things, impairments of only animal names, and selective impairments of naming facial emotional expressions. To further complicate the issue, there is a loose correspondence between lesion location and the nature of the semantic impairment, such that lesions both inside and outside the classical zone of language produce semantic category impairments, for example, an animal-naming deficit associated with the left inferotemporal region and a tool-naming deficit associated with the left parietal region. However, it is difficult to argue that these impairments arise with any consistency when a particular region is involved. Semantic-category impairments are intriguing both as behavioral phenomena and as challenges to models of language

representation in the brain; however, we are still quite a long way from understanding them.

B. Grammatical-Class-Specific Anomias

The early grammatical-class dissociation observed in the anomias was that between content-laden (''substantive'') words and functors, those ''small words'' like ''if,'' ''is,'' and ''who'' that convey primarily syntactic information. Patients with a type of aphasia called agrammatism, of course, produce predominantly substantives and few functors. However, agrammatism itself traditionally has not been considered primarily an anomic problem. Rather, problems with retrieving substantives (nouns, verbs, adjectives, and, perhaps, adverbs) in running speech or on confrontation naming constitute anomia.

Like the category-specific anomias, grammatical-class dissociations in naming have been reported that include noun–verb dissociations and proper-name-retrieval failures. Unlike semantic-category dissociations, however, noun–verb retrieval dissociations are more tightly coupled with particular lesion territories. That is, individuals with selective impairment of object naming have left posterior language-zone lesions of the temporal cortex, whereas individuals with selective impairment of action naming have left anterior language-zone lesions of the frontal cortex. The controversy in the domain of noun–verb naming dissociations lies in what these lesions represent in terms of the language system. If the anterior area is associated with verb problems and the posterior area is associated with noun problems, how can we account for the category-specific problems discussed earlier? One possibility is that they arise from smaller lesions than the one that results in general problems with nouns. A second possible explanation is that both the representations for semantics and phonology remain intact for nouns as well as for other substantives but that there is a disconnection between specific semantic centers and the store of the phonological shapes of names. In addition, this second theory assumes that the distribution of semantic and phonological representations for nouns is separate from that for verbs, nouns being distributed throughout the temporal lobe and verbs being distributed throughout the frontal lobe.

In addition to noun–verb dissociations, evidence for specific deficits in proper-name retrieval has been reported primarily by Semenza and colleagues. Lesions of the left temporal pole are associated with this deficit. In individuals with proper-name impairments, common nouns are largely preserved. Even names that would have been well-known before the aphasia are markedly more difficult to access. Indeed, some patients have been reported to have a dissociation between their ability to retrieve proper names for people and that for geographic locations and landmarks. The explanation offered for such dissociations has been that proper names, especially those for people, have limited associations to other items in the lexicon. For example, the occupation baker has associations to other occupations and to what bakers do and produce, whereas the last name Baker refers more arbitrarily to the individual or family one happens to know with that last name. For geographical locations and landmarks, the name may be associated more strongly with other information about that place, for example, one may associate the Eiffel Tower with everything one knows about Paris and France.

C. Modality-Specific Anomia

Whereas the term anomia has classically been associated with problems locating the names of things presented visually, there is a substantial set of cases that have been reported in the literature in which modality-specific factors are evident in accessing or producing substantives. Among the different modalities of presenting the targets are touch, taste, smell, and hearing (e.g., the sound of a bell ringing to elicit the noun ''bell''). Optic aphasia, for example, nicely demonstrates that naming impairments may be specific to a particular sensory modality. In optic aphasia, the person is unable to name an object presented visually but has minimal difficulty naming that same object presented through another modality. A picture of an apple may be met with naming impairment, but once the person has touched or smelled an apple the name is readily retrieved. As is the case with grammatical-class-specific anomia, a particular lesion along the pathway either from visual association areas and the semantic representations for objects or from visual association areas and motor output areas for speech is posited to explain the phenomenon. In addition, patients have been reported for whom it is not the modality of input that results in a dissociation but, rather, the modality of output. For example, an individual may be markedly more anomic when asked

to speak the name of a target than when asked to write it or vice versa.

Cases of modality-specific anomia have lead to theorizing about the extent to which there is one underlying lexicosemantic system for all modalities. Instances with dissociations among modalities suggest rather that there may be multiple systems of input in order to get to the phonological (or orthographic) shapes of words and for outputting them via speech or writing.

III. ANOMIA IN ALZHEIMER'S DISEASE

Alzheimer's disease (AD) is the most studied of the dementia-producing diseases that can have anomia as a salient symptom. Dementia (see the article on it in this volume) may be defined as a progressive cognitive decline resulting from a number of diseases. Such a cognitive decline characteristically may include more language disturbance (as in Alzheimer's disease in all but very-late-onset instances) or less language disturbance (as in the dementia associated with perhaps one-third of the individuals with Parkinson's disease). When a language disturbance is evident, anomia is invariably a part of it. The naming problems associated with the dementias may or may not be termed anomia by different scholars; however, the phenomenon is quite similar to that found in the aphasias. What underlies the problem, however, appears to be different. First, the cognitive problems underlying the dementia regularly include a variety of memory problems, so one may argue that the difficulty with outputting a name results from the memory problem rather than from a more strictly linguistic problem. At the anatomical level, moreover, the dementias are associated not with sizable, isolatable lesions such as those of the aphasias but rather with multiple lesions at the cellular level, thus suggesting that it is a systemic problem rather than an area contributing to some more or less specific aspect of naming that is impaired when anomia manifests itself.

Linked to the cognitive decline crucial for a diagnosis of dementia is a marked component of semantic impairment evident in the dementias, most particularly in AD. The term "semantic" here is meant to be distinct from the term "phonological" (or "orthographic") in that the borders of meaning of a word seem to become more permeable, or less specific information about the meaning of the word is available. This can be well-demonstrated by a task such as one to probe the semantic attributes of a word. For example, when a patient cannot remember the name penguin, one might ask about superordinate information (e.g., is it a tree? an animal?), coordinate information (e.g., is it a dolphin? a robin?), and subordinate information (e.g., does it eat fish? does it live in a cool climate?). Whereas people with anomia resulting from the aphasias can access all three types of information without difficulty, anomics with AD can, at one point in their cognitive decline, access only the superordinate information and later may even have problems with that. Such semantic problems, however, act in conjunction with more purely lexical-access problems. These may be more severe than those associated with normal age-related problems that the patient would be expected to evidence.

One of the ways one tests for the problem lying at the retrieval stage, as mentioned earlier, is by giving phonemic cues. For patients with AD, these may not be helpful as they are for normal elderly individuals with naming problems. Indeed, patients with AD may appear to "free associate" to the phonological cue: for a picture of a trellis, one says "tre…" and the patient may respond "trend." Of course such a response does not positively assure that the problem lies at the lexical level; rather, it may be due to inattention to the task. An alternate indication is to look at the consistency with which the patient can name an item over time. If the patient cannot name the trellis one day but can the next, this argues that the item's representation itself is not impaired, but rather retrieval of it is. An additional factor that enters into the naming errors of patients with AD is perceptual difficulty. That is, pictures may draw inappropriate answers ("cucumber" for "escalator") as the patient is drawn to the overall shape or to a subcomponent of the picture.

IV. ANOMIA IN NORMAL AGING

Whereas there are clear-cut naming problems associated with old age, the term anomia may be even less appropriate to describe them than the naming problems of Alzheimer's disease, perhaps because they are so much more subtle than the naming problems of aphasia. Indeed the term "dysnomia" (here referring to a transient naming problem rather than a permanent disorder) may more accurately portray the naming problems of aging in that it implies a more fleeting problem rather than an enduring impairment. Phenomenologically, older adults report word-finding

difficulties as they age (including more frequent tip-of-the-tongue experiences), and, in fact, laboratory-based studies have confirmed this report. In particular, older adults find the retrieval of proper names increasingly difficult. Whereas the literature confirms that proper nouns are particularly difficult with advanced age, difficulties with nouns and verbs are evidenced as well. There are some reports in the literature that noun retrieval is disproportionately impaired relative to verb retrieval with age, but more recent evidence suggests that, in carefully matched sets of nouns and verbs, the retrieval deficit associated with age, which becomes significant around age 70 years, is equivalent for both types of substantives.

Cross-sectional studies as a rule demonstrate significant naming problems for groups of older adults starting in the decade of the 70s. Longitudinal studies point to subtle declines in naming as early as the 40s. With normal aging, however, there is substantial variability across individuals within any given age cohort; some 80-year-olds perform like some 30-year-olds in a naming task. Education does seem to have long-term protective benefits in this regard. Results from the ability of older adults to rapidly choose a correct name on a multiple-choice task for items that they were unable to spontaneously retrieve, as well as being able to retrieve the correct name for an object once given a phonemic cue, suggest that the deficit in lexical retrieval for older adults lies in accessing the phonological shape of the target word. However, there is a suggestion that subtle semantic degradation may be involved as well from research examining the consistency of naming a given item over a series of test sessions.

Theoretical accounts for dysnomia in aging have centered around the well-described phenomenon of cognitive slowing. In particular, Burke and her colleagues have put forward an account of age-related naming deficits that combines cognitive slowing with the clear-cut deficit in accessing the phonological shape of a target word. Their model suggests that aging produces slowing, or "transmission deficits," across the board in the semantic and phonological networks that underlie lexical retrieval. The phonological system is relatively more affected, however, because, as activation spreads throughout the network, semantic activation for a particular target converges on the appropriate lexical node, but the activation must then diverge from the lexical node to the many phonological nodes that constitute the word. Because general slowing is at play in aging, the amount of activation that has accrued at the lexical node is less for older adults than for younger adults and, thus, when the activation spreads to the phonological nodes, it expires before crossing the threshold for word production.

V. TREATMENTS AVAILABLE FOR ANOMIA

Naming abilities in aphasics recover over time. Indeed, it appears that there may be slow progressive recovery in the ability to name objects and actions over time even years after the aphasia-producing incident and long after treatment has ended. However, some aphasics who have, according to all aphasia tests, fully recovered from their aphasia, nevertheless report feeling that they cannot always find the words they need in conversation.

Despite the apparent spontaneous recovery of naming abilities over time, treating them is quite difficult. Attempts at drilling items, which work well for second-language acquisition in some normal individuals, have no long-term effects for aphasic patients. More successful, by contrast, are efforts to affect the processes that interfere with naming. For example, perseveration of a previous verbal response may prevent access to an otherwise intact name. Therefore, treatments that focus on reducing perseveration may be helpful in "deblocking" verbal naming. One such treatment developed by Helm-Estabrooks and colleagues involved bringing perseveration to the awareness of aphasic individuals and teaching them to actively suppress the item that they were about to incorrectly produce and to then ask the therapist for a cue as to the correct name for the target. This treatment is not, however, a treatment for naming per se but rather an means to prevent an incorrect response from intruding upon the production of the correct name.

Another deblocking technique for severely impaired aphasics is the Voluntary Control of Involuntary Utterances program also developed by Helm-Estabrooks and colleagues. This method is appropriate for use in individuals who have quite restricted verbal ability, perhaps limited to a few stereotypical responses, but whose productions are inappropriate to the context. This type of off-target responding occurs only in cases of severe verbal production deficit. An example of this would be the case where the patient says "cat" when shown a picture of a pencil. The fact that the patient clearly has access to the name "cat" indicates that it is a candidate for remapping to its appropriate semantic representation. The idea underlying this program is that speech that is generated automatically

and without voluntary control can be retrained to be uttered voluntarily in an appropriate context. The clinician starts with a very small vocabulary of items, perhaps as few as five or six, for which the patient can successfully read the name of an object presented with its picture (e.g., the patient says "cat" to a picture of a cat) and expands the functional vocabulary over time by first introducing emotionally salient words and expanding outward from there to include new items. This technique tends to deblock other vocabulary, enabling the clinician eventually to expand functional naming to as many as a few hundred items.

Techniques have also been put forward to enable a less severely impaired patient to use partial utterances to cue a correct response. Often it is the case that, once a phonemic cue is given, an aphasic patient will be able to use that cue to generate the remainder of the word. Because this technique is only appropriate for a particular target, the goal is to get the patient to apply a systematic approach to retrieving the appropriate cue that will "bootstrap" the target response. An alphabetic search strategy is one means by which to generate the appropriate cue (and many aphasics have spared ability to recite fixed sequences like the alphabet), or mentally generating semantic associates to the target may lead to the recovery of the initial phonemes.

VI. COGNITIVE FRAMEWORK FOR NAMING

In order to arrive at a cognitive framework for naming that accounts for a majority of the available data from aphasia, aging, and Alzheimer's disease, a basic three-stage model of naming like that of Levelt consisting of a visual object-recognition stage, an access-to-semantic-information stage, and finally a phonological-realization stage needs to be expanded considerably. First, a lexical level needs to be inserted between the semantic and phonological levels to satisfy constraints of mapping distributed semantic and phonological representations onto one another. In addition, due to modality-specific deficits such as those found in optic aphasia, it is clear that a single semantic system blind to input modality will not suffice. Therefore, it is necessary to propose two levels of semantic analysis: one dealing with modality-specific features of an object and a second dealing with supramodal features such as an object's function. Moreover, on the basis of data from normal aging, it is necessary to include in a model of confrontation naming a means to describe the process by which the information is passed from one level of representation to another and how that information could be disrupted or affected by changes in the processing abilities of the individual. Finally, information about semantic-category dissociations and the degradation of semantic knowledge with Alzheimer's disease requires that there be a hierarchical representation of semantic attributes. Therefore, what initially appeared to be a simple and straightforward three-stage model of naming is complicated by what we know about how the components and processes of naming are affected by aging, disease, and neurological accident. Even with all of the additions to and modifications of the simpler model, many behavioral phenomena associated with anomia, as well as their neurological underpinnings, are as yet unexplained.

See Also the Following Articles

AGING BRAIN • ALZHEIMER'S DISEASE, NEUROPSYCHOLOGY OF • APHASIA • COGNITIVE AGING • LANGUAGE DISORDERS • SEMANTIC MEMORY

Suggested Reading

Burke, D. M., MacKay, D. G., Worthley, J. S., and Wade, E. (1991). On the tip of the tongue: What causes word finding failures in young and older adults? *J. Memory Lang.* **30,** 542–579.

Chan, A. S., Butters, N., and Salmon, D. P. (1997). The deterioration of semantic networks in patients with Alzheimer's disease: A cross-sectional study. *Neuropsychologia* **35,** 241–248.

Goodglass, H. (1993). *Understanding Aphasia.* Academic Press, San Diego, CA.

Goodglass, H., and Wingfield, A. (Eds.) (1997). *Anomia: Neuroanatomical and Cognitive Correlates.* Academic Press, San Diego, CA.

Helm-Estabrooks, N., and Albert, M. L. (1991). *Manual of Aphasia Therapy.* Pro-Ed, Austin, TX.

Levelt, W. J. M. (1993). *Speaking: From Intention to Articulation.* MIT Press, Cambridge, MA.

Anterior Cingulate Cortex

B. J. CASEY,* NICK YEUNG,† and JOHN FOSSELLA‡,*

Weill Medical College of Cornell University, Princeton University,† and Rockefeller University‡*

GLOSSARY

akinetic mutism The inability to initiate movement following surgical removal of or infarcts to the anterior cingulate cortex.

allele Refers to a sequence variant of a particular gene. Each person carries a genome sequence that is approximately 0.1% different from any other human (excluding identical twins and family members). These differences in the sequence of a gene give rise to multiple alleles of a gene in a population and can lead to variations in observable traits. The sequencing of the human genome shows the most common sequence differences (available at http://www.cshl.snp.org).

automatic processing A term used to describe cognitive processes that can be done rapidly, in parallel, and that require no attentional resources.

cingulate epilepsy syndrome Seizures confirmed in the anterior cingulate cortex that result in excessive activity of this region, impair consciousness, and alter affective state and expression. Patients with such seizures may display psychopathic or sociopathic behaviors.

cognitive conflict Interference due to competing stimuli, memories, decisions, tasks, thoughts, or actions.

cognitive control The ability to resolve conflict among competing or interfering thoughts and actions.

error-related negativity (ERN) A negative deflection in the event-related potential following incorrect responses during choice reaction time tasks observed during scalp electrophysiological recordings.

gene A portion of the human genome that contains the instructions and code for the production of a protein. The publication of the sequence of the human genome reveals approximately 35,000 genes (available at http://genome.ucsc.edu/).

gene association study The statistical methods used to determine whether variation in the sequence of a particular gene correlates with variation in an observable trait.

heritability (h^2) A term that describes the extent to which variation in a trait (e.g., cognitive performance) among members of a population is determined by inherited genetic variation. h^2 can vary between 1 (high) and 0 (low). h^2 can be estimated in a number of ways but most conveniently by comparing the correlation in performance of identical (monozygotic or MZ) vs fraternal (dizygotic or DZ) twins. Higher correlation among MZ twins suggests that genetic factors contribute significantly to the variation among individuals for that trait.

The anterior cingulate cortex is a region of the brain thought in to be involved in actions guided by sensation, cognition, and emotion. This article provides evidence for anterior cingulate cortex function from both human and animal studies using neuroimaging, electrophysiology, lesion, and genetic methodologies.

I. NEUROANATOMY

The anterior cingulate cortex is situated around the rostral portion of the corpus callosum. This region has numerous projections into the motor cortex and, thus, advantageously sits where it may have a significant contribution in the control of actions. Basically, the anterior cingulate has been implicated in sensory, cognitive, and emotionally guided actions. A description of the morphology, cytoarchitecture, and

connectivity of the anterior cingulate cortex is provided in order to help specify biological constraints on any theory of anterior cingulate function.

A. Morphology

The anterior cingulate cortex lies bilaterally on the medial surface of the frontal lobes around the rostrum of the corpus callosum, bounded by the callosal sulcus and the cingulate sulcus (see Fig. 1). The cingulate sulcus is consistently observed and typically remains unsegmented from its rostral extent to its termination at the marginal ramus in the parietal lobe across individuals. However, the precise morphology of the anterior cingulate cortex is extremely variable. In particular, a second paracingulate sulcus is observed to run parallel to the cingulate sulcus in many subjects, more frequently so in the left hemisphere than the right.

On the basis of cytoarchitecture and differential patterns of connectivity, the anterior cingulate cortex has been divided into ventral (Brodmann's areas 25 and 33), rostral (BA 24 and 32), and caudal (BA 24′ and 32′) regions. Area 24 has been further decomposed into a, b, and c subdivisions. Despite the variable morphology of cingulate cortex, there are some consistencies in the distribution of areas on the cortical surface. Thus, area 33 lies in the rostral bank of the callosal sulcus, with areas 25, 24a,b, and 24′a,b typically lying on the gyral surface and area 32 lying in the rostral part of the cingulate sulcus. Caudally, areas 24c and 24′c usually form the ventral bank of the cingulate sulcus, facing area 32′ on the dorsal bank. The paracingulate gyrus, when present, comprises areas 32 rostrally and 32′ caudally.

B. Cytoarchitecture

Brent Vogt and colleagues at Wake Forest University have characterized the cytoarchitectural divisions within the anterior cingulate cortex. In common with cortical motor areas, layer IV, the major input layer from sensory cortices, is absent from the anterior cingulate cortex proper (areas 33, 25, and 24). Because layer IV is also called the internal granular layer the anterior cingulate cortex is described as agranular.

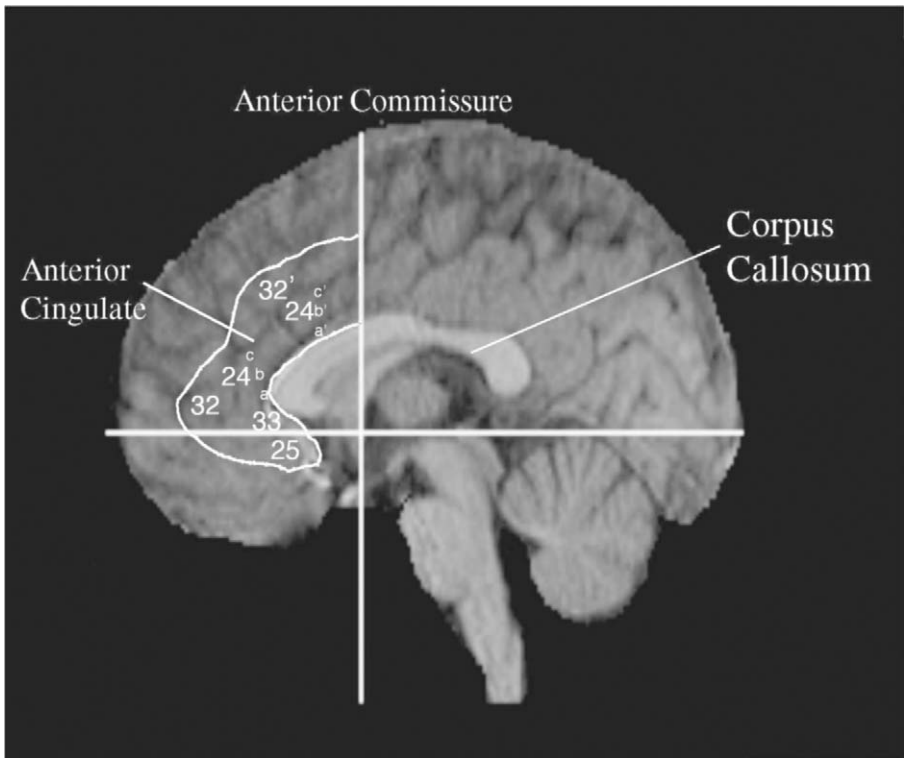


Figure 1 Anatomy of the anterior cingulate cortex.

In contrast, cingulocortical transition area 32 is characterized by a dysgranular layer IV that is attenuated in area 32′. Area 32 can also be distinguished from area 24 by the presence of large pyramidal neurons that form a layer IIIc. In area 32′, these pyramidal neurons are even larger.

Anterior cingulate cortex in general is characterized by a prominent layer V. Ventral areas 25 and 33 are poorly differentiated, although a laminar structure is apparent in a thin layer V and undifferentiated in layers II and III. Area 24 is characterized by a neuron-dense layer Va with prominent pyramidal cells and has a clear division between layers II and III. Area 24′ can be distinguished from rostral area 24 by its lower neuron density and thinner layer V. The a, b, and c subdivisions of areas 24 and 24′ are marked by generally increasing differentiation of cortex away from the corpus callosum. A further division in the most caudal sulcal region of the anterior cingulate cortex, area 24c′g, has been identified as containing gigantopyramidal neurons in layer Vb.

C. Connectivity

Investigations in nonhuman primates have shown the anterior cingulate cortex to be connected with a diverse range of cortical and subcortical areas. This brain region receives afferents from more thalamic nuclei than any other cortical region and also receives diffuse monoaminergic innervation of serotonin (5-HT), dopamine, and norepinephrine from the raphe nucleus, locus ceruleus, and ventral tegmental area, respectively. The overall pattern of connectivity reveals a broad distinction between autonomic functions and affect in ventral–rostral areas and between motor functions and cognition in more caudal areas of the anterior cingulate cortex.

1. Rostral–Ventral Regions

The ventral and rostral regions of the anterior cingulate cortex are closely connected with brain regions concerned with autonomic and visceromotor function. Ventral area 25 projects directly to the medulla, and areas 25, 24, and 32 may also influence autonomic function and affect via their reciprocal connections with the amygdala and insula, regions that project directly to the medulla. Ventral–rostral anterior cingulate cortex connects with other areas implicated in autonomic function—including the periaqueductal gray matter, orbitofrontal cortex, and nucleus accumbens of the ventral striatum—as well as having reciprocal connections with the hippocampus and parahippocampal regions and auditory areas of the superior temporal cortex.

2. Caudal Regions

a. Sulcus In contrast to the visceromotor involvement of ventral–rostral regions, caudal areas within the cingulate sulcus are associated with motor function. Peter Strick of the University Pittsburgh and colleagues have identified three separate cingulate motor areas that differ in cytoarchitecture and connectivity. Two of these regions lie just caudal to the anterior cingulate cortex as defined earlier and lie on the dorsal (CMAd) and ventral (CMAv) banks of the cingulate sulcus. A third, more rostral area (CMAr) is within area 24c.

The cingulate motor areas project directly to the spinal cord, together making up as much as 21% of the total frontal lobe projection. The motor involvement of cingulate cortex is further demonstrated by its projection to ventral motor nuclei of the thalamus and to the dorsal striatum (caudate nucleus and putamen). In addition, all cingulate motor areas have reciprocal, topographically organized connections with primary motor, premotor, and supplementary motor cortices. CMAd has the largest corticospinal neurons and sends the largest projection to the spinal cord and primary motor cortex. It receives input from the pallidum via the thalamus, as well as from parietal area 5. CMAv has smaller pyramidal neurons and receives input from thalamic regions innervated by the cerebellum, as well as from parietal areas 5 and 7, pre-SMA (area 6aβ), and prefrontal area 46. Finally, CMAr has the smallest projection to the spinal cord and primary motor cortex. It receives input from the pallidum (via a different thalamic relay than CMAd), parietal area 7, pre-SMA, and area 46. Thus, the anterior cingulate cortex is advantageously connected in such a way as to have a significant role in the control of actions.

b. Gyrus The connections of the anterior cingulate cortex are not limited to direct autonomic and motor functions. For example, area 24b receives input from the medial thalamic nuclei responsible for relaying nociceptive information from the spinal cord to the cortex. Nociceptive neurons in these nuclei have large, bilateral receptive fields. Area 24 projects, in turn, to the periaqueductal gray matter, both directly

and via the parafascicular nucleus of the thalamus. This pattern of connectivity, together with the high level of opiate receptor binding observed in cingulate cortex, implicates the anterior cingulate cortex in pain processing.

3. Intracingulate Connections

The combination of results from anterograde and retrograde labeling studies reveals strong interconnections within the ventral–rostral and caudal sulcal regions but very few projections between these regions. Thus, injection of a retrograde tracer in area 25 labels neurons in rostral area 24 and vice versa, but neither injection labels neurons in caudal gyral or sulcal areas. Similarly, small injections in the cingulate sulcus label many other neurons in this area, but very few in area 25 or rostral area 24. Reciprocal connections appear to exist between sulcal and gyral regions in caudal anterior cingulate cortex, but these gyral regions have a limited connection with rostral–ventral anterior cingulate cortex.

II. ANIMAL STUDIES

Studies of animal behavior begin to outline the roles of the anterior cingulate cortex in autonomic, affective, pain, and motor functions that were identified earlier with regard to the connectivity of this region.

A. Autonomic Function

Stimulation in area 24 has been shown to elicit almost every type of autonomic response—including changes in blood pressure, heart rate, respiratory rate, pupillary dilation, skin conductance, thermoregulation, and gastrointestinal motility—as well as causing changes in adrenal cortical hormone secretion (ACTH). Stimulation in the cingulate cortex can also lead to vocalizations. The evoked activity is usually linked to the role of the anterior cingulate cortex in visceromotor, as opposed to skeletomotor, function because the vocalizations are limited in number and have affective content.

Although lesions to the anterior cingulate cortex have little or no effect on baseline autonomic function, there is evidence that the anterior cingulate cortex is involved in autonomic conditioning (i.e., the development of autonomic responses to stimuli that are predictive of events, such as electric shocks, that evoke autonomic responses). Neurons in the anterior cingulate cortex show changes in activity as animals learn such contingencies, and lesions of the anterior cingulate cortex greatly reduce autonomic changes induced by the presentation of predictive stimuli. Lesions of the anterior cingulate cortex also change affective responses in more complex situations. Changes following ablation have been characterized in terms of blunted affect, reduced aggression, decreased motivation, and the disruption of mating and social behavior.

B. Skeletomotor Functions

Lesions in the anterior cingulate cortex can cause contralateral motor neglect. Stimulation of the CMAs evoke skeletomotor movements, although with lower probability and a higher stimulation threshold than the primary motor cortex. The evoked movements are typically fast, brief, limited to a single joint, and demonstrate a topographical organization. Complementary to these findings, single-unit recordings have shown that neuronal activity in the CMAs precedes voluntary movements. Neurons in CMAr are more activated during self-paced than stimulus-triggered movements and may precede the movement by long lead times (0.5–2 sec). In contrast, activity in the caudal CMAs shows less specificity to self-paced movements and is characterized by shorter lead times preceding the movement. Findings have linked CMA activity, particularly in CMAr neurons, to reward-based response selection tasks. Different CMA neural populations are active during the learning of new reward contingencies and during the execution of familiar, well-learned responses.

C. Pain

Neurons in area 24b of the rabbit anterior cingulate cortex are activated in response to painful stimuli. Like the nociceptive neurons in the medial thalamic nucleus that innervate the anterior cingulate cortex, area 24b neurons have large receptive fields that are often bilateral. The neurons are specific to nociceptive information, i.e., they are not activated by nonpainful somatosensory stimulation, and they show some specificity to the nature of the painful stimulus. These neurons are also activated during the learning of responses that lead to the avoidance of painful stimuli.

Lesions in the anterior cingulate cortex retard such learning and have been shown to reduce pain sensitivity in monkeys.

D. Learning

Apparent in the preceding discussion is the role of the anterior cingulate cortex in learning novel behaviors, whether as a conditioned response to predictors of painful stimuli, as an instrumental response to avoid such stimuli, or in response to reduced reward. Of interest in this context is the dopamine projection to cingulate cortex from the ventral tegmental area. The ventral tegmental dopamine system is known to be involved in processing rewarding or salient stimuli: Neurons in this area increase firing in response to unpredicted rewards and reduce firing if an expected reward is withheld. Consistent with the notion that dopamine is involved in adaptive behavior in the anterior cingulate cortex, this region, in common with other dopamine-innervated regions, supports self-stimulation. The properties of cortical self-stimulation differ, however, from those of medial forebrain structures—e.g., having longer acquisition times and lower sensitivity to amphetamine modulation—which may question a simple relationship between reinforcement learning in these systems.

III. NEUROIMAGING STUDIES

A small number of studies have used invasive techniques to study anterior cingulate function in humans, recording from or stimulating cingulate neurons during neurosurgery. Stimulation in ventral–rostral areas has been found to evoke autonomic changes in blood pressure and heart rate, visceromotor responses including salivation and vomiting, emotional responses including fear, agitation, and euphoria, and vocalizations with affective content. Anterior cingulate stimulation has also been shown to evoke motor responses of the face, hands, and legs, evidence, perhaps, of areas in the human brain corresponding to the CMAs seen in nonhuman primate anterior cingulate cortex. Recordings in caudal regions have revealed neurons with activity that is modulated during attention-demanding tasks such as arithmetic and generating lists of words. Other neurons, in further caudal and inferior regions, show sensitivity to painful stimuli. Overall, the findings are consistent with those reported in the previous section. However, these invasive studies are necessarily infrequent, and their interpretation is complicated by the neuropsychological condition of the patients involved that necessitated the surgery. Therefore, the preceding findings notwithstanding, much of the current understanding of anterior cingulate function has stemmed from more noninvasive imaging methods that can be applied in neurologically intact subjects.

Most neuroimaging studies of anterior cingulate function have used positron emission tomography (PET) or functional magnetic resonance imaging (fMRI). Both techniques rely on a subtractive methodology, comparing activation in brain areas across different experimental conditions that are designed to isolate specific processes. Such comparisons reveal areas of activation—areas more activated in the experimental than in the control condition—and, less commonly, areas of deactivation. These methodologies have very good spatial resolution on the order of millimeters, but relatively poor temporal resolution on the order of seconds to minutes. Electrophysiological recordings, on the other hand, have poor spatial resolution, but excellent temporal resolution on the order of milliseconds. Findings based on either PET, fMRI, or electrophysiological recordings regarding the role of the anterior cingulate cortex in human behavior are described later by relative region. For the most part, rostral and ventral regions have been linked with affective behavior, and more dorsal and caudal regions have been linked to cognitively driven actions. A major debate regarding the function of this region has resided in the domain of cognition primarily.

A. Affect

Activation in the rostral–ventral anterior cingulate cortex has been observed in normal subjects under a variety of conditions, including (1) when asked to recall sad memories and to view faces with sad expressions compared with recalling neutral memories and viewing neutral faces; (2) when anticipating an upcoming painful electric shock compared with resting activation; and (3) when exposed to scenes or words with emotional content compared with scenes and words with neutral content. Corresponding activations are observed in symptom provocation studies involving phobic, anxious, and obsessive–compulsive patients. These tasks have been characterized as having affective, emotional content, and therefore it is unsurprising that activations in the amygdala and orbitofrontal cortex are commonly observed to

co-occur with those in the anterior cingulate cortex. Often these paradigms require the subject to induce an affective state or think of emotional information that is contrary to the subject's current affective state. Moreover, some of these paradigms require the subject to attend to nonaffective attributes such as the color of an emotionally salient word (e.g., "Murder") rather than the word itself. Performance on such tasks is worst when the stimuli have emotional content versus neutral content and the rostral–ventral anterior cingulate cortex is activated

Deactivation in the rostral–ventral anterior cingulate cortex—often accompanied by deactivation in orbitofrontal and amygdala regions—has been observed in many of the attention-demanding tasks found to activate caudal regions of the anterior cingulate cortex. For example, Wayne Drevets of the National Institute of Mental Health and Marchus Raichle of Washington University have shown decreased activity in the ventral anterior cingulate cortex, amygdala, and orbitofrontal cortex—all areas implicated in emotional processing—during the performance of demanding cognitive tasks. Conversely, they have shown decreases in dorsal anterior cingulate cortex and dorsolateral prefrontal cortex activity—areas implicated in cognitive processing—during emotion-related tasks.

B. Cognition

Caudal anterior cingulate cortex has been found to be activated in a wide range of cognitive tasks. On the basis of their review of 29 PET studies, Peter Strick and colleagues drew a broad distinction between the most caudal area of the anterior cingulate cortex, area 24c′g, lying on and just posterior to the line of the anterior commissure and more anterior regions (areas 24′ and 32′). The former region is commonly activated during simple and/or familiar movements (typically in comparison with resting conditions involving no movements), movements that are associated with activity in the SMA but not lateral prefrontal areas. By contrast, areas 24′ and 32′ tend to be activated during the execution of more complex response selection tasks alongside activations in the pre-SMA and lateral prefrontal areas. On these grounds, Picard and Strick identified the caudal cingulate zone with CMAd observed in primates and tentatively correlated CMAr and CMAv with the anterior and posterior sectors, respectively, of the more rostral region activated by complex tasks.

A variety of complex tasks have been found to be associated with cingulate activity. Tasks that require subjects to respond to one stimulus in the presence of distracting, irrelevant information reliably result in such activation. For example, during Stroop task performance, the anterior cingulate cortex is more activated when subjects are required to name the ink color of a nonmatching color word (e.g. "Blue" written in red ink) than when naming the ink color of a stimulus that is not a color word. The anterior cingulate cortex is also activated when subjects are required to withhold responding in the context of speeded response tasks (e.g., no go and stop signal trials as compared with go trials) and when subjects must saccade away from rather than toward a target stimulus. Activity in corresponding regions is observed during the learning of new motor sequences or new noun–verb associations (compared with activity during the performance of familiar sequences or associations), following a switch to a new task (compared with repeated task performance), during free recall of remembered lists (compared to rest), when generating words that start with certain letters (compared to rest or repeating words seen or heard), and when required to make a decision based on degraded information (compared with undegraded information). As mentioned earlier, the performance of many of the tasks that activate the caudal anterior cingulate cortex is associated with corresponding deactivations in the rostral anterior cingulate cortex. Similarly, deactivations in the caudal anterior cingulate cortex have been observed during the performance of some of the tasks with affective content described earlier.

Attempts have been made to characterize the kinds of tasks that activate caudal anterior cingulate cortex to infer the function of this region. In general, the activating tasks have been categorized as difficult or complex or as eliciting multiple competing responses. Following from these observations, caudal anterior cingulate cortex has been variously attributed a role in attention or executive processes, in the decision process to select or initiate a response to "funnel" to motor areas, and in the monitoring of ongoing performance.

Whereas there is potentially significant overlap in these differing theories—e.g., performance monitoring can be considered as an important executive function in regulating response selection—there is evidence that the anterior cingulate cortex plays a direct role in performance monitoring. First, scalp electrophysiological recordings reveal a negative deflection in the event-related potential following

incorrect responses during choice reaction time tasks, a finding reported independently by Falkenstein and colleagues in Dortmund and by Gehring and coworkers in Illinois. This error-related negativity (ERN or Ne) begins around the time of the response and peaks roughly 100 msec after (see Fig. 2). The ERN has a midline frontocentral distribution on the scalp, and the anterior cingulate cortex is consistently found to be its most likely neural generator. The amplitude of the ERN wave varies as a function of the force with which the error is produced and the probability that the error is corrected, suggesting that the ERN relates somehow to error processing. Crucially, the observation of an ERN following performance feedback and following failures to withhold a response when required—errors that cannot be corrected—suggests that the ERN, a presumed electrophysiological index of anterior cingulate function, is related to performance monitoring rather than to the process of response selection itself.

Caudal anterior cingulate cortex is not uniquely activated following errors, as fMRI studies have shown this region to be activated even on correct trials. When subjects respond correctly, moreover, the activation observed is larger when there is response conflict—e.g., when distractor information presented is associated with a different response than target information—leading to the suggestion that the anterior cingulate cortex monitors for conflict rather than errors per se. Studies of response conflict provide a second line of evidence for the role of the anterior cingulate cortex in performance monitoring. For example, in a given task condition, if attention is effectively engaged to filter out irrelevant information, conflict will be low. In contrast, if attention is low, conflict will be high. When within-condition contrasts are made for subsets of trials with high attention–low conflict and subsets with low attention–high conflict, the anterior cingulate cortex activation is found to

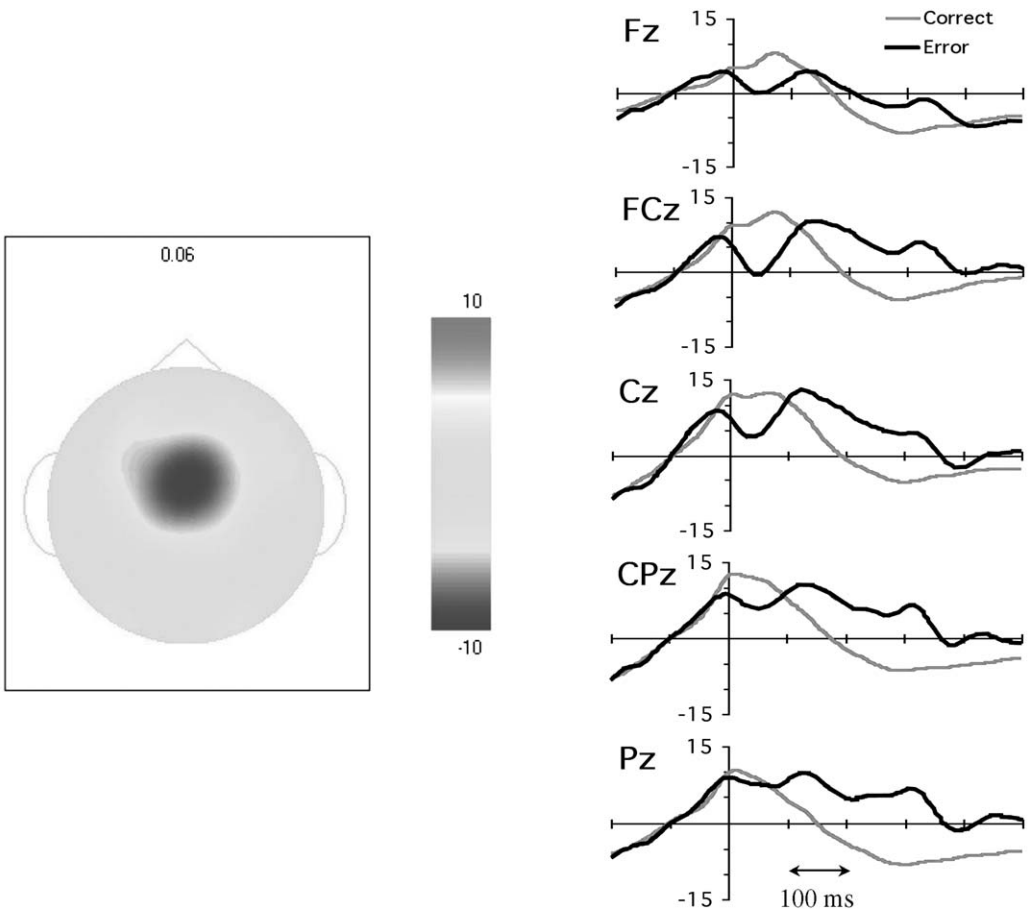


Figure 2 Scalp distribution and electrical tracing of the error-related negativity (ERN).

follow the degree of conflict rather than the level of attention, consistent with the monitoring hypothesis.

C. Pain

Comparisons of activations associated with painful versus nonpainful stimulation commonly reveal foci within posterior portions of the anterior cingulate cortex, in regions caudal and inferior to those observed in complex cognitive tasks. The anterior cingulate cortex is typically thought to be involved in processing the affective or motivational significance of painful stimuli. For example, the anterior cingulate cortex can be selectively activated by an illusion of pain (the "thermal grill," in which spatially interleaved warm and cool bars produce the illusion of noxious cold), compared with activations associated with either warm or cool components in isolation. This result suggests the involvement of the anterior cingulate cortex in subjective pain rather than the sensory coding of intensity. Consistent with this hypothesis is the finding that hypnotic suggestion to selectively manipulate ratings of pain unpleasantness while keeping perceived intensity constant positively correlated with activation in a caudal gyral region of the anterior cingulate cortex.

Studies of pain processing have, on occasion, revealed activation in rostral–ventral anterior cingulate cortex and in the caudal region usually activated during complex cognitive tasks. The former finding may reflect an affective or autonomic response to pain or the anticipation of pain. The latter finding may reflect cognitive processing of pain: such activations are observed, for example, when subjects are required to pay attention to the painful stimuli in order to keep count of the number of changes in stimulus intensity. Alternatively, the activity observed in the anterior cingulate cortex in response to pain may be associated with the representation of competing actions to make in response to the painful stimulus, such as avoidance versus tolerance of pain.

IV. CLINICAL STUDIES

The anterior cingulate cortex—not surprisingly given its involvement in cognitive and affective processes—has been implicated in a number of psychiatric disorders. This region has been shown to play a role in a number of human syndromes and disorders, providing neuropsychological evidence for its function in complex behaviors.

A. Epilepsy

Perhaps the clearest example of support for the role of the anterior cingulate cortex in movement, affect, and social behaviors is that of cingulate epilepsy syndrome described by Brent Vogt and others. Excessive activity in cases with seizures confirmed in the anterior cingulate cortex may impair consciousness and alter affective state and expression. Patients with such seizures may also display psychopathic or sociopathic behaviors.

B. Psychiatric Disorders

A number of psychiatric disorders have been linked with abnormalities in the function of the anterior cingulate cortex. Activity is elevated in this region in obsessive–compulsive disorder, tic disorder, and depression, and normalization of activity in this region occurs with behavioral and pharmacological treatment of these disorders in some cases. With severe forms of these disorders, such as with obsessive–compulsive disorder, cingulotomies have been shown to be effective in relieving the symptoms.

Other psychiatric disorders that have been associated with abnormal functioning of the anterior cingulate cortex include attention deficit hyperactivity disorder (ADHD) and schizophrenia. These patients show reduced activity in portions of the anterior cingulate cortex, particularly in caudal regions. Both attention deficit hyperactivity disorder and schizophrenia have been linked with poor dopaminergic modulation of prefrontal circuitry. Given the strong dopaminergic projections to the anterior cingulate cortex, these findings are consistent with a disruption in the modulation of anterior cingulate activity. Finally, individuals with psychopathic or sociopathic behaviors show less activity in the anterior cingulate cortex following errors in performance than do individuals without these characteristics, as evidenced in the ERN literature. This finding is consistent with the cingulate epilepsy literature mentioned previously.

C. Cingulotomies

Surgical removal of or infarcts to the anterior cingulate cortex provide further evidence for its role in

responsiveness to pain, movement (e.g., akinetic mutism), regulation of autonomic activity, and internal emotional responses, consistent with evidence from animal and imaging studies. Clearly, whether the lesion is more caudal or rostral affects the behaviors observed. Surgical removal of this region has been shown to diminish symptoms associated with severe cases of obsessive–compulsive disorder as described earlier and in cases of chronic and severe pain. One hypothesis for the involvement of anterior cingulate cortex in pain or obsessions is associated with attention to the pain or obsession; however, this interpretation has been called into question specifically with regard to pain.

V. DEVELOPMENT

The anterior cingulate cortex, like other prefrontal regions, undergoes a prolonged developmental process relative to sensorimotor regions. Pasko Rakic, Jean-Pierre Bourgeois, and others at Yale University have shown that, even though neurogenesis is completed in the anterior cingulate cortex (area 24) before the visual cortex, synaptogenesis occurs later. Synaptic density in area 24 then remains at a high level until sexual maturity. Speculations relating the development of this brain region to behavior have been proposed by Michael Posner and Mary Rothbart of the University of Oregon on the basis of behavioral data from infants and children on tasks thought to require the anterior cingulate cortex. B. J. Casey at the Sackler Institute in New York and colleagues have used MRI-based morphometric studies of this region to relate its size and symmetry to behavioral development. More recently, Casey has used functional MRI studies for this purpose.

A. Behavioral Studies

Behavioral studies suggesting development of the anterior cingulate cortex use behavioral tasks shown to activate this region in imaging studies with adults using Strooplike or go–no go tasks. The assumption is that, if a child can perform the task, then the anterior cingulate must be mature enough to support the behavior. Accordingly, one may track the development of behavioral performance on these tasks across ages to determine the degree of development of the brain structure. A substantial increase in the ability to perform these Stroop-like tasks occurs across the ages of 2–5 years. This evidence is inferential and not conclusive of a specific structure–function relation between the development of the behavior and the anterior cingulate cortex. Clearly, maturation of this region alone is not occurring in isolation of other brain systems' maturation.

B. MRI-Based Morphometry

MRI-based morphometric measures show a correlation between size of the right anterior cingulate region and performance on an attention switching task in 6- to 18-year-olds (see Fig. 3). The larger the anterior cingulate region, the faster the performance on the attention switching task that required the subject to switch from attending to color to attending to shape, alternating from trial to trial. This region did not correlate with a simple nonswitching attentional task, suggesting its involvement in tasks that have attentional conflict or require cognitive control. Even after correcting for total brain size and estimated IQ, the correlation remained significant, suggesting that the association was not a general one but specific to this brain region. Yet this association is indirect in that the two measures are assessed at different time points.

C. Functional Neuroimaging Studies

Perhaps the most direct evidence of development of the anterior cingulate cortex and behavioral regulation comes from an fMRI study of children between the ages of 7 and 11 years and young adults. In that study, Casey showed an increase in MR signal intensity in the anterior cingulate cortex as a function of increased number of errors on a go–no go task (Fig. 4). This task required the subject to override a compelling response. Those individuals experiencing the most difficulty with this task showed increased activity in this region. However, in this study, only regions of the prefrontal cortex were examined; thus, again the question may be raised as to how specific this region is to behavioral regulation in children. Other regions including portions of the basal ganglia and projections to and from these regions may be likely candidates. Interestingly, this study demonstrated a dissociation of prefrontal regions from anterior cingulate cortex function, with ventral prefrontal regions correlating with accuracy and the anterior cingulate regions correlating with errors. Thus, whereas the anterior cingulate may index errors or response conflict, the

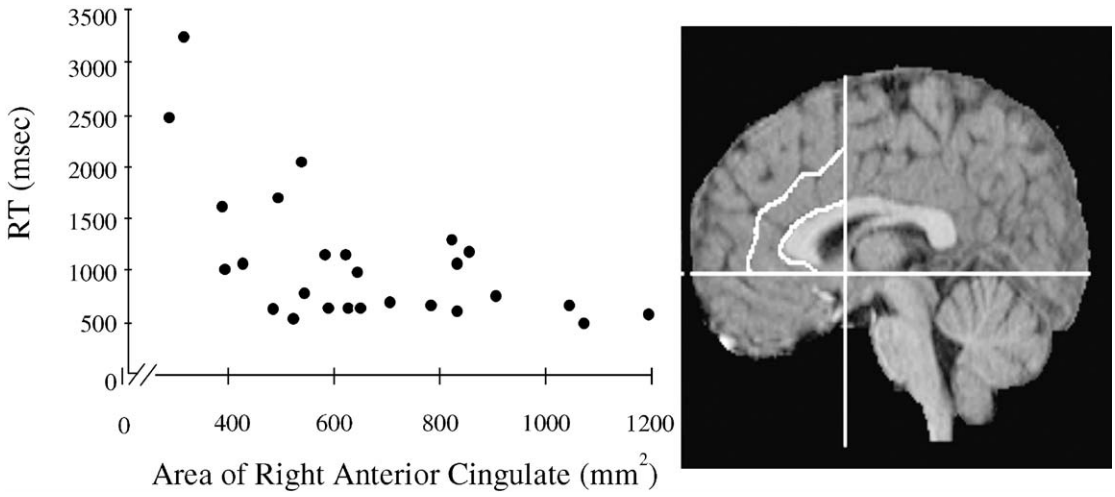


Figure 3 Reaction times during performance of an attention shifting task in children between the ages of 6 and 18 years as a function of the size of the anterior cingulate cortex.

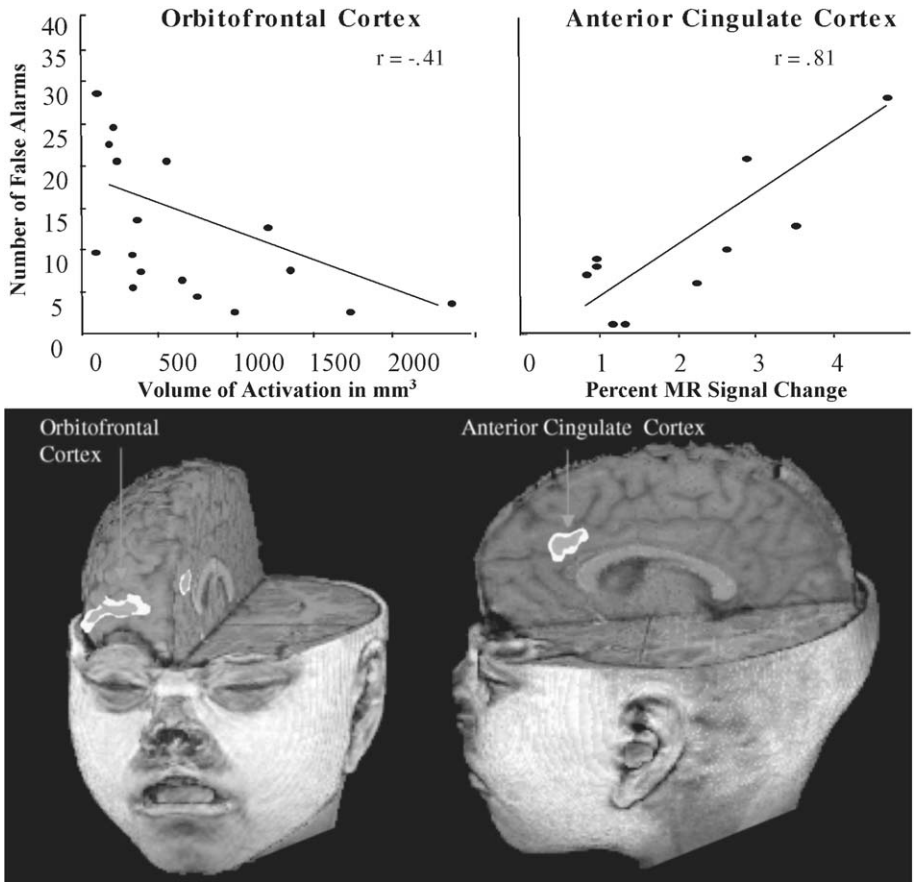


Figure 4 Location, volume, and percent change in MR signal intensity as a function of the number of errors for the ventral prefrontal and anterior cingulate cortices.

ventral prefrontal region indexed the degree of conflict resolution on this task. A similar dissociation has been shown in dorsolateral prefrontal cortex and anterior cingulate cortex on a version of a Stroop task more recently by MacDonald and Carter at the University of Pittsburgh.

VI. GENETICS AND EVOLUTION

As described, the anterior cingulate cortex has been implicated in a number of psychiatric disorders. Interestingly, all of these disorders show familial patterns of inheritance, increased risk among first degree relatives of affected patients, and increased concordance in identical vs fraternal twins. The *heritability* (see Glossary) has been estimated for schizophrenia (0.6), attention deficit hyperactivity disorder (0.79), obsessive–compulsive disorder (0.68), and major depression disorder (0.6). Estimates of heritability have been extended to specific cognitive functions in normal populations as well. Tasks that activate the anterior cingulate cortex, such as spatial working memory, divided attention, and attentional set shifting, have been examined in identical and fraternal twin populations and have high heritabilities. These findings suggest that genetic factors play a role in behaviors associated with anterior cingulate abnormalities and normal anterior cingulate function. These genes may be important for the establishment of proper connectivity during pre- and postnatal development and/or for proper physiological homeostasis under stressful or injurious conditions. Of the approximately 35,000 *genes* (see Glossary) in the human genome, it remains a challenge to identify the genes of critical importance.

A. Candidate Genes

Several candidate genes have already emerged. Gene expression studies in mice and humans have identified many genes whose expression is enriched in the anterior cingulate cortex. For example, the *neurotrophin-3* (*NT3*) gene, a molecule involved in the growth and survival of developing neurons, shows enriched expression in the anterior cingulate cortex during development. Support for the role of *NT3* in anterior cingulate development and function comes from the results of a *gene association study* (see Glossary) showing that variation in the *NT3* gene was associated with schizophrenia. Similarly, the *neurotensin* (*NTS*) gene shows enriched expression in the cingulate cortex at birth. Gene association studies on the promoter of the *neurotensin receptor* (*NTSR1*) have also shown associations with schizophrenia. Other genes whose expression is enriched in the developing cingulate cortex include *Emx2* and *ER-81*. Mutations in the *Emx2* gene have been shown to cause schizencephalies, but it remains to be seen whether less severe genetic variations give rise to subtle cognitive impairments. The *fragile-X mental retardation* gene (*FMR2*) is also enriched in limbic cortex and is associated with mild mental retardation.

B. Dopamine-Related Genes

The expression of the *dopamine D4 receptor* (*DRD4*) gene in the anterior cingulate cortex is of particular interest. As reviewed by Williams and Rakic from Yale University, the anterior cingulate cortex contains high levels of DA innervation in layers II–VI. Dopaminergic axonal fibers arising from the VTA form axon–axononal connections and allow large regions of this cortical area to be globally regulated. The cingulate cortex contains interneuron populations expressing *dopamine D2 receptor* (*DRD2*) and *DRD4* receptors and pyramidal neurons expressing *dopamine D1 receptor* (*DRD1*) receptors. Whereas most DA receptors are expressed widely throughout the brain, *DRD4* shows its highest levels of expression in the frontal and cingulate areas. Interestingly, a polymorphism in the cytoplasmic loop of the *DRD4* gene has been associated with attention deficit hyperactivity disorder. Similarly, a specific polymorphism of the *dopamine D3 receptor* (*DRD3*) and *DRD2* genes was associated with schizophrenia, as were the *dopamine β-hydroxylase* (*DBH*) and *catechol-O-methyl transferase* (*COMT*) genes. Figure 5 illustrates the various alleles (see Glossary) of the *DRD4* gene in normal volunteers based on DNA samples from cheek cells. Basic genotyping studies of this format no doubt will increase given the publication of the human genome in *Science* and *Nature*.

Gene expression studies have also shown that the anterior cingulate cortex is highly sensitive to environmental stress. Anoxia, maternal separation, amyloid protein expression, and drug abuse all induce hypometabolism, gliosis, and programmed cell death in the anterior cingulate cortex. Exposure to stress induces the expression of *glucocorticoid receptor* (*GR*), a transcription factor that mediates the cellular response to stress as shown by Bruce McEwen of Rockefeller University. Stress-induced excitatoxic damage has

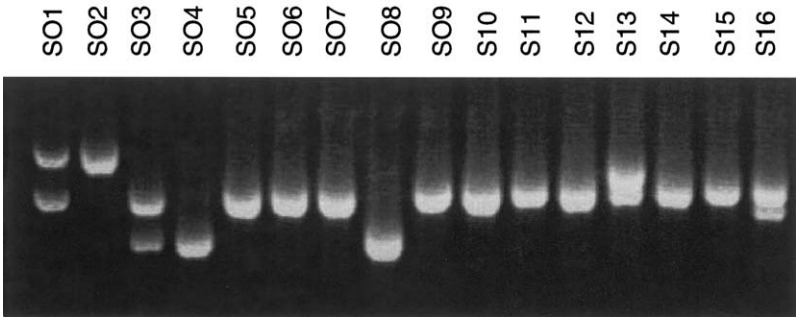


Figure 5 Genotyping of the *DRD4* gene in 16 healthy adults using DNA samples from cheek cells.

been noted in the anterior cingulate cortex in schizophrenia. Specifically, Francis Benes of Harvard Medical School has shown that dopaminergic innervation of interneurons in layers II and V is increased in post mortem analyses of schizophrenia. Under normal conditions, the synaptic contacts on local interneurons that utilize *DRD2* receptors result in inhibition of cell activity. The hyperinnervation of interneuronal *DRD2* contacts is suspected to disable local inhibition of pyramidal cells and lead to excess glutamatergic signaling and excitotoxicity in downstream brain areas. Because this type of excitotoxic damage is mediated by the *GR* gene, it is likely that genetic variation in this gene and/or its downstream targets explains the phenomenon of a gene × environment interaction seen in many psychiatric disorders.

C. Evolution

Knowledge of how genetic factors modulate cingulate development and function may shed light on the unique evolutionary history of this area in humans. The anterior cingulate arises from the medial (limbic) and medial–dorsal telencephalic pallium. These cortical regions show remarkable evolutionary conservation among mammals, birds, and reptiles in contrast to dorsal telencephalic pallium or neocortex, which is highly divergent in volume and microstructure. For these reasons, the cingulate gyrus is often referred to as a phylogenetically "ancient" structure. Whereas this appears to be true in the case of most mammals, analyses of cingulate cortical microstructure in humans show a remarkable leap of recent evolution. Betz, in 1881, first noted the presence of large motor neurons in the cingulate region of humans but not in other great ape species. Similarly, Patrick Hof and John Allman have found spindle cells, so named for their elongate and gradually tapering morphology, predominantly in layer Vb in the medial wall of the cingulate gyrus. These cells are not seen in old world primates but are found in bonobos, chimpanzees, and humans. Phylogenetic studies comparing human and primate genomic sequence data have attempted to explain this recent evolutionary leap. Studies of the role of the *DRD4* gene in human evolution and migration show a high correlation (0.8) between the frequency of the exon III seven-repeat allele and miles of migration. Migration was estimated for ethnic groups around the world as the human population expanded out of Africa. The *DRD4* gene has undergone strong positive selection during primate evolution. Additional studies of genes that contribute to the development and physiology of the cingulate gyrus may shed light on the evolution of this important brain area. Most importantly, many of these genes may serve as useful molecular targets for pharmacologic intervention for the treatment of psychiatric disorders.

VII. CONCLUSIONS

In sum, the anterior cingulate cortex appears to play an important role in autonomic, affective, and cognitive behavior. The precise nature of its role is still uncertain and a number of questions remain unanswered. For example, is the anterior cingulate cortex organized by function or by domain? In other words, do different regions of the anterior cingulate cortex function according to different rules or mechanisms, or is there a common underlying function that applies to all domains of behavior including the domains of sensation (e.g., pain), affect, and cognition? The anterior cingulate cortex appears both anatomically and functionally well-situated to integrate information across these domains. Paus and others have demonstrated somatotopic organization of the cingulate cortex and suggest that this region modulates and funnels cognitive and motor

commands to the motor system. One could interpret this view as the cingulate having a specific function that only differs across regions in terms of the type of information guiding the action. Although Paus' work was specific to motor and cognitive commands, it may be plausible to modify this theory to accommodate the domain of emotion and emotionally guided actions (e.g., to avoid or approach an aversive or appetitive stimulus). A potential caveat in this line of thinking is the inverse relation frequently observed between the affective and cognitive portions of the anterior cingulate gyrus in imaging studies.

Clearly, the anterior cingulate cortex plays an important role in attention and behavioral regulation, and Posner and colleagues have been at the forefront of such speculations. However, the question remains in the cognitive literature as to whether the anterior cingulate's primary function is one of executive control, whereby the cingulate functions as a controller to resolve conflict, or one of detection, whereby the anterior cingulate monitors for errors or conflict. A number of groups have begun to dissociate these functions in adult and developmental neuroimaging studies that attribute conflict detection to the anterior cingulate cortex and conflict resolution to other regions like the prefrontal cortex. Either way, the structure is critical for guiding complex cognitive actions. Its specific role in behavior will be further specified with the refinement of biologically plausible computational models that will help constrain interpretations of the electrophysiological and imaging data. Recently, Matthew Botvinick of the University of Pittsburgh has developed such a model, which seems to clarify the role of the anterior cingulate cortex in cognitive conflict and dissociates conflict from cognitive control. How well this model of anterior cingulate function will hold up within the domains of sensation and affect remains unanswered. However, it is the case that paradigms of affective processing often require the subject to induce an affective state or think of emotional information that is contrary to or in conflict with the subject's current affective state. Moreover, affective paradigms have conflicting responses built in as with the emotional Stroop task, which requires the subject to name the color of the emotionally salient word "Murder" when reading the word is the more compelling response. Likewise, in the case of pain research, one can imagine the inherent response conflict associated with the competing representations to avoid vs tolerate a painful stimulus. Each of these examples provides a plausible role of the anterior cingulate in monitoring unresolved conflict across the domains of emotion and pain. Clearly, exciting work and theory development on anterior cingulate function will continue with the advancements of noninvasive neuroimaging and genetic methodologies in combination with computational and animal models.

See Also the Following Articles

CORPUS CALLOSUM • DOPAMINE • EPILEPSY • MOTOR CORTEX • SCHIZOPHRENIA

Suggested Reading

Benes, F. M. (2000). Emerging principles of altered neural circuitry in schizophrenia. *Brain Res. Brain Res. Rev.* **31,** 251–269.

Botvinick, M., Braver, T. S., Carter, C. J., Barch, D. M., and Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychol. Rev.*, in press.

Bush, G., Luu, P., and Posner, M. I. (2000). Cognitive and emotional influences in anterior cingulate cortex. *Trends Cogn. Sci.* **4,** 215–222.

Devinsky, O., Morrell, M. J., and Vogt, B. A. (1995). Contributions of anterior cingulate cortex to behaviour. *Brain* **118,** 279–306.

Drevets, W. C., and Raichle, M. E. (1998). Reciprocal suppression of regional cerebral blood flow during emotional versus higher cognitive processes: Implications for interactions between emotion and cognition. *Cognition Emotion* **12,** 353–385.

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., and Donchin, E. (1993). A neural system for error detection and compensation. *Psychol. Sci.* **4,** 385–390.

MacDonald, A. W., Cohen, J. D., Stenger, V. A., and Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science* **288,** 1835–1838.

Nimchinsky, E. A., Gilissen, E., Allman, J. M., Perl, D. P., Erwin, J. M., and Hof, P. R. (1999). A neuronal morphologic type unique to humans and great apes. *Proc. Natl. Acad. Sci. U.S.A.* **96,** 5268–5273.

Paus, T., Petrides, M., Evans, A. C., and Meyer, E. (1993). Role of the human anterior cingulate cortex in the control of oculomotor, manual, and speech responses: A positron emission tomography study. *J. Neurophysiol.* **70,** 453–469.

Peyron, R., Laurent, B., and Garcia-Larrea, L. (2000). Functional imaging of brain responses to pain. A review and meta-analysis (2000). *Neurophysiol. Clin.* **30,** 263–288.

Picard, N., and Strick, P. L. (1996). Motor areas of the medial wall: A review of their location and functional activation. *Cerebral Cortex* **6,** 342–353.

Posner, M. I., and Rothbart, M. K. (1998). Attention, self-regulation and consciousness. *Philos. Trans. R. Soc. London B Biol. Sci.* **353,** 1915–1927.

Shima, K., and Tanji, J. (1998). Role for cingulate motor area cells in voluntary movement selection based on reward. *Science* **282,** 1335–1338.

Vogt, B. A., and Gabriel, M. (1991). *Neurobiology of Cingulate Cortex and Limbic Thalamus: A Comprehensive Handbook*. Birkhauser, Boston.

Williams, S. M., and Goldman-Rakic, P. S. (1998). Widespread origin of the primate mesofrontal dopamine system. *Cerebral Cortex* **8,** 321–345.

Anxiety

ERIC VERMETTEN, DENNIS S. CHARNEY, and J. DOUGLAS BREMNER
Yale University

GLOSSARY

amygdala Almond-shaped collection of three nuclei in the anterior portion of the temporal lobe, central portion of the 'limbic' system, also called the 'emotional brain'.

anxiety The apprehensive anticipation of future danger or misfortune accompanied by a feeling of dysphoria or somatic symptoms of tension, when there is no true threat.

fear Feelings of anxiety associated with real external threat.

hippocampus A brain structure consisting of grey matter, receiving highly processed sensory information from the cerebral cortex, playing an important role in learning and memory.

locus coeruleus Dark blue-pigmented cell structure containing norepinephrine in the brain stem above the fourth ventricle projecting to all parts of the CNS, also called the 'alarm bell' of the brain.

panic disorder Disorder characterized by one or more episodes of abrupt, intense anxiety, i.e., a panic attack and persistent apprehension about the recurrence of these episodes.

phobic disorder Disorder of long lasting and persistent anxiety for a thing or situation that compels one to avoid the feared stimulus.

posttraumatic stress disorder Disorder of persistent re-experiencing of a traumatic event in dreams and memories; persistent avoidance of stimuli associated with the event, and increased arousal. The disorder occurs after exposure to a traumatic event that coincided with intense fear, helplessness or horror.

For more than a century scientists have wondered about the basis of fear and anxiety in the brain; however, only recently have clinicians developed neuroanatomical hypotheses to explain specific anxiety disorders. Although anxiety disorders affect a large portion of the population, relatively little is known about the neurobiological correlates of fear, anxiety, and anxiety disorders. Despite the lack of direct neurobiological data regarding anxiety disorders, tremendous advances have been made in understanding the neurobiology of stress and the response to threat. The neurobiology of stress and fear responsivity can give important clues to the neurobiological underpinnings of anxiety. Whether fear is due to an actual external threat (e.g., to be killed) or internal threat (e.g., hypoxia), a neutral but mislabeled cue now categorized as threat related, or an imagined threat, the brain responds in a similar fashion. Researchers in the clinical neuroscience of anxiety disorders have applied animal models of stress to humans with anxiety disorders, testing hypotheses derived from animal studies in human subjects with anxiety disorders. Based on these studies a working model for a neural circuitry of anxiety and fear can be described. This article reviews neural correlates of fear and anxiety and the clinical neuroscience of human anxiety disorders. Connections will be made between neurobiology, functional neuroanatomy, and the clinical presentation of patients with anxiety disorders. This work is

ongoing and many of the models proposed may be subject to modification and revision as our knowledge base in this exciting area continues to expand.

I. THE EXPRESSION OF FEAR AND ANXIETY

Emotion is composed of several elements, including affective valence and arousal. Fear, like anger, is a protective emotion. It is a feeling of agitation and anxiety caused by the presence or imminence of danger and refers to a state or condition marked by this feeling. Fear has evolved as a vitally important physiological response to dangerous situations that prepares one to evade or confront a threat in the environment. Fear has played a critical role in survival, and therefore efficient fear responses have been selected through evolution.

Anxiety is the subjective sensation that accompanies the body's response to threat that is similar to the fear reaction but in which there is no real threat. Panic is the sudden onset of an overpowering feeling of fear or discomfort that occurs suddenly, and it is associated with increased heart rate, sweating, trembling, shortness of breath, feeling of choking, nausea, dizziness, derealization, fear of losing control and fear of dying, numbness, and chills. For some individuals, the frequency, duration, intensity, or context of anxiety are extreme and can interfere with normal development and functioning.

Anxiety differs from fear in that the fear-producing stimulus is either not present or not immediately threatening, but in anticipation of danger the same arousal, vigilance, physiologic preparedness, negative effects, and cognitions occur. Different types of internal or external factors or triggers act to produce the anxiety symptoms of panic disorder (PD), agoraphobia, posttraumatic stress disorder (PTSD), specific phobias, generalized anxiety disorder (GAD), and the prominent anxiety that commonly occurs in major depression.

As one of the most readily accessible and perhaps easily understood of the major symptoms of mental disorders, anxiety may often take the concrete form of intense fear that is experienced in response to immediately threatening experiences, such as being involved in a near deadly car accident or witnessing a violent bank robbery. Experiences like this are typically accompanied by strong emotional responses as well as physical symptoms, such as rapid heartbeat, increased perspiration, cold hands or feet, hot flushes, sensations of shortness of breath, lightheadedness, trembling, restlessness, muscle tention, feelings of fear or dread, of increased gastrointestinal motility, and urge to urinate. Psychiatrists consider anxiety as an excessive alarm response system when there is no real threat—a pathological expression of fear responses.

The appropriate regulation of fear is critical to the survival of virtually every higher organism in every environment. However, the mechanisms that regulate fear may break down in a wide variety of circumstances, leading to excessive or inappropriate expression of anxiety. Specific examples include phobias, panic attacks, and generalized anxiety. In phobias, high-level anxiety is aroused by specific situations or objects that may range from concrete entities such as snakes to complex circumstances such as social interactions or public speaking. Panic attacks are brief and very intense episodes of anxiety that often occur without a precipitating event or stimulus. Generalized anxiety represents a more diffuse and nonspecific kind of anxiety that is most often experienced as excessive worrying, restlessness, and tension occurring with a chronic and sustained pattern. In each case, an anxiety disorder may be said to exist if the anxiety experienced is disproportionate to the circumstance, is difficult for the individual to control, and interferes with normal functioning. In obsessive–compulsive disorder (OCD), individuals experience a high level of anxiety that drives their obsessional thinking or compulsive behaviors. When such an individual fails to carry out a repetitive or ritualistic behavior such as hand washing or checking, there is an experience of severe anxiety. Thus, although the outward manifestations of OCD may seem to be related to other anxiety disorders, there appears to be a strong component of abnormal regulation of anxiety underlying this disorder. Since OCD is in many ways distinct from the other anxiety disorders, it is not addressed in detail in this article. PTSD can occur after an individual has responded with intense fear, helplessness, or horror after having been exposed to a traumatic event that involved actual or threatened death or serious injury or a threat to the physical integrity of oneself or others. The characteristic symptoms that result from such a traumatic event include the persistent reexperience of the event in dreams and memories, persistent avoidance of stimuli associated with the event, and increased arousal.

The clinical acceptance of the heterogeneity of anxiety disorders suggests that there are distinct neurobiological substrates for each. This seems to be most true for PD and PTSD, and clinicians have developed neuroanatomical hypotheses that are related to these specific anxiety disorders.

II. PREVALENCE OF ANXIETY DISORDERS

The American Psychiatric Association first recognized anxiety disorders in 1980 as a separate group of psychiatric disorders. The concept of neurosis (neurasthenic neurosis, anxiety neurosis, phobic neurosis, and obsessive–compulsive neurosis) in previous classifications was abandoned because it was considered too vague. Anxiety disorders now include PD, PTSD, social phobia, specific phobia, OCD, and GAD. Anxiety disorders are by far the most common of psychiatric disorders (25%), followed by affective disorders (17%).

In the past decade, large epidemiological studies have provided information about the prevalence of anxiety disorders in the general population. A landmark epidemiological study in the United States in 1994 found lifetime prevalence rates for all anxiety disorders combined to be 19.2% for men and 30.5% for women. Phobic disorders are the most common diagnosis in broad-based assessments of psychiatric disorders in the community (affecting about 13% of individuals at any point in their lives), whereas PTSD affects 8% of the general population, GAD 5%, and PD and OCD each about 1%. Even apart from the considerable comorbidity between the anxiety disorders, comorbidity rates between anxiety disorders and depressive disorders are high (especially PD with agoraphobia, social phobia, and OCD), ranging from 30% for coexisting in time to 60% lifetime. Comorbidity rates between GAD or PTSD and other psychiatric disorders are even higher, about 80% for GAD and 90% for PTSD (lifetime).

Anxiety disorders have a disabling impact on daily life and are associated with considerable morbidity for the individual patient. They also impose a substantial economic impact on our society as a whole. Twenty-three million people suffer from an anxiety disorder in the United States. The total cost of anxiety disorders in this society is estimated to be $42.3 billion per year as direct and indirect expenses. PTSD and PD have the highest rates of service use. There are no obvious reasons to assume that the picture for other countries is very different (Table I).

Table I
Lifetime Prevelance Rate of Anxiety Disorders[a]

Disorder	Rate (%)
Social phobia	13
Specific phobia	12
PTSD	8
GAD	5
PD	1
OCD	1

[a]Based on results of the National Comorbidity Study, USA (1994).

III. ETIOLOGY OF ANXIETY AND ANXIETY DISORDERS

In broad terms, the likelihood of developing anxiety involves a combination of life experiences, psychological traits, and/or genetic factors. Anxiety disorders are so heterogeneous that the relative roles of these factors are likely to differ. Some anxiety disorders, such as PD, appear to have a stronger genetic basis than others, although no single gene is responsible for any anxiety disorder in the majority of cases. Quantum trait loci (QTL) have been successfully applied in plant and nonhuman genetics, relating dimensional traits to genes. QTL provide a means of combining molecular and quantitative approaches to complex behavior genetics.

Other anxiety disorders are more rooted in stressful life events. It is not clear why females have higher rates than males for most anxiety disorders, although some theories have suggested a role for the gonadal steroids. Other research on women's responses to stress suggests that women experience a wider range of life events (e.g., those happening to friends) as stressful compared with men, who react to a more limited range of stressful events, specifically those affecting themselves or close family members. To delineate the contributions of the environment, population-based research is important in future research.

IV. FUNCTIONAL NEUROANATOMICAL MODELS OF FEAR AND ANXIETY

A. History of Neuroanatomical Modeling of Fear and Anxiety

There has been a long history of hypotheses related to the neurobiology of human anxiety. The central role of a subcortical network of brain structures in emotion in

general was hypothesized by Papez in 1937. In 1949, MacLean coined the term "limbic" system, integrating Papez's original circuit (hypothalamus, anterior thalamus, cingulate gyrus, and hippocampus) and other anatomically and functionally related areas (amygdala, septum, and orbitofrontal cortex). Over the years, various regions have been added or removed from this "emotion" processing circuit. Papez hypothesized that several telencephalic and diencephalic structures which form a border ("limbic" = border) around the diencephalon constituted a circuit, which controlled the emotions. He suggested that blockage of information flow at any point along this circuit would cause disorders of affect (i.e., mood). Removal of the cerebral cortex of the cat, leaving only subcortical regions including amygdala, thalamus, hippocampus, and hypothalamus, resulted in accentuated fearful responses to potentially threatening or novel stimuli, accompanied by signs of diffuse sympathetic activation such as increased blood pressure, sweating, piloerection, and increased secretion of epinephrine from the adrenal medulla. This behavioral response was termed "sham rage". In this situation the animal behaves aggressively with little or no provocation. These experiments led to the hypothesis that subcortical brain structures above the level of the midbrain, such as the hypothalamus, hippocampus, cingulate, entorhinal cortex, and thalamus, mediate human anxiety responses. When the amygdala was also removed, as described by Kluver and Bucy, a typical "psychic blindness" occurred, indicating that these animals had good visual acquity but were "blind" to the psychological significance of stimuli. They showed no fear responses; the lesioned cat remained unprovoked, neither aggressive nor fearful of the appearance of a dog. Thus, the amygdala was added to these limbic brain structures and given a pivotal importance in emotional memory and fear responsivity.

B. Limbic System

The critical role of the limbic system is its participation in the experience of emotions, moods, and consolidation of short-term into long-term memory. It receives input from various regions of cortical structures, such as the gustatory (through the solitary nucleus) and olfactory areas (the olfactory bulb), and major inputs from the brain stem. One of these brain stem inputs comes from the locus coeruleus (LC), a relatively small group of pontine neurons near the central grey. The second area of brain stem input comes from the thalamus.

Since the 1980s, when the prevailing view was that excess discharge of the LC with the acute stress response was the major contributor to the etiology of anxiety, its role has become more balanced and in tune with the contribution of other circuits and systems in fear responsivity. It appeared that the acute stress response relates to arousal and vigilance rather than anxiety. With anxiety, the concern about the stressor is out of proportion to the realistic threat. Anxiety is often associated with elaborate mental and behavioral activities designed to avoid the unpleasant symptoms of a full-blown anxiety or panic attack. Also, anxiety is usually longer lived than arousal. Moreover, anxiety can occur in isolation without exposure to an external stressor.

Although technological developments have contributed to a more differentiated assessment of brain structures involved in anxiety, the past decade has seen considerable research on the role of the amygdala. This has led to hypotheses that the amgygdala plays a critical role in the elaboration of anxiety, panic, and related symptoms of arousal and fearful avoidance. However, the role of mental imagery and conscious memory recall in the development of anxiety suggests that other cortical areas, such as hippocampus and frontal cortex, must play a role in anxiety since the amygdala does not have the capacity for conscious recall, and research findings are consistent with a role for these structures in anxiety disorders. Moreover, the role of specific neurochemical systems has become clearer. Long-term dysregulation of these systems is strongly associated with the development of anxiety disorders, including PD, PTSD, and phobic disorders. In the mediation of symptoms of the anxiety disorders, these neurochemical and neurotransmitter systems act on a substrate of specific cortical and subcortical brain areas.

C. Animal Research in Fear and Anxiety

Animal tests of fear and anxiety are used both to screen new compounds for potential anxiolytic action and to study their neural substrates. Until the mid-1970s, animal tests consisted of delivering shocks as a punishment, most often for an operant lever-press response. These tests were developed as screening tests for the pharmaceutical industry. Matching particular tests of fear and anxiety to particular anxiety disorders

is an extremely difficult task. The social interaction test (placing rats in an unfamiliar or brightly lit environment), the elevated plus-maze (placing the animal on an elevated open arm), predator exposure stress, forced swim, and social defeat or subordination stress are models of fear and anxiety.

In the wide range of approaches used to study fear and anxiety in animal studies, two sets of tests probe their responses. The first set uses models of conditioned fear; the second uses models of unconditioned fear. Both models presuppose that aversive stimuli, such as foot shock or novelty, induce a central state of fear, which can be quantified through specific behavioral and physiological measures (restlessness, avoidant behaviors, body temperature, body posture, tremor, and sweating). Animals display fear responses to aversive events, such as being forced to swim in cold water, predator exposure, or being given electric foot shock.

Typically, if a caged rat is subjected to electrical foot shock, a protective defense system is engaged. It is likely to flee if an exit is available (fear) or attack a cagemate if there is one (anger). If the shocks are repeated randomly and uncontrollably it is likely that it will cower helplessly and later become dull and unresponsive (learned helplessness or depression). On the other hand, in conditioned fear experiments, after repeated pairings of an aversive stimulus with a formerly neutral cue, animals will experience this state of fear even when only the cue is present. The model of conditioned fear provides a critical survival-related function in the face of threat by activating a range of protective behaviors. Classical fear conditioning experiments and experiments in which anxiety is induced by other means (conditioned models and unconditioned models of fear) have expanded our understanding of mechanisms of fear and anxiety and are currently our best approaches as models for human anxiety disorders.

Using animals to measure the effects of chronic stress on neurochemical systems assumes that animal models of the anxiety disorders are directly applicable to human anxiety disorders. However, if there is any validity to our differentiation of unique disorders of anxiety and depression, then it is not possible that, regarding animal models, ''one size fits all.'' These limitations of animal models have to be borne in mind and are useful in guiding research in anxiety disorders. PTSD, PD, and phobic disorders have been found to have many phenomenological and neurobiological characteristics and can benefit from the application of animal models of stress.

V. KEY BRAIN STRUCTURES MEDIATING FEAR AND ANXIETY BEHAVIORS

The brain structures that constitute a neural circuit of fear and anxiety should have the following features:

1. There is sufficient afferent sensory input to permit assessment of the fear- or anxiety-provoking nature of the external threat or internal stress.
2. Neural interactions among the brain structures capable of incorporating an individual's prior experience or memory into the appraisal of stimuli. These interactions are important in the attachment of affective significance to specific stimuli and the mobilization of adaptive behavioral responses.
3. Efferent projections from the brain structures should be able to mediate an individual's neuroendocrine, autonomic, and motor response to threat as well account for the pathological reactions that result in anxiety-related signs and symptoms.

To underscore its survival importance, many brain areas with redundant circuits are involved to subserve this important constellation of behaviors. Critical brain structures capable of incorporating an individual's prior experience or memory into the appraisal of stimuli are amygdala, LC, hippocampus, thalamus, hypothalamus, periaqueductal grey (PAG), and prefrontal cortex. Alterations in neurochemical and neurotransmitter systems that mediate the stress response also play a role in anxiety. Important neurotransmitters are corticotrophine-releasing factor (CRF), adrenocorticotrophic hormone (ACTH), norepinephrine (NE), epinephrine, dopamine, cortisol, benzodiazepines (Bzs), opioids, and other neurochemical systems.

A. Amygdala

The amygdala is a large nucleus (actually a complex of at least three nuclei) that lies in the temporal lobe just lateral to the uncus. All its nuclei have a distinct cytoarchitectonic and chemoarchitectonic organization. The amygdala has a close anatomical and functional relationship with the hippocampal formation (projections to hippocampal neurons) and together they form the two major subcortical telencephalic limbic areas. The central nucleus of the amygdala projects to a variety of brain structures via the stria terminalis and the ventral amygdalofugal pathway. One pathway is from the central nucleus to the brain

stem startle reflex circuit (nucleus reticularis pontis caudalis). Pathways from the amygdala to the lateral hypothalamus affect peripheral sympathetic responses to stress. Lesions of the central nucleus of the amygdala have been shown to completely block fear conditioning, whereas electrical stimulation of the central nucleus increases acoustic startle. Electrical stimulation of the amygdala in cats resulted in peripheral signs of autonomic hyperactivity and fear-related behaviors seen in the wild when the animal is attacked or is attacking, including alerting, chewing, salivation, piloerection, turning, facial twitching, arching of the back, hissing, and snarling, associated with an increase in catecholamine turnover. Electrical stimulation of the amygdala in human subjects resulted in signs and symptoms of fear and anxiety, including an increase in heart rate and blood pressure, increased muscle tension, subjective sensations of fear or anxiety, and increases in peripheral catecholamines. These findings demonstrate that the amygdala plays an important role in conditioned fear and emotional responding. There are also important connections between cortical association areas, thalamus, and amygdala that are important in shaping the emotional valence of the cognitive response to stressful stimuli. In addition to thalamocorticoamygdala connections, direct pathways from thalamus to amygdala, could account for fear responding below the level of consciousness.

The paradigm of conditioned fear has been utilized as an animal model for stress-induced abnormalities of emotional memory and recently further neuroanatomically differentiated. Long-term activation using diffuse cues shows that this relies more on the bed nucleus of the stria terminals, whereas with the use of explicit cues, long-term activation was processed in the central nucleus of the amygdala. This may represent a brain distinction between "real fear" and "anxiety."

The amygdala is also involved in modulating peripheral stress responses. The reciprocal projections to the bed nucleus of the stria terminalis are among the major corticotropin-releasing hormone (CRH) containing systems in the brain. On the basis of chemoarchitecture and hedology, this sometimes is referred to as "extended amygdala."

B. Locus Coeruleus

The LC is located in the dorsal pontine tegmentum. LC neurons utilize norepinephrine (NE) as their neurotransmitter and have been implicated in the control of vigilance. That is, they seem to fire just before an experimental animal begins to pay attention to the novel sensory stimuli. The LC accounts for most of the noradrenergic input to the limbic system. Manipulation of noradrenergic function alters fear and anxiey behaviors, therefore supporting the longstanding notions of the critical role of the LC in anxiety and fear. The LC shows heterogeneity of transmitter content: the catecholaminergic neurons contain noradrenaline and also peptides such as neuropeptide Y (NPY) and galanin. Stimulation of the LC results in a series of responses very similar to those observed in naturally occurring or experimentally induced fear. Drugs have been used (e.g., yohimbine and piperoxone) that activate LC by blocking α_2-adrenergic autoreceptors. Increases in LC function are accompanied by sympathetic activation; the greater the activation, the greater the correlation. Decreasing the function of the LC (interacting with inhibitory opiates, benzodiazepines, and α_2 receptors on LC) results in a decrease in fearful behavior. The LC–NE network helps determine whether, under threat, an individual turns attention toward external sensory stimuli or to internal vegetative states.

C. Hippocampus

The hippocampus plays a key role in memory function and also in the context of fear. An important aspect of the fear response is incorporation of a person's prior experience (memory) into the cognitive appraisal of stimuli. The hippocampus and adjacent cortex mediate declarative memory function (e.g., recall of facts and lists) and play a role in integration of memory elements at the time of retrieval and in assigning significance for events within space and time. The hippocampus is also involved in mediating emotional responses to the context of a stressor—for example, in animal studies timely lesions of the hippocampus disrupted the formation of emotional memories of the context (i.e., the box) where the stressor (i.e., electric foot shock) took place. With long-term storage, memories are believed to be shifted from hippocampus to the neocortical areas, where the initial sensory impressions take place. The shift in memory storage to the cortex may represent a shift from conscious representational memory to unconscious memory processes that indirectly affect behavior. "Traumatic cues" such as a particular sight or sound reminiscent of the original

traumatic event will trigger a cascade of anxiety- and fear-related symptoms, often without conscious recall of the original traumatic event. Lesions in the hippocampus have little effect on classically conditioned modality-specific fear, in contrast to the amygdala. Projections from the amygdala to the hippocampus could permit the experience of fear to modulate or influence short-term memory function, attaching cognitive significance to fear-inducing events and facilitating memory traces that enable the individual to rapidly initiate adaptive behavioral responses. Recent research has shown that the hippocampus is particularly vulnerable to stress possibly through the effects of increased levels of glucocorticoids. Glucocorticoids have also been shown to augment extracellular glutamate accumulation. Other factors besides glucocorticoids, such as brain-derived neurotrophic factor (BDNF), trkB mRNA, and nerve growth factor, which have a regulatory effect on neuronal morphology and proliferation, may mediate stress-induced alterations in hippocampal morphology. Stress-induced damage has been shown to be associated with an increase in levels of CRF mRNA in the paraventricular nucleus (PVN) of the hypothalamus as well as a decrease in the sensitivity of rats to dexamethasone suppression of hypothalamic–pituitary–adrenal (HPA) function.

D. Thalamus

The thalamic nuclei relay sensoric information from auditory, visual, and somesthetic systems to the cortex and limbic forebrain. An exception is the olfactory system, which does not relay information through the thalamus. Processing of threatening stimuli involves the relay of sensory signals to limbic forebrain directly from thalamus and cortex.

E. Hypothalamus

Many of the neuroendocrine and autonomic changes resulting from stress, fear, and anxiety can be understood from the projections that the hypothalamic nuclei receive from many limbic and brain stem structures. Their projection is to sympathetic regions in the spinal cord and medulla. The hypothalamus is important in the regulation of neuropeptide release and sympathoadrenal outflow associated with fear and anxiety. Stress results in increased production and turnover of NE in the hypothalamic nuclei. Stimulation of LC results in increased NE turnover in the PVN and supraoptic nucleus (SON), and helps PVN and SON synthesize enkephalin, vasopressin, and oxytocin—hormones whose levels are increased by stress. There is evidence that stress induces changes in cholecystokinin (CCK) and substance P, concentrated in the hypothalamus.

F. Periaqaeductal Grey

PAG is a key brain region involved in initiating fear-related behaviors. Lesions of the PAG reduce fear-related behaviors. It has been suggested that the amygdala signals the degree of threat to the PAG. Immediate threat may activate vigorous defense behaviors and nonopioid analgesia; less immediate threat or conditioned stimulus (CS) presentation predicting danger may produce freezing behavior and opioid analgesia.

G. Cortex

The cognitive response to threat involves placing the threatening object in space and time. Specific cortical brain areas are involved in these functions; for example, parietal cortex is involved in determining where an object is located in space; posterior portions of the cingulate have connections to parietal cortex, hippocampus, and adjacent cortex (important in visuospatial processing); prefrontal cortex is also involved in memory and cognition and with parietal cortex has important dual reciprocal connections with all the subcortical areas; and the dorsolateral prefrontal cortex has a range of functions, including declarative and working memory as well as planning of action, whereas the parietal cortex plays an important role in spatial memory. The medial prefrontal cortex (mPFC) and parietal cortex probably work in concert in the alerting and planning aspects of the stress response that is critical for survival. mPFC, including anterior cingulate (Brodmann area 32) and subcallosal gyrus (areas 24 and 25), is involved in selection of responses for action and emotion. This area and other medial portions of the prefrontal cortex, including orbitofrontal cortex, modulate emotional and physiological responses to stress, specifically in the effectiveness of the individual's behavior (e.g., the capacity to inhibit and change behavior in the face of threat) and are

discussed in more detail later. Lesions of the mPFC increase resistance to extinction of fear-conditioned responses. The reciprocal interactions between subcortical limbic structures and orbitofrontal cortex may point to interaction in learning and unlearning of the significance of fear-producing sensory events and the choice and implementation of behaviors important for survival. Lesions in the prefrontal cortex cause profound changes in affect, mood, and social behavior.

mPFC areas (areas 24, 25, and 32) modulate emotional responsiveness through inhibition of amygdala function. It has projections to the amygdala, which are involved in the suppression of amygdala responsiveness to fearful cues. Dysfunction of these areas can be responsible for the failure of extinction to fearful cues, which is an important part of the anxiety response. Area 25 also has direct projections to brain stem and is involved in regulation of peripheral responses to stress, including heart rate, blood pressure, and cortisol response. Lesions of this area in animals result in impairments in mounting the peripheral glucocorticoid and sympathetic response to stress. Human subjects with lesions of medial prefrontal cortical areas (e.g., the famous case of Phineas Gage) have deficits in interpretation of emotional situations that are accompanied by impairments in social relatedness. Other case studies of humans with brain lesions have implicated mPFC in "emotion" and socially appropriate interactions. Auditory association areas (temporal lobe) have also been implicated in animal studies as mediating extinction to fear responding.

VI. NEURAL CIRCUITS IN ANXIETY AND FEAR

The major afferent arm of neural circuitry includes exteroceptive sensory systems of the brain, consisting of serially organized relay channels that convey directly or through multisynaptic pathways information relevant to the experience of fear. The sensory information contained in a fear- or anxiety-inducing stimulus is transmitted from peripheral receptor cells in the eyes, ears, nose, skin, the body's own visceral information (e.g., blood glucose, arterial pressure, and CO_2 levels), or any combination of these. Except for olfactory information, which goes directly to amygdala and enthorhinal cortex, these sensory inputs are relayed through the dorsal thalamus to amygdala and cortical brain areas, such as primary visual (occipital), auditory (temporal), or tactile (postcentral gyrus) cortical areas. Input from peripheral visceral organs is relayed through the nucleus paragigantocellularis and nucleus tractus solitarius in the brain stem to LC and from there to unimodal and polymodal cortical association areas. Another projection from the brain stem to the limbic system comes from the midbrain region of the ventral tegmental area. This nuclear group contains some of the few dopaminergic neurons found in the brain. Information that reaches primary sensory areas is then processed by secondary cortical association areas, often physically adjacent to the primary sensory areas from which they receive information. These brain areas send projections to multiple brain structures, including amygdala, hippocampus, entorhinal cortex, orbitofrontal cortex, and cingulate gyrus, that are involved in mediating this visual memory and attached emotion.

As this primary sensory input comes into the brain stem and midbrain, it is matched against previously stored patterns of activation and if unknown, or if associated with previous threat, an initial fear response begins, consisting of affective, behavioral, and somatic responses. A wave of neuronal activation in key brain stem and midbrain nuclei is formed by activation of different neurotransmitters, neuromodulators, and neuropeptides, resulting in patterns of neuronal activation which move from brain stem through midbrain to thalamic, limbic, and cortical areas. At the level of the thalmus and the limbic areas (the amygdala also receives information directly from the thalamus) specific patterns of neuronal activity result in the actual sensation of anxiety. The pivotal role of the amygdala is supported by the fact that it receives afferents from thalamic and cortical exteroceptive systems as well as from subcortical visceral afferent pathways. At the subcortical and cortical level, more complex cognitive associations are made, allowing for interpretation of that internal state of anxiety (e.g., the neuronal interactions between amygdala and orbitofrontal cortex enable the individual to initiate adaptive behaviors to threat based on prior experience and the nature of the threat). Cortical, limbic, midbrain, and brain stem-based neuronal activity may subsequently be involved in various aspects of anxiety regulation or dysregulation.

VII. NEUROENDOCRINE, AUTONOMIC, AND MOTOR RESPONSES

Efferent projections from the brain structures need to be able to mediate an individual's neuroendocrine,

autonomic, and motor response to threat. They also need to account for the pathological reactions that result in anxiety-related signs and symptoms. The organism must rapidly effect peripheral responses to threat, which are mediated by the sympathetic and parasympathetic systems. Structures involved in these responses include the amygdala, LC, hypothalamus, PAG, and striatum. The hypothalamus coordinates the information it receives from various brain regions to patterns of sympathetic responses. Stimulation of the lateral hypothalamus results in sympathetic system activation with increases in blood pressure and heart rate, sweating, piloerection, and pupil dilatation. Activation of the paraventricular nucleus of the hypothalamus promotes release of a variety of hormones and peptides. Its activation in anxiety and fear is thought to rely on stimulation from hypothalamus via projections from the amygdala and LC. The PAG also serves to regulate the sympathetic function, which accounts for parallel activation of the peripheral sympathetic system and the LC. The n. vagus and splanchnic nerves are projections from the parasympathetic system. The n. vagus receives information from the lateral hypothalamus, PVN, LC, and amygdala. The splanchnic nerves receive afferent input from LC. This innervation from the parasympathetic system relates to visceral symptoms associated with fear and anxiety such as gastrointestinal and genitourinary disturbances.

The amygdala has strong projections to most areas of the striatum, including nucleus accumbens, olfactory tubercle, and part of the caudate and putamen. The dense innervation of the striatum and prefrontal cortex by the amygdala indicates that the amygdala can powerfully regulate both these systems. Adaptive mobilization of the motor response to threat probably involves pathways between cortical association areas and the striatum and also the amygdala and the striatum. These interactions between amygdala and the extrapyramidal motor system may be important in generating motor responses to threatening stimuli, especially when they are related to prior experiences.

The cerebellum has a well-known role in motor movement, suggesting that this region is involved in planning for action; however, recent imaging studies are consistent with a role in cognition as well. Connections between parietal and prefrontal cortex are required in order to permit the organism to rapidly and efficiently execute motor responses to threat. These areas have important innervations to the precentral (motor) cortex, which is responsible for motor responses to threat. The striatum (caudate and putamen) modulates motor responses to stress. The dense innervation of the striatum and prefrontal cortex by the amygdala indicates that the amygdala can regulate both these systems. These interactions between the amygdala and the extrapyramidal motor system may be very important for generating motor responses to threatening stimuli, especially those related to prior adverse experiences.

Thus, preparation for responding to threat requires integration between brain areas involved in assessing and interpreting the potentially threatening stimulus. Brain areas involved in response (e.g., prefrontal cortex and anterior cingulate) play an important role in the planning of action and in holding multiple pieces of information in "working memory" during the execution of a response. Parietal cortex and posterior cingulate are involved in visuospatial processing that is an important component of the stress response. Motor cortex may represent the neural substrate of planning for action.

A. HPA Axis

The HPA axis is a component of the response system in fear and anxiety. The HPA axis has important functional interactions with the NE system that facilitate a sophisticated range of responses to stress. Stimulation of the lateral hypothalamus results in sympathetic system activation producing increases in blood pressure and heart rate, sweating, piloerection, and pupil dilatation. Stress stimulates release of CRF from the PVN of the hypothalamus into the portal bloodstream. CRF is transported to the anterior lobe of the pituitary, where it stimulates the release of ACTH and ultimately the release of cortisol from the adrenal cortex. High levels of circulating cortisol act through a negative feedback pathway to decrease both CRF and NE synthesis at the level of the PVN. Glucocorticoid inhibition of NE-induced CRF stimulation may be evident primarily during stressor-induced cortisol release and not under resting conditions. High levels of cortisol likely inhibit the effects of NE on CRF release from the PVN, serving to restrain the stress-induced neuroendocrine and cardiovascular effects mediated by the PVN. CRF increases activity of the LC; CRF injected into the LC intensifies anxiety-related responses. CRF serves as an excitatory neurotransmitter in the LC, which contributes to a pathway for the behavioral effects of CRF.

The mechanism responsible for transient stress-induced hyperadrenocorticism and feedback resistance

may involve a downregulation of glucocorticoid receptors. High glucocorticoid levels decrease the number of hippocampal glucocorticoid receptors, resulting in increased corticosterone secretion and feedback resistance. Following stress termination, when glucocorticoid levels decrease, receptor numbers are increased and feedback sensitivity normalizes.

The effects of chronic stress on ACTH and corticosterone secretion vary depending on the experimental paradigm. In nonhuman primates, adverse early experiences induced by variable maternal foraging requirements result in profound behavioral disturbances (more timid, less social, and more subordinate) years later. Adult monkeys raised in the variable foraging maternal environment were also hyperresponsive to yohimbine and had elevated levels of CSF and decreased CSF cortisol levels in adulthood. These observations suggest that early adverse experience permanently affects the HPA axis.

B. Norepinephrine

Norepinephrine release in the brain represents an important part of the immediate stress response, reacting within seconds. The majority of noradrenergic cell bodies are located in the brain stem, in the LC, with axons that extend throughout the cerebral cortex and to multiple subcortical areas. Neurons in the LC are activated in association with fear and anxiety states and the limbic and cortical regions innervated by the LC are thought to be involved in the elaboration of adaptive responses to stress. Stressors such as a cat seeing a dog result in an increase in firing of neurons in the LC and enhanced NE release in the hippocampus and mPFC. Exposure to chronic stress also results in a potentiation of NE release with subsequent stressors. Consistent with these findings, noradrenergic stimulation resulted in decreased metabolism in hippocampus (consistent with high levels of NE release) and relative failure of activation in mPFC in PTSD patients but not normal subjects. There was also a pattern of a relationship between this metabolic response and increased panic anxiety.

C. Dopamine

The dopamine innervation of the mPFC appears to be particularly vulnerable to stress. Sufficiently low-intensity stress (such as that associated with conditioned fear) or brief exposure to stress increases dopamine release and metabolism in the prefrontal cortex in the absence of overt changes in other mesotelencephalic dopamine regions. Low-intensity electric foot shock increases *in vivo* tyrosine hydroxylase and dopamine turnover in the mPFC but not the nucleus accumbens or striatum. Stress can enhance dopamine release and metabolism in other areas receiving dopamine innervation, provided that greater intensity or longer duration stress is used. Thus, the mPFC dopamine innervation is preferentially activated by stress compared to mesolimbic and nigrostriatal systems, and the mesolimbic dopamine innervation appears to be more sensitive to stress than the striatal dopamine innervation.

D. Serotonin

Although less studied, in conditioned fear experiments animals demonstrated an increase in serotonin (5-HT) turnover in different brain areas, with preferential release in mPFC. Serotonin antagonists produce behavioral deficits resembling those seen following inescapable shock. Chronic stress increases cortical 5-HT_2 receptor binding and reduces hippocampal 5-HT_{1A} receptor binding. Drugs that enhance serotonin neurotransmission are effective in reversing this behavioral "learned helplessness." Injection of serotonin into the frontal cortex after stress exposure reverses behavioral deficits.

The effect of stress in activating serotonin turnover may stimulate a system that has both anxiogenic and anxiolytic pathways within the forebrain. A primary distinction in the qualitative effects of serotonin may be between the dorsal and median raphe nuclei, the two midbrain nuclei that produce most of the forebrain serotonin. The serotonergic innervation of the amygdala and the hippocampus by the dorsal raphe are believed to mediate anxiogenic effects via 5-HT_2 receptors. In contrast, the median raphe innervation of hippocampal 5-HT_{1A} receptors has been hypothesized to facilitate the disconnection of previously learned associations with aversive events or to suppress formation of new associations, thus providing resilience to aversive events.

E. Benzodiazepine

Alterations in Bz receptor function are involved in the stress response and anxiety. Animals exposed to stress

develop a decrease in Bz receptor binding in different brain sites. Decreases in Bz receptor binding are also associated with alterations in memory. Bz receptors are present throughout the brain, with the highest concentration in cortical grey matter. Bzs potentiate and prolong the synaptic actions of the inhibitory neurotransmitter γ-aminobutyric acid (GABA). Central Bz receptors and GABA receptors are part of the same macromolecular complex. These receptors have distinct binding sites, although they are functionally coupled and regulate each other in an allosteric manner. Administration of inverse agonists of Bz receptors results in behavioral and biological effects similar to those seen in anxiety and stress. These effects are blocked by administration of Bzs or pretreatment with the Bz antagonist flumazenil.

F. Neuropeptides

Several neuropeptides also mediate the response to stress. CCK is an anxiogenic neuropeptide present in the gastrointestinal tract and the brain that has recently been suggested to be a neural substrate for human anxiety.

Stress is associated with an increase in endogenous opiate release with decreased density of mu opiate receptors, which may mediate the analgesia associated with stress. Other neuropeptides under investigation that appear to play a role in the stress response are neuropeptide Y, somatostatin, and thyrotropin. Stress also has important effects on the immune system.

VIII. APPLICATION OF THE MODEL OF THE NEURAL CIRCUITRY OF ANXIETY AND FEAR TO ANXIETY DISORDERS

The primary goal of research in the clinical neuroscience of anxiety disorders is to apply findings related to the effects of stress on the brain in animals to patients with anxiety disorders. Different methods contributed to the working model of the neural circuitry of anxiety and anxiety disorders that is presented here. The neural circuits mediating symptoms of anxiety disorders can be studied by measuring neurotransmitters and hormone levels in blood, urine, and saliva; by assessing behavioral and biochemical responses to pharmacological challenge to specific neurochemical systems; by measuring key brain structures with structural neuroimaging; by provoking disease-specific symptoms in conjunction with functional neuroimaging; or by using imaging to measure neuroreceptors.

Among the most characteristic features of anxiety disorders such as PTSD and PD is that "anxiogenic" memories (e.g., of the traumatic experience or first panic attack) can remain indelible for years or decades and can be easily reawakened by all sorts of stimuli and stressors. The strength of traumatic memories relates, in part, to the degree to which certain neuromodulatory systems, particularly catecholamines and glucocorticoids, are activated by the traumatic experience. Release of these stress hormones modulates the encoding of memories of the stressful event. Experimental and clinical investigations provide evidence that memory processes remain susceptible to modulating influences after information has been acquired. Long-term alterations in these catecholaminergic and glucocorticoid systems may also be responsible for symptoms of fragmentation of memories, but also for hypermnesia, amnesia, deficits in declarative memory, delayed recall of abuse, and other aspects of the wide range of memory distortions in anxiety disorders. With long-term storage, memories are shifted from hippocampus to the neocortical areas, where the initial sensory impressions take place. The shift in memory storage to the cortex may represent a shift from conscious representational memory to unconscious memory processes that indirectly affect behavior. Traumatic cues such as a particular sight or sound reminiscent of the original traumatic event will trigger a cascade of anxiety and fear-related symptoms will ensue, often without conscious recall of the original traumatic event. In patients with PTSD, however, the traumatic stimulus is always potentially identifiable. Symptoms of anxiety in panic or phobic disorder patients, however, may be related to fear responses to a traumatic cue (in individuals who are vulnerable to increased fear responsiveness, through either constitution or previous experience), where there is no possibility that the original fear-inducing stimulus will ever be identified.

Thus, patients with anxiety disorders have symptoms that reflect a more or less continuous perception of threat with unconscious processed fear responses. The animal model of contextual fear conditioning represents a good model for these symptoms. Preclinical data suggest that the hippocampus (as well as BNST and PAG) plays an important role in the mediation of contextual fear, and that increased responding to CS is due to hippocampal dysfunction. Hippocampal atrophy in PTSD, as described

previously, therefore provides a possible neuroanatomical substrate for abnormal contextual fear conditioning and chronic symptoms of feeling of threat and anxiety. Interestingly, in light of studies showing abnormal noradrenergic function in PTSD, the BNST has some of the densest noradrenergic innervation of any area in the brain.

The startle reflex has been the subject of the few fear conditioning studies that have been performed in humans. Startle is a useful method for examining fear responding in experimental studies involving both animals and humans that is mediated by the amygdala and connected structures. Patients with combat-related PTSD were found to have elevated baseline startle compared to controls in some studies but not others. In the patient group, there were asymmetry of baseline startle response and increased heart rate responses during measurement of startle. From other studies it is clear that unconscious emotional processes are involved in fear conditioning (e.g., patients with anxiety disorders have demonstrated greater resistance to extinction of conditioned responses to angry facial expressions, but not to neutral facial expressions, compared to controls). In the neural circuitry, damage to the amygdala does not prevent patients from learning the relationship between the CS and the unconditioned stimulus (UCS), but it abolishes conditioned autonomic responses. In contrast, damage to the hippocampus does not affect conditioned autonomic responses but does prevent learning of the CS–UCS association. Further evidence for unconscious processes stems from backward masking techniques, which prevent conscious awareness of a stimulus. Using such a technique, fear conditioning to a certain class of stimuli called fear-relevant stimuli (e.g., spiders and snakes) proves to be mediated by preattentive automatic information-processing mechanisms. These automatic mechanisms may be mediated in part by direct thalamic–amygdala connections. Thalamoamygdala pathways that bypass the cerebral cortex may trigger conditioned responses before the stimulus reaches full awareness, providing an explanation for unconscious conditioned phobic responses to fear-relevant stimuli.

A. Alterations in Neurochemical Stress Response Systems in Patients with Anxiety Disorders

1. CRF/HPA Axis

Anxiety disorder patients have long-term alterations in neurochemical systems that are involved in mediating the stress response and are sensitive to chronic stress. The findings in PTSD and PD (most extensively studied of the anxiety disorders) are summarized in Table II.

Specific alterations in cortisol and HPA axis function are associated with PTSD. An increase in neuronal CRF release is a possible explanation for the clinical findings that have been reported for this disorder, with resultant blunting of ACTH response to CRF, increased central glucocorticoid receptor responsiveness, and resultant low levels of peripheral cortisol due to enhanced negative feedback. Interestingly, nonhuman primates with variable foraging mothers (a model for early life stress) had elevated CSF and CRF and decreased CSF cortisol levels in adulthood, which is more similar to PTSD than to depression.

Evidence for dysfunction of CRF or HPA systems in PD has been inconsistent. Normal levels of CRF in CSF of PD patients have been found, blunted ACTH responses to CRF (indicating chronic elevations in CRF) have been reported in some studies, and both normal and elevated rates of cortisol nonsuppression following dexamethasone have been reported. Urinary-free cortisol results have been inconsistent. Elevated plasma cortisol levels were reported in one study but not in others. In 24-hr secretion of ACTH and cortisol in PD, only subtle abnormalities were seen. Patients had elevated overnight cortisol secretion and greater amplitude of ultradian secretory episodes.

2. Catecholamines

There is extensive evidence indicating that NE plays a role in human anxiety and is dysregulated in anxiety disorders. PTSD and PD seem to have similar alterations in noradrenergic function. However, the causes of the two syndromes may differ, with PD associated more with genetic factors and PTSD with the effects of severe psychological trauma.

Heightened autonomic or sympathetic nervous system arousal was found in children but also in combat veterans with chronic PTSD. They showed higher resting mean heart rate and systolic blood pressure, as well as greater increases in heart rate, when exposed to visual and auditory combat-related stimuli compared with combat veterans without PTSD, patients with GAD, or healthy subjects. The PTSD patients showed hyperreactive responses to combat-associated stimuli but not to other stressful noncombat-related stimuli. They also showed elevated NE and epinephrine in 24-hr urine in comparison to normal

Table II

Evidence for Neurobiological Alterations in CRF/HPA Axis, Catecholaminergic Systems, and Other Neurotransmitter Systems[a]

	PTSD	Panic
CRF/HPA axis function		
Alterations in urinary cortisol	+/−[a]	+/−
Altered plasma cortisol with 24-hr sampling	+ (decrease)	+ (increase)
Supersuppression with DST	+	−
Blunted ACTH response to CRF	+ +	+/−
Elevated CRF in CSF	+	+
Increased lymphocyte glucocorticoid receptors	+ +	NT
Catecholaminergic function		
Increased resting heart rate and blood pressure	+/−	+/−
Increased heart rate and blood pressure response to traumatic reminders/panic attacks	+ + +	+ +
Increased resting urinary NA and Epi	+	+ +/−
Increased resting plasma NA or MHPG	−	−
Increased plasma NA with traumatic reminders/panic attacks	+	+/−
Increased orthostatic heart rate response to exercise	+	+
Decreased binding to platelet-2 receptors	+	+/−
Decrease in basal and stimulated activity of cAMP	+/−	+
Decrease in platelet MAO activity	+	?
Increased symptoms, heart rate, and plasma MHPG with yohimbine noradrenergic challenge	+	+ + +
Differential brain metabolic response to yohimbine	+	+
Other neurotransmitter systems		
Serotonin		
Decreased serotonin reuptake site binding in platelets	+ +	
Decreased serotonin transmitter in platelets	−	
Blunted prolactin response to buspirone ($5HT_{1A}$ probe)	−	
Benzodiazepine		
Increased symptomatology with Bz antagonist	−	+ +
Cholecystokinin		
Increased anxiety symptoms with CCK administration	NT	+ +
Anxiolytic effect of CCK antagonist	−	
Opiate		
Naloxone-reversible analgesia	+	
Increased plasma β-endorphin response to exercise	+	
Increased endogenous opiates in CSF	+	
Altered serotonin effect on cAMP in platelets ($5HT_{1A}$ probe)	−	
Neuropeptide Y		
Altered plasma levels	+ (decrease)	+ (increase)
Thyroid		
Increased baseline thyroxine	+	
Increased TSH response to TRH	+	
Somatostatin		
Increased somatostatin levels at baseline in CSF	+	

[a]Findings of decreased urinary cortisol in older male combat veterans and holocaust survivors and increased cortisol in younger female abuse survivors may be explainable by differences in gender, age, trauma type, or developmental epoch at the time of the trauma. + indicates the availability of studies; NT = not tested.

controls and patients with other psychiatric disorders. Relative elevations of the NE metabolite, MHPG, were found in their nighttime samples. No differences were found in baseline levels of plasma NE when compared with healthy subjects. This noradrenergic hyperreactivity in patients with PTSD may be associated with the conditioned or sensitized responses to specific traumatic stimuli. Women with PTSD secondary to childhood sexual abuse had significantly elevated levels of catecholamines (NE, E, and DA), cortisol and catecholamine metabolites, metanephrine, vanillylmandelic acid, and HVA in 24-hr urine samples.

The release of glucocorticoids or other stress-related factors (e.g., stress-induced decreases in BDNF) is shown to result in hippocampal damage with lasting deficits in verbal declarative memory dysfunction in PTSD. Although hippocampal volume reduction appears to be specific to stress-related anxiety disorders, PD patients have been shown to have alterations of parahippocampal gyrus and other portions of extrahippocampal temporal lobe that may underlie declarative memory deficits that are also seen in PD. Increased cortisol release with stress in both PTSD and PD may result in amnesia and cognitive dysfunction associated with these disorders. Excessive release of NE with stressors in anxiety disorder patients is predicted to result in decreased function of neurons which may be related to both cognitive dysfunction and increased anxiety with stress. In addition, given the known role of the hippocampus in contextual fear, lasting damage to the hippocampus may contribute to excessive anxiety and fear responding in these patients. Finally, since the hippocampus is involved in integration of individual aspects of memory at the time of memory retrieval, hippocampal dysfunction may lead to memory fragmentation and amnesia.

Evidence for noradrenergic responsivity is also manifested on the level of peripheral NE receptor function. A decrease in platelet adrenergic α_2 receptor number as measured by total binding sites for the α_2-antagonist [^{3}H]rauwolscine and greater reduction in number of platelet α_2 receptors after exposure to agonist (epinephrine) may reflect chronic high levels of NE release which lead to compensatory receptor downregulation and decreased responsiveness. Although there is inconsistent evidence for elevations in NE at baseline in PTSD, there is evidence for increased noradrenergic responsivity in this disorder. Studies have also shown alterations in the α_2-receptor and cyclic AMP function in patients with PTSD, which were similar to those in PD. Patients with combat-related PTSD compared to healthy controls had enhanced behavioral, biochemical (NE metabolite MHPG), and cardiovascular (heart rate and blood pressure) responses to the α_2 antagonist, yohimbine, which stimulates central NE release. Moreover, neuroimaging with positron emission tomography (PET) demonstrated that PTSD patients have a cerebral metabolic response to yohimbine, consistent with increased NE release.

Alterations in sleep function may be secondary to altered pontine function and noradrenergic dysregulation in PTSD. Sleep dysfunction has been documented following acute fear and appears to be related to development of chronic anxiety disorders. PTSD patients have been found to have decreased total sleep time, increased "microawakenings," and increases in phasic rapid eye movement activity relative to controls. These abnormalities may play a role in nightmares in this patient population.

Abnormal regulation of brain noradrenergic systems is also involved in the pathophysiology of PD. Evidence for baseline elevations in heart rate and blood pressure, as well as plasma NE, in PD has been conflicting. PD patients had increased heart rate responses to orthostatic challenge when compared to normal subjects. Patients with PD were found to have elevations in baseline plasma epinephrine. One study found increased NE with lactate-induced panic attacks but another did not find an increase with spontaneous panic attacks. PD patients are very sensitive to the anxiogenic effects of yohimbine, in addition to having exaggerated plasma MHPG, cortisol, and cardiovascular responses. Responses to the α_2-adrenergic receptor agonist, clonidine, were also abnormal in PD patients. Clonidine administration caused greater hypotension, greater decreases in plasma MHPG, and less sedation in panic patients than in controls. Clonidine significantly, but only partially, attenuated lactate panic symptoms, indicating that noradrenergic neuronal activity cannot fully account for lactate-induced panicogenesis. Growth hormone response to clonidine (used as a probe of central α_2-adrenergic receptor function) was blunted in PD versus normals. These findings persist despite clinical recovery with fluoxetine treatment.

Few studies have examined noradrenergic function in patients with phobic disorders. In patients with specific phobias, increases in subjective anxiety, and increased heart rate, blood pressure, plasma NE, and epinephrine have been associated with exposure to the phobic stimulus. This finding may be of interest from the standpoint of the model of conditioned fear, in

which a potentiated release of NE occurs in response to a reexposure to the original stressful stimulus. Patients with social phobia have been found to have greater increases in plasma NE in comparison to healthy controls and patients with PD. In contrast to PD patients, the density of lymphocyte α adrenoceptors is normal in social phobic patients. The growth hormone response to intravenous clonidine (a marker of central α_2 receptor function) is blunted in social phobia patients.

There is evidence for a role for dopamine in human anxiety states. A higher incidence of anxiety disorders including panic attacks occurs in patients with Parkinson's disease. A higher concentration of the dopamine metabolite, homovanillic acid, in plasma in PD has been found in patients with high levels of anxiety and frequent panic attacks. Recently, patients with PD were shown to have a greater growth hormone response to the dopamine agonist apomorphine than patients with depression. However, some negative studies also exist. As noted previously, elevations in dopamine were found in 24-hr urine of patients with PTSD. There have been no studies of dopamine function in social phobia.

3. Other Neurotransmitter Alterations

a. Serotonin Clinical studies of serotonin function in anxiety disorders have had mixed results. Although there are only a limited number of studies of serotonergic function in PTSD, there is a large body of indirect evidence suggesting that this neurotransmitter may be important in the pathophysiology of trauma-related symptomatology. In humans, low 5-HT functioning has been associated with aggression, impulsivity, and suicidal behavior. Patients with PTSD are frequently described as aggressive or impulsive and often suffer from depression, suicidal tendencies, and intrusive thoughts that have been likened to obsessions. Patients with PTSD were found to have a predominance of "reducing" patterns to auditory-evoked potential paradigms compared to normal controls. One explanation for these findings is low serotonergic activity in PTSD. This hypothesis gains further support from the observation that serotonin reuptake inhibitors have been found to be partially effective in treating PTSD symptoms such as intrusive memories and avoidance symptoms. A subgroup of patients with PTSD exhibited a marked behavioral sensitivity to serotonergic provocation with *m*-chlorophenylpiperazine (mCPP), a probe of postsynaptic 5-HT_{1C} and 5-HT_2 receptor functions, raising the possibility of pathophysiologic subtypes among traumatized combat veterans.

To date, pharmacologic challenge studies of 5-HT in PD have also been unable to establish a definite role for 5-HT in the pathophysiology of panic. Challenges with the 5-HT precursors, L-tryptophan and 5-hydroxytryptophan (5-HTP), did not discriminate between PD and controls on neuroendocrine measures. Conversely, tryptophan depletion was not anxiogenic in unmedicated PD patients. However, challenge with the 5-HT-releasing agent fenfluramine has been reported to be anxiogenic and to produce greater increases in plasma prolactin and cortisol in PD compared to controls. Studies with mCPP have produced equivocal findings. Increases in anxiety and plasma cortisol in PD patients compared to controls have been reported with oral but not IV administration of mCPP. When the selective 5-HT_{1A} partial agonist ipsapirone was used as a challenge agent, ACTH, cortisol, and hypothermic responses to ipsapirone were blunted in PD patients, but anxiety responses did not differ from controls. These data implicate 5-HT_{1A} receptor subsensitivity in the pathophysiology of PD.

In several case reports the synthetic amphetamine analog MDMA has been associated with panic attacks and in some cases with the induction of PD. MDMA acts as both an indirect dopamine agonist and a potent 5-HT releaser and agonist.

b. Benzodiazepine Despite the preclinical support for the involvement of Bz systems in stress, clinical investigations of the function of this system in patients with anxiety disorders have been difficult to design. The inability to identify variables measurable in living human subjects who reflect central Bz system function has contributed to the paucity of research in this area. However, evidence from clinical studies performed to date suggests a possible role for alterations in Bz receptor function in disorders of anxiety and stress.

Pharmacologic challenge studies with the Bz receptor inverse agonist FG7142 induced severe anxiety resembling panic attacks and biological characteristics of anxiety in healthy subjects. Both oral and intravenous administration of the Bz receptor antagonist flumazenil to patients with PD resulted in an increase in panic attacks and subjective anxiety in comparison to controls. Flumazenil, however, had no anxiogenic effects in PTSD. Bz-induced changes in sedation and cortisol levels, as well as saccadic eye movement velocity, have been suggested to be indicative of Bz receptor-mediated actions. PD patients were found to

be less sensitive than controls to diazepam using saccadic eye movement velocity as a dependent measure, suggesting a functional subsensitivity of the GABA–Bz supramolecular complex in brain stem regions controlling saccadic eye movements. Other evidence for alterations in Bz receptor function in PD patients includes a diminished sensitivity to suppression of plasma NE, epinephrine, and pulse following administration of diazepam in comparison to controls.

Patients with GAD have been found to have decreases in peripheral-type Bz receptor binding as assessed by [^{3}H]PK11195 binding to lymphocyte membranes, although the relationship to central Bz receptor function is unclear. In addition, reduced binding to [^{3}H]PK11195 has been reversed with Bz therapy in patients with anxiety disorders. Healthy subjects exposed to the stress of war have also been found to have a decrease in binding of Bz receptors on peripheral mitochondria during the stressful period before and during war relative to the period after the war, which was correlated with an improvement of anxiety after the end of the war. Mixed findings are reported on peripheral Bz receptor binding in PD patients. Studies examining GABA, which is functionally related to Bzs, have not found differences in plasma levels between PD patients and controls at baseline.

Studies have begun to use neuroimaging to examine central Bz receptor function in anxiety disorders. [^{123}I]Iomazenil binds with high affinity to the Bz receptor and can be used with single photon emission computed tomography (SPECT) to image the Bz receptor in human brain. SPECT has been used to show a decrease in [^{123}I]iomazenil uptake in frontal, temporal, and occipital cortex in PD patients relative to comparison subjects. These studies were limited by the use of nonquantitative methods for estimation of Bz receptor binding, the absence of medication-free subjects, psychiatric comorbidity within PD patients, and/or the use of diseased patients for comparison groups. SPECT iomazenil quantitation of Bz receptor binding, comparing volumes of distribution of receptor (related to the ratio of receptor number and affinity) in patients with PD and controls resulted in a reduction in Bz receptor binding in left hippocampus and precuneus in PD. Elevated levels of anxiety were correlated with decreased binding in the frontal cortex. When PET and [^{11}I]flumazenil were used in the measurement of Bz receptor binding, global decreases in PD patients were found, with the greatest magnitude in the right orbitofrontal cortex and insula. Recently, a decrease was found in Bz receptor binding in prefrontal cortex (area 9) in patients with combat-related PTSD compared to matched healthy controls. These findings were consistent with animal studies of stress showing decreased binding in the frontal lobe and hippocampus.

c. Cholecystokinin The neuropeptide CCK has been shown to be anxiogenic in human subjects. In healthy volunteers intravenous administration of CCK4 (a tetrapeptide that crosses the blood–brain barrier more readily than CCK8) has been shown to induce severe anxiety or short-lived panic attacks. The anxiogenic effect of CCK was blocked by the Bz lorazepam, although this may merely be pharmacological opposition and not true antagonism. PD patients were found to be more sensitive to the anxiogenic effects of CCK4 and a closely related peptide, pentagastrin. These effects were blocked by CCK antagonists. Imipramine antagonizes the panicogenic effects of CCK4 in PD patients. The mechanism is unclear but may relate to the ability of imipramine to downregulate β-adrenergic receptors since propranolol antagonizes the anxiogenic actions of CCK4. Levels of CCK in the CSF are lower in PD patients than in normal subjects, indicating the possibility of enhanced function of CCK receptors. The mechanism responsible for the enhanced sensitivity to CCK4 has not been elucidated. Patients may have an elevated production or turnover of CCK or increased sensitivity of CCK receptors. Since CCK has important functional interactions with other systems implicated in anxiety and fear (noradrenergic, dopaminergic, and Bz), these interactions need to be evaluated in PD patients. CCKB receptor antagonists are now being tested as antipanic drugs.

d. Opiates Only a few studies have examined opiate function in PTSD. Some studies report lower AM and PM plasma β-endorphin levels. In studies that found no differences in plasma levels of methionine enkephalin, degradation half-life was significantly higher in the PTSD group. In a pharmacological challenge of the opiate system, PTSD patients showed reduced pain sensitivity compared to veterans without PTSD following exposure to a combat film. This was reversible by the opiate antagonist naloxone and could be explained by increased release of endogenous opiates with stress in PTSD. Levels of endogenous opiates in cerebrospinal fluid were found to be elevated in combat-related PTSD. Symptoms of avoidance and numbing are thought to be related to a dysregulation of opioid systems in PTSD. The use of opiates in chronic PTSD may also represent a form of

self-medication. Animal studies have shown that opiates, by acting to decrease firing of LC neurons, are powerful suppressants of central and peripheral noradrenergic activity. If, as suggested earlier, some PTSD symptomatology is mediated by noradrenergic hyperactivity, then opiates may serve to "treat" or dampen down that hypersensitivity and accompanying symptoms. On the other hand, during opiate withdrawal when opiates are decreased and noradrenergic activity is increased, PTSD symptoms may become acutely exacerbated. In fact, many symptoms of PTSD are similar to those seen during opiate withdrawal.

e. Neuropeptide Y Low doses of NPY administered intraventricularly have proven to have anxiolytic effects in several animal models of anxiety. Disturbed NPY transmission might have a role in symptoms of anxiety. In PD elevated plasma levels of NPY have been found. In PTSD, lower levels were found.

f. Thyroid In the early part of the twentieth century, Graves first described cases of hyperactivity of the thyroid gland, with neuropsychiatric symptoms of anxiety, palpitations, breathing difficulties, and rapid heart rate, in a series of cases of individuals who were recently exposed to traumatic stress. Although the relationship between stress, neuropsychiatric symptoms, and thyroid disease has continued to be clinically observed, to date there have been no systematic epidemiologic studies of the relationship between stress and thyroid disease. Thyroid hormone has a range of actions, including energy utilization within the cell (important in stress), and stress results in long-lived elevations in thyroid hormone. Elevated levels of T3 were reported in patients with combat-related PTSD.

B. Panicogenic Effects of Lactate

In the late 1960s an observation was made by Pitts and McClure that intravenous infusion of lactate produced panic anxiety in susceptible individuals but not in normal subjects. Subsequently, the reliability of panic provocation by sodium lactate has been well established. Lactate response appeared to be specific for PD compared with other anxiety disorders and psychiatric conditions. Moreover, treatment of panic with imipramine blocked the effects of lactate.

However, the panicogenic mechanism of lactate has not been established. One theory is based on the fact that systemic alkalosis caused vasoconstriction of cerebral vessels, which in turn induced cerebral ischemia, with an increase in the intracellular lactate:-pyruvate ratio. Furthermore, infused lactate resulted in a rapid passive elevation in the lactate:pyruvate ratio in localized brain regions outside the blood–brain barrier, such as the chemoreceptor zones. These two mechanisms lowered the intracellular pH in medullary chemoreceptors. In PD patients there is dysregulation (greater sensitivity to alterations in pH) in this region; thus, a panic response is triggered. This theory predicts that panic could be triggered in any subject if medullary pH was changed sufficiently.

The limitations of the model include the fact that it is not yet known whether the pH changes in the local circulation are mirrored intracellularly. Physiological effects of sodium bicarbonate have revealed a paradoxical intracellular acidosis, so the same may be true of lactate. There has been no clear evidence that intracellular acidosis initiates neural activity, as the theory requires. Second, the model predicts that hypoxia is a profound stimulus for chemoreceptor stimulation, and hyperventilation is belied by experiments in which removal of CO_2 from inspired air leads to loss of consciousness without anxiety or air hunger.

A second major hypothesis is that lactate's panicogenic effect occurs via the induction of a metabolic alkalosis. Infused lactate is metabolized to bicarbonate that is further metabolized to CO_2, which quickly permeates the central nervous system. This central buildup of CO_2 increases the ventilatory rate via a direct stimulation of ventral medullary chemoreceptors. Increasing brain pCO_2 concentration has been shown to be a profound stimulus for LC activation, which could cause panic via central noradrenergic activation.

Although this lactate–CO_2 theory has considerable appeal, initial studies with the isomer D-lactate indicate that this may not be the whole explanation. This isomer is found to also be panicogenic but is not metabolized to CO_2. Comparisons of the behavioral effects of lactate and bicarbonate infusion demonstrate provocation of panic in susceptible patients; however, bicarbonate is less anxiogenic than lactate. This finding argues against alkalosis alone being the panicogenic stimulus. Stimulation of respiratory centers to produce increased ventilation, hypocapnia, and respiratory alkalosis was the common factor in producing panic by both infusions.

Panic can also be provoked by increases in pCO_2 (hypercapnia). This can be done slowly, such as by rebreathing air or by breathing 5–7% CO_2 in air.

Alternatively, panic attacks can be provoked by breathing only one or two deep breaths of 35% CO_2. Hyperventilation and increased CO_2 hypersensitivity have also been posited as an explanation for symptoms of PD. According to the model, elevated levels of pCO_2 lead to activation of the vagus nerve, which through the nucleus tractus solitarius stimulated the LC as well as hyperventilation. Increased tidal volume drove down pCO_2, with increased respiratory alkalosis, and symptoms of panic. Hyperactive chemoreceptors lead to hyperventilation in order to keep down pCO_2, which results in panic symptomatology. A corollary model to the hyperventilation hypothesis stated that PD patients suffered from a physiologic misinterpretation of a suffocation monitor, which evoked a suffocation alarm system. This produces sudden respiratory distress, followed quickly by hyperventilation, panic, and an urge to flee. This model posited that hypersensitivity to CO_2 is due to the deranged suffocation alarm systems. The central problem with these models is that it is difficult to determine whether hyperventilation is a primary or secondary phenomenon in PD. Also, there is no evidence of hyperventilation at baseline in PD patients. There are many potential panicogens in PD, and there is no evidence that hyperventilation is more robust than other agents such as noradrenergic stimulation with yohimbine.

C. Contributions from Neuroimaging Studies

For an overview of findings in imaging studies see Table III.

1. Structural Neuroimaging

Digital processing of magnetic resonance imaging (MRI) of the brain can quantify three-dimensional volumes of brain structures. Nonquantitive studies found evidence of abnormalities in temporal lobe in patients with PD, and one quantitative study found decreased volume in the hippocampus. A reduction in hippocampal volume has been reported in several studies in patients with PTSD (Fig. 1).

This hippocampal volume reduction has been positively correlated with verbal memory deficits. Considering the role the hippocampus plays in the integration of memory elements, which are stored in primary sensory and secondary association cortical areas at the time of retrieval, these findings suggest a possible neural correlate for symptoms of memory fragmentation and dysfunction in PTSD.

2. Functional Neuroimaging

Functional brain imaging studies have examined brain metabolism and blood flow at rest and during stress/symptom provocation in patients with anxiety disorders. Patterns of regional blood flow that are evoked reflect the engagement of neural structures in fear and anxiety responses. PET has been used to measure the closely related processes of brain metabolism (using radiolabeled glucose or [^{18}F]2-fluoro-2-deoxyglucose) and blood flow (with radiolabeled water or [^{15}O]H_2O), whereas SPECT was used to measure brain blood flow (with [^{99m}Tc]HMPAO). An increase in the function of neurons in a specific area is reflected by an increase in metabolism and a shunting of blood flow toward the area that can be measured with these imaging techniques.

Pharmacologic and cognitive provocation of PTSD symptom has been used in order to identify neural correlates of PTSD symptomatology and of traumatic remembrance in PTSD. Administration of yohimbine (which increases NE release) resulted in increased PTSD symptoms and anxiety in the PTSD group. NE had a U-shaped curve type of effect on brain function, with lower levels of release causing an increase in metabolism and very high levels of release actually causing a decrease in metabolism. Yohimbine causing a relative decrease in metabolism in patients with PTSD in orbitofrontal, temporal, parietal, and prefrontal cortex—all brain areas which receive noradrenergic innervation. Normal controls demonstrated a pattern of increased metabolism in these areas. PTSD patients (but not normals) had decreased hippocampal metabolism with yohimbine. These findings were consistent with an increased release of NE in the brain following yohimbine in PTSD and PET metabolism studies showing an inverse-U relationship between anxiety and cortical function, with low levels of anxiety causing an increase in cortical metabolism and high levels causing a decrease in cortical metabolism.

Several studies have used PET in challenge models that vary from reading narrative scripts to exposing patients to slides or slides and sounds of traumatic material. In these studies [^{15}O]H_2O was used to examine brain blood flow during the cognitive challenge which provoked PTSD symptoms and traumatic remembrance. Control subjects typically demonstrated increased blood flow in anterior cingulate during these conditions. When women with histories of childhood abuse were read personalized scripts of their trauma history, blood flow correlates of exposure

Table III

Overview of Findings from Structural Neuroimaging Studies, Functional Studies, and SPECT Bz Studies in Panic Disorder Highlighting Volume Changes, Blood Flow Changes, and Changes in Binding of Receptor Sites

Method	Disorder			
Structural neuroimaging			**Decrease in volume**	**Increase in volume**
MRI	Panic		Temporal lobe abnormalities	
	PTSD		Hippocampus	
Functional neuroimaging			**Decrease in blood flow/metabolism**	**Increase in blood flow**
PET: $[^{15}O]H_2O$ or FDG	PTSD	*Baseline*	Resting flow in temporal and prefrontal cortex	
		Yohimbine	Temporal and parietal cortex	
			Orbitofrontal and prefrontal cortex	
			Hippocampus	
		Traumatic scripts	Medial prefrontal cortex	Right amygdala/insula
			Right hippocampus	Posterior cingulate and motor cortex
			Anterior cingulate	
			Orbitofrontal cortex	
		Combat slides/sounds	Left inferior frontal cortex	
			Anterior cingulate	
			Medial prefrontal cortex	
	Panic	*Lactate infusion*	Left–right parahippocampus	
			Left inferior parietal cortex	
			Left–right hippocampal ratio	
	Phobias	*Exposure*		Temporal pole, orbitofrontal cortex
				Visual association cortex
SPECT: $[^{99}m$ Tc] HMPAO	Panic	*Baseline*	Hippocampal perfusion	
		Yohimbine	Blunted frontal cortical activation	
		Lactate infusion	Blunted global and normalized in occipital cortex	
			Left occipital cortex	
			Right–left inferior frontal asymmetry	
Proton MRS	Panic	*Hyperventilation*		Lactate levels
Neurochemical and neuroreceptor			**Decrease in binding**	**Increase in binding**
SPECT	Panic	*Bz binding*	Left hippocampal	
			Precuneus	
			Prefrontal, orbitofrontal	

to these scripts showed decreased blood flow in mPFC (areas 24 and 25) and failure of activation in anterior cingulate (area 32), with increased blood flow in posterior cingulate and motor cortex and anterolateral prefrontal cortex. These areas are known to modulate emotion and fear responsiveness through inhibition of amygdala responsiveness. At the same time, these women also had decreased blood flow in the right hippocampus, parietal, and visual association cortex. These findings replicated findings in combat veterans with PTSD exposed to combat-related slides and sound (Fig. 2).

One study found amygdalar activation, but overall findings of amygdalar activation are mixed. These findings point to a network of related regions as mediating symptoms of anxiety. Dysfunction of the

Figure 1 Part of an MRI scan of the hippocampus in a normal control and patient with combat-related PTSD. The hippocampus is visibly smaller in PTSD. Overall, there was an 8% reduction in the right hippocampal volume in PTSD.

mPFC areas may represent a neural correlate of a failure of extinction to fearful stimuli in PTSD. Posterior cingulate plays an important role in visuospatial processing and is therefore an important component of preparation for coping with a physical threat. The posterior cingulate gyrus has functional connections with hippocampus and adjacent cortex, which led to its original classification as part of the limbic brain.

Findings from imaging studies may also be relevant to the failure of extinction to fear responding that is characteristic of PTSD and other anxiety disorders. Following the development of conditioned fear, as in the pairing of a neutral stimulus (bright light, CS) with a fear-inducing stimulus (electric shock, UCS), repeated exposure to the CS alone normally results in the gradual loss of fear responding. Extinction to conditioned fear has been hypothesized to be secondary to the formation of new memories that mask the original conditioned fear memory. The extinguished memory is rapidly reversible following reexposure to the CS–UCS pairing even 1 year after the original period of fear conditioning, suggesting that the fear response did not disappear but was merely inhibited. The extinction is mediated by cortical inhibition of amygdala responsiveness. mPFC has inhibitory connections to the amygdala that play a role in extinction of fear responding. Auditory association cortex (middle temporal gyrus) also has projections to amygdala that seem to be involved in extinction. Anterior cingulate activation represents a ''normal'' brain response to traumatic stimuli that serves to inhibit feelings of fearfulness when there is no true threat. Failure of activation in this area and/or decreased blood flow in adjacent subcallosal gyrus may lead to increased fearfulness that is not appropriate for the context.

In PD activation studies have revealed the involvement of many brain areas, depending on the condition and the paradigm. However, the orbitofrontal cortex/anterior insula and the anterior cingulate are implicated in all the studies and may represent the nodal point between somatic and cognitive symptoms of panic but also any other form of anxiety. Most studies of binding at the Bz–GABA(A) receptor are difficult to interpret because of substantial methodological problems; however, regional and/or global reductions are the most consistent finding in PD.

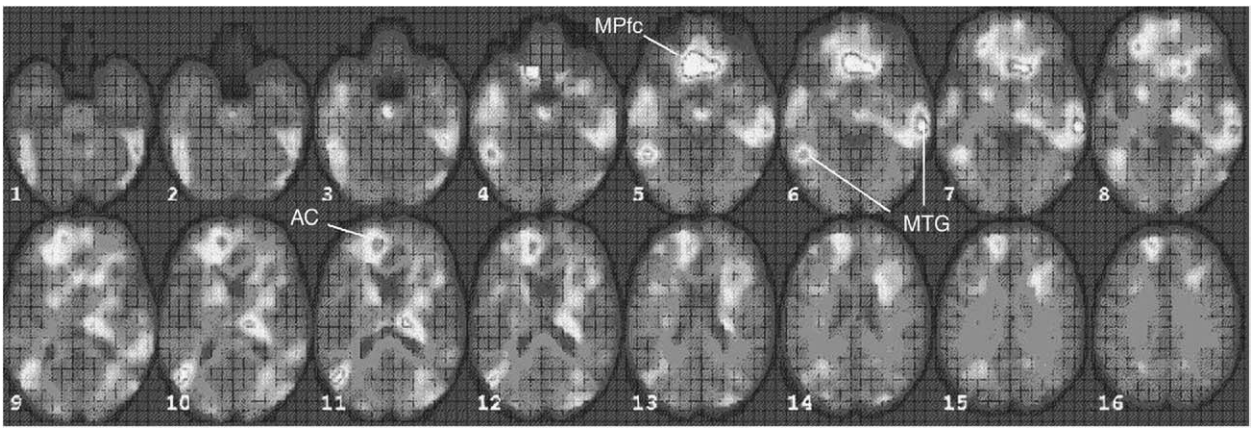


Figure 2 Anxiety response in O^{15} PET scanning in veterans with PTSD during memories of war using audiovisual cues of war. There is decreased blood flow in the medial prefrontal cortex, medial temporal gyrus, and anterior cingulate indicative of a failure of inhibition of limbic structures such as amygdala (areas displayed with z score > 3.00; $p < 0.001$).

IX. A WORKING MODEL FOR THE NEURAL CIRCUITRY OF ANXIETY DISORDERS

Anxiety disorders are characterized by dysfunction of an interrelated neurochemical and neuroanatomical system. PTSD and PD share many biological and phenomenological similarities that allow them to be considered related. Phobic disorders and GAD are still in early stages of investigation. Although phenomenologically they are similar to PTSD and PD, it is premature to include them in a model for human anxiety disorders. PTSD is related more to the deleterious effects of environmental stress, whereas PD is not as clearly related to stress and may be related more to genetic variability in anxiety. In stress-related anxiety disorders (i.e., PTSD), PTSD symptoms as well as cognitive dysfunction associated with PTSD may be linked to hippocampal dysfunction. A model can be created which incorporates informatiom from animal and clinical research relevant to these disorders, keeping in mind that working models are subject to modification with new information, and that generalizations involving causality should be seen as merely speculative when derived from clinical studies that are by their very nature cross sectional. For a schematic model of the neural circuitry see Fig. 3.

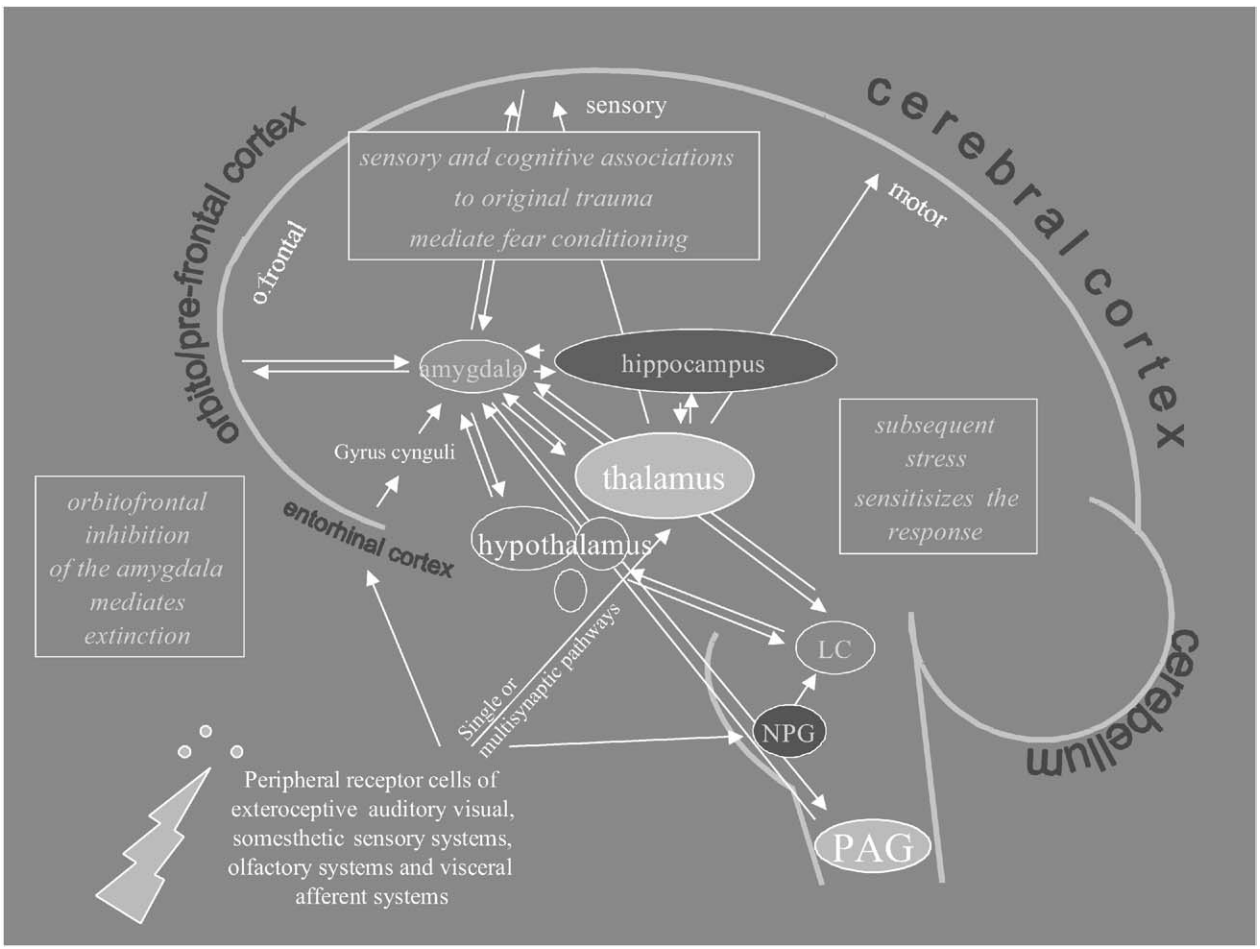


Figure 3 A schematic model of the neural circuits involved in the afferent input of fear- and anxiety-inducing stimuli and the processing of the stimuli. The amygdala plays a pivotal role in the assessment of danger and the response. The LC is a critical component in both afferent and efferent systems, mainly through NE release. The amygdala receives input from PAG, LC, thalamus, hippocampus, association cortices, entorhinal cortex, and visceral pathways. Input from mPFC is involved in determining the significance of fear-producing sensory events, the choice and implementation and the type of behavior, and the extinction of conditioned fear responses. States of stress and fear result in a rapid increase in firing of neurons in the LC, with release of NE transmitter in different target sites in the brain. This results in an increase in attention and vigilance, as well as enhancement of memory recall, which can be life saving in threatening situations. Patients with anxiety disorder, however, develop long-term alterations in the function of this system. The fear response is dependent on previous experience, sensitization, fear conditioning, and extinction of previous fear responses.

A biological model to explain pathological human anxiety should involve both brain stem circuits and cortical and subcortical regions involved in memory and modulation of emotion. The evidence is consistent with chronically increased function of neurochemical systems (CRF and NE) that mediate the fear response in anxiety disorders. Although it is clear that activity at the central portion of the HPA axis is increased, responses at other portions of the HPA axis, including the pituitary and adrenals, and the long-term effects on the hormonal final product (cortisol), are less clear. Increased NE and CRF released in the brain act on specific brain areas, including hippocampus, mPFC, temporal and parietal cortex, and cingulate, that are dysfunctional in human anxiety disorders. Other neurochemical systems, including Bzs, opiates, dopamine, CCK, and NPY, also play a role.

Emotion is a phenomenon uniquely associated with our species. Moving up in terms of species complexity, the most salient change in brain architecture is the massive increase in cortical gray matter, especially frontal cortex. It is therefore not surprising that this frontal lobe plays a role in modulation of emotionality. The medial portion of prefrontal cortex serves to inhibit more primitive limbic processing and thus has an important role in modulation of human emotion in general and also in fear responsivity and anxiety. Amygdala and LC both play a pivotal role in fear processing and anxiety. Hippocampal dysfunction also plays an important role in the development of symptoms of anxiety. mPFC (areas 24 and 25) and anterior cingulate (area 32) have inhibitory inputs that decrease amygdala responsiveness and have been hypothesized to mediate extinction of fear responding. Activation of this area has been shown to be a normal response to stress or increased emotionality. Dysfunction in this area may mediate increased emotionality and failure of extinction to fear inducing cues in anxiety disorders. Evidence to support this idea includes failure of normal activation in this area with yohimbine-induced provocation of anxiety in both PTSD and PD and failure of activation/decreased blood flow with traumatic cue exposure in PTSD. Again, potentiated release of NE with stressors in PTSD and PD is expected to be associated with a relative decrease in function of neurons in this area.

Studies performed to date are encouraging because many findings from animal studies have been successfully applied to human anxiety disorders. The past decade has seen an exciting expansion of research in the field of fear and anxiety. Future research will need to continue to apply findings from the revolution in neuroscience to further understand fear processing, anxiety, and human anxiety disorders.

See Also the Following Articles

CONVERSION DISORDERS AND SOMATOFORM DISORDERS • DEPRESSION • EMOTION • HOMEOSTATIC MECHANISMS • NEUROPSYCHOLOGICAL ASSESSMENT • PSYCHONEUROENDOCRINOLOGY • STRESS

Suggested Reading

Bremner, J. D., Southwick, S. M., and Charney, D. S. (1999). The neurobiology of posttraumatic stress disorder: An integration of animal and human research. In *Posttraumatic Stress Disorder; A Comprehensive Text* (P. A. Saigh and J. D. Bremner, Eds.), pp. 103–144. Allyn & Bacon, Boston.

Bremner, J. D., Staib, L. H., Kaloupek, D., Southwick, S. M., Soufer, R., and Charney, D. S. (1999). Positron emission tomographic (PET)-based measurement of cerebral blood flow correlates of traumatic reminders in Vietnam combat veterans with and without posttraumatic stress disorder. *Biol. Psychiatr.* **45,** 806–816.

Charney, D. S., and Bremner, J. D. (2000). Psychobiology of PTSD. In *Neurobiology of Mental Illness* (D. S. Charney, E. Nestler, and B. Bunny, Eds.), pp. 494–517. Oxford Univ. Press, Oxford.

Charney, D. S., Deutch, A. Y., Krystal, J. H., and Southwick, S. M. (1993). Psychobiological mechanisms of posttraumatic stress disorder. *Arch. Gen. Psychiatr.* **50,** 294–305.

Coplan, J. D., and Lydiard, R. B. (1998). Brain circuits in panic disorder. *Biol. Psychiatr.* **44,** 1264–1276.

Davis, M. (1994). The role of the amygdala in emotional learning. *Int. Rev. Neurobiol.* **36,** 225–266.

Fendt, M., and Fanselow, M. S. (1999). The neuroanatomical and neurochemical basis of conditioned fear. *Neurosci. Biobehav. Rev.* **23,** 743–760.

Hamann, S. B., Stefanacci, L., Squire, L., Adolphs, R., Tranel, D., and Damasio, H. (1996). Recognizing facial emotion. *Nature* **379**(6565), 497.

Kim, J. J., and Fanselow, M. S. (1992). Modality-specific retrograde amnesia of fear. *Science* **256,** 675–677.

LeDoux, J. E. (1998). Fear and the brain: Where have we been, and where are we going? *Biol. Psychiatr.* **44,** 1229–1238.

McEwen, B. S., Angulo, J., Cameron, H., Chao, H. M., Daniels, D., Gannon, M. N., Gould, E., Mendelson, S., Sakai, R., Spencer, R., and Wooley, C. (1992). Paradoxical effects of adrenal steroids on the brain: Protection versus degeneration. *Biol. Psychiatr.* **31,** 177–199.

Phillips, R. G., and LeDoux, J. E. (1992). Differential contribution of amygdala and hippocampus to dues and contextual fear conditioning. *Biobehav. Neurosci.* **106,** 274–285.

Roy-Byrne, P. P., and Cowley, D. S. (1998). Search for pathophysiology of panic disorder. *Lancet* **352,** 1646–1647.

Sapolsky, R. M. (1996). Why stress is bad for your brain. *Science* **273,** 749–750.

Westenberg, H. G. M., Den Boer, J. A., and Murphy, D. L. (Eds.) (1996). *Advances in the Neurobiology of Anxiety Disorders*. Wiley, New York.

Aphasia

MARTHA TAYLOR SARNO
New York University School of Medicine

GLOSSARY

anosognosia A lack of awareness and recognition of a condition or disease.

apraxia of speech An articulatory deficit characterized by awkward and labored articulation, sound distortions, and substitutions in the absence of impaired strength or coordination of the speech organs.

dysarthria A speech impairment resulting from pathology of the motor speech system in which the articulation, rate, resonance, loudness, voice quality, or other acoustic aspect of speech are affected.

neologism A meaningless word.

phonology The system of speech sounds.

prosody The melody, stress, and intonation of speech.

Aphasia is an acquired communication disorder caused by damage to the brain that is manifest in an impairment of the ability to transmit or exchange information and feelings, especially in speaking, and it may also affect writing, comprehension of spoken language, and reading. There are more than 1 million individuals in the United States with aphasia and approximately 84,000 new cases each year. The aphasia population comprises approximately 15% of the adult speech–language impaired population. Aphasia is most prevalent among individuals older than the age of 65 who have suffered a stroke, and in a smaller number it may be the result of head trauma or tumors. Aphasia is also sometimes present in the early stages of Alzheimer's disease.

I. HISTORICAL ASPECTS OF APHASIA

The cerebral lesions that cause aphasia are generally in the left cerebral hemisphere. Some early Greek medical literature noted speech impairment associated with cerebral damage but little was made of the connection between aphasia and the left cerebral hemisphere until the late 19th century when Paul Broca, a surgeon and anthropologist, reported the autopsy findings of the brains of two patients with aphasia. Twenty-five years before Broca made his observations known, Gustav Dax published a paper written by his father, Marc Dax, in which he presented a mass of evidence supporting the discovery that aphasia was due to lesions of the left hemisphere.

Efforts to localize aphasia in its many forms continued throughout the 19th and 20th centuries by medical scientists such as Wernicke and Lichtheim in Germany, Pierre Marie in France, and Henry Head in Britain. Head took issue with the idea that specific areas of the cortex were responsible for the complex functions necessary for human communication. Among those who contributed to the understanding of aphasia in the early 20th century were Kurt Goldstein, a German neurologist; Weisenburg and McBride in the United States; Alajouanine, Ombredane, and Durand in France; and Roman Jakobson, a noted linguist who had emigrated to the United States.

In the early 1960s, Norman Geschwind, a neurologist at the Boston Veteran's Administration Hospital and director of its aphasia center, reinterpreted the works of earlier investigators and stimulated renewed interest in the study of aphasia. His interpretation of aphasia as the result of neural disconnection and concurrent advances in radiologic technology

facilitated the neuroanatomical study of aphasia. In this same era, D. Frank Benson (1967) reported findings from one of the first studies that systematically correlated neuroimaging with the symptoms of aphasia. Benson's work provided radiologic confirmation of the distinction between posterior left hemisphere lesions resulting in "fluent" aphasia and posterior lesions with "nonfluent" aphasia using radiologic data. The emergence of such measures as computerized tomography, magnetic resonance imaging, single-emission computed spectography, positron emission tomography, and functional MRI in the last half of the 20th century increased the understanding of the relationships between lesion localization and the signs and symptoms of aphasia. Although the evolution of neuroimaging techniques has had a great impact on the study of the neural basis of aphasia, the autopsy studies that preceded them have withstood the test of time (Fig. 1).

The contemporary study of aphasia has been influenced by the involvement of many disciplines including the late 20th century emergence of speech–language pathology, neuropsychology, behavioral neurology, and neurolinguistics. Each field has brought its particular perspective, expertise, and body of knowledge to bear on the study of aphasia. The founding of the Academy of Aphasia in 1962 and the publication in this period of new journals that address issues bearing on the study of aphasia, such as *Cortex*, *Brain and Language*, and *Aphasiology*, reflected an increased interest in the condition.

II. SYMPTOMS AND CLASSIFICATION

A. The Nature of Communication Behavior

When an individual generates an idea that he or she wants to verbalize, certain physiologic and acoustic events must take place in order for it to be transformed into words and sentences. The message is converted into linguistic form at the listener's end. The listener, in turn, fits the auditory information into a sequence of words and sentences that are ultimately understood. The system of symbols that are strung together into sentences expressing our thoughts and the understanding of those messages is referred to as language.

Language is made up of phonology, the system of speech sounds that are combined to form the syllables that comprise words; a lexical system, or vocabulary of words, used to communicate information; the grammar, or syntax, that determines the sequence of words that are acceptable as utterances; and semantics, or the meaning system. When communicating using speech, we also employ stress and intonation, referred to as prosody, to help make distinctions among questions, statements, expressions of emotions, shock, exclamation, and so forth.

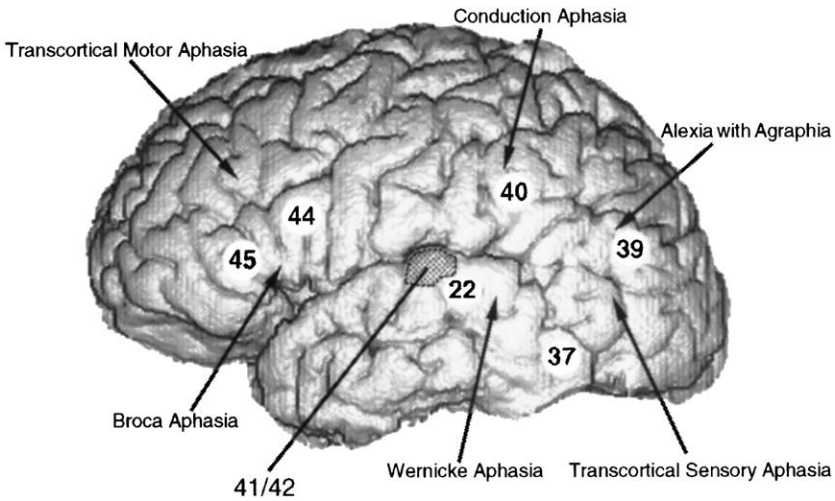


Figure 1 Lateral view of the left hemisphere of a normal adult brain using thin contiguous MR slices and Brainvox. Brodmann's areas 44 and 45 correspond to the classic Broca's area, and area 22 to Wernicke's area. Areas 41 and 42 correspond to the primary auditory cortex; these are located in the depth of the sylvian fissure and cannot be seen in a lateral view of the brain. Area 40 is the supramarginal gyrus; area 39 is the angular gyrus. Area 37, principally located in the posterior sector of the second and third temporal gyrus, does not have correspondence in gyral nomenclature. [reproduced with permission from Damasio, H. (1998). Neuroanatomical Correlates of the Aphasias. In *Acquired Aphasia*, 3rd ed., (M. T. Sarno, ed.), p. 45. Academic Press, San Diego.]

B. Types of Aphasia

Aphasia is generally of sudden onset and is sometimes present during the acute phase of an illness and then disappears in a matter of hours or days. In this context, the term aphasia is used when its symptoms persist for more than 1 month.

An exception to the typically sudden onset of aphasia occurs in primary progressive aphasia (PPA), a diagnosis that is being made with increasing frequency. PPA is usually gradual in onset, sometimes emerging over a period of years, and may evolve into a dementia or Alzheimer's disease.

Individuals with aphasia may show evidence of impairment in any or all language systems ranging from a virtually total inability to communicate using the speech code, but with preserved ability to communicate through the use of gestures, facial expression, and pantomime, to a mild, barely perceptible language impairment.

There is a group of distinct aphasia syndromes that have a high correlation with the location of anatomical lesions. However, it is not always possible to classify patients according to these syndromes. In fact, estimates of the proportion of cases that can be unambiguously classified range from 30 to 80%.

The determination of the aphasia syndrome that best fits depends primarily on identification of the characteristics of speech production combined with a judgment of fluency. Speech output that is hesitant, awkward, interrupted, and produced with effort is referred to as nonfluent aphasia, in contrast to speech produced with ease of articulation, at a normal rate, with preserved flow and melody but that may be lacking in coherence, which is referred to as fluent aphasia. Fluency judgments are generally derived from an extended conversation with a patient.

1. Fluent Aphasia

Fluent aphasia is characterized by fluent speech produced at a normal rate and melody, accompanied by impaired auditory comprehension. It is generally associated with a lesion in the vicinity of the posterior portion of the first temporal gyrus of the left hemisphere. When fluent aphasia is severe, the individual may substitute sounds and words with such frequency and magnitude that speech may be rendered meaningless. Some produce nonsense words, referred to as neologisms or jargon aphasia. Those with fluent aphasia tend to have greatest difficulty retrieving the substantive parts of speech (i.e., nouns and verbs) and also tend to manifest impaired awareness. They do not generally evidence paralysis or weakness of the right arm and leg.

a. Wernicke's Aphasia The most common variety of fluent aphasia is Wernicke's aphasia, characterized by fluently articulated speech sometimes marked by word substitutions and accompanied by impaired auditory comprehension. Individuals with Wernicke's aphasia may produce what appear to be complete utterances, use complex verb forms, and speak at a rate greater than normal that is sometimes referred to as press of speech. Their communication behavior, profile of language impairment, and lack of physical impairment may lead to a psychiatric diagnosis (Fig. 2).

b. Conduction Aphasia Conduction aphasia is also one of the fluent aphasias, but unlike Wernicke's aphasia, auditory comprehension is generally more intact. A deficit of word and sentence repetition prevails and is marked by phonemic paraphasic errors—that is, the production of inappropriate, although precisely articulated speech sounds.

c. Transcortical Sensory Aphasia Individuals with transcortical sensory aphasia have fluent speech output marked by substitutions of words and severe impairment of aural comprehension. Transcortical sensory aphasia is distinguishable from Wernicke's aphasia by a preserved repetition performance. Also, those with transcortical sensory aphasia often "echo" questions, a behavior referred to as "echolalia."

d. Anomic Aphasia The primary characteristic of anomic aphasia is a pervasive difficulty in word retrieval in the presence of fluent, well-articulated, and grammatically correct output and intact auditory comprehension. Anomic aphasia is often the mildest form of aphasia and is sometimes the recovery end point of other types of aphasia. It may also characterize the early stages of primary progressive aphasia.

2. Nonfluent Aphasia

Nonfluent aphasia is characterized by limited vocabulary, sometimes restricted to nouns, verbs, adverbs, and adjectives; slow, hesitant speech production, awkward articulation; and a restricted use of grammar with normal or near-normal auditory comprehension. Individuals with nonfluent aphasia are often referred to as agrammatic since they tend to have a pervasive

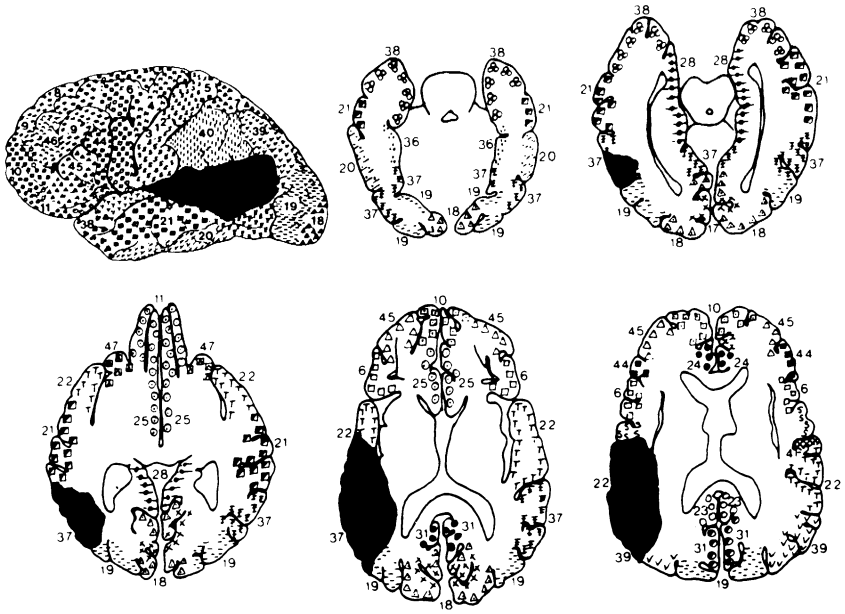


Figure 2 Magnetic resonance template from a patient with Wernicke's aphasia (WG0988). The lesion involved the posterior sector of the left superior and middle temporal gyri but did not extend into the parietal lobe. [reproduced with permission from Damasio, H. (1998). Neuroanatomical Correlates of the Aphasias. In *Acquired Aphasia*, 3rd ed., (M. T. Sarno, ed.), p. 50. Academic Press, San Diego.]

grammatical impairment as the result of difficulty in retrieving less substantive parts of speech (i.e., prepositions, articles, and pronouns).

Nonfluent aphasia is usually associated with a high degree of deficit awareness and impaired motor function on the right side of the body (hemiplegia–paresis). The intact deficit awareness generally present in nonfluent aphasia may produce a significant degree of frustration.

a. Broca's Aphasia The most common type of nonfluent aphasia is Broca's aphasia, which is characterized by awkward articulation, restricted vocabulary, and grammatical forms in the presence of normal or near-normal auditory comprehension. Writing performance generally mirrors speech production and reading may be less impaired than speech and writing. In the majority of individuals with Broca's aphasia, awareness of deficit is good and paralysis on the right side of the body (hemiplegia) is present (Fig. 3).

3. Apraxia of Speech

The term apraxia refers to an impairment in the ability to carry out physical movements in the absence of paralysis or sensory impairment to the body part. Apraxia of speech (AOS), a term synonymous with verbal apraxia and speech dyspraxia, is seldom manifest independent of Broca's aphasia, however mild. It is an articulatory deficit characterized by awkward and labored articulation, distortion of phoneme production, and sound substitutions in the absence of impaired strength or coordination of the motor speech system. Unlike individuals with dysarthria, an impairment of speech associated with pathology of the motor speech system, individuals with AOS do not have difficulty performing nonspeech movements of the oral musculature.

a. Transcortical Motor Aphasia This type of aphasia is characterized by nonfluent speech, intact repetition, impaired auditory comprehension, and a tendency to perseverate and substitute both sounds and words.

4. Global Aphasia

When aphasia is severe in all modes of communication, rendering a person markedly restricted in participating in verbal interactions, it is referred to as global aphasia. Individuals who manifest global aphasia are often able to produce serial or automatic speech (i.e., counting in series, reciting the days of the week, reciting prayers, singing the words to songs, and using common everyday greetings). Global aphasia is generally associated with extensive cerebral damage, often in both hemispheres.

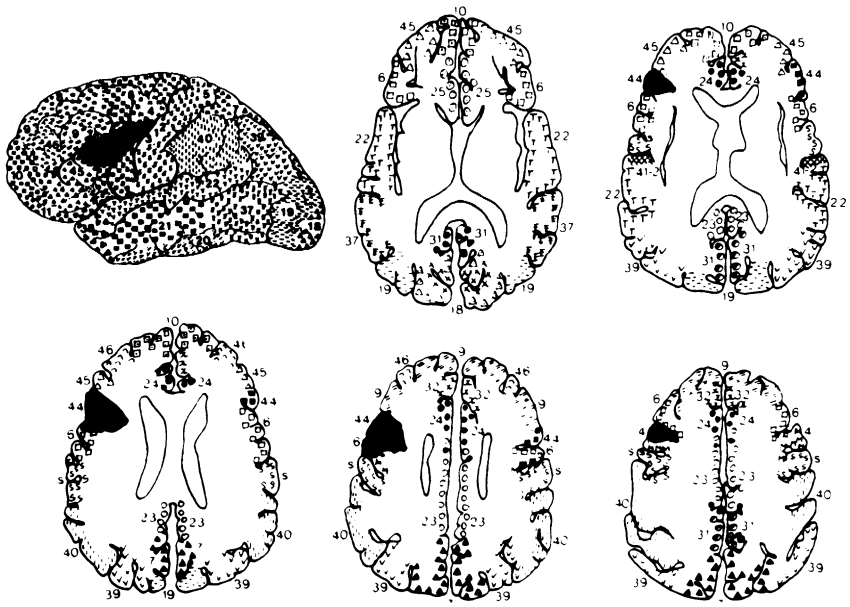


Figure 3 Computerized tomographic template from a patient with Broca's aphasia (MW0018). The lesion involved only the superior sector of Broca's area (area 44) and the premotor region (area 6) immediately above it. [reproduced with permission from Damasio, H. (1998). Neuroanatomical Correlates of the Aphasias. In *Acquired Aphasia*, 3rd ed., (M. T. Sarno, ed.), p. 55. Academic Press, San Diego.]

5. Primary Progressive Aphasia

The onset of language impairment in primary progressive aphasia (PPA), unlike the classic aphasia syndromes, begins gradually. It is often associated with degenerative disease, such as spongiform encephalopathies and Pick's disease. Typically, the condition begins with the gradual onset of word retrieval difficulties, especially for nouns and/or verbs. Although individuals with PPA may appear to be free of cognitive deficits for some time, language deficits other than word-finding difficulties, especially the cognitive symptoms of dementia, usually begin to emerge over time.

C. Classification Nomenclature

Different investigators and clinicians have used different terms to classify the aphasia syndromes, often reflecting a difference in perspective. For example, the classic designation expressive and receptive aphasia was introduced by clinicians Weisenburg and McBride (1935), whereas Luria used the terms efferent, afferent motor, and dynamic aphasia, reflecting a physiological approach. The classification scheme of Goodglass and Kaplan (1972) is currently used (Table I).

III. THE ASSESSMENT OF APHASIA

A. History of Aphasia Assessment

Formal aphasia testing was relatively unknown until the early l920 s when Henry Head and Weisenburg and McBride published the first aphasia tests. Until then, aphasia was assessed primarily on the basis of clinical observations and examination.

B. Purpose of Assessment

The purpose of aphasia assessment varies according to the demands of particular clinical settings and includes the need to

1. Screen patients in the hospital setting in order to determine the presence/absence of aphasia and the need for comprehensive evaluation
2. Perform an in-depth evaluation in order to determine a diagnosis/clinical description
3. Identify areas of deficit and preserved performance for rehabilitation assessment
4. Make recommendations for treatment if indicated
5. Develop a treatment plan
6. Render a prognosis

Table I
Classification Systems Schemes[a]

Boston	Head	Goldstein	Weisenberg and Mcbride	Luria	Wepman
Anomia	Nominal aphasia	Amnesic aphasia	Amnesic aphasia	Acoustic–amnestic aphasia	Semantic aphasia
Wernicke's aphasia	Syntactic aphasia	Sensory aphasia	Predominately receptive aphasia	Acoustic aphasia	Pragmatic aphasia
Broca's aphasia	Verbal aphasia	Motor aphasia	Predominately expressive aphasia	Efferent motor aphasia	Syntactic aphasia
Conduction aphasia		Central aphasia		Afferent motor aphasia	
			Expressive–receptive aphasia		
Transcortical sensory aphasia (isolated speech area syndrome)					
Transcortical motor aphasia				Dynamic aphasia	

[a]Reproduced with permission from Heilman, K. (1979). *Clinical Neuropsychology*. Oxford Univ. Press, Oxford.

The examiner includes an assessment of the individual's motivation, expectations, psychosocial status, and ability to cope with the demands of the rehabilitation process. Baseline measures obtained at the first assessment are used to measure progress over time.

In addition to the described general measures of aphasia, tests of specific linguistic performance, such as the use of syntax, vocabulary, frequency of word usage, pauses, hesitations, length of utterances, and rate of speech, are derived from a free speech sample or a narrative elicited in response to the tasks of describing a picture or telling a story.

C. Aphasia Test Requirements

Adequate measures of aphasia adhere to the same general requirements of all tests; that is, they must be reliable, valid, and standardized. Several issues arise in the design of tests for use with the brain damaged, especially those with aphasia, including the range of item difficulty, overlap of aphasia examinations with intelligence tests, and the effectiveness of the test in measuring progress.

D. Aphasia Measures

Comprehensive language tests designed to assess aphasia are usually structured around specific domains of performance: visual confrontation naming; a spontaneous or conversational speech sample that is analyzed for fluency, articulation, phrase length, prosody, word substitutions and omissions; repetition of digits, single words, multisyllabic words, and sentences of increasing length and complexity; comprehension of single words and sentences that require yes–no responses and pointing on command; word retrieval; reading and writing.

Some of the most widely used aphasia measures are the Boston Diagnostic Aphasia Examination, the Neurosensory Center Comprehensive Examination for Aphasia, the Western Aphasia Battery, the Minnesota Test for Differential Diagnosis of Aphasia, and the Multilingual Aphasia Examination.

Screening tests include the Sklar Aphasia Scale and the Reitan Aphasia Screening test. Tests of specific language performance include the Shewan Auditory Comprehension Test for Sentences, the Boston Naming Test, Controlled Oral Word Association, the Discourse Comprehension Test, the Reporter's Test, and the Token Test. A listing of aphasia assessment instruments is provided in Table II.

E. Functional Communication Assessment

In addition to measuring specific linguistic performance, a comprehensive aphasia assessment also

Table II
Differential Diagnosis of the Main Types of Aphasia[a]

Type of aphasia	Speech	Comprehension	Capacity for repetition	Other signs	Region affected
Broca's	Nonfluent; effortful	Intact or largely preserved	Impaired	Right hemiparesis (arm and leg); patient aware of defect and may be depressed	Left frontal (lower, posterior)
Wernicke's	Fluent; abundant; well-articulated; melodic	Impaired	Impaired	No motor signs; patient may be anxious, agitated, euphoric, or paranoid	Left temporal (posterior and superior)
Conduction	Fluent with some articulatory deficits	Intact or largely preserved	Impaired	Often none; patient may have cortical sensory loss or weakness in right arm; right-sided facial weakness may be seen	Left supramarginal gyrus or left auditory cortex and insula
Global	Scant; nonfluent	Impaired	Impaired	Right hemiplegia may be present without hemiplegia	With hemiplegia: massive left perisylvian lesion Without hemiplegia: separate frontal and temperoparietal damage
Transcortical					
motor	Nonfluent; explosive	Intact or largely preserved	Intact or largely preserved	—	Anterior or superior to Broca's area; may involve part of Broca's area
Sensory	Fluent; scant	Impaired	Intact or largely preserved	—	Area surrounding Wernicke's area, posteriorly or interiorly
Atypical					
"Basal ganglia"	Dysarthric, but often fluent	Impaired	May be intact or Impaired	Right hemiparesis (arm and leg)	Head of caudate nucleus; anterior limb of capsule
Thalamus	Fluent; may be logorrheic	Impaired	Intact or largely preserved	Attentional and memory defects in acute phase	Anterolateral thalamus

[a]Reproduced with permission from Damasio (1992).

includes supplementary measures of "functional communication."

This category of measures provides information about the individual's actual use of residual communication skills in everyday life, in contrast to performance, which is a response to the presentation of specific tasks in a structured testing situation. Ordinary everyday behaviors, such as making telephone calls, using everyday greetings, and reading the newspaper, have been noted to reflect another dimension of communication behavior that is often not elicited in the performance of a specific, structured task. The notion of a "functional" dimension of communication behavior emerged from the rehabilitation medicine setting in the late 1950s when it was observed that many patients with aphasia communicated with different degrees of skill depending on whether they were communicating in a natural, interactive speaking situation or in structured, task-oriented conditions.

Since then, many rating scales designed to assess the functional communication dimension of aphasia have been published, including the Functional

Table III
Aphasia Assessment Instruments[a]

Full title	Abbreviation	Source
Aphasia Screening Test	AST	Reitan (1991)
Appraisal of Language Disturbances	ALD	Emerick (1971)
Arizona Battery for Communication Disorder of Dementia	ABCD	Bayles and Tomoeda (1990)
Auditory Comprehension Test for Sentences	ACTS	Shewan (1988)
Boston Assessment of Severe Aphasia	BASA	Helm-Estabrooks *et al.* (1989)
Boston Diagnostic Aphasia Examination	BDAE	Goodglass and Kaplan (1983)
Boston Naming Test	BNT	Kaplan and Goodglass (1983)
Controlled Oral Word Association	COWA	Spreen and Benton (1977)
Discourse Comprehension Test	DCT	Brookshire and Nicholas (1997)
Minnesota Test for Differential Diagnosis of Aphasia	MTDDA	Schuell (1965, 1973)
Multilingual Aphasia Examination	MAE	Benton *et al.* (1994)
Neurosensory Center Comprehensive Examination for Aphasia	NCCEA	Spreen and Benton (1974)
Pantomime Recognition Test	—	Benton *et al.* (1994)
Phoneme Discrimination Test	—	Benton *et al.* (1983)
Porch Index of Communicative Ability	PICA	Porch (1981)
Reporter's Test	—	De Renzi (1980)
Sklar Aphasia Scale	SAS	Sklar (1973)
Sound Recognition Test	SRT	Spreen and Benton (1974)
Token Test	TT	Many versions
Western Aphasia Battery	WAB	Kertesz (1982)

[a]Reproduced with permission from Spreen, O., and Risser, A. H. (1998). Assessment of Aphasia. In *Acquired Aphasia*, 3rd ed., (M. T. Sarno, ed.), Academic Press, San Diego.

Communication Profile (1969), the Communicative Abilities of Daily Living (1980), the Communicative Effectiveness Index (1989), and the Functional Abilities of Communication Profile (1995).

F. Assessment of Aphasia in Bilingual Speakers

Bilingual speakers who acquire aphasia present special problems with respect to the assessment of the

Table IV
Tests Available in Translation or Adaptation[a]

Test	Language
Bilingual Aphasia Test	French and other languages (Paradis, 1987)
Boston Diagnostic Aphasia Examination	Norwegian (Reinvang and Graves, 1975), Spanish (Garcia-Albea *et al.*, 1986)
Boston Naming Test	Spanish (Taussig *et al.*, 1988)
Communication Abilities in Daily Living	Italian, Japanese (Pizzamiglio *et al.*, 1984; Sasanuma, 1991; Watamori *et al.*, 1987)
Controlled Oral Word Association	Spanish (Taussig *et al.*, 1988)
Multilingual Aphasia Examination	Chinese, French, German, Italian, Portugese, Spanish (Rey and Benton, 1991)
Token Test	Italian, German, Portugese
Western Aphasia Battery	Portugese

[a]Reproduced with permission from Spreen, O., and Risser, A. H. (1998). Assessment of Aphasia. In *Acquired Aphasia*, 3rd ed., (M. T. Sarno, ed.), Academic Press, San Diego.

disorder. Study results differ as to whether or not the older, more frequently used language is more preserved or whether there is little difference in the accessibility of an individual's primary and secondary languages.

Several tests have been translated or adapted for use with bilingual patients and are listed in Table III. The Multilingual Aphasia Examination (MAE) was designed to provide fully equivalent forms in several languages but best results are obtained when it is administered by someone fluent in both languages (Table IV).

G. Other Factors

Aphasia test results can be affected by many nonlanguage factors that need to be considered in the interpretation of test findings. These include acute confusion, agitation, drowsiness, depression, distractibility, lack of cooperation, and attention deficits.

Given the complexity and broad scope of human communication that encompasses both verbal and nonverbal domains (i.e., facial expression and body language), aphasia tests cannot be sufficiently comprehensive or sensitive to the full impact of aphasia on communication behavior. The interpretation of aphasia test results requires an interpretation of the findings in the context of the individual's premorbid communication functioning, his or her perceived impairment, compensatory skills, educational and vocational history, and overall psychosocial status. Most important, the unique, personal nature of communication behavior makes it difficult to make decisions regarding diagnosis, classification, and rehabilitation management based on test findings alone.

H. Related Disorders

Certain associated disorders are generally present in the majority of individuals with aphasia. These include deficits in writing, reading, and calculation. Patients with Broca's aphasia, for example, tend to have difficulty with prepositions, articles, and conjunctions not only in speaking but also when writing or reading, that may range from mild to severe. This include difficulty reading aloud, reading certain parts of speech, greater ease reading longer than shorter words, confusion over word meanings (i.e., reading the word "wife" for "sister"), difficulty copying letters, and writing small words or words with a low frequency of occurrence. Letter deletions, substitutions, and additions may also be present. Those with calculation disorders in addition to aphasia may show deficits in all four basic arithmetic operations.

1. Peripheral Field Deficits

A common physical consequence of damage to the brain is visual deficits involving the peripheral field of vision. The peripheral fields of vision refer to what we see out of the "corners" of our eyes when looking straight ahead. In the case of someone with a left hemisphere cerebral lesion, the individual may have absent vision in the right peripheral field of both eyes and this can further complicate the reading process for someone with aphasia.

I. Intelligence and Aphasia

Aphasiologists continue to debate the relationship between aphasia and intelligence. The boundaries of cognition and language are difficult to determine. Some argue that aphasia is simply a component of the cognitive system. Others view thought and language as independent of each other. Cognitive slowing (i.e., a greater than normal amount of time required to perform cognitive processing) decreased activation and initiation, and response latencies, deficits of spatial relations, abstract reasoning, and visual perception have all been identified as symptoms that may be present in aphasia. Some of these may be coincidental to the anatomical proximity of cerebral areas that govern language and cognition.

From an operational perspective, clinicians generally view aphasia as a language processing disorder that is not a thinking disorder. Those with aphasia often state that they know what they want to say but cannot adequately access language. This dichotomy can be observed in the deaf population, in which normal thought processing is present but the ability to convert thought to language is impaired.

The term anosagnosia refers to a lack of awareness and recognition of a condition or disease and is sometimes present in individuals with aphasia.

J. Artistry and Aphasia

Aphasia affects performance in the arts in different ways. Musical skill is not always affected, probably owing to its cerebral control areas being located in the

right rather than left hemisphere. However, among talented musicians there appears to be different representation in the brain for musical abilities.

In the visual arts, the average individual appears to lose some ability but the very talented continue to function without apparent difficulty. This is not the case among professional writers, in whom aphasia has a profound effect on performance.

K. Aging and Aphasia

Some language changes that are characteristic of aphasia are also present in the normal elderly. For example, the most frequently reported age-related language changes in the normal elderly are word finding difficulties and the comprehension of complex sentences.

Elderly individuals with Broca's aphasia, on average tend to be significantly younger than the median age of people with Wernickes aphasia. Ninety-five percent of aphasias in individuals younger than 30 years of age are of the nonfluent type, whereas only 45% of aphasias are nonfluent in individuals over the age of 60. Results of studies addressing the predictive effect of age on prognosis for recovery are controversial.

L. Acquired Aphasia in Children

Use of the term aphasia has unfortunately been applied to children who have not yet developed age-appropriate language skills. The term developmental aphasia is often used for this population and is more appropriate.

Acquired aphasia in children differs from that in adults in several important ways. When aphasia is acquired in childhood, the child's stage of both central nervous system development and language development at the time of brain damage must be considered. In adults an assumption is made that language was fully developed at the time of disruption, whereas in children a determination must be made as to which of the observed deficits represent aphasia and which represent aspects of language that have not yet emerged.

Children with acquired aphasia generally recover more rapidly and completely than adults. However, there is no single correlation between age at lesion onset and outcome. Furthermore, long-term studies suggest that residual and long-standing linguistic symptoms appear to be present in many children who have acquired aphasia.

M. Aphasia Secondary to Traumatic Head Injury

The majority of patients have been in vehicular accidents (closed head injuries), with a smaller number receiving gunshot or other penetrating wounds (open head injury). Unlike the pathophysiology of stroke, the majority of those who have suffered traumatic head injury do not have focal cerebral lesions but suffer diffuse injury to the cerebral white matter that is apparently the result of shearing and stretching of nerve fibers at the moment of impact.

Classic aphasia symptoms as a result of traumatic head injury are not uncommon, especially when there is a history of loss of consciousness. A characteristic of acute aphasia after trauma is the pervasiveness of anomia. Nonaphasic language processing deficits are a common finding in the closed head injury population. It is generally reported that aphasia secondary to head injury has a better prognosis for recovery than aphasia secondary to stroke.

N. Psychosocial Aspects of Aphasia

The psychosocial consequences of aphasia represent one of the most significant effects of the condition since communication using language, especially speech, is one of the most fundamental of human behaviors. The effect of aphasia on the individual's sense of self, identity, and quality of life is considerable. The ability to initiate and maintain personal relationships, which are dependent on communication, often leads to social isolation for the person with aphasia. Generally, those with nonfluent disorders tend to be more aware, frustrated, and depressed than those with aphasia of the fluent type.

The psychological reactions to aphasia are believed to be influenced to some degree by premorbid personality, level of achievement, and values. As a result of its negative effect on interpersonal activity and quality of life, aphasia is frequently referred to as a social disability. By far the most commonly mentioned psychological reaction is depression. Difficulty in coping with being socially different, feelings of loss, grief, and lowered self-esteem are also pervasive. Family members also suffer from the effects of role changes, caregiving, the impact on the family's sources of gratification, and difficulties in communication.

O. The Natural Course of Aphasia

When aphasia has persisted for several months, a complete recovery to a premorbid level of communication function is unlikely. The temporal course of aphasia from its sudden onset, acute stage, and subsequent phases of recovery evolves along separate pathways depending on whether one is referring to the individual's communication skills, psychological state, social functioning, compensatory skill level, or level of adjustment.

In the period immediately following onset, a degree of natural recovery referred to as spontaneous recovery takes place in the majority of individuals. There is a lack of concensus as to how long this period lasts, ranging from 2 to 6 months postonset.

It is generally agreed that, with some exceptions, major improvements in communication occur in the first 12–24 months following onset. Gradual changes, however, have been reported for many years thereafter, especially in compensatory, alternative skills that are used more effectively over time.

Age, gender, education, and handedness do not appear to affect recovery. Comprehension tends to recover more than expressive communication. Some individuals recover many communication skills but are left with an anomia. Several studies report that for the more severe aphasias recovery of communication skills begins later than for the mild and moderately impaired.

P. Recovery and Rehabilitation

1. History of Aphasia Rehabilitation

Reports of recovery of function were documented as early as the 16th century. In the late 19th and early 20th centuries, the literature related to the rehabilitation of aphasia was almost exclusively in German. Much of this was the result of World War I experience when rehabilitation centers for the brain injured were established particularly in Germany. During World War II, several specialized, comprehensive programs were developed in military medical centers in the United States for veterans with aphasia secondary to head injuries. In the period immediately following World War II, concurrent with the birth of rehabilitation medicine as a medical specialty and the emergence of speech–language pathology as a health-related profession, treatment services for civilians with aphasia began to emerge.

2. Approaches to Aphasia Rehabilitation

Treatment approaches have generally followed one of two models: a substitute skill model or a direct treatment model. Both models are based on the assumption that the processes that subserve normal performance need to be understood if rehabilitation is to succeed. Treatment methods are categorized according to whether they are primarily indirect stimulation–facilitation or direct, structured, and pedagogic in nature. Comprehensive, specialized treatment programs for aphasia generally offer both individual and group therapy using a multimodal approach depending on language symptoms and psychosocial needs.

A primary goal in the management of aphasia is to train the individual to use communication strategies appropriate to the type and severity of impairment in order to substitute or circumvent impaired language processing. However, aphasia rehabilitation goes beyond the development of communication strategies by helping the person and his or her intimates facilitate communication, enhance social interaction, and adjust to the alterations and limitations imposed by the disability. This social disability model considers all the factors that contribute to an individual's life experience, value systems, expectations, needs for fulfillment, level of activation and initiation, social isolation, physical endurance, associated physical and neuropsychological deficits, and emotional state in the intervention process.

Professional responsibility for the language rehabilitation of the individual with aphasia has traditionally been the domain of the speech–language pathologist. In many cases, effective intervention requires a team of several health professionals, including neuropsychologists, occupational therapists, physical therapists, social workers, and vocational counselors. These services are usually available in rehabilitation medicine facilities and centers specializing in the rehabilitation of communicative disorders.

Those experienced in the rehabilitation of aphasia generally agree that no single technique or approach can ameliorate aphasia completely once it is established. Therefore, rehabilitation approaches tend to be eclectic and include a broad spectrum of techniques. Computers are frequently used as a facilitation and training tool for the restoration of language skills by reinforcing through the provision of feedback and frequent repetition and practice. Software designed specifically for the rehabilitation of aphasia and/or facilitation is available. Alternative/augmentative systems are also used for selected individuals.

Meta-analyses of the efficacy of aphasia treatment reported by Robey and Whurr *et al*. indicated (i) that clinical outcomes showed a clear superiority in the performance of those who received treatment by a speech–language pathologist; (ii) the more intensive the treatment, the greater the change; (iii) the outcome is best when treatment is begun in the acute stage of recovery; (iv) treatment for the moderately severe and severe can have a major impact when it is begun in the chronic stage; and (v) treatment in excess of 2 hr per week yields greater gains than less intensive treatment.

Aphasia rehabilitation is viewed as a process of patient management in the broadest sense. The process is complex and encompasses many domains ranging from the neurolinguistic and cognitive to the pragmatic, functional, emotional, and social. As a result, despite several decades of clinical research addressing intervention issues, there is no defined philosophy of aphasia rehabilitation.

Not all individuals with aphasia have access to rehabilitation services, especially in the currently constrained health service delivery situation. As a result, there is a critical need for access to information about aphasia, guidelines and training for the provision of treatment at home by family members, and access to social opportunities and peer support.

A movement to develop advocacy/educational organizations for people with aphasia began in Finland in 1971, followed by Germany in 1978, the United Kingdom in 1980, Sweden in 1981, and Montreal, Quebec, in 1988. In the United States, the National Aphasia Association was founded in 1987 as an advocacy organization that sponsors a national network of aphasia support groups called aphasia community groups, provides educational publications about aphasia, maintains a website (www@aphasia.org), and sponsors conferences and newsletters.

See Also the Following Articles

AGNOSIA • AGRAPHIA • ALEXIA • ANOMIA • APRAXIA • AUTISM • DYSLEXIA • LANGUAGE AND LEXICAL PROCESSING • LANGUAGE DISORDERS • READING DISORDERS, DEVELOPMENTAL • SPEECH

Suggested Reading

Damasio, A. R. (1992). Aphasia. *N. Engl. J. Med.* **326,** 531–539.

Davis, G. A. (2000). *Aphasiology: Disorders and Clinical Practice*. Allyn and Bacon, Boston, MA.

Parr, S., Byng, S., Gilpin, S., and Ireland, C. (1997). *Talking about Aphasia: Living with Loss of Language after Stroke*. Open University Press, Buckingham, U.K.

Robey, R. R. (1998). A meta-analysis of clinical outcomes in the treatment of aphasia. *J. Speech Hearing Res.* **41,** 172–187.

Sarno, M. T. (1993). Aphasia rehabilitation: Psychosocial and ethical considerations. *Aphasiology* **7,** 321–334.

Sarno, M. T. (1997). Quality of life in aphasia in the first poststroke year. *Aphasiology* **11**(7), 665–679.

Sarno, M. T. (Ed.) (1998). *Acquired Aphasia*, 3rd ed. Academic Press, San Diego.

Apraxia

KENNETH M. HEILMAN and LESLIE J. GONZALEZ-ROTHI
University of Florida College of Medicine

GLOSSARY

apraxia A disorder of implementing programs that instruct the motor neurons how to position one's hand and arm to interact with a tool or object, to orient the limb toward the target of the limb's action, to move the limb in space, to determine the speed of the movement, to order a series of acts leading to a goal, and to utilize the mechanical advantage that tools afford.

conceptual apraxia A loss of mechanical knowledge.

conduction apraxia Patients with this form of apraxia are more impaired at imitation than pantomiming to command.

dissociation apraxia Patients have a modality-specific apraxia (e.g., verbal command). Unlike patients with ideomotor and conduction apraxia, patients with this form of apraxia have almost flawless imitation and correctly use actual objects.

ideational apraxia The inability to carry out a series of acts that lead to a goal, an ideational plan.

ideomotor apraxia Patients with ideomotor apraxia make the most severe errors when asked to pantomime transitive acts to verbal command. When imitating their performance may improve but frequently remains abnormal. When using actual tools their performance may improve even further, but often their performance remains impaired. Patients with ideomotor apraxia make primarily spatial and temporal movement errors.

limb kinetic apraxia Patients with limb kinetic apraxia have a loss of the ability to make finely graded, precise, individual finger movements.

The pyramidal motor system, together with motor units, can mediate an infinite number of movements. Therefore, to successfully manipulate environmental stimuli the pyramidal motor neurons must be guided by movement programs. The programs must instruct the motor neurons how to position one's hand and arm to interact with a tool or object, to orient the limb toward the target of the limb's action, to move the limb in space, and to determine the speed of the movement. To successfully interact with the environment, one must also know how to order components of an act to reach a goal and the mechanical advantage that tools afford. Disorders of this control system are called apraxia.

I. HISTORY

In 1871, Heymann Steinthal first used the term "apraxia" for a loss of motor skills. Steinthal thought that apraxia was a defect in "the relationship between movements and the objects with which the movements were concerned." Despite this 19th-century description of apraxia, not many important advances were made until the beginning of the 20th century with the work of Hugo Liepmann, who described three forms of apraxia: limb kinetic, ideomotor, and ideational. After World War I there was a loss of interest in continental European neurology and little interest in apraxia. However, after seeing a patient with a callosal disconnection, Norman Geschwind, in his classic paper "Disconnection Syndromes," renewed interest in apraxia. In this brief review, we describe the three form of apraxia described by Liepmann as well as three other forms of apraxia he did not discuss.

II. APRAXIA TESTING

In order to be classified as being apraxic, the inability to perform skilled purposeful movements should not be caused by sensory loss or by more elemental motor disorders such as weakness, rigidity, tremors, dystonia, chorea, ballismus, athetosis, myoclonus, or seizures. Cognitive, motivational, and attentional disorders may also be associated with an inability to perform skilled acts. Therefore, before diagnosing apraxia, a patient must undergo a detailed neurological examination, including testing of cognitive function. Although the presence of cognitive behavioral disorders does not preclude that the patient may also have apraxia, before diagnosing apraxia the clinician should be certain that these behavioral disorders do not fully account for their patient's inability to perform skilled acts.

Testing praxis involves selectively varying input and task demands. When possible, the same items should be used for all the following subtests. First, patients should be requested to pantomime/gesture. Both transitive (e.g., ''Show me how you would slice bread'') and intransitive gestures (e.g., ''Wave good-bye'') should be tested. Independent of the way in which patients perform to command, patients should be asked to imitate the examiner performing, both with meaningful and meaningless gestures. Independent of the results of the pantomime to command and imitation tests, the patient should also be allowed to see and hold actual tools or objects and to demonstrate how to use them. In addition to having a patient pantomime to verbal command, the examiner may also want to have the patient pantomime in response to seeing pictures of tool or the objects on which tools work. The examiner should learn if the patient can name or recognize transitive and intransitive pantomimes made by the examiner and discriminate between well-performed and poorly performed pantomimes that are performed by the examiner. Lastly, the examiner may want to learn if the patient knows the mechanical advantage that tools afford and if the patient can substitute or even fabricate tools to solve mechanical problems.

III. TYPES OF LIMB APRAXIA

There are six major forms of limb apraxia: limb kinetic, ideomotor, dissociation, conduction, ideational, and conceptual. Each of these forms of apraxia is defined by the nature of errors made by the patient, and each of these disorders also has different mechanisms. In this article, we discuss the clinical presentation of each of these apraxic disorders. Although constructional apraxia and dressing apraxia also involve the limbs, these disorders are strongly associated with visuoperceptual–visuospatial disorders and will not be discussed here.

A. Limb Kinetic Apraxia

In 1920, Liepmann found that patients with limb kinetic apraxia have a loss of the ability to make finely graded, precise, individual finger movements. Tasks such as finger tapping and pegboard may be impaired. The patient may have difficulty picking up a straight pin from the top of a desk using a pincer grasp of the thumb and forefinger. Limb kinetic apraxia usually affects the hand that is contralateral to a hemispheric lesion.

Liepmann thought that limb kinetic apraxia was induced by lesions that injured the primary motor and sensory cortex. In 1968, it was demonstrated that monkeys with lesions confined to the corticospinal system show similar errors.

B. Ideomotor Apraxia

Patients with ideomotor apraxia (IMA) make the most severe errors when asked to pantomime transitive acts to verbal command. When imitating, their performance may improve. When using actual tools their performance may improve even further, but often their performance remains impaired. Patients with IMA make primarily spatial and temporal movement errors. Spatial errors include postural, spatial trajectory, and spatial orientation. One of the most common postural errors is using a body part as a tool, despite being instructed not to use emblems of tools. When not using their body parts as tools, patients with IMA will often fail to position their hands as if they were holding the tool or object.

Whereas normal subjects will orient their hands to an imaginary target of that tool's action, patients with IMA often fail to orient their forelimbs to an imaginary target. IMA spatial trajectory errors are caused by incorrect joint movements. Apraxic patients will often stabilize a joint that they should be moving and move joints that should not be moving. Patients

with apraxia may be unable to coordinate multiple joint movements to get the desired spatial trajectory. It has been noted that patients with IMA may also make timing and speed errors.

In right-handed individuals, IMA is almost always associated with left hemisphere lesions, and in left-handed people IMA is usually associated with right hemisphere lesions. IMA is associated with lesions in a variety of structures, including the corpus callosum, the inferior parietal lobe, convexity premotor area, and the supplementary motor area. In 1907, Liepmann posited that the left hemisphere of right-handed people contains movement formulas (memories of how to make these skilled movements) or what are called praxicons, and that the callosal lesions may disconnect these movement formulas from the right hemisphere's motor areas. Years later, it was proposed that the movement representations first posited by Liepmann were stored in the left parietal lobe of right-handed people and it was demonstrated that destruction of the left parietal lobe produces both a production deficit (apraxia) and a gesture/pantomime comprehension/discrimination disorder. In contrast, apraxia induced by premotor lesions, lesions in the pathways that connect premotor areas to motor areas, and lesions in the pathways that lead to the premotor areas from the parietal lobe may also cause a production deficit. However, unlike parietal lesions, which destroy the movement representations, these more anterior lesions do not induce gesture comprehension and discrimination disorders. The medial premotor cortex or supplementary motor area (SMA) appears to play an important role in mediating skilled movements. The SMA receives projections from parietal neurons and projects to motor neurons. Several patients with left-sided medial frontal lesions that included the SMA demonstrated an IMA when tested with either arm. Unlike patients with parietal lesions, these patients could both comprehend and discriminate between well-performed and incorrectly performed pantomimes.

C. Conduction Apraxia

Patients with IMA are generally able to imitate better than they can pantomime to command. However, one patient was reported who was more impaired when imitating than when pantomiming to command. Because this patient was similar to the conduction aphasic who repeats poorly, this disorder was termed conduction apraxia. Whereas the lesions that induce conduction aphasia are usually in the supramarginal gyrus or Wernicke's area, the lesions that induce conduction apraxia are unknown. The mechanism of conduction apraxia may be similar to that of conduction aphasia, such that there is a disconnection between the portion of the left hemisphere that contains the movement representations (input praxicon) and the parts of the left hemisphere that are important for programming movements (output praxicon).

D. Dissociation Apraxia

In 1973, Heilman described patients who, when asked to pantomime to command, looked at their hand but failed to perform any recognizable actions. Unlike patients with ideomotor and conduction apraxia described previously, these patients' imitation and use of objects were flawless. In addition, when they saw the tool or object they were to pantomime, their performance was also flawless. Later, others not only reported similar patients but also other patients who had a similar defect in other modalities. For example, when asked to pantomime in response to visual or tactile stimuli, some patients were unable to do so, but these patients could pantomime to verbal command.

Patients with dissociation apraxia may have callosal lesions. In those patients with callosal lesions, it has been proposed that, whereas language was mediated by their left hemisphere, the movement representations (praxicons) were stored bilaterally. Therefore, their callosal lesion induced a dissociation apraxia only of the left hand because the verbal command could not get access to the right hemisphere's movement representations. Whereas these patients with callosal disassociation apraxia were not able to correctly carry out skilled learned movements of the left arm to command, they could imitate and use actual tools and objects with their left hand. Using actual objects and imitating does not need verbal mediation and the movement representations stored in their right hemisphere can be activated by visual input.

Dissociation apraxia may also be associated with left hemisphere lesions. These patients probably have an intrahemispheric language-movement formula, vision-movement formula, or somesthesia-movement formula dissociation such that, depending on the type

of dissociation, stimuli from one of these modalities (e.g., language) are not capable of activating the movement representations or praxicon, but stimuli in other modalities (e.g., vision) are able to activate these representations. The locations of the lesions that cause these intrahemispheric disassociation apraxias are not known.

E. Ideational Apraxia

In the early 1900s, the inability to carry out a series of acts, an ideational plan, was studied and termed ideational apraxia. When performing a task that requires a series of acts (such as making a sandwich), patients have difficulty sequencing the acts in the proper order.

It was noted that most of the patients with this type of ideational apraxia have a degenerative dementia. Frontal lobe dysfunction is also often associated with temporal order processing deficits, and the inability to correctly sequence a series of acts may be related to frontal lobe degeneration. Patients who select the wrong movement (content errors) have also been diagnosed as having ideational apraxia, but in order to avoid confusion these errors may be classified as conceptual apraxia.

F. Conceptual Apraxia

To perform a skilled act, two types of knowledge are needed—conceptual knowledge and production knowledge. Whereas dysfunction of the praxis production system induces ideomotor apraxia, defects in the knowledge needed to successfully select and use tools and objects are termed conceptual apraxia. Therefore, patients with ideomotor apraxia make production errors (e.g., spatial and temporal errors), and patients with conceptual apraxia make content and tool selection errors. Patients with conceptual apraxia may not recall the type of actions associated with specific tools, utensils, or objects (tool–object action associative knowledge) and therefore make content errors. For example, when asked to demonstrate the use of a screwdriver by either pantomiming or using the tool, the patient with a loss of tool–object action knowledge may pantomime a hammering movement or use the screwdriver as if it were a hammer. Content errors (i.e., using a tool as if it were another tool) can also be induced by an object agnosia. However, researchers have reported a patient who could name tools (and therefore did not have an agnosia) but often used these tools inappropriately. Patients with conceptual apraxia may be unable to recall which specific tool is associated with a specific object (tool–object association knowledge). For example, when shown a partially driven nail, they may select a screwdriver rather than a hammer. This conceptual defect may also be in the verbal domain such that when an actual tool is shown to a patient, the patient may be able to name it (e.g., hammer), but when a patient with conceptual apraxia is asked to name or point to a tool when its function is described, he or she may not be able to correctly point to it. Patients with conceptual apraxia may also be unable to describe the functions of tools.

Patients with conceptual apraxia may also have impaired mechanical knowledge. For example, if they are attempting to drive a nail into a piece of wood and there is no hammer available, they may select a screwdriver rather than a wrench or pliers (which are hard, heavy, and good for pounding). Mechanical knowledge is also important for tool development. Patients with conceptual apraxia may also be unable to correctly develop tools.

In 1920, Liepmann thought that conceptual knowledge was located in the caudal parietal lobe, but in 1988 researchers placed it in the temporal–parietal junction. Later, a patient was reported who was left-handed and rendered conceptually apraxic by a lesion in the right hemisphere, suggesting that both production and conceptual knowledge have lateralized representations and that such representations are contralateral to the preferred hand. Further evidence that these conceptual representations stored in the hemisphere that is contralateral to the preferred hand derives from the observation of a patient who had a callosal disconnection and demonstrated conceptual apraxia of the nonpreferred (left) hand. Researchers studying right-handed patients who had either right or left hemisphere cerebral infarctions found that conceptual apraxia was more commonly associated with left than right hemisphere injury. However, they did not find any anatomic region that appeared to be critical, suggesting that mechanical knowledge may be widely distributed in the left hemisphere of right-handed people. Although conceptual apraxia may be associated with focal brain damage, it is perhaps most commonly seen in degenerative dementia of the Alzheimer's type. It was also noted that the severity of conceptual and IMA did not always correlate. The observation that patients with IMA may not

demonstrate conceptual apraxia and patients with conceptual apraxia may not demonstrate IMA provides support for the postulate that the praxis production and praxis conceptual systems are independent. However, for normal function these two systems must interact.

See Also the Following Articles

HAND MOVEMENTS • MOTOR CONTROL • MOTOR CORTEX • MOTOR NEURON DISEASE • MOTOR SKILLS • MOVEMENT REGULATION

Suggested Reading

DeRenzi, E., and Lucchelli, F. (1988). Ideational apraxia. *Brain* **113,** 1173–1188.

Gazzaniga, M., Bogen, J., and Sperry, R. (1967). Dyspraxia following diversion of the cerebral commisures. *Arch. Neurol.* **16,** 606–612.

Geschwind, N. (1965). Disconnection syndromes in animals and man. *Brain* **88,** 237–294, 585–644.

Goodglass, H., and Kaplan, E. (1963). Disturbance of gesture and pantomime in aphasia. *Brain* **86,** 703–720.

Heilman, K. M. (1973). Ideational apraxia—A re-definition. *Brain* **96,** 861–864.

Heilman, K. M., and Rothi, L. J. G. (1985). Apraxia. In *Clinical Neuropsychology* (K. M. Heilman and E. Valenstein, Eds.). Oxford Univ. Press, New York.

Heilman, K. M., Rothi, L. J. G., and Valenstein, E. (1976). Two forms of ideomotor apraxia. *Neurology* **32,** 415–426.

Liepmann, H. (1920). Apraxia. *Erbgn der ges Med.* **1,** 516–543.

Ochipa, C., Rothi, L. J. G., and Heilman, K. M. (1989). Ideational apraxia: A deficit in tool selection and use. *Ann. Neurol.* **25,** 190–193.

Ochipa, C., Rothi, L. J. G., and Heilman, K. M. (1992). Conceptual apraxia in Alzheimer's disease. *Brain* **114,** 2593–2603.

Poizner, H., Mack, L., Verfaellie, M., Rothi, L. J. G., and Heilman, K. M. (1990). Three dimensional computer graphic analysis of apraxia. *Brain* **113,** 85–101.

Rothi, L. J. G., Mack, L., Verfaellie, M., Brown, P., and Heilman, K. M. (1988). Ideomotor apraxia: Error pattern analysis. *Aphasiology* **2,** 381–387.

Rothi, L. J. G., Ochipa, C., and Heilman, K. M. (1991). A cognitive neuropsychological model of limb praxis. *Cognitive Neuropsychol.* **8**(6), 443–458.

Watson, R. T., and Heilman, K. M. (1983). Callosal apraxia. *Brain* **106,** 391–403.

Watson, R. T., Fleet, W. S., Rothi, L. J. G., and Heilman, K. M. (1986). Apraxia and the supplementary motor area. *Arch. Neurol.* **43,** 787–792.

Area V2

DANIEL J. FELLEMAN
University of Texas, Houston Medical School

GLOSSARY

callosal connections Pathways that link the two cerebral hemispheres. In the visual system, many callosal pathways link the representations of the vertical meridian in each hemisphere. The representation of the vertical meridian, and thus the location of callosal pathways, forms the border between many cortical areas.

chemoarchitecture The study of the anatomical subdivisions of the brain based on the differential distributions of various neurotransmitters or other chemicals.

cortical area A subdivision of the cerebral cortex that is based on a combination of physiological and anatomical criteria. These criteria in the visual system include a topographic representation of visual space, a distinct cyto-, myelo-, pigmento-, or chemoarchitecture, a distinct pattern of cortical and subcortical connections, distinct receptive field properties of constituent neurons, and a distinct behavioral deficit following its lesion or deactivation.

cytoarchitecture The study of the anatomical subdivisions of the brain based on the size, shape, density, and laminar organization of neurons.

magnetic resonance imaging A collection of imaging techniques that provide either anatomical (MRI) or functional (fMRI) views of the organization of the brain.

myeloarchitecture The study of the anatomical subdivisions of the brain based on the laminar distribution and density of myelinated axons.

pigmentoarchitecture The study of the anatomical subdivisions of the brain based on the size, density, and laminar distribution of cells based on their content of pigment-containing granules.

positron emitted tomography A brain imaging technique that utilizes positron-emitting compounds (often O^{15}-labeled water) to detect functionally active brain regions during the performance of a specific task.

topographic mapping The analysis of the organization of the visual system based on the systematic representation of visual space within each cortical area. Most cortical areas adjoin each other at the representation of either the vertical or horizontal meridian. Two parameters of visual space are extracted for each recording site during topographic mapping: eccentricity and polar angle.

Area V2 is an extrastriate visual area located immediately adjacent to the primary visual cortex in the occipital lobe. This article will describe the known anatomical features of V2 and discuss physiological studies of its topographic organization and functional properties. These studies are discussed with respect to similar studies performed in nonhuman primates.

I. EARLY VIEWS OF THE ORGANIZATION OF HUMAN VISUAL CORTEX

Early views of the functional organization of visual cortex were based largely on two different types of criteria. First were behavioral criteria evaluated following brain injury to large or more restricted portions

of the occipital lobe. Complete lesions of the occipital cortex in monkeys and dogs led to "Rindeblindheit" or complete cortical blindness. However, small lesions of occipital cortex led to "Seelenblindheit" or psychic blindness, in which dogs were capable of seeing and avoiding objects but no longer recognized what they saw. By the end of the nineteenth century, similar observations were made following brain injury in humans.

Second were the anatomical criteria that included the analyses of the cytoarchitecture, myeloarchitecture, pigmentoarchitecture, and, finally, the myelogenesis of the optic radiations. The most widely recognized study of the cytoarchitecture of the occipital lobe was performed by Brodmann, who described three subdivisions of occipital cortex. He identified an area 17, largely contained within the calcarine sulcus, that was identified by a highly differentiated laminar structure in which layer 4 was subdivided into three sublaminae, IVA, IVB, and IVC. Area 17 is also known as striate cortex primarily because of the dense band of myelin, called the stria of Gennari, which is located in layer 4B. Brodmann also described two belts of cortex that surrounded area 17 that he named area 18, the occipital area, and area 19, the praeoccipital area. Unfortunately, the cytoarchitectural characteristics of areas 18 and 19 in humans were not well-specified in this study.

Von Economo and Koskinas described a tripartite organization of occipital cortex that they based on quantitative measurements of laminar widths and cell-packing density. Their areas OC (area striata), OB (area parastriate), and OA (area peristriata) correspond largely with Brodmann's areas 17, 18, and 19. Area 18 is easily recognized at the border with striate cortex (OC) and is characterized by a high cell density in layers II and IIIa, low density in layer V, and large cells in layer 6. A distinct population of very large pyramidal cells in layer IIIc of area OB characterizes the border of area OB with OC. This region, OBγ or limes parastriatus gigantocellularis, is located in the region of OB cortex that contains the representation of the vertical meridian. In macaque monkeys, this region has been shown to contain a population of pyramidal cells that are immunoreactive to SMI-32, many of which were shown to project across the corpus callosum. Furthermore, in humans lacking a corpus callosum, this population of large pyramidal cells in layer IIIc is greatly reduced. The cytoarchitectonic characteristics of area 18 (or OB) at the border with area 17 (OC) are illustrated in Fig. 1A. This border is distinct in many respects, including the changes in

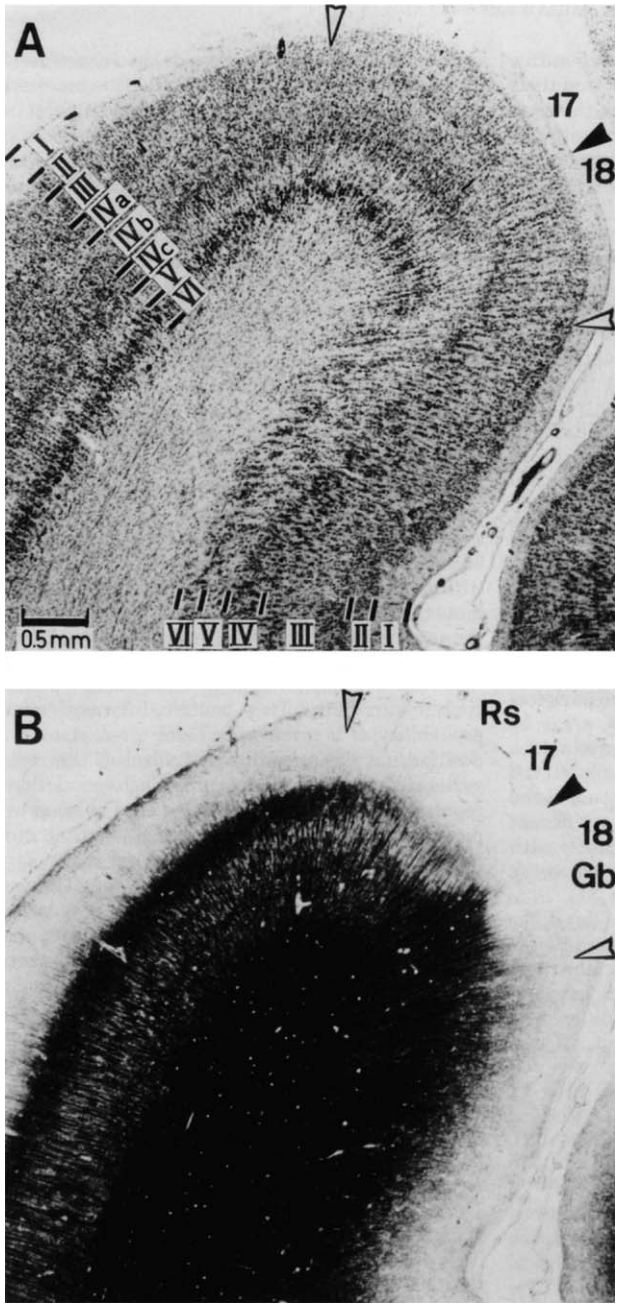


Figure 1 Architecture of the 17 (OC)–18 (OB) border. (A) Cytoarchitectonic characteristics of the 17–18 border region. Black arrow-head indicates the border. Open arrowhead in area 17 indicates the limit of the fringe area, Randsaum. Open arrowhead in area 18 indicates the limit of the border tuft region, Grenzbuschel. Layers indicated according to Brodmann. (B) Myeloarchitectonic characteristics of the 17–18 border. Area 17 is characterized by a dense band of fibers in layer IVB, the stria of Gennari. The inner band of Baillarger begins to emerge within 17 and becomes dense within area 18. A dense border tuft of radially oriented fibers is located in area 18 just across the 17–18 border. From Amunts *et al.* (2000).

width, density, and sublamination of layer IV, the density of layer VI, and increased in density in layer II. This border region is quite complex when viewed in sections stained for myelin (see Fig. 1B). Although the 17–18 border can be recognized by the change in myelin density in the infragranular layers with the emergence of the inner band of Baillarger, the myeloarchitecture of the more superficial layers is more complex. The region of area 17 just within the 17–18 border is known as the fringe area Randsaum (Rs) of Sanides and Vitzthum. Within the fringe area, the inner band of Baillarger, which is not prominent within area 17 proper, begins to emerge. On the other side of the 17–18 border, a narrow band of radially oriented fibers becomes prominent. This region is known as the border tuft, or Grenzbuschel (Gb), of Sanides and Vitzthum. This border tuft is located within von Bonin and Koskinas' area OBγ, which corresponds to the callosal recipient–origin region of OB.

The anterior border of area OB is more difficult to describe, although von Economo and Koskinas were able to use quantitative criteria to describe the overall organization and possible subdivisions within area OA (OA_1, OA_2, and OA_m). More recently, a clear separation of layers II and IIIA in 19 and a decrease in the cell density of layer IIIB in 19 have been used to distinguish area 18 from area 19.

The location and extent of Brodmann's areas 17, 18, and 19 and von Economo and Koskinas' areas OC, OB, and OA are illustrated in lateral and medial views of the human brain in Fig. 2. Similarly, the striate, parastriate, and peristriate subdivisions of occipital cortex, based on the pigmentoarchitectonic of Braak (see later discussion), are illustrated. In the medial view of the hemisphere, Brodmann's area 18 is wider than von Bonin and Koskinas' area OB. Similarly, on the lateral occipital surface, area 18 is described to occupy a $\sim$1 cm wide swath, whereas area OB barely extends onto this surface.

Although the border of area 18 with striate cortex is easily distinguished in fresh brain tissue due to the presence of the densely myelinated stria of Gennari in layer IVB and in Nissl-strained sections due to the characteristic sublamination of layer IV, the anterior border of area 18 has proven difficult to discern using traditional cytoarchitectonic criteria. However, the pigmentoarchitectonic method was able to characterize the lipofuscin distribution in striate cortex and distinguish several subdivisions of extrastriate cortex. This method identified parastriate cortex that forms a belt or horseshoe that surrounds V1 except at its most

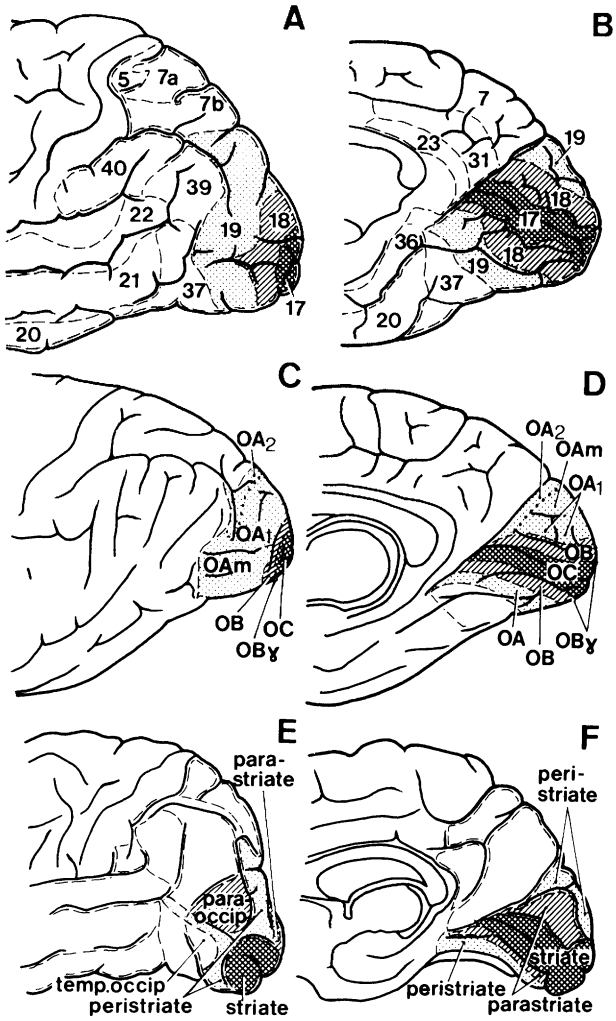


Figure 2 Cytoarchitectonic maps of the human visual cortex. (A, B) Brodmann, 1909; (C, D) von Economo and Koskinas 1925, and (E, F) myeloarchitectonic map by Eliot Smith, 1907. From Zilles and Clarke (1997).

anterior extent. The most striking characteristic of parastriate cortex was the presence of a tripartite external teania (pigmentoarchitectonic layer IV). The upper portion (pigmentoarchitectonic IVA or pIVA) is lightly stained, the middle portion (pIVB) is densely stained, and the lower portion (pIVC) is lightly stained. This pattern is distinctive, forms a precise border with striate cortex, and extends for approximately 10 mm from the V1 border. The pigmentoarchitectonic characteristics of the border between striate and parastriate regions are illustrated in Fig. 3. The parastriate region corresponds largely to area 18 of Brodmann, so it is unclear whether it only contains area V2 or whether it contains additional visual areas

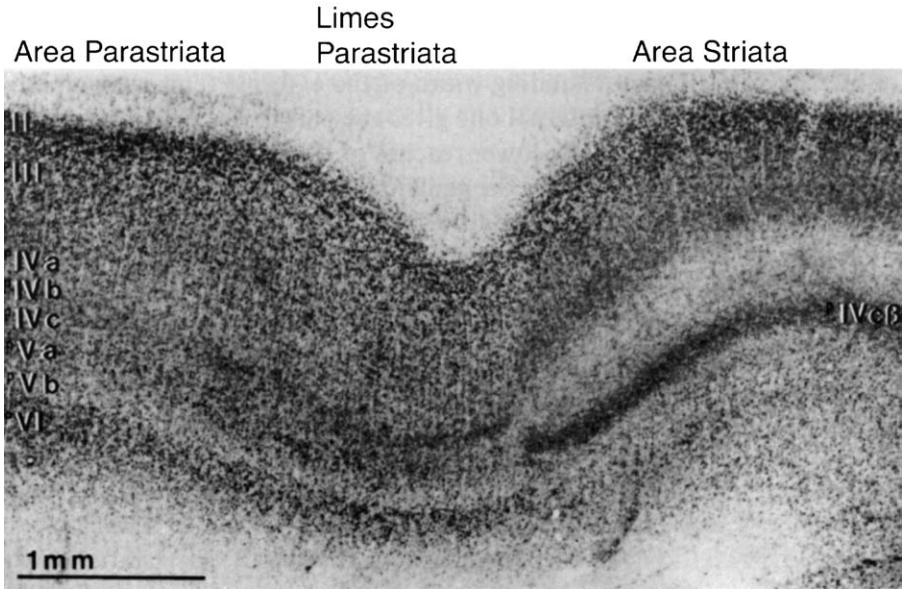


Figure 3 Pigmentoarchitectonics of the border region between area striata and area parastriata. Area striata (right) is characterized by a densely pigmented layer 4Cβ that ends abruptly at the border with area parastriata (left). Area parastriata is characterized by a tripartite layer 4 divided into pIVa, pIVb, and pIVc. The border region (limes parastriatus) contains a collection of densely pigmented cells in layers IVc and Va. From Braak (1980).

such as V3 and VP (see later discussion). In order to address this question, it will be necessary to combine this pigmentoarchitectonic approach with studies of interhemispheric connections in post mortem brain tissue or to combine *in vivo* topographic mapping with post mortem pigmentoarchitectonic analyses.

The third anatomical criterion that distinguished a tripartite subdivision of visual cortex was based on the developmental pattern of myelination of the optic radiation. Three zones were described: a projection zone that was myelinated at birth, an intermediate zone that myelinated 1 month later, and finally a terminal zone that myelinated later yet. The projection zone was shown to correspond to the area containing the stripe of Gennari, the striate cortex. The surrounding two zones together were considered to contain visual association cortex.

II. MORE RECENT VIEWS OF THE ORGANIZATION OF VISUAL CORTEX IN NONHUMAN PRIMATES

The distinction of three subdivisions of visual cortex was a widely held concept until the 1970s and 1980s, when seminal anatomical and electrophysiological mapping studies demonstrated that primate visual cortex consists of a large number of areas that are not simply belts of cortex surrounding striate cortex. They recognized that area V2 is a constant feature of organization of primate (and mammalian) visual cortex. V2 is seen as a belt of cortex that surrounds striate cortex, V1, and appears to correspond largely to area 18 of Brodmann or area OB of von Bonin. However, area 18 is somewhat larger in width than area V2; therefore, it might contain portions of areas such as V3, VP, and perhaps V3A. When electrophysiological mapping is compared with studies of interhemispheric connections, it is revealed that the border between V1 and V2 is coincident with a dense band of callosal connections that demarcates the superior and inferior vertical meridians. Anterior to this dense callosal band is a large callosal-free region that contains area V2, V3 in dorsal cortex, and VP in ventral cortex. A second dense band of callosal connections is observed along the anterior borders of areas V3 and VP, again coincident with the representation of the inferior and superior vertical meridians, respectively. Therefore, the representation of the horizontal meridian is located within this callosal-free zone because areas V2, V3, and VP contain split representations of the horizontal meridian, which form their common border. Similarly, the mapping

studies indicated that the cortex that makes up Brodmann's area 19 contains a number of distinct visual areas rather than a single third-tier area. Specifically, area MT is separated from area V2 by at least one other cortical area, V4 (or DL, the dorsolateral visual area).

Several convergent criteria are generally used to identify cortical subdivisions. Traditionally, a distinctive cyto- or myeloarchitecture has been used to identify specific cortical subdivisions. Second, a representation of visual space, a cortical map, is usually identified with a visual area. Third, areas are defined by a unique set of receptive field properties within their population. Fourth, areas are defined by a specific behavioral loss following their lesion or temporary inactivation. Finally, areas can be defined by a unique pattern of corticocortical and/or cortical–fugal connections.

V2 in nonhuman primates has been characterized by most, if not all, of these criteria. For example, V2 in macaque monkeys has a distinct appearance, most notably its cytochrome oxidase dense stripes. These appear as repeating thin and thick stripes that are separated by cytochrome oxidase pale zones. The distinction of thick and thin stripes is more difficult in macaques than in squirrel monkeys, but the thick stripes can be distinguished by their high immunoreactivity to the CAT-301 antibody. V2 contains an overall representation of the contralateral visual hemifield with the representation of the vertical meridian forming the posterior border with V1 and the representation of the horizontal meridian forming the anterior border with area V3 in dorsal cortex and VP (V3v) in ventral cortex. Furthermore, each of the cytochrome oxidase stripe compartments (thin, pale, and thick) contains a separate representation of the visual hemifield. Macaque V2 contains neurons with a wide range of stimulus specificities. A high proportion of cells located within the cytochrome oxidase thin stripes demonstrate selectivity for stimulus color and lack orientation selectivity, whereas a high proportion of cells located in the pale and thick stripes demonstrate orientation selectivity. Some cells located within the pale and thick stripes are also selective for illusory contour stimuli. In addition, a substantial proportion of V2 cells in awake fixating macaque monkeys exhibit selectivity for complex visual features such as noncartesian gratings.

Macaque V2 makes connections with a variety of cortical areas or modules within areas. The cytochrome oxidase thin stripes of V2 receive a preferential input arising largely from the cytochrome oxidase blobs of V1. Similarly, the cytochrome oxidase pale stripes receive feedforward input from the interblobs of V1. Finally, the V2 thick stripes receive feedforward input almost exclusively from layer IVB of V1. Thick stripes make feedforward connections primarily with areas V3 and MT. V2 makes feedback connections to each of these V1 compartments. Thin stripes and interstripes make dense projections that remain largely segregated into discreet zones of area V4. Areas V3A and VP receive input from V2 compartments that have not been identified. Finally, V2 makes weak projections to area TEO in posterior inferotemporal cortex. Each of these higher areas makes feedback projections to V2, demonstrating that V2 is located at the second level within a complex cortical hierarchy that contains reciprocal pathways.

The behavioral and perceptual contributions of macaque area V2 have been difficult to assess primarily due to the difficulty in limiting lesions to V2 and the need for appropriate controls. These limitations were overcome in one study that made ibotenic acid lesions of parafoveal lower field V2 and compared them to similar lesions in area V1 of fixating, behaving monkeys. In contrast to area V1, lesions of V2 led to no appreciable deficits in acuity or contrast sensitivity. V2 lesions did lead to significant deficits in orientation discrimination of lines made of collinear dots embedded in noise or of texture elements in an array. These results indicate that V2 is not needed for some low-level discriminations but may be essential for tasks involving more complex spatial discriminations, especially those that involve segregating features from a noisy background.

III. ANATOMICAL CHARACTERIZATION OF HUMAN AREA V2

A. Callosal Connections Identify the Borders between Human Visual Areas

Clarke and Mikossy studied the organization of human occipital cortex using cyto- and myeloarchitecture and their relationships to callosal connections. In addition to the easily identified cytoarchitectonic border of V1 with V2, this border is demarcated by a dense band of callosal connections that is bordered anteriorly by a large acallosal zone (see Fig. 4). On the basis of studies in macaque monkeys, these acallosal zones are thought to contain the lower and upper field representations of area V2, as well as the lower field

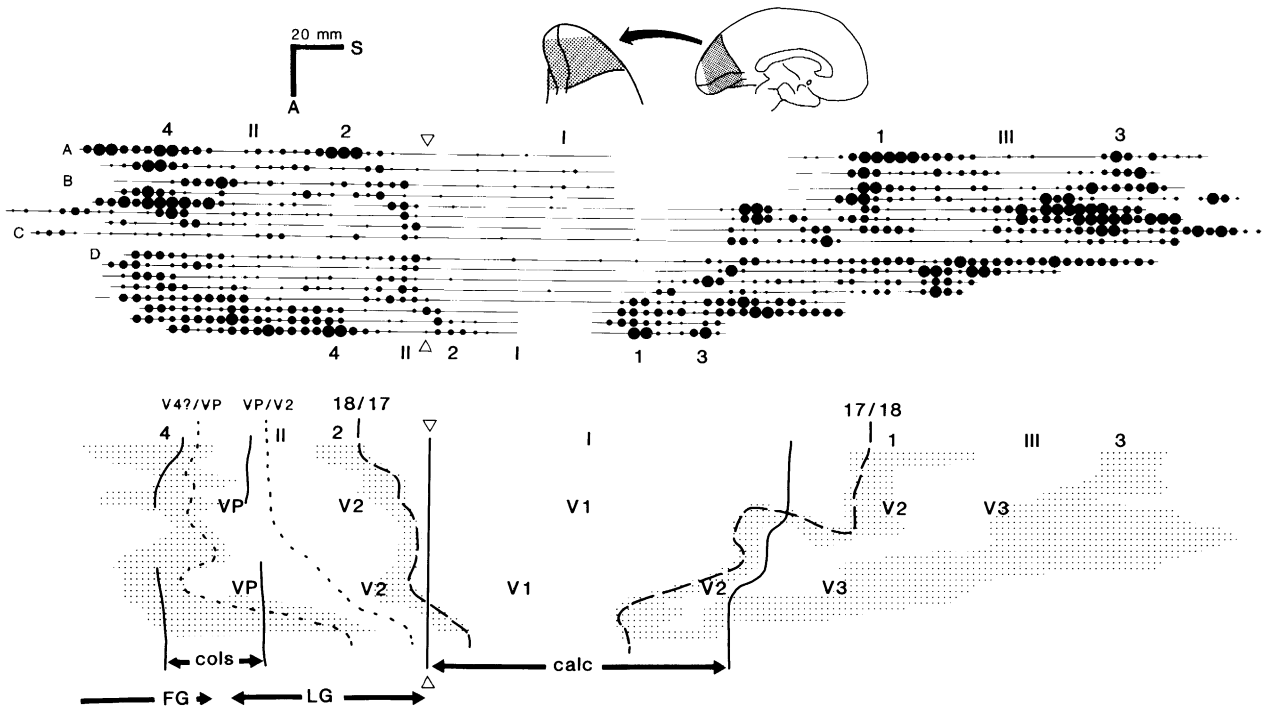


Figure 4 (Top) Flat reconstruction of the medial occipital lobe illustrating the tangential distribution of callosal afferents within the gray matter. The reconstructed region corresponds to the shaded portion of the brain insets and also includes cortex buried within the calcarine and other sulci. Sixteen sections are reconstructed with the callosal afferents projected to a point halfway between the pial surface and white matter. The density of degenerating fibers is represented by the different sizes of filled circles. (Bottom) Stippling shows dense regions of callosal afferents. The large callosal-free region labeled I corresponds to area V1. The 17–18 border (heavy dashed line) is marked to a narrow dense band of callosal afferents (regions 1 and 2). A second callosal-free zone is located anterior to region 2 in ventral cortex (region II) and anterior to region 1 in dorsal cortex (region III). These callosal-free zones is contain representations of the horizontal meridian. Anterior to the callosal-free zones is a second dense band of callosal afferents found in ventral cortex (region 4) and in dorsal cortex (region 3). These anterior bands correspond to a second representation of the vertical meridian. The ventral callosal-free zone contains areas V2 and VP and is located within area 18. Similar arguments can be made for areas V3 and V3A in dorsal extrastriate cortex. From Clarke and Miklossy (1990).

representation in area V3 and the upper field representation in area VP. A second dense callosal band marks the anterior border of this acallosal zone, just as is observed at the anterior borders of areas V3 and VP in macaque monkeys.

The border between V2 and presumably V3 in dorsal occipital cortex was difficult to identify on the basis of cytoarchitecture or myeloarchitecture, although macaque V3 has a distinctive pattern of dense myeloarchitecture. In contrast, the ventral acallosal zone can be subdivided into densely myelinated V2 posteriorly and poorly myelinated VP located anteriorly. This border also corresponds to a decrease in CO activity in VP as compared to area V2. Therefore, area V2 is contained within cytoarchitectonic area OB of von Economo, whereas area VP is contained with cortex of the OA_1 subtype of area peristriata. On the basis of the published map of callosal afferents, area V2 ranges from approximately 10 mm to more than 20 mm in width.

B. Quantitative Cyto- and Myeloarchitecture of V2

In an effort to provide quantitative criteria for the distinction of area V2 from other cortical areas, image-processing techniques have been used to assess the cytoarchitectonic and myeloarchitectonic organizations of area V2 and to compare them with those from areas V1 and V4. According to this approach, Nissl-stained sections are digitized and the profiles of cell bodies are segmented according to gray level. A gray level index is then used to quantify the areal fraction of cortex occupied by stained cell bodies. Figure 5A illustrates the gray level index in areas V1, V2, and V4.

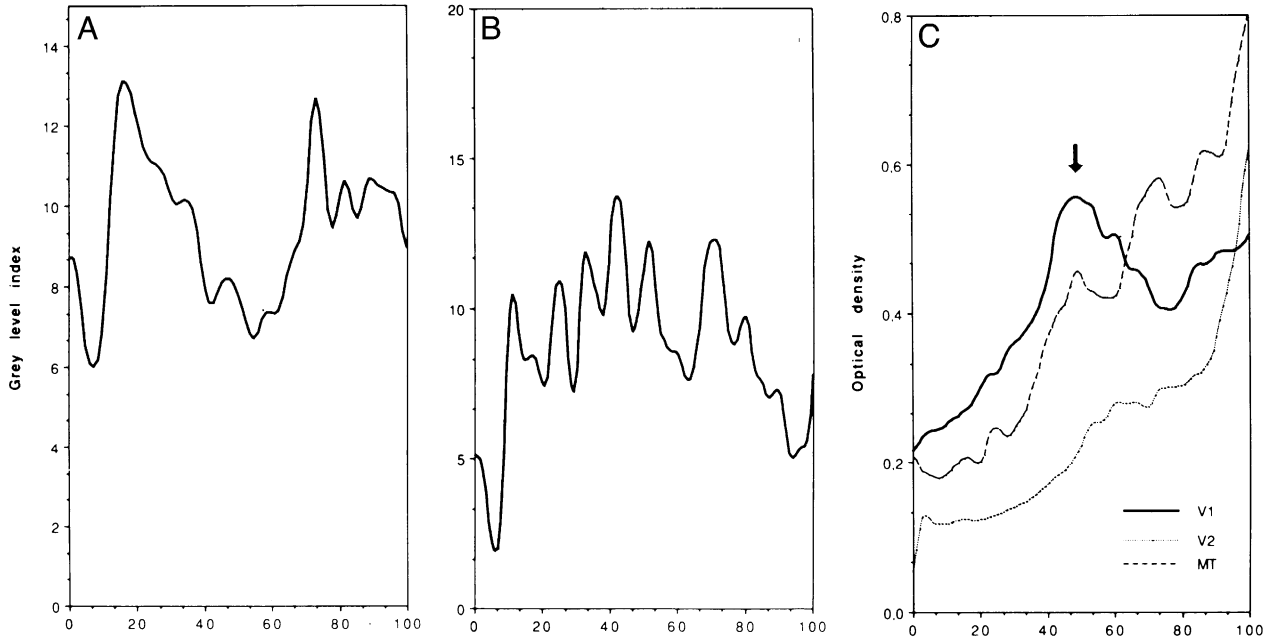


Figure 5 Distribution of the gray level index (GLI) in areas V1 (1) and V2 (2). The differing profiles indicate the different laminar patterns of cytoarchitectonic features in these two areas. Distribution of myelin densities in areas V1, V2, and MT. From Zilles and Schleicher (1993).

V1 is characterized by a bimodal GLI that peaks in layers II–III, drops in layer IVB, peaks again in layer IVC, and finally falls off in layers V and VI. In area 18 (V2, just across the V1 border), a different pattern is observed with multiple peaks extending across the depth of the cortex and no distinctive cell-sparse zone as observed in layer IVB of V1. Finally, a similar multipeaked pattern was observed in area V4 of the fusiform gyrus.

Similar quantitative techniques have been used to quantify the distribution of myelin in sections of areas V1, V2, and MT stained by the Gallyas method for myelin. All three areas show an increase in the density of myelin from the upper to the lower cortical layers. In V1, a distinct peak is found in the middle of cortex corresponding to the stria of Gennari. V2 can easily be distinguished from V1 by the overall lower density of myelin and the clear lack of the stria of Gennari (see Fig. 5B).

Quantitative cytoarchitectonic and myeloarchitectonic methods have been used to determine the variability of areas 17 and 18 when transformed into the Talairach stereotaxic coordinate system. For example, five male and five female brains were obtained at autopsy, fixed in 4% formalin for several months, and imaged with MRI at high resolution. The brains were then cut at 20 μm and sections were analyzed quantitatively every 1200 μm. The borders between areas 17 and 18 and between areas 18 and 19 were determined using the gray level index to quantify neuronal density across cortical layers. These histological borders were then transferred to MRI sections using both linear and nonlinear fluid transformations. Finally, the MRI-reconstructed brains were transformed into a standard format. Figure 6 illustrates the surface rendering of areas 17 and 18 on sagittal views of the MRI reconstructions of the 10 brains used in this study. It is apparent that area 18 is highly variable in location and extent across these 10 brains. This variability can be expressed by the center of gravity and standard deviation of the Talairach coordinates for area 18. Thus, area 18 has a maximum variability of the center of mass of 5 mm in the z plane, and this variability is almost half as large in the x plane.

C. Cortical Connections of Area V2

The organization of corticocortical connections between areas V1 and V2 in human visual cortex has been studied using the neuronal tracer DiI (1,1′-dioctadecyl-3,3,3′,3′-tetramethylindocarbocyanine perchlorate) in aldehyde-fixed post mortem brain tissue. DiI injections

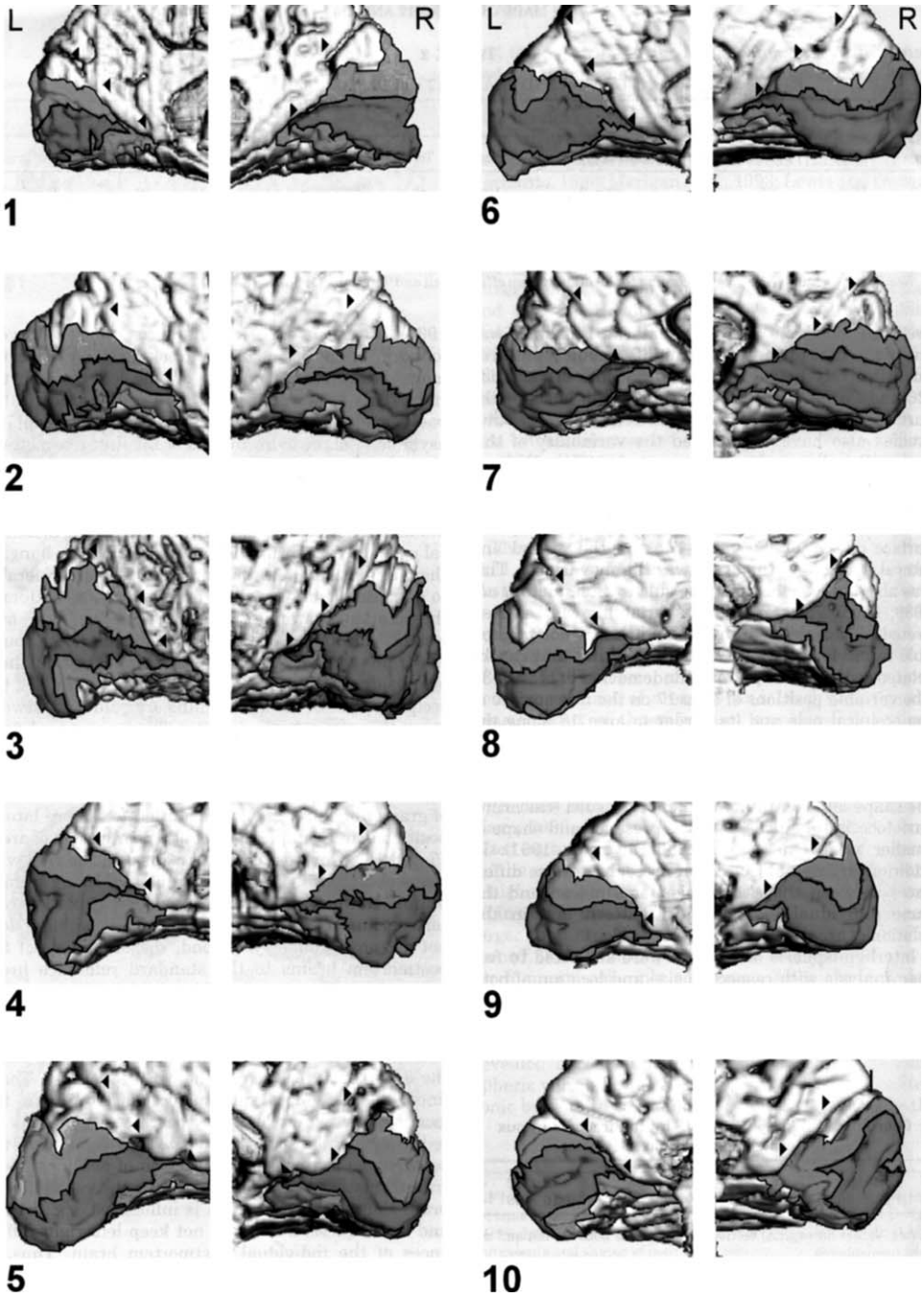


Figure 6 Surface rendering of the two hemispheres from 10 brains illustrating the variability in the locations of areas 17 and 18. From Amunts *et al.* (2000).

were placed either in area V1, identified by the presence of the stria of Gennari and cytochrome oxidase blobs, or in area V2, identified by the presence of cytochrome oxidase stripes. Following V1 injections, labeled fibers enter the white matter and then ascend into the gray matter of area V2. These fibers then either terminate in layers III and IV or travel horizontally through the gray matter to terminate a short distance away in V2 (Fig. 7A). Injections of DiI into V2 led to a different pattern of fiber projections. Rather than coursing through the white matter underlying V1, fibers coursed through the gray matter to terminate in layers I, II, III,

IVB, V, and VI and avoided terminating in layers IVA and IVC (Fig. 7B). The projections to layers IVB and V were highly clustered, forming 0.3-mm-wide clusters with a spacing of 0.6–1 mm. Some of the V2 injections failed to produce labeling of layer IVB. Presumably these V2 injections failed to hit the V2 thick stripes so the projections to layer IVB would not have been expected (Fig. 7C).

DiI injections into area V2 also produce clustered intrinsic projections. Figure 7D illustrates a transverse section throughV2 with an injection site in the superficial layers and clustered projections extending through layers II–V. The spacing of these intrinsic projections is approximately 1 mm, which is smaller than the distance between functional modules identified through cytochrome oxidase histochemistry (see later discussion).

The development of these feedforward and feedback connections between areas V1 and V2 has been studied using the DiI technique in post mortem brains ranging in age from 37 weeks of gestation to 4 postnatal months. At 37 weeks of gestation, injections to area V2 led to the labeling of fibers that originated only from deep cortical layers of V2 and ran through and perhaps terminated in layers V and VI of area V1. At 9 postnatal days, injections of V1 label fibers that are restricted to the deep layers of V2. No fibers were seen to terminate in superficial layers, but a few fibers ended in growth cones in layer IV. Injections of V2 produced both retrogradely labeled cells and anterogradely labeled fibers. Retrogradely labeled cells were observed in layers IVB and VI but not in supragranular layers II and III. Thus, feedforward projections first mature from layers IVB and VI, and the supragranular

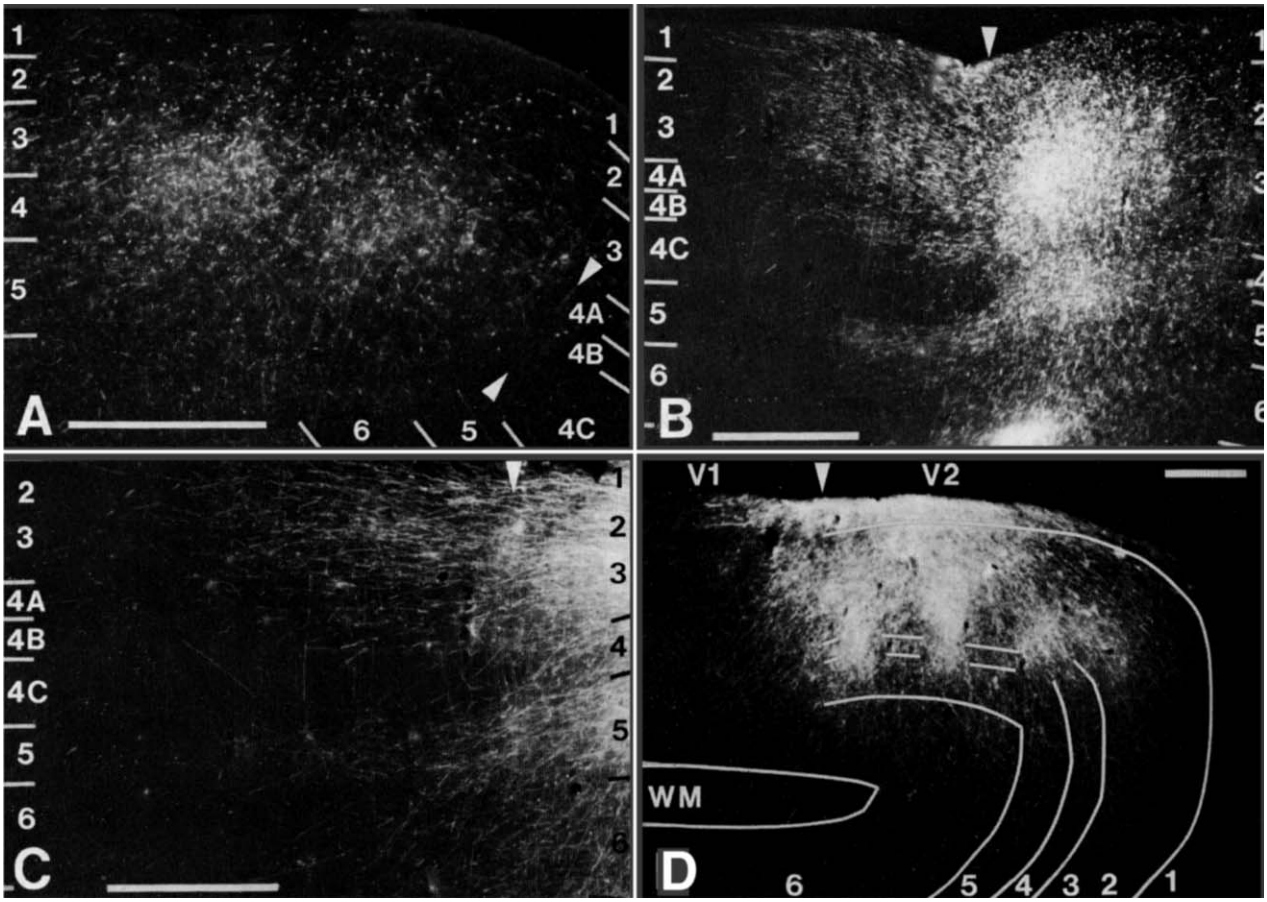


Figure 7 Extrinsic and intrinsic connections of area V2. (A). Transverse section through V1 and V2 showing clustered terminations in layers III and IV of V2 following an injection of DiI into V1. (B) Projections from V2 to V1. DiI injection was made into V2 (right of arrowhead), and labeled fibers and terminals are observed in layers I, II, III, IVB, and V of V1. (C) Projections from V2 to V1 that avoid layer IVB. DiI injection in V2 is located at the right of the arrowhead. Fibers are seen coursing from the injection site to terminate in layers I, II, III, and V of V1. (D) Intrinsic connections of V2. Transverse section showing projections following DiI injection into superficial layers of V2. Clustered projections extend from the injection site to terminate in layers II–V. Bar=1 mm. From Burkhalter and Bernardo (1989).

projections, characteristic of adults, are not yet formed by 9 postnatal days. At 9 postnatal days, feedback projections from V2 to V1 are characterized by horizontally oriented fibers in deep layers V and VI, which contain vertical branches that reach into layer IVB. Some feedback fibers were observed in supragranular layers, but these were usually obliquely oriented fibers that ramify in layer I.

At 7 postnatal weeks, feedforward projections from V1 to V2 begin to show the adult pattern of labeling. That is, labeled fibers enter V2 not only from the deep cortical layers but also from the superficial layers and begin to terminate in layers III and IV. In contrast, the feedback projections from V2 to V1 remain largely immature at 7 postnatal weeks. Feedback fibers terminate in layers I, IVB, V, and VI, just as was observed at 9 postnatal days. Fibers were still not observed in layers II and III, as in the adult. Nevertheless, there was some development of these feedback pathways in that more fibers were observed coursing through layers II and III to terminate in layer I. In addition, fibers in layer IVB were often oriented horizontally and some innervation of layer IVA was observed.

At 4 postnatal months, the projection from V1 to layer IV of area V2 became denser. Feedforward projections from V1 to V2 begin to take on the adult appearance at 4 postnatal months. At this time, injections of V2 produced retrograde labeling of neurons in layers II, III, IVB, V, and VI. In contrast, feedback projections from V2 to V1 are still immature at this time. Fibers were seen to terminate densely in layers I, IVB, V, and VI, but the projections to layers II and III remain weak at this time. Thus, at 4 postnatal months, the feedforward projections from V1 to V2 appear more mature than the feedback projections from V2 to V1. At 2 years of age, feedforward projections from V1 terminate in layers III and IV of V2. This projection is similar to that observed at 4 postnatal months but is more dense. Similarly, the feedback projection from V2 terminated densely in layers II, III, IVB, and V of area V1 and more weakly in layers I and VI. Similar to adults, this feedback projection was highly clustered with $\sim$0.3-mm-wide clusters separated by $\sim$0.3-mm-wide gaps.

The development of long-range *local* connections within areas V1 and V2 has been studied using the DiI technique in fixed prenatal and postnatal post mortem brains. At 37 weeks of gestation, injections of V2 give rise to labeled fibers in layers II, III, and V that extend horizontally for several millimeters. By 9 postnatal days, these fibers become denser and tend to form irregular clusters close to the injection site. In contrast, at 4 postnatal months, local connections within V1 are largely immature. Injections of V1 give rise to horizontally oriented fibers within layers IVB, V, and VI, whereas the labeling of layers I, II, and III is sparse. Thus, although V1 layer II–III neurons begin to project to area V2 by 4 postnatal months, they have yet to form the adult pattern of local horizontal connections within V1.

The intracortical connections of V2 with a more rostral region of human extrastriate cortex have been studied in post mortem brain tissue following a naturally occurring infarct in area 19. A $\sim$1-cm-wide infarct was found in the superior lateral extrastriate cortex of a 92-year-old female who died of natural causes. The brain was recovered shortly after death and was later processed for degenerating axons. On the basis of previous studies of the interhemispheric connections of extrastriate cortex in humans, the infarcted region is located anterior to area V3A in a region that perhaps corresponds to area V6 of macaque monkeys. Dense regions of degenerating axons were found within 2–5 mm of the infarct and were distributed to more distant cortical regions, including areas V1, V2, V3, VP, V3A, and V4. Dense projections were seen in superior portions of V1, V2 and V4 and in areas V3 and V3A, whereas weaker projections were seen in inferior portions of V1, V2, V4, and in area VP. These results suggest that the infarct affected cortex that represented the inferior visual field, yet appeared to have extended into cortex that contained a representation of the superior visual field. In V2, the degenerating fibers were localized primarily in the infragranular layers, but, in its densest region, extended into the middle cortical layers. This pattern of labeling is suggestive of a feedback projection from the infarct to area V2. A similar laminar pattern of labeling was also observed in V1 and in extrastriate areas V3, VP, V4, V3A, and V5.

D. Evidence for a Modular Organization of Human Area V2

The distinction of area V2 from V1 and the identification of functional subunits in both of these areas have been facilitated greatly through the examination of the differential distribution of cytochrome oxidase histochemistry in nonhuman primates. Cytochrome oxidase blobs or puffs have been identified in supragranular striate cortex that represents regions of higher degree metabolic activity than the surrounding

interblob zones. Subsequent investigations using microelectrode, 2-deoxyglucose autoradiography and optical recording techniques have linked these different anatomical modules within V1 with distinct physiological properties. Cells in V1 blobs tend to be monocular, express a lack of orientation selectivity, and possess a higher degree of color selectivity. In addition, V1 blobs tend to be located at the center of ocular dominance stripes.

V2 is also characterized by a unique pattern of cytochrome oxidase histochemistry in a wide range of nonhuman primate species. In V2, increased densities of CO are observed to form elongated stripes that are oriented perpendicular to the V1–V2 border. In squirrel monkeys, these stripes can readily be observed to alternate in a regular thick–thin pattern, separated by intervening interstripe zones of low CO activity. The alternation in stripe width as well as the overall regularity of stripe alternation is less obvious in macaque monkeys as well as in humans (see later discussion).

A periodic pattern of cytochrome oxidase activity in human visual area V2 was first described in 1984. This pattern was similar to that observed in macaque monkeys but differed in size, periodicity, and overall homogeneity. Like nonhuman primates, area V2 can be identified by the presence of CO-dense stripes in layers III–V (see Fig. 8). These stripes form a regular pattern of dense stripes that are oriented perpendicular to the V1 border. The dense stripes are 1.0–2.75 mm in width and are separated by CO-pale zones that are 1.5–2.75 mm in width.

Unlike nonhuman primates, human V2 CO stripes are more globular than homogeneous in structure. These globular zones are 580–1400 μm in width and form stripes that have a spacing of 1500–2300 μm. These CO density dimensions were reported to be rather variable across individuals. Interestingly, CO stripes in V2 can be observed in neonatal tissue at a time when CO blobs are not present in area V1. Furthermore, although the CO densities formed rough stripes oriented perpendicular to the V1 border, the demarcation of thick and thin stripes, which is readily observed in squirrel monkeys and less so in macaque monkeys, generally is not possible in humans. However, it is possible to distinguish these stripes from each other in humans through the use of CAT-301 immunoreactivity.

CAT-301 is a cell-surface proteoglycan that is observed at high density in several levels of the macaque visual system that are related to the magnocellular thalamocortical pathway. The distribution of

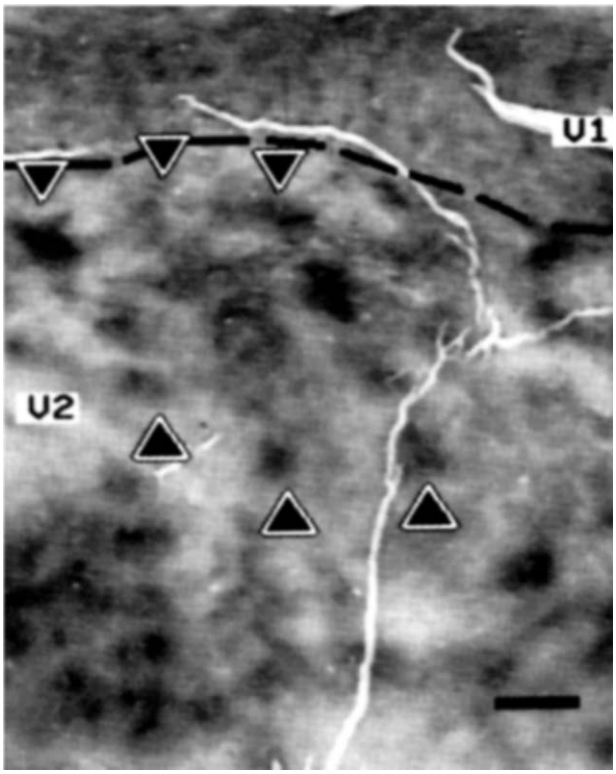


Figure 8 Cytochrome oxidase stripes in human V2. Three prominent stripes are seen in this section through layer 4. Individual stripes consist of dense patches separated by pale zones that are aligned in rows. Bar=2 mm. From Tootell *et al*. (1993).

CAT-301 immunoreactivity has been observed in cross-sectional and tangential tissue blocks of human visual cortex. Labeled cells in V2 are distributed in two distinct bands: one located in layer 3 and a second in layer 5 (see Fig. 9A). The periodicity of CAT-301 is best revealed in tangentially sectioned tissue. CAT-301 stained dense clusters of neurons in V2, which formed stripes that were 1–3 mm wide and 1.5–3 cm long and oriented perpendicular to the V1–V2 border (see Fig. 9B). These dense stripes were separated by approximately 4–8 mm. Because CAT-301 immunoreactivity identifies the CO-dense thick stripes of macaque area V2, it seems likely that this CAT-301-ir in human V2 also recognizes thick stripes. It is important to point out that these CAT-301-dense regions appear distinctly stripelike, whereas studies of V2 using CO tend to identify CO-dense clusters rather than stripes.

In addition to cytochrome oxidase and CAT-301 staining, myelin staining of unfolded, tangentially sectioned portions of V2 with the Gallyas stain reveals stripes that are 6–8 mm apart. This spacing is consistent with the labeling of one set of stripes, either

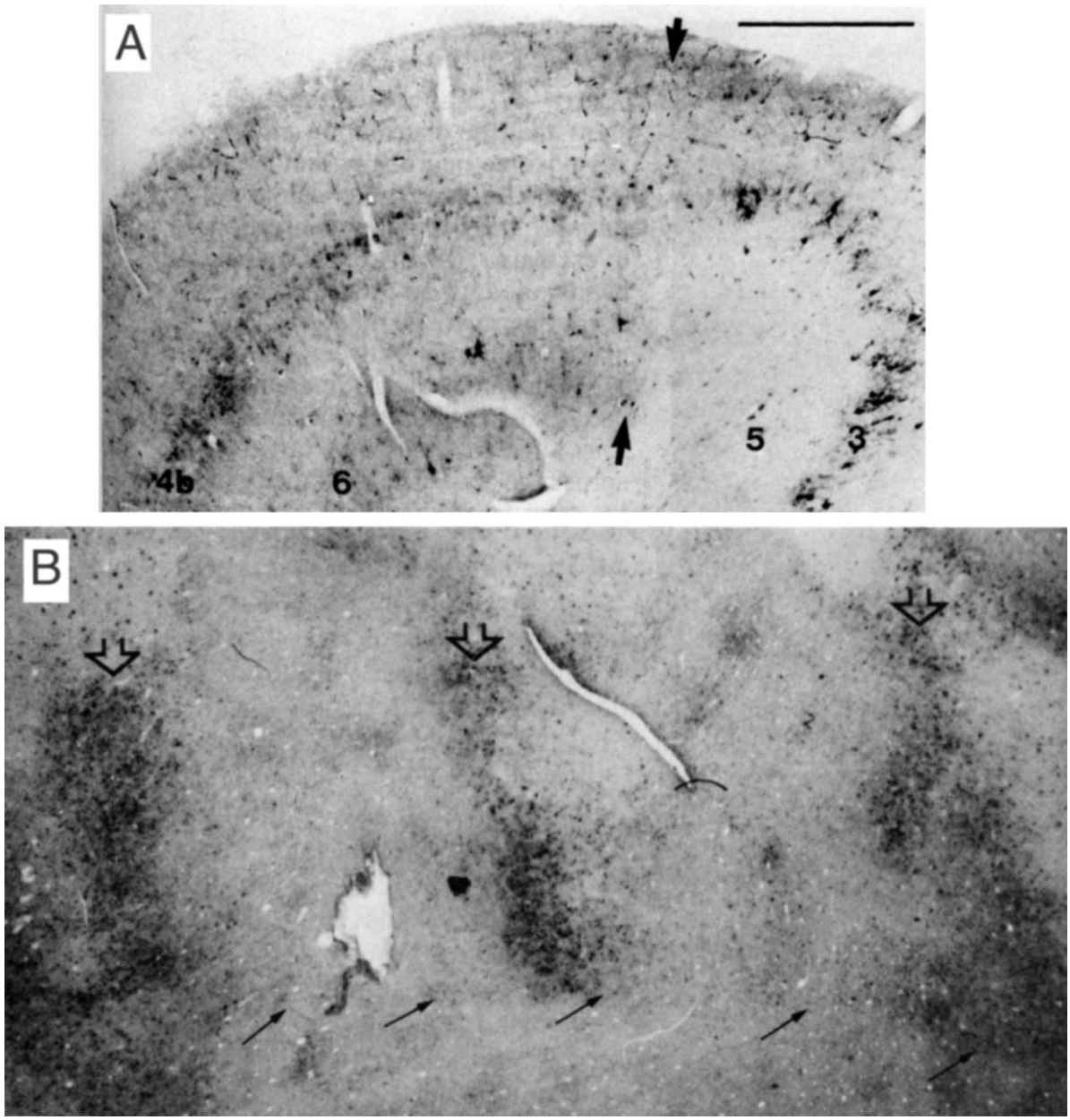


Figure 9 CAT-301 immunoreactivity in human area V2. (A) Laminar distribution of CAT-301-ir neurons in V1 (left) and V2 (right). V1 contains two bands of labeled cells corresponding to layers IV and VI. The laminar pattern changes at the V2 border, where CAT-301-ir cells are located in a dense band in layer III and a lighter band in layer V. Arrows indicate the V1–V2 border. Bar=1 mm. (B) Tangential distribution of CAT-301-ir cells in area V2. Thick bands of CAT-301-ir cells run orthogonal to the V1–V2 border. Bar=1 mm. From Hockfield *et al.* (1990).

thin or thick stripes. Alternatively, the myelin staining may reflect the pale CO compartments that have been reported to contain higher myelin densities in squirrel monkey V2. The pattern of myelin staining in unfolded V2 is illustrated in Fig. 10.

The distinctiveness and regularity of CO densities end at the anterior border of V2 and allow for an assessment of the overall size of V2 in humans. At the foveal representation, V2 is relatively narrow at 1–1.5 cm wide. In both dorsal and ventral cortex, the width

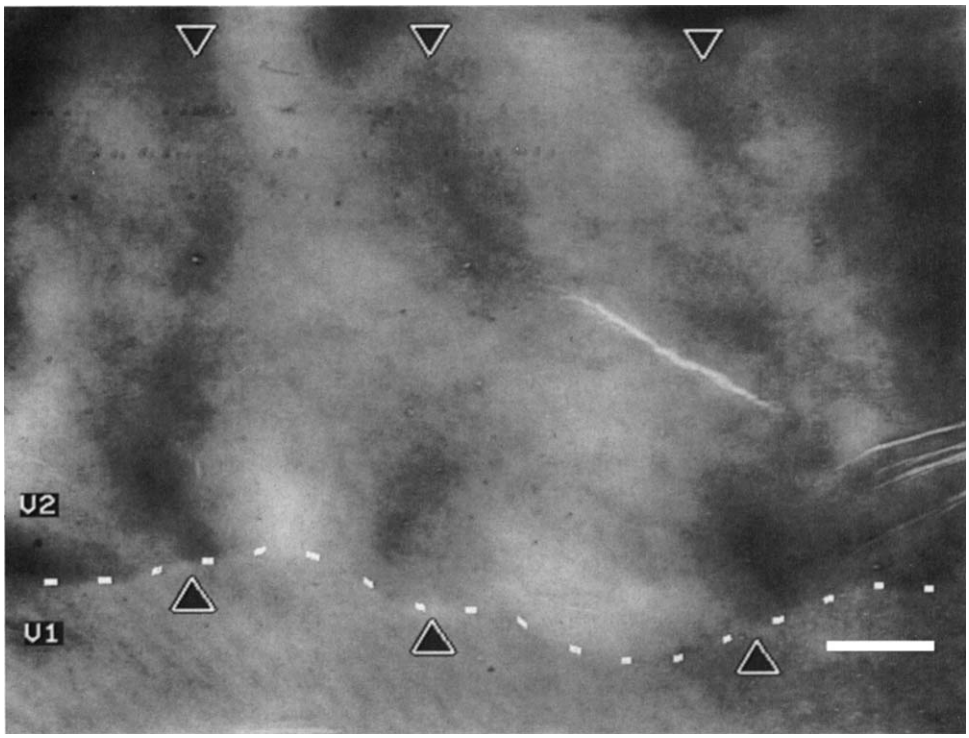


Figure 10 Myelin staining in human area V2. Composite image of four aligned sections spanning layers 3–5. Three myelin dense stripes are observed running orthogonal to the V1–V2 border (dashed line). Bar=2 mm. From Tootell and Taylor (1995).

of V2 increased to 1.8–3.4 cm at parafoveal eccentricities. These features of the modular organization of area V2 are summarized in Table I.

IV. NEUROCHEMICAL CHARACTERIZATION OF HUMAN AREA V2

The distinctive border between areas V1 and V2 has facilitated greatly the determination of the laminar distributions of neurotransmitter receptors in area V2 of humans. [^{3}H] Glutamate and NMDA receptors (labeled by [^{3}H]TCP, [^{3}H]glycine, or [^{3}H]dizocilpine binding) are densely distributed in layer IVC of V1 but are not densely distributed in layer IV of V2. In contrast, kainate receptors (determined through binding of [^{3}H]kainate) are not observed in layer IVC of V1 but are labeled in layer IV of V2. In addition, dense labeling of NMDA and AMPA receptors is observed in layer III of V2, whereas weaker labeling is observed in layers III–IVA of V1. Similarly, weak labeling of kainate receptors is observed in layer III of V2, whereas no labeling is observed in layer III of V1. Finally, whereas dense labeling of kainate receptors is observed in layers V and VI in both areas V1 and V2, weak AMPA receptor labeling is observed in layer V of V1 that is not observed in layer V of V2. These results are summarized in Table II.

In addition to the distribution of glutamate receptors, the distribution of $GABA_A$ receptors distinguishes area V2 from area V1. $GABA_A$ receptors are distributed in a bimodal pattern in V1 and a unimodal pattern in V2. This is due to a lack of $GABA_A$ receptor binding in layer IVB and weak binding in layer III of V1, whereas layers II–IV of V2 show dense $GABA_A$ binding.

Acetylcholine receptors have different laminar distributions in areas V1 and V2. Nicotinic receptors, determined through the binding of [^{3}H]oxotremorine, have their highest densities in layers V and VI in both areas V1 and V2. In V1, binding sites are also apparent in layer IVC, whereas in V2, binding sites are observed in layers I and II. Muscarinic, M_1 receptors, determined through the binding of [^{3}H]*N*-methylscopolamine, show their highest densities in layers II and III of

Table I
Modular Organization of V2[a]

Structure	Size (mm)
Width of dense CO stripes	1–2.75
	0.5–0.7
Width of CO interstripes	1.5–2.75
	0.5–0.7
Diameter of "globular" CO zones	0.580–1.4
Distance between globular CO zones within stripes	0.85–1.85
Distance between CO stripes	1.5–2.3
Length of CO stripes	15–25, 10–15 foveal, 18–34 intermediate, eccentricity
Width of CAT-301 stripes	1–3
Distance between CAT-301 thick stripes	4–8
Length of CAT-301 stripes	15–30
Width of myelin-dense stripes	~2
Distance between myelin-dense stripes	6–8
Length of myelin-dense stripes	12+
Cycle width between like stripes	6–8

[a]This table summarizes the modular features of human area V2 revealed through anatomical methods. An alternating dense and pale stripelike pattern is revealed when V2 is processed for the visualization of cytochrome oxidase (CO). This anatomical feature allows for a comparison of the organization of human V2 with that of nonhuman primates. The length of these cytochrome oxidase stripes provides an anatomical marker for the width of area V2. The pattern of immunoreactivity for CAT-301-provides a means for distinguishing the CAT-301-dense CO thick stripes from the CAT-301-pale CO thin stripes. V2 is also characterized by an array of myelin-dense stripes that may correspond to either the CO thin or thick stripes.

Table II
Neurotransmitter Receptor Distribution in Human V2[a]

Receptor	I	II	III	IV	V	VI
NMDA	–		++	–	–	–
AMPA	+	++	++	–	–	–
Kainate	–	++	+	+	++	++
$GABA_A$	–	++	++	++	–	–
Muscarinic	+	++	++	++	–	–
Muscarinic M1	–	++	++	++	–	–
Muscarinic M2	–	++	++	++	–	+
Nicotinic	+	+	–	–	++	++
5-HT1A	++	++	+	–	–	–
5-HT2	–	+	++	+	++	–
Adrenergic α_1	+	++	+	–	+	–
Adrenergic α_2	–	+	–	+	–	–

[a]Distribution of glutaminergic, gabaergic, cholinergic, serotonergic, and adrenergic receptors as a function of cortical layer in human V2.

both areas V1 and V2. In V1, weaker binding is observed in layers IVC and V, whereas in V2 a high density of binding is observed in layer IV. Muscarinic, M_2 receptors, visualized through the binding of [^{3}H] pirenzepine, show a high density in layers II and III of both areas V1 and V2. A high density of receptors is seen in layer IVC of V1 and in layer IV of V2. Weaker binding is observed in layer V of area V1 and in layer VI of V2.

Adrenergic receptors also show a differential regional pattern between areas V1 and V2. Overall, α_1 receptors have a higher mean density in area V2 than in V1. This uneven distribution is largely due to the high density of α_{1B} receptors in area V2. The peak α_1 density is found in layers II and III, with lower densities in layers I, IV, V, and VI. In contrast, the α_{2A} receptors show a higher density in area V1 than in V2. This difference is largely due to the high density of α_{2A} receptors in area V1. In V2, the highest density of α_{2A} receptors is found in layers II, V, and VI, peaking in layer VI.

The distribution of serotonergic receptors can also be used to distinguish area V2 from area V1. 5-HT_1 binding sites have a higher density in area V1 than in V2. 5-HT_{1A} binding sites have a similar laminar distribution in both areas, peaking in layers I and II. 5-HT_2 binding sites, determined through [^{3}H]keranserin autoradiography, have a high density in layer 4C of V1, and layers II and V have the higher density in area V2. 5-HT_{2A} receptor mRNA has its highest density in layer IVC of area V1, lower densities in layers III and II, and lowest densities in layers IVB, V, and VI. In V2, 5-HT_{2A} receptors show their highest densities in layers III and V, and lower densities were found in layers II and IV.

The distribution of nicotinamide adenine dinucleotide phosphate-diaphorase (NADPH-d) immunoreactivity has been used to distinguish area V2 from area V1. NADPH-d immunohistochemistry labels both the neurophil and several sizes of neurons in both the gray and white matter. In V1, NADPH-d activity in the neuropil is most dense in layer IV, and regions of high activity can be observed to extend into the

supragranular layers. In V2, the NADPH-d-background activity is more diffuse, extending from layer IV into the supragranular layers. NADPH-d-positive cells in areas V1 and V2 can be subdivided into large gray matter cells (16 × 16 μm soma size) with round, oval, or pyramidal cell bodies with sparely spinous dendrites, large white matter cells (12 × 19 μm soma size) with oval cell bodies that are horizontally oriented and contain spinous dendrites, and small gray matter cells (3.6 × 4 μm soma size) with stellate shape. The small cells are distributed in layers II–VI of areas V1 and V2 and they are most numerous in layer IV.

Area V2 also contains a rich complement of neurons and fibers immunoreactive for neuropeptide Y. In both areas V1 and V2, immunoreactive pyramidal and nonpyramidal cells are found in layers I–VI. The nonpyramidal cells are either bipolar, bitufted, or multipolar aspiny cells. In layers II and V, most immunoreactive cells are pyramidal in shape. In addition to these cells located in the cortical gray matter, the majority of neuropeptide Y immunoreactive cells are located in the underlying white matter. Many of these cells are located just below the border of layer VI, whereas other cells are located deeper within the white matter. Immunoreactive fibers are found in two plexuses: one in layers I and II and a second in layer V. This pattern is distinct from that seen in area V1, where a third plexus in layers IVB and IVC is observed. Many of these immunoreactive fibers are found to surround small blood vessels, suggesting that they play a role in the regulation of cerebral blood flow. The localization of neuropeptide Y in both aspinous and spinous pyramidal cells suggests that it plays a role in both inhibitory and excitatory neurotransmission. The localization in layer III and V pyramidal cells suggests that this peptide is involved in cotransmission in long-range corticocortical and corticofugal pathways.

Calcium-binding proteins parvalbumin (PV), calbindin (CB), and calretinin (CR) have been reported to be colocalized in GABAergic neurons in human visual cortex. In general, these GABAergic cells are interneurons of the chandelier and basket cell types located in layers II and IV (PV), double bouquet cells of layers II–III and V (CB), or bipolar and fusiform cells in layers II and III (CR). In addition, some pyramidal cells in layers II–III are immunoreactive for calcium-binding proteins. In some cases, more than one calcium-binding protein has been localized in individual cells in areas V1 and V2 of human visual cortex. In area V2 (area 18 near the V1 border), 3.28% of cells are immunoreactive for both PV and CR, whereas 7.39% of cells are immunoreactive for both CR and CB. The colocalization of PV and CB is rare in both areas V1 and V2. In addition to the single- and double-labeled cells located within the gray matter of V2, single- and double-labeled fibers were observed in the underlying white matter. These labeled fibers may belong to projection neurons from the lateral geniculate nucleus or pulvinar because these cells are often immunopositive for calcium-binding proteins.

Many parvalbumin immunoreactive neurons are surrounded by extracellular lattice coatings called perineuronal nets that are stainable with lectins such as *Wisteria floribunda* agglutinin (WFA). Areas V1 and V2 differ in the density, laminar distribution, and target cells of WFA immunoreactivity. In V1, WFA staining is largely distributed in two bands; one located in layer IVB and the second in layer VI. Lighter staining is observed in layers II and III. In V2 a different laminar pattern of labeling is observed. Dense labeling is observed in layers IIIB–IIIC, and less intense labeling is observed in layer V. In addition to these laminar differences, areas V1 and V2 differ in the size of neurons that were enveloped by WFA finding. In both areas, medium-sized cells (13–21 μm) are the most commonly enveloped cell type. In V1 a greater percentage of small cells (5–12 μm) are labeled in layers II–III, IVA, IVB, and IVC. In contrast, a larger percentage of large cells (22–30 μm) are labeled in layers IIIB, IIIC, and V of area V2. Finally, layer VI of V1 contains a larger proportion of labeled large cells.

The density and laminar distributions of neurons that are immunoreactive to SMI-32 that recognizes nonphosphorylated neurofilament proteins, also provide a basis for distinguishing area V1 from area V2. In area V1, densely stained SMI-32 immunoreactive neurons are found in layer 4B and the Meynert cells of layer VI. In addition to these large cells, smaller immunoreactive cells are found in layers III, V, and VI. In area V2, SMI-32 immunoreactivity is found in the large cells of layers III and V (see Fig. 11A). Area V2 also contains a population of SMI-32 immunoreactive cells that are affected in cases of Alzheimer's disease. Although area V2 contained significantly less neuritic plaques and neurofibrillary tangles than prefrontal and temporal cortex, a significant loss of SMI-32 immunoreactive neurons is found in layers III and VI of area V2 (average loss of 18%). These results reinforce the view that Alzheimer's disease preferentially affects neurons with long corticocortical axons such as those in layers III and VI of V2 that project to distant extrastriate cortical areas or to subcortical targets (Fig. 11B).

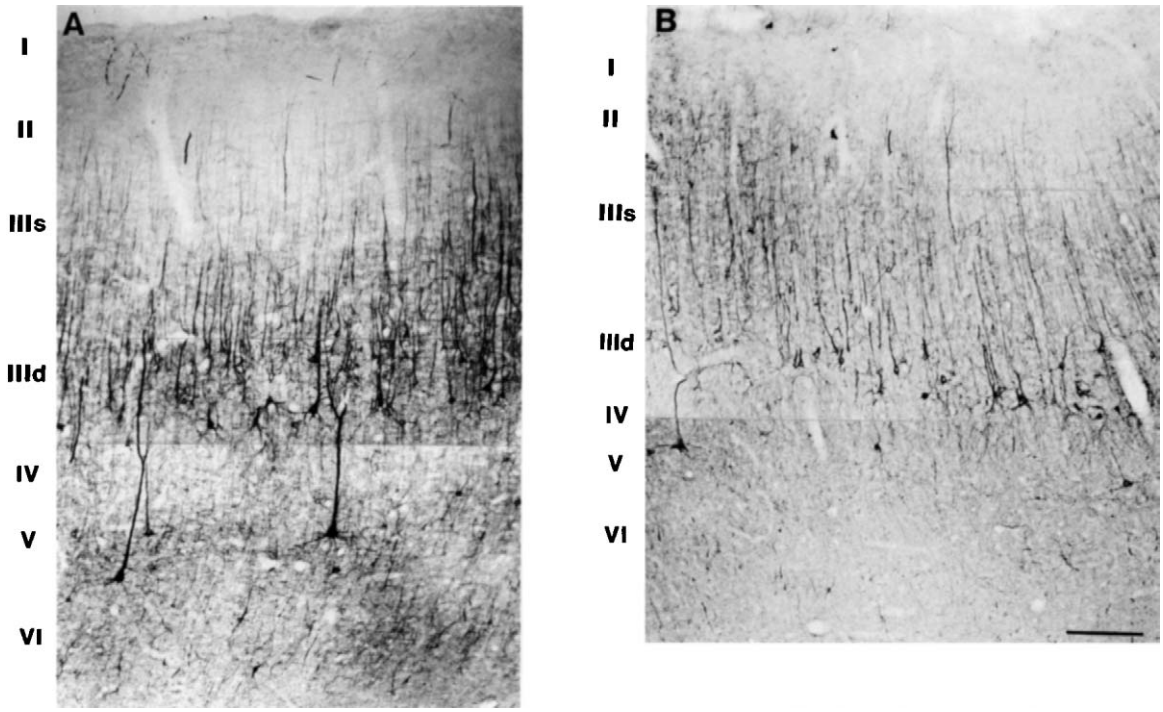


Figure 11 Photomontage of SMI-32 immunoreactivity in human area 18. (A) Normal cortex. Large SMI-32-ir neurons are present in layers IIId and V. (B) V2 in Alzheimer's disease is characterized by a major loss in the density of SMI-32-ir neurons in layers IIId and V. From Hof and Morrison (1990).

V. TOPOGRAPHIC MAPPING OF HUMAN VISUAL CORTICAL AREAS

Functional magnetic resonance imaging (fMRI) has been used to map the location, extent, and topographic organizations of several cortical areas in the human visual system. Signals used by fMRI are thought to represent local changes in blood flow that serve as an indirect marker for the activation of pools of neurons. Eccentricity and polar angle coordinates in several visual areas (V1, V2, V3, VP, and V3A) are usually mapped using phase-encoded, contrast-reversing checkerboard stimuli that were presented as a rotating hemifield (polar angle) or expanding annulus (eccentricity). To maximize attention and arousal and, thus, maximize activity in extrastriate cortex, subjects typically are required to fixate and report changes in the location of a small centrally located stimulus.

Spin echo imaging is used to acquire functional brain maps, whereas gradient-recalled images are generally used to acquire anatomical data that are used for the reconstruction of three-dimensional models of the brain, which then served as input to one of several brain-flattening algorithms to build unfolded cortical maps or inflated brain reconstructions. Figure 12A illustrates the representation of eccentricity in the visual cortex for three subjects in both three-dimensional and two-dimensional reconstructions of the occipital cortex. These cortical maps illustrate the representation of eccentricities ranging from 1.4° to 24° as color-coded bands that extend both dorsally and ventrally from the collateral sulcus (ventrally), to the calcarine sulcus, and then onto the exposed lateral surface of cortex. The representation of polar angle in occipital cortex from these three subjects is illustrated in Fig. 12B. Regions of cortex activated are color-coded yellow when the hemifield checkerboard is located within 5° of the superior vertical meridian, purple when within 5° of the horizontal meridian, and orange when within 5° of the inferior vertical meridian. Because the vertical and horizontal meridians form the borders of many visual cortical areas in macaque monkeys, these data formed the basis for identifying the borders between cortical areas in the human occipital cortex. A large purple region located within the calcarine sulcus corresponds to the horizontal meridian representation of area V1. This region is bordered dorsally by an orange region corresponding to the inferior vertical meridian representation at the border between areas V1–V2. In ventral cortex, V1

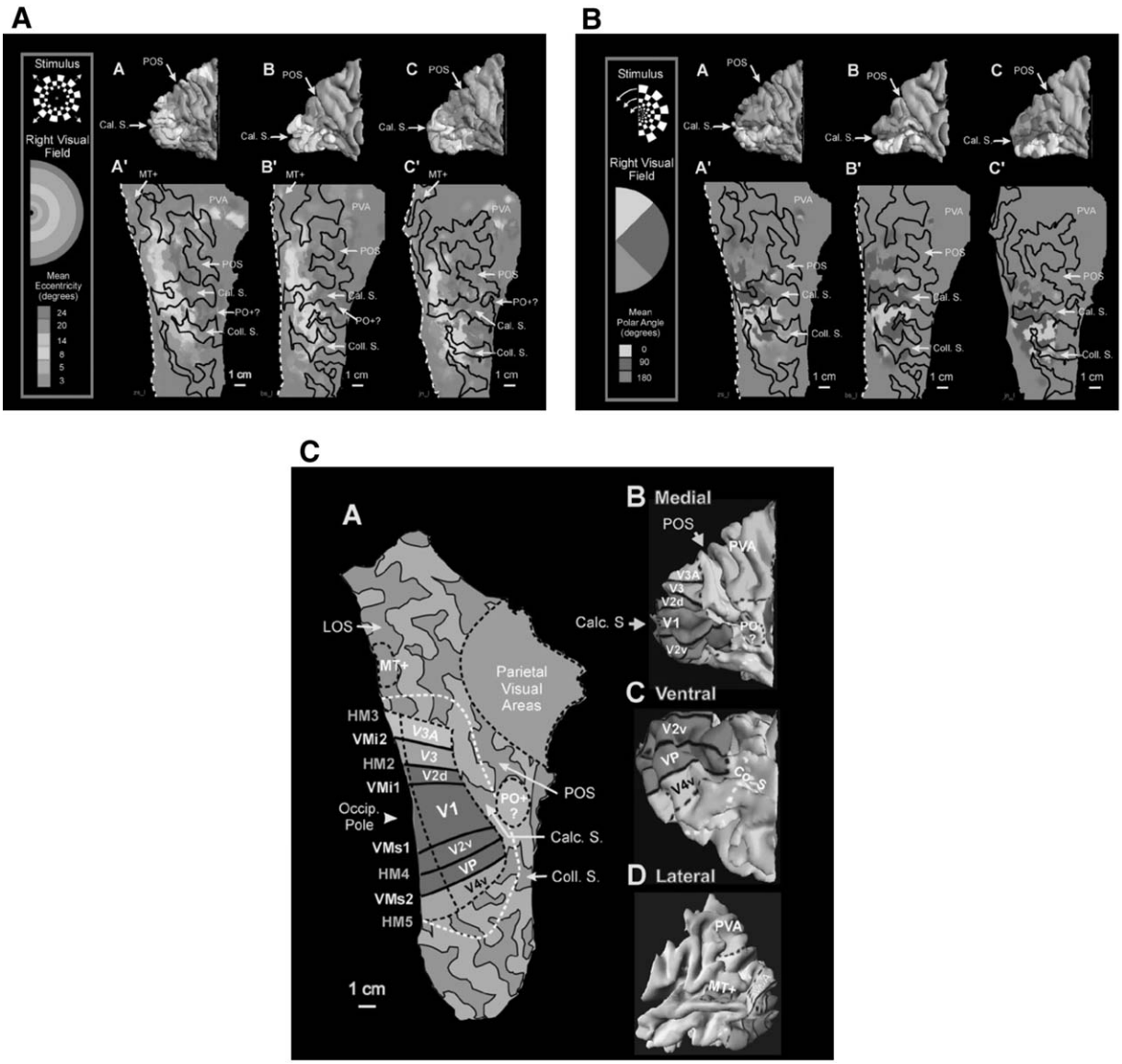


Figure 12 Mapping of visual topography in human visual cortex. (A) Representation of visual field eccentricity from 3° to 30°. Visual stimuli and their representational shades are represented on the left. The locations of activations for three subjects are illustrated on surface reconstructions of the occipital pole (top) and on unfolded cortical maps (bottom). (B) Representation of polar angle determined through activation from rotating checkerboard hemifield (left). (C) Combined visual field topography from three subjects illustrating the multiple representations of the horizontal and vertical meridia found in each subject (left), the localization of the 3–24 isoeccentricity contours (right), and their average overlap (middle). (D) Summary of the topography of human visual cortical areas V1, V2, V3, VP, V3A, and V4v. Area V1 is located in the fundus of the calcarine fissure and is bordered by the representations of the superior and inferior vertical meridia. Area V2 is located both dorsal and ventral to area V1. The anterior border of dorsal V2 is formed by a representation of the horizontal meridian (HM2), and the anterior border of ventral V2 is formed by a second representation of the horizontal meridian (HM4). From DeYoe *et al.* (1996).

is bordered by a yellow band that corresponds to the representation of the superior vertical meridian at the ventral V1–V2 border. The width of V2 can be calculated by determining the location of the next representation of the horizontal meridian that corresponds to the border between areas V2 and V3 in dorsal cortex and between V2 and VP in ventral cortex.

Figure 12C illustrates the average organization of eccentricity and polar angle from the three subjects illustrated in the two previous parts. Area V2

corresponds to a band of cortex that extends from the vertical meridian representation at the border of V1 to the horizontal meridian representations at the border with area V3 in dorsal cortex and VP in ventral cortex. According to this view, area V2 is approximately 1–1.5 cm wide in the region from 1.5° to 24° of eccentricity. This estimate is in agreement with the value of 1–1.5 cm for foveal V2 based on the extent of CO stripes observed in unfolded and flattened human occipital cortex. At more peripheral regions corresponding to intermediate eccentricities, the CO stripes of V2 extend from 1.8 to 3.4 cm. Thus, a potential discrepancy exists between the width of V2 derived from fMRI topographic mapping studies and the width of V2 reported from CO architecture at intermediate eccentricities.

The amount of cortical tissue devoted to the representation of a given portion of the visual field (mm^2/deg^2) is described by the cortical magnification function (CMF). This magnification function varies as a function of eccentricity, with larger values typically found for central visual fields. The calculated CMF for many topographically organized cortical areas is based upon measurements derived from fMRI mapping. The CMF derived for V2 is described by the function $25.19(0.09 + E)^{-1.53}$, where E is eccentricity in degrees.

Areas V1, V2, and several other extrastriate cortical areas have been localized on a surface-based map of the entire cerebral cortex based on the visible man data set. This method provides a convenient method to represent topographic mapping data and functional activation data in a common format that is readily converted to the Talairach stereotaxic space. Figure 13A illustrates the locations of areas V1, V2, V3, VP, V3A, and V4v in the visible man based on topographic mapping data. The uncertainties of the extents of these visual areas are illustrated by question marks on the surface map. Figure 13B illustrates the locations of these areas on a surface model of the visible man's right hemisphere. The lateral view in Fig. 13B illustrates the remaining uncertainty of the foveal projects of these areas. The surface-based atlas also provides a convenient format for the illustration of activation foci from functional studies of color, motion, form, and face processing derived from a wide variety of sources. Corbetta and colleagues have performed a comprehensive study of perceptual processing in the human occipitotemporal cortex. Figure 13C illustrates the location of these activation foci for perceptual tasks involving color, motion, and form processing. Color-processing tasks activated several foci within ventral and dorsal V2; one focus was localized to dorsal V2 in a motion-processing task, whereas three foci were localized to V2 in a form-processing task. Tasks involving the processing of faces do not preferentially activate V2 but instead tend to engage areas in the inferior temporal cortex.

VI. VARIABILITY IN THE LOCATION AND SIZE OF V2 IN HUMANS

PET and MRI imaging have been used to investigate the location and variability of areas V1, V2, VP, V3A, and V4 defined through functional topographic mapping. During PET imaging, 11 subjects viewed dynamic random dot patterns that were arranged as "bowties" along either the horizontal or vertical meridian. Regional blood flow (rCBF) was calculated as deviation from the rest condition for stimulation of either the vertical or horizontal meridian sectors. The resultant PET images were then normalized into the Talairach and Tournoux stereotaxic space. The MRI images from each subject were similarly normalized to Talairach space, and the PET data were coregistered with the MRI anatomical images. The center of area V2 was calculated as the midpoint between the representations of the horizontal and vertical meridia in both dorsal and ventral cortex. The borders of area V2 were consistently observed in the right hemisphere of all 11 subjects and in the left hemisphere in 10 of 11 subjects. Table III summarizes the mean location and standard deviation of the center of gravity of area V2 in the three Talairach coordinates. Overall, the center of gravity of V2 has a mean standard deviation of approximately 5 mm, but it is more consistently localized in the x plane (SD=4.075 mm) and least consistently localized in the z plane (SD=5.975 mm).

VII. VISUAL FIELD DEFECTS ASSOCIATED WITH DAMAGE TO AREA V2 (AND V3)

The organization of area V2 described by functional topographic mapping indicates that area V2 adjoins area V1 along the representation of the vertical meridian in both dorsal and ventral cortex. The anterior border of V2 corresponds to the horizontal meridian, which is split such that the inferior horizontal meridian is located in dorsal cortex and the superior horizontal meridian is located in ventral extrastriate cortex. Area V3 adjoins V2 along the representation of the horizontal meridian in dorsal cortex, whereas area VP adjoins V2 along the representation of the

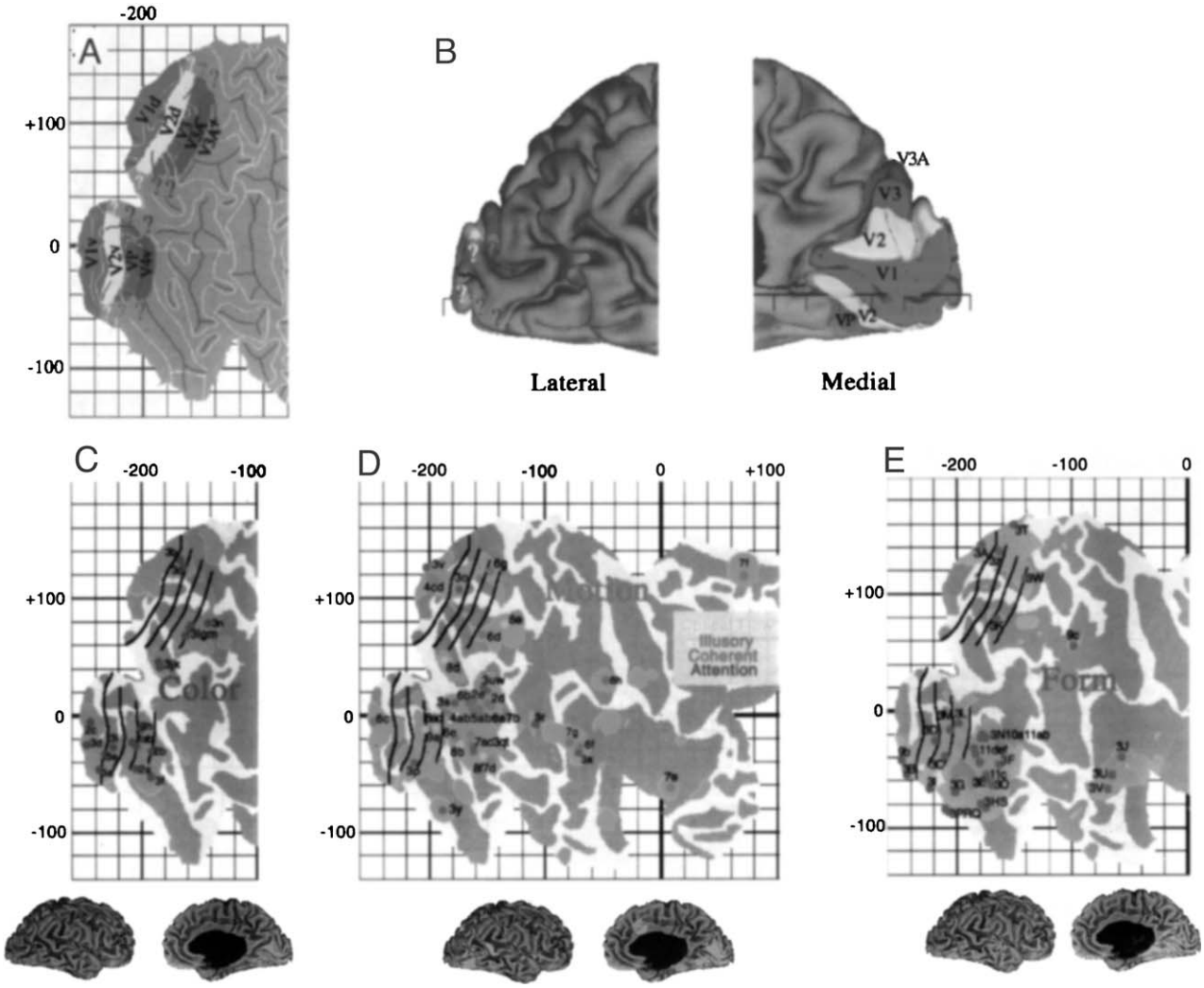


Figure 13 Location and functional mapping onto a surface-based map of human cerebral cortex. (A) Locations of visual areas V1, V2, V3, VP, V4v, and V3A on a surface-based atlas based on fMRI topographic mapping studies. (B) Locations of visual areas projected onto the surface visible man cerebral cortex. (C) Locations of functional activation loci from color-processing tasks (from Corbetta *et al*., 1990). (D) Motion-processing tasks. (E) Form-processing tasks. From Van Essen and Drury (1997).

horizontal meridian in ventral cortex. This is the same topographic organization that has been reported for areas V2, V3, and VP in macaque monkey extrastriate visual cortex. Homonymous quadrantic visual field defects have been observed in two patients with astrocytomas of the extrastriate occipital cortex. Both patients had lower field quadrantanopias that extend up to, but not beyond, the horizontal meridian. Such quadrantanopias have been reported previously and were attributed to lesions of the calcarine cortex. The precise border of the visual field defect at the horizontal meridian is not likely to be caused by lesions of V1 because the border between the lower and upper visual fields is congruent and a lesion is unlikely to precisely respect this border. Instead, quadrantic visual field defects are likely to represent injury to areas V2–V3 (VP) where the horizontal meridian is split into dorsal and ventral halves. Therefore, quadrantic visual field loss may be a hallmark of lesions of early extrastriate cortical areas such as V2, VP (VP), or V4.

In another case, homonymous quadrantanopia was associated with cerebral diplopia (polyopia). This patient had an embolic infarction of the left inferior

Table III
Localization of Human V2 in Talairach Coordinates[a]

V2 component	X plane	Y plane	Z plane
Right V2v	-7.7 ± 4.2	-75.3 ± 5.1	-7.5 ± 4.9
Right V2d	-11.7 ± 4.6	-91.5 ± 4.3	2.7 ± 8.6
Left V2v	7.8 ± 4.0	-71.1 ± 5.9	-3.3 ± 5.9
Left V2d	8.2 ± 3.5	-87.6 ± 4.3	6.1 ± 4.5

[a]Mean and standard deviation of the center of mass of dorsal and ventral V2 in the right and left hemispheres. These values are based on PET activations following topographic mapping of V2. V2 is more consistently localized in the *x* plane and least consistently localized in the *z* plane.

occipital lobe anterior to the primary visual cortex. Similar to the lesions described earlier, this lesion produced a superior visual field quadrantanopia that precisely respected the horizontal meridian. Thus, this lesion is likely to represent damage to ventral extrastriate areas V2 and/or VP. This patient had severe defecits in form and color recognition within the affected quadrant. In addition, the patient reported diplopic images within this quadrant. Although the mechanism of cerebral diplopia is not understood, it is possible that the loss or disruption of feedback signals from extrastriate areas V2 and/or VP to area V1 is responsible for this perceptual anomaly.

VIII. FUNCTIONAL PROPERTIES OF HUMAN AREA V2

A. Disparity Processing

Positron emission tomographic methods have been used to identify visual areas involved in the processing of small horizontal disparities. In one such study, subjects were presented either random noise patterns or Julesz-type stereograms. The difference in the regional blood flow patterns was taken to indicate regions of cortex involved in the analysis of the stereogram patterns. The resultant regional blood flow patterns were then mapped onto the corresponding anatomical sections derived from magnetic resonance imaging. Statistically significant cortical activation was observed over a large region of occipital cortex that involves areas 17, 18, and 19.

B. Chromatic and Motion Processing

Early positron emitted tomographic (PET) studies of the functional organization of visual cortex in humans have concentrated on determining the location and specialization of the human homologs of macaque monkey areas V4 and MT. In a color-processing study aimed at identifying human homologs, regional cerebral blood flow (rCBF) measures were made while subjects viewed either chromatic or gray Mondrian patterns. Two statistically significant foci of activation were observed in the visual cortex: one in the occipital pole corresponding to V1–V2 and a second area in the lingual and fusiform gyrus corresponding to V4. In a motion-processing study aimed at identifying human MT, rCBF measures were made while subjects viewed either stationary or moving small black squares. Again two statistically significant foci were observed: one corresponding to V1–V2 and the second at the junction of the parietal and occipital cortices corresponding to area MT. Given the spatial resolution of the PET technique and the similarity of response properties in V1 and V2, these authors were unable to distinguish activation in V2 from that in V1. Nevertheless, this focus was active in each of the four stimulus conditions tested. Thus, V1–V2 were found to participate in both chromatic and gray scale analyses as well as in stationary and moving square analyses.

C. Illusory Motion Processing

The role of areas V1 and V2 in the processing of illusory motion has been studied using PET to measure rCBF. Subjects viewed two different gray scale images based on *Enigma* designed by I. Leviant. The first consisted of concentric gray rings separated by interrupted black and white spokes. Viewing of this image produces the illusion of rotary motion within the rings. In the second image the spokes continued through the concentric rings and the illusion of rotary motion is not observed. This study revealed strong activation of area MT but little or no activation of areas V1–V2. Thus, these lower cortical areas do not appear to participate in the perception of illusory motion.

D. Higher Order Motion Processing

The role of area V2 in the processing of first- and second-order motion in the human visual cortex has

been investigated using fMRI. Retinotopic mapping of visual cortex was first accomplished using contrast-reversing checkerboard hemifields or wedges. Then various first-order and second-order motion and static control stimuli were presented while subjects fixated on the center of the concentric ring stimuli. Regions of interest (ROI) were then applied to the different cortical areas on two-dimensional cortical maps, and the response magnitude for each stimulus condition was determined for each ROI. Areas V1 and V2 responded to all stimuli that contained dynamic flicker and did not prefer stimuli that contained second-order motion signals. This contrasts with the responses in areas V3, VP, V3A, V3B, and MT that demonstrated strong preferences for second-order motion. Thus, the responses in area V2 were governed by local changes in luminance rather than the higher order motion cues.

E. Illusory Contour Processing

Experiments conducted in macaque monkeys indicate that area V2 contains neurons that are selective for illusory contour stimuli, whereas such responses are weak or not present in area V1. In order to determine which areas in human visual cortex are responsive to illusory contours, fMRI techniques were applied to map the areal distribution of responses to illusory contours of the Kanizsa and displaced grating types. Kanizsa stimuli were compared with aligned and rotated inducers, and the resultant activations were mapped onto retinotopically defined visual areas. For individual subjects, Kanizsa stimuli produced robust activations in intermediate level visual areas V3A, V4, V7, and V8 but not in areas V1, V2, V3, or VP. When an across-subject analysis with restricted regions of analysis (ROIs) was applied to these data, a small, but statistically significant response was also observed in these lower visual areas. When displaced grating stimuli were used to produce illusory contours, strong activations were observed in the intermediate cortical areas and weaker activations were seen in areas V1, V2, V4, and VP. Thus, in agreement with previous single-cell and optical recording in macaque monkeys, area V2 contains neurons responsive to illusory contours based on displaced grating stimuli. Unlike the work in macaques, area V1 contained illusory contour responses that were indistinguishable from those observed in area V2. In addition, these results indicate that human area V2 contains a weak, yet significant signal representing stimuli of the Kanizsa type. In contrast, higher cortical areas contained much stronger responses to both types of illusory contours.

F. Texture Segregation Processing

fMRI has been used to study the role of area V2 in texture segregation. Topographic mapping of human visual cortex was first performed to distinguish topographically organized visual areas V1, V3A, V4, and TEO. In ventral cortex, areas V2 and VP could not be distinguished from each other in all subjects, and, thus, the relevant data must be attributed to this amalgam. Subjects were presented blank displays or oriented line texture displays that were either homogeneous or contained checkerboard patterns that were defined by oriented line element boundaries. In area V4 of all subjects, the checkerboard pattern evoked a larger fMRI response than did the uniform texture display. Similarly, a statistically greater fMRI response to the checkerboard pattern was observed in area V2–VP of one subject, but this significance was not achieved when all subjects were combined ($p < 0.08$). In area V1, the difference in responses between the homogeneous texture and the checkerboard texture was not significant. A texture segregation index was used to characterize the difference in responses to structured and unstructured textures. The greatest differential response to structured textures occurred in higher cortical areas V4, TEO, and V3A. However, the positive finding in V2–VP for one subject and the near-statistical significance in the pooled data suggest that V2 (VP) may be where the analysis of large-scale texture patterns first begins.

G. Global Object Processing

The role of various visual cortical areas in the processing of the global features of objects was studied using fMRI by presenting images of natural objects or highly scrambled versions of the same images. The visual areas of interest were first determined through the traditional mapping of the representations of the horizontal and vertical meridians that delineate the borders between areas. Then a series of full field, blank field, half field (upper or lower), or scrambled images were presented to subjects. Whereas this study concentrated on the sensitivity of the lateral occipital area (LO) to the scrambling of images, important information was acquired about areas V1, V2, and V3. As a

group, these areas were well-activated by the natural object images and did not show a reduction in activity with image scrambling. Thus, these lower tier visual areas appear to be responsive to the local orientation and contrast within a stimulus array rather than selective to the global properties of the natural image.

H. Visual Learning and Imagery

It has been hypothesized that cortical areas involved in visual perception are also involved in the recall of images. This hypothesis has been tested using both fMRI and PET. In one study, subjects alternately viewed LED displays of flashing squares or were asked to imagine the display pattern. A region of high activity during the stimulation condition was observed in the posterior occipital cortex that was reported to include both areas V1 and V2. This same region showed strong, yet weaker activation during the imagine condition in five of the seven subjects tested. These results suggest that areas V1 and V2 are involved in the recall or imagination of visual stimuli.

A second study reached a different conclusion using a different paradigm that allowed the assessment of the role of various cortical regions in visual learning, recall, and recognition. Subjects were presented various colored geometric patterns and were tested for their learning of the pattern contents. PET imaging was used to assess the cortical regions involved in the learning phase as well as in the recall and recognition of these learned patterns. Due to the limited spatial resolution of the PET method and the lack of topographic mapping in these experiments, it was not possible to assign activation foci to specific cortical areas. However, it was possible to compare the activations across learning, recall, and recognition conditions. Early cortical area V1 and adjacent pericalcarine areas (such as V2) were activated in the learning phase of this experiment. In contrast, no cortical regions in the occipital lobe were activated significantly during the recall of learned patterns. In addition, V1 and adjacent pericalcarine fields were activated during the recognition task, where learned patterns were presented intermixed with similar novel patterns. These results suggest that areas V1 and V2 are not involved in visual recall, but instead higher cortical and limbic areas are involved in this complex process. The differences in the results between this study and the first study appear substantial and may be attributable to methodological differences.

I. Effects of Attention

Studies in several areas of occipital and inferotemporal cortices in monkeys have indicated that spatially direct attention can modulate the activity of individual neurons. These studies indicate that, when two stimuli are presented in a neuron's receptive field, the neuron's response is dictated by the attended stimulus. This result has been interpreted as a reduction in the suppressive effect of the second stimulus when attention is directed to the first stimulus in the receptive field. This modulatory effect was observed both in the stimulus-evoked activity and in the spontaneous activity when no stimulus was present.

The role of attention in the activity of area V2 in human visual cortex has been investigated using fMRI with paradigms that are largely identical to those used in single-unit investigations in monkeys used to study stimulus-evoked activity and background modulation. To study the effect of attention on evoked activity, fMRI were investigated in areas V1, V2, VP, V4, and TEO during conditions of sequential or simultaneous stimulus presentation both with and without spatially directed attention. The suppressive effect of simultaneous stimulation with four different stimuli increased from area V1 to area TEO. In addition, the magnitude of the response in both the simultaneous and sequential conditions was increased when attention was directed to the visual stimuli in areas V2, V4, and TEO, but not in area V1. To study the effect of attention on spontaneous activity, the baseline activity prior to stimulus presentation was measured with and without spatially directed attention. A baseline shift index (BSI) was used to quantify this difference in activity during the expectation period prior to stimulus presentation. A statistically significant BSI was observed for all cortical areas studied including areas V1 and V2. Thus, area V2 showed an increase in both stimulus-evoked activity and background activity under conditions of spatially directed attention. Interestingly, area V1 showed a significant effect of attention on background activity but not on stimulus-evoked activity. The increase in background activity during spatially directed attention is similar to the results observed in single-unit recording in areas V2 and V4 of awake macaque monkeys.

Thus, functional imaging techniques have been used to study the role of area V2 in a range of perceptual and cognitive tasks. These functional properties of human V2 are summarized in Table IV. This table represents the beginning of detailed investigations about the function of area V2. Future work will ask more specific

Table IV
Perceptual and Cognitive Functions of Human V2[a]

Task	Functional activity
Attention	+
Color	+
Disparity	+
Form	+
Illusory contour	weak
Global object	−
Motion	+
Illusory motion	−
Higher order motion	−
Texture segmentation	− (weak)
Faces	−
Imagery	?

[a]Summary of the functional properties of human V2 assessed using functional imaging techniques. + indicates function localized to V2; − indicates function not localized to V2. Weak indicates inconsistent or nonsignificant result; ? indicates conflicting results in literature.

questions about the specific roles that V2 plays in these perceptual tasks. Important questions remain concerning the spatial ranges in which V2 performs motion, form, and color analyses and whether these analyses can be attributed to individual functional compartments in V2.

IX. FUTURE INVESTIGATIONS

Future investigations of human area V2 will utilize new techniques to address several anatomical issues related to two- and three-dimensional mapping and several physiological issues related to compartmental organization. Current methods of two-dimensional and three-dimensional mapping techniques can be expected to be used to provide more complete reconstructions of the distributions of several anatomical markers. For example, although cytochrome oxidase stripes have been visualized in limited portions of V2, it has not yet been possible to map the complete distribution of human V2 cytochrome oxidase stripes. Similarly, the distribution of immunoreactivity for CAT-301, characteristic of the magnocellular compartment of V2, has been observed in limited regions of flattened V2, and it will be possible to reconstruct its complete distribution in three-dimensional reconstructions of histological sections and on unfolded cortical maps. These same approaches can be applied to the localization of a variety of other anatomical markers, such as the distribution of callosal afferents and the distribution of immunoreactivity for SMI-32 or various neurotransmitters.

The previously described anatomical studies are primarily concerned with the distribution of anatomical markers in individual brains that have been first imaged by MRI, then sectioned and stained, and then reconstructed back into three-dimensional models and two-dimensional maps. Additional investigations will expand these single-subject reconstructions to build a standardized brain and standard map. These investigations will build upon existing methods of brain warping and map warping to allow for a probabilistic mapping of these anatomical variables.

Future investigations will explore further the connections that area V2 makes with cortical and subcortical targets. One approach might use highly diffusible markers, with greater efficacy than DiI, to trace axonal pathways. Another approach might use physiological methods of microstimulation combined with fMRI to detect cortical targets activated following local magnetic or electrical stimulation. Finally, the more recently developed method of diffusion tensor magnetic resonance imaging might be employed to study the organization and targets of axonal fascicules that leave V2 for other cortical areas.

Future functional studies will provide better insight into the role that human V2 plays in various aspects of visual perception. These studies are likely to proceed along three different fronts. The first of such studies will utilize refined psychophysical methods, in conjunction with fMRI, to better identify the role of V2 in color, form, motion, texture, and illusory contour processing. The overall goal of such experiments will be to distinguish the perceptual contribution of V2 that is distinct from those of area V1 and higher cortical areas. This work will proceed in parallel with similar investigations in macaque monkeys using microelectrode and fMRI techniques.

The second type of future physiological studies will employ fMRI techniques to study the temporal organization of processing in the visual cortex. These studies will explore the temporal sequence of color, form, motion, and texture processing as these signals radiate out of V1. Although it is clear that the visual cortex is organized in an anatomical hierarchy, little is known about the timing of information flow within this hierarchy. Thus, whereas different stimulus attributes are processed by different compartments in V2

and by different cortical areas, it remains unclear, for example, whether fast motion processing via area MT interacts with slower form- and color-processing elements via feedback to area V2.

The third type of future physiological investigation will employ state of the art fMRI techniques to explore the functional properties of the modular compartments of human V2. Previous fMRI studies of human V2 have summed signals across large expanses of cortex to yield a picture of the role V2 plays in perception. Future studies will be directed at the modular segregation of such signals. Current fMRI methods have begun to address the compartmental organization of visual cortex. The visualization of presumed ocular dominance columns in human V1, which are approximately 1 mm in width, increases the likelihood that the larger V2 modules can be similarly visualized.

Finally, better functional localization of visual cortical areas such as V1, V2, V3, VP, V4, and MT in individual brains will allow more highly detailed solid models necessary for the interpretation of multisource electrical or magnetic evoked potentials. The resolution of these multisource models will allow a wide variety of studies that can examine the dynamics of visual perception.

See Also the Following Articles

CEREBRAL CORTEX • MAGNETIC RESONANCE IMAGING (MRI) • MOTION PROCESSING • OCCIPITAL LOBE • VISUAL CORTEX

Suggested Reading

Braak, H. (1980). *Architectonics of the Human Telencephalic Cortex*. Springer-Verlag, Berlin.

Brodmann, K. (1909). Beitrage zur histologischen lokalisation der grobhirnrinde VI. Mitteilung Die cortexgliederung des Menschen. *J. Psychol. Neurol.*, 231–246.

Burkhalter, A. (1993). Sequential development of intracortical processing channels in human visual cortex. In *Functional organization of the Human Visual Cortex* (B. Gulyas, D. Ottoson, and P. E. Roland, Eds.), pp. 151–163. Pergamon Press, Oxford, UK.

Clarke, S. (1993). Callosal connections and functional subdivisions of the human occipital cortex. In *Functional Organization of the Human Visual Cortex* (B. Gulyas, D. Ottoson, and P. E. Roland, Eds.), pp. 137–149. Pergamon Press, Oxford, UK.

Corbetta, M., Miezen, F., Dobmeyer, S., Shulman, G., and Peterson, S. (1991). Selective and divided attention during visual discrimination of shape, color, and speed: Functional anatomy by positron emission tomography. *J. Neurosci.* **11,** 2383–2402.

DeYoe, E. A., Carman, G. J., Bandettini, P., Glickman, S., Wieser, J., Cox, R., Miller, D., and Neitz, J. (1996). Mapping striate and extrastriate visual areas in human cerebral cortex. *Proc. Natl. Acad. Sci. USA* **93,** 2382–2386.

Gulyas, B.(1997). Functional organization of human visual cortical areas. In *Cerebral Cortex* (Rockland *et al.*, Eds.), Vol. 12, pp. 743–773. Plenum Press, New York.

Sereno, M. I., Dale, A. M., Reppas, J. B., Kwong, K. K., Belliveau, J. W., Brady, T. J., Rosen, B. R., and Tootell, R. B. H. (1995). Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science* **268,** 889–893.

Talairach, J., and Tournoux, P. (1988). *Coplanar Stereotaxic Atlas of the Human Brain*. Thieme, New York.

Tootell, R. B. H., Born, R. T., and Ash-Bernal, R. (1993). Columnar organization in visual cortex in nonhuman primates and man. In *Functional Organization of the Human Visual Cortex* (B. Gulyas, D. Ottoson, and P. E. Roland, Eds.), pp. 59–74. Pergamon Press, Oxford, UK.

Wong-Riley, M. T. T. (1993). Cytochrome oxidase studies on the human visual cortex. In *Functional Organization of the Human Visual Cortex* (B. Gulyas, D. Ottoson, and P. E. Roland, Eds.), pp. 165–180. Pergamon Press, Oxford, UK.

Zilles, K., and Clarke, S. (1997). Architecture, connectivity and transmitter receptors of human extrastriate visual cortex. Comparison with nonhuman primates. In *Cerebral Cortex* (Rockland *et al.*, Eds.), Vol. 12, 673–742. Plenum Press, New York.

Arousal

RICHARD T. MARROCCO and BRENT A. FIELD
University of Oregon

GLOSSARY

agonist A drug that activates the same receptor as does the natural neurotransmitter.

antagonist A drug that blocks the activation of a receptor.

electroencephalograph The electrical activity of the brain measured from the surface of the scalp.

G protein-coupled receptors Cell surface receptors that are activated by substances in the synaptic cleft but are coupled to channels by multiple chemical messengers that open the channels from within the cell.

ligand-gated receptors Cell surface receptors that are activated by chemical substances in the synaptic cleft (ligands) and are coupled directly to channels in the cell's membrane.

neuromodulatory transmitter A neuroactive substance found in the peripheral or central nervous system that causes long-lasting changes in excitability of the postsynaptic cell.

Arousal is the ability to mobilize metabolic energy to meet environmental or internal demands on behavior. As such, it is an organismal property that is found throughout the animal kingdom, including those species with very rudimentary nervous systems. For example, in jellyfish, arousal may be produced by the diffusion of activating substances within the body wall that increase metabolic activity. In insects, activating substances may be generally mobilized within ganglia and released onto specific target structures from single neurons. In marine mollusks, repeated application of a noxious stimulus to the skin causes a withdrawal response that grows in magnitude over time. Although termed sensitization, it is associated with global changes in metabolism and is probably the precursor to the arousal response in higher organisms. The anatomy and physiology of arousal are similar among primitive vertebrates and highly complex mammals. Although the arousal response in humans is present at birth, research suggests that some components of the response are recognizable *in utero*. Arousal responses in neonates are frequent and associated with hunger and discomfort. The soothing voice of the parent is usually the first external stimulus to exert control over the arousal response. The major developmental changes in the arousal response are the growth of inhibitory processes that control and shape the response and an increase in the range of internal and environmental stimuli that activate these processes.

I. HISTORY

The concept of arousal dates well back into the histories of many linguistic groups. According to *Webster's Dictionary*, the root *arouse* was used in English as far back as 1593. The first scientific investigation into arousal appeared in the 1860s. Arousal from sleep was measured by determining from how high an object needed to be dropped for the subject to awaken. Kohlshütter reported that sleep depth increased up to 1 hr post-sleep onset and then decreased through the remainder of the night.

Modern work on arousal systems began in the era after World War I, when viral epidemics occurred that caused damage to the brain stem and produced "sleeping sickness," which was characterized by a profound hypoarousal. About two decades later, the first evidence for arousal centers in the upper brain stem was obtained from experimental animals. Perhaps the most cited work in this area of medicine is that of Moruzzi and Magoun. The major conclusion of their work was that brain stem influences modulated thalamocortical neurotransmission. Although the details of this conclusion have changed substantially, the core idea published more than 50 years ago was essentially correct.

In the past two decades, progress in understanding arousal systems has been advanced most dramatically through neurochemistry. It is now clear that the arousing structures of the brain stem and hypothalamus have neurochemically distinct profiles, and the neurotransmitters in each may use distinct intracellular pathways to regulate the excitability of cortical neurons. Most, if not all, of these neurotransmitters show increased turnover during the aroused state and, conversely, drugs that stimulate the release of these substances cause increased cortical arousal. In addition, these neuroactive substances regulate cells of other arousing structures as well, making it all but certain that variations in behavioral states result from a change in the patterns of activity between the arousing structures.

The potential complexity of these patterns of activity is impressive. As will be discussed later, the arousing structures release five major neurotransmitter substances that bind to at least 33 pre- and postsynaptic receptor types. Although found in very low concentrations within the axon terminals, nine additional neuroactive substances, including neuropeptides and adenosine triphosphate (ATP), are coreleased and bind to many postsynaptic sites. Both classes of substances are undoubtedly important for the patterns of arousal. Ultimately, understanding how the arousal response occurs depends on knowing which neurotransmitters are active and where they are exerting their effects.

II. NEUROCHEMISTRY

The major neuroactive substances we discuss are noradrenaline (NA), dopamine (DA), acetylcholine (ACH), serotonin (SE), and histamine (HA). Most of these neuromodulatory transmitters (collectively referred to as NMTs) are located in brain stem or forebrain nuclei and modulate both subcortical and cortical structures. NA (also known as norepinephrine) is contained in many cell groups of the pons but the locus coeruleus is the critical structure for cortical arousal. DA-containing cells reside in the pontine tegmentum and the cortex receives dopaminergic fibers from the ventral tegmental region. ACH neurons are found in the pontine pretrigeminal area and the basal forebrain. SE (also called 5HT) neurons are located along the medullary midline in the raphe nuclei. HA is manufactured by cells of the hypothalamus.

Other neuroactive substances also influence brain arousal, but are different from the NMTs in one or more ways. Adenosine is an amino acid that inhibits other NMTs and is found in many locations throughout the brain. Unlike the other NMTs, however, it is neither a classical neurotransmitter, because it is not released synaptically, nor is it contained in cells that reside in the brain stem. Adrenaline (epinephrine), which resides in brain stem neurons and uses the same receptors as NA, does not project to the cortex and is generally thought to contribute little to conscious arousal. Adenosine and adrenaline will not be discussed further. Glutamate is occasionally listed as an NMT. Glutamate is contained in the terminals of the thalamocortical neurons and provides the synaptic inputs that drive cortical cells. We take the view, however, that arousal is a modulation of the thalamocortical inputs, not the action of the inputs themselves, and we do not include glutamate as an NMT.

This article reviews neuroscientific research conducted within the past 20 years on the roles of each NMT in arousal and identifies, where possible, patterns of activity among the arousing structures that are correlated with behavioral states. We begin by defining the terms that will be used and by describing the physiological indices used to assess arousal tone. We then provide an overview of the anatomical components of the arousal systems and their associated NMTs. For each NMT, its anatomy, its effects on target cells, and its role in specific behaviors will be discussed. We then attempt to integrate the properties of these systems and show how they shape and modulate behavior. We will not discuss the connections between arousal and sexual or consummatory behavior.

III. TERMINOLOGY

A review of the arousal literature reveals that a variety of terms are used to define aspects of arousal. *Arousal*

and *alerting* are often used interchangeably, as are *arousal* and *behavioral state*. In this article, we take the position that the changes in nervous system excitability may differ significantly in duration. We refer to changes lasting a few milliseconds to a few seconds as shifts in *alertness*. We refer to changes lasting from seconds to minutes as *vigilance*, for lack of a better word. Our use of vigilance is different than that in the signal detection literature, in which it refers to the readiness to detect infrequent events in noise. We use vigilance to describe the physiological reactions to environmental events that accompany shifts in focal attention, an increased readiness to respond, and increased activity in the sympathetic nervous system. The events are often cognitive demands on the organism, including emotional valence, task difficulty, and stimulus novelty. Relatively speaking, we consider alerting to be a phasic process and vigilance a tonic process. The term arousal is used in a more general sense to refer to any changes in nervous system excitability regardless of duration.

IV. AROUSAL INDICES

The most common index of arousal is the cortical electroencephalogram (EEG), which can be recorded from the scalps of most mammals. The electrical activity recorded at the scalp represents the summed voltages produced by extracellular neuronal activity beneath the recording electrode. Neuronal electrochemical currents are filtered by the skull and surrounding tissues as the currents spread passively to the recording device. Consequently, the exact location of a particular voltage change is difficult to determine.

The electrical voltage of the EEG ranges from a few microvolts in the aroused state to several hundred microvolts in deep, slow-wave sleep. Therefore, EEG voltage is inversely correlated with behavioral arousal. EEG frequency varies directly with arousal, with the highest values (50–60 Hz) associated with the highly aroused state and the lowest values (2 Hz) seen in comatose states. Both overt behavior and EEG characteristics tend to change gradually during the day. One exception to this is during sleep, when abrupt transitions between slow-wave sleep and rapid eye movement (REM) sleep occur. The EEG characteristics of REM sleep are nearly identical to those of the alert state and reflect episodes of dreaming.

The EEG can also be used to measure alerting and vigilance by measuring changes in electrical activity immediately following an environmental stimulus, a shift of attention, or a change in the task at hand. Although arousal state may be recognizable from a single, 3- to 5-sec epoch of the EEG, it is usually necessary to average tens or even hundreds of short (e.g., 300 msec) segments of the EEG following the external stimulus to measure the change caused by alerting.

Many other measures have been used extensively, including electrodermal responses, changes in blood chemistry, and even early gene expression. A discussion of these issues is beyond the scope of the current review, but excellent treatments may be found in texts on psychophysiology.

There are pitfalls in relying on any single measure of arousal. For example, alerting caused by fear-evoking stimuli causes an increase in heart rate and other autonomic indices. In contrast, phasic alerting caused by orienting toward a nonthreatening stimulus causes a slowing of the heart and other internal organs. Thus, a complete understanding of the arousal response requires a consideration of the pattern of responses and the context in which they occur.

V. ANATOMICAL SUBSTRATES OF AROUSAL

Behavioral arousal is caused by an interplay between the peripheral and central nervous systems. The central nervous system (CNS) evaluates external stimuli for meaning and initiates or modulates activity in the peripheral and central systems. Peripheral arousal is the result of activation of the autonomic nervous system, which affects internal organs directly and skeletal musculature indirectly. One of the main functions of the sympathetic branch of the autonomic nervous system is to liberate adrenaline into the bloodstream. Adrenaline increases alertness and energy expenditure, which are then sensed by the CNS. Thus, peripheral and central arousal are interdependent contributors to the control of behavioral arousal.

A. EEG Arousal Is Controlled by Interconnections between the NMTs, Thalamus, and Cortex

1. Two Modes of Thalamic Activity

Changes in the EEG during arousal are produced by an interaction of the NMTs with the cortex and thalamus. The NMTs control the strength and pattern of the coupling between the thalamus and the cortex.

During rest or sleep, thalamic reticular neurons (Fig. 1) emit rhythmic bursts of action potentials within a narrow range of frequencies (4–8 Hz). The bursting pattern is maintained as long as these cells are slightly hyperpolarized. Hyperpolarization may be caused by the release of SE in the thalamus by the brain stem raphe nuclei. Thalamic burst activity, known as the burst mode, is relayed to the cortex, where it produces cortical burst activity. On the scalp, this pattern is recorded as high-voltage, slow-wave cortical activity, known as the synchronized EEG.

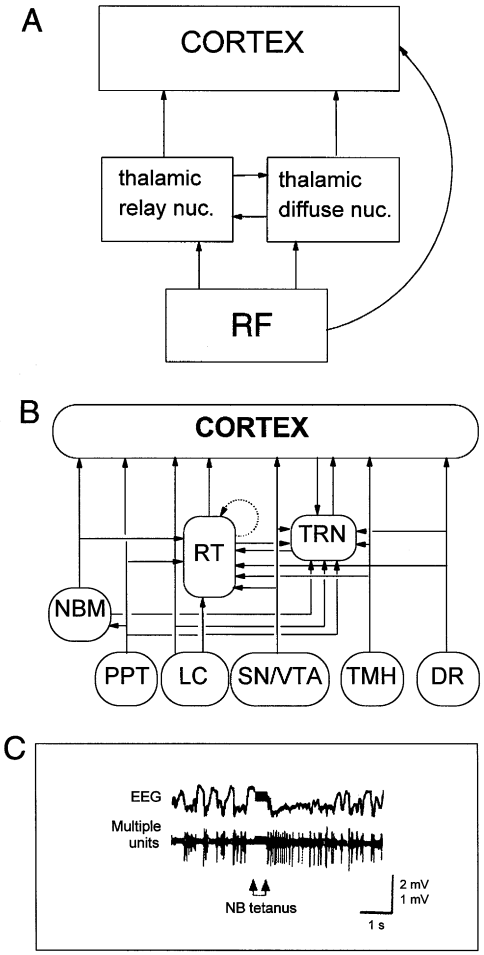


Figure 1 (A) Early model of arousal systems, circa 1972. The RF was considered a unitary structure that modified processing in sensory thalamic relay structures, nonsensory diffuse nuclei, and the cortex. The two groups of thalamic nuclei were mutually interconnected. (B) Contemporary view of arousal systems. Each NMT (bottom) sends influences to both the RT and the TRN. The RT contains oscillatory circuits (dotted line) that produce the rhythmic activity of the brain. (C) Electrical activity recorded in the cortex and the effects of electrical stimulation (tetanus) of the cholinergic NBM. The EEG switches from high-voltage slow waves to low-voltage fast activity. Individual cells show burst activity prior to tetanus and single spike activity for several seconds after tetanus [reproduced with permission from Metherate *et al.*, (1992). *J. Neurosci.* **12**, 4701–4711. Copyright Society for Neuroscience]. RF, reticular formation; NBM, basal nucleus of Meynert; PPT, peduncolopontine tegmental nucleus; LC, locus coeruleus; RT, relay thalamic nucleus; SN/VTA, substantia nigra/ventral tegmental nuclei; TMH, tuberomamillary hypothalamic nucleus; DR, dorsal raphe nucleus; TRN, thalamic reticular nucleus.

In contrast, the low-voltage, fast-wave activity, known as the desynchronized EEG, results when NMTs suppress the hyperpolarizing influences on thalamic neurons and shift the thalamic activity into the single-spike mode. This activity is initiated by ACH that is released from the terminals of basal forebrain neurons. The frequency spectrum of rhythmic activity increases substantially and high frequencies (40–60 Hz) may be frequent. Similar patterns are engendered in cortical neurons and the EEG records of this activity are seen during REM sleep and waking states. Although high-frequency EEG activity has been characterized as desynchronized, it may simply be synchronized to higher frequencies. The mechanisms underlying the burst and single-spike mode are discussed in-depth later.

2. Models of NMT Modulation of Thalamocortical Activity

The mechanisms underlying EEG desynchronization have been of great interest for decades. Early studies in the 1940s and 1950s by Moruzzi and Magoun and others showed that stimulation of the midline thalamus caused a desynchronized EEG pattern in anesthetized cats. Based on these results, two systems for the control of cortical arousal were hypothesized: the generalized thalamocortical system, composed of the midline and intralaminar thalamic nuclei, and the specific (relay) thalamocortical system. The relay system carries sensory information to the cortex, but its cortical influence is modulated by the generalized system. The general system has also been called the ascending reticular system and the diffuse thalamocortical projection system.

The arousal systems must work in unison to control behavior, but the arousal response is not necessarily stereotyped. In some circumstances, the EEG and other indices of arousal may be uncorrelated with behavioral or physiological responses. Specific behavioral contexts are important in determining the patterns of motor, autonomic, emotional responses. For example, intense fear may cause overt, widespread changes in muscular locomotor activity and produce escape responses. Alternatively, there may be a

complete lack of overt activity but widespread covert changes in muscle tone, such as occurs during "freezing" behavior. These qualitative changes may indicate an underlying shift in dominance from one transmitter system to another. Thus, it is more appropriate to view arousal as a flexible response to changing environmental demands rather than as a reflexive, fixed pattern of activity.

B. Arousal State Depends on Excitatory and Inhibitory Actions Mediated by a Variety of Receptors

1. Peripheral Nervous System

In the peripheral autonomic nervous system (ANS), phasic or tonic arousal is associated with a shift from anabolic to catabolic metabolism. Anabolic changes are mediated by ACH, which generally has an inhibitory effect on target organs; catabolic, excitatory changes are mediated by two catecholaminergic transmitters, adrenaline and NA. Thus, the peripheral nervous component of arousal level is determined by an interplay between ACH, NA, and adrenaline.

2. Central Nervous System

There is a more complex interplay between excitatory and inhibitory forces in the CNS compared to the peripheral system due to a greater number of NMTs and their associated receptors. Most of the arousal-related substances may be excitatory or inhibitory or both, depending on the specific receptor–ion channel complex they stimulate. The latency and duration of these effects are quite varied. Another difference is that structures in the CNS are not innervated by two neurotransmitters substances with opposite effects. In most cases, cortical cells are contacted by NMT axons that have similar postsynaptic effects. Opposing actions may be mediated by two NMTs on a single cell but they are likely mediated by inhibitory interneurons.

3. Receptor Families

Transmitters produce receptor-mediated, postsynaptic responses in two ways. First, ligand-gated receptors are directly coupled to ion channels in the postsynaptic membrane. The binding of the ligand produces short latency and short duration effects. Second, G protein-coupled receptors are linked to second messenger systems such as protein kinases or adenylate cyclases. In this type of system, the binding of the ligand can produce a variety of biochemical reactions that result in longer latency and longer duration effects. The effects of ACH 5HT are mediated by either ligand-gated or G protein-coupled receptors. The effects of NA, HA, and DA are G protein coupled only. Next, we provide a more detailed description of the individual neuroactive substances and their interactions with receptors.

C. Noradrenaline

1. Projections of the NA System Are Highly Divergent

NA is supplied to the forebrain by neurons in seven medullary nuclei. The cortex is innervated by one of those nuclei, the locus coeruleus (LC). LC axons travel to the cortex in the dorsal tegmental bundle and show a spectacular pattern of branching that delivers NA to many separate areas. In each area, the fibers approach the cortical mantle from the underlying white matter, project upward, and give off collaterals in almost all laminae.

NA influences cortical neurons through α and β adrenoceptors. These receptors are differentially distributed across cortical laminae and cortical areas. Each receptor is activated by unique ligands, triggers distinct intracellular second messenger systems, and mediates different physiological responses in postsynaptic neurons. In the visual cortex, for example, stimulation of β receptors (β_1 and β_2 subtypes) with sotalol produces mostly short- and long-term inhibitory effects on cells in lower cortical laminae. Stimulation of α receptors (both α_1 and α_2 subtypes) with phentolamine produces only short-duration excitatory effects on cells in the upper cortical laminae.

In the primate visual cortex, greater NA innervation is seen in the superficial and deep laminae than in the middle laminae. This suggests that NA has a greater influence on cortical outflow than cortical inflow. Perhaps NA may be important in regulating the amplification of signals sent to other cortical areas.

2. Cellular Effects of Noradrenaline May Be Excitatory or Inhibitory

In vivo application of NA onto cortical and thalamic relay cells produces a slow depolarization of the membrane and an inhibition of the potassium con-

ductance. This makes the cell react more strongly to its inputs. The effects on the cell's spontaneous activity are more variable. Reductions in activity are seen that effectively increase the salience of evoked responses compared to the unstimulated activity. Thus, the effect of NA is not expressed until a sensory stimulus is present. In other cells, elevations in spontaneous activity are seen. No increase in salience occurs in these cells and decreases in salience may occur. NA's effects on single cells is complex and probably depends on the unique distribution of receptors on each cell. It is to be expected that NA will have complex effects on networks of neurons, but this expectation has not been tested.

What events cause NA to be released onto postsynaptic targets? NA neurons appear to respond phasically to the appearance of nonarousing environmental stimuli and in conjunction with shifts of attention, but they respond more tonically to arousing stimuli (e.g., threatening or appetitive stimuli). In addition, tasks requiring intense focusing of attention are accompanied by a steady rate of firing in LC neurons. When attention wanders, the tonic activity of LC neurons increases and becomes more variable. These results suggest that the rate of LC activity may be positively correlated with attentional lability. Conversely, focused attention reduces the likelihood that the organism will be alerted by unexpected stimuli. This relationship will be discussed later with regard to recent studies of alerting.

The release of NA may also be produced by the application of cholinergic agonist to the LC. This manipulation causes excitation in LC neurons and a reduction in low-frequency EEG activity, similar to that seen during natural arousal. However, it has little effect on the 10- to 30-Hz components that are enhanced during natural arousal. Perhaps isolated stimulation of the LC fails to activate other NMTs that produce increased high-frequency activity during natural arousal.

3. Cortical Arousal May Be Altered by Local Factors

As would be expected from its anatomical organization, stimulation of the LC causes release of NA at many cortical and subcortical sites simultaneously. This has been confirmed empirically and suggests that NA release is governed by activity in LC cell bodies. In many other experiments, however, NA release is regulated by factors near the terminals of LC neurons. In the visual cortex, physiological experiments demonstrate that the release of NA in the visual cortex by visual stimulation depends on the efficacy with which the visual stimulus affects the electrical activity in nearby cortical neurons. The same stimuli are ineffective in evoking LC neuronal activity. These results suggest that the release of NA from visually nonspecific LC neurons may be determined by excitation in nonnoradrenergic, cortical cells. In the frontal cortex, local administration of glutamate causes the release of NA from the terminals of LC neurons.

All these observations can be explained by postulating two kinds of control over the LC neurons. The first is produced by synaptic inputs to dendrites or cell bodies, which cause global release of NA. The second is produced by the local regulation of NA release from individual LC terminals, which would account for the local physiological effect of visual stimulation. Indeed, the second type of control, which occurs in the absence of impulse activity generated in the cell bodies, may be the basis for interactions between many neurotransmitter systems. For example, cholinergic and glutamatergic control of DA and NA release and noradrenergic control of ACH by preterminal modulation of transmitter release have been demonstrated.

4. Behavioral Effects of Altering CNS Noradrenaline Are Diverse

Systemic administration of NA agonists and antagonists has both peripheral and central nervous effects. Here, we limit the discussion to drugs that alter CNS arousal. The stimulants methylphenidate and amphetamine increase the levels of both NA and DA by stimulating α_2, D_1, and D_2 receptors and blocking reuptake of DA and NA into the presynaptic cleft. These drugs also increase alerting and vigilance to environmental stimuli.

The effects of nonstimulant drugs on arousal may be more complex. For example, clonidine, an α_2 agonist, causes deficits in stimulus alerting in nonhuman primates. Subjects are impaired in using the timing of the stimulus onset to help them respond to visual targets in a rapid fashion. This effect is due to stimulation of presynaptic receptors, which reduces brain NA. Clonidine also ameliorates the memory deficits of aged monkeys by increasing the levels of brain NA through stimulation of postsynaptic receptors. It probably increases their alertness as well. The differences in outcomes may be due primarily to the amount of drug present near the synapse. Low doses stimulate postsynaptic receptors and higher doses stimulate postsynaptic and presynaptic receptors.

The results may also depend on the age of the subject and the type of task used.

In humans, the brain locations at which NA regulates changes arousal are of great interest. Recent efforts to identify these sites using functional magnetic resonance imaging of brain activity have shown that the thalamus is especially important in mediating the interaction between changes in arousal and the performance of cognitive tasks requiring sustained attention.

D. Dopamine

1. Dopaminergic Neurons Project Mainly to Subcortical and Frontal Cortical Areas

DA is supplied to the cerebral cortex primarily by cells of the pontomedullary region termed area A10, whose axons travel in the median forebrain bundle and bifurcate to innervate the frontal and motor cortices. There is a rostral–caudal gradient of terminals across the hemispheres, with the highest densities in the frontal lobe and virtually no DA in the occipital lobe. DA neurons from the substantia nigra innervate the corpus striatum. The rest of the brain receives DA from seven other brain stem dopaminergic nuclei.

2. Dopamine's Cellular Effects Resemble Those of Noradrenaline

The cellular effects of DA application are similar to those of NA. There is a generalized reduction of spontaneous activity in neurons with DA receptors and overall increases or decreases in evoked activity. No changes in signal-to-noise ratios are reported. DA neurons typically are active during shifts in behavioral states. Brief bursts of activity are recorded following behaviorally arousing stimuli. Inhibition of activity is reported for periods of sustained, focal attention. Thus, DA neurons are more active during the shifting of arousal or attention than during sustained arousal.

Recent work investigated which of several DA receptors may mediate arousal. D_2 receptor agonists produce cortical desynchronization and an increase in locomotor activity in rats. If both D_1 agonists and antagonists are coadministered, the rat shows no signs of increased arousal. However, administration of a D_1 antagonist blocks the effect of arousing stimuli but does not produce sedation. Thus, an increase in D_1 receptor activity may underlie increases in arousal.

Stimulation of D_2 receptors produces desynchronization but has a less dramatic effect on locomotor activity. These effects are not entirely blocked by coadministration of D_2 antagonists. As in the case of D_1 antagonists, administration of D_2 antagonists alone fails to produce sedation. However, simultaneous blockade of both D_1 and D_2 receptors produces sedation and EEG synchronization.

Thus, it appears that activity mediated by D_1 or D_2 dopaminergic receptors may underlie increases in arousal. Decreases in the activity of either receptor type, however, are not sufficient to decrease the organism's arousal. Rather, activity in both D_1 and D_2 receptors must be decreased simultaneously for a decrease in arousal to be measured.

3. Dopamine Modulates Behavioral Action

The literature linking DA to behavior is vast. Although DA has been implicated in many behaviors, the common theme in behavioral pharmacological studies is motoric. That is, the sensory- and memory-related components of behavior are less affected than the response components by manipulations of DA levels. These observations are consistent with the extensive dopaminergic innervation of the corpus striatum, a key structure for movement. In addition, there are dramatic instances of DA involvement in degenerative diseases that produce major movement deficits. For example, Parkinson's patients, whose midbrain DA neurons have been lost, exhibit severe deficits of voluntary and involuntary action. Also, involuntary movements (tardive dyskinesia) are side effects of drugs that block DA receptors in schizophrenic patients. Lastly, people who abuse stimulants that affect the DA system may show stereotyped, rhythmic hand and head movements. Interestingly, stimulant abuse may also cause a model paranoid psychosis, including delusional behavior, hallucinations, and a hyperattentional state. Clearly, DA's effects may be apparent in brain systems that are not obviously motoric. DA's involvement in other behaviors is seen in normal subjects as well.

DA's effects on attention in normal subjects depend on the task the subject must complete. For example, the shifting of attention may be evoked in humans by cues (arrows) that point in the direction of a location to be attended. Administration of droperidol, a D_2 antagonist, slows overall reaction times to visual stimuli and increases the time required to reorient attention when it is focused on the wrong target location. These results are consistent with the motoric

theme as long as the movement of attention is treated as covert motor activity.

Other human studies use D_2 antagonists to examine the orienting of attention to cues that physically appear near the location at which a target appears. No specific effects on attention are found for these drugs, but there is a generalized slowing of responses to all targets. Studies of animals lacking dopaminergic cortical projections on one side of the brain also show generalized motor deficits in reacting to visual stimuli. Taken together, DA appears to play a role in the motor activity required to react to the targets in general but not in attention *per se*.

E. Acetylcholine

1. Projections of the Cholinergic System Are Relatively Punctate

There are six distinct sources of ACH in the cerebrum. Cortical ACH is supplied primarily from the basal forebrain nucleus and secondarily from the ansa lenticularis and substantia innominata. Recent work has examined the degree of divergence in the axonal projections of basal forebrain neurons. The patterns are distinctly different than those seen for noradrenergic or dopaminergic neurons. Neighboring cortical locations are almost always innervated by different cholinergic neurons. Indeed, the axon terminals of single basal forebrain neurons often synapse entirely within single cortical columns, the smallest anatomical division of the cortex. Although some divergence is occasionally seen, the bulk of the cholinergic projections to the cortical mantle show extremely limited divergence.

The effects of ACH are mediated by nicotinic and muscarinic receptors, which mediate excitatory effects on cortical neurons. Muscarinic receptors are more numerous than nicotinic receptors and are found mostly on postsynaptic cells. Nicotinic receptors are more often presynaptic than postsynaptic. Both types of cholinergic receptors are found throughout the neocortex and are most dense in the lower cortical laminae. Some subcortical areas, such as the optic tectum, seem to contain only nicotinic receptors.

2. The Most Common Cellular Effect of ACH Is Excitation

Both direct application of ACH and electrical stimulation of the basal forebrain nuclei produce rapid firing in cortical neurons (Fig. 1C). Unlike direct application, basal forebrain stimulation causes extracellular ACH to increase simultaneously at widely separated locations in the cerebral cortex. The breadth of this effect is what one would expect from cholinergic anatomy. Is the same pattern evident during natural arousal? The answer is yes: A robust increase in cortical extracellular ACH is measured during the transition between sleep and wakefulness. Does damage to the basal forebrain impair arousal? This answer is also yes because damage to basal forebrain neurons causes synchronized cortical activity. Thus, the release of ACH from the terminals of basal forebrain neurons is both necessary and sufficient for cortical arousal and also mediates the changes in EEG activity.

Knowing how ACH changes the patterns of the cortical EEG is only half the problem. ACH's role in regulating the neural interplay between the thalamus and the cerebral cortex must also be understood. McCormick provides an excellent summary of these regulatory effects. For example, cholinergic stimulation of neurons in the lateral geniculate nucleus (LGN) causes cortical cells to switch from a burst mode to a single-spike firing mode and thereby produces visual cortical EEG desynchronization. This switch is caused by the reduction of three potassium-sensitive currents that excite the cell. In addition, ACH inhibits the LGN interneurons that inhibit the LGN cells that project to the cortex. Thus, cholinergic influences directly excite and disinhibit cortical neurons. During drowsiness, the production of cortical synchrony depends on removal of cholinergic stimulation and stimulation of inhibitory LGN interneurons by noncholinergic NMTs.

The pedunculopontine area of the brain stem is the source of thalamic cholinergic modulation. Stimulation of this area produces arousal-like effects in thalamic cells and a desynchronized EEG pattern in the cortex. Pedunculopontine-mediated cortical arousal is not dependent on the basal forebrain system since blocking the nucleus basalis Meynert does not abolish the effect. Especially intriguing is the specific augmentation of the high-frequency (25–45 Hz) components of the EEG. These components are thought to be important in arousal or attentive visual fixation and pattern recognition, and they are not produced by LC stimulation.

The central role of ACH in arousal is also supported by pharmacological studies. Cholinergic agonists (e.g., carbachol and bethanecol) increase cortical ACH and EEG desynchronization. Interestingly, cholinergic antagonists (atropine and scopolamine) reduce cortical ACH and increase EEG synchrony, but they

produce an alert, wakeful behavioral state. This finding illustrates that there are occasional dissociations between the physiological and behavioral indices of arousal and reiterates that the EEG alone may not give a complete picture of the dynamics of the arousal response.

3. Acetylcholine Improves Cognitive Function

Through the manipulation of the brain's chemical environment, researchers have found that ACH plays a key role in such cognitive functions as learning, memory, and attention. This view is consistent with our view that cholinergic activation, by virtue of its broad excitatory actions, facilitates behaviors by maximizing information processing in brain systems.

As an example of ACH's cognitive benefits, nicotine has been shown to improve visuospatial attention shifting in the cued target detection (CTD) task described previously. Attentional orienting is faster in monkeys given injections of nicotine and in humans who smoke cigarettes. The effect in monkeys is reversed by the nicotinic antagonist mecamylamine, suggesting that it is specific to the nicotinic synapse. Blockade of the muscarinic receptors has a similar effect on the CTD.

If the cholinergic system is involved in arousal, then neurons of the basal forebrain system should become active in tasks that demand increased arousal or attention. In fact, neurons of the basal forebrain nucleus respond during a behavioral choice task that requires either the execution of a hand movement or withholding of the response. Most neurons responded during both tasks, indicating that their activity was linked to some aspect of the decision process or to increased arousal preceding the response.

F. Serotonin

1. Serotonin's Projections to the Brain Are Widespread

There are nine sources of SE within the brains of mammals. All lie in the brain stem and those from the dorsal raphe nucleus comprise 80% of the SE innervation of the cerebral cortex. SE projections contact all the major cortical areas of the brain, but whether individual serotonergic neurons send signals to all areas of the cortex (as do NA neurons) or to more restricted areas (as do cholinergic neurons) is not known. In the visual cortex, the terminals of SE neurons are found in layers avoided by NA neurons.

Current classification schemes list seven main types of serotonergic receptors (5HT-1–5HT-7), each of which has several subtypes. The subtypes appear to be located in different areas of the brain, suggesting that they have different functions. For example, only 5HT-1 receptors are found in large numbers in the hippocampus, suggesting that they are involved in the storage of information.

2. Serotonin's Effects on Cells

The effects of application of SE to cortical neurons depend on the receptor types stimulated. 5HT-3 receptors are ionotropic and applied SE produces immediate excitation of the cell, probably by activation of sodium currents. All other SE receptors are G protein coupled and applied SE produces either excitatory (inhibition of K^+ currents) or inhibitory (activation of K^+ currents) responses that have slower onsets and longer duration than those responses seen in 5HT-3 neuron).

3. Serotonin Plays a Major Role in a Wide Variety of Behaviors

SE is found in the tissues of vertebrates, invertebrates, plants, and fruits. In vertebrates, it is found in many organ systems (e.g., gastrointestinal tract, lungs, and skin) and the peripheral and central nervous systems. It is the precursor for the melatonin produced in the pineal gland.

SE plays a role in many behaviors, including eating, sleeping, and emotional balance. For example, deficits in CNS SE are associated with depression. Interestingly, there is evidence that women have only half as much SE as men, which may correlate with the increased incidence of female depression. Drugs that block the reuptake of SE into the neuron may dull the appetite, alter sleep patterns, and alleviate depression. The range of symptoms successfully treated with SE-altering drugs is remarkable.

Activity in serotonergic neurons of the dorsal raphe nucleus was once thought to trigger slow-wave sleep (SWS). However, recent work has shown that stimulation of the dorsal raphe nucleus increases extracellular SE and its metabolites, electrocortical arousal, and REM sleep. Thus, increased synaptic serotonin appears to be linked with REM sleep, rather than SWS, and perhaps arousal during the awake state. Since SWS alternates with REM sleep, inhibition of the

raphe nucleus or blockade of SE receptors should result in fewer episodes of slow wave cortical activity. This interpretation, however, is inconsistent with sleep patterns observed in mice that are born lacking one subtype of serotonergic receptor (5HT-1B). During their sleep periods, these animals spend more time in REM sleep and less time in SWS than do wild mice. Thus, evidence suggests that SE exerts an inhibitory influence on REM sleep.

G. Histamine

1. The Distribution of Histamine in the CNS Is Widespread

HA neurons, originating in five nuclei within the tuberomamillary nucleus of the hypothalamus, innervate almost every structure in the brain stem and cerebral cortices. The densest innervation is to the hypothalamus and septum. Somewhat lower densities are found in the cortex and amygdala, and the lowest is seen in the hippocampus, brain stem, and cerebellum. At least three types of HA receptors (H_1–H_3) have been identified in the brain. In addition, HA is present in nonneural mast cells of the brain.

2. The Cellular Effects of HA Resemble Those of NA and ACH

The effects of HA on cells can be inhibitory or excitatory, like the majority of NMTs. Although the ionic mechanisms of these effects are less well-known than those of NA or ACH, at least some of the excitation is produced by reductions in K^+ currents, leading to a faster recovery after an action potential. For example, application of H_2 agonists to thalamic and cortical neurons causes excitation by reducing the afterhyperpolarization currents; H_1 agonists may increase excitation by producing slow depolarizations. However, H_3 receptors are probably located on histaminergic terminals and limit HA release by autoreceptor inhibition.

3. Histamine's Action Depends on the Receptors It Stimulates

HA administered intravenously or intracerebrally induces an EEG arousal response and spontaneous locomotor behavior, including exploration and grooming. These changes are blocked by H_1 antagonists, an effect consistent with the cellular effects of these drugs. H_3 agonists cause animals to spend a longer amount of time in SWS and the increased time can be blocked by either H_3 agonists or H_1 antagonists.

The dynamics of the cortical EEG are mediated normally by thalamic mechanisms. Application of HA to thalamic neurons results in a slow depolarization of the membrane produced by reductions in K^+ currents. This switches synchronized activity of individual neurons to the single-spike mode of repetitive firing. Recent evidence indicates that HA's major contribution is to the regulation of arousal rather than the control of alerting or vigilance.

Part of HA's effects may be due to interactions with cholinergic neurons of the basal forebrain and laterodorsal tegmentum as well as the LC. Application of HA excites basal forebrain and tegmental cholinergic neurons through stimulation of H_1 and H_2 receptors. These findings suggest that ACH and HA may synergistically increase cellular and behavioral arousal.

H. Interactions between the NMTs Are Common

Many of the results of the behavioral studies cited previously are interpreted as though a single NMT system was affected by the experimental drug. These interpretations are true in an approximate sense only. Even if drugs bind tightly and selectivity to a single receptor (they frequently do not), more than one neuroactive substance may be affected through simple postsynaptic effects and reuptake mechanisms. For example, if NA is infused into cortical tissue-containing cells and axon terminals of DA neurons, microdialysis of that tissue records increases in both NA and DA. The NA increase is expected since it was added by the experimenter. However, what caused the increase in DA? The increase cannot be caused by the action of NA on receptors on dopaminergic cell bodies because none are present. The answer is that both NA and DA are recycled into the presynaptic terminal of the DA neuron, which causes the synaptic release of DA to increase. Thus, drugs applied by the systemic route are likely to interact with several neurotransmitter systems. Indeed, there are only a few instances in which a single transmitter fails to affect all other transmitters.

Other examples of cross talk can be found between cholinergic and monoaminergic systems. For example, applications of nicotine in the vicinity of the axon terminals of noradrenergic neurons facilitate noradrenergic release and glutamatergic neurotransmission. In addition, GABA, an inhibitory neurotransmitter,

exerts tonic inhibition on the cholinergic system. Lastly, both DA and SE regulate the release of ACH in the cortex.

I. NMT Systems Share Many Properties

Many of the cells and cell groups within the ascending systems share common properties. Indeed, the arousing systems differ in the neuroactive substance they release but otherwise produce similar net effects on cortical tissues. The similarities between NMT systems are summarized as follows:

Projections: The cell bodies reside in the brain stem and project rostrally to cortex and subcortical structures. Each system influences thalamic and cortical structures, and both NA and SE systems project caudally to the spinal cord as well. Several of the systems affect pathways to the cortex at two or more locations.

Anatomical connections: Each system makes synaptic contacts with multiple cortical areas. The noradrenergic and serotonergic neurons have the broadest divergent connections, followed by lesser divergence in the cholinergic, dopaminergic, and histaminergic systems.

Cortical innervation: A given cortical area is usually innervated by two or more systems. In the posterior cortex, the inputs from different NMTs may be interdigitated across the cortical laminae. In more anterior regions, this lamination is much less evident.

Multiplicity of receptors: Each system exerts its effects on families of neurotransmitter receptors that produce a variety of pre- and postsynaptic effects. In the cortex, receptors subtypes are frequently located in different cortical laminae. The action of the transmitter on these receptors produces either excitatory or inhibitory effects, but the onset and duration of these responses vary widely.

Phasic and tonic responses: Many of the cells in the ascending systems register simple sensory events by phasic responses and more complex external and internal events and overt behaviors with variations in tonic activity. In addition, several (e.g., NA, ACH, and SE) have spontaneous activity that increases with increasing arousal.

Global and local effects: Arousing systems are capable of both global and local control of target structures. In other words, they are capable of influencing either narrow or broad areas of the cerebrum as conditions dictate. This is due, in part, to topographic projections within many of these systems so that restricted input leads to restricted output. Generally, global control may be mediated by postsynaptic effects. Within nontopographic systems (e.g., the LC), modulation of the preterminal receptors produces a spatially limited release of neurotransmitter.

Interactions: The arousal systems facilitate not only intrinsic subcortical and cortical activity but also other arousal systems. They may also inhibit activity in another system. For example, activity in the LC is inhibited by the cholinergic system during REM sleep.

J. Despite Similarities, Arousing Systems Also Display Unique Properties

There are considerable differences in the biochemical pathways through which NMTs act on target neurons. NMTs linked directly to ligand-gated mechanisms (e.g., 5HT-1 and glutamate) produce fast postsynaptic responses with short durations. In contrast, those linked to second messenger systems (ACh, NA, DA, SE, and HA) have slower onsets and longer duration. The latter systems may be further divided into classes based on the postsynaptic currents they activate. For example, muscarinic receptor stimulation causes a decrease in K^+ currents, β-adrenergic receptor activation decreases Ca^{2+}-activated K^+ currents (I_{AHP}), and 5HT receptor stimulation causes a decrease in the voltage-sensitive K^+ current I_M.

VI. BEHAVIOR

We now discuss the roles of the NMT systems in the global regulation of brain activity and behavior. To understand these roles, it is necessary to synthesize a large body of data. This is best done by the construction of frameworks or theories that link consistent observations and generate testable predictions. One theory suggests that during arousal, each NMT system provides control over synchronization frequency for different areas of the brain. According to this view, the entire brain shows similar resting frequencies; high-frequency activity appears locally in those parts of the brain that are controlling behavior. Support for this theory is lacking. Studies that examine correlations between neurons during sensory stimulation have

found only brief periods of synchrony. The high-density arrays of EEG electrodes that are currently available or the newer magnetoencephalographic technology can be used to test this hypothesis.

Another hypothesis that has received substantial empirical support is that each NMT system mediates different response components of behavior. This hypothesis arose from research conducted with an attentional task known as the five-choice, serial RT task. This task allows independent assessment of error rate, frequency of premature responses, number of response omissions, and response latency. Using this task, the effects of damage to NMT systems are tested in a highly systematic manner. It was found that cholinergic lesions impair the ability to discriminate visual targets (increased error rate). In contrast, NA lesions impair accuracy of responding, especially in the presence of distracters. Lesions of DA systems impair overall speed and probability of responding. Finally, 5HT lesions produce a reduction in premature responding. It is likely that some of the deficits are due to changes in alerting as well as alterations in attentive behavior.

The primary deficits in one response measure were accompanied by minor changes in the other measures. Perhaps this is due to incomplete independence of the measures. For example, impulsive responding to incorrect stimuli should increase error rates and decrease response latency. In any case, the results are consistent with the central tenet of this review that patterning of activity within NMTs is responsible for distinct aspects of aroused behavior.

A. Properties of Arousal Systems Explain Some Aspects of Behavioral Performance

It is important to understand not only how the NMTs are associated with different kinds of behavior but also how their activity causes variations in a specific behavior. The classic description that relates arousal to cognitive performance is Yerkes–Dodson law, which states that the optimal performance on cognitive and motor tasks depends on a moderate level of arousal or vigilance. Low and high levels produce suboptimal performance. When graphed, Yerkes–Dodson law resembles an inverted U. The decline in performance at high levels of arousal suggests the recruitment of inhibitory processes that reverse the beneficial effects of optimal arousal.

The most likely neural structure that mediates vigilant behavior is the LC. Optimal vigilance means that sensory information is processed efficiently and irrelevant stimuli are ignored. Two properties of the LC/NA system seem well suited to the task. First, application of NA to cortical cells causes an increase in the signal-to-noise ratio, as discussed previously. Second, by increasing NA at sensory sites, the LC filters out signals that differ only slightly from the background activity. Therefore, better filtering of meaningful stimuli is a natural consequence of optimal arousal levels.

LC neurons also show "baseline" activity that varies monotonically with arousal level in the absence of vigilant behavior. One consequence of this tonic activity is that increased release of NA onto thalamic neurons increases their activity in a monotonic manner. These increases are passed on to cortical neurons and presumably account for better performance. What causes the decline in performance with high arousal? There are several possibilities, including higher thresholds for β receptor-mediated inhibition, recruitment of inhibitory GABA interneurons, and stimulation of presynaptic (α_2) autoreceptors. A definitive answer to this question remains to be provided.

Another quantitative aspect of performance to be explained is that in repetitive tasks, performance worsens over time. In one such task, printed letters are rapidly presented to human observers who report when a letter is repeated. Performance of this task varies with arousal level. Early in the task period, subjects are vigilant and performance is high; later, it is lower, presumably a result of the decline in arousal. Drugs that affect NA increase arousal and improve performance, decreasing the number of missed pairs. This suggests that the signal was more detectable, which is consistent with the cellular effects of NA.

B. Recent Studies of Alerting Implicate Noradrenaline

Most of the foregoing discussion focused on the roles of NMTs in arousal and vigilance. However, the LC/NA system is also involved in alerting in that it reacts phasically to unexpected environmental stimuli. If the LC is damaged, alerting responses to intense, novel stimuli are impaired.

A particularly attractive paradigm for studying alerting is the cued target detection task developed by Michael Posner. This task, which was developed to study covert movements of attention, investigates the

effects of visual cues on reaction times to peripheral targets. Two cues, valid and invalid, are included to assess the impact of spatial information on the speed of covert orienting. Two additional cues, neutral and no-cue, are included as controls to specifically assess alerting effects. The difference in reaction times between valid and invalid trials is termed the validity effect, and the difference between neutral and no-cue trials is known as the alerting effect. The question is how altered levels of NA or DA affect the validity and alerting effects. The results confirm that lowered levels of NA produced by α_2 receptor blockers significantly reduce the size of the alerting effect. α_2 agonists have the opposite effects. None of these substances alter the validity effect. These results are also consistent with those of studies using cellular monitoring of LC activity, which suggest that LC activity is associated with attentional lability. That is, when the subject attends to the fixation point, presentation of the peripheral stimulus evokes a major LC response because it is unexpected and located away from the current focus of attention. The presence of the adrenergic agonist, however, reduces LC activity and produces a state of attentional fixity within which reaction times are slowed.

VII. REGULATION OF AROUSAL SYSTEMS

Arousal systems are regulated not only by external stimuli and other arousal systems but also by control systems of the brain. For example, the frontal cortex, particularly the orbitofrontal area, regulates the thalamic reticular nucleus and the cholinergic, basal forebrain structures. Patients with lesions in this area show deficits in arousal. Cortical control is probably not limited to cholinergic modulation. Recent work in the noradrenergic system suggests that the frontal cortex regulates impulse activity in LC neurons as well. The frontal cortex also exerts an influence on the limbic system, which regulates emotional arousal. In the condition known as akinetic mutism, patients may be immobile for hours but when aroused are capable of a wide range of behaviors. This condition is a result of lesions in the mediodorsal frontal cortex and damaged connections with the limbic system. The anterior cingulate region is also important in the self-regulation of arousal through its connections with the cholinergic basal forebrain. In summary, the frontal cortex acts as an executive to regulate the brain stem influences on cortical excitability.

VIII. SUMMARY

Arousal is a physiological and behavioral response to external or internal events (threat, pain, reward, etc.). The arousal response can be divided into three categories based on the duration of response: short-term alerting, long-term vigilance, and longer term arousal. Each category represents an interplay of mechanisms in the peripheral and central nervous systems. Arousal is a result of the release of modulatory neurotransmitters that control target structures (e.g., brain and internal organs) that contain a large variety of receptors. Although all the arousal neurotransmitters can induce EEG arousal, they play different roles in behavioral arousal. The noradrenergic system mediates alertness to sensory events and regulates the vigilant state of the organism. The cholinergic system increases behavioral and cognitive arousal, thereby facilitating attention and the encoding of information. The dopaminergic system facilitates the execution of appropriate behavioral responses to external and internal events. The serotonergic system mediates motivational behaviors such as food consumption and sexual behavior, maintains adequate activation of mood, and regulates REM sleep. Histamine's function is less clear but, like acetylcholine, it regulates cortical and subcortical excitability.

See Also the Following Articles

ALERTNESS • ATTENTION • BIOFEEDBACK • CHEMICAL NEUROANATOMY • ELECTROENCEPHALOGRAPHY (EEG) • PSYCHONEUROENDOCRINOLOGY • STRESS • VIGILANCE

Acknowledgments

Preparation of this article was supported by grants from the National Institutes of Health and the Oregon Medical Research Foundation. We thank Dr. Don Tucker for valuable comments on the manuscript.

Suggested Reading

Broadbent, D. E. (1958). *Perception and Communication*. Pergamon, New York.

Kohlshütter, E. (1862). Messungen der Festigkeit des Schlafes. *Z. Rationalle Med.* **17,** 210–253.

Metherate, R., Cox, C. L., and Ashe, J. H. (1992). Cellular bases of neocortical activation: Modulation of neural oscillations by the nucleus basalis and endogenous acetylcholine. *J. Neurosci.* **12,** 4701–4711.

McCormick, D. A. (1992). Neurotransmitter actions in the thalamus and cerebral cortex and their role in neuromodulation of thalamocortical activity. *Prog. Neurobiol.* **39,** 337–388.

Moruzzi, G., and Magoun, H. (1949). Brain stem reticular formation and activation of the EEG. *Electroenceph. Clin. Neurophysiol.* **1,** 455–473.

Powis, D. A., and Bunn, S. J. (1995). *Neurotransmitter Release and Its Modulation: Biochemical Mechanisms, Physiological Function, and Clinical Relevance*. Cambridge Univ. Press, Cambridge, UK.

Robbins, T. W., and Everitt, B. J. (1995). Attention and arousal. In *The Cognitive Neurosciences* (M. S. Gazzaniga, Ed.). MIT Press, Cambridge, MA.

Artificial Intelligence

DEREK PARTRIDGE
University of Exeter, United Kingdom

GLOSSARY

cognitive level A viewpoint in which structures and mechanisms to support a scientific interpretation of intelligent behavior bridge the gap between neurophysiology and intelligent behavior. Structures and mechanisms at the cognitive level may or may not be found to have a simple mapping to the structures and mechanisms of brain anatomy.

heuristic A rule of thumb used to generate adequate solutions to intractable problems.

knowledge base A set of ''if–then'' rules and facts that captures the information necessary to support intelligent activity within some circumscribed domain.

neural network A computational system composed of highly interconnected simple processing units that operate in parallel to process all incoming activity values, compute an output activity value, and send it to all directly connected neighboring units. A unit's output value transferred by each interunit link is modified by the ''weight'' (a real number) associated with each link. It is this set of link weights that is altered during the initial network training.

Artificial intelligence (AI) has no firm definition. The term refers to attempts to simulate or replicate with computer systems functions that are normally assumed to require intelligence when manifest in humans. The goal of AI is to reproduce some aspect of human intelligence within a computer system. Thus, the bulk of AI work consists of programming computers, although significant engineering can be involved in important subareas such as robotics. Intelligence is thus seen as running the right program on an adequate computer, and the grand goal of AI is to write that program.

I. APPROACHES

There are a variety of approaches to the question of what aspect of human intelligence to attempt to simulate or replicate with a computer system. The basic division, which is particularly germane for brain scientists, is whether the approach is to reproduce functional equivalence with no regard for the mechanisms and structures of the human brain or whether it is important to reproduce structure as well as function.

A. Functional Equivalence

Within a functional equivalence approach to AI the goal is to reproduce (or supercede) some aspect of human intelligence, such as medical diagnosis or chess playing, by using whatever means are available. It is believed, for example, that the best human chess players examine only a few of the many possible move sequences that may be available at any given point in a game. The success of chess-playing computers has been founded on very fast exploration of thousands of possible move sequences. The result is very high-quality chess games from these machines but based on

mechanisms that would not appear to be used by humans who exhibit equally impressive chess-playing behavior.

The bulk of work in AI has taken this purely functional approach for two good reasons. First, the mechanisms underlying human intelligence tend to be options for debate rather than secure knowledge. At the cognitive level, the mechanisms and structures, such as short-term and long-term memories, are by their very nature difficult to describe in detail with certainty. Also, although details of brain physiology may be asserted unequivocally, the uncertainty lies in their roles in the production of the aspects of human intelligence being reproduced. The second reason for neglect of structure and mechanisms is the suspicion that the human brain may not be an optimum architecture on which to build intelligence. Its basic design evolved long before it was required to support intelligent behavior, and there are many seemingly obvious weaknesses in the human brain when viewed as a computational machine, such as error-prone memory retrieval. Therefore, most of the work in AI must be viewed as attempts to reproduce some aspects of intelligence without regard for the way that the human brain might be thought to be operating to produce human intelligence.

B. Structural Equivalence

For those researchers of AI who do claim some measure of structural validity to their work, two levels of structure and mechanisms may usefully be distinguished: the cognitive level and the level of brain anatomy. At the cognitive level, the psychologist posits structures and mechanisms to account for the observed behavior, and although these objects may have no clear implementation in terms of brain anatomy, it is a requirement that such a mapping could be plausibly found. The other level of structure is that of brain anatomy: the structures and mechanisms that can be physically observed, probed, and generally examined. At the cognitive level, we deal with abstract structures such as short-term memory and mechanisms such as forgetting caused by attention shift. At the physical level the structures may be an observed pattern of interconnectivity of neurons, and the mechanisms may be pulse frequency between the component neurons. Neither level is simple, nor are they always distinct, but this twofold division is a useful approximation.

1. Abstract Structure: Cognitive Science

The majority of AI work that has not eschewed structural equivalence has accepted a structural model of the brain at the cognitive level, and much of the AI work in this subarea is also known as cognitive science. Alternatively, most of the AI work in this subarea is aimed at evaluating and exploring cognitive models of human intelligence.

2. Physical Structure: Cognitive Neuroscience

There is a relatively small but growing body of work in AI that does accept constraints that derive from anatomical knowledge of brain structure and function. The impetus for this type of approach to AI has been the surge in interest and activity based on computation with neural networks. A neural network is a computational system that is founded on simple communication between highly interconnected networks of simple processing elements, i.e., a computational basis that appears to reflect some aspects of brain physiology. Conventional programming is also used to construct AI models of this type, but it tends to be restricted to the reproduction of a limited aspect of human intelligence. A system might, for example, attempt to reproduce certain aspects of human learning behavior based on a model of synaptic facilitation at the molecular level.

II. COMPUTATIONAL PARADIGMS

In addition to the type of approach taken, work in AI can be distinguished by the computational paradigm used. By this I mean the style of computational mechanism, both in general and in the specific class of programming language used (e.g., procedural versus declarative languages).

A. Classical Programming

Classical or conventional programming has dominated AI (as all other forms of computer usage). It is characterized by the processes of first deriving an algorithmic procedure for achieving some desired goal within the accepted constraints and then casting this algorithm into a machine-executable form [i.e., the classical task of writing (or coding) a program]. Typically, the elements of the program are the conceptual objects of the problem [e.g., a small array

of storage locations might be labeled short-term memory (STM)], and the algorithmic mechanisms programmed will directly implement the conceptualizations of the way the model works (e.g., subparts of the program using STM will be direct implementations of storing and removing objects from the STM object in the program). In this classical usage of computers, the speed and accuracy of the computer are used to perform a series of actions that could, in principle, be performed by hand with a pencil and paper. The main drawback from an AI perspective is that the programmer must decide in advance exactly what detailed sequence of actions will constitute the AI system.

1. Procedural Programming

Most programming languages are procedural languages; that is, they permit the programming to specify sequences of action, procedures, that the computer will then execute. Several programming languages have a particular association with AI. One such language is LISP, which dominated the early decades of AI. LISP serves the needs of AI by being a general symbol-manipulation language (rather than a numeric computation language such as FORTRAN); it is also particularly useful for AI system development because it demands a minimum of predefined conditions on the computational objects, which permits maximum freedom for changing program structure in order to explore an AI model by observation of the computational consequences. This flexibility minimizes automatic error checking and can lead to fragile programs that may also be large and slow to execute. It is thus not a favored language and method for mainstream computing, in which the emphasis is on fast and reliable computation of predefined procedures.

2. Rule-Based Systems

Early in the history of AI it was realized that the complexity of AI models could be reduced by dividing the computational procedures and the objects they operate on into two distinct parts: (i) a list of rules that capture the knowledge of some domain and (ii) the mechanisms for operating on these rules. The former part became known as the knowledge base and the latter as the inference engine or control regime. The following might be a simple rule in a chess-playing system: IF a potential exchange of pieces will result in the opponent losing the more valuable piece THEN do the exchange of pieces.

The general form of such rules is IF ⟨condition⟩ THEN ⟨action⟩. When the ⟨condition⟩ is met, then the ⟨action⟩ is executed, which generally results in new conditions being met and/or answers being generated. The control regime might be that the first rule whose condition is satisfied is executed, with this simple control repeated until some goal is achieved or no rules can be executed. This strategy for building AI systems became inextricably linked with the AI subarea of expert systems.

3. Logic Programming

Development of the knowledge-base and inference-engine division was formalized in the idea of logic programming. The basis for logical reasoning, which has always been an attractive foundation on which to build AI, is a set of axioms and the mechanism of logical deduction, and this pair map easily onto the knowledge base and inference engine, respectively. This particular development was implemented in detail in the logic programming language PROLOG, which became a popular AI language in the 1970s, especially in Europe and Japan.

B. Neural Computing

Neural networks are computational structures that are composed of many simple, but highly interconnected, processing units or nodes. Each interconnection, or link, is associated with a weight that is a number that controls the strength of the link; for example, with the weight as a simple multiplier, positive weights larger than 1.0 will increase the activity value from one node to the next. There are many varieties of neural network in use and most are trained in order to generate the desired system. Training consists of adapting an initial set of link weights (initially set as random numbers) so that the network reproduces a training set of input–output pairs—the training set, a set of examples of the behavior that we wish to simulate. The training procedure (which consists of adjusting the link weights until the network reproduces the training set) is an automatic algorithmic process. The training cycle typically consists of inputing a training pattern, executing the network, calculating the error between what the network computes and the correct result for this particular input, and adjusting the link weights to reduce this error. It is not usually possible (and not desirable) for a given network to reproduce the training set exactly. Training is thus repeated until a

measure of the error between network performance and training set is acceptably small. The training algorithms usually guarantee that this error will not increase with further training, but there is no guarantee that it will always be reducible to the desired value.

This process of AI system development—random initialization and repetitive training—is clearly different from that of classical programming and it does not always succeed. At best, training only promises a local minimum error, which means that training can become stuck with no further improvement from successive repetitions of the training cycle, but the network is a poor approximation to the training data.

Figure 1 presents a neural network known as a multilayer perceptron. It is a system designed to predict risk of the bone degradation disease osteoporosis. It can be used to prioritize patients for screening in order to improve early detection of the disease. The inputs are 15 risk factors (measured from a given patient) and the single output is a measure of how likely the patient is to develop osteoporosis, with 1.0 the highest likelihood and 0.0 the lowest. Approximately 500 patients both with and without osteoporosis were used to train the network, and it proved to be approximately 72% correct on a set of 200 previously unseen cases.

The network structure and mechanisms illustrated in Fig. 1 bear little more than a tenuous analogical relationship to actual brain structures. Other uses of neural computing have accepted more realistic constraints from neuroanatomy. Figure 2 presents a simplified version of a system that was designed to recognize shape from shading (i.e., to reproduce the human ability to perceive a three-dimensional shape from a two-dimensional pattern of light and dark). The neural net system used was built from hexagonal arrays of formal neurons that were configured to mimic the various known receptor types in the optic tract (e.g., circular receptive fields).

Although AI systems constructed with neural computing methods are necessarily approximations, they offer the advantage of not requiring a prior problem specification. For many AI subareas the details of how the input information is transformed into an "intelligent" output are incompletely understood. In fact, the essence of the problem is to develop such an

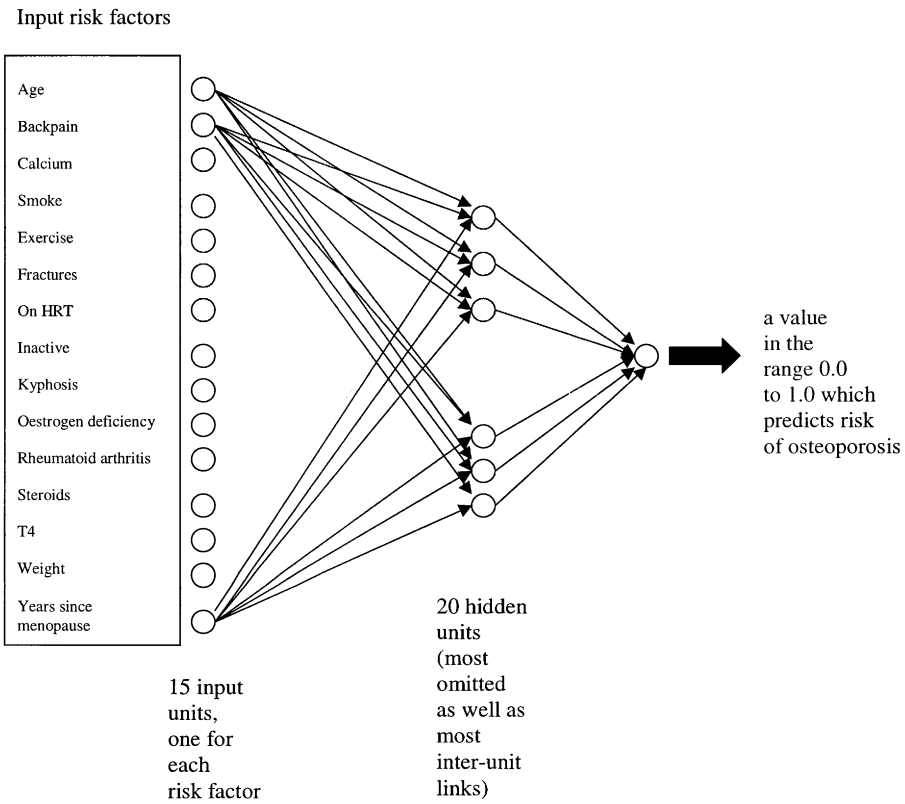


Figure 1 A multilayer perceptron.

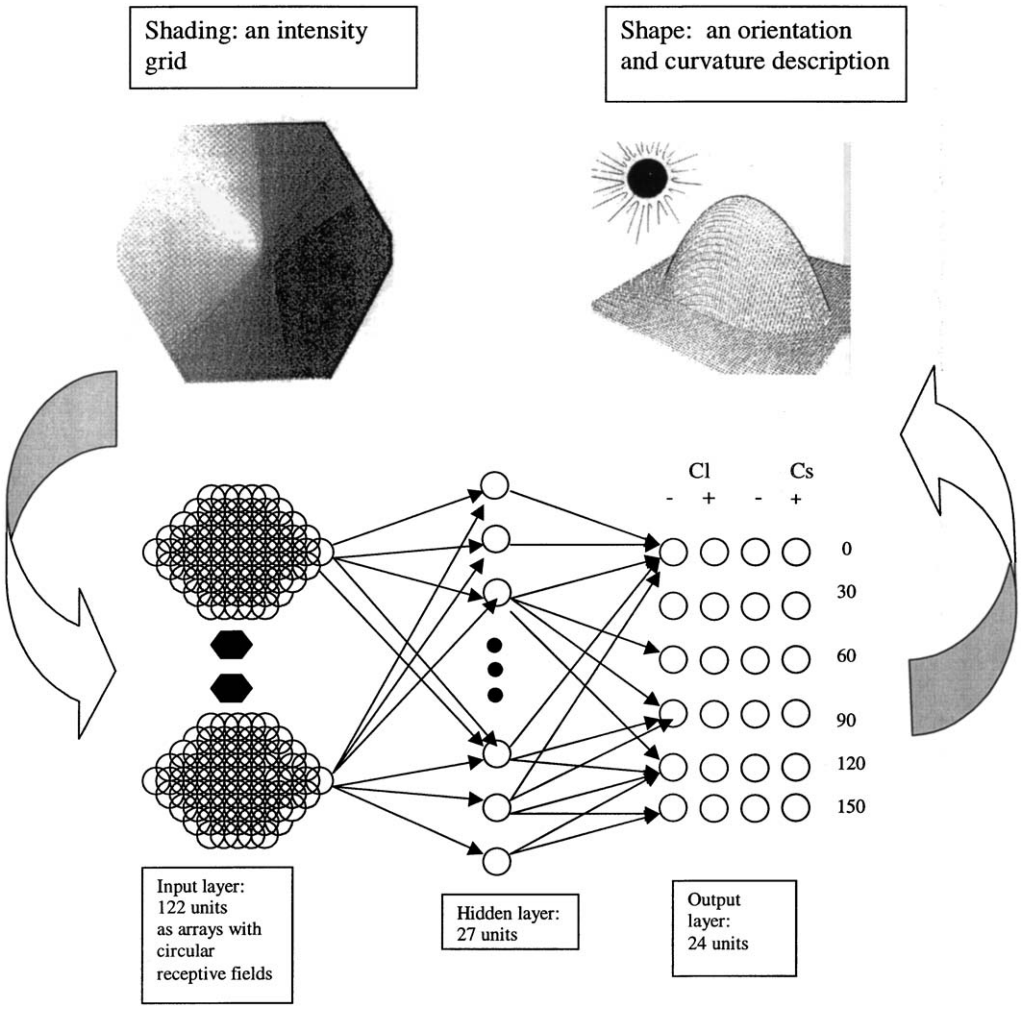


Figure 2 A system designed to recognize shape from shading. [Adapted from Lehky and Sejnowski (1988). Network model of shape from shading: Neural function arises from both receptive and projective fields. *Nature* **333,** 452–454.]

understanding. A data-driven technology, such as neural computing, that enables the construction of a system from a set of examples is thus particularly useful.

Neural computing is a powerful technology but it tends to suffer from lack of transparency: Trained neural networks are resistant to easy interpretation at the cognitive level. In the trained neural network illustrated in Fig. 1, for example, all the individual risk factors are totally and equally (except for the individual link weightings) intermingled in the computation at the hidden layer (all input nodes have a connection to every hidden-layer node). Therefore, how is each risk factor used in the computation of disease risk? It is impossible to determine from the neural net system; its computational infrastructure cannot be ''read off'' as it can be for a classical program. However, different styles of explication are emerging: It is possible, for example, to associate a confidence with each network prediction based on an analysis of the particular input values with respect to the set of values used for training.

Another result of the total intermingling of all the input features and the fact that the internal structure of a neural network is nothing but units with different combinations of weighted links in and out is that no conceptually meaningful representation can be isolated within the computational system. In the osteoporosis system, for example, the input feature ''age'' is assessed in comparison to some threshold in order to add its contribution to the final outcome. However, no comparison operations or threshold values are evident within the trained network. It is usual to say that such

computational mechanisms use distributed representations. In contrast to neural networks, there are other inductive, or data-driven, technologies that operate directly at the cognitive level.

C. Inductive Methods

An inductive method of system building is one that produces a generalized computational system from a set of specific examples. Neural computing offers one general class of inductive method, and rule-induction algorithms provide another. From a set of examples, a rule-induction algorithm will construct a decision tree, which is a computational system that embodies general rules that encapsulate the classifications (predictions, decisions, diagnoses, etc.) provided in the set of examples. Just as with neural networks, a decision tree cannot always be constructed, and when it can it is unlikely to be 100% correct even when classifying the training examples.

Decision trees tend to train faster than neural networks, but they tend to be more dificult to train to low error because they are more sensitive than neural networks to the choice of input features. However, the final decision tree is structured as sequences of decisions at the cognitive level; therefore, these computational systems may admit an interpretation and an understanding of the computational details of the problem in terms of cognitive-level features.

Figure 3 presents part of a decision tree for predicting whether an aircraft will level off at the flight level it is approaching or not. Such trees have been automatically generated using several thousand examples of aircraft leveling and not leveling at various flight levels at London's Heathrow Airport. The examples consisted of radar tracks (sampled at 4-sec intervals) and the particular flight level that was being approached (e.g. 7000-ft level, which is FL7000). Various features are extracted from the radar tracks; for example, in Fig. 3 d45 is the vertical distance traveled by the aircraft between radar points at interval 4 and interval 5 before the flight level is reached.

The internal structure of this (simplified) decision tree is composed of branch choices, and the tree leaves are the predicted outcomes (i.e., level off or non-level off). The branch choice at the top of the illustrated tree tests feature d45: If the value of d45 for the sample to be predicted is less than or equal to 100 feet, then the left branch is taken; otherwise, the right branch is taken and non-level off is immediately predicted.

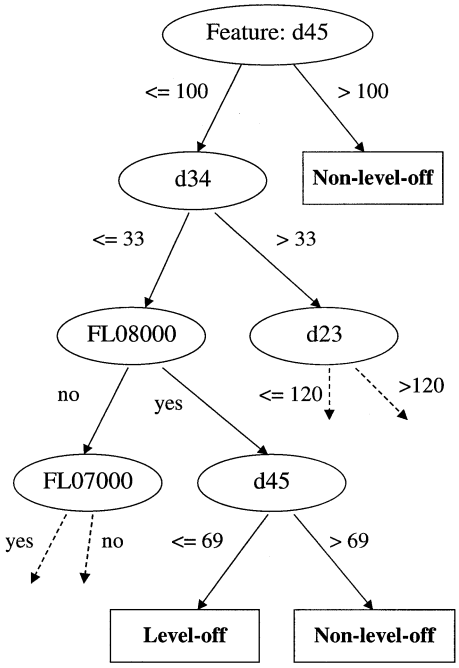


Figure 3 Partial decision tree for predicting whether an aircraft will level off at a certain flight level.

III. SUBDIVISIONS

Intelligence, even if we limit interest to human intelligence rather than intelligent possibilities in some abstract sense, is a broad topic. From the early days of AI (1950s and 1960s) when initial attempts, which dealt with many aspects of intelligence, were revealed to be far too ambitious, the history has been one of scaling down. Consequently, the field of AI has been subdivided into many (often more or less isolated) aspects of intelligent behavior.

A. Knowledge Representation

There is a long-held belief in AI that knowledge is crucial to intelligent behavior. If a system does not have a significant fund of knowledge (about the world or about some particular subarea), then no matter how sophisticated the reasoning, intelligent behavior will not be forthcoming. Quite apart from the difficult questions of what knowledge is needed and how much of it is needed, there is an initial question of how that knowledge must be structured, or represented, within an AI system.

1. Semantic Networks

One enduring belief about how knowledge should be structured is that there should be much and various interconnectivity between the elements of knowledge, whatever they may be. Concepts should be associated with each other with various strengths and according to various relationships. Words, for example, seem to be associated in our brains by semantic similarity and by frequency of cooccurrence.

A representational scheme known as semantic networks is one attempt to provide the necessary associativity. A semantic network is a set of nodes and links between them in which each node is labeled with a concept and each link is labeled with a relationship between the concepts. This scheme can be effective for small and static domains. However, in the wider fields of intelligent behavior it is evident that both the relationships and their strengths are context dependent, which implies that the semantic network must be dynamically reconfigurable. The concept ''chair,'' for example, might be represented in terms of physical relationships as having legs (usually four), a seat, and a back (not always). In terms of functional relationships, it is used for sitting on. However, there are chairs with no legs (e.g., when suspended), and there are chairs that must not be sat on (e.g., in a museum). Neither of these situations would be expected to disrupt normal intelligent behavior, but such exceptional circumstances as well as a detailed computational realization of common qualifiers, such as ''usually,'' cause major problems in AI systems.

Dynamic reconfigurability is the first example of adaptivity, of learning, which seems to be fundamental to intelligent behavior but is a source of great difficulty in computational systems. Neural networks have a fundamental ability to adapt and learn: (they are constructed by adapting the link weights to learn the training examples), but they are severely limited in scope and do not begin to approach the apparent complexities of human learning. In AI, this is considered as yet another subfield, machine learning.

2. Knowledge Bases

A further drawback of the semantic network representation of knowledge is the complexity of the representation. For example, if a node is removed from a network, then all links to it (or from it) must be removed or redirected to other nodes. Therefore, explicit representation of the desired associativity actually works against efforts to introduce dynamic adaptivity.

A knowledge base, or rule base, is a representation of the knowledge as a set of facts and rules that capture much the same information as a semantic network but do so with a set of distinct elements (the individual facts and rules). This does make the addition and deletion of knowledge elements (and hence adaptivity of the knowledge as whole) computationally much simpler. The cost is that relationships between the rules are no longer explicit within the representation. They reside implicitly in the rule ordering in conjunction with the searching strategies.

3. Expert Systems

Despite the loss of explicit representation of the interrelationships between the various elements of knowledge, knowledge bases have become firmly established as a useful way to represent knowledge in relatively limited, fixed, and well-structured domains of human endeavor. They have provided the basis for the large and relatively successful subfield of expert systems.

An expert system reproduces (or even exceeds) human capability in some specialized domain, some area of human expertise. Initially, the use of knowledge bases with appropriate control strategies so dominated the expert systems subfield that this representational scheme was used to define an expert system. Subsequently, it has been accepted that it is the reproduction of intelligent activity within a very limited domain that characterizes an expert system, whatever the means of realization. However, it is not uncommon to find the terms expert system and knowledge base used with virtual synonymity.

B. Machine Learning

The ability to adapt to changing circumstances (i.e., to learn) appears to be fundamental to the nature of intelligence. However, the construction of computer programs is a difficult and demanding exercise that usually requires that much time be spent on what is euphemistically called debugging—tracking down and eliminating errors in a program. The intricacies and complexity of debugging can escalate dramatically if a program is self-modifying (i.e., if it learns), because learning changes some aspects of the original program. This means that when a programmer needs to delve into a learning program to alter or correct its behavior, he or she first has to understand the detailed results of the learning in addition to the complexities of the basic

program as originally written. Consequently, most AI systems have little or no learning behavior built into them. Explorations of mechanisms to support adaptive behavior tend to be pursued as an end in themselves, and these explorations constitute the AI subarea of machine learning.

1. Rote Learning

Rote learning is the process of memorizing specific new items as they are encountered. The basic idea is simple and easy to realize within a computer program: Each time a new and useful piece of information is encountered, it is stored away for future use. For example, an AI system might be designed to recognize faces by extracting a variety of features (such as distance between the eyes) from an image and searching for a match within a database of 1000 stored feature sets. If it finds a match, it has recognized the person; if not, it reports "unknown person." In this latter case, the person or some human operator can provide the necessary details of this unknown person, and the system can enter the details in its database so that next time this person is presented to the system he or she will be correctly recognized.

However, every rote learning event increases the size of the database, and speed of recognition (as well as accuracy if new faces are similar to old ones) will decrease as the size of the database grows. Depending on the application of this system, certain new faces may be seen once and never again, in which case rote learning wastes time and space. Alternatively, this new face may be commonly seen at first but may over time become rare and ultimately no longer seen. In the first case, rote learning is immediately detrimental to system performance, and in the second case it becomes detrimental over time.

Rote learning (and in general any sort of learning) implies some compensating mechanism of forgetting, which suggests that the apparent human weakness of imperfect recall may be a fundamental necessity to long-term intelligence. AI has largely been preoccupied with learning mechanisms usually without the compensating mechanism of forgetting.

The second general point is that most situations change over time, so an AI system with any longevity will have to be able to track and react appropriately to the natural drifts in the problem domain. This is not an easy problem to solve. The fundamental difficulty is one of distinguishing between short-term variation and long-term drift when both short term and long term are ill-defined context-dependent quantities. If after appearing twice a day, a face is not seen for a week, should it be deleted? There is no correct answer and probably no more than guidelines even if the full details of the application of this face recognition system are known.

Therefore, a good rote learning system must exhibit flexibility and sophistication. It cannot simply store every new event and delete every one that is not used during some specified period of time. Human intelligence appears to embody some mechanisms of partial recall that, like forgetting, may not be weaknesses; for example, learned faces may gradually fade over time if not used but may be more quickly and easily reestablished if they reappear before they have been totally forgotten.

Neural networks, in which input information is blended and distributed across the network (distributed representation), do offer hope of a mechanism for partial learning and forgetting. Currently, however, sophisticated control of such abilities proves elusive. Classical programs appear to be less amenable to schemes for partial representation.

2. Parameter Optimization

If a problem-solving procedure can be specified in terms of a set of parameters, then mechanisms of machine learning can be used to reset, or learn, optimum values for these parameters. To grossly oversimplify the face recognition system introduced previously, identification of a face might be based on the distance between the eyes (DE), the length of the mouth (LM), and the depth of the forehead (DH). However, what is the relative importance of each of these features? They are probably not exactly equally valuable in the recognition task. The system can learn this information. Suppose we define recognition score as follows:

$$\text{Recognition score} = X \times \text{DE} + Y \times \text{LM} + Z \times \text{DH}$$

where X, Y, and Z are the parameters that determine the relative importance of the three features for face recognition. By trying different values for these parameters against recognition speed and accuracy for a test set of faces, the system can set them to optimal values. Therefore, from some arbitrary first-guess values for X, Y, and Z, optimal values have been learned and a better recognition score will be computed.

Quite apart from the issue of problem drift, this technique presents many difficulties. First, the form of the initial equation severely constrains what can be

learned, and it might be a poor choice. The most important features might have been omitted because their importance was not recognized. More subtly, there may be interactions between features such that attempts to optimize each parameter independently lead to a highly suboptimal system. Finally, this mechanism offers very little scope for the system to learn anything that was not foreseen explicitly by its designer.

3. Inductive Generalization

As discussed previously, inductive generalization is the process by which a general principle is extracted from a set of instances. It is the basis for several computing technologies (neural networks and rule-induction algorithms). It can also be viewed as a method for machine learning quite independent of the technology used to implement it.

The inductive technologies described in Section II are thus ways of building an AI program by learning from a set of examples. Further interest in this general idea has arisen because the inductive technologies are each based on a specific class of induction algorithm, and many other types of induction algorithm remain to be explored and possibly developed into new computational technologies.

There is an unlimited scope for further developments in this subarea of machine learning because there are an unlimited number of ways to extract generalities from a set of instances, and there is no correct way. Inductive generalization is a fragile procedure. The classic illustrative example derives from philosophy: Suppose you see a large white waterbird and are told it is a swan, and then you see another that is also white, and then another, and so on. How many specific instances of white swans do you need to see before it is safe to generalize to the rule that all swans are white? The answer is that the inductive generalization can never be guaranteed no matter how many white swans you see. However, a single sighting of a black swan can tell you that your generalization is false; it represents an apparent contradiction to your rule.

However, you might choose a different interpretation of the black swan sighting. The first being that it is not a swan because it is not white. In this case you have transformed your initial tentative generalization into a firm rule that is used to further interpret the world. An intelligent agent must do this eventually because this is the whole purpose of generating such generalizations. A different rejection of the black swan sighting might be based on the idea that your generalization is essentially correct, it just needs refining—and this is always possible to do. The black swan sighting might have been on a Saturday, so a new (and now uncontradicted) version of your rule might be that all swans are white except on Saturdays.

This last, somewhat absurd dodge is particularly instructive because it highlights the problem of what features to extract from the instances to include in the generalized rule. We know that a swan's color does not depend on the day of the week, but an AI system, with its necessarily limited knowledge, might not know this to be true.

Therefore, the difficulties in contructing algorithms for inductive generalization are determining what features (from the infinitely many possible ones, such as "3 PM on a cloudy day") to extract to focus the generalization on and determining when features from two instances are sufficiently similar. For example, most swans will not appear pure white, and we must accept that swans that are shades of gray or discolored due to dirt are all still essentially the same color (i.e., white). At what point do we have to admit that a bird's plumage is not sufficiently similar to white, and that the bird is thus an exception to our rule? This, like so much else in AI, is an ill-defined and context-dependent issue.

C. Vision and Perception

Sight, although not essential, appears to be an important component of human intelligence. AI vision systems have many potential applications, from automated security camera systems to automatic detection of pathological images in medicine. Because of the accessibility (and assurance of general functional commitment) of the optic tract, mammalian visual systems offer some of the best opportunities for basing AI systems on neuroanatomical data rather than cognitive-level abstractions. Figure 2 is one example of an attempt to build an AI vision system with regard to neuroanatomy.

1. Pattern Recognition

Pattern recognition usually refers to attempts to analyze two-dimensional images and recognize (i.e., label) within them prespecified subareas of interest. Mathematics and statistics feature strongly in this subarea by providing algorithms for noise reduction, smoothing, and segmentation. Processing of images at the pixel level leads naturally into the classic

bottom-up approach to pattern recognition. In this strategy, the idea is to associate areas in the image of similar texture or intensity with features of interest and to associate discontinuities with boundaries that might be developed to circumscribe features of interest. When the image has been segmented, segments may be collected into recognizable objects or may represent the expected objects.

This final process of labeling image segments may be classically programmed or may be learned through a set of examples and the use of an inductive technology. This latter approach has been found to be valuable when it is difficult to specify exactly what is to be labeled (e.g., a tumor in a radiogram), but many examples are available that can be used to train a system.

It has been argued, however, that an intelligent agent generally sees what it expects to see; this is the top-down approach to vision. In this case, an image is scanned with an expectation of what the system wants to find in the image. The image need only be processed where and to whatever depth necessary to confirm or reject the expectation. In the bottom-up approach, it is anticipated that the features will emerge from the image, and in the top-down approach it is hoped that knowing what one is looking for will facilitate quick and accurate recognition or rejection. Of course, it is possible to combine bottom-up and top-down processing.

2. Image Understanding

One point of distinction in AI vision systems is that of finding and labeling objects in an image (from a stored list of possibilities) as opposed to the development of an understanding of an image. The latter, somewhat ill-defined goal seems to be what an AI vision must achieve, and it probably involves some object labeling as a constituent process. Image understanding is the placement of an image in a global context. A pattern recognition system, for example, might find and label a tree, sky, clouds, and grass in an image, whereas an image understanding system might declare that the image is a countryside scene in some temperate climatic zone, perhaps land used for grazing cows or sheep (an implication of the short-cropped grass).

Sophisticated image understanding has not been achieved, except in some very restricted subdomains (e.g., automatic monitoring of the configuration of in-use and free gates at a specific airport terminal). The inductive technology of neural computing has been trained to assess the crowdedness of specific underground station platforms.

In the latter case, a neural net system can only be trained to generate answers in one-dimension, i.e., it can assess crowdedness but no other characteristic of the people such as category (e.g., business or tourist). Other neural nets might be trained to assess other desired features but this implies that the important features are decided in advance of system construction and that a general understanding will only be gained from a set of trained networks. This approach to image understanding implies that a classically programmed harness system will be needed to collect and combine the outputs of the individual neural networks.

Image understanding, however it is achieved, is expected to require substantial knowledge of the world within which the image is to be interpreted. In the previous example of a rural scene, the only basis for drawing the implication about cows and sheep must be knowledge of why grass is cropped in the countryside (as opposed to a lawn, for which the grazing implication is likely to be wrong) and which grazing animals are likely to be responsible in a temperate climate.

Thus, intelligent image understanding will require image processing, pattern recognition, and a knowledge base together with an appropriate control strategy, and all must be smoothly integrated. Like so much else in AI, image understanding beyond very limited and heavily constrained situations is currently beyond the state of the art.

D. Natural Language Processing

Human language, which is generally recognized to be a phenomenon quite distinct from (and in advance of) any known animal communication mechanism, is often considered to be a characteristic feature of intelligence. The Turing Test for AI hinges on whether a computer system can communicate in natural language sufficiently well that a human observer is likely to mistake it for another person; If so, then we must concede that the system exhibits AI. From this widely accepted viewpoint, natural language understanding, both analysis and synthesis, is the key to AI. This view is defensible because, as with image understanding, natural language understanding implies many other AI subfields—a knowledge base, inferential control strategies, adaptivity, and so on.

Natural language processing (NLP), computational linguistics, computational semantics, cognitive linguistics, and psycholinguistics are interrelated areas within which AI work on language processing may be found.

1. Spoken or Written

A first issue for any proposed AI NLP system is the modality of the input (and output): Will the language be written or spoken. Speech, and thus sound waves, as the input medium is normally viewed as a separate pattern recognition task (i.e., the recognition of words within the sound wave pattern). This is not an easy task but considerable progress has been achieved in recent years, particularly with systems that employ some inductive technology and can be trained to the particular idiosyncrasies of a user's voice. Many relatively cheap systems are commercially available for speech recognition, but they do not extend (other than trivially) to language understanding. This is a different and separate subfield, and it typically requires that the language utterances be typed.

2. Syntactic Analysis and Grammars

The formal basis of language appears to be syntactic structure, and such structure is defined with a grammar. Prior to AI, linguists formulated specific grammars to capture the allowable structures of various languages, and they also formulated classification schemes such that a given grammar (and hence the language it purports to specify) could be assigned to a class.

Formal definitional schemes, such as these grammars, were ideal for computerization and were seized on by the first computational linguists. Simply stated, the expectation was that computers could apply such grammars to check the correctness of a sentence in, for example, English, and also determine the syntactic category of each word. From this point a dictionary of word meanings could be applied, and a meaning for the sentence could thus be generated. Even the most optimistic did not think that this would yield perfect understanding, but they thought it would be good enough, for example, to add the dictionary of meanings from another language and thus translate from one language to another. However, like all other aspects of AI, it was not this simple.

Much human communication, especially spoken communication, is ungrammatical. We often communicate in broken structures; furthermore, neither we nor those we speak to appear to be unsettled by or even notice the ungrammaticalities. In addition, there is no fixed grammar for any living language precisely because any language in use evolves and is modified by those who use it. Further complicating this is the fact that a widely used language develops differently at different points in its range. This leads to dialects and eventually to distinct different languages. These problems do not bode well for any AI system that aspires to deal with the NLP problem beyond the confines of a small, well-defined, and fixed subset of some language.

Consequently, there was a movement among NLP workers to develop systems that were not based on a complete and correct prior syntactic analysis. Efforts in this direction have been founded on finding main verbs and then using these to seek parts of the surrounding sentence that fulfill the anticipated roles. Therefore, for example, if a sentence in English contains the word "hit" the implication will be that the subject of the sentence will be a potential "hitter" (which probably requires an animate object) and that the object of the sentence will be something that "is hit." In addition, the sentence could have the role of "what was used for the hitting" specified and introduced by a word such as "with."

Clearly, some initial syntactic analysis can be used to build a framework of further syntactic expectations that may be further constrained by semantic expectations. Continuing the previous example, the hitter is likely to be an active animate object (a person rather than a ball or a bat, which might also be mentioned in the sentence). A shift to consider such constraints on words, rather than simply grammatical categories such as verb or noun, leads us into the realm of semantics.

3. Semantics

Words have meanings, and so it seems reasonable that the NLP researcher might expect to build up through word meanings to sentence meanings, paragraph meanings, and complete document meanings. However, words sometimes have a variety of meanings. Consider the sentence, "The pig is in the pen". The meaning of the word "pen" here must be "enclosure" rather than "writing implement" or "female swan." The semantics of "pen" is disambiguated by the meaning of the sentence as a whole, so the sentence meaning cannot simply be built up from word meanings.

In general, NLP requires a complex interaction between expectations at all levels that yields a most likely meaning when all (or a maximum) of the mismatches have been eliminated. In the previous example sentence the mismatch between large animal "pig" being inside small writing implement "pen" is eliminated by consideration of the enclosure meaning of "pen." However, suppose this sentence is followed by another: "When will that bird learn to stop

swallowing the children's toys." A whole new semantics for the previous sentence emerges, but only if this latter sentence is taken as a further comment on the first sentence and not as a break to a new topic.

Decades of work on AI NLP systems has resulted in progress, but it has also revealed the complexity of language. The surprise is not that we cannot make computers understand anything beyond small and well-defined language subsets, but that we as humans all intercommunicate freely in an unbounded medium. Part of the answer, of course, is that miscommunication occurs, and that communication is seldom perfect (whatever that might mean) because it does not need to be.

4. Pragmatics

If the NLP problem is not daunting enough already, consider that what we might term the direct or explicit meaning of a string of words may bear little relation to the "real" meaning. Said in an ironic tone, "That's smart" may mean "that was a stupid thing to do." Ultimately, the meaning of a natural language utterance is colored by the predisposition of the listener, then by what the listener thinks the utterer is intending to convey, and then by what the listener thinks the utterer thinks the listener will think. There is, of course, no end to this sequence of presuppositions. The utterer may bring an entirely equivalent but quite different suppositions to bear on what he or she intends to mean. Pragmatics is the collection of all these factors beyond and outside of language that influence meaning. AI NLP systems have hardly come to grips with any aspects of pragmatics other than in the context of theoretical analyses of pragmatic issues.

5. Machine Translation

Due to the commercial value of adequate machine translation (MT), attempts to construct such systems occurred at the birth of AI NLP. Another impetus for MT springs from the potential of such systems to begin to solve the difficult question of whether some text has been adequately understood. For example, in English-to-Russian and a Russian-to-English translation systems, a given utterance in English can be translated into Russian and then translated back to English. In which case, the question of whether the two systems extract accurate meanings can be simplified (without major distortion given the current state of the art) to one of meaning preservation: Does the initial English sentence say the same thing as the final English sentence?

MT systems have been in use for years but only in niche applications either in simple and restricted applications or where a first crude translation is valuable. The former case might be technical manuals that are written according to simple, well-defined structures using a small, well-defined vocabulary. The latter situation might arise as a precursor to human refinement of the documents.

6. Machine-Readable Dictionaries

One spin-off from MT work has been the computerization of dictionaries. Words with their various meanings can be stored in a computer for fast access by any other computer system such as a word processor or by an inquiring human. This is little more than a fast form of the traditional book dictionary, but appropriate computerization can offer much more. A computerized dictionary can be proactive in suggesting words and even phrases, which implies an AI aspect to certain operations.

E. Robotics

The idea of a robot embodies the ultimate goal of AI, and robots have been designed and built since AI's inception.

1. Moving, Sensing, and Reasoning

One feature of the early years of AI was that researchers did not appreciate the depth and difficulty of the problems they faced. Many early robotics systems were mobile machines with a variety of sensors (vision, infrared, tactile, etc.). They were also supposed to analyze their situation and solve problems (such as how to traverse realistically cluttered rooms and corridors). They were conceived as crude simulations of complete humans with the hope that refinements and additions would improve the similarity year by year. This did not happen; instead, each component problem escalated into a full-blown subarea of AI.

2. Hand–Eye Systems

One common simplification of the robotics problem has been to construct hand–eye robots. These systems attempt to integrate visual processing and manipulation of the seen objects. This difficult task requires pattern recognition and some degree of image

understanding to be integrated with fine motor control and tactile feedback.

Industrial companies such as automobile manufacturers employ a dazzling array of robotics systems but these include little or no AI. The elaborate welding systems are preset to weld at specific points in three-dimensional space. The problem of fast, accurate welding is then reduced to one of engineering the means by which the parts to be welded are moved to precisely the correct placement.

3. Cognitive Systems

Some of the most imaginative and ambitious work on AI robots is being done at one of the long-term bastions of AI. At the Massachusetts Institute of Technology, research teams are developing various systems to implement cognitive robotic functions (e.g., an attentional system for social robots).

IV. PHILOSOPHICAL PERSPECTIVES

The field of AI has always spurred extensive philosophical controversy. Can human intelligence be no more than the right algorithm running on a biological computer, sometimes called wetware as distinct from the hardware of modern computers? There are various arguments that human intelligence must be more than this. Is the spiritual or emotional dimension of humanness an essential ingredient of intelligence, or are such aspects merely side issues that humanity happens to exhibit?

Whatever the outcome, some assert that an intelligent machine can never really feel like we do, enjoy the taste of strawberries or the thrills of love, so it must always fall short of full human intelligence. At best, machine will only simulate such feelings, and this is a far cry from actually experiencing them as a human does.

A distinction is sometimes drawn between strong AI and weak AI. The former is a machine intelligence that encompasses all aspects of human intelligence (whatever they may be), whereas the latter category only aims to reproduce limited aspects of intelligence.

Further controversy centers on the notion of intelligence as an abstract concept within which human intelligence is just one (the only one we currently know) particular manifestation. In this view, arguments for the nonoptimality of human intelligence emanate from several sources: Empirical evidence of, for example, imperfect memory suggests that certain specific improvements should be possible, and the realization that the basic ''machinery'' of the brain was not designed to support intelligent activity but evolved for other purposes is also suggestive of nonoptimality. Given a custom-designed machine executing a similarly crafted suite of computational procedures, an AI that supercedes the human version in terms of speed, accuracy, etc. should be possible.

In opposition to this viewpoint, evolution may have produced a set of compromises that is optimal for intelligence on Earth. It is plausible that imperfect memory, for example, is an unavoidable negative outcome from a system that would be seriously disadvantaged by the clutter of useless memories if it forgot nothing. Similarly, feelings and emotions might be vital signals of what courses of action to pursue, and what ones to avoid, in a world full of choices that must be made on incomplete evidence.

What claim can there be that a machine is intelligent just because it behaves as if it is? The famous Turing Test (devised in 1950 by the mathematician Alan Turing) rests on the idea that if a machine behaves as if it is intelligent then we must concede that AI has been achieved. This opens the more general issue of whether and to what extent form can be inferred from function: By observing what a system does, can we determine anything about how it is doing it? If we ignore the physical components of the brain (as most AI work does), then what can we expect to deduce about the detailed mechanisms that constitute intelligence based only on observations of intelligent behavior?

Finally, there are challenges to the accepted wisdom that intelligence is, or can be exhibited by, **any** algorithmic procedure running on **any** modern digital computer. There may be something about the determinism and definiteness of this combination that is totally at odds with the seemingly unconstrained openness of human intelligence. Objections of this nature tend to be founded on either the known limitations of formal systems (such as abstract algorithms) or the apparent difficulties posed by certain aspects of human intelligence (e.g., free will).

V. CONCLUSIONS

In the field of AI, there is a range of often autonomous subfields. This occurred as a reaction to the extreme difficulty revealed by all approaches to the more global goal of complete AI. It is a classic divide-and-conquer approach to excessive complexity, but can it work for AI?

On the negative side, note that investigations within most (if not all) subfields seem to reveal that knowledge of the world and sophisticated contextual awareness are essential for high-quality AI. In other words, real progress in each subfield seems to require major progress in every other as a corequisite, if not a prerequisite.

On the positive side, there are some clear indications that intelligence is not all-or-nothing. There are many examples of intelligent humans who lack sight, speech, and hearing. In addition, there is a case to be made for the modularity of mind both from localization of functional components of intelligence in specific anatomical regions and from observed phenomena such as optical illusions. Many optical illusions remain compelling illusions even when the basis for the illusions is physically demonstrated or explained. This persistence might be interpreted as an absence of close integration between visual processing and reasoning abilities.

One pervasive conundrum is that a broad and deep knowledge of the world appears repeatedly in the subfields of AI as a necessity for intelligent activity, with an associated implication that more knowledge supports higher levels of intelligence. However, from a computational perspective more information (in a larger knowledge base) must lead to more searching for relevant pieces, which takes more time. Stated succinctly, the more you know, the slower you go. This is true for AI systems but not, it seems, for humans. In fact, the growth of a human's expertise is often associated with a faster performance rather than a slowdown. Why? For some, the answer is that classical programming on conventional computers is the wrong basis for intelligence. Therefore, other computational paradigms are being explored. For others, the answer is less radical but no less obscure: Information cannot just be added on, it must be properly integrated and indexed in sophisticated ways.

A final general point of difficulty is that intelligence in humans is characterized by tentative conclusions and sophisticated recovery from errors. We do not always get things right, but we can usually readjust everything appropriately once we accept the error. In AI systems, the compounding of tentative conclusions is difficult to manage and recovery from error is often no easier because we typically have to foresee each source of error and supply a recovery procedure that will always work.

One general conclusion to be drawn from this survey is that AI has not been as successful as is popularly believed. One rejoinder to this common charge is that as soon as an AI problem has been solved by the construction of a programmed system, that problem ceases to be called AI. There is some truth in this, but not much. The real AI problems of adaptivity, natural language understanding, image understanding, and robotics are still out there and waiting to be solved.

Work in AI does not undermine a belief in humanity by aspiring to mechanize it. Quite the contrary—the more one examines the problems of intelligence, the more one's admiration grows for a system that seems to solve all of them at once, often effortlessly.

See Also the Following Articles

CATEGORIZATION • CONSCIOUSNESS • CREATIVITY • EVOLUTION OF THE BRAIN • HEURISTICS • INFORMATION PROCESSING • INTELLIGENCE • LANGUAGE ACQUISITION • LOGIC AND REASONING • MEMORY, OVERVIEW • NEURAL NETWORKS • NUMBER PROCESSING AND ARITHMETIC • PATTERN RECOGNITION • PROBLEM SOLVING • SPEECH

Suggested Reading

Brookes, R. A. (1999). *Cambrian Intelligence: The Early History of the New AI*. Bradford/MIT Press, Cambridge, MA.

Dennett, D. C. (1998). *The Dennett Quartet*. Bradford/MIT Press, Cambridge, MA.

Dreyfus, H. L., and Dreyfus, S. E. (1986). *Mind over Machine*. Macmillan, New York.

Gazzaniga, M. S. (Ed.) (2000). *The New Cognitive Neurosciences*. Bradford/MIT Press, Cambridge, MA.

Heinke, D., Humphreys, G. W., and Olson, A. (Eds.) (1999). *Connectionist Models in Cognitive Neuroscience*. Springer-Verlag, London.

Hinton, G., Dayan, P., Frey, B., and Neal, R. (1995). The wake–sleep algorithm for self-organizing neural networks. *Science* **268,** 1158–1161.

Hutchins, W., and Somers, H. (1992). *Introduction to Machine Translation*. Academic Press, San Diego.

Kistler, W. M., and van Hemmen, J. L. (2000). Modeling synaptic plasticity in conjunction with the timing of pre- and postsynaptic action potentials. *Neural Comp.* **12,** 385–405.

Luger, G. F., and Stubblefield, W. A. (1998). *Artificial Intelligence*. Addison-Wesley, Reading, MA.

Parks, R. W. (Ed.) (1998). *Fundamentals of Neural Network Modeling*. Bradford/MIT Press, Cambridge, MA.

Partridge, D. (1991). *A New Guide to Artificial Intelligence*. Ablex, Norwood, NJ.

Pazienza, M.-T. (Ed.) (1997). *Information Extraction*. Springer, New York.

Rich, E., and Knight, K. (1991). *Artificial Intelligence*. McGraw-Hill, New York.

Searle, J. R. (1992). *The Rediscovery of the Mind*. Bradford/MIT Press, Cambridge, MA.

Wilks, Y., Guthrie, L., and Slator, B. (1996). *Electric words: Dictionaries, Computers and Dictionaries*. MIT Press, Cambridge, MA.

Astrocytes

NICOLE BAUMANN*,‡ and DANIELLE PHAM-DINH†,‡

†*INSERM Unit 495,* †*INSERM Unit 546, and* ‡*University Pierre and Marie Curie, Paris, France*

GLOSSARY

Alexander's disease Alexander's disease is a leukodystrophy in which there is accumulation within the astrocyte cytoplasmic processes of beaded inclusions, i.e., Rosenthal fibers which contain alpha-B crystallin. The relationship between these abnormal astrocytes and demyelination characteristic of this pathology is not yet clear.

blood–brain barrier The blood–brain barrier is a physical and physiological barrier impeding the passive diffusion of solutes from the blood into the extra-cellular space of the central nervous system. It is also a brain–blood barrier. The highly impermeable tight junctions between endothelial cells forming the capillaries in the central nervous system are responsible for the blood–brain barrier functions. Astrocyte perivascular processes or end-feet form a virtually continuous sheath around the vascular walls. Astrocytes are involved in the induction of the blood–brain barrier property of central nervous system endothelial cells.

gap-junctions Gap-junctions are a group of diverse channels that vary in their permeability, voltage-sensitivity and potential for modulation by intra-cellular factors. They provide a pathway for the selective exchange of small molecules. Astrocytes are connected to each other by gap-junctions which are localized between cell bodies, between processes and cell bodies and between astrocyte end-feet that surround blood vessels. Thus activities of neighboring cells can be synchronized.

glial cells Glial cells are present in the central nervous system and distinct from neurons. They include astrocytes, oligodendrocytes and microglia. Interactions between neurons and glial cells are important during neuronal development and for the normal functioning of the nervous system. Dysfunction of glial cells is involved in many degenerative diseases of the nervous system.

glioma Glioma are the most frequent malignant tumors of the brain. Most of them are of astrocytic origin.

gliosis Astrocytes have a capacity to react mainly in relation to an injury in the central nervous system and constitute reactive gliosis. This reaction is characterized by hypertrophy of astrocytes. There is also an increase in astrocyte intermediary filament numbers and in their constituent the glial fibrillary acidic protein GFAP. There may also be a proliferation of astrocytes, occurring in general close to an acute lesion, but it is not constant. Astrogliosis is observed as a secondary process during aging as well as in many pathological conditions such as Alzheimer disease, brain trauma, ischemia, and multiple sclerosis in demyelinated areas. The physiological role of astrogliosis remains controversial with respect to the beneficial or detrimental influence of reactive astrocytes on central nervous system recovery.

radial glia Radial glia are a particular category of the astrocyte family. Radial glia fan out from the ventricular and subventricular zones where the cell bodies reside, and they extend to the pial surface. During development, radial glia are necessary to neuronal migration to the cortex. After migration has begun to subside, these cells assume a variety of transitional forms and transform into star-shaped astrocytes.

The astrocyte is a major glial cell type of the central nervous system. Astrocytes are star-shaped cells which extend

processes terminating in end-feet at the pial surface or on blood vessels. Other cells with a different morphology have key-constitutents of astrocytes, such as glycogen particles and a protein of astrocyte intermediary filaments, the glial fibrillary acid protein (GFAP); thus they are part of the astrocyte family. Among them are embryonic glial fibers (radial glia) necessary for neuronal migration. Development of neuronal pathways and synaptic functions require interactions with astrocytes. Astrocytes help to maintain synaptic functions by buffering ion concentration, clearing released neurotransmitters, and providing metabolic substrates. Their dysfunction is implicated in neurodegenerative diseases. The major brain tumors (glioma) appear to be of astrocytic origin.

I. INTRODUCTION

Virchow (1846) was the first to find cells other than neurons in the brain. He thought that it was the connective tissue of the brain, which he called "nervenkitt" (nerve glue) (i.e., neuroglia). The name survived, although the original concept radically changed. Two main types of glial cells were originally described in 1893 by Andriezen: protoplasmic in gray matter and fibrous in white matter. They were both later proven to be astrocytes and their first clear description was provided by Ramon y Cajal in 1913; he showed that they were independent from neurons and capillaries. In 1928, using staining techniques based on silver carbonate impregnation, as introduced by Golgi, Rio Hortega found two other cell types, the oligodendrocyte (first called interfascicular glia) and another cell type that he distinguished from the two macroglial cells (i.e., macroglia) and that he called microglia.

Astrocytes with oligodendrocytes and microglia constitute the glial cells (Fig. 1) in the central nervous system (CNS). The importance of glial cells is suggested by their increasing number in relation to the evolution of species, from 25% in the *Drosophila* brain to 65% in murine species and 90% in man. Astrocytes

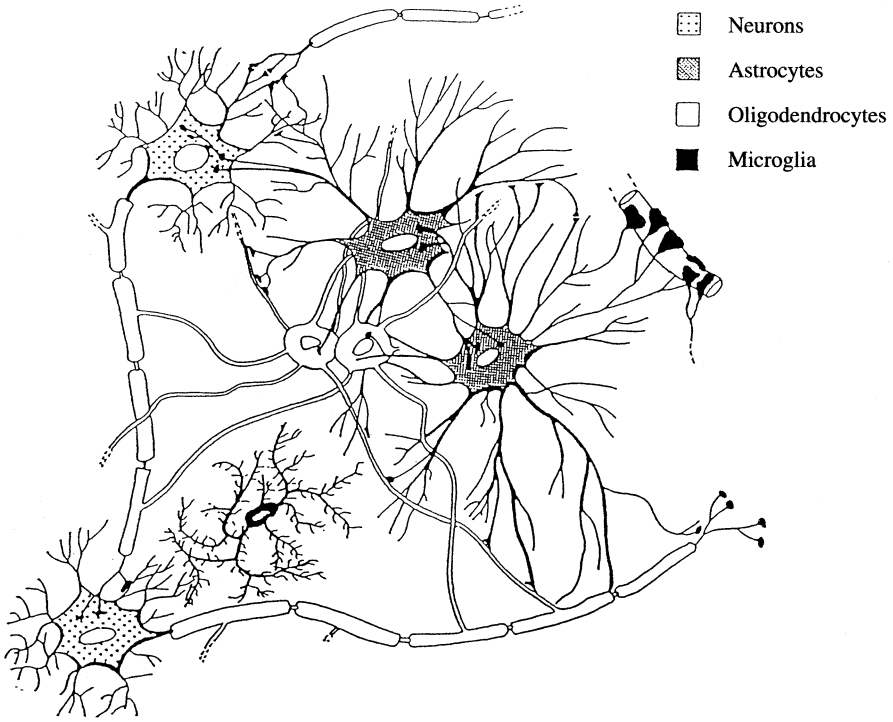


Figure 1 Schematic representation of the different types of glial cells in the CNS and their interactions among themselves and with neurons. Astrocytes are stellate cells with numerous processes contacting several cell types in the CNS: soma, dendrites, and axons of neurons and soma and processes of oligodendrocytes and other astrocytes. Astrocytic feet also ensheath endothelial cells around blood capillaries, forming the blood–brain barrier, and terminate to the pial surface of the brain, forming the glia limitans. Oligodendrocytes are the myelinating cells of the CNS. Many interactions between glial cells, particularly between astrocytes in the mature CNS, are regulated by gap junctions, forming a glial network (adapted from C. Giaume and L. Venance, 1995, *Perspect. Dev. Neurobiol.* **2,** 335–345).

present a wide range of functions: They participate to brain construction and energy metabolism, and they also participate in the formation and maintenance of the blood–brain barrier. Their organization and functions are modulated by circadian rhythms, parturition, lactation, and osmotic stimulation. They respond to many changes in the cellular and extracellular environment. Astrocytes are also directly or indirectly involved in pathological states.

The diverse roles of these cells are still being unraveled and may be more complex than previously thought. Indeed, recently it was shown that astrocytes possess a calcium-dependent form of excitability, and that they are involved in the functional modulation of synaptic activity. The existence of a coordinated and continuous bidirectional neuron–astrocyte signaling provides a new view of the physiology of these glial cells.

II. THE ASTROCYTE FAMILY

Astrocytes are a family of cell types that share certain morphological and functional similarities as well as biochemical and immunological features. Astrocytes are star-shaped cells, the processes of which occupy 25% of the volume of the CNS and extend into the surrounding neuropil (network of intermingled and interconnected processes of neurons). They have an electron-lucent cytoplasm and several distinguishable morphological characteristics. First, as shown by electron microscopy, they contain intermediate filaments of 10 nm (gliofilaments) that were referred to as fibrils in the gold-sublimate preparations. Intermediate filaments are more abundant in the fibrous astrocyte population of the white matter region; they occur in bundles and contain mainly a specific protein, the glial fibrillary acidic protein (GFAP). Second, they contain numerous glycogen particles. Third, their processes terminate on blood vessels (arterioles, veinules, and capillaries) as perivascular end feet, participating in the blood–brain barrier formation, and/or extend expansions toward the surface of the brain where they constitute a glial limiting membrane (the glia limitans). By filling cells intracellulary with horseradish peroxidase or lucifer yellow, it was shown that astrocytes may have 50–60 branching processes that terminate in end feet at the pial surface or on blood vessels. Some of these processes extend to considerable distances (i.e., 300–400 μm). The subpial astrocyte processes present a well-organized palisading pattern in the adult CNS. A basal lamina separates the plasma membrane of the end feet from the blood vessels and also the glia limitans from the pial elements. Fourth, astrocytic processes are connected by gap junctions, mainly at the level of perivascular end feet, or the glia limitans. Finally orthogonal arrays, as observed by freeze-fracture microscopy, are present on the surface of perivascular end feet or subpial astrocytes of the glia limitans; because they are so characteristic of astroglial cell membranes, they are used as a marker in freeze-fracture replicas. Orthogonal arrays may represent sites at which some material is transported across membranes at the interface between the blood and the cerebrospinal fluid compartments.

In addition to the two main classes of astrocytes, protoplasmic in gray matter, and fibrous in white matter, there are also astrocyte subtypes in the olfactory bulb. Embryonic radial cells (radial glia) and cells that may be considered morphologically and biochemically as adult divergent forms of embryonic radial glia (i.e., Müller cells of the retina, Golgi–Bergmann cells of the cerebellum, pituicytes, pineal astrocytes, ependymal cells, and tanycytes) are considered to belong to the astrocyte family. Ependymal cells and tanycytes are considered by others as a separate group of glial cells although certainly closely related to astrocytes.

- Olfactory bulb ensheathing cells are specialized ensheathing cells. They are found in the pathway known as the rostral migratory stream (RMS), in which they accompany the olfactory neurons in their migration from the olfactory epithelium to the olfactory bulb. They remain in the adult since neurons keep multiplying and differentiating in this area of adult neurogenesis. They share both Schwann cells and astrocytic characteristics.
- Embryonic radial glial cells (radial glia) fan out from the ventricular and subventricular zones, where their cell bodies reside, to the pial surface, where they terminate with conical end feet.
- Golgi epithelial cells and their Bergmann glial fibers are generated from radial glia in the molecular layer of the developing cerebellum, where they form characteristic palisades. They may therefore share with radial glia phenotypic traits that allow them to guide the external granular cell migration in the cerebellum. However, in the adult, the Bergmann glia have lost their radial glia characteristics.
- Müller cells are bipolar radial glial cells spanning the entire depth of the retina. They comprise the principal glial cells of the retina in most vertebrate

species. Microvilli from their apical process project into the subretinal space surrounding the photoreceptors. Secondary processes from their main trunk form extensive sheaths that surround neuronal cell bodies, dendrites, and, in the optic fiber layer, the axons of ganglion cells.

- Ependymal cells are located along the internal cavities of the CNS (cerebral ventricles and ependyma). They possess microvilli and ciliated processes that contribute to the movement of the cerebrospinal fluid. In some areas of the ventricular system, ependymal cells give rise to the choroid plexus, which secretes cerebrospinal fluid. Tanycytes are a special category of ependymal cells with numerous microvilli; they are devoid of cilia at their apical surfaces. They are associated with the circumventricular organs (i.e., the choroid plexus, the subcommissural organ, the subfornical organ, neurohypophysis, median eminence, pineal gland, and area postrema).

In fact, there are many astrocyte subpopulations, not only in regard to their shape and ultrastructural characteristics but also with respect to their biochemical properties, such as the receptors they express and the enzymes they carry. There are also morphological variations of astrocytes according to certain physiological states as well as pathological changes (e.g., gliosis, which occurs in gray and white matter regions), mainly in reaction to certain pathological states.

III. CHARACTERISTIC COMPONENTS OF ASTROCYTES

All the biochemical and molecular characteristics of astrocytes will not be described here. We limit discussion to those that have given specific insights toward our understanding of this cell type.

A. Glial Fibrillary Acidic Protein

GFAP is an intermediate filament protein expressed in astroglia. It was first isolated from multiple sclerosis brain, in which astrocytes are very reactive. Intermediate filaments constitute major components of the cytoskeleton of eukaryotic cells as 10-nm filaments. On the basis of differences in size and amino acid sequences, GFAP together with vimentin and desmin are classified as type III intermediate filament proteins. They are composed of three distinct regions: an amino-terminal head region, a central rod region, and a carboxy-terminal tail region. The rod regions are highly conserved among intermediate filament proteins, whereas size and amino acid sequences vary in the other regions. GFAP shows some species-specific amino acid heterogeneity and its molecular weight ranges from 48 to 51 kDa. Despite these minor differences, it is a highly conserved molecule. Soluble forms of the protein are assembled to form dimers; two dimers form protofilaments, which assemble to form the 10-nm filaments. Phosphorylation and dephosphorylation of specific amino acid residues in the head region are involved in the regulation of GFAP assembly. GFAP is observed in all cells of the astrocyte family, including ependymocytes, and is very abundant in tanycytes. Although GFAP has been found in the peripheral nervous system (e.g., in nonmyelinating Schwann cells) as well as in non-CNS tissues, GFAP immunocytochemistry is a main tool to study normal development of astrocytes and their reaction following injury and disease where it is upregulated.

B. Other Specific Components of Astrocytes and Cells of the Astrocyte Lineage

Nestin is a protein the expression of which specifically distinguishes neuroepithelial stem cells (from which the name nestin originated) from other more differentiated cells in the neural tube. Nestin defines a distinct sixth class of intermediate filament protein, closely related to neurofilaments. Nestin is expressed by glial precursors such as radial glia and in the cerebellum by immature Bergmann fibers; it is also expressed by adult Bergmann fibers recapitulating developmental stages when placed in the presence of embryonic neurons. During neuro- and gliogenesis, nestin is replaced by cell type-specific intermediate filaments (i.e., neurofilaments and GFAP, respectively). Nestin expression is reinduced in reactive astrocytes in the wounded adult brain. In adult subventricular zone (SVZ) of adult mice, ependymocytes strongly express nestin and vimentin and lightly express GFAP.

Vimentin, a type III intermediate filament subunit present in most tissues of many vertebrates, is also found in brain. In the mouse, it is expressed in oligodendrocyte precursors, radial glia, and astrocytes before the onset of GFAP expression. Both *in situ* and in culture, vimentin and GFAP may coexist in certain astrocyte subclasses, such as tanycytes and ependymocytes. In vimentin knockout mice, the GFAP

network is disrupted in those astrocytes that normally coexpress vimentin and GFAP (i.e., Bergmann glia of cerebellum and astrocyte subpopulations of the corpus callosum).

Glycogen particles are present in radial glia and mature astrocytes, and they represent the storage form of glucose. Glutamine synthetase (GS) catalyzes the conversion of glutamate to glutamine in the presence of ATP and ammonia. By immunocytochemical methods, its presence has been demonstrated in the cytoplasm of astrocytes. Although some authors have found GS immunoreactivity in a discrete population of oligodendroglia, it remains for most authors an astrocyte marker. Its important function in neuron physiology will be described later.

S100 proteins are Ca^{2+}- and Zn^{2+}-binding proteins found at high concentration in the mammalian brain. It is now known that the S100 family of calcium-binding proteins contains approximately 16 members, each of which exhibits a unique pattern of tissue/cell type-specific expression. There are two 10-kDa isoforms of S100, alpha and beta. In the brain, the beta isoform is most abundant and is synthesized by and released from astrocytes. The S100 beta molecule is considered a calcium-binding protein of astroglia.

IV. ASTROCYTE CELL LINEAGE

Mammalian species acquire their full complement of cortical neurons during the first half of gestation. Radial glia is established early in development at the time of neurogenesis, during the early gestational period, and accompanies neuroblasts during their migration to the cortex. In late gestational and early postnatal mammalian brain, the SVZ is the major source of astrocytes and oligodendrocytes, although astrocytes may also develop from radial glia. Remarkable recent data suggest that cells of the astrocyte lineage could have neural stem cell potential.

A. Embryonic Radial Glia

The radial glia was first revealed by the Golgi silver impregnation at the turn of the twentieth century and later described by Rakic in the early 1970s. They have a highly organized palisading cytoarchitecture, in relation to their role in neuronal migration from the ventricular zone to the cortical layers. In the developing CNS, cell bodies of the radial glia lose their ventricular insertion and migrate toward the brain surface. In the monkey, these glial cells appear during the first third of gestation, first in the spinal cord and then in the diencephalon, the telencephalon, and the cerebellum. During the peak of neurogenesis, there are two distinct classes of proliferative cells: those expressing GFAP that will give rise to radial glia and those not expressing this marker (i.e., neuroblasts). These $GFAP^+$ and $GFAP^-$ cells are intermixed in the ventricular and subventricular zones, which illustrates that glial and neuronal cell lines coexist within the early fetal proliferative zones and that the onset of glial phenotypic expression occurs prior to the last division of neuroblasts. Expression of embryonic radial glial identity requires the presence of extrinsic soluble signals is in embryonic forebrain, including neuronal factors.

By using a Golgi technique, Schmechel and Rakic showed in the monkey brain that radial glia transforms into astrocytes. Radial glial cells assume a variety of transitional forms during the process of this transformation into mature astrocytes. This transformation occurs in different areas of the CNS at a specific embryonic age and is initiated after neuronal migration has begun to subside. The bipolar radial cell becomes transformed into the astrocytic multipolar form; this may occur directly or with a monopolar radial form as an intermediate state of transformation (Fig. 2). The number of astroglial cells increases at an accelerated pace after neurogenesis is complete.

A recent discovery suggests that in the developing CNS, radial glial cells may have the potential to self-renew and to generate both neurons and astrocytes; these data led to the concept that a lineage relationship between these two cell types underlies the radial organization of the neocortex.

B. Astrocytes Derived from SVZ Precursors

In the mammalian cortex, both neurons and glia arise from the proliferating neuroepithelial cells of the telencephalic ventricular and subventricular zone. The SVZ is a mosaic of multipotential and lineage-restricted precursors in which environmental cues influence both fate choice and all surviving cells. In the adult brain, SVZ astrocytes could be neural stem cells.

The generation of astrocytes directly from SVZ cells was suggested by morphological observations

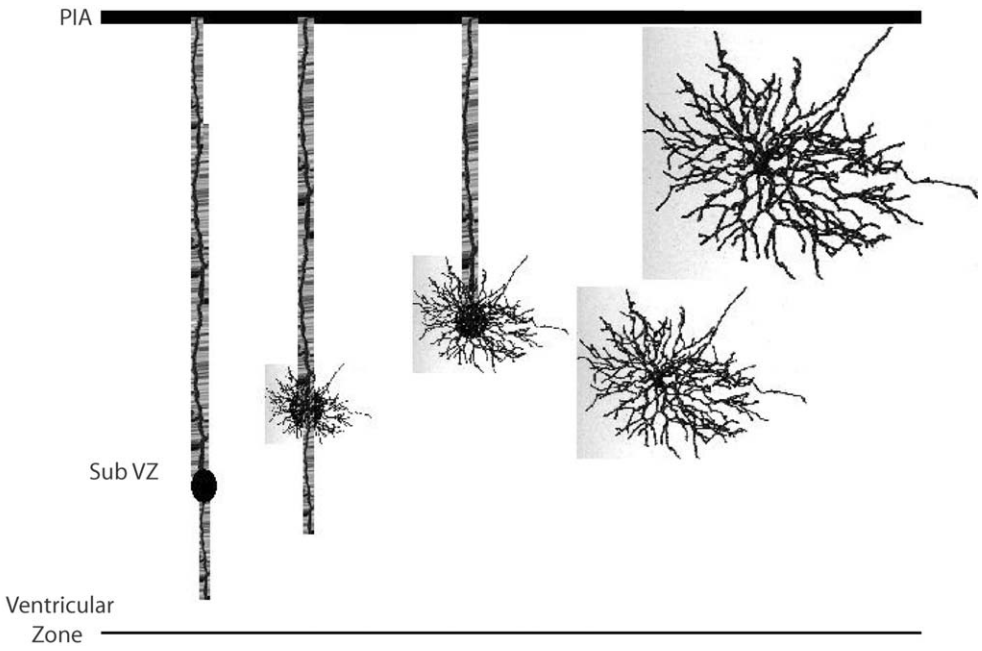


Figure 2 Morphogenetic transformation of fetal radial glial cells into various astrocytic forms [adapted from P. Rakic, 1995, in *Neuroglia* (H. Hettenmann and B. Ransom, Eds.), pp. 746–762, Oxford Univ. Press, New York].

providing evidence for separate astrocyte and oligodendrocyte lineages while these cells are still proliferating. Moreover, because angiogenesis and astrogenesis occur over a similar period of time during neurogenesis, it has been suggested that the growth of blood vessels may even drive the selection of astrocytic fate by a progenitor cell; at the time at which vessel growth is completed, fewer progenitors may be directed toward an astrocytic fate, allowing more to differentiate into oligodendrocytes.

C. Type 2 Astrocyte Phenotype

Since the early 1980s, glial research has centered around an oligodendrocyte type 2 astrocyte progenitor (0-2A) first discovered in culture of the optic nerve. 0-2A progenitors in culture generate oligodendrocytes constitutively but can give rise to process-bearing astrocytes (so-called type 2 astrocytes) when treated with 10% fetal calf serum. Attempts to find cells *in vivo* with the type 2 astrocyte phenotype during normal development have failed. Most astrocytes are generated prenatally and during the first postnatal week, before oligodendrocyte formation, suggesting that there is not a second wave of type 2 astrocytes *in vivo* as suggested by previous *in vitro* studies. Franklin and Blakemore noted that the apparent discrepancies between the data obtained *in vitro* and *in vivo* highlights an important aspect of the approach *in vitro*. When a progenitor cell is grown in tissue culture, the environments to which it can be exposed are restricted only by the imagination of the experimenter. By exposing the cell to a variety of signals, the differentiation potential of the cell can be examined. In contrast, during development, a progenitor cell is exposed to a restricted sequence of signals that are spatially and temporally programmed.

D. Influence of Growth Factors and Hormones

As shown for radial glial cells, neuronal factors influence the development of astrocytes. In tissue culture systems, primary astrocytes respond to the presence of neurons by withdrawing from the cell cycle and extending complex $GFAP^+$ processes. In contrast, in the absence of neurons, astrocytic cells have a flat fibroblastic appearance with little or no processes formation.

Many trophic factors act on cells of the astrocyte lineage *in vivo* such as bone morphogenetic proteins

(BMP), ciliary neurotrophic factor (CNTF), fibroblast growth factor (FGF), epidermal growth factor (EGF), platelet-derived growth factor (PDGF), and leukemia inhibitory factor (LIF). Most of these have been shown to be involved in proliferation and differentiation of cells of the astrocytic lineage and provided by neighboring neurons, with PDGF appearing to be the link in the sequence of cell–cell interactions responsible for matching numbers of neurons, astrocytes, and blood vessels during development.

The thyroid hormone T3 regulates the number of astrocytes and the maturation of Bergman glial cells. Neuroactive steroids also regulate astroglia morphology, at least in hippocampal cultures from adult rats.

V. PARTICIPATION OF ASTROCYTES IN BRAIN CONSTRUCTION

A. Neurons

1. Neuronal Migration and Neurite Formation

In 1911, Ramon y Cajal perceived the functional importance of radial glia as directional guides for migrating immature nerve cells. The concept of glial guidance first described by Rakic in 1971, and derived from the observation of a tight association between migrating granule cells and Bergmann fibers in the developing molecular layer of the cerebellum, underlies a commonly used mechanism for the positioning of young postmitotic neurons in developing brain. Neuronal migration is a key step in neural morphogenesis since inadequately located neurons may not establish the appropriate connections, and this may lead to neuronal death or to functional deficits of synaptic circuits. The close structural relationship between radial glia and migrating neurons consist in a radial unit. Radial glial fibers are present in large numbers during peak periods of neuronal cell migration in each structure of the developing primate nervous system. During the entire period of cell migration to the neocortex, a single radial process can guide several hundreds of neurons that originated from the same position in the ventricular zone, the so-called proliferative unit. Although the migrating neuron encounters myriad processes in various orientations, it remains constantly apposed to radially oriented fibers. Thus, there is a vertical columnar organization of the brain. From counts on samples of ventricular zones, it has been determined that the total number of ontogenetic columns is several millions. The remarkable expansion of the cortical surface during evolution can be explained by an increase in the number of proliferative units, as a single round of cell division doubles the number of ontogenetic columns and therefore the number of cells in the cortex. Within each column, neurons generated earlier occupy deeper positions; therefore, those arriving later have to pass them, along the glial radial fiber, to become situated more superficially. This constitutes an ''inside-out'' sequence of time of origin of neurons in the cerebral isocortex. There is a neuronal leading process along the glial guide, and the neuronal cell body is the site of adhesion of the migrating neuron.

Neuronal cell migration requires the cooperative interaction of adhesion and recognition molecules, which may be expressed by neurons and radial glial cells. Migratory granule neural cells in the cerebellum produce a cell adhesion molecule, astrotactin, that enables them to ride the glial scaffold. The blockage of astrotactin curtails neuronal migration *in vitro*. Although at this stage some of the numerous molecules involved in neuronal migration have been identified, the exact mechanisms involved in the selection of a pathway, migration, and departing from this pathway remain obscure.

In the adult mammalian forebrain, neuronal precursors born in the SVZ of the neonatal and adult rodent brain are able to migrate 3–8 mm from the walls of the lateral ventricle into the olfactory bulb, where they differentiate into olfactory interneurons. This migration depends on a persistent manifestation of the RMS that was first reported in the developing rodent brain. This tangentially oriented migration occurs without the guidance of radial glial or axonal processes. The cells are closely associated, forming elongated aggregates called ''migration chains'' that are ensheathed and moved within channels formed by the processes of specialized astrocytes. However, the role of these astrocytic tunnels is unclear. If the migration of neuroblasts by chain migration is a glial-independent movement, these glial cells may provide *in vivo* a permissive environment and directional cues for migration to the olfactory bulb, restricting the dispersal of neuroblasts outside the migratory stream or isolating the migrating cells from the surrounding parenchyma.

Neurons have the intrinsic property of generating two distinct sets of processes, one axon and multiple dendrites. Some of the mechanisms underlying polarized sorting of membrane and cytosolic proteins are similar in neurons and other polarized cells, such as epithelial cells. The axonal membrane of neurons

shares properties with the apical membrane domain of epithelial cells, whereas the soma and the dendritic arbor of neurons corresponds to the basolateral domain of epithelial cells. Glial cells can control neuronal shape. It was recently shown in culture that the dendritic branching of GABAergic neurons requires signaling from living astrocytes. Among several mechanisms, neurite outgrowth can be promoted by regulating the degradation of the extracellular matrix by protease inhibitors, such as the glia-derived nexin.

2. Astrocyte Boundaries and Neural Pathways

Throughout neurogenesis, the diverse neural cell populations have to establish their correct pathways to find their proper locations and establish correct connections with neural partners. Moreover, the formation of the adult architecture of neuronal networks is essentially based on directed neurite extension. Growth cones have to find their way to their target region by crossing permissive substrates and diffusible factors and avoiding repulsive ones. Glial cells have been proposed to provide some pathways for axon growth in mammalian brain. Therefore, interactions with molecules secreted in the extracellular matrix by glial cells have been implicated in the formation of boundaries separating anatomically defined regions in the CNS. Within the developing nervous system, such boundaries are present in numerous regions of the CNS, such as the diencephalic/telencephalic junction, the optic chiasm, the midline of the developing forebrain, and the cerebral commissures. Among the components of the extracellular matrix, many are synthesized during development by glial cells such as tenascin-C and -R, thrombospondin-1, and laminin as well as a variety of proteoglycans; these molecules can act as either promoters or inhibitors of neural cell migration and neurite outgrowth. When functional patterns have formed and appear to be stabilized, these boundaries are no longer detectable, but can be reexpressed in reaction to injury. Astroglial maturation is accompanied by a decrease in expression of molecules such as laminin, NCAM, L1, and heparan sulfate proteoglycan, which are known to promote axonal outgrowth. In parallel, the maturation of astrocytes is accompanied by an increase in the synthesis of molecules known to inhibit neurite outgrowth, such as chondroitin sulfate proteoglycan and tenascin.

B. Brain Vessels and Blood–Brain Barrier

The blood–brain barrier is a physical and physiological barrier impeding the passive diffusion of solutes from the blood into the extracellular space of the CNS. The blood–brain barrier is also a brain–blood barrier. In mammals and most vertebrates, the blood–brain barrier is a complex cellular system consisting of endothelial cells, pericytes, perivascular microglia, astrocytes, and basal laminae. Interactions of all these cells seem necessary for the induction and maintenance of the specialized functions of the blood–brain barrier which protect the brain from toxin, viruses, and macromolecules. The highly impermeable tight junctions between endothelial cells forming the capillaries in the CNS of higher vertebrates are responsible for the blood–brain barrier functions. Astrocyte perivascular processes or end feet form a virtually continuous sheath around the vascular walls and are nearly as extensive as the vascular endothelial covering. Astrocytes end feet at the pial surface, which form the glia limitans, constitute a boundary around the entire CNS that helps to separate neural elements from mesodermal tissues.

During vascularization in early embryonic development, angioblasts (precurseurs of endothelial cells) invade the CNS neural tissues. When astrocytes differentiate, they extend processes on vessels and they induce blood–brain barrier differentiation in endothelial cells. Conversely, interactions of astrocyte progenitors with developing blood vessels may have a role in differentiation of astrocytes. In this context, it has been demonstrated in the retina that the vascular endothelial growth factor synthesized by astrocytes and Müller cells was an important inducer of endothelial cell growth and blood vessel development; therefore, astrocytes are critical in establishing the correct number of blood vessels in the brain. Recently, it has been shown that astrocytes isolated from GFAP knockout mice are unable to induce blood–brain barrier properties in endothelial cells.

VI. GLIAL NETWORK

Although many astrocytic processes ensheath most of the synapses within the CNS, other astrocytic processes are connected to each other by gap junctions. The main astrocytic gap junction protein is connexin 43, but others may be present. *In situ*, morphological studies have shown that astrocyte gap junctions are

localized between cell bodies, between processes and cell bodies, and between astrocytic end feet that surround blood vessels (Fig. 1). Gap junctions are a group of diverse channels that vary in their permeability, voltage sensitivities, and potential for modulation by intracellular factors. Thus, junctional coupling may provide a pathway for the selective exchange of small molecules (less than 1–1.4 kDa) such as ions, cyclic nucleotides, and small peptides. Gap junctions could potentially give rise to hierarchical signaling systems, in which cells of one class can transfer second messengers to cells of a second class but not vice versa. Although less frequently observed, gap junctions also occur between astrocytes and oligodendrocytes as well as between oligodendrocytes. Thus, the astrocytic syncytium extends to oligodendrocytes, allowing glial cells to form a "generalized glial syncytium." This large glial network extends radially across gray and white matter regions, from the spinal canal and brain ventricles to the glia limitans and the capillary epithelium.

Homologous coupling could synchronize the electrical activity of neighboring cells that serve the same functions, thereby generating a functional network. In this regard, intercellular coupling is believed to facilitate one of the major functions of astrocytes—maintaining the ionic balance of the extracellular fluid—by providing a vast cytoplasmic reservoir for the spatial buffering of ions taken up from the extracellular space. In particular, K^+ released from active neurons would enter neighboring glial cells and be redistributed by current flow through the glial syncytium. Heterocoupling between astrocytes and oligodendrocytes has been proposed to serve K^+ buffering around myelinated axons.

VII. CALCIUM SIGNALING IN THE ASTROCYTE NETWORK AND BETWEEN ASTROCYTES AND NEURONS

A seminal observation was made in the early 1990s by Cornell Bell and collaborators, who found that in the presence of glutamate, Ca^{2+} waves propagate *in vitro* at a velocity of about 15–20 μm/sec following complex routes of hundreds of micrometers through the astrocyte syncytium. Cornell Bell and collaborators proposed that networks of astrocytes may constitute a long-range signaling system within the brain. Since then, several critical steps involved in intercellular calcium signaling have been identified, and it is likely that gap junction communications play a key role in this process. Nevertheless, alternative mechanisms for Ca^{2+} wave propagation may involve extracellular messengers, such as ATP, released by stimulated cells. Propagation of calcium signals from astrocytes to neurons has recently been observed in astrocyte and neuron cocultures as well as *in situ* on acute rat brain slices. Elevation of cytosolic calcium in astrocytes induces glutamate release from neurons, which in turn signals back to neurons. Although different from neuronal excitability, astrocytes do indeed exhibit a form of excitability mediated by intracellular calcium spikes and oscillations as well as intercellular calcium waves that propagate through the glial syncytium in response to firing of glutamatergic neurons. The astrocytic Ca^{2+} excitability may be functionally analogous to the action potential that neurons use for communication, thereby enabling a bidirectional communicating network between neurons and astrocytes.

VIII. PARTICIPATION OF ASTROCYTES IN BRAIN HOMEOSTASIS AND NEURON–GLIA INTERACTIONS

A. Lipids

Apolipoproteins arise from local synthesis or filtration from plasma, although the respective amount remains unclear. According to current views, apolipoprotein E (apoE) is the main apolipoprotein produced and secreted by astrocytes within the brain parenchyma. ApoE is presumably involved in the redistribution of lipids among cells. Other apolipoproteins appear to be involved in brain lipid transport and homeostasis, such as ApoJ and ApoD. ApoD is found mainly in oligodendrocytes and astrocytes. Irregular lipoprotein metabolism, possibly related to special isoforms, may be involved in several neurodegenerative diseases, especially Alzheimer's disease.

B. Energy Metabolism: Glucose

In the brain, astrocytes are situated in a key position between microvessels, neurons, and oligodendrocytes (Fig. 1). If one considers the narrow extracellular space between brain cells, it appears that the substrates for the generation of energy must cross the astrocytes to reach their metabolic destination in neurons. Glucose

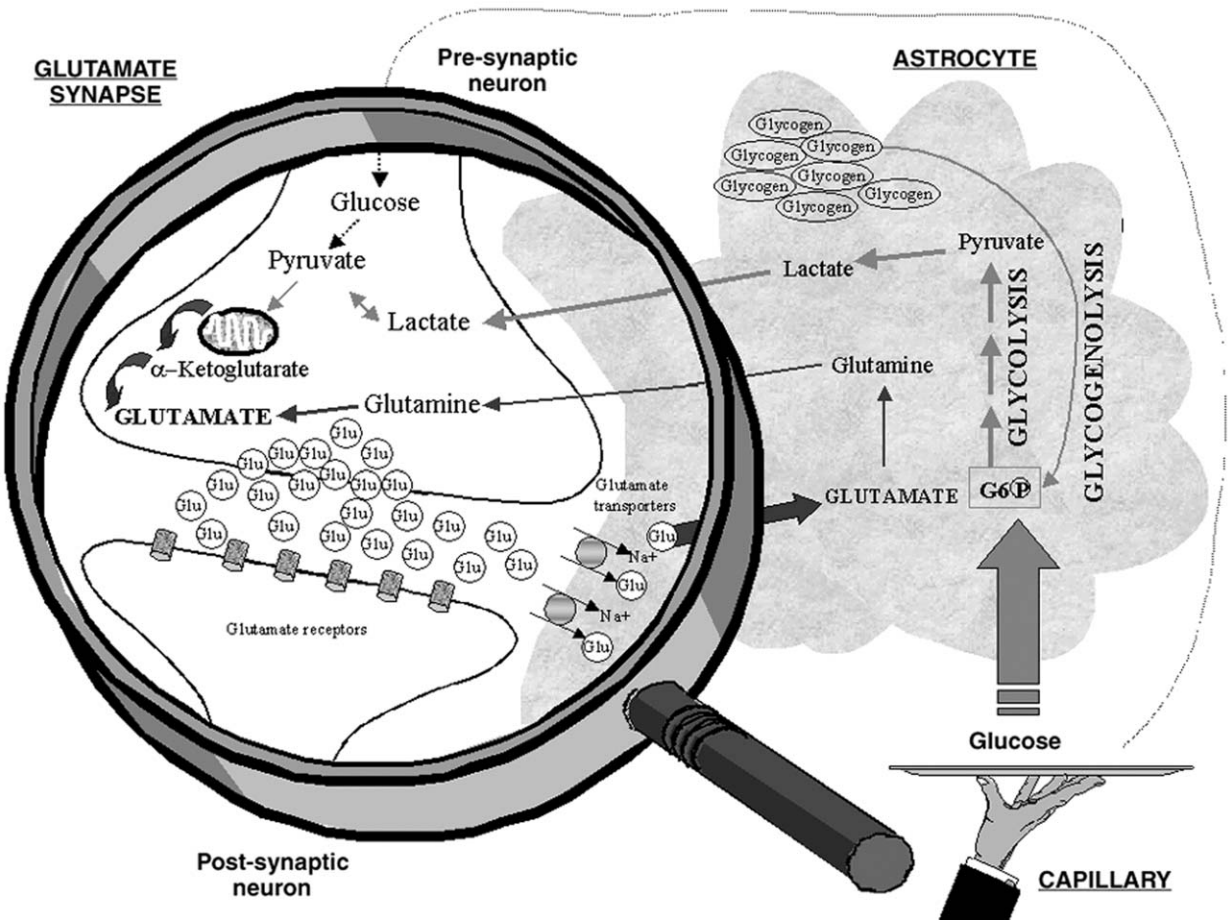


Figure 3 Metabolic coupling between neurons and astrocytes at glutamatergic synapses. Presynaptically released glutamate depolarizes postsynaptic neurons by acting at specific receptor subtypes. The action of glutamate is terminated by an efficient glutamate-uptake system located primarily in astrocytes. Glutamate is cotransported with Na^+, leading to activation of the astrocyte Na^+/K^+-ATPase, which in turn stimulates glycolysis (i.e., glucose utilization and lactate production). Lactate, once released by astrocytes, can be taken up by neurons and serve them as an adequate energy substrate. In accord with recent evidence, glutamate receptors are also shown on astrocytes. Glycogenolysis is also a source of lactate by glycogenolysis to glucose-6-phosphate (G6P), followed by steps of glycolysis. Direct glucose uptake into neurons under basal conditions can also occur. (adapted from Magistretti *et al*., 1999).

is the main energy source in the CNS; although neurons are able to take up glucose and phosphorylate it, at a basic level the tight coupling between the function and the energy metabolism of this cell type requires astrocytes (Fig. 3). First there is a transport of glucose into astrocytes by specific transporters; gap junction permeability also controls the uptake and distribution of glucose in astrocytes and in this way may regulate brain metabolism. Glycogen is also a source of energy in the brain, in which it is localized almost exclusively in astrocytes, to an extent that it can be considered a marker for this cell type. Its level is finely tuned by synaptic activity. Glycogenolysis is also activity dependent. Several neurotransmitters, such as noradrenaline, serotonin, histamine, the vasoactive intestinal peptide, and the purine adenosine, promote glycogenolysis as shown in studies of slices from different brain areas. The glucosyl residues of glycogen appear to also be broken down to lactic acid.

The discovery that glucose is taken up by glial cells was first demonstrated in the CNS model of the honeybee drone retina, in which an analog of glucose [2-deoxyglucose (2-DG)] was quantitatively shown to be taken up and phosphorylated by hexokinase in glial cells during light stimulation. The same coupling also occurs in the mammalian retina. It has been demonstrated in culture of astrocytes that glutamate stimulates 2-DG uptake and phosphorylation and that there

is a tight coupling between Na^+-dependent glutamate uptake and glucose utilization (Fig. 3). Noradrenaline and arachidonic acid also stimulate glucose uptake by astrocytes.

There is mounting evidence that lactate, resulting from the glycolytic processing of glucose in astrocytes, is the preferred energy source for neurons, particularly during situations of high energy demand. In astrocyte cultures, there is a glutamate-evoked lactate release that correlates with glutamate-evoked glucose utilization, indicating the role of glycolysis in this process. A similar process is observed in preparations of freshly isolated Müller cells still attached to photoreceptors, even in the presence of millimolar concentrations of glucose. When such a preparation is maintained in darkness to stimulate neurotransmitter release, lactate derived from glial glycolysis is transferred from Müller cells to photoreceptors, where it serves to fuel mitochondrial oxidative metabolism and possibly the resynthesis of the neurotransmitter pool of glutamate. There is a saturable lactate transporter in neurons. Pyruvate gives rise to lactate in astrocytes through a different lactate dehydrogenase enzyme than that in neurons; thus, there is evidence for an astrocyte–neuron lactate shuttle. In astrocytes, lactate could also be released from the metabolism of amino acids such as glutamate.

In human, glucose metabolism can be studied using (^{18}F) fluoro-2-deoxyglucose (FDG) by positron emission tomography (PET). FDG–PET imaging supports the notion that astrocytes markedly contribute to the FDG–PET signal; this view does not challenge the validity of the 2-DG-based techniques to monitor neuronal activation since the uptake of the tracer into astrocytes is triggered by a neuronal signal (i.e., the activity-dependent release of glutamate).

Although there is no proposal for other neurotransmitters, the well-studied model proposed for glutamatergic synapse can be extended to other neurotransmitters systems, such as the primary inhibitory transmitter γ-aminobutyric acid (GABA) that flows in a similar neurotransmitter cycle between neurons and astrocytes.

C. Neurotransmitter Transporters and Receptors

1. Glutamate Transporters

L-Glutamate is the major excitatory neurotransmitter in mammalian CNS. High-affinity glutamate transporters are believed to be essential both for terminating synaptic transmission and for keeping the extracellular glutamate concentration below neurotoxic levels. Many glutamate transporter subtypes are known and may display different functions in different neural subtypes. GLT-1 and GLAST transporter subtypes were shown to be selective markers of all astrocytic plasma membranes. GLAST transporter is expressed from radial glia through mature astrocytes during spinal cord development. Astrocytic membranes facing capillaries, pia, or dendrites express less glutamate transporters than those facing nerve terminals, axons, and spines, suggesting that the localizations of these glutamate transporters are tightly regulated.

The coexpression of GLAST and GS was specifically demonstrated in Müller cells, astrocytes, and retinal pigment epithelium, three cell types developmentally related in the retina. No neuronal or microglial staining was observed. Astrocytes are known to metabolize extracellular glutamate into glutamine, which constitutes the basis for the ''glutamate–glutamine'' cycle. In the brain, the major site of glutamine synthesis is astrocytes, in which GS is localized. Glutamine synthesis also requires glutamate, ammonia, and ATP. Glutamine released from astrocytes has very low affinity for glutamate receptors and therefore remains available for the synthesis of glutamate in neurons by the phosphate-activated glutaminase. Overall, once degraded in neurons, glutamine provides an important precursor source for amino acids, such as glutamate and GABA. Net synthesis of the neurotransmitter glutamate and GABA can take place from glutamine, α-ketoglutarate, or another tricarboxylic acid cycle intermediate together with an amino acid as the donor of the amino group. The latter may also be synthesized in astrocytes and taken up by neurons.

Glutamine synthesis is important for the detoxification of ammonia. During hepatic encephalopathy and other hyperammonemic states, the ability of the brain to fix excess blood levels of ammonia produced may be limited. Ammonia inhibits glutamine efflux from astrocytes, which probably contributes to astrocyte swelling and alteration of calcium homeostasis. The reduction of the extracellular space following astrocyte swelling may result in an increase in the extracellular concentration of ions, especially calcium that could affect neuronal excitability. Swollen astrocytes also release glutamate, which might also contribute to excitotoxicity. These data suggest that glial glutamate transporters provide the majority of functional glutamate transport and are essential for both maintaining

low extracellular glutamate and preventing chronic glutamate neurotoxicity.

2. Glutamate Receptors

Glutamate receptors were the first neurotransmitter receptors identified in astrocytes by studies showing that excitatory amino acids (glutamate, aspartate, and kainate) directly induce depolarization of astrocytes in culture. Glutamate induces complex changes in $[Ca^{2+}]_i$ characterized by intracellular waves and oscillations, mediated by different ionotropic (AMPA/kainate and NMDA) and metabotropic glutamate receptors (mGluRs).

3. Glycine Receptors

Glycine receptors are also regionally restricted in subclasses of neurons, and their adjacent astrocytes in the spinal cord, suggesting interplay between these two cell types.

4. GABA Transporters

GABA transporters play a key role in the termination of GABA transmission and in the regulation of extracellular GABA concentrations. It was recently shown that astrocytic processes, identified by their immunoreactivity for GFAP, express specific GABA transporters. This specific GABA uptake system is restricted to astrocytes near GABAergic synapses.

5. Benzodiazepine Receptor and Natural Ligands

Mitochondrial benzodiazepine receptors (MBRs) also called peripheral-type benzodiazepine receptors, are extremely abundant in steroidogenic cells and astrocytes. MBR ligands stimulate progesterone synthesis by isolated mitochondria due to MBR translocation of cholesterol, its precursor, from the outer to the inner mitochondrial membrane. The endogenous natural ligands for MBR have been termed "endozepines" because they displace benzodiazepines from their binding sites on GABA A receptors and MBR. All endozepines are so far derived from a 10-kDa polypeptide called diazepam binding inhibitor, which can generate through proteolytic cleavage several biologically active peptides. One of them, an octadecaneuropeptide, is found in high concentrations in glial-rich parts of the brain and is a much more potent inhibitor of [^{3}H]diazepam binding in glial cells than in neurons. MBR and progesterone synthesis are also increased in reactive glial cells.

D. Modulation of Synaptic Activity

As previously mentioned, astrocytes possess many of the characteristics that would allow them to play an active role in the regulation of neuronal activity, particularly in synaptic transmission. Astrocytes tightly surround synapses; they possess ion channels and receptors for various neurotransmitters and are therefore in a good position to regulate the molecular environment of their associated neuronal set. As discussed previously, two mechanisms, not mutually exclusive—gap junction coupling and neurotransmitter release—may mediate the propagation of Ca^{2+} waves in the astrocyte network. Other neurotransmitters, such as norepinephrine and GABA, have also been shown to evoke $[Ca^{2+}]_i$ increases in astrocytes from hippocampal slices, indicating that neuron/astrocyte signaling is not restricted to glutamatergic synapses. Moreover, it has been shown that astrocytes are in turn able to release glutamate and activate the neighboring neurons. Glutamate released from astrocytes can activate either neuronal NMDA receptors or mGluRs and in this way facilitate or depress synaptic activity. Overall, when releasing glutamate, presynaptic neuronal endings signal not only to postsynaptic neurons but also to perisynaptic astrocytes, which can reply by releasing glutamate and inducing $[Ca^{2+}]_i$ changes in the same or in different neurons. This activation of astrocytes by synaptically released glutamate is finely tuned, depending on the intensity and frequency of neuronal activity, indicating plasticity of the astrocyte response.

Other transcellular messengers, such as nitric oxide (NO) and arachidonic acid or its metabolites, are involved in the interactions between neurons and astrocytes. For instance, astrocytes as well as neurons respond to glutamate by releasing arachidonic acid, which in turn participates in the Ca^{2+}-dependent glutamate release by astrocytes. Arachidonic acid and ATP also stimulate glycogenolysis in astrocytes; these molecules are involved in the metabolic cross talk between astrocytes and neurons. Glial cells also produce NO in response to cytokines released in inflammatory conditions, thereby affecting the reuptake of glutamate by astrocyte. In this context, excessive NO production in the brain has been correlated with neurotoxicity and the pathogenesis of

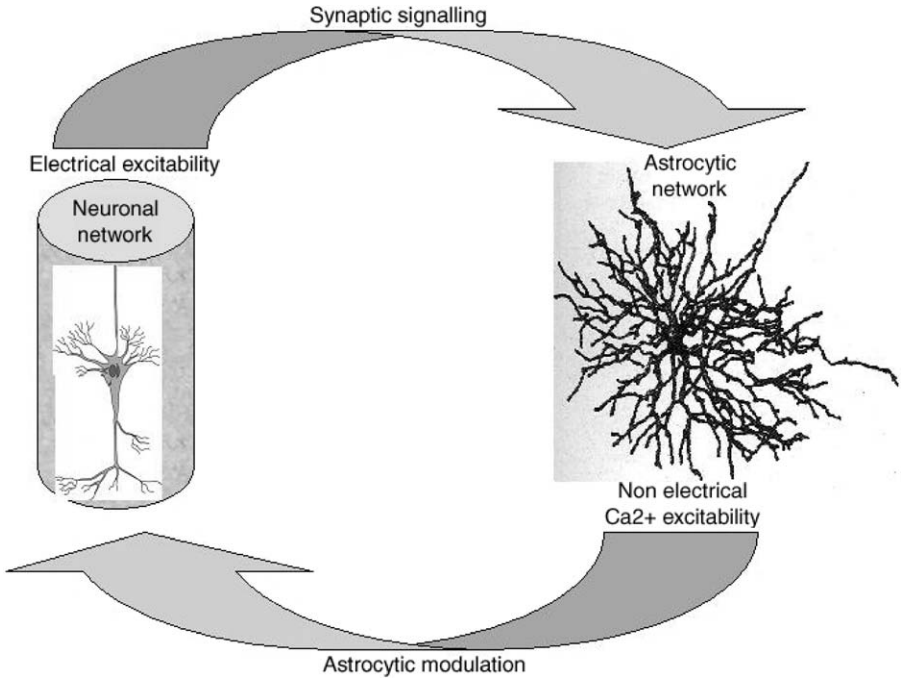


Figure 4 New role for astrocytes: processing of information in the CNS (adapted from S. Smith, 1994, *Curr. Biol.* **9,** 807–810).

several neurodegenerative diseases, such as Alzheimer's disease.

Prostaglandins (PGs) are synthesized in astrocytes and also induce a Ca^{2+}-dependent glutamate release. In many brain pathologies or injuries, PG levels have been reported to increase, suggesting that the dysregulation of PG-dependent glutamate release from astrocytes could participate in neurotoxic cascades. These data again indicate the existence of an integrated glutamatergic cross talk between neurons and astrocytes that may play critical roles both in synaptic plasticity and in neurotoxicity.

Another aspect recently evaluated is the role of glial cells in synaptogenesis. In glia-free cultures, retinal ganglion cells form synapses with normal ultrastructure but display little spontaneous synaptic activity and high failure rates in evoked synaptic transmission. In contrast, when cocultured with glial cells, retinal ganglion cells show an increase in the number, frequency, and amplitude of spontaneous postsynaptic currents and also fewer transmissions failures.

Therefore, although lacking the excitability associated with neurons, glial cells are more actively involved in brain function than had been previously thought. Indeed, it appears that glia is intimately associated with neurons through an exchange of signals regulated by the same types of receptors as those found in neurons (Fig. 4).

IX. PHYSIOLOGICAL PLASTICITY OF ASTROCYTES IN THE ADULT CNS AND NEURON–GLIA INTERACTIONS

In certain regions of the adult CNS, astrocytes display a morphological and functional plasticity in relation to particular physiological states, such as circadian rhythms, parturition and lactation, and osmotic stimulation, as well as to pathological states.

A. Neuron–Glia Interactions in the Hypothalamus and the Pituitary

Although the close anatomical association of glia and neurons is a familiar notion, we are only beginning to appreciate that this association is dynamic and modifiable. A neuronal system, which well exemplifies this is the hypothalamoneurohypophysial system, in which neurons and glial cells undergo physiologically linked morphological changes throughout life. Thus, during conditions that enhance neurohormone secretion

(parturition, lactation, and prolonged osmotic stimulation), glial coverage of oxytocinergic neurons markedly diminishes and their surfaces are left in extensive juxtaposition. Concurrently, there is formation of new synapses, which are predominantly GABAergic and which couple two or more oxytocinergic neurons simultaneously; glutamatergic synapse numbers also increase significantly. The same stimuli induce similar changes in the posterior pituitary, in which retraction of pituicyte processes from posterior pituitary neurons and growth and multiplication of neurosecretory axon terminals lead to an increase neurohemal contact. These structural changes do not permanently modify the anatomy of the system since upon cessation of stimulation neuronal juxtapositions and shared synapses disappear, and they reappear upon new stimulation. Activity-related changes in the conformation of glia, neurons, and synaptic inputs also occur in other hypothalamic–neurosecretory systems, such as the arcuate nucleus in relation to varying estrogen level.

B. Circadian Rhythms

The suprachiasmatic nucleus (SCN) of the anterior hypothalamus is the site of the endogenous clock controlling circadian rhythms in mammals. Within the SCN, there is a circadian fluctuation in GFAP immunoreactivity distribution, in parallel with energy metabolism, electrical activity, and peptide synthesis, thus implicating astrocytes in the mechanism of the circadian clock. Neurons of the SCN display endogenous circadian oscillations of spike-firing rates *in vivo* as well as in slice preparations. Propagating waves of increased Ca^{2+} moving through the astrocytic network were noted in the isolated SCN following glutamate stimulation; glutamate inputs from retinal ganglion cells to the SCN could be the trigger for Ca^{2+} waves in these astrocytes *in vivo*.

X. ASTROCYTE AND PATHOLOGICAL STATES

A. Human Defects in Neuronal Migration

In the human, the major neuronal migrations that form the cortical plate occur by the 16th week of gestation, and late migrations from the germinal matrix into the cerebral cortex continue until 5 months postnatally. The external granule layer of the cerebellar cortex continues to migrate until 1 year of age. Abnormalities of these migratory processes have pathological consequences.

The neural tube of the early human embryo is poorly vascularized. It may rely on the glial processes to transport nutrients between the ventricular system and the glycogen-rich meningeal tissue. Any pathological event, such as ischemia or hypoxia, that interrupts or distorts the radial glial fiber system between the ependymal and pial surfaces of the brain may create abnormalities of neuronal migration. If a premature infant suffers a subependymal hemorrhage, the radial glial process that is guiding neurons to the surface may retract from the cortical surface after its cell somata is destroyed. An immature neuron already in migration along this process may have escaped destruction within the zone of hemorrhage, but it is able to migrate only as far as the end of the glial guide fiber. The migrating cells in the company of several others whose radial fibers are also retracted for the same reason create a group of heterotopic cells that are unable to complete their journey. They mature *in situ* but are unable to establish their intended synaptic relations; consequently, there is a faulty synaptic circuitry. Failure of migrating cells to reach their proper destination may be referred to as neuronal ectopia. Incidental findings at autopsy may be related to the development of an epileptic focus.

In an animal model of Down's syndrome, the trisomy 16 mouse, there is a reduction of radial units in early telencephalic development, explaining the reduction in the final telencephalic size. Most mammals, including all primates, develop convolutions to provide a large cortical surface without incurring a concomitant increase in cerebral volume. Abnormal convolution patterns of the infant brain (i.e., pachygyria, lissencephaly, and polymicrogyria) are the result of faulty neuronal migration and are associated with an abnormally laminated cortex. Some of these are being characterized at the molecular level. The recently sequenced human gene for astrotactin, is a candidate for unsolved neuronal migration disorders in man. Even small perturbations can produce defects; abnormally migrated neurons and glia via a rupture of the glia limiting membrane have been found in the perisylvian cortex of some dyslexic brains.

B. Gliosis

Astrocytes have a capacity to react mainly in relation to an injury to the CNS and constitute reactive gliosis. Although there may also be associated proliferation

and reactivity of microglial cells, the term gliosis is classically related to astrogliosis. This reaction is characterized by hypertrophy of astrocytes, both cytoplasmic and with enlarged nuclei, associated with a profusion of long, thick cytoplasmic processes. There is also an increase in gliofilament number and their constituent, GFAP. The gliotic reaction may also consist of hyperplasia, with the proliferation occurring close to an acute lesion; in other cases, the proliferation of astrocytes may be discrete, even absent, and its degree seems to depend on the kind of lesion, the region, and the state of development of the brain. The proliferation of astrocytes, when it occurs, could originate from multipotential stem cells or glial precursors still widely present in adult CNS. There could also be dedifferentiation of mature astrocytes, which could explain the expression or overexpression of molecules during the first stages of astrogenesis by recapitulation of ontogenic stages.

There are two types of reactive astrocytes. One is isomorphous gliosis, in which there is a regular pattern of astrocytes that are parallel to degenerating axons—for instance, in Wallerian degeneration, in which glial fiber organization preserves a normal structure. This is observed in slowly degenerative lesions or at distance from a lesion. There is also an anisomorphic gliosis in which proximal reactive astrocytes, close to a lesion, form a dense mesh with no discernable pattern. Astrocyte reactivity is more intense in gray matter than in white matter, in which there is already a higher level of GFAP expression. Other markers of astrocytes are also expressed by reactive astrocytes, such as GS and S100 beta. Activated S100 beta$^+$ astrocytes are dramatically increased in the brains of patients with Alzheimer's disease and in the Alzheimer-like neuropathological changes observed in Down's syndrome. This could favor calcium-mediated events in Alzheimer's disease, such as excessive phosphorylation of the tau protein present in neurofibrillary tangles, that could ultimately result in the neuronal cell death characteristic of this disease. The concomitant considerable increase in GFAP, even in regions that do not present the Alzheimer pathology, may reflect the prominent role played by astrocytes during this pathology. Interestingly, in Alzheimer's disease the production of β-amyloid precursor protein (APP) is increased in reactive astrocytes, as is the ApoE isoform, which is associated with the pathophysiology of APP metabolism. In this way, reactive astrocytes may be involved in the processing of APP, perhaps contributing to β-amyloid deposition in Alzheimer's disease.

Gliosis is also a secondary process observed during aging as well as in many pathological conditions affecting neural cells, such as brain trauma, ischemia, experimental autoimmune encephalomyelitis (EAE), and demyelinated areas of multiple sclerosis. The correlation of AIDS dementia with a high level of astrocytic expression of adhesion proteins such as VCAM-1 and ICAM-1 may be involved in cellular dysfunction. In myelin mutants, such as the *jimpy* mouse and the *md* and the *taiep* rat, gliosis is severe; nevertheless, axonal growth is not impaired.

The physiological role of astrogliosis remains controversial with respect to the beneficial or detrimental influence of reactive astrocytes on CNS recovery. On the one hand, the very dense network of processes built up in the scar by reactive astrocytes suggests that the scar tissue may fulfill important functions as a barrier isolating and protecting the intact tissue from the lesions, from which toxic molecules could be released. On the other hand, molecules expressed in lesion scars on the astroglial cell surface or secreted molecules render the reactive astrocyte a less favorable substrate, which could be inhibitory to neuritic outgrowth. Proteoglycans such as chondroitin sulfate proteoglycans may act as inhibitors of neurite outgrowth by attenuating the potential for axon elongation that could occur due to a concomitant expression of growth-promoting molecules such as laminin in regions of reactive gliosis. There also seems to be regional differences in the capacity for the gliotic astrocytes to secrete inhibitory molecules.

In the injured brain, the release of immunoregulatory cytokines by cells around lesion sites may be a mechanism contributing to the induction of gliosis. Molecular signaling may occur between lesioned neurons, glia, inflammatory cells, fibroblasts, and meningeal cells. Among these gliosis signaling molecules are macrophage inflammatory protein (MIP)-1α and MIP-1β, tumor necrosis factor-α, transforming growth factor-β, bFGF, interleukin-1, LIF, and CNTF.

C. Protective Role of Astrocytes

Glial cells may be involved in the differential vulnerability of dopaminergic neurons in Parkinson's disease; this is supported by the inverse relationship between the degree of neuronal loss and the density of astroglial cells initially present in these groups. For instance, even in the substantia nigra, the density of astroglial cells is lowest in the area where dopaminergic

neurons degenerate. Astrocytes may protect dopaminergic neurons from degeneration by metabolizing dopamine and scavenging oxygen free radicals that are associated with dopamine metabolism. Glutathione peroxidase and catalase are involved in the detoxification of H_2O_2 by astroglial cells.

Astrocytes may also protect neurons by secreting neurotrophic factors, although the respective role of astrocytes and neurons in this secretion *in vivo* remains unclear. Astrocyte synthesize metallothionein proteins, which present a neuroprotective role in inflammatory conditions.

According to their role in the reuptake of neurotransmitters, astrocytes are also involved in the control of glutamate-induced neurotoxicity. Astrocyte-specific glutamate transporters play a pivotal role in the clearance of glutamate synaptically released and therefore in the maintenance of glutamate at a physiological level, thus avoiding excitotocity.

D. Alexander's Disease and Its Animal Model

Alexander's disease is a leukodystrophy in which there is accumulation within the astrocytic cytoplasmic processes of eosinophilic, beaded inclusions, the Rosenthal fibers, which contain small heat shock proteins (e.g., α-B crystallin). The pathophysiological mechanisms responsible for this disease, and the relation between affected astrocytes and the demyelination characteristic of this pathology, have not been identified. Astrocytes of transgenic mice overexpressing GFAP contain cytoplasmic inclusion bodies identical histologically and antigenically to the Rosenthal fibers seen in Alexander's disease. Mice in the highest expressing lines die by the second postnatal week. Recently, sequence analysis of DNA samples from patients representing different Alexander's disease phenotypes revealed that most cases are associated with nonconservative mutations in the coding region of GFAP. Alexander's disease represents the first example of a primary genetic disorder of astrocytes, one of the major cell types in the CNS. Therefore, the GFAP-overexpressing transgenic mice provide an animal model for studying the interactions between astrocytes, oligodendrocytes, and myelin. Another important role of astrocyte in myelin maintenance and blood–brain barrier function was revealed by the study of an other animal model, the GFAP-deficient mouse, in which the white matter is poorly vascularized and the blood–brain barrier is structurally and functionally impaired.

E. Gliomas

Malignant gliomas are extremely invasive tumors; they are the most common type of malignant brain tumors in adults (40%). Most of them are of astrocytic origin. Despite aggressive treatment assays such as surgery, chemotherapy, and radiotherapy, mean survival rates for this disease are less than 1 or 2 years, depending on the clinical grade of the disease. Gene therapy has been used to kill tumor cells in animal models; the classic example of this strategy is the retroviral transfer of the herpes simplex virus thymidine kinase (TK) gene to tumor cells followed by treatment with the antiviral compound gancyclovir to kill cells expressing TK. This strategy was successful in rodents due to the bystander effect, in which cytotoxicity is transferred from cells expressing TK to non-TK-expressing cells. Unfortunately, it did not present any therapeutic effect in humans, presumably because of the low infection rate within human tumors and also because of the dissemination of the tumor cells in the human brain, impairing the bystander effect observed in rodent tumors.

Recently, attenuated, nonneurovirulent versions of herpes simplex virus have been used and have been shown to kill glioma cells in culture, and they have also been implanted in rodents. These viruses are currently in clinical trials. The pathophysiological mechanisms of gliomas have not been clearly identified.

XI. CONCLUSIONS

The importance of glia has become increasingly clear with the development of molecular biology and cell culture techniques. With technical progress, the roles of glia in neuronal migration in the development of neuronal pathways as well as in synaptic functions have been deciphered. Increasingly, the molecules involved in developmental processes and in the adult are being identified; molecules necessary for the migration of neurons on radial glia or Bergmann cells are made by neurons or glia with multiple interactions. Molecular studies of developmental mutants and human pathologies have led to the identification of the involvement of glia in numerous defects of migration that lead to microcephaly and other developmental diseases.

In many cases, axonal guidance seems to involve preformed glial pathways that may remain and create glial boundaries. Increasingly these neuroglial interactions are being identified in relation to neuronal

functions. Because of their mobility and plasticity, glial cells appear to be increasingly involved in the functions of the cabled neuronal network (Fig. 4). Synapses throughout the brain are ensheathed by astrocytes. Astrocytes help to maintain synaptic functions by buffering ion concentrations, clearing released neurotransmitters, and providing metabolic substrates to synapses. As recently reviewed, glia should be envisaged as integral modulary elements of tripartite synapses because they are now playing an active role in synaptic transmission and are fully involved in neuron–astrocytes circuits in the processing of information in the brain. They are indispensable in obtaining nutrients from the blood and helping to maintain the blood–brain barrier. For energy metabolism, these glial cells take up glucose from the brain capillaries and transform it into lactate and other fuels absolutely necessary for the neurons to function. The metabolic coupling between glia and neurons is increasingly obvious in view of the development of the methods of investigation, even *in vivo*; astrocytes contribute to the deoxyglucose signal in PET, which may give new insights into the interpretation of this signal in neurological and psychiatric disorders. Astrocytes are necessary to avoid the excitotoxic role of glutamate through the glutamate–glutamine cycle, which is pivotal, as are probably other neurotransmitter cycles. One of the recent developments is the way in which communication can occur through glial cells by calcium waves; this seminal discovery has been followed by a wealth of work demonstrating that calcium signaling can extend even to neurons and can be bidirectional. It is possible that astrocytes may provide new means of communication in the nervous system and new pathways not yet clearly defined. No doubt, there are enormous gaps to fill in relation to their functions *in vivo*: hints have been provided, for example, by the observation that they are modulated by circadian rhythms and hormonal states.

Although myelin repair and synaptic remodeling and regeneration can occur, many enigmas still remain, especially in humans, in which the factors may be different from those in the murine species. Thus, studies in primates and *in vivo* systems cannot be omitted at this stage in view of therapeutic implications.

The dysfunction of glial cells is possibly at the origin of many of the degenerative diseases of the nervous system and the major brain tumors (glioma). Although the neuroimmunological role of astrocytes as antigen presenting cells is still unclear in the CNS under *in vivo* conditions, their role in neurodegenerative diseases seems increasingly evident because they are implicated in the suppression of oxidative stress. No doubt, in the near future, we will understand more about the cross talk between glial cells and neurons in normal and pathological states. Already, abnormal astrocytes and oligodendrocytes appear to be involved in cognitive functions as evidenced from leukodystrophies related to oligodendrocyte or astrocyte genetic disorders. Recently, the primary genetic defect of Alexander disease was demonstrated in astrocytes where mutations of GFAP lead to a secondary demyelinating disease, enlightening the pivotal role of astrocyte on oligodendrocyte and myelin maintenance.

Progress in neuroscience has shown that neurons and glia do not represent just the addition of independent compartments and that the cooperation of both cell populations is essential for development and functions of the nervous system. As mentioned by Peschanski, the time has come for "neurogliobiology" because neurons and glia (including astrocytes, oligodendrocytes, and microglia) in the nervous system are undissociable partners.

See Also the Following Articles

BRAIN LESIONS • CIRCADIAN RHYTHMS • GABA • GLIAL CELL TYPES • NERVOUS SYSTEM, ORGANIZATION OF • NEURON

Acknowledgments

Philippe Mas is greatly acknowledged for providing the figures. The authors' research work was supported by INSERM and VML (to N. Baumann) and ARSEP, ELA, and CNRS (to D. Pham-Dinh).

Suggested Reading

Alvarez-Buylla, A., Garcia-Verdugo, J. M., and Tramontin, A. D. (2001). A unified hypothesis on the lineage of neural stem cells. *Nat. Rev. Neurosci.* **2,** 287–293.

Araque, A., Parpura, V., Sanzgiri, R. P., and Haydon, P. G. (1999). Tripartite synapses: Glia, the unacknowledged partner. *TINS* **22,** 208–215.

Barres, B. A., and Barde, Y. (2000). Neuronal and glial cell biology. *Curr. Opin. Neurobiol.* **10,** 642–648.

Brenner, M., Johnson, A. B., Boespflug-Tanguy, O., Rodriguez, D., Goldman, J. E., and Messing, A. (2001). Mutations in GFAP, encoding glial fibrillary acidic protein, are associated with Alexander disease. *Nat. Genet.* **27,** 117–120.

Fields, R. D., and Stevens, B. (2000). ATP: An extracellular signaling molecule between neurons and glia. *TINS* **23,** 625–633.

Franklin, R. J. M., and Blakemore, W. F. (1995). Glial cell transplantation and plasticity in the O-2A lineage - implications for CNS repair. *TINS* **18,** 151–156.

Gallo, V., and Ghiani, C. A. (2000). Glutamate receptors in glia: New cells, new inputs and new functions. *TiPS* **21,** 252–258.

Hatten, M. E. (1999). Central nervous system neuronal migration. *Annu. Rev. Neurosci.* **22,** 511–539.

Holland, E. C. (2000). Glioblastoma multiforme: The terminator. *Proc. Natl. Acad. Sci. USA* **97,** 6242–6244.

Laming, P. R., Kimelberg, H., Robinson, S., Salm, A., Hawrylak, N., Muller, C., Roots, B., and Ng, K. (2000). Neuronal–glial interactions and behaviour. *Neurosci. Biobehav. Rev.* **24,** 295–340.

Laywell, E. D., Rakic, P., Kukekov, V. G., Holland, E. C., and Steindler, D. A. (2000). Identification of a multipotent astrocytic stem cell in the immature and adult mouse brain. *Proc. Natl. Acad. Sci. USA* **97,** 13883–13888.

Magistretti, P. J., Pellerin, L., Rothman, D. L., and Shulman, R. G. (1999). Energy on demand. *Science* **283,** 496–497.

Malatesta, P., Hartfuss, E., and Götz, M. (2000). Isolation of radial glial cells by fluorescent-activated cell sorting reveals a neuronal lineage. *Development* **127,** 5253–5263.

Momma, S., Johansson, C. B., and Frisen, J. (2000). Get to know your stem cells. *Curr. Opin. Neurobiol.* **10,** 45–49.

Noctor, S. C., Flint, A. C., Weissman, T. A., Dammerman, R. S., and Kriegstein, A. R. (2001). Neurons derived from radial glial cells establish radial units in neocortex. *Nature* **409,** 714–720.

Poitry, S., Poitry-Yamate, C., Ueberfeld, J., MacLeish, P.R., and Tsacopoulos, M. (2000). Mechanisms of glutamate metabolic signaling in retinal glial (Muller) cells. *J. Neurosci.* **20,** 1809–1821.

Vesce, S., Bezzi, P., and Volterra, A. (1999). The active role of astrocytes in synaptic transmission. *Cell. Mol. Life Sci.* **16,** 991–1000.

Attention

SHAUN P. VECERA and STEVEN J. LUCK
University of Iowa

GLOSSARY

attentional blink Following the identification of a target in a rapid sequential stream of stimuli, there is temporary impairment in identifying a second target stimulus that occurs shortly after the first. It is as if attention has ''blinked'' after identifying the first target, preventing subsequent target detection.

binding problem The problem of determining which elementary visual features, such as color, shape, or size, belong to the same stimulus. Spatial attention allows elementary features to be bound or grouped together.

grouped-array representation A spatial representation that contains perceptual grouping information. Image features (e.g., line segments) are grouped to form larger structures, such as surfaces or regions.

neglect (also extinction) A neuropsychological disorder of visual attention that follows damage to the posterior parietal region. Patients with neglect fail to attend to events on the side of space opposite the lesioned hemisphere. As patients recover from neglect, they demonstrate extinction, a form of transient neglect that occurs when events occur on both sides of space simultaneously.

object-centered representation A representation of an object that encodes the properties of the object, such as parts, relative to a reference point on the object itself, such as a principle axis.

visual search The process used to find a visual target (i.e., how you find what you are looking for). Both elementary visual features, such as color, shape, or size, and focal visual attention combine to guide the search for a target.

At any given moment, the sensory systems in your brain are receiving thousands of simultaneous environmental inputs. Some of these inputs are relevant to your current behavior and others are irrelevant. The visual words printed on a page are relevant to reading, but the visual impression of the desk on which the pages lie is irrelevant to reading. Furthermore, some of the inputs in one modality have correspondences with inputs in another modality, as in the link between a person's visual appearance and the speech uttered by that individual. Because we cannot process all inputs simultaneously, there must exist processes that select some inputs and filter out others. These processes collectively are referred to as ''attention.''

The study of attention has a long history in both the cognitive and brain sciences. Although early research implicitly assumed that attention is a single, monolithic process, an emerging view that we endorse is that there are multiple forms of attentional selection. Attention can select stimuli at specific locations in vision, audition, or touch; attention also can select entire objects, not just locations. Furthermore, attention selects not only stimulus inputs but also mental functions such as behavioral goals or tasks: You can attend to the task of reading instead of the task of identifying a font type. To understand attention requires studying these multiple forms of selection, their similarities and dissimilarities, and their neural foundations.

I. CONTROL OF VISUAL SPATIAL ATTENTION

A. Multiple Constraints on Selection: A Framework

Perhaps the most complete understanding of attentional selection is in the visuospatial domain, in which stimuli at specific locations are selected for processing. What are the critical parameters that determine those inputs that receive attention and those that do not? This is the question of attentional control.

There are different parameters or processes that can influence attentional control. Two general classes of control are top-down sources that arise from the current behavioral goals and bottom-up sources that arise from sensory stimuli present in a scene. These two sources can be illustrated by considering visual search, the act of looking for a visual target among distractors (e.g., finding a friend's face in a crowd). In a typical visual search task, observers are instructed to search for a particular target, such as a black vertical line, that appears in a field of distractors (Fig. 1). The target description can be conceptualized as a "template" temporarily stored in memory that can influence visual search in a top-down manner; as an observer, you would actively attempt to look at black and vertical items. The scene presented in a visual search task provides the bottom-up information that is searched; this information indicates where objects are located and which features are present at each location. Effective visual search would require finding a balance between the top-down information and the bottom-up information. An example of an effective search is searching for a single feature, such as a black vertical line among white vertical lines (Fig. 1a). In this example of a feature search, color uniquely defines the target, and the bottom-up information is consistent with the top-down information in constraining where an observer should search. A less efficient search would involve searching for a conjunction of features, such as a black vertical line among black horizontal lines and white vertical lines (Fig. 1b). In this search, any single piece of bottom-up information is not unique to the target item, so the bottom-up constraints are weaker than in the feature search. Top-down constraints would be required to resolve the competition among the input items. In general, the control of spatial attention can be viewed as a "biased competition" model: Competition among bottom-up inputs is biased (i.e., some of the inputs are favored) by top-down inputs, such as a target template.

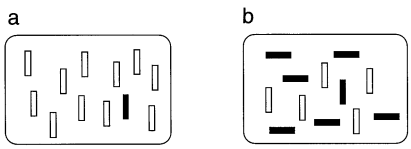


Figure 1 Sample visual search task: search for a black vertical line. (a) Feature search in which the target pops out from a homogeneous background. (b) Conjunction search in which the target does not pop out because the distractors share both black and vertical features with the target.

How is visual search controlled, particularly when search is inefficient and not determined by bottom-up information? Two possible control modes have been hypothesized for visual search. The serial search account proposes a sequential control of attention in which attention shifts from one item to the next until the target is found. The parallel search account proposes that attention is allocated to every item simultaneously, with less attention available for each item when a large number of items must be searched simultaneously. Much debate has surrounded the serial/parallel dichotomy, and current perspectives on the issue focus on the efficiency of search and not on an absolute dichotomy. Although there is evidence for a serial control process, this serial control process must be implemented in the brain's parallel hardware. Furthermore, if multiple attention systems exist, then some forms of attentional control may be more serial and others may be more parallel.

The control of spatial attention has also been examined by directing attention to a location before a target event occurs. In these "spatial precuing" tasks, observers are required to detect or identify a target item that appears at a peripheral location. Before the target appears, one location is precued by an arrow pointing to the location, a flash of light at the location, or some similar means. The subsequent target usually appears at the cued location (a "valid" trial), although it may occasionally appear at an uncued location (an "invalid" trial). Because the cue is usually valid, observers are motivated to attend to the cued location, and observers typically detect valid location targets more quickly and accurately than invalid location targets.

The type of precue used can bias attentional control to favor bottom-up factors or top-down factors. For example, if the precue is an abrupt luminance change (e.g., a flicker in the visual periphery), attention is automatically captured by the bottom-up input, irrespective of the observer's intentions. Such "exogenous" cues are extremely difficult to ignore and they

are not interfered with by concurrent tasks such as a memory task. In contrast, if the precue is a symbol such as a centrally presented arrow that points to a location, attention will move to the cued location only if the observer wants to shift attention, and when attention does move it moves more slowly. These "endogenous" cues are not automatic: They can be ignored, and they are interfered with by concurrent tasks. Because endogenous cues are dependent on task-related goals and observers' expectancies, they involve top-down control processes. In many spatial precuing tasks, the control of attention involves a balance between bottom-up and top-down factors. Although bottom-up onset cues capture attention, they may be influenced by top-down attentional control settings (e.g., expectations of where the target will appear).

B. Neuroanatomy of Control

1. Neuropsychology: Neglect and Extinction

There are undoubtedly several cortical and subcortical areas that participate in the control of spatial attention. The pulvinar nucleus of the thalamus, for example, is involved in filtering or suppressing irrelevant stimuli in a cluttered display. However, the cortical region that plays the most significant role in the control of spatial attention is the posterior parietal region. Damage to the parietal region (especially the right parietal region) in humans results in a profound attentional impairment referred to as neglect. Neuropsychological patients with neglect fail to pay attention to stimuli falling on the side of space opposite the lesion (the contralesional side). For example, a patient with damage to the right parietal lobe may fail to acknowledge a person sitting on the left, may fail to eat food on the left half of a plate, and may fail to read words on the left half of a page. Neglect occurs soon after damage to the parietal region. As a patient recovers and the neglect becomes less severe, patients can process a single stimulus presented in the contralesional visual field. These recovering patients show another disorder, however, referred to as extinction: When two stimuli are presented simultaneously in opposite visual fields, patients will extinguish, or fail to notice, the stimulus in the contralesional field. In other words, extinction patients exhibit neglect of contralesional stimuli only in the presence of ipsilateral stimuli. Both neglect and extinction appear to be disruptions of the ability to control spatial attention and deploy it to the contralesional field.

Neglect can be observed in spatial precuing tasks. In these tasks, extinction patients can detect and identify targets presented in the contralesional visual field; furthermore, these patients can use exogenous precues to allocate attention to the contralesional field prior to the appearance of the target. However, when a precue appears in the ipsilesional field and a target appears in the contralesional field, these patients are much slower to detect the target than when the contralesional field is cued and the target appears in the ipsilesional field. That is, they are impaired primarily when they are cued to the good field and then the target appears in the bad field. A straightforward interpretation of these results is that the contralesional and ipsilesional sides of space compete for attention. A bottom-up factor, such as a spatial precue, can bias the competition in favor of the cued field. The effect of parietal damage is to weaken the ability of the contralesional field to compete for attention. When both the cue and the target are in the same field, there is no competition and the target can be detected quickly. However, when the good field is cued and the target appears in the bad field, the good field wins the competition for attention even though the target is in the bad field, leading to abnormally slow responses.

What aspect of attentional control, bottom-up or top-down, is disrupted in these patients? Although parietal-damaged patients appear to have intact perceptual processing, the disorder of attention appears to involve bottom-up control parameters: Attention is not captured effectively by contralesional inputs. Furthermore, some forms of top-down attentional control appear to be intact in parietal-damaged patients. These patients can make use of top-down expectancies or task-relevant goals. For example, a contralesional stimulus may not be extinguished if the ipsilesional stimulus is irrelevant to a task and the patient is instructed to ignore this ipsilesional stimulus. Presumably, the top-down control of attention is intact in these patients and biases attention to select the relevant item in the bad field.

2. Neuroimaging Studies

Neuroimaging studies also indicate a central role for the posterior parietal lobe in spatial selection. Because the whole brain can be examined using some of these techniques, other brain regions that influence spatial attention can be observed. Observing multiple neural sites simultaneously may be useful for isolating the sources of bottom-up and top-down control.

Separate neuroanatomical sources for two forms of control is suggested by positron emission tomography (PET) studies of performance in the spatial precuing task. Observers were presented with sequences of visual targets that appeared in a predictable sequence that would engage endogenous attentional allocation. The predictable sequences were leftward or rightward appearances of the target; the target first appeared near fixation to the left or right and then continued to move in the same direction in a majority of the trials. For example, if the first target first appeared slightly to the right of fixation, the second target would likely occur to the right of the first target's position; similarly, the third target would likely occur to the right of the second target, and so forth. A predictable sequence allows observers to anticipate the next target location and endogenously allocate attention to that expected location. Thus, the peripheral targets in this task involve both exogenous and endogenous components—the appearance of the target is an exogenous luminance change and the predictable sequence allows observers to anticipate the next target and allocate attention endogenously. Two neural regions of interest exhibited increased blood flow during this task: the superior parietal lobe near the postcentral sulcus (near Brodmann's area 7) and the superior frontal cortex (near Brodmann's area 6).

To distinguish the superior parietal and superior frontal areas by their sensitivity to the exogenous and endogenous components, observers performed a control task. This control task presented the same peripheral targets in a predictive sequence, but instead of detecting these peripheral targets observers detected targets presented at fixation. Exogenous orienting would occur to the peripheral targets, even though these targets were irrelevant to the observers' task. However, endogenous attention would be directed to the central targets because these were the task-relevant stimuli. When endogenous shifts of attention were eliminated in this manner, the superior frontal areas were no longer active, but the superior parietal areas continued to be active. These results suggest that superior parietal regions are involved in the exogenous, bottom-up control of spatial attention and that superior frontal regions are involved in the endogenous, top-down control of spatial attention.

The bottom-up control of spatial attention coordinated by the superior parietal lobe appears to involve spatial selection only. If other visual attributes are selected, such as a color or a shape, the superior parietal lobe does not appear to show increased blood flow as measured by PET, although other extrastriate visual areas are activated in response to different visual attributes. For example, in searching for a target defined by color, blood flow increased in the left dorsolateral occipital lobe and in the left collateral sulcus. Searching for a target defined by movement or by shape resulted in increased blood flow in other extrastriate areas. Thus, there appears to be a large network of extrastriate visual areas that each mediate bottom-up aspects of attention to different stimulus attributes. Furthermore, PET results suggest that each of these areas could receive feedback from frontal lobe areas—areas that may represent task-relevant goals useful for allowing one stimulus attribute (e.g., color) to be selected from an image containing many attributes (e.g., color, shape, and movement).

There may exist some extrastriate visual regions that participate in attentional control across different visual attributes. Recent functional magnetic resonance imaging (MRI) results have found activation in two parietal lobe areas across three very different attention tasks: (i) a spatial shifting task similar to that described previously, (ii) an object matching task in which observers reported whether two attended objects were the same or different, and (iii) a nonspatial conjunction task in which observers searched for a target letter in a sequential stream of colored letters. These two parietal subareas seem to be involved in a wide range of visual selection, contrary to the PET results discussed previously that demonstrated no parietal involvement in visual search for targets defined by color, shape, or motion.

However, there is a resolution to the apparent discrepancy between a general attentional involvement of parietal areas and a spatial-specific role for parietal areas: Parietal lobe attention areas may control the suppression of visual distractors. Functional MRI studies that exhibit parietal activation across attention tasks required irrelevant stimuli to be ignored; PET visual search studies that did not exhibit parietal activation across tasks involved displays containing only task-relevant stimuli. In the biased competition account, parietal lobe areas may receive feedback from frontal lobe areas that allow parietal regions to suppress distractors and act as a "gate" for other extrastriate visual areas.

C. Neurophysiology of Control

As with neuroanatomical studies of attentional control, there has been a substantial amount of research

on single-neuron recordings from a range of brain areas. Those regions relevant to the biased competition framework are regions in the parietal and frontal lobes; other important areas, such as the superior colliculus or pulvinar, will not be reviewed here.

Many neurophysiological studies investigate overt spatial attention, in which the eyes overtly move to an attended location, in contrast to covert spatial attention, in which the eyes do not move. One consequence of overt shifts of spatial attention is that stimuli in a visual scene occupy different retinal locations from one eye movement to the next. Covert spatial attention appears to shift before the overt eye movement, allowing the representation of a visual scene to be updated prior to the eye movement. Neurons in the parietal lobe appear to play a role in controlling the focus of spatial attention.

Evidence for parietal lobe involvement in covert shifts of spatial attention derives from single-unit recordings from the lateral intraparietal (LIP) area in monkeys. Neurons in LIP have topographically mapped receptive fields that respond to visual stimulation (Fig. 2a). When a monkey makes an eye movement to a new location, the receptive field of a LIP neuron will also fall in a new location (Fig. 2c). Prior to the eye movement, however, a LIP neuron will respond to visual stimuli at a location based on the planned eye movement that has not been executed (Fig. 2b). That is, the receptive fields of LIP neurons are remapped in anticipation of an eye movement. This shift of the LIP representation of space may provide the neural mechanism for covert shifts of attention that precede and anticipate overt shifts involving eye movements.

The ability of LIP neurons to update their representation of space implies that this area receives inputs regarding the intended eye movement. This input likely comes from the frontal eye fields (FEFs), suggesting that updating the spatial representation is based on endogenous, top-down factors. Neurons in LIP are also able to alter their firing based on exogenous, bottom-up factors, such as the abrupt appearance of a stimulus in a LIP neuron's receptive field. The appearance of a new stimulus inside a LIP neuron's receptive field results in a large increase in the neuron's firing rate. However, if an eye movement brings a stationary stimulus into the neuron's receptive field, only a weak neural response is produced. The abrupt appearance of a stimulus is the critical parameter for evoking a large neural response: If a stimulus appears shortly before an eye movement (approximately 400 ms), there is a large neural response when the eye movement brings the new stimulus into the LIP neuron's receptive field. The "newness" or salience of the stimulus, indicated by the recency of its appearance, in part controls the response of LIP neurons and, presumably, covert spatial attention.

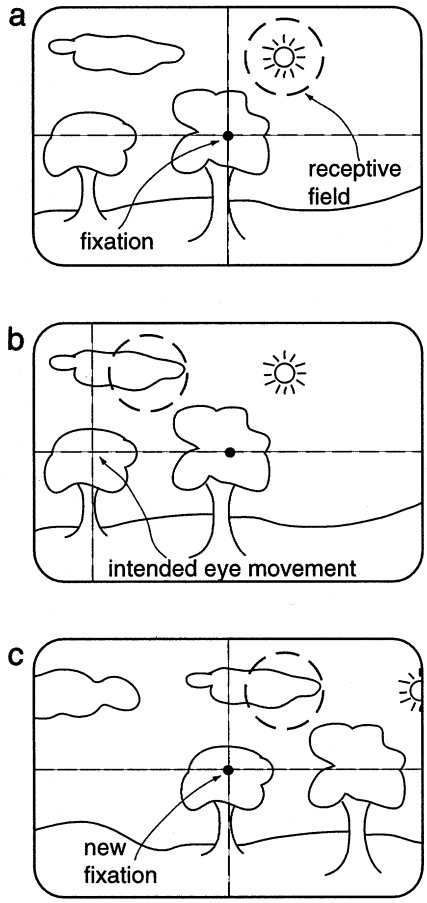


Figure 2 Remapping of receptive fields in area LIP in response to an intended eye movement. (a) The center tree is fixated, and the sun falls within an LIP receptive field. (b) An eye movement to the other tree is planned, and the LIP neuron's receptive field is remapped in accordance with the intended movement. The cloud falls within the receptive field, even though the eyes remain fixated on the center tree. (c) The eye movement is performed, allowing the second tree to be fixated.

Finally, consistent with neuroimaging data, frontal lobe areas participate in the control of attention. In visual search, the FEF is involved in selecting the target to which an eye movement will be directed. Monkeys viewed displays that contained a target that differed from a field of distractors by one feature (Fig. 1a); they were trained to make an eye movement to this target. Prior to the saccade, FEF neurons discriminated target items from distractor items. If the target fell within a FEF neuron's receptive field, the neuron

responded vigorously; if a distractor fell within the receptive field, the neuron responded weakly. Additional studies demonstrated that the enhanced firing of these FEF neurons was not in response to a bottom-up capture of attention by the odd item in the display. Other monkeys were trained to make eye movements to a target defined by color (e.g., make an eye movement to any white target). If a monkey trained to move to a white target viewed Fig. 1a, this monkey would respond by generating an eye movement to any of the white bars. Despite the presence of multiple targets in this situation, FEF neurons continue to show larger firing rates when targets fall in their receptive fields than when the single distractor falls within their receptive field.

In addition to FEF, other areas in prefrontal cortex participate in attentional selection. Many studies implicate dorsolateral prefrontal areas in selection; these studies examined search tasks in which the monkey first sees a cue object that depicts the target for which the monkey must search. Following the presentation of the cue, a display of objects appeared, and the monkey had to search for and remember the location of the target object. During the presentation of the search array, the activity of neurons in the dorsolateral prefrontal cortex was sensitive to the visual attributes of the target only; the distractors were effectively filtered out and did not influence the response of prefrontal neurons. Complementary studies have been performed in extrastriate regions such as the inferior temporal cortex with similar results. The selectivity to target attributes occurs earlier for prefrontal neurons than for inferotemporal neurons, suggesting that target selection first occurs in frontal areas and provides the top-down target template that guides selection in extrastriate areas.

II. EFFECTS OF VISUAL SPATIAL ATTENTION ON PERCEPTION

A. Types of Effects

Having previously discussed how spatial attention can be controlled to focus on an item, we now turn to the effects of attention: How is an attended stimulus processed differently from an ignored stimulus? For example, the representation of an attended item could be either enhanced or suppressed relative to distractor items. An attended item could also integrate all the visual attributes of the attended stimulus (e.g., the color, shape, and size of the stimulus).

1. Locus of Selection: Where Does Attention Have Its Effect?

Since the beginning of attention research, many studies and theories have examined whether attention operates at an early stage of processing (early selection) or at a late stage (late selection). The early selection account proposes that attention operates prior to stimulus identification. Thus, in a visual search task (Fig. 1), an item would be selected first by directing attention to its location and then fully identifying it (e.g., ''black vertical line''). The late selection account proposes that attention operates after all stimuli have been identified to some degree; in visual search, all the items would be recognized in parallel and only the target item would be selected for storage in working memory and for control over behavior.

Under a biased competition account in which multiple forms of attentional selection occur, the early versus late debate has no absolute answer; some forms of attentional selection necessarily will occur earlier or later than others, allowing for both early and late selection. The main issue then becomes determining where effects occur for a given type of attention, such as spatial attention.

Numerous behavioral studies of spatial attention have supported an account in which items are selected at an early sensory level. If a spatial precue orients spatial attention to a region of space, targets appearing at that location (validly cued targets) are detected faster and more accurately than targets appearing elsewhere (invalidly cued targets). Using a signal detection theory approach, performance in such luminance detection tasks can be traced to changes in perceptual sensitivity, independent of changes in response bias. These changes in the sensitivity of perceptual processes also have implications for the second possible effect of spatial attention—that of sensory gain control.

2. Sensory Gain Control

One hypothesized effect of spatial attention is that it may control the gain or amplification of sensory information transmission. Sensory amplification may increase the signal-to-noise ratio between attended and unattended items. As with an early locus for attentional effects, behavioral evidence supports a sensory gain control effect of spatial attention. Because spatial

precues produce a change in perceptual sensitivity, it appears that spatial attention enhances or increases the gain of target items that appear at an attended location.

3. Binding the Attributes of Objects

Another potential effect of spatial attention is that it binds together the various attributes of a stimulus. In cluttered visual scenes that contain many objects (Fig. 1b), the presence of different attributes such as color and orientation is confusing: How does a viewer know that the color "black" belongs to a vertical line when both the color black and vertical lines are present? This question illustrates the binding problem: the difficulty of knowing which attributes belong together by virtue of being aspects of a single object.

Spatial attention may provide the "glue" that binds together the attributes of objects. The probability of binding errors (misconjunctions of feature attributes) can be increased by presenting brief displays while attention is broadly distributed. In such a situation, if a display contained a blue triangle and a red square, observers may occasionally report perceiving a red triangle or a blue square. If observers focus spatial attention on a region in response to a peripheral precue, illusory conjunctions are less likely for objects that appear at the cued location than for objects that appear at an uncued location. Focused spatial attention thus appears to bind together the features of objects.

B. Neuropsychological Evidence

Results from neuropsychological patients with damage to their parietal lobes support the role of spatial attention in solving the binding problem and conjoining the features of objects. In one representative study, an extinction patient was shown two letters in either the contralesional or ipsilesional visual field. One of the letters was a target (F or X) and the other was a distractor (O). The letters were colored, and the patient was instructed to name both the color and the identity of the target letter (i.e., was it F or X and what color was it?). In this task there are two types of errors. The first is a feature error, in which either letter name or color is reported incorrectly. For example, if the patient was presented with a blue F and a red O, reporting a yellow F would be a feature error. The second type of error is a conjunction error, in which a feature of the distractor letter O "migrates" to the target letter, forming an illusory conjunction. For example, if the patient was presented a blue F and a red O and reported a red F, the color of the red O was misconjoined with the target letter F. The extinction patient studied showed many conjunction errors in the contralesional field compared to the ipsilesional field. However, similar numbers of feature errors were made in the contralesional and ipsilesional fields, indicating that feature perception was similar in both fields. Presumably, the damaged parietal-based spatial attention system is unable to permit a correct conjunction of features such as color and shape; the individual features are represented, however, allowing for accurate reports of the features.

Other patients with damage to the parietal lobe attention areas show inabilities to bind features correctly. Patients with bilateral damage to the parietal lobe have Balint's syndrome, which is characterized by an inability to perceive multiple shapes simultaneously (simultanagnosia); Balint's patients can perceive only one object at a time. Recent reports indicated that Balint's patients may show higher than normal rates of illusory conjunctions. If shown a display containing a red X and a blue T and asked to report the name and color of the first letter seen, these patients may often report seeing a blue X and a red T. These misconjunctions even occur when the display is present for several seconds.

C. Physiological Evidence

Neurophysiological and electrophysiological studies have overwhelmingly supported the early locus of selection and sensory gain effects of spatial attention. An early locus for spatial attention has been demonstrated with event-related potential (ERP) studies in which the electroencephalogram is time locked to the appearance of a visual stimulus. The experimental paradigm used most often in these studies is illustrated in Fig. 3a. At the beginning of each trial block, observers are instructed to attend to either the left or the right visual field while maintaining fixation at the center of the screen. Stimuli are then presented rapidly and sequentially to the left and right visual fields, and observers are required to respond when they detect an infrequently occurring target stimulus in the attended field. By maintaining the same sequence of stimuli from trial block to trial block and varying whether the left or right field is attended, it is possible to compare

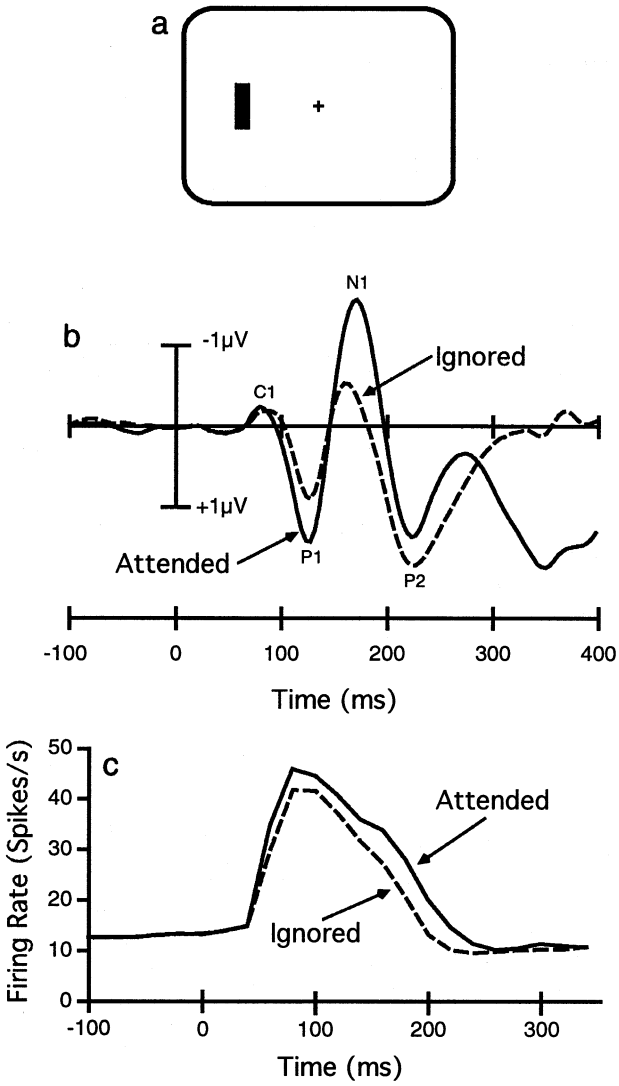


Figure 3 ERP and neurophysiological results from a spatial attention. (a) Observers attend to the left or right side of space, and targets appear at either the attended or ignored location. (b) ERP results show larger P1 and N1 components when targets appear at the attended location than at the ignored location. (c) Neurophysiological results from a representative neuron in area V4. When a target falls within the neuron's receptive field, the neuron fires more vigorously when the target's location is attended than when it is ignored.

the response to the same physical stimulus when it is presented at an attended versus an ignored location.

As shown in Fig. 3b, the ERP waveform recorded over occipital scalp sites in this paradigm consists of a series of positive and negative peaks or components. The earliest component, which is called the "C1 wave" and is observed only under certain conditions, is typically unaffected by attention. Although it is usually difficult to determine the neuroanatomical site at which an ERP component is generated, it is known that the C1 wave is generated in primary visual cortex (area V1), and the finding that the C1 wave is unaffected by spatial attention indicates that the initial volley of V1 activity is not influenced by attention (although V1 activity appears to be modulated by attention at later time points, presumably due to feedback).

The C1 wave is followed by the P1 and N1 waves, both of which are typically larger in amplitude for attended location stimuli than for unattended location stimuli. The P1 effect typically begins before 100 msec poststimulus and combined PET/ERP studies have indicated that it is probably generated in the ventral extrastriate cortex. Moreover, this effect is present for target stimuli, nontarget stimuli, and completely task-irrelevant probe stimuli. Together, these factors indicate that the P1 modulation reflects an effect of attention on sensory processing, supporting early selection models of visual–spatial attention. The N1 effect appears to reflect a modulation of visual discrimination processes, although the precise nature of this effect and its neural origins are not clear.

This paradigm has also been modified for use with single-neuron recordings in monkeys, and a similar pattern of results was obtained (Fig. 3c). Attention had no effect on responses in area V1, but in area V4 (an intermediate visual processing region) attended location stimuli evoked higher rates of neural firing than did ignored location stimuli. Moreover, the effect of attention began at 60 msec poststimulus, which was the same time as the onset of stimulus-related activity in these neurons. However, spatial attention produces no changes in the tuning curves of the neurons. Thus, under certain conditions, attention acts as a preset filter that controls the amplitude of the sensory response in intermediate-level areas of visual cortex.

III. ATTENTION TO OBJECTS

A. Demonstrations of Object-Based Attention

In addition to selecting regions of space, recent research has demonstrated that objects can be attended and selected independently of their locations. Although some investigators have argued that attention is object-based instead of space-based, the emerging consensus is that space-based and object-based attentional systems coexist. A more subtle issue,

however, is whether the objects selected by attention are low-level retinotopically defined regions formed by gestalt grouping processes or higher level invariant objects formed by object representation processes. Again, the emerging consensus is that both types of object selection may exist.

1. Grouped Array Selection

One mechanism for object-based attention involves attending to perceptual objects that are defined in a spatial reference frame—a "grouped array." The grouped array is a spatiotopic map in which locations or features are grouped according to gestalt principles such as similarity (e.g., features similar in color group with one another), closure (e.g., features that form closed shapes group with one another), or figure-ground relations (e.g., figures are closer to the viewer and are more salient than grounds). Selection from a grouped array representation involves attending to the locations of items that are grouped together.

Results from several behavioral studies are consistent with grouped array selection. Two strategies have been used to study this form of object selection. One strategy is to manipulate whether stimulus objects, such as letters, are grouped together based on secondary features, such as a common color or direction of motion. For example, if observers are asked to attend and categorize a target letter presented at fixation, their responses are influenced by adjacent flanking letters. Flanking letters that are consistent with the response to the target decrease observers' response times, and flanking letters that are inconsistent with the response to the target letter increase response times. The effect of flankers depends on object grouping factors: Flankers that group with the target letter influence responses more strongly than flankers that do not group with the target (Fig. 4). Presumably, the gestalt grouping principles define a set of letters as a related group, and this group is attended as a single unit. The flankers within this single perceptual group then decrease or increase response times to the target letter. Grouping effects on targets and flankers have been reported with several gestalt principles, including similarity, common fate, connectedness, and good continuation.

A second strategy for studying object-based attention is to manipulate whether two attended features fall on the same object or on different objects or, alternatively, whether attention shifts within an object or across objects. One widely used paradigm is illustrated in Fig. 5. In this paradigm, observers view

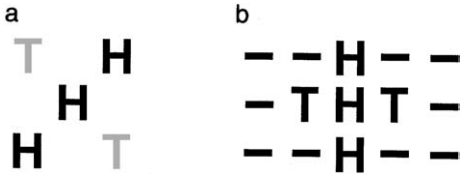


Figure 4 Displays used to study the influence of gestalt organization on visual attention. Observers report whether the central letter is an H or a T. (a) Grouping via similarity; the nontarget Hs group with the target H. Because the flankers that group with the target are compatible with one another, observers would classify the target letter quickly. (b) Grouping via good continuation; the nontarget Ts group with the target H. Because the grouped flankers are incompatible with the target H, observers would classify the target letter slowly.

two rectangles oriented either horizontally or vertically. One end of one of the rectangles is precued with a brief flash, and this precue is followed by a target that requires a keypress response. The target usually appears at the cued location; when it appears at an uncued location, it may appear within the same object as the cued location or in the other object. Both of these uncued locations are the same spatial distance from the

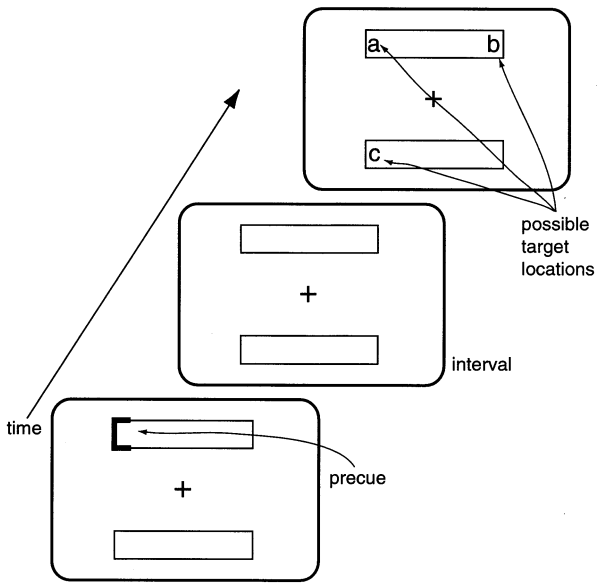


Figure 5 Spatial precuing task used to study object-based attention. Two rectangles appear, and the end of one is precued with a peripheral flash. After a delay, a target appears at one of three locations: a, a validly cued target; b, an invalidly cued target that appears in the cued object; c, an invalidly cued object that appears in the uncued object. Observers more quickly detect invalidly cued targets appearing in the cued rectangle faster than those appearing in the uncued rectangle.

precued region and the target is equally unlikely to appear at either of them, but observers are faster to respond to targets appearing in the uncued end of the cued rectangle than in either end of the uncued rectangle. Thus, attention appears to cover the entire cued rectangle even though only one end was cued. Similar results have been obtained with tasks that do not require spatial precues. In displays containing two rectangles that are overlapped to form an "X," observers are faster to discriminate features on the same rectangle than on different rectangles even though the spatial distances are similar.

Note that spatial location is centrally important in selection from a grouped array because this representation is spatiotopic. A handful of studies have demonstrated the importance of location by manipulating both grouping principles and spatial position. These studies have demonstrated that both grouping principles and spatial position influence attentional selection. In the cued detection task depicted in Fig. 5, moving the rectangles closer to one another reduces the cost of switching attention from the cued rectangle to the uncued rectangle, although attention continues to shift faster within an object than between objects. On the basis of such results, it may be possible to explain many demonstrations of "object-based attention" as occurring within a spatially formatted representation.

2. Object-Centered Selection

A second mechanism for object-based attention involves attending to objects that are defined in an object-centered reference frame, which represents the parts and features of an object in relation to a reference point on the object. Coding parts and features in reference to the object allows the relative positions of the parts and features to be constant as the object changes spatial position and retinal size. A person's head is above the torso irrespective of where the person appears (e.g., left or right visual field) or how distant the person is from the viewer.

Because an object-centered reference frame is relatively insensitive to spatial position, spatial manipulations should not influence this form of object-based selection. There is evidence for selection from a late, object-centered representation from a discrimination task that requires observers to focus attention on a single object or divide attention between two objects —a box and a line (Fig. 6). Observers report the values of two features, such as whether a gap is on the left or right side of a box. Sometimes, the two features are located on the same object (both features on the box) and sometimes on different objects (one feature on the box and one on the line). Performance is object-based in this task because features on the same object are reported more accurately than are features on different objects. However, unlike grouped array selection, performance in this task is not influenced by the spatial distance between the box and line, suggesting the objects are selected and not the locations occupied by the objects.

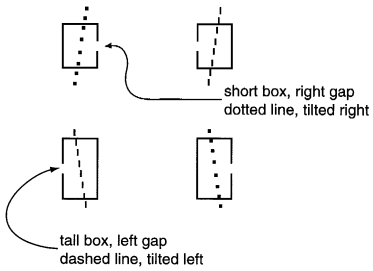


Figure 6 Object stimuli used to study object-based attention. Observers are more accurate in reporting attributes from the same object (e.g., box height and side of gap) than attributes from different objects (e.g., box height and tilt of line).

B. Neuropsychology of Object-Based Attention

1. Object-Centered Neglect

As with most neuropsychological studies of spatial attention, investigations of object-based selection have focused on patients with neglect following damage to the parietal lobe areas. Two key findings from neglect patients have implications for the neural basis of object selection. The first finding is that some of these patients exhibit object-centered neglect. The second finding is a hemispheric difference between object and spatial attention in patients with neglect.

Some patients with right parietal lobe damage (left neglect) neglect not only the left side of space but also the left side of an object, even when that object is located in the good (ipsilesional) visual field. Left neglect of objects can be observed when patients are asked to compare two novel objects and determine if they are the same or different. If the two objects are identical on their right sides but different on their left sides, neglect patients will incorrectly report that the objects are the same. Neglect patients continue to show this error when both objects are rotated 45° clockwise

to place the differing feature in the good right visual field. If neglect occurred for spatial coordinates only, rotating part of the objects into the nonneglected visual field should have allowed the patients to notice the difference between the two objects. Because the neglected region of the object follows the object as it is rotated, the neglect must be defined relative to a reference point on the object, such as the object's principal axis (i.e., midline).

Although object-centered neglect has been the focus of many recent studies, these effects are not present in all neglect patients, which raises two points. First, failure to find object-centered neglect in some patients may occur because visual neglect is not a unitary disorder. Second, and more interesting, object-centered neglect may be unobservable because of the stimulus objects used. Objects that do not posses a strong principal axis or do not have well-defined left and right sides (a "canonical handedness") may not have a single object-centered reference point, which could prevent object-centered neglect. For example, a symmetric letter such as "A", does not have a canonical handedness because the left and right sides are identical; there is no need for the visual system to represent the sides of a symmetric letter differently. Asymmetric letters (e.g., "F") have a canonical handedness, and the visual system must represent the differences between the two sides. Accordingly, object-centered neglect is less likely for symmetric letters than for asymmetric letters, suggesting that the object's handedness may influence the allocation of visual attention.

Finally, what type of object representation is involved in object-centered neglect—a grouped array or an object-centered representation? As noted previously, the common assumption is that object-centered neglect occurs within an object-centered reference frame. However, recent simulation results from connectionist models demonstrate that object-centered neglect could arise from a spatiotopic grouped array representation. These simulations indicate that perceptually organized input to a damaged attentional system may be sufficient to demonstrate object neglect; no object-centered coordinates, principle axis, or canonical handedness need to be computed to explain the findings from neglect patients.

2. Lateralization of Object and Spatial Attention

The second finding from neglect patients relevant to object attention is an apparent hemispheric difference between object-based attention and spatially based attention. Patients with damage to the right parietal lobe most often present with hemispatial neglect in which the left side of space is ignored. (Patients who demonstrate object-centered neglect also exhibit spatial neglect.) Left neglect can also be demonstrated in these patients by using object-based attention tasks. For example, in the task shown in Fig. 5, left neglect patients are slower to detect targets in the contralesional visual field than in the ipsilesional visual field. However, left neglect patients show preserved object-based attention: They detect targets in the cued rectangle faster than targets in the uncued rectangle, and this object-based effect is found in both the contralesional and ipsilesional visual fields. In contrast to patients with left neglect, patients with damage to the left parietal lobe (right neglect) appear to have deficits in object-based attention. Right neglect patients also show slower target detection in the contralesional visual field than in the ipsilesional field. However, these patients exhibit larger object attention effects in the contralesional field; it is more difficult for these patients to switch attention from the cued rectangle to the uncued rectangle in the contralesional field than in the ipsilesional field. These hemispheric differences between object-based and spatially based attention have been supported by research with a split-brain patient. In performing the cued detection task shown in Fig. 5, this patient had greater difficulty switching attention from the cued rectangle to the uncued rectangle when stimuli were presented in the right visual field (left hemisphere) than when stimuli were presented in the left visual field (right hemisphere).

3. Other Patient Groups

Finally, a few patient groups besides neglect patients have been studied to understand object-based attention. For example, studies with a visual form agnosic reported impaired object-based attention but intact spatial attention following diffuse damage to the occipital cortices. Patients with visual form agnosia fail to perceive objects because of damaged early level visual areas. Although these patients have intact sensory processes (e.g., acuity and color perception), they are unable to organize visual features using the gestalt principles. The damage to perceptual organization processes impairs the ability to form perceptual groups, which prevents any object-based component of visual selection. Despite having impaired object-based attention, spatial selection appears to be intact following diffuse occipital damage, suggesting that

object grouping processes appear to be dissociable from spatial selection processes. Thus, the parietal lobe attention system is not the only cortical system involved in mediating object selection.

C. Physiology of Object-Based Attention

In addition to the focus on parietal lobe involvement in neuropsychological studies of object selection, physiological studies have implicated other cortical areas in object attention. Multineuron recordings from macaque primary visual cortex (area V1) have implicated this area in object-based attention. Specifically, attending to one of two objects in a display results in enhanced firing for neurons whose receptive fields contain features of the attended object. Monkeys viewed scenes containing two objects (simple curves); one of the objects was connected to the fixation point, and monkeys were trained to attend to this object. The monkeys' task was to make an eye movement from the fixation point to the opposite end of the attended curve. Segments of the curves fell within receptive fields of V1 neurons. The neuronal responses were larger when a receptive field contained a segment of the attended curve than a segment of the unattended curve. This object-based attentional modulation in area V1 is important because previous studies of spatial attention were equivocal in finding V1 attentional modulation. Neurons in V1 exhibit attentional modulation when an object can act as the recipient or focus of attention; neurons in V1 do not appear to exhibit attentional modulation with blank displays or nonorganized cluttered displays. These neurophysiological findings are consistent with results from visual form agnosia, in which diffuse damage to early cortical visual areas (possibly V1) impaired object-based attention. Thus, there is growing evidence to support a central role for early cortical areas in perceptual organization and object attention.

Neurophysiological results have also suggested that object selection can occur in the oculomotor system as well as in purely sensory areas. Neurons in the supplementary eye field (SEF), located on the dorsomedial surface of the frontal lobes, appear to represent object-centered spatial selectivity for the direction of eye movements. That is, these neurons seem to code for spatial positions within an object, such as the left side of the object. Macaque monkeys were trained to make eye movements to the onset of a target spot. The target appeared in one of three conditions: alone in an otherwise blank display, at the left end of an object (a rectangle), or at the right end of an object. The absolute direction of the eye movement was identical in all three conditions; that is, the monkeys' eyes moved in exactly the same direction and same distance across these conditions. Although the eyes moved identically, a subset of SEF neurons fired at higher rates when eye movements were executed to a specific region of the object, regardless of its absolute spatial location. For example, some neurons responded vigorously to eye movements to the right side of the object; the same eye movement that landed on the left side of the object resulted in a smaller neuronal response. Thus, SEF neurons code for locations within an object; how these locations are coded—in a spatial reference frame such as the grouped array or in an object-centered coordinate frame—is unknown.

Finally, ERP studies with humans have investigated object-based selection. When viewing displays containing two superimposed surfaces (two transparent surfaces of different colored dots that rotate in opposite directions), observers can selectively attend to one of the two surfaces despite their spatial superimposition. If observers are instructed to attend to one of the surfaces, changes to the attended surface will produce evoked potential components with larger amplitudes than stimuli presented on the unattended surface. Specifically, changes on the attended surface generate larger P1 and N1 components compared to changes on the unattended surface. The similarity of these effects to the spatial attention effects described previously suggests that some object-based effects may be generated by neural processes shared with spatial attention. Attentional selection may be occurring from a grouped array in which motion segregation cues allow the two dot surfaces to be separated from one another in depth.

IV. ATTENTION TO OTHER MODALITIES

A. Selection beyond Vision

Beyond locations and objects in the visual modality, stimuli from other modalities can be selectively attended. This observation is evident from studies of selective attention in the 1950s and 1960s in which listeners attended one of two different speech signals arriving simultaneously in each ear. To determine how effectively attention could be restricted to one signal,

listeners were required to repeat (''shadow'') the speech in the attended channel. This shadowing procedure was the main behavioral technique for studying selective attention for decades. Studies of auditory attention supported early selection of stimuli; words spoken into the unattended ear are effectively filtered out. For example, if listeners are instructed to tap a key when they hear the word ''tap,'' they almost always tap when the word is presented to the attended ear and almost never tap when the word is presented to the unattended ear. Words presented to the unattended ear are filtered at an early level of processing prior to word identification; if filtering occurred later, after word identification, listeners should have tapped when the word tap was spoken in the unattended ear. However, the early attentional filter appears to be leaky in that salient material on the unattended channel, such as the listener's name, can deter attention from the attended channel.

An early locus for auditory selective attention has been confirmed with ERP studies. The general procedure is similar to that used to study spatial attention (described previously). Listeners hear a sequence of auditory tone pips, half presented to the left ear and half to the right ear. Listeners pay attention to the tones in one ear and press a button whenever an infrequent target tone is presented to the attended ear. Many studies using this procedure have found that the early ERP waves are larger for tones presented in the attended ear than in the ignored ear, consistent with a sensory gain effect of auditory attention. This enhancement of the voltage amplitude occurs temporally early, with the effect beginning as early as 20 msec after stimulus onset, well within the period of sensory-level processing. Moreover, the attentional enhancement was present for both targets and nontargets presented in the attended ear, indicating that selection occurred before the stimuli were identified. In addition, magnetoencephalographic studies have shown that these effects arise in or near primary auditory cortex. The results from auditory attention experiments are similar with results from visual attention experiments in suggesting an early selection account of attention in both modalities.

B. Cross-Modal Coordination of Attention

The existence of attentional selection in different modalities raises a basic question: Is there a single, supramodal attentional system that mediates selection across multiple modalities, or are there individual attentional systems for each modality that have some degree of cross talk with one another? Results from neuropsychological patients with neglect support a supramodal view of attention. Neglect patients have difficulty attending to both visual and auditory stimuli opposite the lesioned hemisphere, suggesting that parietal lobe attentional processes operate on a representation of space that codes both visual and auditory stimuli. Similar results have been reported from multimodal versions of the spatial precuing task. In the multimodal version of this task, neglect patients are asked to detect a lateralized visual stimulus. This visual target is preceded by a lateralized precue presented in either the visual or the auditory modality. Neglect patients show similar orienting behavior to both types of precues; specifically, neglect patients have difficulty disengaging attention from precues presented on the good (ipsilesional) side of peripheral space irrespective of the cue's modality. These results support a supramodal representation for parietal lobe attentional processes.

Further strengthening the supramodal view of attention, neglect patients also fail to attend to somatosensory stimuli presented on the bad (contralesional) side of peripheral space. However, somatosensory neglect could be explained by contralesional sensory deficits that accompany brain damage; inputs from the contralesional side of the body may be weaker than inputs from the ipsilesional side of the body. A convincing demonstration of somatosensory neglect was provided by testing neglect patients' ipsilesional hands, which have no sensory loss. The patients were touched simultaneously on the left and right sides of their right wrist and asked to report where they were touched. There were two key findings. First, left neglect patients failed to notice being touched on the left side of their right wrists, demonstrating that somatosensory neglect can occur in the absence of sensory deficits. Second, the attentional impairment for detecting contralesional touches occurred whether patients' hands were facing with the palm downward or facing with the palm upward. The neglect did not follow the rotation of the hands but instead remained fixed on the left side of the wrist. These results indicate that the attentional deficit did not occur in somatosensory coordinates, such as always neglecting the thumb side of the wrist or always neglecting the left side of the body. Instead, neglect occurred in abstract spatial coordinates that represent the left and right sides of a limb independent of that limb's orientation. This ''limb-centered'' somatosensory neglect bears a

striking similarity to the object-centered neglect observed in the visual modality: The neglected region is not necessarily defined by the patients' midline but by the midline of a stimulus (an object or a limb).

Finally, both behavioral and ERP studies with neurologically normal observers have extended the operation of selective attention across different modalities. There appear to be strong cross-modal links in spatial attention that allow observers to spatially attend a region and select stimuli occurring in different modalities. In behavioral studies, observers visually cued to one side of space detect more quickly both visual and auditory targets on the cued side of space than on the uncued side of space; this cued location advantage occurs when the modality of the target is unpredictable. Thus, spatial attention appears to operate across modalities. Behavioral studies have also attempted to decouple attentional shifts across modalities by varying the occurrences of auditory or visual targets. Can auditory spatial attention be allocated to a different region than visual spatial attention if an auditory target is expected at one location and a visual target is expected at another location? The answer to this question seems to be "no." If observers voluntarily shift auditory spatial attention to one region to detect a highly probable auditory target, visual spatial attention seems to follow. The same result holds for shifts of visual spatial attention; auditory attention will follow visual attention when visual attention is voluntarily shifted to a region to detect a highly probable visual target. ERP studies have reported similar results. For example, if observers are required to monitor a location for an infrequently occurring visual target, larger evoked responses are generated by both visual and auditory nontarget stimuli at the attended location than at the unattended location.

One caveat from the cross-modal studies of attention is that although visual and auditory attention appear to operate in concert, the effects across modalities are not equivalent to effects within a single modality. Larger attentional effects are found when the cue and target appear in the same modality than in different modalities. Such an observation cannot be explained with a strong supramodal account of spatial attention. The strong supramodal account would predict that each sensory modality could cause an equivalent shift of spatial attention, which would provide enhanced processing at the attended location across all modalities. However, cross-modal studies cannot distinguish between linked unimodal attention systems and a supramodal system that receives differentially weighted inputs from each modality. Furthermore, a supramodal system may be an emergent property of the connections that link modality-specific attentional systems. Consistent with the multiple attentional systems in the visual modality, there are likely to be both unimodal and supramodal attentional mechanisms for the coordination of attention across sensory modalities.

V. ATTENTION TO TASKS

A. Performing Multiple Tasks Simultaneously

In addition to coordinating the processing of sensory stimuli, attention must participate in the coordination of tasks. There are many situations in which humans must perform multiple tasks concurrently, as such having a conversation while driving. Performance is impaired when multiple tasks are performed concurrently, even when the tasks are highly practiced, as with driving and speaking. Attentional processes may be involved in selecting an individual task for current behavior, with a cost in performance occurring when attention is either divided or switched between multiple tasks.

The biased competition model we used to describe spatial attention can also be applied to performing multiple tasks such as the Stroop task, in which different tasks must be performed in different blocks of trials. In the Stroop task, observers view words that name colors (e.g., "red" or "blue") presented in different colors of ink. The words can be written in a compatible ink color ("red" written in red ink) or in incompatible colors ("red" written in blue ink). Observers perform one of two tasks—either reading the word or naming the ink color. Because word reading is more practiced than ink color naming, observers can read words with little effect of the ink color; observers can just as quickly read color words printed in a compatible ink color as color words printed in an incompatible ink color. In contrast, color naming is highly influenced by the word; observers are slower to name ink colors used to print an incompatible word ("red" printed in blue ink) than ink colors used to print a compatible word ("red" printed in red ink). In the Stroop task, the top-down task demands (i.e., the task observers are asked to perform) and the bottom-up stimulus features (the word and ink color) both guide behavior. The color naming task is difficult because the bottom-up inputs are stronger for words

than for ink colors. Thus, there is a stronger bias for word reading than for color naming, allowing the word reading task to compete more effectively for the control of behavior.

One difficulty in using the Stroop task to study attention to tasks is that one of the tasks, word reading, is much easier than the other. An attentional phenomenon that demonstrates competition between two equally difficult tasks is the "attentional blink." In the attentional blink task, observers view a stream of approximately 20 stimuli presented one at a time at a rate of about 10 stimuli per second; observers are asked to detect two targets from this stream. For example, the first target (T1) may be a number that observers must classify as even or odd, and the second target (T2) could be a letter that observers must classify as a consonant or vowel. Observers make both responses at the end of the stimulus stream. Observers often fail to identify T2 if it appears shortly after T1; if T2 appears somewhat later, observers more accurately report its identity. The temporary impairment in identifying T2 is referred to as the attentional blink because it is similar to the consequence of a T1-triggered eyeblink (i.e., a brief period during which subsequent targets cannot be detected).

The attentional blink does not appear to be caused by sensory-level interference between the two targets. Instead, the failure in reporting T2 is more central, resulting from a failure to store T2 in a durable form that can be reported at the end of the stimulus stream. In terms of a biased competition account, top-down task constraints related to the T1 task may bias this item to be coded into working memory. After T1 begins to be encoded into working memory, the task bias can begin to switch to the T2 task, but this reconfiguration takes time. Thus, if T2 appears soon after T1, it may not be efficiently encoded into working memory and may be overwritten by the next item in the input sequence. At the end of the stimulus sequence, reporting T2 is difficult because of this shallow encoding into working memory—T2 is identified but not reported, suggesting that selection may operate relatively late in the visual processing stream (i.e., after stimulus identification).

B. Neural Mechanisms

1. Attentional Blink and Neglect

As with almost every other type of attentional selection, the attentional blink appears to involve the parietal lobe attention system. Patients with damage to the right parietal lobe show a prolonged attentional blink. Compared to normal observers, these patients need longer intervals between T1 and T2 or T2 will be missed. This suggests that neglect is not only a disorder of visuospatial attention but also reflects more general attentional processes. Thus, the parietal lobe attention system may involve several types of attentional "gating," ranging from sensory gating to task gating, which allows the parietal lobes to participate in both early (sensory) selection and late (task) selection.

2. Human Electrophysiology of the Attentional Blink

Perhaps more than any other methodology, ERP studies of the attentional blink have elucidated the mechanisms of this form of selection. Recent ERP studies have demonstrated that the attentional blink occurs late in visual processing, after stimulus identification, consistent with the view that failure to recognize T2 is due to poor encoding of this item into working memory.

To determine if the attentional blink involves sensory-level attention processes, early ERP components were studied. Observers participated in a standard attentional blink task, but a task-irrelevant probe was presented simultaneously with T2. This irrelevant probe, a flashed visual stimulus, elicited the P1 and N1 components that are involved in spatial selection. If the attentional blink is due to suppressed sensory processing, P1 and N1 components should also be suppressed for irrelevant probes presented during the window of the attentional blink. However, no suppression of the P1 and N1 waves was observed, even though behavioral performance exhibited a strong attentional blink. There was no evidence for sensory suppression during the attentional blink period, indicating that the mechanisms of spatial attention are different from the mechanisms of the attentional blink.

Although early sensory processes measured by P1 and N1 are not responsible for the attentional blink, later perceptual processes may be. An additional experiment addressed whether the T2 item was recognized completely; complete recognition of the T2 item would indicate that the attentional blink was not due to any form of perceptual suppression, early or late. This study investigated the N400 ERP component. The N400 is typically elicited by words that mismatch a previously established semantic context. For example, an N400 would be generated by the last word in the sentence "The man drank his coffee with cream and

dog'' because ''dog'' is inconsistent with the semantic meaning of the sentence. If an N400 is elicited by a word, then the word must have been identified; otherwise, a semantic comparison with the context would have been impossible.

In the N400 attentional blink study, the T2 stimulus was a word. Each trial began with a semantic context word, and the T2 word was either semantically related or unrelated to this semantic context word. When the T2 word semantically mismatched the context word, an N400 component was produced, even if the T2 word occurred during the window of the attentional blink such that observers could not report the word. This result indicates that the T2 word was identified during the attentional blink window, and the failure of observers to correctly report the word reflects a failure to store the word in working memory. Thus, the attentional blink occurs after identification, consistent with a late-selection account of attention.

VI. SUMMARY

Attention is necessary for eliminating sensory inputs or behavioral tasks that are irrelevant at a specified time. Although there may be multiple forms of attention, some neural sites, such as the parietal lobe, are involved in many aspects of attention. The control and effects of these multiple forms of selection are quite varied, however. The integration of top-down behavioral tasks or goals and bottom-up stimulus factors may allow the attentional system to be highly flexible and allow a small subset of components to have varied effects.

See Also the Following Articles

CONSCIOUSNESS • INFORMATION PROCESSING • MEMORY, NEUROIMAGING • OBJECT PERCEPTION • PATTERN RECOGNITION • SHORT-TERM MEMORY • SPATIAL COGNITION • VIGILANCE • VISUAL SYSTEM DEVELOPMENT AND NEURAL ACTIVITY

Suggested Reading

Allport, A. (1993). Attention and control: Have we been asking the wrong questions? A critical review of twenty-five years. In *Attention and Performance XIV* (D. E. Meyer and S. Kornblum, Eds.), pp. 183–218. MIT Press, Cambridge, MA.

Colby, C. L., and Goldberg, M. E. (1999). Space and attention in parietal cortex. *Annu. Rev. Neurosci.* **22**, 319–349.

Desimone, R., and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* **18**, 193–222.

Mozer, M. C. (1999). Explaining object-based deficits in unilateral neglect with out object-based frames of reference. In *Disorders of Brain, Behavior, and Cognition: The Neurocomputational Perspective*. (J. Reggia, E. Ruppin, and D. Glanzman, Eds.), pp. 99–119. Elsevier, New York.

Pashler, H. (1998). *Attention*. Psychology Press, Hove, UK.

Auditory Agnosia

HENRY A. BUCHTEL
University of Michigan and VA Healthcare System, Ann Arbor

GLOSSARY

apperceptive agnosia Disturbance of the basic analysis and elaboration of sound energy arriving at the ear and up through sensory pathways.

associative agnosia Loss of the ability to link a stimulus to meaningful associations (meaning).

audiogram A graphical representation of a person's ability to hear sounds of different frequencies.

pure word deafness A form of auditory agnosia in which the disorder is limited to the understanding of spoken language.

Auditory agnosia refers to the loss of the normal ability to derive the meaning from an auditory stimulus. Careful testing is needed before the cause of the agnosia can be determined.

I. OVERVIEW

Auditory agnosia is a rare condition in which the individual has an isolated deficiency in comprehending sounds despite a normal or nearly normal audiogram. The patient is usually unable to recognize both spoken words and environmental sounds. If only the comprehension of spoken language is affected, the disorder is called *verbal auditory agnosia* or *pure word deafness*. If only environmental sounds are affected, the person is said to have *nonverbal auditory agnosia*. Typically, these patients are able to read, write, converse intelligently, and name objects. They can also use prosodic information (e.g., the melody or intonation in a speaker's voice) to identify emotional content and even syntax (e.g., that a person has asked a question rather than having made a statement).

II. VARIETIES OF AUDITORY AGNOSIA

There are two major classes of auditory agnosia. If the disorder affects only the comprehension of spoken language, it is called *pure word deafness*. If the person has difficulty only with environmental sounds, the condition is called nonverbal auditory agnosia.

III. APPERCEPTIVE DEFICITS

If the person with auditory agnosia is unable to extract meaning from a stimulus because the initial analysis (perception) of the stimulus is flawed, then the higher cognitive mechanisms that would ordinarily assign meaning will not have sufficient information to perform adequately. Thus, central mechanisms for extracting meaning could be intact but unable to carry out their functions because the information that they receive is defective. An analogy might be a noisy telephone line. A person trying to understand speech coming over the line would not have a sufficiently clear signal to allow the information transfer to occur.

IV. ASSOCIATIVE DEFICITS

If the initial analysis of the stimulus appears to have been carried out normally and the person is still unable to extract meaning, then it is likely that the person is suffering from an associative disorder. What is lacking is a link between the information and stored knowledge concerning its meaning. The patient's perceptual experience might be similar to that of a person who had previously known a foreign language well but who has now forgotten the meaning of a particular word. Such a person would be able to hear the word clearly and may even recognize that it represented a meaningful collection of sounds, but its meaning would be lost to the listener.

V. NEUROLOGICAL SUBSTRATE OF THE DISORDER

The damage responsible for these disorders is usually caused by cerebrovascular disease, usually embolic stroke, that involves the midportion of the first temporal gyrus bilaterally (at least one case has been reportedly caused by a hemorrhage in an auditory structure in the brain stem, the inferior colliculi). If caused by unilateral damage, the lesion is usually deep in the posterior temporal lobe of the hemisphere dominant for speech (usually the left hemisphere). In such cases, there may be signs that any residual abilities that are present depend on right hemispheric linguistic abilities. For example, the person may report hearing a word that is semantically related to the actual word spoken (''horse'' for ''pony''). Some authors have suggested that damage to the primary auditory cortex produces pure word deafness (agnosia for spoken language only), whereas damage to the auditory association cortex produces difficulties with nonspeech sounds. Damage to both regions would produce global auditory agnosia. The pattern of disabilities in auditory agnosia differs from that of the more common Wernicke aphasia in that the receptive language disturbance is limited to the auditory domain while the patient's spoken language is relatively good. In early accounts of the syndrome, a normal audiogram led to the assumption that basic auditory analyses (apperception) were normal and that the disorder arose because of an inability to link the sound to its meaning (association). This interpretation can still be found in current texts and the syndrome is frequently attributed to a disconnection of auditory cortex from Wernicke's area, with the possible implication that both are capable of normal functioning.

VI. BEHAVIORAL ASPECTS OF AUDITORY AGNOSIA

Although apperceptive functions are usually assumed to be normal in individuals with auditory agnosia, behavioral evidence has been accumulating that the audiogram in patients with auditory agnosia is not an adequate measure of their ability to analyze the kinds of sounds critical for speech. With more sophisticated auditory tests, several abnormalities have emerged. For instance, many such patients have difficulty determining the temporal order of sounds and they may require a greater than normal sound intensity to make very brief sounds audible. Some but not all patients also have difficulty discriminating the loudness of stimuli and their location in space. Disturbances of temporal resolution (order of sounds) and two-click discrimination have also been described. To illustrate some of these associated hearing deficits, the following case is one that might previously have been judged to exemplify an associative auditory agnosia.

A. Case LD

At the age of 49 years, this individual suffered a cerebral stroke in the distribution of the left middle cerebral artery. His right leg and arm were partially paralyzed and his speech was moderately disturbed. The right-sided weakness resolved within several days but the speech deficits slowly resolved over the next year, at which time his naming and ability to understand spoken and written words were normal and he returned to work. About 6 months later, he suffered a second stroke, this time in the right cerebral hemisphere. This stroke evolved over the next 48 hr. Initially, he was globally aphasic and could neither understand commands nor repeat simple sentences. Spontaneous speech was occasionally clear but interspersed with jargon. The motor system appeared to be relatively unaffected apart from an extensor left plantar response, but over the next 24 hr a dense left-sided hemiparesis developed, which resolved by 48 hr. His language disturbances persisted. A computerized tomography scan 3 months later indicated that the infarction had been in the right posterior temporal region. When examined 6 months after the second stroke, the only remaining physical abnormalities were a mild left lower facial weakness, mild left hyperreflexia, and extinction of pin prick in the right face, arm,

and leg with bilateral simultaneous stimulation. Neuropsychological testing at this time demonstrated that his IQ was in the average range. His spoken language was generally correct in content but had a loud, explosive quality. Despite a relatively normal audiogram, he was severely deficient in understanding spoken language and environmental sounds. In fact, in the absence of contextual cues he was never observed to understand a single spoken word. He was also unable to repeat words, letters, numbers, phonemes, or nonsense sounds. He did only slightly better than chance at determining whether two letters spoken successively were the same or different. In contrast, he was able to point correctly to a picture in an array when a word for one of the pictures was spoken. If he heard the name of the object, he pointed correctly to its picture and said the word. If he heard an associated word (e.g., "gambling" when confronted with an array of three objects including a picture of dice), he said the word of the picture while pointing to it ("dice") but never uttered the semantically related word.

The patient gave no indication that his hearing was defective, consistent with the audiogram, and he claimed that he could hear all the sounds presented to him. However, several subtle lower level abnormalities emerged during testing that could account for his auditory agnosia. First, reaction times to brief (30-msec) suprathreshold tones were very slow (averaging about 380 msec), whereas reaction times to longer (100-msec) tones were within the normal range for his age group (averaging about 250 msec). Reaction times to visual stimuli were also in the 250-msec range for both 30- and 100-msec stimulus durations (Fig. 1). Since consonants in speech have durations in the 30- to 40-msec range, slow processing of such sounds could clearly lead to problems understanding speech. Second, the patient required almost a quarter of a second of silence between two brief (30-msec) tones in order to be hear them as separate (Fig. 2). This suggests that the sound of the first tone persisted in his auditory system long after its offset. If so, the resulting "avalanching" of sounds while listening to someone speak would clearly interfere with comprehension.

The previous case is an example of auditory agnosia in which the loss of comprehension may have had an apperceptive rather than associative origin. In this patient the disorder was probably the consequence, at least in part, of a very slow analysis of brief sounds and an independent or derivative slow decay of sounds in the auditory system. Because only supplementary tests demonstrated the possible apperceptive nature of his deficit, it is possible that many other cases of auditory agnosia that were attributed to an associative deficit may also have involved a disruption of lower level language decoding functions.

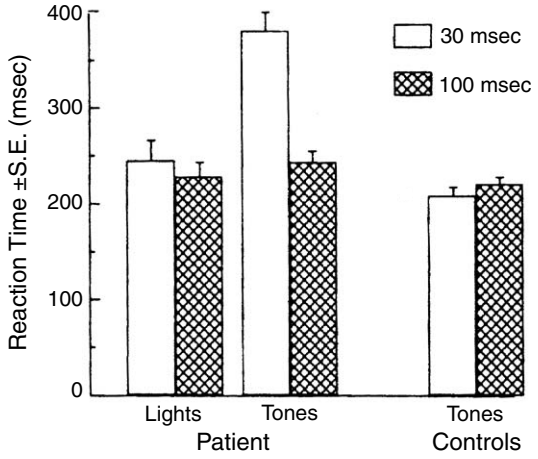


Figure 1 Simple reaction times to tone and light stimuli for the patient (left) and for tones for control subjects at the same intensity above threshold (right). Reaction times were slow for the patient when 30-msec tones were used, whereas changing the duration of light stimuli had only a minor effect. Control subjects have similar reaction times for both 30- and 100-msec tone durations. [From Buchtel and Stewart (1989). Auditory agnosia: Apperceptive or associative disorder? *Brain Language* **37,** 12–25].

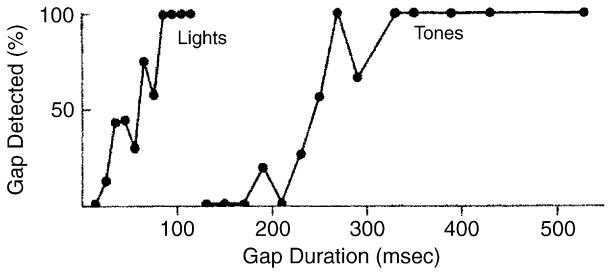


Figure 2 Gap detection: The patient needed almost 300 msec between two 30-msec tones before he could reliably tell that there were two stimuli rather than one. Sensitivity to successive light stimuli was normal. [From Buchtel and Stewart (1989). Auditory agnosia: Apperceptive or associative disorder? *Brain Language* **37,** 12–25].

See Also the Following Articles

AUDITORY CORTEX • AUDITORY PERCEPTION • HEARING • LANGUAGE AND LEXICAL PROCESSING • LANGUAGE DISORDERS • SENSORY DEPRIVATION • SPEECH • TEMPORAL LOBES

Suggested Reading

Buchtel, H. A., and Stewart, J. D. (1989). Auditory agnosia: Apperceptive or associative disorder? *Brain Language* **37,** 12–25.

Auditory Cortex

ROBERT J. ZATORRE
McGill University

GLOSSARY

aphasia A loss or disruption of normal language ability resulting from cerebral damage.

binaural Pertaining to processing of sound combinations coming from the two ears.

formant A resonant frequency band in the acoustic spectrum.

formant transition A change in formant frequency, usually related to speech in which rapid changes in the position of the vocal articulators result in such transitions.

Heschl's gyrus Cortical region in the superior temporal gyrus that contains primary auditory cortex.

planum temporale Cortical area located on the superior temporal gyrus behind Heschl's gyrus.

medial geniculate nucleus Principal thalamic nucleus for the auditory system.

tonotopy The systematic topographical arrangement of neurons sensitive to particular acoustic frequencies.

Wernicke's area A region in the left posterior temporoparietal cortex that plays a role in speech and language processing.

The human auditory cortex may be defined as those regions of the neocortex, situated mostly on the superior and lateral surfaces of the temporal lobe, that contain neurons responsible for processing sound. This article describes the anatomy and physiology of this area, with emphasis on properties that are specifically related to human auditory processing. The role of the auditory cortex in behavior will be considered, particularly in relation to the analysis of speech sounds. Related issues concerning morphological and functional hemispheric asymmetries will also be discussed.

I. ANATOMY

A. Gross Morphology

The auditory cortex consists of several distinct fields that differ in terms of their cytoarchitecture, connectivity, and physiological response properties. In primates, three major subdivisions may be identified, which are usually referred to as core, belt, and parabelt regions. The correspondences and homologies between these regions in human and other primates have not been completely worked out. However, it is well established that in humans the primary auditory cortex is located within Heschl's gyrus (HG), a relatively well-defined gyrus located transversely on the superior temporal gyrus, deep within the Sylvian fissure (Figs. 1 and 2). It is not uncommon for more than one HG to be present in one or both hemispheres, but there is no consistent pattern of asymmetry, as had been once thought. The right HG is shifted anteriorly by approximately 6 mm compared to the left. The primary area is confined mostly to the medial half to two-thirds of HG (the first HG when there is more than one), although in some individuals it may extend into Heschl's sulcus, which forms the posterior border of HG.

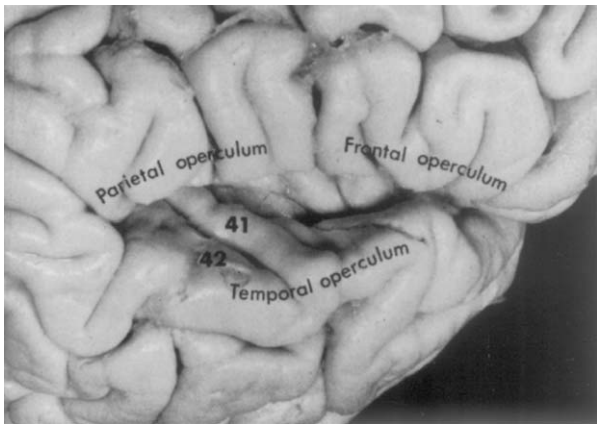


Figure 1 Anatomical specimen showing a lateral view of the right cerebral hemisphere with the lateral (Sylvian) fissure opened to reveal the hidden surface of the superior temporal gyrus. Heschl's gyrus is visible as the prominent transverse convolution, marked 41. The planum temporale is located posterior to Heschl's gyrus and is marked 42. Numbers correspond to cytoarchitectonic definitions according to Brodmann (reproduced courtesy of Charles C. Thomas, Publisher, Ltd., Springfield, Illinois from E. A. Kahn *et al*., 1969, *Correlative Neurosurgery*).

The precise boundaries of the belt and parabelt fields of the human auditory cortex have not been established, but it appears that auditory regions extend from HG both posteriorly and anteriorly along the superior temporal gyrus (STG) as well as onto the lateral aspect of the gyrus and into the superior temporal sulcus, as occurs in the macaque (Fig. 3). There is also evidence that auditory cortex may be found within the temporoparietal opercular area and even in some portions of the posterior insula bordering the medial portion of HG. However, there are no clearly established sulcal or gyral boundaries that correspond to these fields.

Posterior to HG is the planum temporale (PT), an approximately triangular region along the STG, bordering the Sylvian fissure (Fig. 1). This area, defined on the basis of its gross morphology, has been proposed to be important in language processes. Anterior to HG along the STG is an area sometimes referred to as the planum polare, which is likely to contain unimodal auditory cortex as well. The lateral aspect of the STG and the upper bank of the superior temporal sulcus may be considered as part of the parabelt auditory cortex as well. Beyond these regions lie areas that may have multimodal functions, notably cortex within the lower bank of the superior temporal sulcus and/or the middle temporal gyrus that may integrate auditory and visual information.

B. Cytoarchitecture

The core auditory region within HG, which in turn may be further subdivided, consists of koniocortex. Its laminar organization is characteristic of primary regions in other modalities; it is granular, with a thick, well-defined layer IV containing pyramidal cells. It is also densely myelinated and has high reactivity to cytochrome oxidase and acetylcholinesterase. This area is approximately coextensive with area 41 in Brodmann's cytoarchitectonic map of the human brain and with area TC according to the parcellation of von Economo and Koskinas. It is also sometimes referred to as A1 in the neurophysiological literature.

The cytoarchitectonic organization of the rest of the human auditory cortex is not well known, but studies in monkeys have helped to provide a model to which human data may be compared. The core, belt, and parabelt areas mentioned previously have been best defined in the macaque (Fig. 3). In that species the core region may be further subdivided into at least two and likely three separate areas based on the slightly different characteristics of their neuronal organization as well as on physiological response properties. It is likely, but not proven, that a similar organization exists in the human brain.

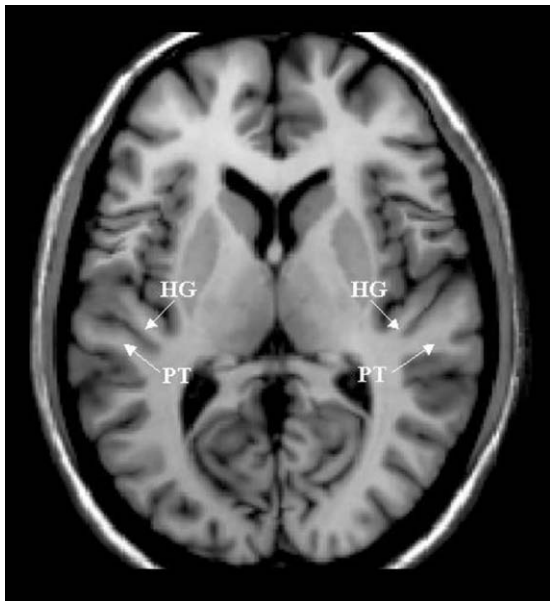


Figure 2 Magnetic resonance image of a normal human brain showing a horizontal section through the region of Heschl's gyrus (HG) and the planum temporale (PT).

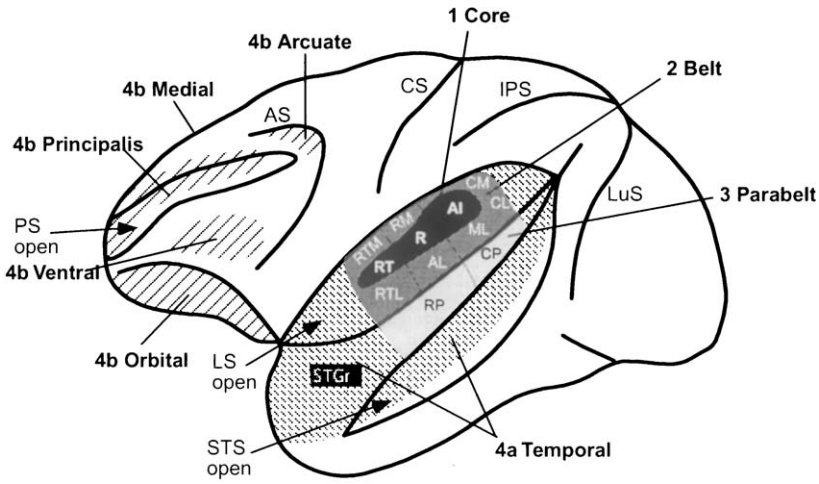


Figure 3 Levels and regions of auditory cortical processing in the macaque monkey. The areas shown are located on the superior temporal gyrus and in the depths of the superior temporal sulcus (STS), which has been opened to show the extension of auditory-related cortex into this sulcus. Level 1 represents the core (darkest shading) and includes tonopically organized regions A1, R, and RT; level 2 represents the belt areas (moderate shading) that encircle the core; level 3 represents the parabelt areas RP and CP (light shading), found in the upper bank of the STS; level 4a (stippling) represents putative auditory cortical areas outside the parabelt region, including the rostral area STGr; and level 4b (light hatching) represents frontal lobe cortex involved in auditory processing (reproduced with permission from Kaas *et al.*, 1999).

C. Connectivity

The input to the auditory cortex in the macaque comes from several subdivisions of the medial geniculate nucleus (MGN), which in turn receives the most important auditory projections from midbrain and brain stem auditory nuclei. The ventral subdivision of the MGN sends projections to all core areas, whereas the belt and parabelt areas receive primary projections from the dorsal and medial divisions of the MGN.

Corticocortical connections exist within and between the core, belt, and parabelt regions in a topographic and hierarchical arrangement. Neurons that respond to specific frequencies tend to be interconnected most heavily with neurons of the same frequency selectivity in other cortical fields. Removal of the A1 core region does not abolish responses in the more rostral core area (field R) in the macaque, indicating that parallel inputs exist to this region. However, belt regions immediately posterior to the core areas in the macaque appear to receive important inputs from the core area since excision of the core region results in abolition of responses to pure tones. Other, more anterior and lateral belt areas do continue to respond to sounds, however, even after removal of the primary region, as shown by evoked-potential responses recorded in human epilepsy patients.

The parabelt area is closely connected to the belt region but receives few if any direct inputs from the core areas. In addition to its thalamic inputs from the dorsal and medial portions of the MGN, the parabelt area receives inputs from other nuclei, including the pulvinar. Although the connectivity of the human auditory cortex is not well-known, it is note worthy that depth-electrode stimulation in human neurosurgical patients indicates that unidirectional connections exist from primary regions to more lateral and posterior areas, in accord with the corticocortical patterns described for monkeys.

Reciprocal cortical connections also exist between regions in the two hemispheres. The principal interhemispheric fibers pass through the midregion of the corpus callosum. Two features of this connectivity are worth noting. First, as with intrahemispheric connections, the projections are topographically organized, with neurons of the same frequency selectivity most highly interconnected. Second, the binaural response properties are important determinants of interhemispheric connectivity: Neurons that respond to stimulation of either ear tend to be heavily interconnected, whereas those units that are inhibited by input to one ear but not the other tend to have sparse interhemispheric projections.

Connections between the auditory cortex and several other cortical and subcortical structures have also been explored, but again little is known about these connections specifically in the human brain. In monkeys, there are important efferent connections between auditory regions and midbrain nuclei, particularly the inferior colliculus, as well as the basal ganglia.

Connections between auditory regions and the frontal lobe are topographically organized, with posterior regions of the STG projecting primarily to dorsolateral frontal cortices, whereas anterior and ventral auditory areas project preferentially to ventrolateral and orbital regions of the frontal lobe. Similar fibers have been described in gross anatomical studies in humans; one of these, the arcuate fasciculus, interconnects posterior STG regions with frontal cortices in and around Broca's area and is thought to play an important role in speech perception and production.

II. PHYSIOLOGY

The vast majority of information concerning neurophysiology of the auditory cortex derives from animal studies. To what extent these findings apply directly to humans is not known; moreover, human auditory cortical neurons may exhibit properties that are not found in other species, considering humans' specific ability to process complex sounds such as speech and music. Nevertheless, it is likely that much of the basic organization of the auditory cortex is at least similar to a first approximation between monkeys and humans.

A. Tonotopy

Tonotopic organization refers to the systematic topographical arrangement of neurons as a function of their response to tones of different frequencies. Tonotopy is a feature of the cortex but is also found throughout the auditory neuraxis, from the most peripheral level, at the basilar membrane of the cochlea as well as throughout most brain stem nuclei, midbrain, and thalamic levels. In primary auditory cortex, A1, single-unit recordings indicate that the majority of neurons demonstrate frequency selectivity, which is typically measured by establishing frequency threshold tuning curves for different frequencies of stimulation. The frequency for which a given neuron responds at the lowest intensity level is referred to as the characteristic frequency or best frequency. These best frequencies have been shown to be organized along a linear dimension of A1. In the human auditory cortex, it has been well established with electrophysiological and functional imaging techniques that high frequencies are represented in the most medial portion of HG, with lower frequencies represented more laterally along the gyrus. In monkeys, the more anterior field R is also known to exhibit tonotopy, but its frequency representation is reversed with respect to A1 so that the boundary between the two corresponds to low frequencies. This arrangement is similar to that of visual cortical regions that are adjacent to one another, which show similar reversals in the retinotopic organization of visual input.

B. Complex Response Properties

Although neurons in A1 typically show clear frequency selectivity, many of these neurons display more complex properties. Even neurons with the same best frequency may display different tuning functions. For example, some have very wide receptive fields, meaning that they respond well to tones over a wide frequency range, whereas others are more narrowly tuned. Still others have multiple best frequencies so that they respond strongly to two separate frequencies but not necessarily to frequencies in between. These features are thought to arise from interactions with convergent excitatory and inhibitory inputs from various afferent sources, as described previously. Thus, unlike auditory nerve fibers, whose tuning curves primarily reflect the mechanical tuning properties of the basilar membrane, cortical neurons display properties that are the result of neural interactions and thus represent higher order processing.

Beyond the core regions, the degree of tonotopic organization appears to break down. Some of the cortical fields adjacent to A1 have some tonotopic organization, but in belt and parabelt areas this organization is either absent or only weakly present. Neurons in these areas tend to have complex response properties and are generally not well characterized by simple frequency tuning curves. In the lateral belt areas, neurons tend to discharge with higher rates to narrowband noise bursts than to pure tones, in contrast with core area neurons whose discharge rates are typically highest for pure tones. In addition, the lateral belt neurons tend to have preferred bandwidths; that is, they respond best to sounds that cover a particular frequency range and less well to sounds that cover greater or lesser ranges. These properties suggest that neural integration over specific frequency ranges is occurring, most likely as a result of converging inputs from neurons in the core areas.

In several animal species, it has been shown that many auditory neurons respond not only to frequency information but also to temporal properties of the

stimulus. There are neurons that respond primarily with excitation to the onset of the stimulus but are insensitive to its offset or vice versa. Other neurons have more sustained responses throughout the period of stimulaion, and still others respond with inhibition or with a combination of these features. Some neurons have been described in unanesthetized monkeys with very complex receptive field properties that take into account both spectral and temporal information. For example, there may be excitation to a particular frequency region during some time period, whereas another frequency region elicits inhibition for a different time period.

There are also neurons sensitive to frequency modulation. These changes in frequency that can be either monotonic (e.g., frequency glides, where the frequency changes in a consistent direction) or periodic (when the frequency moves up and down at regular time intervals, such as in vibrato). Also, neural responses that follow amplitude modulation, or changes to the amplitude envelope of a sound, have been described. These changes may also be of a periodic nature (e.g., in a tremolo), or responses may occur to abrupt changes in envelope, such as occur when an object strikes another object. These abrupt changes are also found in human consonant speech sounds. Because of their dynamic response properties, it is possible that these types of neurons may play a role in processing complex sounds such as vocalizations, which contain time-varying energy distributed in distinct spectral bands (formants). It is highly likely that similar neural responses occur in humans for processing of speech sounds.

In human studies, the distinction between primary and extraprimary fields, as determined from anatomical data such as that reviewed previously, is generally supported by physiological data. For example, the core region located in HG is seen to respond with the shortest latencies to clicks in depth-electrode recordings, whereas surrounding areas have longer latencies, consistent with the idea that information is passed from one region to the next in a hierarchical arrangement. Also, functional imaging shows that the core areas respond well to stimuli such as noise bursts, whereas the surrounding regions generally respond to stimuli with more complex properties. Also relevant are cortical stimulation studies in which the cortex is stimulated while patients undergo neurosurgery. These observations have shown that stimulation of HG generally leads to nonspecific percepts, such as buzzing or hissing. In contrast, stimulation of anterior and lateral STG regions sometimes results in complex auditory hallucinations. In some cases, subjects report hearing music, speech, or voices quite vividly. These observations are consistent with the idea of a hierarchical arrangement between primary and extraprimary areas.

C. Binaural Properties

Many auditory cortical neurons are sensitive to differences in the way a sound reaches the two ears. These interaural differences are primarily of two types: intensity and time of arrival. Both these properties arise as a consequence of the distribution of sounds in space and the acoustics of how such sounds arrive at the two ears. Consequently, these differences are the cues used by the auditory nervous system to compute the spatial position of sounds. Physiological responses that correspond to these two features have been described in several different locations in the auditory core, belt, and parabelt regions. Many cortical neurons are very sensitive to slight differences in interaural intensity, whereas others respond differentially to interaural timing differences. These neurons tend to respond preferentially to sound sources located in the contralateral spatial hemifield, although there are also many units that show broad spatial tuning, extending into both hemifields.

Although there is evidence that neurons sensitive to similar spatial positions are located within cortical columns, there appears to be no topographic representation of space in the auditory cortex. Thus, unlike the visual system in which retinal position, and hence space, is coded by retinotopic mechanisms, the representation of auditory space is not topographically organized. However, recent evidence suggests that auditory space may be represented by population codes based on neural activity rates. Cortical neurons are also sensitive to auditory motion and sometimes respond preferentially to one direction of motion in the azimuthal plane.

III. BEHAVIORAL ASPECTS

The anatomy and physiology of the auditory cortex set the stage for understanding its role in behavior. This role has been traditionally elucidated by studying the effects of cortical damage on various perceptual tasks, either in experimental animals or in patients with brain lesions. Recently, this information has been

supplemented by the development of functional neuroimaging techniques, such as positron emission tomography, functional magnetic resonance imaging (fMRI), and magnetoencephalography.

A. Cortical Deafness

It has been known since at least the late 19th century that lesions to the superior temporal region in dogs, monkeys, and other species entail disorders of auditory perception. Early studies of dogs with cortical damage to this area showed, for example, that they failed to orient to sounds or respond to them normally. However, it was evident even from these early observations that such animals were not really deaf, in the sense that they did show some reaction to sounds. Furthermore, there seemed to be fairly rapid recovery in many cases, making the issue difficult to study. Recently, it has been shown in controlled experimental studies that bilateral ablation of A1 results in a profound loss of hearing sensitivity for a certain period of time, but that recovery does take place so that after a few weeks or months the animals have only a mild disturbance of sound detection. However, problems in more complex domains of auditory processing do remain, as described later.

Corresponding observations in human cases of so-called cortical deafness were also made in the early 20th century; patients with apparently complete loss of the ability to identify or discriminate sounds were described, but these individuals were usually not deficient in detecting the presence of sound. These dissociations have become much better understood in the past two decades. The concept of cortical deafness has been refined and usually refers to patients who suffer damage to auditory cortices in both hemispheres and who have profound difficulties in identifying sounds or recognizing speech. These patients usually have only mild or no threshold detection deficits, in accord with animal studies, and may also show no consistent disturbance in simple discrimination tasks for frequency or intensity. These phenomena indicate a functional dissociation since the inability to identify sounds is independent of the ability to detect the presence of sound. Thus, these cases support the idea that the auditory cortex is necessary for higher order processes but not for simply detecting or reacting to sounds. Cortical deafness contrasts with comparable cases in vision or in the somatosensory domain since lesions to the primary cortices in these modalities typically result in a nearly complete loss of conscious perception of light or touch.

B. Pitch Perception

The role of stimulus frequency in tonotopically organized cortex suggests that the cortex plays an important role in the processing of frequency and, hence, of pitch. However, the relation between pitch and frequency is not simple since pitch is a perceptual entity that is affected by many different aspects of the stimulus. Thus, the effects of lesions of auditory cortex on pitch depend on the nature of the task and stimuli used.

In many animal studies, it has been shown that complete destruction of A1 and surrounding areas bilaterally entails only transient difficulties in simple frequency discrimination tasks. Even when threshold differences in frequency between pure tones are measured, damage to A1 has no effect in tasks in which the animal must discriminate between the same tone presented twice and two tones of slightly different frequency. Similar results have been demonstrated in human patients with lesions of auditory areas. The conclusion from these studies is not that auditory cortex plays no role in pitch perception but, rather, that these simple discrimination tasks do not require the processes carried out by cortical structures.

Many studies have revealed that cortical mechanisms are important for pitch processing under conditions in which stimuli or tasks make certain demands. For example, the phenomenon of residue pitch, or missing fundamental pitch, apparently depends on cortical processes. This effect occurs when a complex tone is presented in which little or no energy is present at the fundamental frequency. Under certain conditions, it is possible to hear the pitch corresponding to the fundamental frequency, even in its absence, as a consequence of the frequencies of the remaining harmonics that are multiples of the fundamental frequency. This percept, which requires integration of information from several harmonics to abstract the missing element, is affected by damage to A1 in cats. It has also been shown in human patients that damage to HG in the right hemisphere leads to disturbances in the same type of pitch task.

Another example of pitch processing that is dependent on cortical function relates to the situation in which the processing of pitch cannot be done on the basis of a simple change detection. In animal

experiments, for example, lesions of at least some portions of auditory cortex impair the ability to respond to the cessation of a pitch from a set (i.e., detecting that ABAB has changed to AAAA), whereas the reverse order is not affected by such lesions. The important factor here is that it is easier to detect a change when a new element is added than when some of the same stimulation remains in both cases. Similarly, cortical lesions impair cats' ability to discriminate simple pitch patterns (e.g., ABA from BAB), whereas there is no difficulty in discriminating AA from AB.

A similar example from human studies is that the threshold for discriminating the pitch of two tones is unimpaired following lesions of HG in a simple same/different task, in which the subject must simply indicate that a change has or has not occurred [i.e., AA (same) vs AB (different)]. However, patients with damage to the right HG suffer a fourfold increase in threshold for a very similar task that requires identifying which of two tones is higher, whereas damage to the left HG has no effect. In this task the two stimuli in a trial are always different (AB or BA); hence, it is insufficient to merely detect that a change has occurred, as in the prior task. Instead, the two stimuli must be represented at a higher level of abstraction because the relation between them on some ordered scale is required to obtain the correct answer. This can only be accomplished if the appropriate cortical mechanisms are intact. Note that in these studies damage to the left HG has little or no effect, reflecting a hemispheric specialization discussed in greater detail later.

C. Spatial Localization

The ability to localize sounds in space is a function that appears to depend on the integrity of the primary auditory cortex. The presence of cortical neurons sensitive to interaural differences would be consistent with this functional role. Experimental studies in animals provide the most consistent data. In many species, unilateral destruction of A1 leads to a disturbance in localizing sound sources in the contralateral field. This deficit can be quite specific; for example, if the damage is to a restricted frequency band in the cortex, the animal will be unable to localize within that frequency band but will be unimpaired at other frequencies. Discrimination of sound sources at different locations is sometimes still possible even after complete destruction of A1, however, if the behavioral response entails a simple discrimination rather than responding to a true position in space. Thus, these findings support the concept that auditory cortex is necessary to construct a representation of auditory space rather than merely computing interaural acoustic differences, which is accomplished at earlier levels of the auditory system.

Data on human auditory spatial localization are not as clear as those from controlled animal studies. Patients with unilateral or bilateral hemispheric lesions have often been described who show deficits in localizing sound sources in space, but the precise nature of the damage and type of deficit are not well correlated. Some studies report greater localization problems in the hemifield contralateral to the lesion, but other findings indicate that deficits may exist in both auditory hemifields after damage to the right cerebral hemisphere, which is known to play an important role in spatial processing generally. However, it is important to distinguish the effects of large hemispheric lesions that affect parietal and frontal lobes as well as auditory cortex from more restricted damage to the superior temporal region. The former are likely to result in generalized spatial processing impairments and hemispatial neglect, whereas the latter result in selective auditory localization problems.

Neuroimaging findings indicate that perception of moving sounds or registering the visual position of auditory targets, engages parietal cortex, in keeping with the known role of these areas in spatial processes. Another important finding is that patients with complete removal of one hemisphere early in life show only very mild disturbances of auditory localization, in contrast with the severe effect of much more restricted lesions of auditory cortex occurring in adulthood. Therefore, there may be considerable plasticity in the functional role of auditory cortices—a finding not suspected from the animal literature, perhaps because most such studies are done only in the acute phase and in adult animals, thus not allowing for plasticity to manifest itself.

IV. COMPLEX FUNCTIONS

A. Pattern Perception

It was previously mentioned that cortical damage in both animals and humans leads to deficits in discriminating pitch patterns, particularly when the elements in the pattern are rearrangements of individual elements. This type of finding is indicative of the

important role played by auditory cortical regions in processing complex sound patterns. In animals, these patterns are especially important for functions such as responding to conspecific vocalizations, but they are also relevant for interpreting many sounds from the environment.

Many aspects of pattern detection appear to occur automatically, as indexed by electrical and magnetic changes measured from STG. Auditory cortical areas appear to be highly sensitive to changes or deviations from a background, even when subjects are not listening to the stimulus but are engaged in another task. The cortex responds to simple deviations in frequency or intensity but also to subtle changes in more complex patterns, indicating that the auditory system is specialized to some extent for ongoing pattern analysis and detection of novelty. Such a mechanism inplies an important survival value since changes in the auditory environment could signal danger or an event needing attention.

In humans, the perception of speech is among the most complex types of patterns that must be processed and will be discussed later. Another example of pattern perception that is related to speech, but is independent of it, pertains to the processing of voices. Human voices contain characteristic acoustic features that are unlikely to exist in other environmental sounds. Recently, it was proposed on the basis of fMRI data that regions in the upper bank of the superior temporal sulcus may be specialized for processing auditory information related to the perception of voices. These parabelt regions would be well situated to extract higher order stimulus features according to the hierarchical nature of the inputs to these areas.

Another important example of complex auditory pattern perception concerns music. Music, which appears to be both unique to and ubiquitous in the human species, offers a particularly rich source of information concerning human auditory cortical processing. Cases of specific loss of musical functions following brain damage have been described dating back to the 19th century. Recently, detailed studies of patients with bilateral damage to auditory cortices have revealed that certain specific aspects of musical function, notably involving discrimination of pitch patterns, may be selectively abolished while leaving other auditory functions, including speech, intact. These findings lead to the conclusion that aspects of music may have a distinct neural substrate, independent of other processing domains.

Studies of patients with unilateral damage to the anterior STG have generally shown that discrimination of melodic patterns is most affected by damage to the right STG, although usually milder disturbance may also be seen after left STG lesions. These findings are in accord with the data reviewed previously indicating that aspects of low-level pitch processing are affected by right HG damage. Thus, whereas more fundamental aspects of auditory processing are dependent on the core regions, damage outside of the core area results in impairments in higher order processes. Deficits in processing tonal patterns are also described following damage to right auditory cortical areas when the stimuli consist of complex spectral elements, such as musical chords, or when the discrimination entails differences in the distribution of spectral energy, such as when changes in harmonic structure affect the timbre of musical instruments. Functional imaging data also suggest that core and belt areas of the right STG are involved in tonal processes. Imaging studies have found neural activity changes, often bilaterally but usually greater on the right, in these areas in tasks requiring analysis of pitch or of pitch patterns, including melodies and chords (Fig. 4).

These findings extend to other types of tasks as well, including situations in which tones must be retained over brief time intervals. In this type of task, working memory mechanisms are involved. Working memory is important for all types of complex auditory processing since sounds necessarily unfold over time, and a mechanism must therefore exist for holding auditory information on-line so that relationships between elements can be appreciated. In the case of melodies, working memory for pitch involves belt areas of the right auditory cortex, as shown both by lesion studies and by functional imaging studies in normal listeners. The latter studies have also shown that pitch judgments depend on interactions between auditory regions and frontal lobe cortices since frontal cortical regions become more active when tones must be retained over time intervals filled with distractor tones. Together, these findings support the idea of a hierarchy of processing, with basic aspects of pitch analysis being carried out in the core areas in HG, and more complex processes important for pattern analysis dependent on cortical areas in the belt region anterior and possibly posterior, to HG. The perceptual analyses carried out within these regions in turn must interact with frontal lobe regions, presumably via the connectivity described previously, for situations in which working memory is important.

Another example of a complex function is auditory imagery, or the ability to imagine sounds in the absence

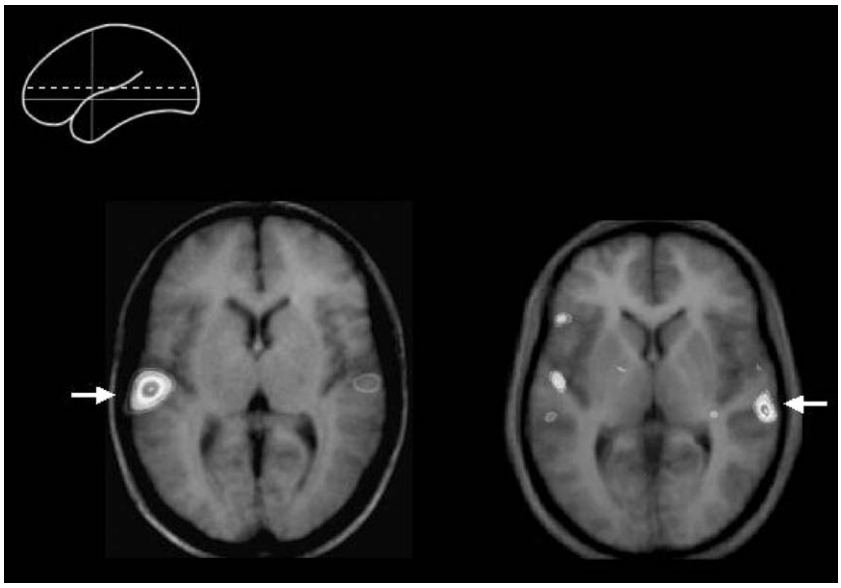


Figure 4 Positron emission tomography scans showing neural activity in superior temporal regions in response to auditory stimulation. Each image shows a horizontal slice through the temporal cortex at the position indicated by the dotted line in the inset. Changes in cerebral blood flow (arrows) are superimposed on structural magnetic resonance images. The left image shows increased activity in the left planum temporale in response to hearing speech sounds compared to a control condition of acoustically matched noise bursts. The right image shows bilateral activity in belt areas of the superior temporal gyrus, with stronger activity on the right side, in response to hearing tonal melodies.

of real auditory stimulation. This phenomenon has been studied in the context of musical imagery as well as imagery for speech. Several studies using functional imaging, as well as behavioral lesion techniques, have shown that auditory cortical areas, especially those in the right STG, are involved in the subjective experience of imagining a melodic pattern. Thus, there is an overlap between the cortical areas recruited for perceiving a sound and imagining it. An example of a similar phenomenon has been described for silent lip reading, which results in cortical activation of HG and surrounding areas, particularly on the left. This functional overlap between cortical zones involved in perception and imagery has also been described in the visual modality. These findings suggest that sensory information may be re-evoked by mechanisms that engage the sensory cortices that are initially involved in processing perceptual information from the environment.

B. Speech Perception

The uniquely human ability to perceive and produce speech sounds is among the most complex cognitive functions to which the auditory system contributes. The precise role of auditory cortical structures in speech is not fully understood, but considerable progress has been made in the past decade. Studies in monkeys have uncovered neural response features that may be considered as evolutionary precursors to human speech analysis—for example, the sensitivity of certain neurons to formant transitions, found in both human speech and monkey vocalizations, or the fact that some neurons respond to specific abrupt changes in sound envelope that correspond to some consonant speech sounds. However, the complexity of speech decoding goes beyond these low-level features since speech is combinatorial in nature, meaning that words in all natural languages consist of permutations and combinations of a limited set of sounds, called phonemes, which are by themselves not meaningful. The ease with which all normal humans understand the complex sound patterns inherent to speech and the fixed developmental sequence for language acquisition, strongly suggest that a specific neural specialization underlies this ability.

It has been clear for more than a century that one aspect of this specialization is that the left cerebral hemisphere plays a more important role in speech processing than the right in the majority of individuals. Classical studies of aphasic populations not only revealed a lateralization but also indicated that cortical

regions in the STG posterior to HG were especially important for the perception of speech sounds since damage to these areas, traditionally referred to as Wernicke's area, typically results in disruption of speech comprehension. The precise role of these areas has not been clear, however, because the precise location of the damage was often not well-known and because patients with such damage often display complex linguistic disorders that go beyond perception of speech sounds. Moreover, damage to frontal lobe language regions can often lead to speech perception deficits. The boundaries of the speech-related areas in the temporal lobe were further defined in the 1950s by Penfield, who tested neurosurgical patients under local anesthesia with the cortical stimulation technique. Penfield and others identified that disruption of speech could often be elicited by stimulation not only of the left posterior STG but also of regions in the middle temporal gyrus as well as the inferior parietal lobe (Fig. 5).

Recently, functional imaging studies have begun to reveal the organization of speech-related auditory cortical areas. The results of most of these studies are in agreement that speech sounds engage auditory areas both anterior and posterior to HG, but that the greatest differential response is in the left posterior temporal area (Fig. 4). The extent to which these specific areas respond uniquely to speech sounds *per se*, as opposed to certain acoustic characteristics that are present in speech, remains a matter of current research. It has also been proposed from imaging research that regions in the left STS and middle temporal gyrus respond specifically to auditory words.

Depending on the nature of the task, different patterns of activity have been noted. In particular, imaging studies have indicated that in addition to posterior temporal areas, left premotor areas may also be active when the task requires discrimination of speech phonemes. This effect has been attributed to the engagement of articulatory processes that may be involved in some aspects of speech processing. According to this view, which is not universally accepted, speech decoding is not solely dependent on analysis of the acoustic waveform but also depends on comparison of the auditory input with internal motor representations of the vocal gestures necessary to produce the sound.

Conversely, functional neuroimaging studies have also suggested that the area of the planum temporale can also be active during articulation even in the absence of auditory stimulation, for example, when auditory input is blocked by masking noise or when viewing lip and mouth movements, as mentioned previously. These findings thus challenge the conventional view that the posterior temporal speech zone is exclusively dedicated to speech perception, and that the frontal region is devoted to speech production. These regions may instead form part of an interactive

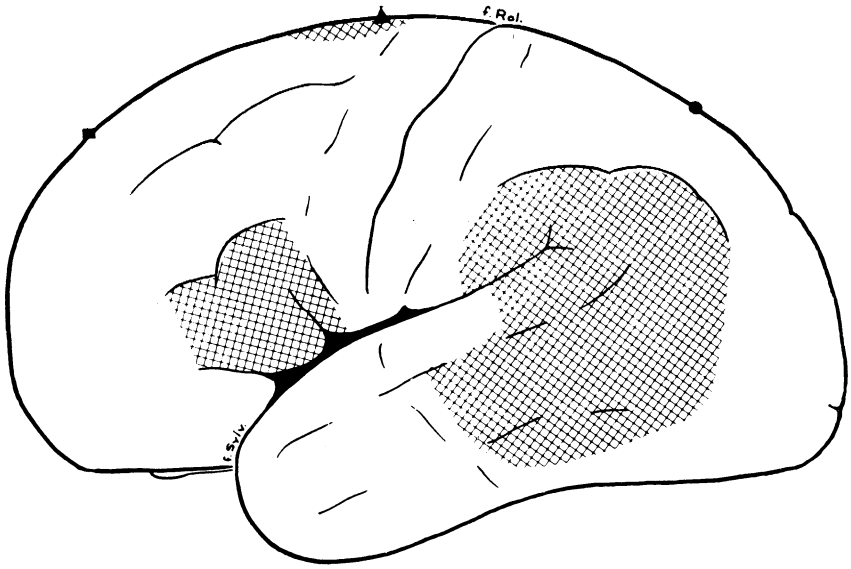


Figure 5 Diagram illustrating areas of the left cerebral hemisphere that cause disturbance of speech when stimulated electrically during neurosurgical procedures (reproduced with permission from W. Penfield and L. Roberts, 1959, *Speech and Brain Mechanisms*. Princeton Univ. Press, Princeton).

functional network, including other brain areas, that is involved in speech processing.

Much recent attention has been devoted to the idea that speech perception may depend on cortical mechanisms specialized for processing rapidly changing acoustic patterns. Because the perception of certain consonants in speech depends on small temporal differences (on the order of 20 msec), it has been proposed that the temporal resolution of the left auditory cortex is higher than that of the right, and that this factor underlies the hemispheric specialization for speech processing. Evidence in favor of this view derives from electrophysiological and functional imaging studies, which tend to show a greater response from the left than the right auditory regions to sounds that are separated by very brief time intervals, that contain short, abrupt changes in stimulus energy, or that contain rapid changes in formant transitions. Also, in at least some studies, patients with damage to auditory regions on the left have increased temporal discrimination thresholds, as do populations of children with certain developmental language acquisition disorders. It is thus possible that at least some aspects of hemispheric specialization for language arise from differences in processing at the level of primary and surrounding auditory cortices in the left and right hemispheres.

V. STRUCTURAL ASYMMETRIES

The considerable body of data indicating that the human auditory cortices are specialized for different types of processing in the left and right hemispheres has led to the search for possible structural correlates of this lateralization. The region of the PT has been one focus of such research since many studies have indicated that it differs across the two cerebral hemispheres. This asymmetry has traditionally been characterized as one of size, with the left PT being larger than the right in approximately 65–80% of brains. It is also known that similar asymmetries exist in the brains of newborns. These findings have been widely interpreted as providing a basis for the specialization of left posterior STG regions for auditory language processing. However, there is only limited evidence directly linking structural to functional asymmetries.

Considerable evidence also exists for hemispheric differences in the shape of the superior temporal region. It is well-known that the right Sylvian fissure angles upward more steeply on the right side, and that the point at which it begins to ascend is more anterior than on the left. Therefore, the apparent size difference of the PT may at least partly be a consequence of hemispheric differences in morphology rather than size.

Evidence exists that certain cytoarchitectonic regions within the PT may be larger on the left, but because there is no established relationship between the cytoarchitecture and the gross anatomical features on which the PT is defined, it is difficult to determine how measurements of the PT are related to cytoarchitectonically defined areas. The search for structure–function correlates is further complicated by the fact that the incidence of structural asymmetry in the PT is far lower than the incidence of left hemisphere language organization, which is estimated to be on the order of 95–98% of the population.

Other structural asymmetries in addition to the PT have also been described in auditory cortices. Recent evidence from both *in vivo* MRI studies and postmortem histological studies indicates that a greater volume of white matter underlies the left HG and surrounding cortex than the equivalent areas on the right (Fig. 2). In addition, there is evidence that myelination of axonal fibers on the left side is thicker, which would be compatible with the hypothesis that the left auditory cortices are specialized for faster neural transmission, as required for perception of speech signals.

VI. PLASTICITY AND REORGANIZATION

In the past several years, much research has revealed that the functional characteristics of the auditory cortex are not necessarily static but instead may be influenced by environmental experience. It has been shown that many properties of cortical neurons may change as a function of such experience, and that reorganization may occur when there is a change in the normal sensory input.

The most dramatic changes in auditory cortex organization are evident when there is loss of a sensory modality, such as occurs in blind or deaf individuals. In experimental animals, it has been shown that early loss of vision entails an increase in auditory-sensitive neurons in some cortical areas. Also, some of these neurons respond with a higher degree of accuracy to sounds in specific spatial positions than do similar neurons in normal animals. This latter observation is most likely responsible for the behavioral

advantage that is sometimes seen in blind humans and other species for navigating in space with the use of auditory cues.

Reorganization of auditory cortex may also occur when it is deprived of its normal afferent input due to cochlear or other damage to the periphery of the auditory system. For example, if the afferent input from a restricted area of the basilar membrane is disrupted, the tonotopic organization of the cortex is also altered, with frequencies corresponding to the damaged region diminishing, while adjacent tonal regions increase their representation. Damage to one ear also changes the response properties of neurons that are normally sensitive to binaural cues.

In cases of complete deafness, on the other hand, there is evidence from both human and animal studies that auditory cortical regions may respond to visual cues, although the nature of the processes involved has not been worked out. However, it has been shown in some animal species that when optic nerve inputs are rerouted to the medial geniculate nucleus during early phases of development, auditory cortical areas begin to show some of the structural and functional characteristics of visual cortex, implying that these features are driven by the nature of the stimulation that is received rather than by epigenetic factors exclusively.

Other evidence of plasticity in the auditory cortex derives from studies examining changes associated with learning and conditioning paradigms. For example, a tone stimulus results in greater cortical activation when it has been conditioned to precede an air puff to the eye than when it is presented in isolation. This increased activity disappears once the tone is no longer paired with the air puff and it is specific to the frequency used for conditioning. Such findings indicate that auditory cortical activity reflects not only the characteristics of the stimulus but also its associative value and hence the importance of the stimulus in a given situation.

In monkeys, cortical tonotopic organization may be influenced by training the animal to respond to a particular frequency, which induces an increase in the representation of that frequency in the tonotopic map at the expense of adjacent regions. The degree of expansion correlates with behavioral performance. These findings indicate that far from being fixed, neural response properties are dynamic and subject to considerable influences based on environmental input. These types of modifications are likely to be important for speech perception as well; infants begin to respond differentially to the phonemes of the language around them by 6 months of age, which is likely related to learning-induced changes in auditory cortical processing. Changes to cortical organization as a consequence of experience have also been noted in studies of musically trained individuals, who show a higher amplitude of cortical magnetic response to piano tones compared to pure tones, whereas musically untrained subjects do not. This type of finding tends to be strongest in individuals whose training started early in life, indicating that plasticity is greatest during development, although it may still occur in the adult organism.

VII. CONCLUSION

The human auditory cortex has attracted considerable attention in the past decade, and much progress has been made. The advent of structural and functional imaging techniques has facilitated the detailed study of the human auditory cortex and has permitted comparison with knowledge derived from other species. It seems clear that distinct auditory cortical areas exist in the human brain, but their boundaries and functional properties have yet to be identified. It also appears that important differences exist in auditory cortical functions in the two hemispheres. Finally, recent work has emphasized the modifiable nature of auditory cortical organization, yielding a complex and dynamic view of this important sensory system.

See Also the Following Articles

APHASIA • AUDITORY AGNOSIA • AUDITORY PERCEPTION • HEARING • LANGUAGE AND LEXICAL PROCESSING • LANGUAGE DISORDERS • SENSORY DEPRIVATION • SPEECH • TEMPORAL LOBES • WERNICKE'S AREA

Suggested Reading

Brugge, J. F., and Reale, R. A. (1985). Auditory cortex. In *Cerebral Cortex. Vol. 4. Association and Auditory Cortices* (A. Peters and E. G. Jones, Eds.), pp. 329–351. Plenum, New York.

De Ribaupierre, F. (1997). Acoustical information processing in the auditory thalamus and cerebral cortex. In *The Central Auditory System* (E. G. Romand, Ed.), pp. 317–397. Oxford Univ. Press, Oxford.

Fitch, R. H., Miller, S., and Tallal, P. (1997). Neurobiology of speech perception. *Annu. Rev. Neurosci.* **20,** 331–353.

Fitzpatrick, K., and Imig, T. (1982). Organization of auditory connections. In *Multiple Auditory Areas* (C. Woolsey, Ed.), Vol. 3, pp. 71–109. Humana Press, Clifton, NJ.

Handel, S. (1989). *Listening*. MIT Press, Cambridge, MA.
Kaas, J. H., Hackett, T. A., and Tramo, M. J. (1999). Auditory processing in primate cerebral cortex. *Curr. Opin. Neurobiol.* **9,** 164–170.
Marin, O. S. M., and Perry, D. W. (1999). Neurological aspects of music perception and performance. In *The Psychology of Music, Second Edition* (D. Deutsch, Ed.), pp. 653–724. Academic Press, San Diego.
Middlebrooks, J. C., and Green, D. M. (1991). Sound localization by human listeners. *Annu. Rev. Psychol.* **42,** 135–159.
Näätänen, R., and Alho, K. (1997). Mismatch negativity—The measure for central sound representation accuracy. *Audiol. Neuro-Otol.* **2,** 341–353.
Rauschecker, J. P. (1998). Cortical processing of complex sounds. *Curr. Opin. Neurobiol.* **8,** 516–521.
Rauschecker, J. P. (1999). Auditory cortical plasticity: A comparison with other sensory systems. *Trends Neurosci.* **22,** 74–80.
Syka, J. (Ed.) (1997). *Acoustical Signal Processing in the Central Auditory System*. Kluwer Plenum, New York.
Whitfield, I. C. (1985). The role of auditory cortex in behavior. In *Cerebral Cortex. Vol. 4. Association and Auditory Cortices* (A. Peters and E. G. Jones, Eds.), pp. 329–351. Plenum, New York.
Zatorre, R. J., and Binder, J. R. (2000). Functional and structural imaging of the human auditory system. In *Brain Mapping: The Systems* (A. W. Toga and J. C. Mazziota, Eds.), pp. 365–402. Academic Press, San Diego.

Auditory Perception

WILLIAM A. YOST
Loyola University, Chicago

GLOSSARY

auditory scene analysis The analysis of multiple sound sources into auditory perceptions of the individual sources using cues such as spatial separation, temporal modulations, spectral profiles, harmonicty and temporal regularity, and common onsets and offsets.

central auditory nervous system The anatomical parts of the brain stem and cortex devoted to auditory processing.

frequency The frequency of a periodic quantity, in which time is the independent variable. Frequency is the number of periods occurring in unit time. Unless otherwise specified, the unit is hertz.

hearing Determining the sources of sound.

intensity The magnitude of sound measured in a specified direction at a point is the average rate of sound energy transmitted in the specified direction through a unit area normal to this direction at the point considered.

loudness Subjective attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from soft to loud.

masking The process by which the threshold of audibility for one sound (signal) is raised by the presence of another (masking) sound.

peripheral auditory system The outer ear, middle ear, inner ear, and auditory nerve bundle.

pitch Subjective attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high.

spatial hearing The ability to locate a source of sound in three-dimensional space based on only acoustic cues.

Hearing, like all of our senses, allows us to determine objects in our world. Objects can vibrate, producing sound, and our sensitivity to that sound allows us to know something about the sound-producing sources. Often, many sources produce sound at the same time, and we parse this complex auditory scene into perceptions of its constituent sound sources. The sounds from these many sources are combined into one complex sound field. When the auditory system receives this complex sound field, the auditory periphery provides a neural code for the physical variables of the sound field: intensity, frequency, and time. This neural code is further processed by the central nervous system providing the neural substrates for the perceptions of the individual sound sources that originally generated the sound field. This article reviews the physics of the three physical variables of sound (intensity, frequency, and time); describes the peripheral auditory system and perceptual data related to processing intensity, frequency, and time; describes the central auditory system; and discusses the variables that influence sound source determination.

I. THE SOUND FIELD

A. Sound

Any object that has inertia and elasticity can vibrate and as a consequence can produce sound. Vibrations have three main variables: amplitude, frequency, and time. The simplest vibration is harmonic motion as described by a sinusoidal relationship between displacement and time shown in Fig. 1. Amplitude describes the distance through which a vibrating object travels, and it is related to intensity. Frequency is the inverse of

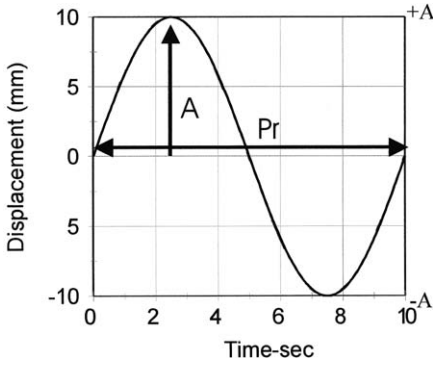


Figure 1 A sinusoidal relation between displacement and time. The period (Pr) is the time it takes to complete one cycle, and A is the peak amplitude. Period is 10 msec, frequency is 100 Hz, peak amplitude is 10, and starting phase is 0°.

period and describes how often per unit of time the object oscillates. Starting phase describes the relative starting point of the vibration.

A powerful theorem, Fourier's theorem, states that any arbitrary physical vibratory pattern consists of a sum of sinusoidal vibrations (Fig. 1). Sinusoidal vibration is a basic unit of vibration, and thus of sound. It is therefore not surprising that much of what is known about hearing has been derived from the study of sinusoidal vibrations and sounds. The sinusoidal relationship is written as $A(t)=A \sin(2\pi ft+\theta)$, where $A(t)$ is the instantaneous amplitude, A is the peak amplitude, f is frequency, t is time, and θ is the starting phase.

1. Intensity

The amplitude of vibration can be expressed as the instantaneous amplitude [$A(t)$], peak amplitude (A), peak-to-peak amplitude, or the root mean square amplitude. When an object vibrates it can produce pressure, which is proportional to amplitude. A vibrating object can also produce work. Work is measured in units of power or energy. Sound intensity is the measure of sound magnitude in units of power or energy. Thus, sound magnitude can be expressed in units of amplitude (displacement), pressure, or intensity (power or energy).

The dynamic range of sound intensity over which the human auditory system functions is approximately 10^{13}. In order to deal with this large dynamic range, sound intensity is often converted to decibels (dB) by $\text{dB}=10 \log_{10}(I_1/I_0)$ or equivalently $20 \log_{10}(p_1/p_0)$ since sound intensity (I) is proportional to pressure (p) squared. Therefore, in decibels the dynamic range of hearing is 130 dB ($10 \times \log_{10}10^{13}=130$ dB). It is common to use 20 (μPa) as the referent pressure (p_0) in calculating decibels since the threshold for hearing sinusoidal sounds is approximately 20 μPa. When this reference is used, the sound is measured in dB sound pressure level (SPL); therefore, a sound level of 50 dB SPL is a sound that is 50 dBs more intense than 20 μPa, (i.e., 50 dB higher than the level of the softest sound humans can detect).

2. Frequency

Frequency (f) is measured in hertz (Hz), where f in Hz=1/Pr, when Pr is the period of vibration and is measured in seconds (Fig. 1). A vibration that oscillates with a frequency of n Hz goes through n complete vibratory cycles in 1 sec. Many animals can hear over a range of sinusoidal frequencies from a few hertz to tens of thousands of hertz. The human range of hearing is approximately 20–20,000 Hz.

3. Starting Phase

The starting phase (θ) of vibration is measured in angular units of degrees or radians. When a sinusoid completed one cycle, it has gone through 360° (2π radians) since sinusoidal motion can be derived from the projection of a point on a rotating circle. The sinusoidal motion that begins at zero amplitude at time zero, as shown in Fig. 1, is defined as having zero starting phase. All other possible starting phases are relative to this zero starting phase condition. Humans are only sensitive to changes in starting phase in certain situations.

B. Sound Propagation

Air is the common medium for sound propagation for human hearing. The molecules of air are in constant random motion and the density of the molecules dictates the static air pressure. If an object is vibrating in this space, the outward vibration of the object causes the molecules that the object initially makes contact with to be pushed in the direction of the object's outward motion. This will cause a condensing of the molecules, thus increasing the air molecule density at the location of the condensation. This increased density leads to an increase pressure. When the object moves in an inward direction air molecules will tend to fill the vacated space evenly (a property of the gas

laws), resulting in a rarefaction of molecules and, hence, a lowering of the density of molecules. The lower density leads to a lower pressure. The distance between successive condensations (or rarefactions) is the wavelength (λ) of sound, such that $\lambda = c/f$, where c is the speed of sound in the medium and f is frequency in hertz.

These areas of increased (condensations) and decreased (rarefactions) pressures of the sound wave radiate out from the vibrating source in a spherical manner. As the sound travels from the source the density at condensations and rarefactions decreases, since the area of the sphere is increasing. This results in a loss of pressure as the sound travels from the source, and this pressure loss is proportional to the distance squared (called the inverse square law).

In most spaces sound will encounter obstacles as it travels from its source to its destination. The sound wave can be absorbed at an obstacle's boundary, transmitted to the medium of the obstacle, reflected from the obstacle's surface, or diffracted around the obstacle. All these have significant consequences for hearing in the real world. For instance, a reflected sound wave may encounter the originating sound wave (or other reflections). The reflected wave may reinforce or cancel the originating sound wave. Locations of reinforcement will cause areas of increased sound pressure, whereas areas of cancellation lead to decreases in sound pressure. Often, the pattern of reinforcements and cancellations sets up a standing wave pattern such the standing wave vibrates at its resonant frequency causing enhanced intensity at this resonant frequency. The resonant frequency is inversely proportional to the size of the environment in which the standing wave occurs. This is the principal on which woodwinds, horns, and organ musical instruments operate. The pitch of the sound is proportional to the resonant frequency of the standing wave set up in the "pipe." Not only can standing waves exist in our acoustic environment but they can also exist in the ear canal and other parts of the auditory system.

In other cases, an object may cause a sound shadow in which there is an area of reduced pressure on the side of the object opposite from the sound source. The size of this sound shadow is a joint function of the wavelength of the sound and the size of the object such that to a first approximation sound shadows exist when the object's dimension are close to those of the sound's wavelength. The fact that the head produces a sound shadow plays a significant role in sound localization.

C. Sound Analysis

The auditory system is a sound analyzer and our models of this analysis are based on physical systems. Frequency is the primary aspect of sound that is analyzed by the auditory system. Consider listening to several notes played together on a piano. Often, it is relatively easy to determine the different pitches. However, the sound that arrives at the auditory system is one complex waveform that is a sum of all the frequencies that make up this composite sound (recall the role of Fourier's theorem). The primary frequencies of this complex sound are those corresponding to the basic vibration of each piano string that is struck when the various piano keys were pressed. These primary frequencies produce the perceived pitches of the various notes. The fact that we perceive these pitches means that the auditory system has determined the various frequency components that made up the complex waveform that arrived at the auditory system as a single sound.

When the waveform is described in terms of the pressure or displacement waveform (Fig. 1), it is being described in the time domain. If it is described in terms of the frequency components that constituent the waveform, it is being described in the frequency domain. Figure 2a shows a complex waveform in the time domain. This waveform is the sum of three sinusoids. The amplitude and phases of these three sinusoids are shown in the amplitude and phase spectra as a function of frequency. These spectra form the frequency domain description of the complex wave.

The most general model for frequency analysis by the auditory system is filtering. Although there are not actual physiological filters in the auditory system, the results of the biomechanical and neural actions of the auditory periphery behave as if there were. The important filter for the purposes of auditory analysis is the bandpass filter, in which all frequency components of a complex sound that are within the passband of the filter are passed unaltered and those components with frequencies higher and lower than the passband have their levels attenuated and their starting phases altered (actually, the components outside the passband are delayed, resulting in a phase shift). The levels will decrease as a constant roll-off ratio in dB/octave changes in frequency from the cutoff frequencies of the passband (an octave is a doubling of frequency).

Bandpass filters can be used to estimate the amplitude spectrum of a complex waveform. If the waveform contains frequency components that are in the

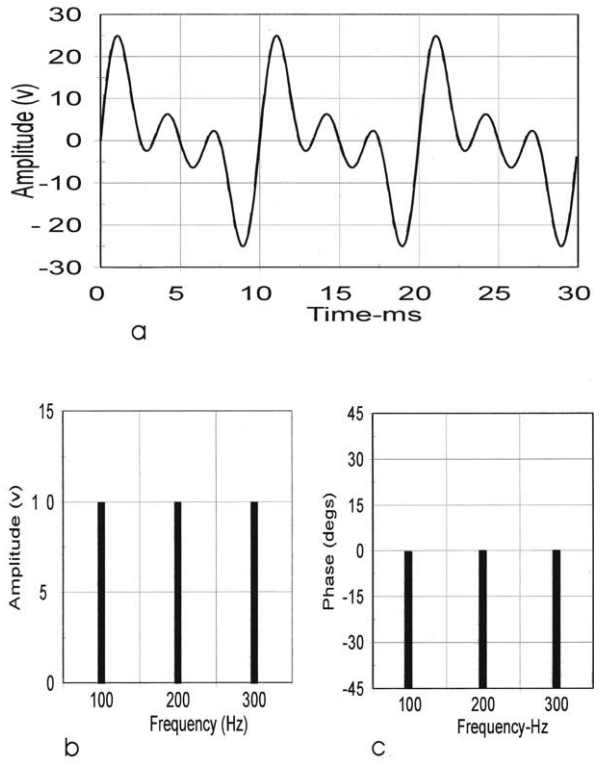


Figure 2 (a) The sum of three sinusoids with frequencies of 100, 200, and 300 Hz produces the time domain waveform. The frequency domain representation is described by the amplitude (b) and phase (c) spectra.

passband of a bandpass filter, then the filter's output will be high compared to the case in which the stimulus does not contain frequency components in the filter's passband. Thus, a series of filters, each with a different passband region (assuming that the passbands are very narrow), can be used to analyze a complex waveform to obtain an estimate of the waveform's amplitude spectrum. Filters with high levels at their output indicate that there are frequency components in the complex waveform near the spectral region of the filters' passbands. Thus, a function relating the magnitude of the filters' output to the spectral region of their passbands is an estimate of the sound's amplitude spectrum. The narrower and steeper the filters' passbands become the better the filters estimate the sound's amplitude spectrum. The equivalent filters in the auditory system have very narrow passbands with steep slopes.

Most real sound analysis systems, including the auditory system, are nonlinear. A consequence of nonlinearity is that frequency components exist at the output of a nonlinear processor that were not present in the input. For instance, if the input consists of the sum of two frequency components, f_1 and f_2, the nonlinear output contains frequencies equal to $mf_1 \pm nf_2$, where $n = m = 1, 2, 3, 4$, etc. The frequency components that are integers of f_1 and f_2 (nf_1 and mf_2) are harmonics, and if they are audible they are called aural harmonics. The terms $mf_1 + nf_2$ are summation tones, and the terms $mf_1 - nf_2$, are difference tones. The cubic difference tone, $2f_1 - f_2$ ($m = 1$, $n = 2$), is a significant nonlinear auditory component. A nonlinear processor will distort its input due to the addition of the nonlinear frequency components to the originating input components.

II. THE PERIPHERAL AUDITORY SYSTEM

Figure 3 displays the major parts of the auditory system: outer ear, middle ear, inner ear, and the central nervous system. The outer ear collects the sound for delivery to the middle ear, which allows air pressure to effectively vibrate the fluids and tissues of the inner ear. The inner ear serves as the auditory biological transducer translating vibration into a neural code for intensity, frequency, and temporal structure. The central nervous system processes this neural code, allowing for sound source determination.

A. Outer and Middle Ears

As sound travels to the outer ear, it passes over the torso, head, and especially the pinna. All these structures attenuate and delay the passage of the sound to the outer ear, and the attenuation and delay depend on the interaction between the size of the structures and the wavelength of the sound, such that high-frequency sounds are affected more than low-frequency sounds. These transformations alter the spectrum of the originating sound and are called head-related transfer functions (HRTFs). HRTFs are significant for sound localization.

The outer ear canal has a resonant frequency near 4000 Hz (i.e., standing waves exist within the outer ear), and as such sounds with frequency components near 4000 Hz are more intense in the ear canal than are other frequency components. The middle ear ossicles (bones) provide further enhancement in sound level in the region of 2000–5000 Hz. All these increases in sound level are necessary if air pressure is to produce an effective vibration within the inner ear. The inner

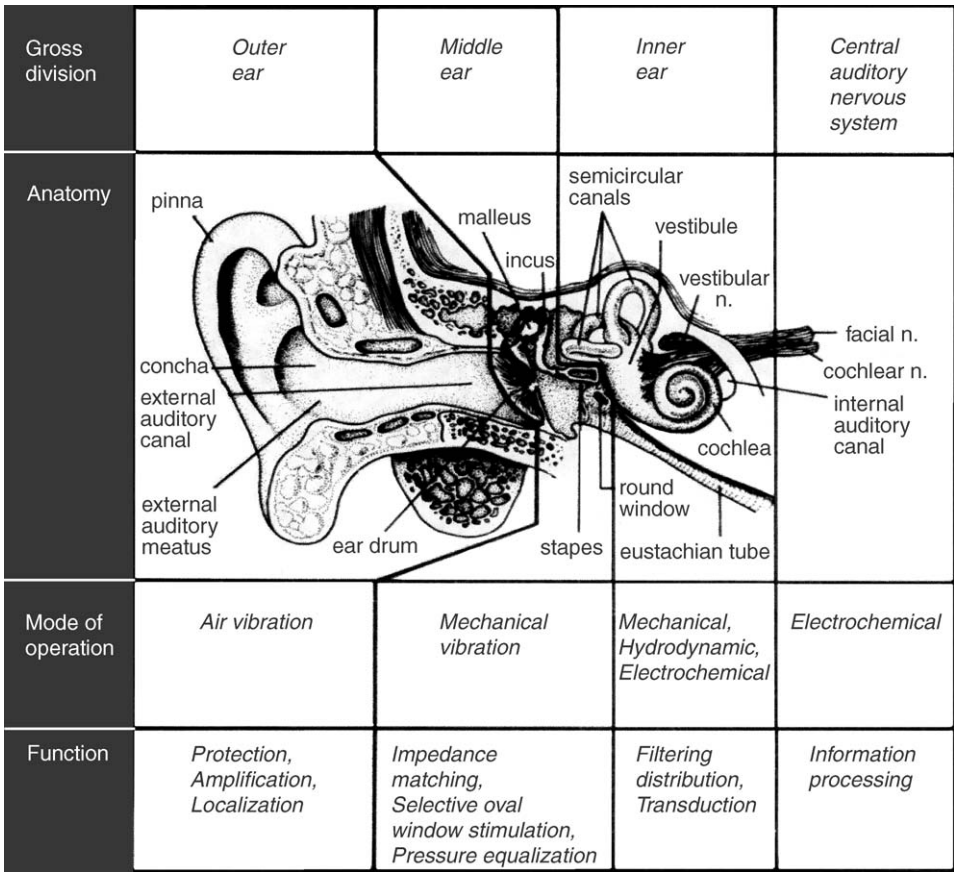


Figure 3 The four major divisions (outer ear, middle ear, inner ear, and central auditory nervous system) of the auditory system, their mode of operation, and their function are shown (from Yost, 2000).

ear is a fluid-filled space containing neural structures and their support. These inner ear structures offer significant impedance to the transmission of vibration from the air-filled outer ear to the inner ear. The resonance of the outer ear and the increases provided by the ossicular chain provide an impedance match between air and the fluids and structures of the inner ear so that over a significant portion of the audible range of hearing, changes in air pressure impinging on the auditory system are efficiently transmitted to the inner ear with no loss in sound level. Damage to the ossicular chain leads to significant hearing loss because of the loss of this crucial impedance matching function.

B. Inner Ear

The inner ear contains the neural structures for both the sense of balance (the vestibular system) and hearing. The auditory part of the inner ear consists of the cochlea, a tube that wraps around itself three or four times (depending on species) in a snail-like coil. The footplate of the stapes (the medial-most ossicular bone) pushes in on the oval widow at the base of the cochlea. The cochlear tube contains an inner tube, the cochlear partition, that is a sealed structure running the length of cochlea except at the top (apex), where there is an opening (the heliocotrema) between the cochlear partition and the end of the cochlea.

Figure 4 shows a schematic diagram of a cross section of the cochlear partition, the inner tube of the cochlea. Along the bottom membrane, the basilar membrane, lie the inner and outer hair cells, the neural elements for hearing, and their supporting structures. Figure 5 shows a view looking down on the top of the hair cells, revealing the hairs (stereocillia) from which the hair cells derive their name. In mammals there are three rows of outer hair cells and one row of inner hair cells. When the fluids of the inner ear are vibrated by

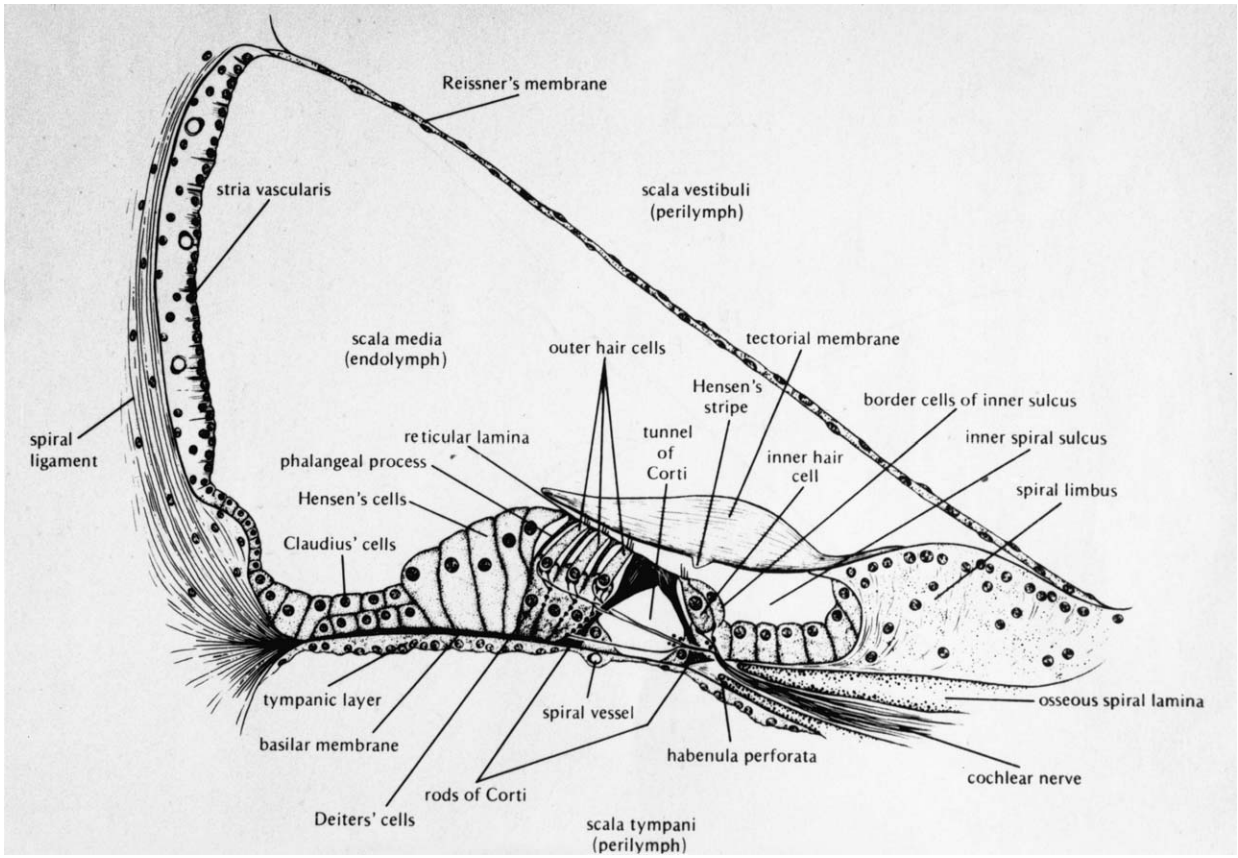


Figure 4 A cross-section of the cochlear partition of the cochlea is shown (from Yost, 2000).

the ossicular chain, a wave motion occurs within the cochlea that causes a differential movement between the basilar membrane on which the hair cells sit and the tectorial membrane at the top of the hair cells. This differential movement causes the stereocillia to shear (a type of bending), and this shearing ignites a generator potential in the hair cell causing an action potential in the auditory nerve fiber attached to the hair cell at its bottom.

This cochlear wave motion is a crucial part of the biomechanical action of the inner ear. The wave is a traveling wave (Fig. 6), such that the region of maximum vibratory displacement along the cochlear partition from base to apex is frequency dependent. High frequencies produce a wave that travels only partway from the base to the apex, thus producing maximum displacement near the cochlear base. Low frequencies cause the wave to travel to the apex, producing maximum displacement at or near the apex. Thus, the area of maximum cochlear partition vibration is distributed along the cochlear partition according to the frequency content of the originating sound. Since the hair cells are distributed from base to apex, those whose stereocillia will be maximally sheared depend on the sound's frequency content. Thus, hair cells are able to signal where, along the cochlear partition, the maximal displacement occurs and, hence, they code for the frequency content of sound. If the stereocilla of hair cells near the base are maximally sheared, the sound contains high frequencies, whereas maximal stereocillia shearing for hair cells near the apex code for low frequencies. Neural discharge rate is directly proportional to the amount of stereocillia shearing and, thus, to the displacement of the cochlear partition.

The frequency selectivity of this vibratory traveling wave is very sharp, the basilar membrane vibration is extremely sensitive, and the vibratory motion is a compressive nonlinear function of sound level. The exquisite frequency selectivity, high sensitivity, and important nonlinear function of the cochlear biomechanical traveling wave are only possible because of the way in which the inner and outer hair cells function. The inner hair cells are the biological transducers that

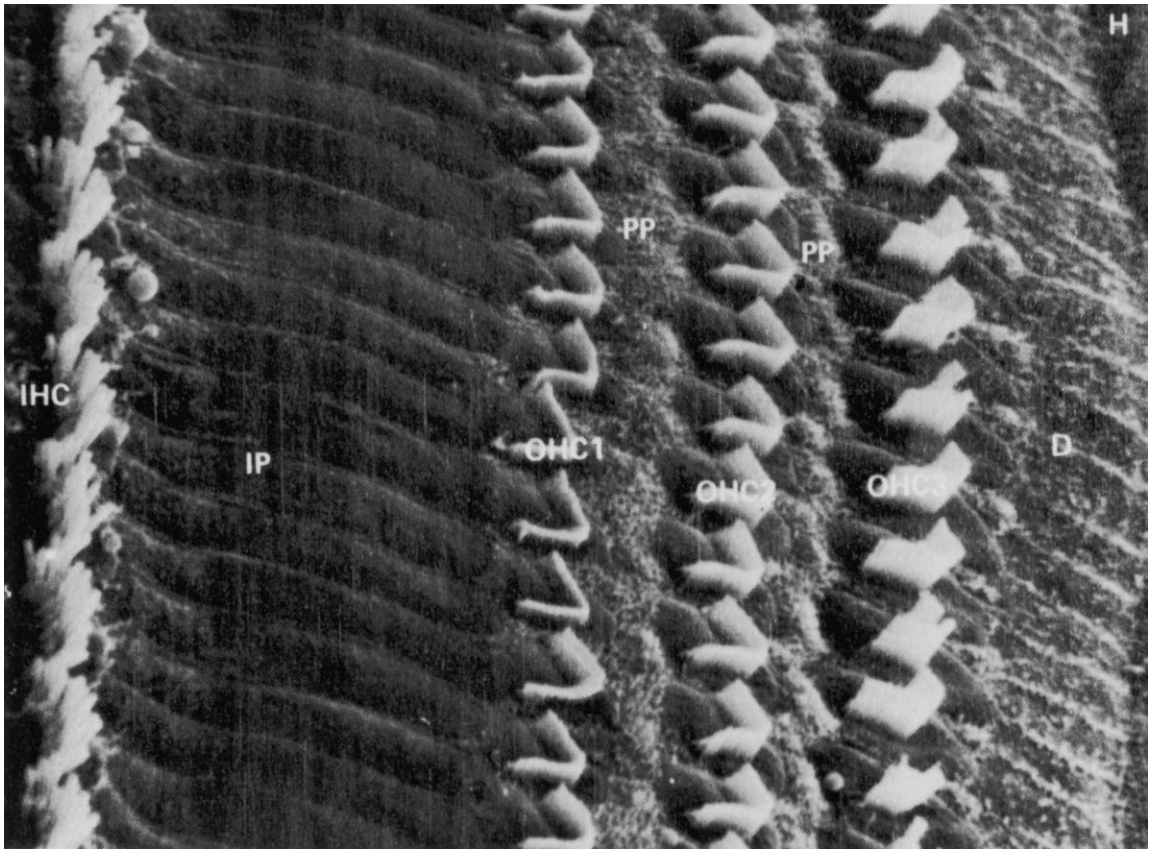


Figure 5 A scanning electron micograph showing the tops of the hairs and the stereocilla of the three rows of outer hair cells (OHC) and the one row of inner hair cells (IHC). H, IP, PP, and D are supporting structures within the cochlea (from Yost, 2000).

provide the neural code transmitted to the central auditory system by the auditory nerve. Each auditory nerve fiber is connected to a few inner hair cells and 90% of the auditory nerve fibers innervate the inner hair cells. The outer hair cells change size (primarily length) as a result of the vibration of the cochlear partition and the consequential outer hair cell stereocillia shearing. Because the length of the outer hair cells changes, this presumably alters the connections between the basilar and tectorial membranes, which would affect the biomechanical vibration of the cochlea. If the outer hair cells are damaged, the frequency selectivity and the sensitivity of the biomechanical vibration are compromised, suggesting the importance of outer hair cell motility in effective cochlear functioning. The fact that the outer hair cells are motile and this motility feeds back into the biomechanical action of the cochlea suggests that the outer hair cells provide an active mechanism that serves as a type of cochlear amplifier allowing the inner hair cells to provide a highly sensitive, very frequency selective, and compressively nonlinear code for the frequency, intensity, and timing of the acoustic input.

C. Auditory Nerve

Each auditory nerve fiber in the auditory portion of the VIII cranial nerve innervates a small number of inner hair cells, with nerve fibers near the center of the auditory nerve bundle coming from the cochlear apex and those on the outside of the bundle coming from the cochlear base. Thus, the fibers in the auditory nerve bundle are topographically organized on the basis of their cochlear innervation, and they transmit the neural code for sound from the cochlea to the cochlear nucleus in the auditory brain stem.

The neural response of auditory nerve fibers increases in firing rate over a 30- to 40-dB range. Each

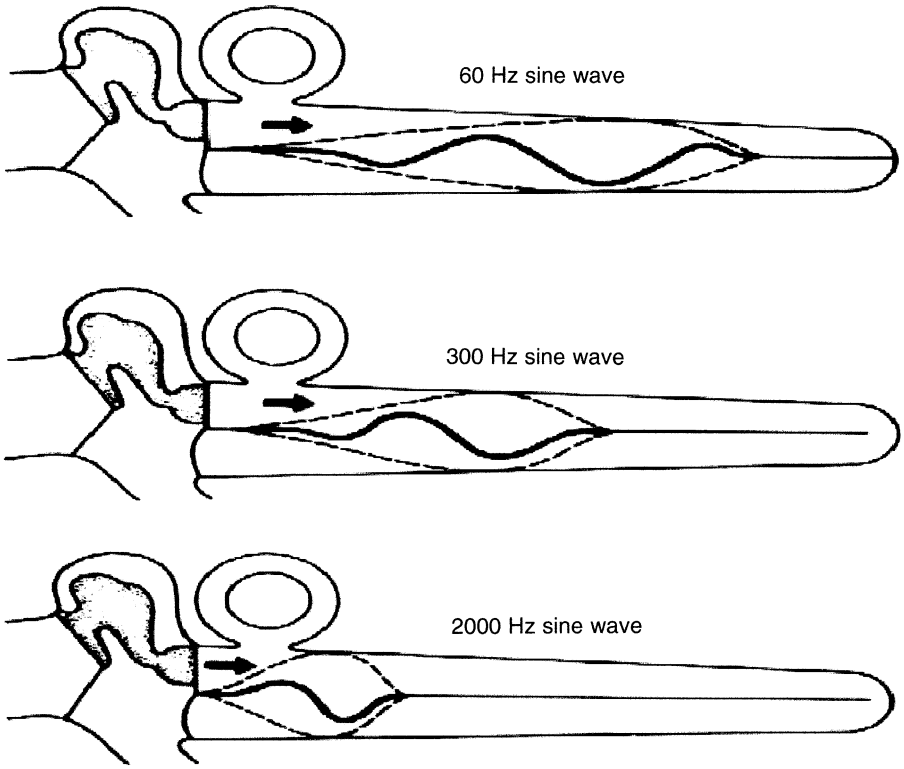


Figure 6 A schematic diagram of the traveling wave for three frequencies. The solid line is an instantaneous traveling wave, whereas the dotted curve represents the envelope of the traveling wave outlining its overall shape. The cochlea is shown as if it were unrolled. For low frequencies, the traveling wave travels to the apex, where maximal stimulation occurs. For high frequencies the wave travels a short distance from the base and maximal displacement is at the base (from Yost, 2000).

nerve fiber is highly selective to the frequency of the sound, reflecting the selectivity of the traveling wave (Fig. 7). Thus, each nerve fiber carries information about a narrow region of the spectrum and as such each fiber is tuned to a particular frequency. Because of the relationship between the location along the cochlear partition that each nerve fiber comes from and the activity of the traveling wave, auditory nerve fibers are tonotopically organized so that fibers near the middle of the nerve bundle carry information about low frequencies and those toward the outside of the bundle carry high-frequency information. The tuning curves of Fig. 7 are obtained by determining for each frequency the tonal level necessary to elicit a threshold neural discharge rate. The frequency that requires the lowest level to reach this threshold (the tip of the tuning curve) is the tuning curve's center frequency (CF).

The nerve fibers discharge in synchrony (Fig. 8) to the pressure waveform, such that the neural output represents a half-wave rectified version of the stimulus waveform. Thus, the neural output of the auditory nerve can follow the temporal structure of the waveform up to frequencies of about 5000 Hz. The upper frequency limitation of such temporal resolution is dictated by the refractory properties of neuronal function.

Figure 9 represents the output of a computational model that simulates the properties of the auditory periphery. Each line represents the aggregate neural response of fibers tuned to particular frequencies (low frequencies for fibers coming from the cochlear apex at the bottom of Fig. 9 and those for high frequencies coming from the cochlear base at the top). The model uses bandpass filtering to simulate the frequency selectivity of the tuning curves (Fig. 7), and a model of the hair cell–nerual interaction that provides a rectified, compressed, and adapted version of the filtered sound is used as the cochlear simulation. Thus, Fig. 9 represents the neural information flowing from the cochlea to the auditory brain stem for the vowel /e/, as in the word bead. The vertical bands of high output represent those fibers that are firing to the regions of

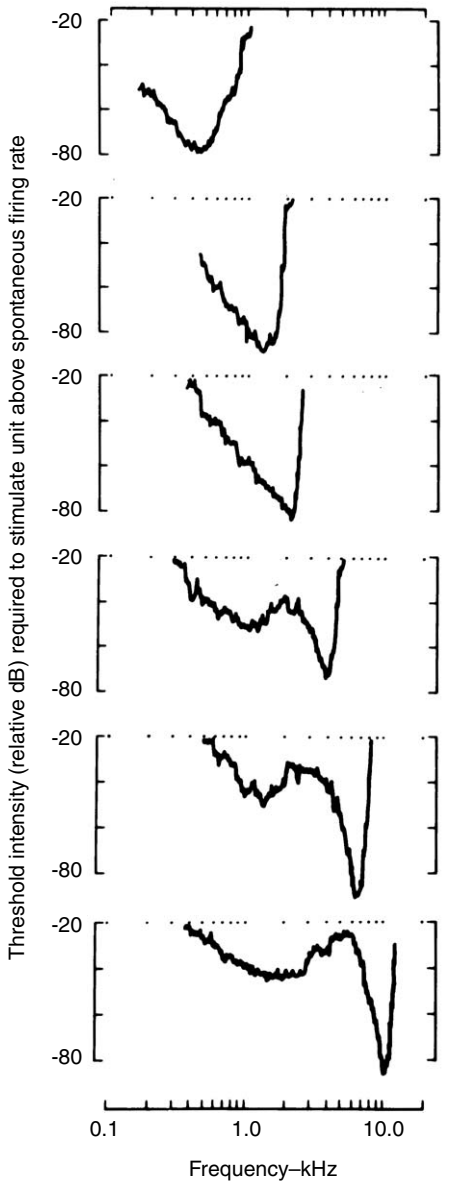


Figure 7 Tuning curves for auditory nerve fibers showing the responsiveness of different auditory nerve fibers to tones of different frequencies (from Yost, 2000).

the spectrum where the most energy occurs in the vowel's spectrum (i.e., the vowel formants), and the changes over time represent the temporal structure of the vowel waveform, most notably the periodic modulation due to the opening and closing of the vocal cords. Thus, Fig. 9 represents a simulation of the neural code for frequency, intensity, and timing provided by the auditory periphery.

III. THE AUDITORY BRAIN STEM AND AUDITORY CORTEX

The bilateral auditory neural pathway goes from the auditory nerve to the cochlear nucleus, the olivary complex, the inferior colliculus, the medial geniculate, and then the auditory cortex. There are numerous connections within and between these nuclei and from nuclei on one side of the brain to those on the other side. There is also a rich array of descending, efferent fibers that modulate the flow of neural information in the auditory pathway.

The tonotopic organization seen in the auditory nerve is preserved throughout the auditory brain stem and cortex, often being represented several times within a nucleus. The phase locking of neural elements to the temporal waveform is preserved in some centers but not in others. The exact nature of the tuning and phase locking can vary from nucleus to nucleus.

The cochlear nucleus (CN) contains three main regions (the dorsal cochlear nucleus, the anteroventral cochlear nucleus, and the posteroventral cochlear nucleus) with a rich array of neuronal cell types and physiological function. There is no detailed understanding of the auditory function of the CN, but there is ample evidence that there are strong inhibitory circuits that might provide a lateral inhibitory network for sharpening spectral contrasts in complex sounds. There are neurons and circuits that appear to integrate information both from auditory nerve fibers that carry similar information about the frequency content of sound (i.e., integrate information from neurons with tuning curves with similar CFs) and from auditory neurons that carry information about different frequencies in a sound (i.e., integrate information from neurons with tuning curves with different CFs). There are also circuits that preserve the exact temporal structure of complex sounds.

The olivary complex represents the first brain stem site at which there is strong bilateral convergence. Thus, the olivary complex [the medial superior olive (MSO) and the lateral superior olive (LSO)] is most likely responsible for the binaural processing of interaural time differences (in the MSO) and level differences (in the LSO) that are crucial for sound localization, especially in the horizontal or azimuthal plane.

The inferior colliculus (IC) is anatomically divided into three areas (central nucleus, dorsal cortex, and paracentral nucei). IC neurons appear to be tuned not only tonotopically but also to other stimulus conditions, such as the amplitude-modulated pattern of

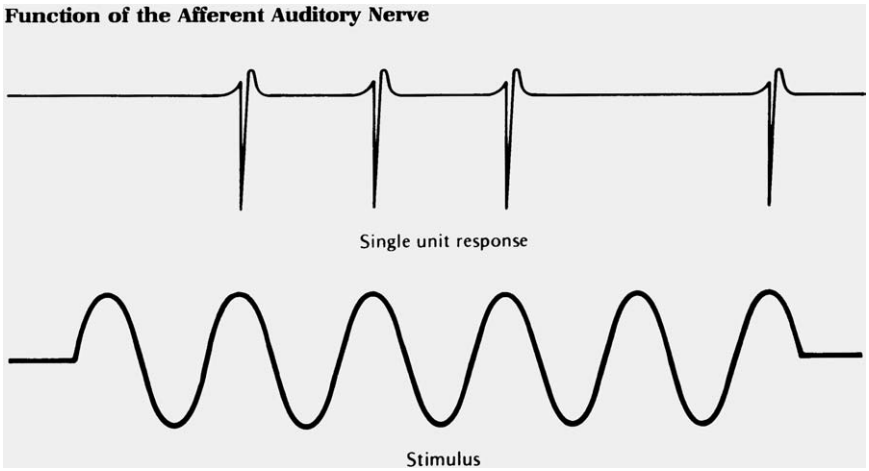


Figure 8 Auditory nerve discharges appear in synchrony with the stimulus waveform, such that when the nerve discharges it does so when the waveform reaches a peak (from Yost, 2000).

complex sounds and to the binaural cues of interaural time and level. It is possible that neural maps of auditory space might exist at the level of the IC. Work in the bat and the barn owl has provided useful models for better understanding the auditory function of the auditory brain stem.

The human auditory cortex is located deep within the Sylvian fissure and contains several areas, with A1 being the major auditory cortical center. As elsewhere in the auditory central nervous system, there is a rich array of fiber types and physiological responses in the cortex. Often, cortical neurons will only respond to particular stimulus waveforms (e.g., to a complex sound but not to a pure tone). However, with the exception of the barn owl and bat, the auditory function of different parts of the auditory cortex has not been well established. There is evidence that, like at the level of the IC, cortical neurons, perhaps organized in columns, are tuned to the amplitude modulation pattern. There is also evidence for auditory spatial maps in the auditory cortex.

The central auditory nervous system is a complicated system of many neural centers and countless number of interconnections that are organized in complex ways. It has been argued that the crucial information for hearing must be computed from the neural code provided by the auditory periphery, and the complexity of the central auditory nervous system represents the hardware for this computation.

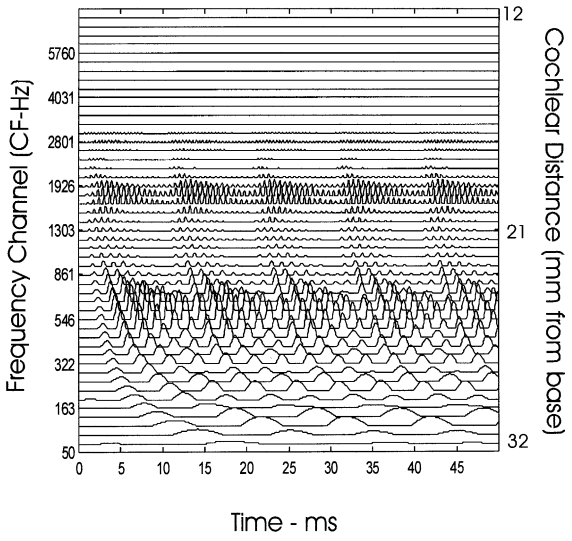


Figure 9 The output of a computational model of the auditory periphery indicating the neural code of the vowel /e/ that flows to the auditory brain stem. Each line represents the aggregate neural response of fibers tuned to the same frequency. Concentration of activity in the vertical direction represents the spectral code for hearing and structure in the horizontal direction represents the temporal code. The model simulates the actions of the outer ear and middle ears, the inner ear, and the auditory nerve.

IV. PERCEPTION OF THE SOUND FIELD

A. Absolute and Difference Thresholds

As noted previously, the key function of the auditory periphery is to code for the frequency, intensity, and temporal structure of sound. Auditory systems of all animals are sensitive to only a limited range of frequencies. Figure 10 displays the auditory thresholds

of humans to sinusoidal sounds of different frequencies. These thresholds of audibility can be obtained in one of two ways: The sounds can be presented over headphones [the minimal audible pressure (MAP) method] or over loudspeakers in a room with the listener [the minimal audible field (MAF) method]. When thresholds of audibility are measured, on average there is a 6-dB difference between MAP and MAF thresholds, which is due to diffraction of sound around the head and the calibration procedures used to obtain the MAP and MAF thresholds. The upper limit of hearing (e.g., the level at which sound becomes uncomfortably loud) is about 130 dB SPL. Thus, the dynamic range of hearing is about 130 dB in the frequency region between 500 and 4000 Hz. Humans can detect sound with frequencies ranging from 20 to 20,000 Hz, although hearing sensitivity, starting with the high frequencies, decreases with age.

The thresholds of hearing depend on the sound's duration. For durations less than approximately 300 msec, the energy of the sound determines threshold. Thus, long sounds require less sound power for detection than do short sounds. For durations greater than 300 msec, the ability to detect sinusoidal sound remains constant for signals of constant power. The duration at which the power of the signal no longer changes for detection threshold is used as an estimate of the integration time for detection.

Humans are sensitive to about a 0.2% change in frequency over a significant part of the range of hearing. For instance, one can discrimination between a 1000- and a 1002-Hz sinwave based on the frequency difference. An approximately 0.5-dB change in level is also discriminable over a significant part of the range of hearing. Thus, sensitivity to frequency and sound intensity (not expressed in decibels) is approximately proportional to frequency or level. This proportionality is referred to as Weber's fraction or law, and to a first approximation the auditory system displays a constant Weber fraction for frequency and level.

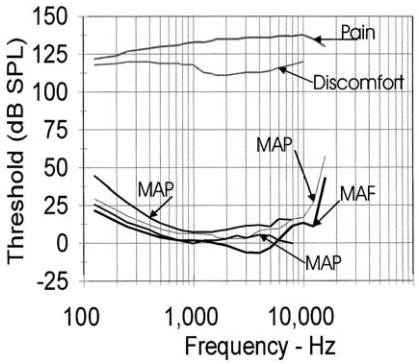


Figure 10 The absolute thresholds of hearing for MAP and MAF measures of audibility. Three different MAP measures are shown, representing measures with three different types of headphones. Thresholds of pain and discomfort are shown as estimates of the upper limit for the dynamic range of hearing.

The ability to discriminate sounds with changing levels from steady sounds is described by the temporal modulation transfer function (TMTF). The TMTF is measured by asking listeners to discriminate between a noise whose amplitudes are sinusoidally modulated and a noise with no amplitude modulation. The TMTF describes the depth of modulation (the difference in the changing amplitude) required to discriminate amplitude-modulated noise from unmodulated noise as a function of the rate of the amplitude modulation. The TMTF has a low-pass characteristic because as the rate of amplitude modulation increases beyond about 50 Hz the threshold for detecting amplitude modulation decreases. That is, it is more difficult to detect changes in the amplitude of modulated noise when the amplitude is changing faster than 50 times per second.

B. Masking

Most sounds do not occur in isolation. Sounds from different sources interact such that sound from one source may interfere with the detection of sound from a target or signal source (i.e., masking occurs). A masker only provides significant masking of the signal when the masker and signal have about the same frequency. Figure 11 shows results from a psychophysical tuning curve masking experiment. In the middle masking contour, the signal was a brief (10-msec), 1000-Hz tone presented at a low level, about 20 dB above its unmasked (absolute) detection threshold. The level of the masker is varied until the listeners in the experiment are at threshold in their ability to discriminate the signal-plus-masker presentation from the masker-alone presentation. At the tip of the psychophysical tuning curve, when the masker and signal have the same or nearly the same frequency, a very low-level masker interferes with discrimination, indicating an effective masker. When the frequency of the masker differs from that of the signal, a greater masker level is required for masked threshold, indicating that for these larger separations in frequency the masker is not an effective masker. Similar functions are

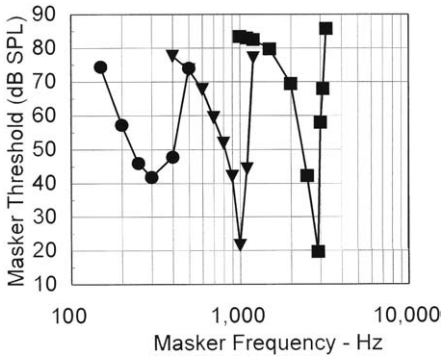


Figure 11 Psychophysical tuning curves for three signal frequencies [250 (●), 1000 (▲), and 3000 (■) Hz] showing the level of the tonal masker required for threshold detection of the tonal signal (from Yost, 2000).

obtained when the frequency of the signal is changed in different experiments.

These experiments suggest that only masker frequencies near that of the signal are critical for masking. A noise masker is a complex sound containing a continuum of frequencies. Experiments have been conducted in which the noise is filtered with a bandpass filter centered on the frequency of the sinusoidal signal and the detection of the signal is measured as a function of narrowing the bandwidth of the filter. The detection threshold remains constant until the filter bandwidth reaches a critical bandwidth, and further decreases in filter bandwidth lower the signal threshold, indicating that the signal is easier to detect. This result is consistent with the assumption that there is an internal filter centered on the signal frequency, and it is the power of the noise masker coming through the internal filter that determines signal detection threshold (Fig. 12). Several different types of noise-masking experiments can be performed to estimate the width of this internal, critical band filter. The width of the estimated critical band is proportional to signal frequency such that critical bandwidth increases with increasing frequency. The critical bandwidth is approximately 130 Hz wide when the signal is 1000 Hz, and it is 1100 Hz wide when the signal is 10,000 Hz. The similarity between the psychophysical tuning curves (Fig. 12) and neural tuning curves (Fig. 7) suggests that the internal, critical band filters are based on the tuning properties of auditory nerve fibers.

Maskers also mask signals occurring before a pulsed signal (forward masking) and after the signal (backward masking). For a fixed temporal separation between pulsed signals and maskers, there is usually more masking in forward than in backward masking.

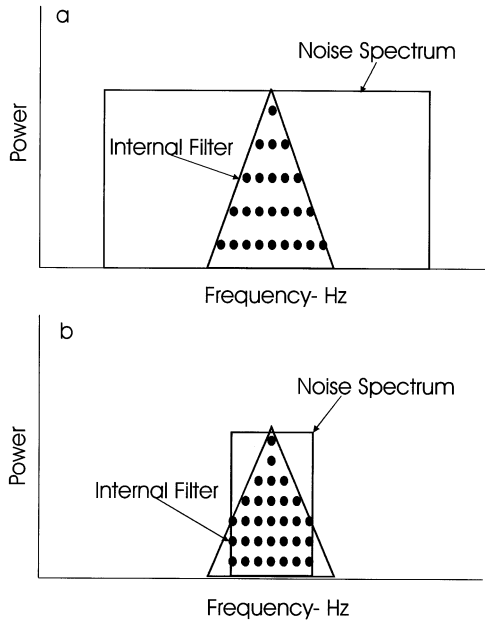


Figure 12 (a) The maximum power comes through the internal filter for a broadband noise resulting in maximal masking. (b) Less power comes through the internal filter for a narrowband noise resulting in less masking.

When the signal is shorter than the masker and when it occurs at the very beginning or the very end of the masker, there is more masking (masking overshoot) than when the brief signal occurs in the temporal middle of the masker.

C. Loudness and Pitch

The physical properties of sound include intensity and frequency. Subjective attributes of sound include loudness and pitch. Although loudness is highly correlated with intensity and pitch is highly correlated with frequency, changes in frequency can cause perceived loudness changes; similarly, changes in intensity can lead to perceptual differences in pitch.

Figure 13 displays equal loudness contours, where each contour represents the level and frequency of tones that are judged equally loud. For instance, the curve labeled 40 phons represents the levels and frequencies of tones that are all equal in loudness to a 40-dB SPL, 1000-Hz tone; the 80-phon curve represents the frequencies and levels required for equal loudness judgments to a 80-dB SPL, 1000-Hz tone. Equal loudness contours are used to define the phon as a decibel measure of loudness level. A tone that is x

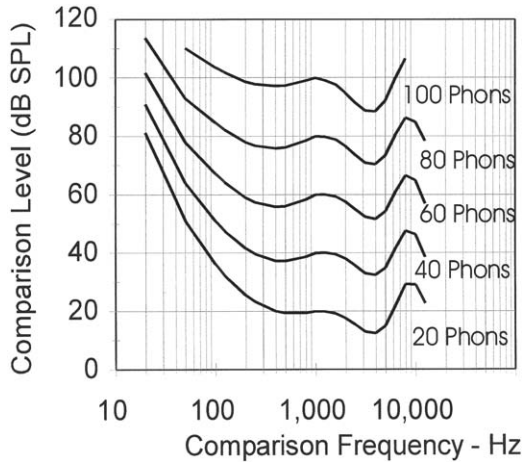


Figure 13 Equal loudness contours showing the level of a comparison tone required to match the perceived loudness of a 1000-Hz standard tone presented at different levels (20, 40, 60, 80, and 100 dB SPL). Each curve is an equal loudness contour. Based on International Standards Organization Standard ISO, R-532 (1981).

phons loud is judged to be equally loud to an x dB SPL, 1000-Hz tone.

Pitch is measured in hertz, and the most often used scales of pitch are musical scales. Musical scales are based on octaves, where an octave is a doubling of frequency (e.g., 440 to 880 Hz is one octave). The octave is divided into 12 equal logarithmic steps (semitones), with each semitone containing 100 cents. Thus, in a musical scale an octave is divided into 1200 logarithmic steps called cents. The musical note scale (A, B, C, D, E, F, G with or without sharps and flats) is also used as a pitch scale.

The perception of pitch exists even when there is no energy in a sound's spectrum at frequencies that correspond to the perceived pitch. These pitches are often referred to as complex, virtual, or missing fundamental pitches. The latter name is derived from the following type of stimulus conditions: a complex sound consisting of six tones added together with equal levels and with frequencies of 400, 500, 600, 700, 800, 900, and 1000 Hz. Independent of the phases of the individual tones, this complex sound has a salient perceived pitch of 100 Hz. Note that all the tones are harmonics of 100 Hz (integer multiples of 100 Hz), but 100 Hz is "missing" (hence the term missing fundamental pitch). Also note that there is no energy in this sound's spectrum in the region of the reported pitch (100 Hz). The temporal waveform of this stimulus can produce an amplitude modulation of 100 Hz, but neither simple measures of the periodicity of the amplitude modulation nor those of the spectral structure of these complex sounds are able to predict complex pitch, suggesting that pitch processing involves more complex operations than simple temporal or spectral mechanisms.

Due to the nonlinear properties of the transduction of sound by the auditory periphery, pitches associated with nonlinear distortion products are often perceived. For instance, a complex sound consisting of 700- and 1000-Hz, equal-amplitude tones may produce pitches in addition to those of 700 and 1000 Hz. In many conditions, listeners also report pitches of 400, 1400, and 2000 Hz for primary stimuli with frequencies of 700 and 1000 Hz. The pitches of 400, 1400, and 2000 Hz are nonlinear distortion products caused by nonlinear peripheral transduction. The 1400- and 2000-Hz pitches are aural harmonics (the second harmonics of 700 and 1000 Hz). The 400-Hz pitch results from the cubic difference tone, which is twice the lower frequency minus the higher frequency ($2f_1 - f_2 = 2 \times 700 - 1000 = 400$). The cubic difference tone is often the most salient distortion product.

Sounds can have many other subjective attributes. The spectral complexity of sound is often correlated with a sound's timbre. Sounds differ in timbre if the sounds are perceived as different even though they have the same perceived loudness, pitch, and duration. Thus, the perceptual difference between a violin and viola playing the same musical note, with the same loudness and duration, is a timbre difference.

V. AUDITORY SCENE ANALYSIS

Previous sections described the perceptual attributes of the sound field but did not discuss the cues used to determine sound sources, especially when there are many sound sources. In addition to each sound source contributing perceptual attributes to the sound field, the spatial location of sound sources can be determine based on sound alone. Clearly, spatial separation could be one cue used to determine sound sources. Other cues include temporal modulation, spectral profiles, harmonicity and temporal regularity, and onsets and offsets.

A. Spatial Hearing

The location of a sound source can be determined entirely on the basis of the sound produced by the source despite the fact that sound does not have spatial properties. A sound source can be located in each plane

in three-dimensional space (azimuth or the left–right plane, vertical or the up-down plane, and distance or the near–far plane). It appears that the auditory system uses different acoustic cues to compute the source's location in each plane.

A sound arriving at a listener's ears from a source lying in the azimuth plane will reach one ear before it reaches the other ear, resulting in an interaural time difference that increases as the sound source is moved further away from midline toward one ear. The sound arriving at the far ear will be less intense than that arriving at the near ear because for a wide range of frequencies the head produces a sound shadow lowering the level at the distal ear. As with the interaural time difference, the interaural level difference also increases as the sound moves in azimuth toward one ear. Thus, interaural time and level differences are the cues used to locate sound sources in the azimuthal plane. At midline, sound sources separated by as little as 1° of visual angle can be discriminated. Interaural time differences as small as 10 microseconds and interaural level differences as small as 0.5 dB can be discriminated. Due to the interaction of sound with the dimensions of the head, the interaural differences are frequency dependent, such that sound sources producing sounds with low frequencies or slow amplitude modulations are located on the basis of interaural time differences, and sounds produced with high frequencies are located on the basis of interaural level differences (the duplex theory of sound localization).

If the head does not move, a sound directly in front of a listener will produce no interaural time and level differences, as would all sounds lying on a plane that runs from in front to directly overhead, directly behind, and directly below (this vertical, midsagital plane is one cone of confusion in which all locations on a cone of confusion produce the same interaural differences). If interaural differences were the only cues used for sound localization, then all sound sources located on a cone of confusion would be perceived at the same location. Although differentiating sound sources located on cones of confusion is more difficult when the head is kept stationary than when it is not, sound sources located at different points on a cone of confusion are perceived at different spatial locations. Thus, cues in addition to interaural differences appear necessary for determining vertical location.

The HRTF described previously occurs because the spectrum of a sound traveling from a source to the tympanic membrane is altered by the body structures (especially the pinna) that the sound must pass over. The spectral alteration and, hence, the spectral characteristics of the HRTF depend on the location of the sound source relative to that of the body and head. The HRTF therefore contains potentially useful information about sound source location. The spectral alterations resulting in the HRTF, especially in high-frequency regions, provide cues for vertical sound localization. In particular, there are HRTF spectral valleys and peaks in regions above 3000 Hz that vary systematically as a function of vertical location. These high-frequency HRTF spectral valleys and peaks are the potential cues for vertical sound location.

Sound location acuity is best in the azimuth plane, poorer in the vertical direction, and even poorer for judgments of distance. The distance of a source can be inferred from loudness differences, assuming one compensates for the other variables that cause a sound's level to change. Distance location is also dependent on the ratio of reflected sound (in reverberant environments) to the amount of sound coming directly from the source. The more relative reflected sound there is, the closer the sound is likely to be.

Most animals locate sounds in reverberant spaces very well, suggesting that the source of the actual sound is not often confused with the location from which reflections (echoes) occur. The sound from the source will reach the listener before that from any reflection because the path of the reflected sound is longer. Thus, our ability to locate sound sources accurately in reverberant environments is most likely due to the earlier arriving sound from the source taking precedent over later arriving reflected sound. Indeed, studies of the precedence effect (or law of the first wavefront) suggest that later arriving reflected sound is suppressed relative to earlier arriving direct sound.

Sounds presented binaurally over headphones are often lateralized within the head as opposed to being localized in the external environment. However, if careful attention is paid to altering the stimuli to reflect the spectral complexity of the HRTF before presenting sounds over headphones, then listeners can perceive virtual sounds in the external environment at locations appropriate for the HRTF-altered stimuli delivered to both headphones.

Segregating different sound sources might be aided by our sound-localization abilities. The ability to determine sound sources based on spatial separation has been referred to as the "cocktail party effect," referring to our ability to attend to one voice at a cocktail party out of many voices and other competing sounds sources. The threshold for detecting a masked signal presented with one set of interaural differences of time and level is lower if the masker has a different

set of interaural differences than if the signal and masker share the same interaural differences. Since interaural differences are used to determine a sound's azimuthal location, the stimulus conditions described previously can be restated: The signal is at one azimuthal position and the masker at another. Thus, the signal is easier to detect if it is at a different location than the masker. The difference in detection thresholds due to interaural differences of time and level between the signal and masker are referred to as binaural masking-level differences (BMLDs or MLDS). MLDs suggest that spatial separation can serve as an aid for sound source determination. However, spatial separation by itself is not a necessary and sufficient condition for sound source determination. The sound of an orchestra recorded by a single microphone played over a single headphone will provide no interaural differences, and yet such a condition interferes little with one's ability to identify the musical instruments (sound sources) in the orchestra.

B. Temporal Modulation

Sounds from most sources vary in amplitude (amplitude modulation) and/or frequency (frequency modulation) over time, and often the modulation of one sound source may differ from that of another sound source. Thus, in some cases the ability to process sound from a source must be resilient to modulations, and in other cases differences in modulation might help segregate one sound source from another. It appears as if frequency modulation per se is not a useful cue for sound source segregation, but amplitude modulation may be.

Much the work on auditory scene analysis has been done in terms of a paradigm referred to as auditory stream fusion or auditory stream segregation. A typical experiment may involve two tones of different frequencies (f_1 and f_2) that are turned on and off (turning a sound on and off is a form of amplitude modulation). Suppose that a tone of frequency f_1 is turned on and off out of phase with a tone of frequency f_2. Thus, the frequencies alternate back and forth between f_1 and f_2. Under some conditions, listeners report that they perceive a single sound source whose pitch is alternating. In other conditions, they report perceiving two separate sound sources, each with its own pulsating pitch, as if there were two auditory streams consisting of the two sources running side by side. The conditions that lead to stream segregation (the perception of two sources) as opposed to stream fusion (one source) provide valuable information about the cues that may be used by the auditory system to sort the various sound sources in a complex auditory scene. Spectral differences are potent for stream segregation, but differences in spatial parameters (e.g., interaural time differences), modulation patterns, and timbre can also support auditory stream segregation.

Sounds that share a common pattern of amplitude modulation (AM) are most likely produced by a single sound source than by different sources. Thus, common AM is a potential cue for sound source determination. When a complex masker contains two spectrally separated masking stimuli (e.g., two narrow bands of noise in different regions of the spectrum), the detection of a signal whose frequency is at the spectral center of one of the maskers is dependent on the similarity of the pattern of AM imposed on the two masking stimuli. If both masking stimuli have the same AM pattern (the maskers are comodulated), then detection threshold is lower than if the AM patterns are not the same (maskers are not comodulated). The difference in detection threshold due to comodulation is called comodulation masking release (CMR), as shown in Fig. 14. Models of CMR often assume that comodulation aids the auditory system in determining the quiet periods in the masker when the signal would be easier to detect.

Another example of the role of common amplitude modulation concerns the detection of a change in the depth of AM (not unlike experiments measuring the TMTF). The ability to detect a change in the depth of AM of a tone with a particular carrier frequency (probe tone) is not changed if a second, unmodulated tone, of a different carrier frequency is added to the probe tone. Since tones of different frequencies do not interfere with each other in masking experiments, this is not a surprising outcome. However, if the two tones are modulated with the same pattern, detection of the depth of AM imposed on the probe tone is more difficult, as shown in Fig. 15. The elevation of the threshold for detecting a change in AM depth due to comodulation is referred to modulation detection interference (MDI). It is as if the common modulation fuses the two tones as being produced by a single source, making it more difficult to process the AM of either component of this single source. If so, making the modulation patterns different should no longer allow the two tones to be fused as if they were produced by a common source and the amount of MDI should decrease, which is what happens as shown on the right of Fig. 15.

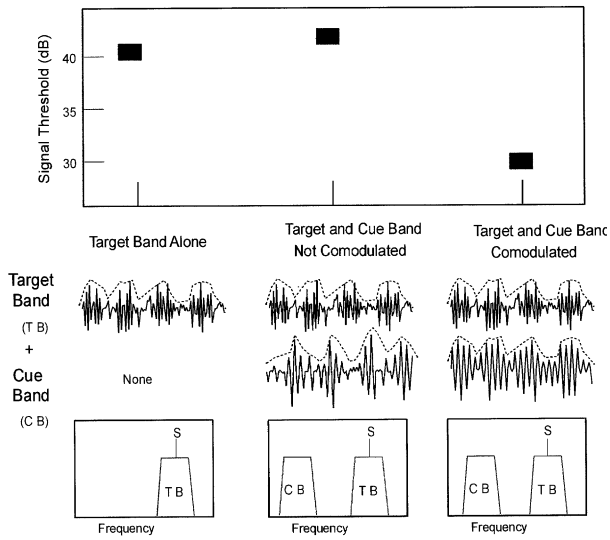


Figure 14 Both the basic CMR task and results are shown. (Bottom) The time domain waveforms for the narrowband maskers (target and cue bands) and the amplitude spectra for the maskers and the signal are shown in a schematic form. The dotted line above each time domain waveform depicts the amplitude modulated envelope of the narrowband noises. The listener is asked to detect a signal (S) which is always added to the target band. In the target band-alone condition, the signal is difficult to detect. When a cue band is added to the target band such that it is located in a different frequency region than the target band and has an amplitude envelope that is different (not comodulated with) from the target band, there is little change in threshold from the target band-alone condition. However, when the target and cue bands are comodulated, the threshold is lowered by approximately 12 dB, indicating that the comodulated condition makes it easier for the listener to detect the signal. The waveforms are not drawn to scale (from Yost, 2000).

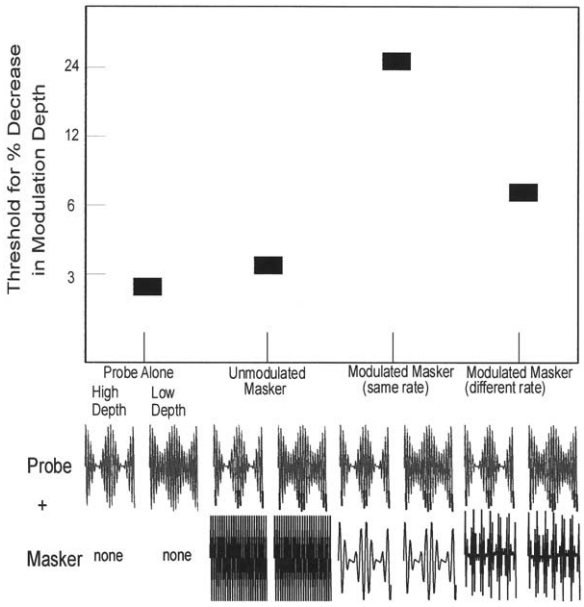


Figure 15 Both the basic MDI task and results are shown. The basic task for the listener is depicted at the bottom. The listener must detect a decrement in the depth of probe amplitude modulation (difference between low and high depth). When just the probes are presented, the task is relatively easy. When an unmodulated masker tone with a frequency different from that of the probe is added to the probes, thresholds for detecting a decrease in probe modulation depth are not changed much from the probe-alone condition. However, when the masker is modulated with the same rate pattern as the probe, the threshold for detecting a decrement in probe modulation depth increases greatly, indicating that modulation depth is difficult to detect when both the probe and the masker are comodulated. When the masker is modulated, but with a different rate (shown as a faster rate in the figure) than that of the probe, the threshold for detecting a modulation depth decrement is lowered. The waveforms are not drawn to scale (from Yost, 2000).

Experiments in CMR and MDI demonstrate the ability of the auditory system to integrate information across a sound's spectrum and experiments in stream segregation describe temporal integration over time. Both spectral and temporal integration are requirements for analyzing complex auditory scenes in the real world.

C. Spectral Profiles

The spectrum of the sound from one source will differ from that of another source. Experiments in spectral profile analysis measure the auditory system's sensitivity to changes in the amplitudes of the spectral components of a complex sound. The schematic diagrams on the bottom of Fig. 16 display different spectra used in a spectral profile experiment. The spectra consist of several tonal components logarithmically spaced in frequency. As can be seen, the center frequency component has a greater level than the flanking components. The listener's task is to discriminate between presentations of spectral profiles in which the center component has a level increment and complex sounds in which all components have the same level. The level of the increment in the center component is adjusted until threshold discrimination is achieved. A key element in these experiments is that the overall level of each complex sound is randomly varied over a very large range. The overall level of randomization makes it difficult for the listener to use the overall change in stimulus level or the change in level of just the level-incremented component (the center component) as the cues for discrimination. The listener appears to use the relative change in level that exists in the complex. That is, the change in level of the center component compared to the other, equal-level

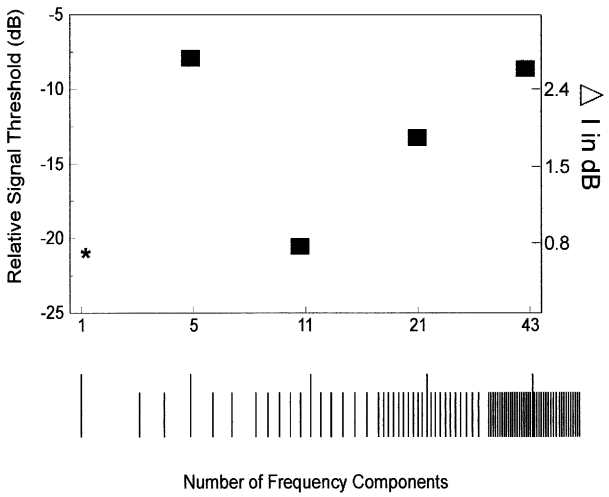


Figure 16 Results from a profile analysis experiment in which the number of masker frequency components surrounding a 1000-Hz signal component increased from 4 to 42. The thresholds for detecting an increment in the 1000-Hz signal component (the center component) are shown in decibels relative to that of the rest of the masker component intensity. The asterisk indicates the typical threshold for detecting a level increment of a single, 1000-Hz tone. The level of the signal is expressed in terms of the signal-to-background level. (from Yost, 2000).

side frequency components is the discrimination cue. As the data in Fig. 16 show, listeners are almost as good at detecting the change in relative level of the center component of the complex sound when there are 11 components as when this center component is presented as a single sound with no flanking components. When there are a few widely spaced components it is more difficult to judge the relative level change. When there are many, densely packed components, discrimination is also difficult, most likely because the components are close enough in frequency to directly mask each other. Studies of profile analysis describe the stimulus conditions that allow the auditory system to detect small changes in spectral shape that would be crucial for sound source determination.

In profile experiments the overall level was randomized. If other forms of randomization are used to make the stimulus conditions vary from trial to trial (e.g., the frequency content of a masker is varied from stimulus to stimulus), there is often considerably more masking than when the uncertainty about the stimuli is small. The extra masking resulting from stimulus uncertainty is referred to as informational masking, distinguishing this form of masking from that which occurs due to the direct interaction of the masker and signal.

D. Harmonicity and Temporal Regularity

Many sound sources vibrate with a fundamental frequency having many harmonics (e.g., voiced speech, musical instrument, and sounds from motors). These harmonically related spectra usually have temporally regular waveforms. As discussed previously, stimuli with harmonic structure or temporal regularity often produce a complex pitch. This unitary pitch can be used to characterize a complex sound. However, it is not always the case that mixing two complex sounds, each with a different fundamental, results in the perception of two pitches that could be used to segregate the two complexes. For instance, a sound source consisting of a 300-Hz fundamental and its harmonics could be mixed with a sound source consisting of a 410-Hz fundamental and its harmonics. In many situations, the auditory system does not analyze the complex mixture into two sources, one with a 300-Hz pitch and one with a 410-Hz pitch. In these cases, the auditory system appears to process the complex mixture synthetically.

In other conditions, a change in the fundamental component of a complex sound can be an aid to sound source segregation. For instance, if two different speech vowels are computer generated so that they have the same fundamental frequency (reflecting the periodicity of the vibrating vocal cords), listeners may have difficulty identifying the two vowels. However, when the two vowels also have different fundamental frequencies, vowel identification is possible. Thus, harmonic relationships and temporal regularity can be cues for sound source determination, but not in all cases.

E. Onsets and Offsets

A powerful cue for segregating one sound source from a mixture of other sounds is temporal asynchrony. If the sound from one source begins or ends at a different time than other sounds from other sources, the differences in onsets or offsets allow listeners to determine the various sound sources. For instance, in the case described previously in which listeners may not perceive two complex pitches (e.g., 300 and 410 Hz) when two different harmonically related stimuli are mixed, listeners report perceiving the two pitches if one complex harmonic stimulus is presented slightly before the other stimulus. Although both stimuli may be played together for most of their duration, a small onset asynchrony can assist complex pitch segregation.

The onset (attack) and offset (decay) characteristics of the playing of many musical instruments are crucial variables for the characteristic timbre of the instruments. These timbrel differences can also aid sound source segregation. The ability to represent the proper attack and decay properties is a key element in musical instrument synthesis.

Thus, the auditory system appears to compare information across the spectrum of sound and over time to process complex sounds so that sound sources can be determined and segregated one from the other. Interaural differences, amplitude modulation, harmonicity, temporal regularity, relative amplitude differences (spectral profiles), and common onsets and offsets are some of the variables that allow the auditory system to determine sound sources in our environment, especially when several sound sources exist at the same time.

See Also the Following Articles

AUDITORY AGNOSIA • AUDITORY CORTEX • BRAINSTEM • HEARING • SENSORY DEPRIVATION • SPEECH • VISION: BRAIN MECHANISMS

Suggested Reading

Blauert, J. (1997). *Spatial Hearing*. MIT Press, Cambridge MA.

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, Cambridge, MA.

Dallos, P., Popper, A. N., and Fay, R. R. (Eds.) (1996). *The Cochlea*. Springer-Verlag, New York.

Fay, R. R., and Popper, A. N. (Eds.) (1992). *The Auditory Pathway: Neurophysiology*. Springer-Verlag, New York.

Hartmann, W. M. (1998). *Signal, Sounds and Sensation*. Springer-Verlag, New York.

Moore, B. C. J. (1997). *An Introduction to the Psychology of Hearing*, 3rd ed. Academic Press, London.

Pickles, J. O. (1988). *An Introduction to the Physiology of Hearing*, 2nd ed. Academic Press, London.

Webster, D., Fay, R. R., and Popper, A. N. (Eds.) (1992). *The Auditory Pathway: Neuroanatomy*. Springer-Verlag, New York.

Yost, W. A. (2000). *Fundamentals of Hearing: An Introduction*, 4th ed. Academic Press, New York.

Yost, W. A., and Gourevitch, G. (Eds.) (1987). *Directional Hearing*. Springer-Verlag, New York.

Yost, W. A., Popper, A. N., and Fay, R. R. (Eds.) (1993). *Human Psychoacoustics*. Springer-Verlag, New York.

Zwicker, E., and Fastl, H. (1991). *Psychoacoustics: Facts and Models*. Springer-Verlag, Berlin.

Autism

ERIC COURCHESNE*,† and KAREN PIERCE*

*University of California, San Diego and †Children's Hospital Research Center

GLOSSARY

cerebellum Also known as the "little brain," the cerebellum is divided by the primary fissure into anterior and posterior lobes. Other fissures further segment these lobes into several smaller lobules. A thin longitudinal strip traverses the midline and is known as the vermis. The cerebellar cortex contains several different neuron types, including Purkinje neurons, the sole output source. Functionally, the cerebellum has been traditionally viewed as involved only in motor control; however, a new understanding has come to include multiple cognitive functions, such as language and attention.

chromosomes Long thread-like associations of genes found in the nucleus of all eukaryotic cells. Chromosomes consist of DNA and protein.

event-related potential technique A neuroimaging method that records electrical activity from the scalp via electrodes as an individual performs a task. Signal amplitudes and latencies are believed to represent net electrical fields associated with the activity of a population of neurons.

functional magnetic resonance imaging A neuroimaging technique that uses the differing magnetic susceptibility of oxyhemoglobin and deoxyhemoglobin in the blood to detect when and where oxygen is used in the brain, typically as an individual performs a cognitive task.

gene A sequence of DNA that indirectly produces a protein and is located on chromosomes.

genetic linkage analysis with disorder loci The preferential association of a particular gene, or set of genes, with a particular disorder than would be expected by chance alone.

gliosis The presence of glial cells in unexpected locations in neuronal tissue usually as the result of injury or insult.

limbic system A constellation of cortical and subcortical brain structures thought to be instrumental in memory and emotion. Traditionally, the limbic lobe includes the parahippocampal gyrus, the cingulate gyrus, and the subcallosal gyrus; subcortical structures include the hippocampus, amygdala, and parts of the hypothalamus.

magnetic resonance imaging A noninvasive technique used to produce high-resolution images of body tissues, especially soft tissues such as the brain. A large magnet is used to polarize hydrogen atoms in the tissues and radio frequencies are used to measure the nuclear magnetic resonance of the protons within the tissue.

pleiotropy The condition in which a single gene or gene pair exerts effects on multiple functions and/or in multiple cell types often at different times.

repetitive behaviors and stereotyped patterns Interests, motor movements (e.g., arm flapping and rocking), and rituals that are performed repetitively and inflexibly and do not serve any apparent function.

Autism is a biological disorder with a clinical onset in the first years of life that persists, to varying degrees, throughout life. It is characterized by abnormality in reciprocal social interaction, communication, and language development as well as by repetitive and stereotyped behavior. These abnormalities are caused by defects in multiple areas of the brain. The prevalence is 1:600, making it one of the most common neurobiological disorders of infancy and early childhood. It is one of five pervasive developmental disorders (PDDs), along with Rett's syndrome, Asperger's syndrome, child-

hood disintegrative disorder, and pervasive developmental disorder not otherwise specified (PDD-NOS), listed in the *Diagnostic and Statistical Manual IV*. Taken together, the prevalence of a PDD is 1:160. The degree to which autism, Asperger's, and PDD-NOS disorders with behavioral profiles that overlap, may be variants on a single biological theme, or constitute unique etiologies, is still under debate.

I. INTRODUCTION

Leo Kanner first described the developmental disorder known as autism more than 50 years ago. Although his central observations of social aloofness, linguistic abnormalities, and restricted interests in individuals with this disorder are still valid, a new understanding of autism has come to include the idea of heterogeneity; differences in symptom profile and severity, onset, collateral conditions, and potential etiologies as well as developmental course and outcome vary considerably both within and across individuals with this disorder. It is this heterogeneity that not only makes autism an interesting and complex disorder but also provides clues to an appropriate scientific path for investigating this disorder. Tying this work together has been the relatively new use of structural magnetic resonance imaging (MRI) and functional neuroimaging technologies [functional MRI (fMRI), positron emission tomography (PET), and event-related potential (ERP)] that afford insights into the neurobiology that mediates the complex symptom profile seen in affected individuals. Data from structural *in vivo* MRI and autopsy studies support the idea that autism is a disorder with multiple loci of neuroanatomic abnormality; some individuals show abnormalities in the amygdala, frontal lobes, parietal lobes, hippocampus, corpus callosum, and brain stem, whereas others do not. Several brain regions, such as temporal and occipital lobes, have yet to be thoroughly investigated. One structure, the cerebellum, has been shown to be abnormal in 95% of the cases studied at autopsy, a consistency not found with any other structure. However, cerebellar defects cannot singly account for the diverse symptoms seen in autism but, rather, may act in concert with dysfunction in other neural systems. It is the idea of multiple damaged systems operating in a behaviorally and biologically heterogeneous disorder that forms the foundation for the discussion of autism that follows.

II. CLINICAL ONSET AND DIAGNOSIS

A. Characteristics in the First Years and Clinical Onset

Despite the almost complete consensus that autism is a disorder with a biological onset prenatally or shortly after birth, behavioral indicators during the first 2 years of life are elusive. Difficulties in developing a clinical profile of autism at such young ages are due to the fact that most children do not receive a diagnosis until age 2–4; therefore, reliable early indicators may be missed. Some information, however, has been made available based on retrospective analyses of home videos taken of children before the diagnosis of autism has been made. Such retrospective videotaped studies have demonstrated that autism can be distinguished from normal at least as early as 1 year based on differences in social behaviors, such as looking at the face of another or orienting in response to his or her name, and in joint attention behaviors, such as pointing, showing objects to others, and alternating gaze between object and person. Given that these types of social and joint attention behaviors may not yet be present or are difficult to reliably assess in young infants (e.g., 3 months of age), scientists are now searching for indicators for infants at risk for autism. It is well-known that the cerebellum is critical for motor control and, combined with the finding that the majority of individuals have cerebellar abnormalities, it has been suggested that one potential early indicator may relate to observable motor executions. Investigations of motor patterns can be easily assessed in newborns and infants, and new research in fact suggests that abnormalities in motor behavior may be seen as early as 4 months of age in infants later diagnosed as autistic. Specifically, deviancies in major motor milestones, such as lying, righting, sitting, crawling, and walking, have been reported based on precise movement analysis systems (e.g., Eshkol–Wachman). Although only one study has utilized movement analyses, results indicated that all autistic infants studied showed some form of movement abnormality. It is quite possible that early behavioral indicators of autism may be found in basic perceptual and motor systems.

1. Neurobiological Findings and Hypotheses

Currently, the strongest evidence addressing the question of the time of biological onset comes from

postmortem and MRI evidence indicating that cerebellar and brain stem anatomical abnormalities most likely have a prenatal onset. Reduction in Purkinje neuron numbers is typically present without accompanying gliosis, a finding indicative of an early (prenatal or early postnatal) onset for this defect. In the first prospective MRI study of autism, Hashimoto and colleagues found cerebellar vermis hypoplasia in infants who were subsequently confirmed to be autistic. Regression analyses of growth curves for the neocerebellar vermis in autistic patients indicate a perinatal or prenatal onset of the vermis hypoplasia. Finally, evidence from a single postmortem autism case has been interpreted as an indication that maldevelopment in the brain stem may begin as early as the first trimester.

B. Diagnosis

Although biological abnormalities are becoming increasingly better understood in autism, currently this disorder is diagnosed exclusively based on behavioral characteristics. Traditionally, the *Diagnostic and Statistical Manual IV* (*DSM-IV*) has been the guide used for making diagnostic decisions about the presence or absence of autism. A burgeoning interest in early identification in autism, however, has prompted the need for instruments that will both reliably and validly diagnose autism at the youngest ages possible. The ability for the *DSM-IV* to meet this need has been questioned by many scientists. For example, part of the *DSM-IV* criteria relates to language abnormalities, and given that language is minimal at young ages (for both autistic and normal children), children suspected of having autism could not be reliably diagnosed with this instrument. In order to meet this need, a relatively new diagnostic tool, the Pre-Linguistic Autism Diagnostic Observation Schedule (PL-ADOS) developed by Catherine Lord and colleagues, has been utilized with children suspected of having autism and can be used to reliably diagnose autism as young as 18 months of age. The PL-ADOS is distinct from the *DSM-IV* because it focuses on behaviors that are indicative of young children with the disorder, such as abnormalities in gesture or joint attention.

Early identification in autism is essential not only for early treatment but also for investigations of neurodevelopmental processes in autism. The overwhelming abundance of past research has concentrated predominantly on samples of adults with autism. The use of diagnostic tools with toddlers suspected of having autism, such as the PL-ADOS, for the first time allows researchers not only the ability to obtain information at young ages in the disorder but also, the opportunity for longitudinal investigations of behavioral and neurobiological growth and development, beginning at a period when brain development in autism is not only quite active but also undiscovered.

III. SYMPTOM PROFILE AND NEURAL BASES

A. Social Abnormalities

1. Characteristics

When Leo Kanner was deciding on a name that might most aptly characterize his original 11 cases, he chose the term autism, derived from the Greek work ''autos'' or self. His choice likely emerged from the strong sense of a preference for isolation, or the self, to the company of others observed in all his cases.

Indicators of social abnormalities are noted early in autism, quite often as early as the first few months of life. Parents often remark that as infants, their autistic children stiffened when held, failed to make prolonged eye contact, and did not cry for help or consolation. A robust early indicator of social abnormality is a near absence of joint attention, or the coordination of attention between object and person, a skill that emerges in normal infants toward the end of the first year of life. As a category, joint attention behaviors may include pointing and showing objects and the use of eye contact and direction of visual gaze to and from an object and person. Normal infants and toddlers use such gestures regularly. Autistic infants and toddlers, on the other hand, rarely use such gestures and, given that joint attention is considered a precursor to normal language development, it is not surprising that this deficit compounds into not only later deficits in social behavior but also severe deficits in language development.

As the autistic infant matures, social abnormalities become even more apparent and include decreased rates of eye contact, difficulties in expressing and identifying emotion, lack of social initiations, lack of maintaining interactions with others, and lack of an ability to demonstrate empathy. Early studies of social abnormality in autism were largely quantitative: Children with autism were shown to initiate interactions, play with peers, make eye contact, or maintain prolonged interactions with others less than normal

children. Current investigations of social behavior have not only refined and brought greater detail to our understanding of the nuances involved in social dysfunction in autism but also aim to test psychological theories or hypotheses about the role that such abnormalities might play in the course of the disorder.

One such example comes from the idea that individuals with this disorder either lack or have an underdeveloped "theory of mind" (TOM), or an inability to develop mental representations for the contents of other people's minds. By age 4, normal children can typically correctly perform simple TOM tasks, whereas most autistic individuals, with mental ages at or above 4 years, cannot, suggesting that this deficit may not be due solely to general developmental delay. Social intelligence, or TOM, comprises our ability to interpret others' mental states (thoughts, desires, beliefs, and intentions), to use this information to interact in groups, to empathize, and to predict others' thoughts and behaviors. The TOM hypothesis of autism, posited by Simon Baron-Cohen, speculates that not only is this deficit central to the symptomatology of autism and may explain social, communicative, and imaginative abnormalities but also it may be independent from general intelligence.

2. Brain–Behavior Findings

Social dysfunction in autism is severe and persistent and likely involves abnormalities in multiple neural regions, including those involved in face perception, emotion processing, orienting to social cues, and attention regulation. New functional neuroimaging studies provide the first evidence that one brain region traditionally involved in face perception (i.e., fusiform gyrus), an important constituent of social behavior, may have abnormally reduced levels of functional activation in autism (Fig. 1). Furthermore, it appears that activation to faces may instead aberrantly occur in neighboring regions or atypical locations within the fusiform gyrus; the reason for such functional relocation is unclear. One possibility is that limited exposure to faces in patients with autism (perhaps due to innate preference, biases of processing style, or learning) results in an underdevelopment or maldevelopment of face processing systems. Interestingly, recent research has reported increased fusiform gyrus activation after normal subjects were trained to be experts in identifying a novel set of nonsense objects, suggesting that the fusiform gyrus may be an area involved in the processing of extremely familiar objects or objects associated with a particular area of expertise, such as faces. In autism, limited social interactions, and eye contact in particular, could then explain decreased rates of fusiform activation. If so, early intervention strategies could be developed to help autistic toddlers and children acquire the necessary "expertise" for more normal utilization of this cortical region.

The amygdala is necessary to recognize the emotion of fear and blends of emotions, information critical to engaging in successful social interchanges. Anatomic abnormality (increased neuron density) of the amygdala has been described by one research group in nine postmortem autism cases, and two *in vivo* MRI studies have reported reduced amygdala volume in a small sample of adolescent patients. Although not conclusive, such initial evidence of structural defect in the amygdala sets the stage for future anatomic research and functional neuroimaging studies aimed at confirming and further elaborating its role in emotional processing deficits in autism.

Anatomic and functional defects in a third structure, the cerebellum, may contribute to deficits in orienting to social cues and regulating attention during rapid social interchanges when the source of critical social, emotional, and linguistic information shifts. Studies of autistic children show that they orient more slowly than normal to social as well as nonsocial stimuli and that the greater the impairment, the greater the cerebellar anatomical abnormality as measured by MRI. fMRI and ERP studies also show that the cerebellum plays a role in the rapidly shifting attention between different sources of sensory information and that, like patients with acquired cerebellar lesions from stroke or tumor, autistic children and adults are slow to shift attention and miss information as a consequence. Moreover, autistic patients, as well as patients with acquired cerebellar lesions, are missing at least two neurophysiological responses associated with redirecting attention—one that signals frontal cortex to orient (Fig. 2) and another that may signal posterior cortical structures to disengage and redirect attention to a new spatial locus or sensory modality.

An important caveat to both current and future fMRI research on social abnormalities in autism is that individuals with autism have less social experience than their normal peers, thus making claims about the functioning of neural regions difficult to interpret. Furthermore, tests of social intelligence frequently have complex attentional demands, yet another area of considerable difficulty for individuals with autism. Nonetheless, because social functions are markedly maldeveloped in autism, brain–behavior studies of this

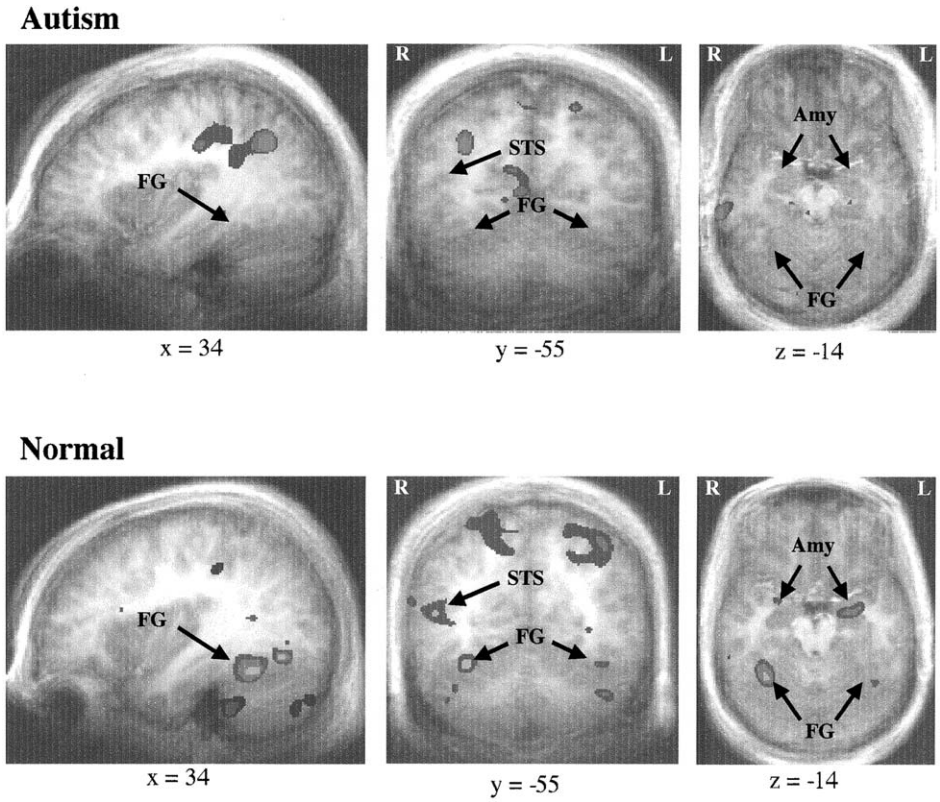


Figure 1 Within group t-maps for both autism and normal groups showing significant regions of activation. Note fusiform, superior temporal sulcus and amygdala activation in normals, in comparison to a lack of positive activation in the autism group. Decreased functional activity in the autism group is likely due to the inconsistent patterns of activation noted across individual autism subjects, which would fail to be seen when results are averaged (from Pierce *et al.*, 2001). (See color insert).

disorder promise to provide special insights into the neural systems underlying human social development.

B. Language and Speech Abnormalities

1. Characteristics

It is the delay or regression of language, usually at approximately age 2, that first prompts parents to seek professional help. Although such delays are significant indicators of problems in child development, they are not specific to autism and are commonly found in many other childhood disorders (e.g., general language delay). By itself, language delay does not account for most of the features of the disorder but can be used as a metric for predicting developmental outcome. For example, several studies have shown that autistic children without speech by age 5 are much less likely to live independently later in life than children with functional speech. Also, the presence or absence of language has often been used to distinguish "high" from "low" functioning autism. The understanding of language abnormalities in autism, however, is critical for the development of appropriate treatment approaches, offers insights into the relationship between various symptoms in the disorder (e.g., between play and language skills), and is an excellent domain for fMRI studies investigating neurofunctional organization in autism.

It is commonly noted that 50% of individuals with autism fail to develop functional speech. For those that do, speech is often characterized as rigid and stereotyped, in which a word or phrase is used in limited contexts and verbal routines may be used to serve the individual's communicative needs. More specific aspects of language, such as syntax, have been examined by investigating the presence of grammatical morphemes during continuous speech. Comparing children with autism to those of similar ages, syntax is

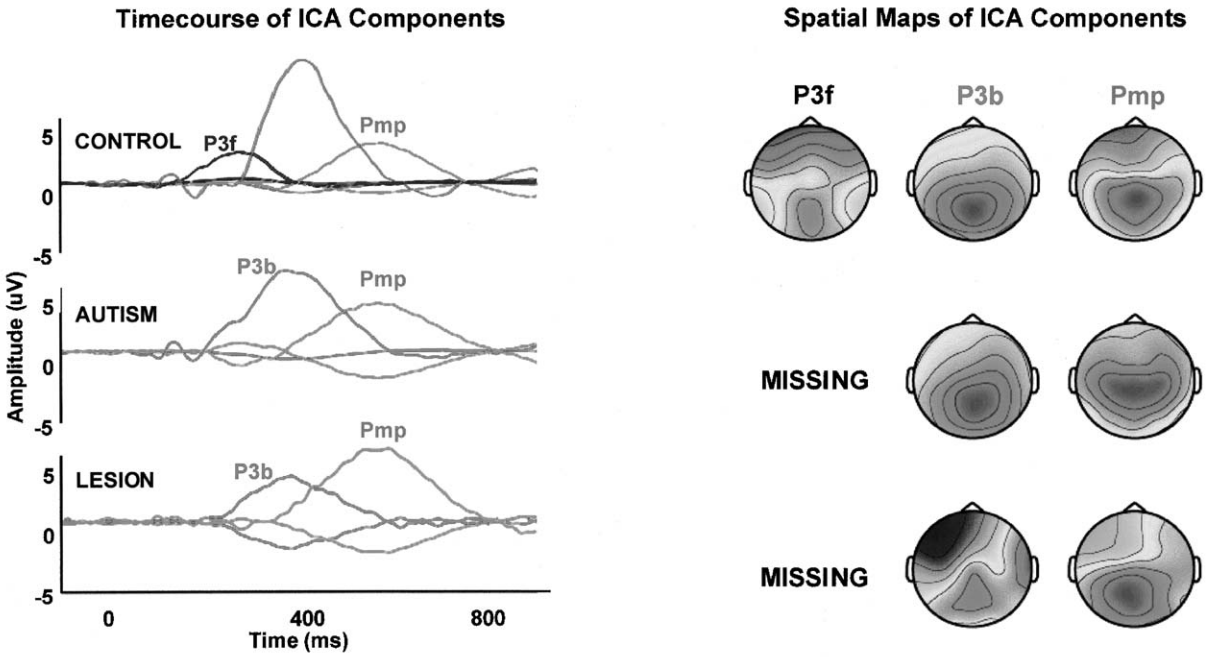


Figure 2 ICA decomposition waveforms and source maps for control, autistic, and cerebellar groups. ICA analyses were performed on grand average ERPs during a spatial attention task. ICA identified the P3f (P3 frontal) occurring at approximately 300 msec, the P3b occurring at approximately 400 msec, and the Pmp (postmotor potenential) occurring at approximately 500 msec for the normal control subjects, but it identified the absence of the P3f for autistic and cerebellar groups (from Westerfield, M., Townsend, J., and Courchesne, E., unpublished data).

severely deviant. Comparisons with mental age matches, however, suggest minimal differences between groups.

Autism is also characterized by several unusual applications of language, such as the repetition of words previously heard (either immediate or delayed) —a characteristic known as "echolalia." For example, an autistic child might repeat a phrase heard earlier on television, such as "You are the next contestant on the Price is Right!" several times throughout the day. Research has suggested that this unique form of communication may in fact serve very specific functions, such as such signifying confusion or requesting clarification. One study found that children with autism were more likely to engage in immediate echolalia in response to questions or commands they did not understand. Once the children learned the correct answer, rates of echolalia dramatically decreased. Other interesting aspects of language use in autism include pronoun reversals (e.g., saying "you want cookies" instead of "I want cookies") and the use of neologisms or nonsensical words. Combined with such unique uses of language are the often noted patterns of odd intonation and tone, usually in the form of monotony.

There is evidence that communication deficits in autism are related to the failure of preverbal mechanisms thought to be important to the development of language pragmatics, such as joint attention described previously. Furthermore, a strong relationship exists between social nonverbal communication and the level of language development in autism. That is, children with autism who display signs of nonverbal communication behaviors during social interactions (e.g., gesture and gaze) develop more receptive language than those that do not. Interestingly, however, the fluid and persistent use of the pragmatics of language persist as an enduring problem, even for autistic individuals who do develop functional speech.

C. Brain–Behavior Findings

There is inconclusive evidence from electrophysiological and PET blood flow studies suggesting that

language deficits in autism may be associated with a lack or reversal of left hemisphere dominance found in most normal adults. However, abnormal asymmetries have been observed in other electroencephalogram, ERP, PET, and SPECT studies that did not involve language tasks. It therefore remains open whether there is any specific link between functional asymmetries and language impairment in autism.

D. Restricted Range of Interests: Repetitive Behaviors and Stereotyped Patterns

1. Characteristics

In addition to salient disturbances in social behavior and language, individuals with autism may also exhibit abnormally intense preoccupations with one subject or activity, exhibit distress over changes in routines, insist on sameness in their environment, and engage in repetitive or ritualistic behaviors. For example, a child may exhibit signs of physical and emotional distress when a new piece of furniture is introduced into the home or if he or she is prevented from taking a traditional route up and down the aisles during grocery shopping. Still another child may center the majority of his or her activities around an obsessional object, such as toilets; he or she may draw pictures of them, talk about them, and flush all toilets in proximity as part of his or her daily life. Finally, others may repetitively bang a spoon or flap their hands for minutes, sometimes hours, during a single day. Such restricted and repetitive behavior may have particularly important developmental consequences for young children with this disorder because it may lead them to miss learning opportunities that fall outside their narrow scope of interest.

2. Brain–Behavior Findings

A novel environment offers an important opportunity for young children to explore and learn. However, when placed in a novel environment, autistic children typically fail to explore or explore only a fraction of available stimuli and this limited exploration is correlated with anatomic abnormality of cerebellar vermis lobules VI and VII (Fig. 3). Additionally, it appears that the more perseverative and stereotyped behavior shown by an autistic child, the greater anatomic abnormality of cerebellar vermis lobules VI and VII.

The correlation between deficits in exploratory behavior and cerebellar anatomic abnormality in autistic children is consistent with a large body of evidence from animal studies. Studies of mice with mutations causing cerebellar anatomic pathology show a reduced tendency to explore novel objects or environments and an increased tendency to move about perseveratively in testing environments. Hypoplasia of vermis lobules VI and VII has been specifically linked to abnormalities in exploratory behavior in two mutant animals, the GS guinea pig and the L1CAM knockout mouse. Potentially, defects in vermis lobules VI and VII may be sufficient to cause deficits in exploratory behavior. Nonetheless, since recent functional neuroimaging studies show that different cerebellar regions have different functional specializations, systematic studies will need to be conducted to learn whether and how various cerebellar regions, in addition to the vermis, play a role in deficits in exploratory behavior in autism. Studies of humans also indicate a role for the cerebellum in exploratory behavior. For instance, humans with acquired cerebellar lesions are inefficient in exploring and learning possible solutions to novel visuospatial puzzles. Interestingly, PET studies show the human cerebellum to be active during mental or ''imagined'' exploration of a spatial configuration.

E. Attention

1. Characteristics

In addition to the severe social abnormalities noted by Kanner, descriptions of attentional disturbances are also prominent features of this first paper on autism. Kanner quotes a father's description of his autistic son: ''He displayed an abstraction of mind which made him perfectly oblivious to everything about him … and to get his attention almost requires one to break down a mental barrier between his inner consciousness and the outside world.'' The first experimental demonstration of overly focused attention or ''stimulus overselectivity'' in the autistic child was by Lovaas, Koegel, and Schreibman in the early 1970s; they showed that autistic children respond to a restricted range of environmental stimuli, suggesting they may miss critical social and nonsocial information. Impairments in attention are now known to be among the most consistently reported type of cognitive deficit in this disorder, whether observations come from parents, teachers, clinicians, or researchers. Also, among cognitive deficits in autism, attention deficits are one

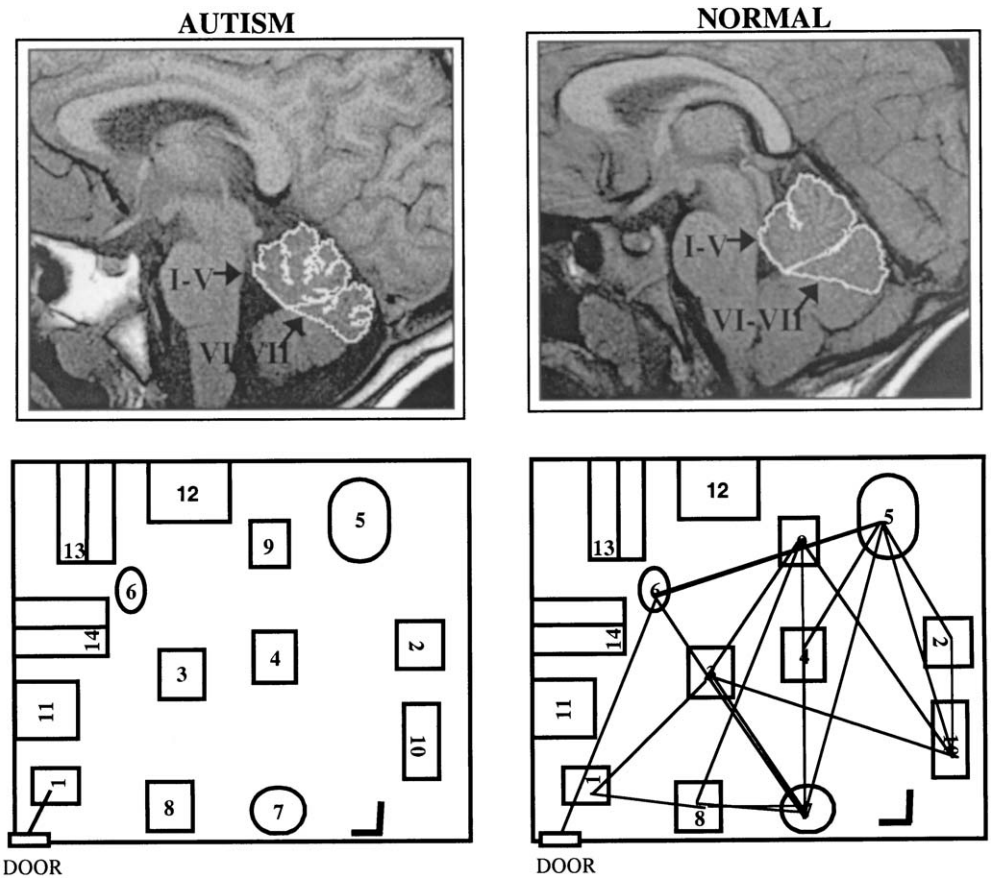


Figure 3 (Top) T1-weighted magnetic resonance image showing midsagittal vermal lobules I–V and VI–VII for a 6-year-old autistic boy and his matched normal control. Note sulcal widening and hypoplasia in the autistic child in comparison to his matched control. (Bottom) Exploration map for the autistic and matched control subject (cerebella for each child shown in the top portion). Lines depict patterns of movement across the entire session. Note the varied exploration pattern of the normal child in comparison to the near absent exploration of the autistic child (reproduced with permission from Pierce and Courchesne, 2001. Evidence for a cerebellar role in reduced exploration and stereotyped behavior in autism. *Biological Psychiatry* **49,** 655–664. Copyright 2001 Society of Biological Psychiatry).

of the most thoroughly examined by anatomical and functional neuroimaging technologies.

Attention abnormalities may contribute to clinical features in autism. As described earlier, a cardinal feature of autism is failure to engage in joint social attention and it is among the first striking deficits noticed in the autistic infant. In normal development, from joint social interactions between infants and mothers spring social knowledge and many higher cognitive, affective, and communicative functions. Tronick writes "successful regulation of joint interchanges … results in normal [cognitive, affective, and social] development. The crucial element is that the infant and mother … share the same focus of attention during the interaction." To do so, an infant must do more than focus his or her attention on a single, captivating aspect of an object or person: The infant must follow the rapid and unpredictable ebb and flow of human social activity, such as words, gestures, touching, postures, facial expressions, and actions on objects. By being able to smoothly, selectively, and rapidly adjust attention, the normal infant is able to combine, as a single entity in memory, the various and separate elements of a social situation. Abnormalities in adjusting attention would place the autistic child at a severe disadvantage in such rapidly changing social situations, and, in conjunction with severe abnormality in social and emotional systems, would add an additional major impediment to engaging effectively in joint social interchanges.

The regulation of attention is likewise important in the adaptive and active exploring of novel environ-

ments, acquiring a wide array of nonsocial knowledge, understanding causal relationships, appreciating the context of discrete events or stimuli, learning multimodality sequential stimulus relationships, and observing relationships among different sensory attributes of objects, people, or spatial arrays. Moreover, the inability to attentionally anticipate or react in a flexible moment-to-moment manner may, in combination with other impairments, increase the autistic child's tendency to remain restricted in interests and activity.

As described next, behavioral performance during tasks requiring rapid and dynamic adjustments in the direction of attention is clearly impaired in autism, but performance during tasks requiring simple sustained or selective attention is often reported to be normal or even supernormal in some autistic patients. One interpretation is that these latter attention functions may be spared, reflecting once again the variability in which brain functions are affected and which are not. Recent knowledge gained from neuroimaging studies, however, has revealed previously unsuspected abnormality in neural functions that mediate these "normal" or "supernormal" attention performances and thereby raises the important issue of neural compensatory reactions in the face of developmental neural defect.

2. Brain–Behavior Findings

Studies of the brain bases of attention functions in autism provide a classic example that normal behavior does not necessarily indicate normal neural function.

a. Sustained and Selective Attention In simple tasks requiring an autistic child to sustain his or her attention to a stream of stimuli (e.g., pictures and shapes) presented one at a time in order to detect a target stimulus, performance can be near normal or normal. More interesting are reports that during some tasks requiring selective attention, such as detecting a specific target stimulus at one of many possible spatial locations, accuracy and speed of pressing a button to the target might even exceed normal. Such behavioral performance, however, belies the presence of underlying abnormal neurophysiological responses to these stimuli.

Thus, although autistic patients seem to be attending and behaviorally performing normally, the stream of stimuli and the targets either fail to elicit attention-related ERP brain responses, such as the Nd, N1, N270, or Nc, or elicit abnormally reduced responses, such as the auditory P3b. Often such brain responses are also atypically located. Moreover, like patients with acquired lesions of the cerebellum, autistic patients performing such attention tasks fail to produce a cerebellar-dependent neural response that may signal frontal cortex (Fig. 2). Also, in cerebellar and autistic patients the attention-related P3b response is abnormal. The P3b is cerebral cortical activity reflecting recognition of task-relevant information and updating of the stimulus context. Unlike the normal brain that produces this response reliably to each single target stimulus, in these patients the P3b response occurs intermittently and with highly variable amplitude and latency to target information. In fact, the challenge has been to find attention-related ERP brain responses that act normally in normally performing autistic patients.

The large disparity between the appearance of normal and intact behavioral performance and the reality of abnormal underlying neural functional activity remains to be experimentally explained. One possibility is that tasks used to date fail to be sufficiently demanding on the neural system supporting it, and properly demanding conditions would reveal behavior deficits consistent with abnormal neural activity. Such a possibility is lent credence by numerous examples of resolution of analogous disparities in studies of other types of patients. Another possibility is that systems mediating these particular attention functions are indeed spared, and the abnormal neural responses recording during such tasks reflect other aberrant operations whose responsibility is to effectively process the information that sustained attention mechanisms "bring in." A third possibility is that the abnormal neural responses reflect the result of successful compensatory brain reorganization in the face of developmental neural damage.

The first possibility is supported by evidence that adds an anatomic twist. Autistic patients with parietal volume loss appear to have a narrow, spotlight-like attention. Therefore, when their task is to detect target stimuli at only one specific location and ignore other nearby stimuli, they behaviorally respond faster to targets than normal, have shorter latency P3b responses to targets than normal, and have larger than normal visual sensory P1 responses to targets. Conversely, when their task is to make a difficult discrimination and they are led to expect it will occur in one location but instead the stimulus occurs in another, they are slower and more inaccurate, missing discrimination opportunities. For other autistic patients who do not have parietal volume loss, the reverse

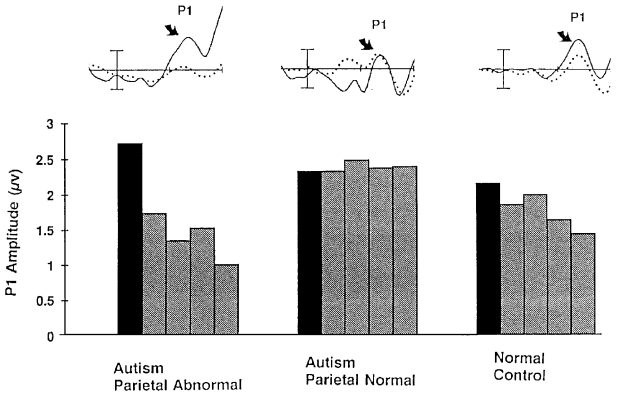


Figure 4 ERP evidence that autistic patients with parietal lobe volume loss have an extremely narrow "spotlight" of visual–spatial attention. Bar graphs are shown for autistic patients with and without parietal abnormalities and normal control subjects. The figure shows the P1 peak amplitude ERP response at occipital scalp sites to visual stimuli at an attended location (dark bar) compared to the P1 ERP responses at that location when attention was focused one, two, three, and four locations away (lighter bars in order left to right). Visual–spatial locations were separated by 2.7° of visual angle. Waveforms at the top show P1 ERP responses at occipital sites to attended locations (solid line) compared to the average of all unattended (dashed line) locations (reproduced with permission from Townsend and Courchesne, 1994; Parietal damage and narrow "spotlight" spatial attention. *J. Cog. Neurosci.* **6,** 220–232, Copyright 1994 by the Society for Neuroscience).

is seen in these tasks with opposite demands for either narrowly or broadly focusing attention (Fig. 4).

Therefore, in autism, it may be that the underlying type and extent of anatomic defect interacts with selective attention demands, such that certain defects may allow or even facilitate successful performance under certain attention conditions but not under others, but entire neural attention functions such as selective or sustained attention are not uniformly spared per se.

b. Dynamic Regulation of Attention: Disengaging, Orienting, and Shifting Even though the autistic child sometimes seems able to sustain attention to something in his or her environment, he or she typically seems impaired in adjusting attention to follow unexpected and rapid changes in the source of information. Especially demanding are the unexpected and quick changes in verbal, gestural, postural, tactile, and facial cues that signal changes regarding where to direct attention in a stream of social information. Much information comes and goes (a momentary facial expression) or moves from place to place (different speakers in a room). Neuroimaging studies of normal subjects show that the rapid, dynamic regulation of attentional resources in demanding attention conditions involves a distributed neural network including the parietal and frontal lobes, thalamus, colliculus, and cerebellum. Multiple operations underlie such regulation utilizing different as well as overlapping neural components.

In autism, three such operations have been thoroughly studied via neuroimaging techniques. First, disengaging attention is a reactive operation implemented when important stimulus information unexpectedly occurs outside the immediate focus of attention. The parietal cortex is an essential neural component in this operation. Autistic patients with parietal cortical volume loss are slow to disengage attention, and the more loss, the slower the implementation of the disengagement so that in patients with the most loss, by the time attention has been disengaged information needed to make critical sensory discriminations has often already disappeared. ERP imaging shows that in patients with parietal cortical loss, there is an abnormally narrow distribution of visual–spatial attention such that neurophysiological responses to stimuli at the center of attention are hypernormal in size while those outside this center are subnormal in size. This suggests that important stimuli unexpectedly occurring outside the immediate focus of attention are less able to rapidly and effectively trigger the disengage mechanisms. Therefore, in autism, parietal structural defects may impair the formation of effective attentional sensory maps of extrapersonal space, impair the operation of mechanisms that implement disengagement of attention, or both.

Second, when attention is not yet engaged and a stimulus occurs that signals the likely location of upcoming important information, attentional resources are normally immediately applied to that location in anticipation of the upcoming information. Successful anticipatory orienting increases the speed and accuracy of information processing of the upcoming information. Patients with cerebellar lesions from stroke or tumor and autistic patients are slow and inaccurate in such cued, anticipatory orienting of attention, and in the autistic child or adult the greater the cerebellar hypoplasia, the slower and more inaccurate the anticipatory orienting (Fig. 5). This effect has also been demonstrated for orienting to social and nonsocial stimuli.

Third, very often a particular stream of information will reach a point when it provides specific information or cues that explicitly direct or signal that attention is

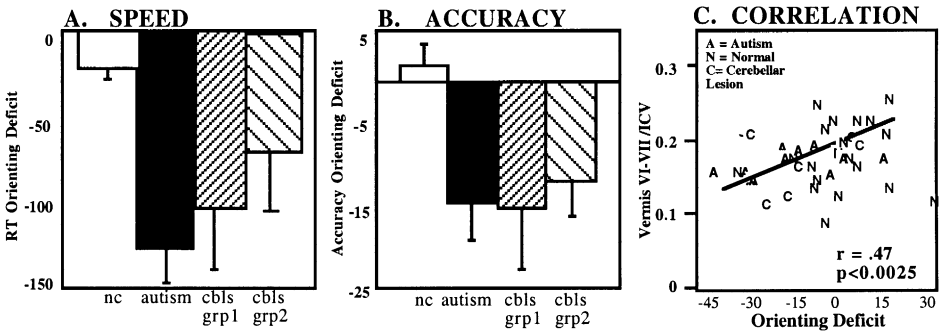


Figure 5 Orienting deficits (RT or accuracy at validly cued location at long cue-to-target delay and at short cue-to-target delay). A more negative orienting deficit indicates slower response or decreased accuracy to targets at cued locations when there is little time between cue and target onset. (A) Autism and cerebellar lesion subjects (cbls groups 1 and 2) showed significantly greater orienting deficits than normal control groups during the spatial detection task. (B) Autism and cerebellar lesion groups displayed significantly greater orienting deficits than normal controls during the spatial discrimination task. (C) Correlation of orienting with vermal lobules VI–VII in 22 normal, 10 autism, and 7 cerebellar lesion subjects. Vermal lobule VI–VII area measures in each subject were divided by that subject's intracranial brain volume (ICV) to control for overall brain size [reproduced with permission from Townsend, J., Courchesne, E., Singer-Harris, N., Covington, J., Westerfield, M., Lyden, P., Lowry, T. P., Press, G. A. (1999) Spatial attention deficits in patients with acquired or developmental cerebellar abnormality. *J. Neuroscience* **19,** 5632–5642.]

now to be shifted to another source of information (e.g., such as when a mother is speaking to her child and then tells the child to look at something). In natural situations, such shifting of attention back and forth between different sources of information occurs often and rapidly, and fMRI studies show that cerebellar and parietal cortex are actively involved in normal people. It is unclear whether operations underlying such moment-to-moment and rapid shifts of attention overlap with the single instance of anticipatory orienting, but this seems likely based on fMRI experiments in normal people. Patients with cerebellar lesions from stroke or tumor and autistic patients are slow and inaccurate in shifting of attention rapidly and frequently back and forth between very disparate (e.g., sights and sounds) sources of information. Therefore, when cerebellar or autistic patients have little time to implement a shift, they often miss detecting important information appearing at the new source of attention and their neurophysiological responses confirm that they failed to mentally register that important information. Stimuli that explicitly signal the need to immediately shift attention elicit an "Sd" brain response, but this is abnormally small or absent in cerebellar or autistic patients.

Although slow and inaccurate in anticipatory orienting and shifting of attention, cerebellar and autistic patients eventually do adjust the direction of their attention if given sufficient time, according to both behavioral and ERP imaging evidence. Thus, it has been a misconception among some researchers that autistic patients are incapable of orienting or shifting attention. Tests requiring conscious and effortful judgments about when to change or shift mental set, such as the Wisconsin Card Sorting Task, are not tests of the cerebellar or parietal involvement in the capacity to make immediate (in the hundreds of milliseconds range) and accurate adjustments in the direction of attention signaled by unexpected stimuli occurring within an attended stream of information. Rather, the successful ability to make such extremely quick and accurate anticipatory attentional adjustments must depend to a large extent on prior learning of associations between the signal to shift and the appearance of subsequent important information. Such automatic and swift neural actions are likely the product of well-known cerebellar association learning mechanisms. Such mechanisms may first learn predictive associations between successive events and then use that information by sending, on a moment-to-moment basis, preparatory signals to appropriate brain systems.

F. Associated Features

1. Mental Retardation

A robust finding is that approximately 75% of individuals with autism also meet *DSM-IV* criteria for mental retardation. For the minority with "normal" intellectual functioning, subtle deficits in social and language behavior clearly set them apart from their typically functioning peers.

There is no consensus regarding the role of mental retardation in autism and scientists have approached the topic in several ways. One thought has been that since mental retardation is not specific to autism (e.g., children with Down's syndrome and cerebral palsy also have mental retardation), attempts to understand its role in the disorder will not elucidate etiological or pathobiological mechanisms. Thus, the role of mental retardation is largely ignored. Another approach has been to investigate only a "pure" form of the disorder and "subtract out" effects due to mental retardation by studying only nonretarded individuals with autism. A further idea has been to control for the presence of mental retardation by comparing autistic individuals with mental age-matched individuals, either normal subjects (usually substantially younger) or subjects from a special population (usually Down's syndrome or general mental retardation). Although each approach has its own unique set of problems (e.g., lack of generalizability of finding when studying only nonretarded individuals), taken together they raise important issues regarding mental retardation and autism. First, autism and mental retardation are inextricably linked and attempts to dissociate the two may not only be impossible but also misleading. Furthermore, it is conceivable that subjects at either end of the mental retardation spectrum—those with profound mental retardation and those without apparent mental retardation—will be important to study because they could potentially represent examples of the extreme forms of biological mechanisms supporting the pathogenesis of the majority of autistic cases. For example, a hypothetical finding of increased brain volume in a severely retarded autistic child could represent the overall phenomenon of excessive brain overgrowth at early ages in autism. It is apparent in this individual because he or she represents the most severe biological version of the disorder. Thus, if such cases of retardation are excluded, important findings may be missed. On the other hand, given the extreme heterogeneity seen in autism, combining subjects with potentially disparate etiologies and courses may also dilute results. Currently, no consistent scientific solutions to the issue of mental retardation in autism have been embraced, but researchers are beginning to analyze data in multiple ways, such as with both group and single subject analyses.

2. Savantism

As is often popularized by the media, a small percentage of people with autism (less than 1%) display a special savant ability. The phenomenon of the savant offers a unique opportunity to study the presence of an extraordinary skill in the background of general cognitive impairment. Savant skills in autism have been reported in artistic talent, musical ability, numeric calculation skills, and "calendar calculation" skills, which involve the generation of the weekday of a given date within seconds. For example, a musical savant might have the extraordinary ability to play any song on the piano after hearing it only once on the radio. Savantism in autism is intriguing because many (although not all) such individuals may not even possess functional communication. Many theories, such as exceptional memories, have been espoused as to why savantism occurs in autism, but none have been definitively supported.

3. Seizures

Approximately one-third of autistic individuals will experience one or more epileptic seizures by adolescence. All types of seizures have been reported in autism, and complex partial seizures with associated centrotemporal spikes are most common, although other spike patterns have been noted. Severity of autism and other phenotypic markers do not seem to be associated with an elevated risk for seizures. Furthermore, the relationship of epileptic activity, if any, in the pathogenesis of autism is unknown.

4. Hyperserotenemia

Approximately one-third of individuals have elevated levels of blood platelet serotonin 5-hydroxy tryptamine (5-HT). This interesting finding has been replicated several times, and scientists have begun to use neuroimaging methods to make inferences regarding serotonergic brain function in autism. One method that has been tried has been the use of a serotonin tracer (α-[^{11}C]methyl-L-tryptophan) in combination with PET. Although a relatively new line of research, reports have indicated asymmetries in serotonin synthesis in autism. For example, decreased serotonin synthesis has been found in frontal cortex, whereas increased synthesis has been noted in the dentate nucleus. As with several other biological systems in autism, inconsistency seems to be a consistent theme. The potential role of the serotonergic system in the etiology of autism is a vigorous line of investigation and points to the serotonin transporter gene as a logical gene candidate. Furthermore, this work has

prompted the use of several pharmacological interventions that specifically target 5-HT systems.

IV. ANATOMICAL BRAIN DEFECTS

A. Multiple Anatomical Abnormalities

Among all types of biological abnormalities in autism, evidence for neuroanatomical abnormality is the strongest. Studies show that in autism, most major brain structures are affected (Fig. 6); these include the cerebellum, cerebrum, limbic system, corpus callosum, basal ganglia, and brain stem. Recent evidence shows that within the cerebellum and cerebrum, there is abnormality in white and gray matter. Such widespread anatomic abnormality explains why autism involves pervasive and persistent neurological and behavioral dysfunction.

Among all anatomic findings, the most consistent is abnormality in the cerebellum. According to MRI data on large numbers of autistic children and adults, there is a reduction in the size of the neocerebellar vermis and the volume of cerebellar cortex. Also in autistic children, there is an inverse relationship between the size of the cerebellar vermis lobules VI and VII and the size of frontal lobes, such that the smaller the vermis, the larger the frontal lobes. In brain autopsy studies, 90% of autistic cases have reduced numbers of Purkinje neurons in the cerebellar vermis and cerebellar hemispheres. The amount of loss typically ranges from about 20 to 60% across different autistic individuals, but in one autistic case Purkinje neuron loss was nearly total throughout the cerebellar hemispheres. Also, Purkinje neuron loss is patchy, and the distribution across the cerebellar hemispheres and vermis differs between individual autistic cases. Autopsy studies have also reported other types of cerebellar anatomic abnormality in other autistic cases. Therefore, cerebellar anatomic abnormality is present in 95% of all autism autopsy cases, making this the most common biological abnormality known for this disorder (Fig. 7).

The limbic system is another common site of anatomic abnormality. In living autistic patients, the amygdala has a reduced volume and the dentate gyrus has a reduced cross-sectional area. In autopsy studies, anatomic abnormality in limbic structures is present in most, but not all, autism cases. When present, limbic system abnormality involves increased density of neurons and reduction in neuron sizes.

Another important site is the cerebrum. A recent MRI study found abnormally increased volume of the

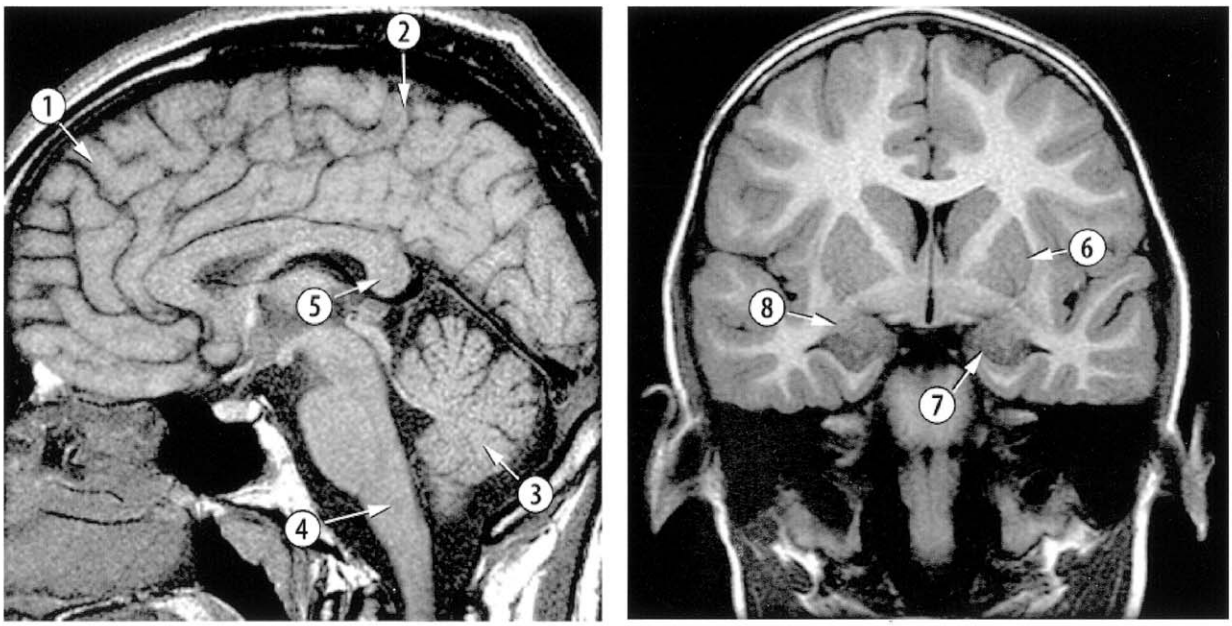


Figure 6 Depiction of cortical and subcortical structures reported as abnormal in autism via autopsy or MRI data from laboratories throughout the United States and Europe. 1, frontal lobes; 2, parietal lobes; 3, cerebellum; 4, brain stem; 5, corpus callosum; 6, basal ganglia; 7, hippocampus; 8, amygdala. (See color insert).

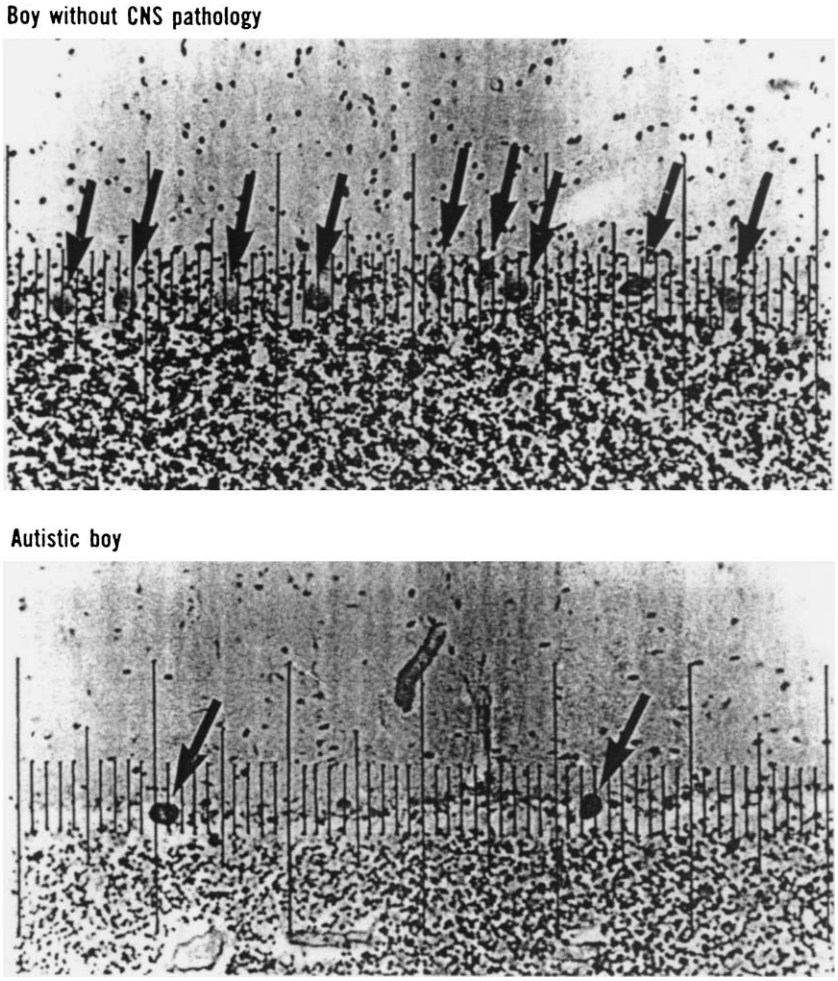


Figure 7 In autism, reduction in cerebellar Purkinje neuron numbers is a common finding across postmortem cases examined to date. (Top) Control case; (Bottom) autism case. Figure shows a comparable span of cerebellar cortex in a control case and in a case with autism. Arrows point to Purkinje cell bodies in these Nissel-stained sections. There also appear to be fewer granule neurons in the case with autism [adapted with permission from Ritvo, E. R., Freeman, B. J., and Scheibel, A. B. (1986). Lower Purkinje cell counts in the cerebella of four autistic subjects: Initial findings of the UCLA–NSAC autopsy research report. *Am. J. Psychiatry* **143,** 862–866.]

cerebrum, but only in 2- to 4-year-old autistic children. A follow-up study found a gradient of abnormality in these very young autistic children, such that abnormal increases in volume were most marked in frontal brain regions and least in posterior regions. In autopsy studies, some autistic patients have abnormally increased thickness of frontal cortical regions; however, other autistic patients have irregular formation of the layers of cortical regions. The corpus callosum is a massive tract of fibers enabling communication between the right and left hemispheres of the cerebrum, and in the posterior portion of it there is a reduction in size.

Anatomic abnormalities in some sites have only been reported from single MRI studies (e.g., parietal lobe and basal ganglia) or in single individuals at postmortem (e.g., superior olive agenesis and facial motor nucleus dysgenesis); how common these sites of abnormality are remains unknown.

In conclusion, anatomical abnormalities in the autistic brain prove that this is a biological disorder, not a psychogenic one. Different brain structures have different types of abnormality (e.g., cerebellum and cerebrum), and even within a particular brain structure (e.g., frontal cortex) more than one type of abnormality can be seen in different autistic patients. Also, among patients with the same type of anatomic defect, there are differences in the amount and exact location of abnormality. Such individual-specific differences in anatomic abnormality in the cerebellum, limbic sys-

tem, cerebrum, and so forth may be the reason for individual differences in initial symptoms and behavioral outcome.

B. Brain Development

Information about what brain abnormalities are present early in development and how they change with age cannot be gained by observations of the brains of older children or adults alone because the older brain is the outcome or end product of developmental transformations. Animal model studies demonstrate that early abnormal perturbations can create complex cascades of structural alterations. Abundant evidence from developmental neurobiology shows that when the developing brain sustains significant neural insult precipitated by environmental or genetic factors, the typical balance of competition that mediates anatomical and functional organization is altered and can result in atypical functional activity and structural growth (reduced or enlarged dendritic arbors, neurons, representational maps, or even whole cytoarchitectonic regions). This neuroplasticity may produce arrangements that successfully compensate for the structural defect but may also produce arrangements that do not (perhaps this partially explains the different degrees of impairment among autistic individuals). If the developing autistic nervous system abides by well-known principles of competition and neuroplasticity in the face of neural defect, then anatomical and functional abnormality at a given site or age will affect growth at other sites or later ages. In other words, under both normal and abnormal conditions, brain development can be thought of as a cascade of events that trigger multiple series of typically nonlinear and interdependent structural and functional transformations from prenatal life through maturity (and even then change continues). Thus, the end product of developmental neuroplasticity will be multiple sites and types of abnormality. To determine the developmental anatomical phenotype in autism, age-related changes in anatomy must be studied directly from the earliest ages possible.

1. MRI Evidence

With *in vivo* MRI, it is possible to study brain development at the time of clinical onset during the first years of life through maturity. Unfortunately, as with postmortem evidence, anatomic evidence from *in vivo* MRI also comes primarily from adults, adolescents, and older children with the disorder. Currently, there is little knowledge regarding what neuroanatomic abnormalities are present at the earliest developmental stages of this disorder (2 or 3 years of age) and what pathological growth trajectories characterize subsequent years of brain development. This is due to the scarcity of studies, small samples reported, and difficulty in both identifying autistic patients at the youngest ages and recruiting normal youngsters. Nonetheless, several important observations about the developmental anatomic phenotype of autism have been made, and, surprisingly, one has to do with neonatal head circumference.

At birth, head circumference in autism is typically normal, which is indicative of normal overall brain volume at that age. By ages 2–4 years, however, a majority of autistic children have a brain volume larger than normal (Fig. 8); in fact, more than one-third of young autistic children have a brain size falling into the definitional zone of developmental macrencephaly. One autistic 3-year-old had a brain volume whose weight we calculated to be 2000 g, which exceeds weights reported on more than 8000 normative child and adult cases in the literature (Fig. 9). One reason for this excessive and premature brain size among autistic 2- to 4-year-olds is premature overgrowth of white matter in both the cerebrum and the cerebellum; the magnitude of this abnormality is the largest for any anatomic defect reported in the autism MRI literature (Fig. 10). Another reason is premature overgrowth of cerebral gray matter. Strangely, after this young age, further growth is retarded in autism, whereas the

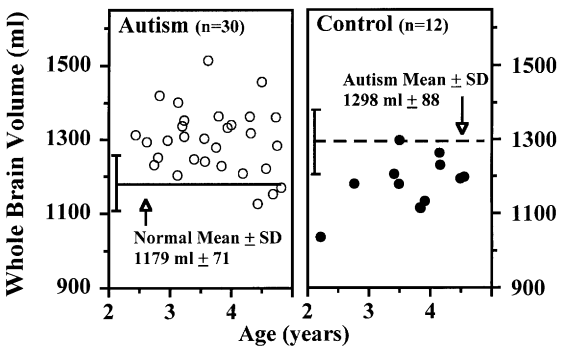


Figure 8 Two- to 4-year-old autistic and normal males (circles) are plotted showing overall whole brain enlargement of the youngest autistic children. As shown, 31 of 36 (86%) of the autistic boys and girls had whole brain volumes larger than the normal mean. In contrast, only 1 of 24 (4%) normal boys and girls in this age range exceeded the autism mean (reproduced with permission from Courchesne *et al.*, 2001).

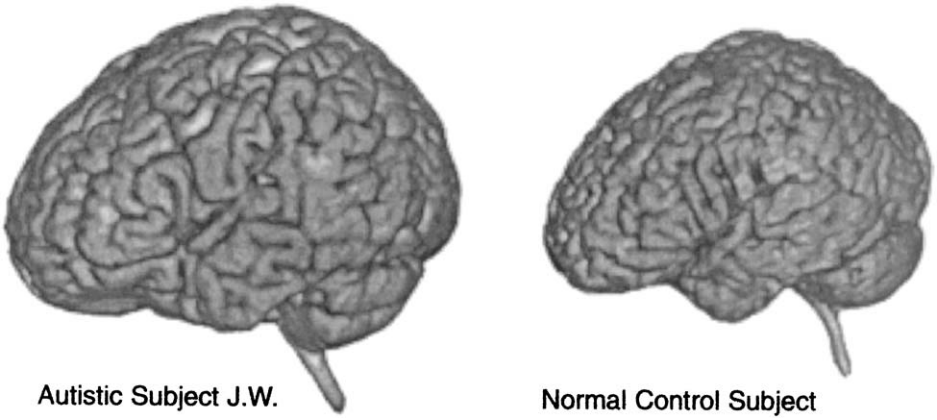


Figure 9 A case of extreme macrencephaly in autism. Three-dimensional image of 3.4-year-old autistic subject, JW (left), compared to that of a normal male child whose brain volume was scaled to equal normal average size (right). The autistic child's brain volume (1816 ml) was more than 6 SDs above the normal average (1162 ml) for his age (from Courchesne *et al.*, unpublished data).

normal child's brain continues to grow at a strong pace. By older childhood and adolescence, the normal child's cerebrum and cerebellum have reached or surpassed sizes achieved at an earlier age by the very young autistic child. This explains why postmortem observations of the weight of the autistic brain show it to be normal in the majority (86%) of older children and adult cases, although a few rare cases of extreme weight at that age do exist.

Apparently, after birth but prior to about 2 or 3 years of age in autism, several cerebral and cerebellar anatomical abnormalities rapidly occur, which means that postnatal, biologically guided intervention prior to the full expression of these abnormalities may be a possibility in the future. These observations suggest autism involves an unusual, perhaps unique, developmental neuroanatomic phenotype.

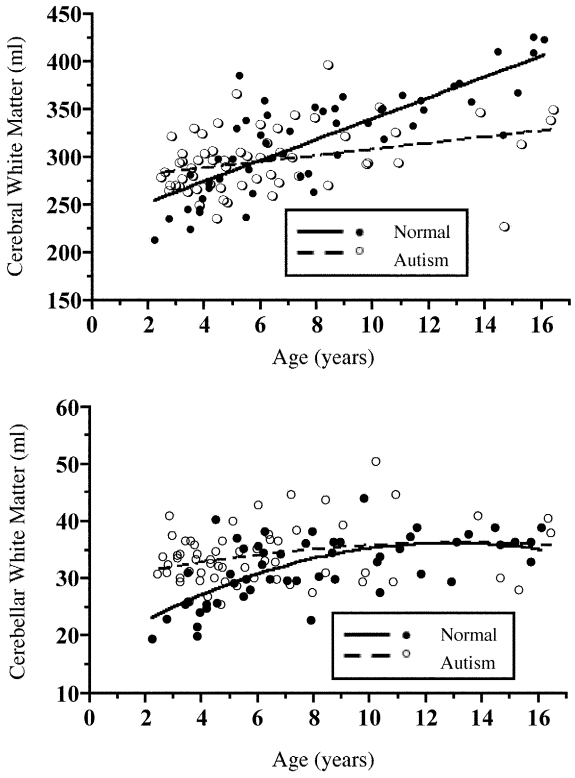


Figure 10 Volumes of cerebral white matter (top) and cerebellar white matter (bottom) are plotted with their best-fit curves for normal and autistic 2- to 16-year-old males. Despite precocious growth by early childhood, in autism, both cerebral and cerebellar white matter growth were smaller than normal by adolescence. (reproduced with permission of Courchesne *et al.*, 2001).

2. Postmortem Evidence

The major strength of postmortem research is that it allows detailed histoanatomical examination unattainable by MRI, and many important observations of the brain of the older child or adult with autism come from such studies. However, currently, there are too few postmortem autism cases in the literature to provide definitive evidence of either age at pathological onset or developmental trajectories for any brain structure. Only six postmortem cases of autism less than 19 years of age have been reported, and only one of these is younger than 9 years of age (i.e., one 4-year-old case). Additionally, postmortem studies of purported "autistic" infants and children under the age of 5 years suffer from one obvious drawback: At that young age, the diagnosis of autism remains uncertain, and obviously no longitudinal observations may be per-

formed to confirm or alter an initial diagnosis of autism suspected at that age.

Nonetheless, some neuropathological features are indicative of early developmental onset. Such features include cells arrested in migration in the inferior cerebellar peduncles, inferior olive malformation, irregular lamina disarray in small focal cerebral regions, reduced numbers of Purkinje neurons in the apparent absence of glial scarring, dysgenesis of facial motor nuclei, and agenesis of the superior olivary nuclei. However, several of these pathologies have only been noted in one or a few cases and not in the majority of postmortem cases, and therefore they do not provide definitive evidence about age of onset of pathology in the majority of autism cases. Interestingly, the different pathological features point to several different prenatal periods of onset.

Other reported pathological features—small cell size, increased cell packing density, and thickened cortices—are not indicative of any particular age of onset, but it is parsimonious to hypothesize onset during early development.

Since cerebellar pathology is ubiquitous among autism cases, knowledge of age of onset would be valuable in seeking causes and effective interventions. Decreased Purkinje neuron number is not accompanied by empty basket cells, which is indicative of early rather than later developmental age at the time of loss. Since the decrease in Purkinje number is evident regardless of history of seizures or seizure medication, the age of onset and cause of cell loss are not explained by seizure onset or seizure treatment. Also, cerebellar folia do not show signs of atrophy (mutant mice with developmental loss of Purkinje cells show cerebellar hypoplasia, not atrophy). In several autism cases, the inferior olives (which are developmentally, structurally, and functionally intimately involved with cerebellar Purkinje neurons) are maldeveloped and neuronal migration errors appear in the inferior cerebellar peduncles. In one postmortem autism case, Purkinje neurons were irregularly aligned, which could not caused by some later postnatal event. The dysplastic olives and migration errors are clear signs of first trimester events but have only been seen in three autism postmortem cases. The developmental ramifications of dysplasia and migration defects are unknown. Despite the reduced number of Purkinje neurons, cerebellar cortex in autism has clearly defined cortical lamina (granule, Purkinje, and molecular layers), even in cases with large distances between single surviving Purkinje neurons. Although Purkinje neuron loss or dysgenesis could occur prenatally along with olivary dysgenesis and the migration errors in the inferior cerebellar peduncle, observations of intact cerebellar lamina seem to argue against a massive dysgenesis or loss of Purkinje neurons before the final phases of granule cell proliferation and migration. Therefore, the possibility cannot be reasonably ruled out that Purkinje neuron loss could occur postnatally, perhaps in the first 1 to 3 years of life.

V. POTENTIAL ETIOLOGIES

Autism is currently thought to be etiologically heterogeneous. Many ideas about the etiology of autism have been presented in the literature; for example, mitochondrial DNA mutation and toxin exposure have been cited as possible causes of autism. Although intriguing, these do not yet account for subsets of persons with autism.

A. Genetic Findings

Autism is among the most heritable of neuropsychiatric disorders. Twin studies report pairwise concordance rates of up to 91% for autism among monozygotic twins and 24% among dizygotic twins, suggesting the disorder is strongly genetic. The increased risk of autism among siblings of autistic probands is 2%, which is about 45 times the rate expected in the general population. For the majority of the autistic population, multiple interacting genes are likely to be etiologically involved. About 5% of autistic patients have abnormalities in one or another of nearly all possible chromosomes; therefore, across autistic individuals with a chromosomal abnormality, the site of abnormality is variable.

Nonetheless, according to many reports, the most common site of chromosomal abnormality is 15q11-13. This abnormality occurs in about 2–4% of all autism cases, and it has been suggested that the autism phenotype occurs with maternal transmission of this particular chromosome defect. Because of the relative frequency of chromosome 15q11-13 abnormalities in autism, several candidate approach studies specifically tested and found significant linkage within the 15q11-13 region; two studies obtained statistically suggestive linkage. One specific location within this 15q11-13 region, the GABRB3 155-CA2 location, was highly significant in linkage with autism in a recent study that used data from all available genetic studies combined.

The first few genomewide searches for potential autism-related effects have only recently been completed. These searches studied families that have multiple affected individuals (which represent 3% of autism cases), and generalizability of their positive as well as negative findings to the larger autism population is uncertain. To date, results have been highly variable and generally not replicable. Some genomewide studies report only weak to modest linkage scores for several different sites; other studies find no significant linkage to any site or weak linkage with a large number of sites from nearly a dozen chromosomes. No reported site has been definitively statistically replicated by other genomewide studies. Genetic studies have been consistent in finding no association between carefully diagnosed autism and the fragile X gene (although autistic features are seen in many fragile X patients).

Because of the link between autism and hyperserotonemia in about one-third of patients studied, candidate gene studies have investigated known serotonin genes. Two such studies found significant linkage with the promotor of the serotonin transporter gene but to opposite alleles of the promotor, whereas another two studies found no linkage.

Despite the fact that autism is likely a complex genetic disorder and modern genetic techniques have only recently been applied to autism, several important observations have been achieved and more success can be expected.

B. Teratogens, Toxins, and Viruses

Specific teratogens, toxins, or viruses have not been shown to account for any substantial subset of the autism population, although clearly in each category there are known members capable of producing developmental brain damage, including damage to regions commonly affected in autism. For example, in humans, prenatal and neonatal exposure to alcohol causes Purkinje cell loss in the cerebellum. Prenatal exposure to the anticonvulsant medication valproic acid has been linked to cases of autism in one family and causes significant reduction of cerebellar Purkinje neuron numbers in rats. Prenatal exposure to sedatives such as thalidomide has been linked to autism in some individuals, and an animal model using rats to study treatment for opiate and cocaine addiction using the indole alkaloids ibogaine and harmaline found degeneration of a subset of Purkinje cells in the vermis.

Cells of the early developing brain are in a state of rapid proliferation and differentiation in the fetus and neonate—conditions ideal for viruses that replicate best in dividing cells. This in turn can lead to severe or lethal birth defects. Lymphocytic choriomeningitis virus infection of 1- to 7-day-old rats interferes with cerebellar development, resulting in cerebellar hypoplasia and permanent ataxia. After developing rats are exposed to Borna virus, affected animals show similar cerebellar pathology as well as abnormal play and other social behavior.

Recent speculation that the measles, mumps, and rubella vaccine may be an etiological factor in autism led to a thorough examination of the evidence by many medical researchers. No scientific support for the speculation was found. Nonetheless, scientists remain open to the possibility that some teratogenic, toxic, or viral exposures might serve as risk factors that add to or compound the etiological effect of genetic factors.

VI. TREATMENT

When the disorder was first described almost 60 years ago, few if any treatment options were available, and parents were left with little hope of cognitive or behavioral advancement for their child. It was common for autistic children of this era to be placed in institutions, largely ignored, grouped together with children with other special needs (e.g., general mental retardation and blindness). Although there is still no cure for autism, modern-day treatment of this disorder affords several possibilities and research has shown that significant advances in cognitive and behavioral domains can be made. In fact, today, most children receive intervention on a daily basis, some as early as age 2. Furthermore, many school-aged children with autism are ''mainstreamed'' or are placed in classrooms with typically developing children for some or all of their school day. The reasons for this dramatic metamorphosis in thinking about the treatment of autism come from several sources. First, the 1940s through the 1970s was a time when autism was thought of as ''psychogenic'' in origin—a disorder resulting primarily from poor parenting, particularly on the part of the mother. The late 1960s and early 1970s, however, brought new scientific information, such as elevated serotonin levels, the presence of seizure activity, and increased rates in monozygotic twins compared to dizygotic twins in autism. Such information presented clear evidence that this disorder was of biological, likely genetic, origin. This belief is still true

today. Increased sophistication of diagnostic tools, heightened media attention of the disorder, federal laws mandating specialized education for children with disorders, and a growing knowledge of biological principles have all contributed to a new understanding of the treatment of autism.

Although intervention in autism is multimodal and typically includes several treatment types (e.g., language therapy and occupational therapy), takes place in several locations (e.g., the home and school), and is implemented by several treatment providers, (e.g., parents, teachers, and speech therapists), by far the most common treatment approach is one that relies on a behavioral model. The behavior analytic view is that autism is a syndrome of behavioral deficits that are amenable to change in response to specific, carefully programmed, constructive interactions with the environment. Within this framework, both global approaches, such as discrete trial training, or specific approaches, such as video modeling, have been used. Behavior analytic treatment focuses on teaching small, measurable units of behavior systematically. Every skill absent in the child's repertoire—beginning from simple responses such as looking at an object to complex ones such as verbally describing an object—is broken down into small steps. Each step is taught to the child by providing consistent prompts or instructions followed by consistent reward for accurate and/or other appropriate behavior. Inappropriate behavior is often ignored or redirected. Teaching trials are repeated many times, often in rapid succession, until the child performs the target behavior in the absence of adult prompting. For example, a child may practice the association between the word "car" and a picture of a car dozens, perhaps even hundreds of times per day. It is this repetition along with gradual increases in demands in skill complexity that characterize behavioral treatment for children with autism.

Based on review of the treatment literature in autism, three general themes emerge: (i) Interventions based on a behavioral model are successful in increasing some functional behaviors (e.g., eye contact or language) as well as decreasing some nonfunctional behaviors (e.g., repetitive motor movements); (ii) the same intervention does not affect all children in the same way, even when factors such as IQ are controlled for; and (iii) the earlier the intervention is provided, the more efficacious for the child. These points raise larger questions regarding the biological mechanisms that may support their conclusions.

First, what are the mechanisms that might support positive behavioral change in autism after treatment? As described previously, in many behavioral approaches the use of repeated contingencies or associations forms the foundation of most teaching interactions. Similarly, repeated environmental stimulation is a common method for initiating plasticity in animals. For example, it is well-known that mice repeatedly exposed to an exercise gym have more motor neurons and increased numbers of synapses than mice without such repeated exposure. The approach of repeatedly pairing a stimulus and response may be particularly effective in autism because it targets associative learning, a function mediated by the cerebellum, a structure consistently cited as abnormal in autism. The importance of the cerebellum in associative learning has been well documented. For example, a traditional cerebellar/association learning experiment might compare cerebellar activity during the presentation of random unpaired tones and air puffs to that during paired conditioning trials. Relative to the unpaired condition, increased glucose metabolism is often found during the paired associative conditions in various cerebellar regions. In autism, perhaps substantial changes in behavioral phenotypes can be more readily achieved when using a method that attempts to remap this potentially deficient function. Another possibility is that during the initial stages of treatment in autism, the environment is typically cleared of all distractions (e.g., the child may receive treatment in a quiet room without the sound of other children in nearby classrooms or potential unexpected sounds, such as the telephone ringing). If it is indeed true that autism involves difficulties in sensory integration and attention, a quiet and distraction-free environment would likely be less stressful for a child than one with ambient noise, etc. Animal studies have in fact shown that neurogenesis in structures implicated in autism, such as the dentate gyrus, is facilitated by an enriched environment and stymied by a stressful one. Therefore, treatments aimed at reducing sensory stress while presenting clear and repetitive learning trials might be the most palatable for the nervous systems of children with autism. It is important to note, however, that as the child progresses in ability and response complexity, the need for long periods of repetition and a distraction-free environment decreases. That is, once the nervous system of the autistic child is able to assimilate basic information, he or she may be more readily able to learn from his or her environment in a way more similar to that of typically developing children.

Second, why does the same treatment fail to initiate the same level of efficacy for each child with autism? In

order to illustrate this point, consider the landmark paper on treatment in autism in which Lovaas reported that 47% of children with autism receiving high-intensity behavioral treatment (i.e., 40 hr per week) were placed in mainstreamed academic environments after treatment, whereas the remaining 53% were not. These differential rates of treatment success cannot be explained by single behavioral phenotypic characteristics (e.g., severity of autistic symptoms, IQ, and language ability) because no significant differences were found between treatment "high responders" and treatment "low responders." The explanation might be found when one takes into account biological substrates that distinguish these two outcome groups. In order to push the child with autism down the most beneficial pathway, treatment efforts must be based on brain–behavior research that provides information regarding the most fruitful starting point from which to intercede *for each particular child*. Although brain-based treatments of this type have yet to be implemented, current evidence supports the conclusion that available experimental and analytic methods can be used to sort patients into biologically distinct groupings on which to test whether different treatment approaches have reliably different efficacies. However, to succeed, it is necessary to first establish direct experimental correlations between specific behavioral and psychological dysfunctions, on the one hand, and specific underlying physiological, anatomical and other biological abnormalities on the other hand. Recent studies have in fact succeeded in doing so. These studies suggest, and new studies may further specify, the distinctly different compensatory mechanisms used by different autistic individuals. Furthermore, such brain–behavior findings will identify specific neurobehavioral operations (e.g., deployment of visual spatial attention, shifting attention, and others) that may be pivotal to the normal development of still more complex, higher cognitive, linguistic, and social functions. The validity of treatment efforts based on epigenetic theories can thus be easily tested: The strength of brain behavior correlations (e.g., overly focused visuospatial attention correlated with parietal abnormalities) should decrease over the course of treatment as compensatory mechanisms are enhanced in the autistic brain.

Third, why does treatment initiated at age 2 seem to have more beneficial effects than treatment initiated at age 6? Again, we turn to well-known principles of neural development provided by animal research. Specifically, animal models of developmental abnormality show the importance of critical periods or "windows of opportunity" for effective treatment. For example, for the kitten, the longer the period of abnormal neural activity caused by monocular deprivation from birth, the more abnormal the neural structural and functional outcome. Abnormality compounds gradually; correspondingly, the possibility for recovery declines gradually. Therefore, the longer one waits before starting treatment, the poorer the recovery, and eventually an age is reached when treatments cease to be effective. Fortunately, the human cerebral cortex has an exceptionally long period of continued growth, providing both the opportunity and vulnerability to change. Therefore, assuming the biological etiologies causing autism begin prenatally or early in infancy, the lesson from developmental biology is to take advantage of these windows of epigenetic opportunity by intervening as early as possible. Early and biologically informed treatments are especially critical when rare human neurodevelopmental disorders are due to mutations of genes with pleiotropic effects. Theoretically, if initiated early enough, a single, biologically precise treatment may prevent the cascade of multiple paths of maldevelopment; if not, multiple types of treatment may be needed to address each of the distinct and unrelated biological and behavioral outcomes that characterize such gene mutations. Not surprisingly, a major goal of autism research is to devise specific biological or behavioral countereffects that avert or significantly mitigate the disorder before they are fully manifested and require multiple, marginally effective treatments. Early intervention research efforts have embraced this notion, with treatment often beginning as early as 14 months of age or as soon as a provisional diagnosis is made. Combined with neuroimaging technique, scientists have for the first time the ability to track potential changes in neurobiology as a result of intensive early intervention in autism.

A. Pharmacological

A wide range of pharmacological agents have been used as viable treatment options for autism, ranging from serotonin uptake inhibitors, such as clomipramine and even the common antidepressant Prozac, to dopamine antagonists such as haloperidol. Overall, drugs have not proven successful at ameliorating the key features of autism but, rather, may have a modest effect on specific features of the disorder for some children. For example, serotonin uptake inhibitors

(known to be effective in the treatment of obsessive–compulsive disorder) have been shown to be effective in reducing repetitive behaviors seen in autism. Side effects common with psychotropic medications (e.g., tardive dyskenisia), however, typically limit the use of such interventions. Most children with autism participate in pharmacological intervention at some point during their treatment course, but it is the minority that maintains drug use across their life span.

B. Other

Behavioral treatment is not the only option for children with autism and it is typically augmented with other approaches, such as occupational therapy and/or sensory integration training, language therapy, play therapy, or pharmacological intervention. Each of these approaches is grounded in a particular theory of autism. For example, sensory integration theory posits that autism is a disorder with impairments in processing sensory information and methods common to this treatment type are aimed at making the absorption and integration of sensory information more palatable for the autistic child. Methods common to this intervention type include "brushing" the child to enhance vestibular awareness and providing controlled tactile stimulation, such as exposure to objects that are extremely course to extremely rough, in an effort to reduce tactile defensiveness. Interestingly, although there is little or no scientific evidence that treatment of this type is effective in autism, it is a very common component of the education of most autistic children, and parents often express that it is both an important and an efficacious aspect of their child's treatment. Finally, it is important to note that not all treatments for children with autism are based on a behavioral model and may be combined with or replaced by treatment approaches based on other models, such as a developmental one. Developmental models have been particularly useful in the treatment of play, language, and social skills in children with autism, although less documentation is available for this approach.

A unique aspect of the treatment of autism throughout the decades has been the presence of "alternative" approaches, such as swimming with dolphins or other animal therapy, holding therapy wherein the child is held tightly for several hours, art therapy, music therapy, facilitated communication, and megavitamin and hormone therapy. Such intervention approaches come and go, and they have yet to make a solid contribution to our understanding of or the treatment of this complex disorder.

VII. CONCLUSIONS

Neuroanatomical, neurofunctional, and neurobehavioral findings in autism all suggest that the disorder cannot be reduced to one primary deficit that underlies the disorder. Thus, there is not just one important neural defect or one important neurobehavioral deficit; there are many. Such neurobehavioral domains include, but are not limited to, deficits in reciprocal social interactions, affective responses, verbal and nonverbal communication, face perception, problem solving, working memory, orienting and shifting attention, spatial attention, sensorimotor learning, auditory processing, and motor coordination. In order to fully appreciate why the autistic child has pervasive and devastating disorder in the development of numerous abilities from social to motor, it is necessary to understand the entire pattern of anatomical and neurobehavioral impairment he or she faces throughout life. This picture of multiple types, times, and sites of neural defect is analogous to that seen in complex genetic neurodevelopmental disorders (e.g., fragile X, Williams' syndrome, CRASH, adenylosuccinate lyase deficiency, and PKU). In such complex genetic disorders, the genetic defects have adverse pleiotropic effects on brain development, such that multiple neural sites and often extraneural systems develop abnormally. Autism is almost certainly a complex genetic disorder, with involvement of one or more genes, and gene defects will likely have pleiotropic effects producing maldevelopment in multiple brain structures and possibly extraneural systems.

The importance of understanding the neural basis of autism cannot be underestimated. It will facilitate, for example, the development of treatment approaches that address individual differences in neural organization, the development of biologically guided measures of the efficacy of treatment approaches that can augment current behavioral measures, and the development of animal models of autism that will be crucial for validating genetic or nongenetic explanations or biological or behavioral treatments. (Note that animal models rely on the availability of brain–behavior linkage data to be modeled.) In fact, the proposed combined behavioral, neurofunctional, and neuroanatomical approach is consistent with the types of approaches currently taken in the study of animal

mutants and knockouts, potentially facilitating the translation of information from these studies to future autism animal model studies.

Moreover, the combination of neuroimaging techniques, such as fMRI, with cognitive and behavioral techniques will help to sharpen diagnostic distinctions not only between autistic and normal individuals but also between biologically different subgroups within the disorder of autism. Finally, as suggested by our preliminary fMRI data, brain–behavior findings on autism may provide insights into specific types of deviant functional organization as well as possible forms of compensatory organization invoked during development in autism.

See Also the Following Articles

ATTENTION • BRAIN DEVELOPMENT • CEREBELLUM • LANGUAGE ACQUISITION • LANGUAGE DISORDERS • LIMBIC SYSTEM • MANIC–DEPRESSIVE ILLNESS • MENTAL RETARDATION • MOTOR CONTROL • SPEECH

Suggested Reading

Allen, G., and Courchesne, E. (2001). Attention function and dysfunction in autism. *Frontiers in Bioscience* **6,** 105–119.

Bailey, A., Luthert, P., Dean, A., Harding, B., Janota, I., Montgomery, M., Rutter, M., and Lantos, P. (1998). A clinicopathological study of autism. *Brain* **121,** 889–905.

Courchesne, E., *et al*. (2001). Unusual brain growth patterns in early life in patients with autistic disorder: An MRI study. *Neurology* **57,** 245–254.

Courchesne, E., Yeung-Courchesne, R., Press, G. A., Hesselink, J. R., and Jernigan, T. L. (1988). Hypoplasia of cerebellar vermal lobules VI and VII in autism. *N. Eng. J. Med.* **318,** 1349–1354.

Kanner, L. (1943). Autistic disturbances of affective contact. *Nervous Child* **2,** 217–250.

Lamb, J. A., Moore, J., Bailey, A., and Monaco, A. P. (2000). Autism: Recent molecular genetic advances. *Hum. Mol. Genet.* **9,** 861–868.

Lewy, A. L., and Dawson, G. (1992). Social stimulation and joint attention in young autistic children. *J. Abnormal Child Psychol.* **20,** 555–566.

Lovaas, I. (1987). Behavioral treatment and normal educational and intellectual functioning in young autistic children. *J. Consulting Clin. Psychol.* **55,** 3–9.

Mundy, P., and Sigman, M. (1989). Theoretical implications of joint-attention deficits in autism. *Dev. Psychopathol.* **1,** 173–183.

Pierce, K., Müller, R.-A., Allen, G., and Courchesne, E. (2001). Face processing occurs outside the fusiform "face area" in autism: Evidence from functional MRI. *Brain* **124,** 2059–2073.

Stone, W. L., Lee, E. B., Ashford, L., *et al*. (1999). Can autism be diagnosed accurately in children under 3 years? *J. Child Psychol. Psychiatr. Allied Disciplines* **40,** 219–226.

Autoimmune Diseases

FELIX MOR and IRUN R. COHEN
Weizmann Institute of Science, Rehovot, Israel

GLOSSARY

autoimmune disease A disease process mediated by self-reacting antibodies or T cells.

Behçet's syndrome A multisystem, chronic recurrent disease characterized by ulceration in the mouth and genitalia, iritis, uveitis, arthritis, and thrombophlebitis.

DNA vaccination The use of DNA-encoding proteins to induce immune responses in subjects inoculated with it.

major histocompatibility complex The set of gene loci specifying major histocompatibility antigens (e.g., HLA in man, H 2 in mice, RLA in rabbits, RT 1 in rats, DLA in dogs, and SLA in pigs).

multiple sclerosis A human inflammatory disease affecting the central nervous system causing motor impairment, characterized by the gradual accumulation of focal plaques of demyelination particularly in the periventricular areas of the brain.

oral tolerance Administration of antigens orally to modulate immune responses to the given antigen.

paraneoplastic diseases A group of diseases in cancer patients affecting organ systems at sites remote from the tumor and mediated by secreted factors or immune reaction to tumor antigens cross-reactive to normal tissue antigens.

systemic lupus erythematosus A human disease characterized by multisystem involvement and the presence of anti-nuclear antibodies.

T cell receptor The antigen-recognizing receptor on the surface of T cells. Heterodimeric (disulfide linked), one member of the immunoglobulin superfamily of proteins, binds antigen in association with the major histocompatibility complex, leading to the activation of the cell.

T cell vaccination The use of T cells to treat autoimmune disease by enhancing the regulation over T cells mediating the disease.

I. INTRODUCTION

Most human brain diseases are not causally linked to the immune system; cerebrovascular and degenerative brain diseases, which account for the greatest toll in human life and suffering, do not directly involve the immune system. A minority of brain diseases are thought to be autoimmune—those caused by a mistaken attack of the immune system against various components of brain tissue. Autoimmune brain damage can be part of a generalized autoimmune disease such as systemic lupus erythematosus or Behçet's syndrome, or it can result from a brain-specific disease such as multiple sclerosis. Paraneoplastic diseases are another group of autoimmune neurological diseases; these conditions are triggered by an immune attack on tumor antigens that are cross-reactive with brain antigens.

Learning about human disease is greatly aided by the analysis of disease models in experimental animals. Models help answer questions related to the development of the disease and to the various mechanisms operating in its inception, perpetuation, and recovery. In addition, a model enables the development of novel therapeutic modalities resulting from our enhanced understanding of the disease process. Our understanding of autoimmune disease develops in parallel with our understanding of the immune system in general.

Since the study of self–nonself discrimination is at the core of immunology, the evolution of our understanding of autoimmune diseases has significantly changed in recent years. A comprehensive review of all autoimmune brain diseases is beyond the scope of this article; we will discuss general principles of autoimmunity. We will draw concrete examples, where needed, using multiple sclerosis (MS) as a prototype autoimmune brain disease, based on our understanding of experimental autoimmune encephalomyelitis (EAE), which is an animal model for MS. The general principles outlined for EAE and MS are applicable to other tissue-specific autoimmune diseases, such as rheumatiod arthritis, insulin-dependent diabetes mellitus, or inflammatory bowel diseases. In other diseases, the pathogenic autoantigen and autoreactive T cells are different but the mechanisms of disease regulation and recovery are similar. Thus, the rules developed and the modalities of treatment can be copied from one autoimmune disease to another.

MS is an inflammatory demyelinating disease (causing the destruction of myelin, the insulating coat of axons) affecting the white matter of the central nervous system. In the United States, more than 300,000 patients suffer from MS. Women are more frequently affected (the female to male ratio is 2:1). The disease often begins in the second to the fourth decade of life and results in paralysis, sensory disturbances, incoordination, and visual impairment. EAE will be discussed later.

II. BENIGN AUTOIMMUNITY

The mammalian immune system is composed of two classes of elements: innate and adaptive. The innate arm of the immune system is encoded in the germline of the species and is inherited by the individual from his or her parents. Innate immunity includes the macrophages, neutrophils, and other inflammatory cells and molecules. The adaptive arm of the immune system contains the T cells and the B cells, which express antigen receptors that are somatically generated in each individual. Activated B cells secrete specific antibodies of various isotypes. The adaptive arm of the immune system recognizes antigens and progressively learns from the person's ongoing immune experience; immune memory is an attribute of the adaptive arm of the immune system.

The individual's immune system develops in such a way that the effort to combat foreign invaders (bacteria, viruses, and other parasites) is usually successful, while the self-tissues are not damaged by autoimmune disease. This is achieved by a process of "education" in the thymus, in which most of the self-reacting T cells are purged from the mature immune cell repertoire. The thymus is an organ high in the thorax that serves as a "factory" for the generation of new T cells. This process of "negative selection" (killing of self-reactive T cells by contact with self-antigens) in the thymus is thought to be important for the prevention of autoimmunity. However, one can easily detect self-reacting antibodies and T cells in both humans and experimental animals. Analysis of normal human serum reveals a multitude of antibodies that bind normal tissue components. Immunization of experimental animals with self-antigens (proteins or peptides) results in the generation of self-reacting T cells that can induce autoimmune disease. Thus, the process of purging the immune repertoire from self-reactivity is far from complete, and there exist effective mechanisms that act to regulate the autoimmune repertoire in most individuals. Therefore, if the individual manages to live at peace with his or her autoimmune lymphocytes, what causes an autoimmune disease?

III. TRANSITION FROM BENIGN TO PERNICIOUS AUTOIMMUNITY

Fortunately, the process of transition from immune recognition of self to the full blown antiself attack is relatively uncommon; autoimmune disease affects about 5% of the population. For a disease to develop, several conditions must be met. The individual must have certain predisposing genetic features, including human leukocyte antigens (HLAs). HLA molecules are a set of surface proteins expressed on immune cells that are important in forming a complex with peptides to be recognized by T cells. In addition to HLA, hormonal factors, tissue susceptibilities, and others, along with an inciting event, trigger the development of the disease. This trigger is often an infection. The infectious agent (viral or bacterial) may have antigenic epitopes that activate T cells capable of attacking the self; the immune attack against a foreign invader starts an immune reaction that spills over and results in mistaken antiself attack. Alternatively, an emotional stress may result in loss of regulation over anti-self-reactivity that may evolve into pernicious autoimmunity. During emotional stress various hormones are secreted, including corticotropin-releasing factor, ACTH, and adrenal steroids, which can induce

apoptosis of both immature and memory lymphoid cells including regulatory cells. In many individuals, this early autoimmune attack activates regulatory immune mechanisms that will prevent future attacks of the disease. Thus, some patients with MS will have a single attack of optic neuritis (inflammation of the optic nerve causing transient blindness) with complete recovery and no further evidence of MS. However, in other patients, probably lacking efficient antidisease regulation, the disease runs a course of recurrent attacks leading to progressive damage. In contrast to pathogenic (disease-causing) T cells, which are well characterized, the regulatory cell population is less clearly defined. This population of cells is probably heterogeneous and contains (i) antiidiotypic T cells, which specifically recognize the T cell receptor (TCR)—the molecule used by T cells to recognize antigenic peptide in the MHC binding groove—of the pathogenic cells and have the ability to suppress them; (ii) antiergotypic T cells, which respond to activated T cells irrespective of their TCR; (iii) antigen-specific Th2 cells, which react to the pathogenic peptide but produce inhibitory cytokines that suppress the pathogenic cells; and (iv) natural regulatory T cells that are characterized by the expression of the α chain of the interleukin-2 (IL-2) receptor (CD25). In addition to regulation mediated by other cells in the pathogenic population, the disease-causing cells may undergo activation-induced apoptosis (cell death) following contact in the brain or other organs with death-inducing molecules, such as tumor necrosis factor and FAS.

IV. EXPERIMENTAL AUTOIMMUNE ENCEPHALOMYELITIS

EAE is a model in experimental animals that, due to similarities in the clinical and pathological features, is considered by many investigators to reflect human MS. EAE is probably the most intensively studied autoimmune disease model. The experimental animal (mouse, rat, guinea pig, or monkey) is injected with a protein from the central nervous system (CNS) myelin (myelin basic protein, proteolipid protein, or myelin oligodendrocyte glycoprotein) in a suitable adjuvant —a material added to the antigen to enhance the immune response (containing oil and bacterial products). Ten to 14 days after the injection, the animal shows signs of neurological damage, including paralysis of the tail and posterior and anterior paws and loss of urinary sphincter control. This neurological disease lasts from several days to several months (depending on the animal strain); in some animals there are repeated attacks of paralysis (similar to the remitting relapsing form of human MS). An identical form of EAE can be induced by injecting a T cell clone reactive to a myelin protein antigen. A T cell clone is composed of a homogeneous population of T cells with identical T cell receptors: All cells in a clone emerge from a single parental cell. This form of adoptive EAE usually starts earlier than actively induced EAE (within 3–5 days after T cell inoculation), and the animal shows identical symptoms to the actively induced disease. Animals that have recovered from EAE are usually resistant to another attempt to induce the disease. Postrecovery resistance to EAE is probably related to induction of efficient regulatory T cells since resistance can be transferred to other animals with T cells. In some studies, anti-MBP antibodies could also transfer resistance.

What have we learned from EAE? First, as in humans, there is a genetic predisposition to the disease; some mouse and rat strains are susceptible to the disease, whereas others are resistant. The cell causing the disease is a T helper type 1 (secreting IL-2 and γ-interferon and mediating tissue damage) reactive to myelin antigens. Upon activation, the pathogenic T cells acquire the capacity to migrate to the CNS and cause inflammation resulting in demyelination and paralysis. Some of the pathogenic T cells then undergo apoptosis (programmed cell death). In some animal models, there is spreading of the autoimmune reaction to other epitopes in the same myelin antigen or to other antigens. This molecular immune spreading may be a reason for additional attacks of the disease. Thus, the crucial elements of the autoimmune reaction in EAE are the trimolecular complex of the TCR reactive to the pathogenic peptide within the appropriate major histocompatibility complex (MHC) class II antigen (molecules, which include HLA in the human, are expressed on the surface of immune cells that in association with peptides form the trigger of TCRs). Following the activation of the T cell, it migrates to the target tissue and produces cytokines and other molecules that cause damage. Equally important is a lack or a failure of the regulator cells that normally operate to prevent autoimmune activation.

In other autoimmune reactions (such as paraneoplastic cerebellar degeneration, a disease of the cerebellum manifested by gait and movement abnormalities in patients with tumors in other tissues), the major players may be antibodies reacting against CNS antigens rather than T cells.

V. THERAPEUTIC INTERVENTIONS

The understanding that the autoimmune disease is caused by an immune response directed against self points to possible modes of therapeutic interventions. Following this understanding, it was possible to prevent or treat EAE by

1. Blocking MHC class II molecules
   - By antibodies that are specific to the MHC or to the complex of particular MHC molecules with the pathogenic peptide embedded in its binding groove.
   - Using peptides that bind the MHC and do not activate the pathogenic T cells; thus, the peptides block the interaction between the pathogenic T cells and the MHC.
   - Using peptides that bind both the MHC and the pathogenic TCR, but the effect of this interaction is silencing of the T cells. Such peptides are termed TCR antagonists. A related group of peptides, called altered peptide ligands, also bind the MHC and TCR, but the result of stimulation modifies or changes the effector function of the T cell, inducing the secretion of different cytokines or the creation of T cells incapable of mediating disease.
2. In addition to the TCR contacting its appropriate peptide in the MHC, efficient T cell activation needs an interaction called "the second signal" or T cell costimulation. In other words, the signal delivered to the T cell by the interaction of its TCR with the peptide–MHC complex is not sufficient to cause activation, and another trigger to a simultaneously bound surface molecule on the T cell is needed to enable the stimulation to proceed. This second signal is termed a costimulatory signal. Blocking this interaction by decoy molecules prevents T cell costimulation. Such interventions in animal models of autoimmune disease result in prevention or cure of the disease.
3. Blocking the TCR, using anti-TCR antibodies, either against a specific disease-causing TCR or against a family of TCRs known to be involved in the disease.
4. Tolerance induction by administering the pathogenic antigen in a form that will abort the immune response to it. An example is administering the antigen orally or intravenously.
5. Recently, several researchers have examined the possibility of using DNA instead of protein or peptides to induce forms of tolerance. Using DNA encoding the TCR V β 8.2, EAE was prevented and the anti-MBP T cell response was shifted from Th1 (secreting interferon γ and IL-2) to Th2 (secreting IL-4 and -5). Others have used DNA-encoding myelin antigens to prevent EAE.
6. After pathogenic T cells have been activated in the periphery, they have to cross the blood-brain barrier to enter the site of the disease. The adhesion of activated T-cells to endothelial cells is mediated by the T cell molecule very late antigen type-4 (VLA-4; a protein expressed on the surface of T cells late after activation) interacting with the vascular cell adhesion molecule (a member of intercellular adhesion molecules) on endothelial cells. Blocking this interaction by monoclonal antibodies can prevent EAE. Clinical studies with a humanized version of anti-VLA-4 are now in phase II clinical trials in MS. The migration of T cells across the extra cellular matrix involves the secretion of enzymes such as matrix metalloproteinases. Inhibition of these enzymes is another way of treating EAE.
7. Once the pathogenic T cells have reached the target tissue, the damage is inflicted in part by cytokines secreted by these cells. Thus, one way to treat autoimmune diseases involves the neutralization of cytokine effects (e.g., anti-tumor necrosis factor antibodies or IL-1 antagonists). These treatments were found to be successful in patients with rheumatoid arthritis.
8. An additional mode of immunointervention is to enhance the immune regulation over the pathogenic T cells by procedures such as T cell vaccination and TCR peptide vaccination or by injection of anti-idiotypic T cell lines (T cells that are specific for the TCR of the pathogenic cell and have the functional capacity to suppress the disease causing-cell).

VI. TREATING THE HUMAN DISEASE

Current treatments for many human autoimmune diseases are largely based on immunosuppression, usually by giving medications such as corticosteroids, cyclophosphamide, azathioprine, or cyclosporin A, which inactivate or kill all classes of lymphocytes. In many patients these medications are effective in the short term only; their administration does not usually alter the natural history of the disease, and the toxicity of these drugs is considerable. This toxicity is related to nonselective immunosuppression, which results in susceptibility of the patient to infections and malignancies. In addition, these medications have considerable side effects (such as in the case of corticosteroid

drugs: elevated glucose levels, hypertension, obesity, osteoporosis, and psychiatric disturbances). Thus, the ideal therapy for autoimmune disease should affect the pathogenic clone or clones specifically without suppressing the entire immune system; it should be devoid of toxicity; and it should be easily administered. Although considerable efforts have been made to improve the treatment of autoimmune disease, none of the current immunosuppressive therapies are satisfactory.

Many of the procedures found to be successful in experimental animals have raised hopes (often unfulfilled) in the scientific community; many successful experimental procedures are being translated to the clinical scene and have undergone clinical trials in patients with MS and other autoimmune disorders.

Oral tolerance to myelin antigens, anti-CD4–T cell monoclonal antibodies, T cell receptor vaccination, T cell vaccination, stem cell replacement, intravenous pooled immunoglobulins, and copaxone have all been developed in EAE and attempted in human MS. Currently, of these novel treatments, the most widely used in MS patients are β-interferon and copaxone. In general, the treatment modalities developed as a result of experimental studies have proven to be safe and show beneficial effects in treated patients. However, since many of the patients enrolled in such studies have failed other treatments and are in an advanced stage of disease, it is difficult to show complete recovery because some of the damage inflicted is not reversible.

VII. FUTURE DIRECTIONS

The major force in the introduction of new and effective treatment for autoimmune diseases is our enhanced understanding of immunology in general and control of immune cell activation in particular. As we discover the key players in T cell activation and the molecular tools to abort an immune response in its early stages without adversely affecting the millions of other beneficial immune cells, we will be able to introduce novel and more effective treatments. Since these diseases cause considerable suffering and loss of life, the sense of urgency has shortened the time in which clinical applications of inventions in the laboratory are carried out.

See Also the Following Articles

BEHAVIORAL NEUROIMMUNOLOGY • HIV INFECTION, NEUROCOGNITIVE COMPLICATIONS OF • PSYCHONEUROIMMUNOLOGY

Suggested Reading

Albert, L. J., and Inman, R. D. (1999). Molecular mimicry and autoimmunity. *N. Engl. J. Med.* **341,** 2068–2074.

Cohen, I. R. (2000). *Tending Adam's Garden: Evolving the Cognitive Immune Self*. Academic Press, San Diego.

Critchfield, J. M., Racke, M. K., Zuniga-Pflucker, J. C., Cannella, B., Raine, C. S., Goverman, J., and Lenardo, M. J. (1994). T cell deletion in high antigen dose therapy of autoimmune encephalomyelitis. *Science* **263,** 1139–1143.

Evavold, B. D., Sloan, L. J., and Allen, P. M. (1993). Tickling the TCR: Selective T-cell functions stimulated by altered peptide ligands. *Immunol. Today* **14,** 602–609.

Gijbels, K., Galardy, R. E., and Steinman, L. (1994). Reversal of experimental autoimmune encephalomyelitis with a hydroxamate inhibitor of matrix metalloproteases. *J. Clin. Invest.* **94,** 2177–2182.

Kuchroo, V. K., Das, M. P., Brown, J. A., Ranger, A. M., Zamvil, S. S., Sobel, R. A., Weiner, H. L., Nabavi, N., and Glimcher, L. H. (1995). B7-1 and B7-2 costimulatory molecules activate differentially the Th1/Th2 developmental pathways: Application to autoimmune disease therapy. *Cell* **80,** 707–718.

Lobell, A., Weissert, R., Storch, M. K., Svanholm, C., de Graaf, K. L., Lassmann, H., Andersson, R., Olsson, T., and Wigzell, H. (1998). Vaccination with DNA encoding an immunodominant myelin basic protein peptide targeted to Fc of immunoglobulin G suppresses experimental autoimmune encephalomyelitis. *J. Exp. Med.* **187,** 1543–1548.

Pisetsky, D. S. (2000). Tumor necrosis factor blockers in rheumatoid arthritis. *N. Engl. J. Med.* **342,** 810–811.

Racadot, E., Rumbach, L., Bataillard, M., Galmiche, J., Henlin, J. L., Truttmann, M., Herve, P., and Wijdenes, J. (1993). Treatment of multiple sclerosis with anti-CD4 monoclonal antibody. A preliminary report on B-F5 in 21 patients. *J. Autoimmunity* **6,** 771–786.

Scaravilli, F., An, S. F., Groves, M., and Thom, M. (1999). The neuropathology of paraneoplastic syndromes. *Brain Pathol.* **9,** 251–260.

Sette, A., Alexander, J., Ruppert, J., Snoke, K., Franco, A., Ishioka, G., and Grey, H. M. (1994). Antigen analogs/MHC complexes as specific T cell receptor antagonists. *Annu. Rev. Immunol.* **12,** 413–431.

Van Parijs, L., and Abbas, A. K. (1998). Homeostasis and self-tolerance in the immune system: Turning lymphocytes off. *Science* **280,** 243–248.

Waisman, A., Ruiz, P. J., Hirschberg, D. L., Gelman, A., Oksenberg, J. R., Brocke, S., Mor, F., Cohen, I. R., and Steinman, L. (1996). Suppressive vaccination with DNA encoding a variable region gene of the T-cell receptor prevents autoimmune encephalomyelitis and activates Th2 immunity. *Nat. Med.* **2,** 899–905.

Weiner, H. L., Friedman, A., Miller, A., Khoury, S. J., al-Sabbagh, A., Santos, L., Sayegh, M., Nussenblatt, R. B., Trentham, D. E., and Hafler, D. A. (1994). Oral tolerance: Immunologic mechanisms and treatment of animal and human organ-specific autoimmune diseases by oral administration of autoantigens. *Annu. Rev. Immunol.* **12,** 809–837.

Zhang, J., and Raus, J. (1995). Clonal depletion of human myelin basic protein-reactive T-cells by T-cell vaccination. *Ann. N. Y. Acad. Sci.* **756,** 323–326.

Axon

KATHLEEN S. ROCKLAND
Brain Science Institute, RIKEN, Saitama, Japan

GLOSSARY

anterograde tracers Substances taken up by and transported from the cell body toward the distal portion of an axon. After appropriate processing, these reveal which target structures receive connections from the injected region.

arbor The distal terminal portion of an axon, which carries the synaptic specializations.

axon a single process, usually emitted from the cell body, with membrane specializations adapted to the conduction of the nerve impulse.

bouton (French "button") General term that refers to the light microscopic appearance (swellings, beads, or stalked endings) of what are usually synapses. Some swellings, however, may be just accumulations of mitochondria, without synaptic vesicles. Boutons can also be called terminal specializations.

dendrites Extensions, usually multiple, of the cell body. Dendrites receive synaptic inputs and are usually much shorter than axons. For some neurons, such as local circuit neurons, dendrites and axons occupy about the same volume.

feedback A type of cortical connection that originates mainly from neurons in layer 6, terminates in layer 1, and has an elongated arbor (>1.0 mm long).

feedforward A type of cortical connection that originates mainly from neurons in layer 3, terminates in layer 4, and has one to four small arbors (≤ 0.2 mm in diameter).

retrograde tracers Substances taken up by synaptic endings and transported back to the cell body. After appropriate processing, these reveal which areas send projections to the injected region.

synapse (Greek "to fasten together") Intercellular junctions that are specialized for the transmission of nerve impulses. Neurotransmitter substances are packaged within synaptic vesicles, as seen by electron microscopy.

Almost all nerve cells, and only nerve cells, have an axon. This structure can be several microns (50–200) to several centimeters in length, and is specialized for conduction of the nerve impulse This article begins with a brief historical survey, followed by a discussion of the cell biology of axons and network properties. Because much of the experimental research on axons employs invasive techniques, this article emphasizes work carried out in animal models. The current understanding, supported by selective verifications, is that the results in most cases will extrapolate to humans, at least from closely related mammalian phyla.

I. INTRODUCTION

Nerve cells consist of three components: the cell soma, dendrites, and axon (Fig. 1). There are usually multiple dendrites and a single axon (there are instances of neurons with no axon and, rarely, of neurons with multiple axons). Dendrites are frequently viewed as extensions of the cell body and are the primary target for synaptic inputs to the neuron. Unlike the dendrites or cell body, the axon does not contain ribosomes, the apparatus of protein manufacture. Via the excitable plasmalemma and other specializations, it is mainly

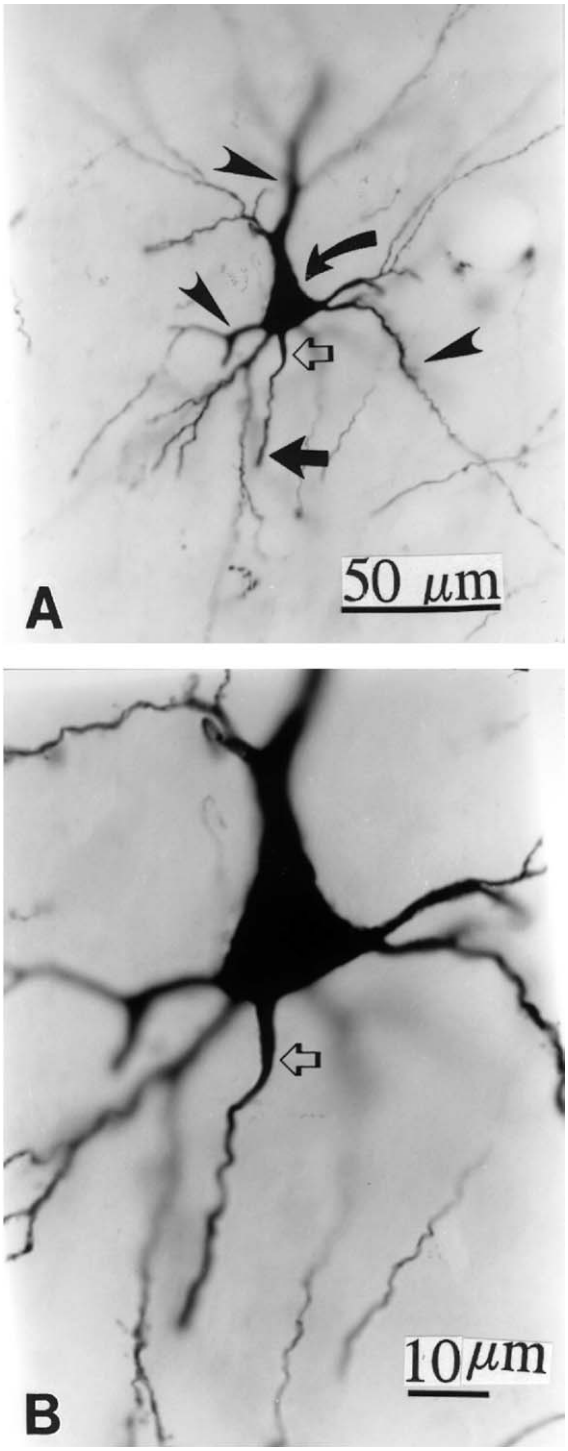


Figure 1 Photomicrographs at lower (A) and higher (B) magnifications of a cortical pyramidal neuron in layer 3, retrogradely labeled from an injection of tracer (biotinylated dextran amine). Curved arrow, cell body; arrowheads, three of the dendrites; open arrow, axon hillock; solid arrow, descending portion of axon (this continues toward the white matter but has been cut off in this section). Open arrows point to equivalent features in A and B.

concerned with conduction and transmission of the nerve impulse and other signals.

Modern investigations of the axon are proceeding on several fronts, namely, the cell and molecular biology of the axon, especially concerning the mechanisms of axoplasmic transport; the specializations and mechanisms of the plasmalemma as an excitable membrane; and the network properties of the axon as a constituent of neural architectures. These issues are of major importance for further understanding of the basic science of neural function and bear directly on pathological conditions, such as peripheral neuropathies, demyelinating diseases, and diffuse axon injury subsequent to trauma. Recent progress has been rapid in all these areas owing to dramatic technical advances. The technical armament includes physiological techniques such as nodal voltage clamping and optical methods for studying electrical activity, sophisticated immunocytochemical and ultrastructural methods, and an array of fine molecular tools.

II. HISTORICAL SURVEY

The early morphological description of nerve fibers has been credited to Antoine van Leeuwenhoek (1718), who studied nerve bundles from cow or sheep spinal cord under the microscope. In his report, however, the individual fibers are described as hollow tubes. In cross section, given the suboptimal preservation of the specimens, the dark core of axoplasm would be shrinken, and myelinated axons would therefore appear hollow.

Further characterization of the axon, through the second half of the 19th century, was hindered by technical limitations as well as by the controversies surrounding the basic organization of nerve cells. One of the most important contributions during this period was made by O. F. K. Deiters. By microdissection of individual nerve cells from histologically treated specimens of spinal cord, Deiters successfully established the continuity of the soma and its processes and distinguished between the protoplasmic processes (''dendrites'') and what he called the more slender axis cylinder. He correctly concluded that the nerve cell is a cell that bears on its soma one axon and several dendrites.

The advent of the Golgi silver stain (1883), which demonstrates a small number of neurons, often in their complete or near complete entirety, stimulated a highly creative period in cellular neuroanatomy, culminating in the Nobel prize awards in 1906 to Camillo Golgi and

Ramon y Cajal. These morphological studies, especially those of Cajal, firmly established the neuron as the basic cellular unit of the brain, as opposed to the "reticularist" position promulgated by Golgi. They also clarified the dynamic polarization of the neuron, as a cell with receiving and transmitting elements (respectively, the multiple dendrites and single axon).

With this work, the investigation of axons rapidly expanded to their functional anatomy and their connectivity and network properties. Axons were defined as two major subtypes, those with local ramifications and those with extensive ramifications (Fig. 2). These became known as Golgi type II and type I, respectively (or local circuit and projection neurons). These two types were later found to correspond generally to inhibitory and excitatory connections, each using different neurotransmitters. In the same period, many of the specific sensory, motor, and associational pathways interconnecting subcortical nuclei and cortical areas were successfully mapped.

III. CELL BIOLOGY OF THE AXON

The fine structural organization of axons has been extensively investigated since the mid-1950s using various forms of electron microscopy (EM), such as standard transmission, high voltage, and EM combined with freeze etching. From such studies, we know that the axon cytoplasm contains many of the same components as the neuronal soma and dendrites; that is, mitochondria, multivescicular bodies, agranular endoplasmic reticulum (ER), neurofilaments, and microtubules. Axons, however, do not have free ribosomes or granular ER and therefore do not engage in protein synthesis (Fig. 3; Table I).

A. Subdivisions

Axon structure can be analyzed in several ways. One way is to distinguish between cytoskeletal and membranous components. The cytoskeleton is considered to comprise three domains linked in a microtrabecular meshwork: the microtubules, neurofilaments, and, adjoining the peripheral membrane, the actin microfilaments. Neurofilaments are a type of intermediate filament, about 10 nm in diameter, which consist of a species-specific ratio of three proteins (68, 150, and 200 kDa). Microtubules are somewhat larger (20–26 nm in diameter), and consist of two globular polypeptides, α and β tubulin, 50 kDa each. Microtubules have

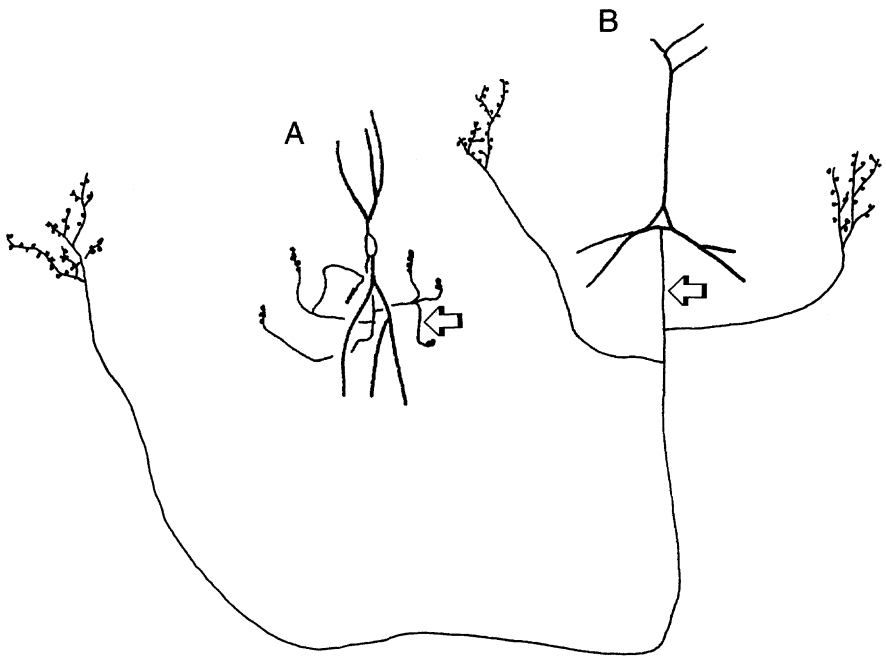


Figure 2 Schematic illustration of (A) local circuit neurons, whose axons arborize in the immediate vicinity of the cell body, and (B) a cortical projection neuron, whose axon typically projects to distant targets, several millimeters or centimeters away. Open arrows point to axons. The caliber of the dendrites has been exaggerated to distinguish these from the axon.

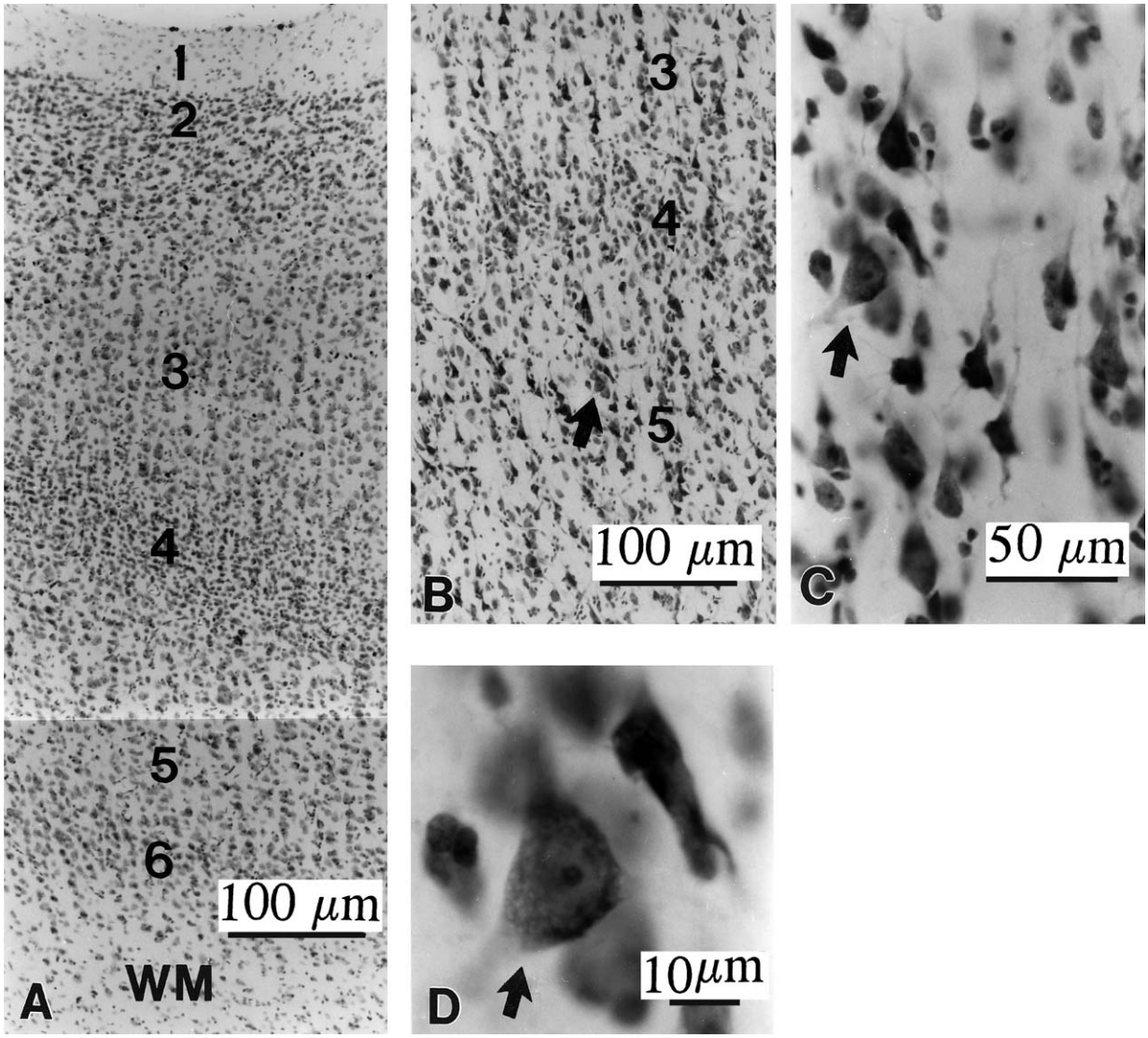


Figure 3 (A) Overview of the six cortical layers in monkey (from visual association cortex). (B) Portion through the middle cortical layers of human temporal cortex from a surgical specimen. (C, D) Progressively higher magnification from the arrow in B. In D, clumps of rER are visible in the cell body. Numbers denote the conventional cortical layers 1–6. WM, white matter.

polarity, by conventional designated as "+" or "−" (respectively nearest the distal or proximal cell body region). Head-to-tail polymerization occurs, with subunits added to the plus end and released from the minus end ("treadmilling"). There are several microtubule-associated proteins (MAPS) that protrude as side arms and contribute to the microtrabecular meshwork.

Neurofilaments predominate in large axons, but microtubules predominate in small axons, and the total number of both structures is proportional to the caliber of the axon (Fig. 4). These structures play a role in skeletal support as well as in intracellular transport of ions, metabolites, and vesicles.

A major membranous component of the axon is the agranular ER, which is thought to consist of two subsystems: (i) clusters of tubules and flattened sacs at the outer wall of the axons and (ii) a complex of narrow tubules and sacs oriented parallel to the long axis of the axon, which is believed to extend the full length of the axon. The agranular ER has been thought to contribute to axonal transport, and it possibly contributes to the sequestration of Ca^{2+} ions and the provision of membrane for forming synaptic vesicles.

Table I
Characteristics of Axons and Dendrites[a]

Axon	Dendrite
1. Extends from either cell body or dendrite	1. Extends from cell body
2. Begins with a specialized initial segment (except dorsal root ganglion cell and autonomic ganglion cell)	2. At least in proximal portions, continues cytoplasmic characteristics of cell body
3. May be absent as in amacrine cells of retina	3. May be absent as in dorsal root ganglion cell
4. Unique in most cells, but there are some examples of multiple origin	4. Usually multiple
5. May be myelinated or unmyelinated	5. Rarely myelinated (and if so, only thinly)
6. Almost never contains ribosomes (Except in initial segment)	6. Contains granular endoplasmic reticulum or ribosomes, diminishing with distance from origin
7. Usually has smooth contours and cylindrical shape	7. Usually has irregular contours and specialized appendages
8. Usually is the thinnest process of the cell at site of origin	8. Usually originates as a thick tapering process
9. Ramifies by branching at obtuse angles	9. Ramifies by branching at acute angles
10. Usually gives rise to branches of the same diameter as parent stem	10. Usually subdivides into branches smaller than parent stem
11. Ramification can be close to cell body or at great distances; may extend long distances away from cell body, even into peripheral nervous system	11. Ramification usually confined to the vicinity of the cell body; if cell body lies in central nervous system, dendrites remain entirely within central nervous system
12. Neurofilaments predominate in larger axons	12. Microtubules predominate in larger stems and branches
13. Capable of generating action potentials, propagating them, and synaptic transmission	13. Conducts in a decremental fashion, but may be capable of generating action potentials
14. Primarily concerned with conduction and transmission	14. Primarily concerned with receiving synapses.

[a]From *The Fine Structure of the Nervous System: Neurons and Their Supporting Cells, Third Edition* by Alan Peters, S. L. Palay, and H. Webster, Copyright © 1990 by Alan Peters, Used by permission of Oxford University Press, Inc.

Many long, type I axons have an associated specialization, the myelin sheath of the plasmalemma, which is elaborated by glial processes. The myelin wrapping is interrupted by nodes of Ranvier, which are involved in the process of saltatory impulse conduction.

Axons can also be subdivided into several specialized portions along their proximal to distal length. The proximal portion, at the juncture with the cell body, is termed the axon hillock (Fig. 1). This region is characterized by fascicles of microtubules and a high density of sodium-dependent channels. The high density of channels accords with the importance of this region as a "trigger zone" for the initiation of axon impulses.

Adjoining the axon hillock distally is the initial segment (20–50 nm in length). This portion has clusters of microtubules and an undercoating of dense material below the plasmalemma. The initial segment leads into the body of the axon. This can range from <0.5 mm in the case of type II, local circuit axons to several centimeters or even meters in the case of axons projecting from motor cortex to the lower spinal cord of larger animals. Axons in some systems can send branches ("collaterals") to different target structures. Finally, at the distalmost portion, axons form terminal arbors (Fig. 5). Arbors consist of preterminal and terminal portions. The preterminal portions are where the axon begins to branch repeatedly, and, if myelinated, loses the myelin sheath. Terminal portions are thin-caliber branches, decorated with synaptic terminal specializations.

B. Axoplasmic Transport

The axoplasm is not static but, rather, participates in a state of dynamic flow. There are at least two reasons for this. One is the necessity for maintenance of the

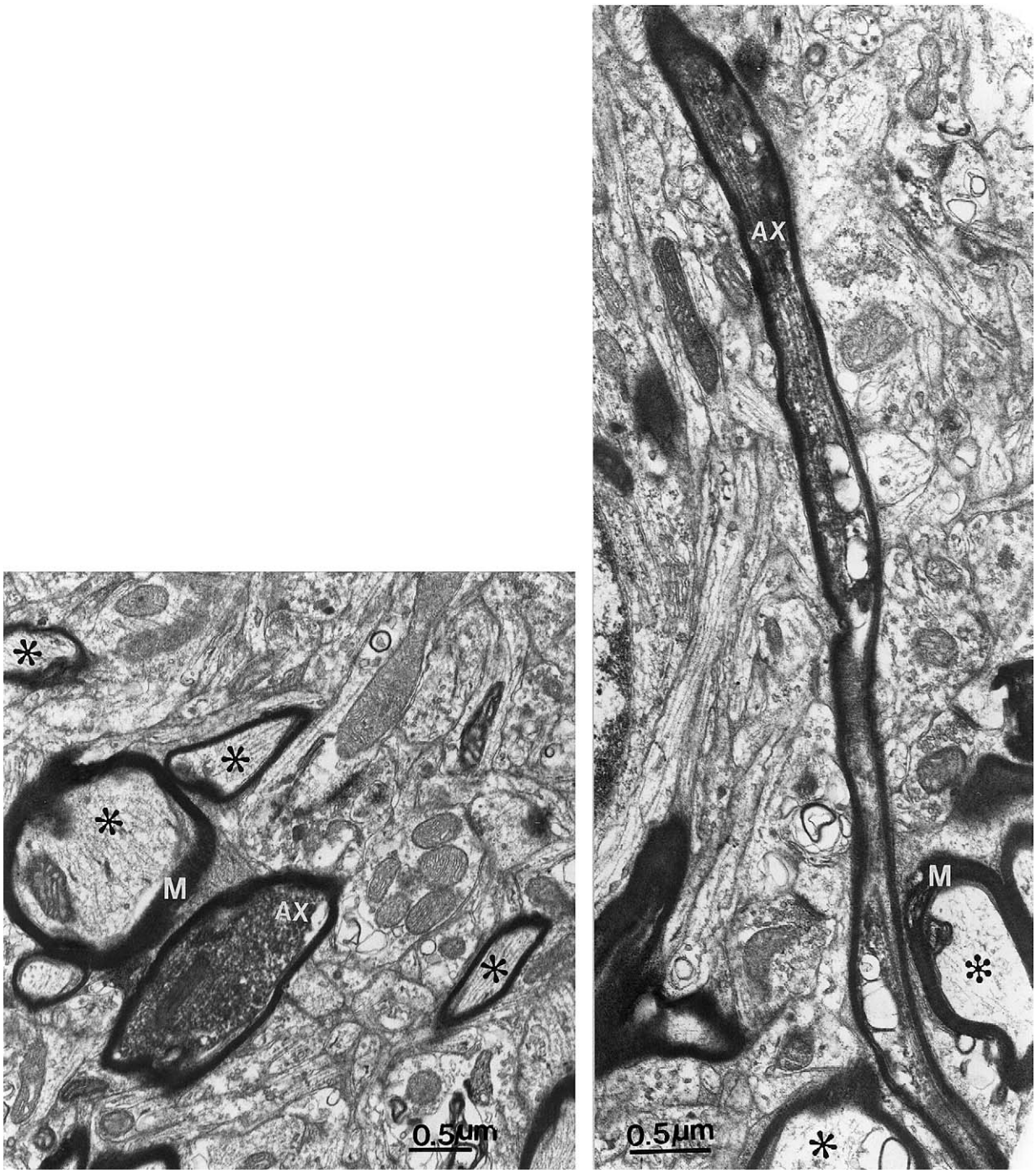


Figure 4 Electron micrographs of monkey cortical tissue. Two axons (AX) have been anterogradely filled with tracer transported from a distant injection site. Several other unlabeled axons occur in the same field (asterisks). M, myelin.

cytoskeleton, plasmalemma, synaptic apparatus, and other organelle components. Since little or no protein synthesis takes place within the axon proper, materials synthesized within the cell body must be transported along the full length of the axon. A second, less understood requirement relates to the encoding and regulatory functions of the axon, whereby conditions at its distal end are signaled back to the cell body.

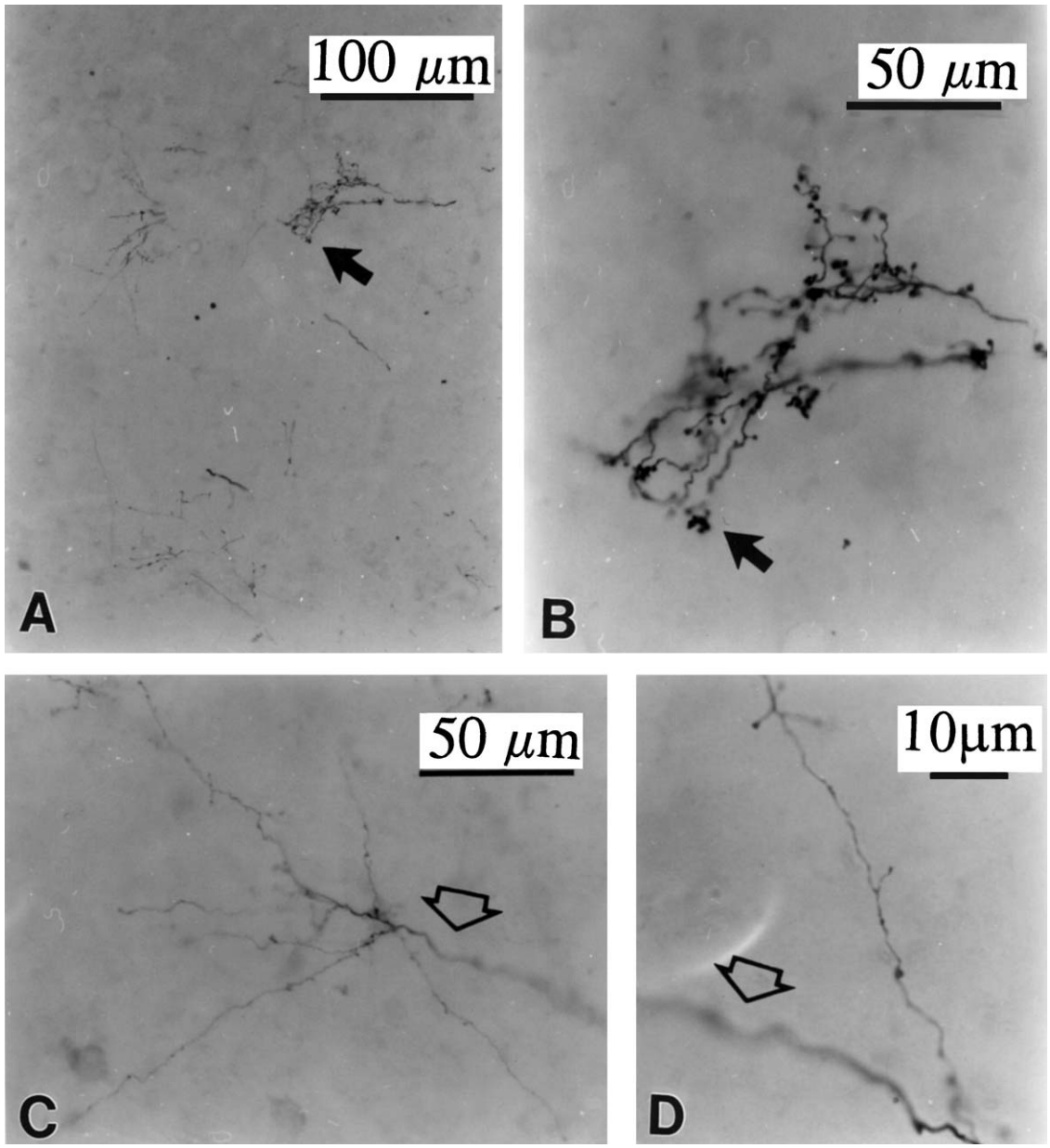


Figure 5 Photomicrographs of two corticocortical axon arbors anterogradely filled with tracer (BDA) transported from a distant injection site. Both arbors are about the same size. One (A, B) has very large terminal specializations. The other (C, D) has delicate beaded and stalked endings and is more typical. B is higher magnification of A. D is higher magnification of C. Arrows point to corresponding features.

Axon transport is a complex process that operates in both the anterograde (toward the distal end with terminal arbors) and retrograde (toward the cell body) directions (Fig. 6). There are several subcomponents that require many different mechanisms. The main distinction is between fast and slow transport. Fast transport operates in both the anterograde and retrograde directions and comprises several rate classes. These may reflect differences in transport mechanisms and/or the sieving action of the cytoskeletal meshwork on organelles of different sizes. Small vesicles and neurotransmitter molecules are conveyed at the fastest velocity (200–400 mm/day), mitochondria at 50–100 mm/day, and various metabolic enzymes at

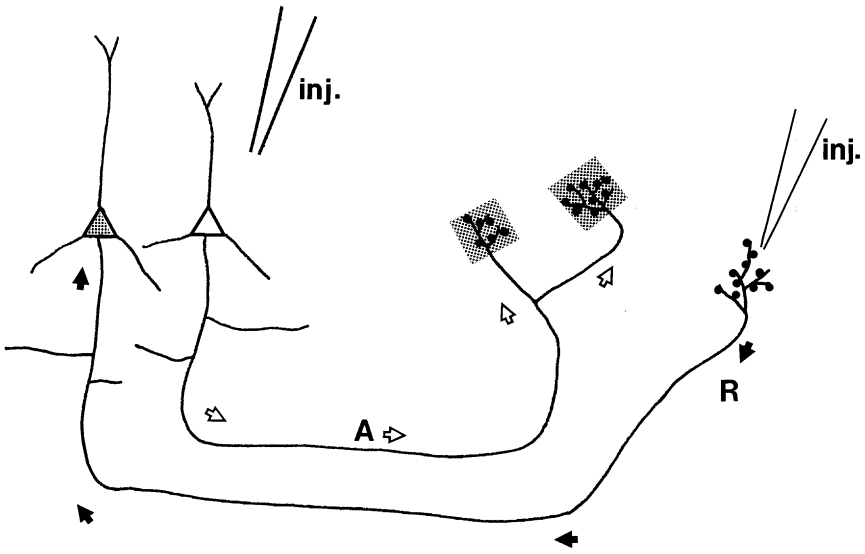


Figure 6 Schematic representation of axonal transport in the anterograde (A) and retrograde (R) directions.

28 mm/day. Microtubules and neurofilaments migrate by slow transport, which proceeds at less than 5 mm/day.

1. Fast Transport

Early work on axon transport took place in the 1940s, in part motivated by interest in axon regeneration during World War II. The phenomenon was simply but convincingly demonstrated by the accumulation of an irregular bulge of material proximal to a ligature placed around a single axon. When the constriction was removed, the accumulated material continued down the axon at a rate of about 1.0 mm/day. Fast transport (i.e., at rates of 50–400 mm/day) was demonstrated several decades later by the uptake of radioactively labeled amino acids by the cell soma. The label is incorporated into proteins and transported down the axon, where the pulse of radioactive material is then sampled by either autoradiography or scintillation counting.

Recently, investigations of axon transport have shifted from elucidation of what materials are transported and at what rate to unraveling the underlying mechanisms. Indirect evidence had already suggested the importance of microtubules. For example, agents such as colchicine that disrupt microtubules, also interfere with transport. More direct evidence of the role of microtubules in fast transport was provided by studies using the compound β, β_1-iminodiproprionitrile, which causes microtubules to segregate from neurofilaments and form several bundles in the center of the axon. Electron microscopy and autoradiography show a clear association of mitochondria and other rapidly transported elements with the central clusters of microtubules. Finally, enhanced videomicroscopy of microtubules extruded from axoplasm indicates that these can transport organelles and vesicles.

In the transport process, microtubules operate in concert with several "molecular motors" (Fig. 7). The energy-transducing protein, kinesin transports organelles anterogradely along microtubules. Dynein, another microtuble-associated ATPase, is implicated in retrograde transport. The role of kinesin has been established experimentally. Injection of antisense oligonucleotides suppresses kinesin heavy-chain expression and concurrently results in abnormal accumulation of protein within the cell body. Additionally, kinesin heavy-chain mutations produce various disruptions in action potential propagation and neurotransmitter release.

Recent studies have identified additional carboxy-terminal-type proteins within the kinesin superfamily that participate in the transport of specific organelles at specific velocities. Most of these putative motors have been characterized by molecular cell biological approaches, such as cloning and sequencing of the genes encoding these proteins, expression and purification of the proteins using the baculovirus Sf9 cell system, observation of molecular structures by EM,

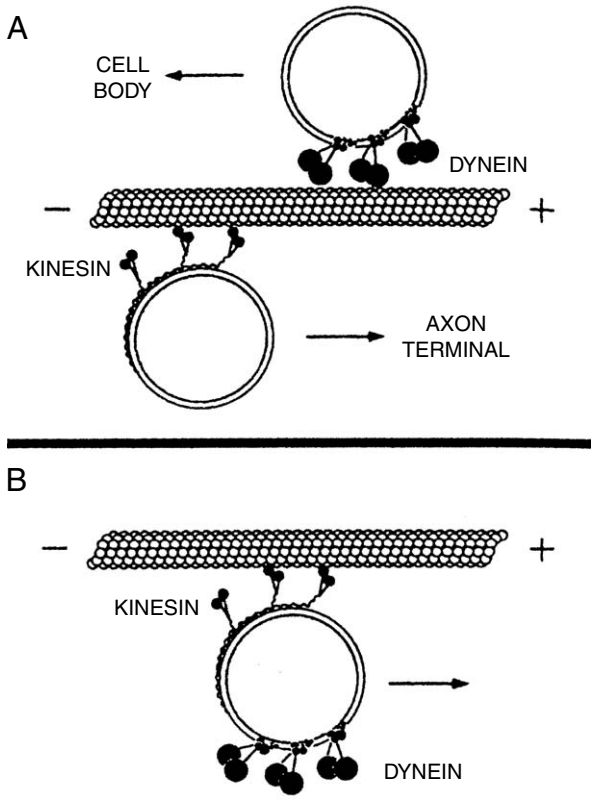


Figure 7 Alternative models for the function of molecular motors in fast axonal transport. (From *The Axon: Structure, Function and Pathophysiology* by S. G. Waxman, J. D. Kocsis, and P. K. Stys. Copyright © 1995. Used by permission of Oxford University Press, Inc.).

in vitro motility assays, immunocytochemistry, and identification of particular cargoes by subcellular fractionation and immunoprecipitation.

2. Slow Transport

Less is known about the mechanisms of slow transport. The polymer sliding model proposes that cytoskeletal proteins, assembled in the cell body, travel down the axon in the form of individual microtubule and neurofilament polymers. Several recent experiments, however, suggest that cytoskeletal proteins are transported as small oligomers along microtubules. For example, in a modern variant of classical metabolic labeling studies, a recombinant adenoviral vector, encoding epitope-tagged neurofilament-M protein, was used to infect neurons of a transgenic mouse line in which axons normally lack neurofilaments. The viral-encoded subunits do not form polymers, but confocal and electron microscopy demonstrated that these nevertheless were transported (at a slow rate of 5 mm/day) and that transport occurred in the close vicinity of microtubules.

IV. METHODS FOR INVESTIGATING CONNECTIVITY

The excitable axonal membrane is specialized for the propagation and conduction of the nerve impulse. Via synaptic connections with dendrites and, to a lesser extent, neuronal somata and other axons, axons are the chief means of "information transfer" within neural networks.

The organization of connectional systems has been investigated by a variety of experimental techniques, mostly in animal models. A large body of work has been devoted to simply mapping the complex interconnections of different structures. Much of this work has utilized extracellular injections tracer substances, which typically cover a tissue volume of 0.1–2.0 mm in diameter.

A. Anterograde Tracers

Anterograde tracers, such as ^{3}H amino acids or WGA-horseradish peroxidase (HRP), are taken up by the thousands of cell bodies within the injection site, transported anterogradely in different compartments through the axon, and accumulated along the axonal membrane and/or at the terminal specializations. Histological processing reveals label in structures that receive connections from the injected region (Fig. 6). Recently, a new generation of glycoconjugate anterograde tracers (biocytin, biotinylated dextran amine, and kidney bean lectin) have been introduced that bind to the plasmalemma and, after appropriate histological processing, result in morphologically detailed images of axon structure, including branch points and fine terminal specializations.

B. Retrograde Tracers

Retrograde tracers such as the enzyme HRP and several fluorescent dyes (including a kind of latex paint) are taken up by synaptic terminations within the injection site. They are transported retrogradely through the axon process to the cell body. Thus, labeled neurons are visualized in those structures that send projections to the injected region.

C. Electrophysiology

Electrophysiological investigations define receptive field and other functional properties. These are carried out by intra- or extracellular recordings, usually from the soma but also from axon processes. Axonal conduction velocity is another important parameter. This is analyzed by measuring the latencies of spikes evoked by antidromic activation of axons by electrical stimulation. Temporal relationships between the firing times of two neurons can also be assessed by cross-correlation techniques, although these are more often used for inferring whether two neurons are connected.

A particularly elegant approach is intracellular or intraaxonal microelectrode recording combined with intracellular microinjections of tracers such as HRP or biocytin. In this way, physiological characteristics can frequently be correlated with morphological specializations. For example, thalamocortical connections to primary visual cortex are subdivided into several functionally distinct categories concerned with form or motion vision. Axons conveying "information" related to form (from the parvocellular layers of the lateral geniculate nucleus) have properties consistent with achieving fine spatial resolution. Among other specialized features, their terminal arbors are smaller than those of axons concerned with motion vision (from the magnocellular layers).

D. Techniques Used in Humans

1. Microstructure

In humans, methods for direct investigation of axon connectivity are restricted to noninvasive approaches. At the microlevel of individual axons, morphological techniques are limited to a small number of methods that can be applied in postmortem tissue or surgical biopsies. First, the classical Golgi stain can demonstrate aspects of axon structure, especially in young tissue, although the identity and full conformation of long axons are difficult to establish. Second, degeneration methods can be used with moderate success. In these methods (the precursors to the more physiological modern methods, which utilize axon transport of injected tracers), tissue is treated in a series of solutions that stain fragments of axons degenerating as a result of damage related to stroke or some other traumatic process. Third, some labeling can be achieved by intracellular fills of neurons or axons in tissue slices *in vitro*, but this technique is necessarily restricted to shorter connections or only incomplete portions of longer connections. Fourth, one class of tracers, lipophilic carbocyanine dyes (DiI and DiO), has been found to bind to the plasmalemma, even in postmortem tissue, and transport by diffusion. These tracers produce high-resolution, Golgi-like labeling, but only within a distance of 2–5 mm of the injection spot. Thus, although evidence supports the applicability of animal data to the human brain, there is a serious need for new techniques that might allow direct investigation of fine axon connectivity in human tissue. Finally, there is an increasing number of antibodies available for immunohistological investigation of neural structures. Many of these (e.g., antibodies against peptides, calcium binding proteins, transmitters, receptors, or cytoskeletal elements such as neurofilaments) can successfully be used in human biopsy or postmortem tissue. Most of these markers, however, are for cellular or subcellular components and not long-distance axons.

2. Macrostructure

In humans, long-distance cortical connections have been easier to investigate at the macrolevel of axon bundles or tracts. Major tracts include cortical commissural fibers crossing in the corpus callosum to the contralateral hemisphere, projection fibers to and from subcortical nuclei (such as the thalamus, basal ganglia, and colliculi), and ipsilateral cortical association fibers.

The various white matter tracts were grossly identified more than 100 years ago by methods such as blunt dissection or myelin staining subsequent to localized lesions. These techniques were successfully used in identifying pathways such as the uncinate or arcuate fasciculi (respectively linking temporofrontal or occipito- and parietofrontal fields) or the cingulum (frontoparahippocampal) bundle.

Recently, specific pathways or tracts have been visualized in their stem portion by diffusion-weighted magnetic resonance (dwMR; "stem" refers to the compact coherent middle portion of an axon bundle, contrasted with its more fanned-out proximal or distal extremes). This technique detects the diffusivity of water molecules in three dimensions. For oriented tissue, such as fiber bundles, this diffusivity tends to be anisotrophic. Images of the location, size, and trajectory of major white matter bundles acquired by dwMR accord well with earlier postmortem studies based on histological myelin stains.

Visualization of white mater bundles in humans is important for issues of both clinical and basic science. In clinical neurology, there are many conditions, such as stroke, multiple sclerosis, amyotrophic lateral sclerosis, and traumatic head injury, that lead to pathological alterations in white matter tracts. Imaging techniques allow improved *in vivo* diagnosis and analysis of these conditions. These techniques also offer immense promise as a probe of normal connectivity in the human brain.

3. Electrophysiology

Electrophysiological studies of axon organization in humans are also highly constrained by the need for noninvasiveness. Considerable information, however, is available from methods such as evoked potentials, which can record populational responses through the skull. Evoked potential recordings can be in response to presented stimuli or, recently, to transcranial magnetic stimulation. Electroencephalogram coherence analysis is another method for studying connectivity in both the human brain and the animal brain.

The need for precise localization during neurosurgical procedures provides another source of data. Surface electrodes are routinely used to map regions involved in cortical language processing, and in certain conditions depth electrodes may be placed in structures such as the amygdala, hippocampus, or substantia nigra.

V. NETWORK PROPERTIES OF AXONS

Long axons form connections between structures. Connections are often schematized by arrows (e.g., 'area A $\rightarrow$ area B'), but this is really a shorthand notation for complex processes, such as convergence (how many axons converge on a given postsynaptic target), divergence (how many postsynaptic neurons are contacted by a single axon), and combinations and integration of multiple inputs.

Network properties are difficult to investigate, even in animal models. To some extent, we know what is connected to what, but only partial and incomplete data exist concerning biophysical parameters and dynamic properties. A recent, powerful approach is the combination of intracellular labeling and EM analyses with dual or triple intracellular recordings *in vitro*. This combined anatomical–physiological approach provides highly precise data regarding the number and location of synapses between interconnected neurons and also regarding the functional dynamics of the identified synapses. For example, in pyramidal-to-interneuron but not pyramidal-to-pyramidal connections, higher presynaptic firing rates result in frequency-dependent, incremental facilitation.

For long-distance connections, network properties have been more difficult to address. This is partially because of their spatial dispersion. It is difficult to trace a single axon along its complete trajectory and through multiple spatially separate arbors. Moreover, there is only a relatively small EM database, especially for human and subhuman primates, and this leaves many unanswered questions: Are there changes in caliber along the course of an axon? What are the postsynaptic targets? How do multiple inputs interact at a given postsynaptic target?

A. Thalamocortical Connections

Such data as are available pertain mainly to thalamocortical systems terminating in primary sensory areas since these can often be identified by physiological and/or structural criteria. For example, the two main groups of geniculocortical axons terminating in the primary visual cortex have been physiologically characterized, labeled by intraaxonal injection of tracers, and then morphologically analyzed by histological reconstructions. Such studies demonstrate that parvocellular axons have a small diameter, small arbors ($\sim$0.2 mm in diameter), and a moderate number of terminal specializations ($<$1500), consistent with their functional specialization for form and color vision. Magnocellular axons, which are associated with motion vision (requiring low spatial but high temporal resolution), have a larger diameter, larger arbors, and a large number of terminal specializations ($>$3000). Since each termination makes on average two synapses, if there is one synapse per postsynaptic neuron, then a reasonable estimate is that a single parvocellular axon diverges to about 3000 postsynaptic elements and a single magnocellular axon to at least 6000. The degree of convergence of parvocellular axons has been estimated to be at least 20–40. These numbers are important for further understanding of neuronal transformations, which are considerably more complicated than simple "relays."

EM studies suggest that both parvo- and magnocellular pathways have similar postsynaptic pools,

which suggests that other aspects of circuitry, such as recombinations with local and other extrinsic connections, contribute to the physiological differences associated with the cortical stages of these two pathways. Further analysis of network properties, at the resolution described previously for dual or triple intracellular recordings, will depend on new technical advances.

B. Feedforward and Feedback Cortical Connections

Beyond the primary sensory areas, axon properties are more difficult to characterize both physiologically and structurally. On the basis of structural evidence, interareal corticocortical connections have been subdivided into two broad categories: feedforward and feedback (Fig. 8). The terms reflect an assumption of serial organization, whereby feedforward connections progress from primary sensory through higher order areas. They are reciprocated in the "reverse" directions by feedback connections. Feedforward axons originate from neurons mainly in layer 3, terminate mainly in layer 4, and have one to four spatially separated arbors, each about 0.2 mm in diameter. Serial section reconstruction of axons labeled with the newer anterograde tracers demonstrates that individual arbors carry 50–400 terminal specializations, for a total number per axon of 400–1000 boutons. Feedback axons originate from neurons mainly in layer 6, terminate heavily in layer 1, and tend to have a single, rod-like axon at least 1.0 mm long, also with 400–1000 terminal specializations. On the basis of these data, if each bouton corresponds to one synapse, and if no more than one synapse is made with any postsynaptic neuron, a divergence factor of 1:400–1000 is suggested. No estimates are available for convergence in these pathways, and there are only sparse data available concerning postsynaptic populations. In primates, feedback inputs to layer 1 are commonly thought to target predominantly apical dendrites since these are a major constituent of layer 1, but the actual mix and identity of the postsynaptic targets are unknown. Similarly, the actual postsynaptic targets of feedforward connections, while probably including apical dendrites of deeper neurons that pass through layer 4 and the basal dendrites of overlying neurons in layer 3, are not known. Confocal microscopy may be an improved means of approaching these questions because it offers the necessary resolution of EM—to verify that boutons close to a structure are actually contacting that structure—without impractically slow EM processing and analysis.

In contrast to thalamocortical connections to primary sensory cortices, the functional properties of feedforward and feedback axons are not known. A convenient assumption is that feedforward axons are relays from early to higher order cortical areas and are

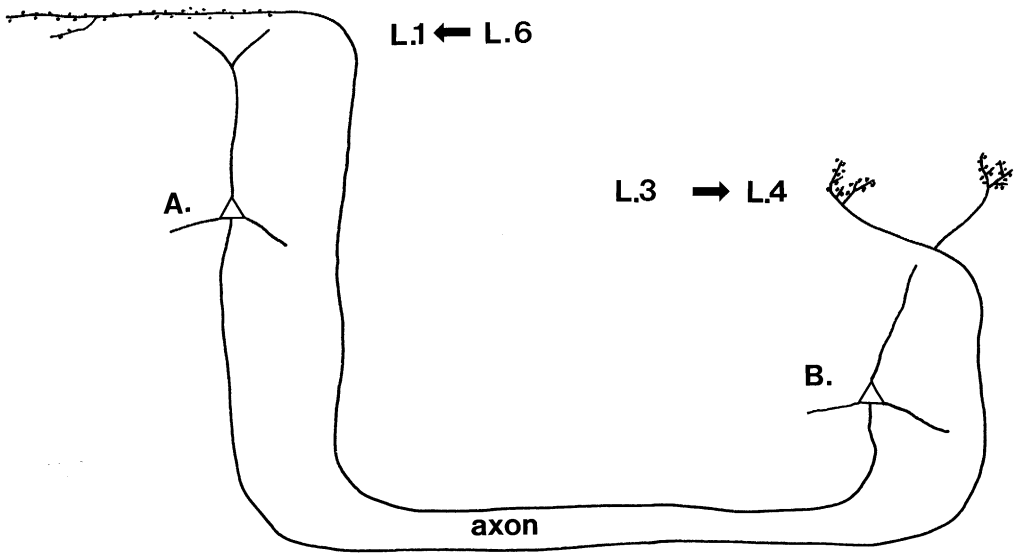


Figure 8 Schematic representation of two distinguishable types of cortical axons: feedforward (A) and feedback (B).

involved in progressive elaborations, whereas feedback axons exert a return, "modulatory" influence. This is clearly an oversimplification, however, in several respects. First, neither of these systems operate in isolation but, rather, are part of a broader network of thalamocortical, callosal, and intrinsic connections, among others. Second, the efficacy of feedback connections, despite their location on distal dendrites in layer 1, may actually be considerable since recent work has identified mechanisms for active boosting of synaptic inputs at dendritic locations distant from the axon hillock. Third, both feedforward and feedback axons, rather than operating as complementary pairs, likely comprise many functionally significant subdivisions.

C. Subtypes of Connections

1. Morphological Evidence

The current evidence for subdivisions within the main connectional groups derives mainly from the characteristics of the cells of origin. Neurons belonging to one projection class (e.g., corticoclaustral vs corticogeniculate projection neurons in layer 6 of primary visual cortex) frequently have a distinctive laminar pattern of their local axon collaterals. Other experiments suggest that projection systems can be distinguished by the distribution of neurofilament protein in the cell bodies. That is, some feedback projections in the visual pathway are reported to exhibit a higher proportion of neurofilament protein-containing neurons that the feedforward projections (70–80 vs 25–36% for projections to area V4). Additional criteria, such as specificity of surface molecules on the cell body, have been more difficult to establish, although it seems likely that this is due to technical limitations rather than any basic uniformity.

There are a few known examples of strikingly distinct structural specialization of axons. For example, corticothalamic projections to association thalamic nuclei (subdivisions of the pulvinar and the mediodorsal thalamus) originate from two groups of neurons: large pyramidal neurons in cortical layer 5 and smaller neurons in layers 5 and/or 6. This has been known since the late 1970s from the results of retrograde tracers injected into the thalamus and transported to cortical areas. Recently, studies using high-resolution anterograde tracers combined with serial section reconstruction have shown that the axons of these two groups differ dramatically (Fig. 9). Axons differ in caliber, being respectively greater than and less than 1.0 μm, and in the configurations of their terminal arbors. The larger axons terminate in small, round arbors with a small number of large terminal specializations. The smaller axons have elongated fields bearing a larger number of slender specializations. The physiological characteristics of these two types are not known, although it seems likely that the larger axons would be faster because of their larger diameter and larger terminations.

Another recognizable axon type is the connections originating from the giant Meynert cells in the deeper layers of primary visual cortex (Fig. 10). The axons of these cells are thick (2 or 3 μm in diameter) and have distinctively large terminal specializations. Both features are suggestive of a fast conduction velocity, although the functional properties of Meynert cells are not known. Some but not all of these axons can branch to multiple targets, including the pulvinar, cortical area MT/V5, and superior colliculus. The latter two structures are known to have a role in motion processing.

2. Physiological Evidence

Other evidence of axonal subtypes comes from physiological investigations of conduction velocity. Conduction velocity depends on several factors: axon caliber (larger axons are faster), myelination (myelinated axons are faster), the existence of branching, synaptic dynamics, and integration delay at the postsynaptic target. Axon length is less important. Conduction velocity is analyzed by measuring the latencies of spikes evoked by antidromic activation of axons by electrical stimulation or by calculating the temporal relationships between the firing times of two impaled neurons in cross-correlograms.

Long-distance projecting axons typically exhibit a spectrum of conduction velocities and caliber (Fig. 11). Projections from cortical motor areas to the spinal cord (corticospinal tract) are subdivided into slow and fast components. Slow fibers, comprising the majority, have an antidromic latency of about 2.6 msec from the brainstem pyramid, whereas for fast fibers the equivalent latency is 0.9 msec. These latencies respectively correspond to conduction velocities greater than or less than 30 $msec^{-1}$, and axon calibers <4 μm or >6 μm (in man). Curiously, in this system there does not seem to be any correlation of fiber size with either phylogenetic status or digital dexterity. From comparative studies, the largest fibers (25 μm) have been reported in the seal.

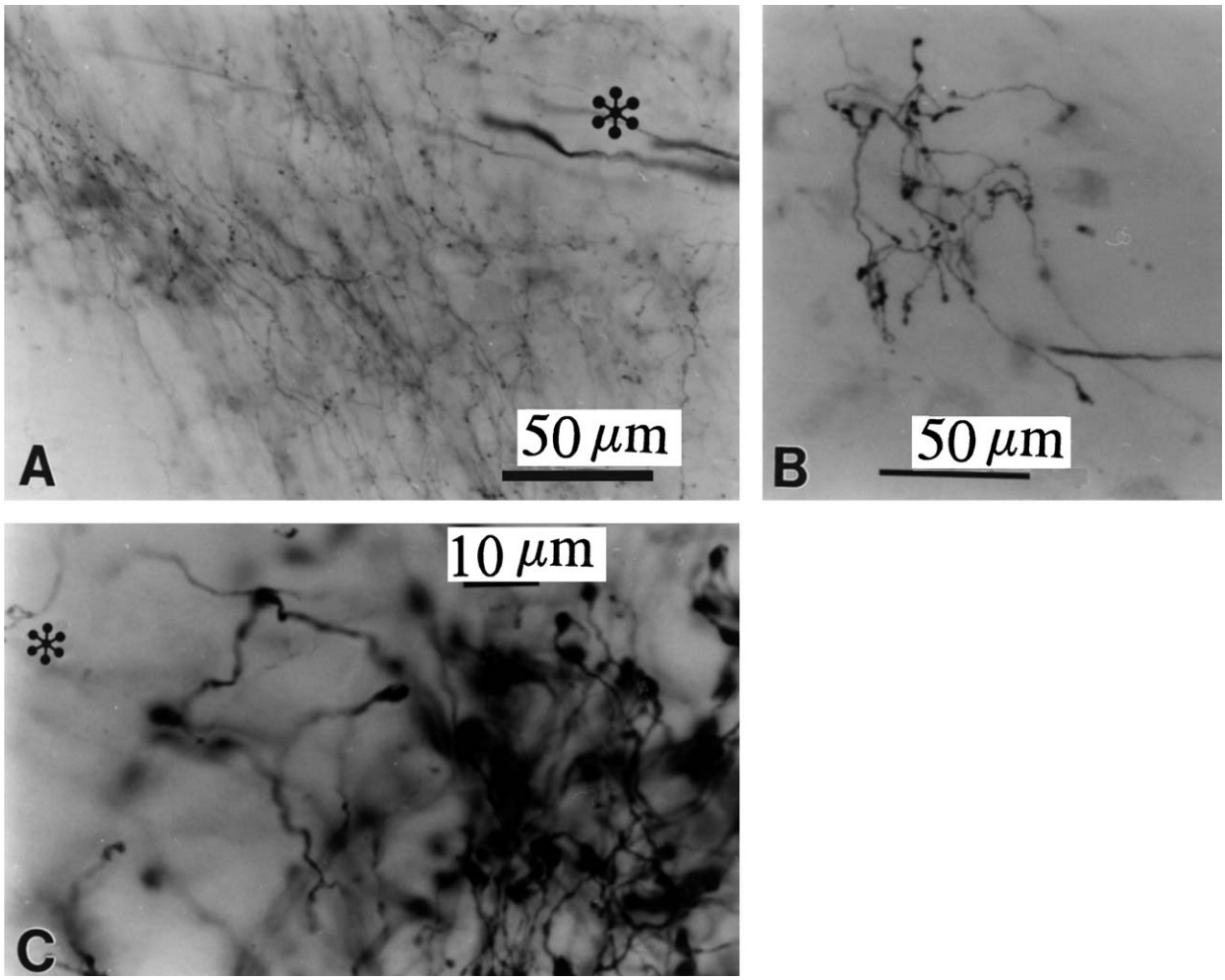


Figure 9 Photomicrographs of two morphologically distinct types of corticopulvinar axons. (A) A field of thin axons with slender, spine-like terminations. Portions of two thicker axons pass in the vicinity (asterisk). (B) A single, round arbor bearing a small number of larger boutons. (C) Higher magnification of a field of large boutons, mixed with a small group (asterisk) of the thinner terminations.

Several studies have investigated the distribution of latencies to visual stimulation in different areas of the visual system of macaque monkeys. These suggest an average latency difference of 10 msec between neurons in successive areas. All areas, however, exhibit a range in latencies, and it has been noted that short-latency neurons in higher order areas in inferotemporal cortex can respond to a given stimulus before long-latency neurons in area V1. The range in conduction velocity may correlate with axon caliber, which for most interareal cortical connections ranges from about 0.5 to 2.0 μm in diameter. How these timing differences contribute to cortical processing is the subject of ongoing investigations.

The temporal–computational characteristics of individual axons have been analyzed by simulating the propagation of an action potential in arbors traced with histological techniques. Simulated results on latencies and conduction velocities are in good agreement with those found electrophysiologically, and this technique may provide an effective tool for probing issues such as how the velocity of spike propagation is affected by changes in axon diameter at branching points or at boutons, the delay in activation latencies between boutons, and how conduction velocity might be modified by the previous occurrence of an action potential. An interesting general result from this work is that different axon geometries can lead to similar patterns of spatiotemporal activation. This raises the issue of how close a correspondence there is between the morphology and computational properties of an axon. A larger sample of different axons and more

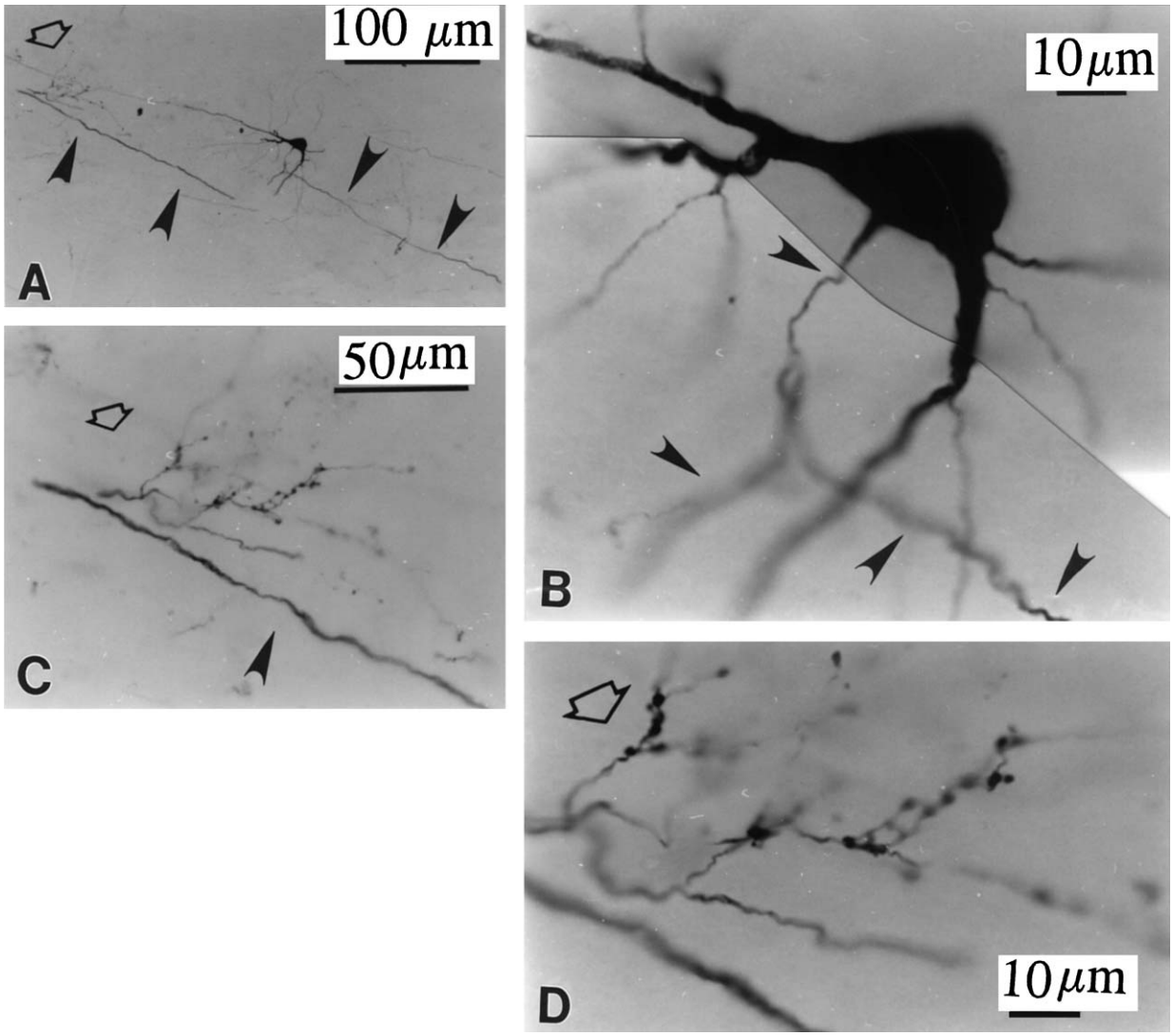


Figure 10 Photomicrographs of axons originating from large Meynert cells of primary visual cortex. These are large caliber and terminate in arborizations with large boutons. [A and (higher magnification) B] Cell body, dendrites, and axon (arrowheads) filled with BDA. [C and (higher magnification) D] One of several terminal arbors. Open arrows point to corresponding features in A, C, and D. (Portions in the deeper planes of the histological section are out of focus.)

precise data on the biophysical properties of the terminal arbor are necessary to extend this line of research.

VI. HOW AXONS RELATE TO CORTICAL LAYERS AND COLUMNS

Two key elements of cortical architecture are layers and modules (''columns,'' ''assemblages,'' or cell groups). Both layers and columns are complicated structures. Layers can be distinguished on the basis of cell sizes and density, and in many cases enzyme concentrations and other markers also show orderly laminar patterns.

As described previously, feedforward systems terminate heavily at the level of layer 4, and feedback systems preferentially target layer 1. Intrinsic collaterals of pyramidal neurons tend to arborize mostly in layers 3 and 5. It is not clear, however, whether laminar specificity implies, as it does in the hippocampal formation, dendritic stratification of inputs. Available evidence suggests, to the contrary, that extrinsic axons

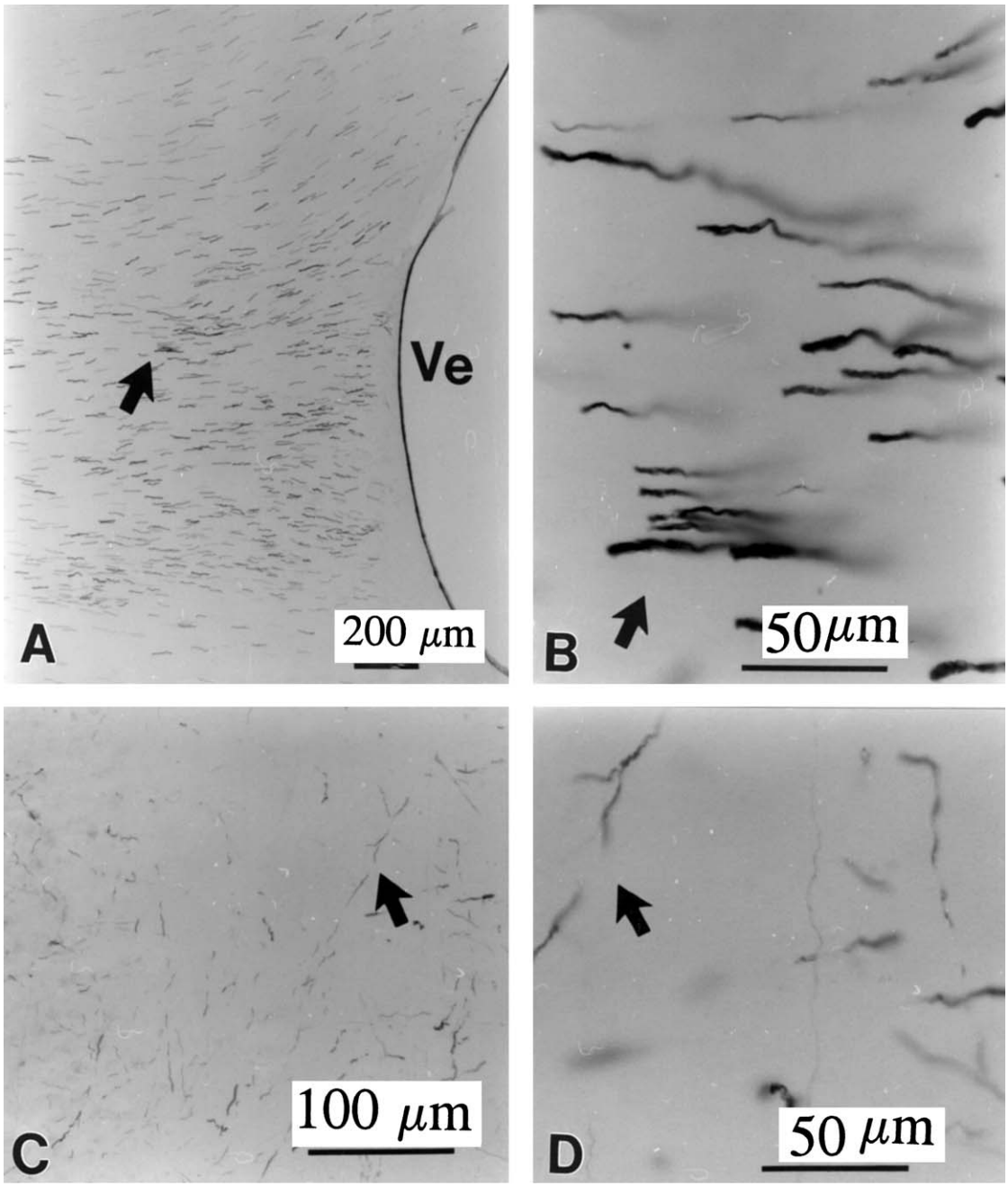


Figure 11 Photomicrographs of axons labeled by BDA anterogradely transported from an injection site. [A and (higher magnification from arrow) B] A field of axons crossing in the corpus callosum. The histological sectioning results in axon fragments that can be identified and traced in sequential sections. Note the range of different fiber diameters. [C and (higher magnification from arrow) D] Field of corticocortical axons subjacent to their target region. Branched profile is evident at the arrows.

synapse on many of the structures within their terminal domain.

The relationship between axon systems and cortical columns is likewise complex. For example, in visual cortex geniculocortical axons derived from the left or right eye segregate into alternating ocular dominance columns. These constitute the primary anatomical substrate for the functional columns that can be demonstrated by microelectrode recordings or by activity-driven imaging (with the metabolic tracer

2-deoxyglucose in animals or position emission tomography or functional magnetic resonance imaging in humans). It is important to remember, however, that the functional columns, that are visualized as extending from pia to white matter through the cortical depth, are actually a combination of direct thalamocortical connections to layer 4 and local, short interlaminar relays from layer 4 to the upper and lower layers.

In nonprimary cortices, the anatomical substrates of columnar architecture are still under investigation. Apparently different from the primary areas, in the association areas many connectional systems directly terminate in an extended column involving several layers. These projection columns, which can range from 0.2 to 0.5 mm in diameter, result from the convergence of hundreds (thousands?) of axons. The individual arbors vary in size and laminar distribution but are usually smaller in diameter than the parent column.

Thus, although cortical architecture is commonly compared to a brickwork of repetitive modules (''ice-cube'' model of the cortex), surprisingly few data are available about columnar microorganization, and the actual structure may be considerably more elaborate.

VII. AXONS AND CONSCIOUSNESS

Several contemporary theories of consciousness give a prominent place to axons. At a large-scale global level, the theory of neuronal group selection proposes that selectionist (as opposed to instructionist) mechanisms govern the formation, adaptation, and interactions of local collections of neuronal groups. Group activity is cojoined or correlated in what is viewed as a dynamic process brought about by reentrant connections. Reentry implies reciprocal feedforward–feedback networks, but the concept is also compatible with the axonal networks in aggregate. A key component is the sheer density of the connectional plexus, where vast numbers of combinations can be effectuated by the massively parallel organization in space and time.

See Also the Following Articles

CRANIAL NERVES • NERVOUS SYSTEM, ORGANIZATION OF • NEURONS • PERIPHERAL NERVOUS SYSTEM • SYNAPSES AND SYNAPTIC TRANSMISSION AND INTEGRATION

Suggested Reading

Almenar-Queralt, A., and Goldstein, L. S. B. (2001). Linkers, packages and pathways: New concepts in axonal transport. *Curr. Opin. Neurobiol.* **11,** 550–557.

Burkhalter, A., and Bernardo, K. L. (1989). Organization of corticocortical connections in human visual cortex. *Proc. Natl. Acad. Sci. USA* **86,** 1071–1075.

Castellani, V., Yue, Y., Gao, P. P., Zhou, R., and Bolz, J. (1998). Dual action of a ligand for Eph receptor tyrosine kinases on specific populations of axons during the development of cortical circuits. *J. Neurosci.* **18,** 4663–4672.

Freund, T. F., Martin, K. A. C., Soltesz, I., Somogyi, P., and Whitteridge, D. (1989). Arborization pattern and postsynaptic targets of physiologically identified thalamocortical afferents in striate cortex of the macaque monkey. *J. Comp. Neurol.* **289,** 315–336.

Gennarelli, T. A., Thibault, L. E., and Graham, D. I. (1998). Diffuse axonal injury: An important form of traumatic brain damage. *The Neuroscientist* **4,** 202–215.

Hirokawa, N. (1997). The mechanisms of fast and slow transport in neurons: identification and characterization of the new kinesin superfamily motors. *Curr. Opin. Neurobiol.* **7,** 605–614.

Innocenti, G. M., Lehmann, P., and Houzel, J.-C. (1994). Computational structure of visual callosal axons. *Eur. J. Neurosci.* **6,** 918–935.

Makris, N., Worth, A. J., Sorensen, A. G., Papadimitriou, B. S., Wu, O., Reese, T. G., Wedeen, V. J., Davis, T. L., Stakes, J. W., Caviness, V. S., Kaplan, E., Rosen, B. R., Pandya, D. N., and Kennedy, D. N. (1997). Morphometry of *in vivo* human white matter association pathways with diffusion-weighted magnetic resonance imaging. *Ann. Neurol.* **42,** 951–962.

Nowak, L. G., and Bullier, J. (1997). The timing of information transfer in the visual system. In *Cerebral Cortex: Extrastriate Cortex in Primates* (K. S. Rockland, J. H. Kaas, and A. Peters, Eds.), Vol. 12, pp. 205–241. Plenum, New York.

Penrose, R. (1994). *Shadows of the Mind (see Ch. 7)*. Oxford Univ. Press, New York.

Peters, A., Palay, S. L., and de Webster, H. (1991). *The Fine Structure of the Nervous System.* Oxford Univ. Press, New York.

Porter, R., and Lemon, R. (1993). *Corticospinal Function and Voluntary Movement.* Clarendon, Oxford.

Rockland, K. S. (1997). Elements of cortical architecture: Hierarchy revisited. In *Cerebral Cortex: Extrastriate Cortex in Primates* (K. S. Rockland, J. H. Kaas, and A. Peters, Eds.), Vol. 12, pp. 243–293. Plenum, New York.

Shepherd, G. M. (Ed.) (1998). *The Synaptic Organization of the Brain*. Oxford Univ. Press, New York.

Van der Loos, H. (1967). The history of the neuron. In *The Neuron* (H. Hydén, Ed.), pp. 1–47. Elsevier, New York.

Waxman, S. G., Kocsis, J. D., and Stys, P. K. (Eds.) (1995). *The Axon. Structure, Function, and Pathophysiology.* Oxford Univ. Press, New York.

Basal Ganglia

BRUCE CROSSON,*,† LEEZA MARON,* ANNA B. MOORE,*,† and LAURA GRANDE*

University of Florida Health Science Center and †VA Medical Center, Gainesville

GLOSSARY

basal ganglia A group of phylogenetically old structures, including the caudate nucleus, putamen, globus pallidus, archistriatum, and ventral pallidum, that are thought to play a role not only in motor functions but also in more complex cognitive operations, such as language, learning, and working memory.

caudate nucleus The nucleus of the basal ganglia that follows the inner curvature of the lateral ventricle from the head of the caudate nucleus (embedded in the frontal horn of the lateral ventricle deep to the frontal lobe) to the body of the caudate nucleus (inferior and adjacent to the body of the lateral ventricle deep to the parietal lobe) and to the tail of the caudate nucleus (adjacent to the temporal horn of the lateral ventricle deep within the temporal lobe). In terms of its histological features and connectivity, it is homologous to the putamen; collectively, the two structures are referred to as the neostriatum.

cortico-striato-pallido-thalamo-cortico loops In 1986, Alexander and colleagues discovered five parallel and segregated loops through the basal ganglia; motor, occulomotor, anterior cingulate, dorsolateral frontal, and orbitofrontal loops. These loops involve glutamatergic input from the cortex to the striatum and GABAergic projections from both the striatum to the globus pallidus and from the globus pallidus to the thalamus. Reciprocal projections between the thalamus and the cortex are glutamatergic and, therefore, excitatory in nature.

executive functions Complex cognitive functions that involve the use or modification of more basic information or the control of other cognitive processes. Examples of executive functions are reasoning, organization and planning, and selection of actions appropriate to an external context.

globus pallidus A cone-shaped structure situated medial to the putamen that consists of an external (lateral) and internal (medial) segment. The globus pallidus is separated from the thalamus by the posterior limb of the internal capsule and from the head of the caudate nucleus by the anterior limb of the internal capsule. Collectively, the globus pallidus and the putamen are referred to as the lentiform nucleus, but the internal structure of the globus pallidus is distinct from that of the putamen. From a standpoint of connectivity, the medial segment of the globus pallidus and the substantia nigra pars reticulata are homologous structures.

Huntington's disease An autosomal-dominant genetic disorder characterized by choreiform movements and later dementia. In the early stages of the disease cell loss can be observed primarily in the head of the caudate nucleus; however, as the disease progresses, cell loss involves other parts of the basal ganglia and cortex, particularly the frontal lobe.

nucleus accumbens The nucleus inferior to the head of the caudate nucleus and putamen. The nucleus accumbens consists of core and shell divisions. The core division is considered part of the archistriatum (or old striatum). Its pattern of connections is similar to that of the neostriatum, except that its input is primarily from limbic cortex instead of the neocortex and its projections are primarily to the ventral pallidum instead of the globus pallidus.

Parkinson's disease A syndrome caused by loss of midbrain dopaminergic neurons in the substantia nigra pars compacta and ventral tegmental area. Parkinson's disease produces disturbances of movement such as tremor, rigidity in movement, difficulty initiating movement, and disturbance of gait.

putamen A basal ganglia structure situated lateral to the globus pallidus and deep to the insula, separated from the insula by the external capsule, the claustrum, and the extreme capsule. Although it is histologically dissimilar from the globus pallidus, the two are often grouped together due to their spatial proximity and are labeled as the lentiform nucleus. Because of their similar internal structures and patterns of connectivity, the putamen and the caudate nucleus are collectively referred to as the neostriatum.

substantia nigra The midbrain nucleus lying dorsal to the cerebral peduncles. Pars compacta is the dorsal portion of this nucleus and contains dopaminergic neurons that project to the caudate nucleus

and putamen. Pars reticulata is the ventral portion of this nucleus and is considered a homologous structure to the medial globus pallidus because of its similar connections.

thalamus An oblong group of nuclei bordering on the third ventricle medially and separated laterally from the lentiform nucleus by the posterior limb of the internal capsule. Thalamic nuclei have extensive, mostly reciprocal connections, with specific cortical regions. Some thalamic nuclei act as relays between lower centers and visual, auditory, and somatosensory information. Other thalamic nuclei receive projections from the globus pallidus and substantia nigra pars reticulata, acting as a part of cortico-striato-pallido-thalamo-cortical loops.

Deep within the cerebral hemispheres lie phylogenetically old structures collectively known as the basal ganglia. These structures are prominent in the reptile brain. When compromised in the human brain, devastating effects on behavior and cognition can result. In some instances, disruption of the basal ganglia produces disturbances of movement characteristic of Parkinson's disease. Symptoms of Parkinson's disease include tremor, rigidity in movement, difficulty initiating movement, and disturbance of gait. In severe cases, patients suffering from this disorder can be almost frozen, barely able to initiate any movement. They require complete care. In the autosomal-dominant disorder of Huntington's disease, cell loss occurs in the basal ganglia (specifically in the neostriatum), producing involuntary movements that resemble portions of volitional movements. The disorder progresses to the point where the patient is totally incapacitated. Both Parkinson's and Huntington's disease patients commonly have changes in mental status. Sometimes in Huntington's disease, cognitive symptoms can precede the movement disorder, and patients may even be psychotic. Given these phenomena, the reader might be surprised to learn the function of the basal ganglia is a matter of debate.

I. MACROSTRUCTURE AND CONNECTIONS OF THE BASAL GANGLIA

Classically, the basal ganglia consist of the globus pallidus, the putamen, and the caudate nucleus. Because of their wedge-shaped appearance in coronal (Fig. 1) and axial (Fig. 2) sections, the putamen and the globus pallidus are sometimes collectively referred to as the lentiform nucleus. The putamen is the more lateral structure of the two, and the globus pallidus is the medial, tapered end of the wedge. The globus pallidus can be divided into a lateral (or external) and a medial (or internal) segment. Despite the fact that they share the same name, the medial and lateral globus pallidus differ in terms of connectivity and function. The medial globus pallidus is in fact frequently considered a homologous structure to the substantia nigra pars reticulata. Even though the putamen and globus pallidus are adjacent and can be referred to with a collective label, they are neither structurally nor functionally homologous. The internal structures of the caudate nucleus and the putamen are quite similar. Although their connectivity is not overlapping, it can be considered to be parallel; this parallelism of connections will become evident when we discuss the connectivity of the basal ganglia. Collectively, the caudate nucleus and the putamen are referred to as the neostriatum or, often, as just the striatum.

When viewed in sagittal section (Fig. 3), the caudate nucleus forms a structure that arches over the lentiform nucleus and the internal capsule. The head of the caudate nucleus is comparatively bulky and is embedded in the frontal horn of the lateral ventricle, lying posterolaterally to this portion of the ventricular system. The caudate nucleus follows the inner curvature of the lateral ventricle posteriorly, from the head (deep to the frontal lobe) to the body (deep to the parietal lobe) to the tail (deep within the temporal lobe), with the tail ending adjacent to the amygdala. The caudate nucleus tapers considerably as it moves posteriorly from the head through the body to the tail, until the tail becomes difficult to discern.

The internal capsule is a band of white matter that molds itself around the lentiform nucleus and the lateral portions of the caudate nucleus. The anterior limb of the internal capsule separates the caudate nucleus from the putamen and globus pallidus, although at points thin bands of gray matter traverse the internal capsule to connect the caudate nucleus and putamen. The posterior limb of the internal capsule separates the putamen and globus pallidus from the thalamus. The posterior and anterior limbs meet at the genu (or knee) of the internal capsule. The internal capsule contains fibers from the frontal cortex to the brain stem and spinal cord (e.g., corticopontine and corticospinal tracts). The internal capsule also contains reciprocal fibers between the cortex and the thalamus (i.e., corticothalamic and thalamocortical fibers). Pallidothalamic fibers compose a portion of the inferior thalamic peduncle and, therefore, traverse an inferior portion of the internal capsule.

Phylogenetically older structures bearing some similarity to the neostriatum, especially from the standpoint of connectivity, are referred to as the

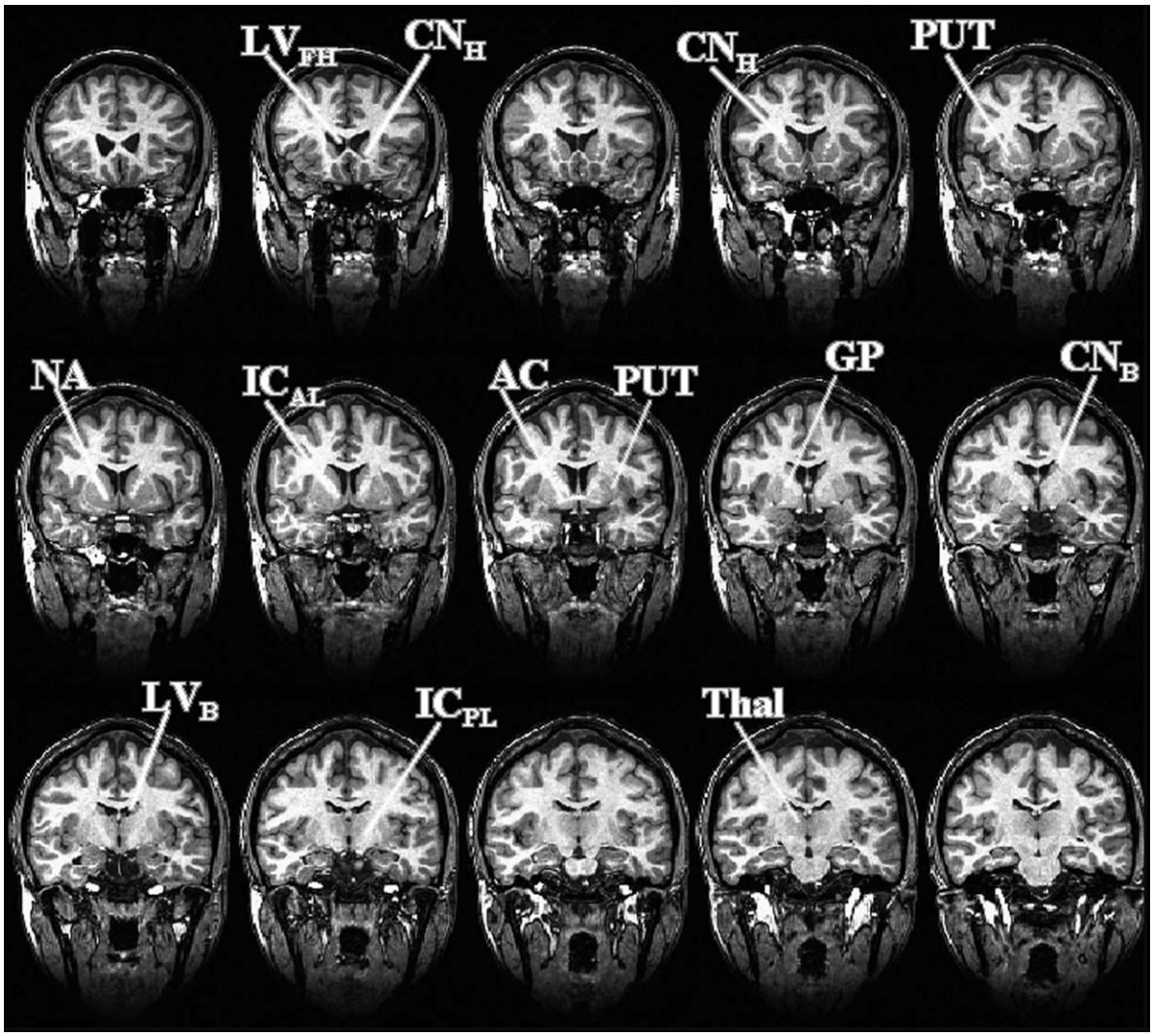


Figure 1 T1-weighted coronal MRI images showing the basal ganglia. Basal ganglia structures can be traced from the more anterior sections (upper left) to the more posterior sections (lower right). The slices are 1 mm thick, with one slice every 3 mm. The frontal horn of the lateral ventricle (LV_{FH}) is already apparent in the first slice of the top row and can be traced through consecutive slices until it becomes the body of the lateral ventricle (LV_B) around the fifth slice of the second row. The head of the caudate nucleus (CN_H) is embedded in the frontal horn of the lateral ventricle; it appears in the second slice of the top row and can be traced through consecutive slices until it tapers into the body of the caudate nucleus (CN_B) around the fourth and fifth slices of the second row. The putamen (PUT) can be seen as early as the third slice of the first row (lateral to CN_H and separated from it by the internal capsule), and remnants of it are visible in the fifth slice of the third row. The nucleus accumbens (NA) is best seen in the first slice of the second row; it occupies a position ventral to CN_H and PUT and seems to join the two. The anterior limb of the internal capsule (IC_{AL}) is the band of white matter between CN_H and PUT; it is labeled in the second slice of the second row but runs from the third slice of the first row through the third slice of the second row. The globus pallidus (GP) first appears in the third slice of the second row and can be seen until the third slice of the third row. The thalamus (Thal) can first be seen clearly on the first slice of the third row. The thalamus is separated from PUT and GP by the posterior limb of the internal capsule (IC_{PL}). The anterior commissure (AC), which connects the left and right temporal lobes, can be seen in the third slice of the second row.

archistriatum, or sometimes as the limbic striatum, because of their connection to the limbic system. The structures that comprise the archistriatum include the olfactory tubercle and the core portion of the nucleus accumbens (Fig. 1). Likewise, a portion of the basal forebrain region inferior to the anterior commissure is designated as the ventral pallidum and has connectivity similar to that of the globus pallidus except with

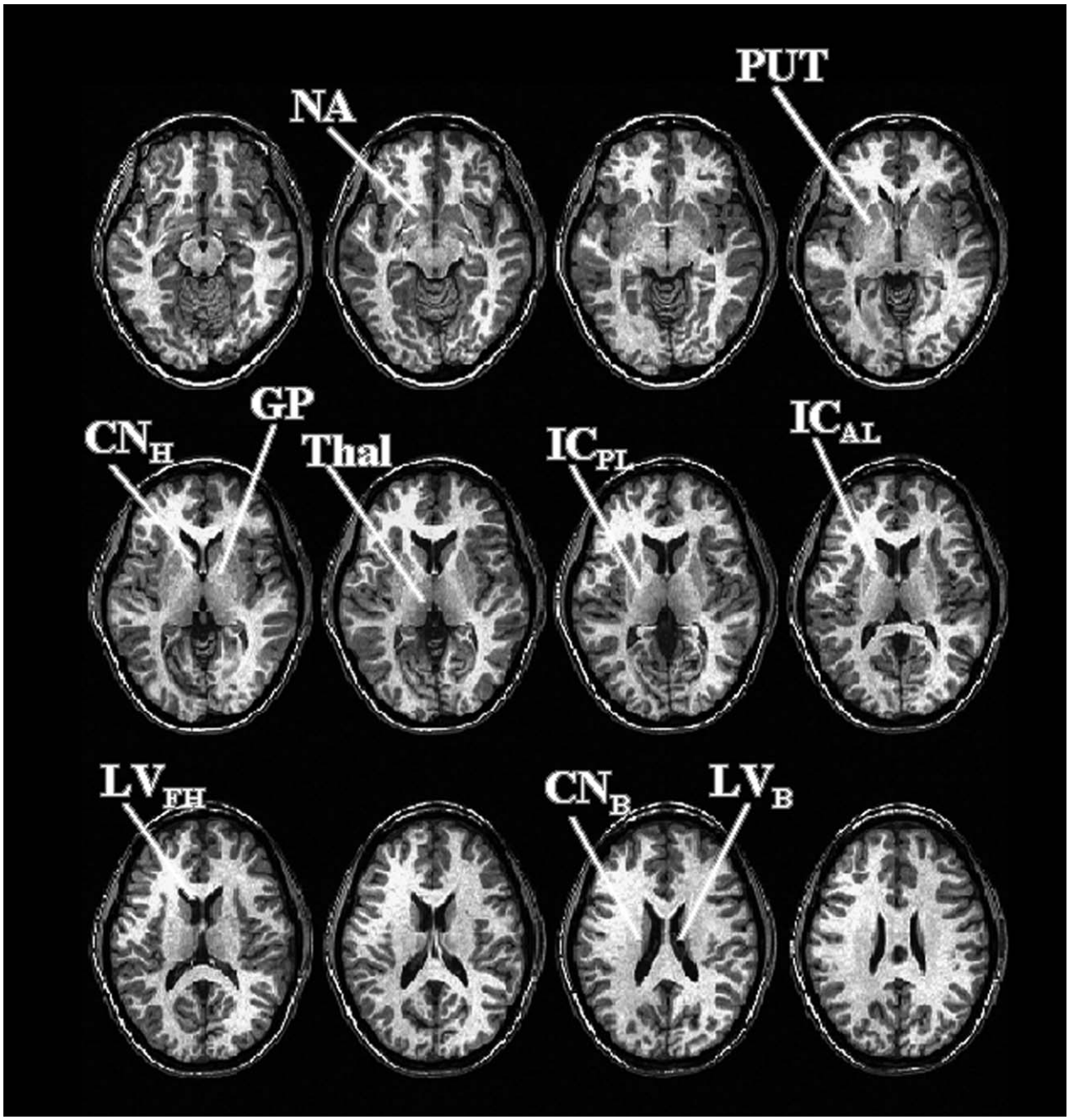


Figure 2 T1-weighted axial MRI images showing the basal ganglia. The basal ganglia can be traced from the more inferior sections (upper left) to the more superior sections (lower right). The slices are 1 mm thick, with one slice every 3 mm. The nucleus accumbens (NA) is best seen on the second slice of the first row; it occupies a position ventral to the head of the caudate nucleus (CN_H) and the putamen (PUT). The PUT, CN_H, frontal horn of the lateral ventricle (LV_{FH}), and globus pallidus (GP) all can be seen beginning on the fourth slice of the first row. The thalamus (Thal) can first be seen clearly on the second slice of the second row and is visible for four or five slices. As the caudate nucleus and lateral ventricle arch over the thalamus, they become the body of the caudate nucleus (CN_B) and the body of the lateral ventricle (LV_B), respectively. Again, the anterior limb of the internal capsule (IC_{AL}) separates CN_H and PUT, and the posterior limb of the internal capsule separates Thal from PUT and GP.

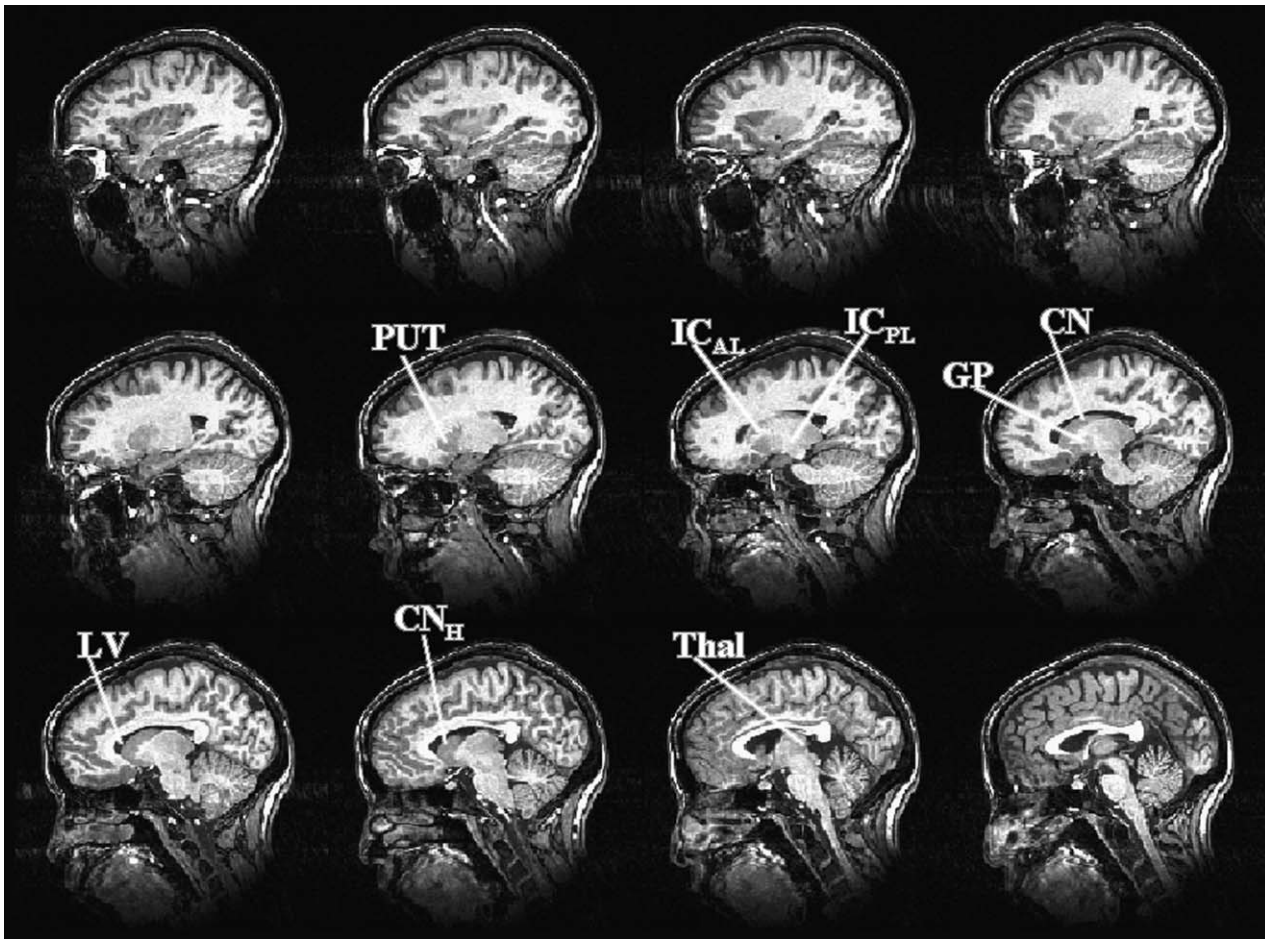


Figure 3 T1-weighted sagittal MRI images showing the basal ganglia. The basal ganglia can be traced from the more lateral sections (upper left) to the more medial sections (lower right). The slices are 1 mm thick, with one slice every 3 mm. The putamen (PUT) is first clearly visible on the first slice of the second row. The caudate nucleus (CN) is first seen on the third slice of the second row, as it follows the lateral ventricle (LV). The globus pallidus (GP) can be seen in the third and fourth slices of the second row. The thalamus (Thal) can seen in the second slice of the second row and runs through the most medial slice of the series. Again, the anterior limb of the internal capsule (IC_{AL}) separates CN_H and PUT, and the posterior limb of the internal capsule (IC_{PL}) separates Thal from PUT and GP.

limbic structures. Although the amygdala and the claustrum are sometimes referred to as part of the basal ganglia, we will not follow this convention here.

The organization of the basal ganglia into circuits with the cortex was described in the late 1980s by Alexander, Delong, and Strick. From the standpoint of connectivity, the basal ganglia are closely related to specific thalamic nuclei and cortical structures. As noted previously, the connectivity of the basal ganglia to these structures follows a predictable pattern that generally involves projections from the cortex to the striatum, from the striatum to the pallidum, from the pallidum to the thalamus, and from the thalamus back to the cortex (Fig. 4). Different portions of the frontal lobe project to different segments of the striatum. Grossly, anterior cingulate cortex projects to the archistriatum, premotor and motor cortex project more heavily to the putamen, and lateral prefrontal cortex projects to the caudate nucleus. The striatum projects to the lateral and medial segments of the globus pallidus. Various segments of the medial globus pallidus project to locations within the ventral lateral, ventral anterior, and dorsomedial thalamic nuclei. These loops involving projections from the cortex to the striatum, from the striatum to the globus pallidus, and from the globus pallidus to the thalamus are

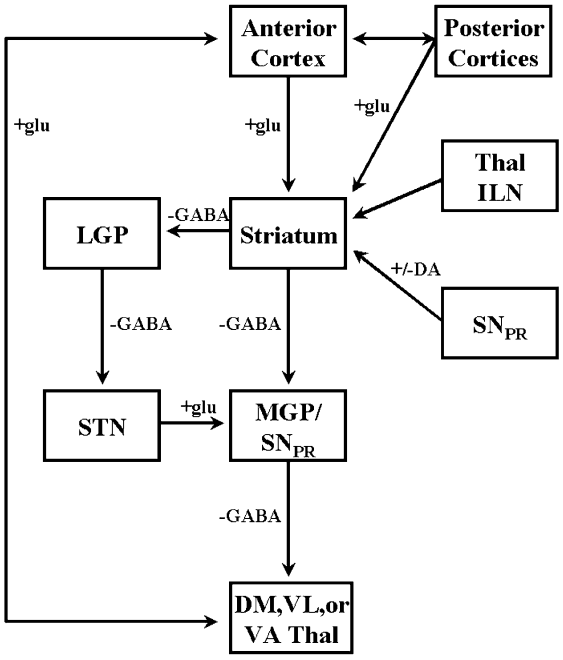


Figure 4 Diagram showing the relationship between components of basal ganglia loops. As conceptualized by Alexander *et al.*, both anterior and posterior cortical areas contributed input to the loops, as shown. More recent conceptualizations of these loops (e.g., Middleton and Strick, 2000) are that they are primarily closed, i.e., they involve a single cortical input that is also the target of the loop. DM, dorsal medial nucleus of the thalamus; ILN, intralaminar nuclei; LGP, lateral globus pallidus; MGP, medial globus pallidus; SN_{PC}, substantia nigra pars compacta; SN_{PR}, substantia nigra pars reticulata; STN, subthalamic nucleus; Thal, thalamus; VA, ventral anterior nucleus of thalamus; VL, ventral lateral nucleus of thalamus; +glu, the excitatory neurotransmitter glutamate; −GABA, the inhibitory neurotransmitter gamma amino butyric acid, +/−DA, dopamine that can have an inhibitory or excitatory impact depending on which output neurons it affects.

considered closed due to the thalamic projections to frontal cortex. It is worth noting that these thalamocortical projections are reciprocated by direct corticothalamic connections. These cortico-striato-pallido-thalamo-cortico loops appear to be separated from one another at all levels. Although the loops were eloquently examined by Alexander *et al.* it is worth noting that the influence of the motor loop on motor disorders was explored at least as far back as 1942 with Bucy's paper. The seminal work of Alexander and colleagues described five cortico-striato-pallido-thalamo-cortical loops: a motor loop, an occulomotor loop, an anterior cingulate loop, a dorsolateral frontal loop, and an orbitofrontal loop. They also predicted that other parallel basal ganglia loops would be discovered. Indeed, Strick and colleagues described supplementary motor area, ventral premotor, and primary motor cortex loops within the motor domain.

Several other aspects of basal ganglia connectivity should be mentioned. First are the midbrain dopaminergic projections to the striatum. Neurons in pars compacta of the substantia nigra project to the caudate nucleus and the putamen. It is the loss of the dopaminergic neurons projecting to the putamen that is responsible for the motor symptoms of Parkinson's disease. Projections from the ventral tegmental area provide dopaminergic input to the archistriatum. Second, the lateral globus pallidus projects to the subthalamic nucleus that lies below the thalamus, separated from it by an inferior portion of the internal capsule. The subthalamic nucleus, in turn, projects to the medial globus pallidus, creating a kind of subloop in the cortico-striato-pallido-thalamo-cortical loops. Third, the intralaminar nuclei of the thalamus project heavily to the striatum, constituting another source of input to this structure.

As noted previously, cortico-striato-pallido-thalamo-cortical loops maintain separation at all levels, including the level of the thalamus. The motor loops include portions of the ventrolateral thalamus. The anterior cingulate loop passes through a portion of the dorsomedial thalamus; the oculomotor, dorsolateral prefrontal, and lateral orbitofrontal loops pass through different portions of the ventral anterior and dorsomedial thalamic nuclei.

To summarize, the basal ganglia are structures deep within the cerebral hemispheres that collectively consist of the putamen, the caudate nucleus, the archistriatum, the globus pallidus, and the ventral pallidum. Organization of the basal ganglia is dominated by multiple parallel but segregated loops that start in the cortex and project to the striatum. The striatum, in turn, projects to the medial globus pallidus and the substantia nigra pars reticulata. The latter two structures project to thalamic nuclei that project back to the cerebral cortex. Recent work suggests that the loops can best be considered as closed because the cortical target of the loops is the same as their cortical input. Multiple frontal loops exist, but temporal and parietal loops recently have been discovered. In a subloop, or indirect pathway, fibers from the lateral globus pallidus project to the subthalamic nucleus, which in turn sends fibers to the medial globus pallidus. This anatomic connectivity helps to determine the role of the basal ganglia in behavior and cognition. The nature of the neurotransmitters between these structures provides another clue to how the loops function.

II. MICROSTRUCTURE AND NEUROTRANSMITTERS OF THE BASAL GANGLIA

Several aspects of the microstructure of the basal ganglia as described by Wilson, Groves, and others are worth specific mention. The vast majority of striatal neurons (90–95%) are medium-sized spiny neurons characterized by the proliferation of spines on the dendrites. These spiny neurons are the output neurons of the striatum. They use gamma aminobutyric acid (GABA) as their primary neurotransmitter and have an inhibitory influence on targets in the medial and lateral globus pallidus and in the substantia nigra pars reticulata, although these neurons also contain neuropeptides such as enkephalin, substance P, and dynorphin. The axons of these medium spiny neurons also give off short collaterals within the striatum that appear to terminate on neighboring neurons forming a network of collateral inhibition. In the past, it was assumed that this network of collateral inhibition accounted for the relatively low basal firing rates of the medium spiny neurons. However, Wilson recently reviewed evidence suggesting that it is not collateral inhibition that keeps the basal firing rates low. Rather, it takes coherent excitatory input distributed over a large portion of the dendritic tree of the spiny neuron to overcome the effects of potassium currents that act to shunt less coherent inputs.

Corticostriatal inputs use glutamate as the neurotransmitter and have an excitatory influence on the vast majority of the target neurons in the striatum. Perhaps half of the input to medium spiny neurons of the striatum derives from corticostriatal projections. Another major input to the striatum derives from the thalamic intralaminar nuclei. Besides medium-sized spiny neurons, neurons in the striatum include large spiny neurons and cholinergic aspiny neurons. The classification of these neurons has been refined since early reports.

The pioneering work of Gerfen, Graybiel, and others in the 1980s indicated that compartmentalization is an important aspect of striatal organization. Early work on striatal compartmentalization indicated that the internal structure of the striatum could be divided into striosomes (or patch compartments) and a matrix compartment. An early discovery about these compartments was that they could be distinguished by the presence of markers for acetylcholine in the matrix compartment and a relative absence of these markers in the striosomes.

Connections of these compartments were also found to vary, with the striosomes projecting to pars compacta of the substantia nigra and the matrix compartment projecting to both segments of the globus pallidus and pars reticulata of the substantia nigra. Since this seminal work, other complexities have been noted. It has been discovered that limbic cortex projects more heavily to striosomes, whereas neocortex projects more heavily to the matrix compartment. Gerfen and others have related this feature to the dorsal–ventral organization of corticostriatal projections; that is, limbic connections predominate in the ventral portions of the striatum (nucleus accumbens), whereas neocortical projections predominate in dorsal portions of the striatum (caudate nucleus). Other characteristics of striosomal compartmentalization include increased μ-opiate receptor binding and increased staining for the neuropeptides enkephalin and substance P. However, inhomogeneities in enkephalin and substance P distribution within the striosomal compartments suggests additional levels of organization within the neostriatum.

In the 1980s, Yelnik, Percheron, and Francois described important aspects of the microstructure of the globus pallidus. Pallidal neurons have sparsely branched dendritic arborizations distributed in disc-shaped fields that are parallel to the lateral border of their respective pallidal segment. Thus, these dendritic fields appear to be oriented in a direction perpendicular to incoming striatopallidal axonal fibers. It can be speculated that this organization leaves pallidal neurons in a position to provide spatial and temporal summation of inhibitory inputs from different striatal inputs. Neurons in the globus pallidus and those within pars reticulata of the substantia nigra that are a part of the basal ganglia loops are primarily GABAergic and, as noted previously, project to specific thalamic nuclei.

Other aspects of neurotransmitter function in the basal ganglia should be briefly mentioned. First, in the subloop from the lateral globus pallidus through the subthalamic nucleus and back to the medial globus pallidus, the pallidal efferents are GABAergic, as are other pallidal projection neurons. The neurotransmitter from the subthalamic nucleus to the medial globus pallidus is excitatory and thought to be glutamate. Second, the thalamocortical neurotransmitter appears to be excitatory and is also thought to be glutamate. Finally, dopaminergic input to the striatum from pars compacta of the substantia nigra seems to have a different impact on striatal neurons projecting to the lateral globus pallidus, which in turn projects to the subthalamic nucleus, than it does on striatal neurons projecting

directly to the medial globus pallidus and pars reticulata of the substantia nigra. Gerfen and others have noted that in striatal neurons projecting to the lateral globus pallidus, D_2 receptors predominate, and dopaminergic activity appears to have an inhibitory impact on these striatial neurons. In those striatal neurons projecting to the medial globus pallidus, D_1 receptors predominate, and dopaminergic activity appears to have an excitatory impact on these striatal neurons.

In summary, the striatum can be divided into striosomal and matrix compartments. Striosomes are dominated by projections from limbic cortex, whereas the matrix compartment is dominated by connections from neocortex. This distinction may be related to the ventral–dorsal division of the striatum. More ventral portions of the striatum, the archistriatum (olfactory tubercle and nucleus accumbens), receive input primarily from limbic cortex, whereas more dorsal portions of the striatum, the caudate nucleus and putamen, receive connections primarily from neocortex, with limbic input primarily reserved for striosomes. Striosomal projections are primarily to the dopamine-rich portion of the substantia nigra, pars compacta, whereas matrix projections go to the lateral and medial globus pallidus and pars reticulata of the substantia nigra. Corticostriatal fibers use glutamate, which has an excitatory influence on striatal projection neurons. Striatal projections are dominated by GABA, an inhibitory neurotransmitter. Striatal neurons project to the medial and lateral globus pallidus and pars reticulata of the substantia nigra; projection neurons in the latter structures appear to integrate input from different striatal areas and are also GABAergic. Fibers from the medial globus pallidus and substantia nigra pars reticulata project to various thalamic nuclei, which then send excitatory (glutamatergic) fibers to the cortical component originally projecting into the loop. An indirect pathway sends inhibitory GABAergic fibers from the lateral globus pallidus to the subthalamic nucleus, which in turn sends excitatory fibers to the medial globus pallidus/pars reticulata of the substantial nigra. Finally, dopamine input to the striatum seems to have an inhibitory impact on striatal fibers projecting to the lateral globus pallidus and an excitatory impact on fibers projecting directly to the medial globus pallidus. These aspects of neural transmission are shown in Fig. 4. Understanding of the basal ganglia loops and their neurotransmitters provides a basis for exploring functions of the basal ganglia.

III. FUNCTIONS OF THE BASAL GANGLIA

Despite all that is known about the macrostructure and microstructure of basal ganglia and about neurotransmitters in these structures, the functions of the basal ganglia in cognition and behavior are not definitively understood. Many theoretical descriptions of basal ganglia functions are based on the knowledge discussed previously. Such theories were devised to explain various aspects of cognition and behavior. A review of such theories could occupy a sizable volume on its own and, therefore, is well beyond the scope of this article. Here, we present brief descriptions of some theoretical propositions about basal ganglia, referring the reader to the suggested readings at the end of the article for greater detail. However, before undertaking these descriptions, a couple of preliminary matters must be addressed. The first is the means of studying basal ganglia functions; the second is general operational characteristics of the basal ganglia.

In addressing basal ganglia functions, we focus primarily on our understanding of human basal ganglia function. Much of the information discussed previously has been obtained from animal models. This method of exploring such issues as anatomical connectivity, neurotransmitter systems, and physiological responses has been long accepted. Although there can be substantial differences between humans' and other animals' brain systems, investigations across species can give us greater confidence in applying findings to humans. However, animal models can have limitations. Nowhere are such limitations more obvious than in the study of complex cognition. In particular, there are no good animal models for complex human language, for which the differences in human and animal brain systems are most important.

To specifically address basal ganglia functions in humans, many methods have been used. One such method is the lesion method. Modern X-ray computed tomography (CT) scanning and magnetic resonance imaging (MRI) have afforded a fair degree of accuracy in locating lesions caused by stroke, tumor, and other pathologies. Extensive behavioral and cognitive paradigms can be applied to patients with such lesions to describe how the lesions have disrupted behavior. However, naturally occurring lesions in humans are rarely limited to the structure of interest. Additional phenomena such as ischemic neuronal dropout and hypoperfusion are not well visualized on structural imaging techniques such as CT and MRI scanning.

This issue is particularly relevant to studies of basal ganglia lesions.

A related method is to study degenerative diseases of the basal ganglia such as Parkinson's and Huntington's diseases. Although the study of both diseases has provided important insights regarding probable basal ganglia functions, both diseases involve structures outside of the basal ganglia, including those of the cerebral cortex. This phenomenon makes it difficult to isolate basal ganglia functions. Another method of increasing importance in the study of human brain functions is functional neuroimaging, such as positron emission tomography or functional MRI. Functional neuroimaging has the potential to enhance our knowledge of brain systems but also has many drawbacks, including difficulty imaging some types of changes, an inability to readily distinguish inhibitory from excitatory neuronal activity, and the ambiguities almost always present in behavioral designs. Techniques such as surface evoked potentials have been of little value in studying basal ganglia function because of their inability to localize to deep structures.

A flurry of studies exploring basal ganglia function resulted from the increasing use of CT scanning in the 1970s and structural MRI scanning in the 1980s. The use of the lesion method is worth discussing in greater detail because of the specific problems related to exploration of basal ganglia function. In this regard, the most common lesions affecting behavioral and cognitive functions have been vascular. In particular, problems have been detailed that arise when striatocapsular infarctions are used to model basal ganglia functions. In the 1990s, Nadeau, Weiller, and others addressed these critical issues for human basal ganglia research. Striatocapsular infarctions are comma-shaped lesions that include portions of the caudate nucleus and putamen, the anterior limb of the internal capsule between them, and often portions of the globus pallidus. Although this region deep within the cerebral hemispheres is supplied by the lenticulostriate arteries, which branch from the initial segment of the middle cerebral artery, the cause of the infarction is often blockage of the middle cerebral artery just after it branches from the internal carotid artery or sometimes even blockage of the internal carotid artery itself. The reason that the whole territory of the middle cerebral artery in these cases does not show cystic infarction, which can be identified on structural imaging (CT or MRI), is that end-to-end anastomotic circulation from other arterial distributions supplies regions in the middle cerebral artery distribution outside the area of cystic infarction in the basal ganglia. The lack of cystic infarction does not mean that these areas are functioning normally, however. It has been demonstrated that patients who show significant and lasting cognitive deficits demonstrate poor anastomotic circulation and later recanalization of the middle cerebral artery after stroke, whereas patients without such deficits tend to have more adequate anastomotic circulation and/or early recanalization. Lasting cognitive deficits could be due to ischemic neuronal dropout or otherwise inadequate function in the territory of the middle cerebral artery outside the region of cystic infarction. The latter problems might be caused by circulation that is inadequate to support normal tissue function but not so inadequate as to cause cystic infarction that can be identified on a CT or MRI scan. Similar problems occur with hemorrhagic lesions where intracerebral hematomas under dynamic tension tend to cause pressure ischemia around the hematomas. In summary, large vascular lesions of the basal ganglia may be accompanied by damage or dysfunction in surrounding areas of cortex that are difficult to identify on structural imaging. In some instances, smaller lacunar infarctions may provide a better model for the study of basal ganglia function because they are not subject to the same vascular dynamics as larger lesions.

The second matter is the operational characteristics of the basal ganglia. In a seminal work in the late 1980s, Divac, Oberg, and Rosenkilde divided neural activity into quantitative and patterned neural activity. Quantitative activity affects the activity level of target structures and involves the number and rate of firing of neurons in the relevant structure; it does not code specific information. Patterned activity involves the temporal firing patterns of neurons and the temporal relationship between the firing of active neurons; such patterns are used to code specific information. Lesions will interrupt both quantitative and patterned neural activity; however, stimulation with implanted electrodes will cause firing of neurons in the stimulated structure that will mimic quantitative neural firing but interrupt specific temporal–spatial patterns of activity characteristic of patterned activity. It is critical to understand the neuronal activity of the basal ganglia if we are to understand its function. If the basal ganglia are limited to quantitative output, it is likely their function will be primarily of a regulatory nature, either on a tonic or phasic basis. However, if the basal ganglia are capable of processing patterned neural activity, then they are likely to be involved in complex information processing functions. As noted later, both positions have been represented by scientists working in the area.

In the following sections, functional considerations of the basal ganglia are addressed. How the basal ganglia operate is still a matter of study, and it may be years before generally agreed on models are developed. In other words, theories of basal ganglia function are still being developed and debated. In the discussion that follows, emphasis is placed on general theoretical considerations. The various functions that are considered include motor functions and movement programs, language, learning, and reward.

A. Motor Functions and Movement Programs

Given the motor symptoms of Huntington's and Parkinson's diseases and the major involvement of the basal ganglia in these disorders, the involvement of the basal ganglia in movement has been taken for granted. Huntington's disease is an autosomal-dominant genetic disorder characterized by choreiform movements (i.e., involuntary movements that resemble segments of voluntary movements). In the early stages of Huntington's disease, cell loss is obvious in the head of the caudate nucleus, but as the disease progresses cell loss becomes obvious in other parts of the basal ganglia and in the cerebral cortex, most conspicuously in the frontal lobe. Dopamine antagonists have been used to treat symptoms of Huntington's disease. Hemiballismus is also a disorder in which abnormal involuntary movements occur on one side of the body. Ballism represents the more forceful, or violent, form of involuntary movement. Discrete lesions of the subthalamic nucleus cause hemiballismus. Parkinson's disease is characterized by resting tremors, difficulty or slowness in initiating voluntary movements (bradykinesia), rigidity, and gait disturbance. Parkinson's disease is caused by loss of midbrain dopamine neurons (the pars compacta of the substantia nigra and the ventral tegmental area), particularly those projecting to the motor portions of the striatum. Theoretical explanations of the involvement of the basal ganglia in movement as far back as Bucy's (1942) paper attempt to explain the symptoms of one or more of these disorders by involvement of the basal ganglia motor loops.

Because of the involvement of the striatum and the subthalamic nucleus in disorders exhibiting involuntary movements (Huntington's disease and hemiballismus, respectively), the indirect basal ganglia loop involving projections from the striatum to the lateral globus pallidus, from the lateral globus pallidus to the subthalamic nucleus, and from the subthalamic nucleus to the medial globus pallidus has frequently been assigned the function of suppressing unwanted movements. One interesting facet of Huntington's versus Parkinson's disease is that the effects of the former can be mitigated with dopamine antagonists, whereas the effects of the latter can be mitigated with a dopamine precursor. These differences in response to dopaminergic agents, as well as the differences in motor symptoms (involuntary movements in Huntington's disease and bradykinesia in Parkinson's disease), have been related to the different phenomena affecting the indirect loop. The loss of striatal neurons influencing the indirect loop in the case of Huntington's disease prevents this loop from suppressing unwanted movements. The loss of inhibitory effects of dopamine on striatal neurons projecting into the indirect loop is purported to cause so much suppression of movement in the case of Parkinson's disease that it is difficult to switch to new movements. Rigidity in both Huntington's and Parkinson's diseases has been related to effects on the direct loop (loss of striatal neurons projecting to the medial globus pallidus in Huntington's disease and loss of dopaminergic excitation of these same neurons in Parkinson's disease). In their seminal 1986 work, Penny and Young gave detailed treatment to these differences.

In a recent review, Jueptner and Weiller described their work regarding functional neuroimaging of the basal ganglia and cerebellum. They concluded that the basal ganglia are involved in the selection of movements, whereas the cerebellum processes sensory information and integrates it with movement. The concept of movement selection is consistent with the suggestion that the basal ganglia enhance desired movements while suppressing unwanted movements.

B. Language

Nowhere has the function of the basal ganglia been more enigmatic than in the area of language. Formulations regarding the involvement of the basal ganglia in language have represented the entire range of possibilities from no involvement to a central role in selection of the words used to express oneself. One reason for skepticism about a role for the basal ganglia in language concerns the vascular lesion cases that have been used as evidence for such a role. As discussed previously, cortical dysfunction that is unseen on

structural CT or MRI images often accompanies basal ganglia infarction or hemorrhage. The variability of symptoms in cases with anatomically similar lesions suggests dysfunction of different cortical regions between such cases. Such cortical dysfunction may account for many of the more obvious language symptoms. A 1997 review by Nadeau and Crosson gave extensive treatment to this topic.

However, it appears likely that the presence of more obvious cortical symptoms masks more subtle deficits that may be due to basal ganglia dysfunction, and some investigators have suggested that core language symptoms can be found with lesions of the dominant neostriatum and internal capsule. In their 1994 work, Mega and Alexander indicated that these core elements include impairment of generative aspects of language (verbal fluency, sentence generation, and extended discourse). Based on evidence that stimulation of the dominant caudate nucleus can evoke language fragments, others have suggested that the basal ganglia may be involved in releasing cortically formulated segments for speech. This evidence was extensively discussed in works by Crosson and colleagues. It is worth noting that such hypotheses essentially propose regulatory functions for the basal ganglia which could be based on quantitative rather than patterned neuronal activity.

Perhaps the most comprehensive work on nonthalamic subcortical language functions was Copland's (2000) doctoral dissertation, which described a series of well-designed studies using cases of nonthalamic subcortical lesion of the left hemisphere as well as cases of Parkinson's disease. Among cases of nonthalamic subcortical lesion, basal ganglia damage was prominent. Various publications by Copland and colleagues are beginning to describe aspects of these studies. Copland's work indicated that patients with nonthalamic subcortical lesion have difficulty with higher level language processes (e.g., providing multiple meanings for semantically ambiguous sentences). A series of priming studies suggested that these patients have difficulty with resolution of lexical–semantic ambiguities under conditions involving controlled cognitive processing, with automatic processing less affected by the lesions. In the context of previous literature, the studies of Copland implied a role for the basal ganglia in executive language functions. His work also indicated that lesion of the basal ganglia and Parkinson's disease, which deprives the basal ganglia of its midbrain dopaminergic input, do not result in the same pattern of language-related deficits.

C. Learning and Memory

For some time, a division between the anatomical structures involved in declarative memory and those involved in procedural memory has been proposed. Declarative memory involves the ability to learn facts or items that can be deliberately recalled; procedural learning involves the ability to master a set of procedures to perform a specific task. Procedural memory has been explored primarily in patients with degenerative diseases of the basal ganglia (i.e., Huntington's and Parkinson's diseases). Although results in the cognitive realm of procedural learning have been mixed, findings have been more consistent with motor skill learning being impaired in patients with these diseases. Nonetheless, even within the realm of motor skill learning, evidence suggests that the basal ganglia are needed more for some skills than for others.

If one reviews the literature on memory deficits in Huntington's and Parkinson's diseases, problems with declarative memory are also evident. A relatively consistent pattern emerges: Patients with these disorders demonstrate maximum deficits on memory tasks requiring recall of information, but they demonstrate normal or near normal memory performance when only recognition of information is required. Some have interpreted better recognition than recall as an indication of difficulty in retrieving information that is actually stored in long-term memory. Problems with retrieval have been linked to the influence of basal ganglia circuits on frontal functions; however, cortical deterioration cannot be entirely ruled out as a cause of these difficulties.

Activity in prefrontal cortex has been associated with working memory functions. Working memory refers to the ability to hold information on line in order to use it for some impending activity. Because of its connectivity to the prefrontal cortex, it has been suggested that the basal ganglia are involved in working memory. However, support for the role of the basal ganglia in working memory is equivocal and the issue deserves further attention. Gabrieli reviewed many of these issues in 1995.

D. Reward

Studies have shown that dopaminergic neurons in the substantia nigra pars compacta and ventral tegmental area respond to administration of rewards and

guidance of behaviors to obtain rewards. In a seminal 1995 work, Houk and colleagues set forth a theory that explains how animals might learn that certain events tend to predict reward. The theory is quite complex and should be read in its entirety. Briefly, this theory suggests that cortical activity associated with events that might predict reinforcement creates a receptivity to dopamine in striatal spiny neurons residing within striosomal units. When reinforcement occurs, dopamine is released and instantiates a change in synaptic strength. This change at the synaptic level is posited to make firing of the striosomal spiny neuron more likely the next time the predictive event occurs. The hypothesis that synapses related to specific events are strengthened in striosomal neurons implies that patterned neuronal activity is required for this basal ganglia function. Continued study of the relationship between reward and dopaminergic activity in the basal ganglia promises not only to shed light on mechanisms involved in motor skill learning but also to reveal underlying motivational mechanisms for a variety of behaviors.

IV. CONCLUSIONS

The basal ganglia are phylogenetically older structures deep within the cerebral hemispheres. Connectivity between basal ganglia structures, the cortex, and the thalamus are organized in parallel but separate cortico-striato-pallido-thalamo-cortical loops. The loops project only back to the cortical region that projects to their striatal component. Although these loops are thought to be critical to understanding basal ganglia functioning, the precise role of the basal ganglia in behavior and cognition is not well understood. Without a doubt, the glutamatergic connections between the cortex and the striatum and between the thalamus and the cortex, and the GABAergic connections between the striatum and the globus pallidus and between the globus pallidus and the thalamus, play a significant role in basal ganglia functions. The inputs from the intralaminar nuclei to the striatum may also play a significant role in basal ganglia functions.

Consensus indicates that the basal ganglia play an important role in motor behavior and learning. Many of the devastating effects of Huntington's disease and Parkinson's disease on movement appear to be due to changes within the neostriatum and closely related structures. The basal ganglia are also thought to play significant roles in various cognitive processes related to executive functions. These roles probably involve selection and initiation of responses that are simultaneously tied to external circumstances and internal motivations. The basal ganglia may also play some role in complex language processes, working memory, motor procedural memory, and reward. Continuing research will provide increasing definition to these roles.

See Also the Following Articles

BEHAVIORAL NEUROGENETICS • COGNITIVE PSYCHOLOGY, OVERVIEW • COGNITIVE REHABILITATION • MOVEMENT REGULATION • NEUROANATOMY • NEUROTRANSMITTERS • PARKINSON'S DISEASE • THALAMUS AND THALAMIC DAMAGE

Suggested Reading

Alexander, G. E., DeLong, M. R., and Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Ann. Rev. Neurosci.* **9,** 357–381.

Bucy, P. C. (1942). The neural mechanisms of athetosis and tremor. *J. Neuropathol. Exp. Neurol.* **1,** 224–231.

Chenery, H. J., Copland, D. A., and Murdoch, B. E. (1999). The processing of lexical ambiguities within a sentential context following nonthalamic subcortical lesions. *Brain Language* **69,** 405–408.

Copland, D. A. (2000). A real-time examination of lexical ambiguity resolution following lesions of the dominant nonthalamic subcortex. Doctoral dissertation, University of Queensland, Australia.

Copland, D. A., Chenery, H. J., and Murdoch, B. E. (2000a). Persistent deficits in complex language function following dominant nonthalamic subcortical lesions. *J. Med. Speech–Language Pathol.* **8,** 1–15.

Copland, D. A., Chenery, H. J., and Murdoch, B. E. (2000b). Processing lexical ambiguities in word triplets: Evidence of lexical–semantic deficits following dominant nonthalamic subcortical lesions. *Neuropsychology* **14,** 379–390.

Crosson, B. (1992). *Subcortical Functions in Language and Memory*. Guilford Press, New York.

Crosson, B. (1997). Subcortical limb apraxia. In *Apraxia: The Neuropsychology of Action* (L. J. G. Rothi and K. M. Heilman, Eds.), pp. 207–243. Erlbaum, East Sussex, UK.

Crosson, B., Zawacki, T., Brinson, G., Lu, L., and Sadek, J. R. (1997). Models of subcortical functions in language: Current status. *J. Neurolinguistics* **10,** 277–301.

Divac, I., Oberg, G. E., and Rosenkilde, C. E. (1987). Patterned neural activity: Implications for neurology and neuropharmacology. In *Basal Ganglia and Behavior: Sensory Aspects of Motor Functioning* (J. S. Schneider and T. I. Lidsky, Eds.), pp. 61–67. Hans Huber, Lewiston, NY.

Gabrieli, J. (1995). Contribution of the basal ganglia to skill learning and working memory in humans. In *Models of Information Processing in the Basal Ganglia* (J. C. Houk, J. L. Davis, and D. G. Beiser, Eds.), pp. 277–294. MIT Press, Cambridge, MA.

Gerfen, C. R. (1992). The neostriatal matrix: Multiple levels of compartmental organization in the basal ganglia. *Ann. Rev. Neurosci.* **15,** 285–320.

Graybiel, A. M., Baughm, R. W., and Eckstein, F. (1986). Cholinergic neuropil of the striatum observes striosomal boundaries. *Nature* **323,** 625–627.

Heimer, L., Harlan, R. E., Alheid, G. F., Garcia, M. M., and de Olmos, J. S. (1997). Substantia innominata: A notion which impedes clinical–anatomical correlations in neuropsychiatric disorders. *Neuroscience* **76,** 957–1006.

Houk, J. C., Adams, J. L., and Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In *Models of Information Processing in the Basal Ganglia* (J. C. Houk, J. L. Davis, and D. G. Beiser, Eds.), pp. 249–270. MIT Press, Cambridge, MA.

Jueptner, M., and Weiller, C. (1998). A review of differences between basal ganglia and cerebellar control of movements as revealed by functional imaging studies. *Brain* **121,** 1437–1449.

Mega, M. S., and Alexander, M. P. (1994). Subcortical aphasia: The core profile of capsulostriatal infarction. *Neurology* **44,** 1824–1829.

Middleton, F. A., and Strick, P. L. (2000). Basal ganglia and cerebellar loops: Motor and cognitive circuits. *Brain Research Reviews* **31,** 236–250.

Nadeau, S. E., and Crosson, B. (1997). Subcortical aphasia. *Brain Language* **58,** 355–402.

Penney, J. B., Jr., and Young, A. B. (1986). Striatal inhomogeneities and basal ganglia function. *Movement Disord.* **1,** 3–15.

Weiller, C., Willmes, K., Reiche, W., Thron, A., Insensee, C., Buell, U., and Ringelstein, E. B. (1993). The case of aphasia or neglect after striato-capsular infarction. *Brain* **116,** 1509–1525.

Wilson, C. J. (1995). The contribution of cortical neurons to the firing pattern of striatal spiny neurons. In *Models of Information Processing in the Basal Ganglia* (J. C. Houk, J. L. Davis, and D. G. Beiser, Eds.), pp. 29–50. MIT Press, Cambridge, MA.

Yelnik, J., Percheron, G., and Francois, C. (1984). A golgi analysis of the primate globus pallidus. 2. Quantitative morphology and spatial orientation of dendritic arborizations. *J. Comp. Neurol.* **227,** 200–213.

Behavioral Neurogenetics

FRANS SLUYTER
Institute of Psychiatry, United Kingdom

ECO DE GEUS
Vrije Universiteit, The Netherlands

GILLES VAN LUIJTELAAR
University of Nijmegen, The Netherlands

WIM E. CRUSIO
Brudrick Neuropsychiatric Research Institute

GLOSSARY

allele Alternative form of a specific gene, sometimes leading to the production of different proteins or RNA products.

chromosome One of the DNA–protein structures that contains part of the nuclear genome of a eukaryote.

epistasis Alteration of phenotypic effects of one gene by the specific alleles of another gene.

gene A stretch of DNA coding for a specific protein or RNA product.

genetic correlation Correlation between the effects of one single gene or a set of genes on two different phenotypes; indicative of pleiotropy.

genome The entire genetic complement of a living organism.

genotype Genetic constitution (set of alleles present in each cell of a particular individual), as contrasted to phenotype.

heritability The proportion of phenotypic variation for a given character that can be associated with genetic variation in a given population. Heritability, commonly denoted h^2, is therefore character and population specific.

homologous recombination A process by which one DNA segment can replace another DNA segment with a similar sequence.

inbreeding Reproduction by the mating of closely related individuals resulting in an increased frequency of homozygotes.

knockout mice A mouse that has been engineered so that it carries an inactivated gene.

locus A location on a specific chromosome defining the position of a specific DNA sequence.

phenotype A measurable characteristic manifested by an organism. Phenotypes can be defined at many levels, ranging from the isoform of a protein to a complex trait such as neuroticism. Two organisms may have the same genotypes but different phenotypes (owing to environmentally induced variation). Also, two organisms may have the same phenotypes but different genotypes (e.g., heterozygotes and homozygotes in the case of dominance).

pleiotropy A single gene affects two or more different phenotypes.

quantitative trait locus (QTL) A chromosomal locus correlated with individual differences in a given trait. Such a locus is believed to contain a gene in which differences in the coding DNA cause quantitative individual differences in that trait. A QTL generally explains only part of the total genetic variation.

recombination The process during germ cell formation that generates sperm or eggs with allele combinations different from those of the individual's own chromosomes.

selective breeding Also called artificial selection. Animals that score high (or low) on a desired trait are artificially selected and mated to produce offspring.

transgenic mice Mice that have a foreign gene inserted into their genome.

wild type The most frequent allelic form of a gene found in natural (or laboratory) populations.

zygosity Twins may be dizygotic (stemming from two eggs), in which case they share on average half of their parental genes identical by descent. Twins may also be monozygotic (stemming from one egg), in which case they share all their genes identical by descent (so-called "identical" twins).

Behavioral neurogenetics is the study of the way in which genes, directly or in interaction with the environment, influence the structure, chemical composition, and/or function of the central nervous system and, through this, cause individual differences in behavior. This article presents a brief overview of the field, providing some examples from human and animal studies.

I. HISTORICAL BACKGROUND

Plato, in *The Republic*, speculated that behavioral and personality characteristics might be heritable, but only in more recent times have such phenotypes become amenable to genetic analysis. In the 19th century, Francis Galton (1822–1911) started the scientific study of human behavioral phenotypes and, in the second half of the 20th century, the field of behavior genetics was born with the publication of John Fuller and Robert Thompson's classic book *Behavior Genetics* in 1960. By 1960, it was quite clear that behavior is influenced by genetic factors, as shown by a wealth of studies on inbred strains and crosses between them, by the generation of artificially selected lines, and by the results of twin studies. However, with the exception of some neurological mutations with rather severe phenotypical effects, the genes remained elusive. Questions concerning where the genes were located on the chromosomes, what proteins they encoded, and where in the nervous system they were expressed remained a mystery. All this changed in the 1970s when a new technology, recombinant DNA, was born that allowed scientists to study the molecular mechanisms underlying neural function. At the same time, a few scientists, such as Benson Ginsburg and Dick and Cynthia Wimer in the United States and Hans van Abeelen in Europe, began to integrate behavioral, neurobiological, and genetic aspects of their research, and this integration has continued in both animal and human studies. This led to the emergence of the new and budding field of behavioral neurogenetics, which is rapidly coming of age with its own specialist society (the International Behavioural and Neural Genetics Society; *http://www.ibngs.org*).

The new genetic technologies that have become available in recent years have created very high expectations for progress in the near future. Although it is doubtful that single gene analyses will help us to elucidate complex cognitive functions, the manifold advantages offered by modern behavioral–neurogenetic analysis should help us unravel the cellular basis of brain–behavior relationships. This creates great potential for the development of new therapeutic tools.

II. FROM TRAIT TO GENE AND BACK: A GENERAL OUTLINE

The fundamental goal in behavior neurogenetics is to understand the genetic basis of the neuronal processes underlying individual differences in behavior. At the outset, the relative contribution of genetic factors to a trait needs to be established. In humans, the similarities in a trait are examined in subjects of different degrees of genetic relatedness. Comparison of monozygotic and dizygotic twins in the classical twin method, for instance, allows a reasonable estimate of the heritability of a trait. In animals, for instance, the ability to selectively breed lines that score either extremely low or extremely high on a trait is a good indication of the amount of genetic influence on that trait. If substantial heritability has been demonstrated for a trait, it becomes feasible to search for the building stones, i.e., the differences in coding DNA sequences underlying the genetic part of the individual differences in that trait.

Finding the actual genes is a formidable task for several reasons. Many of the targeted behaviors (aggression, cognitive ability, and depression) are very complex and can usually be quantified on more than one aspect (often with more than one measure per aspect). As a first step the trait under study is refined by using a combination of multiple measures of the trait that best capture a common underlying genetic factor. A subsequent problem is the vast amount of genes involved. Candidate gene studies can be used when previous studies have identified a specific gene that codes for a protein involved in a pathway known to be relevant to the variation of the trait under study. This applies only to genes with known location and function and to pathways

that are already partially understood. All other situations need a different approach. Whole genomic searches are used to establish the most likely location on the entire DNA of genes that influence the trait under study. These may be genes that were identified but had not been suspected to be linked to the trait, or they may be new genes. Because there is a large degree of homology between animal DNA and human DNA, animal studies can be very helpful to focus genomic searches on only a part of the human DNA (a single chromosome or even a part thereof). Also, after a putative location has been found, animal studies may help to fine map that region to find the actual gene. Gene finding in animals is generally easier because various types of controlled crosses, such as repeatedly backcrossing or intercrossing, can be made at will.

Once the gene has been found, many studies can be performed to elucidate the exact biological pathway by which the allelic variation in the gene influences variation in the neurophysiological or behavioral trait. In animals, genes can be deleted (knockout models) or inserted (transgenic animals) to investigate the scope of the gene's effects and how it operates. By selecting either animals with a known genetic background or humans with known allelic variants for the gene under study, it is possible to do gene-by-environment studies in which the differential effects of manipulations of the environment on different genotypes can be directly tested. Most important, once animals/humans can be characterized by their genetic predisposition for a trait, structural (e.g., size of hippocampal cell population) and functional aspects of the brain (e.g., electrophysiological response to a stimulus) can be compared to lay bare the actual biological pathways connecting gene and behavior.

In the following sections, the most commonly applied research paradigms in behavioral neurogenetics will be discussed and illustrated on the basis of a specific trait. The methods to determine the actual presence of genetic factors—selective breeding and/or comparing inbred strains in animal studies and the classical twin method in human studies—are discussed in Sections III and IV, respectively. The search for genes is discussed in Sections V (linkage, association studies in humans) and VI (intercrosses and backcrosses in animals). Recently developed techniques, such as knockout and transgenic methods in animals, are discussed in Section VII, while interactions between genes and the environment are discussed in Section VIII. Finally, in Section IX we focus on the intermediate brain structures.

In our discussion of the various research paradigms, we use aggression as the main example of a behavioral trait. Aggression is exemplary for complex traits such as depression or cognition, with their multiple aspects and multiple levels of measurement. One of the first problems encountered is the definition and the consequent way of measuring the trait under investigation. In humans, we can use the outcomes of questionnaires as a measure of aggression, but we can also use registered criminal acts or convictions for violence. All three variables might very well yield different results. Aggressive behavior in animals appears to be easier to measure than in humans: We can stage an actual fight between two animals. However, there are many different ways to influence the results of a test—for example, by varying the opponent to be attacked, the environment in which the encounter takes place, or the duration of the test. Also, many different aggression parameters are used, ranging from simple ones such as attack latency, which is the time it takes for the animal under investigation to attack its opponent, to more complex measures that combine different aspects of offensive behavior over time and calculate an overall aggression score. This has created an abundance of aggression tests and measures that are sometimes difficult to compare.

Different measures of aggression may be under distinct genetic control mechanisms. Thus, the way aggression is measured will strongly influence results of behavioral neurogenetic testing. This is not only true for aggression but also for all other complex traits. This cautionary note must be kept in mind.

III. STRATEGIES IN ANIMALS: INBRED LINES AND SELECTIVE BREEDING

In animal studies there are two fundamental strategies to determine whether a trait is under the influence of genetic variation. The first one is the comparison of inbred strains. An inbred strain is generated by repeatedly mating close relatives. Since these relatives (usually brothers and sisters) are approximately 50% genetically alike (counting only those genes for which their parents had different genotypes), individuals of the same sex will be almost genetically identical after 20 generations of inbreeding. Within an inbred strain, nearly all trait variability will be caused by the

environment, whereas differences among strains will be virtually genetic in origin. Therefore, when in a controlled testing environment multiple strains are compared for a specific behavior, the extent to which inter-strain differences exceed the pooled within-strain variability provides a test of the existence of genetic influence. A good illustration of the variation in inbred strains is the data collected on aggressive behavior in the laboratory of Pierre Roubertoux in Orléans, France (Table I). As can be observed in the table, levels of aggressiveness vary enormously among strains. NZB mice always attack (20 of 20 animals; 100%), whereas C57BL/6J males almost never attack (1 of 20; 5%).

Inbred strains are generally stable over time and across laboratories. For instance, NZB mice have been demonstrated to be aggressive in several laboratories on different continents since they were first tested. However, certainly the most important advantage of inbred strains is that they are the conditio sine qua non for the recently developed high-tech molecular–genetic techniques, such as the creation of knockouts and transgenics. A disadvantage of the use of inbred strains is that they do not represent natural populations in which heterozygosity is the rule rather than an exception.

Another useful technique to show that a specific trait is genetically influenced is selective breeding or artificial selection. This method has been used throughout recorded history, even long before people understood how it actually worked, and is currently still being applied extensively. It is based on the fact that the offspring of animals with a desired quality are more likely to demonstrate that quality than the progeny of random individuals. For centuries, farm animals and pets have been bred for desired characteristics, some of them behavioral (e.g., in dogs: hunting dogs, shepherd dogs, etc.). Similarly, mice and rats can be bred for various behaviors, such as learning performances or aggression. Usually, animals are selected for the opposite of the desired behavior (directional selection).

Table I
Proportion of Attacking Males in Seven Inbred Mouse Strains

Strain	% attacking males (*N=20*)
BA	20
BALB/c	40
CBA/H	20
C57BL/6	5
CPB-K	50
DBA/2	50
NZB	100

For aggressive behavior, various pairs of selection lines exist. One of the more conspicuous ones originated from animals from a feral population caught in The Netherlands that were subsequently selectively bred for behavior in a test cage developed by Geert van Oortmerssen. Different chambers of the cage were connected through slide doors such that home territories and a border area were created. Since in the wild most agonistic encounters occur at the border of the territories, this design takes the natural settings of the test animals into account. Moreover, the aggression test takes 3 days, with one encounter every day. This procedure not only reduces the effects of chance but also creates the opportunity to investigate the development of aggression over time. Using the previously mentioned paradigm, van Oortmerssen selected two lines for attack latency (Fig. 1): one highly aggressive line, characterized by short attack latencies (SALs), and one low to nonaggressive line, characterized by long attack latencies (LALs).

Although it is beyond the scope of this article to discuss the theory and practice of selected lines in detail, the following strict conditions must be met in order to reliably reveal the genetic underpinnings of behavior. First, one has to bear in mind the obvious. Behavior without genetic variation underlying it cannot be selected for or against. Therefore, if no differences for the desired trait are observed in a panel of inbred strains, it is theoretically impossible to selectively breed for that characteristic. Second, the simpler the phenotype, the better the experiment generally works. Third, if possible, a genetically defined, replicable founding population should be used. In this case, one can always use the parental population if needed. These conditions can be met relatively easily and without great financial resources. However, a last theoretical requirement results in almost prohibitive costs. Since selection depends on genetic variation, inbreeding is its natural enemy. Inbreeding often leads to mortality or loss of fertility, but it also conflicts with the desired result of selection: to increase the frequency of only those alleles that favor the trait under selection. Inbreeding is not desired for the vast majority of genes that need to be kept segregated. To this end, one would have to maintain an unrealistically large number of animals. In practice, selective breeding leads to the subsequent loss of alleles (genetic drift). However, it is possible to

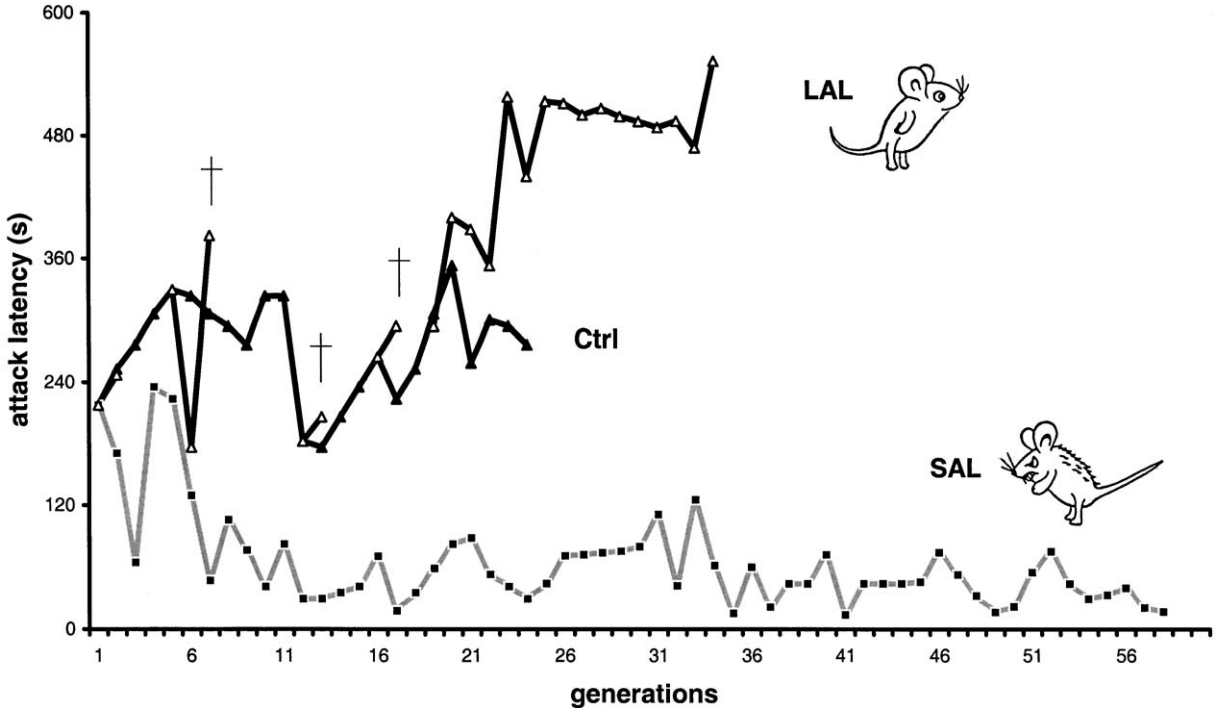


Figure 1 Artificial bidirectional selection for attack latency in male wild house mice. Three lines have been established: one control line (Ctrl) and two selection lines—one highly aggressive line, characterized by short attack latencies (SAL), and one low to nonaggressive line, characterized by long attack latencies (LAL). Differences in the number of generations between SAL and LAL mice originated from difficulties in developing the LAL line per se (†, unsuccessful attempts) and in unequal rates of reproduction, with the SAL females producing both earlier and larger litters.

control to a certain extent for genetic drift through the establishment of separate, replicated lines.

IV. STRATEGIES IN HUMANS: THE TWIN METHOD

The twin method is one of the most powerful and frequently used tools to determine genetic and environmental influences on variation in traits. It compares the resemblance between genetically identical, monozygotic (MZ) twins to the resemblance between fraternal or dizygotic (DZ) twins. MZ twins have the same allelic combinations at all their genes, whereas DZ twins share, on average, 50% of allele combinations (counting only those genes for which their parents had different genotypes). The shared environment for both is very comparable. Hence, the demonstration that MZ twins are more similar than DZ twins for a certain trait points to a genetic contribution to the variation in this trait. Efficient use of the information available in the variances and covariances of genetically related subjects (e.g., twins) can be done with a statistical technique called structural equation modeling. Basically, the total variance in a trait is decomposed into several contributing factors: (i) additive genetic variance, which results from the additive effects of the two alleles of all contributing genes; (ii) dominance genetic variance, which results from the nonadditive effects of the two alleles for all genes showing a dominance effect; (iii) shared environmental variance, which results from environmental events shared by both members of the twin pair (e.g., school and diet); and (iv) unique environmental variance, which results from nonshared environmental effects and also includes measurement error. The influence of confounding variables such as age or socioeconomic status can be incorporated in the model. In addition, a further refinement from univariate to multivariate models allows exploration of the question of whether any covariance between different variables is genetically and/or environmentally determined.

There are some pitfalls in the twin method that need to be avoided. First, there is the matter of zygosity. An underestimation of the number of MZs will lower

estimates of genetic effects because they will raise the fraternal twin correlation and lower the correlation of identical twins. Conversely, underestimation of DZs will raise estimates of genetic influences. Correct zygosity has become even more important with the increasing use of DZ twins for the location of quantitative trait loci (QTL). An accurate zygosity diagnosis is now easier than ever. Usually, six to eight highly polymorphic DNA markers are adequate to discriminate between MZ and DZ twins. The second pitfall is the so-called equal environments assumption, which holds that the degree of environmental similarity is comparable in MZ and DZ twins. At first, this assumption seems quite justified. Both types of twins share the same womb and are reared in the same family, all at the same time. However, the two types of twins may be treated differentially. For instance, MZ twins may be treated more similarly than fraternal twins. If this is true and MZ twins would experience more similar environmental influences than DZ twins, the genetic contribution to the variation of the trait under investigation would be overestimated. Although there is evidence that the equal environment assumption is violated, studies on a range of traits suggest that the violations do not seriously compromise the validity of the twin method. Finally, like any other study, twin studies make sense only when statistical power is adequate. In earlier studies samples were often too small because experimenters thought they could use the same sample size required to compare means. Power studies have demonstrated that large samples are necessary for an accurate estimate of the presence and size of genetic influences, especially if heritability is low.

The use of the twin method in the genetic analysis of human aggression has been demonstrated in a recent study. Using three types of questionnaires, nine different aspects of aggression were measured: physical assault, indirect hostility, irritability, negativism, resentment, suspicion and verbal hostility, trait anger, and type A behavior. All these aspects showed moderate to fair heritabilities, varying from 23% for indirect hostility to 53% for resentment and suspicion. A multivariate analysis revealed two common additive genetic and two common environmental factors. Hence, two sets of genes and two sets of unique environmental factors give an adequate description of the pattern of associations between the nine measured aspects of human aggression. Of course, the nine measures are also individually influenced by specific additive and/or environmental factors. This replicates an extensive literature on twins and aggression or related concepts such as antisocial behavior and hostility (thoroughly reviewed elsewhere). The correlation between MZs is not perfect, showing that unique environmental factors play a role in these types of aggression. In all studies, however, MZs are more similar than DZs, indicating a genetic contribution to the variation in aggressive behavior. Finally, results from a related design, the adoption method, further corroborate the existence of a significant genetic contribution to aggression.

The obvious question remains, which genes are responsible for the variation in aggression? Also, where are they located on the chromosomes? What proteins do they encode? Where are they expressed in the brain? Are they always expressed or only during a certain time period? Actually, there is a gene that is specifically associated with a behavioral phenotype that includes disturbed regulation of impulsive aggression. However, before we discuss this gene, we review the techniques to find "behavioral" genes.

V. CANDIDATE GENE STUDIES AND GENOMIC SEARCHES

The major problem facing gene hunters is not the sheer number of possible genes that need to be examined, of which—despite the recent completion of the Human Genome Project—the majority are unidentified, but the relatively small contribution each gene may make to the relevant trait. As a rule, genes will be found more easily if they explain more of the variance in a trait. Therefore, gene finding is relatively easy if only a single gene affects the trait. In these instances, a simple Mendelian segregation of a limited number of phenotypes is observed for all possible genotypes at a specific locus. Many rare diseases or disorders (but also Huntington's disease) are Mendelian in nature. As a general strategy to find such Mendelian genes, many DNA markers of known location, evenly dispersed throughout the entire genome, are measured in individuals from multiple generations. DNA markers can be mutations in a single base pair [single nucleotide polymorphisms (SNPs)] or a variable number of repeats of two or more base pairs (microsatellites) and need not be part of a functional gene: They are just landmarks in the genome. As with genes, all individuals have the same markers (in this sense the term "marker" may be misleading); it is the allelic variant of the marker that may differ between individuals. When a given marker is situated near the gene influencing the

trait of interest, allelic variants of the marker and the gene will be more likely to be transmitted together to the next generation than if they are distant or on different chromosomes. This so-called cosegregation, however, is not perfect. Sometimes, the marker and the gene may be separated by recombination events during meiosis. The extent to which marker and gene cosegregate is referred to as "linkage," and (this is a crucial assumption) the chance of linkage increases if the marker and the gene are close physically (although not necessarily linearly—not all chromosomal locations have equal chance for recombination). For each marker, evidence for linkage is derived using statistical procedures that trace the cosegregation of the trait (and thus in many instances the gene) and a specific variant of the DNA marker along familial lineages in extended pedigrees. Simply stated, if two children resemble each other for a certain trait and they both received exactly the same variant of a DNA marker from the same parent, that marker might be close to the gene influencing the trait.

Linkage analysis assigns a probability value (LOD score) to all markers, and a LOD score profile is obtained for each chromosome. Evidence for linkage is said to be present when the maximal LOD score exceeds a predefined threshold, which depends on the size of the genome and the number of genotyped markers. The chromosomal region surrounding a marker with a significantly high LOD score will be selected for fine mapping, which is essentially a repetition of the same procedure but now with all markers concentrated in the area of interest on a single chromosome. If the region containing the putative gene is sufficiently small, the DNA in the entire region is sequenced in full for a few persons or animals. Because genes have a specific structure, this identifies all genes in the region. By comparing all base pairs in these genes in many different persons/animals, the sites of allelic variation, also called polymorphisms, within these genes can be identified (mutational analysis). If the trait is a disease or disorder, comparison of the polymorphisms between patients and controls without the disease will ultimately reveal which allelic variant is responsible for the disease. The entire process from significant LOD scores to the actual allelic variants is usually called "positional cloning."

In 1993, using classical linkage in pedigrees, Han Brunner and colleagues discovered a point mutation associated with a behavioral phenotype that includes disturbed regulation of impulsive aggression. This mutation affected a gene that codes for one of the two isozymes [monoamine oxidase A (MAOA)] that are responsible for the breakdown of several neurotransmitters, including dopamine, noradrenaline, and serotonin. Since the MAOA gene is located on the X chromosome and the mutation is recessive, only males that possess the mutant allele show behavioral changes. One should nonetheless be extremely cautious in calling this gene an "aggression gene." The phenotype is not limited to impulsive aggressive behavior but also extends to borderline mental retardation, arson, attempted rape, and exhibitionism. Aggression is therefore merely one of the impulsive behaviors by which these individuals differ from "normal" people.

Because of its power to localize classical Mendelian genes, linkage analysis has been the workhorse for mapping genes for simple monogenic traits/diseases. Unfortunately, most complex traits (depression, intelligence, and aggression) are polygenic, i.e., they are influenced by many different genes, environmental factors, and their possible interactions. These interactions involve gene–gene interactions (epistasis), gene–environment interactions, and environment–environment interactions. Also, the same trait may be brought about by different subsets of genes in different individuals (genetic heterogeneity). Traits that are influenced by many genes and environmental factors are usually quantitative traits, and each of the genes that influence them is called a polygene. The locus where such a polygene can be found is called a quantitative trait locus (QTL). The contribution of a single polygene to the population variance in most complex behaviors is likely to be very small. Statistical power for the detection of such a QTL remains a major concern for the simple reason that only one gene (explaining 30% of the variation in fruit size in tomatoes) has been identified using these methods. Among the various solutions to boost power are the use of isolated populations to reduce genetic heterogeneity, the use of more DNA markers, and the use of selected family members (e.g., sibling pairs with either very high or very low values for the trait).

Two general alternative approaches to find genes for complex traits can be distinguished.

A. Candidate Genes

In some cases, there may be good theoretical reasons to focus on a single candidate gene. For instance, because neurotransmission is crucial to virtually every behavior, all known genes for receptors, transporters, or

synthesis elements for neurotransmitters are usable as candidate genes. The ideal candidate gene has been shown to be functional: It influences the concentration of the (iso)form of a protein, its functionality or efficiency, or, perhaps most important, its responsiveness to environmental factors triggering the expression of the gene. All candidate gene studies are association studies and are similar in design to classic case-control studies in epidemiology. DNA is collected from all participants and the trait is compared across the various allelic variants of the candidate gene. Also, frequencies of the various allelic variants may be compared in subjects with a particular disease to detect an association between a particular allele and the occurrence of the disease.

The main problem of association studies is false positives that arise due to population stratification. The famous example is the "chopstick" gene. Suppose that the genome of random San Francisco inhabitants was used in a study on the complex trait of "using chopsticks to eat" without stratification for Chinese or Anglo-Saxon background. There would be many genes associated with chopstick eating simply because frequencies differ between Chinese and Anglo-Saxon populations for a multitude of genes.

B. Genomic Searches

In most cases, no a priori information is available on the nature or location of the gene in question. A random search of the genome for QTL is then the most viable strategy employed. A first strategy is sib pair analysis. As in classical pedigree studies, several hundred DNA markers are obtained from siblings and (optimally) their parents. Siblings in a sib pair study, however, come from different families rather than from a few single pedigrees. By definition, the differences between two siblings for a trait will be smaller if they share the same variant of a QTL for that trait. Linkage of a marker to a QTL implies that the differences in the trait between the siblings will also be smaller if they share the same variant of the marker, obtained from the same parent. In a regression procedure originally derived by Haseman and Elston, the likelihood that a marker is in linkage with the trait is obtained by regressing the trait difference between siblings on the proportion of marker alleles shared identical by descent (IBD). QTL for complex common diseases can also be mapped by affected sib pair analysis. The probability that siblings share zero, one, or two alleles IBD for any gene is 0.25, 0.5, and 0.25, respectively. However, if a marker locus is close to the QTL this will be detectable as increased allele sharing IBD at the marker by two affected sibs. A major drawback of this method is that it requires large numbers of sibs to detect significant evidence for linkage.

With recent advances in genetic molecular technology, several thousand SNPs can be assessed on a microarray or DNA chip. This allows a second strategy for genomic searches: allelic association to detect linkage disequilibrium. This method is akin to the association approach in candidate gene studies. However, it is not the variation in a known functional gene that is associated to the trait but, rather, the variation in a series of DNA markers (which may but need not be in any functional gene). Linkage disequilibrium occurs when a marker allele and the QTL are so close on the chromosome that they remain linked over many generations of meiotic recombination. The advantage over sib pair analysis is that linkage disequilibrium mapping can detect the region of a QTL with only very small effects on the trait. The disadvantage is that linkage disequilibrium is found only when the marker and QTL are very close. A complete genomic search may require thousands of markers. However, technological barriers have been removed with much more ease than logistical barriers (i.e., the sampling of very large numbers of siblings).

VI. FROM ANONYMOUS QTL TO IDENTIFIED GENE IN ANIMALS

Increasingly, the laborious steps of positional cloning have become easier due to the ongoing sequencing of the entire human genome. A rapidly increasing database on all polymorphic DNA is now available as are appropriate data mining algorithms to combine this wealth of genetic data with existing protein and mRNA databases. Homology searches in DNA, mRNA, or protein databases for possible genes in the region identified by linkage analysis can aid in identification of the candidate genes. Although this will speed up gene hunting immensely, the need to first identify the region of interest in a genomic search and to narrow that region by (repeated steps of) fine mapping remains. Only after the region is sufficiently small (e.g., <100 genes) does the positional candidate gene approach becomes feasible. Repeated fine mapping is expensive and laborious, particularly when the

low statistical power of each repeated search step is taken into account. Not surprisingly, most QTL mapping has been conducted in animals. Animal QTL studies have the clear advantage that the environment can be controlled, and that a sufficient number of progeny and types of crosses can be performed. The latter advantage is of great importance if the chromosomal region containing a QTL is still very large. Various strategies are available, including constructing congenic strains. They are produced by repeatedly backcrossing the strain with the mapped QTL (donor) to another strain (recipient) while checking each backcross for the presence of the QTL using flanking DNA markers. After a number of predefined backcrosses, a strain is developed that is genetically almost identical to the recipient strain except for the QTL area. Phenotypic comparisons between congenic and recipient strains might verify the existence of the QTL, its impact, and possible interactions with other QTLs. Once the existence of the QTL has been proven by means of congenic lines, the actual fine mapping can commence. This is done by phenotyping substrains that are recombinant at various places in the QTL area.

Other strategies to fine map QTLs are the production of recombinant congenic strains, advanced intercross lines, or interval-specific congenic strains.

Also, exclusively in animals, transcript mapping techniques such as reverse transcriptase polymerase chain reaction and exon trapping can be used to locate genes. For a detailed review of these strategies, the reader is referred to the suggested reading.

VII. KNOCKOUT AND TRANSGENIC STRATEGIES

The development of targeted gene disruption has been one of the more important breakthroughs in unraveling the molecular underpinnings of behavior. The aim is to selectively inactivate a gene of interest (i.e., disrupt a targeted gene) and to compare this so-called knockout animal (usually a mouse) with a control or wild-type animal that has all its genes intact. Observed differences can then be attributed to the gene in question. Hence, by comparing the behavior and underlying neuronal processes of knockouts and wild types, one can deduce the function of the gene and determine its effects on complex traits.

It is beyond the scope of this article to discuss the technical details of this technique. Briefly, this technique relies on the fact that embryonic stem cells can be grown *in vitro* and modified by transfection. By homologous recombination the wild-type (or normal) gene in question can be replaced by a disrupted form of the gene that for example, codes for an antibiotic. The latter is used to select for recombined embryonic stem cells. These engineered cells are then injected into a recipient embryo to form chimeric mice (mice that have both cells with the normal gene and cells with the disrupted gene). Next, only those mice that have transmitted the mutation in their germline are used to generate nonchimeric mice carrying this artificial mutation.

To date more than 30 genes that influence male offensive behavior have been identified using the knockout technique (Table II). Interestingly, the involvement of the MAOA gene in abnormal social behavior (including aggression) was confirmed by a knockout study. Mice having a disrupted MAOA gene show more aggressive behavior than control mice. Also, the relatively nonspecific consequences of this gene defect were confirmed: MAOA-deficient mice tremble, have difficulty in righting themselves, and are fearful and blind. Other knockout lines show a more specific aggressive behavioral profile. For instance, mutant mice lacking the 5-HT1B receptor do not exhibit any obvious developmental or behavioral defects but do show enhanced aggressive behavior. In this respect, it is notable that about half of all studies on gene knockouts and aggression indicate that there is either a direct or an indirect relation between aggression and serotonin. Although the exact mechanisms through which serotonin exercises its influence on aggressive behavior is far from clear, data suggest that a low serotoninergic transmission is associated with increased aggressive behavior.

At this point, some comments on knockout studies must be made. First, there is always the possibility that the knockout and the wild type differ in more than one gene. This so-called "flanking gene" problem results from the technical procedure per se and may lead to false positives. A second problem is the genetic background of the knockout, which is either randomized or, at best, homogeneous. In the latter case, the obtained knockout is repeatedly crossed back to mice from the same inbred strain. After many predefined backcrosses, usually 10 or more, in which the presence of the mutated allele is checked every generation, the background is said to be homogeneous. A comparison between the knockout and the inbred strain will then yield information on the effect of the knocked out gene on a specific genetic background. However, it is certainly possible that an inactivated gene

Table II
Genes or Gene Variants Found to Affect Male Offensive Aggression in Knockout Studies

System	Gene	Symbol	Effect[a]
Steroid	Aromatase	Cyp19	–
Metabolism	Estrogen Receptor 1 (alpha)	Esr1	–
	Estrogen Receptor 2 (beta)	Esr2	+(=)
Neurotransmission	Adenosine A2a Receptor	Adora2a	+
	Adrenergic Receptor, alpha 2a	Adra2c	+
	Creatine Kinase, Brain	Ckb	+
	Catechol-O-Methyltransferase	Comt	+(=)
	Dopamine Receptor 2	Drd2	–
	Enkephalin	Penk1	+
	Glutamic Acid Decarboxylase 2	Gad2	–
	Histamine Receptor H1	Hrh1	–
	5HT1B Receptor	Htr1b	+
	5HT1A Receptor	Htr1a	–?
	Monoamine Oxidase A	Maoa	+
	Tachykinin 1	Tac1	–
	Discoidin Domain Receptor Family, Member 1	Ddr1	+
	Nitric Oxide Synthase 1, Neuronal	Nos1	+(=)
	Oxytocin	Oxt	+/–
	Arginine Vasopressin Receptor 1B	Avpr1b	–
	VGF Nerve Growth Gactor Inducible	Vgf	–
Signaling proteins	Calcium/Calmodulin-dependent Protein Kinase II Alpha	Camk2a	+/–
	Regulator of G-protein Signaling 2	Rgs2	–
	Breakpoint Cluster region Homolog	Bcr	+
Growth factors	Brain Derived Neurotrophic Factor	Bdnf	+
Transcription factor	ETS-domain Protein	Pe1	+
Development	Neural Cell Adhesion Molecule	Ncam	+
	Nuclear receptor Subfamily 2, Group E, Member 1	Nr2e1	+
	Dishevelled, Dsh Homolog (Drosophila)	Dv1	–
	Arg, Abelson-related Gene	Ab12	–
Immunological factors	Interleukin 6	I16	+
Miscellaneous	Nitric oxide synthase 3, Endothelial Cell	Nos3	–
	Gastric-Releasing Peptide Receptor	Grpr	+(=)
	Transformation Related Protein	Trp73	–

[a] +, knockout increases aggression; –, knockout decreases aggression; +(=), increase but depending on background/paradigm/test day; –?, possible decrease; +/–, increase/decrease depending on design/type of aggression

dramatically affects a trait on one background, whereas it has no or a different effect on another background. This phenomenon, in which gene(s) influences the effect of another (i.e., the background genes interact with the knockout gene), is called epistasis and has been found in aggression research. For instance, the effect of the Y chromosome on aggressive behavior depends on the genetic background to which it has been backcrossed. Third, traditional knockouts are constitutive: They lack the gene in every cell and tissue from conception on. This means that in practice one cannot study the effects of genes that affect complex traits that are also essential for normal development. Such knockouts simply die at or before birth.

Joe Tsien at Princeton University developed a method to overcome these problems. He encountered this problem when he knocked out various subunits of the NMDA receptor. This receptor is thought to increase the synaptic strength between two nerve cells, a process called long-term potentiation (LTP), which is fundamental for learning and memory. By coincidence, he engineered NMDA knockout mice that lacked the subunit in a specific section of their hippocampus termed the CA1 region, which appears to be essential for memory. Hence, these so-called conditional, regionally restricted knockouts lack an essential "memory" gene, but only in a specific part of the brain and nowhere else in the body. As expected, it appeared that these animals demonstrated not only decreased LTP but also poor spatial memory.

Genetic engineering can be used not only to knock out genes but also to insert extra copies of a gene. This method is called transgenic integration. One of the more convincing behavioral examples comes from the same laboratory that developed the conditional NMDA knockouts. Instead of inactivating a gene, they inserted an extra copy of another memory gene. This gene codes for an NMDA subunit called NR2B, which is more strongly expressed in young people and stays open longer than the "old people's" NR2A, a phenomenon that might explain the age-related differences in learning and memory. Indeed, transgenic mice that had an extra copy of the gene for this receptor learned certain tasks better than did normal mice.

The development of both knockout and transgenic integration techniques has certainly deepened our knowledge about the effects of specific genes on complex traits. However, in addition to the previously mentioned, more pragmatic problems (flanking gene effects, genetic background, and temporal and spatial limits), there is another, more theoretical pitfall. Fundamentally, two types of genes coexist in nature: polymorphic and monomorphic genes. When studying genes that in nature are monomorphic, we generally deal with underlying mechanisms common to most or even all members of a species. In contrast, when studying natural genetic variation, we investigate mechanisms underlying spontaneous individual differences (i.e., polymorphic genes). Analysis of this natural genetic variation may thus enable us to identify genes that modify behavioral and neural function to a degree that is not grossly disadvantageous to the individual that carries such alleles. In short, whereas one type of question addresses, for example, how animals store information, the other type of question asks why some individuals perform better in a given task than others. Artificially induced mutations (e.g., knockouts and transgenes) can be used to study both types of genes, but it should be realized that the results of knockout or transgenesis studies do not contribute to the explanation of naturally occurring interindividual variation in cases in which the genes investigated are monomorphic in nature. In fact, most null mutations are not found to occur spontaneously in natural populations.

VIII. GENE–ENVIRONMENT INTERACTIONS

In the previous sections, we have shown that individual differences in behavior can be partly explained by genetic variation. Obviously, genetic variation is not the only source of variation; differences in the environment also play an important role. This section focuses on the border of both sources of variation—the so-called gene–environment interaction. Generally, the term refers to the phenomenon that the behavioral expression of the genotype depends on the environment.

A recent example of a gene–environment interaction in aggression research is a study that compared the behavior of male mice in which the brain-specific isoform of the creatine kinase gene was knocked out with animals that did not lack the gene (wild types). Both knockouts and wild types were tested for their aggressive behavior in the neutral cage paradigm. In this test the encounters took place on neutral grounds;

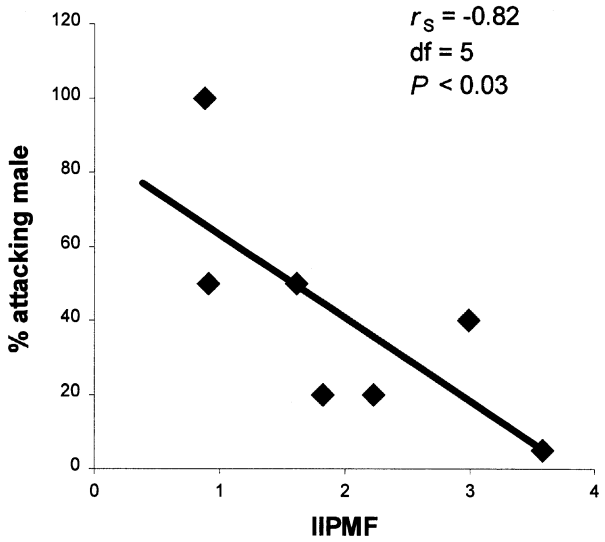


Figure 2 Relationship between the sizes of the hippocampal intra- and infrapyramidal mossy fiber (IIPMF) projections and attack behavior in seven inbred mouse strains. Data from Guillot *et al.* (1994).

that is, neither the animal under investigation nor its opponent were familiar with the test cage. Moreover, each mouse was tested twice on consecutive days against genetically different standard opponents. At first, knockouts did not seem to differ from wild types: Both were similarly aggressive against whichever opponent. A more detailed analysis, however, revealed that knockouts were more prone to attack one of the two opponents, whereas wild types showed no preferences.

Another example showing that the behavioral expression of the genotype depends on the environment comes from mice that lack a specific part of the NMDA receptor in a specific section of their hippocampus. As mentioned previously, when raised under normal laboratory conditions, they perform relatively poorly in learning and memory tasks. However, when they are exposed to an enriched environment each day for an extended period, they improve markedly and do as well as normal mice in various tasks. This behavioral enhancement is reflected anatomically: The number of connections between hippocampal cells actually increases. Hence, in these mice the enriched environment counterbalances a genetically engineered memory defect.

IX. GENETIC ANALYSIS OF BRAIN–BEHAVIOR RELATIONSHIPS

So far, we have not dealt with the intermediate neuronal structures through which genes modulate aggressive behavior. Results from knockout studies have certainly identified a wealth of brain receptors that might be involved in aggressive behavior. However, do they also represent naturally occurring variation in behavior and underlying neuronal structures?

One of the more eye-catching products of genetic analyses of brain and behavior comes from studies on the hippocampal intra- and infrapyramidal mossy fiber (IIPMF) terminal fields. Also, this research clearly demonstrates the importance and strength of classic behavioral genetical techniques.

The hippocampus has been shown by means of lesion studies to be involved in learning and memory. Subsequently, it was shown that the performance of mice in learning tasks dependent on hippocampal integrity is correlated with heritable variations in the size of the IIPMF. Animals with larger IIPMF projections perform better on spatial learning tasks. However, several mossy fiber studies also indicate hippocampal involvement in the regulation of intermale aggression.

The first indication for the involvement of genetics and the IIPMF sizes in the development of aggression came from the two selection lines previously mentioned. (Aggressive) SAL males had shorter IIPMF terminal fields than (nonaggressive) LAL males. Soon after, a study that compared multiple inbred strains for both the sizes of the IIPMF terminal fields and aggressive behavior clearly demonstrated that the SAL–LAL difference was no coincidence. In fact, a strong negative genetic correlation was observed between these two variables: The shorter the IIPMF sizes, the more aggressive the behavior (Fig. 2). These findings strongly suggest that hereditary neuroanatomical variations in a defined brain structure, the hippocampus, may influence intermale attack behavior.

Acknowledgments

FS was supported by PULS Grant 48.001 from the Earth and Life Sciences Foundation, which is subsidized by The Netherlands Organization for Scientific Research. WEC was supported by the Centre National de la Recherche Scientifique (Grant UPR 9074).

See Also the Following Articles

AGGRESSION • BEHAVIORAL NEUROIMMUNOLOGY • BEHAVIORAL PHARMACOLOGY • EVOLUTION OF THE BRAIN • NEUROBEHAVIORAL TOXICOLOGY • PSYCHONEUROENDOCRINOLOGY • STRESS: HORMONAL AND NEURAL ASPECTS

Suggested Reading

Bock, G. R., and Goode, J. A. (Eds.) (1996). *The Genetics of Criminal and Antisocial Behaviour*. Wiley, Chichester, UK.

Crusio, W. E. (Ed.) (1996). The neurobehavioral genetics of aggression [Special issue]. *Behav. Genet*. **26,** 459.

Crusio, W. E., and Gerlai, R. (Eds.) (1999). *Handbook of Molecular–Genetic Techniques for Brain and Behavior Research*. Elsevier, Amsterdam.

Plomin, R., DeFries, J. C., McCleann, G. E., and McGuffin, P. (2001). *Behavioural Genetics*, 4th Ed. Worth Publishers, New York.

Stoff, D. M., and Cairns, R. B. (Eds.) (1996). *Aggression and Violence: Genetic, Neurobiological, and Biosocial Perspectives.* Erlbaum, Mahwah, NJ.

van Abeelen, J. H. F. (Ed.) (1974). *The Genetics of Behaviour*. North-Holland, Amsterdam.

Behavioral Neuroimmunology

DOUGLAS L. DELAHANTY and JULIE K. CREMEANS-SMITH
Kent State University

GLOSSARY

cytokines Protein messenger molecules secreted by cells of the immune system that serve to communicate between immune cells and signal the presence of foreign material within the body.

lymphocyte proliferation assay An assay conducted to examine the ability of lymphocytes to mount a response against one of a number of standard mitogens.

mitogen A plant-like substance that elicits a cascade of immune responses, including the division of lymphocytes.

stress A negative experience elicited by threat, harm, or demand.

Research in behavioral neuroimmunology examines the interaction of central nervous system-mediated behavior and the immune system, two areas that were once thought to operate relatively independently. However, recent interdisciplinary research has revealed that these components can affect and be affected by each other. Specifically, the primary focus of behavioral neuroimmunology research is on the bidirectional relationship between behavior and immunity, examining the impact of health-risk behaviors on the immune system, and the relatively newly discovered impact of immune activation on subsequent behavior.

I. OVERVIEW OF THE IMMUNE SYSTEM

Although a more detailed description of the immune system is provided in other articles of this encyclopedia, it is necessary to briefly review the basics of immunology. This is not meant to be a comprehensive review but, rather, is meant to provide the reader with sufficient background to interpret research in the field of behavioral neuroimmunology.

The general purpose of the immune system is to defend the body against infection and disease by identifying and eliminating foreign "nonself" pathogens and mutated "self" cells. This is accomplished through the activity of two general categories of immunity: innate (also called nonspecific or natural) and acquired (also called specific) immunity. Natural killer (NK) cells, macrophages, and neutrophils are agents of innate immunity. These cells serve a surveillance function and will attack pathogens or cancerous cells without specificity and without requiring prior exposure to the invader. In contrast, acquired immunity involves a specific cascade of reactions to a particular antigen, and the magnitude of response to this pathogen increases with each successive exposure. Acquired immune responses can be classified into two types: cell-mediated and humoral immunity. Cell-mediated immunity is carried out by T lymphocytes (which mature in the thymus), whereas humoral immunity is mediated by antibodies that are produced by B lymphocytes (which mature in the bone marrow in humans).

Perhaps the most efficient way of reviewing the acquired immune response is to trace how the immune system would respond to a given pathogen (Fig. 1). Upon initial infection, a macrophage or other antigen-presenting cell (APC) encounters the pathogen, engulfs it, and digests it. The macrophage then displays

the foreign antigen on its surface so that it can be recognized by antigen-specific T cells. In addition, the macrophage secretes cytokines or chemical messengers such as interleukin-1 (IL-1) that stimulate helper T ($CD4^+$) cells to begin proliferating. Upon encountering the APC, helper T cells secrete IL-2 and interferon (IFN), which stimulate cytotoxic T cells ($CD8^+$) to destroy infected cells. In addition, helper T cells secrete IL-4, which activates the humoral branch of the immune response. IL-4 stimulates B cells either to mature into plasma cells or to become B memory cells. Plasma cells produce large quantities of antibody specific to the antigen that serve to destroy it. B memory cells "remember" the particular antigen so that, upon subsequent reexposure, the speed and magnitude of the immune response will be increased.

II. CONDITIONED IMMUNOSUPPRESSION

Although early research suggested that allergic reactions could be provoked with artificial roses, and hay fever attacks could similarly be stimulated by exposure to pictures of hay fields, Ader and Cohen's (1975) demonstration that immunosuppression could be conditioned to a novel stimulus provided the first empirical investigation of the relationship between behavior and the immune system. In their landmark study, Ader and Cohen paired the novel taste of saccharin with the immunosuppressive agent cyclophosphamide in a taste aversion paradigm. Subsequent presentation of saccharin resulted in an attenuated immune response to sheep red blood cells. These results suggested that the immune system should be viewed as an integrated physiological system subject to the influence of psychosocial and environmental factors and not, as previously thought, an autonomous system operating relatively independently of other physiological processes. These findings were replicated and extended in both human and animal studies and led to the development of the field of psychoneuroimmunology. Subsequent research has provided extensive evidence for the relationship between stress and immune alterations and has examined possible mechanisms through which psychosocial factors could lead to altered immune activity. In addition, moderators of the stress–immune system relationship and the role of stress-induced immunosuppression in various disease processes have been studied.

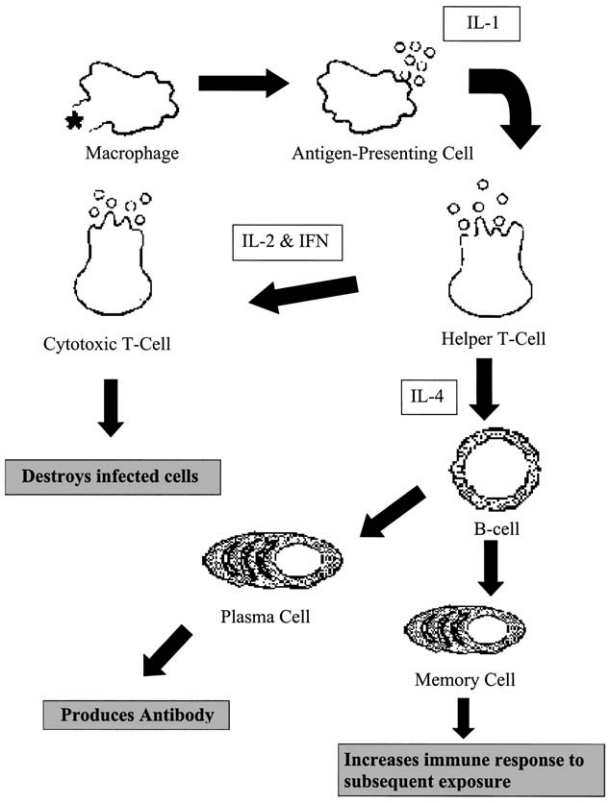


Figure 1 The cascade of responses by the acquired immune system to stimulation by a pathogen. IFN, interferon; IL, interleukin.

III. STRESS AND IMMUNITY

The demonstration of links between the brain and immune system suggested that emotional states such as stress could affect immunity, leading to hundreds of studies that have measured the effects of stress on humans and animals. Most of these studies have examined the impact of acute laboratory or more chronic naturalistic stress on various enumerative and functional indices of immune activity.

A. Acute Laboratory Stressors

1. Animal Research

Animal researchers have typically defined acute stressors as being relatively short in duration or consisting of relatively few exposures to a stressful stimulus. This general definition allows for numerous types of stressors differing in intensity, duration, and effects. Examples of stimuli that have been used to study the effects of acute stress on animals include aggressive confrontations, noise or bright light exposure, passive avoidance tasks, and tail shocks. Early animal studies reported decreased T cell proliferation during and following acute stress exposures concomitant with increases in levels of circulating corticosterone. Subsequent research determined that stress-related

increases in corticosterone provided a mechanism through which stress could lead to suppressed immunity. However, recent research has clouded these conclusions by demonstrating that acute stressors do not always result in immunosuppression. By varying the type, severity, and duration of exposure as well as the time at which immune measures are sampled, acute stressors can result in suppressed, unaltered, or even enhanced immune system activity.

2. Human Research

Acute stress in the human literature is often defined as transient exposure to a variety of distressing or challenging laboratory tasks (e.g., mental arithmetic, speech tasks, and distressing films) or as a short-duration naturalistic event (e.g., first-time tandem parachute jumping). Despite differences in type and duration of stressors used and sampling time during or following stress exposure, research examining the immunomodulatory effects of acute stressors suggests that numbers of peripheral lymphocytes increase during and following acute stress, whereas lymphocyte proliferation to assorted mitogens is typically suppressed. Increased numbers of lymphocytes are thought to be due to a redistribution of cells from lymphoid organs into the peripheral blood. However, these cells are often immature and not fully capable of combating pathogens, thus leading to suppressed proliferation. In addition, acute stress may lead to decreased immune activity by downregulating the activity of existing cells.

Whereas lymphocyte proliferation consistently appears to be suppressed during stress, research examining the effects of acute stress on NK cell activity has yielded more inconsistent results. Some researchers have reported increased activity of NK cells during and following acute stress, whereas others have found decreased activity. Closer examination of the timing of blood samples in these studies suggests that those examining NK activity during or immediately after stress exposure found increases in activity, whereas those examining later time points found decreases in activity. In other words, NK activity may follow a biphasic pattern in response to acute stress, with initial increases in activity being followed by a subsequent decrease to levels below baseline. Increased innate immunity during acute stress may allow an individual to fight off opportunistic agents without having to mobilize the entire cascade of the acquired immune response. Simple opponent processes have been hypothesized to account for the subsequent decrease below baseline.

In summary, results of studies examining the impact of acute laboratory stressors on immunity suggest that acute stress is associated with increases in number but decreases in the functional efficacy of circulating lymphocytes. In addition, NK cell activity may follow a biphasic pattern of initial increases in activity followed by a decrease to below baseline levels upon stressor termination.

B. Naturalistic Examination Stress

Laboratory studies of stress, while providing standardized sampling times and controlling for numerous extraneous factors, suffer from questions of generalizability. Whether responses to laboratory stressors parallel those to naturally occurring acute stressors is questionable, and the extent to which transient changes in immune activity observed in the lab are clinically relevant is unknown and thought to be minimal.

To improve upon some of these shortcomings, Kiecolt-Glaser and Glaser conducted a comprehensive series of studies in which they examined the immunomodulatory effects of a commonplace, everyday stressor: academic examinations. In their standard protocol, medical students were examined at two time points during their first 2 years of medical school; first during a lower stress baseline period in which exams were not being given and then a month later during examinations. Results revealed that exam stress was associated with decreased NK activity, decreased lymphocyte proliferation to mitogens, and decreased production of various cytokines. In addition, exam periods were associated with greater self-reports of illness (primarily upper respiratory infections) suggesting that the observed immune changes may have health consequences. Students' levels of plasma albumin and transferrin fell within normal ranges during the exams, suggesting that immune changes were not due to poor nutrition during the examination period. Subsequent studies revealed that the immunosuppressive effects of exam stress altered the competence of the cellular immune system to control latent herpesviruses and influenced students' responses to a series of hepatitis B inoculations. Less stressed students were more likely than more stressed students to seroconvert after the first inoculation, providing more evidence for health consequences of exam-related immune alterations. Because examination stress was shown to reliably decrease immune efficacy, an intervention was conducted to determine if decreasing stress through hypnosis and relaxation could attenuate these immunosuppressive effects. Although NK activity and

the percentage of T lymphocytes decreased during exam time in both the intervention and the control groups, more frequent practice of the relaxation exercise was associated with higher T cell percentages within the intervention group. This suggested that a stress-reducing intervention could, in part, moderate the immunosuppressive effects of exam stress.

In summary, the studies of Kiecolt-Glaser and Glaser demonstrated that relatively mild, everyday stressors have consistent immunosuppressive effects. It is important to note that the participants in these studies were medical students who had been taking (and typically excelling on) examinations for the majority of their lives, so this stress was not beyond the realm of their normal experience. Furthermore, these studies suggested that the transient changes in immunity exhibited during exam periods could have possible health implications.

C. Chronic Stress and Depression

1. Animal Research

One of the major criticisms of research conducted on the effects of chronic stress on animals is that the stressor is not qualitatively different from that used in acute conditions. That is, acute stress is often differentiated from chronic stress only quantitatively, by the number or duration of exposures to the stimulus. Despite these problems in operationalizing chronic stress, animal research has suggested that the immune system may be differentially affected by acute and chronic stress situations. Studies have shown that exposures to sessions of acute noise stress for less than 1 week are associated with increased NK activity, whereas exposure for 3 weeks or more is associated with decreased NK activity.

Animal studies have also examined the effects of chronic social stress on immunity. In these studies, subjects were exposed to numerous changes of social environment, resulting in struggles to establish dominance. This stressor paradigm is thought to more readily approximate the human chronic stress experience. Results of studies using social hierarchy disruption as a stressor in rats demonstrate that submissive animals have lower immune responses on some immune measures but not others. Studies of social stress in nonhuman primates may more adequately examine the processes underlying human chronic stress, but these studies have also produced mixed results. Some have reported suppressed T cell function in response to chronic social stress, whereas others have found no significant immunological effects. This may be due, in part, to the small sample size examined in primate research and the consequent lack of power to detect significant results. Despite some exceptions, for the most part animal research suggests that prolonged exposure to stressful stimuli or situations results in immunosuppression.

2. Human Research

The results of studies examining immune consequences of chronic stress in humans largely parallel the findings in animals. Research across a range of chronic stress situations, including bereavement, caregiving for a relative with dementia, unemployment, and diagnosis with a life-threatening illness, reveals that stressed individuals have altered immune function in comparison with controls. In contrast to the acute stress literature, chronic stress is typically, although not always, associated with lower numbers of immune cells and weaker immune functioning. In particular, chronic stress associated with the loss or disruption of personal relationships has been shown to reliably alter immune activity. For instance, women who have been separated or divorced for less than 1 year have demonstrated poorer immune function than married women, and both men and women who reported poor marital quality had lower immune function than more happily married individuals. Research has also suggested that recently widowed women whose husbands died of lung cancer had more impaired immune function than women whose husbands were dying of lung cancer and women with healthy husbands.

The effects of losing a loved one on immunity have also been examined prospectively. Men whose wives died of breast cancer had significantly lower lymphocyte proliferation after the death of their spouse than prior to bereavement. Alterations of immune activity in chronically stressed individuals may also have important health consequences. For example, Alzheimer's caregivers have exhibited significantly impaired antibody and T cell responses to an influenza virus vaccine, and poor response to the vaccine has been associated with increased risk of influenza infection and higher rates of clinical illness. Similarly, caregivers have also been shown to take significantly longer to heal from a cutaneous punch biopsy wound than a matched control group.

Chronically stressed individuals exhibit a greatly increased incidence of depression, and the relationship between depression and immunity has also been examined. However, results have been mixed, perhaps due to differing diagnoses and medical treatment protocols among patient samples and to different immune measures being assayed. Depression does not

appear to be consistently associated with alterations in particular lymphocyte subsets, although overall a higher number of leukocytes have been reported in depressed patients compared to controls. Studies of lymphocyte proliferation to mitogens in depressed patients have been particularly mixed, resulting in findings of higher, lower, or no differences in activity between depressed and age- and gender-matched nondepressed patients. Decreased NK activity appears to be the most consistent immunological change linked with depression. In addition, depressed individuals taking selective serotonin reuptake inhibitors demonstrate significant increases in NK activity (NK activity had been significantly lower than that of controls at the start of the experiment) which coincide with alleviation of the depressive symptoms. Although depression appears to be linked with suppression of some immune indices, many variables have been shown to affect the relationship between immune and depression, including age of the patient, severity of symptomatology, hospitalization status, and gender. Therefore, the complex nature of the relationship between depression and immunity makes it inappropriate to conclude that depression, in general, is associated with specific immune consequences without examining the effects of these moderating variables.

D. Traumatic Stress and Posttraumatic Stress Disorder

Although many studies have examined the effects of acute, laboratory stress and chronic, naturalistic stress on immunity, relatively few have studied immune alterations in response to severe, traumatic stressors. Stressors can vary along many dimensions, including the duration of the event, the duration of threat experienced, and the duration of responding to the stressor. Therefore, distinguishing between acute and chronic stressors may be too simplistic and may not encompass the full range of stressful experiences. Sudden traumatic stressors involving life threat and direct bodily harm may be qualitatively different from other types of stress. Often, the traumatic event is acute in duration, but in some individuals distress may persist long after the event is over. A small but significant percentage of victims may develop posttraumatic stress disorder (PTSD), which is characterized by persistent reliving of the event, avoidance of situations that remind one of the event, and hyperarousal in response to reminiscent stimuli.

Evidence suggests that acute-phase responding to naturalistic and human-caused disasters is associated with increases in some immune measures, but the lack of published studies examining immune levels immediately after traumatic events necessitates viewing this conclusion with caution. Longer term follow-up of victims indicates that stress stemming from both human-caused and natural disasters appears to be associated with fewer cells and decreased functional efficacy of a wide range of immune measures. Immune system changes associated with human-caused disasters appear to be particularly persistent. Altered immune activity was found as many as 6 years after a nuclear accident in residents of Three Mile Island. Residents had greater numbers of neutrophils, but fewer B lymphocytes, cytotoxic T lymphocytes, and NK cells than did controls. In addition, residents displayed higher antibody titers to herpes simplex virus than controls, suggesting downregulation of cellular immunocompetence. The longest follow-up of a naturalistic disaster indicated that hurricane victims had significantly lower NK activity than laboratory controls 2 years after the hurricane, but that NK activity did not differ between the two groups 4 years post-hurricane.

Although traumatic stress appears to be immunosuppressive in the majority of trauma victims, findings concerning immune system activity in victims who meet criteria for PTSD are more mixed. Despite equivocal findings, research suggests that war veterans with PTSD may exhibit greater immune activity than similarly exposed veterans who do not meet PTSD criteria. However, these findings must be interpreted cautiously due to several characteristics of the patient sample. For instance, the length of time that has passed between the traumatic event and sampling for immune measures in these studies is often great. PTSD patients in these studies have typically suffered from chronic PTSD, often presenting with symptoms for more than 20 years. This chronicity of symptomatology is unusual and may account for some of the findings. These patients also often suffer from comorbid drug and alcohol abuse, making conclusions regarding immune alterations in PTSD difficult.

One mechanism through which PTSD could lead to increases in immune activity has been examined and provides suggestive results. The majority of studies have found lower 24-h urinary cortisol excretion in patients with PTSD compared to patients without PTSD and normal controls. Cortisol exerts immunosuppressive effects, so persistently low levels could be associated with increased immune activity. However, future research examining immune levels in less chronic PTSD without comorbid diagnoses is necessary before conclusions can be drawn with any certainty about the relationship between PTSD and immunity.

E. Summary

In general, research has produced mixed results with regard to the effects of stress on immune activity. This may be due, in large part, to differences in methodology, subject samples, and especially immune components being examined. The immune system is an extremely complex physiological system with numerous cellular and chemical components that may respond differently to stressful stimuli. Therefore, it is necessary in future research to examine multiple measures of immunity rather than focusing on one particular cell or cytokine. Overall, acute and chronic stressors appear to be associated with immunosuppression, although acute stress has been linked to increases in the number of circulating lymphocytes and increases in NK activity. With regard to traumatic stress, very little research has been conducted examining immune changes during acute phases of responding to a traumatic event. Long-term immune changes appear to parallel the immunosuppression seen in chronic stress and depression; however, research has suggested that chronic PTSD may be associated with immunoenhancement.

IV. MECHANISMS OF STRESS-RELATED IMMUNE ALTERATIONS

The finding that psychosocial stressors can evoke consistent changes in immune measures led researchers to begin searching for the mechanisms through which emotional states could impact immune levels. Results of these studies suggested that stress has both direct and indirect effects on immunity. Direct effects stem from the activation of the two primary stress pathways: the sympathetic nervous system (SNS) and the hypothalamic–pituitary–adrenal (HPA) axis. SNS activation is accompanied by increases in levels of catecholamines in the bloodstream, whereas HPA activation results in increased release of glucocorticoids. These hormones appear to participate in regulating the immune system both through direct neural innervation of lymphoid tissue and through their activity at relevant receptors on lymphocytes. As mentioned previously, cortisol often acts as an immunosuppressive agent, and animal research has suggested that genetically low glucocorticoid responders are more susceptible to autoimmune disorders, whereas high glucocorticoid responders appear to be at increased risk for infectious disease. On the other hand, catecholamine increases are associated with a transient lymphocytosis or an increase in the number of circulating lymphocytes, especially NK cells. This increase in the number of circulating cells appears to be accompanied by a reduction in immunological functional capabilities. Therefore, activation of stress-related endocrine pathways provides a direct route by which stress may induce immunosuppression.

Stress can also affect immunity indirectly by leading to negative health behaviors that can contribute to alterations in immune functioning. Stress is typically associated with riskier health behaviors, including increased drug/alcohol use, increased smoking, poorer sleep, and poorer nutrition. These behaviors can also negatively affect the immune system, although the relationship between individual health risk behaviors and immunity is often not clear. Partial sleep deprivation has been shown to decrease immune functioning in humans and rats, and poor sleep has been hypothesized as one mechanism through which repeated, intrusive thoughts may impact immunity. The immunological effects of cigarette smoking, on the other hand, have been particularly mixed, with research reporting decreases, increases, and no change in immune functioning. Some of this variability may be due to differences in the amount of smoking, acute versus chronic effects of smoking, and the timing of sampling in relation to when the participant last smoked. Many other factors can affect the relationship between smoking and immunity, and recent evidence suggests that cigarette smoking may interact with depressed mood such that depressed smokers exhibit the greatest alterations in immunity.

Use of alcohol also increases during periods of stress, and results of studies examining the immunological effects of chronic alcoholism suggest that immune activity is suppressed while individuals are using alcohol but recovers within months after participants stop drinking. These results must be interpreted cautiously because alcoholism is associated with many other factors that may affect immunity (poor diet, less exercise, etc.). Recent research suggests that minimal intake (one or two drinks per day) may provide some protection against certain chronic diseases. However, chronic and moderate acute alcohol use can increase susceptibility to bacterial and viral infections, although the effects of alcohol on immunity appear to be differentially affected by alcohol dose, extent of alcohol use, and time since use.

V. MODERATORS OF THE STRESS–IMMUNE LINK

Although many situational and personal variables have been shown to affect the relationship between stress and immunity in individual studies, only a few

have been shown to have consistent effects across studies and research groups. These include perceived control over the stressor, the presence of or perceived social support, and personality variables such as optimism.

A. Perceived Control

Perceived control over illness has been shown to be positively associated with better psychological adjustment in many studies of different chronic illnesses, with the strongest relationships being found in patients with more severe disease. In addition, the effects of stressor controllability have been widely researched in animal models. Animals exposed to uncontrollable stress typically demonstrate decreased immune responses and faster disease progression than yoked animals exposed to controllable stress. Human studies of immune alterations to uncontrollable stress have produced mixed results. Whereas some have found reduced lymphocyte proliferation to mitogens in participants exposed to controllable versus uncontrollable stress, the majority have found the opposite. It appears that objective control (actually having control over the stressful stimulus) is not as important as *perceived* control in modulating immune responses. Research examining the impact of perceived controllability over both acute and chronic stressors has found decreased functional and enumerative immune measures in participants perceiving less control over the stressor. These findings persist whether the stressor is actually controllable or not.

B. Social Support

As mentioned earlier, research examining the relationship between quality of personal relationships and immune function has demonstrated consistent immunosuppressive effects in individuals who reported low marital satisfaction or who recently lost a spouse. Similarly, social support has been linked with physiological processes associated with risk for cardiovascular disease and cancer, and less social integration is associated with higher mortality rates from all causes. Individuals with high levels of social support have lower baseline blood pressures than those with lower levels of support, and the presence of a supportive individual has been shown to decrease cardiovascular reactivity during an acute laboratory stressor.

Individuals reporting greater diversity of social support also appear less likely to develop the common cold upon exposure to a virus than those with a less diverse social support network. Consistent with these findings, social support has been shown to moderate the effects of stress on immunity. For the most part, greater social support is related to stronger immune responses. High levels of social support in spouses of cancer patients were associated with greater NK cytotoxicity and lymphocyte proliferation responses. Similarly, levels of social support have been positively related to NK cell activity, and NK activity has been linked to measures of disease progression in breast cancer patients. These results suggest that social support may buffer the immunosuppressive effects of stress, and that the attenuation of these effects may have relevant health consequences.

C. Optimism

Personality variables also appear to moderate the relationship between stress and immunity. Research has suggested that optimism, or the generalized tendency to "look on the bright side of things," affects the way in which individuals perceive stressful situations and is associated with better health and positive health outcomes. Optimism has been linked to better physical health, better quality of life following coronary artery surgery, decreased stress in breast cancer patients, and later symptom onset and longer survival time in AIDS patients. An optimistic outlook has also been shown to buffer the immunosuppressive effects of stress.

Situational optimism has been associated with higher immune activity in first-year law students, although many factors can affect this relationship because optimism has been shown to interact with mood and perceived stress to account for immune changes. Characteristics of the stressor may also determine the efficacy of optimism in attenuating stress-related immune changes. Optimism has been shown to buffer the effects of acute stress on immunity, but in response to high levels of persistent distress, optimists have demonstrated more immunosuppression than pessimists. Although optimism may be beneficial for dealing with transient stressors, persistence of distress is inconsistent with the expectations of an optimistic individual and optimists may therefore face difficulties in coping with persistent distress.

D. Summary

As demonstrated in the previous sections, the relationship between stress and immune changes is complex. It is difficult to interpret mixed or contradictory findings

due to the use of different types of stressors and different immune measures. Overall, stress appears to have an immunosuppressive effect. However, many personal and situational variables have been shown to moderate the relationship between stress and immunity, further complicating interpretation of the research. Perceived control, optimism, and social support all appear to be associated with buffering the effects of stress on immunity, although these variables can interact with each other and other variables to affect this relationship. In addition, although stress has consistently been associated with alterations in immunity, observed immune levels often remained in the normal range, leading many researchers to question the meaningfulness and impact of stress-related immune changes.

VI. STRESS-INDUCED IMMUNOSUPPRESSION AND HEALTH

Stress has repeatedly been linked to poorer health, greater incidence of the common cold and upper respiratory infections, faster progression of chronic diseases such as cancer and AIDS, and exacerbation of autoimmune disorders. Despite these consistent findings of stress-related impairments in health, relatively few studies have demonstrated that stress-related immunosuppression could serve as a mechanism through which stress could lead to disease. Although the extent to which stress-related immunosuppression is clinically relevant is debated, recent research has attempted to prospectively show relationships among stress, immunity, and health.

A. Cancer

Research on the psychoimmunology of cancer presumes that stress may alter the immune system, leading to faster growth and progression of neoplastic disease. However, this presumption is very controversial and the extent to which stress-related changes in immunity result in clinically significant development or growth of malignant tissue is unknown and often considered minimal. The literature is further complicated by contradictory findings in animal models of stress and cancer. In animals, stress has been shown to result in both exacerbation and inhibition of tumor growth and metastases, with differences in results due to timing of tumor induction, type of stressor examined, type of tumor, and species and prior experience of animals examined.

Human research examining the relationship between stress and cancer has also produced mixed results. For example, although initial studies suggested that stress could lead to the development of cancer, contradictory findings also exist. Furthermore, due to methodological limitations and questions concerning the events and timing of initial tumor development, many researchers have concluded that there is no demonstrable relation between stress and cancer onset. However, research examining the effects of stress on progression of cancer is more suggestive, and the immune system has been examined as a mediator of the relationship between stress and cancer progression. Current knowledge regarding the role of the immune system in surveillance against tumors is limited, but the majority of studies have focused on NK cells due to their antitumor properties. As mentioned earlier, NK cells also appear to be very responsive to psychosocial stress, and stress-related alterations in NK activity have been widely reported. Examination of levels of NK activity in breast cancer patients has revealed that measures of distress in patients recently diagnosed with cancer predict current and subsequent NK activity. In addition, NK activity appears to be an important predictor of disease status in breast cancer patients.

Another way to examine the relationship between stress, immunity, and cancer is to reduce stress in some cancer patients and determine any consequent effects on immunity or disease progression. Stress-reducing interventions have been shown to increase vigor and the use of more effective coping strategies in malignant melanoma patients. In addition, at a 6-month follow-up, intervention patients had higher percentages of NK cells and greater NK activity than patients who did not receive the intervention. Six years after the intervention, intervention participants were more likely to have survived than control participants, although NK activity levels were not related to survival.

B. Wound Healing

Perhaps the strongest, albeit indirect, evidence for clinical health consequences of stress-induced immunosuppression derives from studies examining differences in stressed and nonstressed individuals in length of time to heal from standardized wounds. In these studies, women who were caring for relatives with dementia and age- and income-matched controls received a standard 3.5-mm punch biopsy wound resulting in identical tissue damage to all participants. Results revealed that the chronically stressed care-

givers took significantly longer to heal than controls. In addition, stimulated peripheral leukocytes from caregivers produced significantly less IL-1β (a cytokine involved in the initial stages of wound healing) than did leukocytes from controls. Therefore, chronic stress appears to slow wound healing and affect a cytokine that serves multiple functions in early stage wound repair. Subsequent research examined a less chronically stressed sample and found that it took 3 days longer on average for dental students to heal from an oral biopsy wound during examinations than during summer vacation. In addition, IL-1β production was significantly lower during the exam period. These results demonstrate that relatively mild, everyday stressors can also have effects on immune activity and healing. However, neither study examined the relationship between IL-1β levels and healing, making conclusions concerning the role of stress-induced immunosuppression in slowed healing difficult.

C. Summary

Overall, the inability to demonstrate clinical effects of stress-induced alterations in immune activity on health appears to be the major limitation of research in behavioral neuroimmunology. Research has consistently demonstrated stress effects on health and stress effects on immunity; however, immune changes have not been shown to mediate the effects of stress on health. This may be due to methodological difficulties. As mentioned previously, the immune system is very complex and its components may respond differently to stress. Perhaps measuring activity or the number of one or a few immune cells is not a realistic representation of *in vivo* immune responses to stress or illness. Recent results showing stress-induced immune alterations and poorer health are suggestive, but research examining multiple immune indices and testing mediational models is necessary before we can conclude with any certainty that immune changes in response to stress play a role in disease progression.

VII. IMMUNE ACTIVATION AND SICKNESS BEHAVIOR

Although the majority of research in behavioral neuroimmunology has examined the impact of stress and central nervous system (CNS) activation on immune activity, it has become apparent that the immune system can also directly communicate with the CNS and thus alter behavior. Therefore, it is necessary to consider the relationship between the immune system and the brain as bidirectional. Furthermore, recent research has suggested that the immune system might often initiate this bidirectional communication by acting as a sensory organ, informing the brain of infection, and triggering behaviors designed to combat the infection (Fig. 2).

A. Sickness Behavior

The primary illustration of the immune system's, influence on the CNS involves a collection of responses to infection termed sickness behavior. In the early 1960s, researchers examining the activity of various mitogens determined that simple activation of the immune system resulted in a collection of behaviors

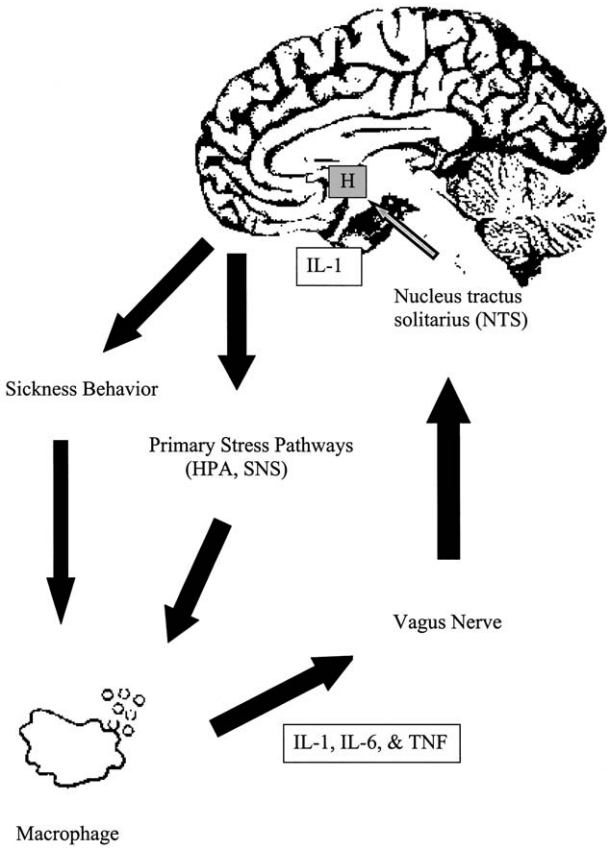


Figure 2 Bidirectional communication between the immune system and the brain begins with the activation of macrophages by foreign particles. Activated macrophages secrete various cytokines that stimulate the vagus nerve. The vagus then sends a neural signal to several brain regions, including the nucleus tractus solitarius (NTS) and hypothalamus (H), whose activation leads to symptoms of illness such as fever and loss of appetite. In addition, the activated immune system also triggers activation of the hypothalamo–pituitary–adrenal (HPA) axis and the sympathetic nervous system (SNS), which results in immunosuppression. IL, interleukin; TNF, tumor necrosis factor.

normally seen in sick animals. In these studies, animals were typically injected with lipopolysaccharide (LPS), a fragment of the cell wall of gram-negative bacteria such as *Escherichia coli* bacteria. LPS does not cause infection in the animal but, rather, triggers the typical cascade of immune responses to infection. Although the animals were not injected with any disease-causing agent, they still exhibited reduced food consumption and decreased levels of general activity, suggesting that these sickness behaviors were triggered solely in response to immune system activation. However, the mechanism through which an exogenously administered mitogen could trigger this type of response was unknown.

A hint of a possible mechanism was provided by early attempts to treat cancer patients with various cytokine therapies. Cancer researchers searching for therapeutic agents that could elevate immune responses and help to combat the disease examined the efficacy of a variety of cytokines [including IL-1, IL-4, IL-6, IL-12, and tumor necrosis factor (TNF)] in slowing cancer progression. Clinical trials conducted with both healthy and ill individuals demonstrated serious side effects of cytokine administration, such as fever, anorexia, general malaise or depression, irritability, and delirium. Again, these symptoms parallel those seen in illness and suggest that cytokine activation may serve a primary role in initiating these sickness behaviors.

Subsequent research supported this hypothesis and demonstrated that sickness behavior results from an initial cytokine cascade triggered largely by the activation of macrophages. As mentioned earlier, macrophages are cells of the innate immune system which engulf foreign particles that have entered the body and release cytokines upon activation. Cytokines relay the message of immune system activation to the CNS, which evokes a number of changes in behavior designed to mobilize the body's resources to combat infection. These behavioral changes are typically referred to as sickness behavior and consist of depressed levels of general activity, decreased bodily care, reduced consumption of food and water, temperature change (usually fever), and increased sleeping.

The symptoms of sickness behavior are the result of the acute-phase immune response to bacteria, viruses, or other stimuli and thus remain constant despite differences in the particular invading pathogen. The primary function of sickness behavior is to conserve energy by limiting expenditure so that energy reserves may be put into generating fever. Fever is an adaptive symptom because it slows the growth of many pathogens and increases the activity of both nonspecifc and specific immune system processes. Because fever is the result of an increase in the hypothalamic set point for body temperature, the organism is motivated to seek out warm places. Other symptoms of sickness behavior, such as increased sleep and decreased levels of general activity, are designed to conserve energy. In addition, a sick individual also experiences anorexia and adipsia, which may appear counterintuitive, because the individual should be motivated to consume in an effort to increase energy reserves that may be used for fever generation. However, an organism that is foraging for food and water is exposed to predation and is exposing body surfaces to the elements, resulting in a loss of heat that has already been generated. The reduction of foraging efforts during sickness allows the organism to conserve energy by reducing activity and complements other symptoms, such as increased amounts of sleep. Anorexia is also beneficial by reducing iron levels in the blood, which can slow pathogen replication. Although sickness behavior is an unpleasant state for an organism, this pattern of behavior allows for more efficient destruction of pathogens and more speedy recovery. Therefore, sickness behavior represents an evolved strategy designed to limit infection, conserve energy, and promote recovery of the organism.

The primary cytokines that have been implicated in the production of sickness behavior are IL-1α/β and TNF-α. Many other cytokines have been shown to play a role in sickness behavior, but IL-1 has been the most studied due to its discovery as the first endogenous pyrogen. In addition to stimulating fever, IL-1 appears to be primarily involved in producing the reduction in social activities that occur during sickness. However, although individual cytokines appear to serve specific roles in sickness behavior, there is much overlap in their activities, and cytokines have been shown to compensate for each other in triggering certain behaviors. For instance, research has demonstrated that intracerebroventricular injection of an IL-1 receptor antagonist did not alleviate the effects of intraperitoneal (ip) LPS on social behavior. This suggests that other cytokines may have compensated for the absence of IL-1 effects by producing a reduction in social behavior, demonstrating the complex nature of cytokine involvement in triggering sickness behavior following immune system activation.

B. The Immune System as a Sensory Organ

The phenomenon of sickness behavior illustrates that the immune system may be considered as another

sensory organ of the body. Following this definition, the immune system's function is to detect "noncognitive stimuli," such as bacteria, viruses, and tumors, and to alert the CNS of their presence. The immune system has particular receptors (or, in this case, different types of cells) that detect specific stimuli, produce signals that can be understood by the CNS, and transmit messages to the brain. As a result of receiving this specific sensory input, the CNS is able to respond by inducing physiological and psychological changes in behavior designed to combat infection.

C. Evidence for Bidirectionality

Despite the fact that research has only recently begun to focus on the sensory capabilities of the immune system, much evidence has been collected that supports the existence of bidirectional communication. First, the immune system and the CNS make use of the same messenger molecules. There is a commonality in their production of and sensitivity to hormones and neurotransmitters. For example, although traditionally research has examined the impact of stress on levels of cytokines, IL-1 has been found to act directly on the HPA axis resulting in the production of corticotropin-releasing hormone (CRH), adrenocorticotropic hormone, and glucocorticoids. Therefore, although we have principally reviewed studies indicating that stress hormones can impact immunity, it appears that the immune system can also directly affect levels of stress hormones.

For the immune system to communicate directly with the CNS and lead to sickness behavior, neural connections must exist through which peripheral immune activity can signal the brain and inform it of infection. However, the pathway through which this is accomplished is still under debate. Although initially it was thought that cytokines merely traveled through the circulatory system into the brain, they are large hydrophilic proteins that are unlikely to be able to penetrate the blood–brain barrier (BBB). Rather, researchers favor the theory that peripheral cytokines stimulate the vagus nerve, which then sends a neural signal to several brain regions involved in producing symptoms of sickness behavior. Attention has focused on the vagus nerve as the primary pathway through which the immune system could stimulate the CNS because it serves as the main afferent route from the abdominal cavity to the brain and because it innervates tissue that plays an important role in immune activation, such as the liver, thymus, and gastrointestinal tract. Afferent vagal fibers have also been shown to terminate in the nucleus tractus solitarius (NTS), a brain region that is highly activated following peripheral immune system activation.

Further evidence for the role of the vagus nerve in relaying the message of immune activation to the brain has been provided by studies of subdiaphragmatic vagotomy. Animals that have had their vagus nerves severed display reduced sickness behavior symptoms in response to peripheral injections of LPS and IL-1 than do animals with intact vagus nerves. The reduction of sickness behavior following vagotomy has been found in several different animal models (including guinea pig, mouse, and rat), but only in response to peripheral immune activation triggered by ip injections of IL-1β or LPS. Vagotomy has no effect on sickness behavior triggered through central immune activation or from intravenous (iv) injections of immunostimulants. Because severing the vagus has no effect on iv injections of immunostimulants, cytokines must be able to relay the message of immune activation to the CNS through some pathway in addition to the vagus nerve. As mentioned previously, vagotomy decreases but does not eliminate the effects of peripheral immune activation on sickness behaviors. Therefore, investigators have examined possible secondary pathways through which the immune system could signal the brain. Research suggests that this secondary pathway may involve the direct passage of cytokines through areas in which the BBB is modified to allow greater contact with substances in the blood. The organum vasculosum laminae terminalis (OVLT) has been indicated as a possible site for this type of signaling. Future research will continue to investigate these and other pathways through which the immune system may communicate with the CNS to affect behavior.

Additional evidence supporting the role of the vagus as a major pathway between the peripheral immune system and the CNS stems from the fact that the brain regions in which the vagus terminates have been found to play an important role in the various symptoms of sickness behavior. The primary area of termination for the vagus nerve is within the NTS. The NTS and the area postrema are activated in response to ip injections of IL-1 and appear to be very sensitive to low levels of this cytokine. Areas within the NTS project to the paraventricular nucleus of the hypothalamus, which controls production and secretion of CRH, one of the hormones found to be produced in response to immune activation. The hypothalamus has been found to play a major role in the production of symptoms such as anorexia and fever. In terms of the specific symptom of reduced consumption, it has been demonstrated that IL-1β suppresses the neural activity of glucose-sensitive neurons in the lateral hypothalamus

and induces excitation in the ventromedial nucleus of the hypothalamus. These areas have been studied extensively in terms of the role that they play in initiating and terminating eating behavior, respectively. Therefore, the pattern of inhibition and excitation produced by IL-1β induces anorexia and fits the profile of sickness behavior. Fever, on the other hand, is thought to be the result of projections from the NTS to the preoptic area of the hypothalamus, where catecholamines trigger the synthesis and release of prostaglandins that act on temperature-sensitive neurons. As a result, the hypothalamic set point for temperature is raised, triggering the activation of other mechanisms, such as shivering and increasing metabolism, that are designed to defend the new set point. To date, the majority of research has focused only on the brain mechanisms involved in the production of anorexia and fever and has not addressed the other symptoms involved in sickness behavior.

D. Summary

Research has demonstrated that the connection between the CNS and the immune system is not unidirectional, as once thought. Rather, evidence suggests that the immune system is able to directly affect the CNS, perhaps most directly shown through the effects of immune activation on sickness behavior. Upon infection, the immune system sends a neural signal of activation to regions of the brain (particularly the hypothalamus) that govern the production of symptoms such as fever and anorexia. The vagus nerve is the most likely pathway through which peripheral immune system activation is relayed to the brain. However, studies of vagotomized animals suggest that secondary pathways exist through which cytokines can affect behavior. Most research has focused on weaker areas of the BBB, such as the OVLT, as the secondary pathway. Research examining the impact of immune activation on behavior is in its infancy and future studies will undoubtedly provide more information concerning the relationship between immunity and sickness behavior and the health implications of immune-mediated behavior change.

VIII. CONCLUSION

The research reviewed in this article indicates that much progress has been made in the field of behavioral neuroimmunology. However, although we currently have a greatly increased understanding of the bidirectional relationship between the CNS and the immune system, the implications of this association for health and disease have not been fully realized. Although many studies have demonstrated that the experience of stress is related to decreased immune functioning, the resulting health consequences of stress-induced immunosuppression have yet to be determined. Also, research on the effects of immune activation on the CNS and subsequent behavior has not fully examined potential health consequences. Therefore, although findings have begun to change the way we think about health and disease, the applied aspects of behavioral neuroimmunology are merely in their infancy. We are only now beginning to translate the meanings of bidirectional communication into investigations of the onset, progression, and outcomes of disease. Future research will continue to provide basic knowledge concerning the relationships between behavior and immunity, and applications of these findings will further our understanding of the vast and complicated interrelationships of health, immunity, and disease.

See Also the Following Articles

AUTOIMMUNE DISEASES • BEHAVIORAL NEUROGENETICS • BEHAVIORAL PHARMACOLOGY • NEUROBEHAVIORAL TOXICOLOGY • PSYCHONEUROENDOCRINOLOGY • PSYCHONEUROIMMUNOLOGY • STRESS: HORMONAL AND NEURAL ASPECTS

Suggested Reading

Ader, R., Felten, D. L., and Cohen, N. (Eds.) (2001). *Psychoneuroimmunology*, 3rd ed. Academic Press, San Diego.

Dantzer, R., Wollman, E. E., and Yirmiya, R. (Eds.) (1999). *Cytokines, Stress, and Depression*. Kluwer, New York.

Glaser, R., and Kiecolt-Glaser, J. K. (Eds.) (1994). *Handbook of Human Stress and Immunity*. Academic Press, San Diego.

Herbert, T. B., and Cohen, S. (1993a). Stress and immunity in humans: A meta-analytic review. *Psychol. Bull.* **55,** 364–379.

Herbert, T. B., and Cohen, S. (1993b). Depression and immunity: A meta-analytic review. *Psychol. Bull.* **113,** 472–486.

Kiecolt-Glaser, J. K. (1999). Stress, personal relationships, and immune function: Health implications. *Brain Behav. Immun.* **13,** 61–72.

Maier, S. F., and Watkins, L. R. (1998). Cytokines for psychologists: Implications of bidirectional immune-to-brain communication for understanding behavior, mood, and cognition. *Psychol. Rev.* **105,** 83–107.

Rothwell, N. J. (Ed.) (1996). *Cytokines in the Central Nervous System*. Chapman & Hall, New York.

Behavioral Pharmacology

PETER B. DEWS
New England Regional Primate Research Center

GLOSSARY

behavior The activities of a reasonably intact subject that affect its environment. "Reasonable" is not a precise specification, but activities of a few cells or even a length of intestine removed from the body would not be considered behavior of a subject. Activities of a subject from whom a length of intestine has been removed would be considered behavior of the subject.

drug A comprehensive term for chemical agents that affect physiological functions of a subject. The term is used most often for effects in animals, especially "higher" animals, including vertebrates, mammals, and humans. Because enough of any agent, even water, can perturb physiological function, it is useful to refer to an agent as a drug only if it has effects at a reasonable dose. ("Reasonable" is not a precise specification, but if more than a few grams per kilogram body weight of an agent is required for an effect it would probably not be called a drug.) Vitamins and other trace elements of the diet cause physiological effects, as do hormones such as epinephrine, insulin, and cortisol. It is simpler to accept them as drugs and to recognize their physiological role than to contrive a definition that excludes them. The term drug should not suggest only substances of abuse such as heroin and cocaine.

operant response (or often just "response") An item of behavior that, when related to a reinforcer, is increased or maintained in frequency in similar circumstances in the future.

receptor Receptors are the special part of a cell that binds a particular drug, with consequences to the function of the cell. Many receptors are molecular patches on the surface of cells. The existence of specialized receptors each selective for a limited class of drugs has been recognized by pharmacologists for more than 100 years, but it is only in the past few decades that the molecular structure of individual receptors has been able to be determined. Now, for increasingly more receptors, not only is their structure becoming known but also it is becoming known how they are changed when molecules of drug bind to them and how these changes initiate a cascade that modifies the function of the cell.

reinforcer An event occurring in relation to an operant response that increases or maintains the frequency of occurrence of the response in similar circumstances in the future.

Behavioral pharmacology is an experimental laboratory science measuring the effects of drugs on specific interchanges of a subject with its environment—that is, the effects of the drugs on aspects of the behavior of the subject. Behavioral pharmacology deals with physical as opposed to subjective phenomena and requires the same standards of evidence for conclusions as the rest of pharmacology. By usage, psychopharmacology has come to mean a largely clinical science, studying the effects of drugs in psychiatry. It seeks to determine the optimum regimes for drugs to maximize clinical benefits and minimize toxicity and side effects. New drugs are evaluated and psychopharmacology helps in the development of new agents and their introduction into clinical use. As with other primarily clinical subjects, including medicine and surgery, extension into experimental laboratory experiments is integral to progress in the subject. Other than anatomy, more has been learned about the human brain from studies in laboratory animals, and certainly more about its workings, than by studies in humans. The following discussion concerns mostly

results from laboratory studies in nonhumans, but the history of validation in humans, where possible, is excellent. Behavioral pharmacology and psychopharmacology are complementary and have a relationship to one another similar to that of other fields of basic pharmacology and their corresponding pharmacotherapies. Well-established principles of behavioral pharmacology will be illustrated by examples in this article; no attempt to be topical will be made.

I. WHAT IS BEHAVIORAL PHARMACOLOGY?

Subjects commonly studied in behavioral pharmacology are people, monkeys, rats, and mice, with rats probably the most studied, but the study of mice is increasing. Larger animals may be required when surgical interventions or electrode implantation in brain are part of the study; otherwise, smaller animals such as mice are easier to house and work with, cheaper, can be studied in larger numbers, and are genetically more homogeneous.

The methods of behavioral pharmacology are numerous and varied. It is usually easy to devise a means of recording objectively a type of behavior of interest, and scores, perhaps hundreds, of methods have been described in the literature. Further, it is the rule rather than the exception for workers in a laboratory to modify published methods to suit their own needs. It is obviously impossible to describe all these methods and variants. Instead, a description of a generic class of methods that have been particularly influential in the development of behavioral pharmacology will be presented. Typically, in such a behavioral pharmacology laboratory, several subjects will be studied each working day, both simultaneously and serially at each of several separate study stations. Each study station has one or more key devices, various means of presenting signals to the subject, and ways of presenting a measured quantity of food or water or other ''reinforcer'' in relation to operations of the key. The key device may be a bar, lever, chain, or string to pull; treadle, switch, or other mechanical device; or even a light beam to a photocell to be interrupted or other nonmechanical detector. They will be referred to generically as ''keys.'' Operation of the key by the subject constitutes a response that generates an electrical signal. In principle, responses could be related to delivery of the reinforcer by mechanical or pneumatic means. Today, almost no one would consider using a nonelectrical system any more than they would consider using smoke signals or semaphore flags to communicate. Examples of signals presented to subjects are colored lights, sounds, and television displays. Occurrences of the signals, recording of the operations of the key, and presentations of the reinforcer are effected by a microcomputer through an interface. The events of the experiment are displayed, tabulated, and stored for later scrutiny, analysis, and summarizing. Sessions for individual subjects may last from a few minutes to many hours, often on a daily basis.

Each subject will work under a consistent program for weeks or months or longer. While a subject is involved in an ongoing experiment, times of feeding, watering, room lights on and off, and other features of the subject's environment during and between the sessions are kept constant. It goes without saying that subjects must be in good health and disposition because a sick or miserable subject must be assumed to have a different behavioral performance and different susceptibility to drug effects. If included, such subjects would bias results; therefore, they are not studied (unless determining the effects on a subject with a particular disease is the object of the study). Repeated sessions allow the range and variability of the control performance, the performance in sessions when no drug or only a drug vehicle such as saline has been given, to be measured with required precision. Effects of drugs are measured as changes from control in the number and pattern in time of responses. Repeated sessions also allow a complete dose–effect curve to be determined in each subject because a series of doses are given over time. The doses are usually given some days apart so the effect of each dose is not influenced by previous doses (not always possible). Cumulative dosing is sometimes employed when incremental amounts of a long-acting drug are given at intervals in a single session. Information on dose–effect relations is an almost universal requirement for interpretable results in pharmacology. Each dose is usually evaluated at least twice. Having dose–effect information on each subject allows both session-to-session variability and subject-to-subject variability to be measured directly. Thus, both sampling errors and systematic and experimental errors are included in the error estimate and not just the theoretical sampling error, which underestimates the actual error, sometimes greatly. At the same time, the shape and position of the dose–effect curve is determined more precisely because the different doses are given to the same subject. Repeated observations on the same subject are quite usual in behavioral pharmacology. Behavioral experiments are generally harmless to the subject.

Repeated use of the same subject is not usual in many fields of pharmacology because measurement of drug effects often requires irreversibly invasive methods (although increasingly more measurements can be made noninvasively or with minimal invasion, e.g., by insertion of a probe).

Much of the previous discussion has to be modified when the subjects are humans. It is rarely possible to make humans work daily for long periods on what are, in the context of human life, usually trivial tasks, so investigators have to be content with much more limited periods of observation than are possible for laboratory animals. Also, it is usually not permissible to give a human subject more than a limited number of drug doses. Further, the environment of human subjects outside the experimental sessions is not able to be controlled as well as the laboratory environment of experimental subjects can be controlled. Therefore, it is more difficult to obtain precise results in humans, but, of course, study them we must. Results obtained in human subjects often require less extrapolation to be applied in the clinic.

The drugs that are studied by behavioral pharmacologists naturally depend on the environment, interests, and purposes of the particular investigators. In academic laboratories, well-established drugs that have been extensively studied pharmacologically and behaviorally continue to be studied in the search for coherent principles of behavioral pharmacology. In recent years, however, more new drugs have been studied that were created to have affinity for particular known receptors or for selected mechanism. Such studies enrich aspects of both neuroscience and behavioral pharmacology. Academic investigators also study new drugs begining clinical use to establish how their behavioral effects fit into the context of what is known about the behavioral pharmacology of previously studied drugs. Finally, there is currently a great effort to understand the behavioral pharmacology of drugs that are important socially because they are abused: cocaine, amphetamines, heroin, and the other street drugs.

In the laboratories of pharmaceutical companies there will usually exist an up-to-date library of the effects of many standard drugs that have been studied in that laboratory in the past to provide reference standards for effects of new chemical entities. New agents that have shown promise in an initial screening procedure are studied further. The behavioral pharmacologist will try to help the discovery team choose the most promising compounds and will provide information about what to expect regarding behavioral effects of the agent if, as all hope, the agent progresses to a clinical trial. As mentioned previously, the behavior that is measured is often simply the pattern in time of objectively recorded responses. It may seem that such a trival sample of behavior—and indeed the key operations are usually deliberately made trivial in terms of effort—is a small basis on which to attempt to understand any significant part of the rich complexity of the behavior of mammals, including humans. It is amazing, however, what a rich variety of behavioral phenomena can be formulated in such a way that they can be illuminated by scrutiny and measurement of patterns in time of responses on keys. In this article, I will only be able to hint at the possibilities. It is well to remember that other sciences have prospered on what seems a trivial sample of their subject matter. All that is known about the universe beyond the solar system is derived from analysis of minute amounts of electromagnetic radiation, and much that is known about subatomic phenomena comes from examination of paths of condensation in bubble chambers and similarly trivial samples. Students of behavior are fortunate in that their subject matter is accessible for experimental probes and repetitions and the information that can be collected is not limited by ineluctable physical constraints, as in astronomy.

II. ORIGINS

The effects of some drugs affecting behavior have been described as far back as there is written record and there are many more agents with an ancient pedigree. Opium, for instance, has been lauded by physicians for millenia. Some have been incorporated into religious rituals (e.g., mescaline for some Amerindians). Until recently, however, knowledge of behavior-affecting drugs was limited to descriptions of what could be seen in people under the influence of the drugs, supplemented by what the subjects told observers. Even when pharmacology developed into a flourishing scientific field, by the end of the 19th century, behavioral pharmacology languished and it was not until the 1950s that behavioral pharmacology became established as a recognized field. Even later, a respected textbook of pharmacology cited Shakespeare on the pharmacology of alcohol. Scientists do not generally cite artistic descriptions as their authority.

As is usual when a new field of science emerges, it is easy to identify prior work that presaged it. For example, the work of Russian Pavlovian physiologists

in the early 20th century was aimed explicitly at developing a science of behavior and extended to studying effects of drugs on behavior. Working primarily on dogs, interesting and influential results were obtained, but the line of work did not lead to a recognized field of pharmacology. The failure was likely due to the rigidity of the framework: All behavior had to be identified as reflexive: unconditioned or conditioned reflexes. That is, behavior consists only of responses to stimuli. Such a view not only is contrary to common experience but also frustrated scientific progress.

Another example is work in Baltimore from 1915 onwards. The vision was clear. A pharmacologist wrote: "The effect of drugs on psychological functions has been the subject of remarkably little investigation on the part of either psychologists or pharmacologists ··· so that the field of 'psychopharmacology' is virgin soil, full of opportunities." Studies on movement capabilities of rats led to methods that were applied to pharmacology. Rats climbed a rope to reach a platform with food and the effects of the drugs on climbing time were measured. Although, again, the field of behavioral pharmacology did not develop in the short run, it is interesting to trace the subsequent history of the methods. After being largely ignored for a couple of decades, they were revived by their originator (and modified, of course) in the 1940s and then applied by scientists at a drug company United States. They were then adapted (and modified) by scientists at a drug company in France. There, they helped in the discovery of the drug chlorpromazine, which in turn was important in the rapid development of behavioral pharmacology in the 1950s.

In retrospect, a paper published by B.F. Skinner and colleague in the psychological literature in 1937 presaged much of the behavioral pharmacology of the 1950s. It passed unnoticed by pharmacologists.

Why was behavioral pharmacology late to develop? Automatic programming was used extensively in behavioral pharmacology from the first. Automatic programming had been in use in industry since at least the 1920s and had been applied to behavioral research in the 1930s, so what eventually started in the 1950s could have started in the 1930s. However, it seemed self-evident to researchers interested in the behavioral effects of drugs that the drugs must affect primarily the "higher" functions of the brain, especially functions that came to be loosely called "cognitive." Learning, in particular, was assumed to be a frequent target of drugs. Higher functions were vaguely associated with higher organisms, particularly humans. (Paradoxically, what we know about learning as a biological phenomenon is compatible with learning being a primitive function with a long evolutionary history: It is not a peculiarly human phenomenon.) However, when researchers tried to measure such effects on higher functions, results were ambiguous, difficult or impossible to interpret, and did not seem to lead anywhere. Therefore, researchers were discouraged and abandoned the line of work, so the field did not develop. The truth is that it is usually bad to approach a scientific field with some of its basics accepted as self-evident because they are quite likely wrong. For example, it was considered self-evident that the earth was flat and stationary at the center of the universe, but approaching the study of the earth and the universe with this conviction was not hopeful. What may seem to be self-evident about behavioral phenomena may be equally false. It was not until behavioral phenomena were approached without prejudice and respected as the appropriate subject matter to be studied for drug effects on behavior that behavioral pharmacology could begin. Preconceptions had diverted attention from the actual subject matter.

It is not clear why behavioral pharmacology started just when it did. It has been suggested that the quantitative graphic method of displaying behavioral phenomena in real time, the cumulative record—the type of display given by a kymograph and as such familiar to pharmacologists of the time—was seminal. Displays of this kind are compelling and informative and do seem to facilitate the development of a field, since physiology was stimulated by the introduction of the kymograph in the mid-19th century. Such displays had been present in behavioral research for a decade and a half. They doubtless facilitated behavioral pharmacology. Whatever the reasons for behavioral pharmacology starting in earnest when it did, it is clear that the effort was launched before chlorpromazine came on the scene. It is equally clear that the great growth of behavioral pharmacology in the next few years was helped by the impact of the introduction of chlorpromazine and one or two other drugs.

III. EXAMPLES AND DISCUSSION OF BEHAVIORAL PHARMACOLOGY STUDIES

A. Example

A method that contributed to the discovery of chlorpromazine was a descendant of the rat rope

climbing studies. A rat cage had a grid floor. A rope hung from the roof of the cage. A rat was put in the cage. A single ring of a bell was sounded and a few seconds later a small AC current was applied across the grid. The current was sufficient to prompt escape by the rat but not, of course, sufficiently strong to cause any harm. If the rat climbed the rope, it removed itself from the current. After a few experiences the rat began climbing the rope promptly when the bell sounded. The effects of drugs on consistency and speed of climbing were studied.

It was known in the 1940s that clinically available antihistamine drugs caused a side effect of "drowsiness" even though the various drugs belonged to many different chemical classes. Sometime before the end of 1950, a group at a French drug company had the insight to consider that perhaps the drowsiness is indicative of interesting new pharmacology. Therefore, starting with an antihistamine, 3277RP (promethazine) marketed by the company, the group helped guide the synthetic program of chemists in the team toward compounds with potent behavioral effects using, inter alia, the method just described as one of their tests. They selected a compound designated 4560RP, which became chlorpromazine. They found many effects of the drug but in particular the following: In the rope (or pole) climbing test, with increasing dose, the latency from the sound to the rat climbing increased progressively until the rat sometimes remained on the grid until onset of the current, whereupon the rat promptly climbed the rope. As dosage was further increased, the rat increasingly did not climb the rope after the sound but continued to do so at onset of the current. The climbing at onset of the current showed that the lack of climbing during the sound was not due to motor incapacity to climb. A barbiturate at increasing dose caused increased latency, but at dose levels at which the rat failed to climb after the sound, it also failed to climb at onset of current. The investigators thought that the dissociation produced by chlorpromazine was indicative of a new and interesting pharmacology. Indeed they were right.

After an inauspicious beginning, chlorpromazine found its way into psychiatric wards in Paris as Largactil. Within a few years it was in use worldwide (as Thorazine in the United States) and had produced the most profound therapeutic revolution that psychiatry has experienced, certainly since the cessation of the punitive treatment of the insane. Chlorpromazine has beneficial effects in psychoses such as schizophrenia and manic-depressive psychoses that previously had no effective treatment. Chlorpromazine is said to have "antipsychotic" properties. What it does particularly well is quieten agitated, disoriented, difficult to handle psychotics. As a result, the large, custodial psychiatric hospitals run by states and cities became quieter and more manageable places. In a relatively short amount of time, a large fraction of the population of hospitals was able to be discharged for care in the community. The hospitals were consolidated and many closed. There have been problems, of course, because many families and communities lack the knowledge, resources, or willingness to continue the level of care that the ex-hospitalized psychotics still require since chlorpromazine is not curative. Still, anyone who remembers the bleak and dreadful hopelessness of the large public so-called mental hospitals can only regard the problem as an acceptable price to pay for the great emancipation.

Chlorpromazine had a major impact on attitude in psychiatry. Hitherto, the mission of psychiatry was to help patients cope with their disorders and be more comfortable. Chlorpromazine came without warning and showed that it was possible to help psychotics with drugs not just by "sedating" them but by moving them toward normality. The question now asked was, if chlorpromazine can do this, for what other beneficial effects can agents be discovered? An optimism about future therapies permeated psychiatry, even though a minority of psychiatrists, most with doctrinaire affiliations with old "schools," resented the advent of pharmacotherapies and regarded help from drugs as failure to persist with persuasion. Because a chemical agent could be antipsychotic, there was new faith that biological bases of psychoses could be discovered and there was increased research activity and funding. The drug companies were galvanized into developing new research and development programs, and many new phenothiazine and other chemical classes of antipsychotic agents were introduced.

B. Methods of Wider Applicability

Operant behavior is defined in *Webster's New Collegiate Dictionary* (1974) as "behavior or responses (as bar pressing by a rat to obtain food) that operate on the environment to produce rewarding and reinforcing effects." Most people have some familiarity with the basic situation but not with its ramifications. It is the type of behavior described in Section I as commonly used in behavioral pharmacology.

In the present context, the word "response" is not used in its dictionary sense but to mean an operant response. According to common usage, a response is a reactionary behavior "to" something. An operant response, however, is simply an elementary unit of behavior that (i) has a selected detectable effect on the environment (e.g., operation of a key) and (ii) can have its frequency of occurrence in similar circumstances in the future increased or maintained by an event called a reinforcer. A reinforcer is recognized because when it occurs in relation to an operant response the frequency of occurrence of the response in similar circumstances is first increased. Then, with repetition of a consistent relationship between responding and reinforcer, a consistent pattern of rates of responding develops and is maintained. Because rates of responding are commonly the focus of interest, a response usually calls for only brief operation of the key, but this is not necessarily so. For example, a response requirement could call for a minimum force to be exerted for many minutes or longer. Obviously, the definitions of operant response and reinforcer are circular, like the definitions of mass and force in Newtonian physics. This means that they have to be dealt with together. The implication in *Webster's* definition that a reinforcer must be rewarding is now known not to be true. There are reinforcers that no one would consider rewarding but that will maintain responding, as will be described later.

Of great importance was the discovery that responding can be maintained with only rare occurrences of the reinforcer, for example, after the elapse of hours with hundreds or even thousands of responses between occurrences. Indeed, the ability to maintain certain types of behavior over long periods depends critically on infrequent occurrences of reinforcers. For many reinforcers, e.g., food or water, frequent presentations lead to temporary loss of efficacy as reinforcers. The program that specifies when the reinforcer will occur is called the schedule of reinforcement. Two basic requirements that can be imposed by schedules are numbers of responses and elapsed time. In its simplest form, the former could specify that the reinforcer would occur when some number of responses has been made since a starting event, usually the last occurrence of the reinforcer (e.g., 30, 100, 300 responses—so-called FR schedules). In the second type, the reinforcer occurs in relation to a response when a certain amount of time has elapsed (e.g., 100, 1000, or 10,000 sec—so-called FI schedules). Under a FI 1000-sec schedule, the reinforcer occurs when a response occurs after the lapse of 1000 sec since the start of timing of the interval. Schedules can have both number and time requirements (e.g., 10 responses with at least 10-sec elapsed time between responses) and sequential requirements. More than one schedule can be programmed in a session either with unique signals (usually lights or tones) paired with each schedule or with no distinctive signals associated with each schedule. More than one schedule can operate simultaneously. It is evident that by combining number and time and sequences and adding second and third keys, adding concurrent schedules and so on, an almost limitless variety of schedules can be devised; in fact, a great many have been studied. Typically, a subject is first exposed to an easy, undemanding preliminary schedule such as the reinforcer occurring immediately when a response occurs. Then parameters are changed incrementally toward the desired schedule parameters. The final schedule is then presented consistently until a similar performance is seen session after session. How long it takes to reach steady state depends on many factors, one being, not surprisingly, the complexity of the schedule. With standard, relatively simple schedules and signals, steady state may be reached in 10–30 sessions and pharmacological interventions can start. The computer programming the schedule can also perform analyses in real time that can be used to modulate schedule parameters within single sessions. The long and detailed description of schedules of intermittent reinforcement has been given because studies on such schedule-controlled patterns of responding have been prominent, even dominant, in the development of behavioral pharmacology. The following example is from work published in the mid-1950s.

C. Example

Steady-state performance of a subject under a schedule was developed and is shown in Fig. 1. The schedule was mult FR30 FI300s, which is technical shorthand for the following schedule: A stimulus, A, starts: When the subject has made 30 responses, a reinforcer is presented (FR30). Then, a stimulus starts again. If it is A, then FR30 is in effect again. If, however, it is a different stimulus, B, then the first response after 300 sec has elapsed leads to the reinforcer. FR and FI components continue in an irregular sequence (Fig. 1), with each component accompanied by its distinctive stimulus. In Fig. 1, each of the four upward-sloping continuous lines shows a complete session. The sessions are spaced some days apart. The record is

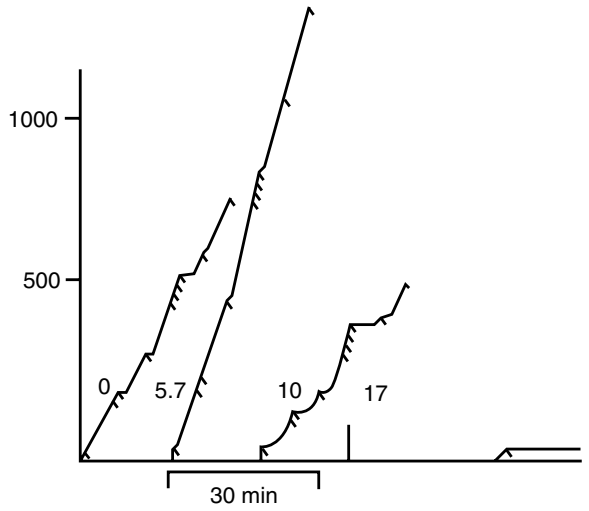


Figure 1 Effects of three different doses of phenobarbital in the same subject. Ordinate is cumulative number of responses in session. For description see text (reproduced with permission from *Annals of the New York Academy of Sciences* **65,** p. 272, Fig. 5, 1956).

made by an ink pen on paper moving from left to right. Each response moves the pen one step up the paper. Results today are not obtained, of course, by measuring a record with a ruler, any more than astronomers hand measure star positions on a photograph. The computer presents results as desired. But graphic displays are still useful in communicating information, and some workers still find the cumulative record a useful real-time monitor.

The leftmost line in Fig. 1 shows the control performance. The remaining three records show the performance of the same subject after doses of 5.7, 10, and 17 mg of phenobarbital, from left to right. (Phenobarbital was widely prescribed in low doses during the first half of the 20th century for its calming effect).

The subject was a pigeon, a frequently studied subject at the time, on a restricted diet; during the session, operation of a key led intermittently to brief food deliveries, the reinforcer. Each session comprised five FR30 segments and five FI 300s segments, each concluded by food delivery. The two distinctive stimuli were visual, different colored lights, consistently accompanying the FR components and the FI components. The sequence of components was the same in all sessions: FR30, FI300 sec, FR30, FI300 sec, FI300 sec, FR30, FR30, FR30, FI300 sec, and FI300 sec. The sessions thus lasted about 30 min. In the first record, labeled "O" (a control session), the FR components appear as two horizontally (i.e., timewise) closely spaced hatch marks with a high constant rate of responding (steep slope). The FI components, in contrast, show low or absent responding at the beginning of the interval, followed by a period of acceleration (short here) to a rate less than that under FR that is then sustained until the interval ends and the reinforcer ensues.

The effects of phenobarbital are major and obvious. Following 5.7 mg, the total number of responses in the session increased to about 1500 (averaging more than 1 per second) compared to about 750 in the control session. Responding under FR is not visibly changed, although the rate is actually somewhat increased. The initial pause in the intervals is markedly attenuated by the phenobarbital so that responding approaches a constant rate through the interval. It should be noted that the 5.7-mg dose is already higher than the usual daily dosage that was given to ambulant patients: The effects, however, are also greater than were sought therapeutically.

Following 10 mg of phenobarbital, FR responding is still not visibly affected, but responding under FI is reduced and almost abolished in the fourth interval. Following 17 mg, responding under FI is abolished, but one FR is completed, albeit at an abnormally low rate of responding.

D. Discussion

A general lesson illustrated by this simple series of experiments is that the effect of a given dose of a drug in a given subject in a single session can be strongly modified by the pattern of responding engendered by the schedule in effect. Specifically, in this example, after an appropriate dose of phenobarbital the rate of responding under FR was slightly increased, while rate of responding under FI was decreased. Responding in a FR component persisted following a dose that abolished responding under FI. As noted, researchers have often approached behavioral research with preconceptions as to what influences on behavior are important (e.g., motivations, emotions, phenomena of learning, and "cognition"). The preconceptions extended to the study of drugs and much ingenuity and effort have been expended in seeking experimental situations that had plausibility as reflecting the influence of interest. For example, if a drug has a different effect on responding in the presence and absence of an influence (e.g., the influence of a putative "emotion" such as "fear"), there was a tendency to immediately attribute the effect to an attenuation (or enhancement) of the influence (e.g., attenuation of fear). The

conclusion was not warranted. It relied on two assumptions: That the change in rate of responding was the effect of fear and that the change due to the drug was due to a selective effect on fear. Either or both assumptions may be incorrect. In any case, it is a basic finding of behavioral pharmacology, illustrated already, that different rates of responding can be differently affected by a dose of drug. If the putative effect of fear was to slow the rate of responding and the drug increased the slowed rate of responding more than the rate in the absence of fear, the parsimonious explanation is that the difference in the rates of responding produces the difference in the effects of the drug. It is not surprising that different rates of responding should be differently affected by a dose of drug. It is commonplace in pharmacology for the effect of drug to be influenced by the physiological state of the system that is affected by the drug. The effect of epinephrine on the smooth muscle of the gastrointestinal tract is to cause relaxation if the tone is already high but contraction if the tone is low. The relationship between effect on drug and the control rate of responding has been extensively investigated. It is surprising, however, that there are many instances in which the effect of the drug, expressed as a ratio to control, is related to the control rate by a simple relationship: Namely, the log of the effect is a descending linear function of the log of the control rate. High control rates may be lowered by a dose of a drug, whereas low control rates are increased. The slopes and intercepts of the relationship vary widely with different drugs and different doses, and the direction of the slope can even be reversed as in the example of phenobarbital.

IV. REINFORCERS

The ability of an event to function as a reinforcer is not an inherent, constant quality of particular stimuli. The ability to function as a reinforcer depends on circumstances. For example, food delivery does not ordinarily function as a reinforcer in a sated subject. Similarly, the ability of fluids, heat, light, sex, and so on to function as reinforcers is dependent on circumstances.

Behavior may also be maintained by schedules that program that responses will postpone, avoid, or turn off a stimulus. The ability of a stimulus to function in this capacity parallels the ability of the reinforcer discussed previously in depending on circumstances. Such stimuli are commonly called aversive stimuli and there are many. Some stimuli can suppress responding, depending on circumstances. A puff of air in the face of a squirrel monkey following a response may suppress responding for food. The same stimulus may (but may not) maintain responding that avoids and suppresses responding that produces it in similar circumstances. None of these findings with so-called positive reinforcers or aversive stimuli that have been described so far are counterintuitive.

In the early days of behavioral pharmacology there was some expectation that the effects of drugs would be different on behavior maintained by positive reinforcement such as food or fluid from that maintained by aversive stimuli. There are instances in which similar behaviors are clearly differently affected by a particular dose of a drug depending on the nature of the reinforcer. In general, however, similar behaviors have been found to be similarly affected, without regard to the maintaining stimuli, which is interesting and somewhat surprising. Amphetamine, for example, may increase response rate both under schedules that program food delivery and under schedules that program termination of lights associated with the occurrence of an aversive stimulus. What is more surprising, however, is that the clear distinction between positive reinforcers and aversive stimuli has broken down. It has been emphasized that the ability of an event to function as a positive reinforcer or as an aversive stimulus depends on circumstances: history, experience, deprivation, ambient influences, and so on. It also depends on the schedule under which the reinforcer is delivered. The same is true of aversive-type stimuli: Their ability to maintain or suppress responding also depends on circumstances and the schedule under which they are delivered. The dependence is not only quantitative but also qualitative.

Monkeys will continue responding when the reinforcer is the delivery of nicotine under, for example, the FI300-sec schedule. Monkeys will also respond to postpone injections of nicotine that would otherwise have occurred. The rate of responding of squirrel monkeys working under FR30 for food was greatly suppressed when the first response in the FR delivered nicotine. The first of these results shows that an injection of nicotine can function as a positive reinforcer, but the second and third show the injection of nicotine to have aversive-type actions. Subjects can be trained to respond regularly to postpone a brief current that would otherwise recur regularly. After the experience of pressing a lever to avoid current, squirrel monkeys were exposed to the following schedule: A

light appeared and remained on for 10 min. The monkey continued to respond, although as for FI 10 min for food, there were no programmed consequences to responses during this time. At the end of 10 min a response produced the same brief current that the subject had previously avoided by responding. The light thereupon went out for 30 sec and then the cycle repeated. The pattern of responding during the 10-min interval came to be similar to that shown in Fig. 1 of the subject responding for food. The aversive-type stimulus was thus functioning as a positive reinforcer. The phenomenon has been independently confirmed in another species and with other types of aversive-type stimuli as well as with drugs such as nicotine.

Thus, an identical physical event, depending on circumstances and schedule, may maintain responding that avoids or postpones the event, may maintain responding that produces the event, or the event may suppress responding maintained by another reinforcer. The schedule and training are the prime determinants of how an aversive-type stimulus will function, and the devil is in the details.

Two caveats are required. First, the phenomena of transmutation of reinforcing function is not a unique peculiarity of application of a particular stimulus. Therefore, seemingly irrational behavior may be maintained by events that would seem to have aversive properties and would not intuitively be expected to be positive reinforcers. When faced with seemingly irrational or self-destructive behavior, this possibility must be considered. The second caveat, however, is that it must not be assumed that any positive reinforcer can be transmuted into an aversive-type stimulus and vice versa according to a standard formula. At the least, the schedule parameters, the environment, and the course of training and experience will be different from agency to agency for it to function, on the one hand, as a positive reinforcer and, on the other hand, as an aversive-type event. At the most, it may simply be impossible to devise means to transmute the reinforcing or aversive function of some stimuli. Insufficient systematic results have been published for it to be possible to delineate what could or could not be accomplished.

V. BEHAVIORAL PHARMACOLOGY OF DRUG ABUSE

Behavioral pharmacology is one of the fields (along with neuroscience, neurochemistry, sociology, psychiatry, epidemiology, etc.) that are pursued in the hope of greater understanding of the phenomena of abuse of drugs. The present discussion will attempt to relate behavioral pharmacological matters to phenomena of drug abuse in society.

It seems self-evident that in the abuse of drugs, such as intravenous injection of heroin or sniffing of cocaine base, the drug represents the reinforcer that maintains the behavior of obtaining the drug. Nevertheless, it was not until it was arranged in the laboratory for an injection of certain drugs to occur occasionally on a response, according to an appropriate schedule, that it was established scientifically that some drugs could function as reinforcers. Responding can be maintained much as with similar presentation of food, fluid, etc. These drugs include most of those that are abused on the streets—notably heroin, cocaine, and amphetamines—but not most of the drugs that are not abused. So-called self-administration has proved to be a useful technique for studying many phenomena relevant to drug abuse, notably circumstances and other drugs that modulate the intake of the drug being self-administered. Combined with "drug discrimination," the study of self-administration can provide laboratory information to help predict abuse liability of agents early in development of a new drug. The following is an example of a drug discrimination situation. A subject responds on two levers. When under the influence of a particular dose of a drug, food is delivered on a schedule of reinforcement of responding on one of the levers, whereas responding on the other lever is without programmed consequences. After saline injection, the consequences of responding on the two levers are reversed: Responses on the second lever are reinforced on the schedule, whereas responses on the first lever are without programmed consequences. A subject may be trained so that while under the influence of the drug almost all responding is on the first lever, whereas when not under the influence of the drug almost all responding is on the second lever. Specifically, a subject can be trained to respond on the first lever when under the influence of heroin. If a new drug in development leads to robust self-administration and to a trained subject under its influence responding on the heroin lever rather than the saline lever, such findings suggest the new agent possesses heroin-like attributes and increase the likelihood that further development of the new agent will be abandoned.

Although self-administration may be an approximate qualitative guide to possible abuse liability, the avidity of self-administration (e.g., the amount of

responding that administration of the agent will maintain and its robustness) is affected by many factors. First, the speed with which the agent can cross the blood–brain barrier and cause its pharmacological effect influences its efficacy as a reinforcer. Generally, rapidity enhances reinforcing effectiveness. Second, agents have pharmacological effects beyond those related to reinforcing effects. For example, on the one hand, cocaine has the ability to increase responding maintained under a variety of schedules of, e.g., food delivery. On the other hand, heroin tends to reduce responding under a variety of schedules and circumstances. Such direct pharmacological effects will inevitably modulate behavior related to self-administration. Such modulations make quantitative comparisons of reinforcing effectiveness difficult, even impossible, without extensive additional investigations. Third, the relative contribution of pharmacological effects to other attributes of abused agents varies widely so that even if the reinforcing effectiveness of the agent were known, it would not generally provide sufficient information to predict abuse liability. This seemingly paradoxical statement can be clarified by examples. For cocaine and heroin, the pharmacology is important for maintenance of their use. In both people and laboratory animals, cocaine or heroin will lead to sustained self-administration with a range of preparations by a variety of routes of administration and under a variety of schedules. Even with heroin and cocaine, however, social context is important in prompting self-administration in a former use; for example, ''hanging out'' with the old gang who are ''doing drugs'' in an environment in which the drug has been taken in the past. There are also reports of people continuing to maintain a habit in a supportive group when the drug they were taking was so diluted as to be essentially without pharmacological effect. With alcohol, the pharmacological aspects are less dominant. The form in which the alcohol comes, the beverage, is more important than is the vehicle for cocaine or heroin. It has been said that a clue as to whether an individual is an alcoholic can be gained by asking ''What do you drink?'' An alcoholic is more likely to name a specific brand than to say whisky or gin. The flavor and bouquet of the beverage are reinforcing, and the appearance of the label and the pouring and mixing are conditioned reinforcers. That is not to say that a Chivas Regal consumer will not drink vodka if nothing else is available, but it will lack some conditioned associations that enhance consumption. With cigarette smoking, the drug nicotine may be even less important in maintaining habitual use. In laboratory animals, self-administration of nicotine could not at first be demonstrated. As the right conditions for maintenance of self-administration were worked out and published, self-administration of nicotine became more reliable across laboratories. It is true that nicotine-free cigarettes do not generally maintain smoking well, but nicotine gum and nicotine patches are easily available but are not appreciably abused. For example, they do not seem to be used by smokers on long nonsmoking flights, when the delay in onset after gum or patch would not be expected to be important. Cigarette smokers generally get little satisfaction from smoking a pipe, although they can readily get as much nicotine from the pipe as from cigarettes. The fragrance and flavor of tobacco smoke and the pleasurable stimulation of the respiratory mucosa in cigarette smoking may be of similar importance to the specific pharmacology of nicotine in maintaining smoking. Cigarette smoking has been dubbed ''nicotine addiction,'' thus associating it with heroin and cocaine, which produce habits that most people have been led to believe are so strong that they cannot be discontinued by the victims' own efforts. Because they have been taught that they are ''addicted'' to nicotine, many cigarette smokers approach the task of discontinuing smoking without confidence that they will be able to do so, which is a good prescription for failure. The role of caffeine in maintaining coffee or tea drinking is even more marginal than nicotine in cigarette smoking. Decaffeinated coffee and tea do maintain their intake, so the fragrance and flavor and fluid intake and, for some, sweetness and ritual are sufficient to maintain consumption.

Drug abuse is not simply a matter of the ability of certain insidious substances to produce a ''euphoria'' so sublime that individuals experiencing it cannot resist the temptation to repeat. Indeed, euphoria was a hypothetical intervening variable whose only postulated properties were to ''explain'' the abuse of the drug. How can the abuse of LSD, MDMA, and other psychedelic intoxicants that produce hallucinations, disorientation, and ''bad trips'' be explained? There may be homology to the laboratory animal with behavior maintained by an aversive-type stimulus. Some habits seem to self-maintain themselves as a quirk of how behavior has evolved to be controlled by historical, contextual, and schedule effects that in most normal circumstances have survival value but can malfunction. The analogy is the immune system, which makes it possible for us to survive in a hostile microbial environment but can produce disease by autoimmune and hypersensitivity reactions.

VI. STATUS OF BEHAVIORAL PHARMACOLOGY

The terms behavioral pharmacology and psychopharmacology both came into common usage in the 1950s. There has never been any doubt about what behavioral pharmacology stood for: rigorous objective assessment of behavioral effects of drugs. Psychopharmacology was more loosely defined at first. In the early days, a session called "psychopharmacology" at a national meeting, such as the Federation of American Societies for Experimental Biology (now Experimental Biology) could have contributions from chemists studying metabolism of catecholamines, behavioral pharmacology studies, and clinical studies. Usage has led to clarification, and as noted psychopharmacology has come to be reserved for primarily clinical science studies.

Early communications in behavioral pharmacology appeared in the *Journal of Pharmacology and Experimental Therapeutics* (JPET) and other pharmacological journals. JPET established specific field editors in 1959 (Vol. 125) to oversee publications in the various fields of pharmacology, one of which was called psychopharmacology. However, it was really behavioral pharmacology, and the name was changed to behavioral pharmacology in 1960 (Vol. 129) and remained until 1998 (Vol. 284), when the system of specific field editors was discontinued. Papers on behavioral pharmacology appear in leading journals of pharmacology throughout the world, including the description of the pharmacology of chlorpromazine in *Archives Internationales de Pharmacodynamie* in 1953. In the early years, there were many behavioral pharmacology papers in the psychological literature, especially in the *Journal of the Experimental Analysis of Behavior*, but fewer behavioral pharmacology publications seem to be following that route today. The Behavioral Pharmacology Society was established in 1957 following a few years of informal meetings, and there is now a European Behavioral Pharmacology Society. In 1999, the American Society for Pharmacology and Experimental Therapeutics established a division for behavioral pharmacology to parallel its division for cardiovascular pharmacology, division for drug metabolism, and several other divisions. In summary, behavioral pharmacology is now a well-recognized scientific discipline with prestigious outlets for its communications.

Psychopharmacology is also well established. Its early development (and also that of behavioral pharmacology) was facilitated by the establishment in 1955 of a Psychopharmacology Service center in the USPH National Institute of Mental Health to foster psychopharmacology (and behavioral pharmacology) and to help collaboration between industry, academic laboratories, and the federal government in developing new agents for psychiatric disorders. It was a manifestation of the wave of optimism for developments in psychiatry that followed the introduction of chlorpromazine. An organization called the Collegium Internationale Neuro-Psychopharmacologicum was founded in 1962 and shortly thereafter established a journal called *Psychopharmacologia* (now *Psychopharmacology*) that has served both behavioral pharmacology and psychopharmacology well. The American College of Neuropsychopharmacology (ACNP) was established in 1961 and started a journal sometime later. The developments in North America followed the pioneering in Europe that gave us chlorpromazine and were paralleled worldwide. ACNP now has approximately 20 sister organizations throughout most of the world. Due partly to the initiatives of such organizations since the 1950s, psychiatric pharmacotherapy, and evaluation of agents both old and new, has changed from being rudimentary in 1950 to as well-developed as any field of pharmacotherapy. It is true that the measuring instruments available in psychiatric pharmacotherapy, mostly scales, inventories, checklists, and the like, are enumerative rather than measurements physically derived in units. However, the care in the design of studies and the means to minimize bias are now of the same standard as prevails in clinical pharmacology of other systems.

Although behavioral pharmacology has been a legitimate science for almost 50 years, it is only in recent decades that it has been able to go beyond concern with behavior and effects of drugs thereon and to establish meaningful relationships with other branches of science. A science is enriched and empowered when it can be related to other scientific disciplines in both reductionist and emergent directions. For example, integrative physiology has been enriched by reductionist analyses. The physiology of Starling's law of the heart, relating the rate of return of blood to the heart to the size and force of the heartbeat, was a legitimate science. However, it becomes much more powerful when the law can be related to the known mechanisms of the molecules that power contraction in the heart cells and how they do it. Physics has been enriched by the feedback from its emergent permeation of astronomy and its reduction to mathematical strings. Attempts of behavioral pharmacology to relate to the effects of drugs on the central nervous system initially had limited success. Not so many years

ago people attempted to relate particular behavioral effects of a drug to the effects of the drug on a particular anatomical part of the brain, e.g., certain drugs that produce behavioral alerting effects were said to be (cerebral) "cortical stimulants." However, neuropharmacologists have not confirmed that their effects are confined to or even predominantly exerted through a discrete gross anatomical part of the brain.

Drugs are chemicals: There is no reason to expect them to respect the gross anatomical contrivances of evolution. However, the chemical anatomy of the brain is another story. Chemicals must do what they do chemically. Although the idea of neurons communicating by extruding chemicals has been mooted by pharmacologists for 100 years, only recently have methods become widely available to analyze the chemical details of how the neurons of the central nervous system communicate. Not surprisingly, it is exceedingly complex. There are many transmitters and modulators and many more receptors as targets so that although the methods for studying them are powerful, the task has only just begun. The ancient Egyptians knew how to survey, but the survey of the world was not completed for several millennia. A powerful tool has even more recently become widely available. It is now possible to selectively delete all the chemicals consequent on the actions of a particular gene by deleting that gene. This approach, so to speak, is orthogonal to traditional methods and such independent flows of information have been valuable historically in the evolution of sciences. Because the ability to modify genetic makeup is more advanced for mice than for other species, it is likely that mice will increasingly become subjects for behavioral pharmacologists. Thus emergent extensions of behavioral pharmacology are also occurring. Funding agencies support behavioral pharmacology for its contributions to mental health research and the problems of drug abuse. Unsentimental drug companies support behavioral pharmacology for its contributions to understanding psychoses and neuroses and to discovery of remedies for their treatment. The future prospects for behavioral pharmacology are bright indeed.

See Also the Following Articles

ALCOHOL DAMAGE TO THE BRAIN • BEHAVIORAL NEUROGENETICS • BEHAVIORAL NEUROIMMUNOLOGY • NEUROBEHAVIORAL TOXICOLOGY • NEUROPHARMACOLOGY • PSYCHONEUROENDOCRINOLOGY

Suggested Reading

Cook, L., and Kelleher, R. T. (1963). Effects of drugs on behavior. *Annu. Rev. Pharmacol.* **3,** 205.

Dews, P. B. (1975). Neurotransmitter balances and behavior. In *Neurotransmitter Balances Regulating Behavior* (E. F. Domino and J. M. Davis, Eds.), p. 125. Univ. Michigan Press, Ann Arbor.

Swazey, J. P. (1974). *Chlorpromazine in Psychiatry. A Study of Therapeutic Innovation*. MIT Press, Cambridge, MA.

Bilingualism

JYOTSNA VAID
Texas A&M University

GLOSSARY

aphasia Deficits in language comprehension or production arising usually after left unilateral damage to the brain following stroke, degenerative disease, or trauma.

bilingual An individual who has knowledge of and regularly uses two languages, although the two languages need not be used in the same contexts or known to the same degree.

cerebral lateralization of function The extent to which particular cognitive skills (such as those underlying language) are functionally mediated more by the left or the right cerebral hemisphere.

crossed aphasia The occurrence of aphasic deficits following injury to the right hemisphere in right-handed individuals.

polyglot aphasia The occurrence of aphasic deficits in one or more languages of individuals who knew two or more languages prior to the brain injury.

Bilingualism has been the focus of a rapidly growing body of research in cognitive neuroscience. This overview summarizes conceptual, methodological, and interpretive issues relevant to understanding the literature on "the bilingual brain" that spans more than 100 years, from early case reports of aphasia in bilinguals and polyglots to recent brain imaging studies of patterns of cortical activation in brain-intact bilinguals. An underlying issue in both the clinical and the experimental research has been to determine whether, or in what circumstances, there are distinct intra- and/or interhemispheric neuropsychological correlates of acquiring and using two or more languages. Four potential sources of evidence on this issue are critically reviewed: (i) language recovery patterns in bilingual or polyglot aphasia, (ii) the incidence of crossed aphasia in bilinguals/polyglots vs monolinguals, (iii) the extent of right hemisphere involvement in language functioning in brain-intact bilinguals vs monolinguals, and (iv) findings from brain imaging studies of bilinguals regarding whether neural regions activated when processing each language are spatially overlapping or distinct. Evidence from each source is discussed in terms of how well it provides a test of the question of whether differential brain localization and/or lateralization can be attributed to knowing and using two languages.

I. WHY BILINGUALS?

Viewed globally, far more individuals understand and use two, if not more, languages in their daily life than those who understand and use only a single language. However, psycholinguistic theorizing has, until recently, held monolingualism as the canonical form of language use to be problematized.

A repercussion of this monolingual-as-norm assumption has been that research on bilingualism has either been neglected or marginalized. To the extent

that bilinguals are viewed as being qualitatively different from monolinguals, they are not seen as being relevant to models addressing the monolingual-as-norm. When included in mainstream research, bilinguals are cast as differing only quantitatively from monolinguals, and they are regarded as valuable for cross-linguistic comparisons (addressing how a particular variable is manifest in one language vs another). There is little room in this view for asking how language processing may be affected by exposure to and coordination of two linguistic systems. Indeed, phenomena that would appear to mark bilinguals as distinct from monolinguals, such as their ability to move from one language to another, have been argued to have counterparts in monolingual usage, such as when a speaker shifts from a formal to an informal speech register. The suggestion that bilinguals are basically the sum of two monolinguals and that there is thus no reason to view them as being qualitatively different from monolinguals may have contributed to the underinclusion of bilingualism in psycholinguistic research.

The view that bilinguals are nothing but the sum of two monolinguals has not gone unchallenged, however. One challenge has taken the form of disputing the monolingual-as-norm assumption on numerical grounds, in view of the fact that bilingualism is globally a more prevalent mode of language experience. As a consequence, to be representative, research and theory would need to address the bilingual situation directly; monolinguals, in this view, would be considered a deviant variety of language use. Adoption of this view would shift what is considered to be interesting in monolingual language use to those phenomena that have clear counterparts in bilingual language use.

Another challenge to the monolingual-as-norm position has been one that privileges neither bilinguals nor monolinguals. Rather, it holds that a truly comprehensive understanding of language functioning depends on enlarging the scope of language experience typically studied to include the full spectrum of language acquisition contexts and language use, rather than the arbitrary designation of only one form of use as appropriate or sufficient for study. Such a view encourages inquiry into the influence of such parameters as the structural properties of the language or linguistic subsystem; when the language was acquired (at birth, before the onset of puberty, or later); what level and type of proficiency was achieved in the language; and how the language was learned and used (formally or informally, primarily in a written mode or in naturalistic discourse, etc.); on how language is perceived, processed, or represented. With the exception of language structural variables, most of these parameters have simply not been examined when only monolinguals have been the subject of study, given that monolinguals tend to be relatively uniform regarding when, how well, and in what circumstances they acquire their language. Bilinguals, however, are quite heterogeneous on these dimensions and thus more readily lend themselves for such study. Insights derived by considering bilingual subgroups varying on these dimensions can thus enrich models of language processing and representation.

Considering the full spectrum of language experience may thus be a more judicious approach to the study of language functioning because it encourages a broader range of questions for study. One fundamental question that has been the subject of much debate and research concerns the cognitive repercussions of exposure to two languages. Vygotsky maintained that the ability to express the same thought in different languages enables a child to view his or her language as one particular system among many and to view its phenomena under more general categories; this leads to awareness of the child's linguistic operations. In his pioneering empirical studies of bilingual memory, Wallace E. Lambert similarly concluded that early exposure to two languages enhances metalinguistic awareness and promotes cognitive flexibility. Lambert further suggested that simultaneous versus consecutive exposure to two languages may have differential consequences, with early, simultaneous exposure having more pronounced cognitive enhancement effects. A vast body of psychological research addressing cognitive concomitants of bilingualism in turn led to questions about neural correlates of variations in language experience.

II. WHY THE BILINGUAL BRAIN?

For a long time, the topic of neurological substrates of bilingualism was simply not addressed in the neuropsychological literature, or it was raised only in obscure outlets. As already noted, a comprehensive account of brain–behavior relations can no longer afford to ignore bilingualism, which represents a particularly prevalent and varied form of language experience. The inclusion of bilinguals broadens the scope of research questions, permitting a fuller examination of how language functioning is influenced both by biological parameters (such as the state of brain

maturation at birth versus at puberty, with obvious implications for differences in the processing of languages acquired early versus late in life) and by cultural parameters (such as how a language is taught and in what context it is primarily used).

The earliest empirical research on the bilingual brain took the form of case reports of selective language loss or recovery following aphasia in speakers of two or more languages. This literature on polyglot aphasia, as it was called, sparked many questions and subsequent investigations of brain involvement in the language functioning of brain-injured and brain-intact bilinguals using a diverse array of methods. Questions raised in the context of the polyglot aphasia literature (and subsequently tested using normative samples) include the following: What factors influence which language is recovered following brain injury in polyglots? To what extent is language impairment parallel or differential across languages? Does brain damage impair the ability to translate or code-switch between languages? Does it produce novel forms of language mixing? How are the different regions of the left hemisphere organized for language in bilinguals? What is the role of the right hemisphere in bilingual versus monolingual language functioning? How does the modality in which the language is learned or other factors in the context of language acquisition influence the pattern of brain involvement in language functioning?

Although there are now several hundred clinical reports of aphasia in polyglots and bilinguals and more than 160 experimental studies on cerebral lateralization of language in brain-intact bilinguals, generalizations from this literature are fraught with difficulty. A major source of this difficulty is the fact that although theorizing about the bilingual brain has been shaped by prevailing conceptions about language, bilingualism, and the brain, each of these domains has undergone considerable changes in conceptualization. Thus, some of the earlier hypotheses examined in the bilingual neuropsychological literature are untenable because they rely on assumptions about the brain, language, or bilingualism that have not been supported. Similarly, some evidence acknowledged to be relevant to testing existing hypotheses has turned out to be inadequate, whether in the sense of being insufficient or unsystematic or in the sense of not providing a true test of the hypothesis under study. Moreover, some phenomena that have been interpreted predominantly in bilingual-specific, neuroanatomical terms (e.g., translation or switching) may alternatively be interpreted in terms of more general cognitive attentional phenomena. Finally, it has become apparent that despite the size of the existing neuropsychological literature on bilingualism, many questions remain understudied.

In light of the complexity of the literature, this review uses a combination of a sociohistorical and a methodological approach, in which the aim is to situate the questions in terms of the prevailing zeitgeist and methodologies available at different points in the research enterprise.

III. HISTORICAL CONTEXT

The association between aphasia and damage to a particular region of the left cerebral hemisphere in the vast majority of right-handed individuals (and, to a lesser extent, among left-handers) has, since it was first noted by the neurologist Paul Broca in the late 1800s, generated an explosion of research into neural bases of language organization and functioning. Much of the early research took the form of clinical case reports of aphasia following various kinds of brain injury. Other clinical research, such as that carried out in the 1950s by the Montreal-based neurosurgeon Wilder Penfield, studied patients scheduled to undergo neurosurgery for treatment of severe cases of epilepsy. These patients first underwent cortical electrical stimulation to map sites responsible for language production. Another technique used with these patients, the Wada technique named after its founder, involves injection of sodium amytal into the left or right cerebral artery, which has the effect of rendering the patient temporarily aphasic, but only when the language-dominant hemisphere has been injected. In contrast to these invasive techniques, which are restricted to patient populations, many noninvasive methods emerged in the 1960s that allowed questions about brain organization of language to be addressed in neurologically healthy individuals. Laterality techniques were used to draw inferences about brain functional asymmetry from performance asymmetries in judgments on dichotic listening or tachistoscopic half-field viewing tasks. These behavioral measures have been supplemented in the past two decades by electrophysiological and neurobehavioral measures based on brain imaging technologies that provide snapshots of cortical and subcortical functioning while participants are engaged in some cognitive task.

The vast majority of clinical reports described aphasic deficits in only a single language. However, as early as 1919, the neurologist Bychowski noted that

"it seems quite improbable that the many aphasics whose extensive case histories have been published should have spoken and understood only the one language of the researcher." What was not known until fairly recently is that many cases were indeed being published documenting aphasia in speakers of two or more languages. Although by no means as numerous as reports of aphasia in (ostensibly) monolingual speakers, reports of aphasia in polyglots have appeared as far back as the 1860s. The fact that as many as 70% of the first 250 cases culled from largely European journals appeared in languages other than English (mostly German or French) made this body of research largely inaccessible to a North American audience until recently, when certain investigators, such as Martin Albert, Loraine Obler, and Michel Paradis, began to make this literature more widely known within the United States through comprehensive monographs and anthologies of this research in English translation.

IV. NEUROANATOMICAL HYPOTHESES ABOUT THE BILINGUAL BRAIN

From the early neurological literature one can extract at least eight different neuroanatomical hypotheses of bilingual language organization proposed during the past century, primarily as explanations for the diverse patterns of language recovery observed in clinical populations.

A. Available Space

According to the available space view, brain damage affects multiple language use by limiting the total brain space available for language functions. Thus, the more extensive the damage, the greater the overall aphasic deficit, regardless of the language. There is no claim as to the deficit being greater for a particular language compared to another.

B. Modality Performance

The modality performance view is often attributed to the neurologist Alexander Luria, who argued that the type of language deficit observed will vary as a function of whether each language was acquired and used primarily in a written or a spoken mode and whether the injury involved areas that selectively disrupted visual or auditory functioning, respectively. Thus, languages used primarily for reading and writing should recover better if the damage spared posterior, occipital cortical areas, whereas languages used primarily in speech should experience more disruption following damage to more anterior, temporal areas.

C. Context/Age of Language Acquisition

According to the context of language acquisition view, proposed by Wallace E. Lambert in the late 1950s, bilinguals who acquired their languages consecutively and in separate contexts (coordinate bilinguals) should be more likely to have differential disturbances in their languages than those who acquired the two languages simultaneously and in similar contexts (compound bilinguals). The assumption is that the earlier a second language is acquired, the more likely it is that it will share the neural substrate with the first language. No claim is made about the locus of the neural substrate. A recent study of 50 Catalan–Spanish bilingual aphasics classified as compound or coordinate did not find a different pattern of language recovery in the two groups; however, as Hamers and Blanc suggest, the criteria used to classify subjects in this study were not the same as those used in the original study by Lambert, and other psychosocial differences between the groups could have been operative.

D. Differential Localization

The differential localization view holds that the different languages of the polyglot are localized in spatially distinct regions of the language-dominant left hemisphere. This view was first articulated by R. Scoresby Jackson in 1867, who, anecdotally describing the case of an Englishman who had lost his knowledge of Greek following head injury, asked, "Where was that gentleman's Greek deposited that it could be blotted out by a single stroke whilst his native language and all else remained?" Scoresby Jackson proposed that since the foot of the third frontal convolution harbors one's native language, any successively learned language might be represented in the rest of that convolution.

E. Selective Inhibition

Pitres, who authored the first monograph devoted to polyglot aphasia in 1895, dismissed speculations of

differential neuroanatomical localization of the languages of polyglots. He argued that it was unnecessary to propose the existence of new centers for each new language learned by polyglots. Instead, he proposed a functional account of language recovery in which the apparent selectivity in language loss and recovery following aphasia reflects inhibitory processes rather than actual neurological destruction of language areas.

F. Differential Intrahemispheric Localization of Proficient vs Nonproficient Languages

On the basis of evidence from naming deficits following cortical stimulation of different sites within the language-dominant hemisphere of bilingual epileptic individuals, H. A. Whitaker suggested that a smaller cortical region is needed to subserve automatized functions (and presumably the language in which one is more proficient) than that needed to subserve functions that are more labored.

G. Differential Lateralization

The notion that the right hemisphere might be more involved in bilingual language use was first proposed by an early neurologist, Gorlitzer von Mundy, based on a single case study. In their monograph on the bilingual brain, Albert and Obler argued for the plausibility of this notion after surveying a vast body of neuropsychological and behavioral research on bilingualism. Much recent laterality work with brain-intact bilinguals (reviewed later) has sought to uncover support for greater right hemisphere involvement in bilingual language functioning.

H. Differential Lateralization of Early vs Late Bilinguals

Chernigovskaya, Balonov, and Deglin presented a theoretical framework based on evidence of language recovery patterns in a Turkmen–Russian schizophrenic undergoing shock therapy. In their proposed framework, semantic representations and surface structures in a second language learned later are held to be localized in the left hemisphere, whereas those in a first language and in a second language acquired in childhood are held to be located in the right hemisphere.

The neuroanatomical hypotheses outlined previously are included here largely for their heuristic value. Many remain to be systematically tested.

V. RELEVANT SOURCES OF EVIDENCE

In the remaining sections, actual clinical and experimental evidence bearing on brain organization of language in bilinguals is reviewed and evaluated. The evidence is examined with respect to the notion of whether bilinguals show differential intrahemispheric localization and/or differential interhemispheric organization of language, either in one of their languages relative to the other or in either language relative to monolingual users or other bilingual users.

There exist many potentially informative sources of evidence. For example, there are four studies of cortical electrical stimulation in a total of 11 polyglot patients, five studies representing a total of 9 bilingual epileptic patients who were administered sodium amytal, two studies of electroconvulsive shock therapy administered to a single bilingual psychiatric patient, and one study of the effects of anesthesia on language recovery in a bilingual patient. While clearly relevant to the present discussion, inferences drawn from these largely clinical studies to language organization in brain-intact bilinguals may be problematic given the very likely possibility that early damage to the brain in many of these patients may have led to functional reorganization of language.

We focus instead on four sources of evidence on which there is much more information. Two of these involve clinical populations and two involve normative samples. The sources to be reviewed include (i) clinical reports of language recovery patterns in polyglot aphasia, (ii) studies of the incidence of crossed aphasia in bilinguals versus monolinguals, (iii) studies of cerebral lateralization of language in brain-intact bilinguals, and (iv) studies of functional neuroimaging in brain-intact bilinguals. For each source, criteria that ideal studies of each type would need to meet to provide a better test of the question are outlined. The extent to which each of these four sources provides an adequate test of the question of differential intra- and/or interhemispheric involvement of language in bilinguals is then addressed.

VI. LANGUAGE RECOVERY FROM APHASIA

The polyglot aphasia literature, consisting in large part of case reports (approximately 400), documents a rich

variety of patterns of language impairment and recovery following unilateral brain injury due to a stroke, tumor, trauma, etc. What is particularly intriguing about many of these reports is that aphasic patients who spoke two or more languages fluently before injury do not necessarily recover their languages at the same rate or to the same degree.

A. Patterns of Recovery Observed in Polyglot Aphasia

Paradis documented six basic patterns of recovery in bilingual or polyglot aphasia: parallel (when both or all languages are similarly impaired and restored at the same rate), differential (when the languages are impaired to a different degree relative to their premorbid proficency), successive (when one language recovers only after another has been maximally restored), antagonistic (when one language regresses as the other improves), selective (when one or more languages do not recover), and blended (when elements of the various languages are mixed or blended inappropriately). These patterns are not mutually exclusive either over time or between languages. Three additional patterns have recently been observed: alternate antagonism (a variant of antagonistic recovery, in which patients alternate in having access to only one of their languages), differential aphasia (where patients present with different types of aphasia in their different languages), and selective aphasia (where there are clear impairments in one language without any measurable deficit in the others).

The existence of these different patterns allows one to infer possible functional models of language organization or functioning in bilinguals. Paradis argues, for example, that cases of parallel recovery indicate that the language faculty as a whole can be affected and thus that it forms a neurofunctional macrosystem. In differential, selective, successive, and antagonistic recovery, one language is clearly selectively impaired, allowing one to presume that each language constitutes a subsystem of the larger language macrosystem.

B. Factors Accounting for Nonparallel Language Recovery

Many factors related to the patients' premorbid language use have been proposed to explain the particular patterns of recovery observed in individual cases. These include familiarity or proficiency (Pitres' rule); primacy (the so-called rule of Ribot); context of use, whereby the language used in the patient's recovery environment was thought to be more likely to return; and affective factors, whereby the language associated with more pleasant experiences in the patient's life was thought to recover better. These factors provide plausible accounts for many observed patterns even though no one principle accounts for the majority of the patterns observed.

The existence of nonparallel recovery in polyglot aphasics, particularly selective or differential aphasia, has also been interpreted to mean that each language is located in a different part of the cortex, whether in the left hemisphere per se or also in the right hemisphere. However, an alternative account first proposed by Pitres is also plausible, according to which each language could be independently inhibited rather than differentially localized. Indeed, antagonistic recovery, and alternating antagonism in particular, cannot easily be explained in terms of differential localization. Pitres' account is particularly pertinent in explaining such cases of temporary inaccessibility. Data from patterns of recovery from polyglot aphasia, in their present form, are equally compatible with a differential localization account as with a functional inhibition account of nonparallel recovery.

C. Limitations of Applicability of Existing Aphasia Evidence

Many limitations of the aphasia literature weaken attempts to argue in favor of differential localization of language in polyglots. First, the accounts of polyglot aphasia are highly varied in level of detail provided both about the nature of the neurological damage and about the language deficits. Indeed, many of the early reports were anecdotal or second-hand accounts. Second, only scattered information is available on the extent of patients' premorbid skills and usage in each of their languages. Third, because of the overrepresentation of single, selected cases in this literature, one does not have an accurate estimate of the relative incidence of the different patterns. The few studies of unselected cases seem to indicate that parallel recovery is the norm. The absence of accurate and meaningful estimates of the probabilities of occurrence of the different patterns of recovery, and their correlation with factors in the bilinguals' language acquisition use prior to insult, makes attempts

to test the various models of brain organization proposed conjectural at best.

During the past two decades, more systematic assessment instruments such as Paradis' Bilingual Aphasia Test, have been developed. The BAT is now available in over 60 languages and 150 specific language pairs. It is hoped that wider use of these instruments will result in more systematic and thorough assessment of unselected cases of aphasia in bilinguals and polyglots.

D. Criteria for Further Evidence Based on Language Recovery from Aphasia

A more direct test of neurological interpretations of differential language recovery patterns in polyglot aphasia would require a more detailed and systematic mapping of aphasic deficits in each of the patients' languages following damage to specific regions of the left hemisphere. One would want to search for consistent patterns of deficits or recovery associated with specific neurological parameters (e.g., the location, size, and etiology of the damage). This approach would require access to comparative data from a large body of unselected cases, diagnosed using a comparable test instrument, with detailed premorbid language use information preserved.

VII. INCIDENCE OF CROSSED APHASIA

Although the predominant focus of investigations of polyglot aphasia has been to document and seek to explain the patterns of language recovery and loss, one aspect of this literature is thought to be particularly relevant to the question of whether there is greater right hemisphere (RH) participation (premorbidly) in bilinguals. Before discussing this evidence, it is instructive to consider what is meant by greater participation in language.

A. Role of the Right Hemisphere in Language Functioning

1. Redundant Role

It is generally acknowledged that the left hemisphere (LH) is the dominant hemisphere for language. Thus, when one talks about greater RH participation in language in bilinguals, it is important to specify what exactly one is hypothesizing because there are at least two senses in which the RH could be said to be more involved in language in bilinguals. The RH could serve as a redundant, additional processor of language available for bilinguals but the LH might still be the predominant hemisphere for language, even in bilinguals. In this view, damage to the RH should not be expected to have differential consequences for bilinguals and monolinguals since the LH would still be intact.

2. Complementary Role

If the RH is instead viewed as being complementary to the LH, mediating functions essential to language processing but that constitute affective and/or pragmatic (rather than phonological or syntactic aspects of language, which may be subserved by the LH), then one would expect that right-sided damage could indeed result in certain kinds of linguistic deficits (i.e., those involving pragmatic or affective components). To the extent that bilinguals rely more than monolinguals on these components in language processing, one would expect bilinguals to be more disrupted than monolinguals by right-sided damage. In this view, it would make sense to compare the relative incidence of aphasia following lesions to the RH in bilinguals and monolinguals. Most research has tended to adopt this latter position in searching for a differential incidence of crossed aphasia in bilinguals and monolinguals.

B. Early Studies of Crossed Aphasia in Polyglots

Crossed aphasia, or aphasia following right-sided lesions in right-handed individuals, is estimated to be very rare, ranging from 2 to 4% of the general population. Three initial studies sought to compare the low estimate of crossed aphasia in monolinguals with that in polyglots, drawing on cases of aphasia in polyglots reported in the published literature in which information about handedness and side of lesion could be extracted. The samples studied ranged in size from 15 to 102. All three studies indicated an incidence of crossed aphasia in right-handed polyglots of 12–15%, nearly three times the estimate for monolinguals.

1. Problems with Early Studies

a. Sampling Bias However provocative the percentages of a higher incidence of crossed aphasia in

polyglots, their reliability has been questioned since the samples on which they are based have an overrepresentation of single or selected cases. Given the rarity of crossed aphasia in the general population, reported accounts of crossed aphasia in bilinguals or polyglots may well have been published because they were unusual. Any conclusions derived from statistical tallies of this literature are therefore open to the criticism of a sampling bias.

b. Variation in Neurolinguistic Assessment Another problem that affects estimates of crossed aphasia obtained by tallying cases drawn from the published literature is that different studies may have used different diagnostic criteria to assess aphasia. Given that neurologists since the mid-19th century have regarded the LH as the "language-dominant" hemisphere (the RH being literally called "mute" or silent), it is entirely likely that the assessment of aphasia due to RH damage may have been less than rigorous in the past, with certain language deficits associated with RH lesions going undetected. In the absence of uniform or comprehensive assessment instruments, variations in diagnostic criteria are likely to add noise to any compilations of data from different laboratories.

C. Criteria for Evaluating the Claim of Greater RH Involvement Based on Crossed Aphasia Evidence

In order to have confidence in the accuracy of the estimates of crossed aphasia in bilinguals, one would want the study to meet the following criteria: (i) The sample should consist of a large number of randomly obtained, unselected cases of monolinguals and bilinguals; (ii) testing should be undertaken at a single location, using a comprehensive testing instrument, with the severity and type of deficit noted along with the occurrence of any language-related deficit; and (iii) cases of left- and right-sided lesions in both aphasic and nonaphasic persons should be included. Comparing bilinguals and monolinguals already diagnosed as aphasic does not provide a true test of the actual incidence of aphasia subsequent to right-sided injury in each group.

D. Recent Studies of Crossed Aphasia in Bilinguals

Three studies meet at least the first two of these criteria. The first study, by Nair and Virmani, examined the incidence of aphasia following right- and left-sided lesions in a large group sample of hemiplegics, many of whom were polyglot, presenting at the researchers' clinic. Although a rather high incidence of aphasia following right-sided lesions was noted (12/24 right handers), a separate breakdown of the relative incidence of crossed aphasia in the polyglots versus the monolinguals in their sample was not provided. Two additional studies have since been published that do provide the requisite breakdown by language status. Both report data from large samples of unselected cases of monolingual and polyglot patients seen at the researchers' clinics in south India.

In a study by Chary, a total of 100 consecutive patients presenting with speech disorders at the author's clinic in Madras, India, were examined (excluding patients with known histories of psychiatric disorders). An aphasia assessment battery administered to the patients revealed that 88 were aphasic. Of these, 31 were polyglot and the remainder monolingual. Of critical interest is the relative incidence of aphasia following injury to the right side. Like that noted in the earlier studies drawn from selected cases, the incidence of crossed aphasia noted by Chary was high for the polyglots (13.6%); however, crossed aphasia was found to be nearly as high in the monolinguals (12.9%).

In the other study of crossed aphasia in bilinguals, by Karanth and Rangamani, two samples were described. One consisted of 48 patients (all right-handed) diagnosed with aphasia at the All India Institute of Speech and Hearing in Mysore in 1984 and 1985. Thirty-two were polyglot and 16 were monolinguals. Four of the patients had crossed aphasia: Two were monolinguals (12.5%) and 2 were polyglot (6.3%).

Note that this sample, as well as that reported by Chary, does not meet criterion (iii) noted previously, inasmuch as patients in these samples already presented with speech disorders. Another sample collected by Karanth and Rangamani that does meet all three criteria is therefore of particular interest. In this sample, all cases of patients treated for stroke at the National Institute of Mental Health and Neurosciences in Bangalore, India, in 1984 were examined. There were 205 cases in all. Of these, 9 were omitted because they had thalamic lesions or dementia. Furthermore, only those patients on whom clear records were available of handedness, side of lesion, and the presence or absence of aphasia (with patients with dysarthria constituting the nonaphasic group) were retained. There were 94 such cases. Of these,

information on language background was available only on 78 patients. Sixty-one of these were polyglots (including only 1 left-hander); the remaining 17 were monolingual (all right-handed). The finding of interest was that the incidence of crossed aphasia was 25% (6/24 patients) among the polyglots; in contrast, none of the 7 monolinguals who had right-sided lesions showed crossed aphasia.

Given the extremely low incidence of reported left-handedness in their samples, Karanth and Rangamani suggest that the higher incidence of crossed aphasia noted in Indian samples as a whole may reflect an overrepresentation of forced right handers, given the strong cultural pressure against use of the left hand. However, even acknowledging that the estimate of crossed aphasia could be influenced by handedness classification, the fact that a higher incidence of crossed aphasia was nevertheless found among the polyglots relative to the monolinguals in the Bangalore sample (which included both aphasic and nonaphasic unselected cases) is intriguing. It remains for further research to determine if this effect is replicable. It would also be helpful for future evidence of this type to provide a more detailed account of the severity of the aphasic impairment in each language following RH damage. Finally, it should be noted that the effort to test the hypothesis of greater RH participation in bilinguals solely by considering the incidence of crossed aphasia may serve to exclude cases of bilateral participation in language, consistent with the other view of RH involvement stated earlier.

E. Overall Assessment of the Generalizability of Clinical Data

Although suggestive, clinical data are not ideal when one wants to extrapolate to healthy populations. In some cases, early onset of brain damage could have resulted in functional reorganization of language functions. Second, assessment of language deficits drawn from studies of aphasics and also from studies of epileptics (e.g., cortical stimulation and amytal data) has tended to weight language production more heavily than comprehension, no doubt because production deficits are easier to observe. Third, interpretation of severity of loss of language functions following aphasia is complicated by the fact that patients may compensate for their deficits by drawing on resources from the nondamaged areas. As a result, it becomes difficult to disentangle premorbid from postmorbid mobilization of brain resources (particularly, RH resources). In light of these problems, if one wants to extrapolate to brain organization of language in healthy individuals, it would be best to study them directly.

Approximately three decades ago it became possible to do just that, due to the refinement of techniques in experimental psychology enabling lateralized presentation of sensory input (whether auditory, visual, tactile, or even olfactory). Most of the techniques available to study brain lateralization in healthy individuals have been applied to bilingual populations.

VIII. CEREBRAL HEMISPHERIC FUNCTIONAL ASYMMETRY

In this section, we consider the available evidence on the question of differential cerebral lateralization of language in brain-intact bilinguals.

A. Overview of the Bilingual Laterality Literature

To date, there are more than 160 studies of brain organization of language in neurologically healthy bilinguals. Nearly one-fourth of these are unpublished studies, including 18 theses and/or doctoral dissertations. The studies may be classified as focusing primarily either on language-specific parameters or on language acquisitional parameters. Studies examining language-specific variables seek to establish whether specific attributes of a language (e.g., a tonal language) are associated with distinct patterns of hemispheric processing compared to a language in which that attribute is absent. In studies of this type, the bilingualism of the participants has been secondary to the goal of examining language-specific differences. Studies examining language acquisitional parameters in turn may be subdivided into those simply trying to establish whether bilinguals as a group differ from monolinguals and those in which specific bilingual subgroups are compared with other bilingual subgroups. In the latter case, the questions of interest center around differences in the context of acquisition of the two languages, i.e., in comparisons between early vs late bilinguals, between formal vs informal learners, between beginning vs advanced second language users, etc.

In an attempt to integrate the various hypothesized outcomes, in 1980 Jyotsna Vaid and Fred Genesee proposed a model according to which,

> *right hemisphere involvement will be more likely the later the second language is learned relative to the first, the more informal the exposure to the second language, and possibly, the earlier the stage of language acquisition. Left hemisphere involvement is more likely the earlier the second language is learned relative to the first, the more formal the exposure to the second language, and the more advanced the stage of acquisition.*

In general, the model posited that the more similar the conditions of first and second language acquisition, all other things being equal, the greater the likelihood that bilinguals will show comparable patterns of hemispheric involvement in processing their two languages, whereas the less similar the language acquisitional conditions, the greater the likelihood of dissimilar patterns of hemispheric involvement in each language, with the specific nature of the pattern reflecting a complex interaction of the effects of the acquisitional variables operative.

1. Languages Studied

The vast majority of bilingual lateralization studies used bilinguals who spoke English as one of their languages. Only 15 studies did not include English as either of the languages studied. Four studies tested speakers from differing languages and 3 studies specifically examined trilinguals. Where English was one of the languages spoken, the five most commonly studied languages paired with it were French (22 studies), Spanish (18 studies), Chinese (15), Hebrew (12), and Japanese (8). Other languages studied with English include (in decreasing frequency), American Sign Language, German, Hindi, Navajo, Portuguese, Crow, Italian, Russian, Yiddish, Arabic, Dutch, Hopi, Korean, Malay, Kannada, Tok Pisin, and Vietnamese.

2. Range of Stimuli and Tasks Used

Stimuli and tasks used have ranged widely from consonant vowel syllables to common single words, word pairs, sentences, and text, presented visually or auditorily. Although many early studies simply measured word recall accuracy, other studies have varied the kinds of judgments to be made on the stimuli, including rhyme, synonym, or semantic and syntactic category judgments. Still other studies have examined shadowing, translation and interpretation accuracy, and Stroop interference.

3. Range of Paradigms Used

The two most commonly used methods have been dichotic listening and tachistoscopic viewing; both of these methods were particularly popular during the 1980s. The next two most commonly used procedures have been the dual-task methodology and electrophysiological measures (which include eletroencephalograph studies, magnetocephalographic studies, and event-related brain potential studies).

4. Range of Outcomes

Studies of lateralization in bilinguals have yielded varied and occasionally contradictory findings: Although some have found group and/or language differences in cerebral lateralization among bilinguals in either or both of their languages, others have reported no differences and still others have found differences that reflect task demands rather than subject variables. However, variability of outcomes should not be surprising to those familiar with laterality studies in general since it is not uncommon to find that even subtle variations in method, task, and stimulus parameters can alter patterns of asymmetries. Adding a set of subject variables related to bilingual language acquisition clearly complicates the situation.

By the early 1980s several reviews of the laterality literature appeared, including methodologically oriented analyses. Aside from variation in methodological parameters, the literature has also been fairly diverse in terms of the particular theoretical rationale underlying the studies. Thus, some researchers have consistently included monolingual and bilingual comparison groups in their design, whereas others have focused on cross-language comparisons within a single bilingual group, and still others have compared the performance of one bilingual group with that of another varying in language acquisition context. These differences make the task of comparing studies more difficult.

B. Vaid and Hall Meta-Analysis, 1991

In an attempt to make sense of this complex literature, in 1991 J. Vaid and D. G. Hall undertook a meta-analytic review of all the bilingual laterality studies on

which the relevant data were available and amenable for analysis at that time. These included a total of 59 studies, one-fourth of which were unpublished. Five hypotheses about presumed effects of bilingualism on brain lateralization as articulated in the laterality literature were tested. The analysis examined comparisons by language, group, and hemisphere; each comparison was further coded on such dimensions as paradigm and response measure, age, sex, handedness, and second language acquisition context (in particular, the relative age, stage, and manner of second language acquisition). Two dependent measures were examined in testing whether a given comparison supported or refuted any of the five hypotheses: an overall hemisphere main effect, indicating the direction and extent of difference between the two hemispheres on the task, and a hemisphere by group or hemisphere by language interaction. Effect size estimates used were Pearson product-moment correlations, which indicated the direction in which and the degree to which scores on a laterality task correlated with involvement of a particular hemisphere or a particular combination of hemisphere by language or group.

1. Hypotheses Tested and their Outcomes

a. The Balanced Bilingual Hypothesis The first hypothesis examined was the most general, comparing bilinguals as a group to monolinguals to test whether bilinguality per se (and the presumed cognitive restructuring it entails) has differential neuropsychological repercussions, reflected perhaps in a greater reliance on RH processing in language, relative to that noted in monolinguals.

A total of 17 studies in the corpus included monolingual controls (English users in the majority of cases). Six of these studies used consonant vowel (CV) syllables as stimuli and were thus analyzed separately, leaving 11 studies. The analysis showed that on their first language bilinguals were not significantly different from monolinguals in laterality patterns. In 8 of the studies for which bilinguals tested in their second language only were compared to monolingual users of that language, bilinguals' performance was found to be significantly more RH lateralized. In the separate analysis undertaken on the CV syllables based on 6 additional studies, all involving speakers of Native American languages (typically Navajo–English users) and English-speaking monolinguals, bilinguals were again found to be significantly more RH lateralized.

b. The Second Language Hypothesis According to this hypothesis, bilinguals should be less lateralized in their second language relative to their first language. This hypothesis was examined in 45 comparisons from 39 studies. The results indicated no significant support for greater RH involvement in the second language. However, several variables moderated the size of the effect: The hypothesis was more likely to be supported under conditions in which subjects received a visual task (tachistoscopic presentation) rather than an auditory task (typically, dichotic presentation) and when response latency rather than accuracy was the dependent measure. Another relevant moderating variable was directionality of the second language: Readers of second languages that are read from right to left (e.g., Hebrew) were more likely to show less lateralized outcomes, reflecting a presumed influence of scanning biases.

c. The Stage of L2 Acquisition Hypothesis According to the stage hypothesis, greater RH involvement should occur in the early stages of second language acquisition than in more advanced stages. Twenty-comparisons drawn from 17 studies formed the basis of this analysis. No support was obtained for the notion that LH involvement increases with greater proficiency.

d. The Manner of L2 Acquisition Hypothesis This hypothesis tested the prediction that a second language acquired in a more informal mode would be more likely to engage RH processing mechanisms than a language learned primarily in a formal context. Only six comparisons drawn from five studies addressed this variable. Although no reliable difference was noted, the effect sizes in five of the six comparisons, though small, were in the direction predicted by the hypothesis. It is possible that a larger set of comparisons may have yielded a significant outcome.

e. The Age of L2 Acquisition Hypothesis Deriving from prior psychological research on the effects of language acquisition context on language processing, this hypothesis proposed that RH participation would differ in early versus late proficient bilinguals (the latter classified as having acquired their second language after the age of 10 years, whereas the former were defined as acquiring both languages before the age of 6 years). The nature of the difference was thought to be such that late bilinguals would show greater RH involvement than early bilinguals.

A total of 16 comparisons drawn from 12 studies (a majority involving French–English bilinguals) tested this hypothesis. Although no early/late bilingual difference was obtained on the first language (L1) comparison, for the second language (L2) and the combined L1/L2 comparisons, the two groups were significantly different. Interestingly, the direction of the difference was that early bilinguals were the less lateralized group.

2. Interpretation of the Meta-Analytic Findings

Viewed as a whole, the meta-analytic results present a mixed picture, not unlike the literature as a whole. Although two hypotheses (e.g., the stage hypothesis and the manner hypothesis) received no support at all, two others received qualified support (e.g., the second language hypothesis and the balanced bilingual hypothesis), and one receiving clear support (the age hypothesis) was supported in an unexpected manner.

It must be noted that the number of comparisons available to test these differing hypotheses ranged widely from a low of 6 (in the case of the manner hypothesis) to a high of 45 (in the case of the second language hypothesis). Moreover, there were surprisingly few studies that included monolingual controls in the corpus at large (and at least 5 that could not be included in the meta-analysis owing to the form in which the data were available).

The lack of support for the stage hypothesis need not be taken at face value if one recognizes that the predominant paradigm used to test this hypothesis (in 11/17 cases) was dichotic listening accuracy. This method is the least sensitive for detecting group differences. According to the stage hypothesis, beginning second language learners are thought to be more attuned to semantic and pragmatic aspects of language input, which is why they are thought to rely more on RH processing. A direct test of this idea would require presentation of tasks that test for differential reliance on these different components of language, which a standard dichotic listening procedure (which more readily taps phonetic components) is not suited to do. It could well be the case that the stage hypothesis has not been appropriately operationalized to date.

a. Task Matters Perhaps the most interesting outcome of the meta-analysis is the one that was not anticipated, namely, that early bilinguals were less lateralized compared to late bilinguals. The age hypothesis arose out of empirical observations that early and late bilinguals differ in what aspects of verbal input they process more readily. Early bilinguals were found to be more inclined to process for meaning, whereas late bilinguals were found to be more adept at processing surface characteristics of words. We now know that tasks requiring semantic processing (relative to those requiring phonetic or syntactic processing) show the most bilateral involvement. Hence, it is not surprising that early bilinguals are the less lateralized subgroup.

In general, the meta-analyis suggests that brain lateralization for one or both languages of certain subgroups of bilinguals may indeed differ from that in other subgroups and monolinguals. A close examination of individual studies in which differences were obtained suggests that differences arise when the task can be performed in more than one way and as such may reflect strategy differences. The importance of including appropriate task measures to directly tap into presumed strategy differences thus becomes evident. Many of the early studies represented in the meta-analysis were not designed to tease out the relative components of language processing or to systematically examine task-related processing demands interacting with individual differences in task performance.

C. Evaluating the Functional Asymmetry Evidence

1. The Validity Issue: When Can One Infer Hemispheric Asymmetry?

Performance asymmetries are intended to signal differences between the hemispheres in underlying specialization of function. Unfortunately, there is not always a direct or simple correspondence between behavioral asymmetries and hemispheric asymmetry: Many artifactual, nonhemispheric factors have been shown to affect the size and even the direction of behavioral asymmetries. In their initial enthusiasm with the availability of noninvasive methods to infer hemispheric lateralization, researchers neglected to search for and rule out noncerebrally based explanations, tending to ascribe any observed asymmetry to a neural source or to interpret any lack of behavioral asymmetry in terms of bilateral hemispheric involvement. The danger in this kind of reasoning is magnified when one is additionally claiming group differences in degree of lateralization. A standard criticism of many bilingual laterality studies has been that in the absence of objectively assessed, systematic criteria for subject

selection, such as language and task proficiency, group differences could simply reflect floor or ceiling effects mistakenly attributed to bilateral hemispheric involvement.

2. The Conceptualization Issue: What Aspects of Language Are Lateralized in Each Hemisphere?

It is now known that language is not uniformly subserved by the LH. When applied to bilinguals, the question of differential lateralization of language becomes one of interpreting what a greater RH effect in bilinguals might mean. Does it mean that aspects of language normally subserved in the LH in monolinguals are bilaterally mediated in bilinguals? Or does it mean that in both monolinguals and bilinguals language lateralization is the same; what differs is the relative reliance of the groups on those aspects of language thought to be under RH control. Although the latter view would appear to be more parsimonious, the research to date is still not at a point where these possibilities can be separated and tested. One can try and disentangle these possibilities by designing studies with processing components presumably mediated by one or the other hemisphere to determine if hemispheric differences will differ by task. In some work that attempted to do this, Vaid found that when a task calls for phonetic processing, a LH superiority was found in both languages of bilinguals, whether early or late bilinguals, as well as in monolinguals. Similarly, when the task called for syntactic processing, a LH superiority was again observed in all groups. On semantic tasks, group differences emerged, with monolinguals and late bilinguals showing a LH superiority and early bilinguals showing either no hemisphere differences or a RH superiority in both their languages.

D. Vaid and Hull Meta-Analysis, 2001

Since the 1991 meta-analysis of Vaid and Hall, the literature on behavioral functional asymmetry in bilinguals has more than doubled. Many of the newer studies do not suffer from at least the more obvious methodological limitations that plagued the earlier studies, although it has also become clear that continued use of standard dichotic listening procedures may not be particularly informative. Many recent studies have begun to consider task-related processing differences in interaction with language and group differences. In view of the expanded literature and its greater sensitivity to task-related variables, a new meta-analysis was undertaken by Vaid and Hull in 2001. This meta-analysis focused only on those studies, whether published or unpublished, in which monolinguals were compared with late bilinguals on their first language or with early bilingual speakers.

In all, 28 studies met the above inclusion criteria. Three findings from Vaid and Hull are of particular relevance to the present discussion. First, there were significant effects attributable to paradigm: dichotic-listening was the most strongly left lateralized, whereas the tachistoscopic and dual task methods were both bilateral. Task was also a significant moderator: Tasks involving surface auditory processing were the most left lateralized, whereas those involving global or semantic processing were the least. Second, there was a clear bilingual/monolingual difference: Monolinguals were significantly more left lateralized than bilinguals, who showed no overall asymmetry as a group. Third, early bilinguals were less lateralized than monolinguals or than late, fluent bilinguals. This group difference was more pronounced among men than women. Additional meta-analyses are planned for those laterality studies in which bilinguals were compared with other bilinguals (i.e., in the absence of monolingual comparison groups). For now, the results of Vaid and Hull corroborate Vaid and Hall's 1991 finding of less pronounced functional asymmetry associated with bilinguality, particularly early bilinguality, relative to monolinguality.

What remains is to consider studies that have used neuroimaging techniques to address the question of differential intra- or interhemispheric activation in bilinguals' language processing.

IX. NEUROIMAGING DATA

Our focus here is on studies using positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) techniques to examine patterns of language activation. These techniques are being increasingly extended to bilingual populations and have been the focus of a recent review by Vaid and Hull. Although PET studies provide good temporal resolution, the results from individual participants are usually not available because group averages are typically reported. fMRI studies, in turn, offer better spatial resolution and permit visualization of data of

individual participants. Studies using each methodology with monolinguals have already uncovered strong evidence for greater LH activation in the so-called classic language areas. However, bilateral activation has also been reported more than might have been expected on the basis of the lesion data.

A. An Overview of Bilingual Neuroimaging Studies

Within the past decade, several studies using PET and fMRI techniques to assess patterns of neural activation in healthy bilingual adults have been performed, including 13 studies using PET and 25 using fMRI.

1. Participant Characteristics

Two of the neuroimaging studies tested bilinguals from diverse languages and two tested trilinguals in all three languages. The predominant language pairs studied to date have beeen Chinese–English (over 10 studies), followed by French–English, Spanish–English, and Sign language–English. Other language pairs studied include German–English, Italian–English, Polish–English, Russian–English, Spanish–Catalan, Russian–German, Japanese–English, and Finnish–English. The majority of the studies used bilinguals who had acquired their second language, on average, after the age of 7 years. The proficiency of these bilinguals varied from moderate to high, although the criteria for assessment of proficiency, when stated, also varied.

2. Tasks

The studies differ in the types of tasks employed and in the modality of the input or output. The range of tasks employed included sentence-level comprehension, listening to stories, thinking about the events of the previous day in one language or another, silent word generation including repetition, word stem completion and generating words beginning with a specified letter ture, cued picture naming, translation, rhyme vs synonym generation, and semantic vs nonsemantic judgments. Only two of the studies kept the stimuli constant, with only the task varying. In one of these, subjects were to make abstract/concrete judgments, as an example of a semantic task, or upper-lowercase judgments based on the same stimuli, as an example of a nonsemantic task. The baseline tasks used (necessary to do the analyses, which are based on a subtractive logic) ranged from silent rest periods (during which subjects were asked not to think at all) to silent word repetition tasks and reading tasks.

B. Patterns of Activation across L1 and L2: Distinct or Overlapping?

The question explicitly addressed in the studies to date that have examined bilingual participants has been whether the two languages of bilinguals activate nonoverlapping, spatially distinct brain regions (cortically or subcortically, in the left or right side) or whether the regions of activation overlap.

With respect to this question, three sets of outcomes have been reported: (i) a language-specific effect, as, for example, noted in the studies comparing users of signed versus spoken languages, whereby the two languages of the bilinguals show overlapping LH regions of activation but where only one language (the signed language), whether in hearing or deaf users, shows greater RH activation; (ii) a pattern of consistent and overlapping activation in, but not restricted to, the classic language areas in the LH and (iii) a pattern of greater variability of activation of the two languages in the L2, both cortically and subcortically, in the left and right sides.

The predominant outcome has been a pattern of overlapping activation for both languages of bilinguals in the classic language areas on the left side and in the pattern of bilateral activation in homologous sites. However, exceptions to this pattern have been reported in a few fMRI studies.

1. Patterns of Nonoverlapping Activation in fMRI Studies

One fMRI study, based on five bilinguals who had studied a third language and were tested in all three on a silent word-generation task, reported that the amount of activation in the three language areas studied was greatest in the least proficient language. Another study found that while early bilinguals showed overlapping activation for L1 and L2 in the left inferior frontal gyrus (Brodman's area 44), late bilingual speakers of different languages showed differential activation within this region for their two languages. A third study reported that activation in the left and right frontal and temporal regions was more widespread in the second than in the first

language of French–English late bilinguals of moderate proficiency.

2. Patterns of Nonoverlapping Activation in PET Studies

An interpretation of a unitary cortical representation of the bilinguals' two languages based on findings of no differences between the languages noted in many of the PET studies has been countered by suggestions that the PET studies, since they rely on averaged data, may not be as sensitive in detecting actual differences as fMRI studies, particularly when these may involve adjacent regions. However, even PET studies have occasionally reported some differences across the languages in patterns of activation.

In two PET studies, both with English–French bilinguals, greater left putaminal activation was observed in the participants' second than first language on word-generation tasks; this effect was interpreted to reflect greater articulatory effort in the second language. However, two subsequent PET studies (one with Chinese–English and one German–English speakers) have not replicated this finding, suggesting that other parameters may also be at play. Variation in activation patterns by language in still other regions has also been reported. In one PET study, greater activation was found for the supramarginal gyrus in German–English late bilinguals perceiving words in their second language; the other study reported greater right middle temporal gyrus activation for L1 and greater right hippocampal and superior parietal lobule activation for L2 in Spanish–Catalan early bilinguals.

C. Summary of Outcomes of Imaging Studies

The predominant outcome from these studies is that of a robust pattern of overlapping activation in users of even highly diverse languages such as Chinese and English, particularly among late bilinguals. However, although most studies stress a strong overlap in patterns of neural activation for L1 and L2, slightly more than half document some variability in activation patterns across the two languages. For example, users of sign languages vs spoken languages appear to show differences in interhemispheric activation. Moreover, studies of Catalan vs Spanish, despite being highly similar languages, have nevertheless shown differences (in the RH) even among early bilinguals. Studies of Chinese–English speakers have tended not to show differential activation. In some cases, the overlap in activation has been attributed to carry-over effects from L1 to L2 processing.

D. Interpretation of Imaging Study Outcomes

1. Nonproficient Bilinguals

Although there has been no direct test of the role of proficiency as a determinant of variability in sites activated in the second language relative to the first, many of the studies propose proficiency (rather than age of onset of bilingualism) as a post hoc explanation of the discrepant results noted in the literature. However, only one study, by Perani *et al.* in 1998, actually compared two groups of late bilinguals varying in second language proficiency. In referring to their finding, Perani *et al.* echo the stage hypothesis:

> *A possible interpretation of what brain imaging is telling us is that, in the case of low proficiency individuals, multiple and variable brain regions are recruited to handle as far as possible the dimensions of L2 which are different from L1. As proficiency increases, the highly proficient bilinguals use the same neural machinery to deal with L1 and L2.*

2. Proficient Bilinguals

The predominant observation of overlapping activation in the two languages of bilinguals raises a question of its own. In the words of Perani *et al.*,

> *How do we reconcile the discrepancy we observe between the imaging data (largely similar activations with L1 and L2 in highly proficient individuals, regardless of age of acquisition) with … behavioral findings [of L1–L1 differences even in early bilinguals]? … We wish to raise the possibility that spatially overlapping networks to process L1 and L2 should not immediately be equated with competence, or performance identity.*

So just how are we to interpret these results? The question resists easy answers, for as Perani *et al.* point out,

> *What our results show is that for the happy few late bilinguals that reach high proficiency, the (macroscopic) brain activation is similar to that of native learners of the language. What we do not know, however, is whether the similarity in brain*

activation is the consequence or cause of learning L2 successfully. Further research is needed to clarify this point.

E. Methodological Concerns

For many reasons, the neuroimaging data in their present form do not provide satisfactory or conclusive answers to the question of whether there is differential hemispheric activation in bilinguals and what overlapping patterns of activation might mean. As already noted, the studies have varied widely in their tasks (including overt and covert production vs comprehension tasks, with the unit of language varying from single words to sentences or discourse); however, task parameters have not specifically been addressed in most of the studies. Although proficiency and age of onset of bilingualism have been proposed to explain discrepancies across studies, only two studies specifically compared early versus late proficient bilingual users (controlling for language used), and only one directly compared proficient and nonproficient late bilinguals.

More important, since only 3 of the 38 imaging studies employed monolinguals and only 4 controlled for whether a particular language was the first or the second language, it is not clear to what extent the patterns of activation observed (whether differential or overlapping) simply reflect language-specific effects (e.g., characteristics of the languages' orthographies, phonologies, or grammar). Another problem in interpreting the studies is that task performance was not consistently monitored behaviorally. It is important to do so both to ensure that subjects were complying with instructions and to determine subjects' level of performance on each language. In some cases, where subjects were reported to be slower in performing the task in the second language, their pattern of brain activation showed no difference across languages. In other studies subjects were pretested to be equivalent in their task performance in the two languages and overlapping patterns of brain activation were again observed.

As Paradis cautions, PET and fMRI data may tell us that some cerebral area is activated during the performance of a particular task, but they do not tell us what function is activated in this area, nor is it possible to determine if the areas that are activated in a particular study are crucial or incidental to language. Furthermore, failure to detect activation in a specific area is not evidence that this area is not active—only that it is not revealed by the procedure that is used. The likelihood of observing a change in activation is influenced by both the study design and the image analysis technique employed. That is, areas found to be significantly activated are activated only with reference to the particular task chosen as a baseline (whether this be attentive silence, periods of rest, random noise, word repetition, or backward speech) and with reference to the statistical tools (and cutoff thresholds for significance) and rationale used, regarding what is to be subtracted from what. Indeed, the appropriateness of the practice of using a subtractive logic to study functional activation of language in general has been questioned.

F. Criteria for Evaluating Bilingual Neuroimaging Studies

Given improvements in the availability and use of neuroimaging technologies, one should expect to see many more studies being conducted with a variety of populations, including bilinguals. However, caution should be exercised in interpreting these studies because they are still at a preliminary stage with regard to researchers' sensitivity to possible artifacts. At the very least, the following set of considerations should be applied when designing or evaluating future neural imaging studies of bilinguals.

1. Adequate assessment of the bilinguals on proficiency and other language acquisition and use parameters,
2. Use of a homogenous language sample rather than speakers of various different language pairs
3. Inclusion of early and late bilinguals, with each group subdivided according to degree of L1 and L2 proficiency
4. Inclusion of monolingual comparison groups for each of the languages under study
5. Monitoring of participants' behavioral responses during the scanning to provide additional evidence regarding task performance and to ensure that participants complied with instructions
6. Inclusion of a measure of stimulus complexity (e.g., word frequency or other measures)
7. Inclusion of some task parameters, ideally in such a way that the stimuli are kept constant, leaving only the task to vary (e.g., by varying the instructional demands)
8. Inclusion of more than one baseline measure for purposes of comparison

9. Examination of all sites where either significant increases or decreases in activation were noted, not just sites in the so-called classic language areas
10. Statistical testing for relative activation in left and right hemisphere sites, both cortical and subcortical, to the extent possible
11. Use of converging measures (e.g., testing the same subjects on different paradigms)
12. Use of statistical analyses that search for interaction effects of language or by group by task or site direction
13. Use of hypotheses that are grounded in prior theory and research in bilingualism.

Despite the rapid proliferation of neuroimaging studies using bilingual participants, the image of the bilingual brain that has emerged to date is neither uniform nor easily interpreted.

X. CONCLUSION: WHAT IS RIGHT? WHAT IS LEFT?

After examining four different sources of evidence bearing on the question of differential intra- or interhemispheric organization of language in bilinguals, one is left with an appreciation of the danger of taking findings at face value. In many cases, the evidence has not been "clean" enough to permit a direct test of the question of interest, allowing for alternative explanations of the phenomena observed. In other cases, the patterns observed appear to be influenced by a myriad of variables, including those specific to the population used (i.e., bilinguals), those specific to the methodologies, and interactions between the two. No one method appears to be ideally suited to address the overall question because each frames the question according to its own assumptions. For example, the laterality data ultimately do not permit examination of nonlateralized differences (e.g., cortical vs subcortical influences or interhemispheric cooperation vs competition effects). The crossed aphasia data do not permit examination of bilateral mediation of language in either bilinguals or monolinguals since they focus on patients with unilateral lesions. The neuroimaging data are constrained by the use of a subtractive logic and by inconsistencies in subject selection and comparison.

At a more subtle level, how one chooses to interpret the patterns observed, or even how one chooses to frame particular questions, may reflect ideological differences regarding bilingualism such as those outlined at the outset. An implicit adoption of the prevailing view of bilinguals as being the sum of two monolinguals may in part account for why many of the behavioral functional asymmetry studies and nearly all the functional neuroimaging ones sought to compare lateralization for L1 versus L2, as though the two languages of bilinguals are indeed separable and independent systems (in some cases, this has meant arbitrarily designating the languages as L1 or L2 as when classifying early bilinguals or trilinguals). In contrast, if one takes the view that the process of acquiring and mastering two or more languages alters how language is processed, one might be led to a different sort of comparison (e.g., not between the two languages so much as between bilinguals and monolinguals or between certain bilinguals and other bilinguals on different tasks).

Yet another variable that has influenced the interpretation of the evidence has been ideological orientations regarding language. For some researchers, language continues to be viewed as a stable, monolithic system, with grammar as its defining core, localized in the LH. However, recent evidence from research with monolinguals suggests a more sophisticated, dynamic, and multidimensional view of language as a system that calls on a number of other processes, some of which may be bilaterally mediated and some subserved by the RH. Clearly, advances in our understanding of language will have distinct implications for how one regards questions about how language is processed in bilinguals, both behaviorally and neurobehaviorally. Similarly, advances in our understanding of bilingualism and in our understanding of the brain will alter the way in which questions about language processing will be framed and interpreted in future research.

Despite the number and perhaps because of the sheer diversity of neuropsychological studies with bilinguals now available, generalizations are difficult to formulate. Progress in this field will depend on attention to reliability and validity of the measures used to infer functional asymmetry, more appropriate and converging methodologies for testing the hypotheses advanced, and a more theoretical grounding of the research undertaken. Further studies need to be tuned to current developments in cognitive and linguistic research on bilingualism and on L2 learning as well as to developments in neuroscience. Only through a joint consideration of cognitive and neuropsychological parameters and assiduous design of experiments can one begin to make sense of the complex and often seemingly contradictory evidence brought to bear on

long-standing questions about language and brain organization in bilinguals and monolinguals alike.

See Also the Following Articles

APHASIA • LANGUAGE ACQUISITION • LANGUAGE AND LEXICAL PROCESSING • LANGUAGE DISORDERS • LANGUAGE, NEURAL BASIS OF • LATERALITY • READING DISORDERS, DEVELOPMENTAL

Acknowledgments

I thank T. Gollan, M. Paradis, and L. K. Obler for comments and for sharing data from unpublished work. Preparation of this article was supported by a Texas A&M University Honors Teacher/Scholar award.

Suggested Reading

Albert, M., and Obler, L. K. (1978). *The Bilingual Brain: Neuropsychological and Neurolinguistic Aspects of Bilingualism*. Academic Press, New York.

Cook, V. (1997). The consequences of bilingualism for cognitive processing. In *Tutorials in Bilingualism: Psycholinguistic Perspectives* (A. de Groot and J. Kroll, Eds.), pp. 279–299. Erlbaum, Mahwah, NJ.

Fabbro, F. (1999). *The Neurolinguistics of Bilingualism: An Introduction*. Psychology Press, East Sussex, UK.

Federmeier, K., and Kutas, M. (1999). Right words and left words: Electrophysiological evidence for hemispheric differences in meaning processing. *Cognitive Brain Res.* **8,** 373–392.

Grosjean, F. (1998). Studying bilinguals: Methodological and conceptual issues. *Bilingualism Language Cognition* **1,** 131–149.

Hamers, J., and Blanc, M. (2000). *Bilinguality and Bilingualism*. Cambridge Univ. Press, Cambridge, UK.

Illes, J., Francis, W., Desmond, J. E., Gabrieli, J. D. E., Glover, G. H., Poldrack, R., Lee, C. J., and Wagner, A. D. (1999). Convergent cortical representation of semantic processing in bilinguals. *Brain Language* **70,** 347–363.

Karanth, P., and Rangamani, G. N. (1988). Crossed aphasia in multilinguals. *Brain Language* **34,** 169–180.

Lambert, W. E. (1969). Psychological studies of interdependencies of the bilingual's two languages. In *Substance and Structure of Language* (J. Puhvel, Ed.), pp. 99–126. University of California Press, Los Angeles.

Paradis, M. (2001). Bilingual and polyglot aphasia. In *Handbook of Neuropsychology* (R. S. Berndt, Ed.), Vol. 3, 69–91. Elsevier, Amsterdam.

Paradis, M. (1999, August). Neuroimaging studies of the bilingual brain: Some words of caution. Paper presented at the 25th Lacus Forum, University of Alberta, Edmonton, Alberte, Canada.

Paradis, M. (2000). Generalizable outcomes of bilingual aphasia research. *Folia Phoniatrica Logopaedica* **52,** 54–64.

Perani, D., Paulesu, E., Sebastian Galles, N., Dupoux, E., Dehaene, S., Bettinardi, V., Cappa, S., Fazio, F., and Mehler, J. (1998). The bilingual brain: Proficiency and age of acquisition of the second language. *Brain* **121,** 1841–1852.

Rapport, R., Tan, C., and Whitaker, H. A. (1983). Language function and dysfunction among Chinese- and English-speaking polyglots: Cortical stimulation, Wada testing and clinical studies. *Brain and Language* **18,** 342–366.

Vaid, J., and Hall, D. G. (1991). Neuropsychological perspectives on bilingualism: Right, left, and center. In *Bilingualism, Multiculturalism, and Second Language Learning: The McGill Conference in Honour of Wallace E. Lambert* (A. G. Reynolds, Ed.), pp. 81–112. Erlbaum, Hillsdale, NJ.

Vaid, J., and Hull, R. (2001, April). *A tale of two hemispheres: A meta-analytic review of the bilingual brain*. Paper presented at the Third International Symposium on Bilingualism, Bristol, UK.

Vaid, J., and Hull, R. (2002, in press). Re-envisioning the bilingual brain: Methodological and interpretive issues using functional neuroimaging. In *Advances in the Neurolinguistics of Bilingualism* (F. Fabbro, Ed.). Udine Univ. Press, Udine.

Zatorre, R. (1989). On the representation of multiple languages in the brain: Old problems and new directions. *Brain Language* **36,** 127–147.

Biofeedback

RICHARD GEVIRTZ
CSPP Alliant International University, California

GLOSSARY

autonomic nervous system (ANS) A branch of the peripheral nervous system that controls any vital functions in the body, especially those involved in adaptation to environmental demands.

biofeedback The process of using information from one's own biological signals to regulate a physiological and/or psychological process.

electroencephalogram (EEG) Electrical activity from the brain is summarized on a graphic display that separates frequency components.

electromyogram (EMG) Electrical activity emitted from the nerves enervating muscle displayed as a raw signal or an integrated average (in microvolts).

neurofeedback Brain wave information is shown and used to achieve a change in function.

parasympathetic nervous system (PNS) One of two branches of the autonomic nervous system usually associated with the restoration of steady states or homeostasis.

respiratory sinus arrhythmia (RSA) The fluctuation of heart rate with normal respiration (heat rate increases with inspiration and decreases with expiration).

sympathetic nervous system (SNS) A branch of the autonomic nervous system usually associated with readiness for fight, flight, or a freeze response.

Biofeedback, the process of using one's own biological signals to achieve a change in physiological functioning, has become increasingly popular for disorders that do not respond well to standard medical treatment, for performance enhancement, and as an adjunct to various meditative or relaxation techniques. This article describes the different ways biofeedback is used in current contexts and briefly summarizes the scientific basis for its efficacy.

I. INTRODUCTION

Humans have demonstrated fascination with the connections between mind and body from our earliest records to the present. As conscious and self-conscious creatures, we have always wondered how our emotions, thoughts, or ideas affect our bodily processes. Scientific progress on this topic had been slow, however, until the latter half of the twentieth century because we were unable to measure physiological processes easily and efficiently. At the same time, the psychological side of the equation was not well-conceptualized until modern psychology abandoned spiritualistic models for cognitive and evolutionary biological ones. In the 1950s, scientists working to delineate the mind–body connection created the term psychophysiology to distinguish themselves from physiological psychology or other early neuroscience subspecialties. As the technology and conceptual

development progressed, psychophysiology also prospered. Typically, early psychophysiological studies manipulated social–psychological variables while measuring physical parameters. For example, early leading researchers (the Lacys, Paul Obrist, or Peter Lang) began to interpret heart rate in light of various cognitive–emotional states. In this way, progress has slowly been made in psychophysiology. At the same time, some psychophysiologists and others asked a different question. If the subject could be made aware of their own physiology, could they actually change it? This "mind over body" idea has always had adherents in religious traditions, eastern meditative traditions, and even medicine (physician as healer). At this point there was a scientific paradigm with the potential to elucidate this concept. Early successes in the laboratory led to an emerging clinical cohort who have tried to use biofeedback to correct problems, enhance performance, or modify experience. Biofeedback is the term coined to indicate the influence of information or feedback of various physiological parameters on the regulation of the same. For example, it was demonstrated that, if a person was given a meter of his or her heart rate and asked to slow or speed it, he or she could in fact accomplish this, at least statistically.

By the 1990s, growing public dissatisfaction with traditional western medical paradigms pushed health providers to offer complementary–alternative medicine services, and biofeedback is often included in this category. Greater sophistication in the neurosciences, together with less expensive instrumentation, has also led to a resurgence of interest in EEG biofeedback, often called neurofeedback. The revolutionary idea that a person could modify his or her own brain wave activity and thereby modify arousal, attention, or seizures has tweaked the imagination of the public.

From these various roots, the field of biofeedback or applied psychophysiology has grown to be a health profession with an international following, professional organizations, and a certification process.

II. BIOFEEDBACK DIRECTED AT MODIFICATION OF A SPECIFIC PHYSIOLOGICAL SYSTEM

There is considerable evidence that subjects or patients can regulate some physiological systems. For systems that fall under the regulation of the voluntary or somatosensory nervous system, the models are not all that revolutionary. For example, feedback from striate muscle with an electromyograph can be used to relax or contract muscles that may have slipped from voluntary control. This kind of work has been done in at least three settings.

A. Rehabilitation

Neuromuscular rehabilitation centers have used EMG biofeedback to successfully aid physical therapy and occupational therapy procedures with stroke, spinal cord, head injury, and various neuromuscular disorder patients. Typically, the feedback is used with stimulation and exercises to recruit neuromuscular pathways that are not up to par. It is hard to pin down the components in this protocol. It may improve patient motivation, reopen neural pathways, help regulate motor homeostasis, or all of the above. John Basmajian, in a pioneering work called *Muscles Alive*, has described this work including the demonstration of conditioning a single motor unit. The feedback appears to potentiate an afferent pathway in a way not possible without it. Within this context, physical and occupational therapists use biofeedback in conjunction with physical hands-on training and stretching techniques (see Fig. 1).

B. Pelvic Floor Muscle Feedback for Incontinence and Vulvodynia

A similar application has strong scientific support: the retraining of pelvic floor muscles using a vaginal or anal EMG sensor. Feedback from the pelvic floor muscles can help the physical therapist and physician diagnose problems such as muscle weakness, spasticity, or poor motor control. Then exercises using visual displays of muscle activity are designed and practiced until mastered. This is another rehabilitation model use of biofeedback, but it appears that the feedback-assisted exercises improve motor control and this leads to a series of positive gains for urinary incontinence or painful intercourse (vulvodynia). A similar application can be used for fecal incontinence. The patient can often regain control of his or her bowel function in a few sessions using an anal probe pressure device with instructions to normalize the pattern of contraction and relaxation of the anal sphincters.

C. Jaw Pain and Dysfunction (Temporomandibular Dysfunction)

Numerous studies have shown that feedback of neuromuscular activity from the masticatory muscles

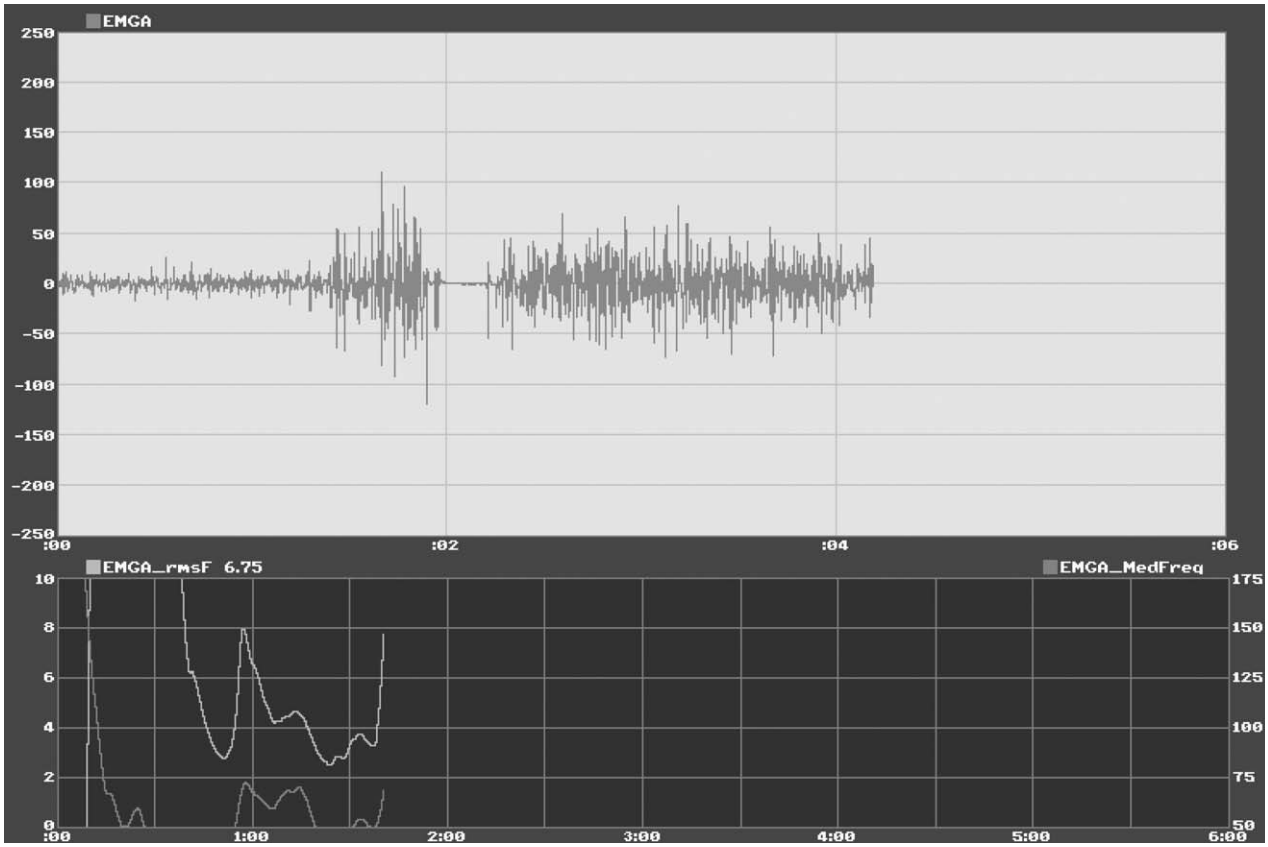


Figure 1 A typical electromyographic (EMG) display with raw activity shown above and RMS integrated levels below (dotted line is a threshold target line).

(jaw, pterygoids, etc.) reduces activity, tension, and pain from these muscles. It is not clear whether the muscle feedback helps reduce muscle activity or simply relaxes the whole muscle system, but the end result is positive. EMG sensors are placed on the jaw muscles and the patient is rewarded for reductions in activity. This is usually accomplished by setting a goal (called a threshold) and supplying feedback (auditory or visual) when the activity in the jaw muscles is reduced below the threshold.

D. Other Muscular Pain Conditions

Biofeedback from other tense muscle sites (low back, head, shoulders, neck) has also been shown to be clinically useful. Typically patients use relaxation techniques in conjunction with EMG feedback to reduce the sensations of pain or tension.

Work by David Hubbard and the author has shown that this is probably due to a reduction in sympathetic flow to muscle spindles rather than to a reduction of voluntary muscle action potentials, but more needs to be done to elucidate the exact mechanism or ''active ingredient'' in the success of EMG biofeedback.

III. AUTONOMIC NERVOUS SYSTEM REGULATION

Perhaps the most paradigm-shifting example of biofeedback comes from the applications to modify autonomic parameters. Until the 1960s it was widely believed that the autonomic nervous system, with its sympathetic branch supporting fight–flight and its parasympathetic branch supporting restoration and homeostasis, perhaps could be modified by Pavlovian or classical conditioning, but not by operant conditioning. It was thought that one might be able to condition a ''reflex'' to a neutral stimulus (bell–meat for a dog) but that the system could not be brought under voluntary control.

In 1969, Neil Miller at Rockefeller University published the results of several studies that showed definitive regulation of autonomic processes such as heart speeding and slowing, urine formation, peripheral blood flow, and blood pressure in lower mammals. The studies were notable because the animals were paralyzed with curare, ruling out any voluntary mediation of the regulation. Miller later showed that quadriplegics suffering from hypotension could raise their blood pressure with a biofeedback protocol in a similar manner. Although the original results have failed to be replicated, the publication of the work produced a shift in paradigm with respect to the autonomic nervous system. Several demonstrations of regulation have been published.

A. Vasodilation and Constriction

It appears that humans (and animals) can regulate vasodilatation and vasoconstriction. Robert Freeman and his colleagues have carefully worked out the mechanisms in finger blood regulation using biofeedback training procedures. First, they demonstrated that volunteers could warm or cool their hands upon command. Then they used a series of pharmacological blockades to show that the ability to cool one's fingers is mediated by the sympathetic nerve, but warming above baseline temperature is blocked by a β-andrenergic blocker in the blood supply. Thus, autonomic voluntary control was demonstrated. Finger warming is a commonly used biofeedback modality.

One particular application of finger warming has a solid scientific basis: Reynauds phenomenon is a multifaceted disorder that produces symptoms of vasospasm in the extremities. Hand warming against a cold challenge appears to be a viable treatment for this serious and sometimes severe problem.

B. Other Autonomic Biofeedback Methods

A few other types of feedback have been reported that seem to indicate learned regulation of the ANS. Some reports of control of sweat gland activity at the extremities appear credible. Hyperhydrosis is a disorder of excessive sweating. Biofeedback is used, with the signal usually taken from the palms of the hands, to reduce excessive activity. Whether the reduction is an example of direct autonomic control or is secondary to learning to dampen the sympathetic response is not known.

Biofeedback of the oculomotor response to improve accommodation has been reported. Special visual devices are used to improve the muscular component of the visual accommodation for improved acuity.

IV. BIOFEEDBACK-ASSISTED CULTIVATED LOW AROUSAL (RELAXATION) PROCEDURES

Biofeedback or applied psychophysiology (AP) has been most frequently used to assist in the attainment of a relaxed physiological and psychological state. Schwartz has labeled the many procedures aimed at accomplishing this task as "cultivated low arousal." The paradigm driving these applications is based on the assumption that many disorders (often called functional disorders) in modern medicine are the result of the sympathetic nervous system and the hypothalamic–pituitary–adrenaline (HPA) system being driven to excessive levels over a long period of time. Disorders included in this list are many: hypertension, headache, irritable bowel syndrome, back pain, asthma, noncardiac chest pain, fibromyalgia, chronic fatigue syndrome, temporomandibular disorder, and perhaps the somatic symptoms of anxiety disorders. Whereas the evidence related to etiology is mixed with regard to most of these disorders, the treatment protocols described next have been fairly successful in reducing or eliminating symptoms.

The biofeedback modalities that have been used include finger temperature warming, muscle activity reductions (usually from the forehead muscles, but also from the trapezii and lower back muscles), respiration training from an abdominal and thoracic strain gauge, skin conductance reduction, and sometimes heart rate reduction. All of these have in common the presumed mechanism of reduction in activity in the SNS with the assumption that the HPA axis will follow. Patients usually report subjective states of low arousal as well. Because these functional or stress-related disorders constitute a large portion of primary medical patients and are accompanied by heavy utilization of medical services, this type of biofeedback application is the most commonly practiced. Most practitioners report good results in 6–10, 1-hr sessions spread out over a few months. The patient is instructed to try to achieve a certain physiological state based on either visual or auditory feedback. Currently, this is delivered on computerized

systems with colorful displays and a variety of available auditory signals. The patient might be instructed to "find a way to warm your hands," or "slow your breath," or "relax your muscles," etc., based on the signals presented.

Because sympathetic activity is not elevated reliably in many of these patients, researchers have been looking for other mediators of symptoms. Respiration has been examined as another alternative to SNS elevation. This work is based on modern knowledge of the respiratory system. Breathing is unique in that it is the one organ system that can be regulated voluntarily but will operate automatically when the organism attends elsewhere. This is probably the reason that meditative traditions throughout human history have incorporated breathing techniques. It appears that humans under threat or stress modify their breathing to a more shallow and rapid style. This depletes the body of CO_2. It turns out that CO_2 regulates the association between the hemoglobin molecule and the oxygen needed by target organs. With prolonged overbreathing and CO_2 depletion, the oxygen is "overbound" to the hemoglobin and therefore unable to adequately fuel the target organ. The resulting anoxia can produce many of the symptoms typical of the disorders listed earlier. For this reason, biofeedback of respiratory parameters is often used to reregulate the gas exchange between oxygen and CO_2. Capnometers that measure expired CO_2 are often used as a feedback device along with indications of breath rate and pattern. Breathing patterns that emphasize slow, diaphragmatic breathing are learned in this way.

V. THE PARASYMPATHETIC BRANCH OF THE AUTONOMIC NERVOUS SYSTEM

Until relatively recently, little attention in biofeedback has been given to the other branch of the ANS, the parasympathetic branch (PNS). Activity from this system is difficult to measure, and it was thought that the sympathetic system was the dominant contributor to stress. However, the evidence to support this idea has not been found. It is difficult to verify sympathetic "overdrive" in the disorders listed earlier. In fact, most patients with functional disorders do not appear to be in obviously stressful circumstances for a sufficient period of time to produce symptoms. This has led some scientists such as Steve Porges at the University of Illinois-Chicago to theorize the importance of the PNS in everyday, nonemergency stress situations. Porges points out that a branch of the Xth cranial nerve (vagus nerve) in the PNS has a distinct anatomy that has evolved in mammals more recently than the more vegatative tracts in the vagal system. This tract seems to be involved in subtle, socially based, heart and lung regulation on an ongoing basis. When a true emergency is perceived, this system moves to the background and allows sympathetic recruitment to occur to meet the challenge. During ordinary experiences when we are negotiating complex social situations, the vagal system is dominant. In biofeedback this can be seen by analyzing the rhythms occurring in the time periods between heart beats. There is a characteristic speeding and slowing of the heart rate with inspiration and expiration called respiratory sinus arrhythmia (RSA). The vagus nerve systematically brakes the heart pacemaker during inspiration and releases the brake during expiration. Biofeedback is based on the attainment of a normal smooth RSA during slow breathing. Patients appear able to learn to produce this pattern by watching their heart rate and respiration on a computer graphic, breathing slowly and effortlessly and finding a "mindful" or blank mental state free of future or past concern, and free of rumination or worry (see Fig. 2). In fact, one group of researchers (Paul Lehrer, Evgeny Vaschillo, and colleagues) has found that 20 min of practice each day at this specific "resonant frequency" improves homeostatic reflexes in the body.

Feedback techniques such as these, which take advantage of newer technologies, may prove to be an important new tool in helping us understand and treat disorders that have puzzled modern medicine.

VI. FEEDBACK FROM BRAIN WAVES

One of the first applications of biofeedback occurred in the 1960s when a great deal of interest in altered states of consciousness arose. A number of researchers in California used newly acquired knowledge from sleep research to study the encephalographic (EEG) aspects of meditative states. It was observed that calm, transcendent states were accompanied by a predominance of α rhythm (8–12 Hz or cycles per second) activity. Biofeedback of this activity was accomplished by providing a tone or other sound when the subject produced a certain amount of the α activity. With shaping, it appeared possible to, in fact, dramatically increase EEG in this band. Many subjects reported meditative types of experiences, and EEG biofeedback

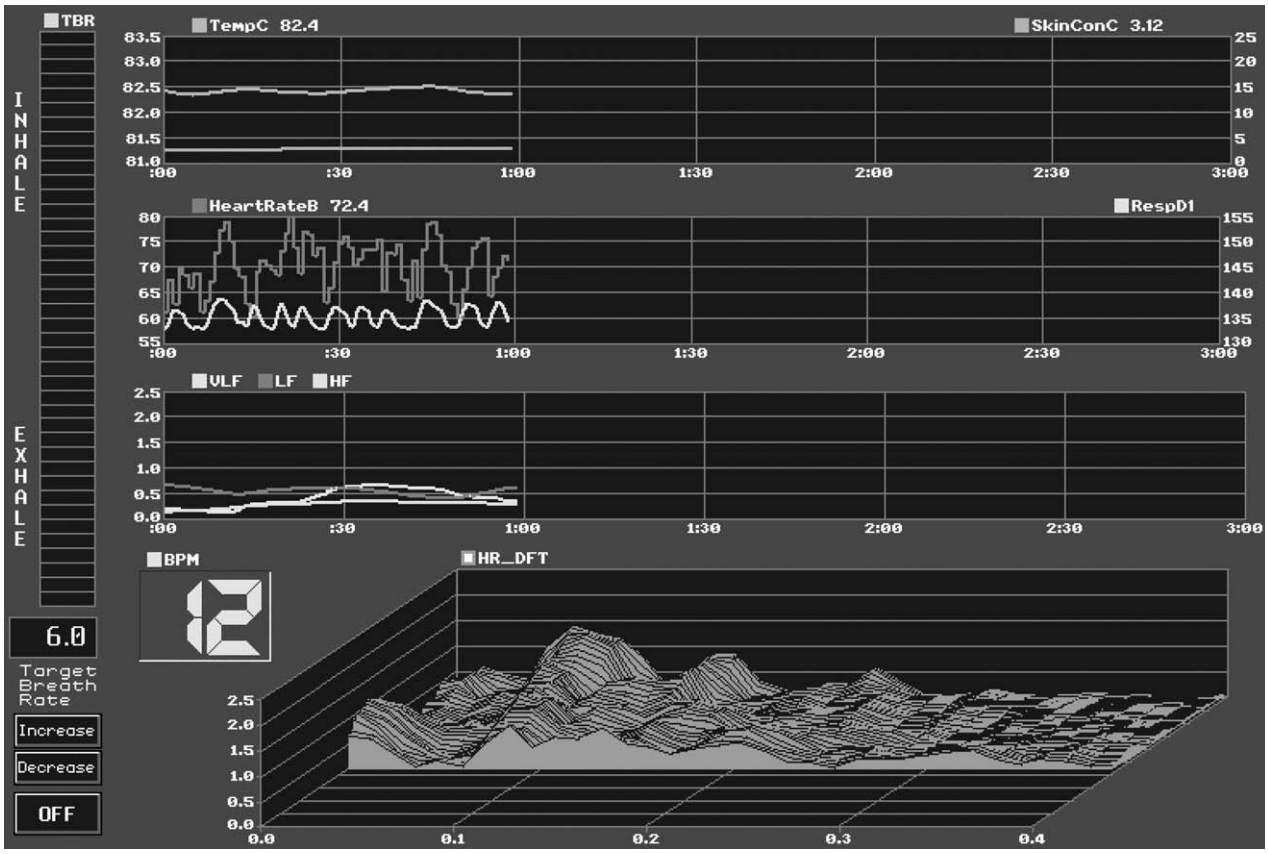


Figure 2 A computerized trace showing (from top to bottom) skin temperature, skin conductance, heart rate, respiration, and spectral derivations of heat rate.

was launched as a number of commercial ventures began producing "Alpha Trainers," promising wondrous results without the time needed to master meditation. Though this trend turned out to be transitory, it did trigger an interest in the use of one's own brain wave information as a way of achieving desired changes. Since then, two applications have established themselves within the field of brain wave biofeedback, often called neurofeedback.

A. Regulation of Brain Activity to Control Siezures

Barry Sterman in Los Angeles and then later Neils Birbaumer in Germany have shown that epileptics whose seizures have not been well-controlled with medication can reduce the frequency and amplitude of their seizures by learning to increase a certain rhythm of EEG over the motor–sensory cortex or by learning to produce a shift in the direct current of part of the brain. As a matter of fact, Birbaumer has convincingly shown that patients who have lost all motor function, including respiration and eye blinks, can learn to communicate by producing a DC shift in brain activity (one patient actually wrote a book this way).

We do not know exactly what mechanisms are involved, but the demonstration of operant control of these rhythms is undeniable.

B. Using Neurofeedback to Change Attentional States

Following the early work on seizures, Joel Lubar and others began to work on a neurofeedback application to help improve the attentional abilities of children and adults with attention deficit disorder (ADD or ADHD). A great deal of research has indicated that ADD is often accompanied by an excess of slow brain waves, especially during less compelling tasks such as

might be found in school. The predominance of θ (4–7 Hz) activity versus β (13–22 Hz) activity at the central vertex of the cranium has been the most common marker. On this basis, θ–β activity is fed back to the ADD participant with rewards for boosting β relative to θ. This is often done using games or other novel stimuli to motivate the learner. Early results indicate that the participants gain attentional abilities that translate into school performance gains. However, the lack of well-controlled trials with placebo or false feedback limits our confidence in these results.

Neurofeedback has become quite popular over the last 5 years and now plays an important role in the field of biofeedback. If research corroborates the impressions of many clinical trials, it could become a common treatment for many prevalent problems.

See Also the Following Articles

AROUSAL • ATTENTION • ELECTROENCEPHALOGRAPHY (EEG) • NEUROFEEDBACK • PSYCHOPHYSIOLOGY • STRESS

Suggested Reading

Cacioppo, J. T., Tassinary, L. G., and Bernston, G. G. (Eds.) (2000). *Handbook of Psychophysiology*, 2nd ed. Cambridge University Press, Cambridge, UK.

Porges, S. (1997) Emotion: An evolutionary byproduct of neural regulation of the autonomic nervous system. In *The Integrative Neurology of Affiliation* (C. S. Carter, I. Lederhendler, and B. Kirkpatrick, Eds.), 807. New York Academy of Science.

Schwartz, M. S. (1995). *Biofeedback: A Practitioners Guide*, 2nd ed. Guilford Press, New York.

Body Perception Disorders

GEORG GOLDENBERG
Krankenhaus München Bogenhausen, Germany

GLOSSARY

apraxia A disorder of action planning that affects object use, performance of meaningful gestures on command, and imitation of gestures.

autotopagnosia Inability to designate body parts on command.

finger agnosia Inability to designate single fingers on command.

heautoscopy Seeing a double of oneself.

hemineglect A failure to attend to one side of space.

optic ataxie Inaccurate reaching for visually presented targets contrasting with accurate reaching for targets from other modalities.

phantom The persistence of proprioceptive sensation from missing or deafferented body parts.

proprioception Somatosensory "selfperception" of configuration and movements of one's body.

The importance of perceiving one's body and of knowing about it appears intuitively evident. This intuition and clinical observations of disturbed body perception following local brain damage led to the proposal that there are brain mechanisms specifically dedicated to body perception and body knowledge. "Body schema" and "body image" have become popular terms for designating mental representations of the spatial structure of our body that have a distinct cerebral substrate and can hence be subject to selective damage. Two kinds of criticism have been raised against this idea: Apparently selective deficits of body perception and knowledge have been criticized as being artificially isolated aspects of more general disorders of perception and knowledge, and a unique mental representation of body has been said to be unable to account for clinical dissociations between different kinds of disturbances of body perception and body knowledge. Whereas the first criticism calls into question the existence of any dedicated mental representation of body, the second posits that there must be more than one of them.

I. INTRODUCTION

In this article, I analyze disorders of human body perception to determine whether perception and knowledge of the spatial structure of one's body have a different neural basis than perception and knowledge of other objects, and whether the neural basis of body perception and knowledge includes a comprehensive "master map" corresponding to the intuitive unity of our body. I first consider the sensory sources, modes, and purposes of body perception.

A. Sensory Sources of Body Perception

Information about the spatial structure and configuration of one's body is provided by vision, somatosensory afferences, and monitoring of motor commands. Vision of one's body enters the brain by the same pathways as vision of any other object. Direct

sight of one's own body is limited to the limbs and the front of the trunk but does not basically differ from vision of other person's bodies. Mirrors, photography, and electronic recording provide ample opportunity for seeing one's own body as completely as it can be seen by anyone else. Vision provides a view of one's own body that is very similar to how it can be seen by external observers.

Somatosensory afferences from cutaneous surface receptors subserve primarily the reaction to and recognition of external stimuli, but tactile sensations can also be analyzed with respect to their localization on the body. One can also touch oneself. Tactual exploration of one's own body differs from exploration of external objects by the coincidence of tactile sensations from touching and touched body parts. This coincidence can serve as a clue for recognizing one's own body as being the object of touch.

In contrast to surface receptors, afferences from mechanoreceptors in deep skin layers, joint capsules, Golgi tendon organs, and muscle spindles are predominantly used for proprioception. Proprioception comes from the interior of the body and differs fundamentally from the way in which one's body can be perceived by another person.

Data from the physiology of motor control and perception strongly suggest that "efference copies" of motor commands can be monitored and compared with proprioceptive feedback from motor execution. In routine behavior we are not aware of motor commands as being different from motor execution, but motor commands may rise to consciousness when there is discrepancy between the command and its execution.

B. Modes of Body Perception and Knowledge

One can walk and at the same time have an intellectually demanding discussion. The motor acts of walking and speaking result in changes in the spatial relationships between body parts. These changes have to be constantly monitored in order to enable adequate chaining of motor acts but should not distract attention from the arguments of the discussion. In other situations—for example, when learning a new dancing step—one may pay attention to the relationships between involved body parts and monitor them consciously, presumably being unable to deal with any intellectual problems simultaneously. There are thus two modes of perceiving and knowing one's own body: One functions in the background of attention and is implicitly used in action control, whereas the other one demands focused attention and gives rise to explicit knowledge about body configurations.

C. Purposes of Body Perception and Knowledge

Intuitively, one may discern three significant purposes of knowledge about one's body. First, it is needed for motor control. Second, it can give rise to conscious awareness and evaluation of the integrity and configuration of one's own body. Third, it can serve social communication and learning. The ability to compare the configuration of one's own body with those of other persons and to adapt one's movements to those shown by other persons is important for skill learning and social communication.

The classification of purposes of body perception will serve as a guideline for the structure of this article. I discuss how disturbed perception or knowledge of one's body can affect each of them, considering experimentally induced disturbances in normal persons and disturbances caused by brain lesions.

II. MOTOR CONTROL

In order to plan the muscular actions for bringing a body part into an intended position, the brain has to be informed about its starting position. Afferent information about the body part's final position is needed to control whether the intended position has been accurately reached.

Demands on information about configuration and position of one's body increase when movements are aimed at external targets. Reaching with the hand for a visually presented object requires a transformation of the target location from retinotopic to body-centered coordinates. This transformation has to take into account the position of the eyes relative to the head and of the head relative to the trunk. Hence, information is needed about the actual position of eyes and head in addition to the initial limb position.

A. Prism Adaptation

In normal subjects, flexibility of transformation from retinotopic to body-centered coordinates has been demonstrated by prism adaptation. When subjects view objects through laterally shifting prisms and try to reach for them, their hands will move toward the shifted location and miss the actual object. Subjects

will see their hand deviating from the object's apparent location even further to the side of prism shift. After some trials, however, the direction of movement will adapt to the visual distortion. Subjects will reach the actual object while seeing their hand reaching its apparently shifted location. The actual movement path of the hand thus deviates from its visually perceived movement path in the opposite direction of prism shift. Consequently, when prisms are removed subjects will initially misreach in this direction. Prism adaptation and aftereffects are restricted to the hand that performed the movements. This means that rescaling affects the transformation from a retinal to a one-limb-centered rather than globally body-centered reference frame.

Prism shift creates several conflicts. First, the visual location of the hand deviates from its proprioceptive location. Second, the visual path of the reaching hand deviates from the intended path. Third, after misreaching the visual location of the hand deviates from the visual location of the object. The motor system could use either of these discrepancies for computing the reaching error and rescaling visuomotor transformation. Subjects' introspections, however, accuse a nonexistent discrepancy between intended movement and proprioceptively signaled movement path of the hand as being the source of errors. They report that their movements aimed at the visually perceived target but were mysteriously pulled to the side of prism shift. At the end of the movement they believed their hands were where they saw them—that is, shifted to the prism-distorted locations. In fact, the hand had faithfully followed the intended path, but vision was distorted. I discuss the phenomenon of visual capture later, but retain now that subjects entertain false beliefs about the perceptual basis of coordinate tranformation rescaling. The brain mechanisms of coordinate transformation seem to work outside the realm of conscious body perception.

B. Somatosensory Deafferentiation

An impressive illustration of the importance of somatosensory proprioception for motor control was provided by two patients who lost all proprioceptive afferences from the body up to the neck as a sequel of peripheral nerve diseases. The most dramatic symptom of complete sensory loss was an inability to make purposeful movements. Lacking somatosensory information about the actual position of moving limbs, the patients produced ill-oriented and ineffective movements of inappropriate strength. Eventually, they learned to replace proprioceptive afferences by visual control of moving body parts. One of them even succeeded in learning to walk again, whereas the other patient remained wheelchair bound. However, movement control remained highly abnormal in both patients. Coordinated movements afforded constant visual monitoring and allocation of attention. Even the ambitious patient who had remastered walking could not entertain a discussion nor admire the beauties of the landscape while walking.

Replacing somatosensory by visual perception of limb position thus seems to be associated with changing the mode of body perception and motor control from one working implicitly in the background of attention to one demanding focalized attention and explicit knowledge about body configurations.

C. Blindtouch

Jaques Paillard proposed to name the ability to point to the location of touch without consciously perceiving that touch "blindtouch." It has been explored in two patients suffering from complete anaesthesia of one half of the body. They were blindfolded and asked to point with their intact hand to the location of touch on their anesthetic hand. The accuracy of pointing was reduced but was definitely above chance, even for discriminations between proximal and distal portions of single fingers. One patient commented, "But I don't understand that! You put something here. I don't feel anything and yet I go there with my finger. How does that happen?" Localization was restricted to direct pointing and did not differ from chance when a patient was asked to indicate the localization of touch by pointing to a picture of a hand or by a selection between verbally proposed localizations. When verbal responses were given simultaneously with direct pointing, the verbal errors captured the motor response and pointing accuracy decreased to chance.

Blindtouch can be considered as being the opposite of movement control in deafferented patients. In blindtouch, the implicit processing of somatosensory information about locations on the body seems to be preserved in the absence of any explicit representation of the same information, whereas in deafferented patients the explicit representation of the body is used for compensating the absence of implicitly processed somatosensory afferences.

D. Optic Ataxia

Patients with optic ataxia cannot accurately reach for visually presented targets. They move their hand to the vicinity of the target and then start searching movements until they hit the target. Visual misreaching contrasts with fast and accurate pointing to parts of own body and to auditorily presented targets. Many patients also have difficulties when asked to explore, compare, and estimate spatial positions without reaching for them, but there is no correlation between the severity of this general visuospatial disorder and the severity of visual misreaching. Single patients with optic ataxia pass all tests of visuospatial estimation and exploration perfectly, and many patients with severe visuospatial problems can accurately reach for visually presented targets.

A plausible interpretation of optic ataxia is that it affects the transformation of retinotopic locations into body-centered reference frames necessary for movement planning, leaving intact motor control not requiring visual input as well as visuospatial processing not requiring transformation from visual to body-centered coordinates.

Optic ataxia can be restricted to one hand or one hemifield or even to a specific hand–hemifield combination, and manual misreaching may contrast with accurate fixation by saccades. These dissociations indicate that transformations from retinotopic to body-centered coordinates are made by mechanisms dedicated to single body parts or sectors of the visual field.

E. Physiology and Anatomy

The majority of fibers carrying information from deep mechanoreceptors end in spinal reflex loops or motor control centers in brain stem and cerebellum. Only a minor portion joins the leminiscal pathway and reaches primary sensory cortex via the ventrolateral thalamus. Both patients who completely lost the ability to exploit somatosensory afferences for motor control had peripheral nerve diseaeses that blocked somatosensory input before it reached even the spinal cord. Of the two patients with blindtouch, one had an extensive contralateral parietal lesion that in all probability destroyed the primary sensory cortex, whereas the other one had a thalamic lesion interrupting all sensory input to primary sensory cortex. Taken together, these observations indicate the extent and importance of sensory motor coordination below the level of cortical processing. The preservation of touch localization in the absence of cortical processing suggests that lower-level topographical body representations suffice for distinguishing locations on the body with above chance accuracy. Conversely, the dramatic effects of peripheral deafferentiation may be due to the deprivation of not only cortical but also lower-level representations of body configuration from proprioceptive afferences.

From primary sensory cortex in the postcentral gyrus proprioceptive input is transmitted rostrally to primary motor cortex and caudally to parietal areas. In the parietal lobe, somatosensory infomation meets visual input that is forwarded from secondary visual cortex along the "dorsal" route of visual processing.

Single-cell recordings in monkey have provided evidence that neuronal networks in parietal areas 5 and 7 are capable of transcoding visual locations from retinotopic to body-centered reference frames. These networks are centered within the intraparietal sulcus, which separates areas 5 and 7. Evidence for a similar function of the human intraparietal sulcus is provided by functional imaging of brain activity in prism adapatation as well as by analysis of lesions in optic ataxia: A positron emission tomography study of normal subjects who were exposed to prism adaptation of one hand demonstrated activation in the contralateral intraparietal sulcus. The lesions causing optic ataxia are also centered around the intraparietal sulcus. Unilateral lesions may cause optic ataxia of the contralateral hand, the contralateral visual field, or both.

The accordance of monkey and human data on the importance of the intraparietal sulcus for visuomotor coordination should not diverge attention from a fundamental difference in the anatomical layout of monkey and human parietal lobes. Human areas 5 and 7 are shifted dorsally to the superior parietal lobe and the intraparietal sulcus separates both of them from the inferior parietal areas 40 on the supramarginal and 39 on the angular gyrus. Lesions causing optic ataxia differ from lesions causing general visuospatial disorders by largely sparing the inferior parietal lobule. I discuss the role of the inferior parietal lobe for body perception later.

F. Representation of One's Body for Motor Control

Perception of body configuration for motor control strongly relies on somatosensory input and predomi-

nantly works outside of attention. It can dissociate from or even contradict the conscious perception of one's body. The central nervous system uses information about body configuration for motor control at multiple levels from spinal cord up to the parietal lobe. Presumably, there are representations of the spatial configuration of body parts at each of them. Evidence from visuomotor coordination suggests that even at the highest level there are multiple task- and body part-specific representations rather than one master map of the entire body.

III. AWARENESS OF ONE'S OWN BODY

Human subjects have a body image depicting both the permanent features and the dynamic configuration of their bodies. They can tell whether any of their body parts are mutilated or even absent, and they can perceive and deliberately control the configuration of moveable body parts.

A. Illusive Body Perceptions

Even in normal subjects the integration of visual perception, somatosensory input, and monitoring of motor intentions into a coherent body image may go astray when there is conflict between the senses.

1. Illusions of Impossible Body Configuration

External vibration of a muscle stimulates deep receptors in much the same way as extension of that muscle does. It therefore gives rise to an illusive feeling of limb movement stretching that muscle. This feeling is strong enough to create a strange sensation of impossible body configurations. For example, vibration of the biceps brachii while the fingers are holding the nose leads to a "Pinocchio effect" of apparent lengthening of one's nose. Conversely, vibrating the triceps brachii creates a sensation of the head being pushed backwards and downwards into the body beyond all limits of anatomical possiblity. Ramachandran and Hirstein produced a Pinnochio effect by seating blindfolded subjects behind another person facing the same direction. The experimenter stood behind the blindfolded subject. With his left hand, he took the blindfolded subject's left index finger and used it to repeatedly tap and stroke the other person's nose while his right hand tapped and stroked the subject's nose in synchrony. After a few seconds, the subjects developed the illusion that their own noses has either been dislocated or stretched out.

Illusions of impossible body configurations demonstrate that conscious body perception will sacrifice the permanence of body shape if it is in conflict with the temporal dynamics of ongoing sensations. The synchrony between holding the nose and stretching the arm or between touching a nose and feeling touch on the nose is most parsimonously accounted for by lengthening of the nose. Finding parsimonious accounts for current events seems to have higher value than fidelity to lifelong experience.

2. Visual Capture

The previous examples concerned conflicts between components of somatosensory afference while visual feedback was excluded. When somatosensory afferences is in conflict with visual perception, vision generally wins. This effect, one example of which is prism adaptation, has been termed "visual capture." Visual capture is not merely due to subjects believing their eyes more than their feeling; if distorts somatic sensations to make them correspond with the visual image. Subjects feel their hand to be as they see it. The subordination of feeling to seeing is even more remarkable because visual perception of one's own body is not very discriminative. It can be deceived by replacing the own hand with a mirror image of the other hand, a gloved hand of another person, a hand made of plaster or rubber, or even a shoe. In 1937, Tastevin reported that subjects who held a 5-cm-wide box beneath a cloth with one hand while they were shown a plaster hand holding a 2-cm-wide box up to 30 cm away from their real hand gradually adapted their proprioception to the artificial display so that they could feel their hand holding the small box in the plaster hand's place. Recently, the credibility of this astonishing report has been endorsed by more rigorously controlled experiments. Botvinick and Cohen simultaneously touched a visible rubber hand and the subjects' hand, which was hidden behind a screen. Subjects developed the illusion that the rubber hand was their own hand, provided that there was strict synchrony between seen and felt touch. Ramachandran seated subjects in front of a table on which a shoe was placed. Subjects saw the experimenter repeatedly and randomly stroke the shoe while their own hand was stroked synchronuously behind a screen. Subjects developed the illusion that their tactile sensations were coming from the shoe, which had become part of their own body.

Monitoring of motor commands shows better resistence against visual capture than somatosensory afferences. This leads to the illusive conflict between motor intention and motor execution, which was discussed previously. The conditions for conflicts between motor intention and visual feedback can be further explored by means of a simple mirror. If the mirror is placed sagitally at body midline and one looks into it from one side, vision of the opposite hand is replaced by the reflected image of the hand on that side. Looking from the left side, one sees a mirror image of the left hand replacing the right hand. If the left hand is moved, a symmetrical movement of the right hand will be shown regardless of what the right hand is actually doing. A compelling feeling of moving the right hand in accordance with its pretended image can occur when the right hand rests immobile, particularly when both hands hold moveable objects of the same shape (e.g., rings pending on a horizontal stick). If, however, the hidden right hand performs active movements incompatible with the mirror image, one feels a conflict between intention and execution similar to the conflict described for prism adaptation rather than an uncontradicted feeling of moving the hand in accordance with its visual appearance.

B. Body Part Phantoms

The occurrence of body part phantoms was among the first and continues to be among the most impressive arguments for the contention that a predetermined body image underlies and modifies the way we experience our own bodies.

The basic experience of body part phantoms is somatosensory: A body part is felt to exist, although it is either absent or cut off from the cerebral cortex by peripheral, spinal, or subcortical interruption of somatosensory afferences. Phantoms have been reported for nearly every part of the body but are most frequent and best explored for limbs and female breasts.

1. Somatosensory Influences on Phantom Sensation

It has long been established that touch of the stump can evoke referred sensations in phantoms of amputated limbs, but only recently it has been shown that referred sensations can originate in body parts that have no anatomical proximity to the amputated body part. In patients with upper limb amputations referred sensations have been evoked from both sides of the chest, both sides of the face, and the contralateral arm. Sensations in phantoms of amputated breasts have been evoked by touch of both sides of the back and the ipsilateral pinna. The presence, extent, and localization of referred sensations vary greatly between patients. In some of them, an exact and reproducible topographical remapping from stimulated to referred locations has been demonstrated that remained stable for up to several weeks. Reexaminations after longer delays, however, have documented radical changes or even complete breakdown of topographical referral without accompanying changes of the phantom's size and shape.

2. Visual Influences on Phantom Experience

Phantoms of amputated body parts clearly contradict the visual perception of their absence and thus challenge the dominance of vision for awareness of one's body, but evidence for an influence of vision on phantom experience derives from patients with spinal cord transsection. Their limbs are anesthetic and plegic but visibly present. Virtually all these patients have phantom experience of the disconnected limbs. During the first days many patients feel the phantom limbs to be flexed although their plegic real limbs are extended, but within a few weeks the phantom position joins the real position. Because the patients' brains have no somatosensory information about the real limb position, they must use vision of the real limbs for adapting the phantoms.

Further evidence for a visual influence on phantom experience may be deduced from adaptation to prostheses. In patients fitted with prostheses, phantoms frequently adapt to their shape. Similar to normal persons who were induced to feel a distant shoe as being part of their bodies, some amputated patients integrate the prosthesis into their bodies and identify it with the phantom. They feel touch directly at the surface of the prosthesis rather than deducing it from the prosthesis' pressure on the stump. Apparently, vision of the prosthesis entertains and shapes the phantom experience and makes possible a meaningful and functionally advantageous stabilization and refinement of sensory remapping from stump to phantom.

Upper limb prostheses seem to have less influence on shape and size of phantoms than do lower limb prostheses, but visual capture of upper limb phantoms has been convincingly demonstrated in a series of elegant experiments by Ramachandran and

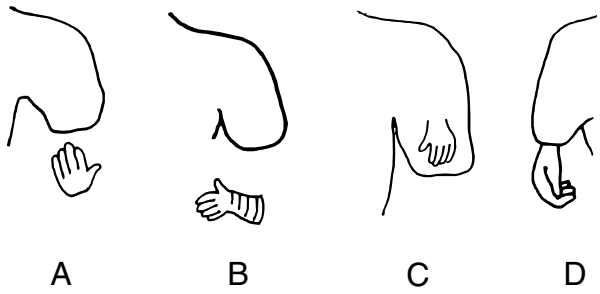


Figure 1 Samples of anatomically impossible or abnormal phantoms after upper elbow amputation. (A and B) The hand is disconnected from the stump. (C) The hand is located within the stump. (D) The extreme shortening of the arm is anatomically possible but corresponds to a deformation that had not been present before amputation (redrawn with permission from sketches made by the affected patients, published in W. E. Haber, Observations on phantom limb phenomena. *Arch. Neurol. Psychiatr.* **75,** 624–636, © 1956, American Medical Asociation).

coworkers. They ''resurrected'' vision of the amputated arm by means of a mirror reflecting the patient's opposite arm or the gloved arm of another person. Movement of the mirror image induced a feeling of phantom movement regardless of whether the patient's or the experimenter's arm had induced the illusion. When the patient's intact arm was touched, patients felt touch at the mirror location on the phantom. This illusion was contingent on the actual application of touch to the intact arm. Seeing touch of the experimenter's hand reflected to the resurrected phantom did not induce a feeling of touch on the phantom. A synchrony between any feeling of touch and vision of phantom touch was necessary for inducing a feeling of touch on the phantom.

3. Phantoms in Congenital Absence of Limbs

The possibility of phantom limbs in persons with congenital absence or very early amputation of limbs has been reliably established, but their frequency is substantially lower than that after later amputation. Permanent phantoms are reported by approximately 10% of persons with congenital absence or very early amputation of limbs compared to approximately 90% of persons amputated after the age of 10 years. The incidence increases to approximately 20% when temporary phantom sensations are considered. Some persons report having had phantoms as long as they can remember, but in the majority phantoms occur only after a delay. The mean time to phantom onset has been calculated to be 9 years in congenital absence and 2.3 years in early amputation. The emergence of the phantom may be triggered by minor trauma to the stump.

Because the affected children had no or only rudimentary opportunity to experience the presence of the now missing limb, the phantom has been said to represent a genetical prefiguration of the mental representation of body shape. However, there is evidence that pre- and postamputation experience can shape phantoms in children. In children with early amputation of congenitally deformed limbs the phantom may replicate the initial deformation rather than restituting a normal limb. This shaping by early experience may contrast with an inability to consciously remember the deformation. Like phantoms of adult patients, those of children can be triggered and shaped by prostheses: Phantoms of congenitally absent or early lost limbs are more frequent in children who have been fitted with prostheses than in those without, and they usually adapt their size and shape to the prosthesis.

4. Anatomically Impossible Phantoms

Over time, the proximal portion of limb phantoms tends to fade. This can lead to the strange sensation of the distal limb being disconnected from but still belonging to the body or to telescoping of the limb (Fig. 1). Telescoping causes a severe deformation of size and shape of phantoms and may result in the anatomically impossible location of fingers inside the stump. Full-sized phantoms may be in unnatural positions violating anatomical constraints. For example, the hand of a phantom arm may penetrate into the chest. There is a report of one girl who, after amputation of her congenitally deformed right leg, developed one phantom adapted to the prosthesis, one reproducing the original deformation, and one consisting of toes fixed to the stump, thus experiencing an anatomically impossible coexistence of three right legs.

5. The Body Representation Underlying Phantoms

The mental representation of one's body underlying the phantom experience seems to be a moldable and fleeting construction aimed at integrating pre- and postmorbid experience, functional affordances, and discrepant sensory afferences. It cannot be reduced to being a replication of a genetically prefigured body image or of the premorbidly intact body shape. It seems doubtful that a genetically predetermined mental representation of normal body shape plays a

role in the genesis of phantoms at all. It is difficult to imagine how a predetermined mental representation can force completion of the body image by phantoms of missing parts but cannot prevent deformation of that image by anatomically impossible phantoms.

C. Personal Hemineglect

Patients with hemineglect fail to direct attention and action toward the hemispace opposite their cerebral lesion. Neglect can affect either side, but it is more frequent, more severe, and more durable after right brain lesions. Hemineglect can disturb attention and action to the left side of extrapersonal space and external objects as well as to the left side of the patient's own body. Patients may fail to wash, comb, shave, and dress the left half of their bodies. They may forget about their frequently plegic left limbs and let them slip into uncomfortable and dangerous positions. When asked to touch their left arm, the reaching movement of the intact right hand may stop at the body midline or at the left shoulder.

It has been suggested that the basic disorder in hemineglect concerns the mental representation of space rather than either perceptual input or motor output. This line of thought leads to the prediction that if there exists a distinct mental represenation of one's own body, hemineglect can be restricted to the patient's own body. Empirical findings on this issue are inconsistent: One study found personal hemineglect without extrapersonal hemineglect in only 1 of 97 right brain-damaged patients, whereas the reverse dissociation occurred in 9 patients. Another study with a smaller sample of patients and using different tests of neglect found that one-third each of patients had predominantly extrapersonal, predominantly personal, and combined hemineglect. These studies were concerned with only relative differences of severity between personal and extrapersonal neglect, but there is one report of a patient showing severe personal hemineglect in testing as well as in spontaneous behavior but no extrapersonal hemineglect.

Although the previous studies invite the conclusion that personal and extrapersonal hemineglect are independent sequels of brain damage that happen to cooccur because of anatomical contiguity of causal lesions, functional interactions between them were demonstrated by varying the relationships between personal and extrapersonal definitions of left and right. In patients with left hemineglect blindfolded detection of touch of the left hand improves when the hand is placed across the body midline into the right hemispace. When the ulnar and radial edge of the left hand are touched simultaneously, neglect will affect the edge that happens to lie on the left side. Ulnar touch will be neglected when the hand is pronated and radial touch is neglected when the hand is supinated. Within a somatotopic representation of one's body, the left hand and its left side would be expected to be coded as being left regardless of their momentary position. The dependence of tactile sensation on hand position suggests an influence of position in peripersonal space on proprioceptive awareness.

In summary, studies of personal hemineglect lend support to the idea that the mental representation of one's own body is distinct from the mental representation of external space but demonstrate close proximity and functional relationships between them.

D. Denial of Ownership of Limbs and Supernumerary Limbs

These delusions are typically though not exclusively observed in patients with left-sided hemiplegia and neglect. Patients may deny ownership of their left arm, claiming instead that the arm to their left belongs to another person or to a dead body. Conversely, they may insist that they have a supernumerary, nonplegic left arm in addition to a plegic one.

Denial of ownership is sometimes embedded in bizarre speculations concerning the origin and purpose of the strange body parts. The basic phenomenon, however, may be traced back to contradictory tactile, proprioceptive, and visual information about the left side of the body. I illustrate this with an example.

I had a conversation with a woman suffering from left-sided hemiplegia, hemianesthesia, and partly recovered hemineglect. During the conversation she recognized her plegic left arm perfectly, but she related an unusual procedure happening to her each night. When preparing her for the night, the nurses would place a strange and lifeless left arm into her bed and ask her to take good care of it. She presumed this to be some new kind of therapy for strengthening her lesioned brain. When asked where her own left arm remained during the night, she guessed that it had always been placed "as it should be." She then described a weird experience that she had the very last night: When grasping the strange arm with her right hand, she felt tension in her left shoulder. She concluded that during that night the nurses must have

fixed the strange arm to her shoulder, perhaps to augment its therapeutic effect. Her bizarre account can be interpreted as an attempt to resolve conflicts between different sources of information about her body: Because of left-sided anesthesia the right-sided sensation of touching an arm was not accompanied by the simultaneous left-sided sensation of being touched and the patient lacked the somatosensory clue to self-touch. When awake in daylight she visually detected the continuity of the touched arm with her own body, but in the dark and with lowered vigilance hemineglect prevented exploration of the strange arm up to its origin at the shoulder. The idea that the strange arm was placed there for therapeutic purposes may have been reinforced by the nurses who probably did not mention that they referred to the patient's own arm when advising her to take good care of the left arm during night. Recovery of proprioception in the left shoulder restituted synchrony between manipulation of the arm and a feeling of being manipulated, but at that stage she preferred to accommodate this isolated piece of counterevidence to the already established belief.

Illusions of supernumerary limbs are regularly associated with a loss of proprioception from the duplicated limb and frequently with the belief that in contrast to the first, plegic limb, the supernumerary limb can actively move. Monitoring of motor commands may play a pivotal role in creating the illusion of having a moving limb supplementing the plegic one. In routine behavior we are not aware of motor commands being different from motor execution. It is only when swift motor execution is prevented that we become aware of the discrepancy between intention and execution by noting a "sense of effort." In hemiplegia, alertness to the discrepancy between command and execution may be affected when sensory disturbances prevent proprioceptive feedback about movement execution from reaching awareness. The patients may react to the lack of an error signal in either of two ways: They may adjust intention to execution and experience a complete inability even to initiate any movement of the affected limb, or they may adjust execution to intention and experience successful execution of the intended movements.

Transient feelings of active movements of a plegic arm are sometimes reported by patients who are perfectly aware of the uniqueness and immobility of their arm. Conversely, some patients experience immobile supernumerary limbs. Illusive movements are not the only factor contributing to supernumerary limbs.

A further contributing factor may be a general incapability to compare and estimate spatial locations. Supernumerary limbs following brain lesions are regularly associated with disturbed localization of stimuli on the body and in external space and may be grossly mislocalized originating, for example, from the middle of the chest or, like a bird's wing, from the back. Reduplication and mislocalization of a part of the patient's body is a manifestation of visuospatial confusion that parallels the patients' distorted copies of complex drawings or three-dimensional constructions. The association between mislocalization of limbs and general visuospatial disturbances is similar to the association between personal and extrapersonal hemineglect and raises the question whether there might be cases in whom mislocalizations are restricted to body parts without affecting stimuli in extrapersonal space.

In summary, I have tried to make the case that the bizarre illusions of having body parts replaced by those of a stranger or of having supernumerary body parts can be traced to failed attempts at reconciling conflicting sensory information about the hemiplegic half of the body. The peculiarity of the resulting beliefs may be due to the sudden occurrence and unfamiliarity of hemiplegia, the distorted nature of incoming information, the absence of adequate exploration of the affected half of the body, the general inability to estimate and integrate spatial relationship, and the adverse influence of brain damage on problem-solving and plausibility control. The observation that these severe disturbances of the body image occur exclusively in combination with disorders of visual exploration or visuospatial processing emphasizes the importance of vision for the perception of one's own body.

E. Heautoscopy

Heautoscopy refers to the strange experience of seeing a double of oneself. The double may attract the person's identity, leading to the weird illusion of being changed into the double and seeing one's left and lifeless former body from outside. The self may vacillate between the two instantations of the same person, possessing each of them in turn. Heautoscopic experiences are often associated with fear and suicidal ideations.

The view of the double may be incomplete, but in most instances it includes the face. The double usually wears clothes that are frequently the same as those that

the person wears when seeing the double. In patients with repeated heautoscopic experiences the dress of the double may thus change clothes between appearances.

Heautoscopic phenomena demonstrate that persons can produce mental images of their own visual appearance. It seems doubtful whether the mental representations underlying these images have much in common with the mental representations involved in phantom experiences and personal neglect. Their emphasis is on clothes and the face much more than on the spatial structure and configuration of the body. In any event, they reflect day-to-day changes of one's visual appearance in addition to permanent knowledge about one's body.

F. Physiology and Anatomy

1. Plasticity of Primary Sensory Cortex after Amputation

Plasticity of somatotopic organization of primary sensory cortex (S1) in patients with phantoms after amputation of the lower arm and hand has been demonstrated by functional imaging with magnetoencephalography. The receptive fields of adjacent regions, devoted to the face on one side and to the upper arm on the other, invade the receptive field originally devoted to the hand. The parallel of remapping of sensations from face and stump to the phantom is striking, but a longitudinal study found that the topography of referred sensations can change without correlated changes in cortical remapping. This finding suggests that a factor besides cortical remapping must contribute to remapping of sensation. Possibly this factor is involved in an interpretative activity of higher brain centers that integrate information from sensory cortex with other sensory afferences to produce a coherent body image.

Cortical remapping in S1 may pose difficulties for the construction of a coherent body image because higher brain centers cannot easily localize the source of sensations signaled from reorganized sensory cortex. Referring them to the phantom is one possibility. Alternatively, they could be correctly assigned to those body parts from which they actually derive. Correct somatotopic assignment of signals from invaded cortex might be used to increase the sensitivity of the stimulated body parts. The reality of such an exploitation of cortical remapping is suggested by observations of increased two-point discrimination on the skin of amputation stumps. Consistency and synchrony with input from other sensory channels as well as possible functional advantages may determine how higher brain centers treat the aberrant signals. Presumably, they will retain and stabilize referral to the phantoms only for sensations from body parts that move in close spatial proximity and temporal synchrony with the phantom. The proximity of face and hand in S1 does not correspond to proximity on the body, whereas the proximity of arm and hand does. Referred sensations from face to hand phantoms are indeed less stable than referred sensations from the amputation stump.

The contention that the sensory homunculus is a source of input to, but not the site of, the neural substrate of a body image is not incompatible with effects of direct stimulation of sensory cortex. In patients in whom phantoms had vanished several years after arm amputation, transcranial magnetic or direct stimulation of primary sensory cortex evoked a resurrection of phantoms. The temporary resurrections of a phantoms in these cases may indicate a rapid reaction of higher brain centers to abnormal signals from primary sensory cortex rather than revealing the cerebral site of the phantom.

2. Parietal Lobe Representations of One's Own Body

The parietal lobes receive afferences from somatosensory and visual cortex, and their close reciprocal connections with premotor cortex enable transmission of information about motor commands. Parietal lobe areas are thus in an optimal position for integrating information from vision, somatosensory afferences, and the motor commands into a coherent image of the body. Clinical evidence confirms this expectation.

There are a few cases in which a cerebral lesion abolished a phantom limb. In all of them, clinical evidence pointed to lesion in the parietal lobe of the opposite hemisphere, and in one of them autopsy confirmed a metastasis in the supramarginal gyrus. In a further case clinical evidence strongly suggestsed sparing of primary sensory cortex. A functional magnetic resonance imaging study found activations in inferior parietal and premotor but not in primary motor or sensory cortex of a patient with phantoms of congenitally absent limbs when she moved her phantom arms.

Lesions at different locations within the right hemisphere can cause hemineglect and visuospatial impairment, but lesions causing severe and permanent neglect and visuospatial impairment usually involve the inferior parietal lobe. A group study found that lesions

that cause personal and extrapersonal neglect are larger than lesions that cause only extrapersonal neglect, but they are equally centered on the inferior parietal lobe, and the patient in whom isolated personal neglect was diagnosed also had a parietal lesion.

In contrast to body perception for visuomotor coordination, perception for awareness seems to depend on integrity of inferior parietal areas 40 and 39 rather than areas 5 and 7. Areas 40 and 39 are linked to the ventral stream of visual processing dedicated to object recognition, and it has been suggested that they build an explicit allocentric representation of space rather than coding egocentric spatial locations for motor control.

3. Laterality of Body Representations

Abolishment of phantoms after parietal lesions was observed with right-sided and left-sided phantoms and was always caused by lesions in the opposite hemisphere. There are no significant differences between the frequencies of left- and right-sided limb or breast phantoms. It can be concluded that each parietal lobe constructs a representation of its opposite hemibody and that both have a similar propensity to replace lost body parts by phantoms. The restriction of disturbed body perception to the opposite half of the body also applies to hemineglect, denial of ownership, and supernumerary limbs, but the distribution of these symptoms is asymmetric. They are mainly caused by right parietal lesions and affect the left half of the body. A possible explanation for the asymmetry is that these symptoms result from an interaction between general visuospatial disorders and disturbed body perception, and that general visuospatial impairment is more severe after right brain damage.

4. The Neural Substrate of Heautoscopic Hallucinations

Heautoscopic hallucinations are associated with temporooccipital rather than parietal lesions. Their brief duration and their occurrence in patients with epileptic seizures suggest that heautoscopic hallucinations are manifestations of partial complex seizures originating from the temporal lobes. Their association with fear and suicidal ideation might be related to involvement of limbic structures in the anterior and mesial temporal lobe. A further clue to the temporal lobe origin of heautoscopy is provided by the frequent inclusion of the face into the heautoscopic hallucination. There is ample evidence that recognition of faces is bound to temporal rather than parietal lobe function.

Heautoscopic hallucinations can be restricted to the visual hemifield opposite the lesion without a clear-cut difference between the right and left sides, but the lateralized hallucinations always include both sides. Either each temporal lobe can produce the visual image of a symmetric body or both collaborate in creating lateralized heautoscopic hallucinations.

G. The Conscious Representation of One's Own Body

The conscious image of one's own body is not necessarily a faithful replication of its actual structure and configuration, nor can deviations from reality be explained as resurrections of a prefigured image of one's body as it should be. It is rather like a fleeting reconstruction aimed at integrating afferences from vision, proprioception, and monitoring of motor commands into a coherent spatial structure. Temporal coherence seems to be a major clue for inferring spatial coherence, and synchronous sensations are assumed to stem from the same location. If synchronous sensations carry different spatial information, the brain uses heuristics for weighting them. Generally, proprioception is assimilated to vision. The subordination of proprioception to vision contrasts with the dominant role of proprioception in body perception for motor control. Monitoring of motor commands resists visual capture but cannot liberate proprioception from being subdued. The conflict between motor commands and vision is therefore experienced as opposing motor commands to proprioception of motor execution.

The construction of a conscious representation of one's own body seems to depend on integrity of the inferior parietal lobes, thus clearly differing from the neural basis of body perception for motor control. Each parietal lobe entertains an image of only the opposite half of the body. The splitting of the neural basis of conscious body representation contrasts with the feeling of a seamless corporal unity.

IV. IMITATION OF BODY CONFIGURATIONS

Imitation is the motor replication of a visually perceived action. The translation of a perceived action into an executed action requires more than a translation from visual percepts to motor commands. It

requires that an action performed by someone else be translated into an equivalent action by oneself. To do this, one must create equivalence between another person's and one's own actions and hence between another person's and one's own body. The need to recognize the communality between another person's and one's own body becomes particularly salient when imitation is examined for pointing to body parts or for meaningless gestures. Because these gestures have neither a conventional shape nor an external referent, they are defined solely as a particular configuration of the body.

In this section, I discuss autotopagnosia, finger agnosia, and defective imitation of meaningless gestures in apraxia. Autotopagnosia and finger agnosia affect the ability to select single body parts corresponding to those shown on a model. Imitation of meaningless gestures tests the ability to translate a configuration of several body parts from a model to one's own body. Although caused by unilateral brain damage, all these disorders affect both sides of the body equally.

A. Autotopagnosia

In 1922, Alois Pick reported on a demented patient who could not point to body parts on command. She searched for her eyes sometimes around her head and at other times on her hands, on the examiner's head, under the bedclothes, or on the ceiling. She also grossly misreached when trying to grasp external objects. Pick nonetheless concluded that this patient had a specific inability to locate parts of her own body and proposed the term ''autotopagnosia'' to designate this disorder.

Modern neuropsychology has retained the idea and the name but has identified it with a much more restricted clinical disturbance: Patients with autotopagnosia commit errors when asked to point to body parts on themselves, on another person, or on a model of the human body. The majority of these errors are contiguity errors in which the patient searches for the body part in its vicinity (e.g., for the wrist on the lower arm). Less frequent is confusion of body parts with similar functions (e.g., wrist and ankle). The error types may combine to lead the patient's pointing response to the most proximate body part with similar function (e.g., to elbow for wrist). Errors affect mainly body parts with ill-defined and inconspicuous borders (e.g., the cheek) and body parts that share many functional properties with similar body parts (e.g., the articulations of the limbs). Localization of individual fingers was repeatedly found to be preserved in autotopagnos ia.

Errors occur not only when the body parts are designated by verbal command but also when they are shown on pictures or when the examiner demonstrates correct pointing and the patient tries to imitate. In contrast, patients are able to name the same body parts when they are pointed to either on their own body or on a model, and they may be able to select a named body part when given an array of drawings of isolated body parts. Nor is there a general inability to point to the parts of their own body: They identify them accurately when asked to indicate the typical location of accessories (e.g., a wristwatch) or the location of objects that had been temporarily fixed to a body part. Some patients were asked to verbally describe body parts. They could describe their functions and the visual appearance but could not relate their location.

In summary, autotopagnosia selectively affects the ability to locate parts of the human body in response to a request that explicitly asks for the locations of the body parts. It has been argued that the disorder is not confined to the human body but also affects pointing to single parts of other multipart objects such as bicycles; however, since this proposal was published several reports have documented patients with autotopagnosia who could locate parts of other multipart objects. The conclusion that the disturbance is restricted to the topography of only the human body is not straightforward. It is questionable whether the structure of bicycles involves a similar amount of subtle distinctions between easily confusable and adjacent parts of the structure of the human body. If such distinctions exist, their cognizance is reserved to experts and is outside the scope of neuropsychological examination. In contrast, subtle distinctions between easily confusable parts of the human body are tested in autotopagnosia and account for the majority of errors. Knowledge about the structure of the human body may be more vulnerable to brain damage because it is more fine-grained and diversified than knowledge about other multipart objects. Few persons have expert knowledge about bicycles, but all have expert knowledge about the human body.

B. Finger Agnosia

In 1924, Joseph Gerstmann reported on a patient in whom an inability to select fingers of either the patient's own hand or the examiner's hand on verbal

command contrasted with otherwise normal language comprehension and with correct pointing to other body parts. Gerstmann noted that the difficulties were most marked for discrimination between the second, third, and fourth fingers. He named the disturbance "finger agnosia," which he considered to be a partial form of autotopagnosia restricted to only the fingers. Gerstmann's patient also had difficulties with writing, calculation, and right–left discrimination. Having observed the same combination repeatedly, Gerstmann speculated that calculation was affected because children learn arithmetic by counting their fingers, writing was affected because it demands differential finger movements, and right–left discrimination was affected because it is made with reference to the hands. None of these speculations are very convincing, and later group studies indeed showed that the correlations of finger agnosia to disorders of writing, calculation, and right–left discrimination were lower than or equal to those of other neuropsychological symptoms that had not been considered as being part of the Gerstmann syndrome.

Like autotopagnosia, finger agnosia is not restricted to verbal testing. Patients also confuse fingers when asked to show on a diagram which of their fingers has been touched or to move a finger as shown on a diagram. Finger agnosia can occur in combination with autotopagnosia but also without it. Because there are cases of autotopagnosia without finger agnosia, it appears that autotopagnosia and finger agnosia represent independent disorders of body part localization. Another association has hitherto been examined only in a few patients and may turn out to be more typical: These patients had similar difficulties with selection of toes as with selection of fingers.

C. Imitation of Meaningless Gestures

Defective imitation of meaningless gestures has traditionally been considered a symptom of apraxia. Other symptoms of apraxia are disturbed production and imitation of meaningful gestures, such as waving goodbye or miming the use of a hammer, and disturbed use of real objects. However, there are, patients with pure "visuoimitative apraxia" in whom defective imitation of meaningless gestures contrasts with preserved production and imitation of meaningful gestures and object use. It thus seems justified to discuss defective imitation of meaningless gestures as a separate disorder.

Defective imitation of meaningless gestures affects not only the translation of gestures from a model to the patient's own body but also translation to other human bodies. Patients who commit errors when imitating gestures also commit errors when asked to replicate a gesture on a mannikin or to select a gesture from an array of photographs showing gestures performed by different persons and shown at different angles.

Similar to the dissociation between autotopagnosia and finger agnosia, defective imitation of meaningless gestures can affect imitation of gestures defined by proximal body parts differently from finger

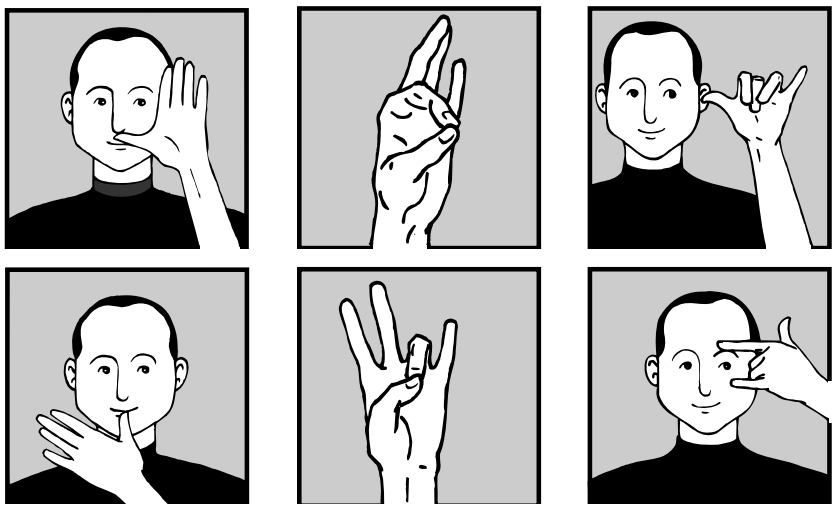


Figure 2 Examples of hand postures, finger postures, and combined gestures (reproduced with permission from G. Goldenberg, Defective imitation of gestures in patients with damage in the left or right hemisphere. *J. Neurol. Neurosurg. Psychiatr.* **61,** 176–180, 1996).

configurations. Figure 2 shows three types of meaningless gestures that have been used to explore these differences: Hand postures specify a position of the hand relative to face and head whereas the internal configuration of the hand remains invariant. Finger postures specify different configurations of the fingers, whereas the position of the hand is not considered for scoring. Combined gestures specify the position of the hand relative to the body as well as the configuration of the fingers. Patients with apraxia have problems with all three kinds of gestures, but the impairment is more severe for hand than for finger postures and more frequently concerns the hand position rather than the finger configuration of combined gestures. There are even single apraxic patients in whom defective imitation of hand postures contrasts with normal imitation of finger configurations. In contrast, patients with visuospatial impairment following right brain damage have problems only with finger configurations.

D. Physiology and Anatomy

''Mirror'' neurons in premotor cortex of the macaque monkey become active when the monkey performs an action and when it observes a similar action performed by the experimenter. The triggering of mirror neurons requires interaction between the agent and the object of an action. When the same movements are performed without the appropriate object or when the object is displayed without an action, the neurons do not react. Mirror neurons may subserve recognition of biologically significant actions. Their existence indicates that the monkey appreciates the communality between actions executed by another subject and by itself and thus possesses an essential prerequisite for imitation, but imitation is not within the behavioral repertoire of these monkeys. It would be a bold speculation to consider mirror neurons as correlates to the neural substrate of human imitation of body-centered gestures.

Lesion data from patients with autotopagnosia, finger agnosia, and defective imitation of meaningless gestures tell a rather straightforward story about the cerebral basis of body perception for imitation. The lesions in cases of pure autotopagnosia are remarkably uniform. They always affect the left inferior parietal lobe. That pointing to proximal body parts depends on integrity of the left hemisphere was confirmed by group studies of patients with left or right brain lesions. Regardless of whether the body parts were designated verbally or nonverbally, only patients with left brain damage committed errrors in pointing to them. In contrast, errors in selecting fingers have been found with about equal frequency in patients with left or right brain damage.

Patients with pure visuoimitative apraxia had either parietal lobe degeneration or vascular lesions in the left inferior parietal lobe. The combination of autotopagnosia and visuoimitative apraxia following left parietal lesions has been documented. The difference between the cerebral substrates of autotopagnosia and finger agnosia is paralleled by a difference between the cerebral substrates of disturbed imitation of hand and finger postures. Imitation of hand postures is disturbed exclusively in patients with left brain damage, whereas imitation of finger postures can be impaired in left and in right brain-damaged patients.

Previously, I identified the inferior parietal lobes as the neural substrate of conscious perception of the opposite half of one's own body. I now ascribe to the left inferior parietal lobe a central role for imitation on both sides of the body. It will be an interesting question for further research whether there is a difference in location between both kinds of body representations within the left inferior parietal lobe.

E. Creating Equivalence between Bodies

The observation that patients with autotopagnosia, finger agnosia, and defective imitation of meaningless gestures commit errors when trying to replicate the demonstrated body configuration on an external model indicates that their difficulties concern the perception and representation of human bodies in general rather than of only their own bodies. This generality is not surprising if one considers that imitation *per se* requires recognition of equivalence between one's own body and that of another person. It would be surprising if a representation of body that applies to two arbitrarily paired persons will not be valid for any other instances of human bodies as well.

A feasible way to create equivalence between human bodies would be to code their configurations with reference to conceptual knowledge classifying significant body parts and specifying the boundaries that define them. Application of this knowledge reduces the multiple visual features of a demonstrated gesture to simple relationships between a limited number of significant body parts. Coding a gesture's visual appearance by classification of body parts produces an equivalence between demonstration and imitation that is independent of the particular angle of view under

which the demonstration is perceived and also independent of accidental minor differences of the exact shapes of demonstrated and imitated body configurations.

The requirements for body part coding of the demonstrated gesture may be different for proximal body parts and the fingers. There are a considerable number and large diversity of proximal body parts, such as forehead, eyebrows, eyes, nose, cheeks, lips, and chin on the face, or shoulder, upper arm, elbow, lower arm, wrist, back, and palm of the hand on the upper extremity. Knowledge about the classification and boundaries of many of these parts is needed for conceptual mediation of proximally defined body configurations. In contrast, the fingers are a uniform set of only five body parts, and the conceptual distinction between them can be reduced to an appreciation of their serial position. Discrimination of proximal body parts may therefore tax conceptual knowledge about the structure of the human body more than discrimination of fingers. Selection of fingers may pose particular difficulties for visuospatial analysis preceding body part coding. The shapes of fingers two, three, and four are nearly identical, and their identity is mainly determined by their spatial position with respect to the other fingers. A careful analysis of their spatial position is necessary for discrimination. A distinction between the index and the middle finger is likely to put higher demand on visuospatial exploration than a distinction between the lips and the chin.

Left and right brain damage may interfere with imitation at different levels: Right brain damage impairs visual exploration and visual analysis of spatial relationships. It affects the discrimination of the spatial position of fingers but not the recognition of perceptually salient proximal body parts. Left inferior parietal damage abolishes body part coding. Selection of fingers can be spared because it puts lower demands on knowledge about body parts and because a direct mapping between the visuospatial configurations of the model's and one's own fingers can partially circumvent body part coding.

Imitation of body configurations is thus subserved by representations that apply equally to one's own and to other persons' bodies and that are not restricted to only one half of the body. Nonetheless, there is no unique neural substrate for all aspects of body configurations. Although the left hemisphere affords classification of body parts, a right hemisphere contribution is needed for spatial discrimination of perceptually similar body parts. A complete representation of the configuration of all parts of the human body requires both aspects and hence involves at least two anatomically separated parts of the brain.

V. CONCLUSIONS

I began this article by postulating three different purposes of body perception and representation: motor control, awareness of the configuration of one's own body, and imitation of the configuration of other persons' bodies. The data support the validity of this classification by showing that the nervous system employs different sensory channels and different central representations for each of these purposes. Perception of one and the same body state can yield different and even contradictory results in representations at different levels of the central nervous system.

There is no such thing as a master map of one's body in the brain. Not only do the neural mechanisms subserving each of the purposes of body perception differ but also there is more than one representation of the body involved in each of them. Most, if not all, of these representations are incomplete and devoted to only limited portions or limited aspects of the body. Some of them seem to be specific to body perception, but others are shared with perception. and representation of other objects. The intuitively given unity and uniqueness of our bodies does not seem to correspond with the neural substrates of body perception.

See Also the Following Articles

AGNOSIA • APRAXIA • CONSCIOUSNESS • HALLUCINATIONS • HAND MOVEMENTS • MOTOR CONTROL • OBJECT PERCEPTION • PHANTOM LIMB PAIN • SPATIAL COGNITION • TACTILE PERCEPTION

Suggested Reading

Berlucchi, G., and Aglioti, S. (1997). The body in the brain: Neural bases of corporeal awareness. *Trends Neurosci.* **20,** 560–564.

Brugger, P., Regard, M., and Landis, T. (1997). Illusory reduplication of one's own body: Phenomenology and classification of autoscopic phenomena. *Cognitive Neuropsychiatr.* **2,** 19–38.

Cole, J., and Paillard, J. (1995). Living without touch and peripheral information about body position and movement: Studies with deafferented subjects. In *The Body and the Self* (J. L. Bermudez, A. Marcel, and N. Eilan, Eds.), pp. 245–266. MIT Press, Cambridge, MA.

Denes, G. (1990). Disorders of body awareness and body knowledge. In *Handbook of Neuropsychology Volume 2* (F. Boller and J. Grafman, Eds.), pp. 207–228. Elsevier, New York.

Frederiks, J. A. M. (1985). Phantom limb and phantom limb pain. In *Handbook of Neurology Vol. 1* (J. A. M. Frederiks, Ed.), pp. 395–404. Elsevier, Amsterdam.

Goldenberg, G. (1999). Matching and imitation of hand and finger postures in patients with damage in the left or right hemisphere. *Neuropsychologia* **37,** 559–566.

Jeannerod, M. (1997). *The Cognitive Neuroscience of Action*. Blackwell, Cambridge, MA.

Lackner, J. R. (1988). Some proprioceptive influences on the perceptual representation of body shape and orientation. *Brain* **111,** 281–297.

Meltzoff, A. N., and Moore, M. K. (1997). Explaining facial imitation: A theoretical model. *Early Dev. Parenting* **6,** 179–192.

Melzack, R., Israel, R., Lacroix, R., and Schultz, G. (1997). Phantom limbs in people with congenital limb deficiency or amputation in early childhood. *Brain* **120,** 1603–1620.

Ramachandran, V. S., and Hirstein, W. (1998). The perception of phantom limbs—The D. O. Hebb lecture. *Brain* **121,** 1603–1630.

Rock, I., and Harris, C. S. (1967). Vision and touch. *Sci. Am*. **217,** 96–104.

Borna Disease Virus

W. IAN LIPKIN, THOMAS BRIESE, and MADY HORNIG
University of California, Irvine

GLOSSARY

dopamine A neurotransmitter important in modulation of movement and behavior.

major histocompatability complex Antigen presenting proteins expressed on the surface of most cells; specific immunity to viruses is dependent on recognition of non self (viral) peptide fragments presented in the context of these proteins.

negative-strand RNA virus A virus with a genome composed of ribonucleic acid that serves as a template for transcription of messenger RNA-encoding viral proteins. Its minimal infectious unit is composed of the viral genome, nucleoprotein, phosphoprotein, and polymerase.

reverse transcription polymerase chain reaction A method for logarithmically amplifying ribonucleic acid sequences to facilitate cloning and/or detection of targets present only at low frequency.

Borna disease virus (BDV) is a novel neurotropic RNA virus that infects a wide variety of warm-blooded animal species. Infection may be asymptomatic or result in a broad spectrum of behavioral disorders, dependent on the age of the host and its immune response. Although the question of human BDV infection remains to be resolved, burgeoning interest in this unique pathogen has provided tools for exploring the pharmacology and neurochemistry of neuropsychiatric disorders potentially linked to infection. Analysis of rodent models of infection has yielded insights into mechanisms by which neurotropic agents and/or immune factors may impact developing or mature central nervous system circuitry to effect complex disturbances in movement and behavior.

I. INTRODUCTION

Borna disease virus (BDV) is the prototype genus (bornavirus) of the family *Bornaviridae*, within the nonsegmented, negative-strand (−strand) RNA viruses (order *Mononegavirales*). The name Borna is derived from its association with the city of Borna, Germany, where an equine epidemic in the late 1800s crippled the Prussian cavalry. This neurotropic virus appears to be distributed worldwide and has potential to infect most, if not all, warm-blooded hosts. BDV is similar in genomic organization to other nonsegmented, (-) RNA viruses; however, its approximately 9-kb genome is smaller than those of *Rhabdoviridae* (about 11–15 kb), *Paramyxoviridae* (about 16 kb), or *Filoviridae* (about 19 kb). The BDV genome is remarkably compact and encodes six major open reading frames (ORFs) in three transcription units (Fig. 1). BDV is distinctive in its nuclear localization of replication and transcription. Although this feature is shared with the plant nucleorhabdoviruses, it is unique among nonsegmented, negative-strand animal RNA viruses. The first transcription unit reveals a simple pattern with only one coded protein [nucleoprotein (N), p40]. The second transcription unit encodes the proteins X (p10) and P (phosphoprotein, p23) in overlapping ORFs. The third unit contains coding sequences for the putative matrix protein (M), type I membrane

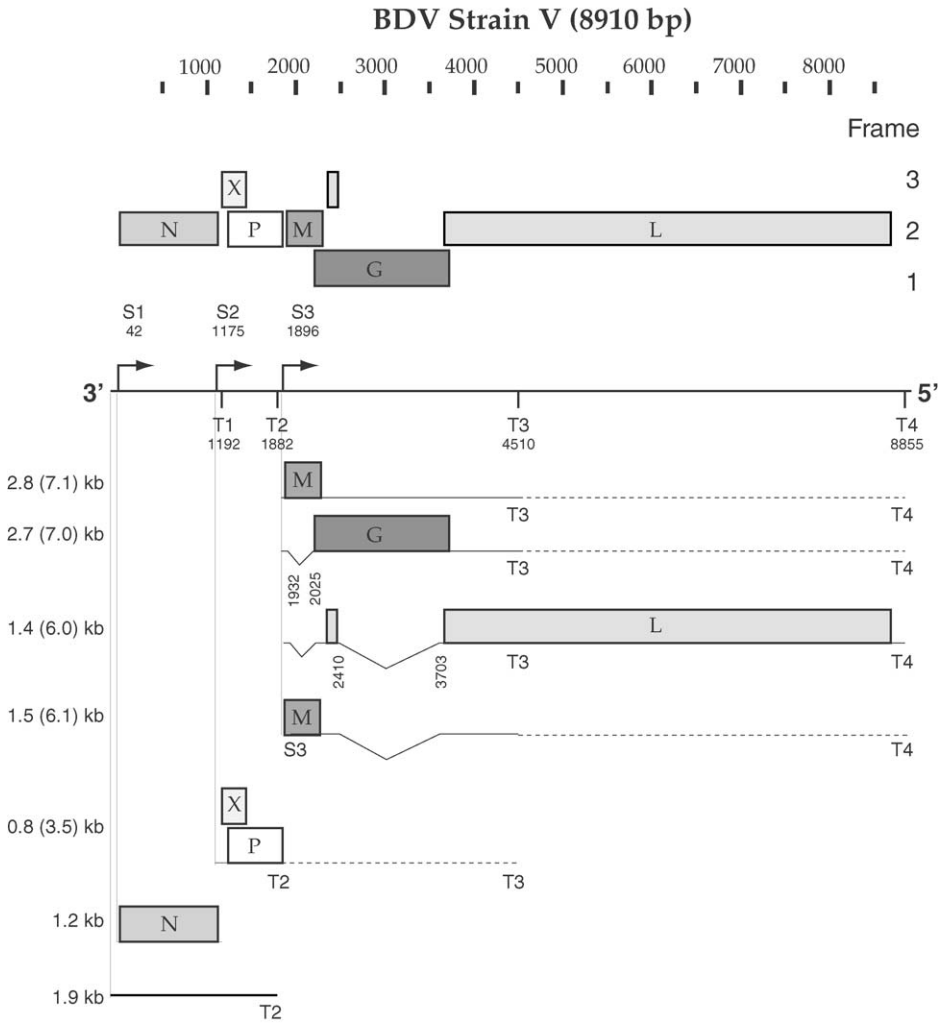


Figure 1 BDV genomic map and transcripts. S1–S3, initiation sites of transcription; T1–T4, termination sites of transcription. Readthrough at termination signals T2 and T3 is indicated by dashed lines.

glycoprotein (G, p57, gp94), and polymerase (L, p190). Elaboration of these proteins is dependent on a variety of mechanisms for transcriptional, posttranscriptional, and translational control of expression, including alternative transcriptional initiation, readthrough of termination signals, alternative splicing, and leaky ribosomal scanning. Although splicing is also found in *Orthomyxoviridae* (segmented, negative-strand RNA viruses), it is unprecedented in *Mononegavirales*.

Virions are stable at 37°C and lose only minimal infectivity after 24 hr of incubation in the presence of serum. Virus infectivity is rapidly lost by heat treatment at 56°C and exposure to pH 5.0, organic solvents, detergents, chlorine, formaldehyde, or UV radiation.

II. EPIDEMIOLOGY OF BDV

A syndrome of progressive meningoencephalitis of horses and sheep consistent with BDV infection was recognized 100 years before the disease received its name from the equine outbreak in Borna, Germany. This syndrome is still considered to represent classical Borna disease; however, infection may also result in asymptomatic carrier status or subtle disturbances in learning and memory, movement, and behavior. Emerging epidemiologic data, including reports of asymptomatic, naturally infected animals, suggest that the host range and geographic distribution of BDV are larger than previously appreciated (Table I). Natural infection has been reported in a wide variety of hosts,

Table I
Animal Species Reported to Be Infected with BDV[a]

Host and country	Seropositive	Viral protein/RNA	Seropositive and viral protein/RNA	Reference
Horse normal USA	3–6% (8–18/295)			Kao *et al*. (1993). *Vet. Rec*. **132,** 241
Horse normal (PBMC) Japan	26% (15/57)	RNA 30% (17/57)	18% (10/57)	Nakamura *et al*. (1995). *Vaccine* **13,** 1076
Horse CNS disease (B) Germany	38% (3/8)	RNA 100% (8/8)	38% (3/8)	Zimmermann *et al*. (1994). *J. Virol. Methods* **46,** 133
Horse CNS disease (B) Germany	100% (4/4)	RNA 100% (4/4)	100% (4/4)	Binz *et al*. (1994). *Virus Res*. **34,** 281
Donkey CNS disease (B) Germany	1/2	RNA 1/2	1/2	Zimmermann *et al*. (1994)
Donkey CNS disease (B) Germany	1/1	RNA 1/1	1/1	Binz *et al*. (1994)
Cattle normal (PBMC) Japan	20% (15/74)	RNA 11% (8/74)		Hagiwara *et al*. (1996). *Med. Microbiol. Immunol. B* **185,** 145
Sheep CNS disease (B) Germany	1/2	RNA 2/2	1/2	Zimmermann *et al*. (1994)
Sheep CNS disease (B) Germany	1/1	RNA 1/1	1/1	Binz *et al*. (1994)
Cat CNS disease Germany	7% (12/173)			Lundgren *et al*. (1993)
Cat CNS disease Sweden	46% (11/24)			Lundgren *et al*. (1993)
Cat Various (PBMC) Japan	8% (7/83)	RNA 13% (11/83)	0% (0/83)	Nakamura *et al*. (1996). *J. Clin. Microbiol*. **34,** 188
Cat CNS disease United Kingdom		RNA 33% (5/15)		Reeves *et al*. (1998). *Vet. Rec*. **143,** 523
Lynx CNS disease (B) Sweden		Protein/RNA 1/1		Degiorgis *et al*. (2000). *J. Clin. Microbiol*. **38,** 3087
Dog CNS disease (B) Austria		Protein/RNA 1/1		Weissenböck *et al*. (1998). *J. Clin. Microbiol*. **36,** 2127
Ostrich CNS disease (B) Israel		Protein 7/13		Malkinson *et al*. (1993). *Vet. Rec*. **133,** 304

[a]Tissue examined for protein/RNA. PBMC, peripheral blood mononuclear cells; B, brain tissue; CNS, central nervous system.

including horses, donkeys, sheep, cattle, dogs, cats, rabbits, and birds. Experimental infection has been achieved in many of these species and also in rodents and primates. Whether BDV naturally infects humans remains controversial; however, there is consensus that all warm-blooded animals are likely to be susceptible to infection. Although central Europe has the highest reported prevalence of Borna disease, natural infection without disease has been described throughout Europe, Asia, and North America. It is unclear whether the apparent increase in host and geographic range of BDV is due to the spread of the virus or enhanced case ascertainment.

Neither the reservoir nor the mode of transmission of natural infection are known. Although experimental animals are most frequently infected by either intracranial or intranasal inoculation, infection can be achieved by virtually any method of parenteral inoculation. Natural infection of horses and sheep is typically sporadic and peaks in spring months; epidemics of disease are infrequent. An olfactory route for natural transmission has been proposed because intranasal infection is efficient and the olfactory bulbs of naturally infected horses show inflammation and edema early in the course of disease. One outbreak in a zoo has been attributed to inadvertent inoculation of BDV. Experimental infection of neonatal rats results in virus persistence and is associated with the presence of viral gene products in saliva, urine, and feces. Such secreta/excreta are known to be important in aerosol

transmission of other pathogenic viruses (e.g., lymphocytic choriomeningitis virus and hantaviruses). Normal adult rats housed in cages separate but adjacent to those of neonatally infected rats can become infected, suggesting that aerosol transmission of BDV is plausible. The observations that rodents can be persistently infected with BDV and excrete virus suggest that they have the potential to serve as both natural reservoirs and vectors for virus dissemination. However, the significance of rodents for transmission of BDV to domesticated animals and humans is unknown. Reports of BDV nucleic acids and proteins in peripheral blood mononuclear cells (PBMCs) also indicate a potential for hematogenous transmission. Vertical transmission of BDV has been reported in horses; this appears to be an infrequent event.

III. NATURAL INFECTIONS

Although infection may be asymptomatic, symptomatic disease typically follows a predictable course. Clinical signs at the onset of disease in horses and sheep are nonspecific: excited or depressed behavior, hyperthermia, anorexia, jaundice, constipation, and colic. Classical disease becomes apparent within 1 or 2 weeks. Animals maintain an upright, wide-based stance with their heads extended. Repetitive behaviors are common and may include vacuous chewing, circular ambulation, and running into obstacles. Horses become paretic in the terminal phases of disease. A distinctive decubitus posture associated with paddling movements of the legs has been described. Frequently, in late disease, the virus migrates centrifugally along the optic nerve to cause retinopathy and visual impairment. Acute mortality may be as high as 80–100% in horses and 50% in sheep. Sheep that survive may have permanent neurologic deficits. Recurrence of acute disease has been described in sheep. Natural symptomatic infection with fatal outcome has also been reported in cattle, cats (including a free-ranging lynx), rabbits, and one dog. Epidemics of paresis in ostriches were attributed to BDV infection; however, these data have not been confirmed.

Viral persistence without apparent disease has been described in naturally infected horses, sheep, and cats in Europe and Asia. There is one report indicating subclinical infection of horses in North America.

IV. EXPERIMENTAL MODELS OF BORNAVIRUS INFECTION

A wide range of animal species have been experimentally infected with BDV. Rats, mice, hamsters, rabbits, guinea pigs, tree shrews, rhesus monkeys, and chickens are all susceptible to classical disease; however, the incubation period, mortality, and severity of disease vary considerably between species and immune status of the host. Whereas in adult immunocompetent hosts infection results in dramatic, immune-mediated meningoencephalitis consistent with the classical syndrome observed in naturally infected horses and sheep, animals tolerant of infection due to immature or compromised immune systems have a more subtle course.

A. Adult Rat Model

Susceptibility to disease varies with host rat strain. Wistar and black-hooded rats can be productively infected but have less severe disease than Lewis rats, a strain with a lesion in the hypothalamic–pituitary adrenal axis associated with enhanced susceptibility to immune-mediated disorders, such as experimental allergic encephalomyelitis and adjuvant-induced arthritis. Resistance to BD is reported to be inherited as a dominant trait in Lewis and black-hooded hybrids in a manner independent of major histocompatibility complex genes. Virulence of viral strains for rats may be enhanced by serial passages of virus in rat brain.

Infection is most rapidly achieved with intracranial and intranasal inoculations. Nonetheless, any route of inoculation that allows virus access to nerve terminals (e.g., intramuscular, intraperitoneal, or footpad injection) ultimately results in central nervous system (CNS) infection and classical disease. Viremia is unlikely to play a significant role in BDV dissemination and pathogenesis. Although viral gene products may be detected in PBMCs of infected animals, intravascular inoculation only infrequently results in productive infection. Several observations indicate that BDV disseminates primarily via neural networks: (i) viral proteins and nucleic acids can be traced centripetally and transsynaptically after olfactory, ophthalmic, or intraperitoneal inoculation; (ii) the onset of disease is delayed with distance of the inoculation site to the central nervous system; and (iii) migration of virus to the CNS after footpad infection can be prevented by sciatic nerve transection.

Irrespective of the route of inoculation, initiation of clinical disease coincides with the appearance of viral proteins in hippocampal neurons and onset of meningitis. It has been proposed that BDV, like rabies virus, spreads as a ribonucleoprotein complex (RNP) within neural networks. Structures consistent with RNPs are described in neurons of experimentally infected animals; nonetheless, the form of disseminating virus is unknown.

In adult-infected rats, disease presents clinically as hyperactivity and exaggerated startle responses 10–14 days after intracerebral infection. The acute phase coincides with infiltration of monocytes into the brain, particularly in areas of high viral burden, including the hippocampus, amygdala, and other limbic structures. Two or 3 weeks later, rats have stereotyped motor behaviors, dyskinesias, dystonia, and flexed seated postures. Five to 10% of animals become obese, reaching body weights up to 300% of normal rats. Some investigators have reported isolation of a BDV strain that causes obesity in more than 50% of infected adult rats.

Disorders of movement and behavior are linked to distinct changes in CNS dopamine (DA) systems. Adult-infected BD animals are more sensitive to DA agonists and antagonists than are normal rats. Administration of the indirect DA agonist dextroamphetamine to infected animals results in increased locomotor and stereotypic behaviors. Similarly, cocaine-mediated inhibition of DA reuptake potentiates DA neurotransmission, resulting in enhanced locomotion and stereotypic behaviors. The movement and behavior disorder is improved following treatment with selective DA antagonists; whereas D2-selective antagonists (e.g., raclopride) do not affect locomotor responses in BD rats, high doses of selective D1 antagonists (e.g., SCH23390) and atypical DA blocking agents with mixed D1 and D2 antagonist activity (such as clozapine) selectively reduce locomotor activity in BD rats but not in controls.

The basis for these functional disturbances appears to be partial DA deafferentation with compensatory metabolic hyperactivity in nigrostriatal and mesolimbic DA systems. At the receptor level, both pre- and postsynaptic sites of the DA transmitter system appear to be damaged in striatum (caudate-putamen and nucleus accumbens). DA reuptake sites are reduced in nucleus accumbens and caudate-putamen. D2 (but not D1) receptor binding is markedly reduced in caudate-putamen; D2 and D3 receptor binding is reduced in nucleus accumbens. It is possible that the differential effects of infection on DA receptors in nucleus accumbens and caudate-putamen may reflect BDV-mediated interference with the cellular transcription and/or splicing machinery. Whereas binding to receptors expressed from spliced messages, D2 and D3, is reduced, binding to D1, a receptor expressed from an unspliced message, is not.

Although the increased locomotor activity, stereotypic behaviors, and dyskinesias of the adult BD model are linked to distinct disturbances in DA pathways, additional neuromodulator abnormalities have also been noted. The expression of genes for neuromodulatory substances and their associated synthesizing enzymes, including somatostatin, cholecystokinin, and glutamic acid decarboxylase, is greatly reduced during the acute phase and recovers toward normal in the chronic phase of adult BD. The cholinergic system, a major component in sensorimotor processing, learning, and memory, also appears to be affected in adult BDV infection. A progressive decrease in the number of choline acetyltransferase-positive fibers in hippocampus and neocortex has been observed to begin as early as day 6 postinfection. Preliminary work on dysregulation of the serotonin (5-HT) and norepinephrine (NE) systems suggests metabolic hyperactivity of 5-HT. There is a modest increase in the 5-HT metabolite, 5-hydroxyindoleacetic acid, in striatum and of the NE metabolite, 3-methoxy-4-hydroxyphenethyleneglycol, in prefrontal and anterior cingulate cortex. These changes may reflect compensatory upregulation following partial loss of DA afferents to these brain regions. Selective effects of BDV on 5-HT and NE pre- or postsynaptic receptors have not yet been investigated. The mechanism by which adult BD rats are rendered hypersensitive to the hyperkinetic and convulsant effects of the opiate antagonist, naloxone, is also unclear. In addition, because adult infected rats have marked CNS inflammation, it has not been possible to determine whether monoaminergic, cholinergic, and opiatergic dysfunction in BD results from direct effects of the virus, virus effects on resident cells of the CNS, or a cellular immune response to viral gene products.

In late disease, BDV disseminates throughout the autonomic and peripheral nervous systems and can be readily detected in autonomic plexi in the lungs and gastrointestinal tract and at the neuromuscular junction. It is also present at lower levels in nonneural tissues, including bone marrow, thymus, and PBMCs. Although numbers of mononuclear inflammatory cells in the CNS are markedly reduced during the chronic phase of disease, there is an elevation in titers of antibodies directed against all BDV proteins (N, X, P, M, G, and L). Antibodies specific for M and G have

neutralizing activity *in vitro*. Although virus is not cleared from the brain, neutralizing antibodies may modulate viral gene expression and limit the infection to the CNS, preventing further dissemination to nonneural organs. The blood–brain barrier remains functional in various assays. Animals may have a normal life expectancy despite loss of up to 50% of brain mass.

B. Neonatal Rat Model

Neonatally infected rats may have a wide range of physiologic and neurobehavioral disturbances. They are smaller than uninfected littermates, display a heightened taste preference for salt solutions, and have altered sleep–wake cycles. There is no apparent alteration of glucose, growth hormone, or insulin-like growth factor-1 or amount of food ingested; thus, the cause of runting remains obscure. Behavioral disturbances are less dramatic in neonatally infected animals than in their adult-infected counterparts. A study of behavioral and cognitive changes in Wistar rats infected in the neonatal period found spatial and aversive learning deficits, increased motor activity, and decreased anxiety responses. Similar deficits in spatial learning and memory were found by Carbone and colleagues in neonatally infected Lewis rats. Play behavior is also abnormal in the neonatally infected rats, with decrease in both initiation of nondominance-related play interactions and response to initiation of play by noninfected, age-matched control animals or infected littermates.

The neonatal infection model has not been studied as extensively as the adult infection model. CNS dysfunction in neonatally infected animals has been proposed to be linked to viral effects on morphogenesis of the hippocampus and cerebellum, two structures in rodents that continue to develop after birth. Carbone and colleagues found a quantitative relationship of hippocampal pathology to behavioral abnormalities in the neonatal infection model; the extent of neuronal loss in dentate gyrus appeared to be correlated with the severity of spatial learning and memory deficiencies in neonatally infected Lewis rats.

Humoral immune responses to BDV in neonatally infected animals are significantly lower than in adult infected animals. There is a transient cellular immune response that peaks approximately 4 weeks postintracerebral infection and dissipates within 10–14 days. This period coincides with the presence of high levels of mRNAs for cytokine products of CNS macrophages/microglia (IL-1α, IL-1β, IL-6, and tumor necrosis factor-α) in hippocampus, amygdala, cerebellum, prefrontal cortex, and nucleus accumbens. The extent to which cellular inflammatory infiltrates contribute to neuropathogenesis is uncertain given that thymectomized animals do not have these infiltrates but have the same behavioral disturbances and histopathology.

Marked astrocytosis has been noted in dentate gyrus and cerebellum. Upregulation of tissue factor (TF)—a member of the class II cytokine receptor family primarily produced by astrocytes that plays important roles in cellular signal transduction, brain function, and neural development through its effects on coagulation protease cascades—has been identified as one mechanism by which BDV may alter CNS development. Additionally, levels of two molecules implicated in brain development and plasticity, growth-associated protein 43 (a presynaptic membrane phosphoprotein found in neuronal growth cones) and synaptophysin (a calcium-binding protein associated with presynaptic vesicles), are reduced in cortex and hippocampus prior to neuronal losses in those structures. However, cerebellar changes cannot be explained by this mechanism because neither TF upregulation nor decrease in levels of growth-associated protein 43 or synaptophysin are observed in cerebellum. Furthermore, BDV infection of astrocytes appears to be required for TF upregulation, and cerebellar astrocytes are reported to be spared from BDV infection, at least for 30 days following neonatal infection. Whether BDV infection influences expression in cerebellum of molecules similar in function to growth-associated protein 43, synaptophysin, or TF is unknown.

Although the highest concentration of virus in neonatally infected rats is found in the CNS, it is readily detected in a wide variety of organs and body fluids (urine, saliva, and feces). Neonatally infected rats can transmit virus to adult animals without direct contact (presumably through aerosols), suggesting the possibility that neonatally infected rats may serve as reservoirs and vectors for transmission of BDV in nature.

C. Mouse Models

Mice are readily infected and have high titers of virus in brain. Disease can be induced by adaptation of virus through multiple passages in mice or infection of

specific host strains during the neonatal period. As in rats, severe clinical disease is mediated by MHC class I restricted cytotoxic T cells. A model of subtle behavioral disease has been reported in mutant mice that lack functional CD8 cells (C57BL/10J mice) wherein maze performance was correlated with IP-10 expression independent of frank histopathology.

D. Tree Shrews

Little is known about BDV pathogenesis in phylogenetically higher species such as nonhuman primates and the prosimian tree shrews (*Tupaia glis*). Intracerebral inoculation of tree shrews leads to persistent infections and a disorder characterized primarily by hyperactivity and alterations in sociosexual behavior rather than motor dysfunction. Disturbances in breeding and social behavior were most profound in animals caged in mating pairs. Females rather than males initiated mating, and infected animals failed to reproduce despite increased sexual activity. Although detailed neuroanatomic studies were not performed, the syndrome was interpreted to be due to neuropathological changes in the limbic system.

E. Nonhuman Primates

The only reported studies of experimentally infected primates employed adult rhesus macaques (*Macaca mulatta*). These animals were initially hyperactive and subsequently became apathetic and hypokinetic. Pathology was remarkable for meningoencephalitis and retinopathy. Recent preliminary data suggest that a subtle or subclinical infection can be achieved in monkeys by neonatal intranasal infection.

V. BDV AND HUMAN DISEASE

Recognition of BDV's broad experimental host range, and the observation that disturbances in behaviors in experimentally infected animals are reminiscent of some aspects of human neuropsychiatric diseases, including major depressive disorder, bipolar disorder, schizophrenia, and autism, led to the proposal that BDV might be implicated in their pathogenesis. Although there is consensus that humans are likely to be susceptible to BDV infection, the epidemiology and clinical consequences of human infection remain controversial. There have been no large, controlled prevalence studies. Furthermore, methods for diagnosis of human infection are not standardized; thus, it is difficult to pursue meta-analysis. Most reports suggesting an association between BDV and human disease have focused on neuropsychiatric disorders, including unipolar depression, bipolar disorder, or schizophrenia; however, BDV has also been linked to chronic fatigue syndrome, AIDS encephalopathy, multiple sclerosis, motor neuron disease, and brain tumors (glioblastoma multiforme) (Tables II and III). The improbably broad spectrum of candidate disorders has led some investigators to propose that infection is ubiquitous, and, that in some disorders, elevation of serum antibody titers or the presence of viral transcripts in peripheral blood mononuclear cells or neural tissues reflects generalized (AIDS) or localized (glioblastoma multiforme) immunosuppression.

There are only infrequent reports of isolation of infectious virus from humans, or detection of BDV gene products in human brain by *in situ* hybridization and immunohistochemistry: a group of four North American subjects with temporal sclerosis and a single Japanese subject with schizophrenia. Methods used most commonly for serologic diagnosis of infection include indirect immunofluorescence with infected cells and Western immunoblot or enzyme-linked immunosorbent assays with extracts of infected cells or recombinant proteins. Infection has also been diagnosed through demonstration of BDV transcripts and proteins in tissues or peripheral blood mononuclear cells. Frequently, detection of viral RNA has been achieved through nested reverse transcription polymerase chain reaction (nRT-PCR), a sensitive method that is prone to artifacts due to inadvertent introduction of template from laboratory isolates or cross contamination of samples. Amplification products representing bona fide isolates and those due to nRT-PCR amplification of low-level contaminants cannot be readily distinguished by sequence analysis because, unlike some other NNS RNA viruses, in which the inherent low fidelity of viral RNA-dependent RNA polymerases results in sequence divergence of 10^3 or 10^4 per site per round of replication, BDV is characterized by extraordinary sequence conservation. Although one report from western Austria described a horse with encephalitis, wherein N and P gene nucleotide sequences differed from other isolates by up to 15%, most studies of BDV isolates revealed variability of up to only 4.1% at the nucleotide level and 3% at the predicted amino acid level. Thus, similarities in sequence between putative new isolates

Table II
Serum Immunoreactivity to BDV in Subjects with Various Diseases[a]

Disease	Prevalence (%)		Assay	Reference
	Disease	Control		
Psychiatric (various)	0.6 (4/694)	0 (0/200)	IFA	Rott *et al*. (1985). *Science* **228,** 755
	2 (13/642)	2 (11/540)	IFA	Bode *et al*. (1988). *Lancet* **2,** 689
	4–7 (200–350/5000)	1 (10/1000)	WB/IFA	Rott *et al*. (1991). *Arch. Virol*. **118,** 143
	12 (6/49)		IFA	Bode *et al*. (1993). *Arch. Virol*. **S7,** 159
	30 (18/60)		WB	Kishi *et al*. (1995). *FEBS Lett*. **364,** 293
	14 (18/132)	1.5 (3/203)	WB	Sauder *et al*. (1996). *J. Virol*. **70,** 7713
	24 (13/55)	11 (4/36)	IFA	Igata-Yi *et al*. (1996). *Nat. Med*. **2,** 948
	0 (0/44)	0 (0/70)	IFA/WB	Kubo *et al*. (1997). *Clin. Diagn. Lab. Immunol*. **4,** 189
Affective disorders	4.5 (12/265)	0 (0/105)	IFA	Amsterdam *et al*. (1985). *Arch. Gen. Psych*. **42,** 1093
	4 (12/285)	0 (0/200)	IFA	Rott *et al*. (1985)
	38 or 12 (53 or 17/138)	16 or 4 (19 or 5/117)	WB (N or P)	Fu *et al*. (1993). *J. Affect. Disord*. **27,** 61
	37 (10/27)		IFA	Bode *et al*. (1993)
	12 (6/52)	1.5 (3/203)	WB	Sauder *et al*. (1996)
	0–0.8 (0–1/122)	0 (0/70)	IFA/WB	Kubo *et al*. (1997)
	2 (1/45)	0 (0/45)	WB	Fukuda *et al*. (2001). *J. Clin. Microbiol*. **39,** 419
Schizophrenia	25 (1/4)		IFA	Bode *et al*. (1993)
	9–28 (8 or 25/90)	0–20 (0 or 4/20)	WB (N or P)	Waltrip *et al*. (1995). *Psychiatr. Res*. **56,** 33
	17 (15/90)	15 (3/20)	IFA	Waltrip *et al*. (1995)
	14 (16/114)	1.5 (3/203)	WB	Sauder *et al*. (1996)
	20 (2/10)		WB	Richt *et al*. (1997). *J. Neurovirol*. **3,** 174
	0–1 (0–2/167)	0 (0/70)	IFA/WB	Kubo *et al*. (1997)
	9 (4/45)	0 (0/45)	WB	Fukuda *et al*. (2001)
CFS	24 (6/25)		WB	Nakaya *et al*. (1996). *FEBS Lett*. **378,** 145
	0 (0/69)	0 (0/62)	WB	Evengard *et al*. (1999). *J. Neurovirol*. **5,** 495
MS	13 (15/114)	2.3 (11/483)	IP/IFA	Bode *et al*. (1992). *J. Med. Virol*. **36,** 309
	0 (0/50)		IFA	Kitze *et al*. (1996). *J. Neurol*. **243,** 660

[a]Abbreviations used: ELISA, enzyme-linked immunosorbent assay; IFA, immunofluorescence assay; WB, Western immunoblot; IP, immunoprecipitation; CFS, chronic fatigue syndrome; MS, multiple sclerosis; N, nucleoprotein; P, phosphoprotein.

and confirmed isolates cannot be used to exclude the former as artifacts.

A blinded international multicenter study composed of more than 2000 subjects recruited at sites in North America, Europe, and Asia using standardized diagnostic clinical instruments is under way to determine the prevalence of BDV infection in individuals with affective disorders or schizophrenia. Results of serology (antibodies to BDV N, P, and M proteins) and real-time polymerase chain reaction analyses (BDV genetic sequences in white blood cells) are anticipated in late 2002.

No specific vaccine or antiviral therapy is established for BDV. Although there is one report in which BDV was found to be sensitive to amantadine *in vitro* and *in vivo*, three other reports found no antiviral activity *in vitro* or *in vivo*. The nucleoside analog ribavirin inhibits viral replication *in vitro*. Whether it has an impact on viral replication *in vivo* or on severity of disease is unknown.

Table III
BDV RNA, Virus or Protein in Subjects with Various Diseases[a]

Disease	Tissue	Prevalence (%)		Divergence (%)[b]	Reference
		Disease	Controls		
Psychiatric (various)	PBMC	67 (4/6)	0 (0/10)	0–3.6	Bode *et al*. (1995). *Nat. Med*. **1,** 232
	PBMC	37 (22/60)			Kishi *et al*. (1995). *FEBS Lett*. **364,** 293
	PBMC	42 (5/12)	0 (0/23)	0–4.0	Sauder *et al*. (1996). *J. Virol*. **70,** 7713
	PBMC-coculture	9 (3/33)	0 (0/5)	0.07–0.83	Bode *et al*. (1996). *Mol. Psych*. **1,** 200 de la Torre *et al*. (1996). *Virus Res*. **44,** 33
	PBMC	2 (2/106)	0 (0/12)		Kubo *et al*. (1997). *Clin. Diagn. Lab. Immunol*. **4,** 189
	PBMC	0 (0/24)	0 (0/4)		Richt *et al*. (1997). *J. Neurovirol*. **3,** 174
	PBMC	37 (10/27)	15 (2/13)		Vahlenkamp *et al*. (2000). *Vet. Microbiol*. **76,** 229
Affective disorders	PBMC	33 (1/3)	0 (0/23)		Sauder *et al*. (1996). *J. Virol*. **70,** 7713
	PBMC	17 (1/6)	0 (0/36)		Igata-Yi *et al*. (1996). *Nat. Med*. **2,** 948
	Brain	40 (2/5)	0 (0/10)		Salvatore *et al*. (1997). *Lancet* **349,** 1813
	PBMC	4 (2/49)	2 (2/84)	0–5.1	Iwata *et al*. (1998). *J. Virol*. **72,** 10044
	Brain (CSF)	5 (3/65)	0 (0/69)	Protein	Deuschle *et al*. (1998). *Lancet* **352,** 1828
	PBMC	2 (1/45)	0 (0/45)		Fukuda *et al*. (2001). *J. Clin. Microbiol*. **39,** 419
Schizophrenia	Brain	0 (0/3)	0 (0/3)		Sierra-Honigman *et al*. (1995). *Br. J. Psychol*. **166,** 55
	CSF	0 (0/8)	0 (0/8)		Sierra-Honigman *et al*. (1995)
	PBMC	0 (0/7)	0 (0/7)		Sierra-Honigman *et al*. (1995)
	PBMC	64 (7/11)	0 (0/23)		Sauder *et al*. (1996)
	PBMC	10 (5/49)	0 (0/36)		Igata-Yi *et al*. (1996)
	PBMC	100 (3/3)		4.2–9.3	Kishi *et al*. (1996). *J. Virol*. **70,** 635
	PBMC	0 (0/10)	0 (0/10)		Richt *et al*. (1997)
	Brain	53 (9/17)	0 (0/10)		Salvatore *et al*. (1997)
	PBMC	4 (3/77)	2 (2/84)	0–5.1	Iwata *et al*. (1998)
	PBMC	14 (10/74)	1 (1/69)		Chen *et al*. (1999). *Mol. Psychiatr*. **4,** 566
	PBMC	0 (0/45)	0 (0/45)		Fukuda *et al*. (2001)
	Brain	25 (1/4)		RNA, virus, protein	Nakamura *et al*. (2000). *J. Virol*. **74,** 4601
CFS	PBMC	12 (7/57)			Kitani *et al*. (1996). *Microbiol. Immunol*. **40,** 459
	PBMC	12 (3/25)		6.0–14	Nakaya *et al*. (1996). *FEBS Lett*. **378,** 145
Hippocampal sclerosis	Brain	80 (4/5)			de la Torre *et al*. (1996)
MS	Brain (CSF)	11 (2/19)	0 (0/69)	Protein	Deuschle *et al*. (1998)
Normal controls	PBMC		5 (8/172)		Kishi *et al*. (1995). *Med. Microbiol. Immunol. Berlin* **184,** 135
	Brain		6.7 (2/30)		Haga *et al*. (1997). *Brain Res*. **770,** 307

[a]Abbreviations used: PBMC, peripheral blood mononuclear cells; CSF, cerebrospinal fluid; CFS, chronic fatigue syndrome.
[b]Divergence of P-gene nucleotide sequence from common BDV isolates (strain V and He/80).

See Also the Following Articles

AUTOIMMUNE DISEASES • DOPAMINE • HIV INFECTION, NEUROCOGNITIVE COMPLICATIONS OF • LYME ENCEPHALOPATHY • PRION DISEASES

Suggested Reading

Briese, T., Hornig, M., and Lipkin, W. I. (1999). Bornavirus immunopathogenesis in rodents: Models for human neurological disease. *J. Neurovirol.* **5,** 604–612.

Dietrich, D. E., Schedlowski, M., Bode, L., Ludwig, H., and Emrich, H. M. (1998). A viropsycho-immunological disease model of a subtype affective disorder. *Pharmacopsychiatry* **31,** 77–82.

Durrwald, R., and Ludwig, H. (1997). Borna disease virus (BDV), a zoonotic worldwide pathogen. A review of the history of the disease and the virus infection with comprehensive bibliography. *Zentralbl. Veterinarmed.* **44,** 147–184.

Gonzales-Dunia, D., Sauder, C., and de la Torre, J. C. (1997). Borna disease virus and the brain. *Brain Res. Bull.* **44,** 647–664.

Gosztonyi, G., and Ludwig, H. (1995). Borna disease—Neuropathology and pathogenesis. *Curr. Topics Microbiol. Immunol.* **190,** 39–73.

Hornig, M., Solbrig, M. V., Horscroft, N., Weissenböck, H., and Lipkin, W. I. (2001). Borna disease virus infection of adult and neonatal rats: Models for neuropsychiatric disease. *Curr. Topics Microbiol. Immunol.* **253,** 157–178.

Jordan, I., and Lipkin, W. I. (2001). Borna disease virus. *Rev. Med. Virol.* **11,** 37–57.

Ludwig, H., and Bode, L. (2000). Borna disease virus: New aspects on infection, disease, diagnosis and epidemiology. *Rev. Sci. Tech.* **19,** 259–288.

Richt, J. A., and Rott, R. (2001). Borna disease virus: A mystery as an emerging zoonotic pathogen. *Vet. J.* **161,** 24–40.

Schneemann, A., Schneider, P. A., Lamb, R. A., and Lipkin, W. I. (1995). The remarkable coding strategy of borna disease virus: A new member of the nonsegmented negative strand RNA viruses. *Virology* **210,** 1–8.

Schwemmle, M. (1999). Progress and controversy in Bornavirus research: A meeting report. *Arch. Virol.* **144,** 835–840.

Solbrig, M. V., Fallon, J. H., and Lipkin, W. I. (1995). Behavioral disturbances and pharmacology of Borna disease. *Curr. Topics Microbiol. Immunol.* **190,** 93–101.

Solbrig, M. V., Koob, G., and Lipkin, W. I. (1999). Orofacial dyskinesias and dystonia in rats infected with Borna disease virus: A model for tardive dyskinetic syndromes. *Mol. Psychiatr.* **4,** 310–312.

Staeheli, P., Sauder, C., Hausmann, J., Ehrensperger, F., and Schwemmle, M. (2000). Epidemiology of Borna disease virus. *J. Gen. Virol.* **81,** 2123–2135.

Stitz, L., Diezschold, B., and Carbone, K. M. (1995). Immunopathogenesis of Borna disease. *Curr. Topics Microbiol. Immunol.* **190,** 75–92.

Brain Anatomy and Networks

M.-MARSEL MESULAM
Northwestern University Medical School

GLOSSARY

cerebral cortex The surface of the forebrain.

heteromodal cortex Neocortical area that is interconnected with association areas serving different sensory modalities.

large-scale network A set of interconnected areas that mediates a specific mental function.

paralimbic areas Cytoarchitectonically transitional areas that are interconnected with the amygdala and hippocampus.

transmodal areas Collective term for all heteromodal, paralimbic, and limbic cortices.

unimodal cortex Neocortical area devoted to the processing of a single sensory modality.

The cerebral cortex of the human brain contains approximately 20 billion neurons spread over nearly 2000 square centimeters of surface area. There is no universal agreement on terminology, no distinct boundaries that demarcate one region from another, and, in most instances, no clear correspondence among lobar designations, traditional topographic landmarks, cytoarchitectonic boundaries, and behavioral specializations. One part of the brain can have more than one descriptive name, and cytoarchitectonic (striate cortex), functional (primary visual cortex), topographic (calcarine cortex), and eponymic (Brodmann area 17) terms can be used interchangeably to designate the same area. The purpose of this article is to provide an introductory guide for navigating the human cerebral cortex from the vantage point of functional specializations.

This article is adapted from Chapter 1 of *Principles of Behavioral and Cognitive Neurology* (M.-M. Mesulam, Ed), pp. 1–120. Oxford Univ. Press, New York, 2000.

I. PARTS OF THE CEREBRAL CORTEX

The absence of clear anatomical boundaries has encouraged the development of numerous approaches to the subdivision of the cerebral cortex. The most commonly used map is based on Korbinian Brodmann's parcellation of the cerebral hemispheres into more than 50 areas, each characterized by a specific pattern of neuronal density and lamination (Fig. 1). These cytoarchitectonic areas can be divided into five major functional zones: *limbic, paralimbic, heteromodal association, unimodal association, and primary sensory–motor*. The principal factual base for this functional categorization is derived from experiments in macaque monkeys. The homologies to the human brain have been inferred from comparative cytoarchitectonics, electrophysiological recordings, functional imaging, and the behavioral effects of focal lesions.

A. The Limbic Zone (Corticoid and Allocortical Formations)

The basal forebrain is usually considered a subcortical structure. However, some of its constituents are

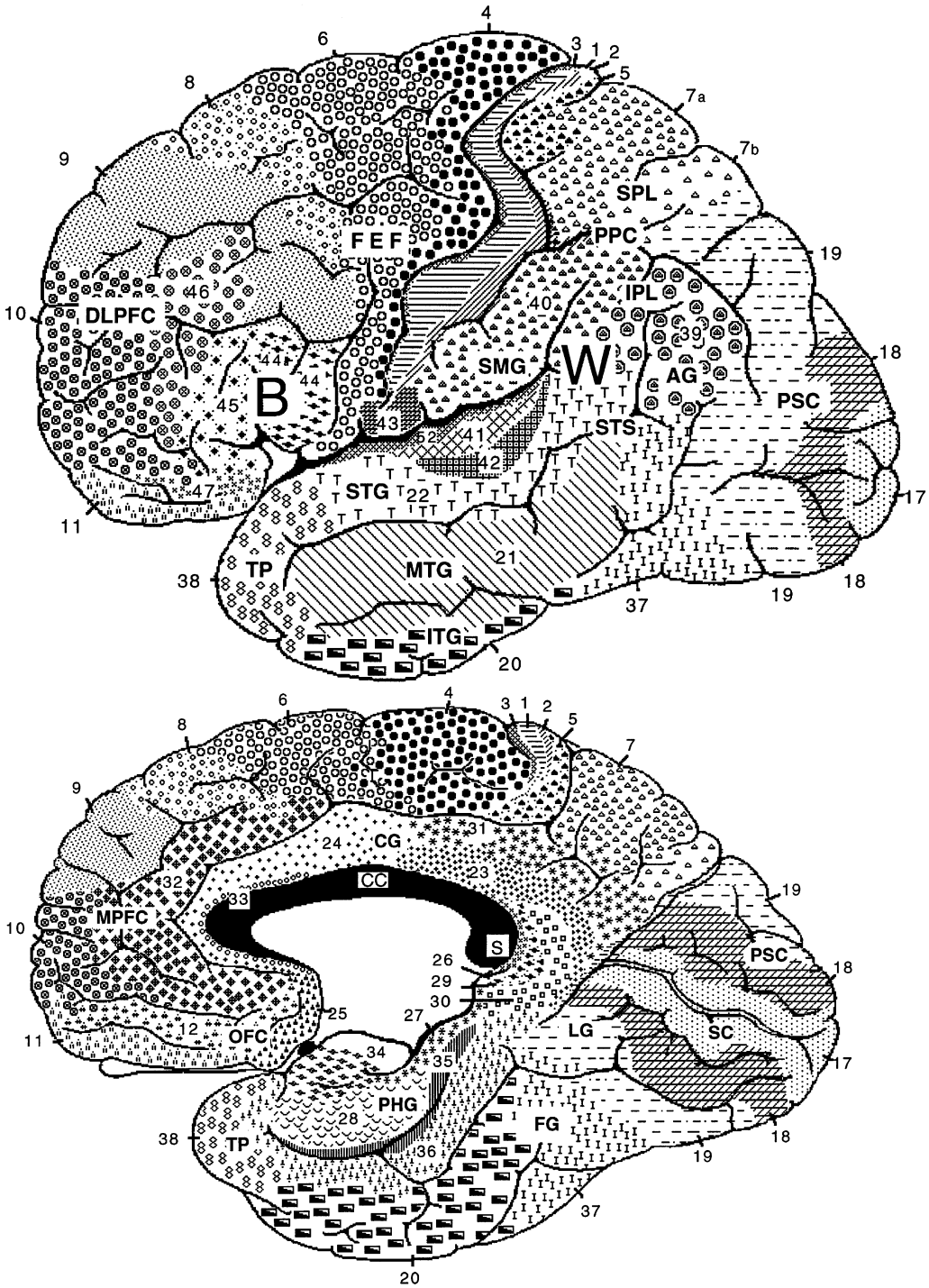


Figure 1 Lateral (*top*) and medial (*bottom*) views of the cerebral hemispheres. The numbers refer to the Brodmann cytoarchitectonic designations. Area 17 corresponds to primary visual cortex, 41 and 42 to primary auditory cortex, 1–3 to primary somatosensory cortex, and 4 to primary motor cortex. The rest of the cerebral cortex contains association areas. AG, angular gyrus; B, Broca's area; CC, corpus callosum; CG, cingulate cortex; DLPFC, dorsolateral prefrontal cortex; FEF, frontal eye fields (premotor cortex); FG, fusiform gyrus; IPL, inferior parietal lobule; ITG, inferior temporal gyrus; LG, lingual gyrus; MPFC, medial prefrontal cortex; MTG, middle temporal gyrus; OFC, orbitofrontal cortex; PHG, parahippocampal gyrus; PPC, posterior parietal cortex; PSC, peristriate cortex; SC, striate cortex; SMG, supramarginal gyrus; SPL, superior parietal lobule; STG, superior temporal gyrus; STS, superior temporal sulcus; TP, temporopolar cortex; W, Wernicke's area. The insula is hidden from view.

situated directly on the ventral and medial surfaces of the cerebral hemispheres and are therefore part of the cerebral cortex. These basal forebrain structures include the septal region, the substantia innominata, the amygdaloid complex, and the anterior olfactory nucleus. Because of their simplified cytoarchitecture, these structures can be designated "corticoid," or cortex-like. In some corticoid areas, such as the septal region and the substantia innominata, the organization of neurons is so rudimentary that no consistent lamination can be discerned and the orientation of dendrites is haphazard. All corticoid areas have architectonic features that are in part cortical and in part nuclear. This duality is particularly conspicuous in the amygdala.

The next stage of cortical organization carries the designation of allocortex. This type of cortex contains one or two bands of neurons arranged into moderately well-differentiated layers. The apical dendrites of the constituent neurons are well developed and display orderly patterns of orientation. There are two allocortical formations in the mammalian brain: (i) the hippocampal complex (i.e., the dentate gyrus, the CA1-4 fields, and the subicular areas), which also carries the designation of *archicortex*, and (ii) the piriform or primary olfactory cortex which is also known as *paleocortex*. The corticoid and allocortical formations collectively make up the limbic zone of the cerebral cortex.

B. The Paralimbic Zone (Mesocortex)

The next level of structural complexity is encountered in the paralimbic regions of the brain, also known as *mesocortex*. These areas are located between allocortex and isocortex so as to provide a gradual transition from one to the other. Allocortical cell layers often extend into paralimbic areas. The sectors of paralimbic areas that abut allocortex are also known as periallocortical or juxtaallocortical, whereas the sectors that abut isocortex can be designated proisocortical or periisocortical. In most paralimbic areas, the transitional changes from periallocortex to periisocortex include

1. Progressively greater accumulation of small granular neurons (star pyramids) first in layer IV and then in layer II
2. Sublamination and columnarization of layer III
3. Differentiation of layer V from layer VI and of layer VI from the underlying white matter
4. An increase of intracortical myelin, especially along the outer layer of Baillarger (layer IV)

In general, the emergence of relatively well-differentiated granular cell bands in layers IV and II, the sublamination of layer III, and the differentiation of layer V from layer VI mark the end of the paralimbic zone and the onset of six-layered homotypical isocortex.

There are five major paralimbic formations in the human brain: the orbitofrontal cortex [posterior parts of Brodmann's area (BA) 11 and 12 and all of BA 13]; the insula (BA 14–16); the temporal pole (BA 38); the parahippocampal cortices, including the pre- and parasubiculum, the entorhinal area, the prorhinal area, and the perirhinal (transentorhinal) area, corresponding to BA 27–28 and 35; and the cingulate complex, including the retrosplenial, ventral cingulate, and subcallosal parolfactory areas, corresponding at least in part to BA 23–26, 29–33.

These five paralimbic regions form an uninterrupted girdle surrounding the medial and basal aspects of the cerebral hemispheres. The olfactory piriform cortex provides the allocortical nidus for the orbitofrontal, insular, and temporopolar paralimbic areas. These are designated olfactocentric paralimbic formations. The hippocampus and its supracallosal rudiment (known as the induseum griseum) provide the allocortical nidus for the cingulate and parahippocampal areas. These are designated hippocampocentric paralimbic formations. The olfactocentric and hippocampocentric sectors of the paralimbic belt merge into each other within the subcallosal, medial orbitofrontal and anterior parahippocampal cortices.

C. Homotypical Association Isocortex (the Heteromodal and Unimodal Zones)

By far the greatest area of the cerebral cortex in the human brain is devoted to six-layered homotypical isocortex (or neocortex), also known as association isocortex. Association isocortex can be subdivided into two major zones: modality-specific (unimodal) and high-order (heteromodal). Unimodal sensory association areas are further divided into *upstream* and *downstream* components: Upstream areas are only one synapse away from the relevant primary sensory area, whereas downstream areas are at a distance of two or more synapses from the corresponding primary area. Unimodal sensory association isocortex is defined by three essential characteristics:

1. The constituent neurons respond predominantly, if not exclusively, to stimulation in only a single sensory modality.
2. The predominant sensory information comes from the primary sensory cortex and other unimodal regions of that same modality.
3. Lesions yield deficits only in tasks guided by that modality.

Unimodal visual association cortex can be divided into an upstream peristriate component that includes areas BA 18 and 19 and a downstream temporal component that includes inferotemporal cortex (BA 20 and 21) in the monkey and the fusiform, inferior temporal, and probably parts of the middle temporal gyri in the human. *Unimodal auditory association cortex* covers the superior temporal gyrus (BA 22) and perhaps also parts of the middle temporal gyrus (BA 21) in the human. The connectivity of the monkey brain suggests that the posterior part of the superior temporal cortex (BA 22) displays the properties of upstream auditory association cortex, whereas the more anterior part of this gyrus and the dorsal banks of the superior temporal sulcus may fit the designation of downstream auditory association cortex.

In the monkey brain, BA 5 in the superior parietal lobule represents an upstream component of *somatosensory unimodal association cortex*, whereas parts of BA 7b in the inferior parietal lobule and the posterior insula may represent its downstream components. In the human, unimodal somatosensory association cortex may include parts of BA 5 and BA 7 in the superior parietal lobule and perhaps parts of BA 40 in the anterior parts of the inferior parietal lobule. The subdivision of unimodal auditory and somatosensory association cortices into upstream and downstream areas in the human remains to be elucidated. Unimodal association areas for olfaction, taste, and vestibular sensation have not been fully characterized. Premotor regions (BA 6, the frontal eye fields, and BA 44) fulfill the role of motor ''association'' areas because they provide the principal cortical input into primary motor cortex.

The heteromodal component of association isocortex is identified by the following characteristics:

1. Neuronal responses are not confined to any single sensory modality.
2. The predominant sensory inputs come from unimodal areas in multiple modalities and from other heteromodal areas.
3. Deficits resulting from lesions in these areas are always multimodal and never confined to tasks under the guidance of a single modality.

Some neurons in heteromodal association areas respond to stimulation in more than one modality, indicating the presence of direct multimodal convergence. More commonly, however, there is an admixture of neurons with different preferred modalities. Many neurons have sensory as well as motor contingencies; others change firing in ways that are responsive to motivational relevance. Defined in this fashion, heteromodal cortex includes the types of regions that have been designated as high-order association cortex, polymodal cortex, multimodal cortex, polysensory areas, and supramodal cortex. The monkey brain contains heteromodal areas in prefrontal cortex (BA 9, 10, 45, 46, anterior BA 8, and anterior BA 11 and 12), the inferior parietal lobule (parts of BA 7), lateral temporal cortex within the banks of the superior temporal sulcus (junction of BA 22 with BA 21), and the parahippocampal region. In the human brain, it is reasonable to assume that the analogous zones of heteromodal cortex are located in prefrontal cortex (BA 9 and 10 and 45–47, anterior 11 and 12, and anterior 8), posterior parietal cortex (posterior BA 7 and 39 and 40), lateral temporal cortex (including parts of BA 37 and BA 21 in the middle temporal gyrus), and portions of the parahippocampal gyrus (parts of BA 36 and 37).

Unimodal and heteromodal areas are characterized by a six-layered homotypical architecture. There are some relatively subtle architectonic differences between unimodal and heteromodal areas. In general, the unimodal areas have a more differentiated organization, especially with respect to sublamination in layers III and V, columnarization in layer III, and more extensive granularization in layers IV and II. On these architectonic grounds, it appears that heteromodal cortex is closer in structure to paralimbic cortex and that it provides a stage of cytoarchitectonic differentiation intercalated between paralimbic and unimodal areas.

D. Idiotypic Cortex (the Primary Sensory–Motor Zones)

Primary visual, auditory, somatosensory, and motor cortices are easily delineated on cytoarchitectonic and functional grounds. Primary visual cortex (also

known as V1, striate cortex, calcarine cortex, or BA 17) covers the banks of the calcarine fissure, primary auditory cortex (also known as A1 or BA 41 and 42) covers Heschl's gyrus on the floor of the Sylvian cistern, primary somatosensory cortex (also known as S1, usually meant to include BA 3a, 3b, 1, and 2) is located in the postcentral gyrus, and primary motor cortex (also known as M1) includes BA 4 and probably also a posterior rim of BA 6 in the precentral gyrus.

There are two divergent opinions about these primary areas. One is to consider them as the most elementary (even rudimentary) component of the cerebral cortex; the other is to consider them as its most advanced and highly differentiated component. The latter point of view is supported from the standpoint of cytoarchitectonics. Thus, the primary visual, somatosensory, and auditory cortices display a "koniocortical" architecture representing the highest level of development with respect to granularization, lamination, and columnarization, whereas primary motor cortex displays a unique "macropyramidal" architecture characterized by highly specialized giant pyramidal neurons known as Betz cells.

The visual, auditory, and somatosensory systems provide the major channels of communication with the extrapersonal world. The information transmitted by these channels plays a critical role in shaping the contents of cognition and consciousness. The primary and unimodal areas related to these modalities are cytoarchitectonically highly differentiated and quite large. The vestibular, gustatory, and olfactory sensations do not have the same type of prominence in the primate brain. The corresponding primary areas are cytoarchitectonically less differentiated, smaller, and closer to limbic structures. In the monkey, primary gustatory cortex is located in the frontoinsular junction in BA 43; the primary vestibular area lies within the Sylvian fissure, where the temporal lobe joins the insula and parietal lobe; and the primary olfactory cortex is a core limbic region located at the confluence of the insular, orbitofrontal, and temporopolar areas. The equivalent areas of the human brain have not been fully identified.

II. CORTICAL ORGANIZATION, CONNECTIVITY, AND TRANSMODAL AREAS

The constituents of the limbic zone have the most extensive hypothalamic interconnections. Through neural and also humoral mechanisms, the hypothalamus is in a position to control electrolyte balance, glucose levels, basal temperature, metabolic rate, autonomic tone, hormonal state, sexual phase, circadian oscillations, and immunoregulation and also to modulate the experience and expression of hunger, aggression, fear, flight, thirst, and libido. In keeping with these functions of the hypothalamus, the constituents of the limbic zone (septal nuclei, substantia innominata, amygdala, olfactory complex, and hippocampus) assume pivotal roles in the regulation of memory, emotion, motivation, hormonal balance, and autonomic function. Many of these behavioral affiliations are related to the upkeep of the internal milieu (homeostasis) and the regulation of genetically programmed basic drives.

At the opposite end of this spectrum of cytoarchitectonic differentiation are the highly specialized primary sensory and motor areas. These parts of the cerebral cortex are most closely related to the extrapersonal space: Primary sensory cortex provides an obligatory portal for the entry of information from the environment into cortical circuitry, and primary motor cortex provides a final common pathway for coordinating the motor acts that allow us to manipulate the environment and alter our position within it. Unimodal, heteromodal, and paralimbic cortices are inserted between these two extremes. They provide neural bridges that mediate between the internal and the external worlds so that the needs of the internal milieu are discharged according to the opportunities and restrictions that prevail in the extrapersonal environment. These three intercalated zones enable the associative elaboration of sensory information, its linkage to motor strategies, and the integration of experience with drive, emotion, and autonomic states. The unimodal and heteromodal zones are most closely involved in perceptual elaboration and motor planning, whereas the paralimbic zone plays a critical role in integrating emotion and motivation with action, perception, and learning.

A. Connectivity of the Cerebral Cortex

Components of each functional zone have *extramural* connections with components of other functional zones and *intramural* connections within the same zone. Experimental evidence, obtained mostly from the monkey, shows that the most intense extramural connectivity of an individual cortical area occurs with components of immediately adjacent functional zones as shown in Fig. 2. For example, although all types of

cortical areas, including association isocortex, receive direct hypothalamic projections, such connections reach their highest intensity within components of the limbic zone such as the septal area, the nucleus basalis of the substantia innominata, the amygdaloid complex, piriform cortex, and the hippocampus. In keeping with this organization, a second major source of connections for limbic structures originates in the paralimbic zone. Thus, the amygdala receives one of its most extensive extramural cortical inputs from the insula, the hippocampus from the entorhinal sector of the parahippocampal region, and the piriform cortex as well as the nucleus basalis from the group of olfactocentric paralimbic areas.

An analogous analysis can be extended to the other zones. Paralimbic areas, for example, have the most extensive extramural connections with limbic and heteromodal areas. Furthermore, the most extensive extramural interconnections of heteromodal areas are with components of the paralimbic zone, on the one hand, and with those of the unimodal zone, on the other hand. Finally, the major extramural connections of unimodal areas occur with primary areas, on the one hand, and heteromodal areas, on the other hand, whereas primary areas derive their major inputs from unimodal areas and the external world. Although some connections cross functional levels, they are not as prominent as those that link two immediately adjacent levels. Thus, the amygdala is known to have monosynaptic connections with unimodal association isocortex and even primary visual cortex, but these are not nearly as substantial as its connections with the hypothalamus and paralimbic regions.

Many cortical areas also have intramural cortical connections within their own functional zones. These are extremely well developed within the limbic, paralimbic and heteromodal zones. In contrast, the intramural connections of primary sensorimotor and upstream unimodal sensory association areas have a particularly restricted distribution. Upstream unimodal association areas in different modalities, for example, have no interconnections with each other. Except for the connection from primary somatosensory to primary motor cortex, there are also no neural projections interconnecting primary areas in different modalities. It therefore appears that there is a premium on channel width within the limbic, paralimbic, and heteromodal regions of the cerebral cortex, whereas the emphasis is on fidelity within the upstream unimodal and primary sensorimotor areas.

B. Intermediary Processing and Transmodal Areas

A fundamental characteristic of the primate brain is the insertion of obligatory synaptic relays between sensation and action and also between the representation of the internal milieu (at the level of the hypothalamus) and that of the external world (at the level of primary sensorimotor cortex). These intercalated synaptic relays collectively provide the substrates for "intermediary" or "integrative" processing. The psychological outcomes of intermediary processing are known as "cognition," "consciousness," and "comportment" and include the diverse manifestations of memory, emotion, attention, language, planning, judgment, insight, and thought. Intermediary processing has a dual purpose. First, it protects channels of sensory input and motor output from being dominated by the emotional and motivational imperatives of the internal milieu. Second, it enables identical stimuli to trigger different responses depending on situational context, past experience, present needs, and contemplated consequences. The neurons that support intermediary processing are located within the unimodal, heteromodal, paralimbic, and limbic zones of the cerebral cortex.

The synaptic architecture of intermediary processing shapes the nature of cognition, comportment, and consciousness. In species with simpler brains, intermediary processing is shallow and does not allow much of a distinction to be made between appearance and significance. The automatic linkage between stimulus and response in such species leads to the many manifestations of instinctual behaviors. A major role of intermediary processing is to transcend inflexible stimulus–response linkages and to enable behavior to be guided by contextual significance rather than appearance.

Unimodal areas contain the initial synaptic relays for the intermediary processing of information emanating from the extrapersonal space. These areas are extremely well developed in the human brain. The absence of interconnections linking a unimodal area in one sensory modality with another in a different modality protects the sensory fidelity of experience and delays cross-modal contamination until further encoding has been accomplished. Unimodal areas are in a position to register the most accurate representation of sensory experience. These areas can encode the perceptual characteristics of specific sensory events, determine if the sensory features of complex entities such as words or faces are identical or not, and even

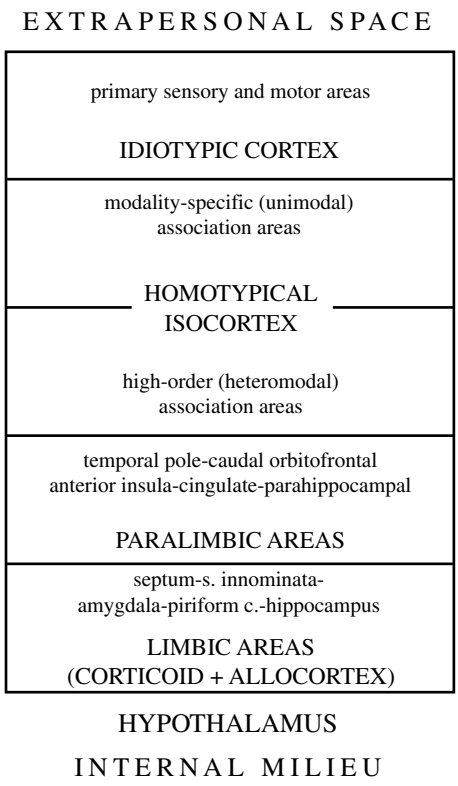


Figure 2 Functional zones of the cerebral cortex.

store all the necessary information in stable memory traces. However, in the absence of access to information in other modalities, unimodal areas do not have the ability to lead from word to meaning, from physiognomy to facial recognition, or from isolated sensory events to coherent experiences. Such integration of sensation into cognition necessitates the participation of *transmodal* areas.

The defining feature of a transmodal area is the ability to support cross-modal integration and, thus, a lack of specificity for any single modality of sensory processing. All components of heteromodal, paralimbic, and limbic zones are therefore also transmodal. A precise localization of transmodal areas became possible initially through the tracing of corticocortical connections in the monkey brain with the use of axonal degeneration methods. Experiments based on this methodology revealed hierarchically organized sets of pathways for linking sensory cortices to primary, secondary, and sometimes even tertiary modality-specific association areas, which in turn sent convergent projections to heteromodal sensory association zones.

The field of neuroscience had been primed to anticipate such a sequential organization through the work of David Hubel and Torsten Wiesel, who had demonstrated a hierarchy of simple, complex, and hypercomplex neurons in primary visual cortex, each successive level encoding a more composite aspect of visual information. The pattern of corticocortical connections in association cortex seemed to be extending this serial and convergent organization from the realms of sensation to those of cognition. A great deal of emphasis was placed on the pivotal role of multimodal convergence in all aspects of mental function, including the storage of memories, the formation of concepts, and the acquisition of language.

Although the importance of serial processing and multimodal convergence to cognitive function was widely accepted, some potentially serious computational limitations of such an arrangement were also acknowledged. Two of these objections are particularly relevant. First, if knowledge of α is to be encoded in convergent form by a small number of neurons, the brain would have to resolve the cumbersome problem of conveying α-related information in all relevant modalities to the one highly specific address where this convergent synthesis is located. Second, the modality-specific attributes of α would succumb to cross-modal contamination during the process of convergence and the sensory fidelity of the experience would be lost. This second circumstance can be likened to the mixing of yellow and blue to obtain green, a process that precludes the subsequent extraction of the original hues from the resultant mixture.

The surfacing of these concerns coincided with the development of newer and more powerful neuroanatomical methods based on the intraaxonal transport of horseradish peroxidase and tritiated amino acids. Experiments based on these methods started to show that the sensory-fugal flow of information was more complicated than previously surmised: There was a central thread of serial processing from one synaptic level to another, but there were also multiple parallel pathways, feedforward and feedback connections, and multiple sites for divergence and convergence. The synaptic templates based on this type of connectivity appeared to have much greater computational power. The objections to the convergent encoding of knowledge, for example, could be addressed by assuming that the principal role of transmodal areas is to create directories for linking (rather than mixing) modality-specific fragments of information into coherent experiences, memories, and thoughts. This alternative

process can be likened to obtaining green by superimposing a blue lens and a yellow lens that can then be separated from each other to yield the original uncontaminated colors. Transmodal areas appeared to fulfill two functions: (i) the establishment of limited cross-modal associations related to the target event, and (ii) the formation of a directory indicating the distributed components of the related information. Transmodal areas could thus enable the binding of modality-specific information into multimodal representations that protect the fidelity of the initial encoding.

Transmodal areas are not centers for storing convergent knowledge but, rather, critical gateways for integrating and accessing the relevant distributed information. They also provide "neural bottlenecks" in the sense that they constitute regions of maximum vulnerability for lesion-induced deficits in the pertinent cognitive domain. Each transmodal area displays a distinctive profile of behavioral specializations that is determined by its overall pattern of neural connections and physiological characteristics. For example, midtemporal cortex plays the role of a pivotal transmodal gateway for face and object recognition, Wernicke's area for lexical labeling, the hippocampoentorhinal complex for explicit memory, prefrontal cortex for working memory, the amygdala for emotion, and dorsal parietal cortex for spatial attention.

III. CHANNEL FUNCTIONS AND STATE FUNCTIONS

Many axonal pathways that interconnect one cortical area with another (or with specific sectors of the basal ganglia and thalamus) are organized in the form of reciprocal point-to-point *channels* where the principal sites of origin and the major fields of termination are of approximately equivalent size. This point-to-point connectivity provides the basic anatomical substrate of specific *channel functions*. Damage to channels such as the splenium of the corpus callosum, the frontotemporal uncinate fasciculus, or the insuloamygdaloid pathway leads to specific impairments such as pure alexia, amnesia, and asymbolia for pain. In addition to these point-to-point channels, each cortical area also receives diffuse modulatory connections. These pathways employ small amines as transmitters and determine the overall *state* of information processing rather than the contents of the information that is being transmitted along the point-to-point channels. These modulatory pathways play important roles in coordinating behavioral states related to arousal, attention, mood, and motivation. At least six such pathways can be identified in the primate brain. Five of these reach the cerebral cortex directly without a thalamic relay, whereas the sixth is relayed through the thalamus:

1. Cholinergic and GABAergic projections from the basal forebrain to the cerebral cortex
2. Histaminergic projections from the lateral and medial hypothalamus to the cerebral cortex
3. Serotonergic projections from the brain stem raphe nuclei to the cerebral cortex
4. Noradrenergic projections from the nucleus locus coeruleus to the cerebral cortex
5. Dopaminergic projections from the substantia nigra and the ventral tegmental area of Tsai to the cerebral cortex
6. Cholinergic projections from the brain stem reticular formation to the thalamus

Each of these modulatory pathways is organized in such a way that a relatively small group of neurons can induce rapid modulations in the information processing state of the entire cerebral cortex. The cholinergic projection from the brain stem to the thalamus, for example, promotes arousal by facilitating the transthalamic passage of sensory information toward the cerebral cortex. The other five modulatory pathways have direct access to the cerebral cortex without any thalamic relay. Each of these pathways displays a slightly different pattern of cortical distribution and physiological specialization.

The cholinergic innervation of the cerebral cortex arises predominantly from the nucleus basalis of the substantia innominata. The neurons of this nucleus are particularly responsive to novel and motivationally relevant sensory events. The major effect of acetylcholine on neurons of the cerebral cortex is mediated through the m1 subtype of muscarinic receptors and causes a prolonged reduction of potassium conductance so as to make cortical neurons more receptive to other excitatory inputs. Cortical cholinergic pathways also promote electroencephalograph desynchronization, long-term potentiation, and experience-induced synaptic remodeling. The ascending cholinergic pathway from the basal forebrain is therefore in a position to enhance the immediate neural impact and long-term memorability of motivationally relevant events. The nucleus basalis also gives rise to a GABAergic projection directed to the cerebral cortex. At least in

the rat, this pathway innervates inhibitory cortical interneurons but its functional specializations remain poorly understood.

The noradrenergic innervation of the cerebral cortex arises from the nucleus locus coeruleus. The neurons of this nucleus are particularly responsive to the motivational relevance of a stimulus. Neocortical norepinephrine increases the signal-to-noise ratio and timing precision of cortical neurons in a way that enhances the specificity of neural responses to significant sensory events.

The dopaminergic projections of the cerebral cortex arise from the substantia nigra and ventral tegmental area of Tsai. These dopaminergic cells are selectively responsive to motivationally relevant stimuli and to cues that signal their existence. In keeping with these characteristics, the dopaminergic projections from the ventral tegmental area to the cerebral cortex and nucleus accumbens appear to play an important role in mediating the neural processes related to substance addiction. Furthermore, both dopamine and acetylcholine promote cortical responses related to working memory.

The serotonergic projection to the cerebral cortex arises from the brain stem raphe nuclei. The electrical stimulation of these neurons can induce an arousal-related pattern of low-voltage fast activity in the cerebral cortex. Serotonin also appears to modulate the sensory gating of behaviorally relevant cues in the environment. The hypothalamus is the source of histaminergic projections to the cerebral cortex. Histamine receptors are widely distributed throughout the cerebral cortex and have been implicated in the regulation of cortical arousal, energy metabolism, autonomic function, and sensitivity to pain.

One anatomical feature common to all of these modulatory corticopetal projections is the absence of equally well-developed reciprocal projections from the cerebral cortex. The nucleus basalis, for example, projects to all parts of the cerebral cortex but receives cortical projections from only a handful of limbic and paralimbic areas. The other cortically projecting cell groups also receive sparse cortical projections, most of which come from limbic and paralimbic areas. This asymmetry of corticofugal versus corticopetal connections allows the neurons of the basal forebrain, substantia nigra–ventral tegmental area, brain stem raphe, and nucleus locus coeruleus to rapidly shift information processing states throughout the cerebral cortex in a way that is responsive primarily to the demands of the limbic system and internal milieu, with relatively little intervention from feedback loops emanating from heteromodal, unimodal, and primary cortices.

Many psychiatric diseases, including mania, depression, paranoia, obsessive–compulsive disorders, and chronic anxiety, are characterized by pathological biases in the interpretation of events and experiences. When compared to control subjects, for example, patients with generalized anxiety disorder show greater metabolic activation of temporal and frontal cortex during the passive viewing of neutral stimuli. Such altered responses, even to neutral stimuli, may indicate the existence of a fundamental processing state abnormality that biases the impact of experience. Indirect evidence based on the pharmacological treatment of depression and anxiety with noradrenergic and serotonergic agents and of paranoia with dopamine blockers suggests that the modulatory pathways of the cerebral cortex may play important roles in setting such fixed attitudinal biases in the processing of sensory experience.

It is unlikely that there will be a one-to-one relationship between any of these modulatory pathways and specific cognitive or comportmental domains. In general, however, the activation of these modulatory pathways provides a mechanism for augmenting the neural responses to novel and motivationally relevant events, facilitating their storage in memory, enhancing their access to on-line processing resources, sharpening the attentional focusing they elicit, and increasing their impact on consciousness. These projections are in a position to alter the tone, coloring, and interpretation of experience rather than its content. In addition to cholinergic and monoaminergic receptors, many areas of the cerebral cortex, especially components of the limbic system, also contain receptors for estrogen, testosterone, and other steroids. Alterations in the circulating level of these hormones, as in puberty or menopause, could influence behavioral states in a manner analogous to the effect of the modulatory projection systems.

The modulatory projections reviewed here are part of the ascending reticular activating system. These pathways highlight the multiple factors that contribute to the neural control of cognition, comportment, and consciousness. Language, spatial orientation, attention, memory, and emotion are all subserved by large-scale networks that contain multiple point-to-point channels. These pathways encode the perceptual, motor, visceral, and affective components of the relevant behavior and the way in which these components are interlinked. The modulatory pathways influence the processing states within which these

domain-specific channels function. In the course of remembering, for example, the content of what is recalled is determined primarily by the information that flows along the point-to-point sensory–limbic interconnections. On the other hand, the speed and efficiency of the recall, and perhaps the perspective from which the information is interpreted, may be regulated by the activity of the modulatory pathways that innervate the relevant regions of limbic and association cortex. In the realm of emotion, the linkage of a specific thought or event with a specific feeling and visceral state is determined by the point-to-point projections of the amygdala and related limbic structures, whereas the intensity of the emotion and its influence on other aspects of mental activity may be determined by the modulatory pathways. Among all the complex factors involved in the neural control of comportment and cognition, those that represent the contributions of these modulatory pathways are the most accessible to therapeutic manipulation by existing pharmacological agents.

IV. DISTRIBUTED LARGE-SCALE NETWORKS AND THEIR EPICENTERS

Transmodal nodes in midtemporal cortex, Wernicke's area, posterior parietal cortex, prefrontal cortex, amygdala, and the hippocampoentorhinal complex link distributed information into coherent multimodal assemblies necessary for face and object recognition, naming, working memory, spatial attention, emotional channeling, and explicit memory. These transmodal areas provide the cortical epicenters of large-scale distributed networks.

The pattern of connectivity summarized in Fig. 3 is based on anatomical experiments related to the network for spatial attention in the monkey and may reflect an organization that is common to all large-scale networks. In Fig. 3, A and B represent two interconnected epicenters of any large-scale neural network. They could represent the frontal eye fields and posterior parietal cortex in the network for spatial attention, midtemporal and temporopolar cortices in the network for face and object recognition, the amygdala and the hippocampoentorhinal complex in the emotion/memory network, Wernicke's area and Broca's area in the language network, and prefrontal cortex and posterior parietal cortex in the working memory/executive function network. Axonal transport experiments in the spatial attention network of the monkey indicate that if one member of such a pair, for example, A, is interconnected with additional cortical areas such as 1–3, then B is also interconnected with the same three cortical areas. Consequently, if A transmits a message, B will receive it directly but also through the alternative vantage points provided by areas 1–3. This arrangement enables parallel processing and contains multiple nodes where transitions

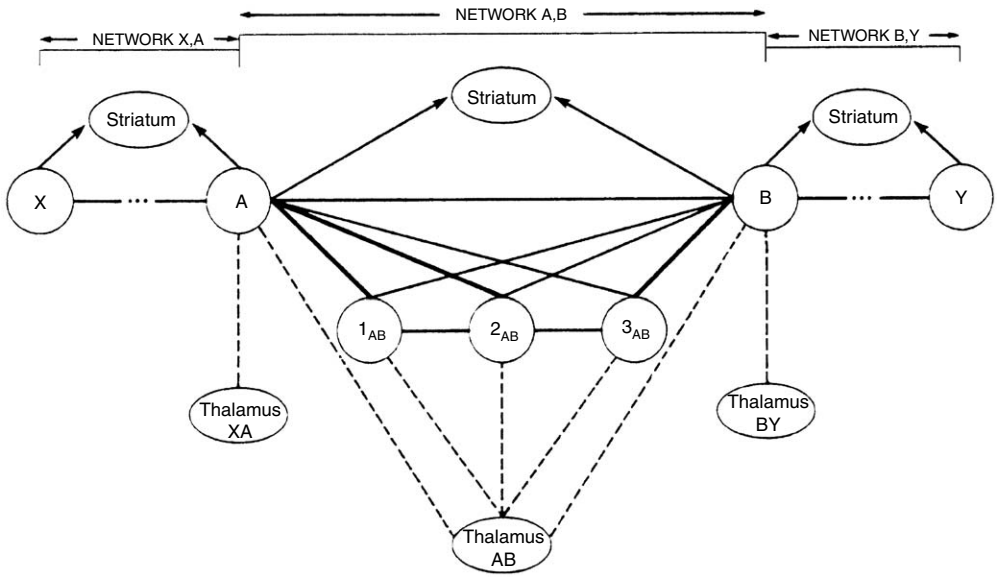


Figure 3 General organizational principles of large-scale neurocognitive networks.

between parallel and serial processing can occur. In resolving a complex cognitive problem such as reconstructing a past memory, selecting words to express a thought, or figuring out the identity of a face, a set of cortical areas interconnected in this fashion can execute an extremely rapid survey of a vast informational landscape while considering numerous goals, constraints, scenarios, and hypotheses until the entire system settles into a state of least conflict that becomes identified as the solution to the cognitive problem.

Because cortical areas tend to have very extensive corticocortical projections, individual sectors of association cortex are likely to belong to multiple intersecting networks. With rare exceptions, however, thalamic subnuclei have almost no connections among each other and some thalamic subnuclei can project to both epicenters of an individual large-scale neural network. Thalamic subnuclei can thus fulfill the very important role of setting coactivation boundaries for individual networks. Neuroanatomical experiments have shown that interconnected cortical areas are likely to send interdigitating projections to the striatum. Since the striatum receives cortical inputs but does not project back to the cerebral cortex, it could serve the role of an efference synchronizer (or filter) for coordinating the outputs of cortical areas in a given network. The human brain contains at least the following five large-scale neurocognitive networks that follow these principles of organization:

1. Dorsal parietofrontal network for spatial orientation: The cortices around the intraparietal sulcus and the frontal eye fields constitute the two major interconnected epicenters. The parietal component displays a relative specialization for the perceptual representation of salient events and their transformation into targets for attentional behaviors, and the frontal component displays a relative specialization for choosing and sequencing exploratory and orienting movements. Additional critical components are located in the cingulate gyrus, striatum, and thalamus. Damage to this network yields deficits of spatial attention and exploration such as contralesional hemispatial neglect, simultanagnosia, and other manifestations of spatial disorientation. Contralesional neglect occurs almost exclusively after right-sided damage to this network, whereas simultanagnosia tends to arise after bilateral lesions.

2. Limbic network for memory and emotion: The hippocampoentorhinal complex and the amygdala constitute the two interconnected epicenters. The former displays a relative specialization for memory and learning and the latter for drive, emotion, and visceral tone. Additional critical components are located in the paralimbic cortices, the hypothalamus, the limbic thalamus, and the limbic striatum. Damage to this network yields deficits of memory, emotion, affiliative behaviors, and autonomic regulation. Severe deficits usually occur only after bilateral lesions. Occasionally, unilateral left-sided lesions give rise to a multimodal amnesia but this is transient. Frequently, unilateral lesions in the left give rise to prominent deficits of verbal memory, whereas unilateral lesions on the right give rise to nonverbal memory deficits that are usually quite mild.

3. Perisylvian network for language: The two epicenters of this network are known as Broca's area and Wernicke's area. Broca's area includes the premotor region BA 44 and the adjacent heteromodal fields of BA 45–47; Wernicke's area includes the posterior part of auditory association cortex in BA 22 and also adjacent heteromodal fields in BA 39–40 and BA 21. Broca's area displays a relative specialization for the articulatory, syntactic, and grammatical aspects of language, whereas Wernicke's area displays a specialization for the lexical and semantic aspects. Additional components of this network are located in the striatum, thalamus, and the association areas of the frontal, temporal, and parietal lobes. Damage to this network yields aphasia, alexia, and agraphia. Such deficits are seen only after damage to the left hemisphere in the majority of the population.

4. Ventral occipitotemporal network for object recognition: The middle temporal gyrus and the temporal pole appear to contain the transmodal epicenters for this network. Additional critical components are located in the fusiform gyrus and inferior temporal gyrus. Damage to this network yields recognition deficits such as object agnosia and prosopagnosia. The lesions that cause such deficits are almost always bilateral. The fusiform gyrus is the most common site of lesions, probably because it is the only part of this network with a vascular supply that makes bilateral damage likely. Occasionally, unilateral left-sided lesions can lead to object agnosia and unilateral right-sided lesions to prosopagnosia.

5. Prefrontal network for executive function and comportment: Prefrontal heteromodal cortex and orbitofrontal cortex are the major cortical epicenters involved in the coordination of comportment, working memory, and related executive functions. The head of the caudate nucleus and the mediodorsal nucleus of the thalamus constitute additional critical components. Deficits of comportment are frequently associated

with lesions of orbitofrontal and adjacent medial frontal cortex, whereas deficits of executive function and working memory are frequently associated with damage to dorsolateral prefrontal cortex. Clinically significant deficits are usually seen only after bilateral lesions. Occasionally, unilateral left-sided lesions give rise to a syndrome of abulia, whereas unilateral right-sided lesions give rise to behavioral disinhibition.

Neuroanatomical experiments in the homologous regions of the monkey brain have shown that the components of these five networks are interconnected according to the pattern shown in Fig. 3. As noted previously, each of these networks receives its sensory information from a common set of unimodal cortical areas. The differences in the resultant cognitive functions are determined by the anatomical location and specializations of the relevant transmodal epicenters. The large-scale network approach predicts that many, if not all, network components will be activated in concert during the performance of any task in a given cognitive domain. In keeping with this prediction, tasks related to spatial awareness, language, working memory, explicit memory, and object identification in human subjects have led to the collective activation of the relevant epicenters noted previously. Functional imaging experiments cannot yet determine whether all network components are activated simultaneously or if the temporal sequence of activation varies according to the nature of the task. Such questions can be addressed by combining functional imaging with event-related potentials.

Large-scale neural networks are organized according to the principles of selectively distributed processing. For example, Wernicke's area occupies the lexical/semantic pole of the language network but also participates in articulation and syntax, whereas Broca's area occupies the articulatory/syntactic pole of the network but also participates in phonological discrimination and lexical access. In the case of spatial attention, the frontal eye fields occupy the motor/exploratory pole of the relevant network but also participate in the compilation of perceptual representations, whereas posterior parietal cortex occupies the sensory/representational pole but also participates in the programming of exploratory movements. In the limbic network, the hippocampal complex is most closely related to explicit memory but also plays a role in emotional modulation, whereas the amygdaloid complex is most closely related to emotional modulation but also participates in the encoding of emotionally salient memories. This organization promotes flexibility without sacrificing regional functional segregation.

At least two levels of connectivity contribute to the functional organization of neurocognitive networks. At one level, genetically encoded and relatively fixed axonal connections specify the type of information that a given region will process. These connections determine the location and functional specialization of network components. At a second level, experience-induced modifications of synaptic strengths enable the gradual accumulation of an experiential base that is unique for each individual. This process is known as neuroplasticity. Although we tend to think of plasticity as a phenomenon confined to early development, dendritic and axonal remodeling occurs throughout life and allows each individual brain to establish new associations, adapt to new situations, and compensate for biological attrition. A deeper understanding of these more dynamic aspects of brain structure is a major goal of behavioral neuroscience.

See Also the Following Articles

CHEMICAL NEUROANATOMY • NERVOUS SYSTEM, ORGANIZATION OF • NEURAL NETWORKS

Suggested Reading

Brodmann, K. (1994). *Localisation in the Cerebral Cortex*. Smith-Gordon, London.

Mesulam, M.-M. (1994). Higher visual functions of the cerebral cortex and their disruption in clinical practice. In *Principles and Practice of Ophthalmology* (D. M. Albert and F. A. Jakobiec, Eds.), pp. 2640–2653. Saunders, Philadelphia.

Mesulam, M.-M. (1998). From sensation to cognition. *Brain* **121,** 1013–1052.

Mesulam, M.-M. (2000). Behavioral neuroanatomy: Large-scale networks, association cortex, frontal syndromes, the limbic system and hemispheric specialization. In *Principles of Behavioral and Cognitive Neurology* (M.-M. Mesulam, Ed.), pp. 1–120. Oxford Univ. Press, New York.

Mesulam, M.-M., and Geula, C. (1988). Nucleus basalis (Ch4) and cortical cholinergic innervation in the human brain: Observations based on the distribution of acetylcholinesterase and choline acetyltransferase. *J. Comp. Neurol.* **275,** 216–240.

Mesulam, M.-M., and Mufson, E. J. (1982). Insula of the old world monkey. I. Architectonics in the insulo-orbito-temporal component of the paralimbic brain. *J. Comp. Neurol.* **212,** 1–22.

Pandya, D. N., and Yeterian, E. H. (1985). Architecture and connections of cortical association areas. In *Cerebral Cortex* (A. Peters and E.G. Jones, Eds.), pp. 3–61. Plenum, New York.

Brain Damage, Recovery from

PAUL BACH-Y-RITA
University of Wisconsin, Madison

GLOSSARY

brain plasticity The adaptive capacities of the brain; its ability to modify its own structure, e.g., organization and functioning.

nonsynaptic diffusion neurotransmission The diffusion through the extracellular fluid of neurotransmitters released at points that may be remote from the target cells, with the resulting activation of extrasynaptic receptors.

late brain reorganization Changes in sensory and motor representation, as well as other cerebral functions, that occur 2 or more years after damage to the brain.

functional rehabilitation Rehabilitation programs for persons with brain damage that are based on motivating activities relating to real-life activities.

receptor plasticity The up- and downregulation of synaptic and/or nonsynaptic neurotransmitter receptors, such as those on the neuron or glia surface.

There is generally at least some recovery after the brain is damaged. Various mechanisms contribute to early and late reorganization of the brain. Much of the recovery, which is maximized by appropriate rehabilitation, occurs within months of the damage, but functional gains can be obtained even many years after the damage has occurred. The degree of recovery depends on many factors, including age, the brain area and amount of tissue damaged, the rapidity of the damage, the brain's mechanisms of functional reorganization, and environmental and psychosocial factors. If recovery is not expected or sought with the active participation of the disabled person, little recovery is obtained.

I. BRAIN REORGANIZATION

The brain can reorganize on the basis of the structures that remain after brain damage, possibly mobilizing mechanisms such as unmasking of previously present but relatively "weak" neural paths, neuronal sprouting, and the up- and downregulation of receptors at synaptic and nonsynaptic sites. Both human and animal studies have demonstrated mechanisms of brain plasticity and the reorganization of function following brain damage. A *Science* Research News article noted that "a striking body of recent work suggests that the adult brain can reorganize itself in areas that were long thought to be completely 'hardwired.'"

The neurosciences have in general been descriptive and have been influenced primarily by concepts of strict localization of function in the brain, the synapse as the only important means of cell-to-cell communication, and the irreversibility (beyond a certain period of time) of functional deficits produced by brain damage. Until recently, evidence supporting plasticity was generally ignored, and plasticity concepts had little effect on the development of clinical procedures. The functional results of ablating some area of cortex, or of eliminating some sensory input, have been interpreted as meaning that that particular part of the cortex is essential for a particular function, or that particular sensory input is required during a particular

developmental stage for a particular function to develop.

Although such studies are necessary for an understanding of the baseline functions of the central nervous system, the following is the critical plasticity-related question: What is the reorganization capacity of the central nervous system (or some particular subset of it)? When faced with the evidence that a particular type of brain injury ''permanently'' eliminates some function, a brain plasticity-influenced approach is to ask the following: What can be done to reorganize the remainder of the brain to enable that function to be regained? Examples of this include the bilateral motor control that can be developed following hemispherctomy and the recovery from facial paralysis that can be obtained with appropriate rehabilitation even many years after the causative event. In the absence of a particular sensory input (e.g., vision through the eyes), what can be done to enable the brain to interpret comparable information (e.g., vision through a tactile vision substitution system using nonvisual pathways to the brain and nonvisual cortex structures to receive the sensory projection)? Not only experimental evidence but also clinical observations can be interpreted within the framework of brain plasticity only when plasticity is part of the conceptual substance of the neurosciences.

Although brain plasticity is generally positive for recovery, some negative effects, such as the development of spasticity and kindling causing epilepsy, can occur. Furthermore, the presence of mechanisms for plasticity does not automatically lead to recovery of function; appropriate rehabilitation and the appropriate physical and psychosocial environments can play decisive roles in the reorganization.

Rehabilitation and the developmental learning process have some aspects in common. Both include important elements of inhibition in regard to selective function and precisely coordinated movements. For example, a child learning to write initially demonstrates electromyographic activity in virtually all the muscles related to the hand. As ability increases, muscle activity decreases progressively until there is a minimum of activity, which is coordinated precisely to produce just the muscle action necessary for writing. It then becomes virtually fatigue-free. Following brain damage, coordinated movements are often disturbed. Patients become fatigued when attempting controlled movements; rehabilitation is then oriented toward the restoration of precise, fatigue-free movements. Comparably, reflexes that are normal shortly after birth (e.g., Babinski's) are inhibited during maturation and can reappear following brain damage. Studies on unmasking may relate to the balance of excitation and inhibition and may involve increased excitability of surviving neural connections following the loss of some inputs due to neural damage.

Most brain-damaged patients receive rehabilitation during a stage in which they would have demonstrated some spontaneous recovery even in the absence of specific rehabilitation, especially if some of the complications had been prevented (e.g., passive movements to prevent contractures, which may have been an important factor in the remarkable recovery in a series of monkey decortication experiments). Therefore, in addition to the difficulties caused by the absence of accurate and reliable functional assessment methodologies, it is difficult to accurately quantify the effect of an early rehabilitation program since its results are achieved simultaneously with the spontaneous recovery. Brain reorganization associated with rehabilitation can be demonstrated with models of late rehabilitation. Both animal and human models have been developed to evaluate functional recovery.

A. Postulated Mechanisms of Reorganization of Function

Undoubtedly, many mechanisms intervene in the reorganization following damage to the brain. In the first stages, it is likely that surviving cells that have been totally or partially denervated may have one or more of three responses: (i) the ''strengthening'' of synapses from secondary connections (unmasking); (ii) the development of extrasynaptic receptors on the membrane of the surviving cells, which respond by means of nonsynaptic diffusion neurotransmission (NDN) mechanisms; and (iii) receptor development related to the new synapses formed by the sprouting of processes from surviving cells. The contributions of stem cells and neurogenesis have yet to be evaluated, but exciting recent findings open the possibility that they may contribute to functional recovery.

Some mechanisms of brain reorganization are briefly discussed in the following sections. Some may be active in both early and late reorganization.

1. Multiplexing and Unmasking

Multiplexing in the brain consists of the multiple uses of neurons and fibers so that they participate in various functions. Many studies have demonstrated multiple sensory and motor representations of a single brain

region and overlap of representation. This may provide the neural substrates for plastic changes with training. Individual neurons in the brain, such as in the nucleus pontis oralis, can respond to many types of sensory stimuli, and there is evidence that activity produced by one input affects the responses to the other potential inputs. It may be a mechanism for minimizing the number of cells needed and may relate to a proposed law of the conservation of space and energy in the brain. Individual somatosensory neurons in the cortex also have the potential to respond to a wide array of inputs from widespread surfaces on the skin. The rapidly occurring changes in the functional topography of the cortex have been proposed to be mediated by the unmasking of the latent excitatory connections via inhibition of local inhibitory synapses. It has been proposed that injury initiates rapid, concurrent, consistent, and reiterative changes in brain stem and cortical maps, and that changes in thalamic maps may also occur.

The role of unmasking may be of considerable importance in functional reorganization. A series of studies have demonstrated the unmasking of normally ineffective connections that may become active if the dominant inputs are put out of action. In one study, peripheral nerves were cut, removing the normal sensory input to a population of spinal cord cells, without anatomical evidence of degeneration or vacated synaptic sites. Large numbers of cells in a region of the cord that were normally dominated by afferents from the foot and toes began to respond to other areas of the leg days or weeks after section of the peripheral axons that previously supplied their excitatory drive. Comparable changes have been demonstrated in monkey somatosensory cortex following restricted deafferentation. Specific finger exploratory training of monkeys produces a marked increase in the cortical sensory representation of the fingers used for the haptic exploration, which appears to indicate the unmasking of previously existing connectivity. That study also reinforces the concept of overlapping representation; other studies have shown that in monkeys the territories controlling different fingers overlap. Control of any finger movement appears to utilize a population of neurons distributed throughout the M1 hand area rather than a somatotopically segregated population. Human functional magnetic resonance imaging studies have strongly suggested that in the primary hand region around the central sulcus, the same neuronal population is active in the three tasks studied (active finger apposition, texture on fingers, and haptic exploration).

The unmasking-related reorganization can be practically instantaneous. During temporary altered sensory input, the injection of local anesthetic into the receptive fields of neurons of the dorsal column neurons uncovered new receptor fields that emerged within minutes, probably due to the unmasking of previously ineffective inputs. Another of the important studies of unmasking demonstrated cortical plasticity related to training: Both blind and sighted persons reading Braille had expanded sensorimotor cortical representation of the reading finger, and cortical representation is affected by learning to play the piano. One of the most dramatic examples showed that after limited sensory deafferentation in adult primates, somatosensory cortical maps reorganize, over a period of years, over distances previously thought to be impossible (more than 1 cm).

Interesting direct evidence for the unmasking of pathways in monkeys has been provided by a study of monkeys that had spent the first year of life with the eyelids sutured. Microelectrode studies of cells in area 19 revealed that 20% of the cells responded to somesthetic stimuli, whereas in the normal monkey no somesthetic responses are recorded. It had previously been shown that almost half of the visual cortical cells that responded to visual stimuli also responded to auditory and/or pinprick stimuli, but these nonvisual responses had considerably longer latencies and were easily blocked. In blind persons, the visual cortex is metabolically very active in response to auditory and tactile stimuli: Such activity may represent the unmasking of nonvisual inputs to the visual cortex. The pathways related to those responses may be similar to those unmasked by the lid suture.

Comparable changes may occur following brain damage. A specific human case of recovered function has been interpreted in the context of unmasking: Following brain damage, the unmasking activated previously existent pathways that (previous to the injury) had not had the same relationship to the function. The extensive damage included the destruction of a pyramidal tract that, at autopsy 7 years after the stroke, was composed of scar tissue except for approximately 3% of normal appearing fibers scattered through the scar tissue. The recovery was interpreted in terms of the possible unmasking of pathways to the dendrites of the cells with long pyramidal tract fibers and their axonal ramifications. It is possible that connections to large numbers of motoneurons that had previously been relatively inactive were rendered active, possibly in part due to postinjury receptor plasticity in those motoneurons.

Unmasking can be produced by some pharmacological agents: In cats anesthetized with chloralose, pyramidal stimulation not only increased the size of the cutaneous receptor fields of the recorded cortical neurons but also revealed responsiveness to sensory modalities that did not produce responses in the absence of the pyramidal stimulation. Chloralose increased the size of cortical evoked potentials in zones of sensory convergence, and in primary visual cortex, the activity of cells before, during, and after the administration of thiopental anesthesia increased the receptive fields and types of stimuli to which cells respond. Furthermore, thiopental made previously unreceptive cells respond to visual stimuli. In all these cases, apparently the drugs unmasked pathways that already existed but were either inhibited or were not sufficiently active to elicit responses before drug administration.

2. Nonsynaptic Diffusion Neurotransmission and Receptor Plasticity

A postulated example of this class of responses is the dopamine receptor upregulation that had been demonstrated in human stroke patients following the destruction of dopamine pathways (and the synaptic and nonsynaptic dopamine release sites). Upregulation of receptors may have occurred on the extrasynaptic membrane (comparable to the response noted on denervated muscle fibers) following the loss of those dopamine release sites. This may have led to supersensitivity to the neurotransmitters in the extracellular fluid, with activation by nonsynaptic transmission.

Information in the brain appears to be transmitted both by synaptic connectivity and by NDN. NDN includes the diffusion through the extracellular fluid of neurotransmitters released at points that may be remote from the target cells, with the resulting activation of extrasynaptic receptors. The existence of many receptor subtypes offers the possibility of selective neurotransmission at a distance by NDN.

Individual movements or functions, such as playing the piano or watching a tennis game, require great selectivity, rapid initiation, and rapid ending; for such functions, synaptic action is essential. However, for mass sustained functions (e.g., sleep, mood, and hunger), sustained, widespread activity (rather than speed and selectivity) is required, which may be largely mediated by NDN. Many functions may be produced by combinations of both types of neurotransmission. For example regarding piano playing, in addition to the relevant synaptic mechanisms, the finger movements can be more precise in the presence of adequate preparation including changes in brain tone (probably mediated by noradrenaline). Also, the visual perception of the tennis game may require neuronal receptivity to be set at a high level, probably involving several neurotransmitters, including nitric oxide and dopamine in the retina, serotonin and histamine in the lateral geniculate nucleus, and noradrenaline in the visual cortex. These effects appear to be primarily nonsynaptically mediated. Some of them have been called modulation; the modulation of synaptic activity by diffusion outside the synaptic gap is also a nonsynaptic, diffusion-mediated activity.

NDN can be modeled by students in a university classroom, who can be equated to neurotransmitter molecules in a vesicle. Upon release, they must go to specific other classrooms throughout the campus (receptor sites). They flow out into the halls and the grounds between buildings (extracellular fluid), where they mix with other students (neurotransmitter molecules) from other classrooms (vesicles) going to other target classrooms. They walk (diffuse) to their specific classrooms (receptors), which they enter (bind). In contrast, "synaptic transmission" students would be propelled along enclosed walkways connecting the point-of-origin classroom with each target classroom.

This conceptual model of information transmission in the brain, involving both synaptic transmission and NDN, may have considerable relevance to the functional reorganization following brain damage. Receptor plasticity, both at synapses and on the cell membrane away from synapses (reached by nonsynaptic diffusion neurotransmission), may play a major role in the reorganization of function following brain damage.

NDN may be the principal means of neurotransmission in the noradrenergic system, which is involved in many activities related to recovery from brain damage. Both acetylcholine and norepinephrine can provide a state of excitability consistent with cognition, which is consistent with inhibition of the locus ceruleus activity during lack of vigilance. These findings may directly relate to the results of rehabilitation programs: When vigilance is high and the patient is actively involved in the rehabilitation program, good results are more likely to be obtained. There may also be a relationship to functional rehabilitation programs that are based on the interests of the individual patient and to the positive results obtained with some home programs. In all these cases, the increased vigilance and participation may lead to greater locus ceruleus production of

noradrenaline. This and other neurotransmitter changes may also be mechanisms by which psychosocial factors influence recovery.

In 1949, Hebb developed the concept of brain "cell assembly," which continues to be a major model of brain function. He considered that any frequently repeated particular stimulation will lead to the slow development of a cell assembly, a diffuse structure capable of acting as a closed system. Hebbian cell assemblies consist of enormous numbers of individual cells, with every cell connected to every other cell. With such an architecture, however, the length of the links among cells and the resulting volume of the assembly would easily exceed space constraints. In contrast, a "wireless" mechanism, such as NDN, might be consistent with volume limitations. However, the space and energy considerations in synaptic and nonsynaptic neurotransmission have been calculated, and it has been shown that synaptic connectivity in cell assemblies would be too costly in terms of the metabolic energy and the space required, thus suggesting NDN to be a less expensive means for intercellular communication.

3. Extracellular Space Volume Fraction

The extracellular space in the brain plays a role in many functions, including nonsynaptic diffusion neurotransmission. In an assembly of cells in the brain, the distance between neurons can be reduced by 50% with neuron activity that causes them to swell. This has an effect on the excitability and metabolism of the cells by means of changes in the distance between the neurons, which produces changes in ionic concentrations and dynamics.

Changes in the size of the extracellular compartment [volume fraction (VF)] may play a role in membrane excitability in pathological brain states such as epilepsy, brain damage, and in the survival of partially denervated neurons during the postinjury period of receptor upregulation that can lead to reorganization of brain function by unmasking and other mechanisms. Under pathological conditions such as anoxia, the extracellular VF is reduced; it is also reduced (by up to 50%) in hyperexcitability, by changes in the concentration of potassium, and with epileptiform discharges. However, it is also possible that hyperexcitability due to a VF decrease, either independently or in combination with excitotoxic activity, may increase secondary cell death following brain damage.

4. Reactive Synaptogenesis

The third class of responses to damage, is usually called "sprouting." However, a case can be made for calling this "reactive synaptogenesis" when considering recovery of function. Although reactive synaptogenesis may participate in the recovery of function in some cases, it may also compete with the process of restoration by introducing aberrant connections.

It has been proposed that the takeover of muscle fibers when motor neuron loss occurs, such as in polio, may occur through vestigial pathways since in the embryonic stage the muscle fibers are polyneuronally innervated. Although all connections except those from one motor neuron disappear shortly after birth, vestigial remains of the polyneuronal pathways could conceivably serve as tracks for the growth of pathways from surviving motoneurons to the muscle fibers denervated by the motoneuron loss. If such pathways could be demonstrated in the nerve–muscle fiber preparation (which has served very well as a model of central nervous system nerve–cell connectivity), it would be interesting to consider the implications for central nervous system repair mechanisms. In particular, reactive synaptogenesis would have to be considered in the context of the reestablishment of connections that had existed during an early stage of brain development.

Strong evidence for reactive synaptogenesis associated (with appropriate rehabilitation) with functional recovery has emerged from a series of studies on hemispherectomized cats. Hemishperectomy has also been a model for recovery of function for other animal and human studies since remarkable recoveries have been recorded both in experimental animals and in young and adult humans. The outstanding feature of neonatally and adult lesioned cats was the presence of terminal fields in the nucleus ruber (RN) of both sides. The terminals in the RN contralateral to the cortical injection of radioactive material (leucine–proline) for autoradiographic labeling of pericruciate cortical projections to the RN revealed a normal pattern of new innervation for the RN partially deafferented by the hemispherectomy. Thus, an extensive reorganization of the brain may account for the remarkable recovery of function.

5. Diaschisis

Diaschisis relates recovery of function to the recovery from the neural depression of sites remote from, but connected to, the site of injury. Changes in

neurotransmitter function following injury have been suggested as a mechanism by which diachisis could operate. Following a unilateral sensorimotor cortex injury in rats, a depression in norepinephrine function occurs in the cerebellum contralateral to the lesion at 1 day postinjury. The norepinephrine depression, and the hemiplegia, can be resolved by infusions of norepinephrine into the contralateral, but not the ipsilateral, cerebellum. The recovery is maintained permanently. The anatomical mechanism may be via the simultaneous projection of individual locus coeruleus cells to both the contralateral cerebellum and the ipsilateral sensorimotor cortex. Damage to the sensorimotor cortex also damages fibers to the locus coeruleus, which may shift the body's energy from neurotransmitter production to protein synthesis for repair of the damaged terminal. While the body attempts repair, the undamaged terminals elsewhere (e.g., in cerebellum) may not be not functioning normally. However, long-term mechanisms are also probably involved since crossed-cerebellar inhibition has also been demonstrated in humans several years after brain lesion by positron emission tomography studies.

6. Trophic Factors

Trophic factors such as nerve growth factor have also been shown to be related to cell survival and recovery after brain damage. Transmembrane glycoproteins (e.g., integrins), gangliosides, and putrecine have been suggested to have roles in recovery of function.

7. Synapsins

Changes in synapsins may provide a mechanism for brain plasticity, based on regulation of neurotransmitter release. Synapsins are neuronal phosphoproteins that coat synaptic vesicles and bind to the cytoskeleton. Membrane depolarization and neurotransmitter release have been correlated with phosphorylation of the synapsins. The synapsins, a novel class of actin-binding proteins, are located in the presynaptic nerve terminal and contribute a significant portion of the total synaptic vesicle protein. They may connect synaptic vesicles to each other, to the cytoskeletal, or both, and they may play a role in positioning synaptic vesicles in the nerve terminal, thus regulating transmitter release. Studies on synapsins may form a framework for understanding how these molecules function in organizing presynaptic intracellular space, and they could provide a mechanism for reorganizing function.

8. Parallel and Adjoining Cortical Representation Areas

Parallel motor areas could substitute for lost function in brain-injured persons. In a study of the anatomy of motor recovery assessing motor function in 23 patients following capsular or striatocapsular stroke, small capsular lesions, which can disrupt the output of functionally and anatomically distinct motor areas selectively, were considered to reveal the capability of the different motor pathways to substitute each other functionally in the process of recovery from hemiparesis. Contribution of the undamaged hemisphere to the process of recovery was suggested by both electrophysiological and metabolic studies, mediated by either bilateral pathways or uncrossed or recrossing pyramidal tract fibers, which may also play an important role. The reorganization of parallel-acting multiple motor areas ipsilateral to the lesion site was considered to be the central mechanism in motor recovery.

Motor and sensory cortical representation can expand into adjoining cortical areas either following a lesion (e.g., limb amputation) or following specific training, such as in monkeys highly trained in tactile tasks. The cortical representational plasticity may play a role in functional recovery following brain damage.

9. Emergent Mechanisms

Recent findings that may increase our understanding of the mechanisms of brain reorganization include the isolation of stem cells, which are the primordial cells from which all others evolve. They can potentially be made into neural cells that can migrate through the brain to where they are needed after a lesion or that can be implanted in the brain. Neural cells grown in labs have been implanted in the brains of persons who have had a stroke, with resulting improved function. It has been shown in both adult and senescent animals, and adult humans, that neurogenesis can occur (at least in the hippocampus). Adult neural stem cells have been shown to have very broad developmental capacities that may potentially be used to generate a variety of cell types for transplantation. Regeneration has been shown to be possible in some brain regions

We are currently at a stage in which, although much new evidence is accumulating regarding these

emerging research areas, conclusions as to their roles in recovery from brain damage cannot be drawn.

B. Psychosocial and Environmental Factors in Recovery of Function

Family support, mood, the environment, resistance to change, the attitudes of the rehabilitation professionals, and hope and expectations for recovery are some of the factors related to recovery. These factors appear to be intimately related to neurotransmitters and to the effectiveness of rehabilitation programs. The type of physical environment in which the brain-damaged person is placed may be closely related to the recovery.

The most important functional gains in the animal model of traumatic brain injury were obtained with an enriched environment. Most animal studies are undertaken with environmentally deprived laboratory animals, which may distort the experimental results. For example, in beam-walking studies, laboratory rats had difficulty crossing a gap on a 2-cm-wide strip of wood, whereas rats in the wild can scamper across much narrower strips. In addition to the physical environment for rehabilitation, and the content and timing of the rehabilitation programs, psychosocial factors and fitness level can affect outcome of rehabilitation programs. Fitness also affects many functional measures: Physically active men have been shown to have shorter event-related cortical potentials, stronger central inhibition, better neurocognitive performance, and better visual sensitivity.

The excellent functional recovery noted in some unusual cases of recovery from brain damage with home rehabilitation programs may be related not only to neuroplasticity factors but also to psychosocial and environmental considerations and to the functionality of the rehabilitation program.

C. Temporal Factors

Lesions of the brain early in life have effects that differ from those that occur in adulthood. Functional results can be either better or worse, depending on the age at injury and on the area injured. Furthermore, the age at which the lesion occurs may influence motor as well as behavior lesion effects. Damage in infancy yields more profound socioemotional effects than does damage in adulthood. Compensatory mechanisms do not always operate to ensure recovery of functions after early brain damage.

Early lesions (such as in persons with congenital hemispherectomy), resulting in acquisition of language in the right hemisphere, may interfere with right hemisphere nonverbal functions due to the "crowding effect"; this may be a case of functional plasticity being of limited advantage to individuals with regard to total cognitive capacities.

Tactile stimulation with premature babies leads to faster growth. Similarly, tactile stimulation of brain-lesioned laboratory rats led to unexpectedly large attenuation of the behavioral deficits, correlated with reversal of atrophy of cortical neurons normally associated with these lesions. Both premature infants and laboratory rats have in common that both are in impoverished environments. Thus, the positive responses with tactile stimulation can be considered to be related to an improvement in the direction of a normal environment.

Reorganization and recovery of function does not cease at any arbitrary time, such as 6 months; the potential can exist for many years after injury. It is becoming an important area in the field of neurologic rehabilitation since it appears that, in humans, specific late (postacute) rehabilitation programs are necessary to exploit that potential.

D. Neuroplasticity in the Aging Brain

Neurotransmitter levels and mechanisms in the aged brain appear to vary from those in young and adult brains in many ways. Implants of young tissue into aged hosts, and those of aged tissue into young hosts, have been studied in order to determine the relative importance of intrinsic versus extrinsic influences in such factors as age-related adrenergic defecits. Receptor plasticity, the up- or downregulation of receptors for specific neuroactive substances, is an important mechanism of neuroplasticity. Many changes in brain function lead to receptor changes since they must result in changes in neurotransmitters and other neuroactive substances. One study of receptor plasticity of the aged human brain demonstrated upregulation of dopamine D1 receptors in the brains of three persons, 80, 81, and 87 years of age who had died respectively 9, 19, and 27 days following a unilateral infarct of the ventral midbrain, producing a relative dopamine depletion on the lesioned side. In the autopsy material, an increase of between 27 and 37%

of dopamine receptors was demonstrated on the lesioned side. The receptor plasticity demonstrated in stroke patients has been discussed in regard to the possible role of NDN in the functional results of that plasticity. The D1 receptor upregulation may occur on the nonsynaptic cell membrane following the loss of synapses due to the partial denervation. That and comparable up- and downregulation may result in adaptive or maladaptive responses to the brain damage: Many neurotransmitter substances have been demonstrated by microdialysis in the extra cellular fluid, and some of these may produce responses in cells hypersensitive due to receptor plasticity following denervation. A possible maladaptive response is convulsive activity; it has been reported that 10–15% of stroke patients have convulsions.

Recovery of function following brain damage in the aged is also evidence for neuroplasticity; unusual cases of recovery stimulate investigators to explore the mechanisms of such recovery and the means of mobilizing those mechanisms to obtain the maximum possible recovery of function following brain damage. Several unusual cases have been described in the literature that reveal plasticity in the aged brain. Among these is the recovery of function in a patient who, at 65 years of age, had major brain damage (demonstrated on autopsy 7 years later) but who made a good recovery with a home-based rehabilitation program during the 5 years following the stroke. Aged patients with facial paralysis showed significant improvement with a rehabilitation program started years after the damage.

II. NEUROLOGIC REHABILITATION

Clinical rehabilitation has developed in an ad hoc fashion. The first formal stroke rehabilitation method, published by Frenkel in the mid- 19th century, emerged from a program that a nonprofessional wife had developed to successfully rehabilitate her husband. Rehabilitation clinicians still rely on nontheory-related methodologies. Little has been written about the fact that rehabilitation as currently practiced has such meager carryover to real-life activities, and even little carryover from one session to another. This has had the effect of reinforcing the widespread view in the physical rehabilitation field that once a patient reaches a plateau, usually 6–12 months after a stroke, further administration of rehabilitation therapy does not have useful results. With standard rehabilitation, a study showed that there was a difference in what the stroke patients could do in the hospital stroke unit and what the patients did at home. Each activity of daily living was less well performed in the home situation in 25–45% of the cases, and in 52% of the cases the chief caretaker claimed that the patient did not do two or more activities at home that the patient was capable of performing in the day hospital.

Ideally, therapy should be based on experimental findings. Carryover to real-life activities requires programs specifically developed to do so. Issues such as the nature of the interaction between behavioral and neural plasticity and the nature of rehabilitation programs that produce functional carryover should be evaluated. Programs based on conditioned responses have no carryover, whereas those based on shaping and on constraint-induced facilitation have been shown to have excellent carryover to real-life tasks.

A cat hemispherectomy model has provided some of the most reliable information on mechanisms underlying recovery of function in adult cats as well as the functional compensation in the developing animal following neonatal lesion. Forced exercise of the impaired limb was effective in reversing paw preference bias in all cases; however, the adult lesioned cats required more trials and extensive food deprivation. All cats continued improved performance indefinitely in their home cages. Thus, an impressive potential for recovery of directed purposeful movements remains after hemispherectomy, and recovery of function can be enhanced by forced exercise. Although 1 month of recovery time was needed in the adult hemispherectomy cases, recovery time for directed self-feeding movements could be reduced with passive mobilization and treadmill walking and running immediately after the experimental surgery. This provides animal experimental evidence that supports clinical rehabilitation applications of passive movements (which prevent contractures that lead to limited limb range of motion) in the early stages following brain damage.

Another productive, widely used model is the differential effect of serial lesions versus the one-stage lesion of the same amount of tissue, from the same brain area, as the total of all the serial lesions. Many studies have shown that the functional recovery is far greater in the animals that have had serial lesions than in those that have had one-step lesions. Reorganization of function, including reorganization in brain tissue different from the eventually removed tissue, must occur between lesions. This provides firm evidence for brain plasticity and recovery of function, even in the adult animal.

In addition to motor recovery with appropriate rehabilitation, animal models have also demonstrated sensory recovery, such as cat visual recovery from ambliopia with training and the ability to regain all fine sensory functions with rehabilitation following primary somatosensory projection cortex ablations in monkeys trained in tactile discriminations prelesion. An incidental finding is that postlesion it is necessary to inhibit negative cues, and that lesion of the ipsilateral cortex can produce bilateral deficits

There are many excellent recent studies on the scientific basis of recovery of function with rehabilitation. The following are some conclusions drawn by the experimenters of these studies. Rehabilitation must be varied and must not be repetitive. Changes in the motor cortex are driven by the acquisition of new motor skills and not simply by motor use, which may indicate that the repetitive, boring activities of standard rehabilitation are virtually useless. Functional plasticity is accompanied by structural plasticity. Unmasking, multiplexing, synaptic plasticity, sprouting, and inhibition are mechanisms of functional reorganization following brain damage. Functional plasticity in intact cortex is initiated immediately after injury. Activity (rehabilitation) results in an increase in the neuropil, in dendritic arborization, in the number of synapses, and in the separation between neurons, resulting in the reduction of the number of neurons per cubic millimeter. Spontaneous motor recovery cannot be explained by substitution of function in the spared motor cortex immediately adjacent to the lesion. The retention of functional representation in tissue adjacent to the lesion requires motor training (rehabilitation), which appears to have a modulatory effect on plasticity in the surrounding tissue. Blocking the NMDA receptors may be neuroprotective in the early postinjury stage, but blocking them later may reinstate deficits. Immobilizing the good limb too soon after brain damage in a rat model can have extremely deleterious effects, both in behavioral responses and in causing a dramatic expansion of the original lesion, which suggests that either too much or too little activity can have profound negative consequences. In rats, forced disuse for 1 week has many measurable negative effects; thus, studies of the negative effect of forced bed rest in brain-damaged humans are necessary. Mild rehabilitation may improve functional outcome, whereas early moderate rehabilitation can have negative effects, including the exaggeration of infarct size. Human stroke patients had worse outcome with forced speech therapy for several weeks than with brief (15-min) conversations.

Many of these conclusions challenge several dogmas of rehabilitation; however, especially since they are based on experimental findings, they should lead to further studies and, undoubtedly, to changes in the practice of clinical rehabilitation. Clinical studies have rarely used prospective randomized methods. One such study, carried out at the Karolinska Institutet in Huddinge (Stockholm), revealed that for patients for whom early stroke rehabilitation was possible, home rehabilitation was effective. However, well-documented individual case studies demonstrating unusual recovery are valuable, especially since they indicate what is possible and they may suggest effective rehabilitation approaches. In one such study, an elderly stroke patient with extensive brain damage (that was documented on autopsy 7 years later) followed a home rehabilitation program and recovered an extrodinary degree of function and returned to full-time work.

Little emphasis has been placed on late rehabilitation programs, possibly because late recovery has not generally been expected. However, reports of late recovery are not new; it was discussed in an article in the *Journal of the American Medical Association* in 1915, and many such reports have emerged from the treatment of war injuries. A case report of late recovery in a quadriplegic patient noted that since most quadriplegic patients are discharged 4 or 5 months postinjury, many patients have not achieved full motor recovery at discharge. It is possible that in many other cases of central nervous system damage the late recovery of function goes unnoticed since the patients have been discharged.

In addition to neural factors, learned nonuse has been demonstrated in human stroke patients. Although both lower extremities are almost always used as soon as possible in the recovery phase following the stroke since both are necessary for gait, the affected upper extremity is often not used, possibly because many tasks can be performed with one hand, leading to the development of learned nonuse. Behavioral training provided even years after the lesion that could involve as few as 3 days of restraint of the normal limb, thus forcing the use of the affected limb, can reverse the learned nonuse, converting a useless limb into a limb capable of extensive movement. It had previously been shown that forced use of the paretic upper extremity of monkeys with experimentally produced hemiplegia (unilateral cortical area 4 ablation) produced significant recovery of function.

A sensory substitution model of late brain rehabilitation has been developed based on the consideration

that a major sensory loss, such as blindness, removes a large part of the input to the brain and, similar to an actual lesion, induces a major reorganization of the brain. Tactile vision substitution systems (TVSS's) have been developed to deliver visual information from a TV camera to arrays of stimulators in contact with the skin of one of several parts of the body, including the abdomen, back, thigh, forehead, fingertip, and tongue. Mediated by the tactile receptors, images transduced from the camera are encoded as neural pulse trains. In this manner, the brain is able to recreate "visual" images that originate in a TV camera. Indeed, after sufficient training with the TVSS, subjects who were blind since early infancy reported experiencing the images in space instead of on the skin. They learned to make perceptual judgments using visual means of analysis, such as perspective, parallax, looming, and zooming, and to make depth judgments. They have been able to perform complex perception and "eye"–hand coordination tasks, including facial recognition, accurate judgment of speed and direction of a rolling ball with more than 95% accuracy in batting the ball as it rolls over a table edge, and complex inspection–assembly tasks. The results have been interpreted as demonstrating the capacity of the brain to reorganize even when the training (rehabilitation) of congenitally blind persons is initiated in adulthood.

Another post-acute program was developed for persons with long-standing facial paralysis due to facial nerve damage during the removal of an acoustic neuroma who had undergone a VII–XII cranial nerve anastomosis (connecting part of the tongue nerve to innervate the facial muscles). In this model, it is clear that the facial muscles are innervated by nerve fibers from structures genetically programmed to move tongue muscles. However, with appropriate rehabilitation, persons recover spontaneous and voluntary bilateral facial symmetrical movements, and they learn to inhibit dyskinetic movements, even many years after the causitive event. The study was designed to evaluate brain plasticity in a human model in which the extent of the lesion is definitely known and in which, due to the complete loss of connectivity from the brain regions genetically programmed to control facial movements, another system (in this case, the brain regions that had previously controlled tongue movements) could be demonstrated to have reorganized to obtain the functional recovery.

Motivating therapy is effective. An example is the ingenious approach taken by a research group in France to obtain eye movement control in children with cerebral palsy who had eye coordination deficits. They noted, as had others before them, that watching a pendulum aided in the training, but they found that the children refused to watch because they found it too boring. They developed a fascinating functional pendulum by projecting children's movies (*Snow White* and *Lassie*) at a galvanometer–controlled mirror, which reflected the image to the back side of a projection screen. The children sat in front of the screen with their heads fixed so that to follow the pendular movements of the image they had to use eye movements. They underwent 6 hr a week of intense therapy (three movies) and within 1 month improved to the point that they could learn to read.

A comparable approach was taken in the early 1970s with the early electronic pong games, which could be connected to home TV sets. One of the joystick controls was replaced with a device used in the clinic for hemiparetic persons to train arm movements. Instead of meaningless exercise, the arm could control a paddle (paddle size and ball speed were varied according to the capabilities of individual patients) allowing participation in a highly motivating game.

III. IMPLANTED AND ATTACHED INSTRUMENTATION

Miniaturized electronics and device technology (e.g., nanotechnology) offer the promise of overcoming some of the deficits produced by brain damage. Sensors and stimulators have been implanted in the brains of blind and paralyzed persons to interface with computers. An array of electrotactile stimulators built into a false palate (similar to an orthodontic retainer) may allow the tongue to act as a human–machine interface for information from many proposed artificial receptors, such as a TV camera for blind persons, a pitch-and-roll sensor for persons who have lost vestibular function, or a robotic hand with position and touch sensors for paralyzed brain and spinal cord-injured persons. These are examples of the application of emerging technologies to compensate for functional loses.

IV. CONCLUSION

The brain is capable of reorganization during the organism's entire lifetime. Functional improvement after brain damage depends on many biological and psychosocial factors.

See Also the Following Articles

AGING BRAIN • BRAIN ANATOMY AND NETWORKS • BRAIN DISEASE, ORGANIC • BRAIN LESIONS • COGNITIVE REHABILITATION • MODELING BRAIN INJURY/TRAUMA • NEUROPLASTICITY, DEVELOPMENTAL • STROKE

Suggested Reading

Bach-y-Rita, P. (Ed.). (1980). *Recovery of Function: Theoretical Considerations for Brain Injury Rehabilitation*. Hans Huber, Bern.

Bach-y-Rita, P. (1995). *Nonsynaptic Diffusion Neurotransmission and Late Brain Reorganization*. Demos-Vermande, New York.

Bach-y-Rita, P. (1999). Theoretical aspects of sensory substitution and of neurotransmitter-related reorganization in spinal cord injury. *Spinal Cord* **37,** 465–474.

Levin, H., and Grafman, J. (Eds.) (2000). *Neuroplasticity and Reorganization of Function after Brain Injury*. Oxford Univ. Press, New York.

Taub, E., and Crago, J. E. (1995). Behavioral plasticity following central nervous system damage in monkeys and man. In *Maturational Windows and Adult Cortical Plasticity* (B. Julesz, and I. Kovacs, Eds.). Addison-Wesley, Redwood City, CA.

Brain Development

HARUN K. M. YUSUF and KHALEDA ISLAM
University of Dhaka, Bangladesh

GLOSSARY

action potential A sudden overshoot of membrane potential of excitable cells caused by a stimulus that is chemical, electrical, or mechanical in nature.

development Increase in weight of the organ due to an increase in size of the cells through synthesis and deposition of cell constituents such as protein, lipid, and RNA (hypertrophy).

gliogenesis Emergence of glial cells from a specific section of the neuroepithelium from where no other cells (neurons) originate.

gray matter Outer layer of cerebrum (cerebral cortex) or cerebellum (cerebellar cortex) where the majority of the neurons (about 75%) reside. The cerebral cortex is thought to be the center of all higher mental functions.

growth Increase in weight of an organ due to an increase in the number of cells through cell multiplication (hyperplasia).

growth spurt The period when the growth of a tissue is at its peak. This is the period when the tissue is most vulnerable to internal as well as external inhibiting agents.

neurogenesis Birth of neurons from the stem cells (neuroblasts) of the neuroepithelium, with each neuroblast generating a specific family of neurons.

neurotransmitter Specialized substances that mediate the transmission of nerve impulse from the presynaptic to the postsynaptic membrane and produce a postsynaptic effect (excitatory or inhibitory) by interacting with specific receptors.

The brain is the most complex of all biological tissues in nature. The complexity lies in the fact that the brain is an extremely heterogeneous tissue, composed of many constituent parts. Each part not only has its own structural organization, cellular makeup, chemical composition, and functional activity but also has a unique developmental characteristic. However, the different parts develop and maintain among themselves a highly organized coordinated system that contributes ultimately to the functioning of the brain seemingly as a single organ. It is estimated that the human central nervous system contains 11–12 billion nerve cells, each of which is capable of making up to 10,000 synaptic interconnections. According to one estimate, the number of intercellular interactions that could possibly occur within a single human brain may be greater than the total number of particles in the whole universe. It is the development of a network of this vast number of interactions that forms the basis of the functioning of the brain. Understanding the complex developmental processes of the brain and brain functions has indeed been one of the greatest challenges for man. This article explores and unveils some aspects of the complex processes of brain development in man and other animals.

I. BASIC STRUCTURE OF THE BRAIN

The brain has undergone a long process of evolution, taking millions of years for it to take the shape and organization currently found in man. During this evolution, the brain has increased not only in size but also in its level of organization, sophistication, and

complexity. The increased size of the brain, however, is not necessarily an indication of increased intelligence. Larger animals, such as the elephant or the whale, have much larger brains than that of man. Also, the basic chemical and metabolic setup of the brain of even the lower mammals and that of man are, in a broader perspective, difficult to distinguish. However, the human brain is very special in relation to culture, consciousness, language, memory, and especially intelligence. It distinguishes itself from even the most highly developed brains of other animals.

The morphology and cytology of the brain tissue are now well understood. A large variety of cell types occur in the brain that can be classified into two major groups according to their general morphology and function: the neuronal cells and the glial cells.

A. Neurons

Neurons are the excitable nerve cells of the brain. They are involved in the initiation and conduction of nerve impulses. Because of this unique property, neurons are considered the functional, computing elements of the nervous system. Two kinds of processes are formed from extensions of the neuronal plasma membrane—the axons and the dendrites (Fig. 1).

In vertebrates, the function of the axon is to conduct the nerve impulse in one direction only, usually away from the neuronal cell body (efferent) to another neuron or another part of the system through intercellular junctions called the synapses. On the other hand, dendrites carry information from the synapse to the cell body (afferent).

B. Glia

Unlike neurons, glial cells are not excitable and do not have any signaling apparatus (axons or synapses). Therefore, these cells do not participate in initiating the nerve impulse or in mediating activities of the nervous system. Their main function is to form myelin (which facilitates nerve conduction) and to support growth and differentiation of neurons. Glial cells, in relation to neurons, become increasingly numerous along the phylogenetic scale, and in mammals they constitute almost half the volume of the brain and greatly outnumber neurons by a ratio of approximately 10 to 1. Thus, as many as 100 billion glial cells are found in the human central nervous system (CNS).

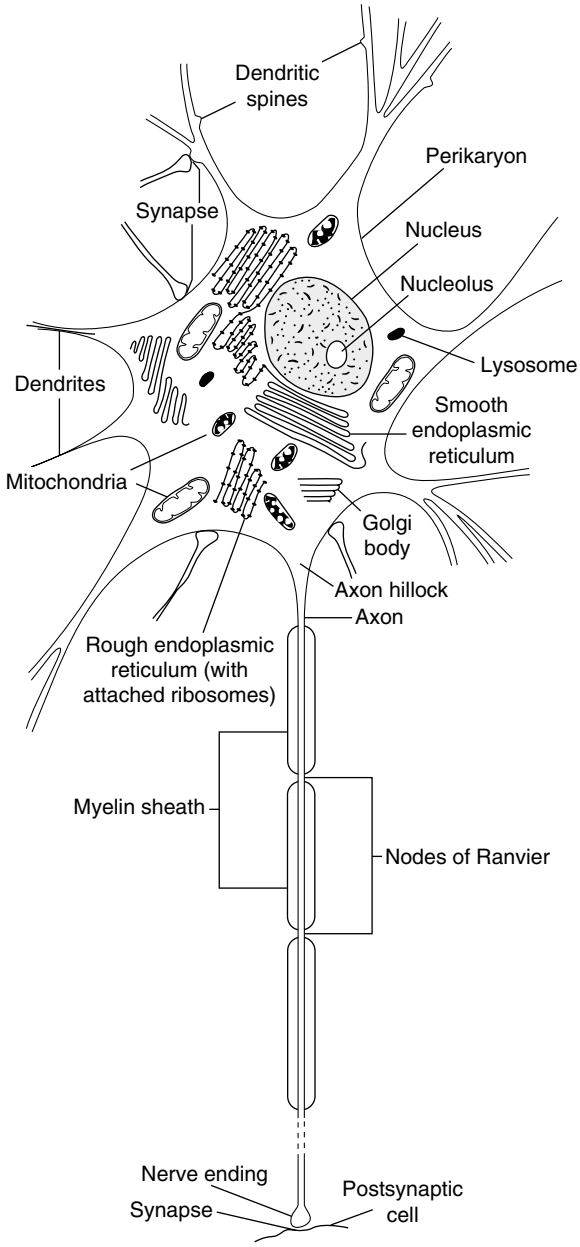


Figure 1 Schematic drawing of a typical mammalian neuron (reproduced with permission of Chapman and Hall).

C. The Synapse

In the nervous systems of both vertebrates and invertebrates, the neurons are in close apposition with each other and they relay information from one to another. The junctions at which the cells are interconnected and through which the impulses are propagated are called synapses, first named by Sherrington in 1906 from the Greek word *synapto* meaning tight

clasping. The membranes that participate in the formation of synapses are those of the axons, dendrites, and the main cell body.

D. Basic Principle of Nerve Action

The basic principle of all nerve impulses and nerve functions is the production of an action potential across neuronal membrane. The action potential is a sudden and transient overshoot of membrane potential of neuronal membrane from the resting value of −60 to −70 to 20 mV or more, caused by an electrical, chemical, or mechanical agent. The action potential is conducted along the axon and then transmitted to the next neuron and finally to the target tissue. The response of the target tissue is specific to the chemical substances that mediate this transmission (the neurotransmitters).

II. HYPERPLASTIC GROWTH OF THE BRAIN

The weight gain of a given organ with time is a measure of growth of that organ. Weight gain is indeed the simplest and the easiest criterion of growth, provided that the presence of an unusual amount of water in edematous condition does not affect the actual situation. However, the weight curve of tissues including the brain, like that of the body, is not linear with time but follows a sigmoidal pattern (Fig. 2), meaning that there is a period of more rapid growth compared with that occurring before or after that period. This period of rapid growth is known as the growth spurt and is a determinant for the so-called "critical" or "vulnerable" period because at this time even a slight disturbance in the growth-supporting environment may have a profound effect on the overall growth and development of the tissue.

The anatomical heterogeneity of the brain makes the developmental study of the tissue particularly difficult because it is almost impossible to dissect its hundreds of tiny functioning parts and study them. However, valuable information has been obtained from the study of the major regions of the brain—the forebrain, cerebellum, and brain stem (Fig. 2)—although such studies are of limited value with regard to the growth spurts of the composite structures in each of these regions.

Generally, the growth process of the brain, like that of any other organ, encompasses two major processes—*growth and development*. The growth process consists of an increase in the number of cells until the

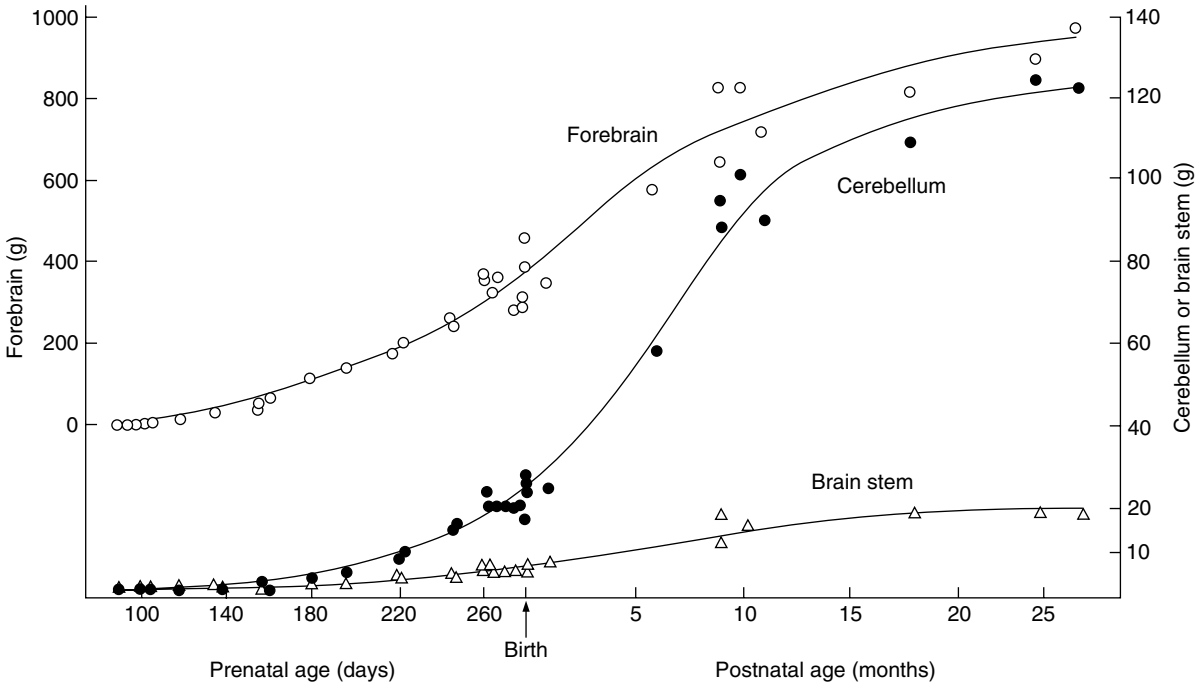


Figure 2 Growth curves of human forebrain, cerebellum, and brain stem.

adult cellular population is achieved in the tissue. This process of growth by cell multiplication is known as *hyperplasia* and is determined by the genetic makeup of the individual organs. The process of development, on the other hand, is one in which the cells obtained through hyperplasia are developed into mature functioning units as a result of deposition of different cellular constituents (e.g., protein and lipid). This process is termed *hypertrophy*. Therefore, mature cells differ from immature ones in both size and chemical composition. For the brain, the major events in hypertrophy are myelination and synaptogenesis.

Brain cells are mostly diploid, and the amount of DNA in all diploid cells of a given species is constant, with the value being slightly different in different species. For example, the value is 6.2 pg for the rat diploid cells and 6.0 pg in humans. Therefore, the total number of cells in a brain can be determined by dividing its total content of DNA by the value per cell.

Once the number of cells of an organ is known, the weight per cell and the protein content, RNA content, or lipid content per cell can be determined from the total weight of the tissue and the total content of protein, RNA, and so on. These values are used as an approximate measure of the size of the cells of that tissue. An increase in these values indicates an increase in cell size (i.e., hypertrophy). However, the general problem associated with this assumption for the brain tissue lies in the different cell types: One cell type is likely to have an amount of protein or lipid very different from that of another cell type.

Within these limitations, the neurochemical approach has made an important contribution toward our understanding of the general pattern of growth and development of tissues including the brain. Based on these principles, Myron Winick proposed that organs grow and develop in three consecutive phases (Fig. 3). During the first phase, growth proceeds entirely by cell multiplication (hyperplasia alone), with a proportional increase in weight and in protein, RNA, or lipid content, so that the cell size, as measured by weight/DNA or protein/DNA, remains constant. At the end of this phase, the rate of DNA accretion gradually slows, but the accumulation of other constituents continues. This constitutes the second phase of mixed hyperplasia and hypertrophy, in which there is an increase in cell size and a smaller increase in cell number. In the third phase, DNA growth stops altogether and the cells develop by increasing in size by continued synthesis and accumulation of proteins and lipids. This is the phase of hypertrophy. Finally, when net protein or lipid synthesis, and therefore the weight of the cells, is established, the tissue attains the state of maturity.

The scheme proposed for the brain by Davison and Dobbing in terms of growth and development of

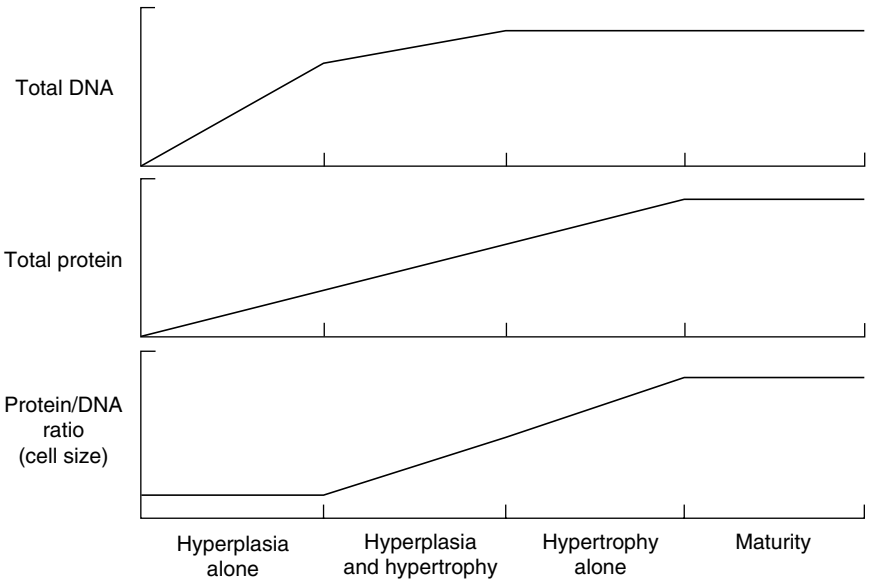


Figure 3 Relationship between DNA, protein, and protein/DNA ratio during the three phases of tissue growth (reproduced with permission of Oxford University Press).

different types of cells and their associated structures fits well with the general scheme of Winick:

Stage I: Organogenesis and neuronal multiplication (hyperplasia alone?)
Stage II: The brain growth spurt, including
II(a). A maturation period of axonal and dendritic growth, glial multiplication, and myelination (hyperplasia plus hypertrophy)
II(b). a later period of growth in size (hypertrophy alone)
Stage III: The mature, adult state (maturity)

According to this scheme, organogenesis, followed by neuronal multiplication, takes place in the first stage. The brain growth spurt occurs in the second stage. It includes maturation of the axons and dendrites, multiplication of cells, and myelination. The third stage consists almost entirely of continued myelination and increase in tissue size, at the end of which the brain is thought to attain structural maturity.

A. Species Differences

The developmental stages outlined previously for the brain proceed in a sequential manner in all species studied. Regarding the physical development of the mammalian brain, one major difference between species is the timing of birth in relation to the time of brain growth spurt. In some animals, brain growth spurt occurs prenatally (guinea pig, sheep, and monkey), in some it occurs perinatally (human and pig), and in others it occurs postnatally (rat and mouse) (Fig. 4). Accordingly, these animals are called prenatal, perinatal, and postnatal brain developers, respectively. Thus, the event of birth in mammals, although representing a major physiological milestone in the development of the respiratory and cardiovascular structures and functions, is apparently of little significance to the development of the brain and behavior.

B. Structural Differences

Studies of the gross anatomical regions of the brain have revealed an extraordinary rapid rate of growth of the cerebellum compared to that of the rest of the brain in all mammalian species examined (Fig. 2). In the race of growth and development, the cerebellum starts slightly later but finishes earlier. This makes this part of the brain especially vulnerable to internal or external influencing agents.

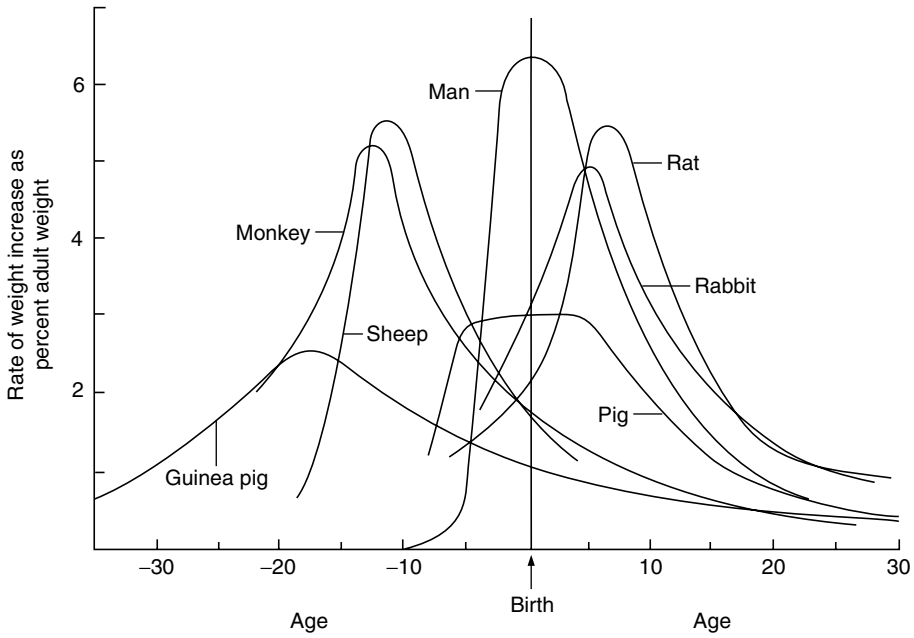


Figure 4 The timing of brain growth spurts in various mammalian species in relation to the timing of birth. The units of age for the different species are as follows: guinea pig, days; sheep, × 5 days; monkey, × 4 days; man, months; rat, days; rabbit, × 2 days; pig, weeks (reproduced with permission from authors and Elsevier Science).

C. Gray and White Matter of the Developing Brain

Marked changes occur in the relative proportions of the gray and white matter of the brain during development. During early developmental stages, the proportion of gray matter increases, and then, when myelination begins, the proportion of the white matter increases. For example, biochemical studies showed that at 2 weeks of age (i.e., at the time when active myelination has just begun), the rat brain is composed mainly of gray matter (85%). During the following 4 weeks of growth, the gray matter proportion decreases rapidly to approximately 70%, whereas the proportion of the white matter increases from the initial value of 15% to approximately 30%. Later, this change in percentage proportions is gradual, and at 52 weeks of age the two matters share the total mass of the brain almost equally (Fig. 5).

However, when the absolute amounts of gray and white matter in the developing brain are considered, the amount of gray matter is found to increase (by approximately 12%) only up to 3 weeks of age and remain almost constant thereafter (up to the age of 9 weeks), whereas the absolute amount of white matter increases steadily throughout this growth period (Fig. 5).

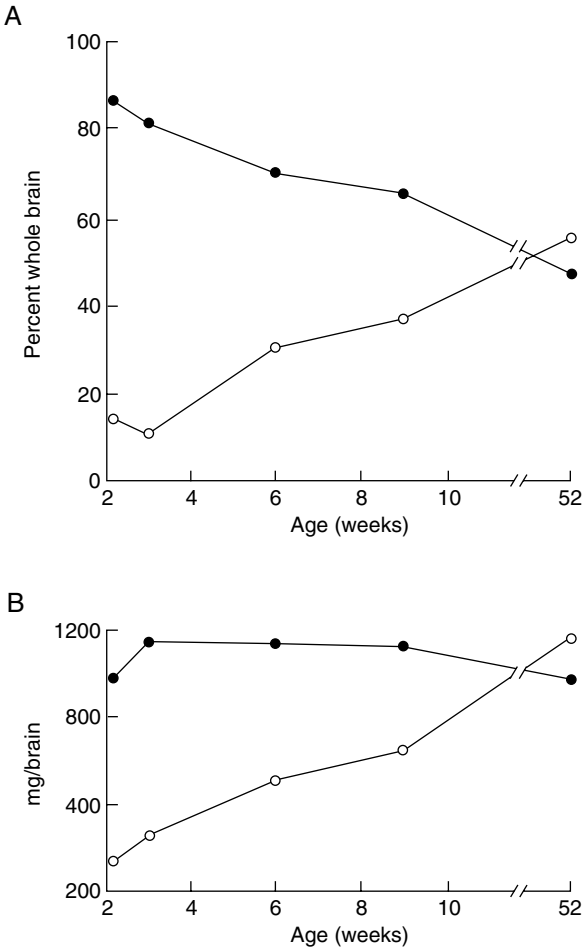


Figure 5 Changes in the (A) proportions and (B) total amounts of gray (●) and white (○) matter of the developing rat brain (reproduced with permission of International Society for Developmental Neuroscience).

D. Growth of Different Cell Types

Not only do the different parts of the brain grow at different times and at different rates but also there are differences in the timing of multiplication of different cell types within a particular brain region. It has been known for a long time that multiplications of the neurons and of the glia are two consecutive processes, the former being followed by the latter.

While studying the cellular growth of the human brain, Dobbing and Sands observed two distinct peaks of DNA accumulation in the forebrain, the first occurring at 18 weeks of gestation and the second at approximately birth (Fig. 6). The first peak was interpreted as corresponding to the peak of neuronal multiplication and the second to that of glial multiplication. However, it is likely that the adult neuronal population is not achieved at 18 weeks of fetal life; some neurons continue to divide beyond this time, but their number is not significant compared to the number achieved by 18 weeks of gestation. Man is thus a fortuitous species in that his brain neuronal population is established as early as mid-fetal life and thus saved from the deleterious effects of even severe nutritional insufficiency in the mother.

E. Neurogenesis

Neurons originate from the stem cells (neuroblasts) of the neuroepithelium. Understanding of the process of neurogenesis was enhanced by studies on relatively simple nervous systems, such as those of the nematode, leech, and insects. During embryonic development of the CNS, an enormous diversity of cellular types are arranged and interconnected in a remarkably precise pattern. For example, within the neuroepithelium in each segment of the grasshopper embryo, 61 neuronal precursor cells (neuroblasts) are arranged in two

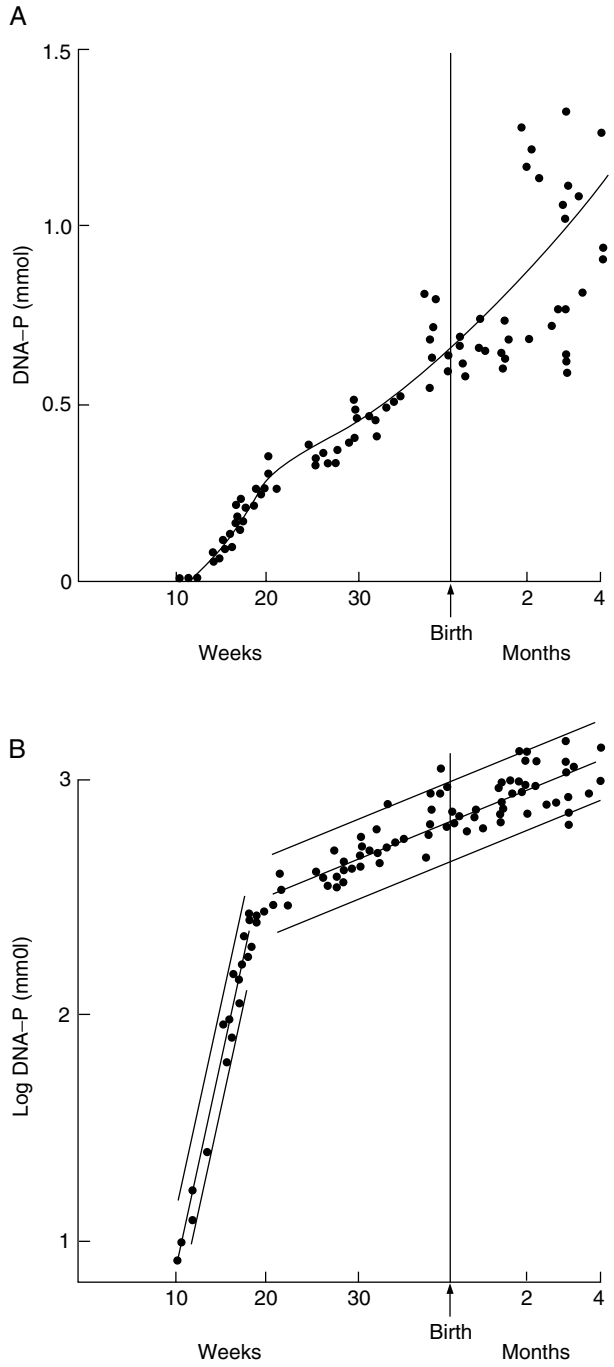


Figure 6 (A) Changes with age in total DNA–phosphate, equivalent to total cell number in the human fetal and infant forebrain. (B) A semilogarithmic plot of the same data as in (A). Regression lines with 95% confidence are drawn in B (reproduced with permission of British Medical Journal).

symmetric plates of 30 neuroblasts each and one median neuroblast (Fig. 7). Each neuroblast can be identified by its position within the neuroepithelium and also from the highly stereotyped family of neurons it produces.

As a neuroblast appears in the neuroepithelium, it divides repeatedly to generate a chain of ganglion mother cells. Each ganglion mother cell then divides once more to produce two ganglion cells in a chain of cell doublets, which then differentiate into a family of neurons. The families of neurons originating from different precursor neuroblast cells differ by their unique morphology, physiology, biochemistry, and function.

Thus, each neuroblast appears to generate a specific family of neurons, and a neuroblast's position in the epithelium is so specific that no other cell, including the ectodermal cell adjacent to it, can take its place, nor can the progeny of one neuroblast be replaced by the progeny of the neighboring neuroblasts. The final differentiation of the individual neurons depends on their mitotic ancestry and their later cell-specific interactions with other neurons in the environment.

F. Gliogenesis

In the vertebrate CNS, the neuroepithelium gives rise to most of the neuronal as well as the macroglial cells found in mature tissue. The mechanism by which the neuroepithelial cells first differentiate and follow the different cell lineages is largely unknown, nor is it clear when the decision for differentiation is made.

As with neurogenesis, most of the current understanding of the process of gliogenesis derives from the study of a relatively simple nervous system, such as the optic nerve. The optic nerve develops from the optic stalk, which is an extension of the original neuroepithelium (Fig. 8). The neuroepithelial cells that form this optic stalk give rise only to glial cells (astrocytes and oligodendroglia) and not to neurons. Neurobiologists have used this characteristic of the cell lineage systems of the optic stalk to study the pattern of gliogenesis.

G. Neuronal Death and Nervous System Development

Neuronal death is a major phenomenon in the normal development of the vertebrate as well as the

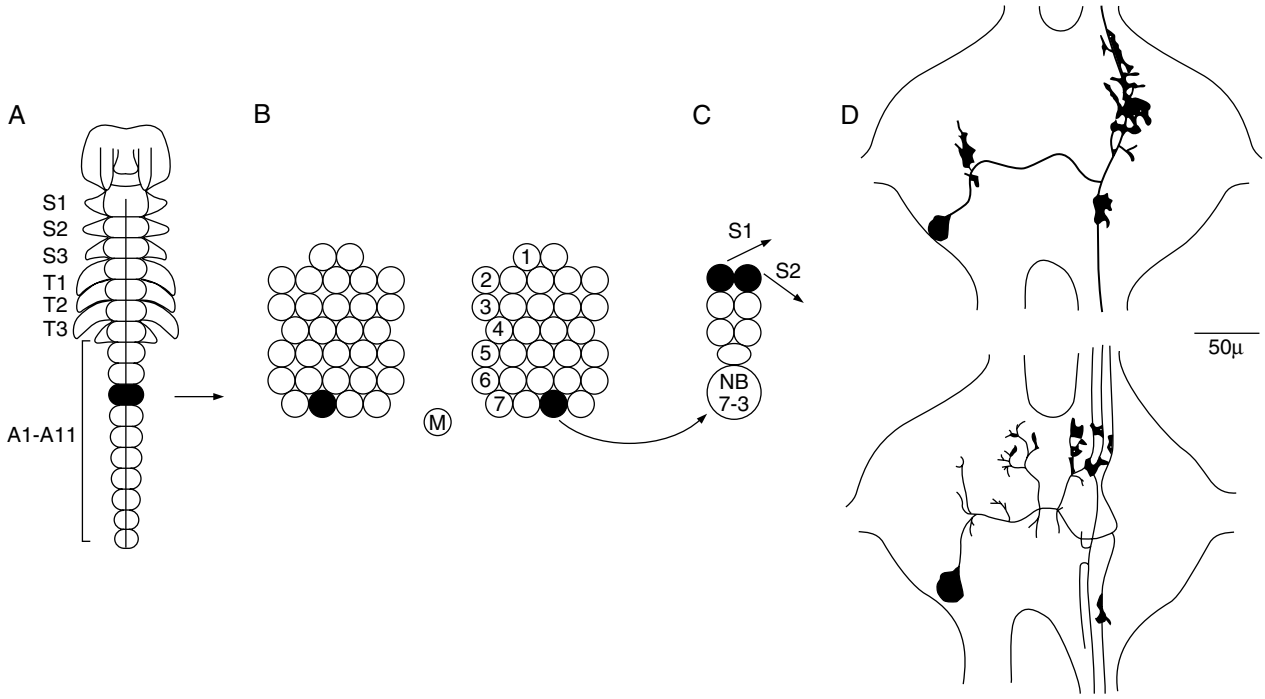


Figure 7 Neurogenesis in the grasshopper central nervous system (CNS). (A) Schematic diagram of a 30% grasshopper embryo (hatching occurs at 100%) showing the segmental structure of the ectoderm. The middle of the ectoderm is a longitudinal strip of neuroepithelium that gives rise to the CNS. The location of the neuronal precursor cells for a single segment (A4) are drawn in black. (B) Pattern of neuroblastomas (NBs) for a single segment (A4) includes two plates of 30 NBs arranged in a precise pattern of seven rows and one median neuroblastoma (MNB), for a total of 61 NBs. (C) Cell lineage of two identified serotonin immunoreactive neurones, S_1 and S_2, from NBs 7–3. NBs generate neuronal progeny by a series of asymmetric cell divisions that produce small, ganglion mother cells. Ganglion mothers cells then each divide once more symmetrically to produce two cells that will both differentiate into neurons. The cell lineage is represented with a ganglion mother cell closest to the NB and older neuronal progeny lying progressively farther away from the NB. (D) The cell-specific morphologies of neurons S_1 (top) and S_2 (bottom) in the A4 ganglion at 70% embryonic development, as revealed by intracellular injection of Lucifer Yellow followed by HRP immunocytochemistry with an anti-Lucifer Yellow antibody (reproduced with permission from Macmillan Journals Ltd).

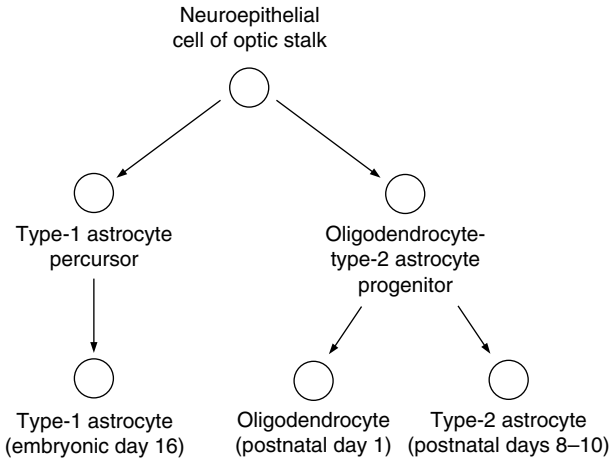


Figure 8 Postulated glial cell lineages in rat optic nerve.

invertebrate nervous systems. The first systematic report in this context was that of Hamburger and Levi-Montalcini, who observed massive cell losses by death in dorsal root ganglia of the growing chick at the upper cervical and thoracic levels. Subsequent studies showed similar neuronal degeneration in other nerve structures to the extent of 25–75% occurring during a well-defined period of days or weeks. For example, 30% of the cells of mouse cerebral cortex or lateral geniculate nucleus are lost between Postnatal Days 5 and 30, and chick trigeminal mesencephalic nucleus loses as much as 75% of the cells between Embryonic Days 11 and 13.

There are many reasons for the natural death of the neurons in the developing nervous system, but the best understood and most documented is the lack of maintenance from the target tissue. The target cells are known to secrete one or more trophic factors, the

so-called neurotrophic factors, of which the nerve growth factor is the best characterized. These trophic factors are essential for the survival of the neurons and are made available to the neurons by retrogradic transport.

Many observations have clearly indicated the phenomenon of "competition" among neurons to gain access to the target tissue and obtain the necessary trophic factors. More active neurons have been shown to win and survive, whereas less active ones make fewer synapses, receive less trophic factors, and eventually degenerate. This is in agreement with the suggestion made more than 100 years ago by Roux that cells, including neurons, follow the Darwinian "struggle for existence—survival of the fittest" theory.

During early periods of development, many diffuse, inaccurate, and aberrant synaptic interconnections are formed in the nervous tissue that have to be removed for a precise and topographically ordered synaptic network characteristic of the nerve tissues to be established. Death of the neurons with such aberrant connections would be a way to achieve this. However, cell death is not the only means by which neuronal connections are eliminated. In many cases, this is achieved by pruning the axons of aberrant neurons without the death of the cells.

H. Regeneration of Brain Cells

One important difference between the nervous systems and most other tissues of the body is that in the nervous tissues, neuronal cell division has "once-and-for-all" characteristics. Once adulthood is reached, neurons stop multiplying in any appreciable number. However, studies of the incorporation of labeled thymidine into DNA show that the glial cells divide even in the normal, mature brain but with negligible rate of turnover. These rates are found to increase in response to traumatic injury of the brain.

III. HYPERTROPHIC DEVELOPMENT OF THE BRAIN

The cellular growth of the brain described in the preceding section is followed by *hypertrophic* development. During this phase, cells are developed by a process whereby the size of the main cell body and the length and size of the cell's different processes increase. This increase in size is accomplished by synthesis and quantitative deposition of proteins and lipids in various cellular membranes, particularly the deposition of lipids around the axons in the form of myelin.

A. Brain Lipids during Development

The brain is the tissue of the body that is most abundant in lipids. With the exception of a small proportion of lipids concerned with cellular dynamic processes such as signal transduction, the bulk of the lipid material is located in structural components of various membranes, especially the myelin. Therefore, the rate of lipid accretion in the brain during development governs the rate of synthesis of the whole component structure of the membranes. From the knowledge of membrane specificity of different lipid substances, it is possible to understand the developmental profile of a particular membranous structure by monitoring the developmental profile of the lipid specific to that membrane. For example, cholesterol is widely used as a marker of myelination; however, cerebrosides and sulfatides are more specific myelin lipids. Gangliosides are generally accepted as a marker of dendritic arborizatin and synaptic development.

The lipid content and composition of the brain at early stages of growth are very similar to those of other tissues. However, due to a rapid synthesis of lipids in the brain during development, lipid content as well as composition of the mature nervous tissue become markedly different from those of tissues in the rest of the body. The brain also differs from other tissues in that it is rich in more complex and polar lipids. On the other hand, the adult brain contains negligible quantities of esterified cholesterol and almost none of the triglycerides that are abundant in almost all other tissues.

It should be noted that the sequence of lipid deposition in the brain at various stages of development does not differ substantially between species. However, within the brain, not all lipids are synthesized and deposited at the same rate. For example, myelination, as evidenced by deposition of cholesterol, proceeds at different times and at different rates in different tracts even within the same region of the brain.

In mammals, the peak rate of cholesterol synthesis, and therefore myelin formation and deposition, occurs sometime after the peak period of glial cell multiplication. Thus, the peak rate of myelin accumulation occurs at approximately 15–18 days of age in the rat brain, at the time of birth in the guinea pig brain, a few

days after birth in the pig brain, and during the first 8 or 9 months after birth in the human brain (Fig. 4). Moreover, the cells that multiply just before the peak cholesterol synthesis are almost all oligodendroglia, the cells that form myelin around the central neuronal axons. The same sequence of cell (oligodendroglial) multiplication and peak cholesterol accumulation is also seen in the brain of nonmammalian species such as the catfish.

Gangliosides, a group of sialoglycosphingolipids, are present in the plasma membrane of all animal cells. Their hydrophobic ceramide portion is embedded in the lipid matrix of the membrane, whereas their hypdrophilic sialoglycosyl chain is oriented toward the extracellular medium. These sugar chains, which are different from one kind of cell to another, are believed to contribute to the cell adhesion and cell–cell recognition properties of living cells.

Gangliosides are most concentrated in the brain compared to the rest of the body. Within the brain, the concentration in the gray matter is several times higher than that in white matter. Within the gray matter, these lipids are most concentrated in the axon terminals and dendrites, the structures that constitute most of the synaptic interconnections. As mentioned previously, gangliosides are thus considered label lipids for synaptogenesis. Based on ganglioside measurement, it has been shown that the period of most rapid synaptogenesis in the human brain is the first 8–12 months after birth.

B. Lipids of Gray and White Matter

The general view of lipid distribution in the brain is that lipids such as cholesterol, cerebrosides, and sulfatides are more concentrated in the white matter than in the gray matter, whereas the gangliosides are gray matter-specific constituents. The phospholipids as a whole are specific neither for the gray nor for the white matter, although ethanolamine phosphoglycerides and sphingomyelin are found more in the white than in the gray matter.

C. Myelin Composition in Developing Brain

Myelin accounts for more than 25% of the weight of the mature brain. Myelination is thus a major event in brain development. Myelin is composed mainly of lipids, proteins, and water, with small amounts of inorganic salts. About 40% of the wet weight of myelin is water. Of the material, 75–80% is lipids, and most of the remainder is proteins. The most extensive studies on myelin composition have been confined to lipids, although attention is now being given to proteins.

1. Myelin Lipids and Proteins

The most prominent changes that take place in the lipid composition of the CNS myelin during development are in relative mole proportions of cholesterol, phospholipids, and galactolipids, which attain a final value of approximately 2:2:1 as opposed to about 8:10:1 in the whole brain.

The content of protein in myelin is much less than that usually found in typical cell-surface membranes. The protein composition of myelin is unique. Most typical membranes contain many different kinds of proteins, none of which supercede the others in quantitative terms. On the other hand, myelin, not only contains less protein than other membranes but also has fewer (four or five) types of proteins. These four or five proteins have been separated by polyacrylamide gel electrophoresis. Of these, three are found to be highly specific for myelin: myelin basic protein (so named because of its relative richness in basic amino acids), proteolipid protein (so named because it is soluble in neutral organic solvents), and the so-called intermediate protein (lies intermediate between basic and proteolipid protein in an electric field). The relative proportions of these and other minor proteins change during myelin maturation.

2. Lipid–Protein Associations in Developing Myelin

At the beginning of myelination, as the flattened surface membrane of the myelin-forming glial cell wraps the axon for the first time, a double-membrane structure, the mesaxon, is produced. This double-membrane mesaxon, in the course of its differentiation into myelin, seems to serve as a "template" (protomyelin lattice) for the accumulation of myelin-specific lipids and proteins leading to the assembly of the myelin structure. The synthesis and deposition of the basic proteins and the lipids is thought to represent an early step in the differentiation process, which is advanced further by the accumulation of large amounts of more lipids and the proteolipid protein. Myelin basic proteins are synthesized on free polysomes in glial processes, whereas the proteolipid protein is synthesized on bound polysomes

(endoplasmic reticulum) in the glial cell body. The basic proteins are more rapidly incorporated into the developing CNS myelin sheath (within 2 min after synthesis), whereas the proteolipid protein appears in the myelin sheath after about 30 min. This assembly of myelin (incorporation of finished products into myelin sheath) is thought to be regulated by axonal signals.

D. Synaptic Development

1. Increases in Dendritic Arborization and Axon Density

One major process in brain development is the establishment of synaptic interconnection between individual neurons or between a neuron and its target cell. After the neuroblasts have divided and been transformed into neurons, the neurons, in order to serve as functional units of the nervous tissue, develop by increasing in size. This increase in cell size manifests itself in an increase in the size of the main cell body as well as in the size and length of the cellular processes—the axons and the dendrites. This in turn results in increases in dendritic arborizaion and axon density. The axons continue to grow in length until they find and form synapses with their target membranes. A target membrane may be the cell soma, dendrite, or the axon of another neuron (the axosomatic, axodendritic, and axoaxonal synapses) within the CNS, or it may be the plasma membrane of an excitable cell such as muscle cell (neuromuscular junction) in the peripheral nervous system. Thus, an increase in axon density is immediately followed by an increase in synaptic number (Fig. 9).

Synaptic activity begins when the presynapic neuron acquires the ability to secrete its neurotransmitter upon stimulation and the neurotransmitter thus released produces an effect, excitatory or inhibitory, on the postsynaptic membrane. The brain and other nervous regions become increasingly active when the number of synapses and their activities increase during development. Concomitant with this increase in synaptic number and synaptic activity is an increase in the content of the neurotransmitters and in the activities of the enzymes involved in their metabolism. In the simple nervous system of the leech, it has been shown that by the time of body closure, the neurons show characteristic action potentials and are therefore connected to functioning synapses. At this time, the synthetic activity for acetylcholine and other neurotransmitters is achieved, but at very low levels. The synthetic activity then increases enormously, 25 times the initial value, when the axons reach the connective and segmental nerves.

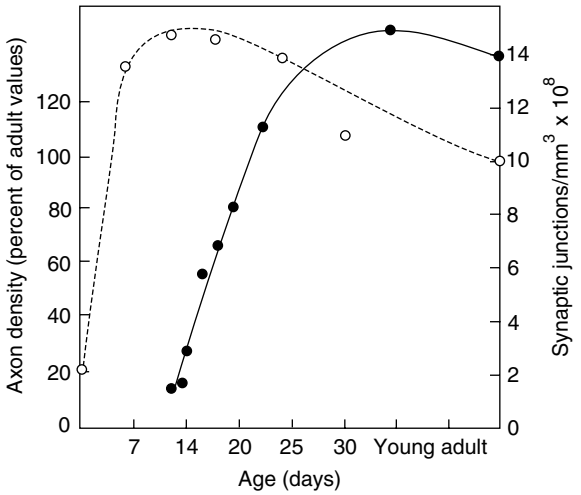


Figure 9 Increases in axon density and in synaptic number in layer 1 of the developing rat cerebral cortex. ●, synaptic number, parietal cortex; ○, axon density, sensorimotor cortex (reproduced with permission of Elsevier Science and Cambridge University Press).

In the vertebrate CNS, the monoaminergic neurons develop very early during ontogeny and are among the first to differentiate, synthesize, store, and release their transmitter substances (dopamine, noradrenaline, or serotonin) long before the maturation of the brain region they innervate. These neurons are therefore thought to play important roles in neurotrophy and plasticity in the developing nervous system.

2. Mechanism of Synaptogenesis

Most of the structural synapses in the brain are formed within the early stage of life, e.g., during the first month after birth in rats and 1 or 2 years after birth in humans. However, determining the exact mechanism whereby the neurons in the developing nervous system form synapses remains a central challenge in neurobiology. Some neurons, such as those in the brain, do not project their axons very long before finding their appropriate synaptic partners. In other neurons, such as the motoneurons of the spinal cord, the axons traverse long distances to find and form synapses with the desired postsynaptic target membrane, bypassing all "inappropriate" membranes along the migration pathway. The question thus arises how the neurons distinguish appropriate from inappropriate

postsynaptic membranes. Questions also arise regarding how the growth of the axons is regulated, what makes them stop growing when they reach their target cells, and what brings about the changes in the ultrastructural organization of the cytoskeleton that results in metamorphosis of the growth cone (axon tip) into the normal nerve terminal. Much research during the past decade has attempted to answer these and other questions relating to the process of synaptogenesis. Today, we have some insight into the mechanism, but we do not know all the details of the mysterious method whereby neurons establish their synaptic interconnections.

Many suggestions have been put forward to explain neuronal specificity in synaptogenesis. Sperry first proposed a chemoaffinity hypothesis that postulates that the specificity of neuronal connections is conferred by specific interactions between molecular addresses in the presynaptic membrane and the complementary molecules residing in the postsynaptic target membrane. This hypothesis is based on studies of presynaptic retinal neurons and the postsynaptic target neurons in the tectum. Hoffman and Edelman discovered a sialic acid-rich neuronal glycoprotein, termed N-CAM (neural cell adhesion molecule), that mediates intercellular adhesion in the absence of Ca^{2+}. This glycoprotein might be a part of the molecular address (mentioned previously) on neuronal membranes that is recognized by specific receptor sites on target neurons, and it may be important in the development of the nervous system by maintaining the topographical relationships between individual neurons and/or their axons in a set of neurons.

During the formation of a synapse, a complex series of ultrastructural changes occur, notably in the synapsing axon and its tip. The net result is that the axon stops growing in length and its tip (the growth cone) takes the shape of a typical nerve terminal.

3. Orchestration at the Synapse

Another important aspect of synaptogenesis is the molecular and structural reorganization in the postsynaptic membrane upon formation of a synapse with a prospective presynaptic nerve terminal. It is conceivable that the receptor proteins in the postsynaptic membrane must be aligned exactly opposite to the sites in the presynaptic membrane where the transmitters are released. Such alignment is maintained as long as the membrane participates in synaptic activity. Indeed, during early periods of development when synapses have not yet formed, the transmitter receptor proteins are distributed throughout the membrane surface of the postsynaptic cell. Also, during nerve degeneration, the receptor proteins disperse and the whole orchestration at the synapse breaks down. On the other hand, during synapse formation as well as during nerve regeneration, the receptor proteins aggregate themselves again to form clusters at the point of synaptic junctions. Thus, some signal mechanism must exist that regulates the behavior of the postsynaptic membrane during synaptogenesis and synaptic degeneration.

IV. ENERGY METABOLISM OF THE DEVELOPING BRAIN

Much like the changes in lipid and protein compositions described earlier, energy metabolism of the brain also undergoes an interesting shift during development. The most dramatic of these changes are changes in blood flow and oxygen consumption and the utilization of glucose as the source of energy. It is well-known from both *in vitro* and *in vivo* studies that oxygen consumption by the cerebrum remains at a low level at birth, although oxygen supply to the tissue may be high. Investigations show that relative to the amount of oxygen consumed, the amount of oxygen delivered to the cerebrum during fetal life exceeds that in the newborn and adult by as much as 70%. This may protect the fetus from the stress of labor and delivery, or it may simply be an obligatory adaptation to the low arterial oxygen pressure in the intrauterine environment. After birth, both oxygen supply and oxygen consumption increase rapidly and reach maximum levels at the time of peak development. Oxygen consumption then slowly decreases to the adult level at maturity.

The primary fuel that fulfills the high demand by the brain for metabolic energy is glucose. However, utilization of glucose and the efficiency of the process are not uniform throughout the developmental program. Early immature brain is less aerobic than the mature adult brain and oxidative processes (mitochondrial metabolism and respiration) are not fully developed in the immature brain. Many years ago, it was shown that newborn rats can withstand anaerobic conditions for as long as 1 hr, but this resistance to hypoxia disappears when the animals are administered iodoacetate, an inhibitor of the glycolytic pathway. This led to the understanding that early in development (before or at birth in rats or humans),

glycolytic breakdown of glucose to pyruvate or lactate is a major mode of glucose utilization. However, as is known, comparatively little energy is produced in this pathway (only 2 net moles of ATP are synthesized per mole of glucose utilized compared with 36 mol in the oxidative pathway). After birth, as development proceeds, respiration becomes increasingly important, and the oxidative pathway (complete oxidation of glucose to CO_2 and H_2O) plays a dominant role in glucose metabolism.

This shift in the metabolic pattern of the brain is understood from the changes in the activities during postnatal development of the enzymes of the glycolytic, TCA cycle, and the electron transport systems, with the latter two representing oxidative mechanisms of glucose utilization. For example, it has been found in the rat that the glycolytic enzymes in the brain, (e.g., glyceraldehyde-3-phosphate dehydrogenase, hexokinase, aldolase, and lactate dehydrogenase) already are considerably active after birth and show relatively small increases during the entire postnatal life, whereas the activities of the enzymes of the other two systems show marked variations. In the case of the rat brain, the latter enzymes generally remain at very low levels during the first 10 days after birth, following which their activities rapidly increase and reach maximum levels at about 40 days of age. Figure 10 shows examples with glyceraldehyde-3-phosphate dehydrogenase (glycolytic) and succinate dehydrogenase (TCA cycle), and Fig. 11 shows an example with cytochrome oxidase (electron transport chain). Activities of many other enzymes of oxidative metabolism (e.g., pyruvate dehydrogenase, citrate synthase, isocitrate dehydrogenase, and fumarase) also change markedly during this period.

The activities of the enzymes of both glycolytic and oxidative pathways develop first in the spinal cord and medulla, then in the hypothalamus, striatum, midbrain, and the cerebral white matter, and finally in the cerebral cortex and cerebellum (i.e., development is in the caudocephalic direction). This is consistent with the embryological and neurophylogenetic development of the brain regions, which is known to proceed from the medulla to the telencephalon (the end brain), and strongly indicates that the development of morphological and neurological maturity in the various regions of the brain is correlated with the development of their aerobic potential. Particularly interesting in this context is the development of the pyruvate dehydrogenase complex since this enzyme has a key regulatory role in controlling the flux of glucose carbon via pyruvate into the TCA cycle for energy metabolism in all tissues including the brain. Indeed, in the brain of species born neurologically mature (the prenatal brain developers such as guinea pig), pyruvate dehydrogenase activity is fully developed at birth, whereas in the brain of the purely postnatal brain developers such as the rat (which are born

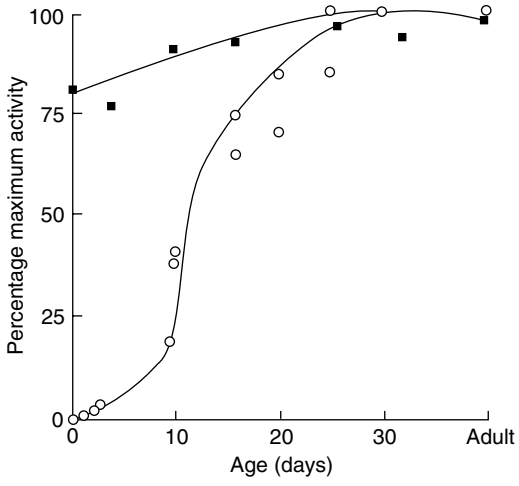


Figure 10 Changes in the activities of glyceraldehyde-3-phosphate dehydrogenase (●, moles substrate/kg protein of forebrain/hr) and succinate dehydrogenase (○, moles substrate/g whole brain/hr) in the developing rat brain. Values are plotted as percentages of maximal activities (reproduced with permission from Pergamon Press).

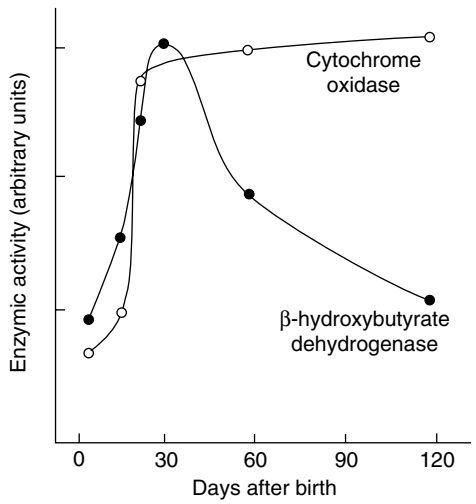


Figure 11 Changes in activities of cytochrome oxidase (○) and β-hydroxybutyrate dehydogenase (●) in the developing rat brain (reproduced with permission of the American Society of Biological Chemists).

neurologically immature), the activity of the enzyme is low. The subsequent neurological development of the latter species is correlated with the development of the pyruvate dehydrogenase activity in their brains.

An important aspect of energy metabolism in the neonatal brain concerns the utilization of ketone bodies, β-hydroxybutyrate and acetoacetate, as an additional source of energy. Probably because of an underdeveloped blood–brain barrier, the neonatal brain can take up ketone bodies and utilize them for the production of energy. During the early postnatal period, the ketone bodies are preferred to glucose as substrates for synthesis of phospholipids and sphingolipids to meet the requirements for growth and myelination. The ketone bodies thus form an important fuel metabolite for the rat brain during early periods after birth.

The pattern of cerebral metabolism of glucose also changes during development. Thus, in the adult brain of most mammalian species, the rate of glucose utilization ranges between 0.3 and 1.0 mmol/kg tissue/min, and most of this metabolism ($>90\%$) is carried through the glycolysis $\rightarrow$ TCA cycle $\rightarrow$ electron transport pathway. In contrast, in neonates, the hexose monophosphate (HMP) shunt represents a significant mechanism for the metabolism of a portion of total glucose. It has been estimated that the HMP shunt is responsible for as much as 50% of total brain glucose utilization during the first 4 weeks of life in the rat. Since production of ribose and NADPH is characteristic of the shunt, it is very likely that this pathway is utilized by the developing brain to meet its own demand for rapid synthesis of nucleic acids and large amounts of lipids, respectively. As the brain matures, the activity of the shunt gradually decreases, as reflected by a decrease with age in the activity of glucose-6-phosphate dehydrogenase, the first enzyme of the shunt, relative to the activities of the glycolytic and the TCA cycle enzymes.

V. INFLUENCE OF ADVERSE FACTORS ON BRAIN DEVELOPMENT

The process of brain development is accomplished by a series of continuous and sometimes overlapping events. These events of chemical and structural significance—organogenesis followed by cellular proliferation and later by synaptogenesis and myelination—proceed for a long time in a unidirectional manner until the mature adult tissue has emerged. These events, all occurring in early periods of life, must therefore be allowed to proceed in the genetically programmed fashion relative to age if a normally functioning tissue is to be formed. Any adverse extrinsic (environmental) or intrinsic (genetic) interference with the program would likely have deleterious effects on the ultimate makeup of the tissue. Because of the once-and-for-all characteristics of most developmental events and because the brain constituents are metabolically relatively inert, some effects are likely to persist throughout life.

During growth and development, the brain is especially vulnerable. Any derangement might have a profound effect on the functioning of the brain and, subsequently, on the behavior of the affected individual. If derangements caused in early life are permanent, the individual would be permanently deficient in behavior and mental functions. It is this apprehension that has aroused keen interest among neuroscientists in their search for teratogenic agents and their effects on brain development. Extrinsic agents include viruses, drugs, various forms of irradiation, nutritional insufficiency (including mineral and vitamin deficiencies), and probably also an inadequate environment. Intrinsic agents, such as defects in the genetic makeup of the brain, cause serious malformation of the brain as well as serious defects in brain function and metabolism. They are commonly referred to as inborn errors of metabolism.

Of the extrinsic factors, only viral infection and exposure to drugs and irradiation can affect organognesis and therefore also the later phases of brain growth and development. On the other hand, numerous studies have shown that nutritional inadequacy in early life can affect only the later processes of brain growth and development. The effects are not as profound and obvious as those found with defects in organogenesis. These studies have thus added a new dimension to the subject of teratology, with the increasing belief that important aspects of brain growth and development, such as cell multiplication, myelination, and synaptogenesis (which occur long after organogenesis is complete), are vulnerable to influences such as nutrition. The subject of nutrition and brain development has therefore become a widely studied field of research during the past few decades. This is not surprising in view of the fact that undernutrition and hunger are a part of life for at least one-third of the world's population. Children are the worst victims.

Indeed, protein-energy malnutrition is prevalent in varying degrees among children and mothers of almost all developing countries. It is also quite common for

poverty, hunger and malnutrition, concomitant with a high infant mortality rate, to also exist among the indigent communities of developed and industrialized countries also. Apart from high mortality rates among these children, particularly those less than 5 years old, there is increasing concern about the possible effects of nutritional insufficiency in early life on the later development of the survivors. The smaller stature, smaller brain, and more important, the apparent lower mental capabilities found in adolescents and adults of poor communities have often been correlated with restricted growth in early life.

VI. CONCLUSION

This article on brain development provides a simplified version of the subject. The process of brain development is much more complex and complicated than depicted, particularly for humans. Familial and social environment has been increasingly implicated in the development of the human brain, both structurally and functionally. Indeed, children raised in an enriched environment do better on intelligence tests than children raised under the same nutritional regime but in a poorer environment. Heredity is another important determinant in brain development.

The whole human situation is extremely complex, compounded by intricate interactions between nutrition, environment, and heredity on the overall development of physical and mental health. The brain stands in the center of this complex milieu, and, as noted by John Eccles, a great pioneer in brain research, we must continue to use our brains to understand how they show plasticity toward these and other possible factors to attain the ultimate potentiality of the development and function of our brains.

See Also the Following Articles

BRAIN ANATOMY AND NETWORKS • EVOLUTION OF THE BRAIN • GLIAL CELL TYPES • NERVOUS SYSTEM, ORGANIZATION OF • NEURON • NEUROTRANSMITTERS • SYNAPSES AND SYNAPTIC TRANSMISSION AND INTEGRATION

Suggested Reading

Adams, R. D., and Victor, M. (1993). *Principles of Neurology*, 5th ed. McGraw-Hill, New York.

Dickerson, J. W. T., and McGurk, H. (Eds.) (1982). *Brain and Behavioural Development*. Surrey Univ. Press, Guildford, UK.

Frackowiak, R. S. J., Friston, K. J., Frith, C. D., Dolan, R. J., and Mazziotta, J. C. (1997). *Human Brain Function*. Academic Press, San Diego.

Nolte, J. (1993). The Human Brain, 2nd ed. Mosby–Yearbook, St. Louis, MO.

Obeso, J. A., DeLong, M. R., Ohye, C., and Marsden, C. D. (1997). *Advances in Neurology*. Lippincott–Raven, Philadelphia.

Pansky, B., Allen, D. J., and Budd, G. C. (1988). *Review of Neuroscience*. Macmillan, New York.

Serafini, T. (1999). Finding a partner in a crowd: Neuronal diversity and synaptogenesis. *Cell* **98,** 133.

Stein, P. S. G., Grillner, S., Selverson, A. I., and Stuart, D. G. (1997). *Neurons, Networks and Motor Behaviour*. MIT Press, London.

Wauben, I. P., and Wainwright, P. E. (1999). The influence of neonatal nutrition on behavioural development: A critical appraisal. *Nutr. Rev.* **57,** 35.

Whitaker, J. N. (2000). Neurobiological understanding of myelination in the 21st century. *Arch. Neurol.* **57,** 57.

Yusuf, H. K. M. (1992). *Understanding the Brain and Its Development: A Chemical Approach*. World Scientific, Singapore.

Zeki, S. (1993). *A Vision of the Brain*. Blackwell, London.

Brain Disease, Organic

KENNETH MAIESE
Wayne State University School of Medicine

GLOSSARY

anoxia A cellular environment that is absent of oxygen.

brain herniation A movement of brain tissue from one compartment of high pressure to another compartment of decreased pressure.

coma A state of unresponsiveness in which the patient remains with eyes closed and is unarousable.

persistent vegetative state A condition following diffuse brain injury that results in the absence of cognitive function but the persistence of sleep–wake cycles.

programmed cell death An active process by an individual cell that requires energy and sometimes new protein synthesis which eventually leads to the directed destruction of the cell.

Organic brain disease is the result of neuronal and vascular injury that can occur as a consequence of several injury paradigms that include anoxia, vascular disease, and metabolic insults. Such insults can then subsequently precipitate both anatomic and cellular neuronal disease. However, it is the downstream cellular and molecular pathways that are increasingly being recognized as vital mechanisms that can both prevent and reverse neuronal injury. This article provides a pathophysiologic approach to the diagnosis and treatment of organic brain disease that incorporates current knowledge from both clinical patient studies and basic science investigations. Further analysis of neuronal injury that relies on a ''bimodal'' clinical and basic approach will foster the development of rationale, efficacious, and safe therapeutic interventions for the treatment of organic brain disease.

I. ANOXIC COMA

A. Historical Background

Initial credit for the recognition of the dependence of the vital organs on the circulatory system can be attributed to Galen in the 2nd century AD. The doctrines of Galenic physiology stated that blood was produced in the liver, flowed to the heart to obtain vital spirits, and subsequently bathed the brain to gain ''animal spirits.'' William Harvey extended this work during the 17th century to illustrate that blood within the human body was under continuous circulation. The vital spirits were later discovered to consist of oxygen.

Oxygen was discovered independently by Schiele in Sweden and by Priestly in England. It was named oxygen (acid-former) by Antoine Lavoisier (1743–1794) of France. Lavoisier made significant medical discoveries concerning oxygen's role in respiration and determined that oxygen makes up only approximately one-fifth of the volume of atmospheric air. In addition, oxygen is the only gas in air that sustains combustion and respiration. In animal experiments, Lavoisier and others discovered that anoxia can rapidly lead to death.

In 1920, Barcroft introduced the terms "anoxic," "anemic," "histotoxic," and "stagnant" to designate the various forms of anoxia. However, clinical (human) anoxia is complex and consists of several components such as decreased oxygen availability, systemic acidosis, hypercapnia, and eventual superimposed ischemia (loss of oxygen and blood flow). The synonymous use of the terms "anoxia" and "ischemia" reflects the fact that hypoxic and ischemic states usually overlap and it is sometimes difficult to determine which predominates.

B. Etiology

Lack of oxygen to the brain can be categorized as anoxic anoxia, anemic anoxia, and ischemia. Anoxic anoxia consists of reduced arterial oxygen content and tension. This may be secondary to decreased oxygen in the environment or an inability for oxygen to enter the circulatory system such as during cardiac arrest or during pulmonary disease. Anemic anoxia consists of low oxygen content in the blood secondary to decreased hemoglobin content. Ischemic anoxia describes a state of insufficient cerebral blood flow. Such decreased flow states may be secondary to cardiovascular collapse or conditions of increased vascular resistance, such as in stroke or migraine.

The brain normally consumes approximately 3.5 ml of oxygen for each 100 g of brain tissue per minute. When this rate declines to 2.5 ml, delirium and subsequent coma can develop. Rates of cerebral oxygen metabolism below 2 ml/100 g per minute are incompatible with a conscious state. The brain cannot store oxygen and survives only for minutes after its oxygen supply is reduced below critical levels. In acute anoxia, consciousness is lost within 15 sec. Pyramidal cells in the CA1 sector of the hippocampus, Purkinje cells of the cerebellum, and pyramidal cells of the third and fifth layers of the cerebral cortex are vulnerable to even moderate degrees of anoxia. Widespread necrosis of the cortex with the brain stem intact produces a vegetative state. More severe anoxia affecting the cortex, basal ganglia, and brain stem results in coma and subsequent death.

Under physiologic conditions, glucose is the brain's only substrate and crosses the blood–brain barrier by facilitated transport. During each minute, the normal brain uses approximately 5.5 mg (31 μmol) of glucose per 100 grams of tissue. If there is hypoglycemia, defined in adults as a blood glucose concentration of less than 40 mg/dl, signs and symptoms of encephalopathy result secondary to cerebral cortex or brain stem dysfunction. The cerebral cortex is more vulnerable to the effects of hypoglycemia, whereas the brain stem and basal ganglia exhibit less histologic damage during periods of reduced serum glucose. Although periods of hypoglycemia may precipitate or confound anoxic coma, serum glucose must be maintained in a closely controlled range to prevent further neurologic disability. Similar to the detrimental effects of hypoglycemia, periods of hyperglycemia (>180 mg/dl) have been shown to worsen neurologic outcome.

C. Clinical Presentation

Coma is a state of unresponsiveness in which the patient remains with eyes closed and is unarousable. During the initial assessment of the patient in coma, the physical examination of the patient becomes vital. Although the etiology of coma may be obvious in cases of diffuse cerebral ischemia following cardiac arrest, the patient should be evaluated for evidence of head trauma, such as scalp laceration, hemotympani, otorrhea, and rhinorrhea. During coma, patients may also suffer from peripheral nerve injuries. Compartment syndromes and compression neuropathies have been described in comatose individuals of duration ranging from 4 to 48 hr.

Although several conditions may precipitate hypoxic–ischemic coma, acute anoxic coma usually is a result of cardiac arrest. Approximately 1.5 million people per year in the United States succumb to a cardiac death. In addition, of the 200,000 attempted cardiac resuscitations, only 70,000 are considered successful. Of those considered successful, only 10% of these survivors return to their former lifestyles. Survival among young adults is also not very promising. In this population of patients, one-fourth of the cases of cardiac arrest are caused by ischemic heart disease and one-fourth are caused by drug overdose with maximum long-term survival estimated at 28%. Following cardiac arrest, individuals admitted to intensive care units in coma suffer a high mortality rate. For example, those in coma for more than 48 hr have a 77% mortality rate. In contrast, individuals not in coma have an 11% mortality rate.

The neurologic examination of an individual in coma consists of an assessment of the level of consciousness as determined by eye opening, verbal responses, and purposive movements in response to noxious stimulation of the face, arms, and legs. Neuroophthalmologic function is assessed by

documenting the pupillary size and response to light, spontaneous eye movements, oculocephalic (doll's eyes), and oculovestibular (ice water caloric) responses. This approach considers the brain in terms of its hierarchical, longitudinal organization into cortical and brain stem functions. Clinical neurologic signs can be correlated with specific anatomic sites to establish the severity and extent of central nervous system dysfunction.

The level of consciousness is determined by the degree of behavioral arousal. Attempts should be made to elicit a behavioral motor response by verbal stimulation alone. If no response follows verbal commands that are known to be clearly audible, noxious stimulation can be applied to the face by digital supraorbital pressure and individually to the arms and legs by gentle compression of distal interphalangeal joints with a nontraumatic object, such as a soft wood tongue blade. Eye opening indicates activity of the reticular activating system, whereas verbal responses indicate cortical hemispheric function.

The fundus of each eye should be examined for signs of increased intracranial pressure that would include papilledema or hemorrhage. Bilateral dilated, fixed pupils indicate enhanced sympathetic nervous system activity due to either an endogenous sympathetic discharge, such as during anoxia–ischemia, or to exogenous catecholamines. Comatose patients may have no spontaneous eye movements. In such cases, doll's eyes responses and an ice water caloric test can be used. Doll's eyes indicate the integrity of proprioceptive fibers from the neck structures, the vestibular nuclei, and the nuclei of the third and sixth cranial nerves. When doll's eyes are absent, it becomes necessary to perform the ice water caloric test. During severe bilateral functional cortical depression that leads to coma, the doll's eyes can disappear before the ice water caloric responses because the latter are produced by a stronger stimulus.

Decorticate or flexor responses occur following damage to the hemispheres or during diffuse metabolic depression of cortical function. Decerebrate or extensor responses correlate with destructive lesions of the midbrain and upper pons but also may be present during anoxic coma. The absence of motor responses, especially if flaccidity and areflexia are present, indicates severe brain stem depression and is frequently found in terminal coma or in severe sedative intoxication. Withdrawal and localizing responses imply purposeful or voluntary behavior. If the individual can follow commands, then this is considered to be a positive response and marks the return of consciousness.

During cerebral anoxia, the brain experiences insufficient delivery of oxygen (Table I). Generalized brain anoxia is usually a consequence of systemic circulatory arrest caused by cardiac arrhythmia. Although other organs, such as the kidney and heart, can tolerate ischemic periods of up to 30 min, the brain can tolerate no more than a few minutes of anoxia. Brief episodes of cerebral anoxia are usually well tolerated, with patients escaping any irreversible deficits. However, an amnestic syndrome may follow transient periods of global ischemia. Patients may experience a severe antegrade amnesia and variable retrograde memory loss with preservation of immediate and remote memory resembling Korsakoff's psychosis. Individuals with anoxic–ischemic coma of more than 6 hr duration, but with unremarkable cranial magnetic resonance imaging (MRI) or cranial computed tomography (CT) imaging, have demonstrated persistent poor learning and recall of paired associations when compared with age- and intelligence quotient-matched controls. During a recovery from the immediate effects of an anoxic insult, a few patients have been known to return to an unconsciousness state. Delayed neurologic deterioration occurs in approximately 1 or 2 individuals per 1000 cardiac arrests and is not predictable by the type of insult, duration of anoxia, or any other identifiable variable.

Severe or prolonged periods of hypotension can result in focal cerebral ischemic lesions that result in coma. Individuals usually remain in coma for at least 12 hr and on awakening experience deficits including partial or complete cortical blindness, bibrachial paresis, and quadriparesis. Cortical blindness is rarely permanent in nature and is a result of ischemia to either occipital pole, which are located in an arterial border zone. Bilateral infarction of the cerebral motor cortex in the border zone between the anterior and middle cerebral arteries appears to be responsible for the syndrome of bibrachial paresis, sparing the face and lower extremities following cardiac arrest.

The spinal cord is considered to be more resistant to ischemic insults than more rostral sections of the central nervous system. However, cases of isolated spinal cord infarction can occur without evidence of cerebral injury. Syndromes of spinal cord ischemia following transient hypotension are characterized by flaccid paralysis of the lower extremities, urinary retention, and a sensory level in the thoracic region with the anterior spinal–thalamic tracts usually more affected than the posterior columns.

Table I
Clinical and Pronostic Correlation Following Cerebral Anoxic Injury

Anoxic insult	Clinical signs	Prognosis
Rapid and less than 5 min	Transient cognitive or memory loss	Complete recovery, rare delayed deterioration
Prolonged and more than 6 hr	*Cerebral hemisphere dysfunction*	Incomplete recovery with deficits
	Cognitive loss	
	Personality changes, depression	
	Brachial paresis or quadriparesis	
	Cortical blindness	
	Seizures	
	Myoclonus	
	Parkinsonism	
	Spinal cord dysfunction	Incomplete recovery with deficits
	Flaccid paralysis of the lower extremities	
	Urinary retention	
	Sensory loss of pain, temperature	
Severe circulatory arrest	Vegetative state	Chronic vegetative state
	Loss of cortical and brain stem function	Brain death

The most devastated group of patients following diffuse anoxic injury suffer widespread destruction of the cerebral cortex and progress to either a persistent vegetative state or death from neurologic complications. Extrapyramidal tract dysfunction following anoxia, such as during cardiac arrest or carbon monoxide poisoning, can produce a clinical syndrome identical to Parkinson's disease. Parkinsonian features may represent only a small part of widespread cerebral injury, but in other instances the clinical presentation of rigidity, akinesia, and tremor may represent the only neurologic disability.

Movement disorders, such as myoclonic jerks and cerebellar ataxia, may also follow episodes of anoxic injury. The "action myoclonus" syndrome can occur following global cerebral ischemia. The myoclonic jerks are frequently stimulus activated by light, sound, or initiation of movement and can incapacitate individuals in their daily living activities. Therapy with various agents, such as serotoninergic agents, clonazepam, and valproic acid, has been used with some success. Cerebellar ataxia, involving the trunk or extremities, is an infrequent postanoxic syndrome and may be secondary to the selective vulnerability of Purkinje cells to anoxia.

D. Clinical Evaluation and Management

Imaging studies such as cranial CT or MRI can assist in determining the presence of complications following an anoxic insult. These studies can differentiate between an ischemic infarct, intracerebral hemorrhage, and a mass lesion involving the cortex or the brain stem. A CT without contrast is helpful in suspected cases of cerebral hemorrhage. Within the first 72 hr of intracerebral hemorrhage onset, a cranial CT usually provides greater resolution than a MRI. In addition, CT and MRI are useful to demonstrate cerebral herniation prior to clinical presentation. Both positron emission tomography (PET) and magnetic resonance spectroscopy have also been used to follow cerebral metabolic function in individuals suffering from cerebral anoxic injury. Recently, MRI with the apparent diffusion coefficient of water has become a sensitive tool of neuronal physiology and may represent a reliable indicator of progressive neuronal injury following cerebral ischemia.

The electroencephalogram (EEG) can be useful in assessing cortical dysfunction and identifying the presence of epileptic activity. The EEG is classified in terms of increasing severity in five categories. Grade I

represents normal alpha with theta–delta activity; grade II is theta–delta activity with some normal alpha activity, grade III is dominant theta–delta activity with no normal alpha activity, grade IV is low-voltage delta activity with alpha coma (nonreactive alpha activity), and grade V represents an isoelectric tracing. In individuals suffering postanoxic coma, grade I is compatible with a good prognosis, grades II and III have no definitive predictive value, and grades IV and V are compatible with a poor prognosis and infrequent recovery.

In addition to the EEG, evoked potentials can provide information regarding the functional state of the cerebral cortex following cerebral anoxia. Somatosensory evoked potentials determine the functional integrity of the spinal cord posterior columns, brain stem medial lemniscus, thalamus, and frontoparietal sensorimotor cortex. During the loss of bilateral cortical responses despite the etiology of the coma, afflicted individuals can experience a high mortality rate. In anoxic coma, patients who maintain normal responses throughout their illness maintain a good prognosis but may have permanent neurologic sequelae. Brain stem auditory evoked potentials correlate with brain stem dysfunction during coma. A simultaneous latency increase of all components is consistent with progressive ischemia of the posterior fossa and a decrease in cerebral perfusion pressure. Although brain stem auditory evoked potentials are rarely modified by exogenous factors, they can be altered by hypothermia, anesthetics, and barbiturates.

Recovery of the comatose patient is dependent on the rapid treatment of the underlying disorder. Prompt attention must be directed to the restoration of respiratory, hemodynamic, and metabolic homeostasis. The respiratory rate and its pattern should be documented prior to therapeutic measures such as intubation and mechanical ventilation. Following initial examination of the respiratory rate, an adequate airway should be obtained. If intubation is required in a comatose patient, one must rule out the existence of a neck fracture prior to hyperextension of the head for endotracheal tube insertion. Arterial blood gases should be obtained to ensure adequate oxygenation (oxygen saturation $>90\%$) and to monitor serum acid/base status.

On the establishment of adequate ventilation, blood should be obtained for determination of serum glucose, routine chemistries, and toxicology. Since patients in coma may have poor nutrition and are susceptible to Wernicke's encephalopathy, initially 100 mg of thiamine should be given intravenously. Bedside stat glucose determinations should be employed to identify hypoglycemia. In such cases, 50 mg of 50% dextrose should be administered. Although administration of one ampule of dextrose is not detrimental in cases of hyperosmolar coma, identification of hyperglycemic states is important since elevated serum glucose may promote ischemic damage in cases of anoxic coma.

The hemodynamics of the patient should be closely controlled. Hypertension may be secondary to Cushing's reflex with increased intracranial pressure or a result of brain stem ischemia. Hypotension may be indicative of myocardial infarction, hemorrhagic shock, sepsis, or sedative–hypnotic drug overdose. Bradycardia associated with elevated blood pressure suggests brain stem compression or increased intracranial pressure. One must note, however, that elevated intracranial pressure does not decrease the heart rate in all instances. Reversible causes of transtentorial herniation, such as subdural hematoma, should be immediately considered before cardiovascular collapse ensues.

Status epilepticus following cardiac arrest can result in progressive anoxic brain damage and requires immediate attention. Following airway stabilization, generalized convulsions can initially be treated with diazepam intravenously in up to a 10-mg total dose. This is to be followed by a phenytoin loading dose of 18 mg/kg (50 mg/min) intravenously. If status epilepticus continues, 20 mg/kg phenobarbital intravenously should be administered. Persistent convulsions at this point require general anesthesia. In the case of generalized convulsions that are not consistent with status epilepticus, phenytoin orally in appropriate daily dose (dependent on body size) should be maintained in individuals with either EEG or CT/MRI evidence of a persistent epileptic focus (hemorrhage, neoplasm, large ischemic infarct, abscess, etc).

Measurement of the patient's rectal temperature is a vital component of the initial evaluation. Hypothermic patients with temperatures below 34°C (93.2°F) should be warmed slowly to a body temperature higher than 36°C (96.8°F). Since hypothermia below 80°F results in coma, resuscitative measures are indicated in all hypothermic patients even if vital signs are absent. Hypothermic patients have recovered following cardiac arrest, presumably because of the protective effects of low body temperature and depressed cerebral oxygen requirements. In addition, hypothermia has been shown to reduce neuronal death in the hippocampus and caudate putamen in animal models with forebrain ischemia and is currently under investigation in the clinical setting.

Some centers advocate aggressive treatment of raised intracranial pressure to significantly reduce mortality. Measurement of intracranial pressure can be performed through the use of epidural monitoring or through intraventricular pressure measurements. Intracranial pressure monitoring can differentiate between active hydrocephalus and mass lesions requiring surgical intervention. Intracranial pressure monitoring has also been linked to prognosis. Most patients with a maximum intracranial pressure increase of less than 30 mmHg experience good recovery, whereas a pressure increase above 25–30 mmHg represents a great risk for brain tamponade.

E. Clinical Prognosis

Several studies have examined the prognosis of individuals who remain in coma for extended periods following cardiac arrest. Patients can suffer memory impairment if coma duration is at least 6 hr. In some groups of patients, a poor prognosis was evident in individuals who lacked a motor response to pain. Poor recovery has also been associated with the presence of generalized myoclonus status, which can be suggestive of diffuse neocortical damage. More comprehensive studies evaluating individuals in coma following diffuse global ischemia have demonstrated that individuals with absent pupillary light reflexes never regain independent daily function. However, following the initial insult, the early onset of incomprehensible speech, orienting spontaneous eye movements, or the ability to follow commands were indicative of a good prognosis. Individuals with the best chance of recovery have preserved brain stem function following the initial insult. The most favorable sign of a good outcome is incomprehensible speech, such as moaning. At Day 1, the following signs are each associated with at least a 50% chance of regaining independent function: any form of speech, orienting spontaneous eye movements, intact oculocephalic or oculovestibular responses, ability to follow commands, and normal skeletal tone.

II. METABOLIC COMA

A. Historical Background

The patient in metabolic coma may be the cause of some of the greatest diagnostic errors ever made. Some historians believe that the description of Jesus resurrecting his friend Lazarus from the dead may represent one of the first reports of an individual in coma that eventually recovered from a metabolic disability. In addition, recent descriptions, documented approximately 150 years ago, discuss the diagnosis of ''apparent death'' as a possible synonym for coma secondary to metabolic illness.

B. Etiology

Although levels of attention and alertness are affected in metabolic coma, each disease process also yields a specific clinical picture. For example, severe anoxic ischemia following cardiac arrest will produce coma, whereas alcohol withdrawal will initially result in an agitated delirium. Metabolic encephalopathy is often reversible if the underlying systemic disorder is corrected promptly.

Under physiologic conditions, glucose is the brain's only substrate and crosses the blood–brain barrier by facilitated transport. Each minute, the normal brain requires 5.5 mg (31 μmol) of glucose per 100 grams of tissue. However, one of the most significant complications of exogenous insulin therapy is hypoglycemic coma. If there is hypoglycemia, defined in adults as a blood glucose concentration of less than 40 mg/dl, loss of cortical function ensues secondary to cerebral cortex and brain stem dysfunction. Neurologic presentation during hypoglycemia can be variable. Some patients will present with focal motor or sensory deficits, whereas others become comatose.

A significant proportion of cases of metabolic coma are a result of drug ingestion, with at least half of the afflicted individuals administering multiple drugs. Proper diagnosis relies heavily on the physical examination since an accurate or complete history may be unobtainable from the patient. Toxic drug screens of blood and urine will assist in the diagnosis. Excessive barbiturate consumption results in hypothermia, hypotension, and possible apnea. An individual's pupils remain small but reactive with intact ciliospinal reflexes.

Alcohol ingestion may be indistinguishable from other metabolic disorders, such as depressant drug intoxication or hypoglycemia. In addition, progressive loss of consciousness may be complicated by underlying cerebral trauma such as a subdural hematoma. Evidence of ''alcohol on the breath'' provides insight into the etiology of the coma but does not distinguish between pure alcohol intoxication or a ''cocktail'' of

alcohol, sedative, and hypnotic drugs. A blood alcohol level and drug screen should be determined in individuals with alcohol and/or multiple drug abuse who present with coma.

Benzodiazepines can result in stupor or coma without respiratory depression. Clinical trials have studied treatment with the benzodiazepine antagonist flumazenil. Flumazenil is also considered to have a diagnostic value in cases of mixed-drug intoxication.

Opiate and heroin overdoses occur by either parenteral injection or sniffing of the agent. Neurologic impairment secondary to opiate abuse is currently on the rise despite a reduction in use during previous years. Systemic complications include hypothermia, hypotension, bradycardia, respiratory slowing, and pulmonary edema. Coma does not require chronic administration and can result following an initial injection of the opiate. Opiate coma is characteristically associated with pinpoint pupils reactive to bright light.

Toxicity from cocaine administration has systemic, psychiatric, and neurologic manifestations. Neurologic complications of cocaine use range from benign headaches to coma. Focal neurologic manifestations include subarachnoid hemorrhage, anterior spinal artery syndrome, lateral medullary syndrome, transient ischemic attacks, and cerebral infarction. Although the exact mechanism of cocaine-related cerebral vascular disease is unknown, adrenergic stimulation and surges in blood pressure may play a significant role.

Overmedication with prescription agents can also lead to coma. The most notable neurologic manifestations of tricyclic antidepressants are seizures and coma. Lithium toxicity can progress to seizures, lethargy, and coma. Treatment consists of supportive care with artificial ventilation, fluid, and electrolyte infusions in conjunction with hemodialysis. Valproate can lead to coma during excessive drug levels or during carnitine insufficiency. Amantadine, an antiviral agent employed in the therapeutic regimes for Parkinson's disease and fatigue syndromes, has been reported to induce coma in the presence of end-stage renal disease. Ibuprofen, a popular over-the-counter analgesic agent, is rarely associated with nervous system toxicity, but abuse of this agent can result in metabolic acidosis and lead to coma.

In addition to drug administration, other metabolic mechanisms that lead to coma should be considered. For example, endocrine dysfunction that occurs with severe hypothyroidism or renal insufficiency can result in coma. Infectious processes, such as cerebral malaria and meningitis, lead to increased intracranial pressure and subsequent coma if allowed to progress without treatment. Less common causes of coma, such as ciguatera food poisoning and hyperammoniemia, should also be considered when the diagnosis appears obscure.

C. Clinical Presentation

Metabolic coma is significant not only in clinical terms but also on economic grounds. For example, individuals admitted in coma with methanol poisoning suffer a significant mortality rate of 64%. The annual cost for this intensive care management of high-mortality patients is more than $40 billion.

The patient in coma as a result of a metabolic disorder requires a detailed physical and neurologic examination. The patient should be evaluated for evidence of head trauma, such as scalp laceration, hemotympani, otorrhea, and rhinorrhea. Blisters of the skin during coma may be suggestive of barbiturate overdose, whereas bullous skin lesions have been associated with antipsychotic drug ingestion. Patients in coma may also succumb to compartment syndromes and compression neuropathies.

Similar to the assessment of individual's suffering from anoxic coma, the neurologic examination consists of an assessment of the level of consciousness as determined by verbal responses, eye opening, and purposive movements. Attempts should be made to elicit a behavioral motor response by verbal stimulation alone. If no response follows even shouted commands, noxious stimulation can be applied to the face by digital supraorbital pressure and individually to the arms and legs by compression of distal interphalangeal joints with a nontraumatic object, such as a soft wood tongue blade. Verbal responses are indicative of dominant hemisphere function, whereas eye opening indicates activity of the reticular activating system.

During the neuroophthalmologic examination, the fundus of each eye should be examined for papilledema or hemorrhage. In coma due to metabolic brain disease, the pupils are generally small but reactive to light. Small, reactive pupils are present in normal individuals during sleep and are a common finding in elderly persons due to degenerative changes in the iris and ciliary muscles. Small, sluggishly reactive pupils that respond to naloxone administration are characteristic of an overdose of opiates. Pinpoint pupils occur in brain stem pontine compression.

Bilateral dilated, fixed pupils may be secondary to endogenous sympathetic discharge following periods of diffuse anoxia or to the release of exogenous catecholamines. Dilated pupils are also seen in glutethimide-induced coma and overdosage with tricyclic antidepressant or other atropine-like agents. In coma due to amphetamine, cocaine, and LSD overdosage, the pupils are large but reactive. Midposition, fixed pupils are indicative of midbrain failure and loss of both sympathetic and parasympathetic pupillary tone, whether caused by structural or metabolic disease. A unilateral, dilated fixed pupil suggests damage to parasympathetic fibers of the external portion of the third cranial nerve.

In comatose patients without spontaneous eye movements, doll's eyes responses and the ice water caloric test can be used to determine the integrity of the eighth, sixth, and third cranial nerves and their interconnecting brain stem pathways. Doll's eyes indicate the integrity of proprioceptive fibers from the neck structures, the vestibular nuclei, and the nuclei of the third and sixth cranial nerves. If the cortical influences are depressed but brain stem gaze mechanisms are intact, the eyes will deviate conjugately to one side when the head is rotated to the opposite side. When doll's eyes are absent, it becomes necessary to perform the ice water caloric test. The caloric response may be absent not only during barbiturate and phenytoin overdosage but also during brain stem lesions and labyrinthine disease.

The motor examination provides insight into the functional integrity of the neuronal networks linking the cortex, brain stem, and pyramidal tracts. Although techniques such as transcranial magnetic evoked potentials are sometimes used to assess the pyramidal tract pathways in comatose patients, the neurologic examination remains the mainstay for assessment of the patient in metabolic coma. The absence of motor response, especially if flaccidity and areflexia are also present, indicates severe brain stem depression and is frequently found in severe sedative intoxication. Decerebrate or extensor responses correlate with destructive lesions of the midbrain and upper pons but may also be present in reversible conditions such as anoxic coma. Decorticate or flexor responses occur after damage to the hemispheres as well as in metabolic depression of brain function. Withdrawal and localizing responses imply purposeful or voluntary behavior. Obeying commands is the best response and marks the return of consciousness. Generalized or focal repetitive movements, not affected by stimuli, usually represent seizure activity.

D. Clinical Evaluation and Management

Imaging with cranial CT or MRI can be useful in metabolic coma to differentiate between an ischemic infarct, an intracerebral hemorrhage, and a mass lesion involving the cortex or the brain stem. However, these imaging studies are often unremarkable during metabolic coma. Patients in coma who present with cranial nerve deficits and posturing likely suffer from a mass lesion involving the cortex or the brain stem. However, patients in coma with unilateral masses may not initially suffer from transtentorial herniation. In this group, CT and MRI demonstrate horizontal displacement at the level of the pineal body.

In some centers, continuous bedside monitoring of regional cerebrocortical blood flow with laser Doppler flowmetry is employed to follow fluctuations in the cerebral microcirculation in comatose individuals. PET and single positron emission computerized tomography (SPECT) provide an alternative method of imaging that also assesses cerebral function. MRI spectroscopy may provide suggestive evidence of neuronal function through *N*-acetylaspartate/creatinine ratios. Blood flow and metabolic studies can also be used as diagnostic aids to differentiate individuals in coma from patients with the "locked-in" syndrome.

Monitoring with EEG is useful in assessing cortical dysfunction since it is sensitive to fluctuations in cerebral blood flow and metabolism. In addition, EEG can detect the presence of occult seizure activity, which reportedly can occur in more than 30% of intensive care patients suffering from metabolic disorders. In patients with overt convulsions, the EEG can also be used to temper anticonvulsant treatment in the comatose patient to prevent or reduce neuronal cell loss. In addition to issues of care, the EEG has been used to assess prognosis. In patients with terminal coma, onset of abnormal EEG changes may sometimes be suggestive of a poor outcome.

Evoked potentials can provide information concerning the functional state of the cerebral cortex and brain stem. Patients with bilateral absence of cortical responses exhibit mortality rates of 73–98%. Others report that motor evoked potentials, although less sensitive than somatosensory evoked potentials during coma, are helpful in directing management. The bilateral absence of motor evoked potentials is suggestive of a poor outcome during metabolic coma. Brain stem auditory evoked potentials correlate with brain stem function and can provide a useful tool for the assessment of brain stem activity. Simultaneous latency increase of all components can be consistent

with progressive ischemia of the posterior fossa and a decrease in cerebral perfusion pressure. Loss of the brain stem auditory evoked potentials is usually suggestive of an incumbent deterioration and death. Although brain stem auditory evoked potentials are not usually modified by exogenous factors, brain stem auditory evoked potentials can be falsely altered by hypothermia, anesthetics, and barbiturates.

Treatment of the comatose patient must be directed toward the restoration of respiratory, hemodynamic, and metabolic function. The respiratory rate and its pattern should be documented prior to therapeutic measures such as intubation and mechanical ventilation. Following initial examination of the respiratory rate, an adequate airway should be obtained. If intubation is required, one must rule out the existence of a neck fracture prior to hyperextension of the head for endotracheal tube insertion. Arterial blood gases should be obtained to ensure adequate oxygenation and to monitor serum acid/base status.

On the establishment of adequate ventilation, blood should be obtained for determination of serum glucose, routine chemistries, and toxicology. Since patients in coma may have poor nutrition and are susceptible to Wernicke's encephalopathy, initially 100 mg thiamine should be given intravenously. Bedside glucose determinations should be employed to identify hypoglycemia. Identification of hyperglycemic states is also important since elevated serum glucose may promote ischemic damage in cases of anoxic coma. Naloxone should be administered intravenously in cases of suspected opiate abuse, and flumazenil should be administered in cases of benzodiazepine-induced coma.

Measurement of the patient's rectal temperature is a vital component of the patient's care. Hypothermia can be due to environmental exposure, near drowning, sedative drug overdose, hypothyroidism, and Wernicke's disease. Hypothermic patients with temperatures below 34°C (93.2°F) should be warmed slowly to a body temperature higher than 36°C (96.8°F). Since hypothermia below 80°F results in coma, resuscitative measures are indicated in all hypothermic patients even if vital signs are absent. The presence of fever in a comatose patient requires investigation for an underlying infection.

Seizure control is also critical to the management of the metabolic comatose patient. Status epilepticus can result in permanent anoxic brain damage and requires immediate attention. Following airway stabilization, generalized convulsions can initially be treated with diazepam intravenously up to 10-mg total dose. This is to be followed by a phenytoin loading dose of 1000 mg (50 mg/min), but may be increased to 1500 mg if required.

E. Clinical Prognosis

In most cases, coma secondary to infection or sedative drug intoxication carries a low mortality rate. If ventilatory and hemodynamic support are supplied without delay, most individuals experience no residual neurologic impairment. However, complications can include cardiovascular collapse secondary to hypothermia, hypotension, dysrhythmias, myocardial infarction, or pulmonary edema.

In cases complicated by cerebral ischemia, patients can suffer permanent neurologic sequelae if coma duration is at least 6 hr. Individuals with absent pupillary light reflexes usually never regain independent daily function, but the early onset of incomprehensible speech, orienting spontaneous eye movements, or the ability to follow commands can be indicative of a good outcome following the initial insult.

Other factors also play a role in the eventual outcome from coma. Metabolic coma complicated by traumatic lesions, such as subdural hematoma, can have less than a 10% recovery rate. Comatose patients with increased plasma glucose, hypokalemia, elevated serum leukocyte counts, or absent P300 event-related potentials also have a poor prognosis.

III. BRAIN HERNIATION

A. Historical Background

In the 17th century, the human body began to be viewed as a system of subunits and independent compartments. This eventually led to the first human anatomical descriptions that mapped the body into different organs and tissues. As a result of this "subunit" or "compartment" theory, the Latin term *herniation* was employed to describe the protrusion of a portion of an organ or tissue through an abnormal passage.

B. Etiology

Brain herniation may result from either supratentorial or subtentorial lesions. Supratentorial masses, such as

those that result from lobar hemorrhage, cause brain shifts that can be termed cingulate, central (transtentorial), or uncal herniation. Cingulate herniation refers to the displacement of the cingulate gyrus under the falx cerebri with subsequent compression of the internal cerebral vein. A mass in the cerebral hemisphere that produces cingulate herniation can compress the ipsilateral anterior cerebral artery, producing subsequent vascular ischemia, edema, and progressive mass effect.

Downward displacement of the hemisphere with compression of the diencephalon and midbrain through the tentorial notch results in central herniation. Lesions of the frontal, parietal, and occipital lobes initially can precipitate cingulate herniation that progresses to central herniation. Displacement of the diencephalon against the midbrain produces hemorrhage in the pretectal region and thalamus. The medial perforating branches of the basilar artery rupture during herniation of the midbrain and pons.

Uncal herniation involves shift of the temporal lobe, uncus, and hippocampal gyrus toward the midline with compression of the adjacent midbrain. During this process, the ipsilateral third cranial nerve and the posterior cerebral artery are compressed by the uncus and edge of the tentorium. Both central and uncal herniation can compress the posterior cerebral artery, resulting in occipital lobe ischemia. Further increased intracranial pressure results from compression of the aqueduct. In this instance, cerebral spinal fluid cannot drain from the supratentorial ventricular system, producing a pressure gradient between structures above and below the obstruction. In addition, expansion of the supratentorial volume can yield pressure necrosis of the parahippocampal gyrus.

Supratentorial herniation results in an orderly progression of neurologic dysfunction from the cerebral hemispheres to the brain stem. Exceptions to this observation exist. Massive cerebral hemorrhage can rapidly flood the ventricular system, compress the fourth ventricle, and result in acute medullary failure. Isolated medullary failure can also occur from withdrawal of cerebral spinal fluid by a lumbar puncture in a patient with incipient central herniation from a supratentorial mass.

C. Clinical Presentation

Untreated or progressive cerebral herniation can result in increased morbidity and mortality rates. For example, in cases of fulminant hepatocellular failure, cerebral herniation is the leading cause of death and disability. Even in controlled environments such as the operating room, cerebral herniation can be the primary cause of death in approximately 15% of patients.

Patients with elevated intracranial pressure resulting in cerebral herniation require a detailed physical and neurologic examination. Similar to patients presenting with coma, individuals should be evaluated for evidence of head trauma, such as scalp laceration, hemotympanum, otorrhea, and rhinorrhea. The neurologic examination consists of an assessment of the level of consciousness as determined by verbal responses, eye opening, and purposeful movements. Attempts should be made to elicit a behavioral motor response by verbal stimulation alone. If no response follows even shouted commands, noxious stimulation can be applied to the face by digital supraorbital pressure and individually to the arms and legs by compression of distal interphalangeal joints with a nontraumatic object, such as a soft wood tongue blade. Verbal responses are indicative of dominant hemisphere function, and eye opening indicates activity of the reticular activating system.

During the neuroophthalmologic examination, the fundus of each eye should be examined for papilledema or hemorrhage. Although episodic anisocoria associated with headaches is usually benign, a unilateral, dilated fixed pupil (sometimes referred to as a "blown pupil") suggests damage to parasympathetic fibers of the external portion of the third cranial nerve as a result of brain herniation. In contrast, pinpoint pupils suggest compression of pontine structures. Midposition, fixed pupils indicate midbrain failure and loss of both sympathetic and parasympathetic pupillary tone, whether caused by structural or metabolic disease.

As previously described for comatose patients without spontaneous eye movements, doll's eyes responses and the ice water caloric test can be used to determine the integrity of the third, sixth, and eighth cranial nerves and their interconnecting brain stem pathways. If the cortical influences are depressed but brain stem gaze mechanisms are intact, the eyes will deviate conjugately to one side when the head is rotated to the opposite side. When the doll's eyes' responses are absent, it is necessary to perform the ice water caloric test.

The motor examination provides insight into the functional integrity of the neuronal networks linking the cortex, brain stem, and pyramidal tracts. Although diminution of brain stem auditory evoked potentials and somatosensory evoked potentials can be sugges-

tive of transtentorial brain herniation, the neurologic examination remains the mainstay for assessment of the patient with increased intracranial pressure. The absence of motor response, especially if flaccidity and areflexia are also present, indicates severe brain stem depression. Decerebrate or extensor responses correlate with destructive lesions of the midbrain and upper pons. Decorticate or flexor responses occur after damage to the hemispheres and in metabolic depression of brain function. Withdrawal and localizing responses imply purposeful or voluntary behavior, but the ability to follow commands is considered to be the best response and marks the return of consciousness.

D. Clinical Evaluation and Management

Supratentorial lesions can be classified as either extracerebral or intracerebral. Extracerebral lesions include neoplasms, infections, and trauma-related injuries such as hematomas. Lesions such as neoplasms or abscesses impair consciousness via the mass effect that they exert. Headaches, seizures, motor sensory deficits, and cranial nerve dysfunction, rather than an altered state of consciousness, are usually the initial symptoms of neoplasms. Occassionally, a progressing frontal lobe lesion can produce behavioral changes prior to brain herniation. Subdural empyema, a process secondary to otorhinologic infection, meningitis, or intracerebral abscess, can present as an extracerebral lesion. Initial presentation includes subdued consciousness, sinusitis, headaches, focal skull tenderness, and fever. Further deterioration can lead to language dysfunction, hemiparesis, seizures, and eventual coma.

Intracerebral disease initially produces focal deficits involving motor, sensation, or language function prior to altering consciousness. As previously described, frontal lobe lesions, as well as intraventricular masses, may present only with changes in personality before progressing to cerebral herniation.

Among supratentorial vascular lesions, hemorrhage into the cerebral parenchyma is the most common cause of altered consciousness. The etiologies of cerebral hemorrhage include rupture of an intracerebral blood vessel or arterial aneurysm rupture, leakage from an arteriovenous malformation, and bleeding into a metastatic tumor. Primary hypertensive cerebral hemorrhages usually involve the putamen, internal capsule, or thalamus. When these structures are involved, patients present with motor sensory deficits, headache, or loss of language function. The size of the hemorrhage usually correlates with the neurologic deficit, but clinical symptoms such as headache intensity may be more dependent on the anatomical location of the hemorrhage. A parenchymal cerebral bleed may extend intraventricularly. Ventricular system hemorrhage alone produces little more than a chemical meningitis. Downward brain herniation can occur when rapid flooding of blood into the ventricles causes a pressure gradient between the anterior and posterior fossae. Large cerebral infarcts can present with obtundation that sometimes results in brain herniation. Cerebral vascular congestion and edema reach a maximum within 4 days following the infarct. At this point, patients are at risk for transtentorial herniation.

Lesions below the tentorium can involve the brain stem reticular activating system directly or impair its function by compression. Coma arises from intrinsic brain stem lesions that destroy the paramedian midbrain or impair its vascular supply. Mass lesions alter consciousness by exerting direct pressure on the midbrain and pons with resultant ischemia and necrosis.

Upward transtentorial herniation of the cerebellum and mesencephalon occurs with enlarging masses of the posterior fossa. During this process, the posterior third ventricle is compressed, obstructing cerebral spinal fluid flow and contributing to increased intracranial pressure. Upward herniation also distorts the vasculature of the mesencephalon, compresses the veins of Galen and Rosenthal, and produces superior cerebellar infarction from occlusion of the superior cerebellar arteries.

Downward herniation of the cerebellar tonsils from a posterior fossa mass or hemorrhage occurs as an isolated process or in conjunction with upward herniation and direct brain stem compression. It is normal for the cerebellar tonsils to extend 2 cm into the cervical canal. Downward compression of the tonsils into the foramen magnum produces ischemia of the cerebellar tonsils, medulla, and upper cervical cord.

Imaging studies, such as cranial CT or MRI, can assist in determining the etiology of cerebral herniation. These studies can differentiate between an ischemic infarct, intracerebral hemorrhage, and a mass lesion involving the cortex or the brain stem. A CT without contrast is helpful in suspected cases of cerebral hemorrhage. As previously noted, CT within the first 72 hr of onset of intracerebral hemorrhage usually provides greater resolution than MRI. Contrast CTs are necessary to visualize the cerebral

vasculature following the exclusion of a cerebral hemorrhage. Patients in coma who present with cranial nerve deficits and posturing are most likely to suffer from a mass lesion involving the cortex or the brain stem. However, patients in coma with unilateral masses may not initially suffer from transtentorial herniation. In this group, CT and MRI may demonstrate early downward displacement of the midbrain. Although the clinical exam remains the gold standard for the initial assessment of cerebral herniation, MRI can be a valuable asset in the early recognition of cerebral herniation.

During the acute management of an individual with suspected brain herniation, the respiratory rate and its pattern should be documented prior to therapeutic measures such as intubation and mechanical ventilation. Following initial examination of the respiratory rate, an adequate airway should be obtained. Alterations in respiration correlate well with brain dysfunction. With supratentorial lesions, Cheyne–Stokes respiration may be evident, with waxing and waning hyperpnea alternating with shorter periods of apnea. Progressive brain herniation involving the midbrain–upper pontine tegmentum yields central neurogenic hyperventilation characterized by an increase in the rate and depth of respiration to the extent that respiratory alkalosis results. Low pontine lesions result in apneustic breathing, which consists of respiratory pauses of approximately 3 sec following full inspiration.

Treatment for elevated intracranial pressure relies on several modalities. The goal is to normalize elevated intracranial pressure by restricting cerebral blood flow or fluid to the brain tissue. Currently used treatments include hyperventilation, diuresis, fluid restriction, blood pressure control, steroids or surgery when indicated, and drug therapy.

Hyperventilation produces hypocapnia, which in turn decreases cerebral blood flow globally through vasoconstriction. Reduction in pCO_2 from 40 to 30 mmHg can reduce intracranial pressure approximately 30%. It is recommended to maintain pCO_2 at 25–30 mmHg since aggressive hypoventilation below 25 mmHg may reduce cerebral blood flow excessively and result in cerebral ischemia.

Fluid restriction and fluid elimination are both employed in the treatment of elevated intracranial pressure. In some cases, hypotensive therapy has been advocated. Normal saline is administered to maintain normal serum osmolality and to prevent systemic intravascular depletion. Restriction of normal saline infusion to one-half maintenance can help reduce intracranial pressure. In addition, osmotic agents can be administered acutely to control intracranial pressure. These drugs do not cross the blood–brain barrier and therefore pull water from cerebral tissue across an osmotic gradient into plasma. Long-term therapy with osmotic agents may not be as efficacious as use on an acute basis. Equilibrium with cerebral tissue is eventually established with these drugs.

Mannitol reduces intracranial pressure and increases cerebral perfusion pressure and cerebral blood flow approximately 10–20 min following infusion. Recommended doses are 0.5–1 g/kg, with higher doses effective for up to 6 hr. Diuretics have also been employed to produce an osmotic diuresis. They function by decreasing formation of cerebrospinal fluid, and through the removal of sodium and water from brain tissue. Renal loop diuretics, such as furosemide and ethacrynic acid, are usually employed in conjunction with osmotic agents when a single agent does not sufficiently reduce intracranial pressure.

There is conflicting evidence concerning the use of steroids to reduce elevated intracranial pressure. Steroids are efficacious in reducing edema from brain neoplasms and in the treatment of such conditions as pseudotumor cerebri. In addition, a relatively small dose should be employed, such as 4 mg dexamethasone per day, since efficacy is not lost but toxicity is minimized. There is no advantage to high-dose dexamethasone in clinical outcome in severe head injury, and corticosteroids have not proved beneficial in the treatment of global brain ischemia. In addition, steroids elevate serum glucose and may worsen the effects of cerebral ischemia.

Barbiturate administration can reduce cerebral blood flow and cerebral metabolic requirements. However, the efficacy of barbiturates in reducing elevated intracranial pressure remains controversial. In addition, prophylactic use of pentobarbital following severe head injury does not decrease the incidence of elevated intracranial pressure or the duration of intracranial pressure hypertension. The most notable complication of barbiturate administration is hypotension. Barbiturates may also yield dysrhythmias, myocardial depression, and impaired uptake or release of glutamate.

Recent clinical studies suggest the use of decompressive craniectomy during brain herniation. Decompressive craniectomy is recommended for the specific indications of increased intracranial pressure such as those following severe head injury. Surgery is not indicated for cases that involve significant primary lesions, cerebral ischemia, or central herniation.

E. Clinical Prognosis

In cases of cerebral herniation complicated by ischemia, patients may suffer permanent neurologic sequelae if coma duration is 6 hr or longer. In addition, individuals with absent pupillary light reflexes never appear to regain independent daily function, but the early onset of incomprehensible speech, orienting spontaneous eye movements, or the ability to follow commands are considered indicative of a good outcome. Recent studies confirm the prognostic value of the neurologic examination that focuses on pupillary and motor function.

Measurement of intracranial pressure has also been linked with prognosis. Most patients with a maximum intracranial pressure increase of less than 30 mmHg experience good recovery. In contrast, a pressure increase more than 30 mmHg, and in some cases an intracranial pressure increase more than 20 mmHg, represents a great risk for the brain. Others have noted that in a limited number of cases aggressive treatment based on intracranial pressure monitoring can reduce mortality when applied to patients in coma secondary to cerebral infarction.

IV. PERSISTENT VEGETATIVE STATE

A. Historical Background

The term ''vegetative'' refers to a passive or involuntary existence with limited cerebral activity. In 1972, Jennet and Plum described the ''vegetative state'' as a chronic condition following diffuse brain injury that resulted in the absence of cognitive function but the persistence of sleep–wake cycles. Individuals could open their eyes to auditory stimuli and were autonomically stable with the maintenance of respiratory and hemodynamic function.

The American Neurological Association Committee on Ethical Affairs and the Quality Standards Subcommittee of the American Academy of Neurology have individually defined the vegetative state as a chronic condition that preserves the ability to maintain blood pressure, respiration, and cardiac function but not cognitive function. Specifically, the individual has no consciousness of self or the environment. The patient does not possess language function and therefore lacks any ability to communicate. Voluntary behavior or movements are absent, but facial expressions such as smiling, frowning, and crying can occur. These are not linked to any external stimulus. Sleep–wake cycles are present but do not necessarily reflect a specific circadian rhythm and are not associated with the environment. Although medullary brain stem functions remain intact to support cardiorespiratory functions, the presence of midbrain or pontine reflexes may be variable. Spinal reflex activity may also be present, but bowel and bladder function are absent.

The diagnosis of vegetative state can be made once the previously mentioned criteria are satisfied. This condition differs from the diagnosis of persistent vegetative state in both clinical and prognostic terms. The persistent vegetative state consists of the vegetative state that continues for a duration of 1 month or longer. Persistent vegetative state does not imply permanent disability since in some cases patients can partially recover from this condition.

B. Etiology

Multiple etiologies may lead to the persistent vegetative state. The end result is severe neuronal loss of the cerebral hemispheres with sparing of at least the lower brain stem. Following an anoxic insult, CT may reveal diffuse cortical swelling and progress to hypodensities in both cerebral hemispheres. In animal models of persistent vegetative state induced by transient global brain ischemia, neurodegeneration of Purkinje cells and the hippocampal CA1 pyramidal cells represents the initial phase of cerebral damage. Postmortem examinations of vegetative patients have demonstrated the loss of forebrain and neocortical structures with an intact brain stem.

Cerebrovascular accidents, such as those secondary to cardiac arrest, or the progressive loss of cortical function, such as dementia, appear to be the most common causes of the persistent vegetative state, accounting for at least 63% of the cases in one study. In younger individuals, cerebral trauma is a common cause of persistent vegetative state. Trauma patients who progress to a vegetative state have been reported to sustain more severe closed-head injury and develop swelling or shift of midline structures. Metabolic disorders such as hypoglycemia can also result in the vegetative state.

C. Clinical Presentation

Limited information on the clinical characteristics of individuals in the persistent vegetative state is available, but it is estimated that approximately 25,000

adults in the United States carry the diagnosis. In one group of nursing home patients, persistent vegetative state was diagnosed in 3% of the individuals. The age of afflicted individuals ranged from 19 to 96 years and the duration of the persistent vegetative state ranged from 1 to 16.8 years. Twenty-five percent of the patients were vegetative for more than 5 years. Other studies have indicated that all vegetative patients require at least one type of functional support, such as respirators, gastrotomy tubes, or intravenous hydration. The average cost of care can be as high as $170,000 per patient.

Individuals in the vegetative state require a multidisciplinary approach to patient care. Their course is frequented by extended hospital stays and multiple complications, such as pneumonia, decubiti, urinary incontinence, and urinary infection. In addition, patients may experience disordered neuroendocrine dysfunction with elevated profiles of growth hormone, prolactin, luteinizing hormone, and cortisol.

As with any neurologic disorder, the patient evaluation should consist of a thorough general physical and neurologic examination. The general physical examination is important for the initial assessment of a patient in the vegetative state. The presence of scalp lacerations, hemotympanum, otorrhea, and rhinorrhea may suggest recent trauma and a treatable condition such as a subdural hematoma. Bedridden patients may also suffer compartment syndromes and compression neuropathies that require decompression. In addition, pressure sores are common in this population of patients.

Eye opening, language function, and purposive movements in response to noxious stimulation should be determined to assess the level of consciousness. The neuroophthalmologic examination evaluates pupillary size and response to light, spontaneous eye movements, and oculocephalic (doll's eyes) and oculovestibular (ice water caloric) responses. In some vegetative patients, spontaneous nystagmus during wakefulness or REM sleep may be present. The respiratory pattern may supply further information concerning the level of the insult. Increased intracranial pressure may be associated with Cheyne–Stokes respiration. In contrast, midbrain pontine lesions may result in central neurogenic hyperventilation, and low pontine lesions can cause apneustic breathing. A common issue concerning the persistent vegetative state focuses on whether patients are able to perceive pain. Careful clinical examination and postmortem studies do not support the concept that these patients experience pain or suffering.

Misdiagnosis of the persistent vegetative state is not uncommon but unfortunately may severely affect patient care. Some individuals initially described as having persistent vegetative state do not meet the strict diagnostic criteria of this syndrome. Patients have been found to be interactive with their environment, to possess alternative diagnoses such as multiple sclerosis, or to be in coma rather than in the persistent vegetative state.

One factor that appears to result in misdiagnosis is the confusion in terminology used to describe vegetative patients. The terms "apallic syndrome," "akinetic mutism," "coma vigil," "alpha coma", and "neocortical death" should not be used to depict individuals in the persistent vegetative state. In addition, the neurologic conditions of coma, locked-in syndrome, and brain death are not synonymous with the persistent vegetative state.

Unlike individuals in the persistent vegetative state, locked-in patients are conscious but lack the ability to communicate with their environment. Loss of motor control of their facial musculature and extremities prevents these individuals from speaking and interacting with their surroundings. Occasionally, even ocular movements may be lost. Usually, these patients are not aphasic and are able to comprehend language. In contrast, patients in the persistent vegetative state are not conscious and cannot communicate with their surroundings.

Brain death is declared when there is loss of function of the entire cortex and brain stem. A brain-dead individual has no response to noxious stimulation. The pupils are fixed to light stimulation. Corneal blink reflex, doll's eyes, and cold water caloric responses are not present. Brain stem reflexes are absent, with loss of spontaneous respiration. The documentation of apnea is vital to the determination of brain death. The diagnosis of brain death requires the death of the brain stem, whereas the persistent vegetative state preserves brain stem function.

D. Clinical Evaluation and Management

Cranial imaging studies, such as CT or MRI, can assist in determining the etiology of the persistent vegetative state. Approximately 48 hr following a prolonged anoxic episode, hypodensities in the cerebral and cerebellar cortices and in the caudate and lenticular nuclei can occur. Days to weeks later, focal infarcts, edema, and atrophy may be evident.

PET and SPECT provide an alternative method of imaging that assesses cerebral function rather than brain anatomy. In vegetative patients, cerebral blood flow and metabolism are globally depressed. These procedures are useful in examining remaining cortical and brain stem function. In some cases, global reduction of cortical blood flow in the persistent vegetative state appears to be a consistent predictor of poor long-term outcome. Blood flow and metabolic studies can also be used as a diagnostic aide in differentiating individuals in the vegetative state from patients with the locked-in syndrome.

In patients with the persistent vegetative state, the EEG is useful in assessing cortical dysfunction and identifying the presence of occult seizure activity. The EEG may show a variety of changes. Some individuals in the persistent vegetative state have an isoelectric EEG or have periodic tracings consistent with REM sleep. If seizures develop and aggressive management is desired for the patient with the persistent vegetative state, then epileptic activity should be treated to reduce further neuronal cell loss in the cortex. Status epilepticus, status myoclonus, and myoclonic status epilepticus are associated with inability to recover consciousness in cardiac arrest survivors. In addition to issues of care, the EEG has been used to assess prognosis. In patients with terminal coma or persistent vegetative state, onset of abnormal EEG changes may be suggestive of a poor outcome.

Evoked potentials can provide information concerning the functional state of the cerebral cortex and brain stem. As previously described, patients with bilateral absence of cortical responses experience mortality rates between 73 and 98%. Brain stem auditory evoked potentials correlate with brain stem function. Simultaneous latency increase of all components can be consistent with progressive ischemia of the posterior fossa and a decrease in cerebral perfusion pressure. Loss of the brain stem auditory evoked potentials is usually suggestive of an incumbent deterioration and death. Although brain stem auditory evoked potentials are not usually modified by exogenous factors, brain stem auditory evoked potentials can be falsely altered by hypothermia, and combined use of anesthetics and barbiturates can induce latency and eventual abolition of brain stem auditory evoked potentials.

Management of the patient in the persistent vegetative state can raise several ethical concerns. The decision to assume an aggressive course or to gradually terminate care must be made after careful consideration of the patient's prior documented wishes. If the patient has not provided advance directive, then the course for further management should be decided between the patient's physician and health care surrogate (guardian, spouse, children, and family members). Decisions for care should focus on the utility of future treatment modalities, quality of life, and possible resource constraints.

Several therapeutic modalities have been proposed for the treatment of patients in persistent vegetative state. Sensory and electrical stimulation is a method that attempts to "increase arousal" in vegetative patients. Deep brain stimulation of the mesencephalic reticular formation increases cerebral evoked potentials and EEG activity, whereas cervical cord spinal stimulation can activate cerebral glucose metabolism and blood flow. Pharmacological manipulation of the central nervous system has also been attempted in several cases of persistent vegetative state. Treatment with amantadine or Sinemet has been reported to improve the level of consciousness in some patients. However, these reports are anecdotal, and prospective, controlled trials with significant numbers of patients are required to determine the true efficacy of these treatment protocols. Nutrition should also be considered a component of the therapeutic regiment. The degree of nutritional support should correlate with the level of care selected.

E. Clinical Prognosis

The ability to predict outcome for individuals in the persistent vegetative state is important for the generation of a reasonable care plan for the patient. Several variables play a role in the prognosis of patients in the persistent vegetative state. One factor that may influence outcome is the etiology of the persistent vegetative state. For example, subarachnoid hemorrhage can have the worst prospects for recovery, whereas insults resulting from hypoxia–ischemia and metabolic causes may have a better outcome.

Duration of coma and the persistent vegetative state following a hypoxic–ischemic event can also influence outcome. Although some case reports document recovery following prolonged periods in the persistent vegetative state, patients can suffer permanent neurologic sequelae if coma duration is at least 6 hr. Most individuals in the persistent vegetative state for a duration of 1 month following a hypoxic–ischemic insult do not recover.

In survivors of cardiac arrest who become vegetative, the prognosis for recovery is not significantly

correlated with age. Younger patients may demonstrate improved ability to recover motor function. However, age is not a factor in the recovery of cognition, behavior, or speech.

Prediction of outcome for traumatic-induced persistent vegetative state usually parallels the criteria used for persistent vegetative state unrelated to trauma. Persistent vegetative state of prolonged duration rarely results in recovery of independent function but does occasionally occur. Features such as young age (less than 40 years old), pupillary reactivity, and eye opening signify a favorable prognosis for individuals in the persistent vegetative state following trauma.

V. CELLULAR AND MOLECULAR MEDIATORS OF NEURONAL DISEASE

A. Pathways That Determine Neuronal Injury

Despite the immediate event, such as cardiac arrest or cerebral trauma, that may ultimately result in organic brain disease, specific cellular signal transduction pathways in the central nervous system ultimately influence the extent of neuronal injury. However, one must remember that it is multiple mechanisms, rather than a single cellular pathway, that determine neuronal survival during organic brain disease. Although neuronal injury associated with several disease entities, such as stroke, Alzheimer's disease, and Parkinson's disease, was initially believed to be irreversible, it has become increasingly evident that either acute or chronic modulation of the cellular and molecular environment within the brain can prevent or even reverse neuronal injury. Pharmacological manipulation of glutamate receptor activity and imidazole receptor binding agents have been shown to reduce the extent of ischemia within the penumbral zone. In addition, muscarinic agonists have been demonstrated to possibly influence neuronal plasticity. In order to develop rational, efficacious, and safe therapy against organic brain disease, it is vital to elucidate the cellular and molecular mechanisms that modulate neuronal injury. Some of these pathways include nitric oxide, programmed cell death, peptide growth factors, and metabotropic glutamate receptors.

B. Free Radical Injury in the Brain

Nitric oxide (NO), a free radical, has been shown to alter neuronal survival. Glutamate release during neuronal injury leads to both calcium influx into neurons and the production of NO. NO is toxic to the neuronal population under certain conditions, such as during glutamate release and during anoxia. As free radical, NO can react with superoxide to produce peroxynitrite, an agent that leads to cell lipid peroxidation. This process may account for the selective vulnerability of neurons in the hippocampus and cortex.

Experimental models have illustrated that ischemic generation of the free radical NO can be a "trigger" for the subsequent induction of neuronal injury in cortical neurons, hippocampal neurons, and mesencephalic neurons. Inhibition of NO production in cell culture systems during anoxia has been shown to be neuroprotective. During a global cerebral insult, production of NO is increased and nitric oxide synthase (NOS), the enzyme responsible for NO generation, is induced in hippocampal astrocytes. Blockade of NO formation by some NOS inhibitors can reduce infarction following middle cerebral artery occlusion and prevent cerebral endothelial cell disease such as atherosclerosis. In addition, inhibition of neuronal NO formation in transgenic animals can reduce infarction following middle cerebral artery occlusion. The cellular pathways generated by NO that result in neuronal injury are varied but include neuronal endonucleases, intracellular acidification, and cysteine proteases. As a result of its close link to the molecular pathways that lead to neuronal injury, NO functions not only as a potential therapeutic target but also as a valuable investigational agent. New clinical treatments, especially for organic brain disease, are now focusing on NO and the subsequent downstream pathways of neurodegeneration.

However, not all studies demonstrate a detrimental role for NO. Some experimental models have argued for the protective effects of NO-induced vasodilatation. Increased production of NO, in some circumstances, has been shown to decrease rather than increase cerebral stroke size in animals. Although these results have been attributed to improved cerebral perfusion to the ischemic penumbra, others have illustrated that improved cerebral perfusion alone to the ischemic zone is insufficient to sustain neuronal survival. It is unclear why certain environmental conditions may predispose NO to function as a neuroprotectant rather than a toxin. Several factors appear to contribute to these divergent observations and involve such parameters as the experimental model, external environmental conditions, duration of the insult, and age of the neuronal system.

C. Programmed Cell Death

Programmed cell death (PCD) is also believed to be one of the contributing factors to neuronal injury and is considered to be an active, directed process that can rapidly lead to the destruction of a cell. In contrast to necrosis, PCD is characterized by the preservation of membrane integrity and internal organellae structure, chromatin condensation with nuclear fragmentation, and the budding of cellular fragments known as "apoptotic bodies." In most cellular systems, the end result of PCD is termed apoptosis. The protein- or gene-mediated mechanisms of PCD form the basis of several organic brain disorders. In some clinical investigations of global cerebral ischemia, the ischemic vulnerability of cerebellar granular cells has been attributed to early apoptotic DNA fragmentation. *In vitro* studies also support a role for PCD during neurodegeneration. Glutamate toxicity in primary neuronal cultures may lead to double-strand DNA breaks and to early single-strand DNA breaks that are consistent with apoptotic neuronal death. The detailed understanding of the cellular mechanisms that modulate PCD may provide the basis for novel therapeutic strategies to prevent or reverse neuronal loss.

Neuronal and vascular PCD is believed to proceed through two dynamic but distinct pathways that involve both DNA fragmentation and the loss of membrane asymmetry with the exposure of membrane phosphatidylserine (PS) residues (Fig. 1). These processes are considered to be functionally independent determinants of neuronal PCD. The internucleosomal cleavage of genomic DNA into fragments may be a late event during PCD and ultimately commit a cell to its demise (Fig. 2). In contrast, the redistribution of membrane PS residues can be an early event during PCD that usually precedes DNA fragmentation and may serve to later "tag" injured cells for phagocytosis (Fig. 3).

However, one of the current central issues surrounding PCD focuses on whether this process, once initiated, is committed in nature to lead to cellular death or is reversible to the extent of preventing further neuronal injury. Experimental studies have illustrated that at least the onset of PCD can be significantly limited during therapeutic modulation, such as during neuronal ischemia, excitotoxicity, and free radical exposure. In addition, these models have demonstrated a prominent link between the generation of the free radical NO and the induction of neuronal PCD during brain injury since NO can result in the rapid induction of PCD.

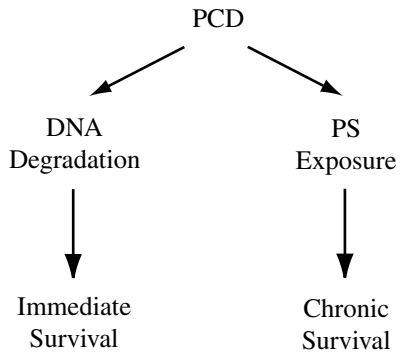


Figure 1 Programmed cell death (PCD) consists of two independent pathways. The potential sequence of events following cellular PCD induction is illustrated. Neuronal and vascular PCD is believed to proceed through two distinct pathways that involve both DNA degradation and the loss of membrane asymmetry with the exposure of membrane phosphatidylserine (PS) residues. The internucleosomal cleavage of genomic DNA into fragments may be a late event during PCD and ultimately commit a cell to its demise. In contrast, the redistribution of membrane PS residues can be an early event during PCD that usually precedes DNA fragmentation and may serve to later identify injured cells for phagocytosis.

Current techniques employed to assess PCD following NO exposure, such as terminal deoxy-UTP nick end labeling (Fig. 2) or transmission electron microscopy, are useful for identifying the extent of PCD induction. However, these procedures lack the ability to assess dynamic changes in PCD in individual cells. As an alternative, a recently developed technique can monitor the induction and change in PCD in individual living cells over a period of time. The method employs the reversible labeling of annexin V to exposed PS residues of cells undergoing PCD, an early primary event during PCD induction (Fig. 3). By exploiting the dependence of annexin V on cellular calcium to bind to exposed membrane PS residues, one can reversibly label individual neurons over time. During the induction of PCD, such as following NO exposure, progressive externalization of membrane PS residues occurs that is independent of the loss of neuronal membrane integrity.

Given the ability to continually monitor changes in the course of PCD, it is now conceivable to assess whether neuronal brain injury during PCD is, in fact, reversible in nature. Studies employing neuroprotective regiments, such as the application of either trophic factors or metabotropic glutamate receptor agonists, have provided evidence supporting the concept of reversible injury during PCD. For example, since cytosolic and nuclear changes associated with PCD are evident within the first hour of NO exposure, the signal

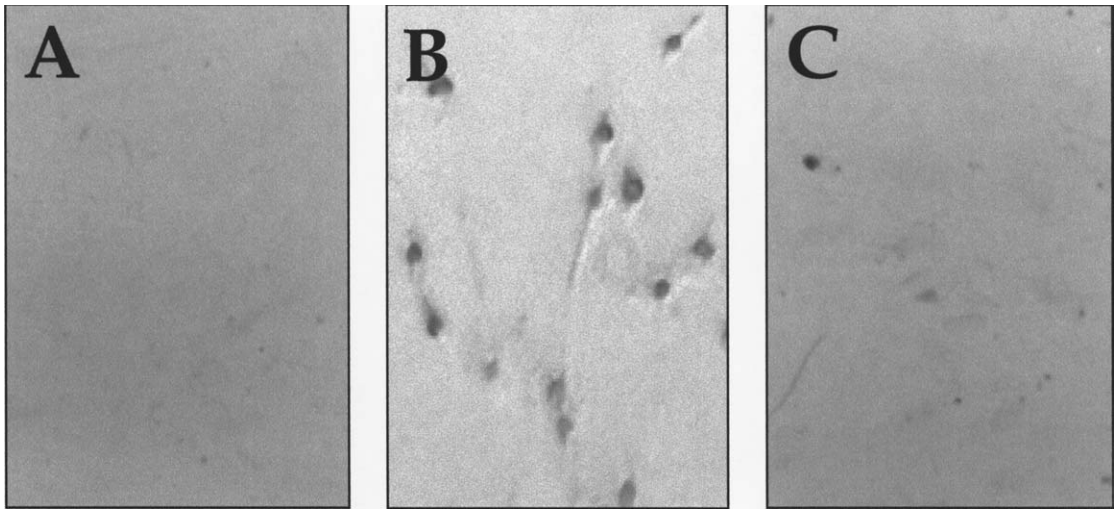


Figure 2 Prevention of neuronal DNA fragmentation following NO exposure. A representative figure is shown of neuronal DNA fragmentation that was assessed using the terminal deoxy-UTP nick end labeling (TUNEL) assay. (A) Untreated neurons (control) are without significant TUNEL labelling. (B) Neurons were exposed to NO and DNA fragmentation was assessed 24 hr later. In the presence of NO, DNA fragmentation was present in approximately 71% of the neurons examined. (C) Treatment with the growth factor, basic fibroblast growth factor, significantly prevented neuronal injury and genomic DNA fragmentation induced by NO exposure.

transduction mechanisms of the metabotropic glutamate system may reverse early steps in PCD.

The ability to prevent genomic DNA degradation and maintain membrane PS asymmetry may be closely linked to the modulation of cysteine protease activity. NO is believed to be one of the signal transduction systems that can elicit cysteine protease activity and can directly stimulate caspase 1- and caspase 3-like activities. Caspase 1 has been linked to the modulation of membrane PS residues through cytoskeletal proteins such as fodrin. Caspase 3 can lead to the direct degradation of DNA through the enhancement of DNase activity. Application of some neuroprotective agents, such as trophic factors or metabotropic glutamate agonists, can directly prevent the activation of caspase 1- and caspase 3-like activities following NO exposure, suggesting that these agents may maintain both genomic DNA integrity and membrane PS asymmetry through the modulation of cysteine protease activity. In addition to directly downregulating cysteine protease activity, specific neuronal enzymes responsible for the destruction of genomic DNA and

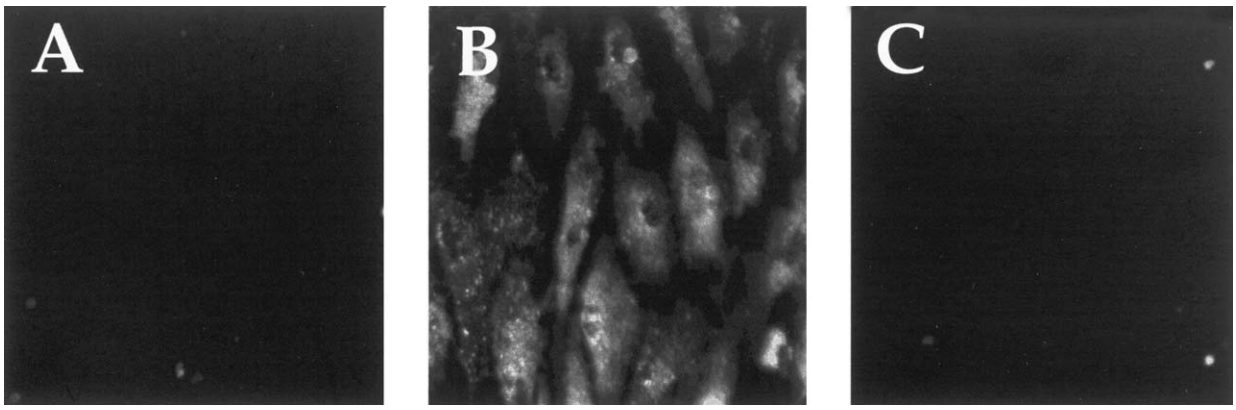


Figure 3 Prevention of endothelial cell NO-induced phosphatidylserine (PS) externalization. A representative figure is shown of cerebral endothelial cells that were identified for PS externalization following exposure to NO with fluorescent microscopy. (A) Untreated endothelial cells (control) are without significant PS exposure. (B) Endothelial cells were exposed to NO and PS exposure was assessed 24 hr later. In the presence of NO, PS was present in approximately 80% of the endothelial cells examined. (C) Treatment with the metabotropic glutamate receptor agonist DHPG significantly prevented endothelial cell injury and cellular membrane PS exposure induced by NO exposure.

their cation modulators have been identified that may contribute to acute neuronal injury. As an alternative mechanism, free radical regulation of intracellular pH can also influence PCD induction. It is hoped that further knowledge of these pathways will assist in the development of therapeutic regiments against organic brain disease.

D. Future Directions for the Treatment of Organic Brain Disease

In terms of preventing or reversing neuronal injury, peptide growth factors are increasingly being investigated as therapeutic regiments. Although currently the most efficacious application of growth factors involves the peripheral nervous system, progress is being made with central nervous system injury during both growth factor application and the generation of neuronal progenitor cells. The mechanisms employed by trophic factors to achieve neuroprotection can be diverse and not well understood. However, peptide growth factors have been shown to prevent neurodegeneration in hippocampal cultures during glutamate toxicity, potassium cyanide administration, hypoglycemia, and NO toxicity. In addition, the neuroprotective effects of peptide growth factors have been intimately linked to the modulation of the signal transduction systems of NO and protein kinase C.

Recently, several cloned metabotropic receptor subtypes have also been linked to the modulation of neuronal survival. They function through signal transduction pathways such as cAMP, protein kinase C, inositol phosphate, ion channel flux, and phospholipase D. Activation of the metabotropic receptors can reduce glutamate toxicity, protect synaptic transmission during periods of hypoxia, and increase neuronal survival during NO exposure. Current investigations have elucidated the role of the metabotropic glutamate system during cysteine protease activation, neuronal endonuclease and pH modulation, and intracellular calcium flux.

With the proper therapeutic intervention, organic brain disease may not only be preventable but also be reversible. The ability to effectively translate therapeutic intervention into the clinical spectrum requires an initial understanding of the predominant cellular mechanisms that may mediate neuronal injury during stroke. In this regard, future work that is directed at investigating the character of the injury, such as necrotic versus apoptotic disease and neuronal versus vascular injury, and the underlying cellular and molecular pathways that can contribute to the injury, such as the activity of specific cysteine proteases, is necessary in order to formulate a better understanding of the ability of a particular neuronal insult to influence both clinical plasticity and functional outcome in the nervous system. Thus, neuroprotective agents should be viewed as possessing two distinct utilities that function not only as agents to prevent or reverse clinical neuronal injury but also as investigational tools to elucidate the cellular pathways that modulate subsequent neuronal function. However, in all aspects, these separate functions for a neuroprotective agent should be considered parallel in nature in order to procure the foundation for the development of more safe and efficacious future neuroprotective strategies for the treatment of organic brain disease.

See Also the Following Articles

BRAIN DAMAGE, RECOVERY FROM • BRAIN LESIONS • CEREBROVASCULAR DISEASE

Suggested Reading

American Neurological Association Committee on Ethical Affairs (1993). Persistent vegetative state: Report of the American Neurological Association Committee on Ethical Affairs. *Ann. Neurol.* **33,** 392–396.

Chen, R., and Young, G. B. (1996). Metabolic encephalopathies. *Baillieres Clin. Neurol.* **5**(3), 577–598.

Edgren, E., Hedstrand, U., Kelsey, S., Sutton-Tyrrell, K., and Safar, P. (1994). Assessment of neurological prognosis in comatose survivors of cardiac arrest, BRCT I Study Group. *Lancet* **343**(8905), 1055–1059.

Fisher, C. M. (1995). Brain herniation: A revision of classical concepts. *Can. J. Neurol. Sci.* **22**(2), 83–91.

Kerr, J. F., Wyllie, A. H., and Currie, A. R. (1972). Apoptosis: A basic biological phenomenon with wide-ranging implications in tissue kinetics. *Br. J. Cancer* **26**(4), 239–257.

Levy, D. E., Caronna, J. J., Singer, B. H., Lapinski, R. H., Frydman, H., and Plum, F. (1985). Predicting outcome from hypoxic–ischemic coma. *J. Am. Med. Assoc.* **253**(10), 1420–1406.

Lin, S.-H., Vincent, A., Shaw, T., Maynard, K. I., and Maiese, K. (2000). Prevention of nitric oxide-induced neuronal injury through the modulation of independent pathways of programmed cell death. *J. Cereb. Blood Flow Metab.* **20**(9), 1380–1391.

Maiese, K., and Caronna, J. J. (1995). Coma. In *Critical Care Medicine: Principles of Diagnosis and Management* (J. E. Parrillo and R. C. Bone, Eds.), pp. 1157–1176. Mosby, Philadelphia.

Plum, F., and Posner, J. B. (1982). *The Diagnosis of Stupor and Coma*. 3rd ed. Davis, Philadelphia.

Shihabuddin, L. S., Palmer, T. D., and Gage, F. H. (1999). The search for neural progenitor cells: Prospects for the therapy of neurodegenerative disease. *Mol. Med. Today* **11,** 474–480.

Brain Lesions

MATTHIAS KEIDEL and PHILIPP STUDE*
*District Hospital of Bayreuth and *University of Essen*

GLOSSARY

astrocyte Control function of the neuronal environment influencing local electrolyte and neurotransmitter concentrations.

contusion Necrotic cortex and white matter with variable quantities of petechial hemorrhages and edema due to head trauma.

Creutzfeld–Jacob disease Transmissible spongiform encephalopathy leading to a widespread cortical atrophy with the prion (proteinaceous infectious particle) as the responsible agent.

cytotoxic edema Accumulation of excess intracellular water.

encephalitis Brain infection commonly due to bacteria, viruses, fungi, or protozoa.

epidural hematoma Collection of blood between the skull and the dura.

focal ischemia Pattern of damage caused by a stroke resulting from a brain infarct.

head trauma Damage of the brain and/or skull due to a trauma.

morbus Alzheimer Most common degenerative brain disease causing a presenile or senile dementia.

multiple sclerosis Demyelinating disorder affecting the central nervous system with predominant white matter lesions.

oligodendrocyte Myelin-forming cell of the central nervous system.

subdural hematoma Collection of blood between the dura and the underlying brain frequently a result of torn veins draining into the dura or dural sinuses.

A lesion is the structural or functional correlate of a disease or illness to which clinical signs and symptoms can be attributed. Specific lesions are related to distinct groups of diseases or types of injuries. In general, a brain lesion is understood as a structural one. However, a lesion may be purely biochemical or systemic; thus, not all diseases have overtly visible lesions associated with them, despite profound consequences for the patient.

I. INTRODUCTION

In this article, we present an overview of possible brain lesions. We focus on commonly occurring or controversially lesions of clinical relevance (Table I). Emphasis is on the correlation between pathogenic aspects, on the one hand, and clinically as well as obvious and diagnosable lesions, on the other hand. Neuropathological and pathophysiological aspects of brain lesions and relevant repair mechanisms are considered.

II. GENERAL MECHANISMS RELATED TO BRAIN LESIONS

A. Structure, Function, and Reaction of Central Nervous System Cells to Injury

1. Neuron

Neurons and their networks of axons, dendrites, and synaptic contacts are the basic elements for perception, conduction, and processing of information. They vary in shape and size and show different changes in

Table I
Overview of Pathogenetic and Clinical Aspects of Brain Lesions

Inflammatory	Infections	Bacterial
		Viral (HSV, HIV)
		Fungal
		Parasitic
		Prions (CJD)
		Immunosuppressed patients
	Immunologic	Multiple sclerosis
Vascular	Cerebral ischemia	
	Intracerebral haemorrhage	
	Vascular malformations	
	Vaskulitis	
Traumatic	Traumatic hematomas	Epidural, subdural, subarachnoidal hematomas
	Parenchymal lesions	Concussion
		Contusion
		Intracerebral hemorrhages
Degenerative	Cerebral cortex	Alzheimer's disease
		Pick's disease
	Basal ganglia and brain stem	Parkinsonism
		Huntington's disease
	Spinocerebellar	Friedreich's ataxia
		Olivopontocerebellar atrophy
	Motor neurons	Motor neuron disease
Neoplastic	Gliomas	
	Tumors of neuronal origin	
	Tumors of primitive or undifferentiated cells	
	Tumors of meninges	
	Lymphomas	
	Metastatic tumors	
Nutritional/metabolic/toxic	Alcohol-related disorders	Wernicke–Korsakoff syndrome
		Central pontine myelinolysis
		Cerebellar atrophy
		Cortical atrophy
	Vitamin B-group deficiency	Wernicke–Korsakoff syndrome
		Subacute combined degeneration of the spinal cord
	Hepatic encephalopathy	Wilson's disease
		Hepatoportal encephalopathy
	Leukodystrophies	Metachromatic leukodystrophy
		Adrenoleukodystrophy
		Krabbe's disease
	Mitochondrial encephalopathy	MELAS, MERRF syndrome

diseases associated with specific etiologic agents or pathologic processes. A noxious condition that affects the nervous tissue results in biochemical and subsequent morphologic alterations. In most cases, no pathologic process is limited to a single cellular element and the subsequent effects result in a constellation of cellular reactions. For our purposes, however, it is convenient to evaluate seperately the changes in neurons, glia, connective tissue, and vascular structures. Some of the more important or commonly encountered ones are listed in Table II.

2. Glia

Astrocytes mediate between the mesenchymal and the neuronal compartment. They surround the vessels, provide a border to the brain at the meninges, and surround the extracellular space of the brain as a supportive framework for other cells in the central nervous system (CNS). Their main function is the control of the neuronal environment influencing local electrolyte (especially potassium) and neurotransmitter concentrations. Astrocytes show changes in response to almost every type of injury or disease in the CNS. Their reactions may take the form of degeneration, hypertrophy, or hyperplasia, although they are less vulnerable to hypoxia than are neurons. Loss of neurons from any cause, breakdown of myelin, or injury to nerve fibers are followed by an increase in the number of astrocytes and glial fibers; this process is called gliosis, which is composed entirely of cellular processes.

Oligodendrocytes are the myelin-forming cells of the CNS. In white matter they are frequently lined up along the myelinated fibers. In gray matter they tend to cluster around the neurons, where they are called satellite cells. Since they are responsible for synthesis and maintenance of myelin, they are affected in disorders that depend on myelin and myelination, mainly the leukodystrophies and demyelinating diseases (e.g., multiple sclerosis). The reaction to injury is limited: The cell turnover is very slow and the presence of axons is necessary for their proliferation.

Microglial cells, which are brain-intrinsic cells, have phagocytic properties and contribute to the CNS's macrophage response to injury. The conditions under which the microglial reaction is found cover a range from mild proliferation around chromatolytic neurons to the removal of degenerating terminals and the invasion of totally necrotic brain as in an infarct. Since they express major histocompatible antigens, microglial cells also act as antigen presenting cells.

Table II
Categories and Types of Neuronal Reactions to Noxious Conditions

Category	Type
Acute necrosis	Ischemic nerve cell change
	Hypoglycemia
Atrophy and degeneration	In many system degenerations without specific change
Abnormal accumulations	Neurofibrillary tangles (M. Alzheimer)
	Argentophilic inclusion (M. Pick)
	Lewy body (M. Parkinson)
	Inclusion bodies of viral diseases (cytoplasmic/nuclear)
	Lipofuscin excess
	Abnormal accumulation of storage diseases
Chromatolysis	Axonal damage

3. Connective Tissue

Connective tissue encloses the meninges, including the dura, arachnoid, and the pia, covering the complete surface of the brain. The properties of separate compartments (subarachnoidal, subdural, and epidural) result in different effects of lesions, even in cases in which they are also the origin and target of disorders (e.g., infections and tumors). Extradural lesions tend to be localized, whereas subarachnoidal lesions (e.g., subarachnoidal bleeding) can expand widely over the surface of the whole brain. An important function of the mesenchyme is the resorption of the cerebrospinal fluid (CSF) into the venous sinus via arachnoid granulation. Interruption of resorption at these sites due to obliteration after inflammation or hemorrhage leads to an accumulation of CSF with the potential risk of increased intracranial pressure and hydrocephalus.

4. Blood Vessels

The blood vessels in the CNS are similar in structure and function to those elsewhere in the body, with the important exception of the capillaries. The capillaries within the CNS differ from most other capillaries because they are not fenestrated. Tight junctions are

present between adjacent cells and the endothelial cell basement membrane is intimately surrounded by a network of astrocytic processes. These special structural features are important constituents of the blood–brain barrier. In general, an inflammation of the CNS (e.g., encephalitis) and/or the meninges (e.g., meningitis) is associated with a disturbance of the blood–brain barrier. Such malfunction of the blood–brain barrier enables the visualization of brain lesions by contrast enhanced functional imaging. Congenital or acquired malformations of the cerebral blood vessels (e.g., angioma, AV fistula, and cavernoma), inflammation (e.g., viral, bacterial, autoimmunological, radiogenic, granulomatous, sarcoidous, and angiitis), structural anomalies (congophilic angiopathy), venous or arterial obstruction (sinus thrombosis and thrombotic or embolic arterial occlusion), or disturbed autoregulation of the diameter of the small brain vessels (e.g., microangiopathy) can cause a variety of brain lesions, such as lacunar or territorial ischemia, hematoma, edema, or calcifications.

B. Common Pathophysiologic Complications

The enclosure of the brain and its coverings in the rigid box of the skull limit the potential for alterations of its volume. This renders the brain susceptible to interrelated pathophysiologic complications that occur in many different pathologic processes.

Any space-occupying lesion that is able to increase the volume of the intracranial contents is also able to produce increased intracranial pressure—for example, hemorrhage, infarct, tumor, cerebral abscess, cerebral diffuse or focal edema, and hydrocephalus.

The edema is either a correlate of the primary brain lesion, such as diffuse swelling due to cerebral hypoxia or due to severe head trauma, or it develops as a secondary perifocal edema surrounding focal brain lesions (e.g., due to hematoma, tumor, abscess, or filiae). The underlying mechanisms of cerebral edema can be divided into a cytotoxic and a vasogenic component. The most common form is the vasogenic edema. This results in an accumulation of a protein containing filtrate of plasma in the extracellular space that settles either on damaged capillaries that have lost their barrier function (infarcts and contusions) or on newly formed capillaries that have not yet established a barrier (primary or metastatic tumors). Predominantly the white matter is affected.

Cytotoxic edema is the accumulation of excess intracellular water that occurs in processes resulting in a breakdown of cerebral energy metabolism. Since the sodium/potassium pumps are altered in their function, this leads to a passive influx of sodium and water into the cells. This type of edema is more pronounced in the gray matter than in the white matter.

An internal hydrocephalus often occurs as a consequence of a focal brain edema with compression of the aqueduct (e.g., due to a cerebellar swelling caused by a ''malignant'' cerebellar infarction) or as a consequence of a blockage of the foramen of Monroi (e.g., due to compression of the third ventricle caused by a hemispherical swelling with a ''midline shift'').

The interaction among the three processes of primary brain lesions, the (focal or diffuse) edema, and the potential development of a secondary hydrocephalus is one of the most frequently encountered and potentially dangerous complications of CNS afflictions.

Depending on the size of the lesion and the acuity of expansion, there are initially three ways to compensate a potential increase in intracranial pressure. The volumes of CSF are reduced by enlarged resorption and reduced production, and the volumes of blood are reduced by vasoconstriction. These compensatory mechanisms allow the brain to expand without serious effects. Pressure-induced atrophy occurs most commonly with slow-growing extrinsic lesions.

After this phase, a critical period occurs in which a further increase in the intracranial contents will cause an abrupt increase in intracranial pressure. Further expansions lead to a shift of intracranial tissue, evident in the shift of midline structures, which is followed by herniation. Depending on the site of the space-occupying lesion, herniations occur mainly at characteristic sites (transtentorial, cerebellar tonsillar, and below the falx cerebri). Since vital respiratory centers in the medulla are compressed, transtentorial herniations are a frequent cause of death.

The impact of brain lesions varies depending on different degrees of size, site, subcortical extent, and temporal evolution. Apart from the functional deficits within the lesion area, local brain lesions involve corticocortical connections as well as afferent and efferent cortical projections. They also affect functionally connected but structurally intact brain regions (diaschisis).

C. Compensatory Mechanisms of Repair

In this section, we describe the processes involved in the restitution of function (requiring tissue survival) or

of functional reorganization (in permanent brain damage) following brain lesions. These become obvious as the functional recovery mechanisms of relearning, within-system recovery, substitution by related systems, or assumption of new strategies take over.

In several experiments, increased and prolonged field potentials due to an increased NMDA/GABA ratio were found in the surrounding areas and these spread widely throughout areas connected to a focal brain lesion. It has been suggested that this hyperexcitability supports synaptic plasticity and reorganization by facilitating heterosynaptic long-term potentation (LTP), a mechanism also involved in learning.

An increasing number of neurotrophic factors have been found to be able to promote neuronal survival (protection against the late retrograde death of distant neurons) and sprouting, synapse formation (stabilization of LTP), glial proliferation and growth of new blood vessels. Reorganization of synaptic networks due to sprouting of new pre- and postsynaptic endings has been demonstrated in cases of regenerative axonal long-distance extension after injury. Even the possibility of induction of new neurons from precursor cells from the temporal lobe subventricular zone has been shown in *in vitro* experiments.

In studies of motor cortex lesions, a lesion-induced reduced reciprocal inhibition of pyramidal neurons that form interconnected patterns has been considered as a potential mechanism to increase the coupling strength between pyramidal cells in order to modify motor cortical output. Various studies concerning cortical neuronal plasticity demonstrated an enlargement of areas representing specific cortical functions directly adjacent to the lesion site.

D. Relation between the Site of Lesion and Functional Consequences

Even though a local lesion causes neurological symptoms that can be attributed both to the site of lesion and to disturbances of network-encoded functions, there are symptoms that implicate a topical classification of the underlying lesion. In many cases, the etiology plays only a secondary role because the combination of symptoms (e.g., caused by a stroke or a tumor of frontal or central areas) can be quite similar apart from their temporal course. An overview of the correlation between the site of the lesion and the related dysfunctions is provided in Table III.

Table III
Topistic Relation between the Brain Lesion Site and the Correlated Functional Deficit

Site of lesion	Functional deficit
Frontal	Personality change (alteration in drive and affectivity, e.g., disinhibition or decreasing concern of family or business affairs, intellectual impairment, memory defects)
	Broca aphasia (dominant hemisphere)
	Epilepsy (focal motor, adversive, status epilepticus)
	Contralateral weakness/hemiparesis
	Loss of micturition control
	Optic atrophy
	Loss of smell
Parietal	Amnestic aphasia, dyslexia (dominant hemisphere)
	Neglect, geographical confusion, apraxia (nondominant hemisphere)
	Epilepsy (sensory seizures)
	Sensomotory hemiparesis (mainly sensory)
	Hemianopia
Temporal	Personality changes (subtle: depressive, irritable, anxious)
	Epilepsy (temporal lobe epilepsy)
	Upper quadrantic hemianopia
	Wernicke's aphasia
Occipital	Dyslexia, visual agnosia
	Epilepsy (seizures with visual aura)
	Homonymous field defects
Basal ganglia	Lack of drive
	Akinetic Parkinsonism
	Hemiparesis
Cerebellum	Limb ataxia, truncal ataxia, intention tremor, optokinetic symptoms
	Signs of raised intracranial pressure
Brain stem	Signs of cranial nerve disturbances, possibly in combination with sensory loss and/or hemiparesis

III. TYPES OF LESIONS

A. Vascular Lesions

Disturbances of cerebral blood flow are one of the most frequently encountered causes of neurological

Table IV
Etiology, Morphology, and Consequences of Vascular Malformations

Etiology	Morphology	Consequences
Brain stem aneurysms (1–2% of the population)	Abnormal, focal or segmental extension of the brain stem arteries in cases of congenital weakness of the blood vessel walls at section with a high hemodyamic load, resulting in sack- or berry-shaped distensions of the walls, especially in ACA, and exit of ACP from ACI and ACM	Rupture, frequently seen as a potentially letal event resulting in SAB, ICB, vasospasms, and residual bleeding
Arteriovenous angioma	Birth defects with the formation of dysplastic arterial and venous vessels as a result of a lack of differential development of the blood vessel plexus in the embryo into a capillary network with persisting arteriovenous short-cuts. This results in convolutions of the blood vessels of various sizes in the cerebral hemispheres. Regressive changes may be found in the blood vessel walls with fibers, calcification, and signs of thrombosis.	Widespread intracerebral bleeding, bleeding in the ventricles, and infarct as a consequence of steal effect accompanying high shunt volumes
Cavernous angioma	Vessels show sinusoidal convoluted distensions of variable size and a lack of the typical layering of the structure of the walls. Frequently clinically apparent are small sources of bleeding surrounded by gliosis.	Residual small sources of bleeding but frequently free of obvious clinical symptoms

illness. The sudden appearance of paralysis to one side of the body is clinically described as a stroke. This term does not immediately imply that the pathology underlying the symptomatology is pathogenetic due to an ischemic lesion or is directly associated with an intracerebral bleeding. Vascular lesions may occur as a result of an occlusion or stenosis in the branches or network of the arteries leading from the origin of the aorta from the left ventricle to the affected region. They can also reflect a chronic or subacute arterial hypertension (e.g., arteriosclerosis, status lacunaris, and hypertonic encephalopathy) or derive from defects in the development of the blood vessels (basal aneurysms, arteriovenous angioma, and cavernous angioma; Table IV).

B. Cerebral Ischemia

In clinical practice, two general types of acute cerebral ischemia can be distinguished: a focal and a global cerebral ischemia. Global cerebral ischemia is a result of cardiac arrest (i.e., following a collapse of the circulation) and leads to diffuse hypoxic brain damage. This is caused by various mechanisms that will not be discussed here. Disturbances in a supply area of a cerebral artery lead to focal perfusion deficits followed by an abrupt or ictal onset of focal or global neurologic symptoms, which can be permanent or reversible.

Clinically, a variety of symptoms may become overt that can be attributed to particular anatomical areas affected by the disrupted blood flow. The cardinal symptom arising from a stroke is a hemisparesis. In conjunction with additional neuropsychological deficits, the affected hemisphere can be identified; if associated with cranial nerve dysfunction, crossed symptoms, or an initial lack of consciousness, then a vertebrobasilar locus can be diagnosed. Lacunar infarcts often lead to monosymptomatic, minor strokes.

Focal ischemia is the term applied to the pattern of damage caused by a stroke resulting from a brain infarct (Fig. 1). Here, it is important to distinguish between the seriously affected *core* of damage and the less severely damaged *penumbra* on the borders. Even within the core of the affected area, a small degree of blood flow remains (<10 ml/100 g/min). This is not enough to supply the tissue with the necessary minimum of oxygen and glucose. In turn, this means that the cells will become necrotic within 1 hr. The ischemic peripheral penumbra is defined by a rate of

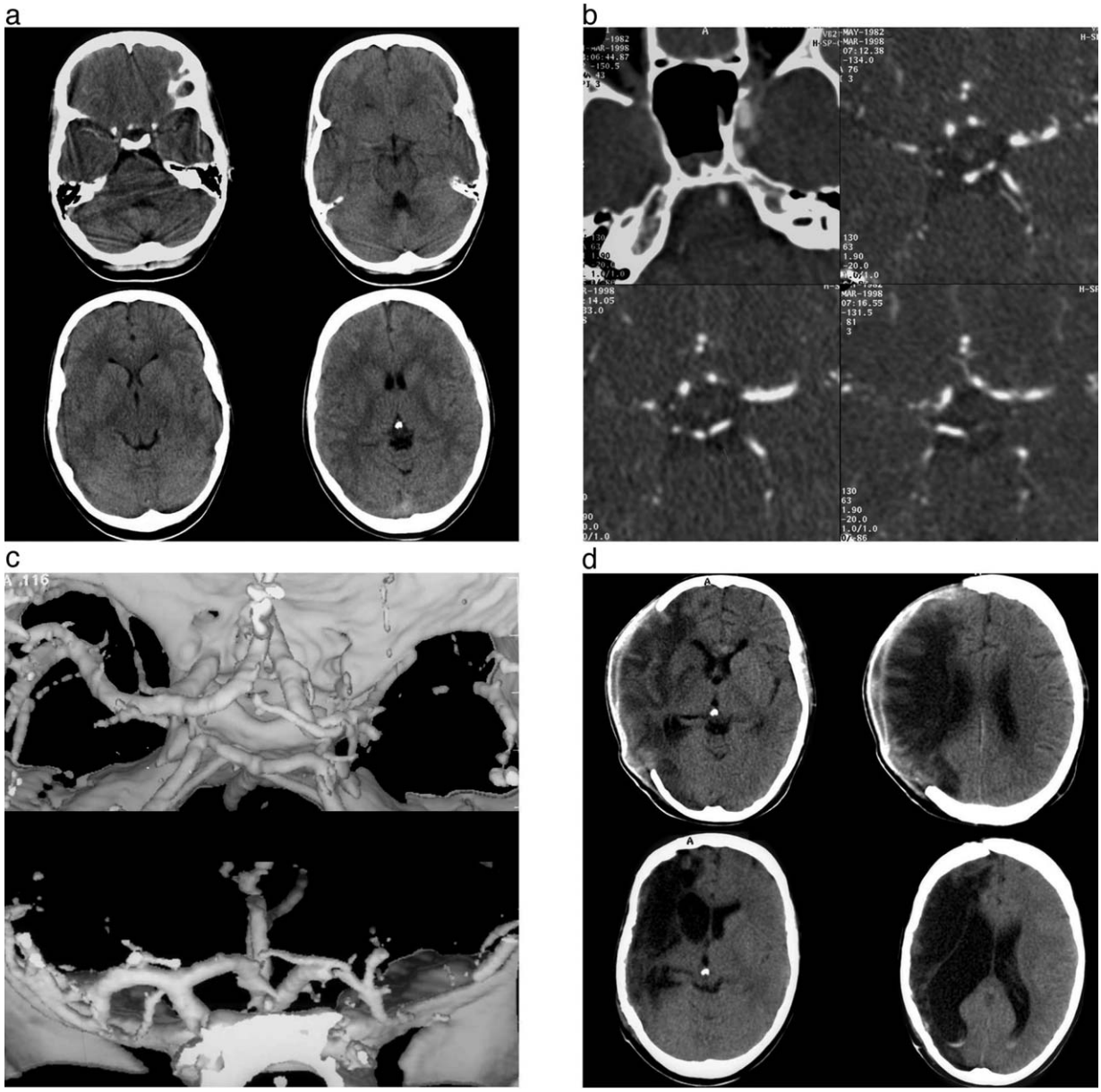


Figure 1 CT scan of a 65-year-old woman with an acute stroke and a left sided hemiparesis: (a) initial CT with (b) contrast-enhanced angio-CT (dense media sign), (c) three-dimensional reconstruction (occlusion of the ACM), and (d) after craniectomy (top) and re-implantation of the bone (bottom).

perfusion of about 10–25 ml/100 g/min. This constitutes more than the minimum required to maintain tissue metabolism in the short term. However, it is not enough to maintain neuronal function in the long term. The period during which the cells will tolerate an ischemia (i.e., those that make up the penumbra) depends on the extent of the ischemia, just as the extent of the ischemia depends on the number of surviving cells. If this critical period becomes too long or the blood flow decreases below the minimum, the consequence is an infarct. The temporal dependence of the infarct gives rise to the concept of a therapeutic window for clinical regimens (thrombolysis and drug treatment).

At the cellular level, the energy deficit abolishes the sodium/calcium exchange at the membrane and thus

leads to an increase in the intracellular sodium levels. Similarly, the energy deficit also impairs glutamate uptake inhibition that naturally leads to excessive levels of extracellular glutamate. Glutamate activates NMDA binding sites, stimulating the flow of calcium into the cell. At the same time, glutamate activates the metabotropic binding sites, leading to the outflow of calcium from intracellular stores. The resultant major increase in intracellular calcium is a central aspect of the cell destruction in ischemia. The high calcium levels activate a series of protein kinases (including protein kinase C) leading to the phosphorylation of cell proteins. Via protease activation (e.g., endonuclease), proteolysis results in the breakup of the DNA and through activation of phospholipases the breakup of lipids. This lipolysis also results in an increase in platelet activating factor and arachidonic acid, which promote the formation of free oxygen radicals. These free radicals interact with a host of related molecules, destroying these and causing more lipolysis. Reperfusion of the tissue brings a supply of oxygen that only serves to increase the formation of more radicals and can cause secondary reperfusion damage.

In the absence of oxygen, glucose is metabolized to lactate. This sets many protons free, leading to a marked acidosis that is neurotoxic and, though inducing vasodilation and reducing the density of NMDA receptors, neuroprotective.

At the level of microcirculation, protective and destructive mechanisms appear. On the one hand, vasodilation results from various neurogenic and metabolic influences. This can improve the remaining blood flow through the ischemic tissue. On the other hand, this increased perfusion is inhibited by a swelling of perivascular cells and endothelia, platelet aggregation, and leukocyte adhesion.

This juxtaposition of factors promoting repair and cell damage provides insight into why one attempts to reduce the damage resulting from a brain infarct by interfering with these circular processes.

Another route to initiate cell death following cerebral ischemia is by so-called "apoptosis." Apoptosis is the genetically regulated form of cell death. It enables the balance between growth and elimination of cells and occurs physiologically during the embryonal development or involution processes. It seems likely that all cells have evolved the capability to undergo apoptosis and that alterations in the environmental conditions can initiate, accelerate, or slow down the process. However, dysregulation of apoptosis can also result in inflammatory, malignant, autoimmune, or neurodegenerative diseases. Furthermore, infectious agents and other cell-damaging circumstances (e.g., traumatic or ischemic conditions) can lead to apoptosis.

Apoptosis takes place in four consecutive stages: stimulation, intracellular response, apoptosis, and phagocytosis. Stimulation is mediated by many different mechanisms that are either receptor mediated or directly act on the cell. Receptor-mediated apoptosis has been demonstrated for several growth factors [transforming growth factor-β and cytokines including tumor necrosis factor (TNF)] acting via a large family of receptors on the surface of target cells (TNF receptor superfamily). Non-receptor-mediated stimuli promoting apoptosis penetrate the cell directly. They include heat shock/stress factors, free radicals, ultraviolet radiation, toxins, numerous drugs, synthetic peptides, and lymphocyte enzymes. Both types of stimuli induce the intracellular response consisting of direct activation of caspases or via the mitochondrial release of cytochrome c, which is also influenced by proapoptotic and antiapoptotic mediators.

Activated caspases are cystein proteinases that form an intracellular proteolytic cascade modulating many cellular events in the apoptotic pathway, including activation of transcription factors and induction of apoptosis-related genes. In the next step of apoptosis, intracellular calcium is released accompanied by a depletion of ATP, degradation of DNA, expression of cell surface phagocyte recognition molecules, and cell dismantling. The last step consists of phagocytic recognition of apoptotic bodies and phagocytosis of the apoptotic cells by phagocytes.

The apoptotic program is not just executed in cases of cell damage; it occurs in cases of missing growth factors and inadequate stimulation of growth. Therefore, apoptosis is also a physiological safety tool eliminating cells with deregulated growth.

In therapeutic approaches the modulation of apoptotic cell death enables new ways to treat disease. Possible mechanisms are induction of apoptosis in malignant cells, prevention of apoptosis in senescence or neurodegenerative disorders, or regulation of tissue regeneration/repair by inducing apoptosis to limit fibroblast activity and scar formation.

Four principles that are important for determining therapy may be derived from the basis of the pathophysiologic precursors described previously: reperfusion, the influence of the blood clotting system, improvement of global cerebral perfusion, and neuroprotection. Substances administered to improve neuroprotection should prolong the tolerance to ischemia in the functionally disturbed neurons and glia.

Alternatively they should protect these cells against toxic metabolites (e.g., free radicals) and excitatory amino acids (excitotoxicity). To date, animal experimentation has provided numerous impressive results that unfortunately cannot be transferred to the clinical situation. Many substances are being tested in clinical trials, whereby the next step must consider the possible combination of these substances with forms of therapy designed to promote revascularization.

C. Traumatic Lesion

1. Head Trauma

The clinical picture of the effects of traumatic head injuries is defined in terms of the morphological damage inflicted on the brain and skull. Either an open or a closed head injury is diagnosed depending on the mechanism that mediated the trauma. An open head injury means that there is an open connection between the subdural and epigaleal space (i.e., between the CSF and the outside world). An open head injury usually results from penetrating brain damage (e.g., from a shooting incident or use of another weapon) but can also arise from a tearing of the dura (e.g., with ventricular fistula following basal skull fractures). An additional categorizatation of the morphology of skull and brain injury indicates the presence or absence of bone damage and whether intracranial lesions are focal or diffuse. The types of head injury are summarized in Table V in terms of the mode of accident, the clinical picture, and the damage as represented by imaging techniques.

Primary brain damage results from the effects of shear, stretch, and pressure on the tissue (including intracranial and intracerebral pressure gradients). Damage is usually a consequence of either extreme pressure on a particular point on the skull (contact injury) or a rapid acceleration/deceleration of the head and the brain mass. Primary damage occurs immediately at the site of the accident and consists of the breaking of the skin and scalp fracture, contusions and crushing of the brain tissue, intracranial bleeding, and/or diffuse damage to the neuronal tissue and axonal network of the brain (Fig. 2).

Secondary brain damage appears with a delay. This manifests itself as a complication of the course and may be based on hypoxia, free radical formation, release of excitatory amino acids, ischemia, a swelling of the brain tissue, increased intracranial pressure, infection, or delayed (subacute) intracranial bleeding.

Table V
Classification of Traumatic Brain Injuries According to Biomechanical and Morphological Aspects[a]

Classification		
	Mechanism of injury	
Closed	High speed (e.g., car accident)	
	Low speed (e.g., a fall with somatic injury)	
Open	Missile injury	
	Other penetrating injuries	
	Morphology of the injury	
Skull fracture	Skull cap	Along the length/cross-sectional with/without impression; open/closed
	Base/basis	With/without CSF fistula; with/without facial paresis
Intracranial lesion	Focal	Epidural
		Subdural
		Intracerebral
	Diffuse	Diffuse axonal injury

[a]Modified from Keidel and Miller (1996).

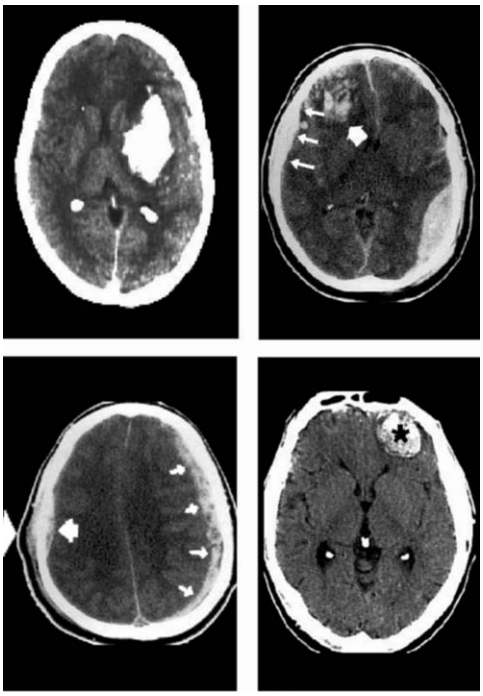


Figure 2 Common types of traumatic brain injuries. (Top, left) Intracerebral hemorrhage. (Top, right) Intracerebral contusion of the right frontal lobe accompanied by an ipsilateral frontotemporal epidural hematoma (coup) and a left-sided temporoparietal epidural hematoma (contre). (Bottom, left) Right-sided temporoparietal epidural hematoma, soft tissue injury, and contralateral subdural hematoma. (Bottom, right) Left frontal traumatic hemorrhagic contusion.

An overview of primary and secondary brain damage is presented in Table VI.

2. Intracerebral Bleeding

Epidural hematoma is a collection of blood between the skull and the dura. It usually results from low-speed head injuries (e.g., from falls or impact with objects). It occurs most commonly in association with linear skull fractures that traverse meningeal vessels and tear them. In about half of the patients, after an initial short period of unconsciousness on impact, a lucid interval can be observed that is followed by a progressive loss of consciousness and contralateral hemiparesis. The underlying mechanism, the expanding mass-producing compression of the corticospinal tract, transtentorial herniation, and diencephalic derangement, is visible in computed tomography (CT) scans as a biconvex structure of the epidural hematoma bounded by dural insertion at cranial sutures (Fig. 2).

Table VI
Type of Lesion and Mechanism of Injury in Posttraumatic Primary and Secondary Cerebral Damage

Lesion	Mechanism
Primary cerebral damage (acute manifestation)	Scalp injury
	Bruising or crushing of the brain
	Intracranial bleeding
	Diffuse axonal brain damage
	Acute intracranial hematoma
Secondary cerebral damage (delayed manifestation)	Hypoxia or ischemia
	Space occupying mass
	Delayed formation of hematomas
	Brain swelling
	Brain edema
	Hyperemia
	Hydrocephalus
	Increased intracranial pressure
	Arterial vasospasm
	Infections
	Cellular mechanisms
	Phospholipid metabolism
	Lipid peroxidation, arachidonic acid
	Prostaglandins, leukotrienes
	Platelet-activating factor
	Free oxygen radicals
	Excitotoxic mechanisms
	Glutmate, acetylcholine
	Neuropeptides
	Endorphines
	Disturbances of Ca^{2+} and Mg^{2+} metabolism
	Lactate acidosis
	Disturbances of axonal transport
Long-lasting sequelae	Focal scar formation
	Secondary degeneration of the neuronal pathways and nuclei
	Posttraumatic Demyelinization
	Brain atrophy
	Hydrocephalus internus

The *subdural hematoma* is a collection of blood between the dura and the underlying brain and frequently the result of torn veins draining into the

dura or dural sinuses. It may be the result of high-speed impacts, such as occur in car accidents. Associated severe brain damage (contusions and edema) is common. In CT scans damaged areas are not bordered by sutures and can cover most of the hemisphere but do not cross the interhemispheric fissure (Fig. 2). The treatment of both types of hematoma is comparable and consists of neurosurgical interventions to reduce the effects of the expanding mass.

The term *contusion* describes necrotic cortex and white matter with variable quantities of petechial hemorrhages and edema. Traumatic contusions show a typical pattern of distribution, covering the poles of the frontal and temporal lobes, the orbital surface of the frontal lobe, and the inferolateral temporal lobes (Fig. 2). Due to the contact of the brain to bony or meningeal structures, several types of contusions can be distinguished: at the site of fracture, at the site of impact (coup), contralateral to the site of impact (contre-coup), or at the site of herniation. All these appear in CT scans as areas of variable densities (increased, decreased, or normal).

Diffuse axon damage commonly occurs in traffic accidents due to shearing injury. Since a strong acceleration or deceleration can elicit these shearing forces and pressure gradients, a direct impact on the skull is not necessary. Often, traumatic hematoma of the corpus callosum and necrotic hemorrhages of the dorsolateral brain stem are associated with diffuse axon damage. Furthermore, it can develop from a trauma-induced sequence of changes of tissue by neurotoxic mediator substances leading to the final breakdown of axonal function. The diagnosis should be considered if, in unconscious patients, no focal lesion is displayed in the CT scan, perhaps merely a small swelling is evident.

D. Inflammatory Lesions

1. Infections

The CNS is protected from the direct spread of infectious agents by bony and meningeal coverings and by the blood–brain barrier. There are four routes of infection in the brain: hematogenous spread, direct implantation, local extension, and the neural route. The most common is the hematogenous route. Most viruses, bacteria, fungi, and protozoa use this route, often after local growth in nonneural tissue (mucous membranes of respiratory or enteric systems) or proliferation in the blood stream. There are several ways to penetrate the blood–brain barrier: Transport in infected reticuloendothelial cells, transport in circulating immune complexes, replication in infected endothelial cells, or passive pinocytotic transfer in the plexus chorioideus have been identified as strategies for viruses.

Direct implantation takes place almost invariably in traumatic lesions, but it is also sometimes iatrogenic (ventriculoperitoneal shunt and lumbar puncture). Furthermore, extensions of local adjacent infections of frontal or mastoid air sinuses or an adjacent osteomyelitis are portals of entry for bacterial infections.

The underlying mechanism of the neural route is bidirectional axonal transport that carries products of axolemmal renewal and the constituents of synapses. Pathologic agents may also use either the centropetal pathway (e.g., herpes simplex virus, rabies virus, or tetanus toxin) or the centrofugal pathway (e.g., reactivated herpes simplex or varizella zoster viruses or rabies virus).

Once an infectious pathogen has crossed the barriers into the nervous system, unique anatomical features are encountered. The brain has no intrinsic system for antibody prduction, no lymphatic system in the usual sense, and few phagocytic cells. The blood–brain area that inhibits invasion also deters clearance and impedes the entry of therapeutic agents. However, since few microorganisms overcome the barriers into the CNS, the invasion of the CNS is still a rare event.

The first responses to invading viruses consist of increased levels of lymphocytes (mostly T cells) and monoytes in the CSF. There may also be a modest increase in protein concentration. The reaction to pyrogenic bacteria is a fast and spectacular increase in polynuclear leukocytes and proteins. Less dramatic changes are seen after more slowly developing and less pyrogenic microorganisms (e.g., tuberculosis or listerial meningitis) invade. The pathological consequences of CNS infection depend on the type of microorganism. Viruses can induce a perivascular infiltration of monocytes and lymphocytes. Indirect inflammatory or direct cytopathic responses can induce neuronal damage until cell lysis along with damage to the myelin and glial cells and small focal hemorrhages. Various associated immunocomplexes also play a role in the response to viral and CNS components. For example, B cells migrate in, producing antibodies against the invading micororganisms, and as a reaction to the microbial antigens, T cells release cytokines that attract and activate other T cells and macrophages. Viruses show different preferences for inflicted cells:

On the one hand, rabies, polio, and herpes simplex and flavin viruses infect neurons, whereas on the other hand, the Jakob–Creutzfeldt virus attacks oligodendrocytes. Pathological changes appear more rapidly after bacterial infection. Here, local responses to bacterial antigens and toxins play an important role.

Brain infections such as encephalitis occur much more rarely than infections of the brain membranes such as meningitis. This means that many other factors must interact to allow relatively common bacteria or viruses to cause CNS infection. Bacterial infections are frequently marked by an inflammatory abscess for which the sources are difficult to differentiate. These arise from the immunological response in the form of fibrin deposits and exudates resulting from inflammation. These take the form of capsules and are mostly surrounded by a large edema. Streptococci and staphylococci can be found along with pneumococci, enterococci, and other anaerobic organisms. These abscesses appear as hyperdense areas in a CT scan: The edema presents itself as a ring of enhanced contrast.

Viruses are a more frequent cause of encephalitis than is bacteria. Indeed, the most frequent cause of severe, sporadic encephalitis is the herpes simplex virus. The virus may originate in the nasal mucus and cross the tractus olfactorius, or it may be reactivated from a persistent cache in the trigeminal ganglion. It may then infect the basal regions of the frontal or temporal lobes of either or both hemispheres. The effect is a dimming of consciousness, aphasia (reflecting a preferred infection of the left hemisphere), moderate hemiparalysis, and perhaps focal if not general epileptic discharges.

The pathological characteristics of herpes simplex encephalitis can be clearly visualized, if somewhat delayed, in radiological images. These consist of the spotty but continuous area of necrotic tissue, reflecting the infiltration of granular lymphocytes, and a marked edema in cases of rapid pressure damage (Fig. 3). Persistent tissue damage with severe signs of a neuropsychological deficit is frequent, despite antiviral therapy.

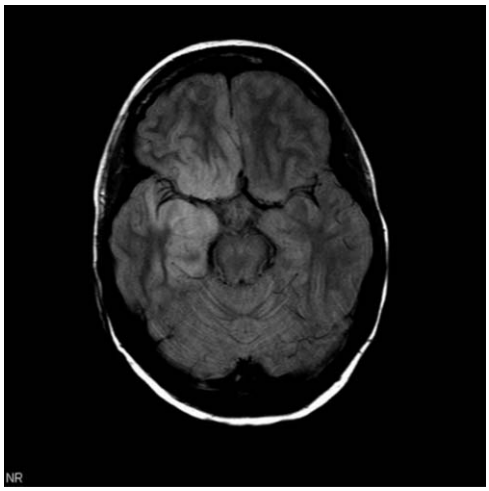


Figure 3 MRI scan of a 32-year-old man suffering from encephalitis caused by herpes simplex virus. Note the typical hyperintense lesion of the right temporal and basal frontal lobe.

2. Creutzfeld–Jakob Disease

Creutzfeld–Jacob disease (CJD) belongs to the group of transmissible spongiform encephalopathies. This group occurs in animals (scrapie in sheep and bovine spongiform encephalopathy) and in man (Kuru; fatal, familial insomnia; Gerstmann–Sträussler–Scheinker disease; and primarily CJD and a new variant nvCJD). They all share the following features: The pathological changes are neuronal degeneration and loss with astroglial proliferation presenting as small cystic (spongiform) areas in gray matter without any inflammatory signs leading to a widespread cortical atrophy. The clinical course includes a prolonged incubation period, possibly a decade or layer, with a subsequent progressive fatal course following the first symptoms. All diseases are transmissible and the responsible agent is the prion (proteinaceous infectious particle). Prions differ from all other microbial agents in that they contain no nucleic acid. This results in no immune response. They are also very resistant to the usual sterilization procedures (heat, radiation, ultraviolet light, and chemical agents). The pathologic prion protein (PrPsc) develops from a physiologic cellular prion protein with the same sequence of amino acids as a result of conformational change with more pronounced beta folding structure. The underlying mechanism of this conformational change is unclear. However, PrPsc is able to induce the same transformation from which it develops.

CJD occurs in sporadic, inherited (several mutations) and iatrogenic forms due to transplants of cornea or dura, insufficiently sterilized neurosurgical instruments, and growth hormone extracted from human hypophyses. The clinical course of all three forms is comparable: progressive dementia often accompanied by myoclonus, visual, cerebellar, or motoric disturbances and finally an akinetic mutism and decerebration. The electroencephalogram shows typical periodic complexes of sharp waves in an underlying decelerated background activity. Since

clinical diagnostic tools are not sufficiently, sensitive, a definite diagnosis depends on postmorteum analysis of cerebral tissue, including neuropathologic, immunhistochemical, and Western blot analysis with antibodies against PrP.

In 1996, a new variant of Creutzfeld–Jakob disease (nvCJD) was described. In contrast to the classical form, it starts at an ealier age (<30 years) with pronounced psychiatric signs. Also, electrophysiological and neuropathologic findings differ from those in sporadic cases. The similarity to experimentally induced CJD analog changes in macaque monkeys infected by transfer of BSE led to the hypothesis that nvCJD is caused by food containing BSE. This hypothesis has not been proven.

3. Multiple Sclerosis

Multiple sclerosis is a demyelinating disorder affecting the CNS. The clinical associations reflect white matter damage: upper motor neuron weakness and paralysis, incoordination, visual disturbances, and paraesthesia. The course is often acute onset followed by a progress of many years with alternating remissions and relapses. Less commonly, there are different variants, such as an acute severe disease that incurs a rapid progress to death, chronic progession without remission, and minimal signs with very long remissions. The typical pathologic findings are so-called ''plaques,'' areas of demyelination characterized inflammatory infiltrates in early stages, profound myelin loss, an almost total absence of oligodendrocytes, and a scattering of astrocytes in the late stages. Although plaques have a predilection for the angles of the cerebral ventricles and the corpus callosum (Fig. 4), they may occur anywhere in the CNS.

Multiple sclerosis is an autoimmune disease in which the patient's immune system reacts against the neuronal myelin sheath. In addition to demyelination, damage in the CNS occurs to the axons and an inflammatory filtrate can be detected. Although numerous aspects of the genetics, histopathology, and other immunological details have been the subject of recent and wide-ranging investigations, the precise pathophysiological course of the illness remains unclear.

Early microscopic studies revealed that the typical periventricular demyelinated foci oriented radially in the blood vessels. There are three recognized histopathological stages of lesion development:

1. Early developing lesions (acute inflammatory reaction of all parts of the myelin system and disrupted blood–brain barrier)

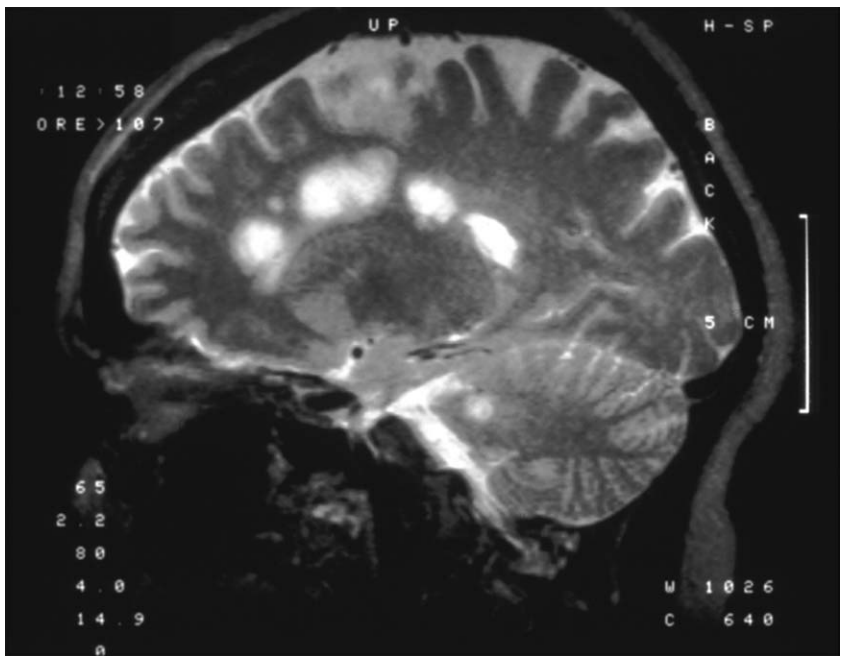


Figure 4 MRI scan of a 25-year-old woman with multiple sclerosis, which was diagnosed at age 22. Note the hyperintense pericallosal white matter lesions.

2. Late developing lesions (marked breakdown of the myelin in the absence of myeline oligodendrocyte glycoproteins or CNPase)
3. Inactive demyelinated lesions

Comparison of histological findings with those from magnetic resonance neuroimaging have shown that due to an impaired blood–brain barrier, the lesions in their early stages can take up much gadolinium. However, surprisingly, T2-weighted images show a remarkably weak signal intensity. Rather, the late developing or old lesions show areas of hyperintensity and a more homogeneous uptake of gadolinium. In contrast, axonal lesions can be shown to be associated more with hypointensities in T1-weighted images.

It is clear that the key immunological role in the development of the illness is played by activated T cells. Activated T helper cells interact with specific proinflammatory cytokines, particularly tumor necrosis factor-α, lymphotoxin, interferon-γ (IFN-γ), and interleukin-2 (IL-2). They can thus activate macrophages and damage the target tissue. Whereas in the peripheral circulation monocytes and macrophages can take up an antigen once detected, T cells are available, and phagocytosis can occur, in the CNS such activities are directed against the microglia, astrocytes, oligodendrocytes, and the myelin they synthesize. The most important target candidates include proteolipid protein (which that makes up about 50% of the protein content of the myelin sheath), basic myeloprotein, myelin-associated glycoprotein, and myelin oligodendrocyte glycoprotein. T2 helper cells destroy the antiinflammatory cytokines such as IL-4 and IL-10 that would otherwise keep the functions of the inflammatory T cells in check. During restitution, this type of cell shows a higher level of activity. Nevertheless, this subtype of helper cell supports antibody synthesis in the β-lymphocytes with the result that the inflammatory process and the underlying mechanisms are perpetuated and the damage becomes chronic.

The course of the inflammatory process may occur as follows: T cells are activated by an unknown antigen in the periphery. This activation is followed by the release of specific cytokines (e.g., the expression of integrins and the intracellular cell adhesion molecule-1 from the superfamily of immunoglobulins) that promote the binding of the activated T cells to the endothelium of the blood vessels. By way of the release of the proinflammatory cytokines, the structure of the blood vessels in the blood–brain barrier disintegrates. This enables the migration of T cells to the structures that would otherwise have been protected. During the course of this migration, mediators of the toxic effect on myelin are released, edemas form, and the constituents of myelin are affected in their function and eventually destroyed. The cytokines of the T2 helper cells (IL-4 and IL-10) are modulators of this process. They help reduce the extent of the inflammation and may even increase in number during the more stable phases of the illness. The extent to which apoptosis (programmed cell death) of the T cells actually influences this process cannot be determined with confidence.

The mainstays of therapy constitute neurophysiologically based practice, ergotherapy, and various measures for support and rehabilitation: the latter include psychotherapy, pharmacotherapy directed toward symptom relief, as well as the putative causes (immunomodulation, inhibition of glial scar formation, and the promotion of remyelination). High-dose corticosteroid therapy in the acute phase can help inhibit inflammation and can stabilize the blood–brain barrier, and it can be directed toward T cell apoptosis. In the form of illness with an episodic course, immunomodulators may also be applied. Thus, recombination IFN-β can antagonize the effects of IFN-γ and the induced suppressor activity, and it can stabilize the blood–brain barrier. Glatirameracetate, a synthetic polypeptide, can promote T cell activation and reintroduction of the disturbed TH1–TH2 balance. Other therapeutic options include the use of mitoxantron, azathioprin, and methotrexate, which can inhibit global proliferation and inflammation.

E. Degenerative Lesions

Use of the rubric of neurodegenerative disease is intended to cover a group of illnesses characterized by a progressive loss of neuronal populations in the CNS. In contrast to the defined metabolic, toxic, hypoxic–ischemic, or infective lesions described previously, the pathogenesis of neurodegenerative illness remains largely unknown. Some of these illnesses belong to the most frequent neurological entities and include Alzheimer's and Parkinson's disease. The affected neuronal populations in some of these illnesses do show a characteristic histologic pattern of damage (e.g., Alzheimer's and Pick's disease). However, in other illnesses there seems to be little more to describe than a general reduction of the number of neurons in

particular parts of the brain. In many cases, the affected brain regions demonstrate a decided atrophy during a long course with an associated pattern of functional disturbance.

Depending on how the degenerating neurons are distributed, one can classify the illnesses according to a diffuse and generalized atrophy (e.g., Alzheimer's disease), or a system (anatomic and function)-specific atrophy (Parkinson's disease). Occasionally a third group of multisystem atrophy is referred to in which several functional systems are affected (olivopontocerebellar atrophy). An overview of degenerative diseases is presented in Table VII.

1. Alzheimer's Disease

Alzheimer's disease is the most common degenerative brain disease and the most common cause of dementia, and it includes a presenile and senile form. Most cases occur sporadically, although a small proportion are familial and inherited. The heritability is distinguished by an autosomal dominant disorder that leads to disease and vulnerable genes that increase the probability of illness. The pathological features of Alzheimer's disease are neuritic plaques that consist of amyloid proteins and neurofibrillary tangles that contain isoforms of the phosphorylized τ-protein. Both are also found in normal aging, although their density and distribution typically differ from those seen in Alzheimer's disease, in which the number of plaques is approximately linearly related to the degree of dementia. On a cellular level, pyramidal cells, the main neuronal component of corticocortical pathways, are the major cortical cell type lost in Alzheimer's disease. In the early stages, the loss is predominantly confined to the amygdala and to the temporal and parietal association cortex. Later, degeneration develops in other mesial temporolimbic structures such as the hippocampus and in cholinergic nuclei in the basal forebrain (nucleus basalis of Meynert and the locus ceruleus). Degeneration here incurs a decrease of activity in transmitter systems (acetylcholine, serotonin, and others). Primary visual and sensorimotor areas remain relatively spared so that sensory perception and movement remain intact, whereas deficits in memory, learning, attention,

Table VII
Etiology, Morphology, and Clinical Correlates of Degenerative Disorders of the Brain

Disorder	Etiological factor	Morphological characteristic	Preferred locus	Clinical association
Alzheimer's disease	A4 amyloid presenilin I/II	Tangles, neuritic plaques congoph; angiopathy	Whole cortex	Progessive loss of higher brain functions: dementia
Pick's disease	Mostly sporadic incidence	Eosinophils, round inclusion bodies (Pick bodies)	Frontal and temporal lobes	Primary frontal lobe syndrome followed by progressive dementia
Chorea Huntington	Autosomal dominant hereditary (4p 16.3) trinucleotide (CAG)-coded protein aggregations	Intranuclear inclusions in interneurons of the n. caudatus with a progressive loss	Atrophy of the n. caudatus, also cortical atrophy in the frontal and temporal lobes	Choreatic motor disturbances with progressive dementia and neuropsychiatric symptoms
Parkinson's disease	Neurotoxin hypothesis: disturbance in the chain of respiratory enzymes in the mitochondria	Loss of pigmented dopaminergic neurons; concentric neuronal inclusion bodies (Lewy bodies of synuclein)	Substantia nigra and to a lesser degree the locus coeruleus	Disturbed extrapyramidal movements (rigor, tremor, hypokinesis)
Olivopontocerebellar atrophy	Autosomal dominant, recessive, sporadic	Degeneration and atrophy of the transverse pontocerebellar pathways, loss of Purkinje and granule cells	Ventral pons, cerebellar stem and cortex, and inferior olive	Progressive cerebellar ataxia

and in the ability to solve complex problems increase. The final state consists morphologically of a generalized atrophy of gray matter and clinically of dementia.

The underlying defect remains unknown. It has been suggested that β-amyloid (β-AP) plays a central role. It is produced by the splitting of a precursor protein, the amyloid precursor protein (APP). It is an integrative membrane glycoprotein that is encoded on chromosome 21 and mainly expressed in the brain and the kidneys. Conversely, APP may be internalized and cleaved by unknown secretases to form β-AP fragments, which aggregate in fibrillar, nonsoluble material to comprise the core of the neuritic plaque. It has also been suggested that amyloid production mediates the pathological phosphorylization of τ-protein, leading to the formation of neurofibrillary tangles. The neurotoxicity results in oxidative stress, with increased intracellular reactive oxygen species, and disruption of structures involved in ion homeostasis. Inflammatory responses with reactive glial cells lead to the production of cytokines and complement.

These pathologic events are the possible pharmacological targets in the treatment of Alzheimer's disease. To date, only a cholinergic enhancement with acetylcholinesterase inhibitors has been shown to provide a symptomatic benefit and may slow disease progression. Antiinflammatory and antioxidant medication and mediation of APP metabolism are also options that are being researched.

Acknowledgments

We thank our collegues M. Forsting, J. Weber, and C. Weihl, who gave consent for publication of the CT and MRI scans. We are indebted to Dr. Oades for editorial work on the manuscript.

See Also the Following Articles

ALZHEIMER'S DISEASE, NEUROPSYCHOLOGY OF • ASTROCYTES • BRAIN DAMAGE, RECOVERY FROM • BRAIN DISEASE, ORGANIC • LYME ENCEPHALOPATHY • MODELING BRAIN INJURY/TRAUMA • MULTIPLE SCLEROSIS • PRION DISEASES • STROKE

Suggested Reading

Adams, R. D., and Victor, M. (1998). *Principles of Neurology*, 4th ed. McGraw-Hill, New York.

Freund, H. J., Sabel, B. A., and Witte, O. W. (Eds.) (1997). *Advances in Neurology. Vol. 73: Brain Plasticity*. Lippincott–Raven, Philadelphia.

Graham, D. I., and Lantos, P. L. (Eds.) (1999). *Greenfield's Neuropathology*, 6th ed. Arnold, London.

Keidel, M., and Miller, J. D. (1996). *Head trauma*. In *Neurological Disorders: Course and Treatment*. (Th. Brandt, L. R. Caplan, J. Dichgans, H. C. Diener, and Ch. Kennard, Eds.), pp. 531–544. Academic Press, San Diego.

Mandell, G. C., Bennett, J. E., and Dobin, R. (1995). *Principles and Practice of Infectious Diseases*, 4th ed. Churchill–Livingstone, Edinburgh, UK.

Witte, O. W. (1998). Lesion-induced plasticity as a potential mechanism for recovery and rehabilitative training. *Curr. Opin. Neurol.* **11,** 655–662.

Brain Stem

WILLIAM W. BLESSING
Flinders University, Adelaide, Australia

GLOSSARY

chronic vegetative state Condition of a person with bilateral forebrain and/or upper brain stem dysfunction that eliminates thinking, purposive behavior, and self-consciousness but leaves intact the basic homeostatic mechanisms necessary for life.

cranial nerves The 12 sensory and/or motor nerves that connect the brain, especially the brain stem, with the peripheral sensory and motor organs in the region of the face and head.

nucleus of the tractus solitarius Secondary sensory nucleus in the dorsomedial medulla oblongata that receives inputs from taste, respiratory, cardiovascular, visceral, and gastrointestinal primary afferents traveling in cranial nerves VII, IX, and X.

parasympathetic motoneurons in the brain stem Preganglionic motoneurons, with axons leaving the brain in cranial nerves III, VII, IX, and X, that innervate cranial, cervical, thoracic, and abdominal postganglionic neurons responsible for controlling visceral functions such as pupillary diameter, salivation, heart rate, bronchosecretion, and gut motility.

premotor neurons Forebrain or brain stem neurons with axonal projections synapsing on somatic, parasympathetic, or sympathetic motoneurons.

presympathetic motoneurons in the brain stem Neurons with axons that descend to the spinal cord and synapse on sympathetic preganglionic neurons in the thoracic and upper lumbar spinal cord.

primary sensory neurons Sensory (afferent) neurons, with cell bodies in the dorsal root ganglia or in peripheral sensory ganglia located in the upper neck or facial regions, that convey information from the periphery to the spinal cord or the brain.

reticular formation Regions of the brain stem containing neuronal cell bodies (motoneurons, premotor neurons, and interneurons) interspersed between bundles of axons; the cell bodies are not gathered into obvious nuclei and the appearance is net-like. The term is also used to signify a brain stem system for alteration of the level of arousal and consciousness.

secondary sensory neurons Brain and spinal cord neurons receiving direct synaptic inputs from the central processes of the primary sensory neurons.

somatic motoneurons Brain stem and spinal motoneurons with a peripherally directed axon that innervates striated muscle; subdivided into general somatic efferent and special visceral efferent.

The brain stem is the portion of the central nervous system rostral to the spinal cord and caudal to the cerebral hemispheres. This article defines the different collections of nerve cells and axonal pathways that constitute the brain stem and summarizes the manner in which brain stem neuronal circuitry mediates the basic bodily homeostatic "housekeeping" functions necessary for our daily lives.

I. NAMES AND THE BRAIN STEM

The brain is such a complex structure that even now we know only a tiny portion of what is to be known about it. One of our human characteristics is the need to feel familiar with the unknown, and one of the ways we achieve this is to give names to things. Thus, a small group of stars in our Southern Hemisphere night sky is named the Southern Cross and recognition of this familiar star patterns is comforting, guiding us on our journeys even if there are no intrinsic properties

connecting the individual stars and demarcating them from surrounding stars. The same principle applies to many of the named regions of the brain. Often, the names are not scientifically justified, and sometimes they are misleading. In the midbrain, the four "colliculi" are named for their resemblance to four "little hills." These same structures also constitute the "tectum," i.e., the dorsal part or "roof" of the midbrain. There is no functional implication in any of this terminology. Thus, traditional names for different brain regions may be a poor guide to the function or even to the structure of the named region.

Thus, when we ask "What is the brain stem?" we need not be too concerned if the answer is a little ambiguous. The underlying rationale for the use of the term is the division of the brain into the cerebral hemispheres (the presumed "flower" of the brain) and the remainder of the brain (the "stem"). The arbitrary dividing line has been drawn at different levels by different investigators. Sometimes the thalamus and the hypothalamus (grouped together as the diencephalon) have been included with the brain stem, especially by physiologists, who developed the concept of the reticular activating system. However, neuroanatomists commonly confine the term to the midbrain, the pons, and the medulla oblongata, without the cerebellum (the little brain). This is the meaning adopted for the term brain stem in this article. In general, it is best to regard the term brain stem as a conveniently loose way of referring to the candal or inferior half of the brain. There are no sharp demarcations between spinal cord, medulla oblongata, pons, and midbrain. The midbrain (the mesencephalon) is so called because as this brain region develops, it is between rostral and caudal flexures in the developing neural tube. The term pons (bridge) derives from the fibers from each side of the cerebellum which sweep around the ventral aspect of the brain stem.

The term "nucleus" (understood in its "collection of neurons" sense, not in its "part of the cell containing the DNA" sense) requires some explanation when it is used with respect to the brain stem. When brain stem sections are examined after a traditional Nissl stain (a technique for demonstrating the cell body portion of the neuron), particular collections of neurons (e.g., the hypoglossal nucleus and the inferior olivary nucleus) virtually define themselves as nuclei. They stand out quite obviously. Other areas of the brain stem appear as a network arrangement of fiber pathways and cell bodies, without clearly definable nuclei. Traditionally, these regions of the brain stem are referred to collectively as the reticular formation.

The brain stem also contains numerous ascending and descending fiber pathways linking the forebrain and the cerebellum with the spinal cord as well as pathways originating in the brain stem. These axonal pathways, called fibers of passage, often pass right through particular brain stem nuclei, dividing them into apparently different subgroups.

The different parts of the brain stem are indicated in a sagittal section (Fig. 1) and in transverse sections of the brain stem from the caudal medulla to the midbrain (Fig. 2).

II. BRAIN STEM NEURONAL ORGANIZATION, INCLUDING THE RETICULAR FORMATION

Portions of the brain stem function as "the spinal cord for the head." The facial, jaw, tongue, pharynx, larynx, and the neck muscles are innervated by efferent axons of brain stem somatic and "special visceral" lower motoneurons (cranial motoneurons), just as the muscles of the arms or the legs are innervated by spinal somatic lower motoneurons. Axons of these cranial motoneurons emerge from the ventral or ventrolateral surface of the brain stem, leave the cranial cavity via specific holes ("foramina") in the skull, and continue on as part of the various cranial nerves. Similarly, on the sensory side, the skin of the face and head is innervated mostly by afferent/sensory axons of trigeminal and upper cervical sensory neurons. The cell bodies of these afferent neurons are in ganglia outside the cranial cavity (i.e., the cavum trigeminale for the trigeminal nerve) and the central processes enter the brain and synapse with secondary sensory neurons in dorsal portions of the medulla and pons, just as spinal afferents synapse in the dorsal horn of the spinal cord.

The rostral end of the neuraxis also receives sensory input from special sensory organs located at the front end of the individual, the portion that first moves into a new environment. Olfactory receptors are stategically placed at the front end of the respiratory system. Eyes and ears are also positioned at the front end, as is the vestibular apparatus located within the inner ear. Taste receptors and their sensory nerves are at the front end of the digestive system. The central processes of sensory neurons carrying information from hearing, vestibular apparatus, and taste receptors enter the brain at the level of the medulla and pons and synapse with second-order sensory neurons located in dorsally situated sensory nuclei. Special sensory neurons also detect internal bodily states (blood pressure, blood oxygen, and carbon dioxide levels; stomach distention;

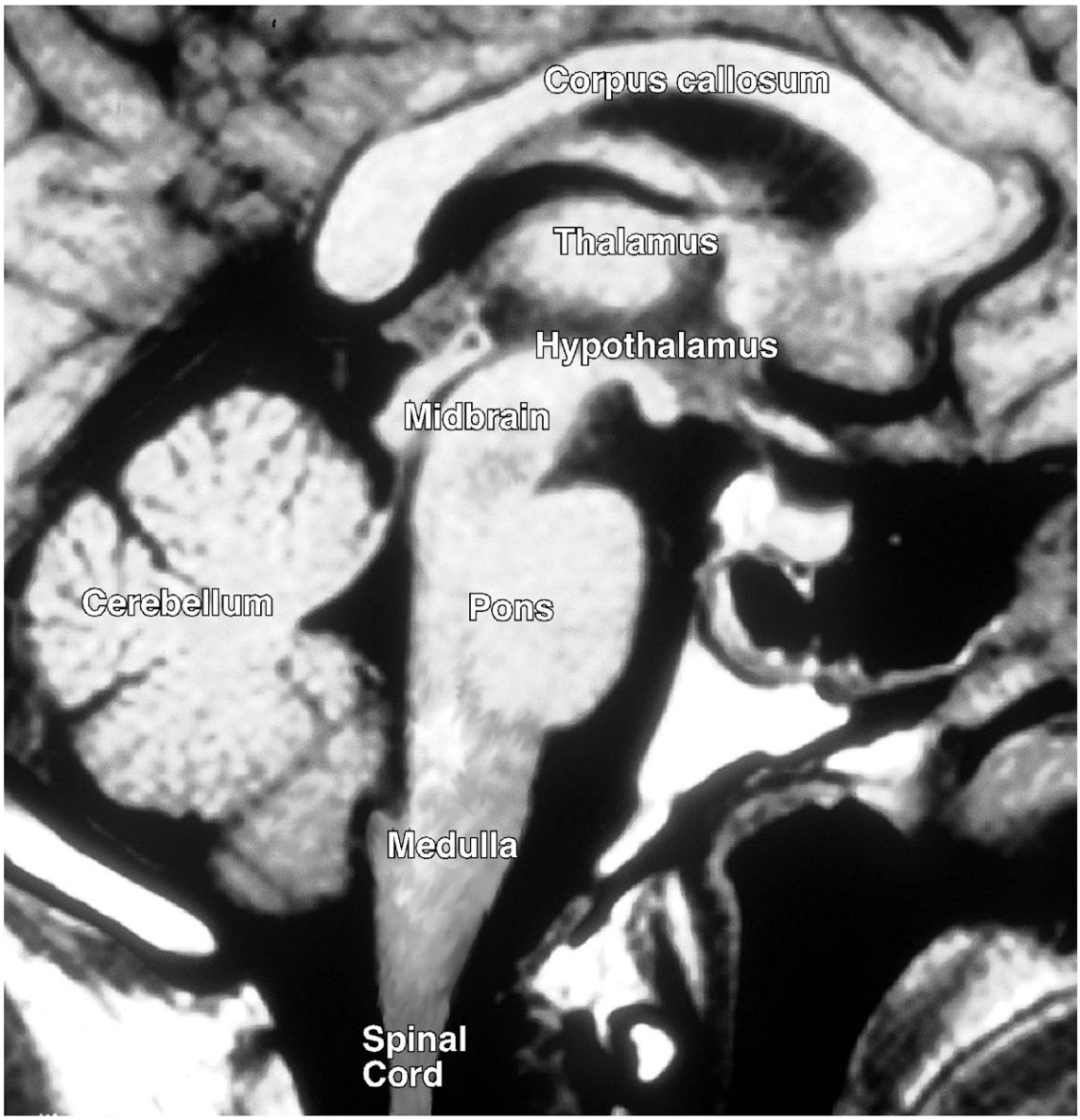


Figure 1 Magnetic resonance image of the sagittal plane of a human brain stem (modified with permission from Blessing, 1997).

and blood sugar level) and the central processes of these cells also reach the brain stem via the lower cranial nerves.

There is no direct communication between different groups of cranial motoneurons, so during, for example, swallowing, coordinated movements of the lips and the tongue are controlled by inputs from integrating premotor neurons with axonal collaterals innervating particular subgroups of lower motoneurons and not by collaterals of the lower motoneurons. Many of the integrative premotor neurons are also located in the brain stem. They help coordinate the behavioral, respiratory, cardiovascular, visceral, gastrointestinal, eliminative, and reproductive events essential for survival of the individual and the species. In turn, the premotor neurons are controlled by interneurons that are controlled by higher order integrative interneurons in the brain stem and in other parts of the brain. These interneurons receive inputs from the afferent side of the nervous system so that their genetically determined patterned output can also reflect the influence of past and present environmental events. Different groups of brain stem premotor neurons and interneurons are described in Table I.

As already noted, in the brain stem not all particular subgroups of neurons are arranged in discrete ''nuclei'' clearly defined with conventional Nissl stains. Certain brain stem neurons (including lower motoneurons, premotor neurons, and interneurons) are arranged diffusely in regions of the medulla, pons, and midbrain, scattered between ascending and descending fiber bundles. The net-like appearance of the regions containing these neurons led to the designation ''reticular formation,'' a term that was originally used in a purely descriptive anatomical sense. However, in the 1950s and 1960s, an additional functional meaning was assigned to the reticular formation so that the name began to refer to a hypothetical anatomical–physiological neuronal entity responsible for regulation of the level of arousal and consciousness as well as the maintenance of bodily posture. Instead of being used in a descriptive anatomical way, the reticular formation was promoted to a functional concept, a brain stem system which, by virtue of its nonspecific connectivity, could act as a kind of volume control for the degree of conscious arousal, among other things. The rostral part of the brain stem reticular formation became the ''ascending reticular activating system'' and loss of consciousness with brain stem injury was attributed to damage to this system. The caudal part of the brain stem reticular formation was seen as a source of descending excitatory or inhibitory inputs to various brain stem centers and to the spinal cord.

The problem was that the very nonspecific and all-inclusive nature of the reticular formation, conceived as a kind of functioning neuronal system, made it difficult to generate specific research hypotheses. This extended use of the term meant that the reticular formation became elevated to a magical entity, an easy shortcut explanation of complex and little understood physiological processes. This was not a helpful strategy. The term reticular formation should be restored to its original neuroanatomical status and emphasis should be placed on specifying the particular group of brain stem lower motoneurons, premotor neurons, and interneurons responsible for the various physiological functions. The cerebral hemispheres and the cerebellum can be viewed as an additional vast collection of interneurons whose coordinated discharge controls less complex sets of interneurons and premotor neurons in the brain stem reticular formation, thereby producing a finely coordinated discharge of brain stem motoneurons. Some axons descending from control neurons in the cerebral cortex synapse directly onto lower motoneurons in the brain stem or in the spinal cord, but a substantial proportion synapse

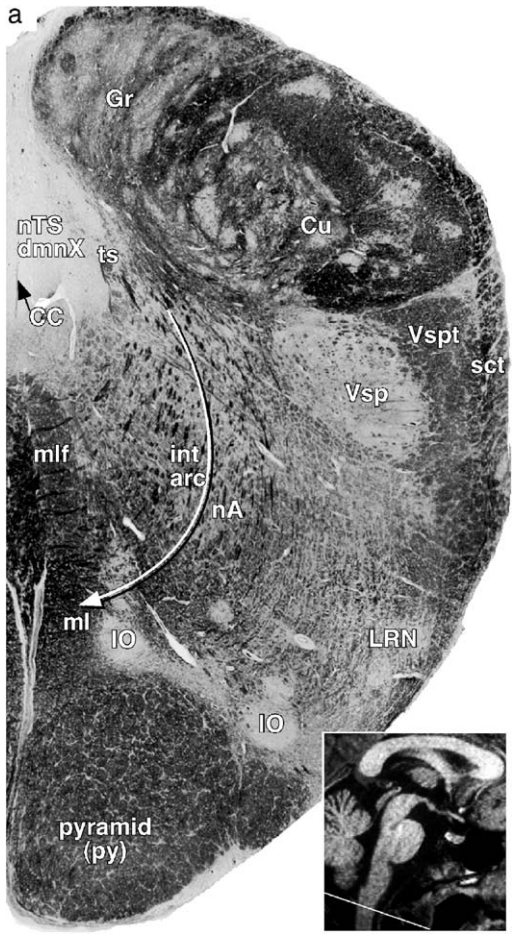


Figure 2 A series of photomicrographs of transverse sections through the human brain stem, stained for myelin by the Weil procedure. The insert in each figure gives the approximate level from which the section is taken Aq, aqueduct; Arc, arcuate nucleus; Coch, cochlear nucleus; Cu, cuneate nucleus; Cuext, external cuneate nucleus; Cun, cuneiform nucleus; dmnX, dorsal motor nucleus of the vagus; DR, dorsal raphe nucleus; Gr, gracile nucleus; IC, inferior colliculus; icp, inferior cerebellar peduncle; Int, nuclues intercalatus; IV, trochlear nucleus; IX fibers, intramedullary fibers of the glossopharyngeal nerve; lat lem, lateral lemniscus; LC, locus coeruleus; LRN, lateral reticular nucleus; LVe, lateral vestibular nucleus; mcp, middle cerebellar peduncle; ml, medial lemniscus; mlf, medial longitudinal fasciculus; MR, median raphe nucleus; MVe, medial vestibular nucleus; nA, nucleus ambiguus; nTS, nucleus of the tractus solitarius; PAG, periaqueductal gray; PB, parabrachial nucleus; PN, pontine nuclei; PPT, pedunculopontine tegmental nucleus; PrPH, nucleus prepositus hypoglossi; py, pyramidal tract; RM, nucleus raphe magnus; scp, superior cerebellar peduncle; sct, spinocerebellar tract; SN, substantia nigra; SpVe, spinal vestibular nucleus; ts, tractus solitarius; V_{fibers}, intramedullary fibers of the trigeminal nerve; VI_{fibers}, intramedullary fibers of the abducent nerve; VII_{fibers}, intramedullary fibers of the facial nerve; VII_{genu}, genu of the facial nerve; V_{mes}, mesenteric nucleus of the trigeminal nerve; V_{sp}, spinal nucleus of the trigeminal nerve; V_{spt}, spinal tract of the trigeminal nerve; XII, hypoglossal nucleus (modified with permission from Blessing, 1997).

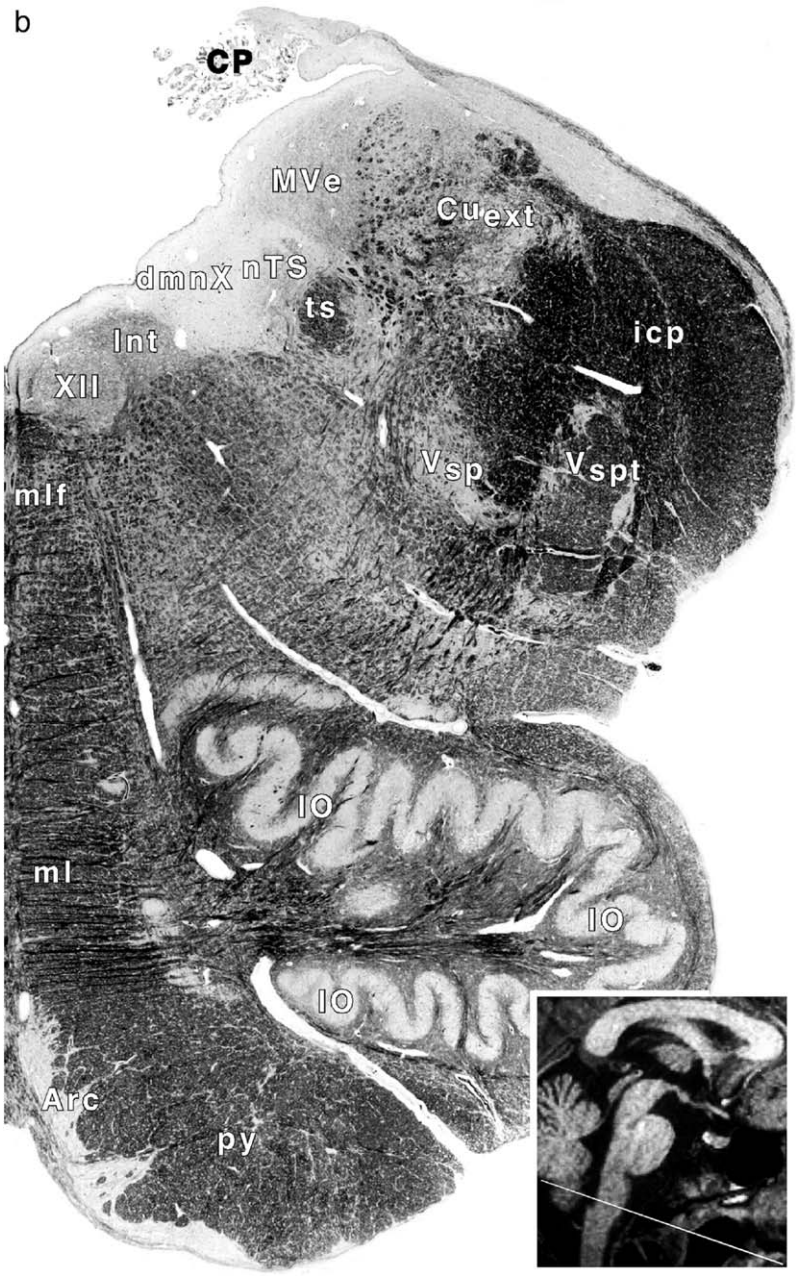


Figure 2 *(continued)*

with relevant brain stem interneurons and premotor neurons.

Much is already known concerning the lower motoneurons and the secondary sensory neurons in the different brain stem regions. Traditionally, these two groups are subclassified into particular categories, as summarized in Table I. Secondary sensory neurons are classified according to whether their peripheral processes terminate in somatic structures (skin and joints), in special sensory structures (hearing, vestibular function, and taste), or from internal receptors in the large arteries (baroreceptors and chemoreceptors) or in the viscera (heart, lungs, stomach, etc). Brain stem lower motoneurons are classed as somatic if their axons innervate striated muscle (e.g., the muscles of the face, jaw, and tongue) and as parasympathetic if their axons

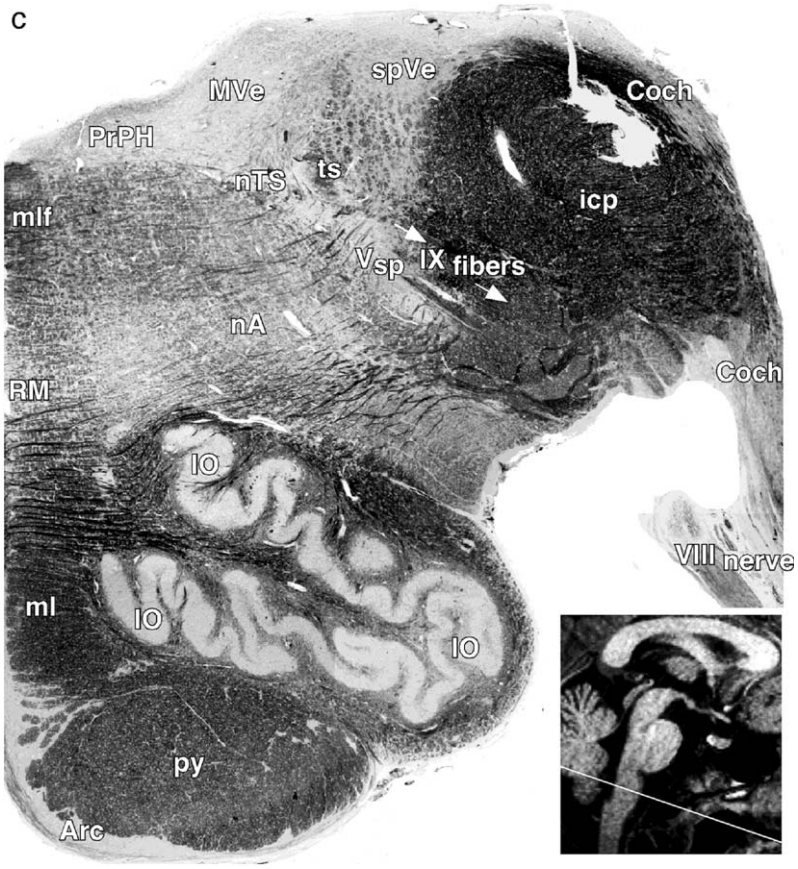


Figure 2 *(continued)*

innervate peripheral ganglionic neurons which, in turn, project to viscera in the head, neck, thorax, or abdomen (e.g., the blood vessels in the head, the salivary glands, the heart, lungs, and the abdominal viscera). Somatic efferents are subdivided (because of embryological considerations) into medially situated general somatic efferent nuclei (innervating extraocular muscles and tongue) and more ventrolaterally situated special visceral efferents, innervating (striated) muscles of chewing, facial expression, and swallowing.

Traditionally, the peripheral parasympathetic cell bodies are designated as "postganglionic" even though they are located within ganglia. It is the axons which are postganglionic. It is best to refer to the parasympathetic cell bodies as final parasympathetic motoneurons. The motoneurons in the brain stem can be referred to as parasympathetic preganglionic motoneurons or, in the appropriate context, simply as brain stem parasympathetic motoneurons. The "preganglionic" terminology is reasonable, but it derives from an earlier period when attention was focused much more on the periphery than on the brain.

Brain stem and forebrain neurons with inputs to parasympathetic motoneurons can be referred to as preparasympathetic motoneurons. The sacral region of the spinal cord also contains parasympathetic preganglionic motoneurons (sacral parasympathetic motoneurons) so that some of the brain stem preparasympathetic motoneurons have long axons descending to the sacral cord from different brain stem groups (e.g., Barrington's nucleus for control of micturition).

A. Cranial Nerves: Associated Motor and Sensory Nuclei

There are 12 paired cranial nerves, numbered in Roman numerals from the most rostral (in the forebrain) to the most caudal (in the lower brain stem). The nerves are named for their particular functions (the "oculomotor" nerves move the eyes) or for some other notable property ("trigeminal," meaning "three twins," is the name given to the fifth cranial nerve; "vagus," meaning "wanderer," is the name for the 10th cranial nerve).

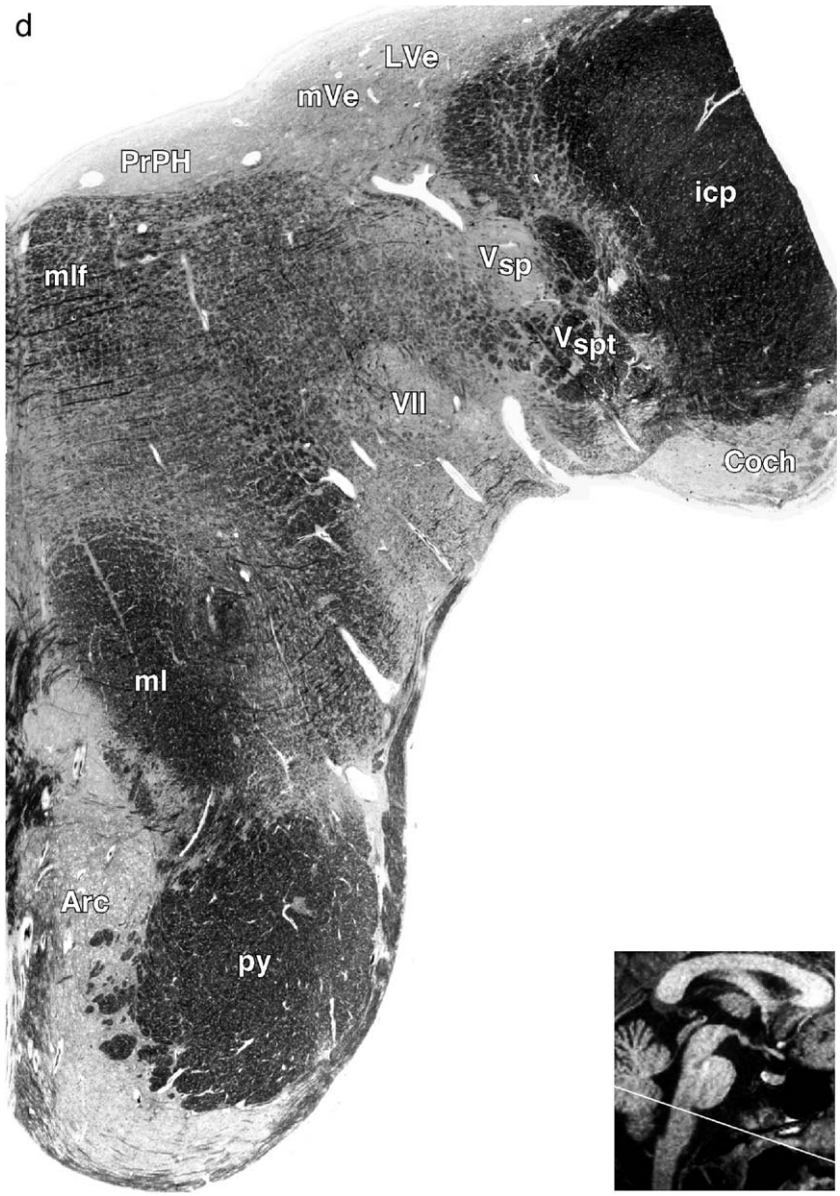


Figure 2 *(continued)*

Some of the cranial nerves are motor, some are sensory, and some are mixed (Table II).

B. Axonal Pathways Traversing the Brain Stem

1. Descending Pathways

The direct motor pathway from the cerebral cortex to the spinal cord comprises myelinated axons of motoneurons in the cerebral cortex. These axons form bilateral thick bundles which traverse midbrain, pons, and medulla. As shown in the different cross sections in Fig. 2, the pathway is often loosely referred to as the pyramidal tract. In the midbrain the anterior region through which the pyramidal tract passes is called the crus cerebri or the basis pedunculi. In the pons the fibers are divided into smaller bundles by bundles of axons which sweep around the front of the brain stem. In the medulla, the corticospinal fibers are again situated medially and anteriorly, and the paired pathways are referred to as the pyramids.

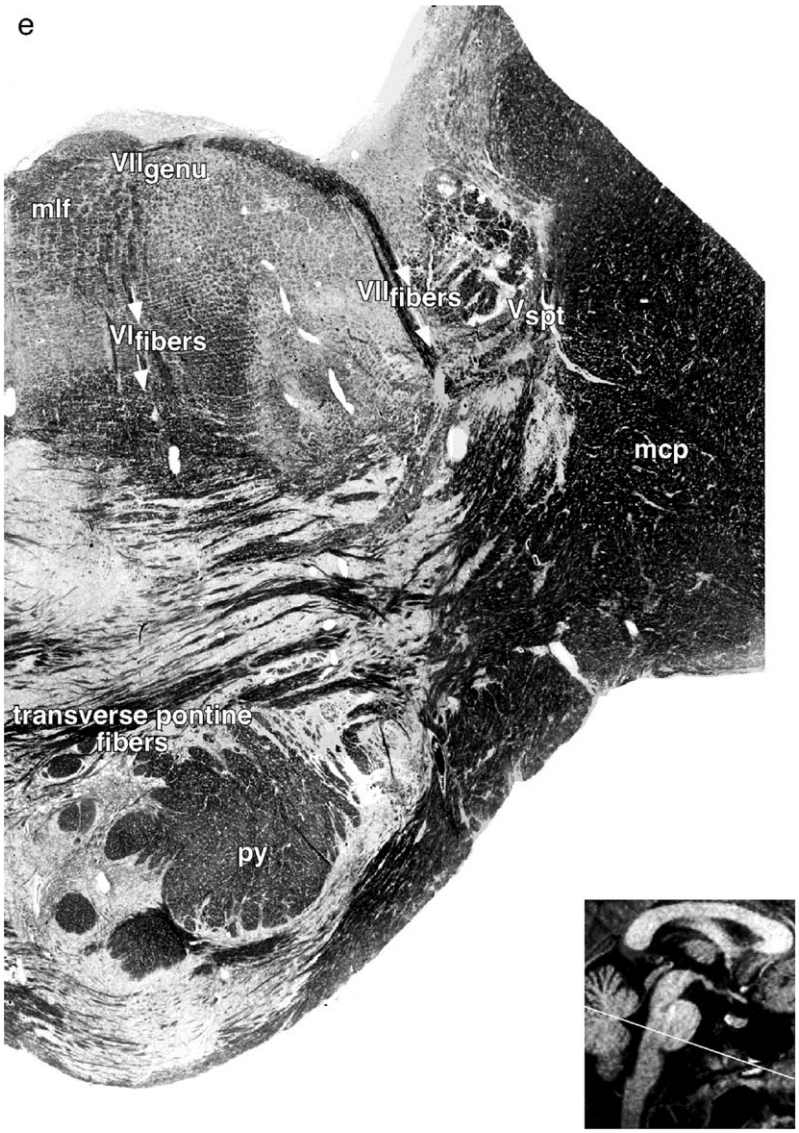


Figure 2 *(continued)*

There are other well-recognized descending pathways whose names provide helpful clues to their neuroanatomy. The rubrospinal tract consists of axons of neuronal cell bodies located in the red nuclei in the midbrain. The vestibulospinal tract consists of axons of neuronal cell bodies located in the vestibular nuclei in the dorsolateral part of the rostral medulla oblongata. Other descending pathways include the so-called ''reticulospinal'' tracts. These axons descend in less well-defined pathways. The hypothalamus and certain regions of the pons and medulla contain neurons (presympathetic neurons) whose axons descend to the spinal cord via the dorsolateral funiculi and innervate sympathetic preganglionic neurons located in the intermediolateral columns of the thoracic and upper lumbar spinal cord.

Other neurons in the medulla and pons function as parasympathetic premotor neurons with axons descending to the sacral portion of the spinal cord to innervate the parasympathetic preganglionic neurons that regulate genital and bowel–bladder function.

2. Ascending Pathways

Cell bodies in the dorsal root ganglia of the spinal cord whose peripheral processes detect light touch,

Figure 2 *(continued)*

vibration, and joint position sense have a centrally directed process that passes into the ipsilateral dorsal column of the spinal cord, ascends to the dorsal region of the medulla oblongata, and synapses on a second-order sensory neuron in the gracile and cuneate nuclei (Figs. 2a and 2b). The axon of the second-order neuron passes ventrally and medially before crossing to the other side and ascending through the medulla, pons, and midbrain in a distinctive pathway known as the medial lemnisus (Figs. 2a–2g). Lemniscal axons synapse in the thalamus.

In contrast, cell bodies in the dorsal root ganglia of the spinal cord whose peripheral processes detect pain and temperature sensation have a centrally directed axon which synapses in the dorsal horn near the level of entry. The axon of the second-order neuron passes ventrally and medially, crossing the midline just ventral to the central canal. The axon then ascends as the spinothalamic tract, through the spinal cord, medulla oblongata, pons, and midbrain, before synapsing in the thalamus.

Other bundles of axons ascend as components of the spinocerebellar pathways, traversing the lateral part of the medulla oblongata and entering the cerebellum via the inferior or superior cerebellar peduncles. In the medulla oblongata, an additional ascending pathway to the cerebellum consists of axons of inferior olivary neurons (Figs. 2b and 2c). These olivocerebellar axons pass medially from the olive, crossing the midline and entering the laterally positioned inferior cerebellar peduncle en route to the cerebellum.

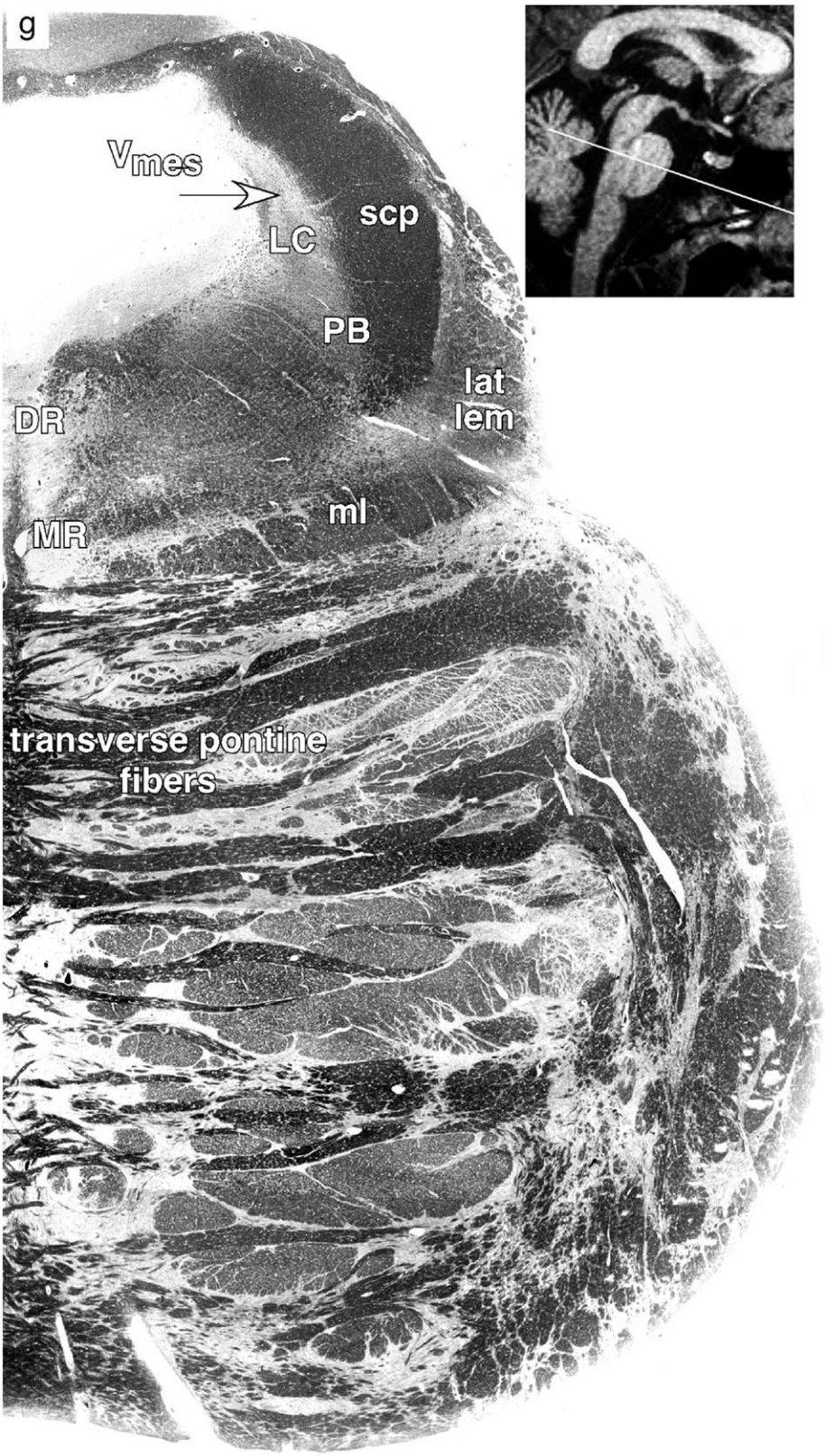


Figure 2 *(continued)*

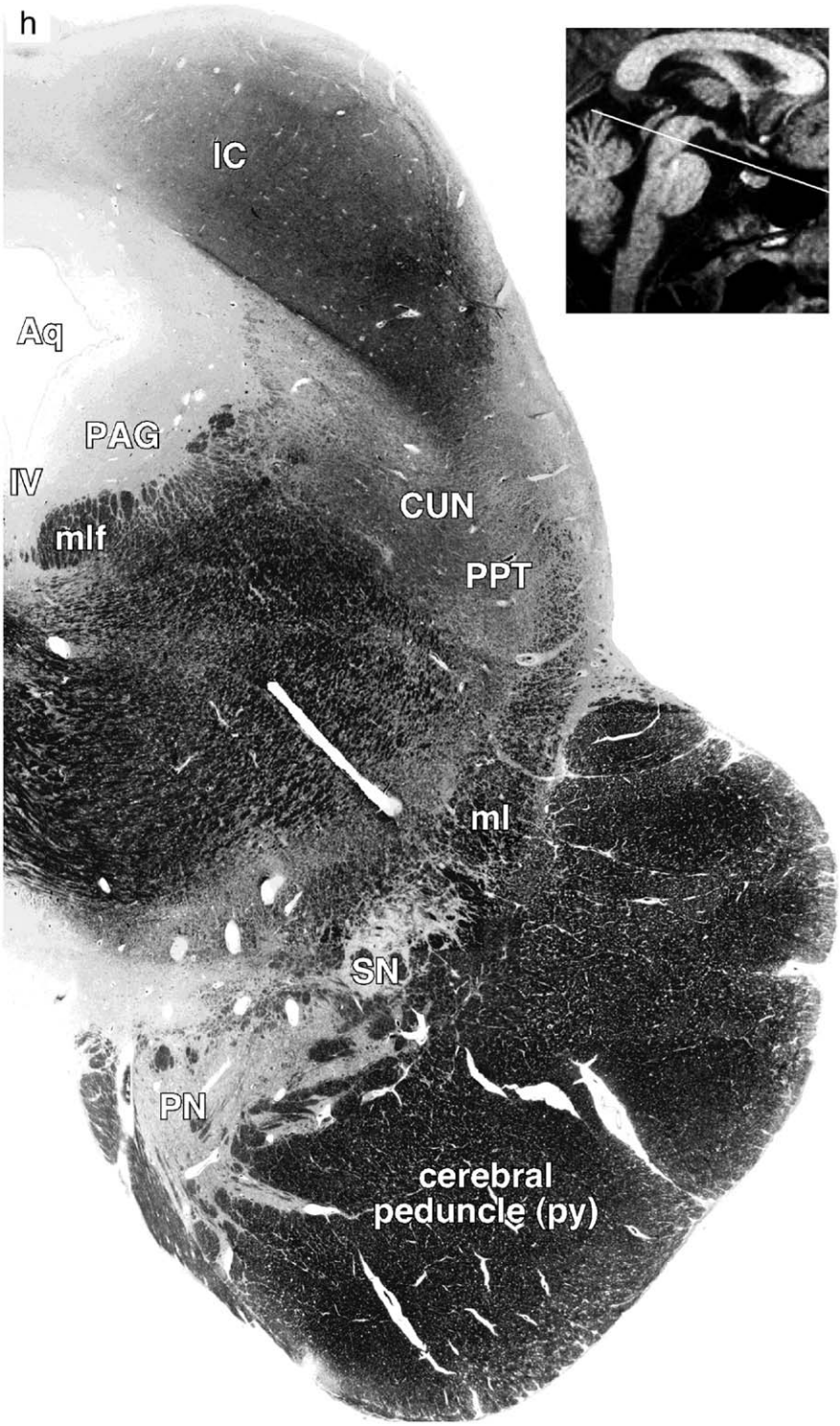


Figure 2 *(continued)*

Table I
Classification of Brain Stem Neurons

Somatic motoneurons

Peripheral striated muscle innervated	Cranial nerve	Brain stem nucleus
Extraocular muscles	III, IV, VI[a]	Edinger–Westphal, trochlear, abducent principal trigeminal motor nucleus
Jaw muscles	V[b]	
Mylohyoid, anterior digastric, tensor tympani, stapedius	V[b]	Accessory trigeminal motor nucleus
Stylohyoid, posterior digastric	VII[b]	Accessory facial nucleus
Facial expression	VII[b]	Main facial nucleus
Swallowing and phonation (muscles of pharynx, larynx, and upper part of esophagus)	IX, X[b]	Nucleus ambiguus (excluding external formation)
Tongue	XII[a]	Hypoglossal
Sternomastoid and trapezius	XI	Accessory

Parasympathetic motoneurons

Peripheral target	Final motoneuron	Cranial nerve	Brain stem nucleus
Ciliary and iris muscles	Ciliary ganglion	III	Edinger–Westphal nucleus
Lacrimal gland	Sphenopalatine ganglion	VII, IX	Neurons dorsal to rostral portion of facial nucleus
Salivary and mucosal glands	Cranial ganglia	VII, IX	Neurons dorsal to rostral portion of facial nucleus
Cranial blood vessels	Cranial ganglia	VII, IX	Probably with salivary and lacrimal preganglionic cells
Lower airways and lung	Ganglia in airways	X	Nucleus ambiguus, dorsal motor nucleus of vagus
Heart	Cardiac ganglia	X	Nucleus ambiguus (external formation)
Abdominal organs	Enteric neurons	X	Dorsal motor nucleus of the vagus

Secondary sensory neurons

Peripheral origin of primary afferent	Location of primary cell body	Cranial nerve	Brain stem nucleus in which primary afferents terminate
General cutaneous from anterior two-thirds of head	Cavum trigeminale	V, IX, X	Principal sensory nucleus of V
Joint position sense	In pons and midbrain	V	Mesencephalic nucleus of V
Nociception, temperature anterior two-thirds of head	Cavum trigeminale	V, IX, X	Spinal nucleus of V, possibly nucleus tractus solitarius
Hearing and vestibular function	Spiral and vestibular ganglia	VIII	Dorsal and ventral cochlear nuclei; medial, lateral, superior, and spinal vestibular nuclei
General cutaneous from posterior one-third of head and rest of body gustation	Dorsal root ganglia	Spinal dorsal roots	Gracile and cuneate nuclei
	Geniculate VII, petrosal IX, nodose X	VII, IX, X	nTS, spinal nucleus of V, and paratrigeminal islands
Baro- and chemoreceptors, receptors in heart, lung, and abdominal organs	Petrosal IX, nodose X	IX, X	nTS, spinal nucleus of V, and paratrigeminal islands

(continues)

Table I *(continued)*

Interneurons, including premotor neurons	
Type of neuron	**Brain stem nucleus**
Interneurons with no primary afferent input and no direct projections to motoneurons	Inferior olive, lateral reticular nucleus, cerebellar nuclei, vestibular nuclei, nuclei pontis, arcuate nuclei, superior olive, nucleus intercalatus, prepositus hypoglossi, locus coeruleus and subcoeruleus, dorsal tegmental nuclei, A7 catecholamine cells, parabrachial, Kölliker–Fuse, pedunculopontine and cuneiform nuclei, PAG
Premotor cells for cranial somatic motoneurons	Ventral pontine nuclei (relay nuclei between cerebral cortex and cerebellum), near various regions of the pons and medulla, many still undefined; neurons loosely scattered between fibers without formation of defined nuclei are referred to as the "reticular formation"
Premotor cells for cervical (phrenic) and thoracic spinal respiratory neurons	Some Kölliker–Fuse and parabrachial neurons, Bötzinger complex neurons, rostral inspiratory and more caudal expiratory neurons in the ventrolateral medulla, some nucleus tractus solitarius; raphe magnus, parapyramidal and more caudal raphe neurons
Presympathetic motoneurons	Paraventricular nucleus of hypothalamus, A5 catecholamine cells, C1 catecholamine cells and other intermingled noncatecholamine cells, raphe magnus, parapyramidal and more caudal raphe neurons
Preparasympathetic motoneurons (cranial outflow)	See details in text
Preparasympathetic motoneurons (sacral spinal outflow)	Raphe and parapyramidal nuclei, rostral ventrolateral medulla, A5 region, Barrington's nucleus, and paraventricular and preoptic nuclei of the hypothalamus
Premotor cells for hypothalamic magnocellular neurons	A1 and A2 catecholamine-synthesizing neurons, and possibly some midbrain raphe neurons

[a]General somatic efferents.
[b]Special visceral efferents.

The medial longitudinal faciculi are paired paramidline tracts containing axons which interconnect cranial nerve nuclei, especially those concerned with horizontal eye movements.

III. BLOOD SUPPLY AND CEREBROSPINAL FLUID CIRCULATION: BLOOD–BRAIN BARRIER, AREA POSTREMA, AND VENTRAL SURFACE OF THE MEDULLA

The brain stem, up to the level of the rostral midbrain, is supplied by the single midline basilar artery, formed by union of the paired vertebral arteries. At its rostral extent, the basilar artery bifurcates into paired posterior cerebral arteries which continue on to supply the medial temporal lobe and the occipital poles of the cerebral hemispheres. This knowledge is very important in clinical practice because patients with brain stem ischemic stroke (vertebrobasilar ischemia) present with a clinical picture (see Section V) quite different from that displayed by patients with ischemia in forebrain territory supplied by the paired internal carotid arteries. The posterior communicating arteries connect the vertebrobasilar arterial supply with the internal carotid supply, forming part of the circle of Willis.

Cerebrospinal fluid (CSF) is filtered from the blood by the choroid plexus in the lateral, third, and fourth ventricles of the brain. Fluid secreted in the forebrain descends through the midbrain aqueduct and enters the large dorsal brain stem pool bounded anteriorly by the dorsal surface of the medulla oblongata and the pons (known as the floor of the fourth ventricle) and posteriorly by the cerebellum and by special folds of the meninges. These folds have openings (formamina of Luschka and Magendie) through which CSF can pass en route to the subarachnoid space surrounding the brain and to the dorsally positioned arachnoid villae which absorb the CSF back into the bloodstream. From the fourth ventricle, the CSF also passes caudally down the central canal of the spinal cord. Occlusion of the aqueduct or of the foramina of Luschka and Magendi obstructs the flow of CSF, causing increased pressure in rostral portions of the system so that the forebrain ventricles enlarge. This condition is known as hydrocephalus and it may be associated with drowsiness or coma.

The "inside" of the brain (the brain parenchyma and the cerebrospinal fluid) is chemically isolated from the "outside" of the brain (the cerebral blood vessels). A particular configuration of the basement membranes in the small cerebal blood vessels results in a blood–brain

Table II
Cranial Nerves

Nerve No.	Nerve name[a]	Site for CNS motoneurons	CNS site for termination of primary sensory axons
I	Olfactory (S)		Forebrain
II	Optic (S)		Forebrain (thalamus)
III	Oculomotor (M)	Dorsomedial midbrain	—
IV	Trochlear (M)	Dorsomedial midbrain	—
V	Trigeminal (S, M)	Pons	Pons, medulla, and upper cervical spinal cord
VI	Abducent (M)	Dorsomedial pons	—
VII	Facial (S, M)	Ventrolateral pons	Nucleus tractus solitarius
VIII	Vestibulocochlear (S)		Vestibular nucleus in dorsolateral medulla
IX	Glossopharyngeal (S, M)	Ventrolateral pons and nucleus ambiguus	Nucleus tractus solitarius
X	Vagus (S, M)	Dorsal motor nucleus of the vagus and nucleus ambiguus	Nucleus tractus solitarius
XI	Accessory (M)	Nucleus ambiguus and upper spinal cord	—
XII	Hypoglossal (M)	Hypoglossal nucleus in dorsomedial medulla	—

[a]M, motor; S, sensory.

barrier that protects the brain from circulating agents which might interfere with normal neurotransmitter action. One particular group of cells on the dorsal aspect of the medulla oblongata forms the area postrema, a specialized circumventricular organ which samples the blood side of the blood–brain barrier and sends action potential messages to the brain, with axons of the area postrema neurons projecting widely within the brain stem, particularly to the parabrachial nuclei.

Cells lining certain portions of the ventral surface of the medulla oblongata, either glial cells or neurons, may be sensitive to different properties of the CSF, especially its hydrogen ion concentration (pH). Some theories of the neuropathological mechanism underlying sudden infant death syndrome (cot death) postulate that the normal sensitivity of these cells is somehow reduced so that when the airway becomes obstructed, affected infants fail to arouse, to turn their head, and to increase their breathing in the normal manner.

IV. FUNCTIONS REGULATED IN THE BRAIN STEM

A. Normal "Vigilant" Consciousness and the Sleep–Wake Cycle

Many lines of evidence point to the importance of the brain stem in maintaining the normal awake state of awareness. Humans with damage to the region of the dorsal pons, midbrain, and thalamus (by trauma, brain tumor, viral or bacterial infection, or ischemic or hemorrhagic stroke) may exhibit an impaired state of alertness, possibly becoming stuporose or comatose. In animals, experimental transection at various levels of the spinal cord and brain stem established that coma ensued after lesions through the level of the colliculi, but not after lesions through the level of the medullospinal junction. In intact anesthetized animals, electrical stimulation in the brain stem caused the neocortical electroencephalogram to change to a pattern normally associated with alertness. This activation probably reflects stimulation of ascending axons arising in noradrenaline neurons in the locus coeruleus, serotonin neurons in the median and dorsal raphe nuclei, and acetylcholine neurons in the pedunculopontine tegmental nuclei (Figs. 2f and 2g).

The locus coeruleus (blue place) consists of a group of pigmented neurons (melanin-containing) located in the dorsolateral portion of the rostral pons. These noradrenaline-synthesizing neurons have extensive efferent connections to the forebrain, including the thalamus and the cerebral cortex, and to the cerebellum and the spinal cord. Activity of neurons in the locus coeruleus is increased when the individual detects a significant, potentially rewarding or threatening stimulus as well as during certain phases of the sleep–wake cycle. It is thought that locus coeruleus inputs to

forebrain neurons increase the efficiency with which the innervated neurons respond to particular environmental stimuli.

B. Control of Pupil Size

Our pupils become larger when we move to a darker environment, and they constrict again when we move back into the light. This light reflex is mediated via contraction and relaxation of the sphincter pupillae muscle, innervated by parasympathetic final motorneurons located in the ciliary ganglia behind the eyeballs. In turn, the final motoneurons are regulated via inputs from brain stem parasympathetic motoneurons in the Edinger–Westfal subdivision of the oculomotor (cranial III) nucleus located in the midbrain, ventral to the aqueduct. The afferent limb of the light reflex involves activity in optic nerve (cranial II) fibers, with a branch to the pretectal nucleus in the superior colliculus of the midbrain and from there to the parasympathetic motorneurons in the Edinger–Westfal nucleus.

The pupils also dilate when strong emotions are experienced, such as when we are surprised, afraid, or sexually aroused. This active dilatation is mediated by sympathetic innervation of the dilator pupillae muscle. The final cell bodies are located in the superior cervical ganglion, near the bifurcation of the common carotid artery. The preganglionic cell bodies are in the upper thoracic spinal cord. The pathway descending from the brain is a principally unilateral direct projection from the paraventricular nucleus of the hypothalamus. Interruption of the pupillodilator pathway in the brain stem, the spinal cord, or in the periphery causes an ipsilateral Horner's syndrome, consisting of a small pupil (meiosis) and slight retraction of the eyeball associated with partial closure of the eyelids. The drooping upper eyelid is referred to as a ptosis.

C. Moving the Eyes

1. Voluntary Movement of the Eyes

The eyeballs move when the extraocular striated muscles contract. Medial movement is mediated by the medial rectus muscle, innervated by somatic motor axons traveling in cranial nerve III (the oculomotor nerve). Lateral movement is mediated by the lateral rectus muscle, innervated by axons which travel in cranial VI (the abducent nerve). Upward movement is mediated by the superior rectus and inferior obliques muscles (both innervated via cranial III) and downward movement is mediated by the inferior rectus muscle (innervated via cranial III) and by the superior oblique muscle (innervated via cranial IV, the trochlear nerve). The brain stem location of the various groups of motoneurons is shown in Fig. 2.

The eyeballs normally move in a mutually coordinated manner. When, for example, the individual voluntarily looks to the right, the lateral rectus muscle of the right eyeball contracts at the same time as the medial rectus muscle of the left eyeball. This horizontal conjugate eye movement depends on coordinated activity in the right abducent nerve (cranial VI) and the relevant fibers in the left oculomotor nerve (cranial III). This coordination is achieved via activity descending from the left cerebral cortex to controlling premotor and interneurons in the brain stem, as outlined in Fig. 3. Vertical eye movements are coordinated by premotor neurons in the upper midbrain.

2. Reflex Movements of the Eyes by the Vestibuloocular Reflex

In a dead person, the eyes stay still in relation to the head so that when the head is held and turned to the left, the eyes move together with the head. In an alive but unconscious person with an intact brain stem, when the examining doctor turns the person's head to the left, the functioning vestibuloocular reflex turns the eyes to the right so that they continue to look straight ahead. This compensatory gyroscope-like adjustment of the eye position is known as a dolls-eye movement. It also occurs in the conscious individual, facilitating fixation of the gaze when the head is moving.

Head movements are detected by the afferent neurons innervating the canal system of the vestibular–labyrinthine apparatus (Fig. 3). For example, if an individual is looking at an object straight in front of him or her and the head is suddenly turned to the left, the left horizontal canal primary sensory neurons will increase their discharge rate. Axons of these neurons enter the brain stem via cranial nerve VIII and excite second-order afferent neurons in the vestibular nuclei of the brain stem (Figs. 2c and 2d). These second-order vestibular neurons project to the contralateral abducens motoneurons (innervation of the right lateral rectus muscle) and to the contralateral abducens interneurons, whose axons cross the midline and ascend in the medial longitudinal fasciculus to the oculomotoneurons (cranial III) that innervate the medial rectus muscle. Thus, turning the head to

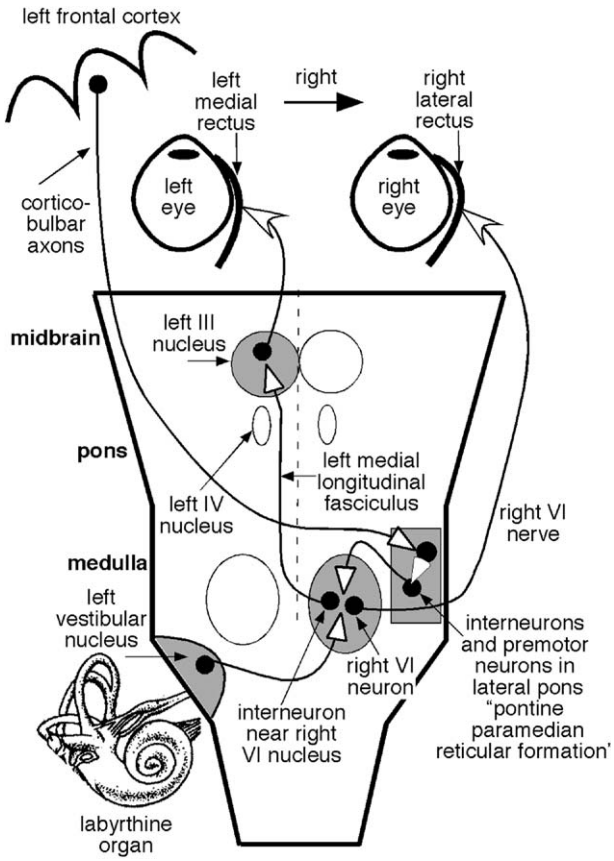


Figure 3 Schematic horizontal section through the human brain stem showing the oculomotor system for voluntary gaze to the right and for movement of the eyes to the right during the vestibuloocular reflex.

the left reflexly leads to contraction of the right lateral rectus and the left medial rectus muscles, turning the eyes to the right.

D. Body Movement and Posture

The expression "running around like a headless chicken" is based on a stark reality. Clearly, many vertebrates can locomote via motor pattern generators intrinsic to the spinal cord. The presence of such segmental and intersegmental spinal circuitry in mammals is evident when a cat with a lower cervical spinal transection is supported and placed on a treadmill. The animal's legs move in a coordinated walking fashion. Control of limb movement is embedded in neural circuitry present at different levels of the neuraxis, with the system operating like embedded procedure loops in a computer program. Each higher level of the nervous system represents a more complex loop, including and controlling the less complex loops. The brain stem, without the forebrain, is capable of coordinating complex movements, including some of the more integrated movements we recognize as behaviors. Decerebrate rats are quite capable of grooming themselves. A decerebrate dog may turn, growl, and bite the fingers holding his hind foot too roughly. Clearly, forebrain neural circuits, including frontal and prefrontal pathways, are necessary if the individual is to accomplish the finely patterned movements of eyes, lips, tongue, and fingers which especially characterize human action. If damage occurs at the midbrain level of the neuraxis (which sometimes occurs with head injury in motor vehicle accidents), the decerebrate individual subjected to a painful stimulus (e.g., a firm pinch on the chest area) may simply increase the breathing rate and internally rotate the extended arms in a purposeless fashion.

Brain stem nuclei with projections to spinal premotor neurons include the red nuclei in the midbrain (rubrospinal), the vestibular nuclei in the rostral medulla and pons (vestibulospinal), the locus coeruleus, and the raphe and parapryamidal nuclei in the pons and medulla. In addition, there are descending motor projections from neurons in various parts of the brain stem reticular formation (collectively named reticulospinal). The brain stem nuclei which innervate spinal motor centers also project to relevant premotor regions in the brain stem, but these pathways are more difficult to delineate and they are usually not referred to by any particular name. The functions of these various brain stem nuclei, and the contribution of each to body posture and movement, are still being defined.

The medulla, pons, and midbrain have numerous interconnections with the cerebellum, a part of the brain intimately involved with coordination of posture and movement. For example, the inferior olivary nucleus in the ventral portion of the medulla oblongata, contains densely packed neurons which project across the midline, enter the inferior cerebellar peduncle, and synapse on neurons in the cerebellar cortex. The information conveyed in these "climbing fibers" derives from various forebrain, brain stem (including the red nucleus in the midbrain), and spinal inputs to the inferior olivary nucleus.

E. Touch, Temperature, and Pain

As noted earlier, the brain stem contains circuitry which subserves "spinal cord functions" for the head

region. General sensation from the face and head region is mediated via trigeminal (cranial V) afferents with perikarya in the trigeminal ganglion. The central process of the trigeminal neuron enters the brain stem at the level of the ventrolateral pons. Myelinated afferents subserving light touch in the face and upper head region synapse in the principal sensory nucleus of V. Light touch and other finely judged senses (vibration and joint position) from the rest of the body are mediated by neurons located in the dorsal root ganglia of the spinal cord. Central processes of these neurons enter the dorsal columns of the spinal cord and ascend to synapse in the gracile and cuneate nuclei in the dorsal medulla oblongata. Axons of the gracile and cuneate cells pass ventrally, cross the midline, and ascend to the thalamus as the bundle constituting the medial lemniscus. Fibers from neurons in the principal sensory nucleus of V also travel to the thalamus as part of the medial lemniscus.

An unusual neuronanatomical arrangement has been described for trigeminal neurons mediating joint position sense for the jaw and the tongue. Perikarya for these primary afferents are located in the brain stem, in the mesencephalic nucleus of V (Fig. 2g).

Pain and temperature sensation is mediated by unmyelinated axons of neurons located in the cavum trigeminale (for the face and head) and in the dorsal root ganglia (for the rest of the body). For the body apart from the face and head, central processes of these neurons synapse in the ipsilateral dorsal horn of the spinal cord. Axons of secondary sensory neurons cross the midline and ascend to the thalamus as the spinothalamic tract. Unmyelinated afferents subserving temperature and pain sensation from the face and head region travel caudally in the spinal tract of V before synapsing in the spinal nucleus of V, sometimes as far caudally as the upper cervical spinal cord. Axons from the secondary sensory axons cross the midline and ascend to the thalamus along with the spinothalamic pathway.

F. Hearing

Sound energy vibrates the tympanic membrane. This movement is amplified by the bones of the middle ear so that vibrations occur in a special fluid-filled organ known as the cochlear. Specialized cochlear receptor cells are innervated by sensory axons of the primary auditory neurons (cranial VIII) located in a peripheral ganglion (spiral ganglion) near the inner ear. Centrally directed afferent axons of these neurons enter the brain stem and synapse with secondary sensory neurons located in the cochlear nuclei on the dorsolateral aspect of the rostral medulla oblongata (Figs. 2c and 2d). The secondary sensory neurons have axonal projections either via the lateral lemniscus directly to the inferior colliculus or via a synapse in the superior olivary nucleus and then via the lateral lemniscus to the inferior colliculus. From the inferior colliculus, ascending axons project to the medial geniculate nucleus of the thalamus and then to the auditory cortex.

G. Sensing Gravity and Movement

The inner ear contains a fluid-filled labyrinth with associated cells sensitive to the orientation and movement of the individual with respect to the outside world. The labyrinthine cells are innervated by sensory axonal processes of nerve cells located in a ganglion near the inner ear (Scarpa's ganglion). Central processes of these afferent neurons reach the brain stem by way of the vestibular branch of cranial nerve VIII and synapse with secondary sensory neurons in the vestibular nuclei located in the dorsolateral portion of the rostral medulla oblongata (Figs. 2c and 2d). In turn, the neurons in the vestibular nuclei project widely to brain stem nuclei (including those regulating eye movements) as well as to the spinal cord motor control regions. The particular vestibular projection which gives rise to the feeling of nausea associated with motion sickness has not been characterized.

H. Lacrimation

The lacrimal (tear) glands and specialized cells in the lining of the nose secrete fluid filtered from the blood, a process regulated by parasympathetic final motoneurons located in the sphenopalatine ganglion. These cells are innervated by axons traveling in the greater petrosal branch of the facial nerve (cranial VII). Brain stem parasympathetic motoneurons are located in the reticular formation, dorsal to the facial nucleus.

I. Tasting

Taste receptor cells in the mouth, tongue, palate, and pharynx are innervated by axons of afferent neurons with cell bodies located in peripheral ganglia

associated with facial, glossopharyngeal, and vagal cranial nerves. Centrally directed axons enter the medulla oblongata and synapse with secondary sensory neurons in the rostral portion of the nucleus of the tractus solitarius (Fig. 2c). Ascending axons of these secondary cells synapse in the parabrachial nucleus (pontine taste center) and also possibly project directly to the thalamus. Neurons in the parabrachial nucleus have extensive rostral connections with forebrain regions, including the amygdala and taste areas of the cortex.

J. Salivation

Saliva is fluid filtered from blood in the various salivary glands and secreted into the mouth via the different salivary ducts. Salivation is initiated by a combination of parasympathetic and sympathetic nerve activity. The parasympathetic final motoneurons are located in ganglia (otic, submandibular, and submaxillary) associated with the facial (cranial VII) and glossopharyngeal (cranial VIII) nerves and along the salivary ducts. The brain stem parasympathetic motoneurons regulating the salivary glands are located in the rostral medulla and caudal pons, diffusely spread in the reticular formation through a region extending dorsally and medially from the facial motor nucleus (cranial VII). Inputs from the forebrain to these parasympathetic neurons presumably mediate the ''conditioned'' salivation that occurs when we catch sight of food or when we think about food. Presumably, there are reasonably direct neural connections between the taste regions of the nucleus of the tractus solitarius and the parasympathetic salivary neurons, but these connections are not yet fully defined.

K. Chewing and Swallowing

Chewing and swallowing of food requires coordinated action of facial, jaw, tongue, soft palate, pharyngeal, and upper esophageal striated muscles. The relevant somatic lower motoneurons are located in a rostrocaudal column of cells, situated in the ventrolateral pons and medulla (VII, IX, and XI), in a more dorsomedial column of cells, just ventral to the central canal and the floor of the fourth ventricle (XII). Details of the premotor neurons and interneurons responsible for coordinating the activity of the different lower motoneurons are sparse. The relevant cells, situated in the reticular formation ventrolateral to the hypoglossal nucleus, presumably have built-in coordinating programs. These neurons must receive input from the forebrain (for voluntary chewing and swallowing) and from the nucleus of the tractus solitarius, the termination site of glossopharyngeal and vagal afferents originating in the pharynx and the esophagus (for reflex swallowing, among many other things).

L. Gastrointestinal Function, Blood Glucose, and Food Intake

Preganglionic axons of parasympathetic neurons in the dorsal motor nucleus of the vagus project via cranial nerve X to ganglia in the wall of the gastrointestinal tract (the enteric nervous system) and to other parasympathetic ganglia in the region of the different abdominal viscera. The vagal motoneurons play a very important role in regulating motility (e.g., stomach contractions) and secretory functions (e.g., gastric acid) of these organs. The central nervous system coordinates the function of the gastrointestinal tract and the other viscera, coordinating the behavioral and the physiological responses associated with nutritional aspects of daily life. As Pavlov demonstrated, the sight or smell of food are powerful stimuli for gastric acid secretion, a process mediated by the appropriate vagal efferents.

Afferent/sensory vagal axons also innervate the various abdominal organs, with the central process of the vagal afferent neurons projecting to the nucleus of the tractus solitarius in the dorsal part of the medulla oblongata. Different vagal afferents are sensitive to stretch or to the local chemical environment. After a meal, products of food digestion activate special intestinal cells containing the hormone cholecystokinin. In addition to its well-known effects on gallbladder contraction (hence its name), cholecystokinin, by a local mechanism, activates vagal afferents in the duodenum. The information which reaches the medulla oblongata initiates a whole sequence of postprandial (after eating) ''satiety'' events. The individual may feel full, stop eating, and go to sleep. Postprandial satiety is not due to an insufficient amount of blood getting to the brain (a commonly held view). It is principally due to the action of specially stimulated vagal afferents.

If blood sugar falls to low values, this is detected by special sensory vagal nerve endings in the periphery (mainly in the liver) and neurons in the nucleus of the

tractus solitarius. One result is the secretion of the hormone adrenaline from the adrenal gland. Control of this secretion is principally via sympathetic nerves, with cell bodies in the thoracic spinal cord, controlled by presympathetic neurons in the rostral ventrolateral medulla oblongata. Neurons in the nucleus tractus solitarius project to the rostral ventrolateral medulla.

M. Vomiting

The act of vomiting involves patterned contraction of striated muscles of the abdominal wall, pharynx, larynx, tongue, and face, together with a reversal of normal smooth muscle action in the upper intestine and stomach and a relaxation of the esophageal sphincter. This complex sequence of events can be initiated either in the gastrointestinal tract or in the brain. Axons of the vagus nerve innervate the upper gastrointestinal tract. Serotonin, released by emetic agents from small cells in the intestinal epithelium, stimulates serotonin receptors on the distal endings of the vagal afferents, causing them to discharge. The central processes of the relevant vagal afferents project into the brain and synapse in the nucleus of the tractus solitarius just rostral to the obex, with probable collateral input to the area postrema. Currently, there is insufficient information concerning central interneuronal pathways linking the secondary sensory neurons in the nucleus of the tractus solitarius with the premotor neurons controlling the relevant motor pathways involved in the act of vomiting. Also, it is not understood how nausea and vomiting due to gastrointestinal upset interact with those which occur as part of motion sickness, when the initial abnormal signal presumably originates in the labyrinthine system.

Activation of special neurons in the area postrema can also elicit vomiting. The nerve cells in the area postrema are in a special relation to a dense network of capillaries so that the usual barrier between blood and brain is not operative. Nerve cells in this region are directly exposed to circulating chemical agents. Many drugs which induce vomiting act by stimulating receptors, especially dopamine receptors, on these neurons. The area postrema cells project widely in the brain, but their medullary connections presumably include either the nTS neurons, which receive the vagal inputs from the upper gastrointestinal tract, or the same interneurons and premotor neurons which mediate vagally induced vomiting.

The survival value of gastrointestinally elicited vomiting is obvious. Why whole body radiation and cancer chemotherapeutic agents affect the serotonin cells in the upper gastrointestinal tract is not clear. Nevertheless, drugs which block serotonin receptors on the distal endings of the vagal afferents have proven of great benefit in reducing the nausea and vomiting associated with anticancer treatments.

N. Passing Urine

Passing urine (micturition) depends on activation of appropriate parasympathetic motoneurons in the sacral region of the spinal cord. Distention of the bladder activates afferent nerves whose spinal input tends to initiate micturition, but efficient emptying of the bladder requires descending activation which originates in brain stem preparasympathetic neurons located in the pontine micturition center, probably comprising a small group of neurons (Barrington's nucleus) located just medial to the locus coeruleus. Normally, micturition is delayed by descending inhibitory activity originating in the cerebral cortex.

O. Blood Flow, Arterial Pressure, and Cardiac Function

1. Presympathetic Vasomotor Neurons in the Brain Stem and the Hypothalamus

When the cervical or upper thoracic spinal cord is transected, the arterial blood pressure decreases to very low levels because the rostral ventrolateral medulla oblongata contains presympathetic vasomotor neurons whose activity tonically excites the spinal preganglionic sympathetic neurons so that regional vasomotor tone and arterial pressure are maintained. The activity of the medullary neurons is maintained either by intrinsic pacemaker activity or by interactions between different neuronal groups in a manner which currently is not understood. The presympathetic vasomotor neurons in the rostral medulla include the C1 catecholamine neurons (with the requisite enzymes to synthesize adrenaline being present in several species, including rats, cats, and humans) and an approximately equal number of noncatecholamine bulbospinal neurons. More medially placed bulbospinal neurons in the raphe and parapyramidal region have a role in regulating the changes in skin blood flow associated with body temperature regulation and with

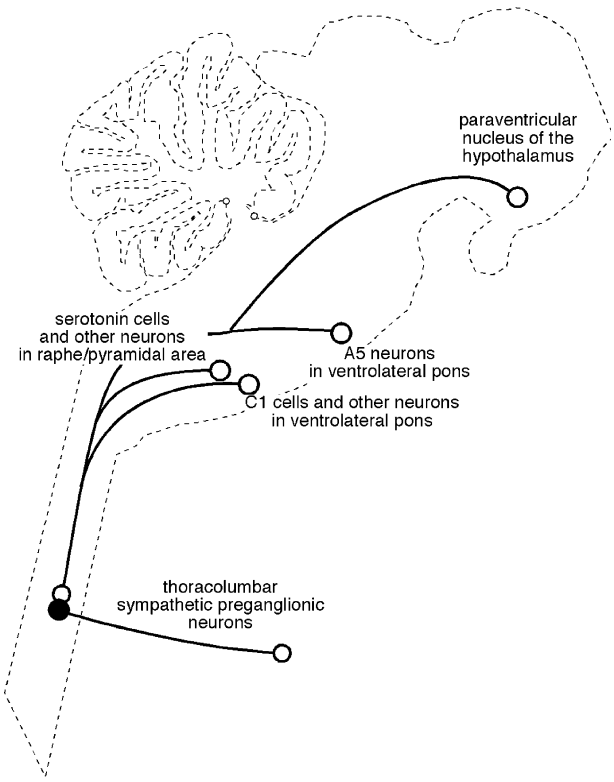


Figure 4 Schematic sagittal section through a rabbit brain showing the distribution of neurons (presympathetic motoneurons) with descending axons that innervate sympathetic preganglionic motoneurons in the intermediolateral columns of the thoracic and upper lumbar spinal cord. (modified with permission from Blessing, 1997).

cutaneous vasoconstriction in response to frightening or painful events. Other regions of the brain stem and the hypothalamus also contain presympathetic motoneurons (Fig. 4), but these cells have not been functionally characterized.

2. Inhibitory Cardiovascular Neurons in the Caudal Ventrolateral Medulla

The caudal ventrolateral medulla contains a group of GABAergic neurons with short axons projecting rostrally to innervate presympathetic vasomotor neurons in the rostral medulla. Discharge of the caudal vasomotor neurons inhibits the excitatory rostral bulbospinal neurons, thereby lowering arterial pressure. Since the caudal depressor neurons are tonically active, interference with their function removes the inhibitory control so that arterial blood pressure rises. It is possible that malfunction of the caudal vasodepressor neurons underlies some forms of hypertension.

3. Cardiovascular Secondary Afferent Neurons in the Nucleus of the Tractus Solitarius

Cranial nerves IX (glossopharyngeal) and X (vagus) include afferent neurons (cell bodies in nodose and petrosal ganglia) with distal processes specialized for the detection of arterial pressure, cardiac atrial pressure, arterial oxygen and content and blood acidity, and other chemoreceptors present in the lung and the heart. The central processes of these neurons project into the medulla oblongata as the tractus solitarius, synapsing principally in the caudal one-half of the nucleus of the tractus solitarius in the dorsal region of the medulla oblongata.

4. CNS Pathway Mediating the Baroreceptor–Vasomotor Reflex

When arterial pressure rises, the baroreceptor afferent neurons increase their discharge rate, thereby increasing the discharge of secondary afferent neurons in the nucleus of the tractus solitarius. These cells project to subclasses of the inhibitory vasomotor cells in the caudal ventrolateral medulla so that activation of the nucleus tractus solitarius cells increases the discharge of the inhibitory vasomotor neurons in the caudal ventrolateral medulla, reducing the activity of the sympathoexcitatory bulbospinal in the rostral medulla neurons and thereby lowering arterial pressure.

P. Breathing

When we breath in, the lungs inflate because contraction of the diaphragm and other respiratory chest muscles means that the pressure around the lungs becomes less than atmospheric pressure. The inspiratory respiratory muscles are innervated by somatic motor nerves (including the phrenic nerve) whose cell bodies are located in the C3 and C4 segments of the cervical spinal cord. These spinal motoneurons have no spontaneous activity. Transection of the upper cervical cord isolates them from descending excitatory control so that breathing ceases and survival means artificial respiration via a tracheal tube (positive pressure to the lung) or with an iron lung (negative pressure around the lung). Stimulating the phrenic nerve via an implanted pacemaker may also be possible.

The neurons that project from the brain to the cervical spinal cord to regulate spontaneous inspiration are located in the medulla oblongata, in the

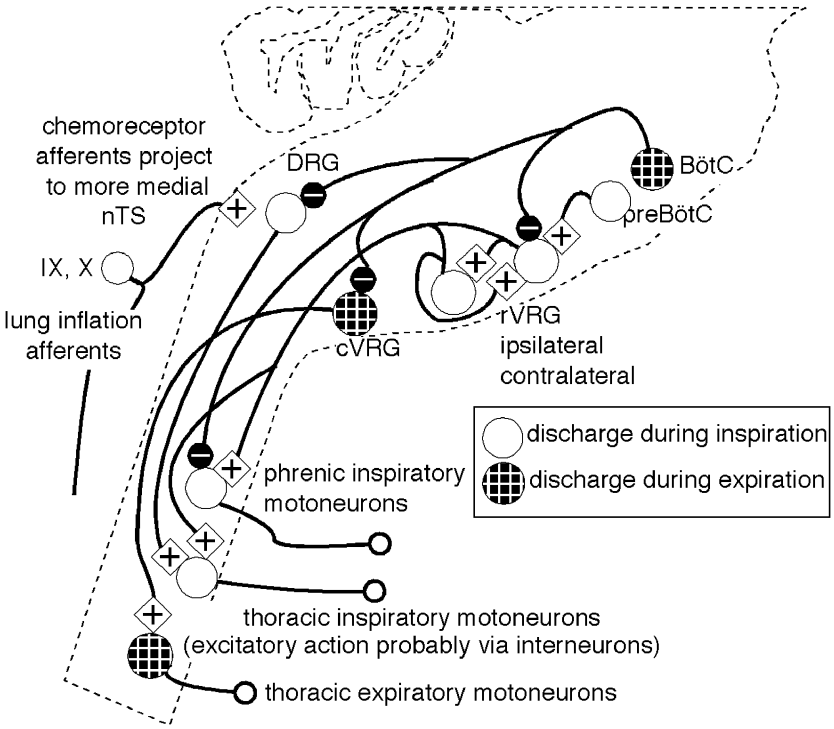


Figure 5 Schematic sagittal section of the rat brain stem summarizing the neuroanatomical and functional organization of the brain stem neural circuitry responsible for breathing (modified with permission from Blessing, 1997).

so-called rostral ventral respiratory group (rVRG). In turn, activity of rVRG neurons depends on net excitation from other medullary neurons either from the combined activity of a network of respiratory neurons or from the pacemaker activity of a particular group of respiratory neurons, probably the pre-Botzinger cells located in the ventrolateral medulla, rostral to the rVRG cells (Fig. 5).

During expiration, activity in rVRG neurons ceases so that the diaphragm and other inspiratory chest wall muscles no longer contract. Since lung tissue is elastic, the lungs deflate by themselves so the first part of expiration is passive. The final part of normal expiration, as well as lung deflation during forced expiration, depends on active contraction of the expiratory muscles in the chest and abdominal regions. These muscles are also innervated by spinal motoneurons activated by descending inputs from special expiratory motoneurons in the medulla oblongata, the Botzinger group, and the caudal ventral respiratory group, as shown in Fig. 5.

Other groups of respiratory neurons in the pons and the midbrain help to coordinate the activity of the medullary network. The cerebral hemispheres must contain the respiratory neurons which are responsible for voluntary breathing and for the voluntary modulation of breathing that occurs during activities such as speaking, singing, clearing one's throat, or blowing out a candle. The descending axons of these cells synapse either on unknown interneurons in the medulla oblongata or perhaps directly on rVRG neurons.

V. BRAIN STEM DYSFUNCTION IN DISEASE

Most people are used to the idea that dysfunction of one cerebral hemisphere can lead to paralysis of the other side of the body. This may occur, for example, with ischemia of brain tissue supplied by the internal carotid artery or the middle cerebral artery. Sudden impairment of the blood supply is also a common cause of brain stem dysfunction, but in this case it is the vertebrobasilar system of arteries that is at fault. Other disease processes (infection, neoplasm, and inflammation) may also affect brain stem function. The neurological symptoms and signs of brain stem dysfunction depend on the precise subregion affected.

Dysfunction of the upper brain stem (especially the more dorsal portions of the rostral pons), the midbrain region just ventral to the aqueduct, and the thalamus can cause the patient to be drowsy, stuporose, or unconscious. This presumably reflects damage to the ascending pathways described in Section IV.A.

Impairment of brain stem eye movement control systems includes supranuclear palsies (upper motoneuron lesions), internuclear ophthalmoplegias, and strabysmus or squint (crooked eyes) reflecting dysfunction of the relevant lower motoneurons or of the oculomotor, trochlear, and abducens nerves (cranial nerves III, IV, and VI) or dysfunction of the muscles moving the eyeballs.

In a supranuclear palsy, as occurs in a neurodegenerative disease known as Steele–Richardson syndrome or progressive supranuclear palsy, both eyes move conjugately (together) but particular directions of movement (particularly looking up) are poorly executed. Internuclear ophthalmoplegia reflects damage to the medial longitudinal fasciculus, a bundle of axons coordinating the relevant pontine and midbrain motoneurons, as may occur with small demyelinating lesions in the disease known as multiple sclerosis. When the patient tries to look to the right, the left eyeball fails to turn inward (adduct) and the outturned right eyeball develops nystagmus.

Young children with a chronic strabysmus must be treated or the brain will suppress the image from one eye (a condition known as amblyopia) so that the individual never develops stereoscopic (binocular) vision. Acute lower motoneuron lesions usually cause double vision (diplopia), which is worse when the patient attempts to look in the direction controlled by the affected nerve. This symptom occurs partially because there is minimal feedback to the brain concerning the actual spatial position of the eyeball (contrast, for example, the accuracy with which we can judge the spatial position of the tip of the tongue). The brain sends the appropriate movement commands to the cranial motoneurons controlling the eyeball muscles and then interprets visual inputs as if the commands have been successfully executed. If one eyeball is misaligned, the brain makes a mistake and the result is double vision, not a sensation that the eyeball is in the wrong position.

A person may suddenly lose the ability to move one side of the face so that it may be impossible to close one eye or to wrinkle up the forehead on one side. The corner of the mouth may droop on one side. This condition may occur with damage to cranial nerve VII (facial) in its course within the pons or, more commonly in the condition known as Bell's palsy, when a peripheral segment of nerve VII is inflamed by a viral infection.

Dysfunction of the vestibular apparatus or its connecting brain stem pathways can induce the sensation of vertigo, sometimes associated with nausea and vomiting. Vertigo, tinnitus, and deafness also occur in Mennier's disease or with an acoustic neuroma, a tumor of the intracranial portion of the vestibulocochlear nerve (cranial nerve VIII). Brain stem dysfunction can also cause difficulty with speaking or swallowing.

The pathways connecting forebrain regions with the spinal cord can also be affected during their course through the brain stem. The major motor pathways from each side of the forebrain travel down the front of the midbrain, pons, and medulla, close to the midline, so that basilar artery ischemia may affect both these "pyramidal tracts," resulting in motor paralysis on both sides of the body. In severe brain stem dysfunction, the patient may be unable to move any one of the four limbs (quadraplegia) or even the head or any of the facial muscles so that the only way the patient can signal "yes" or "no" is by moving the eyes from side to side. Such patients, with normal consciousness and normal higher intellectual function, are described as being "locked in."

Locked in patients are quite different from patients in the so-called "chronic vegetative state." This condition follows damage to the dorsal portion of the upper part of the brain stem with preservation of function in the pons and medulla. Patients in the chronic vegetative state have no voluntary movement and, presumably, no consciousness or thought processes. However, they have reflex movements and they can still breath and swallow. Their cardiovascular, renal, and gastrointestinal function is stable. If such patients are fed with nutritious food and their bowel and bladder wastes are dealt with, they may survive for years. In such patients, the preservation in the limbs of complex motor responses, such as movement of the arms during yawning, may give false hopes to relatives watching at the bedside.

Some patients do recover from a vegetative-like state so it may be difficult to decide whether or not to render life support to a particular patient for an extended time. A useful guide can be obtained from a knowledge of the cause of the brain stem injury, aided by modern investigative techniques such as magnetic resonance imaging (MRI). If a person becomes unconscious from a viral infection in the upper brain stem and the MRI is normal, then there is a chance of

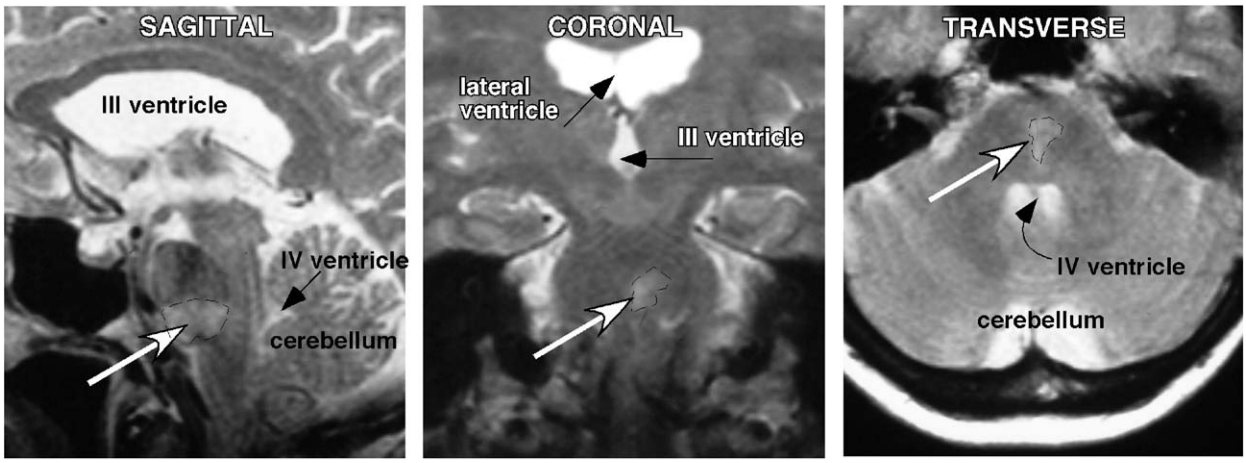


Figure 6 Magnetic resonance images in sagittal, coronal, and transverse planes through the brain stem of a 62-year-old woman with an ischemic stroke in the medial pons (arrows). The infarcted area of the brain is indicated by the dotted lines (MRI courtesy of the Department of Diagnostic Imaging, Flinders Medical Centre, Adelaide).

recovery to a normal life even if the person remains unconscious for months. If a motor vehicle accident or an ischemic stroke is the cause of the upper brain stem dysfunction, and the MRI shows strong evidence of structural damage, then recovery after being unconscious for as long as 1 month or even 1 week is extremely unlikely.

Extensive damage to the lower brain stem by any disease process usually ends the person's life because neural circuitry mediating vital respiratory and/or cardiovascular control no longer functions.

Sometimes, a small area of the brain stem can be affected by a focal disease process. The precise location of the damage can often be specified by particular combinations of physical signs, as in the following case: A 62-year-old woman presented with onset of motor dysfunction of her right hand and right leg (right hemiparesis), occurring in a stuttering fashion over a few hours. She was presumed to have ischemic damage (a cerebral infarct, a type of stroke) to the motor pathways in the left cerebral hemisphere (in the forebrain). However, the next day she developed double vision and was found to be unable to move her left eye to the left (failure of abduction to the left), indicating damage to the intramedullary course of the left sixth cranial nerve (abducent nerve). The combination of right hemiparesis and left sixth nerve palsy implied that the ischemic damage had occurred in the left lower brain stem, anteriorly where the intrapontine fibers of the left sixth nerve pass close to the descending corticospinal motor pathway, above the level where it crosses to the contralateral side (at the lower border of the medulla oblongata). An MRI scan confirmed the presence of a lesion situated anteriorly and medially on the left side of the pons (Fig. 6).

See Also the Following Articles

CEREBELLUM • CRANIAL NERVES • HINDBRAIN • HOMEOSTATIC MECHANISMS • NEUROANATOMY • NEURON

Acknowledgments

I thank my clinical colleagues Richard Burns and John Willoughby and my neuropathological colleagues Peter Blumbergs and Gai Weiping. The National Health and Medical Research Council supported my research on the human brain stem.

Suggested Reading

Blessing, W. W. (1997). *The Lower Brainstem and Bodily Homeostasis*. Oxford Univ. Press, New York.

Brodal, A. (1981). *Neurological Anatomy in Relation to Clinical Medicine*. Oxford Univ. Press, New York.

Kandel, E. R., Schwartz, J. H., and Jessell, T. M. (Eds.) (2000). *Principles of Neural Science*, 4th ed. McGraw-Hill, New York.

Loewy, A. D., and Spyer, K. M. (Eds.) (1990). *Central Regulation of Autonomic Function*. Oxford Univ. Press, Oxford.

Victor, M., and Ropper, A. H. (1997). *Adams and Victor's Principles of Neurology*, 7th ed. McGraw-Hill, New York.

Broca's Area

MICHELLE CRANK and PETER T. FOX
University of Texas Health Science Center, San Antonio

GLOSSARY

Broca's aphasia General term for a group of speech disorders caused by injury or disease in Broca's area of the left hemisphere; selective loss of the ability to produce articulate speech despite preserved speech comprehension and not due to motor impairment.

Broca's area Area of the cerebral cortex responsible for motor speech planning usually located in the frontal lobe of the left hemisphere. The area is named for Pierre Paul Broca, who deemed the left inferior frontal gyrus to be the "seat of articulate language" in 1861. It is the first brain area for which a specific function was identified, "thus the existence of the first localization once admitted, the principle of localization by convolutions would be established."

functional magnetic resonance imaging Use of magnetic resonance imaging to detect physiological events; technique for mapping task-induced changes in regional neural activity by detecting changes in local blood flow, blood volume, and tissue oxygen content triggered by neuronal activity.

positron emission tomography Technique for scanning the brain in which radionuclides are used to determine oxygen utilization, glucose metabolism, and blood flow in the cerebral cortex; used for mapping brain functional organization and for detecting brain disorders.

Broca's area—the posterior, inferior portion of the human left frontal lobe—has the impressive distinction of being the first area of the human brain for which a specific function was proposed: articulatory control of speech. Since its original description in the mid-19th century, Broca's area has been the subject of thousands of scientific articles. Remarkably, present-day conceptualizations of the function of Broca's area are fundamentally similar to those originally proposed more than a century ago. During the past century, the technologies applied to the study of the human brain have steadily advanced. Early clinical and anatomical investigations have been bolstered by cortical electrical stimulation and, recently, by functional brain imaging. Although improvements in technology have provided finer functional descriptions and fractionations, the region of the brain called Broca's area is still firmly rooted in the classical descriptions of the 19th century.

I. CLASSICAL DESCRIPTIONS

A. Broca's Cases

Broca's area is named for Pierre Paul Broca, the French surgeon and anthropologist who identified the area and defined it as the seat of the faculty of articulate language in 1861. In his capacity as a general surgeon, Broca was called upon to treat a case of advanced gangrene of the leg. As Broca treated the patient, Leborgne, he learned that Leborgne was unable to speak due to a chronic brain lesion. Leborgne, who had experienced epileptic attacks since youth, lost his ability to speak at the age of 31. Ten years later he developed paralysis on his right side. Broca, examining him at the age of 51 determined that Leborgne could produce virtually no speech. He often said "tan" (and later became known as "Tan") and occasionally uttered the phrase "Sacre nom de Deiu." Tan's mouth, tongue, and larynx were intact,

indicating a central cause for the lack of speech. Tan could hear and comprehend speech because he responded correctly to questions that he could answer with motions of his left hand, such as indicating the number of years he had been without speech. Because of his loss of speech, Leborgne had been living in the Hospice of Bicetre, where he had been known for his vindictiveness and sometimes objectionable character. He thus seemed to have retained his intelligence despite his lack of articulation, until the paralysis and gangrene confined him to his bed.

One week after Broca's initial examination, Tan died of gangrene. Broca performed an autopsy and found that most of the left hemisphere of the brain was softened from the disease process. The anterior part of the hemisphere contained a large lesion, centered over the posterior aspect of the middle and inferior frontal gyri. The lesion also encompassed part of the anterior insula, the inferior marginal gyrus, and the striate body. Based on an examination of the surrounding, less damaged tissue, and the knowledge that the clinical course was progressive, Broca postulated that the lesion had begun in the third frontal gyrus and extended to surrounding tissue as the disease progressed.

Broca presented Leborgne's brain at a meeting of the Societe d'Anthropologie one day after the patient's death and then deposited the brain in the Musee Dupuytren. Figure 1 shows Leborgne's brain viewed from the left, with a lesion clearly visible in the third frontal convolution. Broca coined the term "aphemia" to describe the condition of loss of articulate speech coupled with the retention of hearing, comprehension, and intelligence. He inferred a correspondence between the anatomical and symptomatological periods of Leborgne's condition: The 10-year period of aphasia occurred while the lesion was confined to the frontal lobe, and as it progressed into the corpus striatum, the symptoms of paralysis of the right extremities developed. Broca concluded that the origin of the lesion, the second or third frontal gyrus, must be responsible for the loss of articulate language.

Even as he described this area and its importance to the study of language in the brain, Broca emphasized that the linguistic aspect of his discovery was less significant than its implications for the general theory of localization of function in the brain. Broca (as quoted in Von Bonin, 1960, p. 55) was well aware that this was the first strong evidence that specific mental functions were specifically localized:

> *The point is cerebral localization, not this or that school of phrenology… . If all cerebral faculties were as distinct and as clearly circumscribed as this one, one would finally have a definite point from which to attack the controversial question of*

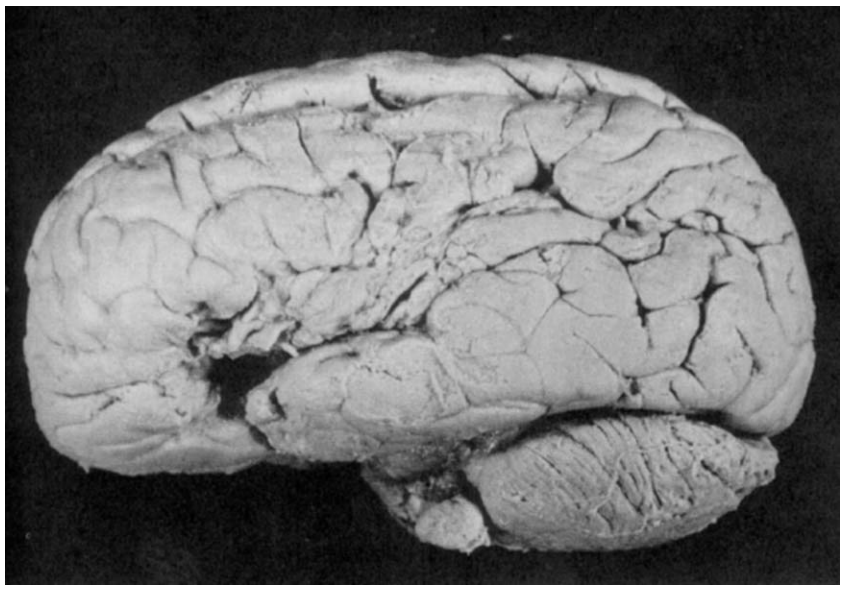


Figure 1 Photograph of Leborgne's brain, the first patient Paul Broca diagnosed with aphemia (later to be known as Broca's aphasia) in 1861. This is a view of the left hemisphere, in which a lesion is visible in the area of the third frontal gyrus, now known as Broca's area (reproduced with permission from Signoret *et al.* (1989), *Brain Language* **22,** 303–319).

cerebral localization. Unfortunately, this is not the case, and the greatest obstacle in this part of physiology comes from the insufficiency and the uncertainty of the functional analysis which necessarily has to precede the search of the organs which are coordinated to each function.

Broca's second case was a man named Lelong, who also had lost his ability to use articulate speech but whose aphemia was not complicated by other symptoms. At autopsy, Lelong's lesion proved much more circumscribed than Leborgne's but was also located in the posterior portion of the third frontal convolution. Broca subsequently chronicled 20 more cases in which patients with aphemia had lesions of the third frontal convolution, lending strong support to his hyopthesis originally based on a single case. Collectively, these cases paved the way for the systematic investigation of the functional organization of the human brain based on the analysis of clinical cases. For decades, lesion-deficit analysis was based on postmortem lesion localization. With the advent of X-ray computed tomography (CT) and magnetic resonance imaging (MRI), lesion localization has become premorbid and quantitative (e.g., describing lesion volume as well as location). This work, summarized in Section II, largely confirms Broca's original observations.

B. Importance to Theories of Localization and Lateralization

Broca advocated the principle of cerebral localization, which at the time was disputed by more traditional supporters of the concept that the brain acted as a whole. Many scientists agreed that the intellectual activities of the mind resided in the convolutions of the brain and that the most advanced functions were located in the frontal lobes, but they were not ready to locate specific functions in specific convolutions. Broca realized that if articulate language were proven to be located invariably in one place in all human brains, it would be a major step toward proving that all functions were localized. In his presentation of Leborgne's brain in 1861, he spoke in general on language and localization as an introduction to his clinical and pathological observations (as quoted in Von Bonin, 1960, p. 57):

There are in the human mind a group of faculties and in the brain groups of convolutions, and the facts assembled by science so far allow to state, as I said before, that the great regions of the mind correspond to the great regions of the brain.

He goes on to specify that the case of language could be evidence enough to draw some conclusions: "Thus the existence of the first localization once admitted, the principle of localization by convolutions would be established" (as quoted in Von Bonin, 1960, p. 58).

Broca was reluctant to state that Leborgne's case proved that language would be invariably located in the third frontal gyrus of the left hemisphere. He merely suggested that the theory of localization was highly probable, and that further observation would be needed to draw firm conclusions. He did, however, distinguish his version of the theory from an earlier phrenological idea, which placed the faculty of articulate language at a fixed point under a certain elevation of the skull. Broca did not support the theory of lateralization, which was even less widely accepted than localization. In fact, he reported it as a coincidence that his later cases were all documented lesions of the left hemisphere: "Remarkable, how in all these patients the lesion was on the left side. I do not dare to draw any conclusion from this and am waiting for new data" (as quoted in Schiller, 1979, p. 194).

He knew that the firm conclusion would be that each side of the brain had different functions, but he did not accept it, and he ceased to mention it in his later research on language. However, his placement of "the seat of articulate language" in the third frontal gyrus of the left hemisphere certainly fueled the debates over localization and lateralization, and his ideas were a starting point for those who did eventually draw firm conclusions about both issues.

C. Brodmann's Description of Broca's Area

In 1909, Korbinian Brodmann published a treatise, "Localization in the Cerebral Cortex," on histological localization in the cerebral cortex. He described his work in the introduction: "Localization which uses exclusively anatomical features as the basis for investigation, in contrast to physiological or clinical aspects" (as quoted in Garey, 1994, p. 117). The lasting result of this work was a scheme for the parcellation of the cerebral cortex into 50 distinct areas based on common anatomical features. The system is still used as a standard for cortical localization.

Broca's area was designated by Brodmann as area 44: "a well-differentiated and sharply circumscribed structural region that on the whole corresponds quite

well to the opercular part of the inferior frontal gyrus: Broca's area'' (as quoted in Garey, 1994, p. 117). Brodmann defined the anatomical boundaries of area 44 as the inferior precentral sulcus posteriorly, the inferior frontal sulcus superiorly, and the ascending ramus of the Sylvian fissure anteriorly, but he amended his description: ''As there is much variability and inconstancy of these sulci one will find rather mixed topographical relationships of these structural areas in individual cases'' (as quoted in Garey, 1994, p. 117).

In addition, Brodmann found enough similarity in the structure of area 44 to that of two other nearby areas, 45 and 47, that he grouped them into a subregion, the ''subfrontal subregion.'' A fourth area, 46, was also more loosely grouped with 44, 45, and 47 in the frontal region, and Brodmann said it was not distinguishable from surrounding areas by its cell structure. These areas are considered to be subdivisions of the more general Broca's area addressed in functional imaging literature today. Brodmann's identification of a cytoarchitectonic area matching the placement of Broca's ''third frontal convolution'' was a confirmation from yet another scientific discipline to add to Broca's own and subsequent clinical and physiological evidence.

D. Anterior/Posterior Dichotomy

That the functions of the cerebrum could be divided into motoric and sensory, with motor functions located in the anterior portion of the brain and sensory functions in the posterior, was a prevailing idea in the 19th century. Broca's postulate concerning the function of the inferior frontal convolution followed this broad division of function in that the portion of language that he had placed in the frontal lobe controlled the output of language. However, Broca considered articulate speech to be an intellectual function and not merely motoric. In this, Broca agreed with the school of thought that placed intellectual functions in the frontal lobes, relegating processes influenced by emotion and passion in the posterior lobes.

In 1874, Karl Wernicke described an area of the posterior, superior temporal lobe as being responsible for speech comprehension, including the understanding of word meaning (i.e., semantics). In doing so, Wernicke undermined Broca's speculations that Broca's areas supported cognitive aspects of language. This set a trend of localization of linguistic semantics to the posterior, superior temporal lobe (Wernicke's area) that has persisted in the clinical literature through the 20th century. Recently, through the use of functional imaging technology, higher order functions related to language have been returned to the frontal lobe. This shift has occurred with the subdivision of the frontal language area into more specialized semantic and articulatory areas. Those that involve semantics activate a distinct area anterior to Broca's classically defined seat of articulate language.

E. Recent Work: Finer Subdivisions

Recent work with clinical studies and functional imaging has shown that the single area Broca called the seat of articulate language can be subdivided into regions active in different types of language activities, such as making semantic versus orthographic or phonological judgments. It seems that three areas, the traditional Brodmann's area (BA) 44, BA 46/47, and the anterior insula, are all activated in speech tasks, but they are activated in different combinations in different speech paradigms. It seems that the higher order processes are activating the left anterior area, BA 46/47, and articulation is activating the more posterior BA 44 and the insula. However, imaging has also shown that these subdivisions work together, as Paulesu suggests in his theory of an articulatory loop for language processing. He suggests that Broca's area (BA 44), superior temporal gyri (BA 22/42), supramarginal gyri (BA 40), and the insulae all bilaterally activated in a phonological short-term memory task constitute a language circuit connecting the posterior and anterior areas. He links Broca's area to the subvocal rehearsal system active in working memory and temporal BA 22/42 to phonological processing independent of higher order memory function. This demonstrates the trend of returning the higher order processes to the frontal lobe. Overall, combining functional imaging with the methods of study already in use, such as Broca's own clinical and pathological observation and electrical stimulation studies, has allowed for a more detailed analysis of the function of the region around Broca's area.

II. ELECTRICAL STIMULATION

Electrical stimulation of the brain in awake subjects undergoing neurosurgery was introduced in the 1930s by Wilder Penfield and colleagues at the Montreal Neurological Institute. Penfield treated patients with

focal brain lesions, operating to remove damaged tissue. In doing so, Penfield performed intraoperative electrical stimulation mapping to identify areas of the cortex whose loss would be debilitating. Of relevance to this article, Penfield mapped language-related brain areas, compiling data from scores of cases to produce functional maps of language areas, including Broca's area. (The reader is encouraged to investigate the numerous journal articles and books published by Penfield and colleagues.) Today, the most prolific applicants of Penfield's intraoperative cortical mapping technique are George Ojemann and colleagues at the University of Washington. In a classic, 1982 review of his cases, Ojemann reported on more than 100 patients, indicating that the location and extent of Broca's area (and other language areas) are more complex and variable than previously thought, at least in a population of persons with chronic brain lesions. The technique of electrical stimulation mapping of the cortex allowed the study of language areas, and Broca's area in particular, to expand into testing the use of normal tissue in specific locations in conscious patients rather than simply documenting the effects of damage to areas of the cortex. Electrical stimulation mapping has emphasized the individual variability of language areas of the brain in clinical populations, and recent work seems to show that the classically defined Broca's area may not be necessary for speech processing. The works of Penfield and Ojemann have made a major contribution to the knowledge of the location and functions of Broca's area.

A. Penfield's Cortical Stimulation Studies

Penfield and Roberts used excision studies coupled with electrical stimulation studies to map language areas, and they concluded that Broca's area plays a role in ideational (as opposed to motor) speech but its excision did not produce permanent symptoms of aphasia; rather, speech is recovered at least to some extent. They used preoperative electrical mapping throughout the mid-20th century to identify areas of the cortex involved in language in order to avoid them when performing cortical excisions to treat epilepsy. When electric current was applied to the cortex at points potentially involved in language, two speech effects occurred that were termed positive and negative. A positive effect was stimulation of speech when motor areas for face and mouth were stimulated, and negative effects included interference of speech when ideational speech areas, such as Broca's, Wernicke's, and supplementary motor areas, were stimulated. Patients were awake, and they were asked either to perform a picture-naming task or to count aloud. The interference was only effective approximately one-half of the time, and movement induced in one spot at one time might not occur upon restimulation of the same spot in the same patient; thus, results were difficult to replicate and inexact. Compiling their data from several patients, Penfield and Roberts found that when Broca's area was stimulated, negative effects such as speech arrest, hesitation, slurring, repetition, confusion of numbers while counting, inability to name pictures with retained ability to speak, and difficulty in reading and writing resulted. They define Broca's area as consisting of the three gyri in front of the lower precentral gyrus. This includes BA 46/47, now separated from Broca's area proper by functional imaging literature. By including this area, Penfield and Roberts decreased the specificity of the area they were mapping. They also found that stimulation of sites outside Broca's area and other traditional language areas, such as Wernicke's area in the temporal lobe and the supplementary motor area, caused similar interruption of language. This suggested expansion of the traditional areas. The majority of their patients did have interference only in left Broca's stimulation, as opposed to its right hemisphere counterpart, which they concluded supported the idea of left hemisphere dominance for language. In observing cases of excision of part or all of Broca's area, Penfield and Roberts concluded that permanent aphasia did not necessarily result, but speech was recovered postoperatively in most cases. This suggests that Broca's area is not necessary for normal language function, and it complicates the picture created by their electrical stimulation studies, which show that Broca's area is a key language area.

B. Ojemann's Localization Mapping

Ojemann compiled data from 117 subjects in an electrical stimulation mapping study of language areas of the left hemisphere, and he concluded that language areas are variable mosaics of 1-cm^2 sites ranging across and beyond classically defined areas such as Broca's and Wernicke's. As in Penfield and Roberts' study, patients were asked to name line drawings of common objects while electrical current was delivered to random sites on the cortex in the area surrounding the proposed excision. Anatomical markers were used to relate the sites of stimulation in each different brain

to an arbitrary grid created using the Sylvian fissure and the rolandic cortex as landmarks. Not all patients had stimulation sites in all regions, but a map of the variability of language localization across all patients was created taking this into account. The most common finding was that sites where stimulation caused errors were separated by less than 1 cm in all directions from sites with no errors. The boundaries of these language areas were sharp in some cases, and in others they were surrounded by areas in which stimulation evoked single speech errors, suggesting a gradation from areas with no role in naming to areas essential for it. In an individual, these language areas were arranged in a mosaic pattern extending 1 or 2 cm^2. Ojemann's study shows the high degree of individual variability that exists for essential language areas coupled with discrete localization in a given individual. The combined area of the inferior frontal mosaics in a given individual is much less than that of traditionally defined Broca's area and less than the corresponding area identified by Penfield and Roberts as essential. Ojemann did not attempt to differentiate among BA 44 and BA 46/47, but he did demonstrate that the most posterior portion of the inferior frontal gyrus was one of the least variable areas. This would correspond to BA 44. His work was limited in the specificity of what part of language function was being deactivated by the stimulation; for example, failures of perception, motor control, naming, or consciousness could all potentially have caused patients to fail at the naming task. In addition, buried sites, such as the insula, could not be directly stimulated and thus may also be essential to language and were not detected by this method. A final limitation is that the technique is necessarily performed on patients with lesions or tumors, which adds the possibility that their language function may be altered from that of normal individuals as a result of their medical condition. Despite the limitations, electrical stimulation does identify areas of the cortex essential for language, and in this case Ojemann demonstrated the great individual variability of those essential language regions. This challenges the assumption that Broca's seat of articulate language is a clearly defined region easily generalized among individuals.

III. CLINICAL STUDIES

Clinical studies of aphasics defined Broca's research and conclusions; since then, clinical work has been expanded and refined, combining functional and anatomical imaging to show that lesions to Broca's area do not necessarily produce symptoms of Broca's aphasia.

A. Mohr's Analysis of Broca's Aphasia and Broca's Area

In 1976, Mohr demonstrated that the symptoms classically defined as Broca's aphasia did not correlate with lesions localized in Broca's area proper, BA 44/6; lesions in that area produced transient symptoms similar to those chronic ones associated with much larger lesions. Mohr began by describing the characteristics of the deficit associated with localized lesions of the third frontal convolution, or Broca's area as Broca defined it. Mohr stated the characteristics: Little persisting deficit in articulation seems to occur, and frequently no significant persisting disturbance in language function is present. He based his claims on personal cases of clinical and pathological correlation and an extensive review of cases reported in the literature since 1861 and more than 1 year of cases of stroke documented by Massachusetts area hospitals. Autopsy was the preferred method of analysis of the location of the lesion, but CT scans were also used in a few cases. He formulated a hypothesis of the function of Broca's area that the clinical data support. His current thesis envisions Broca's area as mediating a more traditionally postulated role as a premotor association cortex region concerned with acquired skilled oral, pharyngeal, and respiratory movements involving speaking as well as other behaviors, but not essentially language or graphic behavior per se. Mohr continued by categorizing Broca's aphasia as it was described originally by Broca (as aphemia) and then repeatedly confirmed by others since 1861. The clinical literature continued to rely on Wernicke's 1874 definition that stated three main aspects: Patients are mute, using only a few senseless syllables or swear words with preservation of muscular speech ability; speech comprehension is maintained, but not for complicated constructions; and written language is lost with spoken language. The lesions that produce these symptoms usually involve Broca's area proper, but they are more extensive. Broca's aphasia reflects a major infarction involving most of the territory of supply of the upper division of the left middle cerebral artery. It is observed only after the initial infarction. The initial syndrome is more severe, described traditionally as golobal aphasia. Mohr notes that Broca's original cases involved more extensive lesions;

however, Broca ignored the size and instead relied on the stroke theory of his time, which stated that large strokes always began as a smaller focus and spread slowly outward. Mohr's dissociation of Broca's area proper from the classically defined syndrome of Broca's aphasia suggested that in the clinical literature too much weight was placed on the original classification of the area with the syndrome, and as a result, too broad a range of function had been attributed to the area. Overall, his investigation proved that more detailed analyses would be necessary to determine the finer points of the function of Broca's area proper. Functional imaging technology has brought those details forward, and now experts continue to analyze the area both in clinical populations and with normal subjects.

B. Clinical Analysis of Syntax

Clinical investigations relating to syntactical analysis have associated some syntactic ability with Wernicke's aphasia that patients with Broca's aphasia do not possess. Even though the extent of lesions causing Broca's aphasia is greater than originally thought, it does include BA 44 and the surrounding areas, including portions of BA 46/47. This can be easily differentiated from the area of the temporal lobe responsible for Wernicke's aphasia. Swinney and Zurif tested Broca's and Wernicke's aphasia patients to compare their success and rates of processing traces in sentences. If Broca's aphasia patients cannot process one of the syntactic components of sentence structure, then the cortical area affected by the lesion in and around Broca's area may be responsible for that component of syntactic processing. In Swinney and Zurif's study, Broca's aphasics were less effective at processing than were Wernicke's aphasics, Swinney and Zurif suggest that a prima facie case can be made that long-distance dependency relations of the sort described here are especially reliant on some form of working memory capacity. This clinical work relates higher order memory and syntactic functions to areas of the frontal lobe responsible for Broca's aphasia and potentially to BA 46/47.

C. Clinical Analysis of Sub-Broca's Area

Clinical data on the anterior insula are sparse. The insula is included in analyses of Broca's area lesions and in descriptions of Broca's aphasia lesions. Clinical observations of selective stimulation of the insula, though few, show that it will cause speech arrest, and thus the area is linked to articulation. The most profound observations of the insula that demonstrate its function more directly have been made possible by functional imaging, and these observations are analyzed in the following section.

IV. FUNCTIONAL IMAGING

Functional imaging technology has freed studies of language in the brain from the lesion deficit framework and has allowed more detailed analyses of the functional organization of language in the brain. Whereas lesion studies may show what areas are necessary for normal language function, position emission tomography (PET) and functional MRI (fMRI) can reveal the complex association of specific areas that are utilized in normal speech and reveal activation of the same areas in other tasks to further specify or broaden their function. Comparison of images obtained from phonological and semantic processing tasks in visual and auditory modalities helps to identify regional cerebral blood flow (rCBF) activation patterns that clarify the role of Broca's area in relation to other language areas. Figure 2 shows a three-dimensional representation of the subdivisions of Broca's area apparent from functional imaging and the motor area for the mouth on the left (Fig. 2, top) and right (Fig. 2, bottom). The average positions (Table I) were calculated from a meta-analysis of functional imaging literature (Table II) reporting activation in standardized space, and relative intensities are based on z scores of activation.

A. Broca's Area Proper

Broca's area proper, BA 44/6, is attributed with a range of functions, but its exact role in language is debated based on the variety of different tasks that activate it. The basic point of agreement in the functional imaging literature is that this region is involved in articulation and transforming phonemes into speech motor plans, as indicated by its activation in overt speech tasks. Figure 3 shows a meta-analysis of the functional imaging literature reporting activation of BA 44/6 in overt or covert speech paradigms plotted as average three-dimensional coordinates. The studies used in the meta-analysis all involved

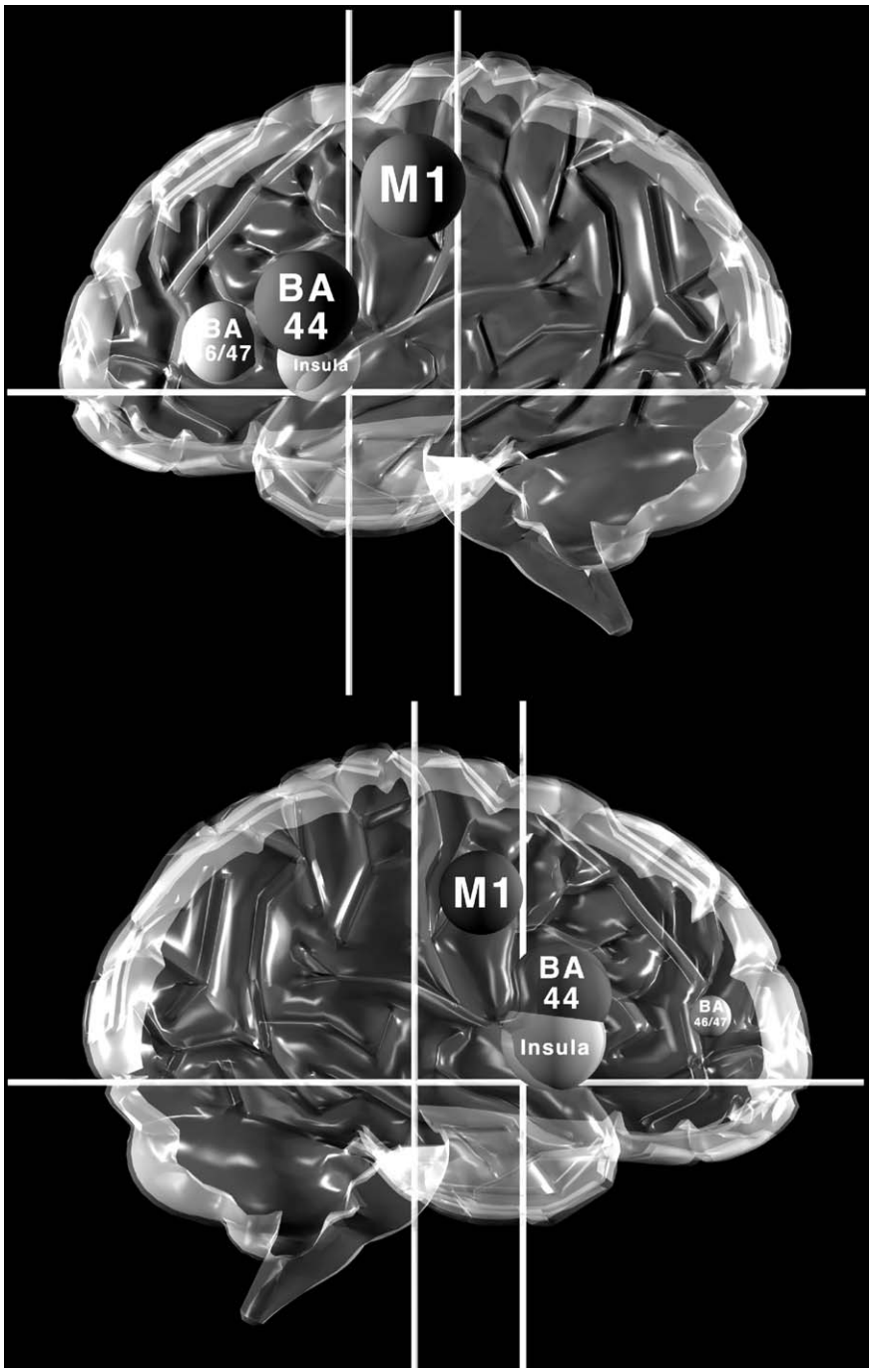


Figure 2 Three-dimensional representations of the three subdivisions of Broca's area and the motor area for the mouth are shown from the left (top) and right (bottom). The average positions (see Table I) were calculated from a meta-analysis of functional imaging literature (see Table II) reporting activation in standardized space, and relative intensities are based on *z* scores of activation. Broca's area proper, or BA 44/6, is active in both hemispheres, but it shows greater activation in the left hemisphere. Sub-Broca's area, or anterior insula, appears more evenly distributed between the hemispheres and demonstrates less intense activation than does Broca's area proper. Pre-Broca's area, or BA 46/47, is significantly lateralized to the left hemisphere, and the area shows the least intense activation of the three subdivisions. The motor area for the mouth is lateralized to the left hemisphere and shows activation on that side comparable to that of Broca's area proper. (See color insert).

Table I
Average Activation Coordinates for Language Areas

Area	*X* (SD)	*Y* (SD)	*Z* (SD)	*N*	Average *z* score
BA 44 left	−45.63 (4.64)	6.73 (10.14)	16.47 (4.74)	134	4.59
BA 44 right	39.5 (9.57)	6 (17.51)	16 (10.33)	33	4.91
BA 46/47 left	−39.61 (7.87)	19.75 (16.88)	10.45 (11.13)	131	4.21
BA 46/47 right	41.14 (5.94)	32.56 (8.28)	11.62 (17.09)	53	2.5
Insula left	−34.65 (3.07)	5.35 (9.09)	6.7 (5.29)	140	4.4
Insula right	39.75 (5.56)	4.5 (9.57)	8.25 (7.68)	36	4.95
M1-mouth left	−47.13 (3.12)	−11.85 (3.08)	36.77 (5.5)	101	5.24
M1-mouth right	49.05 (7.47)	−7.98 (4.55)	34.38 (7.97)	101	3.74

production of overt or imagined speech, although the individual tasks varied from repetition of printed or heard nouns to self-paced counting to narrative speech, and subtractions often did not remove the semantic component of tasks. The studies differentiated among the areas of M1-mouth and BA 46/47 in most cases; therefore, averaging in activation from these distinct areas did not shift the reported coordinates. Activation of Broca's area proper has been observed in other nonlanguage tasks, such as accessing verbal working memory and overt and imagined movement, so its function may need to be generalized from articulatory planning to somatosensory motor planning.

Table II
Meta-analysis of Functional Imaging Literature Showing Activation in Broca's Area

	Broca's Proper BA 44/6		Pre-Broca's area BA 46/47		Sub-Broca's area insula		M1-mouth	
Reference	Left	Right	Left	Right	Left	Right	Left	Right
Becker (1994)	X		X	X	X	X		
Bookheimer (1995)	X	X			X		X	X
Braun (1997)	X		X		X		X	X
Buckner (1995)			X					
Fiez (1996)					X			
Fox (1999)	X				X		X	X
Jennings (1998)		X						
Paulesu (1993)	X	X			X	X		
Paus (1993)			X				X	X
Petersen (1988)	X		X				X	X
Petrides (1993)	X		X				X	X
Petrides (1995)	X		X					
Price (1994)					X	X		
Raichle (1994)			X		X	X		
Rumsey (1997)					X			
Warburton (1999)			X					
Weiller (1995)	X	X	X					
Wise (1991)	X							
Wise (1999)			X		X			

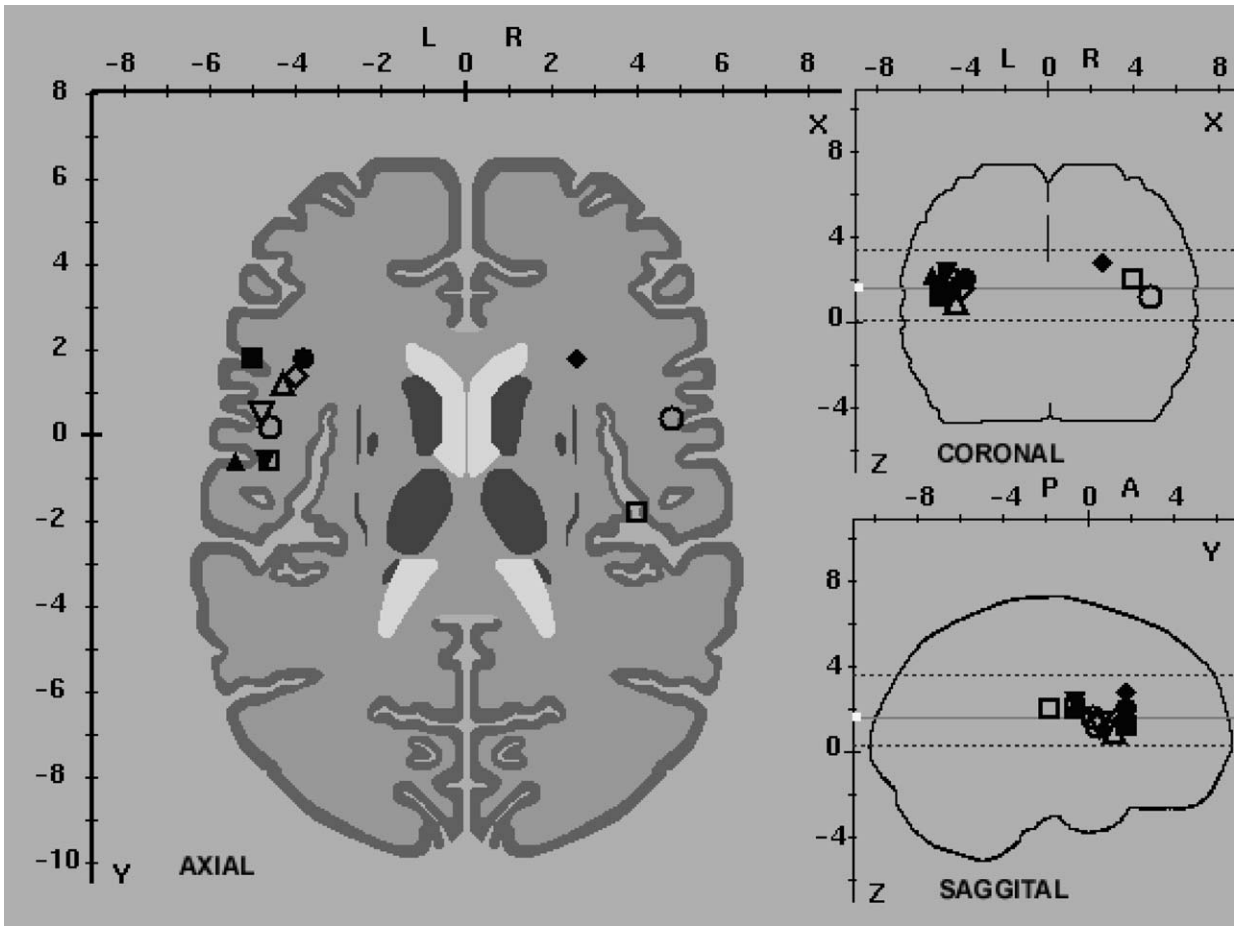


Figure 3 Activations in Broca's area proper, BA 44/6, are shown in a meta-analysis of functional imaging data for overt and covert speech paradigms (see Table II). Mean coordinates are normalized and reported in standard three-dimensional (Talairach) space, plotted with the BrainMap database on an axial cross section at $z=1.6$ cm. Each relevant reference listed in Table II is represented here by a mean coordinate with the following symbols: ■, Becker (1994); □, Bookheimer (1995); ●, Braun (1997); ▲, Fox (1999); ◆, Jennings (1998); ○, Paulesu (1993); ▼, Petersen (1988); △, Petrides (1993); ▽, Petrides (1995); ⊡, Weiller (1995); ◇, Wise (1991). The points are distributed in both hemispheres but are predominantly found on the left, demonstrating the low degree of lateralization of BA 44/6. For all meta-analyses, data were taken at face value, such that those points reported by the author to be within a given subarea are displayed there, although some points may appear to go beyond the Brodmann areas typically included in that subarea.

1. Overt Speech Activation

Posterior Broca's area is activated in fMRI and PET studies when overt speech is produced, specifically in repetition of words presented visually or aurally or generation of verbs or sentences in response to presented nouns. In trying to elucidate the exact reason for its activation, paradigms have been created in a hierarchical scheme. The most complex tasks, such as semantic processing, are performed along with phonological processing and a baseline rest condition so that subtracting one from another will isolate the activation area responsible for the semantic or phonological portions of the task. Often in studies in which Broca's area proper is activated, activation is explained simply by stating that the area is involved in spoken language production. It is credited with a phonological rather than a semantic aspect of language production, but even the nature of that phonological language role is questioned. Petersen found that areas near Broca's area proper are activated bilaterally during movement of the mouth and tongue, and he credits Broca's area with general motor rather than language-specific output planning. In addition to this broadening of the function of Broca's area, one study by Petrides stated that speech can occur without

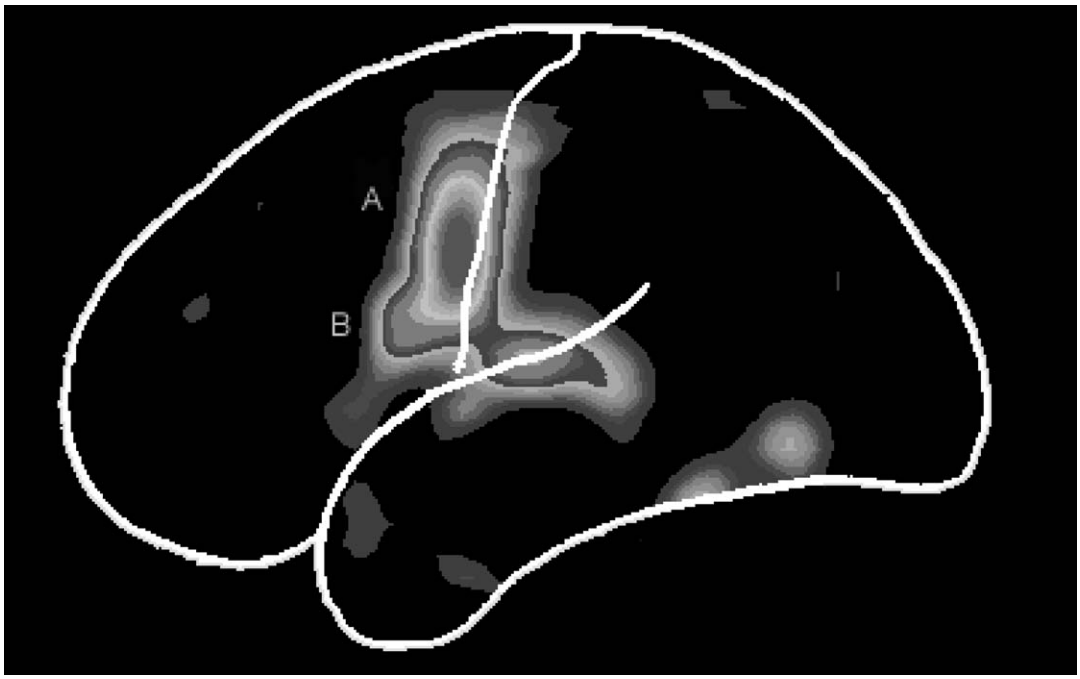


Figure 4 The pattern of activation for overt speech in the left hemisphere is shown through fMRI. The areas most active are Broca's area proper (B) and M1-mouth (A), whereas the prefrontal BA 46/47 shows no significant activation. The motor mouth area can be distinguished from Broca's area proper because the motor area extends well above BA 44/6.

activation of BA 44: It is possible that Broca's area may not necessarily be activated in all speech production tasks. In a study by Braun that compared stuttering to normal speech, a paced speech task and an overlearned speech task showed decreased activation of BA 44 relative to narrative, self-paced speech activation in normal subjects. Braun proposed that in slow or overlearned conditions, it is easier to process and produce language plans. Broca's area, which may be crucial in initial word production or monitoring, is not needed once the given words are learned: Phonological or semantic monitoring may be less critical to the degree that significant engagement of the neocortical language areas is no longer essential. Figure 4 demonstrates the pattern of activation in the production of overt speech, concentrating in BA 44/6 and the M1-mouth motor region.

2. Covert Speech Activation

The description of the function of Broca's area proper expanded from overt speech production when tasks involving inner, or covert, speech revealed activation and complicated the classical description. BA 44/6 is activated under conditions of covert verb generation in response to heard nouns presented at a slow rate. Wise postulated that Broca's area shows activation in an automatic, internal speech response to remembering words even though vocalization is not the goal: The act of retrieving words from semantic memory activates networks concerned with the production of speech sounds. Another aspect of covert speech is subvocal rehearsal, or internal repetition of speech often coupled with phonological working memory. Paulesu described an articulatory loop that he tested with a short-term memory task for letters and rhyming judgment for letters. Both tasks required subvocal rehearsal, and both showed strong activation in Broca's area, whereas only the memory task activated the left supramarginal gyrus. He concluded that Broca's area is crucial to the subvocal rehearsal system. Assuming a connection of Broca's area to subvocal rehearsal, Caplan set out to prove that Broca's area is involved in syntactic processing when subvocal rehearsal is prevented. Subjects judged the goodness of sentences that represented two distinct levels of syntactic difficulty, and as they made judgments they repeated the word "double" timed with a metronome. The task contrast showed increases in rCBF in Broca's area in judging the more difficult type

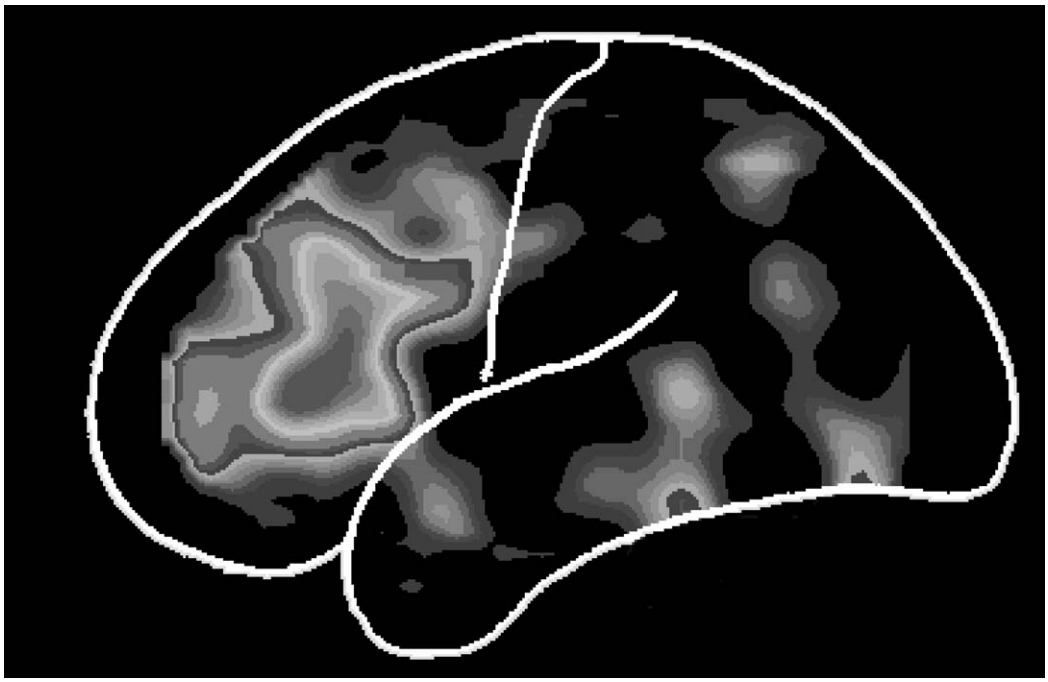


Figure 5 Covert speech, or thinking about words without vocalization, shows heightened activity in left pre-Broca's area, BA 46/47. This indicates that motor aspects of speech are not necessary for this area to be activated, and pre-Broca's is clearly defined in an area forward of Broca's area proper that is probably linked more closely to semantic processing of words than speech or phonological processing.

of sentences: The increase in rCBF in Broca's area reflects the demands of processing the more complex syntactic structure. The possible explanations given are maintenance of the complex structure in working memory or need for more intermediate constructs in computation of the more complex sentences; either explanation supports the addition of syntactic processing to the list of Broca's area functions. Figure 5 shows activation produced when thinking about words, including activation in BA 44/6.

3. Nonlanguage Activation

Movement in imaging tasks that does not incorporate speech shows activation in Broca's area. Both overt and imagined movements of the arm and hand cause activation in BA 44. This activation might be explained by subjects' unintentional performance of covert speech during the movement tasks. Parsons used a task that involved judging whether a presented drawing of a hand was of a right or a left hand, thus engaging the brain to shift the mental representations of its own hands to try to imitate the presented hands. Parsons suggested that since the activation was in the limb-contralateral hemisphere, it was correlated with the actions of the hand covertly moving: The activated BA 44 site subserves the mental simulation of reaching a specific limb into a specific target posture. Also, in a study of saccadic eye movements, activation was found in ventral BA 6, on the edge of Broca's area, when saccades were performed in the dark. This overt movement was performed in proprioceptive space, as opposed to a similar task of saccades to visual targets, performed in visual space, that activated the dorsal premotor area BA 8. This distinction between dorsal and ventral premotor activation suggests that BA 44/6 may be responsible for motor planning in somatic space for movement that cannot be calculated within the visual field. Saccadic eye movements in the dark as well as oral movement must employ somatic planning, and BA 44/6 may be the ventral, somatic counterpart to dorsal premotor areas.

B. Pre-Broca's Area

Semantic function in language, or analysis of word meaning, has been shown by functional imaging to involve the cortical areas immediately anterior to Broca's area proper, prefrontal BA 46/47. Figure 6

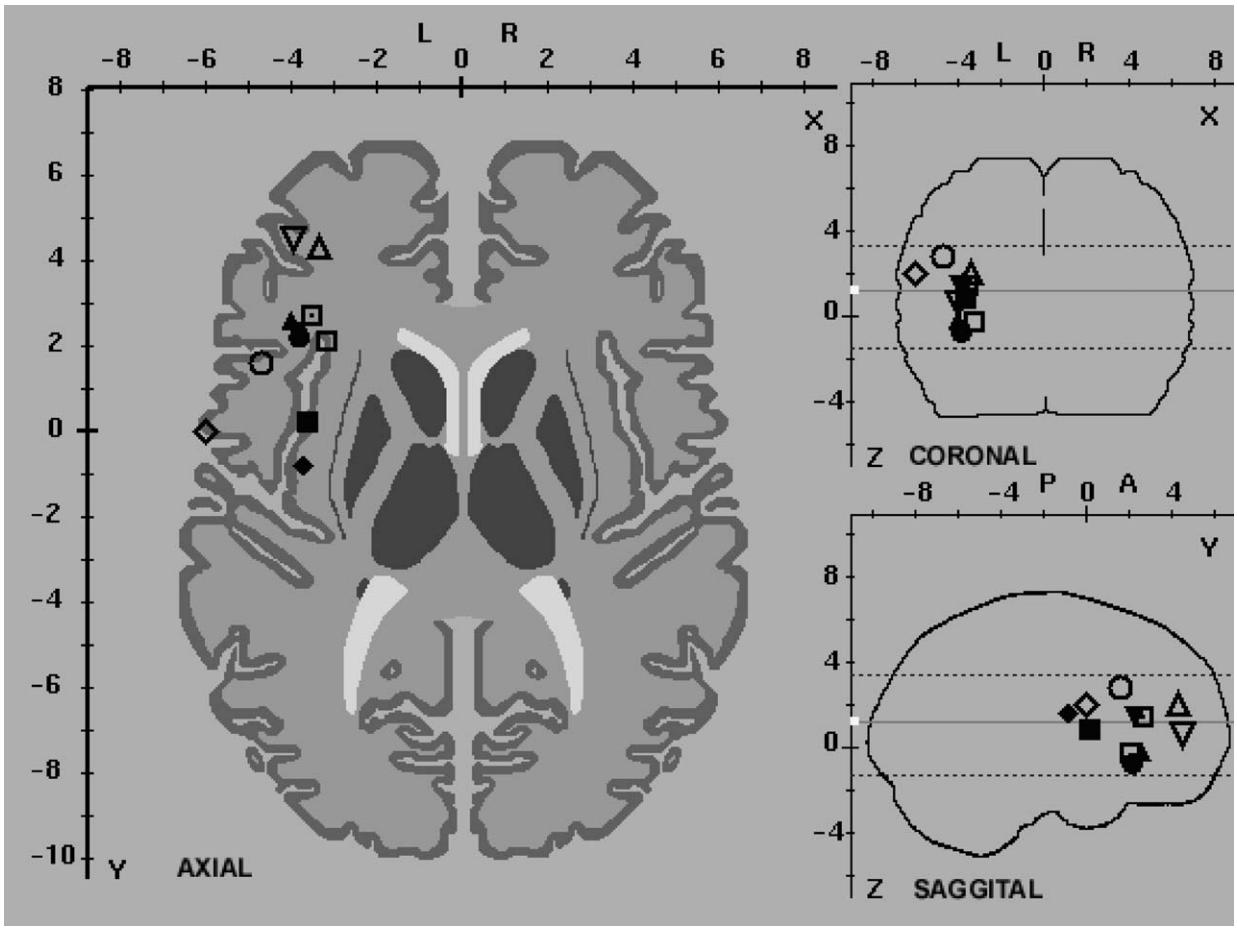


Figure 6 Activations in pre-Broca's area, BA 46/47, are shown in a meta-analysis of functional imaging data for covert speech paradigms. A set of mean coordinates was calculated from data reported from studies in Table II. The coordinates have been normalized and reported in standard three-dimensional (Talairach) space, plotted with the BrainMap database on an axial cross section at z=1.2 cm. Each relevant reference listed in Table II is represented here by a mean coordinate with the following symbols: ■, Becker (1994); □, Paus (1993); ●, Braun (1997); ▲, Buckner (1995); ♦, Raichle (1994); ○, Warburton (1999); ▼, Petersen (1988); △, Petrides (1993); ▽, Petrides (1995); ⊡, Weiller (1995); ◇, Wise (1999). There is a relatively wide spread of data represented in the mean points; this is due to the fact that the data reported by authors were taken at face value, even in cases in which mean coordinates go outside the usual Brodmann areas associated with pre-Broca's area. The data reflect the fact that activation of pre-Broca's area is strongly left lateralized, especially when compared with the activation patterns in Broca's area proper.

shows a meta-analysis of functional imaging data reporting activation in BA 46/47 in language paradigms. This region is often, but not always, activated when language tasks include semantic processing, but it does not become active in phonological tasks. Some studies that report activation in the left frontal lobe do not differentiate between activation in this area and in BA 6/44, and it has been hypothesized that as a result, some coordinates reported for what is traditionally considered Broca's area are shifted forward. However, language tasks are not the only tasks to activate BA 46/47. Episodic memory (i.e., memory of materials encountered in specific episodes) and implicit memory, (i.e., experience-induced changes in performance on indirect tests unrelated to the experience) have each been linked to BA 46/47. Like Broca's area proper, this area cannot be simply classified as a language processing area.

1. Overt Speech Activation

BA 46/47 is activated in tasks that involve narrative speech, verb generation, and stem completion, suggesting that the area plays a role in semantic

processing. Specifically, verb generation in response to visual presentation of nouns activates BA 46/47, but repetition of words does not. In Petersen's study of single-word processing, activation from repetition of visually represented words is subtracted from that of generation of verbs: A left inferior frontal area was identified that almost certainly participates in processing for semantic association. Word repetition activates areas responsible for phonological processing, and subtracting those away from areas highlighted by generation of verbs reveals BA 46/47 as an area responsible for semantic processing. Weiler modified this subtraction by substituting pseudowords in the repetition condition, which might further reduce incidental activation of semantic areas. He drew a similar conclusion from the results: A modulatory, controlling function of nonautomatic, intrinsic generation processing can be attributed to the lateral prefrontal cortex. In a task generating spontaneous narrative speech and one constructing sentences from provided verbs, Braun demonstrated activation of BA 47. In the same study, a paced narrative task and an automatic speech task showed no activation in the area, which Braun attributed to lower semantic processing demands in the structured and automatic tasks. The multiple variations of semantic processing that are required in these tasks provide convincing evidence that BA 46/47 plays a key role in analysis of word meaning.

2. Working Memory Activation

Semantic working memory is another function attributed to the region of 46/47. The concept of semantic processing is basically retrieving word meanings; therefore, the line differentiating semantic working memory from semantic processing is uncertain. Gabrieli offered a view of BA 46/47 activation based on his study of semantic versus phonological analysis of words: Left inferior prefrontal cortex is activated to the extent that semantic information must be held temporarily in working memory to answer a particular semantic question. He observed a decrease in activation in the area, caused by a decreased need for processing, during a semantic priming task in which subjects differentiated between nouns based on semantic categories (living/nonliving or abstract/concrete). Repetition priming refers to the increase in efficiency produced when a task is performed with a set of stimuli already presented versus performance with novel stimuli, and the improvement reflects implicit memory acquired the first time a stimulus is presented. Smith argued that the area is divisible into two sections: area 47, which is responsible for maintaining verbal information in working memory, and area 46, which regulates the manipulation of maintained information. Working memory tasks for different modalities, such as objects or faces, activate different areas of the prefrontal region, but those that deal with verbal, semantic information consistently utilize pre-Broca's.

3. Nonlanguage Activation

Pre-Broca's area is activated during imagined and executed movements when no language is produced, and one study suggests that it is not necessarily activated in semantic processing, suggesting more expanded functions. Actions unrelated to semantics and language, such as overt hand and arm movement and imagined movement, also activate BA 46. Stephan found that both preparation to move and motor imagery activated the area, and Parsons found similar activation in a task judging whether drawings were of left or right hands. Parsons credits the activation to demands on working memory. Bookheimer compared oral object naming, oral word reading, silent object naming, and silent word reading tasks with viewing meaningless line drawings. The oral object naming, silent object naming, and silent reading activated BA 47 significantly, but oral word reading did not. Bookheimer noted that this indicates that semantic processing does not occur in oral reading if BA 47 is responsible for semantics. She suggested an alternative: This region might reflect activation of an articulatory code or motor plan corresponding to the recognized visual form. The variety of activation beyond semantic analysis makes it difficult to conclude that BA 46/47 acts strictly in a semantic capacity when it is engaged, and further evidence is needed to draw firm conclusions beyond the general finding that the area aids in processing and retrieving word meanings.

C. Sub-Broca's Area

Language production tasks often activate the left anterior insula, or sub-Broca's area, located beneath BA 44/6. The anterior insula is most likely activated in phonological and articulatory planning of speech, as opposed to semantic or lexical processing, as a result of its proximity to Broca's area proper. Figure 7 shows a meta-analysis of functional imaging data for activation of anterior insulae, left and right, from paradigms involving language. The posterior portion of the insula

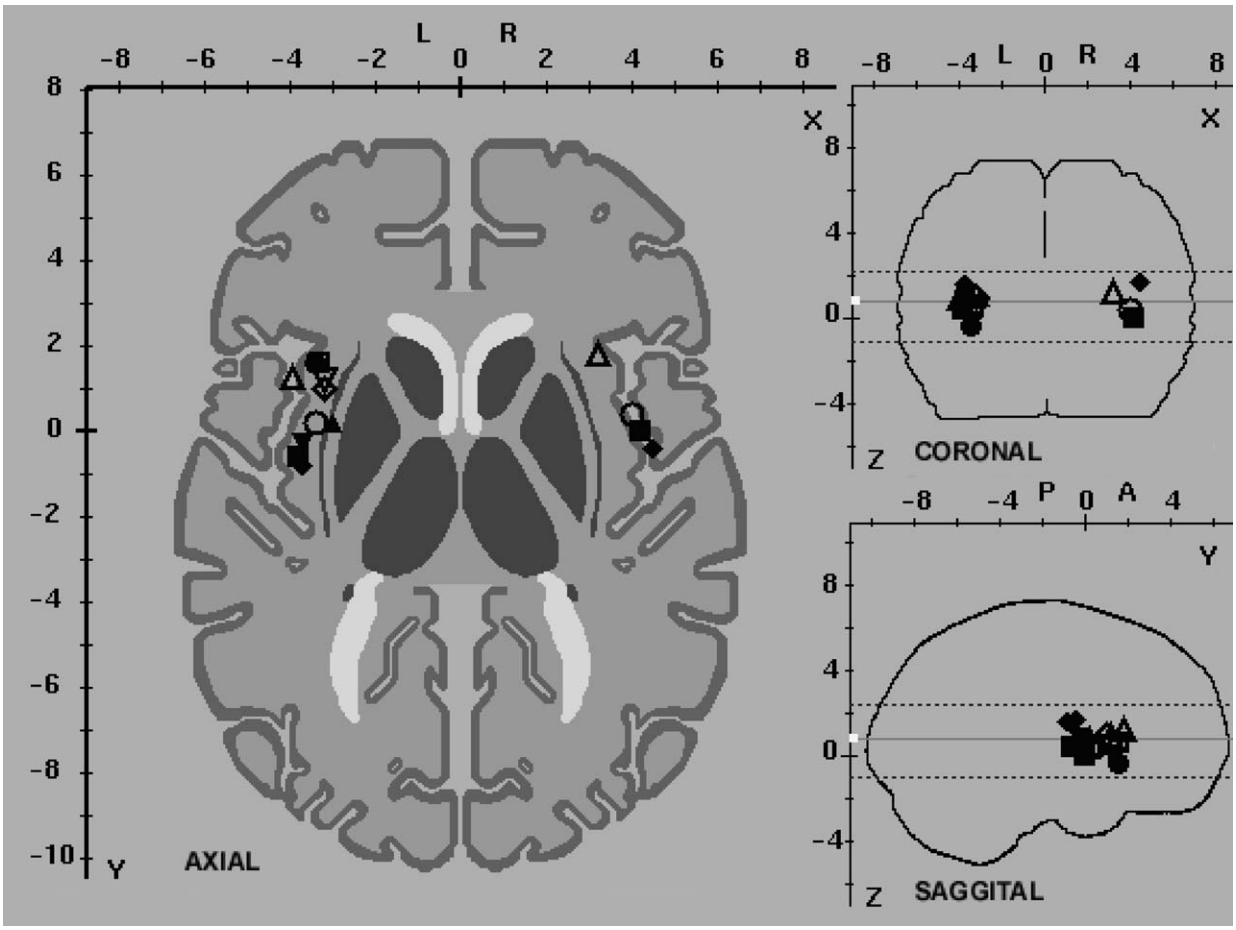


Figure 7 Activations in the anterior insula are shown in a meta-analysis of functional imaging data for left and right hemispheres from paradigms involving language. A set of mean coordinates was calculated from data reported from studies in Table II. The coordinates have been normalized and reported in standard three-dimensional (Talairach) space, plotted with the BrainMap database on an axial cross section at z=0.8 cm. Each relevant reference listed in Table II is represented here by a mean coordinate with the following symbols: ■, Becker (1994); □, Bookheimer (1995); ●, Braun (1997); ▲, Fox (1999); ◆, Raichle (1994); ○, Paulesu (1993); ▼, Fiez (1996); △, Price (1994); ▽, Rumsey (1997); ◇, Wise (1999). The distribution of activation points in both hemispheres shows that insular activation is not as strongly left lateralized as is activation of pre-Broca's area.

is associated with primary auditory and auditory association cortices and has been implicated in processing of auditory input. Braun related this function to a wide range of areas, noting that posterior insula serves as a parallel relay to prefrontal, motor, somatosensory, and cingulate regions of the brain. In addition to its role in speech, anterior insula has also been activated in overt and covert movement tasks. In general, functional imaging of the insulae is less abundant than that of BA 44/6 and 46/47. This trend may be due to the smaller size of the activated area combined with the less sensitive cameras and imaging equipment used in the past. In comparing language tasks that involve semantic, phonological, and articulatory processing, evidence indicates that left anterior insula is responsible for phonological and articulatory planning, but the area's relationship with semantics and the details of its role in articulation are debated.

1. Speech Activation

Left anterior insula is activated, often along with BA 44/6, in language tasks involving picture naming, practiced verb generation, reading aloud, and word repetition. All these tasks connect the area with language production, but the exact relationship of

the insular activation to the tasks follows two different trends of thought. Both agree that insula acts in articulatory planning, but they do not agree on its involvement in phonological and semantic analysis. The first set of opinions differentiated between automatic or rehearsed language tasks and novel tasks and proposed that insula is active in those tasks that are practiced. In a novel verb-generation task in a study by Raichle, the insula was inactivated relative to simple reading of nouns, and after the verb-generation task had been practiced the insula became active. In a silent counting task in a study by Fiez, insula was active, but when a verbal working memory component of retention of novel words was added to the task, insula became inactive. Fiez suggested that silent counting could be thought of as practiced speech. No distinction was made between phonological and articulatory functions in these studies; insular activation is attributed to the rehearsed nature of the speech, irrespective of the exact task being performed.

Another group of studies attempted to specify the purpose of insular activation based on the type of processing it performs or does not perform. Wise contrasted repetition of heard nouns with listening to nouns and anticipation of listening to nouns to differentiate between articulatory and phonological planning. Repetition is a task in which the phonetic plan (the selection and ordering of speech sounds and assignment of syllabic stress) is predetermined by input, and its execution is dependent only on the formulation of an articulatory plan. Left anterior insula was found to be active in both subtractions of listening and repeating nouns, which indicates its articulatory role. Rumsey showed that insula is inactive in orthographic decision making in a task in which subjects had to decide which of two homophones was a real word. She showed it is active in phonological processing in a task in which subjects had to decide which of two pseudowords would be the homophone of a real word. Price demonstrated that left anterior insula is active in a lexical decision-making task in which subjects had to identify real words among words with one incorrect letter substituted but not in a similar feature decision task in which subjects had to find false-font strings with ascending characters. Bookheimer argued that the insula is activated in articulatory coding. She differentiates between two types of motor speech pathways used in creating articulatory plans. Only one activates left anterior insula. She found activation in object naming, which she explained elicited a complete motor response: Subjects select a verbal label and the corresponding complete motor plan associated with that label. The alternate pathway that does not involve insula would be a sequential translation of orthographic units into sound patterns, as in word reading, with modification of the output based on auditory feedback. This differentiation of motor plan could explain Raichle's activation pattern in practiced word production if the practiced words are viewed and programmed as whole motor plans and not modified during production. Overall, activation in sub-Broca's area in speech production tasks indicates that the left anterior insula is active in articulatory planning and works closely with BA 44/6 in language tasks, but some details of the insula's role in language remain unclear.

2. Nonlanguage Activation

Nonlanguage tasks, such as overt and covert movement, activate the left anterior insula and demonstrate that the functions of the area extend to motor planning. Finger opposition and shoulder movement both activated insula in a PET study of movement by Colebatch, and imagined movement of a handheld joystick in a PET study by Stephan showed increased activation over a preparation to move baseline task. Right insula was activated by imagined movement of the hand in Parson's left/right hand decision study. He credited the insula with responding to somatosensory information from the body: The right insula activity observed is related to higher level somatic representation accompanying mentally simulated limb movement. Activation of left and right insulae in overt and covert movement shows that anterior insula is involved in complex nonspeech processes, and it demonstrates the need for more and varied functional imaging investigations of the relationship of language areas of the brain.

V. HEMISPHERIC LATERALITY

Hemispheric laterality of language function, still discussed and debated today, was introduced in the 19th century, and Broca brought it to the attention of the scientific community even though he did not believe in the theory. Although it is now well established that the left and right hemispheres of the brain are specialized for different functions and are anatomically asymmetrical, the exact nature of and reason for the differences have not been agreed on. The left hemisphere is generally described as dominant for

language, although individuals vary in the degree of dominance; however, as clinical, electrophysiological, and functional imaging data have shown, the right hemisphere plays an important role in language even in left-dominant individuals. Finding the basis for the laterality of language and its relationship to other characteristics, (e.g., gender) is another goal of functional imaging technology.

A. General Comments on Laterality

Even though language function has been placed in the left hemisphere historically, the right hemisphere has also been shown to be crucial for language function. One theory from 19th century science was that language dominance was connected with manual dominance since the majority of individuals were right-handed, the left side of the brain was known to control the right side of the body, and most cases of aphasia were documented with left hemisphere lesions. Right-handed individuals with aphasia and right hemisphere lesions in addition to left-handed aphasics with left hemisphere lesions have since been documented, and the connection between manual and language dominance has been severed. Another theory on laterality is that the left hemisphere is responsible for verbal functions, whereas the right is credited with visual and spatial perception. Both of these abilities have to do with language since visual and spatial perception influence visual language perception involved in reading and, as a result, in many functional imaging paradigms. The specific components of language have been placed in different hemispheres based on clinical and functional imaging data. For example, although the left hemisphere has long been thought of as responsible for semantics, the right hemisphere homolog to Broca's area has been shown to play a role in prosody, or the rhythmic and melodic qualities of speech production.

Stuttering studies are linked with the lateralization of language, and theories about reasons for stuttering revolve around hemispheric interactions in language. One long-standing explanation for stuttering is that it is a result of incomplete development of left hemispheric dominance for language, and another explanation is hyperactivity of the right hemisphere (nondominant for language), specifically the premotor cortex. In one PET study, Fox found overactivation in the right hemisphere motor language areas during stuttering and deactivation of left hemisphere auditory and frontal–temporal language production systems. When fluency was induced in speakers who usually stutter, the overactivations on the right decreased or disappeared and the left hemisphere deactivations were eliminated. Stuttering research is revealing complex interactions between the hemispheres in language production, and these interactions in such a clinical population may shed light on normal functional organization.

B. Laterality by Area

In the functional imaging literature, it is no longer a question of whether language is located in the left or right hemisphere but, specifically, which aspects and to what degree are those aspects of language located in which hemisphere. Each subdivision of Broca's area varies in the degree to which it appears on each side of the brain.

1. Broca's Area Proper Is Bilateral

Activation of BA 44/6 is found in both left and right hemispheres in right-handed individuals, although the majority of the activation is on the left and many studies do not focus on the right hemisphere's activation in language paradigms. Those that do report right hemisphere activation in PET and fMRI paradigms in normal subjects state that it is much weaker than corresponding left activation. Right Broca's is activated in some verb-generation and pseudoword repetition tasks as well as in verbal working memory tasks and a continuous performance task of identifying one letter in a stream of others.

2. Pre-Broca's Area Is Left-Lateralized

Few language studies report activation of right hemisphere BA 46/47, although there are a few paradigms that report activation below their individual standards for significance in reporting. This distinct left lateralization has not been explained in relation to more bilateral activation patterns in other language areas.

3. Sub-Broca's Area Is Bilateral

The anterior insula, although not always active along with Broca's area proper on the right, is more frequently active than BA 46/47 on the right. Since the left anterior insula has been linked to activation of Broca's area proper in tasks involving articulation, its

right-sided activation could be explained by a similar proximity argument to right BA 44/6.

C. Gender and Language Laterality

The effect of gender on laterality of speech functions has been and still is debated, with studies supporting male left lateralization and female bilaterality countered by others that demonstrate left lateralization in both genders. One study by Shaywitz shows that phonetic processing involving rhyme judgment clearly demonstrates male left lateralization in the inferior frontal gyrus and female bilaterality on the same task. A larger study of language comprehension by Frost demonstrated that men and women were both strongly left-lateralized in fMRI activation. This contradictory evidence demonstrates the need for continued, deeper investigation of the relationship between gender and hemispheric laterality.

VI. CONCLUSIONS

In closely examining Broca's area, several diverse literatures are available. Historical, electrical stimulation, clinical, and functional imaging studies contribute distinctive and sometimes seemingly contradictory viewpoints on the placement and function of Broca's original seat of articulate language. The way to reach firmer conclusions is to continue to find the discrepancies. In the future, more subdivisions may be created in Broca's area, and further extensions may be made into surrounding areas, for example, higher into BA 46 in order to relate higher order memory functions to current language functions.

See Also the Following Articles

APHASIA • CEREBRAL CORTEX • FUNCTIONAL MAGNETIC RESONANCE IMAGING (fMRI) • LANGUAGE DISORDERS • LANGUAGE, NEURAL BASIS OF • LATERALITY • LEFT-HANDEDNESS • SPEECH • WERNICKE'S AREA • WORKING MEMORY

Suggested Reading

Fiez, J. A., and Petersen, S. E. (1998). Neuroimaging studies of word reading. *Proc. Natl. Acad. Sci. USA* **95,** 914–921.

Gabrieli, J. D. E., Poldrack, R. A., and Desmond, J. E. (1998). The role of left prefrontal cortex in language and memory. *Proc. Natl. Acad. Sci. USA* **95,** 906–913.

Garey, L. J. (Trans.) (1994). *Brodmann's. "Localisation in the Cerebral Cortex."* Smith-Gordon, London.

Geschwind, N., and Galaburda, A. M. (1987). *Cerebral Lateralization: Biological Mechanisms, Association, and Pathology*. MIT Press, Cambridge, MA.

Mohr, J. P. (1978). Broca's aphasia: Pathologic and clinical. *Neurology* **28,** 311–324.

Ojemann, G., Ojemann, J., Lettich, E., and Berger, M. (1989). Cortical language localization in left, dominant hemisphere. *J. Neurosurg*. **71,** 316–326.

Penfield, W., and Roberts, L. (1959). *Speech and Brain Mechanisms*. Princeton Univ. Press, Princeton, NJ.

Petersen, E., Fox, P. T., Posner, M. I., Mintun, M., and Raichle, M. E. (1988). Positron emission tomographic studies of the cortical anatomy of single-word processing. *Nature* **331,** 585–589.

Schiller, F. (1979). *Paul Broca*. Univ. of California Press, Berkeley.

Von Bonin, G. (1960). *The Cerebral Cortex*. Thomas, Springfield, IL.

Young, R. M. (1970). *Mind, Brain and Adaptation in the 19th Century: Cerebral Localization and Its Biological Context from Gall to Ferrier*. Clarendon, Oxford.

Zurif, E. B. (1990). Language and the brain. In *An Invitation to Cognitive Science: Language* (D. N. Osherson and H. Lasnik, Eds.), pp. 177–198. MIT Press, Cambridge, MA.

Cancer Patients, Cognitive Function

CHRISTINA A. MEYERS
University of Texas M. D. Anderson Cancer Center

GLOSSARY

cytokines Molecules that communicate between immune cells and between immunocytes and other peripheral cells, such as fibroblasts and endothelial cells. They are involved in regulating the initiation and maintenance of the immune response.

paraneoplastic syndrome Brain dysfunction directly resulting from a cancer that is not present in the nervous system.

subcortical dementia Dysfunction or injury to the white matter tracts connecting the frontal lobe regions to subcortical targets, causing a clinical syndrome characterized by neurobehavioral slowing, memory loss, frontal lobe dysfunction, and mood changes.

Many cancer patients experience impairments of neurocognitive function, including memory loss, distractiblity, difficulty with multitasking, mood disturbance, and a myriad of other symptoms. Patients may also suffer from symptoms that compromise their ability to function adequately, such as fatigue and pain. The etiologies of these problems are diverse and include the direct effects of cancer within the central nervous system, indirect effects of certain cancers (the paraneoplastic disorders), and effects of cancer treatment on the brain. In addition to these cancer-related causes, patients may have co-existing neurologic or psychiatric disorders that affect their cognition and mood. Careful assessment of patients complaining of neurocognitive or behavioral problems is essential to provide appropriate interventions and to maximize their quality of life.

I. CANCER IN THE CENTRAL NERVOUS SYSTEM

Primary brain tumors, metastatic brain tumors, and leptomeningeal metastases directly alter the functioning of the brain and hence the individual's mental and cognitive functioning. Primary brain tumors are increasing in prevalence, with more than 29,000 new cases diagnosed each year in the United States. In addition, between 20 and 40% of patients with non-central nervous system (CNS) solid tumors develop brain metastases. Cognitive impairments are related to lesion location, tumor growth potential, and host characteristics.

A. Impairments Related to Tumor Site

Although the cognitive symptoms caused by tumors are relatively focal in nature (e.g., the symptoms correspond to expected brain–behavior relationships), the cognitive impairments tend to be less dramatic than those seen in patients with more sudden-onset lesions. Patients with left hemisphere tumors are thus more likely to have impairments of language function, verbal reasoning, and right-sided motor dexterity. Patients with right hemisphere tumors have more difficulties with visual–perceptual skills, visual

scanning, and left-sided motor dexterity. Memory loss is often seen in association with tumors of either hemisphere. Impairments of frontal lobe function (e.g., executive deficits manifested by impairments of cognitive flexibility, abstraction, motivation, planning, and organizational skills and also the ability to benefit from experience, personality changes, etc.) are ubiquitous in brain tumor patients because the frontal lobes comprise one-third of the cerebrum, and thus a large proportion of patients have frontal tumors. However, patients with nonfrontal tumors may also exhibit executive deficits because the frontal lobes have rich afferent and efferent connections with all other brain regions, disrupting the bidirectional modulatory influences on cognition and thus causing impairment in social functioning, motivation, and judgment.

B. Lesion Momentum

Individuals who present with low-grade tumors that have been present for many years may have no detectable changes in brain function due to cerebral plasticity and reorganization, whereas those with very rapidly growing tumors may have widespread impairment due to mass effect on adjacent brain regions. Patients with gliomatosis cerebri, a condition in which tumor cells diffusely infiltrate large brain regions, may have few focal impairments since neuronal function is not substantially altered. Patients with metastatic brain tumors may also present without focal or severe deficits since neuronal tissue is displaced by the lesion and not necessarily damaged.

C. Host Characteristics

The age of the patient is also a factor in the severity of neurobehavioral deficits. The incidence of histopathologically more aggressive tumors, such as glioblastoma multiforme, increases with age. However, histopathologically less malignant tumors, such as anaplastic astrocytoma, also behave more aggressively in the older patient. Older patients are also at higher risk for having other concurrent neurodegenerative illnesses, such as Parkinson's disease or vascular disease. Finally, very old and very young patients may be more sensitive to the toxic side effects of treatment.

II. CNS EFFECTS OF NON-BRAIN CANCER

Certain types of cancers cause brain dysfunction indirectly, causing paraneoplastic brain disorders. It has been estimated that 10% of cancer patients develop such a syndrome. Most commonly associated with small cell lung cancer (SCLC), paraneoplastic brain disorders typically manifest as a diffuse encephalomyelitis, including limbic encephalitis, cerebellitis, brain stem encephalitis, myelitis, and sensory or autonomic ganglionitis. The cognitive deficits associated with these disorders are of two types. Some patients develop a gradually progressive subcortical dementia that is difficult to distinguish from dementia due to other causes. The second form, paraneoplastic limbic encephalitis, is often characterized by the subacute, progressive onset of anxiety, depression, confusion, hallucinations, recent memory loss, or seizures. The onset of symptoms can precede the diagnosis of cancer even by years. These disorders have an autoimmune pathogenesis and are associated with high titer serum autoantibodies, usually of the IgG-1 subtype. The most common antibody identified in patients with paraneoplastic disorders is anti-Hu, which recognizes similar neuronal and autologous SCLC protein antigens. Recently, it has been determined that a large proportion of patients with SCLC experience subclinical alterations in memory and frontal lobe executive function prior to the institution of any treatment. A variety of anti-neuronal antibodies, particularly anti-Hu, have been found in 16% of SCLC patients without florid paraneoplastic neurologic disease, but correlation of subtle cognitive or neurologic dysfunction in SCLC patients with low titer antibody responses has not been performed. Other paraneoplastic syndromes include subacute cerebellar degeneration, which is associated with an anti-Purkinje antibody in ovarian cancer patients, and a rare autoimmune paraneoplastic encephalitis associated with testicular cancer.

Some cancer cells produce ectopic hormones or proinflammatory cytokines. These can have profound effects on brain function and are known to cause cognitive and mood dysfunction. For instance, the cytokines interleukin-1 (IL-1), IL-6, and tumor necrosis factor (TNF-α) are elevated in patients with acute leukemia and myelodysplastic syndrome. A more thorough discussion of the effect of cytokines on brain function is provided in Section III.C. A substantial proportion of untreated patients with these hematologic malignancies have neurocognitive impairments.

III. EFFECTS OF ANTINEOPLASTIC THERAPY

Chemotherapy, radiotherapy, immunotherapy, novel agents, and hormonal treatments all have potential neurotoxic side effects. Cognitive impairments due to treatment effects are generally not due to mood disturbance in the clinical trials that assess both cognition and mood.

A. Chemotherapy

Many cancer patients experience difficulties with short-term memory during cancer therapy. However, approximately 18% of cancer patients who have received standard-dose chemotherapy manifest persistent cognitive deficits after treatment is completed. The risk appears to be greater after high-dose chemotherapy with stem cell rescue—on the order of one-third of patients. This risk is evident 2 years after treatment, but the longer term effects of high-dose therapy remain unknown. Although most impairments related to chemotherapy are relatively diffuse, affecting sustained attention and speed of information processing, some agents have more circumscribed effects on the brain due to their mechanism of action. For instance, a mitotic inhibitor that binds to the colchicine receptor on tubulin and crosses the blood-brain barrier causes a highly specific decline in memory functioning. There are also many case reports of patients developing confusion and behavioral changes after the administration of chemotherapy agents not thought to cross the blood–brain barrier. The mechanism underlying these events is unclear and may reflect second messengers, metabolic disturbance secondary to other organ toxicities, differences between human subjects and the animals that were used for preclinical toxicity assessments, or unanticipated breach of the blood–brain barrier. Finally, some individuals may be at greater risk to develop neurotoxicity from chemotherapy due to preexisting host characteristics. For example, people with dihydropyrimidine dehydrogenase deficiency can develop severe neurotoxic reactions to 5-fluorouracil, a widely used chemotherapy agent.

B. Radiation Therapy

Radiation therapy to the brain is the mainstay treatment for primary and metastatic brain tumors. Due to the high incidence of cancer relapse in the brain, prophylactic brain irradiation is widely used in patients with lung cancer and acute leukemia. Cognitive domains impaired as a result of radiation therapy include information processing speed, executive functions, memory, sustained attention, and motor coordination. Anatomically, periventricular white matter hyperintensities are observed on neuroimaging. The pattern of deficits is consistent with those seen in other subcortical diseases of white matter such as multiple sclerosis. Children are particularly vulnerable to radiation injury to the brain. Even relatively low-dose cranial irradiation can cause mild declines in intelligence in older children, with more severe impairments in memory, acquisition of academic skills, and attentional skills in children treated before age 5. In both adults and children there is conflicting evidence whether concomitant chemotherapy is synergistically toxic. Finally, many patients with head and neck cancers, such as paranasal sinus tumors and anterior skull base tumors, receive local radiation that incidentally injures the brain.

C. Immunotherapy

Immunotherapy with proinflammatory cytokines is widely used in the treatment of chronic leukemia, renal cell carcinoma, and melanoma. As mentioned previously, cytokines have profound effects on brain function. In fact, more than half of patients receiving cytokine treatment have documented cognitive impairments, which is considerably higher than the rate of cognitive dysfunction in patients treated with cytotoxic agents only. Interferon-α(IFN-α) is the most widely used immunotherapy agent and has a wide array of effects on CNS function. Patients on IFN-α may develop cognitive deficits involving information processing speed, verbal memory, and frontal lobe executive functions as well as depression, a pattern suggestive of frontal–subcortical dysfunction.

Impairments of cognition and mood may occur separately or conjointly and often require dose reduction or termination of an otherwise effective treatment. In a few patients cognitive deficits do not appear to be reversible after treatment cessation. The length of treatment, dose, and route of administration are all important factors in the development of neurotoxic side effects. Mechanisms of IFN-α's action on the CNS include both stimulation and inhibition of the hypothalamic–pituitary–adrenal axis and release of neuroendocrine hormones, effects on thyroid function,

induction of secondary cytokines that are neurotoxic, and alterations of neurotransmitter pathways, particularly the opioid–dopamine and serotonergic systems.

Other cytokines have also been used in clinical trials against cancer, and the cognitive impairments tend to be similar to those seen with IFN-α. IL-2 and TNF can cause memory deficits, difficulties with motivation and flexible thinking, motor dyscoordination, depression, and anorexia. Visuoperceptual and language functions tend not to be affected. TNF exhibits dose-dependent toxicity such as decreased attentional abilities, verbal memory deficits, motor coordination impairments, and frontal lobe executive dysfunction. Headache, anorexia, stroke-like events (e.g., transient amnesia), and demyelination in the brain are also adverse effects of TNF. IL-1 and its receptors are found in many areas of the brain, particularly the hippocampus. IL-1 suppresses the influx of calcium into the hippocampus neurons, which may explain the preponderance of memory impairments in patients with IL-1-associated toxicity. The neurotoxic effects of IL-2 appear to be dose related and occur in nearly all patients treated with high dosages. The symptoms range from mild agitation, depression, and forgetfulness to frank confusion, dementia, and paranoia. TNF and IL-2 and TNF and IL-1 have been found to be synergistically toxic with the latter being associated with the development of multiple sclerosis plaques and gliosis. CNS expression of IL-6 has been associated with inflammatory neurodegeneration and learning impairments in mice and in patients with Alzheimer's disease.

In addition to their direct effects on brain function, these cytokines provoke a stress hormone cascade that can affect mood and cognition. The mechanism through which cytokines exert their influence is difficult to discern given their overlapping immunological, hormonal, and inflammatory effects. Cytokines act in cascades, making determination about specific effects of an individual cytokine even more difficult. Cytokines are thought to enter the brain via the circumventricular organs, particularly the organum vasculosum of the lamina terminalis. This structure is associated with the hypothalmus, which has connections with the brain stem, frontal cortex, and the limbic system. Treatments directed against cytokine neurotoxicity are symptomatic and include stimulant therapy for fatigue and neurobehavioral slowing, opiate antagonist therapy for cognitive deficits, and antidepressant therapy for mood disturbance.

D. Hormone Ablation Therapy

Hormone ablation is typically used to treat (i) prostate cancer, via either orchiectomy or androgen antagonists, and (ii) hormone receptor-positive breast cancer, with estrogen antagonists. There is a burgeoning literature on the effects of sex hormones on cognitive functioning. Serum testosterone may be related to spatial ability. Women on gonadotropin-releasing hormone agonists for gynecological problems, which inhibit both testosterone and estrogen release, have reported declines in their memory functioning and mood. These deficits appear to be reversed with estrogen replacement. Tamoxifen is the most widely used hormonal agent in the treatment of breast cancer. The most commonly reported adverse effects associated with Tamoxifen include hot flashes, nausea, and vaginal bleeding. Additionally, fatigue and inability to concentrate have been reported as well as mood changes, including depression, irritability, and nervousness. Although the pharmacodynamic properties of Tamoxifen, especially its CNS effects, are not well understood, there is reason to believe that Tamoxifen therapy is related to the cognitive and emotional difficulties some women experience secondary to its antiestrogen activities. Additionally, *in vivo*, *in vitro*, and animal studies suggest that Tamoxifen affects several other neurotransmitter (e.g., seratonin and dopamine) and cytokine (e.g., IL-1, IL-6, IFN-γ, and TNF) systems that are implicated in cognitive functioning. Major depression (which occurs in 1–15% of patients on Tamoxifen therapy) has been associated with increases in IL-1, IL-6, IL-2, TNF, and IFN-γ and is associated with symptoms such as psychomotor retardation, malaise, lassitude, anxiety, anhedonia, and sleep disturbances.

E. Adjuvant Medical Treatment

In addition to neurotoxic effects of primary cancer therapy, adjuvant medications may also cause neurocognitive symptoms, complicating the assessment of patients who are on multiple medications. Drugs with cognitive and mood effects include steroids, pain medications, psychotropic medications, and antiemetics. Immunosuppressive agents, such as cyclosporine, are widely used in bone marrow transplantation and may also cause adverse CNS effects. Neurotoxicity due to cyclosporine and other immunosuppressive agents may not always be reversible and can cause observable anatomic brain injury.

IV. COMORBID CONDITIONS

A. Neurologic Diseases

The elderly comprise the largest percentage of cancer patients. Patients in older age groups are not only at increased risk to develop cancer but also more vulnerable to age-related neurocognitive disorders unrelated to but coexisting with cancer, such as cerebrovascular disease and dementia of the Alzheimer's type. Cancer patients may also have a preexisting history of traumatic brain injury, developmental disorder, multiple sclerosis, and other conditions or diseases that affect cognitive functioning. These patients may be more vulnerable to develop neurotoxic side effects from cancer and cancer treatment and need to be monitored closely.

B. Sensory Impairment

Elderly patients may have difficulties with hearing or sight or they may suffer from general frailty. Many patients have also experienced ototoxicity and other sensory problems from chemotherapy. These sensory deficits reduce the amount of information patients are able to process and may lead to complaints of memory dysfunction. Some patients even manifest psychiatric difficulties, such as paranoia, because a relative sensory deprivation impedes their ability to appreciate what is going on around them.

C. Functional and Psychiatric Disorders

Stress, anxiety, depression, and other mood disorders can be reactive to the patients' current situation or can predate the cancer diagnosis. In either case, mood and adjustment disorders negatively affect patients' ability to focus, concentrate, and organize activities, leading to complaints of forgetfulness and other cognitive problems. Cancer patients are just as likely as the general population to suffer from major psychiatric disorders, such as bipolar illness, major depression, schizophrenia, and personality disorders. These disorders also have associated cognitive symptoms if they are not under good control with psychiatric management.

D. Fatigue, Pain, and Anemia

Fatigue is extraordinarily common in cancer patients and is now recognized as a more widespread and problematic symptom than pain. Cancer-related fatigue is defined as an unusual, persistent, subjective sense of tiredness related to cancer or cancer treatment that interferes with usual functioning. It can occur during active cancer treatment and it can persist long after treatment has ended. The causes are myriad and overlap considerably with those that can cause cognitive disorders. Fatigue may be physical in that the person has very little stamina or energy to perform usual activities. Fatigue can also be mental. Similar to patients with cognitive dysfunction, patients who suffer from mental fatigue often report that they are easily overwhelmed, that they have difficulty being organized and efficient in their daily activities, and that they have difficulty meeting deadlines or getting things done on time. Activities that used to be automatic now require more effort so that the patients become exhausted even performing routine tasks.

A similar situation exists for patients who experience pain. Patients in pain also suffer from deficits in attention and concentration, multitasking, and speed and efficiency of thinking. They may be on medications that are sedating and contribute to cognitive problems, and they are likely to be suffering from fatigue as well.

Anemia is very common in cancer patients undergoing active treatment. Cognitive deficits have been reported in well-dialyzed patients with end-stage renal disease who are anemic but do not have elevated uremia. Cognitive deficits are also exhibited by patients who have anemia due to iron deficiency. The cognitive problems observed on neuropsychological testing include deficits in attention, perceptual motor speed, memory, and verbal fluency and are accompanied by slowed auditory evoked potentials. There is a high degree of correlation between hemoglobin levels and fatigue. However, patients with anemia may experience cognitive difficulties separately from those due to fatigue alone, possibly due to reduced cerebral blood flow. These cognitive deficits and slowed evoked potentials often improve following reversal of anemia with erythropoietin or blood transfusion.

E. Other Medical Complications

Alterations of thyroid functioning are common in cancer patients and can be associated with cognitive and mood disorders. Patients are at risk for all types of infections and may develop significant cognitive problems or even delirium. Hepatic and renal dysfunction also can cause cognitive problems in acutely ill

patients. In these patients, treatment of the underlying disease will often result in resolution of the cognitive problems as well. Immunocompromised patients, such as individuals undergoing bone marrow transplantation, may be at risk for developing viral encephalitis and persistent amnesia following recovery from an acute viral infection.

V. ASSESSMENT

It is becoming increasingly common to assess cognitive functioning and symptoms experienced by cancer patients in clinical trials of new anticancer therapies. As defined by a working group of members of the Food and Drug Administration, National Cancer Institute (NCI), and the NCI Division of Cancer Treatment Board of Scientific Counselors, net clinical benefit of cancer therapy includes: (i) survival benefit, (ii) time to treatment failure and disease-free survival, (iii) complete response rate, (iv) response rate, and (v) beneficial effects on disease-related symptoms and/or quality of life. Especially for regimens that differ only slightly with respect to response and survival, the rationale for selecting a particular therapy may be highly related to the impact of that treatment on cognitive function, symptoms, and quality of life.

As can be seen from the previous discussion, the specific causes of cognitive dysfunction is critically important to guide interventions because the type of intervention most helpful will be dramatically different depending on the etiology. The specific intervention plan not only takes into account the underlying cause of the complaint but also needs to be individualized for the person because the impact of a cognitive problem will be different for different people. A practical system to evaluate the impact of neurocognitive dysfunction was developed by the World Health Organization, which classifies the impact of an illness in three domains; impairment, disability, and handicap. Impairment is the deficit in function. In the case of neurocognitive symptoms, the deficit is in the function of the brain and manifested by neurologic, cognitive and emotional changes. Formal assessment of neurocognitive function can help determine the etiology of the complaints and the profile of the cognitive changes, and it can also help in the institution of appropriate intervention strategies. Disability is the impact of the deficit in the patient's ability to perform usual work and home activities. The degree of disability for an individual patient will be at least partially related to age, the type of work performed, and the amount of support that is available. Performance and functional status measures may help to define the disability and help determine the need for more comprehensive assessments. Handicap is the impact of the disability on the person's overall satisfaction and well-being. Handicap is generally what is referred to when discussing quality of life (QOL) and is often assessed by QOL questionnaires. Again, handicap is very individual. One person can be handicapped by a relatively minor disability, whereas another individual suffering from a severe impairment may experience little handicap. For example, an impairment in multitasking caused by a difficulty with sustained attention is a common problem for cancer patients. This may not be particularly handicapping to a person who is self-employed and can work at his or her own pace at home. However, it might cause a secretary in a busy office, who needs to answer the phone while word processing, to lose his or her job. Thus, all three levels of function (deficit, disability, and handicap) need to be assessed for appropriate management of the patient.

Many cancer patients have difficulty resuming their normal activities following diagnosis and treatment. Unfortunately, neurobehavioral functioning is often the least addressed aspect of a medical evaluation unless very severe behavioral changes are apparent. Multidisciplinary assessment of neurocognitive complaints can maximize patients' ability to function at the highest level of independence and productivity for the longest duration of time. As cancer treatment becomes more successful there will be increasing numbers of patients who live longer and expect to return to their preillness level of functioning. The risks of treatment and impact on the patient's ability to perform activities of daily living must be addressed more comprehensively. Many intervention strategies are available, including pharmacologic management, behavioral strategies, life-style alterations, formal rehabilitation, and counseling.

VI. INTERVENTION STRATEGIES

A. Pharmacologic Treatment

Fatigue and neurobehavioral slowing are ubiquitous in cancer patients. Stimulant treatment can be very useful in the treatment of concentration difficulties, psychomotor slowing, and fatigue and can help to elevate mood as well. Aggressive treatment of mood disturbance, pain, and fatigue also has the potential to

improve neurocognitive functioning when these symptoms are affecting patient function. New agents are becoming available that intervene on a mechanism-based level, such as cytokine antagonists. In addition, agents that afford neuroprotection from the toxic effects of cancer therapies are currently being developed.

B. Rehabilitation

Physical and occupational therapy can be of tremendous benefit for patients who have developed weakness, generalized deconditioning, sensory neuropathy, or other nerve or musculoskeletal impairments from their cancer and cancer treatment. Both of these modalities can improve mobility and general physical stamina. Patients who have developed impairments of upper extremity function due to peripheral neuropathy from systemic therapy or focal radiation or from central lesions of the spinal cord or brain may benefit from occupational therapy to help improve function. Patients who have developed difficulties with swallowing, articulation, or even primary language problems may benefit from speech therapy.

Many cancer patients who experience neurocognitive declines can improve their function at home and in vocational and leisure pursuits and enjoy an improved level of independence and quality of life given the right support. Given that cancer patients tend to have fairly mild and restricted cognitive problems, they often respond very well to focused rehabilitation efforts. A preliminary study of cognitive and vocational therapy provided to brain tumor patients found that these patients required shorter stays, had less treatment, and had better overall outcome in terms of independence and productivity compared to patients with traumatic brain injuries. Cognitive rehabilitation is designed to improve independence level, whereas vocational rehabilitation is designed to improve productivity, which may include volunteer work, performing household activities, going back to school, working at a modified job, or maintaining competitive employment.

C. Education and Support

Patient and family education is also extremely important. Potential neurobehavioral symptoms may not be explained to the patient, sometimes because the primary physician is not aware of the impact of even subtle symptoms on social and vocational functioning. Patients and families may feel isolated and alone when they experience neurobehavioral symptoms. The more knowledge a patient and family have about the disease, treatment, and expected problems, the more effectively they can deal with the care of the patient. Even simple coping strategies, such as taking intermittent naps, writing notes, and taking special care to plan and organize activities, may be sufficient to effectively cope with symptoms. Support groups and counseling can also be very helpful in assuring patients and families that their experiences are not unusual and in helping them deal with the grief, anger, frustration, and other problems that are frequently manifested over the course of the disease.

It is imperative that in developing new treatments for cancer that provide improved outcomes in terms of survival, there is also consideration about the long-term sequelae of cancer treatment. Cancer survivors should be able to expect to return to a productive and fulfilling life following the cancer experience. Support for this effort will not only reduce disability and improve patient function and satisfaction but also reduce overall burden and cost to society that such disability causes.

See Also the Following Articles

ALZHEIMER'S DISEASE, NEUROPSYCHOLOGY OF • BRAIN LESIONS • BRAIN DISEASE, ORGANIC • COGNITIVE REHABILITATION • DEMENTIA • HIV INFECTION, NEUROCOGNITIVE COMPLICATIONS OF • MEMORY DISORDERS, ORGANIC

Suggested Reading

Crossen, J. R., Garwood, D., Glatstein, E., and Neuwelt, E. A. (1994). Neurobehavioral sequelae of cranial irradiation in adults: A review of radiation-induced encephalopathy. *J. Clin. Oncol.* **12,** 627.

Dantzer, R., Wollman, E. E., and Yirmiya, R. (Eds.) (1999). *Cytokines, Stress, and Depression*. Kluwer, New York.

Erlanger, D. M., Kutner, K. C., and Jacobs, A. R. (1999). Hormones and cognition: Current concepts and issues in neuropsychology. *Neuropsychol. Rev.* **9,** 175.

Lipp, H.-P. (Ed.) (1999). *Anticancer Drug Toxicity: Prevention, Management, and Clinical Pharmacokinetics.* Marcel Dekker, New York.

Meyers, C. A., and Valentine, A. D. (1995). Neurological and psychiatric adverse effects of immunological therapy. *CNS Drugs* **3,** 56.

Osoba, D. (Ed.) (1991). *Effect of Cancer on Quality of Life*. CRC Press, New York.

Ris, M. D., and Noll, R. B. (1994). Long-term neurobehavioral outcome in pediatric brain-tumor patients: Review and metholdological critique. *J. Clin. Exp. Neuropsychol.* **16,** 21.

Vecht, Ch. J. (Ed.) (1997). *Handbook of Clinical Neurology: Neuro-Oncology, Part I*. Elsevier, Amsterdam.

Catecholamines

ARNOLD J. FRIEDHOFF† and RAUL SILVA
New York University Medical Center

GLOSSARY

antidepressant drugs Medications used to treat a variety of conditions, such as depression, panic attacks, and obsessive–compulsive disorder. Generally, there are four different groups: (i) the tricyclic antidepressants, which include imipramine and the related agents amitriptyline and nortriptyline; (ii) the monoamine oxidase inhibitors, such as tranylcypromine and phenelzine; (iii) the newest group, the serotonin reuptake blockers, which include anafranil, fluoxetine, sertraline, fluvoxamine, citalopram, and paroxetine; and (iv) others such as trazadone and bupropion.

antipsychotic drugs Medications used for a variety of conditions. The name is derived from the improvement they produce in certain psychotic behaviors, such as delusions and hallucinations. The first such agent, chlorpromazine, was synthesized circa 1950. New atypical agents include clozapine, risperidone, olanzapine, ziprasidone, and quetiapine. These medications are also called neuroleptics.

catecholamines Three endogenously produced substances—epinephrine, norepinephrine, and dopamine—that serve as neurotransmitters. They are powerful chemicals that can be found in neurons throughout the body. The effects of these compounds are responsible for the functioning of the brain even during the early fetal stages of life. They help regulate an endless number of functions ranging from thinking and mood to motor control.

limbic system A group of structures located in the brain that are involved in regulating emotion and its association with behavioral and mental functioning.

neurotransmitters Compounds that are released into interneuronal junctions called synapses. They are released from the axon of a presynaptic neuron and impact on the receptors of the postsynaptic neuron, the nerve cell on the other side of the synapse. This is the chemical means by which transfer of information occurs in the brain.

†Deceased.

Catecholamines comprise three endogenously produced substances—epinephrine, norepinephrine, and dopamine—that serve as neurotransmitters. They are powerful chemicals that can be found in neurons throughout the body. The effects of these compounds are responsible for the functioning of the brain even during the early fetal stages of life. They help regulate an endless number of functions ranging from thinking and mood to motor control. In this article, we review the structure, the anatomical distribution, and the role that these substances play in functioning and behavior.

I. NATURE OF THE CATECHOLAMINERGIC SYSTEMS

Catecholamines are relativity small organic molecules that function in the brain and elsewhere in the body, primarily in a regulatory or modulating role, to keep various systems functioning smoothly in response to demands of the internal and external environment. The most familiar of the three natural catecholamines is adrenaline, or epinephrine (Fig. 1). Its effects have been experienced by all of us, for example, in response to a frightening experience. Its release into the bloodstream from the neuronal cell bodies located in the medulla of the adrenal gland regulates heart rate and blood pressure and helps to put us into a readiness state for fight or flight. Norepinephrine, the closest chemical relative of epinephrine, is more prominently localized in the brain than epinephrine, but it is also found in so-called peripheral neurons (those neurons found outside of the brain). In the brain norepinephrine regulates mood and level of emotional arousal and

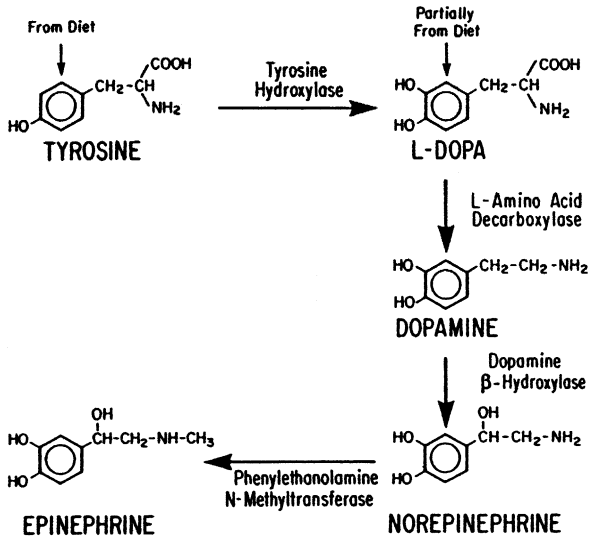


Figure 1 Pathway for biosynthesis of the major catecholamines (reproduced with permission from Friedhoff and Silva, 1993).

alertness. Dopamine, the third catecholamine, is prominently involved in regulating motor or movement functions and also in the coordination of associative thinking and integration of sensory motor function. Thus, key volitional acts such as movement and thinking are fine-tuned, integrated, and given emotional coloration through the actions of the three catecholamines.

A. Neurotransmission

Information transfer in the brain is carried out mainly by synaptic transmission, or the passage of a message across synapses or gaps between communicating cells. This occurs through a combination of electrical transmission that takes place within a neuron and the release of a chemical or neurotransmitter that crosses the synaptic gap and then acts on a postsynaptic neuron via specialized detection sites called receptors. However, there are some exceptions to this general model. For example, some neurons (not catecholaminergic) relate to each other entirely by change in electrical potential. In many cases involving the catecholaminergic system, other substances are coreleased with the neurotransmitter and modify or modulate its effect. The nature of the effector response can vary depending on the type of receptor, location of the membrane, and the nature of the neuromodulators. For example, the stimulation of β_3-adrenergic receptors located in adipose tissues will stimulate the breakdown of fats (lipolysis). This can be contrasted with the stimulation of different α_2-adrenergic receptors, one of which may inhibit the release of certain neurotransmitters at the presynaptic level of adrenergic nerve cells, thereby causing inhibition of norepinephrine release. Stimulation of another α_2-adrenergic receptor located on the membranes of the β cells of the pancreas will cause a decrease in insulin secretion.

The synapse is an important locus for the action of drugs that modify behavior. By blocking reuptake of transmitters, the effect of the transmitter can be enhanced or exaggerated. Conversely, by blocking receptors on postsynaptic cells, transmitter effect can be reduced. A third possibility, which has been exploited pharmacologically, is the modification of the ion exchange involved in electrical transmission. This too can have effects on motor and mental activity.

It should also be kept in mind that neurotransmission in the catecholaminergic system can be altered by impingement of other transmitter systems. One example is the glutamatergic system. Neural projections from this system can modulate dopamine release in the cortex. Pharmacologic antagonists of this system, such as phenylcyclidine, can increase dopamine outflow at the level of the striatum. Strikingly, in the behavioral realm, the administration of this agent produces a clinical picture that mimics the psychotic processes that have been associated with abnormalities in the dopaminergic system.

Finally, when discussing neurotransmission the termination of the signal is of paramount importance. If the transmitter is not removed from the synaptic cleft, a desensitization of the postsynaptic receptor will occur as a result of continued catecholamine exposure. Reuptake of neurotransmitters is a common mechanism for inactivation. This process facilitates the removal of the transmitter and affords the ability to reuse the molecule. There are transporter molecules located on the membranes of terminals that produce high-affinity bonds and facilitate these mechanisms. Theoretically, these molecules are also capable of releasing bonded substances and thus producing postsynaptic activation. Nonetheless, the catecholamine transporter is an exciting area of research.

B. Biosynthesis of Catecholamines

The starting point for the synthesis of all the catecholamines is L-tyrosine, which is a nonessential amino acid that can be found in the diet. Synthesis of these neurotransmitters may vary, depending on dietary

consumption of tyrosine-containing products. L-Tyrosine is hydroxylated (it gains an OH group) to form dihydroxy-L-phenylalanine, which is also known as levodopa or L-dopa. The enzyme responsible for this transformation is tyrosine hydroxylase. In dopaminergic neurons, L-dopa is metabolized to dopamine by means of the enzyme dopa decarboxylase. This enzymatic process occurs in the cytoplasmic component of neurons. In noradrenergic nerve cells and in the adrenal medulla, dopamine is transformed to norepinephrine. This is facilitated by the enzyme dopamine β-hydoxylase. It has been estimated that approximately 50% of the dopamine synthesized in neuronal cytoplasm of noradrenergic cells is metabolized to norepinephrine. Norepinephrine can then be transformed to epinephrine by the addition of a methyl group (CH_3) to its amino group through the action of the enzyme phenethanolamine-*N*-methyltransferase. This last step occurs in certain neurons of the brain and in the adrenal medulla (graphic and schematic representations of the biosynthesis and breakdown of catecholamines can be found in many of the references listed in Suggested Reading). In general, the enzymes described in this section are produced in the neuronal cell bodies and are then transported and stored in nerve endings. Therefore, the process of catecholamine biosynthesis takes place within these terminals. The catecholamines that are synthesized are then taken up and stored in vesicles (chromaffin granules) of the nerve terminals, which are located near the cell membrane. During neural transmission, catecholamines are released from these vesicles into the synaptic cleft. Although certain precursors of catecholamines (such as L-dopa) penetrate the blood–brain barrier, the catecholamines do not. Thus, all the catecholamines found in the brain are produced there.

The amount of catecholamines within the adrenal medulla and the sympathetic nervous system is generally constant; however, there are times when catecholamine levels in the body change dramatically. Initial changes that occur in the synthesis of these substances, in response to changes in demand, occur in minutes, whereas slower adaptational changes occur over much longer periods, even days in some cases. Catecholamines in the body are maintained at constant levels by a highly efficient process that modulates their biosynthesis, release, and subsequent inactivation. One example of a state with abnormal catecholamine levels is the condition known as pheochromocytoma, in which there is a tumor of the chromaffin cells of the medulla in the adrenal glands. It is characterized by hypersecretion of epinephrine, norepinephrine, dopamine, or dopa. In this condition urinary excretion of free catecholamines is also increased. The major clinical manifestations of this illness are high blood pressure, increased heart rate, sweating, rapid breathing, headaches, and the sensation of impending doom.

When an appropriate signal is received by a catecholaminergic neuron, it is transmitted down the axon to the presynaptic terminal, where it initiates the release of quanta of neurotransmitter into the synaptic cleft. The transmitter acts on receptors in postsynaptic neurons, resulting in the activation or inhibition of these cells.

C. Inactivation of Catecholamines

There are two major means for catecholamine inactivation: reuptake and enzymatic degradation. The reuptake system is fast and highly efficient. It operates through a rapid reuptake of released transmitters back into the presynaptic terminal. The involved transporter reuptake has two functions: (i) It rapidly inactivates transmission by removing transmitter from the synapse, and (ii) it conserves transmitter by restoring that which is not used in signal transmission. Catecholamines made in the neuron but not stored in the terminal vesicles are catabolized by a series of isoenzymes known as monoamine oxidases (MAO), which are located in most living tissues. There are two subtypes of MAO: MAO-A and MAO-B. In humans, MAO-B is four times more prevalent in the brain. The amount of MAO seems to increase with age. Another enzyme important in the breakdown of catecholamines released into the synapse is catechol-*O*-methyltransferase. Discussion of all the metabolic steps in degradation of catecholamines is beyond the scope of this article; however, it is important to note that drugs that increase catecholamine levels in the synapse, particularly the level of norepinephrine, are successful antidepressant medications.

The concentration of norepinephrine can be altered by two types of drugs: reuptake blockers, which prolong the life of norepinephrine in the synapse by preventing its reentry into the presynaptic neuron, and monoamine oxidase inhibitors, which interfere with breakdown by monoamine oxidase. From observations of the action of these drugs it has been proposed that depression is the result of low levels of norepinephrine in the brain; however, there is no direct evidence for this proposal. In support of the proposal, the antihypertensive drug reserpine, which depletes

norepinephrine and the other catecholamines, sometimes causes serious depression. Curiously, drugs that increase levels of serotonin, a noncatecholamine found in the brain, are also antidepressants. Norepinephrine and epinephrine also act as hormones when released from the adrenal medulla. Epinephrine is the principal catecholaminergic hormone produced in the medulla. Norepinephrine is the primary neurotransmitter in all postganglionic sympathetic neurons except for those that supply the vasodilator blood vessels of the skeletal muscular system and the sweat glands. The sympathetic nervous system and the parasympathetic nervous system make up the autonomic nervous system, which helps regulate the visceral functions of the body. The autonomic nervous system has control centers that are located in the spinal cord, hypothalamus, the reticular formation of the medulla oblongata, and other regions of the brain system. The centers located in the spinal cord and in the brain stem are regulated by the hypothalamus, which also communicates with the pituitary and cerebral cortex. This interconnection enables the complex orchestration of multiple somatic, visceral, and endocrinological functions.

The noradrenergic system has two major areas of origin in the brain: the locus coeruleus and the lateral tegmental nucleus. The projections of this system extend to all regions of the brain. As explained earlier, dopamine is the precursor in the synthesis of norepinephrine and epinephrine. In addition, dopamine has its own complex system and specialized function.

The dopamine system is composed of three subdivisions: mesocortical, mesolimbic, and nigrostriatal systems. The mesocortical system extends from the ventral tegmentum to a variety of areas, such as the olfactory tubercles, the accumbens, and the prefrontal cortex. The neurons of the mesolimbic system originate in the substantia nigra and the ventral tegmentum and project to the accumbens and amygdala. It is believed that the limbic system is probably more involved in regulating certain mental processes. The nigrostriatal system extends from the substantia nigra to the neostriatial regions. In addition to other functions, the nigrostriatal system is involved in movement. Disturbances of vital structures in this area are related to illnesses such as Parkinsonism.

D. Catecholaminergic Receptor Sites

Catecholamine receptors are proteins embedded in the plasma membrane of a neuron. Activation of these receptors by catecholamines can produce excitatory and/or inhibitory responses. In many cases, receptor number is increased or decreased as an adaptive response. For example, blockage of dopamine receptors by antipsychotic drugs, which are dopamine receptor antagonists, often results in a compensatory increase in the number of receptors. Many types of catecholamine receptors respond to one of the three catecholamines.

E. Dopaminergic Receptors

Five types of dopamine receptors have been identified. They are all called dopamine receptors because they all respond to dopamine and are relatively homologous in structure; however, two types, D_1 and D_2, can be discriminated pharmacologically by both agonists and antagonists. It is very likely that drugs selective for the other three types will also be found. The ability to selectively activate or inactivate different aspects of the dopaminergic system with drugs that act on one receptor type has made it possible to explore the role that the D_1 and D_2 dopaminergic system plays in behavior. The role of the other three types (D_3–D_5) is not clear.

1. D_1 receptors are found in the caudate nucleus and cortex. There are a variety of extraneural sites where these receptors are located, including the vascular structures of the brain, heart, and renal and mesenteric systems.
2. D_2 receptors have been identified in the putamen, caudate nucleus, and striatum as well as in limbic structures and in low density in the cortex. Two subtypes of D_2 receptors have been identified (D_{2a} and D_{2b}), but differences in anatomical location and physiological properties have not been worked out.
3. D_3 receptors have been identified in the limbic system.
4. D_4 receptors have recently been identified in the frontal cortex, basal ganglia, medulla, midbrain, and amygdala. The D_4 receptor demonstrates the greatest similarity to the D_2 family of receptors.
5. D_5 receptors have been identified in the caudate, putamen, olfactory bulb, and tubercle as well as in the nucleus.

F. Adrenergic Receptors

There are two types of adrenergic receptors, with subdivisions within each:

1. α-Adrenergic receptors α_1-Adrenergic receptors are located on postsynaptic effector cells such as those on the smooth muscles of the vascular, genitourinary, intestinal, and cardiac systems. Additionally, in humans these receptors are located within the liver. α_2-Adrenergic receptors inhibit the release of certain neurotransmitters. For example, at the presynaptic level in certain adrenergic nerve cells, these receptors inhibit norepinephrine release, whereas in cholinergic neurons they are responsible for inhibiting acetylcholine release. α_2-Adrenergic receptors are also located in postjunctional sites such as the β cells of the pancreas, in platelets, and in vascular smooth muscle. Although there are at least two subtypes of both α_1- and α_2-adrenergic receptors, the details concerning the actions and localization that would differentiate these particular subtypes have not been worked out.

2. β-Adrenergic receptors β_1-Adrenergic receptors have been located in the heart, the juxtaglomerular cells of the kidney, and the parathyroid gland. β_2-Adrenergic receptors have been identified in the smooth muscles of the vascular, gastrointestinal, genitourinary, and bronchial structures. Additionally, β_2-adrenergic receptors have been located in skeletal muscle and the liver as well as on the α cells of the pancreas, which are responsible for glucagon production. β_3-Adrenergic receptors are reported to be located in adipose tissue.

G. Plasma Catecholamines

The three catecholamines, when found intact in plasma, do not come from the brain because they cannot cross the blood–brain barrier; however, their metabolites can. Thus, metabolites in plasma originate both in brain and in peripheral tissues. Study of these metabolites has provided insight into the role that catecholamines play in behavior. However, direct study of catecholamines in living human brain tissue has not been possible. Fortunately, imaging technologies such as positron emission tomography (PET), magnetic resonance imaging, and single positron emission computerized tomography open up possibilities for visualizing catecholaminergic function in live conscious human subjects during waking hours. In addition, some of these imaging techniques can quantify neurotransmitter receptors and identify concentration differences at the synaptic level following psychopharmacologic administration. A variety of methods are available for measuring catecholamines in plasma.

II. IMPACT OF CATECHOLAMINES ON BEHAVIOR

Much of the information that is available concerning the functions of catecholamines in regulating human behavior directly results from the use of a group of medications often called psychotropic drugs and antidepressant medications called thymoleptics. Other medications that impact on catecholamines include psychostimulants, such as dextroamphetamine and methylphenidate (commonly known by its trade name, Ritalin) and L-dopa (which has been used to treat Parkinsonism), as well as a medication that was initially used to treat hypertension. Most of these drugs affect more than one system (e.g., dopaminergic, noradrenergic, or serotonergic systems). Catecholamines have been proposed as mediators of many psychiatric illnesses, including schizophrenia, Tourette's syndrome, depression, autism, attention deficit–hyperactivity disorder, stereotypic movements, tremors, and substance abuse. More generically, catecholamines also play a critical role in the stress response. Unfortunately, there is no definitive evidence for their role in the etiology of these illnesses. What is definite, however, is the role that catecholamines play in mediating the action of mood-altering and mind-altering drugs. There is speculation about which dopamine receptors these agents block to produce improvement, but the prevailing view is that the more traditional agents block D_2 receptors, whereas the newer atypical agents (such as clozapine) may also block D_4 (as well as D_1–D_3) receptors. These newer agents may also differ from the older agents in their ability to bind to serotonergic receptors of the 2A type. Nonetheless, as mentioned previously, the similarity between the D_2 and D_4 receptor families in terms of their structure and function lends insight into the pharmacologic efficacy of both newer and the more traditional antipsychotic agents on illnesses such as schizophrenia. The fact that agents that block dopamine receptors produce improvement in schizophrenia has led to the proposal that schizophrenia is caused by overactivity of the dopaminergic system. In support of this so-called dopamine hypothesis, one group has reported an increased density of D_2 receptors in brains of schizophrenic patients using PET. Increased density of D_2 receptors in postmortem brain tissue from patients with schizophrenia has also been reported. However, most patients have received neuroleptic treatment, which can cause these changes. Thus, it is not clear whether this increased density is an effect of the pathophysiology or the result of treatment. It is well established that reducing dopaminergic activity

with neuroleptics inhibits hallucinatory activity and normalizes delusional or paranoid thinking. It seems probable that the dopaminergic systems, particularly the D_2 and D_4 systems, have a physiological role in keeping thinking and level of suspiciousness in bounds. Curiously, patients who respond well to antipsychotic medication have a decrease in plasma homovanillic acid (HVA), the principal metabolite of dopamine, during treatment, whereas nonresponders do not. What is odd about this finding is that most plasma HVA does not come from the central nervous system.

Antipsychotics improve other symptoms associated with schizophrenia, such as impaired thought processes and attentional problems. Thus, it seems that the dopaminergic system may also regulate associative processing and attention. Drugs that improve psychotic symptoms have another important effect: They produce emotional blunting or so-called "flat affect." Inasmuch as these drugs reduce dopaminergic activity, it seems that dopamine may play a role in affect regulation.

Another illness that may illuminate the role of dopamine in regulation of behavior is Tourette's syndrome. This is an illness with onset usually between the ages of 4 and 8 years of age; however, it can occur at any time. It is characterized by rapid, repetitive movements known as motor tics that can be as simple as eye blinking or as complex as assuming contorted body positions. In addition to these movements, vocal tics occur ranging from repetitive coughing and throat clearing to shouting obscene words (coprolalia). These utterances can be a great source of embarrassment to the affected individuals. Both the vocalizations and the motor tics respond to antipsychotic drugs that are, of course, dopamine receptor blockers. This effect on Tourette's symptoms occurs even though the patients are not psychotic. Although dopamine is known to play a role in integrating motor movements, there is a distinct possibility that it may also inhibit socially undesirable movements and vocalizations.

It seems that Tourette's syndrome is in some way related to obsessive–compulsive disorder (this latter illness being particularly prevalent in families of Tourette's patients). Obsessive–compulsive disorder is often responsive to drugs that increase serotonergic activity. However, in one study, two dopamine agonists (methylphenidate and dextroamphetamine) were significantly more effective than placebo in reducing the symptomatology of obsessive–compulsive disorder. Thus, there appears to be a complex interaction between the serotonergic and dopaminergic systems in the regulation of psychomotor activity.

The study of psychological depression and its treatment can also help to illuminate the role of catecholamines in the regulation of behavior. Drugs such as the tricyclic antidepressants and the monoamine oxidase inhibitors, both of which increase norepinephrine in the synapse, are useful in treating depressed patients. As a result of these observations, it was first concluded that depression resulted from abnormally low activity of the noradrenergic system. It now appears, however, that increasing norepinephrine levels via drug treatment serves to compensate for unknown pathology in depression, affecting other transmitters besides norepinephrine.

These observations are nevertheless informative. It seems probable that norepinephrine, by regulating its own activity, and in concert with other transmitters, plays a role in the relief and prevention of depression if not in the cause of depression. Norepinephrine may regulate mood, level of emotional arousal, sleep/wakefulness states, and appetite (all of which are often disturbed in depression). However, as mentioned previously, it should be kept in mind that with the advent and demonstrated efficacy of the selective serotonin reuptake inhibitors in the treatment of major depression, the implications of abnormalities within the serotonergic system cannot be ignored.

Autism is a serious psychiatric condition that begins in infancy or early childhood. It is characterized by a qualitative impairment in interaction and socialization. Autistic children appear to be oblivious to their surroundings but ironically can react with a temper tantrum if a single toy is moved from its usual location. They often lack both verbal and nonverbal communication skills. Speech may be limited to repeating a word over and over (perseverations), and they may not even point to something they want in order to obtain it. Autistic individuals exhibit a restriction of activities and engage in a variety of odd behaviors, such as sniffing, twirling and spinning, and inordinate interest in the single function of an object (i.e., staring at a wheel spinning on a toy car for hours). They also sometimes present with violent or self-injurious behavior and temper tantrums. Some patients may possess striking talents beyond their apparent cognitive capacity (often referred to as savant-like traits). A few can masterfully play the piano without ever receiving instructions or memorize an entire city's bus routes.

The pervasive developmental disorders are illnesses that may vary in presentation. They may present with only one feature of autism or most of the features (but by definition not all). Although elevated serotonin levels in whole blood seem to be the most consistent

finding in autism, there have been reports of increased norepinephrine levels in the plasma of these children when compared to normal control groups. Additionally, the effectiveness of dopamine-blocking neuroleptics on attention and improvement of certain behaviors in autistic children cannot be ignored. One investigation of biological markers in children with pervasive developmental disorder reported that the group that responded to treatment had lower initial plasma levels of HVA.

Attention deficit–hyperactivity disorder (ADHD) is characterized by overactivity, fidgetiness, impulsivity, and distractibility. It is more frequently seen in males and there is usually a family history of the disorder. The illness begins in early life but often is not diagnosed until the child is in school because its pathology becomes more evident when more controlled behavior is required. There is strong evidence for the involvement of the catecholaminergic systems in this illness. Prevailing theories propose a decrease in turnover of both dopamine- and norepinephrine-like effects. Oddly, increases in noradrenergic activity in the activating systems of the brain produce emotional arousal and many of the attentional abnormalities of ADHD. In addition, normal adults given the psychostimulants used to treat ADHD in children show the expected activating effects. Interestingly, in the rat, many experimental manipulations have also implicated the dopaminergic system's role in locomotor activity. The quantification of motoric overactivity was directly correlated to two observations within the dopaminergic system. The first was the degree of dopaminergic damage sustained, and the second was onset of D_1 receptor supersensitivity.

Despite the fact that much of our knowledge regarding the impact of catecholamines on behavior has come from observations related to medication trials, recent advances in genetic techniques also further our hopes for future breakthroughs. Certain neurological illnesses, such as hereditary progressive dystonia and supranuclear palsies, have been linked to abnormalities in tyrosine hydroxylase gene expression. Pioneering efforts to investigate the genetic components of other neuropsychiatric conditions are under way. Two such illnesses may be schizophrenia and bipolar disorder. In these illnesses, classical modes of inheritance are not evident. However, there is clinical evidence via adoption, twin, and family studies that implicates a hereditary component to these illnesses. Being able to approach this problem from multiple perspectives employing the evolving fields of molecular biology and genetics may eventually result in the identification of the precise genetic deficits and locations of these complicated illnesses. Currently, several research groups are exploring catecholaminergic-related markers in these fields.

The role of catecholamines in substance abuse has received substantial attention. Most drugs of abuse behaviorally work on the premise of being strong positive reinforcers. These include agents such as nicotine, morphine, cocaine, and the amphetamine group. Increased dopamine levels in mesocortical regions have been identified with the use of amphetamines, cocaine, and even nicotine. In the case of cocaine and the stimulants, the increase is produced by blocking the dopamine transporter system. As discussed previously, dopamine has also been implicated in the production of hallucinations. Many of the agents that increase dopamine levels, such as L-dopa, amphetamines, and cocaine, are also capable of inducing hallucinations.

In the realm of stress, many different corporal systems interact to help the human organism respond. There is usually activation of the adrenergic neurons in the hypothalamus. There is a close link between the cortisol and adrenergic response. In fact, increases in norepinephrine induce cortisol-releasing hormone production. Stimulation of the adrenergic nervous system causes elevations in plasma epinephrine and norepinephrine levels. These produce an increase in glucose. Furthermore, epinephrine induces the breakdown of fats from adipose tissues. In these situations, epinephrine and norepinephrine promote increased heart rate and blood pressure. The purpose of these changes is to facilitate the delivery of needed material for fight or flight. Situations often perceived as stressful to the body include strenuous physical exertion, acute anxiety states, and serious cardiovascular compromise.

III. CONCLUSIONS

Catecholamines in the brain act at the highest levels of mental function. Although their role in specific mental disorders is not entirely clear, there is little doubt that they modulate, if not mediate, functions such as processing of associations, integration of thought processes with movement and speech, emotional tone or affect, mood, appetite, arousal, and sleep/wakefulness state. Most of these functions have not been successfully modeled in nonhuman species, leaving their study to be carried out in living humans. This

limitation has made more than inferential conclusions regarding behavioral and mental function impossible.

New technological advances in functional brain imaging and in studies of gene expression in accessible human cells have opened new windows into the brain, but definitive studies await further advances.

See Also the Following Articles

AUTISM • BEHAVIORAL PHARMACOLOGY • DEPRESSION • HOMEOSTATIC MECHANISMS • LIMBIC SYSTEM • NEUROTRANSMITTERS • SCHIZOPHRENIA • STRESS • TOURETTE SYNDROME AND OBSESSIVE COMPULSIVE DISORDER

Suggested Reading

Anderson, G. H., and Johnston, J. L. (1983). Nutrient control of brain neurotransmitter synthesis and function. *Can. J. Physiol. Pharmacol.* **61,** 271–283.

Axelrod, J. (1987). Catecholamines. In *Encyclopedia of Neuroscience*. (G. Adelman, Ed.), Vol. 1. Birkhauser, Boston.

Breier, A., Adler, C. M., Weisenfeld, N., Su, T.-P., Elman, I., Picken, L., Malhotra, A. K., and Pickar, D. (1998). Effects of NMDA antagonism on striatal dopamine release in healthy subjects: Application of a novel PET approach. *Synapse* **29,** 142–147.

Davis, K. L., Khan, R. S., Ko, G., and Davidson, M. (1991). Dopamine in schizophrenia: A review and reconceptualization. *Am. J. Psychiatr.* **148**(11), 1474–1486.

Friedhoff, A. J. (Ed.) (1975). *Catecholamines and Behavior*, Vols. 1 and 2. Plenum, New York.

Friedhoff, A. J., and Silva, R. (1997). Catecholamines and behavior. In *Encyclopedia of Human Biology*. (M. Yelles, Ed.), Vol. 2. Academic Press, San Diego.

Goldstein, D. S., Eisenhofer, G., and McCarty, R. (1998). Catecholamines bridging basic science. In *Advances in Pharmocology*, Volume 42. Academic Press, San Diego.

Hardman, J. G., and Limbird, L. E. (Eds.) (1996). *Goodman and Gilman's: The Pharmacological Basis of Therapeutics*, 9th ed. McGraw-Hill, New York.

Joffe, R. T., Swinson, R. P., and Levitt, A. J. (1991). Acute psychostimulant challenge in primary obsessive–compulsive disorder. *J. Clin. Psychopharmacol.* **11,** 237–241.

Kaplan, H. I., and Sadock, B. J. (Eds.) (1995). *Comprehensive Textbook of Psychiatry*, 6th ed. Williams & Wilkins, Baltimore.

Silva, R. R., and Friedhoff, A. J. (1993). Recent advances in research into Tourette's syndrome. In *Handbook of Tourette Syndrome and Related Tic and Behavioral Disorders*. (R. Kurlan, Ed.). Dekker, New York.

Wilson, J. D., Baaunwald, E., Isselbacher, K. J., Petersdore, R. G., Martin, J. B., Fauchi, A. S., and Rood, R. K. (Eds.) (1991). *Principles of Internal Medicine*. McGraw-Hill, New York.

Categorization

BARBARA J. KNOWLTON
University of California, Los Angeles

GLOSSARY

category-specific agnosia A neuropsychological deficit in the ability to identify members of a specific natural category accompanied by a near normal ability to identify objects in other categories.

exemplars Items that qualify as category members, especially items that were seen during category learning.

explicit knowledge (declarative knowledge) Learned information that is available to awareness. Individuals are consciously aware that they learned this information, and it can be deliberately applied in order to make judgments.

implicit knowledge (nondeclarative knowledge) Learned information that is not available to awareness. Individuals show that they have learned this information through their performance, but they are not able to describe the information that is guiding their performance.

prototype The most typical example of a category. It is generally considered to be an arithmetic average of all exemplars of the category or an item containing the most frequently occurring features in the category.

Category learning in its broadest sense refers to all learning in which a response is generalized to multiple stimuli. Although all organisms exhibit category learning to some extent, humans in particular have a remarkable ability to learn to classify items based on extremely complex criteria. For example, a wine connoisseur is able to readily classify novel wines by grape, and an art historian can swiftly and accurately judge the period of a newly encountered painting. Human category learning also includes more mundane abilities, such as determining that a papaya is a fruit the first time one is tasted or being able to use nonverbal cues to determine a friend's mood.

I. INTRODUCTION

Given the ubiquity of category learning in human experience, it is likely that there are different types of category learning that depend on different neural systems. A major focus of research in cognitive psychology during the past several decades has been the development of category learning models that can broadly account for the behavior of human participants in category learning experiments in the laboratory. An approach that has recently gained favor is to acknowledge that different types of category learning exist and to more clearly delineate them based on their behavioral and neuropsychological characteristics.

II. RULE-BASED VS EXEMPLAR-BASED CATEGORIZATION

Many common categories can be objectively defined by a set of rules that define membership in the category. For example, a set of criteria must be met to classify an animal as a mammal (e.g., it must be warm-blooded). Given these criteria, an individual should be able to accurately classify new examples. For other categories,

the set of rules that define membership are less clear. Although many people may agree that a particular concerto sounds like a piece by Mozart, it is difficult to create a list of necessary and sufficient conditions that define the category of Mozart concertos other than the fact that they were composed by Mozart. In this case, one gets the sense that classification is occurring based on previous experience with different exemplars of the category. An exemplar-based strategy implies that an individual is classifying new items based on their similarity to previously experienced items in that category.

The fact that some categories can be readily defined by rules and others cannot does not necessarily indicate that two different category learning systems exist. Perhaps people are always using rules to classify, although sometimes those rules are difficult to verbalize and may be imperfect. For example, with enough insight, it seems possible to create a fairly valid set of explicit heuristics that would discriminate Mozart concertos from works by other composers. On the other hand, the existence of explicit rules that define category membership does not automatically mean that people are actually using those rules. People may judge that an animal is a mammal because of its similarity to other mammals rather than by checking whether a list of conditions is satisfied. The fact that people classify a cow as a mammal more readily than a whale demonstrates that the satisfaction of membership conditions is not the only factor influencing classification. Both rule-based and exemplar-based strategies are plausible, but it appears that people are biased to use one strategy over the other depending on the category type and the conditions of learning. Early in training, individuals might use rules to classify items if they are available, or they might try to induce rules if none are given. However, later in training, when they have had extensive experience with examples, they may base their category decisions on their previous experience with these examples. Figure 1 shows examples from two categories that can be classified according to a set of rules or according to similarity to other members of the category.

Neuropsychological data suggest that classification according to rules involves prefrontal cortex. One of the classic tests of prefrontal cortical dysfunction is the Wisconsin Card Sorting Test, in which the subject must induce a sorting rule (categorize by color, shape, etc.). Patients with prefrontal damage are particularly impaired when they must shift from one category to another. Neuroimaging studies have shown dorsolateral prefrontal activity in humans categorizing stimuli according to rules. Also, the application of rules requires working memory, an ability that has been strongly associated with prefrontal (especially dorsolateral) function. In order to determine whether an item satisfies a list of conditions, one must maintain these conditions in a short-term buffer and actively test the item against each condition. Putting demands on working memory by requiring subjects to perform a concurrent task has been shown to induce subjects to shift to an exemplar-based strategy from a rule-based strategy. The fact that young children do not use a rule application strategy as readily as adults do is also supportive of the idea that prefrontal cortex is important for rule-based classification. The frontal lobes mature relatively late in development, continuing into adolescence.

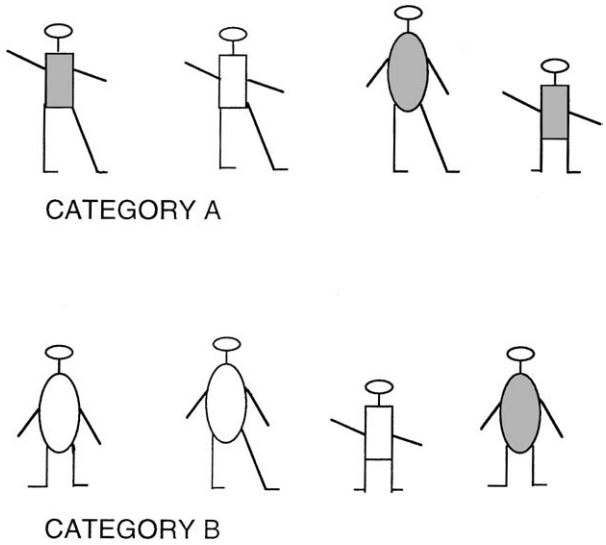


Figure 1 Two categories of stick figures. The figures in category A have at least two of the following three attributes: square body, shaded body, and long legs. The figures in category B have at least two of the following attributes: round body, white body, and short legs. Subjects may be inducing these rules during training, or they may be basing categorization judgments on similarity to previously experienced items in the two categories.

Classification rules tend to focus on individual attributes of stimuli. In addition to working memory, the application of rules requires the ability to selectively attend to attributes of the stimuli and to shift attention between different attributes. Neuroimaging findings are consistent with this idea in that posterior parietal brain regions are active during rule application. These regions are typically active in selective attention tasks and are damaged in patients exhibiting the neglect syndrome, especially when the lesion is on the right side.

III. EXEMPLAR-BASED CATEGORIZATION: GENERAL OR SPECIFIC?

Although people often make classification judgments on their previous experience with examples of the category, the type of exemplar-based information could take different forms. After experiencing many examples of a category, such as chairs, we may have formed an idea of what a typical chair looks like, and we may compare new examples to this typical chair and classify these new items as chairs depending on their similarity to this typical chair. This typical example, or category prototype, represents the average of experienced category exemplars. When additional exemplars are encountered, the representation of the prototype is modified to reflect the new experience. By this view, category knowledge takes the form of abstracted information that is distinct from memories of the individual exemplars. In support of this prototype view is the fact that category prototypes or average members of categories are much easier to classify than members that are very dissimilar from the prototype. People are much faster to acknowledge that a robin is a bird rather than a penguin or chicken. In fact, in artificial categories generated in the laboratory, speed and accuracy of classification can be predicted quite well by the similarity of an item to the prototype (Fig. 2).

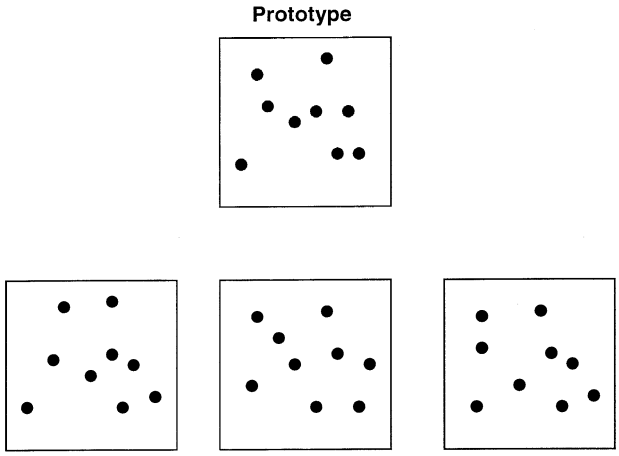


Figure 2 Illustration of a category based on a prototype. The top dot pattern serves as the prototype, and the patterns below are all distortions created by moving each dot in the prototype in a random direction. After studying a series of such distortions, subjects are very likely to classify the prototype as being part of the category.

However, such effects do not necessarily mean that people are using prototypes to make classification judgments. According to another view, people base their classification judgments on the average similarity to previously encountered items stored in memory. This view is appealing because of its parsimony in that one does not need to hypothesize a separate base of category-level knowledge. This view can also account for the superior classification of prototypes because the prototype is generally the item that is similar to the largest number of category members since it is the average of all experienced examples.

Based on experimental data, it is difficult to distinguish between the possibility that people use information general to the category to classify items or that they make comparisons to specific examples stored in memory. These two views often make similar predictions about which items are easiest and most difficult to classify. One approach to addressing this question has been to test patients with amnesia on category learning tasks. Because these patients have sustained damage to medial temporal lobe or midline diencephalic regions, they exhibit a profound deficit in declarative memory. Declarative memory refers to the conscious memory of facts and events. Declarative memory would be accessed on a recognition test of previously seen examples of a category. Because of their declarative memory deficits, amnesic patients should not be able to learn categories normally if this ability depends on declarative memory for examples. If, however, amnesic patients are able to learn to classify items based on category membership despite poor memory for the individual training examples, it would suggest that category-level information could be learned independently of exemplar memory.

In fact, there is evidence that amnesic patients can learn at least some types of categories normally. For example, amnesic patients have been shown to be able to perform normally on an artificial grammar learning task. In this task, subjects view a series of letter strings that were formed according to a finite-state rule system (Fig. 3). This rule system allows only certain letters to follow other letters. After viewing the strings, subjects are told for the first time that the items they had just seen all followed a set of rules and that their task is to classify a new set of items as following these rules or not. Although subjects typically believe that they did not learn anything about the grammatical rules, they nevertheless can reliably classify new items at a level that is significantly higher than chance. Amnesic patients perform as well as normal subjects on this task, although they are severely impaired in recognizing the letter strings that were used during training. Thus, it appears that subjects acquire information

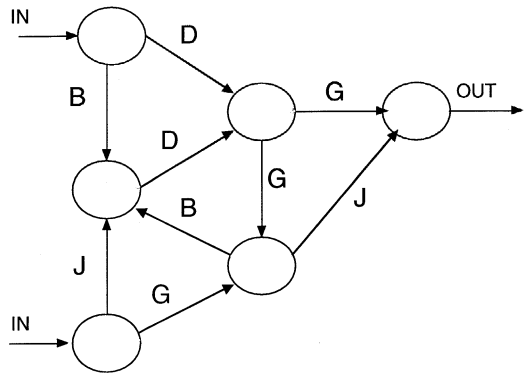


Figure 3 A rule system used to generate an artificial grammar. Letter strings are generated by traversing the diagram along the arrows from the "in" arrows to the "out" arrow. An additional letter is added as each arrow is traversed. For example, DGJ, BDG, JDGBDG, and GBDG are examples of letter strings formed according to the artificial grammar. In contrast, DDJ, BGDJ, and JDGB cannot be formed according to the rule system.

about the grammatical category that is independent of memory for individual exemplars.

The findings from amnesic patients suggest that information about categories can be acquired that is distinct from explicit memory for exemplars. However, these results do not preclude specific exemplar-based classification in other paradigms. It is likely that when subjects are given a small number of highly distinctive and memorizable exemplars, they will consider these in making classification judgments.

IV. EXPLICIT VS IMPLICIT KNOWLEDGE IN CATEGORIZATION

The issue of whether category learning is based on explicit or implicit knowledge has often been conflated with the question of the form of the information acquired. For example, the demonstration that performance on a certain task was mediated by fairly concrete properties of exemplars would often imply that this information was learned explicitly since this information would be apparent to subjects during training. Conversely, if it appeared that subjects were learning complex rules about the category, then it was assumed that this knowledge was implicit since subjects often cannot readily verbalize these rules. However, the issues of rule vs exemplar learning and abstract vs specific exemplar learning are both orthogonal to the question of whether this information is explicitly or implicitly learned. Just as subjects may be fully aware of applying explicit rules to a categorization task, they may also be consciously aware of memories of examples that are being used for comparison to new items. Likewise, category rules or exemplar information could be learned implicitly. Exemplar-derived information such as a prototype seems likely to be learned implicitly since people are not usually aware that they are forming the prototype during the training, and this knowledge is generally expressed only through classification performance. Item-specific information may also be learned implicitly. For example, priming is a form of implicit learning that is specific to particular stimuli and similar mechanisms could contribute to category learning in some circumstances.

Studies of artificial grammar learning illustrate the point that it is important to define what exactly is being learned in order to determine if learning is implicit or explicit in a particular category learning task. In the artificial grammar learning paradigm, several different types of learned knowledge may contribute to classification judgments. Although the category is operationally defined by adherence to the finite-state rule system, it appears that subjects base their judgments to a great extent on whether the test item contains bigrams and trigrams that occurred with high frequency in the training set. Thus, the fact that subjects cannot explicit state the grammatical rules does not necessarily mean that the knowledge they are using to make category judgments is implicit. They may be perfectly aware of which letter bigrams and trigrams were frequent in the training set and are using this information to make their judgments. It is therefore crucial to first understand what learned information subjects use for category judgments in order to attempt to test whether this information is implicitly learned or if subjects are explicitly aware of this information.

V. NEUROPSYCHOLOGICAL STUDIES OF CATEGORY LEARNING

One approach that has successfully demonstrated implicit learning of categories is the use of data from amnesic patients. The fact that amnesic patients are able to exhibit normal category learning in several paradigms suggests that implicit learning can support normal performance. In addition to their intact performance on artificial grammar tasks described previously, amnesic patients have been shown to perform normally on category learning based on similarity to a prototype, category learning based on feature frequency, and category learning based on probabilistic associations. In each case, amnesic

patients are severely impaired at recognizing training exemplars. Importantly, there are circumstances in which amnesic patients do not perform as well as normal subjects, particularly when there are a relatively small number of exemplars used during training that are highly memorizable. In those tasks in which amnesic patients are able to perform normally, it is likely that subjects are not basing their judgments on explicit comparisons to items stored in memory because the exemplars are not easy to memorize (letter strings and dot patterns) or a large number of exemplars are presented during training. These data further indicate that categorization can be accomplished using different kinds of knowledge depending on the task. It can be concluded that category learning does not necessarily depend on those medial temporal lobe structures that are required for explicit memory of exemplars.

If category learning can proceed independently of the brain system that supports explicit memory, it shows that other neural systems are capable of supporting this ability. Because there are multiple types of learned information that are used in different categorization tasks, it is highly likely that there is no single neural system involved in all category learning or even involved in all implicit category learning. Rather, different types of category learning are likely to depend on different brain systems. One system that has been linked with implicit learning is the basal ganglia. Damage to the basal ganglia as a result of Huntington's disease or Parkinson's disease impairs motor skill learning. Recent evidence suggests that the basal ganglia may be involved in nonmotor forms of implicit learning as well. These patients are impaired on classification tasks in which category membership is based on a nonverbal rule or a probabilistic association that is difficult to verbalize. For example, in the probabilistic classification task, a set of arbitrary cues (such as cards with geometric shapes) are probabilistically associated with one of two outcomes (such as rain or shine if a weather prediction cover story is used). Because of the probabilistic nature of the associations, memory for single trials is not very helpful for performing the task. Rather, information gleaned over multiple trials must be used to classify each instance. Patients with basal ganglia disorders exhibit severe impairments in this task. According to one model of implicit category learning, the connections from the striatum to the prefrontal cortex via the globus pallidus and thalamus are the basis for the associations between perceptual inputs and responses that are the basis for nonverbal rule learning. It may be that the striatum actually represents the rules, or that the rules just emerge as a property of these perception–action associations. These nonverbal rules are represented as associations between cues and responses (''If triangles appear, press sun key''). In contrast, for verbal rules that are learned declaratively, there is an explicit association between stimuli and outcomes (''If triangles appear, the weather will be sunny, so I should press the sun key'').

Another potential difference between basal ganglia-dependent category learning and learning that depends on the medial temporal lobe is the fact that the neostriatum may be the actual storage site for learned associations. In contrast, the role of the medial temporal lobe in explicit memory is time limited. Patients with medial temporal lobe damage are able to retain memories from long before their brain injury, demonstrating that the eventual storage sites of these memories are not in the medial temporal lobes. In models of neostriatal function, memories are stored in the connections within the basal ganglia or in the connections within the basal ganglia and thalamus. These models predict that damage to the basal ganglia would disrupt not only new nonverbal category learning but also previously learned nonverbal rules. These patients would not necessarily be impaired at categorizing stimuli based on previously learned explicit rules based on semantic knowledge (sorting fruits and vegetables into separate categories). However, there is some indication that these patients have difficulty with some previously learned implicit rules pertaining to language and social cognition, two domains in which rule learning appears to proceed implicitly.

VI. NEUROIMAGING STUDIES OF CATEGORY LEARNING

Functional magnetic resonance imaging (fMRI) studies also support the idea that there are multiple forms of categorization that depend on different brain systems. Application of explicit rules appears to activate frontal and parietal regions involved in working memory and selective attention. When nonverbal rules appear to guide responses, as in the probabilistic classification task described previously, activation has been reported in the caudate nucleus. Interestingly, this neostriatal activity was accompanied by a decrease in activity in the medial temporal lobe, suggesting that the explicit and implicit learning systems may act in a competitive manner.

Neuroimaging data also support the idea that there are different neural substrates for implicit category learning based on rules vs implicit category learning based on exemplar-based information. In addition to nonverbal rule learning tasks, amnesic patients are able to classify new items based on their similarity to a prototype abstracted from the training exemplars. Both types of category learning can proceed independently of the medial temporal lobe memory system. However, they appear to depend on different systems. Patients with Parkinson's disease are able to learn to classify dot patterns based on a learned prototype despite their impaired performance on nonverbal rule learning. Rather than a neostriatal locus, neuroimaging evidence suggests that learning about these naturalistic categories may rely on regions involved in perceptual processing. After viewing dot pattern stimuli that are all distortions of a prototype, viewing the prototype dot pattern is accompanied by a decrease in blood flow in occipital regions. This blood flow decrease is similar to what is seen in perceptual priming paradigms. Items that had been previously presented trigger less blood flow in occipital regions than do new items. This blood flow decrease can be understood in terms of a decrease in the resources required to process primed stimuli vs new stimuli. Priming has often been considered to result because of increased ease, or fluency, of processing old stimuli relative to stimuli that have not been recently processed. This is thought to occur because of residual activity in neural representations activated by the initial presentation of the item. Categorization judgments may also be influenced by perceptual fluency. Presentation of a set of exemplars would result in residual activity in the neural representations of each individual exemplar. Because the prototype is the average of all the training exemplars, one might expect that the prototype might enjoy the greatest amount of enhanced processing from the summed residual activity of all the training exemplars. Under this scenario, the prototype is not represented separately from the training exemplars but, rather, emerges as a consequence of the overlapping neural representations of category exemplars.

This process may account for categories such as dot patterns that are defined merely by perceptual features. Can a similar perceptual fluency mechanism account for prototype effects seen in semantic categories? For example, it is a well-known finding that subjects are able to confirm that a robin is a bird compared to confirming that a penguin is a bird. It may be that as exemplars of a category are encountered, a link between the representation of the item and its category is formed. These exemplar representations would partially overlap given the degree to which features were shared. When primed with a category name, items near the prototype would benefit from the fact that they are similar to many individual exemplars that have been linked to the category. Thus, semantic priming would result in stronger semantic priming for even novel items near the prototype because of the wide partial activation induced by these items. Again, this phenomenon would occur as a consequence of overlapping semantic representations of individual exemplars.

VII. THE REPRESENTATION OF CATEGORIES IN THE BRAIN

Category learning is accomplished by several different brain systems that are specialized for the acquisition of different forms of information used to make category judgments. As discussed previously, nonverbal rules may be stored in corticostriatal loops. Exemplar-based categories, which form the basis of our semantic knowledge of the world, appear to be stored in the cortex of the temporal lobe. This is particularly true for categories we learn about through visual experience with the world (e.g., animals and plants) and thus may be thought of as the output of the ventral visual stream, or "what" pathway. Categories of actions may be represented in frontal regions associated with movement planning. Data from fMRI studies have revealed that categories such as houses, faces, and chairs activate distinct regions of posterior temporal cortex (Fig. 4). Perhaps categories that are defined more conceptually than visually (animals that live in the jungle and things you can buy at a supermarket) are represented more anteriorly in the temporal lobe. Patients with anterior and lateral temporal lobe damage exhibit a breakdown in semantic knowledge including deficits in sorting items into categories that had been learned previously. This phenomenon is seen most readily in patients with semantic dementia, but it also occurs with the diffuse temporal lobe damage seen in Alzheimer's disease.

The existence of patients with category-specific agnosias has been used as support for the idea that stored semantic knowledge in the brain is organized to some extent by category. Patients have been described with specific deficits in identifying members of such categories as faces, animals, tools, or foods as the result of brain damage. It may be that knowledge about these different categories is stored separately in distinct

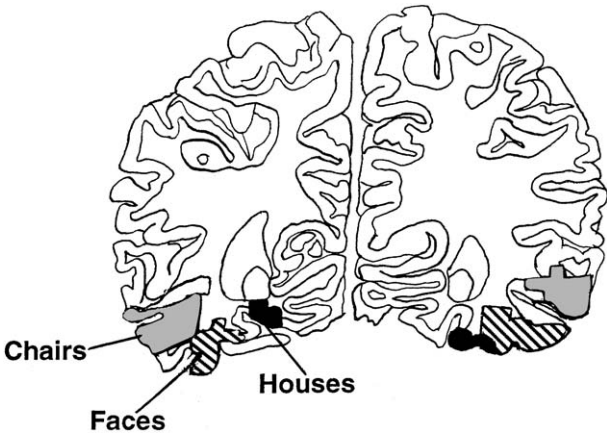


Figure 4 A coronal section showing regions in the ventral temporal lobe that are differentially activated by chairs, faces, and houses.

brain regions. In this view, damage to one region could disrupt knowledge of animals, for example, but not knowledge of other objects.

The view that there are category-specific regions of the brain has been challenged by the fact that different categories may have different processing requirements and thus brain specificity may reflect modules for different types of processing. For example, animals are composed of basically the same parts in the same configuration (eyes, legs, tail, etc.), and they are identified by the relative sizes and shapes of those parts. In contrast, tools generally contain parts that are present in some other tools but not others, and they are identified by the location of the parts relative to each other. A broom and a brush are differentiated by the relative position of the handle and bristles. Thus, brain regions associated with identifying animals may actually be important for processing the metric properties of the parts of items. ''Tool areas'' could actually be involved in processing the relationships between parts. It may be that knowledge about different category members is stored in a distributed fashion throughout the brain, but that specific cortical areas are involved in analyzing particular stimulus properties that are confounded with category membership.

VIII. CONCLUSION

Category learning can occur effortlessly, seeming to arise simply as a consequence of exposure to exemplars. In other cases, category learning results from explicit hypothesis testing to determine diagnostic rules. Because categories can be defined based on a wide range of information, it makes sense that multiple mechanisms would have evolved to allow humans to classify new instances based on previous experience. Being able to make predictions about new stimuli based on experience with members of the same category is highly adaptive. Thus, it is not surprising that multiple mechanisms evolved to enable this ability for a wide range of stimuli. Part of the challenge to cognitive neuroscientists is to precisely define the different forms of information acquired by subjects in category learning tasks in order to guide the search for neural substrates.

See Also the Following Articles

AGNOSIA • ARTIFICIAL INTELLIGENCE • INTELLIGENCE • LANGUAGE ACQUISITION • MEMORY, EXPLICIT AND IMPLICIT

Suggested Reading

Ashby, F. G., and Waldron, E. M. (1999). On the nature of implicit categorization. *Psychonomic Bull. Rev.* **6,** 363–378.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., and Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychol. Rev.* **105,** 442–481.

Caramazza, A. (2000). The organization of conceptual knowledge in the brain. In *The New Cognitive Neurosciences* (M. Gazzaniga, Ed.). MIT Press, Cambridge, MA.

Forde, E. M. E. (1999). Category-specific recognition impairments for living and nonliving things. In *Case Studies in the Neuropsychology of Vision* (G. Humphreys, Ed.). Psychology Press/Taylor & Francis, Hove, UK.

Knowlton, B. J. (1995). Category learning in amnesia. In *Emotion, Memory, and Behavior: Studies on Human and Nonhuman Primates* (T. Nakajima and T. Ono, Eds.). CRC Press, Boca Raton, FL.

Knowlton, B. J. (1999). What can neuropsychology tell us about category learning? *Trends Cognitive Sci.* **3,** 123–124.

Martin, A., Ungerleider, L. G., and Haxby, J. V. (2000). Category specificity and the brain: The sensory/motor model of semantic representations of objects. In *The New Cognitive Neurosciences* (M. Gazzaniga, Ed.). MIT Press, Cambridge, MA.

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In *Essays in Honor of William K. Estes* (A. F. Healy and S. M. Kosslyn, Eds.), pp. 149–167. Erlbaum, Hillsdale, NJ.

Nosofsky, R. M., and Zaki, S. (1999). Math modeling, neuropsychology, and category learning. *Trends Cognitive Sci.* **3,** 125–126.

Shanks, D. R., and St. John, M. F. (1994). Characteristics of dissociable human learning systems. *Behav. Brain Sci.* **17,** 367–447.

Smith, E. E., and Jonides, J. (2000). The cognitive neuroscience of categorization. In *The New Cognitive Neurosciences* (M. Gazzaniga, Ed.). MIT Press, Cambridge, MA.

Smith, E. E., Patalano, A. L., and Jonides, J. (1998). Alternative strategies of categorization. *Cognition* **65,** 167–196.

Cerebellum

MARCO MOLINARI
Rehabilitation Hospital and Research Institute, Santa Lucia Foundation, Rome

GLOSSARY

cerebellar cortex Formed by three cellular layers—the molecular, Purkinje, and granular layers.

cerebellar deep nuclei Three pairs of nuclei embedded in the cerebellar white matter in which the neurons projecting to extracerebellar structures are located. The vast majority of these neurons are excitatory.

climbing fiber Second largest component of fibers that reach the cerebellum. They are formed by axons of inferior olive cells and terminate on neurons located in the deep nuclei as well as on Purkinje cells.

dismetria Disturbance of the control of range direction and force of muscle contraction determining disruption of coordination of multijoint movements.

mossy fiber The largest group of fibers that reach the cerebellum. They terminate on neurons located in the deep nuclei as well as on granule cells.

Purkinje cell Only efferent cell of the cerebellar cortex. It has an inhibitory effect on the cerebellar deep nuclei.

The cerebellum is a portion of the brain that plays a key role in many aspects of human behavior. This article presents available data on the cellular and physiological organization of the cerebellar circuits, stressing from a multidisciplinary standpoint, the new functions that have been recently added to the traditional cerebellar repertoire of motor learning and motor control.

I. INTRODUCTION

The cerebellum (literally, ''little brain'') is located in the posterior cranial fossa. It represents 10% of the total brain volume and contains more than 50% of the total number of neurons of the central nervous system. Its general organization resembles that of the telencephalon with an outer mantle of gray matter, the cerebellar cortex, that covers an internal white matter in which the deep nuclei (i.e., the three pairs of deep cerebellar nuclei) are embedded. The cellular organization of the cerebellum is quite simple; its basic structure has been well-known since the beginning of the 20th century due to the work of Ramon y Cajal (Fig. 1). Because of the simplicity of the cerebellar anatomical organization and its remarkable similarity in all mammals, it seemed to be an easy field for investigating the relationships between brain structure and function; thus, it has always been a field of extreme interest for neuroscientists. Since the pioneering work at the end of the 19th century, the number of studies devoted to the cerebellum and its function has increased enormously and many theories on the cerebellar function have been proposed. Despite all these efforts, we still do not know exactly what the cerebellum is. In recent years, particularly due to unpredicted clinical and functional neuroimaging findings, the classical theories focused on the role of the cerebellum in motor control and motor learning have been reconsidered and new hypotheses have been advanced. In this article, after a general update on the available data regarding the anatomical and functional cerebellar organization, recent theories on the functions of the cerebellum are reviewed.

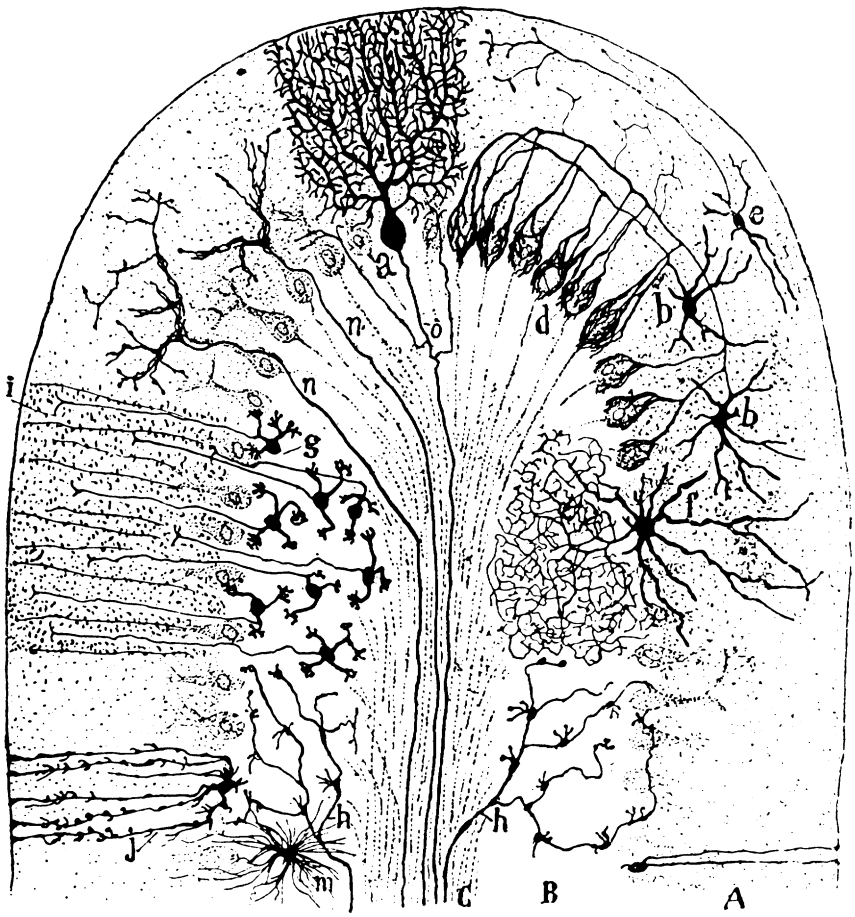


Figure 1 Schematic reconstruction of a section cut perpendicular to the main axis of a cerebellar lobule. Reconstruction of the organization of the cerebellar cortex as drawn by Ramon y Cajal from observations of Golgi preparations. A, molecular layer; B, granular layer; C, white matter; a, Purkinje cells; b, basket cells; d, terminal arborization of the axons of the basket cells; e, stellate cells; f, Golgi cells; g, granule cells; h, mossy fibers; n, climbing fibers [reproduced with permission from Ramon y Cajal (1972). *Histologie du Systeme Nerveux de L'Homme e des Vertebres, Consejo Superior de Investigagaciones Cientificas*. Instituto Ramon y Cajal, Madrid].

II. ANATOMY

When the occipital bone is opened, the dorsal surface of the cerebellum is clearly visible in the posterior cerebral fossa. The most prominent aspect of the cerebellar surface is the presence of many parallel convolutions that run perpendicular to its anteroposterior axis. In saggital sections, this chain of parallel convolutions has a characteristic appearance, like leaves stemming from the same trunk. This feature, defined by early anatomists as arbor vitae (''life tree''), is at the origin of the term folia (''leaves'') applied to the cerebellar convolutions. From medial to lateral, three major subdivisions are easily recognizable: a median one (the vermis) and two lateral parts (the hemispheres). Hemispheres and vermis are clearly separated only in the inferior surface where two longitudinal furrows are evident. In the dorsal surface, there is only a shallow groove between the vermis and the hemispheres and the folia run from one side to the other without obvious interruptions. A further subdivision of the hemispheres in an intermediate and a lateral part has been proposed on the basis of functional and connectional evidence, but it cannot be recognized on the basis of anatomical landmarks. Anatomical subdivision on the anteroposterior axis is mainly based on the features of the cerebellar foliation. Two deep transverse fissures divide the cerebellum into three major lobes. The primary fissure, clearly visible from the dorsal surface, separates the anterior lobe

from the posterior one. On the underside of the cerebellum, the posterolateral fissure separates the floccular nodular lobule from the posterior lobe. The appearance of the human cerebellum on nuclear magnetic resonance (NMR) scans is depicted in Fig. 2. Further subdivisions have been proposed based on smaller fissures or on the difference in the orientation of the folial chains (Fig. 3). A subdivision often applied is that proposed by Larsell in 1952. He divided the vermis into 10 lobules labeled with Roman numerals from I to X and also applied this division to the corresponding lobules of the hemisphere, adding the

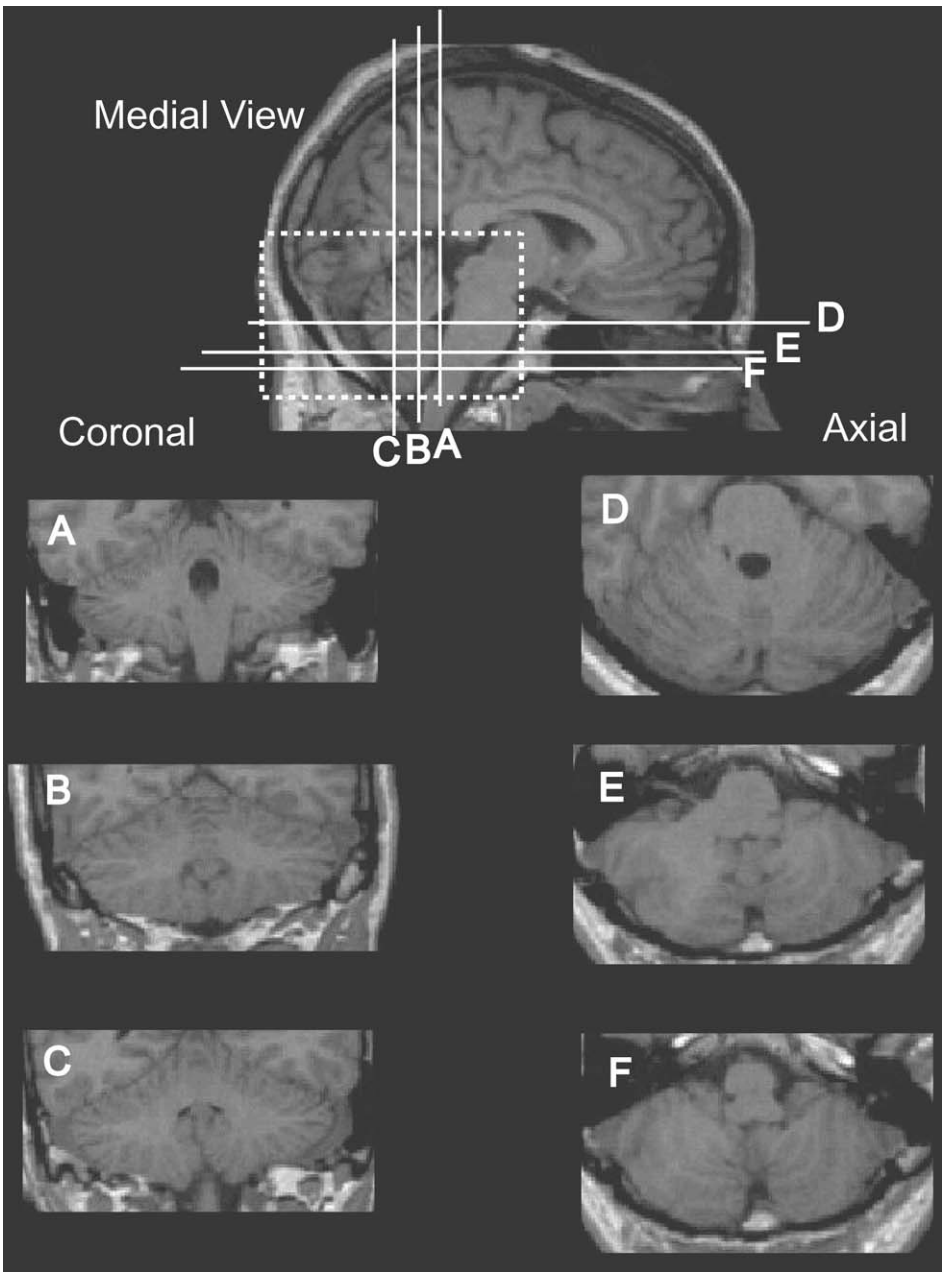


Figure 2 Radiological anatomy of the human cerebellum. Magnetic resonance image (T1 weighted) of the human brain and of the cerebellum. Upper part: medial view of the brain. White lines indicate planes of sectioning. ABC: coronal VIEW of the cerebellum; CDE: axial VIEW of the cerebellum. Courtesy of Dr. Francesco Tomaiuolo, Clinical and Behavioral Neurology Lab, IRCCS S. Lucia, Rome.

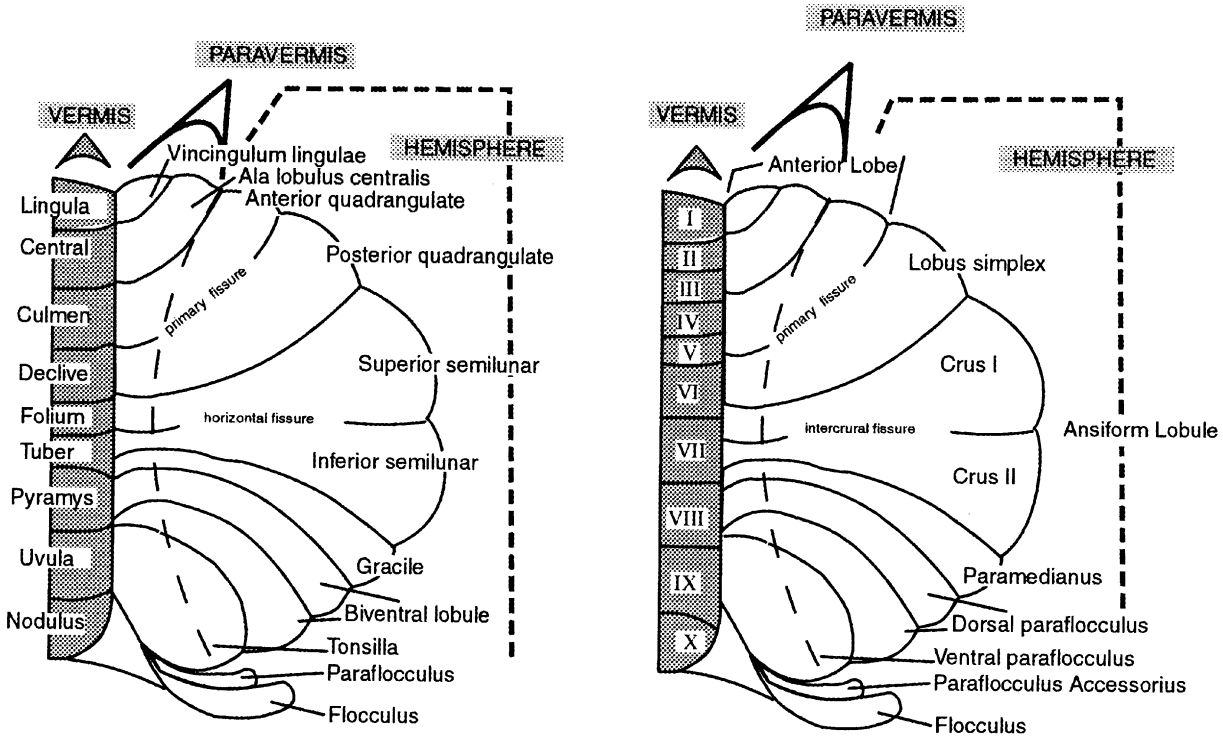


Figure 3 Schematic view of the human cerebellum. (Left) Nomenclature derived from comparative studies. (Right) Nomenclature used in the clinical literature. The Roman numerals indicate the vermal subdivision according to Larsell.

prefix H to the Roman numerals. The cerebellum covers the fourth ventricle and is attached to the brain stem by three pairs of fiber bundles—the inferior, median, and superior cerebellar peduncles.

At the cellular level, the cerebellar cortex is extremely homogeneous in its organization. The intrinsic circuitry is based mainly on five types of neurons organized in three layers: the granular layer, in which granule cells are present, is closest to the white matter; the Purkinje cell layer is the intermediate one and is formed by a single line of Purkinje cells; and the molecular layer is the most external one, in which, beside the dendrites of the Purkinje cells and the parallel fibers of the granule cells, three types of interneurons are present (i.e., the Golgi, stellate, and basket cells).

Granule cells are the only excitatory neurons of the cerebellar cortex and represent by far the largest population of neurons in the cerebellum and possibly the whole brain. The short claw-like dendrites of each granule cell receive many terminals from many different mossy fibers; each terminal forms a complex synaptic arrangement, the glomerulus, in which a terminal from a Golgi axon also participates. The mossy fiber system is the main afferent system. It conveys all information reaching the cerebellum except for that originating from the inferior olives, which reaches the cerebellum through the climbing fibers. These latter fibers terminate directly on the proximal smooth branches of the Purkinje cells. Each Purkinje cell receives only a single climbing afferent, but each climbing fiber contacts up to 10 Purkinje cells. Purkinje cells present a dendritic tree that is relatively smooth in the basilar part but is full of spines in the distal part where terminals from many parallel fibers are present. The general branching of the Purkinje cell dendritic tree is flattened, with its main axis perpendicular to the main axis of the parallel fibers. This arrangement allows each Purkinje cell to be contacted by many different parallel fibers. Each fiber makes very few synapses over the same Purkinje cell but contacts many different ones. The axon of the Purkinje cells is the only way out of the cerebellar cortex and provides a powerful inhibitory control over the deep cerebellar nuclei and some vestibular nuclei.

Whereas the stellate and the basket cells provide a feed-forward inhibition over the Purkinje cells, the axons of the Golgi cells synapse over the incoming

mossy fibers providing a powerful inhibitory feedbackward control over the granule cells. Although all three types of interneurons are GABAergic, only the Golgi cells also colocalize glycine. Recently, a new type of cerebellar interneuron not expressing GABA or glycine was identified—the unipolar brush cell. These cells are located in the granular layer; they receive massive input from a single mossy fiber and their axons contact granule and Golgi cells. Their distribution is not uniform throughout the cerebellar cortex and their function is still a matter of debate. Another cell type recently characterized is that formed by the cells located just under the Purkinje cell layer: the Lugaro cells. These cells present their major cell and dendritic axes parallel to the Purkinje cell layer; they are inhibitory interneurons and their function must still be clarified.

Besides mossy and climbing fibers, a third, often neglected, afferent system has been known since the late 1960s. It is formed by beaded fibers that terminate throughout the cortical layers. This system is considered to play a modulatory role on the activity of the cortical circuits. These fibers originate from different extracerebellar sources of known modulatory function, such as locus ceruleus, raphe nuclei, or brain stem cholinergic cells, and from different nuclei of the hypothalamus.

On their way to the cerebellum, mossy fibers as well as climbing fibers give off collaterals for the deep cerebellar nuclei from which all the cerebellar output fibers originate. On their way to extracerebellar targets, deep cerebellar nuclei axons give off collaterals that contact granule cells or the neurons that give off mossy fibers in precerebellar nuclei. In addition to excitatory projecting neurons, deep cerebellar nuclei also contain small GABAergic cells that project to the inferior olive. The general schema of the cerebellar circuitry is shown in Fig. 4.

The strong anatomical and functional interrelationships between the deep cerebellar nuclei and the cerebellar cortex were clearly defined by Masao Ito, who proposed the term corticonuclear microcomplex (Fig. 5) to describe the basic functional mode of the cerebellum.

Parallel longitudinal zones can be identified in the cerebellar cortex on the basis of output organization and differences in the expression of certain proteins by subpopulations of Purkinje cells. These microzones are perpendicular to the main axis of the folia and they extend across one or more lobules, some of them spanning the entire rostrocaudal length of the cerebellum. These zones were identified and numbered and their connectivity was fully determined, demonstrating that this pattern is preserved in the corticonuclear projections as well as in the olivocerebellar ones. Therefore, each longitudinal zone projects to a specific sector of the deep nuclei and receives the same input from the inferior olive. This basic schema is repeated through the mediolateral extent of the cerebellum and can be identified as the basic functional module of the cerebellum (Fig. 6). The organization of the mossy fiber input is more complex and its relation to the microcomplex organization is still uncertain. Different brain stem and spinal centers project to the cerebellum through the mossy fiber system, terminating in lobule-specific patterns with a roughly preserved somatotopical pattern with the hindlimb located ventrally and face and forelimb located more rostrally in both anterior and posterior lobules. Within this general arrangement, the finer somatotopy is "fractured," as shown by Wally Welker in 1987, without preservation of the precise topographical relation between adjacent receptive fields (Fig. 7).

A. Cerebellar Afferents

1. Mossy Fiber Afferent System

As stated previously, the vast majority of cerebellar afferents reach the cerebellar cortex as mossy fibers. Spinocerebellar and trigeminal cerebellar mossy fibers convey interoceptive, proprioceptive, and exteroceptive information from limbs, trunk, and face. The input is directed to the cerebellum by second-order neurons and not directly by primary sensory neurons. The organization of the vestibulocerebellar system is different. The labyrinth sends afferents directly to the cerebellar cortex; in addition to these primary vestibular afferents, secondary vestibular fibers originating from the four vestibular nuclei also reach the cerebellum. These latter nuclei are also unique in the cerebellar organization since they are the only ones, besides the deep cerebellar nuclei, that receive Purkinje cell axons. This specific organization is confined to the oldest cerebellar region, termed the vestibulo- or archicerebellum.

Many different cell groups located within the reticular formation give origin to mossy fibers. The lateral reticular nucleus is the recipient of mainly spinal afferents but it also receives relevant amounts of fibers from the cerebral cortex and the red nucleus. Mossy fibers from the lateral reticular nucleus terminate bilaterally in the cerebellum with a clear ipsilateral

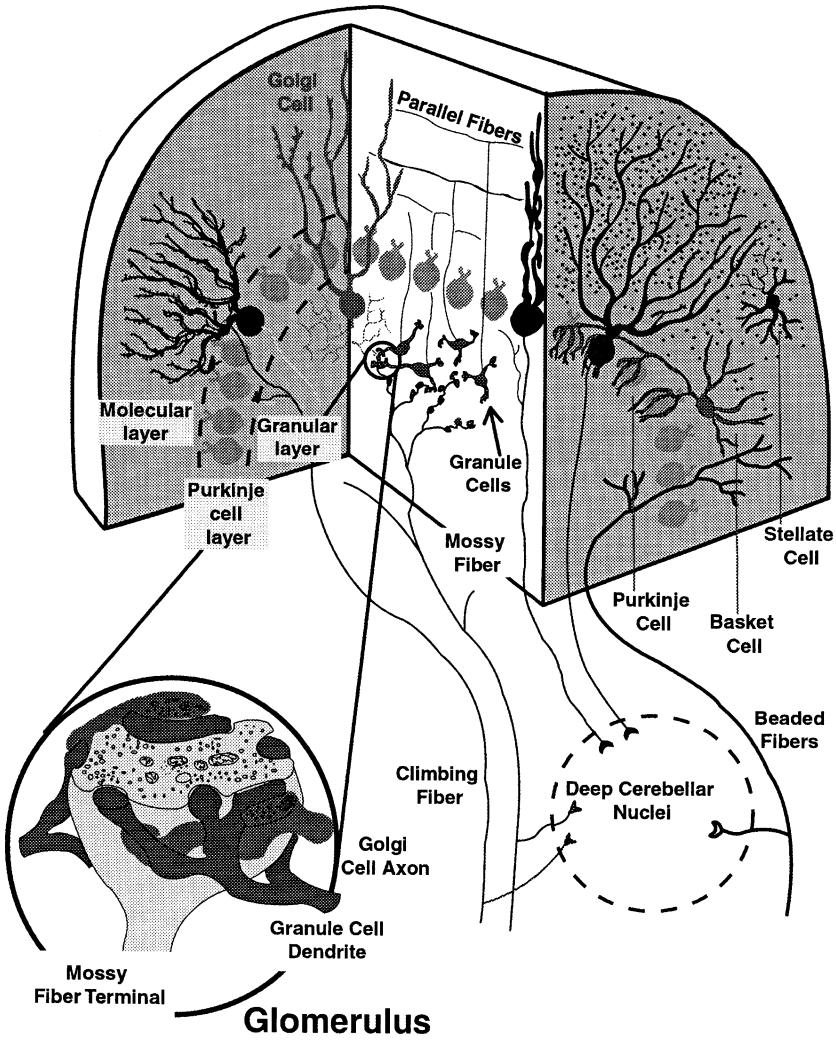


Figure 4 Schematic reconstruction of the circuitry of the cerebellar cortex. A single cerebellar folium has been sectioned vertically, both in longitudinal and transverse planes. (Lower left) Enlargement of a single glomerulus (modified with permission from L. Heimer, *The Human Brain and Spinal Cord*, Fig. 120, p. 213. Copyright © 1983 by Springer-Verlag).

prevalence. The paramedian reticular nucleus conveys information from the somatosensory and frontal cortices mainly to the vestibulocerebellum. The pontine reticular cells provide visual-related information from cortical and subcortical structures mainly to vermal lobules VII and VIII.

Many of the cerebral cortical areas are connected with the contralateral cerebellum through the pontine mossy fiber system, the largest group of mossy fibers. Corticopontine fibers originate from layer V pyramidal cells and their terminal fields are organized topographically with the anterior cortical region located medially and the posterior one located laterally. The topographic organization is maintained in the pontocerebellar projection to the contralateral cerebellum (Fig. 8). However, this projection is not strictly contralateral since both ipsilateral and bilateral fibers have been described.

2. Climbing Fiber Afferent System

There is only one known source of climbing fibers—the inferior olive. This structure receives information from a wide range of brain structures funneling data from the periphery (spinal cord, dorsal column nuclei, trigeminal and vestibular nuclei, and

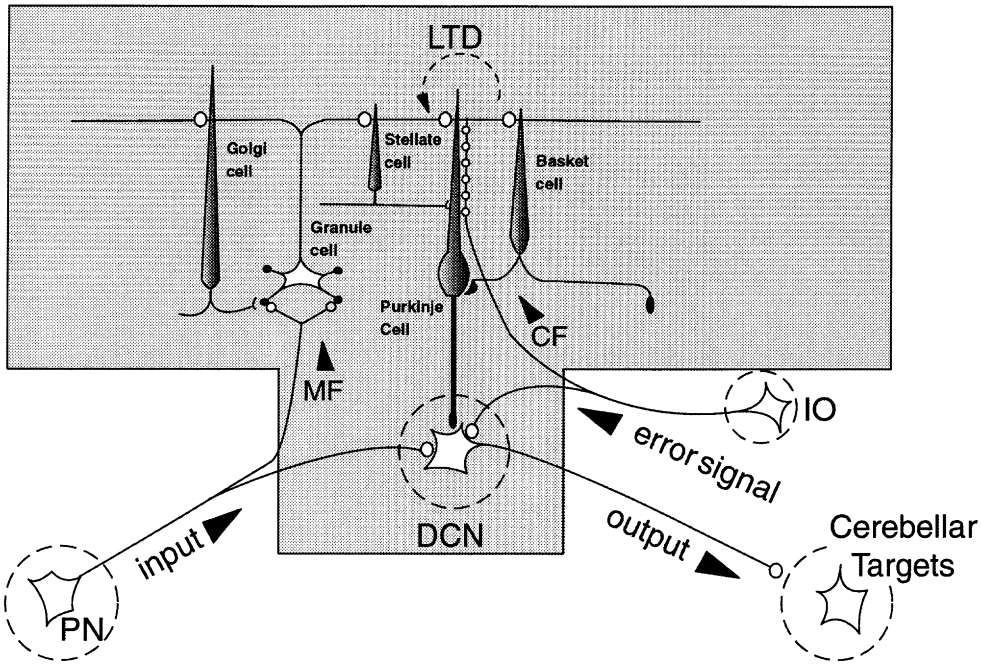


Figure 5 Cerebellar corticonuclear microcomplex as proposed by Masao Ito. DCN, deep cerebellar nuclei; PN, pontine cerebellar nuclei, IO, inferior olive; MF, mossy fiber; CF, climbing fiber; LTD, long-term depression; White cells, excitatory neurons; gray cells; inhibitory neurons; ○, excitatory synapses; ●, inhibitory synapses.

nucleus of the solitary tract) and from higher brain structures (red nucleus, superior colliculus, pretectal area, and cerebral cortex) to the cerebellar cortex and deep nuclei. It is worth noting that a significant number of deep cerebellar nuclei fibers provide a return route from the cerebellum to the inferior olive cells.

Olivo cerebellaris projections are topographically organized in such a way that the different subnuclei of the inferior olive project to different longitudinal stripes of the deep cerebellar nuclei and the cerebellar cortex. This pattern coincides with the neurochemical pattern evidenced with zebrin and described previously. The overall organization is such that each inferior olive fiber sends one collateral to the deep cerebellar nuclei and one to the cerebellar cortex that project to the same sector of the deep nuclei.

B. Cerebellar Efferents

The only output system of the cerebellum is represented by the deep cerebellar nuclei. These nuclei receive massive projections from the Purkinje cells of the overlying cerebellar cortex. This connection is clearly topographic, with a general mediolateral and anteroposterior arrangement. Therefore, more medially placed Purkinje cells project to the medial nucleus, the most lateral ones project to the lateral nucleus, and the anterior lobe projects to the more anterior sectors of the nuclei, and the posterior lobe to the more posterior sectors. However, this general arrangement is not consistently maintained and some divergence and convergence have been reported. Virtually all climbing and mossy fibers give off a collateral for the cerebellar nuclei on their way to the cerebellar cortex. Thus, within the deep cerebellar nuclei there is a convergence of extracerebellar excitatory and Purkinje inhibitory synapses over the same projection neurons. The latter synapses greatly outnumber those of extracerebellar origin and are in a position, over the somata and proximal dendrites, to highly influence neuron firing. From the nuclei, efferent fibers leave the cerebellum mainly through the superior cerebellar peduncle; fewer fibers, mainly originating from the fastigial nucleus, leave the cerebellum through the inferior cerebellar peduncle. Once the brain stem is reached, both efferent systems give off an ascending and a descending branch.

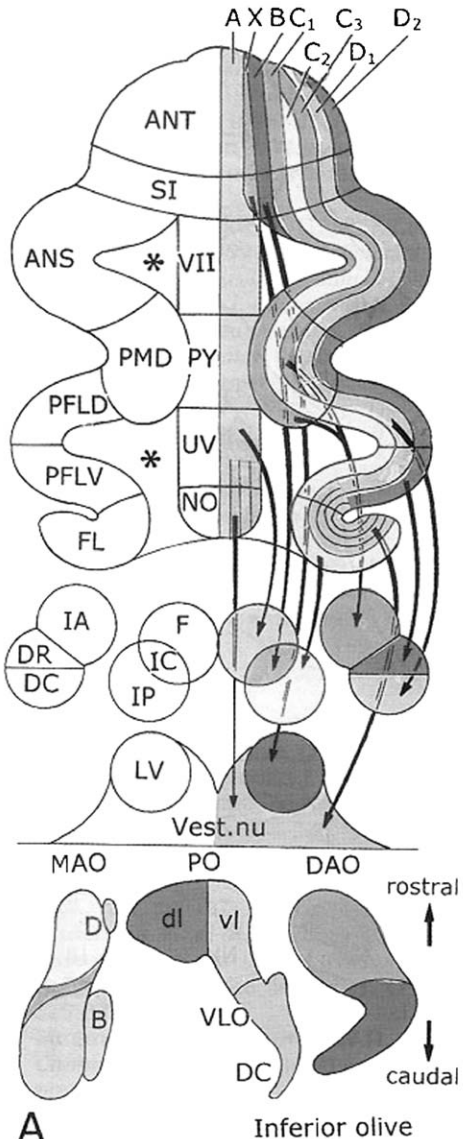


Figure 6 Zonal arrangement of corticonuclear and olivocerebellar projections illustrated on a flattened cerebellar cortex. Three groups of cerebellar nuclei with their corticonuclear projection zones can be distinguished: (i) the fastigial nucleus (F; the target nucleus of the vermal A zone), which is continuous through the intermediate cell group (IC; X zone) with the globose or posterior interposed nucleus (IP; C_2 zone); (ii) the emboliform or anterior interposed nucleus (IA; C_1 and C_3 zones) and the dentate nucleus, which can be subdivided into ventrocaudal (DC) and dorsomedial (DR) parts (target nuclei of the D_1 and D_2 zones, respectively); and (iii) the lateral vestibular nucleus of Deiters (LV; target nucleus of the vermal B zone). Zones in the flocculus and the nodulus project to the vestibular nuclei. The inferior olive is shown at the bottom of the figure in a horizontal projection. The zonal projections of the individual subnuclei are indicated with the same gray shading as shown in the top of the figure (reprinted from J. Voogd and M. Glickstein, The anatomy of the cerebellum. *Trends Neurosci.* **21,** pp. 370–375, copyright 1998 with permission from Elsevier Science).

The descending branch of the fastigial nucleus efferents is formed by ipsilateral, contralateral, as well as bilateral projections directed to many precerebellar nuclei, namely, the vestibular nuclei, the inferior olive complex, and large sectors of the reticular formation. The ascending branch is smaller and reaches several structures of the midbrain, the superior colliculus, and the dorsal thalamus.

Efferents from the interposed nucleus exit through the superior peduncle and are also organized in an ascending and a descending branch. Ascending fibers reach the red nucleus, where they terminate in the magnocellular part in a topographic manner, and the dorsal thalamus terminating in the VL nucleus as well as in the intralaminar nuclei. Descending interposed fibers terminate in reticular and inferior olive precerebellar structures.

The lateral nucleus is the recipient of the projections from the lateral part of the hemisphere and its efferents reach the contralateral parvicellular red nucleus and ventrolateral and intralaminar thalamic nuclei with the ascending branch. The descending branch is formed by fibers directed to the precerebellar stations that reach the lateral cerebellum: reticular formation, inferior olive, and the pontine nuclei. It is worth noting that the deep cerebellar nuclei also give off recurrent projections to the cerebellar cortex.

C. The Transcerebellar Loops

As already mentioned, many of the cerebellar circuits are organized in recurrent loops. This is particularly evident when the rubrocerebellar and cerebrocerebellar circuits are taken into account.

1. Rubrocerebellar Loop

A magnocellular and a parvicellular component can be recognized within the red nucleus. Both receive cerebellar fibers, the former from the interposed nucleus and the latter from the dentate nucleus. The two components project to different targets. Magnocellular cells give origin to rubrospinal fibers, whereas parvicellular cells project to the inferior olivary neurons and thus reverberate the information to the cerebellum through the climbing fibers. Within the red nucleus, there is a convergence of cerebellar and cortical inputs over the same neuron. This is particularly true in the parvicellular part, where cerebellar terminals tend to concentrate in the soma, and

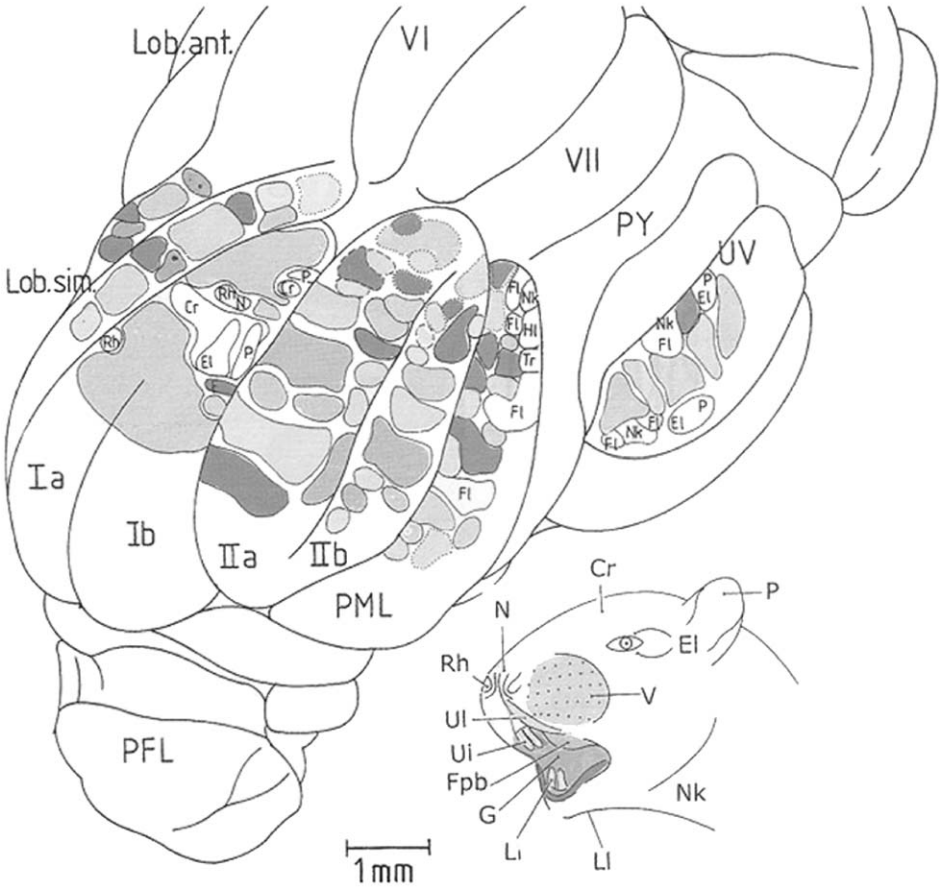


Figure 7 Representation of the fractured somatotopical pattern of multiple patches in the posterior lobe of the cerebellum of the rat. The gray scale code indicates the corresponding receptive fields for the different patches. Cr, crown; El, eyelid; FL, flocculus; Fpb, furry buccal pad; G, gingiva; I–X, lobules according to Larsell; Li, lower incisor; Lob. Ant., anterior lobule; Lob. sim., lobulus simplex; Nk, neck; P, pinna; PFL; paraflocculus; PML, paramedian lobule; PY, pyramis; Rh, rhinarium; Ui, upper incisor; Ul, upper lib; UV, uvula; V, vibrissae [modified with permission from M. Glickstein, C. Keo, and J. Stein (Eds.), *Cerebellum and Neuronal Plasticity*, Fig. 1, p. 114. Copyright © 1987 by Plenum Press].

proximal dendrites while cortical synapses are present in the distal dendritic branches. This arrangement is not fixed but can be functionally modulated as shown by lesion and inactivation experiments.

2. Neocortical Cerebellar Loop

The ascending branches of the efferents from the interposed and lateral nuclei and partially from the fastigial nucleus reach the contralateral thalamus and, in a much smaller amount, recross the midline in the thalamus, terminating ipsilaterally. This latter component is formed by axon collaterals of the main contralateral projection. Within the thalamus the cerebellar fibers terminate in the ventrolateral nucleus and in the intralaminar nuclei. The different cerebellar nuclei have segregated terminal areas within the thalamus, especially in higher mammals. Nevertheless, the precise relationships between cerebellar terminals and cortical and basal ganglia projecting thalamic cells must still be clarified. In general terms, cerebellar recipient thalamic nuclei project mainly to the frontal and prefrontal area, with some components reaching the posterior parietal association areas. The returning loop originates from layer V pyramidal cells located in almost all cortical areas except for the pole of the temporal lobe. These fibers do not reach the cerebellum directly; they terminate in a topographic pattern in the pontine nuclei, which in turn reverberate the information over the cerebellar cortex and deep nuclei.

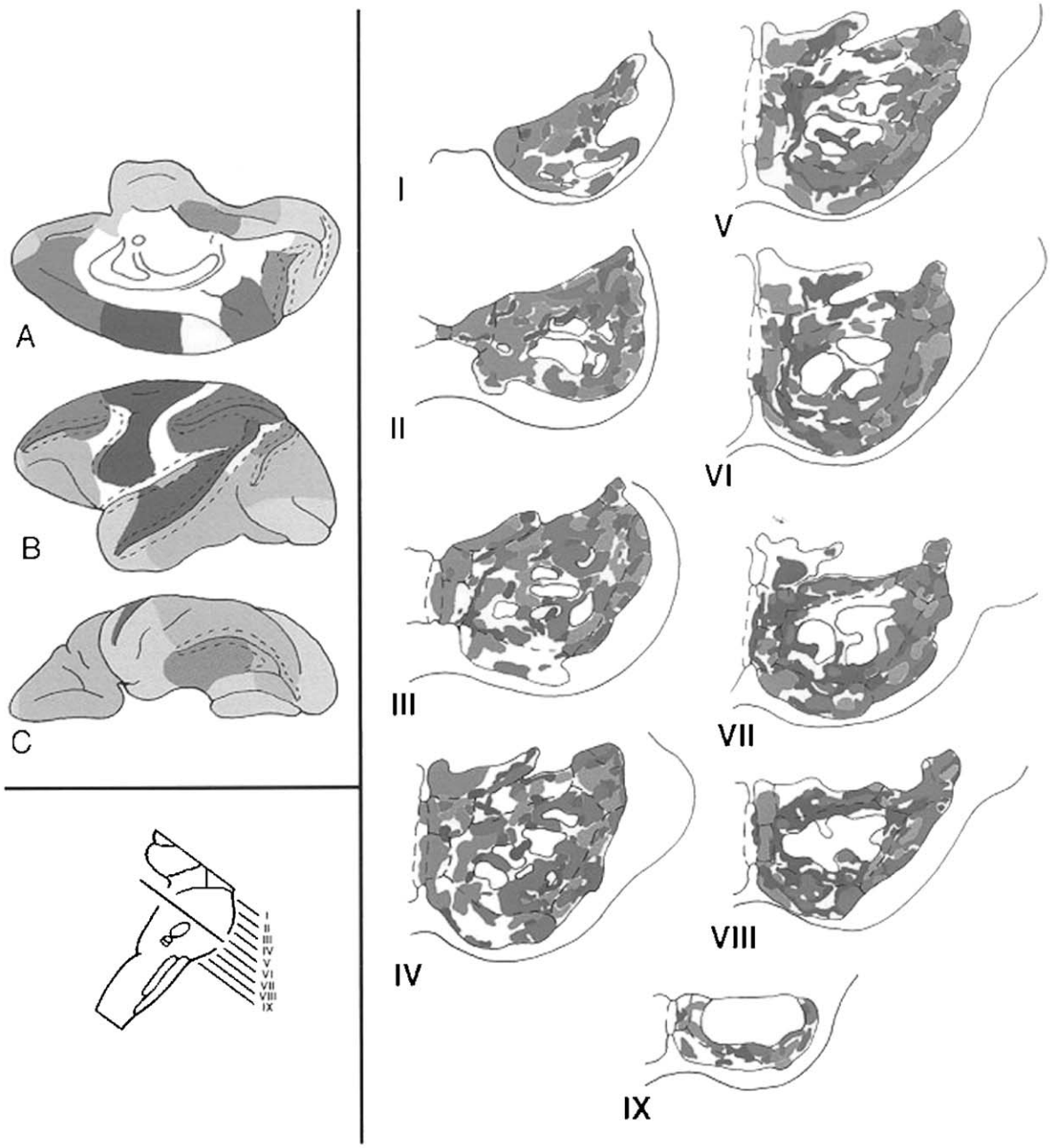


Figure 8 Composite color-coded diagram illustrating the distribution within the basilar pons of the rhesus monkey of projections derived from associative cortices in the prefrontal, posterior parietal, temporal, and parastriate and parahippocampal regions and from motor, premotor, and supplementary motor areas. The medial (A), lateral (B), and ventral (C) surfaces of the cerebral hemisphere are shown in the upper left. The plane of section through the basilar pons is shown in the lower left, and the rostrocaudal levels of pons I–IX are shown in the right. Cerebral areas that have been demonstrated to project to the pons using either anterograde or retrograde tracers are depicted in white, those areas studied with both anterograde and retrograde studies and found to have no pontine projections are shown on the hemispheres in light shading, and those with no pontine projections according to retrograde studies are shaded in gray. The dashed lines in the hemisphere diagrams represent the sulcal cortices. In the pons diagrams the dashed lines represent the pontine nuclei, and the solid lines depict the traversing corticofugal fibers. The associative corticopontine projections are substantial. There is a complex mosaic of terminations in the pons, and each cerebral cortical region has preferential sites of pontine terminations. There is considerable interdigitation of the terminations from some of the different cortical sites but almost no overlap (reproduced with permission from Schmahmann, 1997).

In Fig. 8, the pattern of the corticopontocerebellar loop is shown.

III. NEUROCHEMISTRY

A. Neurotransmission in the Cerebellar Circuits

As mentioned previously, the vast majority of the cortical interneurons are inhibitory and use GABA as a neurotransmitter. Although no other neurotransmitters have been shown for stellate and basket cells, in Golgi cells colocalization of GABA and other putative neurotransmitters has been shown. Approximately 70% of Golgi cells colocalize GABA and glycine, whereas choline acetyltransferase, the synthesizing enzyme for acetylcholine, has been found in approximately 5% of the Golgi cell population. The hypothesis that other neurotransmitters besides GABA are used by Golgi cells is confirmed by the presence of GABA, glycine, and NMDA receptors on granule cells that are the only target for Golgi axons.

Since the 1960s, Purkinje cells have been considered to use GABA as a neurotransmitter notwithstanding the extremely weak GABA or GAD labeling with immunohistochemical techniques. Recent *in situ* hybridization data have confirmed that Purkinje cell bodies and terminals do contain GABA. In particular, Purkinje cell terminals in the cerebellar deep nuclei have been shown to colocalize GABA and taurine, although the function of the latter amino acid is unknown.

The only excitatory circuitry of the cerebellar cortex is the granule cell parallel fiber one, and this circuitry employs glutamate as a neurotransmitter. The excitatory effect of glutamate as well as of parallel fiber stimulation on Purkinje cells has been widely studied and characterized at the molecular level.

Also, the two principal cerebellar afferent systems, the mossy fibers and the climbing fibers, are known to be excitatory in nature. Several lines of evidence indicate that glutamate is the main neurotransmitter of the mossy fibers. On the other hand, there is debate regarding which molecule is used by the climbing fibers; glutamate and aspartate are the main candidates. Other molecules have not been excluded. For instance, mainly on the basis of immunohistochemical evidence, acetylcholine has been proposed as a neurotransmitter for the mossy fibers originating in the medial vestibular nucleus and terminating in the flocculonodular lobe.

The so-called "third afferent system" is heterogeneous with respect to the site of origin and to the neurotransmitter employed. Noradrenergic beaded fibers have been described as originating from the locus ceruleus as well as from A8–A10 adrenergic cell groups in the mesencephalon. A diffuse serotonin system of beaded fibers has been observed immunohistochemically in the cerebellum. Surprisingly, these fibers do not seem to originate in the raphe but in a different part of the brain stem, such as the paramedian and lateral reticular formation, periolivary regions, and the lateral reticular nucleus. On the other hand, no data are available on the nature of the neurotransmitter employed by the raphe cerebellar projection that also sustains the beaded fibers.

B. Chemical Compartmentalization

Different chemical markers are unevenly distributed within the cerebellar cortex in adulthood or transiently during specific developmental phases. These chemical differences clearly challenge the idea of a uniform organization of the cerebellar cortex, and their correlation with functional diversities of the different cerebellar lobules might provide useful information to help clarify the cerebellar function. These many chemical markers include glycolipids as well as proteins. In the adult, these markers are mainly organized in seven sagittal bands of different intensities. In the anteroposterior dimension, chemical patterning of a lower number of markers has been shown. Karl Herrup and Barbara Kuemerle attempted to summarize the available data in a three-dimensional grid of the cerebellar cortex (Fig. 9).

IV. PHYSIOLOGY

The classical definition of the cerebellar function is that of fine-tuning for muscle control. This view derived mainly from the observation that in humans and in experimental animals, cerebellar damage impairs posture and fine coordination of movements. This was supported by a variety of experimental and clinical data. Nevertheless, there is still debate regarding which physiological events support the cerebellar function as motor controller. In general, it is believed that cerebellar control is involved with motor adaptation and motor learning more for compound limb movements than for simple movements.

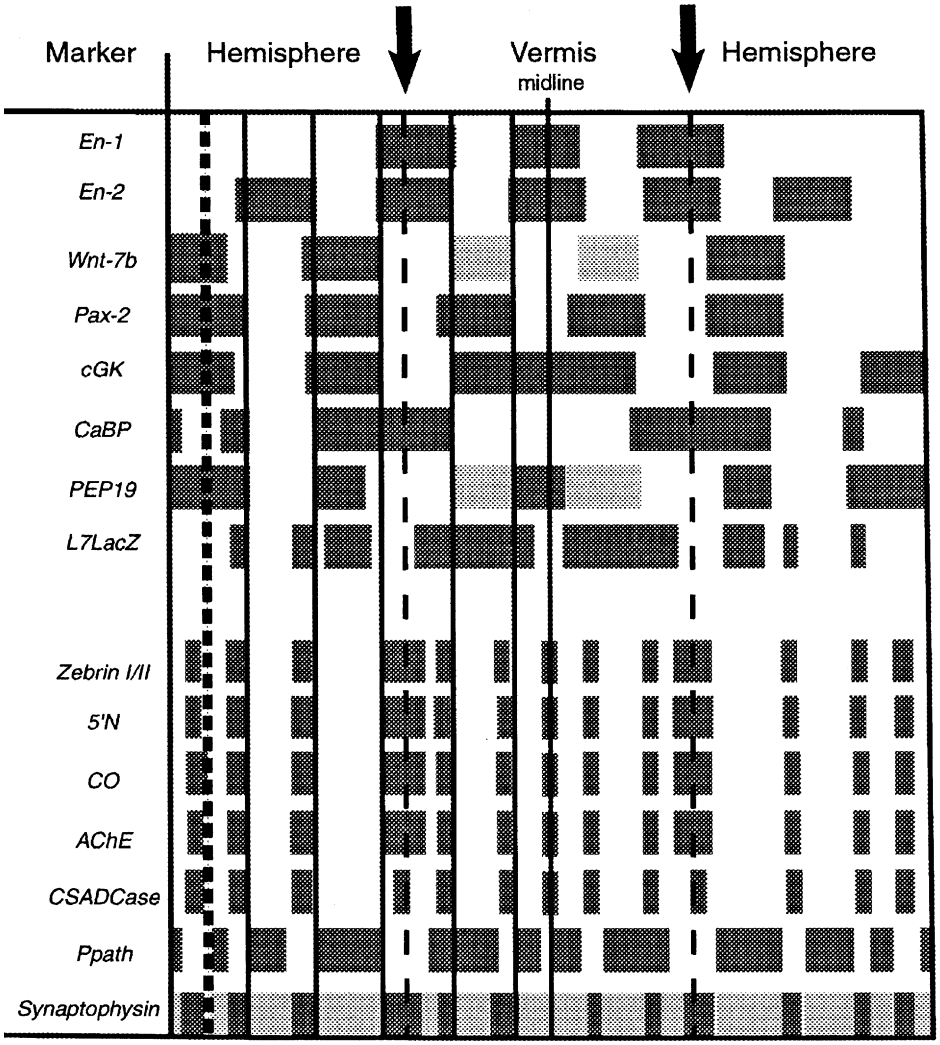


Figure 9 Mediolateral compartment boundaries of the vertebrate cerebellum according to a model proposed by Karl Herrup and Barbara Kuemerle. Six different compartments are located on either side of the cerebellar midline. Shaded boxes denote intensity of gene and antibody expression. Compartmental boundaries are represented by vertical solid lines and vermis–hemisphere junctions are represented by arrows and dashed vertical lines. The top half of the figure represents markers that are expressed transiently during development; the bottom half represents localization of stable markers in the adult cerebellum (from Herrup and Kuemerle, 1997. Reprinted, with permission, from the *Annual Review of Neuroscience*, Vol. 20. © 1997 by Annual Reviews. www.AnnualReviews.org).

For a long time, the most widely accepted theory on how the cerebellum acts was that proposed by David Marr and James Albus and further developed by Masao Ito. According to the Marr–Albus–Ito theory, the main focus of cerebellar activity is acquisition and control of skillful movements through the interactions between climbing fiber (CF) and mossy fiber (MF) inputs in the Purkinje cells (PCs). In resting conditions, PC activity is mainly under the control of the MF–granule cell–parallel fiber (PF) system, whereas the CF system modifies the pattern of PC activation by influencing the strength of the PF/PC synapses. In particular, the convergence of PF and CF activation over the same PC is considered to be capable of inducing a long-term depression (LTD) of the efficacy of PF/PC synapses. Thus, the correct setting for a given movement is coded in the PC–deep nuclei output system by the PF fibers; this pattern is tuned by the so-called "error signal" conveyed through the CF afferents. Although a great deal of experimental evidence can be interpreted according to the Marr–Albus–Ito theory, an increasing body of data challenge the role of LTD in motor learning and the importance of the cerebellum as a motor memory storage site. Most of

the evidence that indicates a primary role of the cerebellum in motor learning derives from experiments examining the substrate responsible for the acquisition of the nictitating membrane reflex in the rabbit or from those responsible for the vestibuloocular reflex (VOR). From these experimental approaches and from evidence derived from patients with cerebellar damage, two major hypotheses about the role of the cerebellum in motor learning have been drawn. First, the cerebellar role in learning is mainly related to participation in the execution of motor behavior (the performance hypothesis). Second, the main cerebellar role is that it acts as a storage site for the motor engrams of different motor tasks (the storage hypothesis). Despite the efforts of many different laboratories, the conflict between the storage and performance hypotheses has not been resolved and a comprehensive theory of the cerebellar role in motor control is far from established.

A different approach to the problem of the cerebellar function focuses on the importance of the cerebellum as a sensory acquisition device rather than a center for motor control and/or learning. This line of thought stems from the anatomical and physiological evidence of massive sensory information that is conveyed to the cerebellum. Recently, it has received important support from functional neuroimaging and experimental data. Within this line is the data acquisition hypothesis, suggested by James Bower and his group. It states that the cerebellum is specifically involved in monitoring and adjusting the acquisition of most of the sensory data on which the nervous system depends. According to this theoretical framework, motor deficits seen after lesion of the cerebellar circuits depend more on the disruption of the inflow of sensory information on which the motor system depends than on a lack of cerebellar control over the motor centers. This approach is in line with the organization of the fractured sensory map described previously. It has been proposed that a key for interpreting the spatial relationships among the different sensory representations is their need to cooperate for better active sensory exploration. The highly expanded and detailed representation of the whisker in the lateral hemispheres of the rat cerebellum seems to support this proposal. The massive cerebellar output toward the motor center is interpreted as the need for direct on-line control of fine movements to optimize the acquisition of sensory information by adjusting the position among the different tactile surfaces and between them and the object being explored.

Marco Molinari and collaborators have also shown the importance of the cerebellum in the acquisition of the sensory information required for implicit learning of a visuomotor task in cerebellar patients. These patients were severely impaired in the acquisition of a task that required a finger-tapping response to the presentation of visual stimuli in a random or fixed sequence. The deficit was particularly due to the inability to recognize the recurrence of the fixed sequence. As shown in Fig. 10, the declarative knowledge of the presented sequence was very poor in all groups of cerebellar patients. On the other hand, a

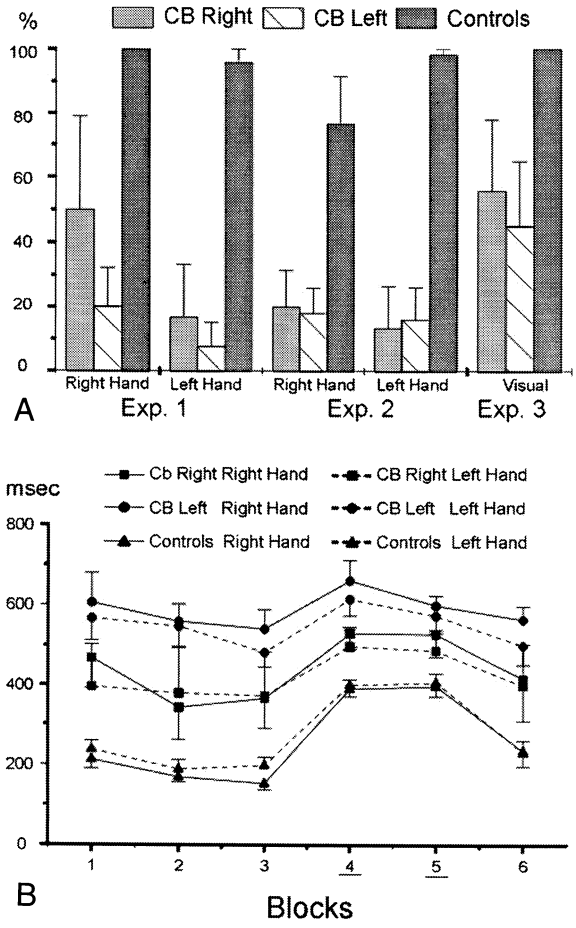


Figure 10 (A) Percentage of sequence items reproduced after a serial reaction time task based on an 8-digit sequence (Exp. 1), a 10-digit sequence (Exp. 2), and after visual presentation only (Exp. 3). (B) Reaction times (msec) after acquisition of a declarative knowledge of the sequence to be reproduced. Note that the presentation of random sequence (underlined blocks) induces an increase in the reaction times in all groups. CB right, group of patients with focal cerebellar lesion on the right side; CB left, group of patients with focal cerebellar lesion on the left side; controls, age- and education-matched control group. vertical bars, standard error (reproduced with permission from Molinari *et al.* (1997). *Brain* **120,** 1753–1762. Reproduced with permission of Oxford University Press).

clear modulation of the response was observed when the subjects received instructions about the digit sequence that would be used in the test. In this condition, a clear improvement in reaction times was observed. This suggests that cerebellar patients are particularly impaired in detecting a sequence and that performance can be improved if knowledge of the sequence has been previously acquired.

The importance of the cerebellum in sensory analysis has also been reported in other forms of learning—that is, in visual perceptual learning by Lucia Vaina and coworkers or in observational learning by Maria G. Leggio and coworkers. In the former functional magnetic resonance imaging (fMRI) study, clear cerebellar activation linked to an early phase of learning a motion perception task was observed. In the latter study, it was reported that lesioning the cerebellum can impair the acquisition of spatial strategies through observation. Both cases clearly demonstrate the role of the cerebellum in processing the sensory information required by the cortical modules during learning.

V. CEREBELLUM AND COGNITION

Possibly the most exciting aspect of cerebellar research in recent years is the converging evidence demonstrating the importance of the cerebellar computational properties for cognition. In 1986, in a seminal work, Henrietta Leiner, Alan Leiner, and Robert Dow challenged the scientific community's generally accepted dogma that the cerebellum is a pure motor structure. Their work was based mainly on anatomic and phylogenetic considerations and they proposed a significant role for the cerebellum in mental functions. Since then, clinical, experimental, and particularly functional neuroimaging data have been reported indicating cerebellar participation in a variety of cognitive functions from mood control to language, from attention to timing, and from spatial data management to memory. This completely new field in cerebellar research has profoundly affected our knowledge of what the cerebellum does and has forced us to reconsider the generally accepted theories on the cerebellar function.

A. Implicit Learning

The classical definition of implicit memory is that of Daniel Schacter—that is, "Implicit memory is revealed when previous experiences facilitate performance on a task that does not require conscious or intentional recollection of those experiences." In contrast to implicit memory, which emphasizes "implicitness" during retrieval, implicit learning primarily refers limited awareness/attention during encoding. This type of learning includes quite different phenomena, such as classical conditioning; learning of skills, procedures, or sequences; priming; and category learning. In many instances, the cerebellum has been considered to be the storage site of such information or directly involved in the acquisition of the implicit competence.

1. Conditioning

Classical eye blink conditioning may be the simplest form of implicit learning, and lesion studies have demonstrated that the cerebellum is essential for this type of conditioning. However, the learning vs performance argument has not been definitively solved in this paradigm either. In this line, although lesion studies in humans have demonstrated eye blink conditioning impairment in cerebellar patients, activation studies in normal volunteers have been controversial in demonstrating cerebellar activation during acquisition of the conditioning. Although this type of learning might appear to be quite different from cognitive functions such as language or working memory, its study is of extreme value since the same basic mechanisms might act independently from the characteristics of the information processed by the cerebellar circuits.

2. Vestibuloocular Reflex

The VOR evokes eye movements in the direction opposite of head movement in order to facilitate vision by minimizing image motion on the retina. This reflex is based on a relatively simple three-neuron arc and its association with the evolutionary old cerebellum, the flocculus, has provided researchers with a simple model for studying the functions of the cerebellum. The gain in the VOR is dependent on visual image motion and head rotation. Thus, by manipulating the ratio between image motion and head rotation it is possible to induce a slip of the image on the retina. If this happens, in order to reduce the retinal slip, adaptation in VOR gain must occur. Since a lesion of the cerebellar flocculus does not abolish the reflex but specifically affects its adaptation, this particular circuit has been identified as the modifiable part of the reflex circuits responsible for its adaptation. Thus, it has

been intensively studied as a model of implicit learning. This model has proven particularly useful in studying the neurophysiological and molecular bases of learning. According to the Marr–Albus–Ito theory, adaptation in the VOR is caused by modification in the output of the cerebellar cortex induced by interaction between the head rotation information carried by the mossy fibers and the image error information carried by the climbing fibers. In particular, this interaction induces LTD in the synapses of the Purkinje cells. Although there is consensus about the importance of the flocculus in the induction of VOR adaptation, its role in the retention of the adaptation is a matter of debate.

3. Motor Skill Learning

As previously mentioned, many theories on cerebellar function focus on motor learning as a key cerebellar function. Many data indicate that the cerebellum supports the acquisition and/or adaptation of simple forms of motor skills. Visuomotor and single joint movements adaptation are impaired in patients with cerebellar damage as well as in animals with permanent or transient inactivation. The cerebellar circuitry and learning-related changes in cerebellar activation have been observed during the acquisition of motor pursuit and trajectory tasks in fMRI studies. On the other hand, experimental evidence indicating a preeminent role of the cerebellar circuits in the acquisition of more complex motor skills is scarce.

In 1990, Jerome Sanes and coworkers reported motor skill learning impairment in cerebellar patients in a mirror tracing task. Recently, the ability of cerebellar patients to improve their performance in generating a trajectory connecting five points was studied by Helge Topka and coworkers. It was found that the patients were able to acquire the task similarly to controls but they were severely impaired in improving their skill at high speed. The authors suggest that these findings indicate a major role of the cerebellum in adaptation learning rather than skill learning.

4. Sequence Learning

Implicit learning of sequences has been investigated by many researchers under a variety of conditions and with many approaches. In general, sequence learning requires making a sequence of motor responses either directly or in response to an external visual or auditory pacing signal. Sequence learning is monitored by assessing performance levels or by recording reaction time from the external cue onset to motor response. In virtually all studies it was found that the cerebellum is required for the acquisition of this skill. However, the significance of the cerebellar contribution to sequence learning is unclear. Different hypotheses have been put forward that mainly stress the role of the cerebellum in timing or processing sensory information or in intervening in sculpturing the motor engram required by the response sequence. Recently, these hypotheses were reviewed by Valentino Braitemberg and coworkers.

B. Cerebellum and Language

Since the beginning of the 20th century, it has been known that cerebellar lesions induce speech deficits. However, these deficits were considered to be dependent on the lack of motor coordination during phonation. Recently, the pure motor nature of the cerebellar influence on speech was challenged on different grounds. Cerebellar activation was documented in language-related tasks independently from motor activity, and cognitive language deficits were also observed in patients with cerebellar damage.

One of the first examples of cerebellar activation in language-related tasks was observed during a verb-generation task. Subsequently, cerebellar activation was confirmed in tasks requiring different types of word generation. Clinically, agrammatic speech and verbal fluency impairment were reported in subjects with cerebellar damage.

Agrammatism is a specific linguistic impairment that does not affect syntactic knowledge; it is often described in patients with left frontal lobe damage. At least two different laboratories reported agrammatism in patients with focal cerebellar lesions. In both cases, the linguistic deficit was very specific. No impairments were detected on an extensive battery exploring general intelligence, orientation, memory, visuospatial skills, praxis, and frontal lobe functions. Language examination was normal for all parameters including sentence comprehension, with the notable exception of dysarthria and agrammatic speech. One group (Marina Zettin and coworkers) interpreted their finding as related to the development of compensatory mechanisms because of the articulatory difficulties due to dysarthric speech; the other group (Caterina Silveri and coworkers) proposed that cerebellar damage might induce agrammatism by affecting the coupling verbal working memory and application of syntactic

rules. According to their interpretation, the cerebellum acts as an "interareal functional coordinator" allowing the production of grammatically correct sentences.

Verbal fluency is the capacity to generate lists of words according to a given rule; it can be either a letter of the alphabet (e.g., retrieval of words that begin with the letter F) or a semantic category (e.g., retrieval of words from the semantic category of "animals"). The existence of verbal fluency deficits in both semantic and letter categories is controversial. Maria Leggio and coworkers reported that cerebellar patients are specifically impaired in the letter category task of word fluency with sparing of the fluency for the semantic category. A qualitative analysis of the strategies used for word grouping during word fluency performances revealed other characteristics of the cerebellar deficit. Although semantic clustering was well preserved in all cerebellar patients, phonological clustering was clearly impaired. The authors interpreted these findings to be due to cerebellar patients' difficulty in acquiring a relatively new verbal skill (phonemic grouping) even though they were not impaired when the verbal fluency was based on a commonly used and well-practiced verbal skill (semantic grouping).

Thus, currently there is little doubt that the cerebellum intervenes in cognitive functions related to language. However, the specific cerebellar contribution is still controversial. Lesion studies suggest that damage of the cerebellar circuits does not directly affect the mental linguistic system but may impair satellite functions, such as verbal working memory or the acquisition of new verbal skills. However, this suggestion needs to be confirmed, possibly with ad hoc fMRI experiments.

C. Cerebellum and Writing

Support for the hypothesis that the cerebellum is a cognitive controller outside the main cognitive processor but necessary for intermodule coordination derives from the observation that cerebellar patients can develop peripheral afferent dysgraphia. According to recent theories on the dysgraphias, peripheral afferent dysgraphia is characterized by two different groups of deficits. The first group, the so-called neglect-related features, is characterized by the tendency to write on the right-hand side of the page and by difficulty in maintaining horizontal lines. Omissions and repetitions of strokes and letters comprise the second group of dysgraphia deficits (feedback-related features). They have been related to defects of visual and proprioceptive feedback during writing movements. Interestingly, in cerebellar patients, in addition to handwriting difficulties consistent with movement dysmetria, some high-order problems are present with the characteristics of the feedback-related features of the peripheral afferent dysgraphia. It has been suggested that the lack of effectiveness of sensory feedback may be due to a complex disorder of sensory feedback related to a form of *inattention* to the feedback. In the case of cerebellar lesions, this *inattention* might be due to the uncoupling of motor planning and proprioceptive feedback.

It is interesting that in all linguistic problems reported, as well as in dysgraphia, an uncoupling of sensory feedback and motor output can be hypothesized. This supports the theory implicating the cerebellum in sensory acquisition and discrimination rather than in motor control.

D. Cerebellum and Working Memory

Working memory has been defined as the ability to maintain and manipulate information "on-line." It has been postulated that this function takes place through the interactions between a central processor (central executive), responsible for manipulating the information, and a rehearsal system, specifically devoted to keeping active the information to be manipulated. It is believed that this latter function is specialized according to different types of information. Furthermore, it is hypothesized that verbal information is retained over a short period of time through a mechanism defined as the phonological loop. Two independent subsystems, the phonological short-term store and the rehearsal system, have been identified within this phonological loop. The phonological short-term store is a limited-capacity store in which the verbal information is held for a short period of time. The rehearsal system is a process that recirculates stored phonological information to prevent its rapid decay. This recirculation is supposed to take place between the phonological short-term store and the phonological output buffer.

fMRI studies have been devoted to the analysis of brain activation during working memory tasks and in particular they have revealed several areas that are specifically active during verbal working memory tasks: left Broca's area, the supplementary motor area, and the cerebellum. It has been hypothesized that

recirculation of the information through these areas may represent the anatomical substrate of the rehearsal system of the phonological loop. A precise definition of the contribution of the different regions to the rehearsal mechanism is still lacking. A recent lesion study by Caterina Silveri and coworkers demonstrated verbal working memory impairment in patients with a right cerebellar lesion. The authors were able to identify the locus of functional damage in the rehearsal system, precisely in the phonological output buffer. This evidence indicates direct involvement of the right cerebellum in silent recirculation of verbal information independently from articulatory functions. This latter finding is in agreement with different fMRI data indicating a larger cerebellar activation in verbal working memory tasks than in silent articulatory rehearsal tasks.

E. Cerebellum and Psychiatric Disorders

Various reports suggesting a possible link between the cerebellum and emotional behavior have been published since the beginning of the 20th century. However, these findings have been overshadowed by the generally accepted view of the cerebellum as a pure motor structure. In the 1970s, the convergence of pathological observations and experimental findings induced some groups to implant a cerebellar pacemaker to treat intractable behavioral disorders. Since then, data have been collected linking gross cerebellar anatomical abnormalities to various psychiatric disorders. Schizophrenia has been linked to different types of abnormalities of the cerebellar vermis; patients affected by attention deficithyperactivity disorder have been shown to have a reduced volume of the posterior vermal lobules (VIII–X). Currently, there is substantial evidence linking the cerebellum with autism, although the precise loci of the cerebellar damage (i.e., vermal lobules VI and VII, dentatothalamocortical pathway, and serotonin innervation) must still be defined. Recently, neuropsychological and neuroimaging data have indicated a possible role of the cerebellum in mood control.

One of the main criticisms of the possible role of the cerebellum in controlling behavior has always derived from the lack of behavioral disorders in patients with cerebellar damage. More careful analysis of cerebellar patients allowed Jeremy Schamahmann to specifically define the cognitive affective pattern of cerebellar patients he defined as having the cerebellar cognitive affective syndrome.

See Also the Following Articles

BRAIN ANATOMY AND NETWORKS • CEREBRAL CORTEX • CEREBRAL WHITE MATTER DISORDERS • CHEMICAL NEUROANATOMY • GABA • MOTOR CONTROL • WORKING MEMORY

Acknowledgments

This work was partially supported by the Italian Ministry of Health, MURST, and CNR grants.

Suggested Reading

Altman, J., and Shirley, A. B. (1997). *Development of the Cerebellar System*. CRC Press, Boca Raton, FL.

Braitemberg, V., Heck, D., and Sultan, F. (1997). The detection and generation of sequences as a key to cerebellar function: Experiments and theory. *Behav. Brain Sci.* **20,** 229–277.

De Zeeuw, C. I., Srata, P., and Voogd, J. (1997). The cerebellum from structure to control. *Prog. Brain Res.* **114**.

Gazzaniga, M. S. (2000). *The New Cognitive Neuroscience*. MIT Press, Cambridge, MA.

Glickstein, M., Yeo, C., and Stein, J. (Eds.) (1987). *Cerebellum and Neuronal Plasticity, NATO ASI series Life Science* Vol. 148**,** Plenum, New York.

Herrup, K., and Kuemerle, B. (1997). The compartmentalization of the cerebellum. *Annu. Rev. Neurosci.* **20,** 61–90.

Jones, E. G. (1985). *The Thalamus*. Oxford Univ. Press, Oxford.

Molinari, M., Leggio, M. G., Solida, A., Ciorra, R., Misciagna, S., Silveri, M. C., and Petrosini, L. (1997). Cerebellum and procedural learning evidence from focal cerebellar lesions. *Brain* **120,** 1753–1762.

Schamahmann, J. (Ed.) (1997). The Cerebellum and Cognition. Academic Press, San Diego.

Silveri, M. C., Di Betta, A. M., Filippini, V., Leggio, M. G., and Molinari, M. (1998). Verbal short-term store-rehearsal system and cerebellum: Evidence from a patient with right cerebellar lesion. *Brain* **121,** 2175–2187.

Trends Cognitive Neurosci. **2**(9) (1998). Special issue: Cerebellum.

Trends Neurosci. **21**(9) (1998). Special issue: Cerebellum.

Cerebral Circulation

JOHN A. JANE, Jr., AARON S. DUMONT, DANIEL E. COUTURE, K. MICHAEL WEBB, DILANTHA B. ELLEGALA, HAYAN DAYOUB, and NEAL F. KASSELL

University of Virginia, Charlottesville

GLOSSARY

aneurysm Focal arterial dilation.

anterior circulation Vessels supplied by the paired internal carotid arteries.

diencephalon Develops from the embryologic prosencephalon and includes the thalamus and hypothalamus.

dural venous sinus Valveless venous structures whose walls are formed by the two dural leaves; it provides the major venous drainage of the cranium.

endovascular therapy Therapy provided from within the vessel lumen utilizing catheter-based technology to treat a variety of vascular disease processes.

intracerebral hemorrhage Denotes the presence of blood within the brain parenchyma; may be spontaneous or secondary to trauma.

posterior circulation Vessels supplied by vertebrobasilar system.

radiosurgery Radiation therapy using stereotactic-guided focused external beam.

telencephalon Develops from the embryologic prosencephalon and denotes the cerebral hemispheres and basal ganglia.

vascular malformations A category of vascular disorders that includes arteriovenous malformations, cavernous malformations, venous malformations, and capillary telangiectasias.

vasculopathy All-encompassing term used to denote any disease of the blood vessels.

The cerebral circulation is formed by a complex vascular network. The cerebral vasculature is subject to a wide range of disorders, including ischemic and hemorrhagic stroke, vascular anomalies, and vasculopathy. Cerebrovascular diseases have the potential to compromise vital cerebral functions, leaving many neurologically devastated. Collectively, cerebrovascular disease poses an enormous societal burden. This article focuses on the specific pathophysiology, presentation, diagnosis, and treatment guidelines for each of these disorders.

I. ANATOMY OF THE CEREBRAL VASCULATURE

A. Normal Anatomy of the Cerebral Arterial Vasculature

1. General Considerations

Although the brain accounts for just over 2% of the body's total weight, it receives nearly 20% of the cardiac output. Whereas larger cerebral vessels are influenced by a balance of sympathetic and parasympathetic tone, vascular tone in medium and small cerebral vessels is altered primarily by the mechanism of autoregulation. Blood flow is tightly regulated at the

arterial level to maintain an average cerebral blood flow (CBF) of 50 ml/100 g/min.

Smooth muscle cells within the cerebral vessels vasoconstrict or vasodilate in response to wall stress and shear to maintain a constant blood flow over a wide range of blood pressures. Within the endothelial cell, a complex balance exists between calcium levels and phosphorylation states of myosin. Central to this balance are endothelium-derived relaxation factors (EDRFs) and endothelial-derived constricting factors (EDCFs). The EDRFs include nitric oxide (NO), prostacyclin, and endothelium-derived hyperpolarization factor; notable EDCFs are endothelin, angiotensin II, prostaglandin $F_{2\alpha}$, and the thromboxanes.

Through autoregulation, local vasculature can also increase CBF to compensate for increased neuronal activity and metabolism (CBF–metabolism coupling). Metabolites produced in the brain and throughout the body can also cause vessel changes and alter CBF. Carbon dioxide has potent vasoactive characteristics. Increased CO_2 causes vasodilatation and a resultant increase in CBF. CBF is less sensitive to arterial oxygen concentrations and increases only at significantly low oxygen levels. Extracellular pH, lactic acid, adenosine, and adenosine triphosphate also have vasoactive properties.

The blood supply to the brain, brain stem, and much of the spinal cord is derived from two paired vessels, the internal carotid and the vertebral arteries. The internal carotid arteries form the anterior circulation and supply most of the telencephalon and much of the diencephalon (Table I). The posterior circulation is a derivative of the vertebrobasilar system and supplies the brain stem and cerebellum as well as parts of the diencephalon, spinal cord, and occipital and temporal lobes (Table II).

2. The Anterior Circulation

a. Internal Carotid Artery The anterior circulation is supplied by the two internal carotid arteries (ICAs). The extracranial ICA is divided into two segments. The most proximal portion of the ICA is dilated and is termed the carotid bulb. The remaining distal portion then ascends through the neck to the skull base without branching. This distinguishes the extracranial ICA from the external carotid artery, which has numerous branches to the face and neck. The ICA then enters the skull base at the carotid canal, traverses the petrous portion of the temporal bone, passes through the cavernous sinus, and finally enters the subarachnoid space at the base of the brain. The ICA can therefore be divided into cervical, petrous, cavernous, and cerebral parts (Fig. 1).

The petrous portion of the carotid is composed of a vertical and horizontal segment. It has two possible branches, the caroticotympanic and vidian arteries, but neither is visualized in the majority of angiograms. The cavernous ICA begins at the level of the petrolingual ligament ascending vertically, then horizontally, and finally vertically again before ending at the level of the anterior clinoid. During its course it gives rise to the meningohypophyseal trunk, the inferolateral trunk, and the capsular arteries of McConnell. The meningohypophyseal trunk divides into the inferior hypophyseal artery, tentorial artery, and small clival arteries. The inferolateral trunk branches into arteries supplying several cranial nerves and the cavernous sinus dura. The capsular arteries, when present, supply the pituitary capsule.

After exiting the cavernous sinus, the supraclinoid ICA gives rise to its first intradural branch, the ophthalmic artery. The ophthalmic artery travels along with the optic nerve through the optic canal to the orbit, where its branches supply various orbital and ocular structures. The supraclinoid ICA also gives rise to the superior hypophyseal artery, which supplies the optic chiasm, pituitary stalk, and anterior pituitary gland. The ICA then proceeds superiorly adjacent to the optic chiasm and bifurcates into its terminal branches, the middle and anterior cerebral arteries.

Before bifurcating, it gives rise to two smaller branches, the posterior communicating artery (PCoA) and the anterior choroidal artery (AChA). The PCoA courses posteriorly, superior to the oculomotor nerve, and joins the posterior cerebral artery, thus creating an anastomosis between the anterior and posterior circulation. The AChA supplies important structures, including the optic tract, cerebral peduncle, internal capsule, thalamus, and hippocampus.

b. Anterior Cerebral Artery The anterior cerebral artery (ACA) runs medially, superior to the optic nerve, and enters the longitudinal fissure, where it arches posteriorly, following the corpus callosum, to supply the medial aspects of the frontal and parietal lobes (Fig. 2). Some of the smaller branches extend onto the dorsolateral surface of the hemisphere.

The proximal ACA gives rise to the medial lenticulostriate and perforating arteries, including the recurrent artery of Huebner, that supply the basal forebrain and optic nerves. The proximal ACAs are connected by the anterior communicating artery

Table I
Major Branches of the Anterior Circulation

Major vessel	Branches and the areas supplied
Internal carotid artery	Petrous portion
	Caroticotympanic artery
	Middle and inner ear
	Vidian artery
	Anastomoses with branches of the external carotid artery
	Cavernous portion
	Meningohypophyseal trunk
	Inferior hypophyseal artery
	Posterior pituitary capsule
	Marginal tentorial artery
	Tentorium
	Clival arteries
	Clivus
	Inferolateral trunk
	Capsular arteries of McConnell
	Anterior and inferior pituitary capsule
	Supraclinoid portion
	Ophthalmic artery
	Optic nerve, choroid, retina, conjunctivae, lacrimal gland, extraocular muscles, falx cerebri, anastomoses with external carotid artery
	Superior hypophyseal artery
	Pituitary stalk, anterior pituitary, optic chiasm
	Anterior choroidal artery
	Optic chiasm and tract, thalamus, internal capsule, cerebral peduncle, choroid plexus, medial temporal lobe
	Posterior communicating artery
	Thalamus, hypothalamus, internal capsule
	Anterior cerebral artery
	Middle cerebral artery
Anterior cerebral artery	Medial lenticulostriate arteries
	Optic nerve, optic chiasm, hypothalamus, fornix, striatum
	Perforating arteries
	Recurrent artery of Heubner
	Basal ganglia, internal capsule, portions of frontal lobe
	Anterior communicating artery
	Infundibulum, optic chiasm, hypothalamus
	Orbitofrontal artery
	Ventromedial frontal lobe, olfactory tract
	Frontopolar artery
	Medial frontal lobe, lateral surface of superior frontal gyrus
	Pericallosal artery
	Callosomarginal artery
	Along with pericallosal artery, their branches supply the anteromedial frontal and parietal cortex

(continues)

Table I *(continued)*

Major vessel	Branches and the areas supplied
Middle cerebral artery	M1 and M2
	Anterior temporal artery
	Anterior temporal lobe
	Lateral lenticulostriate arteries
	Basal ganglia, internal capsule
	M3 and M4
	Orbitofrontal and prefrontal arteries
	Middle and inferior frontal gyri
	Precentral, central and postcentral sulcus arteries
	Precentral, central, and postcentral gyri, anterior parietal lobe
	Posterior parietal artery
	Parietal lobules, supramarginal gyrus
	Angular artery
	Posterior superior temporal gyrus, supramarginal gyrus, angular gyrus
	Temporooccipital artery
	Posterior portions of temporal lobe
	Posterior, medial, and anterior temporal arteries
	Corresponding portions of temporal lobe

(ACoA) near their entrance into the longitudinal fissure. The ACA then gives rise to the orbitofrontal and frontopolar arteries before terminating into the pericallosal and callosomarginal arteries.

c. Middle Cerebral Artery The middle cerebral artery (MCA) begins at the ICA bifurcation and courses into the Sylvian fissure. Before entering the fissure, the MCA bifurcates and these branches ramify over the insula. After emerging from the fissure, the MCA spreads out to supply most of the lateral surface of the cerebral hemisphere (Fig. 2).

This large artery is subdivided into four segments. The first segment, M1, contains the bifurcation and ends at the entrance to the Sylvian fissure. The M1 segment gives rise to the anterior temporal and lateral lenticulostriate arteries supplying the basal ganglia and internal capsule. Within the Sylvian fissure, the M2, or insular, segment ramifies into 6–10 branches that course along the insula. The M3, or opercular, segment begins at the superior portion of the circular sulcus of the insula and ends as the MCA exits the Sylvian fissure to course along the cortical surface. Like the M2 segment, the M3 vessels also ramify and form the end arteries of the M4, or cortical, segment. These cortical vessels supply much of the lateral surface of the frontal, parietal, and temporal lobes.

3. The Posterior Circulation

a. Vertebrobasilar System The posterior circulation denotes the network of the two vertebral arteries (VAs) and the basilar artery (BA). The paired VAs then fuse at the junction of the medulla and pons to form the midline BA, which proceeds rostrally along the anterior surface of the pons (Fig. 3).

b. Vertebral Artery The VA ascends within the transverse foraminae of C6 to C1. After exiting the transverse foramen of C1 it angles posteromedially along the arch of C1 around its superior facet and then turns abruptly again to course rostrally alongside the medulla through the foramen magnum. Along its extracranial course, the VA gives rise to small unnamed muscular and segmental spinal arteries. The distal extracranial VA also supplies meningeal arteries that serve the dura of the posterior fossa.

Before the intracranial VAs join to form the BA, each vertebral artery gives rise to three branches: the posterior spinal, anterior spinal, and posterior inferior

Table II
Major Branches of the Posterior Circulation

Major vessel	Branches
Vertebral artery	Muscular arteries
	Deep cervical muscles
	Segmental spinal arteries
	Vertebral bodies, spinal cord
	Meningeal arteries
	Falx cerebelli, dura around foramen magnum and occipital bone
	Posterior spinal artery
	Posterior third of spinal cord
	Anterior spinal artery
	Anterior two-thirds of spinal cord
	Posterior inferior cerebellar artery
	Lateral medulla; cranial nerves IX, X, and XI; choroid plexus of fourth ventricle; cerebellum (posteroinferior portions, vermis, tonsils)
Basilar artery	Pontine perforating arteries
	Ventral pons and midbrain
	Anterior inferior cerebellar artery
	Cranial nerves VII and VIII, ventrolateral pons, medulla, cerebellum (flocculus, anterolateral portions)
	Superior cerebellar artery
	Deep cerebellar nuclei, superior portions of cerebellum, midbrain
	Posterior cerebral artery
Posterior cerebral artery	Posterior communicating artery
	Perforating arteries
	Cranial nerves III and IV, thalamus, hypothalamus, midbrain, internal capsule
	Posterior choroidal artery
	Choroid plexus of third ventricle, thalamus, fornix, tectum, pineal gland
	Anterior and posterior temporal arteries
	Posterior temporal and anterior occipital lobes
	Medial and lateral occipital arteries
	Occipital lobe and corpus callosum

cerebellar artery (PICA). The posterior spinal artery runs caudally along the dorsolateral aspect of the spinal cord and supplies the posterior third of that half of the spinal cord. The anterior spinal artery joins its counterpart from the opposite side, forming a single anterior spinal artery that runs caudally along the ventral midline of the spinal cord, supplying the anterior two-thirds of the spinal cord. PICA arises at the level of the medulla and, as its name implies, supplies much of the inferior surface of the cerebellum. It also supplies the choroid plexus of the fourth ventricle, cranial nerves (CNs) IX, X, and XI, and much of the lateral medulla.

c. Basilar Artery From its origin at the pontomedullary junction, the BA proceeds rostrally and, at the level of the midbrain, bifurcates into the two posterior cerebral arteries (PCAs). Before this bifurcation, it gives rise to numerous pontine perforating branches, the labyrinthine arteries, and two paired vessels—the anterior inferior cerebellar artery (AICA) and the superior cerebellar artery (SCA). The AICA arises just distal to the basilar origin and supplies the anterolateral cerebellum, the pons, and rostral medulla. The SCA arises proximal to the basilar bifurcation, courses inferior to the oculomotor nerve, and supplies the superior cerebellum, its deep nuclei, and much of the caudal midbrain.

The PCA curves over the oculomotor nerve, around the midbrain, and passes through the superior cistern. The proximal PCA sends perforators to the rostral midbrain, caudal diencephalon, and CNs III and IV. It also gives rise to several posterior choroidal arteries, which supply the choroid plexus of the third ventricle, thalamus, and pineal region. The anterior and posterior choroidal arteries form anastomoses in the vicinity of the glomus. The proximal PCA also gives rise to the posterior communicating artery, connecting the anterior and posterior cerebral circulation. The cortical PCA branches include the anterior temporal, posterior temporal, lateral occipital, and medial occipital arteries. These branches spread out to supply the medial and inferior surfaces of the occipital and temporal lobes. In effect, the PCAs supply areas that serve vision, visual memory, and eye motility.

4. The Circle of Willis

The circle of Willis is a polygonal arcade that connects the two halves of the anterior circulation with the posterior circulation. Its components include both ICAs and ACAs, the ACoAs, PCoAs, PCAs, and the BA (Fig. 4). In cases of major vessel occlusion, either within the circle of Willis or proximal to it, the communicating arteries theoretically permit vital anastomotic flow and prevent neurological damage. In

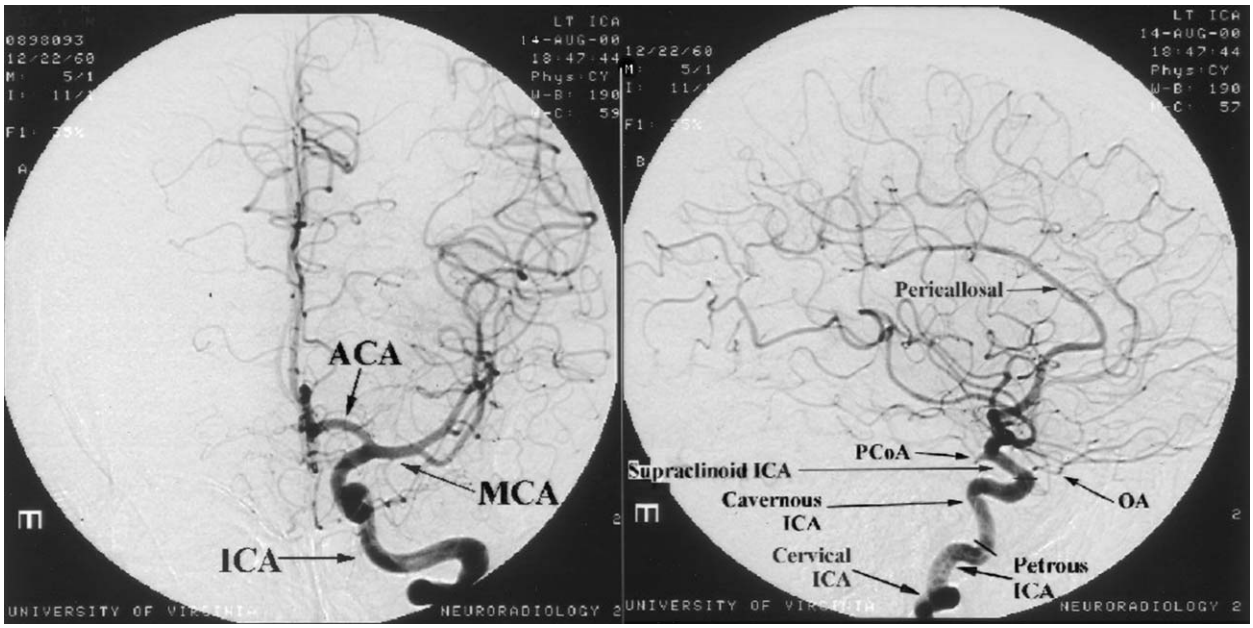


Figure 1 Angiograms after left ICA injection. (Left) Frontal view displays the MCA/ACA origins and major branches. (Right) Lateral view displays the major divisions of the ICA.

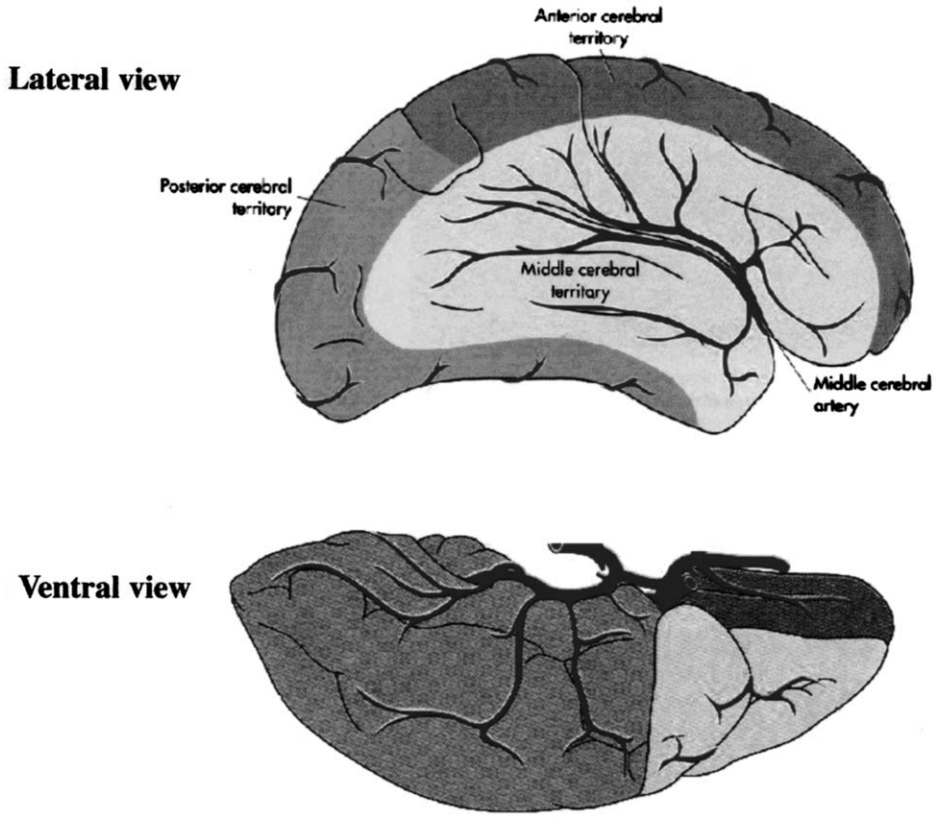


Figure 2 Vascular territories of the cerebral arteries. From Nolte (1993), *The Human Brain: An Introduction to Its Functional Anatomy*, 3rd ed. Mosby, St. Louis. Used with permission.

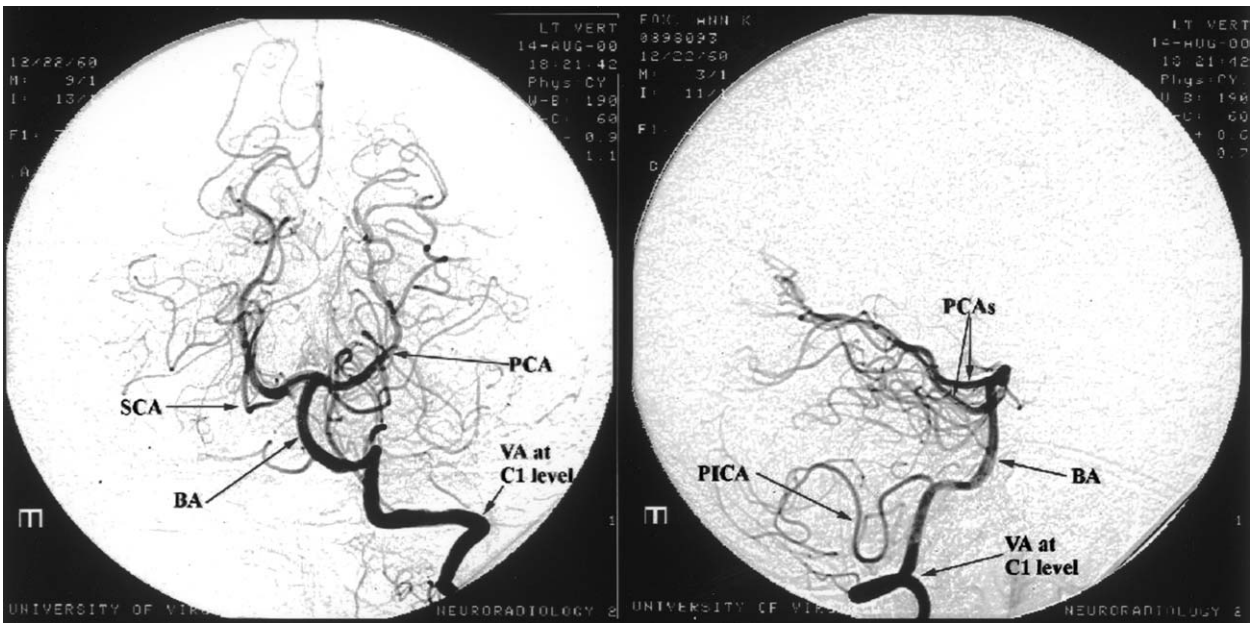


Figure 3 Angiograms after left vertebral artery injection. (Left) Frontal view shows left vertebral, basilar, and posterior cerebral arteries. (Right) Lateral view also reveals PICA.

fact, fewer than half the circles have the classical appearance and asymmetries are common. In rare cases, one of the communicating arteries may be missing, resulting in an incomplete circle.

Other routes of collateral circulation are available. Arterial anastomoses between the extracranial and intracranial carotid exist and may enlarge to a remarkable degree to compensate for slowly developing occlusions. The most important are anastomoses in the orbit between the ophthalmic artery and branches of the external carotid artery. In cases of proximal ICA occlusion, retrograde flow from the external carotid artery through the ophthalmic artery can reach the distal internal carotid territories.

B. Anatomy of the Cerebral Venous System

1. General Considerations

Although the extracranial veins of the scalp and face are not discussed in detail, they do communicate with the intracranial venous system and can have pathological importance. Scalp and facial veins may drain to the dural sinuses via emissary veins that pass through the skull and orbit. These extracranial veins also anastomose with both the cavernous sinus and the basilar, or clival, plexus. They play a relatively minor role in the normal circulatory pattern of the brain but can be important clinically as a nidus for fistula formation, as a path for the spread of infection, or as a collateral drainage system.

The principal route of venous drainage of the brain is through a system of cerebral veins that empty into the dural venous sinuses. The majority of dural sinuses join at the torcula of Herophili (confluence of sinuses) and then split into the left and right transverse sinuses, which flow into the sigmoid sinuses and ultimately into the internal jugular veins. The cerebral veins are conventionally divided into superficial and deep groups.

2. Dural Venous Sinuses

The major dural venous sinuses include the superior and inferior sagittal sinuses, the straight sinus, the superior and inferior petrosal sinuses, the occipital sinus, the transverse sinus, and the sigmoid sinuses (Fig. 5). The superior and inferior sagittal sinuses run in the midline along the superior and inferior edges of the falx cerebri (Fig. 6). The inferior sagittal sinus joins the vein of Galen to form the straight sinus, which then courses along the superior edge of the tentorium cerebelli. The occipital sinus originates at the posterior edge of the foramen magnum and meets the straight sinus and superior sagittal sinus at the torcula of Herophili. The superior petrosal sinus joins the cavernous to the sigmoid sinus and runs along the

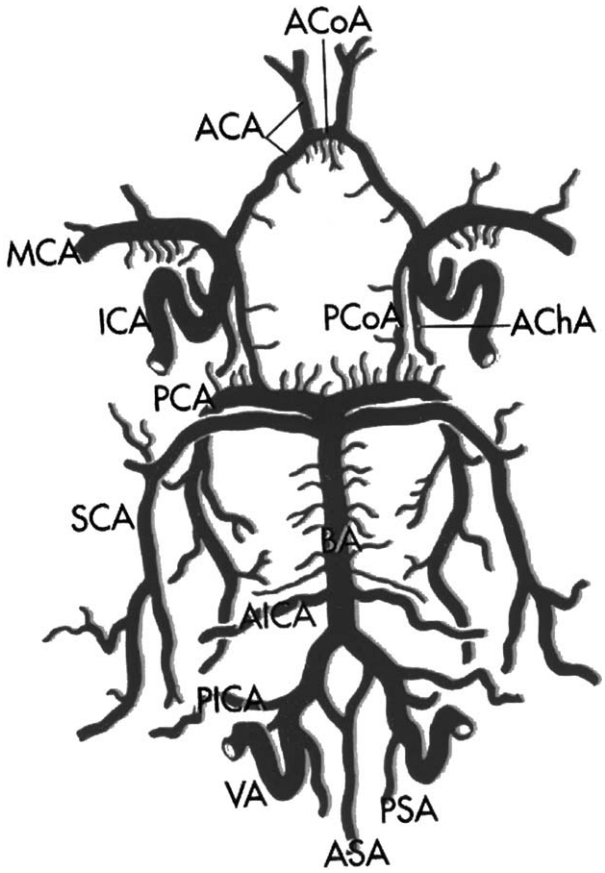


Figure 4 Circle of Willis. From Nolte (1993), *The Human Brain: An Introduction to Its Functional Anatomy*, 3rd ed. Mosby, St. Louis. Used with permission.

superior ridge of the petrous bone. The inferior petrosal sinus has variable tributaries but drains into the internal jugular vein at the jugular bulb.

3. Superficial Venous System

The superficial veins are quite variable and most are unnamed. Only three of these veins are reasonably constant (Fig. 7). The superficial middle cerebral vein runs along the lateral sulcus draining most of the temporal lobe into the cavernous sinus or into the nearby sphenoparietal sinus. The superior anastomotic vein, or vein of Trolard, typically travels across the parietal lobe and drains the superficial middle cerebral vein into the superior sagittal sinus. The inferior anastomotic vein, or vein of Labbe, travels posteriorly and inferiorly across the temporal lobe and connects the superficial middle cerebral vein with the transverse sinus.

4. Deep Venous System

The deep veins are more constant in configuration than are the superficial veins. The major deep vein is the internal cerebral vein, which is formed at the intraventricular foramen by the confluence of the septal vein and the thalamostriate veins (Fig. 5). The septal vein runs posteriorly across the septum pellucidum, and the thalamostriate vein travels in the groove between the thalamus and the caudate nucleus.

Immediately after forming, the internal cerebral vein bends sharply in a posterior direction. This bend is called the venous angle and is used in imaging studies as an indication of the location of the interventricular foramen. The paired internal cerebral veins proceed posteriorly through the transverse cerebral fissure and fuse in the superior cistern to form the unpaired great cerebral vein (or vein of Galen). The great cerebral vein turns superiorly and joins the inferior sagittal sinus to form the straight sinus. Along its short course, the great vein receives the basal veins of Rosenthal. The basal vein is formed near the optic chiasm by the deep middle cerebral vein, which drains the insula, and several other tributaries that drain inferior portions of the basal ganglia and the orbital surface of the frontal lobe.

II. STROKE

Stroke is characterized by the acute onset of a focal neurological deficit referable to an insult of the cerebral vascular system. Stroke is the third leading cause of death in the United States. There are approximately 500,000 new strokes each year causing 200,000 deaths. Many patients have significant functional impairment and require long-term medical or rehabilitative care, thus adding to the socioeconomic impact of stroke. Risk factors for the development of stroke include hypertension, hyperlipidemia, diabetes, cardiac disease (e.g., atrial fibrillation), vascular disease (such as sickle cell disease), and smoking.

Stroke may be classified into ischemic or hemorrhagic subtypes. Ischemic stroke is far more common, comprising approximately 85% of cases (Table III). Ischemic stroke may be either focal (affecting discrete regions) or global (affecting much of the forebrain with predilection for watershed regions between major vessels). Ischemic stroke may occur secondary to embolic or thrombotic phenomenon. Global ischemia usually results from systemic hypoperfusion as seen with cardiac arrest. Hemorrhagic stroke may be due to

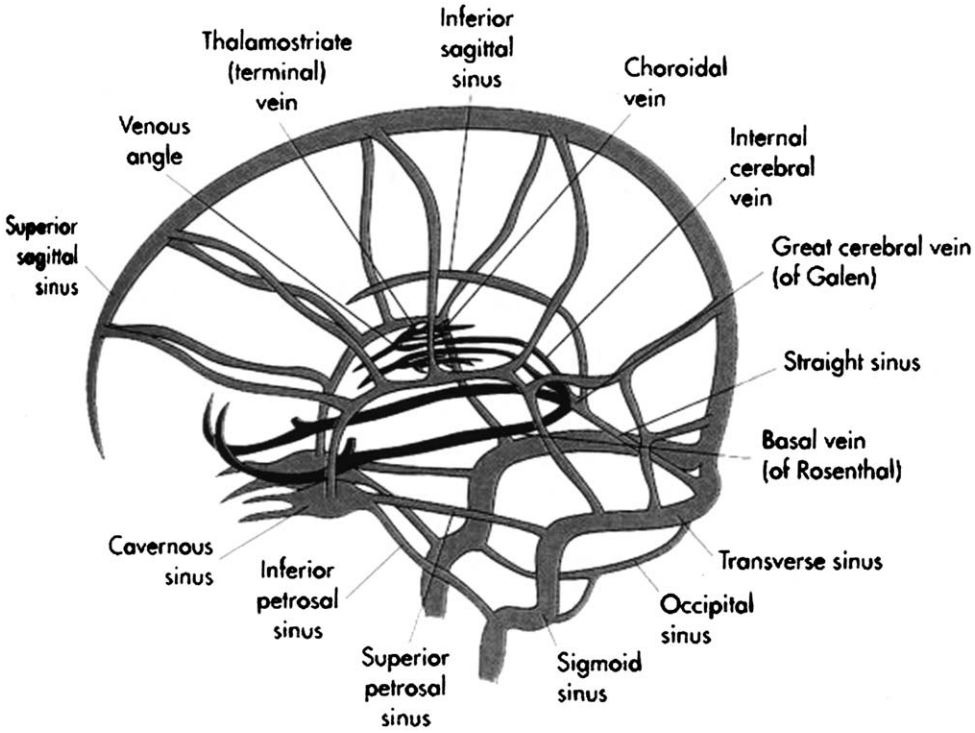


Figure 5 Superficial and deep venous systems. From Nolte (1993), *The Human Brain: An Introduction to Its Functional Anatomy*, 3rd ed. Mosby, St. Louis. Used with permission.

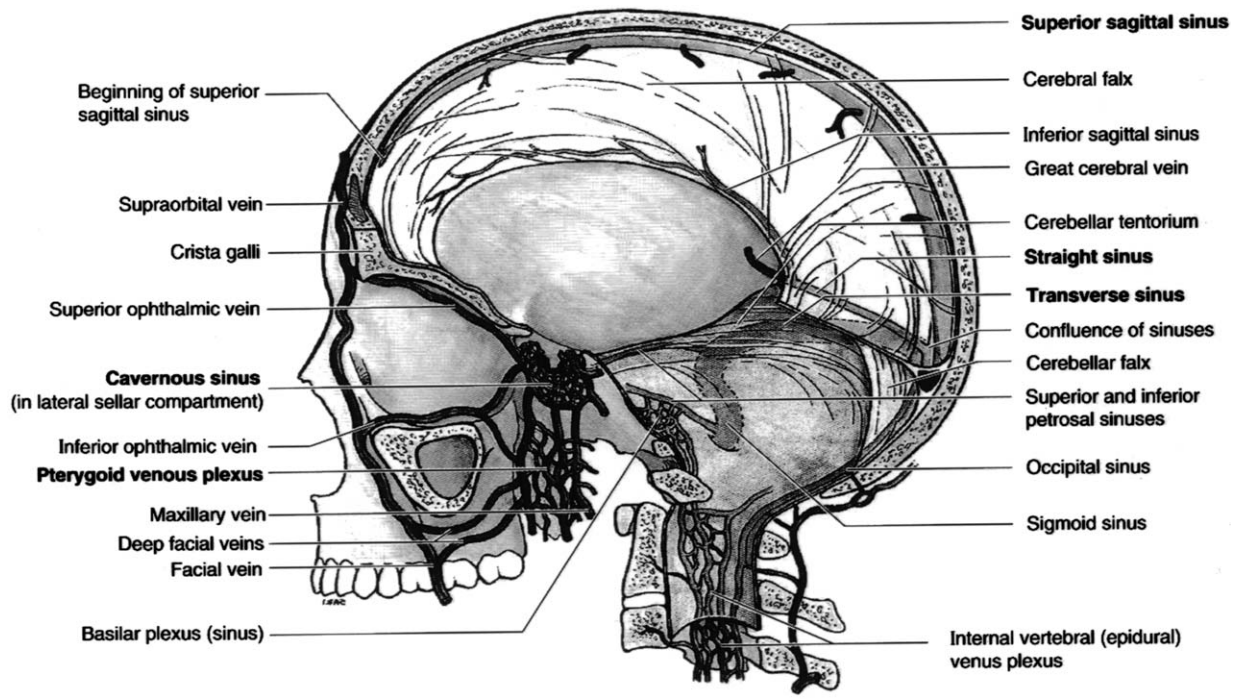


Figure 6 Cerebral venous system. From Nolte (1993), *The Human Brain: An Introduction to Its Functional Anatomy*, 3rd ed. Mosby, St. Louis. Used with permission.

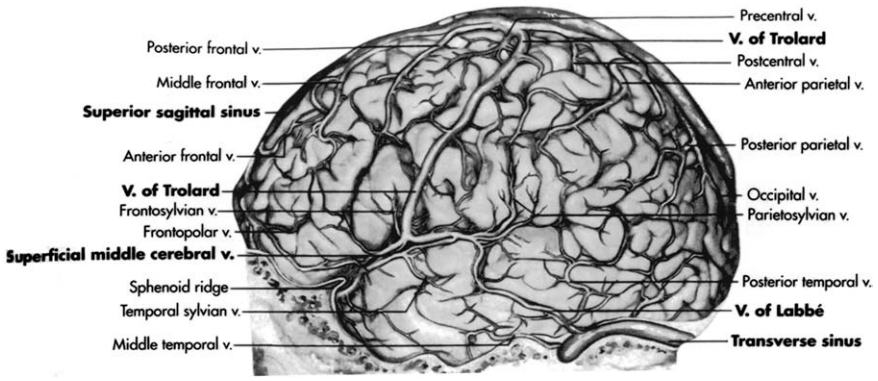


Figure 7 Superficial cerebral veins. (From *Grant's Atlas of Anatomy*, 9th Ed., Copyright © Lippincott Williams & Wilkins.).

Table III
Classification System for Stroke

Ischemic (85%)
Focal
Embolic
Cardiac
Intraarterial
Aortic
Thrombotic
Large artery disease (intracranial and intracranial)
Small penetrating artery disease (lacunar)
Global
Systemic hypoperfusion (e.g., cardiac arrest)
Hemorrhagic (15%)
Subarachnoid hemorrhage
Intracerebral hemorrhage

subarachnoid hemorrhage (SAH) or intracerebral hemorrhage (ICH).

III. ISCHEMIC STROKE

A. Background and Pathogenesis

Ischemia is defined as a reduction in blood flow. Progressive ischemia first disrupts neuronal functioning and later threatens cell viability. Although variability in normal CBF exists, levels below approximately 20 ml/100 g/min result in electrophysiological and functional deficits. CBF levels below 10 ml/100 g/min result in disruption of membrane integrity and cellular death (as seen within an area of completed infarction). Intermediate levels of CBF (between 10 and 20 ml/100 g/min) are seen within the ischemic penumbra, the area of potentially viable tissue immediately surrounding the infarction. Salvaging the potentially amenable ischemic penumbra remains an area of intense research.

B. Patterns of Ischemia

Because cell injury can be reversible depending on the amount of time spent without adequate blood flow and oxygen, three distinct patterns of stroke have been defined. The first pattern is the transient ischemic attack (TIA), which is a transient focal neurological deficit lasting less than 24 hr but usually resolving between 10 min and 4 hr. A common type of TIA, known as amaurosis fugax, causes transient monocular blindness, which is often described as "a shade being pulled" over one eye. The second pattern is the reversible ischemic neurologic defect (RIND). A RIND is defined as a focal neurologic deficit lasting more than 24 hr but completely resolving within 1 week. The final pattern of ischemia is the cerebrovascular accident (CVA). A CVA is a permanent focal neurologic deficit that may improve over time to a limited extent.

C. Pathogenesis

1. Embolic CVA

The most common cause of stroke is the distal embolization of an atherosclerotic plaque within the carotid artery. As early as childhood, a collection of lipid-laden cells known as a fatty streak can be found in the aorta. Over time, and in conjunction with the risk factors mentioned previously, a central core of lipids and cell debris becomes surrounded by a fibrous cap of thickened intima and medial smooth muscle cells. Eventually, this plaque can ulcerate and embolizes to the intracranial circulation (Fig. 8). Also common are emboli from a cardiac mural thrombus in the setting of atrial fibrillation.

2. Thrombotic CVA

Thrombosis may affect both large and small vessels. Thrombosis of large vessels may occur extracranially (e.g., cervical carotid artery occlusion) or intracranially (e.g., MCA occlusion), with resulting systems predictable based on the region served by the affected vessel. Disruption of small penetrating arteries gives rise to lacunar infarction, which typically affects deep brain structures. Lacunar infarctions are thought to arise from chronic hypertension eventually leading to hyalinization and sclerosis of small arteries feeding deep brain neurons.

Figure 8 Carotid artery stenosis: angiogram of left common carotid artery with bifurcation into internal and external branches. Note the area of extreme stenosis of the internal carotid artery.

3. Global Ischemic CVA

Systemic hypoperfusion, as seen during prolonged cardiac arrest, results in diffuse forebrain injury not referable to a discrete vascular territory. Damage is most evident in watershed regions between neighboring vascular territories (e.g., between the ACA and MCA territories).

D. Stroke Syndromes

Ischemia involving individual vascular territories causes specific clinical stroke syndromes. The signs and symptoms that characterize each stroke syndrome can usually predict the specific vessel that is affected. Stroke syndromes have been described for specific regions of the brain and brain stem.

1. Cerebral Hemispheres

Because of its large diameter and high flow, the MCA is most commonly involved in embolic stroke. Occlusion of the superior division of the MCA results in a contralateral hemiparesis of the hand, face, and arm with concurrent sensory deficit. If the dominant cerebral hemisphere is involved, an expressive aphasia may result. The inferior division of the MCA supplies the superior aspect of the optic radiations and the parietal lobe, which regulates higher cortical functions such as spatial thought and recognition of the contralateral body. Thus, infarction results in a contralateral homonymous hemianopsia, combined with astereognosis, agraphesthesia, and contralateral hemineglect with a receptive aphasia if the dominant hemisphere is involved. More proximal lesions of the MCA cause a combination of the previously mention deficits.

Involvement of the ACA is less common and results in a contralateral hemiparesis and hemisensory deficit of the leg. Bilateral ACA infarcts can cause incontinence secondary to the inability to inhibit reflex bladder contractions. A left-sided occlusion of the recurrent artery of Heubner can cause motor aphasia as well.

Occlusion of the posterior cerebral and vertebrobasilar system causes neurologic deficits related to the functions of the occipital and medial temporal lobes, cerebellum, and brain stem. Unilateral occlusion of the PCA results in a contralateral homonymous hemianopsia that is more dense superiorly. Macular vision will often be spared because it is supplied by both the MCA and the PCA. Involvement of the dominant hemisphere results in anomic aphasia (difficulty naming objects), alexia without agraphia (inability to read without impairment of writing), and visual agnosia (inability to recognize objects placed on ipsilateral visual field). Bilateral occlusion of the PCA system can cause cortical blindness and memory impairment.

2. Basal Ganglia and Thalamus

Clinically significant lacunar infarctions typically occur in the internal capsule, thalamus, brain stem, and midbrain. The lesions are small in size but can cause widespread clinical symptoms because of the close proximity of deep brain pathways. As with cortical infarctions, distinct patterns of symptoms are evident and affect the contralateral face, arm, and leg. Lacunar infarcts cause pure motor deficits, whereas infarcts in the thalamus cause pure sensory signs and symptoms. In the internal capsule and thalamus, lacunar infarction typically causes either a pure motor or a pure sensory deficit, respectively, of the contralateral face, arm, and leg.

3. Brain Stem

In the brain stem, infarction may produce an ataxic hemiparesis characterized by contralateral hemiparesis with a cerebellar ataxia of the affected limbs. Midbrain infarction results in an ipsilateral oculomotor nerve palsy with contralateral hemiparesis, also known as Weber's syndrome.

Occlusion of the basilar artery causes neurological deficits referable to dysfunction of the pons and midbrain. With pontine involvement, patients may have hemi- or quadriplegia and constricted pupils, and they will often be comatose. Selective involvement of the dorsal pons will cause additional symptoms of eye bobbing and vertical nystagmus, whereas an infarction of the ventral pons results in the locked-in syndrome, characterized by quadriplegia with intact consciousness. Midbrain involvement results in oculomotor nerve palsy, hemiparesis, quadriparesis, and posturing, usually with impairment in the level of consciousness.

Occlusion of the posterior inferior cerebellar artery results in the lateral medullary syndrome in which a constellation of ipsilateral and contralateral symptoms may occur in various combinations. Ipsilateral symptoms include facial numbness, limb ataxia, Horner's syndrome, hoarseness, and dysphagia. The contralateral body experiences decreased pain and temperature sensation in the arm, trunk, or leg.

E. Treatment

1. Prevention

Treatment of the risk factors includes reduction of blood pressure and cholesterol, control of blood sugar, and cessation of smoking. Since plaque rupture and thrombosis lead to embolism, antiplatelet drugs (i.e., aspirin, ticlopidine, dipyridamole, and clopidrogel), alone or in combination, have been shown to significantly reduce the rate of stroke.

In addition, surgical removal of an atherosclerotic plaque from the carotid artery, known as carotid endarterectomy, has been shown to reduce the incidence of stroke in selected patients. In symptomatic patients with TIA or RIND, carotid endarterectomy has been shown to be more effective in preventing stroke and death than medical management alone in patients with more than 70% stenosis. Although carotid endarterectomy has been shown to reduce stroke rate in asymptomatic patients with severe stenosis, its efficacy against medical management alone and cost-effectiveness remain controversial. Advances in catheter-based technology have made angioplasty of the carotid artery with or without stent placement an option for certain patients unable or unwilling to undergo carotid endarterectomy.

2. Acute Management

Once a stroke has occurred, the immediate objective of treatment is to restore blood flow to the brain. Thrombolytic agents such as recombinant tissue plasminogen activator have been used to break up the occlusive intracranial emboli and restore blood

flow. Thrombolysis has been shown to improve neurologic outcome in patients with stroke if given within 3 hr after the onset of symptoms. However, the risk of intracranial bleeding is increased, thus limiting thrombolysis to patients without conditions predisposed to hemorrhage. Tests to locate the source of emboli, such as echocardiography and carotid ultrasound, are performed. Antiplatelet medicine is started, and anticoagulation with warfarin is begun for patients with clots outside of the carotid system, artificial cardiac valves, atrial myxoma, endocarditits, or atrial fibrillation. After the patient is medically stabilized, rehabilitation at home or specialized facilities can help the patient regain functional ability.

3. New Developments

Many of the developments in stroke treatment have focused on the prevention of embolism and restoration of blood flow after occlusion has occurred. Research has led to advances in preventing or reducing neuronal death after ischemia. The concept of excitotoxicity, by which cell injury and death occur because of excessive stimulation from excitatory amino acids, has provided a new paradigm of cell injury. Agents are being designed to block the receptors through which excitatory amino acids mediate their toxic effects. Other agents may limit injury by reducing neuronal metabolic requirements. These and other future developments may ultimately help to reduce the medical and social impact of this common disease.

IV. HEMORRHAGIC STROKE

Hemorrhagic stroke may result from either intracerebral or subarachnoid hemorrhage. Subarachnoid hemorrhage is most commonly secondary to trauma followed by aneurysmal rupture. Subarachnoid hemorrhage is discussed in the context of cerebral aneurysms. Intracerebral hemorrhage is discussed next.

A. General Considerations

Spontaneous intracerebral hemorrhage (SICH) is defined as hemorrhage in the absence of trauma. It remains a devastating condition that afflicts a significant proportion of the American population. The causes are varied and often associated with other life-threatening pathology. Proper diagnosis requires a thorough knowledge of the possible root causes. Therapeutic goals aim at short-term stabilization and long-term treatment of the inciting diseases.

B. Epidemiology

SICH affects nearly 40,000 Americans annually and is more common among people of Asian and African American descent. SICH in most populations carries a mortality rate of between 20 and 50% in the first month. Morbidity is also high, and only 10–20% live independently following the insult. Chronic hypertension and age are major risk factors. Hypertension is related to the majority of hemorrhages within the deep supratentorial gray matter, cerebellum, and brain stem. Patients presenting with deep-seated SICH tend to be males in their fifth and sixth decades. Younger populations and the elderly more commonly present with lobar hemorrhages.

C. Classification and Etiology

No accepted classification schema exists for SICH. A long list of cerebrovascular disorders are associated with SICH. Hypertensive hemorrhages account for the majority. Other common causes include cerebral amyloid angiopathy, rupture of arteriovenous malformations and aneurysms, postischemic hemorrhage, tumors, and anticoagulation. However, significant overlap exists between these conditions.

The etiology varies with age. Children to middle-aged adults present most commonly with lobar hemorrhages secondary to arteriovenous malformations. Patients in their fifth and sixth decades represent the majority of those presenting with SICH and most suffer from hypertensive hemorrhages. In the elderly population, cerebral amyloid angiopathy is the most common etiology.

In lieu of classification based on the underlying vascular disease or age stratification, ICH is often described based on the location. Hemorrhages can occur above and below the tentorium. Supratentorial hematomas account for the majority of SICH and are classified as lobar or deep (Figs. 9 and 10). Lobar hemorrhages can be secondary to hypertension but are commonly related to vasculopathies (e.g., cerebral amyloid angiopathy and vasculitis), vascular anomalies (e.g., aneurysms and arteriovenous malforma-

Figure 9 Axial noncontrasted CT scan revealing a large left temporoparietal intracerebral hemorrhage.

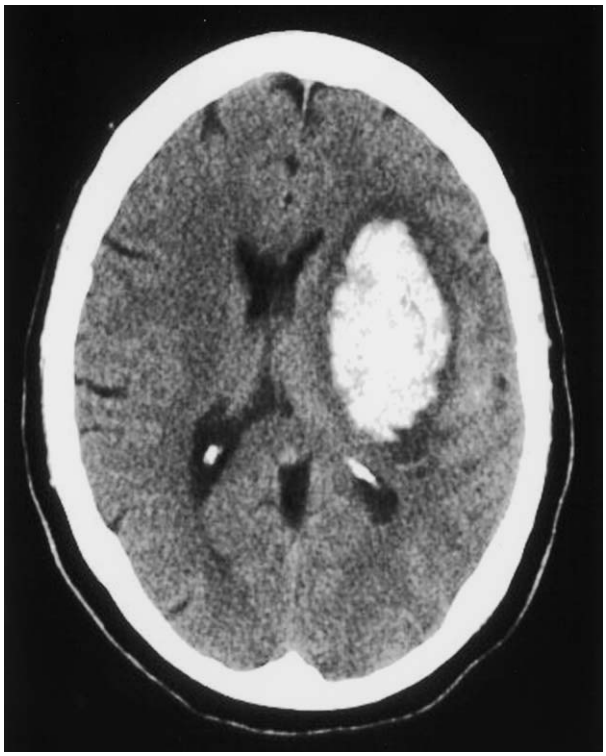


Figure 10 Axial noncontrasted CT scan revealing a typical large hypertensive putaminal intracerebral hemorrhage.

tions), coagulopathy, and tumors. Deep intracerebral hemorrhages are more prevalent than lobar homorrhages and are divided into basal ganglia and thalamic hemorrhages. Infratentorial hematomas are situated either within the brain stem or in the cerebellum.

D. Pathology

The most common etiology of SICH is chronic hypertension, and this population is the focus of this section. The other causes are discussed at length later. Chronic hypertension affects small perforating end arteries deep within the cortex of the brain, especially at the junction between gray and white matter. Lipohyalinosis and microaneurysm formation render the vessels friable and incapable of autoregulation. Abrupt changes in blood flow and blood pressure are thought to incite the hemorrhage. The most commonly affected vessels include the lenticulostriate arteries and perforators of the thalamus and brain stem. Basal ganglia hemorrhages are the most frequent, followed by lobar, thalamic, cerebellar, and pontine hemorrhages.

E. Clinical Presentation

The majority of SICH patients present with a decreased level of consciousness or coma. Other common symptoms include headache, nausea, and vomiting. More focal signs vary depending on location.

1. Supratentorial

a. Lobar Signs and symptoms of lobar hemorrhages depend on the cortical function of the affected region (Fig. 9). Parietal hemorrhages tend to cause contralateral motor and sensor deficits, whereas occipital hemorrhages cause visual disturbances. Dominant hemisphere lesions often disturb language function. Compared to patients with deep supratentorial hemorrhages, these patients are also more often alert and fewer present in coma. Seizure activity is also more commonly experienced in these patients.

b. Deep Deficits from basal ganglia hemorrhages are more often motor than sensory (Fig. 10). In

addition to contralateral hemiparesis or plegia, dominant hemisphere lesions are associated with aphasia. Contralateral neglect is prevalent in nondominant hemisphere hemorrhages. Visual disturbances are also common. Thalamic hemorrhage is also associated with these symptoms, although thalamic ICH typically produces more pronounced sensory findings. Extension of hematoma into the brain stem structures can cause vertical gaze palsy, anisocoria, and nystagmus.

2. Infratentorial

a. Cerebellar Cerebellar hemorrhages present with varying degrees of altered consciousness, headache, ipsilateral gaze palsy, and ipsilateral ataxia. Hemiparesis and hemisensory disturbances are uncommon. If the hematoma extends into the fourth ventricle, patients often present with hydrocephalus. Large hematomas are particularly dangerous because of their propensity to cause herniation and compression of the nearby brain stem. Mortality rates are high for patients who are comatose at presentation.

b. Brain Stem This particularly devastating form of SICH typically occurs in the pons secondary to rupture of the basilar artery pontine perforators. Brain stem hemorrhages cause combined motor and cranial nerve deficits. Patients often experience autonomic dysfunction and dysarthria. Intraventricular hemorrhage into the fourth ventricle is common and results in hydrocephalus.

F. Diagnosis

Patients' past medical histories often reveal the underlying vascular disorder. It is particularly important to pursue evidence of a previous bleeding disorder, malignancy, ischemic strokes, autoimmune disease, and hypertension. Medications should also be reviewed with special attention focused on anticoagulation therapy. A history of drug and alcohol use is also relevant. Certain drugs, such as cocaine and other stimulants, cause severe hypertension that causes SICH. Alcohol abuse can alter liver function and cause coagulopathy.

Laboratory evaluation can also be diagnostic and should include a complete blood count with differential, protime, partial thromboplastin time, serum chemistries, liver function tests, and erythrocyte sedimentation rate (ESR). These studies screen for hematological and coagulation disturbances as well as infectious and inflammatory diseases. The initial radiological test should be computed tomography (CT), which provides rapid and accurate information regarding hemorrhage location, size, and associated mass effect. Little value can be gleaned from magnetic resonance imaging (MRI) or cerebral angiography in elderly patients who have a history of hypertension and who present with a deep intracerebral hemorrhage. Younger patients without a hypertensive history deserve further radiographic evaluation to rule out vascular anomalies, vasculopathy, and tumors. In the remaining patients, MRI and angiography are performed when structural lesions are suspected. Hematoma can obscure the early MRI. To obtain adequate visualization of the affected region, MRI is often repeated after the hematoma has resolved.

G. Treatment

Patients are managed in neurological intensive care units. Control of the blood pressure is of primary importance particularly in the hypertensive patient. Correction of any coagulation disturbance also takes priority. Medical maneuvers to decrease intracranial pressure are performed with caution because hematoma size can actually increase if decompressed as tamponade is diminished. Patients with supratentorial hemorrhages and hydrocephalus are treated with external ventricular drains. Ventricular drains are used with reluctance in cerebellar hemorrhages for fear of causing upward herniation and death. Surgical evacuation is reserved for patients who present in relatively good neurological condition but deteriorate despite medical management, and only if the surgical approach does not require traversing eloquent brain. Hematomas commonly reaccumulate postoperatively and a high index of suspicion is used in these postevacuation patients who subsequently deteriorate. Stereotactic aspiration is another treatment option. This approach benefits from a lower morbidity, but the ability to effectively drain the hematoma has not been clearly established.

H. Summary

Spontaneous intracerebral hemorrhage is a devastating condition that affects a significant number of

Americans annually. Possible causes are protean, but most are secondary to chronic hypertension. Although management of this difficult pathology is primarily supportive, surgical intervention does benefit a subset of patients.

V. ANEURYSMS

A. General Considerations

Intracranial aneurysms represent focal arterial dilations of the cerebral vasculature. They are most commonly situated at major branching points of the intracranial vessels as they navigate through the subarachnoid space. Intracranial aneurysms have a propensity for spontaneous rupture and give rise to SAH. SAH secondary to aneurysmal rupture remains an immense source of morbidity and mortality in contemporary society. Continuing advancements in molecular biology, minimally invasive treatment modalities, and the refinement of microsurgical techniques offer renewed enthusiasm for the treatment of this condition.

B. Classification

Intracranial aneurysms are classified based on pathology, size, and location. Aneurysms may be saccular (berry), fusiform (atherosclerotic or dolechoectatic), miliary (Charcot–Bouchard or microaneurysms related to hypertension), or dissecting (Table IV). Saccular aneurysms comprise the vast majority and are most commonly related to hemodynamically induced degenerative vascular injury arising at the branching points of arteries (developmental/degenerative aneurysms). Saccular aneurysms may also arise from trauma (traumatic aneurysms); infection (mycotic aneurysms); tumors (oncotic aneurysms); high flow lesions, such as arteriovenous malformations (flow-related aneurysms); vasculopathies, such as fibromuscular dysplasia; and drug use such as cocaine.

Aneurysms may also be categorized by size. Aneurysms are denoted as small (<12 mm in diameter), large (12–5 mm), or giant (>25 mm) as described by Dr. Charles G. Drake. Aneurysms may further be described based on location (Table V). Aneurysms are grouped as arising in either the anterior or posterior cerebral circulations and are subsequently subclassified according to the specific vessel of origin.

Table IV
Pathological Classification of Intracranial Aneurysms

Saccular
Developmental/degenerative aneurysms
Traumatic aneurysms
Mycotic aneurysms
Oncotic aneurysms
Flow-related aneurysms
Vasculopathy-related aneurysms
Drug-related aneurysms
Fusiform
Related to atherosclerosis
Miliary
Related to hypertension (Charcot–Bouchard aneurysms)
Dissecting
Spontaneous or related to underlying trauma or vasculopathy

C. Epidemiology

The incidence of intracranial aneurysms varies depending on the source from which the data are gleaned. Large autopsy series indicate that approximately 1–8% of adults harbor intracranial aneurysms. Large proportions of these aneurysms are very small and therefore the prevalence based on autopsy is higher than that detected radiographically. Radiological series derived from cerebral angiographic studies reveal that between 0.5 and 3% of the population is afflicted. Clinical series have confirmed an annual incidence (per 100,000 population) of 6–19. Collectively, these data demonstrate that between 1 and 12 million Americans have intracranial aneurysms, and that approximately 30,000 persons per year suffer SAH.

Aneurysmal SAH has enormous import to society. Aneurysmal rupture is responsible for 0.4–0.6% of all deaths, and despite an overall decrease in stroke rate, the incidence of aneurysmal SAH remains unaltered. Aneurysmal SAH is associated with a mortality rate of more than 25% and a significant morbidity rate exceeding 50%. Furthermore, it is estimated that approximately $1.75 billion and $522 million are spent annually on the care of patients with ruptured and unruptured aneurysms, respectively, in the United States.

SAH from aneurysm rupture most frequently occurs between the ages of 40 and 60. Aneurysms are rarely found in children, which lends credence to the assertion that most aneurysms arise from hemody-

Table V
Intracranial Aneurysms by Location

Location	Incidence (%)
Anterior circulation	
Internal carotid artery	30
Petrous	
Cavernous	
Paraclinoid (ophthalmic)	
Posterior communicating	
Anterior choroidal	
Carotid bifurcation	
Anterior cerebral artery	39
Precommunicating (A1 segment)	
Communicating artery segment	
Postcommunicating segment	
Middle cerebral artery	22
Sphenoidal (lenticulostriate) segment	
Bifurcation segment	
Distal segment	
Posterior circulation (vertebrobasilar)	
Vertebral artery	8
Main trunk	
Posterior Inferior Cerebellar Artery	
Basilar artery	
Anterior Inferior Cerebellar Artery	
Trunk	
Superior Cerebellar Artery	
Basilar bifurcation	
Posterior Cerebral Artery	
Precommunicating segment	
Distal (postcommunicating) segment	

namic vascular injury. Additionally, studies consistently reveal a female preponderance (54–62%).

D. Natural History

Much effort has focused on the natural history of intracranial aneurysms. Traditionally, the risk of rupture for the asymptomatic unruptured aneurysm has been estimated at 1.0–1.4% per year, with cumulative rates of rupture described as 10% at 10 years, 26% at 20 years, and 32% at 30 years. Recently, the rate of rupture of aneurysms less than 10 mm in diameter in patients without a history of other ruptured aneurysms was less than 0.05% per year. The rate for aneurysms less than 10 mm in diameter in those with a history of SAH from a separate aneurysm was 0.5% per year. The annual rupture rate of aneurysms that were 10 mm or larger in diameter was less than 1% per year in those with or without a prior history of SAH from a separate aneurysm.

Ruptured aneurysms have a very high risk of early and late rebleeding. Rebleeding often causes mortal injury and is associated with a 70% mortality rate. The risk of rebleeding is maximal in the first 24 hr (approximately 4%) and diminishes to 1 or 2% per day for the first 2 weeks. After the first month, the rate appears to stabilize at approximately 3% per year.

E. Pathological Characteristics

Aneurysms arise predominantly at the branch points of the major intracranial vessels and rarely occur outside of the cerebral circulation. To explain the observed paucity of extracranial saccular aneurysms, it has been postulated that the intracranial arteries are predisposed as a result of the thinness of their walls (aneurysms have an attenuated tunica media and an absence of an external elastic lamina) combined with minimal perivascular support.

The pathogenesis of cerebral aneurysms is incompletely understood but is likely complex and multifactorial involving a congenital predisposition and superimposed environmental factors. Evidence for a congenital component arises from observations of familial cases and the increased incidence of aneurysms in disorders such as autosomal-dominant polycystic kidney disease, fibromuscular dysplasia, aortic coarctation, and connective tissue disorders such as Marfan's syndrome and Ehlers–Danlos syndrome. In support of an environmental component, data reveal a clear age-related component (increased incidence with increasing age and rarity of aneurysms in children) and the *de novo* development and/or growth of aneurysms after unilateral carotid occlusion. Additionally, cigarette smoking has been consistently reported to confer predisposition to aneurysmal SAH in large series across different populations. Furthermore, injury to the vessel walls by various insults, such as infection, trauma, and neoplastic growth, also gives rise to aneurysm formation.

Histopathological characteristics of aneurysms are well documented. The tunica media is attenuated or absent. Inflammatory and atherosclerotic changes contribute to defects in the tunica elastica and media.

These changes occur primarily at points of marked hypertensive and hemodynamic stress.

As depicted in Table V, aneurysms are most frequently located in the anterior circulation. Regions of particular predilection include the anterior communicating artery complex, the internal carotid artery and posterior communicating artery, or the bifurcation/trifurcation of the middle cerebral artery complex (Fig. 11). Aneurysms arising in the posterior circulation are most frequently found at the BA bifurcation followed by the posterior inferior cerebellar–vertebral artery region (Fig. 12). Multiple intracranial aneurysms are commonly encountered, and data demonstrate that multiple aneurysms are found in 20–30% of patients.

F. Clinical Manifestations

Aneurysms most often present subsequent to SAH. They also may manifest secondary to mass effect (compression of adjacent structures) or cerebral ischemic symptoms, or they may simply be detected as an incidental finding during a test done for other indications.

1. Subarachnoid Hemorrhage

Aneurysmal rupture normally gives rise to blood in and around the basal cisterns in the subarachnoid space (subarachnoid hemorrhage) but may hemorrhage into the ventricular system (intraventricular hemorrhage), brain parenchyma (intracerebral hemorrhage), or, rarely, the subdural space leading to subdural hemorrhage. This rupture presents clinically as the sudden onset of severe headache (often described as the ''worst headache of my life'') and is often associated with signs of meningeal irritation (such as nausea/vomiting, meningismus, photophobia, and phonophobia). Aneurysmal SAH often occurs during straining or exertion, such as with exercise, intercourse, or a bowel movement. SAH may also lead

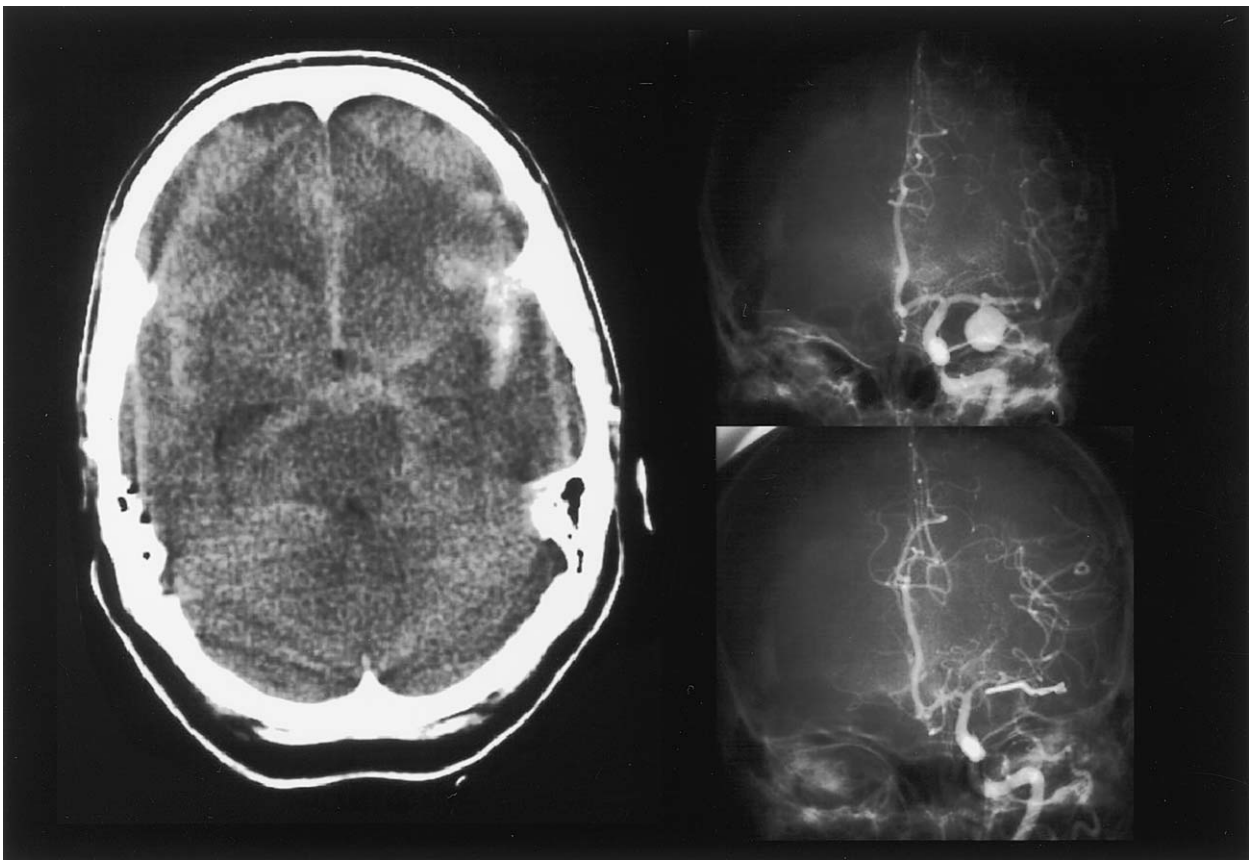


Figure 11 Anterior circulation aneurysm. (Left) Axial noncontrasted CT scan revealing subarachnoid hemorrhage in basal cisterns and Sylvian fissure. Subarachnoid blood predominates in the left Sylvian fissure. (Right, top) Frontal view of angiogram after left ICA injection revealing large MCA bifurcation aneurysm. (Right, bottom) Same view angiogram postoperatively revealing aneurysm clip across aneurysmal neck with no residual aneurysm filling.

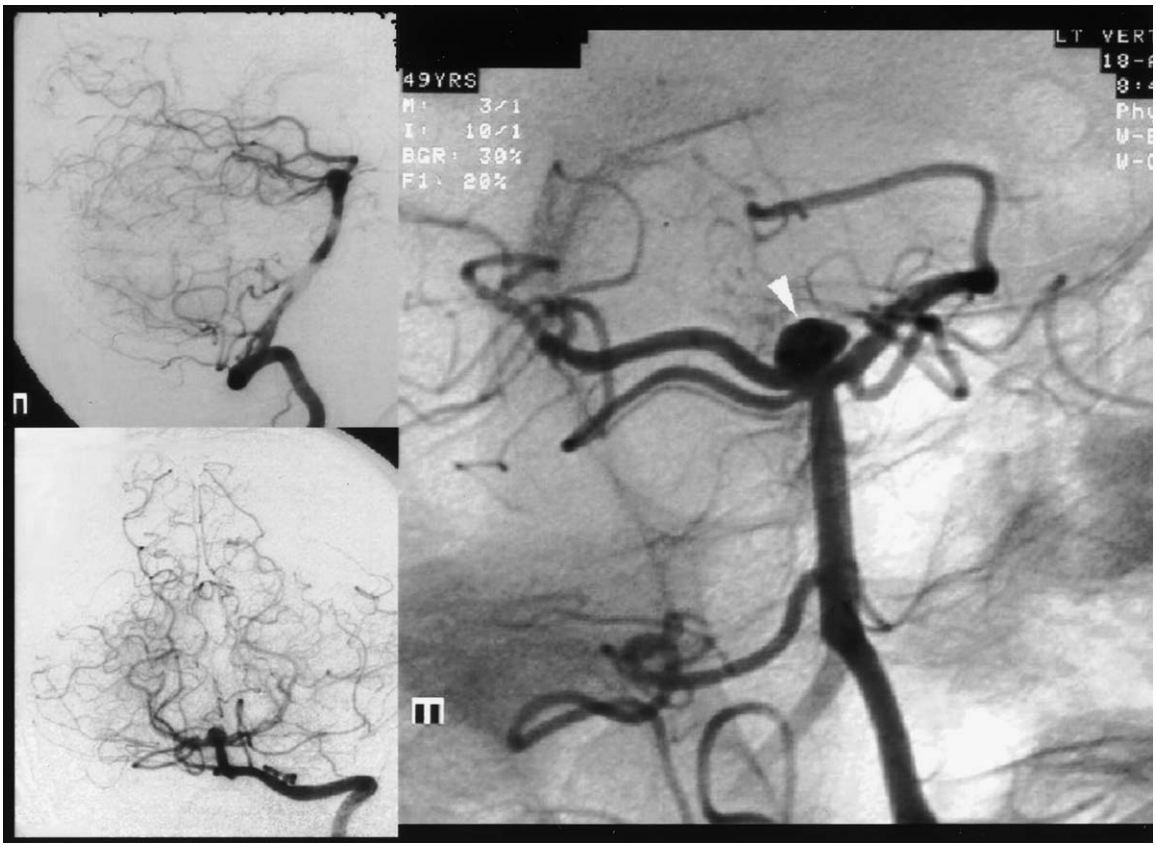


Figure 12 Posterior circulation aneurysm: angiograms after left vertebral artery injection. (Left, top) Lateral projection reveals basilar bifurcation aneurysm. (Left, bottom) Frontal view of same aneurysm. (Right) More superiorly projected frontal view provides optimal view of aneurysm.

to seizures. In more severe cases, SAH may lead to a decreased level of consciousness and various global and focal neurological deficits may be elicited during physical examination. Ophthalmologic examination may reveal unilateral or bilateral subhyaloid hemorrhages in nearly 25% of patients with SAH. These are venous in origin and are situated between the retina and vitreous membrane.

Symptoms preceding the major SAH, such as atypical headaches or neck stiffness, have been ascribed to small hemorrhages and are termed ''sentinel leaks'' or ''warning headaches.'' These symptoms occur in as many as 70% of patients, leading about half of these patients to seek medical attention. Many of these patients are misdiagnosed, which may have future catastrophic repercussions.

SAH may be complicated by cerebral vasospasm and ischemic stroke resulting from the delayed narrowing of intracranial vessels. The incidence of vasospasm begins to increase 3–5 days posthemorrhage, peaks between 7 and 10 days, and diminishes over 2 or 3 weeks. Clinically significant vasospasm occurs in 25–30% of patients, although angiographic evidence is seen in approximately 60–70% of cases.

The most significant predictor of the development of cerebral vasospasm following aneurysmal SAH is the amount and location of subarachnoid blood visualized by CT imaging as popularized by Fischer (Table VI). Approximately 50% of patients with symptomatic vasospasm develop infarction despite therapy. In addition, 15–20% of patients with symptomatic vasospasm will develop a disabling stroke or die of progressive ischemia. Cerebral vasospasm remains the leading treatable cause of death and disability attributed to aneurysmal SAH.

SAH may also be complicated by hydrocephalus. Acute hydrocephalus occurs in 25% of patients with aneurysmal SAH, most often with intraventricular and/or blood in the ambient cisterns. Many patients develop chronic communicating hydrocephalus.

The most important predictor of the outcome of patients suffering from SAH is based on the patient's

Table VI
Fischer Grade Computed Tomographic Scan Classification of Subarachnoid Hemorrhage

Grade	Description
1	No blood detected
2	Diffuse deposition or thin layer of blood, with all vertical layers of blood (interhemispheric fissure, insular cistern, and ambient cistern) <1 mm thick
3	Localized clots or vertical layers of blood 1 mm or more in thickness (or both)
4	Diffuse or no subarachnoid blood but with intracerebral or intraventricular clots

Table VII
Hunt–Hess Clinical Grading Scale for Subarachnoid Hemorrhage

Grade	Clinical conditon
1	Asymptomatic or mild headache and mild nuchal rigidity
2	Cranial nerve palsy, nuchal rigidity, and moderate to severe headache
3	Drowsy, confused, or mild focal deficit
4	Stupor, moderate to severe hemiparesis, and early decerebrate posturing
5	Comatose and decerebrate posturing

Table VIII
World Federation of Neurological Surgeons Clinical Grading Scale for Subarachnoid Hemorrhage

Grade	Glasgow coma scale	Motor deficit
0	15	Absent
1	15	Absent
2	13–14	Absent
3	13–14	Present
4	7–12	Present or absent
5	3–6	Present or absent

condition at presentation. Multiple grading systems have been proposed but the most widely utilized are the Hunt–Hess scale and the scale of the World Federation of Neurological Surgeons (Tables VII and VIII). The World Federation of Neurological Surgeons scale is based on the universally recognized Glasgow coma scale.

2. Mass Effect

Aneurysms, particularly large or giant, may compress adjacent structures and cause neurological deficits. Mass effect commonly causes headache or a third cranial nerve palsy from compression at the junction of the PCoA and the ICA. Other symptoms/signs may include visual field defects, trigeminal neuralgia, brain stem dysfunction, cavernous sinus syndrome (compression of structures traversing the cavernous sinus), seizures, or endocrinologic symptoms due to disruption of the hypothalamic–pituitary axis.

3. Cerebral Ischemia

Embolization of intraaneurysmal thrombus may lead to cerebral ischemic symptoms in the distribution of the cerebral vasculature occupied by the aneurysm. This is a rare presentation of an unruptured aneurysm.

4. Incidental Finding: The Asymptomatic Aneurysm

Widespread adoption of CT and MRI has increasingly led to the discovery of asymptomatic intracranial aneurysms. The discrepancy between the incidence of aneurysms as detected by autopsy and the incidence of SAH apparent in the preceding discussion of epidemiology is resolved by the fact that most aneurysms never rupture.

G. Diagnosis

1. Subarachnoid Hemorrhage

The most common cause of spontaneous (i.e., nontraumatic) SAH is aneurysmal rupture. SAH is first suspected based on clinical presentation. The first diagnostic study performed is usually a noncontrasted CT scan. CT is very sensitive in detecting acute blood and detects between 90 and 95% of patients correctly within 24 hr posthemorrhage. This sensitivity, however, diminishes as blood is rapidly cleared from the subarachnoid space and is approximately 30% sensitive at 2 weeks following hemorrhage. CT can also suggest aneurysm location. For example, blood in the interhemispheric fissure is suggestive of rupture of an ACoA aneurysm, blood in the Sylvian fissure suggests rupture of a MCA aneurysm, and blood in the fourth

ventricle is indicative of a PICA hemorrhage. CT scanning is also used to ascertain the risk of development of cerebral vasospasm. The amount and location of the subarachnoid blood are evaluated according to the Fischer grading system (Table VI).

If CT imaging is negative and the clinical suspicion of SAH is still high, the next diagnostic modality of choice is lumbar puncture. Frank blood in the cerebrospinal fluid (CSF) is not diagnostic and often due to poor lumbar puncture technique. In contrast, xanthochromia (yellow discoloration) of the supernatant which results from the breakdown of blood products within the CSF is diagnostic of SAH. MRI plays a limited role in the diagnosis of SAH and is insensitive in the detection of acute blood. It may be implemented in demonstrating subacute and/or chronic subarachnoid blood after the CT scan normalizes.

2. Intracranial Aneurysms

Three diagnostic modalities are most commonly used to make the diagnosis of intracranial aneurysms: cerebral angiography, MR angiography (MRA), and CT angiography (CTA). Conventional angiography remains the gold standard and provides the most information pertaining to the diagnosis and definitive description of an aneurysm's anatomical configuration (Fig. 11). Conventional angiography, however, is an invasive technique with associated risks, such as stroke, renal failure, and hematoma and/or pseudoaneurysm formation at the puncture site. In contemporary practice, the mortality rate is less than 0.1% and the rate of permanent neurologic injury is approximately 0.5% following cerebral angiography.

MRA is noninvasive and requires no intravascular contrast administration and thereby poses minimal risk to patients. MRA can detect aneurysms as small as 2 mm; however, the critical size for detection from prospective studies is approximately 5 mm. Although MRA may be effectively implemented in the screening for intracranial aneurysms, it is rarely sufficient for adequate surgical planning. CTA has also been the subject of recent interest in the diagnosis of intracranial aneurysms. Helical (spiral) CTA appears to be as sensitive as MRA. It defines the aneurysm's relation to important bony landmarks and is particulary useful when cranial base surgical techniques are being contemplated. CTA is also compatible with older ferromagnetic clips that preclude the use of MRA.

3. Screening for Asymptomatic Aneurysms

Screening for asymptomatic aneurysms in selected patients has been a matter of controversy. A recent large prospective study provided additional guidance in this area. Screening of first-degree relatives of patients suffering aneurysmal SAH was undertaken using MRA. Implementation of a screening program for first-degree relatives of patients with sporadic SAH did not appear to be warranted because the resulting slight increase in life expectancy did not offset the risk of postoperative sequelae.

SAH from aneurysmal rupture is a devastating event. Surgical treatment of unruptured lesions is associated with a relatively low rate of morbidity ($<5\%$) and mortality ($<2\%$) and eliminates the risk of future SAH. Furthermore, evolving minimally invasive strategies, such as endovascular techniques using Guglielmi detachable coils (GDCs) and intracranial stents, offer promise in support of the screening programs and aggressive management of unruptured intracranial aneurysms. Currently, however, widespread screening for asymptomatic intracranial aneurysms does not seem warranted, but future studies may reveal subsets of patients for whom screening has proven efficacy.

H. Treatment of Intracranial Aneurysms

1. General Considerations

Definitive treatment of intracranial aneurysms is aimed at excluding the aneurysm from the cerebral circulation with preservation of the parent artery. Exclusion of the aneurysm from the cerebral circulation, by means of clip placement across the aneurysm neck surgically or with the placement of GDCs endovascularly, theoretically eliminates the risk of future rupture and its associated disability. Alternatively, small intracranial aneurysms, in select locations such as the cavernous sinus, may be observed with caution. Treatment of complications associated with aneurysmal SAH, such as seizures, cerebral vasospasm, and hydrocephalus, is also executed as necessary.

2. Microsurgery

Surgical clipping of intracranial aneurysms remains the definitive treatment and is well validated with proven long-term efficacy (Fig. 11). In experienced

hands, aided by the operating microscope, microinstruments, and microsurgical principles, aneurysm clipping carries relatively low morbidity and mortality directly attributable to surgery. Aneurysms by virtue of their size, anatomical configuration, or location not easily amenable to standard techniques may require specialized adjuncts, including hypothermic circulatory arrest and vascular bypass grafts. Additionally, occlusion of the parent vessel with surgical or endovascular technology may be necessary, with or without concomitant bypass grafting, when clipping cannot be implemented.

The timing of surgery for ruptured intracranial aneurysms remains a matter of controversy. There is general consensus that "good-grade" patients (Hunt–Hess grades 1–3) should undergo early surgery (within the first 48–72 hr) because recurrent hemorrhage is the major cause of death in patients who survive the initial hemorrhage. For "poor-grade" patients and those who are otherwise medically unstable or who bled several days earlier and are experiencing symptomatic vasospasm, surgery is delayed until the patients are stabilized and deemed capable of making a meaningful recovery (or until 10–14 days posthemorrhage in patients experiencing significant vasospasm). However, some centers have advocated early surgery in poor-grade patients and have achieved reasonable results. Some of these patients deemed to be too unstable for surgery may undergo early endovascular treatment.

3. Endovascular Therapy

The 1990s were characterized by the rapid emergence and refinement of endovascular techniques as a minimally invasive means of treating intracranial aneurysms. Endovascular placement of GDCs has been demonstrated to be a safe treatment modality for select patients. Currently, endovascular treatment is utilized predominantly for patients thought to be poor surgical candidates and for aneurysms anticipated to be difficult to treat surgically. GDCs evoke thrombosis within the aneurysmal sac, thereby isolating it from the circulation. Recently, placement of coils concomitantly with intracranial stents to facilitate satisfactory packing and to prevent migration of the coils has been described. Currently, long-term validation of endovascular treatment is not available. With design and implementation of new endovascular techniques and as long-term validation emerges, endovascular therapy is predicted to become an increasingly important treatment modality.

I. Summary

Intracranial aneurysms are a common and potentially devastating disorder of the cerebral circulation. Intracranial aneurysms may be classified by pathological basis, size, or location. The pathogenesis is not completely understood but is likely multifactorial, combining both a genetic predisposition and environmental factors. Aneurysms may present as a result of hemorrhage and its associated complications (such as seizures, vasospasm, and hydrocephalus), from compression of adjacent structures (mass effect), as cerebral ischemia, or may be discovered incidentally. Diagnosis is based on history and physical examination augmented with imaging techniques such as CT or MRI and cerebral angiography. Treatment of intracranial aneurysms endeavors to eliminate the aneurysm from the cerebral circulation while ensuring patency of the parent vessel and surrounding perforating vessels.

VI. VASCULAR MALFORMATIONS

Contemporary understanding and treatment of intracranial vascular malformations have evolved considerably during the past three decades. Advances are attributable to the evolution of neuroimaging, endovascular therapy, microsurgical technique, and stereotactic radiosurgery. Intracranial vascular malformations may be divided into four subcategories: arteriovenous malformations (AVMs), cavernous malformations, venous malformations, and capillary telangiectasias. Rational therapy is contingent on an understanding of the characteristic features of each malformation. A multimodal "team" approach encompasses specialists in microsurgery, radiosurgery, and endovascular therapy and is increasingly required to provide optimal care.

A. Classification

Vascular malformations may be divided into four distinct categories, each with unique anatomical, pathological, radiological, and clinical features. The four traditional subtypes are AVMs, cavernous malformations, venous malformations, and capillary telangiectasias. This classification system is a generally useful template, although there is often considerable overlap between malformations subjected to careful scrutiny.

B. Arteriovenous Malformations

1. Epidemiology and Natural History

AVMs are approximately one-seventh to one-tenth as common as cerebral aneurysms with between 2500 and 3000 new cases presenting each year. An estimated 280,000 patients are afflicted with AVMs in the United States. There appears to be a slight male preponderance between just over 1:1 (male to female) to 2:1.

The natural history of AVMs is incompletely understood. The annual risk of hemorrhage is noted to be 4 and 2% for symptomatic and asymptomatic AVMs, respectively. The average time between initial symptoms (related to hemorrhage or seizures) and rehemorrhage is 8 years. Each hemorrhagic episode appears to be associated with a 20% risk of major neurologic morbidity and a 10% mortality. There is an increased risk of rupture with smaller malformations, AVMs with associated aneurysms, a single draining vein, deep venous drainage, or venous stenosis.

2. Pathological Characteristics

AVMs are congenital lesions consisting of rudimentary arteries feeding directly into veins without intervening capillary or venule networks (Fig. 13). They are normally high-pressure, high-flow malformations with a predilection for damage of surrounding parenchyma upon rupture. Involved veins are "arterialized," becoming dilated, elongated, and tortuous under direct arterial pressure. Vasculopathic changes, such as stenosis or thrombosis of feeding arteries and draining veins, are common. AVMs often assume a conical configuration, with the base located near the cortex and the apex extending to the ventricular system. Gliotic brain tissue may be found encompassed within the malformation. As a consequence of arterial shunting, adjacent brain is subject to poor perfusion and ischemia. Flow-related aneurysms are noted in approximately 8–12% of cases, either on feeding vessels or within the nidus (intranidal). Eight-five percent of AVMs are located supratentorially and 15% are found in the posterior fossa.

3. Clinical Manifestations

AVMs most frequently present as a consequence of hemorrhage in patients in their third and fourth decades. Unlike with rupture of cerebral aneurysms, early rebleeding (within 2 weeks) and cerebral vasospasm are rare because hemorrhage secondary to AVMs is predominantly located intraparenchymally rather than within the subarachnoid space. Neurologic improvement is often seen as the clot regresses and is resorbed, and it provides rationale for delayed surgery 1–4 weeks after hemorrhage when the clot is organized and the brain less friable. Seizures are the second most common presentation of patients with AVMs, being noted in 11–33% of cases. Seizures are thought to be related to gliosis in the adjacent brain resulting from hemosiderin deposition and inflammation. Additionally, patients may present with the new onset of headache (from stretching of the dura, elevated venous pressure, or hydrocephalus) or progressive neurological deficit or cognitive decline (from arterial steal or venous hypertension).

4. Diagnosis

Cerebral angiography remains the gold standard for diagnostic evaluation of putative AVMs (Fig. 13). It provides detailed information regarding the configuration and vascular dynamic properties (such as flow rate, arterial steal/venous hypertension, and collateral flow). MRI and CT imaging are also useful adjuncts. Based on lesion size, location, and pattern of venous drainage, operative risk is assigned to a given patient's AVM via the Spetzler–Martin classification scale (Table IX).

5. Treatment

Comprehensive management of patients harboring AVMs involves three main therapeutic modalities: endovascular therapy, microsurgery, and stereotactic radiosurgery. Endovascular therapy, applying catheter-administered materials for embolization, is a useful adjunct to microsurgery and stereotactic radiosurgery to diminish the degree of arterial shunt. Embolization is rarely curative alone, however. Within 1 week following embolization (to prevent recanalization of embolized vessels), microsurgery may be performed to excise the lesion. Surgical risk is ascertained from the Spetzler–Martin scale preoperatively. In younger patients possessing superficial lesion in noneloquent areas, microsurgery is clearly the treatment of choice (Fig. 13). Microsurgery is also more effective than radiosurgery in ameliorating symptoms of intractable epilepsy and headaches. Stereotactic radiosurgery is reserved for patients with small (<3 cm), unruptured AVMs in eloquent brain substance with deep venous drainage. Stereotactic radiosurgery invokes vascular injury and produces delayed thrombosis after months

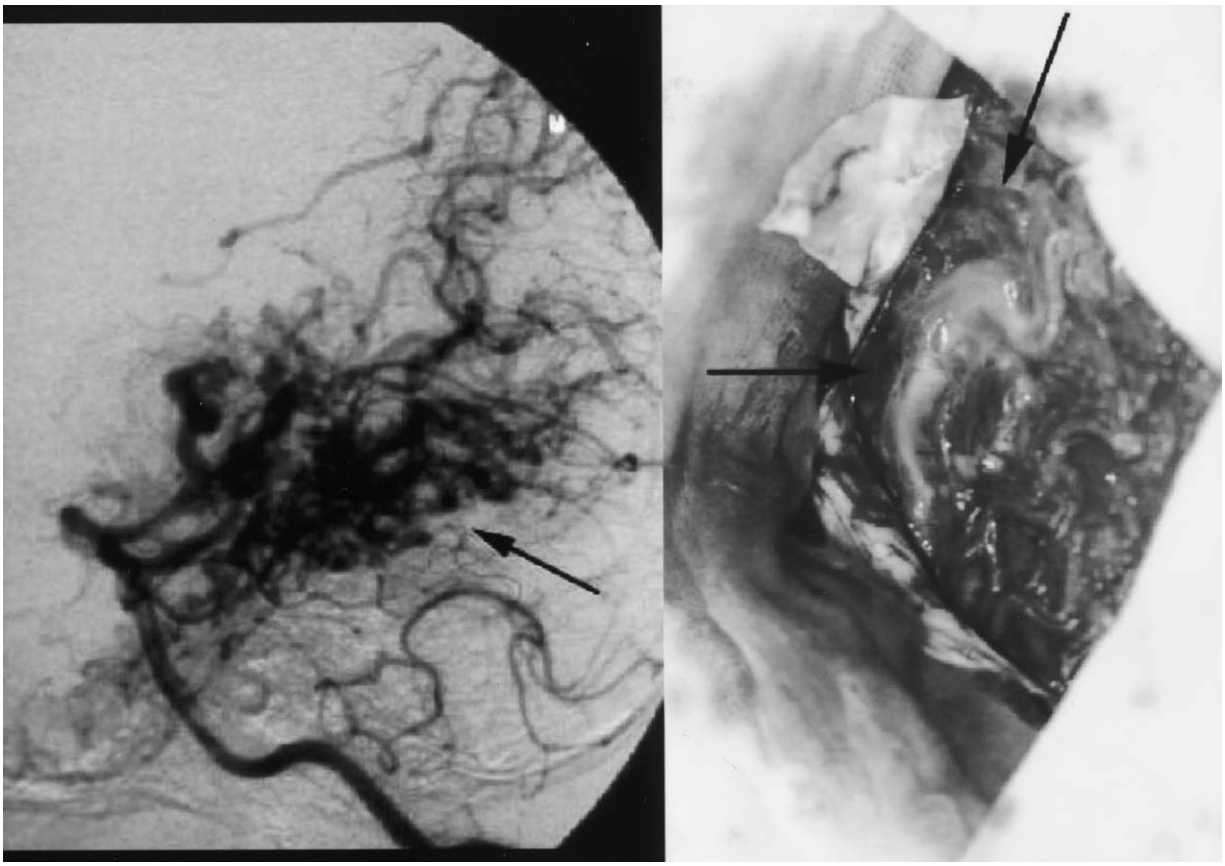


Figure 13 Arteriovenous malformation. (Left) Lateral view of angiogram after left vertebral injection. Arrow indicates AVM. (Right) Intraoperative photograph shows the same AVM immediately evident after durotomy.

Table IX
Spetzler–Martin Classification Scale for Operative Risk in Patients with AVMs[a]

Characteristic	Points assigned
Size	
Diameter <3 cm	1
Diameter 3–6 cm	2
Diameter >6 cm	3
Location	
Noneloquent site	0
Eloquent site (sensorimotor, language, visual, hypothalamus or thalamus, internal capsule, brain stem, cerebellar peduncles or nuclei)	1
Pattern of venous drainage	
Superficial only	0
Any deep	1

[a]Scores of 4 or 5 are associated with the greatest risk of persistent postoperative neurological deficits.

to years. Thus, unlike surgery, protection from hemorrhage imparted from radiosurgery is delayed. Radiosurgery's benefits diminish and its adverse effects (damage to adjacent parenchyma) increase as the size of the lesion increases. Radiosurgery may be a viable option for the treatment of deep, residual AVMs after attempted microsurgical resection.

C. Cavernous Malformations

1. Epidemiology and Natural History

Cavernous malformations (or cavernous angiomas) affect both sexes equally. Estimates of lesion prevalence vary between 0.02 and 0.9% depending on the criteria and methods utilized for definition. Patients typically present in the second to fourth decades of life. Familial forms, characterized by multiple lesions and an autosomal-dominant inheritance pattern, are well described and prevalent in the Hispanic population.

The natural history of cavernous malformations is not well defined. Cavernous malformations are known to be dynamic lesions that change in size and imaging characteristics with time. New lesions are also known to develop over time. Subclinical microhemorrhages are thought to be very common. Previously asymptomatic lesions have a 1% annual incidence of hemorrhage. Once symptomatic, cavernomas rupture at an annual rate of 4.5%. Hemorrhage occurs commonly in the familial forms, deeply located lesions, lesions with associated venous malformations, in pregnant patients, and in patients who have received whole brain irradiation.

2. Pathological Characteristics

Cavernous malformations represent a lobulated collection of dilated endothelial-lined sinusoidal spaces. There is normally no intervening brain parenchyma. They are low-flow lesions that expand by internal thrombosis and hemorrhage within the sinusoidal spaces. Hemorrhage in various stages may be found within the lesion. Neighboring brain is well demarcated by hemosiderin stained, gliotic borders. The walls of the malformation lack smooth muscle and elastic lamina and are thus not true arteries or veins. Venous malformations are found in association with cavernous malformations in 15% of cases. Eighty percent of lesions are supratentorial but may occur at any location. Multiple lesions are demonstrated in 50–80% of cases.

3. Clinical Manifestations

Seizures are the most common manifestation of supratentorial cavernous malformations. Surrounding hemosiderin and gliosis is thought to be responsible for this epileptogenicity. Symptomatic hemorrhage and headache are the next most common presentions. Cavernous malformations are also frequently found incidentally in radiographic studies performed for other indications.

4. Diagnosis

Cavernous malformations cannot be visualized with angiography and are best delineated by MRI. With MRI, cavernous malformations appear as a central focus of mixed signal intensity representing hemorrhage of various stages (''popcorn-like'') surrounded by a hypointense rim of hemosiderin from multiple microhemorrhages (Fig. 14).

5. Treatment

Cavernous malformations that are asymptomatic are not generally treated because of their low risk of hemorrhage ($\sim$1%/year). Microsurgical resection is indicated for symptomatic supratentorial lesions. Brain stem cavernous malformations are considered for surgical resection when there is repeat hemorrhage, progressive neurological deficit, and superficial location. Stereotactic radiosurgery remains an experimental treatment option.

D. Venous Malformations

Venous malformations are composed of histologically normal veins that radially converge on a single draining vein with interposed normal brain (as opposed to cavernous malformations with no interposed normal brain). This pattern has been described as a ''spoke wheel'' or ''caput medusae'' and is best seen in the venous phase of angiographic studies (Fig. 15). There is no arterial inflow. Venous malformations are located within the white matter commonly adjacent to an ependymal surface. Venous malformations are clinically silent and are considered benign. If a hemorrhage or seizure occurs, an associated AVM or cavernous malformation should be suspected.

E. Capillary Telangiectasias

Capillary telangiectasias consist of nests of dilated capillaries with interposed normal brain substance. They are clinically silent lesions and rarely exceed 1 cm in size. They are most frequently seen in the pons and are detected only in contrast-enhanced MRI studies. Capillary telangiectasias may represent a precursor or transitional form of cavernous malformations.

F. Summary

Vascular malformations represent an important disorder of the intracranial vasculature. Vascular malformations may be categorized as arteriovenous malformations, cavernous malformations, venous malformations, and capillary telangiectasias. Each subtype possesses its own unique natural history, clinical manifestations, and imaging characteristics. Different treatment paradigms have been established for each malformation. A multimodal approach to

Figure 14 Cavernous malformation: T1-weighted axial MRI, with arrows marking typical brain stem cavernoma with "popcorn" appearance.

treatment encompasses the disciplines of microsurgery, endovascular therapy, and stereotactic radiosurgery. Effective treatment of the most complex vascular malformations is being increasingly realized.

VII. DURAL ARTERIOVENOUS FISTULAE

Dural arteriovenous fistulae (DAVF) are arteriovenous shunts involving the dura and epidural space. DAVF are located within the dural wall of a venous sinus rather than within the sinus. DAVF are most commonly situated in the transverse and sigmoid sinuses followed by the cavernous sinus. In approximately 7% of cases, multiple DAVF may be encountered. They account for 10–15% of all intracranial AVMs and may be of congenital or acquired origins.

It is hypothesized that acquired DAVF may result from antecedent venous stenosis, thrombosis, or occlusion of a dural venous sinus. With recanalization, normally present microscopic arteriovenous shunts may enlarge, forming multiple microfistulae and giving rise to a DAVF (DAVF have microfistulae and not nidi *per se*). Angiogenesis has also been proposed to play a role in the development of AVF.

DAVF are encountered across all age groups including children. Most symptomatic lesions, however, are seen between the fifth and seventh decades. Clinical signs and symptoms are variable and commonly include headache and bruit (with involvement of the transverse and/or sigmoid sinuses), retroorbital pain, proptosis, chemosis, and ophthalmoplegia (cavernous sinus lesions). Other less common manifestations include cranial nerve palsies, tinnitus, and dementia.

A preponderance of DAVF behave benignly and may even spontaneously regress. In contrast, some DAVF present with hemorrhage or neurological decline. Clinical presentation, neurological status, and long-term prognosis of those individuals harboring DAVF are most closely governed by the pattern of venous drainage.

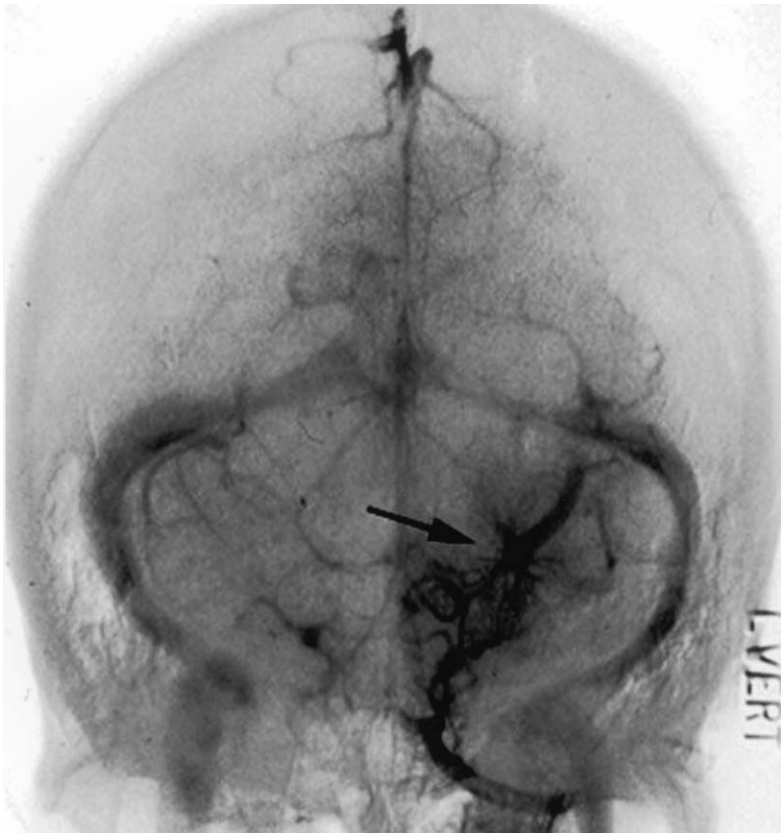


Figure 15 Venous angioma: frontal view of venous phase angiogram with arrow indicating the caput medusa appearance.

Classification schemes have been proposed by Borden and colleagues and Cognard and colleagues. In the Borden scheme, type I DAVF have dural venous sinus outflow only, type II DAVF have dural venous sinus outflow with retrograde leptomeningeal (or cortical) venous drainage, and type III DAVF have retrograde leptomeningeal (or cortical) venous drainage only. Type I DAVF are typically benign, and types II and III are increasingly associated with hemorrhage and aggressive clinical courses.

CT and MRI may help to reveal the lesion (and detect hemorrhage if present). Ultimately, cerebral angiography is necessary to accurately define the anatomic and hemodynamic characteristics.

Treatment is undertaken after careful review of the clinical and radiological characteristics in each case. Treatment options include conservative management (careful observation), microsurgery, radiosurgery, endovascular therapy (transarterial or transvenous embolization), or combinations thereof. The availability of long-term follow-up data will help to validate current treatment paradigms.

VIII. VASCULOPATHIES

Vasculopathies represent a broad category of disorders involving blood vessels. The most common are the vasculitides, moyamoya disease, fibromuscular dysplasia, and cerebral amyloid angiopathy.

A. Vasculitides

1. General Considerations

The vasculitides are a diverse group of infrequent disorders characterized by underlying inflammation and necrosis of the blood vessel wall. They are heterogeneous with respect to presentation, distribution, size of the affected vessel, and underlying etiology. Central nervous system (CNS) vasculitis is a potentially serious disorder with symptoms arising from ischemia or hemorrhage. Rapid diagnosis and treatment are prerequisite to favorably altering the disease course.

2. Classification

The vasculitides are generally divided into primary and secondary disorders (Table X). The primary vasculitides are subdivided based on the size of the vessel affected (i.e., large, medium, or small). Large-vessel vasculitides include temporal arteritis, Takayasu's arteritis, and primary angiitis of the CNS. Polyarteritis nodosa and Kawasaki disease typify the medium-vessel vasculitides. The small-vessel vasculitides include Churg–Strauss syndrome and Wegener's granulomatosis. The secondary vasculitides represent the most frequent cause of CNS vasculitis. They are a varied group and result from underlying autoimmune diseases, intoxications, and infectious and neoplastic processes.

Table X
Classification of Vasculitides Affecting the Central Nervous System

Primary vasculitides	Secondary vasculitides
Large vessel	Autoimmune diseases
Temporal arteritis	Systemic lupus erythematosis
Takayasu's arteritis	Sjogren's syndrome
Primary angiitis of the central nervous system	Behcet's syndrome
Medium vessel	Toxic
Polyarteritis nodosa	Infectious
Kawasaki's arteritis	Neoplastic
Small vessel	
Churg–Strauss syndrome	
Wegener's granulomatosis	

3. Pathogenesis

The pathogenesis of vasculitis is complex and not well understood. There is increasing evidence that antineutrophil cytoplasmic antibodies (ANCAs) play a role in the immunopathogenesis of the vasculitides by activating neutrophils and thereby causing endothelial injury. Various immunological mechanisms have been proposed, but ultimately inflammation causes leukocytes to adhere to the vessel wall. The inflammation is usually segmental in nature, with skip lesions of intense inflammation interrupting otherwise normal vasculature. One or more layers of the vessel wall can be affected. The inflammation tends to resolve, leaving fibrosis and hypertrophy, which cause secondary occlusion.

4. Clinical Presentation

Vasculitis can present with CNS symptoms and signs alone or combined with nonspecific or multiorgan symptoms. The sequelae depend on the number, size, and site of the involved blood vessels. Focal injuries tend to cause wall rupture and hemorrhage. Segmental injuries affect the entire wall circumference and tend to cause stenosis, occlusion, and infarction. Collectively, patients can present with cerebral ischemia, encephalopathy, intracerebral hemorrhage, or seizures.

5. Diagnosis

The diagnosis of vasculitis can be difficult. A complete history and physical examination, augmented with appropriate laboratory tests, are necessary. A complete blood count should be performed and markers of inflammation, such as ESR and C-reactive protein, will be elevated. Other laboratory tests that can help diagnose the cause include liver and renal function tests, complement levels, cryoglobulin levels, Rh factor, anti-nuclear antibody, c- and p-ANCA, serologies for hepatitis B and C, lupus anticoagulant, and anti-cardiolipin antibody. A brain meningeal biopsy is the gold standard in diagnosis and confirmation.

Imaging studies can also facilitate the diagnosis. MRI and angiography may detect CNS lesions but there are no pathognomonic MRI findings. CT and MRI changes will occur in areas of ischemia or hemorrhage. MRI is more sensitive for identifying small foci and for detecting multiple CNS lesions. MR angiography may show narrowing of the larger intracranial vessels but standard angiography is necessary to delineate many of the subtle changes characteristic of vasculitis (Fig. 16). These include multiple arterial occlusions, especially of the small vessels over the convexity, segmental stenosis of intracranial vessels sometimes separated by dilations causing a beaded appearance, and intracerebral aneurysms. However, none of these findings are specific to vasculitis.

6. Treatment

Treatment varies with the specific condition. After the specific syndrome is identified, the inciting agents are avoided and the underlying condition is treated. Steroid administration is the mainstay of treatment and can be guided by the size of the involved vessels. Large-vessel involvement responds well to corticosteroids alone, whereas small- and medium-sized vasculitides respond better to a combination of cyclophosphamide and steroids. Steroids should be

decreased to the lowest effective dosage to minimize side effects. Low-dose methotrexate, V immunoglobulin, or plasmapheresis can be used in difficult or select cases. Hypertension should be treated along with any underlying conditions. Surrogate markers of inflammation, such as erythrocyte sedimentation rate (ESR), C-reactive protein (CRP), antinuclear antibody (ANA), and antinutrophil cytoplasmic antibody (ANCA), can be used to follow the course of the disease.

7. Summary

The vasculitides are a clinically and pathologically heterogeneous group of disorders characterized by inflammation of the blood vessels that may represent a primary or secondary disease process. Additional insight into the pathoetiology will enhance our ability to effectively diagnose and treat these disorders.

B. Moyamoya Disease

1. General Considerations

Moyamoya disease is a rare progressive occlusive cerebrovascular disorder. It is characterized by progressive occlusion of the distal portions of the internal carotid arteries and proximal portions of the anterior and middle cerebral arteries (Fig. 17). A compensatory network of small collateral vessels form in the adjacent areas at the base of the brain producing the characteristic appearance of a puff of smoke (moyamoya is Japanese for "wavering puff of smoke"). Symptoms are due to cerebral ischemia and/or hemorrhage and can occur on alternating sides because of the bilateral nature of the disease. Although some patients present with only headache or visual disturbance, symptoms can include serious neurological deficits such as hemiparesis, motor aphasia, seizure, and mental alteration.

2. Epidemiology

Moyamoya is particularly prevalent in Japan and Korea, where 10% of patients have a family history of the disease. This increased incidence in Asian populations and familial occurrence has also led to the hypothesis of HLA gene involvement. The disease is more prominent in females, with a ratio of 1:1.8. It is rare, with a reported incidence of 0.35 per 100,000 people, but it is probably underdiagnosed because it is

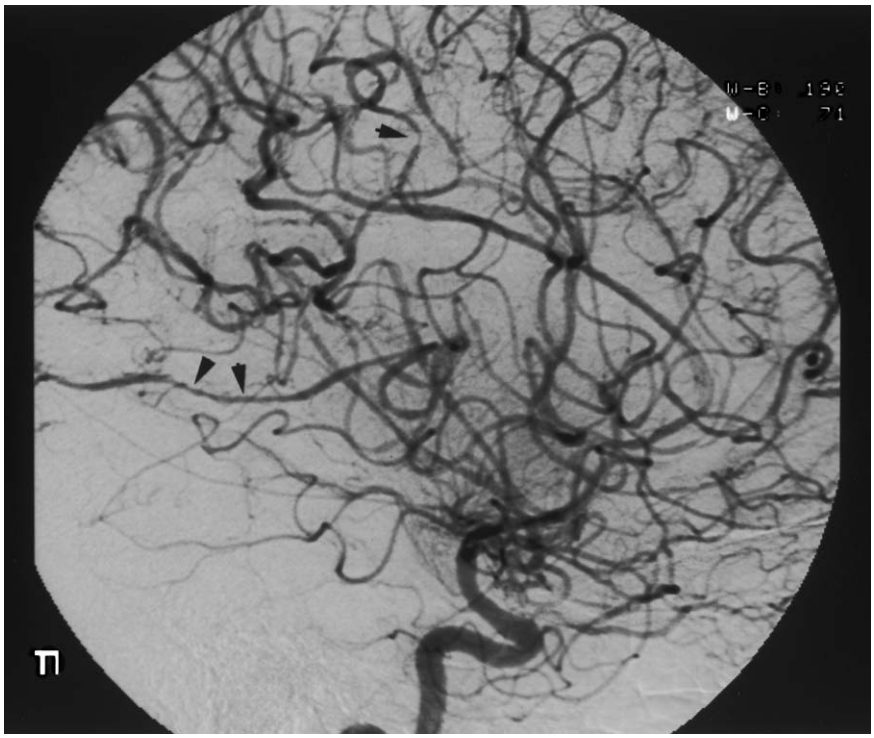


Figure 16 Cerebral angiogram of a patient with vasculitis revealing multiple focal areas of abnormal stenosis followed by dilatation throughout the brain.

often asymptomatic. The age distribution is bimodal, with the first peak occurring between the ages of 10 and 14 and the second between ages 40 and 49.

3. Pathogenesis

Histopathological changes occur in the terminal ICA and proximal ACA and MCA. Vessel walls display multilayered intimal fibrous thickening and markedly wavy, often duplicated or triplicated, internal elastic lamina. The thickened intima contains an increased number of smooth muscle cells but no inflammatory infiltrate, atherosclerosis, or fibrinoid necrosis. The media is usually atrophic. Varying degrees of intimal thickening are evident.

Moyamoya is an entity of unknown etiology. Both a congenital component (due to polygenic inheritance) and an acquired component (due to trauma, infection, brain tumor, radiation, or autoimmune disease) have been proposed. Recently, aberrations of vascular growth factors and cytokines have been implicated. Basic fibroblast growth factor (b-FGF) has been reported to be significantly higher in the CSF of patients with moyamoya disease than in patients with arteriosclerotic cerebrovascular occlusive disease. Increased activity of b-FGF and b-FGF receptors has also been noted in the superficial temporal artery wall.

4. Clinical Features

The clinical characteristics vary depending on the age of diagnosis. Among pediatric patients the initial symptoms are mainly due to cerebral ischemia (TIAs) and are recurrent and progressive. Their symptoms can be induced by crying, coughing, straining, or hyperventilation. In most children, the vascular obstructions rapidly progress and well-developed moyamoya vessels are noted within 2 or 3 years of the first attack. Whereas only 10% of children suffer intracranial hemorrhage, the incidence exceeds 60% in adults. Hemorrhage most often occurs in the basal ganglia, thalamus, or ventricle. Vascular obstructions in adults progress slowly if at all.

5. Diagnosis

The diagnostic criteria of the disease are based mainly on angiographic findings and the following guidelines have been proposed by the Ministry Health Welfare Japan:

- Stenosis or occlusion of the intracranial internal carotid artery or the adjacent ACAs and MCA
- Abnormal vascular network adjacent to the stenosed artery identified during the arterial phase of angiography

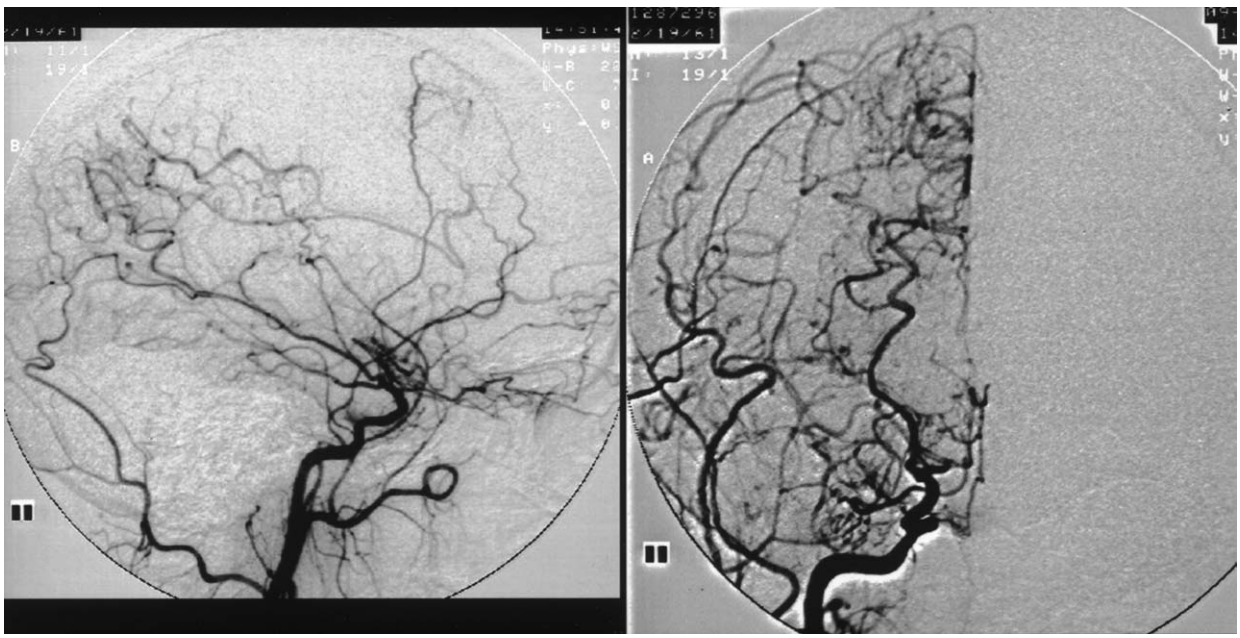


Figure 17 Moyamoya: AP and lateral cerebral angiogram revealing right middle cerebral artery occlusion with prominent collateral lenticulostriate and leptomeningeal vessels giving a characteristic "puff of smoke" appearance.

- Bilateral findings on angiography
- No other identifiable cause

The angiographic findings must be present bilaterally for definite diagnosis but many believe unilateral lesions progress to become bilateral. MRI and MRA have made the diagnosis of moyamoya possible without exposing patients to the invasiveness of conventional angiography. Common MR findings include multiple dilated abnormal vessels at the basal ganglia and/or thalamus, narrowing or occlusion of major arteries of the circle of Willis, parenchymal changes including ischemic infarctions predominantly in watershed areas, intracranial aneurysms (especially in the posterior circulation and moyamoya vessels), and intracerebral hemorrhages. Electroencephalograph (EEG) can also help in diagnosis of children with moyamoya disease. In more than half the cases a characteristic EEG finding termed ''rebuildup'' can be found. The rebuildup is considered to be related to decreased perfusion reserve of the ischemic brain.

6. Treatment

The hemodynamics of moyamoya disease are characterized by low CBF with extremely high vascular resistance in the collaterals at the base of the brain. Revascularization has been shown to be useful in preventing further ischemic attacks but does not prevent rebleeding in the hemorrhagic group of adults with moyamoya disease. Several surgical revascularization procedures are used in moyamoya disease to decrease hemodynamic stress and thereby reduce moyamoya vessel prominence and increase collateral circulation. These include direct, indirect, and combined revascularization techniques. Direct bypass can be accomplished with superficial temporal artery to MCA anastomoses. Indirect bypasses include omentum transplantation, encephalo-galeo-synangiosis, encephalo-myo-synangiosis, encephalo-duro-arterio-synangiosis, encephalo-myo-arterio-synangiosis, and encephalo-duro-arterio-myo-synangiosis. Surgical revascularization is controversial in adult moyamoya disease because it does not necessarily prevent rebleeding.

C. Other Vasculopathies

1. Fibromuscular Dysplasia

Fibromuscular dysplasia (FMD) is a rare vascular disorder of unknown etiology that primarily affects branches of the aorta. It is the second most common cause of extracranial carotid stenosis. The renal artery is often affected as well. Although the specific genetic defect is not known, it has been suggested that FMD is inherited as an autosomal-dominant trait with incomplete penetrance in males.

FMD affects the cervical arteries and occurs mainly in women. Fifty percent of patients present with multiple, recurrent symptoms due to cerebral ischemia or infarction. FMD has been reported to cause ''spontaneous'' dissection in young adults and subarachnoid hemorrhage. Twenty to 50% of patients with FMD are found to have intracranial aneurysms and patients may be at higher risk for carotid dissection, arterial rupture, and carotid cavernous fistula. The most common complaint from patients suffering from FMD is headaches, which are commonly unilateral and may be mistaken for migraines.

Diagnosis is made with angiography. The most common finding on angiogram is the ''string of beads'' sign due to multiple rings of segmental narrowing alternating with dilation in the carotid and vertebral arteries representing fibrodysplasia of the media (Fig. 18). Less commonly, focal tubular stenosis or diverticular outpouchings of the arterial wall can be found.

Anticoagulation or antiplatelet therapy may be helpful in the treatment of FMD. Surgical management is hampered by the often difficult access to, and friability of, the vessels involved. Transluminal angioplasty may also be effective.

2. Cerebral Amyloid Angiopathy

Cerebral amyloid angiopathy (CAA) is a condition characterized by pathological deposition of β-amyloid protein within the media of cerebral and leptomeningeal vessels. The most common presentation is dementia. Severe CAA is associated with vasculopathic changes, vessel rupture, and cerebral hemorrhage. Approximately 10% of patients develop ICH and patients may present with recurrent lobar hemorrhages or a TIA-like prodrome.

The white matter of the brain is most commonly involved and CAA does not affect areas outside of the CNS. There is increased incidence with age, and it is present in nearly 50% of people over the age of 70. It is a part of a family of conditions characterized by amyloid deposition, including Alzheimer's disease. Many believe similar genetic factors are involved (such as the apolipoprotein E ε_2 and ε_4 alleles).

Diagnosis of CAA may be challenging because of the difficulty in distinguishing it from other causes of lobar hemorrhage. Contrast angiography is usually

normal. Gradient-echo MRI may identify multiple petechial hemorrhages and hemosiderin deposits, but this is not specific to the disease. The definitive diagnosis of CAA requires a brain tissue biopsy revealing amyloid deposition. This can be difficult to obtain due to the sometimes sparse and segmental nature of the disorder. Additionally, β-amyloid protein is present in a relatively large percentage of the population, which may further confuse the diagnosis. Patients may develop fibrinoid necrosis of vessel walls, and this is a more specific finding and good indicator of cases at risk for hemorrhage. CAA is often diagnosed in retrospect at autopsy.

The disease is largely untreatable, although there is increasing use of anticoagulant and thrombolytic therapies to prevent recurrent hemorrhage. Current studies are examining techniques to block the molecular and cellular steps involved in the pathogenesis of the disease.

IX. CEREBRAL BYPASS PROCEDURES

Bypass grafts are employed to circumvent restrictions to cerebral blood flow. The more commonly employed conduits include saphenous veins, superficial temporal arteries, and occipital arteries. Bypass procedures may route blood from the extracranial circulation (carotid or vertebral) to the intracranial circulation [extracranial–intracranial (EC–IC) bypass] or directly reconstruct the intracranial circulation (intracranial bypass). Using these procedures, neurosurgeons successfully treat complex vascular lesions that would otherwise result in ischemic morbidity.

Cerebral bypass procedures have been incorporated in the management of patients with difficult aneurysms, neoplastic disease, and traumatic arterial dissection and in patients with ischemic disease refractory to medical therapy. Vascular neurosurgeons employ bypass techniques to redirect flow around giant aneurysms that cannot be directly clipped and require parent vessel occlusion. Cerebral bypass has also been employed in the successful management of symptomatic traumatic internal carotid artery dissection. Skull base surgeons use bypass techniques when faced with tumors encasing large cerebral arteries whose removal requires vessel sacrifice. Despite the negative results of the Cooperative Study of Extracranial–Intracranial Arterial Anastomosis, with appropriate preoperative selection, patients with focal intracranial vascular disease may still benefit from EC–IC bypass. This is particularly true for those patients who have been screened with ancillary tests, such as xenon CT with and without acetozolamide challenge, documenting inadequate cerebrovascular reserve and collateral flow.

Figure 18 Right carotid angiogram revealing extensive fibromuscular dysplasia with a characteristic "string of beads" pattern or alternating zones of luminal dilatation and constriction (arrow).

X. CONCLUSION

The cerebral circulation is eloquently complex and serves as host to many vascular disorders of immense societal significance. Stroke, aneurysms, AVMs, and vasculopathies are potentially devastating cerebrovascular disorders. Advances in our understanding of the molecular and cellular features of cerebrovascular disease, coupled with refinements in multimodal therapies, promise to reduce the enormous personal and societal burden of these disorders.

See Also the Following Articles

BRAIN ANATOMY AND NETWORKS • CEREBROVASCULAR DISEASE • CHEMICAL NEUROANATOMY • FUNCTIONAL MAGNETIC RESONANCE IMAGING (fMRI) • STROKE • VENTRICULAR SYSTEM

Suggested Reading

Batjer, H. H. (1996). *Cerebrovascular Disease*. Lippincott–Raven, Philadelphia.

Drake, C. G., Peerless, S. J., and Hernesniemi, J. A. (1996). *Surgery of Vertebrobasilar Aneurysms. London, Ontario Experience on 1767 Patients*. Springer–Wien, New York.

Executive Committee for the Asymptomatic Carotid Atherosclerosis Study (1995). Endarterectomy for asymptomatic carotid artery stenosis. *J. Am. Med. Assoc.* **273,** 1421–1428.

Ferro, J. M. (1998). Vasculitis of the central nervous system. *J. Neurol.* **245,** 766–776.

Gray, H. (1995). *Gray's Anatomy: The Anatomical Basis of Medicine and Surgery*, 38th ed. Churchill–Livingston, New York.

Greenberg, S. M. (1998). Cerebral amyloid angiopathy: Prospects for clinical diagnosis and treatment. *Neurology* **51,** 690–691.

Grossman, R. G., and Loftus, C. M. (1999). *Principles of Neurosurgery*, 2nd ed. Lippincott–Raven, Philadelphia.

Heimer, L. (1995). *Human Brain and Spinal Cord: Functional Neuroanatomy and Dissection Guide*, 2nd ed. Springer-Verlag, New York.

The International Study of Unruptured Intracranial Aneurysm Investigators (1998). Unruptured intracranial aneurysms—Risk of rupture and risks of surgical intervention. *N. Engl. J. Med.* **339,** 1725–1733.

The Magnetic Resonance Angiography Study in Relatives of Patients with Subarachnoid Study Group (1999). Risks and benefits of screening for intracranial aneurysms in first-degree relatives of patients with sporadic subarachnoid hemorrhage. *N. Engl. J. Med.* **341,** 1344–1350.

Moore, P. M., and Richardson, B. (1998). Neurology of the vasculitides and connective tissue diseases. *J. Neurol. Neurosurg. Psychiatr.* **65,** 10–22.

Natori, Y., Ikezaki, K., Matsushima, T., and Fukui, M. (1997). "Angiographic moyamoya": Its definition, classification, and therapy. *Clin. Neurol. Neurosurg.* **99**(Suppl. 2), 168–172.

North American Symptomatic Carotid Endarterectomy Trial Collaborators (1991). Beneficial effect of carotid endarterectomy in symptomatic patients with high-grade stenosis. *N. Engl. J. Med.* **325,** 445–453.

Osborn, A. G. (1999). *Diagnostic Cerebral Angiography*, 2nd ed. Lippincott Williams & Wilkins, Philadelphia.

Schievink, W. I. (1997). Intracranial aneurysms. *N. Engl. J. Med.* **336,** 28–40.

Cerebral Cortex

DAVID F. CECHETTO and JANE C. TOPOLOVEC
University of Western Ontario

GLOSSARY

agnosia The inability to recognize objects by touch or sight (tactile and visual agnosia) or lack of recognition of what is heard (auditory agnosia).

apraxia The inability to carry out purposeful movements even though there is no muscular weakness.

cortical column Sensory cortical areas are organized in units of neurons from the surface of the cortex to the underlying white matter. Each unit of neurons responds selectively to a specific afferent stimulus.

corticofugal Projections from the cerebral cortex.

corticopetal Projections to the cerebral cortex.

functional neuroimaging Techniques such as positron emission tomography and functional magnetic resonance imaging that utilize cerebral metabolism and blood flow as indicators of activity in specific regions of the brain.

receptive field The receptive field of a neuron is the area in the periphery (area of skin or location in the visual field) that, when stimulated maximally, excites or inhibits that cell.

The cerebral cortex is an extensive layer of gray matter that covers the superior surface of the brain. It receives sensory information from the internal and external environments of the organism, processes this information and then decides on and carries out the response to it. Different regions of the cerebral cortex are specialized for specific functions, such as somatic sensory and motor, visceral sensory and motor, and integrative cognitive functions. All of this information fits within the many grooves and convolutions of the cortex. These folds increase the surface area of the brain while enabling it to fit in the small volume of the skull.

The bulk of the cerebral cortex is comprised of the neocortex. The phylogenetically older parts of the cortex include the paleocortex (olfactory cortex, entorhinal and periamygdaloid areas) and the archicortex (the hippocampal formation). The neocortex is the primary focus of this article.

I. CYTOARCHITECTURE AND NEUROCHEMISTRY

A. Overview

The cerebral cortex is composed of an enormous number of neurons (10^{10}). These neurons in the human are responsible for the high-order cognitive processing or the conscious mind. It is in the cerebral cortex that the sensory signals generated in the body terminate. These sensory inputs are then processed at several different levels and the integrated information is used in the generation of specific actions. In order to accomplish this highest order of processing, the cerebral cortex is composed of a relatively thin layer

of cells and fibers, from 2 to 4 mm thick, that have multiple local interconnections. To receive information regarding the external and internal milieu and to generate commands to control the muscles and organs, the cerebral cortex has both direct and indirect connections with all other regions of the brain and spinal cord. The surface area of the cortex is extensive (approximately 250,000 mm^2) due to the numerous folds in the cortical sheath and it accounts for approximately 40–50% of the mass of the brain. The folds permit the cerebral cortex to contain a large number of neurons and to develop many specialized regions that represent various parts of the body.

B. Cortical Layers

Although there are regional variations, the cerebral cortex is generally organized into six distinct layers based on cell types and the organization of the myelinated axon fibers when examined in stained sections (Fig. 1). The outermost layer, known as the *molecular* layer, is a fiber- and synapse-rich layer with few cell bodies. The cells in layer 1 are primarily neuroglial or nonpyramidal neurons. This layer contains many axons that run parallel to the surface of the cerebral cortex. Layer 2 is known as the *external granular* layer and contains primarily small pyramidal neurons with their apical dendrites directed toward the

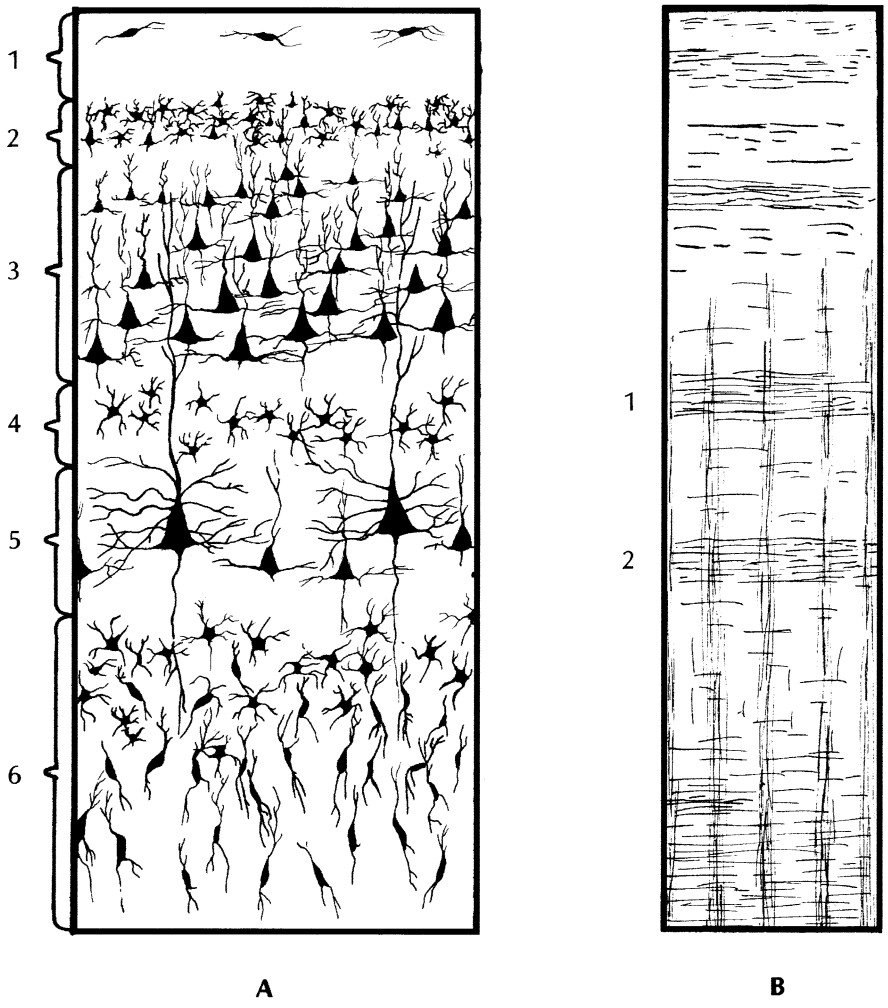


Figure 1 Histology of the six layers of the cerebral cortex illustrating the types and arrangements of neurons (A) and myelinated nerve fibers (B) Numbers 1–6 in A indicate the layers of the cerebral cortex. Numbers 1 and 2 in B indicate the bands of Baillarger (reproduced with permission from J. A. Kiernan, *Barr's, the Human Nervous System*. Lippincott-Raven, Philadelphia, 1998).

surface of the brain. The third layer of cells is quite thick, containing many pyramidal neurons, and its name, the *external pyramidal* layer, is derived from these cells. These pyramidal neurons are of medium size and the size increases toward the deepest part of layer 3. The apices of these pyramidal neurons also extend toward the surface of the brain. The pyramidal neurons in this layer form connections with other regions of the cerebral cortex, including the opposite (contralateral) side of the brain. The *internal granular layer*, or layer 4, contains numerous small cells that are densely packed together. Some of these small cells are stellate (star-shaped), whereas others are small pyramidal cells. As will be discussed later, this layer is particularly prominent in regions of the cortex dedicated to receiving sensory information. The fifth layer of the cerebral cortex, the *internal pyramidal* (or *ganglionic*) layer, contains both medium-sized and large pyramidal neurons that, like other pyramidal cells, have their dendritic apices extending toward the exterior surface of the brain. In the region of the motor cortex some of these cells are very large and are known as the pyramidal cells of Betz. Neurons of layer 5 give rise to the bulk of the axons, corticofugal fibers, projecting to subcortical regions. Sometimes, layers 4 and 5 have a condensation of horizontal fibers that are called the bands of Baillarger. These horizontal fibers are seen in myelin-stained tissue (Fig. 1B). The deepest and final layer 6, the *multiform* (or *polymorph*) layer, is composed of small neurons that can be stellate, pyramidal, or another type of cell with elongated cell bodies called fusiform neurons. This layer also gives rise to corticofugal fibers terminating primarily in the thalamus. Beginning in layer 3 and through layer 6, the pyramidal neurons send axons internally that form bundles of myelinated fibers. As each layer containing pyramidal neurons contributes to these bundles they become progressively thicker.

C. Regional Differences in Cytoarchitecture

Based on microscopic differences in the organization of the six layers, multiple distinct regions of the cerebral cortex can be defined. One of the best known maps of the cytoarchitecture of the cerebral cortex is that of Brodmann, which lists 52 different areas. Although there are subtle differences that are used to distinguish these many areas, the most simplified point of view is to consider only two major types of neocortex. The first is *agranular* cortex in which there is a relative lack of the granular layers 2 and 4, which contain the small, densely packed cells, whereas the third and fifth cortical layers, containing the larger pyramidal cells, are very well developed. This type of cortical layering is most frequently associated with motor regions of the cerebral cortex. The second, or *granular*, type of cortex is characterized by well-developed granular layers 2 and 4 with numerous small, stellate cells. This type of cortical region is usually associated with those areas receiving major inputs from the sensory relay nuclei of the thalamus. In particular, the axons conveying sensory inputs from the thalamus have large terminal plexi in the granular layer 4.

It is the pyramidal cells that send efferent projections from the cerebral cortex to subcortical areas as well as to other regions of the cortex on the same or the contralateral side. The pyramidal neurons of layer 6 are primarily responsible for the cortical outputs to the thalamic nuclei, which in turn send axonal projections back to the same cortical region, terminating in a different layer. However, the pyramidal neurons of layer 5 send projections to other subcortical sites, primarily in the midbrain, brain stem, and spinal cord, whereas the primary source of efferents to cortical regions is layer 3.

D. Neurochemistry

Several classes of neurotransmitters have been identified in the cerebral cortex, including excitatory amino acids, γ-aminobutyric acid (GABA), neuroactive peptides, and monoamines. The excitatory amino acids, glutamate and aspartate, are the primary neurotransmitters of cortical projections. They are quick acting and potent to enable fast transfer of information with high fidelity. These neurotransmitters act primarily on *N*-methyl-D-aspartate (NMDA), α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA), kainate, L-ammonophosphonobutyric acid, and metabotropic receptors, all of which are more prevalent in sensory than motor regions.

GABA is localized in a wide variety of nonpyramidal cells and is often colocalized with neuroactive peptides. It acts on $GABA_A$ and $GABA_B$ receptors to regulate information within small regions throughout the cortex. GABA is the primary inhibitory neurotransmitter of the central nervous system.

Cholecystokinin, vasoactive intestinal polypeptide, neuropeptide Y, somatostatin, substance P, and corticotrophin-releasing factor are all considered to

be neuroactive peptides that influence neurotransmission by modulating the effects of excitatory amino acids and GABA. Moreover, some of these peptides may mediate their effects by altering blood flow and metabolism in local regions of the brain.

The fourth class of chemical mediators found in the cerebral cortex are monoamines and include noradrenaline, serotonin [or 5-hydroxytryptamine (5-HT)], and dopamine. These monoamines are the primary source of input to the cortex from extrathalamic sources and have different patterns of distribution throughout the cortex. Noradrenergic axons (mostly originating from neurons in the locus coeruleus in the pons) innervate primary motor and somatosensory areas and the nearby regions in the frontal and parietal lobes. Noradrenaline acts on α_1, α_2, β_1, and β_2 receptors, which are more prominent in prefrontal, motor, and somatosensory regions. Serotonergic fibers originate primarily from the raphe nuclei and the midbrain tegmentum of the brain stem. They innervate diffuse areas of the cerebral cortex to act on the 5-HT (5-HT_1–5-HT_7) receptor. Dopamine fibers preferentially innervate motor and association cortical areas to modulate the connections between association areas and descending motor systems. These fibers, which act on the dopamine receptor (D_1 and D_2 receptor subtypes), originate in the substantia nigra and ventral tegmental area of the midbrain.

Another class of neurotransmitter is acetylcholine. Cholinergic fibers, which originate in the basal nucleus of Meynert and diagonal band of Broca, innervate diffuse cortical areas, including limbic, primary sensorimotor, and association areas. Acetylcholine acts on muscarinic and nicotinic receptors mainly on pyramidal but also on nonpyramidal cells. The monoamines and acetycholine are not necessarily excitatory or inhibitory; they act to modulate and enhance the inputs from the thalamus and other cortical areas.

II. INTRINSIC CIRCUITS

A variety of techniques have proven useful in delineating the intrinsic organization of cortical circuits. In particular, the recent developments in histochemical methods coupled with intracellular recordings and dye labeling have enabled investigators to unravel the neuronal organization of the local circuitry of the cortex (Fig. 2). As indicated previously, the layers of the cortex can be characterized by the arrangement and density of stellate and pyramidal neurons. The

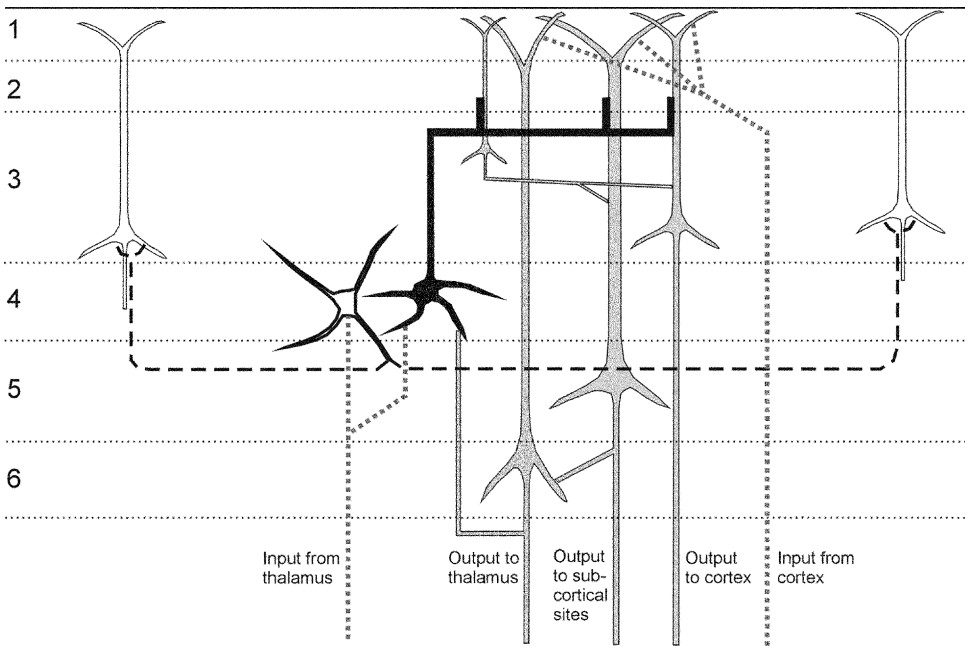


Figure 2 Intrinsic circuitry of the layers of the cerebral cortex. Indicated are inhibitory projections (dashed, heavy line) from aspiny inhibitory neurons (open cell with heavy lines), pyramidal neurons in gray, a stelate neuron in black and pyramidal neurons in adjacent columns (open cells with thin lines.).

pyramidal neurons are responsible for the output of the cerebral cortex, are excitatory in nature, and utilize the excitatory amino acids, aspartate and glutamate, as neurotransmitters.

These projection neurons have apical dendrites that ascend to the outer layers of the cortex and often form apical tufts that receive connections within layers 1 and 2. Other dendrites project from the bases of pyramidal neurons. The stellate neurons, like the pyramidal cells, have spiny dendrites and are excitatory. These spiny stellate cells, primarily seen in layer 4, are local circuit neurons and the axons do not project out of the cortical layers.

Another type of cell that is abundant within the layers of the cerebral cortex is the local circuit neurons with aspiny dendrites. These smooth neurons are inhibitory in nature and have axons that in the neocortex are restricted to a discrete region of the cortical layers. Unlike the stellate and pyramidal neurons, the axons of these neurons can arise from any site on the cell body. The neurotransmitter of these neurons is GABA and these cells often colocalize neuropeptides such as somatostatin, substance P, vasoactive intestinal polypeptide, or cholecystokinin.

The primary corticopetal fibers to the cerebral cortex are from the sensory and motor nuclei of the thalamus. However, there are many other sources for cortical inputs. Basal forebrain nuclei contain acetylcholine neurons with extensive projections to the cerebral cortex. The reduction in this forebrain cholinergic innervation is an important contributor to Alzheimer's disease. The noradrenergic neurons of the locus coeruleus in the pons diffusely innervate the cerebral cortex and appear to play a role in modulating the signal-to-noise ratio of specific cortical inputs. The raphe nuclei of the brain stem are the source of a serotonergic input, which is primarily inhibitory. Finally, the cortex has extensive interconnections that can be from the opposite hemisphere, from adjacent gyri of the same lobe, and from other lobes.

A typical local circuitry in sensory cortex is shown in Fig. 2. The thalamocortical input primarily provides an excitatory termination on the intrinsic stellate neurons in layer 4 and to some extent in layer 6. In fact, in sensory cortex, small columns of neurons are formed such that the cells in the column selectively receive input from a specific sensory stimulus. Inputs from other regions of the cortex or other subcortical sites terminate on pyramidal neuron dendrites in layers 1–3. The stellate neurons in layer 4, receiving the thalamocortical input, terminate on apical dendrites of pyramidal cells in layers 2 and 3. These pyramidal neurons in layers 2 and 3 project back onto layer 5 pyramidal cells, which then project back to pyramidal cells of layers 6, which in turn project back to layer 4, completing a local circuit in the cortical column. In addition, in layer 4, the pyramidal neurons will excite aspiny inhibitory neurons that send an inhibitory impulse to adjacent columns of the cortex.

III. TOPOGRAPHY

A. Sulci, Fissures, and Lobes

In addition to the cytoarchitectonic method of creating divisions, the extensive folding of the cerebral cortex also provides a means of identifying distinct regions in the cerebral cortex (Fig. 3). It is this extensive folding that greatly increases the surface area and thus the volume of the cerebral cortex. Although the folds delineate some of the functional areas of the cortex, the borders between the cytoarchitectonic divisions and the major and minor folds do not necessarily correspond.

The two deepest grooves on the brain are known as fissures. The longitudinal fissure separates the right and left cerebral hemispheres, and the transverse fissure separates the cerebral hemispheres from the cerebellum. The smaller and plentiful grooves on the cerebral hemispheres are known as sulci, whereas the folds or ridges of brain between the sulci are called gyri. Some of the sulci and gyri are relatively consistent in all individuals and are used as anatomical landmarks to identify separate regions of the cerebral cortex. The cerebral cortex is divided into five lobes: the frontal, parietal, occipital, temporal, and insular lobes. The central sulcus (rolandic sulcus) can be seen on the superior surface of the human brain just slightly behind the midpoint between the anterior and posterior poles. This sulcus then descends inferiorly and anteriorly over the lateral surface of the cerebral hemisphere and separates the frontal and parietal lobes. The central sulcus serves as the demarcation between motor and sensory cortices. Immediately anterior to the central sulcus in the frontal lobe lies the precentral gyrus, which is used as a landmark for the primary motor area. The primary somatosensory area is located in the postcentral gyrus, found immediately posterior to the central sulcus in the parietal lobe. The parietooccipital sulcus is located on the medial surface of the cerebral hemisphere and, as the term implies, it is the major landmark demarcating the

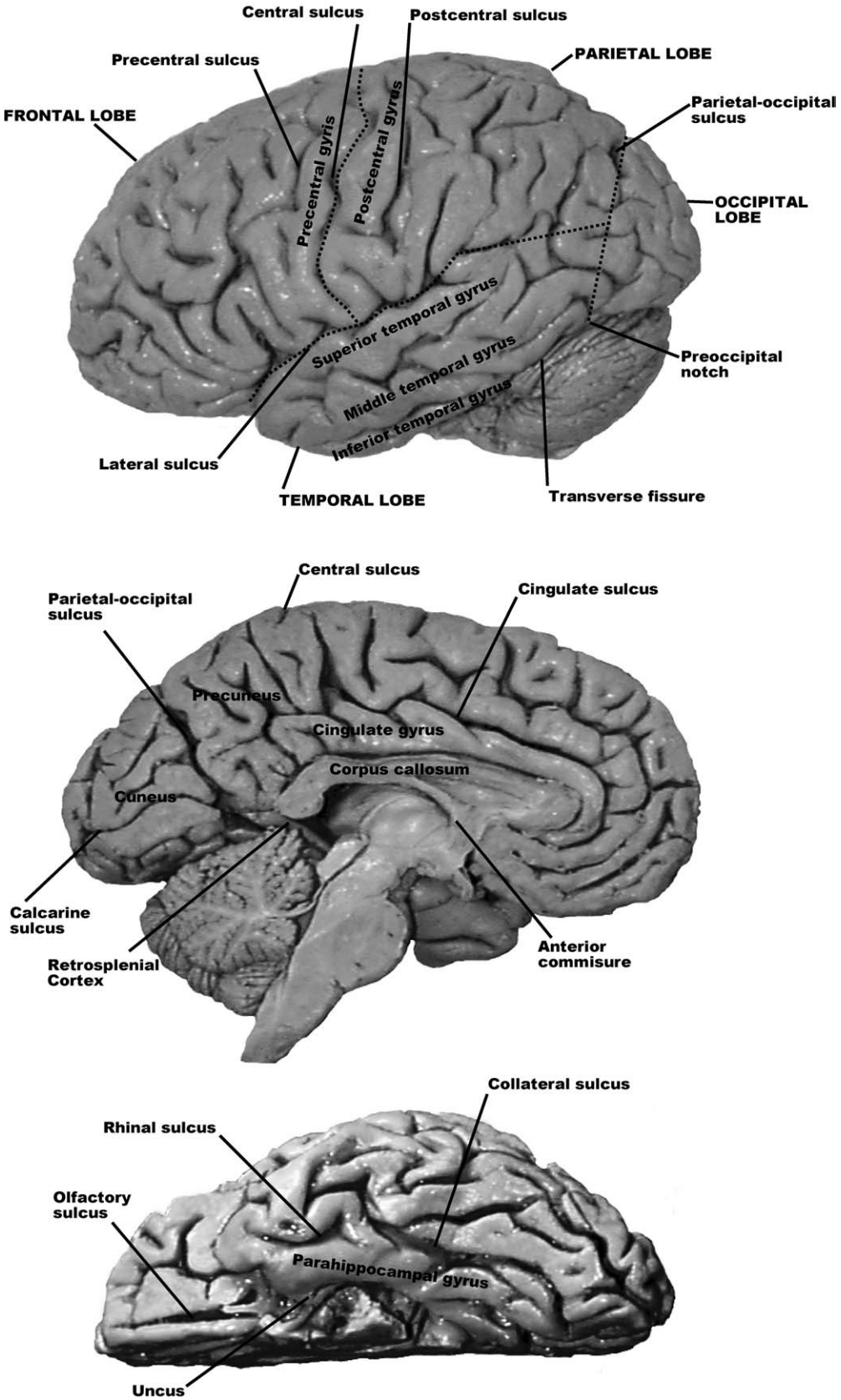


Figure 3 Gyri and sulci on the lateral (top), medial (middle), and inferior (bottom) surfaces of the left cerebral hemisphere.

parietal from the occipital lobes. The parietooccipital sulcus starts at the superior surface of the human brain and extends inferiorly along the medial surface until it merges with the calcarine sulcus, which starts at the most posterior part of the cerebral cortex (the occipital pole) and extends anteriorly along the medial surface of the hemisphere until it meets the parietal–occipital sulcus. The primary visual cortex is located along the calcarine sulcus. On the lateral surface, the occipital lobe is defined by a line joining the parietal–occipital sulcus and the preoccipital notch.

The lateral sulcus (Sylvian fissure) is a very deep fold seen on the lateral surface of the hemisphere running in an anterior to posterior direction and serves to separate the temporal lobe from the frontal and parietal lobes. On the lateral surface of the temporal lobe are three gyri—the superior, middle, and inferior temporal gyri—which are separated by the superior and inferior temporal sulci. The primary auditory cortex is located on the superior surface of the superior temporal gyrus. The final lobe of the cerebral cortex is the insular lobe (island of Reil). It cannot be seen from a surface view of the brain because it is buried deep within the lateral sulcus. In the human brain, the insular cortex is circumscribed by the circular sulcus. The gustatory (taste) discrimination and visceral sensory cortex is located in the anterior part of this lobe.

The "limbic lobe" of the cerebral hemispheres consists of the cingulate gyrus, the parahippocampal gyrus, and the retrosplenial cortex. The limbic lobe is part of the limbic system, which also includes structures such as the amygdala and hippocampal formation (both in the temporal lobe), the mammillary bodies of the hypothalamus, and the anterior nucleus of the thalamus. The cingulate cortex can be seen on the medial surface of the brain, lying above the corpus callosum and beneath the cingulate sulcus. The parahippocampal gyrus can be found on the medial and inferior part of the temporal lobe bounded laterally by the collateral sulcus and the rhinal sulcus more at the anterior aspect. The lateral olfactory area, in which the fibers of the olfactory tract end, is found on the inferior surface of the cortex. This olfactory region includes the uncus, which is the part of the rostral medial parahippocampal gyus that hooks backwards; the entorhinal cortex, which is found in the anterior part of the parahippocampal cortex; and cortex in the region of the limen insulae in the anterior part of the insular cortex. The corticomedial nuclei of the amygdala also receive olfactory tract terminals and are considered to be part of the lateral olfactory area.

B. Fibers

Immediately underneath the cerebral cortex lies a complex array of neural fibers involved in conveying information to and from other cortical and subcortical sites. There are three main types of fiber. The *association* fibers transmit nerve impulses between gyri in the same hemisphere. They can be relatively short, such as those connecting adjacent gyri, or long, connecting gyri in different lobes. Any one gyrus both projects to and receives input from a number of other cortical regions. *Commissural* fibers connect the gyri in one cerebral hemisphere to the corresponding gyri in the contralateral hemisphere. The majority of the commissural fibers are contained within the corpus callosum. However, the temporal lobes utilize both the corpus callosum and the anterior commissure for connecting fibers. The third class are the *projection* fibers, which provide connections between the cortex and subcortical sites such as the basal ganglia, thalamus, brain stem, and spinal cord. The projection fibers from the cerebral cortex initially form a spreading fan shape known as the corona radiata. At a deeper level, the corona radiata coalesces into a compact fiber bundle, the internal capsule. The ascending fibers enter the corona radiata after going through the internal capsule.

IV. LOCALIZATION OF FUNCTION

The term functional localization is used to indicate that certain functions can be localized to particular areas of the cerebral cortex. The mapping of cortical function began with inferences made from the deficits produced by cortical lesions in humans. Subsequently, techniques such as single-cell recording and electrical stimulation of cells in the cerebral cortex have been used in animals, nonhuman primates, as well as humans undergoing surgery for diseases such as epilepsy and Parkinson's disease to map out functional areas of the brain. This research has generated important findings, such as the somatotopic organization of both sensory and motor systems. Recently, research into functional localization in the cerebral cortex has been aided by the introduction of cerebral metabolism and blood flow imaging methods.

Functional neuroimaging techniques such as positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) have been used to confirm previous knowledge about localization of

function within the cerebral cortex as well as to conduct studies in healthy human subjects that were previously not possible. PET has been available for nearly 20 years. It involves monitoring the changes in metabolism or blood flow of a radiolabeled substance in response to a cognitive or sensorimotor task. PET studies have furthered our understanding of the organization of the somatomotor cortex, enabled us to examine visual perception, and begun to shed light on higher cognitive processes such as learning and attention. Moreover, we can localize receptors and determine the sites of action of different drugs in both healthy individuals and those inflicted with disorders such as epilepsy and schizophrenia.

fMRI is a relatively new (approximately 10 years) method that monitors local changes in blood oxygen levels that accompany the performance of a task. With its greater spatial resolution, fMRI has enabled more detailed studies of cortical function. For instance, a recent fMRI study demonstrated that the face representation in the somatosensory cortex, as described by Penfield, is in fact incorrect and is actually inverted along the central sulcus. In addition, it has been demonstrated that the topography in the motor planning areas (Brodmann area 6) is based on function (i.e., learned behaviors localized topographically) rather than on somatotopy. Recently, fMRI has been used to assess complex processes such as pain perception and to assess the effect of analgesic drugs on the areas of the cerebral cortex that are activated (Fig. 4).

Thus, both PET and fMRI have been used to study a wide range of processes, including sensation, perception, and attention in healthy human subjects, as well as changes in cerebral cortex function in patients suffering from neurological disorders. Numerous studies employing these techniques have contributed greatly to our knowledge of functional localization in the cerebral cortex.

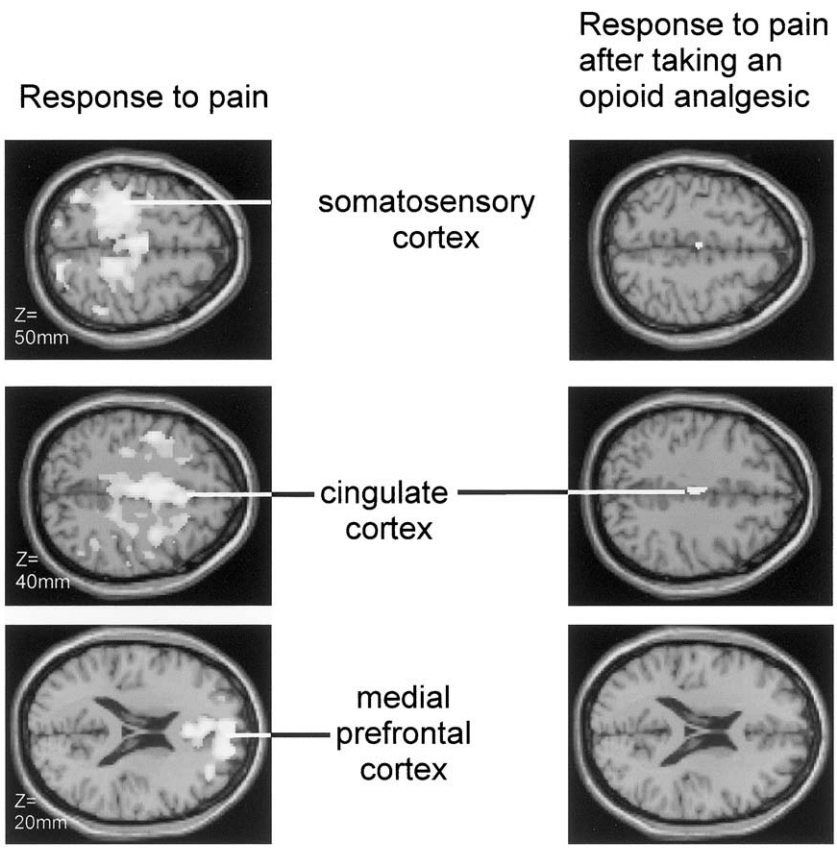


Figure 4 Images obtained using fMRI in response to a painful stimulus induced by stimulation of the ulnar nerve in the right arm. The images, taken at three different levels in the brain, illustrate sites of activation in the cortex (white) to the painful stimulus before (left) and after (right) administration of an analgesic (remifentanil). The top of each image represents the left side of the brain. Note the typical contralateral activation of the somatosensory cortex in response to the peripheral stimulus.

V. SENSORY SYSTEMS

Sensory systems are responsible for detecting changes in an organism's environment. There are three general classes of sensory systems: those concerned with the external environment (exteroceptive), those concerned with the inner environment (interoceptive), and those that monitor the positions of the body (proprioceptive). There are several different sensory systems localized in the cerebral cortex, including the somatosensory, visual, auditory, vestibular, taste, and olfactory systems. Within each of these systems there is a hierarchical organization for information processing such that input from sensory receptors in the periphery is relayed through the thalamus, first to the primary sensory cortex, then the secondary sensory cortex, and finally to the cortical association areas. The primary sensory cortices are involved in detecting, localizing, and discriminating the different properties of a stimulus, be it tactile, visual, auditory, etc. The secondary sensory cortices receive this information and integrate it with previous memories of the stimulus to help identify it. The sensory association areas, in turn, receive and integrate information from different sensory modalities to provide conscious perception of the stimulus and initiate plans for behavioral action in response to it.

A. Somatosensory System

Sensory information (proprioception, touch, pain, temperature, and vibration) from receptors in the skin, joints, and muscles throughout the body is relayed to layer 4 of the primary somatosensory cortex (SI) from the ventral posterior nuclei of the thalamus. SI is located along the postcentral gyrus of the parietal cortex and is divided into four parts based on cytoarchitecture and somatic inputs (Brodmann areas 3a, 3b, 2, and 1). The way in which sensory information (the location and intensity of a stimulus) is encoded and represented in SI is determined primarily by three principles of cortical organization: cortical receptive fields, preservation of somatotopy, and the columnar organization of the cortex.

The pyramidal cells in SI are continuously active, and this activity can be enhanced or inhibited by a stimulus acting at a specific location on the skin. This location in the periphery, where stimulation maximally activates a given cell in SI, is termed the *receptive field* for that cell. The receptive field for each cell in the cortex has a gradient of excitation, such that stimulation at the center of the field maximally excites the cell, and this excitation diminishes toward the outer edge of the field. There is also a less prominent inhibitory gradient that may extend beyond that of the excitatory one (inhibitory surround). This organization of the receptive field is the basis for *lateral inhibition* of sensory input, which enhances the ability to detect the precise location of a stimulus (Fig. 5).

Preservation of *somatotopy* refers to the fact that at all levels of the central nervous system, the organization of somatosensory input from the periphery is maintained. For instance, somatosensory information for discriminative tactile sensation enters the dorsal column of the white matter of the spinal cord in a medial–lateral manner, such that sensory information

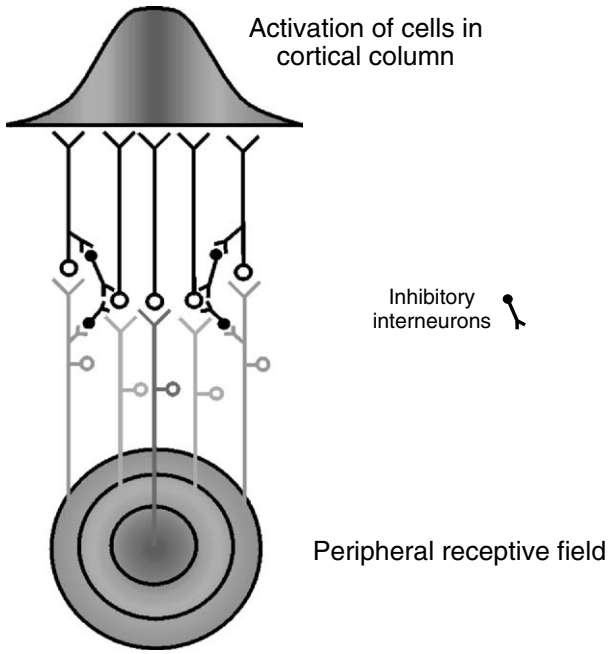


Figure 5 Schematic model of a peripheral excitatory receptive field with inhibitory surround. This provides an anatomical basis for the process of lateral inhibition of sensory input. Stimulation at the center of the receptive field maximally excites the sensory cells, which relay this excitation to a cortical column in the primary somatosensory area. Stimulation of cells in the outer edge of the receptive field suppresses the ascending sensory input by activating cells that synapse upon inhibitory interneurons. This pattern of firing is preserved as the information ascends the neuraxis, such that cortical neurons innervated by input from the center of the receptive field will be maximally activated, whereas the immediately adjacent cells in the cortex will be suppressed. This lateral inhibition acts to enhance the contrast between the stimulation site and peripheral areas to enable precise localization of the stimulus.

from the lower body lies medially and that of the upper body is organized laterally in the dorsal column. This organization is preserved throughout the brain stem, at the thalamic relay nuclei and ultimately in the primary somatosensory area of the cerebral cortex. The somatotopic organization in SI dictates that sensory input from the lower body is represented along the medial surface of the postcentral gyrus, whereas the upper body and head are represented along the lateral convexity of the gyrus. An interesting feature of somatotopy in SI is that those areas of the body that have large numbers of sensory receptors, such as the hands and mouth, have disproportionately much larger areas of SI cortex devoted to them. Moreover, the receptive fields in these areas of the skin are much smaller and more numerous than areas such as the trunk that are not concerned with precise discrimination of the stimulus.

Finally, in addition to its laminar organization as described previously, the cerebral cortex also possesses *columnar* organization, such that the cells in any given column of cortical tissue (from the cortical surface to the underlying white matter) respond maximally to a specific modality of stimulation. For instance, an edge placed on the skin with a specific orientation will optimally activate one cortical column, whereas the adjacent column will maximally respond to a slightly different orientation.

The primary somatosensory cortex uses these features in discriminating the intensity and location of a stimulus or detecting the shape or texture of an object. For example, the activation of a population of neurons on the medial surface of the postcentral gyrus indicates that a precise area of the lower body is being stimulated. The intensity of the stimulus is determined by the frequency of firing of the neurons and by the number of neurons recruited. Finally, the precise nature of the stimulus (pressure, touch, and direction of hair movement) is determined by which cortical columns respond.

Projection fibers from layers 2 and 3 of the trunk representation of SI innervate the contralateral SI as well as the secondary somatosensory cortex (SII). SII is located in the upper bank of the lateral sulcus, just inferior to the central sulcus (Fig. 6A). It is believed to have poor somatotopy, and the cells within SII have bilateral receptive fields, suggesting more of an integrative rather than a discriminative function. Information from SII is relayed to the posterior parietal cortex, which is considered to be the somatosensory association area (Brodmann areas 5 and 7). This area is involved in high-level integration of somatosensory information to enable learning and memory of the surrounding tactile and spatial environment. SI also has direct projections to subcortical structures (the basal ganglia, brain stem, and spinal cord) via pyramidal cells from layer 5 and to the ventral posterior nuclei of the thalamus from layer 6.

B. Visual System

The primary visual cortex (Brodmann area 17) is located in and on either side of the calcarine sulcus, on the medial surface of the occipital lobe. It functions primarily in discerning the intensity, shape, size, and location of objects in the visual field. As with other cortical areas, the primary visual cortex is composed of six layers. Visual input from the lateral geniculate nucleus of the thalamus terminates primarily in layer 4. This input is then relayed through layers 2 and 3 to secondary and association visual areas. In addition, an intracortical loop is established, with information from layers 2 and 3 being transferred to layer 5 and then to layer 6. Visual information is also relayed to the brain stem from the layer 5 of the cortex and back to the lateral geniculate nucleus of the thalamus from layer 6.

As in the primary somatosensory cortex, cells in the primary visual area are continuously active and respond maximally to specific stimuli within their receptive fields. The primary visual cortex also possesses columnar organization with each column of cells preferentially responding to a specific visual stimulus, such as a line of a specific orientation. In addition, it is retinotopically organized with the upper half of the visual field represented in the cortex inferior to the calcarine sulcus and the lower half of the visual field represented in the cortex superior to the calcarine sulcus. Moreover, there is a disproportionately large amount of cortex reserved for the area of the retina with the highest density of receptors (the macula).

The primary visual cortex, or striate cortex, sends information to the secondary visual, or prestriate cortex (Fig. 6B; Brodmann areas 18 and 19), and to the inferior temporal cortex (Brodmann areas 20 and 21). These secondary visual areas are important for color, motion, and depth perception. Information from the occipital cortex is further transmitted to two separate association areas of the cerebral cortex: superiorly to the parietal lobe (Brodmann areas 5 and 7) for processing of spatial features and movement and inferiorly to the temporal lobe for object recognition.

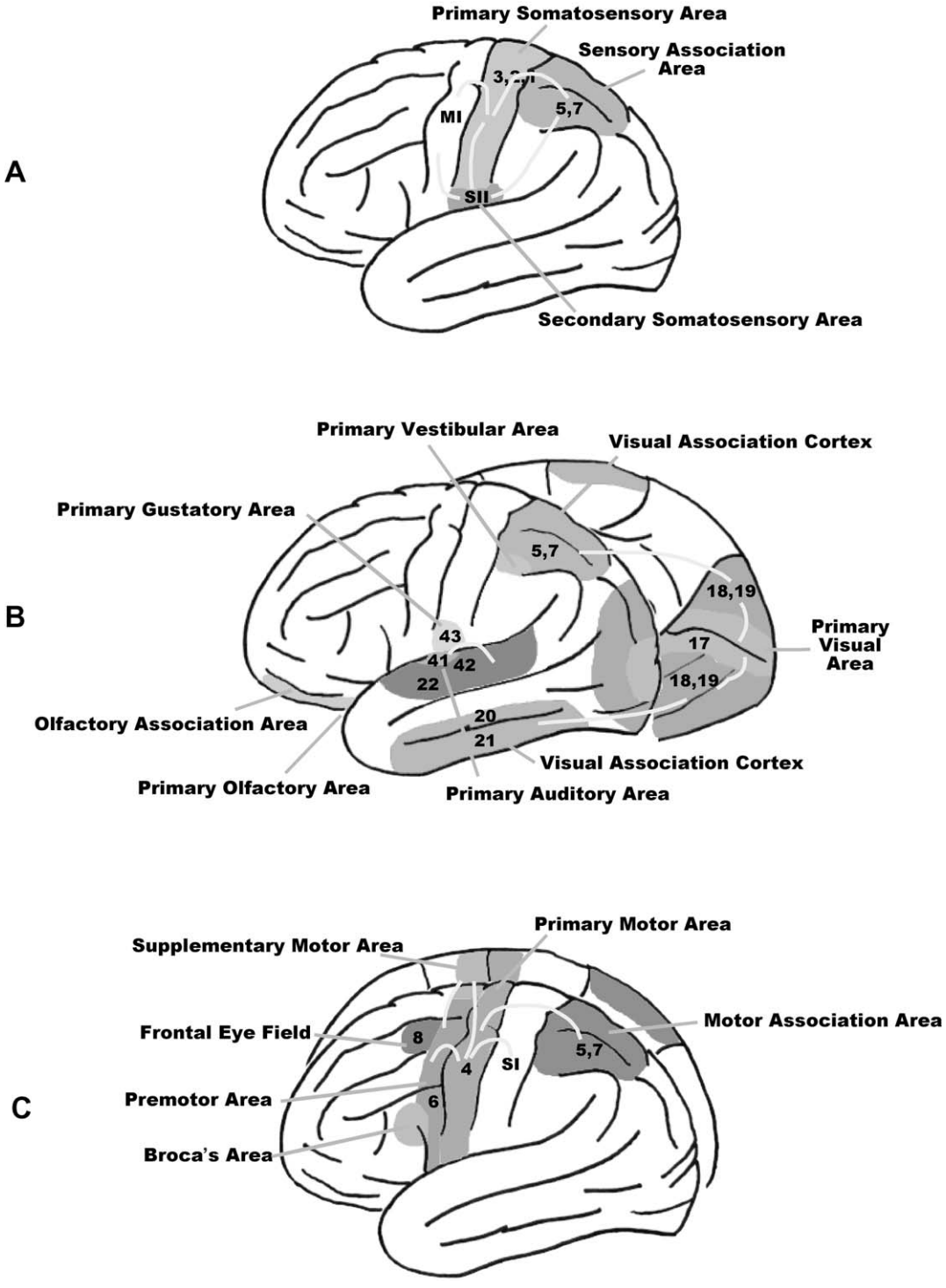


Figure 6 Schematic diagram of the localization of sensory and motor functions in the cerebral cortex. (A) Connections of the somatosensory system. (B) Connections of the visual, auditory, vestibular, gustatory, and olfactory sensory areas. (C) Connections of the somatomotor system. Numbers indicate Brodmann's areas. MI, primary motor area; SI, primary somatosensory area; SII, secondary somatosensory area.

C. Auditory and Vestibular Systems

The primary auditory cortex (Fig. 6B; Brodmann area 41) is located along the upper bank of the superior temporal gyrus, within the lateral sulcus on two gyri known as Heschl's gyri. Similar to other sensory cortices, the primary auditory cortex is organized in columns, such that each column of cells responds maximally to an acoustic stimulus of a specific frequency. Moreover, it is organized tonotopically (the anterior auditory cortex receives signals initiated by sounds with higher frequencies, whereas the posterior cortex receives lower frequency sounds) to enable the discrimination of the pitch of sounds. The auditory association cortex (Brodmann areas 22 and 42), which is concerned with the memory and classification of sounds, is inferior and posterior to the primary auditory cortex, along the superior bank of the middle temporal gyrus.

Movements and positions of the head are monitored primarily by the vestibular system. The primary vestibular cortex in humans is thought to occupy a region just posterior to the primary somatosensory cortex in the parietal cortex.

D. Olfactory System

Fibers from the olfactory bulb terminate in the corticomedial nucleus of the amygdala as well as the entorhinal (Brodmann area 28) cortex of the temporal lobe, comprising the primary olfactory cortex. The information is then relayed to additional areas of the limbic system such as the hypothalamus, which is important for controlling appetite, digestion, and feeding behaviors, and to the mediodorsal nucleus of the thalamus. From the thalamus, the olfactory information is sent to the orbitofrontal cortex (the olfactory association area), which also receives input from other cortical areas such as the somatosensory and visual association areas and insula and limbic system. Although there is segregation of sensory function throughout the cerebral cortex, these systems act in concert to mediate our conscious perception of the surrounding environment.

E. Pain System

Historically, the involvement of the cerebral cortex in the perception of pain (nociception) was an issue of controversy due to failure to elicit a painful response during stimulation of areas of the cortex as well as the lack of analgesia following cortical lesions. However, electrophysiological recording and neuroimaging techniques clearly indicate that numerous areas of the cerebral cortex do act synchronously to mediate the perception of pain and the responses to it. Such techniques have identified several cortical areas that are involved in nociception, including the primary and secondary somatosensory (SI and SII), the inferior and superior parietal, the insular, anterior cingulate, and medial prefrontal cortices.

There are believed to be two pathways, or streams, that mediate the perception of somatic pain, namely the medial and lateral pain pathways. The lateral pain pathway includes the ventral posterolateral and ventral posteromedial nuclei of the thalamus and their projections to SI of the parietal cortex. These thalamic nuclei and SI encode the type, temporal pattern, intensity, and topographic localization of a painful stimulus. Thus, the lateral pain pathway is believed to mediate the sensory-discriminate component of pain perception. Conversely, the medial pain pathway is thought to be concerned with the affective or emotional component of pain. Nociceptive information from the ventral posteroinferior, mediodorsal, and intralaminar nuclei of the thalamus is relayed to cortical areas such as SII, inferior parietal cortex, insular, anterior cingulate, and prefrontal cortices. The neurons in these areas generally have large, bilateral receptive fields and poor, or absent, somatotopic organization, consistent with their role in the affective–emotional dimension of pain perception.

Pain perception at the level of the cerebral cortex is complex and mediated by numerous interacting systems. For instance, all of the previously mentioned areas, particularly those involved in the affective–emotional component of pain processing, also have connections with diffuse areas of the limbic system and influence the ascending pain pathways. These connections likely mediate factors such as cognition, mood, and attention in response to a noxious stimulus.

There is an increasing amount of information regarding the neurochemistry of nociceptive connections at the cortical level. In the somatosensory cortices, excitatory amino acids acting on NMDA, AMPA, and metabotropic glutamate receptors likely mediate the transfer of nociceptive information from the thalamus to the cortex. In addition, there are large pools of GABAergic inhibitory interneurons in the sensory cortices that may modulate this transfer. Moreover, cytokines and neurotrophins may influence

neurotransmission via glutamate NMDA and GABA receptors.

VI. SOMATOMOTOR SYSTEM

Motor function is mediated by several areas of the prefrontal cortex, including the primary motor cortex, the motor planning areas (premotor area, supplementary motor area, Broca's area, and the area corresponding to Broca's area in the right hemisphere), and the cortical eye fields.

The primary motor cortex (MI) functions to mediate control of voluntary movements. It is located anterior to the central sulcus in the precentral gyrus (Fig. 6C; Brodmann area 4) and is the origin for some of the fibers of the corticospinal tract. All these fibers originate in layer 5 of the six-layered primary motor cortex and are pyramidal-type neurons. Primary motor cortex is an example of agranular cortex.

Similar to the primary somatosensory cortex, MI is somatotopically organized such that the lower body representation lies on the medial surface of the cortex, within the longitudinal fissure, whereas the upper body and face representations lie laterally along the precentral gyrus. In addition, there are large, disproportionate areas for the hand and mouth. MI receives input from numerous cortical areas, including SI (Brodmann areas 1–3), SII (Brodmann area 5), and premotor and supplementary motor areas (area 6), as well as input from the cerebellum, basal ganglia, and ventrolateral nucleus of the thalamus. The supplementary and premotor areas are collectively termed the motor planning areas and are responsible for organizing and planning complex movements.

The supplementary motor area (superior Brodmann area 6) is found anterior to the lower leg area of the primary motor cortex (medial surface of the medial frontal gyrus). It receives input from the prefrontal association cortex for planning of complex bimanual and sequential movements and for coordinating motor responses to sensory stimuli. Furthermore, efferent projections from the supplementary motor area are sent to the spinal cord and brain stem in addition to the primary motor cortex.

The premotor area (inferior Brodmann area 6), anterior to the primary motor area on the lateral surface of the prefrontal cortex, receives information from the prefrontal association cortex and sends output to the brain stem, basal ganglia, and cerebellum, in addition to the primary motor cortex. The premotor area is important for planning and control of visually guided movements. Located just rostral to the premotor area in the prefrontal cortex is the frontal eye field (Brodmann area 8), which controls voluntary conjugate movements of the eyes, independent of visual stimuli.

Finally, there are two areas that control the motor planning of speech and gestures associated with speech. Broca's area is located anterior to the premotor area in the inferior prefrontal cortex, usually in the left hemisphere. It is responsible for the motor programming of speech, whereas the corresponding area in the contralateral hemisphere (usually the right) is concerned with the planning of nonverbal communication such as gestures.

VII. CORTICAL ASSOCIATION AREAS

Three areas, the prefrontal, parietal, and temporal cortex, integrate and interpret sensory stimuli of all modalities and plan and execute behaviors in response to the stimuli. The prefrontal cortex corresponds to Brodmann's areas 9–12 and has extensive connections with parietal, temporal, and occipital lobes that provide information on sensory experiences. It is also reciprocally connected to the hippocampus and amygdala and receives a major input from the mediodorsal nucleus of the thalamus. This diverse anatomical arrangement suggests that the prefrontal cortex has multiple complex functional roles. The prefrontal association area, in the anterior prefrontal lobe, mediates personality, self-awareness, and executive functions (the planning and execution of a goal). The main symptoms seen with lesions of the prefrontal association area, as seen after frontal lobotomy, are loss of foresight, failure to anticipate, lack of drive and initiative, lack of spontaneity, apathy, and some subtle cognitive deficits. In addition, dysfunction of the prefrontal cortex has been implicated in schizophrenia. PET and fMRI in schizophrenic patients demonstrate impairments in activation or metabolism in the prefrontal area during certain behavioral tasks. Postmortem studies indicate that there are cytoarchitectural changes as well as neurotransmitter and neurotransmitter receptor deficits in schizophrenic patients.

The posterior parietal cortex consisting of Brodmann's areas 5 and 7 is situated between the somatosensory region of the postcentral gyrus and the visual cortex, from which it also receives inputs. Long

association fibers also connect this region with the premotor and prefrontal association cortices. These anatomical arrangements allow the parietal association cortex to receive and integrate multimodal sensory information and to send this integrated view of the environment to the premotor and prefrontal association areas to initiate the appropriate behavior. Lesions of the superior parietal association cortex usually result in a lack of attention to objects or events in the space contralateral to the damage, a phenomenon known as cortical neglect. Such lesions lead to tactile agnosias, which is the inability to recognize objects by touch. Auditory agnosia, which is the inability to recognize what is heard, occurs when there is bilateral damage to the auditory association area behind the primary auditory cortex. Visual agnosia, which is the inability to recognize objects by sight, is the result of damage to the inferior part of the occipital and temporal lobes with damage to the visual association cortex.

Other lesions, which interrupt connections between the parietal association cortex and the premotor cortex, lead to apraxia or the inability to carry out learned movements even though there is no weakness or lack of sensation. The parietal association area, particularly the inferior region, shows a high degree of lateralization. In most individuals, the left inferior parietal association cortex is concerned with language processing, whereas the right is involved in spatial relationships.

The lateral and ventral temporal cortex is also considered association cortex and demonstrates a high degree of lateralization. Lesions of the left temporal cortex result in deficits in understanding spoken and written language, whereas damage to the right temporal lobe leads to problems using visual, tactile, or auditory information to identify complex shapes or sounds. There is a dorsal–ventral organization within the temporal association cortex such that the auditory representation is located in the more superior regions, whereas the visual component is found in the inferior temporal cortex.

VIII. VISCERAL REPRESENTATION AND FUNCTION

In recent years, considerable progress has been made in determining the role of the cortex in autonomic control in health and disease. Characteristics of the cardiovascular responses elicited from the insular cortex have been determined as well as the efferent and afferent pathways, the nature of the responses, the control of tonic sympathetic activity, and the neurotransmitters mediating these responses. Very early experiments demonstrated that "sham rage" could be induced in cats following crude decortication. Upon innocuous stimulation, such as stroking the fur, cats would demonstrate all the somatomotor and autonomic manifestations (piloerection, dilatation of the pupils, retraction of the nictitating membrane, panting, and tachycardia) of rage. Ablation of the orbitoinsular region of the frontal lobes in cats resulted in a syndrome similar to that of complete decortication or complete removal of the frontal lobes. These studies suggested that the insular cortex might play a role in the tonic regulation of autonomic (particularly sympathetic) responses. Stimulation studies have supported this view. Electrical stimulation of the insular cortex in a variety of mammals, including humans, elicits changes in blood pressure, heart rate, respiration, piloerection, pupillary dilatation, gastric motility, peristaltic activity, salivation, and adrenaline secretion. Phasic microstimulation of the rat insular cortex linked to the R wave of the electrocardiogram (ECG) evokes tachycardia or bradycardia responses without accompanying changes in blood pressure or respiration, and prolonged stimulation in the insular cortex generates progressive degrees of heart block and increased plasma norepinephrine and can cause death in asystole. Accompanying cardiac structural changes are myocytolysis and subendocardial hemorrhages in the vicinity of the origin of the bundle of His, suggesting increased cardiac sympathetic activity. Finally, the efferent pathways and neurotransmitters for these autonomic responses from the insular cortex have been determined. It is clear that a mandatory synapse is located in the lateral hypothalamic area and that the primary neurotransmitter in this region is glutamate acting at NMDA receptors.

The insular cortex is considered to be the visceral sensory cortex on the basis of anatomical and physiological evidence. The earliest demonstration of general visceral input to a cortical level was an investigation in cats that showed that stimulation of the cephalic end of the severed cervical vagus nerve resulted in an increase in the rate and amplitude of cortical electrical activity in the orbitoinsular region. Recent anatomical results in the rat have indicated that the insular cortex might be organized in a viscerotopic manner. The recording of individual neurons in response to activation of specific visceral receptors demonstrated that neurons responding to gustatory inputs were located in the

dysgranular region of the insular cortex. General visceral inputs were primarily located in the granular region of the insular cortex in the rat. In this region, there was a separation between the regions receiving gastric inputs compared to the cardiopulmonary responsive neurons. The distribution of arterial baroreceptor responsive neurons was extensively mapped in the rat insular cortex, which demonstrated that significantly more neurons responding to blood pressure changes were located in the right insular cortex than in the left. This granular insular cortex in which the cardiopulmonary information terminates may be the critical region in the lateral frontal cortex mediating central autonomic manifestations of emotional behavior as predicted by the early experiments demonstrating sham rage.

Gustatory or taste input is considered to be a special visceral sensation. Gustatory input is relayed ipsilaterally from the thalamus to the primary gustatory area, which is located near the face region in the primary somatosensory cortex. Although a cell in the gustatory cortex can be activated by any number of taste stimuli, it will have a preference for one of the four basic taste qualities (sour, sweet, bitter, and acidic). The detection and subsequent identification of a particular taste is not mediated by individual cells in the cortex; rather, the quality of the taste is determined by the pattern of firing in a group of activated cells. The anterior insular cortex (secondary gustatory cortex) also receives gustatory input and this area integrates the gustatory information with input from the olfactory system. Information from both the primary gustatory area and the insula is integrated with other sensory information in the orbitofrontal cortex.

Recent evidence has shown for the first time that it is possible to obtain a clear representation of visceral sensation in the human insular cortex similar to that observed in the rodent. fMRI was used to identify regions of the human brain that were activated in response to a series of tests designed to stimulate cardiopulmonary and gustatory receptors. Cardiopulmonary activation included maximal inspiration, Valsalva's maneuver, and maximal handgrip to elevate arterial blood pressure. These maneuvers consistently resulted in discrete changes in activity in the anterior insular cortex, with a time course corresponding to the changes in arterial blood pressure and heart rate they produced (Fig. 7). Gustatory stimuli, such as salt and sucrose perfusion of the tongue, resulted in activation of the inferior anterior insular cortex.

Several investigations have indicated that the insular cortex may mediate the cardiovascular consequences of stroke. Middle cerebral artery occlusion (MCAO) in the cat and rat results in an increase in blood pressure, norepinephrine level, sympathetic nerve activity, myocytolysis, and death. These changes resemble those seen clinically and are obtained only when the insular cortex is included in the infarct. Asymmetry of responses elicited from the cerebral cortex is an important topic in many types of behavioral investigations. There is evidence that stimulation of the insular

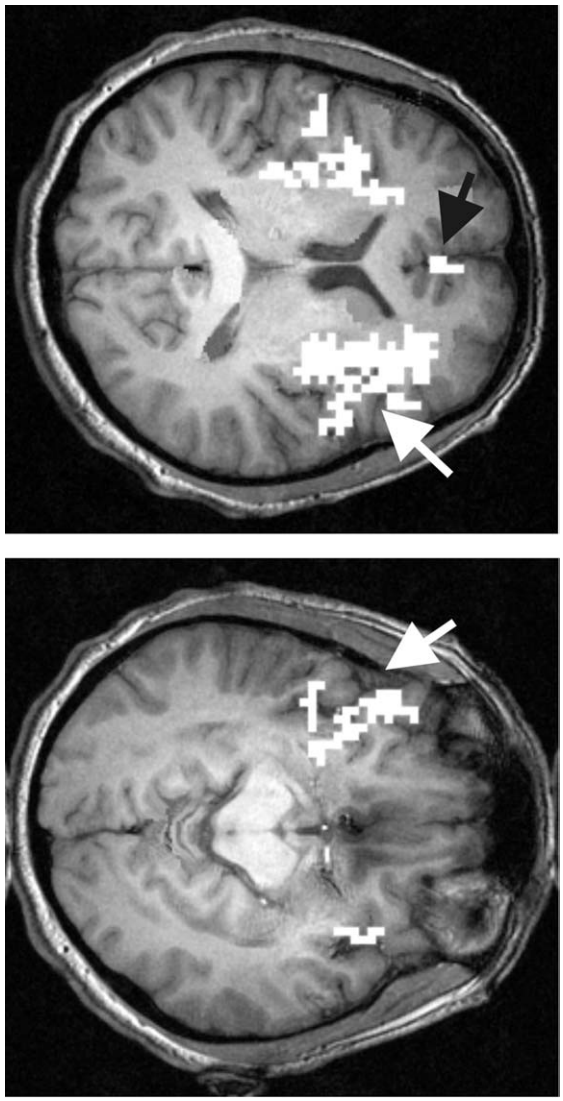


Figure 7 Activation of visceral regions of the cortex in response to an increase in arterial blood pressure. White arrows indicate activation in the anterior insular cortex, and the black arrow shows medial prefrontal cortex activation. The top of each image represents the left side of the brain.

cortex in the human also elicits different results from each hemisphere. In the MCAO model in the rat, right-sided stroke results in a significantly higher increase in mean arterial blood pressure, sympathetic nerve activity, and plasma norepinephrine and an increase in the QT interval of the ECG. Finally, the animal stroke model has clearly shown that the cardiovascular consequences of stroke are more severe with increasing age. The aged animals had significantly increased mortality following the stroke and prior to death exhibited significantly elevated sympathetic nerve activity, plasma norepinephrine concentration, and prolonged QT interval of the ECG. Clinical studies have confirmed the importance of the insular cortex in mediating the cardiovascular consequences of stroke. Plasma norepinephrine levels, changes in circadian blood pressure, and incidence of cardiac arrhythmias correlate highly with percentage insular infarction. These investigations also confirm observations of right hemispheric dominance for sympathetic effects. There is some indication that the left insular cortex is predominantly responsible for parasympathetic effects since left insular stimulation in humans results in more frequent parasympathetic responses (bradycardia) compared to right insular stimulation, which yielded increased blood pressure and heart rate (sympathetic effects). Recently, it was shown in a group of patients with lesions mainly confined to the left insular cortex that there was an increased basal cardiac tone associated with a decrease in heart rate variability.

Anatomical and functional data clearly indicate that another cortical site in the medial prefrontal cortex plays an important role in determining the autonomic responses of the organism to complex behaviors. This medial prefrontal cortical area is considered to be visceral motor cortex. Compared to the insular cortex, considerably less is known regarding the specific role of the medial prefrontal cortex in autonomic control. A variety of autonomic responses can be elicited by stimulation of the medial prefrontal cortex. In fact, complete cessation of heartbeat may occur in monkeys during stimulation of the medial frontal cortex. The medial prefrontal cortex has extensive connections with both the limbic and the autonomic systems of the brain. It receives inputs from the insular and entorhinal cortices, the hippocampus, amygdala, the visceral relay nuclei of the thalamus, and pontine and medullary autonomic control sites. The medial prefrontal cortex has descending projections to many autonomic sites, including the insular cortex, the amygdala, the visceral relay nuclei of the thalamus, hypothalamus, and brain stem autonomic control nuclei.

IX. CORTICAL PLASTICITY

In the past 10 years, many observations have indicated that the cerebral cortex in the adult is not as hardwired as previously thought. Although the concept of plasticity was associated with the developing brain, it was not thought to be possible in the mature central nervous system. It has been demonstrated that a major change in sensory input can result in a reorganization of the cerebral cortex. Either an increase in afferent activation or a deficit in sensory input can cause substantial changes in cortical representation. For example, in blind subjects who rely extensively on Braille for reading, it has been shown that cortical representation for the digits that are utilized can be substantially increased. Deafferentation due to amputation of a limb or digit or severing of afferent nerves leads to an invasion of the dennervated cortical region by an adjacent intact region. For example, amputation of a digit results in encroachment of the cortical region previously representating the digit by the areas representating adjacent digits. In addition, the smaller region that represents the amputated digit has a greater degree of representation by the remaining stump and a much reduced area in the region previously representing the distal skin of the digit. The extent of the invasion can be quite substantial. Deafferentation of an upper limb by severing the dorsal roots results in an invasion of the region representing the face by as much as 15–20 mm.

The immediate response in the cortex is an expansion of representation from an adjacent region into the region previously serving the deafferented limb or digit. At a later stage, the topography in the region is reorganized based on the extent of use or level of afferent input that is being received. The mechanisms responsible for this cortical plasticity are thought to be the same for both the expansion of a cortical region in response to an increase in peripheral input and the invasion of a cortical area following the deafferentation of a peripheral input. In the case of the deafferentation of a digit or limb, there are two distinct phases. For several weeks, the region of the cortex for representation of the deafferented limb or digit becomes totally silent. At a later stage, 6 months or longer, the invasion by the adjacent region takes place.

Currently, there are two proposed mechanisms for how this large amount of cortical plasticity can occur. First, it is known that connections from adjacent regions already exist. These residual connections are present but are not responsive to afferent input. When the cortical region loses its primary input these silent

connections begin to respond to afferent input. This causes changes in neurotransmitters levels and the numbers of neurotransmitter receptors, which in turn results in an increase in the synaptic representation of the alternative region. This increase in synaptic representation may even lead to sprouting of nerve terminals resulting in an even greater degree of spreading of the cortical reorganization. A mechanism restricted to the cortex is supported by evidence that local microstimulation in the cortex is sufficient to produce an increase in cortical representation similar to the effect of an enhancement of peripheral afferent input. However, this may not account for the complete extent of cortical reorganization that occurs. There is good evidence that a great deal of reorganization takes place at subcortical levels. It is postulated that small amounts of reorganization at thalamic and brain stem levels will be amplified at the level of the cerebral cortex.

X. FUTURE DIRECTIONS

This overview of the current state of knowledge on topics related to the cerebral cortex clearly indicates that there are numerous exciting avenues for future research. For instance, there is much to be learned about the molecular nature of processing in the brain. The localization of different receptors and receptor subtypes, the interactions of the neurotransmitters on specific regions of these receptors, and the subsequent intracellular signaling cascades are all important to understanding processing of neural signals in the intrinsic circuitry of the cortical columns and how this is transformed into cognitive processes, such as attention, memory, and personality. Understanding the molecular nature of the cells of the cerebral cortex also has important developmental and clinical implications.

It is also critical to continue to explore mechanisms of cortical plasticity. The specific genes that control plasticity in the developing compared to the mature cerebral cortex need to be determined to provide insight into potential repair mechanisms for neurotrauma or neurodegenerative diseases. In this respect, investigations into therapeutic methods such as the use of neurotrophic factors and neuronal stem cells following stroke or other causes of cerebral cortex injury are providing exciting results. These potential gene therapies will also be of interest to investigators studying the neurodegenerative cerebral cortex disorders such as Alzheimer's disease and schizophrenia.

Finally, neuroimaging methods such as PET and fMRI will continue to be powerful tools to unravel the manner in which the cerebral cortex functions in the human to produce complex behaviors and cognition. This technology will be of fundamental importance in the future for not only localizing areas of the cortex devoted to particular functions but also for identifying the network of connections throughout the brain that mediate the higher cognitive functions that are unique to the human or conscious mind.

See Also the Following Articles

AGNOSIA • APRAXIA • AUDITORY AGNOSIA • BROCA'S AREA • CEREBRAL CIRCULATION • CHEMICAL NEUROANATOMY • CINGULATE CORTEX • GABA • NEUROPLASTICITY, DEVELOPMENTAL • PAIN • VISUAL AND AUDITORY INTEGRATION • VISUAL CORTEX

Suggested Reading

Creutzfeldt, O. (1995). *Cortex Cerebri: Performance, Structural, and Functional Organization of the Cortex*. Oxford Univ. Press, Oxford.

Jones, E. G. (2000). Cortical and subcortical contributions to activity-dependent plasticity in primate somatosensory cortex. *Annu. Rev. Neurosci.* **23,** 1.

Mountcastle, V. B. (1998). *Perceptual Neuroscience: The Cerebral Cortex*. Harvard Univ. Press, Cambridge, MA.

Parent, A. (1996). *Carpenter's Human Neuroanatomy*. Williams & Wilkins, Baltimore.

Peters, A., and Jones, E. G. (Eds.) (1984–1999). *Cerebral Cortex,* Vols. 1–14. Plenum, New York.

Roberts, A. C., Robbins, T. W., and Weiskrantz, L. (Eds.) (1998). *The Prefrontal Cortex: Executive and Cognitive Functions*. Oxford Univ. Press, Oxford.

Sakata, H., Mikami, A., and Fuster, J. M. (Eds.) (1997). *The Association Cortex: Structure and Function*. Harwood Academic, Amsterdam.

Servos, P., Engel, S. A., Gati, J., and Menon, R. (1999). fMRI evidence for an inverted face representation in human somatosensory cortex. *NeuroReport* **10,** 1393–1395.

Cerebral Edema

DAVID S. LIEBESKIND
University of California, Los Angeles

GLOSSARY

cytotoxic Harmful to cells.

diuresis Increased removal of water.

edema The excessive accumulation of fluid in tissue.

ependyma The thin membrane that lines the ventricles of the brain.

extravasation The leakage and spread of fluid from vessels into the surrounding tissues.

herniation The abnormal protrusion of tissue that results in compression of neighboring structures.

interstitium The small tissue space between structural elements.

osmolarity The concentration of a molecule dispersed in a fluid.

osmole A molecule dissolved in a fluid.

parenchyma The functional part of an organ.

pinocytosis A cellular transport mechanism that functions through cytoplasmic engulfment of fluid.

transudation The passage of a fluid through a membrane.

vasogenic Resulting from changes in blood vessels.

Cerebral edema is a heterogeneous condition defined as the volumetric increase in brain tissue resulting from the accumulation of water. Elements of cerebral edema form integral components of the pathophysiology of many neurologic diseases.

I. PATHOPHYSIOLOGY

The concept of cerebral edema has been recognized for more than 2000 years, yet an understanding of the complex physiology of this condition has evolved only within the past 30 years. Ancient Greek authors used the term "$\sigma\iota\delta\eta\mu\alpha$" to describe the swelling of the brain that resulted from compound skull fractures. Hippocrates noted that removal of the overlying skull bones allowed the injured brain to swell outward, thus minimizing compression of normal tissue trapped within the cranial vault. The Monro–Kellie doctrine later recapitulated this concept, affirming that when "water or other matter is effused or secreted from the blood vessels … a quantity of blood equal in bulk to the effused matter, will be pressed out of the cranium." This indiscriminate concept of brain swelling was cited in a diverse range of clinical settings until 1967, when Igor Klatzo defined the modern classification of edema based on pathophysiology. Cerebral edema, according to Klatzo, was defined as "an abnormal accumulation of fluid associated with volumetric enlargement of the brain." This entity was divided into vasogenic edema, characterized by derangement of the blood–brain barrier (BBB), and cytotoxic edema, related to intracellular swelling in the absence of changes at the BBB. The term cytotoxic edema has recently been reserved for states associated with toxin-induced cellular swelling. Klatzo emphasized that these two forms usually coexisted. In 1975, Robert Fishman added interstitial edema as a distinct entity by describing the transependymal flow of cerebrospinal fluid (CSF) into the periventricular white matter in individuals with acute obstructive hydrocephalus; this form was later termed hydrocephalic edema. In similar fashion, ischemic

Table I
Classification of Cerebral Edema

Type	Location	Site	Blood–brain barrier	Pathogenesis	Clinical correlation
Vasogenic	Extracellular	White matter	Disrupted	Vascular permeability	Lead encephalopathy
					Neoplastic disease
					Infection
					Fulminant hepatic failure
					Traumatic brain injury
					High-altitude cerebral edema
Cytotoxic	Intracellular	White or gray matter	Intact	Cellular injury	Toxin exposure
					Ischemia
					Traumatic brain injury
					Fulminant hepatic failure
					Reye's syndrome
Hydrocephalic	Extracellular	White matter	Intact	Transependymal pressure	Hydrocephalus
Ischemic	Intra- and extracellular	White and gray matter	Disrupted	Hypoxia	Ischemia
Hydrostatic	Extracellular	White and gray matter	Disrupted	Hydrostatic pressure	Hypertensive encephalopatby
Osmotic	Intra- and extracellular	White and gray matter	Intact	Osmotic pressure	Hyponatremia
					Dialysis disequilibrium
					Diabetic ketoacidosis

edema was described, incorporating elements of both cytotoxic and vasogenic forms. Additional classifications based on mechanism and location of excess fluid include hydrostatic edema and osmotic edema. Hydrostatic edema results from increased hydrostatic forces with associated disruption of the BBB, as occurs in the setting of hypertensive encephalopathy. Osmotic edema has been ascribed to states of plasma hypoosmolarity that result in cellular swelling with preservation of the BBB. Table I summarizes the nomenclature of cerebral edema. This classification scheme emphasizes the predominant form of edema associated with several clinical scenarios and delineates the corresponding pathophysiology. The remainder of this article utilizes this classification of cerebral edema, emphasizing the role of each form in a variety of clinical encounters.

Consideration of the specific pathophysiologic mechanisms is crucial for establishing a fundamental approach to the various forms of cerebral edema. Each form of cerebral edema results in a net gain of water in the brain, although different mechanisms determine the specific location of water accumulation and resultant pathology. Numerous descriptions fail to distinguish cerebral edema from other causes of intracranial hypertension. Cerebral edema may result in intracranial hypertension, but the two conditions are not synonymous because intracranial hypertension may result from increased cerebral blood volume or increased CSF content. This review specifically focuses on cerebral edema, in distinction from other causes of intracranial hypertension. Various forms of cerebral edema may coexist and are often closely interrelated. One form of edema may also lead to the development of another, particularly in vasogenic and cytotoxic edema. For these reasons it is particularly important to consider the time course involved in specific clinical scenarios. The underlying processes must be delineated to understand and develop rational therapeutic approaches that focus on the particular pathophysiologic mechanism.

A. Distribution of Fluid in the Brain

The normal water content of the human brain is approximately 80% by weight, with higher amounts of

water in the gray matter than in the white matter. Fluid is distributed among the intraluminal cerebrovasculature, the intracellular space of the parenchyma, the extracellular space of the interstitium, and the ventricular system. Therefore, a complex interaction of morphologic features and biochemical processes exists to maintain the normal fluid balance of the brain. An increase in the intravascular plasma volume of the cerebral circulation may lead to fluid engorgement of the brain. Although such an increase in cerebral blood volume is not inherently pathologic, an associated alteration of the BBB may cause fluid to extravasate into the extracellular compartment. Increased water content in the interstitial space may result from obstruction of CSF outflow in the ventricular system with resultant transependymal exudation of fluid. Fluid homeostasis within the intracellular compartment may be altered by many factors, thus leading to cellular edema. To define characteristic features of specific forms of cerebral edema, the location of edema fluid accumulation as extracellular, intracellular, or both proves to be a useful tool, as does the predilection for gray or white matter involvement.

B. Fluid Dynamics of the Intracranial Compartment

Intracranial volume is limited by the confines of the adult calvarium. An increase in the volume of one fluid compartment results in compensatory decrease in the others. Displacement of CSF and venous blood compensates for pressure differentials transmitted between compartments. These volumetric changes in response to pressure differences define the brain's compliance. Under normal physiologic conditions, changes in the intracranial volume are associated with minimal or insignificant fluctuations in the intracranial pressure (ICP), but considerable increases in ICP may occur as the intracranial volume increases and brain compliance decreases. The relationship between intracranial volume and ICP is illustrated in Fig. 1. ICP is commonly evaluated by measuring the opening pressure of the CSF in the lateral decubitus position or with a ventriculostomy or subdural monitor in the neurointensive care unit. ICP normally ranges from 15 to 20 mmHg. Focal elevations in ICP may lead to displacement of brain tissue with herniation, precipitating neurologic deficits that may be fatal. Neuropathologic correlates of herniation are illustrated in Fig. 2.

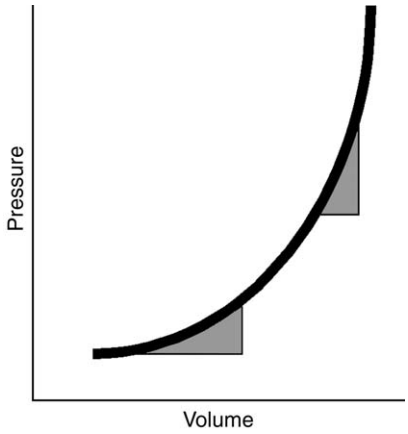


Figure 1 Intracranial volume and ICP correlation, demonstrating that relatively larger pressure increments result from decreasing compliance with volume increases.

Changes in ICP may affect cerebral perfusion pressure (CPP). The CPP is equal to the difference between the mean arterial pressure and the ICP. In normal circumstances, autoregulation of vasomotor tone in the cerebral circulation allows for maintenance of an adequate CPP, coupling brain metabolism with cerebral blood flow. Autoregulation is predominantly mediated at the arteriolar and capillary level under the influence of vasoactive factors and direct innervation from the sympathetic nervous system. Cerebral capillary function is regulated by many substances, including adenosine, vasoactive intestinal peptide, and arachidonic acid metabolites. Numerous disease states including cerebral ischemia and head trauma interfere with this protective mechanism. As the ICP exceeds a critical threshold of 20–25 mmHg, CPP decreases, with significant decrements resulting in cerebral ischemia.

A simplistic approach to fluid dynamics in the intracranial compartment is derived from the application of Starling's law, which states that fluid flux is dictated by a balance between hydrostatic forces and osmotic gradients. The complex structural and functional properties of the BBB result in a highly modified form of Starling's law with dynamic homeostasis influenced by local and systemic mediators, although the concept of balanced hydrostatic and osmotic forces is preserved. Plasma flow across the BBB is driven both by hydrostatic forces determined by the relatively constant CPP and by osmotic forces principally determined by the concentration gradient of sodium ions. Under pathologic conditions, the osmotic forces are modified by the generation of idiogenic osmoles, substances produced by brain parenchyma to prevent

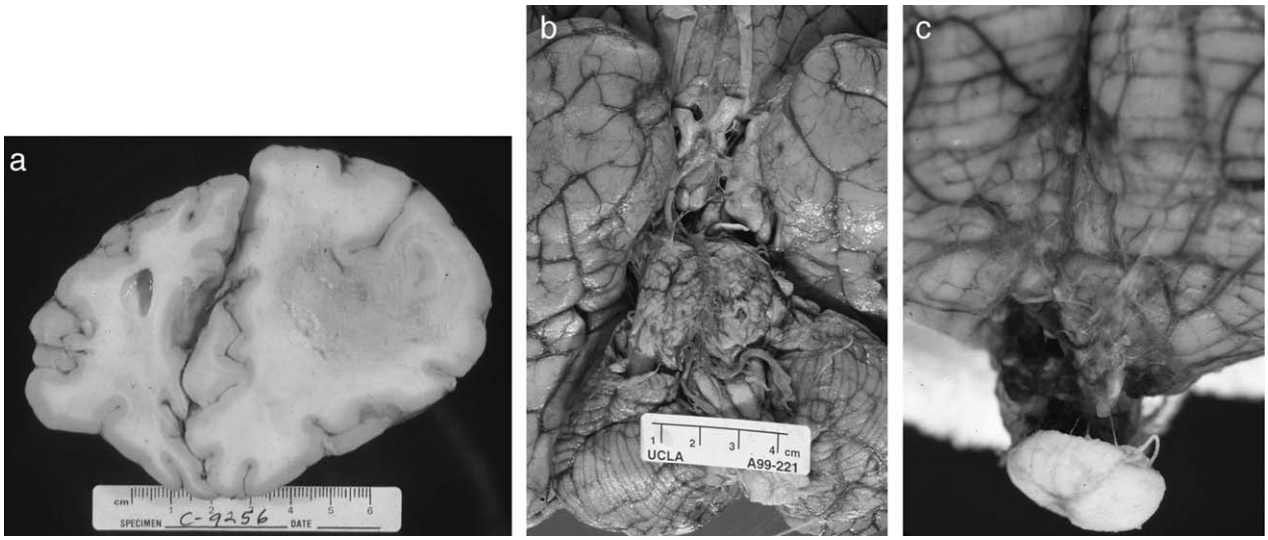


Figure 2 Neuropathologic features of herniation syndromes demonstrated in gross brain specimens. (a) Subfalcine midline shift due to a frontal lobe glioma. (b) Uncal herniation of the temporal lobe over the tentorial edge due to a traumatic hematoma. (c) Compression of the cerebellar tonsils following elevated ICP. (Courtesy of Harry V. Vinters, M.D.)

rapid fluid loss with resultant cellular shrinkage. These idiogenic osmoles included the following: alanine, aspartate, betaine, choline, GABA, glutamate, glutamine, glycerophosphorylcholine, glycine, lysine, *myo*-inositol, phosphocreatine, serine, taurine, and threonine.

C. The Blood–Brain Barrier

The BBB maintains the flux of ions, water-soluble nutrients, and metabolites to titrate the specific composition of extracellular fluid in the central nervous system. The brain parenchyma is contained within an immunologically separate compartment under tight volumetric control. This compartmentalization is achieved and maintained by a combination of specific morphologic features and biochemical factors.

The BBB consists of endothelial cells bridged by tight junctions and adjoined by foot processes of astroglia. Endothelial cells form the inner lining of all vessels and have common properties that include a nonthrombogenic luminal surface, a basement membrane, and the production of active substances. Astrocytic foot processes extend to form a tight network that enmeshes the outer surface of brain capillaries. The endothelial architecture of the cerebral circulation is distinct from other vascular territories. Early in life, the cerebrovascular endothelium becomes almost completely impermeable, thereby isolating the brain from the systemic circulation. The BBB exists throughout the central nervous system, leaving only specific areas, such as the choroid plexus and the circumventricular organs, devoid of this structure. The choroid plexus is responsible for the production of CSF, modifying plasma from the cerebral circulation to maintain the protective fluid compartment of the ventricular system. The CSF circulates from the ventricular system into the subarachnoid space where it functions as a cushion for the brain, protecting it from trauma. The circumventricular organs include chemosensitive and neurosecretory sites that are involved in body fluid turnover, regulation of fluid osmolarity and extracellular sodium content, and cardiovascular function.

Several structural features of the BBB account for the specialized function of this cellular layer. Complex interendothelial junctions prevent even small ions from entering the brain parenchyma. Only lipid-soluble substances are allowed to diffuse through this layer. Pinocytosis, the transport of materials within small intracellular vesicles, is necessary for most substances to traverse the endothelial layer. Electron microscopy of the BBB demonstrates the close opposition of endothelial cells, including a paucity of vesicular transport and a high density of mitochondria. The mitochondrial content of brain endothelia is approximately 10 or 11%, compared to 2.7% in other

blood vessels. This elevated concentration of mitochondria is necessary for the high rate of metabolic activity that occurs. The passage of polar molecules, including many amino acids, is accomplished by several specialized transport systems.

The functional properties of the BBB include a high transendothelial potential and resistance, with a high reflection coefficient and low hydraulic conductivity. The intracellular compartment of the endothelial layer has marked enzyme activity associated with multiple transport systems. Water transport across the BBB is under close neurogenic and endocrine control. Biologically active agents that alter the BBB include substance P, adenosine nucleotides, endothelin-1, calcium entry blockers, metalloproteinases, cytokines, vascular endothelial growth factor or vascular permeability factor (VEGF/VPF), and antibodies to filamentous hemagglutanin. Several cytokines, including interleukins, tumor necrosis factor-α (TNF-α), and platelet-activating factor (PAF), are active at the BBB. Vasoactive substances can increase BBB permeability by activating cyclic nucleotide-generating enzymes or by increasing the rate of lipolysis through phospholipases. Vasodilators such as arachidonic acid, noradrenaline, and histamine may also increase cytoplasmic ionized calcium and activate microvascular guanylate cyclase. These actions result in increased pinocytosis and macromolecular transport. Modulation of BBB permeability has been the focus of recent therapeutic interventions for brain tumors, employing steroidal substances and chemotherapeutic agents.

D. Cellular Elements

Although the BBB serves as the major regulatory site of fluid exchange in the central nervous system, multiple actions at the cellular level in the brain parenchyma further influence fluid homeostasis. These cellular elements include the neuroglia, mainly astrocytes, and the neurons. Neuroglia were once thought to form an inactive architectural framework that supported the neuronal population. Further investigation revealed that neuroglia play an important role in the maintenance of fluid composition and volume. These cells contain numerous membrane surface ion channels that exchange molecules between the intracellular space and the extracellular milieu of the interstitial space and thereby adjust the electrochemical gradients influencing excitatory thresholds and the release of neurotransmitter substances. Glutamate, the main excitatory neurotransmitter of the central nervous system, is ubiquitous in the intracellular space. Efficient uptake systems in glial and neuronal tissue clear glutamate from the extracellular space.

Volumetric changes are energy dependent and modified by fluid osmolarity, principally determined by the sodium concentration. Sodium is exchanged with potassium at the Na^+/K^+ ATPase channel, requiring energy from the release of ATP. This channel is oxygen dependent and therefore sensitive to hypoxic stress. As the astrocytic sodium concentration increases, cellular volume expands and vice versa. Glial swelling represents a normal component of homeostasis, although extreme states occur in pathologic conditions such as hypoxia.

Although sodium shifts are the principal determinant of cellular volume changes, the measurement of sodium concentrations in abnormal hyperosmolar states has revealed that other substances might account for the osmolar increase. Specific identification of these substances has proved difficult, but the concept of idiogenic osmoles has been advanced. These molecules are produced under conditions of hypoosmolarity in order to compensate for volume decreases. During the resolution phase of systemic hyperosmolar states, a differential equilibration of inorganic ions such as sodium relative to idiogenic osmoles results in cellular swelling. Throughout this regulatory volume decrease of brain cells, organic osmolytes leave the cell through a molecular pore or ion channel, termed the voltage-sensitive organic anion channel (VSOAC). Osmotic swelling normally activates the VSOAC, although this pathway may fail under pathologic conditions.

E. Classification of Cerebral Edema

1. Vasogenic Edema

Vasogenic edema results from increased permeability of the BBB. This form of cerebral edema occurs in the setting of brain tumors, abscesses, meningitis, lead encephalopathy, traumatic head injury, ischemia, and intracerebral hemorrhage. Klatzo developed an animal model to study the neuropathology of cold-induced vasogenic edema. A metal probe cooled with liquid nitrogen caused focal swelling of astrocytic foot processes, with the majority of fluid accumulation occurring in white matter.

Several mechanisms have been implicated in BBB disruption associated with vasogenic edema. Controversy has focused on the respective roles of increased

pinocytotic transport, opening of interendothelial tight junctions, and mediators affecting vascular permeability. Pinocytosis refers to the vesicular transport of material across the endothelial cell layer. Cyclic nucleotide-generating enzymes in the endothelial cytoplasm that are linked to receptors on the cell surface trigger this process. Leaky interendothelial tight junctions allow for the exudation of plasma proteins with associated fluid shifts. Extravasation of blood cells occurs only when severe damage occurs. Extravasation of fluid with engorgement of the interstitial space undoubtedly interferes with structural elements of the parenchyma and may lead to further injury due to impaired diffusion of essential substances. The mediator concept of vasogenic edema theorizes that autacoids released from injured brain tissue induce increased vascular permeability. Numerous substances illustrate this capability. Bradykinin and histamine alter the integrity of the BBB by opening tight junctions, increasing vesicular transport, and simultaneously producing relaxation of cerebral arteries with resultant increases in cerebral blood flow. The combination of these effects results in cerebral edema. The effects of bradykinin and histamine have been linked to stimulation of specific receptors. Elevations of bradykinin and histamine also occur in brain tissue following trauma, ischemia, and experimental cold-induced injury. Although inhibition of bradykinin reduces cerebral edema, similar antagonism of histamine has failed to produce a similar effect. Arachidonic acid has also been implicated in the pathogenesis of vasogenic edema. This substance is a free fatty acid that is a major component of brain cells and subcellular organelles. The liberation of arachidonic acid with resultant increases in concentration has been demonstrated following trauma and ischemia, although the mechanism of subsequent BBB breakdown is unclear. Arachidonic acid has a variable effect on the cerebral vasculature. Glucocorticoid inhibition of arachidonic acid production may reduce cerebral edema, but this effect seems to be dependent on etiology. Atrial natriuretic factor (ANF) is a peptide that relaxes smooth muscle and increases renal excretion of sodium and water. Animal models have demonstrated ANF to be effective in reducing cerebral edema. Glucocorticoids are able to regulate the expression of the ANF gene, suggesting a possible role in the response of cerebral edema to steroid therapy.

The roles of nitric oxide and oxygen-derived free radicals have also been the focus of much attention. A free radical is defined as an atom or molecule with an unpaired electron in its outer orbit. Oxygen free radicals play a key role in the development of vasogenic edema. The deleterious effects of oxygen-derived free radicals cause secondary injury following ischemia and head trauma. These substances participate in an iron-catalyzed reaction that leads to lipid peroxidation. The extravasation and subsequent degradation of red blood cells liberate heme compounds and iron, promoting this process. The scavenging action of superoxide dismutase limits the cascade of free radical production. The vasodilatory effects of nitric oxide are well established, and endothelial nitric oxide production causes disruption of capillary integrity. Despite these findings, nitric oxide has not been shown to produce cerebral edema and inhibition of this substance in experimental models of cerebral edema has yielded discrepant results.

2. Cytotoxic Edema

In contrast to the initial events that characterize vasogenic edema, cytotoxic edema occurs in the absence of permeability changes at the BBB. Cytotoxic edema is typified by intracellular accumulation of fluid associated with toxin exposure, ischemia, head trauma, or Reye's syndrome. The experimental animal model of triethyl tin intoxication has been used as the prototypical example of this condition. Cellular swelling affects neurons, glia, and endothelial cells, although the majority of research focuses on astrocytic swelling. The mechanism of astrocytic swelling is complex, involving the activation of ion channels and free radical injury to the cellular membrane. Cytotoxic edema is most commonly precipitated by failure of the Na^+/K^+ ATPase channel due to energy depletion. Secondary events are associated with mediators including glutamate, potassium, free fatty acids, and lactoacidosis. Under normal conditions, the permeability of the astrocytic membrane to sodium is very low and must involve the activation of specific transporters. Several protective mechanisms contribute to an increase in the intracellular sodium concentration when the extracellular fluid composition becomes altered. Increased neuronal metabolic activity and other causes of acidosis encourage the astrocytic uptake of sodium due to activation of the Na^+/H^+ channel. Astrocytes also accumulate sodium in a dependent fashion because of glutamate clearance from the extracellular space. These increases in intracellular sodium result in intracellular fluid accumulation, most prominently at the endfeet that contribute to the BBB. Despite the relative abundance

of oxygen and nutrients at this location, the high density of Na^+/H^+ and Na^+/HCO_3^- cotransport systems can predispose this region to early manifestations of swelling. Astrocytic swelling may have deleterious effects due to the release of excitatory amino acids. Excitatory amino acids, including glutamate, taurine, and aspartate, are released through reversal of normal uptake systems and swelling-induced efflux via anion channels. The release of glutamate stimulates neurons and leads to increases in intracellular calcium and subsequent disruption of normal cellular architecture. Excessive stimulation of excitatory receptor sites may lead to further neuronal injury.

3. Hydrocephalic Edema

Obstruction of CSF outflow from the ventricular system results in hydrocephalus with ventricular dilatation. Progressive CSF accumulation exerts pressure across the ependymal lining of the ventricles. At a critical stage, this force causes the leakage of CSF across the ependyma into the interstitial space. For this reason, hydrocephalic edema is often referred to as interstitial edema. The periventricular white matter tracts are usually involved, whereas the BBB remains intact.

4. Ischemic Edema

Cerebral ischemia triggers a cascade of molecular interactions that culminate in a combination of cytotoxic and vasogenic edema referred to as ischemic edema. This composite of edema patterns evolves with the duration of ischemia. Immediately after the onset of cerebral ischemia, cytotoxic edema develops. The critical step in this process is energy depletion leading to failure of the Na^+/K^+ ATPase pump. The initial increase in astrocytic water content may be transient and fully reversible if reperfusion occurs. If ischemia persists, increasing astrocytic swelling jeopardizes the integrity of the BBB. In cases of permanent ischemia, the breakdown of the BBB occurs after a delay. At the periphery of the ischemic territory, vasogenic edema predominates because acidosis and loss of autoregulatory function produce a reactive hyperemia. Temporary cerebral ischemia results in a biphasic disturbance of the BBB. After a delay, a secondary opening of the BBB occurs due to the release of mediators on the abluminal side from ischemic parenchyma. These substances include kinins, free fatty acids, free radicals, serotonin, and prostaglandins. Early disruption of the BBB is proportional to the intensity of reactive hyperemia. Subsequent changes in vascular permeability also impair the clearance of materials from brain parenchyma. The accumulation of lactic acid may be an important variable in the severity of postischemic edema.

5. Hydrostatic Edema

Elevated hydrostatic pressure due to arterial hypertension and vasodilation may result in increased fluid flux across the capillary wall. Dysautoregulation of the cerebrovasculature limits the ability of blood vessels to vasoconstrict in response to arterial pressure elevations. A transudative fluid with low protein content accumulates in the interstitial space. This process is normally prevented by the opposing ICP and osmotic pressure gradients that retain fluid in the vascular space. The most common clinical example of hydrostatic edema formation is hypertensive encephalopathy. In this condition, intraarterial pressure exceeds the upper limits of cerebral autoregulation. Conditions of systemic hypoosmolarity decrease the osmotic forces that normally retain fluid in the vascular space. Similarly, abrupt decreases in ICP (e.g., mass resection) may precipitate edema formation as the result of acute changes in transmural hydrostatic forces.

6. Osmotic Edema

An alteration in the osmotic pressure gradient with relative hyperosmolarity of the brain parenchyma with respect to plasma encourages the accumulation of fluid in the brain. The development of osmotic edema requires integrity of the BBB. An osmotic discrepancy commonly results from hyponatremic states including water intoxication, the syndrome of inappropriate antidiuretic hormone (SIADH) secretion, cerebral salt wasting, and iatrogenic error with inappropriate administration of hypotonic solution. Although sodium is usually the primary determinant of serum osmolarity, other metabolic derangements may produce a state of relative hyperosmolarity within the brain parenchyma and subsequent formation of osmotic edema. These conditions include dialysis disequilibrium, diabetic ketoacidosis, and rapid correction of hypernatremia. Hyperosmolarity of brain parenchyma has been attributed to the generation of idiogenic osmoles as a protective measure during systemic hyperosmolar states. The temporal profile of changes in osmotic pressure gradients appears to be a critical factor in the development of osmotic edema.

II. CLINICAL MANIFESTATIONS

Cerebral edema is associated with a wide spectrum of clinical disorders. Edema can either result from regional abnormalities related to primary disease of the central nervous system or be a component of the remote effects of systemic toxic–metabolic derangements. In either scenario, cerebral edema may be a life-threatening complication that deserves immediate medical attention. Several challenges surround the management of cerebral edema, because the clinical presentation is extremely variable. This variability reflects the temporal evolution of a diverse combination of edema types because most forms of cerebral edema have the capacity to generate other types. The specific clinical manifestations are difficult to categorize by type and are better described by precipitating etiology. In other words, it is essential to outline the prominent forms of edema that are present in a given clinical scenario. The location of edema fluid determines symptomatology. Focal neurologic deficits result from isolated regions of involvement, whereas diffuse edema produces generalized symptoms such as lethargy. The following sections address the clinical manifestations of cerebral edema that are associated with a diverse range of clinical disorders, first describing primary pathology of the central nervous system followed by the neurologic consequences of systemic disorders.

A. Cerebrovascular Disease

Cerebral ischemia frequently causes cerebral edema. Global ischemia may result from cardiac arrest or sustained hypotension, whereas focal ischemia may be the product of thromboembolic occlusion of a cerebral artery. Tissue hypoxia that results from ischemic conditions triggers a cascade of events that leads to cellular injury. The immediate effect of hypoxia is energy depletion with resultant failure of the Na^+/K^+ ATPase pump. The onset of ischemic edema initially manifests as glial swelling occurring as early as 5 min following interruption of the energy supply. This cytotoxic phase of edema occurs when the BBB remains intact, although continued ischemia leads to infarction and the development of vasogenic edema after 48–96 hr. The development of vasogenic edema usually occurs only after reperfusion or in the setting of transient ischemia. Elevations of systemic arterial pressure with a loss of cerebral autoregulation may further the development of cerebral edema by increasing hydrostatic pressure gradients across a disrupted BBB. Conversely, a reduction in CPP coupled with changes in the microcirculation, including elevation of cerebrovascular resistance, may induce detrimental effects by limiting the distribution of essential nutrients. Secondary ischemia may also result from BBB breakdown or from the effects of free radical-related amplification of neuronal injury. Clinical symptoms are initially representative of neuronal dysfunction within the ischemic territory, although the spread of edema may elicit further neurological deficits in patients with large hemispheric infarction. This clinical syndrome involves increasing lethargy, asymmetrical pupillary examination, and abnormal breathing. The mechanism of neurologic deterioration appears to involve pressure on brain stem structures due to the mass effect of infarcted and edematous tissue. Elevation of ICP may be generalized or display focal gradients that precipitate herniation syndromes. Herniation may lead to compression and infarction of other vascular territories, in turn initiating a new cycle of infarction and edema.

Hemorrhagic infarction is frequently the cause of elevations in ICP, but the contribution of cerebral edema has been controversial. Intracerebral hemorrhage presents with focal neurologic deficits, headache, nausea, vomiting, and/or evidence of mass effect. The edema associated with intracerebral hemorrhage is predominantly vasogenic (Fig. 3), climaxing 48–72 hr following the initial event. Secondary ischemia with a component of cytotoxic edema may result from impaired diffusion in the extracellular space of the perihemorrhage region. Expansion of the hematoma volume or extension of edema may precipitate cerebral herniation. Figure 4 demonstrates uncal herniation that resulted from expansion of a hematoma. Other forms of hemorrhage, including hemorrhagic transformation of ischemic territories and subarachnoid hemorrhage, may be associated with edema that results from the noxious effects of blood degradation products.

Cerebral venous thrombosis may lead to venous stasis and elevations of cerebral venous pressure. These conditions are attributable to the absence of an autoregulatory mechanism within the cerebral venous circulation. The effect of elevated venous pressures is therefore greater than similar increases in the arterial system. The clinical manifestations of cerebral venous thrombosis are highly variable. Individuals may be asymptomatic, and others may suffer a progressive neurologic deterioration with headaches, seizures, focal neurologic deficits, and severe

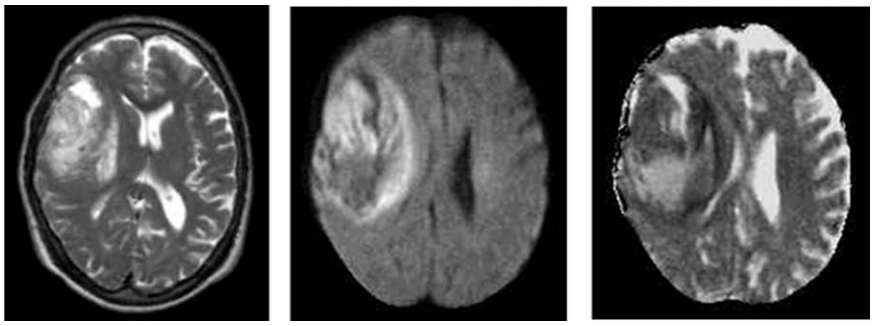


Figure 3 Intracranial hemorrhage depicted by MRI. T2-weighted sequence showing hyperintensity associated with vasogenic edema in the right frontal lobe.

obtundation leading to death. Parenchymal venous congestion and elevation of venous pressures can lead to disruption of the BBB, the formation of vasogenic edema, and perivascular hemorrhage. Further increases in venous pressure lead to a reversal of the normal hydrostatic pressure gradients, and thus reduce capillary perfusion pressure. Fluid distension of the extracellular space also impairs the diffusion of nutrients. These changes induce the formation of cytotoxic edema that may ultimately result in ischemic venous infarction (Fig. 5).

Figure 4 Coronal brain slices illustrating uncal herniation due to hematoma expansion. (Courtesy of Harry V. Vinters, M.D.)

B. Traumatic Brain Injury

Focal or diffuse cerebral edema of mixed types may develop following traumatic brain injury (TBI). Following contusion of the brain, the damaged BBB permits the extravasation of fluid into the interstitial space. Areas of contusion or infarction may release or induce chemical mediators that can spread to other regions. These potentially deleterious mediators include lysosomal enzymes, biogenic amines including serotonin and histamine, excitatory amino acids, arachidonic acid, and components of the kallikrein–kinin system. These factors activated during tissue damage are powerful mediators of extravasation and vasodilation. TBI is associated with a biphasic pathophysiologic response heralded by a brief period of vasogenic edema immediately following injury, followed after 45–60 min by the development of cytotoxic edema. Vasogenic edema may be detected by neuroimaging modalities within 24–48 hr and reach maximal severity between Days 4 and 8. Although cerebral blood flow decreases for the initial few hours following head injury, a reactive hyperemic response that exceeds metabolic demands may exacerbate vasogenic edema. Autoregulatory dysfunction is a common sequela of TBI that may promote the formation of hydrostatic edema in regions where the BBB remains intact.

Recent efforts have also demonstrated a prominent role of cytotoxic edema in head-injured patients.

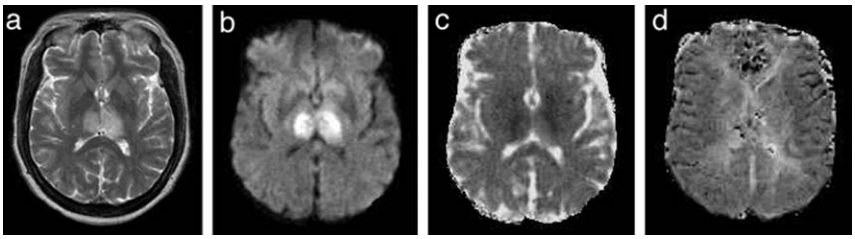


Figure 5 Multimodal MRI of cerebral venous thrombosis. Ischemic edema of the thalami manifest as hyperintensity on (a) T2-weighted sequences and (b) DWI. (c) ADC hypointensity corresponds to the cytotoxic component. (d) Focal hyperintensity suggests increased regional cerebral blood volume.

Tissue hypoxia with ischemic edema formation and neurotoxic injury due to ionic disruption contribute to this cytotoxic component. In addition, osmotic edema may result from hyponatremia, and hydrocephalic edema may complicate the acute phase of TBI when subarachnoid hemorrhage or infections predominate. Conspicuous diffuse axonal injury may produce focal edema in white matter tracts experiencing shear-strain forces during acceleration/deceleration of the head (Fig. 6). Arterial hypotension may aggravate the clinical expression of cerebral edema. Episodic hypotension is a common occurrence in the head-injured patient, exacerbated by hypovolemia and the cardiodepressant effect of sedative medications. When autoregulation is preserved, arterial blood pressure may have a significant effect on ICP levels. As CPP decreases with arterial hypotension, compensatory cerebral vasodilation leads to increased cerebral blood volume and thereby raises ICP. Conversely, therapeutic maintenance of CPP encourages vasoconstriction with subsequent reduction in ICP. These changes in cerebral blood volume influence the severity of edema through modulation of the overall water content of brain tissue. Although the degree of edema has not been directly correlated with outcome, ICP elevations illustrate the ability to predict poor clinical outcome. Raised ICP attributed to cerebral edema is the most frequent cause of death in TBI. The effects of herniation due to a traumatic epidural hematoma and diffuse axonal injury are demonstrated in Fig. 7. The pathogenesis of cerebral edema in the immature brain stands is the subject of much controversy in attempts to address the relatively higher frequency and severity of edema in pediatric head-injury patients. Sustained posttraumatic cerebral hypoperfusion and a greater susceptibility to excitoxic injury have been suggested as alleged factors in this process.

C. Neoplastic Disease

The detrimental effects of cerebral edema considerably influence the morbidity and mortality associated with brain tumors. Cerebral edema commonly occurs with brain tumors and frequently hampers the postoperative course of patients undergoing resection. Postoperative edema may worsen after partial resection, climaxing 36–72 hr later. This exacerbation results

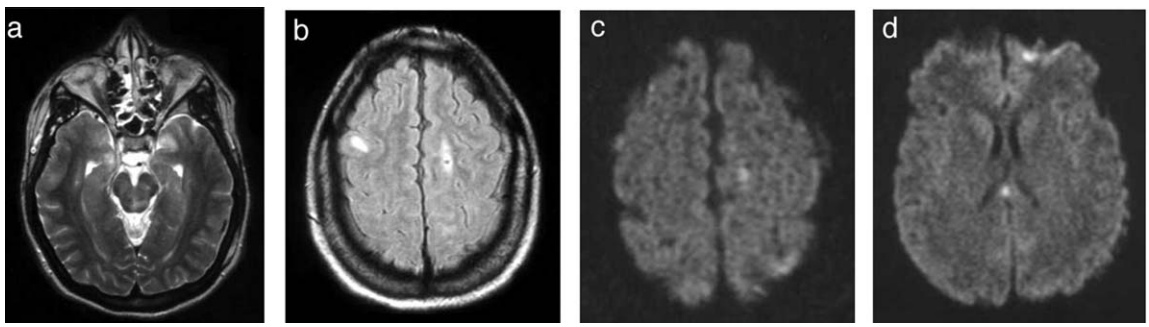


Figure 6 Diffuse axonal injury associated with acute TBI demonstrated by MRI. Hyperintensity of (a) the midbrain on T2-weighted sequences, the frontal lobe white matter on (b) FLAIR imaging and (c) DWI, and (d) the corpus callosum on DWI reveal injury to vulnerable sites.

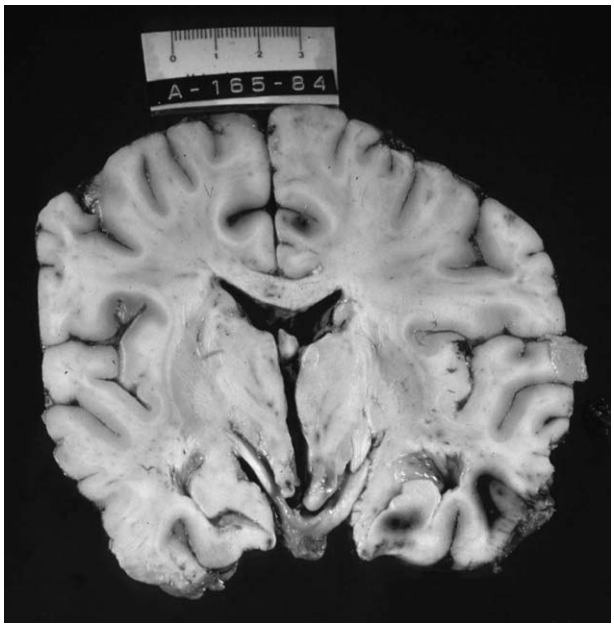


Figure 7 Small punctate hemorrhages due to diffuse axonal injury and evidence of midline shift due to epidural hemorrhage are illustrated on coronal section of gross brain specimen. (Courtesy of Harry V. Vinters, M.D.)

from either hydrostatic pressure changes or a reaction to the distribution of tumor cells and necrotic debris. Chemotherapeutic agents and radiation have also been shown to disrupt the BBB and encourage the formation of edema. The introduction of steroids decreased the operative mortality rate of brain tumors by a factor of 10. Despite frequent references to peritumoral edema, alterations in tissue permeability with fluid accumulation may occur at distances remote from the actual lesion. Tumor-associated edema continues to be a formidable challenge, producing symptoms such as headache and focal neurologic deficits and, occasionally, considerably altering the clinical outcome. The predominant form of tumor-associated edema is vasogenic, although cytotoxic edema may occur through secondary mechanisms, such as tumor compression of the local microcirculation or tissue shifts with herniation. Individuals with hydrocephalus can also develop hydrocephalic edema because of ventricular outflow obstruction. Histologic studies of tumor-associated edema reveal pallor of astrocytes, with swollen and vacuolated myelin sheaths within white matter tracts, whereas gray matter is affected to a lesser degree. Ultrastructural studies indicate that tumor microvessels that significantly differ from normals suggest a possible explanation for the formation of vasogenic edema. Endothelial cells are plump and contain an increased proportion of microvilli, a higher density of pinocytic vesicles, and no tight junctions. Tumor cells and their secretory products alter vascular permeability and induce angiogenesis, the formation of new capillary blood vessels. This complex cascade of events involving biochemical and morphological alterations is guided by the interaction of numerous factors, including fibroblast growth factor, angiogenin, TNF-α, transforming growth factors-α and -β, platelet-derived endothelial growth factor, interleukin-8, and VEGF/VPF. VEGF/VPF interacts with a receptor that is amplified in tumor vessels to cause transient increases in vascular permeability. Immunologic mechanisms may cause cellular injury, thereby modifying microvascular permeability. Inflammatory processes act by macrophage and polymorphonuclear leukocyte release of specific proteolytic enzymes, specialized granules, and PAFs. Superoxide anion radicals and products of the arachidonic acid cascade, including prostaglandins, thromboxane, and leukotrienes, further increase the vascular permeability underlying the vasogenic component of tumor-associated edema. The degree of tumor-associated edema varies with the pathology of brain tumors. Metastatic lesions, most gliomas, and meningiomas cause extensive edema, whereas oligodendrogliomas cause less tumor-associated edema. Diffuse edema associated with glioblastoma multiforme is demonstrated in Fig. 8.

D. Hydrocephalus

Isolated hydrocephalic edema may result from acute obstructive hydrocephalus with impairment of CSF drainage. Transependymal pressure gradients result in edema within periventricular white matter tracts. The rapid disappearance of myelin lipids under pressure causes the periventricular white matter to decrease in volume. The clinical manifestations may be minor, unless progression to chronic hydrocephalus becomes apparent with symptoms including dementia and gait abnormalities.

E. Infections

A combination of vasogenic and cytotoxic edema arises from many infectious processes within the central nervous system. Other forms of edema may also occur in infections, including hydrocephalic edema secondary to CSF obstruction and osmotic

Figure 8 Axial section of a gross brain specimen revealing diffuse edema surrounding the necrotic core of a glioblastoma multiforme. (Courtesy of Harry V. Vinters, M.D.)

edema due to SIADH. Numerous infectious agents have direct toxic effects generating vasogenic edema through alteration of the BBB and cytotoxic edema from endotoxin-mediated cellular injury. Bacterial wall products stimulate the release of various endothelial factors, resulting in excessive vascular permeability. This injurious effect is augmented by a hyperemic response within the first 24–48 hr that elevates ICP to precarious levels. Associated immunologic and inflammatory responses can also engender cerebral edema. Cerebral edema is a critical determinant of morbidity and mortality in pediatric meningitis. Abscess formation or focal invasion of the brain results in an isolated site of infection surrounded by a perimeter of edema encroaching on the neighboring parenchyma. This ring of vasogenic and cytotoxic edema may produce more symptoms than the actual focus of infection. Similar regions of focal or diffuse edema may accompany encephalitis, particularly viral infections such as herpes simplex encephalitis.

F. Hypertensive Encephalopathy

Hypertensive encephalopathy is a neurologic emergency characterized by exceedingly elevated blood pressures with the subsequent development of focal cerebral edema. This potentially reversible condition presents with rapidly progressive neurological signs, headache, seizures, altered mental status, and visual disturbances. The clinical and neuroradiologic presentation of hypertensive encephalopathy shares similarity with the reversible posterior leukoencephalopathy syndrome, a clinical entity that encompasses multiple disorders with a common predisposition for the development of focal cerebral edema. The pathogenesis of edema formation is controversial but is thought to involve elevated hydrostatic forces due to excessive blood pressure, with lesser degrees of involvement attributed to vasogenic edema and secondary ischemic components. The contributions of autoregulatory dysfunction and BBB leakage compared to hydrostatic causes of plasma extravasation continue to be controversial. The rate of blood pressure elevation is a critical factor, because hypertensive encephalopathy usually develops during acute exacerbations of hypertension. Chronic hypertension has induced adaptive modifications of cerebral resistance vessels, thereby minimizing the harmful effects of elevated hydrostatic forces. Early recognition and treatment of hypertensive encephalopathy may reverse cerebral edema, preventing permanent damage to the BBB, and ischemia, although severe cases may be fatal.

G. Seizures

Recent advances in neuroimaging have yielded additional information regarding the development of cerebral edema associated with seizures. Although cerebral edema is unlikely to be clinically manifest following most seizure activity, additional pathophysiologic insight has been gleaned from recent reports of magnetic resonance imaging (MRI) in status epilepticus (Fig. 9). Prolonged seizure activity may lead to neuronal energy depletion with eventual failure of the Na^+/K^+ ATPase pump and concomitant development of cytotoxic or ischemic edema. Unlike ischemia produced by occlusion of a cerebral artery, a more heterogeneous cellular population is affected. The reactive hyperemic response driven by excessive metabolic demands increases the hydrostatic forces across a BBB already damaged by the vasogenic component of ischemic edema. The disruption of normal ionic

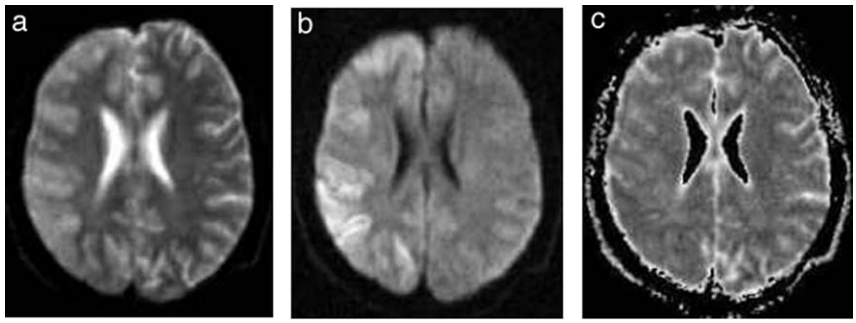


Figure 9 MRI of status epilepticus reveals evidence of cytotoxic edema within cortical structures, illustrated by (a) T2-weighted and (b) DWI hyperintensity, with (c) mild hypointensity on ADC maps.

gradients, extracellular accumulation of excitotoxic factors, and lactic acidosis further exacerbate vasogenic edema. Consequently, cessation of seizure activity usually results in the complete resolution of cerebral edema.

H. Hyponatremia and Hypernatremia

Altered osmotic homeostasis commonly results from hyponatremia and hypernatremia. Osmotic cerebral edema relates to hypoosmolar states such as hyponatremia and the resolution phase of hyperosmolar conditions, including hypernatremia. Hyponatremia may present with alterations in mental status with obtundation and seizures, although there appears to be wide individual variability in adaptation to hyponatremic states. The relative hyperosmolarity of brain parenchyma encourages osmotic edema formation. Dehydration or excessive sodium lead to the compensatory generation of idiogenic osmoles in the brain to maintain cellular volume. Clinical symptoms of hypernatremia include mental status changes with irritability, restlessness, muscle twitching, hyperreflexia, and seizures. During recovery from hypernatremia, the slower dissipation of idiogenic osmoles leads to excessive water shifts into the brain parenchyma. During hyponatremia and the resolution stage of hypernatremia, the rate at which osmotic homeostasis is restored is a critical determinant of clinical sequelae.

I. Dialysis Disequilibrium

The abrupt onset of neurologic symptoms following hemodialysis has been attributed to cerebral edema. Although much speculation focuses on differential fluid shifts related to a rapid decrease in blood urea nitrogen, the specific cause of this osmotic edema remains controversial. A reverse urea effect theorizes that the BBB prevents urea concentrations in the brain from decreasing with serum levels during and immediately following dialysis. A relative hyperosmolar state within the brain encourages the formation of osmotic edema. This fluid shift depends on the rate of dialysis and may be reversed by increasing serum osmolarity. The reverse urea effect hypothesis has been disputed and recent attention has focused on the role of idiogenic osmoles, which has been suggested because urea levels do not fully explain changes in CSF and serum osmolarity measurements. The generation of these organic acids as a protective mechanism against dehydration has been speculated to account for postdialysis decreases in intracellular pH. The clinical manifestations are usually transient and may be avoided or minimized by employing slower rates of solute removal, increasing plasma osmolarity with mannitol, or modifying dialysis solutions.

J. Diabetic Ketoacidosis

Diabetic ketoacidosis may be complicated by rapid neurologic deterioration with severe brain injury due to osmotic cerebral edema. Hyperglycemia produces serum hyperosmolarity with the generation of idiogenic osmoles within brain cells in an effort to prevent cell shrinkage. During the treatment of diabetic ketoacidosis, rehydration with hypotonic solution and administration of insulin result in dramatic fluctuations in osmotic gradients. The osmoprotective molecules within brain cells result in the rapid accumulation of water with cellular volume expansion. The duration of diabetic ketoacidosis may influence

the degree of idiogenic osmole formation. Unfortunately, the consequences of cerebral edema in this setting may be severe with a high mortality rate. Poor outcomes may be avoided by conservative treatment of hypovolemia employing isotonic solution, gradual reduction in serum glucose, and maintenance of normal serum sodium concentrations adjusted for hyperglycemia.

K. Fulminant Hepatic Failure

Intracranial hypertension secondary to cerebral edema frequently complicates the course of patients with fulminant hepatic failure. The devastating effects of cerebral edema contribute to the high mortality of this condition, with cerebral edema noted in 50–80% of cases at autopsy. Although the progressive obtundation of fulminant hepatic failure may be confused with hepatic encephalopathy, important differences exist. The rapid demise associated with fulminant hepatic failure results from a combination of cytotoxic and vasogenic edema (Fig. 10), whereas edema is not seen in cirrhotic patients dying in hepatic coma. The pathogenesis of cerebral edema formation in fulminant hepatic failure has been attributed to the detrimental effects of ammonia. Due to an incomplete urea cycle in the brain, astrocytes detoxify ammonia to glutamine. Excessive production of glutamine leads to excitotoxic neuronal damage and astrocytic swelling, although osmotic compensation occurs in chronic liver disease. This cytotoxic process is aggravated by an influx of water resulting from a combination of augmented cerebral blood flow and disturbed autoregulation. Additional toxic injury is incurred by circulating endotoxins and cytokines, with excessive production of nitric oxide. The necrotic liver releases vasoactive substances producing dramatic cardiovascular instability. These fluctuations in cardiovascular function can complicate surgical transplantation, but autonomic stability is restored following removal of the diseased liver.

L. Toxic Encephalopathy

Although a wide spectrum of substances have toxic effects that can ultimately result in cerebral edema, a direct causal relationship has been established for only a few toxins of clinical significance. Children and adolescents appear to be particularly vulnerable to these toxic exposures. Adolescents ingesting large doses of vitamin A may develop pseudotumor cerebri and visual loss. Organic tin compounds used for the treatment of furunculosis and the use of hexachlorophene as a disinfectant may precipitate cytotoxic edema. Acute intoxication with aluminum, copper, or zinc may mimic dialysis disequilibrium. The majority of substances have been associated with cytotoxic edema, although lead encephalopathy is characterized by diffuse vasogenic edema. Lead encephalopathy causes diffuse vascular injury followed by extravasation of protein-rich fluid into the extracellular space. These changes correlate with the diffuse enhancement noted on contrast computed tomography (CT).

M. Reye's Syndrome

Reye's syndrome manifests with the development of lethargy, irritability, and progressive obtundation in children and young adults following viral infections. The neurologic symptoms have been attributed to diffuse cytotoxic edema documented at autopsy. The

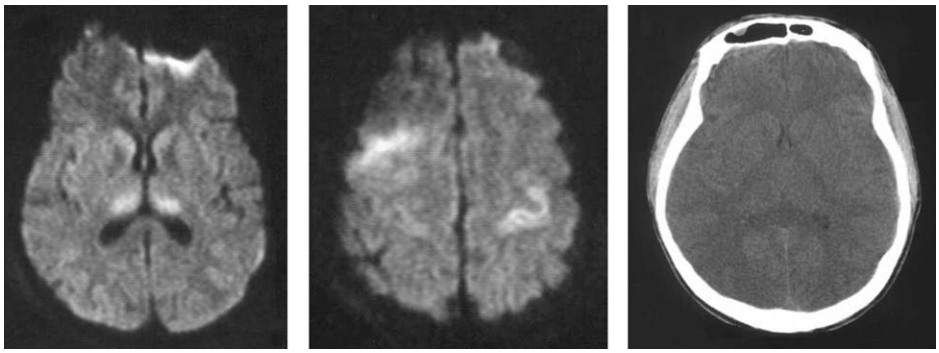


Figure 10 Cytotoxic edema of fulminant hepatic failure illustrated by DWI hyperintensity in (a) bilateral thalami and (b) bilateral frontal cortices. Diffuse vasogenic edema is demonstrated (c) by CT 3 days later.

pathogenesis of this syndrome revolves around the associated hepatic impairment. Toxic effects related to fatty infiltration of the liver and elevated serum ammonia are suspect. Aspirin therapy for fever due to influenza and other viral infections are also implicated as precipitating factors, although the underlying pathophysiology is uncertain.

N. High-Altitude Cerebral Edema

An uncommon presentation of vasogenic edema occurs at high altitudes. Acute mountain sickness and high-altitude cerebral edema (HACE) encompass a spectrum of changes initiated by hypoxia associated with acclimatization. The symptoms of headache, gastrointestinal disturbance, dizziness, fatigue, and insomnia occasionally progress to obtundation and death. The pathophysiology of this potentially reversible form of vasogenic edema is highly complex. Upon arrival at higher altitudes, compensatory elevation of cerebral blood flow with vasodilation gradually decreases over 2 or 3 days in response to increasing arterial oxygen saturation. The hypoxic stimulus also activates angiogenic factors, including VEGF and nitric oxide, leading to increased vascular permeability. The reversible vasogenic edema of white matter structures can evolve to include secondary cytotoxic edema resulting from ischemia. The administration of oxygen and osmotic diuresis may help to decrease the severity of the vasogenic edema, although it may take days for resolution of severe HACE.

O. Hyperthermia

The pathophysiology of this rare cause of cerebral edema is poorly understood. Although the fatal consequences of heat stroke have been recognized since ancient times, the underlying mechanisms await clarification. Scant pathologic material suggests a combination of cytotoxic and vasogenic components, secondary to an increase in BBB permeability due to the release of multiple chemical factors and direct cytotoxic damage. Age and physiologic state of the individual appear to be important determinants of clinical outcome in hyperthermic injury.

III. DIAGNOSIS

Prior to the advent of sophisticated neuroimaging techniques, a precise diagnosis of cerebral edema relied on the histopathologic examination of brain tissue by biopsy or at autopsy. Figure 11 illustrates the gross appearance of diffuse cerebral edema noted at autopsy. Although a demonstration of increased parenchymal water content confirmed the presence of cerebral edema, the underlying pathogenesis was largely obscured by the ultimate changes prior to tissue sampling. The invasive nature of brain biopsy limited the study of the dynamic aspects of cerebral edema. Even the inferential evidence derived from detailed neuroradiographic studies including cerebral angiography and ventriculography demanded an invasive procedure. Radionucleide scanning introduced a diagnostic tool for the demonstration of vasogenic edema of various etiologies. The development of CT further refined the delineation of cerebral edema. Unfortunately, the characteristic CT hypodensity of edematous brain parenchyma may be similar to the appearance of chronic infarction and lipid-containing

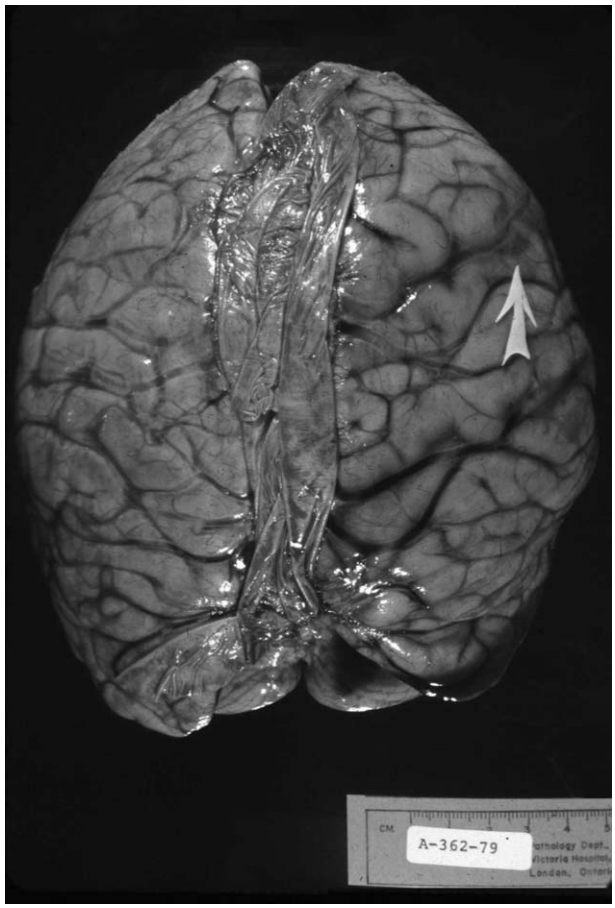


Figure 11 Diffuse cerebral edema with swollen gyri and focal cortical contusion (arrow) noted at autopsy. (Courtesy of Harry V. Vinters, M.D.)

structures. The introduction of MRI revolutionized the understanding of cerebral edema, allowing for the serial study of dynamic pathophysiologic changes associated with various types of edema. Current diagnostic approaches are based on measures that assess ICP, CT, and advanced MRI techniques when available.

A. Intracranial Pressure Monitoring

ICP elevation can infer the presence of cerebral edema. Symptoms of elevated ICP include headache, nausea, vomiting, and altered level of consciousness, which may be superimposed on focal neurologic deficits. When ICP reaches a critical threshold, herniation syndromes may ensue. The initial presentation of this clinical scenario demands an emergent CT scan of the head to identify the underlying process and assess the degree of parenchymal injury. Although neuroimaging methods routinely diagnose the presence of cerebral edema, ICP monitoring devices may detect the evolution of edema in patients expected to develop this complication. A diagnostic lumbar puncture may be performed safely on most patients whose imaging does not indicate a mass lesion to identify elevations of CSF pressure. ICP monitoring by ventriculostomy can be employed to permit drainage of CSF in hydrocephalic conditions. Continuous ICP monitoring is also possible with subarachnoid bolts, fiberoptic catheters, and subdural and epidural devices. These instruments may reflect focal pressure fluctuations and thus may not correlate with global changes in ICP. The utility of additional diagnostic measures such as electro encephalograph, transcranial Doppler ultrasonography, and somatosensory-evoked potential monitoring currently remains under investigation.

B. Computed Tomography

In contrast to the indirect diagnostic information provided by ICP measurements, CT technology may noninvasively illustrate the volumetric changes and alterations in parenchymal density resulting from cerebral edema. Expansion of brain tissue due to most forms of edema may be detected on CT, although diffuse processes including fulminant hepatic failure or osmotic edema may be more difficult to discern. Diffuse swelling may be recognized by a decrease in ventricular size with compression or obliteration of the cisterns and cerebral sulci. Cellular swelling associated with cytotoxic and ischemic edema can manifest as subtle enlargement of tissue with obscuration of normal anatomic features, such as the differentiation between gray matter and white matter tracts. Vasogenic edema may also cause tissue expansion, although the associated density changes may be more prominent. In contrast to the increased volume of brain tissue noted in most forms of edema, hydrocephalic edema may be suspected in cases in which ventricular expansion has occurred. Extensive volumetric changes and the associated pressure differentials resulting in herniation may be noted on CT as shifts in the location of various anatomic landmarks.

The increased water content associated with edema causes the density of brain parenchyma to decrease on CT. The attenuation effects of other tissue contents complicate precise correlation of water content with density on CT. Although slight decrements in tissue density result from cytotoxic and osmotic processes, more conspicuous areas of hypodensity result from the influx of fluid associated with disruption of the BBB in vasogenic edema. The injection of a contrast agent permits enhanced visualization of vascular structures and thereby further defines regions of extravasation where the BBB has been altered. Contrast CT improves the demonstration of infectious lesions and tumors that present with significant degrees of vasogenic edema.

The differentiation of specific forms of edema is limited with CT, but this modality may provide sufficient information to guide therapeutic decisions in many situations. CT perfusion, a new dynamic imaging study, allows for the determination of cerebral blood volume based on serial density changes associated with a contrast bolus injection. CT may be inferior to MRI in the characterization of cerebral edema, but logistic constraints may preclude MRI in unstable trauma patients, uncooperative patients, and patients with contraindications due to the presence of metallic implants or pacemakers.

C. Magnetic Resonance Imaging

Volumetric enlargement of brain tissue due to edema is readily apparent on MRI and the use of gadolinium, an MRI contrast agent, enhances regions of altered BBB. Differences in water content may be detected on MRI by variations in the magnetic field generated primarily by hydrogen ions. T2-weighted sequences

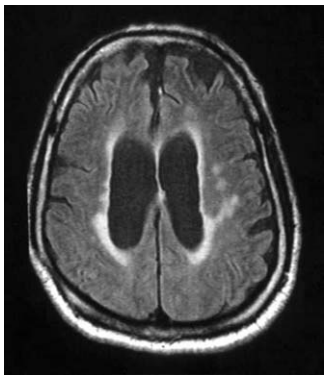


Figure 12 Periventricular FLAIR hyperintensity due to hydrocephalic edema.

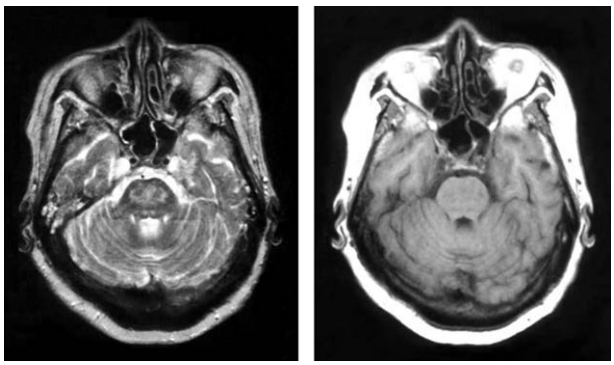


Figure 13 Central pontine myelinolysis illustrated as (a) T2-weighted hyperintensity and (b) T1-weighted hypointensity in the pons.

and fluid-attenuated inversion recovery (FLAIR) images reveal hyperintensity in regions of increased water content. Reversible T2 hyperintensities of the white matter may be seen in HACE and hypertensive encephalopathy. FLAIR images eliminate the bright signal from CSF spaces and are therefore helpful in characterizing periventricular findings such as hydrocephalic edema (Fig. 12). These conventional MRI sequences are more sensitive in the detection of lesions corresponding to hypodensities on CT. MRI is also superior in the characterization of structures in the posterior fossa (Fig. 13). Recent advances in MRI technology make it possible to specifically discern the type of edema based on signal characteristics of a sampled tissue volume. This discriminatory capability resulted from the development of diffusion imaging techniques. The use of strong magnetic field gradients increases the sensitivity of the MR signal to the random, translational motion of water protons within a given volume element or voxel. Cytotoxic edema and cellular swelling produce a net decrease in the diffusion of water molecules due to the restriction of movement, imposed by intracellular structures such as membranes and macromolecules, and diminished diffusion within the extracellular space due to shrinkage and tortuosity. In contrast, the accumulation of water within the extracellular space as the result of vasogenic edema allows for increased diffusion. Diffusion-weighted imaging (DWI) sequences yield maps of the brain, with regions of restricted diffusion appearing bright or hyperintense. The cytotoxic component of ischemic edema has been demonstrated on DWI within minutes of ischemia onset. DWI has also been used to grade the severity of TBI, because cytotoxic injury is thought to accompany severe trauma. Apparent diffusion coefficient (ADC) maps may be generated from a series of DWI images acquired with varying magnetic field gradients. ADC elevations, resulting from vasogenic edema, appear hyperintense on ADC maps, whereas decreases in ADC due to cytotoxic edema appear hypointense (Fig. 14). These maps may be sampled to measure the ADC of a given voxel for multiple purposes, such as differentiating tumor from tumor-associated edema. The vector of water diffusion in

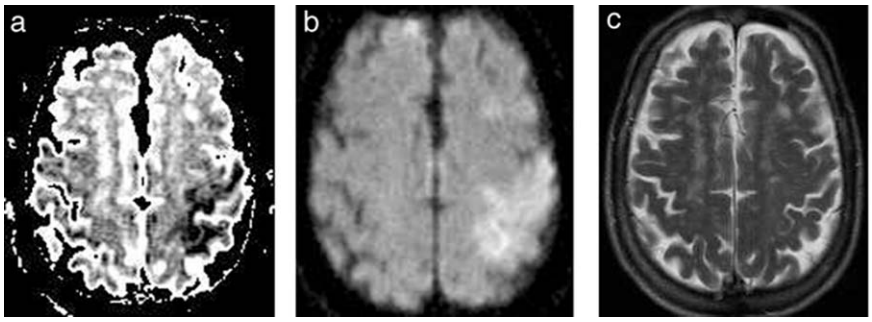


Figure 14 The cytotoxic component of acute cerebral ischemia is demonstrated by ADC hypointensity (a). The ischemic region appears hyperintense on DWI (b), whereas T2-weighted sequences may be unrevealing at this early stage (c).

each of three orthogonal planes also helps to differentiate types of edema. Isotropic diffusion refers to relatively equivalent vectors in all three planes, whereas anisotropic diffusion exists when these effects are disproportionate. Vasogenic and interstitial edema produce anisotropic diffusion patterns. In contrast, cytotoxic edema is less anisotropic. Anisotropic measurements may therefore be helpful in the characterization of cerebral edema.

The recent development of perfusion-weighted imaging (PWI) with MR technology provides parametric maps of several hemodynamic variables, including cerebral blood volume. Elevations in cerebral blood volume associated with cerebral edema are detectable by this technique.

Simultaneous acquisition of multiple MRI sequences enables the clinician to distinguish various forms of cerebral edema. T2-weighted sequences and FLAIR images permit sensitive detection of local increases in water content. Gadolinium-enhanced T1-weighted sequences reveal sites of BBB leakage that may be present surrounding tumors (Fig. 15) or abscesses. DWI localizes abnormal areas of water diffusion, with ADC maps differentiating various forms of edema. PWI can detect regional elevation of cerebral blood volume. The composite interpretation of these studies has revolutionized the diagnosis of cerebral edema. These images often reflect the combined effects of multiple types of edema. For instance, the cytotoxic component of ischemic edema will cause a reduction in the ADC, whereas the vasogenic component will counter this trend. A pseudo-normalization of the ADC may result from these opposing influences. Serial imaging with this noninvasive modality also allows for the temporal characterization of edema evolution. The relative contributions of cytotoxic and vasogenic edema with respect to the ADC during acute ischemic stroke and TBI have been investigated in this manner. The main limitations of this technology logistically relate to cost, availability, contraindications, and its restricted use in critically ill individuals.

IV. TREATMENT

The treatment of cerebral edema should be implemented once the primary disease process has been addressed and the degree of edema warrants intervention. Treatment is usually warranted when the edema is severe and attributable clinical signs and symptoms exceed the manifestations of the primary injury. Because a combination of edema types are usually involved, generalized approaches are often focused on decreasing ICP. Several strategies have been shown to be successful in reducing ICP, although their efficacy varies with the underlying disease, but detrimental effects may also result. The treatment of cerebral edema should therefore consider the primary etiology, the predominant type of edema, and the temporal evolution of the underlying pathophysiologic processes. The following sections summarize the principal therapies for cerebral edema and briefly highlight interventions that focus on specific aspects elucidated from continued research efforts.

A. Head Elevation

Elevation of the head minimizes the contributory effects of increased cerebrovascular venous pressure and outflow resistance. Compression of neck veins can

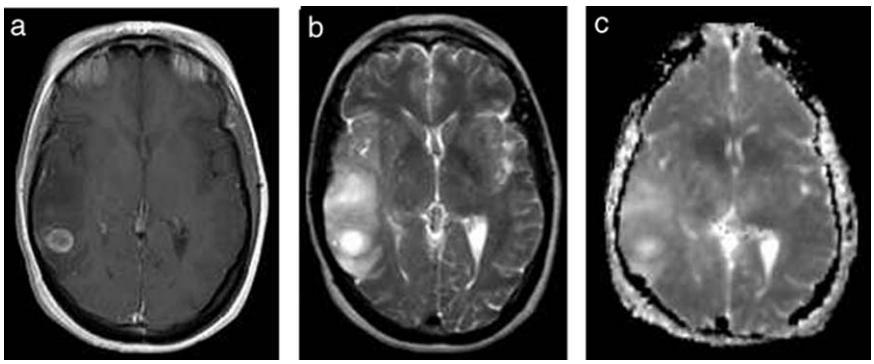


Figure 15 Disruption of the BBB in vasogenic edema associated with a glioma appears hyperintense on gadolinium-enhanced MRI (a). Peritumoral vasogenic edema is demonstrated by hyperintensity on T2-weighted sequences (b) and ADC maps (c).

be avoided by keeping the head in midline position. Prevention of increased intracranial venous pressure may be achieved if the head of the bed is placed at an angle of 30°, although greater angles may theoretically compromise CPP.

B. Diuretics

Rapid and dramatic reductions of most forms of cerebral edema may be accomplished with the use of osmotic diuresis. Osmotic diuretics elevate intravascular osmolarity, minimizing or reversing the osmotic pressure gradients that may participate in cerebral edema formation. Considerable integrity of the BBB is required because extravasation of osmotic agents across this interface may diminish their beneficial effect and potentially cause unintended fluid flux into the brain. Mannitol may also increase cerebral blood flow by reducing blood viscosity, leading to vasoconstriction and a resultant reduction in ICP. Other potentially effective osmotic diuretic agents include urea, glycerol, and hypertonic glucose or saline. However, excessive urea enters cells and causes an unwanted rebound increase in ICP. Chronic or excessive hyperosmolar states may precipitate renal injury. Loop diuretics including furosemide may also be beneficial, although hypovolemia may worsen neurologic outcome. Acetazolamide lowers ICP by decreasing production of CSF. Chronic use of acetazolamide is not recommended because this therapy may result in systemic acidosis.

C. Hyperventilation

Hyperventilation induces rapid, transient reductions in ICP because substantial decreases in pCO_2 cause vasoconstriction and decreased cerebral blood flow. Autoregulatory capability is required for this response. This measure is effective in many forms of edema, although there is a theoretical risk of causing injury due to hypoxia associated with excessively diminished cerebral blood flow. A target pCO_2 of 25–30 mm Hg is recommended when hyperventilation is employed.

D. Cerebrospinal Fluid Drainage

ICP may be diminished by drainage of CSF through a ventricular catheter. Although this reduction in ICP may be beneficial in many forms of cerebral edema, it is particularly effective in the treatment of hydrocephalic edema. Clinicians utilize this intervention when preliminary measures such as osmotherapy have failed. The presence of small ventricles may impair drainage due to the collapse of the ventricular walls around the catheter tip. Lumbar puncture for CSF drainage is not recommended because resulting pressure gradients may precipitate herniation syndromes.

E. Blood Pressure Modification

Alteration of hydrostatic forces in the cerebral circulation is possible through titration of systemic arterial pressure. Reductions of systemic arterial pressure with antihypertensive agents decrease pressure gradients that may precipitate or exacerbate hydrostatic, vasogenic, and ischemic edema. This concept forms the basis for the treatment of hypertensive encephalopathy and related conditions. Although hydrostatic gradients do not mediate the initial changes in vasogenic edema, these forces may encourage extravasation across a damaged BBB. This effect may contribute to the pathogenesis of reperfusion injury following lysis of an arterial occlusion or surgical repair of a hemodynamically significant arterial stenosis. The presence of ischemia limits blood pressure reduction because decreased CPP may precipitate further ischemic injury.

F. Corticosteroids

The use of high-dose glucocorticoids has been shown to produce dramatic reductions in cerebral edema in certain conditions. These agents may stabilize and accelerate repair of the BBB, reduce capillary permeability, improve brain availability of glucose, and alter production of chemical mediators that lead to cerebral edema. Glucocorticoids may also decrease production of CSF, thereby reducing ICP. These agents are particularly effective in edema resulting from tumors and meningitis. Most clinical studies have failed to demonstrate a benefit of glucocorticoids in the setting of TBI. The use of high-dose glucocorticoids may have detrimental effects when used to treat ischemic edema.

G. Barbiturates

Barbiturates reduce the metabolic demands of the brain and cause vasoconstriction in the cerebral

circulation. Preliminary studies suggest that these drugs may effectively lower ICP. Nonetheless, the alleged benefits in clinical outcome have not been demonstrated in studies of TBI and cardiac arrest.

H. Surgical Interventions

The presence of an intracranial mass lesion may require urgent neurosurgical attention. Surgical excision of an edematous lesion or increasing the size of the cranium via hemicraniectomy may reduce further formation of vasogenic edema and eliminate compression of normal brain parenchyma. Rebound edema formation may complicate the perioperative course. Careful control of ventilation and systemic arterial pressure, as well as preoperative administration of glucocorticoids, may minimize the detrimental effects of this complication.

I. Hypothermia

The induction of hypothermia may be an effective treatment for cerebral edema due to vasoconstriction of the cerebral vessels, a reduction in metabolic demands of the brain, and decreased production of CSF. However, the degree of hypothermia may be difficult to control and systemic complications resulting from cardiac arrhythmias, fluid shifts, and increased susceptibility to infection restrict the use of this treatment.

J. Nonsteroidal Antiinflammatory Drugs

Nonsteroidal antiinflammatory drugs (NSAIDs), including ibuprofen and indomethacin, interfere with the production of prostaglandins and have been shown to decrease extravasation in animal models of cerebral edema. Other mechanisms include modulation of endothelial permeability in tumor vessels and peritumoral regions. Clinical studies of NSAIDs for reduction of cerebral edema have been conducted in TBI and tumor-associated edema. These substances may be useful in individuals with tumor-associated edema who are intolerant to steroids or dependent on chronic steroid use.

K. Hyperbaric Oxygen

Increasing inspired pO_2 by administration of hyperbaric oxygen causes vasoconstriction with reduction of cerebral blood flow, cerebral blood volume, and, consequently, ICP. This therapeutic intervention is frequently used in the setting of HACE to diminish the hypoxic stimulus in the formation of cerebral edema.

L. Other Neuroprotective Agents

Many of the previously mentioned interventions have theoretical neuroprotective effects that reduce cytotoxic injury in the formation of cerebral edema. Other potential neuroprotective substances include inhibitors of excitotoxic amino acids, bradykinin release, leukocyte adhesion, and metalloproteinases, modulators of calcium influx, promoters of membrane repair, and free radical scavengers. Preliminary experimental studies of several agents are encouraging, although their efficacy has yet to be demonstrated in clinical trials.

V. CONCLUSIONS

The catastrophic consequences that may result from cerebral edema have provided a perpetual impetus for research in the pathophysiology of this clinical entity. Elucidation of the underlying mechanisms allows for the classification of several forms of cerebral edema. The evolving knowledge of this combined entity emphasizes the critical role of water homeostasis in human physiology and its interaction with the complex biochemical alterations that underlie many neurologic diseases. Noninvasive neuroimaging techniques have revolutionized the identification of specific edema types, but these findings often do not correlate with clinical symptomatology. Therapy is usually reserved for secondary elevations in ICP, at a stage when the consequences prove life threatening and multiple edema types are present. Further diagnostic research efforts may provide early detection and detailed characterization of cerebral edema. This information allows the development of rational therapeutic interventions targeting biochemical factors that are specific to disease-state and temporal patterns. A combination of such interventions that modulate multiple aspects of cerebral edema may prove most efficacious.

See Also the Following Articles

CEREBROVASCULAR DISEASE • EPILEPSY • HYDROCEPHALUS • STROKE

Suggested Reading

Bingaman, W. E., and Frank, J. I. (1995). Malignant cerebral edema and intracranial hypertension. *Neurol. Clin.* **13,** 479–509.

Blei, A. T. (1995). Pathogenesis of brain edema in fulminant hepatic failure. *Prog. Liver Dis.* **13,** 311–330.

Cloughesy, T. F., and Black, K. L. (1999). Peritumoral edema. In *The Gliomas* (M. S. Berger and C. B. Wilson, Eds.), pp. 107–114. Saunders, Philadelphia.

Fishman, R. A. (1975). Brain edema. *N. Engl. J. Med.* **293,** 706–711.

Hackett, P. H. (1999). The cerebral etiology of high-altitude cerebral edema and acute mountain sickness. *Wilderness Environ. Med.* **10,** 97–109.

Hariri, R. J. (1994). Cerebral edema. *Neurosurg. Clin. North Am.* **5,** 687–706.

Katzman, R., Clasen, R., Klatzo, I., Meyer, J. S., Pappius, H. M., and Waltz, A. G. (1977). Report of joint committee for stroke resources. IV. Brain edema in stroke. *Stroke* **8,** 512–540.

Kimelberg, H. K. (1995). Current concepts of brain edema. Review of laboratory investigations. *J. Neurosurg.* **83,** 1051–1059.

Klatzo, I. (1967). Presidential address. Neuropathological aspects of brain edema. *J. Neuropathol. Exp. Neurol.* **26,** 1–14.

Silver, S. M., Sterns, R. H., and Halperin, M. L. (1996). Brain swelling after dialysis: Old urea or new osmoles? *Am. J. Kidney Dis.* **28,** 1–13.

Thapar, K., Rutka, J. T., and Laws, E. R., Jr. (1995). Brain edema, increased ICP, vascular effects, and other epiphenomena of human brain tumors. In *Brain Tumors: An Encyclopedic Approach* (A. H. Kaye and E. R. Laws, Jr., Eds.), pp. 163–189. Churchill–Livingstone, London.

Cerebral Palsy

STEPHANIE WATKINS and ANGELA ROSENBERG
University of North Carolina, Chapel Hill

GLOSSARY

ambulation The act of walking.

Apgar Quantitative assessment of a newborn's medical status after birth.

asphyxia A life-threatening condition in which oxygen is prevented from reaching the tissues due to an obstruction within the respiratory system.

contracture Fibrosis of the muscle tissue with subsequent muscular shortening.

decubitus ulcer Ulceration of the skin due to continuous pressure on one area of the body.

dysarthria The inability to pronunciate words in a clear manner.

dyskinesia Involuntary movements that appear as fragments of normally controlled smooth facial and limb movements.

gait The manner in which an individual walks.

intrathecal Within the meninges of the spinal cord.

meconium The first stool of a newborn infant; may indicate fetal distress if observed in the amniotic fluid.

scoliosis Deviation of the spine that may be congenital or acquired due to abnormalities of the vertebrae, muscles, or nerves.

subarachnoid space The space between the pia and the arachnoid meninges of the spinal cord and brain.

subluxation The partial dislocation of a joint

tone The normal state of partial contraction of a resting muscle.

Cerebral palsy is a disorder of posture and movement that is nonprogressive in nature. The disorder may be viewed as an artificial or umbrella term encompassing a group of symptoms occurring secondary to an insult to the immature brain. This insult may occur pre-, peri-, or postnatally. Sequelae include impairments in the control of selective movement, abnormalities in muscle tone, and impaired activation between agonist/antagonist musculature. Impairments in speech, vision, hearing, cognition, and seizure disorder are also frequently associated with this disorder.

I. EPIDEMIOLOGY

Since the 1950s, the prevalence of cerebral palsy (CP) has fluctuated. During the early 1950s, the prevalence of CP was reported to be between 2 and 2.5 per 1000 live births. A 1996 analysis of the periods between 1954–1958 and 1967–1970 reported the prevalence of CP decreased 40% with concurrent improvements in access to obstetrical and pediatric care as reported by Hagberg and Hagberg. During the period from 1970 to 1982, the prevalence once again increased to the rate of the 1950s. Apparent advancements in medical technology, allowing infants to survive at a lower

gestational age and weight, accounted for this rebound in prevalence of reported cases.

In recent years, the prevalence of CP among live births in developed countries has continued to increase, with the exception of normal birth weight infants, for whom prevalence rates have remained stable. Overall, prevalence rates between 1.06 and 2.5 per 1000 children have been reported. This increase in the existing number of cases has been attributed to increased survival rates of preterm, low-birth-weight infants. A 1996 study by Suzuki *et al.* reported an increase in prevalance of 9-fold among low-birth-weight infants (1500–2499 g) and an increase of 41-fold among very low-birth-weight infants (<1500 g). In 1996, Pharoah *et al.* reported an increase in prevalence of CP among infants weighing less than 1000 g with low-birth-weight babies (<2499 g) accounting for 50% of CP cases. Additionally, perinatal mortality appears to be on the decline, with surviving babies having a higher risk for cerebral damage.

II. ETIOLOGY

In 1862, William John Little, an orthopedic surgeon, was the first to address and hypothesize on the etiology of CP. Little described developmental motor abnormalities in children as "spastic rigidity." He hypothesized that prolonged premature labor, breech delivery, or birth asphyxia were the causes of cerebral damage resulting in impairments in posture and movement. However, since the 1860s, discussion has continued regarding the exact cause of CP because the clinical manifestations of the disorder consistently result in few homogeneous patterns of symptoms.

Today, multiple risk factors contributing to the diagnosis of CP have been identified. However, in most cases, the underlying mechanism remains unknown. Researchers have not been able to link the exact timing of risk factors to resultant lesions in the developing brain. Two etiologies that do produce homogeneous clinical patterns are maternal iron deficiency and Rh incompatibility. Mothers who suffer from an iron deficiency during pregnancy give birth to children who display spastic diplegia with deaf mutism, and infants born with Rh incompatibility experience choreoathestosis and deafness. However, the majority of risk factors that are identified as contributing to CP do not necessarily lead to the disorder.

Risk factors contributing to the origin of CP have been contemplated for both the term and the preterm infant. Risk factors have been identified prior to conception and in the pre-, peri-, and postnatal stages of development (Table I). For full-term infants, the best predictors of the disorder have been reported as neurological abnormalities at birth, including low Apgar scores, abnormalities in respiration, abnormal reflex responsiveness, and neonatal seizures. In many cases, the cause remains unknown or is attributed to prenatal factors. In 1993, Naulty *et al.* reported that for 60% of term infants with CP there appeared to be no identifiable cause.

Since the time of Little, asphyxia during delivery has received much attention from the scientific community as a core etiological factor. However, recent data suggest that most cases of CP do not result from lack of oxygen due to human error during delivery. In fact, the rate of asphyxia at birth has decreased from 40/100,000 births in 1979 to 11/100,000 births in 1996, whereas the incidence of CP has not declined.

The Collaberative Perinatal Study (CPS) of the National Institute of Neurological and Communication Disorders and Stroke, one of the largest longitudinal studies to date, collected data from 43,437 full-term children to identify fetal, intrapartum, and neonatal events in child development. To date, this is the largest and most complete set of prospective data on the study of CP completed in the United States. In 1989, Naeye and colleagues used these data to report that birth asphyxia alone, not attributed to chronic antenatal hypoxia, as a whole accounted for only 6% of all cases of CP (Table II). According to this study, gas anesthesia during labor, maternal seizures, maternal diabetes mellitus, and neonatal hypoglycemia were not associated with a higher than expected frequency of CP. Moreover, congenital disorders accounted for four times as many cases of quadriplegic CP as did birth asphyxia. Congenital abnormalities accounted for 60% of the cases not of this clinical type.

Based on data collected by the CPS, Nelson and Ellenberg supported these findings, reporting that only 9% of CP cases were caused by birth asphyxia alone. Among those cases for which asphyxia may account, it was noted that oxygen deprivation frequently was not measured directly. Rather, asphyxia was inferred from low Apgar scores and signs of fetal distress that may include abnormal respiration, abnormal fetal heart rate, the presence of meconium, low cord blood pH, seizures, and abnormal reflex responsiveness. When investigated, these signs more often related to congenital disorders rather than asphyxia.

Further studies reported similar findings. Congenital malformations of the central nervous system (CNS)

Table I
Risk Factors Associated with Cerebral Palsy[a]

Prenatal	Perinatal	Postnatal
Developmental problems of the fetus	Abnormal presentation of the fetus[b]	Infections (bacterial,viral, parasitic)
Congenital brain malformation	Infection during delivery (hyperbilirubina, blood incompatibility)	Near drowning
Maternal infection	Premature separation of the placenta	Accidents
Maternal seizures	Birth asphyxia: tight nuchal cord	Toxins
Exposure to radiation	Prolonged labor[c]	Brain Trauma
Prescribed thyroid hormones or estrogen in pregnancy	Maternal anesthesia	Birth Weight $<$2001grams
Genetic abnormalities	Gestational age $<$32 weeks	Neonatal seizures
Nutritional deficits		Bilirubin encephalopathy
Preeclamsia[d]		
Chemical teratogenesis		
Maternal mental retardation		
Maternal hemorrhage		
Multiple births		
Long intervals between menstrual cycles prior to pregnancy		
Short ($<$3 months) or long ($>$3 years) interval between pregnancies		

[a]Adapted from Kuban and Leviton, *N. Engl. J. Med.* **330**(3), 188–195 (1994); Nelson and Ellenberg, *N. Engl. J. Med.* **315**(2), 81–86 (1986); and Dzienkowski *et al.*, *Nurse Practitioner* **21**(2), 45–59 (1996).
[b]Abnormal presentation of the fetus may be an indicator of preexisting difficulty as reported by the *New England Journal of Medicine*.
[c]Nelson and Ellenberg reported prolonged labor was not a significant predictor of CP.
[d]In 1999, Nelson and Grether reported premature infants of preeclampsic mothers were less likely to have CP compared to other causes of prematurity.

and non-CNS were reported by Mettau in 1993 as three times more frequent in children with CP than in the general population. In 1990, Torfs *et al.* supported these findings, reporting that 78% of children diagnosed with CP did not have asphyxia at birth.

Table II
Categories of Risk Factors Associated with CP[a]

Chronic antenatal hypoxia	Nonasphyxial
Maternal anemia	Congenital malformation
Continuation of pregnancy $>$42 weeks	Birth trauma
Preeclampsia	Intoxication
Multiple fetuses	CNS infections
Subnormal maternal blood volume expansion in first trimester	Breech delivery
Decrease in maternal blood pressure in third trimester	

[a]Adapted from Naeye *et al.*, *Am. J. Dis. Child* **143,** 1154–1161 (1999).

Aside from birth asphyxia, several other etiological factors have received attention. One-half of all cases of CP have been attributed to prenatal malformations, prenatal strokes, congenital taxoplasmosis, rubella cytomeglavirus, and herpes infections. Maternal infections have been linked to offspring diagnosed with CP, including those for which women have a fever of more than 30°C during delivery, as well as a diagnosis of chorioamnionitis or congenital rubella. In 1997, Grether and Nelson reported a nine-fold increase in the risk of spastic CP in offspring associated with maternal infection. Genetic factors have also been implicated in a small number of cases. Gustavason and colleagues in the late 1960s reported that one-third to one-half of the cases of congenital ataxic CP with associated mental retardation are genetically predetermined.

The majority of events leading to CP are thought to occur during the perinatal period. With the brain still in a stage of rapid growth at birth, complications such as intracranial hemorrhage and periventricular

leukomalacia (PVL) are often associated with a diagnosis of CP. Low-birth-weight infants are at a greater risk of developing PVL with a subsequent diagnosis of CP. PVL has been associated with low blood pressure and neurochemically mediated injury to the white matter. When compared with intracranial hemorrhage, PVL has been reported to be a more accurate predictor of CP. However, questions remain as to whether the resultant CP is attributable to perinatal/postnatal complications or preexisting pathologic conditions contributing to prematurity and potential injury of the developing brain *in utero*.

Despite improvements in medical technology, the rate of CP has continued to rise. With the asphyxial origin of CP being challenged, the debate continues regarding the factors that contribute to CP as well as the developmental stage (pre-, peri-, or postnatal) in which these factors affect the developing brain.

III. DIAGNOSIS

Many behavioral characteristics are indicative for the early detection and diagnosis of CP (Table III). However, the diagnosis of CP is a complicated one, often determined by physical exam in conjunction with a thorough medical history, neurological exam, and the exclusion of differential diagnoses. Currently, there is no one test to determine a diagnosis of cerebral palsy.

Table III
Behavioral Characteristics Associated with Cerebral Palsy[a]

Less than 6 months	6 months to toddler
High-pitched cry	Instability in sitting
Oral hypersensitivity	Oral hypersensitivity
Lethargy	W-sitting
Irritability	Excessive arching of the back
Abnormal primitive reflexes	Inability to perform age-appropriate motor skills (e.g., sitting by 7 months)
Diminished head control	
Stiff or floppy posturing	
Inability to perform age-appropriate motor skills (e.g., rolling by 4 to 5 months)	

[a]Adapted from Dzienkowski *et al.*, *Nurse Practitioner* **21**(2), 45–59 (1996).

In addition, initially identified motor impairments may change across the child's life span, thus adding to the difficulty of diagnosis. It is presumed that maturation of the basal ganglia and neuronal myelination must occur prior to the emergence of spasticity, dystonia, and athetosis. Therefore, the clinical signs displayed by the child may change over time. In the early 1980s, Nelson and Ellenberg found that 118 of 229 children diagnosed with CP at 1 year were free of motor handicaps by age 7. Caution is therefore advised regarding definitive diagnoses and subsequent prediction of outcomes at a young age.

IV. CLASSIFICATION

Cerebral palsy is a global term that requires a classification system to further define the clinical picture. However, classification is often challenging given that a diagnosis of CP represents a collection of many symptoms in the absence of a standard clinical picture.

The most widely used classification system for this disorder was developed by the American Academy of Pediatrics (AAP). The AAP classifies CP using two scales based on symptoms rather than etiology. The *motor classification* scale characterizes the quality of movement, whereas the *topography scale* describes affected body parts.

Characteristics of movement quality are described as spastic, athetoid, rigid, ataxic, hypotonic, or mixed (Table IV). Lesions in specific regions of the brain have been associated with specific abnormalities of movement. The motor area of the cerebral cortex, including the prefrontal cortex, premotor cortex, and primary motor cortex, is commonly affected. Specifically, involvement of the motor cortex, which projects to and from the cortical sensorimotor areas, is associated with a diagnosis of spastic cerebral palsy. Damage to the basal ganglia results in a diagnosis of dyskinesia or athetosis. The basal ganglia functions to regulate voluntary motor function, thus modulating purposeful movement and suppressing unnecessary motion. Lastly, the cerebellum, functioning to control coordination, timing, and sequencing of movement, is associated with an ataxic movement disorder.

Topographically, the individual is described as having one limb involved (monoplegia), one half of the body involved (hemiplegia), both legs involved (diplegia), or both legs and one arm involved (quadriplegia). Therefore, an individual with

Table IV
Definitions of Motor Characteristics[a]

Characteristic	Definition
Spasticity	Characterized by sudden muscle contractions that are convulsive in nature with persistent rigidity; increased resistance to passive movement that is velocity dependent; diminished threshold for stretch response and clonus; Involvement usually is greater in antigravity muscles
Athetosis	Writhing, uncontrolled movements with poor coordination and midrange control
Rigidity	Resistance to slow-speed passive movement from both the agonist and the antagonist; lead pipe resistance throughout the range of motion; cog wheel discontinuous resistance throughout the range of motion
Ataxia	Poor coordination and timing with voluntary movement; deficits in balance, equilibrium, and depth perception
Hypotonia	decreased ability of the muscle to generate force; diminshed muscle tension at rest; excessive joint mobility

[a]Adapted from Minear (1956). *Pediatrics* **18,** 841–852.

velocity-dependent resistance to passive movement involving the right half of the body would be classified as having right spastic hemiplegic cerebral palsy.

Until recently, a classification system was not in place to provide a functional picture of the individual with cerebral palsy. In 1997, Palisano *et al*. developed the *Gross Motor Function Classification System for Cerebral Palsy* to provide a method to standardize clinical observations of child function. In addition to providing a numerical classification of function, the scale standardizes communication between professionals. Palisano *et al*.'s purpose was to describe discrete levels that represent present motor performance and limitations based on self-initiated movement. The classification system emphasizes sitting and walking and ranges from level 1 to level 5, with distinctions between levels based on functional limitations and the need for an assistive device, assistive technology, and/or mobility aides. The classification system uses the concept of disability as defined by the International Classification of Impairments, Disabilities, and Handicaps (ICIDH) of the World Health Organization and the concept of functional limitation as outlined in the disablement model as described by Nagi. The ICIDH defines disability as "the restriction or lack of ability to perform an activity in the manner or within the range considered normal for a human being" (1980). Nagi defines functional limitation as a "limitation in performance at the level of the whole person" (1965).

V. IMPAIRMENTS AND MANAGEMENT

Treatment of the individual with CP is focused on maximizing functional independence while preventing the occurrence of secondary conditions. Associated primary conditions include abnormalities of posture and movement as well as learning disabilities, seizures, and problems with vision, hearing, and speech. Over time, movement abnormalities often place abnormal stresses on joints, resulting in secondary conditions such as joint contractures and bony deformities. Treatment interventions may include but are not limited to pharmacological therapy, surgery, physical therapy, occupational therapy, and speech therapy.

Movement and postural abnormalities may be divided into musculoskeletal impairments and neuromuscular impairments. Musculoskeletal impairments include loss of joint range of motion, bony abnormalities, and strength deficits that frequently develop over time as sequelae of the initial injury. Neuromuscular impairments include abnormal selective muscle control, inefficient regulation of muscular activity, a decreased ability to learn new movements, and abnormalities in tone.

In individuals without CP, muscles are phased in and out with coactivation of synergistic musculature during an activity. However, in the individual with CP, inappropriate muscular sequencing and coactivation of agonist and antagonist musculature occurs during voluntary movement, according to Naher and colleagues and Knutsson and Martensson. As a result, compensatory movement patterns are often seen with movement initiation. Individuals with hemiplegic CP may initiate walking on the uninvolved side of their body, whereas those with diplegia may use their head, neck, trunk, and arms to facilitate movement. A decreased ability to maintain and anticipate postural adjustments is also a common manifestation of the disorder. Muscular contraction for stability in anticipation of postural changes is often inadequate, creating difficulties with postural control.

Despite the variety of impairments noted previously, much of the emphasis of treatment continues to be placed on the management of spasticity. Although the precise mechanism underlying spasticity remains unknown, it has been established that the stretch reflex is exaggerated due to overactivity of the alpha motor neurons. Cortical impairment is thought to interrupt descending modulating pathways to the lower alpha motor neurons that inhibit the motor unit. This prolonged disinhibition of the alpha motor neuron results in a velocity-dependent increase in the tonic stretch reflex.

Spasticity is frequently associated with CP and often results in limited selective muscle control, improper timing and sequencing of movement, and the inability to move against gravity. Therefore, functional independence is affected, often resulting in an altered ability of the individual to sit, stand, crawl, and walk. However, despite negative sequelea, spasticity may serve a functional purpose in some individuals. It is important to evaluate the contribution of spasticity toward individual function and quality of movement prior to initiating treatment focused on spasticity reduction.

VI. PHARMACOLOGICAL INTERVENTION

A. Oral Medications

In the 1990s, Albright reported that oral medications, including benzodiazepines and skeletal muscle relaxants, had been used with limited success to diminish the effects of spasticity. Diazepam, a benzodiazepine, facilitates the inhibitory response of γ-aminobutyric acid (GABA) neurotransmitters by increasing the conductance of chloride ions upon binding to $GABA_A$ receptors. This results in a presynaptic inhibition at the spinal cord level with resultant relaxation of skeletal muscle. Use of this agent has been associated with adverse effects including sedation and a potential for dependency.

Muscle relaxants used in the treatment of spasticity include dantrolene sodium and baclofen. Dantrolene sodium acts directly on the skeletal muscle by inhibiting the release of calcium from the sarcoplasmic reticulum. The result is an interference with the excitation–contraction coupling and a resultant reduction in contractile force of the muscle. Side effects from dantrolene sodium use include drowsiness and a suggested increase in seizure activity in children with CP.

Baclofen is a skeletal muscle relaxant frequently used in the management of spasticity. It acts throughout the central nervous system and has been shown to be most effective when used to treat spasticity of spinal origin. In spasticity of cerebral origin, spasticity reduction has been reported as mild due to poor lipid solubility of the drug. In actuality, only a small amount of baclofen penetrates the blood–brain barrier, requiring higher doses to achieve an effect. Upon entry, baclofen binds to $GABA_B$ receptors to prevent the uptake of calcium necessary for the release of excitatory neurotransmitters, such as aspartate and glutamate. The effect is a decrease in the release of neurotransmitters in the excitatory pathways resulting in a reduction of muscular spasticity. Compared with dantrolene sodium, baclofen use has fewer side effects, with no associated increase in seizure activity. However, drowsiness is frequently experienced.

B. Non-oral Medication

To avoid the side effects frequently associated with oral dosing, other routes of drug administration have been explored including the use of botulinum toxin A injections and intrathecal baclofen pumps.

Botulim toxin A, a toxin produced by *Clostridium botulinus*, is commonly used to inhibit spasticity associated with CP. It acts at the muscle site and is useful in targeting spasticity in isolated muscle groups. The toxin blocks the presynaptic release of acetylcholine from the motor nerve terminal resulting in chemical denervation of the muscle or paralysis. The toxin has a binding affinity for the neuromuscular junction and is injected directly into the affected muscle belly. The degree of paralysis is dose dependent and reversible. Postadministration, the nerve terminal regenerates, requiring the toxin to be reinjected for continued effects.

One side effect of the injection is transient muscle weakness, but this is minimal due to the small amount of toxin actually entering the systemic circulation. The use of botulinum toxin A has been reviewed regarding its efficacy to improve function. Many studies have found the toxin to be effective in reducing spasticity in the upper and lower extremities, with effects beginning 12–72 hr after injection and lasting 2–4 months. In 1995, Denislic and Meh reported improvements in functional hand movement and foot posture. In 1994, Cosgrove and colleagues reported that the injections resulted in few side effects, with improved walking

ability and tone reduction in the lower extremities after 3 days. Despite its reported effectiveness toward spasticity reduction, botulinum toxin A has shown little effect on tremor reduction. Nonetheless, it continues to be used as a conservative approach to delaying surgical intervention or simulating its effects.

Intrathecal baclofen (ITB), approved in 1996 to treat spasticity of cerebral origin, is an alternative formulation to oral blacofen. Utilizing a programmable pump implanted under the skin of the lower abdomen, baclofen is delivered directly into the cerebral spinal fluid (CSF) via a catheter fed posteriorly into the intrathecal (subarachnoid) space of the thoracic spine. The catheter is placed via lumbar puncture at vertebral level L3–L4 or L4–L5 and advances to the L1 vertebral region, where the anatomical curvature of the spine provides a more open area to distribute medication. The drug is delivered to the superficial layer of the spinal cord, binds to the $GABA_B$ receptors, and reduces spasticity using only a few millimeters of active drug. Intrathecal administration delivers four times the concentration of oral baclofen at the site of action at 1/100th of the oral dose. The smaller dose translates into a lower systemic concentration and therefore less systemic side effects such as somnolence.

In addition to the benefit of reducing the effects of somnolence, dosing of ITB may be titrated to accommodate daily variability in patient symptoms. The level of baclofen may be adjusted to moderate spasticity that interferes with gait or other activities of daily living. Furthermore, drug levels may be adjusted to accommodate changes in environmental conditions that may alter spasticity, such as temperature, stress, and hunger. Even though ITB use has fewer systemic side effects when compared with oral dosing, risks associated with ITB use include infection, disconnection and leaks of the catheter, adverse effects (hypotonia, seizures, nausea, and vomiting), and drug overdose.

Spasticity reduction in children with CP using ITB has been consistently reported. However, reports have been inconsistent regarding the effect that ITB has on functional improvements in movement. Further prospective research is needed to scientifically evaluate the effects that ITB has on gait and activities of daily living. The following guidelines have been established to determine individual candidacy for ITB use [from Barry *et al.*, *Pediat. Phys. Therapy* 2000 **12**(2), 77–86].

At least 4 years of age
Adequate body size to accommodate the pump
The presence of hypertonicity
The absence of infection
Medical stability

However, specific selection criteria have not been established.

VII. SURGICAL PROCEDURES

A. Selective Dorsal Rhizotomy

Selective dorsal rhizotomy (SDR) is a surgical procedure that involves severing specific sensory nerves that innervate the lower extremities. This selective denervation reduces abnormal excitatory peripheral sensory input to the spinal cord. Lumbosacral sensory nerves (L2–S2) are exposed and separated into component rootlets. These rootlets are then electrically stimulated and abnormal responses are recorded as well as visually observed. The rootlets producing abnormal responses are then severed, reducing spasticity while preserving tactile and proprioceptive sensation.

The primary purpose of this procedure is to improve or enable ambulation; however, increased joint range of motion and improved ease of positioning may be additional benefits. Ideal candidates for SDR are children between the ages of 4 and 6 who are diagnosed with either spastic hemiplegic or diplegic CP and have the ability or the potential for ambulation.

The procedure is irreversible, with associated risks of paralysis, permanent loss of bowel and bladder control, leakage of spinal fluid, infection, loss of sexual function, and permanent loss of sensation. Temporary side effects, which usually disappear after several weeks, include sensory loss, numbness, or an uncomfortable sensation in the limbs supplied by the severed nerve. Postsurgical treatment involves months of intensive physical therapy to strengthen weak muscles and to develop new movement patterns. Contraindications to SDR suggested by some surgeons include muscle tendon contractures, a history of orthopedic surgery, and hip displacement. The following are guidelines for SDR (from www.ccmckids.org/departments orthopaedics/orthoed4.htm):

Two years of age or older
Diagnosis of spastic diplegic or quadriplegic CP
Difficulty with ambulation due to spasticity
Adequate muscle strength
No previous history of orthopedic surgery

Ability to actively participate in physical therapy
Family that can provide follow-up physical therapy

B. Orthopedic Procedures

Due to immobility frequently associated with spastic CP, many individuals experience loss of joint range of motion and abnormal formation of bony structures. Functional problems often arise as a result of orthopedic impairments. An individual may experience problems with posture and pain as well as difficulty with environmental accessibility, ambulation, and daily hygiene.

Orthopedic surgery is utilized in an effort to correct joint contractures, improve musculoskeletal alignment, and improve joint stability and muscular balance. In contrast to neurological treatment, the neurological patterns of the muscle are not altered, but the influence of the muscle on the joint structure may change.

Orthopedic surgery includes both soft tissue and bony procedures. Soft tissue procedures include tendon lengthening, tendon transfers, and neurectomies. Tendon lengthening is used to increase range of motion within a joint. The tendon is released or lengthened in an effort to increase passive joint range of motion while preserving tension within the muscle. Tendon transfers involve the relocation of spastic muscles to improve functional alignment. Finally, neurectomy, removal of the nerve, is used predominantly for nonambulatory children in combination with other surgical procedures.

In general, bony procedures are indicated for severe deformities that are not correctable by soft tissue procedures. These procedures may be used to improve joint stability, to realign a bone with its muscular attachments, and for bone grafting. Osteotomies (surgical realignment of the bone) and arthrodesis (surgical fusion of a bone over a joint space) are common procedures. Derotation osteotomy (correction of internal rotation deformity of the hip) is a common procedure indicated for children who exhibit excessive femoral anteversion with resultant turning in of their feet during ambulation. Arthrodesis may be indicated in the absence of muscle tone, in the presence of severe deformity, or with gross loss of function (Table V).

Spinal scoliosis and hip subluxation, with subsequent hip dislocation, are common orthopedic concerns. Nonambulatory children diagnosed with spastic CP often exhibit an extensor tonal pattern of hip flexion, adduction, and internal rotation. Suggested contributing factors to subluxation of the hip include the strong pull of the adductor and flexor muscles of the hip joint in conjunction with weak hip abductor and extensor musculature.

Risk factors associated with hip subluxation include an increased femoral neck shaft angle (valgus), femoral anteversion, and a shallow acetabulum. With subluxation of the hip, there is a break in the Shenton line of the femur with exposure of more than a third of the femoral head. However, a portion of the femoral head remains in contact with the acetabulum. Limited hip abduction of less than 45° is often an early clinical sign of hip instability. When left untreated, subsequent dislocation usually follows within 2 years, between 5 and 7 years of age.

Hip subluxation and ensuing dislocation may result in pelvic obliquity, difficulty with ambulation, pain, seating problems, decubitus ulcers, and hygiene

Table V
Common Orthopedic Impairments and Associated Surgical Procedures for Children with CP[a]

Orthopedic problem	Implications	Surgical procedure
Hip adduction contracture	Scissor gait, difficulty with hygiene, predisposes hip to subluxation	Adductor release or transfer
Hip flexion contracture	Crouch gait pattern, lumbar hyperlordosis	Psoas tendon tenotomy
Increased femoral anteversion	Feet turned in during ambulation	Derotational osteotomy
Knee flexion contracture	Crouch gait	Hamstring lengthening
Ankle equinovalgus	Painful calluses and blisters on the inner foot	Arthrodesis, calcaneal osteotomy
Ankle equinus	Lack of heel strike during ambulation	Gastrocnemius tendon lengthening

[a]Adapted from Renshaw (1996).

problems. Therefore, surgical treatment for early signs of hip instability is focused on preventative measures. An adductor tenotomy is often performed to reduce the power of the adductor muscle force on the hip joint. Iliopsoas muscle lengthening may be simultaneously performed. With these simple surgical procedures subsequent hip dislocation may be preventable.

Scoliosis is often a concern with regard to children who are nonambulatory and remain in a seated position for prolonged periods of time. Children with CP who are diagnosed with quadriplegia, exhibit poor sitting balance, pelvic obliquity, multiple spinal curves, and are young in age are at a greater risk for scoliotic curve progression. Nonsurgical management may include the use of orthoses, electrical stimulation, and therapeutic exercise. However, nonsurgical methods are less effective when used to treat CP-related scoliosis compared to scoliosis of an idiopathic origin. Therefore, spinal fusion is warranted in most cases in an effort to balance the spine over a level pelvis.

C. Nonsurgical Treatment

A variety of nonsurgical techniques are employed to prevent secondary problems that frequently interfere with functional movement, such as loss of joint range of motion, muscular weakness, and spasticity. Assistive devices including walkers, crutches, and manual or power wheelchairs may be used to facilitate mobility.

Neuromuscular electrical stimulation (NMES), the use of electrical current to stimulate muscle and nerve fibers, is one method used in an attempt to temporarily reduce spasticity, increase joint range of motion, strengthen the muscle, and facilitate motor control. NMES may also be used for muscle reeducation. In the child with CP, NMES is frequently utilized to facilitate motor performance as well as sensory input to the muscle. For example, electrodes may be placed over a weak muscle to elicit a muscular contraction as the child engages in a purposeful activity. NMES may also be used to treat scoliosis as well as to improve the quality of gait.

Therapeutic electrical stimulation (TES) is an alternative type of electrical stimulation that delivers electrical current at a low intensity as the individual sleeps. The purpose of TES is to improve strength of nonspastic musculature in individuals who undergo a SDR. The current is delivered above the threshold for sensation without producing a muscle contraction and is hypothesized to increase muscle growth with subsequent improvements in strength through increased local blood flow. In 1997, Steinbok and colleagues reported functional improvements as a result of TES, but further empirical research is needed.

To promote proper joint alignment, the individual may undergo assessment and fabrication for an orthosis. The goal of the orthosis is to assist with proper alignment and stability of the foot and leg during weight bearing in an effort to improve efficiency and ease of gait. Among those individuals who are nonambulatory, orthoses may be fabricated to maintain proper joint alignment in an effort to prevent joint contractures.

Orthoses are commonly custom made from plastic material, but lightweight metal and leather are also options. In an effort to provide joint stability, orthoses are available with varying levels of support. Shoe inserts offer the least amount of support, assisting with foot position during ambulation. Ankle–foot orthoses (AFOs) and knee–ankle–foot orthoses (KAFOs) provide additional stability to the ankle and leg. The AFO extends above the ankle to the midcalf, assisting with ankle control, whereas the KAFO extends above the knee to the thigh, providing support to both the ankle and knee joints. KAFOs therefore provide the greatest amount of support to the lower extremity.

AFOs and KAFOs are the most common orthoses prescribed for individuals with CP. The orthoses usually fit inside of the shoe, facilitating stability of the leg, ankle, and foot as the foot contacts the floor while standing or walking. Orthoses are commonly fabricated when the child with CP begins to stand, but they may be utilized at an earlier age to promote joint alignment.

Splinting and therapeutic exercise are also options to maintain or promote functional joint range of motion. In the event that joint motion is limited, serial casting, utilized to lengthen hypoextensible muscles, may be a viable option.

VIII. HABILITATION

Habilitation of the individual with CP is frequently accomplished by using a multidisciplinary approach that involves the input of a variety of professionals in collaboration with the child and the family. Although CP is nonprogressive in nature, the manifestations of the disorder may change over time. As the individual progresses from childhood to adulthood, the goal of habilitation is to prevent secondary conditions such as

muscle contractures or obesity and to assist the individual in interacting independently within many environments. Therefore, habiliatation often emphasizes facilitation of functional activities and the use of environmental adaptations.

A. Speech Therapy

In the rehabilitation of the individual with CP, the primary role of the speech/language pathologist/therapist is to address communication and feeding needs. Common sequalae of the disorder treated by speech therapists include impairments of articulation that may include mild to severe dysarthria, difficulties with swallowing, saliva management, and nutritional problems. Due to poor motor control, individuals may exhibit problems with the coordination of swallowing with breathing, difficulty with collection and organization of food, and problems with drooling.

To facilitate management of swallowing and eating, positioning and oral sensorimotor training may be used. As the child with CP ages, it is hoped that natural speech will improve. However, if a child does not exhibit the ability to functionally communicate and interact with the environment, the use of augmentative or alternative communication may be explored. Augmentative communication involves the use of aids to supplement existing vocal communication. Alternative communication is defined as communication without verbal ability.

Augmentative and alternative communication techniques are widely available and vary in cost and complexity of use. Low-cost techniques include the use of sign language, gesturing, or pointing or eye gaze in conjunction with a picture board. For those individuals possessing the cognitive skills for complex language, computer-driven devices with high-quality voice output and visual imaging systems are available.

B. Occupational Therapy

Occupational therapy facilitates function of the individual to promote accomplishment of occupational pursuits. Occupational pursuits vary depending on individual age or interest and may include hobbies, sports, or vocational pursuits. Occupational therapists frequently assist the individual in performing meaningful tasks in three performance areas: self-care, work, and play. Overall performance level of the task and barriers to performance may be analyzed by the occupational therapist. Specific components of the task, including sensory processing, perceptual processing, attention, neurological status, motor control, cognition, and psychosocial processing, may also be assessed.

Therapeutic techniques implemented to assist the individual with independent task performance may focus on the task, the person, and the environment. For example, the individual may perform muscle strengthening or range of motion exercises to build strength and joint motion required to perform the task. The use of task adaptation may also be addressed. A universal holder to stabilize the hand during writing, switches for communication, and adaptive utensils for feeding are three examples. Environmental modifications, such as the installment of ramps to access a playground, may also be implemented.

In the area of task adaptation, occupational therapists serve a vital role in determining the appropriate type of augmentative/alternative communication aide that may be necessary to accomplish daily tasks. To promote independence in daily activities, electronic aids and/or devices are available to assist the individual to access and control his or her environment. Switches are useful to teach children cause-and-effect relationships and may be used as electronic aids for a variety of daily tasks. Examples include head switches or hand switches to activate a toy, blender, or television. Keyguards preventing unwanted keystrokes, touch screens, mobile arm supports, and computer mice with various methods of activation, are but a few of the devices available to improve computer accessibility for an individual with a disability.

C. Physical Therapy

When providing services for individuals with CP, the primary role of the physical therapist is to facilitate independent mobility. For individuals capable of functional movement, independent mobility may require the use of an assistive device such as a walker or orthoses. For individuals incapable of functional movement, the physical therapist frequently serves to ensure proper positioning to prevent the development of secondary conditions such as joint contractures, bony deformities, or skin breakdown. In the child, treatment is often facilitated through play, with an emphasis on age-appropriate developmental milestones such as rolling, sitting, standing, and walking. Moreover, the focus is on the promotion of movement

patterns to encourage independence at home and at school.

In the adult, emphasis is placed on mobility necessary for the accomplishment of functional activities of daily living. With all individuals and families, the physical therapist aims to facilitate individual assumption of their own health and the prevention of secondary conditions.

In addition to hands-on intervention, physical therapists provide insight regarding environmental accessibility, activity adaptations, and proper positioning to facilitate the use of augmentative/alternative communication devices.

IX. LEGISLATION

During the past decade, several pieces of legislation have served to expand the rehabilitation options and societal opportunities for individuals with disabilities. The Rehabilitation Act of 1973 set a precedent for legislative support regarding the rights of individuals with a disability. The act prohibits discrimination of qualified individuals with disabilities from federally funded programs. Therefore, programs receiving federal money are required to make reasonable accommodations in order that individuals with a disability may participate.

In 1975, the Individuals with Disabilities Education Act was implemented to ensure that children with disabilities have the right to a free and appropriate public education in the least restrictive environment. Prior to its implementation, the rehabilitation or education of children with disabilities primarily occurred in state institutions. Today, the law supports individualized goal planning, written in collaboration with the family and educational staff, supporting a child's full participation in family, school, and community life. Children with disabilities therefore have the right to fully participate with their peers in school-related educational, sport, and recreational activities.

A third piece of legislation, the Americans with Disabilities Act, was implemented in 1990 to ensure environmental accessibility for individuals with disabilities. The law mandates that public or private facilities receiving federal financial support eliminate architectural barriers for public transportation and new facility construction as well as guarantee reasonable accommodations in all aspects of employment. Therefore, these laws have progressively opened doors for enhanced career and recreational opportunities for all individuals with disabilities.

X. COMMUNITY RECREATION/LEISURE

Community recreation and leisure activities can provide individuals with CP a source of enjoyment and a lifelong opportunity for general fitness and wellness. Specifically, community recreational opportunities provide an outlet to engage in physical activity in addition to promoting independence and self-sufficiency.

As individuals with CP age, they experience the normal sequala of the aging process in addition to already existing impairments in movement. Weak muscles may exhibit a decrease in strength, and joints may become less flexible. Therefore, greater effort may be required to perform activities of daily living. Community recreation may assist the individual in retarding this process by providing physical activities that promote muscle strengthening, endurance, and joint flexibility. Finally, involvement in community recreation and fitness can provide a social outlet, improve cardiovascular fitness, assist with weight control, and assist in the prevention of osteoporosis.

In the past, the use of progressive resistive exercise in individuals with CP was a concern. Resistive training was thought to increase recruitment of overactive spastic musculature. However, recent work by Damiano and colleagues has reported resistive exercise to be a safe and effective method to improve strength and motor performance. Through a progressive series of studies, researchers have found that the use of resistive exercise does not elicit or inadvertently strengthen overactive antagonist spastic musculature.

In children with CP, the literature has supported the use of resistive exercise to improve muscle strength and functional movement. Moreover, Damiano and colleagues reported improvements in quadricep muscle strength in conjunction with improvements in walking speed in children with CP. O'Connell and colleagues also documented a correlation between muscular endurance and the ability to perform aerobic and anaerobic wheelchair tasks.

With regard to cardiovascular fitness, individuals with CP frequently exhibit lower cardiovascular fitness levels. Although researchers have shown that cardiovascular fitness can improve in individuals with CP, further research is needed.

As with all individuals seeking to engage in new physical activities, individuals with CP should obtain medical clearance prior to participation in a community recreation or sporting activity. Many individuals pursue inclusive recreational and sport activities through local community organizations such as the

YWCA. In addition, sporting opportunities specifically targeted for individuals with CP are organized by the Cerebral Palsy Athletic Association and the Cerebral Palsy International Sports and Recreational Association. The Cerebral Palsy Athletic Association provides the following sporting opportunities: archery, boccia, bowling, track and field, cross-country, cycling and tricycling, equestrain, powerlifting, slalom, soccer, swimming, table tennis, target shooting, and wheelchair team handball.

The number of nationally recognized recreational opportunities available to individuals with disabilities continues to grow.

XI. CAREERS AND AVOCATIONS

Individuals with CP reflect societal patterns when choosing career and avocational pursuits. While some choose to work in the service sector, others pursue careers in law, education, or medicine. As a result of supportive legislation for individuals with a disability, positive changes in employment rates have occurred. Although competitive employment rates for individuals with CP ranged from only 17 to 35% during the past several decades, a positive change was reported in 2000 by Murphy and colleagues. Fifty-two percent of individuals with CP were employed in competitive work, with 7% employed in semicompetitive work and 18% employed in sheltered work.

With supportive public policy, medical advancements, and assistive technologies, physical and environmental barriers may no longer impede professional success. Currently, individuals with disabilities are becoming autonomous productive members of society as reflected by increases in employment. Thus, the groundwork for vocational opportunities has been laid, empowering individuals with CP to have the freedom to pursue their dreams.

See Also the Following Articles

BRAIN DEVELOPMENT • MOVEMENT REGULATION • SPEECH

Suggested Reading

Broman, S. (1984). The Collaberative Perinatal Project: An overview. In *Handbook of Longitudinal Research* (S. A. Mednick, M. Harway, and K. M. Finello, Eds.), Vol. 2, pp. 185–215. Praeger, New York.

Geralis, E. (Ed.) (1991). *Children with Cerebral Palsy A Parents' Guide*. Woodbine House, Bethesda, MD.

The National Center on Physical Activity and Disability (2000). Cerebral palsy and exercise. http://www.ncpad.org.

Olney, S. J., and Wright, M. J. (1994). *Cerebral palsy*. In *Physical Therapy for Children* (S. K. Campbell, D. W. Vander Linden, and R. J. Palisano, Eds.), 2nd ed., Saunders, Philadelphia.

Renshaw, T. S. (1996). Cerebral palsy. In *Lovell and Winter's Pediatric Orthopedics: Cerebral Palsy*. (R. T. Morrissy, and S. L. Weinstein, Eds.), 4th ed. pp. 469–502. Lippincott–Raven, Philadelphia.

Cerebral White Matter Disorders

CHRISTOPHER M. FILLEY
University of Colorado School of Medicine and Denver Veterans Affairs Medical Center

GLOSSARY

Binswanger's disease A vascular disease of the brain in which ischemic demyelination and infarction primarily affect the cerebral white matter.

demyelination A neuropathological process in which myelin is stripped away from axons, with or without associated axonal and/or neuronal damage.

leukodystrophy A genetic disorder of white matter characterized by abnormal myelin metabolism in the brain.

magnetic resonance imaging A powerful neuroradiological technique that permits detailed noninvasive views of the brain white matter.

multiple sclerosis An idiopathic, inflammatory, demyelinative disease of the central nervous system.

myelin The fatty insulation of many axons in the brain that helps increase the conduction velocity of neurons by the phenomenon of saltatory conduction.

neural networks Integrated systems of gray and white matter in the brain that subserve elemental and higher cerebral functions.

toluene leukoencephalopathy A disorder of brain white matter caused by excessive exposure to toluene vapor.

white matter dementia A dementia syndrome resulting from primary or selective involvement of cerebral white matter.

white matter tracts Collections of myelinated fibers connecting cortical and subcortical gray matter regions in the brain.

The cerebral white matter constitutes approximately one-half the volume of the human brain. White matter establishes anatomic connectivity between cortical and subcortical gray matter regions within and between the hemispheres, and it enables more efficient neuronal transmission and cerebral function. Disorders of the cerebral white matter comprise a diverse group of neuropathologic conditions, of which multiple sclerosis is the best known. Neurologic dysfunction is well-known to be associated with these disorders, and neurobehavioral syndromes including dementia are also being increasingly recognized. Recognition of these clinical phenomena represents an opportunity to improve the care of individuals with these disorders and to expand our understanding of the human brain.

I. CEREBRAL WHITE MATTER

The human brain is that portion of the central nervous system within the cranium. Twelve pairs of cranial nerves provide conduits for special senses, general sensation, and motor function, mainly for the head and neck. The brain is continuous caudally with the spinal cord, which in turn is connected to numerous peripheral nerves that travel to and from the entire body. Thus, the brain, at the highest level of the nervous system both literally and figuratively, exerts control over all aspects of bodily function. The brain is best known for its gray matter—the neocortex and a variety of subcortical structures—but white matter also forms a considerable portion of the brain and has an important role in many of its operations. The white matter of the cerebrum—those fibers and tracts that are found within and between the paired cerebral hemispheres—serves to connect various gray matter regions into functionally unified neural networks. The

study of the organization of these networks, currently understood at only a rudimentary level, is one of the major goals of contemporary neuroscience. Here, the anatomy, physiology, development and aging and neuroimaging of the cerebral white matter are reviewed.

A. Neuroanatomy

White matter is so named because of its glistening white appearance on the cut surface of the brain. This feature is attributable to myelin, a fatty substance that wraps around the axons of cerebral neurons as they course through the brain. Myelin is a complex mixture of lipids (70%) and protein (30%) manufactured by oligodendrocytes, cells of glial origin that are analogous to the Schwann cells of the peripheral nervous system (Fig. 1). At the microscopic level, white matter consists of collections of axons coated with myelin; these axons travel together in various fiber systems called tracts to form structural connections between many different gray matter regions.

At the macroscopic level, white matter makes up about 50% of the adult cerebrum. Three kinds of white matter fiber systems are recognized. First are the projection fibers, which travel corticofugally (from the cortex) or corticopetally (to the cortex) in relation to more caudal destinations. An example of a corticofugal tract is the corticospinal tract, and the optic radiation from the thalamus to the occipital lobe is a corticopetal tract. Second, there are several tracts of commissural fibers that connect the two hemispheres. The most prominent of these is the corpus callosum, a massive structure carrying approximately 300 million axons. Finally, various association fibers serve to connect cortical regions within each hemisphere. There are numerous short association fibers (arcuate or U fibers) that link adjacent cortical gyri, and there are five long association tracts—the superior occipitofrontal fasciculus, the inferior occipitofrontal fasciculus, the arcuate fasciculus, the cingulum, and the uncinate fasciculus—that connect more remote cortical areas (Fig. 2).

Two other anatomic observations are relevant. First is the close relationship of white matter with the frontal lobes, the largest of the human cortical regions: The cerebral white matter is most abundant under the frontal lobes, where all the major association tracts have one of their termini. These features imply that the frontal lobes exert a unique influence in part through the actions of the white matter. Second, there is a higher proportion of white matter compared to gray matter in the right hemisphere than in the left hemisphere. This asymmetry suggests a functional specialization of the hemispheres whereby the white matter in the right cerebrum plays a distinctive role.

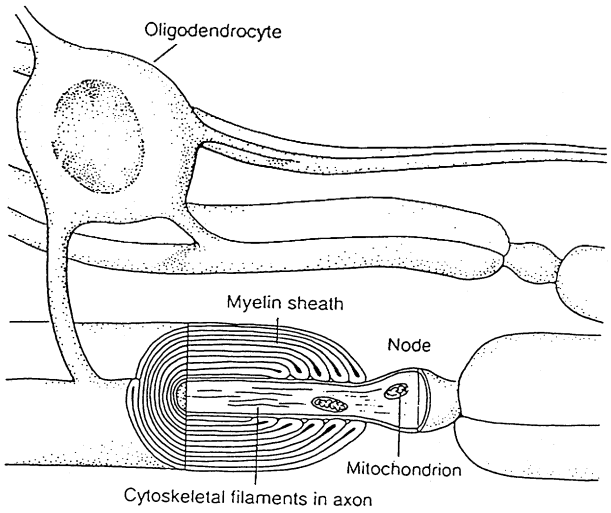


Figure 1 Schematic diagram of the relationship between the oligodendrocyte, myelin sheath, and axon in the brain (reprinted with permission from E. R. Kandel, T. M. Jessell, and J. H. Schwartz, *Principles of Neural Science*, 3rd ed., p. 44. McGraw-Hill, New York, 1991).

B. Neurophysiology

The brain is an electrical organ, and the conduction of electrical impulses from one neuron to another is fundamental to brain activity. A variety of neurotransmitters serve as chemical messengers between neurons, but transmission of impulses within each neuron is equally critical. The continuous process of neuronal communication depends on the integrity of white matter. Myelinated fibers are essential for the rapid and efficient propagation of neuronal information.

The basic physiologic phenomenon in neurons, whether in the central or peripheral nervous system, is the action potential. This is an electrical event in which the neuronal membrane is rapidly and transiently reversed from its resting potential of -70 mV by the influx of positively charged sodium ions. This "all-or-none" potential quickly ends with the efflux of potassium ions, and then follows a short time, the refractory period, during which no impulse can be

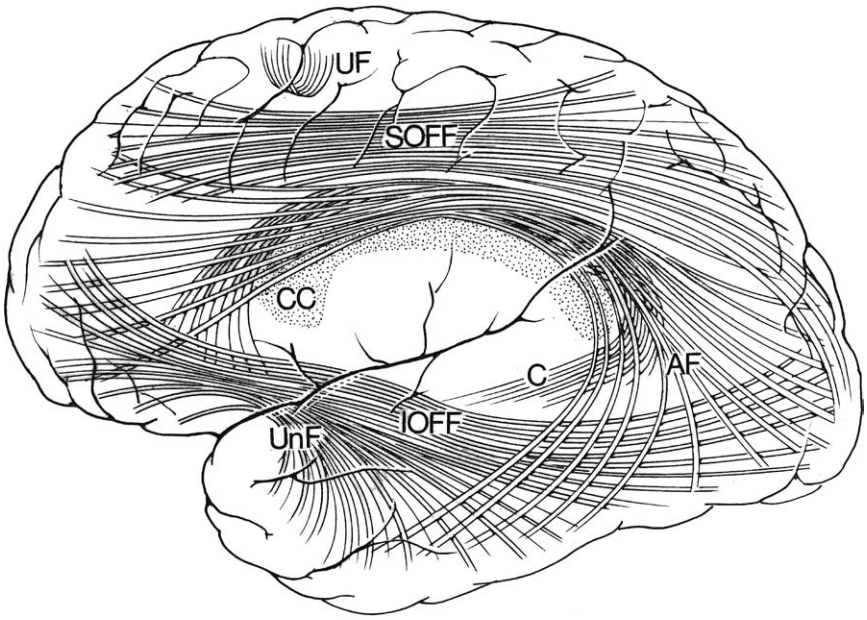


Figure 2 Major white matter tracts in the cerebral hemispheres. UF, U or arcuate fibers; SOFF, superior occipitofrontal fasciculus; IOFF, inferior occipitofrontal fasciculus; AF, arcuate fasciculus; C, cingulum, UnF, uncinate fasciculus; CC, corpus callosum (reprinted with permission from C. M. Filley, *Neurobehavioral Anatomy*, p. 188, Univ. Press of Colorado, Niwot, 1995).

propagated. The resting membrane potential is restored by the action of the sodium–potassium pump, an energy-dependent process requiring the hydrolysis of adenosine triphosphate (ATP), in which sodium is extruded from the cell and potassium taken in.

The primary role of myelin is to increase the conduction velocity of neurons as they propagate action potentials. Myelin is laid down in such a fashion that short segments of the axon, called nodes of Ranvier (or simply nodes), are left unmyelinated (Fig. 1). These nodes contain high concentrations of sodium channels, and the action potential can "jump" sequentially from one node to the next as the depolarizing current shunts the current rapidly forward along the axon (Fig. 3). This phenomenon, known as "saltatory conduction," greatly increases neuronal conduction velocity.

C. Development and Aging

The white matter of the brain undergoes a continuous process of remodeling throughout the life span. It has

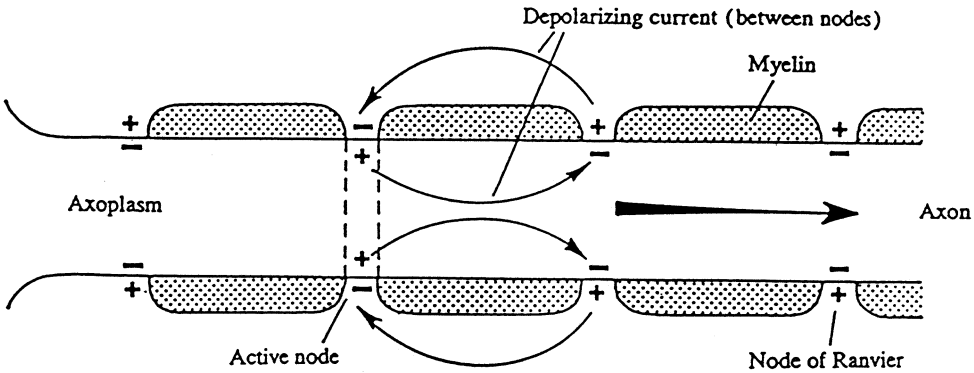


Figure 3 Illustration of saltatory conduction in a myelinated nerve. The action potential is propagated by sodium influx at the nodes of Ranvier, and the impulse "jumps" rapidly throughout the entire length of the axon (adapted from B. Pansky, D. J. Allen, and C. G. Budd, *Review of Neuroscience*, 2nd ed., p. 357. Macmillan, New York, 1988).

been estimated that the ratio of cerebral gray to white matter at age 50 is approximately 1 : 1, whereas in infancy and old age there is a relative preponderance of gray matter. In comparison to gray matter, white matter requires a longer time to develop and then is preferentially lost in senescence.

During development, white matter requires several decades to complete its formation. Myelination of the brain begins *in utero* and continues until adulthood. As a general rule, myelination occurs earlier in areas devoted to elemental motor and sensory functions and later in areas concerned with higher cerebral functions. Thus, the cerebral commissures and the association areas, for example, are not fully myelinated until the third decade of life. Because there is a strong likelihood that the degree of myelination parallels functional maturity, the fact that frontal lobe myelination is not complete until young adulthood suggests that the integrity of white matter contributes importantly to the maturation of personality and comportment.

In contrast, aging in the brain involves not only a loss of neurons and their dendritic trees but also a selective loss of white matter. Recent findings from neuroimaging and autopsy studies have led to a revision of the traditional view that neuronal loss is the major characteristic of brain aging. It now appears that whereas neuronal loss is not as marked as previously thought, gradual loss of white matter in aging is a consistent observation. The implications of this white matter change are not entirely clear, but as in development, there may be a clinical correlation. For example, the cognitive slowing often seen in the elderly may be explained in part by the loss of white matter.

D. Neuroimaging

The study of cerebral white matter has been dramatically facilitated in the past three decades by the development of powerful neuroimaging techniques. Computerized tomography (CT) scanning became available in the 1970s, and for the first time a noninvasive image of the brain *in vivo* became easily obtainable. Magnetic resonance imaging (MRI) was introduced in the 1980s and offered much improved visualization of brain structures, particularly white matter. MRI has been a most significant advance in the understanding of white matter disorders, revolutionizing the diagnosis of multiple sclerosis (MS) and rapidly expanding the list of disorders with white matter pathology that were previously difficult to detect. Another development in the field of MRI is the advent of diffusion weighting, a means by which specific white matter tracts can be identified. Recently, functional imaging techniques have attracted much interest, including positron emission tomography and functional MRI, both designed to evaluate the metabolic activity of brain regions. Although these instruments are most suited to imaging gray matter areas with high metabolic activity, they are likely to assist in the study of white matter function by identifying cortical and subcortical regions participating in neural networks that also include white matter tracts.

II. NEUROLOGIC DISORDERS

Disorders of the cerebral white matter comprise a widely diverse group of neuropathologic conditions. Many of these disorders are well-known in neurology, but others have been recently discovered with the use of MRI, and the number of entries on this list is likely to continue growing. This section reviews the cerebral white matter disorders using a classification system based on known or presumed etiology. Major categories are presented with the disclaimer that some diseases have an unknown etiology and are included in the category that is most appropriate given this uncertainty. Individual disorders are included if there is sufficient neuropathologic or neuroimaging evidence indicating that the cerebral white matter is involved. The intent is to provide an organizational framework for considering these disorders; exhaustive discussions can be found in standard textbooks of neurology.

A. Genetic Diseases

Many genetic diseases can affect the cerebral white matter. These are rare diseases that usually present in infancy or childhood but may occasionally appear in adulthood. Many of these diseases have a progressive and irreversible course, but effective medical treatment is available for some.

1. Leukodystrophies

The best known genetic cerebral white matter disorders are the leukodystrophies, a group of inherited diseases that disturb myelin metabolism and lead to severe neurologic dysfunction and early death. These diseases have been termed "dysmyelinative" because

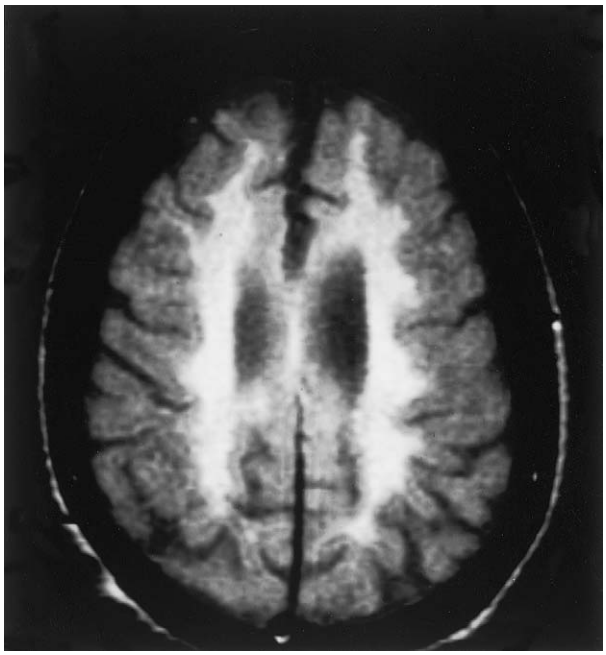


Figure 4 MRI scan of a patient with metachromatic leukodystrophy. There is severe and diffuse dysmyelination of the cerebral white matter (reprinted with permission from *Neuropsychiatr. Neuropsychol. Behav. Neurol.* **6,** 142, 1993).

of the metabolic dysfunction that prevents the normal myelination of the nervous system. The leukodystrophies usually present in early life and have an irreversibly progressive course. The most common of these is metachromatic leukodystrophy (MLD), an autosomal-recessive disease characterized by deficiency of the enzyme aryl sulfatase A; this defect leads to an accumulation of sulfatide in myelin, from which clinical deficits arise. Manifestations of MLD include mental retardation, dementia, spasticity, visual loss, and peripheral neuropathy, and in adults psychosis has been frequently observed. MRI is helpful in identifying the leukodystrophy, which is diffuse and severe (Fig. 4). Leukodystrophies are typically autosomal recessive; an exception is the X-linked disease adrenoleukodystrophy. Treatment has been supportive, since medical interventions have been largely unsuccessful. Recent studies of bone marrow transplantation with hematopoeitic stem cells, however, have shown promising results in MLD and globoid cell leukodystrophy.

2. Aminoacidurias

Phenylketonuria and maple syrup urine disease are inborn errors of metabolism that produce disturbed brain myelination. Cognitive and motor dysfunction are typical, and these have been attributed to white matter pathology. MRI has documented changes in myelination that were previously found on neuropathologic examination. If these diseases are detected in the neonatal period and treated with proper dietary management, an excellent outcome can often be achieved.

3. Phakomatoses

These disorders, also called neurocutaneous syndromes, all involve structural brain lesions in combination with an abnormality of the skin. The three most common phakomatoses are neurofibromatosis, tuberous sclerosis, and the Sturge–Weber syndrome. Although a variety of structural changes can be seen in the brains of these patients, white matter lesions are often seen on MRI scanning and may contribute to neurobehavioral dysfunction.

4. Mucopolysaccharidoses

Of the six major mucopolysaccharidoses, Hurler's syndrome and Hunter's syndrome are the two with the most prominent neurologic sequelae. Cerebral white matter neuropathology is present and may result both from the primary dysmyelination associated with the disease and from associated hydrocephalus.

5. Myotonic Dystrophy

This is an autosomal-dominant disease that has been recognized to affect cognition as well as the muscular system. Recent studies using MRI scanning have disclosed a high frequency of white matter lesions in the brains of myotonic dystrophy patients. Although the neuropathology of these lesions remains obscure, evidence suggests that they contribute to cognitive dysfunction.

6. Callosal Agenesis

This is an uncommon brain malformation that is usually idiopathic but thought to be of genetic origin in some cases. Mental retardation is common, although some individuals have relatively preserved cognitive function. The contribution of the corpus callosum malformation to impaired cognition is difficult to assess because these patients often have other problems, such as additional malformations, epilepsy, and hydrocephalus.

B. Demyelinative Diseases

This group of diseases has long been the most familiar of all the white matter disorders. Multiple sclerosis is the prototype white matter disease of the central nervous system, affecting the spinal cord as well as the brain, and has been a major clinical and research interest of neurologists since the time of Jean Marie Charcot, who made many important observations about the disease in the late 19th century. In part because they tend to affect patients at early stages in life, demyelinative diseases are a significant source of neurologic disability. Study of these diseases has also been stimulated by the likelihood that their pathogenesis can shed light on many aspects of neural dysfunction.

1. Multiple Sclerosis

MS is an inflammatory, demyelinative disease of the central nervous system. Young adults are primarily affected, and women are more vulnerable than men. The etiology of MS is unkown, despite many suggestions that it has a viral cause and that there are genetic influences on susceptibility. The disease is characterized clinically by lesions disseminated in space and time; that is, the lesions affect different regions of the brain or spinal cord and do so at different times. MRI has notably improved the detection of demyelinative lesions in the brain and has evolved to be the most useful method of confirming the diagnosis. Demyelinative plaques in MS are typically periventricular on MRI scanning and become confluent when the disease is severe (Fig. 5). The course of MS is most often relapsing–remitting, implying that patients have episodic exacerbations followed by complete or partial recovery over weeks or months. Others have primay progressive disease, in which continual decline is seen without recovery, and still others experience secondary progressive MS, in which an initial relapsing–remitting pattern is succeeded by a progressive course later in the disease. Common manifestations of MS are paresis, sensory loss, visual loss from optic neuritis, diplopia, vertigo, and bladder dysfunction. Neurobehavioral dysfunction involving memory loss, dementia, depression, emotional incontinence, euphoria, and psychosis are also well described and may be independent of elemental neurologic deficits. Life expectancy is only modestly reduced in MS, but the quality of life may be seriously compromised by loss of the ability to work and the disruption of social relationships. Until recently, treatment of MS was limited to supportive care and the use of corticosteroids to reduce the duration of exacerbations. In the past decade, three new immunomodulatory drugs have been added to the pharmacopoeia of MS patients—interferon-β_{1a}, interferon-β_{1b}, and glatiramer. These agents are given on a regular basis and can reduce relapse rate and possibly functional disability in individuals with relapsing–remitting MS.

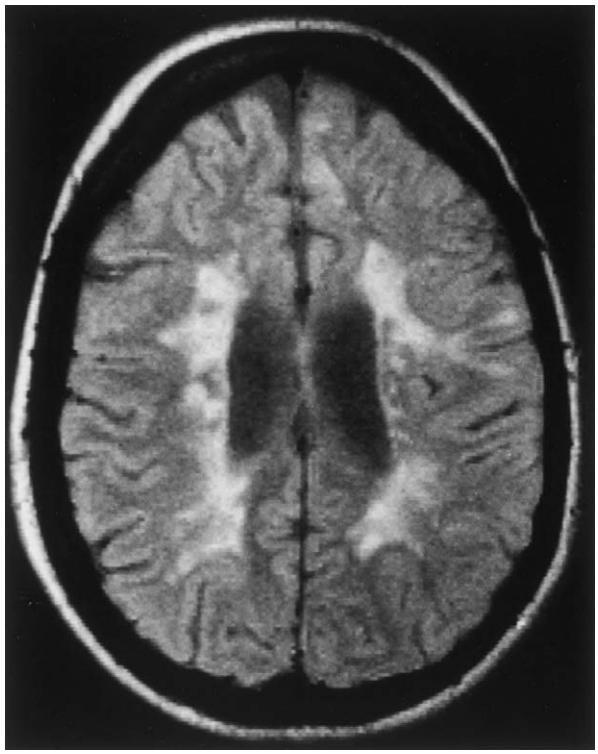


Figure 5 MRI scan of a patient with MS. Multifocal demyelination throughout the cerebral hemispheres is apparent (reprinted with permission from *Neuropsychiatr. Neuropsychol. Behav. Neurol.* **4,** 245, 1988).

2. Acute Disseminated Encephalomyelitis

Acute disseminated encephalomyelitis (ADEM) is a monophasic demyelinative disease of the brain and spinal cord. ADEM typically occurs in children and follows an infection with an exanthematous rash or a vaccination; the presentation is one of rapidly evolving multiple symptoms and signs related to extensive demyelination. The disease may be severe, with a substantial fatality rate and often significant neurologic or neurobehavioral disability in survivors.

3. Neuromyelitis Optica

This disease, also known as Devic's disease, is characterized by the acute onset of unilateral or bilateral optic neuritis in close temporal association with transverse or ascending myelitis. The disease often affects children but can be seen at all ages. The course may range in severity from a single episode from which the patient recovers to a fatal outcome. Neuromyelitis optica is probably an uncommon variant of MS because many survivors go on to develop typical features of the disease as time elapses.

4. Schilder's Disease

In 1912, the entity Schilder's disease was introduced into the literature as a fulminant disease of children with diffusely affected cerebral white matter. Much confusion has surrounded the status of this condition, but later analyses have securely identified three different diseases falling under this eponym: adrenoleukodystrophy, subacute sclerosing panencephalitis, and true Schilder's disease. The disorder is best considered a rare variant of MS usually affecting children and causing severe mental deterioration, progressive disability, and often death as a result of marked cerebral demyelination.

5. Balò's Concentric Sclerosis

This disease is an even rarer MS variant that resembles Schilder's disease clinically but features a different neuropathology. In contrast to widespread areas of myelin destruction, concentric sclerosis manifests as a series of alternating bands of myelin destruction and preservation in concentric rings. Corticosteroid therapy may be helpful in this disease as in Schilder's disease.

C. Infectious Diseases

Many different infectious diseases attack the cerebral white matter. In these diseases, an inflammatory pattern of white matter changes is seen and an infectious agent has been identified in each. The infections discussed later may in some cases involve white and gray matter structures alike, and it may be unclear how much of the clinical picture can be attributed to white matter pathology. However, there is in general a sufficient burden of white matter disease to merit consideration of these diseases in this section.

1. AIDS Dementia Complex

With the emergence of infection with the human immunodeficiency virus (HIV) in the past two decades, neurologic involvement in patients with the acquired immunodeficiency syndrome (AIDS) has been frequently recognized. HIV is a neurotrophic virus, and AIDS may now be considered a disease of the nervous system as well as the immune system. All parts of the nervous system may be affected by AIDS—the brain, meninges, spinal cord, nerve roots, peripheral nerves, and muscle—causing a wide array of neurologic syndromes. One of the most severe of these syndromes is the AIDS dementia complex, a typically late complication of AIDS characterized by progressive and irreversible cognitive and behavioral decline accompanied by a variety of motor signs, including spasticity, ataxia, and gait disorder. In terms of mental function, affected individuals manifest apathy, impaired memory, and cognitive slowness in the setting of progressively enlarging ventricles and patchy or diffuse leukoencephalopathy on neuroimaging studies. Neuropathologic studies have shown diffuse and multifocal rarefaction of the white matter and lesser involvement of subcortical gray matter. The syndrome is likely due to HIV invasion of the brain, although whether the virus exerts its effect directly or indirectly remains uncertain. Treatments for the AIDS dementia complex have been limited, and the onset of dementia usually heralds a fatal outcome soon thereafter.

2. Progressive Multifocal Leukoencephalopathy

Progressive multifocal leukoencephalopathy (PML) is an opportunistic infection of the cerebral white matter that frequently appears in patients with AIDS. The infectious agent is a member of the papavovirus family known as the JC virus. The virus infects oligodendrocytes, and demyelination is seen pathologically. Multifocal nonenhancing white matter involvement is typically seen on MRI scanning. The course of PML is invariably progressive, and attempts to treat the disease pharmacologically have been disappointing.

3. Varicella Zoster Encephalitis

The varicella zoster virus is a member of the herpes family that typically causes a painful disease of the peripheral nervous system and skin known as shingles. In immunocompromised patients, however, multifocal leukoencephalitis with both demyelination and

ischemia of the white matter may occur. Treatment with the antiviral drug acyclovir may be successful.

4. Subacute Sclerosing Panencephalitis

Subacute sclerosing panencephalitis (SSPE) is a rare disease of children and adolescents caused by the measles virus. Although the neuropathology may also involve cerebral cortex, white matter involvement is prominent. It remains unclear why the measles virus reactivates in some children to cause this devastating disease. A subacute course of progressive dementia advancing to death occurs, and no effective treatment to arrest or cure SSPE has been found.

5. Lyme Disease

Lyme disease is a systemic disorder due to infection with the spirochete *Borrelia burgdorferi*. Neurologic manifestations may be central or peripheral, and in the brain a fluctuating meningoencephalitis may occur. Because effective treatment with tetracycline is available, there have been few autopsied cases, but MRI scans have revealed white matter changes in the cerebrum and brain stem.

D. Inflammatory Diseases

These diseases are characterized by an inflammatory process in the brain unrelated to an infection. Because excessive immune system activity is implicated in each, corticosteroids are typically used in their treatment. As is true of infectious diseases, these disorders may involve gray matter in addition to white matter, but prominent white matter involvement has been documented as one aspect of each of the entities discussed here.

1. Systemic Lupus Erythematosus

Systemic lupus erythematosus (SLE), also known simply as lupus, is a multisystem disease with protean clinical manifestations. When SLE affects the brain, causing what has been variously known as lupus cerebritis or neuropsychiatric lupus, a wide range of neurobehavioral features may follow. In many cases, multifocal white matter lesions are seen on brain MRI that likely represent ischemic lesions and infarcts related to the vasculopathy that is characteristic of SLE. Treatment often includes immunosuppressive drugs and anticoagulation, but the efficacy of these modalities is uncertain.

2. Polyarteritis Nodosa

This connective tissue disease most often affects the peripheral nervous system and produces a syndrome of multiple nerve involvement known as mononeuropathy multiplex. The brain is less commonly involved, but there may be white matter infarcts on MRI related to arteritis.

3. Behcet's Disease

This is a systemic vasculitis featuring oral and genital lesions, intraocular lesions, skin and joint involvement, and thrombophlebitis in addition to neurologic manifestations. An acute or subacute meningoencephalitis may occur. MRI has been reported to reveal multifocal white matter lesions.

4. Sjögren's Syndrome

This uncommon autoimmune disease may have neurologic manifestations, the most frequent of which is a sensory neuropathy. The brain may also be affected, and the likely mechanism is small vessel disease from vasculitis or vasculopathy. MRI scans have revealed subcortical white matter hyperintensities in some patients with Sjögren's syndrome.

5. Isolated Angiitis of the Central Nervous System

Despite a lack of systemic markers of vasculitis or inflammation, this autoimmune disease may affect the brain to cause headache, confusion, and multifocal neurologic deficits. The pathogenesis involves a vasculitis that causes multiple strokes affecting white or gray matter.

6. Sarcoidosis

This is a systemic inflammatory disease with the characteristic pathologic finding of noncaseating granulomata. Neurosarcoidosis may affect the central nervous system and cause a wide range of neurologic manifestations. Periventricular abnormalities are seen on MRI in about one-third of patients with neurosarcoidosis; these may represent subependymal granulomas, areas of infarction secondary to granulomatous angiopathy, hydrocephalic edema from basal

meningeal involvement, or opportunistic infection such as PML.

E. Toxic Leukoencephalopathies

The cerebral white matter is vulnerable to a wide range of intoxications from physical and chemical insults. Of all the clinical areas in which MRI has been able to demonstrate white matter lesions, one of the most fruitful has been the toxic disorders. New drugs that affect the white matter, both therapeutic and illicit, are being introduced at a rapid rate into society, and their effect on white matter may be made visible by MRI.

1. Radiation

Cerebral irradiation is an established therapeutic modality in the treatment of primary and metastatic brain tumors. It has become clear in the past two decades that one of the side effects of brain irradiation is leukoencephalopathy. Three phases of white matter disease following irradiation are recognized. The first is an acute reaction, in which confusion or transient worsening of preexisting neurologic deficits occurs during treatment. CT or MRI scans show mild diffuse changes in the white matter that have been attributed to edema. This syndrome is fully reversible. The next phase is the early delayed reaction, which occurs weeks to months after treatment and may involve a variety of neurobehavioral deficits, including somnolence, apathy, memory loss, or dementia. More extensive white matter changes can be seen on neuroimaging studies, and cerebral demyelination is the neuropathologic correlate. This syndrome is also reversible in most cases. The most ominous reaction is the late delayed reaction, which develops 6 months to 2 years after treatment and features a progressive dementia that is due to widespread demyelination and necrosis. Treatment for all these phases is supportive, although corticosteroids are often employed with variable efficacy.

2. Therapeutic Drugs

As the development of new drugs for treating medical diseases steadily advances, so does the potential for toxicity from these agents. Many new drugs have been found to cause leukoencephalopathy, although the mechanism by which this toxicity occurs is obscure.

a. Chemotherapeutic Agents As in the case of cerebral irradiation, the white matter is the target of many of the pharmaceutical agents used in the treatment of cancer. The growing list of antineoplastic agents associated with leukoencephalopathy results in part from improved supportive care of cancer patients that has permitted higher doses and longer survival periods during which adverse effects may appear.

In general, these drugs cause a syndrome that is clinically, neuroradiologically, and neuropathologically similar to radiation leukoencephalopathy. Neurobehavioral features are particularly characteristic and may include lassitude, drowsiness, confusion, memory loss, and dementia. Methotrexate was the first antineoplastic drug found to cause leukoencephalopathy, which is its most common and significant toxicity. 1,3-Bis(2-chloroethyl)-1-nitrosourea is also well recognized to cause white matter changes. Other drugs have often been recognized as white matter toxins because of findings on MRI. Little is known of treatment other than minimization of exposure and supportive care.

b. Immunosuppressive Agents Organ transplanation has become commonplace in recent decades, and immunosuppression has become a necessary component of postoperative care to prevent organ rejection. Two drugs, cyclosporine and tacrolimus, have been associated with leukoencephalopathy. Reversibility has been observed with cessation of therapy.

c. Antimicrobials Two antimicrobial agents may result in leukoencephalopathy. Amphotericin B is a mainstay of fungal infection treatment, and white matter damage is one of its many toxicities. Hexachlorophene is an antiseptic agent that can cause this syndrome if absorbed through the skin in sufficient quantity.

3. Drugs of Abuse

The neurotoxicology of illicit drugs is not well understood in general. Drug abusers often abuse more than one agent, and it is difficult to determine the effects of one drug in isolation. There are examples, however, of relatively pure syndromes of impairment, and certain drugs have a known or putative effect on the cerebral white matter.

a. Toluene This drug, methylbenzene, is an organic compound widely used in industry and is the major solvent found in spray paints. Occupational

Figure 6 MRI scan of a patient with toluene leukoencephalopathy. The cerebral white matter is severely affected, and there is also marked ventricular enlargement and brain atrophy (reprinted with permission from *Neurology* **40,** 533, 1990).

exposure to toluene occurs in painters and other workers, but it is debated whether low-level toluene exposure has any deleterious effects. In contrast, toluene abuse may have dramatic sequelae. Intentional inhalation of toluene vapors from products such as spray paint leads to euphoria, and abuse of toluene over extended periods may produce severe leukoencephalopathy. Clinical studies of individuals who inhaled toluene fumes on a regular basis for months to years have documented dementia, ataxia, and various other neurologic deficits. Cognitive dysfunction is particularly devastating, and the severity of the neurobehavioral deficits approximately correlates with the duration of exposure to toluene. MRI scans reveal diffuse white matter change in the cerebrum (Fig. 6) and cerebellum, and autopsy studies disclose diffuse myelin loss without damage to cortical or subcortical gray matter. Abstinence arrests the neuropathologic process, but there may be little recovery of function if enough white matter damage has occurred. Toluene leukoencephalopathy offers one of the best examples of a major neurologic and neurobehavioral disorder caused by selective white matter involvement.

b. Ethanol The neurology of alcohol has received much attention, and many classic neurologic syndromes—including acute intoxication, withdrawal states, cerebellar ataxia, and peripheral neuropathy—are well recognized in alcohol abusers. However, the pathophysiology of alcohol's effects on the brain are not as well understood. A nutritional deficiency of thiamine (vitamin B_1) is known to cause the Wernicke–Korsakoff syndrome, in which acute confusion, ophthalmoplegia, and a gait disorder can herald the onset of severe amnesia; however, other mechanisms may also contribute to neurobehavioral dysfunction. Substantial evidence, both in humans and in laboratory animals, has been gathered for a direct toxic effect of alcohol on cerebral white matter. Thus, in contrast to the isolated amnesia of the Korsakoff's syndrome, it may be appropriate to consider the syndrome of alcoholic dementia, in which more widespread intellectual deterioration occurs in the setting of toxic white matter damage.

c. Heroin This narcotic drug does not cause leukoencephalopathy when abused by the usual routes of inhalation or injection, but the syndrome has been reported to appear after the inhalation of heroin pyrolysate, a form of the drug resulting from the heating of heroin on tin foil. Memory loss, dementia, akinetic mutism, cerebellar signs, and impaired gait may ensue. Neuroimaging and neuropathologic studies demonstrate white matter degeneration, but the specific toxic agent remains unknown.

d. Methylenedioxymethamphetamine Methylenedioxymethamphetamine (MDMA), also known as ecstacy, has recently become widely popular. Leukoencephalopathy has been reported in individuals exposed to MDMA, possibly related to its effects on serotonergic neurons.

4. Environmental Toxins

Although there are few environmental toxins that affect the white matter, others may be unrecognized. The chemical diversity of the toxins that are known suggests a variety of mechanisms.

a. Carbon Monoxide Carbon monoxide (CO) is a ubiquitous molecule in industrialized society, and human exposure may be accidental or related to suicide attempt. The usual picture of CO intoxication results from cerebral hypoxia that predominantly affects the neocortex, hippocampus, basal ganglia, and cerebellum. A syndrome of leukoencephalopathy may be seen in some patients, however, and typically

develops days or weeks after exposure. This syndrome may produce abulia, parkinsonism, dementia, and akinetic mutism. Recovery is variable, and a decrease in abnormal white matter signal on MRI may parallel clinical improvement.

b. Arsenic Arsenic neurotoxicity is rarely encountered and typically involves the peripheral nervous system. Acute and prolonged encephalopathy may also be seen, and available neuroimaging and neuropathologic information suggests that white matter hemorrhage and necrosis are responsible. Treatment with chelating agents may be effective.

c. Carbon Tetrachloride This is a widely used hydrocarbon in home and industrial settings. The primary toxicity of this chemical falls on the kidneys and liver, but neurotoxicity has occasionally been reported. Severe hemorrhagic white matter changes in the cerebrum, cerebellum, and brain stem may follow carbon tetrachloride exposure.

F. Metabolic Disorders

The cerebral white matter is vulnerable to certain metabolic disturbances. Although this category of disease is more likely to affect gray matter because of the great sensitivity of neuronal cell bodies to metabolic dysfunction, the white matter is particularly susceptible to some metabolic disturbances. The pathogenesis of these disorders is incompletely understood.

1. Cobalamin Deficiency

A lack of cobalamin, or vitamin B_{12}, is a well-known cause of the spinal cord disease known as subacute combined degeneration. Less commonly, cobalamin deficiency may cause neurobehavioral manifestations ranging from personality change to dementia. As in the spinal cord, the brain white matter is affected by vitamin B_{12} deficiency, a finding that may be visible on MRI scanning. Replacement of cobalamin can result in substantial clinical improvement and resolution of neuroradiologic abnormalities.

2. Folate Deficiency

Folate, or folic acid, deficiency is a less common cause of mental status change than cobalamin deficiency. Sporadic case reports indicate that neurobehavioral dysfunction can be seen with folate deficiency, and that replacement of the vitamin can be beneficial. There is also evidence that white matter changes in the brain result from a lack of this vitamin.

3. Hypoxia

As mentioned previously, hypoxia can occasionally cause damage to the cerebral white matter. Diminished oxygenation from any cause can lead to delayed postanoxic demyelination. If the patient survives, treatment is supportive, and the degree of recovery is inversely related to the severity of anoxic injury.

4. High-Altitude Cerebral Edema

This is a potentially fatal disorder seen in persons at high altitude who have acute mountain sickness or high-altitude pulmonary edema. Affected individuals may experience an acute confusional state and ataxia, often in association with white matter hyperintensities on MRI that are particularly prominent in the splenium of the corpus callosum. The syndrome is reversible with removal from altitude and supportive care, and it is thought to be due to vasogenic edema that has a predilection for the white matter.

5. Central Pontine Myelinolysis

In this disorder, most often seen in alcoholics, damage to brain stem and cerebral white matter is seen in patients who have a rapid correction of hyponatremia. Pronounced neurobehavioral and neurologic dysfunction can be observed. The condition is thought to stem from osmotic demyelination. Although early reports suggested a poor prognosis, later experience has demonstrated that recovery can often occur.

6. Marchiafava–Bignami Disease

This disease is also seen most frequently in alcoholics, but its etiology is unclear because nondrinkers may occasionally be affected as well. There is extensive demyelination of the corpus callosum, and often extracallosal white matter damage is also seen. Neurobehavioral deficits are paramount and recovery is variable.

G. Vascular Diseases

All areas of the brain, including both gray and white matter, are vulnerable to vascular insults. Vascular lesions of the cerebral white matter have recently been better recognized because of the ready availability of MRI scanning, and a growing understanding of the importance of these lesions is emerging. Most of the work in this area has been done with older adults, in whom white matter changes on MRI, whether pathologic or associated with normal aging, are frequently encountered.

1. Binswanger's Disease

In this disease, also known as subcortical arteriosclerotic encephalopathy, there is diffuse ischemic damage to the cerebral white matter along with a scattering of lacunar infarcts. Deep gray matter structures are also affected to a variable extent. Nearly all reported patients have had hypertension. Affected persons develop an insidious dementia with prominent neuropsychiatric features, focal neurologic signs, gait disorder, and incontinence. MRI scans have shown marked leukoencephalopathy, and the severity of clinical involvement is generally commensurate with the degree of white matter abnormality. The pathophysiology involves ischemic demyelination as a result of hypoperfusion of the cerebral white matter. Treatment is limited to prevention of disease progression by attention to cerebrovascular risk factors; specific pharmacologic treatment targeted to cholinergic and other systems has been considered, but its efficacy is unknown.

2. Leukoaraiosis

Leukoaraiosis (LA) is the term applied to the frequent finding in older persons of white matter changes on neuroimaging studies that take the form of low densities on CT and hyperintensities on T2-weighted MRI scans. Much controversy has swirled around both the origin and the significance of these findings. Many neurologically normal older people, for example, have apparently incidental white matter changes on CT or MRI. Neuropathologic studies of these changes have suggested that they are ischemic in origin, and that when they are more severe white matter infarction can be seen. Regarding clinical significance, it appears that white matter infarcts, even a single one, may have neurobehavioral sequelae if careful evaluation is undertaken, and that multiple infarcts clearly can lead to dementia. Whether white matter ischemia alone has neurobehavioral effects remains unclear, although evidence is accruing that these effects may be detectable when a certain threshold of involvement is reached. On this basis, it has been suggested that LA lies on the same continuum as Binswanger's disease.

3. CADASIL

Newly described in the past decade, CADASIL (cerebral autosomal-dominant arteriopathy with subcortical infarcts and leukoencephalopathy) is a genetic disorder of cerebral white matter in which ischemia and infarction slowly erode neurologic and neurobehavioral function. The disease is known to be caused by one of several mutations in the notch 3 region of chromosome 19, and characteristic granular osmiophilic material is seen in the walls of arterioles on electron microscopy. The clinical course begins in adulthood and may be characterized by neuropsychiatric dysfunction, migraine, stroke, and dementia. MRI scans regularly show leukoencephalopathy, even in individuals who are asymptomatic. Treatment is limited to supportive care and prevention of cerebrovascular risk factors that can exacerbate the process.

4. Migraine

This is a common syndrome of recurrent headache that is most often seen in young adults who are otherwise healthy. Even though these headaches are sometimes referred to as "vascular" headaches, the pathophysiology may not primarily implicate vascular mechanisms. It has been noted that migraineurs have slightly more MRI hyperintensities than do age-matched control subjects, and it is possible that episodic intracerebral vasoconstriction participates in their development. Clinical mainfestations of these changes have not been established, and no specific treatment other than migraine control is indicated.

5. Periventricular Leukomalacia

Recent improvements in neonatal care have led to the increasing survival of premature babies. One neurologic complication encountered in these very vulnerable infants is periventricular leukomalacia, a destructive disorder of white matter related to perinatal asphyxia. Extensive cerebral white matter injury can be seen that leads to hydrocephalus or brain atrophy. The cortex and deep gray matter are relatively normal, implying a

special vulnerability of white matter to asphyxia. Cerebral palsy and mental retardation are frequent long-term sequelae and may be severe.

6. Eclampsia

This is a disorder of pregnancy involving hypertension, proteinuria, and peripheral edema. In addition to seizures, which are characteristic of the disorder, affected patients may experience headache, visual disturbances, focal neurologic deficits, and mental status alterations including coma. MRI scans reveal white matter hyperintensities, often with a posterior predominance. These changes are reversible, indicating that they result from temporary extravasation of fluid into the white matter from vessels with a damaged blood–brain barrier.

7. Hypertensive Encephalopathy

As in eclampsia, this disorder is characterized by reversible changes in the cerebral white matter that represent edema rather than infarction. Hypertensive encephalopathy develops when blood pressure rises abruptly, and clinical features include headache, papilledema, nausea, vomiting, and mental changes ranging from confusion to coma. The exact pathophysiology is still debated, but control of elevated blood pressure effects prompt recovery.

8. Cerebral Amyloid Angiopathy

This disorder, often but not always seen in patients with Alzheimer's disease (AD), has long been known to be a cause of lobar hemorrhage. It has recently been recognized that deposition of amyloid in cerebral arterioles may also lead to leukoencephalopathy, presumably because of hypoperfusion in arterioles supplying the deep white matter. Thus, white matter changes that are sometimes seen in patients with AD may be related to this condition, and in other patients, the clinical and neuropathologic picture may resemble that of Binswanger's disease.

H. Traumatic Disorders

Physical injury to the brain may take many forms. In recent years, there has developed a better understanding of the range of neuropathology that may occur and the clinical consequences of these injuries. Damage to the cerebral white matter has emerged as a prominent component of brain injury.

1. Traumatic Brain Injury

The problem of traumatic brain injury (TBI) represents one of the most important in neurology and medicine. At least 500,000 new cases of TBI occur annually in the United States, many of which are severe enough to cause lifelong disability. In the majority who have less serious injuries, cognitive and emotional dysfunction often cause significant disruption during the process of recovery. TBI has many effects on the brain, depending on the type and severity of the injury and associated systemic injuries, but the most consistent neuropathology in TBI is a phenomenon called diffuse axonal injury (DAI). Experimental and clinical studies have confirmed that TBI causes shearing injury of axons mainly in the white matter of the cerebrum, corpus callosum, and brain stem. These changes can be seen on MRI in many cases and have been well documented in postmortem brains. DAI is present in brain injuries ranging from concussion to severe TBI, and the clinical severity is largely determined by the degree of shearing injury. DAI is thought to produce prominent disturbances of arousal, attention, memory, executive function, and comportment, all of which play a role in the clinical symptomatology of TBI of any severity. Treatment depends on degree of injury, but in general, rehabilitation measures are the mainstay; these may involve a combination of drug therapy, psychological assistance, and physical, occupational, and speech therapy. The outcome for mild TBI is generally favorable despite often long periods of seemingly slow progress, but moderate or severe TBI often portends a lifetime of problematic deficits in cognition, emotion, and behavior.

2. Shaken Baby Syndrome

In the past three decades, there has been growing awareness of the problem of child abuse involving the violent shaking of babies. This is a special category of TBI in which blunt head trauma may not occur but diffuse axonal injury is nonetheless present. Babies with this injury may present with signs ranging from lethargy and irritability to seizures and coma. Neuroimaging studies show diffuse white matter damage, and autopsy studies have documented diffuse axonal injury resulting from extreme rotational force. If infants survive such a catastrophe, the outcome may be lifelong severe impairment or lesser degrees of disability.

3. Corpus Callosotomy

This procedure causes an iatrogenic lesion and is undertaken in an effort to control intractable epilepsy. The rationale is that seizures that begin in one hemisphere can be contained therein by severing the callosal connections to the other hemisphere. Because naturally occurring lesions of the corpus callosum are uncommon, these patients have provided a source of investigation for callosal function. In general, the effects of such a dramatic procedure are surprisingly inapparent, and patients function relatively normally. The "unity of consciousness" is well integrated despite the two halves of the brain being anatomically separated. However, detailed testing may reveal callosal deficits, such as left unilateral ideomotor apraxia, agraphia, hemialexia, and tactile anomia.

I. Neoplasms

Brain tumors, whether benign or malignant, typically involve widespread cerebral areas that include both gray and white matter. This is particularly true of those tumors that induce surrounding edema and thus affect even more adjacent tissue. Thus, neoplasms in general are not well suited to considering the effects of discrete focal lesions on brain function. Nevertheless, some tumors have a predilection for gray or white matter, at least early in their course, and some preliminary clinical correlations can be made.

1. Gliomatosis Cerebri

This rare neoplasm has the greatest selectivity for the cerebral white matter of any brain tumor. Patients with this lesion present with personality changes, cognitive decline, focal neurologic signs, and seizures, and they are found to have diffuse neoplastic invasion of the white matter on neuroimaging studies. The origin of the tumor is debated because neuropathologic study has not identified any characteristic histopathology. Due to the high malignancy of this tumor, the prognosis is poor.

2. Diffusely Infiltrative Gliomas

The three types of glioma are the astrocytoma, the oligodendroglioma, and the ependymoma. All these malignancies arise from white matter, and a variety of subtle personality and cognitive changes may be detected before seizures, focal neurologic signs, and hydrocephalus develop. The white matter is also implicated in the neurobiology of gliomas by virtue of the fact that they spread throughout the brain along white matter tracts. The prognosis of these tumors is generally unfavorable despite vigorous surgical and medical treatment.

3. Primary Cerebral Lymphoma

In the usual case, this tumor is derived from B lymphocytes and may be either uni- or multifocal. Brain lymphoma tends to occur in patients who are immunocompromised because of AIDS or the chronic use of immunosuppressive drugs. Periventricular white matter involvement is common, and lymphoma often appears in the differential diagnosis of AIDS patients with white matter lesions on MRI. Personality and cognitive changes along with focal motor deficits are common with this lesion. Because of the high radiosensitivity of this tumor, the prognosis is relatively favorable.

J. Hydrocephalus

Hydrocephalus is a condition of an excessive fluid collection in the brain. Normally, the ventricles of the adult brain contain approximately 25 ml of cerebrospinal fluid, and an increase in this amount may be pathologic. There are situations in which hydrocephalus *ex vacuo*—excessive fluid that simply replaces lost brain tissue—may be of little consequence or even normal, but hydrocephalus that compromises brain function is clinically significant. The reason for this is primarily the effect of excess fluid on the cerebral white matter that surrounds the lateral ventricles.

1. Congenital Hydrocephalus

This term refers to hydrocephalus occurring *in utero* or early in the postnatal period that causes an actual increase in head size because the cranial bones of the skull are not fully fused until the end of the second year of life. Because the brain expands under internal pressure and the skull is soft and malleable, the head measurably enlarges. The common etiologies of this problem are germinal matrix hemorrhage in preterm infants, fetal and neonatal infections, Chiari malformation, aqueductal stenosis, and the Dandy–Walker syndrome. Symptoms may be irritability, poor feeding, torpor, and coma. If treatment with the use of a shunting procedure is not undertaken, mental retardation is frequent.

2. Occult Tension Hydrocephalus

In this situation, hydrocephalus becomes evident only after the cranial sutures have closed and the skull is no longer expandable. The disorder may actually be arrested and never clinically apparent, but in other cases there may be headache, cognitive slowing, inattention, perseveration, and other signs of frontal dysfunction. Eventually, gait disorder, dementia, and incontinence may develop. Shunting procedures may be very effective.

3. Normal Pressure Hydrocephalus

Normal pressure hydrocephalus (NPH) was first described in 1965 and generated substantial enthusiasm because it was thought to represent a dementia syndrome that could be treated successfully with shunting procedures. The clinical features of NPH include dementia, gait disorder, and incontinence, reflecting primary involvement of the subfrontal white matter. The disorder sometimes follows TBI, infection, or subarachnoid hemorrhage but is frequently idiopathic. Although undoubtedly some cases exist that have been successfully reversed with a shunt, these cases are disappointingly rare; moreover, complications of the shunt procedure are common. Nevertheless, NPH is a good example of dementia due to potentially reversible structural involvement of the cerebral white matter.

III. FUNCTIONS OF CEREBRAL WHITE MATTER

From the foregoing review of cerebral white matter disorders, it is apparent that a wide range of clinical manifestations can occur. Because disturbed functions can often be correlated with specific lesions, these clinical features provide an opportunity to consider the operations of the cerebral white matter in the normal brain. Although many questions remain unanswered about the role of white matter in the many different spheres of brain activity, it is possible to provide some detail on this topic and, by implication, a overview of the importance of white matter in general. These functions of cerebral white matter are differentiated between those that are neurologic and those that are neurobehavioral. This distinction is used to highlight the emerging understanding that white matter lesions may interfere with higher cognitive and emotional functions of the brain as well as elemental functions, such as vision, motor ability, and somatic sensation.

A. Neurologic Functions

Among its many functions, the brain is well-known to control bodily activities such as vision, movement, and sensation. These functions are also shared by many nonhuman animals, and the descriptor "elemental" is thus appropriate to categorize these in a broad sense. Cerebral white matter disorders clearly disturb these functions. The clinical significance of deficits in elemental function needs little explanation; patients with MS, for example, may suffer greatly from chronic problems, including blindness and paralysis. The neurology of white matter disease has mainly concentrated on these deficits, which may be disabling even in the absence of superimposed neurobehavioral deficits.

1. Vision

Of all the sensory systems, the greatest amount of neural tissue is devoted to vision, and blindness is one of the most disabling sensory losses. White matter plays a crucial role in vision; the optic tracts and the thalamocortical radiations convey visual input from the eyes to the occipital cortex for central processing. A variety of visual deficits may follow cerebral white matter lesions; examples include blindness from bihemispheric demyelinative disease in patients with MS and hemianopia from a vascular lesion in the occipital white matter.

2. Motor Function

The capacity to engage in motor activity is essential for successful existence. Motor function is subserved by a highly organized multilevel neural system in which cerebral white matter plays a vital role. The corticospinal and corticobulbar tracts provide controlling input to lower systems in the spinal cord and brain stem, respectively. Lesions of these tracts may produce paresis, paralysis, spasticity, incontinence, dysarthria, and dysphagia. A wheelchair-bound existence from the effects of motor system lesions is one of the most feared sequelae of cerebral white matter disorders.

3. Somatic Sensation

Just as the capacity for movement is critical, so is the ability to receive stimuli from the external environment. Cerebral white matter participates in the somatosensory system at the level of thalamocortical radiation that transmits somatic sensory information received by the thalamus to the parietal cortex.

Numbness, tingling, and impaired function can occur with lesions involving this sensory system.

B. Neurobehavioral Functions

The higher functions of the brain include the cognitive and emotional aspects of human behavior. These have traditionally been attributed to the activities of the cerebral cortex, and the term "higher cortical function" has become entrenched in much of the literature because of this assumption. However, much evidence has demonstrated that the cerebral white matter is also important in the mediation of higher brain operations. This evidence comes primarily from the study of neurobehavioral dysfunction in individuals with lesions of the cerebral white matter. It may therefore be appropriate to use the term "higher cerebral function" in describing the neurobehavioral repertoire of human beings. The white matter participates in distributed neural networks that mediate all aspects of higher brain function.

1. Cognition

Cognition can be considered generally to be those mental capacities such as attention, memory, language, visuospatial skills, and reasoning that are commonly referred to as "thinking." Whereas these abilities clearly cannot proceed normally without a functioning cerebral cortex, white matter also makes an important contribution to normal cognition by joining cortical and subcortical areas into unified neural networks that subserve neurobehavioral capacities. Therefore, white matter lesions interrupt normal behavior by disconnecting components of these networks, even though these components may themselves be intact. One of the most common neurobehavioral deficits in patients with white matter disorders is cognitive dysfunction, and in many patients this loss of intellectual ability reaches a level that meets criteria for dementia. This syndrome is due to widespread white matter involvement that disrupts many cognitive domains simultaneously. Evidence indicates that a specific neurobehavioral pattern may characterize this syndrome; this "white matter dementia" may involve deficits in sustained attention, memory retrieval, visuospatial skills, frontal lobe function, and psychiatric status, with relative preservation of language, procedural memory, and extrapyramidal function. This profile of deficits and strengths is consistent with that which would be expected from the higher concentration of white matter in the frontal lobes and the right hemisphere. In other cases, in which there is more focal white matter involvement, specific neurobehavioral syndromes have been convincingly described, including amnesia, aphasia, alexia, apraxia, agnosia, executive dysfunction, and callosal disconnection. Thus, depending on its location, cerebral white matter disease may interfere with cognition as a whole or with selected cognitive domains.

2. Emotion

Emotion has long defied precise definition, but it is useful in a general sense to regard emotion as those mental activities approximately equivalent to "feelings." Emotions are powerful motivators in human life and clearly have their origin in brain activity. The neurosciences heretofore have not devoted as much attention to emotions as to cognition, but data are accumulating to indicate that this large group of human afflictions can be understood in terms of disorders of the brain, including those affecting cerebral white matter. White matter systems in the frontal and temporal lobes are most directly involved with emotion, and in particular, those tracts intimately connecting structures of the limbic system are thought to be of primary importance. Many individuals with cerebral white matter lesions experience changes in emotional status. Depression appears to be the most common syndrome, and it may result from white matter lesions interfering with the activity of neurotransmitter systems mediating mood as well as from the patient's psychological reaction to the disease. Other syndromes commonly encountered include bipolar disorder, emotional incontinence, euphoria, and psychosis. Because the study of neurobehavioral aspects of white matter is still in an early stage, more recognition and understanding of these and other syndromes is likely in the future.

IV. CONCLUSION

Disorders of cerebral white matter are common, and their clinical sequelae have been found to range from the subtle to the catastrophic. The number of disorders affecting this subdivision of the brain is substantial, and more are likely to be added as new clinical, neuroradiologic, and neuropathologic information appears. Although elemental neurologic dysfunction has been well established and is important in the

clinical manifestations of these disorders, a growing recognition of neurobehavioral dysfunction has emerged in recent years. Thus, the clinical impact of cerebral white matter disorders may be more significant than is currently appreciated.

Much work remains to be done on the etiology, pathogenesis, and treatment of these disorders. Given the wide variety of neuropathology that can occur, these tasks will require the contribution of neuroscientists and clinicians from many specialty areas. However, as a general rule, the prognosis for white matter disorders appears to be more favorable than that for gray matter diseases. In contrast to many disorders of gray matter that involve irrevocable destruction of neuronal cell bodies, there is frequent sparing of axons despite damage to myelin, and effective treatment seems more feasible. This therapeutic opportunity is another incentive to pursue the study of this group of disorders.

Finally, the cerebral white matter disorders offer a unique insight into brain function. The profound impairments that can attend lesions of white matter serve to illustrate that these tracts are essential components of distributed neural networks that subserve both elemental and higher cerebral functions. This principle is particularly well applied to the phenomena of cognition and emotion, in the operations of which the white matter plays an indispensible role. By considering the deficits that follow white matter lesions, an elegant view of the structure and function of the normal brain can be achieved. The prospects for continued advances in this area are bright, and these advances promise to assist both in the understanding of the human brain and in the alleviation of some its more tragic afflictions.

See Also the Following Articles

AGING BRAIN • BEHAVIORAL NEUROGENETICS • CRANIAL NERVES • DEMENTIA • HIV INFECTION, NEUROCOGNITIVE COMPLICATIONS OF • HYDROCEPHALUS • LYME ENCEPHALOPATHY • MULTIPLE SCLEROSIS • NERVOUS SYSTEM, ORGANIZATION OF • NEURAL NETWORKS • NEUROANATOMY

Suggested Reading

Adams, J. H., Graham, D. I., Murray, L. S., and Scott, G. (1982). Diffuse axonal injury due to nonmissile head injury: An analysis of 45 cases. *Ann. Neurol.* **12,** 557–563.

Adams, R. D., and Victor, M. (1999). *Principles of Neurology*, 6th ed. McGraw-Hill, New York.

Caplan, L. R. (1995). Binswanger's disease—Revisited. *Neurology* **45,** 626–633.

Del Bigio, M. D. (1993). Neuropathological changes caused by hydrocephalus. *Acta Neuropathol.* **85,** 573–585.

Edwards, M. K. (Ed.) (1993). *Neuroimaging Clinics of North America: White Matter Diseases*. Saunders, Philadelphia.

Filley, C. M. (1998). The behavioral neurology of cerebral white matter. *Neurology* **50,** 1535–1540.

Filley, C. M., and Kleinschmidt-DeMasters, B. K. (2001). Toxic leukoencephalopathy. *N. Engl. J. Med.* **345,** 425–432.

Geschwind, N. (1965). Disconnexion syndromes in animals and man. *Brain* **88,** 237–294, 585–644.

Menkes, J. (1990). The leukodystrophies. *N. Engl. J. Med.* **322,** 54–55.

Mesulam, M.-M. (1990). Large-scale neurocognitive networks and distributed processing for attention, language, and memory. *Ann. Neurol.* **28,** 597–613.

Miller, A. K. H., Alston, R. L., and Corsellis, J. A. N. (1980). Variation in age in the volumes of grey and white matter in the cerebral hemispheres of man: Measurements with an image analyser. *Neuropathol. Appl. Neurobiol.* **6,** 119–132.

Navia, B. A., Jordan, B. D., and Price, R. W. (1986). The AIDS dementia complex: I. Clinical features. *Ann. Neurol.* **19,** 517–524.

Nolte, J. (1999). *The Human Brain,* 4th ed. Mosby, St. Louis.

Noseworthy, J. H. (1999). Progress in determining the causes and treatment of multiple sclerosis. *Nature* **399** (Suppl.), A40–A47.

Yakovlev, P., and Lecours, A. R. (1967). The myelogenetic cycles of regional maturation in the brain. In *Regional Development of the Brain in Early Life* (F. Minkowski, Ed.), pp. 3–70. Blackwell, Oxford.

Cerebrovascular Disease

JOHN R. ABSHER
Absher Neurology and Wake Forest University Baptist Medical Center

GLOSSARY

autoregulation The brain's ability to control blood flow across a wide range of systemic blood pressure.

embolism The passage of clot, bubbles, or other discrete objects through the bloodstream.

infarction Irreversible tissue damage resulting from inadequate oxygen delivery, such as that due to hypoxia or ischemia.

ischemia Insufficient blood flow leading to reversible tissue injury that may progress to infarction or resolve without causing tissue damage.

neuroprotection Drug treatments that prevent, inhibit, or block processes that damage the nervous system.

thrombolysis The process of dissolving blood clots in order to restore blood flow.

thrombosis Coagulation of blood, the process of clotting, resulting in thrombus formation.

Cerebrovascular diseases are disorders or diseases of brain (cerebral) blood vessels, cerebral blood flow (CBF), or cerebral oxygenation. The vessels that supply blood to and from the brain constitute the cerebral vasculature. The brain typically derives its blood supply from four large arteries. Smaller arteries and arterioles supply the microvasculature. The microvasculature drains into intracerebral veins that then drain into the venous sinuses. The venous sinuses empty into the deep veins of the neck. Cerebrovascular disorders may originate within any segment of the cerebral vasculature, from cardiac or arterial disease, systemic emboli, hypercoagulable states, hypotension, and many other conditions. Even when CBF is normal and the vessels are healthy, oxygen delivery or oxygen metabolism may be inadequate to maintain brain viability. In this article, current concepts and advances in cerebrovascular disease are summarized.

I. INTRODUCTION

The field of cerebrovascular disease is focused largely on stroke, the sudden interruption of cerebral blood flow (CBF) or oxygen supply that results in brain failure. The field is concerned with chronic, progressive conditions that may limit blood flow and oxygenation (e.g., high blood pressure, pulmonary disease, and arteriosclerosis) as well as acute, devastating events (e.g., sudden occlusion of a cerebral artery due to cardiogenic cerebral embolism or ruptured aneurysm with brain hemorrhage). Risk factors contributing to stroke, such as chronic hypertension and diabetes, are encompassed by the study of cerebrovascular disorders. All events surrounding the actual stroke and its immediate management are thus central to the field. This includes all manner of testing related to the detection, treatment, and understanding of brain damage and dysfunction due to cerebrovascular disease. The prevention and post-acute management of stroke, including risk factor modification and rehabilitation, are also fundamental to the field. The cerebrovascular disorders are ubiquitous and may affect individuals of any age, and the study of cerebrovascular disease is a multidisciplinary medical specialty.

Advances in clinical diagnosis, testing, and therapy have led to substantial gains in the field of cerebrovascular disease, and these gains are emphasized in this article. Large epidemiologic studies have improved our understanding of demographics and stroke risk factors. Knowledge about the underlying causes of stroke and the mechanisms of tissue injury has prompted interest in the primary prevention of stroke through risk factor modification. Neuroprotective interventions are being developed to stop the process of tissue injury during the stroke, and tissue plasminogen activator (tPA) is now available for patients with acute stroke. Multidisciplinary stroke care teams and specialized stroke units have evolved that include interventional neuroradiologists, neurosurgeons, neurointensivists, and neurologists. Functional restoration and cognitive rehabilitation have become important aspects of neurorehabilitation. Novel methods for predicting which patients are most likely to benefit from therapy are emerging, such as transcranial magnetic stimulation and functional magnetic resonance imaging (fMRI). The detection and characterization of cerebrovascular disease are facilitated by the growth of MRI. There are now several advanced neuroimaging techniques for cerebrovascular diseases, including MR angiography (MRA), diffusion weighted imaging, perfusion weighted imaging, and MR venography (MRV). It is now possible to distinguish acute from chronic cerebral ischemia, and to estimate the age of an intracerebral hemorrhage. Cerebral blood flow may be quantitated using computed tomography (CT), single photon emission computed tomography (SPECT), and positron emission tomography (PET). Cerebral blood flow may be estimated using transcranial Doppler (TCD) and MRA. These and other advances have substantially lowered the risk of stroke and improved the outlook for those who experience a stroke.

A. Financial and Societal Significance

Advances in health care policy research have clarified the costs associated with the cerebrovascular diseases. Direct costs include hospital and nursing home care expenditures, physician and other professional service costs, pharmaceutical costs, home health care costs, and the cost of durable medical equipment. Indirect costs include lost productivity from work due to morbidity and mortality. The direct costs of stroke are approximately $30 billion annually, and indirect costs are approximately $20 billion (i.e., $50 billion annually in the United States alone). For Americans aged 40 and older, the average in-hospital and physician costs for stroke were $11,010 based on 1995 data. In 1996, the average cost of a stroke from hospital admission to discharge was $18,244 for people under the age of 65. Table I shows these costs in relationship to other vascular diseases and emphasizes the tremendous financial impact of stroke on society.

B. Cerebrovascular Disease Demographics

Cerebrovascular disease is estimated to affect 600,000–730,000 individuals each year in the United States, and it is currently the third leading cause of death behind cardiovascular diseases and cancer. Seventy-two percent of all persons with stroke are at least 65 years of age, and 29% of these individuals die within a year of the stroke. Although the stroke death rate decreased by 13.9 (per 100,000 people per year) from 1987 to 1997, the prevalence of stroke (the number of people alive who have had one or more strokes) increased by 6.6% during the same period. This paradox relates to the demographic shift in the population, the expansion of the proportion of the population aged 65 and older. Thus, although the age-specific incidence and

Table I
Costs and Demographics of Stroke[a]

	Cost (billions)				
Disease	Direct	Indirect	Total	Annual deaths	Prevalence
Coronary heart disease	55.20	63.00	118.20	466,101	12,200,000
Stroke	30.60	20.70	51.30	159,791	4,400,000
Hypertensive disease	26.10	11.10	37.20	42,565	50,000,000
Congestive heart failure	20.30	2.20	22.50	45,419	4,600,000
All cardiovascular disease	185.80	140.80	326.60	953,110	59,700,000

[a]Data are used with permission from the *2000 Heart and Stroke Statistical Update*. Copyright American Heart Association.

prevalence of stroke is declining (i.e., the incidence and prevalence at a specific age), overall stroke prevalence continues to climb.

The frequency of stroke varies considerably from one geographic area to another. In the Russian Federation, the annual stroke death rate per 100,000 people is 374 for men and 231 for women aged 35–74. This was the highest rate for both men and women among the data reported in the 1998 World Health Statistics Annual. The country with the lowest risk is Switzerland (for both men and women), where the annual stroke death rates are less than 10% of those reported for the Russian Federation. Within the United States, the state with the highest stroke death rate is South Carolina (87.9 per 100,000 people); its rate is nearly twice as high as that of New York, the state with the lowest stroke death rate (45.1). Several southern states besides South Carolina have high stroke death rates, including its closest neighbors, North Carolina (81.5) and Georgia (75.6) (Fig. 1).

C. Stroke Risk Factors

Stroke risk factors are very well understood. Both modifiable and nonmodifiable risk factors are important. The nonmodifiable stroke risk factors are age, ethnicity, and sex. There are several modifiable stroke risk factors, including hypertension, hypotension, diabetes, atrial fibrillation, congestive heart failure, coronary artery disease, smoking, hyperlipidemia, homocysteinemia, obesity, alcohol intake, and sedentary lifestyle. In addition, there are numerous predisposing factors, or stroke etiologies, including genetically inherited hematologic conditions (e.g., sickle cell disease and protein C deficiency), vascular malformation, connective tissue diseases, vasculitis, substance abuse, trauma with arterial dissection, and patent foramen ovale (Table II).

The most important modifiable stroke risk factor, hypertension, is estimated to affect approximately 50 million Americans (23% of the U.S. population), or 44% of Americans at least 65 years of age. Approximately one-third of all strokes directly relate to the impact of hypertension. Hypertension increases the risk of stroke up to 400%; the relative risk of stroke among persons with hypertension is four times higher than that among individuals lacking this risk factor. As many as 246,500 strokes might be prevented by eliminating the adverse impact of hypertension (i.e., 35% of 700,000 annual strokes). The odds that a person with hypertension will die of stroke decreases from 4 at age 50 to 1 by age 90, suggesting that the impact of hypertension declines considerably with age. The prevalence of hypertension among individuals with stroke increases substantially with age, from about 45% at age 50 to 70% at age 70. Thus, if one lives long enough with hypertension, there is a very high probability that one or more strokes will occur, either silently (without the patient's knowledge) or symptomatically (with evident problems such as paralysis, numbness, and unsteadiness).

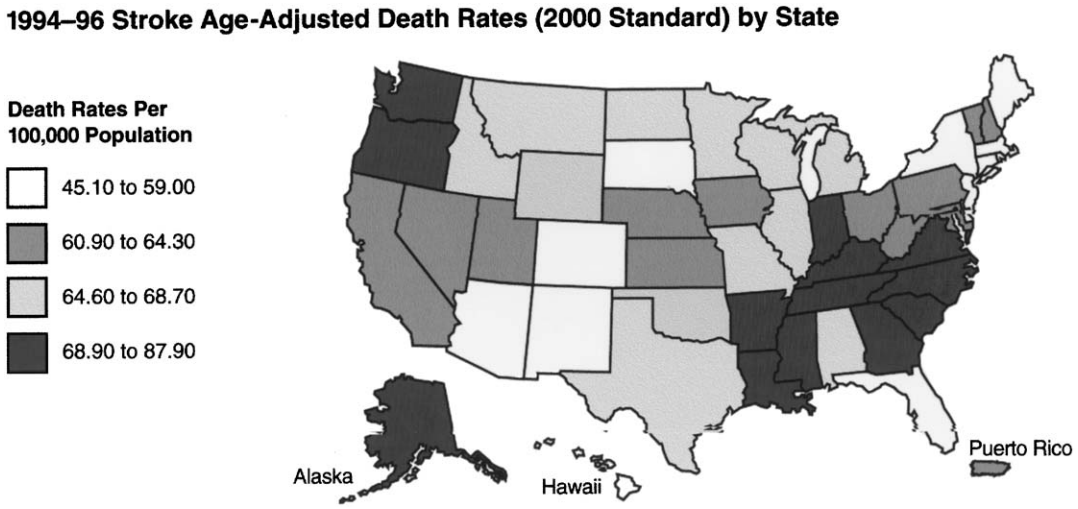


Figure 1 US map showing stroke rates by state. The annual stroke death rates for each state are depicted. Note the "stroke belt" within the southeastern U.S. (reproduced with permission, © *2000 Heart and Stroke Statistical Update*. Copyright American Heart Association).

Table II
Major Causes of Stroke

- Cardiogenic stroke
  - Atrial fibrillation
  - Endocarditis
  - Cardiac valve disease
  - Rheumatic heart disease
  - Cardiomyopathy
  - Acute myocardial infarction
  - Cardiac surgery
  - Congestive heart failure
  - Cardiac arrest
  - Patent foramen ovale
- Extracranial arterial disease and stroke
  - Aortic arch arteriosclerosis
  - Carotid artery arteriosclerosis
  - Vertebral artery arteriosclerosis
  - Carotid artery dissection
  - Vertebral artery dissection
  - Giant cell arteritis
  - Subclavian steal syndrome
  - Fibromuscular dysplasia
- Intracranial arterial disease and stroke
  - Aneurysm
  - Arteriovenous malformation
  - Cerebral angiitis (vasculitis)
  - Arteriosclerosis
  - Cerebral amyloid angiopathy
  - Migraine
  - Vasospasm
  - Toxic vasculopathy
- Extracranial venous disease and stroke
  - Internal jugular vein occlusion
  - Venous thrombosis with paradoxical embolization
- Intracranial venous disease and stroke
  - Central vein thrombosis
  - Sagittal sinus thrombosis
- Respiratory disease and stroke
  - Sleep apnea
  - Respiratory arrest
  - Chronic obstructive pulmonary disease
  - Myasthenia gravis
  - Motor neuron disease
  - Myopathy
  - Myositis
  - Botulism

(*continues*)

(*continued*)

  - Neuromuscular blockade
  - Neuropathy
  - Brain stem stroke
  - Cervical spine trauma
  - Seizures
  - Emphysema
  - Pneumonia
  - Pulmonary embolism
  - Poisoning
- Microvascular disease of the brain
  - Advanced age
  - Chronic hypertension
  - Diabetes mellitus
  - Smoking
  - Hypotension
  - Amyloid angiopathy
- Hematologic disorders and stroke
  - Leiden factor V mutation
  - Protein C deficiency
  - Protein S deficiency
  - Antithrombin III deficiency
  - Anticardiolipin/antiphospholipid antibody syndrome
  - Sickle cell anemia
  - SC disease
  - Polycythemia rubra vera
  - Thrombocytosis
  - Hemophilia
  - Thrombocytopenia
  - Hepatic failure
  - Disseminated intravascular coagulation
- Other
  - Inflammatory conditions
    - Beçet's syndrome
    - Moyamoya disease
    - Wegener's granulomatosis
    - Polyarteritis nodosa
    - Lymphomatoid granulomatosis
    - Takayasu's arteritis
    - Systemic lupus erythematosus
    - Scleroderma
    - Rheumatoid arthritis
- Traumatic
  - Carotid cavernous fistula
  - Air embolism
  - Fat/marrow embolism

(*continues*)

(*continued*)

Increased intracranial pressure
Infection
Herpes zoster
Vascular anomalies
Cavernous sinus syndrome
Atrial septal aneurysm
Cavernous malformations
Telangiectasia
Venous malformations
Neoplastic
Nonbacterial thrombotic endocarditis
Atrial myxoma
Malignant atrophic papulosis
Hypercoagulable states
Cryoglobulinemia

Stroke is more common among individuals with diabetes mellitus. When stroke does occur, it is more likely to be fatal among diabetics than among nondiabetics. After controlling for other stroke risk factors, the relative risk of stroke is 1.5–3.0 among diabetics, and the impact is greatest for elderly women. Diabetes affects about one-fifth of the population, and the attributable risk of stroke is about half that of hypertension.

The prevalence of diabetes differs by racial group. For example, the prevalence of physician-diagnosed diabetes in adults aged 20 or older is 5.4 and 4.7% for non-Hispanic white men and women, respectively; for non-Hispanic blacks, the percentages are 7.6 and 9.5 for men and women; for Mexican Americans, the prevalence of diabetes is 8.1% for men and 11.4% for women. Approximately half of American Indian women aged 45–74 have diabetes. Thus, the risk of diabetes is considerably higher among minority populations than among non-Hispanic whites.

Smoking is the leading preventable cause of death in the United States and a major stroke risk factor. The risk of stroke from any cause is 1.5, but the risk of subarachnoid hemorrhage is 10 times greater among smokers than among nonsmokers. The number of cigarettes per day correlates with stroke risk. After smoking cessation, the stroke risk returns to baseline within 2–5 years.

The Honolulu Heart Program study revealed a continuous and progressive increase in coronary heart disease and thromboembolic stroke rates with increasing cholesterol levels. When the highest and lowest quartiles were compared, the relative risk of stroke was 1.4. This study helped to confirm the long-held suspicion that hyperlipidemia is an independent risk factor for stroke. Two large studies examined a lipid-lowering drug (provastatin) compared to placebo and showed a 19–31% reduction in stroke. Given the strong biological plausibility of the relationship, and the beneficial effects of cholesterol reduction on stroke risk, it seems clear that hyperlipidemia is a significant stroke risk factor.

Low blood pressure is also an important modifiable stroke risk factor. If blood pressure is insufficient to maintain blood flow to the brain, stroke will result. Certain vascular territories may be more prone to cerebral ischemia. For example, intracranial arteriosclerosis impairs the brain's ability to regulate CBF and cerebral perfusion pressure downstream from a significant narrowing (i.e., cerebral autoregulation). Chronic hypertension impairs autoregulation and increases the risk of stroke within the microvasculature. Hypertensive women seem to be more susceptible than men to stroke and silent stroke as a result of nocturnal blood pressure decline, which has been associated with the accumulation of cerebrovascular damage in many studies. Certain types of antihypertensive medications (short-acting calcium channel blockers) are more likely to lead to significant nocturnal blood pressure decline than others. Currently, there is no reliable way to determine whether blood pressure might be low enough to increase stroke risk or to determine whether specific vascular territories might be more sensitive to blood pressure decline than others.

D. Ethnic and Racial Issues

African Americans are the largest minority group in the United States, although by the Year 2009, Hispanic Americans are expected to outnumber African Americans. African Americans are far more likely to die of stroke than are whites or Hispanic Americans, and this is true for both ischemic and hemorrhagic cerebral infarction. Hispanic Americans are no more likely than non-Hispanic whites to die of ischemic infarction, but they are more likely to die of intracerebral hemorrhage. The reasons for these differences in stroke risk are incompletely understood but may relate in part to the higher prevalence of cardiovascular risk factors in the African American population. Other factors, such as socioeconomic status, educational

level, and genetic predisposition, may contribute to the high rate of stroke among African Americans.

In general, African Americans are more likely to have hypertension, diabetes mellitus, obesity, congestive heart failure, and a smoking history than are whites. Furthermore, stroke risk factors differ from one subgroup of African Americans to another. For example, hypertension affects 16% of West African blacks, 26% of Caribbean blacks, and 33% of U.S. blacks. Socioeconomic status is correlated with lifestyle and health outcomes, with lower socioeconomic groups being more likely to have poor diet, poor cardiovascular risk factor control, and lower physical activity level. However, the influence of education, income, and other socioeconomic surrogates of stroke risk is complex and requires further exploration.

Genetic influences are likely to contribute to excess stroke risk among African Americans. Sickle cell disease is a genetically inherited hemoglobinopathy that accounts for a substantial number of strokes, and the disorder is more prevalent among blacks than other racial groups. Ten percent of individuals with sickle cell anemia will have at least one stroke, and 47–93% will have recurrent stroke, with 80% of recurrences occurring within the first 36 months. Three-fourths of the strokes are ischemic, and one-fourth are hemorrhagic. Hemorrhage is more likely to occur in the elderly, whereas ischemic infarction affects children and young adults.

The presence and severity of intracranial atherosclerosis are also likely to relate to genetic factors; African Americans, Hispanic Americans, Japanese, and Chinese individuals are more likely than whites to have stroke on the basis of intracranial atherosclerosis. Extracranial disease (e.g., common carotid artery and internal carotid artery) is more prevalent among whites. Attempts to explain these racial differences have failed to identify other factors, besides perhaps sex.

Although there are clear racial differences in the occurrence of stroke subtypes, including stroke due to intracranial atherosclerosis, stroke risk factor variations across ethnic groups may be partly responsible. For example, intracerebral hemorrhage, subarachnoid hemorrhage, and lacunar infarction (small strokes <2 cm in diameter) are more common among African Americans, whereas cardioembolic strokes are less frequent. Lacunar infarctions directly relate to hypertension, diabetes, and smoking, which are all more common among African Americans than among whites. The risk of intracerebral hemorrhage may also relate to the greater prevalence of hypertension, but the reason for the twofold greater risk of subarachnoid hemorrhage among African Americans remains unexplained. It is also uncertain why cardioembolic strokes are less frequent among African Americans than among whites.

The epidemiology of stroke among Hispanic Americans is quite different from that of African Americans. Overall, Hispanic Americans are less likely than non-Hispanic whites to have a stroke. Death from stroke is more likely among younger Hispanic Americans, and less likely among the elderly Hispanics, in comparison to non-Hispanic whites. Hemorrhages are more common among young Hispanics than among non-Hispanic whites or Native Americans. The presence of hypertension as a stroke risk factor is higher (79%) among Hispanics than among non-Hispanic whites (63%), according to the Northern Manhattan Stroke Study, after controlling for the influence of socioeconomic status. Diabetes is more prevalent among Hispanic Americans (41%) than among non-Hispanic whites (26%), but cardiac disease was nearly twice as common (61 vs 32%) among non-Hispanic whites. Cholesterol levels are generally lower for Hispanic Americans, but alcohol consumption was more prevalent (37 vs 23%). Since alcohol consumption is related to the risk of intracerebral hemorrhage, this difference may partly explain the higher rate of intracerebral hemorrhage among Hispanic Americans.

Several other factors may uniquely affect the Hispanic American population. They have the highest poverty rate (30.3%) and are more likely than non-Hispanic whites to have 5 years of schooling or less (10.3%). Twenty-nine percent of Hispanic Americans speak English poorly or not at all, and 10% speak only Spanish. The language barrier may account for the observation that Hispanic Americans with acute stroke reach the emergency room with the same rapidity as non-Hispanic whites but access to neurological care is delayed. Hispanic Americans are the least likely racial group to be eligible for intravenous thrombolysis, which must be administered within 3 hr of stroke onset.

E. Impact on Health and Quality of Life

Stroke produces devastating effects on quality of life. In one large study, more than half of individuals at risk for stroke perceived "major stroke" to be "equivalent to or worse than death." Even mild language,

cognition, and motor deficits are perceived as severe deficits, greatly reducing quality of life. Initial stroke symptoms include weakness in 88% of stroke survivors, sensory deficits in 50%, and visual, cognitive, or speech impairments in 30–48%. Depression affects almost one-third of patients with stroke and half of those with left frontal lobe stroke. Caregivers are susceptible to anxiety (58%), depression (50%), fear, frustration, impatience, and resentment. Functional impairments following stroke may become the responsibility of these caregivers if other sources of social support are not sufficient, and such caregivers may be forced to leave the workforce temporarily or permanently. These intangible and indirect costs of stroke are among the most devastating to the family and the most difficult to measure. Thus, stroke substantially diminishes caregiver quality of life, and only recently have quality of life assessments begun to reveal the true impact of cerebrovascular disease.

II. OVERVIEW OF STROKE ETIOLOGIES

A. Cardiogenic Stroke

Approximately 20% of ischemic strokes are due to cardiogenic cerebral embolism. When other cardiogenic causes of stroke are included, almost one-third of all strokes relate to some form of cardiac disease. The cardiac conditions most likely to produce cardioembolic stroke are atrial fibrillation, endocarditis, diseased or prosthetic cardiac valves, rheumatic heart disease, dilated cardiomyopathy, and acute myocardial infarction. In addition, cardiac surgery is a cause of cerebral embolism.

Although embolism is the most important cardiogenic stroke mechanism, cerebral ischemia and ''watershed'' infarction may result from impaired cardiac output and low blood pressure. The longer systemic circulation is disrupted, the less oxygen is available to maintain cerebral viability. Consequently, global brain injury and damage are frequent sequelae of cardiac arrest. Cardiac arrest produces the most devastating form of stroke, hypoxic–ischemic encephalopathy with or without coma and brain death. There have been major advances in understanding the process of global ischemia and in predicting the outcome from hypoxic–ischemic coma. As a result of large studies of patients with this condition, the likelihood of meaningful recovery at 1 year can be predicted with substantial accuracy based on the presence or absence of key signs on neurological examination at specific time points following the insult. The ability to provide such prognostic information to families has been a major advance in stroke management because physicians are now better informed during discussions regarding organ donation and end-of-life decision making.

Atrial fibrillation affects 1% of the U.S. population, and the risk increases substantially with age. By age 65, 6% of individuals have atrial fibrillation, and 10% of individuals aged 75 or older are affected. In the absence of valvular disease, atrial fibrillation carries a stroke risk of 5% per year. In patients with concurrent hypertension, diabetes, coronary artery disease, congestive heart failure, or prior stroke, the risk increases to 6% per year. The identification of spontaneous echo contrast (''smoke'') during transesophageal echocardiography indicates a 14% annual risk of stroke, in contrast to a 4% risk when smoke is not evident. Treatment with warfarin decreases the risk of ischemic stroke by 68%. Patients younger than 60 years of age without concurrent stroke risk factors may not require warfarin for their ''lone'' atrial fibrillation. Instead, such individuals are often treated with antiplatelet therapy.

Infective endocarditis produces cerebral embolization and damages cardiac valve leaflets, which may become caked with fibrous tissue and bacteria (vegetations). These vegetations embolize and produce stroke in 15–20% of patients with subacute bacterial endocarditis, the most common form of infective endocarditis. It is now apparent that anticoagulation is contraindicated in subacute bacterial endocarditis. Vegetations may be identified with transesophageal echocardiography, a significant advance in the management of this disorder. Nonbacterial thrombotic endocarditis occurs in association with cancer and other chronic diseases. Platelet and fibrin deposits on the mitral and aortic valves embolize to the brain in a high percentage of cases.

Prosthetic heart valves may become damaged, infected, leak, or dislodge. The risk of cerebral embolization is 3 or 4% per year for patients on anticoagulant therapy with mechanical mitral valves, and 1 or 2% for mechanical aortic valves. The risk of cardioembolic stroke is 0.2–2.9% for patients with bioprosthetic heart valves (e.g., porcine valves).

These risks are small in comparison to the stroke risks associated with severely diseased cardiac valves. Rheumatic heart disease affects the valves and leads to cardiogenic embolism in approximately 20% of cases. In the presence of mitral stenosis, 15–17% of patients

have left atrial thrombus. When combined with atrial fibrillation, the risk of cerebral embolization increases by 17-fold. The severity and duration of mitral stenosis also increase the risk of embolization. Calcific aortic stenosis is found in 1% of patients with transient ischemic attacks or stroke. The emboli are usually small and often lead to transient monocular blindness or retinal ischemia. Mitral annular calcification and mitral valve prolapse have not been definitively linked to cardioembolic stroke risk.

Dilated cardiomyopathy carries a risk of embolization as high as 3.5%. Many causes have been proposed, including infection (e.g., viruses and Chagas disease), immunologic disorders, toxic agents (alcohol and chemotherapeutic agents such as Adriamycin), poor nutrition, pregnancy, and genetic factors. Chronic atrial fibrillation develops in 20–30% of patients. Ventricular thrombi were found in 78% of cases in one autopsy series of 152 patients. Transthoracic echocardiography detects thrombi in 11–58% of cases, and transesophageal echocardiography is even more sensitive.

Acute myocardial infarction and stroke commonly occur together. Left ventricular thrombus, a potent source of cerebral emboli, is evident in 30–35% of cases of acute myocardial infarction. Hyokinesia or akinesia, with injury to the endocardial surface, is more likely to occur with transmural, anterior wall infarcts. Thus, anterior myocardial infarction is most commonly related to stroke. Two or 3% of patients with acute myocardial infarction will have stroke within 1 month, and 10–15% develop stroke within 2 years of the myocardial infarction. The majority of strokes occur within the first 10 days.

Patent foramen ovale is thought to relate to paradoxical cerebral embolization by providing a mechanism for systemic emboli to bypass the lungs and proceed directly to the brain. In patients with stroke from unknown cause, 40% have patent foramen ovale, compared to 10% of controls. When no other cause of stroke can be identified, the occurrence of patent foramen ovale is 54%, suggesting that patent foramen ovale is an independent risk factor for stroke.

B. Large Extracranial Artery Diseases and Stroke

Aortic arch atheromatous plaque thickness directly relates to the risk of stroke. The recurrence rate for stroke is 11.9 per 100 person years when aortic wall thickness reaches 4 mm, in comparison to 3.5 per 100 person years in patients with aortic wall thickness less than 4 mm. Transesophageal echocardiography has substantially improved the detection of aortic arch disease, which had previously been very difficult to assess without angiography.

Extracranial carotid artery disease is an important cause of stroke. Carotid stenosis becomes more likely with advancing age and occurs in approximately 2–7% of individuals aged 50 or older. However, carotid stenosis of over 80% affects only 1% of the general population. Carotid artery dissection and giant cell arteritis may also lead to extracranial stenosis and stroke. Arterial dissection occurs most frequently in the setting of minor or major trauma but may occur spontaneously. The vessel wall develops a hemorrhage, which expands and narrows the lumen. The smooth lining within the vessel (endothelium) often becomes irregular and serves as a nidus for clot formation. The hemorrhage may grow sufficiently to cause stenosis and occlusion of blood flow through the damaged artery, or the clots that form at the site of the endothelial damage may embolize to the brain. In giant cell arteritis, inflammation within the wall of extracranial vessels, most often the carotid arteries, leads to ischemia and stroke. Permanent blindness results in 40–50% of cases because the blood supply to the eye is compromised, and strokes occur in approximately 10% of cases. Headache and jaw pain induced by chewing (jaw claudication) are important clues to the possibility of giant cell arteritis. An elevated erythrocyte sedimentation rate is typically found, but confirmation of the diagnosis requires temporal artery biopsy. The disorder increases in incidence with age, affecting 2.6 per 100,000 persons 50–59 years of age each year and 44.6 per 100,000 over 80 years of age each year. Whites are seven times more likely to be affected than African Americans, and women are three times more likely to have the disease than men.

The vertebral arteries may be affected by both dissection and giant cell arteritis. In addition, vertebrobasilar insufficiency may result from subclavian steal syndrome. This disorder most often occurs in the setting of severe stenosis of the subclavian artery, proximal to the origin of the left vertebral artery, due to atherosclerotic plaque. Blood may be "stolen" from the left vertebral artery, which supplies blood to the left arm rather than to the brain. Transient ischemic attacks and episodic loss of consciousness (syncope) are frequent symptoms of the disorder; large differences in blood pressures are typically encountered when the right and left arms are compared. Surgical

correction of the stenosis may prevent stroke and cure the syncope.

Fibromuscular dysplasia is an uncommon disorder affecting young women more often than men. A muscular band constricts large extracranial arteries, leading to ischemia or infarction. Several areas of narrowing may occur within a single vessel, leading to an appearance resembling a string of beads on catheter angiography. The renal arteries are often involved, and renovascular hypertension may result.

C. Intracranial Arterial Disease and Stroke

Intracranial arterial disease may lead to ischemic and hemorrhagic strokes. Aneurysms and arteriovenous malformations (AVMs), intracranial atherosclerosis, and vasculitis are some of the major diseases of intracranial arteries producing stroke. Approximately 80% of nontraumatic subarachnoid hemorrhage is due to ruptured intracranial aneurysm. Arteriovenous malformations carry a 40–50% lifetime risk of rupture with subsequent fatal or disabling hemorrhagic stroke. Central nervous system vasculitis may occur with or without coexisting systemic vasculitis. Fibromuscular dysplasia rarely affects intracranial arteries. Cerebral amyloid angiopathy may lead to both ischemic and hemorrhagic stroke. Migraine may cause sufficient constriction of intracranial arteries to produce ischemia and infarction. Vasospasm (constriction of vessels due to contraction of vascular smooth muscle) is common following subarachnoid hemorrhage. Toxic substances such as cocaine and other stimulants may lead to arterial spasm and vasculitis, with subsequent stroke.

D. Extracranial Venous Disease and Stroke

There are two situations in which extracranial venous disease may lead to stroke. Internal jugular vein occlusion may prevent drainage of the brain. Venous thromboembolism may cause paradoxical cerebral embolization. These pass through the right ventricle of the heart in most patients to produce pulmonary emboli. In patients with patent foramen ovale or other intracardiac shunt, venous blood and clots may pass from the right side of the heart to the left, where cerebral embolization is possible. This right-to-left shunting is enhanced by increasing intrathoracic pressures (Valsalva maneuver) or right heart pressures (e.g., pulmonary hypertension and right-sided heart failure). Hypercoagulable states may predispose to thrombotic disorders of the extracranial venous system.

E. Intracranial Venous Disease and Stroke

Within the brain, there are several large veins that drain into the venous sinuses. Sagittal sinus thrombosis is the most serious disorder of the intracranial venous pathways. This condition frequently leads to headache (80%), swelling of the optic nerve (papilledema, 50%), motor or sensory deficits (35%), seizures (29%), hemorrhage (50%), and death (5–10%). Contrast enhanced CT scanning allowed superior sagittal sinus thrombosis to be diagnosed by showing the "empty delta" sign. Now, MR is the preferred test, showing increased signal intensity on T1, T2, and proton density weighted images. The clot can be directly visualized using such routine MR techniques, and MRV can easily demonstrate the absence of venous blood flow caused by the thrombus. Although angiography will demonstrate the problem during the venous phase of contrast washout, MRI techniques have reduced the need for angiography for the diagnosis of sagittal sinus thrombosis.

F. Respiratory Disease and Stroke

Respiratory disturbances may cause stroke if the oxygen delivery to the brain is sufficiently impaired. For example, sleep apnea is associated with a tendency to stroke and silent stroke as well as vascular dementia. Occlusion of the airway leads to cessation of air movement (apnea), causing oxygen levels to fall (desaturation) and carbon dioxide levels to increase (hypercarbia). Oxygen desaturation may be so severe that cardiac rhythm disturbances occur. Hypercarbia stimulates respiratory drive and increases CBF, and these responses to the apnea may maintain oxygen delivery within acceptable limits. The recognition that this common sleep disorder increases the risk of stroke and silent stroke has led to improved surveillance and treatment.

Many other disorders may affect oxygen delivery and contribute to stroke. Chronic obstructive pulmonary disease, myasthenia gravis, motor neuron disease, myopathy, myositis, botulism, neuromuscular blockade, neuropathy, brain stem stroke, cervical

spine trauma, seizures, and many other conditions can affect respiratory drive or effectiveness. Even when respiratory mechanisms are intact and adequate oxygen delivery is achieved, gas exchange within the lungs may be impaired so that hypoxia and hypercarbia result. For example, emphysema, pneumonia, pulmonary embolism, and other pulmonary disorders may impair the ability for oxygen to pass from the inspired air into the bloodstream. Carbon monoxide binds to hemoglobin, thus preventing oxygen binding. Hydrogen sulfide and hydrogen cyanide are poisons that can block oxygen delivery and metabolism. Our understanding and awareness of cerebrovascular insufficiency on the basis of respiratory pathology has improved in conjunction with major advances in the fields of pulmonary medicine and critical care neurology.

G. Microvascular Disease of the Brain

Age, hypertension, diabetes, and smoking all cause damage within the microvasculature. These damaged vessels may leak protein that deposits in the walls, producing stiffness and hardening. Microvascular damage makes it difficult for the vessels to dilate and constrict. When flow through these small vessels is sufficiently diminished, a small (lacunar) stroke or silent stroke may result. Risk factor modification is the key to preventing microvascular disease and subsequent stroke. Microvascular disease of the brain affects half of all people aged 70 or older, according to large MR studies that have been done. These areas of MR signal abnormality may be due to demyelination or ischemic damage, and the demyelinating lesions are thought to have an ischemic basis as well (i.e., ischemic demyelination). Microvascular disease of the brain also relates in part to impaired blood flow during periods of relative hypotension, which leads to ischemia in brain areas with disturbed autoregulation. Stroke leads to a breakdown in cerebral autoregulation within the ischemic zone, and vascular damage from hypertension, amyloid angiopathy, atherosclerosis, and other disorders may compromise cerebral autoregulation as well. Research is under way to determine the ischemic threshold in a clinically useful way so that hypotensive ischemia within compromised vascular territories might be avoided. Currently, there is no way to know when blood pressure has been lowered sufficiently to compromise susceptible vascular territories.

H. Hematologic Disorders and Stroke

Disorders of the blood and blood products are a relatively common cause of stroke. Many hereditary conditions lead to a predisposition to the development of thrombosis, including Leiden factor V mutation, protein C deficiency, protein S deficiency, and antithrombin III deficiency. Anticardiolipin antibodies and antiphospholipid antibodies may increase stroke risk as well. Sickle cell disease, SC disease, and polycythemia rubra vera also predispose to coagulation and stroke. An overabundance of platelets (thrombocytosis) may also produce a hypercoagulable state and increase stroke risk. When there are clotting factor deficiencies, such as in hemophilia, there is an increased risk of hemorrhagic stroke. Thrombocytopenia, hepatic failure, and disseminated intravascular coagulation may also lead to brain hemorrhage.

I. Other Causes of Stroke

There are many other causes of stroke, most of which are rare. Inflammatory disorders can cause thickening of the vessel walls, with obstruction of blood flow, hemorrhage, and multifocal infarction. There are several forms of angiitis or vasculitis (Table II), including Moyamoya disease, Wegener's granulomatosis, and polyarteritis nodosa. Some forms of vascular disease are notoriously difficult to diagnose, and the vascular pathology is poorly understood (e.g., the vasculopathy of systemic lupus erythematosus). Traumatic causes of stroke include air, fat, and marrow embolism. Increased intracranial pressure resulting from trauma can lead to reduced cerebral perfusion pressure with watershed ischemia and infarction. Infections may cause stroke due to associated inflammatory changes in the vasculature (e.g., herpes zoster). Developmental vascular anomalies, such as cavernous, venous, or capillary malformations, may lead to stroke. Cardiac abnormalities such as atrial septal aneurysm can predispose to cardiogenic cerebral embolism. Infection or thrombus formation within the cavernous sinus obstructs venous flow leaving the brain and may cause stroke. Neoplasm is commonly encountered in association with nonbacterial thrombotic endocarditis, a notoriously devastating cause of cardiogenic embolism. Atrial myxomas are tumors within the atrium that may produce tumor emboli. Cryoglobulinemia is one of the many causes of hypercoagulable state that may produce stroke. The

list of stroke etiologies is extensive, and Table II is therefore incomplete.

III. DIAGNOSTIC TECHNIQUES

A. Clinical Assessment

The evaluation of stroke has recently become a medical emergency due to the availability of effective acute therapy. Stroke patients who are treated by neurologists fare better than those treated by nonspecialists and have shorter lengths of stay, lower hospitalization costs, improved functional outcome, and improved levels of satisfaction. The neurological history and examination are essential aspects of the acute evaluation. Abnormalities of relevance may be identified literally from head (e.g., papillary asymmetry) to toe (e.g., Babinski sign). Neuroimaging, neurosonology, cardiac evaluation, and hematologic testing are all important aspects of stroke management. A team approach is ideal in such situations.

B. History and Physical Examination

Information may be obtained from the patient, family members, or others during the history. A thorough history includes questions about each stroke risk factor, such as smoking, alcohol use, prior stroke, prior myocardial infarction, and family history of cardiac or cerebrovascular disease. The symptoms being experienced are defined in terms of onset, duration, exacerbating and relieving factors, fluctuations in intensity, associated or antecedent symptoms, and progression. The time of symptom onset is explicitly defined to clarify eligibility for thrombolytic therapy. Patients who awaken after prolonged sleep with stroke symptoms or who are found in an incapacitated state after an unknown length of illness are questionable candidates for thrombolytic therapy, which must be administered within 3 hr of stroke onset. A history of gastrointestinal hemorrhage or recent surgery is especially important since these may preclude the use of thrombolytic therapy. Measurement of blood pressure, respiratory rate, and heart rate (vital signs) is the first step in the examination. The carotid arteries, subclavian arteries, and heart are auscultated. The stool is tested for occult blood. The nail beds, conjunctivae, integument, and extremities are examined. Cranial nerves, motor power and tone, muscle stretch reflexes, coordination, tactile sensation, postural stability, and gait are all carefully assessed. Mental status is assessed, and detailed cognitive testing may sometimes be required.

The fields of neuropsychology, behavioral neurology, neuropsychiatry, and cognitive neurology have been strengthened through the study of patients with cerebrovascular diseases. The observation that specific lesions within the brain are commonly associated with specific patterns of impairment has led to clinically relevant and feasible "neurobehavioral" tests that may be used at the bedside. A skilled neurologist can identify cognitive or behavioral disturbances that provide important information about the site(s) of damage within the brain, and its severity, using such bedside neurobehavioral testing. Sometimes neurobehavioral alterations are the only evidence of stroke apparent on detailed examination. For example, damage within the left temporal lobe may produce severe language comprehension abnormality and rambling, nonsensical speech (Wernicke's aphasia) without other stroke symptoms, such as paralysis, numbness, ataxia, or visual loss. A thorough neurological examination, augmented when appropriate by neurobehavioral testing, is an important aspect of stroke management.

C. Neuroimaging

1. Computerized Tomography

CT is obtained urgently to exclude intracerebral hemorrhage and to identify whether imaging evidence of acute infarction exceeds one-third of the middle cerebral artery territory. In the presence of hemorrhage or large stroke, thrombolytic therapy is contraindicated. A "hyperdense MCA sign" (Fig. 2) may indicate thrombus within the middle cerebral artery and correlates with a worse outcome. Early changes of stroke may not be evident on CT for several hours after the stroke. Sudden deterioration may suggest hemorrhagic transformation of a stroke, a repeated episode of bleeding from subarachnoid hemorrhage, expansion of the size of ischemic infarction, repeated embolic infarction, deterioration due to brain swelling, or many other possibilities. Because CT is widely available and can be obtained very rapidly, it is frequently used for reassessment after such changes. In uncooperative patients, in patients that require ventilatory assistance (or other hardware that may not be exposed to the strong magnetic field of an MR machine), and in emergency situations, CT is still the test of choice.

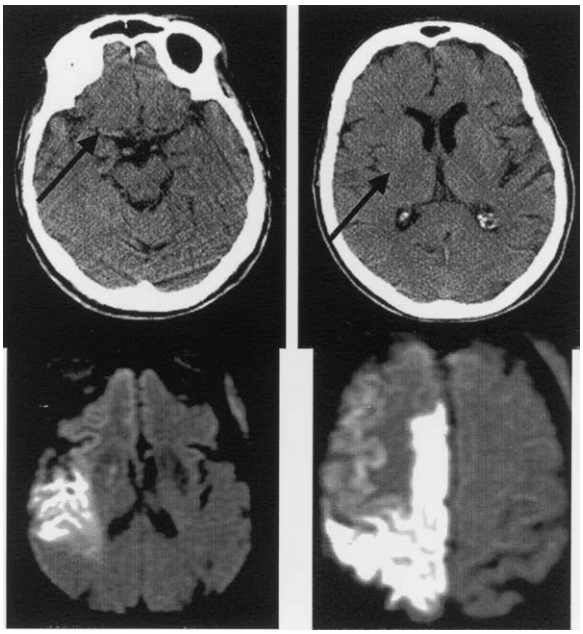


Figure 2 Acute ischemic stroke, CT, and diffusion weighted MR. The top images are slices from a CT scan obtained within hours after an acute middle cerebral artery (MCA) stroke. The "dense MCA sign" can be seen (top left, arrow), but little evidence of stroke is noted (top right, arrow). In contrast, the acute diffusion weighted MRI shows substantial hyperintensity in the ischemic area [adapted with permission from R. Bakshi and L. Ketonen, Brain MRI in clinical neurology. In *Baker's Clinical Neurology* (R. J. Joynt and R. C. Griggs, Eds.). Copyright Lippincott, Williams & Wilkins, 2001].

Three advances in CT have increased its clinical utility. First, contrast agents provide important information about the integrity of the blood–brain barrier. In malignancy, infection, inflammatory disorders, and subacute stroke, contrast material will often escape from the vessels, producing enhancement. Contrast enhancement begins about 3 days after a stroke and may persist for weeks. Second, CT angiography is now available. Vascular contrast agents attenuate the CT signal so that high-quality images of cerebrovascular anatomy can be derived using this technique. Third, contrast materials may be combined with CT to produce quantitatively accurate information about the volume of blood within a specific brain area. The inflow (time to peak) or washout (clearance) of contrast material (e.g., Xe or iodinated contrast agents) provide quantitative information about CBF through a brain area or vessel. These techniques are becoming more widely available and are increasingly important for acute stroke management.

2. Magnetic Resonance

The use of MR to produce brain images has revolutionized the field of cerebrovascular disorders. The underlying physical principles of MR scanners enable the measurement of a large variety of phenomena, depending on the imaging technique that is utilized. Many new MR techniques have been developed and these are summarized in Table III. MRI can be useful prior to stroke (e.g., to detect silent stroke or to reveal

Table III
Neuroimaging Techniques[a]

Neurosonology
Transcranial Doppler
Volume flow imaging
B-mode
M-mode
Computed tomography
Without contrast
With contrast
CT angiography
CT perfusion (e.g., Xe)
Magnetic resonance
T1 weighted
T2 weighted
Proton density
FLAIR
MRA
MRV
DWI
PWI
Spectroscopy
Others (magnetization transfer, gradient echo, ADC, etc.)
SPECT
HMPAO
^{133}Xe
PET
^{15}O CBF
[^{18}F]Fluorodeoxyglucose

[a]Note that some neuroimaging techniques (e.g., PET) are not commonly used for clinical purposes. There are several new neurosonology and magnetic resonance imaging techniques (see text for details) not listed in the table. CT, computed tomography; MR, magnetic resonance; MRA, MR angiography; MRV, MR venography; FLAIR, fluid attenuation inversion recovery; SPECT, single photon emission computed tomography; CBF, cerebral blood flow; HMPAO, [^{99m}Tc]*d*,1-hexamethylpropylene amine oxime; DWI, diffusion weighted imaging; PWI, perfusion weighted imaging; and ADC, apparent diffusion coefficient.

evidence of arterial narrowing predisposing to later stroke), during the acute stages of stroke (e.g., to identify viable brain tissue at risk for infarction that may be salvaged by intraarterial or intravenous thrombolysis, as in Fig. 3), and following a cerebrovascular diagnosis (e.g., to clarify whether a hemorrhagic area was really due to stroke rather than hemorrhage into a rapidly growing brain tumor). These benefits are achieved without undue risks; the presence of a cardiac pacemaker, brain aneurysm clip, or the presence of other metallic foreign bodies (e.g., metal in the eye) are the only contraindications to brain MR. MR is useful in the imaging of both ischemic and hemorrhagic processes (Figs. 3–5).

The earliest sign of ischemic infarction apparent on MR is the loss of the normal intravascular "flow void" due to slow flow or occlusion within an intracranial vessel. This finding is similar to the "dense MCA sign" described in CT images of acute stroke. Arterial hyperintensity may be apparent on fluid attenuation inversion recovery (FLAIR) images almost immediately after stroke. Venous hyperintensity on FLAIR imaging suggests venous thrombosis, which may be confirmed with MRV (Fig. 6). Two to 4 hr following acute ischemic infarction, subtle changes in the shape of cortical gyri may become evident due to cytotoxic edema, which causes neural tissue to swell. Diffusion weighted MR reveals abnormalities within 30 min of acute ischemic infarction and is rapidly becoming a valuable technique for acute stroke imaging (Figs. 2 and 3). The majority of acute ischemic infarcts can be identified using diffusion MR, which reportedly detects 97% while remaining normal in 100% of individuals lacking acute ischemic infarction. Signal hyperintensity on T2 and proton density images becomes apparent within 8 hr of an ischemic infarction as a result of vasogenic edema, the process whereby water extravasates from damaged vessels into the

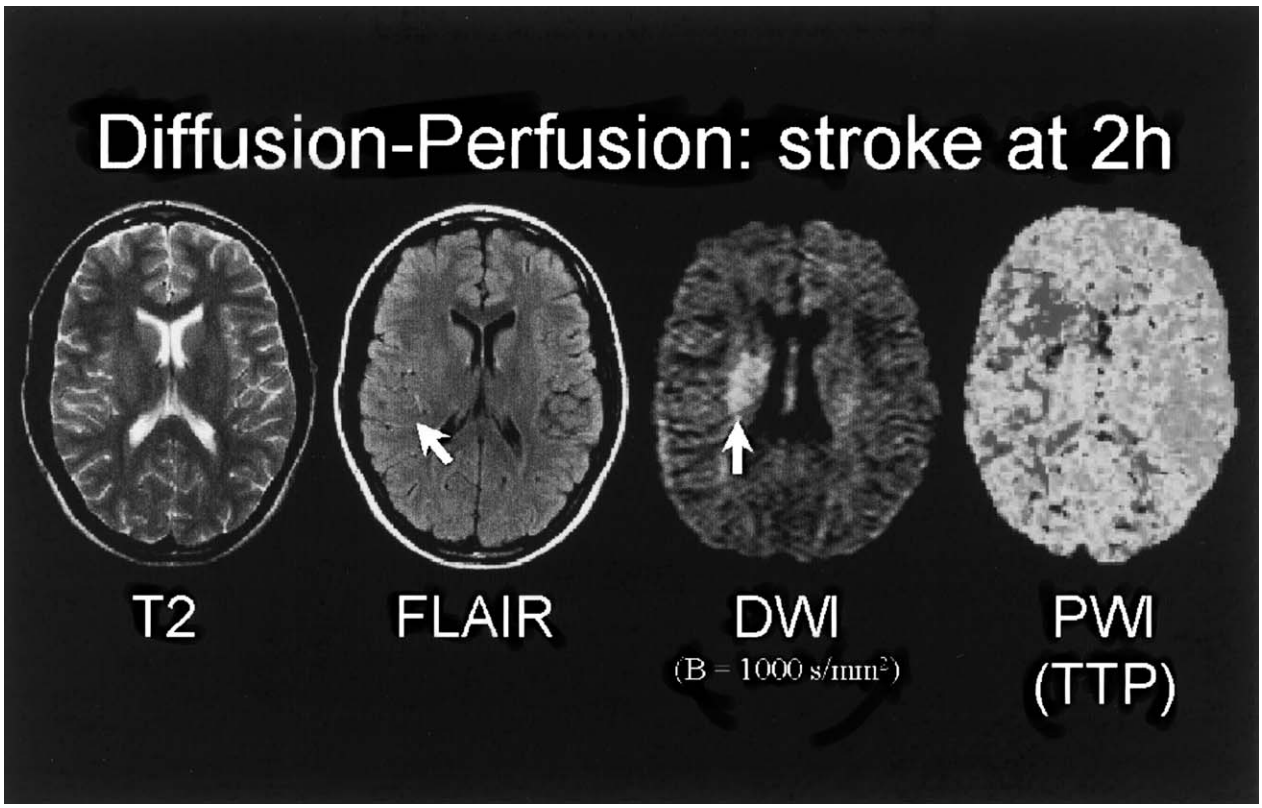


Figure 3 Acute ischemia; MR techniques compared. Two hours after stroke, the T2-weighted MR (left) may be completely normal. Fluid attenuation inversion recovery (FLAIR) imaging may reveal hyperintense vessels in the region of the infarction (arrow). The diffusion weighted image (DWI) reveals an area of infarction (as shown in Fig. 2). Perfusion weighted imaging (PWI) shows a large area of diminished perfusion that has not yet become infarcted, as shown by the discrepancy between the DWI and the PWI. This "mismatch" corresponds to a large area of brain tissue that is potentially at risk for infarction [adapted with permission from R. Bakshi and L. Ketonen, Brain MRI in clinical neurology. In *Baker's Clinical Neurology* (R. J. Joynt and R. C. Griggs, Eds.). Copyright Lippincott, Williams & Wilkins, 2001].

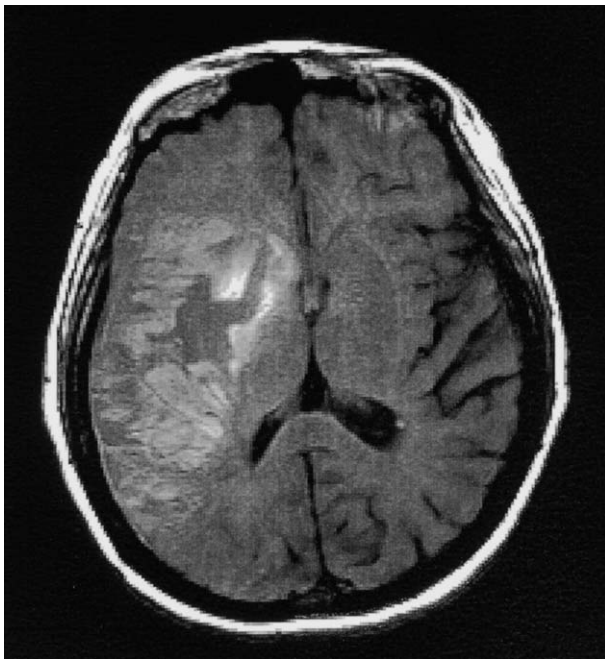


Figure 4 Acute hemorrhage seen on MR. This is a T1-weighted MR, without contrast, demonstrating hemorrhagic transformation within a large middle cerebral artery stroke within 7 days following the stroke [adapted with permission from R. Bakshi and L. Ketonen, Brain MRI in clinical neurology. In *Baker's Clinical Neurology* (R. J. Joynt and R. C. Griggs, Eds.). Copyright Lippincott, Williams & Wilkins, 2001].

surrounding tissues. By 16–24 hr following ischemic infarction, there is usually enough vasogenic edema to cause T1 signal hypointensity. Within 5–7 days after ischemic infarction, there is parenchymal enhancement following injection of gadopentate dimeglumine. The new blood vessels that form within damaged areas surrounding a complete infarction have an imperfect blood–brain barrier, allowing contrast to leak into surrounding tissues.

Hemorrhages may be evident on MRI, but the time course and evolution are much more complex than with ischemic infarction. For example, subarachnoid or intraventricular hemorrhages evolve differently than subdural or epidural hematomas. Intraparenchymal hematomas are different as well. FLAIR images enable detection of oxygenated hemoglobin in acute subarachnoid hemorrhage, and the superiority of CT scanning for the purpose of acute detection of subarachnoid hemorrhage has thus been challenged. Acute intraparenchymal hemorrhages may be missed by T1, T2, and proton density images; these techniques do not show alteration in signal intensity (i.e., decreases) for 1–3 days due to the gradual accumulation of deoxygenated hemoglobin. As red blood cells degrade, deoxygenated hemoglobin is converted to met-hemoglobin, which may appear hyperintense on T1 images 3–7 days after a hemorrhage. There is gradual accumulation of hemosiderin as the breakdown of hemoglobin continues from 1 to 4 weeks. Met-hemoglobin is typically hyperintense on T1, T2, and proton density images, whereas hemosiderin (which accumulates after 4 weeks) produces low signal intensities on T1, T2, and proton density images. Figs. 7 and 8 illustrate the typical progression of brain hemorrhage seen with MR techniques.

MR techniques are important for visualizing the arteries supplying the brain. MRA allows the visualization of intracranial and extracranial arteries of medium and large caliber. Turbulent flow will produce flow signal dropout, which may lead to false interpretations of vessel occlusion. Even relatively large aneurysms may be missed with MRA, and the technique has not yet replaced catheter angiography as the gold standard for defining cerebrovascular anatomy. MRV is sensitive to blood flowing at slower velocities, such as the blood within cerebral veins and sinuses. MRV is thus an important tool for detecting sagittal sinus thrombosis, cortical vein thrombosis, and other disorders of the cerebral veins.

Perfusion imaging (T2*) allows qualitative evaluation of CBF. Areas of relatively reduced CBF can be identified, and a "diffusion–perfusion mismatch" may indicate a large area of brain in danger of complete infarction (Fig. 3). Changes in CBF are being utilized to map the parts of the brain that participate in various kinds of neurological or cognitive functions. Functional MRI (fMRI) may soon become an essential part of the management of cerebrovascular disorders. Not only will it enable detection of diffusion–perfusion mismatch in acute stroke patients eligible for intravenous or intraarterial thrombolysis but also fMRI may facilitate appropriate selection of arteries that might safely be sacrificed in the management of arteriovenous malformation. Magnetic resonance spectroscopy (MRS) allows the breakdown products of tissue destruction to be measured quantitatively. Thus, MRS may become an important method of detecting the presence of injury or measuring the effect of neuroprotective interventions. Its use in clinical practice is currently limited.

3. Neurosonology

The field of neurosonology has emerged as a key aspect of cerebrovascular disorders. Several distinct

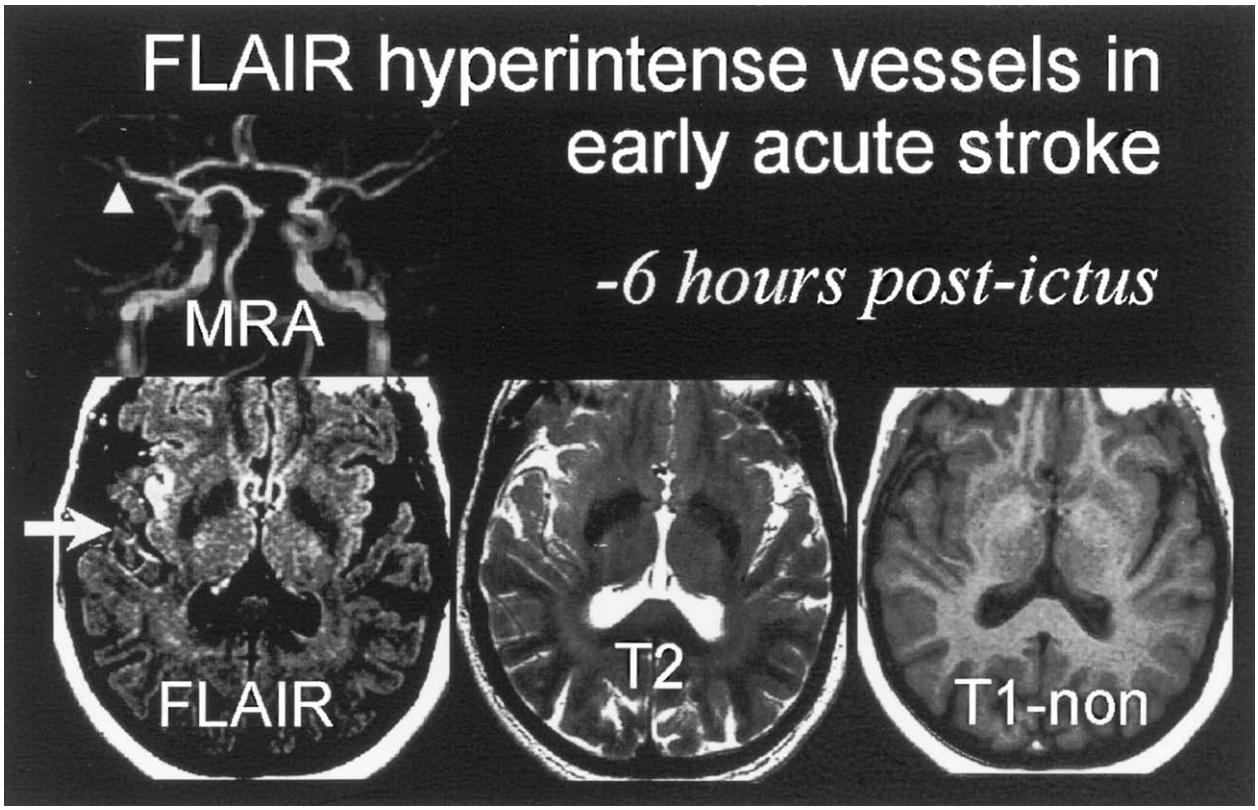


Figure 5 Hyperintense vessel on FLAIR and MRA in acute stroke. The occluded middle cerebral artery is seen on MRA (arrow head, top left) 6 hr following stroke. Hyperintense vessels are noted on FLAIR imaging (arrow, bottom left). The T2 and T1 noncontrast images are normal [adapted with permission from R. Bakshi and L. Ketonen, Brain MRI in clinical neurology. In *Baker's Clinical Neurology* (R. J. Joynt and R. C. Griggs, Eds.). Copyright Lippincott, Williams & Wilkins, 2001].

techniques have been developed to image vascular anatomy and stenosis due to plague, to measure flow velocity by Doppler shift analysis, and to quantitate flow. Neurosonology techniques are noninvasive, safe, reliable, economical, and informative.

a. B-Mode Imaging of the Carotid and Vertebral Arteries Cross sections and longitudinal sections depicting arterial anatomy are readily obtained using B-mode (brightness-mode) ultrasound. Such images depict the thickness of the vessel wall (which correlates with stroke risk in the carotid circulation and in the aorta) and the presence and thickness of plaque. Intraplaque hemorrhages (which increase stroke risk) can be identified with an accuracy of 90%, sensitivity of 96%, and specificity of 88%. Ulceration is difficult to identify with accuracy. Homogeneity can be assessed as well. Heterogeneous plaques are associated with increased stroke risk because they are more likely than homogenous plaques to contain hemorrhage. The characterization of plaque heterogeneity and typology is ongoing, and further research will be required to determine the clinical significance of findings such as ulceration.

b. Doppler Ultrasound of the Carotid and Vertebral Arteries Ultrasound echoes returning from a target, such as blood flowing through a vessel, experience a change in their frequency in direct proportion to blood flow velocity. These Doppler shifts are analyzed using fast Fourier transformation to generate a blood flow velocity spectrum. Mean flow velocity, peak systolic velocity, end diastolic velocity, and a variety of other important indices may be derived. Faster flow velocities generally indicate greater stenosis, unless critical stenosis is present. These principles are used to grade the degree of carotid artery stenosis as A (normal), B (1–15% stenosis), C (16–49% stenosis),

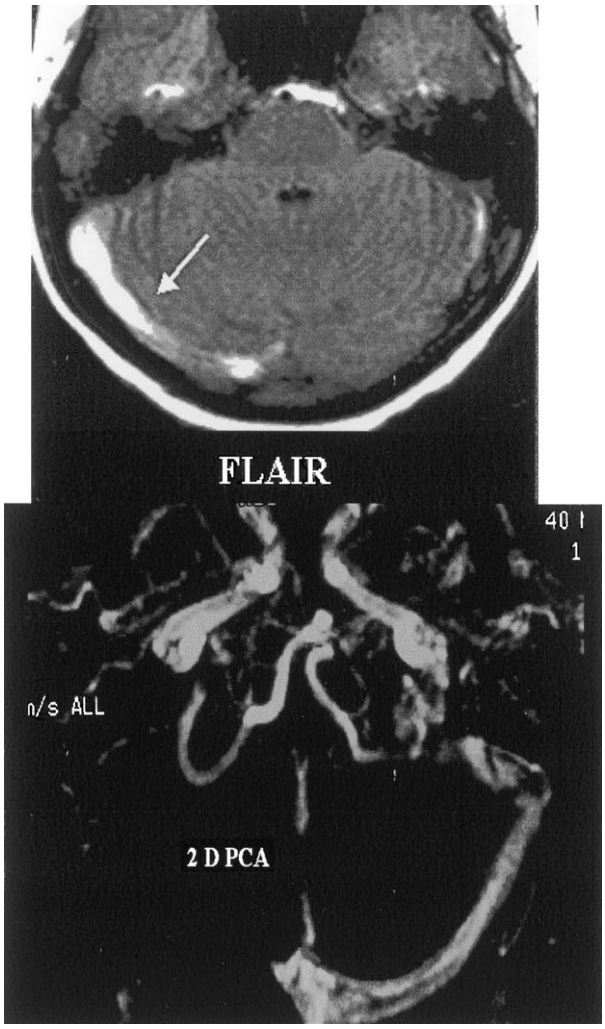


Figure 6 MRV and FLAIR imaging shows transverse sinus occlusion. The top image (arrow) reveals an area of hyperintensity on FLAIR imaging. The hyperintensity corresponds to a thrombus within the right transverse sinus. The bottom image is a magnetic resonance venogram showing lack of venous blood flow through the occluded transverse sinus, manifested as an absence of normal (bright) flow signal. The opposite transverse sinus (shown on the right) has the normal appearance of bright signal, indicating venous blood flow (graphic provided by Rohit Bakshi, M.D.).

D (50–79% stenosis), D+ (80–99% stenosis), or occlusion. Because the degree of stenosis closely relates to stroke risk, noninvasive vascular testing with Doppler ultrasound is extremely valuable in the management of cerebrovascular disease. The combined use of B-mode and Doppler spectrum analysis is called duplex sonography and is currently the preferred tool for extracranial vascular screening in patients with stroke. Testing of the vertebral arteries is limited by the vertebrae but may still provide important diagnostic information (e.g., reversal of flow in the left vertebral artery suggests the subclavian steal syndrome). The direction of flow is easily determined using Doppler ultrasound.

c. Transcranial Doppler Ultrasound of Intracranial Arteries TCD provides a noninvasive method for determining intracranial hemodynamics. The direction of flow within the major intracranial vessels can be determined to rule out significant extracranial stenosis (e.g., subclavian steal) or intracranial disease. Normal ranges for flow velocity have been determined for most large intracranial vessels, and flow velocity changes also may confirm intracranial stenosis. For example, TCD is utilized to detect vasospasm following subarachnoid hemorrhage. TCD can also detect emboli, which appear as high-intensity signals within the Doppler spectrum. Embolus detection is useful for identifying intracardiac shunts, such as patent foramen ovale, and for the detection of microemboli resulting from cardiac surgery. TCD is used for intermittent assessment of vascular patency following thrombolysis, to assess vascular reactivity during provocative tests such as carbon dioxide inhalation or acetazolamide administration, and to continually monitor intracranial flow during surgical procedures. Flow changes have recently been related to cognitive activity.

Vascular reactivity is typically assessed by combining TCD with hyperventilation or carbon dioxide inhalation. The former can reduce middle cerebral artery flow velocity within 15 seconds, and lower it by 35%. In contrast, carbon dioxide inhalation can increase flow velocity by as much as 52.5%. A 1000-mg dose of acetazolamide intravenously can begin to increase flow velocity within 3 min, reaching a peak flow velocity 35% higher than baseline within 10 min. Patients with cerebrovascular risk factors and reduced ability to increase flow velocity in response to carbon dioxide inhalation (i.e., reduced "cerebral perfusion reserve") are at higher risk for stroke than individuals with normal reserve. By lowering blood pressure in a controlled fashion, the ability of the cerebral circulation to maintain CBF during periods of hypotension can be measured using TCD. The blood pressure threshold leading to a sudden and severe decrease in flow velocity is defined as the lower limit of cerebral autoregulation. Future research will determine whether determination of the lower limit of cerebral autoregulation will have clinical applicability in preventing strokes attributable to hypotension.

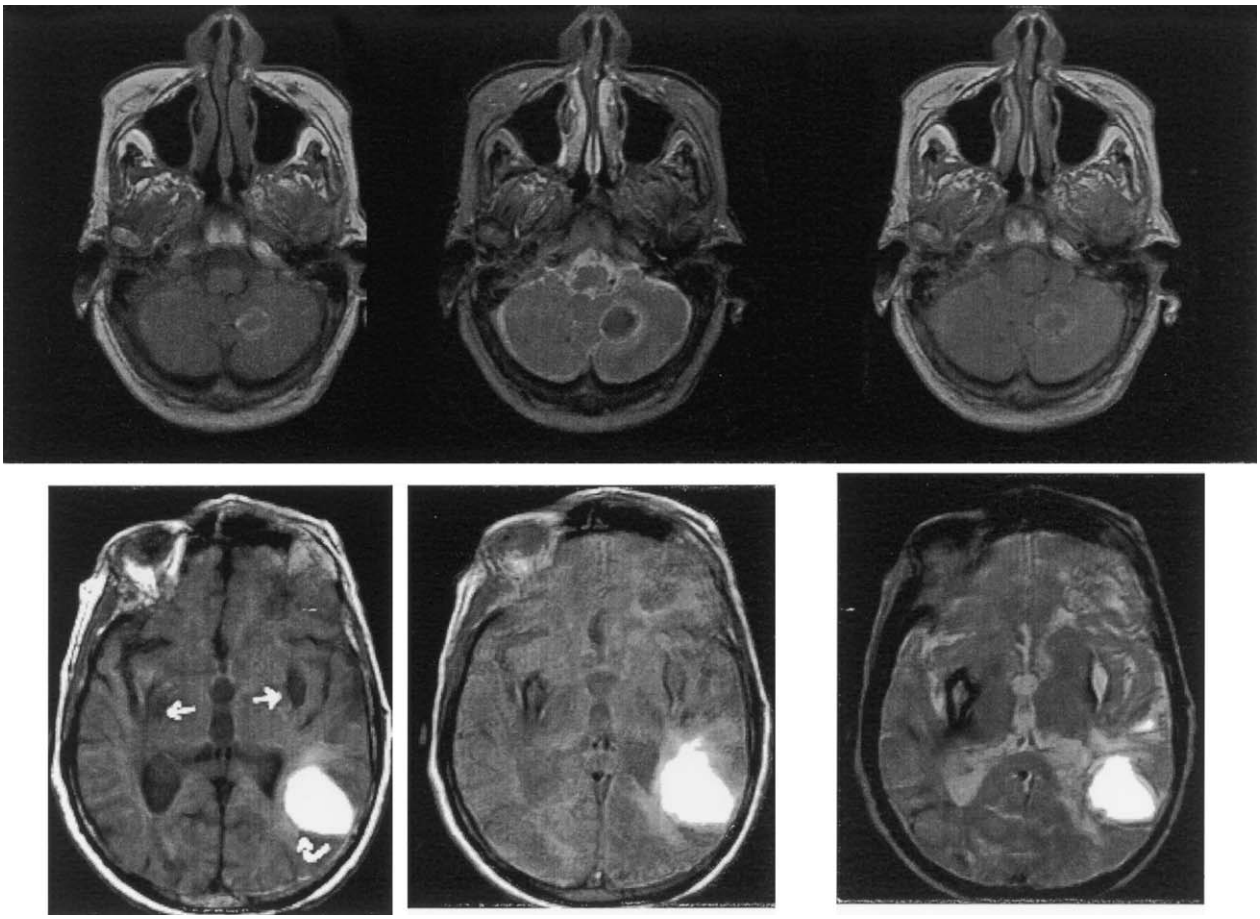


Figure 7 Evolutionary changes of acute and chronic hemorrhage. The top images depict an acute (3 days) cerebellar hemorrhage. Deoxyhemoglobin in the core of the hemorrhage is isointense to slightly hypointense on T1 (left) and proton density (right) images, whereas it is clearly hypointense on T2-weighted imaging (top center). Met-hemoglobin is hyperintense on T1 and proton density imaging but hypointense on T2. Thus, the combination of these techniques differentiates these two hemoglobin degradation forms. The bottom images depict bilateral, chronic (>4 weeks), basal ganglia hemorrhages (straight arrows). A large, late subacute hemorrhage (curved arrow) is shown also. The chronic hemorrhages are hypointense due to a rim of hemosiderin and ferritin. The bright signal within the late subacute hemorrhage (approximately 4 weeks) is due to extracellular met-hemoglobin accumulation. Thus, the age of these hemorrhages can be readily determined using a combination of MR techniques [adapted with permission from R. Bakshi and L. Ketonen, Brain MRI in clinical neurology. In *Baker's Clinical Neurology* (R. J. Joynt and R. C. Griggs, Eds.). Copyright Lippincott, Williams & Wilkins, 2001].

d. New Ultrasound Techniques (Power Doppler, Color Velocity Flow Imaging, Three-Dimensional Doppler Imaging, etc.) A variety of new neurosonology techniques are being developed. Power Doppler measures the intensity of the returning echoes rather than the frequency shifts, and it may be useful for low-flow states. Color velocity flow imaging, or volume flow, computes quantitative blood flow values from mean flow velocity over a fixed period of time through a vessel of known size. Three-dimensional Doppler imaging has been used to noninvasively map aneurysms and other vascular abnormalities. It seems likely that there will be many important clinical applications of these techniques in the future.

D. Nuclear Medicine

1. Positron Emission Tomography

PET can be used to measure brain metabolism, CBF, and the quantity of specific receptor types within the brain. The technique requires administration of a radionuclide that emits positrons during radioactive

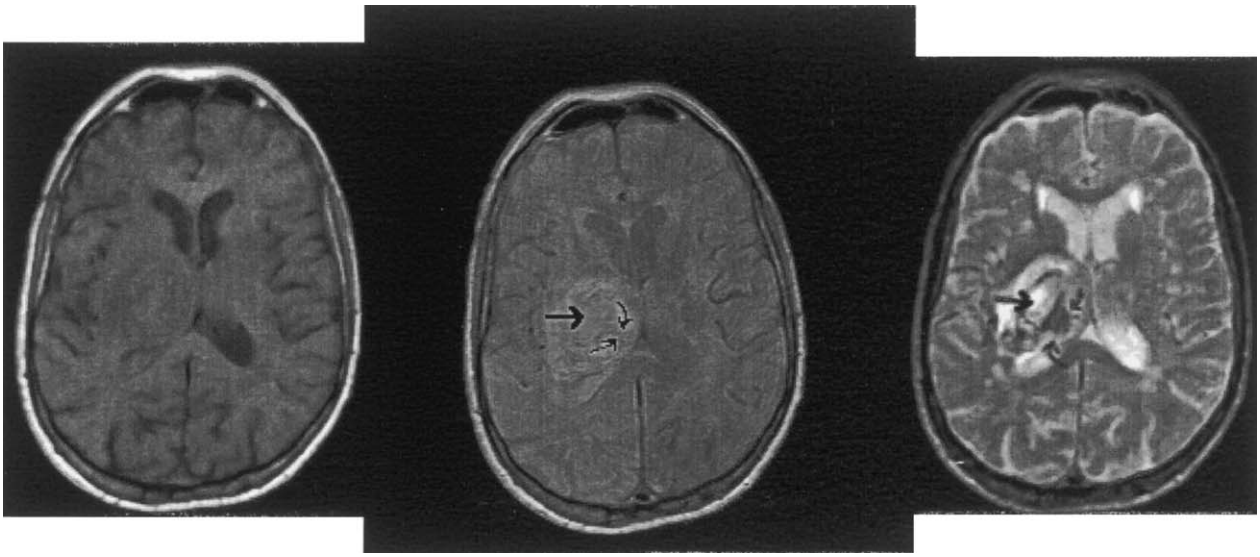


Figure 8 Early acute brain hemorrhage. Oxyhemoglobin and deoxyhemoglobin are mixed in the acute stage of hemorrhage (<24 hr). The T1-weighted image (left) shows only mild hypointensity and mass effect. The proton density image (center) shows hyperintensity in the area of hemorrhage. The T2-weighted image shows hyperintensity (straight arrow) as well as hypointensity (curved arrows), corresponding to oxyhemoglobin and deoxyhemoglobin, respectively. A hyperintense rim of edema is visible around the hematoma on proton density and T2-weighted imaging [adapted with permission from R. Bakshi and L. Ketonen, Brain MRI in clinical neurology. In *Baker's Clinical Neurology* (R. J. Joynt and R. C. Griggs, Eds.). Copyright Lippincott, Williams & Wilkins, 2001].

decay. The scanner detects the gamma rays that are emitted when a positron with an electron and both are annihilated. This technique, annihilation coincidence detection, permits the location and quantity of radioactive decay to be mapped within the brain.

Using ^{15}O tracer PET for imaging, quantitative determination of CBF, blood volume, the rate of oxygen metabolism, and the oxygen extraction fraction can be determined. Brain areas where there is inadequate CBF to satisfy metabolic demands ("misery perfusion"), areas where there is more than enough blood flow to meet metabolic requirements ("luxury perfusion"), and areas of frank ischemia leading to a reduction in brain metabolism can be mapped. Although these techniques revolutionized understanding of the pathophysiology of stroke, they are available only in specialized PET imaging centers that have a cyclotron available to generate ^{15}O. Thus, clinical PET is seldom utilized in the management of acute stroke. [^{18}F]fluorodeoxyglucose PET uses a much longer half-life radionuclide (^{18}F vs ^{15}O) to demonstrate a "patchy" pattern of metabolic compromise in patients with vascular dementia, corresponding to sites of repeated vascular injury.

2. Single Photon Emission Computed Tomography

SPECT requires administration of a radionuclide that emits alpha particles. The alpha particles collide with surrounding electrons, which then emit photons that scatter into detectors. Computers then determine the location of the emitting sources and create tomographic images mapping the quantity and location of emissions. The technique is useful for measuring CBF, both quantitatively and qualitatively. Quantitative measures of CBF may be obtained using the ^{133}Xe inhalation technique, whereas relative measures of CBF are obtained following injection of [^{99m}Tc]*d*, 1-hexamethylpropylene amine oxime (HMPAO).

3. Activation Procedures

Both PET and SPECT may be used to measure a change in CBF. By injecting only a portion of the HMPAO tracer during one condition (baseline) and injecting another dose during a second condition (e.g., following carbon dioxide inhalation or acetazolamide infusion), the brain images obtained during these two

conditions may be compared. The same technique (split-dose injection) may be used in PET scanning. In PET it is more common to utilize a short-lived radionuclide such as ^{15}O so that many separate images of CBF may be obtained during a single scanning session. Each of these CBF maps can be compared, but usually mean CBF images are compared to measure CBF changes from one condition (e.g., rest) to a second condition (e.g., activation). In clinical practice, HMPAO SPECT is often used to distinguish vascular dementia from degenerative dementia (as described previously) or to demonstrate impaired CBF distal to an occluded vessel, as in vasospasm following subarachnoid hemorrhage. HMPAO SPECT demonstrates CBF changes in response to acetazolamide or carbon dioxide, and impaired reactivity is associated with greater stroke risk. Thus, HMPAO SPECT provides clinically useful information about cerebral physiology and may supplement the information obtained from TCD.

E. Catheter Angiography

Cerebral angiography involves insertion of a catheter into the arterial system, injection of iodinated contrast agents into the cerebral circulation, and rapid acquisition of plane radiographs of the head. Vascular anatomy is demonstrated clearly using this technique, which is the gold standard for defining cerebrovascular anatomy. Aneurysms, AVMs, angiomas, and other vascular malformations are detected and classified. Arterial narrowing due to atherosclerosis, vasculitis, vasospasm, arterial dissection, or external compression is visualized. Microvascular disease may not be evident. The risk of complications is approximately 1 in 10,000, and catheter angiography is being utilized less frequently for routine diagnosis as MRA becomes available. Instead, catheter angiography is increasingly restricted to those situations in which interventional radiology procedures or surgery may be needed.

IV. THERAPEUTIC ADVANCES

The various treatments available for stroke are summarized in Table IV. These treatments include agents that directly interfere with coagulation, techniques to obliterate vascular obstruction or structural abnormality, and methods to bypass the obstruction physically or chemically (hyperbaric oxygenation and retrograde cerebral perfusion).

Table IV
Treatments for Stroke

Treatment type	Examples
Antiplatelet agents	Aspirin
	Dipyridamole
	Ticlopidine
	Clopidogrel
Anticoagulants	Heparin
	Low-molecular-weight heparins
	Warfarin
Thrombolytics	Streptokinase
	Urokinase
	Pro-urokinase
	Tissue plasminogen activator
Surgery	Aneurysm clipping
	Arteriovenous malformation resection
	Endarterectomy
	Subclavian thrombectomy
	Valve replacement
	Inferior vena cava filter
	Patent foramen ovale repair
	Atrial septal aneurysm repair
Radiation	Gamma knife
Endovascular therapy	Thrombectomy
	Angioplasty
	Coils
	Stents
	Balloons
	Glue
Cognitive rehabilitation and neurorehabilitation	Memory, language, and attention therapy
	Physical, occupational, speech, recreational therapies
	Assistive devices and adaptive equipment
Other	Hyperbaric oxygen
	Retrograde transvenous neuroperfusion
	Neuroprotective drugs

A. Antiplatelet Agents

Antiplatelet drugs used in the treatment of stroke include aspirin, dipyridamole, ticlopidine, and clopidogrel. Aspirin has been shown to improve the outcome from acute stroke without increasing the risk of brain hemorrhage. The combination of aspirin and

dipyridamole is more effective than aspirin alone in preventing stroke. Antiplatelet agents reduce the risk of nonfatal stroke by 25% overall and by 31% in high-risk patients (those with acute myocardial infarction or prior myocardial infarction, stroke, or transient ischemic attack). A subset of studies revealed a 24% reduction in subsequent disabling or fatal stroke and a 17% reduction in nondisabling stroke among treated high-risk patients. Ticlopidine reduced nonfatal stroke or death rates more than aspirin during the first year after transient ischemic attack or minor stroke, but by 3 years there was no difference between these agents. Clopidogrel is more effective than aspirin, reducing the risk of ischemic stroke, myocardial infarction, or death by 8.7% compared to aspirin. Clopidogrel is recommended for the treatment of stroke patients who also have peripheral vascular disease. Clopidogrel has fewer hepatic and hematologic side effects than ticlopidine, and it is the preferred alternative for patients allergic to or intolerant of aspirin.

B. Anticoagulants

There are no compelling data to support the widespread use of heparin in acute stroke. The Ischemic Stroke Trial examined the outcome of approximately 20,000 patients treated acutely with heparin with or without aspirin. Although a significant reduction in recurrent ischemic strokes was found among patients treated with heparin, this group also experienced significantly more hemorrhagic strokes. The routine use of heparin for acute stroke is thus not supported by the literature, even in specific circumstances such as presumed cardioembolic stroke, progressing stroke, and vertebrobasilar ischemia. Anticoagulation therapy appears to lower the risk of recurrent stroke in the setting of aortic arch atherosclerosis, but cholesterol emboli are often produced by such pathology. Thus, the optimal management of aortic arch disease is uncertain. It is also unclear whether anticoagulation or thrombolytic therapy can prevent cardiogenic embolization in the setting of acute myocardial infarction. In the setting of atrial fibrillation, however, anticoagulation is recommended. Treatment with warfarin decreases the risk of recurrent ischemic stroke by 68%. Patients younger than 60 years of age without concurrent stroke risk factors may not require warfarin for their "lone" atrial fibrillation. Instead, such individuals are often treated with aspirin to prevent future stroke. When acute stroke is complicated by atrial fibrillation, anticoagulation with intravenous heparin is typically utilized, despite relatively limited data supporting this practice. Low-molecular-weight heparins (LMWHs) have recently been utilized in such situations because they have lower rates of hemorrhagic complications and there is less need for blood testing compared to heparin therapy. Efficacy in the prevention of pulmonary emboli and peripheral venous thrombosis has been established. Preliminary research reveals that the rate of death or dependency 6 months following stroke is lower with LMWH treatment than without, and high-dose treatment is more effective than low-dose treatment. Additional research is under way to establish the efficacy of LMWH therapy in acute ischemic stroke. Because LMWHs are given by injection, they are not currently used for long-term stroke prevention, except perhaps for rare patients unable to tolerate all antiplatelet agents and warfarin. Oral forms of LMWH would be a welcome development.

C. Thrombolytics

Thrombolytic therapy has revolutionized the management of acute stroke. Prior to the demonstration that tPA was effective for acute ischemic stroke, aspirin, heparin, and supportive care were the mainstays of acute stroke management. Thrombolysis using tPA may now be done via intravenous injection or intraarterial injection. A favorable outcome is substantially more likely for patients who receive such treatment within 3 hr of stroke onset. The risk of hemorrhagic complications increases greatly when patients are treated more than 4 hr after the stroke, when severe hypertension complicates the treatment, and when more than one-third of the middle cerebral artery vascular territory appears damaged on CT scans. If none of these poor prognostic indicators are present, and if the patient lacks a history of recent surgery, stroke, transient ischemic attack, or anticoagulant therapy, then tPA may be administered acutely.

D. Cardiothoracic Surgery

Surgery on the heart may be helpful in preventing stroke. The aortic and mitral valves are the most common valvular sources of cerebral emboli. Valve replacement can reduce stroke risk. Intracardiac shunts may be repaired to prevent "paradoxical" embolism. Surgery may be done to remove clots from within the left atrium or atrial appendage. Atrial septal

aneurysms may be repaired. Aberrant conduction pathways producing arrhythmias may be ablated or removed. Subclavian stenosis may be surgically repaired, thus eliminating subclavian steal syndrome. Finally, coronary artery bypass grafting reduces the risk of myocardial infarction, a definite cause of stroke. Thus, there are many indications for cardiac or thoracic surgery to reduce stroke risk.

E. Vascular Surgery and Neurosurgery

The most common vascular or neurosurgical procedures done to prevent stroke are carotid endarterectomies. In 1989, there were approximately 70,000 carotid endarterectomies performed in the United States, and approximately half of these were for asymptomatic carotid stenosis. At that time, the value of surgery for stroke prevention was controversial. The Asymptomatic Carotid Artery Study (ACAS), the European Carotid Surgery Trial (ECST), and the North American Symptomatic Carotid Endarterectomy Trial (NASCET) demonstrated the benefit of surgery for both asymptomatic and symptomatic carotid artery stenosis. In the NASCET, carotid endarterectomy decreased the risk of stroke over 2 years by 17% compared to medical management in symptomatic patients with 70–99% stenosis. With greater than 80% internal carotid stenosis, carotid endarterectomy provided an absolute risk reduction of 11.6% in stroke and death end points over a 3-year time period. A smaller benefit (6.5% absolute risk reduction over 5 years) was evident for carotid endarterectomy in symptomatic patients with 50–69% stenosis compared to medical management. The results of the ECST were similar. In asymptomatic individuals with high-grade stenosis ($>70\%$) reported in the ACAS, the relative risk reduction over 5 years was 53%, but the absolute risk reduction was only 1.2% per year. Thus, the number of strokes prevented by carotid endarterectomy in asymptomatic stenosis of $>70\%$ is similar to the benefit seen with symptomatic stenosis of 50–69%.

Surgical repair of other vascular abnormalities is also possible. Extracranial aneurysms of the carotid arteries may be surgically repaired to prevent rupture. Intracranial aneurysms are usually clipped, but endovascular therapies are increasingly utilized for surgically inaccessible or giant aneurysms; detachable coils may be inserted into the vessels (Fig. 9), or balloons may be inserted. These techniques may lead to successful clot formation and obstruction of blood flow into the aneurysm, thus reducing the risk of rupture. Surgical removal of AVMs may be quite successful, although large AVMs may also require adjunctive endovascular therapy. For example, large and complex AVMs supplied by several vessels may need to be obliterated partially by intraarterial glue therapy prior to surgical resection. In patients with paradoxical cerebral embolization from lower extremity deep venous thrombosis, an inferior vena cava filter (Greenfield filter) may be inserted, or the intracardiac shunt may be surgically repaired.

F. Gamma Knife Therapy

AVMs have been treated successfully with focused radiation therapy. The Gamma Knife is a radiation therapy instrument that delivers several beams of radiation to a precise target. Each individual beam delivers a relatively small radiation dose to brain tissue. However, when the beams converge at the site of an AVM, the radiation dose is sufficient to produce damage. The blood vessels within the AVM degenerate and the adjacent tissue may be scarred as well (Fig. 10). This procedure has been used to obliterate AVMs that are located in surgically inaccessible areas of the brain. The risks of surgery, angiography, and endovascular therapy are avoided. Other vascular abnormalities such as aneurysms may be treated using the Gamma Knife. This technique represents a substantial advance in the available options for treating AVMs and other vascular anomalies

G. Endovascular Techniques

Some of the most recent advances in stroke therapy relate to the development of new catheters for interventional neuroradiology procedures. There are now catheters that can inflate a balloon within an area of stenosis to open a lumen within an atherosclerotic area of the vessel (cerebral angioplasty). Angioplasty and intraarterial papaverine injections have recently been used to successfully manage vasospasm following subarachnoid hemorrhage. A small loop or snare may be guided to the site of a thrombus using a catheter so that the thrombus may be removed directly through the catheter. Mechanical disruption of the clot or embolectomy (removal of the embolus) may allow recanalization of the vessel, with restoration of CBF. Thin wires may be inserted through the catheter to achieve recanalization through a recently occluded

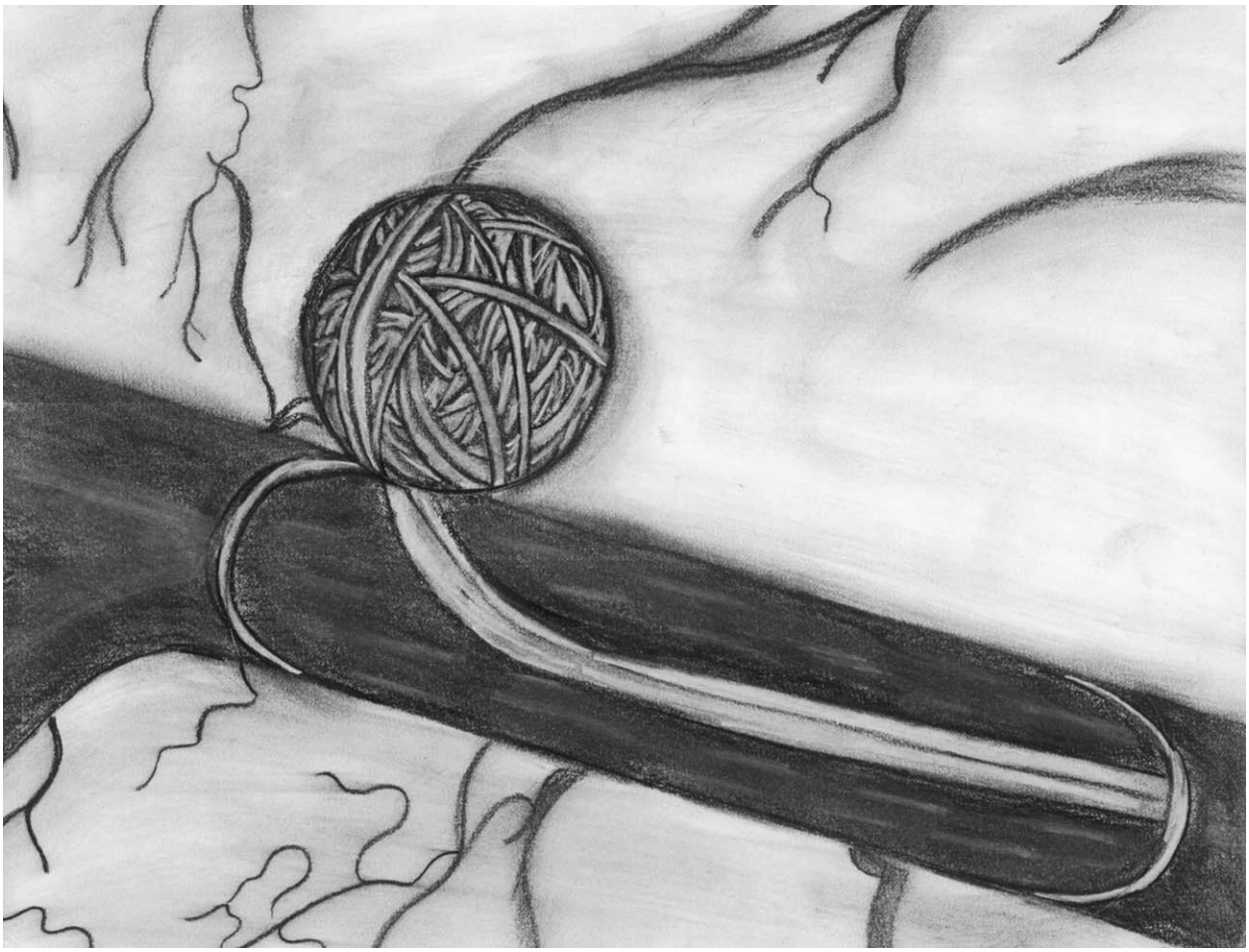


Figure 9 Intravascular coil within an aneurysm. This is a drawing showing how a catheter may be used to insert a detachable coil into an aneurysm. The wall of the vessel is cut away so that the catheter can be seen within the lumen. The coil is inserted through the catheter, and this partly occupies the aneurysm cavity. The coil may be detached so that it remains in the aneurysm permanently. A similar technique may be utilized to insert balloons into giant aneurysms (not shown) (drawing by Adam Bourgoin).

arterial segment. Catheters may deploy intravascular stents that remain in the vessel lumen to maintain vascular patency. Glue may be administered through the catheter to occlude feeder vessels within a large AVM. Test occlusion of such vessels using a balloon (or anesthetic injection into the vessel) may prove that sacrificing the vessel surgically (e.g., to remove an AVM) would not produce unacceptable complications. Balloons or coils (Fig. 9) may be deployed into a large aneurysm to promote coagulation within the lumen of the aneurysm. If blood cannot enter the lumen, the risk of rupture is substantially reduced. Specially formed catheters allow cannulation of small branches of intracranial arteries. Intraarterial thrombolysis may be applied specifically to the vessels that are occluded. Recently, TCD has been used to agitate the clot following thrombolytic administration, and this technique increases the rate of recanalization. These and other evolving endovascular techniques promise to revolutionize the management of acute stroke.

H. Cognitive Rehabilitation and Neurorehabilitation

Neurobehavioral and neuropsychological testing is useful for predicting the outcome of stroke (prognosis) and in selecting the most appropriate forms of therapy to be administered during rehabilitation. Cognitive rehabilitation utilizes neurobehavioral and neuropsychological information in treatment strategies

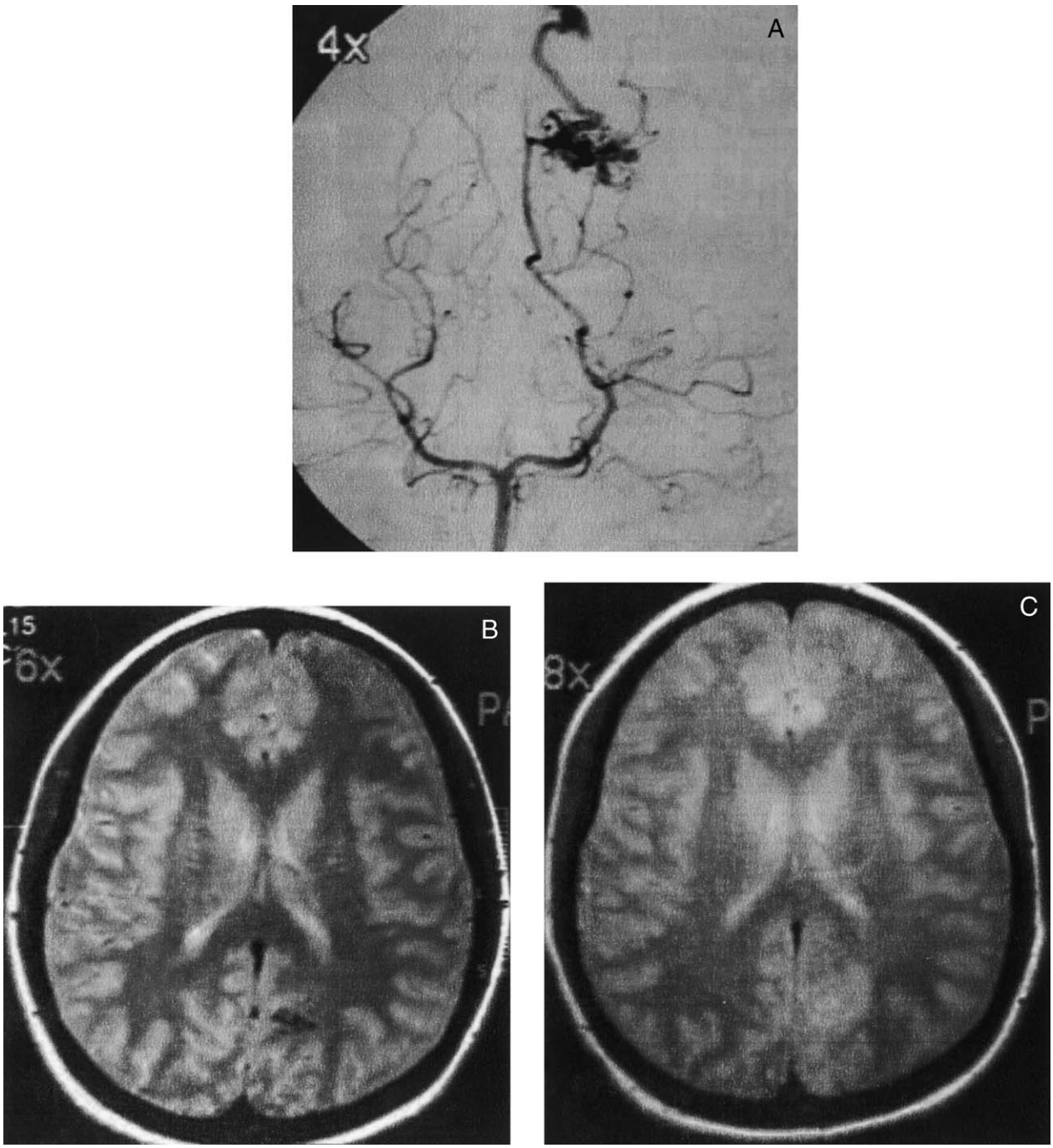


Figure 10 Successful Gamma Knife therapy. A conventional (catheter) angiogram depicts a moderate-sized arteriovenous malformation. This is a cluster of vessels that appears as a dark mass in Fig. 10A. (B) Flow voids (hypointense) corresponding to this AVM. (C) Following treatment with the Gamma Knife, the AVM has been obliterated, leaving only mild hyperintensity in the area that previously revealed flow voids.

designed to enhance functional recovery of memory, language, and attention following stroke. Neuropsychologists, physiatrists, neurologists, speech therapists, occupational therapists, recreation therapists, and other rehabilitation specialists may work together in multidisciplinary cognitive rehabilitation programs.

Functional brain mapping with fMRI, CBF PET, and electrophysiological brain mapping have revealed important information about the mechanisms of functional restoration and recovery. The clinical application of these techniques is just beginning to become integrated into the recovery phases of stroke management.

I. Experimental Therapies

Hyperbaric oxygen therapy has proved to be successful in the treatment of certain forms of stroke. For example, decompression sickness occurs when nitrogen bubbles form within the bloodstream. The hyperbaric chamber causes the nitrogen bubbles to dissolve and delivers oxygen to ischemic brain tissue. The same treatment may be utilized for air embolism, which occurs as a result of trauma, surgery, or vascular procedures. In each of these cases, the cause of the occlusion (i.e., a bubble) is dissolved by the pressure, while hyperbaric oxygen maintains viability within the ischemic brain areas. Now that thrombolysis and endovascular techniques are available to dissolve, recanalize, or remove an intravascular clot, it is possible that hyperbaric oxygen therapy may be combined with such procedures in the treatment of acute stroke. Currently, this hypothesis is under investigation.

By forcing oxygenated arterial blood backward through the venous system, ischemic areas may be supplied with needed oxygen even when the arterial supply is occluded. This technique, called retrograde transvenous neuroperfusion (RTN), is being actively studied to measure its effectiveness as an acute stroke intervention. In a small clinical trial involving eight subjects, RTN has shown promise for the management of acute stroke. Its effect may be almost immediate, in contrast to the effect of thrombolytic therapy, and it may be further increased by simultaneously employing hypothermia to reduce cerebral metabolic demands.

Neuroprotective drugs are under active development. The idea is to arrest the pathophysiologic mechanisms that lead to tissue damage and irreversible injury. If CBF or oxygen delivery could then be restored, a larger volume of intact and viable brain tissue would be available to promote functional recovery. Although the core region of infarction is not likely to be impacted by such an approach, the surrounding "ischemic penumbra" could in theory be rescued completely. Animal studies have shown the feasibility of this approach, but to date there are no effective neuroprotective drugs available for use in humans.

Advanced imaging techniques are being developed and promoted faster than they can be evaluated. Many of these techniques (FLAIR, MRA, MRV, etc.) have rapidly revealed their clinical utility, whereas others (MR spectroscopy and CBF PET) have limited clinical utility due to lack of availability, complexity, expense, or other factors. Despite these limitations, advanced brain imaging techniques enable brain areas at risk of permanent damage to be identified (Fig. 3) quickly and efficiently. This ensures that future research will be able to combine neuroimaging with clinical interventions to demonstrate the relative potencies of various medical therapies for acute stroke, neuroprotective interventions, endovascular techniques, or surgical procedures.

V. CHALLENGES FOR THE FUTURE

The number of people with stroke is increasing, despite the substantial gains in stroke prevention and treatment and reduced stroke risk for all age groups. The demographic shift in the population (i.e., the "graying" of America) accounts for this paradox. The great challenge of the future will be to respond intelligently to this situation. Quality of life will become increasingly important. Elderly individuals may survive a stroke only to be left incapacitated or demented due to vascular damage. Advances in public policy and the ethics of end-of-life care are required. For example, medical futility policies may be adopted to permit a physician to cease health care services that have no meaningful chance of improving quality of life in a hopelessly devastated stroke victim. Public education initiatives could substantially shorten the delay between stroke onset and emergency evaluation. Now that effective acute stroke intervention is available, the public must be educated about the importance of seeking immediate attention for "brain attacks." The reasons for socioeconomic and racial differences in stroke and cerebrovascular death rates must be understood so that preventable morbidity and mortality may be avoided. Such sociocultural developments have far greater potential to influence the future of cerebrovascular diseases than any technical innovation.

The technical challenges within the field of cerebrovascular diseases are considerable. A team approach to the management of cerebrovascular disorders will become the standard of care because the complexity of the field overlaps many clinical specialty areas. Basic

research must achieve meaningful neuroprotection so that the process of brain damage may be temporarily arrested while CBF and oxygenation are restored acutely. The stroke team must learn to mobilize quickly to administer neuroprotection and deliver acute stroke therapy, such as intravenous thrombolysis. These teams must learn to cooperate with regard to research so that the utility of diagnostic and therapeutic approaches can be validated. Idiosyncratic approaches to stroke care will be stifled by the purchasers of health care in favor of evidence-based interventions. The nuts and bolts of stroke care will be studied ever more closely to determine which practice variations are most damaging and which are most helpful. Quality of life assessments will become the standard of care, and the development of advanced directives will be promoted aggressively as a component of routine health maintenance. Approaches to the diagnosis and treatment of stroke that previously might have entered the clinical arena with little or no validation must now be examined to establish safety and efficacy in comparison to accepted standards and with proper attention to functional outcomes, advanced directives, cost, and other sociocultural considerations. Advanced brain imaging techniques will be employed to assist in the process of predicting stroke outcome, and functional brain imaging (in combination with sophistical clinical assessments) will become an indispensable aspect of cognitive and neurorehabilitation. The field of stroke has advanced tremendously during the past 25 years, and the next 25 years will be even more rewarding.

See Also the Following Articles

ALCOHOL DAMAGE TO THE BRAIN • BRAIN DAMAGE, RECOVERY FROM • BRAIN DISEASE, ORGANIC • CEREBRAL EDEMA • COGNITIVE REHABILITATION • RESPIRATION

Suggested Reading

American Heart Association (2000). *Heart and Stroke Statistical Update*. American Heart Association Dallas.

Bakshi, R., and Ketonen, L. (2001). Brain MRI in clinical neurology. In *Baker's Clinical Neurology* (R. J. Joynt and R. C. Griggs, Eds.), Lippincott, Williams & Wilkins, Philadelphia.

Barnett, H. J. M., Mohr, J. P., Stein, B. M., and Yatsu, F. M. (Eds.) (1992). *Stroke, Pathophysiology, diagnosis, and management*, 2nd ed. Churchill Livingstone, New York.

Gomez, C. R. (Ed.) (1999). Endovascular therapy and neurocritical care. *Crit. Care Clin.* **15**(4), 667–869.

Morgenstern, L. B. (Ed.) (2000). Stroke. *Neurol. Clin.* **18**(2), 273–516.

Mazziotta, J. C., and Gilman, S. (1992). *Clinical Brain Imaging: Principles and Applications*. Davis, Philadelphia.

Stark, D. D., and Bradley, W. G. (Eds.) (1992). *Magnetic Resonance Imaging,* 2nd ed. Mosby–Year Book, St. Louis.

Tegeler, C. H., Babikian, V. L., and Gomez, C. R. (Eds.) (1996). *Neurosonology*. Mosby–Year Book, St. Louis.

Chemical Neuroanatomy

YURI KOUTCHEROV, KEN W. S. ASHWELL, and GEORGE PAXINOS
University of New South Wales

GLOSSARY

chemoarchitecture Differentiation between neuronal populations and brain areas based on variation in distribution of chemical compounds.

cytoarchitecture Differentiation between neuronal populations and brain areas based on variation in cell size, morphology, and density.

immunohistochemistry A histological technique that uses specific antibodies for visualizing chemical compounds.

Chemical Neuroanatomy is a modern methodology that identifies and classifies brain structures and pathways by their chemical characteristics. Because chemical characteristics of a brain structure are related to the function of its constituent neurons, chemical neuroanatomy allows insights into the physiological, pharmacological, and clinical significance of neuroanatomy. In other words, chemical neuroanatomy creates meaningful structural references within the general scope of brain and whole body physiology and psychology.

I. CHEMOARCHITECTURE

Chemical neuroanatomy has been used to establish the organizational plan of brain regions in experimental animals and to infer their human homologies. It has also been used to identify chemically specified connections in animals. Finally, it has been used to derive an hypothesis on the function of brain pathways and nuclei. Chemical neuroanatomy has developed as a branch of the structural brain mapping methodology that previously was almost entirely based on cytoarchitectonic consideration of cell shape, size, and density. The insubordination of chemically specified neurons to classic cytoarchitectonic boundaries required a more meaningful delineation of the brain—one that incorporates the information about the distribution of neuroactive substances, connectivity, and function. In this respect, chemical neuroanatomy opened a new dimension in neuroscience and allowed greater precision, resolution, and reliability in differentiating cell groups and brain areas.

Scientists have made use of chemical neuroanatomy in studying affiliations of neurons in experimental animals. Knowing the neurotransmitter carried by a neuronal projection precisely characterizes the pathway. This approach has not been extensively followed in the human; consequently, we only mention it in passing. Once the cells of origin and the terminal fields of a pathway have been chemically specified in experimental animals, then scientist can search for homologous cells and terminals in the human.

Chemical neuroanatomy is used to identify and correlate functional networks across species, particularly when focused on a specific neuroactive circuit

within the brain. Thus, chemical characteristics of brain structure reflect numerous important qualities of the neuronal cell groups, including neurotransmitter or neuromodulator content, receptor arsenal, structural proteins, intrinsic metabolic characteristics as reflected by enzyme activity profiles, and genotype.

In the human brain, chemical neuroanatomy is an essential method of investigation because conventional functional studies as well as neuroanatomical tracing techniques, as routinely performed on laboratory animals, are virtually impossible in humans. Noninvasive imaging techniques, including functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) are useful clinical tools, but their resolution is often disappointingly low for testing modern hypotheses. At the same time, pathological investigations base their conclusions largely on detailed structural and cytoarchitectonic comparisons with normal human neuroanatomy. Importantly, existing presumptions about the organization of the normal human brain are largely made by extrapolating findings from laboratory animals to humans on the basis of structural homology.

For most of the 20th century the understanding of human neuroanatomy was obtained mainly from cytoarchitectonic observations. Thus, the most widely used maps of the human cortex were produced by Brodmann in 1909 on the basis of Nissl substance and myelin staining, whereas the most detailed neuroanatomical description of the human hypothalamus was published by Brockhaus in 1942 and was also based on early cytoarchitectonic techniques. Without detracting from the historical significance of these fundamental studies, it is easy to see the main shortcoming of early neuroanatomical techniques, namely, their distance from mechanisms underlying human brain function. One of the most exciting developments in neuroanatomy was the identification of chemical coding for individual neural pathways and the proliferation of chemoarchitectonic techniques, which allow almost unlimited scope in the classification of neuronal groups. Importantly, chemical neuroanatomy establishes a bridge between structural and functional characteristics of neuronal populations in the brain.

Studies using chemical neuroanatomy were first carried out in rats, in which it was logistically and technically easier to apply. It was not until the 1980s that the chemoarchitectonic techniques of histo- and immunohistochemistry reached sufficient sensitivity that allowed them to be applied in full capacity to human brain tissue. Thus, it became increasingly possible to reveal the distribution of some of the neurotransmitters, receptors, and enzymes of importance in the human brain and then make cross-species comparisons. An advantage of chemoarchitecture is that each chemical substance offers a different view of the organization of the central nervous system, with successive stains revealing more of the areas of interest. Of course, there are significant species differences and any given substance may have inconsistent distributions in otherwise homologous nuclei and areas. Nevertheless, in terms of overall value, chemoarchitectonic delineations have become a preferred method in comparative neuroscience.

Chemoarchitectonic studies are also very important for the understanding, diagnosis, and treatment of neurological and psychiatric disorders, such as obesity, Alzheimer's, disease, Parkinson's disease, Huntington's disease, depression, and schizophrenia. In the past 50 years, researchers have compared the distribution and content of neuroactive chemical compounds of brains from individuals affected with various diseases with those of the brains of control subjects. This has led to development of pathological models of neurochemical imbalance in animals and to extrapolation of these models into the human using chemical neuroanatomy.

In chemical neuroanatomy, the first and most obvious question for the investigator is the choice of the chemical marker. Naturally, the neuroactive profile of neurons offers grounds for determining the organization of neuronal groups within a species and for comparing them across species. For example, dopamine, norepinephrine, epinephrine, and γ-aminobutyric acid (GABA) are neuroactive chemical compounds that can characterize neuronal subgroups. The term chemoarchitecture implies the use of chemical compounds for differentiation between neuronal populations. These compounds are not only neurotransmitters but also can be enzymes, receptors, peptides, and molecules related to neuronal metabolism, such as calcium-binding proteins. Cross-referencing patterns of distribution of many chemicals from different brains or, better yet, studying distribution patterns of different chemicals on the adjacent sections of one brain provide high resolution and reliability in identifying neuronal cell groups.

II. ENZYMES

An important part of chemical neuroanatomy was historically based on the distribution of various enzymes and enzymatic markers. In the mid-20th

century, Koelle and Friedenwald developed a simple and sensitive histochemical method for revealing acetylcholinesterase (AChE), the degradative enzyme for acetylcholine. The application of AChE staining has consequently proven very useful in distinguishing brain areas (Fig. 1), although failure to colocalize AChE-positive neurons with catecholamines, the then darlings of the neuroscience community, dampened the interest of researchers in enzyme expression in the human. A comprehensive delineation of the rat brain by Paxinos and Watson was done largely on the basis of AChE reactivity with Nissl staining used as a second criterion. In the past 30 years, AChE histochemistry was successfully used for delineations of the brain in many mammalian species. Most important, AChE histochemistry works well on the fresh (unfixed) postmortem human brain, which allows this method to be successfully applied to the neuroanatomical delineation of the human brain. For example, in 1995 Paxinos and Huang mapped the entire human brain stem using AChE reactivity as a primary chemoarchitectonic criterion. AChE staining was also used in pathological studies of the brains of patients with Alzheimer's disease and was recently employed as a relatively simple method for revealing the organization of the human hypothalamus (Fig. 1).

Another example in which enzyme distribution was instructive was the identification of the neuronal circuitry containing catecholamines (dopamine, noradrenaline, or adrenaline) by the localization of the enzymes related to catecholamine synthesis. Using immunohistochemical techniques, researchers determined distributions of tyrosine hydroxylase (TH), aromatic amino acid decarboxylase, dopamine β-hydroxylase, and phenylethanolamine-*N*-methyltransferase and were able to the delineate the relevant nuclei and subnuclei in the brain. In contrast, initial studies that aimed at directly localizing catecholamine molecules through chemical reaction with aldehydes were greatly limited by an inability to distinguish between the different catecholamines. The application of enzyme immunohistochemistry allowed the identification of 15 groups of catecholaminergic neurons in the mammalian brain. These cell groups were not confined to traditional cytoarchitectonic boundaries and consequently were termed A1–A16 (there is no A3 cell group), extending throughout the mammalian brain from the medulla to olfactory bulbs. In the human, as in the rat, the majority (A1, A2, and A4–A10) of catecholaminergic neuronal groups were found in the brain stem, where, for example, tyrosine hydroxylase immunostaining has been used to delineate the intermediate reticular zone. Four prominent tyrosine hydroxylase-positive catecholaminergic cell groups (A11–A15) are located in the hypothalamus (Fig. 2) and one (A16) is located in the substantia innominata of the ventral forebrain. The latter cell group is thought to be homologous to the rat's catecholaminergic cell group in the olfactory bulb. Subsequent work has shown that cell groups such as the Al and Cl catecholaminergic neurons are critical for autonomic control in experimental animals and also that these cell groups are strikingly similar in rat and human. Many studies used multiple markers to confirm a high degree of conservation in the chemical identity of brain stem neurons in general between rat, monkey, and human.

Neuronal nicotinamide adenine dinucleotide phosphate diaphorase (NADPH-d) is an enzyme that synthesizes the interneuronal messenger nitric oxide, hence the reason for the current intense interest in this protein. With respect to neuroanatomy, histochemistry reveals a restricted distribution of this enzyme in the brain, a feature that allows accurate identification of exquisitely detailed subnuclear structures. For example, NADPH-d reactivity has revealed the otherwise ambiguous supraoculomotor cap and the dorsolateral subdivision of the periaqueductal gray in the midbrain of the rat, rabbit, cat, monkey, and human. This finding allowed reliable identification of the interspecies homology for these microscopic subnuclear structures. Interestingly, the supraoculomotor cap has also been identified as being AChE positive.

NADPH-d reactivity has also been shown to reveal specific neuronal groups in the hypothalamus, thalamus, and medulla. It is a primary marker for the islands of Calleja in the ventral forebrain and reveals neuronal subpopulations in the human cortex. As an enzymatic neurochemical marker, NADPH-d histochemistry holds significant potential in current neuroanatomy and has been selected for delineation in popular atlases of rat and monkey brains.

III. CALCIUM-BINDING AND STRUCTURAL PROTEINS

There are many protein molecules that bind Ca^{2+} and whose distribution is instructive in brain delineations. The most frequently used are the vitamin D-dependent calcium-binding protein calbindin D-28K, parvalbumin, S-100, and calretinin. The functional significance of these molecules remains unclear, and the main interest in them is derived from the large variability

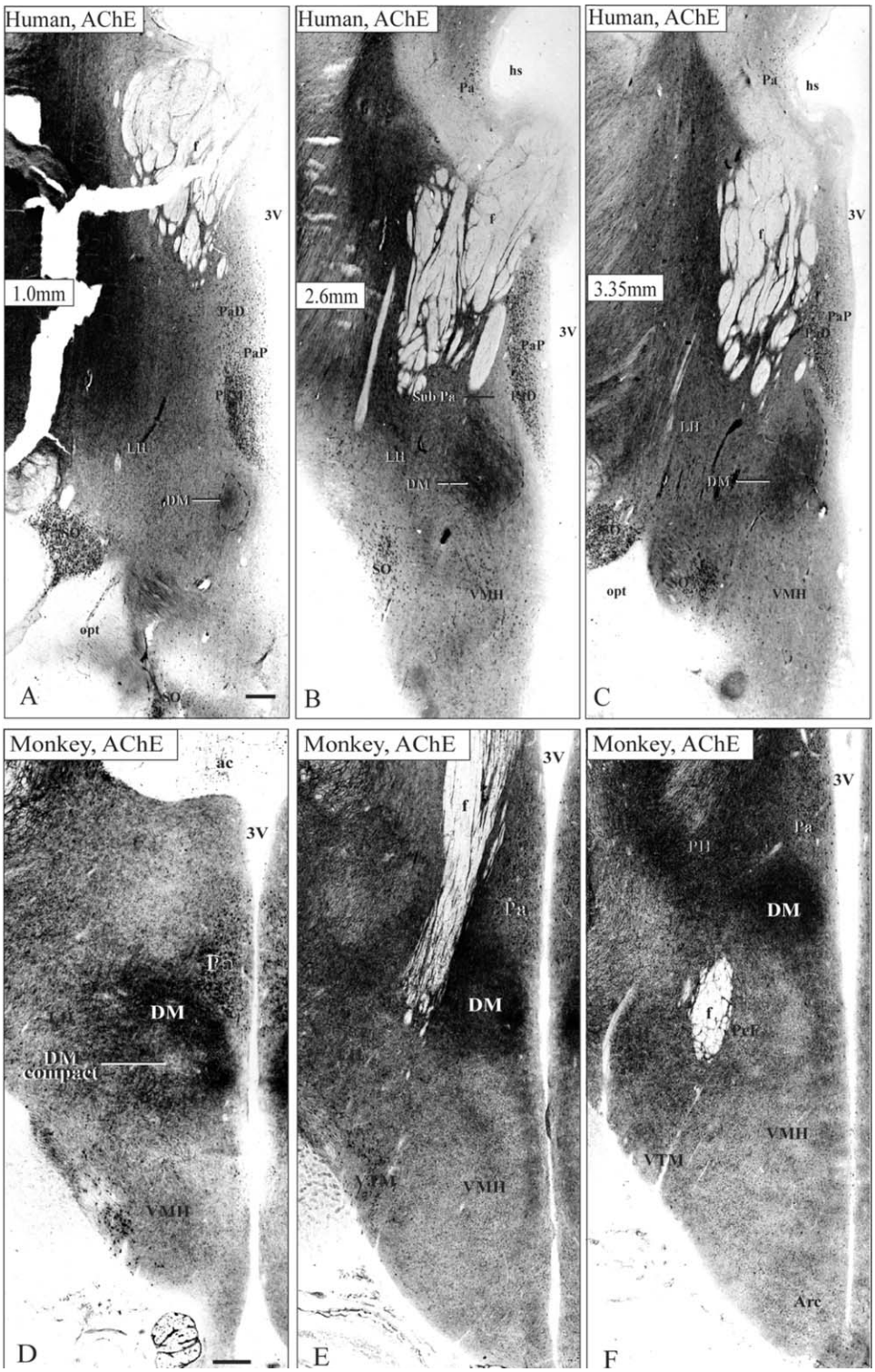


Figure 1 Photomicrographs of a series of coronal sections stained for acetylcholinesterase (AChE) that reveals the other wise vague boundaries of the dorsomedial nucleus at three rostrocaudal levels of the human (A–C) and monkey (D–F) hypothalamus. Stereotaxic coordinates are indicated on the left side of the photomicrographs for the human. Note the area between DM and Pa labeled in B as Sub Pa in the human but absent in the monkey. Also note an AChE-negative compact DM prominent in the monkey (D) but not in the human (A). Scale bar=1 mm.

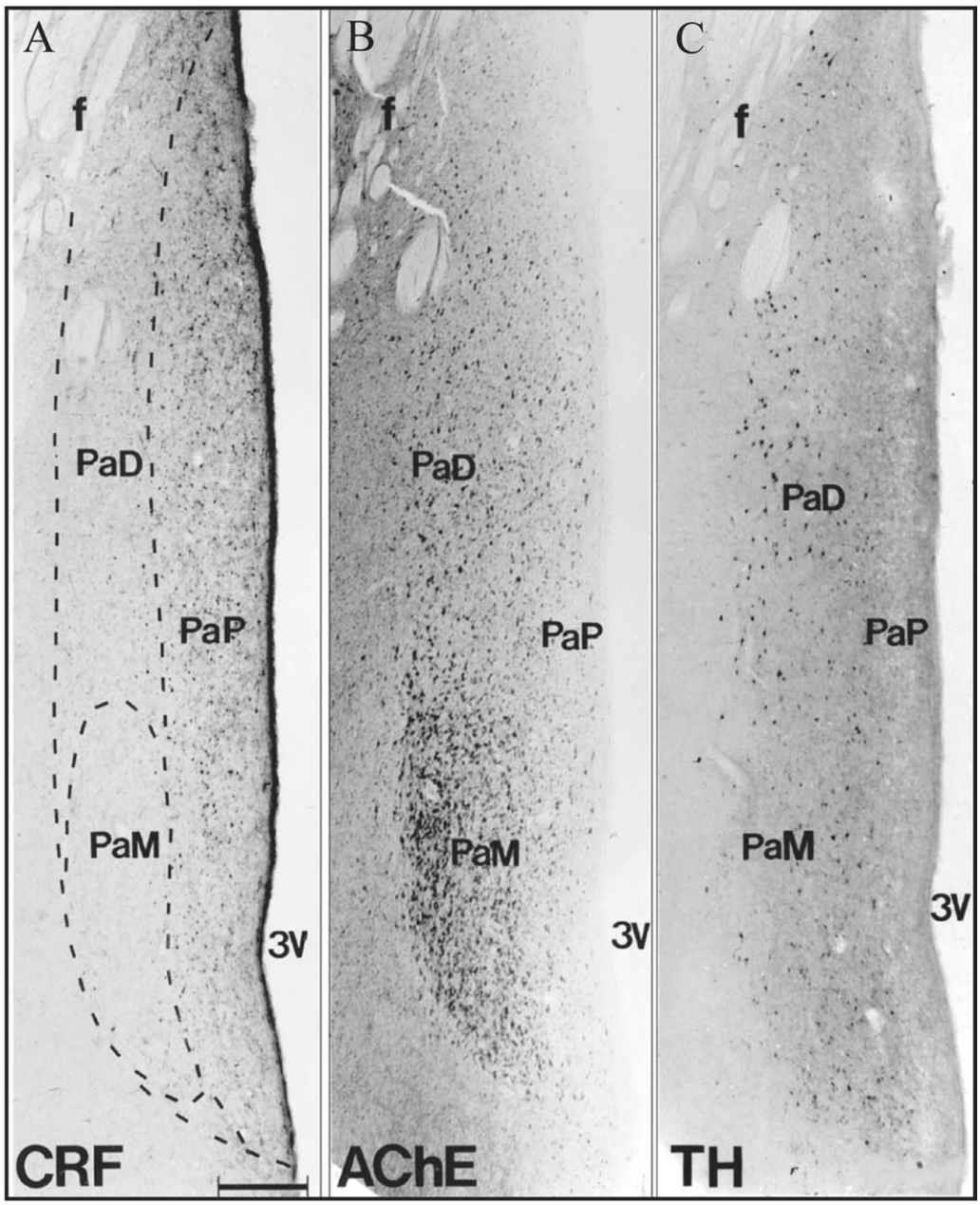


Figure 2 Adjacent coronal sections of the human hypothalamus processed alternately with immuno/enzyme histochemistry for corticotropin-releasing factor (CRF), acetylcholinesterase (AChE), and tyrosin hydroxylase (TH) that reveal the subnuclei organization of the human Pa. In each photomicrograph the third ventricle is to the right. Subnuclear boundaries are indicated by a dashed line in the CRF immunoreacted section. Scale bar=0.5 mm for all photomicrographs.

with which these molecules are distributed in different brain structures. Thus, above all, calcium-binding proteins are chemical markers differentiating brain structures with high resolution and consistency across species.

In the brain, calcium-binding proteins were successfully used to reveal the organization of the human thalamus, striatum, and hypothalamus. In the thalamus and hypothalamus, the content and morphological distribution of calbindin D-28K, parvalbumin, and calretinin varied greatly and consistently between different areas and cell groups. For example, distributions of calbindin and calretinin complement each other in differentiating the shell and core regions of the

accumbens nucleus, whereas parvalbumin is a distinct marker of the ventral pallidum. The distribution of calcium-binding proteins has been utilized in two ways; one related to the presence of the specific protein and another related to the structural boundaries this protein reveals. The first approach infers homology on the basis of similar chemical content of the putative homolog across species. For example, neurons of both the rat and the human medial preoptic nucleus (MPO), which is thought to be important for the regulation of sexual behavior, contain calbindin D-28K. The second approach infers homology on the basis of unambiguous borders as revealed by the distribution of these substances.

Among structural proteins, neuroanatomists often use the distribution of labeling by a monoclonal antibody directed against nonphosphorylated epitopes on neurofilament B protein [Sternberger Monoclonal Inc. 32 (SMI32)]. Immunohistochemical staining for SMI32 has been successfully used as an anatomical marker in the cortex, thalamus, and hypothalamus of the monkey and human. Hof suggested that the large size of SMI32-positive neurons found in the monkey and human cortex reflects the large distances to their neuronal targets. Recently, the pattern of SMI32 staining in the medial preoptic nucleus of the hypothalamus, an important regulatory center for sexual behavior, helped to establish a homology between the lateral MPO in the rat, monkey, and human hypothalamus (Fig. 3).

Occasionally, homologous nuclei in different species do not correspond fully in their chemical profile. For example, no one would dispute the homology of the basal nucleus of Meynert in the rat and human established on the basis of cyto- and chemoarchitecture. However, with respect to some substances, including calbindin D-28K, these cells have opposite profiles in the rat and human. However, the use of multiple markers, including calcium-binding proteins, affords greater confidence about the borders of this nucleus, revealing proportions and position that correspond to the homologous structure in the rat.

IV. PEPTIDES

The development of diverse antisera for a variety of peptides augmented the tool kit of chemical neuroanatomy. Neuropeptides are largely neuron-specific chemical compounds that, depending on the neuropeptide, are characteristic of specific neuronal subgroups. For example, vasopressin is characteristic of large cells in the lateral magnocellular subnucleus of the paraventricular hypothalamic nucleus. The neuropeptides are widely distributed and numerous, and for the sake of nomenclature they have been classified into families. For example, tachykinins (literally "fast acting") constitute a family of small, structurally related peptides that share a similar carboxy-terminal sequence but differ with respect to their function and distribution. The distribution of substance P, the most widely known tachykinin, was useful for the differentiation of four of nine subnuclei of the dorsal motor nucleus of the vagus nerve. Within these subnuclei, substance P staining revealed three different types of neurons. The neuropeptides, of course, are likely to be products of neuronal metabolism and thus a reflection of functional characteristics of the particular neurons. This does not mean that all the peptides are neurotransmitters or neuromodulators, but it does imply that the neuropeptide "alphabet" can be used for chemical coding of the neuronal groups that serve similar functions and are characterized by similar projections. In the *Rosetta stone* Greek was the common denominator between hieroglyphics and English. Likewise, the presence of a neuroactive substance in similar locations in the rat and the human brain can betray a structural homology.

The corticotropin-releasing factor (CRF) (or hormone) is a neuroendocrine peptide found in the cortex, basal telencephalon, brain stem, and hypothalamus. The best known function of CRF is to initiate the response in the hypothalamo–pituitary–adrena1 axis, but it is also implicated in stress responses, metabolic regulation, and food intake. Recent studies have found increases in the number of CRF-positive neurons in the human hypothalamus with age. Another report showed colocalization of CRF with nitric oxide synthase in a subgroup of cells in the human hypothalamus and suggested the possible involvement of these neurons in schizophrenia and depression. The distribution of CRF is very specific. Thus, in neuroanatomy CRF distribution has been used in the human brain to distinguish the subcompartmental organization of specific nuclei in the medulla and hypothalamus. For example, in the paraventricular hypothalamic nucleus, CRF neurons are confined to the parvicellular compartment, whereas the neurons that contain oxytocin are found primarily in the dorsal compartment. Applying these two markers to the same brain allowed researchers to distinguish between these subcompartments, which otherwise appear to be amalgamated, and it also allowed the establishment of subcompartmental homologies between human and rat paraventricular nucleus (Figs. 2 and 7).

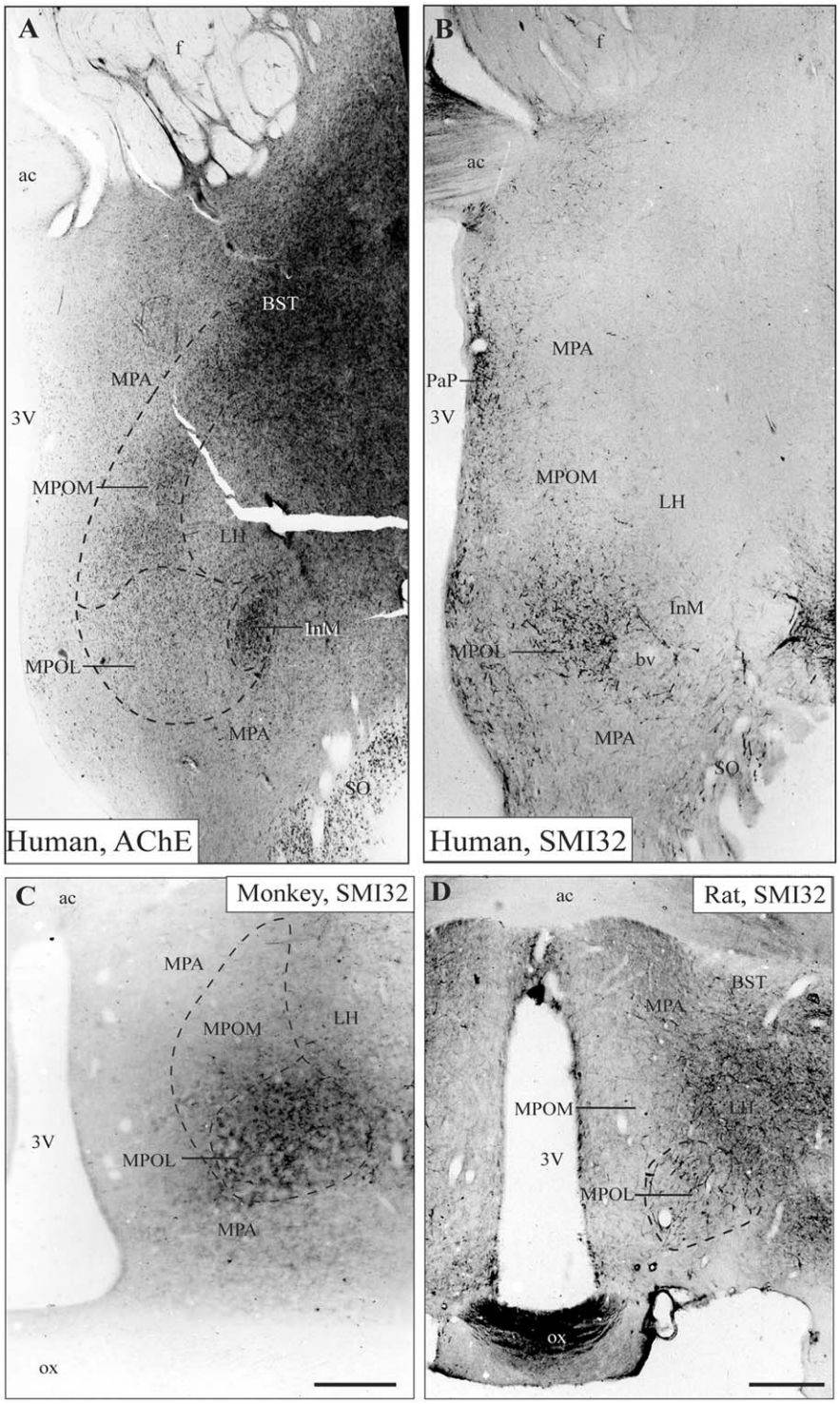


Figure 3 Photomicrographs of coronal sections through the human (A and B), monkey (C), and rat (D) anterior hypothalamus processed enzymatically and immunohistochemically for (A) AChE and (B) SMI32 showing complimentary distribution of these substances and revealing the three constituent subnuclei of the MPO. Sections in A and B are virtually adjacent. In all three species the lateral subnucleus of the medial preoptic nucleus (MPOL) contains SMI32 immunorecative neuropil, whereas the medial MPO (MPOM) is SMI32 negative. Acetylcholinesterase reactivity also reveals the celebrated sexually dimorphic (or intermediate InM) nucleus. Scale bar=1 mm (C) and 0.5 mm (D).

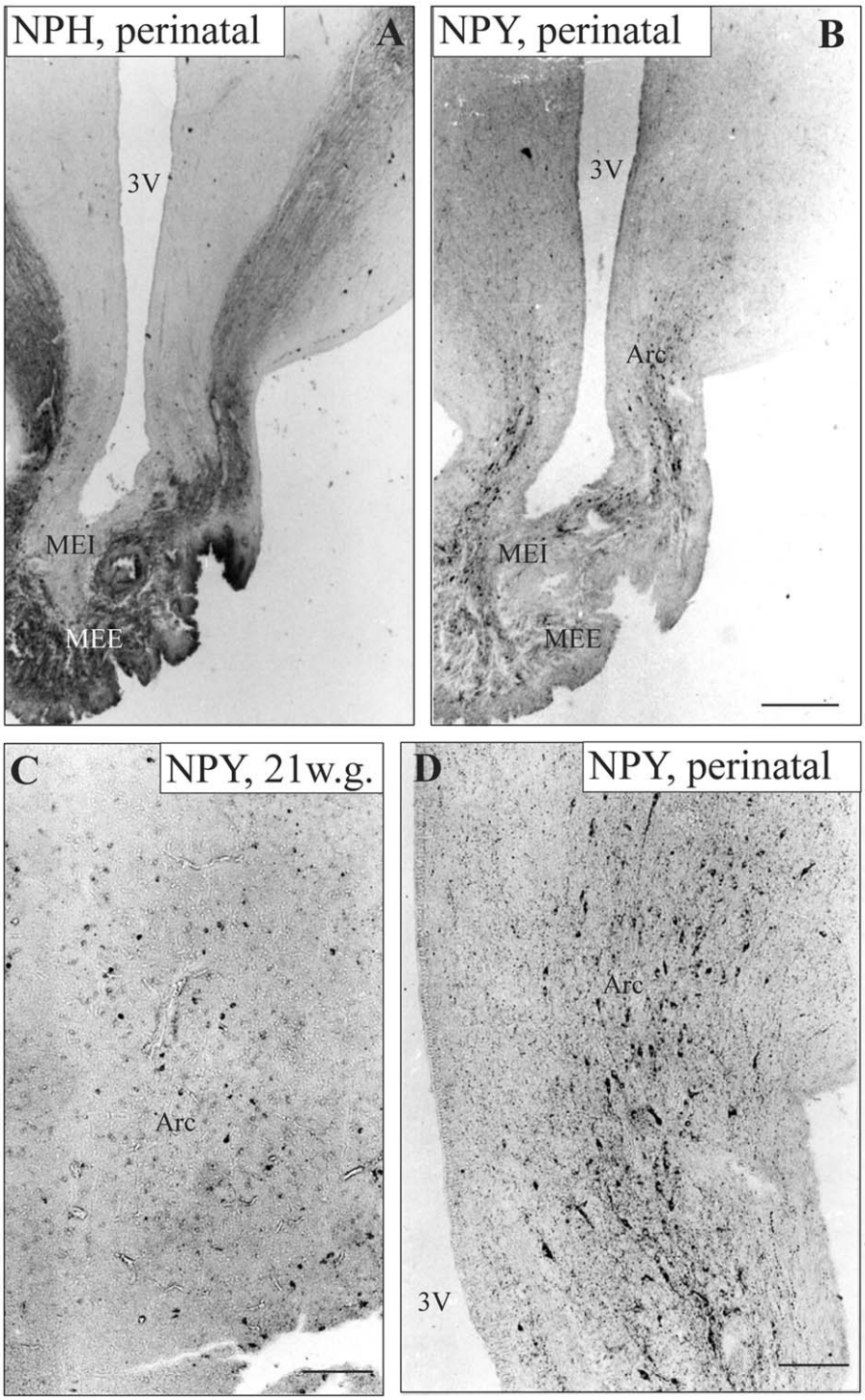


Figure 4 Photomicrographs of adjacent coronal sections through the human hypothalamus (demonstrating differential distribution of NPH-positive fibers (A) and neuropeptide Y (NPY)-positive cells and fibers (B). Immunoreactivity reveals details of intrinsic organization of the arcuate hypothalamic nucleus. (C) First appearance of NPY-positive neurons in the arcuate nucleus at 21 weeks of gestation. (D) Area of arcuate nucleus in B at higher magnification showing a population of morphologically developed NPY-positive neurons. (Reproduced from *J. Comp. Neurol.* with permission from John Wiley & Sons, Inc., 2002).

Neuropeptide Y (NPY) is a peptide that belongs to the pancreatic polypeptide family and is abundant throughout the central and peripheral nervous systems in both human and rat. In the brain, NPY is found at high concentrations in the mesencephalon, medulla, cortex, hippocampus, amygdala, ventral and dorsal striatum, and hypothalamus. As a neuroactive molecule, NPY has been implicated in the regulation of feeding behavior and autonomic and endocrine activity. Particularly instrumental for these ideas was the discovery of NPY-containing networks in the hypothalamus, where this peptide is confined approximately to specific neurons in the arcuate (Fig. 4) and suprachiasmatic nuclei and to fibers and neuronal terminals in the paraventricular and dorsomedial nuclei. Mapping the distribution of NPY in the human brain contributed greatly to the delineations of some functionally significant nuclei and areas. Thus, for example NPY defines the otherwise ambiguous and poorly visible ventral marginal hypothalamic zone, which has been implicated in the regulation of feeding.

V. RECEPTORS

Neuroanatomical delineations based on receptor distributions have been popular in pharmacological and neuroanatomical studies since the late 1960s and 1970s, particularly with the development of autoradiographic techniques. As a clear advantage, receptors were often confined to a small area in the brain and thus differentiated specific neuronal projections with great accuracy. This has not only given neuroanatomical delineations a greater precision and resolution than ever before but also helped our understanding of how drugs work in the brain. For example, neuroanatomical maps of receptors not only reveal the presence of a receptor specific for a certain drug in the brain but also reveal the specific circuitry to which it belongs. This, in turn, points to a mechanism of drug action. For example, Paxinos and Watson originally delineated the intermediate reticular zone in the rat brain on the basis of a band of negative AChE staining between the gigantocellular and parvicellular reticular nuclei. The area remained ambiguous until another group of researchers demonstrated that the zone exclusively contains angiotensin II receptors in the rat and human brain. Intense angiotensin II receptor binding in the intermediate reticular zone and no binding in the neighboring gigantocellular or parvicellular nuclei revealed clear boundaries of the structure and unambiguously pointed to its homology between rat and human.

The neurokinin B receptor (NK3) is an integral component of the neural circuitry of neurokinin B, which in turn is a member of the tachykinin family of peptides (also including substance P and neurokinin A). NK3 is also an important element of the hypothalamic neuronal circuitry regulating blood pressure. Recent studies used immunohistochemistry to reveal the distribution of NK3 in the human hypothalamus and to compare this with the distribution of the receptor in the rat hypothalamus, in which the structure–function associations are better studied. The strongest NK3-like immunoreactivity in the human hypothalamus was found in neurons of the paraventricular nucleus (Pa), specifically in the parvicellular and posterior Pa subnuclei, thus distinguishing these structural subcompartments (Fig. 5). Another prominent population of NK3-positive cells in the human hypothalamus was found in the perifornical nucleus, distinguishing it from the rest of the lateral hypothalamic area. In cross-species comparison there appeared to be a large degree of similarity in the distribution of NK3 in the human and rat hypothalamus.

In the past two decades, the technique of *in situ* hybridization became available, and with its myriad of probes it allowed further characterization of the cell groups (Fig. 6). The presence of mRNA indicates a condition for protein production within specific neurons. A clear advantage of the method is that the mRNA signal is quantifiable. The quantity of signal can then be compared across species or between human brains. The distribution of mRNA generally falls within the boundaries defined by chemoarchitecture but in some rare cases it has been known to differ with the location of the corresponding protein. As an alternative technique to histo- and immunohistochemistry, *in situ* hybridization provides an important confirmation of existing neuroanatomical delineations.

In describing chemical neuroanatomy as a methodology to reveal the organization of the human brain, it is useful to demonstrate the application of the approach in specific areas of the brain.

VI. HYPOTHALAMUS

It became clear from animal experiments that the small area of the hypothalamus contains a structural maze of

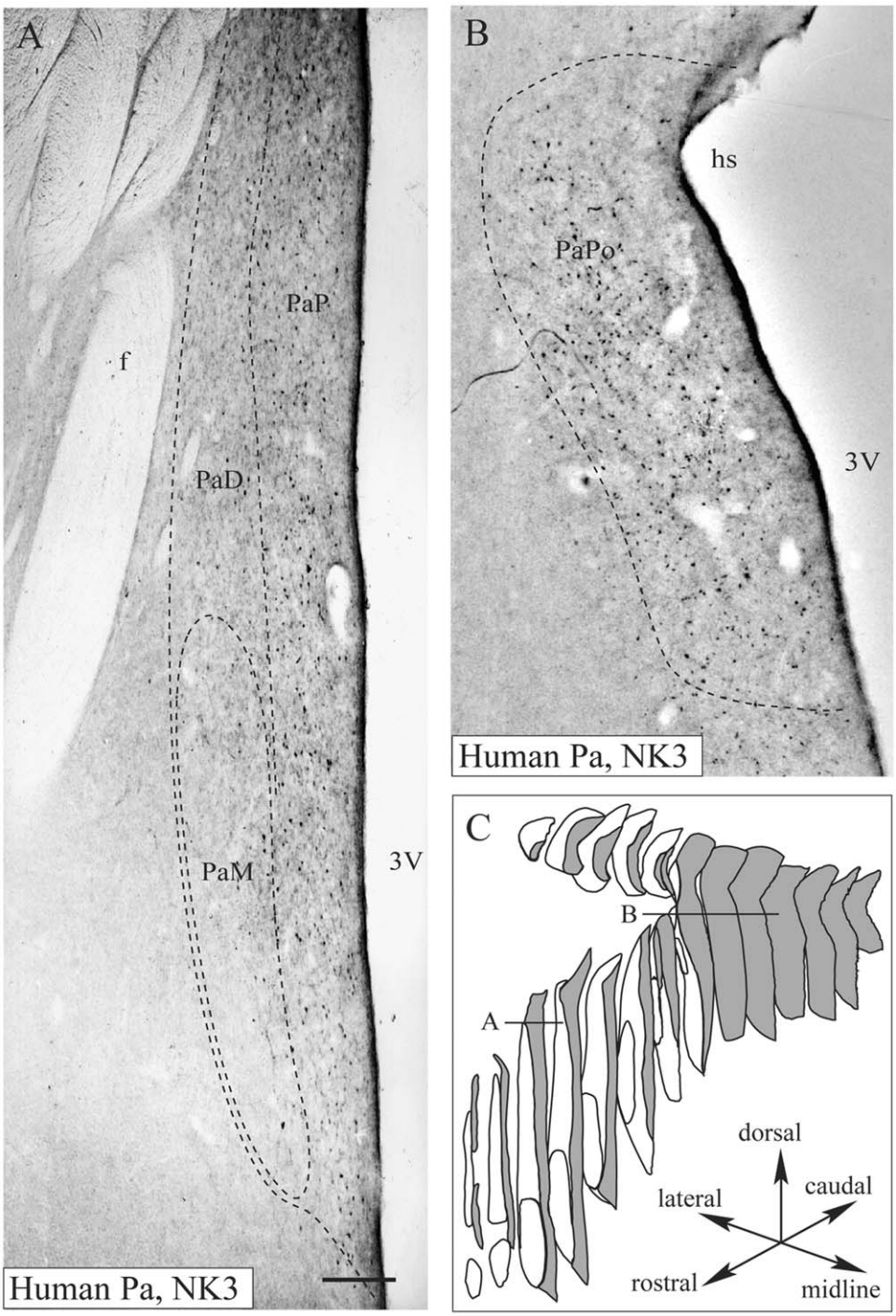


Figure 5 Photomicrographs of coronal sections from the human hypothalamus demonstrating the distribution of NK3 in different subcompartments of the paraventricular nucleus (A and B). Scale bar=0.35 mm. (C) The schematic diagram shows the parts of the Pa from which these sections were taken. (Reproduced from *NeuroReport* **11**(14) with permission from Lippincott Williams & Wilkins, 2000).

neural relays that integrate fundamental autonomic, endocrine, and behavioral responses into an elegant strategy regulating homeostasis and reproduction. These relays are cyto- and chemoarchitectonically distinct nuclei, which also differ with respect to their affiliations and functions. Extrapolations to humans of the conclusions obtained from studies on experimental animals have been based on the rationale that

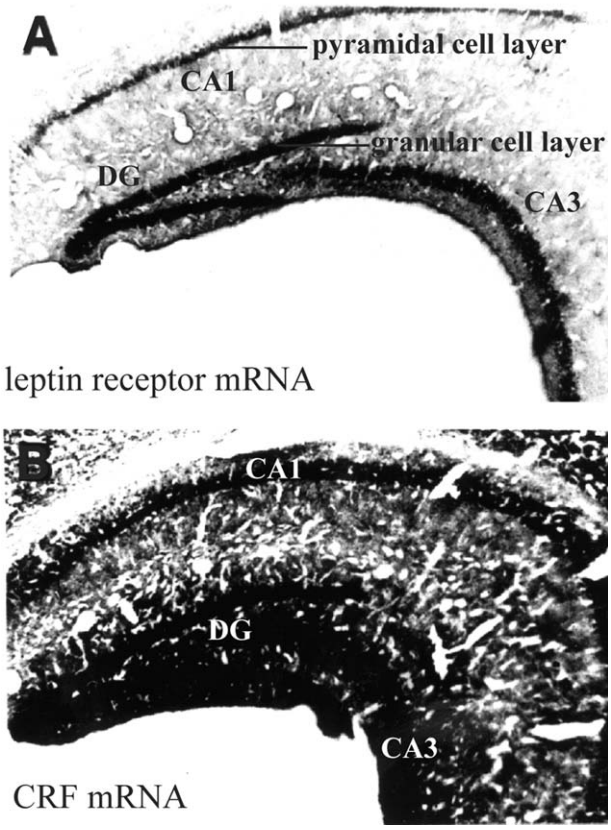


Figure 6 These photomicrographs of coronal sections through the hippocampal region demonstrate an example of nonradioactive *in situ* hybridization signal obtained with two different probes directed against leptin receptor mRNA (A) and against CRF mRNA (B).

homologous structures have analogous function across mammals. The homologies are established on the basis of chemoarchitecture.

The past century has seen comprehensive cytoarchitectonic and comparative studies, which have provided plans for the structural organization of the human hypothalamus. However, apart from the work of Saper and that of Braak and Braak, these studies relied on cytoarchitecture and myeloarchitecture for their delineations. Cytoarchitecture, when used as a single comparative criterion, has shortcomings because the cell groups in the adult human hypothalamus are dispersed and certainly less obvious with respect to their subnuclear boundaries than they are in the rat. As a consequence of ineffectual comparative investigations, there is confusion regarding the terminology of the human hypothalamic cell groups. The basic organizational plan of the hypothalamus is thought to be well preserved throughout the mammalian lineage. It became apparent that to produce a unified nomenclature it is necessary to study human and rat hypothalami in parallel, using the same criteria for the establishment of homologies. Thus, current neuroanatomical studies of the human hypothalamus use cytoarchitecture, chemoarchitecture, topography, subnuclear organization, and pattern of development as criteria for cross-species homologies.

For example, the chemoarchitecture of the human paraventricular hypothalamic nucleus was recently studied with the aid of three-dimensional computer reconstruction (Fig. 7). Chemoarchitecture revealed five subnuclei in the adult human Pa. The most prominent of these is the magnocellular subnucleus (PaM), occupying the ventrolateral quadrant of the Pa and composed of a concentration of large arginine–vasopressin (AVP)- and acetylcholinesterase (AChE)-positive cells and small calbindin (Cb)-positive neurons. Rostrally, the PaM is succeeded by the small anterior parvicellular subnucleus (PaAP), which contains small AChE-, AVP-, and TH-positive cells. Dorsal to the PaM is the dorsal subnucleus (PaD), containing large spindle-shaped TH-, oxytocin (OXY)-, and AChE-positive cells as well as a population of Cb-positive neurons. Abutting the wall of the third ventricle and medial to PaD and PaM is the parvicellular subnucleus PaP. The PaP contains small cells immunoreactive for corticotropin-releasing factor (CRF), NK3, and nonphosphorylated neurofilament protein SM132. The posterior subnucleus (PaPo) is situated posterior to the descending column of fornix; it replaces all previously mentioned subdivisions caudally and is a chemoarchitectonic amalgam that includes dispersed large AChE-, TH-, OXY-, and AVP-positive cells as well as small NK3-, CRF-, SMI32-, and Cb-immunoreactive neurons. The distinctions between Pa subnuclei were further validated by the recent demonstration of different developmental patterns for the subnuclei during fetal gestation (Fig. 8). It is appropriate to mention that the latter study also relied on chemoarchitecture. These findings indicated the homology between the human PaM and PaD and the magnocellular subnuclei of the rat Pa and also between the human PaP and PaPo and the medial parvicellular and posterior subnuclei of the rat.

Apart from the already mentioned CRF, several other releasing hormones have been localized in the human hypothalamus. Most of these molecules are characterized by limited distribution, which benefits the accuracy of identification of the corresponding neuronal structure. Thus, neurons containing growth hormone-releasing hormone (GRH) in the human brain are found exclusively in the arcuate nucleus of

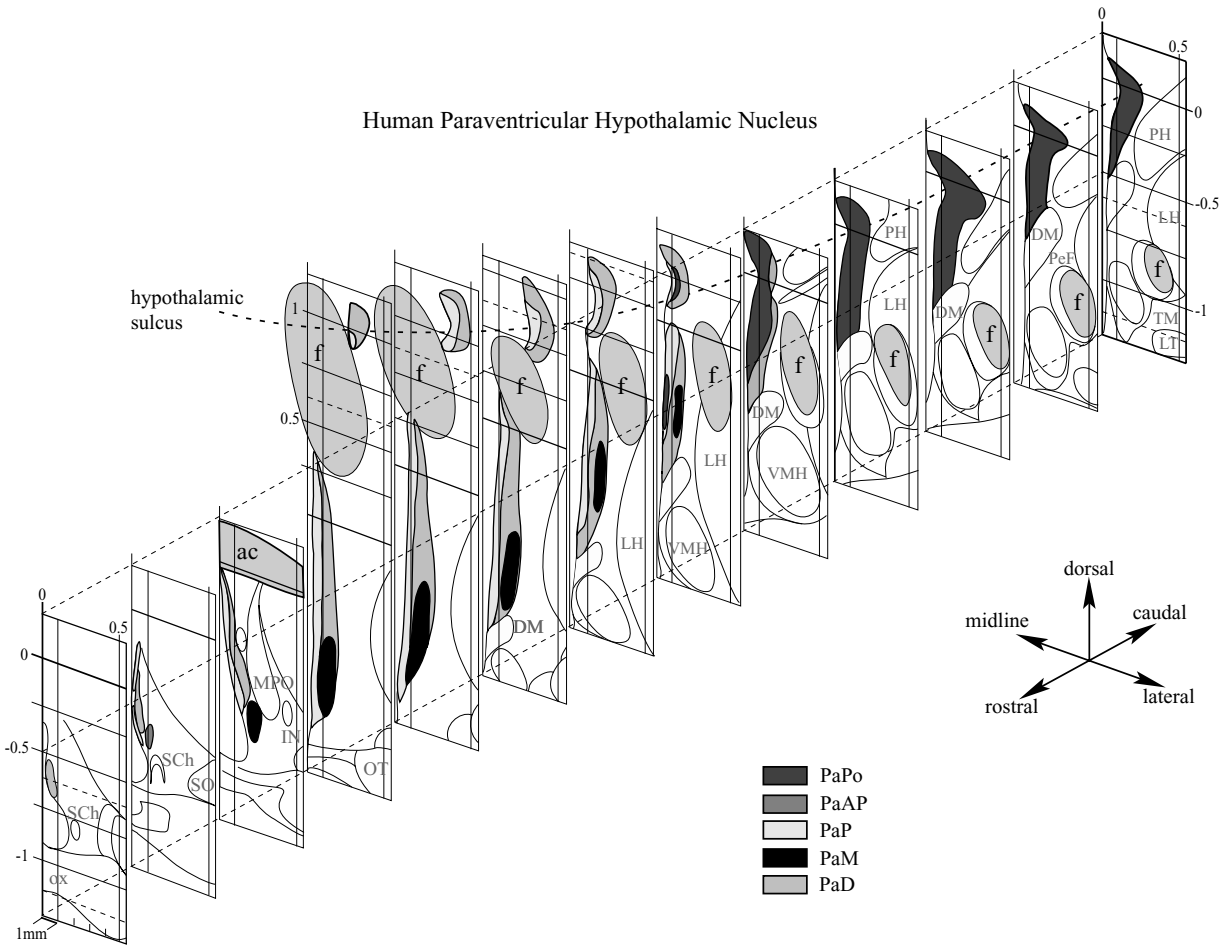


Figure 7 A model of the human Pa showing the subnuclei delineations based on computer-generated plots of six histochemical markers (AChE, SMI32, Cb, TH, NK3, and CRF) and verified with the distribution of AVP, OXY, and NPH. The third ventricle wall lies to the right. (Reproduced from *J. Comp. Neurol.* **423** with permission from John Wiley & Sons, Inc., 2000).

the hypothalamus. Some GRH-immunoreactive fibers are also found dorsally in the periventricular area, dorsomedial hypothalamic nucleus, and on the boundary of the ventromedial hypothalamic nucleus. Luteinizing hormone-releasing hormone (LHRH) has been found in the human hypothalamus within distinct neuronal groups, including the arcuate nucleus, the periventricular nucleus, the medial preoptic area, and the premammillary area.

Neurons of arcuate and periventricular hypothalamic nuclei also contain somatostatin, whereas fibers containing somatostatin are also found extending into the median eminence. Somatostatin is a peptide that inhibits the release of growth hormone (somatotropin) from the anterior pituitary. As such, somatostatin is a release-inhibiting factor. The distribution of somatostatin-immunoreactive neurons extends beyond the hypothalamus, with somatostatin-containing neurons found in the diagonal band of Broca, medial septal nuclei, nucleus basalis of Meynert, striatum, bed nucleus of stria terminalis, amygdala, and as far caudal as the periaqueductal gray and brain stem reticular formation. Concurrent autoradiographic studies revealed somatostatin binding sites in the human brain stem tegmentum, basal forebrain, and striatum. A decade of research on the distribution of the hormone-releasing factors established the location of these distinct chemically coded networks and contributed to neuroanatomical delineation of the human brain.

Recent chemoarchitectonic studies demonstrated the presence of an important feeding regulatory molecule orexin (hypocretin) in the lateral hypothalamic neurons of humans, rats, and mice. The report also distinguished between two populations of neurons

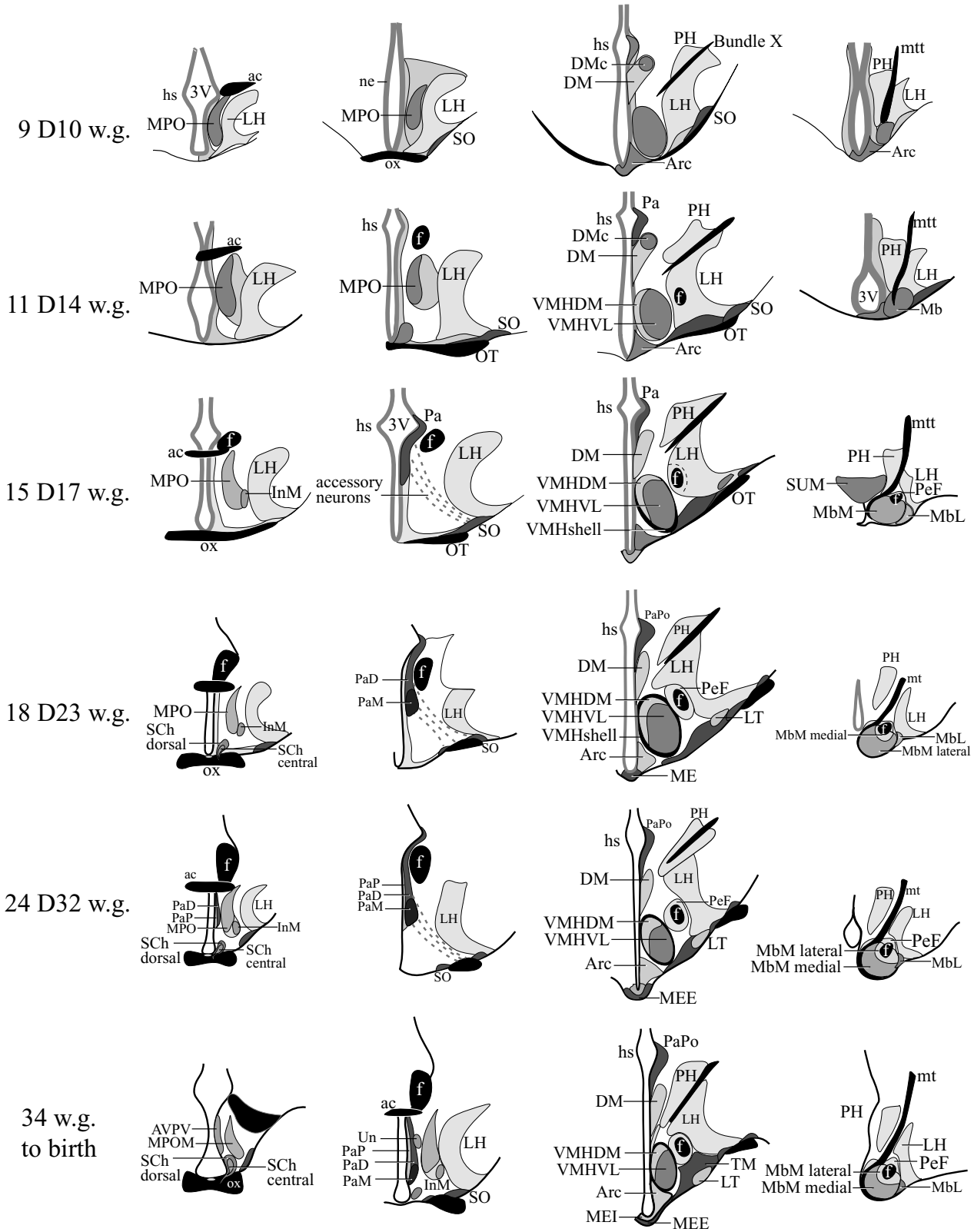


Figure 8 Diagram showing the organization of major cell groups in the developing human hypothalamus depicted at landmark stages of fetal differentiation. Grayscale represents hypothalamic structural entities revealed by cytoarchitecture of the neuroepithelial primordia and transient chemoarchitectonic labeling. Note that these diagrams are not to scale. (Reproduced from *J. Comp. Neurol.* with permission from John Wiley & Sons, Inc., 2002).

in the lateral hypothalamus, one containing orexin- and another expressing melanin-concentrating hormone. Functional evidence as well as the strong relationship of orexin and melanin-concentrating hormone containing cells in the lateral hypothalamus to both the NPY and agouti gene-related protein systems suggest an important role for these neuronal groups among hypothalamic networks regulating feeding. These findings reinforce the importance of comprehensive use of chemoarchitecture in comparative studies of the adult human brain.

Chemoarchitecture is also useful in the study of the fetal brain. Thus, several chemoarchitectonic studies of the human fetal hypothalamus reported on the development of the hypophyseal portal plasma system, CRF-containing circuitry, GRH neurons, and somatostatin-, oxytocin-, vasoperessin-, and neurophysin-containing cells. Another recent study has proved the benefits of using a combination of peptide, receptors, and structural and calcium-binding protein markers on a series of fetal brains at different stages of development to provide insight into the structural development of the human hypothalamus (Fig. 8). Substances that revealed some aspects of synaptogenesis, such as GAP43 and SYN, and an antibody directed against a cell surface membrane glycoconjugate, 3-fucosyl-*N*-acetyl-lactosamine (FAL or CD15) were also found to be useful in the delineation of developing hypothalamic cell groups.

In the fetus, the structural differentiation of the lateral and posterior hypothalamus is apparent at 9 weeks of gestation, when these structures are marked by strong immunoreactivity for GAP43, a nerve terminal membrane phosphoprotein associated with the development and restructuring of axonal connections. This suggests that within the hypothalamus, synaptogenesis is well under way in the lateral and posterior hypothalamic areas as early as 10 weeks of gestation. In addition, GAP43 immunoreactivity clearly reveals, at 13 weeks of gestation, the otherwise ill-defined bundle x, which separates LH ventrally from posterior hypothalamic area (PH) dorsomedially. Following the emergence of GAP43, another significant event in the differentiation of the lateral and posterior hypothalamic areas is the appearance of large Cr- and Cb-immunoreactive neurons at 13 and 16 weeks of gestation, respectively, and their persistence into the postnatal period. The mature morphological appearance of these neurons is reached by approximately 16 weeks of gestation, when they spread far laterally, abutting the *pallidum* and dorsally mingling with the zona incerta. Observations from the human fetus are consistent with reports on Cb immunoreactivity in the developing rat hypothalamus suggesting similarities between the area in the rat and human. In human fetuses chemoarchitecture also reveals distinct differences between the constituent structures of the *lateral hypothalamic zone*. For example, cells in PH, but not LH, show FAL immunoreactivity from 18 weeks of gestation until birth. Furthermore, the LH area contains Cr- and Cb-positive neurons and GAP43 immunoreactivity early in fetal gestation, unlike the adjacent tuberomammillary nucleus, which acquires Cb-positive neurons only late in gestation.

The distinction between the perifornical hypothalamic nucleus and lateral hypothalamic area in the rat has been demonstrated by chemoarchitectonic and functional studies. Until recently, the adult human perifornical nucleus was mainly defined by topographic criteria because cytoarchitectonically it largely resembles surrounding structures. According to a recent chemoarchitectonic study in the human fetus, perifornical hypothalamic nucleus (PeF) is formed as a result of passive displacement of LH neurons medially. The studies' conclusions were based on differentiation patterns of calbindin-, calretinin-, and neuromedin K receptor-immunoreactive neurons. Thus, during gestation, the neurons of the lateral hypothalamus, which develop early, are progressively displaced laterally by the successive waves of neurons of the *midline* and *core* zones, which develop later. In contrast to LH neurons, the neurons of the PeF that originate from the lateral hypothalamic zone remain anchored in the perifornical position, possibly by virtue of their dendrites invading the fornical bundle.

VII. CORTEX

The cortex appears curiously homogeneous given the extensive variety of its functionally diverse areas. Mapping the cortex and distinguishing between the functionally different areas has always been a challenge because of the large size and convolution of the cortical mantle and the subtlety of cyto- and myeloarchitectonic differences between areas. Chemoarchitecture, on the other hand, is a reliable tool that can provide distinct characteristics distinguishing one cortical area from another on the basis of neuroactive content, which in itself is functional evidence. It must be acknowledged, however, that chemoarchitectonic boundaries between areas of the cortex are far less obvious than those in the subcortex.

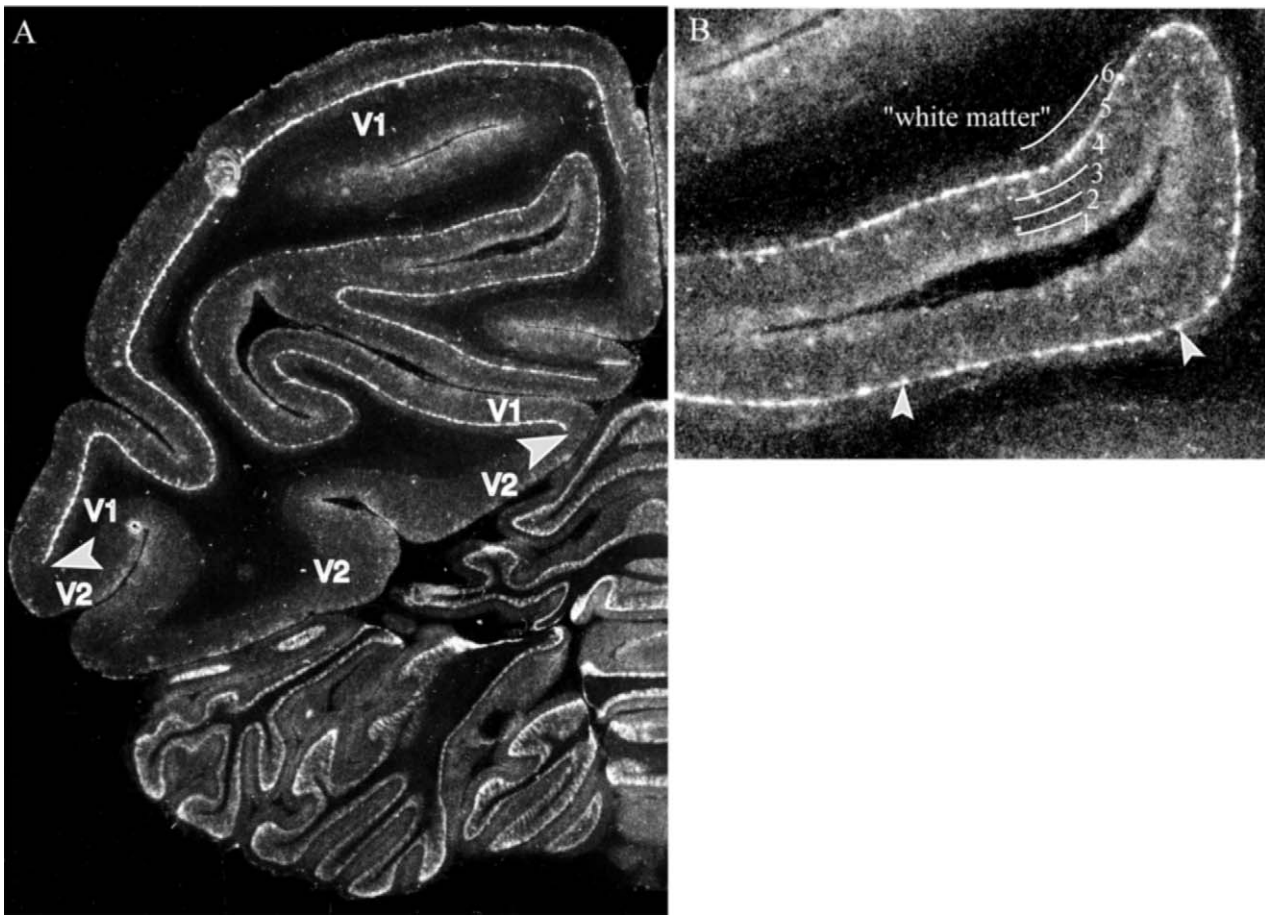


Figure 9 Photograph of an autoradiograph of a coronal section through the visual cortex showing the distribution of ^{125}I-salmon calcitonin binding. Large arrows indicate precise boundaries between V1 and V2. (B) A segment of the autoradiograph at higher magnification demonstrates 200-μm periodicity of binding (small arrows) that may be related to ocular dominance columns or to cytochrome oxidase blobs, which are characteristic of V1. Note how the binding is confined to lamina 5 of the V1.

Therefore, in anatomical studies of the cortex chemoarchitecture is used as an aid to cytoarchitectonic guidelines rather than as a primary delineation criterion. Chemical differences, however subtle, between neuronal populations in different cortical areas and cortical layers are never the less pivotal criteria for understanding the structural organization of the human cortex.

Chemoarchitectonic boundaries often correspond to the boundaries of cortical areas as identified by functional studies. For example, the boundaries of the primary visual cortex (area V1) are apparent by the prominent acetylcholinesterase reactivity that abruptly stops on the border between V1 and V2. Such chemical characteristics bring credibility to the otherwise ambiguous boundaries between these cortical areas. Apart from differentiating areas of the cortex, chemoarchitecture also distinguishes between neuronal populations in different cortical layers. For example, calcitonin receptor binding is abundant and exclusive for V1, but within that area it is confined to neurons of layer 5, which further characterizes these cells (Fig. 9).

As in the subcortex, the best results for structural differentiation of neuronal groups in the cortex are achieved through combining different chemical markers. Strong SMI32 immunoreactivity reveals a distinct population of large pyramidal cells in lower layer III and in layer V, calbindin immunoreactivity distinguishes pyramidal neurons in layers II and III, and immunoreactivity for the glutamate synthesizing enzyme glutaminase is characteristic of pyramidal cells in layers II, III, V, and VI. The combined use of histochemical markers for NADPH-d and the immunohistochemical technique for visualizing NPY revealed two populations of neurons that contain both

markers. This finding prompted the distinction of these groups of nonpyramidal, peptidergic neurons in layers V and VI of the human cerebral cortex.

There are reports of several populations of nonpyramidal neurons containing peptides and characterized by specific morphology and laminar distribution. Chemoarchitectonic characteristics are of assistance in classifying these neuronal groups. For example, immunoreactivity for TH characterizes a population of spindle-shaped neurons found to be numerous in the infragranular layers of the cortex. Cholecystokinin-immunoreactive neurons characterize a population of bipolar cells in the subgranular layers, whereas coexpression of NPY and substance P characterizes another prominent population of multipolar neurons in the deep layers of cortex and in the white matter. It should be reiterated, however, that although chemoarchitecture is informative for differentiation of some areas of the cortex, the difference is generally less striking than in the subcortex.

VIII. CONCLUSION

Chemical neuroanatomy offers a wide range of techniques for identification and classification of neuronal groups in the brain. Chemical markers are a meaningful criterion for characterizing and mapping neuronal groups in the human brain, and chemoarchitecture can be used as a *Rosetta stone* to establish homologs of subnuclei and chemically specific circuitries in the human brain corresponding to those of the rat.

Chemical neuroanatomy is an exciting and quickly developing field of neuroscience. Particularly useful for the study of the human brain and constantly stimulated by new techniques, chemical neuroanatomy continues to evolve both in significance and in sophistication.

See Also the Following Articles

BRAIN ANATOMY AND NETWORKS • HYPOTHALAMUS • NEUROANATOMY • NEUROTRANSMITTERS • PEPTIDES, HORMONES, AND THE BRAIN AND SPINAL CORD • PSYCHONEUROENDOCRINOLOGY • STRESS: HORMONAL AND NEURAL ASPECTS

Suggested Reading

Bouras, C., Magistratti, P. J., Morrison, J. H., and Constantinidis, J. (1987). An immunohistochemical study of pro-somatostatin-derived peptides in the human brain. *Neuroscience* **22,** 781–800.

Brockhaus, H. (1942). Beitrag zur normalen Anatomie des Hypothalamus und der Zona incerta beim Menschen. *J. Psychol. Neurol.* **51,** 1–51.

Halliday, G. M., Li, Y. W., Joh, T. H., Cotton, R. G. H., Howe, P. R. C., Geffen, L. B., and Blessing, W. W. (1988). Distribution of monamine-synthesizing neurons in the human medulla oblongata. *J. Com. Neurol.* **273,** 301–317.

Hornung, J. P., De Tribolet, N., and Törk, I. (1992). Morphology and distribution of neuropeptide-containing neurons in human cerebral cortex. *Neuroscience* **51,** 363–375.

Huang, X.-F., Paxinos, G., Halasz, P., McRitchie, D., and Törk, I. (1993). Substance P—A tyrosine hydroxylase-containing neurons in the human dorsal motor nucleus of the vagus nerve. *J. Comp. Neurol.* **335,** 109–122.

Koutcherov, Y., Ashwell, K. W. A., Mai, J. K., and Paxinos, G. (2000a). Organization of the human paraventricular hypothalamic nucleus. *J. Comp. Neurol.* **423**(2), 299–318.

Koutcherov, Y., Ashwell, K. W. A., and Paxinos, G. (2000b). The distribution of the neurokinin B receptor in the human and rat hypothalamus. *NeuroReport* **11**(14), 3127–3131.

Mendelsohn, F. A. O., and Paxinos, G. (1991). *Receptors in the Human Nervous System*. Academic Press, San Diego.

Paxinos, G., and Huang, X.-F. (1995). *Atlas of the Human Brainstem*. Academic Press, San Diego.

Paxinos, G., and Watson, C. (1986). *The Rat Brain in Stereotaxic Coordinates*. Academic Press, San Diego.

Paxinos, G., Huang, X.-F., and Toga, A. W. (2000). *The Rhesus Monkey Brain in Stereotaxic Coordinates*. Academic Press, San Diego.

Saper, C. B. (1990). Hypothalamus. In *The Human Nervous System* (G. Paxinos, Ed.), pp. 389–413. Academic Press, Sydney.

Spencer, S., Saper, C. B., Joh, T., Reis, D. J., Goldstein, M., and Raese, J. D. (1985). The distribution of catecholamine-containing neurons in the normal human hypothalamus. *Brain Res.* **328,** 73–80.

Tracey, D. J., Paxinos, G., and Stone, J. (1995). *Neurotransmitters in the Human Brain*. Premium Press, New York.

Ziles, K. (1990). Cortex, Hypothalamus. In *The Human Nervous System* (G. Paxinos, Ed.), pp. 757–803. Academic Press, Sydney.

Cingulate Cortex

MICHAEL GABRIEL, LAUREN BURHANS, ANDREW TALK, and PAIGE SCALF
University of Illinois

GLOSSARY

associative attention Attention focused on stimuli that predict the occurrence of reinforcement (reward or pain) and call for action.

brain electrical source analysis A dipole localization algorithm used to model the intracranial sources of electrical fields that generate scalp event-related potentials.

discriminative training-induced neuronal activity Significantly greater neuronal firing frequencies in response to a positive conditional stimulus than to a negative conditional stimulus. Discriminative training-induced neuronal activity occurs in many brain areas as early as 15 msec after conditioned stimulus onset.

duration coding Greater neuronal firing in response to a brief than to a more enduring positive conditional stimulus.

error-related negativity Electrical negativity found in event-related potential recordings from mid-frontal regions of the scalp. The peak of the error-related negativity occurs approximately 100 msec after the onset of an erroneous response.

executive attention A process guided by an individual's plans and goals, which comes into play when routine or automatic processes are insufficient for the task at hand, such as when novel or conflict-laden situations are encountered.

positive conditional stimulus A stimulus employed in studies of classical and instrumental conditioning. The occurrence of a positive conditional stimulus predicts reinforcement (reward or pain). A negative conditional stimulus predicts that no reinforcement will occur.

salience compensation An attentional process of cingulate cortex that amplifies the neural representation of nonsalient but associatively significant stimuli in order to ensure that such stimuli receive processing that is commensurate with their associative significance. Duration coding is an example of salience compensation.

Early and highly influential accounts of cingulate cortical function were presented by J. W. Papez and Paul MacLean, who argued, in essence, that activity in neural circuitry involving cingulate cortex and related structures of the brain's limbic system is the neural substrate of emotional experience. Recent studies have implicated cingulate cortex in pain perception and in the cognitive processes of attention, response selection, learning, and memory. Morever, changes in the cingulate cortex have been found relevant to the etiology of schizophrenia and Alzheimer's disease. This article emphasizes developments concerning the behavioral and cognitive functions of cingulate cortex. A particular emphasis of this article is the integration of many recent findings from laboratories of behavioral neuroscientists concerning neuronal firing patterns and effects of experimental brain lesions on behavior in animals with a large body of recent data from laboratories of cognitive neuroscientists concerning experimental analyses of human cognition using neuroimaging methods such as positron emission tomograpy, functional magnetic resonance imaging, and high-density electroencehpalographic recording. Before addressing these issues, however, some basic facts and definitions are reviewed.

I. CINGULATE CORTEX: WHAT AND WHERE?

Cingulate cortex is located in the medial walls of the cerebral hemispheres and is a subset of "limbic

cortex." Limbic cortex includes all areas that receive fibers of anterior thalamic neurons. By modern convention, the subset of limbic cortex constituting cingulate cortex includes Brodman's areas 24 and 29 in small animals such as rabbits and rats. These and an additional area (23) constitute cingulate cortex in primates. A cytoarchitectural map of the primate limbic cortex is shown in Fig. 1.

Brodman's areas 24 and 29 are often referred to respectively as anterior and posterior cingulate cortex. Neurons in both areas receive afferent fibers from the anterior medial (AM) thalamic nucleus. However, the anterior cingulate cortex (area 24) is also innervated by projections of the medial dorsal (MD) and parafascicular thalamic nuclei. Neurons in the posterior cingulate cortex (area 29) receive projections from all members of the anterior group of nuclei, the anterior ventral (AV), anterior dorsal, and the AM nuclei, as well as the ventral anterior and lateral dorsal thalamic nuclei. Neurons in midline and intralaminar thalamic nuclei send axons to both the anterior and the posterior cingulate cortex.

Cingulate cortical neurons are robustly responsive to multimodal (auditory, visual, somatic sensory, and visceral) stimuli, and they are richly innervated by fibers of the pons and midbrain that distribute the biogenic amines (dopamine, norepinephrine, and serotonin) as well as acetylcholine. Whereas norepinephrine, serotonin, and acetylcholine fibers fairly uniformly innervate both the anterior and the posterior cingulate cortex, only anterior cingulate cortex receives appreciable amounts of dopamine. Many additional afferent systems course to the cingulate cortex, including fibers from visual cortex, hippocampus, subiculum, entorhinal cortex, and amygdala. Cingulate cortical neurons send efferent fibers to most of the aforementioned thalamic areas, the subiculum, entorhinal cortex, and pons and to many areas of the striatal motor system, including the caudate nucleus, nucleus accumbens, and zona incerta. Moreover, in primates, cingulate cortical neurons project to multiple areas of the motor and premotor cortex. Thus, numerous parallel pathways exist whereby cingulate neurons can modulate motor output systems of the brain. Finally, the work of Patricia Goldman-Rakic and colleagues has demonstrated direct reciprocal projections of cingulate cortical neurons in primates to the lateral prefrontal and parietal cortex, areas involved in high-level perceptual and mnemonic functions.

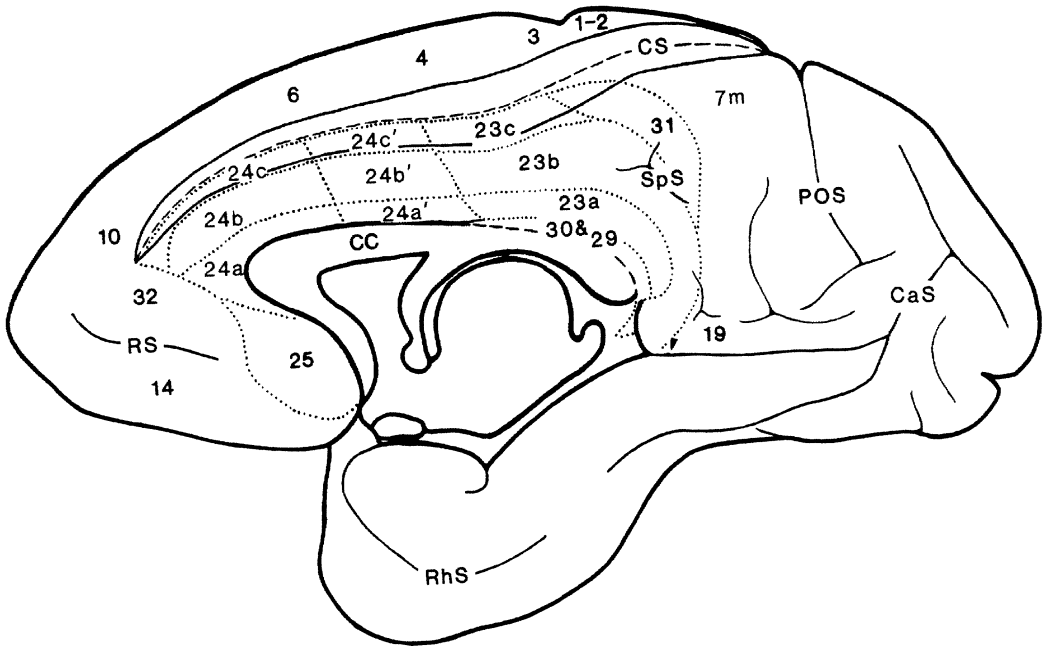


Figure 1 Cytoarchitectural map of the limbic cortex in the rhesus monkey based on Brodman's divisions.

II. FUNCTIONS OF THE CINGULATE CORTEX

A. Associative Attention

As students of cognition are aware, the term attention has multiple interpretations. Research in the area of behavioral neuroscience with rabbits and rats indicates that the attention supported by cingulate cortex is selective attention or attention focused on particular stimuli. The stimuli that are selectively processed by cingulate cortical neurons are those that predict important outcomes, such as reward or aversion. Also, in most if not all cases they are "task-relevant" stimuli (i.e., stimuli that require a particular action).

The selective processing of stimuli by cingulate cortical neurons emerges as the subjects learn that a given stimulus predicts an important outcome and requires action. Thus, the attentional process of cingulate cortex that is invoked by such predictive stimuli is a learned or "associative" process of the brain. Stimuli that signal important outcomes and call for action produce large neuronal activations in cingulate cortex, whereas stimuli that predict that no important event will occur and that no action is needed do not elicit ample cingulate cortical neuronal discharges. An apt characterization is that cingulate cortex mediates associative attention to significant stimuli. Findings in support of these propositions are reviewed next.

1. Cingulate Cortical Neuronal Activity

a. Discriminative Neuronal Activity Data supporting the hypothesis that cingulate cortex mediates associative attention to significant stimuli are afforded by studies of neuronal activity in multiple brain sites during discriminative instrumental learning of rabbits. In these studies the rabbits occupied a large rotating wheel apparatus. They learned to step in response to a tone cue, a positive conditional stimulus (CS^+), in order to prevent a foot-shock scheduled for delivery 5 sec after the CS^+. They also learned to ignore a different tone, the negative conditional stimulus (CS^-), that was not followed by the foot-shock.

Recordings of neuronal activity in multiple areas of cingulate cortex and in related areas of the thalamus exhibited the development, during behavioral learning, of massive discriminative multi-, and single-unit neuronal responses. Discriminative neuronal responses are significantly greater firing frequencies in response to the CS^+ than to the CS^-. They occur at brief latencies (15–100 msec) and persist for varied durations, with some neurons exhibiting discrimination during the full interval preceding response or shock (Fig. 2).

The discriminative neuronal activity represents an associative or learned activity of cingulate cortical neurons. It is not simply a frequency-specific response of the monitored neurons because the discriminative response is observed in many subjects given different (counterbalanced) sets of acoustic frequencies as CS^+ and CS^-. Discriminative activity develops as a result of the associative pairing of the CS^+ with the foot-shock reinforcer. A different neuronal response (either no response or one of significantly reduced magnitude) occurs to the CS^-, which is not predictive of the foot-shock. Thus, the activity of cingulate cortical neurons in these studies may be viewed as a neuronal code for the associative significance of the discriminative cues. It can be said that cingulate cortical neurons "learn" to respond selectively to stimuli that predict the occurrence of significant events and require action, which in these studies is the locomotor response needed to avoid the foot-shock.

Discriminative activity also occurs during learning of a reward-based response, wherein rabbits approach and make oral contact with a drinking spout inserted into the experimental chamber following CS^+ presentation. The rabbits also learn to inhibit responding when the spout is inserted following CS^- presentation and oral contact with the spout does not yield water reward. Once again, neurons in cingulate cortex encode the associative significance of the conditional stimuli, i.e., they produce greater discharges in response to the CS^+ than to the CS^-. These data rule out the possibility that cingulate cortical neurons encode the emotional value (pleasant or unpleasant) of the reinforcer predicted by the CS^+, i.e., differential firing to stimuli that predict pleasant versus unpleasant consequences. Rather, cingulate cortical neurons encode stimuli that predict important events and call for action, whether the events are good or bad.

Discriminative neuronal activity in cingulate cortex has also been reported by Shirley Buchanan, Donald Powell, and Charles Gibbs during classical Pavlovian conditioning of heart rate and eyeblink responses in rabbits. Also, early studies of Menachem Segal and James Olds and recent studies of Takenouchi and colleagues demonstrated the occurrence of neuronal responses in anterior and posterior cingulate cortex that are specific to stimuli that predict reinforcement during appetitive conditioning.

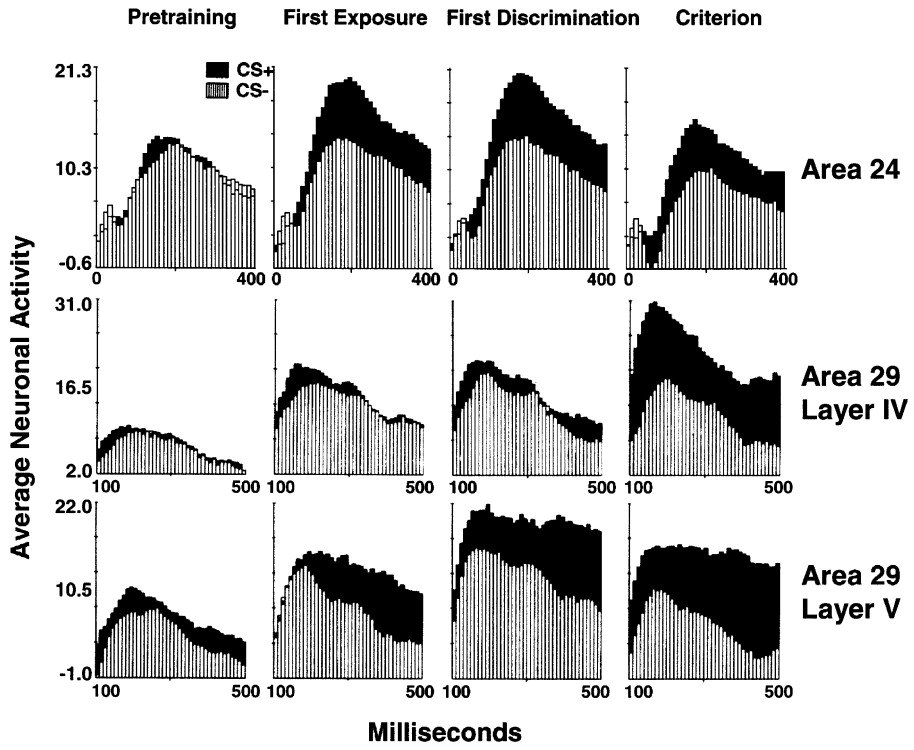


Figure 2 Average anterior (area 24) and posterior (area 29, cellular layers IV and V) cingulate cortical integrated unit activity elicited by CS^+ and CS^- in rabbits during pretraining, first exposure session, first significant behavioral discrimination, and criterial behavioral discrimination in a discriminative avoidance task. The neuronal activity for area 24 is plotted in the form of standard scores normalized with respect to the pre-CS baseline for 40 consecutive intervals following CS onset. Area 29 data are plotted starting 100 msec after tone onset.

The development of discriminative neuronal activity during conditioning supports the idea that cingulate cortex mediates associative attention to significant cues. The fact that the activity is time-locked to the onset of the CS^+ and CS^- and occurs as early as 15 msec after tone onset, at least 2 sec earlier than the learned avoidance response, suggests that it promotes learning-based stimulus selection. The early, stimulus-locked, CS^+-specific brain activation could be involved in capturing attention and thereby enhancing stimulus evaluation and response mobilization processes.

b. Duration Coding and Salience Compensation The brief-latency discriminative neuronal responses are positioned temporally to maximize resources for subsequent processing of and responding to the eliciting cue. Accordingly, Stephen Sparenborg and Michael Gabriel showed that testing with non-salient cues, such as a brief (200-msec) CS^+ and CS^-, is associated with greater brief-latency anterior and posterior cingulate cortical discriminative neuronal responses than are observed when testing is carried out with more enduring and therefore more salient cues (e.g., a 500- or 5000-msec CS^+ and CS^-; Fig. 3). The enhanced neuronal coding of brief stimuli is referred to as duration coding. Duration coding may be thought of as an instance of salience compensation, i.e., a process of cingulate cortex that amplifies the neural representation of nonsalient but associatively significant stimuli in order to increase the likelihood that such stimuli receive processing that is commensurate with their associative significance.

c. Premotor Activity As discussed later, neuroanatomical and neurophysiological data indicate a close association between cingulate cortex and the brain's motor system. One illustration of this involvement in motoric processing is the finding that approximately half of all single neurons studied in anterior and posterior cingulate cortex during discriminative avoidance learning exhibited premotor firing ramps, i.e., progressive increments of firing frequency during the 5-sec interval from CS^+ to the scheduled foot-shock,

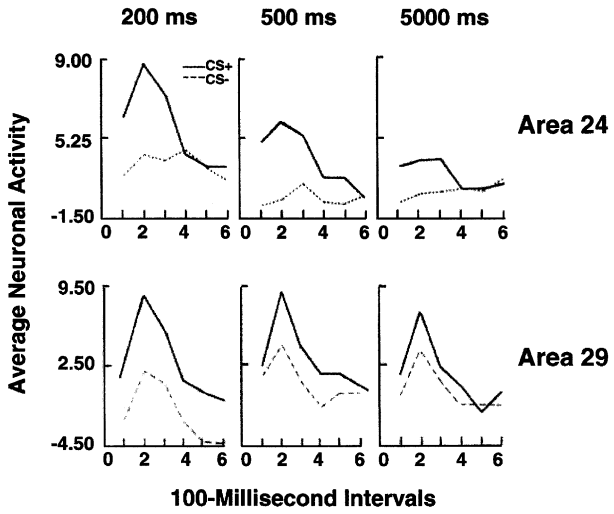


Figure 3 Average multi-unit spike frequency recorded in well-trained rabbits in anterior cingulate (area 24) and posterior cingulate (area 29) during separate counterbalanced sets of three training sessions in which brief (200-msec), intermediate (500-msec), or long (5000-msec) CSs were presented. The neuronal activity following CS onset is in the form of standard scores normalized with respect to the pre-CS baseline in six consecutive 100-msec intervals after CS onset. Data shown for area 29 were obtained from records that exhibited discriminative TIA in the later stages of learning.

in anticipation of the behavioral avoidance response (Fig. 4). Also, in recent studies by Takenouchi and colleagues, neuronal firing in ventral portions of the anterior cingulate cortex was correlated with the onset of consummatory behavior (licking a drinking tube) during appetitive conditioning of rats. This premotor firing of cingulate cortical neurons may represent a neural "command volley" projected from cingulate cortex to cortical and striatal motor areas to trigger the learned response. Thus, in addition to the encoding of significant stimuli, cingulate cortical neurons appear to be involved in the initiation or triggering of learned motor responses.

2. Brain Damage

a. Experimental Lesions If cingulate cortical neurons mediate associative attention to significant stimuli, then damage in cingulate cortex should interfere with learning about such stimuli. This expectation was confirmed by studies that showed that combined lesions of both the anterior and the posterior cingulate cortex severely impaired the ability of rabbits to exhibit discriminative avoidance learning. An equivalent impairment of learning was found following bilateral combined lesions of the anterior and MD thalamic nuclei (i.e., areas of the thalamus that reciprocate connections with cingulate cortex). The interdependence of the thalamic and cingulate cortical processes was indicated by studies that showed that the training-induced, attention-related activity of cingulate cortical neurons did not develop in subjects that had been given thalamic lesions.

The fact that the lesions prevented the rabbits from learning to respond behaviorally to the CS^+ could be interpreted as a motor problem such as an inability to initiate locomotion on cue. That this was not the case was indicated in a study that showed that rabbits with lesions were substantially impaired in learning to inhibit the previously described reward-based approach response to a drinking spout that was inserted into the experimental chamber following CS^- presentation (when approach and oral contact with the spout did not yield water reward). The rabbits with lesions responded equally often to the spout, whether its insertions were preceded by the CS^+ or the CS^-, whereas sham lesion controls discriminated significantly between the CS^+ and CS^-. These results show that cingulothalamic circuitry is critical for rabbits' ability to base behavioral responding on particular discriminative cues (as in the case of discriminative avoidance learning) or learning to inhibit a well-established response on cue (as in the approach learning task). These findings are in keeping with the notion that the cingulothalamic circuitry is not specialized for particular kinds of behavioral outputs, such as active movement or "inhibitory" omission of movement. Rather, this circuitry enables subjects to predicate their performance of context-appropriate, goal-directed instrumental behavior on the occurrence of discrete cues that predict significant outcomes and call for action. A wide variety of distinct forms of behavior can come under stimulus control as a result of cingulate cortical processing. This stimulus-oriented contribution of cingulate cortex is consistent with the hypothesis of an involvement of cingulothalamic circuitry in associative attention to significant stimuli.

b. Exposure to Cocaine *in Utero* Exposure of human fetuses to cocaine during gestation is associated with a variety of developmental, neurocognitive deficits, including impaired attention, habituation, arousal, recognition memory, and language development. Adult rabbits exposed to cocaine *in utero* (4 mg/kg of cocaine given intravenously twice daily to gestating dams) exhibited morphological and

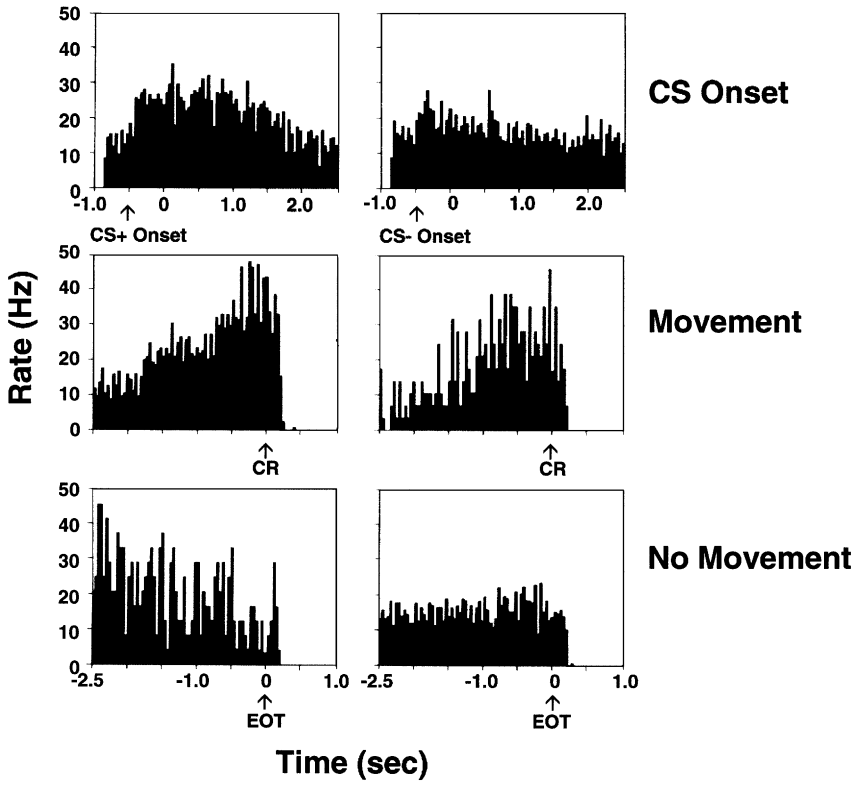


Figure 4 Histograms indicating anterior cingulate cortical single-unit activity related to CS^+/CS^- onset and avoidance responses, where each bar indicates the average firing rate in hertz for the cell during a 40-msec interval. (Top) CS onset-related activity; (middle) premotor discharges preceding the conditioned response (CR); (bottom) neuronal firing on CS^+ and CS^- trials in which no response occurred and the trial terminated (EOT). All histograms represent data obtained from the same neuron.

biochemical abnormalities in the anterior cingulate cortex relative to controls exposed to saline injections. No changes were found in the visual cortices of the subjects exposed to cocaine.

In addition, exposure to cocaine was associated with attenuated anterior cingulate cortical training-induced discriminative neuronal activity and deficient avoidance learning. Specifically, when brief (200-msec) and therefore nonsalient discriminative stimuli (CS^+ and CS^-) were employed for training, cocaine-exposed rabbits performed significantly fewer learned responses than saline-exposed controls in the first session of conditioning. The first-session learning deficit found with brief stimuli could have resulted from compromised salience compensation mechanisms in anterior cingulate cortices of the rabbits exposed to cocaine. Note, however, that with continued training beyond the first session the cocaine-exposed rabbits did attain normal asymptotic performance levels as rapidly as did saline controls. Moreover, learning was entirely normative, even in the first session, in cocaine-exposed subjects when the CS^+ was more enduring (500 msec) and thus more salient.

Very similar results were obtained in studies of Pavlovian conditioning of rabbits' eyeblink response. Rabbits exposed to cocaine *in utero* were able to acquire the conditioned eyeblink response as rapidly as controls when a salient CS^+ and a nonsalient CS^- were used. However, acquisition was significantly retarded in cocaine-exposed rabbits when a nonsalient CS^+ and a salient CS^- were used. These results were obtained when the stimuli were of different modalities (a salient tone and a less salient flashing light) and when they were of the same modality (tones of varying intensity).

The absence of behavioral learning during the first discriminative avoidance conditioning session with the brief (200-msec) stimuli was accompanied by an

absence of discriminative neuronal activity in the anterior cingulate cortex. Moreover, when behavioral discrimination did occur in later stages of training, so also did discriminative neuronal activity in the anterior cingulate cortex in rabbits exposed to cocaine. Thus, the loss of discriminative neuronal activity in anterior cingulate cortex was, arguably, the neural basis of the impaired discriminative behavior. This possibility received further support from previous demonstrations that lesions confined to the anterior cingulate cortex or to the MD thalamic nucleus impaired learning in the early stages of behavioral acquisition, but these lesions also allowed learning to occur to normative asymptotic levels.

A particularly intriguing finding during training of cocaine-exposed rabbits with the brief conditional stimuli was a dramatic alteration of the poststimulus firing profiles of anterior cingulate cortical neurons. Specifically, the neuronal response profiles of the cocaine-exposed rabbits lacked the firing pause, which occurred robustly in controls from 40 to 80 msec after CS onset and which is a consistent feature of the CS^+- and CS^-- elicited poststimulus histograms of neurons in cingulate cortex (Fig. 5). Loss of the brief-latency inhibitory "pause" component of neuronal response was a most robust cocaine-related neuronal phenomenon. Indeed, the magnitude of the inhibitory pause was significantly increased in the saline-exposed rabbits trained with 200-msec stimuli compared to saline-exposed rabbits trained with 500-msec stimuli. These findings suggested that the inhibitory pause is a dynamic feature that reflects the associative attentional processing "demand" that is operating in a given situation. The absence of the inhibitory pause in rabbits exposed to cocaine may thus be a direct neurological indicant of impaired associative attention, which is in turn the basis for the observed retardation of the discriminative neuronal activity and behavioral learning in rabbits exposed to cocaine *in utero*.

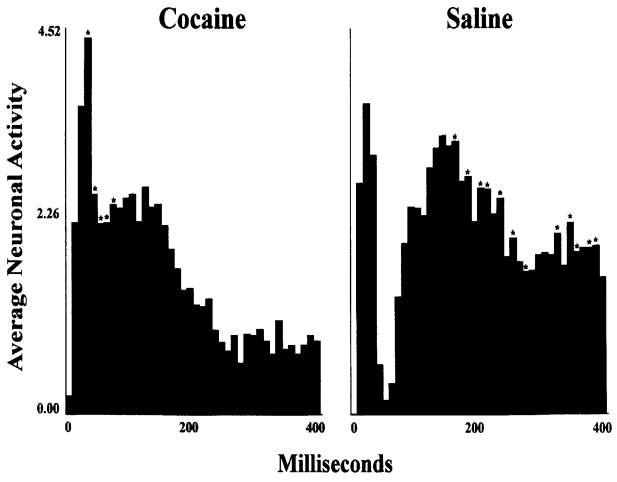


Figure 5 Average anterior cingulate cortical multi-unit spike frequency in 40 consecutive 10-msec intervals after onset of a brief (200-msec) CS in rabbits exposed to cocaine *in utero* and in saline-exposed controls. Asterisks indicate the occurrence of a significantly greater discharge for the indicated interval compared to the discharge in the corresponding interval for the other experimental group (cocaine or saline). Reproduced by permission from Harvey and Kosofsky, 1998, *Cocaine: Effects on the Developing Brain*, Ann. N.Y. Acad. Sci. **846,** 208.

The inhibitory pause in anterior cingulate cortex may function to reset active neurons by halting ongoing firing, thereby maximizing the number of neurons available for processing the incoming stimulus. Inhibitory feedback produced by activation of GABAergic neurons in response to cue-driven inputs is likely to be involved in resetting. The failure of the resetting mechanism in rabbits exposed to cocaine means that neurons already engaged in rapid firing could not contribute to stimulus processing. The resulting reduction in the number of participating neurons could impair processes such as the recruitment of existing modified synapses involved in classifying the incoming stimulus or retardation of synaptic plasticity development needed for the production of discriminative neuronal activity.

Research of Eitan Freidman and colleagues has shown that D_1 dopamine receptors in the anterior cingulate cortex are decoupled from their G proteins in rabbits exposed to cocaine *in utero*. Thus, the failure of the resetting mechanism in exposed rabbits could be a result of impaired activation of GABA neurons normally mediated by stimulation of D_1 dopamine receptors in the anterior cingulate cortex.

B. Executive Attention

New neuroimaging techniques have yielded an explosion of data in cognitive neuroscience concerning brain activation accompanying task engagement of human subjects. Studies employing position emission tomography (PET), functional magnetic resonance imaging (fMRI), high-density electroencephalography (EEG) for these analyses have repeatedly indicated an involvement of anterior cingulate cortex in cognition-relevant processing.

Results and interpretations converge intriguingly with the aforementioned findings in rabbits, rats, and primates in indicating a critical involvement of anterior cingulate cortex in processes subserving attention. Michael Posner and Gregory Di Girolamo provided an integrative account of anterior cingulate cortical function derived from the results of imaging studies and other findings in cognitive neuroscience. Building on the prior theoretical work of Donald Norman and Timothy Shallice, Posner and DiGirolamo proposed that anterior cingulate cortex mediates executive attention, which is part of a more general executive control function. Executive attention comes into play when routine or automatic processing is insufficient for the task at hand, such as when novel or conflict-laden situations are encountered. Guided by the individual's overall plans and goals, executive attention selectively activates and inhibits particular inputs, schemas, and behaviors to deal with the problematic circumstances. Paraphrasing Norman and Shallice, Posner and DiGirolamo argue that executive attention will likely be recruited in situations that involve planning and decision making, error detection, novelty and early stages of learning, difficult and threatening situations, and overcoming habitual behavior.

1. Review of Evidence

Activation of the anterior cingulate cortex as assessed by PET and fMRI is significantly enhanced in response to stimuli that require a particular response from multiple alternative responses, such as when subjects verbally generate uses of visually or acoustically presented words denoting familiar objects (e.g., the response "drive" to the stimulus "car"). In most of these studies a subtraction method is employed whereby the brain activation found during a control condition (merely reading and pronouncing the stimulus words) is subtracted from the scores obtained in the "generate uses" condition. Thus, anterior cingulate cortical activation above and beyond that involved in pronouncing visualized words is associated with the need to generate a response from a set of possible alternative responses. It is argued that executive attention reflected by anterior cingulate cortical activation is brought into play as a result of the conflict created by the multiple response alternatives in the generate uses condition. The activation produced in the generate uses condition declined as the subjects were repeatedly exposed to the same words and thus generated the same uses. According to the theory, the generation of uses became routinized with repetition and thus no longer required executive attention.

Additional support for a role of the anterior cingulate cortex in executive attention is the observed activation of cingulate cortex during performance in Stroop tasks, wherein subjects are required to name the color of visually presented words in a congruent condition (e.g., the word "red" printed in red ink) or in an incongruent condition (e.g., the word "green" printed in red ink). In some of these studies, a neutral condition is also employed in which, for example, subjects must give the ink color of noncolor words. The results of multiple Stroop experiments suggest the following generalization: The anterior cingulate cortex is activated significantly in all three of the aforementioned conditions. However, the extent of activation appears to depend on the degree to which the irrelevant dimension (word meaning) corresponds to the relevant dimension (word color). High correspondence, such as when both dimensions refer to color (as in the incongruent condition), creates maximal conflict and invokes substantial anterior cingulate cortical processing.

In keeping with the indication that the degree of conflict determines the magnitude of anterior cingulate cortical activation, several studies have shown substantial activation in association with dual-task performance, such as when subjects concurrently perform a generate-use task and a motor sequencing task. The results showed that performance in the dual-task situation elicited greater activation of anterior cingulate cortex than did performance in either task singly. Because the two tasks did not conflict at the sensory or motor levels, the dual-task-specific activation was assumed to reflect competition for central processing resources rather than sensory or motor processes. These studies also showed declining anterior cingulate cortical activation as the tasks became well learned, but ample activation was reestablished simply by instructing the subjects to attend to the well-learned tasks. These results are consistent with the hypothesis that anterior cingulate cortical processing is recruited in response to conflict and interference among central aspects of task-relevant processing.

Recent work using event-related potential recordings derived from multiple scalp locations in human subjects has yielded an intriguing result. A marked electrical negativity occurs 100 msec after an erroneous key press in a discriminated reaction time task (Fig. 6). Brain electrical source analysis of data yielded by large arrays of simultaneously recorded scalp EEG records was employed in attempts to localize the brain areas in

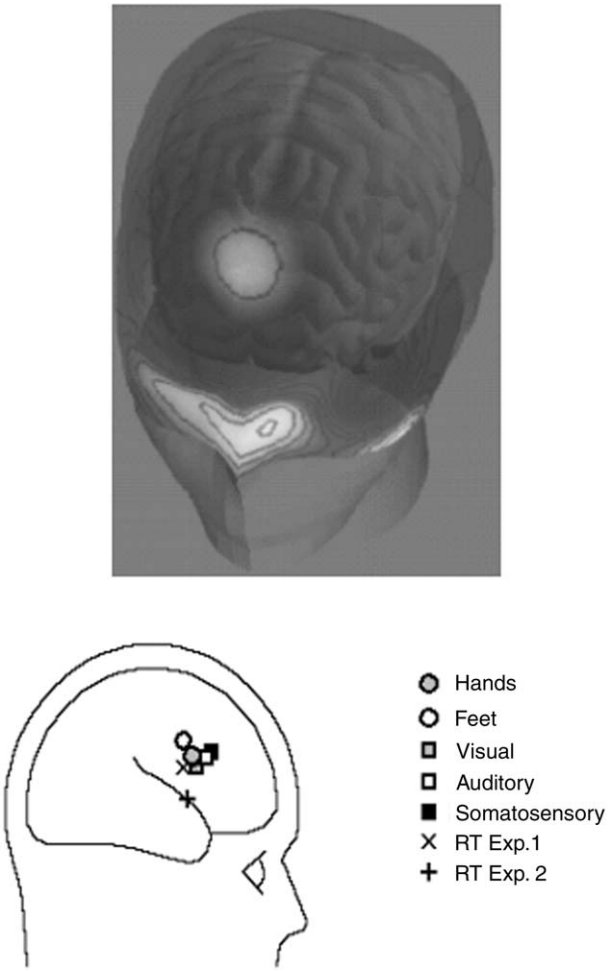


Figure 6 A three-dimensional and sagittal view of the human brain illustrating the source of error-related negativity (ERN) found after brain electric source analysis of event-related brain potentials. (Bottom) The results of several ERN studies that have demonstrated that the source of the ERN is not affected by response modality (subjects responding with their feet or hands) or error feedback modality (visual, auditory, and somatosensory). Also shown is the ERN source for two reaction time experiments, one involving a decision of whether a number was ''smaller than/larger than'' (RT Exp.1) and another involving a classification of words into semantic categories (RT Exp. 2). [Reprinted from Holroyd, C. B., Dien, J., and Coles, M. G. (1998). Error-related scalp potentials elicited by hand and foot movements: Evidence for an output-independent error-processing system in humans. *Neurosci. Lett*. **242,** 65–68, with permission from Elsevier Science].

which the error-related negativity (ERN) is generated. The results of separate studies by William Gehring and colleagues and by Stanislas Dehaene and colleagues indicated that anterior cingulate cortex is a likely intracranial source of the ERN (Fig. 6). These results are consistent with the hypothesis that anterior cingulate cortex is involved in mediating processes of executive attention that are recruited by the occurrence of errors.

2. Executive Attention and Response Selection

One problem with the concept of executive attention is that it includes multiple processes, including selection of input channels, schemas, and responses in accordance with the individual's overall plans and goals. It seems unlikely that all these processes are strictly localized within the cingulate cortex. Even the most cingulate-centered view of the universe must acknowledge the likely importance of continuous exchange of information between cingulate cortex and other brain modules that comprise a larger circuitry and supply information needed for the computation of executive functions. A clear understanding of how the components of executive attention are allocated among cingulate cortex and related areas of brain circuitry remains to be worked out. Nevertheless, a series of recent findings, reviewed in this section, indicate that response selection is an important aspect of cingulate cortical function.

First, there appears to be considerable regional specialization in cingulate cortex with respect to particular response modalities. Studies by Nathalie Picard and Peter Strick have indicated the existence of multiple areas of cingulate cortex, each containing neurons that project to the primary motor cortex in primates. Studies employing PET in human subjects indicate that some of the projecting areas may be associated with distinct response modalities. Similarly, Thomas Paus and colleagues showed that different areas of the anterior cingulate cortex were activated in the same subjects when they performed tasks involving oculomotor, manual, and spoken responses. Also, neurophysiological investigations by Keisetsu Shima and colleagues using primate subjects indicated the existence of two distinct cingulate cortical areas involved in mediating self-paced and stimulus-triggered forelimb movements.

In addition, U. Turken and Diane Swick reported studies of a patient (DL) who had undergone surgery for removal of a brain tumor. The surgery resulted in a circumscribed right hemispheric lesion in a region of the anterior cingulate cortex that had been implicated by neurophysiological studies in hand and arm movements of primate subjects. DL exhibited entirely normal performance in Stroop-like and divided attention tasks when responses were orally reported. However, DL showed a dramatic deficit in the same tasks when manual responses were required. Because

DL showed only a manual impairment, the authors argued that executive functions were intact in DL. Thus, they interpreted the results as favoring the idea that command signals originating in lateral prefrontal areas are sent to motor output areas through the anterior cingulate cortex, where correct response output is facilitated and incorrect response output is suppressed. They concluded that executive functions reside in the prefrontal areas, whereas cingulate cortex performs a "final check" on the already-selected output.

A somewhat different view is afforded by recent findings of Michael Millham, Marie Banich, and colleagues, who employed fMRI imaging and a variant of the color–word Stroop task wherein participants were required to respond selectively to information about the color of a word while disregarding the word's meaning. They sought to address the issue of whether activity in the anterior cingulate cortex is engaged in attentional selection generally or is more specifically related to response conflict. To disentangle these two possibilities, these authors asked whether cingulate cortical processing on incongruent trials is engaged by both response and nonresponse conflict or by only one of these varieties of conflict. Participants indicated the ink color of the word (yellow, green, or blue) via a keypress. Half of the incongruent trials were response eligible in that the word named one of the eligible responses (e.g., the word "blue" printed in green ink). These trials engender conflict at both the response and nonresponse levels. The other half of the incongruent trials were response ineligible in that the word did not name an eligible response (e.g., the word "purple" printed in green ink). These trials engendered conflict at the nonresponse level but not at the response level (Fig. 7).

The pattern of results indicated that the anterior cingulate cortex is specifically involved in detecting the potential for error at the response level. As found in previous studies, both dorsolateral prefrontal regions and the anterior cingulate exhibited greater activation in response to incongruent trials than in response to neutral trials. However, further analysis indicated that the enhanced activation that occurred in anterior cingulate cortex on incongruent trials occurred only on response-eligible trials, not on response-ineligible trials. (In contrast, dorsolateral prefrontal cortex showed the difference on both response-eligible and response-ineligible trials.) These results thus suggest that anterior cingulate activation is driven exclusively by response conflict, and not by conflict related to stimulus-selection processes.

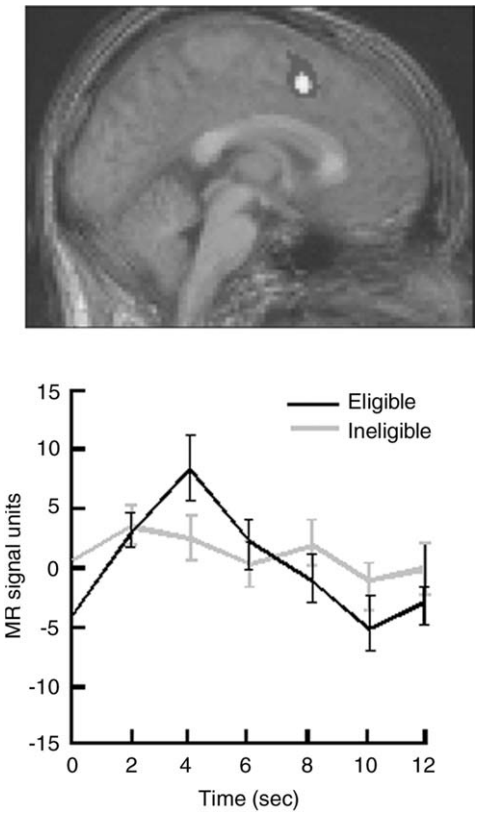


Figure 7 A saggital fMRI image and corresponding line graph depicting the increased activation of the anterior cingulate during response-eligible vs response-ineligible trials in a modified Stroop task.

III. INTEGRATION: ASSOCIATIVE ATTENTION AND EXECUTIVE ATTENTION

Executive attention and associative attention refer to two theoretical accounts of cingulate cortical function. Although they originate from research in cognitive neuroscience and behavioral neuroscience, respectively, the two theories nevertheless converge on an attribution of cingulate cortical involvement in attentional processes. However, the use of the term attention in both instances does not indicate how much convergence has occurred. Here, we attempt an explicit comparison of the two theories and the data on which they are based in an effort to highlight key similarities and differences.

A. Similarities

An important area of similarity concerns the observation that associative attention in animals appears to

become engaged in several of the very situations regarded by cognitive neuroscientists as likely to engender executive attention. These include difficult and threatening situations, the early stages of learning, and overcoming habitual behavior. Specific involvement of associative attention in the early stages of learning and in difficult and threatening situations is indicated by findings that massive and robust training-induced discriminative neuronal activity (TIA) in the anterior cingulate cortex develops in the first session of discriminative avoidance learning and declines in later sessions as discriminative behavior reaches asymptotic levels. This early acquisition of TIA does not occur during reward-based discriminative approach learning, as indicated by recent studies of John Freeman and David Smith in the laboratory of Michael Gabriel. Moreover, various manipulations that eliminated anterior cingulate cortical TIA (lesions in anterior cingulate, amygdala, or auditory cortex and exposure of subjects to cocaine *in utero*) resulted in a failure of rabbits to acquire discriminative avoidance responses in the early sessions of training. Note that despite the early learning deficit that is associated with anterior cingulate cortical lesions and with loss of anterior cingulate cortical TIA, the subjects in these studies did acquire the discriminative behavior to normal asymptotic levels, an accomplishment due at least in part to processes of posterior cingulate cortex. Finally, lesions of the amygdala that eliminated TIA in the anterior cingulate cortex blocked discriminative avoidance learning (which is acquired rapidly) but had no impact on concurrently administered and much more slowly acquired discriminative approach learning in the same subjects.

Evidence that cingulate cortical processes are invoked when animals are engaged in overcoming habitual behavior has been provided in several studies demonstrating a dramatic increase in the magnitude of discriminative TIA during the early sessions of reversal learning, wherein the cues serving as CS^+ and CS^- during original learning to asymptotic levels of discriminative behavior are interchanged. Although the hypothesis that anterior cingulate lesions will specifically impair discriminative instrumental learning of the reversal problem has not been tested, this hypothesis has been confirmed by Shirley Buchanan and Donald Powell in studies of discriminative eyeblink conditioning in rabbits.

A particularly compelling instance of convergence has to do with the general conclusion that executive attention can be invoked to varying degrees, depending on the amount of conflict present in a given situation (see Section II.B), such as when one of many possible responses must be generated in response to a given input or when the ink color and name of a word signal different behavioral responses. These examples generally correspond to instances in which particular inputs pose high processing demands. This facet of executive attention appears to have associative attentional counterparts. The large magnitude of anterior cingulate cortical discriminative neuronal activity during the initial sessions of avoidance training (with declining discharge magnitudes later in training) could represent an increase in associative attention for production of an active response on cue in the early stages of learning, when a clearly conflicting response tendency (to "freeze" in response to a painful shock) is at maximum levels. This example of convergence is in keeping with the "response-selection" notion of cingulate cortical functions (see Section II.B.2).

Also of relevance is the observation of salience compensation (see Section II.A.1.b). In this instance, the anterior cingulate cortex appears to allocate its associative attentional response in accord with an assessment of two factors, stimulus significance and stimulus salience. Only associatively significant stimuli that call for action provoke cingulate cortical neuronal firing. However, associatively significant stimuli of low salience evoke greater cingulate cortical neuronal responses than stimuli of high salience, when both stimuli are of equivalent associative significance. Anterior cingulate cortical neurons provide processing resources to increase the likelihood that a weak but important stimulus will be attended to. This example of convergence is suggestive of cingulate cortical engagement in high-demand stimulus processing situations with no response-selection component.

A prima facie case can be made for convergence of cognitive and behavioral neuroscience data regarding error-related processing. Results analogous to the error-related potential were found in a study of multi-unit activation following various trial outcomes during performance of rabbits in a discriminative avoidance task. Anterior cingulate cortex was the only one of four distinct cingulothalamic areas examined to exhibit a statistically reliable posttrial increase of neuronal firing following "error trials" during avoidance conditioning [i.e., trials in which the rabbits failed to respond (and received shock) following presentation of the CS^+]. Although the observed activation followed the presentation of the shock, it was not a direct sensory response to the shock because it was measured 5–15 sec after shock termination and was not exhibited by cingulothalamic neurons in the other

areas; these other neurons nonetheless exhibited brief, short-latency responses to the shock. In keeping with the intriguing findings of Donald Price and colleagues that anterior cingulate cortex has a unique role in mediating pain aversion, the sustained response to the shock in the anterior cingulate cortex may reflect neural coding of the aversiveness of the shock experience and the perceived failure to avoid the shock. It is possible that errors committed by human subjects give rise to a similar process, error-related pain, and that this is the process detected by studies of the ERN.

A final comment here concerns the apparent disagreement between the executive attention theory and the associative attention theory regarding cingulate cortical functions. Milham and colleagues concluded that anterior cingulate cortical processing is invoked specifically by conflict at the response level. Other aspects of executive attention, such as selection of task-relevant stimulus dimensions, are attributed by these researchers to related areas such as the dorsolateral prefrontal cortex. This idea may seem to be in conflict with findings that inspired associative attention theory—that robust discriminative TIA develops during instrumental avoidance learning in rabbits—results that clearly suggest an involvement of cingulate cortex in stimulus selection processes. It is important to note, however, that discriminative TIA illustrates selection of a specific task-relevant stimulus, not selection of a task-relevant stimulus dimension. The latter process is attributed by Milham and colleagues to the dorsolateral prefrontal cortex. Keeping in mind the presence of cingulate cortex (and the absence of dorsolateral prefrontal cortex) in small mammals such as rats and rabbits, it is worth considering that the selective reinforcement-based coding of significant stimuli in the service of response selection processes is indeed a province of the cingulate cortex, whereas more complex functions, such as selection of task-relevant stimulus dimensions and the control of such processes by linguistic instructions, may be conferred by areas of prefrontal cortex that simply do not exist in the smaller mammalian species. The consequences of dimensional selection in these areas are relayed to cingulate cortex wherein they contribute to response selection. Thus, the cingulate cortex in rats and rabbits is involved in stimulus selection and response selection processes, but stimulus selection in cingulate cortex deals not with selection of stimulus dimensions but rather with specific values of particular physical stimuli. Perhaps the cingulate cortex represents the sole resource for these processes in rats and rabbits. In primates, cingulate cortex is fed by additional areas gained during the course of evolution. These areas confer processing flexibility, including the capacity to form dimensional sets and response sets on the basis of linguistic instructional inputs.

B. Differences

In considering differences between executive attention and associative attention, the term "associative" is critical. This term is used in acknowledgment of the substantial evidence that cingulate cortical associative attention is a product of a learning process and that cingulate cortex is an important substrate of learning and memory. It will be argued specifically that cingulate cortex is involved in response and memory retrieval.

In contrast, executive attention as discussed by cognitive neuroscientists is quite separate from associative learning or memory. For example, it is argued that executive attention is invoked in early stages of learning or when habitual behavior must be revised, but in these instances executive attention acts as an external agent that intervenes to facilitate processing when new learning is needed rather than a process that is a product of learning or that subserves retrieval of learned information. Moreover, cingulate cortex does not appear to be included on the list of brain areas (e.g., hippocampus and prefrontal cortex) that cognitive neuroscientists believe to be involved in learning and memory.

One could argue that no learning process is involved in the evocation of anterior cingulate cortical activation in tasks that require executive attention. The subjects are simply given task-relevant instruction. They perform the experimental task and exhibit differential brain activation depending on the stimulating conditions. However, consideration should be given to the possibility that the task-related instructions instill a memory and a plan of action. The instructional memory functions as does reinforcement in a discriminative conditioning study, informing the subject as to which stimuli are significant and require action. The neural representation of the instructional memory could induce synaptic plasticity fostering incremented neuronal firing in response to the critical stimuli as well as associative pathways for selection of appropriate behavioral output.

Regarding the apparent functional duality being imputed to cingulate cortex, how can one area of the brain subserve both attention and memory retrieval?

The answer proposed is that these processes are fused as a unitary function of cingulate cortex as discussed later. The difficulty lies in the assumption that our human linguistic distinctions denote corresponding, distinct neural functions. We invite the reader to consider the possibility that the brain does not always respect the linguistic categories of neuroscientists.

1. Involvement of Anterior and Posterior Cingulate Cortex Respectively, in Early and Late Stages of Discriminative Learning

Generally supportive of cingulate cortical involvement in response and memory retrieval processes is the necessity of an intact cingulate cortex for a variety of forms of learning, including discriminative avoidance learning (see Section II.A). Also supportive is the massive TIA that develops during learning in cingulate cortex. Since it depends on the prediction of reinforcement by CS^+ and the prediction of no reinforcement by CS^-, the TIA is clearly a product of a learning process.

Particularly compelling is the correspondence between TIA development and the effects of restricted lesions. As indicated previously, discriminative TIA in the anterior cingulate cortex develops early in training, within the first 30 trials of discriminative avoidance learning (15 with CS^+ and 15 with CS^-), and damage in the anterior cingulate cortex or in the related MD thalamic nucleus has been associated in several studies with a deficit of behavioral performance in the early stages of learning. A case in point is described in Section III.A. In contrast, discriminative activity in layer IV of the posterior cingulate cortex and in the AV thalamic nucleus develops late, after the development of significant discriminative behavioral responding. Lesions in these areas are associated with normal behavioral acquisition followed by significant loss of performance efficiency in the later stages of training, when asymptotic performance is first attained in controls, and during subsequent overtraining. The correspondence between the training stage-specific behavioral deficits and the stages of training in which TIA develops in anterior and posterior cingulate cortex is consistent with a role of these areas in learned response retrieval.

Timothy Bussey, Barry Everitt, Trevor Robbins, and John Parkinson and colleagues at Oxford University have provided support for the early and late contributions of anterior and posterior cingulate cortex, respectively, to behavioral learning. Late-stage involvement of posterior cingulate cortex was indicated in studies involving both pre- and postsurgical training in a conditional discrimination task in which rapidly or slowly flashing lights predicted which of two levers, when pressed, would yield a reward. Posterior cingulate cortical lesions did not block acquisition of this discrimination but the lesions did impair performance in the late stages of training. Inexplicably, in this study the rats with anterior cingulate lesions did not differ from controls or were facilitated relative to controls. However, other studies of this group have supported an early contribution of anterior cingulate cortex to learning. These studies involved the Pavlovian conditioning of orienting behavior often referred to as "autoshaping." In Pavlovian conditioning the CS^+ is followed by reward delivery and the CS^- is not, without regard to the subjects' responses. Autoshaping refers to the fact that subjects acquire approach responses to the site of CS^+ delivery, but they do not approach the site of CS^- delivery. The approach responses to the CS^+ occur before the subjects actually approach and consume the reward. The rats given lesions in the anterior cingulate cortex failed to attain a significant level of discriminative autoshaping but they did show a trend toward significance in the late training stages. In a subsequent test session with reinforcement omitted, the control subjects showed robust discrimination and the animals with lesions also showed discrimination, although to a lesser degree. The authors stated that although anterior cingulate lesioned animals were significantly impaired at autoshaping, there was evidence that they were beginning to learn toward the end of the experiment. The aforementioned studies with rabbits suggest that this later learning was a product of posterior cingulate cortical involvement.

2. Complementary Functions of the Early and Late Circuits

It is interesting to consider the functional significance of the respective early and late involvement of anterior and posterior cingulate cortex in learning. Anterior cingulate cortex appears to be specialized for rapid encoding of novel CS^+-reinforcement contingencies in the early stages of training, as indicated by early discriminative TIA. A brain circuitry that is organized for such rapid encoding is not likely to handle well consistent coding of repetitive, enduring CS^+-reinforcement contingencies. For this reason, there exists a complementary system centered in the posterior cingulate cortical and anterior thalamic circuitry that does not encode stimulus reinforcement contingencies

until they have proven reliable through repetition (i.e., this system exhibits late discriminative TIA).

Intriguingly, as shown by John Freeman and Michael Gabriel, repetition of trials is not the only way to produce the late TIA in the posterior cingulate cortical circuit. TIA also develops when there is a substantial (48-hr) delay between an initial training session and a later training session. During this interval behavioral responding also improves significantly, a phenomenon referred to as "incubation." These results suggest that in addition to trial repetition, covert processing during the intersession interval can engender the late-developing TIA in the posterior cingulate cortex and in related anterior thalamic nuclei.

3. The Early Circuit and Working Memory

The rapid coding of associatively significant stimuli exhibited by anterior cingulate cortical neurons is quite similar to neuronal activity recorded in the pioneering studies of Patricia Goldman-Rakic in the lateral prefrontal cortex of primates—activity that Goldman-Rakic interprets as reflecting the operation of a working memory process. The shared properties of the neuronal activity in these regions suggest a role of the anterior cingulate in working memory processes. Clearly, a functional kinship between the anterior cingulate and lateral prefrontal areas would not be surprising because both areas are closely interconnected in primates and share common anatomical associations with the MD nucleus and the hippocampus. Of course, as mentioned previously, the lateral prefrontal cortex is absent in rabbits and rats, suggesting that anterior cingulate cortex may be the sole center of a rudimentary working memory circuit in these animals—a memory circuit that has achieved substantial elaboration during the course of mammalian evolution.

4. Context-Specific Retrieval Patterns in the Late Circuit

Additional evidence in support of the retrieval hypothesis is afforded by consideration of the distinctive training stage-related posterior cingulate cortical and anterior thalamic TIA. Extensive multi-site recording in these areas during discriminative avoidance learning has demonstrated unique topographic distributions of brief-latency CS^+- elicited neuronal activity across distinct nuclei of the anterior thalamus and the layers of posterior cingulate cortex. Certain thalamic nuclei and cortical layers are maximally activated by the CS^+ in the initial session of training, different areas are activated maximally in intermediate training stages, and other areas are activated maximally in late stages of acquisition, as asymptotic discriminative performance occurs. These changing patterns thus afford a "temporal tag" that codes the "age" of a given habit or memory and that could be used as a cue for retrieval.

In a different study it was shown that the same physical cues elicited different patterns of activation depending on whether the subjects were engaged in the performance of the moderately learned discriminative avoidance habit or (in a separate training apparatus) the well-learned discriminative approach habit. Thus, the elicited patterns coded the specific habit being practiced. The fact that different patterns are elicited by the same physical stimuli in the same subjects depending on which of two discriminative tasks is being engaged means that the patterns could uniquely retrieve habit-specific (context-appropriate) memories and learned behavior. Thus, the distributions of activation changed systematically, not only across time (training stage) but also with respect to the training context. Therefore, the patterns afford potentially a neural representation of both spatial and temporal aspects of the training context. These habit-specific patterns, elicited at brief latency (80 msec) by the cue (CS^+) that calls forth the learned response, are arguably the brain's earliest sign (in the millisecond series) of context-specific retrieval. Thus, the patterns could contribute to pattern separation, the property of retrieval whereby different behavioral responses and/or memories are retrieved in response to very similar and thereby confusable eliciting cues. Pattern separation defeats intertask proactive and retroactive interference.

5. Retrieval and Firing Ramps in Cingulate Cortex

Also consistent with the idea of a cingulate cortical role in response retrieval is the finding described earlier that many cingulate cortical neurons show firing ramps or progressive firing rate increments in advance of the performance of learned responses. Thus, whereas the brief-latency cue-elicited patterns of cingulate cortical activation may represent initial stimulus evaluation constituting the earliest sign of retrieval in the brain, the pre-response firing increments may represent the final command volley calling for execution of the learned response. In the language of executive attention, this activity could be viewed as a direct neural indicant of response selection. Thus, the notions of

response selection and response retrieval have a great deal of common meaning.

6. Dependence of TIA Patterns on the Hippocampus

Studies have indicated that the brief-latency, habit-specific patterns of activation found in the posterior cingulate cortex and related anterior thalamic areas depend on input from the subicular complex of the hippocampal formation, which is an origin of substantial fiber projections to the anterior thalamus and to the posterior cingulate cortex. Damage to the subicular complex alters and degrades the exquisite training stage-related patterns of activation found in cingulate cortex, although basic discriminative TIA remains intact in subjects with subicular complex damage. Any role that the stage-related patterns play in response retrieval would presumably be diminished or abolished in subjects with hippocampal formation damage.

Recent experiments have indicated that the context-specific patterns of activation exhibited in the posterior cingulate cortex and related areas of the thalamus are not essential for simple cued instrumental learning, such as discriminative avoidance and approach learning, which are mediated by basic, nonpatterned discriminative TIA. However, the patterns are important when the subject is confronted with a more challenging retrieval problem, such as when the subject is required to perform two discriminative tasks, approach and avoidance, with reversed cues (i.e., the CS^+ in one task is the CS^- in the other and vice versa). In this case, fornix lesions that disconnect the subicular complex and the anterior thalamus significantly impair dual-task performance, whereas performance in either task singly is not affected.

It is important to note that the functions of cingulate cortex are not restricted to simple acts of retrieval such as occur in instrumental approach and avoidance learning. There is evidence of cingulate cortical involvement in more complex forms of learning in animals and in humans. From the pioneering work of James Ranck, John O'Keefe, Lynn Nadel, Philip Best, Bruce McNaughton, Howard Eichenbaum, Carol Barnes, and Jeffrey Taube, the existence of place cells and head direction cells has been well established. Place cells are sensitive to the location of the subject (i.e., a given cell fires only when the subject occupies a particular area of the experimental environment). Head direction cells are sensitive to the points of the compass (i.e., a given cell fires when the subject is oriented in a particular direction). Place cells are abundant in the hippocampus, a structure that is closely interconnected with the cingulate cortex and has been widely implicated in spatial learning in animals as well as in human episodic and declarative memory. Recently, it has been found that head direction cells and place cells are present in cingulate cortex and in related areas of the thalamus. Moreover, damage in cingulate cortex has been shown to disrupt behaviors (such as acquisition in the Morris water maze) that depend on spatial processing. It is generally well-known that visuospatial information that defines a learning environment contributes importantly to the retrieval of context-appropriate memory and behavior. A commonly held interpretation of head direction and place cells is that they reflect neurological processes whereby information about the spatial context is used to retrieve context-appropriate behaviors and memories.

7. Retrosplenial Amnesia

Additional evidence for a cingulate cortical involvement in memory processes is indicated by the report of Dawn Bowers and colleagues on the amnesic patient TR, who sustained circumscribed damage in the left retrosplenial cortex, a subdivision of the posterior cingulate cortex. As a result of the lesion, TR developed retrograde amnesia for events occurring 9–12 months before the lesion and virtually complete anterograde amnesia (amnesia for events after the lesion). Of particular interest is the finding that TR was impaired on making judgments about the temporal order of visually presented sentences and faces, but TR was unimpaired in making recognition judgments of the same items. In other words, he was able to report accurately that he had seen a sentence or face before but could not accurately report the order in which the material had been presented. TR was impaired in making judgments about which of two target words in a list occurred more recently, even when the most recent target was the last stimulus presented, suggesting that the inability to "time-tag" new information took place early in the process of memory acquisition. The deficit in recall of temporal order extended only to recently acquired information because TR performed at control levels on a task that required the temporal ordering of remote memory (memory for public events from 1930 to 1970). The specific deficit in temporal ordering is intriguingly concordant with the hypothesis that the training stage-related peaks of activation in the posterior circuit represent a temporal context code that is used for mnemonic retrieval.

8. Memory Consolidation and the Cingulate Cortex

Memory consolidation refers to processes whereby memories become less susceptible to interference and disruption with the passage of time. It is commonly held that consolidation is based on rearrangements of the circuitry and neuroanatomical substrates involved in memory storage and retrieval. Thus, for example, it has been proposed that the hippocampus and neocortex are critically involved in the initial encoding of complex memory in humans but as the "age" of a given memory increases its dependence on the hippocampus decreases.

Previously, it was indicated that anterior and posterior cingulate cortex are critically involved respectively in mediating the early and late stages of discriminative avoidance learning of rabbits. Therefore, it might be assumed that the posterior cingulate cortex and related thalamic areas are the final sites of habit storage and retrieval. However, it is now known that neither anterior nor posterior cingulate cortex, or the related areas of the limbic thalamus, are involved in mediating discriminative avoidance learning in animals that are highly overtrained. Cingulothalamic lesions that normally block learning and prevent the expression of the learned behavior in subjects that are trained to asymptotic levels before they receive the lesions do not impair performance when they are induced after rabbits have been given extensive post asymptotic overtraining. Thus, in highly overtrained rabbits the brain processes involved in storage and retrieval of the discriminative avoidance habit occur entirely within noncingulate circuitry. These results are in accord with the findings concerning retrosplenial amnesia, suggesting that posterior cingulate cortex is involved in the mediation of memory for recently established memories but not for long-stored, "remote" memories. Similarly, lesions in the hippocampus, a region that is closely interconnected and highly interactive with posterior cingulate cortex, are associated with loss of recent memory but not of remote memory. The areas that mediate storage and retrieval of these remote memories have not been definitively identified, although it is widely assumed that the neocortex plays an important role.

IV. CONCLUSION

Studies carried out by behavioral neuroscientists using rats and rabbits have shown that the cingulate cortex is a critical substrate of instrumental learning of goal-directed behavior. Cingulate cortical neurons in these animals code associatively significant stimuli and exhibit context-specific topographic patterns that could mediate cued retrieval of context-appropriate learned behavior. These functions occur as a result of intimate interactions of hippocampal and cingulothalamic brain regions. Studies of cognitive neuroscientists concerning brain activation during cognitive task performance in human subjects have recently yielded many important findings, promising for the first time major advances in understanding complex cognitive processes of the human brain. These studies have yielded results that are fundamentally in agreement with the results of the studies on rats and rabbits. However, cognitive neuroscientists have discussed cingulate cortex as involved in attentional processes, with particular reference to response selection. They have not explicitly included cingulate cortex as an important component of the brain's memory system. However, given the findings of behavioral neuroscience and the very close neuroanatomical association of cingulate cortex with other structures (e.g., the hippocampus and parahippocampal cortex) that are acknowledged by a consensus of neuroscientists as components of the memory system, it is very likely that there will soon occur an even greater convergence of behavioral and cognitive neuroscience on a common mnemonic interpretation of cingulate cortical function.

Acknowledgments

This work was supported by National Institutes of Health Grant NS26736 to MG .

See Also the Following Articles

ATTENTION • BRAIN ANATOMY AND NETWORKS • BRAIN LESIONS • CEREBRAL CORTEX • CHEMICAL NEUROANATOMY • GABA • MOTOR CORTEX • NEOCORTEX • WORKING MEMORY

Suggested Reading

Bowers, D., Verfaellie, M., Valenstein, E., and Heilman, K. M. (1988). Impaired acquisition of temporal information in retrosplenial amnesia. *Brain Cogn.* **8,** 47–66.

Bussey, T. J., Muir, J. L., Everitt, B. J., and Robbins, T. W. (1997). Triple dissociation of anterior cingulate, posterior cingulate, and medial frontal cortices on visual discrimination tasks using a touchscreen testing procedure for the rat. *Behav. Neurosci.* **111,** 920–936.

Dahaene, S., Posner, M. I., and Tucker, D. M. (1994). Localization of a neural system for error detection and compensation. *Psychol. Sci.* **5,** 303–305.

Gabriel, M. (1993). Discriminative avoidance learning: A model system. In *Neurobiology of Cingulate Cortex and Limbic Thalamus* (B. A. Vogt, and M. Gabriel, Eds.), pp. 478–523. Birkhauser, Toronto.

Gabriel, M. (1999). A tale of two paradigms: Lessons learned from parallel studies of classical eyeblink conditioning and discriminative avoidance learning. In *Engrams: Model Systems of Vertebrate Learning: Feschrift Volume in Honor of Professor Richard F. Thompson* (A. Steinmetz, and M. Gluck, Eds.). Erlbaum, New York.

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., and Donchin, E. (1993). A neural system for error detection and compensation. *Psychological Science* **4,** 385–390.

Goldman-Rakic, P. S. (1988). Topography of cognition: Parallel distributed networks in primate association cortex. *Annu. Rev. Neurosci.* **11,** 137–156.

Harvey, J. A., and Kosofsky, B. E. (1998). Cocaine: Effects on the developing brain. *Ann. N.Y. Acad. Sci.* **846,** 208.

Holroyd, C. B., Dien, J., and Coles, M. G. (1998). Error-related scalp potentials elicited by hand and foot movements: Evidence for an output-independent error-processing system in humans. *Neurosci. Lett.* **242,** 65–68.

Milham, M. P., Banich, M. T., Webb, A., Barad, V., Cohen, N. J., Wszalek, T., and Kramer, A. F. (2001). The relative involvement of anterior cingulate and prefrontal cortex in attentional control depends on nature of conflict. *Cog. Brain Res.* **12**(3), 467–473.

Nadel, L., and Eichenbaum, H. (1999). Introduction to the special issue on place cells. *Hippocampus* **9,** 341–345.

Papez, J. W. (1937). A proposed mechanism of emotion. *Arch. Neurol. Psychiatry* **38,** 725–743.

Paus, T., Petrides, M., Evans, A. C., and Meyer, E. (1993). Role of the human anterior cingulate cortex in the control of oculomotor, manual, and speech responses. *J. Neurophysiol.* **70,** 453–469.

Parkinson, J. A., Willoughby, P. J., Robbins, T. W., and Everitt, B. J. (2000). Disconnection of the anterior cingulate cortex and nucleus accumbens core impairs Pavlovian approach behavior: Further evidence for limbic cortical–ventral striatopallidal systems. *Behav. Neurosci.* **114,** 42–63.

Picard, N., and Strick, P. L. (1996). Motor areas of the medial wall: A review of their location and functional activation. *Cereb. Cortex* **6,** 342–353.

Posner, M. I., and DiGirolamo, G. J. (1998). Executive attention: Conflict, target detection, and cognitive control. In *The Attentive Brain* (R. Parasuraman, Ed.), pp. 401–423. MIT Press, Cambridge, MA.

Posner, M. I., and DiGirolamo, G. J. (2000). Attention in cognitive neuroscience: An overview. In *The New Cognitive Neurosciences* (M. S. Gazzaniga Ed.), pp. 623–724. MIT Press, Cambridge, MA.

Powell, D. A., Buchanan, S. L., and Gibbs, C. M. (1990). Role of the prefrontal-thalamic axis in classical conditioning. In *The Prefrontal Cortex: Its Structure, Function and Pathology*, Vol. 85, (H. B. M. Uylings, C. G. Van Eden, J. P. C. De Bruin, M. A. Corner, and M. G. P. Feenstra, Eds.), pp. 433–466. Elsevier Science, Amsterdam.

Rainville, P., Duncan, G. H., Price, D. D., Carrier, B., and Bushnell, M. C. (1997). Pain affect encoded in human anterior cingulate but not somatosensory cortex. *Science* **277,** 968–971.

Takenouchi, K., Nishijo, H., Uwano, T., Tamura, R., Takigawa, M., and Ono, T. (1999). Emotional and behavioral correlates of the anterior cingulate cortex during associative learning in rats. *Neuroscience* **93,** 1271–1287.

Taube, J. S. (1998). Head direction cells and the neurophysiological basis for a sense of direction. *Prog. Neurobiol.* **55,** 225–256.

Turken, A. U., and Swick, D. (1999). Response selection in the human anterior cingulate cortex. *Nat. Neurosci.* **2,** 920–924.

Circadian Rhythms

CHARLES A. FULLER and PATRICK M. FULLER
University of California, Davis

GLOSSARY

circadian rhythm A term coined by Franz Halberg in 1959 to describe the approximately 24-hr biological cycles that are endogenously generated by an organism (Latin: *circa*=about, *dies*=day).

entrainment When an exogenous stimulus achieves both phase and period control of one or more circadian oscillators, entrainment has occurred.

free-running rhythms Rhythms that persist even when an animal is isolated from external time cues. These rhythms do not damp out, are self-sustained, and have a period that is close to 24 hr.

masking Light and other stimuli can superimpose direct effects on the level of a variable that can alter the amplitude and waveform of circadian rhythms. These exogenous influences are referred to as "masking effects." Masking may cause an observed rhythm to inaccurately reflect the underlying circadian pacemaker.

phase The instantaneous state of an oscillation (i.e., maximum or minimum) defined by the value of a rhythmic variable and its time derivative. Often, in order to facilitate comparisons between rhythms or between rhythms and cyclic time cues, the time of a specific phase is chosen as a cycle reference point.

phase angle The difference in time between the phases of a rhythm and the entraining cycle or between two different rhythms. The phase angle is usually measured in fractions of the entire period.

phase response curve (PRC) A plot of the magnitude and direction of a phase shift as a function of the circadian phase at which a phase-shifting stimulus (typically light) is applied.

zeitgeber German for "time giver;" an environmental time cue such as sunlight, food, noise, or social interaction that is capable of entraining the biological clock to a 24-hr day.

Biological rhythms are oscillations in an organism's physiology and behavior. These rhythms span a wide range of frequencies, ranging from fractions of a second to a year. Many rhythms exhibit a periodicity close to that of the 24-hr geophysical cycle that results from the daily rotation of the earth around its axis. Biochemical, physiological, and behavioral rhythms that oscillate with a period close to 24 hr are termed circadian rhythms. Circadian rhythms are generated by an endogenous pacemaker, are self-sustaining, and persist in the absence of time cues. It has been hypothesized that the original adaptive role for circadian organization was to ensure that phases of DNA replication sensitive to damage by high ultraviolet light levels in the primitive day were protected by being synchronized to night. Examples of human circadian rhythms are shown in Fig. 1. These include the daily oscillations of the sleep–wake cycle, body temperature, growth hormone, cortisol, and urinary potassium excretion. Circadian rhythms provide temporal organization and coordination for physiological, biochemical, and behavioral variables in all eukaryotic organisms and some prokaryotes. Without this important temporal framework, an organism will experience, minimally, a severely reduced homeostatic capacity. Circadian desynchronization can be likened to a symphony without a conductor—each instrument section may be fully functional, but the sections collectively are unable to harmonize. The medical importance of circadian rhythms is just beginning to be appreciated in the clinical arena. For example,

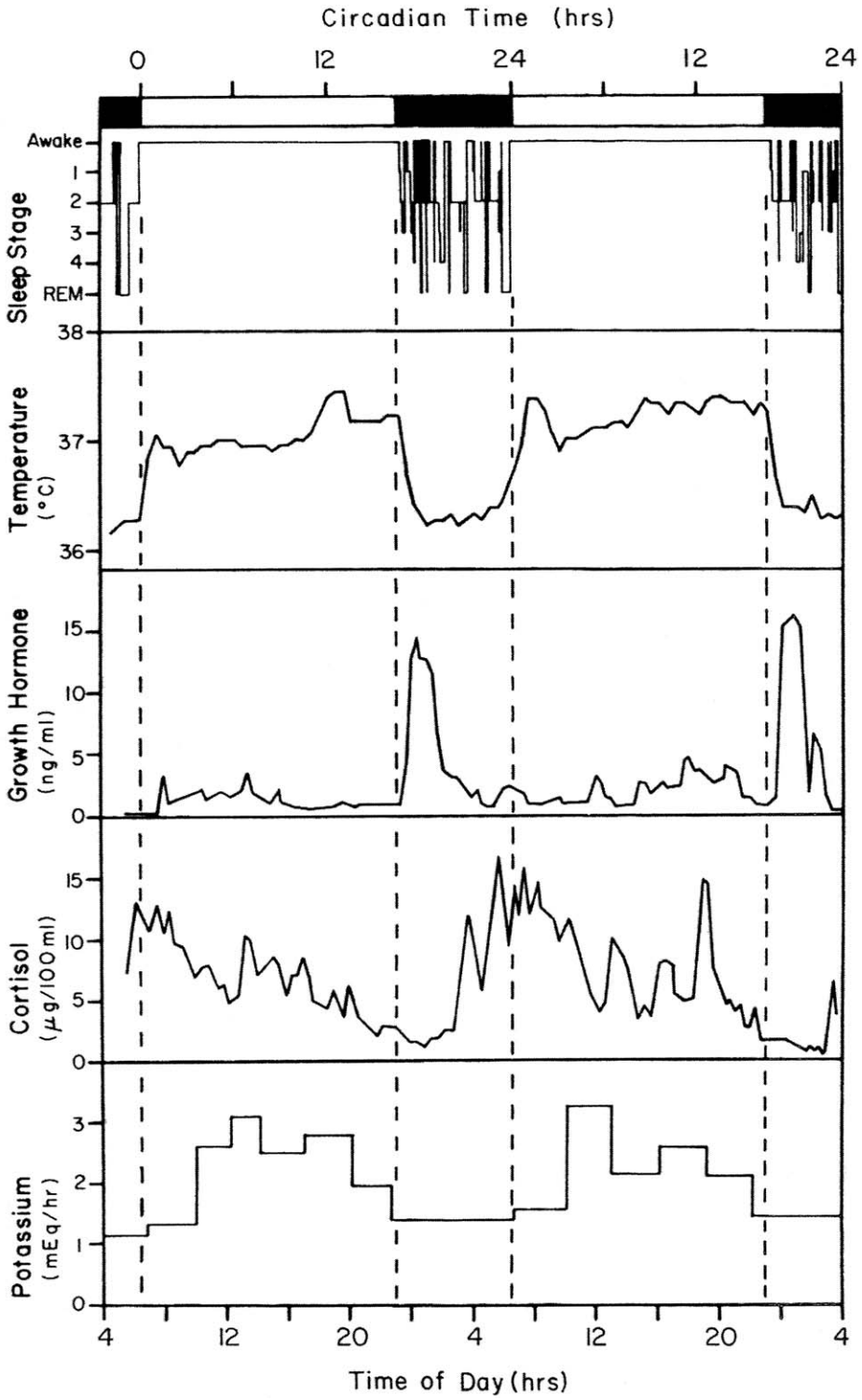


Figure 1 A representative panel of human circadian rhythms. Starting from the top panel: Rhythms in sleep, body temperature, plasma growth hormone, plasma cortisol, and urinary potassium excretion. These data represent 48 hr of data from a subject entrained to a 24-hr day (LD 16:8; 16 hr of light and 8 hr of dark). The light and dark periods are shown at the top of the data panels. [Reproduced by permission from Moore-Ede, M. C., Sulzman, F. M., and Fuller, C. A. (1982). *The Clocks That Time Us: Physiology of the Circadian Timing System*. Harvard University Press, Cambridge, MA].

circadian rhythms are now recognized as important variables in the diagnosis and treatment of many pathophysiological conditions. Many of the properties of circadian rhythms are analogous to physical oscillators, and, consequently, much of the terminology used to describe rhythms is derived from oscillator theory. All biological rhythms are characterized by fundamental parameters, which include period length, rhythm mean, circadian waveform, and circadian amplitude. Each parameter describes an important property of the circadian timing system. Figure 2 illustrates the important parameters of biological rhythms, zeitgebers, and reference time scales. After decades of circadian rhythm research, several generalizations can be made about circadian rhythms: (1) circadian rhythms are ubiquitous, they are present in all eukaryotes and some prokaryotes; (2) circadian rhythms are genetically determined, not learned; (3) circadian rhythms are generated by an endogenous, self-sustained pacemaker; (4) in an environment without external time cues, circadian rhythms persist with a period that approximates 24 hr; (5) the period of the circadian pacemaker is temperature-compensated; and (6) the circadian pacemaker can be entrained to external time cues.

I. INTRODUCTION

As a consequence of the Earth's daily rotation about its axis, all terrestrial organisms have evolved in an environment with alternating cycles of light and dark. During the course of evolution, organisms have adapted to the challenge of living with a light–dark

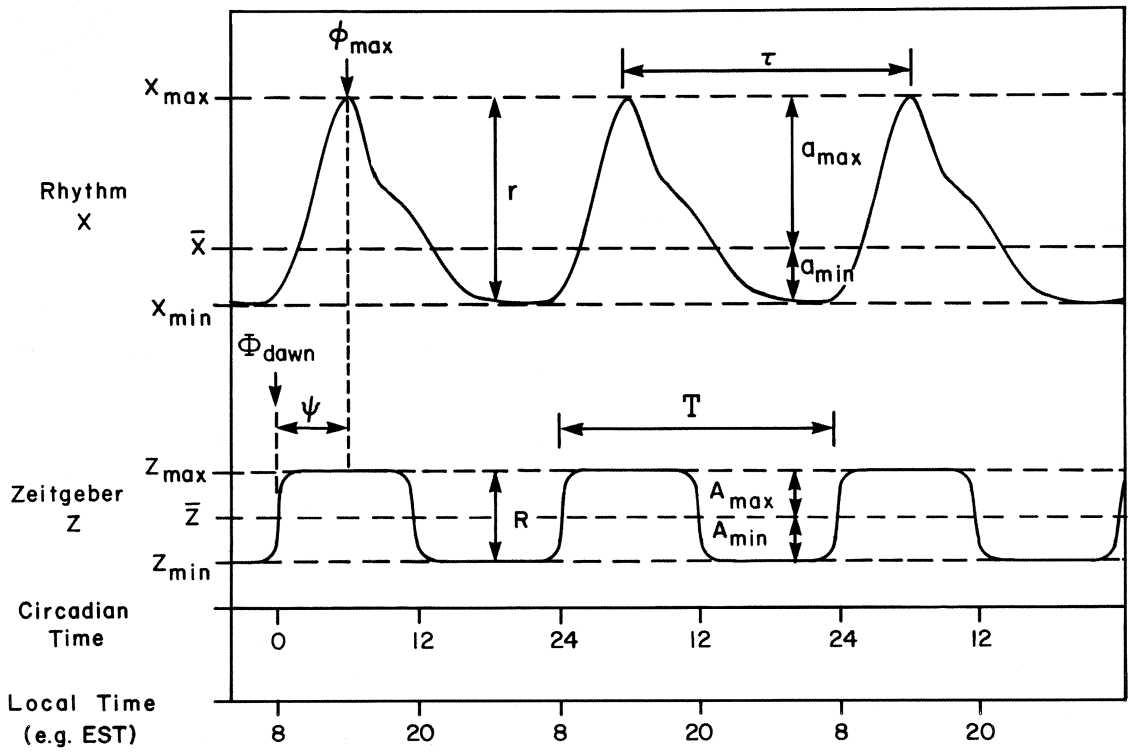


Figure 2 The important parameters of biological rhythms, zeitgebers, and reference time scales. Top panel, rhythm of a biological variable, x; bottom panel, cycle of an environmental zeitgeber, Z, which entrains the rhythm. These variables are plotted against two different time scales. The local time scale indicates the time of day at which the biological measurement was made. The circadian time scale standardizes the relationship of the biological variable to the zeitgeber cycle by defining circadian time 0 as dawn. The parameters for the zeitgeber are given in captial letters and those for the rhythm in lower case. Identified for rhythm and zeitgeber cycles are the mean values of the variable (x and Z), the maximum and minimum values (x_{max}, x_{min}, Z_{max}, Z_{min}), the ranges of the oscillations (r and R), the periods of the oscillations (τ and T), the amplitudes (a_{max}, a_{min}, A_{max}, and A_{min}), and the reference phases (ϕ and Φ). The figure also shows the phase relationship (ψ) between dawn on the zeitgeber cycle and the maximum of the rhythm, but ψ can be defined between any two reference points on the zeitgeber and rhythm waveforms. [Reproduced by permission from Moore-Ede, M. C., Sulzman, F. M., and Fuller, C. A. (1982). *The Clocks That Time Us: Physiology of the Circadian Timing System*. Harvard University Press, Cambridge, MA].

(LD) cycle by occupying temporal niches in which they can optimize life-sustaining activities such as feeding, sleeping, and avoiding predators. It then is not surprising, perhaps, that organisms, ranging from prokaryotes to multicellular eukaryotes, have evolved an internal time-keeping system. In mammals, this internal time-keeping system, the circadian timing system (CTS), provides temporal organization for a host of behavioral, physiological, and biochemical variables. The fundamental adaptive advantage of this temporal organization is that it allows for predictive, rather than entirely reactive, homeostatic regulation of function. For example, prior to waking, body temperature, heart rate, blood pressure, and plasma cortisol increase in anticipation of increased energetic demands. In addition, the CTS, by monitoring day length and its rate and direction of change, contributes to adaptive reproductive and seasonal rhythms.

A. Discovery of Circadian Rhythms

Daily rhythms in the activities of plants and animals have been recognized from the earliest recorded times. It was not until 1729, however, that diurnal rhythmicity was empirically shown to be an endogenously generated phenomenon and not merely a passive response to a cyclic environment. That year, French astronomer Jean Jacques d'Ortous de Mairan studied the leaf movements of the heliotrope plant, *Mimosa pudica*. De Mairan observed that the leaves and pedicels of the plant continued to open during the day and close during the night even when the plant was isolated from the normal light–dark environment. De Mairan's study was not generally accepted as proof that circadian time-keeping was an innate property of the plant because other external (geophysical) oscillations (e.g., temperature or electromagnetic oscillations) had not been ruled out as time cues. Since de Mairan, other scientists have attempted to account for the potential influence of these external oscillations. For example, Hamner and co-workers showed the persistence of circadian rhythmicity in hamsters (*Mesocricetus auaratus*), fruit flies (*Drosophila*), and a fungus (*Neurospora*) when placed on a table rotating counter to the earth's rotation at the South Pole. More recently, Dr. Frank Sulzman and co-workers tested the exogenous vs endogenous hypothesis during NASA's Spacelab 1 mission. For this experiment, the mold *Neurospora crassa* was flown in the orbiting space shuttle. The specimens were maintained in constant conditions (e.g., constant dark conditions). The results showed that *Neurospora* conidiation rhythms persisted though the period of the rhythm was significantly different from the earth's 24-hr geophysical period of rotation. This finding made it clear that living organisms have internal time-keeping devices.

B. Free-Running Rhythms, Entrainment, and Masking

Persisting circadian rhythms in animals and plants isolated from environmental time cues are termed "free-running" rhythms. Every species possesses a characteristic average free-running period (τ) with individual periods distributed normally around the species mean. Most species exhibit a period close to, but not exactly, 24 hr. It was determined that the human pacemaker has an endogenous period very close to 24 hr. A true free-running circadian rhythm will persist, for the most part, indefinitely in an environment without any time cues. In an early study by Curt Richter, a blinded squirrel monkey showed a persisting circadian activity rhythm for as long as it was studied ($\sim$3 years).

If circadian clocks free-ran in the natural environment they would be of little use to the organism. Evolution has ensured that the circadian system can be synchronized to certain environmental time cues or *zeitgebers*. The process by which circadian rhythms are synchronized to periodic environmental time cues is called entrainment. The light–dark cycle is the predominant environmental entraining agent in most plants and animals, including humans. However, daily cycles of food availability, social interactions, and ambient temperature have also been shown to be effective entraining agents in *some* species. Environmental synchronization is adaptive because it confers on an organism the ability to maintain appropriate phase relationships with the external environment throughout the year. From an evolutionary perspective, the ability to perform certain behaviors at the appropriate time of day or the appropriate season confers a strong selective advantage to an organism. In effect, entrainment ensures that an organism does the right thing at the right time. Entrainment, however, is restricted to synchronizers, which cycle close to 24 hr. If the period of the zeitgeber is too short or too long, the circadian system is incapable of being entrained. This phenomenon is referred to as the "range of entrainment."

Any external signal that is capable of entraining a circadian rhythm will set the phase of the pacemaker's oscillation; however, the signal may also affect the overt (measured) rhythm in a more direct way, with or without a relationship to the process of entrainment. Such a direct effect by the signal obscures the rhythm of the oscillation and is referred to as masking. Masking may be adaptive, however. For instance, masking may augment the amplitude of a rhythm and therefore complement clock control, or, alternatively, masking may represent the reactive portion of homeostatic control with the CTS representing the predictive component. Virtually all rhythmic physiological variables can show masking and, therefore, care must be taken when evaluating circadian data.

Light is the most potent zeitgeber and the primary environmental agent for entraining the circadian timing system. Not suprisingly, then, light is typically the principal dependent variable in chronobiological studies. For the purpose of this article, the term light refers to the radiant energy detected by the physiological sensors in humans and animals. Since the late 1970s, Dr. Charles Czeisler and colleagues have performed many experiments to evaluate the ability of light to entrain and phase-shift the human circadian clock. These and other studies have demonstrated the utility of light therapy to reset rhythms in humans suffering from various circadian and sleep–wake disorders.

C. Phase Shifts and Phase Response Curves

The circadian clock is not equally sensitive to a given zeitgeber at all times of day. This fact is a conserved feature in all examined taxa and is graphically summarized in a phase response curve (PRC), shown in Fig. 3, which plots the magnitude and direction of a phase shift as a function of the circadian phase at which a stimulus is applied. The PRC models the mechanisms by which the circadian clock is entrained to environmental cycles. For example, PRCs to light stimuli illustrate the differential effects of light at different circadian phases. Light, the principal environmental zeitgeber, has its greatest effect on the circadian clock when delivered during the subjective night, the time when a species is normally exposed to darkness. Light pulses will phase-delay rhythms in the early subjective night and phase-advance rhythms in the late subjective night. Little phase shifting occurs during the subjective day—the dead zone. These features of the light PRC explain the ability of the circadian pacemaker to entrain to daily light–dark cycles (e.g., entrainment by light-inducing phase advances–delays equals the difference between the free-running period and the light–dark cycle period). In humans the PRC for light is asymmetric; delay shifts are usually larger than advance shifts. The asymmetry of the light PRC underlies the relative ease of adjusting to westward transmeridian jet travel as compared to eastward travel (jet lag).

Nonphotic cues have also been shown to have a strong resetting effect on the clock. It has been suggested that nonphotic cues may provide an additional adaptive advantage to organisms. Food availability, ambient temperature, social interaction, acute exposure to sexual odors, and exercise are all examples of factors that have been shown to phase-shift and entrain the circadian system in many animals, including humans. In general, nonphotic cues cause phase shifts when administered during the subjective day. This phase sensitivity is during a period when the clock is insensitive to light. In fact, the PRC for nonphotic cues is about 180° out of phase with the photic PRC. Nonphotic cues may help to synchronize the free-running clock in blind individuals and attenuate temporal desynchrony in shift-workers. More recent studies have further suggested that events in the environment that reliably precede the onset of light can assume the resetting function of light through conditional stimulus control.

D. Mammalian Circadian Rhythms and the Endogenous Pacemaker

At the turn of the century, any observed oscillation in mammalian physiology or behavior was dismissed as a random fluctuation of little importance. This is largely because physiologists and physicians were grounded in the developing tenet of homeostasis. Homeostasis, a term coined by Dr. Walter B. Cannon, describes the relative constancy of the internal environment or milieu de interior. Although the concept of homeostasis accounted for the presence of daily variations in physiological systems, circadian rhythmicity per se was not recognized as an important biological characteristic. This changed in 1922, however, when Dr. Carl Richter demonstrated the endogenous nature of circadian activity rhythms in rats and, furthermore, showed that the rats were synchronized by both the light–dark cycle and the time of feeding. Later in the

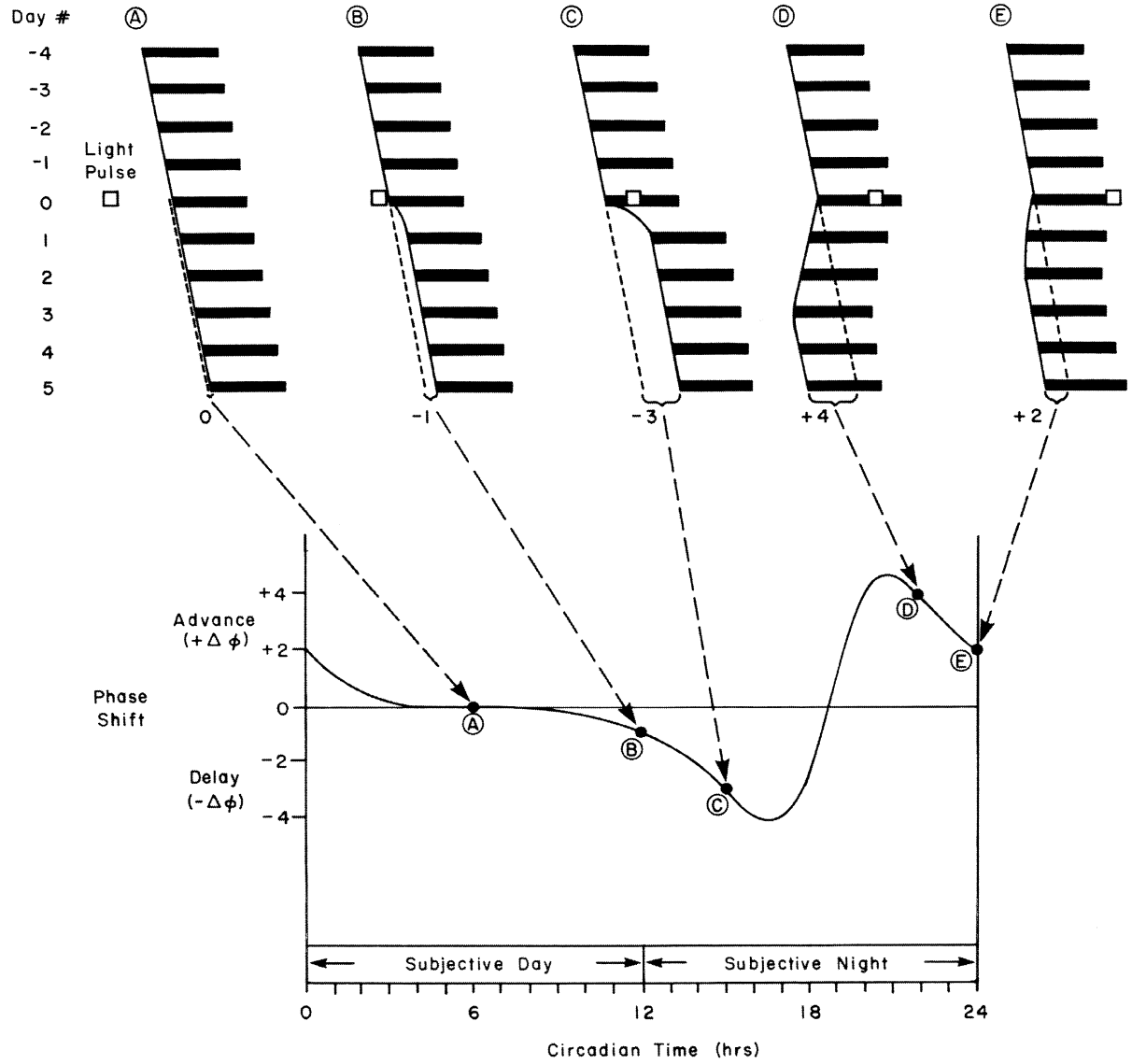


Figure 3 Derivation of a phase–response curve (PRC). (A)–(E) show five sample experiments in which a nocturnal animal, free-running in constant darkness, is exposed to a 1-hr light pulse. A free-running activity rhythm with a period of 25.0 hr is seen on days −4 to −1. On day zero, a light pulse is given at mid-subjective day (A), at late subjective day (B), at early subjective night (C), at late subjective night (D), and at early subjective day (E). The light pulses in mid-subjective day and early subjective night (B and C) produce phase delays of the activity rhythm that are complete within one cycle. The light pulses in late subjective night and early subjective day (D and E) produce phase advances with several cycles of transients before reaching a steady-state shift by day 5. *Lower panel*: Direction and magnitude of phase shifts plotted against the time of light pulses to obtain a PRC. When light pulses are given at frequent intervals throughout the subjective day and night, the waveform for the PRC follows the solid line. In mammals there is normally a gradual transition between maximum phase delay and maximum phase advance, with a point in mid-subjective night (like the one in mid-subjective day, A) where there is no phase shift. [Reproduced by permission from Moore-Ede, M. C., Sulzman, F. M., and Fuller, C. A. (1982). *The Clocks That Time Us: Physiology of the Circadian Timing System*. Harvard University Press, Cambridge, MA].

twentieth century, the endogenous nature of the human circadian timing system was confirmed, and circadian rhythms were documented in hundreds of physiological, biochemical, and behavioral variables. In fact, it became apparent that it was often more significant to find a physiological variable with no circadian rhythm.

The definition of a circadian rhythm may be straightforward; however, defining a biological clock is a bit more difficult. Any biological structure to be

considered a clock must be able to measure the passage of time independent of any periodic input from its environment as well as from timed biological events. In addition, in order to keep time reliably, a clock must have both resolution, an ability to detect the temporal order of two events closely spaced in time, and uniformity, a regularity of period and thus the ability to predict the occurrence of other regularly timed phenomena. Experimentation has revealed that a mammalian clock, which we now know resides in the suprachiasmatic nucleus (SCN) of the hypothalamus, has both relatively high resolution and uniformity. The uniformity of the clock is further augmented by the ability of an organism to "reset" its clock each day using environmental cues.

The location and function of the SCN, as a master pacemaker of the mammalian brain, have been confirmed by an impressive variety of experimental approaches. For example, lesion studies, in which the SCN was ablated, resulted in arrhythmic experimental animals. Neural transplantation of fetal SCN tissue into lesioned animals restored rhythmicity, confirming both the location and function of the circadian pacemaker. Similarly, animals with reciprocal SCN transplants have a circadian period determined by the genotype of the grafted tissue and not that of the host. The SCN also exhibits clear circadian rhythms of gene expression, metabolism, and electrophysiological activity *in vitro*.

Whereas the evidence supports the role of the SCN as a master circadian pacemaker, additional findings suggest that the circadian timing system is multioscillatory. In humans, the primary evidence for multiple pacemakers comes from temporal isolation studies in which some rhythms, such as the sleep–wake rhythm, and other rhythms, such as the body temperature rhythm, sometimes free-run with very different periods. This phenomenon is referred to as internal desynchronization.

The overall organization of the circadian timing system is thought to be hierarchical. That is, the SCN acts as a master oscillator, which in addition to receiving temporal cues from the environment drives a variety of slave oscillators. Some of these slave oscillators are thought to also generate oscillations independently, whereas others, termed passive slaves, do not. A study using fibroblasts in culture demonstrated that nonneural tissue might also possess intrinsic clocks. Serum induction of clock behavior in the fibroblast suggests that the SCN, as the principal oscillator, may simply orchestrate rather than actively drive oscillations in peripheral tissues.

Whereas the number of processes that show circadian fluctuation is large, a few physiological and behavioral rhythms (locomotor activity, body temperature, and melatonin) are commonly used to study the circadian timing system. Furthermore, several neural structures comprise the anatomical circadian timing system. Each will be discussed later in the article.

II. THE FUNCTIONAL AND ANATOMICAL CIRCADIAN TIMING SYSTEM

It is common to describe the anatomical and functional circadian timing system as consisting of three elements: input, pacemaker, and output. Though this simplification is heuristically useful, the reader should note that this descriptive paradigm belies the functional and anatomical complexity of the CTS.

A. The Pacemaker

Much is known about the anatomy and morphology of the SCN. The SCN is located in the anteroventral hypothalamus and consists of a pair of nuclei bordering the third ventricle at midline. Experimental lesions of the SCN result in the abolishment or severe disruption of most circadian rhythms, including locomotor activity, feeding, drinking, body temperature, sleep–wake, cortisol, melatonin, and growth hormone secretion. The unique functionality of the SCN is underscored by the fact that neonatal ablations of the SCN in rats permanently abolish circadian rhythms; thus, no other brain region can take over the function of the SCN. Although SCN lesions do not interfere with thermoregulation, eating, drinking, hormonal secretion, and so on, the temporal organization of homeostatic regulation is lost.

The SCN can be divided into two primary anatomical subdivisions, each with several neuronal subfields. The first subdivision, the shell, receives input from the basal forebrain, thalamus, and brain stem, whereas the second subdivision, the core, receives visual input, both direct, the retinohypothalamic tract (RHT), and indirect, intergeniculate leaflet (IGL), as well as afferents from the midbrain raphe, hypothalamus, and thalamus. The core subdivision, analogous to the ventrolateral SCN (vlSCN), projects densely to the shell, but reciprocal projections from the shell, analogous to the dorsomedial SCN (dmSCN), are sparse. Both the core and shell have commissural connections with homologous areas in the contralateral SCN.

In both the core and shell SCN, distinct subfields of SCN neurons exist. The retinorecipient zone of the vlSCN contains most of the neurons containing vasoactive intestinal peptide (VIP), gastrin-releasing peptide (GRP), and peptide histidine isoleucine (PHI). In contrast, the dmSCN contains most of the neurons containing arginine vasopressin (AVP) and somatostatin (SS). Notably, the VIP neurons form extensive connections both within the SCN and with extra-SCN targets. Photoinduction of c-Fos occurs in VIP neurons; thus, they are implicated in photic entrainment. The AVP- and SS-containing neurons are predominantly found in the dmSCN. These neurons exhibit robust circadian rhythms in peptide synthesis both in constant conditions and in hypothalamic slice preparations. More recently, it has been shown that GABA is colocalized in all neurons throughout the SCN and, thus, appears to be the principal small neurotransmitter in the mammalian SCN. A number of studies have also provided evidence for nitric oxide (NO) as an SCN neurotransmitter. Because the nitric oxide synthase-containing cells are found within VIP neurons, the putative role for NO is in mediating the effects of light.

The SCN also possesses a mechanism for generating autonomous circadian rhythms in individual pacemaker cells and the ability to synchronize these autonomous pacemaker cells. Electrophysiological evidence has strongly suggested that the generation of circadian rhythms within the SCN is by autonomous cell pacemakers and is not an emergent property of a population of cell oscillators as once thought. However, the intracellular mechanisms responsible for circadian rhythms and the mechanisms of cell–cell electrical coupling remain poorly understood.

B. Input

1. Retinohypothalamic Tract

The best characterized projection to the SCN is the retinohypothalamic tract (RHT). The RHT serves as the anatomical and physiological circadian visual system and is distinct from the visual system responsible for reflex oculomotor function and image formation. The RHT originates in a subset of specialized retinal ganglion cells, W-cells, and terminates bilaterally within the vlSCN.

Ablation of the RHT abolishes photoentrainment, suggesting that all photoreceptors responsible for entrainment of the CTS are in the retina. Retinal rod and cone photoreceptors are thought to participate in CTS entrainment. However, other putative photoreceptive elements have been implicated in entrainment, including several non-opsin-based photopigments (light-absorbing heme moieties such as hemoglobin and bilirubin), B2-based photopigments (i.e., cryptochromes), and, more recently, melanopsins. Electrophysiological and retrograde labeling studies suggest a diffuse distribution and high photic threshold for the photoreceptive cells of the RHT. The spectral sensitivity of the photoreceptive system that mediates entrainment of the hamster's circadian rhythms has a maximum sensitivity near 500 nm. Dr. Joseph Takahashi noted two unusual features of this system: the threshold of the response is high, especially for a hamster with a predominantly rod retina, and, second, the reciprocal relationship between intensity and duration holds for relatively long durations (up to 45 min). Largely on the basic of these observations, Takahashi has suggested that the clock's photoreceptive system is capable of integrating total photon exposure over a long period without becoming nonlinear in response. In addition, the unique intensity–response curve and relatively high threshold suggest that the photoreceptor system, at least in the hamster, is specialized for luminance coding in the range of light intensities occurring around dusk and dawn.

In vitro and *in vivo* electrophysiological studies have provided compelling evidence that the RHT utilizes glutamate as its primary transmitter. Both ionotropic and metabotropic glutamate receptors for glutamate have been identified in the SCN. Antagonists, which act upon ionotropic glutamate receptors, both reduce the phase-shifting effects of light on rodent locomotor rhythms and attenuate the photoinduction of c-Fos in the SCN. Additional neurotransmitters, including substance P (SP) and pituitary adenylyl cyclase-activating peptide (PACAP), are also thought to participate in retinohypothalamic transmission. Both SP and PACAP have been shown to modulate glutamate neurotransmission, though it is not currently understood how. Extensive analysis on the organization, synaptology, and morphology of the RHT projection to the SCN has revealed considerable structural plasticity in its synapses with the SCN.

2. Geniculohypothalamic Tract

The intergeniculate leaflet (IGL), a distinct lamina of neurons interposed between the dorsal and ventral lateral geniculate nuclei of the thalamus, gives rise to a dense neuropeptide-Y (NPY) containing projection that terminates bilaterally in the vlSCN. This projec-

tion is the geniculohypothalamic tract (GHT), and the termination in the vlSCN is coextensive with the retinorecipient neurons. This NPY projection is thought to subserve, in part, nonphotic entrainment of the SCN. Lesion of the IGL results in altered phase angles of entrainment consistent with a role of mediating nonphotic input to the SCN. However, some IGL cells are photically responsive. These cells are innervated by retinal afferents, which are collaterals of fibers that also innervate the SCN. Thus, the SCN can receive photic influences via the multisynaptic retina–IGL–GHT–SCN pathway. It is thought that the IGL's primary role is to integrate photic and nonphotic information in order to provide a regulatory influence on the SCN. GABA colocalizes with NPY-containing cells in the IGL and also with a subset of enkephalin-containing neurons, which give rise to a commissural projection to the contralateral IGL.

3. Raphe Afferents

The ventral and medial SCN receive dense innervation from serotonergic (5HT) fibers originating in the median raphe nuclei. The functional significance of this pathway remains unclear; however, evidence suggests that the primary role of the serotonergic innervation of the SCN is to modulate the activity of the retinal afferents (i.e., photic information). Similar to the GHT innervation of the SCN, the 5HT raphe projection to the SCN is coextensive with visual input and also synapses on VIP neurons in the vlSCN. The dorsal raphe nuclei projects to the IGL, providing another pathway to modulate the SCN. Some effects of 5HT are thought to be mediated by both the $5HT_7$ receptors in the SCN and the $5HT_{1B}$ receptors located on retinal afferents to the SCN. However, because several other 5HT receptor subtypes exist in several sites in and around the SCN, the exact mechanisms by which the serotonergic raphe acts on the SCN are unclear.

4. Other Afferents

Whereas the SCN does not appear to contain any intrinsic cholinergic neurons, it receives choline acetyltransferase-immunopositive projections from the pedunculopontine tegmental nuclei, the lateral dorsal tegmental nuclei, and the basal forebrain. These nuclei contribute to sleep and wakefulness and probably other behavioral state changes. *In vivo* administration of carbachol, a cholinergic agonist, advances the phase of the SCN at night via muscarinic acetylcholine receptor-mediated activation of guanylyl cyclase–cGMP–PKG pathways. Nicotinic acetylcholine receptors may also mediate, in part, the cholinergic influence on the circadian system; however, the extent of the mediation remains undetermined.

Many additional hypothalamic projections to the SCN have been identified and characterized both anatomically and pharmacologically. These include hypothalamic projections from the preoptic, arcuate, dorsomedial, and ventromedial nuclei, lateral hypothalamic area, posterior hypothalamus, tuberomammillary nuclei, and the bed nucleus of the stria terminalis. In addition, there are extrahypothalamic projections from the paraventricular thalamic nuclei, lateral septal nucleus, pretectum, infralimbic cortex, zona incerta, and ventral subiculum. The majority of the afferent projections are excitatory, whereas GABAergic projections from the IGL, tuberomammillary, and arcuate nuclei are inhibitory.

SCN afferents utilize a variety of neurotransmitters in addition to acetylcholine as signals, not all of which are fully characterized. In turn, different signals act through various intracellular pathways to adjust clock timing. Notably, there appear to be one or more time domains in which the SCN is refractory or shows altered responses to the different signaling compounds at different times. For example, light and the RHT transmitter glutamate cause phase delays in the early night and phase advances in the late night. Studies have shown that glutamate acts via increased intracellular Ca^{2+}, mediated by activation of ryanodine receptors in the early night and via a cGMP-mediated signaling pathway in the late night. During dusk, the SCN exhibits an increased sensitivity to the hypnogenic pineal hormone melatonin, which appears to act via a protein kinase C pathway. Other putative signal pathways include cAMP, nitric oxide, and PIP2 and suggest that the SCN may engender its temporal sensitivity at the intracellular level.

C. Output

The efferent projections of the SCN are extensive. Dr. Robert Moore has described four primary projections. These include the following: (1) projections to the anterior paraventricular thalamus, ventral lateral septum, and bed nucleus of the stria terminalis; (2) a periventricular fiber system innervating a large portion of the medial hypothalamus from the preoptic region to the premammillary region; (3) a lateral thalamic projection to the IGL; and (4) projections to the posterior paraventricular thalamus, precommisural

nucleus, and olivary pretectal nucleus. The majority of these projections exhibit vasopressin or VIP immunoreactivity. In general, this restricted subset of projections receives a particularly dense innervation from the SCN and therefore is thought to serve as relays for the distribution of SCN information to other neural structures. The subparaventricular zone (SPVZ), a region ventral to the paraventricular nucleus of the hypothalamus, is another structure receiving dense efferent SCN projections. The SPVZ has parallel projections to most of the same sites as the direct efferents of the SCN. With the exception of the SCN commissural connections, the efferents of each nucleus are ipsilateral.

The efferent projection to the preoptic hypothalamus is of particular importance because this area is involved in the regulation of sleep–wake, reproduction, fluid homeostasis, and thermoregulation. The projection to the PVN is part of a multisynaptic pathway regulating melatonin synthesis. Evidence suggests that the SCN, via the PVN, may influence the intermediate nucleus of the solitary tract (NTS), a structure that integrates gustatory, respiratory, and cardiovascular information. How efferent projections from the SCN convey time of day information to other neural networks is largely unclear.

The SCN also has a humoral output, which is known to include a daily rhythm in levels of the neuropeptide arginine vasopressin (AVP). Cerebrospinal fluid levels of AVP are high during the subjective day and low during the subjective night in both nocturnal and diurnal animals. These rhythms also persist *in vitro*. Rhythms of other neuroactive peptides synthesized in the SCN have been measured in the cerebral spinal fluid (CSF) and *in vitro* as rhythms in both mRNA and protein production.

The circadian and noncircadian visual systems may, in fact, interact. Except for the sharing of retinal photoreception, the nature of this interaction is not well-understood. Interconnections of these pathways occur between the superior colliculus and ventrolateral geniculate leaflet and the IGL.

III. MOLECULAR MECHANISMS OF THE CIRCADIAN CLOCK

A. Molecular Biology of the Pacemaker

Until the 1990s, very little was known regarding the intracellular mechanisms in the pacemaker responsible for the generation of circadian rhythms. The first indication that the circadian pacemaker had a genetic underpinning occurred during a selection experiment by Drs. Konopka and Benzer. In their classic experiment, the frequency of pupal hatching in *Drosophila* was studied, leading to the identification of the first single-gene circadian clock mutant, the *Period* (*Per*) gene. To date, three mammalian orthologs, homologs of the fly *Period* gene, have been identified (*mPer1, mPer2*, and *mPer3*). Shortly after the discovery of the *Period* gene, several other mutants exhibiting altered or eliminated circadian rhythm were discovered, including *timeless* (*tim*) in *Drosophila* and *frequency* (*frq*) in *Neurospora*. The first mammalian clock gene, *Clock*, was discovered by Joseph Takahashi using a phenotype-driven *N*-ethyl-*N*-nitrosourea (ENU) mutagenesis screen. The *Clock* gene was identified by positional cloning followed by a functional transgenic BAC rescue, which served to confirm that the *Clock* gene was in fact responsible for the mutant phenotype. More recently, two additional mammalian *Clock* genes, the *Cryptochrome* genes (*mCry1* and *mCry2*), have been described.

The *Clock* gene regulates two properties of the circadian clock: the endogenous circadian period and the persistence of rhythmicity. The protein encoded by the *Clock* gene, CLOCK, is a basic helix–loop–helix (bHLH) PER–ARNT–SIM (PAS) protein that belongs to a family of transcription factors. Shortly after the cloning of *Clock*, a CLOCK-interacting protein, BMAL, was identified using a yeast two-hybrid system. The CLOCK–BMAL heterodimer acts in a *trans*-activating fashion on E-box elements, which contain bHLH DNA-binding domains. Notably, both *Period* and the *Cryptochromes* contain an E-box sequence in the promoter region. By using a luciferase gene reporter assay, it was shown that the CLOCK–BMAL1 heterodimer, via transactivation of the E-box enhancer, drives the positive component of *Per* and *Cry* transcriptional oscillations. In addition, the CLOCK–BMAL1 heterodimer is thought to activate transcription of arginine vasopressin, a primary output marker for the circadian clock. It is not currently known what other *Clock* output genes are controlled by CLOCK–BMAL1; however, it is likely there are many. Both PER and CRY proteins block the ability of the CLOCK–BMAL1 dimer to activate *Cry* and *Per* promoters via the E-box, and, thus, a simple negative feedback loop begins to emerge. The rhythmic expression of PER proteins is thought to constitute the output of the oscillator, leading to the expression of *Clock*-controlled genes (Fig. 4). More recently, it was shown that, in mice, the two *Cryptochrome* genes were

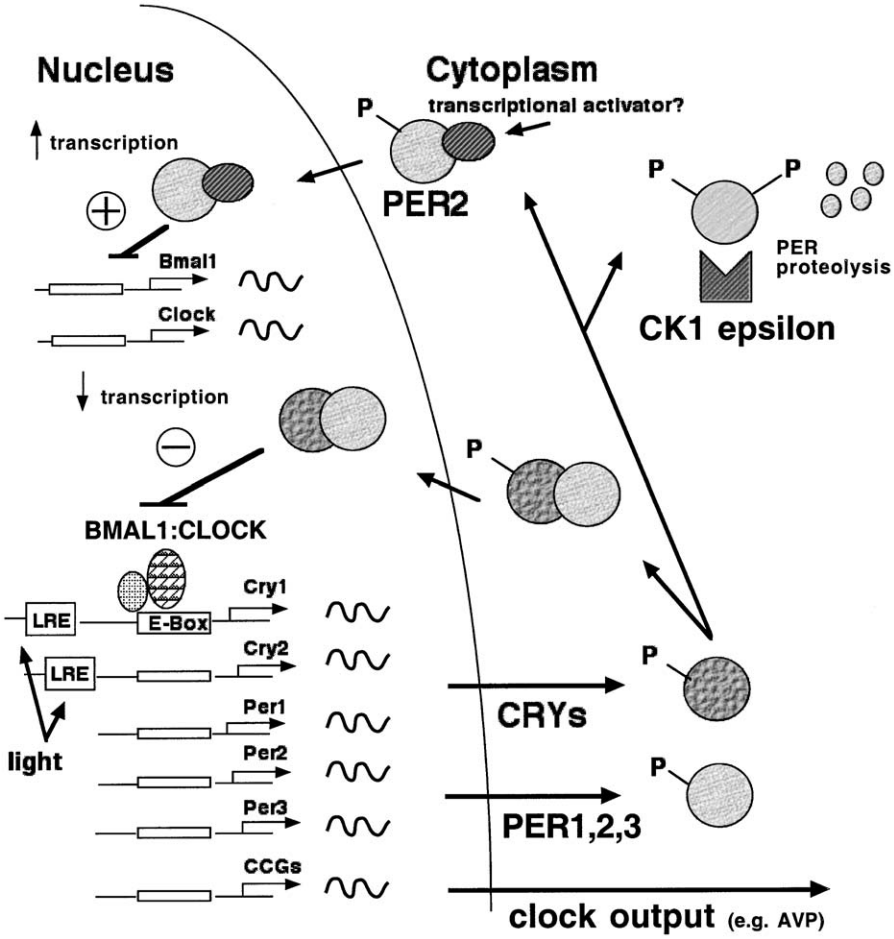


Figure 4 An illustration of the contemporary model for the autoregulatory transcriptional–translational feedback loop thought to constitute the endogenous time-keeping mechanism. The relationship between variables is described in the text. [Modified from Lowrey, P. L., Shimomura, K., Antoch, M. P., Yamazaki, S., Zemenides, P. D., Ralph, M. R., Menaker, M., and Takahashi, J. S. (2000). Positional syntenic cloning and functional characterization of the mammalian circadian mutation tau. *Science* **288,** 483–492.]

essential components of the negative limb of the circadian clock feedback loop. It now appears that the Cryptochrome proteins, CRY1 and CRY2, play an important role in the translocation of PER back to the nucleus, where PER disrupts CLOCK–BMAL1 transcriptional activity. A mammalian circadian mutant, *tau*, has been described. The *tau* mutation is a semidominant autosomal allele, which shortens the period length of circadian rhythms in Syrian hamsters. It was determined that the *tau* locus encodes casein kinase I epsilon (CKIe). CKIe phosphorylates PER proteins, which targets them for degradation. CKIe is the only enzyme identified so far in the mammalian pacemaker and, thus, may be a potential target for pharmaceutical compounds. Interestingly, the *tim* gene does not appear to have a role in mammalian circadian time-keeping.

How do light during the night and behavioral events during the day produce delays and advances? Evidence suggests that both types of shifts are mediated by changes in the levels of the state variables of the clock. For example, it now appears that PER1 protein and mRNA are rapidly down-regulated during behavioral resetting and that PER1 protein and mRNA are rapidly induced after light exposure. Whereas it is clear that much remains to be discovered about the mechanisms by which the clock loop is modulated, it is known that light appears to differentially regulate the *Per* gene. Future studies utilizing conditional transgenic and knockout mice in which genes are altered in a development- or tissue-specific manner should prove to be powerful tools for the dissection of the mammalian circadian clock mechanisms.

B. Molecular Basis of Entrainment to Light

In 1989, several groups of scientists independently reported that light could regulate the expression of the protoncogene *c-fos* in the SCN of rats and hamsters. The protein (c-Fos) belongs to the leucine zipper family of DNA-binding proteins and forms transcriptional regulatory complexes with other binding proteins. That light could physiologically regulate *c-fos* expression suggested that transcriptional control, via transcriptional regulatory proteins, was an important part of photoentrainment and might contribute to the mechanisms of circadian time-keeping. JunB appears to have a temporal pattern of expression similar to that of c-Fos and, thus, may also have an important role in light entrainment.

Following a light pulse administered during the subjective night in constant darkness, both *c-fos* mRNA and c-Fos protein are dramatically elevated in the retinorecipient areas of the SCN. This photoinduction of *c-fos* and c-Fos does not occur during the subjective day and, thus, photoinduction, like phase shifting, is phase-dependent. Circadian phase dependency of *c-fos* and c-Fos expression persists *in vitro* in SCN tissue slices. Thus, the gating of *c-fos* and c-Fos induction does not appear to require an intact retina or other neuronal structure. The mechanism that gates the photoresponsiveness of *c-fos* expression is unknown. However, phase dependency of c-Fos induction implies that the *c-fos* gene is clock-controlled. In contrast, photic stimulation of *c-fos* and c-Fos in another important CTS structure, the IGL, is not dependent on the circadian phase. The response to light in the SCN includes rapid and transient peak expression of mRNA after 30 min and peak expression of immunoreactive c-Fos protein 1–2 hr after onset of light administration.

There is a strong correlation between the photic induction of *c-fos* and the magnitude of phase shifts in behavioral rhythms. Further, the illumination threshold for gene expression is identical to the threshold for phase shifts. This correlation does not apply at high light intensities or with long photoperiods. Experimentation has indicated that about 20% of the total SCN neuron population consists of photoinducible c-Fos cells. These cells are located in the ventrolateral portion of the SCN. The exact mechanism of *c-fos* induction in the SCN remains unclear. However, it appears that an initial intracellular elevation of Ca^{2+} and cyclic AMP results from glutamanergic stimulation from the RHT. This then leads to phosphorylation of Ca^{2+}–cAMP response element-binding protein (CREB), an important transcription factor.

Only when Fos proteins are complexed as heterodimers, particularly with *jun* gene family members (c-Jun, JunB, JunD), will binding to the DNA regulatory element, AP-1, occur. AP-1-binding complexes may have constant or variable protein components, and it has been proposed that light, via c-Fos activation, alters the protein composition of the AP-1-binding complex, a regulator of transcription, altering its stability and binding affinity. The change in binding activity alters the transcription of SCN genes that have AP-1 sites on their promoters. To date, however, the *trans*-activating and *trans*-repressing activities of the various heterodimer complexes are unknown, and the identities of the AP-1-dependent genes remain to be elucidated. Exactly how the IEGs modulate *Clock* genes is not understood.

C. Molecular Basis of Pacemaker Cell Coupling

It is not understood how the activity of autonomous pacemaker cells of the SCN is synchronized to form a unified endogenous oscillator. Unfortunately, empirically distinguishing between endogenous pacemaker components and synchronizing mechanisms is very difficult. Extracellular ion fluxes, small membrane-diffusible molecules, glial regulation, and neural adhesion molecules have been examined as possible synchronizers of SCN neurons. Synaptic transmission and calcium-dependent neurotransmitter release are not essential for circadian time-keeping or pacemaker neuron synchronizing. Nitric oxide, a small membrane-diffusable molecule, is a promising candidate for a synchronizer. However, the fact that nitric oxide synthase is calcium-dependent is a confounding point. Previously, GABA was thought to underlie the generation of circadian rhythmicity via a rhythm in GABA equilibrium potential. However, it has been shown that antagonists to GABA neurotransmission do not prevent circadian rhythms in multiunit activity *in vitro*. Glial processes, gap junctions, adhesion molecules, and regions of membrane appositions have also been proposed as potential mechanisms of intercellular synchronization.

IV. MELATONIN AND BODY TEMPERATURE

Both the body temperature (T_b) and melatonin rhythms are robust and relatively easily measured in mammals. Consequently, both body temperature and

plasma melatonin are routinely utilized as indicators of the phase, amplitude, and period of the circadian clock. As such, it is instructive to examine both rhythms.

A. Melatonin

Melatonin, the "hormone of the night," has received much attention over the past few decades. Circulating melatonin levels at night are elevated by about 10-fold relative to during the day. Physiologically, melatonin acts as a humoral signal to provide a highly accurate indicator of night length. In effect, melatonin acts as an endocrine code for photoperiod information. Melatonin can also modulate the endogenous clock and influence a number of physiological functions. It is also thought to be hypnogenic. In certain vertebrates, lengthening or shortening of the nocturnal melatonin signal can alter reproduction, body weight, pelage, and seasonal behavior.

Melatonin is produced from serotonin in two enzymatic steps. Serotonin is first converted to *N*-acetylserotonin by arylalkamine *N*-acetyltransferase (AANAT), which, in turn, is converted to melatonin by hydroxyindole-*O*-methyltransferase (HIOMT). This conversion takes place in the pineal gland and is regulated by β-adrenergic receptor-mediated regulation of AANAT activity. The pineal gland is located in the epithalamus adjacent to the habenular nuclei. Pineal melatonin synthesis is driven by the SCN via a rather elaborate neural pathway. SCN efferents project to cells in the paraventricular nucleus, which in turn send projections to the intermediolateral cell column (IML) of the spinal cord. The IML projects to the superior cervical ganglia cells, which in turn send noradrenergic sympathetic projections to the pineal gland.

In mammals, the synthesis of melatonin is regulated by the SCN. In turn, melatonin can regulate the phase of the clock in a time-of-day-dependent manner via G-protein-coupled melatonin receptors in the SCN. Additional melatonin receptors are located in the pars tuberalis of the adenohypophysis. Melatonin is thought to act via the pars tuberalis to effect changes in gonadotrophin secretion and the reproductive axis. More recently, melatonin receptors have been discovered in the testis, epididymis, vas deferens, prostate, ovary, and mammary glands of various mammals, showing the multiple sites of action on the reproductive system.

Melatonin is a popular output marker for the circadian pacemaker because melatonin production and levels are not influenced by many exogenous factors. In humans, only very bright light and β-adrenergic drugs are known to suppress melatonin production. Notably, the PRC to melatonin for humans is about 12 hr out of phase with the PRC to light. Exogenous melatonin administration thus has been used clinically, both alone and in concert with bright light exposure, to shift circadian rhythms. Appropriate scheduled melatonin administration has also been used to treat circadian sleep and mood disorders, shift-work adaptation, and jet lag.

B. Body Temperature

The daily oscillation of body temperature in mammals constitutes one of the most regular and predictable of mammalian circadian rhythms. Historically, the body temperature rhythm is of unique importance in the circadian field. In addition, the T_b rhythm is also easily measured, coupled to numerous other body rhythms, and stable. Whereas the T_b rhythm integrates many other aspects of physiology, for the most part it is generated endogenously. The overt or expressed rhythm is a complex product of physiological integration among many control systems, including the thermoregulatory, cardiovascular, sleep–wake, and circadian systems. In addition, the endogenous temperature rhythm can be modified by factors such as activity, feeding, ambient temperature, and light. However, in a controlled environment, the T_b rhythm can provide an extremely accurate picture of the underlying clock activity.

The influence of the CTS on body temperature homeostasis is well-documented. The daily rhythm of body temperature is generated by the relative phasing of the heat production and heat loss rhythms in both entrained and free-running subjects. Changes in the ambient temperature produce coordinated and compensatory changes in heat production and heat loss; however, their phase relationships with the T_b rhythm are conserved. As with all circadian rhythms, the temperature rhythm is controlled by a circadian oscillator. Interestingly, this circadian oscillator may reside in another nucleus separate from the SCN. Some studies support the hypothesis that the thermoregulatory system is controlled by two or more circadian oscillators. The concept of multioscillator control of body temperature is old and is derived, in part, from studies in which the body temperature rhythm persisted following SCN lesions. However, other studies of similarly placed lesions have evidenced a highly

attenuated or absent body temperature rhythm. These findings suggest that a second (or more) oscillator would likely be found very close to the SCN, perhaps in the hypothalamic preoptic area or subparaventricular zone.

V. SLEEP–WAKE CYCLES

According to sleep researcher Alan Hobson, "Sleep is characterized by a recumbent posture, a raised threshold to sensory stimulation, decreased motor output and a unique behavior, dreaming." Sleep is, ironically, one of the least well-understood biological phenomena, yet over one third of our lives are spent in this behavioral state. Various hypotheses of sleep function and the functional significance of sleep have been proposed. These hypotheses have led to the development of a variety of sleep theories, each with evidence of varying quality to support it. For example, there is a restorative sleep theory, an energetic sleep theory, a dream and REM theory, a thermoregulatory theory, an immune theory, and a ethological theory. Whereas each theory most likely holds some truth, a unified sleep theory remains elusive.

The specific physiological processes that control sleep–wake behavior are fairly well-characterized. Collectively, these processes interact to promote and inhibit both sleeping and waking. The nature of these interactions has been described in a number of models for human sleep regulation, of which a circadian consolidation model is one of the most widely accepted at this time. In this model, sleep is timed by the CTS to provide a longer period of uninterrupted sleep. A two-process model of sleep regulation also exists that describes the interaction between a homeostatic sleep drive and a circadian rhythm of alertness. Work in sleep regulation has uncovered a role for the SCN in both the timing of the sleep–wake cycle and the regulation of the internal structure of sleep (Fig. 5). When the SCN was ablated in squirrel monkeys, a loss of sleep–wake consolidation was seen. Furthermore, circadian rhythms in sleep–wake, sleep stages, brain temperature, and drinking were eliminated, and total sleep time was significantly increased in SCN-lesioned monkeys. However, total times in deeper stages of non-rapid eye movement (non-REM; e.g., δ sleep) and REM sleep were not significantly affected by SCN lesions. It appears that the circadian timing system promotes waking and alertness while interacting with a homeostatic sleep drive to promote consolidated sleep. In effect, the interaction between the output (amplitude) of the circadian pacemaker and the homeostatic sleep drive is the primary determinant of vigilance and sleep–wake consolidation.

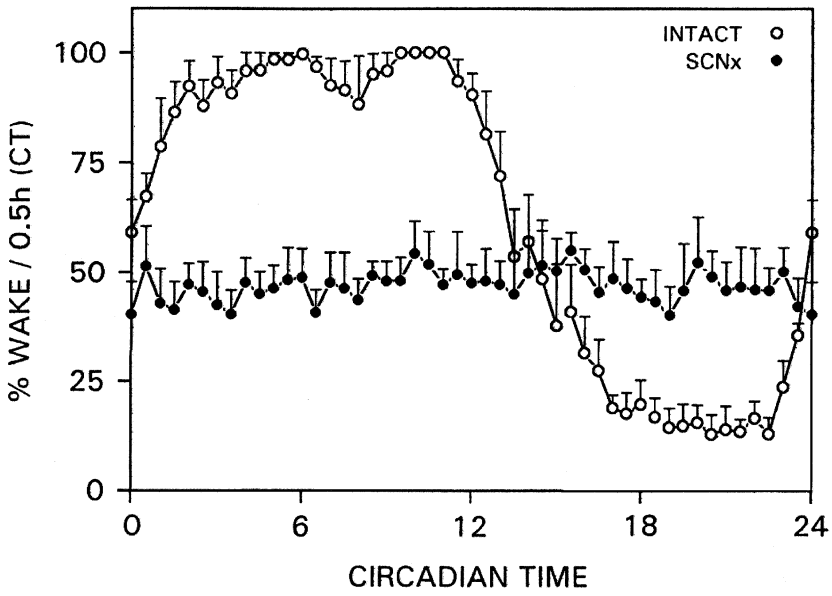


Figure 5 Effects of SCN lesion on sleep–wake behavior in the squirrel monkey. This diurnal nonhuman primate is similar to that of humans in sleep–wake consolidation pattern and sleep architecture. SCN-lesioned monkeys exhibited significantly greater wake and sleep bout counts and a loss of the daily sleep–wake cycle. [Reproduced by permission from Edgar, D. M., Dement, W. C., and Fuller, C. A. (1993). Effect of SCN lesions of sleep in squirrel monkeys: Evidence for opponent processes in sleep–wake regulation. *J. Neurosci.* **13,** 1065–1079.]

The circadian rhythm of sleep and waking is tightly coupled to the circadian rhythm of body temperature. Sleep onset occurs on the descending limb of the body temperature cycle. Notably, studies have shown that human subjects in isolation will enter long sleep episodes only at or near the minimum body temperature. In these studies, entry into sleep was not seen at other circadian phases of body temperature. Experimental paradigms utilizing different photoperiods have allowed investigators to artificially separate the sleep–wake and body temperature rhythms. This process is called forced internal desynchronization and results from the inability of one rhythm, usually that of T_b, to entrain to the LD cycle. This rhythm then becomes free-running, whereas the other remains entrained to external time.

VI. CLINICAL RELEVANCE

A. Circadian Susceptibility

Circadian rhythmicity is a significant source of variation in many important diagnostic variables (Fig. 6). Taking circadian variation into account can only improve the precision and validity of any clinical test. Furthermore, circadian rhythmicity can substantially affect both susceptibility to trauma and toxins and drug effectiveness and toxicity. Examples include cancer chemotherapy, anesthetics and analeptics, corticosteroid therapy, antiasthmatics, cardiovascular medications, antibiotics, and anabolic steroids. From a pharmacological perspective, a circadian variation in therapeutic response and efficacy is of interest. The effectiveness of a drug is largely a function of the rates of absorption, metabolism, excretion, and, finally, target susceptibility. Documented circadian variations in heart rate, acid secretion, glomerular filtration, renal plasma flow, urine production, pH, gastric emptying time, and blood pressure all, to a great extent, determine absorption, metabolism, excretion, and target susceptibility.

It is possible to increase the therapeutic efficacy and minimize the toxic side effects of drug treatments by providing treatment at a certain time of day. One medically relevant disease for chronotherapy is cancer, as many drugs used in chemotherapy affect the mitotic replication of normal and malignant cells. By treating the normal cells at times when DNA synthesis is low, higher levels of chemotherapeutic agents can be tolerated. Conversely, limitation of exposure to chemotherapy drugs during times when DNA synthesis is high can reduce the side effects. Moreover, cancer patients have been receiving adjuvant immunotherapy. Because both the humoral arm and the delayed (cellular) arm of the immune system function in a rhythmic manner, chronodiagnostics and treatments may prove to dramatically increase the efficacy of immunotherapy in cancer treatment. Figure 7A illustrates the effects of cyclophosphamide and 1-β-D-arabinofuranosylcytosine administered to mice previously given injections of L1210 leukemia cells. Mean survival time and cure rate exhibited a marked circadian phase dependency. For example, the cure rate was 94% in mice treated at one circadian phase but only 44% in those treated at another phase.

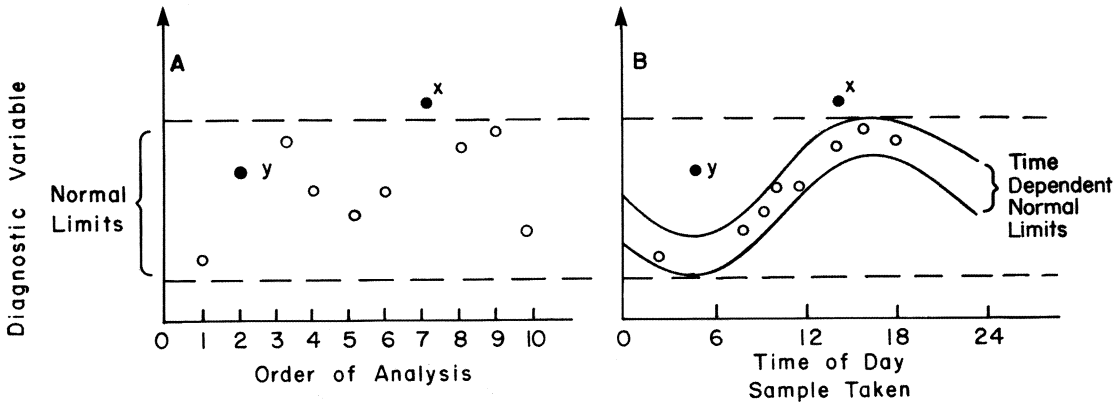


Figure 6 Normal limits for a diagnostically useful variable plotted (A) without and (B) with regard to the time of day. When the time of day is taken into account, the detection of abnormal values is improved, so that not only value *x* but also value *y* can be identified as being outside the normal range. [Reproduced by permission from Moore-Ede, M. C., Sulzman, F. M., and Fuller, C. A. (1982). *The Clocks That Time Us: Physiology of the Circadian Timing System*. Harvard University Press, Cambridge, MA].

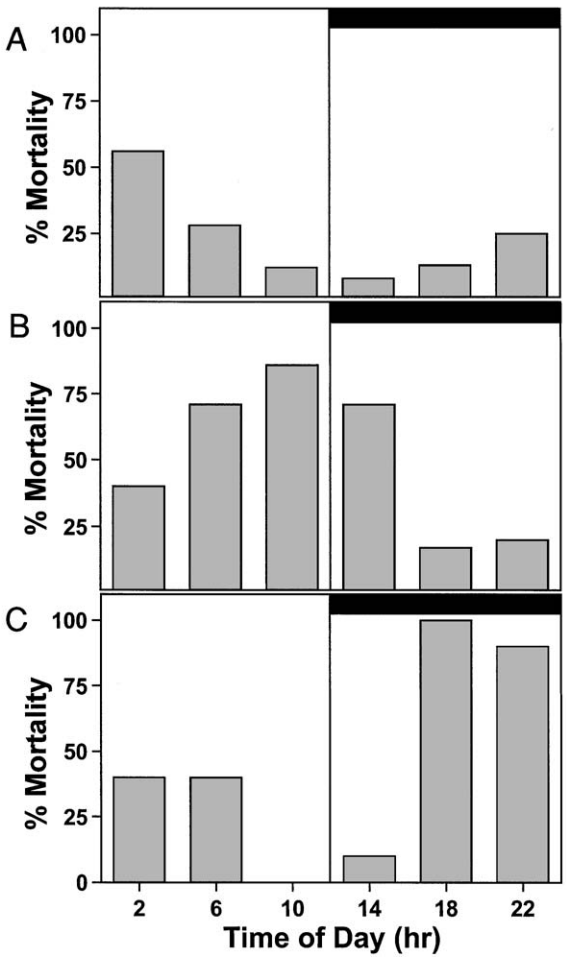


Figure 7 Circadian phase dependency for toxicity and efficacy of exogenous agents. Three different experiments demonstrate the marked circadian variation in the tolerance, susceptibility, and the rate of cure for (A) chemotherapuetic agents, (B) radiation, and (C) bacterial infection. [Modified from Moore-Ede, M. C., Sulzman, F. M., and Fuller, C. A. (1982). *The Clocks That Time Us: Physiology of the Circadian Timing System.* Harvard University Press, Cambridge, MA].

As previously mentioned, circadian variation in susceptibility has been documented for a number of exogenous challenges. Figure 7B illustrates a dramatic example of circadian susceptibility in mice injected with a bacterial endotoxin. Here, over 80% of the mice who received the toxin during the late subjective day died, whereas less than 20% of the mice who received the toxin during the mid-subjective night died. Figure 7C illustrates an even more dramatic example of circadian susceptibility with direct relevance to cancer therapy. In this study, mice were exposed to 550 R of whole-body X irradiation. Strikingly, 8 days after exposure all mice irradiated at the midpoint of the subjective night were dead, whereas all those exposed during the late subjective day were still alive. Similar studies have examined the minimum alveolar concentration (MAC) for effectiveness of inhalant anesthetics. Not surprisingly, both the concentration necessary to achieve an adequate plane of anesthesia and the lethal effects of the same drug vary widely and *predictably* over the 24-hr day. One area of clinical relevance that has not been adequately explored is homeostatic regulatory capacity, which can be severely compromised in conditions without temporal cues (Fig. 8). This condition unfortunately is very similar to the environment experienced by patients in the intensive care units (ICUs) of some hospitals.

The onset and symptoms of pathologies such as myocardial infarction, stroke, acute pulmonary embolism, sudden cardiac death, thoracic aortic rupture, and paroxysmal supraventricular tachycardia all exhibit marked circadian phase dependency. Thus, circadian time is an important parameter to be considered during the diagnosis and treatment of any patient.

B. Circadian Disorders

Disorders of the circadian timing system are more common than previously thought. Circadian dysfunction is thought to be a common contributing factor in both sleep–wake disorders and affective disorders. The etiological bases of most circadian disorders are not known. However, such problems are thought to result from either compromised pacemaker function or faulty signaling to effector systems, or from other factors. Circadian function disorders are often manifest in the elderly, the blind, and individuals with hypothalamic, pituitary, or optic tumors.

Individuals with circadian sleep disorders typically are only symptomatic when forced to conform to a societal schedule. The underlying circadian pacemaker is usually functional but either is unable to entrain or has a deficient capacity for phase delays–advances. The circadian sleep disorders include delayed and advanced sleep phase syndrome and non-24-hr sleep–wake syndrome. Certain forms of sleep–wake disorders such as irregular sleep–wake syndrome are linked to defective pacemaker function.

Shift-work sleep disorder constitutes a significant epidemiological problem. Some 8 million workers in the United States currently work a schedule that requires night, swing or rotating shifts. The night shift sleep–wake schedule is in direct opposition to the

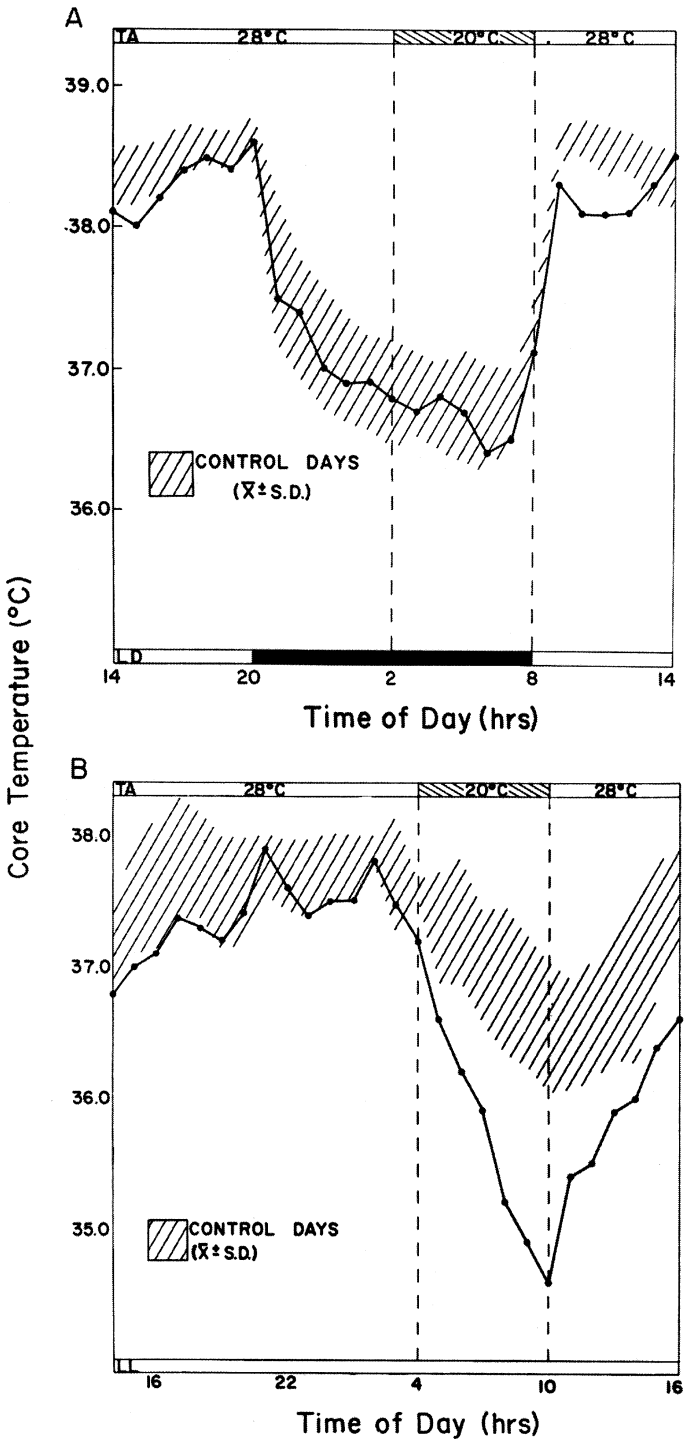


Figure 8 Thermoregulation is impaired in an environment with temporal cues. Effects of 6-hr exposures to cold (20°C ambient temperature) on body temperature in monkeys normally maintained at 28°C when (A) entrained to an LD 12:12 cycle and (B) free-running in constant light (lighting regimen indicated at the base of each graph). The shaded areas show the body temperature rhythm ($x \pm$SD) for the three previous control days, and the solid line is the body temperature on the day of the cold pulse. The cold pulse had virtually no physiological effect in LD but produced a significant fall in core body temperature when the animals were isolated from environmental time cues. [Reproduced by permission from Fuller, C. A., Sulzman, F. M., and Moore-Ede, M. C. (1978). Thermoregulation is impaired in an environment without circadian time cues. *Science* **199,** 794–796.]

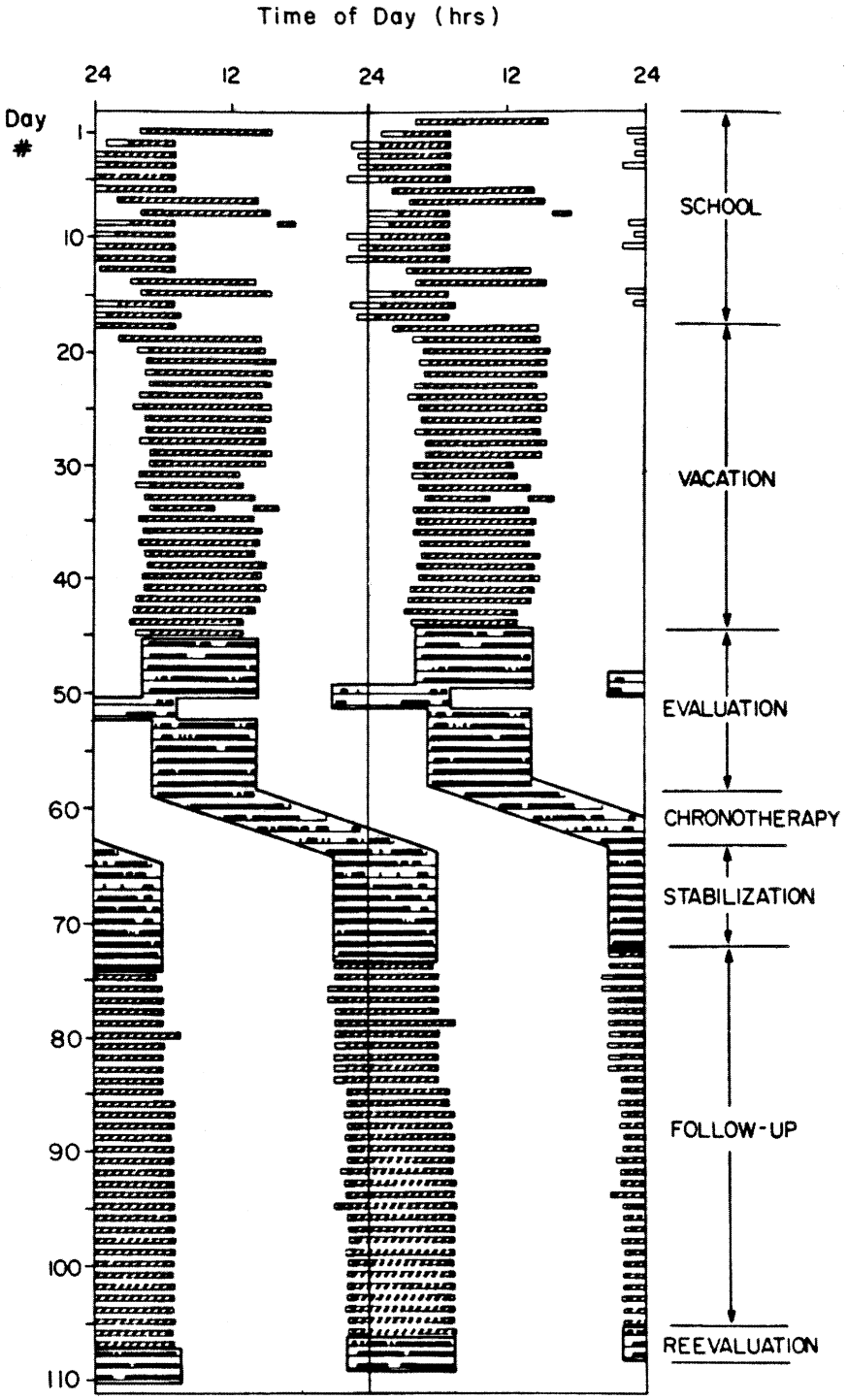


Figure 9 Sleep times as self-reported (striped horizontal bars) and via polygraphic recordings (dark horizontal bars) in an individual with delayed sleep phase insomnia. When attending school 5 days week, days 1–18, she lay awake in bed for several hours (open horizontal bars) before falling asleep but could sleep at a delayed phase with no difficulty on weekends (days 1, 7, 8, 14, and 15) and on vacation (days 19–45). On days 46–59, she was evaluated in the laboratory and on days 51 and 52 was subjected to an acute phase advance, which confirmed her inability to sleep at the desired time (2200–0600 hr). She was treated by chronotherapy on days 58–65 by phase-delaying her bedtime 3 hr each day until the desired phase relationship with the 24-hr clock was obtained. [Reproduced by permission from Moore-Ede, M. C., Sulzman, F. M., and Fuller, C. A. (1982). *The Clocks That Time Us: Physiology of the Circadian Timing System.* Harvard University Press, Cambridge, MA].

sleep–wake schedule dictated by the circadian clock. Similar to jet lag, many workers report insomnia during their off hours and hypersomnolence during their work hours. Consequently, alterness and performance suffer on the job and place the worker at greater risk for sleepiness-related accidents. Most workers also complain of additional symptoms including gastrointestinal discomfort and malaise. It is thought that the extreme work shift schedules of workers are largely responsible for two infamous disasters, the grounding of the oil tanker Exxon Valdez in 1989 and the Chernobyl nuclear plant explosion in 1986. In a similar vein, a 1992 study demonstrated that nurses on shift-work schedules were up to 3 times more likely to misdiagnose and provide improper treatment than their daytime counterparts. Furthermore, a well-documented link exists between shift-work and both gastrointestinal and cardiovascular diseases. Often confounding the treatment of shift-work disorders is the fact that shift-workers tend to implement countermeasures in the form of caffeine, nicotine, alcohol, and sedatives.

Circadian dysfunction is also implicated in a number of affective disorders. One particularly notable example is *winter depression* or seasonal affective disorder (SAD). SAD is characterized by depression, lethargy, loss of libido, hypersomnia, weight gain, carbohydrate cravings, anxiety, and an inability to concentrate and focus. Onset of SAD occurs during the late autumn or winter and occurs 6 months out of phase in the northern and southern hemispheres. SAD patients experience a spontaneous remission in late spring or early summer. SAD is set apart from nonseasonal depression by three atypical symptoms: hyperphagia, carbohydrate craving, and hypersomnia. Bright light therapy has proven extremely effective in ameliorating its symptoms. Administration of a few hours of early morning or evening bright light (>2500 lx) can frequently help to resynchronize biological rhythms. This may be primarily due to the ability of light to entrain the daily melatonin rhythm or to suppress inappropriate daytime melatonin secretion.

Figure 9 demonstrates the efficacy of chronotherapy in an individual suffering from delayed sleep phase insomnia, a common sleep disorder.

C. Aging

The circadian timing system and circadian clock are altered in senescence. Circadian rhythms of body temperature, melatonin, corticosterone, and serum testosterone all show reduced amplitude in old rats. In addition, changes in the intrinsic period and phase angle of entrainment to light–dark cycles have been observed in rhythms of old hamsters. A dramatic age-related decrease in the amplitude of circadian rhythms has also been observed in a number of animal models. Moreover, the effects of age on activity rhythms can be reversed with SCN grafts from young animals. Because mRNA synthesis in mammals has been demonstrated to decrease with age, altered production of clock proteins may contribute to age-related changes. Aging also appears to alter the response to entraining agents such as light. For example, the light-induced increase in the number of c-Fos and JunB immunoreactive SCN cells is significantly attenuated in aged animals. More recently, it has also been suggested that reductions in polysialyted neural cell adhesion molecules on the surface of SCN cells may contribute to the aging-related deficits in circadian function.

VII. SUMMARY

Circadian rhythms touch upon every aspect of human biology. The future of circadian biology will rest heavily on the use of molecular biology techniques to dissect the biological time-keeping processes. For example, we will be able to ask and answer questions such as are circadian disorders and individual variations in period linked to polymorphisms of the clock genes? In addition, advances in circadian biology will allow physicians to develop more efficacious and rational therapies through a greater understanding of temporal variances in sensitivity. Furthermore, an increase our understanding of the "clocks that time us" will aid in the treatment of circadian-related disorders such as depression, sleep–wake disorders, SAD, aging-related circadian problems, and desynchrony (e.g., jet lag).

See Also the Following Articles

AGING BRAIN • DEPRESSION • EVOLUTION OF THE BRAIN • HOMEOSTATIC MECHANISMS • HYPOTHALAMUS • SLEEP DISORDERS • TIME PASSAGE, NEURAL SUBSTRATES

Suggested Reading

Amir, S., and Stewart, J. (1998). Conditioned fear suppresses light-induced resetting of the circadian clock. *Neuroscience* **86,** 345–351.

Czeisler, C. A., Duffy, J. F., Shanahan, T. L., Brown, E. N., Mitchell, J. F., Rimmer, D. W., Ronda, J. M., Silva, E. J., Allan, J. S., Emens, J. S., Dijk, D. J., and Kronauer, R. E. (1999). Stability, precision, and near-24-hour period of the human circadian pacemaker. *Science* **284,** 2177–2181.

Dijk, D. J., and Czeisler, C. A. (1995). Contribution of the circadian pacemaker and the sleep homeostat to sleep propensity, sleep structure, electroencephalographic slow waves, and sleep spindle activity in humans. *J. Neurosci.* **15,** 3526–3538.

Edgar, D. M., Dement, W. C., and Fuller, C. A. (1993). Effect of SCN lesions of sleep in squirrel monkeys: Evidence for opponent processes in sleep–wake regulation. *J. Neurosci.* **13,** 1065–1079.

Fuller, C. A., Sulzman, F. M., and Moore-Ede, M. C. (1978). Thermoregulation is impaired in an environment without circadian time cues. *Science* **199,** 794–796.

Hobson, J. A., Stickgold, R., and Pace-Schott, E. F. (1998). The neuropsychology of REM sleep dreaming. *Neuroreport* **9,** R1–R14.

Koller, M. (1983). Health risks related to shift work. An example of time-contingent effects of long-term stress. *Int. Arch. Occup. Environ. Health* **53,** 59–75.

Konopka, R. J., and Benzer, S. (1971). Clock mutants of *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* **68,** 2112–2116.

Liu, C., and Gillette, M. U. (1996). Cholinergic regulation of the suprachiasmatic nucleus circadian rhythm via a muscarinic mechanism at night. *J. Neurosci.* **16,** 744–751.

Lowrey, P. L., Shimomura, K., Antoch, M. P., Yamazaki, S., Zemenides, P. D., Ralph, M. R., Menaker, M., and Takahashi, J. S. (2000). Positional syntenic cloning and functional characterization of the mammalian circadian mutation tau. *Science* **288,** 483–492.

Miller, J. D., Morin, L. P., Schwartz, W. J., and Moore, R. Y. (1996). New insights into the mammalian circadian clock. *Sleep* **19,** 641–667.

Moore-Ede, M. C., Sulzman, F. M., and Fuller, C. A. (1982). *The Clocks That Time Us: Physiology of the Circadian Timing System*. Harvard University Press, Cambridge, MA.

Pittendrigh, C. S., and Daan, S. (1974). Circadian oscillations in rodents: a systematic increase of their frequency with age. *Science* **186,** 548–550.

Robinson, E. L., and Fuller, C. A. (1999). Endogenous thermoregulatory rhythms of squirrel monkeys in thermoneutrality and cold. *Am. J. Physiol.* **276,** R1397–R1407.

Takahashi, J. S., DeCoursey, P. J., Bauman, L., and Menaker, M. (1984). Spectral sensitivity of a novel photoreceptive system mediating entrainment of mammalian circadian rhythms. *Nature* **308,** 186–188.

Classical Conditioning

ROBERT E. CLARK
University of California, San Diego

GLOSSARY

alpha responses Nonassociative responses that follow the presentation of the conditioned stimulus. They are nonassociative in the same way that unconditioned responses are nonassociative in that they are innate reflexes to the presentation of a stimulus. However, unconditioned responses do not show much habituation. In other words, repeated presentations of the unconditioned stimulus will continue to induce unconditioned responses. Alpha responses quickly habituate, which means that after a few presentations of the conditioned stimulus, they no longer occur. Ideally, conditioned stimuli are chosen because they are neutral (i.e., do not initially elicit a response). However, sometimes it is preferable to use a particular conditioned stimulus even if it temporarily results in alpha responses.

amplitude The magnitude of the conditioned response or unconditioned response. For example, in eyeblink conditioning the amplitude is expressed as the distance the eyelid moves in millimeters. Accordingly, a conditioned response of 2 mm has a greater amplitude than a conditioned response of only 1 mm.

conditioned response A learned response that is elicited by the conditioned stimulus.

conditioned stimulus A stimulus that signals the unconditioned stimulus. Initially the conditioned stimulus does not cause a response, but eventually it elicits a conditioned response.

conditioned stimulus-alone test trial A conditioned stimulus-alone test trial is a trial in which the unconditioned stimulus is omitted, leaving only the conditioned stimulus presentation. Conditioned stimulus-alone test trials are sometimes used because the presentation of the unconditioned stimulus and subsequent unconditioned response can obscure or mask the conditioned response.

extinction training Extinction training can be presented following the acquisition of the conditioned response. Conditioned responses are formed by pairing a conditioned stimulus and an unconditioned stimulus. During extinction training only the conditioned stimulus is presented (i.e., the unconditioned stimulus is omitted). Extinction training will gradually result in the disappearance of the conditioned response. This is referred to as extinction "training" because the conditioned response is not forgotten but inhibited. In other words, extinction is an active process. Evidence that extinction training is different from forgetting comes from paired conditioned stimulus and unconditioned stimulus trials. This results in the rapid reinstatement of the conditioned response. The reinstatement occurs much more rapidly than the original acquisition of the conditioned

response. Therefore, the conditioned response is not forgotten but, rather, inhibited.

interstimulus interval Sometimes called the conditioned stimulus–unconditioned stimulus interval, the interstimulus interval is the amount of time between the onset of the conditioned stimulus and the onset of the unconditioned stimulus.

intertrial interval A classical conditioning trial begins with the onset of the conditioned stimulus and ends with the offset of the unconditioned stimulus. The intertrial interval is the amount of time between each individual trial and is usually an average amount of time. The exact time between trials is usually varied within a restricted range in order to prevent ''time'' from becoming a conditioned stimulus as it is in temporal conditioning.

latency The amount of time between one event and another event. For example, the latency of the conditioned response is the amount of time between the onset of the conditioned stimulus and the onset of the conditioned response. The latency of the unconditioned response is the time between the onset of the unconditioned stimulus and the onset of the unconditioned response.

pseudoconditioning Classical conditioning results in conditioned responses when an association forms between the conditioned stimulus and the unconditioned stimulus. However, sometimes responses that appear to be conditioned responses, in that they follow the presentation of a conditioned stimulus, result from experience with the unconditioned stimulus only and not because of an association between the conditioned stimulus and the unconditioned stimulus. This is known as pseudoconditioning.

spontaneous recovery Spontaneous recovery is related to the process of extinction. Extinction training will result in the disappearance of the conditioned response. However, when the subject is given further extinction training following a break, the conditioned response will initially reemerge or spontaneously recover. Continued extinction training will more quickly result in the disappearance of the conditioned response until spontaneous recovery no longer occurs.

unconditioned response An innate response that is elicited by the unconditioned stimulus.

unconditioned stimulus A stimulus that is signaled by the conditioned stimulus. The unconditioned stimulus always elicits an unconditioned response.

unconditioned stimulus-alone test trial In unconditioned stimulus-alone test trials, the conditioned stimulus is omitted, leaving only the unconditioned stimulus presentation. In many classical conditioning experiments, the amplitude of the unconditioned response is an important measure (e.g., as a test to evaluate if a manipulation such as a brain lesion affects the ability of the subject to make a completely normal response). However, in a subject that is emitting conditioned responses, it is not possible to obtain an accurate measure of the unconditioned response amplitude because of contamination from the presence of conditioned responses. Unconditioned stimulus-alone test trials allow an uncontaminated measure of unconditioned response amplitude in subjects that are emitting conditioned responses.

Classical conditioning is a basic form of associative learning in which the organism learns something about the causal fabric of the environment or, in an experimental setting, the relationship of stimuli. Stimuli can be arranged so that one stimulus provides the organism with information concerning the occurrence of another stimulus. This type of associative learning is most commonly referred to as classical conditioning, but it has also been termed Pavlovian conditioning, respondent conditioning, and conditioned reflex type I conditioning, or type S conditioning.

I. INTRODUCTION

In the most basic form of classical conditioning, the stimulus that predicts the occurrence of another stimulus is termed the conditioned stimulus (CS). The predicted stimulus is termed the unconditioned stimulus (US). The CS is a relatively neutral stimulus that can be detected by the organism but does not initially induce a reliable behavioral response. The US is a stimulus that can reliably induce a measurable response from the first presentation. The response that is elicited by the presentation of the US is termed the unconditioned response (UR). The term ''unconditioned'' is used to indicate that the response is ''not learned'' but, rather, it is an innate or reflexive response to the US. With repeated presentations of the CS followed by US (referred to as paired training) the CS begins to elicit a conditioned response (CR). Here, the term ''conditioned'' is used to indicate that the response is ''learned.''

The most well-known example of classical conditioning comes from the pioneering work of Ivan Pavlov (1849–1936) and his dogs. In this prototypical example, a bell (in reality the CS was usually a metronome or buzzer) is rung just before meat powder is placed on the dog's tongue. The meat powder causes the dog to salivate. Therefore, the meat powder acts as the US, and the salivation caused by the meat powder is the UR. Initially, the ringing bell, which serves as the CS, does not cause any salivation. With repeated pairings of the ringing bell CS and the meat powder US, the ringing bell CS causes the salivation to occur before the presentation of the meat powder US or even if the meat powder not presented. The salivation in response to the presentation of the ringing bell CS is the learned or conditioned response.

II. BRIEF HISTORY

In 1904, Ivan Pavlov was awarded the Nobel prize in medicine for his pioneering work on the physiology of

digestion. Pavlov primarily studied the process of digestion in dogs. This early research set the stage for the discovery of a learning process that would come to be known as classical conditioning. As early as 1880, Pavlov observed that sham feedings, in which food was eaten but failed to reach the stomach (being lost through a surgically implanted esophageal fistula), produced gastric secretions, just like real food did. He immediately understood that this phenomenon must involve the central nervous system and began a program to thoroughly evaluate the parametric features of this learned response.

Pavlov modified his preparation in order to simplify the forthcoming studies. Rather than measure gastric secretions, he began measuring salivation. The first of these studies involved showing the dog a piece of bread as the CS. Pavlov then noticed that in dogs with extensive training, even the act of an experimenter walking into a room could elicit salivation. This finding led to the discovery that a variety of stimuli could induce salivation if paired with meat powder. Accordingly, in a further simplification of his experimental procedures he began using the sound of a bell presented with the meat powder to elicit salivation. It is this preparation that has become synonymous with Pavlov and historical examples of classical conditioning. In fact, Pavlov has become synonymous with classical conditioning because the term Pavlovian conditioning can be used interchangeably with classical conditioning.

Initially, Pavlov referred to the conditioned response as a ''psychic secretion'' to distinguish this type of response from the unlearned physiological secretions that would later come to be known as the unconditioned response. In 1903, a student of Pavlov's published a paper that changed the term psychic secretion to conditioned reflex. The terms conditioned reflex and unconditioned reflex were used during the first two decades of the 20th century, during which time this type of learning was often referred to as ''reflexology.''

Although Pavlov is correctly credited with the discovery of classical conditioning, and with identifying and describing almost all the basic phenomena associated with this form of conditioning, it is worth noting that the phenomenon of classical conditioning was independently discovered by an American graduate student in 1902. Edwin B. Twitmyer made this discovery while finishing his dissertation work on the ''knee-jerk'' reflex. When the Patellar tendon is lightly tapped with a doctor's hammer, it results in the well-known knee-jerk reflex. Twitmyer's work required many tap-induced reflexes for each subject. Twitmyer, like Pavlov, noticed that eventually the mere sight of the doctor's hammer (the CS) could produce a knee-jerk reflex (the CR). This largely forgotten report was the first example of classical conditioning of a muscle reflex. The potential significance of this finding was not apparent to Twitmyer, and the work was never extended or cast in a theoretical framework as Pavlov had done.

Pavlov's work on classical conditioning was essentially unknown in the United States until 1906 when his lecture ''The Scientific Investigation of the Psychical Faculties or Processes in the Higher Animals'' was published in the journal *Science*. In 1909, Robert Yerkes, who would later become president of the American Psychological Association, and Sergius Morgulis published a thorough review of the methods and results obtained by Pavlov. Although these reports provided a true flavor of the potential value of classical conditioning, the method was not immediately embraced by psychologists. This changed when John B. Watson, who is widely regarded as the founder of a branch of psychology known as behaviorism, championed the use of classical conditioning as a research tool for psychological investigations. Watson's presidential address delivered in 1915 to the American Psychological Association was titled ''The Place of the Conditioned Reflex in Psychology.'' Watson was highly influential in the rapid incorporation of classical conditioning, as well as other forms of conditioning, into American psychology. In 1920, his work with classical conditioning culminated in the now infamous case of ''little Albert.''

Albert B. was an 11-month-old boy who had no natural fear of white rats. Watson and Rosalie Rayner used the white rat as a CS. The US was a loud noise that always upset the child. By pairing the white rat and the loud noise, Albert began to cry and show fear of the white rat—a CR. With successive training sessions over the course of several months, Watson and Rayner were able to demonstrate that this fear of white rats generalized to other furry objects. The plan had been to then systemically remove this fear using methods that Pavlov had shown would eliminate or extinguish the CR—in this case, fear of furry white objects. Unfortunately, little Albert, as he as historically come to be known, was removed from the study by his mother on the day these procedures were to begin. There is no known reliable account of how this experiment on classical conditioning of fear ultimately affected Albert B. Nevertheless, this example of classical conditioning may be the

most famous single case in the literature on classical conditioning.

In 1921, a popular textbook on conditioning changed the terms of conditioned and unconditioned reflex to the current terms of conditioned and unconditioned response. This broadened the concept of conditioning to include other behaviors that were not merely automatic reflexes. In 1927, the Anrep (a former student of Pavlov's) translation of Pavlov's *Conditioned Reflexes* was published, thus making all of his work available in English for the first time. The availability of 25 years worth of Pavlov's research, in vivid detail, led to increased interest in the experimental examination of classical conditioning—an interest that has continued to this day.

III. THE DISTINCTION BETWEEN CLASSICAL AND INSTRUMENTAL/OPERANT CONDITIONING

In classical conditioning, no contingency exists between the CR and the presentation or omission of the US. In other words, it does not matter if a CR is made or not; the presentation of the US still occurs and is experienced and processed by the subject. This has been termed stimulus-contingent reinforcement or the law of contiguity. In instrumental conditioning, in contrast, a contingent relationship is arranged between the subject's response and the presentation or nonpresentation of a reinforcer. In other words, instrumental conditioning involves response-contingent reinforcement or the law of effect.

The simplest forms of learning are known as nonassociative learning. Examples of nonassociative learning include habituation (a decrease in a response to repeated presentations of a stimulus) and sensitization (an increase in a response to repeated presentation of a stimulus). Associative learning is a more complex form of learning. Classical conditioning and instrumental conditioning are both examples of associative learning. In classical conditioning, an association is made between two stimuli, the CS and US. This association is manifested by the occurrence of a conditioned response. In instrumental conditioning, an association is made between a stimulus and the outcome of a response. In other words, the organism learns what responses are reinforced given a particular stimulus. Although classical conditioning and instrumental conditioning are both examples of associative learning, classical conditioning is generally viewed as the simpler form of learning. There are many reasons for considering classical condition to be a simpler form of learning. For example, classical conditioning has been demonstrated in organisms that are very low on the phylogenetic scale, such as planaria, slugs, and leeches. Second, the ability to be classically conditioned appears earlier ontogenetically than the ability to be instrumental conditioned. Successful classical conditioning has been reported for chick embryos, neonatal monkeys, goats, and dogs as well as human fetuses *in utero*.

IV. ORIGIN OF THE TERM CLASSICAL CONDITIONING

Often, it is possible to precisely identify the time and reason a new scientific term is introduced. Generally, this is possible because the author of the new term describes the reasons for introducing the term and justifies why the particular term was chosen. For example, the term instrumental conditioning was coined by Clark L. Hull to describe the type of learning in which the subject is "instrumental" in obtaining reinforcement. That is, the animal is the instrument, as in maze learning. Operant conditioning was chosen by B. F. Skinner because the subject must perform a behavioral "operation" to obtain reinforcement. In other words, the animal operates on a manipulandum in the environment, such as pressing a bar to obtain food. However, unlike instrumental and operant conditioning, there does not appear to be an instance in which a single individual coined the term classical conditioning. Therefore, only an inference can be made concerning the origin of the term classical conditioning.

Only in the most rare cases do authors comment on the origin of the term classical conditioning. Often in these cases, the term is inaccurately attributed to John B. Watson. This is understandable because Watson was largely responsible for publicizing the classical conditioning method and outlining how it could be used as a method of scientific investigation. However, Watson never actually used the term classical conditioning, instead referring to it as simply the "conditioned reflex." His 1930 edition of *Behaviorism* was his last significant scientific publication, and although the process of classical conditioning is referenced extensively, the term was never used.

Before the 1940s, the process of classical conditioning was referred to most often as the conditioned reflex, Pavlovian conditioning, or simply conditioning.

However, in the 1930s, scientists began to understand that the laws governing learning, in paradigms in which reinforcement was contingent on the organism's behavior, appeared to be fundamentally different from the laws governing the conditioned reflex. The latter type of learning would come be known as instrumental or operant conditioning. This created a need to distinguish these different forms of conditioning. It is my contention that the term classical conditioning developed as a contraction of the descriptive phase "classical Pavlovian conditioning" that was used to denote the "well-known" (i.e., classical) type of conditioning used by Pavlov (i.e., Pavlovian).

V. TYPES OF CLASSICAL CONDITIONING

Classical conditioning is a generic term that can refer to a variety of different types of classical conditioning procedures and paradigms. Two different parameters can distinguish the type of classical conditioning: (i) the temporal arrangement and spacing of the CS and the US and (ii) the type of response that is measured and conditioned. Some general comments regarding the influence that different types of conditioning have on behavior can be made, but it is impossible to make any specific comments regarding conditioning. The specifics depend on the response system being measured, the nature of the stimuli being used, and the species being conditioned.

A. Temporal Arrangement and Spacing of the CS and the US

The relationship of the CS and the US can be arranged in different ways to produce different conditioning paradigms. These arrangements can greatly influence how the CR develops. Figure 1 illustrates the major classical conditioning paradigms. The upward movement of a line represents the onset of a stimulus, and the downward movement of a line represents the offset of a stimulus.

1. Simultaneous Conditioning

In simultaneous conditioning the CS and the US are presented simultaneously. In this case, because the US always elicits a UR and the CS and US are presented together, it is not possible to determine if CRs are present. In order to determine if a CR has developed, a CS-alone trial must be presented (i.e., the US is omitted). If the CS elicits a response on the CS-alone trial, this response is a CR. This form of conditioning is not commonly used, and it generally yields only weak conditioning, if any at all.

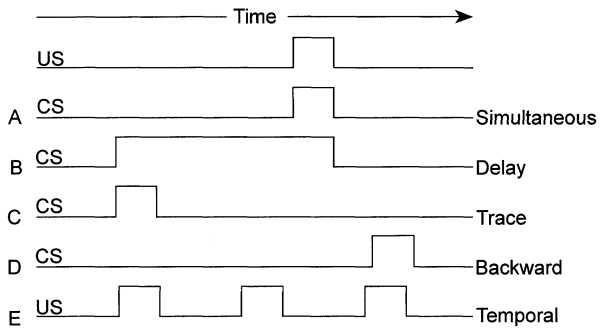


Figure 1 Temporal arrangements of the CS and the US (top trace) used in five classical conditioning paradigms. The upward movement of a line represents the onset of a stimulus, and the downward movement of a line represents the offset of a stimulus. A, simultaneous conditioning; B, delay conditioning; C, trace conditioning; D, backward conditioning; E, temporal conditioning. In the case of temporal conditioning, there is no discrete CS. The interval of time between the USs serves as the CS.

2. Delay Conditioning

In delay conditioning, the CS onset precedes the US onset. The termination of the CS occurs with the US onset, during the US, at the termination of the US, or at some point after the US. This paradigm is called delay conditioning because the onset of the US is delayed relative to the onset of the CS. Generally, responses that develop to the CS and occur before the onset of the US are CRs. This is the most common conditioning paradigm and generally results in the most robust and rapid conditioning.

3. Trace Conditioning

In trace conditioning, the CS is presented and terminated before the onset of the US. The interval separating the CS offset and the US onset is called the trace interval. This paradigm was named trace conditioning by Pavlov because in order for conditioning to occur, the subject (i.e., the subject's brain) must maintain a memory "trace" of the CS. Responses that develop in response to the CS and occur before the onset of the US are CRs. This form of conditioning is

also common and yields good conditioning but generally not as readily as delay conditioning.

4. Backward Conditioning

In the backward conditioning paradigm, the US is presented and terminated prior to the CS. A CR is a response that follows the presentation of the CS. This form of conditioning is not commonly used and in most circumstances does not result in conditioning. However, in some situations backward conditioning can occur.

5. Temporal Conditioning

In temporal conditioning there are no discrete CSs. Instead, the US is presented at regular intervals, and over time the CR will be exhibited just prior to the onset of the US. In this case, the CS is the time interval. Temporal conditioning is possible in some experimental paradigms, but in most classical conditioning paradigms it does not result in conditioning.

6. Differential Conditioning

In differential conditioning, two CSs are used. One of the CSs always precedes and predicts the US. This CS is termed the positive CS or the CS^+. The other CS is not predictive of the US and occurs alone. This CS is termed the negative CS or the CS^-. Differential conditioning is indicated by a greater number of CRs in response to the CS^+ than to the CS^-.

7. Controls for Pseudoconditioning

Classical conditioning results in CRs when an association forms between the CS and the US. However, sometimes responses that appear to be CRs, in that they follow the presentation of a CS, result from experience with US only and not because of an association between the CS and the US. This is known as pseudoconditioning. For example, if a very intense US is presented (e.g., a strong shock), the organism might respond to any subsequent stimulus presentation. The response does not occur because of an association between the CS and the US but, rather, because the US sensitizes the subject, making it more likely to respond to any stimulus presentation. To test for the possibility of pseudoconditioning, the CS and the US can be arranged so that the CS does not predict the US. This is known as unpaired training. There are two basic forms of unpaired training. Explicitly unpaired training is used to describe a situation in which the CS and the US never coincide. Random unpaired training is used to describe a situation in which the CS and the US rarely coincide but occasionally (i.e., by chance) do so. Random unpaired training is generally thought to be superior to explicitly unpaired training because during explicitly unpaired training the animal learns that the CS signals a safety period (i.e., the US will not occur).

Because the CS and the US are never paired, an association between them cannot be made. If responses nevertheless follow the presentation of the CS, it can be concluded that pseudoconditioning has occurred and not classical conditioning. If unpaired presentations do not result in responses to the CS, then any responses that subsequently develop in response to the CS, after paired training is begun, can be considered true CRs.

B. Types of Classical Conditioning Based on the Measured Response

In addition to the conditioning paradigms discussed previously, different types of classical conditioning can be characterized by the measured response. For example, in Pavlov's experiments, he measured salivation, whereas in Twitmyer's experiments, he measured the knee-jerk response. Although both of these basic experiments are examples of classical conditioning, they involve different response systems.

1. Two Fundamental Response Classes

The nervous system can be divided into the central nervous system (CNS) and the peripheral nervous system (PNS). The CNS consists of all the neural tissue that is encased in bone (i.e., the brain and spinal cord). The CNS is discussed later. Because this section focuses on responses, the PNS will be discussed. In order for a response to occur, the PNS must be engaged (excluding responses of neurons that can be recorded from the CNS). The PNS can be divided into the autonomic nervous system and the somatic nervous system. The autonomic nervous system controls the viscera, and the somatic nervous system controls muscles.

a. Autonomic Classical Conditioning Autonomic classical conditioning refers to any classical conditioning paradigm in which the measured response is under

the control of the autonomic nervous system. The autonomic nervous system consists of two divisions, the sympathetic nervous system and the parasympathetic nervous systems. These two systems work together, in opposing directions, to control bodily functions such as heart rate, breathing, dilation and constriction of the pupil, and the control of sweat glands. These systems are not generally under voluntary control but can be modified by classical conditioning. Autonomic conditioning has been used extensively because in addition to involving primarily involuntary responses, autonomic responses can also be used as an index of changes in emotion. Autonomic conditioning is sometimes referred to as conditioned emotional responses because changes in emotion are accompanied by changes in these autonomic measures. For example, many of these measures are also used in polygraph/lie-detection work. The following is a brief overview of the most common types of autonomic conditioning studies.

i. *Galvanic Skin Response/Skin Conductance Response Conditioning* This response is measured by the change in skin resistance to an electrical current. Research using this response measure dates back to the 1880s. For most of this time, the response was termed the Galvanic skin response (GSR) in honor of Luigi Galvani (1737–1798), an early pioneer in research describing the electrical nature of the body. Today, descriptive terms are commonly used, such as the electrodermal response or the skin conductance response (SCR). This response is measured by passing a small amount of electrical current between two electrodes pasted to the skin. The conductance (the reciprocal of resistance) between the two electrodes is measured. Many different stimuli can induce a SCR. For example, a mild shock, or any new stimulus of sufficient intensity as to attract the attention of the subject, can cause an autonomic response in which sweat glands pump extra sweat into sweat ducts located in the skin. The net effect of this action is an increase in skin conductance (i.e., a skin conductance response). In conditioning experiments in which, for example, a tone CS is paired with a shock or loud noise US, an association is made between the CS and the US and the CS begins to elicit a classically conditioned SCR.

ii. *Heart Rate Conditioning* Another common autonomic conditioning paradigm is the classically conditioned heart rate response. In this paradigm the change in heart rate (i.e., heartbeats per minute) is measured with electrodes pasted on the chest. For example, a tone CS can be paired with a mild shock US. Initially, the CS does not cause a change in the heart rate, but the shock will increase the heart rate. With continued pairing of the CS and the US, the CS will elicit a CR—a change in heart rate. Interestingly, the CR in this case is change in heart rate to the CS, but the direction of change (i.e., slower or faster) depends on the animal species being testing. For example, the CR with human subjects is an increase in heart rate, which is called conditioned tachycardia (heart rate increasing). However, the CR with, for example, rabbits, is a decrease in heart rate, which is called conditioned bradycardia (heart rate slowing).

iii. *Other Examples* Another example of autonomic classical conditioning is the conditioned pupillary response, which is a conditioned change in the size of the pupil. This response was first conditioned in 1922, but today it is rarely used because it is a difficult response to condition and to measure, and it is subject to a great deal of noise (i.e., spontaneous responses that are unrelated to the CS or the US). As noted previously, salivary conditioning was used by Pavlov in the initial demonstration of classical conditioning and this response was adapted for work with humans by Karl Lashley (1890–1958). Today, this paradigm is almost never used because salivation is a relatively slow response, difficult to measure, and difficult to condition in humans.

b. Somatic Classical Conditioning Somatic classical conditioning refers to any classical conditioning paradigm in which the measured response is under the control of the somatic nervous system. The somatic nervous system as it relates to classical conditioning controls striate or skeletal muscles. Accordingly, any response that requires a motor movement must be controlled by striate muscles and the somatic nervous system. The following is a brief overview of the most common types of somatic conditioning paradigms.

i. *Eyeblink Classical Conditioning* Eyeblink classical conditioning is by far the most common form of experimental conditioning paradigm with both humans and experimental animals. As early as 1922, eyeblink classical conditioning paradigms were being used. In eyeblink conditioning a CS (typically a tone or light) is paired with a US (e.g., a mild shock or puff of air to the eye). Eyeblink studies have been carried out in a variety of species, including humans, monkeys, rabbits, cats, rats, and mice. The response can be

measured with a minitorque potentiometer. A string is attached to the eyelid and any eyelid movement changes the resistance in the potentiometer, which can then be recorded. Electromyographic (EMG) measures can also be used to record the activity of the muscles controlling the eyeblink. Today, the most common method of measuring eyeblink responses in humans is the use of an infrared reflective sensor. In this method, an infrared beam is directed at the eye (usually mounted in goggles) and the amount of light reflected by the eyelid is recorded. Eyeblinks change the amount of light that is reflected and subsequently detected by an infrared sensor.

ii. *Nictitating Membrane Classical Conditioning* Closely related to eyeblink conditioning, conditioning of the nictitating membrane (NM) is the most frequently used response measure in rabbits, which are the most frequently used experimental subjects for conditioning studies. The NM is often called the "third eyelid" and it consists of a sheet of cartilage located behind the inner canthus of the eye. When the eyeball is stimulated, the eyeball retracts into the eye socket. This causes the NM to passively sweep across a portion of the eyeball. The NM response is popular because it is simple to measure (usually with a minitorque potentiometer) and because the NM cannot completely cover the eye. Thus, it prevents the subject from avoiding the airpuff US (the eyelids are prevented from closing by the experimenter during NM conditioning).

iii. *Leg Flexion Classical Conditioning* In the typical leg flexion experiment, an animal is usually restrained so that the legs hang freely, although it has also been used in freely moving subjects. A CS is paired with a mild shock US to the leg, which causes the leg to flex. With pairing of the CS and the US, a CR develops where the leg flexes in response to the CS. The response can be measured by a minitorque potentiometer or EMG.

iv. *Fear Classical Conditioning* Classical conditioning has been increasingly used to study the learning of fear. This paradigm can be considered a hybrid of autonomic and somatic classical conditioning because fear causes numerous autonomic changes, which could be measured as the CR. However, in the rat, the most common subject for studies of this type, fear can also be measured with the somatic response of freezing. In the typical paradigm, a tone CS is paired with a shock US. The shock US is delivered to the rat through an electrified floor grid. With pairing of the CS and the US, a fear CR develops in response to the CS. In this case, the fear CR is freezing (the rat holds completely still).

VI. ACQUISITION OF THE CR

As noted previously, the specifics regarding any aspect of classical conditioning must be reserved for the particular type of classical conditioning paradigm, the response that is measured, and the species that is used. However, some general features can be noted regarding the acquisition of the CR. The most fundamental element of classical conditioning is the association of the CS and the US that results in the acquisition of the CR. During the initial presentation of the CS and the US, no response to the CS is observed. With continued pairing of the CS and the US (i.e., presentation of additional training trials), CRs begin to develop in response to the CS. This development of the CR is referred to as the acquisition of the CR. It is often referred to as an increasing probability of CRs. In other words, with continued training, the probability of a CR on a given trial increases. In some conditioning paradigms (such as classical conditioning of the NM response in rabbits) this probability can approach 0.99 (i.e., CRs on almost every trial). During the early phases of CR acquisition, CRs are generally small-amplitude responses that gradually (over the course of training) grow larger until they become as large as the UR. The latency of the CR also tends to change during CR acquisition. Initially, the latency of the CR onset begins just before the onset of the US. With continued training the latency decreases so that the onset of the CR begins well before the onset of the US. Finally, the CR is acquired to the extent that the maximum amplitude of the CR (the CR peak) occurs at the time of the US onset. This is sometimes referred to as a "well-timed" CR.

A. Factors That Can Influence the Acquisition of the CR

1. Interstimulus Interval

Every type of conditioning paradigm has an optimal interstimulus interval (ISI). The optimal ISI depends on the particular paradigm, but it is clear that ISIs that are very short or very long and result in little or no CR acquisition. For example, in the rabbit classically conditioned NM response paradigm, the optimal ISI

for delay conditioning is 250 msec. An ISI of 100 msec or less results in poor conditioning, as does increasing the ISI to longer than 250 msec. In this paradigm, an ISI of 2000 msec (i.e., 2 sec) will not result in any CR acquisition.

2. Intertrial Interval

It is a general finding from many conditioning paradigms that increasing the ITI will tend to produce more rapid CR acquisition, although this influence is not robust in most cases. Reducing the ITI to less than several seconds can drastically impair acquisition of the CR. The typical ITI ranges from 30 to 60 sec.

3. Temporal Arrangement and Spacing of the CS and the US

As noted previously, delay conditioning generally results in the most rapid rate of CR acquisition followed by trace conditioning. Simultaneous conditioning generally results in much poorer CR acquisition. In many cases simultaneous conditioning (ISI=0) does not produce CRs. Backward conditioning results in CRs in only a few paradigms and generally does not produce any CR acquisition.

4. CS Intensity

It was initially believed that the CS intensity did not influence that rate of CR acquisition, but recent studies indicate that more intense CSs tend to result in slightly more rapid CR acquisition.

5. US Intensity

An increase in the acquisition rate of the CR has been a consistent finding with increases in the intensity of the US.

VII. EYEBLINK CLASSICAL CONDITIONING IN HUMANS

It was previously noted that different types of classical conditioning paradigms can be distinguished on the basis of the measures response. By the early 1930s, successful classical conditioning had been reported in 23 different response systems (e.g., eyeblink response, skin conductance response, pupillary response, leg flexion response, and salivary response). Of these numerous preparations, the majority were explored for only short periods before being, for the most part, abandoned for one reason or another. For example, the salivary preparation in humans or other experimental animals never flourished because of many methodological difficulties.

Classical conditioning of the skin conductance response has had a long history dating back to the late 19th century. However, skin conductance conditioning failed to flourish in the 1930s–1950s because the physiological basis of the response was poorly understood and difficult to measure accurately with the equipment of the day. A renewed interest occurred in the 1960s due to a better understanding of the physiology of the response and better and more readily available equipment for measuring and quantifying the response. Today, this paradigm still enjoys some popularity with researchers studying cognitive factors of classical conditioning and the neural mechanism underlying conditioned emotional responses.

By the 1940s the human eyeblink classical conditioning paradigm had surpassed all other conditioning paradigms in terms of number of articles published primarily because it was methodologically superior to all other classical conditioning paradigms. From the 1940s through the 1960s, studies of human eyeblink conditioning remained the dominant paradigm for studying the processes and variables that related to classical conditioning. During this time, methodological improvements that included precise measurement of the eyeblink response allowed investigators to explore the effects of such variables as the intertrial interval, interstimulus interval, CS intensity, US intensity, and variable reinforcement schedules. Additionally, cognitive factors were also extensively explored. For example, subjects were asked about the information they acquired during the conditioning session. This information was then compared to how well the subjects acquired the CR. Other studies explored how instruction sets affected CR acquisition rates: that is, how different verbal instructions given to subjects before conditioning affected CR acquisition rates.

VIII. PROBLEMS FOR STUDIES OF HUMAN EYEBLINK CLASSICAL CONDITIONING

In all forms of classical conditioning, the responses are considered primarily reflexive in nature. For example, the UR is an innate, automatic, ''reflexive'' reaction to

the puff of air that is delivered to the eye. With repeated pairing of a CS and a US a learned, conditioned response develops. This response is also thought to be reflexive and, as such, should not require cognitive involvement. Nevertheless, although an eyeblink response can be involuntary, clearly the eyeblink response can also be brought under voluntary control. If individuals become aware of the fact that the CS predicts the US, they are in a position to voluntarily blink their eyes to avoid the airpuff (i.e., blink on purpose). This circumstance has caused concern among experimentalists because it has been argued, at least since the 1940s, that the processes and characteristics of responses that are voluntary are very different from those that are involuntary or reflexive. That is, classical conditioning measures involuntary, automatic responses, not purposeful behavior. Therefore, some means of identifying voluntary responses needed to be developed so that they could either be discarded as contaminants or analyzed separately.

Two criteria emerged for identifying voluntary responses. The first was based on the slope of the response. Voluntary responses were believed to be more rapid and thus to have a characteristically steeper slope than a true CR. The second was based on response latency. Voluntary responses were believed to have a short onset latency (which also involved a steep slope), and eye closure was maintained until the onset of the US. Despite these improvements, there is no consensus about the most appropriate methods for detecting voluntary responses. In fact, the only large-scale study to test the validity of these two criteria (slope and latency) found that neither was fully satisfactory for discriminating voluntary responses from conditioned responses.

By the 1960s, studies of human eyeblink classical conditioning began to wane. This gradual decrease in human eyeblink classical conditioning research was likely due to a combination of at least two factors. First, researchers were never completely successful in identifying and dealing with the issue of voluntary vs conditioned responding. Second, researchers never reached a consensus on what the "standard" conditions should be for conditioning studies. Various laboratories used different CS and US intensities, different ISIs and ITIs, and different criterion for determining if responses were CRs. Consequently, there were persistent problems with obtaining reproducible results between different laboratories, which essentially prevented progress in exploring interesting variables.

IX. THE EMERGENCE OF ANIMAL STUDIES OF CLASSICAL CONDITIONING

Although human eyeblink classical conditioning studies began a steep decline in the 1960s, classical conditioning studies using animals showed rapid growth. This was in part likely due to at least two factors. First, some experimentalists moved from human conditioning research to working with animals, which served to stimulate the field of animal work. Second, although problematic, the human research on classical conditioning nevertheless revealed the potential of classical conditioning to serve as a paradigm for the systematic and thorough analysis of associative learning if it could be appropriately exploited. What was needed was the development of a "model" paradigm of classical conditioning in which all the basic features (e.g., CS and US types and intensities, measurement of the responses, and ISIs and ITIs) would be held consistent across laboratories. A model paradigm must result in robust acquisition of the CR, which is reliable across laboratories. This would allow theoretical questions about learning and memory to be addressed. Additionally, these methods must be economical and relatively easy to implement, and the characteristics of the learned response must not be unique to the experimental circumstances or to the species being tested.

In the early 1960s, Isidore Gormezano and colleagues developed a paradigm in which classical conditioning of the NM response was used with rabbits. The NM is vestigial in humans but is quite pronounced in the rabbit. It consists of a sheet of cartilage located behind the inner canthus of the eye. When the eyeball is stimulated, it retracts into the eye socket. This causes the NM to passively sweep across a portion of the eyeball. This movement is the measured response. The NM response was chosen because it is simple to measure (usually with a minitorque potentiometer) and because the NM cannot completely cover the eye. In this preparation, the rabbit's eyelids are held open with clips to prevent the subject from completely closing its eyelids and avoiding an airpuff US. Because the rabbit cannot avoid the airpuff US, the response cannot be an instrumental response but remains squarely within the domain of classical conditioning.

This paradigm remains the model classical conditioning paradigm. It has endured because it has simply proven to be an ideal paradigm. The result is that the processes and factors that influence the acquisition of the CR are now understood in detail. These details

allow learning theories to be constructed. This model system then allows the hypotheses that are derived from these theories to be rigorously tested and interpreted against an immense background of empirical data. In this respect, the importance of classical conditioning to modern learning theory cannot be overestimated.

X. EYEBLINK CLASSICAL CONDITION AS A TOOL TO STUDY BRAIN FUNCTION

In his preface to the Russian edition of his 1927 book *Conditioned Reflexes*, Ivan Pavlov wrote "At [this] time [I am] am convinced that this method of research [classical conditioning] is destined, in the hands of other workers, and with new modifications in the mode of experimentation, to play a yet more considerable part in the study of the physiology of the nervous system." Pavlov's statement can be viewed not only as prophetic but also as in fact quite understated. Until the past 25 years, Pavlov's enthusiasm for classical conditioning as a method of studying brain function had not been shared by physiologists. As a research procedure, classical conditioning had been almost the exclusive property of psychologists. However, beginning in the mid-1970s, classical conditioning became one of the most important tools for relating learning to specific brain structures.

XI. THE SEARCH FOR THE ENGRAM

The term engram is a hypothetical construct used to represent the physical processes and changes that constitute memory in the brain, and the search for the engram is the attempt to locate and identify that memory. Karl Lashley was perhaps the first person to clearly conceptualize the issue in a framework that would lend itself to experimental analysis. Lashley's primary interest was in localizing the engram to a specific region of the brain—an endeavor that would prove to be extraordinarily difficult and ultimately end in failure for Lashley. A likely reason why Lashley and others have historically had such great difficulty in locating an engram is because the behavior they typically used was an operant form of conditioning—a relatively complex form of associative learning. To make localization more tractable a simpler behavior was needed. Classical conditioning of the NM response in rabbits proved to be the solution.

For many reasons, Classical conditioning is a valuable tool for studying the physical processes of the brain that may be important for forming and storing memory. The presentation of the CS and the US is determined by the experimenter, not by the subject's behavior. This has important implications for the analysis of stimulus selection, which can be precisely manipulated by the experimenter. However, of greater importance is the fact that the CR is time locked to the CS. This allows a temporal analysis of neural events and then correlation of electrophysiological recordings with changes in the CR. The conditioning procedures provide a more adequate control for nonspecific effects of training on biological processes than do operant procedures. For example, the same kind of density of stimulation and number of unconditioned responses can be produced in both experimental and control conditions. Perhaps the greatest advantage is that the effects of experimental manipulations on "learning" rather than "performance" can be easily evaluated. This problem of learning versus performance has plagued the study of brain substrates of learning from the beginning. For example, does a brain lesion impair a learned behavior because it damages the memory trace or because it impairs the animal's ability to respond? This problem can be circumvented by simply comparing the amplitude of the learned or conditioned response to the reflex or unconditioned response amplitude. If the lesion abolishes the CR (the learned response) but leaves the UR (the ability to respond) intact, then it can be concluded that the lesion has impaired learning rather than performance.

The use of the classically conditioned NM response in rabbits offers additional advantages for studying the neural correlates of the learned behavior. Rabbits are docile animals that do not find full-body restraint aversive. They will usually sit quietly for more than 2 hr with little or no struggling. Rabbits also have a very low spontaneous blink rate, which reduces the amount of contamination caused by blinks that are unrelated to the CS and the US. The parametric features of the CR have been well characterized. The behavioral NM response is robust and discrete and the exact amplitude–time course of the response is easily measurable. The CR is acquired, at least to a significant degree, in a single training session, but not all at once. This allows for an analysis of brain substrates over a period of time as learning is proceeding (as opposed to one-trial learning or some operant tasks, which show a rapidly accelerating acquisition curve). The presentation of the CS and the US does not yield sensitization or

pseudoconditioning. That is, unpaired presentations of the CS and the US do not increase the ability of the CS to elicit a response; they also do not increase the spontaneous blink rate. These are the primary reasons that led Richard F. Thompson and associates to choose this paradigm to continue the search for the elusive engram.

Conceptually, to search for an engram, a neural circuit must be identified. No single neurobiological method or technique is sufficient in and of itself to define the characteristics of a neural circuit, including the essential site of plasticity. Therefore, numerous methods must be used to determine the roles that different brain structures play in any learned behavior. To date, the most fruitful methods have been electrophysiological recordings, permanent and reversible lesions, and electrical microstimulation of various nuclei and fiber tracts. These methods are used in combination with classical conditioning and with the assumption that there are neural pathways connecting the CS, the US, and the UR pathways. It is also assumed that it is possible to localize a specific brain region where some essential modification takes place to drive the conditioned response. All these methods have been used successfully in defining the circuit for classical eyeblink conditioning.

XII. THE BRAIN SUBSTRATES OF THE CLASSICALLY CONDITIONED NM RESPONSE IN THE RABBIT

Since the early 1970s, Thompson and many others have used a variety of methods to search for the classically conditioned NM response engram. A neuronal circuit diagram has been systematically constructed during the course of 20 years to represent all the brain structures that are essential for the acquisition and retention of the classically conditioned NM response for the delay conditioning paradigm. This circuit is the most thoroughly investigated and completely understood learning and memory circuit known for the mammalian brain. The essential site of plasticity (i.e., the location of the engram) appears to be the interpositus nucleus of the cerebellum, although the cerebellar cortex is also important.

A. Components of the Circuit

Figure 2 is a neuronal circuit diagram that represents all the brain components that are essential for the acquisition and retention of the CR.

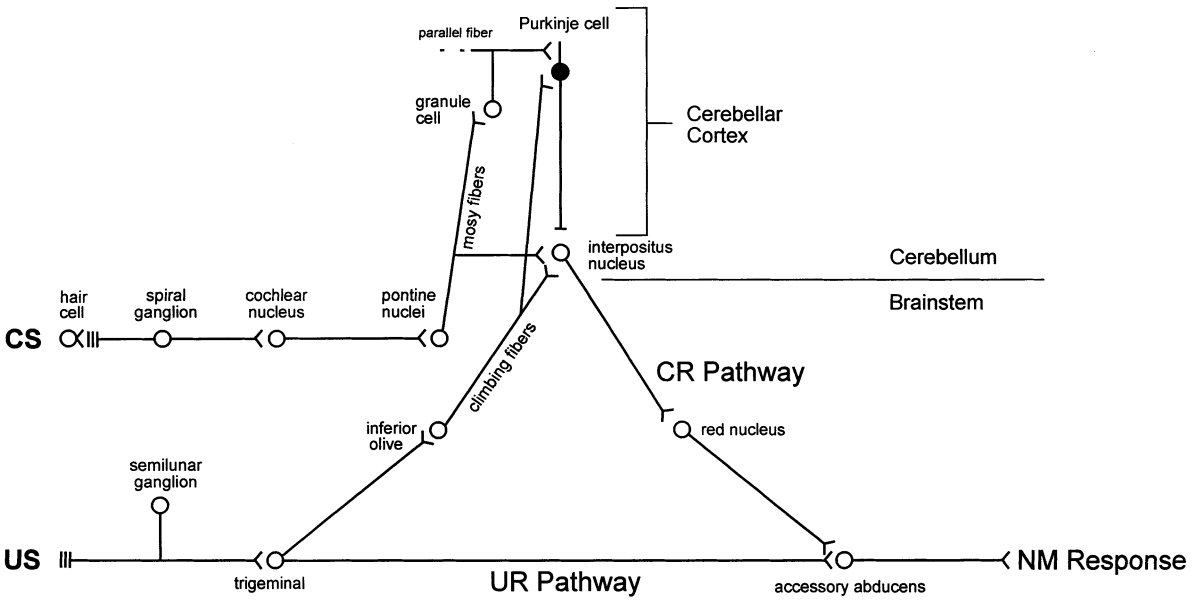


Figure 2 The cerebellar neural circuit for delay eyeblink/nictitating membrane (NM) classical conditioning. The focus of this diagram is on the relationships between the conditioned stimulus (tone CS), unconditioned stimulus (airpuff US), conditioned response (CR), unconditioned response (UR), and their connecting fiber pathways. The cerebellar cortex in the diagram consists of granule cells, parallel fibers, and Purkinje cells. Cerebellar structures in the diagram include the cerebellar cortex and interpositus nucleus. All other structures depicted are located in the brain stem. Pairing of the CS and US results in an essential plastic change in the interpositus nucleus that results in the generation of the CR.

1. The CS Pathway

The tone CS is detected by auditory hair cells, which stimulate spiral ganglion cells. These cells project to the cochlear nucleus, which then projects to the lateral pontine nuclei. The lateral pontine nuclei send mossy fibers to both interpositus nucleus and granule cells in the cerebellar cortex.

2. The US Pathway

The airpuff US is detected by similunar ganglion cells that send somatosensory information to the spinal trigeminal nucleus. This nucleus projects to the dorsal accessory olive. The olive sends climbing fibers to both the interpositus and the cerebellar cortex.

3. The UR Pathway

The sensory trigeminal projects to the abducens, accessory abducens, and facial motor nucleus, which are the motor nuclei that drive the eyeblink and NM response.

4. The CR Pathway

The interpositus nucleus, which receives convergent information concerning the CS and the US, projects to the red nucleus. The red nucleus then projects to the same motor nuclei that drive the eyeblink and NM response (i.e., the abducens, accessory abducens, and facial motor nucleus).

5. Circuit Philosophy and Function

Figure 2 focuses on the relationships of the CS, US, CR, and UR pathways rather than emphasizing the physical position of the nuclei and tracts within the brain. Initially, in a naive subject, the CS does not cause a NM response. The US stimulates the trigeminal nucleus, which then projects to motor nuclei to cause the reflexive UR from the first trial on. During training the interpositus nucleus receives convergent CS and US information. Eventually, this convergent information causes a change (i.e., learning) in the interpositus nucleus and cerebellar cortex. This change enables the CS information to activate the interpositus, which can then activate the red nucleus. The red nucleus then activates the motor nuclei, which then drive the NM response. This response occurs before the delivery of the airpuff US. This is the conditioned response.

B. Brain Structures That Are Critical for Classical Conditioning of the NM Response

1. Interpositus Nucleus

The interpositus nucleus receives convergent CS and US information from the pontine nuclei (CS information) and the inferior olive (US information). The plastic changes that occur in the interpositus as a result of this convergent information are responsible for the CR. The interpositus is considered to be the prime location of the classically conditioned NM response engram. Damage to the interpositus nucleus completely and permanently abolishes the CR without affecting the UR.

2. Cerebellar Cortex

The cerebellar cortex also receives convergent CS and US information. The output of the cerebellar cortex goes through the interpositus nucleus. It has been suggested that the cerebellar cortex might form and store the essential plasticity for the engram and that interpositus lesions only block the expression of the CR. However, large lesions of the cerebellar cortex do not abolish the CR or prevent the acquisition of the CR. The cerebellar cortex likely plays a critical modulatory role in the acquisition and expression of the CR.

3. Pontine Nuclei

The pontine nuclei receive information about the tone CS and relay this information to the interpositus nucleus and cerebellar cortex. Rabbits can be classically conditioning by electrically stimulating the pontine nuclei in place of a tone CS and pairing the stimulation with an airpuff US.

4. Inferior Olive

The inferior olive receives US information from the trigeminal nucleus located in the brain stem. The inferior olive then relays this information to the interpositus and cerebellar cortex. Rabbits can be classically conditioned by electrically stimulating the inferior olive in place of an airpuff US. Pairing a tone CS with inferior olive stimulation as the US results in the development of CRs. In fact, rabbits can be conditioned without tones or airpuffs by stimulating the pontine nuclei as the CS and the inferior olive as the US. In well-trained rabbits (i.e., rabbits showing CRs),

lesions of the inferior olive result in a gradual extinction of the CR as if the US has been removed, even though the US is still being presented. This indicates that the inferior olive sends critical US information to the interpositus and cerebellar cortex.

5. Red Nucleus

The red nucleus is a critical structure in the CR pathway. It receives input from the interpositus, which is outputting a neural signal for driving the CR. The red nucleus then relays this signal to motor nuclei for driving the learned NM response. Lesions of the red nucleus abolish the conditioned response. It is known, however, that the red nucleus is not responsible for forming the CR because reversibly inactivating the red nucleus during training prevents the expression of the CR but does not prevent the acquisition of the CR. When the lesion is reversed (i.e., restored to normal function), CRs rapidly appear. In contrast, when the interpositus is inactivated, there is no acquisition of the CR.

6. Trigeminal Nucleus

The trigeminal nucleus receives information regarding the occurrence of the US. It then sends this information to the interpositus and cerebellar cortex (via the inferior olive) so that an association between the CS and the US can be accomplished. It also sends the US signal to the motor nuclei for driving the UR.

7. Accessory Abducens

The accessory abducens is the motor nucleus that drives the NM response. It receives information from the trigeminal nucleus for driving the UR and information from the red nucleus for driving the CR.

XIII. THE HIPPOCAMPUS IS REQUIRED FOR TRACE CLASSICAL NM CONDITIONING

The circuit just described includes all the brain structures required for the successful acquisition and retention of the CR in the delay classical conditioning paradigm. For example, in a well-trained animal, all the neural tissue above the level of the midbrain can be completely removed without affecting the retention and expression of the CR. In other words, removing the entire neocortex, thalamus, basal ganglia, and limbic system, including the hippocampus, has no effect on the CR.

For trace classical conditioning, these structures are also essential; however, the hippocampus is also required. If an interval of at least 500 msec separates the offset of the CS and the onset of the US, hippocampal lesions prevent CR acquisition and retention.

XIV. THE AMYGDALA IS ESSENTIAL FOR THE ACQUISITION AND RETENTION OF THE CLASSICALLY CONDITIONED FEAR RESPONSE

Although amygdala lesions do not affect the acquisition or retention of eyeblink/NM classical conditioning, the amygdala is essential for the acquisition and retention of the classical conditioned fear response. For example, in rats, when a tone CS is paired with a shock US several times, a conditioned fear response develops. In other words, after pairing the CS will cause a fear response of freezing. The rat that has been conditioned will hold completely still when the CS is presented. Amygdala lesions completely prevent rats from learning or retaining this fear response. Thus, classical condition has revealed that the amygdala is a critical brain structure for emotional fear learning.

XV. BRAIN STRUCTURES INVOLVED IN HUMAN CLASSICAL CONDITIONING

Although studies of classical conditioning in humans began to wane in the 1960s, particularly for eyeblink classical conditioning, within the past 10 years there has been a resurgence of experimental work that can largely be attributed to the success of classical conditioning as a tool to study brain function in the experimental animal. Currently, our understanding of how different human brain structures contribute to classical condition lags far behind what is known in the animal and will likely never approach the level of precision that is possible with animal studies. Nevertheless, there has recently been much progress with relating classical conditioning to brain function in humans.

Recent findings from human studies have been remarkably consistent with previous work with animal studies of conditioning. For example, work in rats has convincingly demonstrated the importance of the amygdala in fear classical conditioning. Studies using

brain imagining methods such as positron emission tomography (PET) and functional magnetic resonance image (fMRI) have shown that fear classical conditioning activates the amygdala. Additionally, humans with bilateral degeneration of the amygdala, as a result of Urbach–Wiethe disease, show an impaired ability to acquire fear classical conditioning.

Results from human studies of eyeblink classical conditioning have also been remarkably consistent with those of studies in rabbits. For example, work with rabbits clearly indicates that the cerebellum is critical for acquisition of delay classical conditioning. Humans with cerebellar damage are also severely impaired on eyeblink classical conditioning. Humans with bilateral hippocampal damage due to anoxia are normal at acquiring CRs in the delay classical conditioning paradigm, but they are impaired when tested on trace classical conditioning. These results are entirely consistent with the animal work. Imagining studies using PET and fMRI have also consistently identified the cerebellum and hippocampus as being activated during classical conditioning of the eyeblink response.

In addition to supporting previous work in animals, human eyeblink conditioning studies have also extended our understanding of brain function in several interesting ways. For example, it has been reported that the knowledge that humans sometime acquire about the stimulus contingencies of the conditioning experiment (e.g., the CS predicts the US) is an important variable for trace conditioning but irrelevant or superfluous for delay eyeblink conditioning. Humans with hippocampal damage fail to acquire this knowledge and accordingly fail to acquire trace conditioning while being unaffected on delay conditioning.

It has repeatedly been reported that subjects with Alzheimer's disease and those with probable Alzheimer's disease are impaired on delay classical conditioning of the eyeblink response. At first, this finding did not appear to be congruent with the animal work because Alzheimer's disease does not affect the cerebellum or other brain stem structures that are critical for delay conditioning. However, work in rabbits has shown that although the hippocampus can be removed without affecting acquisition in the delay paradigm, disrupting the hippocampus by inactivating the cholinergic input to the hippocampus can disrupt delay conditioning. Alzheimer's disease disrupts the septohippocampal cholinergic system and it is this disruption that likely causes the impairment in delay eyeblink conditioning. If fact, this classical conditioning paradigm is so sensitive to the early effects of Alzheimer's disease that, it has been proposed as a simple neuropsychological test for Alzheimer's. Finally, humans with autism, a developmental disorder characterized by severe impairments in communication and social relating and by ritualistic and repetitive patterns of behavior, also show abnormalities in the cerebellum. Subjects with autism also show abnormal acquisition and extinction of the CR. This finding further supports the involvement of the cerebellum in classical eyeblink conditioning.

See Also the Following Articles

BEHAVIORAL NEUROGENETICS • COGNITIVE PSYCHOLOGY, OVERVIEW • INTELLIGENCE • REINFORCEMENT, REWARD, AND PUNISHMENT

Suggested Reading

Gormezano, I., Prokasy, W. F., and Thompson, R. F. (Eds.) (1987). *Classical Conditioning*. Erlbaum, Hillsdale, NJ.

Green, J. T., and Woodruff-Pak, D. S. (2000). Eyeblink classical conditioning: Hippocampal formation is for neutral stimulus associations as cerebellum is for association-response. *Psychol. Bull.* **126,** 138–158.

Kim, J. J., and Thompson, R. F. (1997). Cerebellar circuits and synaptic mechanisms involved in classical eyeblink conditioning. *Trends Neurosci.* **20,** 177–181.

Pavlov, I. P. (1927). *Conditioned Reflexes* (G. V. Anrep, Trans.). Oxford Univ. Press, London.

Woodruff-Pak, D. S., and Steinmetz, J. E. (Eds.) (2000a). *Eyeblink Classical Conditioning: Volume I—Applications in Humans*. Kluwer, Boston.

Woodruff-Pak, D. S., and Steinmetz, J. E. (Eds.) (2000b). *Eyeblink Classical Conditioning: Volume II—Animal Models*. Kluwer, Boston.

Cognitive Aging

NAFTALI RAZ
Wayne State University

GLOSSARY

Alzheimer's disease (AD) An age-related degenerative condition that is associated with excessive deposits of amyloid in the neurons, formation of plaques and tangles, and eventual demise of large swaths of brain tissue. As a rule, AD starts in the hippocampus and the enthorhinal cortex and gradually progresses to engulf the whole cerebral cortex. The cognitive expression of AD is dementia characterized by significant difficulties in delayed recall of information, spatial and temporal disorientation, and progressive language difficulties.

association cortex Regions of the cerebral cortex that are cytoarchitechtonically different from the sensory cortices and perform no modality-specific functions.

basal ganglia A group of subcortical nuclei extensively connected among themselves as well as with the cerebral cortex and the midbrain nuclei. The group includes the caudate nucleus, the putamen, the globus pallidus, the subthalamic nucleus, and nucleus accumbens. Sometimes referred to as the extrapyramidal motor system.

excitotoxicity A chain of biochemical and electrophysiological events that accompanies release of an excitatory neurotransmitter glutamate. Excessive release of glutamate results in the demise of the neurons in which it occurs. Excitotoxicity is believed to be the cause of cerebral damage in hypoxic and ischemic episodes and stroke.

executive functions The term "executive" is amorphous and ill defined. It encompasses a broad range of cognitive skills, such as monitoring one's recent and past performance, generating future goals, inhibiting prepotent overlearned responses, and switching behavioral patterns in response to feedback.

explicit memory A mnemonic activity that relies on effortful conscious processes. It can be considerably improved by an increase in environmental contextual support, additional elaboration of material, and reduction in uncertainty in the target stimuli. It responds negatively to an increase in attentional load. The major subtypes of explicit memory are episodic and semantic. Episodic memory refers to encoding and retrieval of information about personally experienced past events and observed objects, locations, and actions. Semantic memory represents hierarchically organized knowledge about past events common to many individuals, concepts, rules, and scripts.

implicit memory A mnemonic activity that involves retention of information without conscious awareness. Implicit memory can be inferred only from changes in performance rather than from an individual's direct report. A type of implicit memory called repetition priming occurs when a past experience, such as exposure to a stimulus, increases the likelihood of selecting this stimulus beyond the statistically expected level. Implicit memory is insensitive to distraction and cannot be improved by elaboration of encoding procedures.

ischemia The temporary or permanent interruption or significant reduction of blood supply to an organ (e.g., brain). In the brain, it is accompanied by deprivation of delivery of oxygen and glucose to neural tissue. It sets in motion a chain of biochemical events that result in necrosis of living tissue.

magnetic resonance imaging A neuroimaging technique based on recording the response of protons in the living tissue to radiofrequency waves in a strong magnetic field. By converting the values of proton relaxation in circumscribed regions of the brain into grayscale values, an anatomically faithful picture of the brain is created, thus allowing *in vivo* observation of the cerebral structures in intact humans. A variant of this technique, called functional magnetic resonance imaging, is based on measuring changes of blood oxygenation in time (a hemodynamic response) that can be triggered by an externally presented stimuli. By time-locking stimuli to the hemodynamic response sampling, brain response to specified cognitive and perceptual demands can be studies.

neurotransmitter A messenger between two neurons and a substance that is synthesized in the body of a neuron, transported to the presynaptic membrane, released into the synaptic cleft, and

delivered to the postsynaptic terminal where its molecules attach to the molecules of a receptor, thus delivering information between the two neurons. Neurotransmitters that play an important role in aging and age-related pathology include dopamine and norepinephrine (catecholamines), glutamate, and acetylcholine.

positron-emission tomography An imaging technique based on computerized reconstruction of localized brain activity inferred from decay of highly volatile radioactive probe material attached to one of the major fuels of the neurons (oxygen and glucose) or to a specific pharmaceutical ligand that binds to a receptor.

primary cortex A region in the cerebral cortex that contains neurons receiving projections from the periphery of a given sensory system (sensory cortex) or sending projection to the peripheral systems of neurons regulating the effectors (motor cortex).Within primary sensory cortices a high degree of segregation of processing (by retinal location, sound frequency, or skin patch) is observed.

procedural memory Retention of a learned skill (mental, perceptual, or motor) acquired after practice. It is independent of explicit or declarative understanding and recollection of the performed activity and of the acquisition process.

receptor A protein embedded in the membrane of a neuron. The ionotropic receptor forms a fluid-filled pore (ion channel) that allows a regulated flow of ions across the membrane. The metabotropic receptor opens ion channels indirectly through a chain of intracellular messengers. The stereochemical configuration of the receptor and the neurotransmitter determines the likelihood of binding between the former and the latter.

working memory An ability to process information while maintaining intermediate products, goals, and associated strategies of processing online.

Aging is a fundamental biological process with far-reaching cognitive consequences. Deeply rooted in the genetic makeup and metabolic workings of the organism, but sensitive to fluctuations of biological and social environment, aging brings with it complex changes in the workings of the mind. Although multiple age-related changes in brain and cognition have been described in detail, their mechanisms are still poorly understood.

I. INTRODUCTION

After decades of diligent cataloguing of neural and behavioral properties of the aging organisms, the stage is set for a union of neuroscience and cognitive psychology aimed at understanding the relationships between aging brain and cognitive performance. Although a promise of cognitive neuroscience looms large in the minds of the students of cognitive aging, its emergence created a flux of ideas and generated empirical findings that compel us to revise ostensibly well-set concepts and theories on almost a daily basis. Thus, the following article, by necessity, is a snapshot of a powerful stream in its exciting but ill-defined turbulence rather than a detailed picture of a stationary entity.

As often happens in science, the advent of new measurement techniques and invention of new instruments provided an impetus for the current neural revolution in cognitive science in general and cognitive gerontology in particular. In the past two or three decades, with the help of magnetic resonance imaging (MRI), previously inaccessible *terra incognita* of the human brain has been clearly visualized *in vivo* with relatively high anatomical precision ($<1\ mm^3$). Metabolic workings of the brain and local changes in cerebral blood flow can be gauged by positron emission tomography (PET) and functional MRI (fMRI).

II. THE AGING BRAIN

The studies of brain and cognition reveal that in both domains, age-related changes are differential as well as generalized. The results of postmortem and *in vivo* investigations of the human brain demonstrate that although advanced age is associated with reduced total brain weight, nonspecific sulcal expansion, ventricular enlargement (Fig. 1), and a decline in global brain

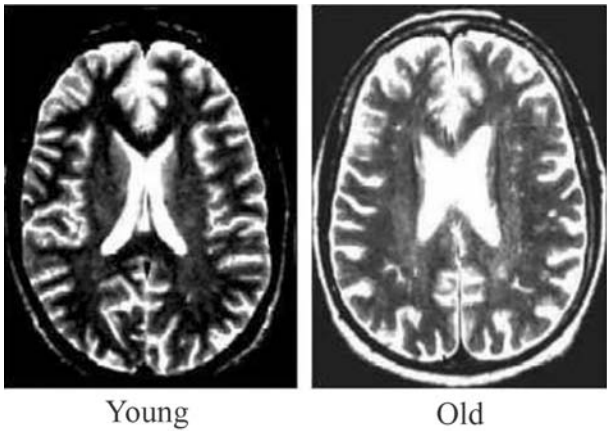


Figure 1 Examples of brain aging as seen on MR images (a 23-year-old female and a 77-year-old female). The images represent a cross section of the brain in a horizontal plane. The parameters used to acquire these images emphasize changes in brain water state and content, thus highlighting the spaces surrounding blood vessels and the cerebral sulci and ventricles filled with cerebrospinal fluid. In this type of MR scan, fluid appears white and myelinated fibers are at the darkest end of the gray scale. Note enlarged ventricles, widened sulci, and the presence of subcortical white matter hyperintensities in the aging brain.

blood flow, the greatest negative impact of aging is evidenced by the prefrontal cortex (PFC) and the neostriatum (Figs. 2A and 2B). Age-related changes in the hippocampus (Fig. 2C) are also significant, although some of them may be attributed to the early signs of age-related pathology and are more prevalent in the oldest old. The cerebellum and the primary sensory cortices show only mild age-related decline, and the lower brain regions such as the ventral pons appear insensitive to the impact of senescence.

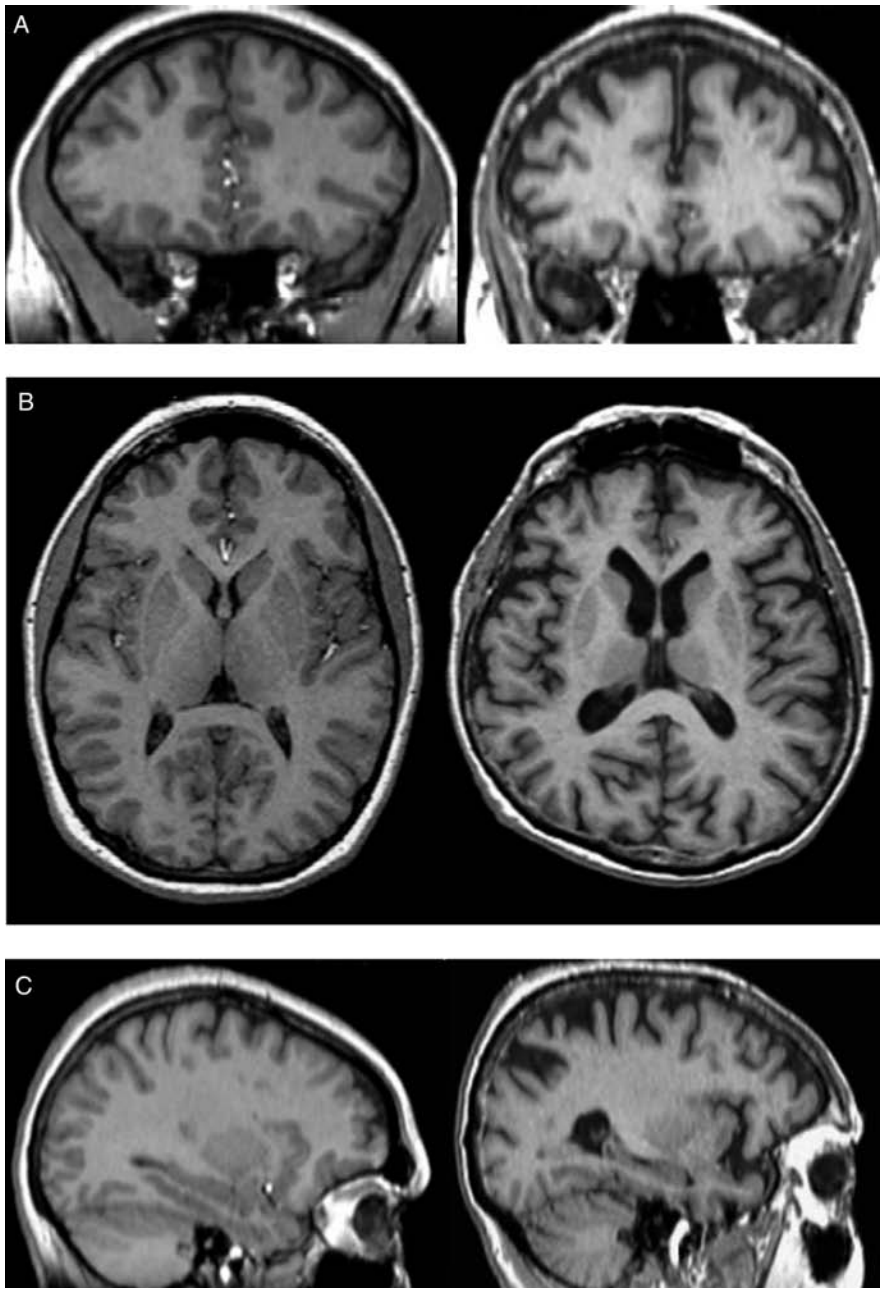


Figure 2 Regional age-related differences in the brain comparison of MR images. (A) Prefrontal cortex in a coronal plane perpendicular to the plane cutting through the anterior and posterior commissures (AC–PC). (B) Basal ganglia (striatum) in a horizontal plane approximately parallel to the AC–PC. (C) Hippocampus (HC; in a sagittal plane, perpendicular to the AC–PC). The images of a 23-year-old female are on the left, and the images of a 77-year-old female are on the right. The image acquisition parameters were selected to emphasize anatomical details and the white–gray matter differences. On these images, the white matter and blood vessels appear in high-intensity white, the gray matter is rendered in a darker end of the gray scale, and the cerebrospinal fluid is black.

The reasons for the observed differential distribution of age-related vulnerabilities are unclear. The pattern of differential brain aging reflects to some extent the outline of age-sensitive neurotransmitter systems and may be associated with age-related changes in the intracellular distribution of calcium ions. The regions and cell ensembles that have higher propensity to calcium-driven plasticity may be more vulnerable to excitotoxic events that cause age-related damage. Although there is a tendency of phylogenetically newer and ontogenetically late to mature regions to show greater negative effects of age, interspecies and even interstrain differences in brain response to aging are plentiful. Most of the age-sensitive regions are also located in the areas of the brain that are prone to suffer negative effects of subclinical ischemia. Although those are only clues, and biological mechanisms that drive differential brain aging remain to be elucidated, the connection between brain aging and cognitive changes of late adulthood is undeniable. The main challenge for cognitive neuroscience is to establish specific relations between age-related transformations of the brain circuitry and altered performance of cognitive operations that depend on those circuits.

III. COGNITIVE AGING AND ITS NEURAL SUBSTRATES

Age-related changes in cognitive performance mirror the complex picture of brain aging as it presents a pattern of selective preservation and decline against the background of generalized changes. A substantial proportion of age-related variance in cognitive performance can be explained by generalized slowing and dwindling of the general pool of cognitive resources. Age-related slowing of response and difficulties in dealing with multiple tasks are easy to spot even by untrained observers. It is difficult, however, to separate cognitive declines that are due to slowing from those that result from a reduction of resources. From the stand-point of the classic information theory, time needed for message processing, signal-to-noise ratio of the message, and the availability of processing elements (resources) are mutually dependent. In principle, deficiency in one of these properties can be compensated by excess in another. However, the process of aging affects all the information processing capabilities and such compensation may no longer be feasible. Thus, aging individuals may require increasingly longer time for processing, can tolerate less admixture of noise in the stimuli, and can deploy fewer resources. As task difficulty and complexity of processing increase, so do age-related differences in speed and accuracy of performance. The debate among scholars of cognitive aging is centered on whether such incremental slowing is proportional and uniform across the tasks or disproportional and differential. Indeed, after accounting for generalized change in speed and resource availability, examples of differential cognitive aging are not difficult to find. Functions associated with episodic memory and executive control of cognitive processes are more sensitive to aging than are semantic memory and verbal reasoning, which rely on stable knowledge structures and well-honed expert skills. The distinction between the tasks dependent on overlearned and automated skills and those that require effortful processing in the absence of environmental support extends into the age-sensitive domains as well.

Because both brain and cognition age differentially, it is plausible that age-sensitive cognitive operations are supported by age-sensitive areas of the brain, whereas age-invariant cognitive abilities are maintained by brain circuits that are not particularly prone to aging. It is also plausible that age-related deterioration of specific circuits that are crucial for normal execution of cognitive operations leads to functional reorganization of the aging brain. As a result of such reorganization, the patterns of brain–behavior relations observed in young adults may become transformed and redefined in older people, thus complicating an already daunting task of understanding brain–behavior relations in the old age. On a more positive note, an important factor in reduction of age-related deficits is expertise. The tasks that are performed repeatedly throughout the life span are less likely to show age-related deficits. Moreover, deliberate prolonged practice that characterizes experts careers acts as a protector against age-related declines even in the realms of behavior that demand considerable exertion even from younger performers. Evidence from multiple domains, such as piano playing and chess supports the notion of expertise as an antidote to aging. One must bear in mind, however, that impressive stability of expert performance shows little if any spillover to other cognitive and behavioral domains. Thus, despite some local successes, the general battle against aging is gradually lost, and determination of the neural harbingers, substrates, and correlates of that process remains the central issue in modern gerontology.

The relationship between global indices of cognitive performance and generalized age-related alterations in brain structure has been examined in several studies. The early studies that relied on neuroimaging techniques of limited resolution revealed moderate general associations between nonspecific age-related atrophy markers (ventriculomegaly and cerebral sulcal expansion) and performance on a broad range of neuropsychological tests in healthy elderly. In some samples, however, global deterioration of the white matter predicts age-related cognitive slowing and declines in executive functions whereas showing no links to other, more age-invariant functions. Although newer investigations in which brain structure and functions were assessed with more sophisticated techniques confirmed the link between the signs of generalized brain deterioration and general indices of cognitive decline, more detailed inquiries into specific relations between cognitive operations and brain structure and function became possible. The findings generated to date by this fruitful approach are summarized next.

1. Motor Functions

Age differences in cognition are inferred from observing a motor (vocal or manual) response to the task stimuli, and age-related changes in the motor system must be taken into account. Although age effects on simple reaction time are mild, aging is accompanied by multiple deficits in planning, control, and execution of movements. Only some of these changes can be attributed to age-related changes in the musculoskeletal system. *In vivo* PET studies of the dopaminergic system of healthy adults suggest that an age-related decline in availability of dopamine receptors in the caudate nucleus and putamen is linked to impaired fine motor performance. In fMRI measures of activation associated with simple motor tasks, older participants show more sluggish hemodynamic response and lower signal-to-noise ratio than their younger counterparts. Notably, some of the age-related motor deficits may be mediated by differences in aerobic fitness and cardiovascular status. Thus, the results of experimental studies of age differences in cognition should be interpreted in the context of age-related changes in the motor response system.

2. Sensory Processes

Aging is known to inflict significant selective changes on the earliest echelons of sensory systems. Thus, the information extracted from the environment reaches the cerebral cortex of an elderly individual in a partial or distorted form. Some functions, such as auditory gain control, high-frequency discrimination, and visual contrast sensitivity, are especially hard hit by age-related processes. In the context of far-reaching changes in the cerebrovascular system, it is not surprising that aging is associated with alteration of cerebral hemodynamics in response to simple sensory stimulation. Because microvascular response of brain vessels is regulated by dopamine, age-related dopaminergic deficits may dampen the hemodynamic response elicited from the aging brain, regardless of cognitive factors. On the other hand, patients with cardiovascular diseases that are common in old age display task-relevant *hyper*activation in the brain regions that are activated by the same task in normal controls. Thus, sensory changes and the health problems frequently associated with aging impose constraints on the interpretation of patterns of brain evoked by cognitively demanding tasks.

3. Perception

A substantial amount of evidence established the existence of two main streams of visual information processing in the brain. According to the two-stream model, perception of objects depends on the ventral stream, incorporating the inferior temporal and the fusiform gyri, whereas identification of spatial location is subserved by the dorsal stream anchored by the superior parietal lobule. Neuroimaging investigations of age differences in visual perception support the double dissociation of ventral and dorsal streams in both younger and older subjects. However, several important age differences have been observed. When confronted with a relatively simple perceptual task, older subjects evidenced increases in rCBF to the amodal and visual association areas, whereas their younger counterparts showed greater activation of the primary and secondary visual cortices.

4. Attention

The current view of the neural substrate of attention distinguishes between two major brain systems guiding deployment, maintenance, focusing, and division of attention: a posterior system mainly devoted to selective deployment and disengagement of visual attention and an anterior system that is responsible for division of attention among multiple cognitive processes. The anterior cortical attentional system

comprises dorsolateral prefrontal cortex (DLPFC) and anterior cingulate gyrus and is critical for executive control of attention, including control of its motivational aspects. The posterior attentional system involves a network of cortical and subcortical components, such as the superior and inferior parietal lobules, secondary association cortices in visual and auditory modalities, and the subcortical pathways guiding saccadic eye movements. A dense network of neural connections between the two systems led to an understanding that they cannot operate in isolation but act as parts of an integrated entity.

Age differences in focused and divided attention appear to stem not only from quantitative declines in the magnitude of cerebral activation but also from discrepancies in the patterns of cortical activation. The younger subjects activate the brain circuitry that subserves specific visual attention processing, whereas their older counterparts appear to increase engagement of the anterior, general attention system while failing to deactivate the regions usually not associated with attentional control.

5. Memory

Perhaps because it is widely believed that mnemonic deficits are the harbingers of dementia, the cognitive aging literature emphasizes age-related differences in memory. Consequently, most of the neuroimaging studies on aging have concentrated on a search for neural substrates of encoding and recall of information. Memory, however, is not a uniform domain, and aging as well as other brain-altering processes affect its distinct systems differentially. The contrast between age-related vulnerability of the episodic and stability of the semantic memories is clear. Moreover, within the realm of episodic memory, age differences in free recall tasks that challenge the participants to effortful retrieval are greater than in the tests of recognition memory, which involve a less demanding selection from a restricted menu of alternatives. In contrast, repetition priming, with its absence of conscious and effortful recollection requirements shows few age-related differences.

It is a matter of broad consensus that explicit (mainly episodic) memory depends on two subdivisions of the central nervous system: the medial–temporal–diencephalic structures and the prefrontal cortex. However, beyond that general accord, there is little agreement regarding the exact role played by specific limbic and anterior association structures in various memory functions.

Studies of experimental lesions in nonhuman mammals and observations on human diseases affecting the hippocampus suggest a central role for the hippocampal formation in mnemonic processes. The hippocampus and affiliated structures, such as the fornix, the entorhinal cortex, the medial dorsal thalamus, and the mammillary bodies, contribute to the consolidation and maintenance of memories. However, human neuroimaging studies suggest that the left prefrontal cortex is preferentially involved in the encoding of episodic memories, whereas its counterpart on the right is dominant in the retrieval of episodically acquired information. However, in addition to the prefrontal and limbic structures explicit memory depends on a wide network of cerebral circuits involving modality-specific association cortices, anterior cingulate gyrus, and selected thalamic nuclei, all of which are affected by normal aging to various degrees. In contrast to explicit memory, implicit memory processes require virtually no limbic–diencephalic support and depend instead on multiple cortical and subcortical systems, including the primary sensory cortices. Procedural learning such as acquisition of a motor or perceptual skill depends on the integrity of premotor, supplementary motor, and secondary visual areas, the neostriatum, the cerebellum, and the connections between them. Finally, simple Pavlovian conditioning relies on circumscribed circuits and nuclei in the brain stem and the cerebellum.

The majority of studies that have addressed the issue of neuroanatomical correlates of age-related differences in mnemonic performance have concentrated on the hippocampus and adjacent limbic structures, although in some samples global indices of brain aging such as measures of the white matter integrity are mildly predictive of declines in delayed recall. In healthy older adults, shrinkage of the hippocampal formation but not of the superior temporal gyrus is associated with poor delayed recall. In some samples, visual memory deficits correlate with reduced volume of the amygdala. Relatively short (2–6 years) longitudinal follow-up studies of healthy older adults reveal that reduction of the hippocampal size predicts the decline in performance on age-sensitive memory tests, and reduction of the whole brain volume predicts decline in verbal memory. Semiquantitative indices of hippocampal atrophy derived from MRI scans also predict poor memory performance in healthy older adults. In some samples, the association between hippocampal shrinkage and memory deficits is observed only in the oldest subjects or it is almost entirely confounded with age. Moreover, sometimes the results

point in the opposite direction, with smaller hippocampi and smaller parahippocampal gyri predicting better delayed recall.

Although structural neuroimaging studies provide insights into relations between long-term changes in the brain and cognitive declines, they cannot reveal brain mechanisms of observed age-related differences in cognition. Functional neuroimaging, with its almost real-time capabilities, is better suited for examination of the impact of age-related changes in brain work on cognitive performance. This area of research is in its inception and only a handful of activation studies of age-related differences in memory have been conducted. Although the stability of observed age differences in the patterns of activation is hampered by the small sample size of a typical study, some general trends have emerged. In neuroimaging studies of encoding, older participants either show reduced activation in the left inferior prefrontal cortex or fail to activate this area at all. In addition, in some samples, age-related reduction of activation in the anterior cingulate gyrus is observed. When encoding is studied within a semantic recognition framework, no age differences in prefrontal activation are observed, in accord with the findings of age-related stability of semantic memory. Also, as predicted on the basis of behavioral studies that reveal no age-related declines in priming, repetition priming produces equivalent deactivation in the occipital cortices of older and younger adults.

The evidence is less consistent in regard to the role of medial temporal activation in age-related memory changes. Although in some samples older subjects evidenced a significantly reduced activation in the hippocampus and the parahippocampal gyrus, this finding has not been consistently observed in other groups of older adults. To complicate the picture further, a recent fMRI study of verbal cued recall suggested that task-relevant medial temporal activation is more widespread in healthy older adults who possess a genetic risk for Alzheimer's disease than in risk-free controls. Thus, one cannot conclude that aging brings with it a decrease in task-related brain work. Older people in general and those predisposed to age-related pathology may use medial temporal and prefrontal structures in ways that are different from how younger people use them.

In the young participants engaged in retrieval of episodic information, activation of the prefrontal cortex predicts the speed of information processing across tasks, whereas in the older people the same region is also associated with task-specific decision processes. Thus, it appears that older brains have to recruit additional resources to manage the executive overhead of the task. It is unclear whether such an adjustment is of compensatory value because it relies on the very circuitry (prefrontal association cortex) that is most affected by the adverse influences of aging.

6. Executive Functions

A growing body of research has revealed that many age-related deficits in memory may stem from declines in the means of strategic management of mnemonic resources. Multiple studies have demonstrated that a significant proportion of age-related declines in cognitive performance can be traced to deterioration of working memory (WM). Specifically, the more prominent the strategic or executive component of the WM task, the greater the negative effect of age on WM performance. Earlier models of WM derived from laboratory experiments in nonhuman primates and human activation studies emphasized its almost exclusive dependence on amodal prefrontal association areas. However, recent data suggest that modality-specific secondary association cortices play an important role in supporting this resource, and coactivation of anterior and posterior association cortices may be critical for maintaining performance during increased load on WM.

In light of the central role played by WM in cognitive aging, the literature on cortical activation patterns associated with performance on WM tasks is surprisingly sparse. Structural MRI investigations of the neural substrates of age-related declines in WM revealed significant but weak links between verbal working memory and the volume of the prefrontal cortex. Functional neuroimaging studies of WM in young and older adults revealed that whereas task-relevant activation in secondary association cortices is equivalent across the age range, greater prefrontal (amodal) cortex activation is observed in the elderly. A typical WM task results in activation of ventral and dorsal prefrontal, anterior cingulate, and posterior parietal cortices as well as the neostriatum in both young and older adults. However, when the mnemonic demands increased, activation extended into a larger region (posterior DLPFC) in young but not old. The latter seemingly fail to respond adequately when the task calls for recruiting additional executive resources. Thus, different aspects of the delayed response task appear to be reflected in activation patterns of young and older people. Whereas at younger ages they may

reflect greater investment in stimulus-related aspects of the tasks, in older people activation may signify recruitment of resources for maintaining general attentional demands.

If the central executive component of WM is preferentially affected by aging, greater emphasis on that aspect of WM in future neuroimaging studies may help to bring more coherence to this field of research. Age-related deficits in WM may stem from declining function of the dopaminergic system and specifically the prefrontal dopamine receptors that are affected by aging. Such an association is especially plausible because activity of the dopaminergic receptors plays an important role in executive functions subserved by the prefrontal cortex. In aging monkeys, administration of low doses of dopamine agonists improves working memory. Furthermore, it has been suggested that optimal catecholaminergic and cholinergic activity in the prefrontal cortex may be the key to superior performance on executive and working memory tasks. Thus, impaired interaction between dopaminergic and noradrenergic systems rather than a deficit in either of them may cause age-related declines in WM and more complex cognitive activities that depend on it. This impairment may or may not be more acute in individuals who experience greater shrinkage of the prefrontal cortex and who show less efficient metabolism of glucose in the affected areas.

Executive tasks that engage WM and additional strategic components, such as planning, feedback-contingent switching, goal management, inhibition, and response selection, are sensitive to aging. Humans as well as nonhuman primates exhibit an age-related increase in perseverative errors (i.e., they demonstrate a propensity not to abandon a familiar response even after receiving information indicating that the response is incorrect). Empirical evidence and formal models suggest that executive functions disproportionately depend on the integrity of the prefrontal cortex.

Structural neuroimaging investigations of age-related differences in executive functions have yielded inconclusive results. A reduction in the volume of prefrontal gray matter is associated with an increase in perseverative behavior, and age-related increases in perseverative errors and other executive deficits are linked to elevated burden of white matter abnormalities in the frontal lobes.

Because prefrontal cortex is extensively connected to the basal ganglia, it is plausible that age-related declines in executive functions stem from age-related neurochemical changes in the neostriatum. In accordance with this hypothesis, a reduction in the availability of dopamine receptors predicts difficulties in inhibiting a prepotent response.

When young and older adults perform a task with a significant executive component, they demonstrate significant activation in the DLPFC and the posterior association cortices in their left hemispheres. In addition, only the young subjects show significant left-side activation of the premotor and the posterior parietal cortices, as well as the left fusiform gyrus and the left cerebellum. Notably, in the areas of activation that were common to both age groups, the elderly subjects evidenced lower magnitudes of activation than their younger counterparts, even after adjustment for individual differences in global CBF. Moreover, the magnitude of activation in the left DLPFC, right parahippocampal gyrus, and left prestriate cortex was negatively correlated with the number of perseverative errors.

Inhibition has been considered one of the key notions in discussions of age-related differences in working memory and higher cognitive functions, and the prefrontal (mainly the orbitofrontal) cortex has been widely considered the substrate of inhibitory control. The results of a PET investigation suggest that although young people respond to the introduction of a task switching requirement by an increase in prefrontal activation, their older counterparts show an equivalent activation boost when the tasks are preformed in a blocked order, without switching.

In summary, age-related shrinkage and reduced task-specific activation and increased nonspecific activation of the prefrontal cortical circuits are associated with a variety of errors in executive control of cognitive performance. The mechanisms underlying these deficits and the relationship among executive functions, working memory, and their cortical substrates are yet to be elucidated.

7. Language

Verbal skills assessed by tests of word knowledge and verbal reasoning are considered virtually age invariant, although age-related differences have been reported in more subtle aspects of language processing. A possible explanation of relative stability of some linguistic skills throughout the life-span lies in the fact that hours of sustained practice in reading and oral expression fill our day. Thus, superior vocabulary skills frequently noticed in older adults can be attributed to their extensive practice in that domain. In addition, natural languages are characterized by such a high degree of redundancy that many linguistic

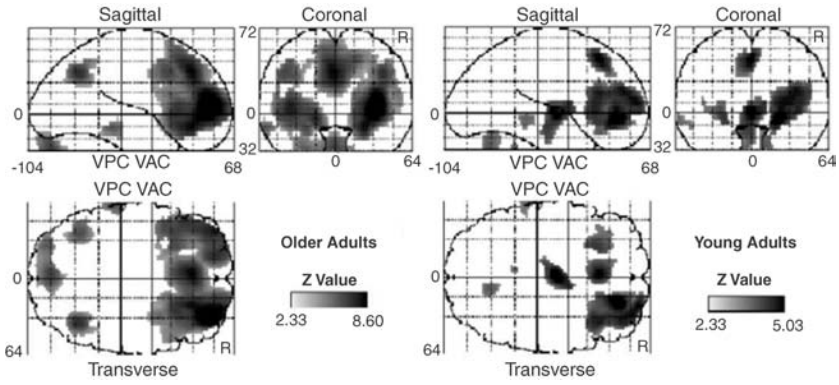


Figure 3 An example of a $^{15}H_2O$ PET activation map in a recognition memory task. (Left) A composite of 12 younger subjects. (Right) A composite of 12 older subjects. The labels VAC and VPC indicate a standard vertical reference plane passing through the anterior and posterior ommissures (AC and PC), respectively. The activation values (z scores) are mapped onto a coordinate grid (Talairach space) anchored at the AC–PC line. Note a substantially broader activation in older adults that in their younger counterparts [images courtesy of D. J. Madden, and first appeared in N. Raz (2000), in *Handbook of Aging and Cognition—II*, (F. I. M. Craik and T. A. Salthouse, Eds.), pp. 1–90. Erlbaum, Mahwah, NJ].

operations can be performed literally with half of the information deleted. However, when a task calls for higher than usual demands on the working memory or speed of processing, age-related declines in linguistic performance become apparent.

Very little is known about neural substrates of age-related differences in linguistic functions. It is unclear whether declines in performance in a given language function reflect changes in general resources such as speed and WM, or specific deterioration of brain circuitry responsible for language processing. An optimistic belief in the lack of age effects on language may be responsible for the dearth of neuroimaging studies of age differences in that domain, although the results of recent investigations that did employ measures of linguistic competence may temper that optimism. For instance, a longitudinal study of a small sample of healthy middle-aged adults showed that reduction of temporal lobe volumes may be associated with mild but statistically significant worsening of performance on language tests.

Neuroimaging findings suggest that although neural mechanisms responsible for semantic processing and filtering of verbal information are unaffected by age, older adults may experience difficulty in recruiting cortical resources for processing of novel pseudolinguistic stimuli.

IV. CONCLUSIONS

Aging is associated with slowing of information processing, reduction in cognitive resources, moderate declines in episodic memory, and increased likelihood of failure of executive or strategic control over cognitive operations. Extensive practice in a specific cognitive domain ameliorates and delays these age-related changes within the restricted domain of practice. Performance on the tasks that depend on overlearned and well-practiced skills such as reading, use of lexicon, and engagement of semantic knowledge network are stable across most of the life span. Brain mechanisms of selective age-related changes in cognition are unclear. A small but steadily growing number of neuroimaging studies indicate that older adults evidence reduced brain activity in the task-relevant and modality-specific cortical areas while increasing the magnitude of activation and broadening the activation regions in the areas that are amodal and irrelevant to the task. An illustration of such dedifferentiation of cortical activation with age is shown in Fig. 3. In an apparent paradox, the brain regions recruited by older people are located in cortical structures that are differentially vulnerable to the effects of aging.

A contrasting interpretation of age-related differences in brain activation patterns has been offered by researchers who proposed to view it as a sign of compensation. One way to decide whether the observed pattern reflects compensation or dedifferentiation is examination of associations between activation and performance. Although the data are still sparse, several studies of healthy adults revealed that more bilateral, wide spread activation patterns is associated with better performance. Clarifying the relationships between structural and functional neuroanatomy and

cognitive performance remains one of the central questions in cognitive aging research.

See Also the Following Articles

AGING BRAIN • ALZHEIMER'S DISEASE, NEUROPSYCHOLOGY OF • BASAL GANGLIA • COGNITIVE REHABILITATION • MEMORY, EXPLICIT AND IMPLICIT • NEUROTRANSMITTERS • STROKE • WORKING MEMORY

Suggested Reading

Cabeza, R. (2000). Functional neuroimaging of cognitive aging. In *Handbook of Functional Neuroimaging of Cognition* (R. Cabeza, and A. Kingstone, Eds.), MIT Press, Cambridge, MA.

Craik, F. I. M., and Salthouse, T. A. (Eds.) (2000). *Handbook of Aging and Cognition—II*. Erlbaum, Mahwah, NJ.

Meyerson, J., Hale, S., Wagstaff, D., Poon, L. W., and Smith, G. A. (1990). The information-loss model: A mathematical theory of age-related cognitive slowing. *Psychol. Rev.* **97,** 475–487.

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychol. Rev.* **103,** 403–428.

Cognitive Psychology, Overview

JOHN F. KIHLSTROM and LILLIAN PARK
University of California, Berkeley

GLOSSARY

counterfactual emotions Counterfactual arguments involve reasoning that makes assumptions contrary to the facts in evidence (e.g., "If I were king, I'd make everyone rich"). Counterfactual emotions are feeling states, such as regret and disappointment, that require a comparison between some state of affairs and what might have been.

dichotic listening A technique in which different auditory messages are presented over separate earphones; the subject is instructed to repeat (shadow) one message but ignore the other.

dissociation A statistical outcome in which one variable, either a subject characteristic (such as the presence of brain damage) or an experimental manipulation (such as the direction of attention), has different effects on two dependent measures (such as free recall or priming).

functional magnetic resonance imaging A brain-imaging technique using magnets to measure the changes in the ratios of deoxygenated to oxygenated hemoglobin due to brain activity.

gambler's fallacy The idea that prior outcomes, such as a string of "red" numbers in roulette, can influence the outcome of some future outcome, such as a "black" number; it is a fallacy because in a truly random game each outcome is independent of the others.

magnetoencephalography A brain-imaging technique using superconducting quantum interference devices to measure changes in weak magnetic fields caused by the brain's electrical activity.

positron emission tomography A brain-imaging technique that uses positrons (positively charged electrons) to measure blood flow, metabolic rate, and biochemical changes in the brain.

priming The facilitation (or, in the negative case, inhibition) of perceptual–cognitive processing of a target stimulus by prior presentation of a priming stimulus.

schemata Organized knowledge structures representing a person's beliefs and expectations, permitting the person to make inferences and predictions.

sensory thresholds In psychophysics, the minimum amount of energy required for an observer to detect the presence of a stimulus (the "absolute" threshold) or a change in a stimulus (the "relative" threshold).

tabula rasa Latin meaning "blank slate"; refers to the empiricist view that there are no innate ideas, and that all knowledge is gained through experience.

Cognition has to do with knowledge, and cognitive psychology seeks to understand how human beings acquire knowledge about themselves and the world, how this knowledge is represented in the mind and brain, and how they use this knowledge to guide behavior.

I. A SHORT HISTORY

Psychology was cognitive at its origins in the mid- to late 19th century. Structuralists such as Wilhelm Wundt and E. B. Titchener attempted to decompose conscious experience into its constituent sensations, images, and feelings. On the very first page of the *Principles of Psychology* (1890), the discipline's founding text, William James asserted that "the first fact for us, then, as psychologists, is that thinking of some sort goes on," and the functionalist tradition that he and John Dewey established sought to understand the role of thinking and other aspects of mental life in our adaptation to the environment. In

the early 20th century, however, John B. Watson attempted to remake psychology as a science of behavior rather than, as James had defined it, a science of mental life.

For Watson, public observation was the key to making psychology a viable, progressive science. Because consciousness (not to mention "the unconscious") was essentially private, Watson argued that psychology should abandon any interest in mental life and instead confine its interest to what could be publicly observed: behavior and the circumstances in which it occurred. In Watson's view, thoughts and other mental states did not cause behavior; rather, behavior was elicited by environmental stimuli. Thus began the behaviorist program, pursued most famously by B. F. Skinner, of tracing the relations between environmental events and the organism's response to them. Psychology, in the words of one wag, lost its mind.

The behaviorist program dominated psychology between the two world wars and well into the 1950s, as manifested especially by the field's focus on learning in nonhuman animals, such as rats and pigeons. Gradually, however, psychologists came to realize that they could not understand behavior solely in terms of the correlation between stimulus inputs and response outputs. E. C. Tolman discovered that rats learned in the absence of reinforcement, whereas Harry Harlow discovered that monkeys acquired general "sets" through learning as well as specific responses. Noam Chomsky famously showed that Skinner's version of behaviorism could not account for language learning or performance, completely reinventing the discipline of linguistics in the process, and George Miller applied Chomsky's insights in psychology. Leo Kamin, Robert Rescorla, and others demonstrated that conditioned responses, even in rats, rabbits, and dogs, were mediated by expectations of predictability and controllability rather than associations based on spatiotemporal contiguity. These and other findings convinced psychologists that they could not understand the behavior of organisms without understanding the internal cognitive structures that mediated between stimulus and response.

The "cognitive revolution" in psychology, which was really more of a counterrevolution against the revolution of behaviorism, was stimulated by the introduction of the high-speed computer. With input devices analogous to sensory and perceptual mechanisms, memory structures for storing information, control processes for passing information among them, transforming it along the way, and output devices analogous to behavior, the computer provided a tangible model for human thought. Perceiving, learning, remembering, and thinking were reconstrued in terms of "human information processing," performed by the software of the mind on the hardware of the brain. Artificial intelligence, simulated by the computer, became both a model and a challenge for human intelligence.

Jerome Bruner and George Miller founded the Center for Cognitive Studies at Harvard University in 1960, intending to bring the insights of information theory and the Chomskian approach to language to bear on psychology. Miller's book, *Plans and the Structure of Behavior* (1960; written with Karl Pribram and Eugene Galanter) replaced the reflex arc of behaviorism with the feedback loops of cybernetics. The cognitive (counter)revolution was consolidated by the publication of Neisser's *Cognitive Psychology* in 1967 and the founding of a scientific journal by the same name in 1970. With the availability of a comprehensive textbook on which undergraduate courses could be based, psychology regained its mind.

II. THE DOMAIN OF COGNITION

Although some philosophers (including Plato, Descartes, and Kant) have asserted that some knowledge is innate, most also agree that at least some knowledge is acquired through experience. Accordingly, theories of human cognition must include some account of the sensory and perceptual processes by which the person forms internal, mental representations of the external world; the learning processes by which the person acquires knowledge through experience; the means by which these representations of knowledge and experience are stored more or less permanently in memory; the manner in which knowledge is used in the course of judgment, decision making, reasoning, problem solving, and other manifestations of human intelligence; and how one's thoughts and other mental states are communicated to others through language. We cannot hope to give a comprehensive analysis of these processes in this article. Detailed treatment is provided by the textbooks listed under Suggested Reading and also in the multivolume *Handbook of Perception and Cognition*, which has appeared serially since 1994. Instead, we seek only to orient the reader to the general thrust of work in this field and to the problems and controversies that occupy its practitioners.

A. Sensation

Philosophers of mind have debated two views about the origins of knowledge: nativism and empiricism. Cognitive psychology, while acknowledging the possibility that some knowledge is innate, favors the empiricist view that most knowledge is acquired through the senses, including our reflections on sensory experience. Therefore, cognitive psychology begins with an analysis of the sensory mechanisms by which physical energies arising from a stimulus are transformed into neural impulses.

Research on sensation is dominated by the definition of the various sensory modalities (vision, audition, etc.), the determination of thresholds for sensory experience, the physical basis of various qualities of sensation (e.g., blue, C-sharp, and sour), and the search for *psychophysical laws* that would relate the physical properties of a stimulus to the psychological properties of the corresponding sensory experience. The most general psychophysical law, Stevens' law ($S=kI^n$), holds that there is some exponent that will relate any physical property of a stimulus to the psychological property of its corresponding sensory experience. The analysis of sensation is so closely tied to the physical and biological sciences that it is often omitted from cognitive psychology textbooks. However, even such "lower" mental processes as sensation and perception do not escape the influence of "higher" mental processes of judgment and decision making.

For example, the early psychophysicists assumed that the detection of an object in the environment was simply a matter of the physical intensity of the stimulus and the sensitivity of the corresponding receptor organs. If a light were of sufficient intensity, given the modality and species in question, it would be detected (a "hit"); otherwise, it would be missed. However, it is also the case that observers will miss stimuli that are clearly above threshold and make false alarms by "detecting" stimuli that are not actually present. Experiments based on signal detection theory use the pattern of hits and false alarms to decompose performance into two parameters: Sensitivity, presumably closely tied to the biology of the sensory system, and bias, or the perceiver's willingness to report the presence of a stimulus under conditions of uncertainty. Interest in most signal detection experiments focuses on sensitivity; bias is a nuisance to be evaluated and statistically controlled. However, the fact that bias occurs at all shows that the perceiver's expectations, motives, and biases influence performance even in the simplest sensory task. Thus, processes involved in reasoning, judgment, choice, and decision making percolate down to even the lowest levels of the information processing system, and they are grist for the cognitive psychologist's mill.

B. Perception

While sensation has to do with detecting the presence of stimuli and changes in the stimulus field, perception has to do with forming mental representations of the objects that give rise to sensory experiences. Much perceptual research focuses on the process by which individuals determine the size, shape, distance, and motion of objects in the environment.

For most of its history, the study of perception has been dominated by the constructivist/phenomenalist tradition associated with Hermann von Helmholtz, Richard Gregory, Julian Hochberg, and Irvin Rock (among many others). The constructivist view assumes that the proximal stimulation impinging on sensory receptors is inherently ambiguous, and that there is an infinite array of distal configurations compatible with any momentary state of proximal stimulation. For example, consider the relationship between the size and distance of an object, on the one hand, and the size of the retinal image of that object, on the other hand. Holding distance constant, the size of the retinal image is directly proportional to the size of the stimulus; however, holding size constant, the size of the retinal image is inversely proportional to the distance between the object and the perceiver. Thus, given the size of a retinal image, the perceiver does not know whether he or she is viewing a large object far away or a small object close at hand.

According to the constructivist view, stimulation of this sort must be disambiguated by inference-like rules that compare the size of the object with the size of its background or make some comparison of the stimulus input with an a priori model of the world that tells us the size of various objects. In either case, perceiving entails thinking and problem solving. Sometimes, the solution can be wrong, such as when the perceptual system overcompensates for distance cues to generate the illusion that the moon on the horizon is larger than the moon at zenith. Helmholtz famously argued that the inferential rules that guide perception are part of our tacit knowledge: They can be discovered by the scientist but cannot be articulated by the perceiver. Because the thoughts that give rise to our percepts are unconscious, perception lacks the phenomenal quality

of thought. However, it remains the case that the final product of perception is a mental representation of the stimulus world, constructed by cognitive operations such as computations and symbolic transformations. We are not aware of the world itself but only of our mental representation of it, which is projected onto the world so that the objects of perception and the objects of the world are coreferential. Even more than sensation, perception from the constructivist/phenomenalist view has cognitive underpinnings that cannot be denied.

Nevertheless, a contrary, noncognitive view of direct realism was proposed by J. J. Gibson in his ecological theory of perception (interestingly, Neisser took a constructivist approach to perception in Cognitive Psychology, but has since embraced a version of direct realism). According to the ecological view, stimulation is ambiguous only at very elementary levels, but there is no ambiguity at higher levels. Therefore, for example, in determining an object's size, the perceptual system extracts information about the ratio of the size of an object to the size of its background; this ratio determines perceived size, not the size of the retinal image of the object alone. Thus, the perceived size of an object remains constant even as its distance from the viewer (and thus the size of its corresponding retinal image) varies. However, this requires no computations, inferences, or a priori models of the world on the part of the perceiver; size is perceived directly from information available in the environment about the ratio of the figure to ground, without need of any mediating cognitive operations. Because perceptual systems evolved in order to support adaptive behavior, Gibson further proposed that we perceive objects in terms of their affordances, or the actions that we can take with respect to them. Thus, in the same way that the ecological view of perception argues that all the information required for perception is "in the light," an ecological view of semantics argues that the meanings of words are "in the world," available to be perceived directly.

The ecological theory of perception proposes that the perception of form, distance, motion, and other stimulus properties is no different from perceiving the hue of a light or the pitch of a sound. In each case, phenomenal experience occurs by virtue of the transduction of stimulation into perception, accomplished in a single step by specialized neuronal structures that have evolved to be selectively sensitive to higher order variables of stimulation available in an organism's environmental niche. The contrast between the constructivist/phenomenalist and direct/realist views dominates much of contemporary perception research, with proponents of the ecological approach conducting clever experiments showing that percepts commonly attributed to computations, inferences, or world models are actually given directly by higher-order variables of stimulation. However, it is one thing to demonstrate that such information is available and quite another to demonstrate that such information actually contributes to perception. The occurrence of visual illusions strongly suggests that we do not always see the world as it really is, and that the perceiver must, in Bruner's famous phrase, go "beyond the information given" by the environment in order to form mental representations of the world.

C. Attention

In many theories, attention is the link between perception and memory: The amount of attention devoted to an event at the time it occurs (i.e., at encoding) is a good predictor of the likelihood that it will be consciously remembered later (i.e., at retrieval). Early cognitive theories considered attention to be a kind of bottleneck determining whether incoming sensory information would reach short-term memory and thus enter into "higher level" information processing. In early attention research there was a major controversy between early selection theories, which held that preattentive processing was limited to "low-level" analyses of physical features, and late selection theories which allowed preattentive processing to include at least some degree of "high-level" semantic analysis. Early selection was favored by experiments showing that subjects had poor memory for information presented over the unattended channel in dichotic listening experiments. Late selection was favored by evidence that such subjects were responsive to the presentation of their own names over the unattended channel.

Definitive tests of early versus late selection proved difficult to come by, and beginning in the 1970s the problem of attention was reformulated in terms of mental capacity: According to capacity theories, individuals possess a fixed amount of processing capacity, which they can deploy freely in the service of various cognitive activities. Various information processing tasks, in turn, differ in terms of the amount of attentional capacity they require. Some tasks may be performed automatically, without requiring any attentional capacity at all; such tasks do not interfere

with each other or with effortful tasks that do make demands on cognitive resources. When the total attentional capacity required by effortful tasks exceeds the individual's capacity, they will begin to interfere with each other. Some automatic processes are innate; however, other processes, initially performed effortfully, may be automatized by extensive practice. Thus, skilled readers automatically and effortlessly decode letters and words, even while they are doing something else, whereas unskilled readers must expend considerable mental effort performing the same task at great cost to other, ongoing activities.

According to one prominent view, automatic processes are almost reflexive in nature (although the "reflexes" in nature are cognitive, not behavioral, and they are acquired, not innate). That is, they are inevitably engaged by the appearance of certain stimuli, and once invoked they proceed inevitably to their conclusion. Because their execution consumes no attentional resources, they do not interfere with other ongoing processes, and they leave no traces of themselves in memory. This "attention-based" notion of automaticity plays a central role in many cognitive theories. According to a revisionist "memory-based" view, however, automaticity has nothing to do with attention but, rather, depends on the way in which skill underlying task performance is represented in memory. Automatization occurs when performance is controlled by procedural rather than declarative knowledge representations. From either point of view, automatic processes are strictly unconscious: We have no direct introspective awareness of them, and we know them only by inference from task performance.

D. Memory

Perceptual activity leaves traces in memory, freeing behavior from dominance by stimuli in the immediate present. The knowledge stored in memory consists of two broad forms: declarative knowledge, which can be either true or false, and procedural knowledge of how certain goals are to be accomplished. Procedural knowledge can be further classified into cognitive and motor skills, such as one's knowledge of arithmetic or grammar, how to tie one's shoes, or how to drive a standard-shift car. Similarly, declarative knowledge can be subdivided into episodic memories of specific experiences that occurred at a particular point in space and time, such as one's memory of eating sushi for dinner at home last Thursday, and semantic memories that are more generic in nature, such as one's knowledge that sushi is a Japanese dish made of rice, vegetables, and fish. In theory, many semantic memories are formed by abstraction from related episodic memories, and much procedural knowledge represents a transformation of declarative knowledge.

Most research on memory has focused on episodic memories for specific events and is based on an analysis of memory into three stages of encoding, storage, and retrieval. Early views of memory that made a structural distinction between short- and long-term stores have now been replaced by a unitary view in which "short-term" (or "working") memory refers to those items that are actively engaged in processing at any moment. Earlier views of forgetting as a product of the loss of memories from storage have been replaced by the view that retention is a function of the extent of processing received by an item at the time of encoding and the amount of cue information available at the time of retrieval. The relations between encoding and retrieval processes are effectively captured by a general principle of encoding specificity (also known as transfer-appropriate processing), which states that the likelihood that an event will be remembered depends on the match between the information processed at the time of encoding and the information available at the time of retrieval.

Most research on memory has employed experimental tasks requiring conscious recollection, or the ability of subjects to recall or recognize past events. However, episodic memory may also be expressed implicitly in tasks that do not require conscious recollection in any form. For example, a subject who has recently read the word "veneer" will be more likely to complete the stem "ven" with this word than with the more common word "vendor." A great deal of experimental research shows that such priming effects can occur regardless of whether the study word is consciously remembered; in fact, they can occur in amnesic patients who have forgotten the study session in its entirety. Similarly, amnesic patients can learn new concepts without remembering any of the instances they have encountered, and they can acquire new cognitive and motor skills while failing to remember the learning trials. Along with the concept of automaticity, the dissociations observed between explicit and implicit expressions of memory have given new life to the notion of the psychological unconscious.

The dissociations observed between explicit and implicit expressions of episodic memory, between

semantic and episodic memories, and between procedural and declarative knowledge are subject to a variety of interpretations. According to the multiple systems view, explicit (conscious) and implicit (unconscious) memories are served by different memory systems in the brain. The multiple systems view, in turn, is compatible with the neuroscientific view of the brain as a collection of modules, each specialized for a particular information-processing task. In contrast, researchers who prefer a processing view, while accepting that there is some degree of specialization in the brain, explain these same dissociations as generated by different processes that operate in the context of a single memory system. For example, according to one processing view, implicit memories are the product of automatic, attention-free processes, whereas explicit memories are the product of effortful, attention-demanding ones. In general, processing views are compatible with computational theories of memory, which typically assume that different memory tasks require the processing of different features of memories stored in a single memory system. One of the interesting features of the debate over explicit and implicit memory is how little contact there has been between neuroscientific and computational views of memory.

E. Categorization

Memory also stores conceptual knowledge about things in general as well as representations of specific objects and events. Bruner noted that this conceptual knowledge plays an important role in perception: In fact, every act of perception is an act of categorization. A great deal of research in cognitive psychology has sought to understand the way in which conceptual knowledge is organized in the mind.

According to the classical view handed down by Aristotle, concepts are represented by a list of features that are singly necessary and jointly sufficient to define the category in question. For example, in geometry, all triangles are closed two-dimensional figures with three sides and three angles, and a sharp boundary divides all triangles from all quadrilaterals. However, in the 1970s it became clear that however satisfying such a definition might be philosophically, it did not reflect how concepts are represented in human minds. When perceivers judge equilateral and right triangles to be ''better'' triangles than isosceles triangles, they are referring to something other than a list of defining features. According to the classical view, all members of a category are equally good representatives of that category. For this and other reasons, the classical view of concepts as proper sets has been replaced with a revisionist probabilistic view of concepts as fuzzy sets. According to the fuzzy set view, features are only imperfectly correlated with category membership, and concepts themselves are represented by prototypes (real or imagined) that possess many features that are characteristic of category members. The probabilistic view permits some instances (e.g., robin) of a category (bird) to be better than others (e.g., emu), even though all possess the same set of defining features. Moreover, it permits the boundaries between categories to be somewhat blurred (is a tomato a fruit or a vegetable?).

Both the classical and the probabilistic view regard concepts as summary descriptions of category members. However, an alternative exemplar view holds that concepts are represented as collections of instances rather than as summary descriptions. Thus, when we seek to determine whether an object is a bird, we compare it to other birds we know rather than to some abstract notion of what a bird is. Just as there is empirical evidence allowing us to firmly reject the classical view of conceptual structure as inadequate, so too there are studies showing that objects are slotted into categories if they resemble particular instances of the category in question, even if they do not resemble the category prototype. Perhaps novices in a domain categorize with respect to abstract prototypes, whereas experts categorize with respect to specific exemplars.

Regardless of whether concepts are represented by prototypes or exemplars, categorization is a special case of similarity judgment: The perceiver assigns an object to a category by matching its features to those of his or her category representation, prototype or exemplar. There is no absolute threshold for similarity, however: Categorization, like signal detection, is always a matter of judgment.

Although categorization seems to be special case of similarity judgment, and the most recent development in theories of concepts has been stimulated by evidence of certain anomalies of similarity. For example, subjects judge gray clouds to be similar to black clouds and different from white clouds, but they judge gray hair to be similar to white hair but different from black hair. The brightness of the color patches is identical, so the judgment must be based on something other than perceptual similarity, such as the perceiver's theory about how hair changes with age or how clouds change with the weather. According to the theory-based view of categorization, concepts are not represented by lists

of features or instances, and categorization does not proceed by feature matching. Instead, concepts are represented by theories that make certain features and instances relevant and that explain how features and instances are related to each other, and categorization proceeds by applying the theory to the case at hand. It remains to be seen, however, whether the theory-based view of concepts and categorization will supplant, or merely supplement, the similarity-based view.

F. Learning

Behaviorism was dominated by an emphasis on learning, but cognitive psychology has not abandoned the question of how knowledge is acquired. After all, although knowledge of such basic categories as time and space may be innate, most knowledge is derived from experience. Learning, then, is the process of knowledge acquisition. In fact, some of the earliest cognitive challenges to behaviorism came through alternative accounts of learning. Even such basic processes as classical and instrumental conditioning are now interpreted in terms of the organism's developing ability to predict and control environmental events. Pavlov's dogs did not salivate to the bell because it occurred in close spatiotemporal contiguity with meat powder, and Skinner's pigeons did not peck at the key because it was reinforced by the delivery of food in the presence of a certain light. Rather, they did so because they expected food to follow the bell and the keypeck. Learning occurs in the absence of reinforcement; reinforcement controls only performance, the organism's display of what it has learned.

The importance of expectancies, and the limited role played by contingencies of reinforcement, is underscored by the development of theories of social learning by Julian Rotter, Albert Bandura, Walter Mischel, and others, who argued that human learning rarely involved the direct experience of rewards and punishments. Rather, most human learning is vicarious in nature: It occurs by precept, in the sense of sponsored teaching, or by example, as in observational modeling. In either case, we learn by watching and listening to other people. Bandura argued that behavior was controlled not by environmental stimuli but by expectancies concerning the outcomes of events and behaviors and also by self-efficacy expectations (i.e., people's belief that they can engage in the behaviors that produce desired outcomes). Some clinical states of anxiety may be attributed to a (perceived) lack of predictability in the environment, whereas some instances of depression may be attributed to a perceived lack of controllability. Interestingly, a capacity for observational learning has been uncovered in nonhuman animals, such as rhesus monkeys, and has been implicated in the genesis of animal "cultures."

Learning processes are obviously implicated in analyses of the encoding stage of memory processing, in the acquisition of procedural knowledge, and in concept formation. At the same time, cognitive psychologists have generally avoided the topic of learning itself. This may partially reflect an overreaction to the excessive interest in learning on the part of behaviorists, and it may partially reflect the influence of Chomsky, who discounted the role of learning in the development of language. Recently, this situation has changed due to the rise of parallel distributed processing, interactive activation, neural network, or connectionist models as alternatives to traditional symbolic processing models of human information processing. In symbolic models, each individual piece of knowledge is represented by a node, and discrete nodes are connected to each other to form a network of associative links. Thus, a node representing the concept *doctor* is linked to semantically related nodes representing concepts such as *nurse* and *hospital*. Such models are very powerful, but they leave open the question of how the knowledge represented by nodes is acquired in the first place. Connectionist models assume that individual concepts are represented by a pattern of activation existing across a large network of interconnected nodes approximately analogous to the synaptic connections among individual neurons (hence, the alternative label). No individual node corresponds to any concept; every concept is represented by a pattern of widely distributed nodes. Instead of one node activating another one in turn, all nodes are activated in parallel, and each passes activation to each of the others. In connectionist systems, learning occurs as the pattern of connections among nodes is adjusted (sometimes through a learning algorithm called backpropagation) so that stimulus inputs to the system result in the appropriate response outputs. Although the link between connectionism and stimulus–response behaviorism is obvious, connectionist theories are cognitive theories because they are concerned with the internal mental structures and processes that mediate between stimulus and response. Compared to traditional symbolic models, they are extremely powerful and efficient learning devices. Unfortunately, they also display a

disconcerting tendency to forget what they have learned as soon as they are asked to learn something new—a phenomenon known as catastrophic interference. Moreover, although connectionist models seem to reflect the neural substrates of learning, it has proved difficult to demonstrate the biological plausibility of specific features such as backpropagation. Accordingly, the future of connectionist models of information processing remains uncertain.

G. Language and Communication

Language is both a tool for human thought and a means of human communication. The ability of people to generate and understand sentences that have never been spoken before is the hallmark of human creative intelligence and arguably the basis for human culture. Also, language permits us to convey complex information about our thoughts, feelings, and goals to other people, and it provides a highly efficient mechanism for social learning.

Language was one of the first domains in which cognitive psychology broke with behaviorism. Early cognitive approaches to language were couched in terms of information theory, but a real breakthrough came in the late 1950s and early 1960s with the work of Chomsky. In his early work, Chomsky distinguished between the surface structure of a sentence, viewed simply as a sequence of words, its phrase structure in terms of noun phrases and verb phrases, and its deep structure or underlying meaning. He also argued that transformational grammar mediated between deep structure and phrase structure. Transformational grammar is universal and innate. Language acquisition consists of learning the specific rules that govern the formation of surface structures in a particular language.

The field of psycholinguistics largely arose out of attempts to test Chomsky's early views. For example, it was shown that clicks presented while sentences were being read were displaced from their actual location to the boundaries between phrases, and that the time it took to understand a sentence was determined by the number of transformations it employed. Nevertheless, over the ensuing years this theory has been substantially altered by Chomsky and colleagues, and in some quarters it has even been discarded. In any event, it is clear that other aspects of language are important besides grammatical syntax. Languages also have phonological rules, which indicate what sounds are permitted and how they can be combined, and morphological rules, which constrain how new words can be formed. Moreover, research on the pragmatics of language use shows how such aspects of nonverbal communication as tone of voice, gesture, facial expressions, posture, and context are employed in both the production and the understanding of language.

A major controversy in the psychology of language concerns speech perception. According to the motor theory of speech perception proposed by Alvin Liberman and colleagues, speech is special in the sense that it is processed by mechanisms that are part of a specifically human cognitive endowment; no other species has this capacity. According to a rival auditory theory of speech perception, understanding speech is simply a special case of auditory perception, and requires no special capacities that are unique to humans. Tests of these theories often revolve around categorical perception, or the ability to distinguish between related sounds such as [b] and [p]. Evidence that there are sharp boundaries between such speech categories is often attributed to innate, specifically human, mechanisms for producing speech, thus favoring the motor theory. On the other hand, evidence of categorical perception in nonhuman species that do not have a capacity for speech favors the auditory theory.

A common theme in the Chomskian approach to language is that language use is mediated by innate rules that cannot be acquired by general-purpose systems that learn solely by virtue of associations among environmental events. Chomsky claims that support for the role of innate rules comes from the errors children and other language learners make in forming the past tenses of irregular verbs; for example, "goed" instead of "went" or "eated" instead of "ate." Because children have never heard such words (the adults they listen to do not speak them), the words cannot have been acquired through experience; rather, they must be produced by a rule that the child has abstracted in the course of learning his or her native tongue. If computers are to have a language capacity, they must be programmed with these kinds of rules; they will never learn language without some rule-based cognitive structure. Recently, however, connectionist models of language have been developed that have no rules of syntax and operate solely by associationistic principles but that make precisely the errors that are traditionally attributed to the operation of grammatical rules. If so, the implication may be that human language is special after all: It is something that can be

done by any associationistic learning system that possesses sufficient computational power. Proponents of rules have criticized these demonstrations as misleading and unrepresentative of people's actual language use. For example, the model makes mistakes that children do not make, and it learns by virtue of inputs that do not resemble children's actual learning environments. Because the capacity for language is so central to our traditional conception of what it means to be human, the debate between rules and connections is likely to be vigorous and protracted.

H. Judgment, Reasoning, and Problem Solving

Thinking played a prominent role in early psychology. The structuralist school of Wundt and Titchener was almost consumed by a fruitless debate over imageless thought, whereas Oswald Kulpe's act psychology attempted to characterize the process of thinking rather than the static elements of thoughts. Like other mentalistic topics, thinking dropped out of sight with the rise of behaviorism, but interest in the topic was preserved by Gestalt psychologists such as Kohler, who worked on problem solving in chimpanzees. Whereas behaviorists construed thinking as a matter of gradual trial-and-error learning, the Gestaltists emphasized sudden insights produced by a cognitive restructuring of the problem at hand. In England, Frederick C. Bartlett argued that perception and memory were essentially exercises in problem solving—the problem being to construct mental representations of the present and to reconstruct mental representations of the past. After World War II, the cognitive revolution was heralded by Bruner's work on concept learning, Jean Piaget's research on the development of thought in children, and the work of Herbert Simon and Allan Newell on computer simulations of human problem solving.

Much early research on thinking was guided, at least tacitly, by a normative model of human rationality that held that people reason according to a logical calculus and make rational choices based on principles of optimality and utility. According to the classical view of concept structure, for example, people categorize objects according to lists of defining features that are singly necessary and jointly sufficient to define category membership. In classical decision theory, individuals calculate the costs and benefits to themselves of various options and then make choices that maximize their gains and minimize their losses as efficiently as possible. Much work on reasoning was dominated by the search for algorithms: Logical, systematic rules, analogous to recipes, that specify how information should be combined to yield the correct solution to whatever problem is at hand.

Although normative rationality provided a reasonable starting point for developing theories of human reasoning and problem solving, many human judgments must be made under conditions of uncertainty, where there is no algorithm applicable or where the information needed to apply an algorithm is unavailable. Other judgments must be made under conditions of complexity, where there are simply too many choices available, or too many factors entering into each choice, to permit evaluation according to some judgment algorithm. Under such conditions, people tend to rely on judgment heuristics, which are shortcut "rules of thumb" that bypass normative rules of logical inference and thus permit judgments without recourse to algorithms. One such heuristic, associated with the work of Herbert Simon on organizational decision making, is satisficing: Instead of conducting an exhaustive search for the optimal choice, a judge may terminate search as soon as the first satisfactory option is encountered. Prototype or exemplar matching constitute heuristics for categorization: Instead of consulting a list of defining features that are singly necessarily and jointly sufficient to assign an object to some category, people compare the object at hand to a "typical" instance, or indeed to any instance at all.

Although appropriate algorithms are guaranteed to deliver the correct solution to whatever problem is at hand, use of heuristics incurs some risk of making an error in reasoning or judgment. Analysis of common judgment errors, such as the "gambler's fallacy" described by Daniel Kahneman, Amos Tversky, and others, has documented many other commonly used judgment heuristics. Representativeness permits judgments of category membership, similarity, probability, and causality to be based on the degree to which an event resembles the population of events from which it has been drawn. Availability permits judgments of frequency and probability to be based on the ease with which relevant examples can be brought to mind, whereas simulation bases judgments on the ease with which plausible scenarios can be constructed. In anchoring and adjustment, initial estimates are taken as reasonable approximations to the final result of some calculation.

These and other effects show that the principles of cognitive functioning cannot simply be inferred from abstract logical considerations; rather, they must be

inferred from empirical data showing how people actually perform. Research shows that people commonly depart from the principles of normative rationality, but what should be made of these departures? Although Aristotle defined humans as rational animals, one possible conclusion from empirical studies is that people are fundamentally irrational—that human judgment, reasoning, choice, and problem solving are overwhelmed by many fallacies, illusions, biases, and other shortcomings. At best, according to this argument, most people are "cognitive misers" who use as little information, and as little cognitive effort, as possible in their lives; at worst, people are just plain stupid—incapable, without extensive instruction (and perhaps not even then), of conforming themselves to the principles of logic and rationality.

This pessimistic conclusion about human nature is reminiscent of Sigmund Freud's argument at the turn of the 20th century that human rationality is derailed by unconscious affects and drives. On the other hand, it is possible that the case for human irrationality has been overstated. For example, the philosopher Jonathan Cohen questioned whether the formal laws of deductive and probabilistic reasoning are properly applied to the problems that people actually encounter in the ordinary course of everyday living. Herbert Simon concluded that human rationality is bounded by limitations on human information-processing capacity (as demonstrated, for example, by George Miller's famous essay on "the magical number seven, plus or minus two"). From this perspective, it is simply unreasonable to hold humans up to an impossible standard of unbounded rationality of a sort that might characterize a computer, which has the capacity to search and calculate for as long as it takes to deliver a "logical" result. Relatedly, Gerd Gigerenzer and colleagues argued that "fast and frugal" judgment heuristics succeed more often than they fail because they are appropriately tuned to the structure of the environments in which people actually operate. From this perspective, most "fallacies" in human reasoning emerge in performance on laboratory tasks that do not adequately reflect the real world in which judgment heuristics work. On the other hand, some evolutionary psychologists have argued that judgment heuristics are part of an "adaptive toolbox" of domain-specific cognitive devices (or modules) that evolved in the "environment of evolutionary adaptedness"—namely, the African savanna of the late Pleistocene era, approximately 500,000 years ago—to help our hominid ancestors solve fundamental problems of survival and reproduction.

III. COGNITIVE DEVELOPMENT

Compared to other mammals, human beings are born with relatively immature brains; moreover, physical development continues even after brain development has essentially completed. These facts raise the question of how the intellectual functions characteristic of the human adult arise in the first place (and, from a life span perspective, whether, how, and to what extent cognitive skills are lost through aging).

A. The Ontogenetic View

From an ontogenetic point of view, tracing the growth of cognition in the individual organism, cognitive development has recapitulated the debate between nativism and empiricism that has dominated cognitive psychology at large. From the empiricist perspective, the child is a *tabula rasa* who acquires knowledge and skills with learning and experience. From the nativist perspective, even neonates possess at least primitive cognitive faculties, which develop further from interaction with the environment. The tension between nativism and empiricism can be clearly seen in the debate, discussed earlier, regarding whether language acquisition is mediated by innate grammatical rules or by a general-purpose associative learning mechanism.

A new perspective on cognitive development, combining elements of both nativism and empiricism, was offered by Jean Piaget, who proposed that children enter the world with a rudimentary set of reflex-like cognitive structures, called sensory motor schemata, through which they interact with the world. Environmental events are interpreted through prevailing cognitive schemata, but they also force these schemata to change in order to cope with an increasingly complicated stimulus environment. Through the cycle of assimilation and accommodation, the child moves through many qualitatively different stages, each highlighted by a particular cognitive achievement. The milestone marking the end of the sensory motor stage, at about 18 months of age, is object permanence—the ability to deal with objects that are not present in the immediate physical environment. The end of the preoperational stage, at about age 7, is marked by conservation—the ability to appreciate that quantities remain constant despite changes in physical appearance. The transition from concrete operations to formal operations, at about age 12, is marked by the

child's ability to comprehend abstract concepts and formal logical relations.

According to Piaget, the child proceeds through these stages of cognitive growth in a strict sequence: Some tasks, requiring abilities characteristic of later stages, are simply impossible for a child who is still at an earlier stage. Such a proposal was bound to be challenged, and indeed later experiments employing extremely subtle measures often showed that, as a 1993 cover story in Life magazine stated, "Babies are smarter than you think." For example, 7-month-old infants may not reach for the spot where a toy has been hidden, but they do stare at it, indicating that they have some sense of object permanence. Similarly, when a mouse hidden behind a screen is joined by a second mouse, 5-month-olds show surprise when the screen is removed to reveal only one mouse, suggesting that they have some ability to conserve number.

Results such as these suggest to some theorists that infants enter the world with a surprisingly sophisticated fund of innate knowledge about the world that is refined and elaborated through experience. What develops, then, is expertise: The infant starts out as a novice with respect to objects, numbers, and so on. Development proceeds with continuous increases in motor control and information-processing capacity and also with increased opportunities for learning through experience. Infants reach for the hidden toy not because they have acquired object permanence but because they have acquired the ability to coordinate their actions with their thoughts. Recent experiments, however, indicate that infants' appreciation of object permanence is incomplete. Developing children acquire new knowledge and skills, not just the ability to use innate knowledge and skills more efficiently and effectively.

Development also entails the acquisition of metacognition—one's knowledge of what one knows, how one's own mind works, and how this knowledge can be deployed strategically in the service of adaptive behavior. Metacognitive knowledge is sometimes characterized as a theory of mind, a phrase that recalls Piaget's argument that children, no less than adults, function as naïve scientists. From the first moments of life, children are constantly trying to understand themselves and the world around them, including other people, by generating hypotheses, testing them empirically, and refining their theories accordingly. The "theory" theory, as it is sometimes called, makes clear that cognitive development is not just something that happens passively to the child by virtue of maturation, learning, and the activities of adults. Rather, the child takes an active role in his or her own development, instigating the very interactions that promote cognitive growth.

B. Language, Culture, and Thought

Cognitive development can also be viewed in cultural terms. Cognitive anthropology had its origins in the efforts of Lucien Levy-Bruhl, Franz Boas, W. H. R. Rivers, and others to determine whether there were differences in the thought patterns characteristic of members of "primitive" and "advanced" cultures. Some early Soviet psychologists, such as Lev Vygotsky and Alexander Luria, attempted to trace the effects of economic development on how people think. In view of the central role of language in culture, considerable effort has been devoted to the question, initially raised by the American anthropologists Edward Sapir and Benjamin Whorf, whether there are cognitive differences between speakers of different languages. This is not a matter of "development" per se because all languages are equally complex: It is merely a matter of language and culture.

The Sapir–Whorf hypothesis takes two forms: That language determines thought or that language influences thought. The former is a much stronger view because it states that one is incapable of understanding a concept for which the language has no name (it also implies that there is no thought without language). There is no empirical evidence supporting the strong version and considerable evidence that thought can proceed without benefit of language. However, the weak version plausibly suggests that different languages can "carve up" the world into different ways or, in other words, that conceptual thinking can be shaped and constrained by available linguistic categories. As Whorf stated,

> *We cut nature up, organize it into concepts, ascribe significance as we do, largely because we are parties to an agreement to organize it in this way—an agreement that holds throughout our speech community and is codified in the patterns of our language.*

There are actually two aspects to the Sapir–Whorf hypothesis: Linguistic relativity and linguistic determinism. Relativity refers to the claim that speakers are required to pay attention to different aspects of the world that are grammatically marked (e.g., shape classifiers in Japanese or verb tenses to indicate time).

Determinism claims that our cognitive processes are influenced by the differences that are found in languages. The most famous example of the Whorf hypothesis, and the most erroneous, is Whorf's observation that Eskimos have many words for snow, implying that because they live in a snowy environment they need to make finer distinctions for the different types of snow. However, American skiers also have different words for snow, so the example is not as remarkable as it first may appear because expertise leads to larger vocabularies for certain domains.

In a classic test of the Sapir–Whorf hypothesis, Paul Kay and colleagues compared English speakers with Tarahumara speakers. Tarahumara is an Uto-Aztecan language of Mexico that does not have a separate color term for blue and green. In the first experiment, the subjects were presented with a blue color chip, a green color chip, and another color chip that was intermediate to blue and green. English speakers sharply distinguished the intermediate color chip into either blue or green by using a naming strategy, whereas the Tarahumara speakers chose randomly. In the second experiment, English speakers were first presented with two color chips and shown that one (intermediate) was greener than the other color chip (blue), and then they were shown that the same intermediate chip was bluer than the other color chip (green). By making the subjects call the intermediate color chip both green and blue, the bias that was demonstrated in the first experiment was removed and the English speakers performed similarly to the Tarahumara speakers.

The influence of language on how we think about the events that happen in our world can be demonstrated in experiments other than those designed to confirm or disconfirm the Whorf hypothesis. Classic work by Leonard Carmichael and colleagues demonstrated that subjects had different systematic distortions in their recall of ambiguous line drawings depending on which verbal label they were given (e.g., dumbbells or eyeglasses). Experiments on eyewitness testimony by Elizabeth Loftus and others showed that by varying the verb (e.g., ''crashed'' or ''hit'') one can manipulate the estimated speed of the traveling car given by the subjects. Whorf became interested in language when he noticed that behavior around gasoline drums changed when the drums were called ''empty'' even though they contained dangerous vapors. Because the word *empty* connotes ''lack of hazard,'' careless behavior from the workers resulted in fires from the tossing of cigarette stubs or smoking by the workers.

Beyond the influence of language on categories, the linguist George Lakoff's work on metaphors offers another way of testing the Sapir–Whorf hypothesis without depending on the idea that language carves the world into different pieces and, as he stated, ''cultures differ only in the way they have their meat cut up.'' Though some metaphors are universal (e.g., love is warmth), not all cultures share the same metaphors. Future research on the Sapir–Whorf hypothesis should be less reliant on the differences between ''exotic'' and ''nonexotic'' languages, a paradigmatic focus which sometimes implies that speakers of the exotic language are cognitively deficient.

C. The Phylogenetic View

Cognitive development can also be approached from a phylogenetic perspective, tracing the relations between the intellectual functions of human children and adults and those of other animals, especially the great apes, whose genetic endowment is similar to our own. There is a vigorous debate over whether chimpanzees and gorillas have anything like the human capacity for language, but it is clear that these and other animals do have the ability to acquire symbolic representations of objects, events, and concepts—similar to semantics, if not syntax as well. Pigeons can be taught to categorize a wide variety of objects, including trees, people (and their emotional expressions), fish, flowers, and automobiles. Studies of mirror recognition indicate that chimpanzees possess a rudimentary concept of self, and the notion of a ''theory of mind'' initially arose from observations that chimpanzees had the ability to attribute mental states to others of their kind. Setting aside the question of whether other species have a capacity for language, it is clear that the behavior of nonhuman animals, especially those who are closest to us in the evolutionary scheme of things, is not just a matter of innate and conditioned responses; some of them, at least, have cognitive capacities not unlike our own.

IV. COGNITION BEYOND PSYCHOLOGY

The cognitive revolution in psychology was paralleled by the development of the field of cognitive science, whose practitioners included philosophers, linguists, computer scientists, neuroscientists, behavioral biologists, sociologists, anthropologists, and psychologists. In some sense, the rise of cognitive science may have been a reaction to the dominance of behaviorism

within psychology: Many who wished to pursue a science of mental life may have believed that they would have to go outside psychology to do so. By the same token, it seems reasonable to hope that the combined efforts of many different disciplines are more likely to yield a better understanding of cognitive processes than any one working in isolation.

Whereas some early cognitive psychologists viewed the computer as a model of the human mind, some early cognitive scientists believed that it offered the prospect of implementing the "mechanical mind" debated by philosophers at least since the time of Descartes. In the formulation of the philosopher John Searle, work on artificial intelligence (AI) takes two broad forms. In "weak" AI, the computer provides a vehicle for writing formal theories of the mind, which can be tested by pitting the results of a computer simulation against the data of actual human performance. In terms of formal precision of its theories, weak AI is cognitive psychology at its best. In contrast, "strong" AI entails the notion that computer programs can, in principle, really think just as humans do. The program of strong AI has its origins in the proposal by Alan Turing that appropriately programmed computers are capable of performing any explicitly stated cognitive task: A machine would pass the Turing test if its responses were indistinguishable from those of a human being. Searle believes that the program of strong AI is seriously misguided—a position that is opposed with equal vigor by other philosophers such as Daniel Dennett.

Recently, research in AI has shifted from an effort to make machines think the way humans do to an effort to allow machines to "think" however they are capable of doing so, regardless of how humans might accomplish the same task. Thus, in May 1997 Deep Blue, a supercomputer programmed by IBM, was able to beat the world champion Gary Kasparov at chess, but nobody claimed that Deep Blue played chess the way Kasparov (or any other human) did. Cognitive psychology remains an important component of cognitive science. However, to the extent that it seeks to develop intelligent machines on their own terms, without reference to human intelligence, cognitive science departs from cognitive psychology.

Cognitive science, which once was dominated by behavioral experiments and computational models, has recently reached "down" to strengthen its connections to neuroscience. At the same time, neuroscience, which once was preoccupied with events at the molecular and cellular levels, has reached "up" to take an interest in the organismal level of experience, thought, and action. Both trends have been aided by the development of brain imaging techniques, such as positron emission tomography, functional magnetic resonance imaging (fMRI), and magnetoencephalography (MEG), which open windows to the brain as it is engaged in complex cognitive activities such as perceiving, remembering, imaging, and thinking. Particularly promising is the combination of the high spatial resolution of fMRI with the high temporal resolution of MEG. Just as earlier investigators discovered specific cortical areas specialized for vision, hearing, and so on, a new generation of brain researchers have uncovered specific areas activated in such activities as language comprehension, mathematical computation, analytical reasoning, and working memory.

The advent of new brain imaging techniques promises to solve the ancient mind–body problem, but they also present the danger of lapsing into a kind of high-technology revival of phrenology. In the final analysis, brain imaging can only reveal which areas of the brain are activated when experimental subjects engage in particular tasks, such as lexical decision and mental arithmetic. Discovering what these areas do requires a careful analysis of the experimental tasks employed in the imaging study, and this is a matter for cognitive psychology. If researchers do not have a correct description of the components of the task at the cognitive and behavioral level, they will reach erroneous conclusions concerning the cognitive functions of various parts of the brain. Solving the mind–body problem is not just a matter of building bigger magnets, resolving the details of neurochemistry, and ruling out physiological artifacts. Genuine advances in cognitive neuroscience depend on continued progress in cognitive psychology so that brain researchers can work with tasks that are well understood.

Even within the social sciences, it is clear that cognition is not just for psychologists anymore (if it ever was). Linguistics, traditionally concerned with the discovery of linguistic regularities and the origins of words, has increasingly worked to understand language as a tool of thought and means of sharing ideas. Economics, once concerned solely with the abstract description of economic systems, has recently turned its attention to individual economic decision making (and, in the process, drawing on the insights of psychologists such as Tversky and Kahneman). For cognitive sociologists, social conventions and norms create a framework in which individuals think their thoughts. By the same token, cognitive anthropologists are willing to entertain (and test) the hypothesis that cultural differences entail differences in modes of

thought as well as differences in beliefs and behavior. Sociology and anthropology have challenged the doctrines of individualism and universalism, which have traditionally dominated psychological approaches to human thought.

V. BEYOND COGNITION: EMOTION AND MOTIVATION

Cognitive psychology is about knowing, but knowing is not all that the mind does. In his Critique of Pure Reason (1791), the philosopher Immanuel Kant proposed that there are three "faculties of mind," knowledge, feeling, and desire. Each of these enters into a causal relationship with behavior, and none is reducible to any other. If Kant is right, then cognitive psychology cannot be all there is to psychology: The principles of cognition must be supplemented by principles of emotion and motivation. In fact, some cognitive psychologists have argued that Kant is wrong, and that our emotional and motivational states are the by-products of cognitive activity. For example, prominent cognitive theories of emotion hold that our emotional states are essentially beliefs about our feelings. In other words, cognitive theories of emotion hold that our emotional states depend on our interpretation of environmental events and our own behaviors. As William James famously stated, we do not run from the bear because we are afraid; we are afraid because we run from the bear. In response, some theorists have argued that emotions are not dependent on cognitive processing but, rather, are governed by their own independent systems. To some degree, such proposals reflect a reaction to the hegemony of the cognitive point of view within psychology. At the same time, the question of the independence of emotion and motivation from cognition is a legitimate one, and has given rise to a new interdisciplinary field, affective neuroscience, proceeding in parallel with cognitive neuroscience.

Regardless of how the independence issue is resolved, it is clear that cognitive processes can influence emotions and motives. Emotions can be induced by remembering past events, and they can be altered by construing events differently. Certain "counterfactual" emotions, such as disappointment and regret, require that the person construct a mental representation of what might have been. Although some emotional reactions may be innate and reflex-like, others are acquired through conditioning and social learning. As noted earlier, there is evidence that some emotional states, such as anxiety and depression, result from the perception that environmental events are unpredictable or uncontrollable; they may disappear when such beliefs are corrected. Surgical patients' fears can be allayed (and the outcome of treatment improved) if their doctors carefully explain what is going to happen to them and why it is necessary. The ability to use cognitive processes to regulate one's own feelings and desires is an important component of emotional intelligence.

On the other hand, it is clear that emotional and motivational states can have an impact on cognition. In an important sense, the "affective revolution" in psychology was initiated by studies of the effects of mood on memory; these led psychologists to become more interested in the nature of the moods. Five such effects have been well documented: The affective intensity effect (better memory for positive or negative events compared to neutral events), the affective valence effect (better memory for positive than for negative events), mood-congruent memory (better memory for material whose affective valence matches the mood in which it is encoded or retrieved), resource allocation effects (depression impairs performance on effortful, but not automatic, aspects of memory function), and mood-dependent memory (memory is better when there is congruence between the emotional state present at the time of encoding and the state present at the time of retrieval). Although clinical lore holds that emotional trauma can render people amnesic, the overwhelming finding in both the clinical and the experimental literature is that traumatic experiences are remembered all too well.

There is also a growing literature on the emotional effects of other cognitive processes, such as perception and judgment. Signal-detection theory has demonstrated that goals and motives can percolate down to affect the most elementary psychological functions. Common metaphors speak of happy people viewing the world through rose-colored glasses and that things look dark when we are unhappy; in fact, mood and emotion do seem to serve as filters on perception, just as they do on memory. Similarly, emotions have a considerable effect on judgment and decision making. Prospect theory, proposed by Kahneman and Tversky as an alternative to rational choice, holds that decisions are affected by the way in which choices are framed, and emotions and motives form an important element in these frames. Happy people are more likely to take risks than are unhappy people. Even if feelings and desires prove to be largely independent of knowledge and belief, the interest of cognitive psychologists

in our emotional and motivational lives gives eloquent testimony to the breadth of the field as it approaches its second half-century.

Acknowledgment

Preparation of this article was supported by Grant MH-35856 from the National Institute of Mental Health.

See Also the Following Articles

BEHAVIORAL NEUROGENETICS • BRAIN DEVELOPMENT • CLASSICAL CONDITIONING • COGNITIVE AGING • COGNITIVE REHABILITATION • CREATIVITY • INTELLIGENCE • LANGUAGE ACQUISITION • LANGUAGE AND LEXICAL PROCESSING • MEMORY, OVERVIEW • NEUROPSYCHOLOGICAL ASSESSMENT • NEUROPSYCHOLOGICAL ASSESSMENT, PEDIATRIC • PSYCHOACTIVE DRUGS • PSYCHOPHYSIOLOGY

Suggested Reading

Anderson, J. R. (2000). *Cognitive Psychology and Its Implications*, 5th ed. Worth, New York.

Baars, B. J. (1986). *The Cognitive Revolution in Psychology*. Guilford, New York.

Barsalou, L. W. (1992). *Cognitive Psychology: An Overview for Cognitive Scientists*. Erlbaum, Hillsdale, NJ.

Benjafeld, J. G. (1996). *Cognition.* Prentice-Hall, Englewood Cliffs, NJ.

D'Andrade, R. (1995). *The Development of Cognitive Anthropology.* Cambridge Univ. Press, Cambridge, UK.

Gardner, H. (1985). *The Mind's New Science: A History of the Cognitive Revolution.* Basic Books, New York.

Gopnik, A., Meltzoff, A. N., and Kuhl, P. K. (1999). *The Scientist in the Crib.* Morrow, New York.

Gould, J. L., and Gould, C. G. (1994). *The Animal Mind.* Scientific American Library, New York.

Medin, D. L., Ross, B. H., and Markman, A. B. (2001). *Cognitive Psychology*. 3rd ed. Harcourt Brace Jovanovich, Ft. Worth, TX.

Park, D. C., and Schwarz, N. (Eds.). (1999). *Cognitive Aging: A Primer.* Psychology Press, Philadelphia, Pa.

Shettleworth, S. J. (1998). *Cognition, Evolution, and Behavior.* Oxford Univ. Press, New York.

Siegler, R. S. (1996). *Emerging Minds: The Process of Change in Children's Thinking*. Oxford Univ. Press, New York.

Sternberg, R. J. (1999). *The Nature of Cognition*, pp. 173–204. MIT Press, Cambridge, MA.

Wilson, R. A., and Keil, F. C. (1999). *The MIT Encyclopedia of the Cognitive Sciences.* MIT Press, Cambridge, MA.

Zerubavel, E. (1997). *Social Mindscapes: An Invitation to Cognitive Sociology*. Harvard Univ. Press, Cambridge, MA.

Cognitive Rehabilitation

STEPHANIE A. KOLAKOWSKY-HAYNER and JEFFREY S. KREUTZER
Virginia Commonwealth University

GLOSSARY

cognition The process of understanding and functioning within the world, including complex skills such as orientation, initiation, attention, concentration, learning, memory, perception, communication, comprehension, self-monitoring, problem solving, and reasoning.

compensatory strategies Skill- and strength-focused rehabilitation technique including the use of external aids to offset cognitive and psychomotor impairments and improve daily living activities.

generalization The ability to apply previously learned material to a variety of novel contexts.

prosthetic device An external aid used to offset cognitive problems arising from neurological impairment; includes use of microcomputer organizers, filing systems, and lists.

supported employment A functional approach to cognitive rehabilitation incorporating a variety of rehabilitation strategies, characterized by the presence of an employment specialist working directly at the job site.

transfer The ability to apply previously learned material to similar contexts; it is *positive* when learning in one context is applied to another context, and it is *negative* when learning in one context, and it is hinders learning in another context.

Cognitive rehabilitation is a comprehensive, holistic approach to improving and restoring impaired mental, psychomotor, and behavioral functioning. Cognitive rehabilitation is most frequently used after postconcussion syndrome, head injury, and other neurological conditions. Clients are made aware of changes in their cognitive abilities and taught to recognize impairments. Cognitive rehabilitation enhances a client's ability to compensate for lasting impairments.

I. INTRODUCTION

Ongoing advances in medical care and medical management, following an insult to the brain (e.g., brain injury, tumor, and stroke), have increased clients' likelihood of survival. Although more clients are surviving, they are faced with numerous enduring physical, cognitive, and behavioral difficulties. Often, life quality is also diminished due to impaired functional skills. Most survivors are unable to return to work or school, live independently, support themselves financially, or participate in social activities. In order to ameliorate functional problems, and increase the likelihood of return to productive living, many clients participate in a program of cognitive rehabilitation.

Cognitive rehabilitation includes a variety of approaches and techniques, including computer training, memory retraining, social skills training, life skills training, use of prosthetic devices, stimulation therapy, process training, attention and concentration training, stimulus response conditioning, strategy training, domain-specific training, cognitive cycle techniques, enhancement of physical and emotional health and social functioning, and nutrient and drug treatment. However, given the numerous approaches, development of a standard treatment protocol has

been unsuccessful. No consensus exists regarding which approaches or techniques should be used.

There are few research studies regarding cognitive rehabilitation techniques, and these have limited generalizability. For example, studies often used single-subject designs, selectively enrolled clients, and showed therapy benefits in terms of only paper-and-pencil tests. Studies have not been consistent regarding efficacy, do not provide long-term outcome information, and do not adequately describe therapy candidate characteristics. Furthermore, most prior research has failed to recognize and control for the factors influencing cognitive rehabilitation, including a person's physiological state, perceptual skills, emotional status, level of motivation, and social states (e.g., marital status, education level, and socioeconomic status).

II. NEUROBEHAVIORAL PROBLEMS

Persons with brain injury and other neurological conditions often experience lasting neurobehavioral problems. Cognitive deficits may include impairments in orientation, initiation, attention, concentration, discrimination, perception, learning, memory, and reasoning. Behavioral problems may include lack of motivation, inhibition, poor self-monitoring, irritability, apathy, aggressiveness, depression, lack of awareness, restlessness, unusual sexual behavior, and substance abuse. Physical difficulties may include fatigue, motor slowness, and diminished coordination.

Neurobehavioral sequelae are a consequence of both the client's reaction to loss of ability and neurophysiological changes in the brain. Researchers have found that behavioral and emotional problems cause greater impairments in family and emotional functioning than neuropsychological and physical disabilities. Behavioral problems related to depression, anxiety, and social withdrawal have been shown to be predominant.

Although some problems are resolved relatively early postinjury, deficits have been known to persist for a decade or more. The most common long-term difficulties include lack of social contact, personality change, slowness, decreased initiation, poor memory, restlessness, irritability, and bad temper. Persons injured before age 20 tend to have significantly more long-term difficulties. Persistent cognitive, behavioral, and physical problems impede employment and productive living after injury or neurologic insult.

III. COGNITIVE REHABILITATION TECHNIQUES

Cognitive rehabilitation techniques are often geared to three broad categories: restore, substitute, or restructure. Techniques for restoring cognitive functions include cognitive training and exercises directed toward strengthening and the reestablishment of impaired abilities. Substitute techniques commonly include external prosthetic devices and compensatory strategies utilized in place of impairments, which do not attempt to ameliorate impairments. Lastly, restructure techniques reorganize the client's environment in order to compensate for impairments. The following are examples of the most common cognitive rehabilitation techniques and approaches:

- ***Computer training*** utilizes multipurpose software for assessment, treatment planning, treatment, data analysis, progress recording, and reporting. Multimedia applications are also used for education and simulation.
- ***Memory retraining*** often involves teaching a client to use rehearsal techniques to maintain information within memory.
- ***Social skills training*** primarily involves training conversation tracking skills, eye contact, initiation, attention, and appropriate behavioral skills.
- ***Use of prosthetic devices*** involves using external aids rather than retraining the cognitive deficit (e.g., memory watch, calculator, tape recorder, pill organizer, and voice recognition software).
- ***Stimulation therapy*** is the oldest type of cognitive rehabilitation. It involves direct retraining through paper-and-pencil exercises and computer programs that require one or more mental skills.
- ***Process training*** is similar to stimulation therapy but is designed to improve only one specific aspect of cognition.
- ***Attention and concentration training*** is designed to improve one's ability to focus attention, maintain vigilance, resist distraction, and perform mental manipulations quickly and efficiently.
- ***Stimulus–response conditioning*** involves the identification of rewards and punishments and then either providing or withholding these in order to affect some desired change.
- ***Strategy training*** involves the teaching of mental sets that are applicable in other situations (e.g., mnemonics, conversation skills, and problem-solving skills).
- ***Domain-specific training*** provides guidance with the use of simulated life experiences or within a specific

functional domain (e.g., computer-simulated driving).

- ***The cognitive cycle technique*** is a five-step process for retraining complex executive functioning skills. The five steps include having the client identify goals, determine methods of achieving the goals, act to achieve goals based on chosen method, assess progress, and, if necessary, repeat the process until the desired goal is achieved.
- ***Enhancement of physical and emotional health and social functioning*** is an indirect approach that focuses on changes to one's overall lifestyle (e.g., stress reduction, better nutrition, exercising regularly, and getting enough sleep and relaxation).
- ***Nutrient and drug treatment*** is a relatively new approach involving the use of various drugs and nutrients believed to improve brain functioning to treat cognitive deficits (e.g., nootropics, vasodilators, mechanism-based drugs, nutrients, herbs, minerals, antioxidants, and vitamins).

IV. PRACTICAL ISSUES

A. Who Provides Therapy?

To date, there are few education programs specializing in cognitive rehabilitation. Persons qualified to provide cognitive rehabilitation include psychologists, rehabilitation counselors, speech pathologists, occupational therapists, neuropsychologists, recreational therapists, rehabilitation nurses, psychiatrists, physical therapists, and other rehabilitation providers. Each specialty provides its own treatment approaches and diverse resources. Deciding which practitioner is best suited to help each client is a matter of personal judgment.

B. Generalization Issues

Generalization can be characterized into three distinct levels. Level 1 involves basic learning over trials, in which performance improves as training continues. The client should demonstrate the ability to transfer learning from one task to another similar task. The second level involves the transfer of learning to performance on psychometric tests measuring similar functional areas. For example, a client is trained to use maintenance rehearsal techniques (simple repetition of information) for memory and then incorporates similar behaviors while completing a verbal memory test. Lastly, level 3 includes transferring learning to general activities of daily living. For example, the client taught to use maintenance rehearsal techniques calls directory assistance for the phone number of a pizza place. He proceeds to repeat the number until he finishes dialing.

Cognitive rehabilitation has been shown to provide adequate training such that clients are able to transfer previously learned material to similar contexts. However, the most controversial issue regarding the efficacy of cognitive rehabilitation is level 3 generalization, the ability to apply previously learned material to a variety of novel contexts. For example, outcome studies often use paper-and-pencil tests to measure the benefits of cognitive rehabilitation, and it has been difficult to determine whether such test results translate into meaningful functional abilities. Furthermore, difficulty arises when trying to measure real-life contexts. Vocational rehabilitation settings provide potential real-world assessment arenas; however, obtaining a truly randomized and controlled sample would be nearly impossible.

C. Assessing Rehabilitation Potential

Who should be a candidate for cognitive rehabilitation? Clinicians often confront a series of questions. For example, is the client out of a coma? Is he or she medically stable and out of crisis? What was his or her preinjury level of functioning? What is his or her current level of functioning? Are his or her perceptual skills intact? Will he or she be able to attend to a therapist for any length of time and, if so, for how long? Is the client combative, apathetic, depressed, or aggressive? Are there any secondary gain issues (e.g., settlement, worker's compensation, and retirement secondary to disability)? What social supports does the client have? Is the client likely to improve with rehabilitation?

Most of these questions are easily answered via medical record review, neuropsychological testing, and client/informant interview. However, there is no clear-cut, standardized method for determining who will benefit from cognitive rehabilitation and who will not. When making such a decision, a clinician must consider the following: First, when improvement is demonstrated, improvement is likely with intervention. For example, if a client performs progressively better on the second and third trials of a task but is still exhibiting impairment relative to preinjury levels, this

area will likely benefit from training. Second, when improvement is not demonstrated, intervention will most likely be ineffective and frustrating. For example, when the client has difficulty learning across multiple trials, additional trials will most likely not improve functioning but will probably diminish enthusiasm and increase aggravation. Alternate training approaches should be taken. Third, start rehabilitating stronger areas, and increase difficulty of tasks with progress. A clinician should determine the client's strengths and weaknesses. Rehabilitation of a semi-intact skill will prove easier than rehabilitation of a completely impaired skill. Fourth, concentrate on steep learning curves, not high performance levels. A steep learning curve exists when a client gradually learns a little more information at each training session or across each trial (Fig. 1). Such a curve suggests the client has a high learning potential and will likely benefit from intervention. Conversely, high performance levels are evident when the client makes a proverbial leap from learning almost nothing to knowing a great deal more and performs relatively consistently over time (Fig. 2). No more and no less learning occurs between trials.

D. Center-Based vs Community-Based Programs

Cognitive rehabilitation has been provided in a variety of settings, including acute hospitals, rehabilitation hospitals, skilled nursing facilities, outpatient rehabilitation clinics, and at clients' homes. Within the past 10 years, the scope of cognitive rehabilitation has grown tremendously. In the past, cognitive rehabilitation was provided primarily through center-based computer training of rote memory tasks. Recently, there has been a shift to real-world community-based programming with a more holistic approach. Currently, the ideal is to provide cognitive rehabilitation sessions along a continuum of care from the acute setting to outpatient and homebound settings.

Both center-based and community-based programs have advantages and disadvantages. Center-based cognitive rehabilitation programs often provide comprehensive, multidisciplinary assessment, treatment, and follow-up. However, they often provide canned programming and computer-based rehabilitation tasks and the programs are often artificial (e.g., clients performing office tasks in a hospital room). Community-based cognitive rehabilitation programs often occur in real-world settings, alleviating the problems with generalizability. However, real-world settings are highly variable and often unpredictable, challenging the effectiveness of therapy.

V. FUNCTIONAL APPROACHES

The main goal of cognitive rehabilitation is to enhance productive living by improving impaired daily living skills. Although the approaches and techniques previously discussed affect overall cognitive functioning,

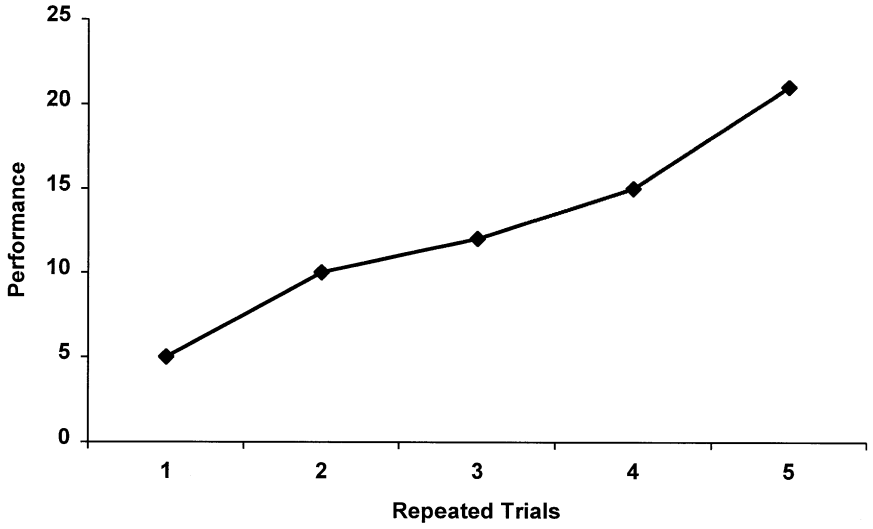


Figure 1 Example of a steep learning curve across learning trials.

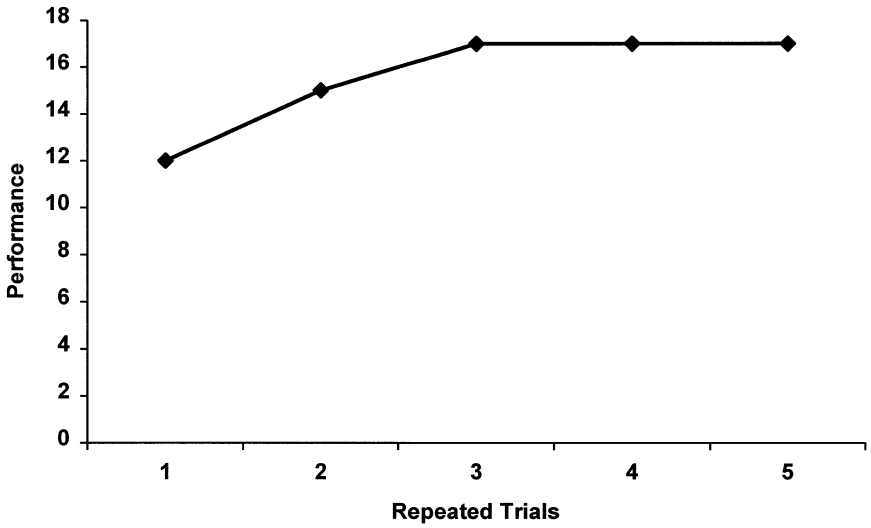


Figure 2 Example of high performance across learning trials.

the generalizability of such training is uncertain. Functional strategies have begun to be developed to assuage the generalizability issue. Within the past 10 years, clinicians and researchers have attempted to design new methods of training to counteract the most problematic issues faced by persons with neurological disorders—living and working independently.

A. Compensatory Strategy Training

Memory and psychomotor problems are extensive following brain injury and neurological impairment. Compensatory strategy training was developed to help persons with brain-related disabilities overcome impairments in daily living skills. People with brain dysfunction are trained to utilize external aids to compensate for impairment rather than working to improve the impaired skill. Compensatory strategy training is often used when clients' cognitive limitations and environmental barriers impede alternative types of treatment.

Compensatory strategy training utilizes and focuses on existing strengths to compensate for what the client cannot do. Clinicians often maximize client involvement in the development and use of strategies because less involvement results in hesitancy and reduces chances of success. For the client, understanding the utility of the strategy is helpful but not necessary. Selection and use of strategies will change over time and strategies are developed along a continuum of intrusiveness and complexity (e.g., the number of external cues is reduced over time following internalization). Clinicians are often cautious to ensure that the client is not taught too many strategies, resulting in inefficiency and confusion.

Compensatory strategy training typically provides rapid improvement in daily living skills, providing clients with increased self-esteem and feelings of accomplishment. The training is cost-effective given that only a limited number of sessions are needed for mastery of skills. Additionally, the minimal number of sessions makes it easier for clients with transportation problems to participate. Since training involves real-world activities, there are minimal assumptions about lack of generalization. Furthermore, client participation in the treatment process is increased and resistance is reduced because benefits are apparent.

As with any type of cognitive rehabilitation technique, clinicians must consider many factors prior to initiating compensatory strategy training. Specifically, one must consider motivation issues; cognitive and neurobehavioral impairments, especially memory dysfunction and adynamia (the inability to complete tasks), often due to lack of energy or interest; the number of environments in which behavior is expected to occur and similarity among environments; relevance to real-world requirements; and need for generalization from training environment. Failure to implement strategies may be due to the fact that the strategy was not matched to client's needs. The client

may not be convinced of the value or may hope that the problem will resolve quickly without intervention. Clients may go from a supportive training session to an unsupportive environment. Additionally, failure may be due to the client's unwillingness to practice the compensatory technique.

Many compensatory strategies are relatively simple and are often used by persons without disabilities. Examples of common compensatory strategies include the use of memory books, "wake up" services, pocket memo recorders, word processors, spell checkers, calculators, phone message pads, schedule books, financial management software, pill minder boxes with daily compartments, "to do" lists, alarm watches, shopping lists, and autodial phones.

B. Supported Employment

Supported employment is an individualized rehabilitation approach developed as a means of addressing the inadequacies of traditional employment methods. Originally developed to assist persons with psychiatric and intellectual disabilities, supported employment was adapted in the early 1980s for use with persons with brain dysfunction. Supported employment has become the standard of practice for many vocational rehabilitation professionals. The approach to vocational rehabilitation is well suited to illustrate the benefits of real-world cognitive rehabilitation efforts and the use of compensatory strategies. Despite progress in rehabilitation program development, return to work remains a formidable challenge, especially for persons with severe cognitive impairments. In many societies, employment is the basis of an adult's personal identity, with higher paying and more prestigious occupations affording higher levels of social recognition. Consequently, many unemployed persons with brain dysfunction view their recovery as incomplete or insignificant.

There are many variations of supported employment. Services are provided by job coaches or employment specialists. In contrast to traditional programs, which focus on job acquisition, emphasis is placed on helping people maintain jobs. Characteristics of the successful model include the following:

- ***Community placement and integration:*** Clients are helped to find and keep jobs in their home community by working in local businesses. In an effort to avoid segregation and promote widespread community integration, one to four clients work alongside nondisabled coworkers.
- ***Competitive hiring, wages, and benefits:*** Clients are hired through the same competitive process as other employees, such as by completing the employer's standard application form and attending an interview. Often, employment specialists help clients complete job applications, provide transportation, or prepare the client for an interview using role-playing techniques. Furthermore, workers with disabilities receive the same pay and benefits as coworkers.
- ***Emphasis on inclusion:*** Rather than excluding clients on the basis of mental health or behavioral or medical problems, an interdisciplinary team follows the client to meet needs for rehabilitation, medical care, and mental health services before and after placement.
- ***Holistic assessment of the client, home environment, and workplace:*** Assessment often includes a series of questionnaires completed by the client and family members; center-based and community-based interviews with clients, family, employers, and coworkers; and a comprehensive neuropsychological evaluation. Additionally, a job analysis is conducted in which the job coach visits each potential employment site and evaluates the characteristics of the workplace. A special effort is made to understand the exact nature of job responsibilities, the degree of interpersonal interaction with customers and coworkers, safety risks, and level of available supervision.
- ***Emphasis on choice and job matching:*** People, regardless of disability, are less likely to succeed at a job that is demeaning, meaningless, or boring. Although the client has the final choice regarding which job to apply for, the job coaches strive to match clients' goals, interests, and abilities with appropriate employment settings based on job analysis.
- ***Emphasis on intervention after placement:*** On average, 32–36 hr of intervention time occur prior to placement, whereas 240–260 hr occur within the first 6 months following placement. Ongoing behavioral and cognitive rehabilitation techniques are utilized within the workplace. Problems with memory are common after brain injury and severely limit generalization from preplacement training. Strategies such as self-monitoring, rehearsal, and reinforcement are used to teach and help generalize skills.
- ***Coworker and employer education:*** To alleviate negative attitudes, prior to placement the job coach provides the employer and coworkers with educational materials regarding the effects of brain injury and information concerning the client's positive

attributes. After placement, the job coach acts as a liaison between clients, coworkers, and employers, maintaining communication between all parties.

- ***Long-term follow-up:*** After placement, the client may face many difficult situations that all people face during employment. For example, there may be changes in personnel or job responsibilities. External life stresses may interfere with productivity. Job support and intervention fluctuate as new problems are encountered and resolved. Feedback from the client, supervisor, and rehabilitation team is considered in the determination of follow-up intensity. Early on, the employment specialist may spend the entire workday with the client to identify problems, develop methods of resolution, provide direct feedback, and reduce the client's stresses. With job mastery, the employment specialist "fades," reducing direct intervention time. After employment stabilization (e.g., 6 months following placement), follow-up may require 2 or 3 hr per month.
- ***Job completion guarantee:*** Prior to employment, job coaches assert that clients will meet work responsibilities. A system of checks and balances is employed by the coach, including having the job coach work alongside the client to make certain job goals are met.
- ***Intensive ongoing analysis of program outcome:*** To ensure efficacious supported employment programming, ongoing data collection and analysis are essential. Program data include typical information such as injury severity, age, time elapsed since injury, employment status, and wages. In addition, the following information is collected: hours and types of intervention, employers' and clients' performance and satisfaction ratings, reasons for job separation, factors promoting successful work outcome, and level of employment stability.

C. Supported Living

The success of supported employment inspired the development of supported living programs. Instead of relying on job coaches, clients are afforded the help of a life skills trainer or specialist. The life skills trainer uses cognitive rehabilitation techniques to teach living skills, which may include cooking, cleaning, financial management, personal hygiene, use of public transportation, time management, first aid, driving, shopping, laundry skills, medication management, and appliance safety. Supported living programs are often centered in apartment buildings or homes that can accommodate 4–10 clients and several life skills trainers. Successful clients are gradually transitioned to greater independence, and many ultimately live on their own.

See Also the Following Articles

BRAIN DAMAGE, RECOVERY FROM • CANCER PATIENTS, COGNITIVE FUNCTION • COGNITIVE AGING • COGNITIVE PSYCHOLOGY, OVERVIEW • HIV INFECTION, NEUROCOGNITIVE COMPLICATIONS OF • MODELING BRAIN INJURY/TRAUMA • STROKE

Suggested Reading

Harrell, M., Parente, F. J., Lisicia, K. A., and Bellingrath, E. G. (1992). *Cognitive Rehabilitation of Memory: A Practical Guide*. Aspen, Gaithersburg, MD.

Kreutzer, J., and Wehman, P. (Eds.) (1991). *Cognitive Rehabilitation for Persons with Traumatic Brain Injury: A Functional Approach*. Brookes, Baltimore.

Kreutzer, J., and Wehman, P. (1996). *Cognitive Rehabilitation for Persons with Traumatic Brain Injury*. Imaginart, Bisbee, AZ.

Parente, R., and Anderson-Parente, J. (1991). *Retraining Memory: Techniques and Applications*. CSY, Houston.

Parente, R., and Herrmann, D. (1996). *Retraining Cognition: Techniques and Applications*. Aspen, Gaithersburg, MD.

Randall, M. C. (1999). *Rehabilitation for Traumatic Brain Injury*. DIANE, Collingdale, PA.

Rosenthal, M., Griffith, E., Kreutzer, J., and Pentland, B. (Eds.) (1999). *Rehabilitation of the Adult and Child with Traumatic Brain Injury*, 3rd ed. Davis, Philadelphia.

Sohlberg, M. M., and Mateer, C. A. (Eds.) (1989). *Introduction to Cognitive Rehabilitation: Theory and Practice*. Guilford, New York.

Toglia, J., and Golisz, K. M. (1999). *Cognitive Rehabilitation: Group Games and Activities*. Academic Press, San Diego.

Wehman, P., and Kreutzer, J. (Eds.) (1990). *Vocational Rehabilitation for Persons with Traumatic Brain Injury*. Aspen, Rockville, MD.

Color Processing and Color Processing Disorders

PETER GOURAS
Columbia University College of Physicians and Surgeons

GLOSSARY

achromatic contrast Gradients of energy in an image resulting in black–white vision.

achromats Subjects who have either only short-wave-sensitive or no cones; they have no color vision.

chromatic contrast Gradients of wavelength in an image responsible for color vision.

color opponency Colors that cancel each other when mixed, such as red and green or blue and yellow.

cone opponency Signals from two spectrally different cones that have opposite effects on the same neuron (i.e., one cone signal excites and the other inhibits the neuron).

deuteranopes Subjects who lack middle-wave-sensitive cones; their color vision is also divariant (about 1% of males).

opsins Proteins in photoreceptors that absorb light and trigger a neural response.

protanopes Subjects who lack longwave sensitive cones; their color vision depends on two variables and is divariant (about 1% of males).

successive color contrast Enhancement of a color appearing just after another color disappears (i.e., red is redder when it appears just after green goes off).

simultaneous color contrast Enhancement of a color when next to another color (i.e., reds are redder when surrounded by green).

tritanopes Subjects who lack short-wave-sensitive cones; they are also divariant but extremely rare.

trivariance The fact that normal human color depends on only three variable—the three different types of cones.

Color vision is a neural process occurring in our brain that depends on a comparison of responses of at least two, but normally three, spectrally different cone photoreceptors. The brain uses these cone systems to detect achromatic and chromatic contrasts. Achromatic contrasts depend on light energy gradients across an image; chromatic contrasts depend on wavelength gradients of light across an image. Achromatic contrast is detected by the two longer but not by the short-wave-sensitive cones; in the fovea, achromatic contrast can resolve the dimensions of neighboring cones (about 1 μm on the retina). Chromatic contrasts are detected using all three types of cones and is done by comparing their responses to the same object. From an evolutionary perspective, the first comparison was between the short- and the longer wavelength-sensitive cones creating blue/yellow color vision. This still occurs in most mammals and in about 2% of human males. In this case, if an object affects the short-wave more than long-wave cones, it appears blue. The converse appears yellow. If the object affects both cone

systems equally, it appears white, gray, or black. In the Land color vision model this comparison between cone signals occurs after each cone system's response is normalized over visual space. The second comparison evolved in primates when the longer wave cones were split into long- and middle-wave cones; this created red/green color vision. If an object affects the long-wave more than the middle wave-sensitive cones, it appears red. The converse appears green. If an object affects both of these cones equally, it appears yellow, white, or bluish depending on the former comparison. Color vision compares responses of groups of cones rather than single neighboring cones; it has a lower spatial resolution than achromatic vision. The neural processing responsible for color perception occurs in visual cortex. The retina and the lateral geniculate nucleus provide the signals from each cone mechanism in a form that allows the comparisons necessary for color vision. Our brain combines cues from both chromatic and achromatic contrast to perceive a particular color. There are several different mechanisms involved. One depends on the blue/yellow and red/green comparison signals and is most related to the hue of a color, reflected by the words red, green, etc. The second is related to the amount of white, gray, or black that is mixed with the previous signal. This confers a quality called saturation of a color; it distinguishes pinks from reds. The third depends on the achromatic system to establish the lightness or darkness of a color. Together these separate operations provide us with about 1 million different colors. All these operations, which involve both achromatic and chromatic contrast detection, provide cues to the form of an object. Although we can mentally separate the form of an object from its color, it is not known where in cerebral cortex this separation of form from color occurs.

I. COLOR VISION

Light has several independent properties, including its energy and its wavelength or frequency of vibration. Color vision takes advantage of both of these properties, allowing us to detect objects and navigate in the external world. Wavelength and energy contrasts vary independently in a visual image and therefore require different neural circuits for their analysis. The secret of color vision is to understand how the brain distinguishes wavelength from energy contrasts. The addition of wavelength contrasts to energy contrasts greatly increases the potential cues to the visual universe. In addition, energy contrasts depend on both the surface properties of an object and its position relative to its illuminant, sunlight for our past evolutionary history. Positional variations can lead to highlights and shading that can obscure the identity or existence of objects. The wavelength reflectance of an object is relatively independent of such effects and therefore a more reliable identifier. For this reason, wavelength contrast and derivative, color vision, have been important to survival. In man it has added an aesthetic dimension that far exceeds the value for survival in modern society.

Figure 1 shows a colorful Renaissance painting, which has been decomposed into its wavelength (chromatic) and achromatic (black-and-white) contrasts. When these two operations are combined in our brain, the picture is seen in color. Energy contrast leads to the perception of achromatic forms of different brightness (white, gray, or black). Wavelength contrast leads to the perception of forms that have a new dimension (color). Energy contrast provides finer detail than wavelength contrast, but the latter adds a new dimension to vision. The combination of both forms of contrast is what we normally experience as color vision.

II. NEURAL BASIS OF COLOR VISION

A. The Photoreceptors

Vision starts in the photoreceptor cells of the retina and their outermost organelle that contains specialized protein molecules, called opsins (from the Greek word "to see"), which absorb and amplify light energy. In the human retina, there are four types of photoreceptors, defined primarily by the opsin they contain. Usually one type of cone contains only one type of opsin. Each type of opsin absorbs a particular waveband of light in the visible spectrum. The wavelength selectivity, defined by an absorption spectrum, is related to the opsin's amino acid sequence interacting with a specific isomer of vitamin A.

One set of photoreceptors, called rods because of the shape of their long outer segments, function only at dim light levels. Their extraordinarily high sensitivity to light drives them into saturation at daylight levels. Rods are very sensitive and numerous in our retina but only work well in moonlight and play little to no role in

Figure 1 A segment of a painting by Jan van Eyck (1434) in which the luminance information has been removed (left) rendering an image based on chromatic contrasts alone. (Right) The chromatic contrast has been removed, rendering an achromatic black-and-white image of the same scene; the latter has more fine detail. In the middle and fusion of these two different forms of contrast produces color vision. (See color insert in Volume 1).

color vision. Rod vision is without color and has become vestigial in modern society.

Color vision is mediated by the other three photoreceptors called cones, which require broad daylight to work well. Figure 2 shows the absorption spectra of the three cones of human vision. Cones are much more important to human vision than rods because if they are lost or fail to function, one is legally blind. The absence of rod function, on the other hand, is only a minor inconvenience. Cones' tolerance of high levels of illumination depends in great part on their ability to adapt to light. As light levels increase, cones reduce their sensitivity and speed up their responses, making them almost impossible to saturate. For color vision, the responses of the three types of cones are compared. Figure 2 shows the spectral colors that result from these comparisons. The colors we see in the spectrum are not near the peak absorption of the cone opsins because they depend on sophisticated neural comparisons.

B. Long-Wavelength Cone System

Long-wavelength-sensitive cones dominate the primate retina, comprising 90% of all the cones. The short-wavelength-sensitive or S cones comprise only 10% of the cones and are absent from the central fovea, the area of highest visual resolution. The long-wave cones are composed of two types of cones, with similar opsins. Both absorb best in the long wave or yellow half of the spectrum. One type, M or "green" cones, absorbs slightly better in the middle or green part of the spectrum, and the other, L or "red" cones, absorbs slightly better in the long or red part of the spectrum (Fig. 2). These L and M cones seem to be

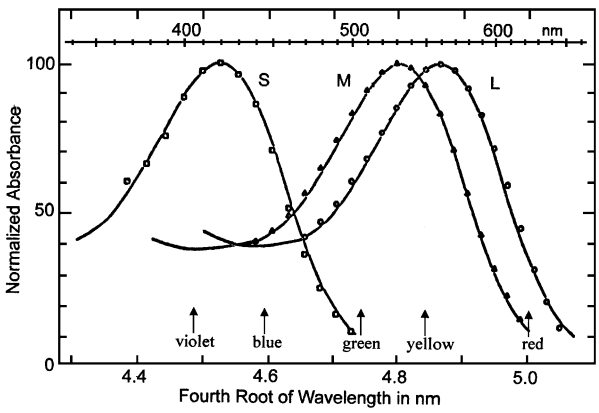


Figure 2 Normalized absorption spectra of the three cone mechanisms of human vision plotted against the fourth root of wavelength, which tends to make these curves independent of their position on the abscissa. An equivalent wavelength abscissa is shown above. Short-wave- (S), middle (M), and long-wavelength (L)-sensitive cones. The colors perceived at different points in the spectrum are shown below.

organized in an identical way, although there may be slightly more L than M cones. The central fovea contains only L and M and no S cones. L and M cones in the fovea are slender, optimizing their ability to sample small areas of visual space.

Each L and M cone synapses with at least two different bipolar cells. One is an on bipolar that is excited (depolarized) whenever its cone or cones absorb light; the other is an off bipolar, which is excited (depolarized) whenever light absorption by its cone or cones decreases (Fig. 3). This is a push–pull system that provides excitatory signals for increments (lightness) in one channel and excitatory signals for decrements (darkness) of light energy in a parallel channel. These signals are used in visual cortex to sense both energy (achromatic) and wavelength (chromatic) contrasts.

In the fovea, each L and M cone on and off bipolar synapses with a single cone (Fig. 3, left). Each single bipolar cell synapses with a single on or off ganglion cell. This is the so-called "midget cell" system discovered by Stephen Polyak. The signals from single L or M cones are transmitted by relay cells in the parvo cellular layers of the lateral geniculate nucleus of the thalamus to striate cortex mediating the high visual resolution of the fovea.

Away from the fovea, the midget arrangement ceases and several L and/or M cones synapse with single on or off bipolar cells. The breakdown of the midget system creates a problem for transmitting information about both energy and wavelength contrast in the same neural channel. For energy contrast, it is best to minimize the area within which the cones are selected. Therefore, selecting both L and M cones would be preferable (Fig. 3, right). For wavelength contrast it is best to select only L or only M cones in any one area. This tends to expand the area from which cones were selected and therefore decrease spatial resolution. It is not clear how this situation is handled.

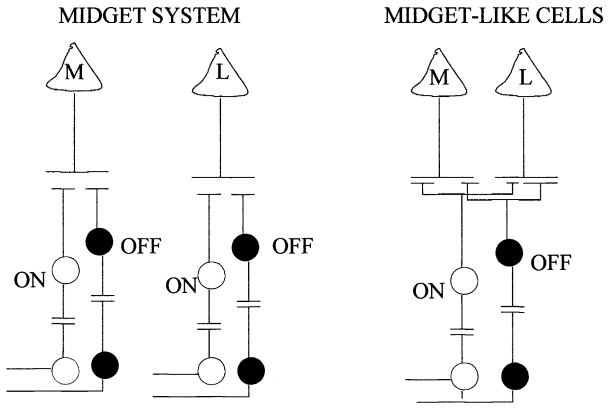


Figure 3 The retinal circuitry of the parvocellular L and M cone system. The midget system on the left is characteristic of the fovea, where each L and M cone has a private on and off bipolar and ganglion cell. Away from the fovea, this private line breaks down. Some cells, midget-like, are thought to preserve the selectivity for only one type of cone, a prerequisite for chromatic contrast.

1. The Parvo System: L and M Cone Antagonism

The selectivity for either an L or M cone input (Fig. 3) exposes antagonistic cone interaction in these retinal ganglion cells that depends on the wavelength of stimulation and is mediated by amacrine and horizontal cell interneurons. Ganglion cells transmitting signals of L cones are antagonized by stimuli that affect M cones and vice versa. This antagonism enhances both achromatic and chromatic contrast.

It enhances achromatic spatial contrast by reducing a ganglion cell's response to large but not small stimuli covering the center of the receptive field of the ganglion cell. It enhances chromatic contrast by reducing a ganglion cell's spectral response to green light if it is transmitting signals of L cones or reducing the cell's response to red light if it is transmitting the signals of M cones. It also enhances successive chromatic contrast, which occurs when a stimulus moves over the retina. The removal of green (antagonizing) light as red light enters the receptive field of an on ganglion cell transmitting the signals of L cones will enhance its response to the following red light. The removal of red (antagonizing) light as green light enters the receptive field of an on ganglion cell transmitting the signals of M cones will enhance its response to the green light. Off cells will also exhibit successive color contrast in the opposite way.

Retinal ganglion cells showing this behavior, called "cone or color opponency," are a system of small cells relatively concentrated at the fovea, including the "midget system." These cells transmit their signals to the parvocellular layers of the lateral geniculate nucleus of the thalamus. This system seems to be involved in both achromatic and chromatic contrast and is responsible for the high spatial resolution of the fovea. These cells receive no input from S cones.

2. The Magno System: L and M Cone Synergism

There is a parallel system of on- and off-bipolars and ganglion cells transmitting signals from L and M cones to striate cortex via relay cells in the magnocellular

layers of the lateral geniculate nucleus. These are larger cells, which have faster conduction velocities and respond phasically to maintained stimuli. They mix synergistic signals of L and M cones. They are only involved in achromatic contrast, showing no color opponency (Fig. 4). This phasic magnocellular system is relayed to layer 4C alpha of striate cortex. The system is not as foveally oriented as the parvosystem. It is not involved in high spatial resolution and color vision; it receives no input from S cones. It seems to play a role in the detection of movement and body orientation and perhaps brightness perception.

C. Short-Wavelength Cone System

The S cones are only involved in chromatic contrast, which has a lower spatial resolution than achromatic contrast. Chromatic aberration makes the short wavelength image out of focus when the image that the long-wave cone system sees is in focus. Therefore, the S cones have been excluded from high-resolution achromatic vision and the central fovea. S cones transmit their signals to the brain by a unique system of retinal ganglion cells. S cones synapse on S cone-specific on bipolars, which excite a bistratified ganglion cell that is also excited by L and M cone off bipolars (Fig. 5). Such a ganglion cell cannot mediate energy contrasts, being excited by both increments and decrements of light energy. It is designed for successive chromatic contrast, being excited when short wavelengths enter and long wavelengths leave its receptive field. The S cone channel is transmitted by relay cells in the parvo and/

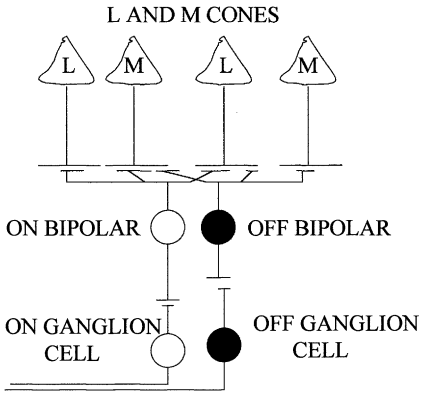


Figure 4 The retinal circuitry of the magnocellular L and M cone system. Here, the L and M cone act synergistically on both on and off channels.

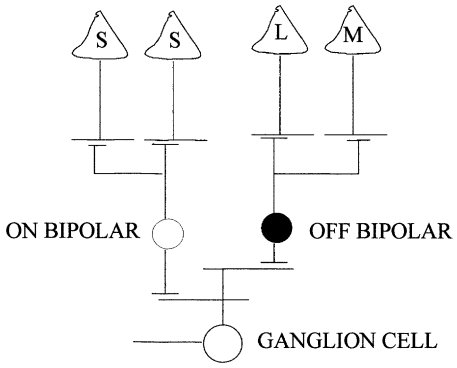


Figure 5 The retinal circuitry of the S cone channel. There are two inputs to a bistratified retinal ganglion cell. One is from on bipolars from S cones and the other is from off bipolars from L and M cones. Both bipolars excite the ganglion cell. The S cones excite their bipolars when there is an increment and the L and M cones excite their bipolar when there is a decrement of light absorption in the corresponding cone. The existence of an antagonistic amacrine cell that interacts off with on bipolars can explain transient tritanopia, a phenomenon of S cone vision.

or the intercalated layers of the lateral geniculate nucleus of the thalamus to striate cortex, layer 4C beta, where it is used to distinguish chromatic contrasts.

There is agreement that this S cone channel, consisting of an on–off ganglion cell, is a major route, possibly the only route for information from S cones to reach striate cortex and visual perception. This channel exhibits relatively little cone opponent behavior. Both white and blue lights excite this cell. Long wavelengths excite the cell when they go off, producing strong responses to successive color contrast. It responds best to white or blue after yellow. This cell informs the brain that the S cone system is or is not absorbing significant light in a particular area of visual space.

Whether there is also an S cone off channel is unclear. Few investigators have detected it. The strongest evidence for its existence comes from measurements with stimuli changing along the tritanopic axis of color space. It is possible that some of the more numerous tonic L and M cone opponent cells could also respond uniquely to such a stimulus. Because S cones are not involved in achromatic contrast, an off channel may not be necessary. Evidence suggesting the absence of an S cone off channel is the subjective phenomenon of transient titanopia, which involves a brief weakening of the appearance of blue whenever a long-wavelength field is turned off. There is no corresponding weakening of yellow when a

short-wavelength field is turned off. S cones appear to lack the ability to antagonize the other cone channels at the retinal level, presumably because their image is so out of focus that it would interfere with achromatic vision.

III. ACHROMATIC VERSUS CHROMATIC CONTRAST

Achromatic contrast detects spatial differences, which depend on the distribution of light energy in the retinal image. For daylight vision, it is mediated by the two longer wavelength-sensitive cones. It is well developed in the fovea, mediating our highest spatial resolution, and represented in relatively large areas of visual cortex. The foveal area of striate cortex is about 36 times larger than that of striate cortex serving peripheral vision. In the fovea, the responses of neighboring cones are compared for achromatic contrast (Fig. 6). Extrafoveally, spatial resolution is reduced because ganglion cells collect synergistic signals from more than one cone. Achromatic contrasts establish local lightness and darkness, which input orientation-selective neurons in striate cortex and undoubtedly contribute to form perception. Lightness or darkness is determined entirely by simultaneous contrast. Absolute values of light energy are discarded by antagonistic interactions between neurons representing different areas of visual space. An object is light or dark depending entirely on its background.

Chromatic (wavelength) contrast is established by eliminating the effects of energy contrast. It compares the responses of one set of cones to those of another in the same area of space (Fig. 6) and then compares this with comparisons obtained in neighboring areas of space (Fig. 7). Chromatic contrast depends on the difference between cone responses in a unit area of chromatic visual space and not on the absolute responses of the cones. Chromatic contrast does not resolve the detail offered by a single cone because a cone cannot provide an unambiguous clue to wavelength contrast as it can to energy contrast.

A cone's response depends on the energy absorbed, regardless of wavelength. Wavelength determines the probability with which a quantum is absorbed. A response to any wavelength can be reproduced by any other wavelength by varying energy. A mosaic of at least two different types of cones must be sampled to distinguish wavelength contrast. Wavelength contrast needs a double comparison. First, the responses of two different sets of cones in one area of visual space must be compared, and then this comparison must be compared with other areas of visual space to establish spatial contrast based on gradients of wavelength. If one compared only two neighboring cones, a gradient of energy contrast would create an ambiguous signal. For this reason, a unit area of chromatic space is larger than that of achromatic space. The lower spatial resolution of chromatic contrast has been demonstrated psychophysically. The spatial resolution of chromatic contrast for all three cone mechanisms appears to be identical. Figure 8 shows the difference between the spatial resolution of chromatic and achromatic

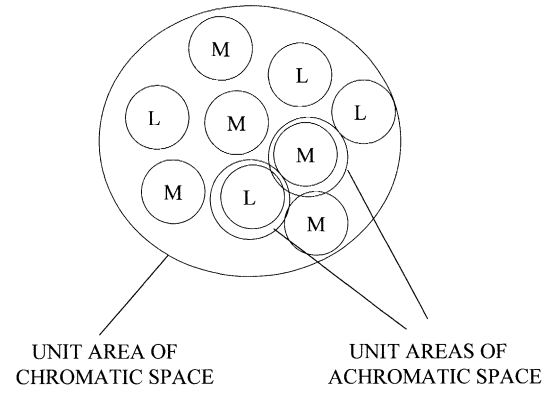


Figure 6 This figure illustrates that the unit areas of chromatic and achromatic space differ because of the nature of the neural comparison. For chromatic contrast, the responses of a group of L cones must be compared with the responses of a neighboring group of M cones in the same area of visual space. For achromatic contrast a single cone can be compared with a neighboring L or M cone.

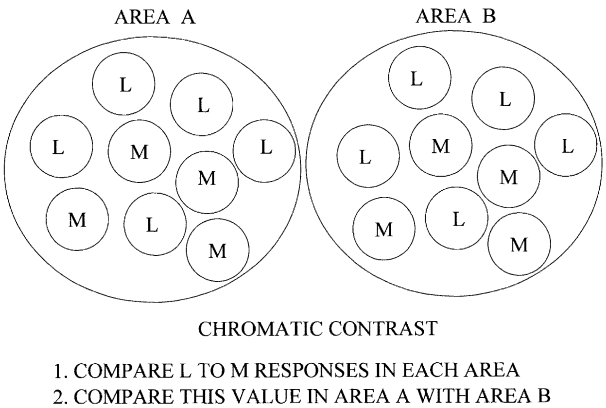


Figure 7 In order to establish chromatic contrast, a unit area of chromatic space must be compared with a neighboring area of chromatic space. Two neighboring L and M cones cannot be compared because a border of any wavelength could create ambiguity.

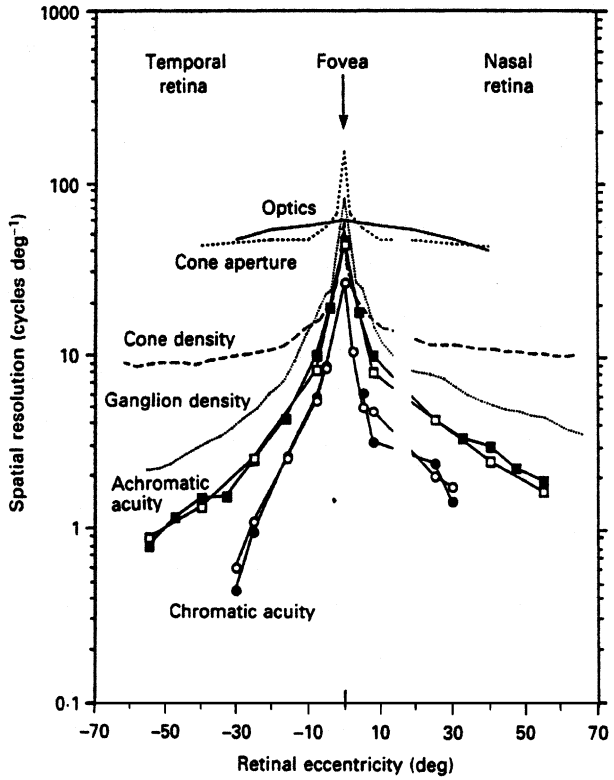


Figure 8 The degree of spatial resolution of the achromatic and chromatic systems of human vision and their distribution across the retina. The symbols represent the data of two normal subjects. They are shown in relationship to the optical properties of the eye and the filtering characteristics of the cones and ganglion cells of the retina (reprinted with permission of Cambridge University Press).

contrast across the human retina. Everywhere the former is lower than the latter, and this increases with eccentricity. Chromatic vision is relatively more developed centrally than achromatic vision.

IV. PSYCHOLOGY OF COLOR VISION

A. Trivariance

In the 1850s, Maxwell in England demonstrated that all human color perception depends on three variables, which he assumed reflected three different types of cones. In Germany, Helmholtz was arriving at a similar conclusion, led by the intuitive insights of Grassmann. The ability to define color by measured amounts of three physically defined variables led to colorimetry. Any color can now be defined by three international standards. The spectral properties of the three-cone mechanisms of human color vision were first derived psychophysically by Stiles in England in the 1950s. Subsequently, microspectrophotometry of single human cones confirmed these measurements. Defining the absorption spectra of the cones is the starting point for understanding human color vision (Fig. 2).

B. Hering Theory of Color Vision

Ewald Hering, a German physiologist, proposed a unique theory of color vision that went beyond trivariance in its insights into how the nervous system compares cone signals to perceive color. He proposed the existence of antagonism between certain pairs of colors, such as blue and yellow and red and green. The brain could not perceive a blue–yellow or a red–green color, whereas it could perceive red–yellow, green–yellow, red–blue, or green–blue colors. Blue and yellow cancel each other so that one cannot perceive bluish-yellow. Similarly, red and green cancel each other so that one cannot perceive reddish-green. This cancellation of one color impression by another was thought to represent neural antagonism. This was a major insight into how the brain compares the three cones in two pairs in order to perceive color.

Hering's theory was based on subjective experience. The discovery of antagonism between cone responses in single neurons of fish retina led to a renaissance of his theory. Linking the physiology of retinal neurons with subjective colors has proven to be difficult, however. The major reason for the difficulty is that color does not appear to be determined at the retinal or geniculate level but depends on additional operations that occur in visual cortex. Antagonism between cones, which occurs in the retina, is not equivalent to antagonism between colors, which occurs in the cortex.

1. Blue–Yellow Opponency

A scheme that combines physiology and Hering's color opponency is shown in Fig. 9. The operations occurring at the retinal and geniculate level are separated from those thought to be occurring in visual cortex. For blue–yellow color vision, the brain combines excitatory signals from S cone on bipolars with excitatory signals from long-wave cone off bipolars at the retinal level. This on–off–S cone channel excites a cell in the cortex that signals blue; this cell also receives antagonistic (inhibitory) signal from L and M

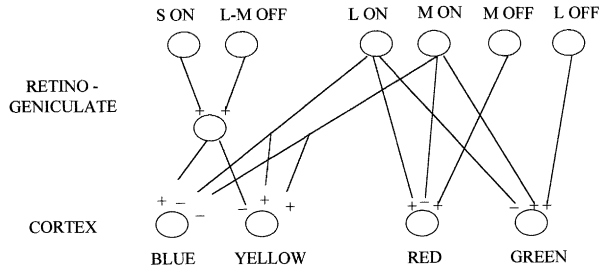


Figure 9 Circuitry that logically organizes the retinogeniculate inputs from the parvocellular system to establish cells responsive to opponent colors in local areas of visual space. (Right) The circuitry of blue–yellow opponent colors in which two different cortical cells receive inputs from the retinal S cone channel, on the one hand, and the long-wavelength tonic system, on the otherhand. The latter excites and the former inhibits the cell detecting "yellow." The converse arrangement affects the cell detecting "blue." If neither cell is excited, achromatic vision determines the color as white, gray, or black. (Left) The circuitry of red–green opponent colors. L cone on and M cone off signals excite and L cone off and M cone on signals inhibit a cell detecting "red"; the converse arrangement detects "green."

cones. This on–off–S cone channel also antagonizes (inhibits) a cell in cortex that signals yellow; this same cell receives an excitatory signal from the L and M cone on-system. These two cells form a blue–yellow opponent system.

If the S cone on-system and the L–M cone on system are excited, the blue and yellow cells are both silent (inhibited) and the system defaults to black-and-white (achromatic) vision. If the S cone on system is excited and the L–M cone on system is not excited (its off system is excited), the color is blue (and dark). If the S cone on system is not excited and the L–M cone on system is, the color is yellow. If both the S cone and the L–M cone on systems are not excited, the color is black (and dark). Whether white and yellow are dark (i.e., gray or brown) is determined by achromatic simultaneous brightness contrast. Humans share this prototypical blue–yellow color vision system, with many other mammals.

The scheme as it stands, however, is deficient in not addressing simultaneous color contrast and in disregarding a fundamental principle in the Land model of color vision, which requires that each cone's system's response be normalized over the visual scene before a comparsion is made for color.

2. Red–Green Opponency

In primates, a second comparison arose with the evolution of two different long-wavelength-sensitive opsins, which split the bright and yellow part of the visible spectrum in two. This provided a new dimension of chromatic contrast (i.e., red–green) (Figs. 9 and 10). The M on and L off channels are not brought together in the same retinal neuron as is the case for the S cone system, presumably because the L and M systems mediate achromatic as well as chromatic contrast. At the cortical level, where many more neurons are available, these systems are brought together in an opponent manner to form cells that respond uniquely to color (i.e., red or green). Again, this scheme disregards simultaneous chromatic contrast as well as the requirement required in the Land model of color vision for normalization of the responses of each cone mechanism over space before color is determined.

This model has another weakness in disregarding a role for the S cone system in determining redness. There is evidence for trivariant interactions in the perception of red–green opponent colors, and this has not been incorporated into this scheme.

There is another controversy in neural modeling of the red–green opponent system. There are two competing models that have been proposed to mediate red–green opponent responses as depicted in Fig. 9. One theory employs all the midget cells of the fovea as well as hypothetical midget-like cells in the parafovea. This is attractive because these cells are very numerous and they possess an essential requirement for transmitting signals for color to the brain. They isolate in one neural channel the signals of one spectral type of cone. However, they have an inappropriate retinal receptive field for a cell mediating color vision. These cells receive excitatory signals from one cone mechanism in

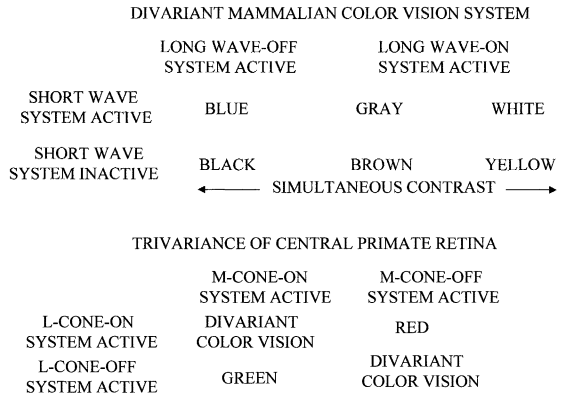


Figure 10 A comparison of the divariant color vision of most mammals and about 2% of human males with that of trivariant color vision.

the center and antagonistic signals from the opponent cone mechanism in the surrounding receptive field. It would be more appropriate to receive the antagonism from the same area of visual space (i.e., its receptive field center rather than from surrounding retina). Nevertheless, this drawback has been disregarded in most models of color vision, which routinely employ this variety of retinal cell. It has been intuitively assumed that visual cortex can correct this deficiency by assembling groups of these cells that subserve coextensive areas of visual space.

A second complexity involved in using the midget system to transmit information about color is that it is also transmitting information about achromatic contrast. Therefore, the brain needs two different detectors to extract the achromatic from the chromatic information from the same neural channel. A scheme suggesting how chromatic and achromatic information, "multiplexed," in a single neural channel is demultiplexed by the brain is shown in Fig. 11. In this scheme synergistic signals from L on and M off channels become logical ways to facilitate the perception of red; similar ones from M on and L off channels are logical to facilitate the perception of green. On the other hand, neurons receiving synergistic inputs from either L on or M on would lose their advantage for chromatic contrast detection but could function to detect achromatic lightness. Similarly, L off and M off channels could function to detect achromatic darkness. By disregarding the source of the cone input achromatic contrast can exploit the fine pixel grain of the fovea. This model does not solve how midget cells with concentric cone opponent fields compare cone responses over coextensive areas of visual space.

A different model has been proposed by Rodieck, who assumes that the midget system does not mediate chromatic contrast. Instead, he postulates a much smaller group of retinal ganglion cells that receive coextensive inputs from L and M cones. These cells would be similar to the S cone retinal channel and perhaps also involve bistratified retinal ganglion cells. There is no anatomical evidence that such retinal cells exist; however, there is physiological evidence that such cells do exist. Further research is required to distinguish which of these two models is correct.

3. Simultaneous Color Contrast

Simultaneous chromatic contrast is not as strong as simultaneous brightness contrast, probably because the comparison of cones in the same area of visual space provides a unique local signal not present in achromatic contrast. Cells sensitive to simultaneous chromatic contrast are first found in visual cortex. Such cells respond best to a red area surrounded by green or the converse. The logical way to establish such cells is to have similar color comparison in one area of space inhibit a similar comparison in neighboring areas or have dissimilar comparisons excite each other (Fig. 12). Cells organized in this manner are first encountered in striate cortex but have also been detected in higher visual areas.

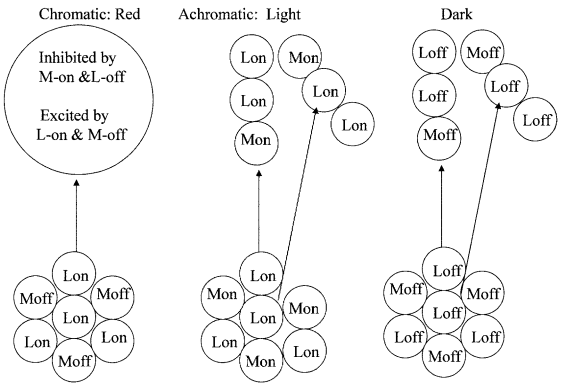


Figure 11 A scheme showing how the same signals from the midget cell system are processed by different sets of parallel circuits, which extract a chromatic signal on the left and an achromatic signal on the right. The chromatic detector mixes synergistic signals from L-on and M-off cells. The achromatic detectors mix only on or only off signals but from either L or M cones.

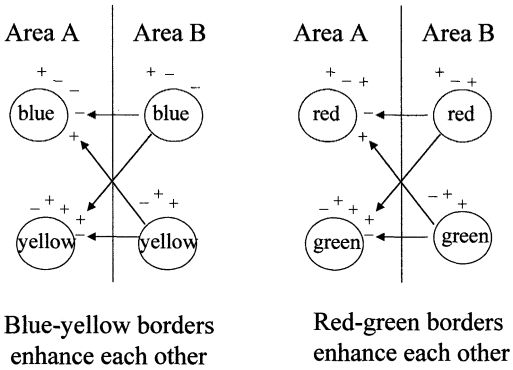


Figure 12 For simultaneous color contrast, the logical arrangement is to have cells responsive to a particular color (in this case "red" or "green" in one area of space) inhibit the same type of neuron in a neighboring area of space. In addition, a cell detecting one color should excite cells detecting the opponent color in neighboring areas of space. In this way, red/green or yellow/blue borders can be enhanced.

C. Land's Retinex Model

Edwin Land emphasized the importance of global influences in color vision. He demonstrated that objects stimulating the retina identically could have totally different colors depending on their surrounding spectral illumination. He proposed a model to handle such global effects called the Retinex model, implying that it depended on both the retina and the cerebral cortex. The model corrects for changes in the spectral characteristics of the illuminant and therefore facilitates color constancy. Color constancy allows us to see a green plant as green in sunlight or artificial light, usually tungsten light, which emits many more long wavelengths than sunlight. The key element in his model is that the signals from each cone mechanism have to be normalized across the visual scene before being compared with another cone mechanism to establish color. It implies that Hering's paired comparisons occur after a stage in which each cone mechanism's response is normalized over a large part of the visual field. Such wide field interactions are quite possible in visual cortex as or before the paired comparisons of cone signals occur.

V. FORM AND COLOR

The most difficult problem in color vision and vision as a whole is to understand how form perception occurs and how once an object's form is determined its color is appended. A reasonable explanation for form perception is Marr's idea of arrays of orientation-selective detectors coding for an object's configuration based on a retinotopic order. The initial detectors of contrast are retinogeniculate neurons with a center-surround receptive field organization based on energy contrast. The mathematical descriptions of such detectors are best represented by a difference of two-dimensional Gaussian-shaped fields. The central field is smaller and overlapped by a larger antagonistic field. The interaction of such detectors can lead to orientation-selective units that detect edges of contrast based on achromatic contrast. The neural machinery for doing this is highly developed in areas of visual cortex serving the fovea. Most of the neurons encountered in striate cortex of primates are sensitive to achromatic (energy) contrast and not selective for color. A smaller amount of neural machinery appears to be devoted to chromatic contrast and color vision in any one area of visual cortex, probably due to its lower spatial resolution.

It is reasonable to assume that chromatic detectors of contrast contribute to form perception in a similar way as achromatic detectors but as a parallel system. Here, wavelength rather than energy contrast determines the activation of the detectors established by cone opponent interactions as independent systems. One system detects blue–yellow chromatic contrasts and the other red–green contrasts. In general, the cues produced by energy contrasts are reinforced by chromatic contrast in detecting the same object. If there are ambiguities, the brain must make a decision favoring one or the other, perhaps favoring chromatic contrast because of its relative independence from shading. Other properties of an object, such as well-defined borders or texture, would be better determined by the achromatic system. In this scheme, three parallel systems for contrast detection are envisaged—one for achromatic (black–white–gray), one for blue–yellow, and one for red–green chromatic contrasts, each arranged with its own retinotopic order. Whether all three systems converge on a single neuron to determine the fused sense of color is unclear. For colors such as magenta, such interactions may occur. In addition, there seem to be neurons that respond to both chromatic and achromatic signals to perform neural functions but are not involved in the perception of color.

A clue to the cortical processing of vision comes from observations of multiple areas of prestriate cortex where the entire visual field is re-represented. These areas send and receive signals from each other. Working together they create a unified and presumably richer impression of the visual world. The actual role of each subarea is poorly understood.

Zeki proposed that one of these subareas, visual area 4, is unique for color processing. He argued that in this area cells correct for color constancy and respond to true color. At earlier stages cells may respond to wavelength contrast but not to color. It has been difficult to prove this hypothesis. Single neuron recordings from most visual areas reveal cells selective for chromatic contrast and color vision but usually in the minority. Visual area 4 has been reported to have a larger proportion of color-selective cells but there is no universal agreement about this.

There is little doubt that the cortex works by dividing different aspects of visual function into anatomically distinct areas and that perception of a unified image depends on the multiple physiological linking of these separate subareas. A critical element in this reasoning is the question of retinotopic order, which seems to be the backbone of form perception.

The integration of form must be based on linking inputs labeled by retinal coordinates. Chromatic contrasts must be handled in a similar way as achromatic contrasts, each with its own retinal coordinates. Does the lower spatial resolving chromatic contrast system get funneled into a separate cortical area for color vision even though color vision depends as much on achromatic contrast as on chromatic contrast? It seems more likely that chromatic and achromatic processing occurs in all the areas to which the parvocellular system projects, and color discrimination as well as color constancy improve as increasingly more visual areas are involved in the processing.

Color discrimination continues to improve with the size of the object being judged. Surfaces subtending 20° and more of visual angle significantly improve color discrimination when compared with those subtending only a few degrees. Color discrimination can integrate over very large areas of visual space and consequently striate cortex. One of the characteristics of neurons in higher visual areas is the relatively large size of their receptive fields. They require inputs from large areas of the retina before deciding to respond. The larger size of visual stimuli activates more antagonistic interactions and therefore could increase the variety of experience. If one is judging a small object, he or she may not be using as much of his or her brain. If the object is enlarged, it begins to activate more cells in higher visual areas and its recognition is improved. This operation may depend on all subareas working in concert through feedback rather than a serial progression from lower to higher areas.

Table I
Classification and Incidence of Color Vision Defects

Type	% Males
Congenital	
Trivariants (three cones present)	
Normal	91.2
Anomalous	
Protanomaly (L cone opsin abnormal)	1.3
Deuteranomaly (M cone opsin abnormal)	5
Tritanomaly (S cone opsin abnormal)	0.001
Divariants (Two cones present)	
Protanopia (L cones absent)	1.3
Deuteranopia (M cones absent)	1.2
Tritanopia (S cones absent)	0.001
Achromats	
Complete achromat (all cones absent)	0.00001
S cone achromat (L and M cones absent)	0.000001
Acquired	
Tritanopia (outer or peripheral retinal disease)	0.01
Protan deuteranopia (inner or central retinal disease)	0.01
Cerebral achromatopsia (brain disease)	0.0000001

VI. DISORDERS OF COLOR PROCESSING

A. Genetic Defects

The major abnormalities of color vision are due to genetic defects (Table I). The most common involve red–green color vision and are due to defects in the genes for L and M cone opsins located on the X chromosome. Therefore, they occur mainly in males. In the most severe cases, both genes fail to function. This leads to achromatopsia, loss of all color vision. Such subjects have only S cones and rods and are called S cone achromats. With only one class of cones, color vision is impossible. Such subjects may experience a primitive form of color vision that depends on a comparison between S cones and rods. Without L and M cones, they also lack high spatial resolution and are legally blind. A second form of achromatopsia is due to a gene defect on chromosome 2 for an ion channel exclusive to cones. These subjects only have functional rods.

In less severe defects, only an L or an M cone gene is functional. Such subjects have blue–yellow color vision obtained by comparing S cones with the remaining long-wave cone mechanism. Subjects who lack L cones are called protanopes. They are insensitive to the long wave end of the spectrum (the red end). Subjects who lack M cones are called deuteranopes. They are sensitive to the entire spectrum. Both protanopes and deuteranopes cannot distinguish reds from greens and have only blue–yellow color vision. They all have high spatial resolution.

In milder defects, both L and M cone opsins are transcribed but the absorption spectrum of one is shifted closer to the other, diminishing the potential for chromatic contrast. Such subjects are called deuteranomalous if the M cone opsin is abnormal and protanomalous if the L cone opsin is abnormal. They have reduced red–green but normal blue–yellow color vision. They all have high spatial resolution. The similarities of colors vary in different ways for each color deficiency (Fig. 13).

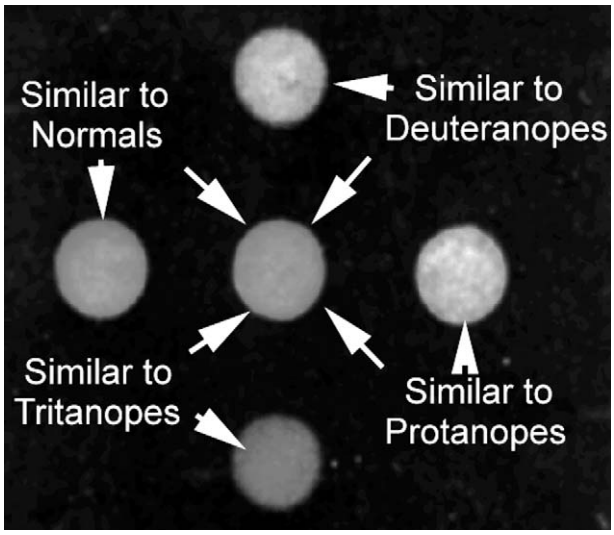


Figure 13 These spots of color appear different to subjects with color deficiencies. Normal subjects see the spot on the left most similar to the central spot. Deuteranopes see the upper, protanopes the right, and tritanopes the lower spot as most similar to the central spot.

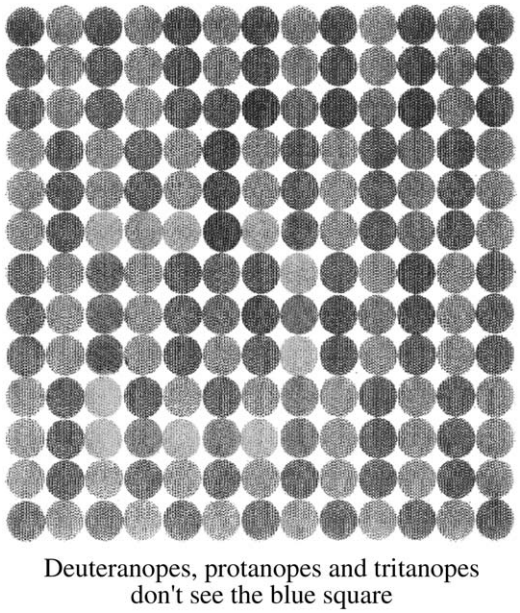


Figure 14 A pseudoisochromatic plate that detects protanopes, deuteranopes, and tritanopes, who fail to see the blue square at the upper right side. (See color insert in Volume 1).

Such color deficiences can also be detected by the use of plates in which spots of different colors that confuse defective subjects are arranged in different sizes so that a figure can be seen by normal subjects but not by defective ones and sometimes vice versa. One that detects both red–green and blue–yellow defective vision is shown in Fig. 14.

Defects in the S cone opsin located on chromosome 7 are rare. They are usually autosomal-dominant mutations. Such subjects are called tritanopes and have only red–green color vision. They retain high spatial resolution. A pseudoisochromatic plate that detects tritanopia is shown in Fig. 15.

Interestingly, there is no gene defect that produces achromatopsia without decreasing acuity, but the converse frequently occurs, including virtually all disorders of the macula.

B. Color Vision Defects Associated with Other Diseases

There is a group of retinal degenerations, primarily genetic in origin such as retinitis pigmentosa, that destroy peripheral vision first and in general slowly progress toward the fovea. These diseases tend to produce tritanopia. This may reflect the possibility that blue–yellow color vision extends further into peripheral retina and therefore is more vulnerable to this disease process. It may also reflect the possibility that S cones are more susceptible to the degeneration than L and M cones.

There is a second class of diseases that tend to affect central vision and are due to defects in either the optic nerve or the macula area of the retina. These can be

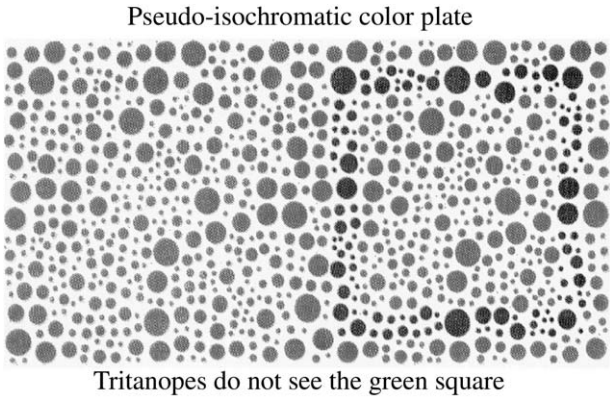


Figure 15 A pseudoisochromatic plate that detects tritanopes, who fail to see the pattern on the right.

associated with red–green color vision defects but they are always accompanied by a diminution in acuity. The vulnerability of red–green defects in these diseases may reflect the possibility that red–green color vision is more developed centrally than blue–yellow vision.

There is a class of autoimmune diseases that arise from cancers that express antigens, which are also expressed by photoreceptors. The immune system detects these antigens and reacts to them both on the cancer cells and on the photoreceptors. This class of diseases has been called cancer-related retinopathies. Some of these diseases are highly specific, involving only cones and not rods. Such subjects can lose all cone vision and consequently color vision. Recently, we found a subject who lost only L and M cones but not S cones or rods. These are examples of acquired achromatopsia.

C. Abnormalities of Acuity and Color Vision

In general, abnormalities of central vision, such as those involving the macula, reduce spatial resolution but have little or no effect on color vision. A remarkable example is ''oligocone trichromasy.'' These subjects have very few cones and greatly reduced acuity but retain normal color vision. This must reflect the fact that color vision integrates over larger areas of the visual field and is disconnected from the high spatial resolution discrimination mediated by achromatic contrast and the ''midget'' system. This does not mean that the midget system does not mediate chromatic contrast and consequently color vision; it must mean that there is considerable overlap in the integration of these midget cell signals that are used for chromatic contrast. There is an alternative view that there are specific ganglion cells divorced entirely from the midget system.

D. Cerebral Achromatopsia

There are rare exceptions-to the tolerance of color discrimination despite large losses in acuity. An occasional subject loses color vision but maintains normal visual acuity after acquired damage of visual cortex. The cases that have been studied most completely with the entire gamut of color testing methods confirm major but not complete loss of color discrimination. They appear to be able to use wavelength contrast to detect objects but are unable to perceive colors from these cues. Such a subject can distinguish the shape and achromatic brightness differences of traffic lights but cannot see them as red, green, or yellow. They appear as washed-out objects of white and gray. Testing reveals a greater tendency for a blue–yellow (tritanopic) than a red–green deficiency. In general, the recovery from damage to visual cortex will usually include a stage in which white and grays are seen first before colors return. The first color to return is usually red. Increasing the size of the stimulus also tends to improve color perception. Nevertheless, such subjects are indeed very deficient in color vision and retain normal acuity.

These clinical findings imply that there is a significant anatomical separation of chromatic and achromatic contrast processing in the cerebral cortex. This is consistent with a cortical area devoted exclusively to the perception of color, but it is not proof. All of these patients invariably have prosopagnosia, the inability to recognize faces, and a scotomatous area, usually in their superior visual field. There are no reports of isolated acquired achromatopsia without other concurrent visual deficits. It is possible that the perception of color requires feedback from prestriate to striate cortex that facilitates the fusion of chromatic contrast with achromatic contrast perception. This feedback may be more important for incorporating the relatively sparse chromatic processing centers in striate cortex into visual perception.

VII. THE FUTURE OF COLOR VISION

Color vision offers a valuable insight into how the cerebral cortex works because it is a well-defined process that depends on three retinal variables serving each area of visual space and is intimately associated with form vision. Many of the logical operations necessary for color vision, which are performed by circuits of nerve cells, are relatively well understood. The major impasse in our understanding of the process is how its clues to contrast are unified into the recognition of form. It is our poor understanding of form vision that makes the final stage of understanding color vision difficult. Understanding how nature interweaves clues from chromatic and achromatic contrast into those for form perception using arrays of neurons and neuronal interactions should lead to a better understanding of not only color vision but also the cerebral cortex in general. The problem of

researching this topic, however, has been that the tools to study it have been and still are too crude. It is necessary to monitor large numbers of single neurons performing their operations within a fraction of a second and to continue doing so during long-term experiments. Such techniques are just beginning to evolve.

See Also the Following Articles

SPATIAL VISION • VISION: BRAIN MECHANISMS • VISUAL CORTEX • VISUAL DISORDERS • VISUAL SYSTEM DEVELOPMENT AND NEURAL ACTIVITY

Suggested Reading

Dacey, D. M. (1999). Origins of spectral opponency in primate retina. In *The Retinal Basis of Vision* (J. I. Toyoda, M. Murakami, A. Kaneko, and T. Saito, Eds.). Elsevier, Amsterdam.

Ehrlich, P., Sadowski, B., and Zrenner, E. (1997). "Oligocone" trichromasy, a rare form of incomplete achromatopsia. *Ophthalmology* **94,** 801–806.

Hubel, D. H. (1988). Eye, brain, and vision, Scientific American Library Series No. 22.

Hurvich, L. (1981). *Color Vision*. Sinauer, Sunderland, MA.

Land, E. H. (1986). Recent advances in retinex theory. *Vision Res.* **26,** 7–21.

Marr, D. (1982). *Vision*. Freeman, San Francisco.

Mullen, K. T. (1990). The chromatic coding of space. In *Vision: Coding and Efficiency* (C. Blakemore, Ed.). Cambridge Univ. Press, Cambridge, UK.

Pokorny, J., Smith, V. C., Verriest, V. G., and Pinckers, A. J. L. G. (1979). *Congenital and Acquired Color Vision Defects*. Grune & Stratton, New York.

Rattner, A., Sun, H., and Nathans, J. (1999). Molecular genetics of human retinal disease. *Annu. Rev. Genet*. **33,** 89–131.

Rizzzo, M., Smith, V., Pokorny, J., and Damasio, A. R. (1993). Color perception profiles in central achromatopsia. *Neurology* **43,** 995–1001.

Rodieck, R. W. (1991). Which cells code for color? In *From Pigments to Perception* (A. Valberg and B. B. Lee, Eds.), pp. 83–93. Plenum, New York.

Stiles, W. S. (1978). *Mechanisms of Color Vision*. Academic Press, New York.

Walsh, V. (1999). How does the cortex construct color? *Proc. Natl. Acad. Sci. USA* **96,** 13594–13596.

Zeki, S. (1993). *A Vision of the Brain*. Blackwell, Oxford.

Zollinger, H. (1999). *Color: A Multidisciplinary Approach*. Wiley–VCH Verlag Helvetica Chimica Acta, Weinheim, Zurich.

Color Vision

GERALD H. JACOBS
University of California, Santa Barbara

GLOSSARY

cones The retinal receptor cells that underlie vision for daylight conditions; there are three classes of cone in retinas of individuals having normal color vision.

metamers Lights of different wavelength compositions that appear identical.

opponent colors Pairs of colors that are perceptually mutually exclusive (e.g., red–green and yellow–blue).

opsin genes Genes specifying the protein component of photopigment molecules; variations in these genes can be mapped directly into variations in color vision.

photopigments Molecules located in photoreceptors that absorb energy from light and initiate the visual process.

principle of univariance Encompasses the observation that the responses of photoreceptors are proportional to the number of photons absorbed by the photopigment they contain independent of the wavelength of the light.

spectrally opponent cells Nerve cells that transmit information useful for the production of color vision; these cells combine excitatory and inhibitory inputs reflecting the activation of different classes of cone.

trichromatic The formal characterization of normal human color vision; derived from the results of experiments in which it is shown that three primary lights can be combined to match the appearance of any other spectral light.

Color vision is the neural process that provides one of the important and dramatic aspects of the brain's interpretation of the visual world. This process is initiated by variation in the spectral distribution of light reaching the eye and culminates in the rich range of human color experience. Fundamental features of color vision are described in this article, including a summary of human capacities to perceive and discriminate color, an outline of the principal biological mechanisms underlying color vision, and a description of variations in color vision among people.

I. THE PERCEPTION AND DISCRIMINATION OF COLOR

At least for simple viewing situations, color experience has been traditionally ordered along three principal perceptual dimensions—brightness, hue, and saturation. The first is considered to be part of both achromatic and chromatic features of color, whereas the latter two are chromatic features. In terms of everyday experience, brightness is that aspect of a percept most closely associated with changes in the intensity of a light yielding visual experience along a dimension that encompasses verbal descriptions running from dim to bright to dazzling. Hue is primarily correlated with the wavelength of light and is typically designated through the use of common color terms—red, green, blue, and so on. Saturation is the degree to which a chromatic sensation differs from an achromatic sensation of the same brightness—for example, yielding sensory experiences that are situated on a continuum changing from white (achromatic) to light pink and to a deep red.

Many aspects of color have been conveniently summarized in geometric configurations. The three

perceptual dimensions of color just described are often represented in a so-called color solid. In this color solid, hues are positioned around the circumference of a circle where they are placed in the order in which they are perceived in a spectrum. The ends of the spectrum (reds and blues) are connected through a series of purple hues. White is located at the center of the hue circle and saturation is then represented as increasing along lines drawn from the center to the various hues located on the circle circumference. The plane surface thus formed encompasses chromatic variation at a single brightness level; increases and decreases in brightness from this level are plotted along a third dimension that runs perpendicular to the plane. Within this color solid any particular color experience can be conceived as representing a location in the three-dimensional space.

A. Environmental Signals

A basic problem for students of color vision is to relate the perceptual dimensions of color to physical features of the environment. Common experience suggests that color is an inherent property of objects: Apples are red; grass is green, and so on. However, is that the case, or is it rather that color is produced by active processes in a perceiver that are initiated by light? Philosophers and others have long debated these alternatives and their arguments continue to reverberate today. Most vision scientists, however, incline to some version of the second alternative and, in so doing, follow the lead of Isaac Newton, the great English scientist whose 16th-century observations caused him to conclude that "the Rays to speak properly are not coloured. In them there is nothing else than a certain Power and Disposition to stir up a Sensation of this or that Colour."

Light reaching the eye from any location in space can vary in intensity (number of photons per unit time) and wavelength. People are sensitive to a narrow band of wavelengths, extending from about 400 to 700 nm (1 nm=10^{-9} m). For most ordinary viewing, the distribution of wavelengths and intensities reaching a viewer depends jointly on characteristics of the source of illumination, on the surface reflectance properties of an illuminated object, and on the geometric relationship between the object and the viewer (Fig. 1). For instance, a snowfield seen in full sunlight yields a very

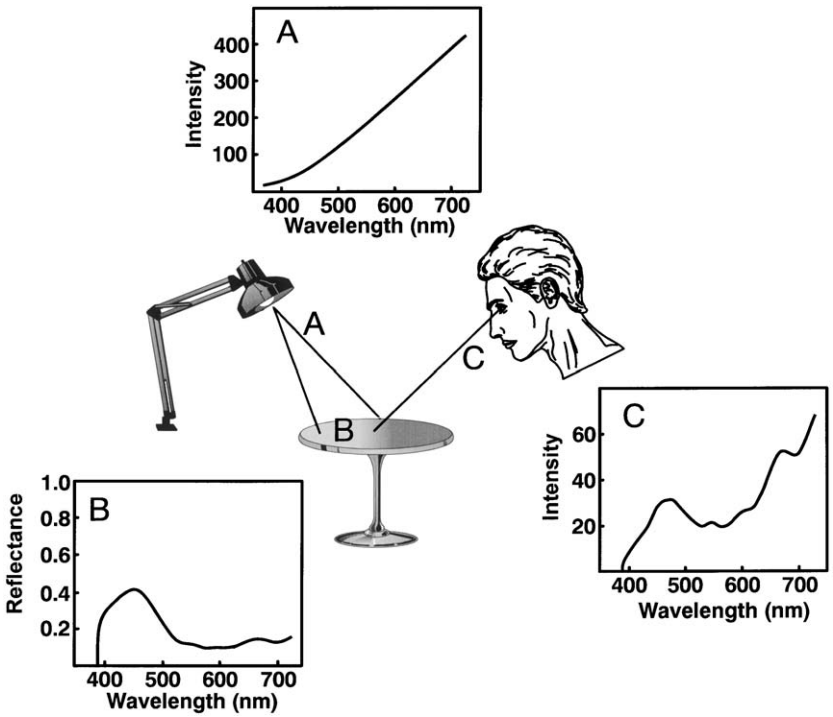


Figure 1 Light reaching a viewer is dependent on the nature of the light source and on the reflectance properties of the object being viewed. In this illustration the spectral distribution of energy emerging from a light source (A) is indicated at the top. The spectral reflectance properties of the surface of the table (B) are given on the left. The spectral distribution of light reaching the viewer (C) that is shown on the right is the product of the spectral distribution of the illuminant and the reflectance property of the object.

different array of wavelengths and intensities for analysis from that provided by an expanse of lawn seen at twilight. Viewed across natural scenes there are usually substantial local variations in the wavelength and the intensity of light, and the number of possible combinations of the two is virtually infinite. All species with sight can exploit intensity variations as an aid to the discrimination of the form, location, and movement of objects. However, only those that also have a capacity for color vision can disentangle the effects of joint variations in wavelength and intensity and in so doing yield the experience of color.

B. The Dimensionality of Color Vision

Color vision has been frequently studied in laboratories by asking observers to judge whether pairs of stimuli—typically viewed as the respective halves of a small, illuminated circle (Fig. 2)—appear the same ("match") or are different. If the two halves are identical in wavelength and intensity content, they of course match. Differences in color appearance may be introduced by changing the intensity or wavelength content of one half or by changing both these features. A change in the intensity of the light in one half of the circle yields a mainly achromatic difference between the two; one side becomes brighter than the other. Changes in the wavelength content of one half may yield a complex of change, in hue and saturation principally but also possibly in brightness. If the observer is then allowed to adjust the relative intensity so as to remove any brightness difference, the two sides will have a pure chromatic difference. It is from this basis that color vision is often studied.

Special cases of this viewing arrangement are those in which the two sides of the small circle differ in wavelength content but have identical color appearances. Pairs of stimuli that differ physically but appear the same are said to be metameric. The occurrence of such metameric matches defines a fundamental feature of human color vision—its limited dimensionality. To better understand this feature, consider color matches obtained when separate lights having fixed spectral content are superimposed on one half of the matching field (Fig. 2). Each of these separate lights is called a primary (P), and the task given to a viewer is to adjust the relative intensity (e) of each of the primaries so as to make that half of the field appear identical to the other half (in turn, traditionally called the test light). The test light can be composed of any fixed combination of wavelengths and intensities. The important result is that most human observers can match the appearance of all test lights by simply adjusting the relative intensities of only three primary lights. Some matches are possible with only one or two primaries, and more than three will work, but three is the minimum needed to complete all matches. Humans are thus said to have trichromatic color vision. There are some qualifications to this conclusion; (i) there are restrictions on the wavelengths that can be employed as primaries, and (ii) for some test lights one of the primaries must be added to the side containing the test light. In any case, the ability of humans to complete matches using only three independent variables means that any color can be represented in a three-dimensional space where the location in that space is specified by the proportions of the three primaries required to capture the appearance of the color. Such spaces are highly useful in the many practical uses of color (e.g., the specification of colored

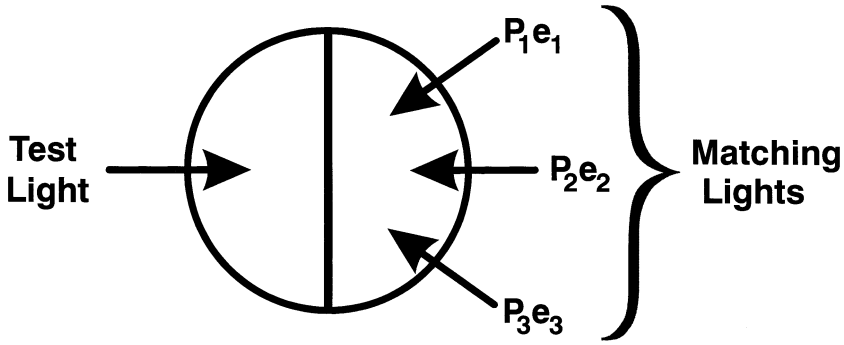


Figure 2 Schematic representation of the color matching test. In this test a subject views a small circular field. The two halves of the field are independently illuminated from different light sources. As shown here, a light of some fixed wavelength and intensity content (a test light) is projected onto the left half of the field. Lights from three separate sources are superimposed on the left. These are primary lights (P_1–P_3) that differ in wavelength. The task of the subject is to adjust the intensities of the three lights (e_1–e_3) until the two halves of the field appear identical. Illustrated is a trichromatic match where combinations of three primary lights are required to complete the match.

signal lights or of the color of a commercial logo) and so have been intensively evaluated.

The mixing of lights reveals another surprising feature of color perception. Many pairs of lights can be added together to yield an achromatic percept. Such pairs are called complementary colors and their occurrence is counterintuitive in that the perceived colors associated with the two wavelengths viewed separately utterly disappear when they are mixed together. In the color solid described previously, complementary colors are placed on opposite sides of the hue circle such that a line drawn between them passes through the location of white. The facts of color mixing make clear that the visual system is not a simple wavelength analyzer. From the viewpoint of the biology of color vision, the fact that people are trichromatic has long been taken to predict the nature of the transformations occurring in the visual system. Indeed, more than two centuries ago it was hypothesized that the fundamental facts of human color perception imply that there must be three kinds of physiological mechanisms in the eye responsible for processing those aspects of light that lead to color. This turned out to be an inspired prediction.

C. Color Discrimination Indices

One way to characterize human color vision is to ask how good we are at making color discriminations. Figure 3 summarizes results from four different tests of human color vision. Just as it does for all other animals, the sensitivity of humans to light varies as a function of the wavelength of the light. The results shown in Fig. 3A illustrate this point. The continuous line is called a spectral sensitivity function and it plots

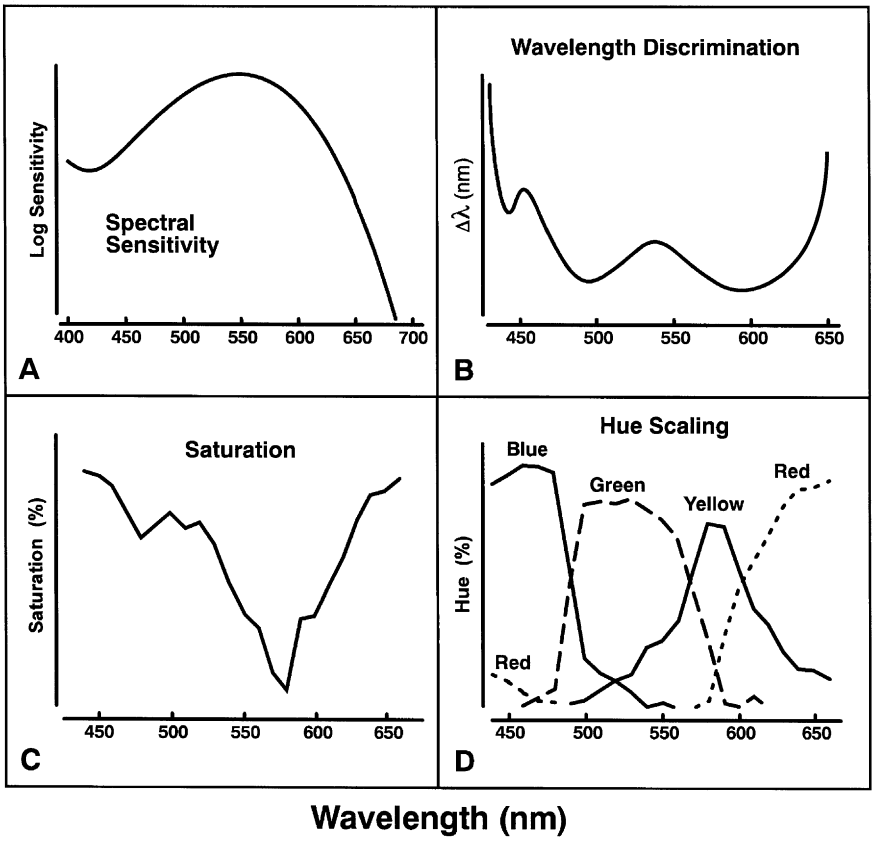


Figure 3 Four measurements of human color discrimination abilities. (A) Spectral sensitivity function. The continuous line plots the reciprocal of the intensity of spectral lights required for detection under daylight test conditions. (B) Wavelength discrimination function. Plotted is a measure of how much wavelength has to be changed in nanometers at each spectral location in order to yield a perceptible color difference. Higher values indicate poorer discrimination. (C) Spectral saturation. The continuous line shows the variation in the saturation of spectral lights as a function of the wavelength of the light. High values indicate greater degrees of saturation. (D) Spectral hues. Plotted as a function of spectral wavelength are the percentages of cases in which each of four separate hue categories (blue, green, yellow, and red) was employed to describe the colors of various lights.

the inverse of the intensity of light required for detection under daylight test conditions. This is a measure of the degree of achromatic variation across the spectrum. The sensitivity variations shown in Fig. 3A predict that colors produced by lights of different wavelengths but equal intensities vary greatly in brightness, and so they do. This characteristic feature of the visual system is so important that a standard spectral sensitivity curve representing the averaged values for many subjects has been derived. This curve has come to be used to provide another specification of the brightness dimension—a metric referred to as luminance. Because luminance has formal properties that brightness does not possess (e.g., luminance units can be additively combined, whereas units of brightness cannot), vision scientists typically prefer luminance as a means of specifying the visual effectiveness of lights.

As noted previously, color depends centrally on the wavelength of light. Figure 3B illustrates how sensitive humans are to changes in wavelength. Plotted is a measure of the size of the wavelength difference ($\Delta\lambda$) required to yield a discriminable change in the appearance of a light for all locations across the spectrum. This result is sometimes called a hue discrimination function, although in fact wavelength change may induce both hue and saturation differences. Note that people are exquisitely sensitive to wavelength changes in two parts of the spectrum. Indeed, under stringent test conditions humans can reliably discriminate some wavelength differences that amount to no more than fractions of a single nanometer. Saturation can be measured in several ways, but they all indicate that spectral lights differ greatly in the degree to which they yield a saturated color. As illustrated in Fig. 3C, lights of both long and short wavelength (usually having an appearance of red and blue, respectively) are seen as highly saturated; lights of about 570–580 nm (seen as yellow) are very unsaturated. The results shown in Figs. 3A–3C imply that people are acutely attuned to detecting differences in several color dimensions when these dimensions are probed separately. However, how good are we at detecting differences when all three perceptual dimensions can vary as they do in normal viewing? One way to appreciate human color vision is to ask how many different colors people can see. A recent estimate suggests that those of us with normal color vision can discern in excess of 2 million different surface colors. Clearly, the human color palette is very extensive.

The results of Figs. 3A–3C are discrimination measures in which people were asked to operate as instruments solely designed to detect differences, entirely ignoring how these lights actually appear. In our ordinary experience with color the rich medium of language is most often employed to give objects color names. It turns out that people can apply color names to lights varying in wavelength and intensity in a sufficiently systematic fashion to generate very reliable indications of color appearance. Figure 3D shows the results from a so-called hue-scaling experiment in which people were asked to name the colors of lights varying in wavelength. They were allowed to use only four different color names (red, yellow, green, and blue), either singly or in pairs; the latter being the case in which one of the four names could be used to modify another (e.g., yellowish red). Not only can people do this reliably but also, perhaps surprising, there is very high consistency of the use of hue names among individuals. There are two important aspects of these results. First, note that in some parts of the spectrum hue changes rapidly with changes in wavelength (i.e., there are well-defined transitions between the color names used by the subjects). Not surprisingly, these regions are those where wavelength discrimination is most acute. Therefore, for example, at approximately 580 nm there are abrupt changes in the hue names given to lights of different wavelengths and this coincides with one of the locations where wavelength discrimination is most acute (compare Fig. 3D with Fig. 3B). Second, there are apparently mutually exclusive categories of hue appearance: A light may be seen to yield both red and yellow components (i.e., reddish yellow or yellowish red) or green and yellow components, but people almost never describe lights as containing both red and green components. The same conclusion holds for yellow and blue. This mutually exclusive pairing of color sensations has long been known and traditionally interpreted as reflecting mutually antagonistic interactions between color processing mechanisms somewhere in the visual system. The antagonistic hue pairs (red/green and yellow/blue) are usually called opponent colors.

The utilization of names to group colors into coherent categories seems universal across all human populations. Results from a survey conducted by Berlin and Kay in 1969, involving an analysis of more than 100 languages, indicated that there are only 11 basic color terms (the English equivalents are white, black, red, yellow, green, brown, purple, pink, orange, and gray). The survey further suggested that these basic color terms have "evolved" among human populations in a reasonably predictable sequence in the sense that rules can be established to specify which

names are present in languages that do not contain the full set of color terms. For example, languages that contain only two basic color terms have black and white, and those with three color terms have black, white, and red. These observations have been taken to imply that basic color terms could reflect universal properties of the organization of the human nervous system. This conclusion has not escaped criticism, particularly from professional linguists, and although attempts have been made to connect these universal color categories to features of visual system physiology, no completely compelling linkages have been established.

D. Context and Constancy

Those aspects of color vision summarized previously were measured with small illuminated spots of specified wavelength and intensity presented in otherwise featureless, usually achromatic, visual fields or with small, carefully calibrated colored papers viewed against an achromatic background. Although ideal for examining human color vision in rigidly controlled circumstances, these conditions are quite unlike everyday visual worlds that are most often rich with both spatial and temporal variegation. From many laboratory studies of the influence of spatial and temporal variables it is clear that perceived color depends not only on the wavelength and intensity of the light reaching the eye from some point in a scene but also on a host of contextual features.

Color context effects have long been appreciated and often exploited for visual effect by painters, tapestry makers, and dyers. For example, the reddishness of a region in a visual scene can be much enhanced if adjoined by a region that appears green. Similarly, surrounding a yellow spot with an expanse of blue can increase the perceived yellowness significantly. These are examples of simultaneous color contrast. Analogous effects on perceived color operate along a timescale. Thus, exposing the eye to a red surface causes a subsequently viewed white surface to take on a greenish cast. Note that these successive contrast effects are similar to the spatial effects in that they tend to enhance the perception of colors that are complementary to those inducing the effect: Yellows enhance blueness; greens enhance redness, and so on. In this sense, they support the idea of a mutual exclusivity of some color sensations similar to that noted previously for the naming of colors.

Seemingly opposed to such contrast effects are instances of what has been called color assimilation. Assimilation occurs when a background color and an interlaced region tend to take on the same color rather than that of a contrast color. For example, consider a white hatched pattern that periodically interrupts a continuous red field. On the grounds of simultaneous color contrast one might expect the white region to take on a greenish appearance, but instead it looks red. It is as if the predominant color has spread into these neighboring regions. The conditions that determine whether contrast or assimilation occurs are incompletely understood, but the relative sizes and configurations of the spatially opposed regions are certainly factors. In any case, both contrast and assimilation effects make it clear that the visual system constructs a color world that is not solely based on the array of wavelengths and intensities coming from each individual region in the scene.

As Fig. 1 illustrated, light reaching the eye from an object depends jointly on surface reflectance of the object and on the nature of the illumination. Surface reflectance is an intrinsic property of objects, but the nature of the illumination is changeable and, consequently, the light reflected from any particular object can vary significantly. For example, the spectral distribution of light reflected from a clump of grass will change drastically as the sun moves from its overhead position to one on the horizon. If color depended solely on the distribution of wavelengths and intensity reaching the eye, one might expect the color of the grass would also change during this period of time. In fact, for the most part, the grass appears to remain resolutely green in appearance. The ability of the visual system to maintain a reasonably consistent object color in the face of these variations in illumination is called color constancy. The occurrence of color constancy implies that somehow the visual system is able to register the nature of the changes in illumination and then to substantially discount these changes. Although the mechanisms in the nervous system that allow this are not well understood, color constancy is supremely important in enhancing stability in our perceptual representations of the color world.

II. COLOR VISION AND THE RETINA

The processing of information that leads to color vision begins in the retina. The retina is an exquisitely organized portion of the nervous system lining the

interior of the back surface of the eye. Each human retina is made up of approximately 115 million cells that act in concert to extract information from the optical image formed on the retinal surface. Our understanding of the processing of color information in the retina and the rest of the visual nervous system is drawn mainly from studies of humans and from our close relatives, the Old World monkeys.

A. Cone Photopigments

There are two classes of photoreceptor—rods and cones. Rods are involved in supporting vision under low-light conditions (technically called scotopic vision). Cones subserve color vision and other characteristic features of daylight vision (photopic vision), such as high spatial and temporal sensitivity. The conversion of energy from light into nerve signals is accomplished in the photoreceptors. The initial step in this transduction process involves the absorption of photon energy by photopigments. Molecules of cone pigment are densely packed in a series of parallel membranes making up one end of the photoreceptor, a physical arrangement that is particularly effective in trapping incoming light. The pigment molecule has two essential components—a protein called opsin and a covalently linked chromophore, the latter being a derivative of vitamin A. Absorption of energy from a photon of light causes the chromophore to undergo a conformational change, an isomerization. This change is virtually instantaneous, complete in no more than 200 fsec (fsec=10^{-15} sec). It serves as a first step in a cascade of molecular changes that produce a modulation in the flow of ionic current across the photoreceptor membrane. This induced electrical change spreads along the length of the photoreceptor and in turn alters the rate of release of neurotransmitter to second-order cells in the retina and thereby communicates a signal onward into the retinal network.

The efficiency with which photopigments absorb light varies continuously as a function of the wavelength of the light. Figure 4 illustrates the absorption spectra for the three classes of cone pigment found in the retinas of people with normal color vision. When properly scaled the shapes of absorption spectra for all photopigments are similar, and thus they can be economically specified by using one number, the wavelength to which they are maximally sensitivity (λ_{max}). The human cone pigments have λ_{max} values at about 420, 530, and 560 nm. The shape and width of the absorption spectrum for the photopigment are dependent on features of the chromophore, whereas the spectral positioning of the pigment along the wavelength axis depends on structural features of the opsin. Note that there is a significant amount of overlap in the wavelengths of light absorbed by these pigments. This is important because it is the comparison of the amount of light absorbed by different pigments that constitutes the basis for the nerve signals that lead to color vision. Among vision scientists, the receptors containing these three pigments are usually termed S, M, and L cones, respectively (shorthand for short-, middle-, and long-wavelength sensitive). The three classes of cone are unequally represented in the retina, having an overall ratio of 1S:3M:6L, with some significant individual variations. As discussed later, this fact is important in understanding some features of human color vision.

A fundamental feature of the operation of pigments is that their response to light is proportional to the number of photons they absorb independent of the wavelength of these photons so that, once absorbed, each photon contributes equally to any generated signal. This means that the signal provided by a given type of receptor contains no information about the wavelength of the light that was absorbed. The blindness of a single type of photopigment to wavelength differences is formally called the principle of univariance. An important functional consequence of pigment univariance is that a retina having only a single type of photopigment could not support any color vision capacity. This is one reason why colors disappear under scotopic conditions when only a single (rod) pigment is operational.

Color matching behavior can be traced directly to the univariant property of the cone pigments. Thus, in a color match of the sort described previously, the subject is actually adjusting the amounts of the primary lights so that the three types of cone will absorb equal numbers of photons from the primary lights and from the test light. A consequence is that color matches are very sensitive to the relative spectral positions of the cone pigments. Indeed, other things being equal, the retinas of two individuals who set different color matches must contain photopigments that differ in their spectral positioning. This single fact constitutes one of the earliest and still most persuasive examples of a compulsive linkage between behavior and nervous system organization. Most humans have trichromatic color vision because their retinas contain three classes of cone pigment, each of which behaves univariantly. It is obvious from the absorption spectra

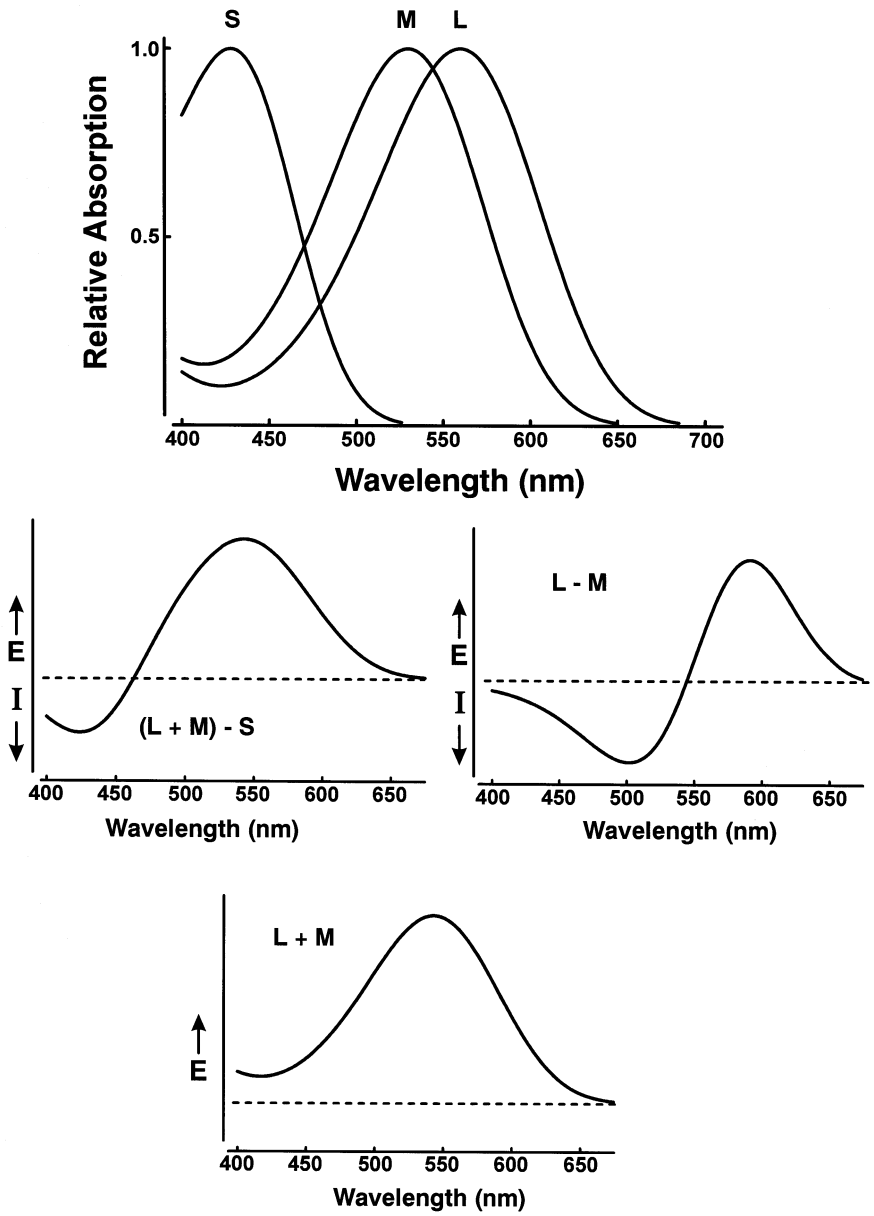


Figure 4 Response properties of photopigments and nerve cells. (Top) The absorption curves for the three classes of cone found in the eyes of people with normal color vision. The peak sensitivities of the three cone types are at 430 nm (S), 530 nm (M), and 560 nm (L). The three graphs at the bottom illustrate the responses of nerves cells in the visual pathway that combine inputs from the three cone types. (Middle) Spectrally opponent cells that show additive and subtractive combinations of cone signals. The inputs are indicated. E, excitation; I, inhibition. (Bottom) Response property of a nonopponent cell that additively combines signals from M and L cones.

of Fig. 4 that S cones absorb very little light from the middle- to long-wavelength portion of the spectrum. A consequence of this is that only two primary lights are required to complete color matches in this part of the spectrum (i.e., over those wavelengths, three-variable, trichromatic color vision gives way to two-variable, dichromatic color vision).

B. Genes and Photopigments

Single genes specify the opsins of each of the types of cone pigment. The human opsin genes were first isolated and sequenced in 1986. As long predicted from observations about the inheritance of color vision defects, the genes for the M and L pigments are located

on the q-arm of the X chromosome where they lie in a head-to-tail tandem array. Each of the X chromosome genes specifies an opsin composed of 364 amino acids, and the two are so similar in structure that the amino acid sequences for M and L pigments are 96% identical. Most individuals (at least 75% of the population) have more than one copy of the M cone pigment gene, although the functional significance of these ''extra'' genes remains debatable. The S cone opsin gene is autosomal, located on chromosome 7. It specifies a photopigment opsin containing 348 amino acids. The sequence of this gene differs from the other two enough to produce an S cone opsin that is 40–45% identical to that of the M and L cone pigments.

The close physical proximity on the X chromosome of the M and L opsin genes and their great sequence similarities makes them particularly susceptible to unequal recombination during meiosis. Such recombination events can result in the loss or the gain of complete copies of the opsin genes, or they can produce novel genes through recombination of partial sequences drawn from the original M and L genes. All of these changes will have an impact on the nature of the photopigments that get produced and hence they will alter significantly the details of color vision in that individual. For example, a loss of either the M or the L pigment gene reduces the retina so that it contains two, not three, types of cone pigment and consequently limits the individual to a dichromatic form of color vision.

C. Evolution of Human Color Vision

In recent years, the study of color vision in nonhuman species has expanded greatly, as has our view of the nature of opsin genes and photopigments in a host of different species. This has begun to allow an understanding of the evolution of human color vision. Comparison of opsin gene sequences indicates that visual pigments have an ancient origin and that there is evidence for the presence of two separate cone pigments very early in vertebrate history (perhaps as long as 400 million years ago). This arrangement would have provided the necessary basis for a dichromatic color vision system, although there is currently no way of knowing if that potential was actually realized. The nature of color vision and the physiology that allows for color vision vary significantly among present-day vertebrates. As previously mentioned, humans normally have three classes of cone pigment and trichromatic color vision. The retinas of many fishes, reptiles, and birds contain at least four classes of cone pigment, and there is evidence to support the conclusion that these may well provide a four-variable form of color vision—a tetrachromacy. On the other hand, the retinas of most mammals (e.g., domestic dogs and cats) contain only two types of cone pigment allowing a basic dichromacy. This difference has led to the hypothesis that the potential for a sophisticated form of color vision was lost sometime early in mammalian history. Among mammals, the presence of three types of cone pigment and trichromatic color vision is restricted to primates. From this fact, it can be concluded that our exceptional color vision sense appeared during the evolution of primates.

Comparison of opsin genes and color vision in Old World primates (in particular, Old World monkeys, apes, and humans) and in New World monkeys indicates that the essential change from a basically dichromatic form of color vision to uniform trichromatic color vision occurred shortly after the divergence of Old World and New World primates approximately 30–40 million years ago. The means for accomplishing this was an X chromosome opsin gene duplication that yielded two separate M and L genes. Whether that duplication arose from identical X chromosome genes that eventually diverged in structure or whether it was built from the baseline of a polymorphism at a single gene site is not clear. In any case, the origins of human trichromacy can be traced to changes that first appeared in the retinas of our early primate ancestors. An important consequence of this is that human color vision is very similar in its detail to the color vision enjoyed by all contemporary Old World monkeys and apes.

D. Neural Mechanisms for Color Vision in the Retina

As noted previously, the human retina contains multiple types of cone photopigment. Multiple cone types are a necessary but not sufficient basis to support a color vision capacity. In addition, there must also be an appropriately organized nervous system. The extraction of information that allows for color vision begins in the neural networks of the retina.

In addition to the photoreceptors, there are four other major classes of nerve cell in the retina: bipolar, ganglion, horizontal, and amacrine cells. These form intricately organized vertical and horizontal

possessing networks arrayed through the thickness of the retina with bipolar and ganglion cells serving as the principal vertical pathway and the other two types providing a rich array of horizontally organized connections. Each of the four types of cell in turn consists of discrete subtypes. Currently, a majority of the latter are poorly defined, both structurally and functionally, but it is believed that it will eventually be possible to characterize as many as 50 distinct types of cells in the primate retina. The main business of the networks formed by these cells is to perform the first stages in the processing of the retinal image. These tasks include the analysis of local spatial and temporal variations and the regulation of visual sensitivity. The processing of color information proceeds in the context of all these other analyses.

Because each cone type behaves univariantly, the extraction of color information requires the comparison of the signals from cone types containing different photopigments. There are two principal ways in which cone signal information is combined in the nervous system: additively (spectrally nonopponent) or subtractively (spectrally opponent). Cells of the former type sum signals from the L and M cone types (L + M). Because they do not respond differentially to wavelength differences irrespective of the relative intensity, cells so wired cannot transmit information useful for the production of color vision. The response properties of spectrally opponent cells are produced by convergence of excitatory and inhibitory signals onto recipient nerve cells. The cone signal combinations are classified into two main groups: L−M and (L + M)−S (for each of these types the signs can be reversed, i.e., there are both L−M and M−L types). Response profiles for opponent and nonopponent cells are shown in Fig. 4. It can be seen that, unlike the nonopponent cells, spectrally opponent cells will yield different responses to different wavelengths of light irrespective of the relative intensities of those lights. They thus can transmit information that may be useful for supporting color vision. With regard to color information, the output cells of the retina (the ganglion cells) are classified into three groups: those that transmit nonopponent information into the central nervous system, those carrying M−L information, and those whose response patterns are (M + L)−S (Fig. 4). Much is known about the anatomy of such cells and the input pathways that yield these different response patterns.

Several structural and functional properties of the retina map are directly mapped into the quality of human color vision. For example, S cones are sparsely distributed across the retina and are absent entirely from its very central portion (the fovea). A consequence is that the color information contributed from S cone signal pathways is lost for stimuli that are very small (i.e., trichromacy gives way to dichromacy under such viewing conditions). The connection pathways from the L and M cones also vary across the retina. Although the picture is still somewhat controversial, in general there is an observed decrease in the relative potency of L/M spectrally opponent signals toward the peripheral parts of the retina. This decrease is presumed to be a factor in the gradual decline in sensitivity of red/green color vision for stimuli located away from the direction of gaze. Finally, due to variations in the neural circuitry, there are significant differences in the spatial and temporal sensitivities of the different types of retinal cells. One consequence of this is that our color vision becomes progressively more restricted for regions in the visual scene that are very small and/or are changing very rapidly. Thus, human color vision is trichromatic only for relatively large and slowly changing stimuli, giving way to dichromatic color vision as the space/time components of the stimulus are increased and, eventually, one can lose color vision entirely (i.e., become monochromatic) for very small and/or very rapidly changing stimuli.

III. THE CENTRAL VISUAL SYSTEM AND COLOR

Behavioral studies of vision have led to a consensus view that the facts of human color vision can best be understood as reflecting the operation of three separate mechanisms. These mechanisms are conceived as implying the presence of three parallel channels of information in the central visual system. Two of these are opponent channels reflecting, respectively, the mutual antagonisms of red and green and of yellow and blue. The third is a nonopponent channel that provides achromatic information (the luminance mechanism). The spectrally opponent and nonopponent cells of the type described previously were first recorded approximately 40 years ago not in the retina but in the lateral geniculate nucleus (LGN), which is a large thalamic structure that serves as a relay site situated in the central visual pathway between the retina and the visual cortex. Following the discovery of these cells it was immediately recognized that there is a compelling analogy between the physiological results and the behavioral conception of the color mechanisms. Thus, the L−M cells and the (L + M)−S cells

(Fig. 4) have many characteristic features that appear to be like those of the behaviorally defined red/green and yellow/blue color channels. Similarly, the L + M cells have the appropriate spectral sensitivity as well as many other features that appear to make them ideal candidates to serve as the luminance mechanism.

Over time, however, closer examination of the relationships between the standard behavioral model and the physiological results has revealed a lack of complete correspondence between the two, and this forced a reevaluation of the idea that the spectrally opponent cells of the retina and LGN directly represent the color mechanisms documented in behavioral experiments. To note just one problem, consider the perceptual observation that the spectrum takes on a distinctly reddish appearance in the short wavelengths (this can be seen in the hue naming results of Fig. 3C). Among color vision theorists, the conventional explanation for this fact is that the red/green color mechanism must receive some signals from the S cones. However, electrophysiological studies indicate that the L−M cells do not have an S cone input so they cannot directly account for this feature of red/green color vision. Many other disparities between the behavioral aspects of human color vision and the physiology of cells located early in the visual pathway have also been enumerated. The inevitable conclusion is that although the spectrally opponent cells clearly transmit the information necessary for color vision, the response properties of these cells cannot directly account for many features of human color vision. This means that there must be some further transformations of the signals provided by spectrally opponent cells and those transformations are assumed to take place somewhere in the visual cortex.

The picture of how color information is processed and elaborated in the visual cortex is still very sketchy. Researchers have tried to understand how color information is encoded by cortical neurons, and they have pursued the hypothesis that color processing might be localized to particular component regions of the cortex. Analysis of the response properties of single cortical cells has been avidly pursued. Unfortunately, the results have led to conflicting views as to how cortical cells encode color information. This is understandable since the task is far from simple. Much of the difficulty arises from the fact that although the responses of cells in the retina and in the LGN are only modestly dependent on the spatial and temporal features of the stimulus, the responses of cortical cells are very much conditioned by these properties. The consequence is that studies utilizing differing spatio-temporal stimulus features have often reached quite different conclusions about the nature of cortical color coding. What is known, both from single cell studies conducted on monkeys and imaging studies of human brains, is that color-selective responses can indeed be recorded in primary visual cortex (area V-1) and at other locations in the extrastriate cortex. Studies also make clear that transformations of the spectrally opponent responses of LGN cells do occur in V-1 and that such transformations arise, at least in part, from dynamic interactions between cortical cells. For example, neural feedback circuits may significantly amplify contributions from signals originating in S cones relative to those S cone signals recorded at more peripheral locations. Such a change brings the physiological picture more in line with behavioral measures of color vision. Similarly, at least some cells in V-1 are known to combine signals from the two classes of LGN spectrally opponent cells. This too brings the physiology closer to the coding scheme inferred from behavioral studies of color vision. Although the relationships between cortical codes for color and psychophysical models of color coding are not clear, significant progress is being made toward their rationalization.

For years it has been argued that the processing of color information can be localized in the extrastriate visual cortex. From studies of both monkey and human brains, regions in the lingual and fusiform gyrus (sometimes characterized as area V-4) have been identified as particularly important for color. In humans this area responds robustly to the presentation of stimuli designed to specifically probe color vision, and it has been reported that cells in monkey area V-4 exhibit some forms of color constancy. Providing additional support for the idea of localized cortical representation of color are clinical descriptions of patients who have suffered a loss of color vision as a result of cortical damage (cerebral achromatopsia). The nature of the color vision loss in cerebral achromatopsia is complex and beyond the scope of this article. For our purposes, the important fact is that although the locus of damage that results in cerebral achromatopsia varies significantly among described cases, it most often does include area V-4.

IV. VARIATIONS IN HUMAN COLOR VISION

In 1794, John Dalton, the celebrated chemist, offered a series of astute observations about his own color

perceptions. He noted, for instance, that the spectrum appeared to him to contain just three colors (yellow, blue, and purple) and that flowers that appeared pink to others seemed to him "an exact sky blue." Dalton eventually discovered that many other individuals shared his atypical color perceptions and he was led to offer a hypothesis to explain the condition based on the presumed presence of some unusual intraocular filter. Retrospective analysis shows that Dalton suffered from deuteranopia, a common congenital color vision defect that affects about 1% of all males. Dalton was not the first to learn of defective color vision among humans, and his explanation of the source of the defect was wrong, but his detailed descriptions were very instrumental in initiating a long series of examinations of color defect that continue to the present day. Studies of color vision defects, and of other individual variations in color vision, have provided significant insights into the biology of color vision. In addition to the congenital color vision defects, color vision change may be acquired through a wide variety of visual system pathologies as well as senescent changes occurring in the visual system.

A. Congenital Color Vision Defects

There are many distinct types of congenital color defect and each yields a characteristic pattern of results on behavioral tests of color vision (these include both laboratory tests of the kinds described previously and the familiar plate tests used for the clinical screening of color vision). Those most severely impacted have dichromatic color vision, like that of John Dalton. When tested in color matching experiments of the sort described previously, such individuals require only two primary lights to complete all the color matches. The number of discrete color discriminations that can be made is also severely reduced. For instance, in the wavelength discrimination test (Fig. 3B), dichromats basically fail to discriminate among wavelengths in the middle to long wavelengths while retaining an ability to discriminate wavelengths shorter than approximately 490–500 nm from wavelengths longer than that value. As a result, they are commonly characterized as "red/green color blind." This is a somewhat misleading description since dichromats are not blind to color in this part of the spectrum: Rather, they simply fail to discriminate among these colors. The full gamut of saturations that dichromats see is also significantly smaller than that seen by people with normal trichromatic color vision. There are two major types of congenital dichromacy, protanopia and deuteranopia. Individuals of the two types differ both in the details of color discriminations they can make and in their spectral sensitivity.

In addition to the dichromacies, there are also trichromatic individuals whose color vision nevertheless differs significantly from normal trichromacy. These people are said to have anomalous trichromatic color vision and, like the dichromats, their color vision defects can be detected by color discrimination tests. Anomalous trichromats are of major two types: protanomalous and deuteranomalous. An essential difference between the two types emerges from a simple color matching task in which an individual is asked to complete a color match by adding green and red lights in proportion to match a yellow test light. Dichromats so tested fail to set a reliable match since they cannot discriminate among colors in this part of the spectrum. All trichromatic individuals can make such matches. Relative to the normal, the protanomalous observer will require additional red light to complete the match, whereas the deuteranomalous individual needs additional green light. A wide range of other discrimination tests can be similarly used to distinguish among the trichromatic subtypes. Although the trichromatic anomalies are usually classified as color vision defects, it is important to note that in fact many anomalous trichromats have quite acute color vision.

Individuals with these congenital defects thus differ significantly from those with normal color vision in their discrimination abilities. What do they actually see? This intriguing question has not been easy to answer. Growing up as a minority in a color-coded world, most color defectives acquire a rich vocabulary of color names that they learn to use in a discerning fashion. They can do this because there are many secondary cues to allow one to apply color names in accord with the majority view (e.g., familiar objects are often characteristically colored and objects of different color often have systematic brightness differences). It is only when these secondary cues become scarce or disappear entirely (as they do in formal tests of color vision) that color defect become readily apparent. Some insights into the color world of the color defective come from examination of the occasional individual who has defective color vision in one eye and normal color vision in the other. Similarly, inferences can be drawn from comparison of the discrimination abilities of the normal and the color defective. These pieces of evidence suggest that the two

major types of dichromat view the world in what to the normal would be varying shades of only two hues—blues and yellows. Furthermore, there are significant losses in the saturation of lights so that the perceptual world of the dichromat is distinctly pallid relative to that of the normal. Recently, computer algorithms have been developed that allow one to transform a digitized colored image to obtain a perceptual prediction of the dichromatic world. Although defective color vision is often treated as a benign condition, there is documented evidence to indicate that those who have defective color vision find that it provides a real barrier in many aspects of normal life—from things as mundane as making judgments about the ripeness of a piece of fruit to issues as far reaching as a choice of a profession.

A vast majority of those with defective color vision are male. The incidence of defective color vision among males is given in Table I. In total, approximately 8% of the in Western Europe and in the United States can be classified into one or another of these diagnostic categories with 5% of all males having deuteranomalous color vision. The incidence of red/green defective color vision is low in females, being approximately the square of the frequency of the corresponding male color vision defects.

There are also rare congenital color vision defects that arise from change or alteration in the ability to discriminate among short-wavelength lights (tritanopia and tritanomaly). The incidences of these defects do not differ for male and females. Finally, in addition to dichromats and anomalous trichromats, a relatively few individuals lack color vision completely (are monochromatic) or only show an ability to discriminate color in very restricted test circumstances.

Table I
Incidence of Congenital Defective Color Vision among Caucasian Males[a]

Diagnostic category	%
Protanopia	1.0
Deuteranopia	1.1
Tritanopia	0.001
Protanomaly	1.0
Deuteranomaly	4.9
Monochromacy	0.003

[a]The incidence values are based on large-scale surveys of defective color vision and should be considered approximate.

B. The Biology of Congenital Color Defects

Protanopia, deuteranopia, protanomaly, and deuteranomaly all arise from changes in the X chromosome opsin genes. As noted previously, unequal recombination events that occur during meiosis can lead to additions and losses of genes from the X chromosome as well as to the production of novel genes (Fig. 5). Thus, a loss of the L cone opsin produces a protanopic dichromacy; loss of the M opsin yields deuteranopia. On the other hand, the anomalous trichromacies reflect the genomic presence of atypical opsin genes. In these cases, a new pigment replaces either the normal M pigment (in cases of deuteranomaly) or the normal L pigment (in protanomaly). The spectrum of this new pigment is shifted so as to be closer to that of the other (normal) M or L pigment. Depending on which of several atypical genes is present, the spectral separation of the M and L pigments is made either greater or smaller and severity of the color vision defect is correspondingly more or less. Only females that are homozygous for these defective X chromosome opsin genes show classical indications of these red/green color defects, thus explaining why the frequency of defective color vision in females is low. Pedigrees of color vision defects have long indicated that in the usual case sons inherit a color defect from mothers who carry a defective gene but who have normal color vision. The tritan defects similarly result from changes in S opsin genes. Since these represent changes in autosomal genes, the incidences of these defects do not differ for males and females. The biology underlying the complete loss of color vision is not straightforward and, in different cases, is apparently traceable to either photopigment or nervous system alterations.

Surveys indicate that the incidence of defective color vision varies significantly among different population groups. The highest frequencies are found in the United States and Western Europe, where approximately 7–9% of all males show a form of red/green color vision defect. Elsewhere, the incidence is often lower. For instance, several populations have been reported to show comparable defects in only about 1% of the males. These regional/ethnic differences have received various interpretations, perhaps the most popular being that the higher rates of defect are found in those populations in which there has been a relaxation of natural selection pressure against color vision defects. According to this view, those societies that are least altered from their hunter/gatherer origins should have the lowest incidences of defective color

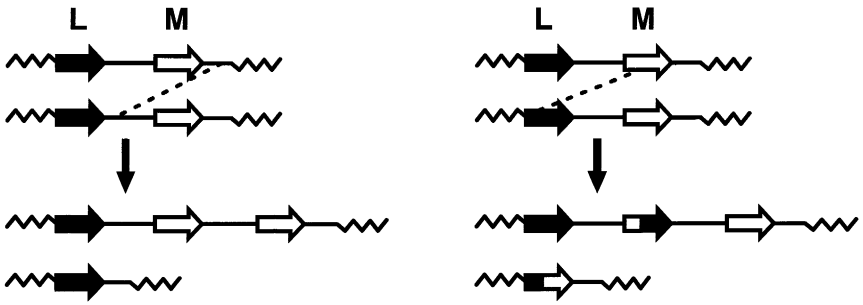


Figure 5 X chromosome gene combinations leading to defective color vision. Arrows represent genes for M and L cone opsins. Illustrated are two examples of unequal recombinations that yield new gene arrays. The dashed lines indicate where crossing over occurs during meiosis. (Left) The unequal crossover yields two new gene arrays, one with an L opsin gene and two copies of the M opsin gene. This genotype is common in individuals having normal color vision. The second result is a deletion of the M cone opsin gene. A male with this genotype would have dichromatic color vision. (Right) The crossovers occur within the gene. This produces new genes whose pigment products will differ from those of the normal M and L opsin genes, and depending on the resultant combination of genes the effects on color vision can be subtle or drastic.

vision. There is no compelling evidence to support this idea, but neither can it be flatly rejected.

C. Subtle Variations in Human Color Vision

The differences in color vision between normal trichromats and various color defectives are major and dramatic, but there are also reliable differences in color vision among individuals within each of these groups and study of these small variations can provide a useful tool for understanding the biological basis of color vision. One particular example comes from the study of variations in the color matching among normal trichromats. In a standard version of the color-matching task, people are asked to adjust the proportions of red and green primaries to match a yellow test light (with lights drawn from this part of the spectrum, trichromats make dichromatic matches). Color discrimination is particularly acute for lights from this part of the spectrum and, consequently, subjects make extremely reliable matches. It has long been known that individual matches may differ in that some people consistently require relatively more red light to complete the match, and others need relatively more green light. As previously explained, differences in color matches mostly arise from differences in the spectral positioning of the retinal photopigments, so an implication is that individuals who make different color matches must have different photopigments.

In recent years, the biological basis of this small variation in color matching has been uncovered. As predicted, it reflects the fact that normal human trichromats show small variations in the spectral positioning of their photopigments. Unexpectedly however, the variation is not continuously distributed in the population but rather appears to have a discrete character. This variation has been shown to result principally from a polymorphism in the gene specifying the L cone opsin. This polymorphism, involving a difference of only a single nucleotide, in turn causes a variation in a single amino acid in the photopigment molecule and thus there are two forms of the L cone pigment. This variation allows the L cone pigment to occupy either one of two spectral positions, with the two varying in their λ_{max} values by about 4 nm. A male who has the longer of the two pigment versions will require systematically less red light in the color match than will a male who has the shorter of the two pigments. These two polymorphic variants of the L cone opsin gene are nearly equally represented in the population. It is remarkable that this very small genetic variation sorts individuals of normal color vision into groups that go through life experiencing the colors of the world as biased to appear either slightly redder or slightly greener.

See Also the Following Articles

EYE MOVEMENTS • MULTISENSORY INTEGRATION • SENSORY DEPRIVATION • SPATIAL VISION • VISION: BRAIN MECHANISMS • VISUAL DISORDERS • VISUAL SYSTEM DEVELOPMENT AND NEURAL ACTIVITY

Suggested Reading

Abramov, I., and Gordon, J. (1994). Color appearance: On seeing red—or yellow, or green, or blue. *Annu. Rev. Psychol.* **45,** 451–485.

Backhaus, W. G. K., Kliegl, R., and Werner, J. S. (Eds.) (1998). *Color Vision—Perspectives from Different Disciplines*. de Gruyter, Berlin.

Dacey, D. M. (1999). Primate retina: Cell types, circuits and color opponency. *Prog. Ret. Eye Res.* **18,** 737–763.

Foster, D. H. (Ed.) (1991). *Inherited and Acquired Colour Vision Deficiencies*. Macmillan, London.

Gegenfurtner, K. R., and Sharpe, L. T. (Eds.) (1999). *Color Vision: From Genes to Perception*. Cambridge Univ. Press, Cambridge, UK.

Gouras, P. (Ed.) (1991). *The Perception of Color*. CRC Press, Boca Raton, FL.

Jacobs, G. H. (1993). The distribution and nature of colour vision among the mammals. *Biol. Rev.* **68,** 413–471.

Kaiser, P. K., and Boynton, R. M. (1996). *Human Color Vision*, 2nd ed. Optical Society of America, Washington, DC.

Komatsu, H. (1998). Mechanisms of central color vision. *Curr. Opin. Neurobiol.* **8,** 503–508.

Lamb, R., and Bourriau, J. (Eds.) (1995). *Colour: Art & Science*. Cambridge Univer. Press, Cambridge, UK.

Martin, P. R. (1998). Colour processing in the retina: Recent progress. *J. Physiol.* **513,** 631–638.

Nathans, J. (1999). The evolution and physiology of human color vision: Insights from molecular genetic studies of visual pigments. *Neuron* **24,** 299–312.

Neitz, M., and Neitz, J. (2000). Molecular genetics of color vision and color vision defects. *Arch. Ophthalmol.* **118,** 691–700.

Wandell, B. A. (1995). *Foundations of Vision*. Sinauer, Sunderland, MA.

Consciousness

MICHAEL S. GAZZANIGA
Dartmouth College

Just what is consciousness? The answer, I believe, is that consciousness is an instinct—a built-in property of brains. Like all instincts, it is just there. You do not learn to be conscious and you cannot unlearn the reality of conscious experience. Someday we will achieve a more mechanistic understanding of its operation, but I warn you now: That won't be especially fulfilling on a personal level. We have to shed our expectation that a scientific understanding of consciousness will sweep away our sense of strangeness, like finding out how ships get in bottles. Take our reproductive instinct. Does it help our sense of desire to understand the role of testosterone when we see a shapely figure across the room? Or take the human instinct for language. Does it help us to enjoy language if we understand that grammar is a universal built-in reflex but that our lexicon is learned? Understanding the problem of consciousness may be essential to our ultimate ability to deal with some mental disorders. Disorders of conscious experience, whether autism or schizophrenia or dementia, will be illuminated by a mechanistic understanding of personal conscious experience.

My own thinking on this topic started early, in Roger Sperry's laboratory at the California Institute of Technology on an afternoon almost 40 years ago, when I first tested a split-brain patient. It seemed that, whatever consciousness was, you could have two of them after the surgical severy of the corpus callosum connecting the two cerebral hemispheres. Mind Left did not appear to know about Mind Right and vice versa. Those first impressions, which still endure, nevertheless left much to be desired as a sophisticated perspective on the question of consciousness. My plight as a researcher echoed Tom Wolfe's admonition to practice writing for 20 years before you seek a publisher.

Classic split-brain research highlighted how the left brain and the right brain serve distinctive functions and led us to believe that the brain is a collection of modules. The left brain (or hemisphere) is specialized not only for language and speech but also for intelligent behavior. After the human cerebral hemispheres are disconnected, the patient's verbal IQ remains intact and his problem-solving capacity (as observed in hypothesis formation tasks) remains unchanged for the left hemisphere. Indeed, that hemisphere seems to remain unchanged from its presurgical capacity. Yet the largely disconnected right hemisphere, which is the same size as the left, becomes seriously impoverished for many cognitive tasks. While it remains superior to the left hemisphere in certain activities (in recognizing upright faces, having better skills in paying attention, and perhaps in expressing emotions), it is poorer after separation at problem solving and many other mental activities.

Apparently the left brain has modules specialized for higher cognitive functions, while the right has modules specialized for other functions. Visuospatial function, for example, is generally more acute in the right hemisphere, but left hemisphere integration may be needed to perform higher order tasks. The use of

This article is reproduced from Gazzaniga, M. S. (1999). The interpreter within: The glue of conscious experience. *Cerebrum* **1**(1). with permission of the Dana Foundation.

tactile information to build spatial representations of abstract shapes appears to be better developed in the right hemisphere, but tasks such as the Block Design test, which are typically associated with the right parietal lobe, appear to require integration between the hemispheres in some patients. Furthermore, even though the right hemisphere is better able to analyze unfamiliar facial information than is the left hemisphere, and the left is better able to generate voluntary facial expressions, both hemispheres can generate facial expression when spontaneous emotions are expressed.

In addition to the skills named previously, our big human brains have hundreds if not thousands more individual capacities. Our uniquely human skills may well be produced by minute, circumscribed neuronal networks, sometimes referred to as ''modules,'' but our highly modularized brain generates a feeling in all of us that we are integrated and unified. If we are merely a collection of specialized modules, how does that powerful, almost self-evident feeling come about?

I. THE ''INTERPRETER''

The answer appears to be that we have a specialized left hemisphere system that my colleagues and I call the ''interpreter.'' This interpreter is a device (or system or mechanism) that seeks explanations for why events occur. The advantage of having such a system is obvious. By going beyond simply observing contiguous events to asking why they happened, a brain can cope with such events more effectively should they happen again.

We revealed the interpreter in an experiment using a ''simultaneous concept test.'' The split-brain patient is shown two pictures, one presented exclusively to his left hemisphere, one exclusively to his right. He is then asked to choose from an array of pictures the ones he associates with the pictures that were presented (or ''lateralized'') to his left brain and his right brain. In one example of this, a picture of a chicken claw was flashed to the left hemisphere and a picture of a snow scene to the right. Of the array of pictures then placed in front of the subject, the obviously correct association was a chicken for the chicken claw and a shovel for the snow scene. Split-brain subject case 1 did respond by choosing the shovel with his left hand and the chicken with his right. Thus each hemisphere picked the correct answer.

Now the experimenter asked the left-speaking hemisphere why those objects were picked. (Remember, it would only know why the left hemisphere had picked the shovel; it would not know why the disconnected right brain had picked the shovel.) His left hemisphere replied, ''Oh, that's simple. The chicken claw goes with the chicken, and you need a shovel to clean out the chicken shed.'' In other words, the left brain, observing the left hand's response, interprets the response in a context consistent with its own sphere of knowledge —one that does not include information about the snow scene presented to the other side of the brain.

One can influence the left-brain interpreter in many ways. As I mentioned, we wanted to know whether the emotional response to stimuli presented to half of the brain would influence the emotional tone of the other half. Using an optical computer system that detects the slightest eye movement, we projected an emotion-laden movie to the right hemisphere. (If the patient tried to cheat and move the eye toward the movie, it was electronically shut off.)

When we did this experiment with case 2, the movie that her right hemisphere saw was about a vicious man pushing another off a balcony and then throwing a firebomb on top of him. The movie then showed other men trying to put the fire out. When V. P. was first tested on this problem, she could not access speech from her right hemisphere. She was able to speak only out of her left brain. When asked what she had seen, her left brain (the half brain that had not actually seen the movie) replied, ''I don't really know what I saw. I think just a white flash.'' When I asked, ''Were there people in it?'' case 2, replied, ''I don't think so. Maybe just some trees, red trees like in the fall.'' I asked, ''Did it make you feel any emotion?'' and V. P. answered, ''I don't really know why, but I'm kind of scared. I feel jumpy. I think maybe I don't like this room, or maybe it's you; you're getting me nervous.'' She turned to one of the research assistants and said, ''I know I like Dr. Gazzaniga, but right now I'm scared of him for some reason.''

This kind of effect is common to all of us. A mental system that is operating outside the conscious realm of the left hemisphere's interpreter generates a mood that alters the general physiology of the brain. Because the alteration in brain physiology is general, the interpreter is able to note the mood and immediately attributes some cause to it. This is a powerful mechanism; once clearly seen, it makes one wonder how often we are victims of spurious emotional/cognitive correlations.

Our recent investigations have looked further at the properties of the interpreter and how it influences mental skills. For example, there are

hemisphere-specific changes in the accuracy of memory processes. Specifically, the predilection of the left hemisphere to interpret events has an impact on the accuracy of memory. When subjects are presented with pictures representing common events (e.g., getting up in the morning or making cookies) and several hours later asked to say if pictures in another series appeared in the first, both hemispheres are equally accurate in recognizing the previously viewed pictures and rejecting the unrelated ones. Only the right hemisphere, however, correctly rejects pictures in the second set that were not previously viewed but were related to pictures previously viewed. The left hemisphere incorrectly "recalls" significantly more of these related pictures as having occurred in the first set, presumably because they fit into the schema it has constructed. This finding is consistent with the hypothesis that a left hemisphere interpreter constructs theories to assimilate perceived information into a comprehensible whole. In doing so, however, the process of elaborating (story making) has a deleterious effect on the accuracy of perceptual recognition. This result has been shown with verbal as well as visual material.

A recent example of the interpreter can be found in studies of case 3, a split-brain patient who can speak out of his right hemisphere as well as his left. His naming of stimuli in the left field seems to be increasing at a rapid rate. Although there is no convincing evidence of any genuine visual transfer between the hemispheres, during trials when J. W. was certain of the name of the stimulus, he maintained that he saw it well. On trials when he was not certain of the name of the stimulus, he maintained that he did not see it well. This is consistent with the view that the left hemisphere interpreter actively constructs a mental portrait of past experience, even though that experience did not directly occur in that hemisphere. This experience was probably caused by the left hemisphere's interpreter giving meaning to right hemisphere spoken responses, possibly by activating the left hemisphere mental imagery systems.

The left hemisphere's capacity for continual interpretation may mean that it is always looking for order and reason, even where there is none. This came out dramatically in a study by George Wolford and me. On a simple test that requires one to guess if a light is going to appear on the top or the bottom of a computer screen, we humans perform in an inventive way. The experiment manipulates the stimulus to appear on the top 80% of the time. While it quickly becomes evident that the top button is being illuminated more often, we keep trying to figure out the whole sequence—and deeply believe that we can. We persist even if, by adopting this strategy, we are rewarded only 68% of the time (whereas if we guessed "top" repeatedly, by rote, we would be rewarded 80% of the time). Rats and other animals are more likely to learn to maximize their score by pressing only the top button! Our right hemisphere behaves more like the rats. It does not try to interpret its experience to find the deeper meaning; it lives only in the thin moment of the present. But when the left brain is asked to explain why it is attempting to psych out the whole sequence, it always comes up with a theory, however spurious.

II. ANOSAGNOSIA AND PARAMNESIA

Neurology yields weird examples of how the interpreter can work, and understanding the interpreter increases our insight into some bizarre syndromes. Take, for example, a malady called "anosagnosia," in which a person denies awareness of a problem he has. People who suffer from right parietal lesions that render them hemiplegic and blind on their left side frequently deny that they have any problem. The left half of their body, they insist, is simply not theirs. They see their paralyzed left hand but maintain that it has nothing to do with them. How could this be?

Consider what may happen as a result of a lesion in a person's optic tract. If the lesion is in a nerve that carries information about vision to the visual cortex, the damaged nerve ceases to carry that information; the patient complains that he is blind in part of his visual field. For example, such a patient might have a huge blind spot to the left of the center of his visual field. He rightly complains. If another patient, however, has a lesion not in the optic tract but in the visual cortex, creating a blind spot of the same size and in the same place, he does not complain at all. The reason is that the cortical lesion is in the place in his brain that represents that exact part of the visual world, the place that ordinarily would ask, "What is going on to the left of visual center?" In the case of the lesion on the optic nerve, this brain area was functioning; when it could not get any information from the nerve, it put up a squawk—something is wrong. When that same brain area is itself lesioned, the patient's brain no longer cares about what is going on in that part of the visual field; there is no squawk at all. The patient with the central lesion does not have a complaint because the part of the brain that might complain has been incapacitated, and no other can take over.

As we move farther into the brain's processing centers, we see the same pattern, but now the problem is with the interpretive function. The parietal cortex is where the brain represents how an arm is functioning, constantly seeking information on the arm's whereabouts, its position in three-dimensional space. The parietal cortex monitors the arm's existence in relation to everything else. If there is a lesion to sensory nerves that bring information to the brain about where the arm is, what is in its hand, or whether it is in pain or feels hot or cold, the brain communicates that something is wrong: "I am not getting input." But if the lesion is in the parietal cortex, that monitoring function is gone and no squawk is raised, though the squawker is damaged.

Now let us consider our case of anosagnosia, and the disowned left hand. A patient with a right parietal lesion suffers damage to the area that represents the left half of the body. The brain area cannot feel the state of the left hand. When a neurologist holds a patient's left hand up to the patient's face, the patient gives a reasonable response: "That's not my hand, pal." The interpreter, which is intact and working, cannot get news from the parietal lobe since the flow of information has been disrupted by the lesion. For the interpreter, the left hand simply does not exist anymore, just as seeing behind the head is not something the interpreter is supposed to worry about. It is true, then, that the hand held in front of him cannot be his. What is the mystery?

An even more fascinating syndrome is called "reduplicative paramnesia." In one such case studied by the author, the patient was a women who, although she was being examined in my office at New York Hospital, claimed we were in her home in Freeport, Maine. The standard interpretation of this syndrome is that the patient has made a duplicate copy of a place (or person) and insists that there are two.

This woman was intelligent; before the interview she was biding her time reading the *New York Times*. I started with the "so, where are you?" question. "I am in Freeport, Maine. I know you don't believe it. Dr. Posner told me this morning when he came to see me that I was in Memorial Sloan–Kettering Hospital and that when the residents come on rounds to say that to them. Well, that is fine, but I know I am in my house on Main Street in Freeport, Maine!" I asked, "Well, if you are in Freeport and in your house, how come there are elevators outside the door here?" The grand lady peered at me and calmly responded, "Doctor, do you know how much it cost me to have those put in?"

This patient has a perfectly fine interpreter working away trying to make sense of what she knows and feels and does. Because of her lesion, the part of the brain that represents locality is overactive and sending out an erroneous message about her location. The interpreter is only as good as the information it receives, and in this instance it is getting a wacky piece of information. Yet the interpreter still has to field questions and make sense of other incoming information—information that to the interpreter is self-evident. The result? It creates a lot of imaginative stories.

III. CREATING OUR AUTOBIOGRAPHY

The interpreter's talents can be viewed on a larger canvas. I began this article by observing our deep belief that we can attain not only a neuroscience of consciousness but also a neuroscience of human consciousness. It is as if something wonderfully new and complex happens as the brain enlarges to its full human form. Whatever happens (and I think it is the emergence of the interpreter module), it triggers our capacity for self-reflection and all that goes with it. How do we account for this?

I would like to make a simple, three-step suggestion. First, focus on what we mean when we talk about "conscious experience." I believe this is merely the awareness we have of our capacities as a species—awareness not of the capacities themselves but of our experience of exercising them and our feelings about them. The brain is clearly not a general-purpose computing device; it is a collection of circuits devoted to these specific capacities. This is true for all brains, but what is amazing about the human brain is the sheer number of capacities. We have more than the chimp, which has more than the monkey, which has more than the cat, which runs circles around the rat. Because we have so many specialized systems, and because they may sometimes operate in ways that are difficult to assign to a given system or group of them, it may seem as though our brains have a single, general computing device. But they do not. Step one is to recognize that we are a collection of adaptive brain systems and, further, to recognize the distinction between a species' capacities and how it experiences them.

Now consider step two. Can there be any doubt that a rat at the moment of copulation is as sensorially fulfilled as a human? Of course it is. Do you think a cat does not enjoy a good piece of cod? Of course it does.

Or a monkey does not enjoy a spectacular swing? Again, it has to be true. Each species is aware of its special capacities. So what is human consciousness? It is awareness of the very same kind, except that we can be aware of so much more, so many wonderful things. A circuit, perhaps a single system or one duplicated again and again, is associated with each brain capacity. The more systems a brain possesses, the greater our awareness of capacities.

Think of the variations in capacity within our own species; they are not unlike the vast differences between species. Years of split-brain research have shown that the left hemisphere has many more mental capacities than the right. The left is capable of logical feats that the right cannot manage. Even with both our hemispheres, however, the limits to human capacity are everywhere in the population. No one need be offended to realize that some people with normal intelligence can understand Ohm's law, while others, such as yours truly, are clueless about hundreds of mathematical concepts. I do not understand them and never will; the circuits that would enable me to understand them are not in my brain.

When we realize that specialized brain circuits arose through natural selection, we understand that the brain is not a unified neural net that supports a general problem-solving device. If we accept this, we can concentrate on the possibility that smaller, more manageable circuits produce awareness of a species' capacities. By contrast, holding fast to the notion of a unified neural net forces us to try to understand human conscious experience by figuring out the interactions of billions of neurons. That task is hopeless. My scheme is not.

Hence step three. The same split-brain research that exposed startling differences between the two hemispheres revealed as well that the human left hemisphere harbors our interpreter. Its job is to interpret our responses—cognitive or emotional—to what we encounter in our environment. The interpreter sustains a running narrative of our actions, emotions, thoughts, and dreams. The interpreter is the glue that keeps our story unified and creates our sense of being a coherent, rational agent. To our bag of individual instincts, it brings theories about our life.

These narratives of our past behavior seep into our awareness; they give us an autobiography. Insertion of an interpreter into an otherwise functioning brain creates many by-products. A device that begins by asking how one thing relates to another, a device that asks about an infinite number of things, in fact, and that can get productive answers to its questions, cannot help giving birth to the concept of self. Surely one question the device would ask is "Who is solving all these problems?" "Let's call it me"—and away it goes! A device with rules for figuring out how one thing relates to another will quickly be reinforced for having that capacity, just as an ant's solving where to have its evening meal reinforces the ant's food-seeking devices. In other words, once mutational events in the history of our species had brought the interpreter into existence, there would be no getting rid of it.

Our brains are automatic because physical tissue carries out what we do. How could it be otherwise? Our brains are operating before our conceptual self knows it. But the conceptual self emerges and grows until it is able to find interesting—but not disheartening—the biological fact that our brain does things before we are consciously aware of them. The interpretation of things that we encounter has liberated us from a sense of being determined by our environment; it has created the wonderful sense that our self is in charge of our destiny. All of our everyday success at reasoning through life's data convinces us of this. And because of the interpreter within us, we can drive our automatic brains to greater accomplishment and enjoyment of life.

See Also the Following Articles

EMOTION • EVOLUTION OF THE BRAIN • UNCONSCIOUS, THE

Suggested Reading

Crick, F., and Koch, C. (1998). Consciousness and neuroscience. *Cerebral Cortex* **8**(2), 97–107.

Dennett, D. (1991). *Consciousness Explained*. Little, Brown, Boston.

Gazzaniga, M. S. (1998). *The Mind's Past*. Univ. of California Press, Berkeley.

Searle, J. (1984). *Minds, brains, and science*. Harvard University Press, Cambridge.

Tononi, G., and Edelman, G. M. (1998). Consciousness and complexity. *Science* **282,** 1856–1851.

Conversion Disorders and Somatoform Disorders

RICHARD J. BROWN and MARIA A. RON
Institute of Neurology and National Hospital for Neurology and Neurosurgery, United Kingdom

GLOSSARY

conversion Putative psychological process whereby emotional distress is converted into physical symptoms to resolve internal conflict.

conversion disorder Disorder characterized by the presence of at least one unexplained neurological symptom thought to be caused by the process of conversion.

dissociation Putative psychological process whereby normally integrated aspects of memory become separated.

dissociative disorder Disorder characterized by at least one unexplained neurological symptom thought to be caused by the process of dissociation.

factitious illness Deliberate simulation of physical symptoms and/or signs to obtain medical attention.

hysteria Label traditionally used to describe either a condition characterized by unexplained physical symptoms or a type of personality characterized by overdramatic expression (i.e., hysterical personality disorder).

malingering Deliberate simulation of physical symptoms for personal gain.

somatization Tendency to experience or express psychological distress as physical symptoms.

somatization disorder A severe, chronic form of somatoform disorder involving multiple unexplained physical symptoms across several bodily systems.

somatoform disorder A group of disorders characterized by the presence of disabling unexplained physical symptoms for which there is no medical explanation.

Individuals who report physical symptoms for which no organic explanation can be found are commonly encountered by medical practice. Such symptoms are often associated with high levels of disability, distress, and medical resource utilization and are attributed to psychological factors. This article addresses the nature of ''unexplained'' physical symptoms, how they are diagnosed and treated, and current ideas concerning the mechanisms involved in their pathogenesis.

I. NOSOLOGY

Attempts to identify the mechanisms involved in the pathogenesis of unexplained physical symptoms extend as far back as ancient Egyptian times, when the prevailing view attributed their origin to a ''hysterical'' process involving abnormal movements of the uterus within the sufferer's body (the ancient Greek *hysteron*=''uterus''). Although the so-called ''wandering

womb'' hypothesis lost favor almost 2000 years ago, the term *hysteria* continued to be used as a generic label for the occurrence of medically unexplained symptoms until 1980, when the terms *hysteria* and *hysterical* were eliminated from the third edition of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-III*) and two new superordinate categories, subsequently refined in *DSM-IV*, were created to capture the basic elements of the hysteria concept (Table I). The *somatoform disorders* category encompasses a range of complaints characterized by the symptoms of somatic illness (e.g., pain, fatigue, and nausea) in the absence of underlying physical pathology. Broadly speaking, *DSM-IV* categorizes somatoform complaints according to the nature, number, and duration of the unexplained symptoms in question. The *somatization disorder* category corresponds to the traditional conception of hysteria as a syndrome characterized by large numbers of unexplained symptoms across multiple bodily systems (also known as *Briquet's syndrome*). A *DSM-IV* diagnosis of somatization disorder (onset before the age of 30) requires a history of unexplained pain in at least four different bodily sites, at least two unexplained gastrointestinal symptoms, at least one unexplained sexual or reproductive symptom, and at least one unexplained neurological symptom. The *undifferentiated somatoform disorder* category includes less severe cases, in which unexplained symptoms have persisted for at least 6 months but are fewer in number and may be less disabling. Similar unexplained symptoms of a shorter duration are categorized as instances of *somatoform disorder not otherwise specified*. Each of the latter categories encompasses all possible physical symptoms, with the exception of unexplained neurological phenomena, which are categorized separately as *conversion disorders*. In addition, the term *pain disorder* is reserved for syndromes specifically characterized by the occurrence of persistent unexplained pain. In all cases, symptoms are classified as unexplained if adequate medical investigation has failed to identify a plausible physical cause for their occurrence; they must also cause clinically significant functional disability or distress to meet diagnostic criteria. Moreover, the symptoms should not be attributable to hypochondriasis or other forms of overt psychopathology, such as anxiety, depression, or psychosis. In addition, *DSM-IV* assumes that somatoform symptoms are not produced intentionally by the person. Rather, ''unexplained'' symptoms that are intentionally produced in order to obtain medical attention (e.g., as in Münchausen's disorder) are classified within the *factitious illness* category; symptoms that are intentionally produced for personal gain (e.g., in the context of a litigation claim or to avoid military service) are labeled as instances of *malingering*.

Table I
Classification of Somatoform and Dissociative Disorders in *DSM-IV*

Somatoform disorders	Dissociative disorders
Somatization disorder	Dissociative amnesia
Undifferentiated somatoform disorder	Dissociative fugue
Conversion disorder	Dissociative identity disorder
Pain disorder	Depersonalization disorder
Somatoform disorder not otherwise specified	Dissociative disorder not otherwise specified
Hypochondriasis	
Body dysmorphic disorder	

The *DSM-IV dissociative disorders* category includes unexplained symptoms involving an apparent disruption of consciousness, such as instances of amnesia and the alteration of personal identity (encompassing *dissociative amnesia, dissociative fugue*, and *dissociative identity disorder*). *Depersonalization disorder*, characterized by persistent and unpleasant feelings of detachment from the self or the world, is also placed within this category. Although, strictly speaking, the dissociative disorders should be viewed as one dimension of the hysteria concept, they are not typically regarded as medically unexplained symptoms and we will not consider them in detail in this article.

The nosological status of medically unexplained symptoms continues to provoke controversy. First, there is very little evidence to suggest that there are natural boundaries between the different categories of somatoform illness. Indeed, evidence suggests that the different somatoform categories should be regarded as points on a continuum of severity, with somatization disorder at the pathological extreme. Second, the persistence and stability of many cases of somatoform illness suggest that these conditions could be more appropriately viewed as a form of personality disorder rather than an acute psychiatric condition in their own right.

II. TERMINOLOGY

Despite being abandoned as a nosological entity, hysteria is still commonly used as a generic label for the

occurrence of medically unexplained symptoms. The continuing popularity of the term is viewed with regret by many. First, it implies that medically unexplained symptoms are an exclusively female phenomenon, despite the fact that somatoform illness is observed in both men and women (although much more commonly in the latter). Second, the terms *hysteria* and *hysterical* have acquired different and pejorative meanings in popular parlance that no longer apply to medically unexplained symptoms. Not surprisingly, many individuals suffering from such symptoms strongly object to the use of the hysterical label.

In addition to hysteria and the diagnostic labels provided in *DSM*, many other terms have been used in relation to medically unexplained symptoms, including *functional*, *nonorganic*, *psychosomatic*, and *psychogenic*, each of which is ambiguous or has unfortunate connotations. For the sake of neutrality and descriptive ease, we use *somatoform illness* and *unexplained medical symptoms* as labels encompassing the range of phenomena described in the somatoform disorder category; we use *unexplained neurological symptoms* as a label for those phenomena specifically included in the conversion disorder category. We use the term *somatization* to refer to the process underlying the generation of unexplained medical symptoms.

III. PRESENTATION

A. Clinical Features

Unexplained symptoms may be related to any bodily system. Indeed, every branch of medicine has its own syndrome characterized by the presence of such symptoms. Pain, fatigue, dizziness, shortness of breath, and general malaise are the most commonly observed symptoms in primary care, although gastrointestinal and sexual/reproductive symptoms are frequently found. Within the neurological domain, the most common symptoms are limb weakness, gait disturbances, abnormal movements, sensory disturbances, amnesia, and cognitive impairment. Contrary to popular belief, somatoform symptoms are subjectively real to the patient—that is, their experience is "as if" they have the symptoms of organic illness.

In some cases, a somatoform diagnosis may be suspected on the basis of the inconsistency or medical implausibility of the symptoms in question, although in other cases they may closely resemble symptoms of organic disease. Symptoms are often internally inconsistent. For example, a patient may have a preserved ability to cough (requiring intact vocal chord function) despite being unable to speak or whisper, or a patient may display weakness for some movements and not others involving the same muscle groups. Other symptoms may not fit with recognized disease patterns (e.g., asynchronous generalized tonic–clonic seizures) and may even be physiologically impossible (e.g., triplopia). Often, symptoms correspond more closely to what the patient believes about the body and illness than to actual physical processes. Indeed, the role of beliefs in determining the nature of somatoform phenomena has been widely recognized since the time of Charcot. It is well evidenced by the fact that individuals prone to the development of somatoform illness often develop symptoms that mirror those observed in others. It is also reflected in the varying prevalence of different unexplained symptoms across cultures. Burning hands or feet, for example, are more common in African and Asian cultures than in North America, reflecting local concerns about health and illness. A core set of somatoform symptoms is nevertheless observed across all cultures (e.g., pain, fatigue, pseudoseizures, and paralysis). A history of other unexplained medical symptoms is also frequently found in these patients.

B. Comorbidity

Comorbid psychopathology is extremely common in somatoform illness, with as many as 75% of these patients meeting criteria for an additional psychiatric disorder. Almost all forms of psychiatric illness are more prevalent in these patients compared to primary care patient averages, with elevated levels of depression and anxiety being particularly common. Psychiatric comorbidity studies suggest that between 40 and 80% of patients with somatoform illness meet diagnostic criteria for an affective disorder, whereas between 15 and 90% meet criteria for an anxiety disorder (particularly panic, phobic, and obsessive–compulsive disorders). Such variations in comorbidity estimates are a function of several factors, including the study sample used and somatoform illness severity. Studies conducted on psychiatric populations, for example, are likely to yield disproportionately high rates of psychiatric comorbidity compared to community, primary care, or other specialist samples. Similarly, patients meeting criteria for somatization disorder are considerably more likely to report psychiatric symptoms than are patients with fewer unexplained symptoms.

The comorbidity between somatoform illness and personality disorders is also much higher than that in patients with other psychiatric diagnoses. About 60% of patients with somatoform illness meet criteria for at least one personality disorder; of these, a significant proportion are likely to meet criteria for two or more such diagnoses. Contrary to historical belief, there is no particular association between medically unexplained symptoms and histrionic (i.e., hysterical) personality disorder. Indeed, evidence suggests that the most common personality disturbances in these patients are of the avoidant, dependent, paranoid, and obsessive–compulsive types. Similarly, there is little evidence in support of the traditional view that antisocial personality disorder is a common concomitant of unexplained medical symptoms.

Somatoform illness is also commonly found in the context of physical illness. In such cases, a somatoform diagnosis is made on the grounds that identifiable pathology is unable to account for the range of symptoms and degree of disability exhibited by the patient. In the neurological setting, as many as 60% of patients diagnosed with somatoform illness have comorbid organic diagnoses, although there is no consistent pattern to the type of neurological pathology found in these patients. In contrast, physical comorbidity is often less than 5% within the general psychiatric setting. Such findings indicate that neurological dysfunction plays an important role in the development of somatoform illness. However, many of these patients have extracerebral comorbid neurological disease, which contradicts the view that altered cerebral function provides a fertile soil for the development of conversion disorder. Rather, it is more likely that exposure to neurological illness provides a model for the development of symptoms and an environment in which illness behavior is rewarded.

IV. DIAGNOSTIC ISSUES

Diagnosing somatoform illness is important in order to offer relevant psychiatric treatment and to avoid unnecessary medical intervention. A much-quoted study by Eliot Slater in the 1960s suggested that the rate of misdiagnosis was very high and questioned the validity of hysteria as a diagnostic entity. Recently, the advent of noninvasive investigative techniques (e.g., brain imaging and video telemetry) and the better characterization of neurological disease patterns have made the exclusion of organic pathology easier and more reliable. Indeed, recent studies of patients with both acute and chronic unexplained symptoms suggest that neurological illness is rarely missed in these patients.

Currently, the gold standard for the diagnosis of somatoform illness involves a combination of careful history taking, the judicious use of investigations to rule out significant organic disease, and the assessment of psychiatric morbidity. The exclusion of organic pathology is essential, although it is often difficult to establish the right balance between the benefits and dangers of further investigations. Moreover, the presence of organic disease does not exclude the possibility of somatization; it is therefore important to assess all patients for unexplained symptoms or disability. The clinical history may reveal many features that support a positive diagnosis of somatoform illness, including the inconsistency and implausibility of symptoms, a previous history of unexplained symptoms, and psychosocial stressors preceding the development of symptoms. The successful alleviation of symptoms in response to placebo, suggestions, or psychological treatment also supports a positive somatoform diagnosis.

In some cases, symptoms are consistent and plausible and there may be few or no additional diagnostic features in evidence. For example, although *DSM-IV* explicitly requires the presence of psychosocial stressors for a firm diagnosis of both conversion and somatoform pain disorder, obvious stressors are absent in many cases. Even when present, establishing a causal link between such stressors and physical symptoms can be extremely difficult and involves considerable subjectivity. Other clinical features previously thought to be characteristic of somatoform illness have been found to have little or no diagnostic validity. For example, historically it has been thought that patients suffering from unexplained symptoms display an unusual lack of concern over the condition, so-called *la belle indifférence*. However, research indicates that patients with somatoform illness show no more indifference to their physical condition than do patients with organic illness. The traditional notion that unexplained medical symptoms are exclusive to individuals of low socioeconomic status has also been disputed.

Even in cases in which positive somatoform features are present, many physicians are reluctant to diagnose somatoform illness for fear of missing an underlying physical illness. However, in the clear absence of pathology following appropriate investigation, and particularly if positive signs are present, a diagnosis of somatoform illness can be safely made. In cases in which somatoform illness is suspected but an organic

explanation cannot be ruled out unequivocally, careful follow-up is often the best diagnostic tool.

V. EPIDEMIOLOGY

Somatoform illness is a ubiquitous phenomenon, representing one of the most common categories of illness encountered within the health care system. Precise prevalence estimates vary enormously due to the different diagnostic criteria employed across studies. In general, prevalence estimates decrease as symptoms become more severe, chronic, and greater in number. Recent evidence suggests that approximately 25% of patients attending primary care have a history of four or more unexplained symptoms. In contrast, studies have estimated the prevalence of *DSM-IV* somatization disorder (defined as a history of eight symptoms across multiple bodily sites) to be between 3 and 5% of primary care attenders. Within the neurological domain, evidence suggests that between 20 and 60% of new inpatient admissions have symptoms that cannot be fully accounted for by organic factors.

The understandable reluctance of many physicians to make formal somatoform diagnoses suggests that current data may underestimate the true prevalence of unexplained medical symptoms. In addition, many patients with diagnoses such as chronic fatigue syndrome, fibromyalgia, and noncardiac chest pain may be more appropriately viewed as having somatoform illness.

VI. RESOURCE UTILIZATION

The health service costs associated with somatoform illness are considerable and are not simply attributable to the ubiquity of these conditions. Rather, patients with somatoform illness often consume a disproportionate amount of resources compared to other service users. Indeed, it has been estimated that the health care costs of patients meeting criteria for somatization disorder are nine times the U.S. national average. The bulk of resource utilization by these patients is on diagnostic tests rather than psychiatric services. The frequency of general practitioner and emergency room visits, as well as the number and length of hospital admissions, also tends to be elevated in these patients. In general, the level of resource use increases with the severity and chronicity of unexplained symptoms. Improved physician recognition of these conditions and the limitation of needless investigations could significantly reduce the health care costs of this group of patients.

VII. MANAGEMENT AND TREATMENT

A. Presenting the Diagnosis

The way in which a diagnosis of conversion or somatoform illness is presented can have a considerable effect on the doctor–patient relationship; it may also be an important factor in whether the patient cooperates with the management plan. Coming to terms with a diagnosis of conversion or somatization is difficult for many patients, who often believe that their suffering is being dismissed as "all in the mind" or, worse, deliberately faked. Understandably, many patients find it difficult to accept that their debilitating symptoms may be caused by psychological rather than physical processes, particularly if they have no obvious psychological symptoms, such as anxiety or depression.

Where somatoform illness is suspected, considerable effort should be made to establish a strong therapeutic alliance between physician and patient while necessary physical investigations are being conducted. This prepares the groundwork for the later exploration of possible psychological factors that may be contributing to the presentation of unexplained symptoms and reduces resistance if a formal somatoform diagnosis is being considered. At all times, it is essential to explicitly acknowledge that the patient's symptoms are subjectively real and that they cause significant distress and disability. Such an approach allows the patient to feel that he or she is taken seriously, thereby reducing the likelihood of subsequent doctor shopping. At the appropriate point, the patient should be reassured that all necessary investigations have been done, that no sinister cause for their symptoms has been identified, and that they are not suffering from a serious physical illness. It should be emphasized that further physical investigation would be of no benefit, that certain physical treatments will not be prescribed, and that current medications might need to be withdrawn. Patients may then be gently introduced to the notion that psychological factors may be important in causing their symptoms. It should be emphasized that such an interpretation does not mean that their symptoms are faked or all in the mind and that they are not insane or crazy; the words hysteria and

hysterical should be avoided. It is also common to introduce the idea that their symptoms may be caused by psychological events of which they are unaware, giving everyday examples demonstrating the close link between psychological processes (e.g., anxiety) and physical symptoms (e.g., tremor and dizziness).

Many practitioners believe that it is unnecessary to label patients with a formal diagnosis of conversion or somatoform illness, preferring to describe their symptoms simply as "medically unexplained." Indeed, if it is possible to make therapeutic progress without psychological referral, the use of an explicit diagnosis may be unnecessary and potentially damaging to the doctor–patient relationship. Many patients with acute symptoms respond to simple measures, such as reassurance and physiotherapy; a greater therapeutic challenge is presented by patients with chronic symptoms, where an explicit diagnosis can help ensure patient cooperation with the management plan. The development of new symptoms, a refusal to accept reassurance, and/or demands for further physical investigation may also necessitate the use of a formal diagnosis. To many people, a diagnosis of "medically unexplained symptoms" implies that further physical investigation could eventually yield a medical explanation for their suffering—a belief that ultimately undermines the appropriate management of somatoform conditions. Indeed, without a firm diagnosis, many patients persist in the belief that "the doctor couldn't find what was wrong with me"—a belief that simply serves to intensify their perceived need for further physical investigation. In contrast, use of the somatoform or conversion label serves to legitimize the patient's complaint, firmly identifying it as a recognizable condition that is commonly encountered within medical care. Where psychological referral is indicated, it may be more appropriate to leave diagnostic labeling to the psychiatrist or psychologist.

B. General Management Principles

The general practitioner plays a central role in the management of somatoform illness, with the often difficult task of controlling the patient's access to further investigation and hospitalization. Referrals should only be made when clearly indicated by the clinical picture, paying particular attention to the signs rather than the symptoms of illness. It is important to remember, however, that individuals with somatoform disorders are no less likely to suffer from physical illness than are other patients; it should not be assumed, therefore, that any new complaints are necessarily somatoform. When further investigations are indicated, their purpose and results should be clearly explained to the patient.

Investigating and treating physicians should liaise with the general practitioner and make him or her aware of any explanation and advice given to the patient. A consistent approach should be adopted where possible, aimed at ensuring that the patient does not receive mixed messages. For this reason, the patient should be strongly discouraged from doctor shopping. One useful approach within primary practice is to schedule regular, time-limited appointments for the patient every 4–6 weeks. Such an approach allows the patient to feel that he or she is being taken seriously and that his or her physical health is being monitored; this is particularly useful for patients who are reluctant to accept a psychological interpretation of their symptoms. It is also preferable to a "return-as-needed" approach, which can encourage the development of new symptoms in order for the patient to obtain physician contact.

It is also important to screen for psychiatric problems such as anxiety and depression, which can prove difficult with this patient group. Often, the patient will deny being anxious or depressed, despite exhibiting clear physiological indicators to the contrary (e.g., sleep and appetite disturbance in depression and palpitations and dizziness in anxiety). Rather than directly asking the patient whether he or she is anxious or depressed, it is useful to question the patient about his or her psychological reaction to his or her symptoms. Such an indirect approach often elicits useful information concerning the patient's mental state while maintaining a strong therapeutic alliance.

C. Treatment

The way in which somatoform illness is treated depends largely on the nature of the individual and his or her symptoms. Where appropriate, efforts are made to resolve any psychosocial problems that may have provoked the presenting symptoms. Symptomatic treatments, such as physiotherapy, orthopedic treatment, occupational therapy, and speech therapy, are often used to reduce loss of function and disability and to prevent secondary damage. Other approaches, such as pain management techniques, are used to ameliorate the distress caused by symptoms. Where appropriate, depression and anxiety are treated using

medication. Often, the use of these and other techniques, combined with a careful management approach, is sufficient to provide the patient with lasting symptomatic relief. In some cases, however, more specific psychological treatment is necessary. In such cases, cognitive-behavioral therapy is most commonly used, typically targeting maladaptive beliefs about health and illness, behavioral patterns that maintain symptoms, and social factors that reinforce them. Cognitive-educational therapy, typically aimed at informing patients about the relationship between symptoms and psychological processes, is also sometimes used, often in a group context. Insight-oriented psychotherapy may be used in cases in which emotional issues appear to be a significant pathogenic factor. In terms of efficacy, evidence suggests that a cognitive approach is often effective in the treatment of somatoform illness; however, very few data exist on the efficacy of psychotherapy in these conditions.

In cases in which symptomatic and psychological interventions fail to provide complete resolution of the patients' symptoms, it may be necessary to adjust the therapeutic goal from curing the patients to simply caring for them. Accordingly, management should aim to contain the patients' symptoms and distress, limit their health care utilization, and prevent iatrogenic damage.

VIII. OUTCOME

The majority of studies addressing outcome in somatoform illness have concentrated either on conversion disorder or on somatization disorder, and research explicitly addressing other forms of somatoform illness is rare. It is nevertheless likely that studies investigating the outcome of conditions such as chronic fatigue syndrome, fibromyalgia, irritable bowel syndrome, and chronic pain conditions are implicitly addressing other aspects of somatoform illness. A review of the substantial literature concerning these conditions is beyond the scope of this article.

One of the earliest follow-up studies of unexplained symptoms was conducted in the mid-19th century by Paul Briquet, who studied 430 patients with a variety of symptoms over a period of 10 years. In about half of these patients, symptoms followed a course of progressive, chronic deterioration and recurrent episodes were common; in other cases, recovery occurred after 3–6 months. Briquet identified many factors associated with outcome in these patients. For example, onset during adolescence carried a poor prognosis, whereas a favorable change in social circumstances was associated with a positive outcome.

Several recent studies have provided information concerning outcome in adult patients with unexplained neurological symptoms and identified many prognostic indicators. Given the heterogeneity of the samples studied, it is not surprising that reported remission rates vary considerably; that notwithstanding, a relatively good prognosis has been found in most studies. On average, complete recovery of the index symptom is observed in about two-thirds of acute cases, whereas little or no improvement is observed in up to 20%. The chances of recovery in acute patients are increased in cases in which symptoms have been precipitated by traumatic events. Evidence suggests that between one-fourth and one-third of patients with chronic neurological symptoms receiving inpatient treatment show complete remission of the index symptom; moreover, between 20 and 60% show some improvement at follow-up. Evidence suggests that patients are most likely to recover during the period of hospital admission; after this time, recovery is much less likely. The findings of such outcome studies should be interpreted with caution, however. Most studies conducted in this domain report improvement for the index symptom only; information concerning other symptoms and subsequent referral patterns is rarely provided.

Younger patients tend to have a better prognosis than older patients; indeed, children with unexplained neurological symptoms of recent onset have a particularly favorable outcome. Other factors predicting outcome include the presence of comorbid psychopathology and personality disorders. The presence of comorbid personality disorder typically carries a poor prognosis, although there is no apparent association between outcome and any specific personality type. Conversely, the presence of comorbid anxiety and depression are associated with a good prognosis.

Outcome is broadly similar across the range of unexplained symptoms; the same prognostic factors (e.g., duration of symptoms) are relevant in all cases.

IX. PSYCHOPATHOLOGICAL MECHANISMS

Current theorizing concerning the psychopathological mechanisms of somatoform illness is dominated by the concepts of dissociation, conversion, and somatization; these models, and the evidence cited in their

support, are described here. Currently, no one model provides a completely satisfactory explanation of how it is possible for the symptoms of illness to be experienced in the absence of organic pathology. Although current models have shed light on many aspects of somatoform illness, they are conceptually underspecified and, in certain respects, not well supported by empirical research. Although it is likely that a complete account of unexplained medical symptoms will incorporate aspects of the dissociation, conversion, and somatization models, the development of such an account ultimately requires a more detailed examination of psychological and neuroscientific concepts not currently considered in this domain.

A. Dissociation and Conversion

The term *dissociation* originated in the 19th-century work of Pierre Janet, who proposed one of the earliest systematic accounts of the psychopathological mechanisms underlying somatoform phenomena. Although more than a century old, many of Janet's ideas concerning the organization of mental processes remain popular (albeit implicitly) within contemporary cognitive psychology. According to Janet, personal knowledge is represented by an integrated network of associated memories, which are accessible to consciousness through the operation of attention. This process of attentional selection is responsible for synthesizing the contents of conscious experience by which the execution of volitional actions is coordinated (Fig. 1). In this view, the synthetic functions of attention are disrupted in hysterical patients, rendering them susceptible to a breakdown in psychological integration in the face of extreme trauma. Traumatic experience initiates the separation or *dissociation* of memories from the main body of knowledge by an amnesic barrier. Although these dissociated memories or "fixed ideas" are prevented from entering consciousness by the amnesic barrier, they may be automatically activated by external events. The activation of dissociated memories in this way is responsible for the generation of hysterical symptoms (Fig. 2) that, because they are produced without conscious control, are essentially nonvolitional phenomena. According to Janet, this automatic activation of fixed ideas is a process of suggestion, with the underlying deficit in attention being akin to a state of hypnosis.

Shortly after the publication of Janet's account of hysteria, the dissociation model was extended by Josef Breuer and Sigmund Freud with the introduction of the *conversion* concept. According to Breuer and Freud, the process of dissociation occurs when the subject attempts to regulate his or her experience of negative affect by defensively suppressing (or *repressing*) the conscious recall of memories associated with personal trauma. In the conversion model, this repression of negative affect serves as the primary determining factor in the generation of hysterical symptoms rather than the dissociation of memories *per se*. According to Breuer and Freud, emotions are associated with high levels of neural energy that must be discharged if the energetic balance of the brain is to be preserved. By avoiding negative affect through the repression of traumatic memories, however, this discharge of neural energy is prevented. In order for it to occur, negative affect is transformed or "converted" into a somatic (i.e., hysterical) symptom, which allows the individual to discharge emotional energy without recalling the traumatic memories

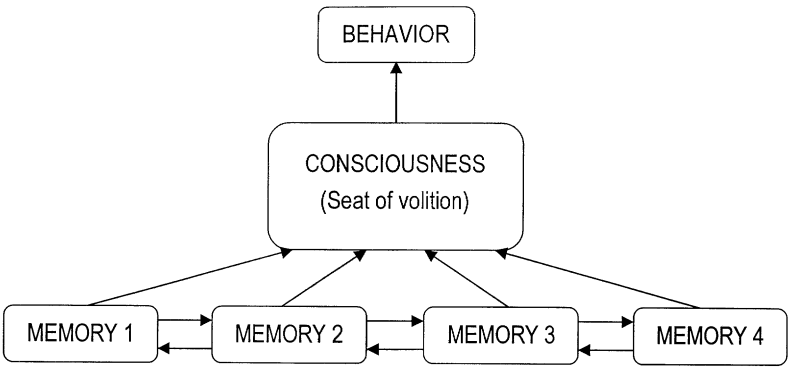


Figure 1 Organization of the cognitive system as described by Janet (1924).

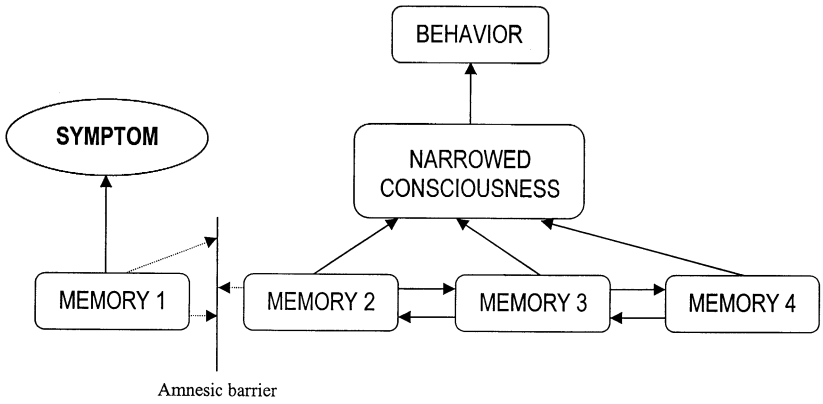


Figure 2 Generation of an unexplained symptom through the dissociation of a memory from consciousness.

giving rise to it. By this account, therefore, hysterical symptoms serve an important psychological purpose (i.e., defense), with the discharge of emotional energy representing the *primary gain* from symptoms. In this model, symptoms generated by the process of conversion correspond to sensations present at the time of the underlying trauma, or they are a symbolic representation of it.

Although more than a century old, the theoretical analyses of hysteria offered by Janet, Breuer, and Freud continue to influence nosology, theory, and clinical practice concerning medically unexplained symptoms. Several lines of evidence provide information concerning the validity of the dissociation and conversion models.

1. Attention

Many theoretical models have endorsed Janet's idea that somatoform illness involves an alteration in attention that prevents processed information from entering conscious awareness. Recent electrophysiological research indicating that conversion disorder is associated with normal early evoked potentials but a deficit in the later P300 component provides strong support for such a view. Several cognitive and psychophysiological studies have also found evidence for a diffuse attentional deficit in individuals with conversion symptoms, with patients showing decrements on tasks assessing vigilance, habituation, cognitive flexibility, set shifting, and mental transformation. However, the precise nature of the attentional deficit underlying conversion remains unclear; further research is required if it is to be described in greater detail.

2. Hypnosis

Janet's proposal that hysterical and hypnotic phenomena share similar psychological mechanisms also continues to attract support. In line with this hypothesis, many studies have shown that individuals with somatoform and conversion illness tend to exhibit high levels of suggestibility. Moreover, recent imaging studies using positron emission tomography have provided limited evidence indicating that similar neuroanatomical substrates may be involved in both conversion and hypnotic paralysis. Currently, however, the link between hypnosis and somatoform illness is still largely theoretical; further empirical evidence based on larger sample sizes is required before firm conclusions can be drawn in this regard.

3. Psychological Trauma

The view that unexplained symptoms are related to traumatic experiences, central to both the dissociation and the conversion models, has also been widely adopted. Indeed, current diagnostic criteria require that clear psychosocial precipitants be present for a diagnosis of conversion disorder or somatoform pain disorder. Although there is substantial evidence to suggest that many instances of somatoform illness are either preceded by psychosocial precipitants or associated with significant early trauma, it is clear that traumatic precipitants are absent in many cases. Moreover, many traumatized individuals with somatoform illness have not experienced amnesia for their trauma, and symptom resolution is not guaranteed by the recovery of previously forgotten traumatic memories. As such, trauma cannot play the primary

pathogenic role in the generation of medically unexplained symptoms suggested by Janet, Breuer, and Freud, although it is clearly relevant in many instances.

4. Alexithymia

Although the concept of emotionally derived neural energies has long since been abandoned, the work of Freud and Breuer still remains influential. The notion that a reduction in anxiety is the primary gain associated with unexplained symptoms is largely responsible for the continuing popularity of *la belle indifférence* as a diagnostic indicator of somatoform illness. As noted previously, however, patients with somatoform illness show no more indifference to their condition than do patients with comparable physical illnesses. Moreover, it is clear that many individuals with somatoform illness show high levels of anxiety, demonstrating that any conversion process that is occurring is far from effective in controlling this emotion. However, there is evidence that somatoform illness is associated with an inability to identify and report on one's emotional states (so-called *alexithymia*). Such evidence could be interpreted as indicating that the process of conversion occurs as a means of discharging unexpressed emotion. These findings must be interpreted with caution, however, because there is evidence that existing measures of alexithymia are confounded by psychopathology in general.

5. Symptom Laterality

The Freudian view that unexplained symptoms are the expression of unconscious emotional conflict has led many to conclude that the right cerebral hemisphere (traditionally viewed as one of the neural sites responsible for emotion) plays an important role in the pathogenesis of unexplained symptoms. Several studies have provided evidence suggesting that unexplained symptoms are more common on the left than the right, apparently providing support for this hypothesis. However, recent research has shown that left- and right-sided symptoms are equally common; bilateral symptoms are also frequently found.

B. Somatization

Dissociation and conversion represent the main theoretical precursors to contemporary accounts of the somatoform disorders based on the concept of *somatization*. The term somatization originated in the psychoanalytic literature of the early 19th century as a label for the hypothetical process whereby bodily dysfunction (i.e., unexplained symptoms) was generated by "unconscious neurosis." Since the 1960s, however, the work of Zbigniew Lipowski has encouraged many within the field to adopt a more descriptive usage of the term. According to this approach, somatization may be broadly defined as the tendency to experience or express psychological distress as the symptoms of physical illness. Unlike previous approaches, which attempted to identify neuropsychological processes underlying the occurrence of medically unexplained symptoms, the somatization model places explanatory emphasis on the entire biopsychosocial context surrounding the experience of physical and mental illness. In this respect, the somatization model is influenced by the concept of "illness behavior"—that is, the way in which we perceive, evaluate, and react to our physical and psychological states in relation to socially sanctioned models of health and illness (the so-called "sick role"). According to this model, individuals suffering from somatoform illness are said to display *abnormal* illness behavior—that is, a tendency to adopt the sick role that is inappropriate given the absence of identifiable physical pathology. Many features of somatoform illness have been identified as instances of abnormal illness behavior. For example, many somatoform disorder patients dispute the diagnoses they have been given, refuse to accept that they have been investigated adequately, or fail to comply with treatment. Indeed, many such patients are perceived as "difficult" by the treating physician, and the doctor–patient relationship is often less than satisfactory.

The somatization model views the development of abnormal illness behavior as a multifactorial process in which social, cultural, cognitive, perceptual, personality, and physiological factors are all implicated.

1. Sociocultural Factors

Research and theory suggest that abnormal illness behaviors may pass from generation to generation through early social learning within the familial context. Disproportionate parental concern over a child's physical symptoms, the misattribution of normal sensations to pathological causes, and inappropriate help-seeking behavior have all been linked to subsequent bodily preoccupation, which may serve as a risk factor for the development of somatoform

illness. Moreover, some studies have shown that children exposed to abnormally high amounts of family illness are more likely to experience somatoform illness as adults. Early exposure to pathology may reinforce illness behavior and provide a model for the subsequent development of symptoms. In addition, repeated exposure to illness provides misleading information concerning illness base rates, leading those that are exposed to overestimate the likelihood of certain forms of pathology.

The dynamic relationship between illness behavior and culture in general may also play an important role in the development of somatization. In most cultures, there is a stigma attached to mental illness that is not associated with physical forms of infirmity; as such, expressing emotional distress somatically offers a socially acceptable way of gaining support via the sick role. Cultural factors pertaining to the health care system are also likely to be influential in shaping the nature and occurrence of somatization. Some studies, for example, have implicated iatrogenic factors in the maintenance of many medically unexplained symptoms, often involving physical misdiagnoses that could serve to perpetuate symptoms by legitimizing the individual's somatic interpretation of their condition. Moreover, the biological emphasis of modern health care systems encourages individuals to communicate nonspecific distress somatically rather than psychologically.

2. Perceptual and Cognitive Factors

What we know and believe about bodily states is inextricably bound with our experience of physical sensations as well as our behavioral response to them. For example, some studies have suggested that the acquisition of maladaptive beliefs about health and illness may be related to the development of somatization. Thus, the somatizing individual may hold the mistaken belief that health is a state devoid of physical symptoms, and that symptoms necessarily imply the presence of disease. Recently, Arthur Barsky and colleagues argued that such beliefs lead the somatizing individual to develop a preoccupation with his or her bodily states and a tendency to misinterpret them as pathological. Once symptoms are attributed to illness, further attention may be directed toward the body, with subsequent physical events being perceived as evidence in support of a pathological interpretation. Evidence demonstrating that hypochondriasis is associated with body-focused attention provides support for this model, as do several studies indicating that self-focused attention is positively related to somatic symptom reports in nonclinical populations. Clinical experience strongly suggests that body-focused attention is also a central aspect of somatoform illness.

3. Personality Factors

It is likely that cognitive and perceptual factors precipitate and maintain somatization by maximizing the degree to which physical states are experienced as aversive. Research suggests that the tendency to experience physical and psychological states as aversive is also related to the personality dimension of negative affectivity (NA), characterized by a negative self-concept and a trait-like tendency to experience distress and dissatisfaction. Indeed, subjective symptom reports consistently correlate with NA, despite there being no correlation between NA and objective health markers. Individuals high in trait NA are hypervigilant about their bodies and have a lower threshold for perceiving physical sensations, so-called somatic amplification. Such individuals are also more likely to interpret physical symptoms as the signs of serious disease and seek medical attention accordingly.

4. Physiological Factors

Unlike previous theories, the somatization model embraces the idea that normal physical processes (e.g., the physical component of an emotional state) and minor pathological events may contribute to the development of unexplained symptoms. For example, anxiety is typically associated with increased autonomic arousal that may result in physical changes such as shaking, sweating, and tachycardia; moreover, fear-related hyperventilation can produce symptoms such as breathlessness, chest pain, and fatigue. Similarly, the sleep problems and physical inactivity often associated with depression may give rise to fatigue, pain, and the feeling that increased effort is required to execute everyday tasks. Other physical processes unrelated to emotional states may also contribute to the development of medically unexplained symptoms. For example, muscle wasting resulting from illness-related inactivity may produce fatigue that perpetuates itself by preventing the resumption of physical exercise after illness remission.

A recent model of the interaction between these different factors has been described by Lawrence Kirmayer and colleagues (Fig. 3). According to this model, illness, emotional arousal, and everyday physiological processes produce bodily sensations that

capture the individual's attention to varying degrees. These sensations may be interpreted as indicators of disease, an attribution that can serve to generate illness worry, catastrophizing, and demoralization. As a result, individual may adopt the sick role by pursuing assessment and treatment for his or her putative condition, thereby exposing himself or herself to social forces that may reinforce their illness behavior and experience. This process may be moderated by many situational and dispositional factors, including previous illness experience, environmental contingencies that reward illness behavior, the response of significant others to illness worry, and individual differences in personality, attentional set, coping behavior, and autonomic reactivity.

Despite their obvious explanatory power, there are certain problems with models based on the concept of somatization. First, such models obscure potentially important differences between the various forms of medically unexplained symptoms, such as those that are the physical concomitants of conditions such as anxiety and depression, normal physical sensations or minor pathological events that are mistakenly attributed to serious illness through hypochondriacal misinterpretation, and those characteristic of the conversion and somatoform disorders. Although there is considerable overlap between these conditions in clinical practice, these different forms of somatization can be distinguished both conceptually and empirically. Although the somatization model offers a powerful account of the medically unexplained symptoms associated with depression, anxiety, and hypochondriasis, it fails to provide an adequate account of the mechanisms underlying conversion and somatoform symptoms. Arguably, a proper understanding of somatoform illness requires a theoretical approach that is specific to this form of somatization—an approach that is inherent to dissociation and conversion models.

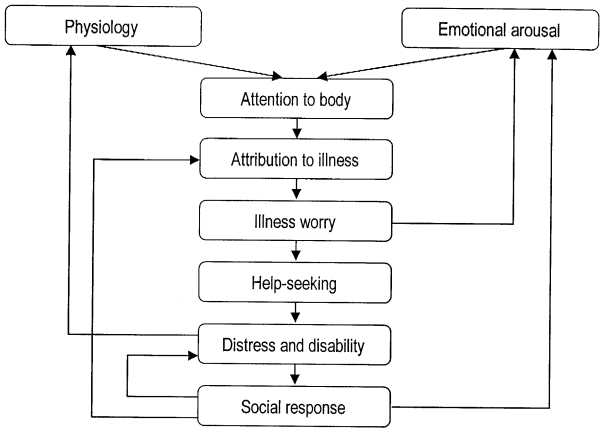


Figure 3 A multifactorial model of somatization (based on Kirmayer and Taillefer (1997)).

Second, the model of somatization described here assumes that medically unexplained symptoms are necessarily the product of physiological processes, such as the physical components of emotional states, and minor pathological events. Although such processes might play an important role in the generation of certain somatoform symptoms, it is difficult to understand how unexplained neurological (i.e., conversion) symptoms can be explained in this way.

Third, as with theories based on the concepts of conversion and dissociation, the somatization model assumes that unexplained symptoms are necessarily the expression of psychological distress. Although this may be true in many cases, particularly those associated with anxiety, depression, or hypochondriasis, it is apparent that such an assumption may be inappropriate in many other cases.

Finally, it may be misleading (and, indeed, pejorative) to identify somatoform illness as necessarily involving "abnormal" illness behavior. In our view, seeking help for subjectively compelling and debilitating symptoms is more appropriately regarded as *normal* illness behavior, regardless of whether an underlying pathophysiological basis for those symptoms can be found. Similarly, it is unclear what constitutes a "normal" illness response to repeatedly negative physical investigations despite the persistence of symptoms, particularly when disability is high (e.g., as in paralysis). As such, it may be more appropriate to reserve the concept of abnormal illness behavior for those cases in which the problem appears to involve more than just a poor doctor–patient relationship or the presentation of unexplained symptoms *per se*.

X. SUMMARY

In summary, the following points should be noted:

1. Somatoform illness is one of the most common forms of psychiatric disorder encountered within the health care system and is associated with high levels of resource utilization.
2. Almost every symptom of organic illness has an "unexplained" (i.e., somatoform or conversion) counterpart.
3. In some cases, unexplained medical symptoms have positive features (e.g., internal inconsistency) that

are indicative of a somatoform diagnosis; however, other somatoform symptoms closely resemble the symptoms of organic disease.
4. Comorbid psychopathology, particularly depression, anxiety, and personality disorders, is extremely common in patients with unexplained medical symptoms; comorbid organic pathology is also common in these patients.
5. The diagnosis of somatoform illness involves careful history taking, the exclusion of physical illness through appropriate investigation, and the assessment of psychiatric morbidity.
6. Management of somatoform illness involves containing the patient's symptoms and distress, limiting their access to investigations and hospitalizations, and preventing iatrogenic damage; careful consideration must be given to whether a formal diagnosis of somatoform illness should be used and, if so, how such a diagnosis should be presented.
7. A typical therapeutic approach involves the use of reassurance, symptomatic treatments and, where necessary, psychological intervention; the prognosis of somatoform illness is quite good, with acute symptoms being more responsive to treatment than chronic symptoms.
8. Current theorizing concerning unexplained medical symptoms is dominated by the concepts of dissociation, conversion, and somatization. Although there is limited evidence in support of these concepts, they fail to provide a complete account of the psychopathological mechanisms underlying these conditions.

See Also the Following Articles

ANXIETY • ATTENTION • DEPRESSION • NEUROPSYCHOLOGICAL ASSESSMENT • PAIN AND PSYCHOPATHOLOGY

Suggested Reading

Barsky, A. J. (1979). Patients who amplify bodily symptoms. *Ann. Internal Med.* **91,** 63–70.

Breuer, J., and Freud, S.(1893–1895/1955). Studies on hysteria. In *The Standard Edition of the Complete Psychological Works of Sigmund Freud* (J., Strachey and A., Strachey, Eds.), Vol. 2. Hogarth Press/Institute of Psycho-Analysis, London, (English translation, 1955).

Brown, R. J., and Trimble, M. R. (2000). Dissociative psychopathology, non-epileptic seizures and neurology. *J. Neurol. Neurosurg. Psychiatr.* **69,** 285–291.

Crimlisk, H. L., and Ron, M. A. (1999). Conversion hysteria: History, diagnostic issues and clinical practice. *Cognitive Neuropsychiatr.* **4,** 165–180.

Crimlisk, H. L., Bhatia, K., Cope, H., David, A., Marsden, C. D., and Ron, M. A. (1998). Slater revisited: 6 year follow up of patients with medically unexplained motor symptoms. *Br. Med. J.* **316,** 582–586.

Iezzi, A., and Adams, H. E. (1993). Somatoform and factitious disorders. In *Comprehensive Handbook of Psychopathology* (P. B. Sutker and H. E. Adams, Eds.), pp. 167–201. Plenum, New York.

Janet, P. (1924). *The Major Symptoms of Hysteria*, 2nd ed. Macmillan, New York.

Kirmayer, L. J., and Robbins, J. M. (Eds.) (1991). *Current Concepts of Somatization: Research and Clinical Perspectives*. American Psychiatric Press, Washington, DC.

Kirmayer, L. J., and Taillefer, S. (1997). Somatoform disorders. In *Adult Psychopathology and Diagnosis* (S. M. Turner and M. Hersen, Eds.), 3rd ed. pp. 333–383. Wiley, New York.

Lipowski, Z. J. (1988). Somatization: The concept and its clinical application. *Am. J. Psychiatr.* **145,** 1358–1368.

Ludwig, A. M. (1972). Hysteria: A neurobiological theory. *Arch. Gen. Psychiatr.* **27,** 771–777.

Martin, R. L., and Yutzy, S. H. (1999). Somatoform disorders. In *The American Psychiatric Press Textbook of Psychiatry* (R. E. Hales, S. C. Yudofsky, and J. A. Talbot, Eds.), 3rd ed. pp. 663–694. American Psychiatric Press, Washington, DC.

Mayou, R., Bass, C., and Sharpe, M. (Eds.) (1995). *Treatment of Functional Somatic Syndromes*. Oxford Univ. Press, Oxford.

Pilowsky, I. (1978). A general classification of abnormal illness behaviours. *Br. J. Med. Psychol.* **51,** 131–137.

Ron, M. (1994). Somatisation in neurological practice. *J. Neurol. Neurosurg. Psychiatr.* **57,** 1161–1164.

Corpus Callosum

KATHLEEN BAYNES
University of California, Davis

GLOSSARY

agenesis of the corpus callosum A chronic condition in which the corpus callosum fails to develop. The condition has been increasingly recognized due to the widespread use of imaging techniques that reveal the distinctive ventricular pattern that occurs when this major fiber tract fails to develop. Persons with this condition may be cognitively normal but show increased interhemispheric transfer times on a variety of tests.

alien (anarchic) hand sign A condition in which one hand performs complex motor acts outside of the person's conscious control. It can occur in either the dominant or nondominant hand. One form of anarchic hand is thought to occur following a callosal lesion. The other type occurs after disruption of the motor system due to medial frontal lobe damage. Both types are usually intermittent and transitory.

anterior commissure A fiber bundle that connects the two hemispheres. It is inferior to the corpus callosum, near the rostrum. It is thought to connect the anterior temporal lobes, but in humans it may carry fibers from more widely distributed areas.

body of the corpus callosum The central fibers of the corpus callosum. The fibers lie between the genu and the isthmus.

callosotomy The surgical section of the corpus callosum. It is usually accomplished in two stages. The anterior two-thirds of the callosum is sectioned first, followed by the section of the splenium if satisfactory seizure control has not been achieved. In humans, this procedure is only used in the treatment of intractable epilepsy.

commissurotomy The surgical section of the corpus callosum, the anterior commissure, and the hippocampal commissure. This more radical procedure was introduced when the first callosotomies appeared to be ineffective. It remains an alternative to the callosotomy.

contralateral This term refers to the hand, visual field, etc. that is on the opposite side from the structure under discussion. The left hemisphere controls the right or contralateral hand.

fiber tract A group of axons that follows the same path through the brain.

genu The anterior fibers of the corpus callosum that appear to bend like a knee in sagittal section before becoming the more horizontal body of the corpus callosum.

hippocampal commissure One of the fiber tracts inferior and dorsal to the callosum that is usually cut during the posterior section of the callosum. It carries fibers from the hippocampus, a structure important in forming new memories.

homonomous hemianopsia Cortical blindness in one complete visual field. It commonly results from destruction of the right or left occipital lobe, causing blindness in the contralateral visual field.

ipsilateral This term refers to the hand, visual field, etc. that is on the same side as the structure under discussion. The left hemisphere does not control the left or ipsilateral fingers well, although it can participate in ipsilateral limb movement.

isthmus The portion of the corpus callosum just anterior to the splenium.

laterality This term refers to the tendency for one hemisphere to perform a particular cognitive or motor task better than the other.

rostrum The most anterior portion of the corpus callosum.

splenium The most posterior portion of the corpus callosum. It conveys visual information between the hemispheres.

split-brain A term used to refer to a person who has undergone either commissurotomy or callosotomy for the treatment of intractable epilepsy.

The corpus callosum is one of the most anatomically prominent structures in the human brain. It is composed of approximately 200 million fibers that course across the brain's midline to connect the two cerebral hemispheres. Despite its structural prominence, which led early investigators to believe it played an important role in cognition and behavior, identifying the role of callosal connections in modifying behavior has proven difficult. This article discusses the anatomy of the callosum, the changing view of its role in behavior, the effects of surgical and other lesions on different aspects of behavior, studies of callosal transfer times, and theories of the role of the callosum in human laterality, cognition, and consciousness.

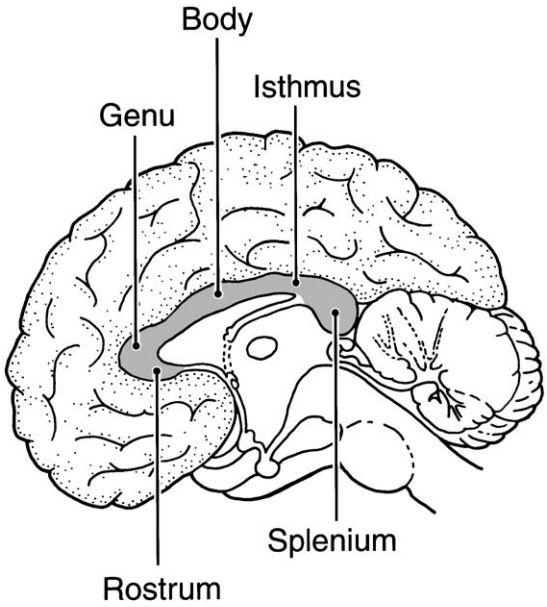


Figure 1 Sagittal section of the brain with major divisions of the corpus callosum labeled.

I. ANATOMY

The corpus callosum is composed of millions of nerve fibers that connect the two halves of the brain. These fibers traveling together from one cerebral hemisphere to the other form a brain structure easily visible to the beginning student of neuroanatomy. Figure 1 shows a saggital section of the brain, which is a slice that runs from front to back on a vertical plane. This slice passes through the midline. The large curved structure in the middle of the brain is the corpus callosum.

Although there is considerable variability in the size and shape of the corpus callosum in humans, it is known that it contains approximately 200 million fibers that carry neural signals from one side of the brain to the other. Although most of these fibers are thought to be excitatory, their effect may be inhibitory due to the activity of inhibitory interneurons. Approximately half of these fibers are small and unmyelinated. These fibers transmit information more slowly than the larger myelinated axons, which are capable of extremely rapid transmission of information. Some of the fibers connect to similar areas in the right and left hemispheres; other fibers go to areas in the contralateral hemisphere that are analogous to areas that have dense ipsilateral connections with their area of origin. A final group of fibers are diffusely connected to the contralateral hemisphere. If there is an analogy in the human brain with findings reported in the animal literature, some neurons may cross the callosum and descend to the subcortical structures before terminating.

Fibers from different areas of the cortex cross the callosum in discrete locations. To understand this phenomenon, it is necessary to examine the divisions of the corpus callosum (Fig. 1). The most anterior part of the callosum is the rostrum. Just behind the rostrum, the callosum bends to form the genu (or knee) and then extends posteriorly in the body. The body constricts slightly to become the isthmus and finally terminates in the slightly bulbous splenium. There is a great deal of individual variation in the shape and thickness of the different parts of the callosum. There have been attempts to understand differences in the anatomy of the right and left hemispheres and in the lateralization of function with regard to the morphology of the corpus callosum. As behavioral claims regarding the contributions of different areas of the callosum become more precise, it will become more crucial to be able to carefully define the areas of the callosum in which particular fibers cross. Currently, however, the areas are conventionally defined as proportions of the length of the callosum. There seems to be minimal difference whether the curvature of the callosum is taken into account or simply maximum anterior and posterior extension is used for the partition. The anterior one-fourth of the callosum is considered the genu. The rostral body begins directly behind the genu, extending back to include the anterior one-third of the callosum. The center one-third of the callosum is split into two equal sections, the anterior and posterior midbody. The isthmus extends from the posterior one-third to the posterior one-fifth of the callosum. Finally, the most posterior one-fifth is considered the splenium.

Because these definitions are arbitrary, they may differ in detail from investigator to investigator, depending on the investigator's emphasis.

Generally, however, it is believed that the fibers that pass through the different regions of the corpus callosum represent the anterior-to-posterior organization of the cerebral cortex (Fig. 2). Deepak Pandya and Benjamin Seltzer demonstrated an anterior-to-posterior organization of the callosal fibers of passage in the rhesus monkey. Prefrontal cortex and premotor cortex axons cross in the rostrum and genu. Motor and somatosensory axons are found primarily in the body. Auditory and association areas are represented in the isthmus and visual areas in the splenium. Stephen Lomber and colleagues document a similar arrangement in the corpus callosum of the cat, although auditory fibers are found throughout the body and dorsal splenium, largely overlapping with limbic and visual fibers except in the very ventral sections of the splenium, which are entirely visual. The same anterior-to-posterior organization has been demonstrated by Marie Christine de Lacoste and colleagues in the human brain. That is, the right and left prefrontal, orbital–frontal, and frontal language areas are connected in the rostrum and the genu. The anterior part of the body carries fibers from the sensory and motor regions that abut the central sulcus. It appears that fibers from the area around the Sylvian fissure, associated with language function in the left hemisphere, cross in the posterior part of the body or isthmus. Finally, the splenium carries fibers connecting the visual areas of the occipital lobe.

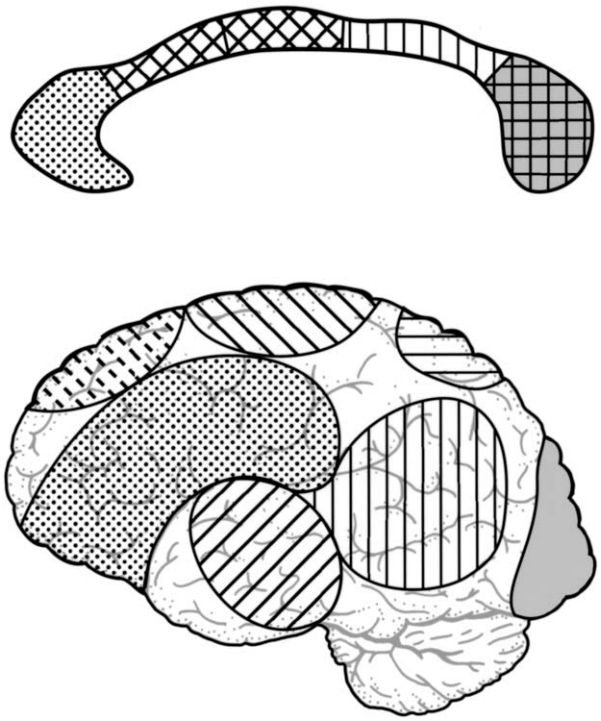


Figure 2 Areas of the cortex where fibers of the corpus callosum originate are coded to match the section of the corpus callosum where those fibers cross to the other hemisphere (adapted from M. C. DeLacoste, J. B. Kirkpatrick, and A. D. Ross, Topography of the human corpus callosum, *J. Neuropathol. Exp. Neurol.* **44,** 578–591, 1985).

Although it is now known that there are distinctive behavioral deficits associated with lesions to some regions of the callosum, contributions of the corpus callosum to behavior have not been easy to observe and understand. Some scientists have argued that evolutionary evidence suggests that the corpus callosum plays a distinctive role in human behavior. The callosum is not present in some primitive marsupials and takes on greatly increased prominence in the human species. The prominence of the corpus callosum appears to increase as lateralization increases, but this relationship is currently very speculative. In fact, James Rilling and Thomas Insel argue that, based on studies of 11 primate species, only the size of the splenium increases in proportion to increasing brain size, whereas the size of the corpus callosum as well as that of the anterior commissure are actually reduced relative to the size of other structures in higher primates. This observation suggests that any changes in the callosum related to human evolution are specific to particular areas of the callosum such as the splenium.

Developmentally, the callosum is a small structure in the neonatal brain that increases in size and prominence as the fibers myelinate (Fig. 3). At approximately 7 weeks of gestation, the lamina terminalis begins to thicken forming the commissural plate. By approximately 9 weeks cells begin to form the massa commissurelis, which supports the growth of the first commissural fibers at approximately 12 weeks. The basic structure of the callosum is present by 20 weeks and it continues to thicken and develop until birth, with development of the genu and body occurring before that of the splenium and rostrum. The myelinization of the corpus callosum is not believed to be complete until puberty, and recent evidence suggests that it reaches its maximum size at approximately age 25. It has been argued that the distinctive evolutionary and developmental patterns suggest a special role for the callosum in human behavior. Sandra Witelson has argued strongly that the anatomy of the callosum may be key to

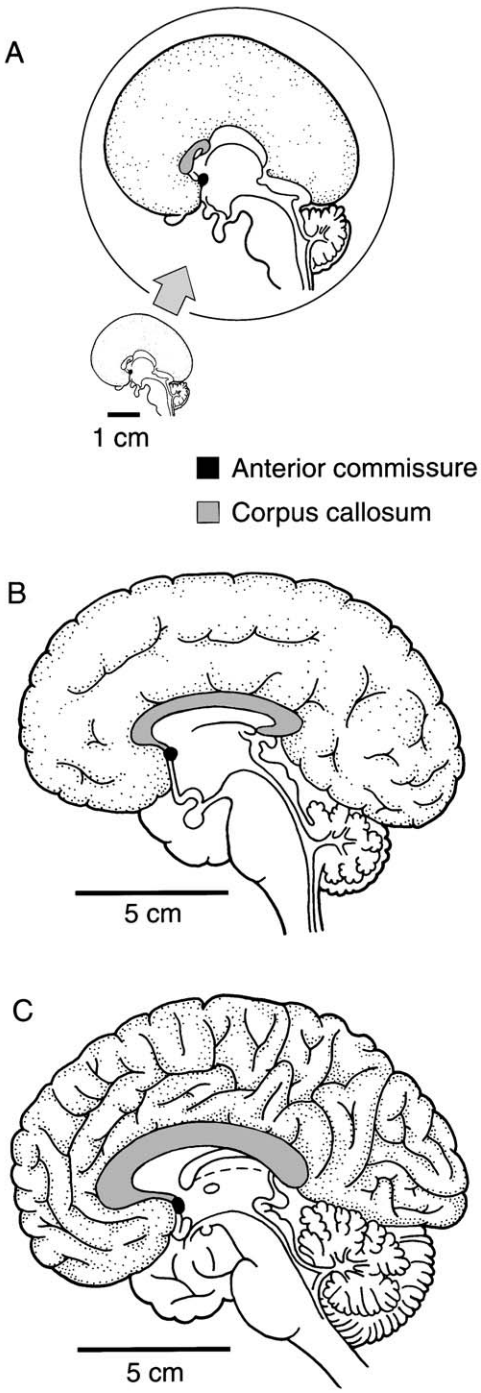


Figure 3 Relative size of the corpus callosum is seen in at 16 weeks (A), 40 weeks (B), and at adulthood (C) (adapted from S. P. Springer and G. Deutsch, *Left Brain, Right Brain: Perspectives from Cognitive Neuroscience*. p. 260. Freeman, New York, 1997).

understanding developmental issues of lateralization and hemispheric specialization. Deliniating that role has proven difficult.

II. HISTORY

The previously mentioned observations, specifically that of the sheer size of the fiber tract and its position as a unique midline structure, were important in suggesting to early scientists that it would be crucial to understand the callosum's role in behavior to fully understand the organization of the brain. In the early literature, it even competed with the pineal gland as a potential seat of the soul. Responding to an increasing belief that the corpus callosum played a major role in the integration of brain activities, Thomas Huxley called the corpus callosum "the greatest leap forward anywhere made by Nature in her brain work."

However, in the 20th century, studies of the callosum suggested that it was of much less interest than its prominent anatomy might indicate. Work in Pavlov's laboratory by Konstantin Bykov and Aleksei Speransky demonstrated that the transfer of conditioned learning from one side of the body to the other in dogs was abolished after section of the corpus callosum. However, this interesting work failed to gain significant recognition and was overshadowed by other animal work that did not document an important role for the corpus callosum in behavior. In the 1940s, William Van Wagenen and Robert Herren resected the callosum in a series of patients as treatment for epilepsy. These patients were extensively tested for psychological and behavioral changes. Although some patients improved, there did not appear to be consistent benefits in control of epilepsy nor consistent changes in their cognitive ability as a result of the severing of this enormous band of fibers. This series of surgeries did not continue and the interest in callosal function waned. Psychologist Karl Lashley was so unimpressed with the effect of severing the corpus callosum that he concluded that the corpus callosum played only a minimal role in psychological function.

It was the remarkable developments in animal research in the 1950s that paved the way for a greater understanding of the function of the corpus callosum in humans. Roger Sperry and Ronald Meyers, after splitting both the corpus callosum and the optic chiasm in the cat, showed that if one eye of the cat was covered while it learned a task, when that eye was covered and the other eye uncovered the animal acted as if it had no knowledge of the task. When exposed to the same learning trials, it had to learn again from scratch with no significant benefit from the prior exposure. In this case, the hemisphere of the brain that had not been able to see the task being learned had to

complete the task and showed no evidence of having learned it previously. The important principle that their approach uncovered was that care had to be taken to introduce information only to one hemisphere of the brain and to test that hemisphere independently as well. For the first time, the discovery that learning of a task could occur independently in the separate hemispheres of the brain was widely appreciated.

In the 1960s, Norman Geschwind published his influential review of disconnection syndromes. Disconnection syndromes are patterns of behavior that occur when the fibers that connect areas of the brain responsible for different aspects of a task are damaged, preventing communication needed to complete complex behaviors. Perhaps the most striking of these syndromes is alexia without agraphia—that is, the ability to write in the absence of the ability to read. Patients with this striking disorder can write words and sentences spontaneously or to dictation, but they are unable to read what they have written. All reading is severely impaired in this group due to a disruption of the fibers that carry visual information to the area of the left hemisphere that decodes words into sound and meaning. However, that decoding area is intact and, hence, the ability to write and spell is intact. The existence of such syndromes suggested that a complete section of the corpus callosum should produce other dramatic behavioral changes.

Geshwind's comprehensive review of the animal and human literature emphasized the importance of the corpus callosum and other fiber tracts in producing complex behaviors. The corpus callosum was obviously the largest of the known fiber tracts and Geshwind's review suggested that it played an important role in transmitting information between the two hemispheres of the brain, despite the discouraging results from the series of split-brain surgeries performed by Van Wagenen. Pursuing this idea, Geschwind and colleague Edith Kaplan were among the first to observe disconnection symptoms in a patient with a naturally occurring lesion of the corpus callosum.

Meanwhile, physicians continued to consider a possible role for the corpus callosum in the spread of epileptic seizures. After a careful review of the Van Wagenen series and with the new observations of Geschwind in mind, Philip Vogel and Joseph Bogen decided once again to attempt to resect the corpus callosum as a treatment for epilepsy, but with some differences in procedure. The corpus callosum is not the only fiber tract that connects the left and right hemispheres. Many smaller tracts, the anterior commissure in particular, that were not severed in the Van Wagenen series also cross the midline and could serve as an alternative route to spread seizure activity. Bogen and Vogel suspected that some of the original surgeries had not alleviated the epilepsy because these alternate tracts served to spread the seizure activity in the absence of the callosum. Hence, they initiated a new series of surgeries that resected not just the corpus callosum but also additional tracts, including the anterior commissure.

Guided by the exciting observations made in the Sperry laboratory with cats and by the renewed interest in disconnection syndromes, plans were made for a more intensive study of the disconnection effects in humans after commissurotomy. A team consisting of Joseph Bogen, Roger Sperry, and Michael Gazzaniga assembled the tasks and methods of response they thought would be necessary to demonstrate the effect of split-brain surgery in humans. To understand their approach and observations, some knowledge of functional anatomy is necessary.

The most crucial problem the researchers faced was that the left hemisphere in most people is the only one that controls speech. Hence, if only a verbal response to a task is accepted, the mute right hemisphere will not be able to respond and may appear unable to do simple tasks. They realized that if right hemisphere skills were to be probed, a manual response was required. Tasks had to be designed to allow a button press response or some other tactile response.

Second, because there is some ipsilateral control of motor output, knowledge being tested had to somehow be isolated in one hemisphere or the other to prevent interference between the two hemispheres in controlling a tactile response. Because of the unique anatomy of the visual system in humans, it was possible to isolate visual words and pictures to one hemisphere but only under very special conditions. Figure 4 shows the organization of the visual system. Information from the right side of space first reaches the cortex in the back of the left side of the brain or occipital lobe and vice versa. Unlike in the cat in the earlier Sperry experiments, covering one eye does not isolate information to one hemisphere, so more elaborate procedures are necessary. In most people, the information about the two sides of space is quickly woven together into a seamless visual world through the neural transmission across the callosum. Once these fibers are severed, information in one visual field is isolated in the contralateral hemisphere. This only holds true, however, if the eyes remain focussed on a single central location or fixation point. In everyday life, though, our eyes are always in motion and are

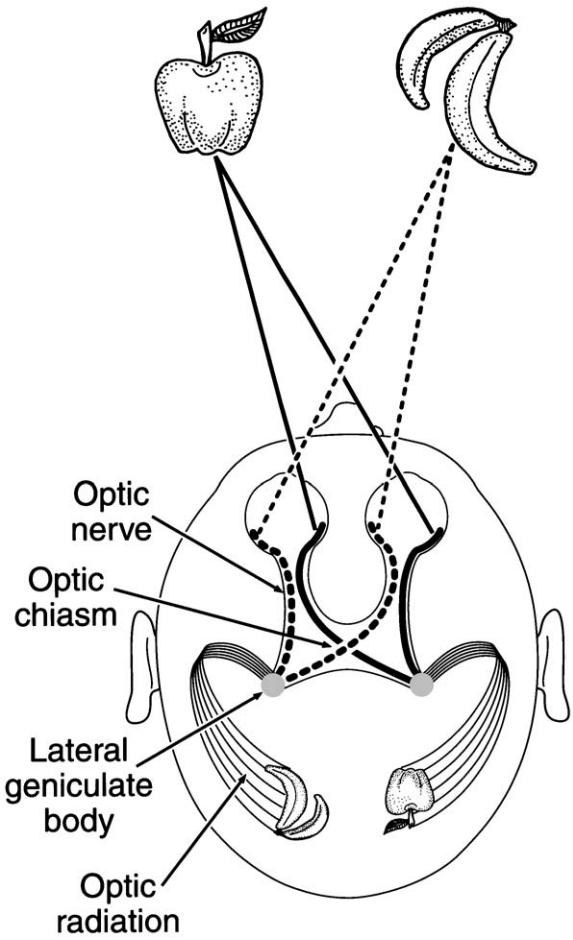


Figure 4 After the corpus callosum is cut, the unique anatomy of the human visual system displays material presented on one side of space only to the contralateral hemisphere. When the callosum is intact, this visual information is shared between the hemispheres via the fibers of the splenium.

drawn quickly to changes in the visual environment. Hence, to an investigator who wishes to have visual information presented in the left visual field seen only in the right hemisphere, this reflexive orienting movement presents a challenge. To prevent eye motion from interfering with lateralization, words or pictures had to be presented for 150 msec or less so that there was not time for the eyes to move from the fixation point to the stimulus display. The early devise used to ensure brief presentations was known as a tachistoscope, but today many investigators control the length of the stimulus display with a computer.

Although the visual system provides clean lateralization of information in a split-brain subject, the auditory system does not. Fibers from the auditory system of one ear reach both hemispheres of the brain, with about 60% of the pathway arriving in the contralateral hemisphere and 40% in the ipsilateral hemisphere. In order to test auditory comprehension separately in the left and right hemispheres, an auditory stimulus had to be compared with some completely lateralized visual stimulus.

One other neural pathway, although it has some ipsilateral representation, provides relative isolation of information to the contralateral hemisphere. Somatosensory information from the hands is predominently transferred to the contralateral hemisphere. Hence, when real objects are placed in the hands and palpated to aid in identification, the somatosensory information about the object remains in one hemisphere. The ipsilateral fibers do provide the ipsilateral hemisphere with some basic perceptual information but do not generally provide enough cues for identification of the object.

Bogen, Sperry, and Gazzaniga used these basic facts about the anatomy of the nervous system to guide their investigations of the changes that follow section of the corpus callosum. By carefully isolating the hemispheres via these methods, they were able for the first time to map the profound changes that do occur after callosotomy. They examined the subjects in this new series before and after surgery and were able to confirm one of the early observations: These patients appear quite normal after surgery. To the casual observer, there appears to be very little difference in the presentation of the person before and after surgery. After the initial few weeks following surgery in which there may be symptoms such as mutism and intermanual conflict, the split-brain subjects experience the world as unified and converse and interact normally. The subjective response to what might be expected to be a radical and frightening change in one's inner world appears to be minimal. Gazzaniga observed, "Indeed, one would miss the departure of a good friend more, apparently, than the left hemisphere misses the right."

However, when observations were made under conditions that allowed only one hemisphere access to information and gave the mute right hemisphere a means of response, the now well-known hemispheric disconnection syndrome was elicited. The researchers gave patients objects to palpate in one hand at a laboratory table that screened the hand and the item from the subject's view. When an item was palpated by the right hand, there was no difference from normal behavior. The patient was easily able to name the item because all the tactile information from the right hand was available to the speaking left hemisphere. In contrast, when the same common objects were placed

in the left hand (and kept out of view), the talking left hemisphere was not able to identify the item and could not name it. The subject appeared to be anomic. However, despite being unable to name the object being grasped, the left hand of the subject could often demonstrate how the object was used or identify it from a group of objects. The somatosensory information from the left hand allowed the right hemisphere to identify but not to name the object. Because the information about the object could not be transmitted across the callosum, the talking left hemisphere could not help and supply the spoken name of the item. This was a vivid demonstration of what is now generally accepted about lateralization of function in right-handed people. The right hemisphere may possess knowledge about objects and their use in the world, but it lacks knowledge of how to produce the spoken name. This information is represented only in the left hemisphere, and once the callosum is cut the right hemisphere is left mute.

Another way to demonstrate the inability to transfer somatosensory information between the hemispheres in the absence of the callosum is to lightly touch different points on the patient's hand when it is out of view. If the right hand is touched in different positions, the right thumb can accurately point to each of the stimulated positions, but if the patient is asked to respond to the right-sided touch on the homologous area of the left hand, the task is impossible. This is equally true in reverse. The left hand can point to stimulated areas on the left hand but cannot transfer this information to the right hand. This task is trivial for people with an intact callosum, and if you are in doubt, try it. With your eyes closed, have a friend tap different areas on the palm and fingers of each hand. You will be able to respond easily with either the ipsilateral or the contralateral hand to the light touches.

Perhaps the most striking change demonstrated by Gazzaniga and Sperry was the inability to transfer visual information between the hemispheres. When two visual stimuli, either words or pictures, were lateralized one to each hemisphere, simple same/different judgments were performed at chance levels. That is, the patients could not accurately decide if the two words or pictures were identical or different when one was seen by the right hemisphere and one by the left hemisphere. However, within a hemisphere, both the right and the left hemispheres not only were able to decide if two stimuli were identical but also to match words and pictures. The ability to match words and pictures within the right hemisphere was particularly exciting because it showed that the isolated right hemisphere was able to read for meaning at least at the single word level (although it could not say the words out loud).

These basic observations ignited a period of rapid and productive investigation of the capacities of the two hemispheres. In the following sections, only the work that bears on a closer examination of the functional significance of the callosum will be presented.

III. ANATOMY AND BEHAVIOR

As the resurgence of interest in the results of callosotomy helped to elucidate the functional capacities of each hemisphere, questions arose regarding the specificity of callosal function. In an early review of the literature on callosal organization, Georgio Innocenti offered some general principles regarding the topography of callosal fibers. He considered the organization of the callosum in humans to be an interesting question because of the demonstration by Sidtis, Gazzaniga, and colleagues in the human split-brain subject that semantic and sensory aspects of visual stimuli were transferred in different parts of the callosum. The Gazzaniga laboratory and others have continued to report very specific limitations on transfer after partial lesions to the callosum.

Nonetheless, Francisco Abolitz and Eran Zaidel argue that there is great equipotentiality across the callosum. There is other anatomical support for this view. The de Lacoste work demonstrates substantial overlap between temporal lobe and superior frontal lobe fibers in the body of the callosum. The superior parietal lobe, temporal parietal junction, and occipital lobe all have some fibers passing in the splenium. Behaviorally, inconsistencies reported later suggest minimally that there may be notable individual differences.

One factor that can make the specific role of the callosum difficult to isolate is the myriad of additional interhemispheric tracts and other interactions that occur in the midbrain and brain stem. There are numerous other commissures not routinely part of split-brain surgery in humans, including the posterior commissure, the habenular commissure, the commissures of the inferior and superior colliculi and the massa intermedia, and the thalamic commissure (which is not present in all brains). As fibers descend into the brain stem, there is much less segregation and many pathways may share information across the

midline. These commissures are so deep in the brain that they are never severed during split-brain surgery in humans because of the devastating effects this would have on life-sustaining behavior. Hence, when an unexpected transfer of information occurs, the possibility that there is some subcortical communication that accounts for it is often raised.

Nonetheless, there are consistent deficits after complete section of the corpus callosum. Joseph Bogen outlined the principal deficits associated with lesions of the corpus callosum. He enumerated 10 symptoms that should be tested to confirm cases of callosal disconnection in normally lateralized right-handed subjects. In all cases, information is isolated in one hemisphere due to the loss of callosal fibers, resulting in the following symptoms:

1. Unilateral "verbal anosia": If an odor is presented unilaterally to the right nostril, it cannot be named, although the left hand can pick out the item associated with the odor.
2. Double hemianopsia: If responses are permitted from one hand at the time, the patient will appear to have a homonomous hemianopsia (i.e., an apparent blindness for one-half of space) in the field ipsilateral to the response hand. When responses are required from the opposite hand, the side of the "blind" field will change.
3. Hemialexia: If words are presented rapidly in one visual field, the ipsilateral hemisphere cannot give any sign that it has read or even seen the words.
4. Auditory suppression: Information presented from one ear is suppressed or extinguished by the ipsilateral hemisphere.
5. Unilateral (left) ideomotor apraxia: Because the left hemisphere is dominant not just for language but also generally for motor planning, the left hand is unable to carry out actions to command. It is essential to demonstrate that the deficit is not due to weakness or problems with coordination or ataxia.
6. Unilateral (left) agraphia: Most right-handed people have some ability to write with their left hand, although the letters may be less fluent and well formed than those made by their dominant hand. After callosal disconnection, the left hand loses this ability.
7. Unilateral (left) tactile anomia: Objects placed in the left hand but not seen cannot be named. Often, the function of the object can be demonstrated or a related object can be selected.
8. Unilateral (right) constructional apraxia: Because the right hemisphere is better at visual–spatial problems, the right hand will have difficulty executing complex drawings or manipulating three-dimensional objects or puzzles. However, it should be noted that the Gazzaniga laboratory has observed bilateral constructional apraxia in several patients. Gazzaniga has argued that these skills depend on interhemispheric integration in some patients, and therefore can be observed to decline for either hand after callosal section in these patients.
9. Spatial acalculia: The degree of visual–spatial impairment may be sufficiently great that patients are more successful at solving verbal arithmetic problems mentally than at writing them down with paper and pencil due to the distortion introduced by the act of writing. However, observation of this symptom may vary in the same way that it does for constructional apraxia.
10. Inability to transfer somesthetic information: As discussed previously, sensory and position information cannot be passed from hand to hand after callosotomy.

These distinctive sequelae of the disruption of cortical fibers had been noticed in part by the great 19th-century neurologists who laid the groundwork for our modern understanding of brain function. Sadly, many of their insights were lost or obscured by later experiments that failed due to shortcomings in observational techniques.

IV. SPECIFICITY

Here, we examine evidence from partial split-brain surgery, agenesis of the corpus callosum, and callosal lesions caused by stroke to better understand if different portions of the corpus callosum can be associated with different symptoms. After complete callosotomy, the whole range of symptoms discussed by Bogen should be found, with exceptions due to individual differences in lateralization. However, partial lesions may result in a subset of symptoms occurring, and these natural lesions provide another window into callosal function.

The discoveries of the Sperry, Gazzaniga, and Bogen group led to the identification of a new syndrome, alien (or anarchic) hand syndrome, by the French neurologists Serge Brion and C. P. Jedynak. They recognized that the unusual behaviors of a series of patients were related to naturally occurring callosal lesions. These lesions were the result of cerebrovascular disease or tumor. The patients shared a partial

lesion of the corpus callosum and the experience of a left or nondominant hand that appeared to be completing complex motor behaviors outside of the control of the dominant left hemisphere. Although Gary Goldberg subsequently showed that medial frontal lobe damage can result in anarchic hand sign that has both a dominant and a nondominant form, patients with no evidence of frontal damage represent what Todd Feinberg terms the callosal type of anarchic hand syndrome. Comparison of the symptoms and lesion location of patients with callosal anarchic hand sign provides another window into the location of the information that transfers in the callosum.

First, we review the observations made based on partial callosotomies, and then we use reports from patients with natural lesions of the callosum to refine these observations regarding specific transfer of information. We know from the differences observed in spilt-brain patients who had either anterior or posterior section of the callosum first that the splenium or posterior callosum is important in the transfer of visual information. This is expected given the function of the occipital lobes and the preservation of the anterior/posterior cortical organization of the cortex in the position of the callosal fibers.

Gail Risse and colleagues examined the transfer of visual, somatosensory, kinesthetic, auditory, and motor information in seven patients with section of the anterior callosum sparing the splenium and varying amounts of the posterior section of the body. By comparing transfer ability across these patients, they were able to confirm that intact splenial fibers permitted transfer of visual information for both naming and same/different judgments. However, patients that had only splenial fibers intact were unable to name objects palpated by the left hand. This suggests that the information transferred in the splenium is specifically visual and does not include higher order information synthesized from sensory information in another modality (i.e., it is generally observed that the right hemisphere is able to identify objects palpated by the left hand). If palpated objects cannot be identified by name in subjects with an intact splenium, it suggests that the tactile information necessary to build a semantic representation cannot cross to the left hemisphere in the splenium and neither can higher order information representing the semantic identification. In contrast, patients with the posterior one-third to one-fifth of the body of the callosum intact were able to name these objects relatively well. Generally, this patient group could also match limb position across hemispheres and could complete tests of apraxia with both the right and the left hand. Some difficulty with intermanual point localization was observed. A suppression of left ear auditory stimuli under dichotic conditions was noted for all but one subject. Although prior work indicates a role for the anterior callosum in praxis (due most likely to an interruption in the pathway necessary to carry the left hemisphere to ''translation'' of verbal commands for the right hemisphere to perform), there was little evidence of limb apraxia in this study except in the patients with sections that extended to the splenium. Another point made by Risse is that the left suppression on dichotic listening tasks in patients with good somatosensory function suggests that auditory fibers may be crossing anterior to somatosensory fibers, contrary to the results of Deepak Pandya and Benjamin Seltzer.

Although many sources confirm the importance of the splenium in the transfer of visual information, there may be even more specific function within the splenium. One of the patients investigated by the Gazzaniga laboratory, V.P., has shown some very specific transfer abilities. Margaret Funnell, Paul Corballis, and Michael Gazzaniga demonstrated that although V.P. is unable to transfer information about color, size, or shape across these spared fibers, she does show evidence of some access to words displayed to either hemisphere in the other, despite MRI confirmation of her status as a ''complete'' split. This remarkable specificity is supported by the work of Kyoko Suzuki and colleagues, who studied a young man with a small ventroposterior lesion to the callosum who could not read words in the left visual field. He could name pictures, however, suggesting that the anterior to middle section transfers picture information and the ventroposterior region is specific to letter transfer. Such precise lesions are rare but are revealing when carefully investigated. Of course, further work to confirm these observations is required, but they suggest that we are moving toward a much more specific understanding of the nature and location of the information transmitted in the callosum.

There remains disagreement regarding the relation between somatosensory and auditory fibers. A patient with a very discrete callosal lesion following a head injury showed increased suppression of left ear stimuli in dichotic testing when investigated by Michael Alexander and colleagues. His lesion appears to coincide with the posterior one-third to one-fourth of the body observed to be intact in some of Risse's patients. However, Risse found no auditory suppression in patients with good somatosensory function, suggesting auditory fibers were crossing anteriorly.

This was not true of Alexander's patient, who showed mildly impaired praxis and somatosensory function. (Unfortunately, transfer could not be tested because there were too many errors within a single hand to make it possible to observe decline in transfer.)

Early dichotic studies were in agreement that the portion of the callosum that caused greater left ear suppression was anterior to the splenium and was made up of the posterior one-half or one-third of the body of the callosum or was in the area of the isthmus. This was consistent with anatomical observations that indicated superior temporal lobe fibers crossed the callosum in this position. However, Sugushita and colleagues' careful examination of six patients with varied abnormalities of the corpus callosum suggested that damage to the splenium and posterior trunk leads to chronic left ear suppression. The patients in this study often had tumors that resulted in partial callosal resections, so there may have been some reorganization of function in response to tumor growth.

Given the individual differences in gyri, it may be that splenial fibers might better be defined by a common point of origin or by functional criteria rather than by a mathematically defined proportion. One paradox is that although some very precise disruptions in callosal transfer have been documented, there is significant variation in the areas where this transfer occurs. If this represents individual differences in the organization of the callosum, mathematical precision in separating the sections may not be helpful. In the meantime, the Risse observation remains difficult to reconcile with our current understanding of anatomy, and further observations will be necessary to understand the nature of the differences.

Another important function of the callosum is the transfer of information about an item identified either visually or tactilely in the right hemisphere to the left for oral naming. In a systematic investigation of the clinical signs associated with callosal transfer during a 12-month review of 282 new cases of cerebral infarction, Giroud and Dumas noted only 1 case of tactile anomia, and this patient had the most posterior callosal lesion, including the anterior one-third of the splenium. A patient investigated by Kathleen Baynes, Mark Tramo, and Michael Gazzaniga corroborates the observation that the anterior one-third of the splenium may be crucial for naming of items palpated by the left hand. This may represent transfer of word information as in the patient V.P., discussed previously.

However, all these observations await a better method for defining the regions of the callosum and tracing the origin and destination of fiber tracts than is currently available in order to be adequately confirmed or denied. It is also necessary to consider differences between right-to-left and left-to-right transmission because there are consistent behavioral advantages associated with direction of transfer.

V. CALLOSAL TRANSFER

Many research groups have been predominantly concerned with the speed and direction of callosal transfer or interhemispheric transfer time (IHTT). Perhaps the oldest method of estimation was developed by Albert Poffenberger in the early 19th century. He used lateralized visual displays to compare reaction time to material in the fields ipsilateral and contralateral to each response hand. Material in the contralateral field should be processed in the same hemisphere that initiates the manual response without need for callosal transfer. In contrast, ipsilateral displays must be processed in the opposite hemisphere and the signal must be transferred into the responding hemisphere. He subtracted crossed from uncrossed reaction times to yield a crossed–uncrossed difference as an estimate of how much time was required for this transfer. His results yielded an estimate of 2 or 3 msec for healthy adults. Such short transfer times are thought to implicate the large myelinated fibers of the callosum.

The advent of averaged evoked potentials to measure electrophysiological responses of populations of neurons provided another way of measuring IHTT. By comparing the time course of ipsilateral and contralateral peaks to lateralized visual or somatosensory stimulation, another estimate of IHTT has been obtained. Clifford Saron and Richard Davidson found about a 12-msec IHTT to visual stimuli. In contrast, somatosensory stimulation yielded estimates from 8 to 26 msec. The differences in even these simple stimulus-response IHTTs suggest to some that there is more than one route of callosal transfer, which would be expected if different sources of information indeed cross in different sections of the callosum. However, anatomical evidence of widespread callosal distribution of fibers from particular cortical loci also presents the possibility of multiple routes of crossing for specific types of information.

There have been many studies of callosal transfer that suggest that this transfer of information may not be symmetric. Carlo Marzi performed a meta-analysis that indicated that for right-handed people, responses

to displays in the left visual field with the right hand are faster than the converse. One implication of this finding is that the right hemisphere transfers information to the left hemisphere more rapidly than the left hemisphere transfers it to the right hemisphere. Because this advantage does not appear in left-handed people (although determination of hemispheric dominance is more problematic in this population), it may not mark a consistent or important principle of brain organization. It is nonetheless interesting to note that the nondominant hemisphere appears to be more adept at "reporting" to the dominant hemisphere, at least in simple reaction time tasks.

VI. CORPUS CALLOSUM AND COGNITION

Despite the remarkably normal presentation of split-brain subjects, profound changes can be observed when the appropriate experimental controls are in place to observe them. However, there remain many claims regarding the role of the corpus callosum in normal and abnormal behavior that are intriguing but require further study to determine if they have merit. Such diverse mental characteristics as lateralization of language, disposition of attention, mnemonic processing, conscious behavior, disorders of learning, and schizophrenia have all been linked to callosal changes.

The relatively large size of the corpus callosum in humans compared with other primates and the prominent cerebral lateralization in humans suggest that corpus callosum may play a role in the development or maintenance of lateralized behaviors. The first premise has been challenged, but despite increasing evidence of some lateralization of function in other species, it remains true that the lateralization of language in the human is one of the most striking examples of cerebral specialization. In a twist on this logic, Sandra Witelson suggested that within the human species, greater lateralization of function may lead to decreased callosal size because fewer fibers are necessary to connect bilateral areas with similar functions. Therefore, if right-handed people are more clearly lateralized than left-handed people, as data from examinations to determine lateralization of language and memory prior to brain surgery indicate, one would expect to see greater relative size of the corpus callosum in left-handed people. Likewise, there have been claims that males are more clearly lateralized than females, so females should as a group have larger callosums. Moreover, because language is the most clearly lateralized cognitive skill, and the axons from the language areas cross the midline primarily in the isthmus of the callosum, that structure is most likely to vary with relatively strong and weak lateralization.

With the advent of magnetic resonance imaging (MRI), visualization of the corpus callosum and measurement of its relative size in the living brain have become more straightforward. This has stimulated many attempts to test hypotheses regarding relative callosal size in right- and left-handed people, males and females, etc. The conclusions remain far from clear, although Witelson's work found a strong relationship between the site of the isthmus and handedness in men. Jeffrey Clark and Eran Zaidel, among others, have reported greater isthmus size in right-handed males compared to left-handed males. This effect is not found in females, although it is not known why handedness should have a different effect on lateralization and isthmus size in females. The Clark and Zaidel laboratories have performed many studies attempting to link behavioral measures to the size of the isthmus, with some success.

Although split-brain patients do not display the profound memory deficits associated with amnestic disorders, there do appear to be changes in the ability to integrate new memories in both laboratory and real-life settings. Most researchers with a long history of investigating these patients would agree that they exhibit a stereotypical conversational pattern that repeatedly returns to the same stories and episodes. This repertoire can expand to include information about things that happened postsurgically (unlike the dense memory deficits associated with hippocampal damage), but there is a lack of richness in the quality of reminiscence. These are, of course, subjective observations, and a more systematic evaluation of the clarity and accuracy of the pre- and postsurgical memory for life events in this population has yet to be completed.

There is evidence that clinical measures of memory show declines in some patients. Dahlia Zaidel observed declines on memory scales postsurgery, but this does not appear to be the case for all of the split-brain patients examined by the Gazzaniga laboratory. Elizabeth Phelps observed that memory scores decline after the posterior but not the anterior section of the corpus callosum, suggesting that perhaps inadvertent damage to the hippocampal commissure during surgery may contribute to the decline in memory skills. Amishi Jha and colleagues found that split-brain patients have difficulty "binding" both visual and verbal memory traces perhaps due to the dependence of some lateralized aspects of the memory trace

formation on interhemispheric communication. Hence, it seems likely that some combination of extracallosal damage as well as the resection may result in some decline in memory function that implicates the callosum in the formation of a complete memory trace.

Work in the laboratories of Marie Banich and Jacqueline Liederman has indicated that greater task difficulty leads to a greater advantage for bilateral presentations. The conclusion is that for tasks that are easy, one hemisphere is able to accomplish them with the same speed and accuracy as two. When tasks become more difficult, bilateral presentations can be more advantageous. Banich employed many cleverly designed and careful experiments to demonstrate the complexity of such interactions, with the corpus callosum providing a crucial mechanism for the division of labor across the hemispheres. Essentially, attention is a resource that depends at least partially on the corpus callosum for allocation. Since most tasks can be accomplished by either hemisphere, there must be a flexible neural mechanism for determining whether a single hemisphere or both are necessary under different task demands. Liederman stressed the corpus callosum as a mechanism to "shield" incompatible neural processes. Although it is clear that the corpus callosum plays some role in attentional processing because deficits in both selective and sustained attention have been observed in split-brain subjects, a detailed model for this role remains a goal for the future.

A theoretical model that depends heavily on callosal inhibition as a mechanism for some aspects of lateralized behavior has been proposed by Norman Cook. Although it is generally agreed that most callosal fibers are excitatory, their effect can be inhibitory via interneurons. Cook stresses the inhibitory processes in describing the process of homotopic callosal inhibition. Excitation of an area in one hemisphere leads to inhibition of the analogous area in the other hemisphere and excitation of immediately surrounding areas in a sort of center-surround arrangement. He also posits a bilateral neural semantic net such that when an item such as "farm" is excited in one hemisphere, the analogous item in the opposite hemisphere is inhibited. Immediately related items (tractor, pig, cow, sheep, and plow in one of his examples) are excited via center-surround mechanisms. This arrangement is the underlying cause of the denotative language capacity of the left hemisphere and the connotative language of the right hemisphere. He presumes this mirror-symmetric activation pattern to be a general mechanism, applicable to more than semantics. As Joseph Hellige rightly points out, the best evidence for interhemispheric inhibition is at the cellular level, whereas Cook often appears to be talking about a more functional role. Moreover, as his model stands, it would seem as if the hemispheric activation patterns could be easily reversed. This does not appear to be the case, at least for language functions. Despite the current lack of empirical support, this remains an interesting model of callosal function.

Despite many attempts to link morphological callosal differences to sex differences in lateralization, learning disabilities, and schizophrenia, definitive work in these areas remains elusive, and they will not be discussed here. The possibilities are intriguing, but the problems at this point contain too many degrees of freedom. The callosum appears to be a lever that could unlock our knowledge of brain function, but choosing the correct place to stand to use that leverage remains a problem.

VII. CORPUS CALLOSUM AND CONSCIOUSNESS

The study of split-brain patients during the past 40 years has helped change our understanding of the nature of consciousness. It has offered a prime example of the modularization of cognitive processes and documented the distinctions between a dominant and nondominant hemisphere. It has raised the question of whether the callosum may have played a unique role in the development of human consciousness. One of the key observations made regarding the Vogel and Bogen series of commissurotomies was that severing the callosum seemed to yield two separate conscious entities with the ability to respond independently. The idea of a "dual consciousness" was embraced by some scientists such as Pucetti, who hypothesized that the human condition was always made up of dual-consciousnesses that were only revealed after the section of the callosum. Others rejected the status of the right hemisphere as conscious. Daniel Dennett concluded that the right hemisphere had, at best, a "rudimentary self."

Michael Gazzaniga, Paul Corballis, and Margaret Funnell, recently proposed a new role for the corpus callosum as "the prime enabler for the human condition." They suggest the corpus callosum allows the brain to be more efficient, allowing hemispheric specialization but permitting integration of specialized functions as needed. In their view, lateral

specialization reflects the emergence of new skills and the retention of others. An advantageous mutation that changes the function of one hemisphere can be maintained and flourish while established functions continue without disruption. The callosum permits a reduction of redundant function and the easier acquisition of new skills. They suggest that the corpus callosum facilitated the development of a ''theory of mind,'' the skills that support the ability to understand the point of view of another creature, by permitting ''this extended capacity [to arise] within a limited cortical space.''

The best examples of behavior that appear to represent a separate consciousness in the right hemisphere come from split-brain patients with at least some language capacity. The newest completely split patient with normal intelligence, V.J., is anomalous in that she controls written and spoken output with different hemispheres. She is also unique in another way. She is the first split in my experience who is frequently dismayed by the independent performance of her right and left hands. She is discomforted by the fluent writing of her left hand to unseen stimuli and distressed by the inability of her right hand to write words she can read out loud and spell. In the myriad of articles discussing duality of consciousness, consciousness is sometimes considered as arising from the need for a single serial channel for motor output. In normally lateralized persons, the left hemisphere maintains control of output of both speech and writing. In V.J., there are two centers of control of the motor output of language, one partially disabled but still functional. One problem of this view point is that some split-brain patients, notably J.W. and V.P., also have some control of motor speech from either hemisphere. However, in both of these cases, the control of spoken language developed after the surgical intervention and this sequence of events may have different consequences for the conscious experience of it. If serial control of output is an important determinant of the function experienced as consciousness and the fluent shifting of control of output from one system to another is a part of that function, we may still have a good deal to learn from the split-brain model.

VIII. FUTURE DIRECTIONS

Giorgio Innocenti suggests that the complexity of callosal connections promises that these functional problems will remain fascinating but frustrating for some time. He points out that there is not yet an adequate understanding of the neural parameters that contribute to differences in size and shape of the callosum, including the number and size of axons that make up the callosum, the proportion of those axons that are myelinated, the thickness of the myelin sheath, and the number and size of blood vessels and other supporting elements. Although work in numerous laboratories is focused on answering some of these questions, until they are answered, studies correlating the size of the callosum or any of its sections to behavior are likely to continue to be frustrating.

Many of the fascinating but unproven hypotheses discussed previously could be more clearly addressed or ruled out if we had a better understanding of the neurophysiology of the human corpus callosum. Relatively little work has been done on the human callosum, although there is a large body of correlative data on differences in the morphology of the callosum derived from MRI and various aspects of behavior. However, there is still a very limited understanding of the anatomy of the callosum, where the crossing fibers originate and terminate, and the differences in the proportion of large and small fibers in different areas. Little is known about the connectivity of interneurons and how it affects the nature of the information transferred.

The advent of MRI led to a major increase in the number of studies examining gross correlations between anatomy and behavior, but the lack of a better understanding of the anatomy of callosal fibers and the basic mechanisms of transmission allowed too much freedom of interpretation. Hence, many of these studies reached inconsistent or conflicting conclusions.

A new method of image analysis known as diffusion tensor weighting is allowing researchers to examine the direction of movement in individual fiber tracts during different cognitive tasks. As this method is refined, and if it can be combined with a more explicit knowledge of callosal anatomy, the plethora of theories regarding the function of the callosum and its contribution to cognition may at last be open to more satisfying investigation.

See Also the Following Articles

ANTERIOR CINGULATE CORTEX • EPILEPSY • NEUROANATOMY

Suggested Reading

Beaton, A. A. (1997). The relation of planum temporale asymmetry and morphology of the corpus callosum to handedness, gender,

and dyslexia: A review of the evidence. *Brain Language* **60,** 255–322.

Bogen, J. E. (1993). *The callosal syndromes*. In *Clinical neuropsychology* (K. M. Heilman and E. Valenstein, Eds.), 3rd ed. pp. 337–407. Oxford Univ. Press, New York.

Clarke, J. M., McCann, C. M., and Zaidel, E. (1998). *The corpus callosum and language: Anatomical–behavioral relationships*. In *Right Hemisphere Language Comprehension: Perspectives from Cognitive Neuroscience* (M. Beeman and C. Chiarello, Eds.), pp. 27–50. Erlbaum, Mahweh, NJ.

Cook, N. D. (1986). *The brain Code: Mechanisms of Information Transfer and the Role of the Corpus Callosum*. Methuen, London.

Gazzaniga, M. S. (1970). *The Bisected Brain*. Appleton–Century–Crofts, New York.

Gazzaniga, M. S. (2000). Cerebral specialization and interhemispheric communication Does the corpus callosum enable the human condition? *Brain* **123,** 1293–1326.

Harris, L. J. (1995). *The corpus callosum and historic communication: An historical survey of theory and research*. In *Hemispheric Communication: Mechanisms and Models* (F. L. Kitterle, Ed.), pp. 1–59. Erlbaum, Hillsdale, NJ.

Hellige, J. (1993). *Hemispheric Asymmetry: What's Right and What's Left*. Harvard Univ. Press, Cambridge, MA.

Hoptman, M. J., and Davidson, R. J. (1994). How and why do the two cerebral hemispheres interact? *Psychol. Bull.* **116,** 195–219.

Innocenti, G. M. (1994). Some new trends in the study of the corpus callosum. *Behav. Brain Res.* **64,** 1–8.

Reeves, A. G., and Roberts, D. W. (Eds.) (1995). *Epilepsy and the Corpus Callosum 2*. Plenum, New York.

Cranial Nerves

ANN B. BUTLER
George Mason University

GLOSSARY

branchial arches The third and several additional visceral arches (usually five in all), which in fishes are in the region of the gills.

epimere The dorsal portion of the mesodermal layer in the developing body wall and head; also called paraxial mesoderm. It forms a segmental, rostrocaudal series of mesodermal masses called somites in the body and the caudal part of the developing head. Further rostrally in the head, it forms incompletely divided, segmental masses called somitomeres.

hypomere The ventral portion of the mesodermal layer in the developing body wall; also called lateral plate mesoderm. It gives rise to the smooth muscle of the gut and to the cardiac muscle of the heart.

mesomere The middle portion of the mesodermal layer in the developing body wall; also called the nephric ridge. It gives rise to the kidneys and gonads.

neural crest Ectodermally derived cells initially located at the lateral edge of the invaginating neural tube that subsequently migrate and contribute to numerous parts of the developing nervous system and body, including some of the bipolar neurons that lie in the ganglia of most sensory cranial nerves, all of the bipolar neurons of the sensory spinal nerve ganglia, the postganglionic neurons of the autonomic nervous system, most of the cranium, and the visceral arches.

neurogenic placodes Thickened regions of the epidermis that give rise to many of the bipolar neurons that lie in the ganglia of most of the sensory cranial nerves. Placodes occur only in the developing head region.

somites The segmental, mesodermal masses of the developing body wall and caudal head that are derived from epimere. In the body, somites give rise to the striated skeletal muscle of the body wall and limbs. In the caudal head, the somites give rise to most of the branchial arch muscles of the palate, pharynx, and larynx (innervated by cranial nerve X) and to the hypobranchial muscles of the tongue (innervated by cranial nerve XII).

somitomeres The incompletely divided, mesodermal masses of the developing head that are derived from epimere. They give rise to the striated muscles of the eyes (innervated by cranial nerves III, IV, and VI) and of the first three visceral arches (innervated by cranial nerves V, VII, and IX).

visceral arches A series of skeletal arches that occur in the region of the jaw and throat (pharynx) in mammals; they include the mandibular arch that forms the jaw, the hyoid arch, and the branchial arches. Most of the tissues of these arches are in fact somatic rather than visceral in developmental origin and function.

The cranial nerves provide the sensory and motor interfaces between the brain and the structures of the head. They supply the sensory inputs from our more than five senses and the motor (effector) innervation of muscles and glands. Like spinal nerves, the cranial nerves have sensory, or afferent, components that innervate structures in the head as well as the viscera of the thorax and abdomen and motor, or efferent, components that innervate muscles and glands in the head and the viscera. Three additional "special" components of cranial nerves are commonly recognized that spinal nerves lack; however, insights into the embryological derivation of sensory structures and muscles in the head allow us to discard this special category. Considering the sensory cranial nerves, humans arguably have at least 13 different senses, but even so we lack some additional senses that are present in other vertebrates. Some tetrapods, including most mammals, have an accessory olfactory (vomeronasal) system, for example, that is present in humans only

transiently during embryological development, and humans (along with other mammals) entirely lack the lateral line mechanoreceptive and electroreceptive systems of most aquatic vertebrates. We do not possess the infrared-receptive system of snakes or the independently evolved electroreceptive and mechanoreceptive systems via the trigeminal nerve that platypuses have. With regard to these other systems, we are, as William Wordsworth wrote, "creature[s] moving about in worlds not realized." Nevertheless, our set of sensory and motor, cranial and spinal nerves is the essential and only connection that the human central nervous system has with the external world. Via these nerves, all the sensory stimuli that one can detect are brought into the brain, and all the motor actions that one makes are commanded.

I. INTRODUCTION

Some of the cranial nerves are purely sensory, others are purely motor, and the rest have both sensory and motor components. Twelve cranial nerves have traditionally been recognized in humans, which are designated by Roman numerals as well as by descriptive names (Table I). The first two cranial nerves, the olfactory nerve (I) and the optic nerve (II), are purely sensory and innervate the nasal mucosa for the sense of smell (olfaction) and the eye for the sense of sight (vision), respectively. The 10 cranial nerves of the midbrain and hindbrain collectively comprise 23 individual, traditionally recognized components.

Table I
Traditional 12 Cranial Nerves

Nerve	Name
I	Olfactory
II	Optic
III	Oculomotor
IV	Trochlear
V	Trigeminal
VI	Abducens
VII	Facial
VIII	Vestibulocochlear (or statoacoustic)
IX	Glossopharyngeal
X	Vagus
XI	Spinal accessory
XII	Hypoglossal

Three purely motor nerves that innervate the muscles of the eye are collectively called oculomotor nerves and comprise the oculomotor nerve (III), the trochlear nerve (IV), and the abducens nerve (VI). The 4 nerves with both sensory and motor components that innervate the jaws, face, throat, and the thoracic and abdominal viscera are the trigeminal nerve (V), the facial nerve (VII), the glossopharyngeal nerve (IX), and the vagus nerve (X). Cranial nerve VIII, the vestibulocochlear (or statoacoustic) nerve, innervates the inner ear organs for the auditory and vestibular senses. Cranial nerves XI, the spinal accessory nerve, and XII, the hypoglossal nerve, are purely motor and innervate the muscles of the neck that are used to turn the head (the sternocleidomastoid and upper part of the trapezius) and the muscles of the tongue, respectively.

Most of the sensory cranial nerves are formed by the processes of bipolar (or pseudounipolar) neurons whose cell bodies lie within one or two sensory ganglia located on the nerve in the peripheral part of the nervous system. The sensory neurons each have a distal portion that either innervates a separate receptor cell or has a modified ending that itself is the receptor. Sensory transduction—the translation of the sensory stimulus into neuronal activity—involves a variety of physical and chemical mechanisms. The proximal portions of the bipolar sensory neurons project to a group of multipolar neurons that lie within nuclei in the central nervous system and in turn project to other groups of multipolar neurons in the sensory pathway. For each sensory system pathway, the bipolar-receptive, multipolar neurons are referred to here as first-order multipolar neurons since they are the first of several groups of multipolar neurons in the pathway. The sensory nuclei that contain the first-order multipolar neurons are named for the name of their major cranial nerve input (vestibular, cochlear, and trigeminal nuclei), the particular sensory system (gustatory nucleus), or for their appearance (solitary nucleus).

The motor cranial nerves innervate either muscles or glands. Most of the nerves that innervate muscles have cell bodies within their respective nuclei in the brain stem; their axons exit the brain in the cranial nerve and terminate directly on the particular muscle. Most of the nuclei of these nerve components have the same name as the nerve (oculomotor, trochlear, trigeminal motor, abducens, facial motor, and hypoglossal nuclei) or arise from a nucleus named for its indistinct appearance (nucleus ambiguus). Glands are innervated via a two-neuron chain of neurons that belong to the parasympathetic division of the autonomic

nervous system. The cell bodies of the first neurons in the chain, called preganglionic parasympathetic neurons, lie within nuclei in the brain stem, and their axons exit the brain in the cranial nerve. These axons terminate on a second set of neurons that lie in a ganglion located close to the target organ and are called postganglionic parasympathetic neurons in reference to their axons, which exit the ganglion and innervate the gland. Three small muscles within the eye are also innervated by the autonomic nervous system. Two of these intraocular muscles (for pupillary constriction and control of the shape of the lens) are innervated by the axons of postganglionic parasympathetic neurons, whereas the third (for pupillary dilation) is innervated by postganglionic axons that are part of the sympathetic division of the autonomic nervous system, which arises from neurons located within thoracic and upper lumbar spinal cord segments. The parasympathetic cell groups of the brain stem have a variety of names, including the eponymic Edinger–Westphal nucleus, the superior and inferior salivatory nuclei, the dorsal motor nucleus of X, and some of the neurons in nucleus ambiguus.

The motor nuclei of the brain stem all receive a variety of inputs from other neuron cell groups in the brain. These inputs include relatively local connections with reticular formation neurons and long, descending projections from motor regions of neocortex. Since the latter connections arise from neurons located above (rostral to) the cranial nerve nuclei, they are referred to as supranuclear connections. The majority of supranuclear inputs to cranial nerve nuclei are bilateral.

II. TRADITIONAL FUNCTIONAL COMPONENTS

Each of the various components of most of the cranial nerves has traditionally been classified as somatic (referring to body wall structures) or visceral (referring to internal organs), afferent (sensory) or efferent (motor), and general or special. The latter pair of terms are used in reference to earlier beliefs concerning the embryological derivation of some of the sensory structures and muscles in the head. Recent new information will allow us to discard these categories later in this article, but their current widespread use necessitates discussing them here.

The first two pairs of classification terms for cranial nerves correspond to the four components of spinal nerves. Sensory nerve components are thus either somatic afferent or visceral afferent, and motor components are likewise either somatic efferent or visceral efferent. Spinal nerve components and some cranial nerve components are additionally classified as general, so they are designated general somatic afferent (GSA), general visceral afferent (GVA), general somatic efferent (GSE), and general visceral efferent (GVE). The sensory GSA components innervate the skin of the face and position sense (proprioception) receptors in head musculature, whereas GVA components innervate the viscera of the thorax and abdomen and a few structures in the head and neck, such as the mucous membranes of the oral cavity. The motor GSE components innervate extraocular eye muscles and the muscles of the tongue, whereas GVE components supply parasympathetic innervation to the thoracic and abdominal viscera and to glands and intraocular muscles in the head.

Two sensory cranial nerve components are categorized as special and thus designated special somatic afferent (SSA) and special visceral afferent (SVA). The SSA category is applied to the auditory and vestibular senses, whereas the SVA category is applied to the sense of taste (gustation). The two most rostral cranial nerves of the traditional 12, for olfaction and vision, are frequently not categorized at all. One cranial nerve motor category is also designated as special—the special visceral efferent (SVE) components of cranial nerves V, VII, IX, and X. Cranial nerve XI is sometimes not categorized but can be included in this traditional SVE category as well. SVE components of cranial nerves innervate muscles of the face, the mandibular and hyoid arches, the throat, and the neck, which all develop embryologically from muscles of the visceral arches that include the mandibular arch for the jaw, the hyoid arch, and a series of branchial arches. The latter give rise to the gill region in fishes and to components of the throat in tetrapods. The variously recognized traditional components of the midbrain and hindbrain cranial nerves are summarized in Table II, whereas the newer, revised classification is presented in Table III.

In the spinal cord, the two sensory afferent and two motor efferent components of the spinal nerves have their central cell groups organized in a dorsal to ventral order of GSA, GVA, GVE, and GSE. These rostrocaudally running functional columns extend into the brain stem so that the central cell groups of the cranial nerves in the hindbrain and midbrain lie in a similar topographic order. Due to the geometry of the brain's ventricular system, these cell columns lie in a lateral to medial order of GSA, GVA, GVE, and GSE. The two additional "special" sensory components of the cranial

Table II
Traditionally Recognized Components of Midbrain and Hindbrain Cranial Nerves[a]

Cranial nerve	Component	Sensory ganglion or motor ganglion and/or nucleus
Oculomotor (III)	GSE	Oculomotor nucleus
	GVE	Edinger–Westphal and anterior medial nuclei/ciliary ganglion
Trochlear (IV)	GSE	Trochlear nucleus
Trigeminal (V)	GSA	Trigeminal ganglion
	SVE	Trigeminal motor nucleus
Abducens (VI)	GSE	Abducens nucleus
Facial (VII)	GSA	Geniculate ganglion
	SVA	Geniculate ganglion
	GVE	Superior salivatory nucleus/pterygopalatine and submandibular ganglia
	SVE	Facial motor nucleus
Vestibulocochlear (VIII)	SSA	Spiral and vestibular ganglia
Glossopharyngeal (IX)	GSA	Superior (jugular) ganglion
	GVA	Inferior (petrosal) ganglion
	SVA	Inferior (petrosal) ganglion
	GVE	Inferior salivatory nucleus/otic ganglion
	SVE	Nucleus ambiguus
Vagus (X)	GSA	Superior (jugular) ganglion
	GVA	Inferior (nodose) ganglion
	SVA	Inferior (nodose) ganglion
	GVE	Dorsal motor nucleus of X and nucleus ambiguus/parasympathetic ganglia of thoracic and abdominal viscera and heart
	SVE	Nucleus ambiguus
Spinal accessory (XI)	SVE	Cervical anterior horn cells
Hypoglossal (XII)	GSE	Hypoglossal nucleus

[a]Abbreviations used: GSA, general somatic afferent; GVA, general visceral afferent; SSA, special somatic afferent; SVA, special visceral afferent; GSE, general somatic efferent; GVE, general visceral efferent; SVE, special visceral efferent.

nerves (SSA and SVA) have central cell groups that lie near the GSA nuclei. The SVE cell column lies in a more displaced, ventrolateral position, which is a legacy of its early evolutionary origin. The SVE cranial nerves and the visceral arch muscles that they supply evolved in the earliest vertebrates before the GSE column and its associated set of muscles were gained.

A. Cranial Nerves of the Medulla

Four of the cranial nerves are present in the medulla: the hypoglossal nerve (XII), the spinal accessory nerve (XI), the vagus nerve (X), and the glossopharyngeal nerve (IX). Cranial nerves XII and XI are simple nerves that have only a single component, which is motor. Cranial nerves X and IX are the two most complex cranial nerves, each with five traditional functional components.

1. Hypoglossal Nerve

Cranial nerve XII innervates the muscles of the tongue. A fleshy tongue and thus a distinct hypoglossal nucleus and nerve are present only in tetrapods. The tongue develops embryologically from hypobranchial musculature that lies ventral to the visceral arches of the pharyngeal region. The only component present in the hypoglossal nerve is the GSE. The nerve arises from motor neurons in the hypoglossal nucleus, which lies in a medial position in the dorsal part of the medulla and is the caudalmost portion of the GSE column in the brain stem. Hypoglossal nerve fibers run ventrally from the hypoglossal nucleus and exit the medulla on its ventral surface between the medially lying pyramidal (corticospinal) tract and the more laterally lying inferior olivary nucleus. Hypoglossal innervation of the tongue is ipsilateral.

Damage to the ventromedial medulla may thus result in the syndrome called inferior alternating hemiplegia, which is motor impairment on the contralateral side of the body (due to damage to the pyramidal tract) combined with weakness of tongue muscles on the ipsilateral (i.e., alternate) side due to a lower motor neuron lesion of the nerve. When protruded, the tongue deviates toward the side of the lesion. Supranuclear innervation of the hypoglossal nucleus is bilateral, so lesions that interrupt this input result only in mild weakness of the tongue musculature on the side opposite to the lesion. During normal usage, the pattern of tongue movements is modulated by inputs to the hypoglossal nucleus relayed from other cranial nerve nuclei, including those of V for

Table III
New Classification of Cranial Nerve Components in Humans[a]

Cranial nerve	Traditional classification	New classification	Embryological derivation or site of innervation
Terminal	—	SA	Olfactory placode
I	—	SA	Olfactory placode
II	—	NTA	Neural tube
Epiphyseal	—	NTA	Neural tube
III	GSE	SE	Epimeric muscles
	GVE	VE	Ciliary ganglion
IV	GSE	SE	Epimeric muscle
V	GSA	SA	Neural crest and trigeminal placode
	SVE	SE: branchial motor	Epimeric muscle
VI	GSE	SE	Epimeric muscle
VII	GSA	SA	Neural crest and/or ventrolateral placode
	SVA	VA	Neural crest and/or ventrolateral placode
	GVE	VE	Neural crest
	SVE	SE: branchial motor	Epimeric muscles
VIII	SSA	SA	Dorsolateral placode
IX	GSA	SA	Neural crest and/or ventrolateral placode
	GVA	VA	Neural crest and/or ventrolateral placode
	SVA	VA	Neural crest and/or ventrolateral placode
	GVE	VE	Neural crest
	SVE	SE: branchial motor	Epimeric muscle
X	GSA	SA	Neural crest and/or ventrolateral placode
	GVA	VA	Neural crest and/or ventrolateral placode
	SVA	VA	Neural crest and/or ventrolateral placode
	GVE	VE	Neural crest
	SVE	SE: branchial motor	Epimeric muscles
XI	SVE	SE: branchial motor	Epimeric muscles
XII	GSE	SE	Epimeric muscles

[a]Abbreviations used: GSA, general somatic afferent; GVA, general visceral afferent; SSA, special somatic afferent; SVA, special visceral afferent; GSE, general somatic efferent; GVE, general visceral efferent; NTA, neural tube afferent; SVE, special visceral efferent; SA, somatic afferent; VA, visceral afferent; SE, somatic efferent; VE, visceral efferent.

proprioception and IX and X for stimuli from the mucosal lining of the oral cavity.

2. Spinal Accessory Nerve

Cranial nerve XI (the accessory nerve) has traditionally been considered to comprise two parts, a cranial part that arises from neuron cell bodies in the caudal part of nucleus ambiguus (which lies in the lateral region of the medulla) and a spinal part that arises from neuron cell bodies located in the lateral part of the gray matter within upper segments of the cervical spinal cord. The laterally displaced position of the latter cell group is similar to that of the corresponding cell column in the brain stem. The so-called cranial part of XI (not included in Table II) is in fact merely the more caudal part of the vagus nerve. It supplies innervation to the intrinsic muscles of the larynx on the ipsilateral side. The spinal part of XI innervates the ipsilateral sternocleidomastoid muscle and the upper parts of the trapezius muscle. Contraction of the sternocleidomastoid muscle on one side causes the mastoid process (which lies behind the ear) to approach the sternum and medial part of the clavicle

(''cleido''), thus causing the head to turn toward the contralateral side. Innervation of the spinal nucleus of XI from motor cortex appears to be predominantly ipsilateral. Damage to the spinal accessory nerve results in weakness of the ipsilateral muscles, which is revealed when trying to turn the head toward the opposite side against resistance.

The accessory nerve is sometimes unclassified, but it is a branchial motor nerve and belongs in the traditional category of SVE. The sternocleidomastoid and trapezius muscles evolved from the levator muscles of the branchial arches.

3. Vagus Nerve

Cranial nerve X has five components and by this measure joins cranial nerve IX in the distinction of being most complex. It has three sensory components GSA, SVA, and GVA and two motor components GVE and SVE. The GSA component is a minor one; its fibers innervate the region of the external auditory meatus and the tympanic membrane. These GSA fibers have cell bodies located in the superior (jugular) ganglion of X and enter the brain stem via the vagus nerve. They terminate in a rostrocaudally elongated group of first-order multipolar neurons that receive most of their input from GSA fibers of the trigeminal nerve. This group of neurons is called the spinal, or descending, nucleus of V since it extends from the pontine levels caudally into the upper part of the spinal cord, ending at the level of the second cervical segment. The ascending sensory pathway that arises from the spinal nucleus of V is discussed later.

The GVA and SVA sensory components of the vagus nerve are more substantial than the GSA component. The GVA component has its bipolar neurons located in the inferior (nodose) ganglion of X and provides sensory innervation to the viscera of the thorax and abdomen as well as to mucous membranes lining the pharynx, larynx, trachea, and esophagus. In the brain stem, these GVA axons enter a fiber bundle that is surrounded by its nucleus and thereby appears to be isolated from neighboring structures. This bundle is called tractus solitarius, and its nucleus, which contains the first-order multipolar neurons that receive the visceral sensory input, is called nucleus solitarius (the nucleus of the solitary tract). Nucleus solitarius projects to multiple brain stem sites, including the dorsal motor nucleus of X and nucleus ambiguus.

Additional visceral sensory innervation is via afferent fibers that traverse sympathetic and parasympathetic nerves and then terminate on spinal cord dorsal horn neurons, which in turn project to the dorsal thalamus (ventral posterolateral and intralaminar nuclei). This information is then relayed to the neocortex. This ascending system accounts for most of the sensations from thoracic, abdominal, and pelvic viscera that are consciously experienced. Since the same pool of dorsal horn neurons also receives inputs from the somatic part of the body at each segmental level, this pattern of innervation is responsible for the phenomenon of ''referred pain,'' in which visceral sensations are perceived as originating from the body wall.

The SVA component of the vagus nerve has bipolar neurons located in the inferior (nodose) ganglion of X. The cell group that receives taste afferents via this SVA component of the vagus nerve (and the SVA components of the glossopharyngeal and facial nerves as well) has previously been viewed as the rostral part of nucleus solitarius but is now recognized as a distinct and separate entity, the gustatory nucleus. The first-order multipolar neurons located in this nucleus receive taste sensation via the superior laryngeal branch of the vagus nerve from taste buds located on the epiglottis and esophagus. An ascending pathway originates from the gustatory nucleus and projects to a nucleus in the dorsal thalamus, specifically the parvicellular portion of the ventral posteromedial nucleus (VPMpc), which in turn projects to two neocortical gustatory regions, one in the insular region and the other in parietal opercular cortex. The gustatory nucleus has other projections, including those to nucleus ambiguus and to the hypoglossal and salivatory nuclei for throat, lingual, and secretory reflexes.

The two motor components of the vagus nerve are both traditionally classified as visceral components. The GVE component provides innervation to the parasympathetic ganglia that supply the viscera of the thorax and abdomen. (The parasympathetic ganglia for pelvic viscera receive their parasympathetic innervation from preganglionic neurons located in the sacral spinal cord.) Preganglionic parasympathetic GVE neurons are mostly located in the dorsal motor nucleus of X, which lies immediately dorsolateral to the hypoglossal nucleus. Some GVE neurons also lie within nucleus ambiguus and specifically supply additional parasympathetic innervation to the heart. The vagus nerve exits the lateral surface of the medulla, and its GVE fibers distribute to parasympathetic ganglia for the esophagus, trachea and lungs, heart, and most of the abdominal viscera. Parasympathetic activity results in decreased heart rate, bronchial

constriction, increased blood flow to the gut, and increased peristalsis and secretion.

The SVE, branchial motor component of the vagus nerve arises from neurons in nucleus ambiguus and innervates muscles of the palate, pharynx, and larynx. These muscles are embryologically derived from muscles associated with the branchial arches. Nucleus ambiguus receives bilateral corticobulbar input and input from other cranial nerve nuclei for coordination of movements involved in reflex actions (such as coughing or vomiting) and for phonation and swallowing. The latter, for example, is a complex action that also involves motor neurons of the trigeminal (V), facial (VII), and hypoglossal (XII) nerves.

Damage to nucleus ambiguus or to the fibers of the vagus nerve results in reduction of the arch of the palate on the ipsilateral side, difficulty in swallowing, and hoarseness. Hoarseness is one of the manifestations of Parkinson's disease. It is due to supranuclear effects on the function of SVE vagal neurons. Hoarseness can also result from a lung tumor on the left side that involves the left recurrent laryngeal nerve; the latter arises from the vagus nerve in the upper part of the thorax, an anatomical relationship that does not occur on the right due to a higher origin of the recurrent nerve.

4. Glossopharyngeal Nerve

Like the vagus nerve, cranial nerve IX has five components, three of which are sensory and two of which are motor. The GSA component is a minor one, which innervates a small region of skin behind the ear. The GSA fibers of the glossopharyngeal nerve have cell bodies located in the superior (jugular) ganglion of IX, and they traverse the spinal tract of V to terminate in the spinal nucleus of V with similar GSA afferents from cranial nerves X, VII, and V. The GVA component of IX has a much more restricted distribution than the GVA component of X. Glossopharyngeal GVA fibers innervate the mucosal membranes that line the internal surface of the middle ear and cover the posterior third of the tongue, the tonsils, the posterior and upper surfaces of the pharynx, and the Eustachian tube. These GVA fibers have cell bodies in the inferior (petrosal) ganglion of IX. They enter the tractus solitarius and terminate in its nucleus. The latter projects to sites that include nucleus ambiguus (for throat reflexes), the dorsal motor nucleus of X, and the reticular formation. The carotid sinus reflex depends on some additional glossopharyngeal GVA fibers that innervate the carotid sinus (which is located at the bifurcation of the internal and external carotid arteries) and are stimulated by increases in arterial blood pressure. These fibers terminate on solitary nucleus neurons that project to the dorsal motor nucleus of X, which, via its projection to the parasympathetic ganglia of the heart, decreases heart rate and the arterial blood pressure. The SVA component of IX innervates taste buds located on the posterior third of the tongue. These fibers have cell bodies located in the inferior (petrosal) ganglion of IX and project to first-order multipolar neurons in the gustatory nucleus, which, as discussed previously, projects to VPMpc in the dorsal thalamus for relay to gustatory cortical areas.

The glossopharyngeal nerve has two efferent components, GVE and SVE. In the latter, branchial motor fibers arise from neurons in nucleus ambiguus and innervate the stylopharyngeus muscle (which arises from part of the styloid process on the inferior surface of the temporal bone of the skull and inserts along the side of the pharynx and part of the thyroid cartilage). The stylopharyngeus muscle is embryologically derived from the muscles of the branchial arches. The GVE component of IX arises from a nucleus that lies rostral to the dorsal motor nucleus of X but is displaced laterally away from the ventricular surface; it is called the inferior salivatory nucleus. As its name implies, its neurons give rise to preganglionic parasympathetic fibers that project via the tympanic and lesser superficial petrosal nerves to the otic ganglion, which in turn gives rise to postganglionic parasympathetic fibers that innervate the largest of the salivary glands—the parotid gland. Parasympathetic stimulation increases the secretory activity of the parotid gland. The glossopharyngeal nerve is rarely involved clinically in isolation, and deficits are predominantly sensory. Damage to cranial nerve IX results in loss of the pharyngeal (gag) reflex due to loss of the sensory input, a similar loss of the carotid sinus reflex for blood pressure regulation, and loss of taste to the posterior third of the tongue. Loss of salivatory activity by the parotid gland results from damage to the glossopharyngeal GVE component.

B. Cranial Nerves of the Pons

Four cranial nerves are present at pontine levels: the vestibulocochlear nerve (VIII), the facial nerve (VII), the abducens nerve (VI), and the trigeminal nerve (V). Cranial nerve VIII has a single sensory component that

comprises two divisions, and cranial nerve VI has only a single motor component. Cranial nerve VII is second in complexity only to the vagus and glossopharyngeal nerves, with four traditionally recognized functional components, whereas cranial nerve V is a mixed nerve with two components, one sensory and one motor.

1. Vestibulocochlear Nerve

The eighth cranial nerve traditionally has been classified in one of the special categories, SSA. It comprises the two senses of audition and the relative position and motion of the head in space. The eighth nerve innervates the inner ear organs—the spiral-shaped cochlea, which transduces auditory stimuli via vibration of its perilymphatic fluid, and the vestibular labyrinth, which transduces the effects of gravity and acceleration on its receptor apparatus.

a. Cochlear Division The bipolar neurons of the cochlear division of VIII innervate the hair cells of the cochlear organ of Corti. The apical surfaces of the hair cells are studded with stereocilia and border the inner chamber of the cochlea, the scala media, which is filled with endolymph. A tectorial membrane overlies the stereocilia. An outer chamber, formed by the scala vestibuli and scala tympani, contains perilymph, and the scala tympani portion is separated from the organ of Corti by a basilar membrane. Sound waves cause vibration of the perilymph and resultant displacement of the basilar membrane, which in turn pushes the stereocilia against the tectorial membrane, bending them and opening ion channels. Influx of potassium from the endolymph through the opened channels causes depolarization of the receptor cell. The frequency of the auditory stimuli is tonotopically mapped along the length of the cochlear spiral, with best responses to high frequencies occurring at its base and those to lower frequencies at more apical locations. The bipolar neurons preserve the tonotopic map for relay to the cochlear nuclei and then throughout the ascending auditory pathway. They also encode intensity by their discharge rate.

Cell bodies of cochlear bipolar neurons lie within the spiral ganglion, named for the shape of the cochlea. Their central processes enter the lateral aspect of the brain stem at a caudal pontine level and terminate in the dorsal and ventral cochlear nuclei. The cochlear nuclei project to multiple sites, including the superior olivary nuclear complex in the pons and the inferior colliculus in the midbrain roof. The superior olivary complex (SO) receives bilateral input mainly from the so-called bushy cells of the ventral cochlear nucleus; SO neurons are coincidence detectors that utilize the time delay between the inputs from the two sides in order to compute the location in space of the sound source. The ascending auditory projections predominantly originate from pyramidal neurons within the dorsal cochlear nucleus and from the superior olivary complex and pass via the lateral lemniscus to the inferior colliculus, which in turn projects to the medial geniculate body of the dorsal thalamus. The latter projects to auditory cortex in the temporal lobe. Damage to the cochlear division of VIII results in dysfunction (such as experiencing a buzzing sound) and/or deafness.

b. Vestibular Division The bipolar neurons of the vestibular division of VIII innervate hair cells within several parts of the vestibular labyrinth. The fluid-filled labyrinth consists of three semicircular canals that transduce rotational movements (angular acceleration) of the head and two chambers, the saccule and utricle, that constitute the otolith organ and transduce linear acceleration and gravitational force. The three semicircular canals are arranged at right angles to each other for responses in the various planes of space, and each contains a dilated area, the ampulla, that contains a ridge, the crista ampullaris. Vestibular hair cells on the surface of the crista have stereocilia and a longer kinocillium that are displaced by movement of an overlying gelatinous cupula, which moves due to fluid displacement within the canal caused by rotational motion. Similarly, the saccule and utricle contain hair cells in an area called the macula. The cilia of these hair cells are displaced by movement of an overlying gelatinous mass, the otolith membrane, that contains the otoliths (or otoconia), which are small particles of calcium carbonate that are denser than the surrounding fluid and are displaced by gravity when the head is tilted or during linear acceleration.

Cell bodies of vestibular bipolar neurons lie within the vestibular (Scarpa's) ganglion. The central processes of these neurons enter the lateral aspect of the pons along with the cochlear nerve fibers; most terminate in the superior, inferior, medial, and lateral vestibular nuclei, which lie in the floor of the fourth ventricle. A small number of vestibular afferent fibers bypass the vestibular nuclei and project directly to the cerebellar cortex, particularly within its flocculonodular lobe, which is concerned with the maintenance of

posture, balance, and equilibrium. The vestibular nuclei also receive a substantial input from the cerebellum via its deep nuclei.

Descending projections from the vestibular nuclei form the medial and lateral vestibulospinal tracts that innervate spinal cord neurons for extensor muscles involved in postural maintenance and related reflexes. Unlike other brain stem nuclei that give rise to descending spinal projections, the vestibular nuclei do not receive any direct input from the cerebral cortex. Other major vestibular connections are with the oculomotor nuclei of cranial nerves III, IV, and VI via the medial longitudinal fasciculus for the vestibuloocular reflex—the stabilization of eye fixation on a target while the head is moving—and other vestibular–oculomotor interactions. The ascending pathway for conscious perception of vestibular stimuli arises from neurons in the superior, lateral, and inferior vestibular nuclei and terminates in several dorsal thalamic nuclei, including part of the ventral posterolateral (VPLc) nucleus and the ventral posteroinferior (VPI) nucleus. Thalamocortical vestibular projections are to ventrally lying parietal cortical areas (designated 2v and 3a), which are adjacent to the head representation of somatosensory and motor cortices, respectively. Damage to the vestibular division of VIII results in dizziness, pathologic nystagmus (an involuntary, repeated eye movement pattern consisting of conjugate movement of the eyes to one side followed by a rapid return to the original position), nausea and vomiting, vertigo (the sensation of rotation in the absence of actual movement), and other related symptoms.

2. Facial Nerve

Cranial nerve VII contains four of the five components present in the vagal and glossopharyngeal nerves, namely, GSA, SVA, GVE, and SVE. The seventh nerve can be divided into two parts—the intermediate nerve (of Wrisberg), which comprises the GSA, SVA, and GVE components, and the facial branchial motor nerve, which is the SVE component. The two sensory components of VII both have cell bodies located in the geniculate ganglion. As is also the case for cranial nerves IX and X, the GSA component of VII is minor, innervating a small area of skin behind the ear and the external auditory meatus. The GSA fibers terminate in the spinal nucleus of V, for which the ascending sensory pathway is discussed later. Facial SVA fibers innervate taste buds that lie on the anterior two-thirds of the tongue via the chorda tympani nerve. These neurons project to the gustatory nucleus, which, as discussed previously, projects to VPMpc in the dorsal thalamus for relay to gustatory cortical areas.

Special visceral efferent fibers arise in the branchial motor nucleus of VII, which lies ventrolateral to the abducens (VI) nucleus in the pons. The axons of motor VII initially follow a dorsomedial course within the pons and then turn laterally and curve over the dorsal surface of the abducens nucleus, forming the internal genu of the nerve. This course is due to ventrolateral migration of the neuron cell bodies during embryological development. The facial motor fibers then traverse the pons in a ventrolateral direction to exit on its lateral surface near the junction of the pons with the medulla. The facial branchial motor nerve innervates muscles embryologically derived from the muscles of the branchial arches, including the platysma (a superficial muscle of the skin of the neck), buccinator (which forms the cheek), stapedius (which inserts on the stapes bone within the middle ear), auricular and occipital muscles, and the muscles of facial expression. An additional component supplies efferent innervation to the inner ear. The branchial motor nucleus of VII receives afferent input from structures that include the spinal nucleus of V for corneal and other trigeminofacial reflexes, auditory input from the superior olivary complex for facial reflexes to loud noise (including closing the eyes and the stapedius reflex to damp sound transmission through the middle ear ossicles), and corticobulbar input, which is bilateral to the facial motor neurons for the forehead and upper face but predominantly contralateral for the lower face and mouth region.

The GVE component of the facial nerve arises from neurons in the superior salivatory nucleus. Some preganglionic parasympathetic fibers project via the greater petrosal nerve to the pterygopalatine ganglion for innervation of the lacrimal gland. Other preganglionic fibers project via the chorda tympani nerve to the submandibular ganglion for innervation of the sublingual and submandibular salivary glands. Damage to cranial nerve VII results in deficits according to the location of the lesion and the components involved. Sensory loss of taste to the anterior two-thirds of the tongue occurs with involvement of the SVA component, and loss of production of tears from the lacrimal gland and saliva from the sublingual and submandibular salivary glands occurs with involvement of the GVE component. Damage to the nucleus or nerve for the SVE motor VII component results in paralysis of all the muscles of the ipsilateral face,

whereas a supranuclear lesion affecting corticobulbar input results in paralysis of only the lower part of the face on the contralateral side. Motor VII damage also results in sounds being abnormally loud due to loss of the stapedius reflex.

3. Abducens Nerve

Cranial nerve VI is one of the set of three ocular motor (oculomotor) nerves (III, IV, and VI) and innervates one of the six extraocular muscles of the eye, the lateral rectus muscle. The abducens nerve is a purely motor nerve with only a GSE component. The nerve arises from motor neurons in the abducens nucleus, which lies in a medial position in the dorsal pons and, along with the other oculomotor nuclei and the hypoglossal nucleus, forms the GSE column of the brain stem. Abducens nerve fibers run ventrally from the nucleus and exit the brain stem on its ventral surface at the junction of the pons with the medulla. The abducens nerve runs forward along the side of the brain stem, traverses the cavernous sinus, and passes through the superior orbital fissure (along with the other oculomotor nerves) to innervate the ipsilateral lateral rectus muscle. Contraction of the lateral rectus causes abduction (lateral movement) of the eye. Damage to the ventral region of the brain stem at the level of the abducens nerve may result in the syndrome called middle alternating hemiplegia, which is motor impairment on the contralateral side of the body (due to damage to the corticospinal tract) combined with medial deviation of the eye on the ipislateral (i.e., alternate) side due to a lower motor neuron lesion of the abducens nerve. In this situation, the eye deviates medially due to the unopposed contraction of the medial rectus muscle, which is innervated by cranial nerve III. Supranuclear innervation of the oculomotor nuclei derives from the superior colliculus and from cortical eye fields in the frontal and parietal lobes, and it is relayed to the oculomotor nuclei via gaze centers in the brain stem.

Connections from the gaze centers and vestibular nuclei and among the oculomotor nuclei via the medial longitudinal fasciculus serve to coordinate the actions of the extraocular muscles so that, for example, when a person moves both eyes to the right, the right lateral rectus contracts to move the right eye laterally, the right medial rectus does not oppose the action, and the actions of the medial and lateral recti of the left eye are similarly coordinated to produce the conjugate movement. Damage to the medial longitudinal fasciculus results in internuclear ophthalmoplegia; this condition is characterized by weakness in adduction of the eye on the side ipsilateral to the lesion during attempted conjugate eye movements toward the opposite side accompanied by nystagmus in the contralateral, abducting eye. Bilateral internuclear ophthalmoplegia can occur due to multiple sclerosis.

4. Trigeminal Nerve

Cranial nerve V has only two components, GSA and SVE. The trigeminal nerve is named for its three divisions: The ophthalmic division provides sensory innervation to the upper face, including the forehead, upper eyelid, and cornea; the maxillary division to the region of the upper jaw, including the upper lip, jaw, and cheek, parts of the nose, and upper teeth; and the mandibular division to the region of the lower jaw, including the lower lip, jaw, cheek, lower teeth, and anterior two-thirds of the tongue. The mandibular division also provides motor innervation to the muscles of the jaw used for mastication (chewing). The trigeminal nerve emerges from the lateral aspect of the brain stem at a midpontine level.

The GSA fibers of the trigeminal nerve have most of their cell bodies located within the trigeminal ganglion and innervate the face and upper head region for fine (discriminative) touch, position sense, vibratory sense, pain, and temperature. GSA fibers for the modalities of fine touch, vibration, and position enter the brain stem and terminate in the principal sensory nucleus of V in the lateral part of the pons, which in turn projects to the ventral posteromedial nucleus (VPM) in the dorsal thalamus. The latter projects to the face representation within somatosensory cortex. GSA fibers for pain and temperature enter the brain stem and distribute to the spinal (or descending) nucleus of V, which also receives inputs from the GSA components of the facial (VII), glossopharyngeal (IX), and vagus (X) nerves. Like the principal sensory nucleus of V, the spinal nucleus projects to VPM for relay of its inputs to somatosensory cortex. (Trigeminal innervation of the mucous membranes of the nose and anterior tongue region has traditionally been lumped with the GSA components, but since these membranes are embryologically derived from endoderm, this component is technically general visceral afferent.)

A third trigeminal sensory nucleus, the mesencephalic nucleus of V, is present along the lateral border of the central gray matter encircling the upper part of the fourth ventricle and the cerebral aqueduct of the midbrain. This nucleus comprises bipolar neurons that

lie within the central nervous system rather than in the cranial nerve ganglion. The neurons' peripheral processes form the mesencephalic tract of V and carry pressure and position information from structures including the teeth, the palate, and the muscles of mastication. The mesencephalic V neurons project to multiple sites, including the motor nucleus of V for reflexive control of jaw position.

SVE, branchial motor fibers arise in the motor nucleus of V, which lies medial to the principal sensory nucleus and receives bilateral corticobulbar input. The axons exit the brain stem on its lateral surface in the trigeminal nerve head and distribute via the mandibular division to the jaw muscles used in mastication—the temporalis and masseter muscles and the medial and lateral pterygoids, which are embryologically derived from the muscles of the branchial arches. Additional branchial arch-derived muscles innervated by the branchial motor component of V are the tensor veli palatini muscle of the palate, the tensor tympani muscle of the middle ear (which inserts on the malleus), and two suprahyoid muscles, the anterior belly of the digastric and mylohyoid.

Damage to the GSA components of cranial nerve V results in sensory loss to the face and loss of many reflexes due to lack of the sensory part of the reflex arc, including the corneal reflex (closing the eyes in response to light touch to the cornea) via the facial motor nucleus (via the reticular formation), the tearing reflex via the superior salivatory nucleus, the sneezing reflex via nucleus ambiguus, the vomiting reflex via vagal nuclei, salivatory reflexes via the superior and inferior salivatory nuclei, and the jaw-jerk reflex via the mesencephalic nucleus of V. The sensation for hot pepper (capseisin) on the anterior two-thirds of the tongue is also lost. Damage to the SVE component of V results in paralysis of the jaw muscles on the ipsilateral side, whereas a supranuclear lesion produces only moderate weakness due to the bilaterality of the supranuclear input. Changes in the loudness of sounds can also result from loss of innervation to the tensor tympani muscle.

C. Cranial Nerves of the Midbrain

Only two cranial nerves are present in the midbrain: the trochlear nerve (IV) and the oculomotor nerve (III). Cranial nerve IV has only a single component, which is motor. Cranial nerve III has two motor components.

1. Trochlear Nerve

Cranial nerve IV is one of the set of three oculomotor nerves (III, IV, and VI) and innervates one of the six extraocular muscles of the eye, the superior oblique muscle. The trochlear nerve is a purely motor nerve with only a GSE component. The nerve arises from motor neurons in the trochlear nucleus, which lies in a medial position in the dorsal part of the caudal half of the midbrain tegmentum. The trochlear nerve is unique among cranial nerves in that it decussates to the contralateral side (through the superior medullary velum that forms the roof of the fourth ventricle), and its point of exit is through the dorsal surface of the brain. The trochlear nerve thus innervates the superior oblique muscle of the contralateral eye.

The superior oblique muscle arises from the dorsomedial surface of the orbit and ends in a tendon that bends through a connective tissue pulley and then passes laterally to insert on the dorsal part of the eye bulb lateral to its center. Contraction of the superior oblique muscle causes intorsion (rotation of the eyeball around a horizontal, anteroposterior axis through the pupil such that the top moves medially), depression, and abduction. Deviation of the eye due to isolated damage to the trochlear nucleus or nerve is not readily apparent due to the actions of the other extraocular muscles, which mask most of the deficit. Supranuclear innervation of the trochlear nucleus is via the gaze centers in the brain stem, and the gaze centers, the three oculomotor nuclei, and vestibular nuclei are interconnected via the medial longitudinal fasciculus.

2. Oculomotor Nerve

Cranial nerve III innervates four of the six extraocular muscles of the eye as well as the levator palpebrae superioris muscle of the eyelid and, via projections to the ciliary ganglion, the small intraocular muscles that control the constriction of the pupil and the shape of the lens. Unlike the other two oculomotor nerves—the abducens nerve, which innervates the lateral rectus muscle, and the trochlear nerve, which innervates the superior oblique muscle—the oculomotor nerve has more than one functional component. It contains a GSE component for innervation of the extraocular and levator palpebrae superioris muscles and a GVE

parasympathetic component for innervation of the ganglion that in turn innervates the intraocular muscles. The oculomotor nerve fibers traverse the tegmentum and exit the brain in the interpeduncular fossa medial to the crus cerebri and then pass rostrally to the orbit with the other oculomotor nerves.

The GSE component of cranial nerve III arises from motor neurons in the oculomotor nucleus proper, which lies in the dorsomedial part of the rostral half of the midbrain tegmentum. This component innervates four muscles that insert on the globe of the eye: the inferior oblique muscle, which extorts, elevates, and abducts the eye; the medial rectus muscle, which adducts the eye; the inferior rectus muscle, which depresses, extorts, and adducts the eye; and the superior rectus muscle, which elevates, intorts, and adducts the eye. (Extorsion is rotation of the eyeball around a horizontal, anteroposterior axis through the pupil such that the bottom moves medially; intorsion is rotation in the opposite direction.)

The oculomotor nuclear complex comprises several nuclei in addition to the GSE main nucleus, including the Edinger–Westphal nucleus and several accessory oculomotor nuclei—the nucleus of Darkschewitsch, the interstitial nucleus of Cajal, and the nuclei of the posterior commissure. These accessory nuclei receive inputs from a variety of visual- and vestibular-related sources and contribute to vertical and smooth-pursuit movements of the eyes directed by the GSE component. The Edinger–Westphal nucleus and a more rostral anterior median nucleus contain preganglionic parasympathetic neurons (GVE) that project via the oculomotor nerve to the ipsilateral ciliary ganglion, which lies deep to the posterior boundary of the eye. Postganglionic parasympathetic fibers from the ciliary ganglion supply the ciliary muscle, which affects the shape of the lens for focusing on near objects, and the sphincter (or constrictor) pupillae muscle. The consensual pupillary light reflex of evoking bilateral constriction of the pupils when light is shined in one eye depends on a pathway from the retina to the olivary pretectal nucleus, which in turn projects bilaterally via the posterior commissure to the anterior median and Edinger–Westphal parasympathetic nuclei. In this context, it should be noted that dilation of the pupil is accomplished by the action of sympathetic innervation that arises from preganglionic neurons in the intermediolateral column of the spinal cord and affects the dilator pupillae muscle fibers via postganglionic neurons of the superior cervical ganglion.

Damage to the ventral region of the midbrain may result in the syndrome called superior alternating hemiplegia, which is motor impairment on the contralateral side of the body (due to damage to the corticospinal fibers in the crus cerebri) combined with lateral and downward deviation of the eye on the ipsilateral (i.e., alternate) side due to a lower motor neuron lesion of the oculomotor nerve. In this situation, the eye deviates laterally due to the unopposed contraction of the lateral rectus muscle, which is innervated by cranial nerve VI. Supranuclear innervation of the oculomotor nuclei is mediated by cortical eye fields and the superior colliculus via gaze centers in the brain stem and is coordinated among the oculomotor, trochlear, and abducens nuclei and the vestibular system via the medial longitudinal fasciculus. Damage to the oculomotor nuclear complex or nerve also results in drooping of the eyelid (ptosis) and pupillary dilation.

D. Cranial Nerves of the Forebrain

Two of the traditionally recognized cranial nerves are present in the forebrain: the optic nerve (II) and the olfactory nerve (I). Both of these nerves are purely sensory, but they are usually left unclassified according to the traditional scheme.

1. Optic Nerve

Cranial nerve II is composed of the distal parts of the axonal processes of retinal ganglion cells. These axons course caudally and medially from the retina to the optic chiasm, where some decussate (cross over) to the opposite side. The proximal parts of these same axons are then called the optic tract as they continue caudally and laterally from the chiasm to sites in the diencephalon and midbrain.

The retinal ganglion cells receive input from retinal bipolar cells, which in turn receive input from the receptor cells of the retina (rods and cones). Rods are predominantly located in the peripheral parts of the retina, whereas cones are densely packed in the central part of the retina, particularly within the fovea. Rods transduce light stimuli of a broad range of wavelengths, whereas cones are of three types for color vision, each transducing a different part of the spectrum. The transduction process is a complex series of biochemical events initiated by the absorption of a photon by pigment within the receptor cells. The visual

world topologically maps in precise order onto the retina, and this map is preserved throughout the system.

The retinal bipolar neurons correspond to the bipolar neurons that lie within the ganglia of other sensory cranial nerve components in terms of their relative position in the sensory pathway. The retinal ganglion cells are the first-order multipolar neurons of the pathway. At the optic chiasm, optic nerve fibers that arise from retinal ganglion cells in the nasal (medial) retina decussate, whereas axons from retinal ganglion cells in the temporal (lateral) retina do not. The net result is that the axons in the optic tract on the right side, for example, receive input that initiates from stimuli in the left half of the visual world. Thus, the right brain ''sees'' the left visual world and vice versa.

The optic tract projects to multiple sites in the diencephalon and midbrain. The major visual pathway for conscious vision is to neocortex via the dorsal lateral geniculate nucleus in the dorsal thalamus. This nucleus contains two large-celled (magnocellular) layers; they receive input relayed from rods via ganglion cells in the peripheral parts of the retina that conveys the general location of stimuli and their motion. It also contains four small-celled (parvicellular) layers that receive fine form and color input from cones. The dorsal lateral geniculate nucleus projects to primary (striate) cortex, which lies in the caudal and medial part of the occipital lobe. The spatial information from retinal ganglion cells via the magnocellular layers of the dorsal lateral geniculate nucleus is relayed from striate cortex via multiple synapses predominantly to posterior parietal cortex, which is involved in spatial cortical functions, whereas the form and color information via the parvicellular geniculate layers is likewise relayed predominantly to inferotemporal cortex, which is involved in numerous complex functions including the visual recognition of objects and individuals.

Midbrain visual projections are to the superficial layers of the rostral part of the midbrain roof, the superior colliculus, in which visual information is mapped in register with similar maps of somatosensory and auditory space that are projected into its deeper layers. The superior colliculus visual input is relayed to part of the pulvinar in the dorsal thalamus, which in turn projects to extrastriate visual cortical areas, which border the primary visual cortex in the occipital lobe. The midbrain visual pathway is concerned with the spatial orientation of the visual world.

Damage to the retina or optic pathway causes loss of vision in part or all of visual space (the visual field) depending on the location and extent of the lesion. Although damage to the retina or optic nerve results in blindness for the eye on the same side, damage located in the thalamocortical part of the pathway causes a deficit for the visual field on the opposite side. Damage to the central part of the optic chiasm, as can occur with a pituitary gland tumor in that region, causes ''tunnel vision''—loss of the peripheral parts of the visual field on both sides.

2. Olfactory Nerve

Cranial nerve I is composed of the set of axonal processes that arise from the olfactory bipolar cells that lie within the nasal mucosa. These sensory cells lie in a sheet-like formation across the mucosal surface rather than being condensed into a ganglion as is the case for other peripheral nerves. Unlike the situation with the visual, vestibulocochlear, and gustatory systems, there are no separate receptor cells for the olfactory system. The distal ends of the olfactory bipolar cells have the molecular machinery to bind oderant molecules and transduce the signal via various complex biochemical events. These bipolar neurons terminate on first-order multipolar neurons called mitral cells that are located within the olfactory bulbs, which lie ventral to the frontal lobes.

The bipolar cell axons form complex synaptic contacts with the mitral cell processes in a series of spherical structures called glomeruli. The glomeruli form a layer superficial to the layer of mitral cell bodies in the olfactory bulb. Different oderant molecules stimulate different sets of glomeruli, but a simple oderant ''map'' across the region of different oderant molecules has not been found. The mitral cells project via the olfactory tract to multiple sites in the ventral region of the telencephalon collectively referred to as the olfactory cortex. An olfactory pathway to part of neocortex is via a relay from olfactory cortex to the mediodorsal nucleus of the dorsal thalamus and then to prefrontal cortex, which occupies a large portion of the frontal lobe and is concerned with a multitude of complex, higher cognitive functions.

Damage to the olfactory system results in the loss of the sense of smell. The appreciation of the sensation of the flavors of food is also impaired since flavor is a complex sensation based on olfactory and gustatory interactions at the cortical level. In some cases of epilepsy caused by lesions involving the ventral, olfactory-related regions of the cerebral hemispheres (particularly a structure called the uncus), an aura

involving an olfactory hallucination may preceed seizure activity.

III. NEW INSIGHTS FROM EMBRYOLOGY AND NEW CLASSIFICATION OF THE CRANIAL NERVE COMPONENTS

Studies of embryological development, most within the past decade and including the work of LeDouarin, Gilland and Baker, Noden and colleagues, Northcutt, Gans, and others, require several changes to our treatment of cranial nerves. Two additional cranial nerves need to be recognized in humans—a terminal nerve (T) and an epiphyseal nerve (E)—and the two diencephalic cranial nerves (E and II) can be classified as neural tube afferent (NTA). The traditional classification of cranial nerve components needs to be revised and simplified. The distinctions between the "general" and "special" categories can now be eliminated, and the components of spinal nerves and most cranial nerves can be reduced to the four simple categories of somatic afferent (SA), visceral afferent (VA), somatic efferent (SE), and visceral efferent (VE). The branchial motor nerves form a subset of the SE column since the branchial muscles are derived from the same mesodermal source as are the extraocular muscles and the muscles of the tongue.

A. Terminal Nerve

The terminal nerve is arguably the most rostral cranial nerve. Its bipolar neurons have free nerve endings that are located in the nasal mucosa, but their modality (possibly chemosensory) has not been established. These neurons project to several sites in the ventral and medial regions of the forebrain, including the septum, olfactory tubercle, and preoptic area. The terminal nerve neurons contain the reproductive hormone, luteinizing hormone-releasing hormone, and may be involved in the regulation of reproductive behavior.

B. Sensory Cranial Nerves Derived from the Neural Tube

The optic nerve (II) and neural retina develop as an outgrowth from the diencephalon. The optic nerve is thus in fact not a true cranial nerve but rather a tract of the central nervous system. Unlike most sensory cranial nerves that have receptors and/or bipolar neurons in the peripheral nervous system, the receptor rods and cones of the retina and the retinal bipolar cells are within the central nervous system. Nevertheless, the optic nerve is so firmly and universally considered to be a cranial nerve that eliminating it from the list is not an option. A second, similar tract of the central nervous system that innervates a diencephalic outgrowth, the epiphyseal nerve, thus also needs to be included in the list of cranial nerves for completeness.

The epiphysis forms from the roof plate of the diencephalon and comprises a variety of structures in various vertebrates, including a frontal organ and pineal gland in frogs, a parietal eye and pineal gland in lizards, and a pineal gland in mammals. The pineal gland does not receive light directly, as do the neural retina and the parietal eye, but it is influenced by light via a pathway from the suprachiasmatic nucleus of the hypothalamus to the sympathetic intermediolateral column of the spinal cord and then via postganglionic sympathetic fibers that arise from the superior cervical ganglion and travel on branches of the internal carotid artery to reach the pineal gland, where the sympathetic input inhibits the conversion of serotonin to melatonin. In mammals, the pineal gland contains neurons that project to the medial habenular nucleus and part of the pretectum. Pineal projections to other brain sites are minor in humans, but recognition of an epiphyseal cranial nerve is justified based on its similarities of development and organization with the optic nerve. Neither the optic nerve nor the epiphyseal nerve can be categorized as somatic or visceral sensory, however. Since both derive from the neural tube, they can be classified as NTA. They are unique among the cranial nerves in this regard.

C. Afferent Components

Components of six cranial nerves could arguably be assigned to the traditional category of special afferent. The facial (VII), glossopharyngeal (IX), and vagus (X) nerves have SVA components for taste sensation. The olfactory nerve (I) is sometimes classified as SVA, and the terminal nerve could be provisionally included in this category as well. The vestibulocochlear nerve (VIII) is traditionally classified as SSA. Of the nondiencephalic sensory cranial nerves, only the trigeminal nerve (V) and small components of VII, IX, and X have been classified as GSA. The fourth traditional afferent category of GVA has been

assigned to the neurons of the glossopharyngeal and vagus nerves that predominantly innervate the visceral structures of the body cavities. What all these sensory cranial nerve components have in common is that their bipolar neurons are embryologically derived from the same two tissues: neural crest and/or placodes.

1. Neural Crest and Placodes

During embryological development, the neural tube forms by a process of invagination of the ectoderm along the dorsal midline of the body and the head. At the edges of the invaginating neural tube tissue, a population of neural crest cells arises from the dorsomedial wall of the neural folds and migrates laterally and ventrally to contribute to many different structures of the adult, including the dorsal root ganglion cells (bipolar neurons) for all the spinal cord sensory nerves. In the head region, neural crest cells likewise migrate away from the region of the neural tube and give rise to multiple structures. A second tissue present only in the head region is a set of neurogenic placodes, which are thickened patches of ectodermal cells. The bipolar sensory neurons for all the cranial nerves except II and the epiphyseal are derived from the neurogenic placodes and/or neural crest tissues.

In the region of the nose, an olfactory placode forms that gives rise to the bipolar neurons of the olfactory and terminal nerves, and additional sets of placodes form along the sides of the head. Neural crest and/or placodally derived cells give rise to the bipolar neurons for the sensory ganglia of the trigeminal (V), facial (VII), vestibulocochlear (VIII), glossopharyngeal (IX), and vagus (X) nerves. A trigeminal placode contributes neurons to the trigeminal ganglion. The sensory neurons within the vestibular and spiral ganglia for the vestibulocochlear nerve arise from a dorsolateral, otic placode; this is a much simpler situation than in fishes, in which a series of dorsolateral placodes gives rise to the vestibulocochlear nerve and to the set of mechanoreceptive and electroreceptive lateral line nerves. The receptor hair cells for the vestibulocochlear system likewise are derived from this dorsolateral placode.

The gustatory system is more variable across vertebrates in terms of both its peripheral and its central components. Mammals have a relatively limited gustatory system. A set of ventrolateral, or epibranchial, placodes contributes sensory neurons to ganglia of cranial nerves VII, IX, and X, but, unlike the situation in the vestibulocochlear system, the taste receptor cells of the taste buds are all derived locally from the pharyngeal endoderm in the oral cavity. This system thus is classified here as visceral. Fishes, in comparison, have much more elaborate taste systems than mammals, and in some groups, such as catfishes, the taste receptor cells are distributed over the entire body surface. The gustatory afferent fibers project into elaborate, laminated lobes in the brain stem and preserve a topographic map of the inputs.

2. Somatic Afferent and Visceral Afferent Cranial Nerve Components

Due to these recent findings on the embryological derivation of these cranial nerve components, their former classification into four separate categories of SVA (I the terminal nerve, and the components of VII, IX, and X for taste sensation), SSA (VIII), GSA (V, VII, IX, and X), and GVA (IX and X) can be revised (Table III). Since the bipolar sensory neurons for all these nerves are derived from the same embryonic source (i.e., neural crest and/or placodes), the "special" versus "general" distinction is not warrented. We are thus left with the resolution of classifying most of the cranial nerve sensory components (i.e., the traditional GSA and SSA components plus I and the terminal nerve) simply as somatic afferent (SA). Due to the derivation of the receptor cells from endoderm, the taste components of VII, IX, and X are assigned to the simple category of visceral afferent (VA). The remaining sensory category of GVA for innervation of the thoracic and abdominal viscera and pharyngeal region by the vagus nerve and of the tongue and pharyngeal region by the glossopharyngeal nerve can likewise be revised to the VA category.

D. Efferent Components

Some of the motor components of four cranial nerves have traditionally been assigned to the category of SVE—the trigeminal (V), facial (VII), glossopharyngeal (IX), vagus (X), and spinal accessory (XI) nerve components that innervate muscles derived from the branchial arches. Four purely motor cranial nerves classified as GSE—III, IV, VI, and XII—innervate eye and tongue muscles. Recent embryological findings, particularly from the work of Noden and collaborators, have shown that the nerve components of the SVE and GSE categories both innervate striated skeletal muscles that are all derived from the same

embryonic muscle masses in the developing head and neck regions. Components of the third efferent category, GVE, are present in cranial nerves III, VII, IX, and X and give rise to preganglionic parasympathetic fibers for innervation of the ganglia that supply glands in the head and the viscera of the thorax and abdomen.

1. Epimere versus Hypomere and the Development of Striated Muscles in the Head

To appreciate the resolution of the problem of head musculature and its innervation, one must consider it in the context of the pattern of development of the muscular components of the body. The body is essentially a tube with three layers—an outer layer of ectoderm that forms the central nervous system, the neural crest, the epidermis, and other tissues; an inner layer of endoderm that forms the lining of the gut and related organs; and an intermediate layer of mesoderm that forms the bones, muscles, blood vessels, dermis, and other structures of the body. The mesoderm has three dorsal-to-ventral components. The epimere (or paraxial mesoderm), the most dorsal component, forms somites, which are segmental mesodermal masses that give rise to the striated, skeletal muscles of the body wall and the limbs. The intermediate component, mesomere (or nephric ridge), gives rise to the kidneys and gonads, whereas the ventral component, hypomere (or lateral plate mesoderm), gives rise to the smooth, visceral muscles of the gut as well as to the cardiac muscle of the heart.

In the head, the embryological derivation of the various groups of muscles was until recently incorrectly understood, particularly concerning the pharyngeal region. In addition to giving rise to multiple components of the peripheral nervous system and to most of the skull, neural crest cells give rise to mesenchymal tissue that forms the so-called ''visceral'' arches of the pharynx. The muscles of this region were long thought to arise from a rostral continuation of hypomeric muscle—the ventral part of the mesodermal tissue that in the body gives rise to the smooth muscle of the gut. Muscles derived from the visceral arch region, which in mammals are the muscles of the mandibular arch (lower jaw muscles innervated by the trigeminal nerve), hyoid arch (facial and other muscles innervated by the facial nerve), and the several branchial arches (throat and larynx muscles innervated by the glossopharyngeal and vagus nerves), were thus presumed to be visceral muscles. In conflict with this view was the resemblance of the visceral arch muscles to the striated, skeletal muscle of the body wall and of the tongue and extraocular muscles. The histological structure of the visceral arch muscles and their single-motor neuron innervation pattern are both markedly different from the histological structure and two-neuron chain, parasympathetic innervation pattern of the smooth muscle of the gut. The designation of ''special'' acknowledged these differences.

Recent work by Gilland, Noden, Northcutt, and others has revealed a different embryological source for the muscles of the visceral arches and has resolved the previous confusion. Rather than being derived from hypomere, the muscles of the visceral arches are in fact derived from the epimeric muscle of the head, the caudal continuation of which gives rise to the striated, skeletal muscles of the body wall. Somites and somitomeres, which are incompletely separated somites, form in the head region and give rise to the extraocular muscles, the muscles of the visceral arches, and the muscles of the tongue. Thus, none of the visceral arch muscles are actually visceral, and there is nothing special about them. They are embryologically derived and innervated in the same manner (by neural tube-derived motor neurons) as the extraocular and tongue muscles. The visceral arch muscles probably evolved before the extraocular muscles were acquired by early vertebrates, however, and long before the muscular tongue was gained by tetrapods, and this difference in history may account for the different position within the brain stem of the two columns of motor neurons that innervate them.

2. Somatic Efferent and Visceral Efferent Cranial Nerve Components

Due to these recent findings on the embryological derivation of the visceral arch muscles, the former classification of their motor cranial nerve components of SVE can be revised. These components can be classified as somatic, just like those that supply the extraocular muscles and the tongue. Since the classification of visceral is invalid, the distinction of special (for V, VII, IX, X, and XI) versus general (for III, IV, VI, and XII) is also moot. All these motor components are simply somatic efferents (Table III). For the remaining motor category, GVE, the parasympathetic components of cranial nerves III, VII, IX, and X are indeed visceral efferents, and with the elimination of the special visceral category the use of the term general is not needed here. These parasympathetic components can simply be classified as visceral efferents.

See Also the Following Articles

BRAIN ANATOMY AND NETWORKS • NERVOUS SYSTEM, ORGANIZATION OF • NEUROANATOMY • PERIPHERAL NERVOUS SYSTEM

Suggested Reading

Butler, A. B., and Hodos, W. (1996). *Comparative Vertebrate Neuroanatomy: Evolution and Adaptation*. Wiley–Liss, New York.

Gilland, E., and Baker, R. (1993). Conservation of neuroepithelial and mesodermal segments in the embryonic vertebrate head. *Acta Anat.* **148,** 110–123.

Haines, D. E. (1997). *Fundamental Neuroscience*. Churchill–Livingstone, New York.

Liem, K. F., Bemis, W. E., Walker, W.F., Jr., and Grande, L. (2001). *Functional Anatomy of the Vertebrates: An Evolutionary Perspective*, 3rd ed. Harcourt, Fort Worth, TX.

Nieuwenhuys, R., Ten Donkelaar, H. J., and Nicholson, C. (1998). *The Central Nervous System of Vertebrates*. Springer, Berlin.

Noden, D. M. (1991). Vertebrate craniofacial development: The relation between ontogenetic process and morphological outcome. *Brain Behav. Evol.* **38,** 190–225.

Noden, D. M. (1993). Spatial integration among cells forming the cranial peripheral nervous system. *J. Neurobiol.* **24,** 248–261.

Northcutt, R. G. (1993). A reassessment of Goodrich's model of cranial nerve phylogeny. *Acta Anat.* **148,** 150–159.

Northcutt, R. G. (1996). The origin of craniates: Neural crest, neurogenic placodes, and homeobox genes. *Israel J. Zool.* **42,** S273–S313.

Northcutt, R. G., and Gans, C. (1983). The genesis of neural crest and epidermal placodes: A reinterpretation of vertebrate origins. *Q. Rev. Biol*. **58,** 1–28.

Parent, A. (1996). *Carpenter's Human Neuroanatomy*, 9th ed. Williams & Wilkins, Baltimore.

Webb, J. F., and Noden, D. M. (1993). Ectodermal placodes: Contributions to the development of the vertebrate head. *Am. Zool.* **33,** 434–447.

Young, P. A., and Young, P. H. (1997). *Basic Clinical Neuroanatomy*. Williams & Wilkins, Baltimore.

Creativity

MARK A. RUNCO
California State University, Fullerton, and University of Hawaii, Hilo

GLOSSARY

alexithemia Lack of affect; low emotionality. This often characterizes the individual with a "split brain" and seems to inhibit his or her creativity.

creativity complex Influences on creativity reflect personality, attitude, affect, cognition, and metacognition. Creativity is not unidimensional or tied to a single domain.

divergent thinking Problem solving and ideation which moves in different directions rather than converging on one correct or conventional idea or answer.

intrinsic motivation The personal interests that often lead to creative work and are independent of extrinsic influences, such as rewards and incentives.

range of reaction Genetic potential delimits the range of potential; environmental influences determine where within that range the individual performs.

split brain Lay term for the result of a commisurotomy, which surgically separates the two hemispheres of the brain.

Creativity is important, and probably vital, for innovation, technological progress, and societal evolution. It may be more important now than ever before because it plays a role in adaptations. Creativity also benefits the individual, providing coping skills and a means for self-expression. Very likely the creativity of individuals determines the creative potential of society. Just as likely, the creative potentials of individuals is dependent on brain structure and process.

I. INTRODUCTION

The brain and creativity are typically studied from very different perspectives. Simplifying, the human brain is a topic for the hard sciences, including neurology and medicine. Creativity, on the other hand, is traditionally studied by social and behavioral scientists. This makes it more difficult to bring the two topics together and may explain why there is not more literature on the brain and creativity. Efforts are being made, especially very recently, to bridge the two topics. In fact, researchers are discovering advantages to this kind of cross-disciplinary investigation.

One advantage is conceptual. Useful concepts and ideas can be borrowed from one field and applied to the other. Consider the concept of a range of reaction. This was originally proposed in the hard sciences (genetics) but is vital for our understanding of virtually all behavior (psychology), including creative behavior. As is the case with the biological contributions to other human characteristics, biology contributes to creativity by providing a range of possible expressions. This range delimits our potential, but it guarantees nothing. The environment determines how much of that potential we fulfill. Family experiences seem to influence the fulfillment of creative potentials, just to name one example of how experience interacts with biology.

The range of potential delimits the levels of performance in simple terms, the amount of creative talent the individual will express but also the domain of expression. This is an important point because creativity in the visual arts is very different from musical creativity, and both of these are very different from mathematical creative talent. There are many domains and possible areas of expression, and each very likely has particular biological underpinnings. An individual may inherit the potential to excel in music but not mathematics or to excel in languages but not the performing arts. It is quite possible that the different domains of creative expression are associated with different brain structures.

Such domain specificity further complicates matters, especially for investigations focused on creativity. This is because creativity is often a kind of self-expression, and it is frequently maximal when the individual is intrinsically motivated. This simply means that the individual is expressing himself or herself in an original fashion or putting the effort into solving a problem in a creative manner, not because of incentive or reward (both of which are extrinsic) but because of personal interests. However, what if those interests do not fit well with the biological potentials? What if the individual is intrinsically motivated to study music but biologically predisposed to excel in a field that involves very different talents, such as science? We can hope that most persons inherent a potential to be interested in the same domain in which they have inherited the potential for relevant skills, but we should also accept the strong likelihood that some persons have unfulfilled creative potentials. This is all the more reason to examine the brain and creativity.

Creativity is multifaceted. There are cognitive skills involved in all creative acts, but these are also attitudes that can facilitate or inhibit creative performance. Personality has long been recognized as an important determinant, and increasingly more is being done with the emotions that play significant roles in creative work. Creativity is often operationalized for research or practice (e.g., education or organizational efficacy) in a fairly simplistic fashion, with an emphasis on divergent thinking or the production of original ideas. However, definitions that focus on one or two traits or abilities do not do justice to the creative efforts that have literally changed each of our lives through technological innovation or the artistic perspective. Often, in the research creativity is called a syndrome or complex because of its multifaceted nature. We must keep this in mind as we discuss the specific research findings on the brain and creativity. Most of the research focuses on one kind of creativity or one of the relevant skills. I previously noted that biological potentials do not guarantee that the talent will be expressed, and the same thing must be said about any connection we discover between the brain and a particular facet of the creativity complex: That is no guarantee of actual creative performance.

The creativity complex comprises cognitive, attitudinal, and emotional components. These are described in this article and tied to brain structure and function.

II. SPLIT BRAIN

Probably the best known research on the brain and creativity is that of Roger Sperry. He was awarded the Nobel prize for his research with individuals who had commisurotomies (i.e., the corpus callosum is cut, thereby functionally disconnecting the two hemispheres of the brain). Sperry did not administer any tests of creativity per se to the commisurotomy patients, but his research is very relevant because he found that certain important alogical and simultaneous processes tended to be controlled by the right hemisphere. Logical and linguistic processes seemed to be controlled by the left hemisphere. This is quite telling, although we must be careful not to make the mistake that is made in the popular press.

Sperry studied more than two dozen individuals, and not all of them demonstrated the dramatic differences between the right and left hemispheres. Additionally, Sperry worked with epileptic individuals who had the surgery because of this disease. It was an attempt to reduce the number of grand mal seizures. Surely epileptic persons have atypical nervous systems. They also had commissurotomies, further precluding generalizations from Sperry's research. Finally, it is simplistic and inappropriate to view creativity as a solely right hemisphere function. Clearly, individuals cannot use one hemisphere or the other. None of us can do that, unless of course you are one of Sperry's patients and have had a commissurotomy. Nor should we try to rely on one hemisphere: Creativity involves both logic and imagination, divergent and convergent thinking.

Sperry's patients have been studied many times. Klaus Hoppe, for example, reported what may be some of the most relevant findings about individuals with split brains—namely, that they often have

trouble understanding their own emotional reactions, if they have them. In "Dual Brain, Creativity, and Health," he and Neville Kyle referred to this as alexithymia. It is especially significant because creativity so often involves enthusiasm, excitement, and the like.

Hoppe and Kyle's interpretation focuses on the corpus callosum. They described how

> *commisurotomy patients, in comparison with normal controls, use significantly fewer affect laden words, a higher percentage of auxiliary verbs, and applied adjectives sparsely revealing a speech that was dull, uninvolved, flat, and lacking color and expressiveness. … Commissurotomy patients tended not to fantasize about, imagine, or interpret the symbols, and they tended to describe the circumstances surrounding events, as opposed to describing their own feelings about these events. … Commisurotomy patients, in comparison with normal controls, symbolized in a discursive, logically articulate structure, using mainly a secondary process, as opposed to a presentational structure as an expression of a predominantly prominent process. They also showed a concreteness of symbolization, emphasizing low rather than creative capacity… showed a relatively impoverish fantasy life, and tended not to be able and convey symbolic meanings.*

III. ELECTROENCEPHALOGRAM PATTERNS

Electroencephalogram (EEG) data were also collected. These indicated that the subjects had low activity in the right temporal area (T4) when listening to music and viewing an emotionally stimulating film. They also had low levels of activity in the left hemisphere, particularly in Broca and Wernickes' areas (F3 and T3). Hoppe and Kyle believed that this suggested a lack of inner speech. The same patients had high levels of activity in the left parietal area (P3).

Perhaps most important was the fact that the patients had what Hoppe called a high coherence between the left parietal area (P3) and the right frontal area (F4). This was interpreted as "a possible interhemispheric aspect of inhibition of expression." The control subjects who were more emotional and expressive, in contrast, had coherence between the left temporal (T3) and the right frontal (F4) areas, which suggested "a possible mechanism facilitating the transformation of the effective understanding in the right hemisphere into verbal expression of the lower left hemisphere."

Hoppe and Kyle believe strongly that communication between the hemispheres is necessary for creative thinking. Earlier, Salvatore Arieti described "a magic synthesis," the title of his 1976 book, and Arthur Koestler proposed in his book *The Art of Creation* that creativity could be understood as bisociation. Both of them were pointing to the need for communication between and collaboration by the two hemispheres of the brain. There are other relevant data. Most, however, involve indirect measurement.

For example, there are studies of handedness, which is supposedly an indication of hemispheric dominance and preference. Yet other studies use dichotic listening tasks wherein particular messages are played to one ear or the other, and thereby presented to one hemisphere or the other. In a 1999 review of this research, Albert Katz indicated that the findings were not all that convincing, although they were in the direction in what would be expected with a right hemisphere contribution to creative thinking. It is very likely that such indirect measures will not be used very frequently or very much longer, given new technologies for brain imaging and the like. Importantly, Katz, like Hoppe and others, concluded that

> *creative activity cannot be localized as a special function unique to one to the cerebral hemispheres. Rather, productive thought involves the integration coordination of processes subserved by both hemispheres. … There appears to be privileged role in creativity to the cognitive functions associated with the right hemisphere. … There is some evidence that different creative tasks may differentially call on cognitive resources for which the two hemispheres were specialized. … Finally, there is some evidence that creativity related hemispheric asymmetries can be found both online (as the person performs a task) and as a consequence of habitual patterns of behavior.*

IV. STAGES OF THE CREATIVE PROCESS

Colin Martindale, like Hoppe, employed EEGs. Martindale examined particular brain wave patterns and their association with specific stages within the problem-solving process. Findings indicated that EEG patterns vary as a function of stage of the creative process, at least in creative individuals. These phases are usually labeled preparation, incubation,

illumination (or inspiration), and verification. Creative individuals have higher alpha during the inspirational phase of creative work than during elaboration phases. There are no differences in basal alpha. EEG was measured at right posterior temporal areas.

Another technique for the study of the biology of creativity involves individuals who have had accidental trauma to the nervous system. In his 1982 book, *Art*, *Mind*, *and Brain*, Howard Gardner tells how through use of this technique he was able to infer much about processes that are very important for creativity, such as the use of metaphor. This kind of research is useful because it identifies particular parts of the brain. However, the locations studied are not chosen by the researchers but of course are the result of various unfortunate accidents.

V. MAGNETIC IMAGING

We probably need not look to accidents much longer. We now have systematic techniques, such as magnetic imaging. In a 1995 investigation, Thomas Albert, Christo Pantey, Christian Wiendruch, Bridget Rockstroh, and Edward Taub found that the cortical representation of the digits of the left hand of string players was larger than that in controls. The effect was smallest for the left thumb and no such differences were observed for representations of the right hand digits. The amount of cortical reorganization in the representation of the fingering digits was correlated with the age at which the person had begun to play. These results suggested that the representation of different parts of the body and the primary somatosensory cortex of humans depends on use and changes to conform to the current needs and experiences of the individual.

Recall the idea introduced earlier in this article, that performance often reflects both biological potential and experience. Note also the fact that this research focused on individuals within a very specific domain of performance (stringed instruments). Generalizations cannot be applied to other instruments, let alone other kinds of musical creativity or other kinds of art. Finally, it is critical to keep in mind that although the participants of this study were musicians, their actual creativity was not assessed. Music may be an unambiguously creative domain, but it would be interesting to know the actual level of creative skill of the individuals and to correlate that with the size of the cortical representations.

VI. EINSTEIN'S BRAIN

Given the complexity of creativity and the diversity of definitions that have been proposed, some scholars prefer to focus on unambiguously creative persons. The advantage of this approach is that the creativity of the research subjects (e.g., Darwin, Picasso, and Einstein) is beyond question.

This approach is often used in the psychological research but has also been employed in physiological research, such as the 1985 study of the brain of Albert Einstein by Marian Diamond, Arnold Scheibel, Greer Murphy, and Thomas Harvey. These authors found that Einstein's brain had a significantly smaller mean ratio of neuron to glial cells (connections) in "area 39" of the left hemisphere than did control scientists. This was not the case in three other areas. It was not true of the right hemisphere. The interpretation focused on the "metabolic need" of Einstein's cortex and the role of the cortex in associative thinking. Diamond and his colleagues concluded that the exceptionality of Einstein's brain may have given him outstanding "conceptual power."

VII. CONCLUSIONS

One area of the brain may be involved in creativity but has received little attention. More than one theorist has pointed to the prefontal lobes. Salvatore Arieti did so long ago, and others have echoed it, often citing tendencies involved in creativity and controlled by the prefrontal lobes. Elliott, for example, argued that creative behavior is a product of the human capacity to will. He believed that this implies that the prefrontal lobe must be engaged, "thereby facilitating the harmonious functioning of the entire brain (left–right, top–bottom, front–back) and thus regulating all psychological functions associated with the creative process." In 2000, Torsten Norlander came to a similar conclusion but drew from what is known about the prefrontal dysfunction and hypofrontality of schizophrenic patients. (Data include the cerebral regional blood flow of schizophrenic patients.) Although largely inferential, this line of thought is noteworthy in part because it is consistent with the idea of balanced processes contributing to creative thinking. The balance may require communication between the left and right hemispheres of the brain, and between logic and imagination (the magic synthesis), but it may also benefit from collaboration with the prefrontal lobes.

Empirical research on the prefrontal lobes and their role in creative thinking is needed. Fortunately, we can

expect a great deal of progress to be made in the very near future. There are two reasons for this. First, the field of creative studies in growing explosively. The implications of creativity for health and learning are now widely recognized, as is the role of creativity in innovation, business, and technological advancement. Even more important may be the technological advances that are being made in medicine. These provide sophisticated methods for studying the brain. Soon, we should understand how the different parts and processes of the human brain correspond to the various components of the creativity complex.

See Also the Following Articles

ARTIFICIAL INTELLIGENCE • CONSCIOUSNESS • EMOTION • HUMOR AND LAUGHTER • INTELLIGENCE • LANGUAGE ACQUISITION • LOGIC AND REASONING • MUSIC AND THE BRAIN • PROBLEM SOLVING • UNCONSCIOUS, THE

Suggested Reading

Albert, T., Pantey, C., Wiendruch, C., Rockstroh, B., and Taub, E. (1995). Increased cortical representation of the fingers of the left hand in string players. *Science* **270,** 305–307.

Arieti, S. (1976). *The Magic Synthesis*. Basic Books, New York.

Diamond, M. C., Schiebel, A. B., Murphy, G. M., and Harvey, T. (1985). On the brain of a scientist: Albert Einstein. *Exp. Neurol.* **88,** 198–204.

Elliott, P. C. (1986). Right (or left) brain cognition, wrong metaphor for creative behavior: It is prefrontal lobe volition that makes the (human/humane) difference in release of creative potential. *J. Creative Behav.* **20,** 202–214.

Hoppe, K. D., and Kyle, N. L. (1990). Dual brain, creativity, and health. *Creativity Res. J.* **3,** 150–157.

Katz, A. (1997). *Creativity in the cerebral hemispheres*. In *Creativity Research Handbook* (M. A. Runco, Ed.), pp. 203–226. Hampton Press, Cresskill, NJ.

Martindale, C., and Hasenfus, N. (1978). EEG differences as a function of creativity, stage of the creative process, and effort to be original. *Biol. Psychol.* **6,** 157–167.

Runco, M. A., and Albert, R. S. (2001). *Theories of Creativity*, 2nd ed. Hampton Press, Cresskill, NJ.

Dementia

ALFRED W. KASZNIAK
University of Arizona

GLOSSARY

aphasia Impairment in the expression and/or comprehension of language.

cognitive functions Distinct domains of intellectual ability, including memory, language, visuospatial ability, judgment, and abstraction, among others.

depression A clinical disorder characterized by at least five of the following nine symptoms present during at least a 2-week period: depressed mood, markedly diminished interest in activities, significant weight loss or gain, difficulty sleeping or excessive sleeping, agitation or slowing of thought and action, fatigue or loss of energy, feelings of worthlessness or guilt, diminished ability to think or concentrate, and suicidal ideation.

executive functions That subgroup of cognitive functions necessary to plan and carry out complex, goal-oriented behavior. Executive functions include planning, organizing, sequencing, and abstracting abilities.

neuroimaging Technologies available for the detailed visualization of brain structure and/or function. These technologies include computerized tomography (CT), magnetic resonance imaging (MRI), functional magnetic resonance imaging (fMRI), positron emission tomography (PET), and single photon emission computed tomography (SPECT).

neuropsychological assessment An approach to the evaluation of dementia and other neurobehavioral disorders involving structured interviewing, behavioral observation, and the administration of standardized tests of cognitive and emotional functions.

perseveration Persistence in a particular action or thought after task demands have changed and the action or thought is no longer appropriate.

prosopagnosia Deficit in the ability to recognize familiar faces.

verbal fluency Ability to quickly generate words from within a particular semantic category (e.g., animal names, items that can be found in a grocery store) or beginning with a particular letter of the alphabet.

The clinical syndrome of dementia is characterized by impairment of multiple cognitive functions (e.g., memory, language, visuospatial abilities, judgment), typically due to chronic or progressive brain disease. The cognitive impairments are frequently accompanied by personality changes, including deficits in motivation and emotional control. Several different illnesses can cause the dementia syndrome, and the pattern of cognitive deficits and personality changes may differ by the type of illness as well as by the specific brain structures and systems most affected. This article will present an overview of the more common dementia syndromes, emphasizing their epidemiology, clinical presentation, and pattern of neuropsychological deficits. Neuropathological characteristics and neurobiological research concerning causes and treatments are mentioned only briefly because these aspects are specifically reviewed in other articles (e.g., Alzheimer's Disease) within this volume.

I. CHARACTERISTICS AND EPIDEMIOLOGY OF DEMENTIA

Most diagnostic criteria for the dementia syndrome stipulate that two or more cognitive functions must be sufficiently impaired so as to interfere with social and/or occupational functioning and that there must not be clouding of consciousness. Clouding of consciousness, particularly in the context of a sudden onset of confusion, disorientation, hallucinations, disturbance in attention, or marked behavior change, is typically indicative of delirium (also termed acute confusional state). Delirium may be caused by an acute or chronic systemic illness (e.g., bacterial infection, hypoglycemia), adverse effects of medication, or serious neurological event requiring immediate medical attention. Untreated, these illnesses may result in death or irreversible impairment.

The risk of developing dementia increases markedly with age in later adulthood and is the single most prevalent category of mental illness for older persons. Worldwide, between 10 and 15% of persons over the age of 65 show at least mild dementia and approximately 6% show severe dementia. After age 65, dementia prevalence doubles approximately every 5 years. Overall, dementia is somewhat more common among women and may be more common in community-dwelling older black Americans. Prevalence for other nonwhite ethnic and racial groups is difficult to estimate due to the small number of adequate epidemiological studies. Annual incidence of new cases of dementia is also age-related, estimated to steadily increase from 0.33% at age 65 to 8.68% for those 95 years of age and older. A family history is a risk factor for dementia, increasing with the number of first-degree relatives similarly affected. Low educational background or low lifetime occupational attainment also increases the risk of dementia diagnosis. Studies have suggested that higher levels of education and higher levels of linguistic ability in early adulthood may be protective against the development of dementia, although controversy concerning this conclusion still remains.

A. Prevalence of Specific Dementia Types

More than 50 different illnesses can produce the symptoms of dementia. Available studies indicate that, on average, 5% of all causes of dementia are reversible and 11% have some specific treatment available, although not typically resulting in symptom reversal. Among the more common potentially reversible causes are those due to prescription and nonprescription drug toxicity, metabolic disorder, brain tumors, subdural hematoma (a collection of blood under the outermost meningeal covering of the brain), and depression. The more common dementia types that are presently irreversible include Alzheimer's disease, Parkinson's disease, dementia with Lewy bodies, Huntington's disease, frontotemporal dementias, vascular dementia, and traumatic brain injury.

Autopsy studies have varied in their estimates of the relative proportion of all dementia cases accounted for by each of these causes or types, although all agree that Alzheimer's disease (AD) is the most common. The neuropathological characteristics of AD consist primarily of neuritic plaques and neurofibrillary tangles, as well as neuronal loss. The neuropathology appears to be focused in the medial temporal lobe regions (including the hippocampus, entorhinal cortex, and subiculum) early in the illness course, with progression to the frontal, temporal, and parietal association cortices. Neuronal loss is also seen within the subcortical neurons of the nucleus basalis of Meynert (a major source of cholinergic neurotransmitter input to widespread cortical areas) and in the locus ceruleus (a major source of widespread noradrenergic neurotransmitter projections). Eventually, diffuse neuronal loss results in gross cerebral atrophy, which can easily be seen in autopsied brains, and in structural neuroimaging evidence of ventricular and sulcal enlargement, such as that provided by computerized tomography (CT) and magnetic resonance imaging (MRI; see Fig. 1). Across several studies, AD alone has been found to account for between 53 and 66% of total dementia cases and may be present in combination with the neuropathology of other diseases in as many as 87%. Vascular disease, typically in the form of multiple infarctions of blood vessels (blockages, with resultant tissue death in the area supplied by the vessel), had previously been thought to be the second most frequent cause. However, more recent studies suggest that multiple infarctions alone may account for less than 2% of all dementias, although contributing to dementia severity, along with coexistent AD, in approximately 13%.

Lewy body disease has only relatively recently been recognized as a significant cause of dementia, and earlier autopsy studies typically did not include it. Lewy bodies are microscopic intracellular abnormalities seen in the brain stem structures of patients with Parkinson's disease but are also found distributed

Figure 1 Magnetic resonance image (MRI) showing the greatly enlarged lateral ventricles characteristic of cerebral atrophy in later stage Alzheimer's disease.

diffusely throughout the cortex and subcortex when associated with dementia. Autopsy studies have suggested that dementia with Lewy bodies (DLB) might account for 15–30% of all cases of dementia, which would make it the second largest subgroup after AD.

The frontotemporal dementias (FTD) include both Pick's disease and non-Pick's neuropathology and account for about 3.5% of all dementia cases. Pick's disease is characterized by Pick's cells (enlarged neurons with displaced nuclei) and Pick's inclusion bodies (round inclusions within the neuron cell body) in frontal and temporal cortical areas. Non-Pick's neuropathology of FTD includes microvacuolation (small holes or vacuoles) or spongiosus (spongelike softening), predominantly in the frontal cortex.

Other causes, including Huntington's disease, Parkinson's disease, corticobasal degeneration (all considered among the frontal–subcortical dementias), depression, and several very rare causes, account for the remaining percentage of dementia cases. Within the following sections, the clinical and neuropsychological features characterizing the major dementia types will be described.

II. ALZHEIMER'S DISEASE

Alzheimer's disease (AD) is a chronic, progressive degenerative brain disease. The most striking behavioral changes of AD include increasing difficulty in learning and retaining new information, handling complex tasks, reasoning, spatial ability and orientation, language, and social behavior. In 1984, a task force assembled by the United States Department of Health and Human Services developed the diagnostic criteria for AD shown in Table I. These criteria now form the basis for standard AD diagnosis in research and have been incorporated into widely used clinical diagnostic criteria (e.g., those of the *American Psychiatric Association Diagnostic and Statistical Manual of Mental Disorders*, 4th ed.). These criteria have been shown to achieve good sensitivity (against the criterion of neuropathological confirmation), although more limited specificity, in the clinical diagnosis of AD.

Early in the course of AD, the most sensitive indicators of cognitive impairment are standardized tests of memory (particularly delayed recall) for recently presented verbal or nonverbal information. This is not surprising given that neuropathology in hippocampal and entorhinal cortices appears early in the illness course and that rapid forgetting is the hallmark of focal damage to these structures in both human and nonhuman animal studies. Several studies now provide evidence that AD has a long, preclinical (i.e., prior to the time when the patient meets standard diagnostic criteria) course, during which subtle memory changes occur. Mild memory impairments, measurable with sensitive neuropsychological tests, thus have been documented in apparently well-functioning persons several years prior to their diagnosis of AD. Similar memory impairments are seen in those presently healthy persons at genetic risk (i.e., showing the apolipoprotein e4 gene allele, a known risk for AD). Of persons demonstrating mild cognitive impairment (MCI), typically defined as standardized memory test performance that is one standard deviation below age- and education-specific normative expectations, between 6 and 25% per year progress to meet diagnostic criteria for dementia (most typically AD). However, preclinical AD is not the only cause, because those with MCI also show higher rates of overall poorer health, cerebrovascular disease, and disability and are more often depressed than those with normal memory functioning.

Also relatively early in the course of AD, individuals tend to show decreased verbal fluency, best documented by an ability to produce words within a given

Table I
Alzheimer's Disease (AD) Diagnostic Criteria

A clinical diagnosis of *probable* AD is made when:

(1) Dementia is established by clinical examination and documented by performance on a standardized mental status examination, *and*

(2) confirmed by neuropsychological testing, documenting deficits in two or more areas of cognition, *and*

(3) characterized by a history of progressive worsening of memory and other cognitive deficits, *with*

(4) no disturbance in level of consciousness, *and*

(5) symptom onset between the ages of 40 and 90 years, most typically after age 65, *and*

(6) there is an absence of systemic disorders or other brain diseases that of themselves could account for the progressive deficits.

The diagnosis of probable AD is further supported by evidence (i.e., from neuropsychological reexamination) of progressive deterioration of specific cognitive functions, impairment in activities of daily living (ADLs), a family history of similar disorders, and results of particular laboratory tests (e.g., a normal lumbar puncture, normal pattern or nonspecific changes in the electroencephalogram, and evidence of cerebral atrophy on structural neuroimaging).

The clinical diagnosis of *possible* AD is made when:

(1) the syndrome of dementia is present, *and*

(2) there is an absence of other neurologic, psychiatric, or systemic disorders sufficient to cause the dementia, *but*

(3) variations are present from criteria for probable AD in the onset, presentation, or clinical course. Possible AD may also be diagnosed when a systemic brain disease sufficient to produce a dementia is present but (for various reasons) is not considered to be the cause of the patient's dementia.

semantic category (e.g., names of animals or items that can be found in a supermarket) that is significantly below normative expectation. With disease progression, difficulties in spatial orientation and visuospatial abilities (e.g., ability to accurately draw a clock face), facial recognition, and language comprehension all occur. These impairments appear to reflect increasing neurofibrillary tangle and neuritic plaque involvement of the parietal and temporal association cortices. In addition, increasing problems with focusing and shifting attention and "executive functions" are seen with disease progression. Executive functions refer to those various abilities (including volition, planning, organizing, sequencing, abstracting, and performance self-monitoring) that contribute to the capacity to plan and carry out complex, goal-oriented behavior. Apathy, difficulty in shifting behavioral set to accommodate new task demands, impaired abstract reasoning, difficulty in the correct sequencing of behavior, and poor awareness of cognitive deficits are all aspects of executive dysfunction and are also correlated with the degree of behavioral disinhibition and agitation among AD patients. The increasing difficulties in attention and executive functioning appear to reflect the frontal cortex neuritic plaque and neurofibrillary tangle density increase that occurs with AD progression.

Apathy and other aspects of executive dysfunction have been found to correlate with functional neuroimaging measures (see Fig. 2) documenting decreased blood flow in the anterior cingulate, dorsolateral frontal, orbitofrontal, and anterior temporal cortical regions. In general, functional neuroimaging [e.g., positron emission tomography (PET)] studies have shown that, in AD patients with mild dementia (relatively early in the illness course), cerebral metabolism is decreased in temporal and parietal cortices, with progression to metabolic decreases in the frontal cortex as dementia severity worsens. There is also evidence for heterogeneity in the relative severity of various cognitive deficits among persons with AD. If we equate for overall severity of dementia, those who show the greatest visuospatial impairments demonstrate the lowest right-hemispheric metabolism, and those with the greatest language impairments have the lowest left-hemispheric metabolism. By the late stage of severe dementia in AD, only sensory and motor cortical areas show preserved metabolic rates.

III. FRONTAL–SUBCORTICAL DEMENTIAS

The most common clinical features of the frontal–subcortical dementias are bradyphrenia (a slowing of thought processes and information processing speed), memory problems, impairment of executive functions, mood changes, and sometimes visuospatial deficits.

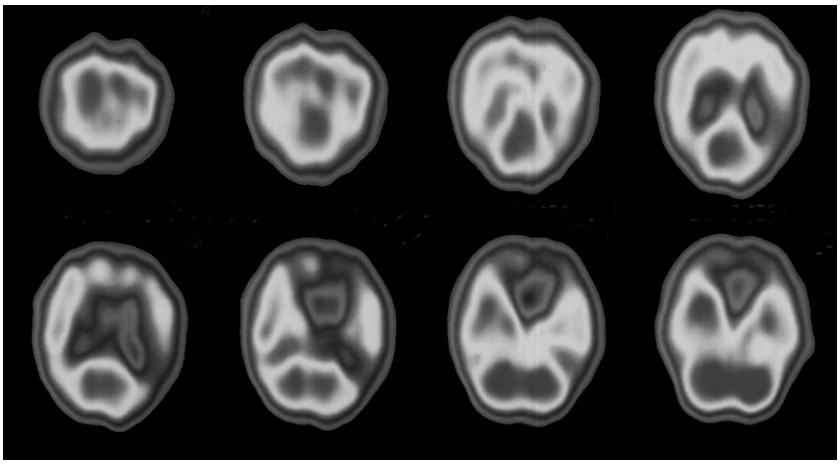


Figure 2 Single photon emission computed tomographic (SPECT) image of the brain of an individual with Alzheimer's disease, showing decreased blood flow (darker areas) in frontal regions as well as in temporal–parietal association cortices.

Those diseases that can cause this dementia syndrome include progressive supranuclear palsy (PSP), Huntington's disease (HD), Parkinson's disease (PD), dementia with Lewy bodies (DLB), and corticobasal degeneration (CBD). Damage to the rostral brain stem, thalamus, basal ganglia, and/or reciprocal connections between these subcortical regions and frontal cortical areas appears to be the common factor in these dementias. These structures all participate in several distinct, semi-closed-loop-circuits that subserve particular motor, cognitive executive, and emotional functions.

A. Progressive Supranuclear Palsy (PSP)

PSP is also known as Steele–Richardson–Olszewski syndrome. It is a degenerative disorder of subcortical nuclei, including the subthalamic nucleus, substantia nigra, globus pallidus, caudate, putamen, and periaqueductal gray. Clinically, persons with PSP show paralysis involving motor neurons of the eye, leading to impaired downward ocular gaze and other ocular symptoms. Clinical diagnosis requires documentation of at least two of the following signs: axial dystonia (abnormal pattern of muscle tonus in the axial musculature) and rigidity (especially of the neck), pseudobulbar palsy (manifested as sudden crying or laughing without apparent cause or accompanying experienced emotion), bradykinesia (slowness in the initiation of movement), signs of frontal lobe dysfunction (e.g., motor perseveration, difficulty in behavioral set-shifting in response to changing task demands), postural instability with backward falling, and sometimes dysarthria (difficulty in the motor aspects of speech). Cognitive deficits, documented by neuropsychological testing, include memory loss, slowed information processing speed, apathy and depression, irritability, and executive function deficits.

B. Huntington's Disease (HD)

HD is an autosomal-dominant genetic disorder marked by initial abnormal choreiform (dancelike) movements, slowed voluntary movements, depressive and manic emotional disturbances, and eventual dementia. HD manifests in an insidious onset, usually in young to middle adulthood. The dementia of HD is characterized neuropsychologically by impaired attention and concentration, executive function deficits, and impaired encoding and retrieval of new information. However, if care is taken to make sure that the person with HD adequately encodes the information (e.g., by having the individual semantically categorize it) and procedures are used that do not require effortful retrieval strategies (e.g., recognition memory rather than a free recall task), then it can be shown that the retention of new information is more intact in HD than in AD patients with comparable overall dementia severity. The neuropathology of HD includes cell loss and atrophy that begins in the medial caudate nucleus and eventually includes the entire caudate and putamen. Some degree of cortical atrophy is also

typically seen in the brains of persons with HD at autopsy.

C. Parkinson's Disease (PD)

PD is primarily a progressive, extrapyramidal (involving motor systems of the basal ganglia rather than the pyramidal tract) disorder with clinical motor features of stooped posture, bradykinesia (slowing of movement initiation), tremor (when the affected limb is at rest, but typically not during movement), cogwheel rigidity (i.e., rigidity that seems to catch and release, much like a cogwheel, when the clinician attempts to passively move the patient's limb), and festinating gait. Dementia has been estimated to occur in approximately 35–40% of persons with PD. The neuropsychological impairments manifested in the dementia of PD include psychomotor slowing, impaired cognitive tracking and set-shifting flexibility, visuospatial deficits, diminished learning and retrieval (although, as with HD, often there is evidence of more intact information retention shown by recognition memory tasks), decreased verbal fluency, and abstract reasoning impairment. Neuronal degeneration (with Lewy bodies; see later discussion) occurs primarily in pigmented cells of the substantia nigra (particularly the pars compacta region) and other brain stem structures (e.g., locus ceruleus, dorsal vagal nucleus). The substantia nigra degeneration causes a severe reduction in dopamine neurotransmitter projections to the striatum, responsible for the observed extrapyramidal motor disorder and contributing to the cognitive impairments.

D. Dementia with Lewy Bodies (DLB)

In addition to occurring in brain stem structures in persons with PD, Lewy bodies are the neuropathological characteristic of DLB, where they appear diffusely distributed throughout the neocortex, diencephalon, brain stem, and basal ganglia. Because of the brain structures involved, DLB is sometimes considered to be among the frontal–subcortical dementias. The core clinical features of DLB are the presence of dementia, gait–balance disorder, prominent hallucinations and delusions, sensitivity to neuroleptic antipsychotic drugs, and fluctuating alertness. A consensus has been reached concerning the clinical criteria for diagnosing DLB (see Table II). These criteria have demonstrated high specificity, although relatively low sensitivity, against neuropathological findings.

Illness progression in DLB is often rapid, with severe dementia and the motor features of PD present

Table II
Criteria for Dementia with Lewy Bodies (DLB)

The central required feature is progressive cognitive decline of sufficient magnitude to interfere with normal social or occupational function.
(1) Prominent or persistent memory impairment may not necessarily occur in early stages but usually is evident with progression.
(2) Deficits on tests of attention, frontal–subcortical functions, and visuospatial abilities may be especially prominent.
Core features (two required for probable DLB and one for possible):
(3) Fluctuating cognition with pronounced variations in attention and alertness.
(4) Recurrent visual hallucinations that are typically well-formed and detailed.
(5) Spontaneous motor features of parkinsonism.
Features supportive of the diagnosis:
(6) Repeated falls
(7) Syncope
(8) Transient loss of consciousness
(9) Neuroleptic drug sensitivity
(10) Systematized delusions
(11) Hallucinations in other modalities
Diagnosis of DLB is less likely in the presence of:
(12) Evidence of stroke from focal neurological signs or brain imaging.
(13) Evidence of any physical illness or other brain disorder sufficient to account for the clinical features.

within 1–5 years of diagnosis. It has been suggested that a diagnosis of DLB should require that cognitive dysfunction occur within 12 months of the onset of parkinsonian motor features. Patients with PD who develop dementia more than 12 months after the onset of their motor signs are diagnosed as having PD with dementia. The neuropsychological test profiles of persons with DLB show marked deficits in attention and visual–constructive skills (e.g., ability to copy a geometric pattern with blocks), frequently with relative sparing of memory. However, neuropsychological testing has not been able to reliably differentiate DLB from either AD or vascular dementia, given the clinical variability and overlap of the cognitive impairments in these dementia types. Structural neuroimaging (CT or MRI) also has not been successful in differentiating DLB from AD.

E. Lewy Body Variant (LBV) of AD

The Lewy body variant of AD (sometimes also referred to as ''common'' DLB, in contrast to ''pure'' DLB) is characterized by the typical neuritic plaques and neurofibrillary tangles of AD, the subcortical changes of PD, and the presence of diffusely distributed cortical Lewy bodies. Persons with the LBV of AD, compared to persons with ''pure'' AD, show a greater proportion of mild parkinsonian or other extrapyramidal motor findings, a typically fluctuating cognitive impairment, visual or auditory hallucinations, and frequent unexplained falls.

F. Corticobasal Degeneration (CBD)

CBD is a rare disorder that presents clinically as an asymmetric, akinetic rigid syndrome with apraxia (impaired voluntary execution of symbolic movements and gestures), myoclonus, and sometimes dementia. When dementia is present, the pattern of neuropsychological test performance resembles that of PSP, except that apraxia is typically more severe in CBD. Symptoms of cortical sensory loss (e.g., impaired visual or tactile discrimination) are also common early in CBD. Persons with CBD often show the ''alien limb'' syndrome, in which the individual reports that their limb is moving without their volitional control, and often are mildly depressed. CBD is associated with markedly asymmetric patterns of cortical metabolism alteration (as measured by PET). Cortical metabolism contralateral to the affected limb is typically lower in inferior parietal, lateral temporal, and sensory motor cortices than that seen in PSP.

IV. FRONTOTEMPORAL DEMENTIAS

FTD includes both Pick's disease and non-Pick's frontotemporal dementia (also called frontal lobe dementia, FLD), motor neuron disease, and progressive subcortical gliosis. Although estimates of the prevalence of FTD remain somewhat controversial, some investigators think that these disorders have been underestimated and could account for 10–19% of all demented cases. The usual age of onset is between 45 and 65 years, with equal sex incidence. A family history of the illness is present in approximately 50% of persons with FTD. FTD includes subtypes of semantic dementia, primary progressive aphasia, and progressive prosopagnosia, as described later.

Pick's disease is characterized by Pick's cells (enlarged neurons with displaced nuclei) and Pick's inclusion bodies (round intraneuronal cellular inclusions) in frontal and temporal cortical areas. In Pick's disease, marked atrophy (termed ''knife edge'' due to the extreme thinning of the cortical gyri) is found in the frontal and anterior temporal lobes, with some patients also showing atrophy in the basal ganglia and caudate nucleus. The gross neuropathology of non-Pick's FTD is similar to that of Pick's, with atrophy primarily affecting the frontal lobes. However, those with non-Pick's FTD do not typically show Pick's cells and bodies. Some show microvacuolation or spongiosus (often termed the frontal lobe dementia type, FLD). Degeneration of bulbar cranial nerve nuclei and anterior horn cells may also be present in some persons with FTD. Chromosome 17 abnormality with τ protein aggregation and decreased τ binding to neuronal cell microtubules has been found in the brains of persons who died with FTD.

Personality and behavioral changes, executive dysfunction, and language disturbances are predominant early signs of FTD. Language deficits are progressive in nature and may eventually lead to mutism. Neurological signs including akinesia, rigidity, masked facies (markedly decreased spontaneous facial expression), gait disturbance, dysarthria, dysphagia (swallowing difficulty), and tremor are present throughout the disease for some patients and only in advanced disease for others.

A. Behavioral and Emotional Characteristics of FTD

The most common behavioral changes that accompany FTD are loss of social awareness and insight, personal neglect, disinhibition, impulsivity, impersistence, inertia, aspontaneity, mental rigidity and inflexibility, motor and verbal perseveration, stereotyped activities and rituals, utilization behavior (tendency to compulsively use whatever object is placed before the individual), and hyperorality (tendency to frequently mouth objects). Emotional changes of FTD include unconcern, apathy, emotional shallowness and lability, loss of empathy and sympathy, and a fatuous jocularity. Consensus clinical criteria, based on these behavioral and emotional characteristics, have been developed for FTD (see Table III). However, as yet there is insufficient research comparing these criteria to neuropathological findings, preventing adequate assessment of sensitivity and specificity.

Neuropsychological testing of persons with FTD typically reveals deficits in verbal fluency, abstraction ability, and other areas of executive functioning. However, because some AD patients also demonstrate substantial executive function deficits, this pattern of neuropsychological test performance cannot be considered specific to FTD.

B. Semantic Dementia

Semantic dementia is a subtype of FTD characterized by fluent, anomic aphasia (marked difficulty with word-finding and naming), with impaired verbal comprehension and loss of semantic knowledge. The neuropathology in semantic dementia is typically most marked in the left (or bilateral) temporal pole and inferolateral cortex.

C. Progressive Nonfluent Aphasia

Another subtype of FTD, known as progressive nonfluent aphasia, is characterized by nonfluent, hesitant, distorted speech with preserved verbal comprehension. The neuropathology is typically most marked in the left perisylvian cortical area.

D. Progressive Prosopagnosia

The progressive prosopagnosia subtype of FTD is characterized by impaired identification of familiar faces, followed by loss of person knowledge. The neuropathology is typically most marked in the right temporal pole and inferolateral cortex.

V. VASCULAR DEMENTIA

In the presence of cerebrovascular disease, dementia may be observed when there are multiple infarctions within both cerebral hemispheres. The clinical features that accompany multiple infarctions of relatively large cerebral blood vessels (termed multi-infarct dementia, MID) depend upon the particular brain regions

Table III
Criteria for Frontotemporal Dementia (FTD)

Core diagnostic features:

(1) Insidious onset with gradual progression

(2) Early decline in interpersonal conduct

(3) Early impairment in the regulation of personal conduct

(4) Early emotional blunting

(5) Early loss of insight

Supportive diagnostic features:

(6) Behavioral disorder, including decline in personal hygiene and grooming; mental rigidity and inflexibility; distractibility and impersistence; hyperorality and dietary changes; perseverative and stereotyped behavior; and/or utilization behavior.

(7) Speech impairment, including altered speech output (either aspontaneity and economy of speech or press of speech); stereotype of speech; echolalia; perseveration; and/or mutism.

(8) Physical signs, including primitive reflexes; incontinence; akinesia, rigidity, and tremor; and/or low and labile blood pressure.

(9) Investigations, including significant impairment on neuropsychological tests of executive functioning in the absence of severe amnesia, aphasia, or visuospatial disorder; normal electroencephalogram despite clinically evident dementia; and brain imaging (structural and/or functional) showing predominant frontal and/or anterior temporal abnormality.

affected. Dementia may also be the result of multiple small subcortical (termed lacunar) infarctions. As previously noted in this article, although vascular disease was earlier thought to be the second most frequent cause of dementia, more recent autopsy studies have indicated that dementia most typically occurs when both multiple infarctions and the neuropathology of AD or Lewy bodies are present. Consensus criteria for the diagnosis of VAD have been proposed (see Table IV). Evaluation of these criteria against neuropathological findings has shown relatively high specificity, but low sensitivity.

In studies of neuropsychological test performance, persons with pure VAD (i.e., those without coexistent AD pathology or Lewy bodies) show less memory impairment than persons with AD matched for overall dementia severity. In contrast, persons with VAD may be relatively more impaired that those with AD on tests of executive functioning. However, the considerable variability of the neuropsychological test performance in both VAD and AD limits the specificity of these features for differential diagnosis.

VI. DEMENTIA ASSOCIATED WITH DEPRESSION

Depression in older age has been given increasing clinical and research attention. Although the prevalence of depression (according to standard diagnostic criteria) is not higher for older than for younger adults, older persons with depressive symptoms have health care costs about 50% higher than those without depressive symptoms. These increased costs are not accounted for by mental health service utilization. Depression frequently coexists with neurological disorders among older adults (e.g., stroke, AD, PD) and is often underdiagnosed in older persons, especially when physical illness is present.

Further complicating both accurate diagnosis and treatment is the fact that various prescription and nonprescription medications (and their interactions) can cause, aggravate, or mimic depression symptoms. In addition, persons with depression show impairment on a range of neuropsychological measures of cognitive functioning, particularly on speeded tasks, vigilance tasks, and tasks with pleasant or neutral (in contrast with unpleasant) content. The magnitude of cognitive impairment tends to increase with older age, severity of depression, and history of electroconvulsive treatment (ECT). Previously, the term "pseudo-dementia" was often used to describe persons with depression and cognitive impairment, reflecting the expectation that the cognitive deficits would reverse along with effective treatment of the depression. However, this term is no longer used because several

Table IV
Criteria for Vascular Dementia (VAD)

The clinical diagnosis of *probable* VAD is made when:

(1) Dementia is present, defined by (a) cognitive decline from a previously higher level of functioning; (b) impairment of memory and two or more other cognitive domains (preferably established by clinical examination and documented by neuropsychological testing); (c) sufficient severity as to interfere with activities of daily living; and (d) not due to the sensorimotor effects of stroke alone, *and*

(2) the patient does *not* show disturbance of consciousness, delirium, psychosis, severe aphasia, sensorimotor impairment sufficient to preclude neuropsychological testing, systemic disorders, or other brain diseases (e.g., AD) that could account for the cognitive deficits, *and*

(3) cerebrovascular disease is present, defined by (a) the presence of focal neurologic signs (e.g., hemiparesis, sensory deficit, hemianopsia, Babinski sign, etc.); (b) evidence of relevant features consistent with cerebrovascular disease on brain imaging (CT or MRI), *and*

(4) a relationship is established between the dementia and the cerebrovascular disease inferred by one or more of the following: (a) onset of dementia within 3 months following a recognized stroke; (b) abrupt deterioration in cognitive functioning; or (c) fluctuating, stepwise progression of cognitive deficits.

The diagnosis of probable VAD is further supported by (a) the early presence of a gait disturbance; (b) a history of unsteadiness and frequent falls; (c) early urinary urgency, frequency, and other urinary symptoms not explained by urologic disease; (d) pseudobulbar palsy; and (e) mood and personality changes, psychomotor retardation, and abnormal executive function.

The clinical diagnosis of *possible* VAD is made when:

(1) The patient meets criteria for dementia with focal neurologic signs, *but*

(2) brain imaging studies to confirm cerebrovascular disease are missing, *or*

(3) when there is absence of a clear temporal relationship between dementia and stroke, *or*

(4) in patients with subtle onset and variable course (improvement or plateau) of cognitive deficits and evidence of relevant cerebrovascular disease.

studies have indicated that substantial numbers of these individuals show progressive dementia that does not reverse with depression treatment. Such studies have indicated the need for both caution and persistent follow-up in any attempts to clinically differentiate dementia associated with depression from that due to AD or other dementia syndromes.

Although the severity of cognitive impairment in depression is typically less than that in dementia syndromes such as AD or VAD, differential diagnosis can sometimes be difficult. Depression can coexist with AD, VAD, and other dementia types, complicating differential diagnosis when dementia severity in these disorders is relatively mild. Further, research has shown that those with a first onset of depression in older age (in contrast to older persons with depression onset earlier in life) often show clinical and/or structural neuroimaging evidence of cerebrovascular damage. Persons with later-age-of-onset depression and MRI evidence consistent with relatively large areas of cerebrovascular damage in the subcortical white matter are also more likely to show executive function deficits on neuropsychological testing and are less responsive to antidepressant pharmacotherapy than older adults with depression onset in younger adulthood.

Research comparing the neuropsychological test performance of persons with depression to those with AD has suggested that the following features can be helpful in differential diagnosis. First, as already noted, the cognitive deficits of depression tend to be less severe and extensive than in AD. Second, consistent impairment across various memory tests (particularly involving delayed recall and recognition memory) is more consistent with AD than with depression. Persons with depression are also less likely than those with AD to show impaired naming ability, verbal fluency, and visuospatial ability. Finally, those with depression may appear to exert less effort than those with AD in the performance of various neuropsychological tests and may complain more about their cognitive difficulties (even when testing does not document significant deficits).

VII. NEUROPSYCHOLOGICAL ASSESSMENT IN DEMENTIA DIAGNOSIS

Accurate diagnosis of dementing illness often necessitates the efforts of multidisciplinary teams of physicians (particularly neurologists and psychiatrists), neuropsychologists, and other mental health professionals (e.g., social workers). As noted throughout the previous sections, neuropsychological assessment, utilizing standardized tests of cognitive and emotional functioning, makes an important contribution to the clinical identification of the dementia syndrome and the differentiation of various dementia types.

Brief mental status questionnaires can provide a quick and valid documentation of the presence of dementia. However, such questionnaires generally are not sufficiently sensitive to detect very mild dementia, particularly in persons with high premorbid intellectual functioning. This limitation in sensitivity reflects the fact that these questionnaires are composed of relatively easy questions and memory items. In addition, they lack the specificity to assist in differential diagnosis of dementia types. In comparison, standardized neuropsychological tests typically contain a range of task difficulty, so that test scores approximate a normal distribution when administered to individuals in the general population. This can provide for greater sensitivity to mild or subtle cognitive deficits than what is possible through the use of mental status questionnaires, for which normative performance distributions are markedly skewed due to most individuals achieving near-perfect scores.

Appropriate neuropsychological tests thus may reveal subtle and circumscribed cognitive impairments in patients who show no evidence of cognitive deficit on commonly employed mental status questionnaires. The ability of neuropsychological test batteries to examine the pattern of performance across different, reliably measured domains of cognitive functioning may be particularly important in assessing persons with high premorbid intellectual functioning. This is due to the fact that the different dementia types do not manifest with equivalent impairment across all cognitive functions. Thus, performance on those cognitive tests that are likely to be affected can be compared to those that are likely to remain intact as an aid in improving detection of relatively circumscribed cognitive decline.

In selecting, administering, and interpreting neuropsychological tests, several factors need to be considered. First, specific tests should be selected on the basis of adequate standardization, including demonstrated reliability and normative data appropriate to the age, educational background, and other demographic characteristics of the person being examined. In addition, the tests should have demonstrated sensitivity and specificity for the diagnostic

possibilities to be differentiated. Fortunately, there is a growing empirical literature that allows for a comparison of the expected test performance of persons with differing dementia types. In addition to test standardization, sensitivity, and specificity, the impact of various patient characteristics must be considered. Age-related sensory acuity changes and response slowing can influence test performance, as can the physical limitations of such prevalent illnesses as arthritis.

In addition to a diagnostic role, neuropsychological assessment provides a foundation for the accurate provision of information to patients, family members, and health care providers concerning specific strengths and deficits. Accurate information can reduce ambiguity and anxiety (for both patients and their caregivers), and the identification of relatively intact areas of functioning facilitates the maximizing of patient independence. Identification of intact functioning helps in developing plans for sharing daily responsibilities between the patient and others in his or her environment, directing caregivers to areas in which the patient will require additional supervision or assistance. Early identification of dementia and accurate feedback to the patient and family members allows time for them to make future plans (e.g., disposition of estate, medical and long-term care wishes) while the patient is still able to communicate his or her wishes. Neuropsychological tests are also a critical component in the assessment of treatment effects or disease progression, and they provide guidance for treatment and management of cognitive and behavior problems. Tracking of disease progression, through periodic reexamination, is important for guiding caregivers and adjusting the goals of clinical management. Research has also shown that feedback to patients, based on neuropsychological test performance, results in the improved utilization of coping strategies in persons with preserved awareness of their cognitive deficits.

Evaluation of behavior problems, utilizing standardized and validated informant questionnaire and behavioral observation instruments, can be particularly helpful in efforts to assist the caregivers of dementia patients. Apathy, agitation, depression, delusions, and diurnal rhythm disturbance (e.g., wandering and agitation at night and somnolence during the day) are common in persons with AD and other dementias. These behavior problems are associated with increased caregiver distress, higher rate of nursing home placement, and more rapid disease progression. Once identified and accurately described by careful assessment, family members and other caregivers can be instructed in approaches to minimizing the occurrence and frequency of these behavior problems.

VIII. TREATMENT AND CLINICAL MANAGEMENT OF DEMENTIA

Pharmacological treatment options for the dementias presently are limited, although numerous experimental treatments are being evaluated. The one well-validated approach to symptomatic treatment of AD is based upon the documented deficit in acetylcholine neurotransmitter levels, due to the loss of cholinergic projecting neurons within the nucleus basalis of Meynert. It has been shown that this deficit can be partially compensated for by interfering with the action of that enzyme (cholinesterase), which breaks down acetylcholine after it enters the synaptic junction, allowing the constituent molecules to be taken back into the presynaptic neuron to synthesize additional acetylcholine. There are four such cholinesterase inhibitors that have been approved by the U.S. Food and Drug Administration: Tacrine, Donepezil, Rivastigmine, and Galantamine. These drugs have all been shown to result in some temporary improvement of function in at least some AD patients. However, none of these drugs interrupts the progression of the illness, and the search continues for treatments that can alter the fundamental neuropathological processes of AD. Vitamin E (α-tocopherol) has been shown to delay the time to institutionalization and the loss of ability in common activities of daily living in moderately severe AD patients; thus, it is often prescribed along with cholinesterase inhibitors.

Some of the behavioral problems of dementia (e.g., aggression) may be treated with antipsychotic drugs (e.g., Risperidone, Haloperidol). However, caution must be exercised because those with DLB may be excessively sensitive to such neuroleptic drugs, and several deaths have been reported within weeks of starting such drugs in these patients. Antidepressant drugs (e.g., tricyclics, monoamine oxidase inhibitors, selective serotonin re-uptake inhibitors) may also be useful in the treatment of depression in persons with dementia.

In addition to pharmacological treatment, behavior management strategies can assist in minimizing the consequences of cognitive deficits and managing the behavior problems of dementia. Memory aids and household memory cues (e.g., reminders posted on the

refrigerator or above the toilet) can partially compensate for memory deficits. There are also interventions available to reduce behavioral problems in dementia. Small-scale empirical studies of various behavior management procedures have shown some efficacy in reducing select behavior problems (e.g., aggressiveness, agitation, incontinence). A larger scale study has documented the effectiveness of a behavioral therapy taught to caregivers in reducing coexistent depression in AD. Other research has shown that intensive, long-term education and support services for caregivers can delay the time to nursing home placement for AD patients. Even short-term programs that educate family caregivers about AD may improve caregiver morale and satisfaction. Finally, education about AD and other dementias for the staff of nursing homes and other long-term care facilities may help to reduce the use of unnecessary antipsychotic medications.

Thus, although there is no available treatment capable of reversing or arresting the progression of dementia, both symptomatic treatment and interventions to reduce behavior problems are available. The vigor with which experimental treatments are being evaluated creates hope for a future in which the dementias can be more effectively treated, and someday perhaps even cured.

See Also the Following Articles

AGING BRAIN • AGNOSIA • ALZHEIMER'S DISEASE, NEUROPSYCHOLOGY OF • APHASIA • CEREBRAL WHITE MATTER DISORDERS • COGNITIVE REHABILITATION • DEPRESSION • HIV INFECTION, NEUROCOGNITIVE COMPLICATIONS OF • LANGUAGE, NEURAL BASIS OF • NEUROPSYCHOLOGICAL ASSESSMENT • PARKINSON'S DISEASE • PICK'S DISEASE AND FRONTOTEMPORAL DEMENTIA

Suggested Reading

Bondi, M. W., Salmon, D. P., and Kaszniak, A. W. (1996). The neuropsychology of dementia. In *Neuropsychological Assessment of Neuropsychiatric Disorders* (I. Grant and K. H. Adams, Eds.), 2nd ed., pp. 164–199. Oxford University Press, New York.

Christensen, H., Griffiths, K., Mackinnon, A., and Jacomb, P. (1997). A quantitative review of cognitive deficits in depression and Alzheimer-type dementia. *J. Int. Neuropsychol. Soc.* **3,** 631–651.

Cummings, J. L., and Benson, D. F. (1992). *Dementia: A Clinical Approach*, 2nd ed. Butterworth-Heinemann, Boston.

Doody, R. S., Stevens, J. C., Beck, C., Dubinsky, R. M., Kaye, J. A., Gwyther, L., Mohs, R. C., Thal, L. J., Whitehouse, P. J., DeKosky, S. T., and Cummings, J. L. (2001). Practice parameter: Management of dementia (an evidence-based review). *Neurology* **56,** 1154–1166.

Duke, L. M., and Kaszniak, A. W. (2000). Executive control functions in degenerative dementias: A comparative review. *Neuropsychol. Rev.* **10,** 75–99.

Green, J. (2000). *Neuropsychological Evaluation of the Older Adult: A Clinician's Guidebook*. Academic Press, San Diego, CA.

Kaszniak, A. W. (1996). Techniques and instruments for assessment of the elderly. In *A guide to Psychotherapy and Aging* (S. H. Zarit and B. G. Knight, Eds.), pp. 163–219. American Psychological Association, Washington, DC.

Khachaturian, Z. S., and Radebaugh, T. S. (Eds.). (1996). *Alzheimer's Disease: Cause(s), Diagnosis, Treatment, and Care*. CRC Press, Boca Raton, LA.

Knopman, D. S., DeKosky, S. T., Cummings, J. L., Chui, H., Corey-Bloom, J., Relkin, N., Small, G. W., Miller, B., and Stevens, J. C. (2001). Practice parameter: Diagnosis of dementia (an evidence-based review). *Neurology* **56,** 1143–1153.

Nussbaum, P. D. (Ed.). (1997). *Handbook of Neuropsychology and Aging*. Plenum, New York.

O'Brien, J., Ames, D., and Burns, A. (Eds.). (2000). *Dementia*, 2nd ed. Arnold, London.

Petersen, R. C., Stevens, J. C., Ganguli, M., Tangalos, E. G., Cummings, J. L., and DeKosky, S. T. (2001). Practice parameter early detection of dementia: Mild cognitive impairment (an evidence-based review). *Neurology* **56,** 1133–1142.

Depression

ESTEBAN V. CARDEMIL
Brown University School of Medicine and Rhode Island Hospital

GLOSSARY

anhedonia One of the core symptoms of depression; defined as the loss of interest in, and inability to derive pleasure from, activities that were previously considered interesting and pleasurable.

catecholamine A group of neurotransmitters that includes norepinephrine, epinephrine, and dopamine.

concordance rates The proportion of twins in a sample that both possess the disorder of interest.

heritability The proportion of the variance of a disorder in the population that is due to genetic factors.

indoleamines A group of neurotransmitters that includes serotonin and histamine.

lifetime prevalence The proportion of individuals in an epidemiological sample that have ever experienced a disorder at some point in their lives.

Depression is a psychiatric disorder with emotional, cognitive, physiological, and behavioral symptoms. This article presents both biological and psychological theories and treatments of depression.

I. SYMPTOMS OF DEPRESSION

Depression is a disorder characterized by emotional, cognitive, physiological, and behavioral symptoms. The primary emotional symptom is a profound sense of sadness and low mood. Irritability, frustration, and anger often accompany this low mood. Cognitive symptoms include a sense of hopelessness and helplessness, worthlessness, and guilt. Depressed individuals often have difficulty concentrating or making simple decisions. Physiological symptoms include changes in appetite and sleep, fatigue, and concerns about aches and pains. Diminished sexual interest is also commonly reported in depressed individuals. Behavioral symptoms include decreased activity, often the result of anhedonia, which is the loss of interest in and an inability to derive pleasure from activities that previously were interesting and pleasurable.

Depression tends to be classified into two major categories: major depressive disorder and dysthymic disorder. The principal differences between major depressive disorder and dysthymic disorder are severity and chronicity. Major depressive disorder is more severe and characterized by discrete major depressive episodes, whereas dysthymic disorder is less severe and characterized by a more chronic course. These disorders are not mutually exclusive; a major depressive episode superimposed on a chronic course of dysthymic disorder is conceptualized as double depression.

The American Psychiatric Association, in its *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)*, defined the symptomatic criteria for a major depressive episode (Table I). Once the criteria for a major depressive episode have been met, and the appropriate diagnostic rule-outs have been made (e.g., other psychiatric disorder and general medical condition causing the depressive episode), a diagnosis of major depressive disorder can be made.

Table I
***DSM-IV* Criteria of a Major Depressive Episode**

Five or more of the following symptoms must be present during the same 2-week period and represent a change from previous functioning. At least one of the symptoms must be either depressed mood or loss of interest or pleasure.

1. Depressed mood most of the day, nearly every day
2. Markedly diminished interest or pleasure in all, or almost all, activities most of the day, nearly every day (anhedonia)
3. Significant weight loss when not dieting or weight gain (e.g., a change of more than 5% of body weight in 1 month) or decrease or increase in appetite nearly every day
4. Insomnia or hypersomnia nearly every day
5. Psychomotor agitation or retardation nearly every day (noticeable by others)
6. Fatigue or loss of energy nearly every day
7. Feelings of worthlessness or retardation nearly every day (noticeable by others)
8. Diminished ability to think or concentration, or indecisiveness, nearly every day
9. Recurrent thoughts of death (not just fear of dying), recurrent suicidal ideation without a specific plan, or a suicide attempt, or a specific plan for committing suicide

A. Subtypes of Depression

DSM-IV distinguishes among several subtypes of depression in order to capture some of the different patterns of depressive symptoms with which individuals present. Melancholic depression is characterized by a loss of interest or pleasure in all, or almost all, activities. Individuals with melancholic depression do not show reactivity to normally pleasurable stimuli and do not show even temporary improvements in mood following good events. In addition, at least three of the following symptoms are present: the sensation that depression is worse in the morning than in the evening, significant weight loss or loss of appetite, insomnia characterized by early morning awakening, psychomotor agitation, and excessive or inappropriate guilt.

In contrast, individuals with atypical depression show mood changes in response to actual life events. Their mood might temporarily brighten with the development of good news, for example. In addition, individuals also demonstrate two or more of the following symptoms: significant weight gain or appetite increase, hypersomnia, feelings of heaviness in the limbs, and a long-standing pattern of sensitivity to interpersonal rejection.

Individuals who display a characteristic onset and remission of depressive episodes at specific times of the year are given the specifier "with seasonal pattern." Most cases of depression with seasonal pattern occur in the winter months and tend to be found in locations where winter seasons are long and accompanied by days with short exposure to sun. The symptoms include increased fatigue, hypersomnia, weight gain, and a craving for carbohydrates.

Women who develop a major depressive episode within 4 weeks of giving birth are given the specifier "with postpartum onset." The symptoms do not seem to differ from nonpostpartum depressive episodes but may be more fluctuating in course. Often, the focus of the depression is related to the newborn and can be accompanied by obsessional thoughts of harming the child, suicidal ideation, and psychomotor agitation. Researchers currently disagree about the extent to which depression with postpartum onset is actually distinctive enough to warrant a separate category.

II. EPIDEMIOLOGY OF DEPRESSION

Up to 17% of the general population in the United States will meet criteria for a major depressive episode at some time. The lifetime prevalence for dysthymia is approximately 6%, and that of depression with a seasonal pattern is approximately 0.4%. Also, many more people will have some experience with lower, subclinical levels of depressive symptoms. Children and adolescents also experience depression and its associated symptoms: Estimates of the number of children who will experience a depressive episode by the end of high school range as high as 20%.

The annual economic consequences of depression in the United States have been estimated at $44 billion (e.g., medical expenses and lost work hours). In addition to the financial burdens to society associated with depression, there are tremendous psychological

and emotional consequences of depression. Almost 75% of individuals who reported a lifetime history of depression also reported experience with one or more other psychiatric disorders, including anxiety disorders (58%) and substance use disorders (39%). Up to 15% of individuals with severe major depressive disorder will attempt to commit suicide.

A. Demographic Risk Factors for Depression

The prevalence rates of depression are approximately twice as high in women as in men. At some point during their lives, as many as 20% of women will experience a major depressive episode. Prevalence rates of depression range from 0.4 to 2.5% in children and are between 0.4 and 8.3% in adolescents. Epidemiological studies in the United States have consistently found lower lifetime prevalence rates of major depression among African Americans than among Caucasians and similar lifetime prevalence rates between Latinos and Caucasians. Although comparable rates of depression have been found in many non-Western cultures, there is evidence that depression may be more prevalent in Western societies. In particular, it appears that the increased prevalence of depression in women may be more characteristic of Western cultures.

Other demographic risk factors that have been consistently associated with depression include a family history of depression, low socioeconomic status, and specific stressful life events (including assaults and robberies, serious marital problems, divorce, loss of employment, serious illness, and significant financial problems). Research on the role of stressful life events has also found that specific buffers may protect against the development of depression. Some of the buffers include the existence of supportive relationships, the presence of three or fewer children, employment, self-identification with a religion, and the possession of clear roles in life.

B. Course of Depression

Major depression can begin at almost any age, but the average age of onset is in the mid-20s, although data are emerging that suggest that the average age of onset is steadily dropping. Once they begin, depressive symptoms can develop over the course of days or weeks. An untreated depressive episode usually lasts between 6 and 10 months and is often followed by a return to normal functioning with the complete absence of symptoms. However, in a significant minority of cases (20–30%), depression either does not abate at all or significant symptoms persist that continue to interfere with normal functioning. After 2 years, approximately 20% of patients will continue to meet criteria for depression, 12% will meet criteria at 5 years, and 7% will continue to meet criteria 10 years after the onset of the episode. Currently, there are few diagnostic predictors that can distinguish among individuals who return to functioning and those who experience a more chronic course of depression.

Depression is also a recurring disorder: Recent estimates have found that approximately 50–85% of individuals who meet criteria for a depressive episode will experience multiple episodes, and between 25 and 40% of patients will have a second depressive episode within 2 years of their first. Moreover, the risk for relapse appears to increase as time goes on; in some studies the rate increases to 60% and higher after 5 years. Given this high risk for relapse, depression is now more commonly understood as a recurrent and potentially lifelong disorder.

III. BIOLOGICAL THEORIES OF DEPRESSION

Depression can be caused by specific biological conditions, including strokes, nutritional deficiencies, and infections. In these cases, the diagnosis mood disorder due to a general medical condition is given. Depression can also be the result of alcohol or substance abuse, often associated with the symptoms of withdrawal and intoxication. In these cases, the diagnosis substance-induced mood disorder is appropriate. Apart from these two categories, however, the role of biological causes in the development of depression is less definitive. Researchers view the role of biology as contributing one important element to the development of depression.

A. Genetic Theories

To date, researchers have been unable to identify specific genes, acting singly or in combination, that account for the expression of depression. As such, the evidence that has supported the proposition that depression is a heritable disorder comes from family studies, twin studies, and adoption studies. Unfortunately, this research has been unable to definitively

separate the respective contribution of genetics and environment.

For example, depression is up to three times more common in first-degree biological relatives of persons with the disorder than among the general population, and it is almost twice as common in first-degree biological relatives of persons with bipolar disorder. Moreover, children of parents with other psychiatric disorders, including anxiety and substance use disorders, are more than twice as likely to develop depression than are children of parents without psychiatric disorders. However, this support for a genetic contribution to depression is confounded with the increased likelihood of shared environment of first-degree biological relatives.

Twin studies have found mixed evidence for heritability in depression. Concordance rates of depression in identical twins tend to be up to twice as high as the concordance rates in fraternal twins, suggesting a genetic component. However, with heritability estimates only reaching 50%, the genetic contribution must also be considered in the context of environmental factors. Adoption studies have yielded similarly equivocal findings with respect to the heritability of depression. One large-scale study reported that biological relatives of individuals with depression have up to eight times the risk of adoptive relatives, but two other studies did not replicate these findings.

Taken together, the evidence suggests that genetic factors play a contributory role in the etiology of depression, but nowhere near as primary a role as in some other psychiatric disorders (e.g., bipolar disorder and schizophrenia).

B. Neuroanatomical Theories

Early neuroanatomical conceptualizations of depression focused on the limbic system, given its predominant role in normal mood and affect regulation. These early conceptualizations have been bolstered by recent functional imaging studies, which are conducted while depressed patients perform a variety of cognitive tasks. The functional imaging research has generally found abnormal patterns of limbic activity in depressed patients when compared to normal controls, although no coherent pattern has emerged across studies. For instance, some research has noted decreased activation in the midcingulate gyrus; other studies have found reduced activity in the inferior frontal gyrus. Researchers have also reported decreased cerebral blood flow in brain structures other than the limbic system, including the frontal and prefrontal lobes, the cerebrum, and the cerebellum. This hypometabolism is often in direct proportion to the severity of the depressive symptoms.

Given both the correlational nature of this neuroanatomical research and the fact that it is still in its relative infancy, researchers have been unable to determine the extent to which these neuroanatomical dysfunctions are the result or the cause of depression. One recent study found intriguing similarities between depressed and remitted (previously depressed) patients in that both groups displayed decreased prefrontal and limbic activation compared with normal controls. This finding raises the possibility that a neuroanatomical dysregulation may produce a vulnerability to depression and/or relapse.

Taken as a whole, neuroanatomical theories of depression have yet to produce reliable and specific insights into the causes and treatment of depression. However, given the power of functional imaging research, it is likely only a matter of time before researchers begin to clarify the neuroanatomical substrates of depression.

C. Neurotransmitter Theories

Most of the biological research on depression has focused on the role of dysregulation of neurotransmitters and has progressed in step with the development of antidepressant medications that affect these neurotransmitters. The two primary classes of neurotransmitters that have been implicated in depression are the catecholamines (primarily norepinephrine) and the indoleamines (primarily serotonin).

The observation that medications with depressive side effects depleted norepinephrine (NE) in the brain led researchers to hypothesize that catecholamines played a primary role in depression. Reserpine, a medication prescribed for hypertension in the 1950s, was found to produce depression in a significant minority of cases and subsequently found to reduce NE levels in the brain. Moreover, medications that increased the availability of NE, including amphetamines, produced antidepressant effects. These discoveries led to the catecholamine theory of depression, which posits that deficits in NE are the cause of depression. This theory led to the development of antidepressant medications that increased the availability of NE in the brain: the tricyclic antidepressants (TCAs) and the monoamine oxidase inhibitors (MAOIs).

Researchers soon noted that in addition to affecting NE, both TCAs and MAOIs appeared to also increase the availability of serotonin (5-HT) in the brain. This discovery led to the indoleamine theory of depression, which hypothesized that deficits in 5-HT were the root cause of depression. As a result, researchers soon developed a class of drugs that primarily increased the availability of 5-HT: the selective serotonin reuptake inhibitors (SSRIs). More than 14 subtypes of 5-HT receptors have been subsequently identified, and evidence is emerging that has linked 5-HT_{1A} and 5-HT_2 receptors with mood. Given that the definition of depression includes a combination of cognitive, physiological, and behavioral symptoms, future research is likely to identify more 5-HT receptor subtypes that play a significant role in depression.

Neither the catecholamine nor the indoleamine theories of depression account for some important clinical observations, however. For example, the biochemical action of antidepressant medication (in particular the TCAs) occurs quickly, whereas the clinical response takes several weeks to develop. That is, the increased availability of both NE and 5-HT precedes the amelioration of depressive symptoms by several weeks. In addition, TCAs vary in their ability to increase the availability of NE and 5-HT, and SSRIs vary in the extent to which they increase the availability of 5-HT, but all antidepressants have generally the same effectiveness. These observations, among others, have led researchers to doubt that simple deficits in either NE or 5-HT can explain the development of depressive symptoms. Instead, researchers are considering more complicated models that involve the dysregulation of one or more neurotransmitter systems and the various interactions across neurotransmitter systems to explain the development of depression.

D. Neuroendocrine Theories

Approximately 40–60% of depressed individuals produce and secrete excessive amounts of cortisol, primarily during the afternoon and evening. In these patients, cortisol secretion returns to normal levels after the depressive episode remits. This excessive cortisol is thought to be due to the overproduction by the hypothalamus of corticotropin-releasing hormone (CTRH), a compound that is stimulated by norepinephrine and acetylcholine, leading some to believe that CTRH and the noradrenergic system may be interconnected. Administration of dexamethasone, a chemical that temporarily suppresses the production of cortisol in nondepressed adults, was at one point believed to be a useful tool in the diagnosis of depression since many depressed individuals did not demonstrate this cortisol suppression. Recent evidence has indicated that many nondepressed psychiatric patients also fail to display a response to the dexamethasone suppression test, thus limiting its utility in the diagnosis of depression.

IV. PSYCHOLOGICAL THEORIES OF DEPRESSION

Despite the prominent role of biological theories of depression, psychological theories continue to play a significant role in understanding depression. There exist many different psychological approaches to understanding depression that can be considered to emerge from three primary theoretical orientations. Psychoanalytic and psychodynamic theories postulate that depression is the result of unconscious developmental processes that lead to anger being turned inwards. Behavioral theories maintain that depression is the result of excessive behaviors and activities that contribute to low mood. Cognitive theories of depression posit that maladaptive thinking puts individuals at greater risk for developing a depressive episode when faced with negative life events.

A. Psychoanalytic and Psychodynamic Theories

Sigmund Freud and his followers developed the first psychological theories of depression. According to Freud, depression (or melancholia) was comparable to normal grief following the loss of a loved one. Depression, however, would occur in certain people who experienced a loss or disappointment at an early age. These individuals would actually feel rage at the lost loved object, but since part of their personality had become identified with this lost loved object, the individuals would then direct this rage at the self. Often, this redirected self-hatred was safer than rage at the lost loved object. Thus, according to the psychoanalytic model, depression is actually anger turned inwards toward the self. This ''anger turned inwards'' is the source of low self-esteem, feelings of guilt and worthlessness, and sense of deserving punishment, and it may ultimately lead to suicide.

The psychoanalytic model was not originally developed to explain the development of discrete psychological disorders. Rather, it was developed as a comprehensive model of human personality and

development. As such, until recently its relationship to the development of depression was not the subject of much empirical investigation; therefore, few data exist to support its perspective.

Modern psychodynamic theories of depression, while tied to the original psychoanalytic theories, tend to emphasize the role of maladaptive social relationships in the development of depression. In general, these maladaptive relationships mirror significant early childhood relationships. Some theorists, including John Bowlby and Harry Stack Sullivan, have argued that the quality of young children's attachments to their mothers will influence the development of future relationships. Some empirical evidence exists, including some animal data, to support the role of maladaptive relationships in depression. The extent to which these early maladaptive relationships play a primary causal role is disputed, however.

B. Behavioral Theories

Behavioral theories of depression focus on the links between behaviors and mood. These dysfunctional connections lead to an excess of behaviors and activities that produce depressed mood and a deficit in behaviors and activities that produce positive mood. This disconnect between behavior and positive mood can occur because of insufficient engagement in pleasurable activities (e.g., living in an impoverished environment) and/or deficiencies in interpersonal skills (e.g., excessive shyness). Depressed individuals thus find themselves engaging in fewer activities and relationships that promote positive mood. This withdrawal and avoidance of pleasurable activities results in increased depressed mood, which then contributes to a cycle of continued lack of engagement and skill deficits.

C. Cognitive Theories

Cognitive theories of depression emphasize the role of accessible cognitive processes in the development and maintenance of depression. Maladaptive cognitions contribute to the etiology of depression by making individuals susceptible to depression in the face of significant negative life events. Moreover, these maladaptive cognitions and behaviors play a critical role in the maintenance of the depressive state by preventing depressed individuals from considering any alternatives to the pervasive hopelessness that dominates their thinking.

Two influential cognitive theories of depression are Aaron T. Beck's negative cognitive triad and Martin E. P. Seligman's learned helplessness theory.

1. Beck's Negative Cognitive Triad

In his 1967 book, *Depression: Causes and Treatments*, Beck first proposed that depression was the result of the activation of overly negative views (or schemas) of oneself, one's world, and one's future. A negative view of oneself would produce low self-esteem, a negative view of the world could produce helplessness, and a negative view of the future would lead to hopelessness. These depressive schemas begin to develop in childhood as individuals begin to form beliefs about their place in the world. They are maintained and reinforced through a system of cognitive distortions, whereby individuals lend excessive credence to evidence that supports their negative beliefs and selectively ignore evidence that contradicts these beliefs.

Extensive research supports the presence of negative schemas and cognitive distortions in depressed individuals. The extent to which these negative cognitions precede and cause the development of depression is less clear. Some researchers argue that negative cognitions should be considered symptoms of depression rather than causes.

2. Seligman's Learned Helplessness Model

The learned helplessness model posits that individuals become depressed and helpless if they experienced a disconnect between their behavior and life outcomes. This experience with uncontrollable outcomes leads to expected noncontingencies between future responses and outcomes. This theoretical model evolved from animal research in which dogs that were exposed to inescapable shock demonstrated helplessness deficits when they were later exposed to escapable shock. This learned helplessness, which included motivational and emotional symptoms, appeared to mimic many of the motivational, emotional, and cognitive deficits found in individuals experiencing depression.

In 1978, Lyn Abramson, Martin Seligman, and John Teasdale reformulated the original learned helplessness model because the theory was unable to explain why not everyone who was exposed to uncontrollable negative life events would become helpless and depressed. The reformulated learned helplessness model proposed that individuals have habitual ways of explaining the stressors that occur in their lives. This tendency to explain stressors in a characteristic

manner was termed attributional or explanatory style. There are three dimensions along which explanatory style can be measured: internality, stability, and globality. Internality refers to the extent to which the cause of an event is due to something about the individual or something outside of the individual (e.g., other people or luck). Stability refers to the extent to which the cause of the event will remain stable over time or is more transient in nature. Globality refers to the extent to which the cause of the event affects many different life domains or those most immediately related to the stressor. A pessimistic explanatory style is the tendency to explain negative life events with internal, stable, and global causes, and according to the reformulated learned helplessness model it puts individuals at risk for developing depression when exposed to uncontrollable life events.

Considerable evidence exists to support the role of a pessimistic explanatory style in depression, both in children and in adults. In particular, the role of a pessimistic explanatory style as a psychological risk factor for depression in the face of negative life events has received much support.

V. BIOLOGICAL TREATMENTS FOR DEPRESSION

The primary forms of biological treatments for depression are the antidepressant medications. Electroconvulsive therapy (ECT) is generally used with only severely depressed individuals. Alternative biological treatments (e.g., herbal supplements) are also discussed.

A. Antidepressant Medication

Given the emphasis on neurotransmitter theories of depression, the primary biological treatments of depression are antidepressant medications, all of which appear to effectively treat depression in approximately 60–70% of individuals. Antidepressants typically improve mood, energy, and sleep and reduce anhedonia, and they generally accomplish these changes by increasing the availability to the brain of norepinephrine and serotonin. More recent antidepressants also increase dopamine levels in the brain. Because research has not been able to definitively identify the relationship between antidepressant chemical structure and function, antidepressants tend to be classified by their action on neurotransmitters. As such, antidepressants can be categorized into three primary classes: those that inhibit the reuptake of neurotransmitter, those that inhibit the degradation of neurotransmitter, and those that act at specific receptor sites to induce greater production of neurotransmitter (Table II).

1. Reuptake Inhibitors

There exist four types of reuptake inhibitors. The first antidepressants to be identified were the tricyclic antidepressants that inhibited the reuptake of both serotonin and norepinephrine (serotonin–norepinephrine reuptake inhibitors). A second generation of tricyclic antidepressants was developed that predominately inhibited the reuptake of norepinephrine (selective norepinephrine reuptake inhibitors). By interfering with the normal reuptake process by which the synapse cleanses itself of excess neurotransmitter, the tricyclic antidepressants increased the availability of norepinephrine and serotonin in the synapse. Both sets of tricyclic antidepressants interact with other neurotransmitter system receptor sites, including histamine, acetylcholine, and epinephrine, producing a wide range of unwanted side effects (e.g., dry mouth, dizziness, blurred vision, constipation, orthostatic hypotension, and cardiovascular effects). In addition, given tricyclics' high solubility in lipid tissue, and subsequent difficult extraction in emergency, risk of death from overdose remains a serious concern.

The effort to develop effective antidepressants that produced fewer and less serious side effects than the tricyclics led researchers to those that selectively inhibit the reuptake of serotonin (e.g., the SSRIs). SSRIs are well tolerated, provide safety from overdose, and do not appear to increase the risk for seizures. The most common side effects associated with SSRIs include some anxiety and agitation, nausea, sleep disruption, sexual dysfunction, gastrointestinal cramps, diarrhea, and headache. SSRIs are currently the most commonly prescribed antidepressant medication.

Recently, researchers have developed several new classes of antidepressants. Venlafaxine is a serotonin–norepinephrine reuptake inhibitor, similar to the first generation of tricyclic antidepressants, but it does not affect the other neurotransmitter systems and thus has significantly fewer side effects. Buproprion is a novel antidepressant that inhibits the reuptake of norepinephrine and dopamine while having no effect on

Table II
Antidepressant Medications

Class	Mechanism of action	Example drugs	Common side effects
Tricyclic antidepressants I	Block reuptake of both norepinephrine and serotonin	Amitryptaline (Elavil), doxepin (Sinequan), imipramine (Tofranil)	Dry mouth, dizziness, blurred vision, constipation, orthostatic
Tricyclic antidepressants II	Selectively block reuptake of norepinephrine	Desipramine (Norpramin), nortryptaline (Pamelor), protriptyline (Vivactil)	Hypotension and cardiovascular effects
Monoamine oxidase inhibitors	Inhibit degradation of norepinephrine, serotonin, and dopamine	Phenelzine (Nardil), tranylcypromine (Parnate)	Dizziness, sleep disturbances, sedation, fatigue, general weakness, hyperreflexia, dry mouth, and gastrointestinal disturbances; important to avoid food rich in tyramine to avoid hypertensive crises or seizures
Selective serotonin reuptake inhibitors	Block reuptake of serotonin by interfering with the serotonin transport system	Citaprolam (Celexa), fluoxetine (Prozac), fluvoxamine (Luvox), paroxetine (Paxil), sertraline (Zolaft)	Some anxiety and agitation, nausea, sleep disruption, sexual dysfunction, GI cramps, diarrhea, headache
Non-tricyclic serotonin–norepinephrine reuptake inhibitor	Block reputake of serotonin and norepinephrine	Venlafaxine (Effexor)	Nausea, dizziness, drowsiness, possible sexual dysfunction; at high doses, risk for hypertension, sweating, and tremors
Norepinephrine–dopamine reuptake inhibitor	Block reputake of norepinephrine and some dopamine	Buproprion (Wellbutrin)	Dry mouth, dizziness, constipation, nausea/vomiting, blurred vision, agitation, seizure risk at high doses (0.4%)
Serotonin antagonist and reuptake inhibitors	Block 5-HT2a receptor and block reuptake of serotonin	Trazodone (Desyrel), nefazodone (Serzone)	Dry mouth, drowsiness, dizziness or lightheadedness, nausea/vomiting, constipation, blurred vision, priapism (1 in 15,000) with Trazodone
Norepinephrine antagonist and serotonin antagonist	Block α_2-adrenoreceptor and 5-HT_2 and 5-HT_3 receptors which leads to increased norepinephrine and specific serotonin production	Mirtazapine (Remeron)	Drowsiness, increased appetite, weight gain

serotonin. Our current understanding of the role of dopamine in depression is poor.

2. Monoamine Oxidase Inhibitors

The antidepressants that inhibit the degradation of neurotransmitter act by interfering with the enzyme monoamine oxidase, which destroys excess norepinephrine, serotonin, and dopamine in the synapse and the presynaptic terminal. Like the tricyclics, MAOIs have unpleasant side effects, including dizziness, sleep disturbances, sedation, fatigue, general weakness, hyperreflexia, dry mouth, and gastrointestinal disturbances. More serious is the fact that MAOIs can interact with tyramine (a protein building block for norepinephrine that is found in many common foods) and produce lethal hypertension. As such, individuals taking MAOIs must adhere to a tyramine-free diet, avoiding such foods as cheese, smoked meats, wine and beer, and yeast.

3. Serotonin and Norepinephrine Antagonists

Mirtazapine (Remeron) is a novel antidepressant that appears to produce antidepressant effects by blocking both the α_2-adrenoreceptor and two postsynaptic serotonin receptors, 5-HT2 and 5-HT3, which increases norepinephrine activity and specific serotonin activity in the brain. Common side effects of mirtazapine include drowsiness, increased appetite, and weight gain.

Currently, practice guidelines recommend SSRIs and some of the newer antidepressants (e.g., buproprion, nefazodone, and venlafaxine) as first choice antidepressants given their relatively modest side effect profile. For nonresponders to the first choice antidepressant, current guidelines suggest consideration of tricyclic antidepressants followed by MAOIs.

B. Electroconvulsive Treatment

ECT is the second form of biological treatment for depression and is currently used predominately with individuals experiencing severe and/or intractable depression that has not responded to antidepressant medications and psychotherapy. ECT involves the induction of a general seizure via the application of a brief electrical stimulus through electrodes that are placed on the scalp. Individuals are placed under general anesthesia and given a muscle relaxant in order to prevent discomfort and injury resulting from the seizure.

It is the seizure that is believed to produce the antidepressant effect, although the mechanism of action remains unclear. Animal studies suggest that ECT exerts its effects via the enhancement of norepinephrine and serotonin systems, paralleling the effects of antidepressant medications. In addition, ECT appears to affect the dopaminergic, cholinergic, GABA, and opioid systems.

Considerable research has supported the short-term efficacy of ECT in alleviating depressive symptoms, particularly those that are accompanied by psychotic symptoms: ECT produces significant improvement in approximately 80–90% of individuals with severe depression. Research has shown that both the dosage of the electrical current and the placement of the electrodes (unilateral or bilateral) are related to the alleviation of depressive symptoms and the production of unwanted side effects (mostly short-term memory loss and other cognitive impairments, including mild disorientation). Bilateral stimulation produces greater and more rapid improvement from depressive symptoms than does unilateral stimulation; however, bilateral stimulation also produces greater short-term disorientation and retrograde amnesia. High-dose unilateral stimulation, while producing fewer side effects than bilateral stimulation, appears to improve depressive symptoms more than low-dose unilateral stimulation, although not to the extent of bilateral stimulation.

C. Alternative Treatments

Little research exists to support the efficacy or mechanism of action of alternative biological treatments, generally in the form of herbal supplements. For example, St. John's Wort, a popular herbal supplement, appears to inhibit the reuptake of serotonin, norepinephrine, and dopamine. Several studies have shown it to be relatively effective in alleviating mild to moderate depression; however, given the significant methodological flaws present in these studies, no general consensus currently exists regarding the effectiveness of this or any other particular supplement.

D. Limitations of Biological Treatments

Although considerable evidence exists demonstrating the effectiveness of antidepressant medication in the alleviation of depressive symptoms, antidepressant medications are not a panacea. Approximately 30–40% of individuals with depression will not respond to antidepressants. In addition, no empirical data exist to guide practitioners in the selection of specific antidepressants, or even classes of antidepressants, over another. Practice guidelines suggest that failure to respond to an antidepressant in one class will likely predict failure to respond to other antidepressants of the same class, but no empirical data exist to conclusively support this proposition.

Furthermore, the relapse rates are very high for individuals who have been treated exclusively with biological treatments. Approximately 40–60% of individuals will relapse if antidepressant medication is discontinued within the first few months of a response. Up to 60% of individuals treated with ECT will relapse in the following year, particularly if they do not continue on antidepressant medication after the ECT regimen. Given the increasingly accepted conception of depression as a chronic, relapsing disorder, practitioners have begun moving toward

prescribing biological treatments over many years, often beyond the length of the original depressive episode.

Research on the use of antidepressant medication with children and adolescents remains sparse, and the few well-designed experiments that have been conducted have not shown that antidepressants are more effective than placebo. There also exists little research that has investigated the safety of these medications with pregnant women or women who are breast-feeding.

VI. PSYCHOLOGICAL TREATMENTS FOR DEPRESSION

There currently exist many different psychological treatments with varying levels of empirical support for their efficacy in treating depression in adults and a few that have been evaluated with children and adolescents. Although each has been developed and studied in its pure form, most psychotherapy available in the community incorporates some elements from more than one theoretical orientation.

A. Psychodynamic Psychotherapy

Psychodynamic psychotherapies evolved from psychoanalytic therapies, and as such they were originally designed to assist patients in the modification of their personality. This task occurs via the uncovering and bringing to awareness of unconscious conflicts that interfere with functioning. Recent adaptations of psychodynamic psychotherapies that have focused on depression emphasize more active approaches while continuing to uncover unconscious conflicts. Psychodynamic psychotherapies tend to emphasize the development of a therapeutic alliance that can increase patients' self-efficacy with respect to problem solving. Once the therapeutic alliance is developed, patients are then better able to gain insight into their problems by learning more about their relationship patterns, which then leads to increased potential for change. Some of this insight is developed via an exploration of the therapist–patient relationship and an examination of the ways in which this relationship mirrors the patients' real-world relationships.

When compared with other forms of psychotherapy for adults, psychodynamic psychotherapy tends to perform equivalently. Unfortunately, few studies have successfully compared psychodynamic psychotherapy with placebo conditions; therefore, the extent to which psychodynamic psychotherapy offers benefits that are particular to the psychoanalytic orientation remains in dispute.

B. Interpersonal Therapy

Interpersonal therapy (IPT) attempts to reduce depressive symptoms by focusing on current interpersonal problems. Specifically, IPT examines grief, role conflicts in relationships, role transitions, and social deficits, all in the context of problematic relationships. Depressed individuals are asked to pay close attention to all of their social interactions and social disappointments. By carefully examining their own role, patients become better able to reconstruct (or construct new) relationships more productively.

Empirical support exists for the efficacy of IPT in the treatment of depression in adults, and preliminary evidence suggests that it may be useful with adolescents. Several studies have found that it consistently outperforms control conditions and produces results at least equivalent to those of cognitive therapy and antidepressant medication.

C. Behavior Therapy

The primary goal of most behavior therapies is to increase the amount of pleasurable activity in patients' lives. Patients learn to monitor the fluctuations in their mood over the course of the week. They are taught to note what events produce positive changes in mood and what events bring about negative changes in mood. The therapist and patient then attempt to institute changes in the patient's life that would bring about more positive mood states. In situations in which the patient lacks the skills necessary to bring about changes in behavior, the therapist and the patient work together to enhance these skill deficits. Social skill training, relaxation skills, and assertiveness training are all examples of skill-building exercises in which patients may engage.

Empirical support exists for the effectiveness of behavior therapy in the treatment of depression in adults, although it has not been as extensively evaluated in clinical populations. Several studies have found results equal to those of other forms of psychotherapy, and one study found that behavior therapy produced comparable results to those of the antidepressant amitriptyline.

D. Cognitive Therapy

Cognitive therapy reduces depressive symptoms by addressing the maladaptive and pessimistic thinking in which depressed individuals engage. Working collaboratively, the individual and therapist attempt to identify and then change those elements of a patient's thinking and behavior that are contributing to the depressive symptoms. Once patients learn about the relationships among life events, thinking, and mood, they are taught to recognize the presence of "negative automatic thoughts" and pessimistic thinking that contribute to the maintenance of their depressed mood. These pessimistic thinking styles are examined together by the therapist and patient in order to consider their validity and utility. Patients learn how to generate alternative explanations for events, search for relevant information, and then decide on the most realistic explanation. Often, the most realistic explanation is less hopeless than the original belief and can lead to behavior change. When the most realistic explanation continues to be depressive, the patient and therapist work together to generate realistic solutions to the problems.

Considerable evidence exists to support the efficacy of cognitive therapy in the treatment of depression in both adults and adolescents, and many researchers consider cognitive therapy to be superior to many other forms of therapy for depression. Researchers are currently investigating the extent to which cognitive therapy can produce results equivalent to or better than those produced by antidepressant medication. Long-term data suggest that cognitive therapy is effective in the prevention of relapse, even more effective than that afforded by antidepressant medication.

E. Alternative Models of Therapy

Although different psychological orientations have led to the development of discrete psychological treatments, alternative models of psychotherapy complement these existing theoretical orientations. For example, cognitive and behavioral therapy techniques are often consolidated into an overarching cognitive–behavioral therapy. Moreover, modern cognitive therapy approaches are beginning to focus more on the therapist–patient relationship in ways that were originally developed by psychodynamic psychotherapies.

In addition to developments in individual-based psychotherapy, alternative models of therapy have also effectively expanded this focus to include multiple participants. For example, research suggests that group psychotherapy for depression can be as effective in treating depression as individual psychotherapy, particularly those group therapies that utilize a cognitive or behavioral approach. Family and marital treatments for depression also have considerable support for their effectiveness in treating depression in adults. In general, both group and family-based approaches work within their respective theoretical orientations to enhance the social support of the participants. Group therapy accomplishes this task by providing a novel social support system, and family-based approaches attempt to modify existing social networks.

F. Limitations of Psychological Treatments

One of the primary limitations of psychological treatments is the difficulty that researchers have in identifying the specific ingredients of the treatment that are most responsible for the alleviation of depressive symptoms. Very few psychological treatments have been compared against "placebo" psychological treatments. This limitation, coupled with the fact that placebo psychological treatments tend to produce some reduction in depressive symptoms, prevents researchers from definitively knowing what aspects of their treatment are acting to reduce depressive symptoms. A second important limitation lies in the differences that exist between the treatment providers in the community and those utilized in treatment studies. The treatment providers in depression treatment research are rigorously trained and supervised while providing the treatment. As such, it is plausible that their skill level is not representative of the skill level of the average community treatment provider, for whom there currently exists a wide range of formal training. A third limitation is the dearth of rigorous, controlled studies evaluating the effectiveness of psychotherapy with children. Although evidence supports the efficacy of cognitive–behavioral treatment with adolescents, less is known about the efficacy of other forms of treatment with adolescents and younger children.

VII. COMBINED BIOLOGICAL AND PSYCHOLOGICAL TREATMENTS

Given the strengths of both the biological and the psychological perspectives on depression, many

clinicians and researchers have assumed that combining different forms of treatments would provide individuals with the most benefit. Antidepressant medications could reduce many of the physiological symptoms of depression (e.g., sleep and appetite dysregulation) while psychotherapy could reduce the maladaptive cognitions and behaviors. However, the research evidence supporting the efficacy of combined treatment for depression has produced mixed results. Adding psychological treatments to biological treatments has been shown to reduce relapse rates; however, various studies examining the combination of antidepressant and psychotherapy from the outset of treatment have not found significant advantages of using either antidepressant medication or psychotherapy alone.

At this point, it is unclear to what to attribute this lack of a consistent effect, although some have argued that the modest advantages produced by combined treatment may not yield statistically significant effects in single research studies, and thus more sophisticated research investigations are warranted. For example, analyses that combine results from single studies into a larger meta-analytic study have demonstrated the advantages of combined treatment for depression, particularly for patients with more severe levels.

Thus, researchers have begun to more closely examine patient variables that might contribute to improved response from combined treatment. For example, a recent large-scale study conducted by Martin Keller and associates clearly demonstrated the advantages of combined treatment over single modality treatment for chronic depression. In this study, the researchers examined the efficacy of combining nefazodone and a variant of cognitive–behavioral treatment for patients with chronic depression (defined as having significant depressive symptoms for more than 2 years). Results showed that the combination of treatments was significantly more effective than either treatment alone. Among the 519 subjects who completed the study, the rates of response after 12 weeks (as defined by either a complete remission or a significant reduction in depressive symptoms) were 55% in the nefazodone group, 52% in the cognitive–behavioral group, and 85% in the combined treatment group.

Currently, the evidence is strongest for the proposition that combined treatment is more effective for individuals with more severe and chronic levels of depression. However, there remain considerable gaps in our knowledge of the effects of combined treatment on depression.

VIII. SUMMARY AND FUTURE DIRECTIONS

Depression is a serious disorder with significant economic and social consequences that appears to be increasing in prevalence. Attempts to understand the causes of depression must encompass both biological and psychological perspectives. Treatments of depression offer significant hope: Up to 70% of individuals will respond to antidepressant medication; specific forms of psychotherapy produce similar rates of symptom alleviation. ECT is even more effective: Up to 90% of people will respond to treatment. However, depression appears to be a chronic, relapsing disorder. The risk for relapse remains high in both treated and untreated individuals.

Future research will likely continue to link the biological and psychological nature of depression. Linkage of specific receptor sites with specific symptoms of depression, imaging studies that examine the extent to which psychotherapies produce noticeable biological changes, and more studies that explore the potential for joint biological and psychological treatments are all on the horizon. In addition, more research is needed to better understand the extent to which current theories of depression apply to children and adolescents.

See Also the Following Articles

ALCOHOL DAMAGE TO THE BRAIN • ALZHEIMER'S DISEASE, NEUROPSYCHOLOGY OF • ANGER • APHASIA • BASAL GANGLIA • BEHAVIORAL NEUROGENETICS • COGNITIVE PSYCHOLOGY, OVERVIEW • DEMENTIA • HUMOR AND LAUGHTER • MANIC–DEPRESSIVE ILLNESS • NEUROPSYCHOLOGICAL ASSESSMENT • PSYCHOACTIVE DRUGS • PSYCHONEUROENDOCRINOLOGY • PSYCHOPHYSIOLOGY • SUICIDE

Suggested Reading

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders,* 4th ed. American Psychiatric Association, Washington, DC.

Blazer, D. G., Kessler, R. C., McGonagle, K. A., and Swartz, M. S. (1994). The prevalence and distribution of major depression in a national community sample: The National Comorbidity Survey. *Am. J. Psychiatr.* **151,** 979–986.

Consensus Development Panel (1985). Mood disorders: Pharamacologic prevention of recurrences. *Am. J. Psychiatr.* **142,** 469–476.

Consensus Development Panel (1986). Report on the NIMH–NIH Consensus Development Conference on electroconvulsive therapy. *Psychopharmacol. Bull.* **22,** 445–502.

DeRubeis, R. J., and Crits-Cristoph, P. (1998). Empirically supported individual and group psychological treatments for adult mental disorders. *J. Consulting Clin. Psychol.* **66,** 37–52.

Elkin, I., Shea, M. T., Watkins, J. T., Imber, S. D., Sotsky, S. M., Collins, J. F., Glass, D. R., Pilkonis, P. A., Leber, W. R., Docherty, J. P., Fiester, S. J., and Parloff, M. B. (1989). National Institute of Mental Health treatment of depression collaborative research program. *Arch. Gen. Psychiatr.* **46,** 971–982.

Fava, G. A., Rafanelli, C., Grandi, S., Canestrari, R., and Morphy, M. A. (1998). Six-year outcome for cognitive behavioral treatment of residual symptoms in major depression. *Am. J. Psychiatr.* **155,** 1443–1445.

Horst, W. D., and Preskorn, S. H. (1998). Mechanism of action and clinical characteristics of three atypical antidepressants: Venlafaxine, nefazodone, buproprion. *J. Affective Disorders* **51,** 237–254.

Keitner, G. I., and Miller, I. W. (1990). Family functioning and major depression: An overview. *Am. J. Psychiatr.* **147,** 1128–1137.

Keller, M. B., McCullough, J. P., Klein, D. N., Arnow, B., Dunner, D. L., Gelenberg, A. J., Markowitz, J. C., Nemeroff, C. B., Russell, J. M., Thase, M. E., Trivedi, M. H., and Zajecka, J. (2000). A comparison of nefazodone, the cognitive–behavioral–analysis system of psychotherapy, and their combination for the treatment of chronic depression. *N. Engl. J. Med.* **342,** 1462–1470.

Kessler, R. C., Nelson, C. B., McGonagle, K. A., Liu, J., Swartz, M., and Blazer, D. G. (1996). Comorbidity of DSM-III-R major depressive disorder in the general population: Results from the National Comorbidity Survey. *Br. J. Psychiatr.* **168**(Suppl. 30), 17–30.

Luborsky, L., Mark, D., Hole, A. V., Popp, C., Goldsmith, B., and Cacciola, J. (1995). Supportive–expressive dynamic psychotherapy of depression: A time-limited version. In *Dynamic Therapies for Psychiatric Disorders (Axis I)*. (J. Barber and P. Crits-Cristoph, Eds.). Basic Books, New York.

Mesulam, M. M. (2000). *Principles of Behavioral and Cognitive Neurology*, 2nd ed. Oxford Univ. Press, Oxford.

Paykel, E. S., Scott, J., Teasdale, J. D., Johnson, A. L., Garland, A., Moore, R., Jenaway, A., Cornwall, P. L., Hayhurst, H., Abbott, R., and Pope, M. (1999). Prevention of relapse in residual depression by cognitive therapy. *Arch. Gen. Psychiatr.* **56,** 829–835.

Stahl, S. M. (1998). Mechanism of action of serotonin selective reuptake inhibitors: Serotonin receptors and pathways mediate therapeutic effects and side effects. *J. Affective Disorders* **51,** 215–228.

Dopamine

LISA A. TAYLOR and IAN CREESE
Rutgers University

GLOSSARY

avoidance learning A behavioral test in which subjects can avoid a noxious stimulus if a correct response is elicited.

G protein A guanine nucleotide-binding regulatory protein that mediates the interaction between extracellular receptors and intracellular effector molecules. The Gi protein is the inhibitory G protein that reduces subsequent intracellular effects. The Gs protein is the stimulatory G protein that increases subsequent intracellular effects.

intracranial self-stimulation A task in which subjects are rewarded with electrical stimulation to brain areas such as the medial forebrain bundle.

operant task A task in which an animal is rewarded for making a particular motor response such as a bar press.

Dopamine (3-Hydroxytyramine) was initially characterized as a neurotransmitter in the central and peripheral nervous system during the 1950s. Although comprising less than 1% of the neurons in the brain, dopamine mediates a variety of physiological functions and modulates many behavioral states. Given the paucity of noninvasive techniques to study dopamine in the central nervous system (CNS) of humans, most studies characterizing the role of dopamine in the brain have been carried out in experimental animals or tissue culture expression systems or by extrapolation from human responses to dopaminergic drugs. With the relatively recent development of more sophisticated neuroimaging techniques, however, studies of dopamine function in the human brain are increasing. Since the focus of this encyclopedia is the human brain, studies using humans are emphasized. Data from animal studies or tissue culture expression systems are presented when there is a lack of data from human subjects or when specific examples illuminate the human literature.

I. ANATOMICAL DISTRIBUTION IN THE CENTRAL NERVOUS SYSTEM

There are several dopamine-containing pathways in the CNS. The nigrostriatal dopamine pathway accounts for approximately 70% of the dopamine in the brain. Cells bodies in this pathway are located in the substantia nigra pars compacta and project to the caudate, putamen, and the globus pallidus (Fig. 1). An interesting characteristic of these dopamine neurons is that they contain extensive dendritic trees, which extend ventrally into the substantia nigra pars reticulata. Dopamine release occurs from these dendrites in addition to the axon terminals. Deterioration of the nigrostriatal pathway underlies Parkinson's disease (PD).

The mesocorticolimbic dopaminergic pathway originates in the ventrotegmental area (VTA) and innervates the olfactory tubercle, nucleus accumbens, septum, amygdala, and adjacent cortical structures (medial frontal, anterior cingulate, entorhinal, perirhinal, and piriform cortet pallidus) (Fig. 1).

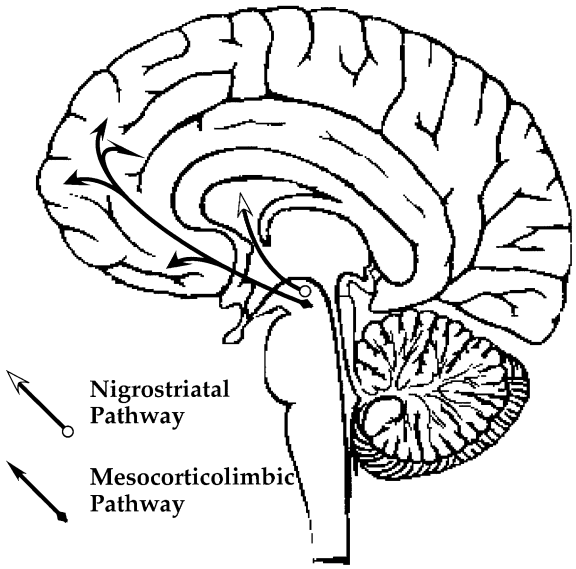


Figure 1 The nigrostriatal and mesocorticolimbic dopamine pathways.

The substantia nigra and VTA dopamine cell bodies are often referred to as the A-9 and A-10 nuclear groups, respectively, following the original designation of Dahlsstrom and Fuxe from their pioneering rodent studies using a novel technique that made dopamine neurons fluorescent. However, more detailed immunohistochemical studies suggest that the A-9 and A-10 nuclear groups are a continuum, with laterally situated cells innervating the striatum and medial cells innervating mesolimbic and mesocortical areas.

The tuberoinfundibular dopamine pathway originates in the arcuate and periventricular nuclei of the hypothalamus and projects to the intermediate lobe of the pituitary and the median eminence. Dopamine released from these neurons is secreted into the hypophysed and portal blood regulates prolactin secretion from the pituitary through inhibitory D_2 receptor on nanotrophic cells.

Other pathways containing dopamine include (i) the incertohypothalamic neurons, which connect the dorsal and posterior hypothalamus with the dorsal anterior hypothalamus and lateral septal nuclei; (ii) the medullary periventricular group, which includes dopamine cells of the dorsal motor nucleus of the vagus nerve, the nucleus tractus solitarius, and the tegmentum radiation in the periaqueductal gray matter; (iii) the interplexiform amacrine-like neurons, which link the inner and outer plexiform layers of the retina; and (iv) the periglomerular dopamine cells in the olfactory bulb, which link mitral cell dendrites in adjacent glomeruli.

II. SYNAPTIC DOPAMINE

A. Synthesis

In dopamine-secreting cells, the first step in dopamine synthesis is the conversion of dietary tyrosine into L-3,4-dihydroxyphenylalanine (L-DOPA) (Fig. 2). This reaction is catalyzed by the rate-limiting enzyme tyrosine hydroxylase (TH). L-DOPA is then converted to dopamine via the enzyme L-aromatic amino acid decarboxylase.

TH is composed of four identical subunits and contains iron ions, which are required for its activity. The cofactors oxygen and tetrahydrobiopterin are also required for its activity. A single gene encodes TH, although in humans four isoforms have been shown to result from alternative splicing of the primary transcript. TH is present in both soluble (cytoplasmic) and membrane-bound forms.

Under basal conditions TH is nearly saturated by tyrosine. The observation that pharmacological agents

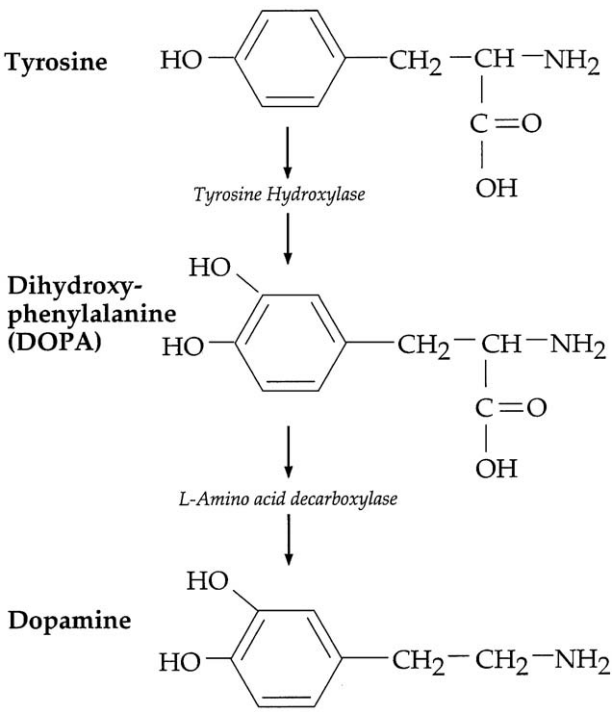


Figure 2 The dopamine synthesis pathway.

known to block TH activity have greater effects on extracellular dopamine levels than agents that block dopa-decarboxylase indicates that the rate-limiting step for dopamine synthesis is tyrosine hydroxylation of tyrosine to L-DOPA by TH. Thus, increasing levels of tyrosine by dietary modifications may also regulate dopamine synthesis.

The conversion of L-DOPA to dopamine by dopamine β-hydroxylase results from the removal of a hydroxyl group and requires pyridoxal 5-phosphate (vitamin B_6) as a cofactor.

Dopamine synthesis is regulated in a variety of ways. End product inhibition is the major regulator when dopamine neuronal activity and release are low. In contrast, when dopaminergic fibers are electrically stimulated, TH activity is increased. This increase appears to be a function of enhanced enzyme substrate kinetics, in part caused by TH phosphorylation. This results in a net decrease in affinity of TH for dopamine, which overrides end product inhibition.

B. Storage and Release

There are two release mechanisms for dopamine. The first is calcium-dependent, tetrodotoxin (TTX)-sensitive, vesicular release at the axon terminal that occurs following an action potential. The second is calcium and TTX independent and occurs following the administration of stimulant drugs that reverse the direction of the dopamine transporter (DAT). Under normal, nondrug conditions the DAT carries released dopamine from the extrasynaptic space back into the terminal region.

Pharmacological studies indicate that dopamine exists in three pools or compartments within the axon terminal. Two of these are vesicular dopamine stores, one containing newly synthesized dopamine and the second containing a longer term store of dopamine. A cytoplasmic pool has also been identified, and it consists of dopamine newly taken up by the dopamine transporter.

C. Inactivation

Dopamine inactivation is accomplished by a combination of reuptake and enzymatic catabolism. Dopamine uptake is an energy-dependent process that requires sodium and chloride. Catabolism occurs through two enzymatic pathways (Fig. 3). Although it is not clear how much dopamine is catabolized in each of these

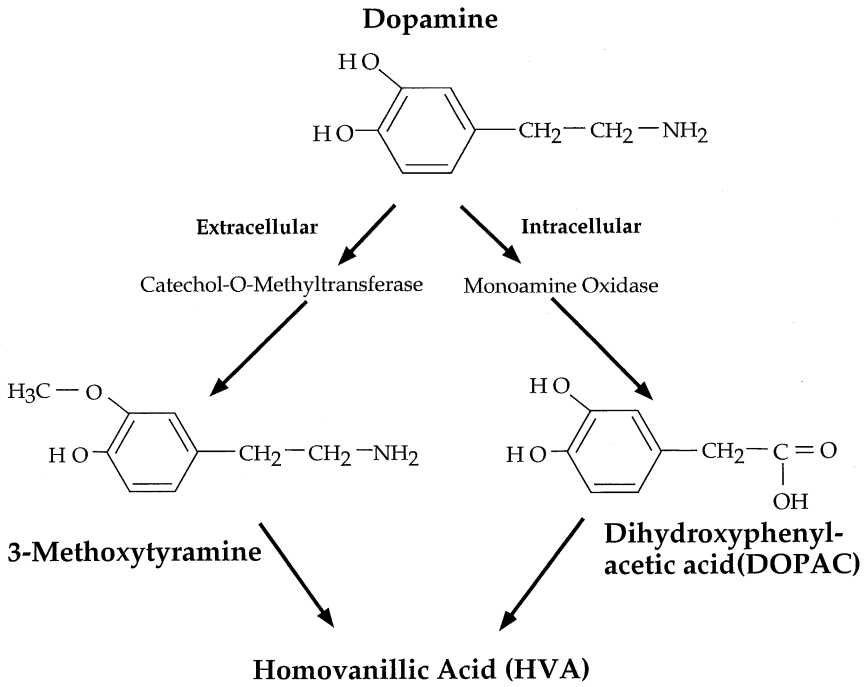


Figure 3 The dopamine metabolic pathway.

pathways in the human brain, almost 90% of catabolism in the rat striatum takes place via the monoamine oxidase (MAO) pathway. In the rat, the level of 3,4-dihydroxylphenylacetic acid is thought to reflect catabolism of intraneuronal dopamine, which includes dopamine that is taken back up by the dopamine transporter, whereas 3-methoxytyramine levels are thought to reflect metabolism of extracellular dopamine. Cerebrospinal fluid (CSF) levels of homovanillic acid (HVA) are often used as an indicator of dopaminergic activity in humans.

III. DOPAMINE RECEPTORS

A. Subtypes

Initial biochemical and pharmacological studies indicated the presence of two dopamine receptor subtypes with differential coupling to G proteins: D1 receptors stimulate adenylate cyclase through Gs protein, and D2 receptors inhibit adenylate cyclase through Gi protein. The recent application of molecular cloning techniques has revealed the existence of five dopamine receptor genes, each encoding a distinct seven-transmembrane spanning receptor. These receptors have been classified according to their structures and divided into two subfamilies, the D2-like receptors (D_2–D_4) and the D1-like receptors (D_1 and D_5). In the case of the D_2 receptor, two isoforms (D_{2L} and D_{2S}), generated by alternative splicing of the primary transcript, have been identified. Given their structural similarity, members of each subfamily exhibit similar pharmacological characteristics.

B. Distribution

The distribution of messenger RNA (mRNA) encoding all five dopamine receptors has been determined in the rat and human brain by *in situ* hybridization. High levels of D_1 mRNA are found in the caudate–putamen, nucleus accumbens, and olfactory tubercle; lower levels are found in lateral septum, olfactory bulb, hypothalamus, and cortex. In contrast, D_5 mRNA is predominantly localized to the hippocampus and the parafascicular nucleus of the thalamus.

The highest concentration of D_2 mRNA is present in the neostriatum, olfactory tubercle, substantia nigra, ventral tegmental area, and the nucleus accumbens. The levels of D_3 and D_4 receptor mRNAs are much lower in the brain relative to the level of D_2 mRNA. D_3 receptor mRNA is predominantly expressed in limbic areas, such as the nucleus accumbens, Islands of Calleja, bed nucleus of the stria terminalis, hippocampus, mammillary nuclei, and the substantia nigra. The areas with the highest levels of D_4 mRNA include the olfactory bulb, hypothalamus, and thalamus, with lower levels present in the hippocampus, cortex, and basal ganglia.

There is evidence in the rat that D_1 and D_2 receptor expression is segregated at the level of the striatum (caudate and putamen in the human). *In situ* hybridization analysis has reported two distinct post synaptic populations of striatal neurons that express D_1 or D_2 receptors: the enkephalinergic striatopallidal neurons, which express D_2 receptors, and the substance P, dynorphin-positive, striatonigral neurons, which express D_1 receptors. However, other groups using different methodologies have reported considerably more overlap in the expression of D_1 and D_2 mRNA in the striatum. They found significant coexpression of D_1 and D_2 mRNAs in the same striatal neurons. These studies suggest that at least a portion of D_1 and D_2 receptors is colocalized in the striatum.

Electrophysiological and neurochemical data suggest that a portion of dopamine receptors are located presynaptically on the axon terminals of striatal dopaminergic neurons where they serve as autoreceptors. These terminal autoreceptors modulate the synthesis and release of dopamine. These activation inhibits production and release of dopamine. These autoreceptors were originally classified as D_2 receptors. However, since D_2 and D_3 mRNA are present in nigral neurons D_3 as well as D_2 receptors are implicated as autoreceptors. Somatodendritic autoreceptors, responding to dendritically released dopamine, have also been observed. Activation of these receptors reduces the firing rate of dopamine neurons.

C. Signal Transduction Pathways

Like other G protein-coupled receptors, dopamine receptors transduce their signals through second messengers regulated by G proteins. In general, D1-like receptors stimulate adenylate cyclase via the Gs, and the D2-like receptors inhibit adenylate cyclase through the Gi. The signaling cascade is initiated by dopamine binding to the extracellular portion of the receptor. The binding of dopamine alters the receptor so that it activates a G protein present on the intracellular face of the neuron cell membrane. Once activated, G proteins modulate adenylate cyclase

activity. G proteins bind guanosine triphosphate (GTP) or guanosine diphosphate (GDP). Dopamine binding to its receptor causes GDP to be replaced with GTP. Then, in the case of the D1-like receptors, a GTP–Gs complex is formed and associates with the catalytic subunit of adenylate cyclase, stimulating the conversion of ATP to cyclic (AMP) adenosine monophosphate. The GTP–Gs protein and the catalytic subunit of adenylate cyclase together constitute the active form of adenylate cyclase. The association of the G protein with the catalytic subunit of the cyclase also results in the hydrolysis of GTP to GDP. This hydrolysis causes the G protein to dissociate from the catalytic subunit and to reassociate with the receptor. This reaction terminates the synthesis of cyclic AMP. Therefore, the duration of cyclic AMP synthesis is regulated by GTPase activity. In the case of the D2-like receptors, receptor activation causes GTP binding to Gi, and this complex inhibits cyclase activity.

It should also be noted that several other second messenger systems are involved in D2-like receptor-mediated signal transduction, such as the stimulation of aracadonic acid metabolism and the activation of potassium channels.

D. Regulation

Most studies concerned with the regulation of CNS dopamine receptors *in vivo* have examined the effects of long-term treatment with receptor agonists, antagonists, or denervation induced by 6-hydroxydopamine, a selective neurotoxin for catecholaminergic neurons, on striatal D2-like dopamine receptors.

Studies examining chronic agonist exposure have yielded mixed results. Repeated treatment with amphetamine, which increases synaptic levels of dopamine, both increased and had no effect on the density of D2-like receptors in the striatum, probably depending on dose and timing. Treatment with systemic L-DOPA, which also increases synaptic levels of dopamine, has been reported to decrease the density of D2-like receptors in the striatum.

Treatment with D1-like or D2-like antagonists has been shown to increase the density of D1-like and D2-like receptors, respectively, without changing the affinity of these receptors for dopamine. Similarly, denervation following 6-hydroxydopamine administration also leads to an increased dopamine D2-like, but not D1-like, receptor density as well as an enhanced behavioral responsiveness to dopamine. Thus, it appears that dopamine D2-like receptors upregulate in response to reduced levels of dopamine. It is not clear, however, if the molecular mechanism responsible for lesion-induced upregulation of dopamine receptors is the same as that underlying the upregulation observed after long-term antagonist treatment. The fact that D1-like receptors upregulate after long-term antagonist treatment, but not after denervation, suggests that the mechanisms must be different, at least for D1-like receptors. In the case of long-term antagonist treatment, it is also possible that the intrinsic properties of the antagonist may play a role in the upregulation of dopamine receptors.

The pathway of protein synthesis, from the initial stage of gene transcription to mRNA splicing, translation, protein processing, and insertion into the membrane, provides a number of potential points of control for the regulation of dopamine receptor number. The rate of receptor removal from the membrane and the recycling or degradation rates are also factors that contribute to the number of receptors available for binding.

It should be noted that the molecular mechanisms underlying dopamine receptor trafficking (another mechanism regulating dopamine receptor expression) are only beginning to be studied.

IV. DOPAMINE AND NEUROPSYCHIATRIC DISORDERS

Many lines of evidence suggest a role for dopamine in neuropsychiatric disorders such as PD, schizophrenia, attention deficit hyperactivity disorder (ADHD), and drug abuse. The first evidence for dopamine's involvement in these disorders came from either postmortem histological/neurochemical studies or the observation that the drugs used to treat these disorders either increased or blocked dopamine neurotransmission in the brain. Since the advent of neuroimaging techniques such as positron emission topography (PET), a majority of the current human studies use this technique to further define the role of dopamine in these disorders.

PET enables the direct measurement of components of dopamine neurotransmission in the living human brain by using radiotracers, which label dopamine receptors, dopamine transporters, and precursors of dopamine or compounds that have specificity for the enzymes that degrade dopamine. Certain types of PET

studies also provide information on regional brain metabolism or blood flow, thus, PET can be used to assess the functional consequences of changes in brain dopamine activity.

A. Parkinson's Disease

PD is a neurodegenerative disorder of unknown etiology that is associated with the degeneration of dopamine neurons primarily in the substantia nigra and manifested by disturbances in the motor system. There is evidence for genetic and environmental causes. The symptoms of PD can be classified as positive or negative. The positive symptoms of PD consist of tremor, muscular rigidity, and involuntary movements. Negative symptoms include bradykinesia, postural disturbances, and cognitive impairments. During the early course of the disease, unilateral symptoms may initially appear. Often, the disease begins with a mild tremor and some muscular rigidity. As the disease progresses, the symptoms present bilaterally, the tremor is exacerbated, and bradykinesia appears. During the later stages of the disease patients are incapacitated and are usually confined to a wheelchair.

L-DOPA is probably the single most effective medication for controlling the early symptoms of Parkinson's disease. L-DOPA is transported to the brain and taken up by the remaining dopaminergic cells, where it is converted into dopamine and increases its synaptic availability. L-DOPA therapy is effective in most patients for several years. However, as the loss of nigrostriatal neurons increases, symptoms continue to worsen and the dose of L-DOPA must be increased. As the dose is increased, it is not uncommon for patients to develop adverse effects, which may not be tolerable. In such cases, a MAO inhibitor is sometimes coadministered with a lower dose of L-DOPA to prolong the actions of dopamine at the synapse. Alternatively, direct-acting dopamine agonists, such as bromocriptine and pergolide, can be used to treat PD.

B. Schizophrenia

Schizophrenia is a serious psychiatric disorder affecting approximately 1% of the world population. Symptoms usually appear in late adolescence or early adulthood. This disease has a devastating impact on the lives of patients as well as their families. Positive symptoms include hallucinations and delusions, whereas negative symptoms consist of flattened affect social withdrawal and cognitive deficits. Several lines of evidence emerged in the 1970s that suggested alterations in dopamine transmission might constitute the pathophysiological basis for schizophrenia. Amphetamine and cocaine, which enhance dopamine transmission in the brain, can produce paranoid psychosis similar to that observed in patients with paranoid schizophrenia and exacerbated symptoms in schizophrenics. Also, a significant positive correlation between the clinical potency of neuroleptic drugs and their ability to block D2-like dopamine receptors has been consistently observed.

Although early studies suggested a postsynaptic locus for the disease because increased D2-Like receptors were observed in postmortem studies evaluating dopamine receptors, it seems likely that these results may have been caused by prior neuroleptic drug treatment. Neuroimaging studies also indicate that increased dopaminergic neurotransmission is involved in schizophrenia and is associated with activation of psychotic symptoms. Furthermore, enhanced dopamine transmission is detected in drug-naïve patients experiencing their first episode of the illness and is not detected in patients during remission, suggesting that the hyperdopaminergic state associated with schizophrenia fluctuates over time.

With respect to the negative symptoms of schizophrenia, PET studies evaluating regional brain glucose metabolism in untreated schizophrenic patients have reported reduced rates of glucose metabolism in neocortical areas, although these differences did not always reach statistical significance. Importantly, there may be a tendency toward a more pronounced reduction in metabolism in patients exhibiting negative symptoms. Since the premise underlying these studies is that glucose consumption in nerve cells is directly proportional to the impulse activity of neurons, it was postulated that decreased dopamine transmission in the mesocortical dopaminergic pathway, which is known to modulate the activity of prefrontal cortical neurons, might underlie the negative symptoms observed in schizophrenia. Support for this hypothesis has been obtained from a variety of studies. For example, animal studies have demonstrated that lesions of the prefrontal cortex in nonhuman primates lead to cognitive disturbances and poor social skills that resemble some of the negative symptoms observed in schizophrenic patients. In addition, many patients with primarily negative symptoms exhibit ventricular enlargement suggesting a loss of neurophil.

C. Attention Deficit Hyperactivity Disorder

A role for altered dopamine neurotransmission as the underlying cause of ADHD has been suggested by several observations. First, drugs used to treat ADHD, such as methylphenidate (Ritalin) and amphetamine (Adderal), increase synaptic levels of dopamine in experimental animals and in human subjects. In patients with ADHD, the maximal therapeutic effects of these drugs occur during the absorption phase of the kinetic curve, which parallels the acute release of dopamine into the synaptic cleft. These drugs are said to have "paradoxical" effects in ADHD children because they cause hyperactivity in normal children but have a "calming" or cognitive-focusing effect on ADHD children. Second, molecular genetic studies have identified genes that encode proteins involved in dopamine neurotransmission as candidate genes for ADHD. Third, neuroimaging studies have shown reduced activation in the striatum and frontal cortex of ADHD patients that is reversed by the administration of methylphenidate at least in a subset of children. Recently, functional magnetic resonance imaging studies have demonstrated differences between children with ADHD and normal controls in the degree of corticostriatal activation during a stimulus-controlled go/no-go task and its modulation by methylphenidate. Off drug, ADHD children showed impaired inhibitory control on this task and reduced striatal activation relative to the control subjects. Administration of methylphenidate significantly increased inhibitory control and frontostriatal activation in ADHD patients. These observations indicate that an optimal level of corticostriatal activation is necessary for subjects to display normal inhibitory control. Since methylphenidate increased both corticostriatal activation and inhibitory control, it follows that decreased corticostriatal activation and the poor inhibitory control may be due to reduced dopamine tone in the brain, perhaps in the striatum. However, it is equally likely that the mesolimbic dopamine system may mediate these effects by virtue of its afferent projections to the prefrontal cortex.

D. Psychostimulant Drug Abuse

Amphetamine and cocaine are psychostimulant drugs that are abused by humans because they produce feelings of euphoria. Both of these drugs increase extracellular levels of dopamine. Cocaine blocks the dopamine transporter and amphetamine primarily reverses it. Studies indicate these drugs are also primary reinforcers in animals because they will self-administer these drugs, and their propensity to do so is an excellent predictor of abuse liability in humans.

Many animal studies using a variety of different paradigms to study the role of dopamine in the rewarding properties of the psychostimulant drugs have suggested that dopamine is critically involved in this process in experimental animals. In the case of cocaine, studies carried out in human subjects have also supported the idea that enhanced dopamine transmission is responsible for the euphoric effects of this drug in humans. Dopamine appears to be a critical mediator of reward in the brain.

Dopamine also appears to play a role in craving, which often leads to relapse in abstinent human substance abusers. Brain imaging studies have identified the amygdala and the dopamine-rich nucleus accumbens as putative neuroanatomical substrates for cue-induced craving. Nucleus accumbens dopamine levels increased withdrawn from cocaine when they were exposed to cues that were previously associated with cocaine intake. Studies in humans examining the level of the dopamine metabolite HVA have reported that craving during abstinence is associated with increased HVA.

It also appears that the dopamine D_3 receptor may be involved in cocaine craving. Animal models of cocaine craving have shown that the D_3 selective partial agonist, BP 897, attenuates craving while lacking any intrinsic, primary rewarding effects.

A very interesting aspect of repeated amphetamine or cocaine administration is behavioral sensitization. Behavioral sensitization or reverse tolerance refers to the progressive augmentation of drug affects that are elicited by repeated administration of the drug. Sensitization to psychostimulants was characterized in experimental animals as early as 1932 and in humans in the 1950s. Behavioral sensitization develops when psychostimulant administration is intermittent. Tolerance develops during continuous administration.

V. DOPAMINE AND LEARNING

Animal research shows that dopamine is involved in the acquisition of operant tasks, avoidance learning, and intracranial self-stimulation. It is thought that dopamine in mesolimbic regions (most notably the nucleus accumbens) is involved in both the acquisition

and the maintenance of these behaviors. This conclusion was based primarily on results from studies using dopamine-depleting agents and dopaminergic drugs to evaluate their effects on acquisition or maintenance of reinforced behaviors. The results of acquisition studies showed that administration of dopamine antagonists or the dopamine-depleting agent, 6-hydroxydopamine, prevented the acquisition of these tasks. Reinterpretation of studies evaluating the effects of these agents on the maintenance of such behaviors has not been so clear-cut. For example, the acute administration of a dopaminergic antagonist after acquisition of an operant task often has no effect or increases the level of operant responding. This has been interpreted to be the result of a decrease in the perceived "salience" of the reward, whereas chronic administration of these agents blocks these behaviors. Recent *in vivo* voltammetry studies, which directly measure fast changes in dopamine release at the level of the synapse, carried out by Garris, Wightman, and colleagues have shown that evoked dopamine release in the nucleus accumbens is associated with the acquisition but not the maintenance of reinforced behavior, at least in the case of intracranial stimulation.

VI. DOPAMINE AND WORKING MEMORY

The prefrontal cortex receives elaborate dopamine inputs from the VTA and an optimal level of dopamine in the prefrontal cortex appears necessary for cognitive performance in experimental animals. Significant increases in dopamine levels in the dorsolateral prefrontal cortex have been observed in monkeys performing a delayed alternation task. In addition, reduced levels of dopamine in the prefrontal cortex have detrimental effects on spatial working memory tasks. It has also been reported that very large increases in dopamine levels in the prefrontal cortex can cause deficits on spatial working memory tasks. These deficits are ameliorated by the administration of dopamine D1-like drugs. Taken together, the previously mentioned observations support the idea that an optimal level of dopamine is required to perform spatial working memory tasks. If that level is reduced or increased, performance will be negatively affected.

Behavioral studies conducted in humans also indicate a role for prefrontal D1 receptors in working memory modulation in humans, although one study has also shown that the administration of the D2-like antagonist sulpride produced a dose-dependent impairment in spatial working memory.

Studies examining the performance of PD patients on working memory tasks also support the idea that dopamine may be involved in working memory in humans as well as in experimental animals. Imaging studies in PD patients have shown reduced fluorodopa uptake in the caudate nucleus and frontal cortex, which correlates with deficits in working memory as well as attention.

Some studies have also evaluated the effects of dopamine receptors in the ventral hippocampus on working memory tasks in rats. These studies suggest that D2-like but not D1-like receptors in the ventral hippocampus modulate spatial working memory.

See Also the Following Articles

CATECHOLAMINES • CHEMICAL NEUROANATOMY • ENDORPHINS AND THEIR RECEPTORS • MANIC–DEPRESSIVE ILLNESS • NOREPINEPHRINE • PARKINSON'S DISEASE • SCHIZOPHRENIA • WORKING MEMORY

Suggested Reading

Goldman-Rakic, P. S. (1998). The cortical dopamine system: Role in memory and cognition. *Adv. Pharmacol.* **42,** 707–711.

Laruelle, M. (2000). Imaging synaptic neurotransmission with in vivo binding competition techniques: A critical review. *J. Cereb. Blood Flow Metab.* **20**(3), 423–451.

Neer, E. J. (1994). G proteins: Critical control points for transmembrane signals. *Protein Sci.* **3**(1), 3–14.

Schultz, W., Tremblay, L., and Hollerman, J. R. (1998). Reward prediction in primate basal ganglia and frontal cortex. *Neuropharmacology* **37**(4–5), 421–429.

Sibley, D. R., Ventura, A. L., Jiang, D., and Mak, C. (1998). Regulation of the D1 dopamine receptor through cAMP-mediated pathways. *Adv. Pharmacol.* **42,** 447–450.

Spanagel, R., and Weiss, F. (1999). The dopamine hypothesis of reward: Past and current status. *Trends Neurosci.* **22**(11), 521–527.

Tedroff, J. M., (1999). Functional consequences of dopaminergic degeneration in Parkinson's disease. *Adv. Neurol.* **80,** 67–70.

Vallone, D., Picetti, R., and Borrelli, E. (2000). Structure and function of dopamine receptors. *Neurosci. Biobehav. Rev.* **24**(1), 125–132.

Dreaming

MARIO BERTINI
University of Rome

GLOSSARY

bizarreness The presence of highly unlikely elements in the dream narrative. Improbable events are usually associated with discontinuities of time, place, person, object, and action as well as incongruities of these features.

REM sleep The stage of sleep defined by the concomitant appearance of relatively low-voltage, mixed frequency electroencephalograph activity with episodic rapid eye movement (REM) and low-amplitude electromyogram.

Dreaming is a term interchangeably used to indicate both the function of generating a dream and the dream itself. That is a special kind of sleep mentation usually characterized by visual imagery, bizarreness, and hallucinatory and story-like quality.

I. HISTORICAL BACKGROUND

Dreaming is a complex psychobiological function, that eludes a precise and universally shared definition. Research into this important dimension of the human mind is continually evolving, and despite the fact that much progress has been made, we are still a long way from a satisfactory understanding of its biological and psychological implications.

Given the peculiar methodological limitations of the research in this area, it is worth presenting an overview of our current knowledge in a historical perspective. I hope it will allow the reader to better judge the validity and the possible lines of development of the field.

In the antiquity, sleep had a dual significance. On the one hand, it symbolized death, in the abandonment of the limbs and in shutting oneself away from the experience of the senses (Hesiod's Theogony depicted Sleep and Death as two brothers); on the other hand, sleep was like a door that, through the secret pathways of dreams, led to a dimension beyond intelligence and enabled humans to come into contact with the supernatural.

Even on a philosophical level, there were two somewhat opposite orientations that can still be recognized in some of today's scientific views. Some thinkers viewed sleep as a passive withdrawal from consciousness and dreaming was thus an illusion or a nonsense; according to others, sleep represented the theater of a different form of mental activity and dreams bore meanings.

In *Sleep and Wake* (453 BC), Aristotle defined sleep, and what takes place in it, as a "privation of wakefulness," i.e., as a purely negative concept—a pause in human sensory experience and intellectual functioning. During sleep, either intelligence is silent and everything is dark or it is deceived by a flow of disconnected fleeting images that give rise to dreams.

In *Republic*, Plato discussed the obscure depths of the psyche that are revealed through nocturnal dreams. Not only may the gods reveal themselves to man through dreams but also in the same way man reveals himself to himself. Even Heraclitus seems to hint at a different orientation of the psyche in the transition from wakefulness to sleep: "He who is awake lives in a single world common to all; he who is asleep retires to his own particular world."

Over the centuries, dreaming aroused more the curiosity than the systematic interest of researchers. The views of 19th-century neurophysiologists, strongly influenced by positivistic thought, essentially adhered to the original Aristotelian lines.

As a matter of fact, until recently, states of consciousness were believed to vary along one continuous dimension in parallel with behavioral performance, from deep coma to the opposite extreme of manic excitement. In this view, rational thought appeared as a form of mental activity exclusively belonging to waking states, whereas dreaming was simply the expression of the fragmentation and dissolution of thought, in close relationship with the lowering of vigilance and the deepening of sleep.

Only toward the end of the 19th century were some contributions made that considered dreaming an event deserving direct attention. The observations of two French scholars, the Marquis d'Hervey de Saint-Denis and Alfred Maury, are worth mentioning here, along with those of the initiator of experimental psychology, the German Wilhelm Wundt. However, the contribution that truly catalyzed the attention of contemporary culture was the brainchild of a scholar, Sigmund Freud, who had no real connection with the emerging experimental psychology approaches. Freud pointed out a methodological flaw in the way those who considered dreaming a completely meaningless activity examined the oneiric content. Researchers were trying to understand dreams according to the same register that was used for comprehending "normal" language instead of searching for the appropriate key to decrypt a language that is quite peculiar. According to Freud, this was like expecting to understand a rebus, with its apparently bizarre compositions, by reading it as a common rational message.

Freud's book on the interpretation of dreams, which was published in 1900, is a cornerstone in psychoanalytical theory and a significant historical event in contemporary culture. However, it is well-known that Freudian studies mainly focused on the contents of the phenomenon, whereas the proposed functional explanations suffered from the poor knowledge at the time of the central nervous system's physiology and of the organization of the mind in general, matters brought to the forefront of research by modern neuroscience.

In order to probe the functional basis of dreaming, it was necessary to shed light on objective, measurable aspects, such as the frequency and periodicity of dreaming, its universality, episodic and total duration throughout the night, and especially its neurophysiological regulation. The first step in solving these and other general problems was the discovery of some objective indicators that could reveal the presence of the dream itself.

As sometimes happens in science, the first curtain protecting the mystery of dreaming was lifted almost by chance by a young student working on his doctoral thesis in the early 1950s. Studying attention in children, under the supervision of Nathaniel Kleitman at the University of Chicago, Eugene Aserinsky observed that sometime after the onset of sleep, the subjects' eyes could be observed, under the closed eyelids, moving left and right and up and down while all the major body movements ceased. After several minutes, these eye movements would stop altogether, only to resume again with the same characteristics at more or less regular intervals throughout the night. Since the pattern of eye movements resembled that of an awake subject exploring a visual scene with his or her gaze, the hypothesis was put forward that dreaming occurs in these periods.

To test the hypothesis, the experimenters set out to awaken and question subjects at various times during the night. In most cases, subjects were able to report dream content when awakened during periods of eye movements, whereas they could rarely recall dreams if awakened while their eyes did not move. The results seemed to clearly confirm the conjecture, and Aserinsky's accidental observation took on the proportions of a profound scientific discovery.

II. THE SLEEP AND DREAMING LABORATORIES

Further confirmation of such association was obtained in the Chicago sleep laboratory. By the end of the 1950s, William Dement and Nathaniel Kleitman had obtained vivid, detailed reports of dreams with an incidence of 80% when awakenings occurred during periods of sleep characterized by rapid eye movements (REM sleep), whereas dreams could be reported in only 7% of the cases when subjects were awakened

while no rapid eye movements could be observed (non-REM sleep). Various demonstrations of proportionality between objective duration of REM sleep periods and subjective length of dreaming further confirmed the relation between the two events and prompted the hypothesis that dream activity continuously progresses throughout the length of a REM phase.

Beginning in the 1960s, many other sleep and dreaming laboratories were set up throughout the world in which subjects were monitored with electroencephalographic (EEG) and polygraphic recordings throughout their sleep. The observations made in these laboratories offered a convincing picture of the physiological processes accompanying the REM stage of sleep. The term REM soon came to represent a particular sleep stage that appeared to be the privileged place for dreams. Besides rapid eye movements, this phase is characterized by fast EEG activity, resembling that of the wake alert state, contrasted by muscle tone so relaxed that the sleeper is virtually paralyzed. Heart and breathing rate are higher and more irregular than in the other sleep stages. Penis erection in males and vaginal moistening in females have also been highlighted in connection with the REM stage of sleep.

Continuous laboratory monitoring has shown that in adult humans REM phases last, on average, 20–30 min each and reoccur regularly at intervals of approximately 90–100 min throughout sleep. Thus, each subject has an average of four or five REM episodes a night that recur cyclically. These episodes, on the whole, comprise approximately 20–25% of an adult's sleep.

In the words of William Dement, "the discovery of REM sleep was the breakthrough, the discovery that changed the course of sleep research." As a matter of fact, the emphasis attributed to the REM stage has been such that the other four stages of sleep are often defined simply as non-REM sleep. As discussed later, this extreme simplification, including the exclusive attribution of dreaming to REM sleep, was later criticized in various ways.

III. PSYCHOPHYSIOLOGICAL FEATURES OF DREAMING

A. Eye Movements and Dream Imagery

The discovery of characteristic bursts of eye movements in sleep led to the hypothesis of a direct relationship with the oneiric imagery, as some anecdotal reports seem to suggest. For example, one subject—who was awakened after a period 15 min of eye inactivity followed by some large shifts from right to left—reported having dreamed of driving a car while staring at the road ahead of him until, coming to a crossroads, he was struck by the sudden appearance of another car approaching quickly from the left. Some empirical studies, albeit not completely exemplary on a methodological level, seemed to support the hypothesis of a close connection between eye movements and dream images. According to some researchers, rapid eye movements are indeed the result of the actual scanning of the dreamed "scene."

According to other researchers, however, eye movements represent random, inevitable bursts of motor activity. Eye movements could merely be one of a variety of physiological signs of the peculiar pattern of activation processes of the organism that coexist with dreaming. If a correlation between REMs and dream content were to be verified, it would still be unclear whether chaotically generated eye movements forced appropriate dream content or dream content forced appropriate eye movements; correlation is not causation. The study of subjects blind since birth soon appeared to be the most logical approach to test the "scanning hypothesis."

B. Dreaming in the Blind

Blind people do have dreams that include visual imagery, but only if blindness occurs later than 6 or 7 years of age. It has been known, however, that dreams in congenitally blind people are nonvisual, and that their content is mostly linked to other sensory modalities. If the scanning hypothesis were true, then one might argue that subjects who have never had the opportunity to gaze upon their surroundings should not show rapid eye movements while dreaming. Sleep laboratory studies, however, have demonstrated quite the opposite (i.e., the presence of eye movements in the REM sleep of congenitally blind subjects). An even greater difficulty in accepting the hypothesis that eye movements follow dream scenes arises when we consider the dreaming of newborn babies or even animals deprived of the cerebral cortex.

As discussed later, newborn infants (and even prematurely born infants) show the presence of all those physiological features that accompany the subjective phenomenon of dreaming in adults,

including eye movements. Furthermore, a persistence of eye movement bursts, although more disorganized, has been found even in decorticated cats and humans. It would certainly be difficult in these cases to postulate their relation to any kind of imaginative activity.

A different interpretation of the phenomenon, which takes into account the apparently contrasting findings, may be considered within a developmental view. Eye movement may be considered as a physiological mechanism included since birth in the constellation of functional characteristics of the REM stage. The overall and experiential maturation of the organism involves a change in these automatic movement bursts along the lines established by the interaction with psychic phenomena as they occur in dreaming processes. After all, the relationships between eye movement and dream imagery may be another example of various physiological functions controlled by lower brain centers that are coordinated in the area of higher brain processes as they mature.

C. Dream Recall

Dreaming in the course of a night takes up a considerably longer amount of time than was previously supposed. Each one of us has 1 or 2 hr per night of REM sleep; however, even those who remember dreams best generally recall only one or at most two fragments per night, and most people remember much less. Even the few dreams that are well remembered after awakening normally fade from memory after a few minutes or a few hours unless special effort is made to keep them in mind. These considerations give rise the following questions: Is it really true that everyone dreams in the proportions reported by the studies carried out in the sleep laboratories? Why are dreams so much more difficult to remember with respect to wake experiences?

An answer to the first question was provided by a study conducted in 1959 at the Downstate Medical Center in New York. Donald Goodenough and coworkers used the EEG technique to record the pattern of brain activity during sleep in two groups of subjects. The first group was composed of people claiming to dream at least once a night, whereas subjects in the second group claimed they dreamed less than once a month or even not at all. These researchers showed that (i) REM periods occurred with the same frequency in both groups; (ii) all subjects, even those who claimed they had never dreamed before, were able to report at least one dream during the nights spent in the laboratory when awakened during REM sleep; (iii) the so-called "dreamers" reported a significantly higher number of dreams than did "non-dreamers". These findings, widely confirmed in later studies, led to the conclusion that, albeit with considerable individual differences, everyone dreams. The distinction between dreamers and nondreamers should therefore be reformulated as those who remember dreams and those who do not.

This brings to the forefront the question of why there is widespread forgetting of nocturnal dreaming experiences. Even "good dreamers" do not normally remember as many dreams as they do if awakened toward the end of a REM phase. Thus, even taking into account individual differences, the basis for this forgetting should be searched for in some ubiquitous physiological factors. The reasons for the failure of transferring sleep mentation from short- to long-term memory are still not completely clear, and the most popular hypotheses present differences that depend on the stress placed on biological vs psychological factors.

One repeatedly put forward hypothesis, which has little credibility today, maintains that the periods of non-REM sleep that follow REM are responsible for forgetting, as if non-REM sleep contained conditions that were somehow incompatible with the long-term consolidation of dreams. Another biologically oriented theory, which finds greater support, considers the particular biochemical substratum underlying the REM phase as responsible for forgetting. According to Allan Hobson and Robert McCarley of Harvard University, dream amnesia depends on the cutoff during REM sleep of a special class of neurotransmitters (namely, the monoamines noradrenaline and serotonin) that are needed to convert our immediate- or short-term memories into long-term storage. According to this hypothesis, when we awaken from a dream the noradrenergic and serotonergic neurons, located in many subcortical centers and widely projecting to the cerebral cortex, turn on and give our brain a shot of these transmitters. If a dream experience is still encoded in activated networks of neurons, it can be reported, recorded, and remembered.

Researchers oriented more toward analyzing the psychological factors tend to minimize or even to deny a substantial subcortical influence on the production and forgetting of dreams. While rejecting Freud's original censorship theory, most dream scientists in the modern cognitive research area would probably accept the general idea of a relationship between the specific

process of dream construction and the easiness of dream forgetting. David Foulkes, for instance, believes that it is the very nature of the dream production process (i.e., the limited presence of the self and the quality of the encoding competence) to make remembering difficult: "When we are dreaming we are not deliberately selecting and organizing the contents of our conscious thoughts and we are not able to reflect on them in a self-conscious way."

Allan Rechtschaffen, another outstanding researcher in the field, has described this peculiarity of our dreaming state as "single-mindedness." While dreaming, there are no alternative lines of thought as in wake, when we are self-reflective and contextually aware; in a sense, our consciousness, functions as a "single channel." A connection can be seen between the tendency to forget and the absence of simultaneous channels to monitor the oneiric mentation and to put it in the more general context of our previous knowledge.

A satisfactory clarification of the reasons why dreams are so easily forgettable will have to wait for a deeper understanding of both the psychological and the biological correlates of sleep and, above all, of their dynamic integrative constraints and interactions.

IV. DEVELOPMENTAL ASPECTS OF DREAMING

A. REM Sleep at the Beginning of Life

The discovery of a sleep stage accompanied by dreaming and characterized by distinct physiological correlates had scientific implications of undoubted significance. It inaugurated investigation of the neurophysiological substratum and functional significance of REM sleep as well its ontogenetic and phylogenetic features.

The developmental aspects of sleep could certainly not be ignored and very soon the sleep and dreaming laboratories started to study newborns and even premature infants. Sleep in these subjects was monitored using the classical polygraphic recording techniques. With some surprise, it was found that the REM phenomenon is clearly present from the beginning of life. However, at variance with what is observed in older children and in adults, REM sleep in the neonate is characterized by the presence of a variety of body movements. In particular, a surprising range of facial expressions can be observed that are not present in the other sleep and wake states. In fact, in the REM state almost all the typical motor components of adult emotional face expressions have been observed since the first days of life. Because of the characteristic presence of facial and body motility, the immature REM sleep of the infant is generally called "active sleep," whereas the rest is known as "quiet sleep." Another characteristic difference from the adult is that the REM phase in the newborn is usually the first stage of sleep. However, perhaps the most relevant difference from adult REM is a quantitative one: whereas adults spend about 20–25% of their total sleep time in the REM stage, and this is reduced to about 15% in people older than 60 years of age, in newborn babies this percentage is about 50%, and it is even higher in prematurity. In various recording experiments, it was found that the proportion of REM to total sleep time reaches 70% in 7-month premature babies. These values are even more impressive if one considers the absolute amount of sleep time of children in the months following birth with respect to that of adults.

B. Do Infants Dream?

The physiological signs that accompany the dreaming experience in the adult are also present in the early months after birth; indeed, they are present in higher proportions. The question of whether infants dream, however, still cannot be answered directly. With regard to subjective characteristics and content, we can make inferences on the basis of indirect evidence whose interpretations, however, vary in relation to the theoretical assumptions and the definition of dreaming adopted by various researchers.

According to David Foulkes, a dream is not a mere analog and pictorial representation of the outside world as perceived by the subject but, rather, a product of psycholinguistic processes that are largely similar to those used in the normal waking state. The peculiar quality of the dream mentation stems from the fact that the cognitive organizing processes are applied to uninhibited memory stores, rather than to the events of the outside world and that such processes occur in the particular situation of loss of voluntary control and of self-regulation induced by sleep, particulary in the REM stage. Thus, dream would be the product of a complex symbolic activity that is not possessed at birth but acquired, in parallel with the development of cognitive–linguistic development. Supported by a longitudinal study conducted in the sleep laboratory, Foulkes claims that "young children may fail to report

dreams because they are not having them, rather than because they have forgotten them or are unable to verbalize their contents.''

This view is considerably influenced by cognitive approaches oriented more toward studying the mind as an organ of pure cognition. Other researchers place greater emphasis on ''sentient'' aspects of the mind, and in any case believe it more parsimonious to postulate the ability of dreaming also in the first stages of life. Following this line of reasoning, the researcher is prompted to consider different levels of development of dreaming ''skills'' and to accept the idea that an infant can experience emotions and fictive visual and kinesthetic sensations. At any rate, the use of narrative as the largely prevalent method to report experiences that may have multimodal origins, though understandable, poses a strong limitation to the interpretation of the data available in the literature.

V. DREAMING IN ANIMALS

Is the REM stage something peculiar to human beings or does it exist in other animal species? After the first successful study carried out in the 1950s by William Dement on cats, research on animals spread rapidly; studies have been carried out on dogs, rats, rabbits, sheep, goats, monkeys, donkeys, chimpanzees, elephants, and even on the opossum (one of the most primitive mammals). Practically all the studied mammals have shown the presence of the REM stage, and its cyclical occurrence is now considered a fundamental feature of sleep in mammals. Apart from mammals, studies seem to indicate the existence of a REM stage, at least in a rudimentary form, in birds but not in reptiles. However, it has been found that at least one reptile, the chameleon, has periods of rapid eye movements when asleep. Phylogenetic considerations therefore lead us to hypothesize that the REM stage appeared for the first time on our planet more than 150 million years ago.

Since it is certain that many animals have recurring periods of REM sleep, is it also reasonable to assume that they dream in the sense that humans do? Anyone with a pet dog, cat, or other animal would swear that their pets dream, judging by the muscle twitches of their limbs, head, and ears that occur ocassionally when asleep. Even Lucretius, in the second century BC, having observed this kind of muscle activity in horses, suggested the possibility that it was linked to their dreaming. According to some, from the developmental evidence in children we must assume that animals do not dream; according to others, this competence cannot be excluded.

Bearing in mind that, in any case, there is no direct evidence for what takes place in animals' minds, the problem cannot merely imply a straight yes or no answer; even at the purely speculative level, hypotheses for and against dreaming in animals must first specify what is meant by dreaming within the context of the animal species under consideration.

VI. NEUROPHYSIOLOGICAL AND NEUROCHEMICAL MECHANISMS OF DREAMING

The discovery of a REM stage in animals opened up new possibilities of experimental research obviously unfeasible in humans. In the early 1960s, transection studies in cats conducted by the French sleep scientist Michel Jouvet and coworkers pointed to structures in the brain stem as the source of REM sleep. More precisely, these researchers discovered the presence of both a trigger mechanism and a REM sleep clock in the pons. Investigating the pontine reticular formation through the recording microelectrode technique, McCarley and Hobson reported a series of findings that led to the formulation of a ''neurochemical'' model of the rhythmic appearance of REM sleep. They called this hypothesis the reciprocal interaction model because of its central concept: During REM sleep, aminergic subsets of neurons (the so-called REM-off cells) were inhibited, while cholinergic (REM-on) cells actively discharged. The opposite pattern of neuronal activity was observed during wake.

The relevance of the pontine brain stem in REM sleep generation has been recently confirmed using neuroimaging techniques. Studies based on positron emission tomography (PET), for example, have shown activation of the pontine tegmentum during REM sleep. Significant activations, however, were also seen in many other brain areas, including limbic and paralimbic forebrain regions that are thought to mediate emotion, an important aspect of dreaming. Finally, PET studies have found, during REM sleep, an increase in the activation of unimodal associative visual (Brodmann areas 19 and 37) and auditory (Brodmann area 22) cortices, whereas heteromodal association areas in the frontal and parietal cortex were deactivated. The frontal deactivation could be responsible for another important aspect of dreaming—the bizarreness of dream content. Despite the

general deactivation in much of the parietal cortex, an activation of the right parietal operculum, a neural structure that is important for spatial imagery construction, has been reported.

Although in REM sleep the global cerebral metabolism tends to be equal to or greater than that of waking, on the whole these findings suggest that the regional activation during dream mentation is very different from that of waking mentation. In particular, the reduced involvement of the cortical areas devoted to processing sensory information in the waking state, as reported in recent imaging studies, appears to further confirm what was already known—that is, a block against external sensory inputs during REM state and the resorting to systems connected to the internal sphere of emotions and memories.

Having recognized that REM sleep originates in neural structures located in the hindbrain, and not in the forebrain, the following was next question to address: To what extent do these centers control dreaming as well? Evidence in this respect could in theory be gained by studying brain-damaged patients. In fact, if dreaming is caused by neural activity in the brain stem, then brain stem lesions should eliminate both REM and dreaming. The available evidence in this respect is not conclusive, mostly because brain stem lesions that are extensive enough to prevent REM render the patients unconscious. Researchers who do not believe in the causal link between the brain stem and dreaming find indirect support to their claim from evidence in the literature that shows that dreaming is eliminated by forebrain lesions that completely spare the pontine brain stem. However, arguments of this sort rely on an unjustified assumption: They equate the process of dream generation to the process of dream recall and report. Cessation or reduction of dream recall after brain damage may indicate an impairment in any of the processes allowing the production of the dream experience, its recall, and its report and should not be attributed exclusively to the process underlying its production.

The question therefore remains largely unanswered despite major effort. Trying to trace the origin of dreaming to one specific brain area or another may represent too simplistic an approach that does not take into account the complex relationships between higher and lower brain centers. A highly rigid separation between the different sleep and wake phases and between rational thought and the oneiric experience is another source of bias that may have a negative impact on the progress of research. The complex psychophysiological nature of dreaming is increasingly evident, to the point that it has been questioned whether dreaming is a mental experience that can only be linked to sleep.

VII. DOES DREAMING BELONG TO SLEEP?

At the beginning of the 1960s, the interest evoked by the discovery of the REM stage was such that a proposal for calling it ''the third state of consciousness'' was often advanced. Instead of ''wake'' and ''sleep,'' it was popular at that time to speak of ''wake,'' ''dreaming,'' and ''nondreaming'' stages. Meanwhile, the occasional reports of dreaming outside of REM sleep were attributed either to recall of REM material or to the subject's waking confabulations.

Later, however, on the basis of consistent empirical findings of dream-like mentation in non-REM sleep, the notion that thought modalities are completely segregated within the different physiological states was seriously questioned. It was shown that non-REM reports could not be safely ascribed to recall of mental events taking place during REM. A considerable amount of mental activity often indistinguishable from REM dreams was found even at sleep onset (i.e., *before* the first occurrence of a REM phase). The debate between those favoring and those rejecting the hypothesis that dreaming is an exclusive production of REM is ongoing.

It seems clear that some form of mental activity is present in non-REM sleep, although in the past, there was substantial agreement on the fact that REM reports are more frequent, longer, bizarre, and more visually animated and emotional compared to non-REM reports, which appear substantially more mundane and thought-like. Later, however, some researchers challenged this notion, claiming that the special quality of REM imagery largely depends on the greater length of reports due to the stronger and more widespread brain activation present in the REM state; when data are statistically checked for report length, qualitative differences would tend to diminish and often disappear. These findings support the so-called ''single dreaming generator'' hypothesis—that all sleep mentation derives from a common imagery source that is driven by different levels of brain activation.

Those who oppose such a hypothesis maintain that qualitative differences are not the result of longer report length. On the contrary, report length would depend on the characteristic features of dreaming

experiences, which require a greater number of words to explain them, as demonstrated especially by Hunt. If we consider the relations between REM and non-REM stage phenomenology and the neurochemical substrates underlying them, we can see how the debate is not a purely academic one but is quite important in the context of a general psychophysiological theory of dreaming.

On the whole, there is substantial agreement on the fact that when a subject is awakened during the REM stage of sleep, chances of obtaining a more precise report on dream-like mentation are highest. This appears to be the main stage for dreaming activity, but not the exclusive one. Indeed, nobody today would claim that dreaming is absolutely unique to REM sleep. Dream-like mentation, in fact, can be reported in every phase of sleep.

Dream-like mentation has been demonstrated even in certain waking situations. Several years ago, a series of experiments conducted at McGill University in Canada showed that subjects placed in conditions of isolation and sensory deprivation begin, after a certain time, to experience quite vivid pseudohallucinatory-type imagery. An abundance of dream-like mentation was elicited through a so-called "reverie technique" derived from the McGill studies and developed by Mario Bertini, Helen Lewis, and Herman Witkin at the Downstate Medical Center in New York. When put in a particularly monotonous and perceptually destructured environment, awake subjects often experience scattered images and even articulated scenes with a clear dream-like character. Such images, often as complex and bizarre as dreams, seem to appear spontaneously to the subject, sometimes beyond any voluntary effort or control.

Furthermore, this type of imagery can be experienced and reported during wakefulness even without using techniques aimed at facilitating its onset. David Foulkes had subjects lie in laboratory beds during the day and at regular intervals asked them to describe their experience. Most subjects seemed to forget their surroundings and reported vivid and bizarre episodes described as "hallucinatory daydreams" that were so compelling that they were briefly experienced as real. These and other empirical observations justify a growing interest in cutting across the traditional divisions between sleep and waking states and devoting more attention to understanding the functional unifying principles, and organizational coherence, of similar experiences in different states and, specifically, of dream-like experiences through REM, non-REM, and wake states.

VIII. THE FUNCTIONS AND THE MEANING OF DREAMING

Dreaming researchers are usually asked the following fundamental questions: What is the purpose of dreaming? Do dreams have any meaning?

A. The Functions of Dreaming

In answering the first question, today's dream researchers can take advantage of a wealth of precious information available on the physiological characteristics of the stage in which dreaming mainly takes place. This is important information, if only for the opportunity offered from a methodological standpoint, for anyone studying dreaming in a laboratory and from the perspective of one's own specific level of analysis. A problem arises when considering the usefulness of findings at a biological level for understanding the nature and function of dreaming at the mental level. If, as it obviously seems, an accurate description of the psychological sphere of dreaming cannot explain the underlying REM physiology, then the opposite must also be considered true: Finding a physiological basis for dreaming is not the same as explaining it.

While greatly respecting the separate levels of analysis and avoiding any dangerous reductionism toward one side or the other, I believe that a specific analysis of the interface between levels may be useful both for a flexible reformulating of the respective theories and for proposing a general theory of dreaming.

1. Dreaming and Memory Systems

The consistent presence in most dream reports of past memories intermixed with what Freud named "day residues" prompts the commonsense idea that dreaming plays a role in the organization and consolidation of what is experienced during waking.

The hypothesis that memories are consolidated during sleep is a long-standing one. In particular, the diffuse brain activation that characterizes REM sleep could provide the appropriate context for the modifications of synaptic strength within the circuits activated during the wake experience. Such changes are thought to represent the neural basis of memory consolidation.

The evidence in favor of this hypothesis, based mostly on animal experimental models, has been methodologically criticized and is countered by some negative findings. In particular, no damage to memory and learning has occurred after the use of major antidepressant drugs that cause a marked suppression of REM sleep. However, functional brain imaging studies have recently shown that the waking experience influences regional brain activity during subsequent sleep. Specifically, several brain areas that are activated during the execution of a task during wakefulness were significantly more active during REM sleep in subjects previously trained on the task than in nontrained subjects. These results support the hypothesis that memory traces are processed during REM sleep in humans.

Thus, the relation between memory and REM is probably more complex than indicated by the idea of a consolidation function. Special attention to REM state ontogenetic development may offer a first guideline to a better understanding of dreaming function.

2. The Relevance of REM Sleep in the Beginning of Life

As already mentioned, the quantity of REM is particularly high in the early stages of life, in which it is accompanied by the appearance of new behaviors (e.g., facial emotional expressions such as smiling, anger, or fear). The fact that this pattern manifests itself at a time of maximum development of the learning processes leads to a reflection on the role of REM sleep in the structural development of behaviors that are essential for survival.

Howard Roffwarg, Joseph Muzio, and William Dement first proposed that neonate REM sleep provides endogenous nervous system stimulation, which facilitates the maturation and differentiation of sensory and motor brain areas in the absence of external stimuli. Although not above criticism, this view has found credit with many scholars who, albeit with some variations, consider REM sleep to be functional to the general processing and adaptive reprocessing of learning and memory systems.

According to Jouvet, REM sleep would serve the maturation of species-specific basic behavioral patterns at the beginning of life, when there is the highest need for brain programming, as well as the modification and continuous revision of the programs after a certain age, when plasticity in interacting with environment is increasingly required. The shift from the substantial amount of REM sleep in the prenatal and early neonatal period to a sharp downward trend, observed between the second and third postnatal months, should constitute a critical period in the development of interaction between the infant and the environment.

The exchange between what is genetically preordained and rooted in the species' history and what is epigenetic and belonging to the subject's recent and distant past appears to be a suggestive hypothesis for the function of dreaming.

3. REM as a Paradoxical Event

It is possible to deeper into this field of complexity by reflecting on the physiologically peculiar characteristics of the REM stage that have suggested the adjective "paradoxical" for this phase (in fact, this phase appears to be a specific psychophysiological condition in which elevated EEG activity paradoxically coexists with deep muscle atonia and high threshold awakening).

If we monitor an adult human subject's nocturnal sleep by using polygraph recordings, we will find the following sequence of events: Before falling asleep, the EEG signal is mostly composed of low-amplitude, high-frequency waves that are typical of a state of alert activity of the cerebral cortex. Upon sleep onset the signal tends to become increasingly more ample and slower in frequency in direct relation to sleep depth and, therefore, to reduced vigilance. However, with the onset of the REM stage (after approximately 90–100 min) and for the whole duration of it, the EEG trace looks flatter and faster again, with characteristics similar to those of the waking state, even though the subject remains in a state of complete nonresponsiveness.

The paradoxical aspect of the transition from non-REM to REM sleep is that, in parallel with EEG signs of a generalized cortical reactivation, a further "detachment" from the outside world can be clearly observed. A mechanism of presynaptic inhibition originating in the brain stem reduces the input of sensory information from the environment (hence, the heightened awaking thresholds). Furthermore, with the exception of the oculomotor and middle ear muscles, a descending mechanism of postsynaptic inhibition causes a generalized tonic paralysis. Thus, during REM sleep the human organism appears disconnected from the outside world, but brain activity is quite intense, especially in certain regions.

Some researchers view paradoxical sleep and wakefulness as almost identical intrinsic functional states,

both able to generate subjective awareness. The difference is in the orientation of the attention. According to Rodolfo Llinas, REM sleep can be considered as a modified attentive state in which attention is turned away from the sensory input and toward memories. The idea of an attentive state of REM sleep receives further support from the similarities between some physiological features of this phase and the classic physiological counterparts of the wake Pavlovian orientation reflex.

Thus, the brain in REM sleep appears to shift to an off-line status where inhibition of the sensory and motor systems corresponds to an activation of limbic and paralimbic regions that neuroimaging studies have shown to be connected with memories and emotions. The subjective experience of dream imagery characteristic of this phase could be the result of such a shift.

The idea of a duality between executive control (i.e., the waking state) oriented toward the outside and a state that favors internal reprocessing (i.e., the sleep state) has prompted several researchers to use the metaphor of the computer, which at times is in the state of executing programs and at other times undergoes reprogramming. To a certain extent, the same concept can be found, in a less prosaic style, in Heraclitus' thinking: ''He who is awake lives in a single world common to all; he who is asleep retires to his own particular world.''

Some experiments in animals may provide further insight into the processes that take place in the REM phase. Michael Jouvet and subsequently Adrian Morrison obtained some interesting results when certain pontine nuclei known to be responsible for motor paralysis during REM sleep were lesioned in cats. Each time the animal entered REM sleep, it showed an astonishing pattern of highly organized but entirely automatic motor behaviors, such as orienting and exploring its territory, using the cat litter, licking itself, attacking an imaginary prey, and having a fit of rage. The animal seemed to act out basic instinctive and highly species-specific behavioral patterns. These experimental observations reinforced the previously mentioned hypothesis of a role of REM sleep in programming such patterns.

On the whole, psychophysiological data on the ontogenetic development and phenomenological characteristics of REM sleep have opened interpretational scenarios centered on the idea of a privileged function of the REM phase with respect to the adaptive organizing and reorganizing of memory systems. This function does not necessarily imply the subjective experience of dreaming; in principle, it is not necessary for the dream to be recalled in order to accomplish it.

B. Does Dreaming Have Any Meaning?

Everything that has been said so far about the REM stage (i.e., the long philogenetic history and the presence of a physiological mechanism guaranteeing its regular ultradian cycle) amply justifies the recognition of the important biological value attributed to a sleep stage that is also called dreaming stage because of its strong links with dream mentation. Moreover, the hypothesis that this biological function may be linked to processes that are crucial for the development and adaptive functioning of the mind is a reasonable one, even though not all researchers agree with it.

Even the most compelling evidence in favor of a primary role for the REM stage in mammalian physiology would not in itself justify the attribution of meaning to the subjective experience of dreaming. In principle, dreaming could be a mere by-product of a fundamental but still largely unknown biological function.

There is no doubt that all cultures throughout history have posited the meaningfulness of dreams. Modern neurophysiology as well as cognitive neuroscience have focused more on the dreaming process than on its content. In the most entrenched positions of these approaches, a question such as ''Is it possible to recognize the presence of a message or the intention to communicate in dreaming?'' would normally receive a negative answer.

The activation–synthesis theory supported by Hobson and McCarley, for example, maintains that dreaming is the result of a ''bad job'' carried out by the cerebral cortex when, intensely activated by random stimulation coming from the brain stem, it tries to provide a plausible narrative context for the stimulation. For more cognitiveley oriented scholars such as Foulkes, the precondition of dreaming is the relinquishment of the voluntary self-regulation and the consequent diffuse activation of mnemonic systems; the organization imposed by a dream production system on the mnemonic units is syntactic in nature, not semantic.

According to both views, dreaming is devoid of any meaningful intention. At the formal level of dream mentation, however, the interpretations differ. Hobson and MacCarley, more anchored to the biological side, believe that the formal aspects of dreams are

isomorphic expressions of the peculiar physiological state of REM sleep. According to Foulkes dream mentation is substantially independent from the REM physiology and under the direct control of the same narrative-linguistic generator that controls cognitive processes during wakefulness. In both cases, we are evidently a long way from the Freudian view, whereby the manifest content of dreams conveys a censored deeper latent message to be decoded that strongly points to the dreamer's unconscious problems.

It must be stressed, however, that even in modern neuroscientific literature, there is wide agreement that dreams do convey some sort of meaning and reveal something about the dreamer. The information may be of two types. The first kind concerns the peculiar characteristics of the processes with which the dreamer constructs the dream. This aspect is particularly meaningful for cognitive sciences. Authoritative scholars hold that, applying the same cognitive–linguistic methods of analysis used in waking or stimulus-dependent narrations, the study of self-organizational mechanisms of dreaming may yield information of great value not only concerning dreaming but also on all those forms of stimulus-independent thought occurring spontaneously over the 24-h period. In this sense, although it is true that the application of the cognitive approaches is certainly useful for the study of dreaming, we should not overlook the importance of studying the mind's involuntary organizing system, in REM as well as in other sleep and wake conditions, in order to broaden our knowledge of the overall mental functioning.

A second, generally accepted notion concerns dreams as sources of information on the dreamer's personality. For example, the fact that some people may experience more frequent dreams of certain expressive or conflictual contents rather than others may, with suitable care and attention to context, provide indirect indications of the dreamer's personality. In this view, which implies the possibility of diverse clinical applications, the dream is "informative" and not "intentional" in nature.

Other perspectives that are worth considering view dream not as "random nonsense" but as a different form of intelligence with respect to that expressed in the rational, logical forms more typical of wakefulness (i.e., not alternative but complementary to it). In line with certain anticipatory views expressed by Jung, Harry Hunt summarizes this concept as follows:

A host of contemporary "dream-workers" holds that the emergent sources of dreaming, at least much of the time, are not "primitive" or "disruptive" at all. Rather, they constitute the spontaneous expressions of a symbolic intelligence alternative to standard representational thought and variously termed "imagistic," "affective," or "presentational". … I would not want … to imply an adherence to rigid dual code models of human intelligence, which tend to distinguish an abstract verbal capacity from a more rudimentary imagistic one. Not only is the latter fully capable of its own line of abstract development, as seen in certain dreams, alterations of consciousness, and aesthetics, but each of these "frames" seems capable of endlessly, complex interactions at all levels of their potential unfolding.

IX. CONCLUDING REMARKS

Studies carried out during the past 50 years, particularly those using recent neuroimaging techniques, show that the areas of the central nervous system that seem more active in REM sleep are found in the brain stem, in the limbic system, and in secondary, associative cortical areas, at variance with the alert waking state, during which the frontal lobes and the primary sensory and motor cortices are privileged. Thus, the stage of sleep in which dreaming has its greatest expression appears to be characterized by a prevalent orientation toward the inner world of memories and emotions. In line with this view, the adjective "paradoxical," as applied to the REM stage, may sound improper. Its physiological characteristics, in fact, appear perfectly orthodox if considered as an expression of a positive and qualitative shifting toward a greater resonance of the inner world—a shifting, however, that although paradigmatic for the REM state, may be available throughout the sleep and wake cycle.

As discussed earlier, a variety of findings encourage stepping beyond the Manichean barriers separating the states of wakefulness, REM sleep, and non-REM sleep in order to appreciate the continuity of the mind across these states. For instance, although it may be true that mentation during REM is largely bizarre, systematic demonstrations indicate that REM dream content also shows a heavy reliance on representations of familiar elements from the waking environment. By the same token, if it is true that during the awake state resources are usually employed in relating to the external environment, there are still opportunities for coding operations and processing of internal material. Consider, for instance, the fantasy activity conceived

as ongoing information processing. As Jerome Singer states, "the popular notion of a daydream represents only one manifestation of a more general class of phenomena perhaps best described as internally generated diversion from an ongoing motor or cognitive task." There is evidence that some type of coding and restructuring of long-term stored material goes on a great deal of the time while the organism is awake.

The critical point that needs further examination is the psychophysiological significance of this shifting. The first question is whether accessing and activating the previously mentioned brain areas in the REM stage—besides serving a possible, although not yet known, biological function—involves any direct functional value for the mind. Second, the problem arises of the semantic meaning of dream and dream-like mentation; as stated earlier, dreaming activity could in principle produce some objective organizational benefit to the mind, independently from the subjective experiencing of dreaming and from the content of the experienced dream, and that organizational benefit may be augmented by dream recall and interpretation.

An overview of the different positions on these issues shows that scholars in the broader field of dreaming can be divided into approximately two groups. On the one hand are those who, although not rejecting the adaptive hypothesis, from various angles and with different reasoning tend to characterize dream mentation as a degradation of the self-conscious, self-controlled thought of wakefulness. For example, sleep physiologist Giuseppe Moruzzi, while maintaining that sleep has a function coherent with the need to restore the myriad of cerebral neural networks, also believed dreaming to be the expression of an "impoverished conscience." Taking into account the intense activity of the brain in REM sleep, he stated that "the plastic debt accumulated during wakefulness could be paid while the cortical neurons continued to discharge impulses, even if in a different way than usual." Loss of consciousness could thus be the result not of a silence but of the temporary loss of a meaningful dialogue between nerve cells.

On the other hand, there are scholars who view REM sleep as a privileged theater for the type of plastic revision on which the adaptive flexibility of our central nervous system is based. In their view, accessing memory stores may set the potential for creating new associations, rearranging past patterns, and reformulating future plans in accord with experiences of the day. Rather than "temporary loss of a meaningful dialogue," dreaming represents the "temporary emergence of a differently meaningful dialogue." In such a perspective, at least some dreaming would be the expression of an alternative form of "intelligence," common to a variety of thinking modalities leaning toward the imagery and metaphoric side of the human vast symbolic capacity. A "symbolic–presentational" form of intelligence (Susanne Langer's terminology), although semantic and communicative, resists a complete narrative formulation when compared to the "symbolic–representational" intelligence that prevails in the vigilant waking state.

In 1890, William James stated: "Our normal waking consciousness, rational consciousness as we call it, is but one special type of consciousness, whilst all about it, parted from it by the filmiest of screens, there lie potential forms of consciousness entirely different." Thus, even though the age-old dilemma of dreaming has been engaged by modern science, we still remain within the interpretational dualism that has characterized the history of thought on dreaming. The work carried out during the past 50 years, however, seems to justify a certain optimism with respect to a converging of views in the near future.

See Also the Following Articles

COGNITIVE PSYCHOLOGY, OVERVIEW • CONSCIOUSNESS • CREATIVITY • DEMENTIA • HALLUCINATIONS • NEUROIMAGING • SLEEP DISORDERS • UNCONSCIOUS, THE

Suggested Reading

Antrobus, J., and Bertini, M. (Eds.) (1992). *The Neuropsychology of Sleep and Dreaming.* Erlbaum, New York.

Aserinsky, E., and Kleitman, N. (1953). Regularly occurring periods of ocular motility and concomitant phenomena during sleep. *Science* **118,** 1150–1155.

Dement, W., and Kleitman, N. (1957). The relation of eye movements during sleep to dream activity: An objective method for the study of dreaming. *J. Exp. Psychol.* **55,** 339–346.

Foulkes, D. (1985). *Dreaming: a Cognitive–Psychological Analysis.* Erlbaum, New York.

Hobson, A. (1988). *The Dreaming Brain.* Basic Books, New York.

Hunt, H. (1989). *The multiplicity of dreams.* Yale Univ. Press, New Haven, CT.

Sleep and dreaming [Special issue] (2000). *Behav. Brain Sci.* **23**(6).

Dyslexia

H. BRANCH COSLETT
University of Pennsylvania School of Medicine

GLOSSARY

central dyslexia A reading disorder caused by impairment of the "deeper" or "higher" reading functions by which letter strings mediate access to meaning or speech production mechanisms.

deep dyslexia A disorder characterized by semantic errors (e.g., "king" read as "prince"), poor reading of abstract compared to concrete words (e.g., "idea" vs "desk"), and an inability to "sound out" words.

peripheral dyslexia A reading disorder attributable to a disruption of the processing of visual aspects of the stimulus that prevents the patient from reliably processing a letter string.

phonologic dyslexia A disorder characterized by a largely preserved ability to read familiar words but an inability to read unfamiliar words or nonword letter strings (e.g., "flig") because of an inability to use a "sounding out" strategy.

surface dyslexia A disorder characterized by reliance on a "sounding out" strategy with the result that words that have atypical print-to-sound correspondences are read incorrectly (e.g., "yacht" read as "yatchet").

Unlike the ability to speak, which has presumably evolved over tens of thousands of years, the ability to read is a relatively recent development that is dependent on both the capacity to process complex visual stimuli and the ability to link the visual stimulus to phonologic, syntactic, and other language capacities. Perhaps as a consequence of the wide range of cognitive operations involved, reading is compromised in many patients with cerebral lesions, particularly those involving the left hemisphere. The resultant reading impairments, or acquired dyslexias, take many different forms, reflecting the breakdown of specific components of the reading process. In this article, I briefly review the history of the study of acquired dyslexia and introduce a model of the processes involved in normal reading. Specific syndromes of acquired dyslexia are discussed. I also briefly discuss connectionist accounts of reading and the anatomic basis of reading as revealed by recent functional imaging studies.

I. HISTORICAL OVERVIEW

Perhaps the most influential early contributions to the understanding of dyslexia were provided in the 1890s by Dejerine, who described two patients with quite different patterns of reading impairment. Dejerine's first patient manifested impaired reading and writing subsequent to an infarction involving the left parietal lobe. Dejerine termed this disorder "alexia with agraphia" and attributed the disturbance to a disruption of the "optical image of words" that he thought to be supported by the left angular gyrus. In an account that in some respects presages contemporary psychological accounts, Dejerine concluded that reading and writing required the activation of these optical images

and that the loss of the images resulted in the inability to recognize or write even familiar words.

Dejerine's second patient was quite different. This patient exhibited a right homonymous hemianopia and was unable to read aloud or for comprehension but could write. This disorder, designated "alexia without agraphia" (also known as agnosic alexia and pure alexia), was attributed by Dejerine to a "disconnection" between visual information presented to the right hemisphere and the left angular gyrus that he assumed to be critical for the recognition of words.

After the seminal contributions of Dejerine, the study of acquired dyslexia languished for decades, during which time the relatively few investigations that were reported focused primarily on the anatomic underpinnings of the disorders. The field was revitalized by the elegant and detailed analysis by Marshall and Newcombe that demonstrated that by virtue of a careful investigation of the pattern of reading deficits exhibited by dyslexic subjects, distinctly different and reproducible types of reading deficits could be elucidated. The insights provided by Marshall and Newcombe provided much of the basis for the "dual-route" model of word reading.

II. READING MECHANISMS AND THE CLASSIFICATION OF DYSLEXIAS

Reading is a complicated process that involves many different procedures, including low-level visual processing, accessing meaning and phonology, and motor aspects of speech production. Figure 1 provides a graphic depiction of the relationship between these procedures. This "information processing" model will serve as the basis for the discussion of the mechanisms involved in reading and the specific forms of acquired dyslexia. It must be noted, however, that the dual-route model of reading that will be employed for purposes of exposition is not uncontested. Alternatives to the information processing accounts, such as the triangle model of Plaut, Seidenberg, McClelland, and colleagues, will also be discussed.

Reading requires that the visual system efficiently process a complicated stimulus that, at least for alphabet-based languages, is composed of smaller meaningful units—letters. In part because the number of letters is small in relation to the number of words, there is often a considerable visual similarity between words (e.g., "structure" vs "stricture"). Additionally, the position of letters within the letter string is also critical to word identification (consider "mast" vs "mats"). In light of these factors, it is perhaps not surprising that reading places a substantial burden on the visual system and that disorders of visual processing or visual attention may substantially disrupt reading.

Normal readers recognize written words so rapidly and seemingly effortlessly that one might suspect that the word is identified as a unit, much as we identify an object from its visual form. At least for normal readers under standard conditions, this does not appear to be the case. Instead, analyses of normal reading suggest that word recognition requires that letters be identified as alphabetic symbols. Support for this claim comes from demonstrations that presenting words in an unfamiliar form—for example, by alternating the case of the letters (e.g., wOrD) or introducing spaces between words (e.g., f o o d)—does not substantially influence reading speed or accuracy. These data argue for a stage of letter identification in which the graphic form (whether printed or written) is transformed into a string of alphabetic characters (W-O-R-D), sometimes referred to as abstract letter identities.

The mechanism by which the position of letters within the stimulus is determined and maintained is not clear. Possibilities include associating the letter in position 1 to the letter in position 2, and so on; binding each letter to a frame that specifies letter position; or labeling each letter with its position in the word. Finally, it should be noted that in normal circumstances letters are not processed in a strictly serial fashion but letter strings are processed in parallel (provided they are not too long). Disorders of reading resulting from impairment in the processing of the visual stimulus or the failure of this visual information to contact stored knowledge appropriate to a letter string are designated as peripheral dyslexias.

In dual-route models of reading, the identity of a letter string may be determined by many distinct procedures. The first is a "lexical" procedure by which the letter string is identified by means of matching the letter string to an entry in a stored catalog of familiar words, or visual word form system. As indicated in Fig. 1 and discussed later, this procedure, which in some respects is similar to looking up a word in a dictionary, provides access to the meaning, phonologic form, and at least some of the syntactic properties of the word. Dual-route models of reading also assume that the letter string can be converted directly to a phonological form by means of the application of a set of learned correspondences between orthography and phonology. On this account, meaning may then be accessed from the phonologic form of the word.

Support for dual-route models of reading comes from a variety of sources. For present purposes, perhaps the most relevant evidence was provided by Marshall and Newcombe's groundbreaking description of "deep" and "surface" dyslexia. These investigators described a patient (GR) who read approximately 50% of concrete nouns (e.g., "table" and "doughnut") but was severely impaired in the reading of abstract nouns (e.g., "destiny" and "truth") and all other parts of speech. The most striking aspect of GR's performance, however, was his tendency to produce errors that appeared to be semantically related to the target word (e.g., "speak" read as "talk"). Marshall and Newcombe designated this disorder deep dyslexia. These investigators also described two patients whose primary deficit appeared to be an inability to reliably apply grapheme–phoneme correspondences. Thus, JC, for example, rarely applied the rule of "e" (which lengthens the preceding vowel in words such as "like") and experienced great difficulties in deriving the appropriate phonology for consonant clusters and vowel digraphs. The disorder characterized by impaired application of print-to-sound correspondences was termed surface dyslexia.

On the basis of these observations, Marshall and Newcombe argued that the meaning of written words could be accessed by two distinct procedures. The first was a direct procedure whereby familiar words activated the appropriate stored representation (or visual word form), which in turn activated meaning directly; reading in deep dyslexia, which was characterized by semantically based errors (of which the patient was often unaware), was assumed to involve this procedure. The second procedure was assumed to be a phonologically based process in which grapheme-to-phoneme or print-to-sound correspondences were employed to derive the appropriate phonology (or "sound out" the word); the reading of surface dyslexics was assumed to be mediated by this nonlexical procedure. Although many of Marshall and Newcombe's specific hypotheses have subsequently been criticized, their argument that reading may be mediated by two distinct procedures has received considerable empirical support.

The information processing model of reading depicted in Fig. 1 provides three distinct procedures for oral reading. Two of these procedures correspond to those described by Marshall and Newcombe. The first (Fig. 1A), involves the activation of a stored entry in the visual word form system and the subsequent access to semantic information and ultimately activation of the stored sound of the word at the level of the phonologic output lexicon. The second (Fig. 1B), involves the nonlexical grapheme-to-phoneme or print-to-sound translation process; this procedure does not entail access to any stored information about words but, rather, is assumed to be mediated by access to a catalog of correspondences stipulating the pronunciation of phonemes. Many information processing accounts of the language mechanisms subserving reading incorporate a third procedure. This mechanism (Fig. 1C), is lexically based in that it is assumed to involve the activation of the visual word form system and the phonologic output lexicon. The procedure differs from the lexical procedure described previously, however, in that there is no intervening activation of semantic information. This procedure has been called the direct reading mechanism or route. Support for the direct lexical mechanism comes from many sources, including observations that some subjects read aloud words that they do not appear to comprehend. For example, Coslett reported a patient whose inability to read aloud nonwords (e.g., "flig") suggested that she was unable to employ print-to-sound conversion procedures. Additionally, she exhibited semantic errors in writing and repetition and performed poorly on a synonymy judgment task with low imageability words, suggesting that semantic representations were imprecise. Semantic errors and imageability effects were not observed in reading; additionally, she read aloud low imageability words as well as many other words that she appeared to be unable to comprehend. Both observations suggest that her oral reading was not mediated by semantics. We argued that these data provided evidence for a reading mechanism that is independent of semantic representations and does not entail print-to-sound conversion—that is, a direct route from visual word forms to the phonological output representations.

III. PERIPHERAL DYSLEXIAS

A useful starting point in the discussion of acquired dyslexia is the distinction made by Shallice and Warrington between "peripheral" and "central" dyslexias. The former are conditions characterized by a deficit in the processing of visual aspects of the stimulus that prevents the patient from reliably matching a familiar word to its stored visual form or visual word form. In contrast, central dyslexias reflect impairment to the "deeper" or "higher" reading functions by which visual word forms mediate access to meaning to speech production mechanisms. In the

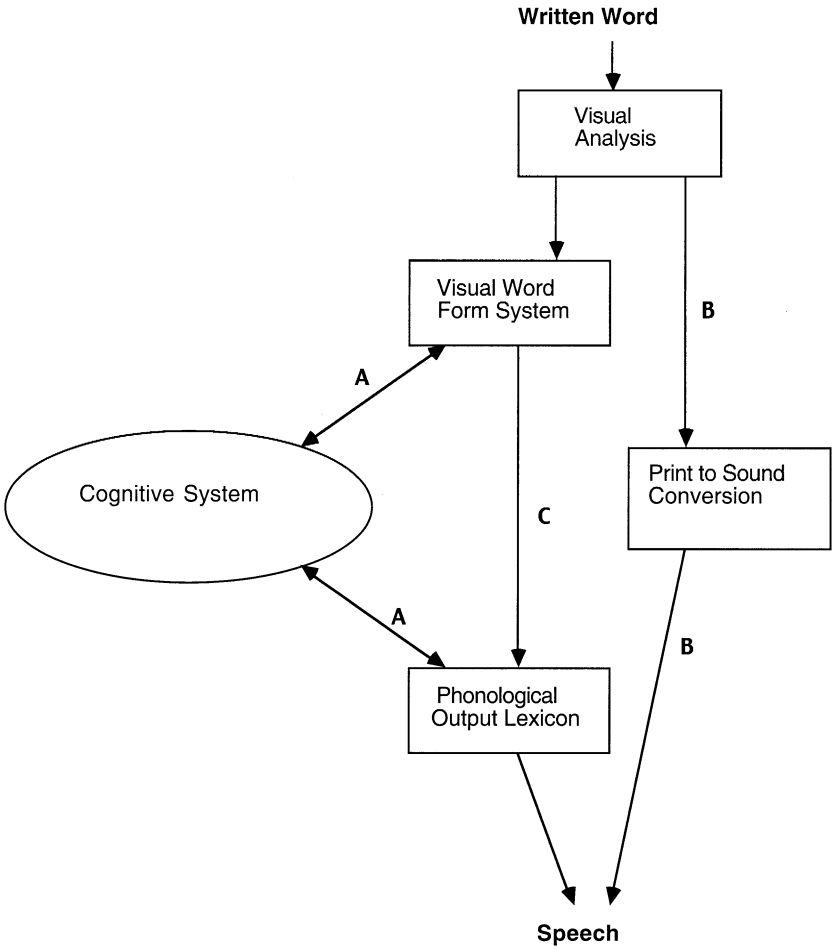


Figure 1 The relationship between the procedures involved in reading.

following sections, I discuss the major types of peripheral dyslexia.

A. Alexia without Agraphia (Pure Alexia; Letter-by-Letter Reading)

This disorder is among the most common of the peripheral reading disturbances. It is associated with a left hemisphere lesion affecting left occipital cortex (responsible for the analysis of visual stimuli on the right side of space) and/or the structures (left lateral geniculate nucleus of the thalamus and white matter, including callosal fibers from the intact right visual cortex) that provide input to this region of the brain. It is likely that the lesion either blocks direct visual input to the mechanisms that process printed words in the left hemisphere or disrupts the visual word form system. Some of these patients seem to be unable to read at all, whereas others do so slowly and laboriously by means of a process that involves serial letter identification (often termed "letter-by-letter" reading). At first, letter-by-letter readers often pronounce the letter names aloud; in some cases, they misidentify letters, usually on the basis of visual similarity, as in the case of N → M. Their reading is also abnormally slow.

It was long thought that patients with pure alexia were unable to read except letter by letter. There is now evidence that some of them do retain the ability to recognize letter strings, although this does not guarantee that they will be able to read aloud. Several different paradigms have demonstrated the preservation of word recognition. Some patients demonstrate a word superiority effect of superior letter recognition when the letter is part of a word (e.g., the "R" in "WORD") than when it occurs in a string of unrelated

letters (e.g., "WKRD"). Second, some of them have been able to perform lexical decision tasks (determining whether a letter string constitutes a real word or not) and semantic categorization tasks (indicating whether or not a word belongs to a category, such as foods or animals) at above chance levels when words are presented too rapidly to support letter-by-letter reading. Brevity of presentation is critical in that longer exposure to the letter string seems to engage the letter-by-letter strategy, which appears to interfere with the ability to perform the covert reading task. In fact, the patient may show better performance on lexical decision at shorter (e.g., 250 msec) than at longer presentations (e.g., 2 sec) that engage the letter-by-letter strategy but do not allow it to proceed to completion. A compelling example comes from a previously reported patient who was given 2 sec to scan the card containing the stimulus. The patient did not take advantage of the full inspection time when he was performing lexical decision and categorization tasks; instead, he glanced at the card briefly and looked away, perhaps to avoid letter-by-letter reading. The capacity for covert reading has also been demonstrated in two pure alexics who were completely unable to employ the letter-by-letter reading strategy. These patients appeared to recognize words but were rarely able to report them, although they sometimes generated descriptions that were related to the word's meaning (e.g., "cookies" → "candy, a cake"). In some cases, patients have shown some recovery of oral reading over time, although this capacity appears to be limited to concrete words.

The mechanisms that underlie implicit or covert reading remain controversial. Dejerine, who provided the first description of pure alexia, suggested that the analysis of visual input in these patients is performed by the right hemisphere as a result of the damage to the visual cortex on the left. It should be noted, however, that not all lesions to the left visual cortex give rise to alexia. A critical feature that supports continued left hemisphere processing is the preservation of callosal input from the visual processing on the right. One possible account is that covert reading reflects printed word recognition on the part of the right hemisphere, which is unable either to articulate the word or (in most cases) to adequately communicate its identity to the language area of the left hemisphere. By this account, letter-by-letter reading is carried out by the left hemisphere using letter information transferred serially and inefficiently from the right. Furthermore, this assumes that when the letter-by-letter strategy is implemented, it may be difficult for the patient to attend to the products of word processing in the right hemisphere. Consequently, performance on lexical decision and categorization tasks declines. Additional evidence supporting the right hemisphere account of reading in pure alexia is presented later.

Alternative accounts of pure alexia have also been proposed. Behrmann and colleagues, for example, proposed that the disorder is attributable to impaired activation of orthographic representations. By this account, reading is assumed to reflect the residual functioning of the same interactive system that supported normal reading premorbidly.

Other investigators have attributed pure dyslexia to a visual impairment that precludes activation of orthographic representations. Chialant and Caramazza, for example, reported a patient, MJ, who processed single, visually presented letters normally and performed well on a variety of tasks assessing the orthographic lexicon with auditorily presented stimuli. In contrast, MJ exhibited significant impairments in the processing of letter strings. The investigators suggest that MJ was unable to transfer information from the intact visual processing system in the right hemisphere to the intact language processing mechanisms of the left hemisphere.

B. Neglect Dyslexia

Parietal lobe lesions can result in a deficit that involves neglect of stimuli on the side of space contralateral to the lesion, a disorder referred to as hemispatial. In most cases, this disturbance arises with damage to the right parietal lobe; therefore, attention to the left side of space is most often affected. The severity of neglect is generally greater when there are stimuli on the right as well as on the left; attention is drawn to the right-sided stimuli at the expense of those on the left—a phenomenon known as "extinction." Typical clinical manifestations include bumping into objects on the left, failure to dress the left side of the body, drawing objects that are incomplete on the left, and reading problems that involve neglect of the left portions of words (i.e., neglect dyslexia).

With respect to neglect dyslexia, it has been found that such patients are more likely to ignore letters in nonwords (e.g., the first two letters in "bruggle") than letters in real words (compare with "snuggle"). This suggests that the problem does not reflect a total failure to process letter information but, rather, an attentional impairment that affects conscious recognition of the letters. Performance often improves when words are

presented vertically or spelled aloud. In addition, there is evidence that semantic information can be processed in neglect dyslexia, and that the ability to read words aloud improves when oral reading follows a semantic task.

Neglect dyslexia has also been reported in patients with left hemisphere lesions. In these patients, the deficiency involves the right side of words. Here, visual neglect is usually confined to words and is not ameliorated by presenting words vertically or spelling them aloud. This disorder has therefore been termed a "positional dyslexia," whereas the right hemisphere deficit has been termed a "spatial neglect dyslexia."

C. Attentional Dyslexia

Attentional dyslexia is a disorder characterized by at least relatively preserved reading of single words but impaired reading of words in the context of other words or letters. This infrequently reported disorder was first described by Shallice and Warrington, who reported two patients with brain tumors involving (at least) the left parietal lobe. Both patients exhibited relatively good performance with single letters or words but were significantly impaired in the recognition of the same stimuli when presented as part of an array. For example, both patients read single letters accurately but made significantly more errors naming letters when presented as part of 3×3 or 5×5 arrays. Similarly, both patients correctly read more than 90% of single words but read only approximately 80% of words when presented in the context of three additional words. Although not fully investigated, it is worth noting that the patients were also impaired in recognizing line drawings and silhouettes when presented in an array.

Two additional observations from these patients warrant attention. First, Shallice and Warrington demonstrated that for both patients naming of single black letters was adversely affected by the simultaneous presentation of red flanking stimuli and that flanking letters were more disruptive than numbers. For example, both subjects were more likely to correctly name the black (middle) letter when presented "37L82" compared to "ajGyr." Second, the investigators examined the errors produced in the tasks in which patients were asked to report letters and words in rows of two to four items. They found different error patterns with letters and words. Whereas both patients tended to err in the letter report task by naming letters that appeared in a different location in the array, patients often named words that were not present in the array. Interestingly, many of these errors were interpretable as letter transpositions between words. Citing the differential effects of letter versus number flankers as well as the absence of findings suggesting a deficit in response selection, these investigators attributed the disorder to a failure of transmission of information from a nonsemantic perceptual stage to a semantic processing stage. Another patient, BAL, was also reported by Warrington and colleagues. BAL was able to read single words but exhibited a substantial impairment in the reading of letters and words in an array. BAL exhibited no evidence of visual disorientation and was able to identify a target letter in an array of "X"s or "O"s. He was impaired, however, in the naming of letters or words when these stimuli were flanked by other members of the same stimulus category. This patient's attentional dyslexia was attributed to an impairment arising after words and letters had processed as units.

Recently, Saffran and Coslett reported a patient, NY, with biopsy-proven Alzheimer's disease that appeared to selectively involve posterior cortical regions who exhibited attentional dyslexia. NY scored within the normal range on verbal subtests of the WAIS-R but was unable to perform any of the performance subtests. He performed normally on the Boston Naming Test but performed quite poorly on a variety of experimental tasks assessing visuospatial processing and visual attention. Despite his visuoperceptual deficits, NY's reading of single words was essentially normal. He read 96% of 200 words presented for 100 msec (unmasked). Like previously reported patients with this disorder, NY exhibited a substantial decline in performance when asked to read two words presented simultaneoulsy. He read both words correctly in only 50% of 385 trials with a 250-msec stimulus exposure. Most errors were omissions of one word. Of greatest interest, however, was the fact that NY produced a substantial number of "blend" errors in which letters from the two words were combined to generate a response that was not present in the display. For example, when shown "flip shot," NY responded "ship." Like the blend errors produced by normal subjects with brief stimulus presentation. NY's blend errors were characterized by the preservation of letter position information; thus, in the preceding example, the letters in the blend response ("ship") retained the same serial position in the incorrect response. NY produced significantly more blend errors than did five controls whose overall level of performance had been matched to NY's by virtue of

brief stimulus exposure (range, 17–83 msec). A subsequent experiment demonstrated that for NY, but not for controls, blend errors were encountered significantly less often when the target words differed in case (''desk FEAR'').

Saffran and Coslett considered the central deficit in attentional dyslexia to be impaired control of a filtering mechanism that normally serves to suppress input from unattended words or letters in the display. Specifically, they suggested that as a consequence of the patient's inability to effectively deploy the ''spotlight'' of attention to a particular region of interest (e.g., a single word or a single letter), multiple stimuli fall within the attentional spotlight. Because visual attention may serve to integrate visual feature information, impaired modulation of the spotlight of attention would be expected to generate word blends and other errors reflecting the incorrect concatenation of letters.

Saffran and Coslett also argued that loss of location information also contributed to NY's reading deficit. Several lines of evidence support such a conclusion. First, NY was impaired relative to controls with respect to both accuracy and reaction time on a task in which he was required to indicate if a line was inside or outside a circle. Second, NY exhibited a clear tendency to omit one member of a double-letter pair (e.g., ''reed'' → ''red''). This phenomenon, also demonstrated in normal subjects, has been attributed to the loss of location information that normally helps to differentiate two tokens of the same object. Finally, it should be noted that the well-documented observation that the blend errors of normal subjects and attentional dyslexics preserve letter position is not inconsistent with the claim that impaired location information contributes to attentional dyslexia. Migration or blend errors reflect a failure to link words or letters to a location in space, whereas the letter position constraint reflects the properties of the word processing system. The latter, which is assumed to be at least relatively intact in patients with attentional dyslexia, specifies letter location with respect to the word form rather than space.

D. Other Peripheral Dyslexias

Peripheral dyslexias may be observed in a variety of conditions involving visuoperceptual or attentional deficits. Patients with simultanagnosia, a disorder characterized by an inability to ''see'' more than one object in an array, are often able to read single words but are incapable of reading text. Other patients with simultanagnosia exhibit substantial problems in reading even single words.

Patients with degenerative conditions involving the posterior cortical regions may also exhibit profound deficits in reading as part of their more general impairment in visuospatial processing. Several patterns of impairment may be observed in these patients. Some patients exhibit attentional dyslexia with letter migration and blend errors, whereas other patients exhibiting deficits that are in certain respects similar do not produce migration or blend errors in reading or illusory conjuctions in visual search tasks. It has been suggested that at least some patients with these disorders suffer from a progressive restriction in the domain to which they can allocate visual attention. As a consequence of this impairment, these patients may exhibit an effect of stimulus size such that they are able to read a word in small print but when shown the same word in large print see only a single letter.

IV. CENTRAL DYSLEXIAS

A. Deep Dyslexia

Deep dyslexia, initially described by Marshall and Newcombe in 1973, is the most extensively investigated of the central dyslexias and, in many respects, the most compelling. Interest in deep dyslexia is due in large part to the intrinsically interesting hallmark of the syndrome—the production of semantic errors. Shown the word ''castle,'' a deep dyslexic may respond ''knight''; shown the word ''bird,'' the patient may respond ''canary.'' At least for some deep dyslexics, it is clear that these errors are not circumlocutions. Semantic errors may represent the most frequent error type in some deep dyslexics, whereas in other patients they comprise a small proportion of reading errors. Deep dyslexics also typically produce frequent ''visual'' errors (e.g., ''skate'' read as ''scale'') and morphological errors in which a prefix or suffix is added, deleted, or substituted (e.g., ''scolded'' read as ''scolds'' and ''governor'' read as ''government'').

Additional features of the syndrome include a greater success in reading words of high compared to low imageability. Thus, words such as ''table,'' ''chair,'' ''ceiling,'' and ''buttercup,'' the referent of which is concrete or imageable, are read more successfully than words such as ''fate,'' ''destiny,'' ''wish,'' and ''universal,'' which denote abstract concepts.

Another characteristic feature of deep dyslexia is a part of speech effect such that nouns are typically read more reliably than modifiers (adjectives and adverbs), which are in turn read more accurately than verbs. Deep dyslexics manifest particular difficulty in the reading of functors (a class of words that includes pronouns, prepositions, conjunctions, and interrogatives including "that," "which," "they," "because," and "under"). The striking nature of the part of speech effect may be illustrated by the patient who correctly read the word "chrysanthemum" but was unable to read the word "the." Most errors to functors involve the substitution of a different functor ("that" read as "which") rather than the production of words of a different class, such as nouns or verbs. Because functors are in general less imageable than nouns, some investigators have claimed that the apparent effect of part of speech is in reality a manifestation of the pervasive imageability effect. There is no consensus on this point because other investigators have suggested that the part of speech effect is observed even if stimuli are matched for imageability.

All deep dyslexics exhibit a substantial impairment in the reading of nonwords. When confronted with letter strings such as "flig" or "churt," deep dyslexics are typically unable to employ print-to-sound correspondences to derive phonology; nonwords frequently elicit "lexicalization" errors (e.g., "flig" read as "flag"), perhaps reflecting a reliance on lexical reading in the absence of access to reliable print-to-sound correspondences.

Finally, it should be noted that the accuracy of oral reading may be determined by context. This is illustrated by the fact that a patient was able to read aloud the word "car" when it was a noun but not when the same letter string was a conjunction. Thus, when presented the sentence "Le car ralentit car le moteur chauffe" ("The car slowed down because the motor overheated"), the patient correctly pronounced only the first instance of "car." Recently, three deep dyslexics were demonstrated to read function and content words better in a sentence context than when presented alone.

How can deep dyslexia be accommodated by the information processing model of reading illustrated in Fig. 1? Several alternative explanations have been proposed. Some investigators have argued that the reading of deep dyslexics is mediated by a damaged form of the left hemisphere-based system employed in normal reading. In such a hypothesis, multiple processing deficits must be hypothesized to accommodate the full range of symptoms characteristic of deep dyslexia. First, the strikingly impaired performance in reading nonwords and other tasks assessing phonologic function suggests that the print-to-sound conversion procedure is disrupted. Second, the presence of semantic errors and the effects of imageability (a variable thought to influence processing at the level of semantics) suggest that these patients also suffer from a semantic impairment. Lastly, the production of visual errors suggests that these patients suffer from impairment in the visual word form system or in the processes mediating access to the visual word form system.

Other investigators have argued that deep dyslexics' reading is mediated by a system not normally used in reading (i.e., the right hemisphere). Finally, citing evidence from functional imaging studies demonstrating that deep dyslexic subjects exhibit increased activation in both the right hemisphere and nonperisylvian areas of the left hemisphere, other investigators have suggested that deep dyslexia reflects the recruitment of both right and left hemisphere processes.

B. Phonological Dyslexia: Reading without Print-to-Sound Correspondences

First described in 1979, phonologic dyslexia is perhaps the "purest" of the central dyslexias in that, at least by some accounts, the syndrome is attributable to a selective deficit in the procedure mediating the translation from print to sound. Thus, although in many respects less arresting than deep dyslexia, phonological dyslexia is of considerable theoretical interest.

Phonologic dyslexia is a disorder in which reading of real words may be nearly intact or only mildly impaired. Patients with this disorder, for example, correctly read 85–95% of real words. Some patients read all different types of words with equal facility, whereas other patients are relatively impaired in the reading of functors (or "little words"). Unlike patients with surface dyslexia described later, the regularity of print-to-sound correspondences is not relevant to their performance; thus, phonologic dyslexics are as likely to correctly pronounce orthographically irregular words such as "colonel" as words with standard print-to-sound correspondences such as "administer." Most errors in response to real words bear a visual similarity to the target word (e.g., "topple" read as "table").

The striking and theoretically relevant aspect of the performance of phonologic dyslexics is a substantial impairment in the oral reading of nonword letter

strings. We have examined patients with this disorder, for example, who read >90% of real words of all types but correctly pronounce only approximately 10% of nonwords. Most errors to nonwords involve the substitution of a visually similar real word (e.g., ''phope'' read as ''phone'') or the incorrect application of print-to-sound correspondences [e.g., ''stime'' read as ''stim'' (to rhyme with ''him'')].

Within the context of the reading model depicted in Fig. 1, the account for this disorder is relatively straightforward. Good performance with real words suggests that the processes involved in normal lexical reading (i.e., visual analysis, the visual word form system, semantics, and the phonological output lexicon) are at least relatively preserved. The impairment in nonword reading suggests that the print-to-sound translation procedure is disrupted.

Recent explorations of the processes involved in nonword reading have identified many distinct procedures involved in this task. If these distinct procedures may be selectively impaired by brain injury, one might expect to observe different subtypes of phonologic dyslexia. Although the details are beyond the scope of this article, there is evidence suggesting that different subtypes of phonologic dyslexia may be observed.

Lastly, it should be noted that several investigators have suggested that phonologic dyslexia is not attributable to a disruption of a reading-specific component of the cognitive architecture but, rather, to a more general phonologic deficit. Support for this assertion comes from the observation that the vast majority of phonologic dyslexics are impaired on a wide variety of nonreading tasks assessing phonology.

In certain respects, phonologic dyslexia is similar to deep dyslexia, the critical difference being that semantic errors are not observed in phonologic dyslexia. Citing the similarity of reading performance and the fact that deep dyslexics may evolve into phonologic dyslexics as they improve, it has been argued that deep and phonologic dyslexia are on a continuum of severity.

C. Surface Dyslexia: Reading without Lexical Access

Surface dyslexia, first described by Marshall and Newcombe, is a disorder characterized by the relatively preserved ability to read words with regular or predictable grapheme-to-phoneme correspondences but substantially impaired reading of words with ''irregular'' or exceptional print-to-sound correspondences. Thus, patients with surface dyslexia typically are able to read words such as ''state,'' ''hand,'' ''mosquito,'' and ''abdominal'' quite well, whereas they exhibit substantial problems reading words such as ''colonel,'' ''yacht,'' ''island,'' and ''borough,'' the pronunciation of which cannot be derived by sounding out strategies. Errors to irregular words usually consist of ''regularizations''; for example, surface dyslexics may read ''colonel'' as ''kollonel.'' These patients read nonwords (e.g., ''blape'') quite well. Finally, it should be noted that all surface dyslexics reported to date read at least some irregular words correctly; patients will often read high-frequency irregular words (e.g., ''have'' and ''some'') but some surface dyslexics have been reported to read such low-frequency and highly irregular words as ''sieve'' and ''isle.''

As noted previously, some accounts of normal reading postulate that familiar words are read aloud by matching the letter string to a stored representation of the word and retrieving the pronunciation by means of a mechanism linked to semantics or by means of a ''direct'' route. A critical point to note is that because reading involves stored associations of letter strings and sounds, the pronunciation of the word is not computed by rules but is retrieved, and therefore whether the word contains regular or irregular correspondences does not appear to play a major role in performance.

The fact that the nature of the print-to-sound correspondences significantly influences performance in surface dyslexia suggests that the deficit in this syndrome is in the mechanisms mediating lexical reading—that is, in the semantically mediated and ''direct'' reading mechanisms. Similarly, the preserved ability to read words and nonwords demonstrates that the procedures by which words are sounded out are at least relatively preserved.

In the context of the information processing discussed previously, how would one account for surface dyslexia? Scrutiny of the model depicted in Fig. 1 suggests that at least three different deficits may result in surface dyslexia. First, this disorder may arise from a deficit at the level of the visual word form system that disrupts the processing of words as units. As a consequence of this deficit, subjects may identify ''sublexical'' units (e.g., graphemes or clusters of graphemes) and identify words on the basis of print-to-sound correspondences. Note that semantics and output processes would be expected to be preserved. The patient JC described by Marshall and Newcombe exhibited at least some of the features of this type of surface dyslexia. For example, in response to the word

"listen," JC said "Liston" (a former heavyweight champion boxer) and added "that's the boxer," demonstrating that he was able to derive phonology from print and subsequently access meaning.

In the model depicted in Fig. 1, one might also expect to encounter surface dyslexia with deficits at the level of the output lexicon. Support for such an account comes from patients who comprehend irregular words but regularize these words when asked to read aloud. For example, MK read the word "steak" as "steek" (as in seek) before adding, "nice beef." In this instance, the demonstration that MK was able to provide appropriate semantic information indicates that he was able to access meaning directly from the written word and suggests that the visual word form system and semantics are at least relatively preserved.

In the model depicted in Fig. 1, one might also expect to observe semantic dementia in patients exhibiting semantic loss. Indeed, most patients with surface dyslexia (often in association with surface dysgraphia) exhibit a significant semantic deficit. Surface dyslexia is most frequently observed in the context of semantic dementia, a progressive degenerative condition characterized by a gradual loss of knowledge in the absence of deficits in motor, perceptual, and, in some instances, executive function.

Note, however, that the information processing account of reading depicted in Fig. 1 also incorporates a lexical but nonsemantic reading mechanism by means of which patients with semantic loss would be expected to be able to read even irregular words not accommodated by the grapheme-to-phoneme procedure. Surface dyslexia is assumed to reflect impairment in both the semantic and the lexical but nonsemantic mechanisms. It should be noted that in this context the triangle model of reading developed by Seidenberg and McClelland provides an alternative account of surface dyslexia. In this account, to which I briefly return later, surface dyslexia is assumed to reflect the disruption of semantically mediated reading.

V. READING AND THE RIGHT HEMISPHERE

One important and controversial issue regarding reading concerns the putative reading capacity of the right hemisphere. For many years investigators argued that the right hemisphere was "word blind." In recent years, however, several lines of evidence have suggested that the right hemisphere may possess the capacity to read. Indeed, as previously noted, many investigators have argued that the reading of deep dyslexics is mediated at least in part by the right hemisphere.

One seemingly incontrovertible finding demonstrating that at least some right hemispheres possess the capacity to read comes from the performance of a patient who underwent a left hemispherectomy at age 15 for treatment of seizures caused by Rasmussen's encephalitis. After the hemispherectomy the patient was able to read approximately 30% of single words and exhibited an effect of part of speech; she was unable to use a grapheme-to-phoneme conversion process. Thus, as noted by the authors, this patient's performance was similar in many respects to that of patients with deep dyslexia, a pattern of reading impairment that has been hypothesized to reflect the performance of the right hemisphere.

The performance of some split-brain patients is also consistent with the claim that the right hemisphere is literate. For example, these patients may be able to match printed words presented to the right hemisphere with an appropriate object. Interestingly, the patients are apparently unable to derive sound from the words presented to the right hemisphere; thus, they are unable to determine if a word presented to the right hemisphere rhymes with an auditorally presented word.

Another line of evidence supporting the claim that the right hemisphere is literate comes from the evaluation of the reading of patients with pure alexia and optic aphasia. We reported data, for example, from four patients with pure alexia who performed well above chance on many lexical decision and semantic categorization tasks with briefly presented words that they could not explicitly identify. Three of the patients who regained the ability to explicitly identify rapidly presented words exhibited a pattern of performance consistent with the right hemisphere reading hypothesis. These patients read nouns better than functors and words of high imageability (e.g., "chair") better than words of low imageability (e.g., "destiny"). Additionally, both patients for whom data are available demonstrated a deficit in the reading of suffixed (e.g., "flower") compared to pseudosuffixed (e.g., "flowed") words. These data are consistent with a version of the right hemisphere reading hypothesis postulating that the right hemisphere lexical–semantic system primarily represents high imageability nouns. By this account, functors, affixed words, and low imageability words are not adequately represented in the right hemisphere. An important additional finding is that magnetic stimulation applied to the skull, which disrupts electrical activity in the brain, interfered with

the reading performance of a partially recovered pure alexic when it affected the parietooccipital area of the right hemisphere. The same stimulation had no effect when it was applied to the homologous area on the left. Additional data supporting the right hemisphere hypothesis come from the demonstration that the limited whole-word reading of a pure alexic was abolished after a right occipitotemporal stroke.

Although a consensus has not been achieved, there is mounting evidence that, at least for some people, the right hemisphere is not word blind but may support the reading of some types of words. The full extent of this reading capacity and whether it is relevant to normal reading, however, remains unclear.

VI. THE ANATOMIC BASIS OF DYSLEXIA

A variety of experimental techniques, including position emission tomography, functional magnetic resonance imaging (fMRI), and evoked potentials, have been employed to investigate the anatomic basis of reading in normal subjects. Although differences in experimental technique and design inevitably lead to variability in reported sites of activation, there appears to be at least relative agreement regarding the anatomic basis of several components of the reading system.

As previously noted, most accounts of reading postulate that after initial visual processing, familiar words are recognized by comparison to a catalog of stored representations that is often termed the visual word form system. A variety of recent investigations involving visual lexical decision with fMRI, viewing of letter, and direct recording of cortical electrical activity suggests that the visual word form system is supported by inferior occipital or inferior temporooccipital cortex.

Additional evidence for this localization comes from a recent investigation by Cohen *et al.* of five normal subjects and two patients with posterior callosal lesions. These investigators presented words and nonwords for lexical decision or oral reading to either the right or the left visual fields. They found initial unilateral activation in what was thought to be V4 in the hemisphere to which the stimulus was projected. More importantly, however, for normal subjects activation in the left fusiform gyrus (Talairach coordinates −42, −57, and −6) that was independent of the hemisphere to which the stimulus was presented was observed. The two patients with posterior callosal lesions were impaired in the processing of letter strings presented to the right compared to the left hemisphere; fMRI in these subjects demonstrated that the region of the fusiform gyrus described previously was activated in the callosal patients only by left hemisphere stimulus presentation. As noted by the investigators, these findings are consistent with the hypothesis that the hemialexia demonstrated by the callosal patients is attributable to a failure to access the visual word form system in the left fusiform gyrus.

Deriving meaning from visually presented words requires access to stored knowledge or semantics. Although the architecture and anatomic bases of semantic knowledge remain controversial, investigations involving semantic access for written words implicate cortex at the junction of the superior and middle temporal gyrus (Brodman areas 21, 22, and 37).

VII. ALTERNATIVE MODELS OF READING

Our discussion has focused on a "box and arrow" information processing account of reading disorders. This account not only has proven useful in terms of explaining data from normal and brain-injured subjects but also has predicted syndromes of acquired dyslexia. One weakness of these models, however, is the fact that the accounts are largely descriptive and underspecified.

In recent years, many investigators have developed models of reading in which the architecture and procedures are fully specified and implemented in a manner that permits an empirical assessment of their performance. One computational account of reading has been developed by Coltheart and colleagues. Their dual-route cascaded model represents a computationally instantiated version of dual-route theory similar to that presented in Fig. 1. This account incorporates a lexical route (similar to C in Fig. 1) as well as a nonlexical route by which the pronunciation of graphemes is computed on the basis of position-specific correspondence rules. The model accommodates a wide range of findings from the literature on normal reading.

A fundamentally different type of reading model was developed by Seidenberg and McClelland and subsequently elaborated by Plaut, Seidenberg, and colleagues. This model belongs to the general class of parallel distributed processing or connectionist models. Sometimes referred to as the triangle model, this approach differs from information processing models in that it does not incorporate word-specific representations (e.g., visual word forms and output phonologic representations). In this approach, subjects are assumed to learn how written words map onto spoken

words through repeated exposure to familiar and unfamiliar words. Learning of word pronunciations is achieved by means of the development of a mapping between letters and sounds generated on the basis of experience with many different letter strings. The probabilistic mapping between letters and sounds is assumed to provide the means by which both familiar and unfamiliar words are pronounced. This model not only accommodates an impressive array of the classic findings in the literature on normal reading but also has been "lesioned" in an attempt to reproduce the patterns of reading characteristic of dyslexia. For example, Patterson *et al*. attempted to accommodate surface dyslexia by disrupting the semantically mediated reading, and Plaut and Shallice generated a pattern of performance similar to that of deep dyslexia by lesioning a somewhat different connectionist model.

A full discussion of the relative merits of these models as well as approaches to the understanding of reading and acquired dyslexia is beyond the scope of this article. It appears likely, however, that investigations of acquired dyslexia will help to adjudicate between competing accounts of reading and also that these models will continue to offer critical insights into the interpretation of data from brain-injured subjects.

Acknowledgment

This work was supported by Grant RO1 DC2754 from the National Institute of Deafness and Other Communication Disorders.

See Also the Following Articles

AGRAPHIA • ALEXIA • LANGUAGE ACQUISITION • LANGUAGE AND LEXICAL PROCESSING • LANGUAGE DISORDERS • READING DISORDERS, DEVELOPMENTAL

Suggested Reading

Coltheart, M. (1996). Phonological dyslexia: Past and future issues. *Cognitive Neuropsychol.* **13** (Special issue).

Coltheart, M. (1998). Letter-by-letter reading. *Cognitive Neuropsychol.* **15** (Special issue).

Coltheart, M., and Rastle, K. (1994). Serial processing in reading aloud: Evidence for dual-route models of reading. *J. Exp. Psychol. Hum. Perception Performance* **20,** 1197–1211.

Coltheart, M., Patterson, K., and Marshall, J. C. (Eds.) (1980). *Deep Dyslexia.* Routledge Kegan Paul, London.

Coslett, H. B. (1991). Read but not write "idea": Evidence for a third reading mechanism. *Brain Language* **40,** 425–443.

Coslett, H. B., and Saffran, E. M. (1989). Evidence for preserved reading in "pure alexia". *Brain* **112,** 327–359.

Coslett, H. B., and Saffran, E. M. (1998). *Reading and the right hemisphere: Evidence from acquired dyslexia.* In *Right Hemisphere Language Comprehension* (M. Beeman and C. Chiarello, Eds.), pp. 105–132. Erlbaum, Mahway, NJ.

Fiez, J. A., and Petersen, S. E. (1998). Neuroimaging studies of word reading. *Proc. Natl. Acad. Sci. USA* **95,** 914–921.

Funnell, E. (Ed.) (2000). *Case Studies in the Neuropsychology of Reading*. Psychology Press, Hove, UK.

Marshall, J. C., and Newcombe, F. (1973). Patterns of paralexia: A psycholinguistic approach. *J. Psycholinguistic Res.* **2,** 175–199.

Patterson, K. E., Marshall, J. C., and Coltheart, M. (Eds.) (1985). *Surface Dyslexia.* Routledge Kegan Paul, London.

Patterson, K. E., Plaut, D. C., McClelland, J. L., Seidenberg, M. S., Behrmann, M., and Hodges, J. R. (1997). *Connections and disconnections: A connectionist account of surface dyslexia.* In *Neural Modeling of Cognitive and Brain Disorders* (J. Reggia, R. Berndt, and E. Ruppin, Eds.), World Scientific, New York.

Plaut, D. C., and Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychol.* **10,** 377–500.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., and Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychol. Rev.* **103,** 56–115.

Saffran, E. M., and Coslett, H. B. (1996). "Attentional dyslexia" in Alzheimer's disease: A case study. *Cognitive Neuropsychol.* **13,** 205–228.

Electrical Potentials

FERNANDO H. LOPES DA SILVA
University of Amsterdam and Dutch Epilepsy Clinic Foundation

GLOSSARY

alpha rhythms Electroencephalogram (EEG)/magnetoencephalogram (MEG) rhythmic activity at 8–13 Hz occurring during wakefulness, mainly over the posterior regions of the head. It is predominant when the subject has closed eyes and is in a state of relaxation. It is attenuated by attention, especially visual, and by mental effort. There are other rhythmic activities within the same frequency range, such as the mu rhythm and the tau or temporal alphoid rhythm, that in MEG recordings have been associated with cortical auditory function, but these are less clear in EEG recordings.

beta/gamma rhythms These rhythmic activities represent frequencies higher than 13 Hz; typically, the beta rhythmic activity ranges from 13 to approximately 30–40 Hz and the gamma extends beyond the latter frequency. Rhythmical beta/gamma activities are encountered mainly over the frontal and central regions and are not uniform phenomena. These rhythmic activities can occur in relation to different behaviors, such as during movements or relaxation after a movement.

desynchronization A state in which neurons are randomly active. Desynchronization is reflected in the absence of a preferred EEG/MEG frequency component. Event-related desynchronization of ongoing EEG/MEG is a state characterized by a decrease in the power of spectral peaks at specific frequency components, elicited by an event.

electroencephalogram, magnetoencephalogram Electrical potentials or magnetic fields recorded from the brain, directly or through overlying tissues.

excitatory synaptic potential or current An active electrical response of the postsynaptic membrane of a neuron to the release of a neurotransmitter that consists of a local, graded depolarization or of the corresponding ionic current.

event-related potential or field A change in electrical or magnetic activity related to an event, either sensory or motor. The event may precede or follow the event-related potential or field.

evoked potentials or fields A change in electrical or magnetic activity in response to a stimulus. Typically, these transients last tens or hundreds of milliseconds.

inhibitory postsynaptic potential/current Consists of a local, graded hyperpolarization or the corresponding ionic current.

mu or rolandic (central) rhythm Rhythmic activity within the same frequency range as the posterior alpha rhythm but with a topographic distribution that is predominant over the central sensorimotor areas. It is attenuated or blocked by movements.

nonlinear dynamics and brain oscillations Systems with nonlinear elements, such as neuronal networks, may exhibit complex dynamics. Typically, these systems may have different kinds of evolution in time and may switch from one oscillatory mode to another. A qualitative change in the dynamics that occurs as a system's parameter varies is called a bifurcation. Some brain oscillations appear to be generated by systems of this kind, with complex nonlinear dynamics.

sleep or sigma spindles Waxing and waning spindle-like waves occurring during the early stage of sleep at 7–14 Hz within sequences lasting 1 or 2 sec. Sleep spindles typically recur at a slow rhythm of about 0.2–0.5 Hz.

synchronization State in which neurons oscillate in phase as a result of a common input and/or of mutual influences. Event-related synchronization of the ongoing EEG/MEG signals is a state characterized by an increase in power of specific frequency components.

theta rhythms or rhythmical slow activity This term denotes rhythmic activity in the frequency range from 4 to 7 Hz in humans, although macro-osmatic animals show a powerful limbic, and especially hippocampi, rhythm from 3 to 12 Hz that is activated by arousal and motor activity.

The recording of the electrical activity of the brain [i.e., the electroencephalogram (EEG)], either the ongoing activity or the changes of activity related to a given sensory or motor event [the event-related potentials (ERPs)], provides the possibility of studying brain functions with a high time resolution but with a relatively modest spatial resolution. The latter, however, has been improved recently with the development of the magnetoencephalogram (MEG) and more sophisticated source imaging techniques. These new methods allow an analysis of the dynamics of brain activities not only of global brain functions, such as sleep and arousal, but also of cognitive processes, such as perception, motor preparation, and higher cognitive functions. Furthermore, these methods are essential for the characterization of pathophysiological processes, particularly those with a paroxysmal character such as epilepsy. In this article, the possibilities offered by EEG/MEG recordings to analyze brain functions, and particularly the neural basis of cognitive processes, are discussed. In this respect, special attention is given to the basic mechanisms underlying the generation of functional neuronal assemblies (i.e., the processes of synchronization and desynchronization of neuronal activity) and their modulation by exogenous and endogenous factors. Therefore, a brief overview of the neurophysiology of the dynamics of neuronal networks and of the generation of brain oscillations, an account of the phenomenology of EEG/MEG activities, and a discussion of the main aspects of ERPs and/or magnetic fields are presented. The article concludes with a short discussion of the roles that brain oscillations may play in the processing of neural information. In this respect, the need to understand what happens at the level of neuronal elements within brain systems in relation to different types of oscillations is emphasized. Only in this way is it possible to obtain insight into the functional processes that occur in the brain, at the level of information transfer or processing, during the occurrence of brain oscillations and the transition between different states. The relevance of electrophysiological studies of the human brain regarding neurocognitive investigations is also discussed in light of recent advances in brain imaging techniques.

I. THE ELECTROPHYSIOLOGICAL APPROACH VERSUS OTHER APPROACHES OF STUDYING BRAIN FUNCTIONS

In the past decade, the emergence of brain imaging techniques, such as positron emission tomography and functional magnetic resonance (fMRI), has made important contributions to our understanding of basic brain processes underlying neurocognitive functions. Notwithstanding the fact that, in particular, fMRI can provide maps of brain activity with millimeter spatial resolution, its temporal resolution is limited. Indeed, this resolution is on the order of seconds, whereas neurocognitive phenomena may take place within tens of milliseconds. Therefore, it is important to use electrophysiological techniques to achieve the desired temporal resolution in this kind of investigation, although this is done at the cost of spatial resolution. The essential problem of electro/magnetoencephalography is that the sources of neuronal activity cannot be derived in a unique way from the distribution of EEG or MEG activity at the scalp; thus, the inverse problem is an ill-posed problem. However, new possibilities are emerging to solve this problem. A promising solution is to solve the inverse problem by imposing spatial constraints based on anatomical information provided by MRI and additional physiological information from metabolic or hemodynamic signals, derived from fMRI, that are associated with the neuronal activity. The precise nature of this association, however, is not trivial and it is a matter of current investigation. This implies that we may expect that in the near future technical advances in combining fMRI and EEG/MEG measures of brain activity will contribute to a better understanding of neurocognitive and other brain functions.

II. HISTORY OF THE ELECTROPHYSIOLOGY OF THE BRAIN

The existence of the electrical activity of the brain (i.e., the EEG) was discovered more than a century ago by Caton. After the demonstration that the EEG could be recorded from the human scalp by Berger in the 1920s, it was slowly accepted as a method of analysis of brain functions in health and disease. It is interesting to note that acceptance came only after the demonstration by Adrian and Mathews in the 1930s that the EEG, namely the alpha rhythm, was likely generated in the occipital lobes in man. The responsible neuronal

sources, however, remained undefined until the 1970s, when it was demonstrated in dog that the alpha rhythm is generated by a dipole layer centered on layers IV and V of the visual cortex. It is not surprising that the mechanisms of generation and the functional significance of the EEG remained controversial for a relatively long time, considering the complexity of the underlying systems of neuronal generators and the involved transfer of signals from the cortical surface to the scalp due to the geometric and electrical properties of the volume conductor (brain, cerebrospinal fluid, skull, and scalp).

The EEG consists essentially of the summed electrical activity of populations of neurons, with a modest contribution of glial cells. Considering that neurons are excitable cells with characteristic intrinsic electrical properties and that interneuronal communication is essentially mediated by electrochemical processes at synapses, it follows that these cells can produce electrical and magnetic fields that may be recorded at a distance from the sources. Thus, these fields may be recorded a short distance from the sources [i.e., the local EEG or local field potentials (LFPs)], from the cortical surface (the electrocorticogram), or even from the scalp (i.e., the EEG in its most common form). The associated MEG is recorded usually by way of sensors placed at a short distance around the scalp.

In order to understand how the electrical and magnetic signals of the brain are generated, it is necessary to examine how the activity of assemblies of neurons is organized both in time and in space and which biophysical laws govern the generation of extracellular field potentials or magnetic fields.

III. THE GENERATION OF ELECTRIC AND MAGNETIC EXTRACELLULAR FIELDS

It is generally assumed that the neuronal events that cause the generation of electric and/or magnetic fields in a neural mass consist of ionic currents that have mainly postsynaptic sources. For these fields to be measurable at a distance from the sources, it is important that the underlying neuronal currents are well organized both in space and in time. The ionic currents in the brain obey Maxwell's and Ohm's laws.

The most important ionic current sources in the brain consist of changes in membrane conductances caused by intrinsic membrane processes and/or by synaptic actions. The net membrane current that results from changes in membrane conductances, either synaptic or intrinsic, can be either a positive or a negative ionic current directed to the inside of the neuron. These currents are compensated by currents flowing in the surrounding medium since there is no accumulation of electrical charge. Consider as the simplest case that of synaptic activity caused by excitatory postsynaptic currents (EPSCs) or inhibitory postsynaptic currents (IPSCs). Because the direction of the current is defined by the direction along which positive charges are transported, at the level of the synapse there is a net positive inward current in the case of an EPSC and a negative one in the case of an IPSC. Therefore, extracellularly an active current sink is caused by an EPSC and an active current source by an IPSC. Most neurons are elongated cells; thus, along the passive parts of the membrane (i.e., at a distance from the active synapses) a distributed passive source is created in the case of an EPSC and a distributed passive sink in the case of an IPSC. In this way, a dipole configuration is created (Fig. 1). At the macroscopic level, the activation of a set of neurons organized in parallel is capable of creating dipole layers. The following are important conditions that have to be satisfied for this to occur: (i) The neurons should be spatially organized with the dendrites aligned in parallel, forming palisades, and (ii) the synaptic activation of the neuronal population should occur in synchrony.

Lorente de Nó named the type of electric field created in this way an "open field," in contrast to the field generated by neurons with dendrites radially distributed around the soma which form a "closed field." In any case, as a result of synaptic activation, extracellular currents will flow. These may consist of longitudinal or transversal components, the former being those that flow parallel to the main axis of a neuron and the latter flow perpendicular to this axis. In the case of an open field, the longitudinal components will add, whereas the transversal components tend to cancel out. In the case of a closed field, all components will tend to cancel, such that the net result at a distance is zero.

The importance of the spatial organization of neuronal current sources for the generation of electric and/or magnetic fields measurable at a distance can be stated in a paradigmatic way for the cortex. Indeed, the pyramidal neurons of the cortex are lined up perpendicular to the cortical surface, forming layers of neurons in palissade. Their synaptic activation can occur within well-defined layers and in a synchronized way. The resulting electric fields may be quite large if the activity within a population of cells forms a

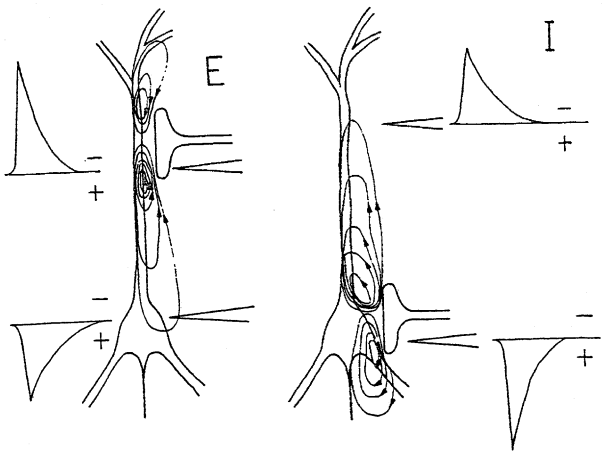


Figure 1 Model cortical pyramidal cell showing the patterns of current flow caused by two modes of synaptic activation at an excitatory (E) and an inhibitory (I) synapse. Typically, the apical dendrites of these cells are oriented toward the cortical surface. E, current flow caused by the activation of an excitatory synapse at the level of the apical dendrite. This causes a depolarization of the postsynaptic membrane (i.e., an EPSP), and the flow of a net positive current (i.e., EPSC). This current flow creates a current sink in the extracellular medium next to the synapse. The extracellularly recorded EPSP is shown on the left. It has a negative polarity at the level of the synapse. At the soma there exists a distributed passive current source resulting in an extracellular potential of positive polarity. I, current flow caused by activation of an inhibitory synapse at the level of the soma. This results in a hyperpolarization of the postsynaptic membrane and in the flow of a negative current. Thus, an active source is created extracellularly at the level of the soma and in passive sinks at the basal and apical dendrites. The extracellularly recorded IPSP at the level of the soma and of the apical dendrites is shown. Note that both cases show a dipolar source–sink configuration, with the same polarity, notwithstanding the fact that the postsynaptic potentials are of opposite polarity (EPSP vs IPSP). This illustrates the fact that not only does the nature of the synaptic potential determine the polarity of the potentials at the cortical surface but also the position of the synaptic sources within the cortex is important (adapted with permission from Niedermeyer and Lopes da Silva, 1999).

coherent domain (i.e., if the activity of the neuronal sources is phase locked). In general, the electric potential generated by a population of neurons represents a spatial and temporal average of the potentials generated by the single neurons within a macrocolumn.

A basic problem in electroencephalography/magnetoencephalography is how to estimate the neuronal sources corresponding to a certain distribution of electrical potentials or of magnetic fields recorded at the scalp. As noted previously, this is called the inverse problem of EEG/MEG. It is an ill-posed problem that has no unique solution. Therefore, one must assume specific models of the sources and of the volume conductor. The simplest source model is a current dipole. However, it should not be considered that such a model means that somewhere in the brain there exists a discrete dipolar source. It simply means that the best representation of the EEG/MEG scalp distribution is by way of an equivalent dipolar source. In the sense of a best statistical fit, the latter describes the centroid of the dipole layers that are active at a certain moment. The estimation of equivalent dipole models is only meaningful if the scalp field has a focal character and the number of possible active areas can be anticipated with reasonable accuracy. An increase in the number of dipoles can easily lead to complex and ambiguous interpretations. Nevertheless, methods have been developed to obtain estimates of multiple dipoles with only the a priori information that they must be located on the surface of the cortex. An algorithm that performs such an analysis is multiple signal classification (MUSIC), which is illustrated in Fig. 2 for the case of the cortical sources of the alpha rhythm. An alternative approach is to use linear estimation methods applying the minimum norm constraint to estimate the sources within a given surface or volume of the brain. Currently, new approaches are being explored that use combined fMRI and EEG/MEG recordings in order to create more specific spatial constraints to reduce the solution space for the estimation of the underlying neuronal sources. In general, the problems created by the complexity of the volume conductor, including the scalp, skull, layer of cerebrospinal fluid, and brain, are easier to solve in the case of MEG than EEG since these media have conductivities that affect the EEG much more than the MEG. The major advantage of MEG over EEG is the relative ease of source localization with the former. This means that when a dipole source algorithm is used on the basis of MEG recordings, a simple homogeneous sphere model of the volume conductor is usually sufficient to obtain a satisfactory solution. The position of the sources can be integrated in MRI scans of the brain using appropriate algorithms.

IV. DYNAMICS OF NEURONAL ELEMENTS: LOCAL CIRCUITS AND MECHANISMS OF OSCILLATIONS

Here, I discuss the dynamics of the electrical and magnetic fields of the brain, i.e., their time-dependent properties. These properties are determined to a large extent by the dynamical properties of ionic currents.

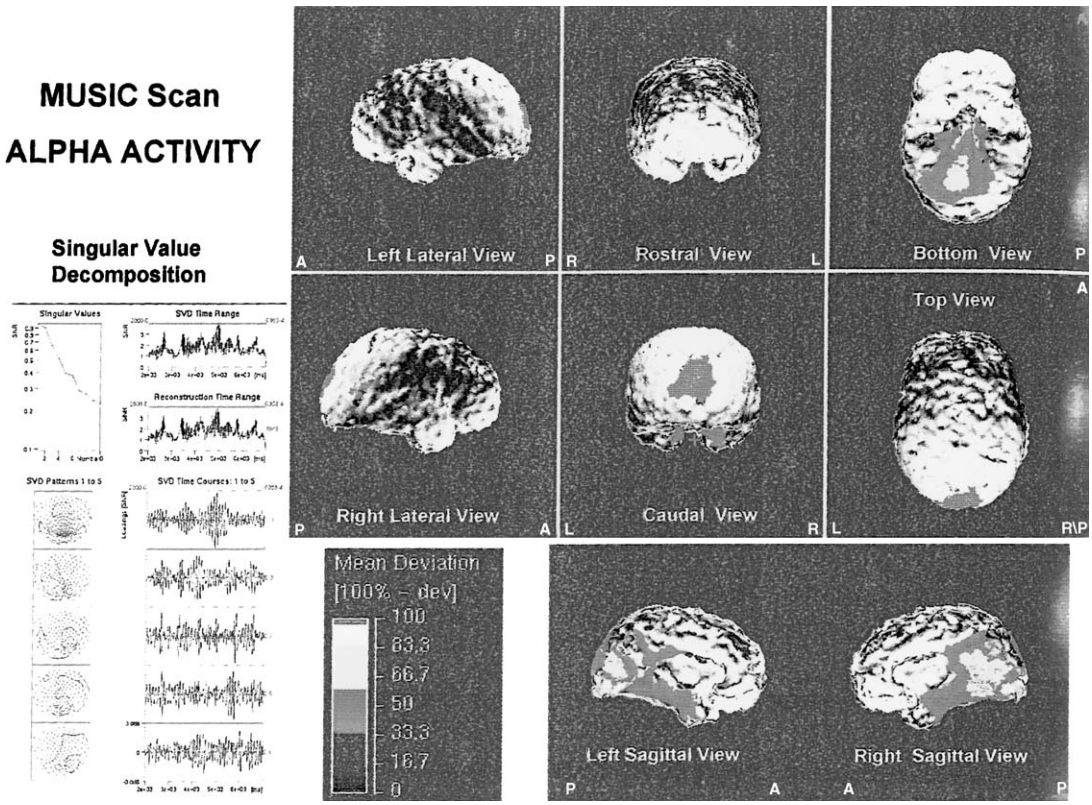


Figure 2 Results of a multiple signal classification analysis (MUSIC) scan of a 5-sec-long epoch of alpha activity recorded using a whole-head 151 MEG system during the eyes-closed condition. A singular value decomposition of the MEG signals yielded four factors. The scale indicates the cortical areas where the most significant sources were estimated (reproduced with permission from Parra *et al.*, *J. Clin. Neurophysiol.* **17,** 212–224, 2000).

The most elementary phenomenon to be taken into account is the passive time constant of the neuronal membrane. This is typically about 5 msec for most cortical pyramidal neurons. Therefore, the activation of a synapse causing an ionic current to flow will generate a postsynaptic potential change that decays passively with such a time constant. However, there are other membrane phenomena that have much longer time constants. Some of these are intrinsic membrane processes, whereas others are postsynaptic. Among the intrinsic membrane phenomena, a multitude of ionic conductances, distributed along both soma and dendrites, enable neurons to display a variety of modes of activity. Among the synaptic phenomena, some present much longer time constants than the membrane passive time constant. This is the case, for example, for the $GABA_B$-mediated inhibition that consists of a K^+ current with slow dynamics. The synaptic actions of amines and neuropeptides, in contrast with those of amino acids such as glutamate and GABA, also have slow dynamics. The effect of acetylcholine (ACh) is mixed since fast and slow actions occur, depending on the type of receptors to which ACh binds. Furthermore, note that the dendrites do not consist of simple passive membranes, as in the classical view; rather, the dendrites also have voltage-gated ion conductances that can contribute actively to the electrical behavior of the whole neuron. Thus, a neuron must be considered as a system of functional nodes, where each node represents a chemical synapse, an active ionic conductance, a metabolic modulated ion channel, or, in some cases, a gap junction (i.e., a direct electrical coupling between adjacent cells through a low-resistance pathway).

The sequence of hyperpolarizations and depolarizations caused by inhibitory and excitatory synaptic activity, respectively, can induce the activation or the removal of inactivation of intrinsic membrane currents. This aspect is difficult to analyze in detail under natural conditions in intact neuronal populations. However, much was learned about this kind of phenomena from studies carried out *in vitro* both in

isolated cells and in brain slices. To illustrate this, consider the typical behavior of thalamic neurons that participate in the thalamocortical circuits with feed-forward and feedback connections (Fig. 3). Even in isolation, these cells may show intrinsic oscillatory modes of activity, either in the alpha range of activity (namely, in the form of spindles at 7–14 Hz) or in the delta frequency range (0.5–4 Hz). To understand how

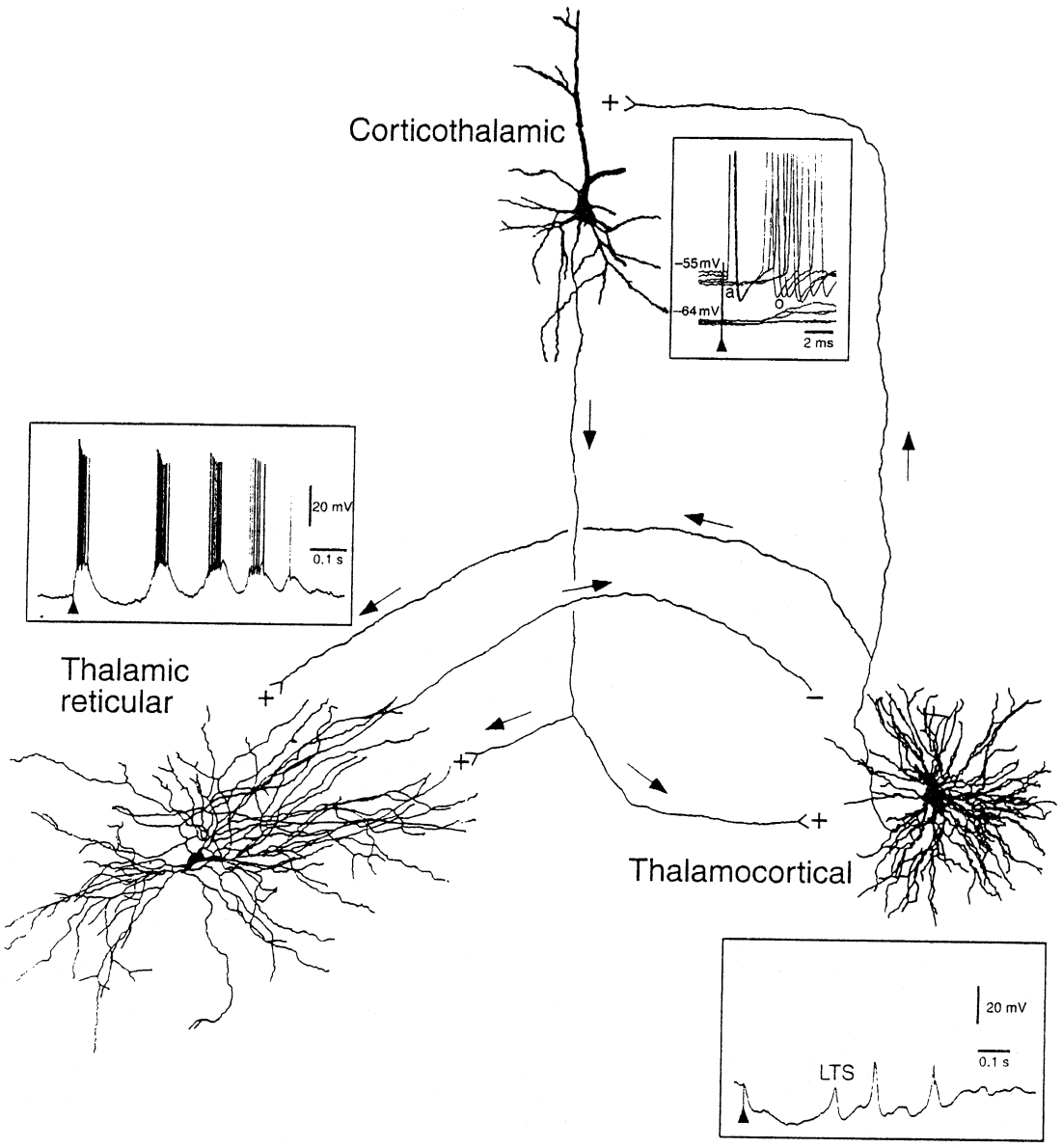


Figure 3 Corticothalamic circuits and different types of neuronal activities. A thalamocortical relay cell is shown on the lower right-hand side, a neuron of the reticular nucleus (GABAergic) is shown on the left, and a cortical pyramidal cell is shown on top. The direction of the flow of action potentials along the axons is indicated by arrows. Positive signs indicate excitatory synapses and negative signs indicate inhibitory synapses. The insets show cellular responses obtained intracellularly. In this case, the activity of the cortical neuron was obtained by electrical stimulation of the intralaminar nucleus of the thalamus. At the membrane potential of −55 mV the response consisted of an antidromic spike (a) followed by a bursts of orthodromic spikes (o). At a more hyperpolarized level (−64 mV), the antidromic response failed but the orthodromic response consisted of a subthreshold EPSP. The responses of the two thalamic neurons were obtained by stimulation of the cortex. The response of the neuron of the reticular nucleus consists of a high-frequency burst of spikes followed by a series of spindle waves on a depolarizing envelope. The response of the thalamocortical neuron consists of a biphasic IPSP that leads to a low-threshold calcium spike (LTS) and a sequence of hyperpolarizing spindle waves (reproduced with permission from Steriade, 1999).

oscillations can take place in these circuits it is important to realize that these cells have, among other kinds of ionic currents, a low-threshold Ca^{2+} conductance (I_T) that contributes to the low-threshold spike. This I_T conductance is de-inactivated by a previous membrane hyperpolarization and causes sufficient depolarization of the cell for the activation of a persistent (non-inactivating) Na+ conductance. These cells also have a nonspecific cation "sag" current (I_h) that has much slower kinetics than I_T and is activated by hyperpolarization. The alpha range oscillatory mode depends essentially on the low-threshold Ca^{2+} current I_T. In addition, other currents contribute to the oscillatory behavior—namely, a fast transient potassium current (I_A) that has a voltage dependence that is similar to that of I_T, a slowly inactivating potassium current, and calcium-dependent potassium currents that need increased intracellular Ca^{2+} concentration for activation—and are mainly responsible for after-hyperpolarizations. In this way, a sequence of hyperpolarizations followed by depolarizations tends to develop. This could be sufficient for such cells to behave as neuronal "pacemakers." Under the *in vitro* conditions in which these basic properties have been studied, the initial hyperpolarization that is necessary for the removal of the inactivation of the Ca^{2+} current is provided artificially by an intracellular injection of current. However, under natural conditions in the intact brain, the oscillatory behavior cannot occur spontaneously (i.e., it is not autonomous). For the de-inactivation of the low-threshold Ca^{2+} current to occur, initially the cell has to be hyperpolarized by a synaptic action mediated by GABAergic synapses that impinge on these cells and stem from the GABAergic neurons of the reticular nucleus of the thalamus. Therefore, the notion that these thalamo-cortical relay cells behave as pacemakers, in the sense of generating pure autonomous oscillations, *in vivo* is only a relative one. Indeed, they need a specific input from outside to set the conditions under which they may oscillate. Thus, *in vivo* the alpha spindle and delta waves result from network interactions. Nevertheless, the intrinsic membrane properties are of importance in setting the frequency response of the cells and of the network to which they belong.

In general terms, it must be considered that neurons, as integrative units, interact with other neurons through local circuits, excitatory as well as inhibitory. In a neuronal circuit feedforward and feedback elements have to be distinguished. In particular, the existence of feedback loops can shape the dynamics of a neuronal network affecting its frequency response and even creating the conditions for the occurrence of resonance phenomena and other forms of oscillatory behavior. Furthermore, note that the transfer of signals in a neuronal network involves time delays and essential nonlinearities. This may lead to the appearance of nonlinear oscillations and possibly even to what is in mathematical terms called chaotic behavior.

In recent years, much has been learned about these issues with the advent of the theory of nonlinear dynamics. This has resulted in the application of nonlinear time series analysis to EEG signals, with the aim of estimating several nonlinear measures to characterize different kinds of EEG signals. This matter is the object of recent interesting studies reported in specialized publications in which the question "chaos in brain?" is discussed. I stress here only that a theoretical framework based on the mathematical notions of complex nonlinear dynamics can be most useful to understanding the dynamics of EEG phenomena. In this context, insight may be obtained into how different types of oscillations may be generated within the same neuronal population and how such a system may switch from one type of oscillatory behavior to another. This occurs, for example, when an EEG/MEG characterized by alpha rhythmic activity suddenly changes into a 3-Hz spike and wave pattern during an absence seizure in epileptic patients. Based on the use of mathematical nonlinear models of neuronal networks, it is possible to formulate hypotheses concerning the mechanisms by means of which a given neuronal network can switch between these qualitatively different types of oscillations. This switching behavior depends on input conditions and on modulating parameters. Accordingly, such a switch can take place depending on subtle changes in one or more parameters. In the theory of complex nonlinear dynamics, this is called a bifurcation. In this respect, the most sensitive parameter is the neuronal membrane potential that in the intact brain is modulated by various synaptic inputs. Typically, the change of oscillation mode may be spectacular, whereas the initial change of a parameter may be minimal.

The frequency of a brain oscillation depends both on the intrinsic membrane properties of the neuronal elements and on the properties of the networks to which they belong. Considerations regarding the membrane conditions that determine which ionic currents can be active at a given time lead to the conclusion that the modulating systems mediated by several neurochemical systems are of utmost importance in setting the initial conditions that determine the

activity mode of the network (i.e., whether it will display oscillations or not and at which frequency). Different behavioral states are characterized by the interplay of different chemical neurotransmitter and neuromodulator systems.

A fundamental property of a neuronal network is the capacity of the neurons to work in synchrony. This depends essentially on the way the inputs are organized and on the network interconnectivity. Thus, groups of neurons may work synchronously as a population due to mutual interactions. The experimental and theoretical work of Ad Aetsen and Moshe Abeles showed the existence of precise (within 5 msec) synchrony of individual action potentials among selected groups of neurons in the cortex. These synchronous patterns of activity were associated with the planning and execution of voluntary movements. These researchers noted that during cognitive processes the neurons tend to synchronize their firing without changing the firing rates, whereas in response to external events they synchronize firing rates and modulate the frequency at the same time. In addition, they also showed that the synchronization dynamics are strongly influenced by the level of background activity. This indicates the importance that the ongoing electrical activity can have in setting the activity climate in a given brain system.

A fundamental feature of the cortex is that groups of neurons tend to form local circuits organized in modules with the geometry of cortical columns. The basic cortical spatial module is a vertical cylinder with a 200–300 μm cross-section. There exist different systems of connecting fibers between cortical columns, namely, (i) the collaterals of axons of pyramidal neurons that may spread over distances of approximately 3 mm and are mostly excitatory; (ii) the ramifications of incoming terminal axons that may extend over distances of 6–8 mm along the cortical surface; and (iii) the collaterals of interneurons, an important part of which are inhibitory, that may branch horizontally over 0.5–1 mm within the cortex. These systems range over distances on the order of magnitude of hundreds of micrometers, and this determines the characteristic length of intracortical interactions.

It is not simple to directly relate the dynamic behavior of a neuronal network to the basic parameters of neurons and synapses. In order to construct such relationships, basic physiological and histological data have to be combined like pieces of a puzzle. However, the available knowledge of most neuronal networks in terms of detailed physiological and even histological information is still incomplete. To supplement this lack of specific knowledge, a synthetic approach can be useful. This implies the construction of connectionist models of neuronal networks using all available and relevant data. Such models can be of practical use to obtain a better understanding of the main properties of a network. Indeed, a model offers the possibility of studying the influence of different parameters on the dynamic behavior of the network and of making predictions of unexplored properties of the system, which may lead to new experiments. In this way, EEG signals, particularly local field potentials (LFPs), can also be modeled. Thus, hypotheses concerning the relevance of given neuronal properties to the generation of special EEG features may be tested.

V. MAIN TYPES OF EEG/MEG ACTIVITIES: PHENOMENOLOGY AND FUNCTIONAL SIGNIFICANCE

A. Sleep EEG Phenomena

In the neurophysiology of sleep two classic EEG phenomena have been established: the spindles or waves between 7 and 14 Hz, also called sleep or sigma spindles, which appear at sleep onset, and the delta waves (1–4 Hz), which are paradigmatic of deeper stages of sleep. Recently, the work of Mircea Steriade and coworkers in Quebec described in animals another very slow oscillation (0.6–1 Hz) that is able to modulate the occurrence of different typical EEG sleep events, such as delta waves, sleep spindles, and even short high-frequency bursts. This very slow oscillation was recently revealed in the human EEG by Peter Acherman and A. Borbély and in the MEG by our group.

The basic membrane events that are responsible for the occurrence of sleep spindles and delta waves were described previously. The sleep spindles are generated in the thalamocortical circuits and result from the interplay between intrinsic membrane properties of the thalamocortical relay neurons (TCR) and the GABAergic neurons of the reticular nucleus, and the properties of the circuits to which these neurons belong. It is clear that the spindles are a collective property of the neuronal populations. Experimental evidence has demonstrated that the sleep spindle oscillations are generated in the thalamus since they can be recorded in this brain area after decortication

and high brain stem transection. The very slow rhythm (0.6–1 Hz), on the contrary, is generated intracortically since it survives thalamic lesions, but it is disrupted by intracortical lesions. Interestingly, note that the rhythmicity of the very slow oscillation appears to be reflected in that of the typical K complexes of human EEG during non-REM sleep.

How are these oscillations controlled by modulating systems? It is well-known that sleep spindles are under brain stem control. It is a classic neurophysiological phenomenon that electrical stimulation of the brain stem can block thalamocortical oscillations causing the so-called "EEG desynchronization," as shown in studies by Moruzzi and Magoun. This desynchronization is caused by the activation of cholinergic inputs (Fig. 4) arising from the mesopontine cholinergic nuclei, namely, the pedunculopontine tegmental and the laterodorsal tegmental areas. Indeed, both the reticular nucleus and the TCR neurons receive cholinergic muscarinic synapses. Cholinergic activation of the reticular nucleus neurons elicits hyperpolarization with a K^+ conductance increase that is mediated by an increase in a muscarinic-activated potassium current. In contrast, it causes depolarization of TCR neurons.

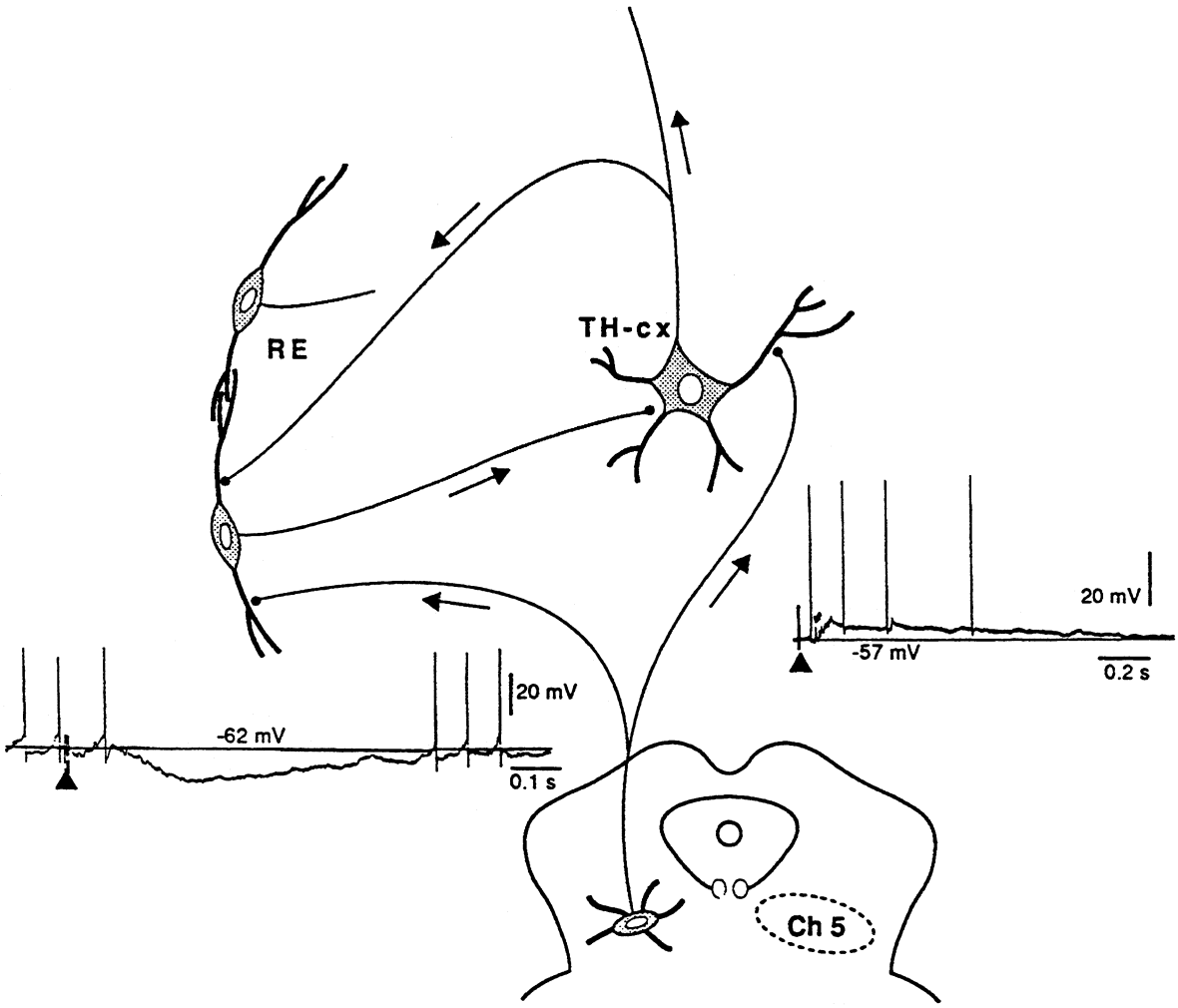


Figure 4 Basic thalamic network responsible for the generation of spindles at 7–14 Hz. A thalamocortical (TH-cx) neuron and two neurons of the reticular nucleus (RE), which are interconnected by mutual inhibitory synapses, are shown. This network is under the modulating influence of mesopontine afferents arising from the cholinergic (Ch5) neurons of the pedunculopontine tegmental nucleus. The stimulation of Ch5 neurons causes depolarization of TH-cx neurons and hyperpolarization of RE neurons. In this way, the occurrence of spindles is blocked (i.e., desynchronization of the corresponding oscillation takes place) (reproduced with permission from Steriade *et al.*, *Electroencephalogr. Clin. Neurophysiol.* **76,** 481–508, 1990).

Furthermore, the reticular nucleus receives inputs from the basal forebrain that may be GABAergic and can also exert a strong inhibition of the reticular neurons leading to the subsequent suppression of spindle oscillations. In addition, monoaminergic inputs from the brain stem, namely those arising at the mesopontine junction (i.e., from the noradrenergic neurons of the locus coeruleus and the serotoninergic neurons of the dorsal raphe nuclei), also modulate the rhythmic activities of the forebrain. These neuronal systems have only a weak thalamic projection but have a diffuse projection to the cortex. Metabotropic glutamate receptors also appear to exert a modulating influence on the activation of thalamic circuits by descending corticothalamic systems.

Because this point is often misunderstood, it is emphasized that slow-wave sleep, characterized by typical EEG delta activity, does not correspond to a state in which cortical neurons are inactive. On the contrary, in this sleep state cortical neurons can display mean rates of firing similar to those during wakefulness and/or REM sleep. Regarding the neuronal firing patterns, the main difference between delta sleep and wakefulness and REM sleep is that in the former the neurons tend to display long bursts of spikes with relatively prolonged interburst periods of silence, whereas in the latter the firing pattern is more continuous. The functional meaning of these peculiar firing pattern of delta sleep has not been determined.

In general, EEG signals covary strongly with different levels of arousal and consciousness. The changes of EEG with increasing levels of anesthesia are typical examples of this property.

B. Alpha Rhythms of Thalamus and Neocortex

Alpha rhythms recorded from the occipital areas occur in relaxed awake animals and show a typical reactivity to eye closure. Although the frequency range of alpha rhythms overlaps that of sleep spindles, these two phenomena differ in many aspects. Namely, the behavioral state in which both types of oscillations occur is quite different, and their distribution over the cortex and thalamus also differs considerably. The basic mechanisms responsible for alpha oscillations at the cellular level have not been described in detail due to the inherent difficulty in studying a phenomenon that by definition occurs in the state of relaxed wakefulness under conditions that allow measuring the underlying membrane currents under anesthesia, or, optimally, in slice preparations *in vitro*. To overcome this difficulty, some researchers assumed that spindles occurring under barbiturate anesthesia were analogous to alpha rhythms. However, this analogy was challenged on experimental grounds because a comparative investigation of alpha rhythms, obtained during restful awakeness at eye closure, and spindles induced by barbiturates, recorded from the same sites over the visual cortex and lateral geniculate nuclei in dog, presented differences in frequency, spindle duration, topographic distribution, and amount of coherence among different cortical and thalamic sites. Investigations using multiple electrode arrays placed on the cortical surface, depth intracortical profiles, and intrathalamic recordings from several thalamic nuclei elucidated many elementary properties of alpha rhythms:

- In the visual cortex, alpha waves are generated by a current dipole layer centered at the level of the somata and basal dendrites of the pyramidal neurons of layers IV and V.
- The coherence between alpha waves recorded from neighboring cortical sites is greater than any thalamocortical coherence.
- The influence of alpha signals recorded from the pulvinar on cortical rhythms can be conspicuously large, depending on cortical area, but intercortical factors play a significant role in establishing cortical domains of alpha activity.

These experimental findings led to the conclusion that in addition to the influence of some thalamic nuclei, mainly the pulvinar, on the generation of alpha rhythms in the visual cortex, there are systems of surface-parallel intracortical connections responsible for the propagation of alpha rhythms over the cortex. These oscillations appear to be generated in small patches of cortex that behave as epicenters, from which they propagate at relatively slow velocities (approximately 0.3 cm/sec). This type of spatial propagation has been confirmed, in general terms, by experimental and model studies. A comprehensive study of alpha rhythms in the visual cortex of the cat showed characteristics corresponding closely to those of alpha in man and dog. It was found that this rhythmic activity was localized to a limited part of the primary visual cortex area 18 and at the border between areas 17 and 18. In this context, additional insight into the sources of alpha rhythms in man has been obtained using MEGs integrated with anatomical information obtained from MRI. Different sources of alpha rhythms, the so-called "alphons," were found

concentrated mainly in the region around the calcarine fissure, with most sources located within 2 cm of the midline. A typical distribution of sources of alpha rhythmic activity over the human cortex was recently obtained using whole-head MEG and analyzed using the MUSIC algorithm as shown in Fig. 2.

In addition to the alpha rhythms of the visual cortex, rhythmic activities with about the same frequency range (in man, 8–13 Hz, in cat; 12–15 Hz) have been shown to occur in other cortical areas, namely in the somatosensory cortex (SI areas 1–3). These activities are known as "rolandic mu rhythms" or "wicket rhythms," named after the appearance of the records on the scalp in man, and they have a typical reactivity since they appear when the subject is at rest and they are blocked by movement. The mu rhythm is particularly pronounced in the hand area of the somatosensory cortex and it reacts typically to the movement of closing the fists. In the cat, there is no significant coherence between the mu rhythm of the SI cortex and the alpha rhythm of the visual cortex, which supports the general idea that these two types of rhythms are independent. Furthermore, mu rhythms of the SI area also differ from the alpha rhythms of the visual cortex recorded in the same animal in that the former have systematically higher frequencies than the latter, the difference being about 2 Hz. Mu rhythms were also recorded in thalamic nuclei, namely in the ventroposterior lateral nucleus. The mu rhythm has also been identified in MEG recordings over the rolandic sulcus, particularly over the somatomotor hand area. The reactivity of the EEG/MEG mu rhythm to movement and other behavioral conditions has been analyzed in detail in man using advanced computer analysis. In addition, another spontaneous MEG activity, the so-called tau rhythm, was detected over the auditory cortex. This rhythmic activity was reduced by sound stimuli. This MEG tau rhythm, first described by the group of Riitta Hari in Helsinki, is apparently similar to an EEG rhythm that was found using epidural electrodes over the midtemporal region by Ernst Niedermeyer in Baltimore.

C. EEG Activities of the Limbic Cortex

Two main types of rhythmic EEG activities can be recorded from limbic cortical areas characterized by different dominant rhythmic components, one in the theta and one in the beta/gamma frequency range. The former was first described by Green and Arduini in 1954, and in several species it may cover the frequency range between 4 and 12 Hz. It is common practice to call this activity theta rhythm since in most species it is dominant between 4 and 7.5 Hz; however, since in rodents it can extend to 12 Hz, it is preferentially called rhythmic slow activity (RSA). The brain areas in which RSA is most apparent are the hippocampus and parahippocampal gyri, although it also occurs in other parts of the limbic system. It has been questioned whether RSA also occurs in humans. However, RSA was incidentally recorded in the human hippocampus and was clearly demonstrated in the hippocampus of freely moving epileptic patients using spectral analysis. The relative difficulty in recording RSA in the human hippocampus may be related to the decrease in amplitude and regularity of RSA encountered in higher primates. Therefore, it can be concluded that the human hippocampus is no exception among mammals with respect to the occurrence of RSA. Using MEG imaging in normal human subjects, Claudia Tesche showed that the activity centered on the anterior hippocampus consisted of spectral components lower than 12 Hz that included task-dependent peaks.

It is generally accepted that hippocampal RSA depends on intact septo-hippocampal circuits. The experimental evidence is based on the fact that lesions of the septal area result in the disappearance of RSA from the hippocampus and other limbic cortical areas. Furthermore, it was shown that subpopulations of medial septal/diagonal band neurons, particularly those lying within the dorsal limb of the diagonal band and in the ventral part of the medial septal nucleus, discharge in phase with hippocampal RSA. This population of septal neurons is capable of sustaining burst firing within the RSA frequency range even when disconnected from the hippocampus. These experimental findings have led to the general idea that in the septal area some neuronal networks can work as pacemakers of RSA. However, it must be emphasized that the hippocampal neurons do not act as simple passive followers of the septal neurons. The neuronal networks of the hippocampus and of other cortical limbic areas also contribute to the characteristics of the local RSA. Indeed, even hippocampal slices kept *in vitro* are able to generate RSA-like rhythmic field potentials on the application of the cholinergic agonist carbachol in a large concentration. Theta frequency oscillations were also recorded in hippocampal neurons, maintained in *in vitro* slices, induced by depolarization by current injection of long duration. This cholinergic activation is mediated by muscarinic receptors since it is blocked by atropine, and RSA

elicited in hippocampal slices involves recurrent excitatory circuits of the CA3 region, mediated by glutamatergic (non-NMDA) synapses. It should be emphasized that this form of *in vitro* RSA mimicks the so-called atropine-sensitive RSA of behaving rats that appears during motionless behavior and REM sleep. However, the atropine-insensitive RSA that appears during motor activity certainly involves other processes and depends on other neuromodulating systems.

In addition to the rhythmic activity in the theta frequency range, one can also record from the paleo- and archicortex other EEG activities within the frequency range from 30 to 50 Hz (i.e., beta or gamma activities). An interesting feature of these rhythmic activities of the basal forebrain is that they have a relatively large value of coherence over a wide range of brain areas involved in olfaction. This coherent domain of beta activity reflects the fact that there is a rostrocaudal spread of this activity along the olfactory bulb, prepyriform, entorhinal cortices.

The cellular origin of the beta activity of the olfactory areas (paleocortex) has been revealed by many comprehensive investigations in which both single and multiple neuronal activity or LFPs were recorded, and their relationships were analyzed using signal analysis techniques by Walter Freeman. A conclusion of Freeman's studies is that the interactions at the neuronal population level are mediated by synaptic somadendritic mechanisms, involving electrotonic spread of activity and the transmission of action potentials, that have nonlinear dynamical behavior. It is also assumed that the oscillations result primarily from synaptic interactions among interconnected populations of neurons and not from intrinsic neuronal oscillations since the paleocortex isolated from extrinsic connections remains silent.

It is noteworthy that computer simulation studies of Walter Freeman and collaborators showed that the typical complex nonlinear (possibly chaotic) dynamics of the olfactory cortical areas appear to depend on the presence of interactions between the olfactory bulb, the anterior olfactory nucleus, and the prepyriform cortex, which are interconnected by short and long pathways with the corresponding time delays. This implies that the dynamical behavior of neuronal networks should not be assumed to be determined only by local properties but also to depend on the interactions between networks at different locations, through reentrant circuits, such that new emerging properties arise from these functional assemblies of corticocortical and subcorticocortical systems.

D. Beta/Gamma Activity of the Neocortex

The identification and characterization of high-frequency rhythms in the neocortex has been concentrated mainly in two neocortical areas—the visual cortex and the somatomotor cortex. Here, some of the properties of these beta/gamma rhythmic activities for these two areas are discussed.

Commonly, the EEG or LFP of the visual cortex is associated with the alpha rhythm, with typical reactivity with closing and opening of the eyes as described previously. However, other types of rhythmic activities can be present in the same cortical areas, namely within the beta frequency range. In the dog, it was shown that the EEG spectral density was characterized by peaks within the beta/gamma frequency range while the animal was looking attentively at a visual stimulus. These findings in the awake animal are in line with the demonstration in the bulbospinal transected preparation that brain stem electrical stimulation causes not only desynchronization of alpha spindles but also the appearance of fast rhythms in the cortical EEG. Recently, Walter Freeman and Bob van Dijk found in the visual cortex of a rhesus monkey that fast EEG rhythms (spectral peak of 30 ± 3.7 Hz) occurred during a conditioned task to a visual stimulus. A possibly related finding is the discovery by the group of Charles Gray and Wolf Singer and by Eckhorn and collaborators of oscillations within the beta/gamma frequency range (most commonly between 30 and 60 Hz) in the firing of individual neurons of the visual cortex in response to moving light bars. Using auto- and cross-correlation analyses, it was demonstrated that neurons tended to fire in synchrony, in an oscillatory mode, within cortical patches that could extend up to distances of about 7 mm. The oscillations in neuronal firing rate were correlated with those of the LFPs. The cortical oscillations are modulated by the activation of the mesencephalic reticular formation (MRF), but the stimulation of the MRF alone does not change the pattern of firing of the cortical neurons. However, MRF stimulation increases the amplitude and coherence of both the LFP and the multiunit responses when applied jointly with a visual stimulus.

In the somatomotor cortex, beta/gamma oscillations of both neuronal firing and LFPs were described in the awake cat by the group of Rougeul-Buser, particularly when the animal was in a state of enhanced vigilance while watching an unreachable mouse. Also, in monkey during a state of enhanced attention, fast oscillations were found in the somatomotor cortex. Oscillations of 25–35 Hz occurred in the sensorimotor

cortex of awake behaving monkeys both in LFPs and in single- /multi-unit recordings. They were particularly apparent during the performance of motor tasks that required fine finger movements and focused attention. These oscillations were coherent over cortical patches extending at least up to 14 mm that included the cortical representation of the arm. Synchronous oscillations were also found straddling the central sulcus, so they may reflect the integration of sensory and motor processes. The LFP reversed polarity at about 800 μm under the cortical surface, indicating that the source of the LFP is in the superficial cortical layers. It is noteworthy that at least some of the cortical beta/gamma rhythmic activities appear to depend on projecting dopaminergic fibers arising in the ventral tegmental area, but it is not clear to what extent the beta rhythms of the somatomotor cortex are related to thalamic or other subcortical activities.

It is relevant to correlate the characteristics of EEG beta/gamma activities found in experimental animals with those recorded from the scalp in man. Beta/gamma activity was reported by DeFrance and Sheer to occur over the parieto-temporo-occipital cortex in man, particularly in relation to the performance of motor tasks.

With respect to the origin of beta/gamma rhythmic activity, several experimental facts have led to the interpretation that these rhythmic activities are primarily generated in the cortex: (i) the fact that oscillations in the beta/gamma frequency range were easily recorded from different cortical sites but not from simultaneously obtained recordings from thalamic electrodes; (ii) the observation that in the visual cortex there are neurons that show oscillatory firing rates with a phase difference of about one-fourth of a cycle, indicating that a local recurrent feedback circuit can be responsible for the oscillations; and (iii) the finding of intrinsic oscillations in cortical neurons from layer IV of the frontal cortex of guinea pig *in vitro*. Nevertheless, it is possible that thalamic neuronal networks also contribute to the cortical beta/gamma rhythmic activity since about 40-Hz oscillatory behavior has been observed by Mircea Steriade and collaborators in neurons of the intralaminar centrolateral nucleus that projects widely to the cerebral cortex. The question cannot be stated as a simple alternative between a cortical versus a thalamic rhythmic process, both considered as exclusive mechanisms. As discussed in relation to other rhythmic activities of the mammalian brain, both network and membrane intrinsic properties cooperate in shaping the behavior of the population, including its rhythmic properties and its capability of synchronizing the neuronal elements.

VI. EVENT-RELATED PHENOMENA: EEG DESYNCHRONIZATION AND SYNCHRONIZATION

The electrical activity of the brain is ever changing, depending on both exogenous and endogenous factors. The spatiotemporal patterns of EEG/MEG activity reflect changes in brain functional states. These patterns are apparent as a sequence of maps of scalp activity that change at relatively short time intervals, on the order of seconds or even fractions of seconds. Dietrich Lehmann identified these series of maps as reflecting brain "microstates" and proposed that different modes of mentation are associated with different brain EEG microstates.

A classic phenomenon is the EEG/MEG activity that occurs in response to sensory events, the so-called sensory evoked potentials (EPs). In addition, another class of EEG phenomena of the same kind should be distinguished that consists of those changes of the ongoing EEG activity that precede motor acts, such as the readiness potential, or "bereitschafspotential" of Kornhuber and Deecke, and the premotor potentials. In order to detect EPs or, in more general terms, event-related potentials (ERPs), computer-averaging techniques are commonly used. The basic model underlying this approach is that the evoked activity is time and phase locked to a given event, or stimulus, while the ongoing EEG activity behaves as additive noise. Here, I do not consider this kind of phenomena in detail since this forms a specialized EEG field. Nevertheless, it should be noted that the discovery of evoked activity related to cognitive events has resulted in important contributions to neurocognitive research. One example is the contingent negative variation (CNV), first described by Grey Walter in 1964. The CNV is a slow potential shift with negative polarity at the cortical surface that precedes an expected stimulus and that is related to motivation and attention. Another example is a component of EPs that peaks at approximately 300 msec, with surface positivity, after infrequent but task-relevant stimuli that was discovered by Sutton, Zubin, and John in 1965. This component is called the P300, and it depends more on the meaning of the stimulus and the context of the task than on the physical properties of the stimulus. Still another EP phenomenon with relevant cognitive connotations is the so-called processing negativity

described by Näätänen that is a large surface negative wave that can begin as early as 60 msec and can last for 500 msec. It is a sign of selective attention. Neurocognitive studies using ERPs and, recently, event-related magnetic fields have been successfully carried out. Such investigations have benefited much from the approach developed by Alan Gevins and collaborators, the so-called EP covariance methodology, that has provided interesting results concerning the cortical processes involved in working memory and in planning of movement. In these studies, the recording of EPs at different brain sites during the sequential processing of cognitive tasks allowed researchers to follow sequential and/or parallel activation of different cortical areas as cognitive tasks evolved. This approach has been particularly successful in studies of brain processes underlying language functions, such as the seminal investigation of Riitta Salmelin and collaborators in Helsinki. Using a whole-head MEG and event-related magnetic fields, they were able to trace the progression of brain activity related to picture naming from Wernicke's area to the parietal-temporal and frontal areas of the cortex. Subsequent MEG studies of the same group in collaboration with that of Pim Levelt of Nijmegen revealed a more refined pattern of the dynamics of cortical activation associated with the successive stages of a psychological model of spoken word generation. These neurocognitive processes were approached using the novel methodology of combining fMRI and MEG in order to obtain high-resolution imaging, both in space and in time, of cortical activity during semantic processing of visually presented words. These studies confirmed that in general there is a wave of activity that spreads from the occipital cortex to parietal, temporal, and frontal areas within 185 msec during picture naming. Furthermore, they indicated that the effects of word repetition are widespread and occur only after the initial activation of the cortical network. This provides evidence for the involvement of feedback mechanisms in repetition priming (Fig. 5).

Since the focus of this article is the ongoing EEG/MEG activity, I consider in more detail a class of EEG/MEG phenomena that are time locked to an event but

Neuron
62

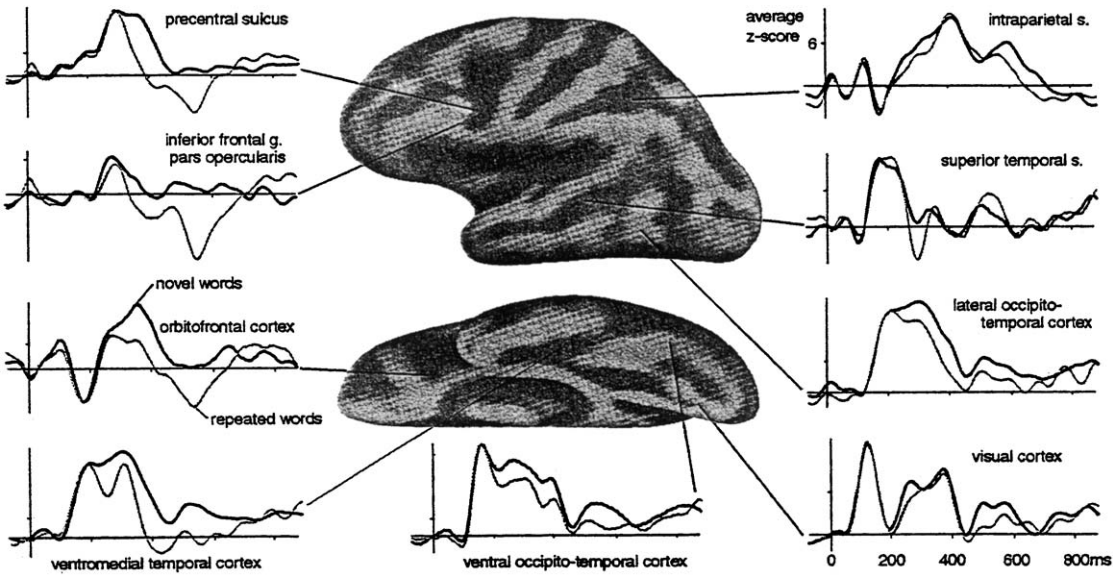


Figure 5 Estimated time courses of MEG signals corresponding to different cortical areas involved in processing of novel and repeated words. The MEG sources were estimated using noise normalization and were fMRI biased toward hemodynamically active cortical areas. The results represent averages across four subjects. Black lines show responses to novel words, and gray lines indicate the responses to repeated words. Waveforms are derived from single cortical locations (each representing 0.5 cm^2). Vertically, the z scores are shown: a score of 6 corresponds to a significance level of $p < 10^{-8}$ (reproduced with permission from Dale *et al.*, 2000).

not phase locked. These phenomena cannot be extracted by simple averaging but may be detected by spectral analysis. An event-related phenomenon may consist of either a decrease or an increase in synchrony of the underlying neuronal populations. The former is called event-related desynchronization (ERD), and the latter is called event-related synchronization (ERS) (Fig. 6). Gert Pfurtscheller and collaborators extensively studied these phenomena. When referring to ERD or ERS, one should specify the corresponding frequency band since, for example, there may be ERD of the alpha band and ERS of the beta band at the same time. The term ERD is only meaningful if the EEG activity during the baseline condition shows a clear spectral peak at the frequency band of interest, indicating the existence of a specific rhythmic activity. Complementarily, the term ERS has meaning only if the event results in the emergence of a rhythmic component, and therewith of a spectral peak that was not detectable under baseline conditions. In general, ERD/ERS reflect changes in the activity of neuronal networks that take place under the influence of specific and/or modulating inputs, which alter the parameters controlling the oscillatory behavior of the neuronal networks.

Even within the alpha frequency band, ERD is not a unitary phenomenon since we have to distinguish at least two patterns of alpha ERD: Lower alpha (7–10 Hz) desynchronization is found in response to almost any kind of task and appears to depend mainly on task complexity. Thus, it is unspecific and tends to be topographically widespread over the scalp. Upper alpha (10–12 Hz) desynchronization has a more restricted topographic distribution, particularly over the parieto-occipital areas, and it is most often elicited by events related to the processing of sensorisemantic information, as shown by the investigations of Wolfgang Klimesch. In general, many psychophysiological variables that cause ERD of rhythms within the alpha frequency range are related to perceptual and memory tasks, on the one hand, and voluntary motor actions, on the other hand. Voluntary movements can result in

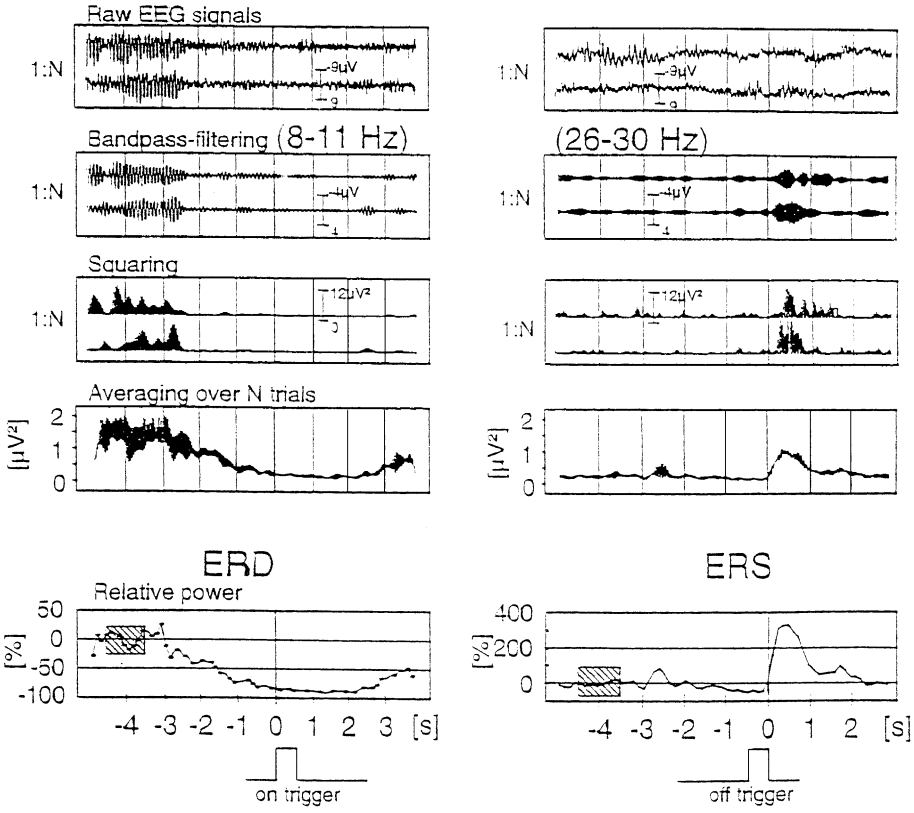


Figure 6 Principle of ERD (left) and ERS (right) EEG processing. A decrease of power within a given band (8–11 Hz) indicates ERD and an increase of band power (26–30 Hz) indicates ERS. Note that ERD precedes the trigger, a finger movement, and that ERS follows the trigger (i.e., it occurs at the cessation of the movement) (adapted with permission from Pfurtscheler and Lopes da Silva, 1999).

an ERD of the upper alpha (also called mu rhythm) and lower beta bands localized close to the sensorimotor areas. This desynchronization starts about 2 sec prior to movement onset, over the contralateral rolandic region in the case of a unilateral movement, and becomes bilaterally symmetrical immediately before movement execution. It is interesting to note that the ERDs for the different frequency bands have specific topographical distributions, indicating that different cortical populations are involved. For instance, in relation to a hand movement, the 10- to 11-Hz mu ERD and the 20- to 24-Hz beta ERD display different maxima over the scalp, although both activities are localized around the central sulcus. The mu rhythm ERD has maximal magnitude more posteriorly than the beta activity, indicating that it is generated mainly in the postrolandic somatosensory cortex, whereas the low beta activity is preferentially generated in the prerolandic motor area. In addition, after a voluntary movement the central region exhibits a localized beta ERS that becomes evident in the first second after cessation of the movement, at a time when the rolandic mu rhythm still presents a desynchronized pattern. The exact frequency of this rebound beta ERS can vary considerably with subject and type of movement. This beta ERS is observed not only after a real movement as shown in Fig. 6 but also after an imagined movement. Furthermore, ERS in the gamma frequency band (approximately around 36–40 Hz) can also be found over the central egions during the execution of a movement, in contrast with the beta ERS that has its maximum after the termination of the movement. A prerequisite for detecting this gamma ERS is that alpha ERD takes place at the same time.

A particular feature of ERD/ERS phenomena that we have recently analyzed is that under some conditions one can find a localized ERD at the same time as ERS in a neighboring region. This antagonistic ERD/ERS phenomenon can occur between two different modalities—for example, ERD of the alpha over the occipital region elicited by visual stimulation accompanied by ERS of the mu rhythm of the central somatosensory region—but it can also occur within the same modality. For example, a voluntary hand movement can result in an ERD over the cortical area representing the hand and simultaneously an ERS over the cortical area representing the leg/foot (Fig. 7). The opposite can be seen in the case of a voluntary foot movement. We interpret this ERD/ERS antagonistic phenomenon as indicating that at the level of the thalamic reticular nucleus, cross talk between the neuronal networks processing different inputs takes place. This may occur as follows: The specific movement would engage a focal attentional process that results in a desynchronization of a module of thalamocortical networks called the target module. This attentional signal is most likely mediated by the activation of modulating cholinergic inputs. These cholinergic inputs hyperpolarize the inhibitory neurons of the reticular neurons and depolarize the thalamocortical relay neurons of a given module. Consequently, the thalamocortical feedback loop responsible for the rhythmic activity becomes open, which is reflected in the ERD that is recorded over the corresponding cortical projection areas. At the same time, the neurons of the reticular nucleus that are adjacent to those of the target module become disinhibited. This results in an increase in the gain of the feedback loops to which the latter belong, thus resulting in an increase in the magnitude of the corresponding rhythmic activity that is reflected at the cortex by ERD. In this way, the analysis of ERD/ERS phenomena has led to the formulation of a hypothesis concerning the neurophysiological processes underlying the psychological phenomenon of focal attention/surround inhibition.

In short, ERD can be interpreted as an electrophysiolgical correlate of activated cortical areas involved in the processing of sensory, motor, or cognitive information. The mirror image of alpha ERD, of course, is alpha ERS (i.e., a pronounced rhythmic activity within the alpha frequency range), indicating that the corresponding neuronal networks are in a state of reduced activity. Thus, these rhythmic activities are sometimes called ''idling rhythms,'' although one should be cautious about the literal interpretation of this term since the underlying neuronal populations are not really ''idle''— they are always active but may display different dynamical properties. An important point is that ERD and ERS cannot be considered as global properties of the brain. Indeed, ERD and ERS phenomena can be found to coexist in neighboring areas and may affect specific EEG/MEG frequency components differently.

Understanding the significance of ERS of the beta frequency range that typically occurs after a movement has been aided by the observation that at the time that this form of ERS occurs the excitability of the corticospinal pathways decreases, as revealed by transcranial magnetic stimulation. This supports the hypothesis that the postmovement beta ERS corresponds to a deactivated state of the motor cortex. In contrast, the ERS in the gamma frequency band appears to reflect a state of active information

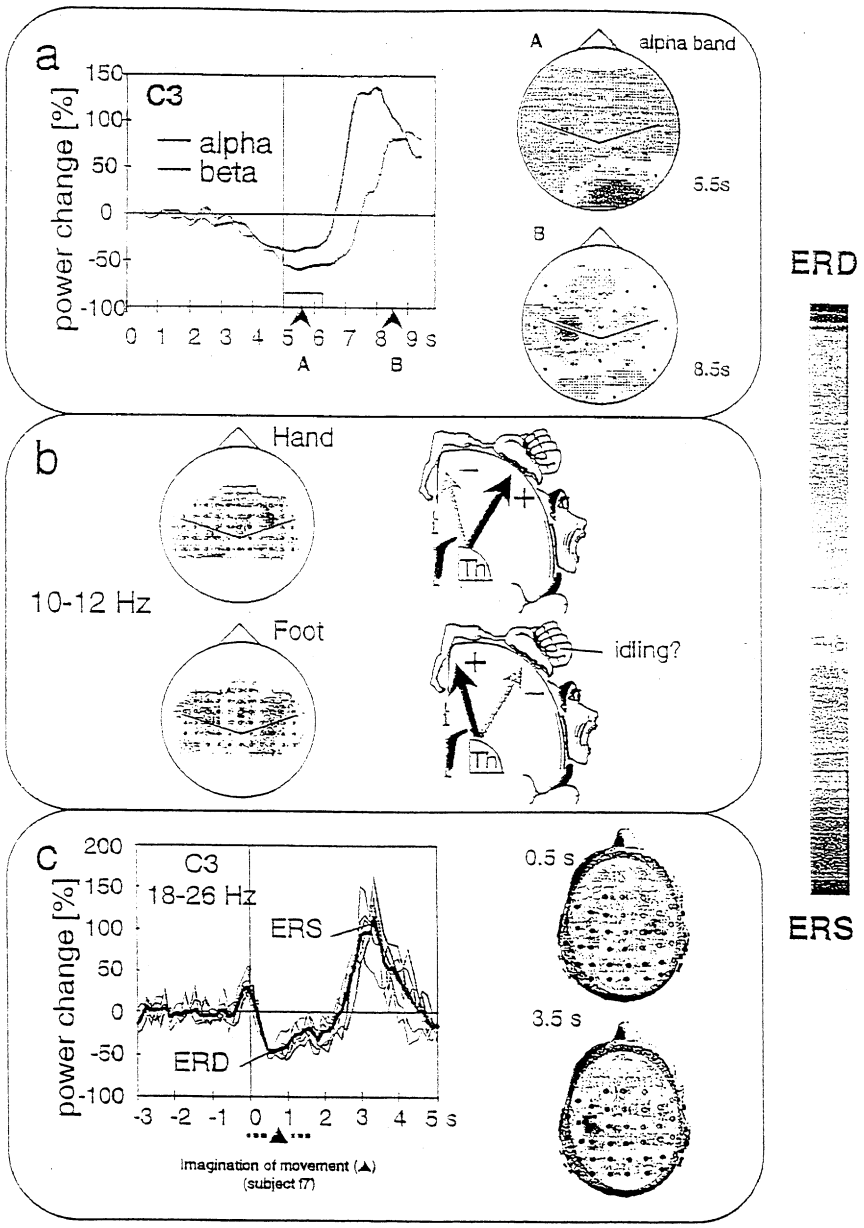


Figure 7 Event-related desynchronization (ERD) and event-related synchronization (ERS) in relation to movement. (a) Average ($n = 9$) of ERD curves calculated in the alpha and beta bands for a right-hand movement task; recording from C3 (left). The maps were calculated for a 125-msec interval during movement (A) and after movement offset in the recovery period (B). (b) Maps displaying ERD and ERS for an interval of 125 msec during voluntary movement of the hand and movement of the foot. The motor homunculus model is shown on the right-hand side to give an indication of the localization of the hand and the foot cortical areas. (c) (Left) Superimposed ERD curves and beta ERS rebound from eight sessions with right-hand motor imagery in one subject within the frequency band 18–26 Hz and for electrode C3. The average is superimposed. (Right) ERD maps displaying simultaneously occurring ERD (contralateral) and ERS (ipsilateral) during imagery and ERS (contralateral) after motor imagery. Color code: dark areas indicate power decrease (ERD) and light areas power increase (ERS) (reproduced with permission from Pfurtscheller and Lopes da Silva, 1999).

processing. Indeed, recordings by Bressler and collaborators from the monkey motor cortex during the performance of a visual-guided motor task showed increased neuronal activity at relatively high frequencies that corresponds in time to the gamma ERS found in human. Furthermore, these gamma band activities recorded from the striate and motor cortex were correlated when an appropriate motor response was

made but were uncorrelated when no response occurred. Similarly, the group of Pfurtscheller found that there was an increase in coherence between the EEG recorded from the sensorimotor and supplementary motor areas over one hemisphere, within the gamma range, during the performance of contralateral finger movements.

VII. WHAT ARE THE ROLES OF OSCILLATIONS IN THE PROCESSING OF NEURAL INFORMATION?

This question is sometimes answered in a negative way with the implication that brain oscillations are irrelevant for the functioning of the brain and should be considered simple epiphenomena. This point of view, however, is superficial since is does not take into account what is happening at the neuronal level during EEG oscillations. In order to answer the question, it is necessary to identify the specific neuronal processes underlying such oscillations. Only in this way we may reach an understanding of the functional implications of an EEG oscillation. In other words, we have to analyze the state of a neuronal population that displays an oscillatory mode in terms of the membrane potentials of the neurons that belong to the population. Of course, during an oscillation the membrane potentials of the constituting neurons vary in synchrony. Since the membrane potential controls the transfer of information of a neuron, we must analyze what happens to the membrane potential of the main neuronal population (i.e., the collective membrane potential) during specific oscillatory modes. Two essentially opposite states of the collective membrane potential may occur during an oscillatory state: The membrane potential of the main population changes in either a hyperpolarized or a depolarized direction with respect to the resting membrane potential.

The former case occurs during a sleep spindle or during a burst of alpha or mu waves since in these circumstances the mean membrane potential of the thalamocortical neurons is in a hyperpolarized state. The main neuronal population displays phased inhibitory potentials with only a few occasional action potentials. Consequently, the functional unit formed by this population is in an inhibitory mode and the transfer of information is blocked. In other words, a gate is closed for the transfer of information. Thus, the oscillatory mode of activity characterized by a dominant alpha frequency represents a ''closed gate'' functional state. In cases in which the mean membrane potential of the main neuronal population is hyperpolarized at still deeper levels, the population will present oscillations at lower frequency in the delta or ultra-delta frequency range, as it occurs during deep sleep.

In the opposite case, the mean membrane potential of the main neuronal population is displaced in a depolarized direction and it shows oscillations phased with the occurrence of action potentials, sometimes in the form of bursts. The implication is that the probability that series of action potentials of different sets of neurons are time locked increases. This can be the basic mechanism ensuring that neuronal firing becomes synchronized (i.e., that binding occurs). This form of oscillation is usually manifest at relatively high EEG frequencies in the beta/gamma band. Note that this does not imply that oscillations at about the same frequency will have the same functional connotation at the neuronal level. This can be illustrated with two examples that show that different types of beta oscillations may correspond to different functional states of corticospinal pathways. In one case, the beta burst that occurs over the central areas of the scalp at the end of a hand movement corresponds to a state of decreased excitability of the corticospinal pathways. In contrast, the MEG beta activity that occurs over the motor cortex during isometric contractions of the muscles of extremity is directly related to the EMG of the corresponding muscles. A common denominator of both cases, however, is that during the oscillatory mode the main neuronal population displays coherent activity. We may hypothesize that this coherent activity serves a goal in binding the neurons within the underlying population so that they may form a functional unit. The ultimate behavioral result corresponding to the EEG oscillatory mode will depend not only on the state of the membrane potentials of the neurons within the population but also on their connectivity with other networks. The complexity of the circuits that are involved in this process makes it difficult to assign a well-defined functional state to a given EEG/MEG oscillatory mode without precise knowledge of the related processes at the level of neurons and of interconnections between subsystems.

Nevertheless, two main functional connotations of brain oscillations can be put forward: (i) binding or linking of single neurons within a population such that they can form a functional unit of information processing, and (ii) gating the flow of information through given circuits. In addition, we should also consider the possibility that oscillatory modes of activity may contribute to other aspects of brain functions. One aspect is that neuronal networks in

general have frequency-dependent properties (e.g., they may show resonance behavior or frequency selectivity). This implies that the transfer of information between two neuronal networks coupled by way of anatomical pathways is likely to depend on the frequency of the sets of action potentials that are transferred between both. This means that oscillatory modes of activity can have another function, namely that of matching, which is the transfer of information between networks from distinct but related areas such that this transfer may be facilitated or optimized.

In this context, it has been demonstrated that in the hippocampus during the theta rhythmic mode of activity, the transmission of impulses through the synaptic path from the CA3 to the CA1 region is reduced. During the theta mode of activity, impulses arising in the entorhinal cortex may pass the two first synapses of the trisynaptic pathway (i.e., to the dentate gyrus and CA3 area) but cannot reach the CA1/subiculum areas. As a consequence, they will be rerouted and will be lead from the CA3 area to target structures of the forebrain, such as the septal area, instead of back to the entorhinal cortex through the subiculum.

Another functional implication of brain oscillatory modes of activity derives from the observation that the strength of synaptic transfer can change, either in the sense of potentiation or depression, depending on frequency of stimulation. This has been evidenced mainly in the case of the theta rhythm of the limbic cortex, particularly of the hippocampus, since it was shown that electrical stimulation at the theta frequency can induce long-term potentiation, whereas stimulation at much lower frequencies may lead to long-term depression. This implies that brain oscillatory activities may be associated with modulating processes of synaptic transmission and plasticity, even promoting the latter in specific synaptic systems.

One general principle of the functional organization of the brain is that it must keep track of different flows of information in the appropriate circuits while it establishes dynamical associations between several sets of processes through multiple reentrant pathways. This means that multiple neuronal networks, at both the cortical and the subcortical levels, are simultaneously active (i.e., that most cognitive processes involve the parallel activity of multiple cortical areas). In this way, different rhythmic activities may be present simultaneously in distinct brain areas or systems.

Another general principle is that processing of information takes place in populations of neurons organized in networks or assemblies of neurons and not just in single neurons. In this respect, evidence obtained recently from studies in which recordings of multiple neurons and LFPs in the visual cortex were obtained simultaneously is strongly indicative that coherent oscillations in a population of neurons could be the basic mechanism to ensure feature binding in the visual system. To underscore this finding, the concept of the linking or association field has been introduced, which is the equivalent—at the level of a neuronal population—of the classic receptive field of the single neuron. The linking field represents the area in visual space in which a stimulus can induce synchronized oscillations in assemblies of cortical neurons. In general, synchronized oscillations occur in different cortical populations if they have similar, or overlapping, receptive field properties and if the appropriate stimulus is present. Therefore, stimulus-induced synchronized oscillations in a given neuronal network may be the basic unit from which attentive percepts that require iterative processing between different neuronal groups are formed.

A third principle of organization, mentioned previously, is that the functional assemblies are interconnected by multiple reentrant connections. The experimental finding that the rhythmic firing of most neurons in an assembly tend to oscillate with zero-degree phase shift results from this property of multiple-level feedback and feedforward.

In summary, note that the change in the activity of a neuronal population from a random mode of activity to an oscillatory mode yields two important consequences for the functioning of the corresponding systems. First, it changes the value of the mean membrane potential in a relatively large group of neurons simultaneously: thus, it sets a bias on the activity of a neuronal population. Accordingly, the transfer function of the population can change in a dynamic way. It appears that the occurrence of synchronized oscillatory activities provides an efficient way to switch the behavior (i.e., to cause a qualitative bifurcation within an assembly of neurons) between different modes of information processing. In this way, neuronal groups with a similar dynamical functional state (linking fields) can be formed. Second, the oscillatory activity may have a frequency-specific role. This role is evidenced by two types of phenomena: (i) matching between interconnected neuronal networks in order to facilitate the transfer of information and (ii) modulation of synaptic plasticity. It is likely that this optimal transfer of information takes place bidirectionally between any two populations of neurons,

ensuring what Gearld Edelman called the reentry process. In this way, large groups of neurons can interact optimally. We may speculate that these processes of facilitating the transfer of information between populations of neurons distributed over different brain areas, such that they can form functional units (although with distinct patterns of local activity), could constitute the physiological basis of consciousness. Under specific conditions consciousness would be impaired, as occurs when these brain areas are recruited into a generalized oscillatory mode, such as the 3-Hz oscillations typical of an epileptic absence seizure.

Finally, the fact that neuronal networks tend to oscillate in a synchronized way, depending on local circuit properties and intrinsic membrane mechanisms, is gradually receiving more attention from physiologists and theoreticians. This is important since it is becoming apparent that the functional units of the brain, by means of which information is processed and transmitted, are dynamical assemblies of neurons. It is indeed necessary to determine how such functional networks are formed in order to be able to understand how the brain subserves cognitive functions. To reach these goals, it is clear that new techniques are necessary that combine the analysis of elementary physiological properties of neurons, *in vitro* and/or *in vivo*, with the study of groups of neurons working as dynamical systems in the intact brain.

VIII. CONCLUSIONS

Knowledge of the electrical and magnetic fields generated by local neuronal networks is of interest to neuroscientists because these signals can give relevant information about the mode of activity of neuronal populations. This is particularly relevant to understanding high-order brain functions, such as perception, action programming, and memory trace formation, because it is becoming increasingly clear that these functions are subserved by dynamical assemblies of neurons. In this respect, knowledge of the properties of the individual neurons is not sufficient. It is necessary to understand how populations of neurons interact and undergo self-organization processes to form dynamical assemblies. The latter constitute the functional substrate of complex brain functions. These neuronal assemblies generate patterns of dendritic currents and action potentials, but these patterns are usually difficult to evaluate experimentally due to the multitude of parameters and the complexity of the structures. Nevertheless, the concerted action of these assemblies can also be revealed in the local field potentials that may be recorded at the distance of the generators. However, extracting information from local field potentials about the functional state of a local neuronal network poses many nontrivial problems that have to be solved by combining anatomical/physiological with biophysical/mathematical concepts and tools. Indeed, given a certain local field potential, it is not possible to precisely reconstruct the behavior of the underlying neuronal elements since this inverse problem does not have a unique solution. Therefore, it is necessary to assume specific models of the neuronal elements and their interactions in dynamical assemblies in order to make sense of the local field potentials. This implies that it is necessary to construct models that incorporate knowledge about cellular/membrane properties with that of the local circuits, their spatial organization, and how they are modulated by different mechanisms.

I have presented many arguments and suggestions in line with the concept that the synchronized activity of populations of neurons in general, and the occurrence of rhythmic oscillatory behavior of different kinds in particular, is not an epiphenomenon of the functional organization of neuronal networks; rather, brain oscillatory activities can play essential functional roles within the brain.

See Also the Following Articles

ACTION POTENTIAL • CIRCADIAN RHYTHMS • ELECTROENCEPHALOGRAPHY (EEG) • EVENT-RELATED ELECTROMAGNETIC RESPONSES • GABA • ION CHANNELS • LIMBIC SYSTEM • NEOCORTEX • NEURAL NETWORKS • SLEEP DISORDERS • THALAMUS AND THALAMIC DAMAGE

Suggested Reading

Basar, E., and Bullock, T. H. (Eds.) (1992). *Induced Rhythms in the Brain*. Birkhäuser, Boston.

Dale, A. M., Liu, A. K., Fischl, B. R., Buckner, R. L., Belliveau, J. W., Lewine, J. D., and Halgren, E. (2000). Dynamic statistical parametric mapping: Combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron* **26,** 55–67.

Gray, C. M. (1999). The temporal correlation hypothesis of visual integration: Still alive and well. *Neuron* **24,** 31–47.

Hari, R. (1993). Magnetoencephalography as a tool of clinical neurophysiology. In *Electroencephalography. Basic Principles, Clinical Applications, and Related Fields* (E. Niedermeyer and

F. H. Lopes da Silva, Eds.), pp. 1035–1061. Williams & Wilkins, Baltimore.

Lamme, V. A., Supèr, H., and Spekreijse, H. (1998). Feedforward, horizontal, and feedback processing in the visual cortex. *Curr. Opin. Neurobiol*. **8**(4), 529–535.

Lehnertz, K., Arnhold, J., Grassberger, P., and Elger, C. E.*Chaos in Brain?* World Scientific, London.

Lopes da Silva, F. H. (1991). Neural mechanisms underlying brain waves: From neural membrane to networks. *Electroencephalogr. Clin. Neurophysiol*. **79,** 81–93.

Niedermeyer, E., and Lopes da Silva, F. H. (Eds.) (1999). *Electroencephalography. Basic Principles, Clinical Applications, and Related Fields*, 4th ed. Williams & Wilkins, Baltimore.

Nunez, P. L. (1995). *Neocortical Dynamics and Human EEG Rhythms*. Oxford Univ. Press, New York.

Pfurtscheller, G., and Lopes da Silva, F. H. (1999). Event-related EEG/MEG synchronization and desynchronoization: Basic principles. *Clin. Neurophysiol*. **110,** 1842–1857.

Regan, D. (1989). *Human Brain Electrophysiology*. Elsevier, New York.

Singer, W. (1989). Neuronal synchrony: A versatile code for the definition of relations? *Neuron* **24,** 49–65.

Steriade, M. (1999). Coherent oscillations and short-term plasticity in corticothalamic networks. *Trends Neurosci*. **22,** 337–345.

Steriade, M., Jones, E. G., and Llinás, R. R. (1990). Thalamic Oscillations and Signaling. Wiley–Interscience, New York.

Electroencephalography (EEG)

PAUL L. NUNEZ
Tulane University

GLOSSARY

cerebral cortex The outer layer of mammalian brains; in humans, approximately 2–5 mm thick with a surface area of approximately 1500–3000 cm^2.

coherence A squared correlation coefficient measuring the phase consistency between two signals, expressed as a function of signal frequency.

conductivity (ohm^{-1} mm^{-1}) The property of a material (e.g., living tissue) that determines the ease with which charges move through the medium to produce current.

current density ($\mu A/mm^2$) The flux of positive plus negative charge passing through a cross-sectional area.

electric potential (μV) The negative gradient of the local electric field vector.

electrocorticogram Electric potential recorded directly from the brain surface.

electroencephalography Electric potential generated by brain tissue, recorded from locations outside of nerve cells, either from scalp or from inside the cranium.

event related potential Brain electric potential produced by a combination of sensory stimulus and cognitive task.

evoked potential Brain electric potential produced as the direct result of a sensory stimulus.

Ohm's law for a volume conductor Vector current density equals the product of conductivity and vector electric field. Ohm's law simplifies to the usual scalar expression, voltage change = current × resistance, for current flow confined to one spatial direction, for example, in electric circuit wires.

The first recordings of electrical activity [electroencephalography (EEG)] from the human scalp were obtained in the mid-1920s. EEG provides functional, as opposed to structural, brain information and has important applications in medicine and cognitive science. The field of EEG encompasses evoked and event-related potentials and spontaneous EEG. This article focuses on spontaneous EEG recorded from human scalp.

I. WINDOW ON THE MIND

Human electroencephalography (EEG) provides a convenient but often opaque "window on the mind," allowing observations of electrical processes near the brain surface. The outer brain layer is the cerebral cortex, believed to be largely responsible for our individual thoughts, emotions, and behavior. Cortical processes involve electrical signals that change over times in the 0.01-sec range. EEG is the only widely available technology with sufficient temporal resolution to follow these quick dynamic changes. On the other hand, EEG spatial resolution is poor relative to that of modern brain structural imaging methods—computed tomography, positron emission tomography, and magnetic resonance imaging (MRI). Each scalp electrode records electrical activity at large scales, measuring electric currents (or potentials) generated in cortical tissue containing approximately 30–500 million neurons.

Electrodes may be placed inside the skull to study nonhuman mammals or human epilepsy patients. Such intracranial recordings provide measures of cortical dynamics at several small scales, with the specific scale dependent on electrode size. Intracranial EEG is often uncorrelated or only weakly correlated with cognition and behavior. Human "mind measures" are more easily obtained at the large scale of scalp recordings. The technical and ethical limitations of human intracranial recording force emphasis on scalp recordings. Luckily, these large-scale estimates provide important measures of brain dysfunction for clinical work and cognition or behavior for basic scientific studies.

EEG monitors the state of consciousness of patients in clinical work or experimental subjects in basic research. Oscillations of scalp voltage provide a very limited but important part of the story of brain functioning. For example, states of deep sleep are associated with slower EEG oscillations of larger amplitude. More sophisticated signal analyses allow for identification of distinct sleep stages, depth of anesthesia, epileptic seizures, and connections to more detailed cognitive events. A summary of clinical and research EEG is provided in Fig. 1. The arrows indicate common relations between subfields. Numbers in the boxes indicate the following:

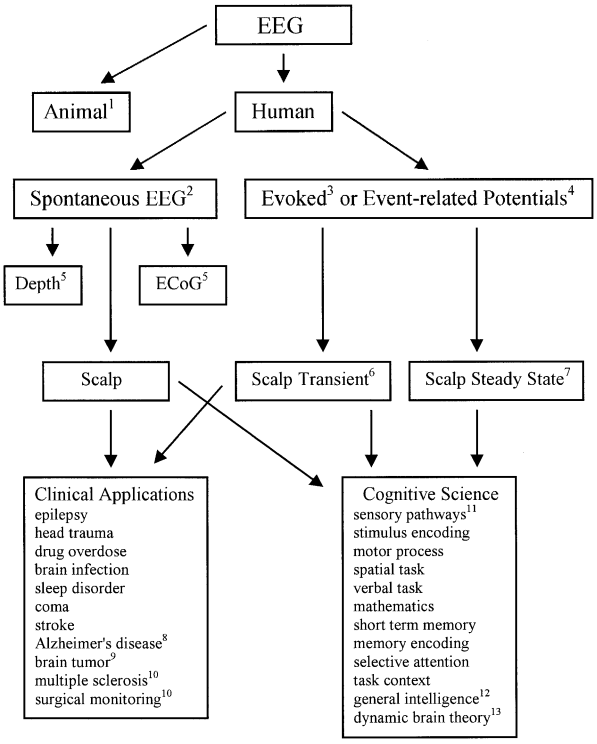


Figure 1 Common relationships between EEG subfields. Clinical applications are mostly related to neurological diseases. EEG research is carried out by neurologists, cognitive neuroscientists, physicists, and engineers who have a special interest in EEG.

1. Physiologists record EEG from inside skulls of animals using electrodes with diameters ranging from approximately 0.001 to 1 mm. Observed dynamic behavior generally depends on measurement scale, determined by electrode size for intracranial recordings. In contrast, scalp-recorded EEG dynamics is exclusively large scale and mostly independent of electrode size.
2. Human spontaneous EEG occurs in the absence of specific sensory stimuli but may easily be altered by such stimuli.
3. Averaged evoked potentials (EPs) are associated with specific sensory stimuli, such as repeated light flashes, auditory tones, finger pressure, or mild electric shocks. They are typically recorded by time averaging to remove effects of spontaneous EEG.
4. Event-related potentials (ERPs) are recorded in the same way as EPs but occur at longer latencies from the stimuli and are associated more with endogenous brain states.
5. Because of ethical considerations, EEG recorded in brain depth or on the brain surface [electrocorticogram (ECoG)] of humans is limited to patients, most of whom are candidates for epilepsy surgery.
6. With transient EPs or ERPs the stimuli consist of repeated short pulses. The number of pulses required to produce an average EP may range from approximately 10 to several thousand, depending on the application. The scalp response to each pulse is averaged over the individual pulses. The EP or ERP in any experiment consist of a waveform containing a series of characteristic component waveforms, typically occurring less than 0.5 sec after presentation of each stimulus. The amplitude, latency from the stimulus, or covariance (in the case of multiple electrode sites) of each component may be studied, in connection with a cognitive task (ERP) or with no task (EP).
7. Steady-state EPs use a continuous sinusoidally modulated stimulus (e.g., a flickering light) typically superimposed in front of a TV monitor showing the cognitive task. The brain response in a narrow frequency band containing the stimulus

frequency is measured. Magnitude, phase, and coherence (in the case of multiple electrode sites) may be related to different parts of the cognitive task.

8. Alzheimer's disease and other dementia typically cause substantial slowing of normal alpha rhythms. Traditional EEG has been of little use in dementia because EEG changes are often only evident late in the illness when other clinical signs are obvious. However, recent efforts to apply EEG to early detection of Alzheimer's disease have shown promise.
9. Cortical tumors that involve the white matter layer (just below neocortex) cause substantial low-frequency (delta) activity over the hemisphere with the tumor. Application of EEG to tumor diagnosis has been mostly replaced by MRI, which reveals structural abnormalities in tissue.
10. Most clinical work uses spontaneous EEG; however, multiple sclerosis and surgical monitoring are exceptions, often involving EPs.
11. Studies of sensory pathways involve early components of EPs (less than approximately 50 msec) since the transmission times for signals traveling between sense organ and brain are short compared to the duration of multiple feedback associated with cognition.
12. The study of general intelligence, often associated with IQ tests, is controversial. However, many studies have reported substantial correlation between scores on written tests and different quantitative EEG measures.
13. Mathematical models of large-scale brain function are used to explain or predict observed properties of EEG in terms of basic physiology and anatomy. Although such models represent vast oversimplifications of genuine brain function, they contribute to a general conceptual framework and may guide the design of new experiments to test this framework.

II. RECORDING METHODS

A. EEG Machines

Human EEG is recorded using electrodes with diameters typically in the 0.4- to 1-cm range, held in place on the scalp with special pastes, caps, or nets as illustrated in Fig. 2. EEG recording procedures are

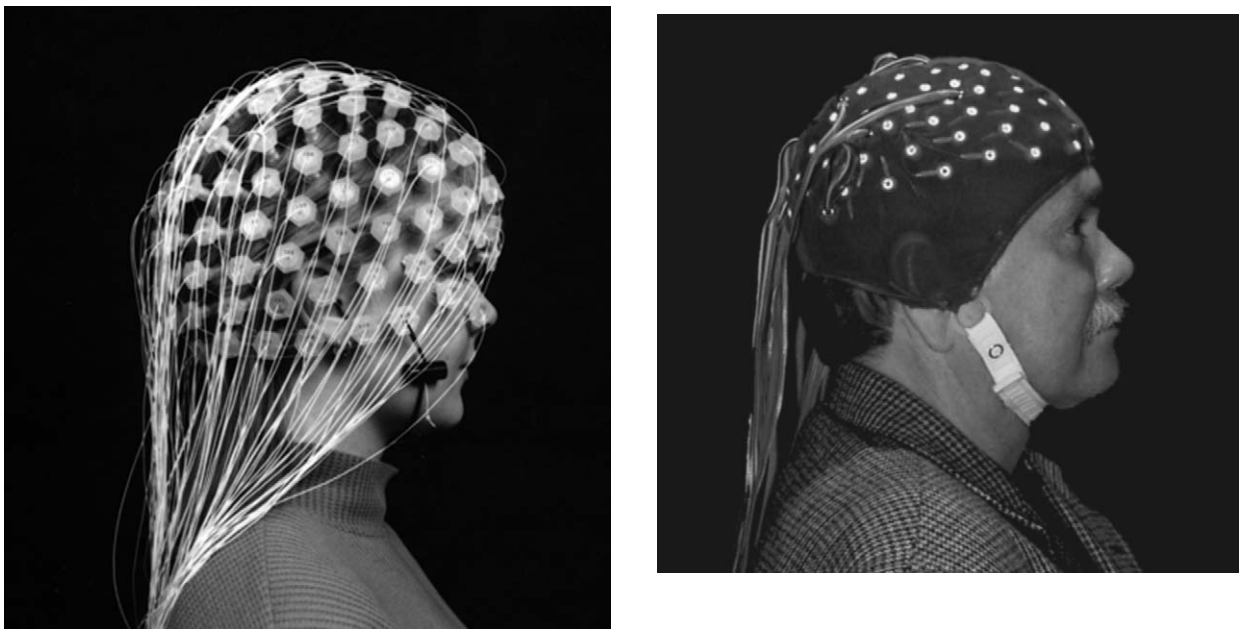


Figure 2 Two kinds of EEG scalp electrode placements are shown in which electrodes are held in place by tension in a supporting structure. (Left) A geodesic net with 128 electrodes making scalp contact with a sponge material (courtesy Electrical Geodesics, Inc.). (Right) An electrode cap containing 131 metal electrodes (courtesy Electro-Cap International, Inc., and the Brain Sciences Institute, Melbourne, Australia). Alternate methods use special pastes to attach electrodes.

noninvasive, safe, and painless. Experimental subjects used in research laboratories are often the same students or senior scientists conducting the research. Special gels are applied between electrodes and scalp to improve electrical contact. Wires from scalp electrodes connect to special EEG machines containing amplifiers to boost raw scalp signals, which are typically in the 5- to 200-μV range or approximately 100 times smaller than EKG (heart) signals. With older EEG machines, analog signals are displayed by rotating ink pens writing on chart paper that moves horizontally across machine surfaces. Modern machines typically replace such paper tracing with computer displays (*digital EEG*) and provide software packages to analyze unprocessed data.

B. Electrode Placement

In standard clinical practice, 19 recording electrodes are placed uniformly over the scalp (the *International 10–20 System*). In addition, one or two reference electrodes (often placed on ear lobes) and a ground electrode (often placed on the nose to provide amplifiers with reference voltages) are required. Potential differences between electrode pairs are recorded with EEG machines containing amplifiers, filters, and other hardware. In *referential recordings*, potentials between each recording electrode and a fixed reference are measured over time. The distinction between "recording" and "reference" electrodes is mostly artificial since both electrode categories involve potential differences between body sites, allowing closed current loops through tissue and EEG machine. *Bipolar recordings* measure potential differences between adjacent scalp electrodes. When such bipolar electrodes are placed close together (e.g., 1 or 2 cm), potential differences are estimates of tangential electric fields (or current densities) in the scalp between the electrodes. Electrode placements and the different ways of combining electrode pairs to measure potential differences on the head constitute the *electrode montage*.

Many research and some clinical laboratories use more than 21 electrodes to obtain more detailed information about brain sources. However, more electrodes may add very little useful information unless supplemented by sophisticated computer algorithms to reduce raw EEG data to a manageable form. Often, 48–131 recording electrodes are used in research; laboratories may soon use as many as 256 channels. The resulting multichannel data are submitted to computer algorithms that estimate potentials on the brain surface by accounting for distortions caused by intervening tissue and the physical separation of electrodes from brain. The combined use of high electrode density and computer algorithms providing such "inward continuation estimates" to the brain surface is called *high-resolution EEG*. Another approach using sophisticated computer methods is *dipole localization*. This method can estimate the location of source regions in the brain depths in a few specialized applications in which EEG is generated mainly in only one or two isolated source regions. However, in most applications, the sources are distributed over large regions of cerebral cortex and possibly deeper regions as well.

C. Artifact

Potentials recorded from the scalp are generated by brain sources, environmental and hardware system noise, and biological artifacts. Biological artifacts often contaminate EEG records and generally pose a more serious problem than environmental or system noise. Common artifact sources include whole body movement, heart, muscle, eyes, and tongue. EEG records containing large artifacts are often discarded. Artifact removal by computer is typically successful only for the largest artifacts. Such automated artifact editing is severely limited because the frequency bands of biological artifacts substantially overlap the important EEG bands, making distinctions between artifact and brain signal difficult.

III. TIME DEPENDENCE

A. Oscillatory Waveforms

Voltage traces of EEG signals recorded from each electrode pair oscillate with mixtures of component waveforms. Each component may be defined in terms of three parameters: its amplitude (A_{nm}), frequency (f_{nm}), and phase (ϕ_{nm}). The subscript n denotes the frequency component and the subscript m indicates the electrode pair. The electrical power associated with each frequency component is proportional to the square of the corresponding amplitude. One may express any physical waveform as a sum of components with different frequencies, amplitudes, and phases called a Fourier series. Fourier series are

analogous to expressions of music or other sounds as compositions of tones or of white light composed of many colors. The EEG voltage $V_m(t)$ recorded from any electrode pair m may be expressed generally as a sum over frequency components:

$$V_m(t) = \sum_{n=1}^{N} A_{nm} \sin(2\pi f_{nm} t - \phi_{nm}) \qquad (1)$$

Waveform frequencies f_{nm} are expressed in terms of the number of cycles per second (or Hz). EEG frequency ranges are categorized as *delta* (1–4 Hz), *theta* (4–8 Hz), *alpha* (8–13 Hz), and *beta* (>13 Hz). Very high frequencies (typically 30–40 Hz) are referred to as *gamma* activity. These distinctive labels correspond approximately to frequency ranges (or bands) that often dominate particular human brain states. For example, delta activity with frequencies lower than about 1 or 2 Hz provides the largest EEG amplitudes (or power) during deep sleep and in many coma and anesthesia states. Alpha, often mixed with low-amplitude delta, theta, and beta, is typically predominant in awake–resting states. It also occurs in *alpha coma* and is superimposed on delta activity during some sleep stages. Distinct rhythms can also occur in the same frequency range. Disparate rhythms may be associated with behavioral or cognitive state, brain location, or by other criteria. Thus, the plural terminology (alpha rhythms, beta rhythms, etc.) appropriately describes the wide variety of EEG phenomena.

B. Alpha Rhythms

Alpha rhythms provide an appropriate starting point for clinical EEG exams. The following are some initial clinical questions. Does the patient show an alpha rhythm, especially over posterior scalp? Are its spatial–temporal characteristics appropriate for the patient's age? How does it react to eyes opening, hyperventilation, drowsiness, etc.? For example, pathology is often associated with pronounced differences in EEG recorded over opposite hemispheres or with low alpha frequencies. A resting alpha frequency lower than about 8 Hz in adults is considered abnormal in all but the very old.

Alpha rhythms may be recorded in approximately 95% of healthy adults with closed eyes. The normal waking alpha rhythm usually has larger amplitudes over posterior regions, but it is typically recorded over widespread scalp regions. Posterior alpha amplitude in most normal adults is in the range 15–50 μV; alpha amplitudes recorded from frontal electrodes are lower. A posterior rhythm of approximately 4 Hz develops in babies in the first few months of age. Its amplitude increases with eye closure and is believed to be a precursor of mature alpha rhythms. Maturation of the alpha rhythms is characterized by increased frequency and reduced amplitude between ages of about 3 and 10.

Normal resting alpha rhythms may be substantially reduced in amplitude by eye opening, drowsiness, and, in some subjects, moderate to difficult mental tasks. Alpha rhythms, like most EEG phenomena, typically exhibit an inverse relationship between amplitude and frequency. For example, hyperventilation and some drugs (e.g., alcohol) may cause reductions of alpha frequencies together with increased amplitudes. Other drugs (e.g., barbiturates) are associated with increased amplitude of low-amplitude beta activity superimposed on scalp alpha rhythms. The physiological bases for the inverse relation between amplitude and frequency and most other properties of EEG are largely unknown, although physiologically based dynamic theories have provided several tentative explanations.

C. Spectral (or Fourier) Analysis

The modern methods of time series analysis are often used to simplify complicated waveforms such as EEG. Many industrial applications involve such methods as electric circuits, signal processing (television, radar, astronomy, etc.), and voice recognition. Most time series analyses are based on spectral (or Fourier) methods. Computers extract the amplitudes A_{nm} and phases ϕ_{nm} associated with each data channel (m) and frequency (n) from the often complicated EEG, represented by Eq. (1). The computer "unwraps" the waveform $V_{\mathrm{m}}(t)$ to reveal its individual components. Such spectral analysis is analogous to the physical process performed naturally by atmospheric water vapor to separate light into its component colors. Each color is composed of electromagnetic waves within a narrow frequency band, forming rainbows.

D. Alpha Spectra

Figure 3 shows a 4-sec period of alpha rhythm recorded from four scalp locations. The subject is a healthy waking adult, relaxed with eyes closed. Amplitude spectra recorded from left frontal (top left), right frontal (top right), left posterior (bottom left),

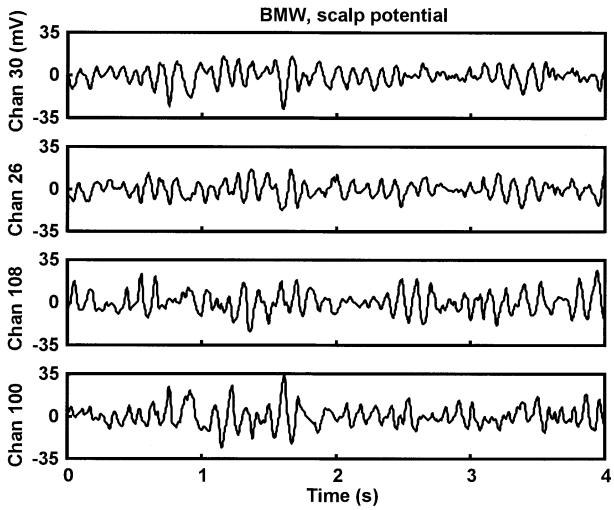


Figure 3 Alpha rhythm recorded from a healthy relaxed subject (age 25) with closed eyes using an electrode on the neck as reference. Four seconds of data are shown from four scalp locations (left frontal channel 30, right frontal channel 26, left posterior channel 108, and right posterior channel 100). The amplitude is given in microvolts. This EEG was recorded at the Brain Sciences Institute in Melbourne, Australia, using the electrode cap shown in Fig. 2 (right).

and right posterior (bottom right) scalp based on 5 min of EEG are shown in Fig. 4. The amplitude spectra show mixtures of frequencies that depend partly on scalp location; however, alpha rhythm is dominant at all locations in this typical example. This subject has two frequency peaks near 10 Hz, a relatively common finding. These two alpha oscillations are partly distinct phenomena as revealed by their separate distributions over the scalp and their distinct behaviors during mental calculations.

E. Alpha, Theta, Cognitive Tasks, and Working Memory

Two prominent features of human scalp EEG show especially robust correlation with mental effort. First, alpha band amplitude normally tends to decrease with increases in mental effort. Second, frontal theta band amplitude tends to increase as tasks require more focused attention. In addition to amplitude changes, tasks combining memory and calculations are associated with reductions in long-range coherence in narrow (1 or 2 Hz) alpha bands, whereas narrowband theta coherence increases. Large alpha coherence reductions at large scalp distances (e.g, >10 cm) can occur with no appreciable reduction in alpha amplitude and simultaneously with increases in short-range (<5 cm) alpha coherence.

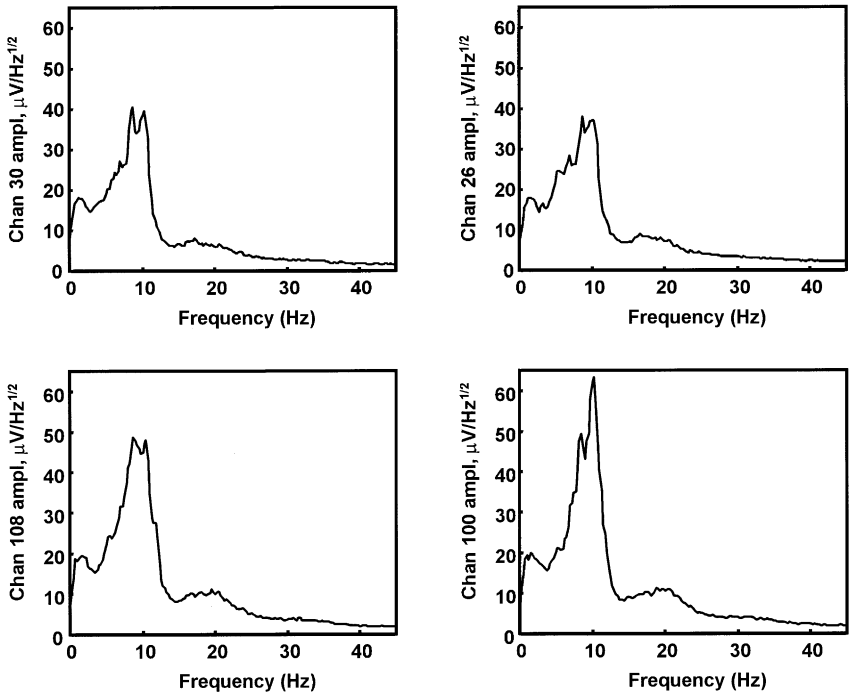


Figure 4 Amplitude spectra for the same alpha rhythms shown in Fig. 3 but based on the full 5-min record to obtain accurate spectra. Frequency resolution is 0.25 Hz. The double peak in the alpha band represents oscillations near 8.5 and 10.0 Hz.

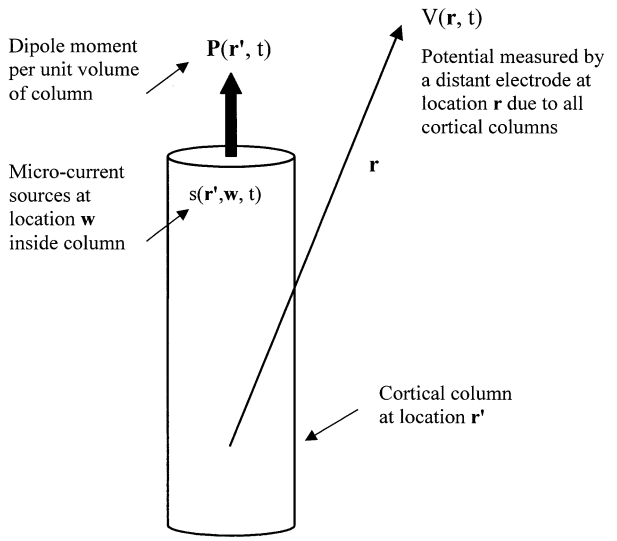


Figure 5 A single column of cerebral cortex with height equal to cortical thickness (2–4 mm) and diameter in the approximate range 0.03–1 mm (the mesoscopic scale). The volume microcurrent sources $s(\mathbf{r}', \mathbf{w}, t)$ ($\mu A/mm^3$) are generated by synaptic and action potentials at cell membrane surface elements, located by the vector $\mathbf{w}$ inside the column. A part of the column (e.g., the center) is located at $\mathbf{r}'$. The microcurrent sources integrated over the volume of the column produce a dipole moment per unit volume of the column $\mathbf{P}(\mathbf{r}', t)$, given by Eq. (2) and expressed in $\mu A/mm^2$. In Eq. (3), the electric potential $V(\mathbf{r}, t)$ at any tissue location $\mathbf{r}$ external to columns (including scalp) is due to the summed (cortical volume or surface integral) contributions from all column sources $\mathbf{P}(\mathbf{r}', t)$.

A common but oversimplified view of alpha rhythms is one of brain idling. However, upper and lower alpha band amplitudes may change independently, depending on scalp location and task. Some tasks cause lower alpha band amplitude to decrease while upper alpha band amplitude increases. Alpha amplitude reductions may be local cortical phenomena occurring in task-relevant brain areas, whereas task-irrelevant regions may be unchanged or even produce larger alpha amplitudes. Another (possibly complementary) hypothesis is that increases in higher frequency alpha amplitudes reflect a specific memory processing function and not simple idling.

Generally, intracerebral electrodes record a variety of alpha rhythms. Some intracerebral alpha rhythms are blocked by opening the eyes and some are not. Some respond in some way to mental activity and some do not. The alpha band rhythms recorded on the scalp represent spatial averages of many alpha components. For an alpha rhythm of a particular type to be observed on the scalp, it must be synchronized (roughly in phase) over a large cortical area.

F. The Rhythmic Zoo

Human EEG exhibits many waveforms, especially in the experience of clinical electroencephalographers (neurologists with specialized training). Some EEGs have known clinical significance and some do not. Any complicated waveform can be described as a mixture of oscillations with different frequencies and amplitudes, as indicated by Eq. (1). However, more picturesque descriptions are often preferred by electroencephalographers to characterized the "zoo" of EEG waveforms. Such labels include *paradoxical alpha*, *spike and wave*, *delta focus*, *sharp transient*, *sleep spindle*, *and nonspecific disrhythmia*.

Cortical EEG (ECoG) typically consists of complex waveforms composed of rhythms with different frequencies, locations, and spatial extent. This normal ECoG differentation between cortical areas is eliminated by anesthesia, suggesting a transition from more locally to more globally dominated brain dynamics. Highly localized cortical rhythms are not recorded on the scalp. Cortical beta rhythms are often strongly attenuated between cortex and scalp because they are more localized than some of the alpha band activity. EEG during sleep, coma, and anesthesia typically exhibits large scalp amplitudes, implying widely distributed cortical source activity.

IV. TOPOGRAPHY

A. Dynamic Measures that Depend on Spatial Location

EEG recorded from a single electrode pair is fully characterized by its time dependence, as in Eq. (1). In spatially extended systems, however, dynamic behavior generally depends on both time and spatial location, the usual independent variables of dynamical systems. Thus, multichannel recordings potentially introduce many new measures of brain dynamic behavior. Amplitude, phase, and frequency may vary with scalp or cortical spatial location, for example.

B. Spatial Distribution of Alpha Rhythms

Alpha rhythms have been recorded from nearly the entire upper cortical surface (ECoG), including frontal and prefrontal areas. High-resolution EEG scalp recordings also show widespread distribution of alpha rhythms over nearly the entire scalp in healthy, relaxed subjects. EEG clinical populations differ, typically

involving patients who are older, have neurological problems, and may be anxious during recording. These factors all tend to work against production of robust, widespread alpha rhythms. Second, the clinical definition of alpha is based on raw waveforms rather than spectra. Often, alpha is identified simply by counting the number of zero crossings of recorded waveforms. This can sometimes provide a misleading picture because raw EEG composed of broad frequency bands can appear very "non-alpha" to visual inspection, even though its amplitude spectrum shows substantial contribution from the alpha band. Such alpha rhythms may consist of mixtures of both localized and widely distributed activity.

Larger amplitude frontal alpha often occurs as subjects become more relaxed, for example, by employing relaxation or meditation techniques. Alpha rhythms of unusually large amplitude or exhibiting frontal dominance may be associated with mental retardation and some types of epilepsy. Large amplitude and dominant frontal alpha rhythm may also be recorded in some coma and anesthesia states. In summary, frontal alpha rhythms of moderate amplitude are common in healthy relaxed subjects with closed eyes, but very large frontal alpha is associated with disease or anesthesia. The physiological relationships between these disparate alpha phenomena are unknown, but they appear to share some underlying physiological mechanisms since their frequencies and widespread distributions are similar.

C. Coherence

Other dynamic measures involve a combination of location and time measures. For example, the (normalized) covariance of two signals is a correlation coefficient expressed as a function of time delay for characteristic waveforms recorded at the two locations. Covariance is used in ERP studies of cognition. A measure similar to covariance is the coherence of two signals, which is also a correlation coefficient (squared). It measures the phase consistency between pairs of signals in each frequency band. Scalp potential (with respect to a reference) recorded at many scalp locations, for example, over 1-min record may be represented by Eq. (1). Consider any two locations with time-dependent voltages $V_i(t)$ and $V_j(t)$. The methods of Fourier analysis may be used to determine the phases ϕ^p_{ni} and ϕ^p_{nj} associated with each 1-sec period or *epoch* (indicated by superscript p) of the full 60-sec record. The frequency component is indicated by subscript n and the two electrode locations are indicated by subscripts i and j. If the voltage phase difference $(\phi^p_{ni} - \phi^p_{nj})$ is fixed over successive epochs p (*phase locked*), the estimated EEG coherence between scalp locations i and j is equal to 1 at frequency n. On the other hand, if the phase difference varies randomly over epochs, estimated coherence will be small at this frequency.

An EEG record involving J recording electrodes will generally provide $J(J-1)/2$ coherence estimates for each frequency band. For example, with $J = 64$ electrodes and 1-sec epochs, coherence estimates may be obtained for all electrode pairs (2016) for each integer frequency between 1 and 15 Hz. The generally very complicated coherence picture may be called the *coherence structure of EEG*. This dynamic structure provides information about local versus global dynamic behavior. It provides one important measure of functional interactions between oscillating brain subsystems. EEG coherence is a different (but closely related) measure than EEG "synchrony," which refers to sources oscillating approximately in phase so that their individual contributions to EEG add by superposition. Thus, *desynchronization* is often associated with amplitude reduction. Sources that are *synchronous* (small phase differences) over substantial times will also tend to be coherent. However, the converse need not be true; coherent sources may remain approximately 180° out of phase so their individual contributions to EEG tend to cancel.

V. SOURCES OF SCALP POTENTIALS

The *generators* of scalp potentials are best described as microcurrent sources at cell membranes. Relationships between such very small-scale sources and macroscopic potentials at the scalp are made easier by employing an intermediate (*mesoscopic*) descriptive scale. This approach makes use of the columnar structure of neocortex, believed to contain the dominant sources of spontaneous scalp potentials. For macroscopic measurements, the "source strength" of a volume of tissue is defined by its electric dipole moment per unit volume:

$$\mathbf{P}(\mathbf{r}', t) = \frac{1}{W} \int \int_W \int \mathbf{w} s(\mathbf{r}', \mathbf{w}, t) dW(\mathbf{w}) \qquad (2)$$

Here, the three integral signs indicate integration over a small, local volume W of tissue, where $dW(\mathbf{w})$ is the tissue volume element. $s(\mathbf{r}', \mathbf{w}, t)$ is the local volume source current ($\mu A/mm^3$) near membrane surfaces inside a tissue volume with vector location $\mathbf{r}'$. $\mathbf{w}$ is the

vector location of sources within $dW(\mathbf{w})$ as indicated in Fig. 5. The current dipole moment per unit volume $\mathbf{P}(\mathbf{r}', t)$ in a conductive medium is fully analogous to charge polarization in a dielectric (insulator). Macroscopic tissue volumes satisfy the condition of electroneutrality at EEG frequencies. That is, current consists of movement of positive and negative ions in opposite directions, but the total charge in any mesoscopic tissue volume is essentially zero. Cortical morphology is characterized by its columnar structure with pyramidal cell axons aligned normal to the local cortical surface. Because of this layered structure, the volume elements $dW(\mathbf{w})$ may be viewed as cortical columns with height $\approx$2–5 mm, as shown in Fig. 5. For purposes of describing scalp potentials, the choice of basic cortical column diameter is somewhat arbitrary. Anything between the cortical *minicolumn* ($\approx$0.03 mm) and *macrocolumn* scales ($\approx$1 mm) may be used to describe scalp potentials.

The microsources $s(\mathbf{r}', \mathbf{w}, t)$ are generally mixed positive and negative due to local inhibitory and excitatory synapses, respectively. In addition to these active sources, the $s(\mathbf{r}', \mathbf{w}, t)$ include passive membrane (return) current required for current conservation. Dipole moment per unit volume $\mathbf{P}(\mathbf{r}', t)$ has units of current density ($\mu A/m^2$). For the idealized case of sources of one sign confined to a superficial cortical layer and sources of opposite sign confined to a deep layer, $\mathbf{P}(\mathbf{r}', t)$ is approximately the diffuse current density across the column. This corresponds to superficial inhibitory synapses and deep excitatory synapses, for example. However, more generally, column source strength $\mathbf{P}(\mathbf{r}', t)$ is reduced as excitatory and inhibitory synapses overlap along column axes.

Increased membrane capacity tends to confine the microsources $s(\mathbf{r}', \mathbf{w}, t)$ within each column to produce smaller effective pole separations—that is, smaller strengths $\mathbf{P}(\mathbf{r}', t)$. However, capacitive effects at macroscopic scales are negligible in normal EEG frequency bands. Also, tissue conductivity is only very weakly dependent on frequency. As a result of these two properties, a single dipole source implanted in the brain generates a time dependence of scalp potential that is identical (except for amplitude attenuation) to that of the source. Amplitude attenuation is independent of source frequency in the EEG range if the sources are equally distributed in location. The selective attenuation of different EEG frequency bands occurs as an indirect result of distinct spatial distributions of the sources.

Human neocortical sources may be viewed as forming a large *dipole sheet* (or *layer*) of perhaps 1500–3000 cm^2 over which the function $\mathbf{P}(\mathbf{r}', t)$ varies continuously with cortical location $\mathbf{r}'$, measured in and out of cortical folds. In limiting cases, this dipole layer might consist of only a few discrete regions where $\mathbf{P}(\mathbf{r}', t)$ is large, consisting of localized or *focal sources*. However, more generally, $\mathbf{P}(\mathbf{r}', t)$ is distributed over the entire folded surface. The question of whether $\mathbf{P}(\mathbf{r}', t)$ is distributed or localized in particular brain states is often controversial. The averaging of EPs over trials substantially alters the nature of this issue. Such time averaging strongly biases EPs toward (trial-to-trial) time stationary sources (e.g., sources confined to primary sensory cortex).

VI. VOLUME CONDUCTION OF HEAD CURRENTS

Scalp potential may be expressed as a volume integral of dipole moment per unit volume over the entire brain provided $\mathbf{P}(\mathbf{r}', t)$ is defined generally rather than in columnar terms. For the important case of dominant cortical sources, scalp potential may be approximated by the following integral of dipole moment over the cortical volume Θ:

$$V(\mathbf{r}, t) = \int_{\Theta} \int \int \mathbf{G}(\mathbf{r}, \mathbf{r}') \cdot \mathbf{P}(\mathbf{r}', t) d\Theta(\mathbf{r}') \qquad (3)$$

If the volume element $d\Theta(\mathbf{r}')$ is defined in terms of cortical columns, the volume integral may be reduced to an integral over the folded cortical surface. Equation (3) indicates that the time dependence of scalp potential is the weighted sum (or integral) of all dipole time variations in the brain, although deep dipole volumes typically make negligible contributions. The weighting function is called the vector Green's function, $\mathbf{G}(\mathbf{r}, \mathbf{r}')$. It contains all geometric and conductive information about the head volume conductor. For the idealized case of sources in an infinite medium of scalar conductivity, σ, the Green's function is

$$\mathbf{G}(\mathbf{r}, \mathbf{r}') = \frac{\mathbf{r} - \mathbf{r}'}{4\pi\sigma/\mathbf{r} - \mathbf{r}'/} \qquad (4)$$

The vector $\mathbf{G}(\mathbf{r}, \mathbf{r}')$ is directed from each column (located at $\mathbf{r}'$) to scalp location $\mathbf{r}$, as shown in Fig. 5. The numerator contains the vector difference between locations $\mathbf{r}$ and $\mathbf{r}'$. The denominator contains the (scalar) magnitude of this same difference. The dot product of the two vectors in Eq. (3) indicates that only the dipole component along this direction contributes to scalp potential. In genuine heads, $\mathbf{G}(\mathbf{r}, \mathbf{r}')$ is much more complicated. The most common head models consist of three or four concentric spherical shells,

representing brain, cerebrospinal fluid, skull, and scalp tissue with different electrical conductivities σ. More complicated numerical methods may also be used to estimate $\mathbf{G}(\mathbf{r}, \mathbf{r}')$, sometime employing MRI to determine tissue boundaries. The accuracy of both analytic and numerical methods is limited by incomplete knowledge of tissue conductivities. MRI has also been suggested as a future means of estimating tissue conductivities.

Despite these limitations preventing highly accurate estimates of the function $\mathbf{G}(\mathbf{r}, \mathbf{r}')$, a variety of studies using concentric spheres or numerical methods have provided reasonable quantitative agreement with experiment. The cells generating scalp EEG are believed to have the following properties: First, in the case of potentials recorded without averaging, cells generating EEG are mostly close to the scalp surface. Potentials fall off with distance from source regions as demonstrated by Eq. (4). In genuine heads, tissue inhomogeneity (location-dependent properties) and anisotropy (direction-dependent properties) complicate this issue. For example, the low-conductivity skull tends to spread currents (and potentials) in directions tangent to its surface. Brain ventricles, the subskull cerebrospinal fluid layer, and skull holes (or local reductions in resisitance per unit area) may provide current shunting. Generally, however, sources closest to electrodes are expected to make the largest contributions to scalp potentials.

Second, the large *pyramidal cells* in cerebral cortex are aligned in parallel, perpendicular to local surface. This geometric arrangement encourages large extracranial electric fields due to linear superposition of contributions by individual current sources. Columnar sources $\mathbf{P}(\mathbf{r}', t)$ aligned in parallel and synchronously active make the largest contribution to the scalp potential integral in Eq. (3). For example, a 1-cm^2 crown of cortical gyrus contains about 110,000 minicolumns, approximately aligned. Over this small region, the angle between $\mathbf{P}(\mathbf{r}', t)$ and $\mathbf{G}(\mathbf{r}, \mathbf{r}')$ in Eq. (3) exhibits relatively small changes. By ''synchronous'' sources, it is meant that the time dependence of $\mathbf{P}(\mathbf{r}', t)$ is approximately consistent (phase locked) over the area in question. In this case, Eq. (3) implies that individual synchronous column sources add by linear superposition. In contrast, scalp potentials due to asynchronous sources are due only to statistical fluctuations—that is, imperfect cancellation of positive and negative contributions to the integral in Eq. (3). Scalp potential may be estimated as approximately proportional to the number of synchronous columns plus the square root of the number of asynchronous columns. For example, suppose 1% ($s_1 \approx 10^3$) of the gyrial minicolumns produce synchronous sources $\mathbf{P}(\mathbf{r}', t)$ and the other 99% of minicolumns ($s_2 \approx 10^5$) produce sources with random time variations. The 1% synchronous minicolumn sources are expected to contribute approximately $s_1/\sqrt{s_2}$ or about three times as much to scalp potential measurements as the 99% random minicolumn sources.

Third, the observed ratio of brain surface (dura) potential magnitude to scalp potential magnitude for widespread cortical activity such as alpha rhythm is approximately in the 2–6 range. In contrast, this attenuation factor for very localized cortical epileptic spikes can be 100 or more. A general clinical observation is that a spike area of at least 6 cm^2 of cortical surface must be synchronously active in order to be identified on the scalp. Such area contains about 700,000 minicolumns or 70 million neurons forming a dipole layer. These experimental observations are correctly predicted by Eq. (3).

Finally, for dipole layers partly in fissures and sulci, larger areas are required to produce measurable scalp potentials. First, the maximum scalp potential due to a cortical dipole oriented tangent to the scalp surface is estimated to be about one-third to one-fifth of the maximum scalp potential due to a dipole of the same strength and depth but orientated normal to the surface. Second, tangential dipoles tend to be located more in fissures and (deeper) sulci and may also tend to cancel due to opposing directions on opposite sides of the fissures and sulci. Third, and most important, synchronous dipole layers of sources with normal orientation covering multiple adjacent gyri can form, leading to large scalp potentials due to the product $\mathbf{P}(\mathbf{r}', t) \cdot \mathbf{G}(\mathbf{r}, \mathbf{r}')$ having constant sign over the integral in Eq. (3).

VII. DYNAMIC BEHAVIOR OF SOURCES

EEG waveforms recorded on the scalp are due to a linear superposition of contributions from billions of microcurrent sources or, expressed another way, by thousands to millions of columnar sources $\mathbf{P}(\mathbf{r}', t)$ located in cerebral cortex, as indicted by Eq. (3). However, the underlying physiological bases for the dynamic behavior of the sources are mostly unknown. The 10-Hz range oscillations of alpha rhythm, the 1-Hz range oscillations of deep sleep, and other waveforms in the EEG zoo must be based on some sort of characteristic time delays produced at smaller scales. Such delays can evidently be developed in *neural networks* that cover a wide range of spatial scales.

Locally generated activity in small networks and more globally generated activity involving spatially extensive networks up to the global scale of the entire cerebral cortex may be reasonably assumed. The local network category includes so-called *thalamic pacemakers* that could possibly impose oscillations in specific frequency ranges on cortex (*local resonances*). Other possible mechanisms occur at intermediate scales between local and global. These involve feedback between cortex and thalamus or between specific cortical locations. Preferred frequencies generated at intermediate scales may be termed *regional resonances*. At the global scale, the generation of resonant frequencies (*global resonances*) due to *standing waves* of synaptic action has been proposed.

Delays in local networks are believed due mainly to *rise and decay times of postsynaptic potentials*. In contrast, global delays occur as a result of *propagation of action potentials* along axons connecting distant cortical regions (*corticocortical fibers*). Delays in regional networks may involve both local and global mechanisms. A working conjecture is that local, regional, and global resonant phenomena all potentially contribute to source dynamics. However, the relative contributions of networks with different sizes may be quite different in different brain states. The transition from awake to anesthesia states is an example of a local to global change. The ECoG changes from rhythms depending strongly on location to rhythms that look similar over widespread cortical locations. Another example is desynchronization (amplitude reduction) of alpha rhythms that occurs with eye opening and certain mental tasks.

Several mathematical theories have been developed since the early 1970s to explain the physiological bases for source dynamics—that is, the underlying reasons for specific time-dependent behaviors of the source function $\mathbf{P}(\mathbf{r}',t)$. Distinct theories may compete, complement each other, or both. Some common EEG properties for which plausible quantitative explanations have emerged naturally from mathematical theories include the following observed relations: frequency ranges, amplitude versus frequency, spatial versus temporal frequency, maturation of alpha rhythm, alpha frequency–brain size correlation, frequency versus corticocortical propagation speed, frequency versus scalp propagation speed, frequency dependence on neurotransmitter action, and mechanisms for cross-scale interactions between hierarchical networks. Because the brain is so complex, such theories must involve many approximations to genuine physiology and anatomy. As a result, verification or falsification of specific theories for the physiological bases for EEG is difficult. However, such mathematical theories can profoundly influence our general conceptual framework of brain processes and suggest new studies to test these ideas.

See Also the Following Articles

CEREBRAL CORTEX • ELECTRICAL POTENTIALS • EVENT-RELATED ELECTROMAGNETIC RESPONSES • IMAGING: BRAIN MAPPING METHODS • MAGNETIC RESONANCE IMAGING (MRI) • NEOCORTEX

Suggested Reading

Braitenberg, V., and Schuz, A. (1991). *Anatomy of the Cortex. Statistics and Geometry*. Springer-Verlag, New York.

Ebersole, J. S. (1997). Defining epileptogenic foci: Past, present, future. *J. Clin. Neurophysiol.* **14**, 470–483.

Gevins, A. S., Le, J., Martin, N., Brickett, P., Desmond, J., and Reutter, B. (1994). High resolution EEG: 124-channel recording, spatial enhancement, and MRI integration methods. *Electroencephalogr. Clin. Neurophysiol.* **90**, 337–358.

Gevins, A. S., Smith, M. E., McEvoy, L., and Yu, D. (1997). High-resolution mapping of cortical activation related to working memory: Effects of task difficulty, type of processing, and practice. *Cerebral Cortex* **7**, 374–385.

Klimesch, W. (1996) Memory processes, brain oscillations and EEG synchronization. *Int. J. Psychophysiol.* **24**, 61–100.

Malmuvino, J., and Plonsey, R. (1995). *Bioelectromagetism.* Oxford Univ. Press, New York.

Niedermeyer, E., and Lopes da Silva, F. H. (Eds.) (1999). *Electroencephalography. Basic Principals, Clinical Applications, and Related Fields*, 4th ed. Williams & Wilkins, London.

Nunez, P. L. (1981). *Electric Fields of the Brain: The Neurophysics of EEG.* Oxford Univ. Press, New York.

Nunez, P. L. (1995). *Neocortical Dynamics and Human EEG Rhythms.* Oxford Univ. Press, New York.

Nunez, P. L., Srinivasan, R., Westdorp, A. F., Wijesinghe, R. S., Tucker, D. M., Silberstein, R. B., and Cadusch, P. J. (1997). EEG coherency I: Statistics, reference electrode, volume conduction, Laplacians, cortical imaging, and interpretation at multiple scales. *Electroencephalogr. Clin. Neurophysiol.* **103**, 516–527.

Nunez, P. L., Wingeier, B. M., and Silberstein, R. B. (2001). Spatial–temporal structures of human alpha rhythms: Theory, microcurrent sources, multiscale measurements, and global binding of local networks. *Human Brain Mapping* **13**, 125–164.

Nuwer, M. (1997). Assessment of digital EEG, quantitative EEG, and EEG brain mapping: Report of the American Academy of Neurology and the American Clinical Neurophysiology Society. *Neurology* **49**, 277–292.

Sato, S. (1990). *Advances in Neurology, Vol. 54, Magnetoencephalography*. Raven Press, New York.

Scott, A. C. (1995). *Stairway to the Mind.* Springer-Verlag, New York.

Srinivasan, R., Nunez, P. L., and Silberstein, R. B. (1998). Spatial filtering and neocortical dynamics: Estimates of EEG coherence. *IEEE Trans. Biomed. Eng.* **45**, 814–825.

Uhl, C. (Ed.) (1999). *Analysis of Neurophysiological Brain Functioning*. Springer-Verlag, Berlin.

Emotion

RALPH ADOLPHS and ANDREA S. HEBERLEIN
University of Iowa College of Medicine

GLOSSARY

amygdala (from Greek *amygdala* = "almond"). A collection of nuclei located deep in the medial temporal lobe bilaterally. Extensively connected with cerebral cortex, it projects also to hypothalamus and brain stem nuclei.

basic emotions A limited set (usually six) of emotions thought to be primary; exhibited and recognized cross-culturally and observed early in human development. The standard list of basic emotions is anger, fear, sadness, happiness, disgust, and surprise.

emotional reaction The physiological components of an emotion, including but not limited to changes in heart rate, blood pressure, and piloerection.

feeling The subjective experience of emotion.

orbitofrontal cortex (used interchangeably here with ventromedial frontal cortex). The area of cerebral cortex on the ventral and medial side of the frontal lobes. This area is heavily connected to the amygdala.

periaqueductal gray matter Several columns of cells in the midbrain, surrounding the aqueduct.

social emotions A set of emotions thought to exist only in social species, in which emotional states are elicited by specific social situations and require some awareness of other individuals. Social emotions include embarrassment, pride, and guilt.

valence The pleasant or unpleasant aspect of an emotional feeling.

ventral striatum Components of the basal ganglia, including the nucleus accumbens septi and ventral parts of the caudate nucleus. These structures receive input from areas including amygdala and orbitofrontal cortex and project to the ventral globus pallidus (which projects via other structures back to amygdala and orbitofrontal cortex).

Emotions are internal states of higher organisms that serve to regulate in a flexible manner an organism's interaction with its environment, especially its social environment. Emotions can be divided into three functionally distinct but interacting sets of processes: (i) the evaluation of a stimulus or event with respect to its value to the organism, (ii) the subsequent triggering of an emotional reaction and behavior, and (iii) the representation of (i) and (ii) in the organism's brain, which constitutes emotional feeling. On the one hand, emotions are continuous with more basic motivational behaviors, such as responses to reward and punishment; on the other hand, emotions are continuous with complex social behavior. The former serve to regulate an organism's interaction and homeostasis with its physical environment, whereas the latter serves an analogous role in regard to the social environment. Studies in animals have focused on the first of the two previously mentioned aspects of emotion and have investigated the neural systems whereby behavior is guided by the reinforcing properties of stimuli. Structures such as the amygdala, orbitofrontal cortex, and ventral striatum have been shown to play critical roles in this regard. In humans, these same structures have been shown to also participate in more complex aspects of social behavior; additionally, there are structures in the right hemisphere that may play a special role in the social aspects of emotion. Emotions influence nearly all aspects of cognition, including attention, memory, and reasoning.

I. INTRODUCTION

In order to interact flexibly with a changing environment, complex organisms have evolved brains that construct an internal model of the world. Two necessary components of such a model are (i) a representation of the internal environment (i.e., a self-model) and (ii) representations of the external environment, including the other individuals constituting the social environment. An organism must continuously map self and external environment as the two interact in time. Emotion refers to a variety of different aspects of nervous system function that relate representations of the external environment to the value and significance these have for the organism. Such a value mapping encompasses several interrelated steps: evaluation of the external event or situation with which an organism is confronted, changes in brain and body of the organism in response to the situation, behavior of the organism, and a mapping of all the changes occurring in the organism that can generate a feeling of the emotion.

A. Historical Overview

William James, writing in the middle of the 19th century, made the somewhat counterintuitive claim that the subjective experience, or feeling, of emotion is caused by and follows the bodily changes of emotion. Thus, for example, one sees an angry animal approaching quickly, and one's gut tenses, one's heartbeat rises, and one's hair stands on end, all *before* one feels afraid. In fact, James argued, we depend on these physiological changes *in order* to have a feeling of an emotion.

Charles Darwin focused on emotional expressions and described similarities between human emotional expressions, such as smiles and frowns, and nonhuman animal reactions to positive and negative situations. These similarities supported his claim that human emotional expressions were innate and had evolved from once-adaptive muscle movements. Darwin also emphasized the social communicative function of emotional expressions—for example, between mother and infant or between fighting conspecifics.

These historical viewpoints have counterparts in our commonsense, or folk-psychological, concepts of emotion: Important components of emotion include observable expressions of emotion and perceptions of our own feelings, both of which follow the perception and evaluation of an emotionally salient stimulus. These components are also important in most modern theories of emotion, which share the view that emotions are adaptations sculpted by evolution, and that emotions in humans are on a continuum with emotions in animals.

B. The Functional Structure of Emotion

Emotion can be broadly viewed as a relation between an organism and its environment, pertaining both to the evaluation of external stimuli and to the organism's dispositional and actual action on the environment in response to such evaluation. To analyze this view in more detail, we can identify three components of emotion: (i) recognition/evaluation/appraisal of emotionally salient stimuli; (ii) response/reaction/expression of the emotion (including endocrine, autonomic, and motor changes); and (iii) feeling (the conscious experience of emotion). These components, and some of the neural structures that participate in them, are schematized in Fig. 1.

However, it is important to keep in mind that emotion is a broadly integrative function for which recognition (appraisal), experience (feeling), and response (expression) typically all overlap and influence one another. Several lines of evidence point to a correlation between the experience and expression of emotion, at least in many circumstances. For example, production of emotional facial expressions and other somatovisceral responses directly causes changes in emotional experience, brain activity, and autonomic state. Additionally, viewing emotional expressions on others' faces can cause systematic changes in one's own facial expression and emotional experience.

C. Basic Emotions and Facial Expression

It is widely thought that a small number of emotions are basic or primary. Data suggest that there are six basic emotional expressions: happiness, surprise, fear, anger, disgust, and sadness. However, these categories are without clearly demarcated boundaries and show some overlap (e.g., facial expressions can be members of more than one category). The conceptual structure of emotions may thus bear some similarity to the conceptual structure of colors. As with primary colors, there are basic emotions, and, like colors, an emotion can be a blend of other emotions. Basic emotions correspond closely to the emotions signaled from human facial expressions. The basic emotional expressions are recognized easily by normal subjects and are recognized consistently across very different cultures, as shown in the work of the psychologist Paul Ekman.

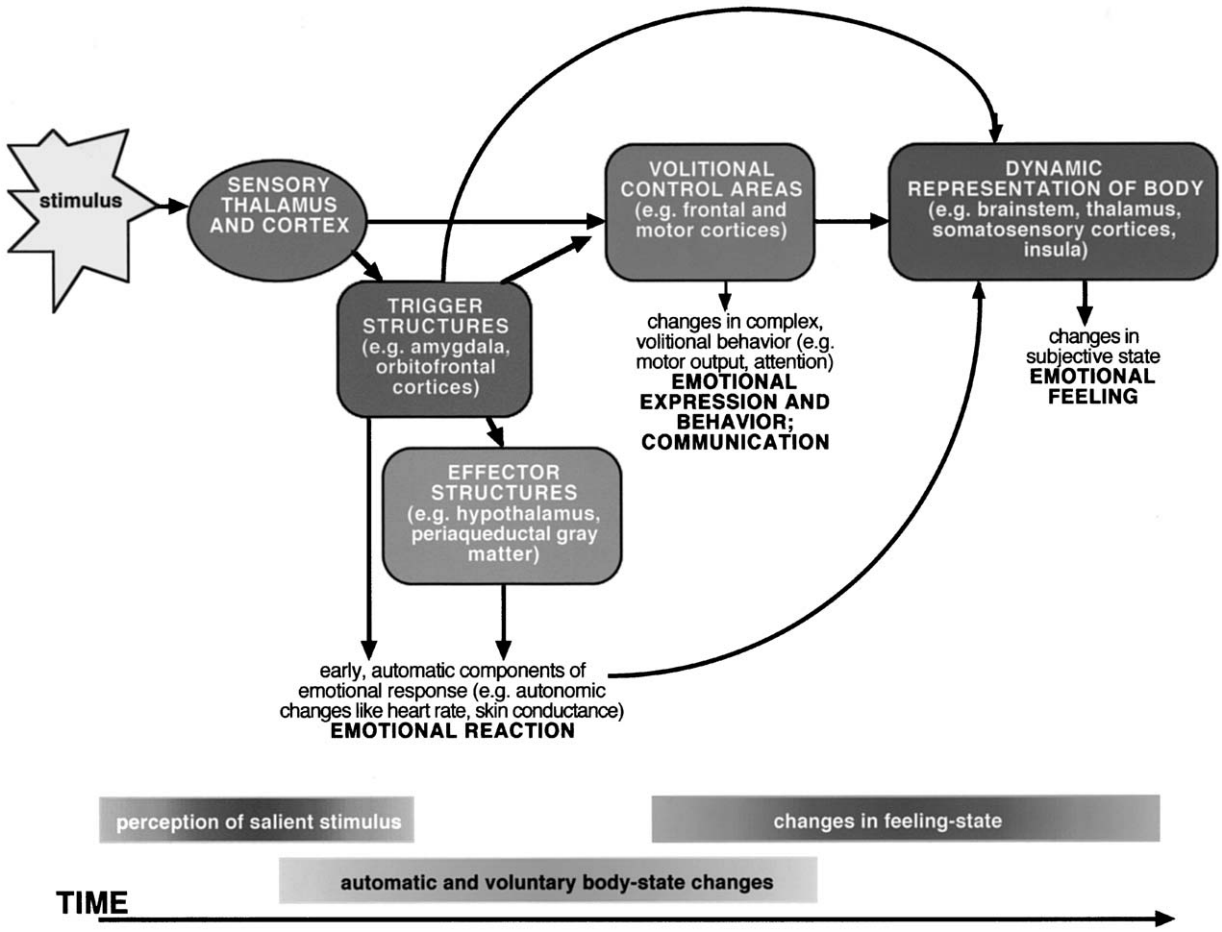


Figure 1 Time line relating the different components of emotion. A stimulus is perceived via a sensory organ, in most cases relayed through the thalamus to sensory cortex. The amygdala and other trigger structures receive projections from thalamus and from sensory cortex at many different levels. These trigger structures then project both to effector structures, such as hypothalamus, that effect changes in autonomic and endocrine activity and to volitional control areas, which are responsible for motor behavior and higher order cognition. All of these areas also project to areas which represent the body state and contribute to the subjective experience, or feeling, of emotion.

Although basic emotions may rely on largely innate factors, they do not appear immediately in infancy. Rather, like the development of language, emotions mature in a complex interplay between an infant's inborn urge to seek out and to learn certain things and the particular environment in which this learning takes place. Considerable learning important to emotion takes place between an infant and its mother. Some emotions that are present very early on, such as disgust, can be elaborated and applied metaphorically to a very large number of situations in the adult. For instance, all infants make a stereotyped face of disgust (as will most mammals) when they have ingested an unpalatable food. In adulthood, both the lexical term ''disgust'' and the facial expression are applied more broadly, for example to include responses to other people whom one finds distasteful. Although the circumstances in which an emotional expression may be elicited can be complex and can depend on the culture, a basic core set of emotional reactions are likely shared across different cultures.

Each of the basic emotions is distinct at the level of concept, experience, and expression, but psychological studies have also examined the possibility that there might be factors shared by these emotions. There is evidence from both cognitive psychology and psychophysiology that valence (pleasantness/unpleasantness) and arousal are two orthogonal factors that may capture the entire spectrum of basic emotions. Data from normal subjects show that emotions, as depicted both in facial expressions and in verbal labels, can be represented on a two-dimensional grid with valence

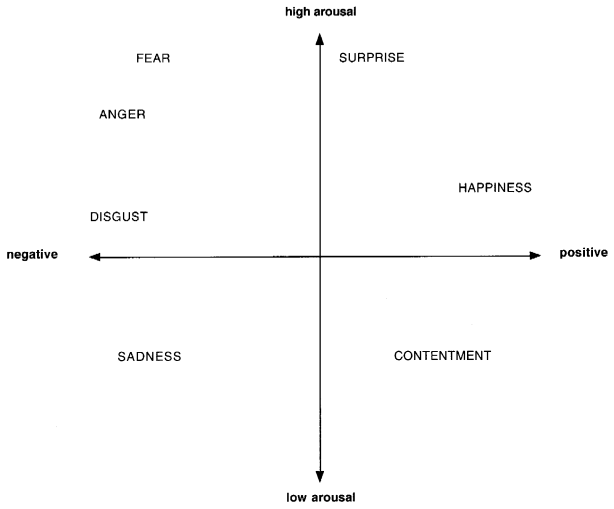


Figure 2 Valence and arousal emotion coordinates. Basic emotions can be placed on this two-dimensional grid.

and arousal as orthogonal axes (Fig. 2). Although such a representation makes sense intuitively, one need not conclude that emotions are analyzed completely in terms of their valence and arousal. A more pragmatic view might treat valence and arousal as two attributes of emotions that are useful for investigation but that may not exhaust all there is to know about emotion.

In addition to the previously mentioned ways of conceptualizing emotions, there are also the terms in our language by which we describe emotions. Certainly, these are much more numerous than just terms for basic emotions, and these include many combinations between emotions and varying shades of intensity of an emotion. Furthermore, we include so-called social emotions, such as guilt, embarrassment, and shame, which may be found only in very social animals and which may be distinct from basic emotions. Table I provides a summary of some of the different schemes used to classify different emotions. It remains an open question how these different schemes are related to one another. Currently, there is no clear classification scheme that stands out as uniquely suited for scientific investigation: In all likelihood, we will have to use different schemes for different purposes, and we will have to be able to translate research findings from one scheme to provide answers to questions posed under a different scheme. This issue becomes especially pressing when we are using data obtained from studies in nonhuman animals, which often investigate emotion at the level of reward and punishment, to provide insights into emotion at the psychological level in humans. Here, we provide an overview of research findings from both animals and humans that considers some of these issues.

II. ANIMALS

A. Historical Research Findings in Animals

Early studies of the neural substrates of emotion-related behavior in nonhuman animals used both lesion and stimulation methods. In the 1920s, Philip Bard and Walter Cannon observed behaviors in decorticate cats that appeared similar to extreme anger or rage. Brain transections produced increased autonomic activity and behaviors including tail lashing, limb jerking, biting, and clawing. They termed this *sham rage* because they thought it occurred without

Table I
Some Schemes Used to Classify Emotions

Classification schemes	Theory/function
Reward/punishment, approach/withdrawal, and positive/negative valence and high/low arousal	All emotions are reducible to and/or based on basic and opposing principles
Basic emotions: happiness, sadness, anger, fear, disgust, and surprise (sometimes others)	Evolved for adaptive value relative to commonly encountered situations
Basic/primary emotions: desire, anger, fear, sadness, sexual lust, joy, and maternal acceptance/nurturance	Come from subcortical brain mechanisms; evolved for adaptive value; there are several categories of emotions involving progressively greater levels of higher cognitive involvement
Social emotions: guilt, embarrassment/shame, and pride (possibly flirtation)	Occur in addition to basic emotions; elicited only in social situations
All the emotions we have words for in our language	Probably many more distinctions made in language than there are separable brain systems subserving emotions

conscious emotional experience and because it was elicited by very mild stimuli, such as light touches. Further research showed that when the lateral hypothalamus was included in the transected region, sham rage was not observed. More focal lesions of the lateral hypothalamus were found to result in placidity and lesions of medial hypothalamus to result in irritability. This observation correlated with the results of stimulation experiments performed at approximately the same time by Walter Hess. Hess implanted electrodes into different areas of the hypothalamus and observed different constellations of behaviors depending on electrode placement. Stimulation of the lateral hypothalamus in cats resulted in increased blood pressure, arching of the back, and other autonomic and somatic responses associated with anger. These results led to a concept of the hypothalamus as an organizer and integrator of the autonomic and behavioral components of emotional responses.

A second key neuroanatomical finding in animals was that by Klüver and Bucy in 1937, who showed that large bilateral lesions of the temporal lobe, including amygdala, produced a syndrome in monkeys such that the animals appeared unable to recognize the emotional significance of stimuli. For instance, the monkeys would be unusually tame, and would approach and handle stimuli, such as snakes, of which normal monkeys are afraid.

On the basis of these animal findings, as well as on the basis of findings in humans, early theorists proposed several neural structures as important components of a system that processes emotion. One of the most influential of these, put forth by Paul MacLean in the 1940s and 1950s, was the notion of a so-called "limbic system" encompassing amygdala, septal nuclei, and orbitofrontal and cingulate cortices. This system was interposed between, and mediated between, neocortical systems concerned with perceiving, recognizing, and thinking, on the one hand, and brain stem and hypothalamic structures concerned with emotional reaction and homeostasis, on the other hand.

Although the concept of a specific limbic system is debated, the idea is useful to distinguish some of the functional components of emotion, as we have done previously. A key insight is the need for specific structures that can link sensory processing (e.g., the perception of a stimulus) to autonomic, endocrine, and somatomotor effector structures in hypothalamus, periaqueductal gray, and other midbrain and brain stem nuclei. In the next section, we discuss various structures that play a role in either linking sensory processing to emotional behavior or effecting the body state changes of emotion.

B. Brain Structures Studied in Emotional Behavior in Animals

1. Amygdala

The amygdala is a collection of nuclei deep in the anterior temporal lobe which receives highly processed sensory information and which has extensive, reciprocal connections with a large number of other brain structures whose function can be modulated by emotion. Specifically, the amygdala has massive connections, both directly and via the thalamus, with the orbitofrontal cortices, which are known to play a key role in planning and decision making. The amygdala connects with hippocampus, basal ganglia, and basal forebrain—all structures that participate in various aspects of memory and attention. In addition, the amygdala projects to structures such as the hypothalamus that are involved in controlling homeostasis and visceral and neuroendocrine output. Consequently, the amygdala is situated so as to link information about external stimuli conveyed by sensory cortices, on the one hand, with modulation of decision-making, memory, and attention as well as somatic, visceral, and endocrine processes, on the other hand.

Although the work of Klüver and Bucy implicated the amygdala in mediating behaviors triggered by the emotional and social relevance of stimuli, by far the majority of studies of the amygdala in animals have investigated emotion not at the level of social behavior but at the level of responses to reward and punishment. These studies have demonstrated the amygdala's role in one type of associative memory—the association between a stimulus and the survival-related value that the stimulus has for the organism. The amygdala is essential to link initially innocuous stimuli with emotional responses on the basis of the apparent causal contingencies between the stimulus and a reinforcer. Although such a mechanism is in principle consistent with the amygdala's broad role in real-life social and emotional behaviors, it has been best studied in the laboratory as "fear conditioning." Fear conditioning uses the innate response to danger, which is similar in many mammals and includes freezing, increase in blood pressure and heart rate, and release of stress hormones. This response is elicited in fear conditioning by a noxious stimulus such as an electric shock to the foot. If this shock is preceded

by an innocuous stimulus such as a bell, the bell comes to be associated with the shock, and eventually the subject will exhibit the fear response to the bell. Joseph LeDoux and colleagues used this paradigm to study the neural substrates of conditioned fear responses in rats. By using anatomical tracing techniques to determine candidate structures and then lesioning these to observe changes in conditioning behavior, LeDoux determined that two nuclei in the amygdala are vital for associating an auditory stimulus with a fear response. The lateral nucleus receives projections from the auditory thalamus and cortex, and the central nucleus coordinates the response in various effector systems (freezing, increase in blood pressure, etc.).

The amygdala's role in associating sensory stimuli with emotional behaviors is also supported by findings at the single cell level. Neurons within the amygdala modulate their responses on the basis of the rewarding or punishing contingencies of a stimulus, as shown in the work of Edmund Rolls and others. Likewise, the responses of neurons in primate amygdala are modulated by socially relevant visual stimuli, such as faces and videos of complex social interactions. Again, it is important to realize that the amygdala is but one component of a distributed neural system that links stimuli with emotional response. There are several other structures, all intimately connected with the amygdala, that subserve similar roles (Fig. 1).

2. Orbitofrontal Cortex

Lesions of the orbitofrontal cortex (discussed in more detail later in regard to humans) produce impairments very similar to those seen following amygdala damage. As in the amygdala, single-neuron responses in the orbitofrontal cortex are modulated by the emotional significance of stimuli, such as their rewarding and punishing contingencies, although the role of the orbitofrontal cortex may be more general and less stimulus bound than that of the amygdala. Amygdala and orbitofrontal cortex are bidirectionally connected, and lesion studies have shown that disconnecting the two structures results in impairments similar to those following lesions of either structure, providing further support that they function as components of a densely connected network.

3. Ventral Striatum

Structures such as the nucleus accumbens also receive input from the amygdala and appear to be especially important for processing rewarding stimuli and for engaging the behaviors that cause an organism to seek stimuli that predict reward. Amygdala, ventral striatum, and orbitofrontal cortex all participate jointly in guiding an organism's expectation of reward on the basis of prior experience. Recent elegant single-unit studies by Wolfram Schultz and colleagues dissect some of the specific component processes. An important neurochemical system subserves the functional connectivity between ventral striatum and frontal cortex: the neurotransmitter dopamine. This system and the specific neurotransmitters involved are currently being intensively investigated as models of drug addiction.

4. Other Trigger Structures

There are several other structures that link stimulus perception to emotional response. Work by Michael Davis and colleagues has highlighted nuclei situated very close to the amygdala, such as the bed nucleus of the stria terminalis, and emphasized their role in anxiety. Other structures in the vicinity, such as nuclei in the septum, are also important and may mediate their effects through the neurotransmitter acetylcholine.

There are a collection of nuclei in the brain stem that can modulate brain function in a global fashion by virtue of very diverse projections. The locus ceruleus, a very small set of nuclei, provides the brain with its sole source of noradrenergic innervation. Similarly, the Raphe nuclei provide a broad innervation of serotonergic terminals. These neuromodulatory nuclei, together with the dopaminergic and cholinergic nuclei mentioned previously, are thus in a position to alter information processing globally in the brain. It is important to emphasize that these changes in the brain's information processing mode are just as important as the somatic components of an emotional reaction —and just as noticeable when we feel the emotion.

5. Effector Structures

The structures involved in emotional reaction and behavior include essentially all those that control motor, autonomic, and endocrine output. Some of these structures have further internal organization that permits them to trigger a coordinated set of responses. For instance, motor structures in the basal ganglia control some of the somatic components of emotional response (facial expressions in humans), and distinct regions in the hypothalamus trigger concerted emotional reactions of fear or aggression, as mentioned previously. Another important structure

is the periaqueductal gray matter (PAG), which consists of multiple columns of cells surrounding the aqueduct in the midbrain. Stimulation of these areas has long been known to produce panic-like behavioral and autonomic changes in nonhuman animals as well as reports of panic-like feelings in humans. Moreover, different columns within the PAG appear to be important for different components of emotional response. In a recent functional imaging study, structures in brain stem and hypothalamus were active when human subjects were experiencing emotions, further supporting the roles of brain stem and hypothalamic structures in the coordination of emotional reaction and behavior.

III. HUMANS

Not surprisingly, the neural structures that are important for emotional behavior in nonhuman animals are also important for emotional behavior in humans (Fig. 3). Again, these can be divided into (i) structures important for homeostatic regulation and emotional reaction, such as the hypothalamus and PAG, and (ii) structures for linking perceptual representation to regulation and reaction, such as the amygdala and orbitofrontal cortex. We will discuss the role of frontal and parietal regions in representing the organism's own changes in body state, focusing on right frontoparietal cortex as well as other somatosensory structures such as the insula. It is again important to remember, however, that all these structures are heavily interconnected, and that most play at least some role in multiple components of emotion. Next we discuss those structures for which the most data are available from humans: the amygdala, the orbitofrontal cortex, and right somatosensory-related cortices.

A. The Human Amygdala

Data on the amygdala in humans have come primarily from lesion studies and from functional imaging studies [e.g., positron emission tomography (PET) or functional magnetic resonance imaging] that image brain activity in neurologically normal individuals. These studies have provided evidence that the amygdala responds to emotionally salient stimuli in the visual, auditory, olfactory, and gustatory modalities. Lesion studies involve patients who have had amygdala damage because of encephalitis (such as Herpes simplex encephalitis) or other rare diseases or who have had neurosurgical resection of the amygdala on one side of the brain to ameliorate epilepsy.

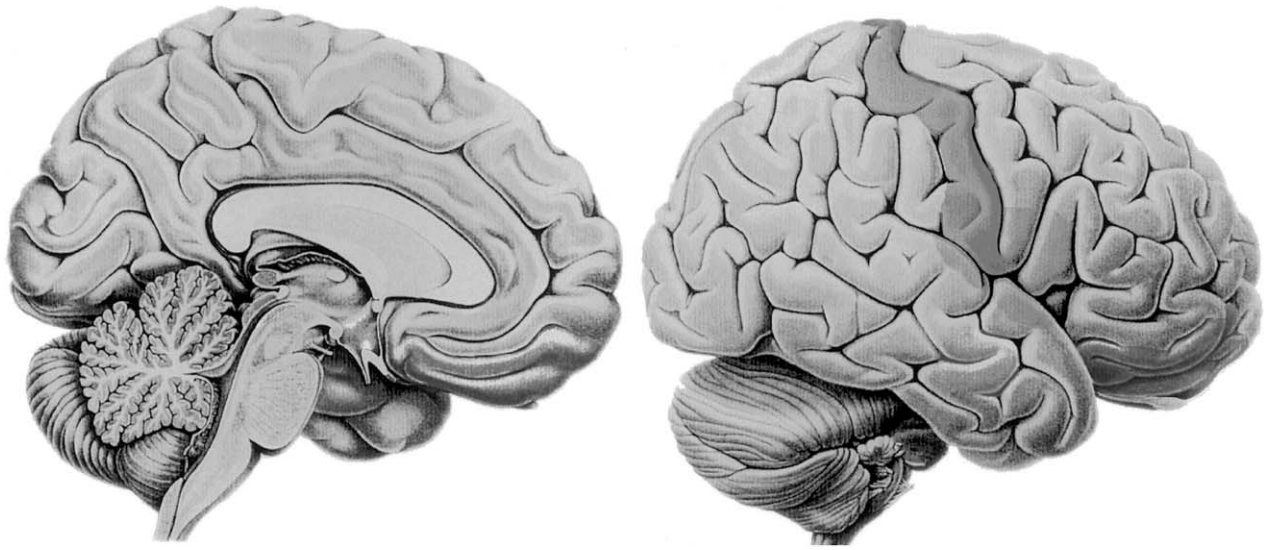


Figure 3 Some of the human brain areas important for emotion. (Left) Medial view of left hemisphere showing orbitofrontal cortex and, deep within the temporal lobe, amygdala, as well as hypothalamus and periaqueductal gray. (Right) Lateral view of the right hemisphere showing orbitofrontal cortex and, buried in temporal cortex, amygdala and also somatosensory cortex and, buried under overlying cortex, insula. As shown schematically in Fig. 1, amygdala and orbitofrontal cortex receive perceptual information and project to effector structures including, principally, hypothalamus and periaqueductal gray. These latter structures effect automatic changes in body state, including blood pressure and heart rate etc. The representation of body state (feeling) depends critically on somatosensory areas, including somatosensory cortex and insula. (See color insert in Volume 1).

As in animals, the human amygdala appears to be important for fear conditioning—for associating conditioned sensory stimuli with an aversive unconditioned stimulus. A variety of neuropsychological tasks have been used in humans to investigate in more detail both the recognition of emotion from stimuli (such as the recognition of emotions from viewing facial expressions of other people) and the experience of emotion triggered by emotional stimuli or emotional memories. Studies of the amygdala's role in emotion recognition have primarily used photographs of emotional facial expression (such as those developed by Paul Ekman). One subject with selective bilateral amygdala damage has been studied extensively with regard to her recognition of emotion in these photographs of facial expressions. This subject, SM046, has been shown in several different tasks to be specifically and severely impaired in regard to faces of fear. When rating the intensity of emotions in facial expressions, SM046 consistently failed to rate the emotions surprise, fear, and anger as very intense. She was particularly impaired in rating the intensity of fear, on several occasions failing to recognize any fear whatsoever in prototypical facial expressions of fear.

SM046's spontaneous naming of the emotions shown in faces in a labeling experiment using identical stimuli was impaired relative to normal controls: She virtually never used the label "fear," typically mislabeling such faces as surprised, angry, or disgusted. Thus, subject SM046's impairment in recognizing emotional facial expressions is disproportionately severe with respect to fear. However, she also has lesser impairments in recognition of highly arousing emotions that are similar to fear, such as anger. This is consistent with a more general impairment in recognition of negative emotions observed in other subjects with bilateral amygdala damage, and it leads to the question of how specific the amygdala's role is in recognition of certain emotions. Interestingly, SM046 is also impaired in her ratings of the degree of arousal present in facial expressions of emotion. When asked to place photographs of emotional facial expressions on a grid with valence (positive/negative) and arousal (low/high) as orthogonal axes, SM046 was normal in her valence ratings but abnormal in the level of arousal that she assigned to negative facial expressions. Thus, it is not the case that bilateral amygdala damage impairs all knowledge regarding fear; rather, it impairs the knowledge that fear is highly arousing.

A likely interpretation of the previous results from SM046, in conjunction with results from other subjects with bilateral amygdala damage, is that the amygdala is part of a more general neural system for recognizing highly arousing, unpleasant emotions—in other words, emotions that signal potential harm to the organism—and rapidly triggering physiological states related to these stimuli. Such physiological states involve both specific sets of behavioral responses and the modulation of cognitive processes, including those involved in knowledge retrieval necessary for normal performance on the previously mentioned tasks. In animals, the amygdala may trigger predominantly behavioral reactions; in humans, it may trigger both behavior and conscious knowledge that the stimulus predicts something "bad." How it is that conceptual knowledge about the arousal component of emotions comes to depend on the amygdala, in addition to emotional arousal itself depending on the amygdala, is a key issue for future research.

Functional imaging studies in normal individuals have corroborated the lesion studies implicating the amygdala in recognition of signals of unpleasant and arousing emotions. Visual, auditory, olfactory, and gustatory stimuli all appear to engage the amygdala when signaling unpleasant and arousing emotions. These studies have examined the encoding and recognition of emotional stimuli, as well as emotional experience and emotional response, but it has been exceedingly difficult to disentangle all these different components. Although there is now clear evidence of amygdala activation during encoding of emotional material, it is less clear whether the amygdala is also activated during retrieval. Two findings suggest that the amygdala's role may be specific to linking external sensory stimuli to emotion and not for triggering emotional responses that are internally driven. First, subjects with bilateral amygdala lesions can volitionally make facial expressions of fear. Second, when normal subjects induced a subjective experience of fear in themselves while undergoing a PET scan, no activation of the amygdala was observed.

Further insight has come from studies that used stimuli that could not be consciously perceived. Amygdala activation was observed when subjects viewed facial expressions of fear that were presented so briefly they could not be consciously recognized, showing that the amygdala plays a role in nonconscious processing of emotional stimuli. In summary, an important function of the amygdala may be to trigger responses and to allocate processing resources to stimuli that may be of special importance or threat to the organism, and ecological considerations as well as

the data summarized here all appear to argue for such a role especially in regard to rapid responses that need not involve conscious awareness.

B. Orbitofrontal Cortex

The importance of the frontal lobes in social and emotional behavior was demonstrated in the mid-1800s by the famous case of Phineas Gage. Gage, a railroad construction foreman, was injured in an accident in which a metal tamping rod shot under his cheekbone and through his brain, exiting through the top of his head. Whereas Gage had been a diligent, reliable, polite, and socially adept person before his accident, he subsequently became uncaring, profane, and socially inappropriate in his conduct. Extensive study of modern-day patients with similar anatomical profiles (i.e., bilateral damage to the ventromedial frontal lobes), has shed more light on this fascinating historical case. These patients show a severely impaired ability to function in society, even with normal IQ, language, perception, and memory. The work of Antonio Damasio and others has illuminated the importance of ventromedial frontal cortices (VMF; we use ventromedial frontal cortex and orbitofrontal cortex interchangeably here) in linking stimuli with their emotional and social significance. This function bears some resemblance to that of the amygdala outlined previously but with two important differences. First, it is clear that the ventromedial frontal cortices play an equally important role in processing stimuli with either rewarding or aversive contingencies, whereas the amygdala's role, at least in humans, is clearest for aversive contingencies. Second, reward-related representations in VMF cortex are less stimulus driven than in the amygdala and thus can play a role in more flexible computations regarding punishing or rewarding contingencies.

Antonio Damasio and colleagues tested VMF-lesioned patients on several types of tasks involving the relation of body states of emotion to behavioral responses. When patients with bilateral VMF damage were shown slides of emotionally significant stimuli such as mutilation or nudity, they did not show a change in skin conductance (indicative of autonomic activation). Control groups showed larger skin conductance responses to emotionally significant stimuli, compared to neutral stimuli, suggesting that VMF patients are defective in their ability to trigger somatic responses to stimuli with emotional meaning. In a gambling task in which subjects must develop hunches about certain decks of cards in order to win money, VMF-lesioned patients made poor card choices and also acquired neither subjective feeling regarding their choices nor any anticipatory autonomic changes before making these poor choices. All these findings support the idea that the VMF cortices are a critical component of the neural systems by which we acquire, represent, and retrieve the values of our actions, and they emphasize the close link between emotion and other aspects of cognitive function, such as reasoning and decision making. Damasio presented a specific neuroanatomical theory of how emotions play a critical role in reasoning and decision making—the *somatic marker hypothesis*. According to this hypothesis, our deliberation of choices and planning of the future depend critically on how we feel about the different possibilities with which we are faced. The construction of some of the components of an emotional state and the feeling that this engenders serve to tag response options with value and serve to bias behavior toward those choices associated with positive emotions. This set of processes may operate either under considerable volitional guidance, and as such may be accessible to conscious awareness, or it may play out in a more automatic and covert fashion.

Both the amygdala and the orbitofrontal cortex function as components of a neural system that can trigger emotional responses. The structure of such a physiological emotional response may also participate in attempts to reconstruct what it would feel like to be in a certain dispositional (emotional or social) state and hence to simulate the internal state of another person. In the case of the amygdala, the evidence thus far points toward such a role specifically in regard to states associated with threat and danger; in the case of the orbitofrontal cortex, this role may be somewhat more general.

C. The Right Hemisphere

One generally defined neural area whose importance in emotional behavior and perception has been explored in primates much more than in other animals is the right hemisphere. Both clinical and experimental studies have suggested that the right hemisphere is preferentially involved in processing emotion in humans and other primates. Lesions in right temporal and parietal cortices have been shown to impair emotional experience, arousal, and imagery. It has been proposed that the right hemisphere contains systems specialized for computing affect from

nonverbal information; these may have evolved to subserve aspects of social cognition.

Recent lesion and functional imaging studies have corroborated the role of the right hemisphere in emotion recognition from facial expressions and from prosody. There is currently controversy regarding the extent to which the right hemisphere participates in emotion: Is it specialized to process all emotions (the *right hemisphere hypothesis*), or is it specialized only for processing emotions of negative valence while the left hemisphere is specialized for processing emotions of positive valence (the *valence hypothesis*)? It may well be that an answer to this question will depend on more precise specification of which components of emotion are under consideration.

Recognition of emotional facial expressions can be selectively impaired following damage to right temporoparietal areas, and both PET and neuronal recordings corroborate the importance of this region for processing facial expressions of emotion. For example, lesions restricted to right somatosensory cortex result in impaired recognition of emotion from visual presentation of face stimuli. These findings are consistent with a model that proposes that we internally simulate body states in order to recognize emotions.

In contrast to recognition of emotional stimuli, emotional experience appears to be lateralized in a pattern supporting the valence hypothesis, in which the left hemisphere is more involved in positive emotions and the right hemisphere is more involved in negative emotions. Richard Davidson posited an approach/withdrawal dimension, correlating increased right hemisphere activity with increases in withdrawal behaviors (including feelings such as fear or sadness, as well as depressive tendencies) and left hemisphere with increases in approach behaviors (including feelings such as happiness).

D. Neuropsychiatric Implications

Emotion is a topic of paramount importance to the diagnosis, treatment, and theoretical understanding of many neuropsychiatric disorders. The amygdala has received considerable attention in this regard and has been shown to be involved in disorders that feature fear and anxiety. Moreover, specific neurotransmitters, acting within the amygdala and surrounding structures, have been shown to contribute importantly to fear and anxiety. Anxiolytic drugs such as valium bind to GABA-A receptor subtypes in the amygdala and alter the neuronal excitability. Corticotropin-releasing factor is an anxiogenic peptide that appears to act in the amygdala and adjacent nuclei. Several functional imaging studies have demonstrated that phobic and depressive symptoms rely on abnormal activity within the amygdala, together with abnormalities in other brain structures.

With regard to depression, the frontal lobes have also been investigated for their contribution to emotional dysfunction in psychiatric disorders. Evoked potential recordings and functional imaging studies have revealed their participation in depression, which may engage regions below the frontal end of the corpus callosum. Moreover, individual differences in affective style, independent of any overt pathology, may rely on hemispherically asymmetric processing within the frontal lobes.

IV. DEVELOPMENT AND EVOLUTION OF EMOTION

Both the developmental and the evolutionary aspects of emotion remain important issues for further research. A large body of findings, primarily from developmental psychology, has shown that the highly differentiated sets of emotions seen in adult humans develop over an extended time course that requires extensive interactions between an infant, its parents, and its cultural environment. The importance of many of the structures discussed previously in the development of emotional behavior is underscored by findings that damage to these structures relatively early in development causes more severe impairments than damage during adulthood. Although newborns do show some relatively undifferentiated emotional responses (such as general distress), and although they have an innate predisposition to respond to emotionally salient stimuli (such as the mother's face), the subsequent development of emotion depends both on the presence of critical neural structures and, crucially, on the child's environment. As with language, humans are predisposed to have a rich and complex set of emotional processes, but the precise details require maturation and learning in a socially rich environment.

Phylogenetically, human emotion depends on the more basic sets of emotions and motivational processes that we share in common with other animals. Clearly, the neural circuitry that subserves

the processing of reward and punishment must be in place before more differentiated emotions can evolve. However, higher mammals, and especially highly social mammals such as primates, did evolve additional circuitry in order to permit them to respond in a more flexible and adaptive manner to environments that change rapidly in time. The most dynamic environment of all, of course, is the social environment, and keeping track of and responding rapidly and appropriately to numerous conspecifics requires a rich repertoire of emotional regulation.

The comparative and developmental investigations of emotion raise important questions about how emotion contributes to behavior and about how emotion contributes to other aspects of cognition. A point of fundamental importance is that more complex behavior, and more complex cognition, requires more complex and differentiated emotions.

V. THE FUTURE

Emotion is now a hot topic in cognitive science in general and in neuroscience in particular. Future directions can be classified under two general topics: (i) development of a theoretical framework for thinking about emotion and for generating hypotheses regarding its component processes and (ii) further empirical investigations using new methods and using new combinations of methods and species. Especially important will be studies that combine different techniques, such as functional imaging and lesion methods, and studies that combine the same paradigms in different species, such as infant humans and monkeys. Such a multifaceted approach to investigating emotion is in fact being pursued by many laboratories. Currently, neuroscientists, psychologists, and anthropologists are collaborating on several of these issues.

A very difficult, but very important, problem to address in the future is the relation between emotion and consciousness. Recent proposals, for instance, by Antonio Damasio and Jaak Panksepp, have stressed that a proper understanding of emotion may in fact provide the key to understanding one particular feature of conscious experience: the fact that consciousness is always experienced from the particular point of view of the subject. The subjectivity of conscious experience shares in common with the feeling of an emotion that it requires a neural instantiation of a subject; that is, both require a set of structures in the brain that map and represent the organism and its ongoing state changes as the organism interacts with its environment. This proposal is in line with findings that damage to right hemisphere structures involved in self-representation also impairs the ability to experience emotions. The further investigation of the neural basis of such a mechanism, and of its enormous elaboration in humans, may provide us with a better understanding not only of emotion but also of the nature of conscious experience and its role in human cognition.

See Also the Following Articles

ANGER • AGGRESSION • BEHAVIORAL NEUROGENETICS • COGNITIVE PSYCHOLOGY, OVERVIEW • CREATIVITY • EVOLUTION OF THE BRAIN • HUMOR AND LAUGHTER • INHIBITION • PSYCHONEUROENDOCRINOLOGY • SEXUAL BEHAVIOR

Suggested Reading

Adolphs, R. (2002). Recognizing emotion from facial expressions: psychological and neurological mechanisms. *Behav. Cogn. Neurosci. Rev.* **1,** 21–61.

Adolphs, R. (1999b). Social cognition and the human brain. *Trends Cognitive Sci.* **3,** 469–479.

Aggleton, J. P. (Ed.) (2000). *The Amygdala: A Functional Analysis*. Oxford Univ. Press, New York.

Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. Grosset/Putnam, New York.

Davidson, R. J., and Irwin, W. (1999). The functional neuroanatomy of emotion and affective style. *Trends Cognitive Sci.* **3,** 11–22.

Davis, M. (1992). The role of the amygdala in fear and anxiety. *Annu. Rev. Neurosci.* **15,** 353–375.

LeDoux, J. (1996). *The Emotional Brain*. Simon & Schuster, New York.

Panksepp, J. (1998). *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford Univ. Press, New York.

Rolls, E. T. (1999). *The Brain and Emotion*. Oxford Univ. Press, New York.

Endorphins and Their Receptors

CATHERINE ABBADIE and GAVRIL W. PASTERNAK
Memorial Sloan–Kettering Cancer Center

GLOSSARY

alternative splicing A way in which different portions of a gene are put together to yield more than a single protein.

analgesic A substance capable of relieving pain without interfering with other sensations.

endorphins Peptides naturally present in the brain with morphine-like actions. They function by activating a family of opioid receptors, which are also responsible for the effects of drugs such as morphine, methadone, and heroin.

opiates Compounds acting on the opioid receptors. Initially defined by their pharmacological similarity to morphine and other analgesic alkaloids present in opium, a product of the poppy plant.

respiratory depression A decrease in breathing.

The opiates, initially derived from opium obtained from the poppy plant, have opened new insights into many aspects of functioning of the brain. The opiates have been used since ancient times and represent one of the most important classes of medications currently in use. In ancient times, opium and extracts of opium were primarily used, although opium also was smoked. Morphine and codeine were isolated and purified from opium in the 1800s, providing physicians with a pure drug with a constant level of activity. Since then, thousands of analogs have been developed in an effort avoid side effects. The structures of these agents vary greatly: however, they share many pharmacological properties.

I. INTRODUCTION

The opiates have a wide range of actions. The opiates remain the most widely used drugs for the relief of moderate to severe pain (Fig. 1). Their actions on pain perception are unique. Rather than blocking the transmission of pain impulses into the central nervous system, like local anesthetics, or working on the sensitization of pain fibers, like the antiinflammatory drugs, opiates relieve the ''suffering component'' of pain. Patients often report that the pain is still there, but it just does not hurt anymore. This ability to relieve the pain without interfering with other aspects of sensation provides a major advantage. Furthermore, opiates do not display a ''ceiling effect,'' implying that even very severe pain can be successfully relieved at sufficiently high drug doses. Unfortunately, side effects become increasingly troublesome as the dose is increased and the ability of the patient to tolerate these unwanted actions may interfere with the ability to administer adequate doses of the drug.

The actions of the opiates are not limited to pain. The additional actions most often encountered when treating patients are constipation, sedation, and respiratory depression. All these actions are mediated through opioid receptors and can be reversed by opioid antagonists, although evidence is mounting that they may be produced by different opioid receptor subtypes. Opiates influence gastrointestinal transit both centrally and peripherally. Although this action can be troublesome in pain management, it is valuable in the treatment of conditions associated with increased gastrointestinal motility, such as diarrhea.

Respiratory depression is another important opioid action. Opiates such as morphine depress respiratory

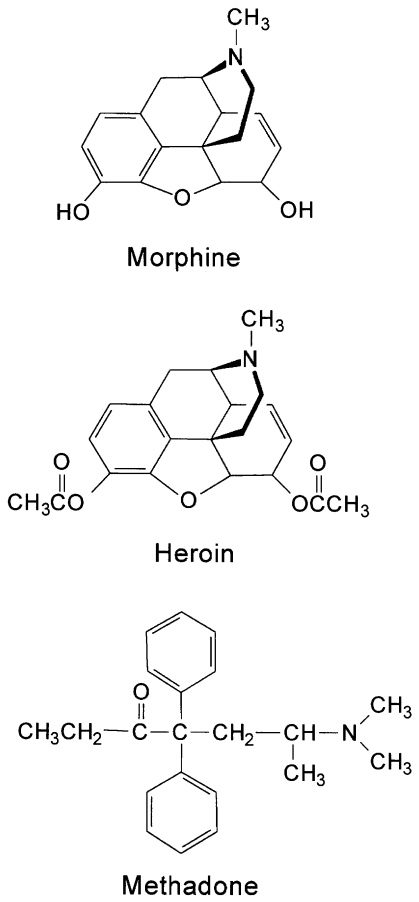


Figure 1 Structures of selected opiates.

depression in a dose-dependent manner and at sufficiently high doses breathing stops. The potential of respiratory arrest is a concern with very high doses of the drugs in naive patients, but typically it is not an issue with patients chronically on the drugs due to the development of tolerance to the action.

Sedation is also a common effect of morphine and related drugs. Indeed, morphine was initially named for Morpheus, the god of sleep. Like other opioid actions, tolerance will develop to sedation, but it may develop more slowly and to a lesser degree than tolerance to the analgesic actions of the drug. Opioids have many other actions. Their effects on the endocrine system are extensive and recent studies have also implicated them in immune function.

Morphine and related opiates produce their actions by activating receptors in the brain, mimicking a family of peptides with similar actions that are naturally occurring within the nervous system, termed the endorphins (Table I). The concept of peptide neurotransmitters has expanded dramatically, the endorphins are now only one family of many active endogenous peptides. This article discusses the four major groupings of opioid and opioid-like peptides and their receptors and also their physiological and pharmacological importance.

II. THE OPIOID AND RELATED PEPTIDES

Soon after the initial description of opioid binding sites using morphine-like radioligands, several laboratories identified endogenous peptides within the brain that bind to these opiate receptors (Table I). These peptides were then termed "endorphins," connotating their endogenous morphine-like character. Although their actions are likely to be very diverse, the only ones examined in detail involve analgesia and all the opioid peptides are active analgesics.

A. Enkephalins

The first peptides identified were the two enkephalins (Table I). These pentapeptides shared the same first four amino acids, differing only by a leucine or methionine at the fifth position. They are the endogenous ligand for the delta opioid receptor. The enkephalins are present throughout the brain and have been implicated in many actions. It is interesting that the enkephalins are present at high concentrations in the adrenal medulla, where they are colocalized with adrenalin. They also have been identified in other, nonneuronal tissues, including immune cells and the testis.

B. Dynorphins

Additional peptides with opioid-like actions were subsequently isolated. Dynorphin A is a heptadecapeptide that contains the sequence of [Leu^5]enkephalin at its amino terminus (Table I). Dynorphin A is the endogenous ligand for the kappa$_1$ receptor, although it retains high affinity for mu and delta receptors as well. Its actions were difficult to evaluate until highly selective, stable drugs were synthesized. These agents have established an important role for dynorphin A and its receptors in pain perception.

Table I
Opioid and Related Peptides

[Leu^5]enkephalin	**Tyr-Gly-Gly-Phe-Leu**
[Met^5]enkephalin	**Tyr-Gly-Gly-Phe-Met**
Peptide E (amidorphin)	**Tyr-Gly-Gly-Phe-Met-**Lys-Lys-Met-Asp-Glu-Leu-Tyr-Pro-Leu-Glu-Val-Glu-Glu-Glu-Ala-Asn-Gly-Gly-Glu-Val-Leu
BAM 22	**Tyr-Gly-Gly-Phe-Met-**Lys-Lys-Met-Asp-Glu-Leu-Tyr-Pro-Leu-Glu-Val-Glu-Glu-Glu-Ala-Asn-Gly-Gly
BAM 20	**Tyr-Gly-Gly-Phe-Met-**Lys-Lys-Met-Asp-Glu-Leu-Tyr-Pro-Leu-Glu-Val-Glu-Glu-Glu-Ala-Asn
BAM 18	**Tyr-Gly-Gly-Phe-Met-**Lys-Lys-Met-Asp-Glu-Leu-Tyr-Pro-Leu-Glu-Val-Glu-Glu-Glu
BAM 12	**Tyr-Gly-Gly-Phe-Met-**Lys-Lys-Met-Asp-Glu-Leu-Tyr
Metorphamide	**Tyr-Gly-Gly-Phe-Met-Arg-Val**
Dynorphin A	**Tyr-Gly-Gly-Phe-Leu-**Arg-Arg-Ile-Arg–Pro-Lys–Leu-Lys-Trp-Asp-Asn-Gln
Dynorphin B	**Tyr-Gly-Gly-Phe-Leu-**Arg-Arg-Gln-Phe-Lys-Val-Val-Thr
α-Neoendorphin	**Tyr-Gly-Gly-Phe-Leu-**Arg-Lys-Tyr-Pro-Lys
β-Neoendorphin	**Tyr-Gly-Gly-Phe-Leu-**Arg-Lys-Tyr-Pro
β_h-Endorphin	**Tyr-Gly-Gly-Phe-Met-**Thr-Ser-Glu-Lys-Ser-Gln-Thr-Pro-Leu-Val-Thr-Leu-Phe-Lys-Asn-Ala-Ile-Ile-Lys-Asn-Ala-Tyr-Lys-Lys-Gly-Glu
Endomorphin-1	Tyr-Pro-Trp-Phe-NH_2
Endomorphin-2	Tyr-Pro-Phe-Phe-NH_2
Orphanin FQ/nociceptin	**Phe-Gly-Gly-Phe-**Thr-Gly-Ala-Arg-Lys-Ser-Ala-Arg-Lys-Leu-Ala-Asp-Glu
Orphanin FQ2	Phe-Ser-Glu-Phe-Met-Arg-Gln-Tyr-Leu-Val-Leu-Ser-Met-Gln-Ser-Ser-Gln
Nocistatin	Thr-Glu-Pro-Gly-Leu-Glu-Glu-Val-Gly-Glu-Ile-Glu-Gln-Lys-Gln-Leu-Gln

C. β-Endorphin

The third member of the endogenous opioid family is β-endorphin, a 31-amino acid peptide that is localized primarily in the pituitary and the cells of the arcuate nucleus within the brain. The initial amino terminus of β-endorphin is identical to that of [Met^5]enkephalin, raising the question as to whether the enkephalins might simply be degradation products of longer peptides. However, this is not the case, as clearly demonstrated with the cloning of the peptide precursors.

The enkephalins and dynorphins are rapidly broken down by peptidases. Their extreme lability leads to very short durations of action, hampering early work on their pharmacology. However, substituting the glycine at the second position with a D-amino acid stabilizes the enkephalins and many stable derivatives have now been synthesized. The selectivity of these synthetic peptides for the opioid receptor classes can also be dramatically affected by their amino acid sequences. β-Endorphin, on the other hand, is more stable with a more prolonged duration of action. When steps are taken to minimize degradation, all three families of peptides share many actions, including the ability to produce analgesia. These actions are reversed by opioid-selective antagonists, confirming an opioid mechanism of action for the peptides.

D. Processing of Opioid Peptide Precursors

All these opioid peptides are generated by processing longer precursor proteins, which have been subsequently cloned. There are three distinct genes responsible for generating these peptides (Fig. 2).

1. Pre-Proenkephalin

The pre-proenkephalin gene contains four copies of Met-enkephalin and one copy each of Leu-enkephalin, the heptapeptide Met-enkephalin-Arg-Phe, and the octapeptide Met-Enkephalin-Arg-Gly-Leu (Fig. 2A). No dynorphin sequences are present within this precursor. However, the sequence of the gene predicts many additional putative peptides that also contain the sequence of met-enkephalin as the N terminus and thus might have opioid activity and be physiologically important. For example, peptide F contains [Met^5]enkephalin sequences at both its N terminus and its C terminus. The carboxy-terminally amidated peptide

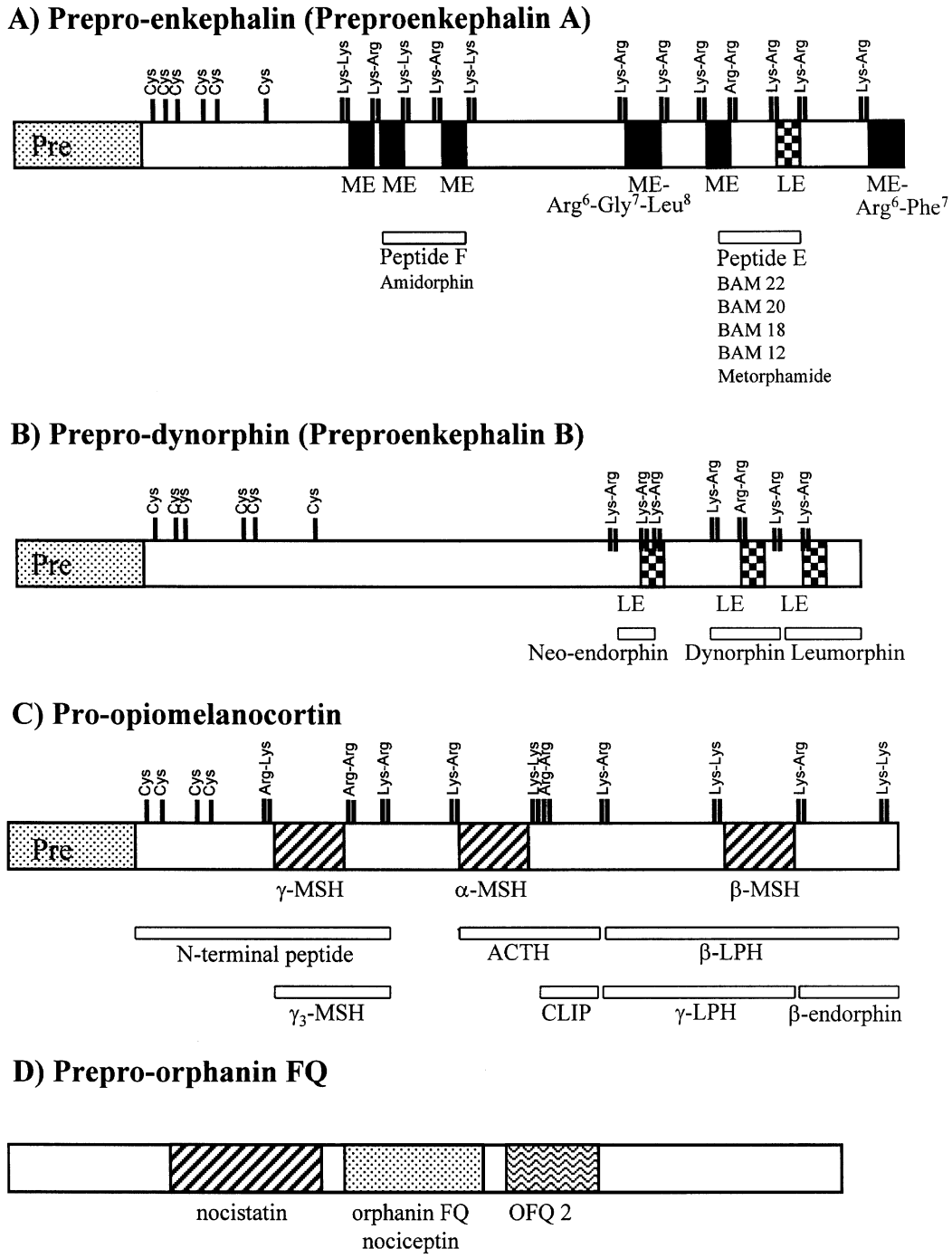


Figure 2 Structures of the precursors of the opioid peptides. Schematics of the precursor proteins for (A) the enkephalins, (B) the dynorphins, (C) β-endorphin, and (D) orphanin FQ/nociceptin.

comprising the first 26 amino acids of peptide F is named amidorphin. Peptide E has a [Met^5]enkephalin sequence at its N terminus and that of [Leu^5]enkephalin at its C terminus. Several peptides in which peptide E has been truncated at the C terminus have also been isolated: BAM 22, BAM 20, BAM 18, BAM 12, as well as an amidated octapeptide corresponding to the first 8 amino acids of peptide E termed metorphamide or

adrenorphin. The physiological significance of these additional peptides remains unclear. They may represent distinct neuropeptides with their own actions. However, the presence of dibasic amino acids following the enkephalin sequence in most of them raises the possibility that they might simply be further processed to the enkephalins. Although they were described many years ago, little work has been reported on these compounds for many years.

2. Pre-Prodynorphin

The pre-prodynorphin gene encodes a larger precursor that has many additional putative opioid peptides (Fig. 2B). The dynorphin precursor is quite distinct from the enkephalin precursor. It contains three [Leu^5]enkephalin sequences, each flanked by pairs of basic amino acids. If Lys–Arg pairs were the only processing signals, pre-prodynorphin would be cleaved into three larger opioid peptides: β-neoendorphin, dynorphin A, and leumorphine. However, several other peptides derived from pre-prodynorphin have been identified. Thus, the formation of dynorphin B results from the cleavage of leumorphin, whereas dynorphin A (1–8) is generated from dynorphin A. In addition, several larger peptides have been identified as putative processing products: a peptide containing dynorphin A at the N terminus and dynorphin B at the C-terminal end have been isolated. Dynorphin 24 contains dynorphin A with a C-terminal extension of Lys–Arg and the sequence of Leu–Enk. There is also evidence for a peptide comprising dynorphin A and leumorphin. In addition to dynorphin A (1–17), the truncated dynorphin A (1–8), and dynorphin B, this precursor also generates α-neoendorphin and β-neoendorphin. Again, the significance of these different peptides remains uncertain. Pharmacologically, they have opiate-like actions, but their physiologically relevance has not been proven.

3. Pre-Opiomelanocortin

β-Endorphin has the most interesting precursor peptide, pre-opiomelanocortin (Fig. 2C). Unlike the other opioid precursor peptides, the β-endorphin precursor makes many important, biologically active peptides that are not related to the opioid family. The precursor for β-endorphin also generates ACTH, an important stress hormone, α-melanocyte-stimulating hormone (MSH), and β-MSH. The association of β-endorphin with stress hormones is intriguing in view of the many associations between stress and a diminished perception of pain. In the pituitary, stimuli that release ACTH also release β-endorphin.

E. Endomorphins

For many years, the endogenous ligand for the mu opioid receptor was unknown. The recent identification of two tetrapeptides, endomorphin 1 and endomorphin 2, has resolved this issue. These peptides have a very interesting pharmacology and display the selectivity expected of a morphine-like, or mu, opioid peptide. It is assumed that the two peptides are also generated from a larger precursor, but this protein has not been identified. Although the evidence supporting the endomorphins is quite strong, the identification of its precursor is needed to fully establish its importance pharmacologically.

F. Orphanin FQ/Nociceptin

Recently, an opioid-related peptide has been identified termed orphanin FQ or nociceptin (OFQ/N). OFQ/N is the endogenous ligand for the fourth member of the opioid receptor gene family, ORL-1. Although it has some similarities to the traditional opioid peptides, it also has many significant differences. Like dynorphin A, OFQ/N is a heptadecapeptide (Table I) and its first four amino acids (Phe-Gly-Gly-Phe-) are similar to those of the traditional opioid peptides (Tyr-Gly-Gly-Phe-). However, OFQ/N has very poor affinity for the traditional opioid receptors, whereas the opioid peptides show no appreciable affinity for the ORL-1 receptor. Functionally, OFQ/N is also quite distinct from the opioid peptides. Although high doses of OFQ/N can elicit analgesia, low doses given supraspinally functionally reverse the opioid analgesia. The precursor for OFQ/N has been cloned, and it too contains additional putative neuropeptides, including nocistatin and OFQ2 (Fig. 2D).

III. OPIOID RECEPTORS

The opioid receptors were first demonstrated in 1973 with the binding of radiolabeled opiates to brain membranes. To date, three distinct families of opioid receptors have been identified that share many features. On the molecular level, they all are members of the G protein-coupled receptor family, which is a large

family of receptors located within membranes. Members of this receptor family span the membrane seven times, providing both intracellular and extracellular domains (Fig. 3). The opioids and opioid peptides bind to the extracellular component of the receptor and activate G proteins inside the cell. The opioid receptors are almost exclusively coupled to inhibitory systems, primarily G_o and G_i. Each receptor is encoded by a separate gene, but there is high homology between them.

A. Mu Receptors

Morphine and most clinical drugs act though the mu opiate receptors, making these receptors particularly important. Morphine has many actions, including analgesia, respiratory depression, and constipation. Early studies raised the possibility that these actions may be mediated though different subtypes of mu opioid receptors based on results using highly selective opioid antagonists. Antagonists have been developed that selectively block morphine analgesia without influencing respiratory depression and the inhibition of gastrointestinal transit. Another antagonist has also been reported that can reverse the actions of heroin without interfering with those of morphine. Thus, there is extensive pharmacological evidence for multiple subtypes of mu receptors.

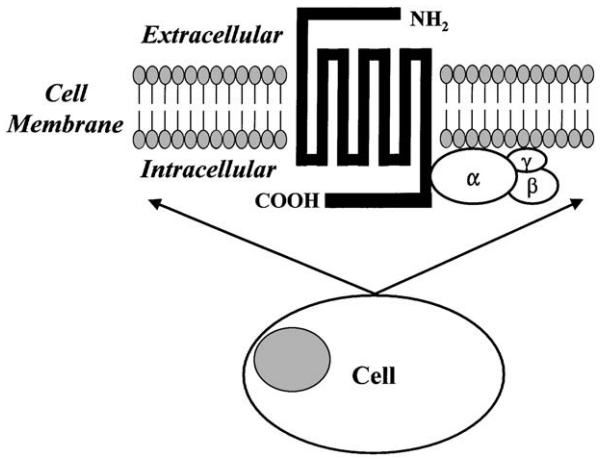


Figure 3 Structure of G protein receptors. G protein-coupled receptors span the cell membrane seven times, with the amino terminus located extracellularly and the carboxy terminus inside the cell. They are coupled to G proteins, which are composed of three subunits (α, β, and γ) and located on the inside surface of the cell. When the receptor is activated, it changes the G proteins, which then influence transduction systems.

Recently, a mu opioid receptor gene has been identified, *MOR-1* (Fig. 4). Additional cloning work has isolated seven different splice variants of this gene, implying the existence of at least seven different mu opioid receptor subtypes at the molecular level. Thus, the molecular biological approaches have uncovered far more mu receptor subtypes than even suggested from pharmacological studies. Correlating these variants with specific opioid actions is not simple. We know that a major disruption of the *MOR-1* gene eliminates all mu analgesia. However, more subtle disruptions of the gene leading to the elimination of only some *MOR-1* variants block the analgesic actions of some, but not all, mu analgesics. At this point, it appears that most mu analgesics will activate more than one mu receptor subtype. Thus, differences among the mu analgesics may reflect differences in their pattern of receptor activation.

B. Delta Receptors

Delta receptors were first found following the discovery of the enkephalins, their endogenous ligand. Early attempts to define the pharmacology of the delta receptors were limited by the lability of the enkephalins and their limited selectivity. The development of highly selective delta peptides, and recently organic compounds, has revealed important actions. Like mu receptors, activation of delta receptors produces analgesia but with far less respiratory depression and constipation. There is evidence for subtypes of delta receptors based on a series of antagonists, but functional splice variants have not been reported. The full functional significance of delta receptor systems is not entirely established.

C. Kappa$_1$ Receptors

Kappa receptors were first proposed from detailed pharmacological studies using a series of opiates long before the identification of its endogenous ligand, dynorphin A. Like the delta receptor, efforts to define kappa receptor pharmacology have been difficult due to the stability and selectivity of dynorphin A. The synthesis of highly selective ligands has helped to define the kappa$_1$ receptor, both biochemically and pharmacologically. Kappa$_1$ receptors can produce analgesia, but they do so through mechanisms different from those of the other receptor classes. However, kappa receptor ligands have also been associated with

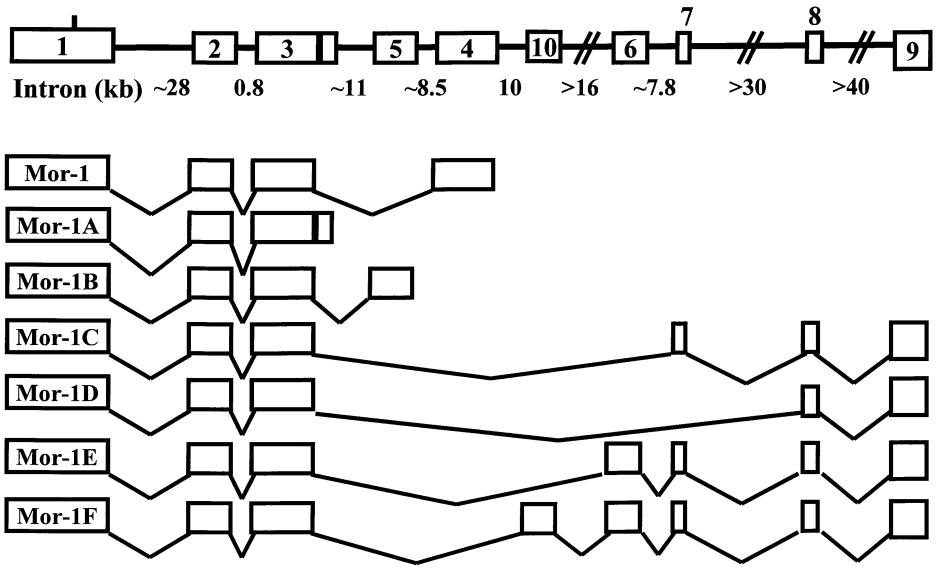


Figure 4 The murine *MOR-1* gene has been extensively studied. Ten different exons, the DNA sequences that end up in mRNA, are spliced together in different patterns to make seven different mu receptor variants. Although the human gene has not been explored as completely, there is little reason to believe that it differs significantly.

a variety of psychomimetic effects, making their clinical utility somewhat limited.

A kappa receptor with the appropriate pharmacological profile has been cloned, KOR-1. It has high homology with the other receptors and is localized within the brain. However, $kappa_1$ receptors also appear to be present in a variety of immune cells, as demonstrated by binding and molecular biological approaches.

D. The Opioid Receptor-like Receptor

A fourth member of the opiate receptor family has been cloned that is highly selective for OFQ/N. As noted previously, OFQ/N has many similarities to the opioid peptides, but it has very poor affinity for the traditional opioid receptors. Similarly, the opioid peptides do not label the ORL-1 receptor. ORL-1 has an interesting pharmacology. Evidence from both binding and pharmacological studies has suggested multiple subtypes of receptors.

The relationship of ORL-1 to the opiate family is unusual. Although there is high homology at the molecular level, traditional opiates do not bind to this site, and its endogenous ligand, OFQ/N, has poor affinity for all the opiate sites as well. It is interesting, however, that OFQ/N is also a heptadecapeptide. Functionally, the ORL-1 receptor has some unusual actions. Depending on the dose and the site of administration, OFQ/N can be a potent antiopioid peptide, functionally reversing the actions of morphine and other opioid analgesics. However, at higher doses the peptide is analgesic. Antisense mapping studies have suggested that the $kappa_3$ receptor is related to the ORL-1 receptor, but they are not identical. Thus, the pharmacology of OFQ/N and its receptor is quite complex and many issues remain to be evaluated.

E. $Kappa_2$ and $Kappa_3$ Receptors

Several other kappa receptor classes have been proposed. U50,488H is a potent and highly selective $kappa_1$ receptor agonist. Binding studies revealed additional kappa receptor binding sites insensitive to U50,488H. The first site was termed $kappa_2$ to distinguish it from the U50,488H-sensitive $kappa_1$ site. The difficulty with defining this site pharmacologically resulted from the lack of highly selective ligands. Recently, it has been suggested that the $kappa_2$ receptor may actually represent a dimer consisting of a $kappa_1$ and a delta receptor. Together, the receptors display binding characteristics quite distinct from either $kappa_1$ or delta receptors alone and may correspond to the $kappa_2$ receptor originally observed in binding studies.

The kappa$_3$ receptor was originally proposed using naloxone benzoylhydrazone (NalBzoH), which is a potent analgesic. Its actions are not reversed by highly selective mu, delta or kappa$_1$ receptor antagonists, giving it a very unique selectivity profile. Recently, its actions have been associated with the ORL-1 receptor. Antisense mapping studies revealed that KOR-3, the cloned mouse homolog of the ORL-1 receptor, and the kappa$_3$ receptor are both encoded by the same gene but are not identical and may be splice variants of the *KOR-3*/*ORL-1* gene. The exact relationship between the kappa$_3$ receptor and the *KOR-3*/*ORL-1* gene has not been completely defined.

IV. CONCLUSION

The opioid system, composed of a family of receptors and their endogenous peptide ligands, is quite important within the central nervous system. It has many functions, some of which are readily demonstrated. However, the complexity of the system is quite extensive and a full understanding is not available. Many questions remain, including the role of the many unexplored opioid peptides postulated to be generated within the brain. Furthermore, splicing appears to play a major role in generating multiple subtypes of the various opiate receptor families. Although our knowledge of this system is growing, many unknowns remain.

See Also the Following Articles

CHEMICAL NEUROANATOMY • DOPAMINE • GABA • PAIN • PEPTIDES, HORMONES, AND THE BRAIN AND SPINAL CORD • PSYCHOACTIVE DRUGS • PSYCHONEUROENDOCRINOLOGY • RESPIRATION • STRESS: HORMONAL AND NEURAL ASPECTS

Suggested Reading

Evans, C. J., Hammond, D. L., and Frederickson, R. C. A. (1988). The opioid peptides. In *The Opiate Receptors* (G. W. Pasternak, Ed.), p. 23. Humana Press, Clifton, NJ.

Hughes, J., Smith, T. W., Kosterlitz, H. W., Fothergill, L. A., Morgan, B. A., and Morris, H. R. (1975). Identification of two related pentapeptides from the brain with potent opiate agonist activity. *Nature* **258,** 577.

Pan, Y. X., Xu, J., Bolan, E. A., Abbadie, C., Chang, A., Zuckerman, A., Rossi, G. C., and Pasternak, G. W. (1999). Identification and characterization of three new alternatively spliced mu opioid receptor isoforms. *Mol. Pharmacol.* **56,** 396.

Reisine, T., and Pasternak, G. W. (1996). Opioid analgesics and antagonists. In *Goodman & Gilman's: The Pharmacological Basis of Therapeutics* (J. G. Hardman and L. E. Limbird, Eds.). McGraw-Hill, New York.

Snyder, S. H. (1977). Opiate receptors and internal opiates. *Sci. Am.* **236,** 44.

Epilepsy

BETTINA SCHMITZ
Humboldt University, Berlin

GLOSSARY

antiepileptic drugs (anticonvulsants) Drugs with different modes of action that prevent epileptic seizures.

epilepsy surgery A surgical treatment that includes the removal or disconnection of epileptogenic tissue and that aims at complete seizure control.

focal (partial) seizures Epileptic seizures with clinical or electroencephalographic changes indicating initial activation of neurons limited to one part of one cerebral hemisphere.

generalized seizures Epileptic seizures in which the first clinical changes indicate involvement of both hemispheres.

hippocampal sclerosis Typical pathology in mesial temporal lobe epilepsy characterized by loss of neurons in the CA1 region and endfolium (CA3/4) but with relative sparing of the C2 region.

Epilepsy is a heterogeneous condition characterized by recurrent seizures. Because the epileptic seizure is a nonspecific neurological symptom, it is important to distinguish epilepsy from isolated seizures and acute symptomatic seizures provoked by acute systemic or cerebral disorders. Epilepsy is one of the most common neurological disorders, with peak prevalence rates in early childhood and old age. Two international classification systems are used to differentiate seizure types and epileptic syndromes. The major distinction is between localization-related and generalized seizures. The former are most often symptomatic and caused by a circumscribed lesion, and the latter are typically idiopathic and associated with an inherited imbalance in excitatory and inhibitory mechanisms. There may be an association with learning disability or other neurological problems, particularly in symptomatic epilepsies. Idiopathic epilepsies are characterized by a specific age of onset, a good response to drug treatment, and a relatively benign course. The genes responsible for some rare syndromes have recently been identified; however, the suspected genetic background of the more common idiopathic epilepsies is complex and likely involves multiple susceptibility genes. The differential diagnosis of epilepsy includes a large spectrum of medical, neurological, and psychiatric disorders. The diagnosis of epilepsy is primarily based on history. Technical investigations include electroencephalography, functional and structural imaging techniques, and neuropsychological assessments. Treatment of epilepsies involves anticonvulsant drugs and behavioral recommendations and also surgery in suitable patients who are resistant to conservative strategies. Anticonvulsant drugs can be categorized into three groups according to their mode of action: GABAergic drugs, antiglutamatergic drugs, and membrane-stabilizing drugs that act through the modulation of the sodium channel. The choice of antiepiletic drugs depends on the type of epilepsy. The aim of drug treatment is complete seizure control without causing any clinical side effects.

Complications of active epilepsy are not only injuries but also psychiatric disorders such as depression. The social impact of epilepsy is more significant than with most other neurological disorders because of stigmatization and discrimination of seizure disorders. Professional training and employment are further limited by restrictions with respect to driver's licenses, working night shifts, and handling potentially dangerous machines.

I. CLASSIFICATION

How best to classify epileptic seizures and epileptic syndromes has occupied epileptologists for centuries. In the middle of the 19th century the epilepsy literature was dominated by psychiatrists who tried to systematize epilepsy based on their experiences with chronic patients in asylums. They spent more time on the psychopathological classification of psychiatric symptoms in epilepsy than on describing epileptic seizures. Specific psychiatric syndromes, especially short-lasting episodic psychoses and mood disorders, were regarded as being of equal diagnostic significance for epilepsy as convulsions.

With the introduction of electroencephalography (EEG) in the 1930s many episodic psychiatric states were identified as nonepileptic in origin. Epileptic seizures, however, showed very different ictal EEG patterns. Since then, epileptologists have concentrated on the electroclinical differentiation of epileptic seizures. Problems of terminology became obvious with increasing communication between international epileptologists in the middle of the 20th century. The first outlines of international classifications of epileptic seizures and epileptic syndromes were published in 1970. The proposed classifications by the Commission on Classification and Terminology of the International League against Epilepsy (ILAE) from 1981 and 1989 (Tables I and II) are based on agreements among international epileptologists and compromises between various viewpoints. They must not be regarded as definitive; a revision is currently being developed.

A. Classification of Seizures

Most authors in the 19th century simply differentiated seizures according to severity ("petit mal" and "grand mal"). Hughlings Jackson was the first to recognize the need for an anatomical description, physiological delineation of disturbance of function, and pathological confirmation. In the 20th century clinical events could be linked to ictal electroencephalographic findings. More detailed analysis of seizures became possible with simultaneous EEG video monitoring and ictal neuropsychological testing. In the former, patients are monitored by video cameras and continuous EEG recordings for prolonged periods. The data can be simultaneously displayed on a split-screen television monitor (Fig. 1).

The International Classification of Epileptic Seizures (ICES) from 1981 is based on this improved monitoring capability, which has permitted more accurate recognition of seizure symptoms and their longitudinal evolution. The current classification of seizures is clinically weighted and gives no clear definitions in terms of seizure origin. The ICES does not reflect our most recent understanding of the localizing significance of specific seizure symptoms, which has grown significantly since 1981 due to increased data from intensive monitoring and epilepsy surgery. Some parts of the ICES are therefore outdated and in need of revision. Complex focal seizure types, for example, are not yet distinguished according to a probable origin in the frontal or the temporal lobe.

The principal feature of the ICES is the distinction between seizures that are generalized from the beginning and those that are partial or focal at onset and may or may not evolve to secondary generalized seizures (Table I). In generalized seizures there is initial involvement of both hemispheres, reflecting an epileptogenic generator in subcortical structures. Consciousness may be impaired and this impairment may be the initial manifestation. Motor manifestations are bilateral. The ictal EEG patterns are initially bilateral. Spikes, spike–wave complexes, and polyspike–wave complexes are all typical (Fig. 2).

Focal seizures are those in which the first clinical and EEG changes indicate initial activation of a system of neurons limited to a part of one cerebral hemisphere. The other important feature of the ICES is the separation between simple and complex partial seizures depending on whether there is preservation or impairment of consciousness.

B. Classification of Syndromes

"Epilepsy is the name for occasional sudden, excessive, rapid, and local discharges of the gray matter." This simple definition was formulated by Jackson in 1866

Table I
International Classification of Epileptic Seizures[a]

Clinical seizure type	Ictal EEG
Focal (partial, local) seizures	
A. Simple partial seizures	Local contralateral discharge starting over corresponding area of cortical representation (not always recorded on the scalp)
1. With motor symptoms	
a. Focal motor without march	
b. Focal motor with march (Jacksonian)	
c. Versive	
d. Postural	
e. Phonatory (vocalization or arrest of speech)	
2. With somatosensory or special sensory symptoms (simple hallucinations, e.g., tingling, light flashes, and buzzing)	
a. Somatosensory	
b. Visual	
c. Auditory	
d. Olfactory	
e. Gustatory	
f. Vertiginous	
3. With autonomic symptoms or signs (including epigastric sensation, pallor, sweating, flushing, piloerection, and pupillary dilatation)	
4. With psychic symptoms (disturbance of higher cortical function); these symptoms rarely occur without impairment of consciousness and are much more commonly experienced as complex partial seizures.	
a. Dysphasic	
b. Dysmnesic (e.g., déja vu)	
c. Cognitive (e.g., dreamy states and distortions of time sense)	
d. Affective (fear, anger, etc.)	
e. Illusions (e.g., macropsia)	
f. Structured hallucinations (e.g., music and scenes)	
B. Complex focal seizures (with impairment of consciousness; may sometimes begin with simple symptomatology)	Unilateral or frequently bilateral discharge; diffuse or focal in temporal or frontotemporal regions
1. Simple partial onset followed by impairment of consciousness	
a. With simple partial features (as in A, 1–4) followed by impaired consciousness	
b. With automatisms	
2. With impairment of consciousness at onset	
a. With impairment of consciousness only	
b. With automatisms	
C. Focal seizures evolving to secondarily generalized seizures (this may be generalized tonic–clonic, tonic, or clonic)	Above discharge becomes secondarily and rapidly generalized
1. Simple partial seizures (A) evolving to generalized seizures	
2. Complex partial seizures (B) evolving to generalized seizures	
3. Simple focal seizures evolving to complex focal seizures evolving to generalized seizures	

(continues)

Table I *(continued)*

Clinical seizure type	Ictal EEG
Generalized seizures	
A. Absence seizures	Usually regular and symmetrical 3-Hz but may be 2–4 Hz spike and slow-wave complexes and may have multiple spike and slow-wave complexes; abnormalities are bilateral
1. Absence seizures	
a. Impairment of consciousness only	
b. With mild clonic components	
c. With atonic components	
d. With tonic components	
e. With automatisms	
f. With autonomic components	
2. Atypical absence a. Changes in tone that are more pronounced than in A.1 b. Onset and/or cessation that is not abrupt	EEG more heterogeneous; may include irregular spike and slow-wave complexes, fast activity, or other paroxysmal activity; abnormalities are bilateral but often irregular and asymmetric
B. Myoclonic seizures, myoclonic jerks (single or multiple)	Polyspike and wave, or sometimes spike and wave or sharp and slow waves
C. Clonic seizures	Fast activity (10 c/sec or more) and slow waves; occasional spike and wave patterns
D. Tonic seizures	Low-voltage, fast activity or a fast rhythm of 9–10 c/sec or more, decreasing in frequency and increasing in amplitude
E. Tonic–clonic seizures	Rhythm at 10 or more c/sec decreasing in frequency and increasing in amplitude during tonic phase, interrupted by slow waves during the clonic phase
F. Atonic seizures	Polyspikes and wave or flattering or low–voltage fast activity

[a]From the Commission on Classification and Terminology of the International League against Epilepsy (1981).

long before the introduction of electroencephalography. It has not lost its justification today. Recurrent epileptic seizures are pathognomonic for all types of epilepsies. The clinical spectrum of epilepsy, however, is much more complex and an epileptic syndrome is characterized by a cluster of signs and syndromes customarily occurring together; these include such features as type of seizure, etiology, structural lesions, precipitating factors, family history, age of onset, severity, chronicity, diurnal and circadian cycling, and prognosis.

In contrast to a syndrome, a disease is characterized by a specific etiology and prognosis. Some recognized entities in the International Classification of Epilepsies and Epileptic Syndromes (ICEES) are diseases and others are syndromes, some of which may turn out to be diseases—a specific etiology may still be discovered.

The ICEES distinguishes generalized and localization-related (focal, local, and partial) epilepsies. Generalized epilepsies are syndromes characterized by generalized seizures in which there is involvement of both hemispheres from the beginning of the seizure. Seizures in localization-related epilepsies start in a circumscribed region of the brain. The other important classification criterion refers to etiology. The ICEES distinguishes idiopathic, symptomatic, and cryptogenic epilepsies. Idiopathic means that a disease is not preceded or occasioned by another. The major pathogenetic mechanism is genetic predisposition. Symptomatic epilepsies and syndromes are considered the consequence of a known or suspected disorder of the central nervous system. In cryptogenic disorders, a cause is suspected but remains obscure, often due to limited sensitivity of diagnostic techniques.

Table II
International Classification of Epilepsies and Epileptic Syndromes[a]

- Localization-related (focal, local, and partial) epilepsies and syndromes
  - Idiopathic (with age-related onset)
  - Currently, the following syndromes are established, but more may be identified in the future:
    - Benign childhood epilepsy with centrotemporal spikes
    - Childhood epilepsy with occipital paroxysms
    - Primary reading epilepsy
  - Symptomatic
    - Chronic progressive epilepsia partialis continua of childhood (Kozhevnikov's syndrome)
    - Syndromes characterized by seizures with specific modes of precipitation
    - Temporal lobe epilepsy
      - With amygdala–hippocampal seizures
      - With lateral temporal seizures
    - Frontal lobe epilepsy
      - With supplementary motor seizures
      - With cingulate seizures
      - With seizures of the anterior frontopolar region
      - With orbitofrontal seizures
      - With dorsolateral seizures
      - With opercular seizures
      - With seizures of the motor cortex
    - Parietal lobe epilepsies
    - Occipital lobe epilepsies
- Generalized epilepsies and syndromes
  - Idiopathic, with age-related onset, listed in order of age
    - Benign neonatal familial convulsions
    - Benign neonatal convulsions
    - Benign myoclonic epilepsy in infancy
    - Childhood absence epilepsy (pyknolepsy)
    - Juvenile absence epilepsy
    - Juvenile myoclonic epilepsy (impulsive petit mal)
    - Epilepsy with grand mal seizures (GTCS) on awakening
    - Other generalized idiopathic epilepsies not defined previously
    - Epilepsies precipitated by specific modes of activation
  - Cryptogenic or symptomatic (in order of age)
    - West syndrome (infantile spasms, Blitz–Nick–Salaam Krämpfe)
    - Lennox–Gastaut syndrome
    - Epilepsy with myoclonic–astatic seizures
    - Epilepsy with myoclonic absences
  - Symptomatic
    - Nonspecific etiology
    - Early myoclonic encephalopathy
    - Early infantile epileptic encephalopathy with suppression burst
    - Other symptomatic generalized epilepsies not defined previously
  - Specific syndromes
    - Epileptic seizures may complicate many disease states. Under this heading are included diseases in which seizures are a presenting or predominant feature.
- Epilepsies and syndromes undetermined as to whether they are focal or generalized
  - With both generalized and focal seizures
    - Neonatal seizures
    - Severe myoclonic epilepsy in infancy
    - Epilepsy with continuous spike waves during slow-wave sleep
    - Acquired epileptic aphasia (Landau–Kleffner syndrome)
    - Other undetermined epilepsies not defined previously
  - Without unequivocal generalized or focal features

[a]From the Commission on Classification and Terminology of the International League against Epilepsy (1989).

1. Localization-Related Epilepsies and Syndromes

Seizure symptomatology and ictal EEG findings are the most important criteria for anatomical classification of localization-related epilepsies. As mentioned, the 1981 version of the ICES, with its emphasis on the formal structure of seizures, is of limited help in the detailed localization of seizures required in the ICEES from 1989. The 1989 classification has been criticized because reliable localization often requires invasive EEG techniques. It is hoped that with more data from taxonomic studies, clinical symptoms or symptom clusters will be identified that eventually will allow classification on clinical grounds in most cases.

Temporal lobe epilepsies are classified as those with lateral temporal seizures with auditory hallucinations, language disorders (in the case of dominant hemisphere focus) or visual illusions, and those with amygdala–hippocampal (mesiobasal limbic or rhinencephalic) seizures. The latter are characterized by simple seizure symptoms, such as increasing epigastric discomfort, nausea, marked autonomic signs, and other symptoms including borborygmi, belching, pallor, fullness of the face, flushing of the face, arrest of respiration, pupillary dilatation, fear, panic, and

Figure 1 Video recording of a complex partial seizure with secondary generalization (frontal lobe epilepsy): (a) interictal normal behavior; (b) patient pushes alarm button; (c) patient describes aura of fear; (d) automatisms: hand rubbing; (e) alternating tapping of legs; (f) patient does not respond when asked to name a pair of glasses; (g) tonic elevation of right arm; (h) generalized myoclonic movements; (i) reorientation; (j) end of seizure, patient answers adequately.

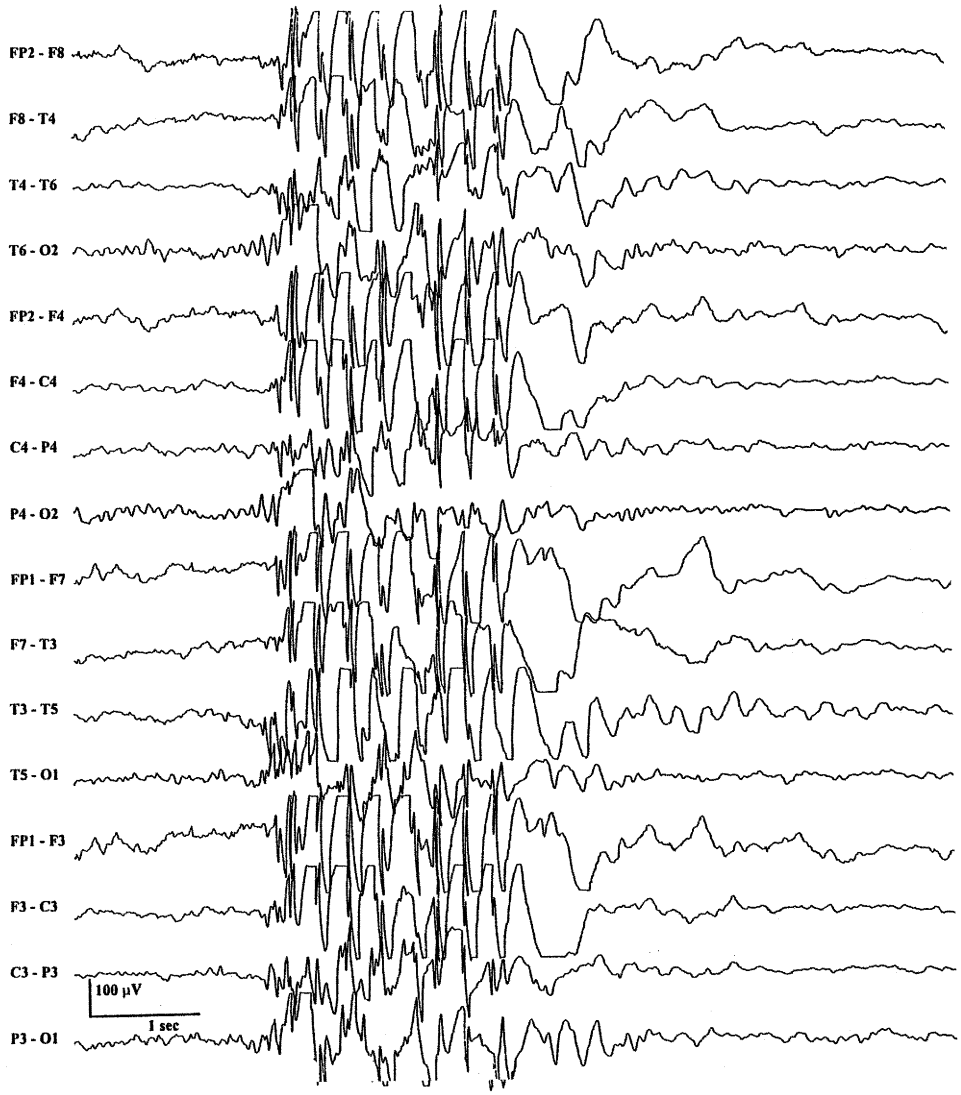


Figure 2 EEG in idiopathic generalized epilepsy (juvenile myoclonic epilepsy) with generalized polyspike–wave pattern.

olfactory hallucinations. Complex focal seizures often begin with a motor arrest, typically followed by oroalimentary automatisms. The duration is typically more than 1 min and consciousness recovers gradually.

Frontal lobe epilepsies are characterized by seizures of short duration, minimal or no postictal confusion, rapid secondary generalization, vocalization, prominent motor manifestations that are tonic or postural, complex gestural automatisms, and nocturnal clustering. Ictal scalp EEGs may show bilateral or multilobar discharges or are often normal. Accurate localization of frontal lobe epilepsies may therefore be difficult.

2. Generalized Epilepsies and Syndromes

Idiopathic generalized epilepsies are characterized by an age-related onset that reflects the ability of the brain to produce certain seizure types depending on cerebral maturation. In general, patients are asymptomatic between seizures. Radiological investigations are negative. Frequently, there is an overlap of idiopathic generalized epilepsies, especially of those manifesting in later childhood and adolescence. Response to antiepileptic drug treatment and psychosocial prognosis are good.

Symptomatic generalized epilepsies and syndromes usually start in infancy or early childhood. In most children several seizure types occur. EEG discharges are less rhythmical and less synchronous than in idiopathic generalized epilepsies. There are neurological, neuropsychological, and radiological signs of diffuse cerebral disease. The only difference between cryptogenic and symptomatic syndromes is that in cryptogenic syndromes the presumed cause cannot be identified.

3. Epilepsies and Syndromes Undetermined as to Whether They Are Focal or Generalized

There are two groups of patients who cannot be classified as focal or generalized. The first group consists of patients with both generalized and focal seizures (e.g., patients with both focal seizures and absence seizures). The second group comprises patients without unequivocal generalized or focal features (e.g., patients with nocturnal generalized tonic–clonic seizures).

4. Specific Syndromes

These are isolated seizures and situation-related syndromes, such as febrile seizures and seizures occurring only when there is an acute metabolic or toxic event due to factors such as alcohol, drugs, eclampsia, or hyperglycemia.

II. EPIDEMIOLOGY

A. Prevalence and Incidence

Epidemiological studies of epilepsy are often difficult to compare because of different study designs and definitions of epilepsy. Calculated incidence rates of epilepsy range between 20 and 70/100,000 per year. The incidence is age dependent, with a maximum in early childhood and lowest rates in early adulthood. Incidence rates rise again in older age groups, probably due to the higher prevalence of cerebrovascular disease. The overall risk for epilepsy is slightly higher in males than in females. The point prevalence of active epilepsy is of the order of 3–5/1000. The cumulative lifetime prevalence has been estimated to be 3.5%.

There is thus a 10-fold difference between the point prevalence and the lifetime prevalence, suggesting that the disease remains active in only a small proportion of cases. The prevalence of epilepsy is higher in third world countries, with rates of up to 37/1000 in Africa. This is probably related to a higher frequency of infectious diseases of the nervous system diseases such as neurocysticercosis and the higher risk of perinatal complications.

B. Relative Frequency of Epileptic Syndromes

Studies of the distribution of epileptic syndromes and epileptic seizures in the population have had conflicting results. In population-based studies, epilepsies with complex focal and focal seizures secondarily evolving to tonic–clonic seizures are most frequent, occurring in 69% of all patients. This is followed by primary generalized tonic–clonic seizures in 30% and absence or myoclonic seizures in less than 5%. In hospital-based studies, the numbers of subtle generalized seizures (absences and myoclonic seizures) are generally higher, almost certainly related to the increased accuracy and sensitivity of diagnosis.

C. Etiology

In most patients with epilepsy, the etiology remains unclear. According to an epidemiological survey in Rochester, New York, 65% of cases are idiopathic or cryptogenic. The most frequent cause of epilepsy is vascular disorder (10%), followed by congenital complications (8%), brain trauma (6%), tumor (4%), degenerative disorders (4%), and infectious diseases (3%). The etiology of epilepsy is largely dependent on the age of onset, with congenital disorders dominating in childhood epilepsies, trauma and tumors being most common in early adulthood, and cerebrovascular and degenerative disorders becoming more frequent with older age. Recent studies using advanced imaging techniques have revealed a high number of patients with subtle cortical dysgenesis or hippocampal sclerosis—patients who were formerly classified as cryptogenic.

III. PROGNOSIS

A. The Natural Course of Epilepsy

The natural course of untreated epilepsy is unknown. For obvious ethical reasons no one has ever conducted

a controlled study, and there are no systematic data from the days before antiepileptic drugs were introduced. Nineteenth century epileptologists emphasized the poor prognosis in epilepsy, but their experience was limited to institutionalized patients. Gowers (1885) believed that seizures may beget more seizures: "The tendency of the disease is to self-perpetuation, each attack facilitates the occurrence of another, by increasing the instability of nerve elements." Gowers studied the recurrence of seizures in 160 cases. A second seizure followed the first within 1 month in one-third of patients and within 1 year in two-thirds.

B. Recurrence Risk after a First Seizure

In an unselected sample of the population, isolated seizures occur in 20–40% of people with one or more seizures. It is methodologically extremely difficult to investigate the recurrence risk after an initial epileptic seizure. Estimates in the literature range between 27 and 71%, depending on inclusion criteria, duration of follow-up, and whether or not patients are treated after a first seizure. The most important source of error has been the exclusion of patients with early recurrences of seizures before presentation. The longer the interval between first seizure and inclusion into the study, the lower the recurrence rate.

The recurrence risk within 24 months of a first seizure calculated from a meta-analysis of 14 studies was 42%, with a higher risk in the group of patients with symptomatic seizures and pathological EEG findings (65%) compared to those patients with idiopathic seizures and a normal EEG (24%).

C. Prognostic Studies

Recent prospective and population-based studies have challenged previous views that epilepsy is likely to be a chronic disease in as many as 80% of cases. In a population-based survey in Rochester, 20 years after the initial diagnosis of epilepsy 70% of patients were in 5 years of remission and 50% of patients had successfully withdrawn medication. In an English study, 15 years after diagnosis 81% of patients were seizure free for at least 1 year. In another study of 104 patients who were followed up after onset of treatment, 60% were in 1-year remission after a follow-up period of 24 months. By 8 years of follow-up, 92% had achieved a 1-year remission. It is recommended that antiepileptic drugs be withdrawn after a minimum seizure-free period of 2 years or 5 years in severe cases. Relapses occur in 12–72% of patients after a 2-year remission and in 11–53% after a 3-year remission.

Approximately 5–10% of all patients with epilepsy eventually have intractable seizures despite optimal medication; most of these patients have complex partial seizures. Despite the overall favorable prognosis of epilepsy and the good response to treatment, the mortality rate of epilepsy is 2.3-fold higher than that in the general population, and it is 3.8-fold higher in the first years of the illness. The incidence of sudden unexpected death in epilepsy was estimated to be in the order of 1/525.

The prognosis of epilepsy largely depends on the syndromatic diagnosis. Idiopathic localization-related epilepsies such as Rolandic epilepsy have an excellent prognosis in all respects. Prognosis in terms of seizure remission, social adjustment, and life expectancy, on the other hand, is extremely poor in symptomatic generalized epilepsies such as West syndrome and in progressive myoclonic epilepsies.

Several studies have examined prognostic factors independent of the syndromatic diagnosis. Most studies have consistently shown that diffuse cerebral damage and neurological and cognitive deficits are associated with a poor outcome. There has been less agreement on the significance of other possible risk factors for poor prognosis, such as EEG features and positive family history of epilepsy. Whether early treatment and medical prevention of seizures improves the long-term prognosis as suggested by Gowers and indicated by experimental data from animal epilepsy models is not clear and subject to controversy. A recent Italian study showed that the treatment prognosis risk after a first seizure is not lower in patients treated with antiepiletic drugs (AEDs) compared to patients not treated after a first seizure.

IV. BASIC MECHANISMS

Epileptic syndromes are characterized by a tendency to paroxysmal regional or generalized hyperexcitability of the cerebral cortex. Because of the phenomenological diversity and the etiological heterogeneity of epilepsies, it is likely that there are multiple underlying cellular and molecular mechanisms.

The mechanisms responsible for the occurrence of seizures (ictogenesis) and the development of epilepsy (epileptogenesis) have been studied in animal models and by *in vitro* studies of surgical human brain tissue. The exact mechanisms remain to be clarified. They

represent complex changes of normal brain function on multiple levels, involving anatomy, physiology, and pharmacology. There are categorical differences in the pathophysiology of idiopathic generalized and symptomatic focal seizures. The latter are caused by a regional cortical hyperexcitation due to local disturbances in neuronal connectivity. Synaptic reorganization may be caused by any acquired injury or congenital abnormalities, such as in the many subforms of cortical dysplasia. Some areas of the brain seem more susceptible than others, the most vulnerable being mesial temporal lobe structures. The pathophysiology of hippocampal sclerosis, the most common etiology of temporal lobe epilepsy, has been a topic of much controversy. It is most likely that the hippocampal structures are damaged by an early trauma, typically prolonged febrile seizures. The process of hippocampal sclerosis involves a synaptic reorganization with excitotoxic neuronal loss, loss of interneurons, and GABA deficit.

Epileptic neurons in an epileptogenic focus may produce bursts of action potentials that represent isolated spikes in the EEG as long as they remain locally restricted. Depending on the failure of local inhibitory mechanisms, these bursts may lead to ongoing and repetitive discharges. If larger neuronal networks are recruited in this hypersynchronous activity, this leads to an epileptic seizure that is either focal or secondary generalized depending on the extent of seizure spread. The local seizure threshold is regulated by influences on excitatory and inhibitory postsynaptic potentials. The membrane excitability is regulated by ion channels that are modulated by excitatory transmitters, particularly glutamate and aspartate, and the inhibitory transmitter GABA. These three levels are the major targets for antiepileptic drugs (Table III).

Primary generalized seizures are accompanied by a bilateral synchronous epileptic activity. Therefore, they cannot be explained by cortical dysfunction alone. They are caused by an imbalance of pathways between the thalamus and the cerebral cortex. These thalamocortical circuits are modulated by reticular nuclei. Excessive GABAergic inhibition and dysfunction of thalamic calcium channels also play a role in the pathogenesis, at least in the pathogenesis of absence seizures. These cortical–subcortical circuits are also responsible for circadian rhythms, which may explain the increased seizure risk after awakening and following sleep withdrawal in idiopathic generalized epilepsies. The appearance and disappearance of primary generalized seizures is strongly age dependent, a phenomenon that has been explained by disturbances in brain maturation processes.

V. GENETICS

Genetic factors play a major role in the etiology of idiopathic epilepsies. Progress in molecular genetics has revealed several susceptibility loci and causative gene mutations in many rare human epilepsies with monogenic inheritance, such as the progressive myoclonus epilepsies. Mutations in the gene encoding the α_4 subunit of the neuronal nicotinic acetylcholine receptor gene have been identified as predisposing to autosomal-dominant nocturnal frontal lobe epilepsy. Mutations of two genes encoding potassium channels cause benign neonatal familial convulsions. A gene encoding the β subunit of the voltage-gated sodium channel has been identified in a rare syndrome called generalized epilepsy with febrile seizures plus. These findings suggest that at least some epilepsies are related to ion channel dysfunction.

Table III
Mode of Action of Old and New Antiepileptic Drugs

	Sodium channel[a]	GABA[b]	Glutamate[c]	Calcium channel[d]
Old AED				
Benzodiazepines	+	+ +	0	0
Carbamazepine	+ +	±	±	?
Ethosuximide	0	0	0	+
Phenobarbitone	+	+	+	0
Phenytoin	+ +	±	±	0
Valproate	+ +	+	±	0
New AED				
Felbamate	+	+	+ +	?
Gabapentin	±	+	±	0
Lamotrigine	+ +	0	±	0
Levetirazetam	0	0	0	0
Oxcarbazepine	+ +	0	?	?
Tiagabine	?	+ +	?	?
Topiramate	+	+	+	?
Vigabatrin	0	+ +	?	?
Zonisamide	+	±	?	?

[a]Blockade of voltage-gated sodium channels.
[b]Potentiation of GABAergic mechanisms.
[c]Blockade of glutamatergic mechanisms.
[d]Blockade of thalamic calcium channels.

In the common idiopathic generalized epilepsies, positional cloning of susceptibility genes was less successful due to the underlying complex genetic disposition. Many chromosomal regions (on chromosomes 6 and 15) are thought to harbor susceptibility genes for generalized seizures associated with juvenile myoclonic epilepsy.

In the genetic counseling of patients with epilepsies with complex modes of inheritance, empirical risk estimates are used. The recurrence risk for the offspring of probands with idiopathic generalized epilepsies is on the order of 5–10%.

VI. DIAGNOSIS

A. Clinical Diagnosis and Differential Diagnosis

Epilepsy is a clinical diagnosis, defined by recurrent epileptic seizures. The most important tool for the accurate classification and optimal diagnosis is the clinical interview. This should cover seizure-related information, such as subjective and objective ictal symptomatology, precipitation and frequency of seizures, history of seizures in first-degree relatives, and also information relevant for etiology, such as complications during pregnancy and birth, early psychomotor development, and history of brain injuries and other disorders of the central nervous system. Other important information that should be obtained refers to doses, side effects, and efficacy of previous medical or nonmedical treatment; evidence of psychiatric complications in the past; and psychosocial parameters, including educational and professional status, social independence, and psychosexual history.

The neurological examination may reveal signs of localized or diffuse brain damage. One should also look for skin abnormalities and minor stigmata suggestive of genetic diseases and neurodevelopmental malformations. There may be signs of injuries due to epileptic seizures, such as scars from recurrent falls, burns, and tongue biting. Hirsutism, gingival hyperplasia or acne vulgaris are indicative of side effects of long-term antiepileptic medication.

The clinical interview and the neurological examination are often sufficient to distinguish between epilepsy and its wide spectrum of differential diagnoses (Table IV) in most cases. However, there is a substantial problem with pseudoseizures.

For the correct interpretation of functional and structural diagnostic techniques it is important to understand that in focal epilepsies different concepts of pathological cerebral regions have to be distinguished. The epileptogenic zone is defined as the region of the brain from which the patient's habitual seizures arise. Closely related but not necessarily anatomically identical are the irritative zone, defined as the region of cortex that generates interictal epileptiform discharges in the EEG; the pacemaker zone, defined as the region of cortex from which the clinical seizures originate; the epileptogenic lesion, defined as the structural lesion that is usually related to epilepsy; the ictal symptomatic zone, defined as the region of cortex that

Table IV
Differential Diagnosis of Epilepsy

Neurological disorders
Transitory ischemic attacks
Migraine
Paroxysmal dysfunction in multiple sclerosis
Transient global amnesia
Movement disorders (hyperexplexia, tics, myoclonus, dystonia, paroxysmal choreoathetosis)
Drop attacks due to impaired CSF dynamics
Sleep disorders
Physiologic myoclonus
Pavor nocturnus
Somnambulism
Enuresis
Periodic movements in sleep
Sleep talking
Bruxism
Nightmares
Sleep apnea
Narcolepsy (cataplexy, automatic behavior, sleep attacks, hallucinations)
REM behavior disorder
Psychiatric disorders
"Pseudoseizures"
Panic attacks and anxiety disorders
Hyperventilation syndrome
Dissociative states, fugues
Episodic dyscontrol, rage attacks
Catatonia and depressive stupor
Medical disorders
Cardiac arrhythmias
Syncope (cardiac, orthostatic, reflex)
Metabolic disorders (e.g., hypoglycemia)
Hypertensive crisis
Endocrine disorders (e.g., pheochromocytoma)

Table V
Localization of Pathological Zones in Focal Epilepsies[a]

Epileptogenic zone	Ictal EEG with special electrode placements
Irritative zone	Interictal EEG, MEG
Pacemaker zone	Ictal EEG with special electrode placements
Ictal symptomatic zone	Ictal EEG with special electrode placements
Epileptogenic lesion	CCT, MRI
Functional deficit zone	Functional imaging, neurological examination, neuropsychological testing, nonepileptiform interictal EEG abnormalities

[a]From Lüders and Awad (1991).

generates the ictal seizure symptomatology; and the functional deficit zone, defined as the region of cortex that in the interictal period is functioning abnormally. The diagnostic techniques applied in epilepsy are characterized by a selective specificity for these various pathological regions (Table V).

B. EEG

The interictal surface EEG is still the most important method in the diagnosis and assessment of all types of epilepsy. A routine EEG is recorded over 30 min during a relaxed condition, including photic stimulation procedures and 5 min of hyperventilation. Paroxysmal discharges strongly suggestive of epilepsy are spikes, spike waves, and sharp waves. These epileptiform patterns, however, are not specific for epilepsy. They may be observed in patients suffering from nonepileptic neurological diseases and even in a small proportion of normal subjects.

The sensitivity of the routine EEG is limited by restrictions of spatial and temporal sampling. About 50% of patients with epilepsy do not show paroxysmal epileptiform discharges on a single EEG recording. Their detection depends on the epileptic syndrome and the therapeutic status of the patient. In untreated childhood absence epilepsy the EEG almost always shows generalized spike–wave complexes either occurring spontaneously or provoked by hyperventilation. In mild cryptogenic focal epilepsies, on the other hand, the interictal EEG is often negative. The temporal sensitivity can be increased by repeating the EEG or by carrying out long-term recordings with mobile EEGs. Paroxysmal discharges may also be identified by performing an EEG after sleep deprivation while the subject is asleep (Table VI).

Simultaneous video EEG recordings of seizures are useful for differentiating between different types of epileptic seizures and nonepileptic seizures. Ictal EEGs are also required for exact localization of the epileptogenic focus when epilepsy surgery is considered.

Surface EEGs record only a portion of the underlying brain activity. Discharges that are restricted to deep structures or to small cortical regions may not be detected. The spatial resolution of the EEG can be improved by special electrode placements, such as pharyngeal and sphenoidal electrodes.

Invasive EEG methods with chronic intracranial electrode placement are necessary for complex analysis in cases in which there are discordant or multifocal results of the ictal surface EEG and imaging techniques. These include foramen ovale electrodes positioned in the subdural space along the amygdala–hippocampal formation, epidural and subdural strip electrodes, and grids to study larger brain areas. Stereotactic depth electrodes provide excellent sensitivity for the detection of small areas of potentially epileptogenic tissue. The definition of exact location and boundaries of the epileptogenic region, however, is limited by the location and number of electrodes placed. Because of the limited coverage of implanted electrodes it may be difficult to distinguish whether a seizure discharge originates from a pacemaker zone or represents spread from a distant focus. Complications, the most severe being intracerebral hemorrhages, occur in 4% of patients.

Table VI
EEG Methods in the Diagnosis of Epilepsy

Interictal EEGs
Routine surface EEG
EEG after sleep withdrawal, during sleep
Mobile long-term EEG
Ictal EEG
Long-term video EEG
Special electrode placements (in increasing order of invasiveness)
Nasopharyngeal electrodes
Sphenoidal electrodes
Foramen ovale electrodes
Epidural electrodes (strips, grids)
Subdural electrodes (strips, grids)
Depth electrodes

C. Magnetoencephalography

Multichannel magnetoencephalography (MEG) is in some centers used in the presurgical assessment as a supplementary method to EEG. The electrical activity that can be measured by EEG produces a magnetic field perpendicular to the electric flow. This magnetic signal can be measured by MEG. In contrast to EEG, MEG is not influenced by intervening tissues, with the advantage of noninvasive localization of deep electric sources. Disadvantages are high costs and the susceptibility to movement artifacts, which makes it almost impossible to perform ictal studies.

D. Structural Imaging

Imaging studies should always be performed when a symptomatic etiology is suspected. Cranial computed tomography (CCT) is a quick, easy, and relatively cheap technique. The sensitivity can be improved by scanning in the axis of the temporal lobe (in cases of Temporal Lobe Epilepsy) and by using intravenous contrast enhancement. Except for a few pathologies such as calcifications, magnetic resonance imaging (MRI) is superior to CCT in terms of sensitivity and specificity in detecting epilepsy-related lesions such as malformations, gliosis, and tumors. With optimized MRI techniques, including T2-weighted images, inverse recovery sequences, and coronal images perpendicular to the hippocampus and thin sections, the sensitivity in depicting mesial temporal sclerosis reaches 90% (Fig. 3). Diagnosis of hippocampal pathology can be improved by quantitative MRI techniques such as T2 relaxometry and volumetric studies.

MR spectroscopy (MRS) is a noninvasive method for measuring chemicals in the body. MRS does not produce images but instead generates numerical values for chemicals. With phosphate spectroscopy it is possible to study energy metabolism in relation to seizure activity. Proton MRS is a technique that measures neuronal density, which has been found to be significantly decreased in the mesial temporal lobe of patients with mesial temporal sclerosis. MRS is a time-consuming procedure and therefore not routine.

E. Functional Imaging

Single photon emission computed tomography (SPECT) in epilepsy has mainly been confined to the imaging of cerebral blood flow in focal epilepsy. The tracer most widely used is ^{99}Tc-HMPAO. Interictally, there is localized hypoperfusion in an area extending beyond the epileptogenic region. Initially, there was considerable skepticism about the clinical value of the technique because the early studies were of low resolution and the correlations with electroencephalographic findings were imprecise. The sensitivity of interictal focus detection of SPECT in the literature ranges from 40 to 80%. In recent years there have been major technical developments in instrumentation. Using brain-dedicated multiheaded camera systems, the sensitivity of SPECT is similar to that of [^{18}F] fluorodeoxyglucose (FDG) positron emission tomography (PET).

^{99}Tc-HMPAO is distributed within a few minutes after injection in the brain, where it remains fixed for about 2 hr. If the radioisotope is injected during or shortly after an epileptic seizure, scanning can be carried out postictally without problems due to involuntary movements. Postictal and ictal SPECT is more sensitive than interictal SPECT and typically shows hyperperfusion ipsilateral to the epileptogenic

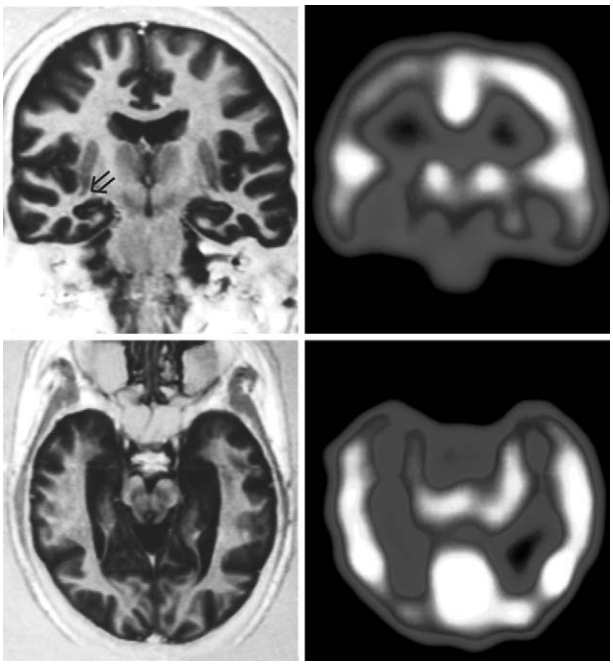


Figure 3 Imaging findings in temporal lobe epilepsy with hippocampal sclerosis. Left: MRI coronal (top) and horizontal (bottom), inversion recovery sequence. Hippocampal sclerosis (arrow). Right: Interictal HMPAO SPECT shows extensive ipsilateral hypoperfusion.

focus. Another ligand used for SPECT is ^{123}I-iomazenile, which is used to demonstrate benzodiazepine receptor binding, which is decreased in the epileptogenic region.

Compared to SPECT, PET is superior with respect to spatial and contrast resolution. PET, however, is expensive and requires an on-site medical cyclotron. Ictal studies are difficult with PET because of the short half-life of positron emitting radioisotopes. PET has mainly been used to study interictal blood flow with nitrogen-13-labeled ammonia and oxygen-15 and also glucose metabolism with FDG in focal epilepsy. Localized hypoperfusion and hypometabolism in the epileptogenic area as shown by PET are seen as a reliable confirmatory finding in the presurgical assessment of temporal lobe epilepsy. PET has also been used for imaging of benzodiazepine receptor binding and opiate receptor binding in epilepsy.

The major application of functional MRI in epilepsy is the noninvasive mapping of eloquent cortex in the presurgical evaluation of patients based on activation studies. The clear advantage over PET is the superior spatial resolution, which is on the order of 3 mm. The coregistration of a high-resolution MRI allows excellent localization of activated regions.

F. Neuropsychology

Identification of neuropsychological deficits is important for optimizing education, professional training, and rehabilitation in patients with epilepsy. Another aim of neuropsychological testing is to establish cognitive effects of antiepileptic drugs, seizure frequency, and ''subclinical'' EEG activity.

In the presurgical assessment neuropsychological evaluation is used to identify localizable deficits that can be related to the epileptogenic lesion. Crucial for lateralizing temporal lobe epilepsies is the function of verbal and nonverbal memory. Another neuropsychological task in the presurgical assessment is forecasting postsurgical cognitive outcome, which sometimes requires the intracarotid sodium amylobarbital procedure (Wada test).

G. Magnetic Cerebral Stimulation

Recently, magnetic cerebral stimulation has been used to lateralize the epileptogenic region in focal epilepsies. The technique of repetitive stimulation is also being evaluated for antiepileptic properties.

VII. TREATMENT

Once a diagnosis of epilepsy has been made and the decision to treat has been taken, many options are available. Pharmacological, surgical, and behavioral approaches may be taken. The majority of patients with epilepsy are treated with anticonvulsants. Recently, there has also been renewed interest in attempts at behavioral and nonpharmacological approaches to the management of seizures. A minority of patients with drug-resistant epilepsy may proceed to have surgery.

These strategies should not be viewed as being mutually exclusive. Hence, although in the following section, these treatments will be described separately, in practice they may be combined. Indeed, if surgery is pursued it is very likely that patients will receive antiepileptic medication both before and after the operation.

A. Antiepileptic Drugs

Ideally, patients should be managed on a single drug that leads to complete seizure freedom without causing any side effects. Approximately 75% of patients with epilepsy can be fully controlled on monotherapy (50% with initial monotherapy and 25% with second monotherapy), with the choice of agent determined by the epilepsy syndrome and seizure type (Table VII). Carbamazepine and valproate are the recommended substances for simple, complex partial, and secondary generalized tonic–clonic seizures. For most generalized seizures, sodium valproate is the most useful treatment. These drugs are generally thought of as the first-line treatments. Of the remaining patients not controlled on monotherapy, addition of another first-line drug will gain control in 15%. However, some patients will develop chronic seizures unrelieved by these treatments. In such circumstances, alternative monotherapies or adjunctive therapies will be considered. Many of the recently introduced anticonvulsants, often considered as second-line treatments, may be introduced, either alone or in combination with a first-line agent.

In Europe and the United States, nine novel anticonvulsants have been introduced in the past decade: vigabatrin, lamotrigine, felbamate, tiagabine, gabapentin, oxcarbazepine, topiramate, levetiracetam, and zonisamide. Because of serious side effects, two of these drugs are used only in a highly selected group of patients. Felbamate has caused fatal hepatotoxic and

Table VII
Selection of Antiepileptic Drugs According to the Epileptic Syndrome

Type of syndrome	First-line drugs	Second-line drugs
	Focal epilepsies	
All seizure types	Carbamazepine	Lamotrigine
	Valproic acid	
		Gabapentin
		Topiramate
		Tiagabine
		Levetiracetam
		Oxcarbazepine
		Phenytoin
		Clobazam
		Primidone
		Acetazolamide
		Vigabatrin
		Felbamate
	Generalized epilepsies	
Idiopathic		
Absences	Ethosuximide	Lamotrigine
	Valproic acid	
Myoclonic	Valproic acid	Lamotrigine
		Primidone
		Levetirazetam
Awakening grand mal	Valproic acid	Lamotrigine
		Primidone
		Topiramate
Symptomatic		
West syndrome	Vigabatrin	Clobazam
	ACTH	
	Valproate	
Lennox–Gastaut syndrome	Valproic acid	
		Lamotrigine
		Clobazam
		Topiramate
		Felbamate

hematotoxic adverse events, and vigabatrin leads to irreversible visual field constriction in 30% of patients.

An important feature in the clinical use of the first-line treatments is therapeutic drug monitoring (TDM). This should not be done as a routine procedure, but it may be useful in certain circumstances (e.g., to detect noncompliance and to explore changes in pharmacokinetics). There are significant interindividual variations in the relationship between drug levels and clinical side effects. Therefore, the measurement of serum concentrations should not be used for the optimization of treatment (with few exceptions); decisions to alter dosages should be based on efficacy and tolerability of AEDs. For appropriate agents, indications for TDM may include drug initiation or dose change, investigation of the absence of or change in therapeutic response, investigation of suspected toxicity, and situations in which pharmacokinetics may change such as during pregnancy and related to renal or hepatic disease. Sampling should occur once steady state has been achieved, four to six half-lives after treatment has been modified or introduced. TDM is particularly important for phenytoin because its hepatic metabolism is saturable and therefore a small increase in dose can result in a disproportionate and unpredictable increase in serum concentration.

B. Behavioral Treatment

Pharmacological treatments of epilepsy are not uniformly successful. Even if good seizure control is obtained, many patients experience troublesome side effects of treatments that must often be continued for many years. The need to take medication on a long-term basis has obvious implications for women wishing to become pregnant. In addition, many individuals describe feelings of oppression and an increased fear of being labeled as ill because of their ongoing need for regular drug taking. Surgical treatment is only an option for a minority of patients and is also not without physical and psychological sequelae. That alternative treatment approaches should be sought is therefore not surprising.

It has been suggested that many patients with epilepsy have a mental mechanism that they use in an attempt to inhibit their seizures. In one study of 70 patients, 36% claimed that they could sometimes stop their seizures. A behavioral approach to the treatment of epilepsy is based on observations that epilepsy can be manipulated in a systematic way through environmental, psychological, and physical changes. The initial stage in this approach is a behavioral analysis of the ways in which environmental and behavioral factors interact with seizure occurrence.

It has been demonstrated that significant reductions in seizure frequency may be achieved by teaching patients a specific contingent relaxation technique that they must be able to employ rapidly when they identify

a situation in which they are at high risk of having a seizure. In a subsequent study of three children with intractable seizures it was found that contingent relaxation alone did not significantly reduce seizures but that such a reduction was obtained following the addition of specific countermeasures aimed at changing the arousal level relevant to and contingent on early seizure cues—for instance, suddenly jerking the head to the right when it would habitually move to the left with a feeling of drowsiness at the onset of a seizure.

Some patients suffer from reflex seizures in that their seizures are precipitated by external stimuli. A proportion of people can identify specific environmental or affective triggers and may be able to develop specific strategies to abort or delay a seizure. These methods may involve motor or sensory activity or they may be purely mental. However, in one study it was found that 50% of patients who inhibited their seizures at times had to "pay the price in subsequent discomfort."

Primary seizure inhibition describes the direct inhibition of seizures by an act of will; for instance, a man whose seizures were precipitated by a feeling of unsteadiness dealt with this by keeping his gaze fixed on a point when walking down an incline. The nature of the successful act varies from person to person, and if this treatment approach is to be pursued then it must be tailored to each individual, based on an analysis of their seizures and of any actions they may already have noticed modify their seizures.

The term secondary inhibition is employed by Fenwick to describe behavioral techniques that are thought to act by changing cortical activity in the partially damaged group 2 neurons around the focus without deliberately intending to do so, thereby reducing the risk both of a partial seizure discharge and of a generalized seizure discharge that may otherwise follow recruitment of surrounding normal brain by group 2 neurons firing abnormally. An example of this is the act of maintaining alertness by a patient whose seizures appear in a state of drowsiness. Treatment in this case starts with trying to identify situations in which the subject reliably tends to have seizures or, alternatively, to be free of them.

In addition to these seizure-related approaches, more general psychological strategies have also been investigated. Several anecdotal reports have been published demonstrating benefit from reward programs that aim to reward seizure-free periods. Based on the observation that some patients with olfactory auras can prevent progression of the seizure by applying a sudden, usually unpleasant olfactory counterstimulus, "aromatherapy" techniques have been studied in the control of epilepsy. Currently, it is not clear what the relative contributions of specific olfactory stimuli and the general relaxation that is a part of the treatment are to any clinical benefit that might be observed.

Specific biofeedback techniques have also been explored. Measurement of scalp electrical activity has demonstrated that there is an increase in surface-negative slow cortical potentials (SCPs) in the seconds before a seizure occurs. These SCPs represent the extent to which apical dendrites of cortical pyramidal cells are depolarized and hence indicate neuronal excitability. Studies using visual feedback of this effect have demonstrated that some patients are able to modulate cortical electrical activity with an associated decrease in seizure frequency. However, it appears that patients with epilepsy are less able than normal controls to regulate their cortical excitability. This impairment can be minimized by extending the amount of training received by those with epilepsy. In a study that gave 28 1-hr training sessions to 18 patients followed up for at least 1 year, 6 became seizure-free. However, not every patient who achieved reliable SCP control experienced a reduction in seizure frequency.

It has been noted that the teaching of any of these methods of self-control of seizures may increase morale not only by reducing seizures but also by providing patients with a sense of control over their epilepsy. An important aspect of many "nonmedical" treatments of epilepsy is that although still of very limited proven benefit, they aim to consider seizures in the wider setting of the patient's life. In mainstream clinical management it is sometimes easier to focus purely on seizure response to the latest change in anticonvulsant therapy.

C. Surgical Treatment

In all patients with persistent epilepsy unrelieved by AED treatment, it is appropriate to consider surgical intervention. The aim of epilepsy surgery is the removal of the epileptogenic brain tissue in order to achieve seizure control without causing additional iatrogenic deficits. Only in rare cases, such as those with diffuse pathology and catastrophic seizures, are palliative surgical procedures performed. Approximately one-third of patients with refractory epilepsy are suitable for epilepsy surgery. The prerequisites of surgery are frequent epileptic seizures and a minimum

of 3–5 years of unsuccessful AED treatment including at least two first-line AEDs in monotherapy and combination. This is a minimal condition that only applies to patients who are ideal candidates for epilepsy surgery. These are patients with unilateral mesial temporal lobe epilepsy and hippocampal sclerosis, extratemporal epilepsies with localized structural lesions, and some types of catastrophic epilepsies in childhood. The poorer the individual prognosis of surgery, the longer the requested presurgical attempts to achieve seizure control with AEDs and complimentary nonsurgical treatments.

1. Presurgical Assessment

Treatment centers that engage in routine surgical management of epilepsy generally have a standardized assessment process for patients being considered for such treatment. Although the program may vary between centers, the general procedure is similar. A first hypothesis with respect to the suspected epileptogenic region is based on clinical history, neurological examination, and interictal EEG. These procedures focus particularly on searching for etiological factors, evidence of localizing signs and symptoms, and a witnessed description of the seizures. In addition, psychosocial information must be gathered relating to education, employment, social support, and past and present mental state findings. In a second step of presurgical assessment, all patients undergo more specific investigations. In general, these include several days of continuous video telemetry using surface electrodes. The aim of telemetry is to obtain multiple ictal recordings, which give more valuable localizing information than interictal records. Often, patients will reduce their AED prior to telemetry in order to facilitate the occurrence of seizures. Recent advances in structural and functional neuroimaging have made invasive EEG recording less necessary. MRI using optimized techniques has increased the sensitivity for the detection of subtle structural lesions not seen on CCT and standard MRI and is now used routinely. Functional imaging techniques using SPECT and PET are used to further delineate the epileptogenic region. If results from all investigations are concordant, patients will have surgery. If results are discordant, invasive EEG recordings using subdural grid electrodes or intracerebral depth electrodes may be applied in order to identify the critical epileptogenic region.

All patients being considered for epilepsy surgery should have a neuropsychological assessment. This is important both to detect focal brain dysfunction and to predict the results of surgery, especially temporal lobe surgery. The intelligence quotient may be measured using the Wechsler Adult Intelligence Scale. In some centers a score of less than 75 has been taken as evidence of diffuse neurological disorder and, hence, as a relative contraindication to surgery. It is important before proceeding to surgery to investigate the hemispheric localization of language and memory. This is generally performed using the Wada test. Sodium amylobarbitone is injected into one internal carotid artery, and while that hemisphere is briefly suppressed language and memory tests are performed. The procedure is then repeated for the other hemisphere.

2. Surgical Procedures

Many surgical procedures have been developed. Most patients who undergo surgery suffer from temporal lobe epilepsy. Most of these patients undergo one of two standard procedures: two-thirds anterior resection or amygdala hippocampectomy. In extratemporal epilepsies, lobectomies, lesionectomies, or individually "tailored" topectomies are performed. The latter techniques may involve preoperative stimulation in order to identify eloquent brain tissue that should be preserved from resection. A palliative method that can be applied in functionally crucial cortical regions is multiple subpial resections, which interrupt intracortical connections without destroying the neuronal columns that are necessary for normal cerebral function. Another palliative method that is performed mainly in patients with epileptic falls is anterior or total callosotomy.

3. Prognosis and Complications

The outcome of epilepsy surgery is classified according to four categories: class 1, no disabling seizures; class 2, almost seizure free; class 3, clinical improvement; and class 4, no significant improvement. According to a meta-analysis of 30 surgical series with a total of 1651 patients, seizure outcome is as follows: class 1, 59%; class 2, 14%; class 3, 15%; and class 4, 12%. In this study, predictors of a good surgical outcome were febrile seizures, complex partial seizures, low preoperative seizure frequency, lateralized interictal EEG findings, unilateral hippocampal pathology on MRI, and neuropathological diagnosis of hippocampal sclerosis. Predictors for a poor outcome were generalized seizures, diffuse pathology on MRI, normal histology of removed tissue, and early postoperative seizures.

Relative contraindications for temporal lobe resections are extensive or multiple lesions, bilateral

hippocampal sclerosis or dual pathology within one temporal lobe, significant cognitive deficits, interictal psychosis, multiple seizure types, extratemporal foci in the interictal EEG, and normal MRI. The ideal patient has mesial temporal lobe epilepsy due to unilateral hippocampal sclerosis of the nondominant hemisphere. A negative MRI decreases the chances of surgery for seizure control and a focus in the dominant hemisphere increases the risk of postoperative neuropsychological deficits.

The nature of potential perioperative complications and postsurgical neurological, cognitive, and psychiatric sequelae depends in part on the site of surgery. Operative complications occur in less than 5% of patients. In two-thirds of patients who undergo temporal resection of the dominant hemisphere, verbal memory is impaired. Another possible complication of epilepsy surgery is psychiatric disorder. Only rarely does this manifest as *de novo* psychosis. Many patients, however, will go through a phase of depression and increased anxiety following surgery. Therefore, most surgery centers include a psychiatric assessment in the preoperative phase and also a psychiatric follow-up in order to avoid catastrophic reactions (including suicide) to either the failure of surgery or the success of surgery, which requires far-reaching psychosocial adjustment.

4. Vagal Nerve Stimulation

A different surgical approach is that of vagal nerve stimulation (VNS) using an implanted stimulator. This approach has the drawback that while the nerve is being stimulated, usually for 30 sec every 5–10 min, the voice changes. More intense stimulation may be associated with throat pain or coughing. Nevertheless, in one series of 130 patients, mean seizure frequency decreased by 30% after 3 months and by 50% after 1 year of therapy. Altogether, 60–70% of the patients showed some response, but seizure freedom has rarely been achieved. Therefore, VNS is only indicated in patients who are pharmacoresistant and who are not suitable for resective epilepsy surgery.

VIII. STATUS EPILEPTICUS

Status epilepticus (SE) is defined as a condition in which a patient has a prolonged seizure or has recurrent seizures without fully recovering in the interval. With respect to the most severe seizure type (primary or generalized tonic–clonic) a duration of 5 min is sufficient for diagnosis of status epilepticus. SE may occur in a person with chronic epilepsy (the most frequent cause being noncompliance) or in a person with an acute systemic or brain disease (such as hypoglycemia or stroke). The classification of SE follows the international classification of epileptic seizures. In clinical praxis, often only two major types are distinguished: convulsive and nonconvulsive status epilepticus. SE, particularly convulsive SE, is a life-threatening condition that requires emergency treatment. Prolonged seizure activity causes systemic complications and brain damage.

The incidence of SE is 50/100,000. In all studies the convulsive type is most common (80%). However, nonconvulsive SE is difficult to recognize and may therefore be underrepresented in epidemiological studies. Mortality of SE is on the order of 20%. Prognosis and response to treatment largely depend on the etiology of SE. SE occurring in the context of chronic epilepsy has a much better outcome than SE complicating acute processes (such as hopoxemia). SE should always be managed on an emergency ward. Traditionally, drug treatment of SE includes the immediate intravenous administration of a fast-acting benzodiazepine such as diazepam plus phenytoin. Alternatively, lorazepam may be used alone because its pharmacokinteics are such that the medium-term control is comparable to that of phenytoin. If SE persists, phenobarbitone is an alternative, with the next step being a general anesthesia using pentobarbital or propofol .

IX. PSYCHIATRIC AND SOCIAL ASPECTS OF EPILEPSY

A person with epilepsy, particularly if seizures are not immediately controlled, carries an increased risk of developing psychiatric complications. These may be related to the psychological burden of a disorder that is characterized by unpredictable loss of control and is often stigmatized in modern societies. Other risk factors are biological and include the consequences of recurrent epileptic cerebral dysfunction as well as the potentially negative psychotropic effects of AEDs. Psychoses in epilepsy can occur in direct relationship to seizure activity, most often triggered by a series of seizures. They may also occur when seizures are controlled—so-called "alternative psychoses" due to "forced normalization." More common than psychoses are depressive syndromes, and suicide accounts for about 10% of all deaths in persons with epilepsy. Other psychiatric disorders that are common in

epilepsy are anxiety disorders and the coexistence of epileptic and nonepileptic pseudoseizures.

Epileptic seizures have a significant impact on the social life of a person with epilepsy. Depending on the type and frequency of seizures, patients are restricted in terms of their professional options as well as their leisure time activities. In all countries there are restrictions on driving. Patients and relatives often have false ideas about the nature and course of epilepsy and therefore need to be informed about individual dangers, restrictions, and liberties. Therefore, optimal treatment of epilepsy is a multidisciplinary and enduring task, in which not only the doctor but also the social worker and the psychologist play a significant role.

See Also the Following Articles

ANTERIOR CINGULATE CORTEX • BIOFEEDBACK • BRAIN DISEASE, ORGANIC • CEREBRAL CORTEX • GABA • MOOD DISORDERS • MOVEMENT REGULATION

Suggested Reading

Annegers, J. F., Hauser, W. A., and Elveback, L. R. (1979). Remission of seizures and relapse in patients with epilepsy. *Epilepsia* **20,** 729–737.

Berg, A. T., and Shinnar, S. (1991). The risk of seizure recurrence following a first unprovoked seizure: A quantitative review. *Neurology* **41,** 965–972.

Commission on Classification and Terminology of the International League against Epilepsy (1981). Proposal for revised clinical and electroencephalographic classification of epileptic seizures. *Epilepsia* **22,** 489–501.

Commission on Classification and Terminology of the International League against Epilepsy (1989). Proposal for revised classification of epilepsies and epileptic syndromes. *Epilepsia* **30,** 389–399.

Engel, J., and Pedley, T. A. (Eds.) (1997). *Epilepsy: A Comprehensive Textbook*. Lippincott–Raven, Philadelphia.

Fenwick, P. (1991). Evocation and inhibition of seizures: Behavioural treatment. *Adv. Neurol.* **55,** 163–183.

Lüders, H. O., and Awad, I. (1991). *Conceptual considerations*. In *Epilepsy Surgery* (H. Lüders, Ed.), pp. 51–62. Raven Press, New York.

Lüders, H. O., and Noachtar, S. (2001). *Atlas of Epileptic Seizures and Syndromes*. Saunders, Philadelphia.

Musicco, M., Beghi, E., Solari, A., and Viani, F. (1997). Treatment of first tonic–clonic seizure does not improve the prognosis of epilepsy. First Seizure Trial Group (FIRST Group). *Neurology* **49,** 991–998.

Schmitz, B., and Trimble, M. R. (2001). *Psychobiology of Epilepsy*, Cambridge Univ. Press, Cambridge, UK.

Shorvon, S. D. (1990). Epidemiology, classification, natural course and genetics of epilepsy. *Lancet* **336,** 93–96.

Shorvon, S. (1994). *Status Epilepticus*. Cambridge Univ. Press, Cambridge, UK.

Temkin, O. (1971). *The Falling Sickness*. Johns Hopkins Univ. Press, Baltimore, MD.

Event-Related Electromagnetic Responses

JOHN S. GEORGE
Los Alamos National Laboratory

GLOSSARY

electroencephalography (EEG) The noninvasive measurement of electrophysiological activity of the brain using electrodes attached to the scalp surface.

endogenous response An event-related response associated with voluntary movement, decision-making, or cognitive or other neural activity associated with internal processes of the experimental subject.

event-related potential (ERP) A spatiotemporal pattern in EEG data associated with a discernible event such as delivery of a stimulus or voluntary motion. The MEG analogs of these responses are sometimes called event-related fields.

event-related synchronization–event-related desynchronization (ERS–ERD) Neural population responses apparent as a resetting or reorganization of ongoing oscillatory activity.

evoked response An event-related response associated with sensory information processing of an external stimulus.

forward problem Computation of the physical consequences (e.g., an observable magnetic field or potential distribution at the head surface) associated with an assumed pattern of neural activation and a model of the properties of the head.

inverse problem Estimation of patterns of neural activation that can account for an observed pattern of experimental measures.

latency The timing of a response component or other feature relative to a defining event. Originally used to denote the period before any response was apparent (i.e., the latent phase of the response), in present usage often used to identify the peak of a response component.

magnetoencephalography (MEG) The magnetic analog of EEG, measuring magnetic fields associated with neural currents using ultrasensitive superconducting instrumentation.

response components Reproducible features in an event-related response. Typically identified on the basis of peaks and valleys in the response waveform in early work, components may be discriminated on the basis of response field topographies, consequences of stimulus properties, effects of behavioral modulations, or other criteria.

source localization Identification of the regions of neural activation that give rise to externally observable electromagnetic responses or other evidence of neural activation.

Event-related responses are spatial–temporal patterns of physiological responses associated with neural population activity, elicited by external stimuli or internal imperatives. Because these signals typically are much smaller than the ongoing activity, signal averaging, correlation, or related signal processing strategies are employed to recover the response.

I. INTRODUCTION

The function of neural systems, from the feeding and avoidance behaviors of the simplest multicellular organisms to the highest cognitive functions of the human brain, depends on dynamic spatial and temporal patterns of activation within linked networks of excitable cells. In the human brain, most purposeful function is mediated by correlated activity in substantial populations of neurons. The physical

and physiological consequences of this activity can be detected with noninvasive measurement techniques, including electroencephalography (EEG) and magnetoencephalography (MEG). These techniques measure the integrated activity of thousands to hundreds of thousands of neurons. Many cells operating in synchrony are required to generate fields that can be detected centimeters away from the source. Fortunately, neural activation typically involves large clusters of neurons with similar response properties.

The information processing activities of individual neurons depend on a chain of biophysical processes: the integration of synaptic input (both excitatory and inhibitory) throughout the dendritic tree; electrical excitation mediated by the biophysical properties of ionic channel proteins in the cell membrane; transmission mediated by active and passive physical processes in the neuronal axon; and chemical or electrical relay of activation at the synapses on target cells. By linking together collections of excitable cells, networks can achieve complex behaviors beyond the capacity of individual cells.

MEG and EEG are not the only methods that can be employed with event-related and evoked response techniques. Essentially any technique that records a consistent transient response to neural activation can be used. Optical imaging methods employing sensitive video cameras have been used to record event-related responses from exposed brain tissue in experimental animals and in humans undergoing neurosurgery. With small animals such as rats or mice, it is possible to acquire images of reasonable quality though the skull. Most studies to date have exploited changes in blood flow and oxygenation associated with neural activation—the same processes that serve as the basis of fMRI. Studies have shown that it is possible to image fast intrinsic responses of neural tissue with high-performance video technology *in vivo*. Fast optical responses tightly coupled with the electrophysiological processes of neural activation were described at least 30 years ago, and the roots of such work go back much farther.

Other investigators have recorded optical evoked responses from human subjects using noninvasive methods, i.e., by injecting and recording light at the head surface to detect changes in the optical properties of tissue buried deep beneath the skull. Impedance tomography techniques have been used in a similar way to detect transient changes in tissue conductivity associated with neural activity. Even sensitive temperature measurement techniques, based on thermal emission of infrared photons, or, more recently, MRI-based methods have been used to record event-related responses associated with neural activation.

Event-related experimental paradigms have been demonstrated with functional MRI (fMRI), in spite of the facts that the earliest detectable hemodynamic responses require several hundred milliseconds and that the response peaks several seconds after neural electrophysiological activation. It is possible to use sophisticated selective averaging or correlation techniques to identify fMRI evoked responses that are highly overlapping due to rapid stimulation rates.

MRI has other important and unique roles in functional neuroimaging. MRI is used to identify and visualize the anatomical substrate of functional activation through coregistration with other imaging modalities. MRI can be used to define the computational geometries required to model the physics of techniques such as MEG and EEG, facilitating three-dimensional (3D) source localization on the basis of data that is surface-based and, thus, topographic. By defining the geometry of cortex, MRI provides useful constraints on ill-posed source localization procedures. Functional MRI provides an alternative measure of neural activation that can increase confidence in the results of comparative or integrated analysis based on multiple imaging modalities. Event-related fMRI techniques allow the same experimental paradigms to be used for NEM and fMRI experiments, a useful feature for most purposes and a prerequisite for simultaneous measurements, for example, combining EEG and fMRI.

We will consider a range of methods within this article, but the major focus will be on neural electromagnetic measurement (NEM) techniques (MEG and EEG) and noninvasive measurements of neural population responses. These methods have challenges and limitations for localizing the source of neural responses, but the excellent temporal resolution of these responses can be exploited by clever experimental paradigms to probe the dynamic interactions between multiple cortical regions. These interactions serve as the basis of information processing and control by the human brain.

II. PHYSIOLOGICAL AND PHYSICAL BASIS OF NEM RESPONSES

A. Single-Unit Electrophysiological Responses

Much of our understanding of neural function stems from electrophysiological studies of single neurons. By

placing microelectrodes made of wire or drawn glass capillary tubing into or next to neurons, it is possible to record signals generated by one or a few cells. Because of the stereotypical shape and amplitude of action potentials (spikes) recorded from a given cell in a given recording configuration, it is often possible to sort a complex record into contributions from a small set of cells. The intensity of neuronal activation is typically assessed on the basis of firing rate; in sensory systems, the intensity of a stimulus is often encoded logarithmically in the firing rate within the sensory nerve. The response properties of neurons are often defined in terms of the *receptive field*, the region of sensory parameter space (e.g., location on the retina or stimulus properties of an auditory or visual stimulus) that can influence the firing rate of a particular neuron. A growing body of evidence shows that the temporal pattern of firing may also encode information. For example, in measurements across an ensemble of auditory nerve fibers in response to a pure tone stimulus below 1 kHz or so, spikes tend to occur in phase with the stimulus, i.e., a temporal code captures the temporal structure of the stimulus. In the visual system, phase-locked oscillatory activity in widely spaced cells appears to encode higher order features, such as coherent motion, beyond the spatial limits of the conventional receptive field.

B. Multiple Areas, Discrete Sources

Macroscopic electrophysiological techniques can resolve the course of neural population activation with sub-millisecond temporal resolution. This is adequate to detect the synchronous volley of action potentials that can result from electrical stimulation or sharp physiological activation and allows the characterization of oscillatory activity that can emerge within a network or by interactions between brain regions. In practice, most responses measured noninvasively do not disclose significant structure at time scales below a few milliseconds. A few specialized methods such as auditory brain stem evoked responses or electrical stimulation of the median nerve elicit a tightly correlated volley of action potentials that can be measured externally. However, most NEM evoked responses are a few milliseconds to tens of milliseconds wide and occur tens of milliseconds to hundreds of milliseconds after a stimulus event.

Dynamic responses of neural tissue involve a number of distinct processes. In response to upstream neural activity (or by endogenous processes in certain specialized sensor neurons, such as retinal photoreceptors or hair cells of the auditory system), the neuron generates an electrical response. This response is typically initiated by the flow of ionic currents though transmembrane protein channels in the plasma membranes of the neuronal dendritic tree. These currents produce a change in the standing potential across the membrane and produce potential gradients along the neuron. These potential gradients give rise to passive currents that flow along dendritic processes to rapidly equilibrate the membrane potential. Most neurons contain other channel proteins whose properties are voltage-sensitive; an excursion of the membrane potential from its typical resting value across some threshold produces a conformational change that opens an ion-specific conductance. This transient conductance change gives rise to an action potential or spike. Subsequent responses by other voltage-sensitive channels within the neuronal membrane shut down the action potential and allow recovery of the membrane potential to resting levels.

C. Propagation and Transmission

Unlike the passive conduction processes of the dendrites, action potentials are active processes that are essentially regenerated locally as the response moves across the cell body and along the axon. This allows the action potential to propagate for long distances without significant degradation of response amplitude. In some neurons the axons are sheathed with an insulating material called myelin. The sheath is interrupted periodically at nodes where membrane proteins responsible for excitation are concentrated. Excitation in myelinated axons appears to hop from node to node at rates considerably faster than transmission in uninsulated axons.

Because a patch of membrane is insensitive or refractory for a period after firing and given the stereotyped spatial pattern of dendritic integration, action potentials normally propagate in one direction from the cell body to the far reaches of the axon. Electrical simulation can elicit backward—antidromic—transmission, and there is evidence that similar processes may operate in the recovery of dendrites. Electrophysiological transmission in some specialized neural tissue is mediated by direct conductive interconnections (gap junctions), but most interneuron transmission is mediated by electrochemical transmission at specialized sites known as synapses. At the presynaptic terminal, the arrival of an action potential

causes ionic changes (including an influx of calcium) that lead to the release of a chemical neurotransmitter into the synaptic cleft. The neurotransmitter molecules diffuse across the synapse to bind to receptor proteins at the postsynaptic terminal. These receptors in turn activate ionic channels, which alter the local ionic composition and/or membrane potential within the postsynaptic dendrite, leading to another cycle of integration, excitation, and transmission.

D. Neural Population Responses

These processes are the basis of information processing by neural networks. The patterns of connectivity between cells define which cells can influence which other cells and, thus, define the architecture of information processing. Spikes signal the generation of a response by a single neuron and trigger the transmission of that information to the next cell in the chain. However, most neurons are interconnected in such a way that a spike in a single presynaptic neuron is not sufficient to elicit a response. Given the transient nature of postsynaptic responses, the efficacy of excitation within any particular neuron is critically dependent on the detailed temporal dynamics of activity within the population of neurons that converge on it.

Macroscopic electrophysiological responses depend on the activation of a significant population of neurons activating in concert. This is primarily a consequence of the distance separating the neuronal current generators and the electrodes or magnetic field detectors at the head surface. Electric and magnetic fields drop off

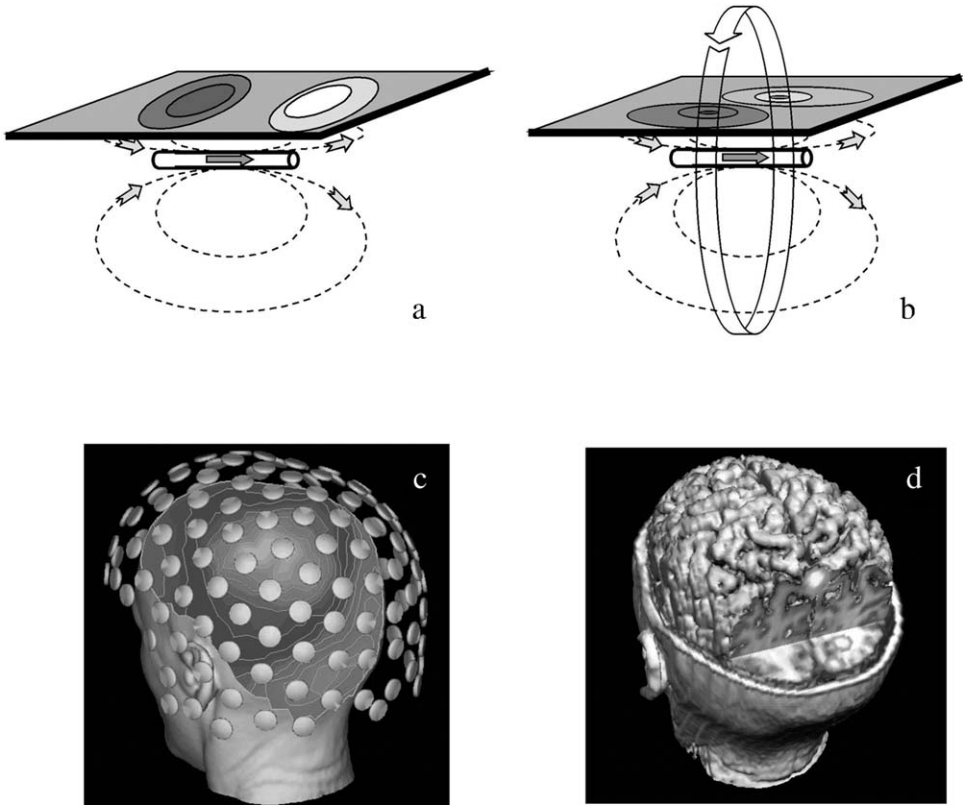


Figure 1 The physical basis of neural electromagnetic source localization. (a) Intracellular currents in tangential neural processes (parallel to the head surface) give rise to extracellular volume return currents. These currents interact with the head volume conductivity to set up a potential distribution, with surface extrema aligned with the current. (b) A detectable magnetic field is associated with the intracellular current. Extracellular currents tend to cancel in a spherical conducting volume. The extrema of the observed magnetic field distribution straddle and are orthogonal to the source current element. (c) An array of electrodes or SQUID-based magnetic field detectors are positioned on or over the surface of the head. Potential and field distributions consistent with one or more simple dipole-like sources are often observed. (d) Source localization based on a time-varying set of equivalent current dipoles. A simple source model is fit to the observed field distribution using nonlinear optimization techniques. In this figure, the uncertainty of source localization due to noise was estimated using Monte Carlo techniques, and a 3D histogram of dipole location was constructed.

rapidly with distance. EEG is further compromised by the insulating properties of the skull, which attenuate and diffuse the potential distributions that can be measured at the surface of the brain. Magnetic measurements are much less sensitive to the conductivity properties of the head, but the sensitivity of the method is reduced by the use of gradiometer sensors. Measurement of field gradients makes MEG less sensitive to interference from environmental influences, such as electric equipment, passing vehicles, or moving metal objects such as gurneys, at the expense of absolute sensitivity to neural responses.

E. Neural Electromagnetics

NEM responses are governed by the same physical processes that give rise to electric and magnetic fields in other systems. The vector currents set up by potential differences along cellular processes give rise to an electric field aligned with the current and an orthogonal magnetic field that encircles the current element. The sense (polarity) of the magnetic flux is predicted by the right-hand rule, which also predicts the direction of current flow induced in a coil penetrated by the flux. In a simple medium, charge is neither created nor destroyed; thus, all currents flow in closed loops. These relationships are summarized in Fig. 1.

Longitudinal, intracellular current flowing in the neuron during activation is matched by return current flowing through the extracellular space. These currents flow passively throughout the entire volume conductor, driven by the spatial potential gradient and limited by the resistance of the medium. In a spherical volume conductor, the magnetic contributions of these volume return currents integrate to zero, so that the measurement is dominated by the coherent currents flowing within neurons. The magnetic field is not strongly affected by the properties of biological tissue. EEG depends on the measurement of potential distributions set up by charge migration within the electric field and are critically dependent on the properties of the volume conductor. The potentials observed at the head surface arise from volume or capacitance currents that penetrate the highly resistive skull.

F. Spikes and Dendritic Responses

Although many investigators consider action potentials to be the most relevant observable consequence of neuronal activation, a number of factors limit their contribution to field and potential distributions measured at the head surface. The transmembrane currents that generate the action potential are radially symmetrical. The field cancellation associated with such distributions prevents their observation at a distance. The action potential itself is biphasic in both time and space: a depolarizing current flows along the axon, ahead of the region of excitation, and a repolarizing current flows in the opposite direction. This so-called quadrupolar current pattern is not readily detected in field measurements at a distance. Finally, the biphasic, temporally transient nature and stochastic timing of action potentials may limit their contribution to time-locked averages typically used in event-related or evoked response paradigms.

A number of lines of evidence suggest that electrophysiological responses observed at the head surface are dominated by postsynaptic dendritic currents. Such currents are typically gated by neurotransmitters, although they may have voltage-dependent characteristics. These channels often have a defined ionic specificity; some ions such as calcium may have powerful and specific effects on the response characteristics of the cell. Potentials set up by postsynaptic ion channels drive passive (i.e., ohmic) conduction along the dendrites.

G. Superposition and Integration

Electric and magnetic fields and the associated potential distributions obey the principle of superposition: the field distribution associated with a complex 3D pattern of currents is the integral over the contributions of all of the source currents flowing within the conducting volume. Within a small volume element (voxel), the distribution of cellular and volume currents can be modeled adequately by an equivalent point current with some orientation and magnitude, and the field of the entire volume can be computed as the sum over the contributions of all of the voxel currents.

Although the dendritic tree of a single neuron is often contained within a single voxel of an anatomical magnetic resonance image, each segment of current within the dendritic processes makes a contribution to the observed magnetic field and potential distributions. If the dendritic tree has a radially symmetrical geometry, the contributions may cancel and produce no effect observable at a distance. Cells of this configuration have been termed ''closed field''

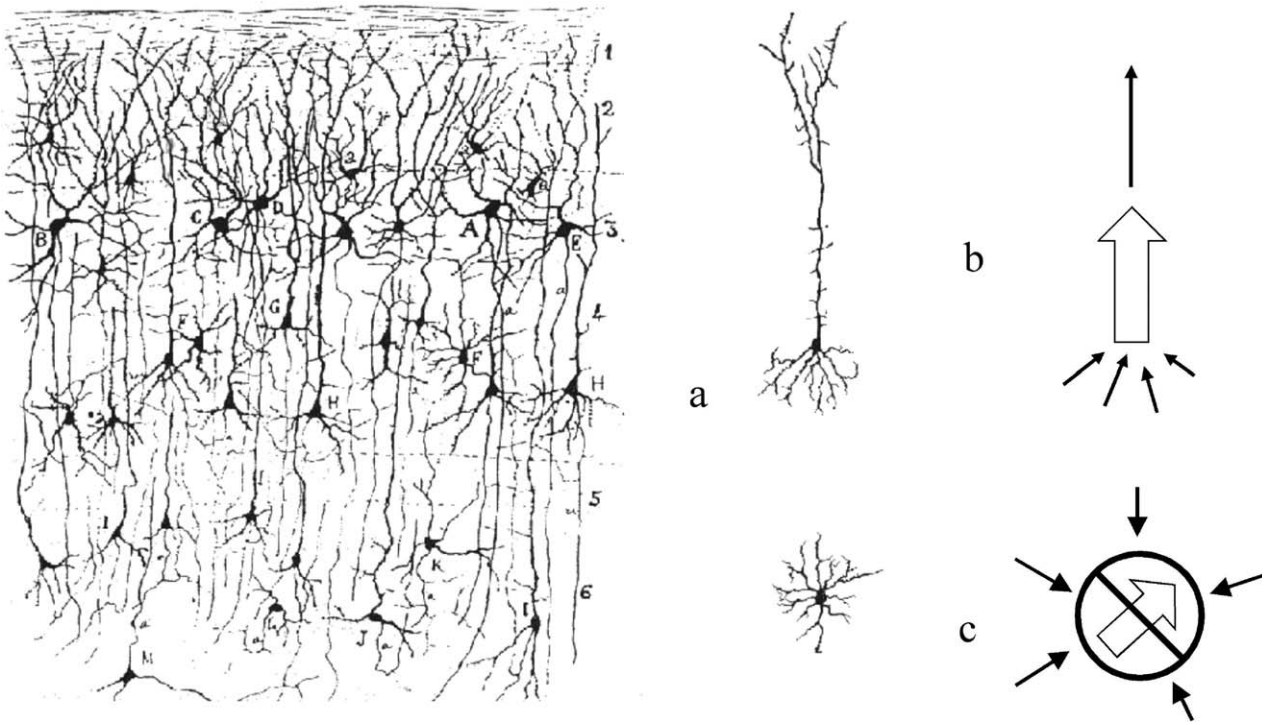


Figure 2 Vector summation of neuronal intracellular currents. (a) The regular array of cortical neurons is evident in this classic drawing of Golgi-stained neural tissue. (b) The partial symmetry of an open-field neuron (pyramidal cell) gives rise to a net intracellular current vector aligned normal to the cortical surface. (c) The radial symmetry of a closed-field interneuron produces current cancellation resulting in no net intracellular current vector.

neurons. Neurons with an asymmetric dendritic arborization have a net current vector and are termed "open field." These configurations are illustrated in Fig. 2. The neurons of neocortex have a net asymmetry that is normal (i.e., perpendicular) to the local cortical surface. Experimental validation of this prediction comes from experimental observations from linear arrays of microelectrodes penetrating the cortex. Such measurements find layered distributions of current sources and sinks across the thickness of cortex, whereas tangential potential differences are small within an extended area of activation. This basic prediction also is consistent with NEM measurements of responses from systematically arrayed sources in somatosensory or primary visual cortex.

III. THE METHODS

A. Measurement Technologies

Both MEG and EEG are passive technologies; they use sensitive instrumentation to detect the tiny electrical and magnetic perturbations associated with neural activity. Although some aspects of biological electricity can be measured directly even with primitive galvanometers, the practical application of EEG required the development of sensitive amplifiers. In particular, amplifiers with high input impedance [such as devices incorporating field effect transistors (FETs) on the front end] allow precise potential measurements without drawing significant current. This strategy has a number of advantages, including reduced sensitivity to high impedance electrodes or connections to the scalp.

EEG techniques traditionally employ a modest number of electrodes that are applied individually by hand. The most commonly used system (particularly for clinical practice) is the international 10–20 system —a montage of around 20 electrodes placed over the surface of the scalp with reference to anatomical landmarks. Electrodes are affixed to the scalp with conductive gel or other adhesives after preparation of the surface, typically by mild abrasion. This procedure produces lower contact resistance but is labor-intensive. Advances in source localization techniques, to

some extent driven by MEG, have provided motivation to increase the density of potential sampling across the scalp surface. Over the past decade typical research systems have grown from 32 channels to 128 channels or more. The development of electrode arrays based on caps or mechanical tension structures (see Fig. 3) has made the application of such arrays considerably more practical, requiring minutes instead of the hours required to apply dense arrays using conventional techniques. Although it is not yet clear how many electrodes are required to capture the nuances of the scalp potential map, studies of the spatial Nyquist frequency for EEG suggest that useful new details are available using 128-channel electrode arrays and perhaps even larger arrays.

MEG is based on ultrasensitive magnetic field measurement technology, incorporating superconducting quantum interference devices (SQUIDs). The field sensors are typically superconducting pickup coils consisting of multiple loops configured for sensitivity to field gradients. Both SQUIDs and gradiometers can be constructed from high-temperature superconductors. In designs to date, the increase in noise associated with higher temperatures has proven unacceptable for brain measurements, although such systems are adequate for cardiac applications. From early in the history of MEG, it has been clear that sensor arrays of 256 channels or more would be required to adequately sample field distributions associated with superficial cortical sources. However, the first systems consisted of detector arrays of limited extent. Mapping studies involved multiple placements of the sensor array in the context of event-related or evoked response paradigms. Most current MEG and EEG systems still sacrifice spatial sampling in order to provide whole-head coverage with a practical and more affordable number of sensor channels.

MEG sensor arrays are housed in a dewar, essentially a double-walled, vacuum-insulated flask, designed to hold cryogenic fluids such as helium. The dewar limits the proximity of sensors to the head surface; most superconducting sensor arrays are separated from the outside world by 1 cm or more. With advanced designs and considerable care, this distance can be reduced to closer to 1 mm for some applications. The standoff distance produces a significant loss of sensitivity, but this is not the limiting factor for most human brain measurements. Sensor arrays must be constructed with a rigid housing designed to accommodate the majority of heads, and smaller heads will involve greater separations over at least portions of the sensor array. Even the most superficial areas of cortex are on the order of 1 cm below the head surface, and some subcortical structures of interest may be

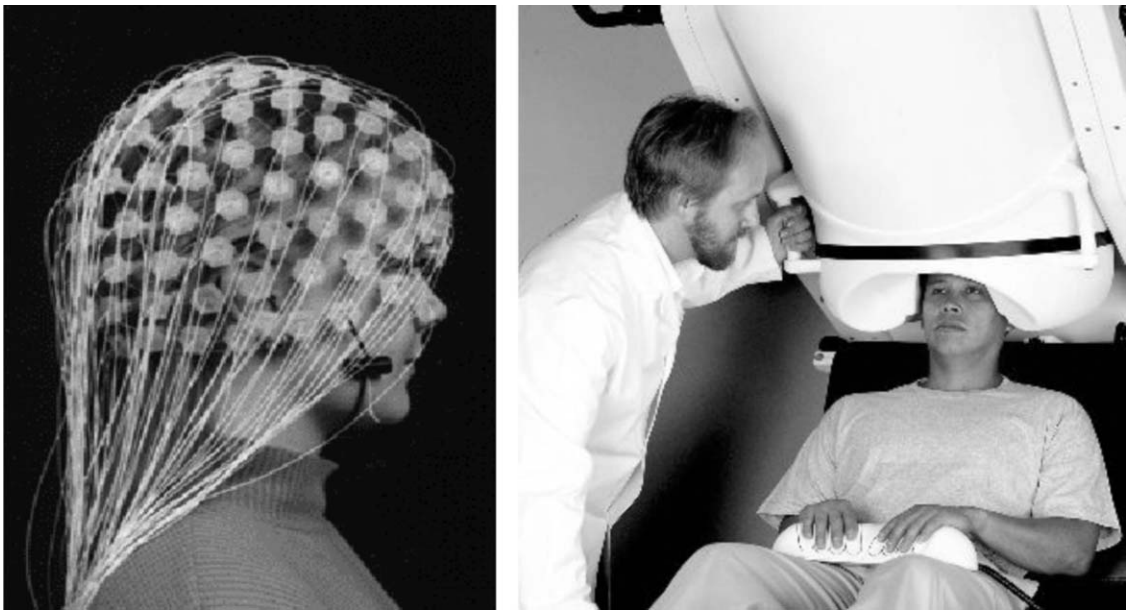


Figure 3 NEM measurement technologies. (Left) An EEG electrode array based on a mechanical tension structure (figure courtesy Electrical Geodesics, Inc.). (Right) A whole-head MEG system showing the helmet-shaped dewar that contains cryogenic fluids (figure courtesy of 4D Neuroimaging).

5–10 cm below the scalp. The fixed geometry of the MEG sensor array, the noncontact nature (and associated efficiency of subject preparation), and the relative insensitivity of MEG to head conductivity properties are all practical advantages of MEG for brain mapping applications. However, MEG sensor arrays providing full head coverage together with the magnetically shielded room used by most existing systems can cost as much as or more than a typical clinical MRI system with capabilities for functional neuroimaging.

B. Experimental Methodologies

Ongoing spontaneous activity can be recorded at the surface of the human head using MEG or EEG. Such activity typically is characterized by regions of relatively large amplitude oscillatory patterns that vary as a function of position on the head and state of the subject. Pathological responses such as certain forms of slow oscillation or the spikes and waves associated with epileptic activity often can be clearly resolved in the ongoing EEG record. The signals associated with responses to individual stimuli or other punctate cognitive or control processes typically are much smaller and require specialized experimental paradigms and signal processing techniques to pull the signals out of the noise.

1. Stimulus Evoked Responses

Brief presentations of sensory stimuli elicit a sequence of responses in the chain of specialized cortical and subcortical areas that process sensory information. In the primate visual system, over three dozen areas have been identified that are involved in the processing of visual information. Other sensory modalities such as auditory and somatosensory systems involve smaller numbers of areas and less cortical real estate but still employ distributed processing strategies.

In order to enhance the consistent aspects of the sensory response while suppressing the contribution of other physiological processes or environmental noise, most investigators employ averaging of temporal sequences time-locked to the stimulus. In this sort of *evoked response* paradigm, individual stimuli are typically presented in isolation. Different examples of a class of stimuli are often presented in a random sequence. The interstimulus interval typically is varied within some limits to minimize habituation and thwart the generation of temporal expectations by the subject; such effects can influence the amplitude and timing of certain components of the response. The time course of the electrophysiological response to a single stimulus ranges from tens of milliseconds to a few hundred. Thus, by presenting stimuli at intervals of 500 msec or more, it is possible to examine the entire time course of the response with little or no overlap from preceding or subsequent responses. Figure 4 illustrates an example of some of the techniques used for visualizing a somatosensory evoked response.

2. Steady-State Paradigms

An alternative approach for evoked response paradigms is to present stimuli at high rates, accepting the response overlap. Instead of discrete stimulus presentations, such paradigms often impose rapid changes on an ongoing stimulus, such as amplitude modulation of a continuous tone or contrast reversal of a patterned video display. Because the response does not return to baseline between stimuli (i.e., residual activation is maintained between trials), this class of techniques is referred to as *steady-state* methods. However, when using typical AC-coupled recording methods, the response must be modulated in order to be detectable. The time-locked response is typically isolated using Fourier transform techniques. The amplitude at the stimulation frequency is taken as a measure of the evoked response, whereas the phase is taken as a measure of the temporal delay (or latency) of the response.

The primary advantage of this strategy is speed; steady-state methods provide an efficient way of collecting and analyzing topographic data produced by MEG or EEG. However, such methods also have disadvantages. Fast stimulus presentations produce a measure of sensory overload, often leading to habituation or reduced levels of attention to stimuli. Steady-state methods may also introduce phase ambiguities. For example, visual evoked responses show evidence of an initial activation in layer 4 of the primary cortical visual area, V1, followed by activation in other layers and eventually by feedback activation in V1 from higher visual areas. At high stimulation rates the temporal relationship between various phases of the V1 cortical response may be obscured. Further, the subsequent activation of other visual processing areas may produce field or potential topographies that overlap with the responses of earlier areas. The loss of timing information removes an important tool for the identification of sources and for studies of the dynamics of information processing.

Experimental studies employing MEG have demonstrated a clever application of high-frequency stimulation techniques. By modulating the stimulus at frequencies that may be too high to consciously perceive, it is possible to frequency-tag the downstream response. Such methods can be used to tag the hemifield of stimulation in a wide-field visual stimulus or to identify the stimulated ear in a dichotic listening paradigm. Residual modulation at the tag frequency can be used to identify the origin of a response even after the arrival of the signal at a higher sensory processing area with convergent bilateral inputs.

By presenting fast pseudo-random sequences of individual stimuli, it is possible to achieve much of the efficiency of steady-state techniques while avoiding several of the problems. For example, a video display can be divided into elements that are turned on and off in an apparently random sequence, such as an m-sequence. The temporal activation sequences are designed to be orthogonal, i.e., each element has its own unique activation sequence. This allows correlation techniques to be used to extract the spatial and temporal patterns of response to each element of the display. This method can be very efficient because many stimulus elements can be presented simultaneously. However, this leads to stimuli that are decidedly nonphysiological and may be a bit disconcerting. At present, the method appears to be more useful for rapid mapping of the systematic parametric organization of primary sensory areas (i.e., retinotopic, tonotopic, or somatotopic projections) than as a tool for probing higher cognitive processes.

Sensory evoked response paradigms provide a powerful and robust tool for probing the functional architecture and dynamics of the neural systems devoted to the processing of sensory information. However, this is only one of several classes of functional activity within the human brain.

3. Motor Control

The control of voluntary movement is another critical capability of higher organisms. Although the coordination and fine-tuning of movement appear to be distributed through several brain centers, the initiation of movement is a function of the motor cortex, located adjacent to primary somatosensory areas in humans. Some investigators have used cued trials to study motor function, hoping that the temporal jitter in the reaction time does not wash out the targeted motor response. This strategy also produces a time-locked response to the sensory cue, which can interfere with the analysis of the motor response. An alternative strategy is to use self-paced tasks and to time-lock averaging to the motor response. For simplicity, the response can be a button press or similar action that is readily converted into an electronic timing signal. A bit more sophisticated strategy is to base experimental timing on a measured physiological response. For example, it is possible to record an electromyogram signaling the activation of specific muscle groups by using the same basic technology as EEG. These are large, robust responses that can be detected in a continuous recording, often using simple threshold techniques. This sort of activity, which may not be evoked directly by external stimulation but is associated with externally observable consequences, is referred to as an *event-related response*. In addition to motor processes or simple behavioral trials such as reaction time detection tasks, event-related experimental designs are often used to probe high order cognitive processing activity. In some cases, the "event" is only apparent as an internal state, e.g., conjunction of a particular stimulus with a specific behavioral or cognitive task.

4. Cognitive Processes

The general strategy for event-related cognitive studies is to employ a set of tasks designed to isolate and contrast the processes of interest. Measures derived from MEG or EEG are used as an index of activation and, thus, of the underlying cognitive processes. Such studies often employ well-balanced control trials, which account for sensory or motor components of a response while manipulating the relevance or difficulty of the cognitive component. Such strategies have been used to isolate and probe various aspects of selective attention. For example, many designs used for attention studies employ a cue stimulus presented before the probe trial to direct attention to one region of the visual field or another. The subject might be instructed to respond to a particular type of stimulus only when presented at the attended site. Thus, in the same experiment, a given physical stimulus might serve as the response target, an inappropriate stimulus at the attended site, or an irrelevant stimulus that can be ignored.

Other designs tap the endogenous cognitive skills of the subject. For example, a list of real words might be presented visually along with interspersed pseudowords (i.e., pronouncible constructs that look like words but have no meaning) or nonwords. The nature of the observed response varies as a function of the

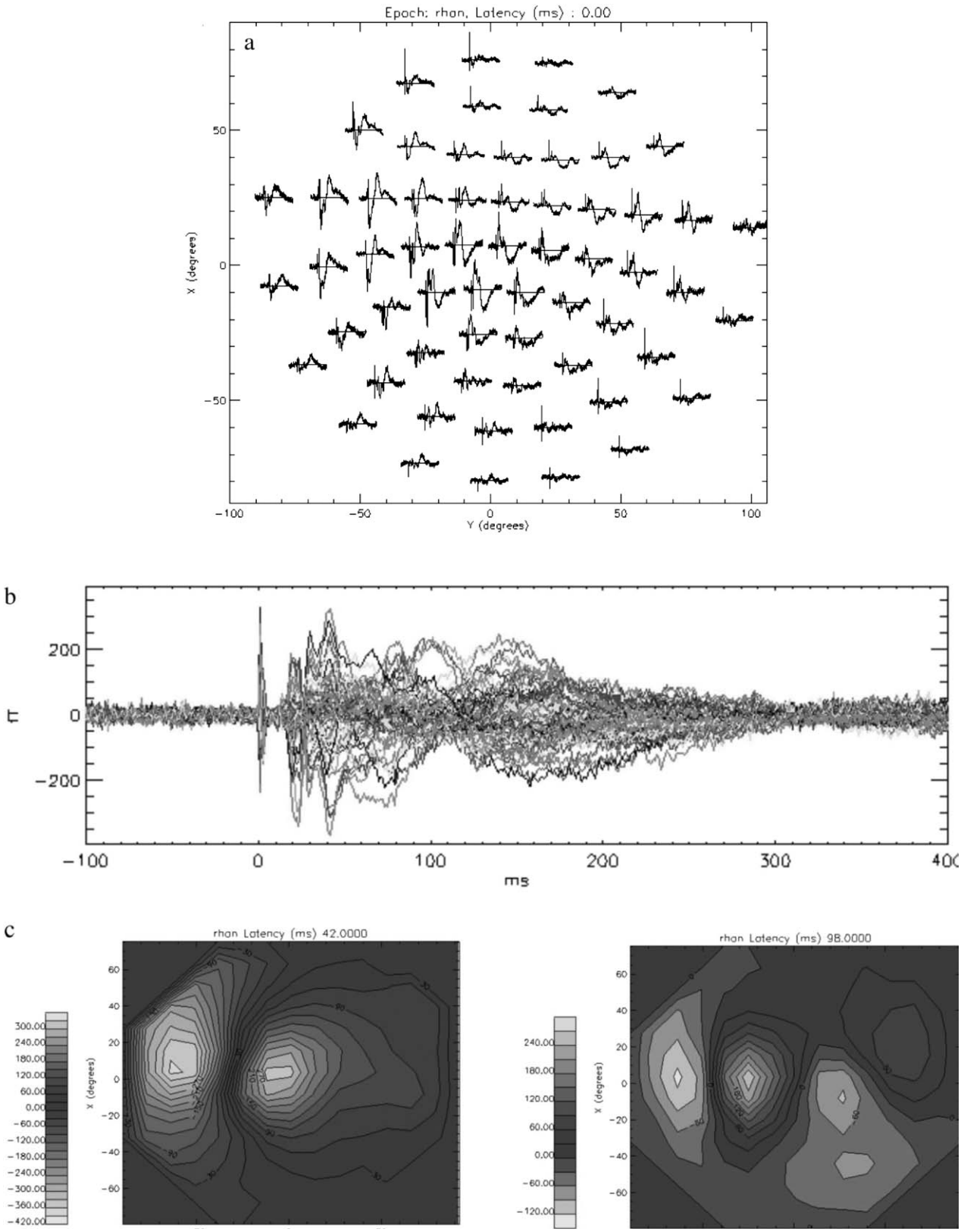

Epoch: rhan, Latency (ms) : 0.00
a
X (degrees)
Y (degrees)
b
ms
c
rhan Latency (ms) 42.0000
rhan Latency (ms) 98.0000
X (degrees)

Figure 4a–c

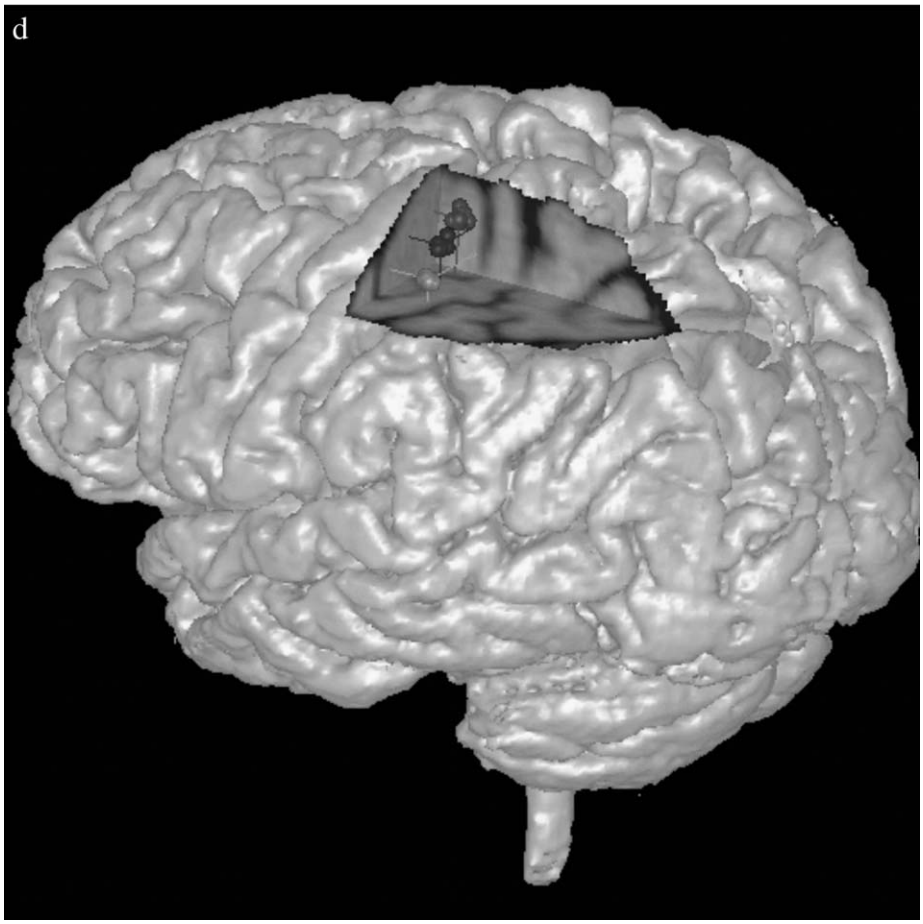


Figure 4 Several views of somatosensory evoked response recorded with MEG. (a) A depiction of the evoked response waveforms recorded over the surface of the head. Waveforms are positioned according to the polar projection of the corresponding sensor locations. (b) An envelope or "butterfly" plot of the stacked waveforms reveals the balanced polarity of MEG responses, allowing appreciation of temporal relationships between waveform features. (c) Contour plots of the field maps observed at two latencies of the evoked response. Field maps are plotted in polar coordinates. The left panel suggests a simple activation pattern early in the response. The right panel suggests the presence of an additional source in the contralateral hemisphere. (d) Equivalent current dipole sources estimated from early latencies localize to the posterior bank of the central sulcus.

lexical class of stimulus. A spoken sentence might be presented in which the final word either completes a reasonable semantic construct or renders the sentence nonsensical. Such paradigms can tap the mechanisms by which we comprehend language.

C. Signal Processing

Event-related and evoked responses in MEG and EEG are typically quite small compared to ongoing neural activity or other physiological activity or environmental noise. Effective signal processing strategies are essential for isolating the desired signals from other observed signals not of interest.

Temporal filtering is a ubiquitous initial step in the signal processing chain. Virtually all MEG and EEG systems employ some sort of analog filtering on the front end to reduce the signal bandwidth to avoid aliasing in digitization, if for no other reason. Many conventional EEG systems provide an extensive set of analog filtering capabilities, including high-pass and low-pass filters as well as notch filters (often tuned to the 50- or 60-Hz frequency of the power grid), to eliminate major sources of environmental interference. If the frequency content of the targeted signal is known or can be determined empirically, reduction of the bandwidth of the measurement reduces noise and can increase the efficiency of signal acquisition. However, as high-density sensor systems and digital acquisition

and processing become more standard, instruments are increasingly designed with fixed analog front ends. Signal processing is increasingly performed by digital signal processing subsystems, general purpose computers embedded in the data system, or workstations used by investigators for subsequent analysis.

Digital filtering is cheaper and in many respects technically superior to the use of analog filters. Methods based on the fast Fourier transform (FFT) greatly simplify the algorithms required to implement digital temporal filters. Related techniques are used for the frequency domain analysis of the major rhythms characteristic of spontaneous EEG or MEG recording or for the extraction of steady-state responses at the stimulation frequency. Short time base analyses based on wavelets or special FFT algorithms may be used to identify periods of transient synchronization (or desynchronization) that are associated with (but not time-locked to) external stimuli or perceptual states. Such analyses are often based on trial by trial analysis of continuously acquired data. The averaging techniques used for most event-related response work would attenuate or eliminate such responses because they are not time-locked to the stimulus and may vary in time and phase from one response to another.

Eye blinks or other movements of facial muscles can produce strong signals in MEG or EEG. The startup of an electric motor or power-hungry instrument in the vicinity (or sharing a power circuit) can also introduce large artifacts. A single large-amplitude transient may leave significant residue in an averaged event-related response even after many trials. For this reason, *artifact rejection* is another common step applied in the analysis of MEG and EEG data. The digital data stream is monitored for values that exceed some criterion threshold. If the threshold is exceeded, the offending channels or, more commonly, the entire trial is excluded from further analysis. For some clinical applications, artifact rejection is typically accomplished by visual inspection. In epilepsy, the pathophysiological responses of interest may exceed the size of signals considered artifacts under other circumstances, and inspection by a trained clinician may be the most efficient and effective way to identify both artifacts and epileptiform activity. However, as data streams become more dense and algorithms become more sophisticated, there is increasing reliance on software that can categorize events in the experimental record on the basis of temporal waveforms, spatial topographies, or both. Such systems are often used to preprocess an extensive record, bringing interesting or suspicious events to the attention of the reviewer.

Time-locked selective averaging is the mainstay of most existing work with event-related or evoked responses. As a first step, epochs in the data are identified and characterized according to the nature of the stimulus, the response, or the task and its performance. Averaging can even be undertaken relative to a reproducible endogenous transient, such as an epileptiform spike. Segments of the waveform data across all channels are selected relative to the timing event, and within each channel the corresponding time points in the segment are averaged. In general, averages are segregated according to the particulars of the trial, e.g., the identity of the stimulus or the accuracy of task performance, although in many cases averages are collapsed across conditions that are considered irrelevant for a particular experimental question. Many studies report "grand averages" constructed by averaging responses across many subjects in order to increase the power of statistical inference, although this practice probably precludes reliable source localization.

The construction of *difference waves*, by subtracting a control condition from a particular response, is another time-honored method for the analysis of event-related response data. For example, such methods clearly disclose the increase in response amplitude associated with selective attention to a particular location in the visual field and isolate responses associated with certain endogenous cognitive responses. In many cases such records are reduced to a single response waveform, for example, by averaging (or summing power) across all channels or a selected subset. However, an increasing number of investigators construct difference topographies by subtracting control responses from the corresponding signal channels. These distributions are often analyzed in the same way as the underlying event-related signals, for example, subjected to source localization procedures.

Such analyses should be approached with caution. The methods will work, in principle, if the only differences between conditions are due to the strength of the underlying sources or the appearance of a new source under particular experimental conditions. Unlike PET and fMRI, which in effect produce an estimate of the distribution of source activity, EEG and MEG produce topographic maps with a complex relationship to source activity, driven by the physical properties of the measurement instrument and the system under study. Minor changes in the location or extent of activation (especially in cortical regions of high curvature) can produce big changes in the

observed field topography and, thus, significant changes in computed difference fields.

Single-pass analytical methodologies are increasingly applied to the analysis of event-related response data. Frequency decomposition techniques (described previously) have been used to explore the putative role of transient phase-locking of oscillatory activity in certain perceptual processes. Correlation techniques can be applied to continuous evoked response data to identify consistent features in a manner analogous to signal averaging. Spatial filtering techniques compute a linear transform (based on a computed or assumed source model) that can be applied to the data to estimate the activation time course of the source. Several investigators suggest that techniques in this class (such as minimum variance beam-forming) are most useful if applied to single trial data rather than averaged responses. Some new methods such as independent component analysis (ICA), synthetic aperture magnetometry, or magnetic field tomography by the nature of the algorithm are most effectively applied to the analysis of single trial data.

IV. RESPONSE DYNAMICS

The principal strengths of neural electromagnetic methods stem from their capacity to define the dynamics of neural population activity. Even a single electrode pasted to the scalp may disclose a complex temporal waveform consisting of a series of peaks and valleys. In some cases, a peak in a waveform at a particular latency is observed across a large subset of the channels in a whole-head sensor array, even though the amplitude or the polarity of the peak changes. This pattern is characteristic of a single anatomical source or a set of sources acting in synchrony. In some cases, a close examination of the waveform montage collected from a sensor array discloses that what appears as a single peak in one waveform can be resolved as multiple overlapping temporal peaks observed at other locations. In such cases, the topographic map of the evoked response typically has apparent features that appear to shift systematically over the course of the response. A simple-minded analysis may suggest that the source is a single focus of activity moving through the brain volume. A more sophisticated source model may allow the same response to be decomposed into two or more component sources with stable NEM topographies and distinct but overlapping time courses.

In the absence of effective source localization techniques, there was a tendency in early work to focus on the peaks in the waveform as the unitary building blocks—the components from which complex event-related responses were built. Components were given names on the basis of the polarity and latency of the waveform peaks: e.g., the N100 (a negative-going response component peaking around 100 msec poststimulus) or the P300 (or P3, a positive peak in the response waveform around 300 msec poststimulus). As the characterization of response components proceeded, descriptions of the component scalp topography were sometimes added to aid identification and discrimination of named components. For endogenous cognitive response components, identifying criteria often included the nature of experimental manipulations required to elicit or enhance a particular peak in the waveform. Whereas such information is critical for investigators attempting to reproduce or extend a particular observation, it complicates the business of component quantification.

In an effort to address this concern, some investigators turned to blind decomposition techniques such as *principal components analysis* (PCA). PCA is a linear technique based on eigen analysis and singular value decomposition. The method attempts to find a set of basis functions—in this case, field or potential topographies—that can be used to reconstruct the original experimental data. In order to make the decomposition unique, principal components are constrained to be mutually orthogonal. Each principal component has an associated weighting vector that quantifies the representation in the data of each component as a function of time. In some well-behaved cases, principal components correspond to response components identified by other criteria. However, in general, the requirement for orthogonality precludes the proper identification of more than a few components. An alternative decomposition strategy has been developed that appears to hold an alternative decomposition strategy that appears to hold significant promise for functional neuroimaging applications. Independent component analysis identifies a basis set in which components are statistically independent though not necessarily orthogonal. Initial results with the algorithm are promising, although it will certainly be possible to find pathological cases that cause the algorithm to fail. It is not yet clear how effectively the algorithm will identify proper components in routine applications to event-related response data.

In a few cases, the idea that components reflect the successive activation of links in a processing chain

appears basically correct. For example, in the *auditory brain stem evoked response* (ABER), the peaks in the waveform are associated with specific structures in the early auditory pathways, and the waveform morphology can be used to assess the integrity of the relay and processing circuitry. Similarly, the earliest components of somatosensory responses evoked by electrical stimulation appear to be associated with specific anatomical loci. In contrast, for visual evoked responses the situation is considerably more complex. Although some investigators have reported an early, relatively small EEG component (N70) that was identified as the initial activation of cortical V1, most analyses have focused on the more robust P100 component. A variety of source analysis techniques indicate that this component is a complex consisting of temporally overlapping responses from several distinct though nearby visual areas. Although there is an element of sequential processing in the early visual system as activation spreads through the information processing tree, there is also considerable parallel processing. There are also forward and feedback links that skip over portions of the schematic processing hierarchy and delays within areas that can further complicate the simple orderly picture of temporal response dynamics.

V. SOURCE LOCALIZATION

The existence of detectable magnetic fields and electric potential distributions at the head surface is a consequence of the physics of electromagnetism. Given an adequate description of the source currents and the conductivity properties of the medium, it is possible to compute the anticipated field topographies. Calculations to solve this so-called *forward problem* can employ models of greater or lesser detail, depending on the complexity of the system and the required degree of accuracy. Computation of the forward problem for more complex source distributions—extended regions of activation or multiple active sources—is more time-consuming, but not significantly more difficult. The principle of superposition tells us that the contributions for multiple source currents will sum linearly.

In principle, it may be feasible to invert the process—to compute the currents that give rise to an observed field or potential distribution at the head surface. Unfortunately, this *inverse problem* is not well-behaved. In general, many different current distributions can produce the same set of surface measurements. To appreciate the problem intuitively, consider a homogeneous spherical volume that is conductive. It is possible to account for any given potential or magnetic field distribution measured at or above the surface, with a suitable collection of currents limited to the surface of the sphere. However, we can define another spherical shell 1 cm below the surface and derive another current distribution to account for the same set of observations. Given this fundamental ambiguity, how can we hope to reconstruct the proper set of current sources buried deep in the brain from the data available at the head surface?

The general strategy is to build a model of the sources that might produce the observed responses. The source model defines the structure of the solution. Model parameters define the details. In some cases, source models are very restrictive, so that a single, best-fitting set of model parameters can be found. In such cases, the accuracy of the solution depends on the applicability of the source model. As the complexity of the source model increases, generally the number of parameters also increases. This allows more complex source distributions to be modeled, but tends to increase the ambiguity of the reconstruction problem. By defining the criteria that we prefer in an acceptable solution it is generally possible to find a solution, but again, the accuracy of the solution depends on the validity of the assumptions.

A. Model-Based Approaches

Thus, source localization depends on the use of nested computational models that describe the distribution of neural currents and that predict the observable consequences of those currents. These models are based on implicit or explicit knowledge and assumptions about the nature of the system. The first 50 years of work with EEG involved little quantitative effort to localize the sources of observed topographies in the surface potential data. The development of MEG and the recognition that many observed field distributions could be explained by a simple forward model led to advances in procedures that have subsequently been applied to EEG data.

B. Forward Modeling

For electrical or magnetic measurements at a distance significantly larger than the extent of the source, the spatial fine structure of the field distribution is not

detected, and the dominant contribution is the dipole associated with longitudinal intracellular currents. In MEG, the effects of ohmic currents through the head volume are minimal, and a reasonable analytical solution can be derived by treating the head as a homogeneous conducting sphere. The radius of the sphere is chosen to approximate the inner surface of the skull, based on individual anatomical images or external measures of head shape coupled with a knowledge of average anatomy. Unlike MEG, EEG signal topographies are strongly influenced by the conductivity properties of the head. Even simple models of elemental EEG signals must take into account the presence of multiple tissue layers of differing conductivity. Dipole response topographies for EEG can be computed in a volume model consisting of multiple spherical shells, with layers corresponding to brain, cerebral–spinal fluid, skull, and scalp, using a truncated series of Legendre polynomials. Many researchers in the field have argued that this class of model is more than adequate for source localization, given the uncertainties associated with the inverse problem. However, as inverse procedures have improved, a number of studies have underscored the value of more sophisticated forward models for source localization accuracy. Some of these approaches are summarized in Fig. 5.

A significant improvement in the accuracy of the forward calculation can be achieved by *boundary element* calculations incorporating the geometry of the major tissue classes within the head. Surface meshes are constructed to approximate conductivity boundaries. Because the conductivity of the skull is significantly lower than that of other tissues in the head, it is particularly important to capture the geometry of the skull near sensors. Most calculations of this sort are limited to simple topologies, e.g., nested compartments without intersections or penetrations. Conductivity values are typically taken from the literature and correspond to measurements originally made in cadaver tissue. Because there is no basis for further subdivision, conductivity is taken as homogeneous within a compartment. Although the boundary element method employs simple geometries, the calculations are relatively time-consuming, because the solution matrix typically contains terms for the interactions between every pair of nodes in the mesh. Numerical studies have demonstrated that it is possible to achieve accuracy approaching that of the boundary element methods at much lower computational cost by approximating the skull boundary with local spheres selected for each sensor.

3D methods provide an alternative strategy for high-resolution forward calculations. In *finite difference* (FD) or *finite element* (FE) forward calculations, the head volume is divided into a collection of volume elements that form the computational mesh. Potentials are computed at nodes, typically associated with boundaries or vertices of the volume elements. These calculations assume a current source that generates the potential and account for the conductivity properties of the volume elements that strongly influence the spread of potential. FE methods typically employ an irregular mesh composed of tetrahedral elements. This allows the mesh to closely match conductivity boundaries within the medium that may give rise to sharp gradients in the potential distribution, providing greater accuracy with a smaller number of volume elements. However, most present methods for mesh generation and refinement require considerable human intervention. FD methods typically employ a regular (e.g., rectangular) mesh. Segmentation or voxel classification schemes can be applied to regular volumetric data such as MRI to produce a data volume that can be used for calculations without constructing specialized meshes.

3D methods are most useful if we have access to external information that can be used to define the geometry or electrical properties of the conductive medium. MRI provides the most accessible and flexible measure of tissue properties, although X-ray CT provides a better definition of the geometry and microstructure of the skull. New and evolving MRI techniques will eventually provide even more useful information for forward modeling. *Current density MRI* operates by applying external currents at the head surface and measuring the perturbation of the image by local volume currents. This allows an estimation of head conductivity on a voxel by voxel basis. *Diffusion tensor MRI* measures the magnitude and characterizes the direction of diffusion of water within tissue. Structures such as white matter tracts give rise to anisotropic physical properties that can be described by a tensor, e.g., the apparent diffusion coefficient of water and the measured conductivity are much higher along a fiber tract than transverse to it. MRI can be used to estimate the anisotropic conductivity of such tissue. FE and FD methods can accommodate anistropic conductivity (though again the FD application is simpler). By passing currents through pairs of surface electrodes and estimating the induced potentials at other electrodes in the array, electrical impedance tomography (EIT) allows an estimation of the bulk properties of major tissue

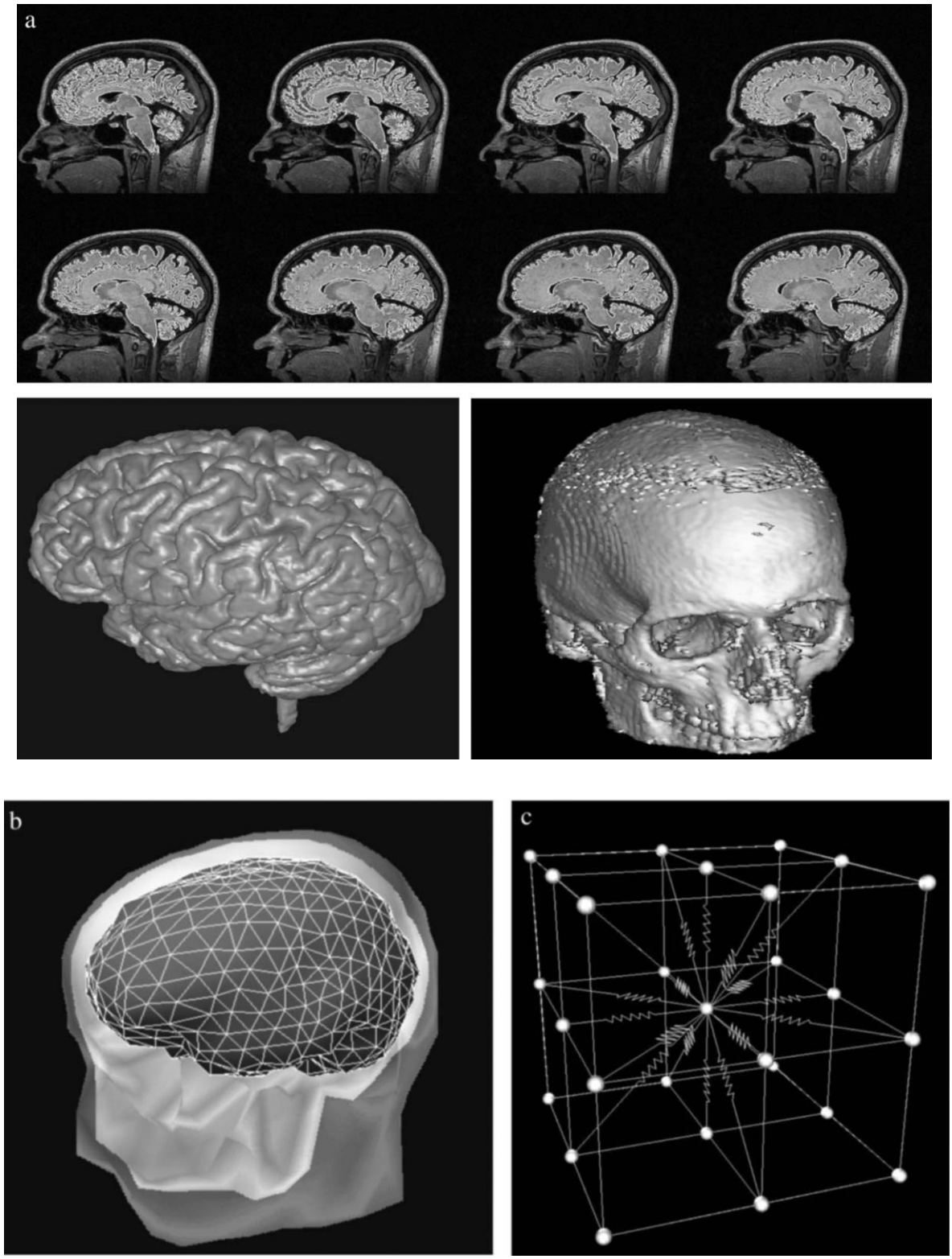


Figure 5a–c

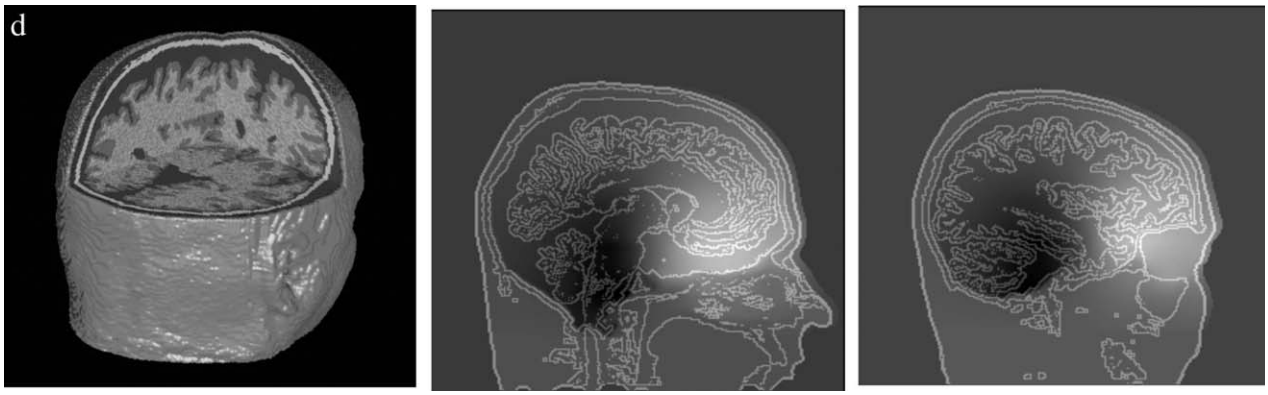


Figure 5 NEM forward modeling in realistic geometries. (a) Computational tools for interactive and semiautomatic segmentation of cortical anatomy allow extraction of computational geometries. Upper panel: Region-growing algorithms with adaptive criteria perform segmentation of white matter and identification of gray matter by dilation. Lower left: 3D rendering of the cortical surface identified by an automatic algorithm. Lower right: Rendering of the skull segmented by region-growing techniques from 3D MRI data. (b) Boundary element mesh based on simplified skull and scalp geometry derived from MRI volume imagery. (c) The regular computational mesh employed for finite difference computations in anisotropic media. The nodes for potential computation are at the corners of the spatial volume elements. (d) Finite difference calculation of potential distribution, using a detailed computational geometry derived from MRI. The current source is located within the temporal lobe, with a posterior to anterior orientation. The slices from the computed potential distribution show evidence of current leakage through the skull penetrations of the optic nerve.

classes within the head and even a measure of 3D reconstruction. Tomographic reconstruction of head conductivity requires 3D computational techniques and is greatly facilitated by accurate geometrical information drawn from MRI.

C. Source Model Estimation

Source localization from EEG and MEG began with educated inspection of surface field topographies. In EEG, a radially oriented current will produce a potential extremum over the source. In MEG, a radial current source produces little externally detectable magnetic field, but the tangential component of a neural source produces a field distribution with extrema that straddle the source. The distance separating the field extrema allows an estimation of the source depth, given assumptions regarding the nature of the current source. In both EEG and MEG, many observed response topographies can be explained by an *equivalent current dipole* (ECD) source, i.e., an isolated point current with a given location, orientation, and amplitude. Theory suggests that such a model provides a reasonable estimate of the field or potential distributions due to a small cluster of oriented neurons measured at a distance. Even extended patches of activated cortex often produce a dipole-like distribution, although the estimated location and strength of the ECD will contain systematic errors. An extended patch of parallel current elements produces an ECD estimate that is deeper and stronger than the center of mass and integral current estimated from the actual source distribution.

Because field topographies are typically diffuse, source estimation by eye is a rather inexact process. Some investigators have employed image processing techniques to allow for easier detection of features in the field topography. For example, computation of the spatial derivative (the Laplacean) of the observed magnetic field or potential topography tends to place maxima over the current sources. Indeed, one form of MEG sensor—a first-order planar gradiometer—effectively implements this transform in the detector coil configuration.

Although *inspection of response topographies* is a useful starting point for finding NEM sources, model-based parameter estimation procedures provide a more objective strategy that allows quantitative assessment of the goodness of fit. Nonlinear optimization procedures such as simplex or gradient descent allow the estimation of a dipole source; current orientation and strength are linear parameters that can be optimized separately. If a single focal current source is active in a simple medium, such procedures can localize it very precisely. However, at any given

instant in time, many sources may contribute to a given response topography. Because of superposition, a combination of several sources may give rise to a distribution that is adequately modeled by a single source at an entirely different location. Knowledge of the number and nature of active sources can reduce the ambiguity of the source localization problem. Spatiotemporal source localization techniques attempt to find a minimal set of sources, each with an associated time course, that explains the observed distribution across a defined interval within an event-related response. If the sources are activated asynchrously, this strategy can be very effective for decomposing and localizing the contributions of individual sources.

Nonlinear optimization methods can be used to find locations for a collection of simple current sources that might explain changing field topographies observed across time. Because the time course is an estimate of the source current amplitude as a function of time, linear methods can be used to optimize the estimates between iterative nonlinear optimization steps. This general strategy has been the most widely used approach for NEM source localization for over a decade, but modifications, extensions, or alternatives to the method can provide enhanced performance. A common problem with nonlinear methods is that they require a starting estimate of model parameters. This may be provided by an informed analyst or by a stochastic process. If the starting estimate is close enough, an optimization procedure that follows the error gradient can find the best-fitting model. If the starting estimates are far off and the error surface is complex (often the case with multiple-source models), the procedure may fall into a local minimum that is not globally optimal. A conscientious analysis can combat this problem by running the optimization procedure with multiple sets of starting parameters. *Multistart* procedures automate this process, employing a numerical algorithm to generate random starting parameter estimates. Such methods often find a consistent set of best-fitting source models that are globally optimal. Other methods such as *simulated annealing* or *genetic algorithms* employ alternative strategies to address the problem of local minima within the model parameter error space. Examples of these approaches are illustrated in Fig. 6.

The multiple signal classification (MUSIC) algorithm operates within the same framework (i.e., a multiple-dipole spatiotemporal model) but employs a more systematic strategy for finding sources. The array of measured potential or field values at any given instant in time is treated as a multidimensional vector in the space of possible measurements. Similarly, the topography associated with any given dipole is another vector in the same space. The algorithm operates by systematically stepping through a set of possible sources (e.g., a grid of locations within the brain) and evaluates the match between the source field vector and the collection of signal vectors across time. The sources that most closely match the observed signals across time are considered the most likely. The method is very effective and can be exhaustive, avoiding problems of local minima seen with nonlinear optimization techniques. Multiple sources with highly correlated time courses create problems for the algorithm, but enhanced methods such as recursively applied (RAP) MUSIC address this and other concerns.

A second general strategy is to use *linear inverse techniques* to solve a large and general source model. The reconstruction space is defined by a regular grid, a collection of voxels, or vertices from a computational mesh. One to three current elements are associated with each possible source location. The reconstruction procedure employs the Moore–Penrose pseudo-inverse to assign a current value to each model element. This procedure, based on singular value decomposition, estimates a source distribution with minimum power over the collection of driving currents. A number of variants on this *minimum norm* procedure have been described, mostly based on different strategies for weighting the lead field basis matrix in order to select a solution with desired properties. Anatomical constraints based on cortical geometry can improve accuracy and efficiency. However, even with substantial reductions in the source space based on anatomy, the inverse is a highly underdetermined problem. There are many more source model parameters to estimate than the number of independent measures available from MEG or EEG.

The major problem with minimum norm procedures is that there is no guarantee that the solution of minimum Euclidean norm (i.e., the sum of squared currents) will be representative of the true solution. Because of the strong dependence of measured magnetic field on distance from the source, the basic minimum norm procedure tends to produce diffuse, superficial current reconstructions, even when the reconstruction space is constrained to the cortical surface. Currents closer to the sensor array can account for more power in the field map with less current and therefore are favored by the method. However, more current elements are required to account for the shape of a field distribution that may actually arise from a more focal but deeper source.

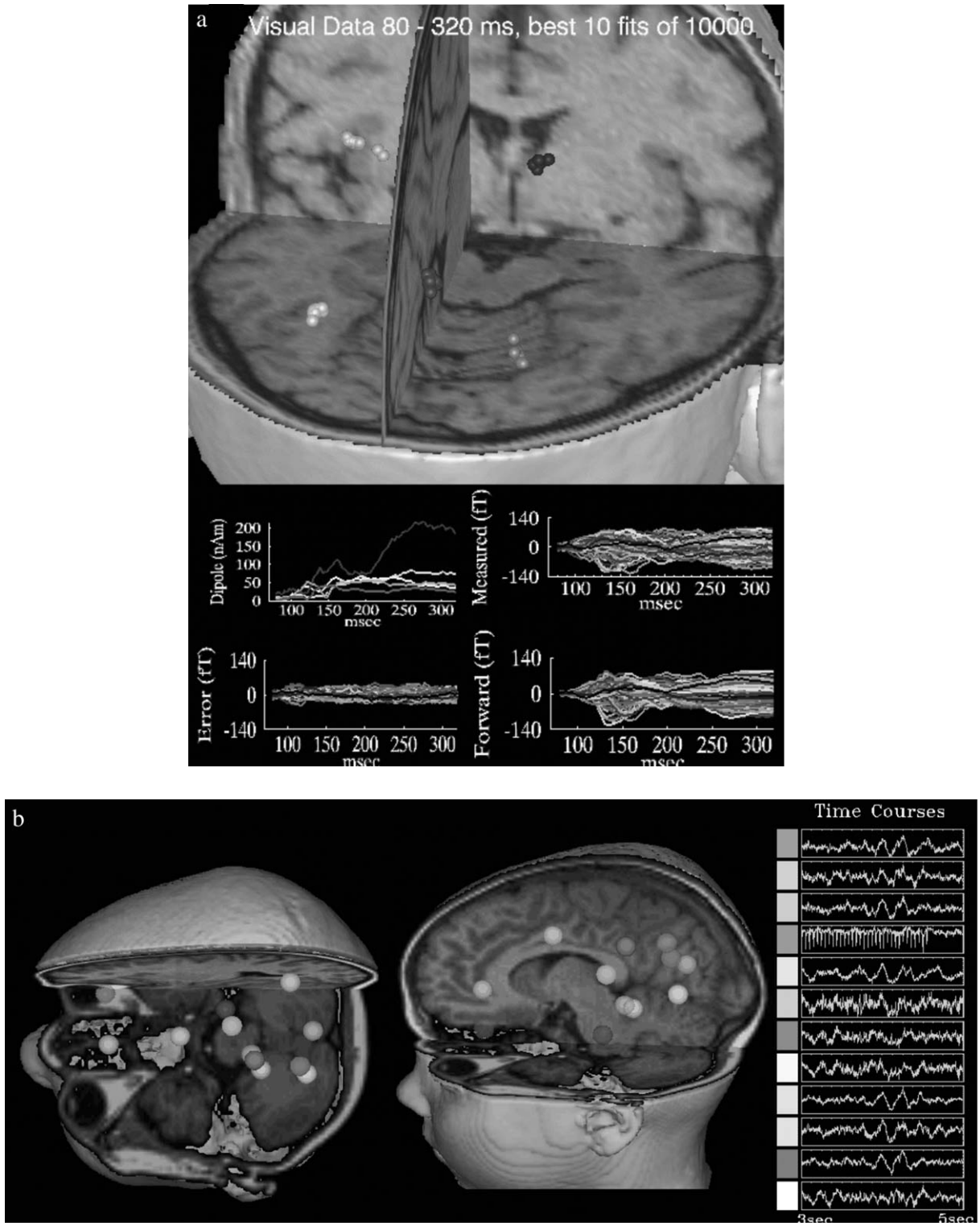


Figure 6 Source localization by multiple dipole spatiotemporal estimation procedures. (a) Dipole source locations and time courses estimated from visual evoked response data. This analysis used a multistart algorithm; the best 10 solutions from 10,000 generated are tightly clustered, suggesting that the algorithm has found a global best fit. (b) Dipole locations and time courses associated with an epileptic response from a photosensitive child. Dipole locations were estimated with a genetic algorithm. Note that the magnetic response of the eye contains evidence of the strobe flashes that triggered the response.

In order to combat this tendency, it is possible to scale the field distributions (or, alternatively, the strength of the unit currents) in order to normalize the field power associated with each elemental source. Pseudo-inverse procedures based on a normalized basis matrix offer some improvement in the fidelity of reconstructions. Explicit or implicit basis matrix weighting procedures have proven to be a useful general strategy for modifying the properties of reconstruction algorithms. The FOCUSS algorithm employs an iterative reweighting procedure to derive sparse reconstructions based on focal activated patches. The LORETA algorithm uses an alternative weighting scheme to find current reconstructions that are maximally smooth.

Given the fundamental ambiguity of the inverse problem and the complex error surface associated with the parameter space, there is no guarantee that the proper form of source model (e.g., the number of active sources) can be determined or that a single global minimum will be found. The estimated parameter values critically depend on model assumptions and may vary widely as a function of small amounts of noise in the data.

D. Bayesian Methods

Bayesian analysis techniques provide a formal method for integration of prior knowledge drawn from other imaging methods. In pure form, Bayesian techniques estimate a posterior probability distribution (a form of solution) based on the experimental data and prior knowledge expressed in the form of a probability distribution. In additional to providing a flexible mechanism for multimodality integration, these techniques allow rigorous assessment of the consequences of prior knowledge or assumptions about the nature of the preferred solution. Several investigators have explored traditional Bayesian methods, seeking a single "best" solution that satisfies some criterion, such as maximum likelihood, maximum *a posteriori* (MAP), or maximum entropy solution. However, any given single solution is effectively guaranteed to be inaccurate, at least in its details.

A technique for Bayesian inference has been developed that addresses this concern by explicitly sampling the posterior probability distribution. The strategy is essentially to conduct a series of numerical experiments and determine which solutions best account for the data. To make the method efficient, a Markov chain Monte Carlo (MCMC) technique is employed. After the algorithm identifies regions of the source model parameter space that account for the data by a stochastic process, the algorithm effectively concentrates its sampling in that region. Thus, in the end, samples are distributed according to the posterior probability distribution—a probability distribution of solutions upon which subsequent inferences are based. The Bayesian inference method does not employ optimization procedures and does not produce an estimate of the best-fitting solution. Instead, it attempts to build a probability map of activation. This distribution provides a means of identifying and estimating probable current sources from surface measurements while explicitly emphasizing multiple solutions that can account for any set of surface EEG–MEG measurements.

This method for Bayesian inference uses a general neural activation model that can incorporate prior information on neural currents, including location, orientation, strength, and spatial smoothness. Instead of equivalent current dipoles, the method uses an extended parametric model to define sources. An active region is assumed to consist of a set of voxels identified as part of cortex and located within a sphere centered on cortex or a patch generated by a series of dilation operations about some point on cortex. In a typical analysis, 10,000 samples are drawn from the posterior distribution using the MCMC algorithm. Despite the variability among the samples, several sources common to (nearly) all are often apparent. Features such as these are associated with a high degree of probability. By keeping track of the number of times each voxel is involved in an active source over the set of samples, it is possible to build a probability map for neural activation and to quantify confidence intervals. In addition to information about the locations of probable sources, the Bayesian inference approach also estimates probabilistic information about the number and size of active regions. Figure 7 illustrates several aspects of this approach to Bayesian inference.

VI. MULTIMODALITY TECHNIQUES

A growing body of evidence suggests that there is a good if imperfect correspondence between neural electrical activation and the fMRI BOLD response and that convergent information can be used to improve the reliability of macroscopic electrophysiological techniques. Because of the ambiguity associated with the neural electromagnetic inverse problem,

a number of investigators have pursued the strategy of using fMRI to define the locations of activation while using MEG or EEG to estimate time courses. Although this approach may have considerable value, it also has its pitfalls. In general, there is no guarantee that activation seen in one modality will be apparent in the other. The relationship between the precise areas of increased bloodflow and electrophysiological activation is not certain. If we assume that anatomical MRI constrains the location and orientations of possible source currents and that fMRI provides an estimate of the identity and relative strengths of active voxels, it is possible to compute the field topography associated with an extended source of arbitrary shape and size. Alternatively, fMRI can be employed as a method to *seed* dipole source estimates, which are optimized subsequently by using standard nonlinear procedures. This strategy provides a measure of flexibility to account for mismatches between assumptions and source model.

Other investigators have employed a form of weighted minimum norm to combine fMRI and NEM data. The inverse solution is constrained to lie within the cortical surface, and source current orientation may be constrained to lie normal to the local surface. By weighting the reconstruction according to the spatial pattern of apparent activation disclosed by fMRI, it is possible to guide the minimum norm reconstruction to preferentially place current in those regions. Because Bayesian methods explicitly employ prior knowledge to help solve the inverse problem, they provide a natural and formal method to integrate multiple forms of image data. The simplest strategy is to use fMRI data as a prior. This method can profitably employ strategies for the analysis of fMRI data that quantify the probability of activation in any particular voxel on the basis of fMRI data. Bayesian methods will also benefit from the efforts to develop probabilistic databases of functional organization. Figure 8 illustrates two approaches employed for the integration of MEG and fMRI data.

VII. TYPES OF EVENT-RELATED POTENTIALS

Model-based source localization, computational decomposition of complex responses, and quantitative estimates of the activation time courses of identified brain regions are all relatively new tools for the analysis of event-related neural electromagnetic data. However, 50 years of macroscopic electrophysiological study of neurological, psychological, cognitive, and behavioral processes of the human brain have produced a rich legacy of experimental observations and interpretation. Many of these results are based on the analysis of features in the response waveform, often averaged across many sensor locations and sometimes across many individual subjects. A comprehensive summary of such results is well beyond the scope of this article. Excellent review articles and even entire volumes have been devoted to small corners of the field. We will briefly review some of the major classes of event-related responses that have been identified and characterized.

A. Evoked Responses

Evoked responses are elicited by sensory stimulation and are usually recovered by signal averaging time-locked to stimulus delivery. Response waveforms consist of a stereotypical series of peaks and valleys observed between 2 and 250 msec poststimulus. Such features reflect the sequential and parallel activation of multiple specialized processing waystations within the targeted sensory pathway, although individual response components may represent simultaneous activity in several distinct areas. The responses are a product of the network architecture and the dynamics of participating neurons and are reasonably robust. Sensory evoked activity can often be recorded even under anesthesia.

Evoked responses have been used to probe the architecture and the information processing activities of sensory systems. Such work has provided evidence for specialized processing modules, corresponding to the discrete areas identified through invasive physiology, and has disclosed systematic projection of the sensory parameter space onto cortex. Other studies have provided evidence for early modulation of sensory responses by selective attention and demonstrated dynamic plastic changes in the functional organization of cortex, based on patterns of sensory coactivation or neglect.

1. Somatosensory Evoked Responses

The somatosensory system can be activated through tactile stimulation of almost any body surface. Particularly robust responses can be elicited by electrical stimulation, for example, of the median nerve of the forearm. The earliest responses to somatosensory

stimulation are observed from 9 to 15 msec poststimulus and represent responses generated in the spinal cord and brain stem. Initial responses in primary somatosensory cortex are observed around 20 msec poststimulus. An evolving complex of responses typically lasts over 150 msec. Source modeling studies identify at least 5–6 discrete regions of activation. Studies of the organization of the primary somatosensory area (S1) disclose a systematic projection of the body surface onto cortex adjacent to the central sulcus,

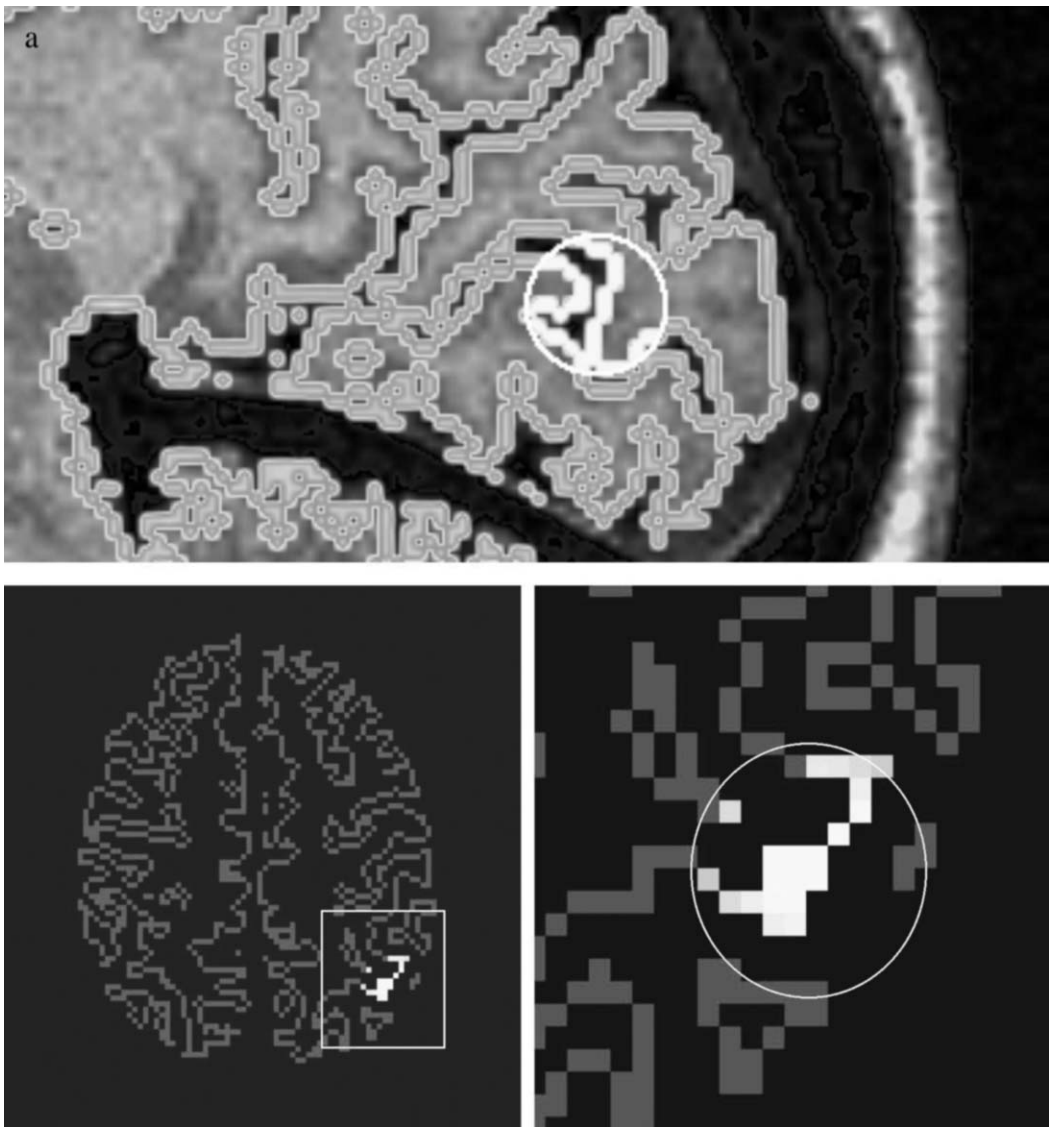


Figure 7 Analysis of MEG evoked responses by Bayesian inference. (a) Extended parametric source models used for Bayesian inference. Upper panel: A source defined by the intersection of cortex with a sphere centered on cortex. Note that adjacent sides of a sulcus or gyrus are often labeled together for extended sources. Lower panels: A source defined by a patch grown on the cortical surface. A location on cortex is seeded, and adjacent bands of voxels are labeled in a series of dilation operations. (b) A series of sample solutions from the posterior probability distribution. After 1000 iterations the MCMC algorithm found the same set of three sources in almost every sample, although additional extraneous sources also appear in some solutions. These data were simulated and thus known to contain three sources in the locations suggested by Bayesian inference. (c) Interactive visualization of spatial–temporal source probability maps coregistered with anatomical MRI for the same subject. The anatomical data set was used in the Bayesian inference procedure to constrain sources to lie in cortex. (d) Source probability maps estimated for visual evoked response data. Four views of a region found to contain activity at a 95% probability level. This example is for left visual field stimulation. For right field stimulation, the most probable source is lateralized to the calcarine fissure in the left hemisphere.

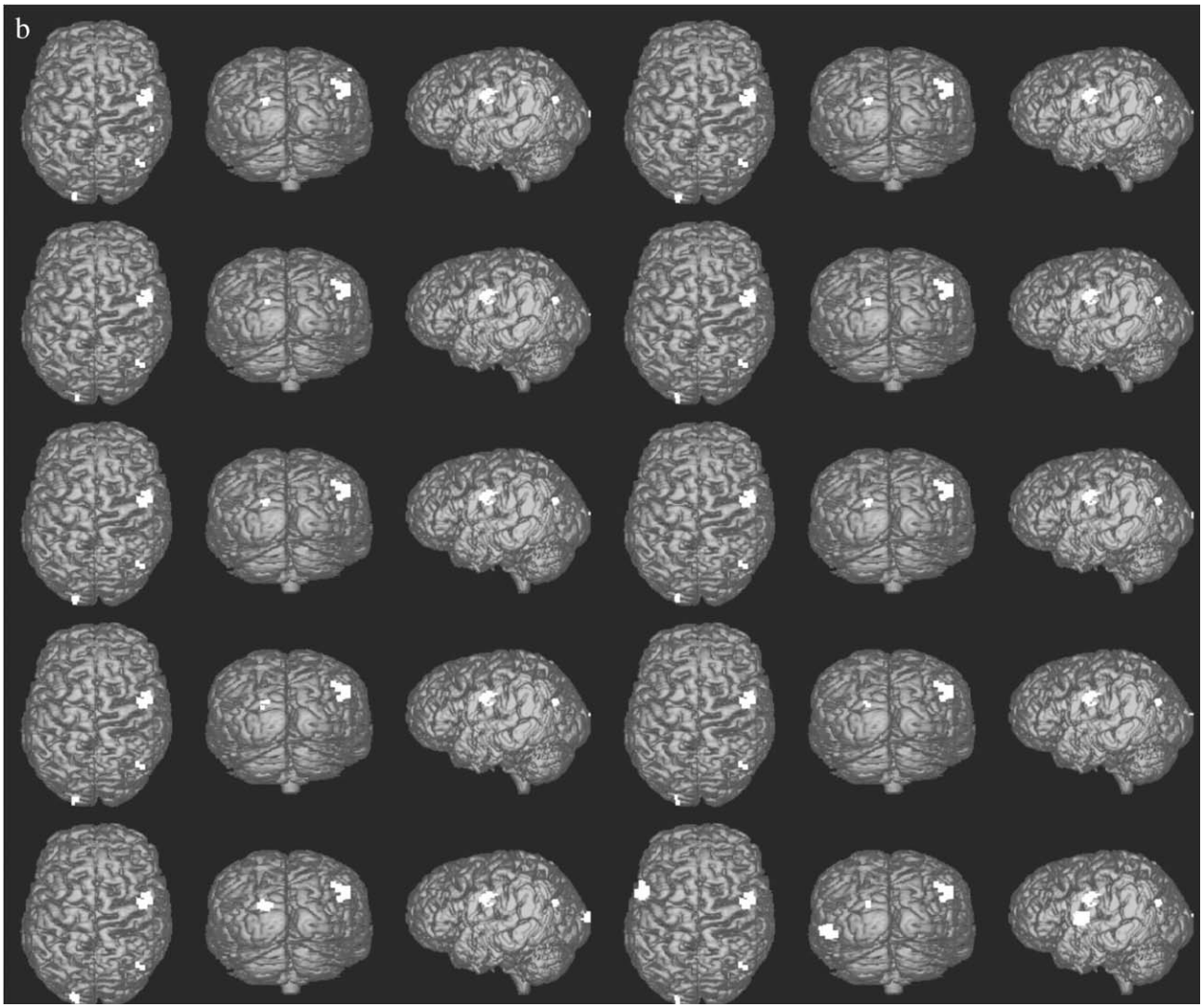


Figure 7b

generally consistent with the classic homunculus described by Penfield, based on intraoperative electrical stimulation.

2. Auditory Evoked Responses

Auditory evoked responses are typically elicited by clicks or tone bursts, which may be delivered to one or both ears. With appropriate electrode placement, it is possible to noninvasively measure electrical responses of the cochlea, including receptor potentials, and signals that reflect from neural encoding. From 1 to 12 msec poststimulus auditory brain stem responses can be measured. These responses consist of a series of mostly discrete waves labeled I–VII. Although the sources of these components are still a matter of some debate, there is general agreement that wave I is due to a compound action potential or graded dendritic potentials at the dictal (cochlear) end of the acoustic nerve. Waves III–V are generated in the brain stem. Waves V and VII are associated with higher brain stem structures—perhaps the medial geniculate body. Auditory brain stem responses are useful for hearing assessment in infants, uncooperative adults, and cases of functional deafness, as well as for evaluating brain stem function in suspected multiple sclerosis. The general temporal structure of auditory evoked responses is illustrated in Fig. 9.

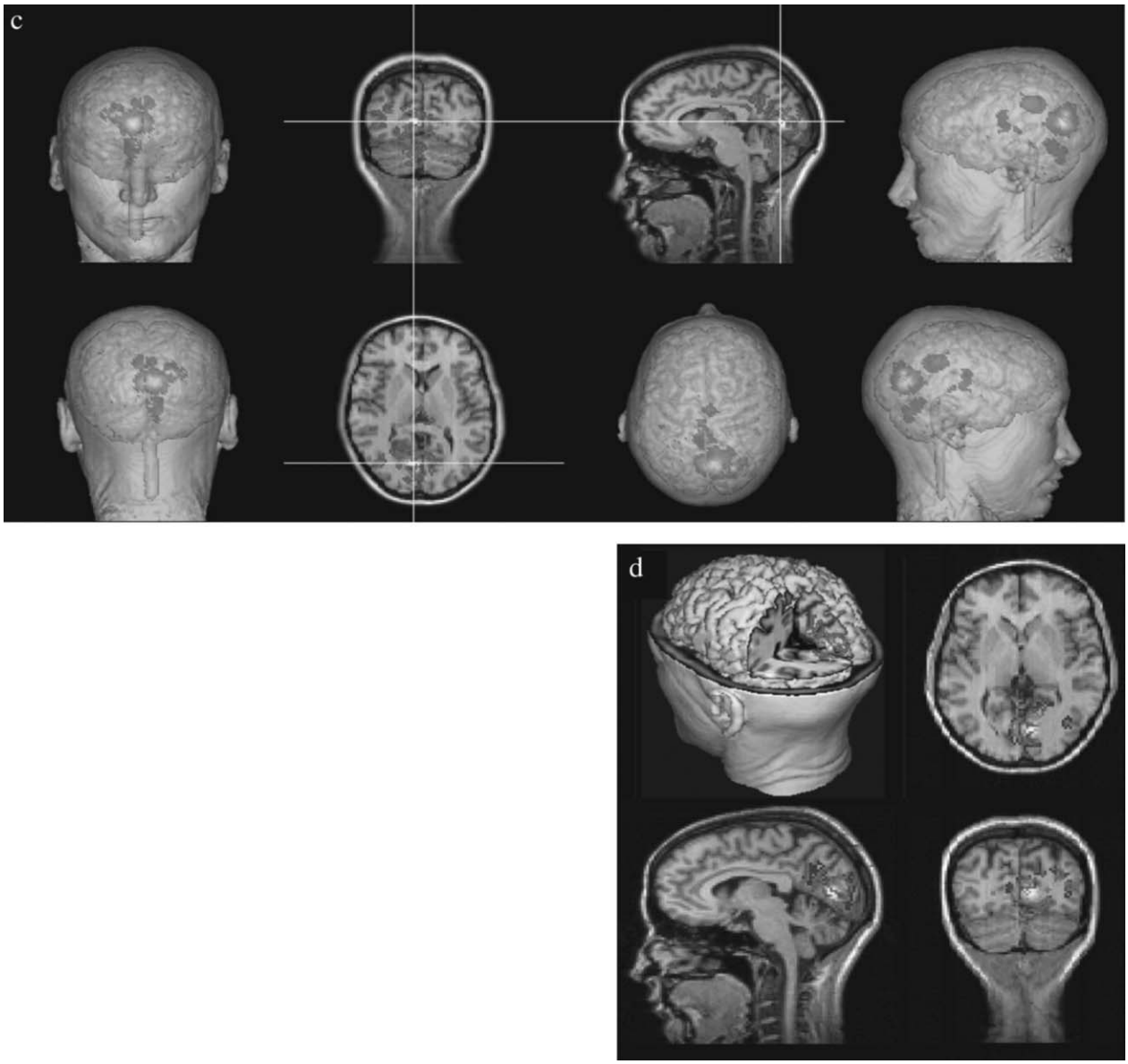


Figure 7c–d

Middle latency auditory evoked responses are typically observed from 12 to 50 msec poststimulus and are considered to represent subcortical activation.

Late auditory evoked responses (50 msec or more after the stimulus) are generally a product of neocortex. Such responses are best evoked by tone bursts and in EEG recordings show the highest amplitude over the vertex. MEG studies lave localized these responses to primary sensory areas along the Sylvian fissure and to nearby association areas. Such studies have also demonstrated clear tonotopic organization in primary cortical areas. With dipole localization techniques, very fine-grained discrimination of relative locations is possible. Auditory stimuli are often used for language studies, and in this role they may elicit an interesting array of endogenous responses that reflect the neural processing of language. However, at least one endogenous response appears purely acoustic: the mismatch negativity is observed when a repetitive auditory stimulus is briefly altered and may serve as an orienting response to cause a shift in the focus of attention.

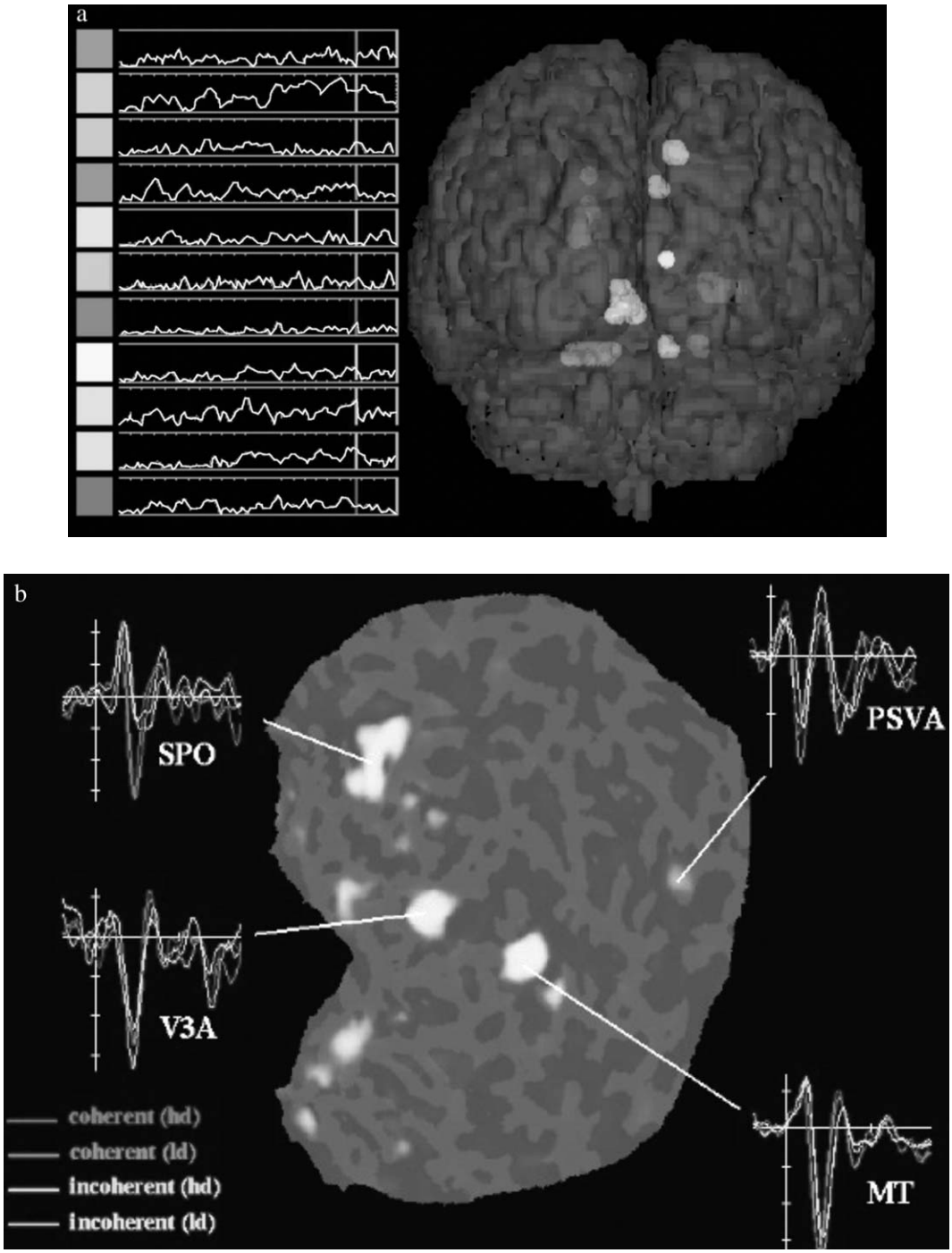


Figure 8 Integrated analysis of fMRI and MEG. (a) Time courses of fMRI equivalent sources estimated from MEG data. fMRI visual data were acquired using blocked steady-state stimulation, using the same video display from a previous MEG experiment. Currents were assumed to vary within the source according to the distribution of functional MRI activation. Currents were constrained to lie normal to cortex as indicated by anatomical MRI. Topographies were derived for each of the assumed sources and used as basis functions for a linear decomposition of the time-varying field maps. Estimated time courses for 11 areas are coded in color. (Figure courtesy of Dale *et al.*) (b) MEG time courses of fMRI sources using a weighted minimum norm procedure. Areas of activation from a visual fMRI experiment (involving visual motion) are shown on an unfolded cortex. A 0.9:0.1 weighted pseudo-inverse procedure was applied to field maps at each time point. Estimated time courses for activation in four identified visual areas are illustrated. Differences as a function of stimulus type are coded in color. (Figure courtesy of Dale *et al.*)

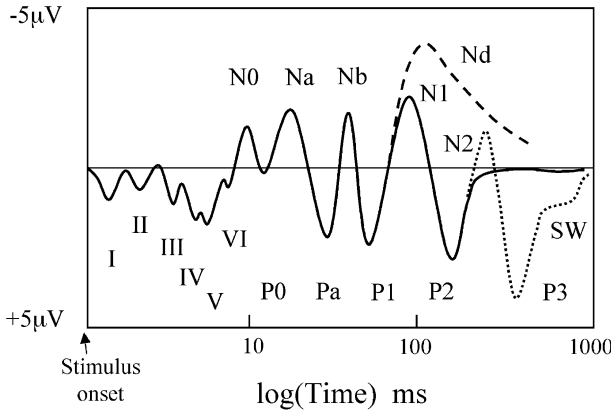


Figure 9 Component structure of the auditory event-related potential. The trace schematically represents the averaged evoked response of the auditory system to a brief stimulus such as a click or a tone. A logarithmic time scale allows visualization of the major response component peaks in a single trace. Components include auditory brain stem responses (I–VI), early positive (P) and negative (N) cortical components (Na, Nb, Pa, P1), and late cortical components (N1, P2). Other components that vary as a function of cognitive or attentive states are shown with dashed lines. (Figure courtesy of Hillyard and colleagues)

3. Visual Evoked Responses

In primates, the visual system is the largest and most distributed of the sensory modalities, consisting of over three dozen discrete areas and spanning at least one-third of the neocortical surface area. The system has been studied extensively with invasive electrophysiological techniques, as well as with MEG and EEG (and fMRI). The comparatively small signals and their dynamic complexity make this system a major challenge for sensory evoked response studies.

As in the auditory system, it is possible to measure the electrical response of the sensory organ—the eye. The electroretinogram (ERG) is typically measured using a contact lens electrode referred to a reference on the head surface. The response consists of receptor and neural components. Because the retina is a relatively accessible outpost of the brain, it has been the target of a number of studies of information processing by neural networks. In the future, optical techniques may allow noninvasive characterization of retinal network dynamics.

The visual evoked response observed with surface sensor arrays is dominated by primary cortical areas. The initial cortical activation in layer 4 of striate cortex probably occurs around 70–80 msec poststimulus, although this component is often small and difficult to detect. Other cortical responses are observed in striate and nearby areas from 90 to 120 msec poststimulus, and evoked activity often lasts through 250 msec. Source localization studies have demonstrated the anticipated retinotopic organization of primary visual cortex as well as extrastriate areas. The visual field is systematically projected onto striate cortex, mostly buried in the fissure between the hemispheres along the calcarine fissure. The central field representation is found near the posterior pole of occipital cortex and may extend onto the posterior surface. The lower quadrants of the visual field are mapped onto the upper banks of the contralateral calcarine and interhemispheric fissure; the upper field projects to the lower banks, and the horizontal meridian projects to the depths of the calacrine in a scheme summarized by the cruciform model (due to its appearance in coronal section). Noninvasive studies have confirmed the outlines of this model, although individual departures appear common. Such studies also support the idea of the cortical projection factor: the cortical area devoted to a given size patch of the visual field systematically decreases from the center to the periphery of the visual field.

Noninvasive techniques so far have largely been used to confirm in humans results suggested by invasive studies in animals. Thus, a number of specialized areas have been identified in humans analogous to those identified in electrophysiological studies in nonhuman primates, including areas specialized for processing visual motion, color, texture, and even faces. The visual system appears to be organized into two major processing chains or streams. The dorsal stream, arrayed mainly across occipital cortex and the upper surface of the parietal lobe, operates in low contrast and is involved in processing visual motion. The system is probably involved in orientation and allocation of attention and may interact with motor control. The ventral stream flows along the base of the occipital and temporal cortices, and is involved in the processing of color, texture, and other detailed attributes of visual information. This system probably interacts with language processing centers. Some investigators have dubbed these streams the *what* and *where* systems.

B. Motor Control

The control of voluntary movement can also be studied with event-related response techniques. In a typical experiment, the subject is instructed to perform a series of self-paced voluntary movements, and the

signal is averaged relative to the movement as registered by a button press or an electrical response recorded from muscle. The response appears as a slowly developing negative potential shift somatotopically arrayed along the central sulcus, starting approximately 1 sec before movement. This response is called *the readiness potential* and is taken as an index of motor preparation; the amplitude of the response is correlated with the complexity of the subsequent movement as well as the force and speed developed. The readiness potential preceding a lateralized response (such as a hand movement) is maximal over the contralateral hemisphere. Some investigators use signal subtraction techniques to remove the ipsilateral contribution to the signal. In addition to the readiness potential, other movement-related responses can be resolved, as well as somatosensory and proprioceptive feedback generated as a consequence of the movement.

C. Cognitive Event-Related Responses

The readiness potential is considered an *endogenous response* because no external (exogenous) stimulation is required to elicit it. Although motor cortex is certainly involved in responses that involve sensory processing, typical paradigms do not employ sensory cues because stimuli would elicit sensory evoked responses and trigger a set of cognitive processes that might complicate interpretation. Cognitive responses to assigned tasks involving search, discrimination, classification, or decision processes also have been studied extensively, often in the absence of any observable (behavioral) response.

1. CNV

Studies of the effects of classical conditioning led to the identification of the *contingent negative variation* (CNV), one of the first of the endogenous responses clearly linked to a cognitive process. In these experiments, an initial stimulus was delivered, followed by a delay and then a second stimulus. During the interval between the stimuli, a slowly growing negative potential was observed at the scalp. In many experiments the second stimulus was intended to cue a behavioral response, although a response is unnecessary and, as outlined above, would presumably elicit motor potentials that would complicate interpretation. Studies with variable intervals between the stimuli suggest that the CNV reflects at least two processes: an *orienting response* (o-wave) associated with the initial (warning) stimulus and an *expectancy response* (e-wave) that develops in anticipation of the imperative stimulus.

2. P300

One of the most extensively studied cognitive event-related responses is the P300 (or P3) complex, sometimes referred to as the family of late positive responses. The form of this response is relatively independent of the sensory modality used to elicit it. The generic paradigm associated with the response is an oddball discrimination task: for example, two or more stimuli are presented in a random series so that one occurs infrequently. This oddball elicits a response that begins around 300 msec poststimulus and may last 100 msec or more. Since the initial reports, the response has been resolved into at least two components that are differentiated on the basis of scalp topography and sensitivity to paradigm manipulation. The P3A is larger in amplitude over the central and frontal electrode sites. The response may be an alerting response arising in frontal cortex; in any case it is the component most strongly associated with the oddball response. The P3B appears to reflect subsequent allocation of attentional resources and encoding of stimulus memory. Some authors have argued that the response reflects information transfer via the corpus callosum with subsequent activation of hippocampal and parietal processes.

3. Selective Attention

The processing of information from the environment can be effectively suppressed or significantly enhanced depending on the state of attention. Event-related responses have been used extensively to study the time course and anatomical basis of *selective attention* in the human brain. Such experiments are typically conducted with contingencies that are manipulated within or between blocks of stimuli, so that the same physical stimulus may be the target of a discrimination task, may share some features with the target, or may be irrelevant. Data are typically analyzed by taking differences between responses to the same stimuli that differ in assigned task. The so-called negative difference or processing negativity is constructed by subtracting the response to an unattended stimulus from the response to the same stimulus while attended. In focused auditory attention, this effect may be observed as early as 20 msec poststimulus and is most apparent

around the latency of the N1; however, the time course does not reflect a simple scaling of the N1.

In visual evoked responses, tasks involving spatial selective attention, spatial cueing, or visual search produce enhanced P1 and N1 components, although controversy remains regarding whether the sources of the difference component are the same as for the underlying evoked response component. Visual selection on the basis of features such as shape, texture, or color elicits a difference signal termed the selection negativity and is observed from 150 to 250 msec poststimulus with a maximum over posterior electrodes. Taken together, these observations suggest that attention to location in the visual field is an early process, perhaps mediated by facilitation and/or suppression in prestriate relays such as the thalamus. Attention to other features or conjunction of features probably is mediated by specialized visual areas further along the processing chain.

4. ERN

In discrimination trials in which a fast choice is enforced by a reaction time task, an *error-related negativity* (ERN) is observed in trials in which the wrong response is performed. This response, predominantly observed over midline frontal areas, appears to reflect a process of rechecking that is initiated in parallel with the response. In many cases the ERN appears before the behavioral response; presumably when the response is optimized for speed, the motor system is committed before the decision–recheck cycle is complete.

5. Language-Related Responses

Event-related potential studies have been used to identify a number of responses associated with language processing. Word stimuli presented in spoken or written form elicit a response peaking around 280 msec (N280 or *lexical processing negativity*) typically characterized by a negative peak over left frontal regions of the scalp. This is considered to be a generic correlate of word processing, though in an experimental context the latency can be manipulated as a function of the frequency of the word. Another prominent response is elicited by semantic incongruity in written text or speech. The N400 is a right posterior scalp negativity the begins around 200 msec and peaks around 400 msec following the presentation of a word that violates semantic expectation. Because the response is associated with meaning and context, the N400 is considered an index of higher order language processing. A number of response components have been associated with processing of the form and structure of language. The *syntactic positive shift* (P600) is a late positive response elicited by certain types of grammatical errors (such as subject–verb agreement), even though the meaning of such a sentence may be clear. Figure 10 illustrates an example of a language-related ERP.

6. ERS–ERD

Event-related or evoked responses are generally considered to reflect transient, discrete processes associated with responses to external stimuli or internal imperatives. The time courses of many event-related response components look suggestively like the postsynaptic potentials that are recorded directly from neural tissue, perhaps reinforced by population activation and broadened by timing jitter. However, in some cases the response of neural populations appears to reflect a reorganization of ongoing activity; either an increase in synchronized oscillatory activity (*event-related synchronization*) or a decrease (*event-related desynchronization*). Although most authors consider these to be distinct phenomena, others have argued that all event-related responses might reflect transient phase locking of spontaneous activity. For example, the waveforms associated with some auditory evoked responses look like a damped oscillation. Perhaps the most interesting responses of this class are transient oscillatory population responses that may reflect phase locking within a network. Such processes are not phase-locked with respect to the stimulus and appear to reflect detection or "binding" of features such as coherent motion or a shared contour that may extend well beyond the receptive field of any individual neuron.

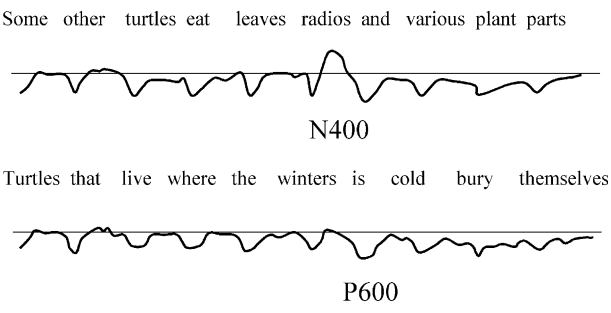


Figure 10 ERP effects elicited during language processing. Average ERP data from written sentences read one word at a time for comprehension. Upper trace: An N400 elicited by semantic violation. Lower trace: P600 elicited by grammatical violation. (Figure courtesy of Kutas and Dale.)

VIII. APPLICATIONS

A great deal of scientific effort has been devoted to the study of event-related neural responses: characterizing the spatial distribution and temporal structure of responses, probing the underlying physiology and biophysics, dissecting the consequences of cognitive and behavioral manipulations, and developing the analytical techniques and computational tools that allow more powerful inferences based on such measures. Event-related and evoked responses have recognized utility for a variety of applications in clinical and basic research.

A principal clinical application of evoked responses has been *for assessment of the patency of sensory information processing pathways*. Measurements of the ABR and early VEP have been used as objective diagnostics for the integrity and proper development of the auditory and visual pathways in infancy and childhood. The latency of the VEP has been used for the assessment of multiple sclerosis. Demyelination associated with this progressive degenerative disorder causes decreases in conduction velocity apparent as increases in the latency of cortical response components. Endogenous response components such as the P300 have been used as a generic probe of psychological and cognitive processes in disorders ranging from schizophrenia to Alzheimer's to alcoholism. In many cases the link between the diagnostic measure and the underlying pathology is tenuous, but statistical differences between normal and affected individuals can be observed.

The development of source localization techniques has led to applications for *presurgical mapping of eloquent cortex*. When neurosurgeons need to resect portions of brain tissue to remove a tumor or an epileptic focus, a principal requirement is that they spare the cortical tissue responsible for language and movement. Otherwise the pathology is often considered inoperable, even if life-threatening. The deleterious consequences for quality of life are considered too severe. MEG has been used to map the cortical regions responsible for these functions, in some cases even when the anatomical substrate has been distorted or displaced by the disease process. Such studies are typically undertaken as part of the process leading to a commitment to surgery; the conclusions are typically confirmed by conventional procedures during surgery. Because the questions are fundamentally issues of static functional architecture, fMRI is increasingly used for such purposes.

NEM techniques have clear advantages for *presurgical mapping of certain cortical pathology*. Disorders such as epilepsy are fundamentally disorders of dynamic neural function. Although the generator region is sometimes associated with an obvious lesion, sometimes the area appears normal on anatomical MRI. Typically, patterns of seizure and other epileptiform activity are first studied with EEG. MEG is used if available in an attempt to better localize the initial focus of the time-evolving response. In present practice, localization is usually confirmed by grids of electrodes placed over the surface of cortex or depth electrodes inserted in key locations. Many physicians believe that source localization based on noninvasive methods will eventually be accurate and reliable enough for surgical decisions.

Seizures can be triggered in some individuals, often children, by stroboscopic visual stimulation, providing a useful measure of experimental control. However, most epileptic seizures are not scheduled and may be relatively infrequent. Capture of an ictal event during an experimental session is, to some extent, a matter of chance. For practical reasons, many clinical researchers have studied interictal activity. Such studies suggest that interictal activity often arises at or near the locations that can trigger a seizure. Other researchers have noted a strong correlation between computed sources of slow waves with the margins of lesions that give rise to seizures. Slow waves are also associated with regions of closed head trauma. Abnormal low-frequency activity is sometimes seen in such cases even when the tissue appears normal in MRI and no significant cognitive or behavioral deficits can be detected.

NEM measures and event-related response methods have been used for other applications in *physiological and behavioral research*. EEG has been used extensively in human factors studies, for example, to assess the effects of workload, stress, and sustained effort on intellectual performance and attention. Although many of these studies have employed correlation or coherence analysis of ongoing EEG activity, others have employed sensory or cognitive probe tasks to assess neurological performance. Several investigators have explored the use of EEG to generate control signals for electromechanical systems. Prosthetic devices for amputees have been built with electronic control systems, although to date the most successful have employed control signals derived from remnant muscle. Proof of principle has been demonstrated for the use of EEG for the control of vehicles or mobility aids for patients with little or no voluntary control of muscle function.

Basic neuroscience and cognitive research have become major applications of event-related neural

response techniques. The advent of neural electromagnetic source modeling techniques has allowed localization of specialized sensory processing areas and disclosed interesting aspects of functional organization within areas. Such methods have disclosed evidence for neural plasticity in the form of use-dependent changes in the strength and extent of activation associated with stimulation. The greatest value of electromagnetic techniques is likely to come in studies of the dynamics of neural information processing within and between areas. To date, such studies have been largely qualitative and observational. However, by coupling the physical models used for electromagnetic source localization with computational neural network models, it should be possible to predict integrated responses of complex networks from the interaction dynamics of populations of model neurons. Such methods will allow us to frame and answer questions about the role of dynamic, spatiotemporal processes in the encoding and processing of information by neural systems.

See Also the Following Articles

ELECTRICAL POTENTIALS • ELECTROENCEPHALOGRAPHY (EEG) • INFORMATION PROCESSING • MAGNETIC RESONANCE IMAGING (MRI) • NEURAL NETWORKS • NEURON

Suggested Reading

Dale, A. M., and Sereno, M. I. (1993). Improved localization of cortical activity by combining MEG and EEG with MRI cortical surface reconstruction: a linear approach. *J. Cogn. Neurosci.* **5**(2), 162–176.

George, J. S., Aine, C. J., Mosher, J. C., Ranken, D. M., Schlitt, H. A., Wood, C. C., Lewine, J. D., Sanders, J. A., and Belliveau, J. W. (1995). Mapping function in the human brain with MEG, anatomical MRI, and functional MRI. *J. Clin. Neurophysiol.* **12**(5), 406–431.

Hamalainen, M. S., Hari, R., Ilmoniemi, R. J., Knuutila, J., and Lounasmaa, O. V. (1993). Magnetoencephalography—Theory, instrumentation and applications to noninvasive studies of the working human brain. *Rev. Mod. Phys.* **65**(2), 413–497.

Hillyard, S. A., Mangun, G. R., Woldorf, M. G., and Luck, S. J. (1995). Neural systems mediating selective attention. In *The Cognitive Neurosciences* (M. S. Gazzaniga, *et al*., Eds.), pp. 665–681. MIT Press, Cambridge, MA.

Kutas, M., and Dale, A. (1997). Electrical and magnetic readings of neural function. In *Cognitive Neuroscience* (M. D. Rugg, Ed.), pp. 197–242. MIT Press, Cambridge, MA.

Naatenen, R. (1990). The role of attention in auditory information processing as revealed by event-related potentials and other brain measures of cognitive function. *Behav. Brain Sci.* **13**(2), 201–232.

Niedermeyer, E., and Lopes da Silva, F. (1999). *Electroencephalography: Basic Principles, Clinical Applications and Related Fields*. Williams and Wilkins, Baltimore, MD.

Regan, D. (1989). *Human Brain Electrophysiology: Evoked Potentials and Evoked Magnetic Fields in Science and Medicine*. Elsevier, New York.

Schmidt, D. M., George, J. S., and Wood, C. C. (1999). Bayesian inference applied to the electromagnetic inverse problem. *Human Brain Mapping* **7,** 195–212.

Evolution of the Brain

HARRY J. JERISON
University of California, Los Angeles

GLOSSARY

allometry Measured comparisons of two organs or organ-systems. In brain-body allometry, usually the regression of brain size on body size. Size is scaled logarithmically, and the allometric equation is usually a power function.

encephalization The degree to which actual brain size in a species is greater or less than that expected according to an allometric analysis. It is often measured as an "encephalization quotient" (EQ), which is the residual of the regression of log brain size on log body size.

neocortex A characteristically layered portion of the cerebral cortex unique to mammals, usually six-layered.

There are about 50,000 species of vertebrates, each unique in many traits but sharing "primitive" traits with close and distant relatives. The pattern of unique traits defines each species, and the suite of shared (older, more primitive) traits, depending on the size of the suite, helps define them as members of a higher taxon: genus, family, etc. This is as true for neural traits as it is for other traits that differentiate animal species. All biologists accept the evolutionary dogma that the uniformities in traits across species are due either to common ancestry or to convergent evolution, and that the diversity of species should usually be explained by adaptations to specialized environmental niches.

This article is updated from one originally published in D. W. Zaidel, (Ed.), *Neuropsychology*. Academic Press, New York. (1994). For those with access to the Internet, an unusual collection of data is available for inspection and analysis at a site supported by the U.S. National Science Foundation, including photographs of brains and serial sections of brains in many living species of mammals. Fossil evidence is displayed at the same site and may be accessed at http://www.neurophys.wisc.edu/brain/paleoneurology.html.

I. GENERAL CONSIDERATIONS

This article emphasizes inferences from allometric analysis of brain/body relations and encephalization, the latter being a complex trait often attributable to convergent evolution. Although the diversity in organization of brains is at least as important, especially for understanding the phylogenetic trees, an adequate discussion of the evolution of diversified brain organization requires a more detailed review of comparative anatomy and physiology than is possible in a single article, and the conclusions, though important, are easily summarized for evolutionary neurobiology: Brain structure is appropriate to function, and specialized functions are appropriate to the environment (i.e., structure and function are adaptive). In short, the results are consistent with adaptation as a biological principle. Applied to the sizes (weight, volume, or surface area) of the subsystems in the brain, such as cortical projection areas and thalamic nuclei, this is the principle of proper mass.

Despite their simplicity, allometry and encephalization provide more unusual evolutionary insights. Allometry helps us understand the biological role of size; encephalization does the same for understanding neural information-processing capacity and its evolution. It will be enough to review the diversity of

organization by citing a few examples, the reports of which are extremely well documented.

The issues considered in this article are also relevant for the evolution of invertebrate nervous systems. The neuron, for example, probably appeared as a specialized cell early in metazoan evolution, more than 600 million years ago (Ma), and many of its features are identical in all instances in which it functions in a synaptic nervous system. This is evidently true for small networks of neurons as well as for isolated cells. Much of what is known about neural functions as single units and in small networks was learned from giant neurons of horseshoe crabs and from networks of cells in sea slugs and roundworms. The early appearance of the adaptation is deduced from a cladistic analysis of the time of divergence of species in which it is identifiable. It is most likely that the adaptation first appeared in a pre-Cambrian metazoan species that is the "common ancestor" of all the living species that share the adaptation. From an evolutionist's perspective, however, very complex behavior requiring integrated neural activity and involving more extensive neural circuitry that is common to vertebrates and invertebrates is as likely to be analogous ("homoplastic") rather than homologous. It may have evolved in independent evolutionary paths in the vertebrate and invertebrate groups in which it occurs.

A. Brain Structure and Function

Every vertebrate brain is hierarchically organized into forebrain, midbrain or mesencephalon, and hindbrain. The forebrain can be further divided into telencephalon and diencephalon, and hindbrain can be divided into rhombencephalon and myelencephalon. Brain tissue in all animals consists of neurons as information processing units and glia and other cells that are, in effect, supporting tissue. Neurons are often specialized with respect to neurotransmitters, shape, and size. Sizes, for example, range from the granule cells of the cerebellum (soma less than 10-μm diameter) to the giant Mauthner cells (soma about 100-μm diameter) that mediate startle responses in fish and in amphibian tadpoles. The full size of a nerve cell includes the arborization of axon and dendrites, which may account for 95% or more of the volume of a neuron and which varies enormously in pattern both within a brain and between species.

Underlying this diversity, there is surprising uniformity about principles of nerve action in the transmission of information, which makes it possible to use almost any neuron from any species as a model for neuronal action. There is, furthermore, a uniformity at the level of networks of cells in vertebrate brains, evident even in the neocortex in mammals, which encourages one to emphasize information-processing capacity for the brain as a whole as well as in its specialized component systems, such as those for color vision, binocular vision, sound localization, and olfaction.

B. Evolution

The facts of evolution are first, that it occurred and second, that it could occur because of the genotypic and phenotypic diversity both within species and between species. Charles Darwin's great contributions were to recognize the diversity and to explain it by the theory of natural selection. As currently understood, the theory is that given the variety of phenotypes in a species, some individuals will be more successful than others in surviving to produce offspring. Reproduction is the measure of success. The mean of the gene pool of the next generation shifts toward the mean of the successful phenotype. As the environment changes, the characteristics required to be successful change, and there is natural selection of individuals with those characteristics. This is a theory of the origin of species because there will eventually be enough change in the genotypic population to designate it as a new species.

There is so much support for the theory, in laboratory experiments and from field observations, that one might prefer a stronger word than "theory" to describe Darwin's integration. But there are disagreements among evolutionists, of course, which are sometimes taken incorrectly to be challenges to the credibility of the theory as a whole. The controversies are mainly about the relative importance of selection as opposed to random genetic drift, about the merits of various approaches to determining phylogenetic trees (cladistics), and about the rate of evolutionary change (gradualism versus punctuated equilibria). Despite their use and misuse in popular polemics, the controversies are on fairly technical questions and not on the fact of evolution.

II. THE EVIDENCE

A. Fossil Brains

The fossil record of the brain is from casts ("endocasts") that are molded by the cranial cavity of fossil

skulls. Natural endocasts are made by the replacement of soft tissue in the skull by sand and other debris that eventually fossilizes. Artificial endocasts can be made by cleaning the cavity and filling it with a molding compound such as latex, from which plaster casts can be made. Errors in identifying "brain" areas in endocasts of birds and mammals are likely to be about the same as in brains when superficial markings rather than histological or physiological evidence are the basis for the identification. Endocasts of some fossil animals are so brainlike in appearance (Fig. 1) that they are often referred to as fossil brains.

Figure 1 presents lateral views of the endocast and brain from the same domestic cat, *Felis catus* (Figs. 1A and 1B), and a copy of a natural endocast from a fossil sabretooth (Fig. 1C). The sabretooth is *Hoplophoneus primaevus*, which lived in the South Dakota Badlands during the Oligocene epoch of the Tertiary Period, about 30 Ma. Although no more than a piece of rock, its endocast is unmistakably a picture of its brain as it was in life; it was clearly appropriate to name its parts as brain areas in Fig. 1D following the nomenclature for cat brains.

There are several lessons to be learned from Fig. 1. First, an endocast can provide an excellent picture of the whole brain. This is evident when comparing Fig. 1A with Fig. 1B: The endocast of the domestic cat provides an excellent picture of external features of its brain and correctly estimates its size. (The estimation from the endocast is as "correct" as that from the brain, which is probably slightly shrunken by fixation.) Second, the convolutional pattern in an endocast may be fairly constant in related species. Thus, despite their separation by 30 million years of felid evolution, the endocasts of the living cat and of the sabretooth (Figs. 1A and 1C) are clearly similar. This lesson is especially important because convolutions map the way a brain is organized, at least in a general way. A third lesson, therefore, is that the felid brain of 30 Ma was probably organized in a way similar to that of living felids. Finally, as counterpoint to the lesson of uniformity of organization, there is a lesson of diversity: Two gyri, the coronal and sigmoid gyral complexes, are differentiated in living domestic cats but are undifferentiated in the sabretooth. Increases such as these in the apparent complexity of the brain in felid evolution may be related to increases in information processing in the expanded areas in later species compared to earlier species.

The quality of an endocast as a model of the brain differs in different taxa. Endocasts from fish, amphibians, and reptiles (except in very small specimens) are poor models, useful mainly to estimate total brain size after suitable corrections, because the brain fills only a

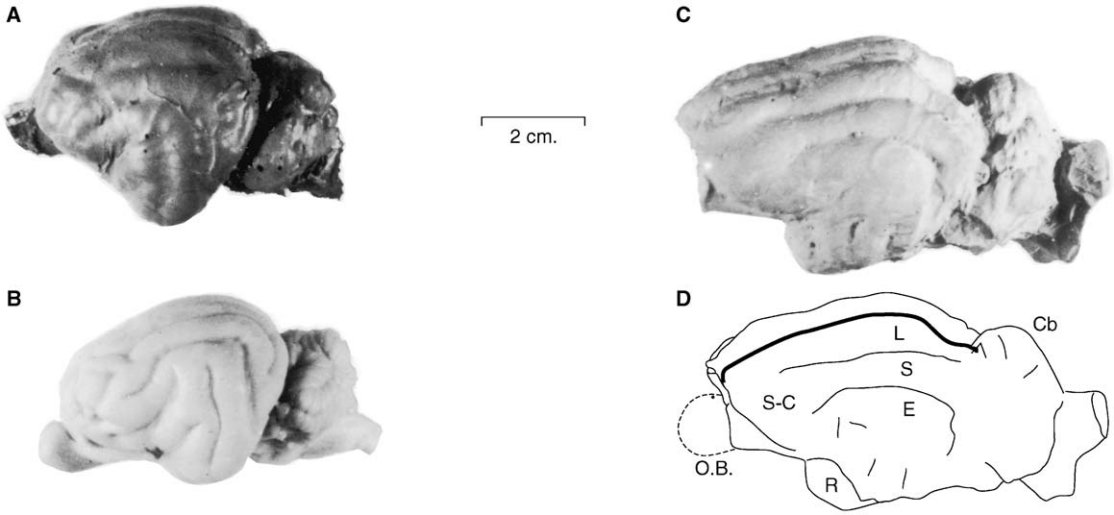


Figure 1 Brain and endocasts of felids. (A) Endocast of domestic cat (volume = 30 ml). (B) Brain of same cat (weight = 29.1 g). (C) Endocast of Oligocene sabretooth, *Hoplophoneus primaevus*, of 30 million years ago (volume = 50 ml; Specimen No. USNM 22538 at the United States National Museum, Smithsonian Institution). (D) Tracing of the endocast of Hoplophoneus with labels for several structures: Cb, cerebellum; E, ectosylvian gyrus; L, lateral gyrus; R, rhinal fissure; S, suprasylvian gyrus; S-C, sigmoid and coronal gyral complex (undifferentiated). Both endocasts are rotated about the anterior–posterior axis, exposing the longitudinal fissure (heavily inked in D). Olfactory bulbs (OB) are sketched in on the basis of more complete endocasts (e.g., AMNH 460 at the American Museum of Natural History). The unlabeled gyrus above the lateral gyrus is the lateral gyrus of the right hemisphere.

fraction of the cranial cavity. In mammals and birds, on the other hand, the brain actually helps shape the cranial cavity during development, and endocasts are usually excellent pictures of the outside of the brain. Olfactory bulbs, forebrain, and hindbrain are readily identifiable, as are most of the cortical gyri and sulci that are seen when a brain is first removed from the skull. Certain large-brained living mammals—namely, cetaceans, elephants, great apes, and humans—are exceptions to this rule, with little or no impression of their convolutions on their endocasts; even the boundary between cerebrum and cerebellum may be unclear.

Fossils provide other information for understanding brain evolution and for extrapolations to behavior. Body size, for example, estimated from postcranial skeletal data, is used to analyze encephalization of fossil vertebrates. Details of structure, such as the shape of teeth, forelimbs, and hindlimbs, can enable one to analyze feeding habits, gait, and other behavior. There is even fossil evidence on social behavior, for example, in dinosaurs, which has been reconstructed from the aggregation of fossils, their eggs, and their foot and tail prints. Perhaps most important for an analysis of brain evolution, there is a fossil record of sensory structures that is useful in reconstructing the information available to fossil animals. Olfactory bulbs, of course, are visible on the endocast, and their size is related to the evolution of the sense of smell. There are fossil middle ear bones and cochlea important for the analysis of the evolution of hearing; the orientation of the orbits of the eye provides evidence on the evolution of binocular vision, and the placement of the hyoid bones on fossil humans has been the basis for speculations on the evolution of the voice box and of articulated speech.

B. The Living Brain

Most of the evolutionary evidence on living brains is from anatomical and physiological studies of brain tracts and regions compared for unique and common features across species. There is growing interest in molecular evidence (e.g., on neurotransmitters), and one can anticipate increasing emphasis on that kind of information.

Braitenberg and Schüz have published a straightforward anatomical monograph, noteworthy for the quantitative analyses of the cerebral cortex of the mouse. Though not specifically concerned with evolutionary issues, they provided exemplars of data necessary for an evolutionary analysis. The most striking facts are on the amount of information processing machinery in the mouse, with some suggestions on the human brain. There are about 40,000,000 neurons in the 0.5-g brain of a mouse; more astonishing, there are about 80,000,000,000 synapses in its neocortex. Taking into account the packing density of neurons and synapses, they reached the conclusion that a particular volume of cortex processes the same amount of information, whether it is in a mouse or a man. This is an outstanding uniformity for evolutionary analysis since it validates the use of brain size as a "statistic" to estimate the total information processing capacity of a brain.

Uniformity is balanced by diversity. All species differ in the details of the organization of the component systems of their brains. The raccoons and their relatives (family Procyonidae) provide an outstanding example reported by W.I. Welker. The fish-handling raccoon has a much enlarged forepaw projection area in its somatosensory neocortex, with separate representation in the brain for each of the pads on the forepaw. The coati mundi, kinkajou, and most other procyonids obtain this kind of information by nosing about, exploring their environment by touching things with the sensory skin around the nostrils. Their neocortical projections from that region are comparably expanded and their forepaw projection areas are much less extensive and not as differentiated as in the raccoon. The conclusion is inescapable that reorganization of the brain, like the differentiation of the behavior that it controls, occurred as part of the speciation of procyonids as they evolved, and that raccoons branched away from the main line by their specialized adaptations in their use of forepaws. Data like these can be used for formal cladistic analyses. The mammalian phylogeny constructed from brain features is essentially the same as that based on a more complete suite of traits.

I depend on the comparative quantitative data laboriously accumulated by Stephan and his colleagues on the volumes of many brain structures in "primitive" species represented by insectivores and their relatives and in "advanced" species (primates) for many of my analyses of allometry and encephalization. Theirs are the most complete data of this sort currently available. In their sample of 76 species, there were 26 from the order Insectivora (shrews, moles, and hedgehogs), 2 Macroscelididae (elephant shrews), 3 Scandentia (tree shrews), and 45 primates, of which 18 are from the suborder Prosimii (lemur-like species) and 27 from the suborder Anthropoidea (simian species,

including humans). The brain structures are listed in Table I. These data are especially useful because of the large number of species that are in the sample and the good sample of brain structures on which measurements were taken.

III. QUALITATIVE ANALYSIS

There have been a number of outstanding evolutionary analyses of classic issues in neurobiology, and I describe three of these very briefly to suggest the topics and flavor: those by Ebbesson, by Karten, and by Killackey (complete citations for this discussion are in the chapter referred to in footnote on first page). Ebbesson provided a superb case history on the difficulties in interpreting and reasoning from available anatomical data on the brain. Ebbesson argued that connections are created by a process of "parcellation (segregation–isolation)" that occurs ontogenetically as well as phylogenetically, with originally diffuse and extravagantly proliferating neurons and connections eventually becoming reduced and segregated from one another during the course of development. Northcutt's commentary was noteworthy, pointing out not only the problems with the data used to support the position but also the semantic and philosophical difficulties: the need for rigorous specification of homologies and homoplasies in using comparative data for cladistic analysis.

Karten analyzed the origin of neocortex as a "uniquely" mammalian brain system, pointing out that neocortex is not functionally unique since its connections are comparable to those of the "neostriatum" in birds (and even in reptiles). The enlarged neostriatum in birds is homologous with mammalian neocortex. His microscopic analysis of these forebrain systems points to their comparability with respect to information processing, despite the very different ways that the brain is organized in these classes of vertebrates. This is in agreement with data that show birds and mammals to be comparable in "grade" of encephalization.

Killackey argued from ontogenetic data on the sequence of appearance of various neocortical regions, making the important point for evolutionists that the detailed organization of the neocortex is established to a significant extent by experience, and the evolution of its organization is therefore likely to be difficult to specify with standard genetic models.

Although informed by modern evolutionary theory, with the exception of Northcutt's commentary the discussion that I just reviewed was traditional in its evolutionary approach. The concern was to develop insight into the origin of neural systems and to the degree of specialization in different species. The analysis proceeded from data on morphology and development, and from educated intuition rather than from the rigorous application of cladistic methodology. All would agree that the nervous system parallels other systems in the body in reflecting adaptations to various environmental niches. Killackey and Ebbesson emphasized the lability of the fate of neural structures, and there is a consensus about the significant extent to which use determines fate for the circuitry of the brain—that brains can be normal only if they develop in a normal environment.

Comparative brain data have also been used for more formal analyses of relationships, either with the methods of modern cladistics or with other multivariate methods. The results of these analyses can be summarized in a few sentences. Performing a factor analysis on brain traits in fish helps to clarify issues on the classification of particular groups of fish. The most helpful traits were the size of the olfactory apparatus. The contribution, however, is primarily to taxonomic issues rather than to neurobiology. The cladistic analyses, using only data on brain traits in constructing a species-by-traits matrix, produce essentially the same phylogenetic tree as when a full suite of traits is used. The diversity of species as determined rigorously by a full suite of traits predicts the measured diversity as determined from brain traits. The similarity between *Hoplophoneus* and *F. catus* in Fig. 1 is the expected finding in any comparative analysis of mammalian brains and confirms the taxonomic conclusion that these are relatively closely related species despite their separation by 30 million years of evolution. The brain can serve as well as other organs of the body for evidence on phylogenetic relationships.

IV. QUANTITATIVE ANALYSIS: LIVING VERTEBRATES

Darwin's theory of natural selection emphasized evidence of selection as practiced by animal and plant breeders. The term natural selection implied that nature, like breeders, worked to select the "most fit" individuals relative to some criterion. In a breeder's case, the criterion might be plumpness, large size, docility, and so forth. Nature had other criteria, and animals well endowed on nature's criteria would be

more likely to survive to produce more offspring than would less well-endowed animals. The unnatural docility of domestic animals suggested some deficit in their brains.

Darwin could explore the relationship by comparing brain size in wild and domesticated populations of animals known to be related to one another. In what was probably his only contribution to neurobiology, Darwin was the first to observe that the brain in domesticated rabbits was smaller than that in their wild cousins. This is evidently a general principle on the effect of domestication on brain size. It may even be true for human brain evolution if we think of ourselves as domesticated and our ancestors as savage or feral, although the available sample size is too small for a clear test. The earliest *Homo sapiens* were the neandertals, and they were slightly larger brained, on average, compared to their living conspecifics.

A. Uniformities in Structure in Living Brains: Allometry

Darwin's publications inspired several generations of comparative neuroanatomists to provide detailed pictures of the diversity of brains. The effort included studies of brain size and of brain-body relations, some of which stand up under appropriate analysis today. In this tradition, Brodmann tabulated data on brain surface area in mammals, and his results are included as 33 of the 50 data points in Fig. 2A. The validity of his work is attested to by its consistency with data from more recent studies that used different methods of measurement. A single regression line fits the entire data set remarkably well. Both Fig. 2A and Fig. 2B are examples of uniformities of organization of the brain in mammals.

Analyses such as those in Fig. 2 are "allometric" in that they display the relationship between different morphological features, such as height and weight, as they can be measured in any animal. The relationships displayed in Fig. 2 transcend species and are so strong that they appear to reflect a fundamental feature of the body plan (*Bauplan*) of mammals. Evolutionists call shared ancestral features plesiomorphies, and although the term is not usually applied to functional relationships such as those shown in Fig. 2, the idea fits. The relationships should be thought of as representing a primitive feature in mammalian evolution.

Because of its unusually diverse sample of species, Fig. 2A provides important justification for using total brain size as a statistic that estimates the total neural information processing capacity of a brain, between species, in the mammals as a class. To understand this further, consider some candidates for the role of processing unit. A frequent candidate is the cortical column, and its cross sectional area appears to be relatively uniform across species. The neuron is another candidate, and the number of neurons under a given surface area of neocortex is more or less constant across species. Finally, the synapse is a candidate. Braitenberg and Schüz observed that the number of synapses per unit volume of cortical tissue is

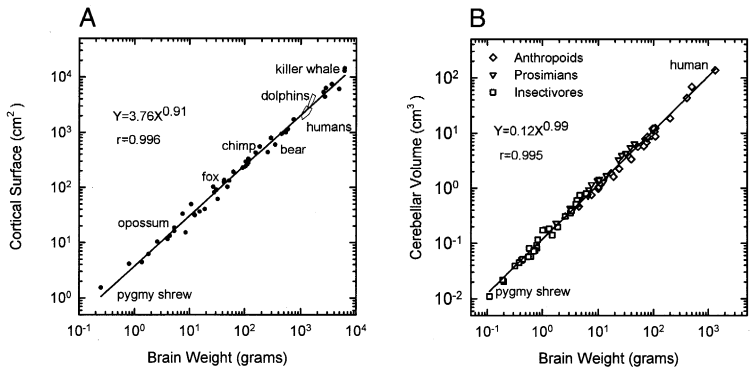


Figure 2 (A) Cortical surface area as a function of brain size in 50 species of mammals, including orders Monotremata, Marsupialia, Artiodactyla, Carnivora (including pinnipeds), Cetacea, Perissodactyla, Primates, and Xenarthra. Minimum convex polygons enclose individual human ($n = 20$) and dolphin (*Tursiops truncatus*: $n = 13$) data and indicate within-species variability. Some species are named to suggest the diversity of the sample. (B) Cerebellar volume as a function of brain size in 76 species of mammals (insectivores: 26 Insectivora, 2 Macroscelididae, and 3 Scandentia; primates: 18 Prosimii, 27 Anthropoidea (redrawn with permission from Jerison, 1991).

constant across species. There are qualifications to these generalizations, but they are reasonable first approximations. They reinforce the conclusion about the use of brain size as a statistic.

Figure 2B is another allometric analysis. It demonstrates the uniformity of cerebellar size in mammals—that, independent of species, if you know the size of the brain you can make a very good estimation of the size of the cerebellum. Although this analysis does not have the obvious theoretical significance of Fig. 2A, it demonstrates the fundamental orderliness of the construction of the brain. It validates the use of brain size as a statistic to estimate the size of other brain structures. To the extent that these other structures can be assigned special functional significance, one may be able to use quantitative data on brain size to assess the evolution of brain functions. This has been done by M.A. Hofman and by me for the analysis of the control of social behavior and of other neocortical functions. In general, anatomists emphasize the differences among the species that they examine, but I have been even more impressed by the uniformities. The example in Fig. 2B of the relationship between cerebellum and brain size is just one of these. Table I summarizes a multivariate analysis of an entire data set: 12 morphological measures in 76 species.

The most important fact in Table I is that just two factors were enough to account for all but 1.5% of the variance, and that 86% of the variance is explained by a single "size" factor. The size factor is a "general" factor in the sense that it is strongly represented in all the brain structures with the exception of the olfactory bulbs, and it is also represented in body weight. The loading of cerebellum on this factor (0.983, accounting for 97% of the variance in cerebellar volume) in conjunction with the even higher loading of total brain weight reflect the high correlation shown in Fig. 2B. All the measures with higher loadings would have produced bivariate graphs such as Fig. 2B. The mammalian brain hangs together well, and when one part is enlarged the rest of the brain tends to be correspondingly enlarged.

Perhaps the most surprising extension of this conclusion is the tentative discovery that even the size of prefrontal neocortex, the "executive organ" of the mammalian brain, appears to be determined by the size of the whole brain in mammals. The conclusion is tentative because it is based on evidence from only four primate species (marmoset, rhesus, orang, and human) and the laboratory rat, but there was an almost perfect correlation between neocortex volume and brain volume in these five species (logarithmic measures; $r = 0.999$). Developmental and morphological evidence obtained by B.L. Finlay and her associates, essentially supporting the conclusions of the factor analysis, has resulted in much more extensive additional evidence for the fundamental uniformity of structure in the mammalian brain, that is, for the extent to which it "hangs together."

Table I

Factor Loadings and Percentage Variance Explained by Two Principal Components (Factors) in Brain and Body in 76 Species of Mammals[a]

	Factor 1 (general brain size)	Factor 2 (olfactory bulbs)
Neocortex	0.991	0.059
Total brain weight	0.989	0.137
Diencephalon	0.987	0.144
Basal ganglia	0.987	0.133
Cerebellum	0.983	0.168
Mesencephalon	0.972	0.196
Medulla	0.966	0.224
Hippocampus	0.962	0.239
Schizocortex	0.954	0.274
Body weight	0.939	0.285
Piriform lobe	0.899	0.399
Olfactory bulbs	0.157	0.985
Percentage total variance	85.855	12.668

[a]Varimax rotation. Reproduced with permission from Jerison (1991).

The second factor in Table I, accounting for 12.7% of the variance, is an olfactory bulb factor. It is represented primarily by the olfactory bulbs, with a modest representation in the parts of the brain that are classic "rhinencephalon" (piriform lobes, schizocortex, and hippocampus) and in body weight. Factor analyses are notoriously susceptible to artifacts of sampling, and I believe that the high fraction of variance accounted for by the olfactory bulb factor is such an artifact. Simian primates have much reduced olfactory bulbs, whereas insectivores are normal mammals in this regard. The distribution of olfactory bulb size in the sample on which the factor analysis was performed is, therefore, seriously bimodal. This inflates the measured variance of the size of the olfactory bulbs and enlarges its fraction of the total variance compared to what would be expected in a more representative sample of mammals. In any case, the important feature of the multivariate analysis is the

almost uniformly heavy loadings of the other measures on the general size factor.

B. Diversity in Living Brains: Cladistics

Having emphasized uniformities so forcefully, I must warn against underestimating the importance of diversity. All brains are different, and there are major differences both within and between species. Differences in brains within species are often difficult to measure with conventional anatomical and physiological methods, but since the brain is the control system for behavior, behavioral differences are evidence of differences among brains. Differences among species, of course, are much more dramatic.

The qualitative differences among the procyonids described by Welker could be presented as quantitative differences as well by measuring the amount of tissue in, for example, forepaw and rhinarial projections to the neocortex of procyonids. Differences among orders of mammals or classes of vertebrates are even more striking, but they too have not been quantified, perhaps because they are so obvious. And even when differences are great, they may be surprisingly difficult to describe quantitatively. One recognizes in an instant that the human brain is an unusual primate brain, for example, but the analysis of the relative size of its major parts usually shows it to be a perfectly normal primate (or mammalian) brain. In Fig. 2 there is nothing other than gross size to distinguish the human data, which fell on or near the regression lines determined for all the mammals in each sample. Also, as mentioned earlier, even the size of the prefrontal neocortex, often assumed to be uniquely important in human performance, is exactly as large as expected for a mammalian brain the size of a human brain. The uniqueness of human behavior is related to the size of the entire neural control system, and the correct conclusion about prefrontal executive function is that its size is appropriate to the very large brain systems that it controls.

A rigorous, though not really quantitative, analysis of the diversity of organization of the brain has been in the application of cladistic methodology. As indicated earlier, the results of this kind of analysis with brain features are essentially the same as those when other morphological features are used in the traits-by-species matrix that provides primary data for the analysis.

A cladistic methodology was applied by P.S. Ulinski, who took as his goal the reconstruction of probable features of the internal anatomy of brains at nodal evolutionary points in the evolution of reptiles, birds, and mammals. He first used the results of cladistic classification to determine nodes at which branching occurred when an ancestral species split into two daughter species. Second, he examined the brains of living representatives of the daughter species (or higher taxa). Finally, he constructed a hypothetical ancestral brain as a kind of lowest common denominator of the brains of the daughter species. The approach can be applied only to nodes in which surviving species from both branches exist. For example, taking living turtles and crocodiles to represent surviving species from the node of the early reptilian branching that led to these species, the ancestral brain can be constructed as having only those features that turtles and crocodiles share. With this procedure Ulinski could suggest various details about the ancestral brain of birds, crocodiles, lizards, and turtles. (From a cladistic perspective birds may be thought of as specialized reptiles derived from dinosaurs.) The nodal point in the history of the mammals, unfortunately (for this approach), is late in a "reptilian" synapsid lineage, represented today only by mammals. The mammal–reptile transition brain could be reconstructed only if synapsids at a reptilian grade of brain evolution had survived; but none have. The reconstruction of the brain at the reptile–mammal transition is, therefore, impossible using his procedure.

It would be possible to temper this conclusion, which is based on qualitative internal features of the brain, by analyzing the superficial anatomy using data on fossil endocasts, although such data are sparse for synapsids. Where they exist, in therapsids, they suggest a size pattern comparable to that in living reptiles. The transition from mammal-like reptiles to mammals is documented in the endocasts, and at present it seems to be reflected primarily as both enlargement of the brain (encephalization) and a major reduction in body size.

In his analysis of mammalian brain evolution, particularly the diversity of organization of somatosensory neocortex, J.I. Johnson presents an impressive catalog of detail on differences but concludes that "a great many features are constant across all mammals, from platypus to monkey, rat, cat, and sheep." The major variations "include [amount of] multiplication of representations of certain body parts" and details of the representations. He also notes few general trends of organization but comments that the appearance of "association cortex" intercalated between somatosensory and visual neocortex is haphazard across species,

and that its appearance seems "to have something to do with the use of limbs as information-gathering and manipulating organs." Johnson's conclusions are consistent with the notion that the pattern of organization of the brain in a species that differentiates it from other species follows no general principles in the mammals but is part of the specialization of each species. In cladistic analyses, Johnson and his associates found that the phylogenetic tree in mammals deduced from 15 brain traits was essentially the same as that deduced from other traits.

The quantification of diversity depends on the measurement of size. The evolution of brain size in mammals has led to the diversification that was already evident in the data of Fig. 2, with species having brains as small as 0.1 g (pygmy shrew) and as large as 8 kg (killer whale). These all evolved from a single species of mammal (according to the monophyly accepted today) that lived more than 200 Ma. I outline the history later, but we must first understand how the diversity of size is analyzed.

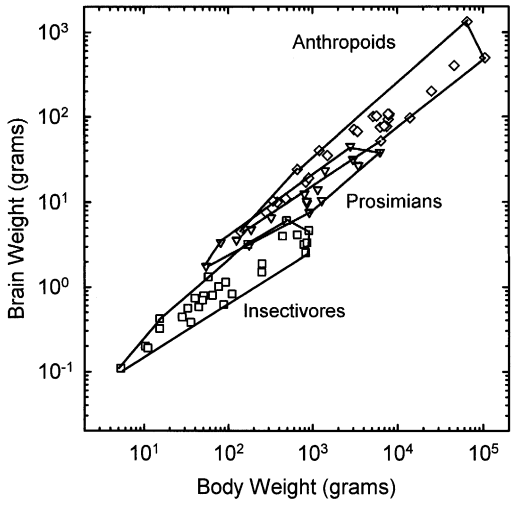


Figure 3 Convex polygons to differentiate insectivore, prosimian, and anthropoid data on brain weight and body weight (from a chapter for Zaidel; see footnote on first page).

C. Allometry and Encephalization

Body size accounts for 80–90% of the variance in brain size between species, a relationship described by an allometric equation: the regression of the logarithms of brain size on body size. The distance of a species from the regression line is a measure of its encephalization. Because the scales are logarithmic, this distance, or residual, is an encephalization quotient—the ratio of actual brain size to expected brain size. Encephalization is a characteristic of a species; it is usually meaningless to discuss differences within a species in encephalization.

Allometry and encephalization do not have to be defined by regression equations and residuals, but most recent work on brain evolution involving brain body allometry uses this approach, which might be called "parametric" since it involves the estimation of the parameters of a normal probability distribution. Instead of the regression, the data can be described with minimum convex polygons enclosing the data points of the groups to be compared, but there are currently no quantitative methods to analyze the polygons.

Minimum convex polygons described the location of human and dolphin data in Fig. 2A, and the brain/body data for the same insectivores, prosimians, and anthropoids as in Fig. 2B are graphed in Fig. 3, with polygons drawn around each group to compare them with respect to encephalization. (Recall that Fig. 2B related the size of the cerebellum to the size of the whole brain; in Fig. 3 the relationship is of the whole brain to the size of the body.) It is not difficult to distinguish relative brain size among the groups since there is little overlap. All the polygons are oriented upward. There is slight overlap between the insectivores and prosimians and a bit more overlap between prosimians and anthropoids. From Fig. 3, one would describe the order of encephalization of these groups as follows: insectivores are least encephalized; prosimians are intermediate, and anthropoids are most encephalized. These data are also described by regression equations in Fig. 4.

The work by Stephan's group is especially relevant for evolutionary analysis because of the species they used. They worked with insectivores to represent a primitive grade of brain evolution and to provide an evolutionary perspective on the human brain. The issues are more complex, of course, but insectivores are reasonable models for the base group from which most placental species evolved. They resemble the earliest mammals both skeletally and in their endocasts. Although primates are currently a highly encephalized order of mammals, they are also a very ancient order, probably derived during the Late Cretaceous period from a species comparable to living insectivores or tree shrews. Comparisons between insectivores and primates are thus very appropriate for our topic.

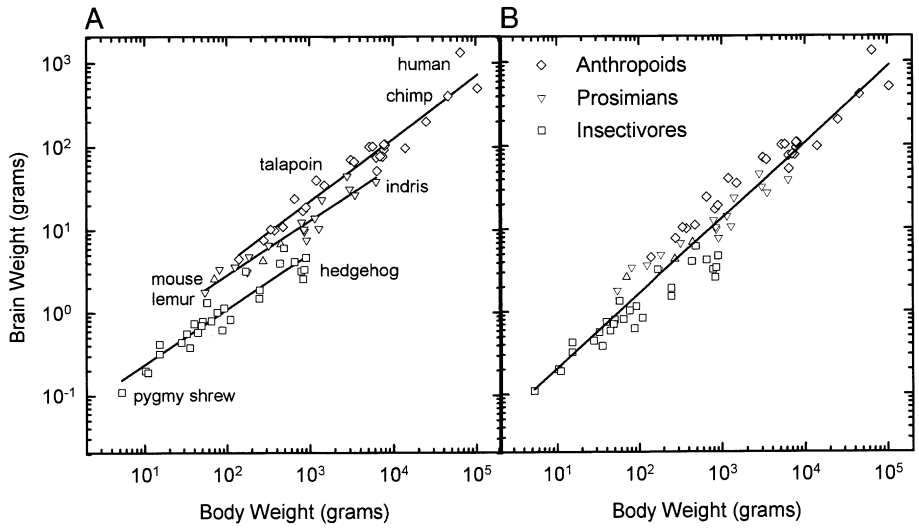


Figure 4 Regression analysis of the data shown in Fig. 3. Some species are named to indicate diversity of sample. (A) Separate regressions and correlation coefficients for the three groups: insectivores: $Y = 0.05\ X^{0.67}$, $r = 0.946$; prosimians: $Y = 0.14\ X^{0.66}$, $r = 0.960$; anthropoids: $Y = 0.13\ X^{0.75}$, $r = 0.972$. (B) Lumping the data for an overall regression for all 76 species: $Y = 0.05\ X^{0.91}$, $r = 0.966$. (redrawn with permission from Jerison, 1991).

D. A Bit of Theory

Issues in parametric quantification of encephalization as they apply to insectivores and primates are suggested in Fig. 4. The two graphs present the same data, fitted by straight lines in different ways. Fig. 4A shows the regression of log brain size on log body size computed separately for the three groups; Fig. 4B is a single regression for all 76 species. The three regression lines in Fig. 4A provide the same information as the polygons in Fig. 3. But if one is interested in curve fitting all the regression lines fit remarkably well ($r > 0.94$) despite their different slopes. These slopes on log–log axes are the exponents of the equations written as power functions, and the value of a "true exponent" has been the subject of considerable debate during the past decade. This is where a little theory may help.

The emerging consensus is that an exponent of 3/4 is the correct value. I have quarreled with this view, arguing in favor of a 2/3 exponent, which has theoretical significance for dimensional analysis of the brain's work in mapping information from the external environment. It is true that empirical analyses of large enough samples of species, or of properly sampled groups of species, lead to the 3/4 exponent when the fit is statistical, but I believe that the theoretical value of 2/3 is nevertheless correct. The point is that the 2/3 value is required by the dimensional problem in order to convert data about a surface into data about a volume (a "mapping"). It reflects the fact that our information about the external world is spread across a surface consisting of sensory cells distributed throughout the body (skin, retina, organ of corti, olfactory epithelium, etc.) and that information is pumped up to neurons distributed through a kind of conceptual surface in a brain. I have assumed a fixed cortical thickness as representing that brain surface, and that the measure of brain volume in brainbody allometry is converted into a measure of that surface area. However, since the conversion is by a physical system that takes up space, one has to take into account the thickness of the map formed by the cortical "surface." To explain the difference between a 2/3 exponent required for the mapping and the 3/4 exponent found empirically, I have argued that this thickness as estimated by the thickness of neocortex is known to be greater in larger brains, varying approximately with the 1/9 power of body size. The value 3/4 is approximately the sum of $2/3 + 1/9$. The theoretical value of 2/3, which is meaningful for the brain's mapping function, thus leads to an expected empirical value of 3/4.

E. Encephalization

The fact of encephalization is evident in the vertical displacement of the lines that are fitted to the three groups (Fig. 4A) or of the minimum convex polygons

(Fig. 3). Since the polygons do not require the dubious assumptions of statistical curve fitting, they may be preferred for describing encephalization. They are certainly preferred if they are adequate for answering questions about whether groups are equal or differ in encephalization.

The degree of encephalization in living vertebrates is summarized in Fig. 5. The polygons enclose all the available data on the indicated classes. The data were assembled from a variety of sources. The main inference from Fig. 5 is that one can characterize birds and mammals jointly as "higher" vertebrates, and reptiles, bony fish, and amphibians can be characterized as "lower" vertebrates. The polygons do an adequate job, although the addition of data on cartilaginous fish (sharks, rays, and skates), jawless fish (agnathans), and electric fish (bony fish: Mormyriformes) makes it difficult to distinguish the groups by inspection. The additions are of relatively few species. As mentioned previously, the present consensus recognizes about 50,000 vertebrate species. Of these, to the nearest thousand, about 25,000 are bony fish, 6000 are reptiles, 4000 are amphibians, 10,000 are birds, and 5000 are mammals. There are about 800 cartilaginous fish species and 70 agnathan species.

From the evidence presented earlier in Fig. 2A, it is appropriate to assume that the amount of encephalization measures the information processing capacity of a brain, adjusted for body size. It is therefore appropriate to consider the ecological requirements met by increments in processing capacity in different groups of species. Not much debate is required to see mammals and birds as higher vertebrates in this regard, given the normal complexity and plasticity of behavior observed in these groups. No reasonable speculations have been offered for the position of the cartilaginous fish as overlapping the higher and lower groups, but the place of electric fish reflects an unusually enlarged cerebellum in these species, related to processing information from their electric organs. It is unclear why that processing should require as great an investment in neural machinery. The position of jawless fish has been placed below the bony fish polygon, leading to speculations that there may have been a reduction in brain size related to the parasitic habits common in this group, particularly among lampreys. However, as evident from Fig. 5, agnathans, though relatively small-brained, fall more or less within the fish polygon, making such speculations unnecessary.

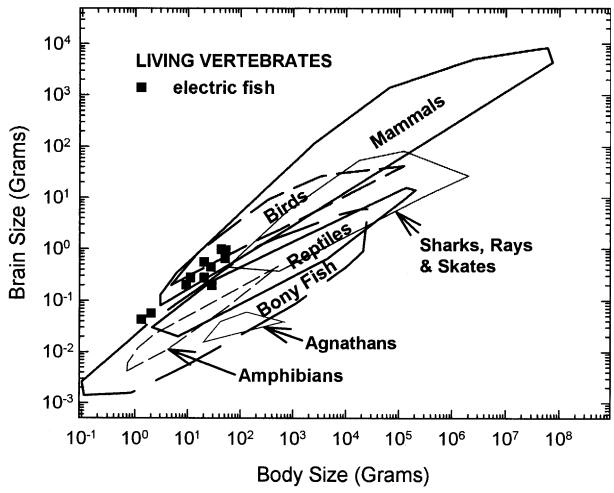


Figure 5 Brainbody relations in 2019 living vertebrate species enclosed in minimum convex polygons. The samples are 647 mammals, 180 birds, 1027 bony fish, 41 amphibians, 59 reptiles, 59 chondrichthyans (sharks, rays, and skates), and 6 agnathan fish (redrawn with permission from Jerison, Roth and Wulliman, 2001).

The approach signaled by Fig. 5 enables us to evaluate fossil endocasts with respect to encephalization, providing a direct evolutionary window to the patterns of change that led to the current diversity in brain size. I present such a nonparametric analysis as well as a parametric (regression) analysis of neocorticalization in the next section. In the analysis by convex polygons, I will be concerned with the evolution of birds and mammals from the reptiles and the utility of the method for some conclusions about dinosaur brains.

V. QUANTITATIVE ANALYSIS: FOSSILS

A. Vertebrate History

Vertebrates first appeared during the past 500 million years of the earth's 4.5 billion (4.5×10^9) year existence, and Table II provides a synopsis of their history. Here are some points to remember.

First, the world was very different in the distant past compared to the present. During the Paleozoic Era, there were times when there was only a single global continent (Pangea), but landmasses joined and separated with the passage of time. The global map was significantly different during the Mesozoic, with major masses (Gondwanaland and Laurasia) during the Paleozoic and Mesozoic Eras. There were warmer and more stable climates during the Mesozoic, and the continents were drifting toward their present loca-

Table II
Synopsis of Vertebrate Evolution

Era	Period and epoch	Age (years $\times 10^6$)	Fauna (first appearance)
Cenozoic	Quaternary		
	Holocene	0.01–	No new megafauna
	Pleistocene	1.8–0.01	*Homo erectus, H. sapiens*
	Tertiary		
	Pliocene	5–1.8	Hominids: *Australopithecus*, *Homo habilis*
	Miocene	25–5	Hominoids (apes)
	Oligocene	35–25	"Progressive" brains
	Eocene	55–35	Progressive ungulates, Anthropoids
	Paleocene	65–55	Primates[a] and carnivores
Mesozoic	Cretaceous	140–65	Marsupials, Placentals
	Jurassic	210–140	Birds, mammal endocast
	Triassic	250–210	Mammals
Paleozoic	Permian	285–250	Primitive dinosaurs
	Carboniferous	360–290	Reptiles
	Devonian	410–360	Bony fish and amphibians
	Silurian	440–410	Jawed fish
	Ordovician	500–440	Jawless fish
	Cambrian	550	First chordates

[a]Primate teeth reported in late Cretaceous deposits. Paleocene primate identification is controversial. There is consensus recognizing early Eocene tarsier-like species, middle Eocene lemur-like species, and recently discovered late Eocene simian species.

tions. The Cenozoic was more variable in every way, with more diverse and sometimes chilling climates and periods of major mountain building. A burgeoning animal and plant life is evident in fossils from sediments laid down during all of these periods.

Second, there were several mass extinctions, with the greatest, at the end of the Permian Era, signaling the beginning of the Mesozoic. The most famous extinction, attributed to impact by a small asteroid, occurred at the end of the Mesozoic (the K–T, or Cretaceous–Tertiary boundary) 65 million years ago. Niches, emptied of their otherwise well-adapted organisms that could not survive the environmental catastrophe, could then be filled by suitably adapted birds, mammals, teleost fish, and snakes.

Third, although mammals were present during much of the 185 million years of the Mesozoic Era, all were small-bodied, none larger than living cats. They were probably nocturnal in their habits. Only during the Cenozoic did very large mammal species appear, and even today the average mammalian species is about cat size and nocturnal. Humans are giant vertebrates, physically larger and heavier than 90% or more of living species. Anthropoid primates are an unusual group of mammals; species of the suborder Anthropoidea (monkeys, apes, and humans) are diurnal and are well-adapted for color vision.

Finally, a major environmental event in human history may have been the Pliocene drying of the Mediterranean about 5 million years ago, which probably contributed to natural selection among chimpanzee-like primates for a species that became the earliest hominid. Extensive glaciation characterized the Pleistocene Epoch and may have driven the evolution of the human species to its present grade.

B. Fossil Brains Revisited

From the history of the brain in fish, amphibians, and reptiles as available from the fossil record, the most unusual inference is that these can all be treated as lower vertebrates in encephalization (Fig. 5). The exceptions are sharks and, perhaps, the ostrich-like dinosaurs (ornithomimids). Here is a list of a few more outstanding discoveries.

First, from the evidence of small (<15 cm long) Carboniferous (350 Ma) fossil fish, the diversity in living teleosts was probably foreshadowed by some of the earliest bony fish. They had optic lobes enlarged in ways comparable to those of living fish, such as trout, that feed at or near the surface of the water and rely on visual information.

Second, although sharks and other cartilaginous fish are often considered primitive, they are not primitive with respect to their brains. Many sharks are big-brained, overlapping the distributions of relative brain size of the lower vertebrates, on the one hand, and of birds and mammals, on the other hand. There is one uncrushed endocast known in a Paleozoic shark, the earliest evidence of encephalization beyond the grade of living bony fish. The species was comparable in both brain and body size to the living horned shark (*Heterodontus*).

Third, dinosaurs continue to be libeled with the walnut-size brain label despite evidence to the contrary. Their brains were within the expected size range for reptiles. The Tyrannosaurus 404-ml endocast implies a brain in the size range of those of living deer—small for an elephant-sized mammal but impressive for a reptile. The basic data are shown in Fig. 6.

Fourth, the major transition from water to land in the amphibians more than 350 Ma was accomplished without enlargement of the brain, and there has been no enlargement since if one compares present amphibians with their fossil ancestors. Of course, there was and is, considerable diversity in relative brain size within each of the classes of living lower vertebrates, just as there is in birds and mammals, and there has been significant reorganization of the brain across species of lower vertebrates, especially between classes but also within classes. By any standard, one must recognize the brain in living reptiles as more specialized than that of fish or amphibians.

Finally, of the major lateralization of function in the living human brain, and the recognized lateralization in the brains of many other living species, there is little or no evidence at a gross level. There is no good fossil evidence for such lateralization, although some has been claimed and evidence may be forthcoming. The problem is that asymmetries are difficult to establish, even in living brains, and are almost impossible to establish in fossils, which are often asymmetrically distorted and partially crushed.

These statements sum up the evidence on brain evolution in about three-fourths of the vertebrate species. There remains the story of about 10,000

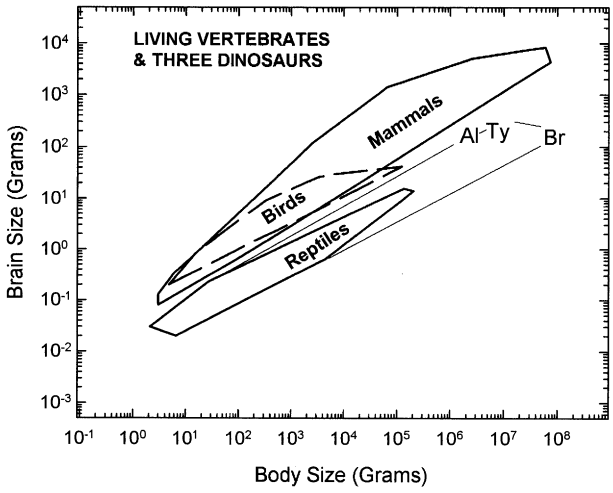


Figure 6 Data from Fig. 5 for mammals, birds, reptiles, and three dinosaurs (Al, *Allosaurus*; Ty, *Tyranosaurus*; and Br, *Brachiosaurus*), showing extension of reptilian minimum convex polygon by adding dinosaur data.

species of birds and 5000 species of mammals, in which encephalization is a major feature. As background, and for a better sense of the potential and limitations of the method, I discuss the relevant data from Fig. 5, namely, those on reptiles, birds, and mammals. These are presented in Fig. 6, in which distracting information from the three classes of fish were removed and to which I have added data on three dinosaurs.

The three added species are the large well-known carnivorous dinosaurs, *Tyrannosaurus* and *Allosaurus*, and the largest of all dinosaurs in which there are reliable body size estimates, *Brachiosaurus*. To illustrate the method and its use, I will review the status of the "ostrich dinosaurs" from this perspective. One of these is a relatively small late Cretaceous carnivorous dinosaur, *Troodon*, which has been described as encephalized as large living birds and in the range of encephalization of living mammals. The analysis depends on estimating body size as well as brain size. There is a reliable estimate of *Troodon*'s body size as about 45 kg. Its brain size was first estimated by inspection and by perceived similarity to an ostrich's brain size of 40 g. However, with the help of a computer program recently developed and a qunatitative analysis of its brain size, I now estimate its brain as about 20 ml in volume. This is somewhat smaller than the albatross's 27-ml brain as determined by the same computer analysis. If *Troodon* had a 20 ml brain in a 45 kg body, it would fit within the reptile polygon of Fig. 6. More analysis is needed, but it is evident that

reports of the past 30 years of these large-brained dinosaurs need to be reevaluated. Other dinosaur data, such as those in Fig. 6, appear to be reasonably reliable, and provide an acceptable basis for evaluating dinosaur brain evolution. The ostrich that provided data for Fig. 6 weighed 133 kg and had a 42 g brain. It is the end-point of the avian polygon, which falls somewhat below the corresponding point on the polygon for living mammals. I have not entered data points for albatross at a body size of 12 kg, but it would fit comfortably within the avian polygon. *Troodon* fits within the reptilian polygon as extended in Fig. 6 to include a few dinosaurs. *Troodon* was indeed among the larger brained dinosaurs, but it probably did not reach either the mammalian or avian polygon.

The analysis with convex polygons is, in short, based on inclusion or exclusion of a particular species within the taxon described by a particular polygon. In Fig. 5, it was this approach that demonstrated the anomalous situation for living electric fish and cartilaginous fish and the "normal" position for agnathans. In Fig. 6, it indicates that dinosaurs were reptilian rather than avian in relative brain size.

C. Birds and Mammals

The history of the bird brain is not as well-known as that of mammals, and the present diversity in brains in birds does not appear to be as dramatic. However, the brain of *Archaeopteryx* was bird-like primarily because it filled the cranial cavity and was larger than that in comparable reptiles. There was no Wulst—that is, the dorsal enlargement of the forebrain in living birds that functions equivalently to primary visual cortex in mammals. The next significant evidence in the history of birds is from an Eocene whimbril-like bird, *Numenius gypsorum*, in which the brain is somewhat smaller than in comparable living birds of its body size but, most dramatically, its forebrain is clearly much smaller so that its optic lobes are more fully exposed than in living birds. Endocasts of later birds are indistinguishable from those of their living relatives.

In their early history, the mammals were small, probably nocturnal insectivorous creatures. Their adaptation for life in nocturnal niches could have been the major selection pressure to explain the "advance" to a mammalian grade of brain morphology and encephalization. The characteristic morphological feature of the brain of living mammals is the presence of neocortex, the six-layered neuronally rich outer covering of the forebrain. Its presence can often be established on an endocast because a major fissure, the rhinal fissure, is its ventral boundary.

The earliest mammalian brain is known from an Upper Jurassic endocast of *Triconodon mordax* and is about 150 million years old. The lateral surface of this endocast is not preserved well enough to indicate whether or not there was a rhinal fissure; thus, positive evidence regarding the presence of neocortex is not available. In encephalization, however, its brain was comparable to that of small–brained living species such as opossums and hedgehogs, in which neocortex is present. It is therefore likely that neocortex appeared at least 150 Ma. The best assumption from available information is that neocortex is, in fact, part of the suite of traits that characterized the mammals from the beginning of their evolution, at least 50 million years earlier.

Mammals in which the endocasts are sufficiently complete to show a rhinal fissure, if present, are about 75 Ma. They are from a unique assemblage of Late Cretaceous mammals from the Gobi desert, which includes early placentals. The most common mammals of the time, the multituberculates, were unrelated to any living species. Superficially, multituberculates probably looked like living rodents or insectivores. Their life span as an order was about 120 million years, between about 150 and 30 Ma, a very long span for a mammalian group. The specimen in which the endocast is best known, *Chulsanbaatar*, weighed no more than about 15 g, a small mammal even for the Mesozoic and smaller than most living species of mice. There is a suggestion of a rhinal fissure in its endocast, although there is disagreement about where it is located. Whether or not one can see a rhinal fissure, from its grade of encephalization it is very likely that neocortex was present in its brain.

D. Neocorticalization

Neocorticalization is a concept in comparative neuroanatomy, based on comparisons among living species. For example, it describes the fact that primates have relatively and absolutely more neocortex than insectivores. One makes the statement: Primates are more neocorticalized than insectivores. Of course, living insectivores did not evolve to change into living primates under natural selection. But there is an evolutionary translation of the statement: The ancestors of living insectivores were members of an order (or other taxon) of mammals that probably included at least one species from which primates evolved. This

species has not been identified, but as evolutionists positing a relationship between insectivores and primates as "sister" groups, we have to assume that it existed. Neocorticalization is then understood as part of the differentiation of daughter species from parent species: One daughter species of a fossil insectivore, which had relatively more neocortex than the parent species, was the ancestral primate species. These statements are almost parodies of evolutionary analysis, but they approximate a correct analysis.

The concept of neocorticalization can also be used in another sense—as describing the history of a trait in successive populations of species. If we sample a broad range of species across geological time and determine that later species had relatively more neocortex than earlier species, we could state that neocorticalization had occurred, even though we would not be able to determine its phylogenetic history. Such a discovery would be enough to suggest that there was a selective advantage in an increase in neocortex. The analytic model to be applied might be an elaboration of simple models of phenotypic evolution within lineages. It would require theorizing about selective advantages above the species level—about broad evolutionary "landscapes" that are contexts for interaction among species.

Neocorticalization in this sense can be quantified as a feature of the history of the mammals. It is more or less evident from a simple inspection of endocasts, but the quantitative effect can only be demonstrated with the help of some statistical analysis. I present such an analysis in Fig. 7. For the analysis, I measured the planar projection of neocortex and olfactory bulbs in 59 species of living and fossil ungulates and carnivores. The sample included 38 Carnivora, 7 Creodonta, 4 Condylarthra, 5 species from extinct South American (neotropical) ungulate orders, 1 Eocene perissodactyl (*Hyrachyus*, an ancestral rhinoceros), and 4 progressive ungulates (artiodactyls). The complete sample consisted of 35 fossil species and 24 living species.

The results were analyzed as neocortical and olfactory bulb quotients, determined by partialling out the effect of body size much as in encephalization quotients. A quotient of 1.0 means that the size was as expected for a species at the centroid, 0.5 means it was half as big, and so on. As presented in Fig. 7, the results lead to four conclusions. There is first the fact of neocorticalization. Later species tended to have relatively more neocortex than did earlier species. This is the meaning of the significantly positive slope of the regression line of the neocortical quotient against geological time.

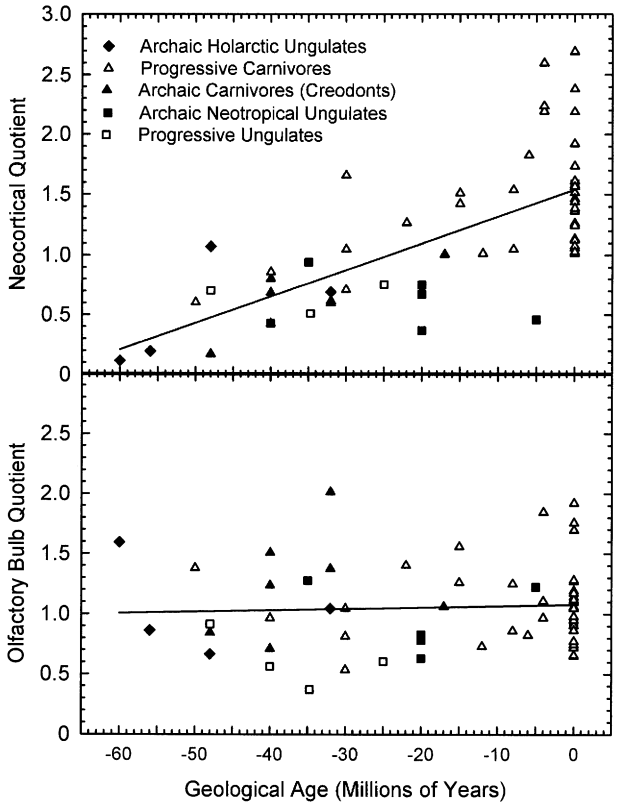


Figure 7 (Top) Change in relative neocortical surface area (neocortical quotient) as a function of geological age. "Progressive" change noted here (positive slope of regression line) indicates increased neocorticalization over time. Each point is a species. (Bottom) Absence of change in relative surface area of the olfactory bulbs as a function of geological age. (Redrawn with permission from Jerison, in Jones and Peters, 1990).

Second, species from archaic orders (Fig. 7, filled symbols) tended to have less neocortex than did species from progressive orders (Fig. 7, open symbols). Thus, 3 of the 4 archaic ungulate species, 5 of the 7 archaic carnivore (creodont) species, and 4 of the 5 Neotropical ungulate species (also archaic in that their orders are extinct) are below the regression line. Twelve of the 16 archaic species thus had less neocortex than would be predicted for their geological age by an unbiased regression analysis. For those who enjoy playing with statistics, a chi-square analysis contrasts this with an expected even split: $\chi^2 = 4$, $df = 1$, $p < 0.05$.

The third result in a comparison between progressive and archaic species limited to fossil Carnivora versus Creodonta. It is in two parts. First, the Carnivora points appear generally to be higher than those of the Creodonta. Second, the Carnivora points seem to show more "progress" over time than do those

of the Creodonta. It is not possible to test the first part properly, because the species are from different geological times, and there is no obvious way to control the time variable. The second part, however, can be tested by simple regression analysis. The correlation between age and neocortical quotient for 15 Carnivora species was $r = 0.72$ ($p < 0.01$). For seven Creodonta it was $r = 0.42$ ($p > 0.05$). Only the Carnivora were demonstrably progressively neocorticalized.

The result is important for evolutionary analysis of the relations between true carnivores and creodonts. The argument is summed up by R. L. Carroll as follows: "Romer (1966) and Jerison (1973) stigmatized the creodonts as archaic and small brained, but Radinsky (1977) demonstrated that relative brain size increased as rapidly among creodonts as it did in the early members of the Carnivora, together with an increase in the extent of the neocortex." The quantitative analysis supports Romer's view as mentioned by Carroll. (My contribution in 1973 was mainly to quote Romer and to provide very limited quantitative data. The current confirmation of our older view is possible because of the additional data collected by Radinsky, which permitted a statistical test.)

The final conclusion is that unlike neocortex, the olfactory bulbs did not change in relative size with the passage of time. The correlation between the olfactory bulb quotient and geological time was $r = 0.1$, which is not significantly different from zero. This is an important point because it shows that this approach is fine enough to discriminate between the presence and absence of change. It also helps quash a myth about what is "primitive" and "progressive" in brain evolution.

Although careful students do not make the error, some neurobiologists assume that having large olfactory bulbs is a primitive mammalian trait and that the olfactory bulbs became relatively smaller as the mammals evolved. Fig. 7 corrects this error by showing that olfactory bulbs have been a stable feature of the brain in Tertiary carnivores and ungulates. The misconception is a bit of primate chauvinism, as it were. Primates (at least the anthropoids) are neocortical specialists, but they are deficient mammals in olfactory development. A reduced role for olfaction is part of the adaptive mosaic of the adaptive zone of simian primates and is not a broad feature of mammalian evolution. (It also complicated the factor analysis presented earlier.) Neocorticalization, on the other hand, appeared as a general trend in many mammalian groups and its relative absence in the insectivores and many marsupials is correctly recognized as a primitive feature in these groups.

VI. CONCLUSIONS

First on neocorticalization. The fossil evidence indicates that there was neocorticalization in mammals and that it can be detected even in samples as small as 15 species of Carnivora. Evolution of the carnivores involved neocorticalization in another sense, in that the two great orders of Tertiary carnivores (Creodonta and Carnivora) differed in the extent of neocorticalization. This could have been a factor in the survival of true carnivores. In any event, the history of neocorticalization indicates that there was almost certainly some benefit derived from the expansion of neocortex. The fossil evidence therefore confirms conventional wisdom in neurobiology that it was a progressive thing for mammals to evolve neocortex and (perhaps within limits) more is better.

If the conclusions on neocortex are expected, those on the olfactory bulbs are not. These structures are surprisingly constant features in mammalian brains. There is no reason to have predicted that they would not evolve to large or small size relative to other parts of the brain, depending on the extent of olfactory specializations in particular species. But according to current evidence from fossil brains, the olfactory bulbs have been constant and relatively unchanging features that make a brain a mammal's brain. They are not unusually enlarged in any species; most mammals are olfactory specialists. Evolutionary changes in the olfactory bulbs occurred mainly in a negative way, by reduction. The reduced state of olfactory bulbs in humans and other primates (and their complete absence in some cetaceans) merely reflects the extremes of diversity that are possible as the brain evolved to control the activities of mammals in the variety of niches in which they function.

From these and related data it seems likely that encephalization in mammals was driven by neocorticalization. One mammalian trend was toward enlarged neocortex, and since neocortex is a fairly fixed fraction of total brain size the enlargement of the brain was presumably correlated with the increased size of neocortex. Because neocortical function is deeply involved with cognitive functions—knowledge of "reality," expanded neocortex would be associated with more elaborate cognition, a major suggestion about the evolution of mind.

It would be appropriate to look more closely at neocortical functions and assume that the evolutionary advantage conferred by these functions was the engine driving progressive brain evolution in mammals. However, it would be a mistake to make much of such an idea of progress. It is true that neocorticalized species are more prevalent now than in the distant past. It must also be true that some fitness is associated with this aspect of the brain's evolution. But there are many successful living species that are at a very ancient grade of mammalian neocorticalization. Hedgehogs in Europe and opossums in America are outstanding examples because they are very fit in the evolutionary sense. They may litter our highways because of their "stupidity" in refusing to yield the right of way to cars and trucks, but the litter is part of the evidence of their reproductive success. And they manage this at a grade of neocorticalization and encephalization that some mammalian species reached 150 Ma.

The analysis of neocorticalization suggests that comparable advances occurred in birds with the expansion of their forebrains. It is certainly true that the avian forebrain is much enlarged compared to that of reptiles, and the grade of encephalization in birds is probably related to forebrain enlargement. The optic lobes (midbrain, homologous to the superior colliculi in mammals) also seem much larger in birds than in reptiles, at least to an analysis by eye.

The final conclusion, however, must be that animals do not live by brains alone. The majority of living vertebrates get along with about as much brain as was present in their earliest ancestors. Adaptation to one's niche can be accomplished in many ways, and to adapt behaviorally by brain enlargement is expensive energetically. The brain is profligate in its use of energy, and almost any other solution to an adaptational problem is less costly. In those groups that adopted encephalization as an adaptive strategy, however, there was evidently a real gain in fitness. Encephalization appeared in many very distantly related species of birds and mammals. It is a general rather than specific adaptation for increased total information processing capacity. It is sometimes considered as the brain correlate of intelligence. If that is true, it must mean that there are many intelligences that evolved, since encephalization is an overall sum of enlargements of constituent regions within the brain. Because the different regional enlargements are correlated with different behavioral capacities, there would be different kinds of intelligences in different species.

See Also the Following Articles

BRAIN ANATOMY AND NETWORKS • BRAIN DEVELOPMENT • CIRCADIAN RHYTHMS • INTELLIGENCE • LANGUAGE ACQUISITION • LATERALITY • NEOCORTEX

Suggested Reading

Braitenberg, V., and Schüz, A. (1998). *Cortex: Statistics and Geometry of Neural Connectivity*, Second Ed., Springer Verlag, New York.

Butler, A. B., and Hodos, W. (1996). *Comparative Vertebrate Neuroanatomy*. Wiley–Liss, New York.

Carroll, R. L. (1988). *Vertebrate Paleontology and Evolution*. Freeman, New York. [References to Jerison, Romer, and Radinsky.]

Dawkins, R. (1987). *The Blind Wachmaker*. Norton, New York.

Deacon, T. W. (1997). *The Symbolic Species: The Co-evolution of Language and the Brain*. W.W. Norton, New York.

Falk, D., and Gibson, K. (Eds.) (2001). *Evolutionary Anatomy of the Primate Cerebral Cortex*. Cambridge University Press, Cambridge, UK. [Chapters by Finlay and by Jerison.]

Farlow, J. O., and Brett-Surman, M. K. (Eds.) (1997). *The Complete Dinosaur*. University of Indiana Press, Bloomington Indiana.

Harvey, P. H., and Pagel, M. D. (1991). *The Comparative Method in Evolutionary Biology*. Oxford Univ. Press, Oxford.

Jerison, H. J. (1991). *Brain Size and the Evolution of Mind*, 59th James Arthur Lecture on the Evolution of the Human Brain. American Museum of Natural History, New York.

Jones, E. G., and Peters, A. (Eds.) (1990). *Cerebral Cortex: Comparative Structure and Evolution of Cerebral Cortex (Parts I and II). Volumes 8A and 8B*. Plenum, New York. [Chapters by Jerison, Johnson, Ulinski, and Welker.]

Martin, R. D. (1990). *Primate Origins and Evolution: A Phylogentic Reconstruction*. Chapman & Hall, London.

Novacek, M. (1996). *Dinosaurs of the Flaming Cliffs*. Doubleday, New York.

Roth, G., and Wulliman, M. F. (Eds.) (2001). *Brain Evoluation and Cognition*. Wiley & Sons, New York. [Chapters by Hofman, Hodos, Jerison, and Schuez.]

Stephan, H., Baron, G., and Frahm, H. D. (1991). *Insectivora: With a Stereotaxic Atlas of the Hedgehog Brain, Comparative Brain Research in Mammals*, Vol. 1, Springer-Verlag, New York.

Taquet, P. (1992). Dinosaures et Mammifères du Désert de Gobi. Muséum National d'Histoire Naturelle, Paris.

Eye Movements

CHARLES J. BRUCE and HARRIET R. FRIEDMAN
Yale University School of Medicine

GLOSSARY

frontal eye field Motor cortex for eye movements. In man and monkey, the frontal eye field lies in premotor cortex, just rostral to the hand representation in motor cortex.

neural integrator Brain stem circuit that maintains the current position of the eye by integrating all eye velocity commands. Its tonically spiking output neurons (tonic neurons) are located in the nucleus prepositus hypoglossi, the interstitial nucleus of Cajal, and the medial vestibular nucleus.

nystagmus Any to-and-fro movement of the eyes. Nystagmus can be normal (e.g., continued optokinetic stimulation causes optokinetic nystagmus, such as watching a train go by) or pathological, both congenital and acquired. The waveform of pathological nystagmus often indicates the damaged structure (e.g., sawtooth nystagmus indicates vestibular system damage).

optokinetic reflex/nystagmus Smooth, relatively slow eye movements elicited by movement of the whole visual field. Prolonged optokinetic reflex becomes optokinetic nystagmus when the slow movement alternates with quick-phase movements in the opposite direction.

response field The spatial parameters for the sensory stimulation and/or motor behavior associated with the spiking (or maximal spiking) of a particular neuron.

saccade generator Circuitry in brain stem reticular formation that underlies all rapid eye movements. Its principal output cells are short-lead excitatory burst neurons (EBNs): Horizontal EBNs are in the caudal paramedian pontine reticular formation near the abducens nucleus, and vertical EBNs are in the midbrain near the oculomotor nucleus. Other key cell types are inhibitory burst neurons, long-lead excitatory burst neurons, and omnipause neurons.

saccades Short-duration, high-velocity (rapid) eye movements that quickly move the eyes to a new position; can include both voluntary saccades and the quick phase of vestibular or optokinetic nystagmus. Special saccade types include memory saccades, in which the eye movement is directed to the location of a remembered target not currently visible, and antisaccades, in which the movement is deliberately directed opposite to the target's location.

smooth pursuit Smooth, continuous eye movements used to track a moving target.

superior colliculus Midbrain visuomotor structure for triggering visually guided saccades.

vestibuloocular reflex Smooth, relatively slow eye movements elicited by the vestibular sense that serve to cancel movements of the visual stimulus on the retina (retinal slip) caused by head movements.

This article on the human brain's widely distributed circuitry for controlling the eyes stresses the functional significance of eye movements. It takes an evolutionary perspective by emphasizing the primacy of the image-stabilization system (vestibuloocular reflex, optokinetic reflex, and quick phases of nystagmus) and how the relatively new eye movements of primates (saccades, smooth pursuit, fixation, and vergence) serve the primate's high-resolution foveal vision by engaging the far older image-stabilization circuits. The anatomy and physiology of principal oculomotor structures are reviewed, with emphasis on the frontal eye field region of primate neocortex.

I. INTRODUCTION

Processing the visual sense is one of the most considerable endeavors of the human brain. One

measure of the complexity of this function is that more than 30 distinct "visual" brain regions have been identified as contributing to the visual percept. Although eye movements are not required for visual function, they immensely improve the ability of primates to gather relevant, high-quality visual data for processing by this large expanse of neocortex. Thus, it is not surprising that a sizeable complement of cortical and subcortical circuits of the human brain is concerned with moving the eyes, nor that eye movements are often closely related to cognitive behavior.

This article is organized around three fundamental questions:

Why move the eyes? (What are the functions of eye movements?)
How do the eyes move? (What is the phenomenology and kinematics of eye movements?)
What moves the eyes? (Which neural structures and circuits effect eye movements?)

Table I
Principal Functions of Eye Movements

Image stabilization: Keep the visual image stationary on the retina
Purpose: To prevent blurred vision caused by movements of the head and body
Movements
Vestibuloocular reflex in response to vestibular signals
Optokinetic reflex in response to whole-field visual motion
Quick phases of vestibuloocular and optokinetic nystagmus
Foveation: Capture and keep particular stimuli on the foveal part of retina
Purpose: To facilitate fine, detailed viewing and visual analysis of important objects
Movements
Saccades to rapidly foveate new of different stimuli
Smooth pursuit to track moving visual stimuli
Fixation to maintain foveation of stationary visual stimuli
Binocular alignment: Foveate the same object/point in both retina
Purpose: To facilitate single vision and steropsis in order to extract relative depth (3D) perception from fine disparities between the 2D images in the two eyes
Movements
Disjunctive eye movements, both convergent and divergent (other eye movements are conjugate/conjunctive, i.e., equal in the two eyes)

II. WHY MOVE THE EYES?

Eye movements serve vision in three principal ways (Table I): to stabilize images on the retina, especially against movements of the head and body (image-stabilization system); to image the important details of the visual world on the most sensitive part of the eye, the fovea (foveation system); and to align the retinal images in the two eyes in order to promote single vision and stereopsis (vergence-accommodation system). Eye contact and avoidance also have very important roles in interpersonal behavior.

A. The Image-Stabilization System

Ironically, one of the principal reasons for the eyes to move is to keep the direction of gaze stationary—that is, fixed in space. The underlying objective is to keep the visual image of the external world stationary on the retina despite movements of the head and body because even slight movements of this image across the retina (i.e., *retinal slip*) cause considerable visual blur. For this reason, seeing animals need image-stabilization strategies and reflexes, and for vertebrates this principally involves extraocular muscles that rotate the eyeball within its bony orbit so as to cancel or minimize the retinal slip consequent to movements of the head. Furthermore, the anatomical and physiological substrates of image-stabilization eye movements are remarkably similar across vertebrate species, from primitive fish to modern apes.

Three distinctive types of eye movements are associated with image stabilization: the vestibuloocular reflex (VOR), the optokinetic reflex (OKR), and the extremely high-speed resetting movements called the "quick phase" (of nystagmus). The VOR is a very short latency reflex instigated by the vestibular system in response to body/head rotation. The OKR, on the other hand, is triggered by visual motion. In the laboratory, VOR and OKR can be separately studied; however, in the real world both are activated whenever the head moves, and they synergistically sum to compensate for most head movements and thereby minimize movement of the visual image on the retina. In contrast, the quick phase is an anticompensatory movement that very rapidly returns the eye to a more central position whenever it is taken to eccentric orbital positions by the vestibulooptokinetic reflexes.

These image-stabilization eye movements are reflexive and automatic. From an evolutionary perspective, they are the most basic, primary eye movements. Indeed, for most vertebrate species VOR, OKR, and the quick-phase movements of nystagmus are the only eye movements.

B. The Foveation System

As in most primates, the human retina has a tremendously specialized central zone, the fovea, where visual acuity is 1000 times better than vision just 10° eccentric. Hence, to "look at" something is effectively to foveate it. However, the fovea subtends only 1° of visual angle (equivalent to the full moon's subtend). Therefore, the second principal function of eye movements is to foveate important parts of the visual scene.

Foveation is accomplished by a triad of voluntary eye movements: saccades, fixation, and smooth pursuit. Of these, saccades are the most conspicuous because humans incessantly make saccades in order to maximize retrieval of high-quality visual information by foveal viewing. Although the saccadic movement is ballistic and stereotyped, saccades are voluntary with regard to the choice of whether or not to make a saccadic movement (when to look) and what saccadic vector to program (where to look). For example, reading is accomplished with a succession of voluntary saccades that march across each line (Fig. 1), and during each fixation (between saccades) the reader processes the foveated word(s) while simultaneously programming the next saccade.

Smooth-pursuit eye movements, the third member of the foveation triad, assists the viewing of objects moving in space by matching eye velocity to target velocity, up to 50–100°/sec. A record of smooth pursuit is shown in Fig. 2. Pursuit is sometimes confused with the OKR; however, they are very different movements, as described later, and pursuit often must override the combined vestibulooptokinetic reflexes (e.g., during combined head–eye tracking of a moving stimulus). Instead, pursuit is more equivalent to the continued fixation of a target that happens to be moving.

Whereas image stabilization is effected by a largely reflexive brain stem system, the more voluntary, cognitive foveation system and its constituent eye movements (saccades, fixation, and pursuit) involve many areas of neocortex, most notably the frontal eye field (FEF) region of the frontal lobe.

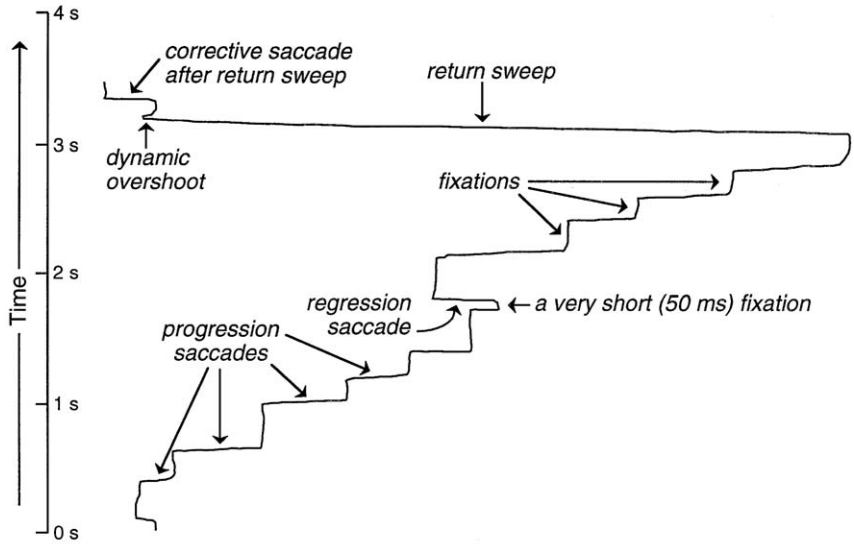


Figure 1 Eye movements during reading. Eye movement records showing the pattern of saccades and fixations made while reading the sentence shown in the abscissa. Records were obtained using the infrared photoelectric scleral reflection technique (at 100 Hz). Time increases along the ordinate; therefore, to recreate the eye's movement the traces should be read from bottom to top. Notice that the sentence/line was read with 10 progressive (rightward) saccades and one regressive (leftward saccade), and the majority of the time spent reading the line is during the 11 fixations between the saccades (adapted from J. K. O'Regan, *Eye Movements and Their Role in Visual and Cognitive Processes* (E. Kowler, Ed.), copyright 1990, with permission from Elsevier Science).

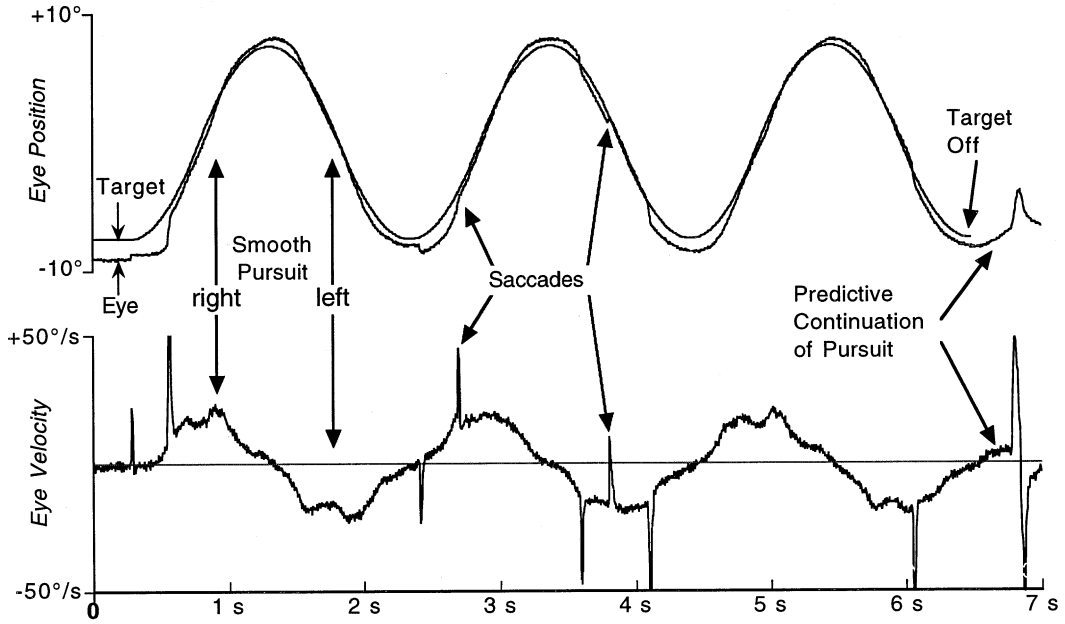


Figure 2 Smooth-pursuit eye movements. Horizontal eye position (top) and eye velocity (bottom) records of a monkey's smooth-pursuit tracking of a small spot with sinusoidal motion (amplitude $\pm 7.5^{\circ}$, frequency 0.5 Hz). Tracking sinusoidal motion is equivalent to tracking the bob of a pendulum, a classic pursuit test. Maximum target velocity is 23.6°/sec, and maximum pursuit velocity was $\sim 25^{\circ}$/sec. Saccades are easily identified by their spikes in the eye velocity traces. A total of seven saccades were made during the 6 sec of sinusoidal motion, starting with a large catch-up saccade shortly after the start of the target motion. At the end of the record smooth pursuit continued for part of a fourth cycle (predictive pursuit), even though the target was extinguished after three cycles (C. J. Bruce, unpublished data).

C. The Vergence-Accommodation System

As a consequence of frontally placed eyes (found in all primates and in predatory species of other vertebrate orders), the visual fields of the two eyes have considerable overlap. Stereoscopic depth can be extracted from the small differences in the these overlapping images, yielding detailed information about the three-dimensional (3D) structure of the visual world. However, good stereopsis requires that both eyes be directed at (i.e., foveate) the same object in visual space. Vergence (or disjunctive) eye movements provide such binocular alignment in response to changing fixation target distances, which necessitate that the two eyes point in different directions. In contrast, all of the aforementioned eye movement types are conjugate (or equivalently, conjunctive). Thus, the two eyes move as one during saccades, pursuit, VOR, etc. but not during vergence. The basic neural mechanisms of vergence and accommodation are in the brain stem, especially in midbrain regions immediately adjacent to the oculomotor nucleus (n. III). However, the vergence-accommodation system is also dependent on neocortex, especially primary visual (striate) cortex and the frontal lobe cortex.

III. HOW DO THE EYES MOVE? PHENOMENOLOGY OF EYE MOVEMENTS

The oculomotor response can be reduced to seven basic types of eye movements (Table I) that achieve the three principal functions of image stabilization, foveation, and binocular alignment. Although these movements differ in many regards (e.g., whether they are fast or slow, voluntary or reflexive, as well as which of the three functions is served), the kinematics of eye movements are generic. All eye movements share a common final path represented by three cranial nerve nuclei and the six pairs of eye muscles that they control. This section describes the basic mechanics of eye movements and the muscles, nerves, and motor nuclei involved.

A. Eye Movements Are Three-Dimensional Rotations of the Eye in the Orbit

Eye movements can generally be regarded as rotations of the eye about the center of the eyeball. Like a ball in a socket, the eyeball spins inside its bony orbit in the skull. As shown in Fig. 3, any particular eye position can logically be obtained by a combination of rotations about the three axes through the center of rotation.

B. Each Axis of Rotation Is Controlled by a Pair of Opposing Muscles

Eye rotations are effected by six extraocular muscles that are organized as three opposed pairs. Each of these muscles has a major action: The medial and lateral rectus muscles effect horizontal eye movements, with the lateral rectus pulling the eye temporally (abduction) and the medial rectus pulling nasally (adduction). The primary actions of the superior and inferior rectus muscles are elevation and depression, respectively. The primary actions of the superior and inferior oblique muscles are torsional rotations of the eye about the line of sight, with the superior oblique effecting intortion, and the inferior oblique effecting extorsion. However, the actions of the superior and inferior recti and the superior and inferior obliques are better regarded as "mixed" in that both pairs actually effect combined vertical and torsional eye movements. Vertical and torsional eye movements are further discussed later.

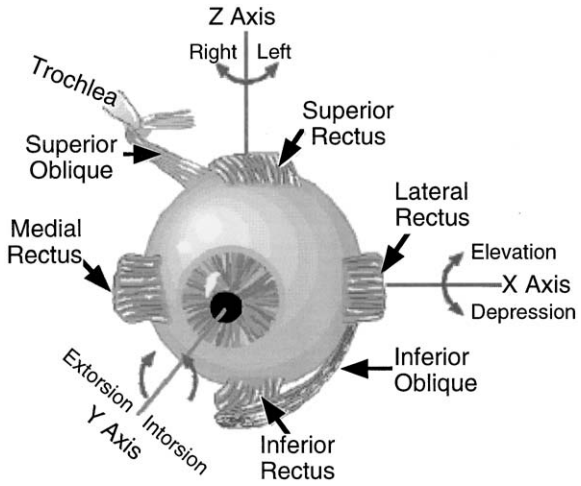


Figure 3 Axes of rotation and muscles of the eye. The eye rotates about three axes to effect horizontal, vertical, and torsional movements. These rotations are implemented via six extraocular muscles for each eye, as shown. Adduction and abduction about the z axis are accomplished by the lateral and medial rectus muscles. Raising and lowering the eyes (x axis rotations) are accomplished by the inferior and superior rectus muscles in association with the inferior and superior obliques. Likewise, torsional movements (y axis rotations) require the inferior and superior obliques together with the superior and inferior rectus muscles.

C. The Oculomotor Cranial Nerves and Their Brain Stem Nuclei

The six extraocular muscles are innervated by cranial nerves III (oculomotor), IV (trochlear), and VI (abducens). Nerve IV innervates the contralateral superior oblique, nerve VI the ipsilateral lateral rectus muscle, and nerve III the remaining four extraocular muscles as well as the eye lid. The neurons giving rise to these general somatic efferents are termed oculomotor neurons (OMNs). Their cell bodies are located in three eponymous nuclei that lie in pairs along the medial longitudinal fasciculus near the midline of the brain stem: the oculomotor, trochlear, and abducens nuclei, which are often grouped and simply termed the oculomotor nuclei.

D. Neural Activity in Oculomotor Neurons

In the early 1970s, several pioneering laboratories recorded the activity of single neurons in oculomotor nuclei while monitoring the eye movements of alert monkeys. They found that every OMN had the same characteristic firing pattern, organized about the eye's position and motion in the plane of its target extraocular muscle (Fig. 4). The clarity and constancy of the OMN responses galvanized sensorimotor neurophysiology, setting the stage for the analysis of circuits responsible for OMN responses and for tracing the antecedents of these activities further into the brain stem and then to the cerebral and cerebellar cortices.

1. Coding of Eye Position by Oculomotor Neurons

The relationship between the discharge rate of any OMN and the static position of the eye is a linear function of the eye's coordinate, θ, as measured along the pulling direction of the OMN's target muscle: $R_{OMN}=A\cdot(\theta-\theta_0)$, where R_{OMN} is spikes/sec, θ_0 is the

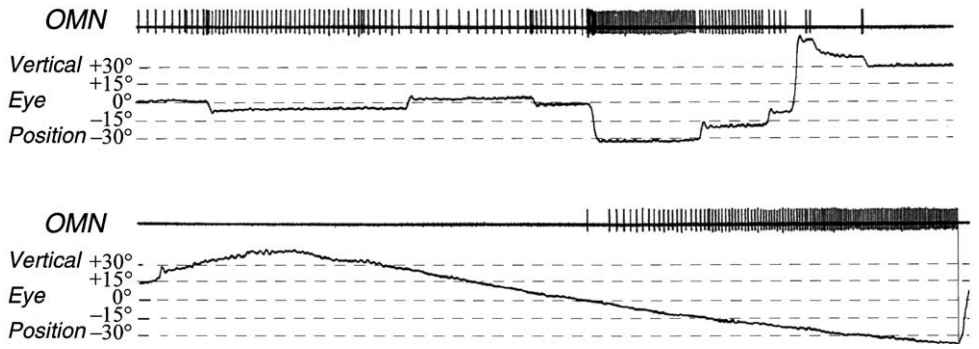


Figure 4 Discharge characteristics of a neuron in the oculomotor nucleus (OMN). The firing rate of this neuron increased in association with downward eye movement. The upper set of traces shows spontaneous saccadic eye movements with intervening periods of fixation. The lower set shows smooth pursuit brought about by moving an object in front of the monkey. In each part the top trace has the neuron's spikes during the eye movement; the **lower** trace (solid line) is the vertical eye position. The dashed lines represent degrees of deviation from straight-ahead gaze; upward deflection corresponds to elevation and downward deflection to depression of the eye (adapted with permission from P. H. Schiller, *Exp. Brain Res*. **10,** 347–362, 1970. copyright © 1970 by Springer-Verlag).

OMN's recruitment position, or threshold, and a typical value for factor A is 4 Hz/deg. Most θ_0 are more than 15° in the "off" direction (i.e., $\theta_0 \leq 15°$), meaning that most motor units are active during central fixation. Two nonlinear constraints on R_{OMN} are necessary, however: $R_{\mathrm{OMN}}=0$ for all $\theta \leq \theta_0$, because neurons cannot have a negative rate, and $R_{\mathrm{OMN}} \leq 500$ because OMNs typically saturate at 500 Hz.

This linear relation reflects the passive physical properties of the eyeball in the orbit, namely the elastic forces that are constantly pulling the eye toward its central position with a force proportional to its eccentricity. Thus, larger deviations from the central position require proportionally larger muscular forces, which in turn require larger rates in the agonist muscle's OMNs. Larger eccentricities also require more relaxation in the antagonist muscle; however, the same equation, with the direction of θ reversed, provides for relaxation of the antagonist OMNs as well.

2. Coding of Eye Velocity by Oculomotor Neurons

For the eye to be rotated at any appreciable velocity, its extraocular innervation must also overcome the viscous drag of the orbit, which is approximately proportional to rotational velocity. Indeed, OMNs have significantly higher or lower spike rates when crossing a given eye position as a function of tracking direction and velocity. Thus, a velocity factor must be added to the OMN equation, yielding $R_{\mathrm{OMN}}=A\cdot(\theta-\theta_0)+B\cdot(d\theta/dt)$. A typical value for velocity factor B is 1 Hz/deg/sec. The magnitude of this factor, like the threshold position (θ_0) and position factor (A), varies across the motor pool of each muscle; however, all OMNs require such a velocity term, as well as a position term, in their rate equation.

This basic equation holds for all OMNs during all types of eye movements. The velocity term has its most dramatic effect when the eyes move via saccades or quick phases (Fig. 4, top) because such rapid eye movements can approach 1000°/sec. Likewise, the nonlinear stipulation that $0 \leq R_{\mathrm{OMN}} \leq 500$ is very important in relation to rapid eye movements; in fact, most OMNs are saturated at their maximal rate during large saccades in their on direction, and most are briefly silenced during saccades in their off direction.

3. Causality

The rate equation for OMNs is instructive, but it is causally backwards because the data are from recordings in alert monkeys making natural eye movements, not from direct experimental manipulation of eyeballs. Presumably, the brain arrives at a goal for θ and $d\theta/dt$ and then provides appropriate rates for all OMNs, including the few from which neurophysiologists record. Consequently, the extraocular muscles contract and the eye moves to position θ at velocity $d\theta/dt$. In other words, the causal equation is $\theta(t)=F(R_{\mathrm{OMN}}(t-\tau))$, where τ is the $\sim$5-msec delay from OMN spikes to actual eye movement. The substantive questions are (i) how an appropriate eye position and velocity are reckoned by a hierarchy of oculomotor structures distributed across the brain and

then (ii) how the appropriate R_{OMN} for effecting that position and velocity are computed.

E. How the Eyes Are Coordinated and Constrained

The complexity of controlling 12 extraocular muscles is dramatically mitigated by several "laws" that reduce the degrees of freedom for controlling the eyes. These laws are implemented in the oculomotor anatomy near the output stage, and thus nearly all eye movements obey them.

1. Descartes–Sherrington's Law of Reciprocal Innervation

This law asserts that the six eye muscles of each eye act as three agonist/antagonist pairs (as discussed earlier), thus reducing the oculomotor system's degrees of freedom from 12 (the total number of extraocular muscles) to 6. Reciprocal innervation was implicit in the R_{OMN} equation because the agonist and antagonist muscles of each pair lie in the same plane with reversed sign.

2. Hering's Law of Motor Correspondence

Hering's law halves the oculomotor systems degrees of freedom from 6 to 3 because it asserts that the two eyes act in unison (conjugately) for all eye movements excepting vergence. Thus, the brain operates as if there is only one eye, a cyclopean retina. This motor correspondence is accomplished by yoking pairs of muscles in the two eyes. For example, the lateral rectus of the right eye and the medial rectus of the left eye comprise a yoked pair. They receive the same commands and both effect a rightward movement.

The brain stem circuits for both reciprocal innervation and motor correspondence in horizontal eye movements (medial and lateral rectus) are shown in Fig. 5 for rapid eye movements, such as saccades, and in Fig. 6 for smooth/slow eye movements, such as VOR. The yoking for horizontal movements is straightforward: Signals entering the right abducens nucleus not only innervate motor neurons of the right eye's lateral rectus but also innervate interneurons within the abducens nucleus. These interneurons relay this signal, via the medial longitudinal fasciculus, to the contralateral oculomotor nucleus, synapsing on OMNs of the left eye's medial rectus. For reciprocal innervation, an inhibitory copy of the command signal is routed to the contralateral abducens, which relaxes the contralateral lateral rectus and, by the yoking circuit, relaxes the ipsilateral medial rectus as well.

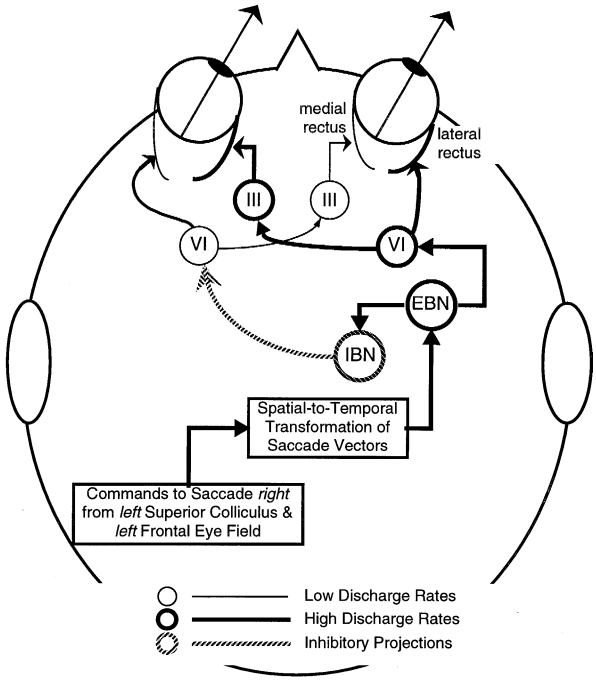


Figure 5 Neural circuit for a conjugate rightward eye movement. Making a conjugate eye movement, as in this figure, invokes activity in the diagrammed neural circuit that implements both Hering's law (motor correspondence) and Descartes–Sherrington's law (reciprocal innervation). For a rightward eye movement, the yoked agonist muscles contracted are the right lateral rectus and left medial rectus, and the antagonist muscles relaxed/inhibited are the left lateral rectus and right medial rectus. In the example shown, a signal to make a rightward saccade originates in the left hemisphere (cortex and colliculus) and is transmitted via the excitatory burst neurons (EBN) to the motoneurons of the right lateral rectus in n. VI (abducens). In order to also contract the yoked muscle in the left eye (the medial rectus), and thereby make the movement conjugate, a set of non-motor neurons in n. VI relay the movement signal to contralateral n. III (oculomotor). Reciprocal innervation (inhibition of the two antagonist muscles) is accomplished by inhibitory burst neurons (IBN) that project to the left n. VI and thereby both directly inhibit motoneurons of the left lateral rectus and indirectly diminish contraction in the right medial rectus via the n. VI to n. III projection. Similar circuits accomplish motor correspondence and reciprocal innervation for the other extraocular muscles in order to carry out conjugate vertical and torsional movements.

For the remaining eight muscles, yoking is less obvious. Each superior oblique and the opposite inferior rectus constitute one yoked pair, and each inferior oblique and opposite superior rectus another.

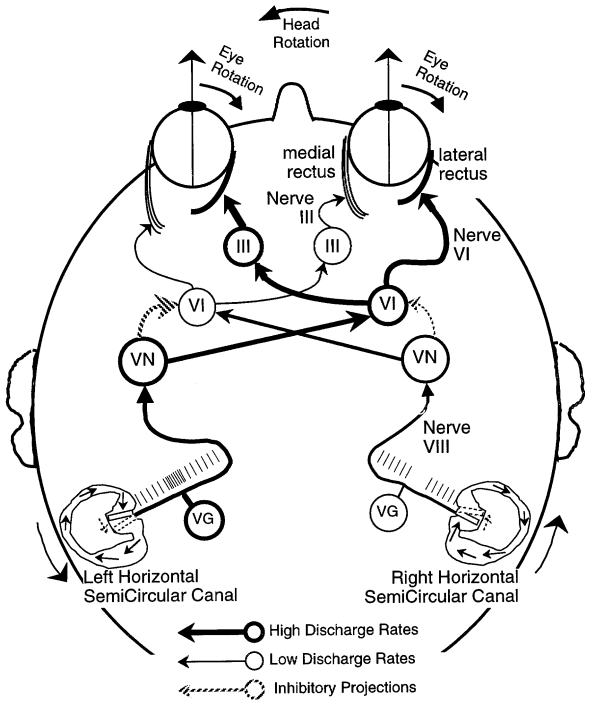


Figure 6 Neural circuit for the horizontal vestibuloocular reflex (VOR). The diagram depicts the rightward horizontal eye movements that compensate for leftward head rotation. The basic connectivity shown here serves all conjugate rightward eye movement (e.g., Fig. 5), regardless of the origin of the eye movement command signal. For the VOR, this signal originates in the semicircular canals and is communicated, via bipolar neurons in the vestibular ganglion (VG), to the vestibular nuclei (VN). The VN sends an excitatory projection to the contralateral oculomotor neurons in n. VI (abducens) and also sends an inhibitory projection to ipsilateral n. VI. The consequences of the increased activity from the left horizontal canal are reinforced by the decreased activity from the right horizontal canal, illustrating the push–pull operation of the canal pairs. An additional excitatory pathway (not illustrated) from the VN directly to the medial rectus motoneurons in ipsilateral n. III (oculomotor) gives the medial rectus its three-neuron VOR drive in addition to its four-neuron VOR drive via the contralateral n. VI.

These pairings reflect the true rotational axes of these muscles and their relation to the vestibular canals (Fig. 7).

3. Listing's and Donders' Laws and the Loss of Torsion

In primates torsional movements are severely constrained. For example, the human eye torts only a few degrees in response to a large sideways tilt (technically a "roll") of the head, whereas the rabbit eye torts up to 70°. The absence of significant torsional movement has an evolutionary advantage for stereoscopic vision.

Ocular torsion is classically summarized by the laws of Donders and Listing. Donders' law asserts that each eye direction (the combination of a particular azimuth and a particular elevation) has a unique torsion always associated with it, regardless of the sequence of eye movements used to achieve that particular azimuth/elevation. This has important implications. First, although it takes three coordinates to specify a general rotation, Donders' law means that the eye has only 2 degrees of freedom, and thus only two angles (azimuth and elevation) need be specified. This is analogous to navigation; longitude and latitude suffice, even though locations on the ocean surface have three coordinates in space. Second, whereas 3D rotations are not commutative, Donders' law stipulates that human eye movements effectively are commutative.

Listing's law asserts Donders' law and more because it specifies the particular torsion associated with any final eye direction, whereas Donders' law only ensured that this torsion is unique. To find the torsional position that the eye will always assume at a particular azimuth and elevation, Listing's law instructs to find the axis in the frontal plane through the center of the eye (Listing's plane) such that a single rotation about that axis takes the eye from the primary position to the particular azimuth and elevation in question. This single rotation will obtain the orientation that the eye will always have at that azimuth and elevation, no matter the sequence of eye movements actually used to get there. This means that the human eye has zero torsion following purely horizontal or purely vertical movements from the primary position (these are called "secondary" positions). However, for a typical oblique direction (a "tertiary" eye position) the eye appears to have undergone a torsional rotation because real-world vertical lines no longer fall along the vertical meridian of the retina when fixated. This is often called a "false torsion" because the eye did not have to make a true torsional movement about the line of sight but, rather, only a single rotation about an axis in Listing's plane, which is orthogonal to the torsional axis.

The role of the extraocular muscles in controlling ocular torsion is complex. The medial and lateral rectus have minimal torsional movements; however, as noted earlier, the four remaining muscles (inferior oblique, superior oblique, superior rectus, and inferior rectus) all have both vertical and torsional actions ("mixed" actions). Consequently, as Ewald Hering stated, elevation of gaze is effected "only by the cooperation of the superior rectus and the inferior

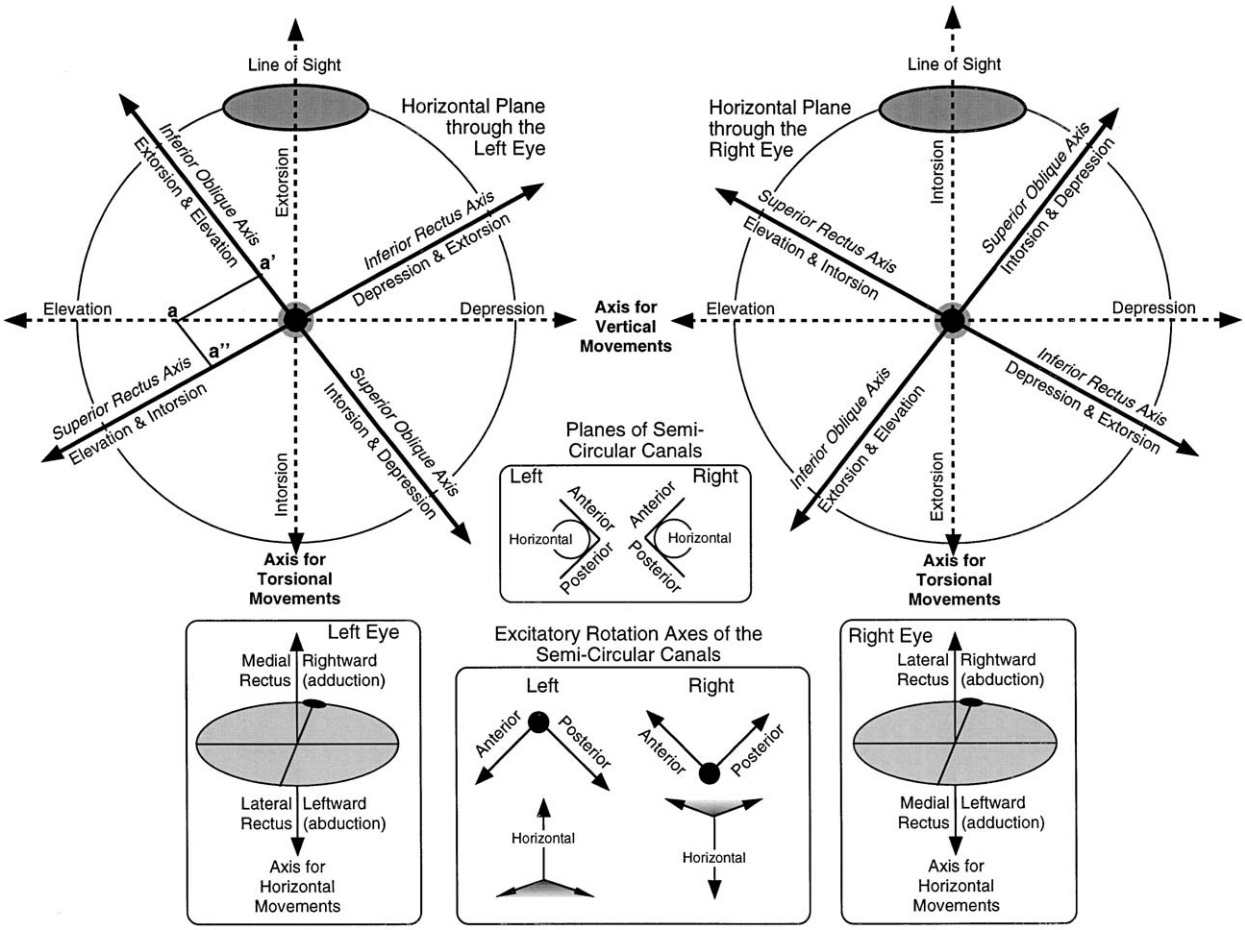


Figure 7 Rotational axes of the eye muscles and canals. (Top, left and right) The axis of rotation for all extraocular muscles, except for the medial and lateral rectus (bottom, left and right). For each muscle, the direction, or sign, of the eye rotation is given by the left-hand rule: Point the thumb along the arrow and the curve of the (slightly closed) fingers indicates the eye rotation. The rotation axes of the six semicircular canals (bottom middle) are perpendicular to the canal planes depicted above (upper middle) and show (by the left-hand rule) the eye rotation evoked by stimulation of individual canals. It is easy to see which muscle axis in each eye best aligns with each canal's axis; those muscles receive the three-neuron excitatory projection from that canal. Although in the primate, the eyes have migrated from the side of the head to the front, the canals are little changed, and the primate's extraocular muscles have maintained their rotation axes in correspondence with the canals axes. The figure also graphically depicts the rationale of yoking of the inferior oblique with the contralateral superior rectus and the superior oblique with the contralateral inferior rectus. The angular difference in both cases is far less than the difference between the left and right obliques axes or the left and right superior–inferior rectus axes. This figure was inspired by Fig. 51 in Hering (1942). His axis values, which were used here, have the axis of the superior–inferior rectus as being 29° off the axis for purely vertical movements and the axis of the superior–inferior obliques as being 38° off the torsional movement axis. The [a, a′, a″] parallelogram (top, left) is his geometrical construction of a pure elevation movement (with no torsion) from simultaneous contractions of the inferior oblique and superior rectus: Add the rotation axis vectors of these two muscles to obtain the axis for combined rotation.

oblique." Likewise, depression is effected by "cooperation of the inferior rectus with the superior oblique." This is analogous to lifting a barbell: If both arms lift with the same force, then the barbell will remain level, whereas if one arm exerts too much force then the barbell will tilt (tort) as it rises. Likewise, when the superior oblique and the inferior rectus are cocontracted, their downward forces are additive, but an extorsional force from increased tension in the superior oblique is canceled by an intorsional force from increased tension in the inferior oblique. Furthermore, the increased tension of the superior oblique and inferior rectus is always reciprocated by decreased tension of the inferior oblique and superior rectus, respectively. This results in a decrease in upward force, but with decreases in extorsion and intorsion canceling.

Thus, the eyes can move downward (or upward) with very little torsion, even though all four muscles involved have significant torsional actions.

Donders' and Listing's laws are not absolute or inherent in the passive properties of the eye. Large departures accompany convergence, head tilt, sleep, and certain neural pathologies. Torsional nystagmus, a clear violation of Donders' and Listing's laws, can be obtained with electrical stimulation of particular midbrain nuclei or individual canals and can also be a consequence of central disease.

IV. NEURAL CIRCUITS FOR EYE MOVEMENTS

A. Neural Circuitry of the Image-Stabilization System

The brain circuitry underlying image stabilization is organized around a vestibular apparatus in the temporal bone that can sense head rotations in any direction. The basic VOR is augmented by the OKR and several other specialized circuits, located in the brain stem and midline cerebellum, to become a very sophisticated system for minimizing retinal slip. These additional circuits are schematized in Fig. 8 and described next.

1. Circuit 1: The Three-Neuron Vestibuloocular Arc

How the brain fashions image-stabilization eye movements based on the vestibular sense has been extensively studied. Of prime importance is the vestibular sensory organ and the three-neuron arc composed of neurons in the vestibular ganglion, the vestibular nuclei, and the oculomotor nuclei.

On each side of the head within the labyrinth of the inner ear lie three semicircular canals filled with endolymph. Sensory receptor cells project hairs into a gelatinous mass, the cupula, that extends across the canal. When the head moves the bony canals move with it, but the endolymph lags behind, bending the cupula and the hairs embedded in it. Deforming the hairs hyperpolarizes their receptor cells for one direction of head rotation and depolarizes them for the opposite direction, and the resting discharge rate of bipolar neurons of the vestibular ganglion that innervate the receptor cells is modulated up or down by the receptor polarization (Fig. 6).

The central course of the bipolar cells' axons is the vestibular component of nerve VIII. They synapse in the vestibular nuclei, located in the medulla. Each canal projects to both the medial and the superior vestibular nucleus, and some vestibular afferents go directly to the cerebellum.

The three semicircular canals (horizontal, anterior, and posterior) lie in three different planes at approximately right angles to each other so that rotations of the head about any axis can be sensed. The planes of the three semicircular canals approximately correspond to the planes of action of the three pairs of extraocular muscles (Fig. 7), and the three-neuron pathway connects each canal to the extraocular muscles that move the eye in the canal's plane. This pathway was illustrated for the horizontal plane in Fig. 6. Horizontal head rotation excites the three-neuron pathway leading from the horizontal canal to the medial and lateral rectus, thus effecting an appropriate horizontal counterrotation of the eyes. Each canal also has a three-neuron inhibitory pathway to the corresponding antagonist muscles via inhibitory relay neurons in the vestibular nucleus. There are left and right sets of canals, and the symmetry of the VOR circuitry effectively uses the difference of the opposing signals from the left–right canal pair in each plane to increase the sensitivity of the VOR response, as shown in Fig. 6 for the horizontal canals.

The same principles hold for the anterior and posterior canals and the remaining four extraocular muscles, but in a less straightforward manner. The anterior and posterior canals lie in approximately vertical planes that are midway between the sagittal and coronal planes (Fig. 7). Thus, they do not correspond to the x–y–z axes of eye rotations described previously. Each anterior canal is paired with the contralateral posterior canal because they are in approximately the same plane and have the opposite responses to rotations in that plane. Also, each anterior–posterior pair controls a yoked pair of extraocular muscles that comes closest to its plane. For example, as can be deduced from Fig. 7, the excitatory three-neuron pathway of each anterior canal leads to the ipsilateral superior rectus muscle and the contralateral inferior oblique.

2. Circuit 2. Position Holding Mechanism (the "Neural Integrator")

The first problem for the basic three-neuron VOR circuit occurs when the head stops moving after a brief

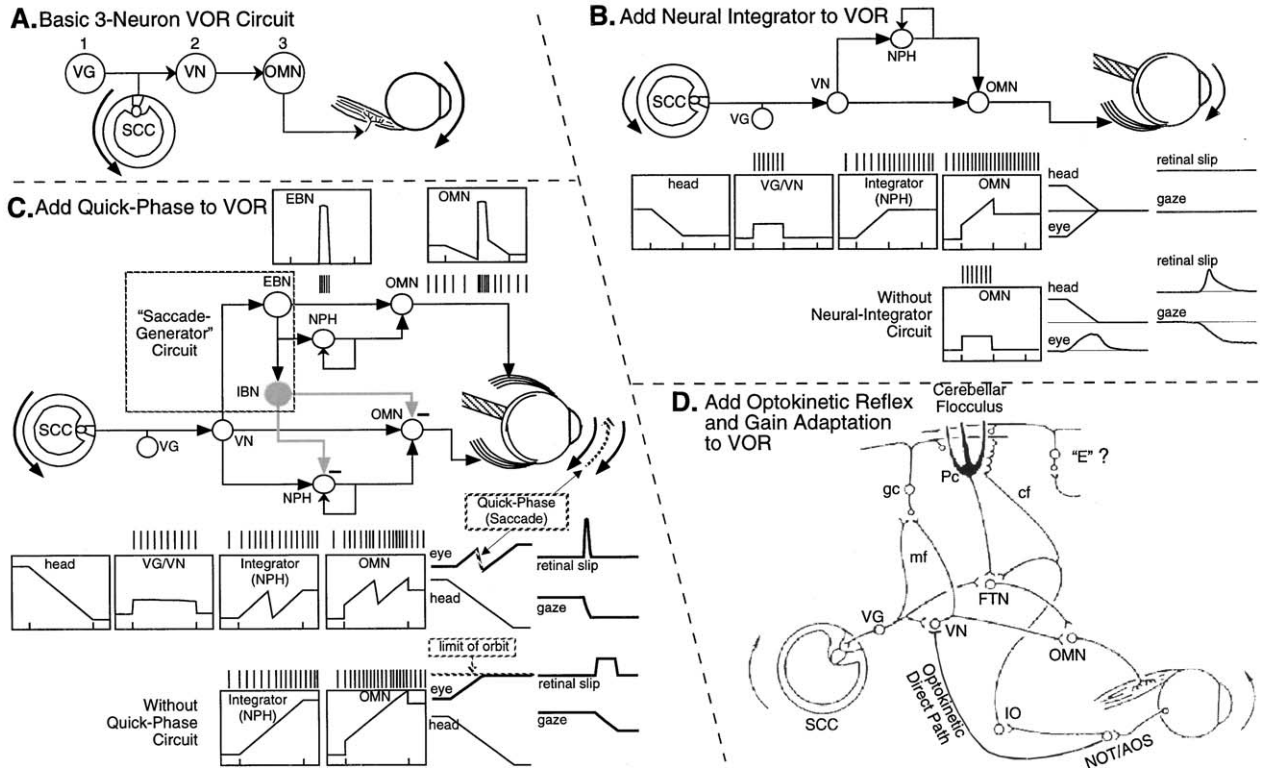


Figure 8 Neural circuits for image stabilization. (A) Basic three-neuron VOR circuit. Head rotations transduced by the semicircular canal (SCC) activate a three-neuron circuit that rotates the eye in the opposite direction. The circuit is as follows: neurons in the vestibular ganglion (VG) to neurons of the vestibular nuclei (VN) to oculomotor neurons (OMN) in one of the three oculomotor nuclei. (B) The VOR with the addition of a neural integrator. (Top) The neural integrator for horizontal eye movements is located in the n. prepositus hypoglossi (NPH) (and in the medial vestibular nucleus, which is not illustrated). Its integration function is symbolized by the recurrent connection of the NPH onto itself. (Middle) The plots show idealized spike rate functions for the components of this circuit, with idealized spikes above the plot boxes. With the neural integrator, retinal slip is minimal as gaze is held steady during and after the head movement. (Bottom) Without the integrator, the OMN does not cancel the elastic forces of the eye; thus, the eye slides back to the center of its orbit, both during and after the head movement, resulting in considerable retinal slip. (C) The quick-phase addition to the VOR. The quick-phase addition is symbolized by the inclusion of the saccade generator circuit with its excitatory (EBN) and inhibitory burst neurons (IBN) that enable the eye to be rapidly "reset" to the central position. Again, the plot boxes and spikes are hypothetical neural responses of the components of this circuit. (D) Addition of the optokinetic reflex (OKR) and a circuit for VOR gain adaptation. Visual motion signals from the nucleus of the optic tract (NOT) and the accessory optic system (AOS) go directly to the VN to add an OKR to the image-stabilization system. This OK signal is also sent to the inferior olivary nucleus (IO), and then to the cerebellar Purkinje cells (Pc) via their climbing fibers (cf). Likewise, the canal signal is relayed to the cerebellum's floccular Pc on mossy fibers (mf) that are axon collaterals of vestibular neurons via granule cells (gc) and their parallel T fibers. These inputs enable the flocculus to adaptively adjust VOR gain via its direct projections to the VN; the VN cells that receive this inhibitory projection are termed floccular target neurons (FTN). Pcs are also thought to receive a copy of the eye velocity signal, hence "E?" (adapted with permission from A. E. Luebke and D. A. Robinson, *Exp. Brain Res.* **98,** 379–390, 1994. Copyright © 1994 by Springer-Verlag).

turn. The canal signal quickly returns to baseline when the head stops moving because the cupula is no longer distended. However, if the innervation of the extraocular muscles also returns to its baseline level, then the elastic forces of the orbit would pull the eye back to the central position, thus negating the benefits of the VOR because vision would be blurred throughout this prolonged return.

Thus, for stable vision, the eye should not only stop rotating when the head stops turning but also must be held stationary at its current position and not slip back toward the central position. This is elegantly accomplished by the brain stem's neural integrator, a hypothesis of David Robinson, who reasoned that a command for eye position could be obtained by integrating eye velocity commands. Robinson's integrator now has a specific location, a neural signature (tonic activity coding static eye position), and distinctive neurological consequences of injury to it (e.g., gaze-evoked nystagmus).

The stabilization system, based on the VOR with the addition of the neural integrator, is shown in Fig. 8B, with the neural integrator represented by a local recurrent connection. Such a local recurrent connection was perhaps the earliest neural mechanism proposed for a short-term memory by Rafael Lorente de Nó and others. As the VOR "eye velocity signal" moves the eye, the integrator output grows, continually signaling where the eye should be held both during the VOR movement and after it. Thus, the neural integrator serves to exactly overcome the elastic force that accompanies increasingly eccentric eye rotations. Referring back to the equation $R_{OMN}=A\cdot(\theta-\theta_0)+B\cdot(d\theta/dt)$, the head velocity signal from the vestibular sensory nucleus is fed to the oculomotor nuclei as a command for eye velocity ($B\cdot(d\theta/dt)$) [where B is the synaptic strength] and the neural integrator obtains its eye position signal by integrating the eye velocity command, $d\theta/dt$, to obtain the $A\cdot(\theta-\theta_0)$ term for the OMN. This scheme is used for all types of eye movements: The command signal is formulated as a desired eye velocity (the velocity command hypothesis), and a common neural integrator integrates velocity commands of all types to continuously maintain the eye position signal.

3. Circuit 3. Quick Phases and Nystagmus

What if a large head rotation is made, or there are successive turns in the same direction? The eye can move only about 40–60° in any direction from its central position in the orbit. Thus, the VOR/neural integrator circuit will quickly drive the eye to the extreme of the orbit and the retinal image will begin to slip. Furthermore, in extreme orbital positions the eye's visual field is partially occluded.

Nature's solution is to periodically reset the eye back to the central position (usually slightly past it) during prolonged or large head rotations. Moving back to center results in considerable blur; however, the vertebrate solution is for the reset mechanism to recenter the eye as fast as possible in order to minimize the duration of the temporary blindness due to the high-velocity smearing of the retinal image. This requires near-maximal contraction of the agonist muscles coupled with complete relaxation of the antagonist muscles in order to produce eye movements approaching 1000°/sec and directed opposite the compensatory vestibuloocular movement.

The resets are obvious when prolonged rotation occurs, the resulting to-and-fro pattern of movement is called nystagmus (Fig. 9). In vestibular nystagmus, fast movements alternate with slow ones, yielding a characteristic sawtooth waveform. The slow phase is the compensatory reflex, and the rapid resets are called the quick phase. The quick-phase circuit is indicated by the addition of "burster" neurons in Fig. 8C. The excitatory burst neurons (EBNs), with their high spiking rate, project to the OMNs to realize the quick-phase movements. Actually, these EBNs are the output neurons of the "saccade generator" circuit, so labeled because it is appropriated by the foveation system to effect voluntary saccadic eye movements that are completely independent of the VOR. Moreover, the saccade generator is usually studied in the context of voluntary saccadic eye movements, in part because rotating subjects about different axes to achieve vestibular nystagmus can be a daunting experiment. Therefore, the saccade generator is further elaborated later in conjunction with voluntary saccades.

4. Circuit 4. The Optokinetic Reflex Complements the VOR

The VOR handles brief head movements very well, but it has trouble with prolonged rotation. This is because, with continued rotation of the head, the endolymph in the semicircular canals begins to catch up with the movement of the head, the cupula returns to its resting position, and hence the vestibular afferents return to their baseline rate, falsely indicating that the head is no longer rotating. Consequently, for continued rotation in the dark (or with the eyes closed) vestibular nystagmus will gradually slow down and completely stop after 30–60 sec. One solution to this problem of vestibular transducer adaptation is to let retinal slip help in driving this compensatory reflex. This is OKR, which is added to the basic VOR circuit in Fig. 8D. The principal sources of this retinal slip signal are two lesser known targets of the optic nerve: the nucleus of the optic tract (NOT) and the accessory optic system (AOS). Cells in these brain stem nuclei are tonically driven by the movement of large patterned visual stimuli, the best stimulus for eliciting OKR, with different neurons tuned to different directions of motion. In primates, the responses of these cells depends not only on their direct retinal inputs but also on pathways that involve neocortex, especially primary visual cortex and extrastriate motion areas. The direct and indirect projections of AOS/NOT to the vestibular nuclei complete the OKR circuit, which means that neurons in the vestibular nuclei also respond to visual (OK) stimuli. Moreover, despite

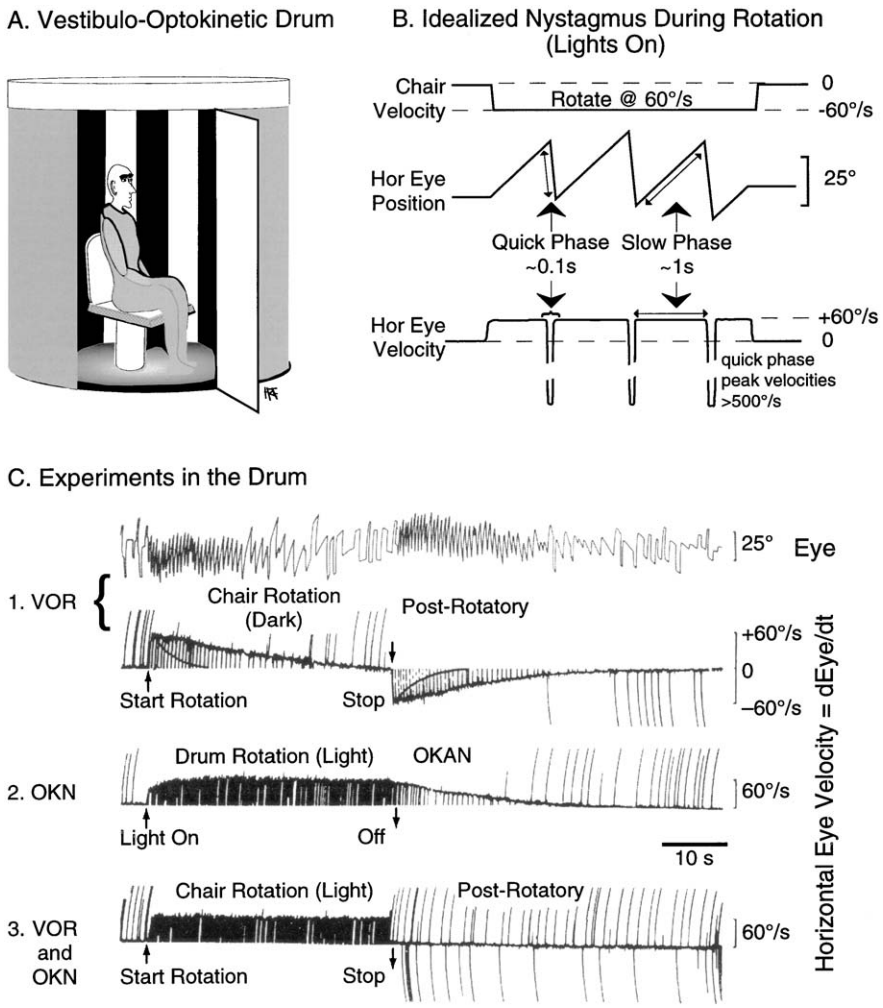


Figure 9 The optokinetic reflex and nystagmus. (A) A depiction of a rotating chair and drum (with striped inner walls) used for testing the vestibuloocular (VOR) and optokinetic reflexes (OKRs), together and separately. Rotation of the drum (surround) alone, with illumination, elicits the OKR in a stationary subject. Rotation of the chair in the dark elicits the VOR. Rotation of the chair in the light activates both reflexes. (B) Idealized nystagmus during chair rotation with full illumination. (Top) The chair velocity plot indicates leftward rotation. (Middle) The horizontal (HOR) eye position record shows slow-phase movements, directed opposite the chair rotation, which compensate for the chair rotation, combined with quick-phase movements in the direction of rotation that periodically reset the eyes toward their central positions in the orbit. (Bottom) The "envelope" of the velocity trace also shows the rightward direction of slow-phase motion, which is quite distinct from the brief, high-velocity leftward-moving quick phase of the nystagmus. (C) Experimental nystagmus data (1) (Top trace) Horizontal eye movement record showing vestibuloocular nystagmus elicited by chair rotation in the dark. (Bottom trace) Eye velocity record from the same trial (see Fig. 9B). The velocity envelope shows that slow-phase rightward movements are elicited with a very short latency after rotation begins but taper off as rotation continues (adaptation). Note that the postrotatory response is in the opposite direction. The exponential dark lines, in both the adaptation and the postrotatory epochs plot the theoretical strength of the vestibular canal signal, which adapts much faster than the slow-phase velocity adapts. (2) Eye velocity during drum rotation in the light. Optokinetic nystagmus (OKN) refers to the combination of the slow-phase OKR response and the quick-phase resets. The OKN velocity envelope during rotation is similar to that of the VOR except that (i) there is no diminution in the response to ongoing rotation in the light and (ii) there is a more gradual rise time to the maximal slow-phase velocity. Also notice that the optokinetic after nystagmus (OKAN) is in the same direction as the OKN, whereas the postrotatory response was opposite the VOR. (**3**) Chair rotation in the light elicits both VOR and OKN responses. There is a quick rise in slow-phase velocity because of the VOR, no adaptation because of the OKN, and no postrotatory nystagmus because the OKAN cancels the postrotatory nystagmus from the canals (Fig. 9C adapted with permission from T. Raphan, V. Matsuo, and B. Cohen, *Exp. Brain Res.* **35,** 229, 1979. Copyright © 1979 by Springer-Verlag).

the addition of many refinements to the VOR, and the addition of new types of eye movements, the vestibular nuclei remain the principal gateway to the oculomotor nuclei.

Since the OKR is indefatigable, why not dispense with the VOR and base image-stabilization eye movements solely on the OKR? The reason is that the OKR is much slower to react than the VOR. OKR latency is 50–100 msec (reflecting the slow pace of visual processing), whereas VOR latency is $\sim$10 msec. To demonstrate that OKR alone is poor, rotate the head left and right while reading. The text is legible because image stability is provided by the VOR. Now move the book left and right with the head still; the page will be blurred despite the OKR. This test hints at another reason the VOR is needed: Once a visual image is moving very fast across the retina, then it is blurred and hence is a poor stimulus for engaging the OKR in order to decrease its retinal slip and blur.

5. Circuit 5. Velocity Storage

VOR adaptation during prolonged rotation in the dark reflects the adaptation of the vestibular transducer in the semicircular canals. However, closer inspection of this phenomenon reveals that the vestibular apparatus adapts rapidly, with a time constant of $\sim$6 sec, but the reflexive eye movement adapts with a much longer time constant (15–30 sec) (Fig. 9C.1). The reason is that a central neural process stores up the vestibular signal, thereby providing a more veridical representation of head motion than is provided by the raw VIIIth nerve signal.

This velocity-storage circuit stores optokinetic as well as the vestibular velocity (because they are merged in the vestibular nucleus). Indeed, the OKR does not stop immediately with the cessation of OK stimulation but continues, in the same direction, for several seconds after the lights go off (Fig. 9C.2). This continuation is called optokinetic afternystagmus (OKAN). Thus, the OKR is like charging a battery: It takes several seconds of stimulation for the OKR to reach its asymptote velocity (Fig. 9C.2) and to ensure robust OKAN after the optokinetic stimulation ceases.

This velocity storage underlies another aspect of the symbiosis between the VOR and the OKR regarding cessation of prolonged head rotations. If the VOR is adapted when rotation stops, the canal endolymph continues to move and produces a strong VOR in the opposite direction, called "postrotatory nystagmus" (Fig. 9C.1). It is accompanied by a loss of balance and vertigo and lasts several seconds. However, postrotatory nystagmus is best demonstrated by rotation in the dark so there can be no optokinetic stimulation. Indeed, there is usually no problem after rotation in the light (Fig. 9C.3) because OKAN serves to cancel postrotatory nystagmus.

6. Circuit 6. Gain Adjustment Mechanism for the VOR

Whereas the OKR is a closed-loop system in that the response cancels its own input signal, the VOR is an open-loop system because the counterrotation of the eye has no direct effect on the semicircular canal head-rotation signal. Also, because it is an open-loop system, the VOR reflex strength is critical. Ideally, VOR gain is exactly 1.0; however, it would seem too much for the genome to so completely specify the VOR circuitry so as to yield this ideal gain. Instead, an adaptation circuit continually adjusts VOR reflex strength so as to minimize the retinal image slippage that accompanies head movements. For example, wearing 2 $\times$ magnifying goggles for a few days will drastically increase VOR gain because 2 $\times$ goggles require that the VOR gain be 2.0 to cancel retinal slip when the head moves. In real life, less drastic but nevertheless critical adjustments of VOR strength need to be made—for example, when the head changes size during development, when people don spectacles (which more modestly magnify or minify the visual world), or whenever vestibular hair cells die or extraocular muscles weaken because of old age or other factors.

The crucial brain structure for VOR gain adaptation is the cerebellum, specifically the flocculonodular lobe that is interconnected with the vestibular nuclei that mediate the VOR (Fig. 8D). Monkeys without a cerebellum still have a robust VOR; however, VOR gain does not adapt in response to experimental goggles and other manipulations that induce adaptation in normal subjects. Other structures critical for VOR gain adaptation include the NOT and AOS, which provide an optokinetic signal not only directly to the vestibular nuclei but also to cerebellar cortex via the dorsal cap of the inferior olive. The overall goal of the gain-adjustment circuit (Fig. 8D) is to minimize the optokinetic signal that the cerebellum "sees"; the better the VOR works, the less the OKR is needed. Thus, the optokinetic signal serves two important purposes: It effects the OKR to assist the VOR and it provides the error signal for continually fine-tuning VOR gain.

B. Neural Circuitry of the Foveation System

The image-stabilization system (ISS) functions to ensure that the visual image as a whole is stationary on the retina. This suffices for many species, but with the evolution of the fovea came the need to deliberately shift the direction of gaze, independent of the ISS, in order to purposively foveate different stimuli. Because the foveal specialization covers only $\sim 1/10{,}000$ of the visual field, the eyes must be moved accurately. Furthermore, they must be moved intelligently to selected targets because systematic scanning of the whole visual field (like radar or television) would require hours and mostly be a waste of time. Therefore, the foveation system (FS) orchestrates eye movements that provide the fovea with an endless succession of the most important, informative, and pleasing visual stimuli.

As mentioned earlier, the primate FS uses three distinct eye movements: saccades, smooth pursuit, and fixation. These eye movements are voluntary in that they are controlled by, or made in the service of, the subject's choice to foveate a particular stimulus. In contrast, the ISS reflexively responds to the aggregate of vestibular and visual motion sensations and requires only a state of wakeful alertness.

Although choosing to foveate a particular stimulus is a voluntary action, doing so automatically activates the appropriate foveation subsystem(s). For example, after choosing to fixate a particular stationary target, (i) if the target moves quickly to a new location (e.g., a "step" motion), then a saccadic movement will be made in order to foveate it at its new location; (ii) if the target begins to smoothly move (e.g., "ramp" motion), then smooth-pursuit movements will match the target velocity and thereby both maintain foveation and reduce retinal blur; and (iii) if the target remains stationary, then fixation continues.

The cerebral cortex has a large role in these foveating eye movements. This seems obvious considering that occipital, temporal, and parietal neocortex are all heavily engaged in visual processing. Nor is involvement of the frontal lobe surprising, because the FS concerns voluntary movement. Although saccadic and smooth-pursuit eye movements are represented in separate cortical areas that have different subcortical projections, it is still likely that a common network of cortical areas mediates the target choice decisions for the FS as a whole.

Despite their functional differences, the FS and ISS engage the same basic brain stem neural circuits to effect eye movements. More precisely, the FS engages OMNs only indirectly via the ISS circuits: Voluntary saccadic eye movements exploit the quick-phase generating circuitry in the midbrain and pontine reticular formation and smooth-pursuit eye movements target the vestibular nuclei where the vestibulooptokinetic velocity commands are assimilated. By engaging these established motor (as opposed to sensory) aspects of the ISS circuitry, foveating eye movements automatically achieve proper connectivity to the neural integrator in order to overcome elastic forces in the orbit and hold the eye steady wherever it has moved. Note that an important implication of this strategy is that the FS, like the ISS, ultimately formulates both its fast and slow eye movement objectives as eye velocity commands.

Usually the FS and the ISS work together; for example, while walking through a garden, saccades might foveate different flowers, with the VOR holding the image of each flower stationary on the retina despite movements of the head while walking. However, it is easy to pit the systems against each other, and in such situations the FS usually prevails. For example, given a small stationary spot in front of a background of moving stripes, one can elect to fixate the spot and thereby subdue the OKR ordinarily evoked by a large moving pattern. Conversely, given a moving spot on a stationary background of stripes, one can choose to pursue the spot smoothly across the stripes and the ensuing retinal slip of the pattern, caused by the pursuit eye movements, would be ignored. Finally, the FS can even suppress vestibular signals to move the eyes. For example, by looking at an object fixed relative to the head (e.g., the brim of a cap) and then rotating the head (like in Fig. 10), one can keep the brim fixated and thus keep the eye stationary in its orbit, despite the vestibular signals that indicate head rotation.

C. Neural Circuitry of Voluntary Saccades

The principal eye movement of the FS is the voluntary saccade. Humans average about two saccades/sec while awake, and thus most of the day is spent in brief fixations of different parts of the visual world, continually interrupted by saccades. This incessant visuomotor activity, the processing of foveal visual data during a fixation, as well as the planning and execution of the next saccade, occupies much of the human brain, as exemplified by the expansive zones of

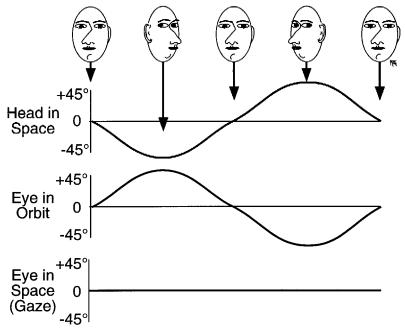


Figure 10 Retinal image stabilization during head rotation. As the head rotates about its vertical axis, the eyes reflexively rotate in the opposite direction to stabilize gaze, the direction of the eye in space. The two sinusoidal plots show the angular direction of the head in space (top) and the eye in its orbit (middle). The flat summation of these changes (bottom) signifies a steady gaze despite movements of the head.

cerebral activation revealed by functional magnetic resources imaging during simple and complex visuomotor tasks.

A few behavioral facts about saccades must be presented before considering their neural circuitry. First, saccades are open loop (i.e., ballistic and preprogrammed). As Gerald Westheimer stated, "once initiated, saccadic movements complete their predetermined course and cannot be modified or countermanded." In fact, 50–100 msec prior to its start, the saccade generally cannot be canceled or redirected on the basis of new sensory information. Second, saccades have long reaction times. Saccadic latency, the time from the appearance of an unpredictable visual target to the start of the movement, is 100–400 msec (200 msec being typical). Third, saccades have a long refractory period. It is usually difficult to initiate a second saccade for 100–200 ms after the previous saccadic movement ends. Fourth, saccades are highly stereotyped movements. Thus, the saccadic waveform and its parameters (duration, velocity, etc.) are almost completely determined by the dimensions of the movement vector being programmed. Finally, although saccades to foveate a stimulus can seem reflexive, and have been termed the visual grasp reflex, one can make accurate and advantageous predictive saccades that anticipate where a stimulus will appear. Thus, saccades can be guided by our experience, memories, guesses, purposes, and strategies as well as by overt visual stimulation.

1. The Saccade Generator

The first four properties mentioned previously reflect the fact that all saccades are ultimately "generated" or programmed in a very rigid, mechanistic fashion by a specialized circuit in the brain stem. This final common pathway for all rapid eye movements, called the saccade generator, is a network of several distinctive types of neurons embedded within the reticular formation near the oculomotor nuclei. This network originally evolved to generate the quick-phase movements of vestibulooptokinetic nystagmus. Thus, voluntary saccades of primates are accomplished by the relatively new FS triggering this ISS circuitry. If the saccade generator is damaged, then all rapid eye movements, both reflexive and voluntary, are disabled.

To accomplish a saccadic eye movement, the OMNs of the agonist muscle(s) need a special waveform of innervation: a pulse followed by a step. The pulse is a brief period of spiking at a very high rate, which is used to move the eye very quickly to its new location. The step is the new level of tonic discharge needed to hold the eye at its new location in the orbit. This pulse–step innervation is inherent in the OMN equation given earlier, and is illustrated in Fig. 11. Not shown is the fact that the pulse briefly but completely silences most OMNs of the antagonist muscles via inhibitory interneurons termed inhibitory burst neurons. This brief silence is also inherent in the basic OMN equation reviewed previously.

The pulse is provided by the principal output neuron of the saccade generator—the excitatory burst neuron (EBN). EBNs discharge with a high-frequency burst of spikes in conjunction with all saccadic eye movements and are silent between saccades. Their bursts begin $\sim$12 msec before the eye starts to move (they have also been called short-lead bursters) and end just before the eye completes the saccade. Horizontal EBNs are located in the caudal portion of the paramedian pontine reticular formation, conveniently near the abducens nucleus. Their spike counts during each burst determine the horizontal displacement of the saccade (in the ipsilateral direction). Vertical EBNs are located in the rostral interstitial nucleus of the medial longitudinal fasciculus, with separate sets of neurons for the upward and downward components. Other EBNs here also represent torsional saccadic eye movements.

The tonic "step" change in the OMN innervation is provided by the neural integrator that adds the pulse (EBN spikes) to its current value and then maintains that new value. Thus, when each saccade is made, the

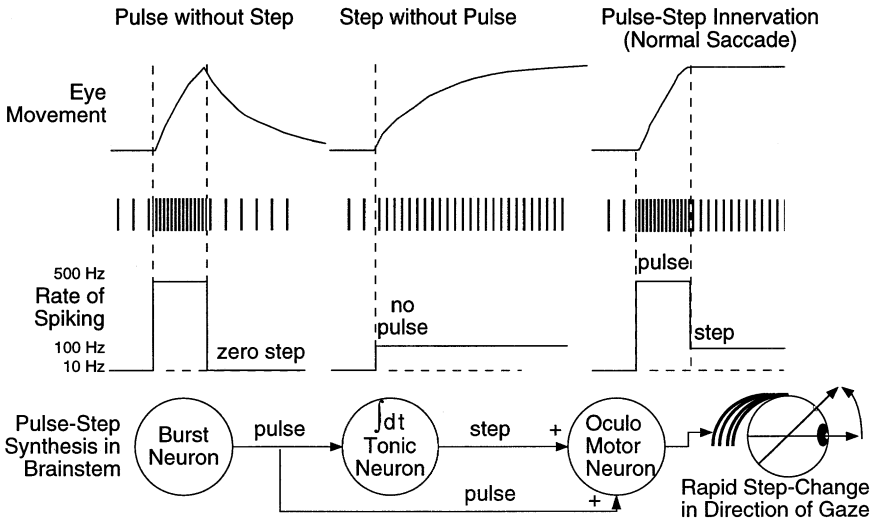


Figure 11 Pulse-step innervation of oculomotor neurons. A schematic showing eye position (top) and spike rates of burst and tonic neurons (middle) for real and hypothetical conditions producing abnormal and normal saccades. (Left) Given pulse activity but no tonic activity (step) to defend the new eye position, the eye slides back to its starting position (e.g., gaze-evoked nystagmus). (Middle) In the hypothetical case of a damaged pulse generator, the tonic/step activity alone causes very slow eye movements that "glide" exponentially to a target. (Bottom) Circuit diagram for pulse–step innervation of oculomotor neurons and the generation of saccades. The excitatory burst neurons are activated by signals from higher levels (e.g., the superior colliculus and frontal eye field). The activity of tonic neurons reflects ongoing integration of all eye velocity commands by neural integrator circuits in the pons and midbrain. Oculomotor neurons sum these phasic and tonic inputs to produce a high-velocity saccade and then to hold this new eye position.

neural integrator quickly "steps" from its old presaccadic level of activity to a new postsaccadic level. In contrast, the response of the neural integrator to a low-velocity long-duration signal from the VOR command is a "ramp" change in output rate. This short-term memory mechanism has been successfully modeled with recurrent neural networks and thus is schematized as a single recurrent feedback connection (Fig. 8C).

Several cell types antecedent to the EBNs have been identified. One of the most remarkable types is the omnipause neuron (OPN). The tonic activity of these inhibitory (glycine) neurons has the critical job of keeping all EBNs totally silent during the intervals between saccades. If EBNs had even a low spontaneous rate when not bursting, then the neural integrator would generate random "walks" between saccades instead of providing steady fixations. Only when OPNs temporarily stop spiking (which they briefly do in conjunction with all saccades, regardless of saccade size or direction) are EBNs given an opportunity to discharge. In fact, electrical stimulation at the OPN locus (nucleus raphe interpositus along the midline of the pons) during a saccadic eye movement will immediately brake the eye and (prematurely) end the saccade, and continuous stimulation there prevents all rapid eye movements, but not slow eye movements, such as the VOR or smooth pursuit. Detailed consideration of OPNs and the other cell types that complete the saccade generator circuit are beyond the scope of this article.

2. Visually Guided Saccades and the Superior Colliculus

How can the brain stem's saccade generator be activated to produce useful saccades in the absence of head movements and VO/OK stimulation? The exemplar structure for this function is the superior colliculus (SC), which forms the roof of the midbrain. The SC is the mammalian version of the optic tectum, which is the vertebrate brain's prototype sensorimotor structure. The SC receives a strong projection directly from the retina as well as afferents from most other senses (auditory, somatosensory, etc.), and its efferents to the brain stem and spinal cord serve to orient the head and body toward localized sensory inputs. In the primate, and other mammals with a foveation system, a major role of the SC is to move visual stimuli appearing in the visual periphery into the foveal region

of the retina by triggering a saccadic eye movement of the appropriate size and direction. This automatic, visually guided foveating saccade has been called the visual grasp reflex.

The functional anatomy of the primate SC that affords such visually guided foveation is forthright (Fig. 12): Neurons in the superficial layer of the primate SC constitute a topographic map of the contralateral visual hemifield (Fig. 12B; interestingly, this is only true of primates—for all other vertebrates each SC represents the entire contralateral eye). Neurons in the deeper, intermediate layer of the primate SC provide a topographic map of all contralaterally directed saccade vectors. These sensory and

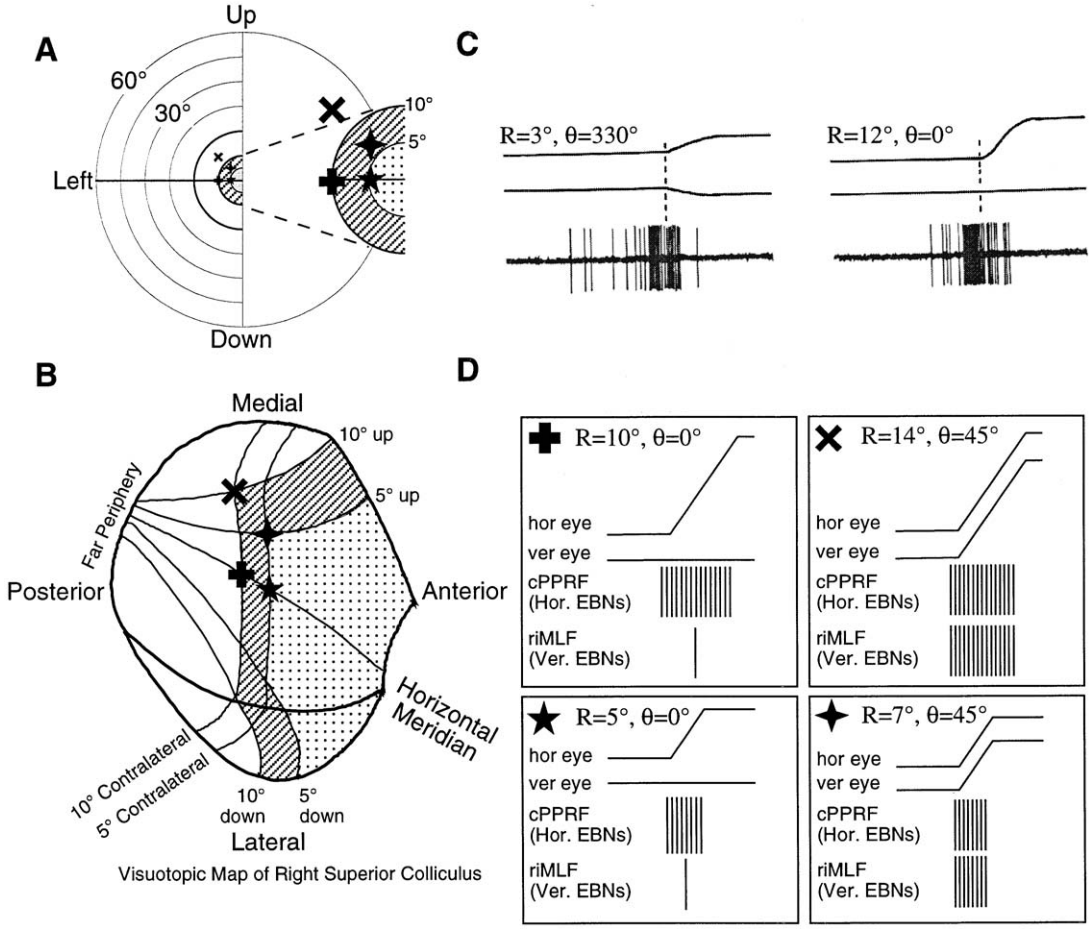


Figure 12 Spatial-to-temporal transformation of saccade vectors. Depiction of the necessity of a spatial-to-temporal transformation as saccade commands are relayed from the superior colliculus (SC) to the saccade generator, but the figure does not show how it is accomplished. (A) Spatial representation of four hypothetical visual stimuli. The stimuli are within the central-most 10° of the diagram and the dashed lines lead to enlargement of this region. (B) A visuotopic map of the SC showing the idealized representations of the visual stimuli depicted in A on the surface of the SC. Note that more than half the SC is used to map the central 10° (adapted with permission from M. Cynader and N. Berman, *J. Neurophysiol.* **35,** 187, 1972). (C) Saccade-related burst cells in the deeper SC layers code saccadic eye movements via their location in its spatial map of the contralateral visual hemifield. The top two traces are horizontal and vertical eye coordinates, and the bottom traces are spikes. These cells burst most robustly prior to a saccade matching their preferred vector (reprinted from D. L. Sparks, *Brain Res.* **156,** 1, copyright 1978, with permission from Elsevier Science). (D) Hypothetical response of the excitatory burst neurons (EBNs), the output stage of the saccade generator, to the visual stimuli shown in A and B. Top traces are eye position, and lower traces are the bursts of horizontal (Hor) EBNs [in the caudal paramedian pontine reticular formation (PPRF)] and vertical (Ver) EBNs [in the rostral interstitial nucleus of the medial longitudinal fasciculus (riMLF)]. Saccade dimensions are a quasilinear function of the number of spikes in the bursts. Since EBNs usually burst at near-maximal rates, their spike count is largely a function of burst duration, just as saccade size is largely a function of saccade duration. In contrast, the burst duration of the saccade-related bursters in the SC is independent of saccade size or duration.

motor maps are in perfect register for making foveating saccades. For example, if recording from a superficial layer neuron located near the center of the left SC, the visual response field center would be 10° right of the fixation point (cross in Fig. 12B). If the microelectrode were then advanced into the underlying intermediate layer, it would record a "saccade-related burst neuron" that responds optimally (i.e., has the most spikes in its burst) immediately prior to 10° rightward saccades (Fig. 12C), exactly the saccade needed to foveate visual stimuli appearing in the superficial (layer) neuron's visual response field (RF). Moreover, electrical stimulation there would yield a 10° rightward saccade as well.

3. Transformation from Spatial Code to Temporal Code

Saccades are clearly "spatially" coded in the SC in that the saccade vector is determined by which part of collicular "space" has active saccade-related bursters. In contrast, the output cells of the saccade generator, the EBNs, use a "temporal" code for saccade metrics: Saccade size is coded by the spike counts in their bursts (Fig. 12D) and not by which (or where) EBNs are spiking most. Exactly how this spatial-to-temporal transformation is carried out is unclear. Moreover, the cortical eye field signals need the same spatial-to-temporal transformation, and neither the cortex nor the colliculus projects directly to the EBNs. Another class of saccade generator neurons, the long-lead burst neurons (LLBNs), seem to be an intermediate stage in this transformation. LLBNs do receive direct projections from the SC and FEF. Moreover, individual LLBNs often prefer saccades with specific oblique directions and amplitudes, like SC and FEF neurons but unlike EBNs.

D. Cortical Eye Fields and Saccades

1. Cortical Pathways for Visually Guided Saccades

In primates, the cerebral cortex is a very important part of the saccadic circuitry. Although the SC seems to be the premier structure for visually guided saccades, it is not a critical component of saccade generation because the effects of SC lesions are largely transient, whereas the saccade structures downstream are vital. Thus, lesions of the saccade generator structures in the pontine and midbrain reticular formation permanently eliminate all rapid eye movements, including visually guided saccades, whereas SC lesions do not eliminate visually guided saccades at all. Following an acute period of visual neglect, there are few lasting oculomotor deficits following SC lesions in monkeys and man, with the most consistent lasting deficit being a modest increase in latency of visually guided saccades.

Visually guided saccades survive SC lesions because the primate also has an elaborate neocortical network for visually guided saccades. The connectivity diagram of Fig. 13 summarizes these cortical pathways and helps explain most effects of experimental lesions. For example, the sparing of visually guided saccades following SC lesions is mediated by FEF projections to the brain stem saccade generator; however, the SC normally provides a shorter path via its direct retinal projections, which explains the increase in saccade latency after SC damage. FEF lesions alone also spare visually guided saccades; however, FEF lesions combined with SC lesions eliminate most visually guided saccades. Thus, the SC and FEF provide parallel pathways for visual stimuli to activate the brain stem saccade generator for the purpose of accurate, foveating saccades. It is also the case that visually guided saccades are spared following lesions of primary visual cortex (V1) even though conscious awareness of the visual world is lost. However, combined V1 and SC lesions eliminate visually guided saccades most likely because, as Fig. 13 indicates, V1 removal eliminates most visual inputs to FEF and, hence, renders FEF incapable of triggering visually guided saccades.

2. Frontal Eye Field Anatomy and Physiology

David Ferrier discovered ($\sim$1875) that electrical stimulation in the frontal lobe of macaque monkeys deviated the eyes toward the contralateral side, and it was soon confirmed that many primate species, including man, have such an FEF. These electrically elicited eye movements are indistinguishable from naturally occurring saccadic eye movements, and each site in FEF yields saccades of a characteristic direction and amplitude, with the set of all possible contralaterally directed saccades represented in each hemisphere's FEF. The macaque FEF lies primarily in the anterior bank of the arcuate sulcus; the human FEF lies in the precentral sulcus, behind the middle frontal gyrus and in front of the hand representation in the precentral gyrus (Fig. 14).

FEF is not the only cortex specialized for eye movements. There is also the parietal eye field (PEF),

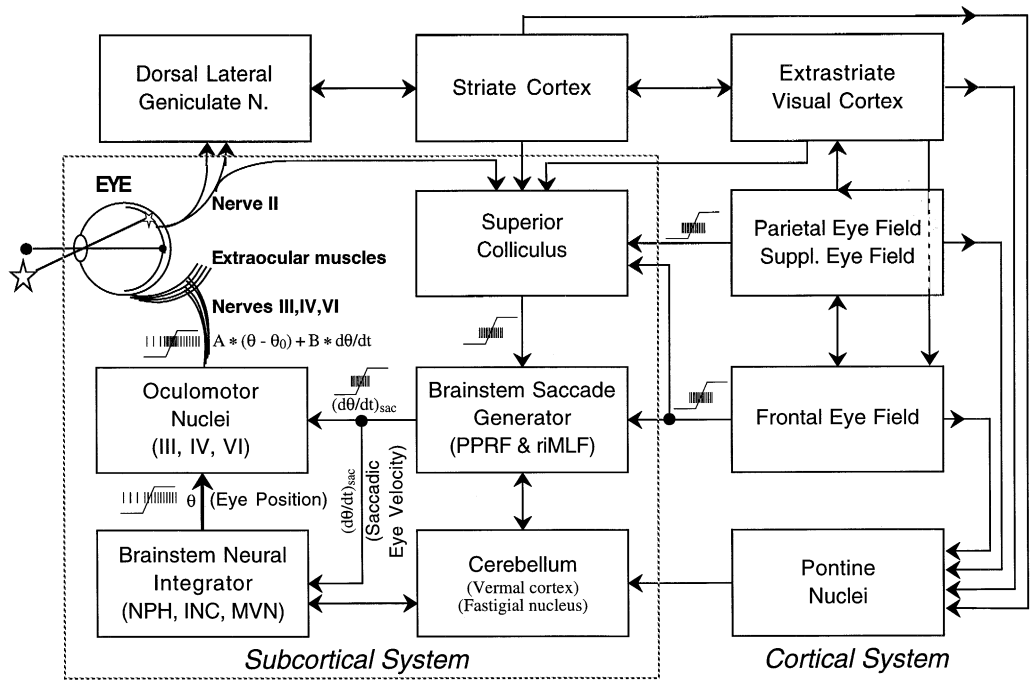


Figure 13 Pathways for visually guided saccades. Anatomical connections between cortical and subcortical structures involved in the control and generation of saccadic eye movements. The frontal eye field (FEF) receives visual information via pathways originating in the striate cortex, as does the parietal eye field (PEF) and the supplementary eye field (SEF). Notice that all these cortical areas project to the superior colliculus (SC), with FEF, PEF, and SEF all projecting primarily to its intermediate layers. In contrast, the superficial layers of SC receive direct visual projections from the retina and indirect visual projections from the striate and extrastriate cortices. The brain stem saccade generator is in the paramedian pontine reticular formation (PPRF) and in the rostral interstitial nucleus of the medial longitudinal fasciculus (riMLF). The brain stem neural integrator is in the medial vestibular nucleus (MVN), the adjacent nucleus prepositus hypoglossi (NPH), and the interstitial nucleus of Cajal (INC). Additional structures and pathways involved in saccades, such as the thalamus and basal ganglia, are omitted for simplicity.

located in the lateral bank of the intraparietal sulcus in the macaque, and the supplementary eye field (SEF), located in the frontal lobe near the midline. FEF, SEF, and PEF are reciprocally interconnected with each other; however, FEF seems to be the principal cortical eye field. FEF has the lowest thresholds for electrically elicited saccades, oculomotor behavior after FEF lesions is generally more impaired than after lesions of the other eye fields, and FEF is indispensable for visually guided saccades if the SC is damaged.

3. Effects of FEF Lesions

Gordon Holmes found that patients with frontal lesions had difficulty moving their eyes in response to verbal commands, even though they could follow visual objects and understood the verbal commands. In a 1938 lecture, he concluded that "the frontal centers make possible the turning of gaze in any desired direction and the exploration of space, but they also keep under control, or inhibit, reflexes that are not appropriate." In 1985, Daniel Guitton and colleagues used the antisaccade paradigm to demonstrate that "frontal lobe lesions in man cause difficulties in suppressing reflexive glances and in generating goal-directed saccades." In other words, their subjects could not launch saccades in the direction opposite the visual target (antisaccades), even though they understood that to be the task. Instead, they made inappropriate saccades toward the visual targets (prosaccades), exactly what they were instructed not to do. Similarly, subjects with FEF lesions have difficulty making memory saccades and in making predictive saccades to square-wave target motion. Thus, lesion studies show that FEF is important for "purposive" saccades, particularly when there is no visual target present to trigger the visual grasp reflex.

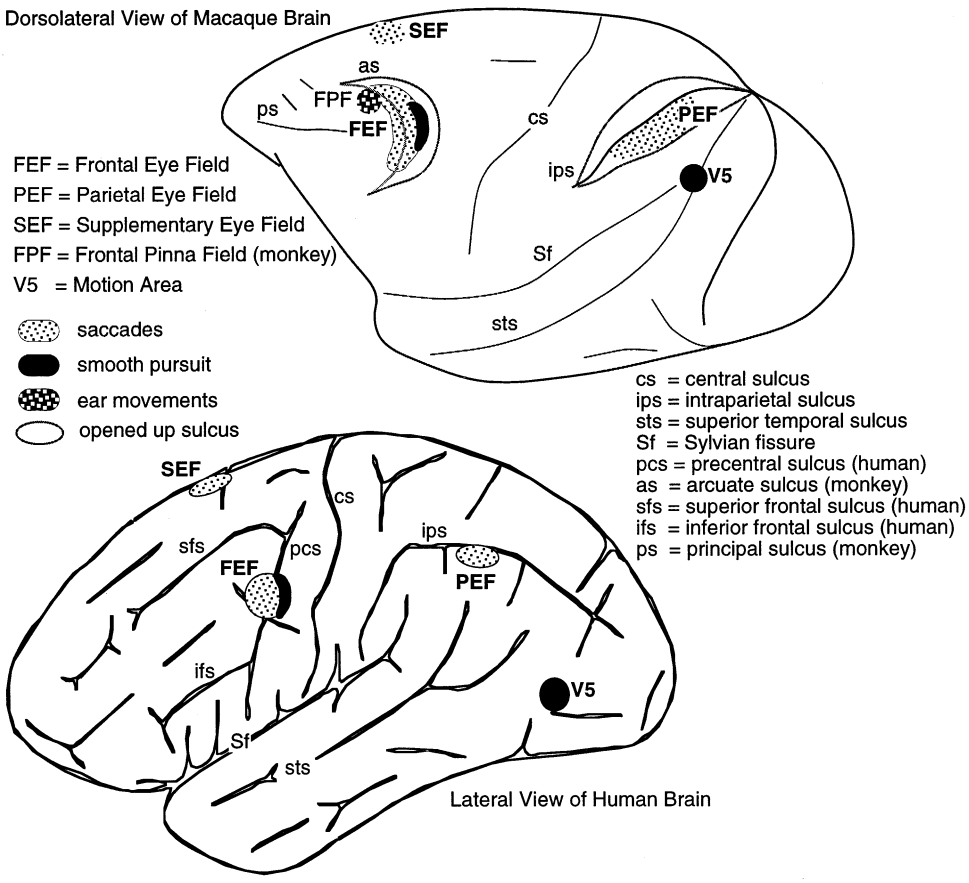


Figure 14 Cortex for eye movements in man and monkey. Cortical regions important for saccade and smooth-pursuit eye movements are highlighted on lateral views of a monkey brain (top) and human brain (bottom). In both monkey and man, FEF is in front of premotor cortex for the hand and neck and mostly lies within the sulcus marking the anterior limit of the precentral gyrus. In both species, the smooth-pursuit region of FEF is just posterior to the saccadic region of FEF. A dorsolateral view is used for the monkey brain in order to minimize distortion of the frontal lobe sulci.

4. FEF Bursts Precede All Types of Purposive Saccades

During the past three decades, single-neuron recordings in trained macaque monkeys have resulted in a detailed picture of FEF activity during all manner of voluntary oculomotor behavior. Presaccadic bursts are the signature activity of FEF and are manifest in more than 30% of FEF neurons. These bursts begin prior to saccade initiation, usually end sharply just after the saccade is completed, and are always tuned for particular saccade vectors, similar to saccade related bursters in the SC and vectorial LLBNs in the pons. Presaccadic bursts seem to constitute the FEF command to the saccade generator (both directly and through the SC), providing both an impetus to saccade and a saccade vector specification. Indeed, electrical stimulation through a recording microelectrode in FEF elicits natural-looking saccades that closely match the vector for the optimal presaccadic burst of nearby cells. Moreover, FEF-elicited saccades are very insistent and are still elicited with low currents even when subjects are intently fixating a stationary light.

a. Visually Guided Saccades FEF neurons have robust presaccadic bursts in conjunction with visually guided saccades. The average response of 51 FEF cells recorded in a single monkey during a "stable-target" type of saccade task is shown in Fig. 15A. This task was chosen to separate any phasic visual response to the appearance of the peripheral target from the presaccadic burst. This provides a baseline for assessing presaccadic bursts made without an overt target.

b. Memory Saccades Presaccadic bursts of FEF neurons in conjunction with saccades made to remembered targets are generally equivalent to bursts associated with visually guided saccades on the stable-target task. In this paradigm, as shown for a representative visuomovement neuron in Fig. 15B, a peripheral cue appears only briefly, and the saccade is made some time later and hence must be guided by a short-term memory of the cue.

c. Antisaccades FEF neurons also have robust presaccadic bursts in conjunction with antisaccades, as shown in Fig. 15C for another representative visuomovement neuron. The cell discharged preceding antisaccades into its visuomovement RF, even though the visual cue had been on the opposite side and thus not at all in the cell's RF.

d. Other Purposive Saccades FEF cells have also been demonstrated to reliably burst for some other types of purposive saccades (e.g., saccades made to the locations of sounds). It is interesting to speculate that FEF lesions might disrupt socially motivated saccades and that FEF cells would burst in conjunction with such saccades, but this has not been tested.

e. Spontaneous Saccades In contrast with most purposive saccades, FEF bursts are usually weaker in conjunction with spontaneous saccades made in the dark, presumably because such saccades are usually not purposive.

5. FEF Activities and Circuits

Just as the basic VOR reflex has a set of associated assisting circuits, a diverse set of functional activities and cortical circuits underlie FEF's programming of purposive saccades in the monkey. These serve to facilitate the generation of appropriate presaccadic bursts in diverse situations and paradigms.

a. Visual Activity More than half the neurons in FEF are visually responsive. Typically, they have large RFs centered in the contralateral hemifield and respond to the appearance of any stimulus within their RF, without much selectivity for color or form. Moreover, FEF visual responses do not require overt attention to the stimulus or the RF location or that the stimulus has functional significance to the monkey.

b. Alignment of Visual and Presaccadic Movement Fields Visuomovement FEF cells have both visual and presaccadic burst activities, and their visual RF generally corresponds with the optimal saccade vector for their burst (i.e., movement field) and also to the electrically elicited saccadic eye movement vector obtained at the cell's location. Thus, the FEF default is a foveating saccade. However, the presaccadic burst is independent of the location and/or the presence of RF stimulation as shown by the memory saccade and antisaccade tasks (Fig. 15B,C). Moreover, a minority of FEF cells are discordant with nonmatching, or even nonoverlapping, visual and movement fields.

c. Tonic Visual Activity The strongest aggregate visual response in FEF is to the initial appearance of visual targets, and many visual cells only respond to this appearance. However, other visual cells tonically respond as long as the target remains in their RF. Notice in Fig. 15A that the composite spike rate was elevated throughout the wait period (between the phasic visual response and the eventual presaccadic burst). Thus, FEF visual activity can guide saccades to old, "stable" visual targets as well as newly appearing targets.

d. Mnemonic Activity Usually tonic visual FEF activity is maintained even after the visual cue is extinguished, and this activity could provide a short-term memory of the visual cue location in the memory saccade test. Many individual FEF cells have robust mnemonic responses (Fig. 15B). However, when many FEF cells are averaged, the mnemonic signal is only a modest elevation of the overall spike rate, especially when compared to the size of the presaccadic burst. Tonic neural activity has a high metabolic cost; therefore, it is economical for overall tonic visual and tonic mnemonic activity to be minimal—just robust enough to inform spatially appropriate saccades whenever the go signal finally arrives.

e. Postsaccadic Activity Coding Executed Saccades (Efferent Copy) Postsaccadic activity in the FEF was first described by Emilio Bizzi, and $\sim$25% of FEF neurons are excited after particular saccadic eye movements. This postsaccadic activity seems to be an efferent copy of saccades actually executed because it reliably follows every saccade made into the cell's postsaccadic movement field, even spontaneous saccades made in the dark or rapid phases of nystagmus. A timely efferent copy of saccadic displacements, as coded by postsaccadic activity in FEF, is critical for several of the following circuits. Interestingly, many FEF cells with presaccadic (visual, movement, or

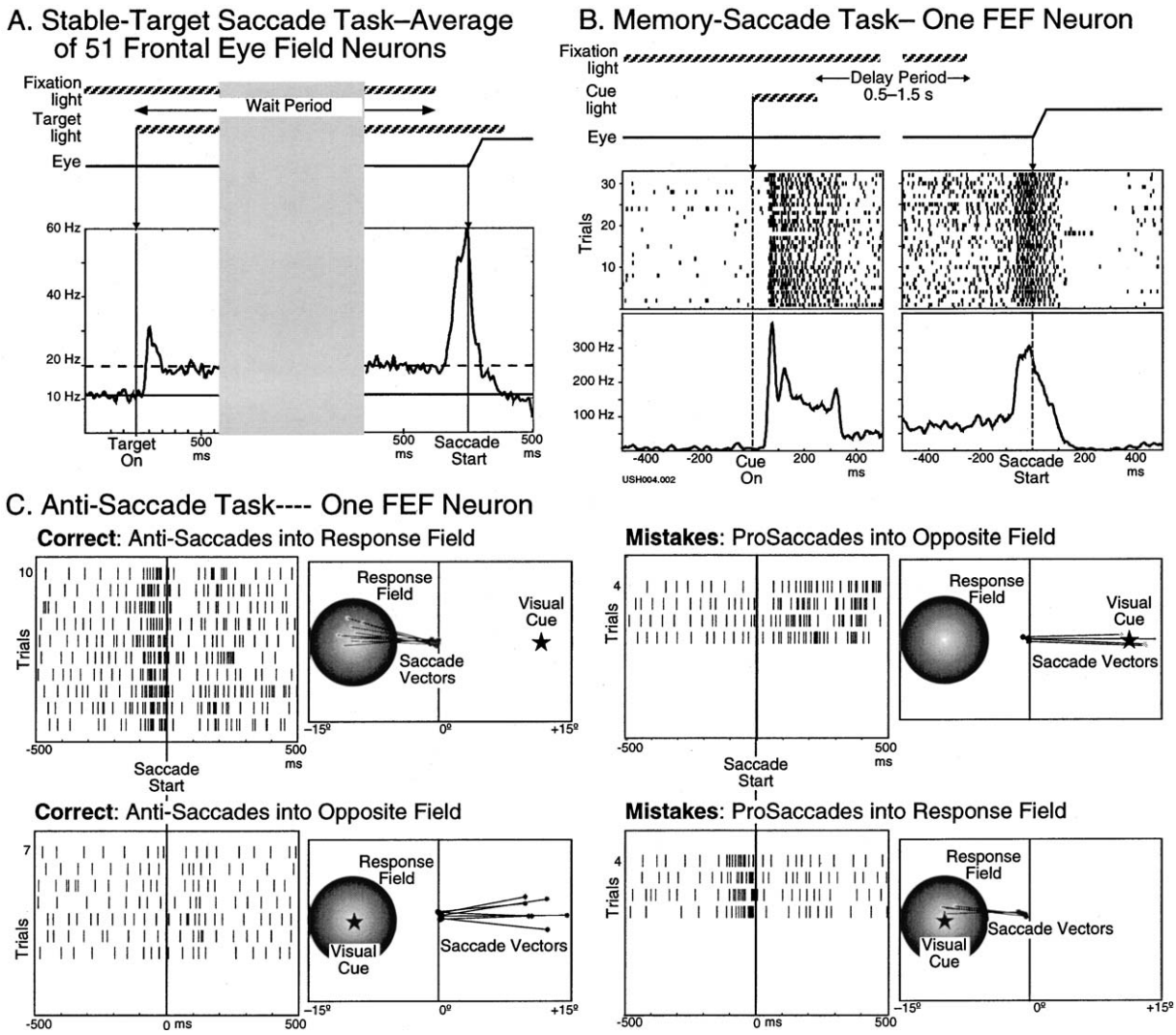


Figure 15 Frontal eye field activity for purposive saccades. (A) Aggregate activity of 51 FEF neurons recorded from one monkey during the stable-target type of visually guided saccade task. (Top) Task events: Shortly after the monkey fixates a central light, a peripheral target appears in the neuron's response field (RF). This target remains on for the remainder of the trial, but no saccade is permitted until the fixation light is extinguished at the end of the wait period. Thus, the target is a stable presence at the time of the saccade. (Bottom) The histogram aligned on the target appearance (left) shows the aggregate visual response. The histogram aligned on the saccade start (right) shows the large aggregate burst that starts just prior to the saccade. Thus, FEF activity manifests both visual (phasic and tonic) and movement activities (H. R. Friedman and C. J. Bruce, unpublished data). (B) Activity of a visuomovement FEF neuron from a second monkey tested on memory–saccade task. (Top) Task events: Shortly after the monkey fixates a central light, a peripheral cue briefly appears. Then, after a delay period, the fixation light is extinguished and the monkey must saccade to the location where the peripheral cue had been shown earlier. Thus, unlike in A, there is no visual target present at the time of the saccade. (Bottom) Rasters (by trial) and histograms of spike activity aligned on cue onset (left) show the burst of activity elicited by the visual stimulus and aligned on saccade onset (right) show the large burst of activity preceding the saccades. Note that the visual response has three components: a phasic high-rate burst to the initial appearance of the visual stimulus (with a latency of $\sim$50 msec), a robust tonic visual discharge while the cue remained on, and then, starting $\sim$50 msec after the cue was extinguished, a medium-level tonic mnemonic response that was maintained above the cell's baseline level of activity throughout the delay interval (M. S. Kraus, H. R. Friedman, and C. J. Bruce, unpublished data). (C) Activity of a visuomovement FEF neuron, from a third monkey, tested on an antisaccade task (memory version). As in B, the monkey must remember the position of a brief visual cue across a delay interval. However, unlike in B, the correct response (once the fixation light goes off) is a saccade to the location opposite to where the cue was shown. The visual RF of this cell was on the left when tested with conventional "prosaccade" tasks and its presaccadic bursts were maximal for leftward saccades. (Top, left) The cell discharged preceding leftward antisaccades, even though the visual cue had been in the right side and thus not in its RF. (Bottom, left) The neuron was silent before rightward antisaccades even though the visual cue had been in its RF. Interestingly, its bursts were completely predictive of erroneous prosaccades mistakenly made on some trials (right). (H. R. Friedman and C. J. Bruce, unpublished data).

both) activity also have postsaccadic activity for saccades directed opposite their presaccadic RF. This provides a mechanism for readily returning to the previous fixation (i.e., glances).

f. Suppression of Presaccadic Activities by Saccade Execution A striking aspect of FEF presaccadic activity of all types (e.g., anticipatory, visual, mnemonic, and movement) is that it quickly ceases upon the execution of a saccade into the RF. Notice in Fig. 15 that saccade execution actively suppresses both tonic visual (Fig. 15A) and mnemonic (Fig. 15B) activity as well as the presaccadic bursts.

This suppression could come from the postsaccadic coding of prior saccades (efferent copy) previously described. Such suppression is very important because visual or mnemonic activity coding a peripheral cue location becomes invalid once the monkey foveates the peripheral location. Without prompt suppression, persistent activity could lead to multiple triggering of the same saccade, much like the "staircase" of saccades evoked by continued electrical stimulation in FEF.

g. Fixation Status Signals (Tonic Foveation and Eye Position Activity) Some FEF cells provide tonic signals concerning the current fixation target rather than pertaining to possible saccade targets. One class is excited by fixation (foveal) stimulation; their activity could play a role in suppressing other saccade cells in FEF and elsewhere in the interest of maintaining fixation. Another class has the inverse activity, being suppressed by foveal stimulation and active thereafter, and thus signaling the extinction of the current fixation light. A small minority of cells tonically respond as a linear function of absolute eye position (e.g., elevation); they could be receiving an efferent copy from the common neural integrator, and some have foveal responses that are modulated by the current eye position.

h. Other Saccade-Related Activities and Responses Many FEF cells show anticipatory activity for predictable saccadic situations. This activity could decrease saccade latency by biasing FEF in favor of the predicted saccade dimensions. Sometimes, anticipatory activity is purely guessing. For example, in the "gap" paradigm, wherein the signal to saccade (e.g., fixation light off) precedes the peripheral target appearance by 100–200 msec, FEF neurons with presaccadic motor bursts will first discharge at an intermediate level in response to the fixation light extinction and then drastically accelerate or suppress their rate after the peripheral target comes in the neuron's RF or opposite it. Saccade latencies in the gap paradigm are typically shorter than in conventional saccade tasks and are termed express saccades.

Cells in and near the medial FEF have responses to sound. Their auditory RFs are partially remapped from a craniocentric to a retinocentric framework, which should facilitate FEF bursts known to precede aurally guided saccades. Moreover, there is a pinna movement region adjacent to the medial FEF of the monkey (Fig. 14).

Finally, saccades to moving targets are usually directed at a predicted target location, based on both retinal position and velocity. Visual and movement RFs of FEF neurons evidence this predictive process, indicating that target motion information is being utilized.

E. Neural Circuitry of Smooth Pursuit

1. Tracking with Pursuit and Saccades

Smooth-pursuit eye movements support scrutiny of objects moving in space by matching eye velocity to target velocity in order to both reduce retinal blur of the moving object and facilitate its continued foveation. Smooth pursuit occurs when the FS selects a moving target or when a previously selected stationary target starts to move. However, target selection for pursuit also activates the saccadic system; hence, moving targets are usually tracked with a combination of smooth pursuit and saccades, with these two eye movement systems operating independently but synergistically to track the same chosen target (Fig. 2). Their synergy reflects control by separate parameters of the target's trajectory. The principal impetus for smooth pursuit is target velocity (i.e., retinal slip), and the pursuit system continuously endeavors to eliminate retinal slip by matching eye velocity to target velocity. In contrast, the principal concern of the saccade system is target position (i.e., retinal error), and saccades are intermittently generated to eliminate retinal error by foveating the target.

However, this division of labor is not absolute. Smooth pursuit is modestly affected by positional errors: Ongoing pursuit accelerates in response to small retinal positional errors, and it is even possible to initiate smooth pursuit with an afterimage placed near the fovea (although eccentric afterimages are usually tracked with a succession of saccades). The pursuit system also responds to the rate of change in retinal slip

(i.e., acceleration). Thus, pursuit is a function of the zero-, first-, and second-order derivatives of the target's retinal image.

Conversely, the saccadic system attends to target velocity as well as location. Saccadic latency is shorter for targets moving centrifugally (away from the fixation point) and longer for targets moving centripetally. Moreover, saccades are usually directed to a predicted target location based on its position and velocity as acquired 100–200 msec before the saccadic movement starts.

Pursuit velocity ranges up to $\sim 100^{\circ}$/sec; however, pursuit gain (defined, like VO and OK gain, as eye velocity/target velocity) is generally poor for target velocities above 25°/sec. When the pursuit gain is low, the eye will persistently fall behind the target and frequent, large "catch-up" saccades will be made; however, if gain is high (~ 1.0), then only a few, small saccades may be needed.

Interestingly, low smooth-pursuit gain is the principal symptom of the eye tracking dysfunction (ETD) of schizophrenia. Subsequent research has shown a cluster of oculomotor impairments that covary across the schizophrenic patient population and are often also present in first-degree relatives of schizophrenic patients. On the basis of the ETD and other cognitive aspects of schizophrenia, it has been hypothesized that schizophrenia reflects diminished frontal lobe function in general, and that the ETD specifically reflects impaired function in both the saccadic and the smooth-pursuit regions of the FEF.

2. Smooth Pursuit versus Optokinetic Following

Smooth pursuit can be confused with the OKR because both provide smooth, nonsaccadic ocular following in response to visual motion. However, the OKR is an automatic response to motion of large parts of the visual world (e.g., watching a train go by), whereas pursuit is the voluntary tracking of a discrete moving object (e.g., a bird flying across the sky).

In typical situations, smooth pursuit is in direct opposition to the OKR (and often to the VOR as well), and we choose to maintain foveation of the pursued target stimulus and reduce its retinal slip at the expense of increasing retinal blur of the rest of the visual field. For example, pursuit will need to override the OKR whenever tracking a target moving across a patterned background because as soon as pursuit starts the background necessarily becomes a stimulus for OKR in the opposite direction. Similarly, during combined head–eye tracking of a moving stimulus, the combined vestibulooptokinetic reflex must be overcome.

3. The Neural Pathway for Pursuit

The major pathways of the smooth-pursuit system are shown in Fig. 16. On the sensory and decision-making side, smooth pursuit is very much a neocortical behavior. It begins with the high-quality visual map in V1 that sends visual motion information into area V5 and other immediate extrastriate areas. Motion information is relayed to several other areas in the parietal and temporal lobe and to the smooth-pursuit zone of FEF. These cortical areas relay their visual motion signals and commands to oculomotor parts of the cerebellum, principally by way of their projections to the dorsolateral and medial pontine nuclei.

On the efferent side of the pursuit pathway, smooth movements are created by engaging the brain stem substrate for the slow-phase, compensatory part of VOR, much as visually guided saccades are made by engaging the mechanism that generates the quick phases of vestibuloocular nystagmus. Specifically, pursuit uses the VOR adaptation circuit that was shown in Fig. 8D. The key neurons are the Purkinje

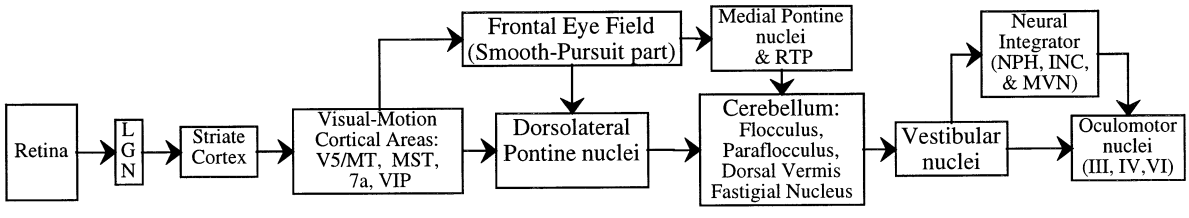


Figure 16 Pathways for smooth-pursuit eye movements. The flow-chart shows that the smooth-pursuit part of the frontal eye field receives visual motion information from extrastriate motion areas, e.g., from the middle temporal area (MT, V5) and the medial superior temporal area (MST) and from parietal regions such as 7a and the ventral intraparietal area (VIP). INC, interstitial nucleus of Cajal; LGN, lateral geniculate nucleus; MVN, medial vestibular nucleus; NPH, nucleus prepositus hypoglossi; RTP, nucleus reticularis tegmenti pontis.

cells in the floccular–nodular lobe and the paraflocculus regions of the cerebellar cortex. They carry a smooth-pursuit signal and effect pursuit by inhibiting vestibular nucleus neurons that in turn project to extraocular motoneurons. This pathway is both indirect, via the fastigial cerebellar deep nucleus, and direct from the Purkinje cells. In the context of the VOR, these projections from the cerebellum to the vestibular nuclei serve to adapt the VOR gain and offset. The smooth-pursuit system uses this pathway to temporarily create a vestibular imbalance in favor of desired pursuit direction, and thereby create a smooth eye movement. Cerebellar outputs may also aid pursuit by suppressing any opposing OKR evoked by the slip of the visual background once pursuit is underway as well as by suppressing any opposing VOR during combined head and eye pursuit.

4. Effects of Lesions in the Smooth-Pursuit Pathway

Neocortex is critical for smooth pursuit. Hemidecortication or large unilateral cortical lesions can cause a profound and permanent deficit of ipsilaterally directed pursuit in man and monkeys. Unilateral loss of V1 produces a permanent pandirectional loss of pursuit in response to motion in the contralateral hemifield. Discrete lesions in motion area V5 produce temporary pursuit deficits in restricted parts of the visual field (pursuit scotomas).

FEF lesions cause permanent ipsilateral pursuit deficits because each hemisphere's FEF is principally concerned with ipsilateral pursuit. Similar ipsiversive deficits, with profound reductions in both acceleration during pursuit initiation and steady-state pursuit velocity, follow lesions or temporary inactivation in the dorsolateral pontine nuclei. Finally, cerebellar lesions can cause severe and permanent pursuit deficits; in fact, smooth pursuit is the only eye movement type that is lost following cerebellectomy; other types of eye movements may be profoundly disturbed but are nevertheless still realized.

5. Sensory-to-Motor Transformation for Pursuit

Neurons throughout the pursuit pathway have been studied in the rhesus monkey. Similar to the saccadic system, neuronal activity in structures closer to the retina have obligatory responses to visual stimuli, but closer to the oculomotor nuclei the neural activity is more aligned to motor behavior. For example, neurons in V5 are highly selective for the direction of stimulus motion in their RFs. At the other end, the smooth eye velocity signal is elaborated in several pontocerebellar circuits, including the dorsolateral and the medial pontine nucleus and the nucleus reticularis tegmenti pontis, together with cerebellar pursuit areas (the flocculus, paraflocculus, and dorsal vermis). Many neurons in these structures respond as a function of pursuit velocity.

The critical step in this sensorimotor transformation underlying pursuit is provided by the neurons that respond to target motion, but only if it is selected for foveation/pursuit. The smooth-pursuit zone of FEF (Fig. 14) is a candidate site. It lies downstream from V5 and other areas that provide most of the visual motion information for smooth pursuit, but upstream from pontocerebellar and brain stem circuits that effect pursuit movements. Microstimulation in the pursuit FEF elicits smooth eye movements (usually ipsiversive), and many FEF pursuit cells discharge to the motion of a pursuit target over most of the visual field but discharge little to moving targets that are not pursued. An FEF pursuit cell is shown in Fig. 17; notice that it also discharged during predictive pursuit at the end of sinusoidal tracking and that reversible inactivation at this FEF site caused a immediate deficit in ipsilateral smooth pursuit.

F. Neural Circuitry of Fixation

When an interesting or important stimulus is foveated, it may be important to maintain its fixation and to temporarily suppress the generation of saccades to other stimuli. For example, during reading (Fig. 2), the average fixation time depends on the overall difficulty of the material, and individual fixation durations depend on the complexity of each word fixated. Thus, controlling fixation duration is an important component of eye movements.

The rostral end of the SC (adjacent to the representation of the smallest saccades) is specialized for fixation. Cells there project to the omnipauser region of the pons, and electrical stimulation at the rostral pole of the SC prevents saccades. Because most visually responsive neocortex projects strongly to the SC, this rostral fixation zone provides a pathway for cortical areas that process foveated stimuli to temporarily suppress saccades. Stimulation at some FEF sites can inhibit saccades as well, and foveally responsive FEF cells resemble cells at the SC rostral pole.

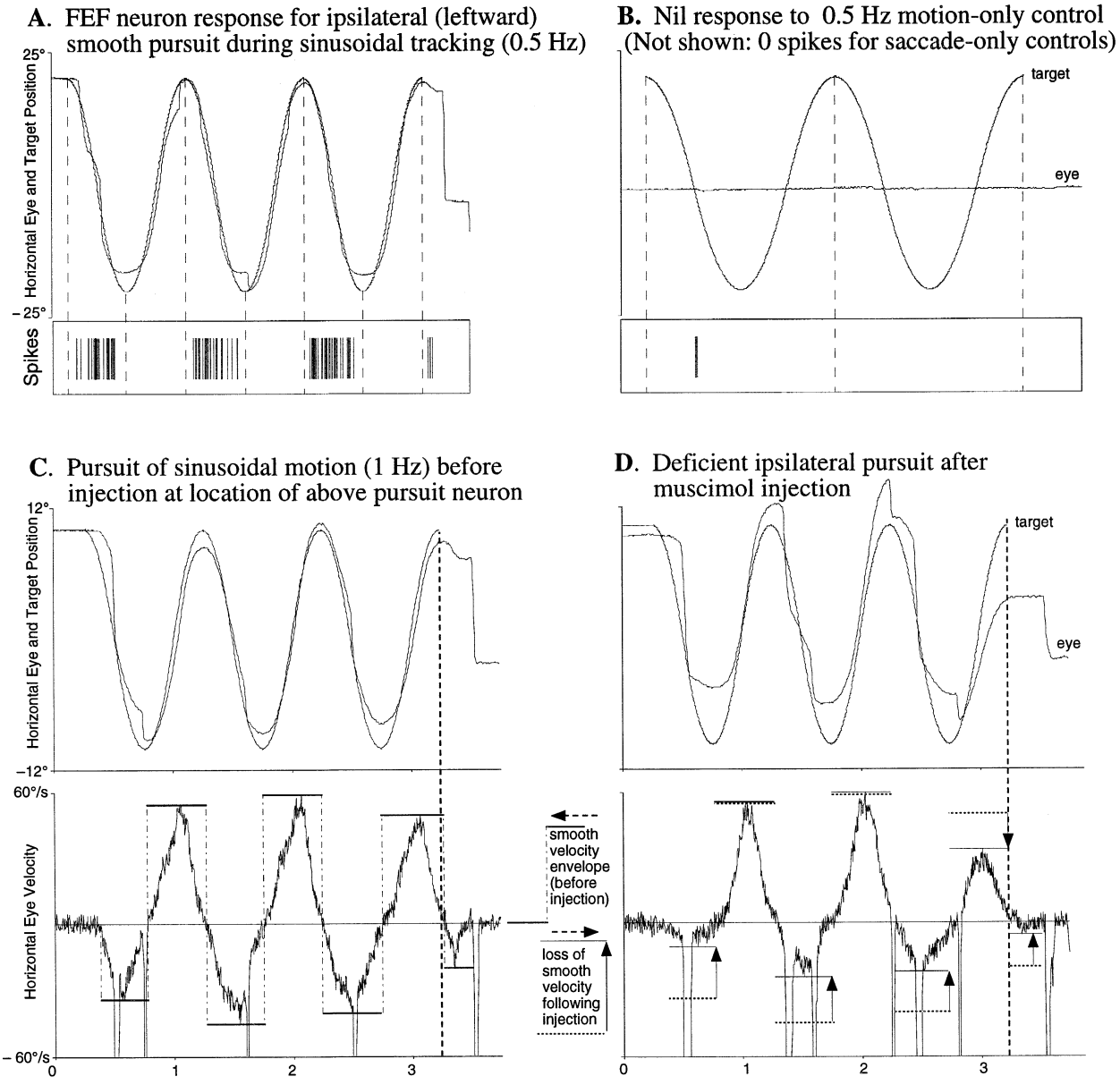


Figure 17 Activity of FEF pursuit neuron and pursuit deficits after FEF deactivation. (A) The pursuit neuron, recorded in the left hemisphere FEF, was active in conjunction with leftward (negative moving on the traces) smooth pursuit and responded throughout the leftward phase in all three cycles of sinusoidal motion (0.5 Hz, $\pm 20^{\circ}$A). Notice that it also discharged prior to leftward predictive pursuit that occurred at the end of sinusoidal tracking (see also Fig. 4). (B) The neuron responded very little to visual motion that was not pursued and was silent, except for three spikes on the initial cycle, when the monkey fixated a stationary point while viewing the same sinusoidally moving stimulus used in A. (C) Sinusoidal (1.0 Hz, $\pm 10^{\circ}$A) smooth pursuit of this monkey immediately before the injection. (D) Deficient smooth pursuit immediately following a small muscimol injection at the site of the pursuit neuron shown in A and B. Leftward smooth eye velocity was much less on every cycle, especially the first, and there was no longer predictive pursuit after the target was extinguished (adapted with permission from D. Shi, H. Friedman, and C. Bruce, *J. Neurophysiol.* **80,** 458–464, 1998).

G. Neural Circuitry of the Vergence-Accommodation System

The primate has frontally placed eyes with highly overlapping visual fields (i.e., a large binocular field). The two retinal images are combined (fused) in the cyclopean retina (effectively located in V1). Fusion, however, requires that both eyes look in approximately the same direction so that visual objects fall on corresponding points of the two retinae. Consequently, most eye movements are conjugate because both eyes move in synchrony during nystagmus, saccades, and pursuit in order to maintain fused vision and obtain stereopsis, the extraction of relative depth from slight differences in the images of the two eyes. As discussed earlier, conjugate movement is accomplished by equal innervation of yoked muscle pairs in the two eyes (Hering's law).

However, when the distance of the fixation object varies, adjustments are needed that involve converging or diverging the relative horizontal directions of the two eyes. Cells in a region immediately lateral to the oculomotor nucleus are active during such disjunctive eye movements. These cells bilaterally innervate the nearby OMNs of the medial rectus and thus add a convergence command to the conjugate eye movement signal that medial rectus OMNs receive from interneurons in the abducens nucleus.

The two principal cues for vergence movements are retinal disparity, which occurs when a fixation target is not on the fovea of both eyes, and retinal blur, which occurs when a target is not in focus (as well as when it has retinal slip). Because vergence is tightly linked to accommodation, either cue evokes both responses. The "near triad" is a popular term for the response to viewing a near object: (i) The lens is compressed to bring the object into focus (accommodation), (ii) the eyes converge to image the object on the fovea of both eyes, and (iii) the pupils constrict to increase the depth of field. As noted earlier, cranial nerve III also has the efferents (from the Edinger–Westphal nucleus) that effect the lens accommodation and pupillary reflex components of the triad.

Although the complete circuitry of vergence is not known, it is clear that cortical visual areas, especially V1, are critical for processing disparity and blur. Paul Gamlin finds that a small region of frontal cortex, located just anterior to the saccadic FEF in the monkey, has cells that respond specifically during vergence movements and electrical stimulation there can elicit vergence eye movements. Moreover, cells throughout FEF can be sensitive to the depth location (or binocular disparity) of visual targets as well as to their azimuth and elevation. This depth information may be critical in allowing changes in vergence to be made simultaneously with saccades to targets differing in depth as well as in visual direction. Vergence movements are generally very slow and can last nearly a second. This slowness reflects in part the difficult task of processing vergence cues; however, saccadic vergence is accomplished much faster than pure vergence movements. Vergence changes can also accompany smooth pursuit in depth, and vergence could even be considered as part of the FS rather than as a separate, third oculomotor system or function.

See Also the Following Articles

COLOR PROCESSING AND COLOR PROCESSING DISORDERS • EVOLUTION OF THE BRAIN • NEUROIMAGING • SPATIAL VISION • SUPERIOR COLLICULUS • VISION: BRAIN MECHANISMS • VISUAL SYSTEM DEVELOPMENT AND NEURAL ACTIVITY

Acknowledgments

We gratefully acknowledge the critical comments of Gregory B. Stanton and Micheal S. Kraus.

Suggested Reading

Becker, W. (1991). *Saccades*. In *Vision and Visual Dysfunction* (R. H. S. Carpenter, Ed.), Vol. 8, pp. 95–137. CRC Press, Boca Raton, FL.

Bruce, C. J. (1990). *Integration of sensory and motor signals for saccadic eye movements in the primate frontal eye fields*. In *Signal and Sense, Local and Global Order in Perceptual Maps* (G. M. Edelman, W. E. Gall, and W. M. Cowan, Eds.), pp. 261–314. Wiley–Liss, New York.

Carpenter, R. H. S. (1988). *Movements of the Eyes*. Pion, London.

Emery, N. J. (2000). The eyes have it: The neuroethology, function and evolution of social gaze. *Neurosci. Biobehav. Rev.* **24,** 581–604.

Goldberg, M. E. (2000). *The control of gaze*. In *Principles of Neural Science* (E. R. Kandel, J. H. Schwartz, and T. M. Jessell, Eds.), pp. 782–800. McGraw-Hill, New York.

Henn, V. (1992). Pathophysiology of rapid eye movements in the horizontal, vertical and torsional directions. *Baillieres Clin. Neurol.* **1,** 373–391.

Hering, E. (1942). *Spatial Sense and Movements of the Eye*. American Academy of Optometry, Baltimore, MD.

Holzman, P. S. (2000). Eye movements and the search for the essence of schizophrenia. *Brain Res. Rev.* **31,** 350–356.

Kowler, E. (1990). *The role of visual and cognitive processes in the control of eye movement*. In *Reviews of Oculomotor Research 4. Eye Movements and Their Role in Visual and Cognition* (E. Kowler, Ed.), pp. 1–70. Elsevier, New York.

Leigh, R. J., and Zee, D. S. (1999). *The Neurology of Eye Movements*. Oxford Univ. Press, New York.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychol. Bull.* **124,** 372–422.

Robinson, D. A. (1987). The windfalls of technology in the oculomotor system. *Proctor lecture. Invest. Ophthalmol. Vis. Sci.* **28,** 1912–1924.

Schall, J. D., and Thompson, K. G. (1999). Neural selection and control of visually guided eye movements. *Annu. Rev. Neurosci.* **22,** 241–259.

Vilis, T. (1997). Physiology of three-dimensional eye movements: saccades and vergence. In *Three-Dimensional Kinematics of Eye, Head, and Limb Movements* (M. Fetter, T. Haslwanter, H. Misslisch, and D. Tweed, Eds.), pp. 57–72. Harwood Academic Publishing, Amsterdam. (See also http://www.med.uwo.ca/neuroscience/vilis/courses.htm)

Wurtz, R. H. (1996). Vision for the control of movement. The Friedenwald Lecture. *Invest. Ophthalmol. Vis. Sci.* **37,** 2130–2145.

Forebrain

LUIS PUELLES* and JOHN RUBENSTEIN†

*University of Murcia, Spain and †University of California, San Francisco

GLOSSARY

prosencephalon Refers to the anteriormost major subdivision of the neural tube that gives rise to the forebrain. Consists of several major parts. The caudal part is the caudal diencephalon. The rostral part is the secondary prosencephalon; this region consists of the telencephalic and optic vesicles, and the rostral diencephalon.

prosomere Prosencephalic (forebrain) segment or neuromere. A transverse subdivision of the forebrain containing all the primary longitudinal subdivisions.

rostral diencephalon Unevaginated part of secondary prosencephalon, divided into hypothalamus proper (basal plate and floor plate) and prethalamus (alar plate).

telencephalon Large evaginated dorsal subdivision of the forebrain that contains primarily the cerebral cortex (pallium) and basal ganglia (subpallium).

The forebrain is the rostralmost portion of the central nervous system (CNS). It is an extremely complex assembly of structurally and functionally diverse structures that regulate most aspects of cognition, homeostasis, and behavior. In this brief overview, we describe the major components of the forebrain using a developmental viewpoint. Although the adult forebrain seems hopelessly complicated to most neuroanatomy students, a rudimentary understanding of its development provides a framework that simplifies its comprehension. Thus, we describe forebrain organization from a neuroembryological perspective. Most of the information that we will describe is valid for tetrapods (mammals, birds, reptiles, and amphibians). Available evidence supports the idea that the fundamental organization of the human forebrain differs little from that of less complex vertebrates. Most apparent differences are due to more extensive growth and larger morphogenetic deformation in the human brain of structures already found in other vertebrate brains. These differential aspects will be noted whenever this is appropriate.

I. EARLY DEVELOPMENT

The forebrain is derived from the anteriormost transverse domain of the neural plate that is generated during gastrulation (Fig. 1). Tissues adjacent to and within the neural plate produce molecules that regulate regional specification and morphogenesis of the CNS. Prospective subdivisions of the brain are specified through several mechanisms. Anteroposterior patterning generates transverse subdivisions: forebrain, midbrain, hindbrain, and spinal cord (Fig. 1A). Dorsoventral patterning generates longitudinally aligned domains, called floor plate, basal plate, alar plate, and roof plate (Fig. 1B). Within the forebrain, neuromeric theories postulate that there are further transverse subdivisions, which will be discussed later. Finally, regionally distinct molecular properties of the neuroepithelium and local signals from adjacent

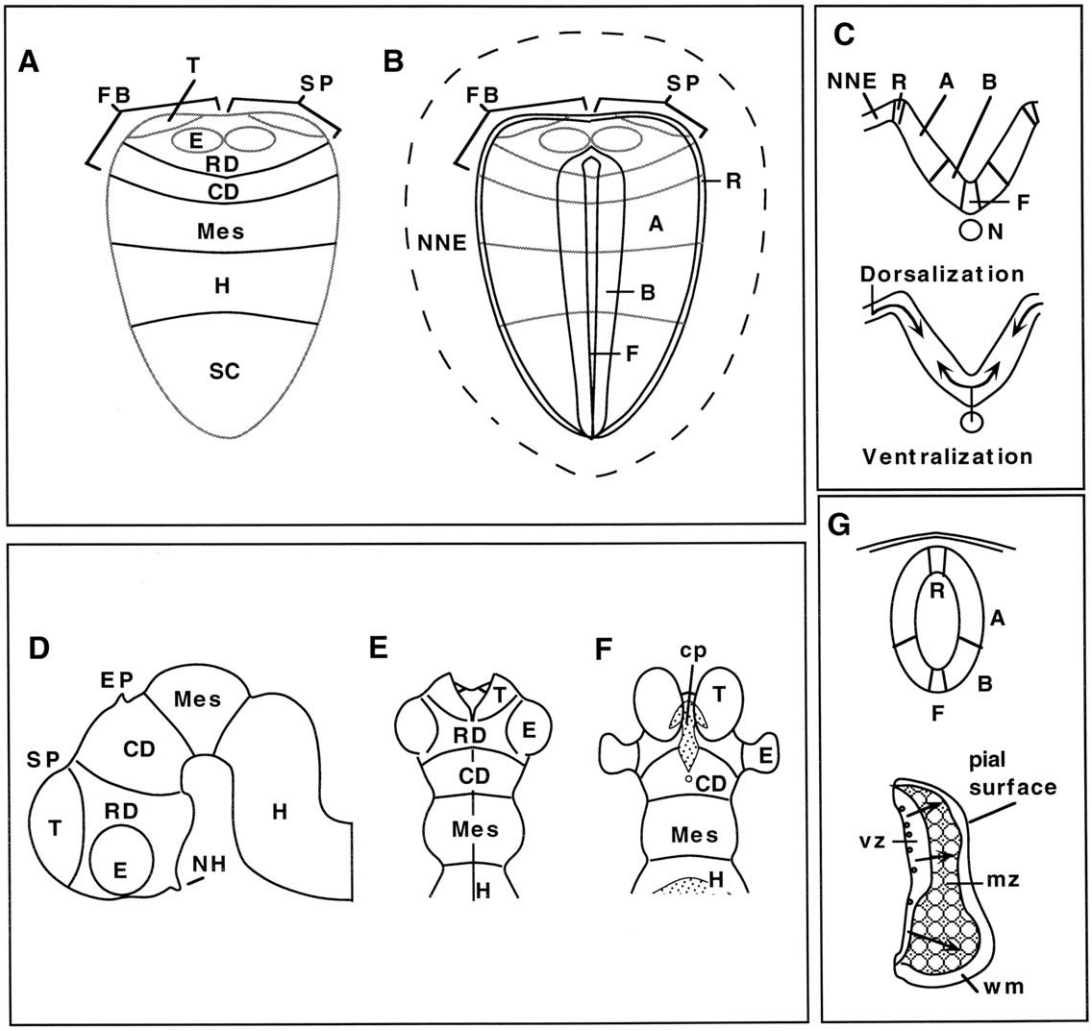


Figure 1 Early development of the forebrain. (A) Schema of transverse domains specified at the neural plate stage; the transverse boundaries are represented by black lines. Note that telencephalic and eye vesicles are separate outgrowths. (B) Schema of longitudinal domains at the same stage (boundaries in black; transverse and other limits in gray). (C) Schema of cross section through the neural plate showing the longitudinal domains and the relative position of the main tissues that regulate dorsoventral patterning, the notochord and the nonneural ectoderm. These exert ventralization and dorsalization effects, respectively, which are symbolized below. (D) Side view of the closed neural tube. The forebrain appears rostral to the mesencephalon and is composed of the caudal diencephalon, rostral diencephalon, eye vesicle, and telencephalon; the latter is shown as it starts to evaginate. (E) Dorsal view of the same stage as in D showing unfinished anterior closure of the neural tube and the bulging eye vesicles. (F). At a later stage, a dorsal view illustrates the stalks of the eye vesicles, the bilateral telencephalic vesicles, and the choroidal plexus primordium at the forebrain roof (cp; stippled area). (G) Schema of the closed neural tube in cross section and its longitudinal domains. Development of its wall is schematized below; a proliferating ventricular zone, a differentiating mantle zone, and a subpial layer of white matter (growing axons) can be distinguished.

tissues regulate the outgrowth of vesicles from the forebrain, such as the bilateral eye and telencephalic vesicles (secondarily also the olfactory bulbs), the neurohypophysis, and the epiphysis (pineal gland) (Figs. 1D–1F).

Dorsoventral patterning is regulated by nonneural tissues flanking the neural plate [the dorsalizing nonneural ectoderm (NNE) in Figs. 1B and C] and below the neural plate midline (the ventralizing notochord and prechordal mesendoderm) (N; Fig. 1C). Anteroposterior patterning is less fully understood, but it involves tissues underlying the neural plate, tissues at the anterior and posterior limits of the neural plate, as well as a later-appearing patterning

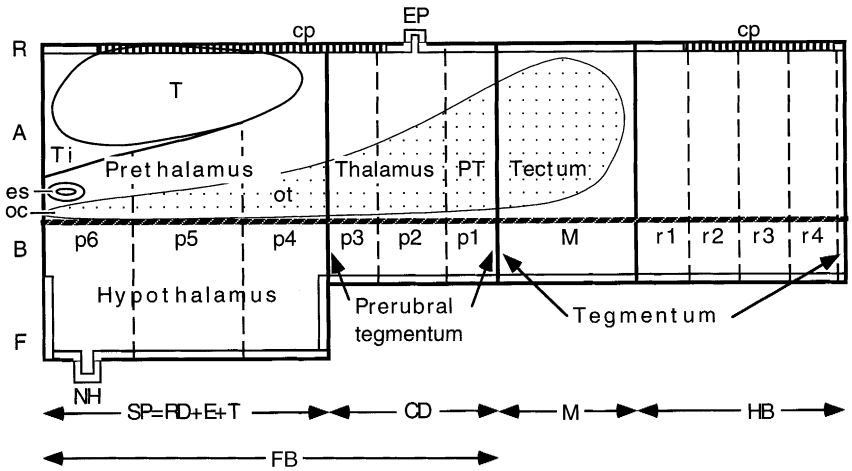


Figure 2 Topological schema of transverse and longitudinal subdivisions of the vertebrate forebrain (FB), midbrain (MB), and rostral hindbrain (HB). The longitudinal zones are indicated at the far left (R, roof plate; A, alar plate; B, basal plate; F, floor plate); the alar/basal boundary is represented by the thick black line with oblique white stripes. The transverse boundaries are vertical lines; black lines separate the main brain vesicles (identified underneath) and dashed lines separate the neuromeric subdivisions (prosomeres p1–p6), which are identified under the alar/basal boundary line. Various other specific names for different alar or basal territories are indicated. The choroidal plexi (cp) are marked by vertical stripes at the top. Note the position of the evaginated telencephalic vesicle (T) and the unevaginated telencephalon impar (Ti). The optic tract (ot) is schematized as a longitudinal domain extending from the optic chiasm (oc) to the optic tectum. See Table I for other abbreviations.

center at the midbrain–hindbrain junction. These stepwise processes generate a two-dimensionally regionalized specification map of prospective forebrain subdivisions (Fig. 2) Table I.

During the patterning stage, morphogenetic processes start to generate the shape of the forebrain. Neurulation folds the neural plate into the neural tube (Figs. 1C and G) and rapid cell divisions expand its surface area. The position-specific production of neurons and glia increases the thickness of the neural tube along the ventriculopial axis, which is the third dimension of the brain (Fig. 1G, arrows). Note that the internal, fluid-filled cavity of the neural tube and adult brain is called ventricular space, which is lined by the pseudostratified neuroepithelium; the outside of the neural tube contacts through a basement membrane the surrounding mesenchyme, which matures into a thin meningeal sheet called the "pia." Thus, the term "pial" means superficial, whereas "ventricular" means internal; most cell divisions occur in the "ventricular zone" (vz) of the neuroepithelium (Fig. 1G). Cell-type specification and differentiation processes generate postmitotic cells that migrate away from their site of origin in the vz toward the pial surface, under which they form the mantle zone (mz; Fig. 1G). Here, nuclei and laminar structures are gradually assembled. Many neurons and some glia cells only undergo radial migrations to the mz (Fig. 1G, arrows), thus maintaining a fixed position relative to their site of origin. This is probably essential for the generation of topographic connectivity maps between different brain regions. On the other hand, some neurons and oligodendrocytes undergo tangential migrations away from their primary sites of entrance in the mantle layer, frequently into specific target loci. Tangential migrations enable certain cell types that can only be formed in specific neural tube areas to become functionally integrated in local circuitry elsewhere.

II. PRINCIPAL COMPONENTS

In this section, we describe the major components of the forebrain in the context of their developmental origins and following the framework of the prosomeric model of Puelles and Rubenstein (Fig. 2). Although this model is still changing as new data accrue (in fact, we introduce some changes here), we suggest that it provides a useful conceptual format to integrate developmental mechanisms with the complex morphology of the forebrain. It is important to note that other morphological models of the forebrain have been postulated, including neuromeric and nonneuromeric

Table I
Anatomical Abbreviations

Abbreviation	Definition
A	Alar plate
Ac	Accoustic cortex
ac	Anterior commissure
Ac1-2	Primary and secondary accoustic cortical areas
ACC	Accumbens nucleus
AEP	Anterior entopeduncular area
AH	Anterior hypothalamus
AM	Amygdala (subpallial part)
B	Basal plate
BL	Basolateral amygdala
BM	Basomedial amygdala
BST	Bed nucleus of the stria terminalis
CAU	Caudate nucleus
cc	Corpus callosum
CD	Caudal diencephalon
Ce	Central amygdala
CLdl	Claustrum, dorsolateral part
CLvm	Claustrum, ventromedial part
cp	Choroid plexus tissue
DP	Dorsal pallium
DT	Dorsal thalamus
E	Eye vesicle
EMT	Eminentia thalami
EP	Epiphysis
es	Eye stalk
ET	Epithalamus
F	Floorplate
FB	Forebrain
FP	Frontal pole of telencephalon
GP	Globus pallidus
H	Hindbrain
hc	Habenular commissure
hic	Hippocampal commissure
iml	Internal medullary lamina
INS	Insular cortex
L	Lateral amygdala
LOT	Lateral olfactory tract
Lp	Lateral pallium
Mes	Midbrain
M	Motor cortex
M1–M3	Primary and secondary motor cortical areas
MAM	Mammillary region
Me	Medial amygdala
MP	Medial pallium
mz	Mantle zone
N	Notochord
NH	Neurohypophysis
NNE	Non-neural ectoderm
OB	Olfactory bulb
oc	Optic chiasm
OP	Occipital pole of telencephalon
OT	Olfactory tuberculum
ot	Optic tract
p1–p6	Prosomeres 1–6
p1c	Floor commissure of p1
p3c	Floor commissure of p3
Pal	Pallium
pc	Posterior commissure
PEP	Posterior entopeduncular nucleus
PIR	Piriform (olfactory) cortex
POA	Anterior (telencephalic) preoptic area
poc	Postoptic (supraoptic) commissure
POP	Posterior (prethalamic) preoptic area
PT	Pretectum
PUT	Putamen nucleus
PV–SO	Paraventricular-supraoptic nucleus
R	Roof plate
r1–r4	Rhombomeres 1–4
RD	Rostral diencephalon
RM	Retromammillary area
rt	Retroflex tract
S	Somatosensory cortex
S1–S3	Primary and secondary somatosensory cortical areas
SC	Spinal cord
SCH	Suprachiasmatic nucleus
Se	Septum
SIA	Subincertal area
sm	Stria medullaris
SP	Secondary prosencephalon
Sp	Subpallium
ST	Striatum
T	Telencephalon
Ti	Telencephalon impar
TM	Tuberomammillary region
TP	Temporal pole of telencephalon

(continues)

(continues)

Table I *(continued)*

Abbreviation	Definition
TU	Tuberal region
V	Visual cortex
V1–V4	Primary and secondary visual cortical areas
VP	Ventral pallium
VPA	Ventral pallidum
VST	Ventral striatum
VT	Ventral thalamus
vz	Ventricular zone
wm	White matter
zl	Zona limitans

ones of Herrick and Kuhlenbeck. The primary prosencephalon, or early forebrain vesicle, soon divides into two principal transverse subdivisions: caudally, the caudal diencephalon, and rostrally, the secondary prosencephalon (sum of rostral diencephalon and telencephalon; Figs. 1D–1F, 2). In our description, the conventional hypothalamus is constituted exclusively by parts of the rostral diencephalon and is independent from the caudal diencephalon.

The caudal diencephalon abuts caudally the midbrain; it has three transverse subdivisions, postulated to be prosencephalic neuromeres (prosomeres p1–p3) (Figs. 2 and 3). These contain the pretectal (p1), dorsal thalamic/epithalamic (p2), and ventral thalamic (p3) regions. Each prosomere has alar (dorsal) and basal (ventral) components; the aforementioned three large regions are the respective alar components of p1–p3 (Fig. 3). The basal components jointly form the prerubral tegmentum.

The transverse organization of the secondary prosencephalon (SP) is more hypothetical, but it has been postulated to contain three prosomeres as well (p4–p6), which are readily distinguishable in the rostral diencephalon, though it is not clear how they relate to the overlying telencephalon (Fig. 2). The telencephalon is a dorsal evaginated region within the SP (T in Figs. 1D–1F), and lies strictly dorsal to the rostral diencephalon. The latter also appears divided into alar and basal plates, like its caudal counterpart. The alar plate extends from the telencephalon down to approximately the ventral limit of the optic tract. Thus, the optic stalk and the optic tract lie in a longitudinal alar plate domain that traverses the rostral diencephalon, thalamus, pretectum, and midbrain (Fig. 2). The alar rostral diencephalon consists of three prosomeric subregions labeled caudal (p4), intermediate (p5), and rostral (p6) prethalamic areas (Fig. 4). The optic vesicles evaginate out of the rostral area at early neurulation stages (Figs. 1A–1F). Part of the prethalamus, particularly the rostral part, has been attributed to the hypothalamus, but it is convenient to restrict the term ''hypothalamus'' to the basal part of the rostral diencephalon. This includes mammillary/subthalamic (p4), tuberomammillary (p5), and tuberal hypophyseal regions (p6) (Fig. 4). In this way, the rostral diencephalon divides into optic vesicles and prethalamus (alar plate) and hypothalamus (basal and floor plates).

III. THE CAUDAL DIENCEPHALON

The pretectum (alar p1) is the caudalmost forebrain region. It is characterized by the posterior commissure, whose fibers cross the pretectal dorsal midline and then course transversally through the alar plate, just in front of the diencephalomesencephalic limit, before spreading longitudinally in the basal plate (pc in Fig. 3). The function of the posterior commissure is unclear. This region contains various pretectal nuclei involved in visual processing, including the centers for the pupillary and optokinetic eye reflexes. The subcommissural organ is a dorsal midline specialization that secretes glycoprotein (Reissner's fiber) into the ventricular fluid. The basal plate has dopamine-containing neurons that form part of the substantia nigra in addition to diverse reticular cell populations involved in motor circuits, like the parvocellular nucleus ruber (origin of the rubroolivary tract) and the interstitial nucleus of Cajal. The latter is a rostral source of descending preoculomotor axons in the medial longitudinal fasciculus, a tract that coordinates eye movements.

Anterior of the pretectum lies the dorsal thalamus and epithalamus complex (alar p2). Its dorsal midline includes the epiphysis (pineal gland), the habenular commissure, and choroid plexus rostrally. The epithalamus (ET or habenula) is the most dorsal nuclear complex; these neurons relate to the longitudinal stria medullaris tract (sm), which crosses the dorsal midline in the habenular commissure (hc in Fig. 3). The habenular nuclei produce the retroflex tract (habenulointerpeduncular tract). This compact transverse fascicle is a landmark that remains just anterior to the p1/p2 boundary throughout its dorsoventral extent; near the diencephalic floor it bends caudally (retroflects), continuing longitudinally across p1 and

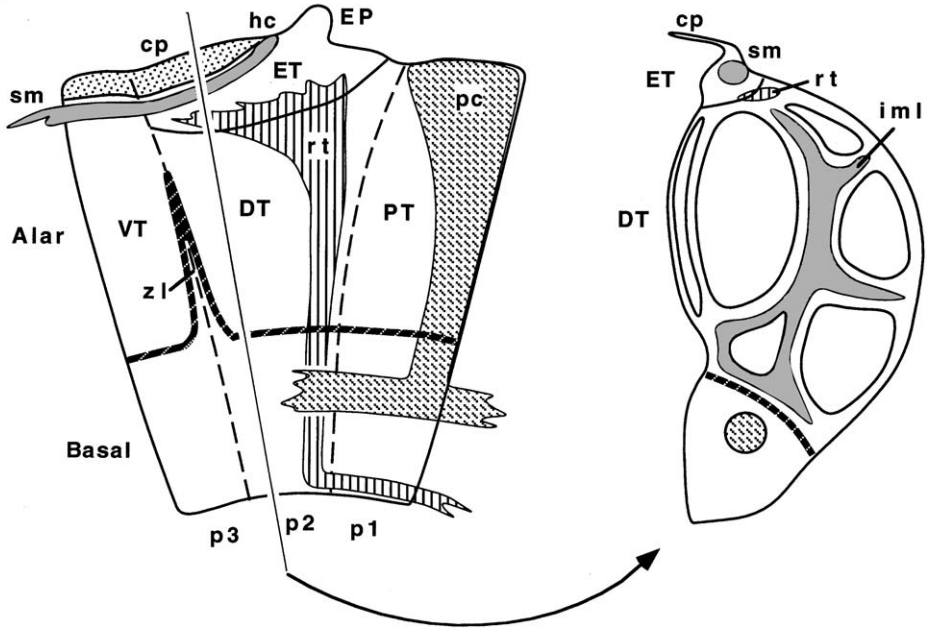


Figure 3 Schema of subdivisions of the caudal diencephalon and relevant anatomical landmarks in lateral view (rostral to the left; alar/basal limit as in Fig. 2): The pretectum (PT) in alar p1 contains the posterior commissure (pc); the dorsal thalamus (DT) and epithalamus (ET) in alar p2 contain to the retroflex tract (rt). The zona limitans (zl) separates dorsal and ventral thalami (VT) and seems to contain an expansion of the basal plate (according to gene expression data). The stria medullaris tract (sm) courses longitudinally near the roof choroidal plexus (cp). (Right) A schematic cross section through the dorsal thalamus, whose section level is indicated in the schema on the left, shows various nuclear groups (each of them is further subdivided into several nuclei), separated by the axon-rich internal medullary lamina (iml). See Table I for other abbreviations.

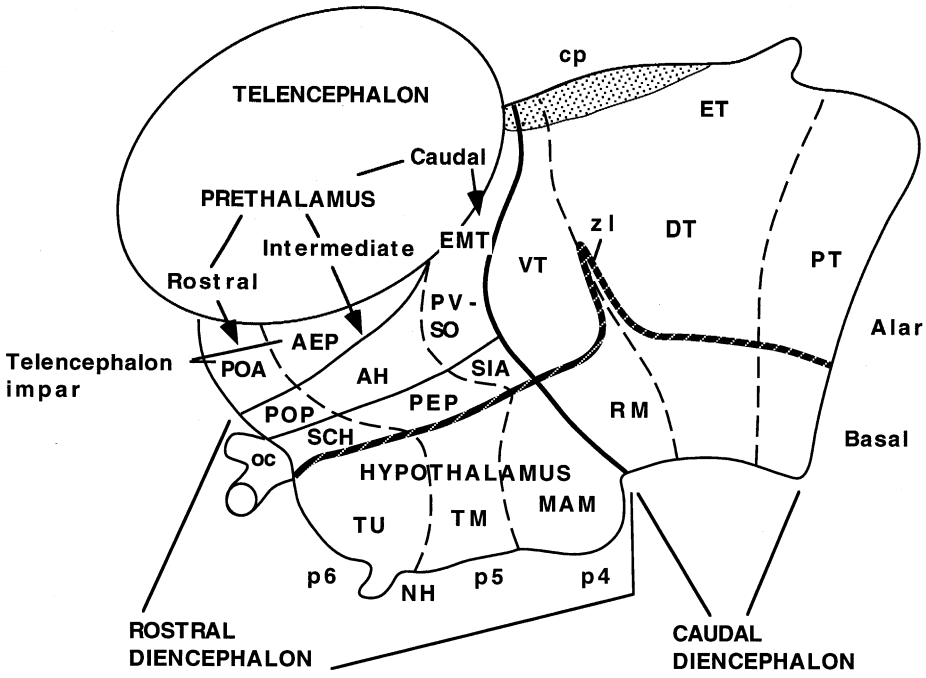


Figure 4 This schema highlights subdivisions in the rostral diencephalon in the context of the overlying telencephalon and caudal diencephalon. We call "hypothalamus" the basal part of the rostral diencephalon and "prethalamus" the alar part. These are both divided into three prosomeres (p4–p6) and various smaller areas. See Table I for abbreviations.

the midbrain into the interpeduncular nuclear complex in the isthmic floor plate (rt in Fig. 3).

The dorsal thalamus (DT) lies ventral to the epithalamus and consists of an elaborate complex of nuclei that generally project to the telencephalon (targeting both subcortical and cortical structures) (Fig. 3 and arrow 11 in Fig. 5A). These nuclei serve as relay centers for numerous telencephalopetal pathways with diverse functional implications. The telencephalic cortex sends topographically ordered projections to the respective thalamic relay centers. The main subdivisions of the DT contain groups of related nuclei. The *intralaminar* nuclei lie within the fiber-rich internal medullary lamina, which separates the other nuclear groups (iml in Fig. 3). These intralaminar nuclei form a separate projecting system with a modulatory role over the cortex and basal ganglia (embodying the final stage of the ascending activating system for mental arousal). The *sensory* nuclei process somatosensory and viscerosensory information (ventrobasal complex), visual input (dorsal lateral geniculate nucleus), and auditory input (medial geniculate nucleus), which are then relayed by these nuclei to the primary sensory areas of the isocortex via axons in the internal capsule (telencephalic peduncle; arrow 11 in Fig. 5A). Other thalamic nuclei are part of the *motor control* system (ventral anterior and ventrolateral nuclei); these process inputs from the globus pallidus and the cerebellar dentate nucleus and send efferents to the motor and premotor cerebral cortex. The anterior and periventricular thalamic nuclei are the *limbic* part of the thalamus, receiving hypothalamic (mammillary and other) projections and projecting to the limbic cingulate cortex, hippocampus, septum, and amygdala. Another important group of dorsal thalamic nuclei is conceived as the *associative* group (lateral dorsal, lateral posterior, pulvinar, and medial nuclei). These nuclei receive their major inputs from parts of the secondary sensorimotor, limbic, and associative cortex and project again into higher order cortical areas of the frontal, parietal, and temporal lobes. The associative nuclei are particularly developed and become secondarily subdivided and specialized in man.

The basal plate of p2 is poorly understood; it resembles that of p1 in that it contains a part of the dopaminergic neurons of the substantia nigra and of the median ventral tegmental area, in addition to some reticular populations participating in preoculomotor functions (i.e., the rostral interstitial nucleus).

The transition from p2 to p3 is characterized by several major changes in molecular and developmental properties. The p2/p3 limit appears as a fiber-rich gap between dorsal and ventral thalami; it is known as the zona limitans intrathalamica, or external medullary lamina (zl in Figs. 3 and 4).

Alar p3 is also known as the ventral thalamus for historical reasons, even though it is anterior to the dorsal thalamus (VT in Figs. 2 and 3; compare resulting position in bent forebrain, as in Fig. 1D). The choroid plexus of p2 extends through p3 and p4 into the telencephalon (cp; dotted roof domain in Figs. 1F, and 3 and 4; striped roof in Fig. 2). A choroid plexus is a neuroepithelial specialization where blood plasma is filtered into the brain ventricles as cerebral spinal fluid, which circulates and then flows out from holes in the hindbrain roof. Alar p3 consists of the ventral lateral geniculate nucleus (a superficial reflex visual center), the reticular thalamic nucleus (traversed by the thalamocortical fiber bundles), and a periventricular zona incerta. The latter two nuclei contain many inhibitory neurons, whose projections spread widely ipsi- and contralaterally in the alar and basal plates of the caudal and rostral diencephalon, perhaps reaching the midbrain. Inhibition exerted by the reticular nucleus on the dorsal thalamus organizes rhythmic electrical activity characteristic of sleep.

The basal plate of p3 consists of the so-called posterior hypothalamic area and the retromammillary area, which is frequently named the "supramammillary area." The latter is characterized by some dopaminergic neurons and other reticular cell populations.

IV. THE ROSTRAL DIENCEPHALON

As outlined previously, we use the term rostral diencephalon to refer to the extratelencephalic part of the secondary prosencephalon. We hypothesize that it is divided in basal and alar plates across three prosomeres (p4–p6), with its alar plate forming the prethalamus and its basal/floor plates representing the hypothalamus, as originally defined for human embryos by His in 1890 (Figs. 2 and 4). Although the diencephalic prosomeres p1–p3 represent the part of the extratelencephalic forebrain machinery that regulates information processing (and subsequent motor output) of signals coming in from the external world, the rostral diencephalic prosomeres p4–p6 attend mainly to the inner world of viscera, neurohumoral functions, and internal homeostasis. They also provide ascending (motivating) input into the limbic parts of

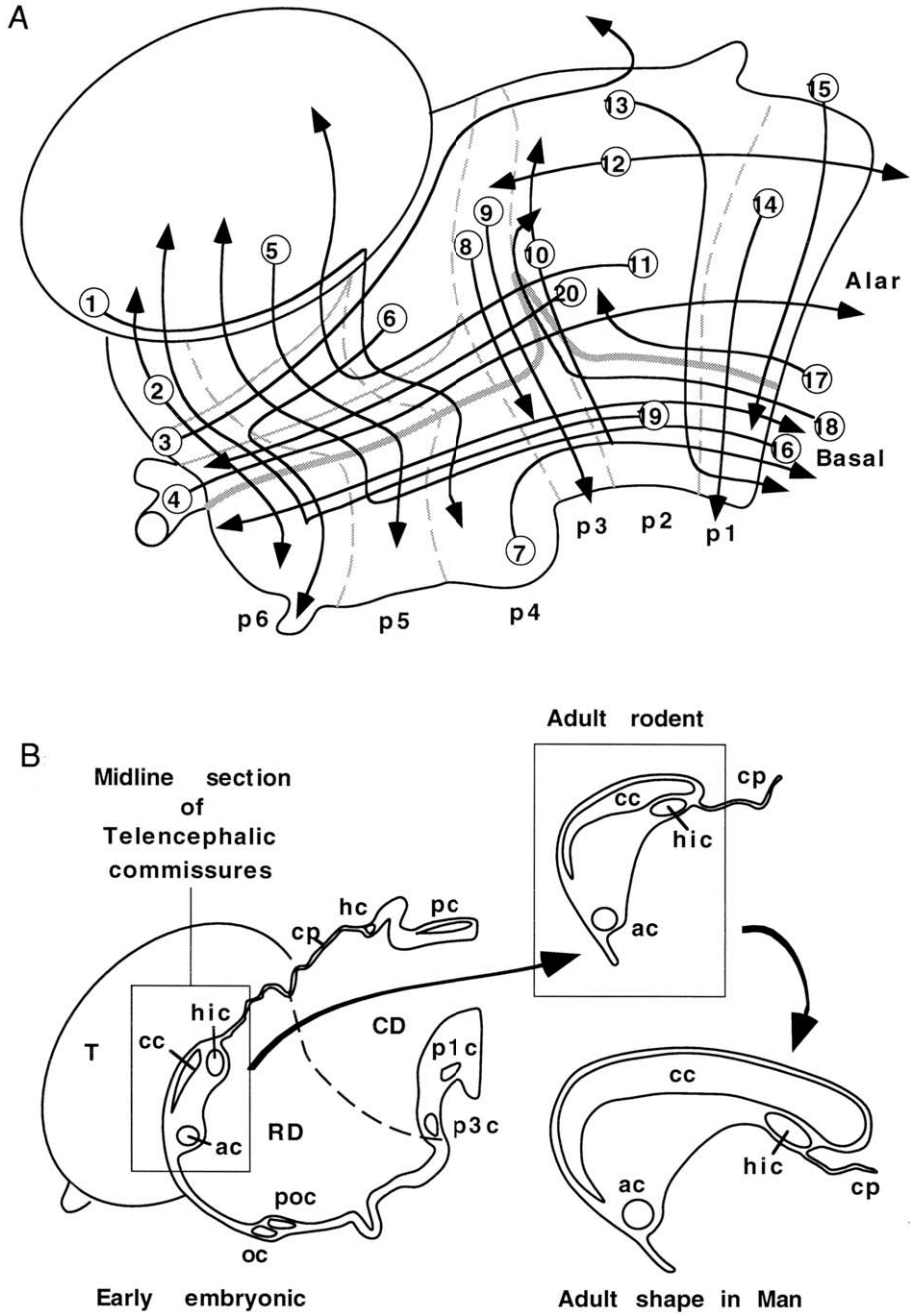


Figure 5 Schema of connectivity in the rostral and caudal diencephalon. (A) In general, tracts can be mapped within the prosomeric model as aligned with the transverse or longitudinal dimensions of the forebrain components. There are thus basically transverse tracts or tract portions and longitudinal tracts or tract portions. Individual tracts may change direction at specific decision points. This schema does not describe all the tracts present in the forebrain. It only exemplifies the previous general principle by illustrating 20 assorted tracts: 1, Fornix (hippocampomammillary) tract; 2, septohypothalamic tract; 3, stria medullaris (into habenular commissure); 4, optic tract; 5, amygdalohypothalamic tract; 6, supraoptohypophyseal tract; 7, mammillotegmental tract; 8, incertotegmental tract; 9, commissural ventral thalamic tract; 10, mammillothalamic tract; 11, thalamotelencephalic tract; 12, longitudinal alar interconnections; 13, retroflex tract; 14, commissural anterior pretectal tract; 15, posterior commissure; 16, nigrostriatal tract; 17, medial lemniscus; 18, superior cerebellar peduncle; 19, medial forebrain bundle; 20, intergeniculate–suprachiasmatic tract. (B) The various commissures present in the forebrain are illustrated in a median sagittal section. Emphasis is placed in the area where the telencephalic commissures develop, showing in the two schemas on the right the progressive relative increase in size of the corpus callosum over developmental time and from adult rodent to adult man. See Table I for abbreviations.

the telencephalon through the medial forebrain bundle, whereas the caudal diencephalon is connected to the telencephalon via the lateral forebrain bundle (internal capsule). The prethalamus and hypothalamus accordingly are a collection of neural centers that are involved in regulation of homeostasis, reproduction, the autonomic nervous system, and emotional states.

The caudal part of the hypothalamus (p4) includes the mammillary region, the subthalamic nucleus, the lateral hypothalamic area, and part of the posterior hypothalamus (MAM in Fig. 4). The prethalamic (alar) region of p4 reaches dorsally the forebrain roof (choroidal plexus) via the eminentia thalami at the caudal end of the telencephalic stalk (EMT in Fig. 4); this area possibly extends into the telencephalon through the medial amygdala. The p4 alar plate probably includes among its derivatives the subincertal area, the paraventricular and supraoptic nuclei (SIA and PV–SO in Fig. 4), and the perireticular nucleus (not shown), with the latter lying just in front of the ventral thalamic reticular nucleus. The eminentia thalami received its name due to its protrusion at the back of the interventricular foramen (tight passage interconnecting the prethalamic ventricular space with that of the telencephalic vesicle, or lateral ventricle) and the idea that it belongs to the thalamus. Its neuronal derivatives participate in the bed nuclei of the stria terminalis (posterior, or medial, parts) and the stria medullaris. This domain is characteristically traversed superficially by the telencephalic peduncle as it exits the telencephalon, incorporating as well the thalamotelencephalic projections. The fornix tract —hippocampal fibers passing from the caudal aspect of the commissural septum to the mammillary region —follows the dorsoventral dimension of p4 at an intermediate depth (arrow 1 in Fig. 5A).

The intermediate hypothalamus (p5) is represented by the tuberomammillary region, which also contains the premammillary nuclei and the dorsomedial hypothalamic nucleus (TM in Fig. 4). The corresponding alar or prethalamic area extends through the posterior entopeduncular area, the conventional 'anterior hypothalamus,' and anterior entopeduncular area into the telencephalic stalk (PEP, AH, and AEP in Fig. 4).

The rostral hypothalamus (p6) starts ventrally with the neurohypophysis, median eminence, and arcuate nucleus (floor plate derivatives) and includes the basal plate tuberal area, with the ventromedial hypothalamic nucleus and the anterobasal nucleus (retrochiasmatic area) (NH and TU in Fig. 4). The rostral prethalamic alar plate continues through the suprachiasmatic and posterior/anterior preoptic regions into the telencephalic stalk (SCH, POP and POA in Fig. 4).

The mammillary area is a complex of nuclei with a variety of inputs, including the fornix fibers from the hippocampus, as a part of the limbic circuit of Papez. These nuclei send their major outputs through the basal plate of p3, p2, and p1 to the brain stem tegmental nuclei (mammillotegmental tract), with a collateral projection to the anterior dorsal thalamus, via the mammillothalamic tract (arrows 7 and 10 in Fig. 5A); these are also part of the Papez limbic circuit, which returns from there to cingulate and hippocampal cortex.

The tuberal/infundibular region and pituitary stalk are traversed by neuroendocrine fibers carrying vasopressin and oxytocin from magnocellular secretory cells in the overlying alar plate (supraoptic and paraventricular nuclei) to the neurohypophysis (arrow 6 in Fig. 5A); the neurohypophysis liberates these substances into the bloodstream. The median eminence contains a profusion of nerve terminals that secrete the proteins that regulate the release of various hormones from the anterior pituitary (corticotropin, thyrotropin, gonadotrophin, and growth hormone) into the hypophyseal portal capillaries. These regulator proteins arise from discrete groups of parvocellular neurosecretory neurons in the anterior hypothalamus and preoptic area. Other modulation of the pituitary occurs directly via axon terminals of dopamine neurons in the arcuate nucleus. The anterobasal nucleus bridges the retrochiasmatic midline area as a bed nucleus of the postoptic (supraoptic) commissures (poc in Fig. 5B) and is the rostralmost component of the basal plate.

The prethalamus can be subdivided into two superposed longitudinal tiers across prosomeres p4–p6. The lower tier coincides with the subpial course of the optic tract and the optic chiasm (Figs. 2 and 5A). It contains the suprachiasmatic nucleus (p6; involved in the control of circadian rhythms in homeostatic functions), the posterior entopeduncular area (including the migrated posterior entopeduncular nucleus; p5) and the subincertal area (p4). Little is known about the functions of the more caudal area. The upper tier is the so-called optoeminential domain, which lies just under the telencephalic stalk and telencephalon impar (unevaginated parts continuous with telencephalic structures, such as the telencephalic preoptic and median septal areas; Ti in Fig. 2). The optoeminential domain starts rostrally (p6), with a prethalamic preoptic area found around the optic stalk recess (POP in Fig. 4). It

includes the anterior hypothalamic area (p5), dorsal to the PEP, and the supraoptic and paraventricular nuclei, together with the derivatives of the eminentia thalami (p4; posterior bed nucleus striae terminalis and bed nucleus striae medullaris). At the boundary with the telencephalic stalk, this domain is traversed longitudinally by the stria medullaris, which continues caudalwards into the diencephalic roof (epithalamus) (arrow 3 in Fig. 5A). The overlying telencephalic stalk domain (telencephalon impar) consists of cell groups found at the peduncular transition into the evaginated telencephalic vesicles. From caudal to rostral, these contribute to the amygdala, the extrapyramidal system [entopeduncular nucleus (or internal segment of globus pallidus in highly evolved mammals), substantia innominata, and basal nucleus of Meynert], the anterior preoptic area/diagonal band formation, and the telencephalic preoptic area.

V. BASIC CIRCUITRY IN THE EXTRATELENCEPHALIC FOREBRAIN

The whole lateral wall of the rostral and caudal diencephalon is traversed by numerous transverse, longitudinal, and commissural fibers that interconnect the diverse prosomeric centers into interactive circuitry (Figs. 5A and 5B). These interconnecting fiber systems often reach the midbrain alar and basal plates (and extend from there into the rest of the brain stem), and they converge rostrally at the postoptic (so-called supraoptic) commissure or diverge into the telencephalic stalk. The commissures allow the left and right halves of the brain to interact for coordination of both analytical (alar) and motor/neurovegetative/neurohumoral (basal) functions.

Another pervasive fiber system, the optic tract, is subpial and is found largely in the alar plate (there is also an accessory basal optic tract targeting specific terminal centers in the basal and alar parts of p1). The main optic tract courses longitudinally from the optic chiasm, where half of its fibers decussate (in man), through the prethalamus, thalamus, and pretectum and up to the midbrain roof (tectum)—the superior colliculus. Along its way, the optic tract gives out collaterals or terminal fibers to prethalamic areas (such as the suprachiasmatic nucleus) and to thalamic (ventral and lateral geniculate nuclei, intergeniculate leaflet, and pulvinar nucleus) and various pretectal centers receiving topographically ordered (retinotopic) projections (anterior pretectal nucleus and nucleus of the optic tract) or nonordered retinal projections (olivary pretectal nucleus). The retinorecipient posterior pretectal nucleus lies in the midbrain.

There are also many characteristic transverse fiber tracts normally coursing close to a given interprosomeric boundary (Fig. 5A), although other systems of transverse fibers course sheet-like throughout a given portion of the wall. Most of these fibers change course (i.e., become longitudinal) at a given decision point, generally coinciding with entrance into the basal plate; a few proceed into commissures across the roof or floor plate (mainly in p1 and p3; Fig. 5B).

Given the specialization of alar centers in analytic signal decoding/locating and associated relay functions, and the specialization of the basal centers in motor patterns, longitudinal alar or basal fibers crossing diverse neuromeres tend to integrate separately such functions (bilaterally, due to the commissures). Conversely, fiber tracts coursing transversally convey the diverse segmentally analyzed (and longitudinally cross-correlated) alar outputs into the responding basal plate net of reticular and motor neurons. Much of our subconscious brain activity and reflex behavior depend on this subtle multimodal neuronal machinery for its precise situation- and aim-dependent adaptations. Note that this system, which is able to respond in a large extent to external and/or internal multimodal stimuli autonomously, is also controlled by the telencephalon for a higher degree of contextual integration, or volitional control, as indicated by separate descending telencephalic projections to many alar and basal plate centers.

VI. THE TELENCEPHALON: BASIC PARTS AND MORPHOGENESIS

The telencephalic hemispheres evaginate bilaterally from the dorsalmost alar plate of the secondary prosencephalon (Fig. 1). This process includes within each vesicle a portion of the roof plate choroidal tissue, which later builds the choroidal plexi of the lateral ventricles (Figs. 1 and 2F). Rostrally, a thick median wall portion is the site where the telencephalic commissures develop (fibers interconnecting both vesicles; there are three major commissural pathways in mammals—anterior commissure, hippocampal commissure, and corpus callosum; Fig. 5B).

There are two principal subdivisions of the telencephalic vesicle: the roof, or pallium, and the basis, or subpallium (Fig. 6A). The subpallium consists

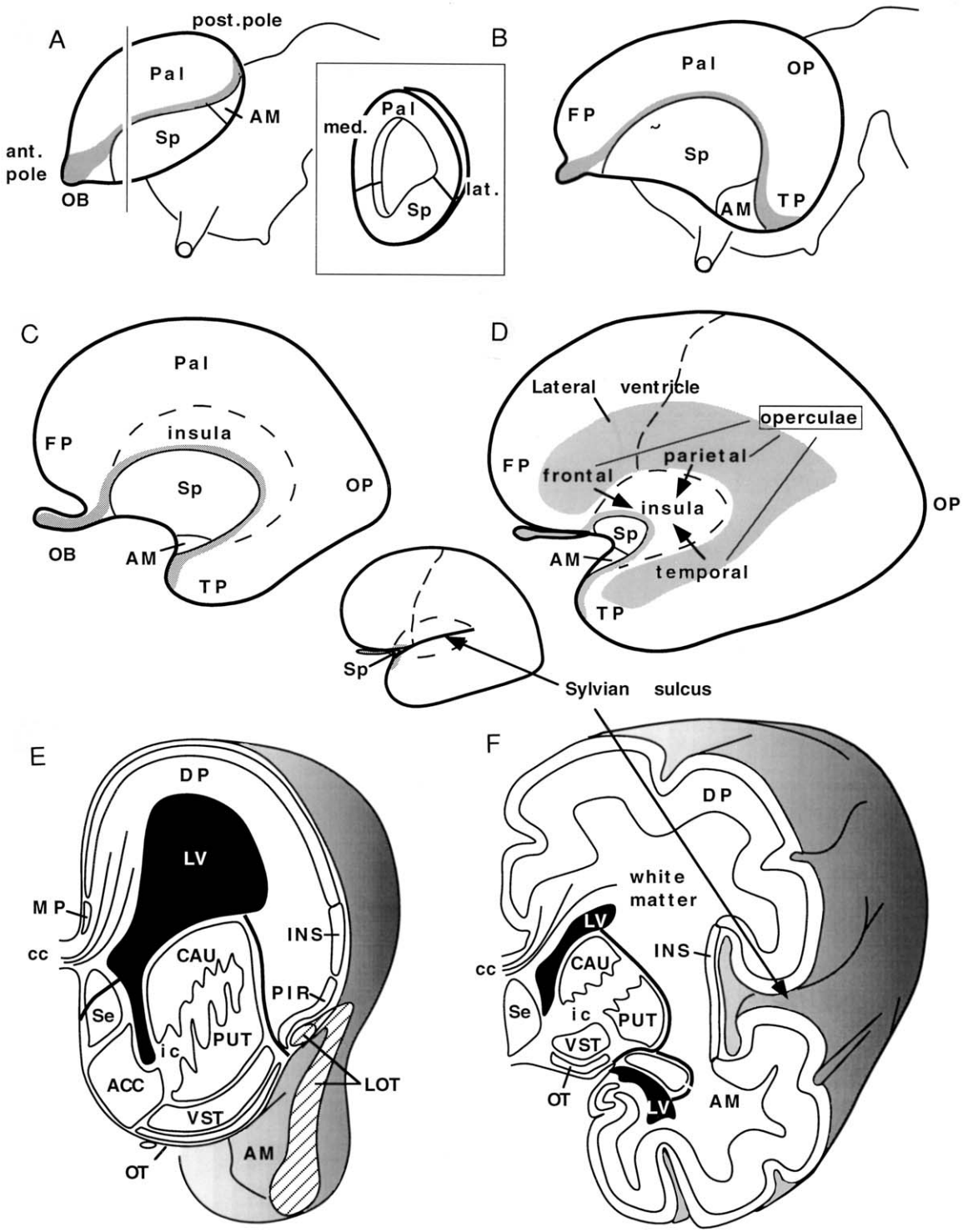


Figure 6 Morphological development of the telencephalon. (A–D) These drawings show lateral views of the left telencephalon at four developmental stages (see text). The inset in A shows a cross section and internal division into pallium (Pal) and subpallium (Sp). The olfactory tract is shaded in gray and helps to visualize the enormous growth of the pallium relative to the subpallium. (E) This drawing represents a section through the frontal part of the telencephalon, seen from the front; it shows in perspective the more caudal parts and the position of the temporal lobe. Within the section, the thick black lines mark the palliosubpallial boundary. Many subdivisions are indicated, as well as the lateral olfactory tract (LOT). (F) Similar to the drawing in E, representing a later stage in which further cortical expansions have introduced gyrification, internment of the insular cortex in the Sylvian sulcus (arrows) under the opercular overgrowths (compare with D and inset), and entrance of the temporal lobe (including its ventricular cavity) into the section plane (compare full shape in D). See Table I for abbreviations.

primarily of nuclei, the so-called basal ganglia (Figs. 7 and 8; note that use of the term "basal" here, or with regard to the amygdala, does not mean an embryological origin in the neural tube basal plate but, rather, a rough topographic indication of intratelencephalic position; all telencephalic parts are alar). The pallium contains primarily cortical structures but also some pallial nuclei (Fig. 8).

The area occupied by the basal ganglia in the hemisphere differentiates early and then its growth slows, whereas the overlying pallium is capable of prolonged surface growth. The increasing disproportion between these two parts leads to a characteristic morphogenetic deformation that in highly evolved mammals consists of a progressive incurvation and relative diminution in size of the subpallium, forced by anteroposterior and mediolateral expansion of the pallium (Fig. 6). The early forming posterior pole of the vesicle is converted into the temporal pole, which protrudes laterally and rostralwards (TP in Figs. 6B–6E). New anterior and posterior poles appear in parallel, forming the definitive frontal and occipital poles of the hemisphere (FP and OP in Figs. 6A–6E). An outgrowth at the early forming anterior pole forms the olfactory bulb, which gradually becomes displaced under the new frontal pole or orbitofrontal cortex (OB in Figs. 6A–6D). The lateral olfactory tract projecting from the olfactory bulb to the primitive posterior pole (gray in Fig. 6A) is transiently visible at the surface along the whole telencephalon, always just lateral to the palliosubpallial boundary (gray in Figs. 6C and 6D).

Internally, the subpallium forms an intraventricular bulge that affects the shape of the lateral ventricle

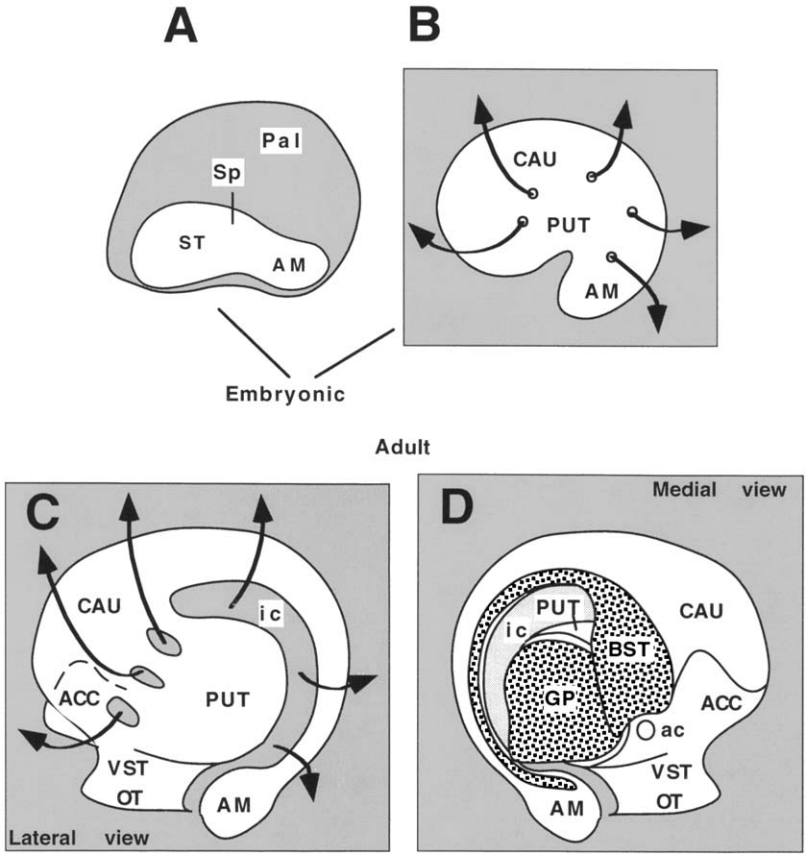


Figure 7 Development and subdivisions of the basal ganglia. (A). The common mass of subpallial formations is shown under the pallium. Its caudal part contains prospective subpallial amygdala (AM) and its rostral part the striatum (and pallidum, internally). (B). Early passage of fibers from the internal capsule (arrows) starts to divide the subpallial nuclei. (C, D). At the final stage, external (C) and internal (D) views show the different main portions that are distinguished. The stippled formations in D represent the pallidal components. See Table I for abbreviations.

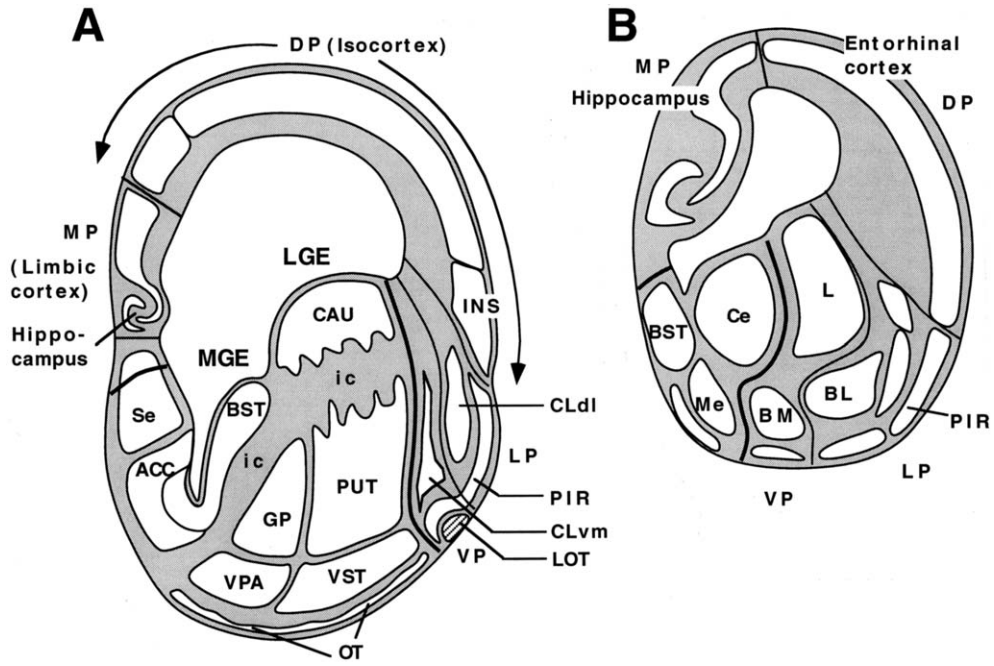


Figure 8 Pallial and subpallial subdivisions shown in schematic cross sections through the middle sector of an undeformed telencephalon (A) and through the amygdaloid complex at the temporal pole (B). The thick black line represents the palliosubpallial boundary. Note the pallium is subdivided into ventral, lateral, dorsal, and medial portions. See Table I for abbreviations.

(Sp in Figs. 6A, 6E and 6F, 7, and 8). The basal ganglia are traversed by the radiating fibers of the internal capsule (bidirectional thalamocortical axons) and become secondarily subdivided into several portions (Figs. 6E and 6F, 7B–D, 8A, and 8B). Pallial overgrowth also occurs mediolaterally, both in the main frontooccipital body of the hemisphere and in the temporal horn. This leads to the formation of the Sylvian (lateral) fissure and the progressive internment under the frontoparietal and temporal operculae of the piriform (olfactory) and insular cortexes (these are the earliest formed cortical parts, covering laterally the basal ganglia) (Figs. 6D–6F).

Additional adjustments of the pallial surface to more localized bouts of final surface growth lead to the partial or total burial of the oldest formed cortex (anchored by its more advanced connections) in the depth of other fissures (hippocampal, collateral, internal parietooccipital, and calcarine) and constant or variable sulci (central, frontal, parietal, temporal, cingulate, and rhinal). The intervening gyri protrude superficially in interlocked shapes, adapting to available space in the cranium (Fig. 6F). The main sulcal formations serve to separate the frontal, parietal, occipital, temporal, insular, and cingulate/entorhinal lobes; the latter includes the hippocampus. The relative amount of gyrification increases in evolutionarily more advanced mammals. The increase in cortical surface that is found in complex mammals is thought to occur without fundamental changes in the radial organization of the cortex, which is divided into columnar modules of constant dimensions and cell density across mammals. Thus, one evolutionary trend tends to increase the number of cortical columnar modules (presumably improving the analytical capacity of the animal). At the interhemispheric surface, gyrencephalic animals also show a correlatively larger corpus callosum, the main interhemispheric commissure (Fig. 5B).

VII. TELENCEPHALIC COMPONENTS

A. The Pallium

Recently, the pallial cover of the telencephalic vesicle has been shown to be divided molecularly and structurally into four pallial territories, that are common to all vertebrates: the medial, dorsal, lateral,

and ventral pallial domains (Fig. 8). The medial and dorsal pallial parts mature structurally into purely cortical centers. The medial pallium primarily forms the hippocampal cortex (allocortex; three-layered), although parts of the surrounding transitional cingulate/entorhinal cortex (mesocortex; four- or five-layered) may have the same origin. This functionally connected complex is subsumed under the term "limbic lobe" and is important in the evaluation of experiences, motivation of behavior, emotions and memory.

The dorsal pallium forms the isocortex (six- or seven-layered; predominant in relative surface extent). This domain specializes into separate sensory or motor cortical areas. Primary sensory cortex processes relatively simple sensory data; secondary cortex generates more complex perceptual constructs of the external world and the body. Associational cortical areas are where higher mental functions are believed to occur (Figs. 9B and 9C). The ancestral dorsal pallium appears as a variably sized island surrounded by the rostrally and caudally interconnected medial and lateral pallia; the reduced sensory/motor and associative representations partly overlap (Fig. 9A). The dorsal pallium is the portion that expands more in mammalian evolution so that the isocortex represents the largest and most variably subdivided cortical domain in mammalian species (Figs. 9B and 9C). Major axonal outputs of the isocortex (apart from the massive layer II/III corticocortical projections, both within the same hemisphere and projecting into the contralateral hemisphere using the corpus callosum) are the corticoclaustral, corticostriatal, corticothalamic, corticopretectal–collicular, corticoreticular, corticopontine, corticonuclear, and corticospinal fiber pathways. They jointly constitute a potent means for

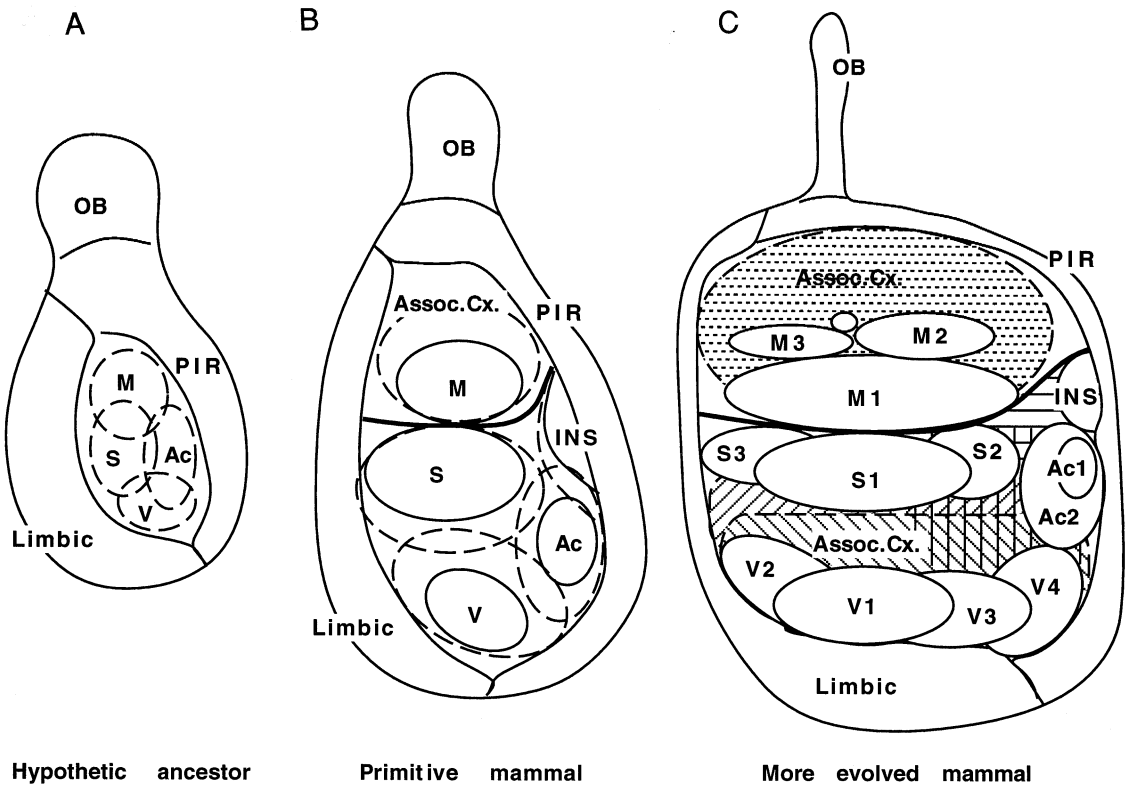


Figure 9 Drawings of flattened cortical surface, with hypothetical three stages of cortical evolution in mammals. The progressive expansion and diversification of specialized and associative fields within the isocortex is put in the context of its invariant topology, both externally [with respect to the surrounding limbic and piriform (PIR) cortexes] and internally (with respect to each area's topological relationship to the other isocortical areas). Dashed lines indicate imprecise boundaries. The thick black lines in B and C indicate the boundary between frontal lobe (action planning and motor functions) and the parietal, occipital, and temporal lobes (sensorial analysis and abstraction). The different patterns of lines in C symbolize associative multimodal intermixing of analysis properties. In man, the number of identified visual areas is approximately 20, and similar numbers may exist for the other specialized regions; this indicates the relative complexity of the functions performed in these domains. See Table I for abbreviations.

descending modulation of subjacent alar and basal centers, and they arise primarily from layer V pyramidal neurons (except those to claustrum and thalamus, which arise from layer VI neurons). The acquisition of direct cortical innervation of motoneurons seems to be a relatively recent evolutionary addition since it is absent in nonmammals.

The lateral and ventral pallial domains both give rise to portions of the olfactory cortex. In addition, each develops underlying pallial nuclei, which have been classified tentatively according to their molecular profiles (Figs. 8A and 8B). The lateral pallial nuclear formation consists rostrally of the dorsolateral claustrum and caudally of the endopiriform nucleus and basolateral parts of the amygdala (plus associated cortical amygdala domains). The ventral pallial nuclear formation is represented rostrally by the ventromedial claustrum and caudally by the lateral and basomedial parts of the amygdala (and associated cortical parts). Little is known of the functions of the claustrum, although the dorsolateral part has a bidirectional topographically ordered projection with the whole isocortex. The ventromedial claustrum may receive multimodal thalamic connections. The pallial nuclei of the amygdala are known to integrate multiple inputs from the isocortex, apparently integrating recognition of contextual Gestalts (characteristic sense data configurations) with entrainment of appropriate instinctive or learned (associated) behavior through motivational (limbic) and direct descending efferents onto extratelencephalic circuitry. It should be noted that the amygdala was long held to be only a part of the basal ganglia, but it has recently emerged as a heterogeneous complex of interactive pallial and subpallial elements (Fig. 8C). The same can be said for the telencephalic septum, found at the interhemispheric or medial wall, which also shows pallial and subpallial portions, although in this case the pallial part (the dorsalmost part of the septum) is much smaller than its subpallial constituents.

B. The Subpallium

The subpallium may be divided for descriptive purposes into a middle sector and two transitional domains at its rostral and caudal ends. The wall of the middle sector bulges into the ventricular cavity early in development, forming two ridges called the medial and lateral ganglionic eminences (MGE and LGE); these transient structures are perforated by the major tract interconnecting the diencephalon with the telencephalon, the internal capsule (ic; arrows in Figs. 7B and 7C; Fig. 8A).

The mantle layer under the MGE matures earliest and gives rise to the globus pallidus (GP) or pallidum (there are dorsal and ventral pallidal formations; the dorsal pallidum divides into internal and external segments in hominids; in other mammals, the internal pallidal segment is called entopeduncular nucleus; Fig. 8A). The internal capsule fibers separate the GP from the periventricular pallidal derivative, which forms a curved complex called the bed nucleus of the stria terminalis (BST in Figs. 7D, 8A and 8B). The stria terminalis is a tract interconnecting the amygdala with the anterior commissure and surrounding areas (Fig. 7D). Together with the LGE, the MGE also produces numerous cohorts of neurons that migrate tangentially to invade the striatum and the whole cortical pallium (they largely become incorporated as cortical inhibitory interneurons).

The mantle developing under the LGE is the striatum [laterally there are dorsal and ventral striatum parts and a rostromedial part, the accumbens nucleus (ACC); Figs. 6E, 7C–7D, and 8A]. In man, the internal capsule partially divides the dorsal striatum into a peripheral portion called putamen and a periventricular portion called caudate nucleus (Fig. 8A). An obsolete anatomical conception artificially joined the putamen to the medially adjacent dorsal pallidum under the concept of ''lenticular nucleus''; this term nevertheless still appears in clinical usage and in some tracts projecting out of the pallidum (lenticular tract and ansa lenticularis).

The respective dorsal and ventral parts of the striatum and pallidum are separated (arbitrarily) by the plane in which rostral and caudal components of the anterior commissure collect and course toward the midline commissural plate. At the subpial (ventral) surface of the subpallium, various superficially migrated cell populations form the laminated olfactory tuberculum. Toward the telencephalic stalk, this ventral area is contiguous with the substantia innominata, which contains both dispersed and dense groups of cholinergic neurons, such as the basal nucleus of Meynert. These constitute the main source of cholinergic input to the overlying cortex.

The striatopallidal complex is integrated into a circuit that modulates the motor output of the motor and premotor cortex. The cortex projects excitatory axons to specific areas of the striatum complex (characteristic afferents: accumbens and ventral striatum, limbic; caudate, associative; and putamen,

sensorimotor). The striatum projects separately to the external and internal parts of the pallidum, where its effect is inhibitory, phasically breaking a tonic inhibitory activity of the pallidum over its targets. The external pallidal part interacts bidirectionally with the subthalamic nucleus, whose fibers counteract striatal inhibition in the pallidum. The internal part of the globus pallidus projects its tonic inhibition on the motor dorsal thalamic nuclei, which connect with the motor and premotor cortex. Phasic striatal inhibition of the pallidal tone thus decreases inhibition on thalamocortical neurons and thereby facilitates movement. Thus, motor plans are filtered through the basal ganglia to modulate the outputs of the motor cortex. It is thought that this circuit is important for regulating the dimensions of movements (e.g., tracing small or larger versions of one's signature) and for improving fluidity in opposite or alternating motions. There is an important ascending activation of striatal activity by nigrostriatal dopaminergic axons, in parallel with an important dopaminergic innervation of the premotor cortex. This input decreases in Parkinson's disease due to degeneration of the dopaminergic neurons in the substantia nigra and ventral tegmental area (partly in midbrain and partly in the basal/floor plate of diencephalon). These patients have hypokinesia (difficulty initiating movements), rigidity, and inexpressive faces. On the other hand, lesion of the subthalamic nucleus (p4 basal plate) diminishes the tonic inhibitory output of the pallidum, which leads to hyperactivity symptoms (involuntary movements are released, e.g., tics, hemibalismus, and choreoathetosis).

The amygdala is caudal to the striatum and pallidum. This massive nuclear formation, formed at the primitive caudal telencephalic pole and brought to the tip of the temporal lobe during morphogenesis (Figs. 6 and 7), can be divided into pallial and subpallial parts according to the expression of genetic markers and other characteristics (Fig. 8B). The numerous amygdaloid subnuclei traditionally have been grouped into central, medial, basal, lateral, and cortical groups (Fig. 8B). The pallial amygdaloid components have considerably increased in size in highly evolved mammals. The subpallial part of the amygdala is represented by the central amygdaloid nucleus, which shares some striatal characteristics, and the medial amygdaloid (plus amygdaloid BST) nuclei, which share some pallidal characteristics.

A recent variant conception of the subpallial amygdala adds the so-called "extended amygdala," which includes all the intratelencephalic supracapsular and subcapsular BST formation (relations to internal capsule) and other specific cell groups related to the anterior commissure. The extended amygdala, however, would belong to the middle and perhaps also the rostral subpallial sectors since it converges on the anterior commissure. According to our embryological formulation, the extended amygdala largely corresponds to pallidal areas surrounding the globus pallidus (Fig. 7D).

Finally, the rostral subpallial sector is constituted largely by the septum, found in front of or partially mixed with the telencephalic commissures in the medial hemispheric wall (See in Figs. 6E, 6F, and 8A). This nuclear complex can again be divided by use of genetic markers and other characteristics into subpallial and pallial portions (the pallial one being very small). The subpallial portion subdivides into an upper, striatum-like part and a ventral, pallidum-like part. Traditionally, the septum has been divided into "medial" and "lateral" septal nuclei, which embryologically correspond to superficial and periventricular strata, respectively. It is unfortunate that the term "lateral" was associated with a periventricular formation since this is somewhat counterintuitive (most other lateral structures in the brain are superficial, i.e., subpial).

See Also the Following Articles

EVOLUTION OF THE BRAIN • HINDBRAIN • HYPOTHALAMUS • NEOCORTEX • NEUROANATOMY

Suggested Reading

Ahlheid, G. F., de Olmos, J. S., and Beltramino, C. A. (1995). *Amygdala and extended amygdala*. In *Rat Nervous System* (G. Paxinos, Ed.), 2nd ed., pp. 495–578. Academic Press, San Diego.

Anderson, S. A., Eisenstat, D., Shi, L., and Rubenstein, J. L. R. (1997). Interneuron migration from basal forebrain: Dependence on Dlx genes. *Science* **278,** 474–476.

Medina, L., and Reiner, A. (2000). Do birds possess homologues of mammalian primary visual, somatosensory and motor cortices? *Trends Neurosci.* **23,** 1–12.

Puelles, L. A. (1995). Segmental morphological paradigm for understanding vertebrate forebrains. *Brain Behav. Evol.* **46,** 319–337.

Puelles, L., and Medina, L. (1994). Development of neurons expressing tyrosine hydroxylase and dopamine in the chicken brain: A comparative segmental analysis. In *Phylogeny and Development of Catecholamine Systems in the CNS of Vertebrates* (W. J. A. J. Smeets and A. Reines, Eds.), pp. 381–404. Cambridge Univ. Press, Cambridge, UK.

Puelles, L., and Rubenstein, J. L. R. (1993). Expression patterns of homeobox and other putative regulatory genes in the embryonic mouse forebrain suggest a neuromeric organization. *Trends Neuro. Sci.* **16,** 472–479.

Puelles, L., and Verney, C. (1998). Early neuromeric distribution of tyrosine hydroxylase-immunoreactive neurons in human embryos. *J. Comp. Neurol.* **394,** 283–308.

Puelles, L., Kuwana, E., Bulfone, A., Shimamura, K., Keleher, J., Smiga, S., Puelles, E., and Rubenstein, J. L. R. (2000). Pallial and subpallial derivatives in the embryonic chick and mouse telencephalon, traced by the expression of the Dlx-2, Emx-1, Nkx-2.1, Pax-6 and Tbr-1 genes. *J. Comp. Neurol.* **424,** 409–438.

Rakic, P. (1995). A small step for the cell, A giant leap for mankind: A hypothesis of neocortical expansion during evolution. *Trends Neurosci.* **18,** 383–388.

Rubenstein, J. L. R., and Beachy, P. A. (1998). Patterning of the embryonic forebrain. *Curr. Opin. Neurobiol.* **8,** 18–26.

Rubenstein, J. L. R., Shimamura, K., Martinez, S., and Puelles, L. (1998). Regionalization of the prosencephalic neural plate. *Annu. Rev. Neurosci.* **21,** 445–478.

Shimamura, K., Hartigan, D. J., Martinez, S., Puelles, L., and Rubenstein, J. L. R. (1995). Longitudinal organization of the anterior neural plate and neural tube. *Development* **121,** 3923–3933.

Smith-Fernandez, A., Pieau, C., Repérant, J., Boncinelli, E., and Wassef, M. (1998). Expression of the *Emx-1* and *Dlx-1* homeobox genes define three molecularly distinct domains in the telencephalon of mouse, chick, turtle and frog embryos: Implications for the evolution of telencephalic subdivisions in amniotes. *Development* **125,** 2099–2111.

Swanson, L. W., and Petrovich, G. D. (1998) What is the amygdala? *Trends Neurosci.* **21,** 323–330.

Frontal Lobe

LADA A. KEMENOFF, BRUCE L. MILLER, and JOEL H. KRAMER
University of California, San Francisco

GLOSSARY

dysarthria Disturbance of speech articulation or impairment of the speech mechanism, including muscle weakness. It is manifested by slurred pronunciation.

dysprosody Loss of the normal rhythm, melody, and articulation of speech.

gliosis Glial cells migrate to and proliferate in areas of neural tissue where damage has occurred.

hemiplegia Paralysis of one side of the body.

hypophonia An abnormally weak voice resulting from uncoordination of speech muscles, including weakness of muscles of respiration.

paraphasia The production of unintended syllables, words, or phrases during speech.

regional cereberal blood flow (rCBF) Amount of blood flow in a region of the cortex is positively correlated to the metabolic activity of that region. Imaging of the rCBF by scintigraphy with inhaled xenon-133 can be combined with psychological testing during the measurement of blood flow, allowing assessment of the effects of cognitive activation procedures on blood flow in specific regions of the cortex.

Commonly described as the anatomic seat of human self-awareness, the frontal lobes are the most evolutionarily advanced components of the human brain. Scientific advancements during the past decade have considerably improved our understanding of the frontal lobes and their complex role in cognition, personality, and neurological disease. This article presents a contemporary perspective on frontal lobe neuroanatomy; neuropsychological functions; and frontal lobe disorders.

I. NEUROANATOMY OF THE FRONTAL LOBE

The frontal lobes comprise the most anterior portion of the cerebral hemispheres. They are demarcated posteriorly by the central sulcus, laterally by the Sylvian fissure, and medially by the cingulate sulcus. The frontal lobes are anatomically and functionally heterogeneous, and the lateral surface of the frontal lobes can be divided into three major functional sectors: primary motor, premotor, and prefrontal cortex.

A. Primary Motor Cortex

The most posterior region of the frontal lobe, the precentral gyrus, represents the brain's primary motor area. This region forms a narrow strip of tissue along the lateral surface of the frontal lobe and continues down around the medial bank of the cortical apex. Primary motor cortex gives rise to the corticobulbar and corticospinal tracts and is responsible for the mediation of movement.

B. Premotor Cortex

Positioned immediately anterior to the primary motor region, the premotor cortex is composed of several functional areas. The lateral premotor area appears to be involved in the integration of motor skills and

learned action sequences. The supplementary motor area on the medial surface in the superior frontal gyrus appears to mediate the initiation and programming of body movements. Broca's area is situated in the inferior, posterior frontal gyrus and is responsible for controlling voluntary speech.

C. Prefrontal Cortex

The prefrontal cortex is composed of the anterior portion of the frontal lobes. This region has robust connections with limbic and subcortical areas. In 1993, Jefferey Cummings described three frontal–subcortical circuits that mediate cognitive, motivational, and emotional processes: the dorsolateral prefrontal circuit, the orbitofrontal circuit, and the anterior cingulate circuit.

The dorsolateral prefrontal circuit includes the lateral convexity of the frontal lobe, dorsolateral caudate, portions of the globus pallidus and substantia nigra, and ventral anterior and dorsomedial thalamic nuclei. The dorsolateral circuit subserves executive functioning abilities, including response inhibition, fluency, working memory, and retrieval from long-term memory.

Jefferey Cummings also proposed an orbitofrontal circuit composed of two parallel subcircuits: the lateral and medial orbitofrontal circuits. The lateral orbitofrontal circuit projects from orbitofrontal cortex to ventral portions of the caudate; the medial circuit projects to the ventral striatum. Both circuits then project to medial portions of the globus pallidus, midbrain structures, and ventrolateral and dorsomedial thalamus. The orbitofrontal circuit mediates the modulation of social behavior; lesions can produce personality changes, including indifference to others, irritability, tactlessness, and impulsive behavior.

The anterior cingulate circuit projects from the anterior cingulate cortex and incorporates the ventromedial caudate, ventral putamen, and nucleus accumbens. This circuit is thought to mediate motivation; lesions are associated with apathy and disinterest.

II. NEUROPSYCHOLOGICAL FUNCTIONS OF THE FRONTAL LOBE

A. Executive Functioning

Damage to the dorsolateral prefrontal cortex has been associated with compromised performance on neuropsychological measures sensitive to executive functioning. According to Muriel Deutsch Lezak, the term "executive functions" can be defined as the capacities that enable a person to engage in purposive, independent, and self-serving behavior. These functions include the ability to plan, to disengage from the immediate environment, to show flexibility of thinking, to fluently generate concepts, and to inhibit responses to overlearned patterns of behavior. Many researchers contend that executive functioning abilities remain among the most highly developed of human frontal lobe accomplishments.

B. Mental Flexibility

Mental flexibility requires the capacity to shift a course of thought or action according to rapidly changing situational demands. The Wisconsin Card Sorting Test (WCST) is a popular neuropsychological measure used to assess concept formation, abstract reasoning, and the ability to shift cognitive strategies in response to changing environmental contingencies. The subject is presented with four stimulus cards depicting figures of varying forms, colors, and numbers of figures (one red triangle, two green stars, three yellow crosses, and four blue circles). The task is to match each consecutive card from a deck of similar stimulus cards with one of the four key cards. In response to each matching, the subject is told only whether his or her choice was correct or incorrect. After the subject makes a specified number of correct matches, without warning the sorting strategy is changed. The subject is therefore required to use the examiner's feedback to develop a new sorting strategy.

Patients with frontal lobe damage have been found to perform more poorly on the WCST task than patients with nonfrontal damage. The required shifting response is particularly challenging for these patients because it entails the use of mental flexibility and reasoning skills. Frontal lobe patients commonly make perseverative errors by continuing to sort by a certain principle (e.g., by color) long after that sorting principle has been changed (e.g., to form). In some cases, patients appear almost oblivious to feedback and continue to make erroneous perseverations, despite their ability to verbalize the correct sorting strategy. Recent work by Kyle Boone suggests that perseverations on the WCST are more strongly associated with right frontal damage than left frontal damage.

The California Card Sorting Test is another useful measure of concept formation, shifting, and reasoning.

This unique card-sorting task was designed to isolate and measure specific components of problem-solving ability. The subject is asked to sort six cards spontaneously into two groups of three cards each, according to as many different rules as possible, and to report the rule after each sort. In another condition, subjects are required to report the rules for correct sorts performed by the examiner. This test examines several executive components, including the ability to generate and initiate different sorts, the ability to verbalize the principles of accurate sorts, and the ability to inhibit perseverative sorts. In 1992, Dean Dellis and colleagues found that patients with focal frontal lobe lesions and patients with Korsakoff's syndrome were impaired on eight of nine components of the task. Based on these results, the authors suggested that a wide array of deficits in abstract thinking, cognitive flexibility, and use of knowledge to regulate behavior contribute to the problem-solving impairment of patients with frontal lobe dysfunction.

C. Response Inhibition

Another important aspect of executive functioning is the ability to inhibit responses to established patterns of behavior. The Stroop Test is considered a measure of a person's ability to inhibit a habitual response in favor of an unusual one. During the interference condition of the Stroop, subjects are presented with a list of colored words (blue, green, red, etc.) printed in nonmatching colored ink. For example, the word *blue* may be printed in red ink, and the word *green* may be printed in blue ink. The subject's task is to name the ink color in which the words are printed as quickly as possible. This challenge involves suppressing the strong inclination to read the color name. Many patients with frontal lobe damage are unable to inhibit reading the words and thus show impairment on this task.

D. Verbal Fluency

Frontal lobe impairment can also produce deficits in a person's ability to rapidly generate words. The Controlled Oral Word Association Test (COWAT), also known as the "FAS," is a commonly used neuropsychological measure of verbal fluency. The COWAT consists of three word conditions. The subjects' task is to produce as many words as he can that begin with the given letter (F, A, or S) within a 1-min time period. Subjects are also instructed to exclude proper nouns, numbers, and the same word with a different suffix.

The COWAT and other measures of verbal fluency have proven to be sensitive indicators of frontal lobe dysfunction. In 1989, Jerry Janowsky, Arthur Shimamura, and Larry Squire found that patients with circumscribed left or bilateral frontal lobe lesions produced significantly fewer words than did control subjects. Other researchers found that left frontal lesions resulted in lower word production than right frontal ones. Similarly, regional cerebral blood flow findings have shown left-sided frontal activation during the performance of verbal fluency tasks.

Frontal lobe damage can also impair performance on visual or design fluency tasks. Design fluency tests were developed as visual analogs to verbal fluency measures such as the FAS. The subjects' task is to generate as many unique designs as they can within a given time period. Although the left prefrontal region appears to be specialized in using verbal material, Christina Elfgren and Jarl Risberg reported bilateral frontal lobe activation during the performance of visual generation tasks.

E. Planning

Patients with frontal lobe dysfunction have been reported to demonstrate impairments in the ability to plan. A wide spectrum of neuropsychological measures have been designed to assess numerous aspects of planning behavior. The Porteus Maze Test is a maze tracing task commonly used to assess planning and foresight. The subject's task is to trace the maze without entering any blind alleys. Performance level is usually measured on the basis of completion time and the test age level of the most difficult task the subject is able to successfully complete. In 1991, Harvey Levin and colleagues reported that patients with frontal lesions solved the Porteus mazes more slowly than severely injured nonfrontal head trauma patients and control subjects.

A number of tower puzzles have been designed to gauge more abstract forms of planning ability. Some of the most popular are the Tower of London, Tower of Hanoi, Tower of Toronto, and Tower of California. In all these tasks, the subject sees a set of pegs on which a number of beads or disks are placed in an initial starting position. The subject is instructed to move the disks to the appropriate pegs in order to reach a predetermined goal state. Two common test rules

include only moving one disk at a time and never placing a larger disk on top of a smaller disk.

In 1991, Tim Shallice and Paul Burgess found that patients with predominantly left anterior lesions performed more poorly on the Tower of London test than patients with posterior lesions and normal controls. Guila Glosser and Harold Goodglass reported similar results using the Tower of Hanoi: Patients with anterior lesions performed worse than the patients with posterior lesions.

Planning ability can also be measured by an individual's ability to prepare and execute target behaviors required for simulated real-life situations and events. In 1998, Eliane Miotto and Robin Morris developed the Virtual Planning Test with the intention of investigating the planning and organizational abilities of patients with frontal lobe neurosurgical lesions. The simulated planning tasks involved preparations for a fictional "trip" abroad or planning events that related to the subject's immediate environment. The frontal lobe patients were found to be impaired on this task and showed a tendency to select inappropriate activities associated with their immediate environment.

F. Memory

The frontal lobes play an important role in adequate memory functioning. Conventional research suggests that frontal lobe lesions do not produce a primary deficit in memory per se, but that they interfere with critical memory processes. Although encoding deficiencies can be observed in frontal lobe patients, they usually occur secondarily to executive functioning impairments. Thus, relatively speaking, frontal lobe damage does not significantly impair the ability to memorize material but does interfere with the ability to organize, attend to, and spontaneously retrieve information.

G. Retrieval

On measures of recent verbal memory, patients with frontal lobe impairment often demonstrate a pattern of poor recollection or retrieval of information in the context of relatively preserved recognition memory. In 1994, Donald Stuss and colleagues examined the memory abilities of patients with prefrontal, temporal, and diencephalic lesions. Patients with medial temporal and diencephalic lesions performed most poorly and did not improve with cueing. Patients with lateral temporal and prefrontal pathology scored worse in the free recall condition than the controls, but they improved with cueing. In addition, their recognition memory performance was relatively better than their free recall performance.

In 1995, Felicia Gershberg and Arthur Shimamura found similar results in their examination of free recall strategies in a group of patients with unilateral frontal lobe lesions. An example of a typical memory strategy is the rehearsal of associations between words presented in a list learning task. This kind of organizational ability is thought to require considerable executive control. As such, one would expect diminished use of such strategies on the part of patients with frontal lobe impairment. Consistent with their hypothesis, Gershberg and Shimamura found that the frontal lobe patients demonstrated impaired free recall ability and reduced use of subjective organization strategies. The frontal patients also benefited from strategy instruction (e.g., category cues) at either study or test, suggesting that both encoding and retrieval processes may be impaired by frontal lobe damage. Based on these findings, the authors suggested that retrieval impairments by patients with frontal lobe lesions might be partly due to deficits in the use of organizational strategies.

Impaired encoding and retrieval functions in prefrontal syndromes can also be related to deficits in working memory. Allan Baddeley uses the term "working memory" to describe the "scratch pad" of the human memory system—the place where information can be held and manipulated while new information is being processed. Researchers suggest that the prefrontal cortex plays a critical role in these complex working memory abilities. In support of this view, several functional imaging studies have reported activation of the dorsolateral prefrontal cortex during both auditory and visual mental working tasks.

H. Temporal Sequencing

Frontal lobe patients have also demonstrated deficiencies in their ability to integrate temporally separated events. Temporal sequencing can be tested by assessing an individual's ability to recall or recognize the temporal order of recently presented lists of words, abstract designs, or pictures. Patients with prefrontal cortex damage have been shown to demonstrate impairment in recalling the temporal order of such items but have no item recognition problems for the same list of words, designs, or pictures. In 1994,

Raymond Kesner, Ramona Hopkins, and Bonnie Fineman conducted a study in which a group of prefrontal cortex-damaged patients were tested for item and order recognition memory for spatial location, word, abstract picture, and hand position information. Compared to controls, frontal patients showed severe deficits on all order recognition tests. However, relative to controls, the frontal patients showed no deficits in their ability to recognize these same items with the exception of a deficit for hand position. With regard to laterality, order recognition memory deficits for spatial location occurred for right and bilateral but not left prefrontal cortex-damaged patients. Both right and left prefrontal cortex subjects showed order recognition memory deficits for words, abstract pictures, and hand position.

I. Source Memory

In addition to deficits in sequencing temporal information, patients with frontal lobe dysfunction have been reported to show deficits in identifying the source of their knowledge. Thus, an individual with source amnesia may be able to remember a fact but will forget where and when that fact was learned.

In a 1989 study, Jerry Janowsky and colleagues found that patients with frontal lobe lesions exhibited impaired source memory for facts acquired in a recent test session, even though their memory for the facts was normal. During the first phase of the experiment, subjects were asked to learn some general information facts (not known prior to the study session). After a 6 to 8 day retention interval, subjects were asked to answer some additional questions. Some questions were the previously nonrecalled items from the initial study phase, some were new items from the same level of difficulty, and others were easy questions that had not been presented previously. When subjects correctly answered a question, they were asked to recollect where and when the information had been learned. Compared to their age-matched controls, the frontal lobe group made significantly more errors by attributing learned information to an incorrect source. Results of this study suggest that the frontal lobes play an important role in one's ability to associate information in memory to the context in which it was acquired.

J. Autobiographical Memory

The active process of recollecting information from one's earlier life requires the use of executive abilities such as attention, flexible searching, and organization. Given the dysexective syndrome characteristic of frontal lobe patients, it is not surprising that individuals with prefrontal cortex damage occasionally exhibit impaired autobiographical memory. Assessments of autobiographical memory usually involve a lengthy series of questions covering the major periods of a typical life span (childhood, adolescence, and adulthood). Subjects are asked to describe specific events in great detail, including the dates and the importance of the events, and to indicate the names of friends and family members who may have participated in the events. For example, a subject may be asked to remember the date of his or her wedding or to provide the name of the preschool that he or she attended. The subject's family members later corroborate all reported information.

In a 1993 study, Sergio Della Salla and fellow researchers examined autobiographical memory retrieval in a group of patients with frontal lobe lesions. The battery of tests administered included an autobiographical memory enquiry and a series of executive functioning measures. Six of 16 frontal lobe patients were impaired on the autobiographical memory measure. Moreover, poor autobiographical retrieval correlated significantly with impaired performance on the executive functioning measures. Thus, impaired ability to retrieve information from autobiographical memory may have been related to the frontal lobe patient's inefficient organizational and searching ability.

K. Language

Frontal lobe lesions can result in a variety of language disturbances, including loss of grammar (aggramatism) and the production of unintended syllables, words, or phrases during speech (paraphasic errors). The following sections highlight some of the most extensively researched frontal lobe language disorders.

L. Broca's Aphasia

Broca's aphasia is one of the most commonly known syndromes of frontal language disorder. The core features of this syndrome include nonfluent, effortful speech production, semantic and phonemic paraphasias, articulatory errors, agrammatism, and relatively preserved comprehension. Widely accepted definitions of Broca's aphasia also include poor repetition, reading, and writing ability.

The lesion in classical Broca's aphasia involves the left posterior, inferior frontal gyrus. With the advent of sophisticated neuroimaging techniques, researchers have discovered that circumscribed damage to Broca's area does not necessarily result in the complete syndrome of Broca's aphasia. Moreover, some individuals with Broca's aphasia do not have lesions in Broca's area. Therefore, it seems that the underlying pathology in Broca's aphasia can be relatively extensive and varied. Regions including the inferior central Rolandic area, the insula, subcortical regions, and the anterior parietal regions have also been implicated in this language syndrome.

M. Pure Motor Aphasia

Verbal apraxia or pure motor aphasia refers to the articulatory and prosodic disturbance of language output in the absence of the agrammatic component. The underlying lesion is said to involve the left lower motor cortex and posterior operculum. Recently, Nina Dronkers emphasized the relationship between articulatory deficits and damage to the precentral gyrus of the insula.

This clinical syndrome is characterized by impaired articulation, slow and effortful speech, segmentation, phonemic paraphasias, dysprosody, and occasional hypophonia. The outcome for pure motor aphasia ranges from full recovery to a status of normal language with persistent dysarthria and/or dysprosody.

N. Transcortical Motor Aphasia

Transcortical motor aphasia (TCMA) involves lesions of the left frontal lobe–supplementary motor area (SMA), just anterior and superior to Broca's area. During acute phases, patients may initially present as mute but later develop a clinical profile characterized by normal repetition and comprehension, with limited, slow, and perseverative spontaneous speech.

Some researchers hypothesize that the SMA represents the "starting mechanism" or center for initiation of speech. Others suggest that damage to the SMA results in a lack of a plan or program to carry out voluntary speech. In their 1984 study, Morris Freedman and colleagues studied 15 patients with TCMA or near variants of the syndrome. Consistent with TCMA literature, these researchers concluded that small lesions to SMA cause a pure disorder of speech initiation. Furthermore, damage to fibers from SMA to premotor cortex may disconnect the limbic starter mechanism of speech from the cortical regions that control the motoric aspect of speech.

O. Frontal–Subcortical Aphasias

Language deficits can also result from lesions to frontal–subcortical brain regions. Lesions of the putamen and internal capsule will produce a nonfluent language disorder that at times resembles Broca's aphasia. In 1982, Margaret Naeser and her colleagues examined nine cases of subcortical aphasia with capsular/putaminal (C/P) lesions documented by computed tomographic scans. Patients with C/P lesion sites with anterior-superior white matter lesion extension had good comprehension, grammatical but slow dysarthric speech, and lasting right hemiplegia.

III. NEUROPSYCHIATRY OF THE FRONTAL LOBE

A. Personality

One of the most well-known early cases of behavioral change following frontal lobe damage is that of Phineas P. Gage. On September 13, 1848, this young man became a victim of a bizarre accident in which a tamping iron rod caused severe damage to his frontal lobes. Prior to sustaining this injury, Phineas was described as a reserved, intelligent, and responsible individual. Following the mishap, he transformed into an irreverent, irresponsible, and careless human being. This famous case and many modern counterparts that followed sparked considerable interest in the wide spectrum of personality changes seen in frontal lobe pathology.

The diminished capacity to modulate emotional behavior has frequently been associated with damage to the frontal lobes. In the context of traumatic brain injury, the changes customarily involve either an exaggeration or muting of emotions. In some cases, previously outgoing and sociable individuals become apathetic, emotionally flat, disengaged, and withdrawn. On the opposite end of the spectrum, patients can demonstrate uncharacteristically aggressive, impulsive, disinhibited, and socially inappropriate behavior. The apathetic syndrome has been localized to damage to the convexity and medial aspects of the frontal lobes, whereas the aggressive and disinhibited

syndrome has been attributed to the orbitofrontal regions.

A disturbance in self-awareness represents another common feature associated with frontal lobe damage. Deficient awareness may be reflected by the inability to appreciate performance errors on neuropsychological testing or to recognize the impact of one's impairments on others. Patients with compromised frontal lobe functioning have also been reported to demonstrate a failure to appreciate social and interpersonal norms. In an attempt to explain these deficits, some authors have suggested that frontal lobe pathology may actually impair an individual's personal consciousness or sense of self. Endel Tulving and colleagues used the term "autonoetic awareness" (awareness of oneself as a continuous entity across time) to describe this phenomenon. According to Donald T. Stuss, in this context disturbed self-awareness is "not lack of knowledge but impaired judgment of the objective facts in relation to one's own life."

IV. DISEASES OF THE FRONTAL LOBE

A. Frontotemporal Dementia

Frontotemporal lobar degeneration (FTLD) is a common cause of early onset dementia. There are three clinical syndromes that occur in FTLD: frontotemporal dementia (FTD), progressive nonfluent aphasia, and semantic dementia. The following sections are limited to a discussion of the core behavioral, neuropsychological, and pathological features of FTD.

1. Behavioral Features of FTD

The early stage of FTD is characterized by marked changes in personality and behavior. Patients typically become disinhibited, apathetic, and irritable. Gross disinhibition often leads to impulsivity, increased aggressiveness, and antisocial behavior. Antisocial acts can range from inappropriate comments to various forms of criminal behavior. In a 1997 study, Bruce Miller and colleagues compared the presence of antisocial conduct in FTD versus Alzheimer's dementia (AD) patients. Almost 50% of the FTD patients were involved in behaviors such as stealing, physical assault, public urination or masturbation, and unethical job conduct. In contrast, only one of the AD patients manifested these types of behaviors.

According to the FTD consensus criteria proposed by David Neary and colleagues in 1998, declines in social graces and decorum are commonly manifested in tactlessness and a loss of interpersonal etiquette and social awareness. For example, FTD patients often demonstrate increased talking, inappropriate laughter, singing, and invasion of personal space. Improper hygiene, grooming, and other aspects of personal awareness are also observed. These deficits in social and personal awareness are further complicated by the FTD patients' lack of insight into their behavioral disturbance and its impact on others.

FTD patients often lack empathy and are typically described as emotionally indifferent and apathetic. Some patients become self-absorbed and exhibit less concern about the welfare of friends and family members. Other characteristic affective symptoms in FTD include depression, anxiety, and psychotic features. Patients with FTD can also demonstrate hyperorality and dietary changes. Hyperorality may manifest in overeating, bingeing, and excessive consumption of liquids and/or alcohol. Some patients develop a strong preference for sweets and may also place inanimate objects in their mouths.

FTD patients progressively develop changes in speech production and structure. During the early phases of the disease patients may demonstrate excessive talking, frequent use of stereotyped phrases, and echolalia. Later stages may be characterized by mutism.

2. Neuropsychological Features of FTD

According to the 1998 consensus criteria proposed by David Neary and colleagues, the neuropsychological profile for FTD is characterized by significant impairments on frontal lobe measures in the context of relatively preserved memory, language, and visuospatial skills. Frontal lobe measures are tests designed to assess executive functioning abilities, such as planning, mental flexibility, verbal fluency, abstraction, and the ability to inhibit responses to overlearned patterns of behavior. Popular standardized neuropsychological measures of executive functioning include the WCST, Stroop Test, and Trail Making Test.

A number of studies have examined the patterns of executive functioning deficits found in FTD. In a 1996 investigation, Nancy Pachana and colleagues documented the neuropsychological changes that distinguish FTD patients from AD patients. They found that AD patients exhibited relatively greater

impairment on memory than executive tasks, whereas the FTD group showed the opposite pattern.

FTD patients may perform poorly on measures of recent memory, language, and visuospatial tasks. However, these deficits are thought to occur secondarily to executive dysfuntion, such as inattention, inefficient organizational strategies, and poor self-monitoring. For example, on measures of recent verbal memory, FTD patients tend to have difficulty spontaneously recalling information (free recall), whereas their ability to recognize these same items is relatively preserved.

3. Pathological Features of FTD

According to the pathological criteria proposed by Arne Brun and colleagues in 1994, FTD results from bilateral and for the most part symmetrical distribution of pathology in the frontal and anterior temporal lobes. Upon inspection, mild atrophy of these brain regions is commonly observed. At the microscopic level, a degenerative process is noted in the frontal or frontotemproal cortex, characterized by microvaculation (spongiosis), neuronal loss, and gliosis. The presence of argyrophillic, ubiquitin positive inclusion bodies (Pick bodies) and swollen, achromatic neurons (Pick cells) can also be observed. The clinical syndrome of FTD can be seen with and without the presence of Pick bodies and Pick cells.

B. Traumatic Brain Injury

Brain damage following traumatic brain injury, particularly closed head injury (CHI), frequently involves the orbital and polar aspects of the frontal lobes. Commonly observed behavioral features of CHI following frontal lobe damage include posttraumatic amnesia, attentional deficits, and changes in personality. Posttraumatic amnesia (PTA) is thought to result from a disconnection or damage of the basal forebrain and orbitofrontal cortex. PTA usually occurs during the period of acute recovery following a severe head injury. Patients typically experience confusion and disorientation, and tend to confabulate, and have difficulty learning and remembering information. The duration of PTA usually depends on the severity and the outcome of the brain injury.

Attentional deficits are a relatively common behavioral consequence of severe head injury. Distractibility, problems with concentration, inability to focus, and an overall slowing are commonly reported complaints. Attention is considered a complex cognitive process composed of a number of components, including phasic alertness, selective attention, and sustained attention. Several studies of CHI patients suggest that individual attentional components can be selectively impaired following frontal lobe damage. For example, investigators have linked the midfrontal regions to deficits in sustained attention. Others attribute impairments in selective attention to damaged frontal–thalamic and subcortical structures.

Traumatic brain injuries that damage the frontal and temporal lobes can also result in profound changes in personality and social adjustment. Victims of traffic and other common head injury accidents frequently exhibit reduced drive, decreased awareness, blunted affect, and a lack of initiative. This apathetic syndrome is attributed to damage of the medial frontal lobes. Closed head injuries can also produce what is described as a euphoric syndrome, characterized by impulsivity, sexual disinhibition, socially inappropriate behavior, and aggressiveness. The orbitofrontal regions have been implicated in this class of personality deficits.

See Also the Following Articles

APHASIA • EVOLUTION OF THE BRAIN • LANGUAGE, NEURAL BASIS OF • MEMORY NEUROBIOLOGY • MODELING BRAIN INJURY/TRAUMA • MOTOR CORTEX • NEUROANATOMY • PHINEAS GAGE • SPEECH • TIME PASSAGE

Suggested Reading

Benton, A. L. (1968). Differential behavioral effects in frontal lobe disease. *Neuropsychologia* **6**(1), 53–60.

Brun, A., Englund, B., Gustafson, L., Passant, U., Mann, D. M. A., and Snowden, J. S. (1994). Clinical and neuropathological criteria for frontotemporal dementia. *J. Neurol. Neurosurg. Psychiatr*. **57,** 416–418.

Damasio, A. R., and Anderson, S. W. (1993). *The frontal lobes*. In *Clinical neuropsychology* (E. Kenneth, M. Heilman, E. E. Valenstein, *et al*., Eds.), 3rd ed., pp. 409–460. Oxford University Press, New York.

Damasio, H., Grabowski, T., Frank, R., Galaburda, A. M., and Damasio, A. R. (1994). The return of Phineas Gage: Clues about the brain from the skull of a famous patient. *Science* **264,** 1102–1105.

Levin, H. S., Goldstein, F. C., Williams, D. H., and Eisenberg, H. M. (1991). *The contribution of frontal lobe lesions to the neurobehavioral outcome of closed head injury*. In *Frontal Lobe Function and Dysfunction* (E. Harvey, S. Levin, E. Howard, M. Eisenberg, *et al*., Eds.), pp. 318–338. Oxford University Press, New York.

Levine, B., Black, S. E., Cabeza, R., Sinden, M., McIntosh, A. R., Toth, J. P., Tulving, E., and Stuss, D. T. (1998). Episodic memory and the self in a case of isolated retrograde amnesia. *Brain* **121**(10), 1951–1973.

Lezak, M. D. (1995). *Neuropsychological Assessment*, 3rd ed. Oxford University Press, New York.

Miller, B. L., and Cummings, J. L. (1999). *The Human Frontal Lobes: Functions and Disorders*. Guilford, New York.

Miller, B. L., Cummings, J. L., Villanueva-Meyer, J., Boone, K., Mehriinger, C. M., Lesser, I. M., and Mena, I. (1991). Frontal lobe degeneration: Clinical, neuropsychological and SPECT characteristics. *Neurology* **42,** 1374–1382.

Miller, B. L., Chang, L., Mena, I., Boone, K. B., and Lesser, I. (1993). Progressive right frontotemporal degeneration: Clinical, neuropsychological and SPECT characteristics. *Dementia* **4,** 204–213.

Neary, D., Snowden, J. S., Gustafson, L., Passant, U., Stuss, D., Black, S., Freedman, M., Kertesz, A., Robert, P. H., Albert, M., Boone, K., Miller, B. L., Cummings, J., and Benson, D. F. (1998). Frontotemporal lobar degeneration: A consensus on clinical diagnostic criteria. *Neurology* **51**(6), 1546–1554.

Perecman, E., and the Institute for Research in Behavioral Neuroscience (1987). *The Frontal Lobes Revisited*. IRBN Press, New York.

Prigatano, G. P., and Schacter, D. L. (Eds.) (1991). *Awareness of Deficit after Brain Injury: Clinical and Theoretical Issues*. Royal Chem. Soc., London.

Stuss, D. T., and Benson, D. F. (1984). Neuropsychological studies of the frontal lobes. *Psychol. Bull.* **95**(1), 3–28.

Stuss, D. T., and Benson, D. F. (1986). *The Frontal Lobes*. Raven Press, New York.

Functional Magnetic Resonance Imaging (fMRI)

ROBERT L. SAVOY
MGH/MIT/HST Athinoula A. Martinos Center

GLOSSARY

block design Experimental design for functional neuroimaging in which an attempt is made to put the subject's brain in a steady state of activity by using the same type of task for an extended period of time (typically 20–60 sec) and then comparing the brain activation during that block with other blocks that use a different task.

blood oxygen level dependent (BOLD) Refers to a general method of magnetic resonance imaging (MRI) for detecting changes in the nuclear magnetic resonance (NMR) signal that are caused by the varying concentration of deoxyhemoglobin, locally, in the blood near a part of the brain.

event-related design Experimental design for functional neuroimaging in which individual, brief (typically 1–2 sec in duration) stimuli of different types are presented in random order, and where the evoked responses for many such trials of a given type are averaged together to detect a measureable response.

flow via alternating inversion recovery (FAIR) Refers to a specific method of MRI for detecting changes in the NMR signal that are caused by the varying flow, locally, of blood in arteries near a part of the brain.

functional magnetic resonance imaging (fMRI) The use of MRI to detect changes in blood flow and blood oxygenation associated with local changes in neuronal activity in the brain.

gradient magnets Part of the technology of MRI used for supplying strong, operator-controlled linear gradients of magnetic field to enable the generation and detection of the NMR signal associated with a specific point in three-dimensional space.

hemodynamics Changes in the properties (volume, flow rate, and chemical composition) of blood over time.

magnetic resonance imaging (MRI) The use of a variety of operator-controlled electromagnetic fields to generate a NMR signal that can be associated with a particular point in space.

nuclear magnetic resonance (NMR) The physical phenomenon of absorption and reemission of electromagnetic energy associated with the quantum mechanical spin and magnetic field of the nuclei of some atoms.

principle component analysis (PCA) The re-representation of multidimensional data into a collection of components (sometimes called eigenimages and eigenvectors) via an algorithm that accounts for the most variance by the first principal component, the second most by the second component, etc.

retinotopy The regular spatial arrangement of the receptive fields of cortical neurons in many parts of the visual cortex that follows, in a systematic way, the two-dimensional spatial arrangement of the retina.

talairach coordinates The most widely used convention for orienting and scaling human brains to facilitate the averaging and/or comparing of data across multiple subjects.

Functional magnetic resonance imaging (fMRI) refers to the use of the technology of magnetic resonance imaging (MRI) to detect the localized changes in blood flow and blood oxygenation that occur in the brain in response to neural activity. This article presents the basics of fMRI-based research, including the physical and biophysical bases of the signals, the current

developments in experimental design and data analysis, and other practical considerations attendant to the technique, and also provides an overview of the broad range of scientific and clinical questions to which fMRI is being applied.

I. INTRODUCTION

It has long been known that there is some degree of localization of function in the human brain, as indicated by the effects of traumatic head injury. Work in the middle of the 20th century, notably the direct cortical stimulation of patients during neurosurgery, suggested that the degree and specificity of such localization of function was far greater than had earlier been imagined. One problem with the data based on lesions and direct stimulation was that the work depended on the study of what were, by definition, damaged brains. During the second half of the 20th century, a collection of relatively noninvasive tools for assessing and localizing human brain function in healthy volunteers led to an explosion of research in what is often termed "brain mapping." The tool that has been developing the most rapidly, and the tool that currently supplies the best volumetric (three-dimensional) picture of activity in the human brain, is fMRI.

Functional MRI uses the physical phenomenon of nuclear magnetic resonance (NMR) and the associated technology of MRI to detect spatially localized changes in hemodynamics that have been triggered by local neural activity. It has been known for more than 100 years that neural activity causes changes in blood flow and blood oxygenation in the brain, and that these changes are local to the area of neural activation. Techniques using radioactive tracers were developed in the mid-20th century to detect metabolic activity correlated with neural activation and to detect blood volume changes correlated with neural activation. In the early 1990s the technique of MRI was successfully adapted to measuring some of these effects noninvasively in humans. The development of fMRI led to a dramatic increase in neuroscience research in human functional brain mapping across the spectrum of psychological functions—from sensation, perception, and attention to cognition, language, and emotion—in both normal and patient populations.

Functional MRI makes the future of functional brain imaging particularly exciting for at least three reasons. First, fMRI does not involve ionizing radiation, and therefore it can be used repeatedly on a single subject and even on child volunteers. This permits longitudinal studies and it permits improvement in signal-to-noise ratios if the task being used elicits the same general response when repeated multiple times. Second, technical improvements in fMRI (due to more powerful magnets, more sophisticated imaging hardware, and the development of new methods of experimental design and data analysis) promise to yield improvements in spatial and temporal resolution for the technique. Third, there is a growing effort to integrate the findings based on fMRI with those from other techniques for assessing human brain function, such as electroencephalography (EEG) and magnetoencephalography (MEG), which inherently have much greater temporal resolution. It is likely that major advances in functional brain imaging will be made in the near future, but the associated technologies are complicated. In particular, to understand the technique of fMRI, one must consider a collection of interrelated issues, from physics and physiology to the practicalities of experimental design, data analysis, safety, and costs.

II. PHYSICS AND PHYSIOLOGY

MRI operates by creating and detecting a signal generated by the physical phenomenon of NMR. Although an in-depth explanation of the physics of NMR and the technology of MRI is beyond the scope of this article, a basic understanding of how a MR scanner operates is useful in discussing fMRI in its applications to psychology and medicine. The following discussion is based on a description of the "classical electrodynamics" picture of NMR and MRI. A more rigorous and complete treatment of NMR would require quantum mechanics, but MRI is best understood in terms of classical electrodynamics.

A. The NMR Signal

To begin an MRI session, the subject is placed horizontally into the bore of a high-field magnet. In a typical clinical MRI system this magnet has a field strength of 1.5 T. A small fraction of the hydrogen nuclei (single protons) of the water molecules in the body of the subject become aligned with the field of this magnet. (Many other atoms and nuclei with magnetic moments are also aligned, but for the purposes of this article and almost all fMRI applications, it is the hydrogen nuclei of water molecules that are the source

of the signal.) The hydrogen nuclei are oriented in a random collection of directions, relative to the main magnetic field, but there is a small statistical preference to have the longitudinal component of their orientations (i.e., the component in the direction of the main magnet) to be aligned with the main field. To the extent that a proton is not perfectly aligned with the main magnet, it will precess (spin) around the orientation of the main magnet. However, because the orientation of each proton is random (except for the component along the direction of the main field), and because it is only the randomly oriented transverse components (i.e., the components perpendicular to the main field) that generate a signal, the collection of spinning protons do not yield a net, detectable magnetic field.

Application of a radio frequency (RF) pulse of magnetic energy, presented at the frequency of the precession (i.e., the resonant frequency), causes all the hydrogen nuclei to change orientation (nutate). By controlling the power and duration of the RF pulse, the nuclei can be rotated to any desired angle relative to the main magnetic field. Typically, the parameters of the device are set so that the protons are rotated 90°. When the protons continue to precess in this new orientation, the net magnetic field that was originally induced in the body by the main magnet, and that was previously aligned with the main magnet, is now oriented 90° away and thus generates a detectable, changing magnetic field as it spins (precesses). A coil of wire around the subject will have a current generated within because of the changing magnetic field (Faraday's law). This current is the raw signal detected in a MRI scanner (Fig. 1).

B. Relaxation

The signal thus generated decays exponentially over time due to a number of processes. If the raw, exponential decay of the signal is measured (without doing anything else to create images), the process takes about 100 msec and the exponential time constant associated with that decay is conventionally called "T2*" (read "tee-two-star"). This decay in the measured signal is driven by a number of different physical processes.

First, the protons slowly (on the time scale of seconds for most brain tissue) realign with the main magnet. This is called longitudinal relaxation, and the time constant associated with this exponential process is called T1. Second, the signal generated by the collection of precessing protons is weakened by the fact that each individual proton experiences a slightly different *local* magnetic field due to interactions with nearby water molecules and other biological tissues and thus precesses at a slightly different frequency from its neighbors. With time (typically on a scale of tenths of seconds for most brain tissue) these protons get out of phase so that their respective magnetic fields are no longer lined up and therefore do not generate a detectable, macroscopic signal in the surrounding coil of wire. This is sometimes called the spin–spin component of transverse relaxation because it is based on the interaction of the spins (which imply magnetic fields) of nearby nuclei. If the magnetic field were perfectly uniform, then the net decay rate of the signal would be given by the exponential decay rate T2, which is driven by the combination of spin–spin transverse relaxation and the T1 longitudinal component. In the brain tissue of interest, T2 is almost entirely determined by the spin–spin relaxation.

In reality, there are other sources of magnetic field nonuniformity. Imperfections in the main magnet, variable magnetic susceptibility of the differing parts of the human body that has been inserted into the magnet, and changes in blood chemistry caused by externally injected "contrast agents" all contribute to nonuniformities in the magnetic field experienced by the precessing protons. Most important for fMRI, some chemicals that occur naturally in the body also distort the magnetic field. Deoxyhemoglobin is such a molecule, and as its local concentration is varied, the amount of distortion also varies. The rate of exponential decay of the NMR signal is influenced by all of these factors (Fig. 1).

C. Creating an Image

The preceding description yields a single number: T2*, the decay rate of the net NMR signal. This signal is conventionally called the free induction decay (FID) of the NMR signal because it is elicited when nothing else is done to the protons—that is, if the system were left free to decay at its own rate. However, this signal represents the net effect of inducing an NMR signal from the *entire volume of tissue* being subjected to the main magnet and the orientation-flipping RF pulse. To create an *image* in which different NMR signals are measured for different points in the three-dimensional volume, nonuniform magnetic fields are applied intentionally. The basic idea is to use linear magnetic field gradients, applied at various times and in various orientations, to distinguish the NMR signals arising

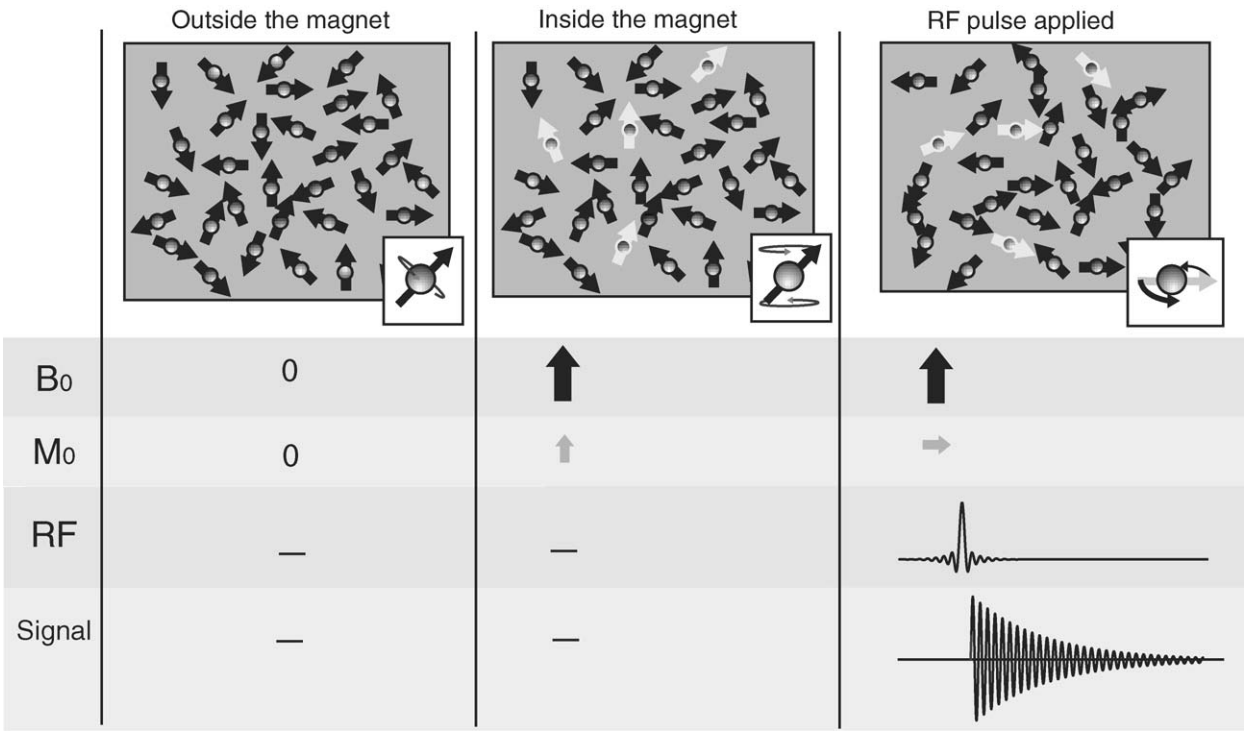


Figure 1 The basics of the NMR signal. The discussion of Section II.A is represented here in a simplified, cartoon form, with the story proceeding from left to right. Protons can be thought of conceptually as positively charged spheres that are always spinning, and this spin about an axis gives the proton an inherent orientation as well as a net magnetic moment along the axis of the spin. Before entering the magnet, the protons (with their magnetic field and spin direction indicated by arrows) are randomly oriented. There is no main field surrounding the body (B_0=0) and there is no induced field within the body (M_0=0). The body is placed in the main magnet ($B_0 \gg 0$). All the protons immediately start processing in a direction around B_0, but because the orientation and phases are random there is no net signal. After a few seconds, a small fraction of the protons change orientation to line up with B_0, which results in the creation of a net magnetic field in the body ($M_0 > 0$) oriented in the same direction as B_0. The individual protons are still precessing, now with a common net orientation, but the components of that rotation perpendicular to B_0 are still random so there is no detectable signal. A radio frequency (RF) pulse is applied for a brief period of time, causing all the protons to change their orientation by 90° so that the net induced magnetic field M_0 is now perpendicular to B_0. Now the individual precessing protons are aligned in such a way that the common component of orientation (M_0) goes around B_0 in a perpendicular direction and generates a macroscopically detectable current—the signal. The strength of that signal decreases exponentially with time for a variety of reasons described in Section II.B.

from different points in the three-dimensional volume. (The application of these nonuniform magnetic fields causes the raw NMR signal to decay even more rapidly than the FID, but the signal can still be measured, and various procedures can be used to create "echos" that enable the recovery of more information.)

The technology of MRI is based on the flexible (but complicated) application of multiple RF pulses and multiple gradients, synchronized precisely (and typically described in a pulse sequence diagram). The pulse sequence diagram indicates how a given slice of the brain is selected for imaging, how individual volume elements (voxels) are detected within each slice, and how the resulting signals are preferentially selected to obtain information about arterial blood flow or about the concentration of deoxyhemoglobin in venous blood flow. Some imaging pulse sequences use multiple RF-pulses in the generation of NMR signals that will yield a single plane of imaging data. Some imaging pulse sequences [such as echo-planar imaging (EPI)] generate data for an entire plane from a single RF pulse. EPI is rapid (with an entire plane collected in less than 50 msec), but it is associated with more expensive hardware and various limitations in spatial resolution or susceptibility to imaging artifacts and distortions.

D. Contrast in an Image

As with any imaging modality, the key variable in producing a meaningful image is contrast. The signal measured at one point in space or time must be higher

or lower than the signal at another point, and the variation in signal intensity across the image should systematically follow some variable of interest. In the endeavor of brain mapping, the ultimate variable of interest is neural activity. In order to measure local brain activity with MRI, one must exploit a chain of indirect linkages from neural activity (a constellation of electrical and chemical events) to changes in brain physiology and metabolism and finally to changes in the magnetic properties of substances within the brain.

Anything that causes a change in the NMR signal from a given voxel relative to other voxels at the same time is a source of *contrast* in the image. The density of protons in a given voxel (due to chemical composition) is one such source of contrast, though not an important one in fMRI. More commonly, it is the variation in rates of relaxation from voxel to voxel that generates contrast in the image.

In the earliest fMRI studies, exogenous contrasts (chemicals injected into the bloodstream of the subject) were used to obtain contrast. These blood-borne chemicals locally distorted the magnetic field, thus allowing increased blood perfusion to be detected. Subsequent studies demonstrated that endogenous contrast agents (i.e., naturally occurring molecules in the body, such as the concentration of deoxyhemoglobin in the blood) could also yield sufficient contrast between different states of neural activity. The use of endogenous contrast agents obviated the need for injecting foreign molecules into the bodies of normal (healthy) subjects, and this is one of the key reasons that fMRI has become so popular as a technique for assessing human brain function. In the short history of fMRI, a wide variety of techniques that produce various contrasts have been developed for detecting changes in brain physiology.

E. Neural Activation MRI

When neurons are active in the brain, there is an increase in blood flow and blood volume local to that region of activity. MRI can be used to detect the change in blood flow directly. The idea is that when fresh blood flows into the slice of the brain that is being imaged, it will have a different "spin history" (i.e., it will not have recently been hit by an orientation-flipping RF-pulse) and will thus have a greater degree of alignment with the main magnet. When another RF pulse is applied, the fresh blood will have a greater concentration of aligned protons to flip and will thus yield a greater NMR signal. The imaging of this signal happens on a timescale that is rapid with respect to the blood flow, so the change is detected. This phenomenon is the basis for one kind of imaging in fMRI. It is largely sensitive to changes in arterial blood flow (where flow is the fastest).

There is a second, and more commonly used, process that yields an fMRI signal. The neural activity that elicited the local increase in blood flow and blood volume surprisingly does not elicit a correspondingly great increase in oxygen utilization. That is, although the neural activity leads to a small increase in oxygen utilization, it is dwarfed by the increase in blood flow. Thus, there is an increase in oxygenated hemoglobin in the venous portion of the circulatory system near the site of neural activity (as well as downstream from that site). The combination of increased oxygenated hemoglobin and increased blood flow results in a *decrease* in the instantaneous concentration of deoxygenated hemoglobin on the venous side of the capillaries. Deoxygenated hemoglobin (unlike oxygenated hemoglobin) is a strongly paramagnetic biological molecule, and it distorts the magnetic field locally. Thus, a *decrease* in the local concentration of deoxyhemoglobin leads to a *more* uniform magnetic field locally and to a longer time period during which the orientations of precessing protons stay in phase. Thus, the NMR signal in a region of decreased deoxyhemoglobin concentration *increases* relative to its normal (neuronally resting) state. This phenomenon is called the blood oxygen level dependent (BOLD) effect. It is the major source of contrast in most fMRI experiments.

F. Other Technical Issues in MRI

Operationally, fMRI differs from conventional MRI in two basic respects. First, it is tailored to be sensitive to contrasts in blood flow and/or oxygenation that reflect neural activity. Second, it is typically conducted with special hardware that permits the very rapid variation of the magnetic field gradients that are needed to create images. This permits much more rapid acquisition of whole-brain volumes than is conventionally done in MRI. This rapid data collection is crucial in most modern fMRI-based experiments.

Functional MRI is made practical and powerful by virtue of special pulse sequences (such as echo planar and spiral scanning) and hardware that permit the encoding of a brain slice while using a single RF pulse, allowing the entire brain to be imaged in a few seconds. A wide variety of different pulse sequences are used in fMRI, and this remains an area of continuous

innovation. Moreover, the versatility of MRI for neuroscience extends beyond fMRI, and MR can also be used to assay various aspects of brain chemistry through a technique known as magnetic resonance spectroscopy (MRS). Because some variants of MRS can measure the presence of brain metabolites at temporal resolutions on the order of minutes and spatial resolutions similar to those of BOLD fMRI, MRS is in many ways conceptually related to fMRI.

G. Summary

A strong, spatially uniform magnetic field aligns a small but significant fraction of the hydrogen nuclei of water molecules in the brain. A carefully controlled sequence of gradient fields and RF pulses is used to generate NMR signals that can be reconstructed to form a three-dimensional image in which contrast is dependent, in part, on the blood flow and/or oxygenation changes caused by neural activity. Thus, MRI can be used noninvasively to detect changes in local neural activity in the human brain.

III. EXPERIMENTAL DESIGN

Designing experiments for fMRI-based studies presents unique opportunities and challenges. First, fMRI [like position emission tomography (PET) using O^{15}] depends on the indirect signals generated by hemodynamic changes (i.e., changes in blood flow and/or blood chemistry) rather than the more direct electrochemical changes associated with neural activity. Second, there are numerous technical challenges that follow from the particular physics of MRI when used for high-speed imaging of the human brain. Third, there are a number of practical considerations associated with both safety and the physical requirement for minimum movement of the subjects in fMRI studies that add to the challenges of fMRI experimental design. All these factors affect (and, in turn, are affected by) the current practical and future potential limits of spatial and temporal resolution associated with fMRI. Finally, as with any experimental approach to important questions concerning human psychology, the most fundamental and difficult problems arise in choosing tasks and stimuli that allow one to be convincing to one's self and to one's audience that the psychological question being asked is truly addressed by the experiment being performed. Can you convince your audience, for example, that when you use fMRI to measure changes in brain activity during the color Stroop task you are actually studying some general attribute of inhibition and higher cognitive function?

A. Experimental Design and Hemodynamics

Functional MRI is dependent on hemodynamic changes rather than the electrical consequences of neural activity. The spatial and temporal characteristics of these hemodynamic effects must be taken into account when designing experiments and analyzing the data from these experiments. The spatial characteristics arise from the underlying vasculature; the temporal characteristics include a delay in the onset of detectable MR signal changes in response to neural activity and a dispersion of the resulting hemodynamic changes over a longer time than that of the initiating neural events.

With regard to the temporal aspects of the hemodynamics, fMRI experiments can be classified into two broad categories: block designs and event-related designs. In block designs, the experimental task is performed continuously in blocks of time, typically 20–60 sec in duration. The idea here is to ignore the details of the temporal characteristics by virtue of setting up a "steady state" of neuronal and hemodynamic change. The fact that there is a brief delay before the MR signal changes are detected is often unimportant when analyzing a long block of steady-state activity. This approach is conceptually simple; it is analogous to older PET experimental designs, and it is of great practical importance for fMRI because it is the optimal technique for *detecting* small changes in brain activity. The major weakness of block design is the requirement that all the stimuli or task characteristics remain unchanged for tens of seconds, precluding the use of many classic psychological paradigms (such as the "oddball" scheme).

The other major approach, event-related design, makes use of the details of the temporal response pattern in the hemodynamics as well as the largely linear response characteristics associated with multiple stimulus presentations. In event-related designs, the different stimuli are presented individually in a random order (rather than in blocks of similar or identical stimuli) and the hemodynamic response to each stimulus is measured. Event-related designs are further subdivided into spaced single-trial designs and rapid single-trial designs. In spaced single-trial designs, stimuli are presented with a long interstimulus interval

(ISI) relative to the hemodynamic response to a single stimulus. Specifically, an ISI of at least 10 sec, and more typically 12–20 sec, is used in an effort to allow the hemodynamic response to each stimulus to return to its resting state before the next stimulus is presented. This approach is conceptually simple but very inefficient in its use of imaging time, much of which is spent collecting data when the MR signal variation due to hemodynamcs is small or not detectable. (This is not only wasteful of expensive imaging time but also boring for the subject, who is only doing something approximately once every 15 sec.)

In contrast to *spaced* single-trial designs, *rapid* single-trial designs take advantage of the linearity and superposition properties of the hemodynamic responses to neural activity. To a first approximation, the hemodynamic changes associated with multiple stimulus presentations are additive and when presented at different times, are simple time shifts of each other. This permits the much more efficient design of experiments in which novel stimuli appear in quasi-random order and with variable ISIs (typically presenting a new stimulus every 1–3 sec). The associated data analysis is more difficult because the hemodynamic responses to the different stimuli overlap in time (and there are consequent weaknesses relative to block designs in terms of the detection of small effects) but the rapid single-trial designs are particularly powerful and useful when it is essential to have random order in the presentation of individual stimuli (i.e., in the situation in which a block design with long periods of the same type of stimulus would not permit the desired comparisons for neural activations). It is also more efficient in the use of imaging time and more engaging (less boring) for the subject.

One final approach to experimental design should be mentioned. All of the previously discussed techniques typically make use of averaging over multiple instances of a given trial type. In block design the trials all occur together, so the averaging is done as much by the hemodynamic and neural systems as by any data analysis software. In event-related designs the averaging over the effects of multiple stimulus presentations is done explicitly in software during data analysis. It is possible, however, to analyze spaced single-trial data on the basis of activation from a *single event* (rather than averaging over multiple instances of the same trial type). This technique has not been widely applied, primarily because the elicited signals to single stimulus events are generally weak. However, high-field MRI systems, and the selection of experimental paradigms that elicit strong, focal neural activity, have been used to demonstrate the feasibility of single-event fMRI.

B. Spatial and Temporal Resolution in High-Speed MRI

The physiology of the circulatory system and the physics of the MRI devices constrain the spatial and temporal resolution of fMRI. Today, it is routine to obtain $1 \times 1 \times 1$-mm structural MR images and $5 \times 5 \times 5$-mm functional MR images. The temporal resolution of fMRI is on the order of 1–3 sec. Neither the spatial nor the temporal resolution numbers are indicative of absolute limits in terms of the physiology or the imaging hardware. Rather, they represent a snapshot in the development of ever-improving resolutions. Moreover, at any give stage of technical development in MRI, the various imaging parameters can be manipulated to emphasize one aspect of resolution in exchange for another.

When investigators approach experimental design in fMRI, they must recognize that the key physical variables—spatial resolution, temporal resolution, brain coverage, and signal-to-noise ratio—are quantities whose values can be manipulated by trading one off against the others. For example, extremely high spatial resolutions are possible, but the techniques needed to achieve them involve reduced temporal resolution, limited brain coverage, and/or decreased signal-to-noise ratio. Alternately, extremely rapid imaging can be performed, but at the cost of spatial resolution and/or brain coverage. Trade-offs will continue to exist even as the overall power of scanning technology improves.

As an indication of the numbers associated with these issues, and the manner in which they are changing, consider the issue of the rate at which individual images can be collected. The first whole-body high-speed (EPI) fMRI system could collect 20 images per second for about 1 min, and then it would overheat. At a slower operating rate of 10 EPI images per second (which was still very fast in 1992), there was no overheating, but the subject (in an fMRI experiment) had to wait a long time between scans for reconstruction of the images from the raw data. At an even slower rate of 5 EPI images per second, the scanner could operate continuously and reconstruct the images in real time, but the memory for buffering those images would fill after about 2000 images. In contrast, modern machines can operate at 20 images per second continuously.

Analogous improvement is ongoing in all of the mentioned domains. Higher field MRI (from 3 to 8 T) will improve spatial resolution and will yield the added signal that may improve the practicality of single-event fMRI designs and the possible use of the "initial dip" in oxygenation to improve temporal resolution. However, these high-field machines are not as widely available as 1.5 T machines, and they are considerably more expensive, difficult, and dangerous to use.

Currently, the message for experimental design is simply that these resolution limits must be taken into account. There is little point to designing a conventional experiment to detect changes in a structure that is much smaller than one's spatial resolution permits, nor in designing a study that requires the detection of temporal changes that are too rapid for one's current technology. At the same time, because these imaging parameters can be traded off, one should not dismiss difficult-sounding experiments too quickly.

C. Practicalities: Psychophysiological Laboratory in the Magnet, Safety, and Costs

The physical properties of MRI, as well as the financial costs, place a number of practical constraints on the design and execution of fMRI-based studies and thereby impact experimental design.

As indicated previously, the experiments take place in the bore of a large, powerful magnet. The presence of this large, static magnetic field precludes the participation of subjects who would be adversely affected by it. For example, subjects with pacemakers or other forms of implanted metalic devices that would be subjected to strong forces by the magnet are clearly ineligible. The fact that subjects must lie in a relatively confined space rules out volunteers who suffer from claustrophobia. The physical position and limitations of the subject's movement also constrain various experimental procedures that would be simple in an ordinary behavioral laboratory (Fig. 2).

In addition to the main, static magnetic field, there are strong varying electromagnetic fields from the gradient magnets (for generating images) and from the RF oscillator (for flipping the protons to obtain the NMR signal). Each of these fields has associated safety considerations. For the most part, this is a minor issue at 1.5 T. It can be more of an issue for some pulse sequences at higher fields. The manufacturers of MRI scanners are required to build in various safety measures to protect subjects (such as calculating the heating effects of the RF pulses for a person of a given size and weight). Nonetheless, each imaging facility normally includes a screening form (and sometimes a metal detector) to prevent the inadvertent harming of subjects from a collection of (sometimes not so obvious) potential dangers.

In addition to these safety considerations, the physics of MRI has other practical consequences. All of the devices used to present stimulation (visual, auditory, etc.) and to obtain behavioral and physiological response measures (button pushes, breathing and heart rate, etc.) must be constructed in an MR-compatible manner. As fMRI has become a more widespread enterprise, various companies have made it a business to supply such equipment. However, custom design of devices for specialized experiments is still common.

The single most vexing problem in the practical application of fMRI is head movement. Although pulse sequences have been developed to collect an entire slice of brain data in less than 50 msec, and multiple slices (for entire brain coverage) can be collected in 2–3 sec, the amount of information in each such image is limited. That is, the amount of *functional contrast* in the images—the differences in the signals between two experimental states—is small. To make up for this, many images are collected over extended periods of time—at least minutes and sometimes hours. During these time periods it is important that the subject's head move as little as possible.

A variety of techniques are used to encourage subjects to keep their heads as motionless as possible, but none is perfect. For well-motivated healthy adult subjects, this is usually not an insurmountable problem. For children and older patient populations, it can be the main reason that data are discarded. Although there are data analytic procedures for transforming images of moving heads back to a fixed position, these procedures are limited. Indeed, because the moving head actually distorts the main magnetic field in different ways, no motion correction algorithms can fix the problem perfectly. There are many extra coils in an MRI scanner that are used to make the main magnetic field as uniform as possible, despite the irregularities introduced by the presence of a human head in the bore of the magnet. These "shimming" coils are supplied with electric current designed to minimize the magnetic field distortions introduced by the head at the beginning of a scanning session. However, they are not modified on the fly, during the session, so any subject head movement results in more than just a displacement of the image; it also causes distortions that are much more difficult to correct.

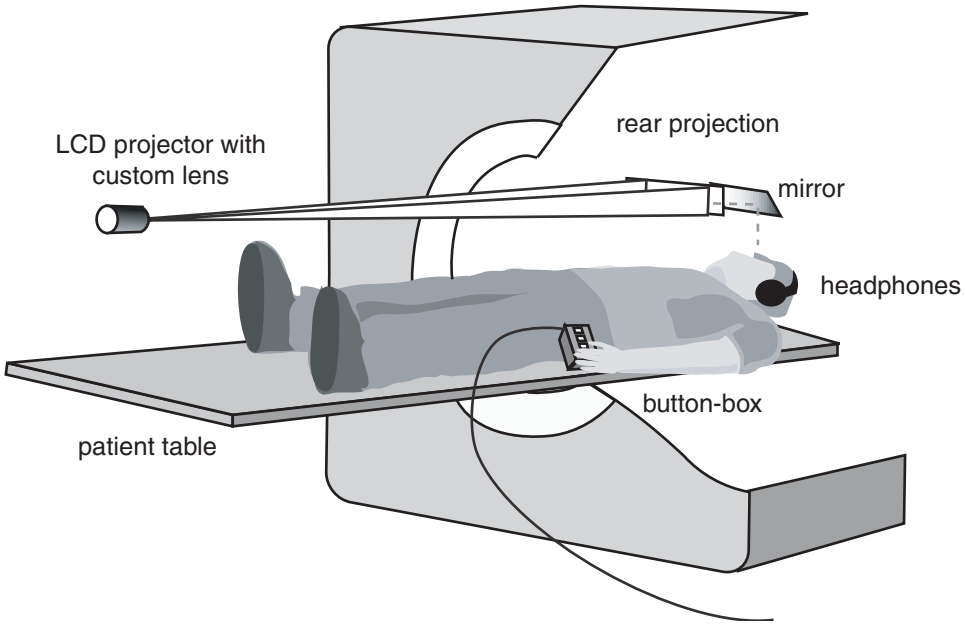


Figure 2 Subject in the magnet. Functional MRI is conducted in the environment of a MRI suite. There are many ways to present stimuli and obtain responses from the subject in such a suite, but all must be compatible with the difficult and hostile electromagnetic environment of MRI. In the example shown, visual stimuli are projected onto a rear-projection screen and are viewed by the subject via a mirror. Auditory stimuli can be presented via MR-compatible headphones or by the speaker systems typically included in MRI scanners. Finally, subject responses can be collected using an MR-compatible button box. A variety of commercial and custom approaches to the problems of presenting stimuli and recording responses in fMRI have been developed.

Finally, it should be noted that MRI time is not cheap. Charges for an hour of clinical imaging are in the hundreds of dollars. Therefore, when designing a study, the total number of imaging minutes is one of the parameters that must also be considered in the trade-offs.

D. Comparing Activation States

Fundamental to the understanding of fMRI as a tool for representing the localization of brain function is the idea that a single image, in isolation, conveys little if any useful information. Rather, it is the comparison of multiple images that are collected during different states of neural activity that supplies interpretable data. Note that this statement is *not* true for structural MR images. A single structural image conveys a great deal of useful information because data about *change* is not sought (except on a much longer timescale, as in developmental and longitudinal studies of brain structure). In contrast, functional imaging data are almost exclusively about *changes* in neuronal activity.

One might ask, Why isn't a single image, collected during rest, a useful definition of the "resting," "neutral," or "idling" state of the brain? In some ways, a single image might be interpretable this way. Indeed, some variants of PET can be used to yield a single snapshot of the metabolic state of the brain. However, the variation in local activity in the brain during "rest" is not very meaningful—the demonstration that one portion of the brain is more active during rest than another has limited value. On the other hand, the demonstration that a particular manipulation (of the stimulus or task requirements for the subject) causes a localized change in neuronal activity is far more useful.

The art of fMRI experimental design lies largely in the creation of tasks that accurately probe the cognitive function of interest. One natural way to design an experiment in functional neuroimaging is to create two tasks that are identical except for one minor difference. This is the basis of the classic subtraction method originally delineated by Donders and widely used in cognitive research. Such experimental designs are sometimes called "tight" task comparisons. The difference between experimental conditions in a tight task comparison is either in the *stimulus alone* (while keeping the response task of the subject fixed) or in the *response task alone* (while keeping the stimulus fixed).

Such an approach is particularly useful for testing specific hypotheses about the activation pattern in a single brain region.

However, there are practical and theoretical reasons for including experimental conditions that are more broadly different from the main conditions of interest. Frequently, this is accomplished via the use of a low-level control task, such as simple visual fixation or rest. This has sometimes been called a "loose" task comparison. It is particularly useful for seeing the simultaneous activation of many areas of the brain. The loose task comparison not only provides an internal check for the integrity of the data collected (because it typically includes robust activations of no direct experimental interest, but the absence of which could indicate a problem with the subject, the machine, or the data analysis) but also serves as an important point of reference for observed differences within the tight task comparison. For instance, a difference between two conditions in a tight task comparison could reflect either an increase in activity in one condition or a decrease in activity in the other. The addition of a loose task comparison provides a means of disambiguating such a situation by providing a baseline against which the two tight task conditions can be compared.

More generally, it is essential to have at least two conditions to be compared, but the power of fMRI-based experiments to test interesting theories is greatly enhanced by the presence of more conditions in the design. Sometimes these multiple conditions are qualitatively different (as indicated previously), but increasingly subtle experiments are being done that make use of quantitative (parametric) variation in the experimental conditions. In general, when attempting to model and understand the networks of the brain, all types of experimental sampling are needed.

Finally, the critical importance (and occasional irrelevance) of behavioral measures must be discussed. It might seem obvious that obtaining observable behavioral responses could only be a good thing in functional neuroimaging research. Certainly most investigators try to have an observable behavioral measure for their tasks when possible. (In "imagination" studies, such as imagining visual images or imagining performance of a motor task, it is sometimes impossible to have an observable *behavioral* response measure, but even in the context of something as "unobservable" as mental rotation, investigators have sometimes found ways to obtain associated reaction times and accuracy measures.) On the other hand, at least one prominent psychologist has argued against the necessity of behavioral response measures, suggesting that the imaging data are sufficient and that adding irrelevant behavioral tasks will only confuse the issue by eliciting neural activity unrelated to the particular cognitive task of interest. Also, several researchers have commented that, independent of anything else, it is good to have a behavioral task associated with the imaging study because it will help keep the subject awake in the scanner.

Each of these observations has merit, but there are more important uses for behavioral response measures in most studies, and for some studies they are critical. Specifically, a number of studies have made the analysis of the imaging data depend crucially on the observed *behavioral* responses. Examples from the study of memory and from the study of the effects of cocaine on brain activity are described in Section I.

E. Summary

Experimental design in fMRI-based experiments is challenging but rewarding. Practical limitations related to the safety and behavior of human subjects, technical limitations in MRI devices, underlying properties of the spatial arrangement of blood vessels, and the temporal characteristics of the coupling between neuronal activity and blood flow are all intertwined. In addition to these technical considerations there is the art of experimental design associated with any question in understanding human psychology. The application of the technology of fMRI to neuroscience entails a collection of trade-offs. Nonetheless, the current strengths of fMRI-based investigations include the best spatial and temporal resolutions for noninvasive, volumetric brain mapping, the most flexible experimental designs, and a constantly improving set of instrument-based limits to the technology's sensitivity and resolution.

IV. DATA ANALYSIS

A typical fMRI scanning session lasts 1–2 hr and results in the collection of hundreds of megabytes of data. The theory and practicalities associated with processing that data are complex and evolving. In contrast to functional neuroimaging associated with PET—in which the total amount of data is much smaller, the understanding and agreement about the sources and nature of noise in the data are well established, and there is a consequent widespread

agreement about the basic issues in data analysis—the situation with fMRI data is much more complex. The sources of machine-related noise in the raw MR images are relatively well understood. However, the general consensus regarding noise in fMRI data is that the most important sources are physiologically based (in the subject) rather than machine based (from the scanner). There is less agreement about the details and the consequences of modeling these noise sources in terms of the practical consequences for data analysis.

Perhaps even more important, the present (and future) spatial and temporal resolution of fMRI data encourages modeling of brain systems at a level that may substantially exceed that of previous volumetric imaging systems. Some of these advances (e.g., the ability to obtain precise delineation of multiple visual areas in occipital cortex by virtue of their retinotopic regularities) require different kinds of data analysis and different kinds of visualization tools than made sense in the context of systems with poorer spatial resolution. Finally, the ability to image the same subject multiple times, and the associated potential for the collection of many kinds of functional data from that same subject, encourages novel approaches to data analysis.

Data analysis is a critical, time consuming, and sometimes controversial part of fMRI-based experimentation. Although the nature of many of the problems is well defined, the appropriate solutions are not. There is general agreement on how to handle some of the issues associated with data analysis (e.g., algorithms to detect head movement and correct for head movement) but there are no universally agreed on approaches to many other issues (e.g., the appropriate statistical tests to define the detection of neural activation, the best way to compare data across different subjects, and the best way to visualize and report the results of data analysis). There are a host of software tools for data analysis, each having its strengths and weaknesses. Because of the rapid development in all aspects of fMRI-based research, no *de facto* standard approach to data analysis has emerged. Figs. 3 and 4 indicate some of the procedures that are discussed in more detail in the following sections.

A. Preprocessing

Before the essential part of data analysis can begin, a number of preliminary steps are typically taken. Some spatial smoothing and temporal smoothing may be applied, but the first and most important step is the assessment of subject head movement during the imaging session.

The problem of subject motion is a pervasive one in fMRI and arises not only in constraining experimental design but also in the analysis of images of a brain that may have moved over the course of an experiment. The high-speed imaging techniques used in fMRI typically minimize the effects of movement in any one image. However, many images are collected in each run, which are typically 2–8 min long, and there are many runs in a typical 2-h session, representing more than 1000 brain volumes per session. Because the fMRI-based signal modulation is intrinsically small (typically 0.5–5%), data from all the images collected in these long runs are normally needed to gain statistical power. Thus, it is important that the images within a run and across runs are properly aligned. Head motion makes this process a challenge. In addition, even if the skull were perfectly immobile, the brain still moves. The pulsatile flow of arterial blood causes movement virtually everywhere in the brain, particularly in subcortical structures.

For all these reasons, the data analytic approach to motion detection and motion correction has been based on the brain images rather than on monitoring head movement externally. Efforts are made to *minimize* subject head movement, and it is not currently possible to correct for severe or rapid movement. All the current algorithms for correcting head movement assume rigid motion of the head. Although a single slice of brain imaging data is collected very rapidly compared to most head movement, the time needed to collect an entire brain volume—consisting of 20 or more slices—is much longer than many head movements. Such motion cannot be corrected with these algorithms. However, if the movement is not too great in amplitude and not too rapid, there are algorithms available in most fMRI data analysis packages that are adequate to detect the motion and to transform the data in an attempt to compensate for the effects of that motion.

A key feature of these algorithms is that they automatically reveal many kinds of movement, including stimulus-correlated movement. If a subject moves every time he or she is supposed to start a task, the movement could create MR signal artifacts that appear as a false activation signal. There is no good way to correct for such data, and these date must be detected and discarded.

Subject movement is generally regarded as the major problem for getting consistent data in fMRI-based experiments. Experienced, well-motivated subjects

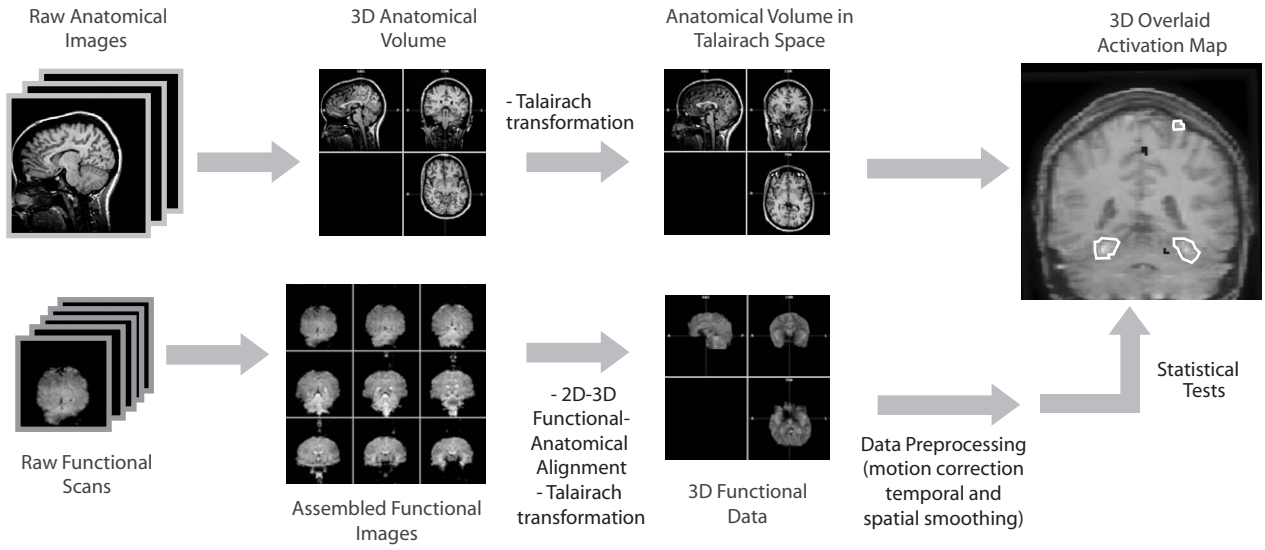


Figure 3 Highlights of an fMRI data processing stream. Data from fMRI-based experiments are analyzed in many steps. The number and order of these steps are still a topic of controversy, with variations across laboratories and software packages. This is a simplified representation of the generic steps. Both high-resolution structural MRI images, and lower resolution functional images pass through a variety of preprocessing steps. Early steps may transform the raw anatomical images from individual (two-dimensional) slices of the brain into volumetric (multislice, three-dimensional) arrays that are more suitable for the detection of head movement. At the same time, data are often transformed to a standard three-dimensional orientation and overallsize scale (Talairach coordinates). In addition to the essential step of *detecting* the presence of head motion in the data, several other (arguably optional) steps may be applied. Motion correction algorithms can help if head movement is not too much (although these algorithms are sometimes unnecessary or counterproductive if motion is little). There are both theoretical and practical reasons for smoothing the data in the spatial and/or temporal domains, although some investigators argue against performing these steps. Finally, statistical tests are performed contrasting the data collected during different experimental conditions. The resulting statistical maps are typically thresholded and overlaid on anatomical images, as indicated in Fig. 4. Further considerations are attendant to the comparisons across the brains of different subjects, as discussed in the text.

who use bite bars in the scanner can routinely be expected to yield data free of serious motion artifact. In contrast, in studies with clinical patients or other difficult subjects, as much as 20–30% of data may need to be discarded because of subject motion. There are methods for helping to minimize physical motion during data acquisition and to detect and possibly compensate for it after data acquistion, but none are fully satisfactory.

B. Basic Detection of Change

The first goal of any analysis of fMRI-based data is to determine whether the experimental manipulation has resulted in a measurable change in the MR signal and to specify where in the brain and when (in time) that change has occurred. In principle, any statistical method that can be applied to a time series can be used with fMRI data. In practice, the demands of the experimental paradigm, limitations of the tool, and the capabilities of distributed software packages constrain the sorts of analyses that are typically performed. A few broad classes of common data analysis options are detailed here, although the presentation is not comprehensive. With the exception of principle component analysis (PCA) and other multivariate techniques, each of these tests is applied at the voxel level. When these statistics are computed for each voxel in the brain, and the resulting collection of statistics is presented in the form of an image in which color or intensity is used to represent the value of that statistic, the result is called a statistical map of brain activation.

High-speed imaging (e.g., EPI) is used to collect many images of the brain during each of the experimental conditions designed. The simplest (and, historically, the first) comparison was obtained by subtracting the average of all the images collected during one condition from the average of all the images collected during another condition. The resulting difference image clearly showed areas that were brighter, indicating greater MR signal during one condition than another.

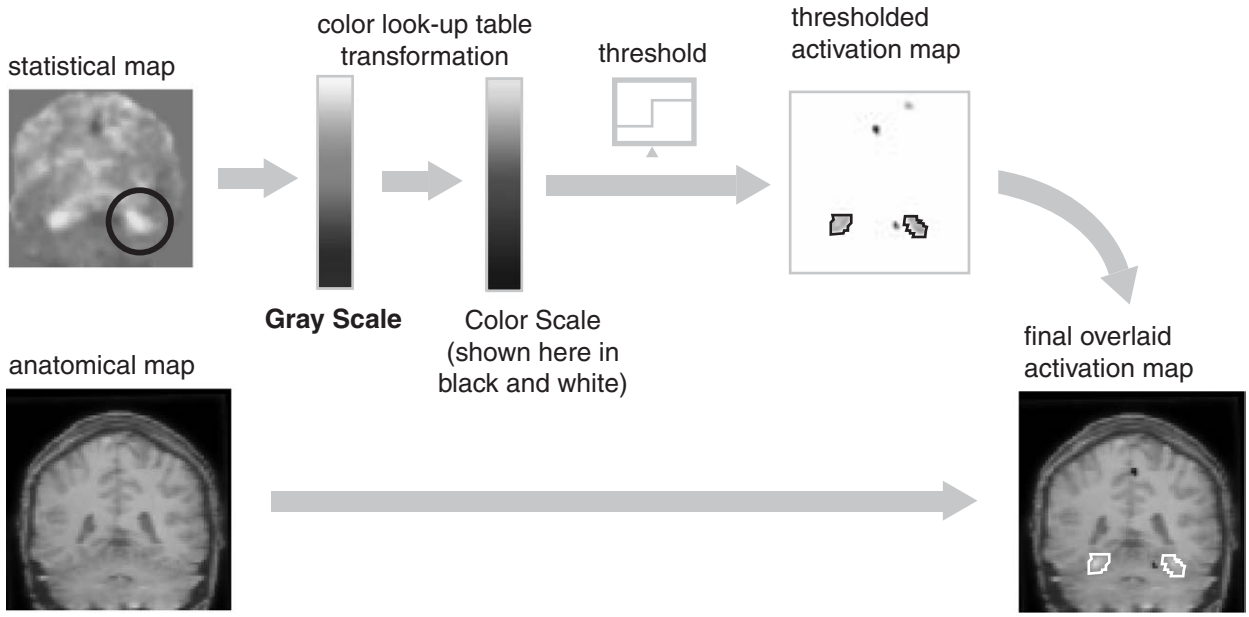


Figure 4 Production of a color–coded activation map. After preprocessing, comparisons of the values of the functional MR images collected during different experimental conditions are used to generate a statistical map. The underlying question is whether the collection of values from one condition is likely to have been generated by statistically different levels of brain activity, when compared with the collection of values obtained during a second condition. These statistics are computed on a voxel-by-voxel basis, resulting in a spatial map (left, top) in which gray-scale intensity indicates magnitude of the statistic. The midgray area represents little difference between the two experimental conditions in question; bright regions occur where the first condition elicited much stronger MR signals than the second condition; and darker regions occur where the first condition elicited much weaker MR signals than the second condition. Because of the need to collect many functional images in a short period of time, functional images typically have substantially less spatial resolution than structural (anatomical) images. (These differences are a consequence of the different pulse sequences that are used to collect functional versus structural MRI data.) In order to combine information from the statistical map with the higher resolution structural images, two transformations are applied to the statistical map. First, the gray-scale intensities used to represent the statistics are mapped into a color scale. Second, a threshold is applied so that statistical values that are within a user-defined range close to zero are mapped to transparency. This permits the combining of the two maps (thresholded pseudocolor map of statistics and high-resolution gray-scale map of anatomy) by overlaying the color map on top of the structural map. In this way, it is possible to get a better sense of the location of the changes in neural activity associated with the different experimental conditions.

This sort of comparison is the only one that, strictly speaking, is "subtraction," although the term is sometimes used informally during discussions of contrasts between conditions, even if those contrasts are based on some other statistic. Instead of subtracting (on a voxel by voxel basis) the averaged data from different conditions, the more general approach is to compute a statistic based on the collection of values at each specific spatial voxel, collected across the times of the many images. Such generic statistical computations on grouped images are more accurately described as "contrasts" or "comparisons" rather than subtractions.

C. Systematic Detection of Change

The most obvious and simple statistical test that can be used in fMRI data analysis is student's *t* test. This test assumes that each number in each group is independent, and that the underlying distribution of numbers is Gaussian (i.e., it is a parametric test). In fact, both of these assumptions are often violated in actual fMRI data. Nonetheless, parametric statistics like the *t* test are the most widely used measures of the difference between the groups of numbers collected in fMRI images across conditions. Other statistical tests are possible and sometimes used.

The most commonly used approach to the detection of systematic effects in fMRI data is the general linear model, which uses correlational analysis. Here, the fMRI data are compared with some kind of reference temporal function to determine where in the brain there are high correlations between the reference function and the MR data. The reference function is obtained from the experimental design. For example, because the brain's hemodynamic response follows a fairly consistent profile, a boxcar function defining the

experimental paradigm is often convolved with an estimated hemodynamic response function to yield the reference function. The resulting reference function is smoother than a boxcar and better takes into account the shape of the hemodynamic response, generally resulting in better correlation between the MR signal time courses and the regressor time course. Several functions have been historically used to model the hemodynamic response, including a Poisson function and, recently, a gamma function. Often, a single canonical hemodynamic response function is used across the entire brain and across subjects, though there is evidence for variation in hemodynamic response shape across subjects and brain regions. Some software packages make provisions for this, allowing for independent modeling of the hemodynamic response function on a voxelwise basis.

There are a number of variations on this general scheme. For multiple experimental conditions, the previously mentioned scheme can be easily extended using multiple regression. In addition, it is also possible, though less common, to perform nonlinear regression on fMRI data, given some nonlinear prior model of expected brain response.

As with any statistical test, one must exercise some caution when using correlational analysis to ensure that incorrect inferences are not made due to violation of the assumptions inherent in the statistical test. In particular, the assumption of independence of consecutive samples is sometimes badly violated in fMRI data, inflating estimates of significance.

All of the preceding approaches make the assumption that the variations of interest in the data are those that occur in temporal synchrony with the experimental variations built in to the design. These tests cannot detect novel temporal variations triggered by the experiment but not part of the design. For instance, if a change was triggered at stimulus onset and stimulus offset, most standard data analytic packages, as they are typically employed, could not detect that response. In contrast, various multivariate approaches (such as PCA) seek regularities in the spatiotemporal structure of the fMRI data that are not specified beforehand. Such techniques typically detect the experimental variation that was designed by the experimenter as well as some physiological variations (such as those due to breathing or heartbeat). The challenge is to refine these tests so that it is easy to interpret the regularities that are detected.

Principal component analysis is not really a statistical test. Rather, it is a re-representation of the data that condenses as much of the variability in that data as possible into a small number of eigenimages, each of which is associated with an eigenfunction that specifies a temporal fluctuation for the entire image. Thus, instead of one temporal variation for each voxel in a brain volume, there are a small number of volumetric images, each of which varies as a single relative image according to some time course. The key virtue of PCA is that it has the power to pick out particular areas in the brain that exhibit a time course similar to that in the experimental design without the experimenter ever having specified that design to the analysis procedure. Similarly, it can detect temporal changes that are different from the ones built into the experiment. On the other hand, there is no obvious way to *know* which eigenimages and eigenfunctions actually correspond to an important or interpretable variation. PCA and related techniques have great theoretical appeal, but they have been rarely used in practical fMRI data analysis. A related multivariate technique, called independent component analysis (ICA), is designed to help with the interpretation of the data. Instead of projecting data into a lower dimensional space that accounts for the most variation (as PCA does), ICA finds a space in which the dimensions are as independent as possible, thus facilitating interpretation of the components.

D. Comparing Brains

Nearly all fMRI studies use multiple subjects and perform statistical analyses across data collected from multiple subjects. This practice introduces a number of practical problems that must be addressed.

A first, relatively simple step toward comparing activity across the brains of multiple subjects is to transform the representations of those brains so that they are similar in overall size and similarly oriented in space. To accomplish this, the brain images are rigidly rotated and linearly scaled into a common "box." The Talairach stereotactic coordinate system is the most widely used standard coordinate system box for comparing brains. In the full Talairach transformation, a rigid rotation and translation to a standardized orientation is followed by a piecewise linear scaling of the anterior, middle, and posterior portions of each hemisphere, independently. The standard orientation in three dimensions is determined by the line between two interhermispheric fiber bundles—the anterior commissure (AC) and the posterior commissure (PC)—and the plane between the two hemispheres. The anterior, middle, and posterior portions of each

hemisphere are defined in terms of the AC–PC line: The portion of each hemisphere in front of the anterior commisure is the anterior, the portion of each hemisphere between the AC and PC points is the middle, and the portion behind the posterior commisure is the posterior.

In some software packages an abbreviated form of the Talairach transformation is performed in which the brain is scaled as a whole, without piecewise linear portions for each hemisphere. This has the advantage of being much faster and simpler to implement. Indeed, some packages compute this transformation automatically by comparing the given brain to a standard "average" brain that was generated by transforming the anatomical MR images of 305 brains to Talairach coordinates and averaging. This process eliminates the tedious and often tricky steps of finding the AC–PC line and other landmarks for each individual brain. Despite the fact that the individual anatomy of the AC–PC line is ignored and the transformation has fewer degrees of freedom than the full Talairach transformation, this automatic process yields data that are adequate for most purposes.

The more serious problems with either the simplified Talairach transformation or the full Talairach transformation are caused by the fact that real brains are not rigid transformations of one another. The simplified Talairach transformation, being strictly linear, cannot account for these differences. Even the full Talairach transformation, which is only piecewise linear with a small number of pieces, is clearly inadequate for dealing with these individual differences in brain anatomy. More powerful nonlinear approaches have therefore been developed.

There are a wide range of approaches to more general transformations of brains to facilitate data display and intersubject comparison. One approach is to permit complicated nonlinear warping procedures based on sulcal and gyral landmarks to guide computer-generated distortions of one brain into another (or to a standard). Another approach, which is also based on nonlinear transformations, is to try to match the perimeters of given brain slices between different brains. Perhaps the most widely used alternative to Talairach and these other nonlinear transformations is to "inflate" the brain as a means of removing all sulci and gyri. This inflation is often followed by cutting the inflated brain in a small number of places to permit "flattening."

The goal of inflating is to obtain a three-dimensional, smooth, nonconvoluted surface representation of cortex. The goal of flattening is to lay that surface representation on a flat plane. Given that a brain hemisphere (when inflated) looks like an ellipsoid, it is necessary to cut it in one or more places to allow it to be flattened. Qualitatively, this is similar to the need to cut a globe to get a flattened representation of the world. Quantitatively, however, for an inflated brain, the analogy is closer to a cylinder. When a globe is cut and flattened, it is necessary to create many cuts or else there will be some very large distortions. On the other hand, when a cylinder is cut, the resulting surface can be flattened with virtually no distortions. In these terms, the inflated cortex is more like a cylinder than a sphere, and the distortions are not terribly large.

Flattened representations are visually and logically very appealing. Unlike Talairach coordinates, however, they are not three dimensional and only apply to the cortical surface. Subcortical structures cannot be represented. (Talairach invented his system for subcortical structures, although it does not include the cerebellum.)

Given the good spatial resolution of fMRI and the ability to detect activations in individual subjects, some researchers eschew averaging across subjects. Their position seems to be that the right way to compare across subjects is to look at *each individual's functional map* (preferably in a flattened brain format to facilitate intersubject comparison). Elaborations on this approach can include warping within the flattened space and could therefore eventually include averaging in that space.

Recent developments in brain comparisons involve the use of more sophisticated algorithms to inflate the brain to a sphere but then warp the surface borders on that sphere (associated with major and almost universal sulcal/gyral landmarks) toward a common standard. This transformation permits a smoother and more effective comparison of activation sites across the brains of different subjects than do the Talairach transformations.

E. Comparing Groups

One of the most obvious and important classes of questions that are addressed with human functional brain imaging is the search for differences between groups. For example, can fMRI be used to detect the early onset of Alzheimer's disease? Does a remedial training program in reading cause changes in brain activity preferentially for one diagnostic classification of dyslexia versus another? Does a given drug

treatment lead to greater area of functional brain activity? All these questions have as an essential component the attempt to make quantitative distinctions between different groups of subjects.

Functional MRI (and functional brain imaging more broadly) can be used to address at least two types of questions. One question might be thought of as the attempt to represent "typical" brain function and associated networks of activity. In that context, collecting increasingly more data about a single brain doing a single task might be useful because the error bars associated with any particular aspect of the brain activity might be expected to decrease with increased measurement. In statistics, this is called a fixed effects model. On the other hand, to determine whether there are differences in brain function and networks of activity between two putatively different *groups* of subjects, it is important to sample many of members of each group, even if the individual measurement of any one member of the group is noisy. In particular, knowing with extreme precision that two members of one group differ from two members of another group is only useful if the within-group variation (i.e., between brains) is as small as the within-brain variation (i.e., between multiple measurements of the same brain). If not, then the exceptional precision of the measurement of the small number of subjects is not useful. In statistics, this is the random effects model.

The practical implication of the fixed versus random effects model of variance for functional neuroimaging is that it is better to have measurements on many brains if the goal is to claim group differences. On the other hand, it may be better to have many measurements on a few brains if the goal is to delineate functional systems as precisely as possible.

V. RESEARCH APPLICATIONS

Human fMRI based on the endogenous contrast agent (deoxyhemoglobin) was first reported in 1991. In the ensuing 10 years the growth of fMRI-based research applications has been explosive. The easiest (and probably the most accurate) way to summarize the range of fMRI-based research applications is "all of psychology and neuroscience". In addition to widespread reports of results of fMRI-based research in general scientific and popular journals, there are two journals devoted exclusively to the technical developments and applications of human brain mapping, for which fMRI is a primary tool. Research applications range from the classical psychophysical questions of sensation, perception, and attention to higher level processes of cognition, language, and emotion, in both normal and patient populations. Even domains ranging from psychotherapy to genetics are beginning to make use of fMRI-based experiments. The impact of fMRI on a few select areas of research is elaborated next.

A. Retinotopy/Multiple Cortical Visual Areas

The first application of fMRI-based research was in the domain of the early stages of visual processing. Indeed, the very first human fMRI study involved the demonstration that a region of the brain associated with early visual processing—occipital cortex in the calcarine fissure—yielded an NMR signal that varied as flashing lights were presented (or not) to a subject.

This demonstration was exciting, but the excitement was limited for several reasons. First, nothing "new" had been demonstrated about human visual cortex. Second, there were a host of technical concerns that, had they been correct, would have meant that the spatial resolution obtainable with fMRI would be seriously compromised. Finally, most of the following simple advances would not go beyond what we already know from (invasive) single cell recordings in nonhuman primates.

However, the development of fMRI in the ensuing years for the study of early visual processing addressed all these concerns and went far beyond them. First, retinopy was demonstrated for area V1 at a level of spatial resolution that exceeded any previously demonstrated with a noninvasive technique. Second, retinotopy was used to delineate multiple visual areas. Differences between the layout of human visual areas compared with those of other primate species were demonstrated, and new visual areas apparently unique to humans were claimed.

At the same time, some classic psychological effects, such as the motion aftereffect, were seen to be associated with detectable brain activity localized to specific parts of the cortex associated with visual motion processing.

The early dependence on the connection to known primate neurophysiology is being lessened. Several laboratories are now developing fMRI suites designed specifically to study nonhuman primates. The idea is to use the invasive technologies such as single cell recording, adapted for the MR environment, to obtain a deeper understanding of both the functional brain structures and the relationship between neural activity

and hemodynamics using methods that would be unethical with human subjects.

B. Modulatory Effects of Attention

One early fMRI-based study demonstrated that the use of voluntary attention (deciding whether to attend to a subset of moving dots or to a subset of stationary dots in a field of moving and stationary dots) caused detectable changes in MR signals associated with a visual motion processing area in cortex (Fig. 5). This study did not have an overt behavioral measure to provide external evidence that subjects were actually performing their assigned tasks. However, the data were sufficiently clean and unambiguous that this study was published and gained considerable attention.

During the ensuing years the study was replicated and extended in a number of ways by different laboratories throughout the world. The initial basic demonstration of attentional modulation became the starting point for much more subtle experiments—experiments that were more tightly tied to behavioral measures. Importantly, both the qualitative and the quantitative measures of attentional modulation were replicated. For instance, the motion processing area was active whenever there was visual movement present, but that activity increased by about 50% when the subject was attending to the movement compared to when the subject was not attending to the movement. The studies that used analogous tasks as part of their design found quantitatively similar changes.

The basic paradigm was adapted to more complex stimulus situations in which subjects could attend to various different aspects of a complex scene. This permitted the testing of specific hypotheses about the allocation and connection of visual attention to different aspects of a stimulus. For example, it was demonstrated that when an object was attended because of one attribute (e.g., motion), there was increased processing of other attributes of that object (e.g., whether it was a familiar face or represented a familiar location) even though the other attributes were irrelevant to the attentional task. Thus, fMRI-based experiments were being applied to theoretical questions of long-standing interest in cognitive psychology.

C. Use of Behavioral Responses

The use of behavioral responses in fMRI-based studies began as a comforting demonstration that subjects were doing what the experimenter had asked them to do. However, behavioral measures can be much more useful. Two studies of memory, for example, made use of behavioral data collected after the MRI scanning session was over and the subject was out of the magnet to retroactively specify the data analytic process. One study of the effects of a drug used the subject's behaviorally reported mental state to obtain a temporal function that could be correlated with the brain imaging data.

The two memory studies each made use of the visual presentation of stimuli during the fMRI scanning session. In one case the stimuli were pictures; in the other case they were words. In both cases the subjects had an irrelevant task to perform related to these stimuli. After the scanning session was over, the subjects were given a memory task (without prior warning) to determine which of the test stimuli they recalled seeing. The key idea here was to group the imaging data according to whether the data were collected during the presentation of stimuli *that were subsequently remembered* versus the presentation of stimuli that were subsequently forgotten. The expectation (which was confirmed) was that various parts of the brain associated with long-term memory and encoding would have been more active on those trials in which the stimulus was subsequently recalled.

The drug study used behavioral measures more directly in analyzing the fMRI data. The study involved the challenging task of measuring fMRI changes during the administration of a psychoactive drug, cocaine. Subjects were regular cocaine users who had declined treatment but who had volunteered for a study. During the course of an imaging session they were given either a placebo or an injection of cocaine. (There were two imaging runs, so each subject received the cocaine on one run and the placebo on the other.) During the runs subjects regularly reported on their subjective state of "high," "low," "rush," and "craving"—terms that were known to be associated with cocaine experiences.

Note that there are unique technical challenges in this study. In addition to being a psychoactive drug, cocaine is also a cardiovascular stimulant. Therefore, before considering the possible effects of cocaine in its psychoactive and addicting role, it was necessary to demonstrate that such effects would not be masked by the circulatory effects of the drug. To accomplish this, the investigators used a standard stimulus (flashing light) to calibrate neuronally triggered hemodynamic responses. Also, they used two imaging pulse sequences: one that was sensitive to BOLD effects and

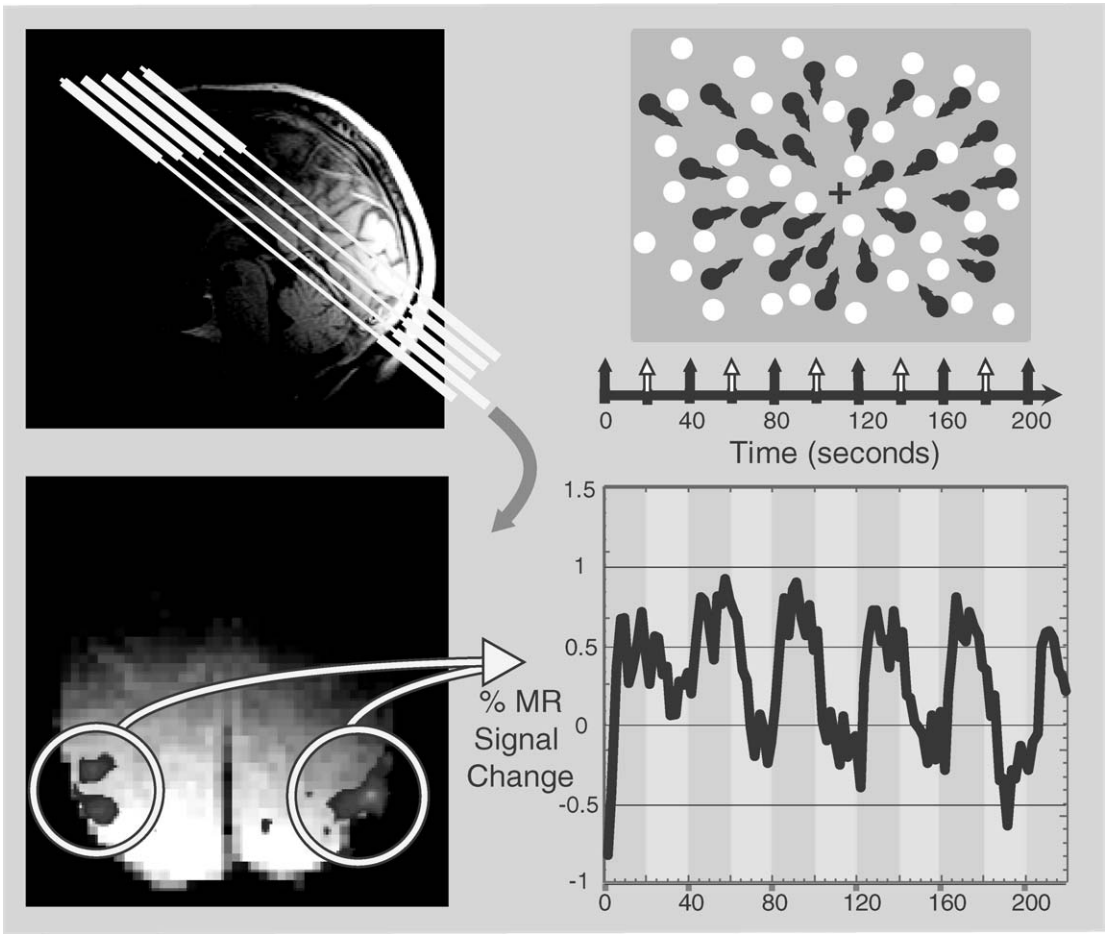


Figure 5 Voluntary attention modulates activity in human visual cortex. This collection of images summarizes an early fMRI-based experiment demonstrating the detection of neural modulation due to the exertion of voluntary attention. Subjects viewed a continuous movie consisting of a cross in the middle of the visual yield (on which they were instructed to fixate) and moving black dots and stationary white dots in the periphery. The cartoon at the upper right represents one frame of this movie, with arrows indicating the direction of motion. In the actual stimuli, of course, there were no arrows present—only moving, stationary dots and the fixation cross. The motion was always radial, toward the fixation point, to make it easy for the subjects to maintain fixation. Additional testing with an MR-compatible eyetracker revealed that subjects could maintain fixation. While looking at this movie for several minutes, subjects received verbal instructions to attend to one collection of colored dots or the other. These instructions alternated every 20 sec, as indicated in the figure by the black (''attend black'') and white (''attend white'') arrows on the timeline, below the visual stimulus. When the subjects heard ''attend black,'' they would continue to fixate the central cross, but they were supposed to pay more attention to the black (moving) dots than the white (stationary) dots during that time. Similarly, when they heard ''attend white,'' they were supposed to attend more to the white (stationary) dots. The imaging data were collected with a small coil of wire (about 13 cm in diameter) placed near the occipital cortex (the back of the head). This yielded a stronger signal in the brain regions of interest for the experiment, but it yielded weak signals from the rest of the brain, as indicated in the structural image shown at the upper left and the functional image shown in the lower left. Data for five slices, oriented parallel to the calcarine fissure (primary visual cortex), were collected (as indicated by the lines in the upper left). Data from one of those slices is shown in the lower left. As described for Fig. 4, a pseudocolor representation of the results of a statistical test comparing the MR data collected during one condition (''attend black'') versus the other condition (''attend white'') is displayed. There was a clear increase in activity on both the left and the right sides of this brain, in a region that corresponds anatomically to a known visual motion processing area of the cortex. Data from the voxels of the brain whose statistic exceeded the threshold used to specify the color map were averaged; the results are plotted as a function of time in the graph in the lower right. Data collected during the time that subjects were attending to moving stimuli (light background portions of the graph) were clearly higher in amplitude than the data collected during the time that subjects were attending to the stationary stimuli (dark background), *even though the visual stimulus was unchanged throughout the entire scanning period.* This experiment represents a simple, dramatic demonstration of the ability of fMRI to detect the neural consequences of changes in cognitive state (data and analysis courtesy of Kathleen O'Craven).

one that was sensitive to flow effects. This combination allowed them to demonstrate that cocaine influenced the flow-dependent MR signal changes but not the BOLD MR signal changes.

Thus, they could use the BOLD changes to study the effects of cocaine in its role as a psychoactive stimulant relatively independent of its role as a cardiovascular stimulant. Using the time course profile obtained from the behavioral ratings (specifically, the temporal modulation of craving and rush) they could find brain areas whose activity followed a similar profile, allowing them to conclude that those areas were implicated in the experience of these sensations.

D. Emotional Affect versus Cognitive Processing Load

A unique strength of functional brain imaging is the ability to test various intuitions and hypotheses about our mental activities by virtue of the quantitative nature of the MR signal changes. Sometimes the most salient aspect of a stimulus (such as its emotional valence) may be less cognitively engaging than the lack of that cue. Specifically, although the localization of function repeatedly found in studies of the low-level aspects of sensory processing appears to have analogs in other cognitive and emotional tasks, the tasks and stimuli that most effectively activate those areas may be counterintuitive. Higher contrast in visual stimuli generally evokes stronger modulation of early visual processing areas in the brain, but high contrast in an emotional domain does not always evoke the strongest variation in the brain areas associated with processing those stimuli. The data and experiment presented in Fig. 6 exemplify this idea.

Subjects were required to classify stimuli along an emotional scale as "neutral," "positive," or "negative." The stimuli were individual words (e.g., "calm," "delighted," and "insecure") intermixed with individual pictures of human faces (Fig. 6, top). There is ample evidence, both from the domain of human functional neuroimaging and from the literature of human brain lesions, that words and faces are exceptionally good stimuli for the activation of specific brain regions. Pictures of faces are good stimuli to activate the fusiform gyrus, especially on the right side. Single words, when associated with a semantic (rather than a purely perceptual) task, are effective stimuli for activating the inferior frontal region, almost always most strongly on the left side. One might reasonably expect that variation along the important dimension of emotional valence would modulate the strength of activation in these two areas. Given the demonstrated power of emotional stimuli to activate areas of the brain associated with general arousal, one might also predict that the most emotionally powerful stimuli (the positive and negative stimuli) would elicit greater neural activity than the neutral stimuli.

The stimuli were presented for 2 sec each, with 8 sec between the onset of each stimulus. This design, as one might deduce from Section II.A, is less than optimal because it does not take advantage of the power of rapid single trials, nor does it completely isolate the hemodynamic responses from the different stimulations. However, it was sufficient to address the preceding question, and it is useful in illustrating the way in which event-related fMRI data are often reported–via overlapping plots of the time course of the hemodynamic responses to the different stimulus types.

The stimuli were effective in eliciting emotional responses. Not surprisingly, the emotional faces (happy, in particular, and, to a lesser extent, sad) evoked stronger amygdala responses than neutral faces, with a weaker finding (but in the same relative order of strength) for words. Also not surprisingly, faces of any type elicited minimal response in left inferior frontal cortex and words of any type elicited minimal response in right fusiform cortex.

However, in the context of this categorization task, neutral stimuli (both faces and words) were far more effective activators of the regions known to be especially responsive to such stimuli (right fusiform and left inferior frontal, respectively) than were the emotionally powerful stimuli. The graphs in Fig. 6 show the time course of activity for each stimulus type in the relevant regions of the brain. The amplitude of the hemodynamic responses to the neutral stimuli was approximately twice as large as that to the positive stimuli, and it exceeded the response to the negative stimuli even more dramatically.

Interpretation of this result is not trivial. Behavioral measures of reaction time and accuracy for the words suggest one possible explanation. In the case of the word stimuli, the reaction times were longest, and percentage correct was lowest, for the neutral words. The fact that it was faster and easier to categorize positive and negative words could mean that the relevant brain area (left inferior frontal) was active for a shorter period of time during the emotional words, resulting in less time for the hemodynamic signal to grow. However, this kind of explanation cannot be the

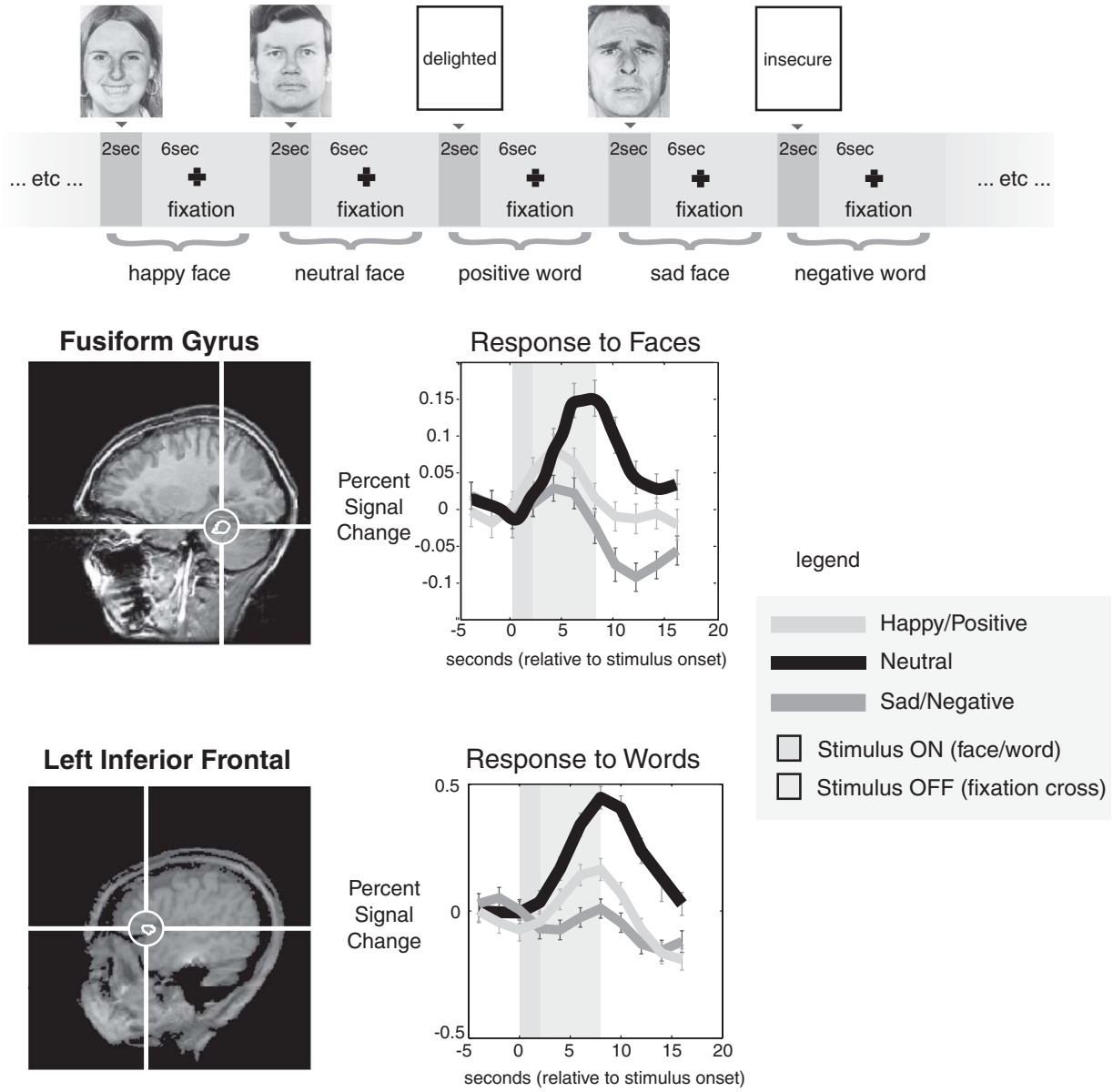


Figure 6 Strength of affect versus difficulty of processing. As described in Section II.A, detection of the hemodynamic responses to different types of events can reveal unintuitive effects. The stimuli and timing for this experiment are indicated in the top row. The stimuli were "positive," "negative," and "neutral" faces and words. One of these was presented for 2 sec every 8 sec in random order. Subjects were required to categorize each stimulus presentation as being positive, negative, or neutral. Two of the resulting activation foci are presented. A region of the brain known to respond strongly to face stimuli (the fusiform gyrus on the right side of the brain) and a region known to respond strongly to words (the inferior frontal region on the left side of the brain) were detected, and their locations are indicated by the crosshairs in the anatomical images shown on the left. The graphs on the right indicate the hemodynamic responses to the three classes of stimuli (positive, negative, and neutral). For words in the frontal region and for pictures of faces in the fusiform region, the most activity was elicited not by the most emotionally salient and affective stimuli but by the neutral stimuli. Apparently, categorizing happy or sad stimuli places a smaller processing load on these areas than categorizing a neutral stimulus (data and analysis courtesy of Patricia Deldin and David Cox).

whole story. In the case of face stimuli, it was the negative (rather than neutral) faces that had the longest reaction times and lowest percentage correct. Face stimuli were generally classified more quickly than words, but neutral faces may still evoke more processing because the observer tries harder to find positive or negative nuances (albeit quickly). However, these are speculations.

The point of recounting this study is to emphasize both the complexity of interpreting fMRI data and their potential nonintuitiveness. To an experimenter, it can naturally seem that positive and negative stimuli will, in one sense, evoke stronger responses than neutral stimuli. For some brain areas, this is no doubt true. However, the constraints of a classification task are different. In that context, the fact that neutral stimuli elicit more activity from their respective processing areas (for words and faces) can be plausibly interpreted as an indicator of greater processing effort.

VI. CLINICAL APPLICATIONS

The ability of fMRI to image brain activity *in vivo* makes it a promising tool for the diagnosis, interpretation, and treatment evaluation of clinical disorders involving brain function. A great deal of effort is currently being exerted to develop concrete clinical applications for fMRI as well as to use fMRI to better understand various psychiatric and neurological disorders. Until we better understand the wealth of data being generated from neurologically intact individuals, however, the use of fMRI-based data in actual clinical applications is likely to be limited.

The development of clinical applications of fMRI is of great importance to the fMRI community for many reasons. Functional MRI-based research has historically been closely tied to the medical community. Indeed, one reason why research using fMRI has been able to expand so rapidly and extensively is the availability of conventional MRI machines. MRI machines are very expensive, and any new, clinically relevant application would help subsidize costs of existing machines and the purchase of new machines (or the upgrading of existing machines) for fMRI use. Additionally, such demand encourages manufacturers to invest in the development of scanners that are more specifically designed with fMRI applications in mind. The widespread acceptance of the development costs for higher field (3 and 4 T) scanners by virtually all MRI manufacturers is almost certainly due to the potential promise of the neural-activation MRI and related imaging techniques, such as diffusion-weighted MRI and diffusion-tensor MRI (which are more relevant to the analysis of stroke and white matter anatomy, respectively).

There is at least one area of clinical importance in which fMRI is already playing an active role: presurgical planning. In situations in which a surgeon is going to be removing portions of a patient's brain, it is critical to know exactly where various motor and sensory functions are mapped in that individual's brain. Some information can be obtained during surgery via direct cortical stimulation or prior to the main surgery via a separate surgical procedure for the implantation of electrode arrays in the brain. Functional MRI, however, is a nonsurgical technique for obtaining similar information well before any surgical intervention starts. Furthermore, it is almost always the case that a structural MRI will be obtained prior to surgery, so it may be a relatively small incremental cost to obtain functional information during the same MRI session.

Many other areas of obvious *potential* clinical importance (in psychopharmacology, neurology, stroke treatment and recovery, drug addiction, and psychiatric disorders) are still largely in the stages of initial research. Although many studies have been performed, there are no specific clinical interventions that are driven by fMRI. A brief summary of some of the ideas in this area is presented in the following sections.

A. Preoperative Planning

One promising clinical application for fMRI is in preoperative planning and risk assessment. Since fMRI allows for the creation of maps of brain activity corresponding to the performance of a specific task, one natural use of fMRI is in identifying areas of the brain that are functionally important and therefore should be carefully avoided during neurosurgery. It should be emphasized that it is the creation of maps specific to the individual patient in question (rather than maps based on averages across many subjects) that is of use to the surgeon. Functional MRI is particularly well suited to the creation of such individualized maps.

Intractable focal epilepsy is sometimes treated by excising the epileptogenic focus surgically. In this context, ''intractable'' means ''not successfully treated with drugs,'' and ''focal'' means that there is a single site that appears to be the source or trigger for the epileptic seizures.) The areas to be removed are often close to brain areas critical to language (so-called ''eloquent'' cortex), so surgeons must exercise extreme caution lest their patients acquire severe language impairments as a result of the operation. For most adults, language function is strongly lateralized,

meaning that damage to one hemisphere (typically the left) has much greater consequences for language comprehension and production than damage to the other (typically the right) hemisphere. Approximately 95% of right-handed individuals are left lateralized for language, whereas left-handed individuals show a much more variable language lateralization (both in left versus right and in the degree of lateralization). Historically, the best procedure for assessing the side and degree of language lateralization in an individual patient has been the Wada test. In this procedure, amobarbitol is injected unilaterally into the left or right carotid artery, resulting in the anesthetization of the corresponding hemisphere of the brain. Both sides are tested in this way. If the subject is relatively unimpaired on language and memory tasks performed during the anesthetization of one hemisphere, then it is deemed safe to operate on that hemisphere. Functional MRI offers several advantages over the Wada test: It is completely noninvasive, and it provides a better estimate of areas important to language within each hemisphere. In the early days of attempting to use fMRI to assess language lateralization, the results were equivocal. Today, fMRI is a robust measure of language lateralization. On the other hand, fMRI has not been demonstrated to be as sensitive a measure of memory function as the Wada test, but progress is being made. At least one hospital in the United States has replace the Wada test with fMRI for presurgical planning, but use of this alternative is not widespread. Notwithstanding this slow start, it can be speculated that the first generally accepted clinical application of fMRI will be as a replacement for Wada test.

More detailed functional maps are being developed for presurgical planning and risk assessment relevant to other brain areas (besides eloquent cortex). For instance, during the surgical removal of tumors and other abnormalities near the central sulcus, one risk is the unnecessary removal of portions of primary motor cortex, resulting in serious impairments of motor control. Functional MRI has been used to map the boundaries of primary motor cortex in order to help weigh the risks and benefits of various treatment options. Obtaining an accurate, *individualized* functional map in this instance is particularly important because brain abnormalities in question often displace structures and make anatomical landmarks ambiguous as well as potentially disturbing the underlying functional maps of normal brain structures. Depending on the type of tumor, if fully resecting a tumor would endanger primary motor cortex, then partial resection might be in the patient's best interest. If, on the other hand, the entire tumor can be safely removed without endangering primary motor areas, then it is in the patient's best interest to remove the entire tumor.

B. Pharmacology

Pharmacology is another area in which fMRI has great potential. Although fMRI is poorly suited to identifying the binding sites of a drug (due to its inherent lack of sensitivity to chemicals in such relatively low concentrations), its good spatial resolution and moderate temporal resolution make it quite well suited to identifying which functional brain systems a drug influences. Studies of the action of clinically and socially significant drugs (such as addictive drugs) have revealed specific patterns and locations of activation via fMRI. Such studies (which include cocaine and nicotine and a growing list of psychoactive pharmaceuticals) are being conducted to learn more about how these drugs affect the brain. A better understanding of the anatomy and physiology of addiction may eventually lead to more effective treatments.

More generally, an important possible use of fMRI could be the determination, on an individual patient level, of whether or not a drug is affecting the appropriate brain systems and the quantification of the strength of that effect and thus potentially guide dosage. Since it is difficult to predict the dose–response effects of a drug on any given individual, fMRI could potentially speed the process of prescribing effective drugs in appropriate doses. Similarly, fMRI has the potential to aid drug development by quickly identifying the brain areas on which a drug acts, increasing knowledge about an existing drug, or helping to identify potential uses of a new drug. Finally, because fMRI has a temporal resolution that is rapid compared to most of the effects of psychoactive drugs, it is possible to use fMRI to follow the pharmacokinetic profile of such drugs.

C. Understanding Neurological and Psychiatric Disorders

In contrast to presurgical planning and some pharmacology, the application of fMRI-based studies to neurological and psychiatric disorders might better be characterized as occurring in the developmental rather than application stages. The primary thrust is in the area of refining diagnosis. The wealth of studies

using neurologically intact subjects supplies a natural baseline for using fMRI to derive more sensitive and/or more specific diagnostic criteria. For every robust finding in the functional localization of tasks involving frontal cortex with healthy subjects, there will eventually be a comparison study involving patients with all manner of neuropsychiatric disorders, from Alzheimer's disease to psychosis, schizophrenia, and autism. Many such studies have been conducted already.

The strength of fMRI is in obtaining spatially localized maps of function, especially in the context of purely cognitive function. These may be of value in diagnosing a variety of disorders. On the other hand, a weakness of fMRI is its extreme lack of sensitivity for directly detecting the presence and concentration of specific drugs in the brain—the discussion so far has referred to the detection of huge numbers of hydrogen atoms in water molecules. In contrast, PET detects virtually every decay of a radioactive atom in a single molecule and is consequently an exquisitely sensitive measure for localizing the sites of activity for suitably labeled psychoactive drugs.

Functional MRI is an *imaging* modality—it generates pictures that have a great deal of spatial specificity. Magnetic resonance spectroscopy (MRS), in contrast, sometimes collects data from the whole head to detect the presence and concentration of some of the body's more plentiful chemicals (such as lactate). There is a trade-off between detecting weak signals (there are far more water molecules than lactate molecules in the body) and being able to make spatial maps. As the MRI devices are made more sensitive (via stronger main magnets, better coils, etc.), the "whole head" MRS data can be refined to yield greater spatial resolution. The clinical relevance of MRS will increase in the future, and some of that improvement will be associated with blood flow changes and hence to fMRI.

D. Dyslexia

A particularly active area of clinically relevant research using fMRI is the study of developmental dyslexia. What makes this effort especially promising is that reading is of such fundamental importance to society and education, but there is limited understanding of the development and actions of the relevant cognitive systems underlying reading, even in normal readers. Psychophysical studies of temporal processing, not only in the auditory domain but also in the context of motor activity and coordination and in the visual perception of motion, have all been implicated, by at least some researchers, in the etiology of dyslexia. Coupling fMRI with behavioral studies in this context has especially rich appeal.

Again, as with some of the preceding potential clinical applications, the likelihood is not for a treatment based on fMRI, but, rather, an improved ability to differentially diagnose different types dyslexia and to monitor the effectiveness and modes of action of any behavioral or drug-based treatments. Dyslexia is a complex disorder, with an etiology that is likely to include low-level physiology, high-level cognitive structures, and other levels as yet undetermined. It is not clear whether fMRI-based research will be able to tap into all aspects of dyslexia, but initial work is encouraging. The discussion of clinical applications of fMRI closes with a simpler disorder than dyslexia, but one that couples directly with the particular strengths of hemodynamically based fMRI.

E. Migraine

A recent tour de force in fMRI-based experimentation brings together some of the most elegant work in a research application context (retinotopic mapping of visual cortex) and a long-standing phenomenon of clinical importance (migraine headaches). Migraines are an intense form of headache, often associated with visual auras (i.e., the perception of various strange visual patterns, typically around a circular arc or perimeter of some portion of the visual field, bilaterally) and an associated temporary blindness (a temporary scotoma) within that perimeter. The fact that these auras and scotomas appear to both eyes at the same portion of the visual field is very strong evidence that the underlying effect is being controlled at the cortical level, where these corresponding portions of the visual field share the same physical location in the brain. Moreover, migraines have long been understood to be associated with changes in dilation and constriction of the cerebral vasculature.

It is very difficult to study this phenomenon using fMRI both because it is relatively short-lived (sometimes 30–60 min and sometimes 2–4 hr) and because it is associated with aversion to loud noises and bright lights (on the part of the sufferer). Therefore, it is difficult to get migraine sufferers to volunteer for an fMRI study; even if they were willing, it would be rare that they got a migraine while they were near the scanner. One research group was fortunate enough to find a volunteer who predictably and regularly triggered his own migraine headache by dint of intense athletic activity

(playing basketball). He was therefore available for repeated (schedulable) scanning immediately before and during the onset of his migraine attacks.

The investigators were experts in visual retinotopy and they designed a protocol that revealed, in exquisite detail, the neurological correlate of the patient's visual symptomotology. As the scotoma grew and as the aura changed in size (both of which phenomena could be reported subjectively by the patient), fMRI data revealed the location on the cortex and the functional variation in amplitude of response to a flickering checkerboard of visual stimulation. Combining this data with previously obtained retinotopic maps of the subjects visual cortex permitted a precise connection between measurable function and subjective loss. Although this study does not immediately suggest a treatment for migraine attacks, it certainly demonstrates a method for objectively assessing the effectiveness of candidate therapies.

VII. CLOSING

Given the strong technical components of fMRI, perhaps it is appropriate to close with a discussion of the ultimate spatial and temporal resolutions that might be obtainable in the near future. The temporal resolution of fMRI is not likely to be limited by the imaging tool but, rather, by the vagaries of the hemodynamic response. We are probably close to that limit already. Although there are aspects of temporal properties of hemodyamics that we can detect on the timescale of 100 msec, for most practical purposes the fMRI temporal resolution limit is likely to remain approximately 1 sec. Significant improvements in temporal resolution *associated* with fMRI are likely to result from integration with modalities such as EEG and MEG, whose intrinsic temporal resolution is the millisecond range.

Technological developments, especially in terms of higher field magnets and more sensitive and versatile imaging coils, will increase the effective spatial resolution of fMRI. Because higher field strength increases the signal-to-noise ratio, less averaging of signals across space is required, allowing for smaller voxels and thus greater spatial resolution. Although there are some safety issues that arise as field strength increases, there is no known reason why MRI cannot be done with humans at much higher field strength than the conventional 1.5 T. An increasing number of 3- and 4-T scanners are in routine use, and a handful of 7- and 8-T scanners for use on humans are either currently in use or are being built. Even without the higher field strengths, it would be possible to increase spatial resolution by using longer imaging times. However, the ultimate spatial resolution limits will not be determined by the MR scanner. Rather, they will be determined by hemodynamic limits, i.e., by the spatial resolution of the smallest vessels that show local changes with neural activity. Based on the available evidence, this limit is approximately the size of a cortical column—between 0.1 mm and 1 mm in linear dimension. If that spatial resolution could be achieved in routine fMRI-based experiments, it should represent a dramatic leap in the ability to develop and test far more interesting and explicit models of functioning neural systems in the human brain.

See Also the Following Articles

CEREBRAL CIRCULATION • DYSLEXIA • ELECTROENCEPHALOGRAPHY (EEG) • EVENT-RELATED ELECTROMAGNETIC RESPONSES • HEADACHES • IMAGING: BRAIN MAPPING METHODS • MAGNETIC RESONANCE IMAGING (MRI) • PSYCHOPHYSIOLOGY • RECEPTIVE FIELD

Suggested Reading

Beauchamp, M. S., Cox, R. W., and DeYoe, E. A. (1997). Graded effects of spatial and featural attention on human area MT and associated motion processing areas. *J. Neurophysiol.* **77**(7), 516–520.

Breiter, H. C., Gollub, R. L., Weisskoff, R. M., Kennedy, D. N., Makris, N., Berke, J. D., Goodman, J. M., Kantor, H. L., Gastfriend, D. R., Riorden, J. P., Mathew, R. T., Rosen, B. R., and Hyman, S. E. (1997). Acute effects of cocaine on human brain activity and emotion. *Neuron* **19**(9), 591–611.

Brewer, J. B., Zhao, Z., Desond, J. E., Glover, G. H., and Gabrieli, J. D. E. (1998). Making memories: Brain activity that predicts how well visual experience will be remembered. *Science* **281,** 1185–1187.

Bushong, S. C. (1996). *Magnetic Resonance Imaging: Physical and Biological Principles*, 2nd ed. Mosby–Year Book, Boston.

Calvin, W. H., and Ojemann, G. A. (1995). *Conversations with Neil's Brain: The Neural Nature of Thought and Language*. Perseus, Reading, MA.

Gollub, R. L., Breiter, H. C., Kantor, H., Kennedy, D., Gastfriend, D., Mathew, R. T., Makris, N., Guimaraes, A., Riorden, J., Campbell, T., Foley, M., Hyman, S. E., Rosen, B., and Weisskoff, R. (1998). Cocaine decreases cortical cerebral blood flow but does not obscure regional activation in functional magnetic resonance imaging in human subjects. *J. Cereb. Blood Flow Meta.* **18,** 724–734.

Hadjikhani, N., Sanches del Rio, M., Wu, O., Schwartz, D., Bakker, D., Fischl, B., Kwong, K. K., Cutrer, M. F., Rosen, B. R., Tootell, R. B. H., Sorensen, A. G., and Moskowitz, M. A. (2001). Mechanisms of migraine aura revealed by functional MRI

in human visual cortex. *Proc. Natl. Acad. Sci. USA* **98**(8), 4687–4692.

Moonen, C. T. W., and Bandettini, P. A. (Eds.) (1999). *Functional MRI*. Springer-Verlag, Berlin.

O'Craven, K. M., and Kanwisher, N. (2000). Mental imagery of faces and places activates corresponding stimulus-specific brain regions. *J. Cognitive Neurosci.* **12**(6), 1013–1023.

O'Craven, K. M., Downing, P. E., and Kanwisher, N. (1999). fMRI evidence for objects as the units of attentional selection. *Nature* **401,** 584–587.

O'Craven, K. M., Rosen, B. R., Kwong, K. K., Treisman, A., and Savoy, R. L. (1997). Voluntary attention modulates fMRI activity in human MT/MST. *Neuron* **18,** 591–598.

Savoy, R. L. (2001). History and future directions of human brain mapping and functional neuroimaging. *Acta Psychol.* **107**(1–3), 9–42.

Toga, A. T., and Mazziotta, J. C. (Eds.) (2000). *Brain Mapping: The Systems*. Academic Press, San Diego.

Tootell, R. B. H., Dale, A. M., Sereno, M. I., and Malach, R. (1996). New images from human visual cortex. *Trends Neurosci.* **19**(11), 481–489.

Wagner, A. D., Schacter, D. L., Rotte, M., Koutstaal, W., Maril, A., Dale, A. M., Rosen, B. R., and Buckner, R. J. (1998). Building memories: Remembering and forgetting of verbal experiences as predicted by brain activity. *Science* **281,** 1188–1191.

GABA

SOFIE R. KLEPPNER and ALLAN J. TOBIN
UCLA Brain Research Institute

GLOSSARY

apoenzyme An enzyme without its obligate cofactor, which is therefore inactive.

carrier-mediated release Nonsynaptic release (usually of a neurotransmitter) by a plasma membrane transporter.

GABA γ-Aminobutyric acid; the major inhibitory neurotransmitter of the central nervous system; also found in extraneuronal tissues (e.g., the pancreas and reproductive tracts).

GABA receptor A protein that binds GABA and initiates downstream effects.

GABA-T GABA-transaminase; enzyme that degrades GABA.

GABA transporter A protein that moves GABA across either the plasma membrane or the vesicular membrane.

GAD Glutamic acid decarboxylase; enzyme that converts glutamate to GABA.

holoenzyme An enzyme bound to its obligate cofactor, which is therefore active.

inhibitory neurotransmitter A neurotransmitter that, when it activates its postsynaptic receptor, decreases the probability that the postsynaptic cell will fire.

phasic inhibition Short-term (milliseconds) decrease in excitability of a cell, usually synaptically mediated.

pyridoxal phosphate Vitamin B6; a cofactor required for GAD activity.

receptor agonist A substance that binds to a receptor and mimics or enhances the receptor response.

receptor antagonist A substance that binds to a receptor and blocks or decreases the receptor response.

tonic inhibition Long-lasting decrease in the overall excitability of a cell or cells, usually extrasynaptically mediated.

GABA (γ-aminobutyric acid; 4-aminobutyric acid) is an inhibitory amino acid neurotransmitter. Three separate groups published the first reports of GABA in the brain in 1950, although its function as a neurotransmitter was not recognized until later. Eugene Roberts, the author of one of the first reports, identified GABA in the course of chromatographic studies on the amino acid profiles of murine neuroblastomas. GABA from potatoes was used as a standard and comigrated with an unknown substance isolated from the tumors. Roberts rigorously pursued his studies of GABA synthesis and degradation in the brain and GABA is now recognized as the most prominent inhibitory neurotransmitter in the central nervous system (CNS).

I. INTRODUCTION

Approximately 30% of neurons in the brain produce GABA, and almost every neuron can respond to GABA. Some nonneural cells also make GABA, including the cells of the endocrine pancreas and the reproductive tracts. GABA in the pancreas presumably acts as a signaling molecule, in a similar manner to CNS signaling. GABA function in the reproductive tracts is unknown.

The molecules associated with GABA synthesis, degradation, transport, and signaling include two synthesizing enzymes, one degrading enzyme, two transporting proteins, and two classes of receptors. These molecules, their functions, and their putative locations are listed in Table I and illustrated in Fig. 1.

GABA acts by binding to GABA receptors, which are located on both pre- and postsynaptic cells. Generally, GABA receptor binding hyperpolarizes the cell, producing an inhibitory effect. In some circumstances, however, GABA can exert excitatory effects by depolarizing the membrane. Excitatory GABA effects occur most notably during development but also in the rodent suprachiasmatic nucleus, where it may be related to circadian rhythms.

Table I
GABA-Related Proteins, Their Locations, and Their Functions[a]

Protein	Location	Function
GAD	Intracellular	Synthesizes GABA
GAD_{65}	Axon terminal	
GAD_{67}	Cytosol	
GABA-T	Mitochondria	Degrades GABA
GAT		Transports GABA across plasma membrane
GAT1	Neurons and glia	
GAT2	Glia and periphery	
GAT3	Glia and periphery	
GAT4	Brain and periphery	
vGAT	Synaptic vesicle membrane	Packages GABA into vesicle
$GABA_A$	Postsynaptic membrane	Ionotropic receptor (chloride channel)
α_{1-6}		
β_{1-4}		
γ_{1-3}		
δ		
ρ_{1-3}		
π		
ε		
$GABA_B$	Pre- and postsynaptic membrane	Metabotropic receptor (G protein linked)
R1		
R2		

[a]Abbreviations used: GAD, glutamic acid decarboxylase; GABA-T, GABA-transaminase; GAT, GABA transporter; vGAT, vesicular GABA transporter.

Beyond neurotransmission, GABA also acts as a signaling molecule during development, both in the CNS and elsewhere in the embryo. In the CNS, GABA can influence cell migration, neurite extension, differentiation, and synapse formation. GABA also plays an important role in the actions of steroid hormones in the brain. Outside the CNS, GABA is associated with developmental functions including palate formation.

GABA is found in organisms that lack synapses (e.g., bacteria and plants); therefore, it may have an important role outside of intracellular signaling. GABA can be degraded into succinate, which participates in the tricarboxylic acid cycle. Thus, GABA can provide energy for the cell, and this may be part of its role in nonsynaptic functions.

Alterations in GABA production, degradation, response, and transport have tremendous effects. Within the brain, changes in GABA signaling contribute to epilepsy, movement disorders (e.g., Huntington's disease and Parkinson's disease), anxiety, and panic disorder. Injury can influence the expression of glutamic acid decarboxylase, the enzyme that synthesizes GABA, suggesting that GABA may play a role in recovery. Outside of the brain, a lack of GABA in the embryo is associated with cleft palate, a craniofacial malformation. Autoimmunity to glutamic acid decarboxylase is associated with the death of pancreatic β cells and the development of insulin-dependent diabetes mellitus.

II. GABA PRODUCTION AND DEGRADATION

GABA is unique among neurotransmitters because it can be synthesized by either of two closely related enzymes. Almost all GABA derives from the decarboxylation of glutamate, catalyzed by glutamic acid decarboxylase [glutamate decarboxylase (GAD); L-glutamate 1-carboxylase, E.C. 4.1.1.15]. Pyridoxal phosphate [PLP; derived from pyridoxine (vitamin B_6)] is an obligate cofactor for GAD, as it is for many other decarboxylases. GABA can also be synthesized from ornithine via ornithine decarboxylase, although this is a minor route of synthesis. The GABA-synthesizing pathway is illustrated in Fig. 2.

A. GABA Synthesis

GAD consists almost entirely of homodimers of two distinct polypeptides—GAD_{65} (with M_r of 65,000)

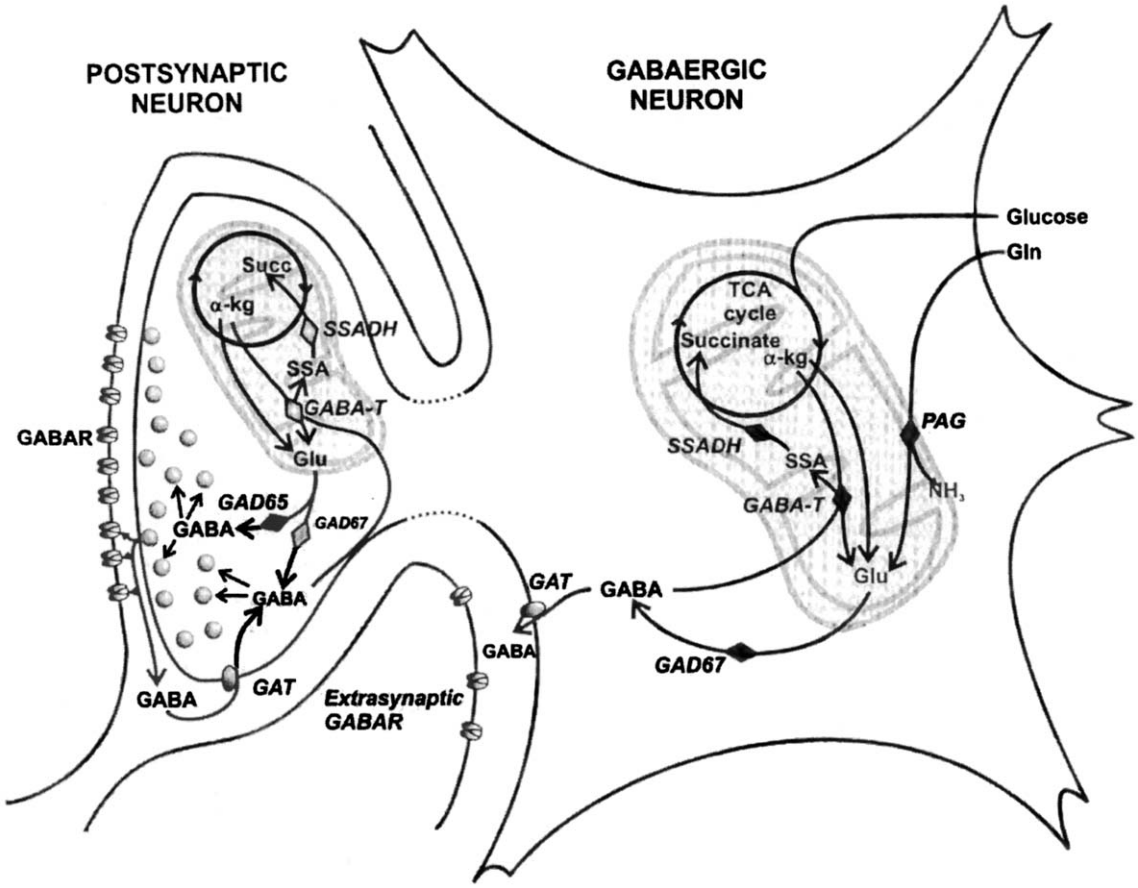


Figure 1 Schematic of GABA-related proteins and their locations within the cell. GAD_{65} or GAD_{67} synthesizes GABA from glutamate. GABA can be packaged into vesicles for release (vesicular GABA transporter not shown) at the synapse, where it binds to GABA receptors. GABA receptors are located on the presynaptic and postsynaptic neuron and can be found outside of the synapse. The plasma membrane GABA transporter (GAT) takes up unbound GABA from the synapse. GABA can also exit the cell via GAT. The GABA shunt is depicted within the mitochondria. Also shown are two sources of glutamate: α-Ketoglutarate can be converted to glutamate by GABA-T, and glutamine, provided by astrocytes, can be converted to glutamate by phosphate-activated glutaminase (PAG). GABAR, GABA receptor; GAD, glutamic acid decarboxylase; GABA-T, GABA-transaminase; α-kg, α-ketoglutarate; Succ, succinate; SSA, succinic semialdehyde; SSADH, succinic semialdehyde dehydrogenase; Gln, glutamine [from D. L. Martin and A. J. Tobin, Mechanisms controlling GABA synthesis and degradation in the brain. In *GABA in the Nervous System: The View at Fifty Years* (D. L. Martin and R. W. Olsen, Eds.). Lippincott, Williams & Wilkins, Philadelphia, 2000].

and GAD_{67} (with M_r of 67,000). The two GADs account for essentially all GAD activity in tissue extracts. Although other purified proteins reportedly have GAD activity, none is known to contribute substantially to GABA synthesis *in vivo*.

The two GAD polypeptides are the products of distinct genes and differ in sequence by about 35%. The two genes have identical exon–intron organizations, suggesting that they share a relatively recent common ancestor. The GAD_{65}s of humans, mice, and rats are 97% identical, as are the GAD_{67}s of humans, rats, and cats. Clearly, the two GADs have been under strong selective pressure during the 150–200 million years since the beginning of mammalian radiation.

Almost every neuron that produces one form of GAD also produces the other, suggesting that the two genes share some transcriptional regulatory mechanisms. During the development of specific brain structures, GAD_{67} usually appears earlier than GAD_{65}. In the hippocampus and the spinal cord, however, GAD_{65} appears earlier in development than GAD_{67}. The ratio of GAD_{65} to GAD_{67} mRNA increases dramatically during synapse formation, both in the striatum and in the cerebellum.

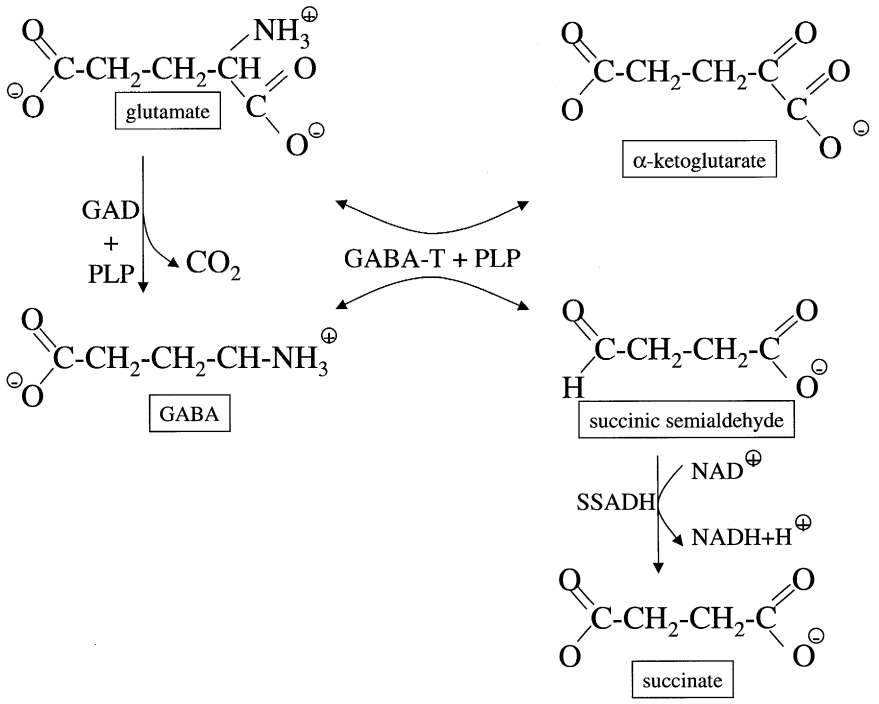


Figure 2 GABA is synthesized from the decarboxylation of glutamate. GABA is degraded to succinic semialdehyde and succinate by GABA-T and succinic semialdehyde dehydrogenase (SSADH). GABA-T can also convert α-ketoglutarate to glutamate, completing the first step of the GABA shunt.

The GAD_{67} gene produces two alternatively spliced transcripts during early development. The two embryonic forms of GAD are GAD_{25} and GAD_{44}, but only GAD_{44} is enzymatically active. GAD_{25} is not enzymatically active because it lacks the PLP binding site. As the embryo develops, the shorter transcripts are replaced by the mature GAD_{67}. GAD_{65} has no identified alternatively spliced transcripts.

GAD_{65} and GAD_{67} differ in their enzymatic characteristics. GAD_{67} exists mainly in the active, holoenzyme form (bound to PLP) and is less sensitive to small changes in neuronal GABA concentration than is GAD_{65}. Most GAD_{65} in the brain is in the apoenzyme form (inactive, not bound to PLP), and the GAD activity of brain extracts increases two- or threefold upon the addition of PLP. The loss of PLP from GAD is not a simple dissociation but a catalytic misstep that results in the formation of succinic semialdehyde and pyridoxamine phosphate (rather than GABA and PLP). This reaction is illustrated in Fig. 3. Pyridoxamine phosphate dissociates from GAD, generating an apoenzyme that lacks enzymatic activity until it again combines with PLP to reform holoGAD_{65}.

The interconversion of apoGAD and holoGAD is sensitive to ATP, phosphate (Pi), and GABA levels. These influences may regulate GABA production in response to altered neuronal activity within GABA neurons. ATP favors the formation of apoGAD and inhibits the formation of holoGAD. Pi enhances the formation of holoGAD. GABA also favors apoGAD formation in the absence of PLP because the GABA-forming steps are readily reversible. Consequently, GAD_{65} is sensitive to small changes in neuronal GABA concentration. GAD has a K_m for glutamate of approximately 0.45 mM and a K_m for GABA of approximately 16 mM. Therefore, given similar concentrations of glutamate and GABA, glutamate is converted to GABA much faster than the reverse reaction.

GAD_{65} and GAD_{67} also differ in their subcellular locations. Both are present in cell bodies and axon terminals, but GAD_{65} is usually more concentrated in axon terminals, whereas GAD_{67} is more concentrated in cell bodies. GAD_{65} associates with vesicle membranes both in neurons and in pancreatic cells, and it is characterized by punctate immunostaining in mature neurons. GAD_{67} generally shows no such association with vesicles, and it is diffuse throughout the cell when examined by immunostaining.

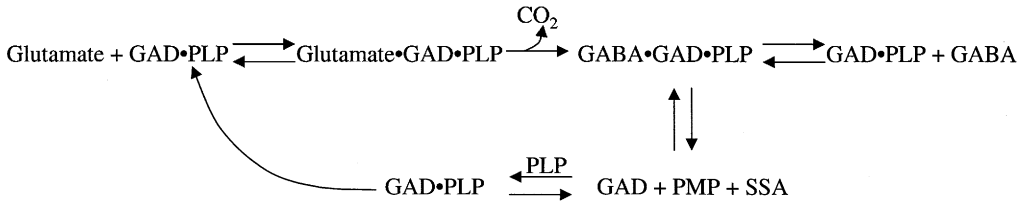


Figure 3 Interconversion of the GAD apoenzyme and holoenzyme. GAD can be converted from its active holoenzyme form, bound to pyridoxal phosphate (PLP), into its inactive apoenzyme form by a catalytic misstep that results in the formation of pyridoxal monophosphate (PMP) and succinic semialdehyde (SSA). GAD must reassociate with PLP to be active [adapted from Brain Glutamate Decarboxylase. In *Neurotransmitter enzymes* (Boulton, Baker, and Yu, Eds.) Humana Press, Clifton, NJ].

GAD_{65} undergoes at least two types of reversible posttranslational modifications—palmitoylation and phosphorylation. These modifications can anchor GAD_{65} to internal membranes, even in cells not specialized for exocytosis, but they are not required for membrane association. Palmitoylation and phosphorylation are limited to a distinct GAD_{65} polypeptide whose electrophoretic mobility is slightly less than that of GAD_{65}, but the structural basis for this difference in mobility is not well understood.

Although most GAD molecules are homodimers, GAD_{65} and GAD_{67} also form heterodimers. The association of some GAD_{67} with synaptic terminals and with the Golgi complex in genetically modified cells appears to depend on its association with membrane-targeted GAD_{65}. The presence of GAD_{67} in a restricted subset of synaptic boutons, observed in the mouse hippocampus, may reflect the distribution of GAD_{65}–GAD_{67} heterodimers.

B. GABA Degradation: The "GABA Shunt"

GABA-transaminase (GABA-T) is the main degradative enzyme for GABA, although steady state GABA levels are normally controlled by GAD. GABA is degraded to succinic semialdehyde (SSA) by GABA-T, a mitochondrial enzyme. Since SSA can be converted to succinate by succinic semialdehyde dehydrogenase (SSADH), GABA can serve as one step in a shunt that bypasses α-ketoglutarate dehydrogenase in the tricarboxylic acid (TCA) cycle. The GABA shunt begins with the conversion of α-ketoglutarate to glutamate by GABA-T, other transaminases, and glutamate dehydrogenase. Glutamate is then decarboxylated by GAD to form GABA, which is degraded to SSA by GABA-T. Finally, SSA is converted to succinate by SSADH. These reactions are summarized in Fig. 4. Normally, the conversion of α-ketoglutarate to succinate produces one NADH and one GTP. The GABA shunt therefore produces approximately 8% less total energy compared to the TCA cycle, although it provides 10–20% of the TCA cycle activity in most brain regions.

III. GABA RESPONSE

Responses to GABA depend on two classes of GABA receptor—$GABA_A$ receptors, which are located on the postsynaptic cell and produce fast (millisecond) responses, and $GABA_B$ receptors, which are found on both pre- and postsynaptic cells and produce slower (second) responses that depend on second messengers. A third class of GABA receptors, the $GABA_C$ receptors, is also ionotropic and is now generally considered to be a subset of $GABA_A$ receptors. The differences in GABA receptor structure, function, and location allow fine-tuning of GABA signaling and response.

A. $GABA_A$ Receptors

Most neurons express $GABA_A$ receptors. $GABA_A$ receptors are GABA-gated chloride channels located on the postsynaptic membrane. Upon binding GABA, the channel opens and chloride flows along its concentration gradient. The extracellular chloride concentration is usually higher than the intracellular concentration, so chloride usually flows into the cell. The inward chloride flux results in hyperpolarization of the postsynaptic membrane and a concomitant decrease in the probability of cell firing. In some cases, however, GABA evokes a depolarizing response, either because of high intracellular chloride concentrations resulting in an outward flow of chloride through the open channel or because of the flow of bicarbonate ions through the channel.

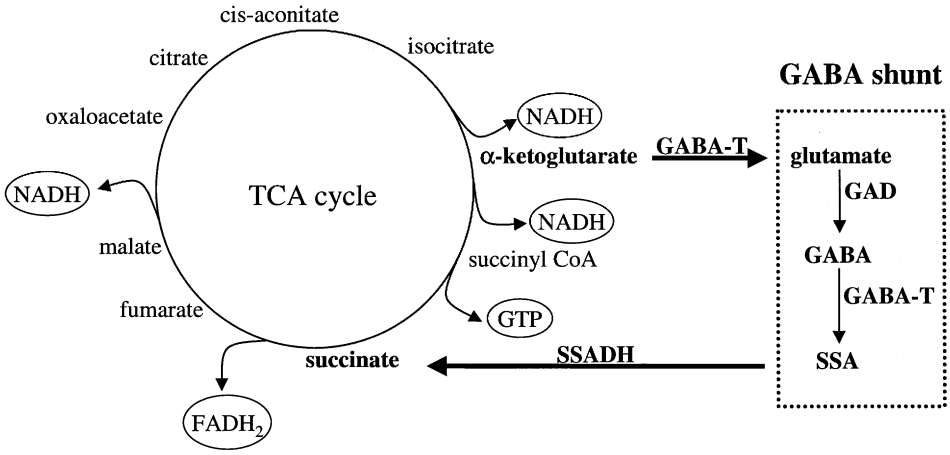


Figure 4 The GABA shunt. GABA can contribute to the TCA cycle via the GABA shunt, which converts α-ketoglutarate to succinate. The GABA shunt provides less energy than the complete TCA cycle because it bypasses the formation of one NADH and one GTP. Nonetheless, approximately 10–20% of TCA cycle activity in most brain regions is provided by the GABA shunt. SSA, succinic semialdehyde; SSADH, succinic semialdehyde dehydrogenase.

1. $GABA_A$ Receptor Topology

$GABA_A$ receptors are highly diverse. Each $GABA_A$ receptor consists of five transmembrane polypeptide subunits. At least 19 subunits exist, named $\alpha(1–6)$, $\beta(1–4)$, $\gamma(1–3)$, δ, ε, $\rho(1–3)$, and π. The recently identified subtype θ may be identical to β_4. Splice variants of several subunits exist as well. Receptor heterogeneity results from at least 20 different combinations of five subunits.

Each subunit consists of a long extracellular amino-terminal region, four transmembrane domains, and a large intracellular loop between the third and fourth transmembrane domains (Fig. 5A). This motif is shared by the channel superfamily that includes nicotinic acetylcholine receptors, glycine receptors, and serotonin 5-HT_3 receptors. Five subunits form a complex with a central pore (Fig. 5B). The amino-terminal domains of the α and β subunits are believed to be exclusively responsible for the GABA binding site. Receptor gating is almost certainly mediated by the M2 regions of all five subunits, which presumably line the pore, as is the case in related receptors.

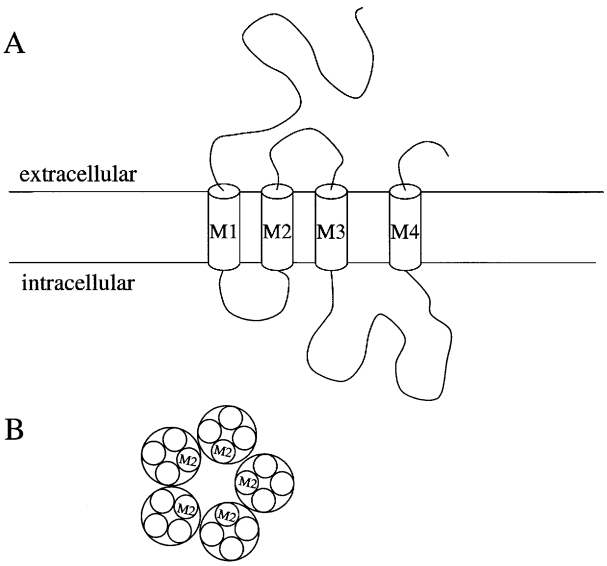


Figure 5 $GABA_A$ receptor subunit and complex topology. (A). Each $GABA_A$ subunit has four transmembrane domains (M1–M4). The amino terminus is extracellular and the carboxy terminus is intracellular. (B). Top view of five subunits that form the receptor complex with a central pore, which is likely lined by the M2 domains of each subunit [reprinted from Cherubini and Conti, *TINS* **24**(3), 155–162, 2001, with permission of Elsevier Science].

2. $GABA_A$ Subunit Composition

The $GABA_A$ subunit genes are located on several chromosomes. For example, human α_1 is located on chromosome 5, α_2 on chromosome 4, and α_5 on chromosome 15. Many of the subunit genes occur in clusters. For instance, $\alpha_1/\alpha_6/\beta_2/\gamma_2$ are clustered on chromosome 5, whereas $\alpha_2/\alpha_4/\beta_1/\gamma_1$ are located on chromosome 4. The chromosomal arrangement of the subunit genes suggests that both gene and cluster duplication occurred during evolution.

Despite the number of potential subunit combinations, there are some common combinations, which are summarized in Table II. Both development and brain region regulate $GABA_A$ subunit composition and presumably each combination has unique properties based not only on each subunit but also on the assembly as a whole. The functional consequences of specific combinations are not well understood but can affect both receptor function and location.

The subunit composition affects receptor binding of GABA and other ligands, but it is also important in receptor targeting, assembly, and clustering. For instance, γ_2 knockout mice show reduced $GABA_A$ receptor clustering as well as reduced ligand binding. α_6 knockout mice also demonstrate decreased δ subunit expression, despite normal mRNA levels, indicating posttranscriptional control by the subunits.

3. Pharmacology of $GABA_A$ Receptors

The $GABA_A$ receptor is affected by a myriad of drugs, including those that both cause and prevent convulsions, those that relieve anxiety, and those that relax, sedate, and anesthetize. Receptor subunit composition is critical in determining drug action. The specific binding sites for many but not all compounds have been identified.

Benzodiazepines (BZs) constitute a class of drugs that act on the $GABA_A$ receptor with sedative, anxiolytic (antianxiety), muscle relaxant, and cognitive effects. BZs bind to the external surface of receptors with specific combinations of subunits: A combination of γ_2 with α_1, α_2, α_3, or α_5 and any β subunit confers BZ binding. Each α subunit may have its own BZ binding affinity. The BZ binding site is thought to be between the γ_2 and α subunits and is highly homologous to the GABA-binding site located between the α and β subunits. Recent work has shown that the α_1 subunit mediates the sedative but not the anxiolytic effect of BZs.

Anticonvulsant and anesthetic drugs can positively or negatively modulate the receptor response to GABA. For example, barbiturates can directly enhance the GABA response either by opening the chloride channel or by increasing the time for which it remains open after binding to GABA. Very high concentrations of barbiturates can block the channel entirely, however. Barbiturates bind within the $GABA_A$ receptor pore and can also act at other receptors, including glutamate and acetylcholine receptors. This may reflect conservation of specific residues within the subunits. The anesthetic binding site is different from the GABA binding site, and it is probably located within the pore.

Table II
Common $GABA_A$ Receptor Subtype Combinations and Their Locations[a]

Subtype	Comments
$\alpha_1\beta_2\gamma_2$	Approx 50% of the total $GABA_A$ receptors; widespread; GABAergic interneurons
$\alpha_2\beta_3\gamma_2$	Approx 15–20% of total $GABA_A$ receptors; cortex, hippocampus, amygdala, septum, hypothalamus
$\alpha_3\beta_3\gamma_2$ ($\alpha_1\alpha_3\beta_3\gamma_2$)	Approx 15–20% of total $GABA_A$ receptors; cortex, amygdala, septum, raphe; monoaminergic, serotonergic neurons
$\alpha_4\beta_x\gamma_2$	$<5\%$ of total $GABA_A$ receptors in whole brain; cortex, hippocampus, thalamus, striatum
$\alpha_4\beta_x\delta$	$<5\%$ of total $GABA_A$ receptors in whole brain; cortex, hippocampus, thalamus, striatum
$\alpha_5\beta_3\gamma_2$	$<5\%$ of total $GABA_A$ receptors in whole brain; hippocampus (approx 20% of total $GABA_A$ receptors), cortex
$\alpha_6\beta_x\gamma_2$ ($\alpha_1\alpha_6\beta_x\gamma_2$)	$<5\%$ of total $GABA_A$ receptors in whole brain; cerebellar granule cells (30–40% of total $GABA_A$ receptors)
$\alpha_6\beta_x\delta$ ($\alpha_1\alpha_6\beta_x\delta$)	$<5\%$ of total $GABA_A$ receptors in whole brain; cerebellar granule cells (20–30% of total $GABA_A$ receptors); extrasynaptic
$\alpha_2\beta_1\gamma_1$ $\alpha_2\beta_1\gamma_1\theta$	$<10\%$ of total $GABA_A$ receptors in whole brain; limbic regions, basal ganglia, tyrosine hydroxylase-positive neurons, Bergmann glia
γ_3-containing receptors	$<5\%$ of total $GABA_A$ receptors in whole brain; widespread, low abundance; little data available
ε-containing receptors	$<5\%$ of total $GABA_A$ receptors in whole brain; hippocampus, hypothalamus; little data available

[a]Adapted from P. J. Whiting, K. A. Wafford, and R. M. McKernan, Pharmacologic subtypes of $GABA_A$ receptors based on subunit composition. In *GABA in the Nervous System: The View at Fifty Years*. (D. L. Martin and R. W. Olsen, Eds.). Lippincott, Williams & Wilkins, Philadelphia (2000).

Alcohol acts on the $GABA_A$ receptor, although its route of action and potential binding site are unclear. In some cases, alcohol potentiates GABA function at the $GABA_A$ receptor, but this effect may depend on receptor subtypes or on an indirect action. Currently, there appear to be alcohol-sensitive and alcohol-insensitive $GABA_A$ receptors, but the basis for that sensitivity is unclear.

B. $GABA_B$ Receptors

$GABA_B$ receptors are located both pre- and postsynaptically and they appear to be largely extrasynaptic. $GABA_B$ receptors are G protein-coupled receptors that interact with a number of proteins on K^+ and Ca^{2+} channels as well as with adenylate cyclase. They belong to the class C metabotropic receptor family, which includes the metabotropic glutamate receptors. The $GABA_B$ receptor has only recently been cloned, and our understanding of this receptor is therefore more limited than that of the $GABA_A$ receptor.

$GABA_B$ activation is inhibitory both pre- and postsynaptically. Presynaptic $GABA_B$ receptor activation causes Ca^{2+} channels to close, which decreases Ca^{2+} conductance. Decreased Ca^{2+} conductance at the nerve terminal reduces exocytotic release of neurotransmitter and also decreases the action potential duration. Postsynaptic $GABA_B$ receptor activation causes inwardly rectifying K^+ channels to open, which increases K^+ conductance. Increased K^+ conductance causes an increase in extracellular K^+ and a concomitant hyperpolarization, rendering Na^+ channels inactive. Thus, the net result of GABA binding to $GABA_B$ receptors is a decrease in the probability of cell firing.

1. Receptor Topology

$GABA_B$ receptors are unique within the metabotropic receptor family because they exist as heterodimers. Functional $GABA_B$ receptors require the GBR1 and GBR2 subunits in order to be expressed at the cell surface, to bind ligand, and to mediate action. Both GBR1 and GBR2 have seven transmembrane domains, with extracellular amino termini and intracellular carboxy termini (Fig. 6). The amino terminus of GBR1 provides a GABA binding site, whereas the intracellular loops and carboxy terminus participate in G protein binding. GBR2 does not appear to interact with GABA. GBR1 and GBR2 interact via the carboxy domain, and this interaction is necessary in order to direct GBR1 expression to the cell surface.

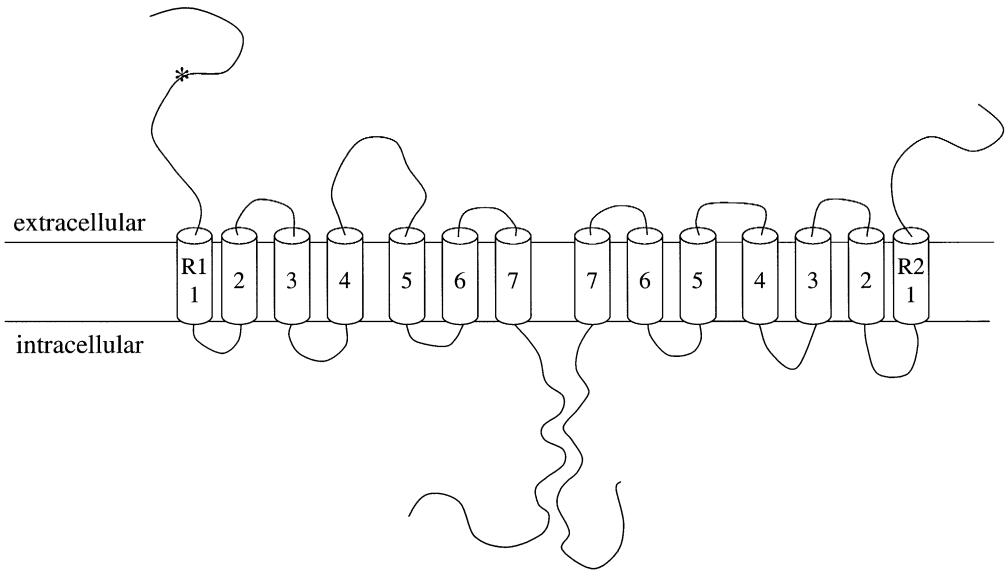


Figure 6 $GABA_B$ receptor subunit and complex topology. The $GABA_B$ receptor is composed of two similar subunits, each with seven transmembrane domains, extracellular amino termini, and intracellular carboxy termini. The putative GABA binding site, denoted by an asterisk, is on the amino terminus of $GABA_B$R1. The subunits interact via their carboxy domains.

2. Subunit Composition

There are at least four isoforms of GBR1, GBR1a–GBR1d, which range in size from approximately 100 to 130 kDa. The GBR1a isoform may be more prevalent at presynaptic sites, whereas the GBR1b isoform is predominantly found postsynaptically. GBR1 and GBR2 share 35% homology with each other but have little homology with other metabotropic receptors.

IV. GABA TRANSPORT

A. Plasma Membrane GABA Transporters

GABA is actively transported across the plasma membrane by GABA transporters (GATs). The four GATs identified so far (GAT1–3 and GAT4 or BGT-1) belong to the superfamily of Na^+- and Cl^--dependent transporters that include transporters for norepinephrine, serotonin, dopamine, and glycine but not for glutamate. The four GATs differ in pharmacology and in cell expression. For instance, GAT1 is inhibited by nipecotic acid and is predominantly found on neurons, although it is also expressed on some glia. GAT2 and GAT3 can transport both GABA and β-alanine. GAT3 is primarily a glial transporter, whereas GAT2 is found on neurons but its cell specificity remains unclear. BGT-1 was first identified in the kidney and transports GABA, betaine, and taurine. GAT2 and GAT3 are also found in peripheral tissues, including the liver and kidney.

Na^+- and Cl^--dependent transporters rely on the ionic gradients across the membrane. GABA transport is an electrogenic process that cotransports approximately two Na^+ ions and one Cl^- ion with each GABA molecule. Since the ionic gradients normally favor inward transport, GABA usually is transported into the cell. Under conditions of high intracellular Na^+ or membrane depolarization, however, GABA can be released through the transporter.

1. Topology

The GATs are glycoproteins with a molecular weight of 70–80 kDa. The presumed topology predicts that both the amino and the carboxy termini are cytoplasmic and that there are 12 transmembrane domains. The GAT1 topology is illustrated in Fig. 7. There are three glycosylation sites between the third and fourth transmembrane domains. The fourth through sixth extracellular loops form a putative GABA-binding site.

2. Carrier-Mediated Release

The GABA transporters reverse to release cytosolic GABA when the Na^+ gradient across the membrane is perturbed and also when the membrane potential increases. Although these changes occur rapidly during an action potential, GAT-mediated GABA release (or carrier-mediated release) is probably not a

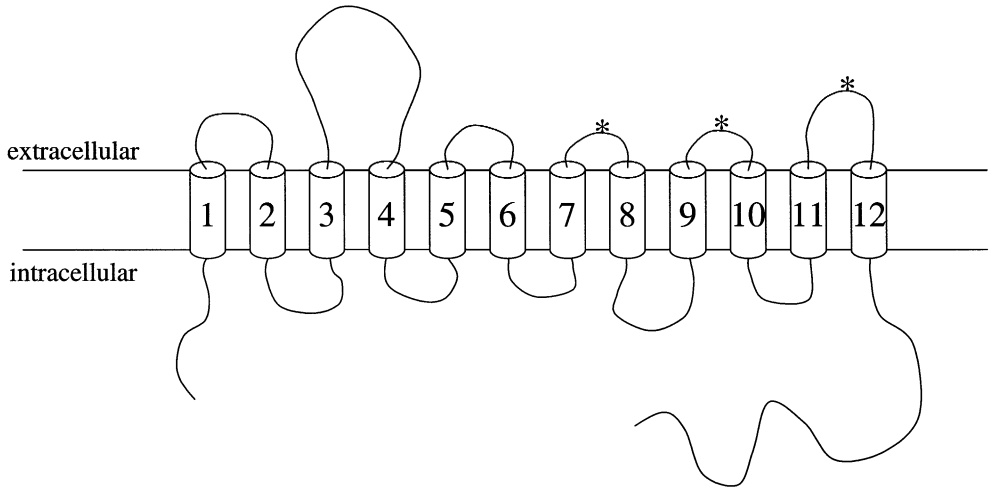


Figure 7 GAT has 12 transmembrane domains, and both the amino and carboxy termini are intracellular. The putative GABA binding site denoted by asterisks, is on the fourth through sixth extracellular domains.

major route of egress during cell firing because the changes in membrane potential are rapid and transient. However, several studies suggest that activation of the glutamate NMDA receptor causes intracellular Na^+ levels to increase sufficiently to produce GAT-mediated GABA release. Small perturbations in the extracellular K^+ concentrations, and concomitant membrane depolarization, can also cause carrier-mediated GABA release at levels sufficient to activate nearby $GABA_A$ receptors. Seizure activity and the accompanying changes in the extracellular ionic environment may therefore stimulate carrier-mediated GABA release.

B. Vesicular GABA Transporters

GABA is packaged into vesicles at the synapse by the vesicular GABA transporter (vGAT), which was recently cloned. Although vGAT transports GABA across a membrane, it bears no resemblance to the GATs. Both the amino and carboxy termini of vGAT are cytoplasmic, and there are 10 predicted transmembrane domains (Fig. 8).

vGAT, with a K_m for GABA in the millimolar range, depends on both pH and electrochemical gradients (ΔpH and $\Delta\Psi$) across the vesicular membrane to concentrate GABA into vesicles. These gradients are maintained by Mg^{2+}-activated ATPase, which belongs to the class of vacuolar ATPases. Vesicular transporters for other neurotransmitters have a different dependence on the same gradients. For instance, the vesicular dopamine transporter depends mostly on ΔpH, whereas the vesicular glutamate transporter depends mostly on $\Delta\Psi$.

Glycine, an inhibitory neurotransmitter, can also be transported by vGAT. vGAT mRNA distribution in neurons coincides with other glycinergic and GABAergic markers. Immunohistochemical studies demonstrate vGAT in terminals and the neurotransmitter phenotype of the particular terminals can be GABA, glycine, or mixed.

V. GABA FUNCTION

The GADs differ in their PLP dependence, encoding gene, membrane association, and subcellular location, which suggests that the GADs may synthesize separate pools of GABA that serve different functions. Although GABA is a neurotransmitter, it also functions as a developmental molecule and can contribute to the tricarboxylic acid cycle to provide energy to the cell. Studies of GABA in transgenic mice deficient in one GAD support the hypothesis that GAD_{65}-synthesized GABA is predominantly packaged for neurotransmission, whereas GAD_{67}-synthesized GABA remains in the cytosol and is available for alternate functions.

A. Effects of GAD Deletions in Knockout Mice

Mice lacking GAD_{67} die shortly after birth, presumably as a result of cleft palate. Their brains contain only 7% of the GABA concentration of control brains and about 20% of the GAD activity. Morphological

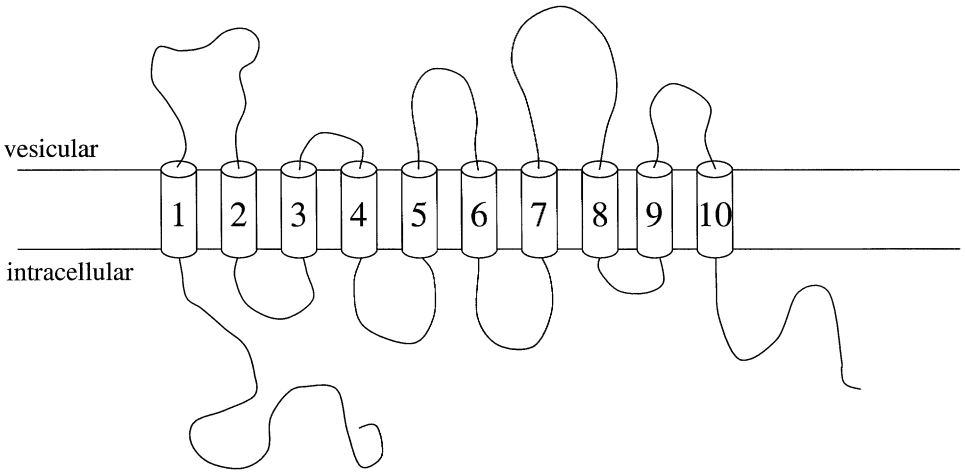


Figure 8 VGAT has 10 transmembrane domains, and both the amino and carboxy termini are intracellular (cytosolic).

analyses of their brains reveal no major abnormalities, but no detailed anatomical or electrophysiological studies of embryonic structures from GAD_{67} knockouts have been reported.

Mice lacking GAD_{65} appear to develop normally but are abnormally sensitive to seizures, particularly in a mouse line that is genetically susceptible to insulin-dependent diabetes mellitus. Perhaps significantly, these mice exhibit both humoral and cellular immune responses to both GAD_{65} and GAD_{67}.

Mice lacking GAD_{65} have a lowered capacity for depolarization-dependent GABA release *in vivo*. The GAD_{65} knockout mice also show impaired experience-dependent plasticity, as determined in monocular deprivation experiments. GAD_{65} may be selectively important in providing GABA to fill secretory vesicles and support exocytotic release of GABA. For example, GAD_{65} knockout mice cannot release normal amounts of GABA during and immediately after sustained stimulation of the retina or hippocampus. Moreover, in contrast to wild-type mice, these GAD_{65} knockout mice show no increase in the probability of GABA release after tetanic stimulation of the hippocampus. Although GAD_{67}-synthesized GABA can evidently be packaged into release vesicles under conditions of low demand, GAD_{67} alone seems to be unable to support the high-efficiency reloading of GABA vesicles required for normal function during conditions of high demand.

Mice lacking both forms of GAD die at birth, presumably from cleft palate. Although these mice have no detectable GABA, there do not appear to be any gross histological abnormalities in their brains. Therefore, GABA's role as a developmental molecule may be redundant, at least in some respects.

B. GABA as a Paracrine "Reset" Signal

GABA-producing cells increase GAD_{67} and GAD_{67} mRNA levels in response to injury, whereas GAD_{65} and its mRNA are usually unchanged. Injuries that produce increased GAD_{67} include chemical lesions of the substantia nigra and hippocampus, neuroleptic drugs, spinal cord transection, and acute stress. The increase in GAD_{67} can be observed as soon as 1 hr after insult, implicating immediate early genes in GAD_{67} regulation.

Most GAD_{67} is not associated with synapses, and GAD_{67} may be responsible for carrier-mediated GABA release. This direct release of cytoplasmic GABA occurs in cultured hippocampal neurons as well as in cells transfected with the GABA transporter and in a pancreatic cell line. Evidence from knockout mice suggests that GAD_{67}-synthesized GABA may not be effectively packaged into vesicles. Therefore, it either remains in the cytosol (where it is available for the GABA shunt) or crosses the plasma membrane via the GABA transporter.

Extrasynaptic GABA (either released via GAT or leaked from the synapse) could interact with extrasynaptic GABA receptors. Such diffuse GABA would cause inhibition diffuse with kinetics different from those at the synapse. Such inhibition would depend on the kinetics of GABA release, diffusion, reuptake, and degradation as well as on receptor composition.

Diffuse GABA could act in a paracrine manner on extrasynaptic $GABA_A$ receptors. Such receptors have been documented electrophysiologically in studies of GABA "spillover" in cerebellar synapses and microscopically by EM immunocytochemistry. The effect of such paracrine action would be to suppress neuronal firing in cells within damaged circuits. Sustained inhibition could allow the plastic remodeling of neural circuits, such as occurs during the retraining of the spinal cord. According to this view, paracrine GABA, synthesized primarily by GAD_{67}, could serve a "reset" function, giving the cells of damaged but plastic circuits the time to recover from challenge or injury. Observed parallel increases in GAD_{67} and extracellular GABA in the hippocampus of kainate-treated rats are consistent with this view. Therefore, GAD_{67} may contribute to a relatively slow mode of paracrine function, whereas GAD_{65} may be mainly responsible for the synthesis of GABA that participates in rapid point-to-point signaling.

C. How Much GABA Is Needed to Evoke a Response?

The cloning of $GABA_A$ and $GABA_B$ receptors has enabled researchers to study the properties of receptors of known covalent structure. In most of these *in vitro* studies, the K_D of the cloned receptors for GABA is in the micromolar range, but *in vivo* GABA often has an EC_{50} in the nanomolar or even subnanomolar range.

Within a synaptic cleft, the concentration of GABA is estimated to be in the millimolar range, enough to saturate postsynaptic $GABA_A$ (and $GABA_B$)

receptors. The occurrence of $GABA_A$ and $GABA_B$ receptors outside of synapses suggests that GABA may also act outside of synapses. The ability of cultured cells and partially purified receptors to bind GABA at nanomolar concentrations is consistent with a role for nonsynaptic GABA, particularly in response to injury, seizures, and other stresses.

D. Neurosteroids and GABA

GABA is involved in the actions of neurosteroids, which are steroid hormones that are synthesized in the brain and are particularly relevant to reproductive behavior and stress responses. Neurosteroids can enhance GABA release. For example, some GABAergic neurons are stimulated by testosterone, via androgen receptors, and may mediate the action or secretion of other hormones. GABA can affect lordosis, the steroid-modulated behavior exhibited by sexually receptive female rats. GABA facilitates or inhibits lordosis depending on the specific region of the brain that is stimulated.

Neurosteroids can directly bind to and activate $GABA_A$ receptors to produce sedative, anticonvulsant, and anxiolytic effects in animals. Neurosteroids can also influence $GABA_A$ subunit expression. Anesthetic effects from neurosteroids are modulated by a different site on the receptor than other types of anesthetics, but the binding site is probably also within the pore. The α, γ, and δ subunits affect neurosteroid response but not necessarily GABA binding.

E. GABA Affects Overall Excitability

GABA is synthesized in both projection neurons (e.g., neurons within the striatum that project to the substantia nigra and globus pallidus) and interneurons (e.g., the basket cells of the dentate gyrus). Projection neurons are thought to play a major role in initiating postsynaptic activity, whereas interneurons are thought to modulate postsynaptic activity. Interneurons within the hippocampus appear to also form large circuits that underlie the theta rhythms in the brain. These slow (4–9 Hz) rhythms are associated with learning and memory and may reflect an overall control of excitation that provides a more amenable environment for synaptic plasticity. Thus, GABA as a neurotransmitter not only provides straightforward neuron-to-neuron signaling but also affects the overall excitability of large regions of the brain.

GABA can initiate two types of responses. Phasic responses occur when GABA binds to postsynaptic receptors. Phasic responses are characterized by fast postsynaptic changes. Tonic responses result from ongoing activation of GABA receptors outside of the synapse. This activation produces continuous inhibition of the postsynaptic cell, thereby controlling its overall excitability.

Tonic and phasic inhibition differ not only because of the receptor location but also in the receptor subunit composition. The extrasynaptic receptors that mediate tonic inhibition have a higher affinity for GABA than do the receptors responsible for phasic inhibition. Different affinities reflect the GABA levels to which the two sets of receptors are exposed. Extrasynaptic GABA levels are lower than those found within the synapse; therefore, extrasynaptic receptors must be sensitive enough to detect low levels, whereas synaptic receptors must not desensitize upon exposure to the high levels of GABA found within the synapse. The role of $GABA_B$ receptors, which are primarily extrasynaptic, has not been elucidated in these two types of inhibition.

VI. GABA AND DEVELOPMENT

GABA functions as more than a neurotransmitter. A growing body of evidence suggests that it also plays an important role in development. GABA appears in most CNS neurons at the beginning of neurogenesis (before synapse formation), as does the $GABA_A$ receptor. During this time, GABA is excitatory because a developmentally regulated chloride transporter maintains high intracellular chloride levels.

The pattern of GAD expression changes during development, again supporting a role for GABA at this time. In the spinal cord, the change from embryonic to mature GAD_{67} transcripts echoes the ventral to dorsal maturation of the cord. GAD immunoreactivity is diffuse early in development, suggesting that there is no vesicle-associated GAD present (which would be characterized by punctate staining) and therefore no synaptic GABA.

GAT-mediated GABA transport decreases from birth through adulthood, whereas vGAT-mediated transport increases from the time of synapse formation through adulthood. The differences in GABA trafficking may account for the developmental properties of

GABA because it does not necessarily act synaptically throughout development. GABA can stimulate both directed and random movement of young neurons, possibly through the $GABA_B$ receptor. However, GABA may not be required during development since GAD knockouts have undetectable GABA but develop anatomically normal brains.

Outside of the CNS, compromised GABA signaling during development produces cleft palate, a craniofacial abnormality in which the roof of the mouth fails to close completely. Animal studies suggest that the barbiturate diazepam, which interacts with $GABA_A$ receptors, can cause palatal defects. In addition, mice lacking GAD_{67} are born with cleft palate and die shortly after birth. The route by which GABA may be associated with cleft palate is not well understood.

VII. GABA AND DISEASE

Defects in GABA synthesis, release, and response have serious consequences. Since GABA is a ubiquitous neurotransmitter, it can be argued that it has a role (direct or indirect) in most neurological diseases. The direct loss of GABAergic neurons in the striatum, for instance, is a hallmark of Huntington's disease. Parkinson's disease, however, results from a loss of dopamine neurons in the substantia nigra. The lost neurons normally project to GABA neurons in the striatum, and those lost projections affect GABA signaling in Parkinson's disease.

A. Changes in GABA Signaling

Epilepsy describes a set of diseases characterized by hyperexcitability and synchronous firing of large groups of neurons. Although there are many different kinds of epilepsy, temporal lobe epilepsy (TLE) is the most common type in adults and is often difficult to resolve. TLE seizures usually begin in the hippocampus and spread to involve large parts of the brain. Genetic mutations, injury, or tumors can all result in TLE. TLE usually causes extensive changes within the hippocampus, including the death of subsets of neurons, neuritic sprouting, and sclerotic lesions. The hyperexcitability characteristic of TLE suggests that this disease compromises GABAergic inhibition.

GAD levels increase in parts of the TLE hippocampus, but GABA levels decrease. Since GAD increases as a result of many types of injury, the changes observed in TLE may not serve to provide more GABA for signaling. Instead, the increase may reflect a general cellular response to injury, and the resultant GABA may be diverted to the GABA shunt to provide energy for the cell.

Hyperexcitability could reflect the loss of GABA neurons, but the neurons that die in TLE are mainly excitatory cells. Only a small subset of GABAergic neurons, which coexpress somatostatin, is lost. Several hypotheses address the compromised inhibitory capability of the TLE hippocampus. One possibility is that the excitatory inputs to some GABAergic neurons are lost, thus decreasing their inhibitory signaling. Another suggestion is that some GABAergic neurons lose their targets and redirect their outputs to many cells. The result is that one inhibitory cell now drives many excitatory cells, causing the synchronized firing that is a hallmark of TLE. A third hypothesis suggests that GABAergic neurons redirect their output to other GABAergic neurons, thus providing a net excitatory output relative to the normal state.

The multitude of changes within the TLE hippocampus may mean that one comprehensive treatment approach is impossible. Although GABA neurons are mainly intact, changes in connections may either increase or decrease net inhibition. GAD levels are increased, GABA receptors are increased in some cells but decreased in others, the GABA transporter is decreased only in some areas of the hippocampus, and overall GABA is decreased. Nonetheless, many of the drugs that are effective for treating TLE increase GABA action through potentiating $GABA_A$ receptors (thus creating a more powerful response to endogenous GABA) or by interfering with GAT1 or GABA-T to increase available GABA.

B. Defects in GABA Receptors

Angelman syndrome is linked to deletions in human chromosome 15, specifically in the area that codes for the cluster of $GABA_A$ receptor subunits $\alpha_5/\beta_3/\gamma_3$. Severe mental retardation, epilepsy, movement disorders, inappropriate laughter, and craniofacial abnormalities characterize this disorder. A similar syndrome, Prader–Willi syndrome, shows mild mental retardation, hypotonia, hyperphasia, and hypogonadism. The two syndromes are linked to the same gene; however, Angelman's is linked to the maternal gene, whereas Prader–Willi is linked to the paternal gene.

Epilepsy and craniofacial abnormalities are characteristic of other GABA-related disorders, highlighting GABA's role in both development and signaling, although the connection between gene deletions and symptoms is unclear.

C. Loss of GABA Neurons

Huntington's disease is an autosomal-dominant disease in which GABAergic projection neurons in the striatum are destroyed. Early symptoms include mood changes progressing to uncontrolled movements (chorea) and eventually to loss of movement and death. Disease progression can be correlated with the loss of three separate populations of GABAergic striatal neurons, although these are not the sole casualties. Cortical neuronal loss occurs later in the disease as well.

The gene for Huntington's disease codes for the huntingtin protein, which contains a long polyglutamine stretch. Normally, the number of glutamines in this protein is less than 40. Huntingtin mutations that include more than 40 glutamines are pathogenic. The number of glutamines is inversely related to the age of onset, and people with longer repeat lengths develop Huntington's disease at earlier ages. The function of the normal huntingtin protein is unknown, as is the reason for the susceptibility of GABAergic projection neurons.

The Huntington's disease gene is expressed throughout life, but disease symptoms normally begin in the fourth decade. This delay in symptom onset may reflect the ability of the brain to compensate for dysfunction for a period of time. Currently, there is no cure for Huntington's disease, and treatments are limited to palliative care.

D. Defects in GAD

Autoimmunity to GAD is a hallmark of two related conditions—insulin-dependent diabetes mellitus (IDDM, juvenile diabetes, or type 1 diabetes) and a rare neurological disorder called stiff-man syndrome (SMS).

1. Diabetes

IDDM is an autoimmune disease in which T cells mistakenly destroy pancreatic β cells, which are the sole producers of insulin. Several β cell proteins are targets of autoantibodies, including GAD_{65}, GAD_{67}, and insulin. These autoantibodies often arise several years before disease onset, possibly providing a window of opportunity for intervention.

The β cells of pancreatic islets contain high levels of GABA and GAD, comparable to those in neurons. GAD_{65} and GAD_{67} are both present in the pancreas of rats and mice, with GAD_{67} and its mRNA at higher levels than GAD_{65} and its mRNA. In humans, however, GAD_{65} greatly predominates and may also be present in pancreatic α cells. Species differences may reflect differences in gene regulation during the development of islet cell precursors. GABA from β cells is thought to act on $GABA_A$ receptors to inhibit glucagon release by α cells.

Autoantibodies to GAD, sometimes in combination with other autoantibodies, provide a highly specific and highly sensitive prediagnostic test for individuals at high risk for developing IDDM, sometimes many years before actual disease onset. GAD autoimmunity also provides a diagnostic marker that distinguishes between adult-onset non-insulin-dependent diabetes mellitus, which does not involve autoimmunity, and late-onset IDDM, which does. Whether it appears in children or adults, IDDM is an autoimmune disease, with GAD_{65} as the most common early target. Autoimmunity to GAD_{65} in IDDM could merely reflect the unmasking of a self-antigen after β cell destruction, but it may well have a causal role in the destruction of β cells.

2. Stiff-Man Syndrome

SMS is a rare neurological disease characterized by muscle rigidity that results from the simultaneous contraction of antagonistic muscle groups. As in the case of spasticity, GABA agonists can ameliorate symptoms.

About 60% of SMS patients have autoantibodies to GAD, but the causal role of GAD remains an open question since there is no apparent difference in the signs and symptoms of SMS patients with and without anti-GAD antibodies. Most SMS patients with GAD autoantibodies also have other autoimmune disorders (such as Hashimoto's thyroiditis and Grave's disease), whereas only a few percent of the GAD antibody-negative patients have any evidence of autoimmune disease. Despite years of attempts, the administration of GAD cannot provoke SMS in experimental ani-

mals, suggesting that GAD autoimmunity in SMS is probably not causal.

VIII. SUMMARY

In the 50 years since its identification as a neurotransmitter, GABA has emerged as a major force in normal CNS function, development, and disease. In addition to neurotransmission, GABA can provide energy through the GABA shunt and can mediate developmental events. Outside of the nervous system, GABA is involved with pancreatic glucagon release and is required for palate formation.

The variety of functions that GABA can perform may be governed in part by the different location and regulation of the two GABA-synthesizing enzymes. GAD_{65} appears to synthesize GABA largely for neurotransmission, whereas GAD_{67} synthesizes GABA that is available for the GABA shunt and also for release directly through GAT. The manner by which GABA leaves the cell may also be important in determining its function. GABA released via exocytosis is largely confined to the synapse, whereas that released through GAT is available to extrasynaptic receptors.

GABA receptors are an additional modulator of GABA function. Different subunit compositions impart different sensitivity, kinetics, and location. The $GABA_A$ receptor can confer inhibition or excitation, depending on the chloride gradient across the membrane. The $GABA_B$ receptor can exert both presynaptic and postsynaptic effects. The diversity of GABA actions is due in no small part to the great variation within the GABA receptor family.

GABA is the major inhibitory neurotransmitter and changes in GABA signaling have widespread effects. A role for GABA dysfunction can be argued even in diseases for which the primary pathology lies elsewhere. The molecules that synthesize, transport, respond to, and degrade GABA are all potential targets for therapy. Drugs that can selectively activate specific receptor subtypes or that can interfere with the kinetics of GABA transport and degradation will certainly be helpful in treating GABA-related diseases. Alternate approaches, including gene therapy, to control local GABA concentrations are also promising. Our understanding of the mechanisms of GABA regulation and signaling will therefore lead to new and effective therapies for a variety of diseases.

See Also the Following Articles

ALCOHOL DAMAGE TO THE BRAIN • DOPAMINE • ENDORPHINS AND THEIR RECEPTORS • EPILEPSY • NEURON • NEUROTRANSMITTERS • NOREPINEPHRINE • PEPTIDES, HORMONES, AND THE BRAIN AND SPINAL CORD

Suggested Reading

Barnard, E. A., Skolnick, P., Olsen, R. W., *et al.* (1998). International Union of Pharmacology. XV. Subtypes of gamma-aminobutyric acid A receptors: Classification on the basis of subunit structure and receptor function. *Pharmacol. Rev.* **50**(2), 291–313.

Blein, S., Hawrot, E., and Barlow, P. (2000). The metabotropic GABA receptor: Molecular insights and their functional consequences. *Cell Mol. Life Sci.* **57**(4), 635–650.

Martin, D. L., and Olsen, R. W. (Eds.) (2000). *GABA in the Nervous System: The View at Fifty Years*. Lippincott Williams & Wilkins, Philadelphia.

Jursky, F., Tamura, S., Tamura, A., *et al.* (1994). Structure, function and brain localization of neurotransmitter transporters. *J. Exp. Biol.* **196,** 283–295.

Kaufman, D. L., Clare-Salzler, M., Tian, J., *et al.* (1993). Spontaneous loss of T-cell tolerance to glutamic acid decarboxylase in murine insulin-dependent diabetes. *Nature* **366**(6450), 69–72.

Kaupmann, K., Huggel, K., Heid, J., *et al.* (1997). Expression cloning of GABA (B) receptors uncovers similarity to metabotropic glutamate receptors. *Nature* **386**(6622), 239–246.

Kaupmann, K., Malitschek, B., Schuler, V., *et al.* (1998). GABA (B)-receptor subtypes assemble into functional heteromeric complexes. *Nature* **396**(6712), 683–687.

Liu, Q. R., Lopez-Corcuera, B., Mandiyan, S., *et al.* (1993). Molecular characterization of four pharmacologically distinct gamma-aminobutyric acid transporters in mouse brain. *J. Biol. Chem.* **268**(3), 2106–2112.

Martin, D. L. (1986). Chapter 10. In *Neuromethods Series I: Neurotransmitter Enzymes* (A. A. Boulton, G. B. Baker, and P. H. Yu, Eds.). Humana Press, Clifton, NJ.

McIntire, S. L., Reimer, R. J., Schuske, K., *et al.* (1997). Identification and characterization of the vesicular GABA transporter. *Nature* **389**(6653), 870–876.

Mohler, H., Fritschy, J. M., Luscher, B., *et al.* (1996). The $GABA_A$ receptors. From subunits to diverse functions. *Ion Channels* **4,** 89–113.

Mohler, H., Benke, D., and Fritschy, J. M. (2001). GABA (B)-receptor isoforms molecular architecture and distribution. *Life Sci.* **68**(19–20), 2297–2300.

Olsen, R. W., DeLorey, T. M., Gordey, M., and Kang, M. H. (1999). GABA receptor function and epilepsy. *Adv. Neurol.* **79,** 499–510.

Smith, G. B., and Olsen, R. W. (1995). Functional domains of $GABA_A$ receptors. *TIPS* **16**(5), 162–168.

Soghomonian, J.-J., and Martin, D. L. (1998). Two isoforms of glutamate decarboxylase: Why? *TIPS* **19,** 505.

Glial Cell Types

CONRAD A. MESSAM, JEAN HOU, NAZILA JANABI, MARIA CHIARA MONACO, MANETH GRAVELL, and EUGENE O. MAJOR

National Institute of Neurological Disorders and Stroke

GLOSSARY

astrocyte The most abundant neuroglial cell type in the brain; named for its characteristic star-like shape due to processes extending radially from the cell body.

ependymoglial cells Glial cells that extend processes to the ventricular lumen in the brain and establish contact with the apical surface of neural tissue.

glial fibrillary acidic protein Type III intermediate filament protein used as a marker to identify mature cerebral astrocytes.

glial limitans Continuous layer at the surface of the cortex and cerebellum formed by the endfeet of radial glial cells.

gliosis Also called reactive astrocytosis; astrocytic response to injury or insult, marked histologically by the accumulation of glial fibers composed of glial fibrillary acidic protein.

microglia Resident macrophage of the brain; primary immune effector cell of the brain.

multipotential stem cells Cells derived from the neuroepithelium of the developing central nervous system that can give rise to different cell types.

myelin Specialized tissue high in lipid content that insulates the axons of neurons; produced in the brain by oligodendrocytes.

neuroglia The cells of the brain, excluding neurons; basic subtypes include astrocytes, oligodendrocytes, microglial, and ependymoglial cells.

oligodendrocyte Neuroglial cell type responsible for the production of myelin.

There are four major types of glial cells in the human brain—astrocytes, oligodendrocytes, microglia, and ependymal cells. This article will review the morphology and normal physiology of the various glial subtypes as well as their involvement in human brain disorders. Although glial cells comprise greater than 50% of the total population in the brain, historically they have been thought of solely as support cells for neurons. However, it has become apparent that the various glial cells perform critical functions during the development and normal functioning of the brain. Additionally, glial cells are central participants in almost all CNS disorders, taking part in both the protection or damage of brain tissue.

I. INTRODUCTION

More than a century and a half ago, Virchow (1846) introduced the word "neuroglia" to describe the tissue that fills the space between the nerve elements in the brain. Although simplistic when compared to our current concepts of neuroglia, the initial definition was correct. Almost four decades passed before the pioneering staining techniques developed by Golgi (1885) paved the way for others in the early 1900s to describe the morphology of the basic cell types that make up the neuroglia: astrocytes, oligodendrocytes, microglia, and ependymal cells.

With regard to establishing functions for neuroglia, His (1889) suggested that embryonic glial cells were responsible for guiding the migration of developing neurons to their final destination within the brain. In 1907, Lugaro proposed that adult astrocytes police the

interstitial milieu and maintain it in a state compatible for neuronal function. He also postulated that astrocytes were the cells responsible for the chemical degradation and uptake of substances released by neurons, which allowed communication and excitation, thus establishing the basis for synaptic transmission of nerve impulses. The concept that glial cells and their processes insulate nerve fibers to enhance neural impulse transmission was proposed in 1909 by Ramón y Cajal. He further postulated that glial cells fill the spatial void created by pathologic neuronal death, setting the precedent for gliosis.

In the 1920s, del Rio Hortega, by use of an innovative and selective silver impregnation technique, was the first to give a detailed morphological description of oligodendrocytes. He also identified oligodendrocytes as the cells that produce the myelin sheath that enwraps the axons of central nervous system neurons. In 1932, by use of the same silver impregnation technique, Rio Hortega also gave the first morphological description of microglia. Unlike astrocytes and oligodendrocytes, microglia have markers normally associated with hematopoietic monocytes. The origin, lineage, and mode of differentiation of microglia are incompletely understood.

Beginning in the 1950s, electron microscopy yielded information about the ultrastructural characterization of the various organelles in glial cells. From these studies came the finding that fibrous and protoplasmic astrocytes contain gliofilaments, which were subsequently determined to be intermediate filaments containing glial fibrillary acidic protein (GFAP) and vimentin. Although not all astrocytes in the normal brain immunochemically stain positive for GFAP, especially those in the gray matter, GFAP immunoreactivity is a major characteristic used to identify astrocytes.

Later in the 20th century, the focus of neuroglial study shifted toward the understanding of various interactive processes that occur between different glial cell types and neurons. These interactions are required to establish, maintain, and regulate normal brain functions and to determine how they may contribute to pathology. In synaptic and nonsynaptic transmission of nerve impulses, neurons and glia were shown to communicate reciprocally. In nonsynaptic regions of the brain, neuron to glia and glia to glia signals were shown to be mediated by neurotransmitters. Prominent among other molecules that are involved in interactive signaling between glia and other cell types of the brain are growth factors, neurotrophic factors, hematopoietic factors, cytokines, and chemokines. In the following sections, the morphology and functions of the subtypes of cells, which make up the "neuroglia," are described along with cellular activities associated with pathology or specific diseases.

II. ASTROCYTES

Astrocytes are generally divided into three subtypes, classified by morphology and anatomic location in the brain: radial, fibrous, and protoplasmic. All astrocytes express varying levels of GFAP, an intermediate filament that bundles together to form gliofilaments. The gliofilaments, 8–12 nm in diameter, are found mainly in the processes and less in the perinuclear region. Astrocytes can be visualized by Golgi impregnation; intracellular dye injection, which allows viewing of the entire cell body and processes; or immunocytochemical staining. Astrocytes are interconnected by gap junctions, which allow intercellular passage of ions and small molecules. The resulting cytoplasmic continuity suggests that astrocytes form a functional syncytium. Gap junctions are also observed between astrocytes and oligodendrocytes but not between astrocytes and neurons. Other characteristic features include the dense granules of glycogen in the cytoplasm and intramembrane "assemblies," observed by freeze fracture electron microscopy. These assemblies consist of 7-nm subunits bundled to form an array in the cell membrane, the function of which is not known. The endfeet of astrocyte processes terminate on the subpial glial limitans, blood vessels, nodes of Ranvier, and axons. Astrocytes are the most abundant cell type in the brain, comprising approximately 55% of the total population, and function in the maintenance of interstitial homeostasis and modulation of synaptic function. Astrocytes respond to almost any central nervous system (CNS) insult by activation and gliosis, which can limit edema, isolate damaged areas, and initiate an immune response.

A. Morphology and Subtypes

1. Radial Astrocytes

Radial astrocytes extend a single or group of long, thick, longitudinal processes that extend from the ventricle toward the surface of the brain. Radial astrocyte processes span the entire white matter and abut on the pia mater, the outermost surface of the brain. Because of their location in the ventricle and

bipolar morphology, some of these cells are sometimes considered ependymal cells. There are three major subtypes of radial astrocytes. First, *Radial glial cells* are primarily observed during development. They extend processes from the ventricle throughout the neural tissue. The main function of these cells is to act as "cables" on which CNS progenitor cells can migrate to their final destination. Radial glial cells are transient since they are scarce in the adult brain. It is thought that some of these cells can later become other astrocyte subtypes. Second, *Bergman glial cells*, also called Golgi epithelial cells, are the radial astrocytes of the cerebellum. The cell bodies are located between the Purkinje neurons and send processes radially through the molecular layer of the cerebellar cortex. Endfeet of Bergman glia processes terminate at the surface of the cerebellum and form a continuous layer known as the glial limitans. Figure 1 demonstrates the immunohistochemical staining for GFAP on human fetal brain. Third, *tanycytes* are radial astrocytes found mainly in adult brain, the nuclei of which are located within or just below the ependymal, or innermost, lining of the ventricle. These cells extend processes radially through the white matter.

2. Fibrous Astrocytes

Fibrous astrocytes are located in the white matter of the brain, characterized by long, unbranched processes that radiate in all directions from the cell body but rarely reach the pia mater. Fibrous astrocytes can be found in the cortex and cerebellum, and they express high levels of GFAP in their processes. The endfeet of these processes terminate on blood vessels and synapses as well as on the cell bodies and processes of neurons.

3. Protoplasmic Astrocytes

Protoplasmic astrocytes are mainly found in the gray matter of the brain. These astrocytes possess numerous short, thin, ramified processes that radiate in all directions from the cell body. These processes express little GFAP and are better observed by Golgi

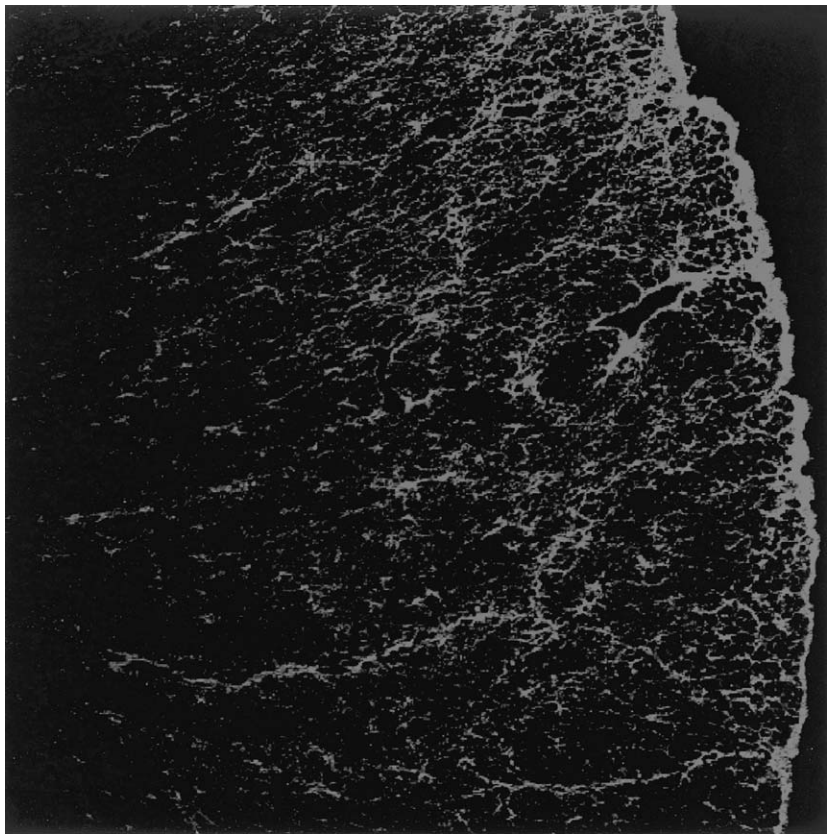


Figure 1 Immunohistochemical staining of adult human brain sections for GFAP. Note the radial morphology of the astrocyte tracts toward the outer surface of the brain. The area of intense GFAP staining at the periphery of the section is the glia limitans.

impregnation or dye injection. The endfeet terminate on blood vessels, synapses, and neuronal cell bodies, and the processes form the subpial glia limitans.

4. Astrocytes *in Vitro*

Astrocytes have varying morphologies in culture that may not fully reflect the range of morphologies found in the brain. *In vitro*, astrocytes can be separated from other neural cell types, thereby generating 95–99% astrocyte-enriched cultures. Figure 2 represents the immunofluorescent staining of cultured human astrocytes. In culture, most astrocytes derived from fetal brain appear fibroblast-like and immunostain with GFAP. These *in vitro* cultures maintain most of the functional properties of astrocytes *in vivo*. Cultured astrocytes are sometimes divided into two subtypes, type 1 and type 2, based on phenotypic characteristics as identified in rodent tissue cultures. Type 1 astrocytes stain with GFAP, whereas type 2 astrocytes costain for GFAP and A2B5, a sialoganglioside. The majority of astrocytes in culture are type 1. The type 2 astrocytes found in culture have not been observed in human brain tissue *in vivo* and may be an artifactual property of culturing.

B. Lineage and Development

Most of the information on development and lineage has been obtained from rodent brain developmental studies, with some confirmation using human fetal brain tissue. Astrocytes are derived from the neuroectodermal tissue of the neural crest. In the brain, the ventricular zone consists of a column of cells lining the ventricle that possesses the highly proliferative multipotential CNS stem cells. These stem cells give rise to progenitor cells of glial and neuronal lineage. Astrocytes are thought to originate from cells in the ventricular zone that migrate to the subventricular zone before traveling to the final destination in the brain. The process by which astrocytic progenitor cells migrate to specific brain regions is poorly understood. Radial glial cells, however, are recognized as the first glial subtype to appear in the brain. They express GFAP and nestin, an intermediate filament characteristic of immature cells. Radial and Bergman glia extend processes from the subventricular zone to the meninges at the outer surface of the brain and serve as the scaffold on which precursor cells migrate to their final location. The production and release of adhesion molecules and chemokines by astrocytes is also

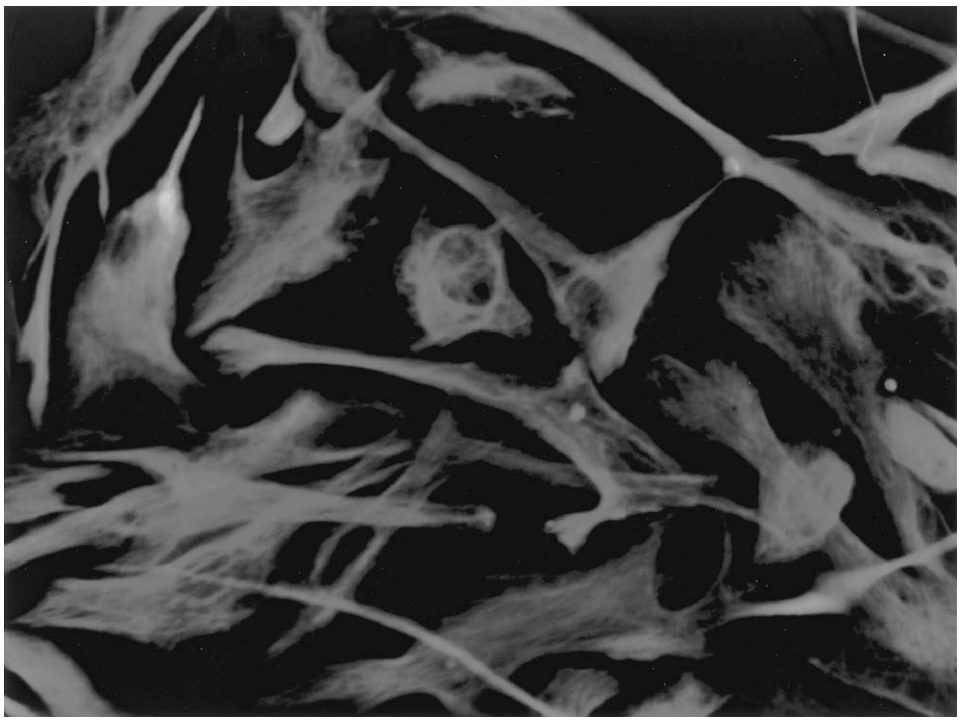


Figure 2 Immunofluorescent staining for GFAP in cultured human astrocytes. Note the filamentous ultrastructure of the cytoplasm. Magnification, × 200.

thought to contribute to the migratory process of progenitor cells during brain development. Additionally, astrocytes are known to produce a variety of trophic factors that assist in the maturation of other neural cells. Studies in rodent brain suggest that radial glia give rise to some astrocytes found in the gray and white matter. Little is known about the lineage progression of astrocyte progenitors and the factors responsible for directing the differentiation of the various astrocyte cell subtypes.

C. Normal Physiology and Function

1. Structural Functions

Astrocytes have several structural functions. They are necessary in conjunction with endothelial cells for the formation of the blood–brain barrier, an anatomic and metabolic barrier at the level of the capillary endothelial cells. The endfeet of astrocytes help to maintain a continuous layer between brain tissue and blood vessels, forming the perivascular glial membrane. The blood–brain barrier is the major obstacle preventing foreign compounds and toxins in the bloodstream from entering and damaging the brain. In effect, this barrier functionally and structurally sequesters the brain from the rest of the body. This protective barrier, however, also hinders the administration of drugs or other therapeutic compounds into the parenchyma of the brain. Another structural function of astrocytes is the formation of the glia limitans, a continuous lining of astrocyte processes, covered by a basal lamina, which is formed between the brain and the meninges. Again, the formation of the glia limitans by astrocytes helps to form the barrier to isolate the brain from other extraneural tissues. With their great abundance, astrocytes and their processes fill the extracellular space of the brain and insulate the various cell types from each other. This also occurs in brain injury, in which astrocytes hypertrophy and increase the size and number of processes at the site of injury to replace degenerated cells. The astrocytic response to CNS injury, termed gliosis, forms a scar that mechanically stabilizes the area of neuronal cell loss.

2. Maintaining Extracellular Homeostasis

Astrocytes contribute to the homeostasis of the extracellular fluid of the brain by helping to regulate pH and the extracellular concentration of K^+. There are several electrical properties characteristic of astrocytes that aid in maintaining K^+ homeostasis: electrical inexcitablity, higher membrane potential than neurons, precision in sensing and maintaining interstitial K^+ levels, and electrical linkage with neighboring astrocytes by gap junctions. The Na^+/K^+ pumps are highly concentrated on the endfeet of astrocyte processes in the vicinity of neurons. The pumps take up excess K^+ after neuronal depolarization, which can then be redistributed to more distal locations via gap junctions. The regulation of interstitial K^+ may limit excitation and be involved in fluid regulation. It is also hypothesized that astrocytes contribute to interstitial pH homeostasis. The Na^+/H^+ exchanger and the Cl^-/HCO_3^- exchangers regulate intracellular pH and may help to regulate extracellular pH as well. Astrocytes may also maintain homeostasis of the extracellular fluid by actively transporting material through the quasisyncytial network from capillary to brain tissue and vice versa.

3. Support and Interaction with Other Neural Cells

Astrocytes serve some essential functions for glutamatergic neurons and the glutamate synapse. The glutamate synapses involve the coordination between the presynaptic terminal of the axons, postsynaptic membrane, and surrounding astrocytes. The processes of astrocytes located around synapses possess high-affinity glutamate and GABA transporters that remove excess released neurotransmitters in order to limit neuronal excitation. Additionally, glutamate synthetase, found only in astrocytes, catalyzes the conversion of glutamate to glutamine, thereby providing neighboring neurons with the substrate for the production of glutamate. Therefore, glutamate synthetase also functions to detoxify neurotoxic levels of glutamate as well as ammonia. Glutamate dehydrogenase, another enzyme found in astrocytes, catalyzes the formation of glutamate from α-ketoglutarate and ammonia. Therefore, astrocytes are able to directly produce glutamate. Astrocytes may also modulate neuronal electrical activity through multiple interactions with neuronal cell bodies and nodes of Ranvier.

Astrocytes perform several functions that can modulate activity, survival, and development of other neural cells. During CNS development, radial glial cells form the highways from the subventricular zone to the parenchyma of the brain, on which the progenitor cells migrate to their final destination. The glycogen stores in radial glia are thought to provide energy for the migrating progenitor cells.

Additionally, astrocytes produce a variety of neurotrophic factors, cytokines, and adhesion molecules that can contribute to neuronal maturation and survival in the normal brain. Table I provides a partial secretory profile of astrocytes during normal function and gliosis. In response to injury, astrocytes can release protective neurotrophic factors to promote neuronal survival or, in other cases, release factors that may exacerbate damage and contribute to neuronal death. Also, the termination of astrocytic endfeet upon the blood vessels facilitates the transport of nutrients from systemic circulation through gap junctions, thereby redistributing the nutrients to neural tissue.

Although there are several studies examining the cellular and biochemical properties of astrocytes, relatively little is known about the molecular factors involved in astrocyte differentiation and normal function. Elucidating the molecular characteristics of astrocytes will be an area of active and intensive research for the future.

Table I
Growth Factors, Cytokines, Chemokines, and Adhesion Molecules Produced by Astrocytes

Factor	Full name	Normal or gliosis[a]
FGF-1 (aFGF)	Acidic fibroblast growth factor	N
FGF-2 (bFGF)	Basic fibroblast growth factor	N
PDGF	Platelet-derived growth factor	N
NGF	Nerve growth factor	N
CNTF	Ciliary neurotrophic factor	N
TGF-α	Transforming growth factor-alpha	N
IGF-1	Insulin-like growth factor	N
GMF	Glia maturation factor	N
TNF-α	Tumor necrosis factor-alpha	R
LIF	Leukemia inhibitory factor	R
TGF-β	Transforming growth factor-beta	N/R
IL-1B	Interleukin-1 beta	N/R
IL-3	Interleukin-3	N/R
IL-6	Interleukin-6	N/R
G-CSF	Granulocyte colony-stimulating factor	N/R
M-CSF	Macrophage colony-stimulating factor	N/R
GM-CSF	Granulocyte and macrophage colony-stimulating factor	N/R
IP-10	Gamma-interferon inducible protein	N/R
MCP-1	Macrophage chemoattractant protein	N/R
IL-8	Interleukin-8	N/R
ICAM-1	Intracellular cell adhesion molecule-1	N/R
VCAM-1	Vascular cell adhesion molecule-1	N/R

[a]Abbreviations used: N, released by normal astrocytes; R, released by reactive astrocytes; N/R, released by both normal and reactive astrocytes.

D. Contribution to Disease

1. Edema

Edema is one of the most common responses to brain injury resulting from astrocytic swelling. There are two types of edemas. The first is cerebral edema, which occurs after a disruption of the tight endothelial junctions of the brain vasculature. This causes fluid influx in the brain parenchyma, resulting in increased intracranial pressure. The second type of edema is cellular edema, which is astrocytic swelling after injury that may not result in a net influx of fluid into the interstitial space. Both types of edema may impair transporter function and the ability of astrocytes to maintain K^+ homeostasis. Acute abnormalities that cause edema include hemorrhage, trauma, and brain infarct. Subacute and chronic conditions such as status epilepticus, malignancy, Reye's syndrome, and encephalitis can also cause edema.

2. Gliosis

Gliosis, also called astrocytic gliosis or astrocytosis, is a common term that refers to the reactive astrocytic response to a brain injury or insult. Almost all brain lesions have a component of gliosis, even with different glial pathologies. Gliosis is a secondary event to CNS damage and may persist for weeks or months after brain injury. This condition occurs after infarct and is associated with infections and neoplasm as well as with demyelinating, toxic, and metabolic diseases. In gliosis, astrocytes hypertrophy, the nuclei become enlarged, and the chromatin becomes less dense while nucleoli become more prominent. There is an increased number of organelles and higher production of the intermediate filaments GFAP, nestin, and vimentin, which results in greater and more highly condensed glial processes and fibers. The increased glial processes replace injured CNS cells and form a gliotic scar. It is thought that the glial scar limits edema and prevents neuronal regeneration in the CNS by blocking regenerating axons from entering the damaged areas. With gliosis there is the release of cytokines, growth factors, and extracellular matrix proteins, which may be

involved in immune response, neuroprotection, or possible further damage. The term ''reactive gliosis'' normally refers to massive hypertrophy of astrocytes; however, it is apparent that gliosis is inherently reactive.

3. Neoplasm

Astrocytomas are neoplastic cells largely derived from astrocytes. They are classified into four categories based on the level of malignancy, with grade I being the least malignant and grade IV being the most malignant. All astrocytomas express some level of GFAP, but less malignant tumors generally express higher levels. The more malignant tumors express higher levels of nestin than GFAP. *Grade I astrocytomas*, called pilocytic astrocytomas, are gray, firm, and often cystic, with tumor cells that have long, hair-like cellular processes. These tumors most often occur in children or young adults. The boundaries are clearly distinguishable from normal brain, and when amenable to surgical removal the prognosis is good with little chance of recurrence. *Grade II astrocytomas* are the low-grade astrocytomas, which occur most often in young adults. They are solid, gray homogeneous tumors, characterized by astrocytes with pleomorphic nuclei and dark, condensed chromatin. These tumors diffusely infiltrate the brain parenchyma, which allows only partial surgical removal. There is variable prognosis with these tumors, with average survival of 3–5 years. *Grade III astrocytomas*, called anaplastic astrocytomas, resemble grade II astrocytomas but have more pleomorphic nuclei and high mitotic activity. Even with surgical removal and radiation therapy, the prognosis is still variable, with average survival of 1–3 years. *Grade IV astrocytomas* are known as glioblastoma multiforme and are the most malignant and frequent brain tumors in adults. These tumors have exaggerated features of grade III tumors with additional coagulation necrosis, vascular proliferation, and occasional hemorrhage. These tumor cells have a very high mitotic rate and the poorest prognosis for survival (less than 1 year after diagnosis).

4. Astrocytes

a. Immune Involvement Astrocytes are immunocompetent cells that participate in local immunological reactions. At the site of CNS damage, these cells can phagocytose dead cells and act as an antigen presenting cell in the initial phase of the immune response. Activated astrocytes express MHC II, which is involved in antigen presentation. Activation after CNS damage or pathogen infection results in the production of a variety of cytokines, interleukins, adhesion molecules, and chemokines. These released factors facilitate recruitment and activation of leukocytes and microglial cells to the area of insult and may contribute to disease development.

b. Involvement in Viral and Other Infectious Agents A variety of neurotropic viruses that cause neuropathogenesis in the human brain also infect astrocytes. Although in many cases infected astrocytes are not a primary target resulting in neuropathy, they contribute to viral latency and amplification of pathology through the release of cytokines. A good example is human immunodeficiency virus (HIV), the virus that causes acquired immunodeficiency syndrome (AIDS). HIV enters the brain and infects microglia and astrocytes, although astrocytic infection is nonproductive. It is hypothesized that astrocytes are a latent reservoir for HIV, complicating the task of complete virus eradication. Astroctyes, in conjunction with microglial cells, contribute to AIDS neuropathy by releasing cytotoxic factors that can amplify the immune response and further contribute to neuronal toxicity.

III. OLIGODENDROCYTES

Oligodendrocytes have a very clear functional definition. They are the cells in the CNS responsible for the production and maintenance of myelin, the insulating layer that surrounds the axons of neighboring neurons. As such, they are considered the most metabolically active cells in the brain. Due to an inherent resistance to early staining techniques, oligodendrocytes were the last cell type of the CNS to be discovered and characterized. Oligodendrocytes develop from the neuroepithelial cells of the neural tube in defined temporal and spatial patterns. In the gray matter of the adult human brain, oligodendrocytes are called satellite cells and function mainly in fluid and respiratory exchange, whereas oligodendrocytes found in the white matter are involved in the synthesis of myelin sheaths. In contrast to the functional diversity among astrocytes, oligodendrocytes appear to be a fairly homogenous population, with a unified purpose of myelinating axons. Structurally, however, oligodendrocytes exhibit a wide polymorphism, and classifying different variants has proven to be a difficult task.

A. Morphology and Subtypes

As a whole, oligodendrocytes are very refractive to stains. *In vivo*, the structure of the oligodendrocyte cell body is shrouded to some extent by the vast networks of myelin it produces as well as by its close association with neighboring axons. In fact, a cell may be identified as an oligodendrocyte based on its continuity with the outermost layer of the myelin sheath. The cells are morphologically diverse and can be classified into four subtypes. Type 1 oligodendrocytes have spherical or slightly polygonal cell bodies. They have several thin processes that emerge from the cell body in the direction of nerve fibers. This subtype can be found in the forebrain, cerebellum, and spinal cord, and these oligodendrocytes are usually arranged around blood vessels, neurons, and fiber tracts. Type 2 oligodendrocytes are polygonal or cuboid in shape, with fewer and thicker processes than those of the Type 1 subgroup. These cells are only found in the white matter and are closely associated with nerve fibers. Type 3 oligodendrocytes have even fewer processes directed toward nerve fibers and are found in the cerebral and cerebellar peduncles, medulla oblongata, and the spinal cord. Finally, type 4 oligodendrocytes are found near the entrance of nerve roots into the CNS and in association with large axons. These categories are arbitrary in nature because oligodendrocytes rarely fit perfectly into any one subgroup. Oligodendroglial cells may also be classified as interfascicular, perivascular, or perineuronal satellite cells, depending on their location.

Although the gross structure of oligodendrocytes varies widely, little difference in the fine structure has been observed in the CNS. The oligodendroglial cell body and the associated myelin membrane are enriched with sphyngoglycolipids such as galactosylceramide and its sulfated form, sulfatide. Immunocytochemical staining of the cell body *in vivo* may prove difficult because the vast network of myelin results in poor antibody penetration. Furthermore, the cell surface marker repertoire changes with different stages of differentiation. However, *in vitro*, isolated oligodendrocytes can be visualized by immunofluorescent staining, as shown in Fig. 3. Similarly, intracellular injection of dyes has revealed much about the ultrastructure of oligodendrocytes. The cytoplasm of the cells has well-developed Golgi apparati and abundant ribosomes, both free and bound to an extensive endoplasmic reticulum. This is to be expected since high-scale myelin protein synthesis and transport are the two main functions of this cell type. The greater density of the cytoplasm and nucleus, as well as the absence of glycogen granules or bundles of specific intermediate filaments, can distinguish oligodendrocytes from astrocytes. Oligodendrocytes do not contain intermediate filaments in the cytoplasm but, rather, actin microfilaments and a high content of microtubules, particularly in the processes. By electron microscopy, oligodendrocytes can be further subdivided into light, medium, and dark, distinguishable by decreasing size and increasing cytoplasmic density. It has been proposed that these structural differences are correlated to functional differences as well. For example, the light oligodendrocytes appear to be highly involved in the production of myelin, whereas the dark oligodendrocytes may be involved with myelin maintenance.

B. Lineage and Development

Most developmental studies of oligodendroglial lineage and development have been conducted in rodent models due to difficulty in culturing human oligodendrocytes. The maturation of oligodendrocyte precursors into mature oligodendrocytes is characterized by a distinct temporal expression of cell surface receptors and response to different growth factors. In mammals, oligodendrocytes originate from multipotential neural stem cells derived from the neuroepithelium of the developing CNS. In the adult brain, mature

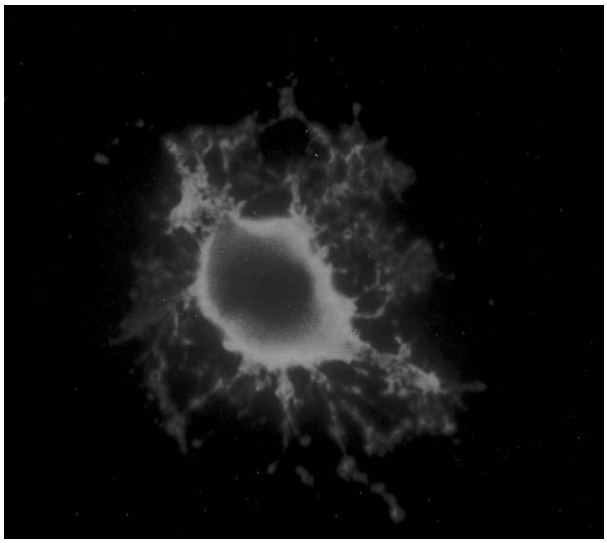


Figure 3 Immunofluorescent staining of a mammalian oligodendrocyte. Stained for Gal-C, a galactocerebroside expressed during the final stages of oligodendroglial development. Note the highly branched processes and intense staining of the cell membrane.

oligodendrocytes are distributed throughout the white and gray matter. In the developing CNS, however, induction of oligodendroglial precursors occurs in spatially restricted areas. Chemical factors released into localized areas of the extracellular environment can initiate changes in morphology and expression of growth factor receptor mRNA. In rodent studies, it is thought that the precursors are initially restricted to the ventral ventricular zone of the developing neural tube. The notochord and surrounding floor plate may release factors such as sonic hedgehog, which results in migration of the precursors dorsally and radially to populate the white matter. Furthermore, oligodendrocyte precursors located in the developing midbrain and forebrain may migrate into the thalamus and hypothalamus later in development, as well as to more dorsal regions of the cerebral cortex.

Induction of the oligodendrocyte progenitors from neuroepithelial stem cells in discrete locations must be followed by an active and large-scale migration throughout the CNS in order to myelinate all the white matter tracts. The cellular terrains or soluble factors utilized during this long-distance migration are unknown. The traveling precursors may interact with radial glial cells or may even utilize axons to guide them to their final destination. The oligodendrocyte precursors do possess a variety of integrin receptors, which may play important roles in migration and differentiation over various extracellular matrix components.

Extensive proliferation occurs in the white matter, and the expansion of the oligodendrocyte population is most likely regulated by a variety of growth factors. Once sufficient progenitor populations have been achieved, proliferation is downregulated and shifted toward differentiation. Again, maturation of the oligodendrocytes is highly influenced by growth factors and hormones. As the precursors mature, they will lose their motility and various cell surface receptors. Table II summarizes the proposed stages of oligodendroglial development and progressive changes in cell surface markers, morphology, and motility. Terminal differentiation of oligodendrocyte precursors and initiation of myelination requires dramatic and highly coordinated changes in the pattern of gene expression. The molecular mechanism by which oligodendrocyte-specific gene expression is regulated is most likely due to the combined action of multiple transcription factors. These include members of the homeodomain proteins in undifferentiated oligodendrocytes and zinc finger proteins in mature oligodendrocytes. Oligodendroglial survival factors secreted at the final destination induce the synthesis of myelin-specific mRNA, such as myelin basic protein (MBP) and proteolipid protein (PLP). MBP is actually a family of seven protein isoforms produced from a single gene by alternative splicing. Little is known about the precise structure of MBP in its native environment, although it is believed that the proteins are self-associating and may exist in an oligomeric form at the cell membrane.

Table II
Oligodendroglial Development[a]

Stage	Specific cell surface marker	Morphology	Motility
Preoligodendroglia, oligodendrocyte precursor (OP)	PDGFRα, O4 immunoreactivity		Highly motile
Immature oligodendrocyte	O1 immunoreactivity		Less motile
Mature oligodendroctye	CNP, PLP, MBP, MOG		None

[a]Abbreviations used: PDGFRα, platelet-derived growth factor receptor alpha; O4, antibody that recognizes cell surface constituents specific for oligodendrocyte precursors; O1, antibody against galactosylcerebroside; CNP, 2′,3′-cyclic nucleotide 3′-phosphotidase; PLP, proteolipid protein; MBP, myelin basic protein; MOG, myelin oligodendrocyte glycoprotein.

PLP is the most abundant protein in CNS myelin and also has alternative splice variants. With several transmembrane domains, functional studies have implicated a role of PLP as an ion channel.

C. Normal Physiology and Function

In the CNS, oligodendrocytes are responsible for the synthesis and maintenance of the myelin that surrounds the axons of neighboring neurons. The purpose of the myelin sheath is to allow saltatory propagation of nerve impulses along the length of the axon, resulting in a faster and more efficient neural impulse than in uninsulated nerve fibers. The exact cellular mechanisms responsible for the process of myelination are unclear. In humans, oligodendrocytes emerge several days or weeks before they actually start to synthesize myelin, and myelination takes place principally within the first year after birth. Recent studies have shown that initiation of myelination may be partially dependent on the activity of protein kinase C (PKC), a family of phospholipid-dependent enzymes ubiquitously present in the CNS. Not only do myelin-associated proteins appear to be excellent substrates for PKC-mediated phosphorylation but PKC activity also increases gradually after birth, coinciding with the temporal pattern of myelination in the human brain. Indeed, in cultures of human adult brain-derived oligodendrocytes, treatment with PKC agonists resulted in the increased synthesis of MBP as well as causing process extension, which are both myelogenic events.

Chemical or structural signals from neighboring axons are most likely the initiators of myelination by oligodendrocytes. An oligodendrocyte process extends from the soma and engulfs a segment of the axon, forming an inner and outer layer. As the cytoplasm is eliminated from the layers, the myelin condenses into compact myelin. The process will continue to encircle the axon, forming layers known as lamellae that abut one another and form a continuous spiral. Cross-linking of MBP oligomers on adjacent layers believed to condense the processes into thin, apposing sheets. Interestingly, the chemical composition of compact myelin is different from the chemical composition of the originating oligodendroglial cell membrane. Myelin has a much higher lipid content than the membrane of the cell body, despite the fact that compact myelin is continuous with the oligodendrocyte cell membrane. This is not surprising, however, since the insulating value of myelin is conferred in large part by its high percentage of lipid constituents. The myelin sheath as a whole is discontinuous because each oligodendrocyte process furnishes the myelin for only one segment of the axon, leaving small areas of exposed axonal membrane called the nodes of Ranvier. The action potentials skip over the myelinated areas and are repropagated only at the nodes, greatly increasing the velocity at which impulses travel. Each oligodendrocyte is capable of myelinating up to 50 internodes at a significant distance. In fact, during periods of active myelogenesis, an oligodendrocyte is capable of producing up to three times its own weight in myelin a day.

Other than the obvious role in the CNS, oligodendrocytes have also been shown to be neurite inhibitors. They prevent abnormal axonal sprouting outside the already established nerve tracts, and they also lend structural integrity to the CNS. However, oligodendrocytes that prohibit axon growth can actually enhance axonal regeneration in instances of neural tissue damage, further contributing to the complexity in function and intercellular relationships of these cells. Satellite oligodendrocytes located in the gray matter may function in regulating ion concentrations and pH levels in the extracellular space by fluid and ion exchange. Further studies must be conducted on mechanisms responsible for the oligodendroglial ability to successfully remyelinate demyelinated regions resulting from traumatic injury or disease. Once elucidated, these cells may be critical for therapeutic purposes in treating demyelinating diseases of the CNS.

D. Contribution to Disease

Damage to oligodendrocytes can occur in a variety of ways, including microbial infections, injury, autoimmunity, genetic defects, inflammation, and exposure to toxins. Although great strides have been made in understanding the core features of many of these demyelinating actions, the molecular events leading to damage of oligodendrocytes and, in many cases, dysmyelation or demyelination are not totally understood.

1. Multiple Sclerosis

Multiple sclerosis (MS) is a major demyelinating disease with pathological features similar to those of the experimental animal model, experimental allergic encephalomyelitis (EAE). Blood vessels in the CNS of MS patients characteristically have inflamed

perivascular cuffs containing T lymphocytes and monocytes recruited from peripheral circulation. Although much studied, the cause of oligodendrocyte death in MS is not clear. Some investigators have presented evidence that oligodendrocytes in acute and chronic demyelinated lesions undergo apoptotic death, whereas many others have found evidence of only necrotic cell death. Plaques of demyelination are present in the white matter, with chronic plaques devoid of both oligodendrocytes and myelin, as shown in Fig. 4. Reactive astrocytosis is also a prominent feature of MS lesions.

Many factors have been linked etiologically with MS. $CD4^+$ and $CD8^+$ lymphocytes have both been reported to lyse oligodendrocytes, with $CD4^+$ cells doing so by a non-MHC-restricted mechanism involving perforin release and $CD8^+$ by a MHC-restricted mechanism. Furthermore, $\gamma\delta$ T lymphocytes found in MS lesions may damage oligodendrocytes by a non-antigen-specific necrotic pathway. Activated macrophage/microglia also have the capacity to necrotically kill oligodendrocytes. Although oligodendrocyte injury and death can be mediated by both antibody-dependent and antibody-independent complement pathways, as well as by exposure to nitric oxide, none of these factors can clearly be shown to be the sole cause of MS.

2. Human T Cell Lymphotropic Virus, Type 1

Human T cell lymphotropic virus type 1 (HTLV-1) is a retrovirus that causes human adult T cell leukemia. Frequently, however, it causes the demyelinating neurological disease tropical spastic paraparesis (TSP), so named because it was thought only to occur in tropical geographic areas of the world. However, it has been found multiracially, in many countries, and in a wide range of climates. Because TSP occurrence has also been documented in individuals residing in countries with temperate climates, the name TSP is considered a misnomer and it has been proposed that it be called HTLV-1-associated myelopathy (HAM).

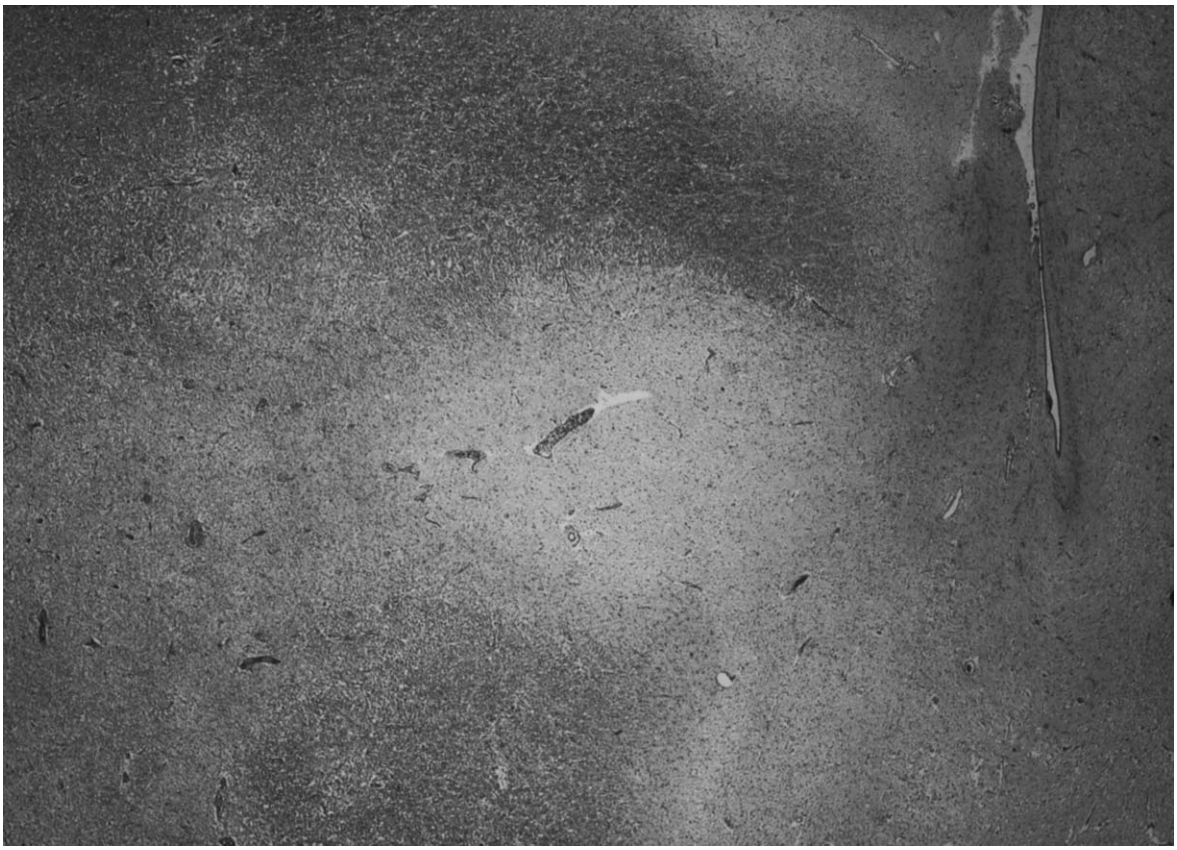


Figure 4 Demyelinated plaque occurring in the brain of a patient with multiple sclerosis. Tissue section staining with Luxol fast blue visualizes myelinated areas (dark) and regions devoid of myelin (light).

Symptoms associated with the chronic form of MS are similar to those of TSP/HAM, particularly in those patients exhibiting cerebellar signs. The occurrence of MS is rare in tropical regions, and the few patients from HLTV-l-endemic regions who were tested for HTLV-1 antibodies were negative. However, some individuals in temperate regions originally diagnosed as having MS in reality had TSP/HAM.

3. Progressive Multifocal Leukoencephalopathy

JC virus (JCV), a human polyomavirus closely related to BK virus and simian virus 40 (SV40), is the cause of the fatal demyelinating disease progressive multifocal leukoencephalopathy (PML), the only human disease known to be caused by infection with JCV. JCV infects 80% or more of the world's population, with initial infection occurring predominantly during childhood. The target of this ubiquitous virus is the myelin-producing oligodendrocyte. The virions can be detected in the nuclei of infected oligodendrocytes in a dense, crystalline arrangement called inclusion bodies. Prior to the AIDS pandemic, PML was a relatively rare disease affecting primarily immunosuppressed cancer patients with lymphoma or leukemia and transplant recipients receiving immunosuppressive therapy. It has been estimated that PML now occurs in about 5% of HIV-1-infected AIDS patients and is a major cause of death. Treatment of PML has been elusive. Some AIDS patients with PML receiving high-intensity antiretroviral therapy, including protease inhibitors, have shown clinical improvement, probably because the drug treatment regimen improved their general immunocompetence. Other AIDS patients with PML receiving high-intensity antiviral drug therapy showed no survival benefit from the treatment.

4. Other

Altered homeostasis of the neurotransmitter glutamate has also been postulated to cause excitotoxic death of oligodendrocytes. This depends on whether glutamate homeostasis is transiently or chronically altered. For example, exposure to short ischemic periods followed by transient increases in extracelluar glutamate may produce only limited damage to oligodendrocytes that can be repaired through the differentiation of oligodendrocyte precursors. Conversely, extended disturbances in glutamate signals may result in progressive oligodendrocyte cell death that exceeds the intrinsic capacity for oligodendrocyte repair, causing permanent damage to the oligodendrocyte population.

Toxic factors and genetic defects such as those seen in the leukodystrophies and Pelizaeus Merzbacher disease can also cause dysmyelination or demyelination. Exposure to free radical donors such as superoxide (O_2^-) or nitric oxide (NO), to free radical generating systems such as catecholamines, heat shock, and irradiation, and to agents that increase intracellular calcium concentrations, such as calcium ionophores, kainite, and myelin basic protein, has also been shown to kill oligodendrocytes.

A small percentage of all gliomas, approximately 5%, are oligodendroglial in lineage. These relatively avascular oligodendrogliomas initially form in the white matter and grow into the gray matter as they progress. The tumors can be grouped into two categories: the low-grade, less aggressive oligodendrogliomas or the more aggressive, anaplastic oligodendrogliomas. As a rule, a tumor in the low-grade category tends to be a pure oligodendroglioma, whereas the anaplastic tumors are a mixture of astrocytomas and oligodendrogliomas, also known as oligoastrocytomas. Although oligodendrogliomas are seen mainly in adults, they can also occur infrequently in children. Treatment strategies for oligodendrogliomas include surgical removal, radiation therapy, and chemotherapy.

IV. MICROGLIAL CELLS

Microglia are the resident macrophages of the brain, comprising 10–20% of all glial cells. They have active functions similar to those of other tissue macrophages, including phagocytosis, antigen presentation, and the production of cytokines, eicosanoids, complement components, excitatory amino acids (glutamate), proteases, and oxidative radicals. The antigenic plasticity of certain microglial populations, the cross-reactivity of the antibodies for microglia and other tissue macrophages, and the lack of a fully microglia-specific antibody argue for a monocytic derivation of these cells. However, a direct lineage relationship between microglia and myelomonocytic will be debated until quantitative and qualitative differences in the expression of some cell surface proteins common to both cell types and in their functions can be evidenced.

A. Morphology and Subtypes

"Brain macrophage" is a general term that comprises several subtypes of cells based on morphology,

localization, surface antigen markers, and function. Perivascular microglia, with an elongate shape, are located around blood vessels in adult tissue. They are thought to be regularly replenished by peripheral monocytes that infiltrate the CNS through the hemoencephalic or blood–brain barrier. The surface antigen profile is similar to that of circulating monocytes. *In vivo*, perivascular microglia are considered the most important antigen presenting cells, given their diversified, anatomical location. Intraparenchymal microglia, or resident microglia, are more numerous in gray matter than in white matter. They are maintained as a pool with a low turnover rate in normal adult brain. Parenchymal microglia can be subdivided into two populations. The first group includes ramified or resting microglial cells, which are highly branched cells with a small amount of perinuclear cytoplasm and a small, dense, and heterochromatic nucleus. The second population consists of ameboid microglial cells, which display migratory capacity and phagocytosis. These cells are particularly present during embryonic development and after CNS injury. In fact, ramified and retracted microglial processes can become ameboid-like, forming reactive microglia during brain injury. The best defined functions of microglial cells are related to microglial activation during pathological processes and include antigen presentation, cytotoxicity, neurovascularization, and phagocytosis. Some of these functions are also important for the normal physiology of the adult brain as well as the developing brain.

B. Lineage and Development

The origin of microglial cells has been debated since the initial silver staining by del Rio-Hortega. Although most macrophages are derived from monocytes, a subset can form locally by division of preexisting macrophages or from other progenitor cells. Here two contrasting theories are discussed in support of a mesenchymal or neuroectodermal origin of microglia.

During the development of the CNS, microglial cells are found after formation of blood vessels. During the vascularization of the CNS, the penetration of the endothelial cells is followed by focal degeneration of subadjacent glial endfeet, which attract monocytes from peripheral circulation. In fact, monocytes can infiltrate the CNS and have the potential to transform into macrophagic cells, giving rise to perivascular macrophages. However, there is not enough evidence that resident mature microglia in the parenchyma are derived from monocytes. They are believed to belong to the mononuclear phagocyte system because they express Fc and CR3 receptors and are capable of phagocytosis. Furthermore, cytoplasmic enzymes expressed by microglial cells, such as lysozyme, nonspecific esterase, and peroxidase, are also found in cells of the mononuclear phagocyte system. Microglia can be identified in human fetuses as early as 13 weeks of gestation. An important accumulation of ameboid cells is observed in the ventricular zone (germinal matrix) of human fetuses at 13–24 weeks of gestation. Human fetuses less than 28 weeks of gestation also have a significant concentration of macrophages in the ventricular zone, perivascular sites, leptomeninges, and subependymal regions. However, there is no correlation between the density of the ramified microglia and areas where there has been a high incidence of cell death. Indisputably, hematopoietic macrophages are present and play an important role in the developing CNS. Nevertheless, as the system develops, the number of dying cells decreases, as does the number of hematopoietic macrophages. The development of ramified microglia seems to be an independent process because they appear later in development, proliferate to occupy the parenchyma, and develop close functional interactions with neurons and other glial cells.

In attempts to demonstrate that microglia can be replaced by hematopoietic monocytes, the use of bone marrow radiated chimeric animals suggests that monocytes–macrophages in leptomeninges, choroid plexus, and perivascular areas are indeed replaced by cells from the bone marrow. In contrast, parenchymal ramified microglia are not replaced by bone marrow-derived marked cells in a significant number. Thus, the majority of microglial cells come from locally present precursor cells, most likely of neuroectodermal origin. The theory of a neuroectodermal origin of microglia assumes that microglia originate from a common glial stem cell, the glioblast, in the ventricular zone and later in the subventricular zone of the developing neural tube. Thus, some of the precursor cells responsible for the turnover of macroglia might be responsible for the turnover of microglia. Another possibility is that the CNS is colonized by hematopoietic stem cells very early during development, and that stem cells become resident cells of the CNS. These cells are then responsible for the regular turnover of the intraparenchymal microglia in normal brain and increased numbers of microglia in pathologic conditions. To date, searches for hematopoietic stem cells in the CNS have been unsuccessful. Attempts to develop cultures

of microglia initiated from the neuroepithelium of mouse embryos prior to vascularization and the appearance of monocytes–macrophages in the yolk sac strongly suggest that at least some microglial cells can originate from the neuroepithelium of the neural tube, as do other glia.

C. Normal Physiology and Function

The morphology and branching patterns of microglial cells display heterogeneity between different brain regions. Microglia in gray matter tend to be ramified, with processes extending in all directions. Cells in the white matter are bipolar and often align their cytoplasmic extensions in parallel, at right angles to nerve fiber bundles. Thus, the morphology of microglia adapts to the architecture of the brain region they populate, whereas their phenotype appears to be influenced by the chemical composition of the microenvironment. For example, MHC class II-positive and CD4-positive microglia are localized mostly in the white matter of normal brain. Regions lacking a blood–brain barrier show microglia and microglia-like cells with a different phenotype, suggesting that serum proteins influence the phenotype.

The microglial plasma membrane contains a large number of receptor and adhesion molecules as well as a wide variety of enzymatic activities (Table III). Therefore, a large number of antibodies can be used to stain microglial cells. Microglia in the human brain can be visualized using analogous antibodies against typical macrophage surface receptors. In addition to the expression of Fc and complement receptors, other cell adhesion molecules are expressed constitutively on resting microglia in normal brain. Belonging to the integrin superfamily of adhesion molecules, these include typical lymphocytic antigens, such as lymphocyte function antigen (LFA), CD4 antigen, and leukocyte common antigen. B lymphocyte antigens are also present on human microglia. Thus, the microglial surface membrane bears molecules associated with white blood cells. It is well documented that the normal brain contains MHC antigens and that the principal MHC-expressing cell type is the microglial cell. The constitutive expression of MHC antigens in the brain is not limited to microglial cells, however; it also includes endothelial cells and certain cell types located in the wall of cerebral blood vessels. MHC antigen expression is considerably increased in pathologic conditions or after systemic administration of cytokines, such as interferon-γ (IFN-γ). The

Table III
Phenotypic Characteristics and Secretory Activity of Microglial Cells[a]

	Antigen expression			Soluble mediator production
	Resting	Activated	Cell specificity	
Fc receptors	±	+	Macrophage	M-CSF
CD68	+	+	Macrophage	G-CSF
Lectin	+	+	Macrophage	TNF-α
CR3 complement receptor	+	+	Leukocytes	IL-1
MHC class I	±	+	Leukocytes	IL-6
MHC class II	−	+	Leukocytes	IL-10
CD4	±	+	T cells, monocytes	IL-12
LCA (CD45)	−	+	Leukocytes	TGF-β
ICAM-1 (CD54)	−	+	Leukocytes	Chemokines
VCAM-1 (CD106)	−	+	Leukocytes	NGF
LFA-1 (CD11a/CD18)	−	+	Leukocytes	NT-3
LFA3 (CD58)	+	+	Leukocytes	Prostaglandins
				Superoxide anions
				Matrix metalloproteinases

[a]Abbreviations used: ICAM, intercellular adhesion molecule; IL, interleukin; G-CSF and M-CSF, macrophage and granulocyte colony-stimulating factors; LCA, leukocyte common antigen; LFA, leukocyte function-associated antigen; MHC, major histocompatibility complex; NGF, nerve growth factor; NT-3, neurotrophin-3; TNF, tumor necrosis factor; TGF, transforming growth factor; VCAM, vascular cell adhesion molecule.

intermediate filament vimentin, which is evident in activated microglia and in brain macrophages, is absent from resting microglia. However, it can be upregulated rapidly in response to neuronal injury. Lectin histochemical staining of nervous tissue revealed that the B4 isolectin derived from *Griffonia simplicifolia* and the lectin from *Ricinus communis* resulted in the selective visualization of microglia. Both lectins have similar sugar-binding characteristics and recognize anomeric forms of galactose in the oligosaccharide side chains of nervous system glycoproteins in the microglial plasma membrane.

Perhaps the most well-known biological function of microglial cells is their role during the development of the CNS. Microglial cells remove dead cell fragments and eliminate transitory or aberrant axons. They can also play a more active role during cell degeneration by inducing the death of certain cells. Microglia support the development and normal function of neurons and glia by producing trophic factors as well as by participating in the growth and guidance of neurites within the developing CNS. They also promote the proliferation of astrocytes, increase myelinogenesis, and stimulate vascularization of the CNS. Many of these microglial effects are apparently mediated by active substances. Microglial cells in culture secrete a number of factors, such as nerve growth factor, neurotrophin-3, chemokines, macrophage and granulocyte colony-stimulating factors (M-CSF and G-CSF), interleukin-1 (IL-1), IL-6 and tumor necrosis factor α (TNF-α). In addition to acting on other populations of the nervous system, tissue culture studies have revealed microglial responsiveness to growth factors such as GM-CSF and CSF-1, which are potent inducers of microglial proliferation and function.

D. Contribution to Disease

Several arguments support a key role for microglia during immunopathologies. They are potentially phagocytic and have a pronounced microbicide and cytotoxic potential. Upon activation, microglia can rapidly upregulate the expression of several immunomolecules, such as MHC I and II, the CD4 antigen, and adhesion molecules. They are the most efficient and the most promptly inducible antigen presenting cells of the brain parenchyma, assuming an active immune surveillance in the CNS. Resting microglial cells stimulated with IFN-γ express MHC II products and display a capacity to present protein antigens in the molecular context of MHC II to CD4-positive T lymphocytes. In the context of MHC I expression, they can also become targets of cytotoxic CD8-positive T lymphocytes. Furthermore, microglia secrete as well as respond to several cytokines.

One of the characteristic features of microglia is the rapid activation in response to injury, inflammation, neurodegeneration, infection, and brain tumors. Microglial activation occurs after injury or changes in microenvironment, even before pathological changes or in the absence of obvious neuropathic changes, as part of an early CNS immune defense system. In terms of structural changes, many intermediate morphologies of activated microglia exist. Microglial hypertrophy begins with the formation of several stout processes, but they do not become phagocytic. If neuronal degeneration occurs in the brain parenchyma, activated microglia proliferate and transform into phagocytic cells. As the primary immune effector cells of the CNS, microglia respond to traumatic insult or the presence of pathogens by migrating to the site of injury, where they may proliferate. Activated microglia at the site of inflammation express increased levels of MHC antigens and become phagocytic. Like other tissue macrophages, microglia release inflammatory cytokines and mediators that amplify the inflammatory response by recruiting effector cells to the site of injury. In addition, microglia can release neurotoxins that may potentiate damage to CNS cells. The intense secretory activity of these cells is associated with diseases such as trauma, stroke, epilepsy, AIDS, and MS, in which the microglial response is prominent and deleterious to the brain tissue. The secretory activity of microglia has also been related to the neuronal destruction seen in Alzheimer's disease (AD).

1. Acquired Immunodeficiency Syndrome

HIV-1 causes an AIDS-associated psychomotor complex in a number of patients, who eventually develop either encephalitis or leukoencephalopathy. The main target cells of HIV-1 infection in the brain are microglia and macrophages, with a very limited infection of astrocytes and endothelial cells. Infected microglial cells can be detected by the presence of HIV antigens. The two pathological hallmarks of the HIV-1 infection of the CNS are multinucleated giant cells as a result of cell-to-cell fusion and microglial nodules (Fig. 5). Although microglial cells and macrophages seem to be the only cell types productively infected by HIV-1 within the brain, the replication rate of the virus remains relatively low in CNS tissue compared to other

tissues. Thus, the neuropathological alterations in AIDS are more likely due to the neurotoxicity of certain viral products, toxic factors, and cytokines released by infected microglia and macrophages. Some of these putative toxic factors, such as viral proteins, are specific to HIV-1 infection of the CNS, whereas other potentially toxic factors are also involved in other diseases in which activation of microglia plays a key role in the pathological process. Among viral proteins, the viral surface glycoprotein gp120 can be released by infected microglia and macrophages. This protein induces excitoxicity in neurons via activation of glutamate receptors, it can inhibit myelin formation in oligodendrocytes, or it can alter Na^+/H^+ ion transport in astrocytes, leading to an increased secretion of glutamate and potassium. The infection of microglial cells and macrophages, and their subsequent activation, also results in the generation of a wide variety of secretory factors that are potentially neurotoxic, such as TNF-α, cytokines, chemokines, arachidonic acid metabolites, and nitric oxide. TNF-α released by infected microglia is particularly toxic to oligodendrocytes and can be an important factor in myelin damage. Arachidonic acid and its metabolites act mainly via potentiation of glutamate receptors on neurons, leading to an increase in intracellular calcium levels and neuronal death. They can also impair the transport of glutamate in astrocytes.

2. Multiple Sclerosis

Pathological studies of MS lesions suggest an important role of macrophages and microglia in MS demyelination. Active or recent plaques have areas of myelin degradation, infiltration by inflammatory cells, and collections of lipid-containing macrophages that may stain for myelin proteins such as myelin basic protein. In chronic plaques, macrophage-like cells remain lipid-laden but no longer express immunoreactive myelin proteins. Some of these macrophages represent phagocytic microglia and some may be of hematogenous origin. MS lesions are frequently localized in perivascular regions and contain not only demyelinated axons and microglia but also lymphocytes and plasma cells. The presence of MHC II molecules in macrophages and microglia in MS lesions indicates that microglia are activated and can exert phagocytosis and antigen presentation functions. They also express Fc receptors and have an increased number of chemotactic receptors, including C5a and IL8.

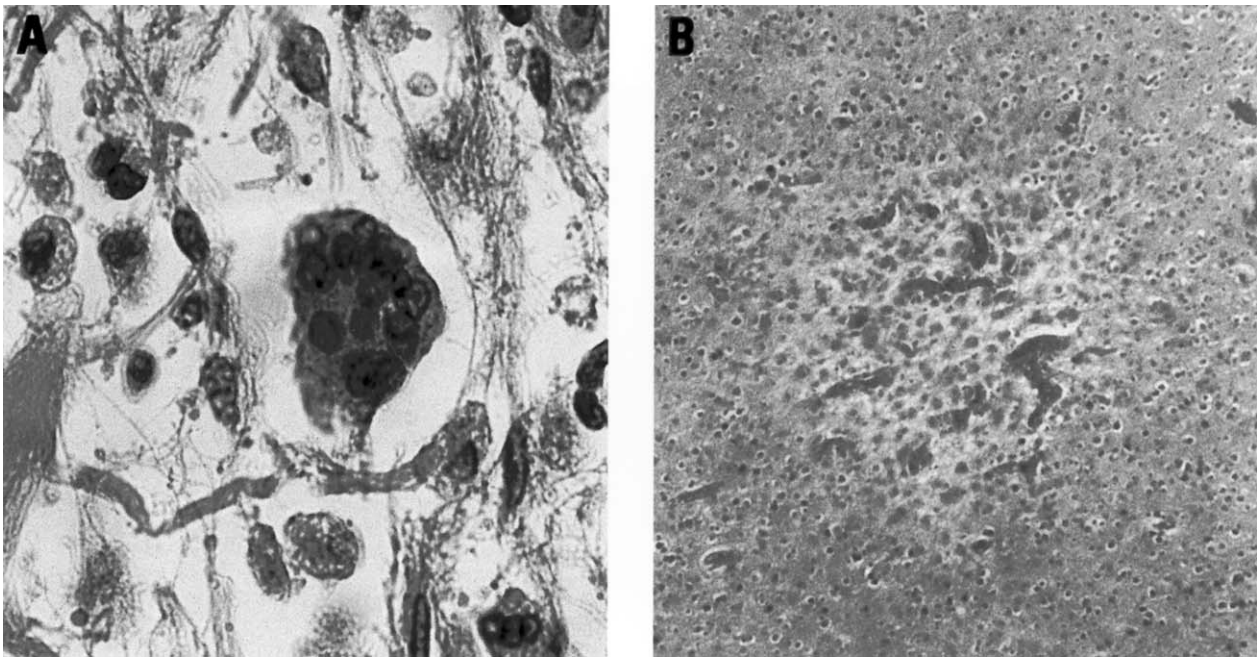


Figure 5 Neuropathological hallmarks of HIV infection. (A) Multinucleated giant cells are considered to be histological lesions directly attributed to HIV infection. (B) Microglial nodules, frequently observed in the brain of HIV-infected patients, are associated with opportunistic cerebral infections.

3. Alzheimer's Disease

Alzheimer's disease is a degenerative disorder characterized by memory loss eventually leading to dementia. Pathologically, AD is characterized by the presence of insoluble structures or depositions in cortical regions of the brain, namely β-amyloid (Aβ)-containing extracellular plaques and intraneuronal neurofibrillary tangles. The presence of large numbers of activated microglia and reactive astrocytes in brain tissue from AD patients has been interpreted as a secondary event. However, evidence suggests a significant role of these cells in the progression of the disease. Microglia cluster around Aβ-containing senile plaques with markedly enhanced MHC II protein expression on microglial cells associated with areas of degenerative pathology. Cytokines such as IL-1, IL-6, and TNF-α also have increased expression in microglia in the vicinity of senile plaques. These cytokines can potentially coordinate the majority of inflammatory changes found in AD brain tissue; however, a classic immune response as defined by the involvement of T cells or immunoglobulins does not appear to occur. The presence of microglia in the vicinity of amyloid plaques suggests a function in phagocytosis and plaque removal. Microglia can also uptake and degrade Aβ with a limited rate. In culture studies, synthetic Aβ not only recruits and activates microglia but also induces the secretion of cytotoxic products by these cells. Thus, rather than being protective, the activation of microglial cells may result in further neurotoxicity. Moreover, microglia can synthesize and secrete Aβ. Therefore, the stimulus of Aβ production may be Aβ.

V. EPENDYMOGLIA CELLS

The introduction of the term "ependyma" was meant to describe the cell layer lining the ventricles of the brain. However, a recent definition or classification of ependymoglial cells has been under debate. Some argue that certain members should be classified as astrocytes, whereas others argue that ependymoglial cells are distinct since they have many characteristics not shared by typical astrocytes. An ependymoglial cell is usually defined as a glial cell extending at least one of its processes to the ventricular space and having physical contact with the outer surface of neural tissue. Ependymoglia are made up of different members, including fetal radial glia, tanycytes, and Müller cells. Other cell types, such as Bergmann glia, have radial processes that extend to the ventricles of the brain. However, since they do not actually establish contact with the apical surface of neural tissue, they are considered to be astrocytes and not ependymoglia.

A. Morphology and Subtypes

Structurally, ependymoglial cells are characterized by different types of processes that are determined by contact with various microenvironmental compartments. The type I process is a feature of every ependymoglia cell. The endfoot comes into contact with a fluid or space into which extend many microvilli. The apical pole contains abundant mitochondria, which indicates a high level of metabolic activity. Another characteristic of apical processes in some but not all ependymoglia is the presence of kinocilia, a simple cilia consisting of a ring of nine pairs of tubules. Type 1 processes are interconnected by various types of apicolateral junctions. In regions where no endothelial blood–barrier exists, ependymoglia cells form a cerebrospinal fluid (CSF) barrier by the expression of tight junctions. Ependymoglial type 2 processes are characterized by basal mesenchymal contact and a cytoskeleton with abundant intermediate filaments. These filaments consist primarily of vimentin when the endfoot is in contact with CSF, but they consist primarily of GFAP when in contact with a blood vessel. The basal processes extend to the basal lamina of the mesenchymal layer underlying the nervous epithelium, forming processes of type IIa. Processes extending to the basal lamina of blood vessels are termed type IIb. Type III processes are defined by contact with neurons. These processes are characterized by the formation of flat or lamellar sheaths that enclose the neuronal somata (type IIIa), synapses (IIIb), or axonal internodes (IIIc1) or by finger-like extensions that contact the nodal specializations of axons (IIIc2). These processes may act as stores of sodium or calcium ions and cannot be elaborated until neurons have completed their differentiation.

B. Lineage and Development

In the human brain, ependymoglial cells are thought to be derived from the developing neuroepithelium along a caudal-to-rostral gradient. Radial glial cells are the first subtype of neuroglia to appear during human fetal brain development. They stain intensely with GFAP and nestin early in development but lose

immunoreactivity with time. It is thought that ependymoglia arise from a radial progenitor in a pathway that can also give rise to astrocytes. The radial progenitors differentiate into radial glial cells, which can then morphologically transform into the subtypes of ependymoglial cells mentioned earlier as well as into astrocytes.

C. Normal Physiology and Function

The function of ependymoglia cells is unclear. Physiological stimulation of brain compartments evokes specific reactions among ependymoglia. It has been reported that fetal ependymal cells play an important role during development of the nervous system by forming the cellular highway for the migration of progenitor cells. Furthermore, they arrest neurogenesis, facilitate motor neuron differentiation, and aid in transport of nutrients before development of capillary networks. Evidence of ependymal involvement in neuroendocrine function has also been reported as changes in ependymal cell morphology, coincident with changes in pituitary hormone secretion. In the adult brain, these cells provide only limited transport of ions, small molecules, and fluids between the brain and the CSF. The ependymal cells lining the ventricles act as a barrier for filtration of brain molecules and for the protection of the brain from potentially harmful substances in the CSF. From these examples, it is apparent that ependymoglia cells are continuously adapting to the changing needs of the neuronal tissue.

D. Contribution to Disease

Ependymomas and ependymoblastomas develop from the ependymal cells surrounding the ventricles and the central canal of the spinal cord. Intracranial ependymomas occur predominantly in children and tend to fill the ventricular lumen. The mean age at diagnosis of ependymoma and ependymoblastoma is 5 years. The incidence of these ependymal tumors in males and females is approximately equal. Clinical symptoms, although related to the location of the tumor, are usually due to blockage of CSF fluid flow, which may lead to a sudden increase in intracranial pressure. Ependymal tumors are sensitive to radiation.

Subependymomas are slow-growing, benign neoplasms originating from the subependymal glial matrix, consisting of a mixture of astrocytic, ependymal, and transitional cell clusters surrounded by their fibers. They generally project into the ventricular lumen. About one-fourth of the symptomatic intracranial tumors have mixed tumor cell populations and consist of a mixture of ependymomas and subependymomas. The prognosis in these cases is worse than that of a pure ependymoma. Another factor that affects the prognosis is the size of the neoplasm; symptomatic tumors tend to be large. Subependymoma is more common in men and has been reported in all decades of life, although no cases of subependymomas have been reported in children less than 2 years of age. The mean age for symptomatic tumors is 39 years, whereas that for asymptomatic lesions is 59 years. Microscopically, subependymomas consist of a nest of glial cells separated by glial fibers. Histologic studies have shown that some of the cells within these glomerate nests have attributes of ependymal cells and others of astrocytes; still others are transitional between the two.

VI. CONCLUSIONS

Neuroglial cells were described initially as specialized cells surrounding the neurons to provide structural support and insulation. In fact, neuroglial cells comprise a wide variety of phenotypes and functions. It is now known that complex intercellular communication exists not only between glial cells and neurons but also among the neuroglial members. This relationship becomes evident during pathological conditions of the brain. Selective damage of any one cell type can have severe ramifications on the functions of others as well as on the neuronal population. In the normal human brain, glial cells and neurons exist in a very delicate and highly coordinated balance. The diversity of neuroglial cells serves to contribute to the complexity of the human CNS.

See Also the Following Articles

ASTROCYTES • HOMEOSTATIC MECHANISMS • MICROGLIA • MULTIPLE SCLEROSIS

Suggested Reading

Adelman, G., and Smith, B. H. (Eds.) (1996). *Encyclopedia of Neuroscience,* Vols. 1 and 2. Elsevier, Amsterdam.

Beneviste, E. N. (1998). Cytokine actions in the central nervous system. *Cytokine Growth Factor Rev.* **9,** 259–275.

Bruni, J. E. (1998). Ependymal development, proliferation, and function: A review. *Microscopy Res. Techni.* **41,** 2–13.

Cuadros, M. A., and Navascues, J. (1998). The origin and differentiation of microglial cells during development. *Prog. Neurobiol.* **56,** 173–189.

Kettenmann, H., and Ransom, B. R. (Eds.) (1995). *Neuroglia.* Oxford Univ. Press, New York.

Kreutzberg, G. W. (1996). Microglia: A sensor for pathological events in the CNS. *Trends Neurosci.* **19,** 312–318.

Landis, D. M. D. (1994). The early reactions of non-neuronal cells to brain injury. *Annu. Rev. Neurosci.* **17,** 133–151.

Miller, R. H., Hayes, J. E., Dyer, K. L., and Sussman, C. R. (1999). Mechanisms of oligodendrocyte commitment in the vertebrate CNS. *Int. J. Dev. Neurosci.* **17**(8), 753–763.

Norenberg, M. D. (1994). Astrocyte responses to CNS injury. *J. Neuropathol. Exp. Neurol.* **53**(3), 213–220.

Stoll, G., and Jander, S. (1999). The role of microglia and macrophages in the pathophysiology of the CNS. *Prog. Neurobiol*. **58,** 233–247.

Wilkens, R. H., and Rengachary, S. S. (Eds.) (1996). *Neurosurgery,* Vol. 1. McGraw–Hill, New York.

Zhang, S. C., Ge, B., and Duncan, I. D. (2000). Tracing human oligodendroglial development in vitro. *J. Neurosci. Res.* **59,** 421–429.

Hallucinations

JANE EPSTEIN, EMILY STERN, and DAVID SILBERSWEIG
Weill Medical College of Cornell University

GLOSSARY

delirium A state of altered attention, arousal, and thought, often caused by an acute medical condition.

delusion A false belief, not generally endorsed by the individual's culture, held despite contradictory evidence.

epilepsy A chronic disorder consisting of intermittent episodes of excessive neuronal electrical discharge (seizures).

ictal Relating to a seizure.

partial seizure An episode of excessive electrical discharge originating from a discrete region of cerebral cortex and remaining confined to one part of the cortex.

psychosis A condition in which unreal beliefs or experiences are believed to represent reality.

Hallucinations are involuntary sensory experiences perceived as emanating from the external environment, in the absence of stimulation of relevant sensory receptors. They were first defined in this manner in 1837 by Esquirol, who differentiated them from illusions, which are perceptual misinterpretations of existing external stimuli. Hallucinations can occur in a variety of contexts but are perhaps most striking and debilitating in the setting of schizophrenia, in which they are combined with a failure to realize that they do not represent reality. In this instance, they are generally experienced as real, emotionally significant, and related to concurrent delusions, and they represent a form of psychosis. Hallucinations can occur in any sensory modality or can involve multiple modalities, with auditory hallucinations most common in schizophrenia and other illnesses traditionally termed psychiatric and visual hallucinations most common in illnesses termed neurologic. This article presents a functional neuroanatomic approach to hallucinations, describing and analyzing them in terms of disorders of sensory input, midbrain/thalamus, and higher brain regions, including cortical sensory, limbic, and frontal regions. It also discusses other investigational approaches to hallucinations as well as treatment considerations. The focus is on visual and auditory hallucinations because they occur most frequently and have been most thoroughly investigated.

I. FUNCTIONAL NEUROANATOMIC APPROACH TO HALLUCINATIONS

The variety of forms, contents, and settings of hallucinations can be described and analyzed in a number of ways, each with its own strengths and weaknesses. Of these, a functional neuroanatomic approach, based on evolving data, is perhaps most heuristically satisfying. In order to present such an approach, we first review the functional neuroanatomy of normal sensory perception as set forth by Mesulam.

A. Functional Neuroanatomy of Normal Perception

External auditory, visual, tactile, gustatory, and olfactory stimuli are first detected by modality-specific

receptors in the periphery. Information from most of these sensory receptors converges on modality-specific nuclei in the thalamus, where extraneous information is filtered out or "gated" and relevant information relayed to various parts of the cortex. At the cortex, the first regions to receive these inputs are the primary sensory areas: visual at the occipital pole and banks of the calcarine fissure, auditory at Heschl's gyrus on the posterior supratemporal plane, and somatosensory at the postcentral gyrus. These highly differentiated, or distinctly structured, regions carry out the most specialized processing within each modality, extracting basic features of sensory information. Output from these regions flows to adjacent unimodal association cortices, where modality-specific sensory elaboration occurs. Output from multiple unimodal association areas converges on heteromodal association cortices in temporoparietal and prefrontal regions, where unimodal percepts are linked with associated information to form multimodal constructs.

The path of information flow can be illustrated by considering the well-studied visual system. In this sensory modality, unimodal processing is mediated by regions stretching from the occiput anteriorly, with information flowing along two major pathways. The first, known as the ventral pathway, begins in association areas adjacent to primary visual cortex, where individual features of visual information, such as color and shape, are processed in separate subregions. As the information moves along the lower surface of the brain toward anterior temporal regions, these features are integrated and complex patterns are extracted, permitting discrimination of individual objects such as faces. Information from anterior temporal regions flows to heteromodal cortex, where highly processed visual percepts are integrated with input from other sensory modalities and with stored information about their characteristics, history, and emotional/motivational relevance, leading to object recognition. Integration is achieved via interconnections with limbic and paralimbic regions involved in mnemonic and emotional processing. The second, dorsal pathway leads from occipital unimodal association areas on the upper side of the brain to adjacent heteromodal regions in the posterior parietal cortex that mediate object localization. As information flows from more differentiated primary sensory areas to less differentiated heteromodal regions, specialized perception merges into complex cognition. Higher level cognitive processes are mediated by prefrontal heteromodal cortices involved in monitoring, categorizing, modifying, and integrating multiple streams of information processing to form an overview of the current situation and generate a relevant plan of action. This serial, hierarchical flow of information occurs within a context of bidirectional feedback projections and concurrent parallel processing.

Although much remains to be learned about the nature and precise localization of the neural phenomena that give rise to hallucinations, the existing data, along with our knowledge of the pathways involved in normal perception, allow for the categorization and analysis of hallucinations based on their probable neuroanatomic substrates.

B. Disorders of Sensory Input Associated with Hallucinations

Hallucinations produced by disorders of the peripheral sensory system appear to result from ongoing cortical sensory processing in the setting of degraded or absent sensory input. Sometimes referred to as "release" phenomena, the use of the term in this context can be misleading because it has traditionally been used to connote neural activity released by the failure of higher level inhibitory centers rather than by disordered lower level input. The probable mechanism by which these hallucinations are generated is best understood by considering the interplay between peripheral and central processing in normal perception. Although the previous description of sensory processing emphasizes the flow of information from periphery to thalamus to cortex ("bottom-up" processing), the connections between thalamus and cortex are in fact bidirectional, as noted. This pattern of connectivity enables the cortex to play a role in selecting from among the massive array of inputs to the thalamus those most likely to be relevant in light of past and current experience. Thus, perceptions arise from an interplay between cortically generated expectations ("top-down" processing) and data (confirmatory or otherwise) from peripheral sensory receptors. In this setting, a dearth of peripheral input might give rise to perceptions dominated by expectations rather than current environmental conditions.

Hallucinations caused by disordered peripheral input are most frequently seen in the visual system. The term Charles Bonnet syndrome has been applied to such phenomena, but without a consistent definition. The hallucinations are most often vivid, colorful, and complex representations of people, animals, scenery, trees, buildings, or flowers that fill the entire

visual field. They frequently appear smaller than normal, or Lilliputian, and may move. Hallucinations tend to have an abrupt onset, can last seconds to hours, and may disappear with movement or closure of the eyes. They can occur in the setting of acute blindness following altered blood flow or trauma to the eye, a phenomenon sometimes referred to as phantom vision, or during a gradual visual decline. In either case, they may fade as the visual disturbance continues. Notably, the individuals experiencing the hallucinations are aware that they do not represent reality and generally have no strong emotional association or reaction to them. Although primary abnormalities within the central nervous system may increase the risk of developing Charles Bonnet syndrome (a hypothesis supported by its greater prevalence in the elderly, who are at increased risk of subtle brain dysfunction), it can occur on the sole basis of peripheral visual system dysfunction. Indeed, its prevalence rate of 20% in people who develop blindness is similar to the 19% of normal individuals who experience visual hallucinations during sensory deprivation experiments—a clear example of hallucinations caused solely by disordered peripheral input.

Conditions such as stroke, that involve destruction of primary visual cortex or the cerebral pathways leading to it can also lead to complex visual hallucinations. Although the lesion in this instance is central, the phenomenology and mechanism are similar to those seen with peripheral lesions because primary visual cortex provides input to the unimodal association areas involved in the generation of complex hallucinations. The major phenomenologic difference derives from the organization of visual processing in primary cortex, where the right visual field is mediated by left occipital cortex and vice versa. When central lesions are limited to one hemisphere, hallucinations occur only in the affected contralateral visual field.

In the somatosensory system, a striking example of hallucinations caused by disordered sensory input occurs in the phantom limb syndrome, as described by Melzack. After amputation of a limb, approximately 95% of adults experience the limb as still present, able to move in space, and to feel pain or tingling. Over time, the perception weakens such that the proximal part of the limb no longer seems to exist, leaving the hand or foot hanging in midair, or the entire limb seems to "telescope" into the body, leaving the hand or foot directly connected to the stump. Phantom experiences have also been reported after the loss of eyes, teeth, external genitalia, and breasts. They are not dependent on sensory input from the residual scar and can occur even in those with congenitally absent limbs, suggesting a central representation of the body that is at least partly innate.

In the auditory system, individuals with peripheral dysfunction can develop complex hallucinations, such as music (either instrumental or vocal) or voices, or simple hallucinations, such as ringing, buzzing, or isolated tones of various pitches. Although more common with bilateral dysfunction, hallucinations can accompany unilateral peripheral auditory disease, in which case they are experienced as emanating from the affected side.

C. Midbrain/Thalamic Disorders Associated with Hallucinations

Hallucinations similar to those produced by peripheral lesions can occur with lesions of the upper midbrain and adjacent thalamus. First described in 1922 by Lhermitte, who attributed them to a lesion in the midbrain peduncular region, they remain known as "peduncular" hallucinations. Like Charles Bonnet hallucinations, they are often vivid, complex visual hallucinations (although other sensory modalities may be involved), frequently of people or animals (sometimes Lilliputian) often engaged in animated activities. Unlike those produced by peripheral lesions, peduncular hallucinations are generally associated with disturbances in sleep and arousal and may at times be interpreted as real.

These disturbances in sleep and arousal provide clues to the mechanism by which hallucinations are generated by midbrain and thalamic lesions. Frequency-specific oscillations in thalamocortical circuits have been associated with the temporal binding of perception and with dreaming. As pointed out by Manford and Andermann, brain stem connections to the thalamus are important in switching thalamic relay nuclei out of waking relay mode, in which they faithfully transmit sensory inputs to the cortex, into sleeping mode, in which they do not. This is accomplished via neurotransmitters, notably acetylcholine and serotonin, whose cell bodies lie within the basal and brain stem reticular regions involved in modulation of arousal, selective attention, and cortical processing. Abnormalities of acetylcholine and serotonin transmission brought on by disease, medication, or drug use (including such hallucinogens as LSD) are frequently accompanied by hallucinations. Similarly, transitions between states of sleep and wakefulness are

associated with hallucinations, usually in the setting of sleep disorders such as narcolepsy or insomnia. Hypnagogic hallucinations occur prior to falling asleep, whereas hypnopompic hallucinations occur upon awakening. Both are generally multimodal, vivid, and emotionally charged. Common examples are the feeling or experience of being about to fall into an abyss or attacked, of being caught in a fire, or of sensing a presence in the room. Hallucinations in the settings of delirium and sedative drug withdrawal are also associated with disturbances in sleep and arousal.

Abnormal thalamic activity was also present in a functional neuroimaging study, performed by our group with colleagues (as described below), of hallucinations in the setting of schizophrenia. This may relate to the modulatory and integrative roles of thalamocortical circuits in the generation of perceptual experience. A possible role for the thalamus in this setting is supported by a number of electrophysiologic, neuropathologic, structural, and functional imaging studies revealing sensory gating deficits and thalamic abnormalities in individuals with schizophrenia.

D. Disorders of Higher Brain Regions Associated with Hallucinations

Hallucinations are also associated with primary pathology at higher levels of the brain. Most prominent in this category are those that occur in migraine, epilepsy, and schizophrenia. Investigation of cerebral activity associated with hallucinations in these settings has been aided in recent years by the development of techniques such as electroencephalography (EEG), evoked potentials, single photon emission computed tomography (SPECT), positron emission tomography (PET), and functional magnetic resonance imaging (fMRI). Combined with data from other avenues of investigation, studies employing these tools have implicated a number of higher brain regions in the generation of hallucinations, corresponding to their form, content, and setting.

1. Cortical Sensory Activity Associated with Hallucinations

Regardless of the mechanism by which they are generated (primary peripheral, midbrain/thalamic, or higher brain disturbances), hallucinations appear to be associated with activity in cortical sensory regions corresponding to their modality and complexity. The hallucinations described previously may be described as complex or formed. Noncomplex hallucinations are referred to interchangeably as simple, unformed, or crude. In the visual system, these are known as photopsias, and they occur frequently with migraines. In this setting, the most common forms are colorless glittering spots and black-and-white zigzag patterns known as fortification lines. They often occur unilaterally but may fill the entire visual field. Photopsias can also occur, briefly, at the onset of partial seizures and for the first few days following an infarction of the central visual system. In both settings they tend to be brightly colored and unilateral. In addition, disorders of visual input may give rise to photopsias. Simple hallucinations are believed to reflect activity in primary sensory or adjacent early unimodal association areas and to correspond in form to the area's functional specialization. For example, colored photopsias would be associated with activity in occipital subregions involved in color processing, as described previously.

Complex hallucinations are associated with activity in sensory association areas, with or without involvement of primary sensory cortex. As with simple hallucinations, their form and content correspond to the location of activity. We investigated neural activity associated with auditory/visual hallucinations in a 23-year-old man with schizophrenia. The subject underwent PET scanning while experiencing frequent hallucinations of colored, moving scenes containing disembodied, rolling heads that spoke to him in a derogatory fashion. Using techniques that allow for identification of activity simultaneous with hallucinatory events, we detected activations in occipital and temporal visual association cortex (higher order visual perception), temporal auditory–linguistic association cortex (speech perception), and temporoparietal and prefrontal heteromodal association cortex (intermodal processing) (Fig. 1). Activations were bilateral but more extensive on the left, perhaps reflecting the dominance of that hemisphere for language.

2. Limbic/Paralimbic Activity Associated with Hallucinations

The previously mentioned study included five other subjects, all of whom experienced frequent auditory/verbal hallucinations in the setting of schizophrenia. Although each had a somewhat different pattern of sensory cortical activation, perhaps reflecting differences in the form and content of their hallucinations, group analysis revealed a highly significant pattern of common activations in thalamic, limbic, and

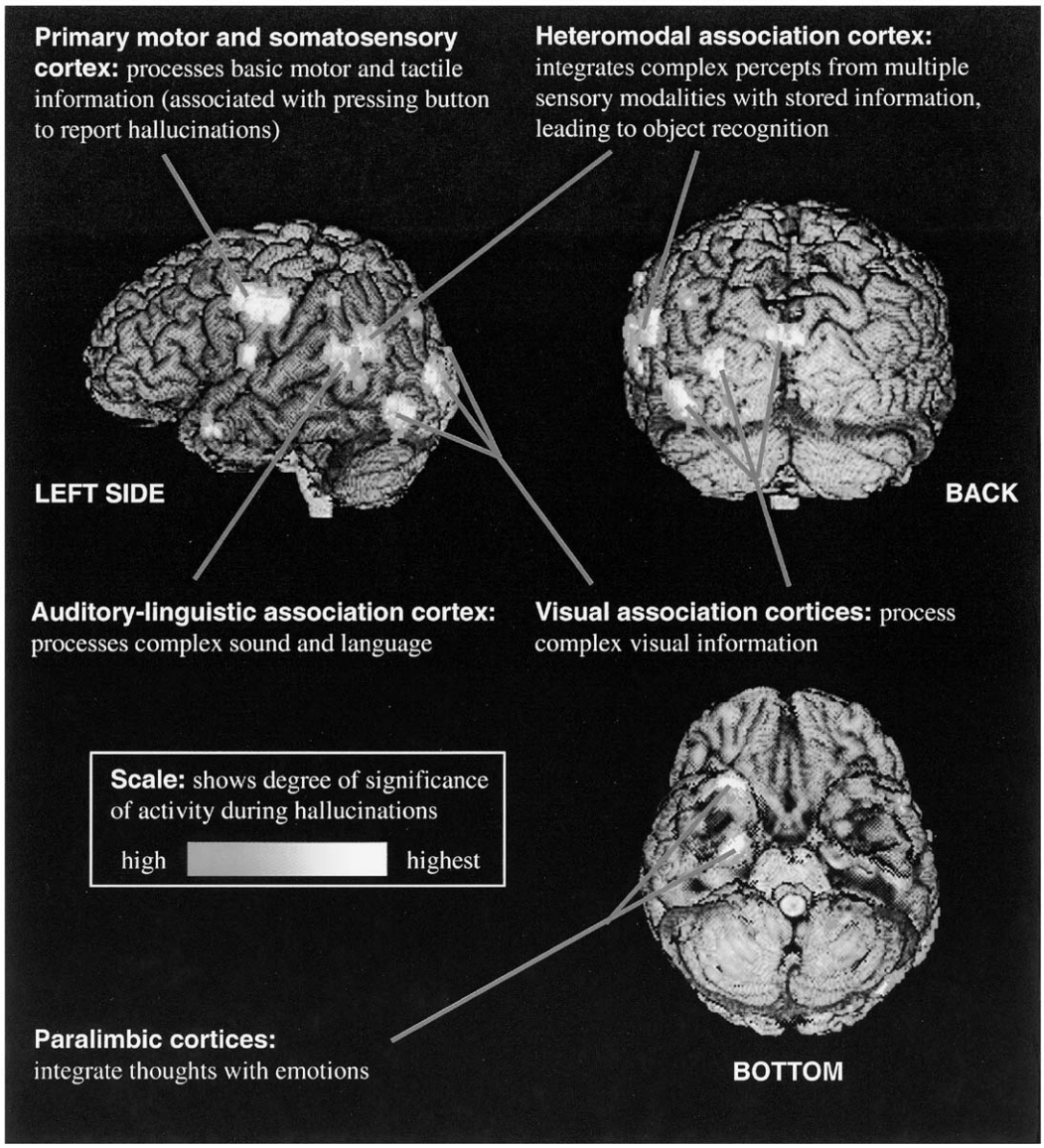


Figure 1 Brain regions active in a schizophrenic patient experiencing auditory–visual hallucinations of disembodied, rolling heads speaking to him. The functional PET results are superimposed on the subject's own structural brain MRI scan. The bright areas pinpoint regions of heightened cortical activity associated with hallucinatory events [reproduced with permission from D. A. Silbersweig *et al.* (1995), A functional neuroanatomy of hallucinations in schizophrenia. *Nature* **378,** 176–179].

paralimbic regions that may be involved in the generation or modulation of hallucinations (Fig. 2). Limbic structures, the least differentiated, or least distinctly structured, older regions of the cortex, are involved in the linking of drives with experience and the processing of emotion. Paralimbic regions are intermediate in structure between, and interconnected with, limbic and heteromodal association areas and serve to integrate emotion and drive with highly processed sensory information. Because limbic and paralimbic structures are closely interconnected and functionally integrated, they are often referred to collectively as the limbic system.

In our study, activation in limbic regions involved hippocampi (extending to the adjacent amygdalae) and ventral striatum. The hippocampal formation, a convoluted structure within the medial temporal lobe, is involved in memory and the processing of contextual

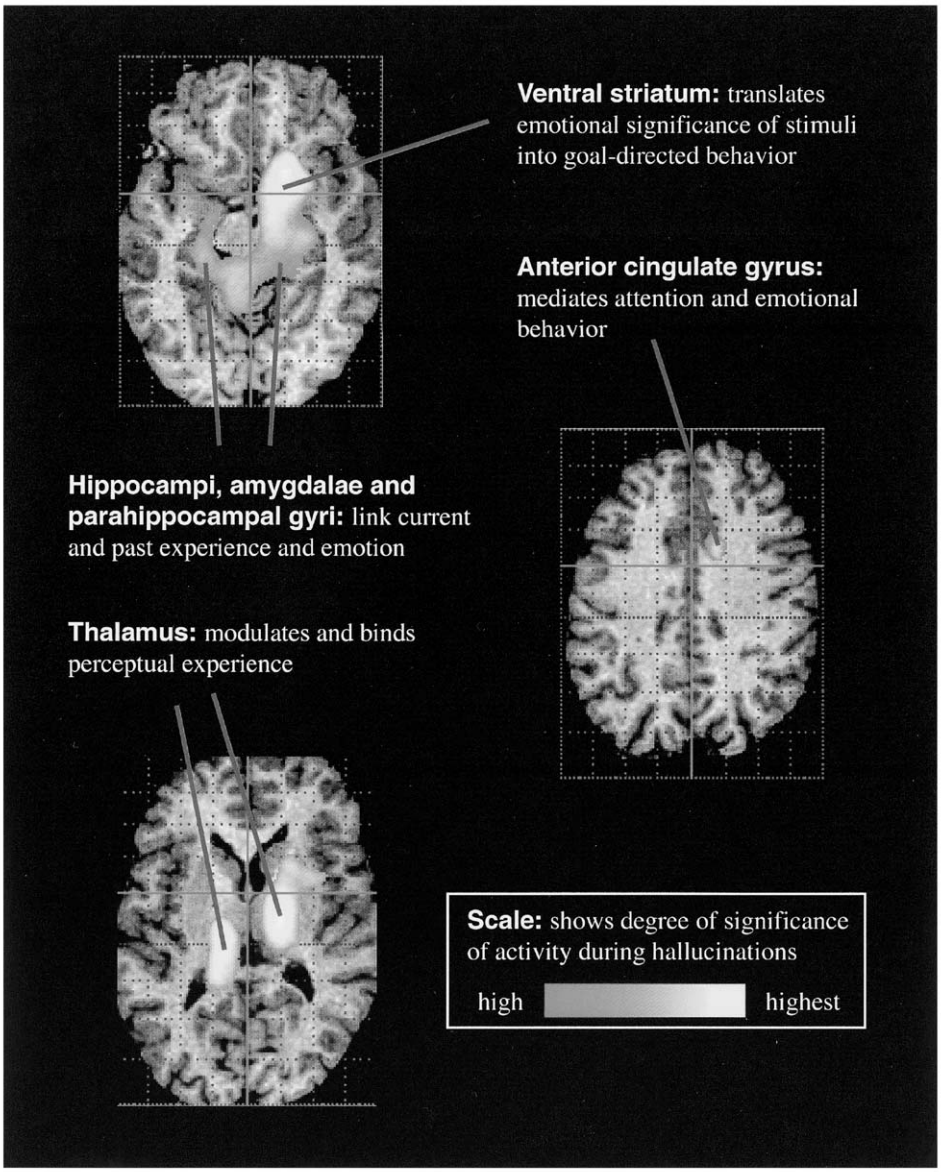


Figure 2 Common areas of brain activity in a group of schizophrenic patients experiencing auditory/verbal hallucinations. The functional PET results are superimposed on an anatomical MRI template. The bright areas pinpoint regions of heightened thalamic, limbic, and paralimbic activity associated with hallucinatory events [reproduced with permission from D. A. Silbersweig *et al.* (1995), A functional neuroanatomy of hallucinations in schizophrenia. *Nature* **378,** 176–179].

aspects of emotional evaluation. The amygdaloid complex, a collection of nuclei adjacent to and interconnected with the hippocampus, plays a central role in evaluating the emotional significance of internally and externally generated stimuli. Both of these structures send output to the ventral striatum, the limbic portion of the subcortical basal ganglia, where emotional significance is translated into goal-directed behavior. Activation in paralimbic regions involved parahippocampal gyri, anterior cingulate, and orbitofrontal cortex. The parahippocampal gyrus, which lies on the medial surface of the temporal lobe, integrates sensory output from heteromodal and more complex unimodal association areas with limbically processed information. The anterior cingulate, the frontmost portion of a band running from medial frontal to parahippocampal regions, is involved in attention and social/emotional behaviors. The orbital frontal cortex

is located in the medial ventral frontal lobes. Like the amygdala, it participates in the evaluation of emotional significance and sends output to the ventral striatum. In contrast to the amygdala, the orbital frontal region is able to modulate emotional responsivity and readjust behavioral responses to stimuli when their reinforcement value is changed or when a more complex assessment of the current context suggests the need for modification.

Just as abnormal activity in cortical sensory regions is correlated with the form and content of hallucinations, it is likely that aberrant activity in limbic/paralimbic regions gives rise to the marked emotional significance of hallucinations in the setting of schizophrenia. Further evidence of a role for limbic system dysfunction in the generation of schizophrenic symptoms is provided by postmortem, neuropsychological, and neuroimaging studies that reveal structural and functional abnormalities of limbic regions in individuals with schizophrenia, including hyperactivity of temporal regions (left greater than right) associated with psychosis. Activity of the limbic system is closely interconnected with that of dopamine, a neurotransmitter implicated in the generation of hallucinations and delusions in the settings of schizophrenia, medication toxicity, and drug abuse. Dopaminergic activity is regulated, in part, by input from limbic system structures and in turn appears to modulate the responsiveness of ventral striatal neurons to limbic inputs. Recent work suggests that glutamate, an excitatory neurotransmitter, may also play a role in both limbic dysregulation and schizophrenia. Hallucinations that occur in the context of severe emotional stress may also involve abnormal limbic activity.

Temporolimbic structures also play a role in the generation of hallucinations associated with epilepsy. In addition to photopsias, the onset of partial seizures can be accompanied by simple hallucinations in any modality, reflecting ictal discharges in primary sensory areas, or by complex hallucinations, reflecting discharges in limbic and sensory association areas. These often involve temporal regions including hippocampus and amygdala, which have the lowest seizure thresholds of all brain structures, as well as sensory association areas. Like the complex hallucinations seen in schizophrenia, these are often emotionally charged. Unlike those seen in schizophrenia, they are more often visual than auditory and are not usually believed by the person experiencing them to represent reality. Relatedly, electrical stimulation of temporal lobe regions, including the amygdala, can give rise to hallucinatory experiences. In addition to hallucinations experienced during seizures, there is evidence that individuals who have suffered from epilepsy for more than 10 years may develop hallucinations between seizure episodes. These are more likely to resemble fully those seen in schizophrenia as they are often emotionally charged, are as likely to be auditory as visual, are accompanied by delusions, and are believed to represent reality. As in schizophrenia, they appear to be associated with temporal lobe abnormalities (left more often than right).

3. Frontal/Executive Activity Associated with Hallucinations

The lack of awareness that hallucinatory experience does not correspond to reality is a striking feature of schizophrenia. In addition to temporal lobe abnormalities, numerous studies have revealed frontal dysfunction and abnormal frontotemporal connectivity associated with schizophrenia. The frontal lobes, in concert with interconnected regions, mediate the higher, more complex aspects of cognition, such as judgment, insight, and self-monitoring. These are termed executive functions. Although relevant studies have produced mixed results, there is evidence that frontal dysfunction may contribute to the inability of individuals with schizophrenia to identify the internal origin of their hallucinatory experience and its relation to their illness. Temporal lobe epilepsy may also be accompanied by executive as well as other forms of cognitive dysfunction and by abnormalities of frontal activity.

II. OTHER INVESTIGATIONAL APPROACHES TO HALLUCINATIONS

The descriptive and analytic framework employed in this article represents one approach to hallucinations. Others tend to be complementary, rather than exclusive, with their boundaries increasingly blurred as convergence and integration occur. The resulting interdisciplinary synthesis has enhanced our understanding of hallucinations. At the cognitive level of analysis, numerous mechanisms have been posited to play a role in the generation of hallucinations, with support derived from psychologic, electrophysiologic, and animal studies. Several theories focus on abnormalities in the processing of input or its comparison with past experience. These can be seen to dovetail with the sensory input, midbrain/thalamic, sensory cortical,

and limbic/paralimbic disturbances described previously. Others focus on cognitive abnormalities that may give rise to deficits in the ability to discriminate between external and self-generated events. These theories, in turn, are related to frontal/executive disturbances. At the neurochemical level, much work has been done on the relation between disturbances in neurotransmitter systems, such as dopamine, serotonin, and glutamate, and hallucinations. This is clearly relevant to the limbic and midbrain/thalamic disturbances noted previously. At the computational level, neural networks models have been used to investigate links between posited cognitive or physiologic abnormalities and hallucinations, with intriguing results. At the social/psychological level, investigations and clinical observations suggest that personal, social, and cultural factors play a role in the development and content of hallucinations—a role likely to be mediated by limbic and cortical brain regions involved in learning and complex cognition.

III. TREATMENT OF HALLUCINATIONS

For hallucinations in the setting of schizophrenia, medications that alter transmission of dopamine and related neurotransmitters, termed neuroleptics, are the mainstay of treatment. In other settings, the first step in the treatment of hallucinations is to address the condition that underlies their existence. When this is impossible or ineffective, neuroleptic medications may be tried. However, these tend to be less effective in settings that do not involve limbic, striatal, or dopaminergic pathology. Fortunately, hallucinations in the setting of sensory input disorders, where neuroleptics are least effective, are often less disturbing to those experiencing them, as described previously. Such hallucinations sometimes respond to carbamazepine, a medication often used for seizure prevention, mood stabilization, or control of pain originating in the nervous system. This is consistent with models of aberrant neural activity described previously.

When hallucinations are distressing and unresponsive to medication, psychological treatments may be helpful. Cognitive–behavioral approaches involve distracting activities or sensory input as well as behavioral and cognitive tasks. Supportive approaches involve helping patients understand their condition, solve problems, and adapt to reality. Psychological approaches tend to decrease distress associated with hallucinations and improve overall functioning rather than ameliorate hallucinations per se.

Future developments in the treatment of hallucinations are likely to be guided by the functional neuroanatomic approach. Recent investigations into the mechanism of action of antipsychotic medications have increasingly focused on the specific cerebral regions modulated by relevant neurotransmitters. In addition, a recent study examined the efficacy of transcranial magnetic stimulation (TMS), a novel technique for altering focal cortical activity through application of a magnetic pulse, in the treatment of persistent auditory hallucinations in schizophrenic patients. The results suggest that administration of TMS to the left temporoparietal regions noted (in the PET study discussed previously) to be active during auditory hallucinations can markedly decrease the severity of such events.

IV. FUTURE DIRECTIONS

The approach to hallucinations presented in this article represents an attempt to organize and synthesize current knowledge from a functional neuroanatomic perspective. It should be regarded as a framework on which further data from disciplines relevant to neuroscience can be laid. As new data emerge and interdisciplinary integration proceeds, our understanding of hallucinations will undoubtedly gain increased specificity and complexity and grow in new directions.

See Also the Following Articles

BODY PERCEPTION DISORDERS • DREAMING • EPILEPSY • LIMBIC SYSTEM • MIDBRAIN • PHANTOM LIMB PAIN • SCHIZOPHRENIA • SENSORY DEPRIVATION • STROKE • THALAMUS AND THALAMIC DAMAGE • VISUAL DISORDERS

Suggested Reading

Amador, X., and David, A. (Eds.) (1998). *Insight and Psychosis.* Oxford Univ. Press, New York.

Devinsky, O., and Luciano, D. (1991). Psychic phenomena in partial seizures. *Sem. Neurol.* **11**(2), 100–109.

Esquirol, J. (1837/1965). *Mental Maladies.* Hafner, New York.

Frith, C. (1998). The role of the prefrontal cortex in self-consciousness: The case of auditory hallucinations. In *The Prefrontal Cortex: Executive and Cognitive Functions* (A. Roberts, T. Robbins, *et al.*, Eds.), pp. 181–194. Oxford Univ. Press, New York.

Hoffman, R., Boutros, N., Berman, R., Roessler, E., Belger, A., Krystal, H., and Charney, D. (1999). Transcranial magnetic

stimulation of left temporoparietal cortex in three patients reporting hallucinated "voices". *Biol. Psychiatr*. **46,** 130–132.

Krystal, J., Abi-Dargham, A., Laruelle, M., and Moghaddam, B. (1999). Pharmacologic models of psychoses. In *Neurobiology of Mental Illness* (D. Charney, E. Nestler, and B. Bunney, Eds.), pp. 214–224. Oxford Univ. Press, New York.

Lhermitte, J. (1922). Syndrome de la calotte pedonculaire. Les troubles psychosensorielle dans les lesions du mesencephale. *Rev. Neurologique* **38,** 1359–1365.

Manford, M., and Andermann, F. (1998). Complex visual hallucinations: Clinical and neurobiological insights. *Brain* **121,** 1819–1840.

Melzack, R. (1990). Phantom limbs and the concept of a neuromatrix. *TINS* **13**(3), 88–92.

Mesulam, M.-M. (2000). Behavioral neuroanatomy: Large-scale networks, association cortex, frontal syndromes, the limbic system, and hemispheric specializations. In *Principles of Behavioral and Cognitive Neurology* (M.-M. Mesulam, Ed.), pp. 1–120. Oxford Univ. Press, New York.

Silbersweig, D., and Stern, E. (1996). Functional neuroimaging of hallucinations in schizophrenia: Toward an integration of bottom-up and top-down approaches. *Mol. Psychiatr*. **1,** 367–375.

Trimble, M., and Schmitz, B. (1998). The psychoses of epilepsy: A neurobiological perspective. In *Psychiatric Comorbidity in Epilepsy: Basic Mechanisms, Diagnosis, and Treatment* (H. McConnell, P. Snyder, *et al*., Eds.), pp. 169–186. American Psychiatric Press, Washington, DC.

Hand Movements

J. RANDALL FLANAGAN* and ROLAND S. JOHANSSON†
*Queen's University, Canada and †Umeå University, Sweden

GLOSSARY

grasp stability control The control of grip forces such that they are adequate to prevent accidental slips but not so large as to cause unnecessary fatigue or damage to the object or hand.

haptic perception Perception through the hand based on tactile and somatosensory information.

internal models Neural circuits that mimic the behavior of the motor system and environment and capture the mapping between motor outputs and sensory inputs.

precision grip The grip formed when grasping an object with the distal tips of digits. Usually refers to grasping with the tips of the thumb and index finger on either side of an object.

sensorimotor control The use of both predicted and unexpected sensory information in the control of action.

The human hand and the brain are close partners in two important and closely interconnected functions: exploration of the physical world and reshaping of parts of this world through manipulation. The highly versatile functions of the human hand depend on both its anatomical structure and the neural machinery that supports the hand. This article focuses on the sensorimotor control of hand movements in object manipulation–a hallmark of skilled manual action. The article also examines relationships between the two main functions of the hand–object perception and object manipulation.

I. THE ACTING AND PERCEIVING HAND

Many of our cultural and technological achievements that mark us as human depend on skilled use of the hand. We use of our hands to gesture and communicate, make and use tools, write, paint, play music, and make love. Thus, the human hand is a powerful tool through which the human brain interacts with the world. We use our hands both to perceive the world within our reach (haptic perception) and to act on this world. These two functions of the hand, which are largely accomplished by touching and manipulating objects in our environment, are intimately related in terms of sensorimotor control. Haptic perception requires specific hand movements that are tailored to the kinds of information the perceiver wishes to extract. For example, to obtain information about the texture of an object, people rub their fingertips across the object's surface, and to obtain information about shape they trace the contour of the object with their fingertips. Conversely, in object manipulation sensory and perceptual information is critical for precise motor control of the hands. The fact that individuals with numbed digits have great difficulty handling small objects even with full vision illustrates the importance of somatosensory information from the fingertips.

To control both the exploratory and manipulatory functions of the hand, the brain must obtain accurate descriptions of various mechanical events that take place when objects are brought into contact with the

hand. Mechanoreceptive (tactile) sensors in the glabrous skin of the volar aspect of the hand play an essential role in providing such information. The density of mechanoreceptors increases in the distal direction of the hand and is exquisitely high in the fingertips. As a perceptual organ, the hand has several advantages over the eyes. The hand can effectively "see around corners," allowing us to explore all sides of an object, and it can directly appreciate object properties such as weight, compliance, and slipperiness.

The numerous skeletal and muscular degrees of freedom of the hand, orchestrated by highly developed neural control systems, provide for tremendous dexterity that allows for both delicate exploration and versatile manipulation of objects. With approximately 30 dedicated muscles and approximately the same number of kinematic degrees of freedom, the hand can take on all variety of shapes and functions, serving as a hammer one moment and a powerful vice or a delicate pair of tweezers the next. The utility of hand movements is further enhanced by our ability to amplify the functions of the hand by using tools.

Different primates have very different hand movement capacities, with humans demonstrating the greatest dexterity. For example, true opposition between the thumb and index finger is only observed in humans, the great apes, and Old World monkeys. New World monkeys can manage pseudo-opposition, but prosimians are only capable of crude grasping. It seems improbable that the tremendous dexterity of the human hand can be explained solely by differences in anatomical factors given that the structural anatomy of the hands of different primates seems similar. This is not to say, however, that anatomical differences do not contribute. For example, the human thumb is much longer, relative to the index finger, than the chimpanzee thumb. This allows humans to grasp small objects precisely between the distal pads of the thumb. Similarly, the greater independence of finger movements in humans compared to monkeys arises, in part, from differences in the passive biomechanical connections among tendons. Humans have more individuated muscles and tendons with which to control the digits.

In addition to structural factors, a major contributor to differences in hand movement capacity among primates, and between primates and lower mammals, is the neural machinery underlying hand movement. Compared to lower mammals, primates have evolved extensive cerebral cortical systems for controlling the hand and the corticospinal pathways have taken on an increasingly dominant role in controlling movement. Moreover, in primates the corticospinal tracts include direct connections between neurons in cortical motor areas and spinal motorneurons. Through these corticomotoneuronal connections, the cerebral cortex possesses monosynaptic control over motorneurons whose axons connect, in particular, with the hand muscles. In effect, these direct connections have moved the hand "closer" to the cerebral cortex. Furthermore, through cortical motor areas the corticospinal tracts provide rapid access to the hand from most other cortical areas and from subcortical structures, including the cerebellum and the basal ganglia, tightly involved in motor control.

The development of cortical systems for controlling the hand in primates parallels the evolution of the arm from a prop for balance and locomotion (in four-legged mammals) to a free and dexterous tool for sensing and acting on objects in the environment. The denser neuronal substrate for hand control provides more flexibility in the patterning of muscle activation and supports the ability to perform independent finger movements. Interestingly, across primates, there is a linkage between the number of corticomotoneuronal connections and manual dexterity in terms of performing tasks that require independent finger movements. Although there are many advantages in terms of control, the reliance on cortical control comes at a cost. Lesions to the motor cortex or corticospinal pathways due, for example, to cerebral vascular accident can be particularly devastating in humans.

The importance of the cortical involvement in fine fingertip control can be further appreciated by considering parallels between the ontogenetic development of central neural pathways and that of hand function. The efficacy of the corticomotoneuronal system can be probed using transcranial magnetic stimulation (TMS) of the brain. TMS applied over the hand area of the motor cortex activates muscles of the contralateral hand. During development the latency of this activation, and the stimulation strength required to elicit a response, decreases as the corticomotoneuronal connections are established. The conduction delays in these motor pathways, as well as in the somatosensory pathways conveying signals from the sensors of the hand, rapidly decrease during the first 2 years after birth and thereafter remain constant at adult values. Responses within the adult latency range appear during the age range in which young children demonstrate important improvements in their ability to grasp objects using the tips of the index finger and thumb. Similar parallels between hand function and corticomoto neurone (CM) system development have

been demonstrated in monkeys using various electrophysiological and anatomical techniques.

II. SENSORIMOTOR CONTROL OF HAND MOVEMENTS IN OBJECT MANIPULATION

To understand and appreciate how the brain controls movements of the hand, it is best to study the natural behavior of the hand in everyday manipulatory tasks. During the past 20 years, the sensorimotor control of the hand in precision manipulation task has been investigated in great detail. In this section, we review what has been learned about the sensorimotor control of natural hand movements when grasping and manipulating objects with the fingertips.

The remarkable manipulative skills of the human hand are the result of neither rapid sensorimotor processes nor fast or powerful effector mechanisms. Rather, the secret lies in the way manual tasks are organized and controlled by the nervous system. Successful manipulation requires the selection of motor commands tailored to the manipulative intent, the task at hand, and the relevant physical properties of the manipulated object. For instance, most tasks require that we stabilize the object within our grasp as we move the object or use it as a tool. To prevent slips and accidental loss of the object we must apply adequately large forces normal to the grip surfaces (*grip forces*) in relation to destabilizing forces tangential to the grip surfaces (*load forces*) (Fig. 1). At the same time, excessive grip forces must be avoided because they cause unnecessary fatigue and may crush fragile objects or injure the hand. Hence, the term grasp stability entails prevention of accidental slips as well as excessive fingertip forces.

When grasping and manipulating objects, the forces needed to ensure grasp stability depend on the physical properties of the object. Object properties such as weight, slipperiness, shape, and weight distribution all impose constraints on the fingertip forces (including their magnitudes, directions, and points of application) required for stability. Thus, a basic question for understanding the control in manipulation is how do people adapt their fingertip forces to the constraints imposed by various object properties. Although visual information about object properties may be helpful in terms of force selection, ultimately people adapt to such constraints by using sensory information provided by digital mechanoreceptors. Individuals with impaired digital sensibility have great difficulty performing manipulation tasks even under visual guidance. For instance, they often drop objects, may easily crush fragile objects, and have difficulties in dressing themselves because they cannot complete such apparently simple tasks as buttoning a shirt. Thus, it is clear that critical sensorimotor control processes required for manipulation are lost with impaired digital tactile sensibility.

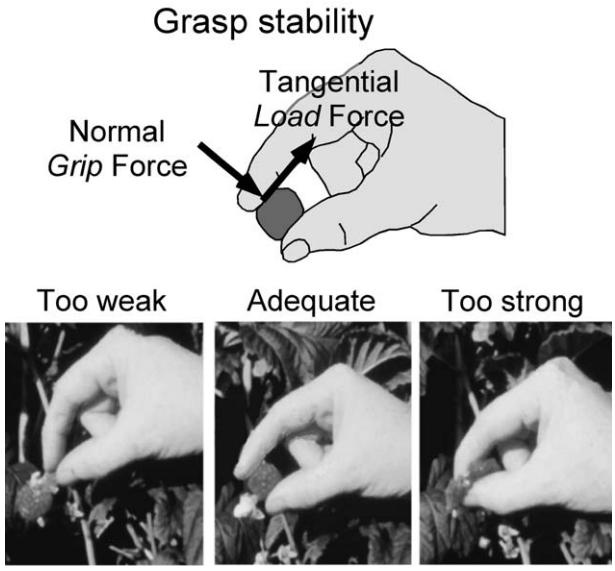


Figure 1 When manipulating objects grasped with a precision grip, we must carefully control the balance between grip force, normal to the contact surfaces, and load force tangential to the grasp surfaces. If grip force is too weak for a given load force, we risk having the object slip from our grasp. If grip force is too strong, we may crush the object or damage our hand and we waste energy.

The control of grip and load forces in object manipulation involves subtle interplay between two types of control: reactive control based on sensory feedback and predictive or feedforward control. These two control mechanisms are closely linked. On the one hand, reactive control mechanisms are invoked when errors arise between actual sensory feedback and the expected sensory feedback predicted from feedforward mechanisms. On the other hand, errors in sensory prediction are not only used for feedback control but also used to update feedforward mechanisms to reduce future prediction errors. In the following sections, we consider these two control processes in detail.

A. Feedback Control Based on Digital Sensors

One way to use digital sensors to adjust the force output would be to engage these sensors in feedback

loops. However, such loops imply large time delays. These time delays arise from impulse conduction time in peripheral nerves, conduction and processing time in the central nervous system, and the inherent sluggishness of muscles. In humans, these factors sum to at least 100 msec for the generation of a significant force response. Consequently, closed-loop feedback is not effective for rapid movement involving frequencies above 1 Hz. In natural manipulation tasks, movement frequency components up to 5 Hz can be observed. Thus, feedback control alone cannot support control of grip force for grasp stability in these movements.

Despite these control limitations, feedback control is essential in certain types of manipulative tasks. For example, feedback control is required in reactive tasks in which we restrain "active" objects that generate unpredictable load forces tangential to the grip surfaces. Examples of tasks in which we must deal with active objects are holding a dog's leash, restraining a child by holding his or her arm, or operating power tools. Consider the situation depicted in Fig. 2A

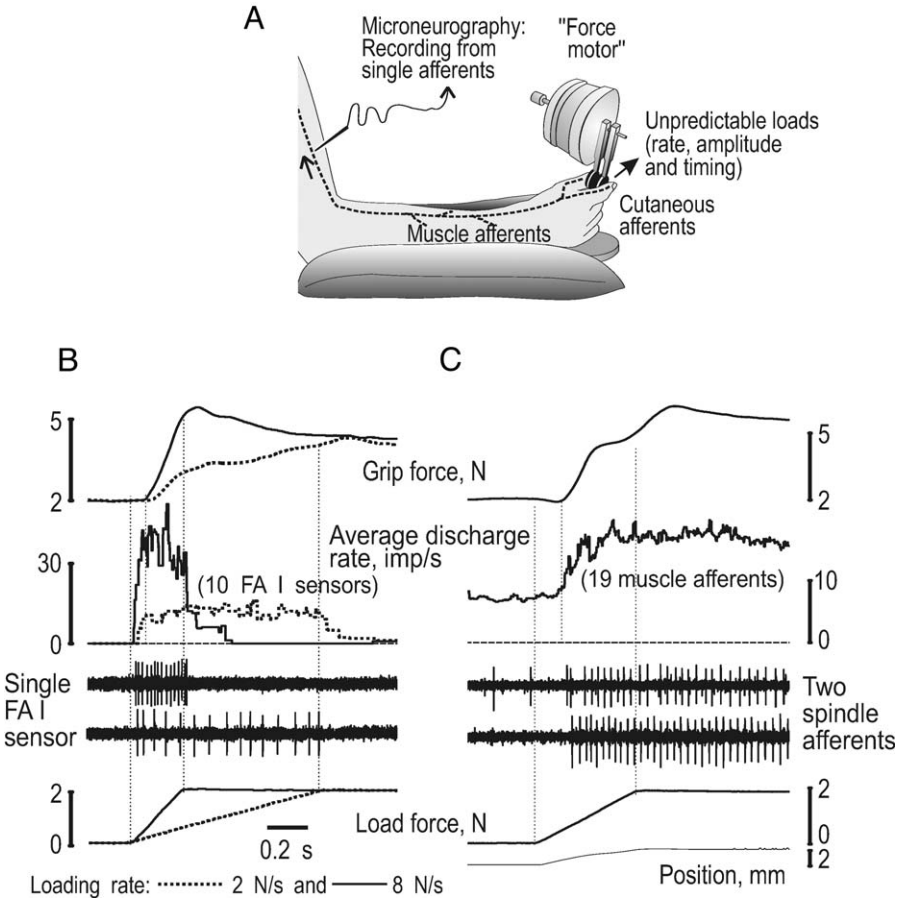


Figure 2 Peripheral afferent and reactive grip force responses to unpredictable loading of the precision grip by a pulling force. (A) The subject grasped the manipulandum with the tips of the thumb and index finger contacting parallel grip surfaces 25 mm apart. The force motor could deliver load forces pulled away from or pushed toward the hand. The grip and load forces, normal and tangential to the grip surfaces, respectively, and the position of the manipulandum were recorded. Afferent activity was recorded from the median nerve, with percutaneously inserted tungsten needle electrodes impaling the nerve about 10 cm proximal to the elbow. (B) Grip responses and average discharge rate of 10 FA I sensors to 2 N pulling loads delivered to the receptor-bearing digit at 2 N/sec (dashed lines) and 8 N/sec (solid lines). The two traces of single unit recordings are examples of responses in a single FA I sensor during load trials at 8 N/sec (upper trace) and 2 N/sec (lower trace). (C) Grip response and average discharge rate of 19 muscle afferents located in the long flexor muscles of the index, middle, or ring finger to 2.0 N pulling loads delivered at 4 N/sec. The single unit recordings are examples of responses in two different muscle spindle afferents. (B and C) The averages of forces and discharge rates are synchronized to the onset of the loading ramp; discharge rate represents average instantaneous frequency (adapted with permission from Macefield, V. G., Häger-Ross, C., and Johansson, R. S., *Exp. Brain Res.* **108,** 155–171, 1996; and Macefield, V. G., and Johansson, R. S., *Exp. Brain Res.* **108,** 172–184, 1996. Copyright © 1996 by Springer-Verlag).

in which an individual grasps an object attached to a force motor using a precision grip with the tips of the thumb and index finger on opposing vertical surfaces. The motor is used to generate increasing load forces (tangential to the grip surfaces) that are unpredictable in terms of onset time, amplitude, and direction (loading and unloading). To prevent the object from slipping, people automatically respond to increases in tangential load by increasing grip force normal to the grip surfaces in parallel with the load force changes (see load and grip force signals in Figs. 2B and 2C). When the load stops increasing, the grip force also stops increasing and may decrease slightly. Importantly, the changes in grip force lag behind the load force changes because they are reactively generated. A reactive grip response is initiated after a delay of approximately 100 msec but this varies with the load force rate. Because of this time lag, the object will slip from grasp unless the background grip force prior to a load increase is strong enough to meet the initial load increase. Indeed, following slips and trials with a high rate of load force increases, people learn to increase the initial background grip force as an adaptation to the expected range of loadings.

Figure 2A also shows signals, recorded using the technique of microneurography, from single nerve fibers of the median nerve that supply cutaneous, muscle, and joint sensors. Experiments with cutaneous anesthesia have demonstrated that reactive fingertip force responses are driven primarily by digital cutaneous inputs. Signals from fast adapting (FA I) cutaneous afferents seem most important, but slowly adapting cutaneous afferents may also contribute. As illustrated in Fig. 2B, the intensity of the cutaneous afferent responses is scaled by the rate of load force increase, and the afferent responses commence before the onset of the grip response. Furthermore, the size and duration of the grip force increase is scaled with the intensity and duration of the afferent response. This scaling is an attractive feature for feedback-based control.

Whereas cutaneous afferents contribute to the initiation and initial scaling of grip force responses, afferents from intrinsic and extrinsic hand muscles and interphalangeal joints do not respond to load increases early enough to allow them to contribute to the initiation of these grip responses. The muscle afferents respond reliably after the onset of the reactive grip force response and their discharge rates are related to changes in force output and, hence, to muscle activity (Fig. 2C). Thus, these muscle afferents are primarily concerned with events in the muscle itself rather than functioning as exteroceptors sensing mechanical events at the fingertips.

B. Feedforward Control Processes

Almost everyone will recall having fallen victim to an older sibling, cousin, or friend who passed us an empty box while pretending it was very heavy. When we took the box, our arms flailed upwards. This trick demonstrates that when we interact with objects, we anticipate the forces required to complete the task. Although it may occasionally result in large movement errors, anticipatory or feedforward control is essential for skilled object manipulation. Feedback control is important when our predictions are erroneous or, as in reactive tasks, when predictions are unavailable. However, because of the long time delays, feedback control cannot support the swift and skilled coordination of fingertip forces observed in most manipulation tasks that involve ordinary "passive" objects. Instead, the brain relies on feedforward control mechanisms that take advantage of the stable and predictable physical properties of these objects. These mechanisms parametrically adapt force motor commands to the relevant physical properties of the target object.

Figure 3 illustrates parametric anticipatory adjustments of motor output to object weight, friction between the object and skin, and shape of the contact surface. The task is to lift a test object from a support surface, hold it in air for a couple of seconds, and then replace it. To accomplish this task, the vertical load force increases until liftoff occurs, stays constant during the hold phase, and then starts to decrease when the object contacts the support surface during replacement. When lifting objects of different weight (Fig. 3A), people scale the rate of increase of both grip force and load force to object weight such that lighter and heavier objects tend to be lifted in about the same amount of time. The scaling occurs prior to liftoff–before sensory information about object weight becomes available–and is therefore predictive. To deal with changes in friction, the motor system adjusts the balance between grip force and load force. As shown in Fig. 3B, when lifting equally weighted objects of varying slipperiness, people scale the rate of increase of grip force while keeping the rate of change of load force constant. Thus, the ratio of these force rates is a controlled parameter that is set to the current frictional conditions. A similar scaling of the grip-to-load force ratio is observed when object shape is varied. A larger

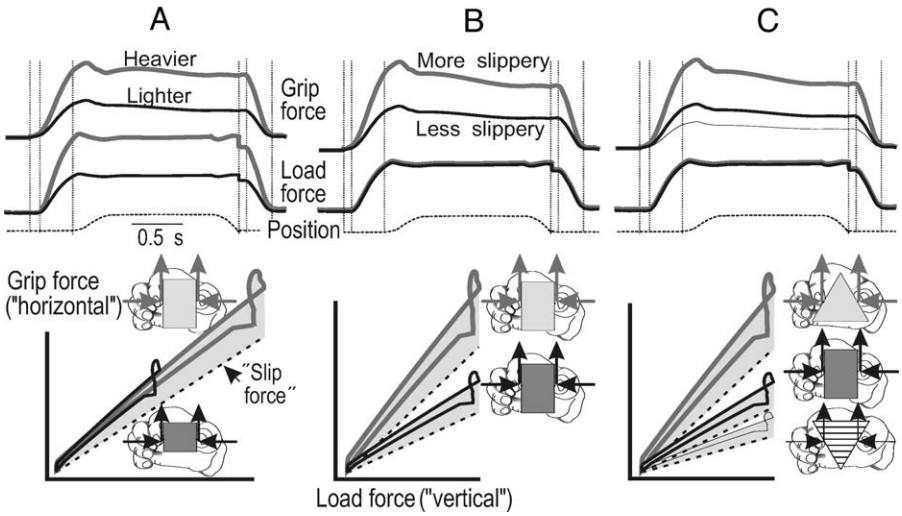


Figure 3 Feedforward adjustments of motor output to object weight (A), frictional conditions (B), and object shape (C) in a task in which a test object is lifted with a precision grip, held in air, and then replaced. The top graphs show horizontal grip force, vertical load force, and the vertical position of the object as a function of time for two superimposed trials. The bottom graphs show the relation between load force and grip force for the same trials. The dashed line indicates the minimum grip-to-load force ratio required to prevent slip. The gray area represents the safety margin against slip. After contact with the object (left most vertical line, top), grip force increases by a short period while the grip is established. A command is then released for simultaneous increases in grip and load force (second vertical line). This increase continues until the load force overcomes the force of gravity and the object lifts off (third vertical line). After replacement of the object and table contact occurs (fourth line), there is a short delay before the two forces decline in parallel (fifth line) until the object is released (sixth line) (adapted with permission from Johansson, R. S., and Westling, G., *Exp. Brain Res.* **56,** 550–564, 1984 by Springer-Verlag; Johansson, R. S., and Westling, G., *Exp. Brain Res.* **71,** 59–71, 1988. Copyright © 1988 by Springer-Verlag; and Jenmalm, P., and Johansson, R. S., *J. Neurosci.* **17,** 4486–4499, 1997 Copyright © 1997 by the Society for Neuroscience).

ratio is used when the grip surfaces are tapered upward compared to downward (Fig. 3C).

In each example shown in Fig. 3, grip force increases and decreases in phase with (and thus predicts) changes in vertical load force. This parallel coordination of grip force and load force ensures grasp stability. The grip force at any given load force is controlled such that it exceeds the corresponding minimum grip force, required to prevent slip, by a small safety margin (gray areas in the bottom of Fig. 3). This minimum grip force depends on the weight of the object, the friction between the object and skin, and the shape (e.g., angle) of the contact surfaces.

This parallel coordination of grip force and load force is a general feedforward control strategy and is not specific to any particular task or grip configuration. Parallel force coordination is observed when grasping with two or more digits of the same hand or both hands, when grasping with the palms of both hands, and even when gripping objects with the teeth. Moreover, it does not matter whether the object is moved by the arm or, for example, by the legs as when jumping with the object in hand. Importantly, the parallel coordination of grip and anticipatory load force is not restricted to common inertial loads. People also adjust grip force in parallel with load force when pushing or pulling against immovable objects and when moving objects subjected to elastic and viscous loads. Figure 4 illustrates parallel coordination of grip and load forces under varying load conditions. People alternately pushed and pulled an object instrumented for force sensors and attached to a simple robot that could simulate various types of opposing loads acting tangential to the grasp surfaces (Fig. 4A). Figures 4B and 4C show kinematic and force records obtained under three different load conditions: an acceleration-dependent inertial load, a velocity-dependent viscous load, and an elastic load that largely depended on position but also contained viscous and inertial components. In all three cases, the grip force normal to the grasp surfaces changes in parallel with the magnitude of the load force tangential to the grasp surface. Importantly, the relationship between arm movement motor commands and the load experienced at the fingertips depends on the type of load being moved. Thus, to adjust grip force in parallel with load

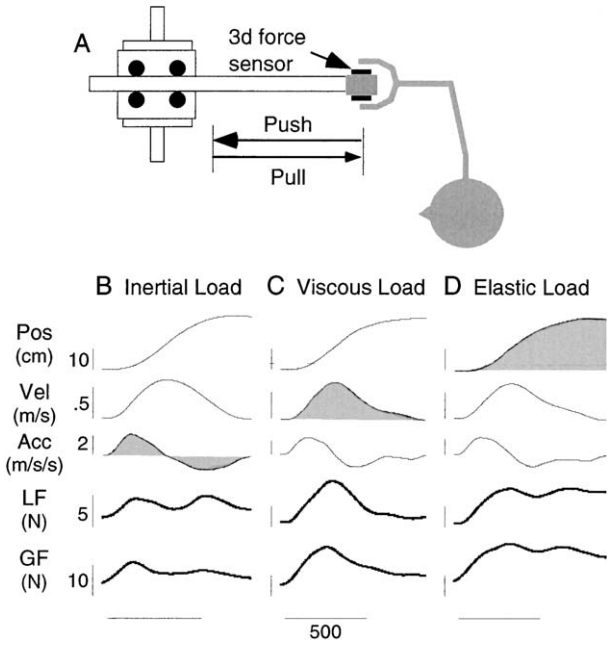


Figure 4 Kinematic and force records from one subject under the three load conditions. Shaded regions indicate the primary kinematic variable on which load depended. Under all three load conditions, grip force (GF) is adjusted in parallel with fluctuations in load force (LF), with the resultant load tangential to the grasp surface. The dashed vertical lines indicate movement onset (modified with permission from Flanagan, J. R., and Wing, A. M., *J. Neurosci.* **17,** 1519–1528, 1997. Copyright © 1997 by the Society for Neuroscience).

force under the different load conditions, people had to alter the mapping between the motor command driving arm movement and that driving the grip force.

In most everyday tasks, destabilizing loads acting on the grasp include not only linear load forces but also torques tangential to the grasped surfaces. Such torsional loads occur whenever we tilt an object around a grip axis that does not intersect the vertical line through the object's center of mass. In addition, torque loads arise in many natural manipulatory tasks due to changes in the orientation of the grip axis with respect to gravity. For example, this occurs when we hold a book flat by gripping it between the fingers beneath and the thumb above (vertical grip axis) and then rotate it by a pronation movement to put it in a bookshelf (horizontal grip axis). Because we rarely take a book such that the grip axis passes through its center of mass, a torque will develop in relation to the grasp. Importantly, the sensorimotor programs for object manipulation account for torsional loads by predicting the consequences of object rotation both when we rotate objects around the grip axis and when we rotate the grip axis in the field of gravity. Rotational slips are prevented by automatic increases in grip force that parallel increases in tangential torque. The sensorimotor programs thus model the effect of the total load in terms of linear forces, tangential torques, and their combination.

C. Internal Models Underlying Predictive Force Control

As illustrated in Fig. 3A, with objects of different weight, people use different rates of force increase prior to liftoff. Since there is no sensory information available about object weight until liftoff, this behavior indicates that people predict the final force requirements. Likewise, with objects of different friction (Fig. 3B) and shape (Fig. 3C), the force output is tailored to the properties of the object from the start of the initial force attack, well before sensory information from the digits obtained after contact with the object could have exerted any influence. Thus, in all three cases, the motor controller operates in a feedforward fashion and uses motor command parameters determined by internal models that capture the physical properties of the object. Figure 4 further illustrates that such internal models also capture dynamic properties of objects. The question arises as to how such models are selected and updated for different objects and after changes in object properties.

1. Prediction Based on Object Shape

Figures 5A and 5B show three consecutive trials taken from a series of lifts in which the angle of the grasped surfaces was changed between trials in a pseudorandom order. The sequence is 30°, −30°, and −30° and thus includes a transition from an upward tapered object (30°) to a downward tapered object (−30°). In the trials preceding this sequence, a 30° object was lifted. First consider the trials in which vision of the objects is available (Fig. 5A). When the shape of the object is changed, the grip force is adjusted from the very start of the lift in anticipation of the lower grip force required to lift the object. In particular, grip force is now increased more slowly before sensory feedback from the digits could have influenced the motor output. The predictive adjustment in grip force observed in the first trial after the switch in object shape is very accurate. Indeed, no further adjustment is

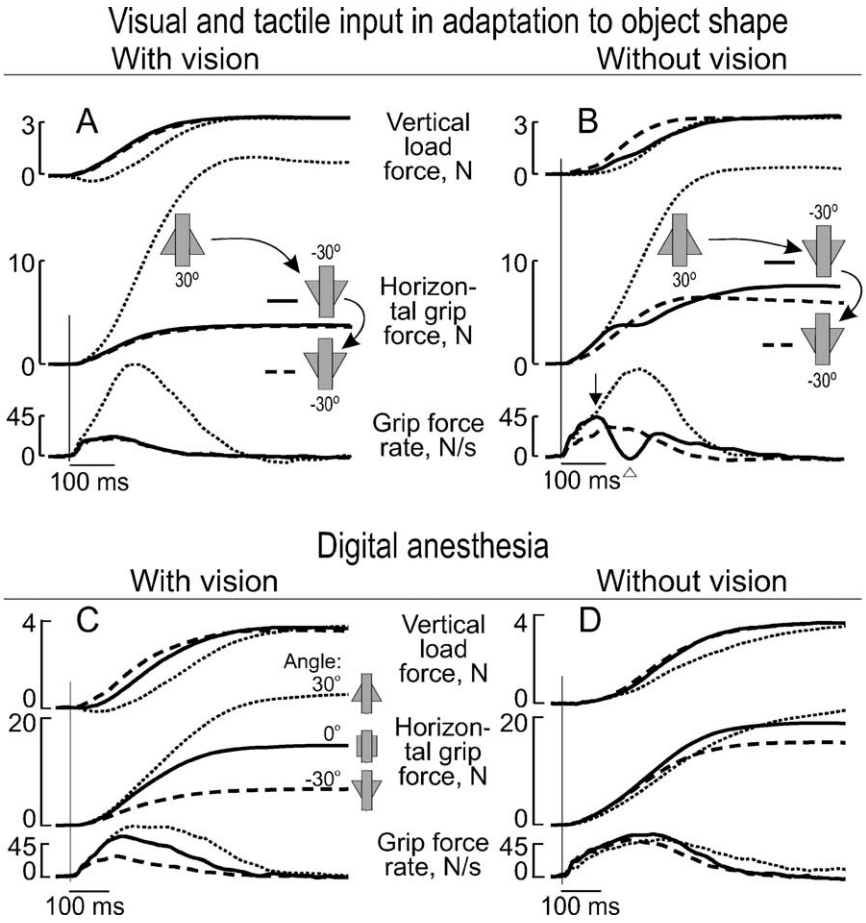


Figure 5 (A and B) Force adjustments to changes in surface angle during lift series in which surface angle was unpredictably varied between lift trials. Vertical load force, horizontal grip force, and grip force rate shown as a function of time for trials with (A) and without (B) vision and with normal digital sensibility. The dotted curves are from the last trial before the switch with the 30° object. The solid curves show the next trial with the −30° object. These curves illustrate adjustments to the smaller angle. The dashed lines show the following trial again with the −30° object. The downward arrow in B indicates the point in time when the new surface angle was expressed in terms of motor output. (C and D) Adaptation to surface shape during digital anesthesia with (C) and without (D) vision. Vertical load force, horizontal grip force, and grip force rate as a function of time for trials with 30° (dotted lines) 0° (solid lines and −30°, (dashed lines) surface angle (modified with permission from Jenmalm, P., and Johansson, R. S., *J. Neurosci.* **17,** 4486–4499, 1997. Copyright © 1997 by the Society of Neuroscience).

observed on the second trial after the change when information about shape has been obtained through tactile sensory signals. These results demonstrate that visual geometric cues can be used to efficiently specify the force coordination for object shape in a feedforward manner. These cues are used to parametrically adapt the finger force coordination to object shape in anticipation of the upcoming force requirements.

When vision of the object is not available, a very different pattern of force output is obtained. On the first trial after the switch to the −30° object, grip force develops initially according to the force requirements in the previous trial. This indicates that memory of the previous surface angle determines the default force coordination in a feedforward manner. However, about 100 msec after the digits contacted the object, the grip force was modified and tuned appropriately for the actual surface angle (see first trial with the −30° in Fig. 5B). This amount of time is required to translate tactile information into motor commands, a process that likely involves supraspinal processing. By the second trial after the switch, the force output is appropriately adapted to the −30° surface angle from the onset of force application. Thus, an internal model related to object shape determines the force coordination in a feedforward fashion and tactile sensory information obtained at initial contact with the object mediates an updating of this model to changes in

object shape. Furthermore, a single trial is enough to update the relevant internal model.

Sensors in the digits are thus used to update the force coordination for object shape when visual cues are unavailable or misleading. When digital sensibility is removed by local anesthesia, leaving neither visual nor somatosensory cues about shape, the adaptation in force output is severely impaired (Fig. 5D). Although grip force and load force still change in parallel, force output is no longer updated following contact. People adapt to the loss of both visual and tactile sensory cues about shape by applying strong grip forces regardless of surface angle. When vision is available during digital anesthesia, people are able to adapt their forces to object shape with only minor impairments (Fig. 5C). Thus, visual geometric cues can be used effectively for feedforward control even in the absence of somatosensory cues about shape.

The curvature of the grasp surfaces is another aspect of object shape. Surprisingly, the curvature of spherically curved symmetrical grasp surfaces has little effect on grip force requirements for grasp stability under linear force loads. However, it becomes acute in tasks involving torsional loads. The relationship between the grip force and tangential torque is parametrically scaled by surface curvature: For a given torque load, people increase grip force when curvature increases. As with linear force loads, this scaling of grip force is directly related to the minimum grip force required to prevent slip. Under torsional loads, people maintain a small but adequate safety margin against rotational slip. As with surface angle, visual information about surface curvature can be used for feedforward control of force. Likewise, people use cues provided by tactile afferents to adapt force once finger contact is established.

2. Prediction Based on Object Weight

When we manipulate familiar or common objects that we can identify either visually or haptically, we are extremely adept at selecting fingertip forces that are appropriately scaled to the weight of the object. That is, during the very first lift of a common object, before sensory information related to weight becomes available at liftoff, the force development is tailored to the weight of the object. This indicates that we can use visual and haptic cues to select internal models that we have acquired for familiar objects and can use these models to parametrically adjust our force output to object weight. For ''families'' of familiar objects that vary in size (e.g., screwdrivers, cups, soda cans, and loafs of bread), we can exploit size–weight associations, in addition to object identity, to scale our force output in a feedforward fashion. However, as we have all experienced, our force output may sometimes be erroneous. Such situations can be created experimentally by unexpectedly changing the weight of a repeatedly lifted object without changing its visual appearance. In such cases, the lifting movement may be either jerky or slow. For example, if the object is lighter than expected from previous lifting trials, the load force and grip force drives will be too strong when the load force overcomes the force of gravity and liftoff takes place. Although somatosensory afferent events, evoked by the unexpectedly early liftoff, trigger an abrupt termination of the force drive, this occurs too late (due to control loop delays) to avoid an excessively high lift. Burst responses in FA II (Pacinian) afferents, which show an exquisite sensitivity to mechanical transients, most quickly and reliably signal the moment of liftoff. Conversely, if the object is heavier than expected, people will initially increase load force to a level that is not sufficient to produce liftoff and no sensory event will be evoked to confirm liftoff (Fig. 6A, solid curves). Importantly, this *absence* of a sensory event at the expected liftoff causes the release of a new set of motor commands. These generate a slow, discontinuous force increase until terminated by a neural event at the true liftoff (Fig. 6A, afferent response during the 800-g lift following the 400-g lift). Taken together, these observations indicate that control actions are taken as soon there is a mismatch between an expected sensory event and the actual sensory input. Thus, the absence of an expected sensory event may be as efficient as the occurrence of an unexpected sensory event in triggering compensatory motor commands. Moreover, this mismatch theory implies that somatosensory signals that represent the moment of liftoff are mandatory for the control of the force output whether or not the weight of the object is correctly anticipated. Finally, once an error occurs, the internal model of the object is updated to capture the new weight. In natural situations, this generally occurs in a single trial. As shown in Fig. 6A, in the trial after the switch trials when the weight of the object was unexpectedly increased from 400 to 800 g, the forces were correctly scaled for the greater weight (dashed curves).

3. Prediction Based on Friction

Whereas people use visual information about object size and shape to scale fingertip forces, there is no

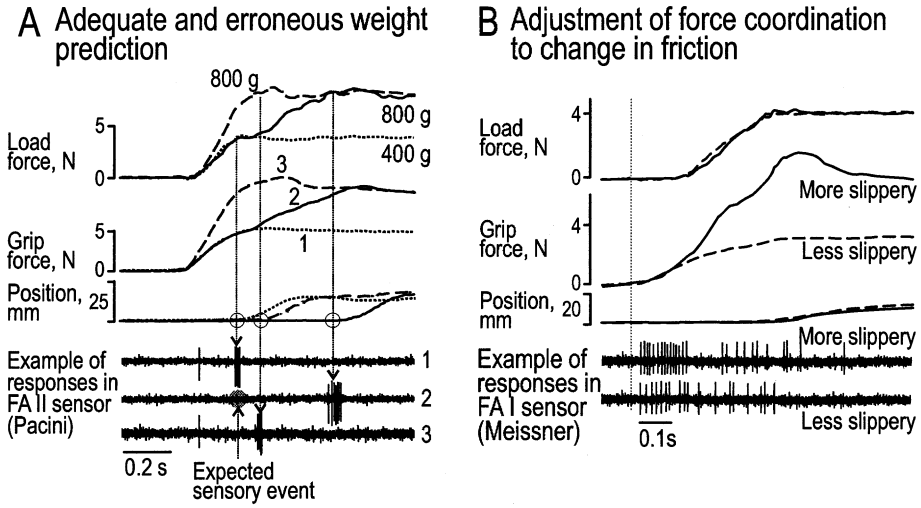


Figure 6 Single unit tactile afferent responses and adjustments in force to changes to object weight (A) and to the frictional condition between the object and the digits (B). Data are from single lift trials. (A) Three successive trials in which the subject lifted a 400-g object (dotted curves), an 800-g object (solid curves), and then the 800-g object again (dashed curves). The forces exerted in the first lift are adequately programmed because the subject had previously lifted the 400-g object. The forces are erroneously programmed in first lift of the 800-g object because they are tailored for the lighter 400-g object lifted in the previous trial. The vertical lines with arrowheads pointing downward indicate the moment of liftoff for each trial and they indicate the evoked sensory events exemplified by signals in a single FA II afferent. The absence of burst responses in FA II afferents at the expected point in time for the erroneously programmed 800-g trial is used to initiate a new control mode. This involves slow, discontinuous, and parallel increases in grip force and load force until terminated by sensory input signaling liftoff. (B) The influence of friction on force output and initial contact responses in a FA I unit. Two trials are superimposed, one with less slippery sandpaper (dashed lines) and a subsequent trial with more slippery silk (solid lines). The sandpaper trial was preceded by a trial with sandpaper and therefore the force coordination is initially set for the higher friction. The vertical line indicates initial touch (modified with permission from Johansson, R. S., and Westling, G., *Exp. Brain Res.* **66,** 141–154, 1987. Copyright © 1987 by Springer-Verlag; and from *Curr. Opin. Neurobiol.* Johansson, R. S., and Cole, K. J., **2,** 815–823, Copyright © 1992, with permission from Elsevier Science).

evidence that they use visual cues to control the balance of grip and load force for friction. However, tactile receptors in the fingertips are of crucial importance. The most important adjustment after a change in friction takes place shortly after the initial contact with the object and can be observed about 100 msec after contact (Fig. 6B). Prior to this force adjustment, there are burst responses in tactile afferents of different types but most reliably in the population of FA I (Meissner) afferents. The initial contact responses in subpopulations of excited FA I afferents are markedly influenced by the surface material as exemplified in Fig. 6B with a single afferent. The adjustment of force coordination to a change in frictional condition is based on the detection of a mismatch between the actual and an expected sensory event. This adjustment involves either an increase in the grip-to-load force ratio if the surface is more slippery than expected (as shown in Fig. 6B) or a decrease in the ratio of the surface if less slippery than expected. The adjustment also includes an updating of the internal model so as to capture the new frictional conditions between the object and the skin for predictive control of the grip-to-load force ratio in further interactions with the object. However, sometimes these initial adjustments to frictional changes are inadequate and an accidental slip occurs at a later point, often at one digit only. Burst responses in dynamically sensitive tactile afferents to such slip events promptly trigger an automatic upgrading of the grip-to-load force ratio to a higher maintained level. This restores the grip force safety margin during subsequent manipulation by updating the internal model controlling the balance between grip and load force.

In summary, skilled manipulation involves two major types of control processes: *anticipatory parameter control* and *discrete event, sensory-driven control.* Anticipatory parameter control refers to the use of visual and somatosensory inputs, in conjunction with internal models, to tailor finger tip forces for the properties of the object to be manipulated prior to the execution of the motor commands. For familiar objects, visual and haptic information can be used to

identify and select the appropriate internal model that is used to parametrically adapt motor commands, prior to their execution, in anticipation of the upcoming force requirements. People may also use geometric information (e.g., size and shape) for anticipatory control, relying on internal forward models capturing relationships between geometry and force requirements. There is ample evidence that the motor system makes use of internal models of limb mechanics, environmental objects, and task properties to adapt motor commands.

Discrete event, sensory-driven control refers to the use of somatosensory information to acquire, maintain, and update internal models related to object properties. This type of control is based on the comparison of actual somatosensory inflow and the predicted somatosensory inflow—an internal sensory signal referred to as corollary discharge. (The somatosensory input provided by tactile signals in the digital nerves is obviously critical in the control of skillful manipulation.) Thus, when we lift an object, we generate both efferent motor commands to accomplish the task and this internal sensory signal. Together, these are referred to as the sensorimotor program. Predicted sensory outcomes are produced by an internal forward model in conjunction with a copy of the motor command (referred to as an efference copy). Disturbances in task execution due to erroneous parameter specification of the sensorimotor program give rise to a mismatch between predicted and actual sensory input. For example, discrete somatosensory events may occur when not expected or may not occur when they are expected (Fig. 6A). Detection of such a mismatch triggers preprogrammed patterns of corrective responses along with an updating of the relevant internal models used to predict sensory events and estimate the motor commands required. This updating typically takes place within a single trial. With respect to friction and aspects of object shape, the updating primarily occurs during the initial contact with the object. In trials erroneously programmed for object weight and mass distribution, the updating takes place when the object starts to move (e.g., at liftoff in a lifting task).

III. ONTOGENETIC DEVELOPMENT OF SENSORIMOTOR CONTROL IN MANIPULATION

The ability to grasp using a precision grip involving the tips of the thumb and index finger first emerges in humans at approximately 8–10 months of age. However, fully mature patterns of grasping, lifting, and holding objects are not observed until about 8 years of age. During this period, there is gradual improvement in grasping behavior as well as qualitative improvements in the capacity to produce independent finger movements. These changes parallel the gradual maturation of the ascending and descending neural pathways that link the hand with the cerebral cortex. These observations strongly suggest that the control of the skilled precision lifting and manipulation relies to a large extent on cerebral processes.

As noted previously, when adults lift objects, they increase grip force and load force in phase such that the two forces increase and decrease together. As a consequence, a linear relationship between these forces is observed (Figs. 3B, 3C, and 7B). The motor system adapts the slope of this relationship to factors such as the frictional conditions and the shape of the contact surfaces but robustly maintains this force synergy (Figs. 3B and 3C). However, before 18 months of age, children do not exhibit such parallel control of grip and load forces (Fig. 7). Instead, they tend to increase grip force in advance of the load force in a sequential fashion. The transition from sequential force coordination to the mature parallel coordination is not completed until several years later. Young children also produce comparably slow increases in fingertip force before liftoff and these increases are discontinuous, featuring multiple peaks in force rate (Fig. 7A). In contrast, adults smoothly increase grip force and load force with a single peak in force rate. The discontinuous or start-and-stop force increases observed in young children suggest that they employ a feedback control strategy rather than feedforward control. That is, they continue to increase force in small increments until liftoff occurs. It is not until they receive somatosensory information that liftoff has occurred that they stop these increases. This feedback strategy is similar to that observed when adults underestimate the weight of an object and then have to increase force again until liftoff occurs (Fig. 6B, solid lines). These observations suggest that young children may not have the cognitive resources for accurate feedforward control.

In addition, very young children appear to be relatively inefficient at integrating sensory information into sensorimotor programs. In precision lifting, people start to increase grip force and load force soon after the digits contact the object. Signals from tactile afferents related to object contact trigger the next phase of the lift. In very young children, there is a

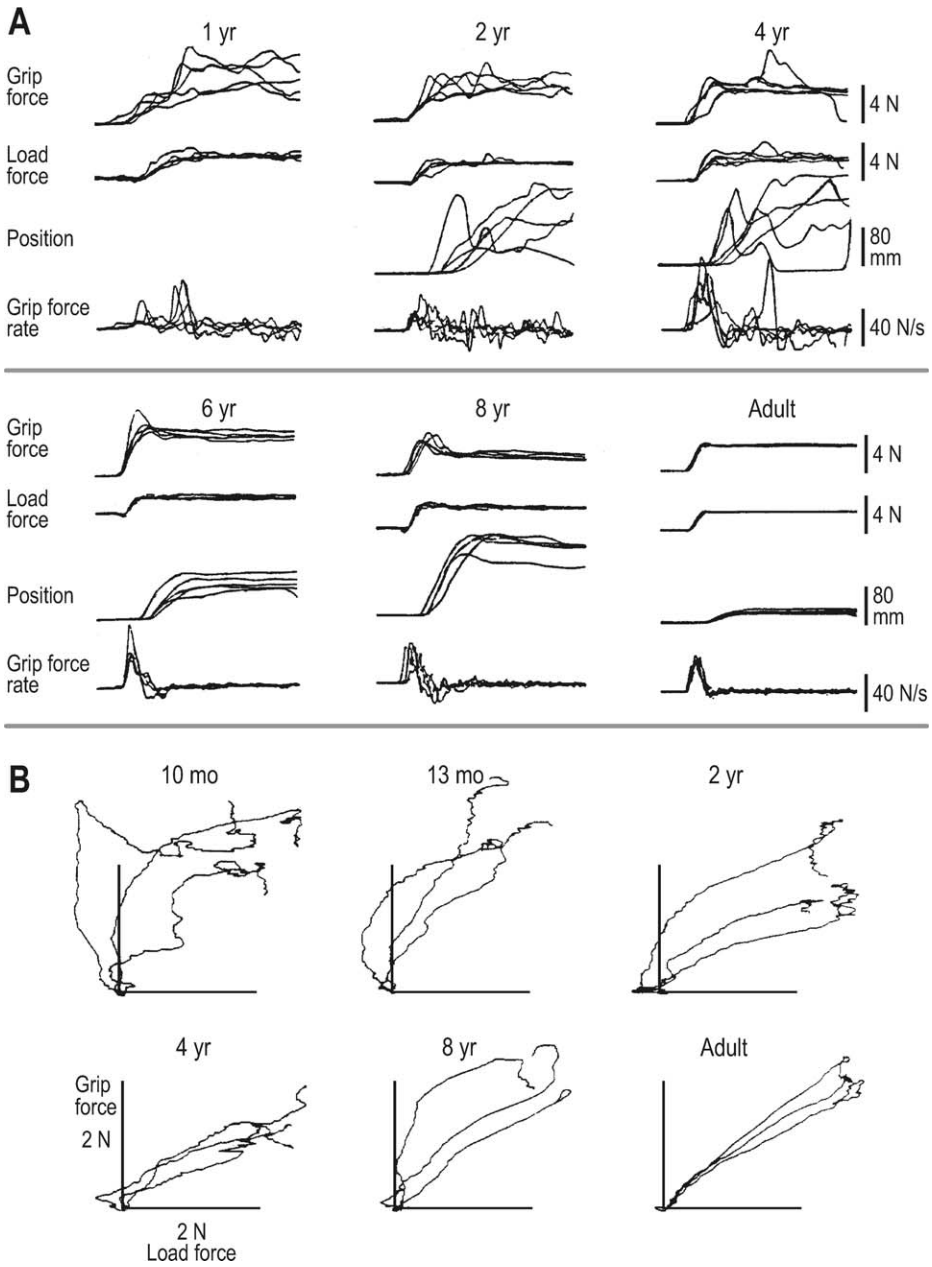


Figure 7 Ontogenetic development of the coordination of grip and load forces during precision lifting. (A) Grip force, load force, and grip force rate as a function of time during several consecutive trials (superimposed) for individual children of various ages and an adult. Note the large variability and excessive grip forces used by young children compared to the adults. (B) Relationship between grip force and load force during the initial parts of lifting trials by children of various ages and an adult. Note the nonparallel increase in grip and load forces for young children compared to adults. (A and B) Surface material and object's weight are constant (adapted with permission from Forssberg, H., Eliasson, A. C., Kinoshita, H., Johansson, R. S., and Westling, G., *Exp. Brain Res.* **85,** 451–457, 1991. Copyright © 1991 by Springer-Verlag).

relatively long delay between initial contact and the onset of increases in grip and load force. This long delay indicates immature control of hand closure and inefficient triggering of the motor commands by cutaneous afferents. The decrease in this delay during subsequent years parallels a maturation of cutaneous reflexes of the hand as assessed by electrophysiological methods.

During the latter part of the second year, children begin to use sensorimotor memory, obtained from

previous lifts, for scaling forces in anticipation of object weight. However, adult-like lifting performance with precise control of the load force for smooth object acceleration does not appear until 6–8 years of age. At about 3 years of age, children start to use vision for weight estimation through size–weight associations for classes of related objects. Thus, additional cognitive development is apparently required before the necessary associative size–weight mapping can take place. Unlike adults, once children begin to use visual size cues, they are unable to suppress adequately their influence when the cues are misleading (i.e., in situations in which weight and size do not reliably covary). This observation is consistent with the view that vision has a particularly strong influence on motor coordination in children. Thus, the context-related selective suppression of visual cues appears to require even further cognitive development.

Young children display a limited capacity to adapt the ratio of grip force and load force to frictional conditions. These children use unnecessarily high grip forces in trials with high friction (or low slipperiness) and their behavior is reminiscent of that of adults with impaired digital sensibility. This increased grip force may be a strategy to compensate for immature tactile control of precision grip because overgripping will prevent slips when handling slippery objects. Nevertheless, even the youngest children (1–2 years) show some capacity to adjust grip force to friction if the frictional conditions are kept constant over several consecutive precision grip lifts. The need for repetitive lifts suggests a poor capacity to form sensorimotor memory related to friction and/or to use this memory to control force output. Older children require fewer lifts to update effectively their force coordination to new frictional conditions, and adults require only one lift.

IV. DISSOCIATIONS AND INTERACTIONS BETWEEN PERCEPTION AND ACTION

An important concept in neuroscience is the idea that sensory information is processed in multiple pathways for different uses. For example, in the visual system, there is strong evidence that neural systems that process visual information for use in guiding action are at least partly distinct from neural systems involved in processing visual information for perception and cognitive reasoning. Similarly, there is evidence that sensory information obtained from the hand can have differential effects on action and perception. Here, we discuss evidence for a dissociation between perception and action related to hand movement. However, first we discuss how manipulatory actions can influence perception.

A. Influences of Action on Weight Perception

Because haptic perception of objects generally involves manipulation, the question arises as to whether the perception of particular object properties is influenced by other object properties or by the way in which the object is handled. For example, does the perceived weight of an object depend on the angle of its contact surfaces or the friction between the object and the digits, both of which influence the grip force required to lift the object? Here, one question is whether the grip forces in lifting influence weight perception even though the grip forces are not directly involved in overcoming the force of gravity. For example, does the greater effort required to lift a slippery object give rise to the perception of it being heavier than a less slippery object of the same weight?

More than 150 years ago, Ernst Heinrich Weber observed that the ability to discriminate weight is better when the weights are actively lifted by the hand than when they are supported by a passive hand. This observation suggests that a sense of effort, associated with voluntary muscular exertion, contributes to the perception of weight. Although afferent signals contribute to weight perception, at least under some conditions there is ample evidence that effort, defined as the level of central or efferent drive, contributes to weight perception. The idea is that when we generate motor commands to lift an object, a copy of the commands (efference copy) generates an internal sensation (corollary discharge) that influences perceived weight. The centrally generated sensation is referred to as the sense of effort.

Figure 8A shows the results of an experiment in which people were asked to compare the weights of a reference object and a series of randomly presented test objects of varying weight both heavier and lighter than the weight of the reference. The test objects had the same size and shape as the reference object, and the objects were lifted using a precision grip with the tips of the index fingers on either side. In one condition, the reference object was covered in less slippery sandpaper and the test objects were covered in more slippery satin (Fig. 8, solid circles and solid curve), whereas in a second condition the reference object was covered in satin and the test objects were cover in sandpaper

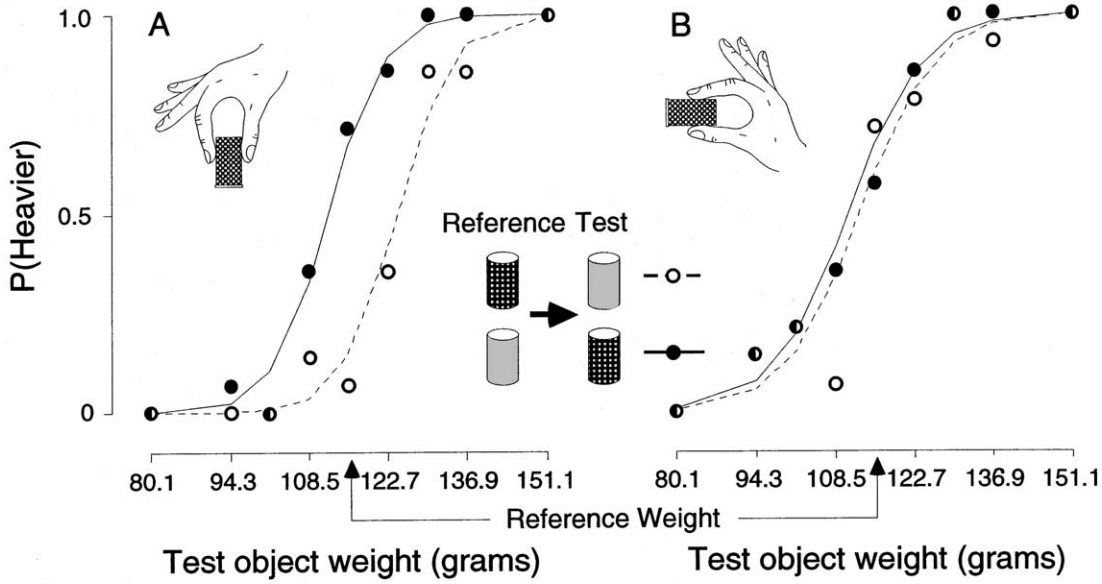


Figure 8 Probability (n=14) of responding that the test canister is lighter than the previously lifted reference canister as a function of the test canister weight. In different experiments, the canisters were lifted with either a vertical (A) or horizontal (B) precision grip. Open circles and dashed lines code the condition in which the test canister was covered in less slippery satin, and the closed circles and solid lines code the condition in which the test canister was covered in less slippery sandpaper. The triangles indicate the reference weight (modified with permission from Flanagan, J. R., Wing, A. M., Allison, S., and Spencely, A., *Perception Psychophys.* **57,** 282–290, 1995).

(Fig. 8, open circles and dashed curve). Figure 8A shows the probability of judging the test object to be heavier than the reference as a function of the weight of the test object. In both conditions, when the test object is much heavier (151.1 g) than the reference (115.6 g) the test object is always judged to be heavier. Conversely, when the test object is much lighter (80.1 g), it is never judged to be heavier. However, in between these extremes, the probability of judging the test object to be heavier is greater when the test object is covered in slippery satin. (Note that there is a general tendency to judge the second of two successively lifted weights, in this case the test object, to be heavier.) This indicates that when lifting with the fingertips on the sides of the object, a more slippery object is judged heavier than an equally weighted object that is less slippery. One interpretation of the results shown in Fig. 8A is that humans judge the more slippery object to be heavier because the grip force used in lifting is greater. When people hold the reference and test objects with a horizontal grip (Fig. 8B), in which surface slipperiness has little influence on the required grip force, there is no effect of surface slipperiness on weight perception.

The results shown in Fig. 8A suggest that people fail to fully distinguish between the effort related to grip force and that related to load force when judging weights lifted with a precision grip. However, this overflow effect may only pertain to muscle actions that are functionally related. Support for this view comes from the observation that the perceived heaviness of a given weight, lifted by one digit, increases if a concurrent weight is lifted by any other digit of the same hand. When the foot or other hand lifts the concurrent weight, the perceived heaviness is not affected.

Although differences in grip force influence weight perception when these differences are determined by frictional conditions, grip force does not appear to influence perceived heaviness when it is manipulated by changing surface shape. When people compare the weights of triangular blocks lifted either on the angled or flat side, there is no effect of angle of perceived weight. It may be that when the grip force requirements strongly match those prescribed by visual cues, people suppress the effort related to grip force differences in evaluating weight. Recall that visual cues related to surface angle can be used effectively for feedforward force control but that there is no evidence that visual information related to frictional condition can be exploited for anticipatory force control.

B. Independent Sensorimotor and Perceptual Predictions of Weight

As discussed previously, people use visual information about object size and shape to estimate parametrically the impending force requirements in manipulation. Thus, people will increase grip and load force more rapidly when lifting a large object than a similar looking small object. This feedforward strategy takes advantage of the link between size and weight that normally pertains to a class or family of similar objects; for example, big cups should weigh more than small ones. However, it fails when this link is altered. In such a case, people must rely on reactive control mechanisms to correct for their erroneous prediction and on feedback mechanisms to tune the internal models used for predictive control. Such a situation arises in the classic size–weight illusion in which people are asked to compare the weights of two equally weighted objects of similar form but unequal size. This illusion, first documented more than 100 years ago, refers to the fact that people reliably judge the smaller of the two objects to be heavier when lifted, even after many lifting trials.

A leading theory of the size–weight illusion is that the illusion arises from a mismatch between predicted and actual sensory feedback. The idea is that when we lift the smaller object, the actual sensory feedback about liftoff will not occur when predicted and the object will thus be judged heavier. Conversely, the larger object, which is lighter than expected, will be judged heavier.

The sensory mismatch seems entirely plausible when one considers lifting the two equally weighting objects the very first time. Here, visual size cues will be misleading and we would expect people to use too much force for the larger object and too little force for the smaller object. However, we also know that people acquire sensorimotor memory related to object weight over repeated lifts. The question arises whether people will continue to misjudge the force required when repeatedly lifting large and small objects of equal weight. Figure 9 reveals the answer. People were asked to repeatedly lift a small and a large cube (Fig. 9A) in alternation. Predictably, when the two objects are lifted for the first time, the forces required for the large object are overestimated and the forces required for the small object are underestimated (Fig. 9B, left). Compensatory, reflex-mediated adjustments in force are triggered in either case. When lifting the small object, the initial increase in grip force and load force is too small and liftoff does not occur when expected. As a result, the forces increase again until liftoff is achieved. When lifting the large object, overshoots occur in the grip and load forces and liftoff occurs earlier than expected. The unexpected early liftoff

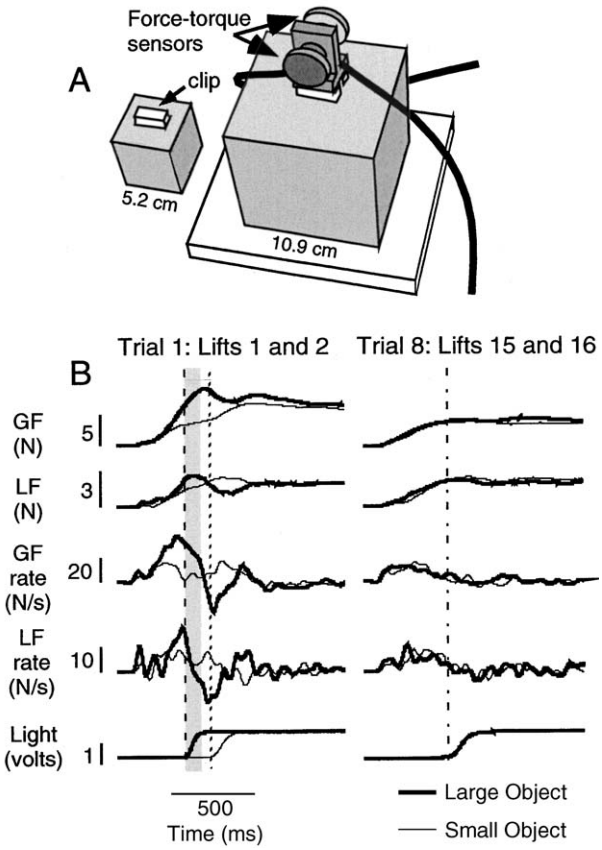


Figure 9 Independent sensorimotor and perceptual predictions of weight. (A) Drawing showing the relative sizes of two equally weighted cubes. Subjects lifted the cubes using a precision grip with the tips of the index finger and thumb on either side of a handle. The handle was attached by clips located on top and in the center of each object. The handle was instrumented with two sensors that measure the forces and torques applied by each digit. Plastic contact disks (3 cm in diameter) were mounted on each sensor and covered in medium-grain sandpaper. A light-sensitive diode embedded into the center of the lifting platform recorded liftoff. (B) Grip force (GF), load force (LF), grip and load force rates, and light-sensitive diode recorded in the first trial (lifts 1 and 2) and the eighth trial (lifts 15 and 16). The subjects lifted the large object (thick traces) and then the small object (thin traces) in each trial. In all trials, subjects grasped the object and increased grip and load force together until liftoff, signaled by the light diode, occurred. In the first trial, peak grip and load force rates were scaled to object size, whereas by the eighth trial the peak force rates were similar for the two objects and appropriately scaled to object weight. Although the subjects adapted their motor output to the true object weights, they still reported verbally that the small object was heavier (adapted with permission from Flanagan, J. R., and Beltzner, M. A., *Nature Neurosci.* **3,** 737–741, 2000).

triggers a decrease in force approximately 100 msec later. However, a very different pattern of force output is observed by the time the cubes are lifted for the eighth time (Fig. 9B, right). Now the force and force rate functions for the small and large cubes are very similar and liftoff occurs at about the same time for both cubes. In contrast to the initial lift trials, grip and load force neither overshoot nor undershoot their final levels, and no corrective adjustments in force are observed. These results illustrate that people adapted their force output, and thus their sensory predictions used for force control, to the actual object weights. Thus, sensorimotor memory about object weight, obtained from previous lifts and based on somatosensory information, comes to dominate visual size cues in terms of feedforward force control.

Although the motor system gradually adapts force output to the true, equal weights of the size–weight stimuli, the perceptual system that mediates awareness of object weight does not adapt. After lifting the two cubes 20 times each, people still reported that the small object was heavier. Moreover, the strength of the size–weight illusion—measured using magnitude estimation techniques—is equally strong. That people experience the size–weight illusion while accurately predicting the fingertip forces required for lifting clearly debunks the theory that the perceptual illusion is accounted for by a sensory mismatch. Instead, the results indicate that the illusion can be caused by high-level cognitive factors. Although the size–weight illusion occurs while there is no evidence of mismatch at the sensorimotor level, the mismatch theory may still operate at a purely perceptual level. For example, people may continue to make erroneous *perceptual* predictions about weight based specifically on visual size cues. A mismatch between these perceptual predictions and actual sensory feedback may give rise to the size-weight illusion. This implies separate comparison processes for perceptual and sensorimotor predictions.

The finding that people continue to experience the size–weight illusion even though they learn to make accurate sensorimotor predictions about object weight indicates that sensorimotor systems can operate independently of perceptual systems. This idea is supported by a growing body of research on visuomotor control showing that partly distinct neural pathways are used depending on whether the sensory information is used to control actions or make perceptual judgments.

See Also the Following Articles

LEFT-HANDEDNESS • MOTION PROCESSING • MOTOR CONTROL • MOTOR SKILL • NEUROFEEDBACK • OBJECT PERCEPTION • SPATIAL VISION • TACTILE PERCEPTION • VISUAL AND AUDITORY INTEGRATION

Suggested Reading

Flanagan, J. R., and Beltzner, M. A. (2000). Independence of perceptual and sensorimotor predictions in the size–weight illusion. *Nature Neurosci.* **3,** 737–741.

Johansson, R. S. (1998). Sensory input and control of grip. In *Sensory Guidance of Movement. Novartis Foundation Symposium, 218* (pp. 45–59). Wiley, Chichester, UK.

Jones, L. A. (1986). Perception of force and weight: Theory and research. *Psychol. Bull.* **100,** 29–42.

Lemon, R. N. (1993). The G. L. Brown Prize lecture. Cortical control of the primate hand. *Exp. Physiol.* **78,** 263–301.

MacKenzie, C. L., and Iberall, T. (1994). *The Grasping Hand*. North-Holland, Amsterdam.

Napier, J. R. (1980). *Hands*. Allen & Unwin, London.

Porter, R., and Lemon, R. N. (1993). *Corticospinal Function and Voluntary Movement*. Oxford Univ. Press, Oxford.

Wing, A. M., Haggard, P., and Flanagan, J. R. (Eds.) (1996). *Hand and Brain: Neurophysiology and Psychology of Hand Movement*. Academic Press, San Diego.

Headaches

SEYMOUR DIAMOND and GEORGE J. URBAN
Diamond Headache Clinic and Finch University of Health Sciences, Chicago, Illinois

GLOSSARY

aura A complex of focal neurological symptoms mostly in a visual form preceding some migraine attacks.

basilar Originating from the basilar artery at the base of the skull, supplying the brain stem.

biofeedback A behavioral treatment, relaxation technique that involves displaying, on monitors, some physiologic functions with the goal of attaining voluntary control over them.

hemiparesis Weakness of one side of the body.

hemiplegic Related to total paralysis of one side of the body.

ophthalmoplegia Weakness of one or more ocular nerves leading to double vision.

osmophobia Sensitivity to odors.

phonophobia An increased sensitivity to sound.

photophobia An increased sensitivity to light.

postdrome A set of symptoms occurring following remission of the headache.

prodrome A set of symptoms starting as early as 48 hr before the onset of a headache.

rebound headache A headache caused by or worsened by the withdrawal of analgesics, ergotamine, or caffeine.

status migrainosus Migraine attack lasting more than 72 hr.

Headache is defined by *Webster's* as "a pain in the head" and "a vexation or baffling situation or problem." Attempting to understand the source of headaches, and efforts at selecting appropriate treatment, is indeed a vexing problem. Headache is one of the oldest and the most common medical complaints, with references to it as early as 4000 BC. In the United States, the prevalence of headaches is about 70%. Headache has an enormous socioeconomic impact on society and personal lives. Headache can be a primary disorder or a symptom of another disease of benign or malignant origin, and it can occur at any age. The most simple classification divides headache types into three major categories; vascular (migraine and cluster), tension type, and traction and inflammatory (organic). Every major type of headache is defined, and the clinical picture, pathophysiology, and therapy are described.

I. VASCULAR HEADACHES

A. Migraine

By definition, migraine is an idiopathic, recurring headache disorder manifesting in attacks lasting 4–72 hr (untreated or unsuccessfully treated), usually unilateral location, of pulsating quality, and of moderate to severe intensity that may inhibit or prohibit daily activities. Pain is aggravated by routine physical activity and is associated with nausea and/or vomiting, photophobia, and phonophobia. History, physical, and neurological examinations do not suggest a secondary headache due to other disorders.

Migraine is an inherited neurological condition. A parental history can be obtained in 50–60% of patients with migraine. The form of inheritance has not been identified, although some genetic studies suggest an autosomal-dominant type in familial hemiplegic migraine. For the more frequent types of migraine, genetic influence is less clear.

Migraine has a marked impact on the economy and society. Surveys show that 8% of men and 14% of women miss all or part of a day of work or school in any given month. In the United States, the annual cost of lost productivity due to migraine has been estimated between $5.6 and $17.2 billion.

1. Clinical Features

Typically, there are five stages of a migraine attack: the prodrome, the aura, the headache, resolution, and the postdrome. The two major types of migraine are distinguished by the occurrence of an aura, a set of warning symptoms, that can precede the headache by 10–60 min. Migraine with aura was previously known as classical migraine, and migraine without aura was common migraine.

The aura is a complex set of focal neurological symptoms that precedes or accompanies an attack and may occur in almost 40% of patients, although not necessarily before each acute migraine headache. The frequency of aura may vary. The most common and most recognizable aura is visual occurring in more than 90% of patients with aura. There are three groups of visual aura: positive—fortification spectra or zigzag lines, flashing of lights and colors, stars, circles, and angles mostly appearing in both visual fields and migrating across the visual field at the rate of 3 mm per minute; negative—scotoma or blind spots, transient loss of vision, grayout, whiteout, heat waves, or the opaque glass-like experience; and metamorphopsia —illusion of distorted size, shape, and location of fixed object and Alice in Wonderland syndrome consisting of complex visual hallucinations. Other auras may consist of sensory phenomena (occurring in 30% of patients), such as numbness in the extremities and/or face, which may be fixed or may migrate. Motor disturbances (6%) and difficulty speaking or understanding language (18%) can also occur. The headache following the aura has rapid onset and graduation, is more lateralized, and has the same quality as the migraine without aura.

During the prodrome phase, patients with migraine with aura or migraine without aura may experience vague symptoms. These forerunners may precede the attack by up to 48 hr and gradually increase in intensity up to the headache onset. These signs are usually ambiguous and consist of mental, neurological, or generalized constitutional symptoms. Mental symptoms include euphoria, depression, restlessness, irritability, mental slowness, fatigue, drowsiness, and hyperactivity. Neurological symptoms may include hypersensitivity to light, sound, and smells. Constitutional symptoms may present with a feeling of cold, sluggishness, thirst, increased urination, fluid retention, decreased appetite or food craving, diarrhea or constipation, or equivocal symptoms of non-well-being.

The onset of headache is usually gradual. The pain is unilateral in 60% and on both sides in 40% of cases. The pain is localized most commonly in the temple, but it may radiate to the upper or posterior part of the head, including the neck and shoulders. Pain may travel from one part to the other or from one side to the other, or it may become generalized. Scalp tenderness may also develop in about two-thirds of patients. The headache usually last from 4 to 72 hr; in children, the headache may be briefer. The pain peaks and then subsides, but in some patients it may plateau for longer intervals. The pain varies in intensity, from annoying to debilitating, and is described as throbbing, pulsating, deep-seated pressure or aching and can be aggravated by routine physical activity such as walking or moving the head. During an attack, patients prefer to remain still in a dark and quiet room.

Accompanying symptoms are very common and are one of the criteria for migraine. The most common associated symptoms are gastrointestinal disturbances, including anorexia, nausea in more than 80%, vomiting in more than 50%, and diarrhea in 16%. Stomach motility is reduced as well as the absorption from the stomach including medications. Neurophysiological accompaniments include photophobia in 82%, phonophobia, blurred vision, osmophobia, lightheadedness in more than 70%, vertigo in one-third of patients; short-lasting numbness and weakness, and mood and mental changes. Visible distension of veins and arteries on the forehead and temples is not unusual. Fluid retention with edema and weight gain may occur in a weight-dependent part of the body, such as the feet, ankles, legs, face, and around the eyes. Increased urination may occurs during and after the headache.

Vomiting and sleeping may be the culmination of the migraine attack leading to the recovery phase. Following the headache, during the postdrome, patients may feel tired, irritable, exhausted, and washed out, may describe hangover-like symptoms, and may experience impaired concentration, muscle aching, and anorexia.

About one-third of migraineurs report severe disability with their attacks. It is estimated that in the United States approximately 30 million individuals experience severe migraine headache. Twenty-five

percent of migraine sufferers experience four or more severe attacks per month, 35% experience one to three severe attacks per month, and the remaining migraine sufferers complain of one or less severe attacks per month.

2. Diagnostic Features

Initial onset of migraine usually occurs during adolescence or the early twenties, but it can start in childhood. In a classic study from Sweden, headaches of all types increase in a stepwise manner between ages 7 and 15. By age 15, 5% of all children have experienced a migraine, and another 15% complain of daily or almost daily tension-type headaches. In a follow-up completed 40 years later, in girls from age 11 years, there was a gradual increase in migraine headache. The original subgroup of 73 children with migraine reported being headache-free by age 25, with a more significant number of males reporting complete migraine remission. The data from an epidemiological study in Washington County, Maryland, revealed that 56% of boys and 74% of girls between the ages of 12 and 17 complained of a headache during the past month. Migraine prevalence (i.e., the proportion of the population with the disease) increases from 12 to 38 years of age in both females and males and is highest between ages 25 and 55 years. The lifetime prevalence of migraine ranges between 14 and 18%. However, many migraine sufferers remain undiagnosed. The highest incidence of new cases of migraine peaks between ages 12 and 17. In older patients, the prodromal aura symptoms may continue after the headaches disappear. Initial onset of migraine is rarely reported after age 50 years.

In children under the age of 14 years, no significant difference in migraine prevalence has been demonstrated. The female-to-male ratio is suggested to be between 2:1 and 3:1. This proportion peaks at about age 42, but thereafter it declines.

Migraine can start at any time of the day, and in many cases the headache will be present upon awakening. In some patients, the acute migraine attack will occur after a stressful episode or on weekends or vacations—the so-called ''let down'' headaches. Migraine patients may be hyperresponsive to various internal and external stimuli that, under certain conditions, can provoke a migraine attack. Environmental precipitants include weather changes (in 50% of migraineurs), high humidity, glare, bright light, fluorescent lighting, high altitude, cigarette smoke, and pungent odors. Nitrites, monosodium glutamate, tyramine, aspartame, histamine, and other chemicals contained in certain foods or beverages are recognized migraine precipitants. Alcohol, chocolate, aged cheeses, pickled, cured and fermented products, caffeinated drinks, and nuts are frequently named offenders for susceptible migraine sufferers. Too much or lack of sleep as well as jet-lag, alternating shift work, and physical exertion can provoke migraine.

Emotional stress of positive or negative character, anxiety, anticipation, anger, fear, and worry are common precipitants. Menstruation or other hormonal changes, including hormonal treatment (oral contraceptives or hormone replacement therapy), play a major role in migraine in 60% of female migraineurs. Some medications, including nitroglycerine, blood pressure agents, antidepressants, and arthritis agents, as well as vitamins and herbal remedies in excess, can provoke an acute migraine attack.

3. Pathophysiology

The mechanism of the attack is still poorly understood. Three basic pathophysiological changes take place during the process of migraine: vascular, neurogenic, and neuropeptides abnormalities or dysfunction. It is unclear which process is primary and which is secondary. In the brain stem of migraine patients, there is evidence of a dysfunctional area of neurons, a so-called ''central migraine generator,'' with abnormal reactivity to varying internal or external stimuli. This dysfunction is a result of genetic abnormalities and leads to a cascade of pathophysiological alterations producing the migraine attack. Initially, vasoconstriction occurs in the blood vessels of the brain, during the prodromal or aura stage. When a stimulus crosses a pain threshold, which differs in every individual and is determined by a complexity of physiological and emotional conditions, the migraine generator sends signals that lead to dilation of certain blood vessels in the brain. Migraine pain is thought to arise when intracranial blood vessels become dilated and as a consequence activate the terminals of trigeminal sensory nerves that surround them. Activation of those terminals leads to release of vasoactive neuropeptides, such as calcitonin gene-related peptide, substance P, bradykinin, prostaglandin, and other neurokinins. Vasoactive neuropeptides exacerbate blood vessel swelling and increase pain transmission back to the brain stem. From there, the signals are relayed to higher cortical centers, where the migraine pain is registered and other symptoms are produced, including nausea, vomiting, and photo- and

phonophobia. Serotonin and serotonin receptors play a major role in this cascade and mediate many of these processes.

4. Therapy

Migraine therapy can be classified into four types: general measures, abortive therapy, pain relief measures, and prophylactic (preventive) therapy. Overall, the treatment consists of behavioral and pharmacological interventions and is individualized for each patient. General measures includes behavioral therapy, such as recognition and avoidance of migraine triggers, regulation of sleep and meal schedules, dietary modification, stress management, exercise, and biofeedback and other relaxation techniques.

The other three types of migraine treatment include pharmacological agents. Abortive medications used to stop the process of migraine in its development include the triptans, ergotamine preparations, and isometheptene. These drugs act on serotonin receptor sites, reducing activation of neuropeptides as well as constricting dilated intracranial blood vessels.

If abortive therapy is unsuccessful, the pain relief measures are indicated to reduce the symptoms of migraine, such as pain, nausea, vomiting, irritability, and anxiety. The agents used for symptomatic treatment include single or combined analgesics, antinauseants, nonsteroidal antiinflammatory agents (NSAIDs), sedatives, and tranquilizers.

Prophylactic treatment is instituted to reduce the frequency, duration, and intensity of migraine attacks. It may be considered if the patient is experiencing more than two migraine attacks per month or if the severity of the headache and its associated symptoms impact on the individual's ability to function. Beta blockers, methysergide, and divalproex sodium have been approved for the indication of migraine prophylactic therapy. Other medications that have been beneficial in migraine treatment include the calcium channel blockers, anticonvulsants, NSAIDs, and antidepressants.

5. Other Forms of Migraine

Migraine with prolonged aura, perviously termed complicated or hemiplegic migraine, is characterized by one or more aura symptoms lasting more than 60 min and less than 1 week. Any of the various forms of aura may occur. This type of migraine is relatively rare.

The headache usually starts within 1 hr of aura onset, becomes progressively more intense, and may linger for a prolonged period. Different forms of aura can be experienced at the same time. The intensity of the pain is usually less than that in the more common types of migraines. The aura persists into the headache stage and may continue after the pain subsides. The typical clinical features of the headache are the same with this complicated form of migraine.

The etiology or mechanism of migraine with prolonged aura and related symptoms is unclear. It has been assumed that a neurological deficit of longer duration is caused by prolonged vasoconstriction or limited spasm of a cerebral artery occurring as a part of the migraine syndrome. It appears that the constriction is probably secondary as the part of a complex cascade of migraine process. Computed tomography (CT) scan and magnetic resonance imaging performed during the attack of a migraine with prolonged aura show no abnormalities.

In patients with this form of migraine, risk factors such as smoking and estrogen use should be avoided. Hypertension and diabetes should be treated vigorously. Vasodilatory medications, such as calcium channel blockers mainly of a rapid onset of action (e.g., sublingual nifedipine) and sublingual nitroglycerine, have been shown to be effective in prompt alleviation of prolonged aura or neurological symptoms. Similar effects have been demonstrated with isoproterenol, administration of CO_2, and papaverine. These vasodilators are known to cause headache. However, interestingly, if they are used at the onset of an attack, the headache does not occur or is lessened. The use of ergotamine and triptans is controversial and generally not recommended due to the increased risk of complications such as stroke.

a. Hemiplegic Migraine Hemiplegic attacks usually start during childhood. The headache may occasionally precede the hemiparesis or be absent. The onset of hemiparesis may be abrupt, lasting for days, and may imitate a stroke. The headache is on same side as the weakness in about 20% of patients, on the other side in almost 50%, and can be generalized in 30%. Nausea and vomiting are unusual. Speech difficulty occurs in about 44% of patients in association with hemiparesis, and confusion can occur in one-third. The diminished sensitivity of the half of the body or extremities on one side accompanies the weakness. Visual aura and hemiparesis occur together in 88%. In most patients, the weakness continued for up to 1 hr, in 14% it lasted for up to 3 hr, in 12% from 3 to 24 hr, and in 16% of patients between 1 day and 1 week. The longer lasting episodes are associated with more

profound weakness. A higher incidence of neurological symptoms accompanying the migraine attacks has been noticed in middle-aged patients. Hemiplegic migrainous attacks occur infrequently and irregularly. Typically, the patients experience complete recovery.

b. Ophthalmoplegic Migraine In this form of complicated migraine, repeated attacks of headache are associated with paresis of one or more of the ocular cranial nerves in the absence of demonstrable intracranial lesions. It is relatively rare, and its victims are usually children, affecting males six times more often than females. Adults may experience their first attack during the fourth and fifth decade. The most commonly affected nerve is the third cranial nerve (oculomotor), occasionally the fourth and sixth nerves are affected, and the ophthalmic division of the fifth cranial nerve may also be involved. The headache that precedes the ocular symptoms is of migrainous quality, usually localized behind or around the affected orbit with radiation to the same side of the head. The pain may last 1–4 days, and as it subsides other symptoms appear—the ptosis (droopy eyelid) on the same side, diplopia (double vision), and blurred vision. The ophthalmoplegia may persist for several weeks but usually completely resolves. After repeated attacks, some weakness of the ocular muscles may persist with permanently dilated pupil and deviated eye.

The isolated involvement of the cranial ocular nerves suggests that the anatomic localization of the lesion is the cavernous sinus. Research demonstrated narrowing in the internal carotid artery, probably due to edema of the wall. The swelling of the carotid wall is probably responsible for compression of the involved nerve or reduced blood flow to the nerve through the nourishing arteries.

These attacks are treated symptomatically with analgesics and antinauseants. Corticosteroids in large doses, instituted at the beginning of the attack of ophthalmoplegia, may reduce inflammation and edema and hasten the recovery. Ergotamine agents and triptans should not be used during the acute attacks as abortive drugs because it is unknown if the vasoconstricting effect of these medications would not further reduce the blood flow to the affected nerve. Prophylactic drugs are indicated if the attacks occur frequently. Neuroimaging is imperative to exclude an aneurysm or tumor.

c. Familial Hemiplegic Migraine This form of migraine with aura, which may be prolonged, includes some degree of hemiparesis and is characterized by the patient having at least one first-degree relative with identical attacks. The pattern of occurrence in family members suggests that familial hemiplegic migraine is an autosomal-dominant genetic disease. Researchers have found a locus on chromosome 19 in the majority of the investigated families.

d. Basilar Migraine This type of complicated migraine is also known as basilar artery migraine, Bickerstaff's migraine, and syncopal migraine. Its symptoms clearly originate from the brain stem or from both occipital lobes. To fulfill the criteria, two or more aura symptoms must be of the following types: visual symptoms in both the temporal and nasal fields of both eyes, difficulty articulating, vertigo, tinnitus, decreased hearing, double vision, ataxia (difficulty controlling bodily movements), numbness on both sides of body, weakness and numbness on both sides of the body, and decreased level of consciousness.

Originally considered to be mainly a migraine disorder of adolescent girls, it affects all age groups and both sexes. These attacks often occur with relationship to menstruation. The aura commonly lasts less than 1 hr. The neurological symptoms, of various duration, are consistent with dysfunction of the posterior part of brain supplied by the vertebrobasilar artery system. These symptoms usually precede the headache but may also begin during the actual headache. Typically, the numbness is bilateral, starting in the periphery of all four extremities, with slow and gradual radiating up the limbs. Confusion and disorientation are not uncommon. Syncopal episode occurs in 7% of the patients with this form of migraine. Various levels of decreased consciousness, including stupor and coma, can last for hours or days. The headache of basilar migraine is often bilateral and localized in the posterior part of the head, but it may radiate to the temporal and frontal areas. The pain is of a pulsatile quality and moderate in intensity. Associated nausea may become severe and progress to prolonged and intractable vomiting. Commonly, the attack ends with sleep. Typical attacks of basilar migraine infrequently occur in these patients, and more common forms of migraine occur in between the attacks. Basilar migraine is very rare after the fourth decade, and any episode in this circumstance should be considered nonmigraine in origin. Symptoms of basilar migraine are consistent with a disturbance in the vertebrobasilar artery supplying the territory of the brain stem and brain and are probably the result of ischemic changes.

6. Migraine Aura without Headaches

Another term for migraine equivalent is acephalic migraine, which is a migrainous aura unaccompanied by headache. It is quite common in individuals with migraine with aura to experience the aura with the headache absent. As patients age, the headache may lessen in frequency and eventually disappear, even if the aura continues. Recurring symptoms of aura for which no underlying organic cause has been found should be considered a possible manifestation of migraine aura without headache or a migraine equivalent. These symptoms may occur in a person with a strong family history of migraine or who has previously experienced attacks of migraine. Symptoms of aura without headache may occur in 20% of all migraineurs and in more than 40% of patients who have migraine with aura. Symptoms usually last about 20–30 min and only occasionally for longer periods. The most common symptoms are visual. They have a similar presentation as the typical aura. Patients with monocular visual defects should be evaluated to exclude the presence of ocular or other diseases. The recurrent abdominal pain and cyclic vomiting may also represent a migrainous phenomenon but only after organic disease has been excluded. The neurological episodic symptoms may represent this form of migraine and include numbness, weakness, vertigo, confusion, and amnesia lasting 1 hr or less. The sensory symptoms continue as long as a typical aura. These symptoms may slowly spread up one arm from the fingers to the body and descend the opposite arm. This march is slower than the abrupt onset of similar symptoms occurring with a typical transient ischemic attack. A neurogenic and vasoconstricting origin has been implicated.

Treatment is not necessary if symptoms of migraine aura without headache do not occur often and do not cause significant discomfort. With frequent attacks, the usual prophylactic agent such as the beta blockers, calcium channel blockers, and GABA-receptor agonist anticonvulsants (divalproex sodium, gabapentin, and topiramate) may be utilized. Sublingual nifedipine and sublingual nitroglycerine have been shown to be effective in prompt alleviation of prolonged aura. Isoproterenol, administration of CO_2, and papaverine have demonstrated similar efficacy.

a. Migraine with Acute Onset Aura In migraine with acute onset aura, the warning symptoms fully develop in less than 5 min. The subsequent headache has all the features of migraine. Transient ischemic attack and other intracranial lesions should be ruled out by appropriate investigation. The neurological symptoms develop within 5 min and last as a typical aura for 20–60 min. The presence of a typical headache phase is required to establish the diagnosis, and previous migraine attacks of another type or a strong family history of migraine support diagnosis. Extensive investigations are necessary to rule out a transient ischemic attack.

b. Childhood Periodic Syndromes These syndromes may be precursors to, or associated with, migraine. The syndromes are characterized by multiple, repeated brief attacks of neurological symptoms without headaches, or the headaches cannot be detected because of the young age. Abdominal migraine and cyclic vomiting are included in this group.

c. Benign Paroxysmal Vertigo of Childhood This disease is characterized by brief attacks of vertigo in otherwise healthy children. Neurological examination and electroencephalogram are normal. It has been found that in 13% of reported cases of benign paroxysmal vertigo of childhood, the individual subsequently developed migraine. Multiple, short-lasting, sporadic spells of disequilibrium are often associated with nystagmus, nausea, vomiting, and anxiety. In 14%, vertigo is accompanied by headache. The pathophysiology is unknown.

Treatment of an acute attack is symptomatic. Beta-blocking agents, as well as cyproheptadine, may reduce the frequency of attacks.

7. Complications of Migraine

a. Status Migrainosus Previously, this condition was called intractable migraine or persistent (pernicious) migraine. It is distinguished as a migraine attack with the headache phase lasting more than 72 hr despite treatment. A headache-free interval of less than 4 hr may occur. Any episode of migraine, in any form of migraine, may evolve into an intractable, daily, continuous headache attack, unresponsive to standard treatments. The headache may be unilateral or global, pulsatile or pressure-like, or may have characteristics of both migraine and tension-type headaches. The headache progressively intensifies to a debilitating pain, accompanied by the usual characteristics of migraine. Typically, the associated nausea and vomiting are severe, leading to osmophobia, dehydration,

refusal to eat, and prostration. The photophobia, phonophobia, and headache exacerbated by any movement forces the patient to remain in a dark and quiet room, unable to function at even a basic level. Some patients will even wear dark sunglasses indoors because of excessive sensitivity to light. Dehydration and anorexia may cause electrolyte disturbances, further complicating their condition. Emotional despair and depression with suicidal ideation are generally present. Status migrainosus is considered "headache urgency" requiring immediate care, preferably in an inpatient setting for rehydration, pain control, and reversal of continuous headache.

Status migrainosus, is often iatrogenically induced due to overuse or inappropriate use of analgesics, ergotamine preparations, narcotics, caffeine, or triptans or due to inadequate treatment of migraine. In susceptible patients, high stress, anxiety, and poor sleeping and eating habits may lead to this condition. Rebound headaches, transformed migraine, mixed headache, and chronic daily headaches may ultimately cause intractable debilitating headache. Status migrainosus is thought to be due to a sterile inflammation of the intracranial blood vessels involved in migraine process. The vasodilatation and inappropriate release of vasoactive neuropeptides, some of which are very potent vasodilators, as well as other active peptides that mediate neurogenic inflammation are self-perpetuating processes leading to endless activation of the trigeminal neurovascular system and fueling the central migraine generator. The plausible mechanism responsible for the refractoriness of status migrainosus is that the constant activation and release of neuropeptides downregulates serotonergic receptors and depletes endorphins.

Patients with acute status migrainosus may require hospitalization, particularly if the condition was induced by dependency on medication, is accompanied by dehydration, or if the patient is depressed or has a prior experience of adverse reactions to medications (Table I). The offending medication causing rebound headache phenomenon must be withdrawn. The withdrawal is usually done in an abrupt manner, but all precautions to prevent seizures and/or other withdrawal reactions should be instituted. Treatment for patients with status migrainosus should be aggressive and includes rest; rehydration and electrolyte replacement; detoxification; round-the-clock parenteral analgesic therapy; symptomatic treatment of nausea, anxiety, insomnia, and withdrawal symptoms; concurrent initiation of prophylactic therapy; and behavioral treatment. Corticosteroids and NSAIDs are used to reduce the neurogenic inflammation. Phenothiazine-based neuroleptics are utilized to control nausea and vomiting, reduce pain perception, and induce sedation. A series of intravenous dihydroergotamine (DHE) administered every 8 hr for 3 days is very effective in interrupting the painful cycle. It cannot be used when the patient is rebounding from ergotamine preparation or when DHE is contraindicated. Pain control is achieved by scheduled administration of parenteral narcotics, neuroleptics, benzodiazepines, and ketorolac. "As needed" pain control is not recommended because of reinforcement of dependency on analgesics. Appropriate prophylactic therapy, education, psychotherapy, and biofeedback should be concomitantly instituted.

Table I
Criteria for Admission to Inpatient Headache Unit

Prolonged, unrelenting headache with associated symptoms, such as vomiting, that if continued would pose a further threat to the patient welfare
Status migraine
Dependence on analgesics, caffeine, narcotics, barbiturate, and tranquilizers
Habituation to ergots, with rebound headache
Pain accompanied by serious adverse reactions from therapy
Pain occurring in the presence of significant medical disease
Chronic cluster unresponsive to therapy
Treatment requiring copharmacy with drugs that may cause a drug interaction and necessitating careful observation within a hospital environment
Patients with probable organic cause to their headache requiring appropriate consultations, diagnostic testing, and perhaps neurosurgical intervention

b. Migrainous Infarction Occasionally, a migraine attack with one or more migraine aura symptoms is not fully reversible within 7 days and/or is associated with an abnormal neuroimaging test, confirming ischemic brain infarction. To fulfill the criteria, cerebral infarction must occur during the course of a typical migraine attack and other causes of infarction must be ruled out. The prevalence of migrainous infarction is about 3.36 cases in 100,000 adult migraineurs. Some studies indicate that women are more in risk, particularly those who are smokers, on oral contraceptives, and experiencing migraine with regular and prolonged aura. Neurological symptoms may occur abruptly or more slowly, may present as aura preceding the headache, or may start during the

headache phase. Various symptoms and the intensity are dependent on anatomical location of infarction. The stroke must be confirmed by neuroimaging, and other causes of stroke must be ruled out. Prolonged, decreased blood flow during the migraine attack may cause ischemia in a susceptible migraineur. Abnormal blood coagulation, platelet changes, and other unrecognized factors also have a causative role.

Treatment consists of reducing the known risk factors, such as smoking, high cholesterol, and oral contraceptives. Supportive and analgesic treatment is indicated, and ergotamine and triptan preparations are contraindicated. Corticosteroids may reduce concomitant neurogenic inflammation and swelling. Calcium channel blockers facilitate vasodilatation and increase of blood flow in affected areas. Aspirin has been found to reduce risk of further stroke. Rehabilitation should be initiated in early stages.

B. Cluster Headache

This form of vascular headache has been known as histaminic cephalalgia, Horton's headache, migrainous neuralgia, sphenopalatine neuralgia, petrosal neuralgia, red migraine, Raeder's syndrome, Sluder's syndrome, erythromelalgia, and Bing's erythroprosopalgia. The defining characteristic of cluster headaches is their occurrence in cycles (clusters) that occur and disappear spontaneously. There are two forms of cluster headache—episodic and chronic. The majority of patients with cluster headaches experience the episodic form, in which the headache cycles or series last for several weeks or months and then may disappear for years. For those unfortunate few with the chronic form, headache remission is briefer than 14 days, or the cycle of headaches is continuous, without any headache-free intervals.

1. Clinical Features

The head pain and the associated symptoms characterize the acute cluster attack. During a cluster series, the acute headaches usually occur several times per day. The acute episodes are characterized by their brief duration (compared to migraine), usually lasting 15–180 min.

The headache is very severe, localizing at the orbital or supraorbital regions, and the patient may have temporal pain. The pain is strictly unilateral without alternating sides during a series, with a slight predominance of right-sided headaches. The side shift may occur between cycles. The most common localization of pain is around and behind the eye, temporal and frontal lobes, and upper and lower jaw. The pain begins abruptly without any warning or just with a slight "awareness" or mild pain. It peaks in intensity within a few minutes and may last from 5–10 min up to more than 3 hr, but in most cases the average duration is 30–60 min.

The intensity of pain is excruciating, almost unbearable; it is said to be probably the most severe type of head pain known. The pain is described as boring, burning, pulsating, squeezing, deep knife-like pain, stabbing, piercing, or as a combination. In most cases, the frequency of attacks ranges from less than one to three per 24 hours up to six to eight per 24 hours. Attacks may occur at any time but usually happen in clockwork regularity at the same time each day or night. There is a high preponderance of nocturnal attacks, with frequent headaches occurring 2 or 3 hr after retiring to sleep. During an attack, patients tend to be restless, unable to remain still; many pace the floor, constantly moving, rocking, and engaging in bizarre activities and behavior that sometime lead to self-mutilation and even suicide.

Typically, the pain is accompanied by various autonomic phenomena. The most common, in descending order, are lacrimation, redness of the conjunctiva, nasal congestion, nasal discharge, forehead and facial sweating, small pupil, droopy eyelid, and eyelid swelling on the same side of the pain. Nausea and photophobia may present in some patients.

The most common triggers during a series are alcohol, nitroglycerine, and hypoxia. During the cluster cycle, patients will voluntarily avoid alcohol consumption and will not note any headache provocation by alcohol during the remission intervals.

2. Diagnostic Features

Cluster headache is the least prevalent of the primary headache syndromes, with an occurrence of about 1:1000 persons. The migraine-to-cluster ratio ranges from 7:1 to more than 20:1. Cluster headache is predominantly a male headache disorder, with the male-to-female ratio ranging from 4.5:1 to 6.7:1. The mean age of onset is approximately 27–30 years, with the youngest 6–8 years old. The incidence of the disease as well as the episodes of cluster headaches decrease after the age of 60.

The mean duration of a cluster cycle is approximately 6–8 weeks, with only about 3% of patients having bouts of 1-week duration and 4% having cycles

lasting more than 26 weeks. The frequency of periods is less than one per year in 13%, one per year in 40%, two per year in 30%, and three per year in 8%. Cycles of one or two per year occur in 60% of patients. There appears to be a cyclic periodicity to the "clusters," with two seasonal peaks in early spring and early autumn. Rhythmicity in daily occurrence has also been noted, with attacks usually occurring 24 hr apart and frequently presenting at the same time of the day or night. During remission periods attacks spontaneously cease. The most frequently reported remissions last 7–12 months in 48%, 1–6 months in 20%, and 2 or more years in the remaining 32%. During the remission period the cluster attack cannot be provoked.

In the chronic form, there are no cycles, and by definition remission does not occur, nor does it last longer than 14 days. Patients with chronic cluster headaches have a tendency to habituation problems because of the continuous nature of the headache. The chronic form has similar clinical features as the episodic type. Attacks of cluster headache recur for a period of time at least 50 weeks in 1 year. The age of onset seems to be higher, with a mean onset of 39 years. Male preponderance of the chronic forms is at least as significant as that in episodic cluster. The mean frequency of attacks and duration are slightly higher. Pain localization, quality, and associated symptoms are the same as those in the episodic form.

3. Pathophysiology

The mechanism of cluster headache is unclear and involves vascular changes and impaired autonomic neuronal activity as well as biochemical, hormonal, and chronobiological changes. The abnormal autonomic neuronal activity is most likely due to a primary disorder of the central regulation of autonomic functions controlled by the hypothalamus. Several vasoactive substances, including histamine, serotonin, and substance P, have been observed to be activated during the cluster attack. Hormonal level changes, including those of testosterone and melatonin, have also been noted. Circadian and circannual rhythmicity suggest the pathological involvement of the "biological clock" in hypothalamus. Furthermore, hypoxia and the role of the carotid body (the most sensitive chemoreceptor for hypoxia) have been discussed in regard to the pathomechanism of cluster attacks. It appears that the origin of cluster headache is in the hypothalamus, with abnormal activation of the autonomic nervous system causing vascular and biochemical changes.

4. Therapy

Treatment for both the episodic and the chronic forms is similar and consists of abortive and prophylactic therapy. Because of the brief nature of acute cluster attacks, medication used to stop the pain requires rapid onset of action (within 30 min). The most effective and safest method is self-administration of 100% oxygen, via facial mask, at high flow of 8–10 liters per minute for 10–15 min. Dramatic termination of an acute attack has been achieved within 10–15 min in 80% of patients. Other effective abortive therapies include parenteral and intranasal administration of sumatriptan or dihydroergotamine; sublingual ergotamine; and intranasal instillation of viscous lidocaine, a local anesthetic, to the affected nostril. Parenteral ketorolac, chlorpromazine, and narcotic analgesics may help control the pain.

The prophylactic treatment should be initiated at the very onset of a cluster cycle and continued at least 2–4 weeks after the last acute cluster episode. Corticosteroids may hasten the interruption of the cycle. Verapamil (a calcium channel blocker), lithium, methysergide, ergotamine, and divalproex sodium have been found to be effective in reducing and preventing acute cluster attacks during the episodic cycle or the chronic form. In some inpatient headache centers, intractable cluster headaches have been successfully treated with intravenous histamine desensitization treatment.

5. Chronic Paroxysmal Hemicrania

Chronic paroxysmal hemicrania (CPH) is a rare disorder. It has the same characteristics as cluster headache, including similar associated symptoms. These episodes are briefer, more frequent, occur mostly in females, and responsive to indomethacin.

The patient with CPH will characteristically complain of 10–20 brief, intense focal episodes of head pain, localized mostly in the temporal, ocular, frontal, and upper jaw area. The pain has the same quality as cluster headache pain, but of even shorter duration (an average of 10–20 min). CPH attacks are associated with autonomic symptoms and signs that are characteristic for cluster headache. In some patients, head movement or pressure on certain points in the neck can trigger attacks. About 70% of diagnosed patients are female and the mean age of onset is 34 years. The pathogenesis is unknown, but it is considered to be a cluster variant. One of the diagnostic criteria for CPH is absolute responsiveness to indomethacin, an

NSAID. Indomethacin selectively stops the attacks, usually within 2 days of treatment.

6. Cluster Headache Variant

These headache attacks are believed to be a form of cluster headache or CPH but do not meet their criteria. Cluster headache variant, originally described by Diamond and Medina, is a syndrome consisting of a triad of symptoms: atypical cluster headaches, multiple jabs, and background continuous headache. The atypical cluster headache is irregular in location, duration, and frequency, occurring several times a day. Multiple jabs are sharp, variable, painful episodes, lasting only a few seconds and occurring several times a day. Background headaches are chronic, continuous, often unilateral, sharply localized and of variable severity, and have vascular features—throbbing and exacerbated by physical exertion. The pathophysiology is unknown. The therapy consists of indomethacin or lithium.

II. TENSION-TYPE HEADACHE

Tension-type headache was previously known by several terms: muscle contraction headache, stress headache, ordinary headache, essential headache, psychogenic headache, or psychomyogenic headache. It is defined as recurrent episodes of headache lasting minutes to days. There are two primary forms—episodic and chronic. The pain is bilateral, with pressing or tightening (nonpulsating) quality of mild-to-moderate severity. The headache is not aggravated by routine physical activity. Photophobia or phonophobia may be present, and nausea may occur in the chronic form. The episodic type has been experienced by almost everyone, is usually relieved by over-the-counter analgesics, and does not require a physician's intervention. The chronic type is daily or almost daily, and the victim is prone to dependency problems with analgesics, tranquilizers, or sedatives.

A. Diagnostic Features

The prevalence in the general population ranges from 30 to 80%. The 1-year prevalence of episodic tension-type headaches is about 55% in men and 70% in women; in the chronic form the prevalence is only about 2–5%. The prevalence decreases with increasing age.

The pain is usually described as steady, nonpulsatile, band-like, vise-like, tightness, ache, soreness, and pressure in a hat-like distribution. Feelings of increased tension may be also felt in the occipital and cervical areas. Muscles of the scalp and neck may be tender to the touch, with palpable, sharply localized nodules. Combing or brushing the hair or wearing a hat may elicit soreness. Electromyographic recording from the scalp and cervical muscles may or may not register increased activity. In many individuals, the headache starts during the afternoon or evening but rarely impacts on the sufferer's daily activity.

The various psychological factors associated with chronic tension-type headaches include anxiety, fear, hostility, and, very frequently, depression. Typically, in the chronic form, the victim may complain of a sleep disturbance. In chronic tension-type headache due to anxiety, the patient may complain of insomnia (difficulty falling asleep). However, in patients with chronic tension-type headaches due to depression, early or frequent awakening may be prominent complaints. Physical exercise may alleviate the headache. Chronic tension-type headache is often associated with concomitant cervicogenic disease. Trigger and aggravating factors are usually identifiable and may include stress in marital, social, occupational, sexual, and interpersonal relationship; fatigue; overwork; insomnia; personality traits; and methods of handling stressful situations.

B. Pathophysiology

The mechanism of tension-type headache is poorly understood. In the episodic form, more peripheral mechanisms seem to be involved, including slightly increased electromyographic activity during resting conditions that are due to insufficient relaxation or reflex contraction. Slightly decreased electromyographic activity during maximal voluntary contraction may be due to impaired recruitment of muscle fibers at maximal activity. In the chronic form, it appears that sensory information is abnormally processed. Plausible mechanisms have been proposed, including sensitization of peripheral myofascial nociceptors and spinal and trigeminal neurons; decreased antinociceptive activation from supraspinal structures; and increased sensitivity of supraspinal pain perception. Furthermore, neurochemical processes involved in depression are also thought to be part of the pathophysiology of chronic tension-type headache. There is strong evidence that depression lowers tolerance to pain.

C. Therapy

The patient should avoid identifiable trigger factors. The episodic form responds well to simple analgesics, muscle relaxants, relaxation techniques, biofeedback, and exercise. Mild headaches of shorter duration are self-limited and in order to prevent analgesic dependency do not need to be treated. Patients with the chronic form need to be evaluated for depression and other psychological and psychiatric comorbid conditions. The daily use of analgesics is inappropriate because of the potential for dependency problems or a rebound phenomenon.

The antidepressant drugs provide the most effective treatment for this condition. Antidepressant prophylaxis should be used even if depression is not detected. Only a few antidepressants in the tricyclic category have been tested in placebo-controlled studies for chronic tension-type headaches. However, clinical experience suggests that other tricyclic and nontricyclic antidepressants, including the newer serotonergic types, are similarly effective. There is a minimal difference in their efficacy, but the adverse reactions can be significant. The choice of antidepressants is therefore made on the avoidance or use of certain side effects. For instance, tricyclic antidepressants such as amitriptyline or doxepin, which have sedative effects, can be used in the patient who has a sleep disturbance. The average dose of antidepressant for treatment of chronic tension-type headache is lower than that used for depression. Pain control is achieved with muscle relaxants, simple analgesics, and NSAIDs. Narcotic-type analgesics and tranquilizers, particularly the benzodiazepine types, are not recommended because of their high propensity of dependency, abuse, and addiction, particularly in a patient with coexisting chronic tension-type headache and depression. Psychotherapy, stress management, and cognitive therapy are essential parts of successful treatment.

III. TRACTION AND INFLAMMATORY HEADACHES (ORGANIC)

Headaches due to organic disease are rare, occurring in about 2% of headache patients. The term traction and inflammatory headaches refers to any headache resulting from inflammation, traction and displacement, and distortion of the pain sources of the head, such as the cranial vessels. Headaches associated with traction include hematomas, abscesses, aneurysms, brain tumor, and nonspecific brain edema from lumbar puncture. Examples of inflammatory headaches include eye infection, iritis, and glaucoma; ear diseases; and acute sinus infections.

A patient presenting with recent onset or change in character of headaches should alert the physicians to a possible organic cause. It is essential to rule out possible morbid causes of the headache. The danger signals for headaches that may suggest the presence of serious illness include

- Headaches that do not fit a recognizable pattern or a pattern that is easily identified
- Headaches occurring for the first time in childhood or after age 50
- Headaches occurring for the first time that rapidly increase in frequency and intensity
- The presence of neurological symptoms, such as dizziness, blurred vision, or memory loss
- A patient who feels sick or "not right" with his or her headaches
- Abnormal physical symptoms, for example, heart murmurs or kidney problems
- Any rigidity (stiffness) of the neck accompanying a headache may indicate an infection or inflammation of the spinal fluid

These signals are important for all headache sufferers—even those with a prolonged history of migraine headaches.

A. Mass Lesions

This category includes brain tumors, hematomas, brain abscesses, and expanding aneurysms. Headache is one of the cardinal signs of a mass lesion but may not present until the size is sufficiently large or the lesion expands to press on structures that will trigger a headache and associated neurological symptoms. Headache will occur if the lesion is expanding rapidly and is producing traction on one of the pain-sensitive structures of the head, especially if the ventricular system is compromised, with obstruction of absorption or flow of cerebrospinal fluid (CSF) and a resulting hydrocephalus.

The headache resulting from a brain mass is usually associated with neurological signs, including seizures (focal or general), progressive loss of neurological function, and mental symptoms. Headache onset is usually intermittent, with dull and aching pain. The frequency and duration will progressively increase, and the headache will sometimes be altered by changes

in posture and tone. A rapid increase in intracranial pressure will usually manifest as a headache.

The headache associated with a brain tumor may facilitate locating the tumor. If the tumor is above the tentorium, the pain is frequently at the vertex or in the frontal regions. If the tumor is below the tentorium, the pain is occipital, and cervical muscle spasm may be present. Headache is almost always present continually with posterior fossa tumor. If the tumor is midline, the pain may be increased with cough or straining or sudden head movement. However, this exacerbation also occurs with migraine. If the tumor is hemispheric, the pain is usually felt on the same side of the head. If the tumor is chiasmal, at the sella, the pain may be referred to the vertex. Colloid cysts of the third ventricle can cause an acute increase in intracranial pressure when the patient assumes certain positions and the tumor blocks the flow of CSF. The pain is relieved when the patient moves the head back to a position in which the flow of CSF is unobstructed. The management of mass lesions requires neurosurgical intervention.

B. Post-Lumbar Puncture Headache

If a patient complains of headache following a lumbar puncture, it is probably related to a loss of CSF secondary to leakage through a dural defect. The headache is often exacerbated in the upright position and relieved with recumbency. The pain has been described as a dull ache that may become throbbing. The headache onset starts within hours to days after the procedure and may persist for 2 to 3 weeks. The symptoms usually subside spontaneously.

Prevention is the key and the use of smaller needles has been recommended to decrease the incidence of these headaches. Treatment of the post-lumbar puncture headache consists of bed rest in the horizontal position. A blood patch to stop the leak may be beneficial.

C. Headaches of Ocular Origin

Headache is rarely due to the eye, with the exception of obvious ocular pathology. Photophobia, associated with migraine, is rarely caused by diseases of the eye, eye muscles, or the optic nerves. Reading, eye strain, eye muscle imbalance, or refractive errors are rare causes of headache.

The pain of glaucoma is due to an increased intraocular pressure within the globe. The severity is more directly related to the rate of increase of the intraocular pressure rather than the absolute pressure. Glaucoma can be easily classified as: (i) open if the anterior angle filtration is patent, (ii) closed if the chamber is blocked, or (iii) combined if the chamber is patent and blocked. The type of glaucoma with the most severe pain is caused by acute closure of the angle in the anterior chamber. On exam, the orbit is ''rock hard'' and immediate ophthalmologic referral is necessary.

D. Sinus Headache

Sinus headache is an often cited complaint of many patients, although the acute headache due to actual sinusitis occurs less frequently than the rate quoted by the advertising media. Acute sinusitis presents with fever, pain triggered by pressure or direct percussion, and headache. Fever is the cardinal sign of this infective process. The pain associated with sinus diseases is a constant, dull ache. If the patient is suffering from acute sinusitis, the headache will typically increase in intensity as the day progresses. To confirm the diagnosis, sinus X-rays or sinus CT should be performed. Treatment consists of antimicrobial therapy and decongestants.

E. Facial Pain

The facial pain syndromes are a group of disorders that usually occur in paroxysms and are characterized by pain of severe intensity. The most common types are trigeminal and glossopharyngeal neuralgia. Trigeminal neuralgia, also known as tic douloureux, is usually seen in the elderly. The nature of the pain and its intensity may cause the patient to wince or twitch. The patient can identify trigger points, and avoidance of these trigger points is a diagnostic feature. The patient avoids these points by circumventing them during shaving, washing the face, or applying makeup. Glossopharyngeal neuralgia is similar to trigeminal neuralgia in its presentation. However, the site of pain is located in the ear, tonsils, or pharynx and is triggered by swallowing, yawning, or eating.

The prophylactic treatment for facial neuralgias consists of the anticonvulsants, and carbamazepine has produced a dramatic response. Its initial dose is 100–200 mg, two or three times daily. Most physicians would use gabapentin or baclofen after carbamazepine. Neurosurgical intervention is indicated for

approximately 25–50% of patients who are refractory to therapy. Due to the chronic nature of the pain, habituating analgesics should be avoided.

F. Temporal Arteritis

Temporal arteritis is caused by an inflammatory process to the cranial arteries, and headache is the most common presenting complaint. It is also associated with night sweats, weight loss, aching of joints, low-grade fever, and jaw claudication. The headache pain is usually localized to the affected scalp vessels. Many patients will complain of pain on chewing. Temporal arteritis should be ruled out in any patient over the age of 50 who presents with an initial onset of headache and who was previously asymptomatic. The female-to-male ratio is 2:1. The area around the temporary artery is tender and the skin may appear red, depending on which artery is involved. Diagnosis is suggested by an elevated sedimentation rate by the Westergren method and confirmed by temporal artery biopsy.

It is essential that a diagnosis be established early and treatment started immediately because 50% or more of untreated cases result in irreversible blindness. Therapy with corticosteroids can be initiated while awaiting the results of the biopsy to avoid delays in treatment.

G. Headache with Vascular Disorders

Fortunately, headaches due to disorders of the cerebral vessels are rare. Because of the gravity of these disorders, these potentially life-threatening disorders should be quickly diagnosed and appropriate treatment started immediately.

1. Epidural Hematoma

Bleeding into the cranium may occur into the epidural, subdural, or subarachnoid space or directly into the parenchyma of the brain. Severe headaches are commonly manifested by these hemorrhages. However, the headache may not be a dominant feature or may be absent in some cases. An epidural hematoma is usually caused by a tear in the middle meningeal artery as it passes under the surface of the temporal bone. Because the dura is sensitive to stretching as it is being torn away from the bone, headache often occurs. Restlessness and combativeness also occur and progress to an altered consciousness. Neuroimaging will confirm the diagnosis. Prompt neurosurgical referral is essential.

2. Acute Subdural Hematoma

Subdural hematoma can be due to either an acute or a chronic event. In acute subdural hematoma, a brief lucid interval occurs between the head trauma and the patient becoming comatose, although the patient is usually comatose from the time of trauma. The bleed may be unilateral or bilateral and is often accompanied by lacerations of the scalp and contusions to the brain and parenchyma. A CT scan will confirm the diagnosis in up to 90% of cases, although angiography may be necessary. Neurosurgical intervention is required.

Chronic subdural hematoma is usually precipitated by minor head injury, which is often forgotten by the patient, and the trauma may have occurred several months previously. This form of hematoma occurs more commonly in the elderly and in patients receiving anticoagulant therapy. In addition to the headache, the patient may have decreased mentation, confusion, and drowsiness. The headache is considered secondary to the stretching of the tributary veins that drain the vessels of the cerebral hemispheres into the sagittal sinuses. Neuroimaging will establish the diagnosis, although angiography may be required. Treatment consists of surgical burr holes and evacuation of the clot.

3. Subarachnoid Hemorrhage

Most subarachnoid hemorrhages (SAHs) are caused by a ruptured berry aneurysm. The patient with SAH will describe the ''worst headache of my life.'' Other causes of SAH include arteriovascular malformations, bleeding disorders, and miscellaneous or cryptogenic causes. A berry aneurysm results from a congenital weakness in the arterial wall, usually occurring in the vessels of the circle of Willis and at these vessels' bifurcations. Frequently, patients are drowsy, may vomit, and have meningeal signs on physical exam. If SAH is being considered, an unenhanced CT scan should be performed and immediate neurosurgical consultation should be provided. A four-vessel cerebral angiogram is usually the next stage of the workup and serves as a road map for surgical intervention.

4. Arteriovenous Malformations

Unless there is active bleeding, arteriovenuous malformations (AVMs) do not usually cause headaches. If

there is bleeding, the AVM can mimic the symptoms of SAH. A slow leak from an AVM could possibly occur and irritate the meninges, thus causing headache.

5. Hypertensive Headache

Hypertensive headache is typically bilateral and often presents in the occipital region. Characteristically, the headache is worse upon awakening and gradually improves throughout the day. The diagnosis can only be confirmed with a diastolic blood pressure of 110 mmHg or higher.

See Also the Following Articles

BIOFEEDBACK • BRAIN DISEASE, ORGANIC • BRAIN LESIONS • DEPRESSION • FUNCTIONAL MAGNETIC RESONANCE IMAGING (fMRI) • MILD HEAD INJURY • PAIN • PAIN AND PSYCHOPATHOLOGY • STRESS

Suggested Reading

Dalessio, D. J., and Silberstein, S. D. (1993). *Wolff's Headache and Other Head Pain*, 6th ed. Oxford Univ. Press, New York.

Diamond, M. L., and Solomon, G. D. (1999). *Diamond and Dalessio's the Practicing Physician's Approach to Headache*, 6th ed. Saunders, Philadelphia.

Diamond, S. (1998). Headache. In *Conn's Current Therapy* (R. A. Rakel, Ed.). Saunders, Philadelphia.

International Headache Society (1988). Classification and diagnostic criteria for headache disorders, cranial neuralgias and facial pain. *Cephalalgia* **8** (Suppl. 7).

Sjaastad, O. (1992). *Cluster Headache Syndrome*. Saunders, Philadelphia.

Raskin, N. H. (1988). *Headache*, 2nd ed. Churchill Livingstone, New York.

Hearing

JOHN F. BRUGGE
University of Wisconsin

MATTHEW A. HOWARD
University of Iowa

GLOSSARY

characteristic frequency The frequency of a tone (kilohertz) that excites an auditory neuron at the lowest threshold (decibel sound pressure level). Sometimes called the "best" frequency.

cytoarchitecture The structural organization of neuronal cell bodies in the brain as revealed by specific histological staining methods. The cerebral cortex in particular may be subdivided based on cytoarchitectural differences.

decibel (dB) The unit of sound intensity usually expressed as sound pressure level (SPL). Each 10-fold change in sound pressure is a 20-dB change in SPL.

Hertz (Hz) The unit of the frequency of a sound, defined as the number of cycles that a periodic signal undergoes each second. 1 kilohertz (kHz) = 1000 Hz.

Heschl's gyrus A gyrus (or in some cases multiple gyri) on the superior temporal plane that is the location of primary auditory cortex in the human.

response area The area within the threshold tuning curve.

superior temporal plane The dorsal surface of the superior temporal gyrus, buried in the Sylvian fissure, that contains much of auditory cortex in humans and in nonhuman primates.

threshold tuning curve A plot of an auditory neuron's sensitivity to sound frequency. Acoustic threshold (dB) is plotted as a function of stimulus frequency (kHz). The tip of the tuning curve is at the neuron's characteristic frequency.

tonotopy The orderly representation of stimulus frequency in the auditory system; often studied by mapping with a microelectrode the distribution of neurons' characteristic frequencies within an auditory structure.

transduction As related to hearing, it is the process by which sound energy is changed (transduced) into electrical energy (nerve impulses) within the inner ear.

Hearing engages a set of complex processes by which humans and other animals detect and discriminate sounds in the environment and determine the directions in space from which they arise. It also involves perceptual and cognitive processes that allow for sound source identification, for species-specific communication, and, in the apparently unique case of humans, for speech and language. Normal hearing is made possible through efficient sound transmission by the external and middle ears, transduction of sound energy into nerve impulses in the inner ear, transmission to the brain by the auditory nerve of information-bearing impulse trains, transformations of the transmitted information along the central auditory pathways, and integration of this information by widely distributed neuronal assemblies within and outside of the classical auditory pathways of the brain. Knowledge of the mechanisms underlying these processes in the human auditory pathways has been obtained from comparative anatomical, physiological, and behavioral studies in laboratory animals, from psychophysical experiments and noninvasive measurements of central

auditory function in normal human subjects, and from intraoperative and chronic electrophysiological studies of auditory function in neurosurgical patients.

I. HUMAN HEARING

Human hearing operates over a wide range of frequency and intensity (Fig. 1), and within this range humans have a remarkable capacity to detect, discriminate, and locate sounds. A young listener hears sounds ranging in frequency from about 20 to 20,000 Hz, and over a considerable portion of this range a change in frequency as little as 0.15% is detectable. This same listener detects sounds that move the eardrum over a distance no greater than the diameter of a hydrogen atom, and yet hearing remains quite clear as sound pressure level is then raised by a factor of 106 or more. Within this dynamic range of 120 decibels (dB), a change in sound intensity of 1 or 2 dB is easily detected. Listeners also detect with uncanny accuracy the location of a sound in space and discriminate between two speakers located within a few degrees of each other on the horizontal plane. Periodic envelope fluctuations, some highly complex, are common and important information-bearing features of natural sounds including speech. The even more complex process of speech communication, which in order to develop properly requires hearing ability, is normally acquired very early in life and shortly thereafter is carried out effortlessly even in environments filled with competing sounds.

II. TRANSMISSION OF SOUND BY THE EXTERNAL AND MIDDLE EARS

The functions of the external and middle ears are to capture sound energy and transmit it efficiently to the receptor organ of the inner ear. Figure 2A illustrates the structural relationships that underlie these functions.

The external ear consists of a highly convoluted protrusion, called the pinna (or auricle), and an ear canal (or external acoustic meatus). Together the pinna and ear canal amplify sound waves and guide them to the tympanic membrane (or eardrum).

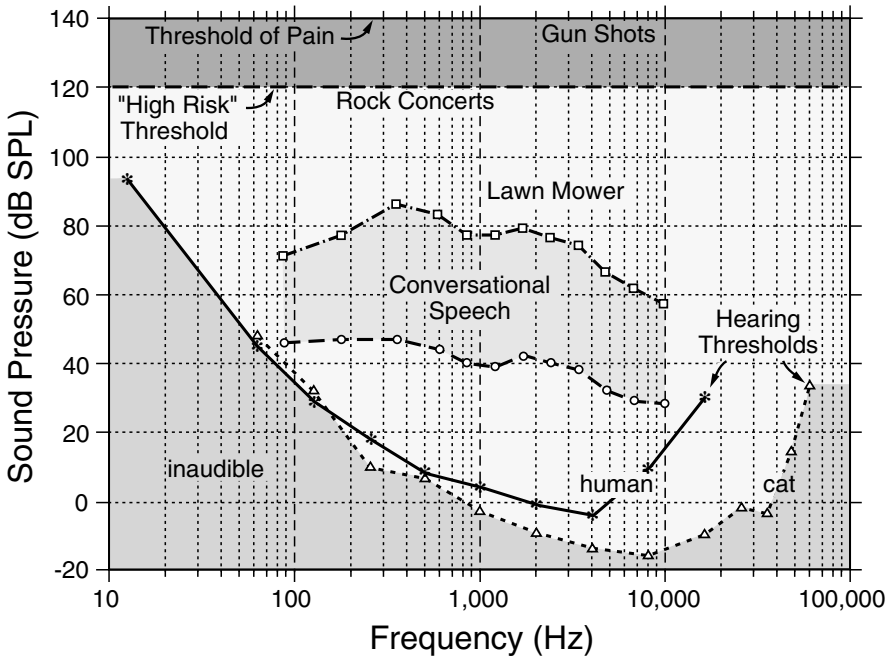


Figure 1 Hearing threshold and range of hearing for human listeners. Shown also are the ranges of frequency and sound pressure level of common environmental sounds, including human speech. The most intense sounds are capable of damaging the inner ear receptor organ. The hearing sensitivity of the cat, a laboratory animal commonly used in studies of the peripheral and central auditory systems, is illustrated as well (adapted with permission from Geisler, C. D, *From Sound to Synapse*. Oxford Univ. Press, New York, 1998).

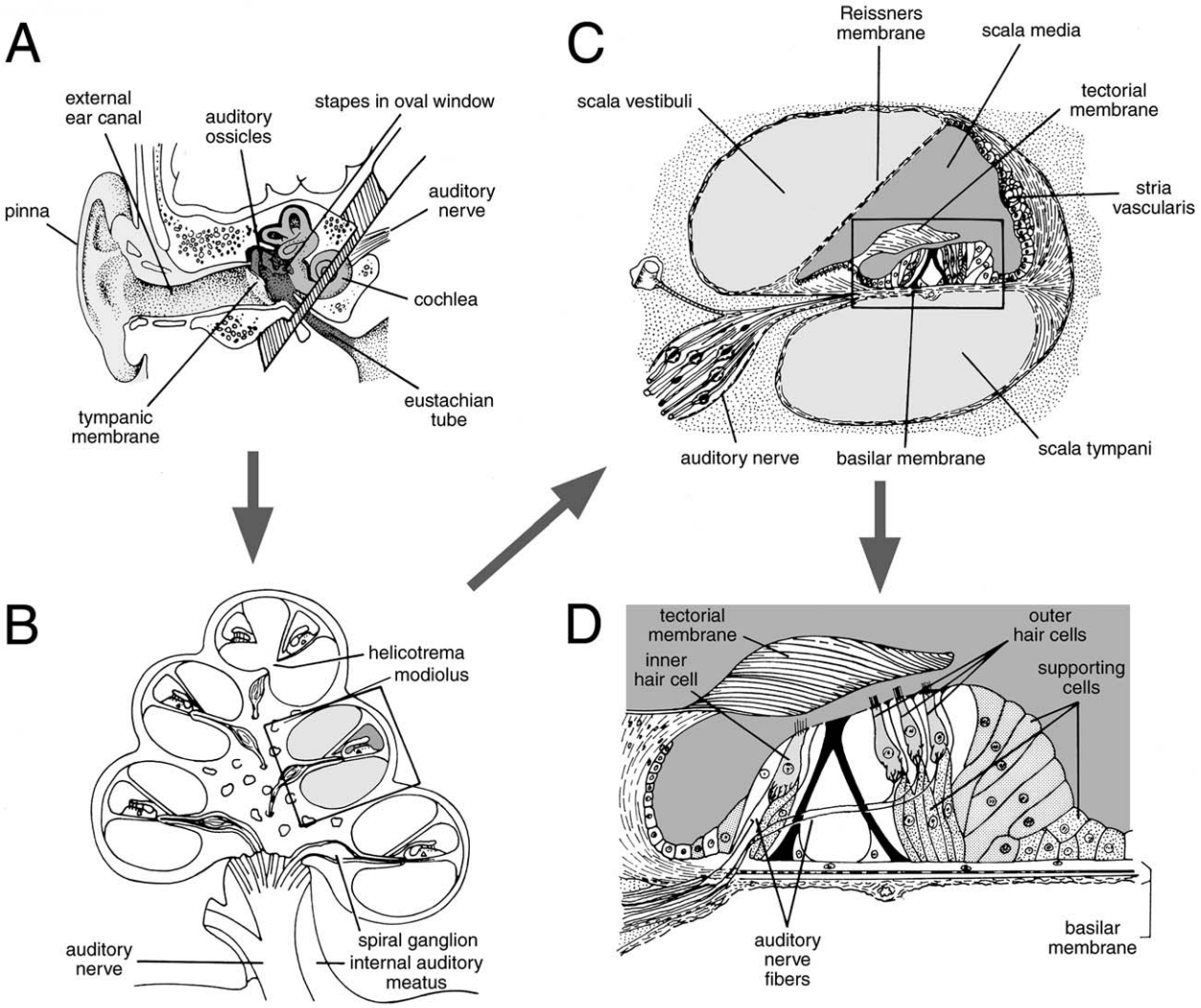


Figure 2 Structures of the external, middle, and inner ears [adapted with permission from Brugge, J. F., Auditory system. In *Encyclopedia of Neuroscience* (G. Adelman, Ed.). Birkhauser, Boston, 1987].

Amplification of sound ranges from 5 to 20 dB for frequencies within the speech range (about 1.5–7 kHz). This improvement in signal-to-noise ratio, which is due mainly to the resonant properties of the external ear canal, provides a way of enhancing speech intelligibility in the presence of many unwanted competing sounds. The pinna, on the other hand, acts as a directional amplifier for high frequencies. Spectrally transformed sound reaching the tympanic membrane provides cues for localizing the source of a sound in space, especially when the sound is on the midsagittal plane, where interaural time and intensity differences are small or nonexistent. Sounds originating from sources on either side of the midline reach the near ear before reaching the far ear, thereby creating an interaural time difference (ITD). The head also acts as an acoustic barrier at high frequencies, creating an interaural intensity difference (IID). The magnitudes of the ITD and IID depend on the location of the sound on the horizontal plane, and neural circuits in the auditory central nervous system have evolved to detect them.

The middle ear cavity is located just behind the tympanic membrane. It is normally air-filled and in equilibrium with atmospheric pressure due to the periodic opening and closing of the Eustachian tube connecting the middle ear cavity with the nasopharynx. Three auditory ossicles (malleus, incus, and

stapes) connect the tympanic membrane with the oval window of the inner ear. Reflex contraction of muscles attached to the stapes and malleus stiffens the ossicular chain and thereby reduces transmission of potentially damaging low-frequency sounds. The real need for a middle ear arises because the auditory receptor organ is an "underwater receiver" operating in the fluid environment of the inner ear. If sound waves in air were to strike this fluid boundary, 99.9% of the energy would be reflected. This interruption in the flow of sound to the inner ear would result in a conductive hearing loss, possibly as much as 30 dB. Thus, the role of the middle ear is to overcome this impedance mismatch between air and fluid and to transfer to the inner ear as efficiently as possible sound energy that impinges on the tympanic membrane.

The first, and most important, mechanism used to overcome impedance mismatch relies on the relatively large area of the tympanic membrane compared to the oval window into which the stapes footplate exerts pressure on the fluid of the inner ear. The force acting on the tympanic membrane is concentrated through the ossicles onto a small area of the oval window resulting in a pressure increase proportional to the ratio of the areas of the two membranes (approximately 20:1). Second, the lever arm of the malleus is longer than that of the incus with which it articulates, giving an additional mechanical advantage of about 1.3. Third, the conical shape of the tympanic membrane imposes additional force on the malleus. Sound energy transmitted to the inner ear is then transduced, through a cascade of mechanical and electrical events, to electrical nerve impulses in axons of the auditory nerve.

III. TRANSDUCTION OF SOUND BY THE INNER EAR

A. Structure of the Inner Ear

The inner ear, located in the temporal bone of the skull, is a complex structure serving both hearing and balance. Figures 2B–2D illustrate in detail the structures that make up that portion involved in hearing known as the cochlea (Fig. 2A). The cochlea is a coiled, tapered, and fluid-filled chamber divided along almost its entire length by a membranous partition. The two spaces thus formed, the scala vestibuli and scala tympani, are filled with a fluid called perilymph. The two scalae communicate with each other through an opening at the top (apex) of the cochlea called the helicotrema. At the base of the cochlea each scala terminates at a membrane that faces the middle ear cavity. The scala vestibuli ends at the oval window into which the footplate of the stapes rocks when the ear drum moves; the scala tympani ends at the round window, a structure that provides a pressure relief for movement of the cochlear fluid. The partition that divides the cochlea lengthwise is a fluid-filled tube called the scala media or cochlear duct. Its fluid, endolymph, is chemically different from perilymph. The cochlear duct is bounded on three sides by a bed of capillaries and secretory cells (the stria vascularis), a layer of simple squamous epithelial cells (Reissner's membrane), and the basilar membrane, on which rests the receptor organ for hearing—the organ of Corti.

The organ of Corti contains the auditory receptor cells (hair cells) and a system of supporting cells that hold them in place. Hair cells are modified epithelial cells with hairs (stereocilia) protruding from their apical ends (Fig. 2D). Figure 3 is a drawing of a stereotypic ear hair cell. A kinocilium, which seems to play no active role in the transduction process, is seen throughout life in the vestibular epithelium but is no longer present in the adult cochlea. Within the organ of Corti two kinds of hair cell—inner (IHC) and outer (OHC)—are distinguishable by location, morphology, and connections with the auditory nerve. Approximately 3500 IHCs form a single linear array, from base to apex (Fig. 2D). About 12,000 OHCs are arranged in three or four rows, parallel to the IHCs (Fig. 2D). Approximately 100–150 stereocilia form a V or W pattern on each OHC, whereas approximately 40–50 stereocilia, arranged in a U shape, adorn each IHC. Displacement of stereocilia is the adequate stimulus for generating receptor currents in hair cells that eventually lead to action potentials in auditory nerve axons.

Spiral ganglion cells, located in the bony core of the cochlea, are the bipolar first-order neurons of the auditory pathway. Their distal processes make synaptic contact with the base of hair cells. The central axons, which in the human number approximately 35,000, form each auditory nerve bundle. The majority ($\sim$95%) of the central processes of spiral ganglion neurons originate at the base of IHCs. Thus, axons connected to IHCs transmit to the brain trains of nerve impulses that encode essentially all the acoustic information eventually perceived by a listener. The innervation is highly focused: Each auditory nerve is connected to one IHC, whereas each IHC is innervated

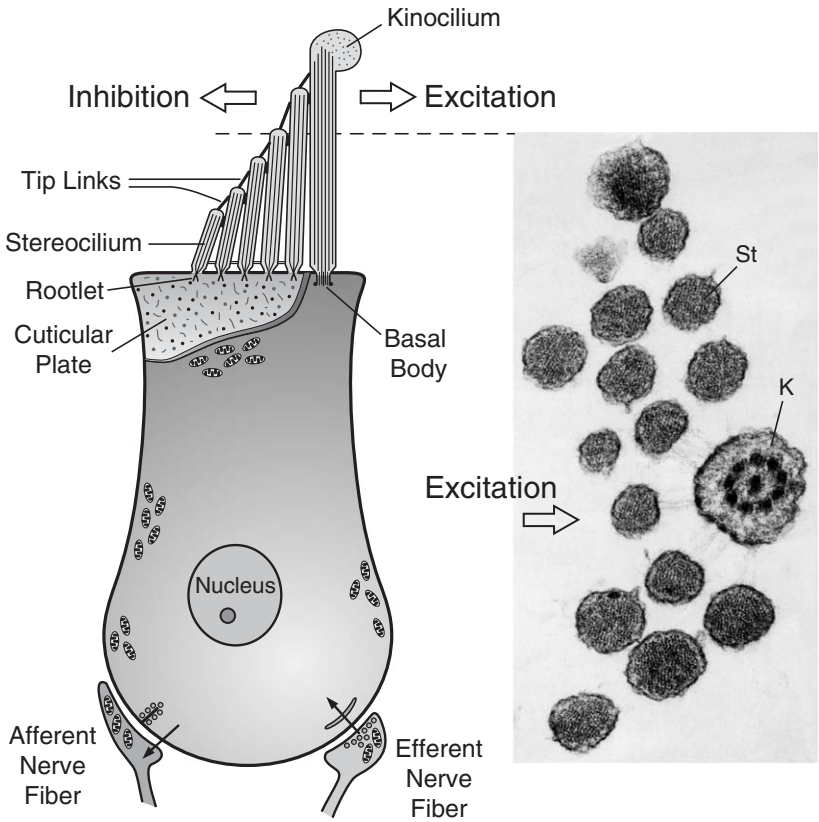


Figure 3 Stylized hair cell of the vertebrate inner ear (left). Deflection of stereocilia toward the kinocilium and basal body results in hair cell depolarization and excitation of auditory nerve fibers, whereas movement in the opposite direction leads to hyperpolarization and inhibition. Cross section of hair bundle is shown at the right (adapted with permission from Geisler, C.D, *From Sound to Synapse*. Oxford Univ. Press, New York, 1998).

by no more than approximately 10–20 spiral ganglion cells. The remaining 5% of axons of the auditory nerve are efferent fibers arising from cells in and around the superior olivary complex of the brain stem. Upon entering the organ of Corti they branch profusely, with each axon reaching many OHCs over considerable distance.

B. Mechanical Motion in the Inner Ear in Response to Sound

The basilar membrane is a resonant structure varying systematically in width and stiffness. It is wider (0.42–0.65 mm) and more flaccid at the cochlear apex than at the base (0.08–0.16 mm). When a sound wave is transmitted to the fluid of the inner ear, the basilar membrane is set in motion. Basilar membrane motion is best described as a traveling wave of deformation, which begins at the cochlear base and moves apically toward a frequency-dependent place of maximal amplitude (Fig. 4). When very high-frequency sound waves reach the ear, only the region nearest the cochlear base vibrates. As the frequency of the sound is lowered, the place of maximal amplitude of vibration shifts toward the cochlear apex. Because of this resonance gradient, the basilar membrane is said to be "tonotopically" organized. Consequently, complex sound (e.g., speech) entering the inner ear is resolved into its component frequencies. This physical separation of sound energy into its spectral components, coupled with the focused innervation of the auditory nerve array, provides an orderly and spectrally segregated projection of the nerve into the auditory brain stem, thereby setting the stage for tonotopic organization along the entire central auditory pathway.

Normal cochlear vibration is not simply the result of passive mechanical resonance as once thought, but rather it involves active processes. Under certain conditions, OHCs are capable of changing shape,

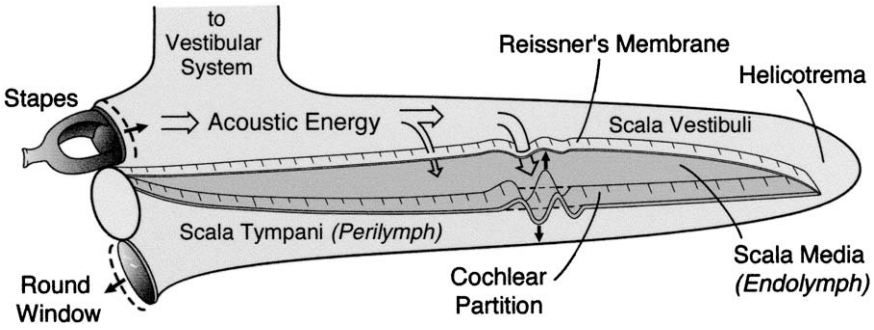


Figure 4 Stylized mammalian cochlea, shown uncoiled to illustrate the flow of energy and pattern of vibration of the cochlear partition in response to a midfrequency tone of modest intensity. The scalae vestibuli and tympani are assigned the same shading as in cochlear cross sections shown in Fig. 2. [Adapted with permission from Geisler, C.D. (1998). *From Sound to Synapse.* Oxford Univ. Press, New York].

which feeds energy back into the organ of Corti and alters the mechanical properties of the cochlear partition and possibly the transduction process. This active process may be controlled, in part, by feedback via olivocochlear axons originating in the auditory brain stem and profusely terminating at the base of OHCs. Apparently as the result of an active process, the organ of Corti acts not only to receive sound but also to generate it, as an otoacoustic emission (OAC) recorded by a microphone in the ear canal. There are several categories of OACs, reflecting perhaps more than one nonlinear active process in the cochlea. OACs are proving useful as an objective tool for diagnosing sensorineural hearing loss.

C. Transduction Process in Cochlear Hair Cells

Stereocilia of IHCs are in functional contact with an overlying auxiliary structure called the tectorial membrane. The base of the hair cell is in synaptic contact with the distal ends of auditory nerve axons. Sound waves that reach the inner ear set the basilar membrane, and hence the organ of Corti, into motion. This causes a shearing motion between the tectorial membrane and the tops of the hair cells that, in turn, displaces the stereocilia and triggers the flow of transducer currents. These changes in the receptor potential are mediated by the opening and closing of mechanically gated ion channels (transduction channels) at the tips of the stereocilia. This action leads to opening and closing of voltage-gated ion channels distributed over the basolateral surface of the cell body and then to the release of neurotransmitter at the afferent synapses at the base of the hair cell. The hair bundle is ''polarized,'' which means that displacement of the stereocilia in the direction of the kinocilium (or basal body) results in hair cell depolarization and an increase in firing of auditory nerve fibers, whereas displacement in the opposite direction leads to hyperpolarization and a decrease in firing (Fig. 3). Thus, the modulation of neurotransmitter release, and, as a consequence the pattern of action potentials in the auditory nerve, is linked tightly to the intensity, frequency, and temporal structure of sound waves entering the ear.

Inner ear structures are easily damaged by intense sound, drugs, viruses, and bacteria, and there are genetic causes of inner ear malformation. The resulting hearing loss is called sensorineural, and in such cases no treatment has been found to fully restore normal inner ear function. Some functional hearing may be restored, however, by electrically stimulating surviving spiral ganglion neurons through a cochlear prosthesis.

IV. TRANSMISSION OF SOUND INFORMATION BY THE AUDITORY NERVE

The term ''code'' as applied to the auditory system is simply a way in which information about a sound is represented in impulse activity of neurons. As in other regions of the central and peripheral nervous systems, the auditory system exhibits a variety of neural activities and, hence, a variety of candidate coding mechanisms. These include labeled lines (place), firing frequency (rate), temporal patterning, and ensemble firing. Each auditory nerve encodes the frequency, intensity, and temporal pattern of the sound reaching the ear. Coding for the direction of sound in space involves more complex processing in the central auditory pathway, where inputs from the two ears converge and where sound direction is computed from ITD and IID cues.

Auditory nerve fibers exhibit a frequency specificity that reflects the mechanical tuning properties of the basilar membrane and the highly focused innervation of IHCs. The discharge threshold of an auditory nerve fiber occurs around a particular frequency, referred to as the fiber's characteristic frequency (CF). Thus, a fiber's CF is directly related to the location along the basilar membrane innervated by that fiber. The fact that auditory nerve fibers and many central auditory neurons are frequency selective has been taken as evidence to support a place theory of hearing. The place theory is a labeled-line theory stating that the pitch of a tone (i.e., the psychological attribute associated with frequency) is determined by those fibers in the auditory nerve array (and, by extension, central auditory neurons) excited by that sound.

Not all sounds having pitch quality exhibit spectral energy at the pitch frequency; pitch is also perceived when a sound waveform is modulated in amplitude. Hence, frequency (or pitch) information may be transmitted from the ear to the brain in the temporal pattern of nerve impulses in auditory nerve fibers. Action potentials evoked by low-frequency tones (below about 4 kHz) or by temporally modulated sounds (below about 1000 Hz) occur at preferred times on a stimulus cycle, a phenomenon known as phase locking. Phase locking is the direct result of to-and-fro displacement of IHC stereocilia in response to low-frequency modulation of the basilar membrane. As a consequence of phase locking, nerve impulses tend to occur at integral multiples of the period of the stimulus time waveform, a behavior that is predicted by the volley theory of hearing. The volley theory states that the pitch of a sound is determined by the temporal rhythm of the discharges of an ensemble of auditory nerve fibers. High-fidelity transmission of temporal information from the cochlea to the brain serves at least two purposes. First, because much of human speech (especially vowel sounds) has its acoustic energy concentrated below about 4 kHz and contains amplitude-modulated components, phase locking serves to preserve temporal information found in human speech sound. Second, it serves to transmit temporal information to neural circuits of the brain stem capable of detecting ITD cues used by listeners for localizing the source of a sound in space. Neither the place theory nor the volley theory alone, however, fully account for a listener's ability to distinguish one tone from another, to encode the entire spectrum of speech, or to localize equally well both low- and high-frequency sounds. Hence, a duplex theory of hearing was postulated that states that temporal coding operates at low spectral frequency and that rate and place coding operate at high spectral frequency.

Sound intensity is likely encoded, in part, by the firing rate of auditory nerve fibers. As stimulus intensity is raised, the number of discharges steadily increases up to a plateau approximately 30–40 dB above threshold. Because a typical single auditory nerve fiber responds more or less linearly only over a 30- to 50-dB range of intensity, a second mechanism is needed to achieve the nearly 100 dB of dynamic range experienced by human listeners. This mechanism involves recruiting additional fibers with different discharge threshold into the active population.

Through the coding mechanisms outlined previously, the cochlear output faithfully reflects the time structure and intensity of speech and other naturally occurring sounds, with the spectrotemporal components of these coded messages distributed across the tonotopic array of primary afferent fibers. The representation of speech in the auditory periphery thus involves the full complement of temporal, rate, place, and ensemble encoding mechanisms operating in concert. Which mechanism dominates depends on the CF and spontaneous rate of the auditory nerve fiber and, in each instant in time, on the acoustic properties of the utterance and of the ambient environment.

Information transmitted in the auditory nerve array is received and transformed by neurons and neuronal circuits of the cochlear nuclei, and the resulting outputs are distributed over various pathways of central auditory system for further processing.

V. STRUCTURE AND FUNCTION OF THE CENTRAL AUDITORY PATHWAYS

The central auditory pathways have been studied extensively in many primate and nonprimate species using a wide range of anatomical and physiological methods. There are far fewer comparable studies in humans, but from them we may conclude that much of the neural circuitry and many of the neural mechanisms underlying mammalian hearing are, to the first approximation, shared by humans and nonhumans alike. Functional studies of the human central auditory system are discussed later. The top of Fig. 5 is a dorsal view of the human brain stem, midbrain, and thalamus showing the relative positions of the major nuclei in the ascending and descending mammalian auditory system. Auditory cortex on the temporal lobe is shown in cross section. The major connecting pathways are illustrated schematically below.

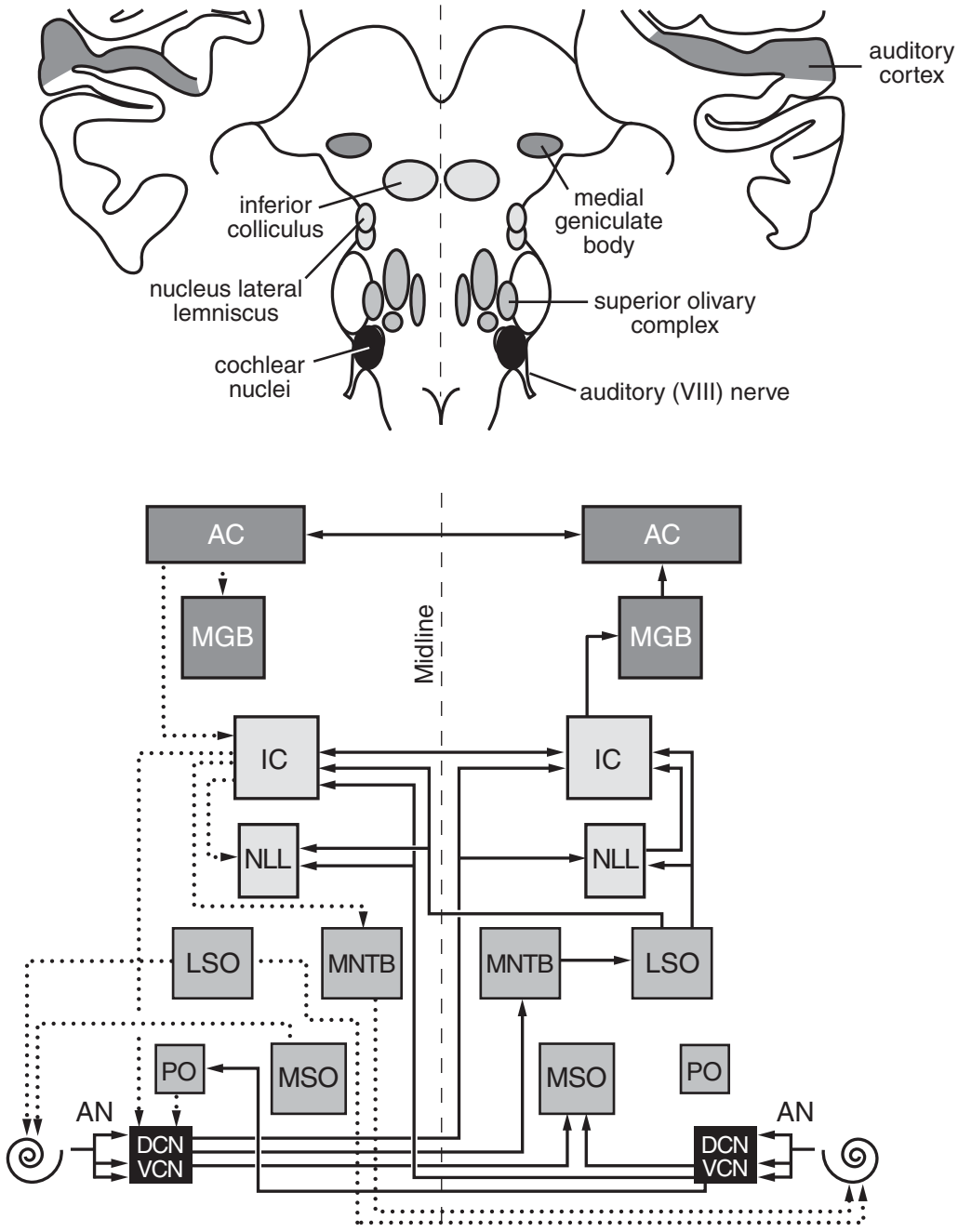


Figure 5 (Top) Dorsal view of the human brain stem, midbrain, thalamus, and cross section of the cortex showing the relative locations of the major nuclei in central auditory pathways. (Bottom) The major ascending (black lines and arrows) and descending (dotted lines and arrows) pathways connecting the major nuclei.

A. Cochlear Nuclear Complex: Initial Transformation of Sound Information Reaching the Brain

The cochlear nuclei, located bilaterally at the pontomedullary junction, are the first synaptic stations in the central auditory pathway. The cochlear nuclei are obligatory relays, meaning that all information flowing from the cochlea must here encounter at least one synapse before being transmitted to higher auditory centers. In humans, as in all other mammals studied, two major cochlear nuclei are recognized: a dorsal

cochlear nucleus (DCN) and a ventral cochlear nucleus (VCN). The latter is often subdivided into anterior and posterior divisions. Auditory nerve fibers penetrate this nuclear complex, bifurcate in an orderly (tonotopic) way, and terminate on neurons in both the DCN and the VCN. In addition to its major afferent supply from the inner ear, the cochlear nuclei receive input from a variety of other sources both intrinsic and extrinsic to the complex. Extrinsic sources include the other major brain stem nuclei. The role of these inputs in modulating the activity produced by primary afferent input from the ears is essentially unknown.

Cochlear nuclear neurons have been categorized based on their synaptic structure, cell morphology, pharmacology, connectivity, and acoustic response properties. Incoming spike trains that reach CN neurons are transformed by combinations of synaptic convergence and physiology, intrinsic membrane properties of target neurons, and feedback of excitation and inhibition via intrinsic or extrinsic circuitry. Cells of each category tend to extract particular features of sound that are transmitted by auditory nerve fibers, such as intensity, spectral composition, modulation amplitude and frequency, low-frequency time structure, and the timing of acoustic transients. Information received about human speech, for example, is thus distributed in such a way that one cell type preserves the temporal representation of vowels, whereas another cell type represents vowel information best in its rate of discharge. Similarly, other cell types extract formant transitions and still others transients (e.g., voice onset). Temporal information needed to detect ITD is preserved by still other cell types in the VCN having a highly specialized synaptic relationship (end bulb of Held) with auditory nerve fibers. Transformed information is sent to higher auditory centers for further processing over three pathways (stria).

Most axons leaving the VCN form a broad pathway that crosses the brain stem ventrally as the ventral stria or trapezoid body. A much smaller pathway, the intermediate acoustic stria, also leaves the VCN and joins the trapezoid body as it approaches the superior olivary complex, where many of its axons terminate. Axons of projecting neurons of the DCN form the dorsal acoustic stria, which reaches primarily the contralateral dorsal nucleus of the lateral lemniscus and the central nucleus of the inferior colliculus. At the level of the cochlear nuclei the input from the two ears has, for the most part, remained separated. One of the next levels of processing takes place in the superior olivary complex, where the inputs from the two ears converge and interact to encode sound direction.

B. Superior Olivary Complex: A Major Site for Integrating Input from the Two Ears

The superior olivary complex (SOC) comprises several interrelated nuclear groups located symmetrically on either side of the brain stem. There are three major SOC nuclei in most mammals studied: the lateral superior olivary nucleus (LSO), the medial superior olivary nucleus (MSO), and the medial nucleus of the trapezoid body (MNTB). In higher primates, including the human, the MNTB tends to be represented more by a collection of scattered neurons than by a coalesced structure. A number of periolivary nuclei (PO) are recognized as well. The major nuclei are the first stations in the auditory pathway that receive substantial bilateral input, and thus the SOC plays crucial roles in extracting and integrating information it receives from the two ears. Localizing the source of a sound in space is crucial to the survival of most animals, and to do so they rely on the fact that the positions of the ears on opposite sides of the head provide frequency-dependent ITD or IID. For human listeners, sound localization also plays an important role in identifying a talker in the presence of competing talkers (the so-called cocktail party effect). Sound localization in the horizontal plane is accomplished by engaging neural circuits in the SOC that compare these interaural time and intensity differences.

Principal neurons of the MSO have been shown in animal studies to receive direct input from particular cell types in the bilaterally placed ventral cochlear nuclei that preserve and transmit precise time information about a sound source in space. To a first approximation, these MSO neurons perform a cross-correlation on the spike trains arriving from the left and right ears. The projections from each ear form what may be considered a neural "delay line." As a result, phase-locked spikes transmitted over the two monaural channels arrive simultaneously at different MSO neurons depending on the interaural time delay and, hence, on the azimuthal (horizontal) location of a sound source. Next, the azimuthal location of a sound source is computed and is now represented at a "place" in the MSO.

In contrast, LSO neurons receive bilateral input from a different population of cochlear nuclear neurons. The ipsilateral input to the LSO is direct and excitatory, but the contralateral input involves an

inhibitory interneuron in the MNTB. This arrangement renders LSO neurons particularly sensitive to IIDs. Hence, at this point in the central auditory system information concerning two of the major cues used by listeners in localizing the source of a sound in space has been extracted and to some extent segregated along two ascending pathways.

Periolivary nuclei in the human tend to encircle the MSO and LSO and are not easily distinguishable into subnuclei as they are in several other mammalian species. PO neurons in experimental animals have been shown to project profusely back to the cochlear nuclei as well as forward to midbrain auditory structures.

A bundle of axons arising in the SOC as well as passing axons form the CN is referred to as the lateral lemniscus. Two major cell groups embedded in this band of axons form the dorsal (DNLL) and lateral (VNLL) nuclei, respectively, of the lateral lemniscus. A third, intermediate (INLL) nuclear group has been identified in several animal species but not in humans. The VNLL receives its major input from the contralateral VCN via the trapezoid body and from the ipsilateral SOC. In contrast, the DNLL receives convergent input from a wide variety of sources, including the contralateral DCN and DNLL and ipsilateral SOC and VNLL. Ascending axons originating in the DNLL and VNLL contribute to the lateral lemniscus, which is now destined for the inferior colliculus.

Although the functions of the SOC and LL nuclei have not been studied directly in humans, synchronous firing of neurons in and around these auditory brain stem areas is believed to be the source of the wave IV–V complex in the sound-evoked averaged brain stem potential recorded from the scalp (Fig. 7).

C. Olivocochlear Efferent System: Central Modulation of Inner Ear Transduction

The cochlea, in addition to supplying afferent input to the brain stem, receives efferent projections that originate in the SOC. In animal studies it has been shown that two basic groups of brain stem neurons provide efferent innervation to the organ of Corti. Based on the general locations of their cell bodies in the SOC, they form the lateral (LOC) or medial (MOC) olivocochlear system. The LOC originates from AChE-positive neurons in and around the LSO and terminates primarily in the ipsilateral cochlea, at the base of inner hair cells. Small AChE-positive cells, believed to be LOC neurons, are found in this location in the human brain stem. The MOC arises for the most part from neurons medial to the MSO and reaches the base of outer hair cells mainly in the contralateral inner ear. Collaterals of olivocochlear axons terminate in the cochlear nuclei. Both LOC and MOC neurons receive input from the ventral cochlear nucleus, with the MOC also receiving input from the inferior colliculus. Cochlear efferents are activated by sound and thereby provide reflex feedback to the cochlea. The consequence of this feedback is increased acoustic threshold due to reduced OHC amplification, which helps protect the cochlea from potentially damaging loud sounds and improves detectability of wanted sound in the presence of unwanted background masking noise. Sound induced reflex contraction of the middle ear muscles serves to complement the action of the olivocochlear system.

D. The Auditory Midbrain: Major Integrating Centers

The inferior and superior colliculi collectively form the roof of the midbrain. Both are major integrating centers, receiving converging input from a wide variety of sources representing the auditory, visual, and somatosensory systems.

Anatomically, the IC is subdivided into a central nucleus (ICC) and a surrounding cortex (divided into external and pericentral nuclei). The ICC forms an essential link in the mainline lemniscal auditory system and thus is critically involved in the transmission of auditory sensory information to the forebrain from lower auditory centers. The ICC also receives a rich supply of afferents from many other sources and directs its outputs to a variety of targets. This results in numerous feedback loops (Fig. 5) that involve essentially all major auditory areas of the forebrain, midbrain, and brain stem of the same side of the brain. In addition, several major axonal bundles (commissures) connect auditory areas of one side of the brain with their counterparts on the other. Hence, it is not surprising that essentially all information transmitted to the auditory forebrain is first transformed within the ICC. Transformations involve changes in the balance of excitation and inhibition exerted by converging afferent inputs—changes that are reflected in such fundamental attributes of coding as frequency tuning curves, timing of spike discharges, spontaneous background activity, sensitivity to modulations in frequency and intensity, binaural interactions, and spatial tuning. Remarkably, the diverse

parallel input and output, and the transformations associated with them, are highly organized within the ICC tonotopic map, which is best described as being made up of frequency-specific layers or bands. Within each band are overlapping functional maps related to a wide array of response features and stimulus attributes such as amplitude modulation, pitch, intensity, threshold, onset latency, sharpness of frequency tuning, and binaural sensitivity. It is thought that these functional maps may provide the early substrates from which sound percepts are eventually derived. Information transformed by the ICC is carried to the auditory forebrain via the brachium of the IC.

The superior colliculus (SC) is a layered structure involved primarily in the control of movements of the eyes to an auditory or visual target. Thus, in the classical sense it is not a "sensory" structure. Auditory input carries information on the direction of a sound source to deep SC layers. Sound-source direction is represented by SC neurons having spatial receptive fields arrayed to form a map of auditory space. Visual input is received in the upper layers, and information on target direction is provided by its position on the SC visuotopic map. The auditory and visual maps are normally in alignment, presumably to coordinate visual and auditory directional information so that the eyes can move quickly and accurately to an object that appears in the visual and sound fields.

VI. FUNCTIONAL ORGANIZATION OF THE AUDITORY FOREBRAIN

The auditory forebrain consists of auditory thalamic relay nuclei and those cortical areas with which they have intimate reciprocal relationships. Together these forebrain areas carry out the highest levels of information processing in the auditory pathway.

A. Thalamus

The medial geniculate body (MGB) is a complex of nuclei that receive massive input from the IC and thus serve as major synaptic stations in the pathways for information reaching auditory areas of cerebral cortex. The MGB nuclei can be differentiated from one another on the basis of cytoarchitecture, chemoarchitecture, tonotopy, connectivity patterns, and acoustic response properties. Neighboring nuclei of the posterior complex and pulvinar also receive auditory input and project upon auditory cortex. The auditory thalamus receives its input over pathways that originate in brain stem nuclei contributing to the lateral lemniscus as well as pathways that are nonlemniscal in origin. The cellular architecture of the human MGB is remarkably similar to that of OldWorld and NewWorld monkeys. Although the relative size and strength of each auditory cortical projection differ from one MGB subdivison to the next, each auditory cortical field receives highly convergent input from a subset of auditory thalamic nuclei and each auditory thalamic subdivision projects in a topographic way to a subset of auditory cortical fields. Thus, the auditory thalamocortical system, like auditory brain stem circuitry, exhibits widespread convergence and divergence. Tones evoke a characteristic pattern of cortical activation that reflects the thalamic projection to that active site. The earliest electrical response is thought to represent monosynaptic activation of neurons in cortical laminae 3 and 4, whereas later activity is considered polysynaptic activation through supragranular cortical layers. In addition to receiving massive ascending input from brain stem and midbrain auditory structures, the MGB also receives a substantial projection from those areas of cerebral cortex to which it projects. This reciprocal relationship provides opportunity for feedback control of ascending auditory input.

The ventral divison (MGBv) displays a high degree of tonotopy imposed by input from the similarly organized ICC. Nonauditory input to the ventral division is derived from the thalamic reticular and ventrolateral medullary nuclei. The MGBv projects heavily on the core areas of auditory cortex, including the primary field. Neurons in the MGBv preserve and convey to cortex with high fidelity the temporal and frequency-specific properties exhibited in the auditory brain stem. In the human, the dorsal division (MGBd) is larger and structurally more heterogeneous than the ventral division. It receives its auditory input from neurons primarily in the pericentral nucleus of the IC. Other afferents arise from the thalamic reticular nucleus, ventrolateral medullary nucleus, nucleus sagulum, superior colliculus, and brachium of the inferior colliculus. Neurons of the MGBd, in contrast to those of the MGBv, are poorly responsive to most sounds and are broadly tuned for frequency; hence, tonotopy is not a remarkable feature. The main targets of MGBd neurons are the belt areas of cortex. The medial, or magnocellular, division (MGBm) also exhibits diverse cellular architecture, and it receives input mainly from the external nucleus of the IC. Tonotopy is not a remarkable feature of this nucleus,

again reflecting the organization of its main afferent source. Neighboring thalamic nuclei, including the posterior group and the pulvinar, receive auditory input and send projections to auditory cortical fields.

MGB neurons in all divisions appear to preserve ITD or IID information extracted at brain stem levels and widely disperse this information to auditory areas of cerebral cortex. Neurons within the MGB and auditory cortical fields respond best when the sound source is in the contralateral acoustic hemifield. There is no obvious anatomical specialization in MGB in humans for speech, nor is there evidence for bilateral asymmetry in the size of the MGB.

B. Auditory Cortex

1. Auditory Cortex Is Made up of Multiple Fields to Carry out Parallel and Serial Processing

In all mammals studied, auditory cortex is made up of multiple fields, which are distinguished from one another on both anatomical and functional grounds. The number of such fields varies among species studied from as few as 2 in rodents to as many as 15 in the rhesus monkey. The number, location, and organization of such fields in the human are not fully known. It would be highly desirable, however, to know the functional and structural counterparts of human and monkey auditory cortex. Although much less is known about the functional organization of temporal auditory cortex in the human than in the monkey, from available data there are some striking anatomical similarities between the two species. These data, together with modern imaging [functional magnetic resonance imaging (fMRI) and positron emission tomography (PET)] and direct [electrocorticography (ECoG)] and indirect [electroencephalography (EEG) and magnetoencephalography (MEG)] electrophysiological recording data in humans, enable us to tentatively apply to the human a model of functional organization of auditory cortex developed for the monkey. Figure 6C shows a schematic representation of monkey auditory cortex based on modern anatomical and physiological studies. Primary auditory cortex is combined with adjacent cortex to form a core area. Surrounding the core auditory cortex is an auditory belt area, and around that is a parabelt region. The belt and parabelt areas are often referred to as secondary or associational areas. Figures 6A and 6B show two views of the human brain showing the general locations and extent of what may be equivalent core, belt, and parabelt areas in the human. There is general agreement about the location, extent, and tonotopic organization of the human primary auditory field (AI), on the mesial aspect of Heschl's gyrus (HG). There is less agreement about the organization of surrounding auditory. Whether this monkey model is adequate to describe the human auditory cortex is yet to be determined. Currently, however, it provides a tool for further exploration using modern recording and imaging technology.

We know from animal studies that auditory fields receive ascending input from the auditory thalamus, that they are richly interconnected on the same and opposite cerebral hemispheres, and that they provide afferent input to the MGB and IC as part of a massive parallel descending system of pathways that eventually reaches the lower auditory brain stem and the cochlea (Fig. 5). These auditory cortical areas are confined to temporal cortex and serve primarily a sensory function. This means that their neurons and neuronal circuits are sensitive mainly to changes in physical parameters of a suprathreshold acoustic signal reaching the ears, and thus they are designed to detect, discriminate, identify, and localize sound sources. Multiple auditory areas are thought to represent hierarchical processing levels. Some areas involved in higher order processing of acoustic stimuli, such as speech, lie outside of these "classic" auditory fields, on the frontal and parietal lobes.

2. Functional Organization of Primary Auditory Cortex: Maps of Stimulus Features

The primary auditory field (AI) has been identified anatomically and studied physiologically in a wide range of mammalian species, including humans. It is reciprocally and topographically tied to MGBv. Within AI, neurons are typically sharply tuned for frequency. Like other sensory cortices, AI exhibits a columnar representation of the peripheral sensory epithelium: Neurons occupying a vertical column tend to have the same or very similar CF. A tone just above threshold would activate a relatively small population of neurons in a cortical band having length, depth, and width (a cortical "sheet"). AI is thus said to be organized into isofrequency bands or sheets. The organization of the AI tonotopic map varies from one subject to the next, possibly suggesting that environmental factors may be involved in the formation and shaping of functional maps. It is now known that the AI tonotopic map exhibits plasticity because it undergoes dramatic change following a cochlear lesion. One can only

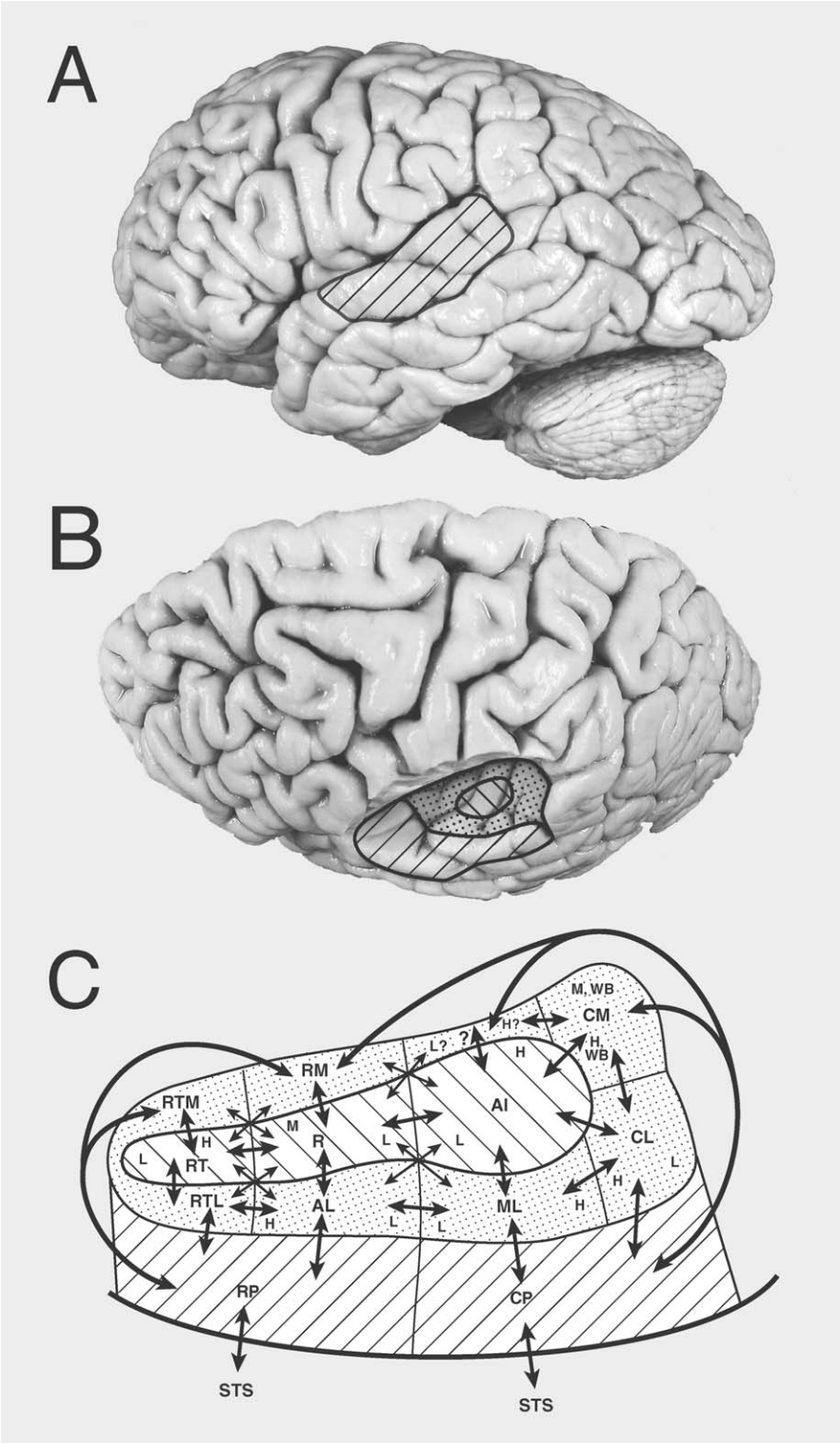


Figure 6 (A and B) Lateral and dorsal views, respectively, of the human brain showing the approximate extent and locations of known auditory fields on the lateral and superior surfaces of the superior temporal gyrus. (C) Diagram of the organization of auditory cortical fields on the same temporal lobe areas of the rhesus monkey (adapted with permission from Hackett, T. A., Stepniewski, I. and Kaas, J. H., *J. Comp. Neurol.* **394,** 475–495, 1998).

wonder if, for example, the improvement in hearing performance over time exhibited by subjects with a cochlear prosthesis may, to some degree, be attributed to auditory cortical plasticity.

In addition to a tonotopic representation within AI, studies of experimental animals have revealed spatial maps related to other functional properties associated with pure tones, ripple spectra noise, broadband transients, frequency sweeps, and binaural stimuli. The picture that emerges from this work is one of primary auditory cortex being made up of overlaid topographic representations of numerous independent stimulus features. To complicate matters even further, these representations, which are based on neuronal firing rate and spike timing, depend on stimulus intensity. Thus, individual stimulus features *per se* are probably not coded at the cortex simply by some fixed place within the primary auditory field. Instead, they may be coded by spatiotemporal activation patterns created by neuronal assemblies within cortex that change in dynamic ways to reflect the many and changing acoustic features that make up complex natural sounds, including speech.

Several auditory cortical maps represent fundamental acoustic features of a stimulus (e.g., spectrum and noise bandwidth), whereas others represent derived properties (e.g., binaural interactions) or perceptual qualities (e.g., pitch). The binaural organizational patterns distributed across AI are examples of computational maps representing the direction of a sound source in acoustic space. Because there is no extraction of spatial direction by the cochlea, such maps can only result from neural interactions taking place in the lower auditory brain stem, where spike trains arriving over the two monaural channels converge. Similarly, a cortical map of pitch may be laid out orthogonal to the tonotopic map. Pitch is the psychological attribute associate with fundamental frequency even in the absence of spectral energy at the pitch frequency. Hence, its representation on cortex would provide evidence for a higher order cortical representation of a percept rather than simply a sensation tied directly to the physical attributes of a complex stimulus.

AI cortical neurons may encode the frequency of temporal modulation of the sound envelope in the phase-locked activity of cortical neuronal assemblies. Indeed, the phase-locked thalamocortical input may provide the upper frequency limit of pitch encoding (approximatley 400–600 Hz) based on temporal mechanisms. Studies of cortical coding of species-specific vocalization and human speech in monkey have revealed no simple correlation between a single cortical neuron's response properties and a particular utterance. Thus, auditory cortical neurons may be specialized to respond more on the basis of the presence or absence of certain acoustic components embedded in a vocalization rather than on a unique vocalization per se. fMRI studies in humans have also revealed that activation of the core auditory area differed little from the surrounding cortex when speech and nonspeech sounds were presented to normal-listening subjects, again indicating that these fields are more involved in detecting the acoustic rather than linguistic parameters of speech. Indeed, the tonotopic constraints imposed on AI and the highly individualistic responses of its neurons to complex sounds suggest that specific mechanisms for identification of individual phonemes, for example, would not reside in this field and, moreover, that such specificity to speech or other species-specific communication sounds is more likely accomplished by ensembles of cortical neurons rather than by single cells. On the other hand, fields on the ventral aspect of the temporal lobe and on the temporal–parietal boundaries of the left cerebral hemisphere show greater activation to speech than to nonspeech sound.

3. Temporal Lobe Lesions Impair Acoustic Processing

Lesion behavioral studies, mainly in rat, cat, and monkey, have provided limited information on the function and organization of auditory cortex. Differences in species and in experimental paradigms have led to conflicting results across studies. One major impairment associated with auditory cortical lesions in cat and monkey, however, is localization of brief sounds in space. Lesioned animals exhibit an inability to localize sound in the acoustic hemifield opposite the side of the lesion. Additional consequences of auditory cortical lesions are impaired discrimination and retention of certain temporal patterns, including species-specific vocalizations.

Lesion studies in experimental animals usually aim to produce damage confined to a particular cortical field or portion of that field, often guided by electrophysiological mapping. This may be done repeatedly in a series of animals in which well-controlled behavior studies are carried out before and for varying period of time after the lesion is made. Far fewer systematic studies of human patients with auditory cortical lesions have been reported. Under ideal conditions a human subject communicates and cooperates well with the examiner in performing complex psychophy-

sical tests. Specific deficits in auditory processing can then be quantified and correlated with anatomical data now obtainable from brain MRIs. Rarely, however, is this ideal achieved. For the most part, patients present with clinical symptoms resulting from large, spontaneously occurring lesions; hence, no two patients experience lesions of the same size and location. The damage is rarely, if ever, confined to a single cortical field. Patient–subjects are scattered geographically and thus examined by different investigators having different specific scientific interests using different experimental methods. Possible exceptions to this are studies of epilepsy patients undergoing temporal lobectomy. Even here, however, a unilateral temporal lobe resection varies in size from one patient to the next and usually does not extend into the core auditory cortex on the mesial aspect of HG. When a cortical lesion is made for clinical purposes, only rarely are preoperative baseline psychophysical data obtained. Even when this is done, however, attention is not always paid to the fact that following an acute brain lesion considerable recovery in function occurs over time. Futhermore, many cases may be overlooked because marked, clinically significant deficits in basic auditory functions are observed only in patients who have sustained bilateral temporal lobe damage. Hence, results of psychophysical testing in the chronic state may not provide a comprehensive picture of the functional significance of regions of auditory cortex affected by a lesion.

With these caveats, it is clear that apart from speech function, which strongly lateralizes to one (usually the left) hemisphere, basic auditory functions are subserved by both hemispheres. However, there does appear to be a degree of lateralization of musical processing to the right hemisphere. Bilateral lesions approximately confined to the core auditory areas are reported to result in transient auditory agnosia, which is a lack of awareness of auditory stimuli combined with abnormal pure tone thresholds in the absence of peripheral or brain stem damage (sometimes referred to as cortical deafness). With time, this condition may evolve into auditory agnosia for speech (pure word deafness), whereby patients regain near normal auditory acuity but remain impaired in their ability to interpret speech sounds. It is hypothesized that speech decoding requires more precise temporal analysis than is necessary for the detection and identification of most nonverbal sounds. This capacity to represent auditory stimuli with a high degree of temporal resolution is dependent on having intact core auditory cortex in at least one hemisphere. Conversely, the core regions can be destroyed, but recognition of nonspeech sounds can remain intact, presumably mediated by surrounding auditory cortical fields that are functionally activated via parallel thalamic afferent inputs that do not synapse, or course through, the auditory core region. Anecdotal observations of localization ability in patients with temporal lobe lesions have suggested repeatedly that humans and nonhuman species suffer impairment in sound-localization ability. Recent and carefully controlled experiments indicate that such impairment is less marked than previously believed.

A small number of patients who have bilateral temporal lobe lesions that spare at least one core region while destroying portions of the middle and anterior regions of the superior temporal gyri (but not Wernicke's speech area) display a strikingly different pattern of auditory impairment than those with bilateral core lesions. Receptive speech function is preserved but there is a marked dysfunction of musical processing and of recognition of nonverbal auditory stimuli. It seems clear that lesions of auditory association cortex that spare core auditory regions are sufficient to impair melody discrimination and the perception of speech prosody. The existence of cases in which each one of these categories of nonverbal auditory processing is selectively impaired suggests that different subregions of auditory association cortex may serve a specialized role in representing different attributes of nonverbal auditory stimuli. It is not clear how these subregions, whose existence is implied by lesion study results, might correlate with cytoarchitectural or electrophysiologic findings in humans.

4. Functional Studies of the Human Central Auditory System

Studies of the neural mechanisms underlying hearing in humans have, with few exceptions, necessarily involved noninvasive and hence indirect methods of measurement of functional brain activity, including EEG, MEG, PET, and fMRI. These methods have proved to be most effective when used in complementary ways.

a. Electrophysiological Measures: EEG and MEG An electrode placed on the scalp records the summated electrical potentials generated by large groups of neurons in the brain. These neuronal assemblies at each station along the auditory pathway fire sychronously and in sequential order in response to a brief and abrupt acoustic stimulus. This orderly

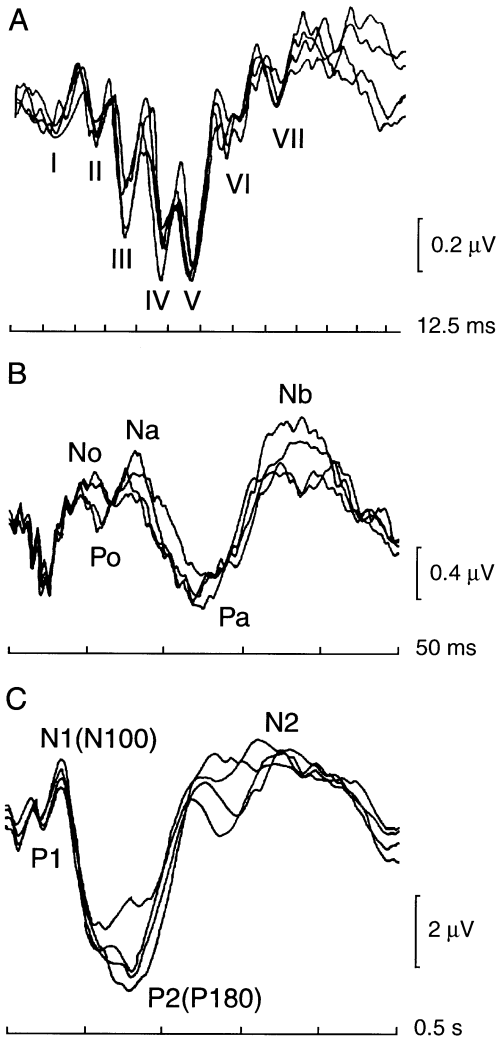


Figure 7 Auditory event-related potentials (ERPs) recorded from the scalp of a human subject in response to the presentation of brief sounds. The ERP is divided into three epochs: a short latency auditory brain stem response (ABR) (A), a middle latency response (MLR) (B), and a late, long-latency response (C) (adapted with permission from Picton, T.W., *et al.*, *EEG Clin Neurophysiol.* **36,** 179–190, 1974).

firing pattern is reflected in the sequence of peaks and valleys that make up the resulting evoked potential waveform, referred to as the event-related potential (ERP). The ERP typically is of very low amplitude and usually indistinguishable from the ongoing EEG. Thus, in practice, averaging the responses to hundreds or thousands of stimulus presentations is needed to reveal the ERP temporal waveform. The ERP is traditionally divided into three time epochs (Fig. 7). The seven waves arising during the first 10–15 msec (Fig. 7A) after stimulus onset constitute the averaged brain stem response (ABR). Intraoperative recording from the human auditory nerve, cochlear nuclei, and inferior colliculi has provided the most compelling evidence that waves I and II are generated by the auditory nerve, wave III by the cochlear nuclei, and wave IV by the SOC. Wave V probably originates in or below the inferior colliculus, whereas waves VI and VII originate in the ICC or thalamocortical projection system. Thus, the temporal waveform of the ERP, its spatial distribution on the scalp, and its sensitivity to various acoustic parameters can provide information about the transmission and encoding properties of large neuronal assemblies at all levels of the central auditory pathway. The ABR is highly stable under a wide range of conditions, including various stages of consciousness, attention, sleep, wakefulness, and sedation. It is present at birth and mature by approximately 18 months. Today, it is used routinely to screen for hearing impairment in newborn babies. Overall, the ABR has proven to be a very useful noninvasive and objective way to test for hearing impairment and other disorders of the peripheral and central auditory pathways, especially in those subjects who are unwilling or unable to participate.

A cluster of three major wave peaks with onset latency beyond the ABR components (out to about 50–60 msec) are widely recorded over the frontal and temporal cortex and referred to as the middle latency response (MLR; Fig. 7B). The generation of the MLR may involve the interaction of many brain structures within and outside of the classic lemniscal auditory pathways. Lesions of auditory cortex in humans and animals disrupt MLR waveforms. MLR is influenced by the state of arousal of a subject, suggesting involvement of the reticular formation. The MLR is used clinically to assess hearing thresholds in the low-frequency range and to evaluate auditory pathway function in hearing individuals and in subjects who have cochlear implants.

When trains of brief stimuli are presented the MLR components become time locked to the individual stimuli at frequencies of approximately 40 Hz. This 40-Hz response to the appropriate stimulus is exhibited by other sensory systems as well, indicating that it represents a general mechanism for recognizing a sensory event. Spontaneous electrical oscillation in the human brain at frequencies of approximately 40 Hz and its resetting by a sensory stimulus have been postulated to reflect cortical mechanisms involved in temporal binding of sensory stimuli and perceptual scene segregation.

Finally, a series of peaks and valleys is recorded with latencies exceeding 50 msec (Fig. 7C). The amplitudes of the waves are higher but more variable than earlier ones, depending on the conscious state of the subject and the stimulus paradigm employed to evoke the waveform. This has led to the suggestion that the late components represent the convergence of input from a number of forebrain systems whose interactions are related to the attentive or cognitive state of the subject.

The currents that give rise to the electrical voltage recorded on the scalp also give rise to weak magnetic fields. These weak fields can be measured by an array of sensitive magnetometers (superconducting quantum interference devices) surrounding the head. The method is known as MEG, and the response to a sound is referred to as the auditory-evoked magnetic field (AEF) (Fig. 8). AEF data reveal many of the same cortical processes as the ERP. Both methods yield similar waveforms with excellent response times capable of tracking with high fidelity neural events in time. In addition, the MEG offers relatively high spatial resolution, on the order of millimeters. Because of its differential sensitivity to currents flowing tangentially to the scalp, the MEG is particularly suited for noninvasive study of the cortex buried within fissures, including auditory cortex within the Sylvian fissure on the superior temporal plane. Because any electrical potential or magnetic signal may have more than one source in the brain (the so-called inverse problem), a source model is applied in which the orientation, strength, and location of the equivalent current dipoles are best accounted for on statistical grounds.

The ERP and AEF methods have been used effectively to study the neural activity associated with

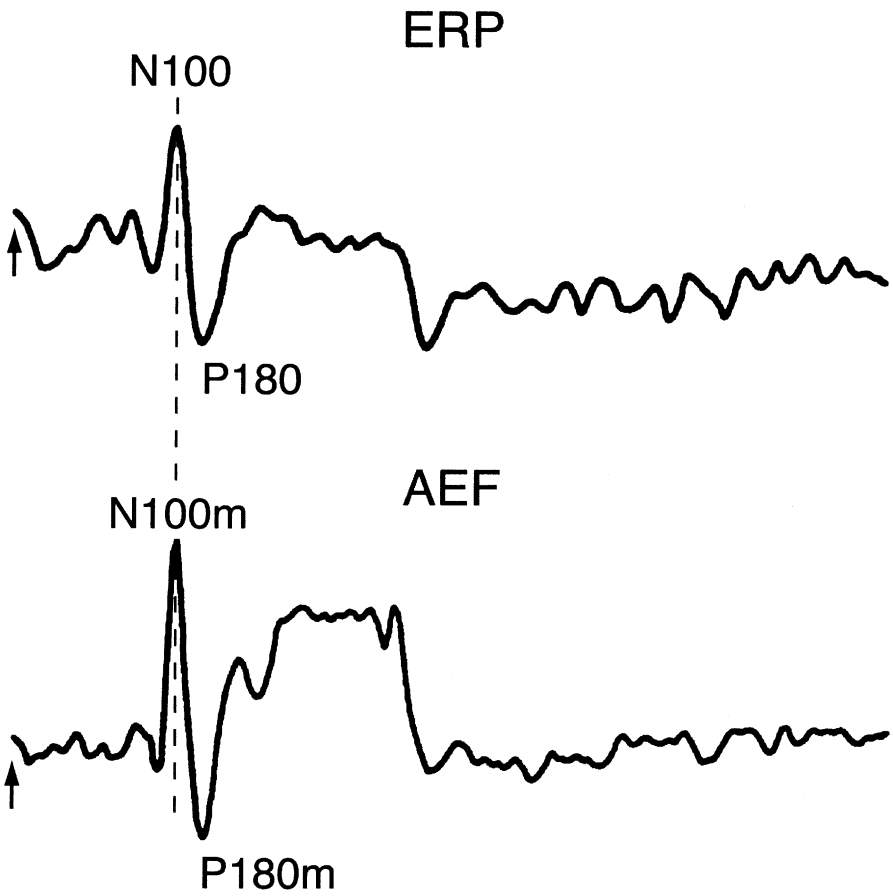


Figure 8 Electromagnetic waves generated by the cortex and recorded on or near the head in response to an acoustic stimulus. Event-related potential (ERP) represents electrical potentials recorded with an electrode in contact with the scalp. The auditory-evoked magnetic field (AEF) is the magnetic field recorded from a detector very near the head. The major component is a middle latency response because it occurs approximately 100 msec after stimulus onset (adapted with permission from Hari, R., *Adv. Audiol.* **6,** 222–282, 1990).

attentional and cognitive processes. Figure 9 (top) shows an example of the ERP obtained when a subject heard the same sound repeatedly and then when a rare, or odd-ball, stimulus was introduced at random times. The major changes associated with the waveform occurred relatively late, after about 200 msec. Similarly, the AES waveforms (Fig. 9, bottom) were obtained under conditions in which the subject was awake but reading and when the subject was attending to the stimulus. Like the ERP, the late components are the ones most affected by this difference in state of attention. These late components are thought to reflect the processing of auditory stimuli.

b. Functional Imaging: PET and fMRI In recent years, the use of noninvasive functional imaging techniques—fMRI and PET—has provided a wealth of new information concerning the location and extent of regions of the human brain that are metabolically activated by acoustic stimulation. Both methods provide indirect evidence of neuronal activation based on the facts that when neurons become more active they require more oxygen and glucose and that this increased demand is met by a compensatory increase in regional cerebral blood flow. Using these indirect imaging methods, investigators have noted that broad regions of the supratemporal plane and lateral superior temporal gyrus are bilaterally activated by a wide

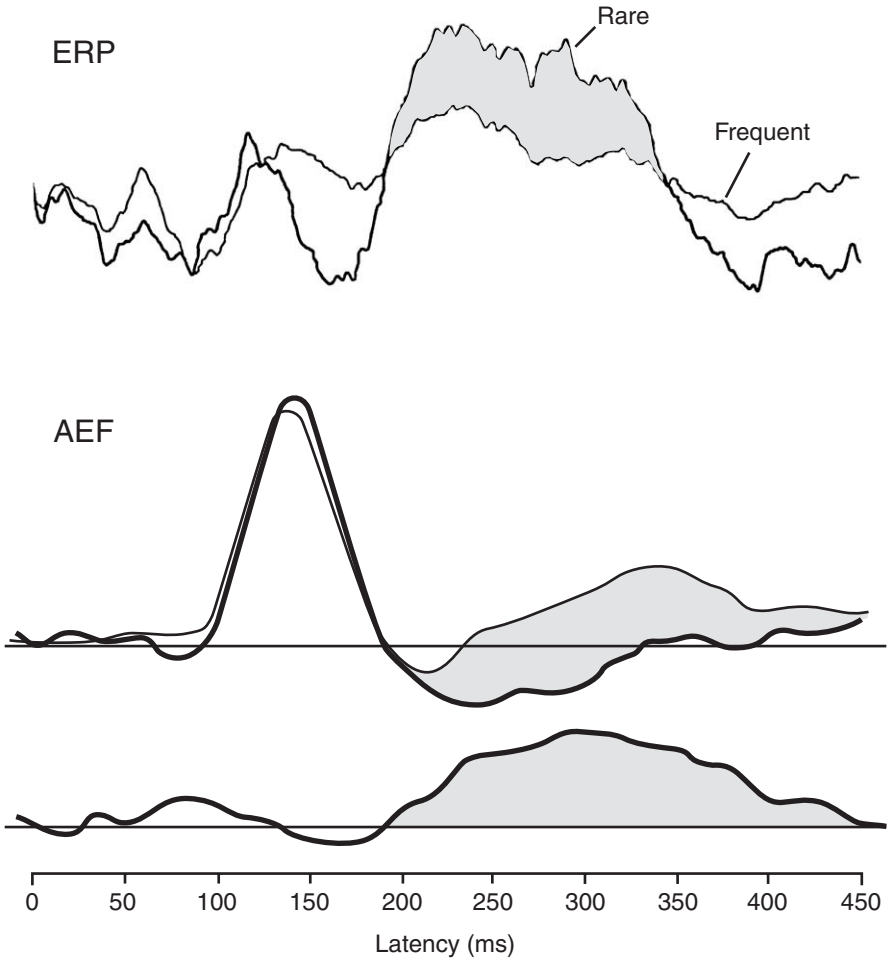


Figure 9 Sound-evoked event-related potential (ERP) and AEF are sensitive to unexpected changes in the sound and to shifts of the listener's attention. (Top) ERPs obtained when the same sound was repeated many times (frequent) and when that stream of repeated sound was interrupted occasionally by a different (rare) sound [adapted with permission from Kraus, N., and McGee, T., *The Mammalian Auditory Pathway*: Neuroanatomy (A. N. Popper and R. R. Fay Eds.). Springer-Verlag, New York, 1992]. (Bottom) The auditory-evoked magnetic field (AEF) obtained when listener attended (counting stimuli) or did not attend (reading) to the stimulus (adapted with permission from Hari, R., *Adv. Audiol*. **6,** 222–282, 1990). The main effect of both activities was seen in the late waves, beyond the 100-msec component.

range of acoustic stimuli, including speech sounds. Tones of different frequency induce localized changes in blood flow within subregions of HG, indicating a tonotopic organization, with higher frequency tones activating the more posterior–medial portion of HG. More complex tasks have been used in conjunction with fMRI to search for evidence of regional functional specialization within auditory cortex. Results from these investigations indicate that the right temporal lobe may play a more important role in the processing of musical stimuli (e.g., pitch and rhythmic temporal patterns) than the left. There have also been reports suggesting functional specialization of right-sided temporal (and parietal) lobe regions in processing information concerning the movement of sound sources. Both fMRI and PET methods have been used extensively to map patterns of cortical activation during speech sound processing.

Inherent limitations of these methods have prevented direct extrapolation of data obtained with these methods to results obtained in nonhuman primates using microelectrode recording techniques. In the future, however, the spatial resolution of fMRI may improve to the extent that the functional organization of human auditory cortex can be studied in far greater detail. Because activity-induced changes in blood flow patterns occur over a period of seconds, it is unlikely that fMRI techniques alone will be capable of delineating the fine temporal patterns of activity that characterize coding and processing of information at the level of auditory cortex.

c. Direct Recording (ECoG) and Stimulation of Auditory Structures in Human Direct electrical-stimulation mapping of the cortex during surgery to relieve medically intractable epilepsy or to remove a tumor is commonly carried out as a way of guiding the surgeon's decision on the location and extent of brain tissue to excise. When the primary auditory field on HG is stimulated, patients report hearing sounds, which are often referred to the ear contralateral to the stimulated cortex. Stimulation of the belt of cortex surrounding the primary field may result in the perception of more complex sounds, although it is now thought that some of this—especially the so-called experiential hallucinations—may be the result of spread of stimulus current to underlying limbic structures. In the caudal region of the superior temporal gyrus (Wernicke's area), and on the angular and supramarginal gyri electrical stimulation may result in the arrest of speech, which is similar to the results of stimulation of the classic Broca's area on the inferior frontal gyrus.

Direct recording from the human auditory cortex has also been carried out in neurosurgical patients both acutely in the operating room and chronically under more controlled experimental conditions. This recording is referred to as the ECoG to differentiate it from the scalp recorded EEG. Results from these experiments show acoustically evoked activity on the superior temporal plane and the lateral surface of the superior temporal gyrus. A field on HG is distinguished from a field on superior temporal cortex based on the HG tonotopic map and the properties of acoustically evoked potentials. Thus, cytoarchitectonic, electrophysiologic, and functional imaging data leave little doubt that the cortex on mesial HG is the primary auditory field of the human and the homolog of AI in nonhuman primates and other mammals. The cortex of the lateral surface of the superior temporal gyrus represents one or more separate auditory fields. Intraoperative electrophysiological studies show that single neurons in this cortex respond vigorously to complex sound, including speech—a finding similar to that in the rhesus monkey.

See Also the Following Articles

AUDITORY CORTEX • AUDITORY PERCEPTION • BRAIN STEM • FOREBRAIN • SENSORY DEPRIVATION • TEMPORAL LOBES • VISION: BRAIN MECHANISMS

Suggested Reading

Crocker, M. J. (Ed.) (1997). *Encylopedia of Acoustics*. Wiley, New York.

Ehret, G., and Romand, R. (1997). *The Central Auditory System*. Oxford Univ. Press, New York.

Geisler, C. D. (1998). *From Sound to Synapse. Physiology of the Mammalian Ear*. Oxford Univ. Press, New York.

Gilkey, R. H., and Anderson, T. R. (Eds.) (1997). *Binaural and Spatial Hearing in Real and Virtual Environments*. Erlbaum, Mahway, NJ.

Harrison, R. V., Kraus, N., Lütkenhöner, B., Rajan, R., and Schreiner, Ch. (Eds.) (1998). *Functional Organization and Plasticity of the Auditory Cortex. Audiology and Neuro-Otology*, **3,** 73–223. Karger, Basel.

Moore, J. K. (1987). The human auditory brain stem: A comparative view. *Hearing Res.* **29,** 1–32.

Moore, J. K. (1994). The human brainstem auditory pathway. In *Neurotology* (R. K. Jackier and D. E. Brachmann, Eds.), pp. 1–17. Mosby, St. Louis.

Nadol, J. B., Jr. (1988). Comparative anatomy of the cochlea and auditory nerve in mammals. *Hearing Res.* **34,** 253–266.

Peters, A., and Jones, E. G. (Eds.) (1985). *Cerebral Cortex: Association and Auditory Cortices*, Vol. 4. Plenum, New York.

Popper, A. N., and Fay, R. R. (Eds.) (1992). *The Mammalian Auditory Pathway: Neurophysiology*, Vol. 2. Springer-Verlag, New York.

Toga, A. W., and Mazziotta, J. C. (2000). *Brain Mapping, The Systems*. Academic Press, San Diego.

Webster, D. B., Popper, A. N., and Fay, R. R. (Eds.) (1992). *The Mammalian Auditory Pathway: Neuroanatomy*, Vol. 1. Springer-Verlag, New York.

Yost, W. A. (1994). *Fundamentals of Hearing. An Introduction*. Academic Press, New York.

Yost, W. A., Popper, A. N., and Fay, R. R. (Eds.) (1993). *Human Psychophysics*, Vol. 3. Springer-Verlag, New York.

Heuristics

RALPH HERTWIG and PETER M. TODD
Max Planck Institute for Human Development, Berlin

GLOSSARY

algorithm A strategy for solving a problem that guarantees solution in a finite number of steps if the problem has at least one solution. An example of a very simple algorithm is that for obtaining temperature on the Fahrenheit scale when the value for the centigrade scale is known: Multiply the known value by 1.8 and add 32.

bounded rationality Principles underlying nonoptimizing adaptive behavior of real information processing systems working under conditions of limited time, information, and computational capacities. Among those principles specified by Herbert Simon is satisficing.

cognitive illusions or biases Systematic deviations of human judgments from rules of probability theory and logic. The occurrence of cognitive illusions is attributed to some judgment heuristics that are often useful but sometimes lead to predictable errors.

heuristic An approximate strategy or "rule of thumb" for problem solving and decision making that does not guarantee a correct solution but that typically yields a reasonable solution quickly or brings one closer to hand.

noncompensatory heuristic A heuristic is noncompensatory if, once it has used a piece of information to make a decision, no further information in any combination can undo or compensate for the effect of the original information. In contrast, a heuristic is compensatory if there is at least one piece of information that can be outweighed by other pieces of information. A compensatory strategy integrates at least some of the available information and makes trade-offs between the relevant pieces of information to form an overall evaluation of each of the available alternatives or options.

satisficing According to Herbert Simon, satisficing is using experience to construct an expectation (or aspiration level) of a reasonable solution to some problem and stopping the search for solutions as soon as one is found that meets the expectation.

unbounded rationality Decision-making strategies that have no regard for constraints of time, knowledge, or computational capacities. Modern mainstream economic theory is largely based on unbounded rationality models that portray economic agents as fully rational Bayesian maximizers of subjective utility.

Many decisions faced by people cannot be made in an optimal way because optimal solutions may take too much computation to find or may not even exist. Instead, real decision makers must often take shortcuts and use heuristics that yield reasonable solutions in a reasonable amount of time, even if they do not guarantee always reaching a good decision. These heuristics are thus an essential aspect of human intelligence, leading to adaptive behavior despite the challenging conditions of limited time, knowledge, and computational capacity under which people have to solve problems. Heuristics are most commonly studied in psychology, particularly within the domains of judgment and decision making, and in computer-based applications in artificial intelligence and operations research. This article focuses on research in psychology that has proposed heuristic models of how people search for information and make decisions and choices.

I. A SHORT HISTORY OF THE CONCEPT "HEURISTIC"

Recent research on decision heuristics descends from earlier schools of thought. For this reason, understanding current thought can be aided by first considering the history of the concept. In 1905, the 26-year-old Albert Einstein published his first fundamental paper in quantum physics, titled "On a Heuristic Point of View Concerning the Production and Transformation of Light." In that Nobel prize-winning paper, Einstein used the term heuristic to indicate that he considered the view he presented therein as incomplete, even false, but still useful. Einstein could not wholeheartedly accept the quantum view of light that he started to develop in this paper, but he believed that it was of great transitory use on the way to building a more correct theory. As used by Einstein, a heuristic (a term of Greek origin meaning "serving to find out or discover") is an approach to a problem that is necessarily incomplete given the knowledge available, and hence unavoidably false, but that is useful nonetheless for guiding thinking in appropriate directions.

A few decades later, Max Wertheimer (a close friend of Einstein's), Karl Duncker, and other Gestalt psychologists spoke of heuristic reasoning, but with a slightly different meaning from that of Einstein. Gestalt psychologists conceptualized thinking as an interaction between inner mental processes and external problem structure. In this view, heuristic methods such as "looking around" and "inspecting the problem" are first used to guide the search for appropriate information in the environment, which is then restructured or reformulated by inner processes.

Heuristic methods also play a prominent role in George Pólya's approach to mathematical problem solving. According to the Hungarian mathematician, effective problem solving consists of four main phases—understanding the problem, devising a plan, carrying out the plan, and looking back—all of which can incorporate heuristics. Devising a plan, for instance, can include heuristic methods such as "examine a simpler or special case of the problem to gain insight into the solution of the original problem," "work backward," or "identify a subgoal."

In the 1950s and 1960s, Herbert Simon and Allen Newell started to develop heuristics for searching for solutions to problems. They replaced the somewhat vague notion of heuristic reasoning of the Gestalt school and of Pólya's with much more precise computer-based models (e.g., in the General Problem Solver system) of human problem solving and reasoning largely based on the means–ends analysis heuristic. This heuristic found some way to reduce the distance between the current state and the goal state. With the advent of information processing theory in cognitive psychology, a heuristic came to mean a useful shortcut, an approximation, or a rule of thumb for guiding search through a space of possible solutions, such as a strategy that a chess master uses to explore the enormous number of possible moves at each point in a game.

Such general-purpose or "weak" methods as the means–ends analysis heuristic, however, proved insufficient to deal with problems other than artificial and well-defined mathematical problems or the games of chess and cryptoarithmetic that Newell and Simon investigated. As a consequence, research in artificial intelligence (AI) in the 1970s turned to collecting domain-specific rules of thumb from specialists in a particular field and incorporating these into expert systems. At approximately the same time, mathematicians working in operations research began dealing with new results from computational complexity theory indicating that efficient algorithmic solutions to many classes of challenging combinatorial problems (such at the traveling salesman problem) might not be found; as a consequence, they too turned to the search for problem-specific heuristics, although through invention rather than behavioral observation.

In psychology after 1970, researchers became increasingly interested in how people reason about unknown or uncertain aspects of real-world environments. The research program that spurred this interest was the heuristics-and-biases program initiated by Amos Tversky and Daniel Kahneman. This program's research strategy has been to measure human decision making against various normative standards taken from probability theory and statistics. Based on this strategy two major results about people's reasoning under uncertainty emerged: a collection of violations of the normative standards (that in analogy to perceptual illusions are often called "cognitive illusions" or "biases") and explanations of these illusions in terms of a small number of cognitive heuristics. According to Kahneman and Tversky, people rely on a limited number of heuristics—most prominently representativeness, availability, and anchoring and adjustment—that often yield reasonable judgments but sometimes lead to severe and systematic biases. Diverging from earlier usage, the term heuristics now gained a different connotation: fallible cognitive shortcuts that people often use when faced with

uncertainty and that can lead to systematic biases and lapses of reasoning indicating human irrationality. This more negative view of heuristics—and of the people who use them as "cognitive misers" using little information or cognition to reach biased conclusions—has spread to many other fields, including law, economics, medical decision making, sociology, and political science.

Recently, however, a new appreciation is emerging that heuristics may be the only available approach to decision making in the many problems for which optimal logical solutions do not exist (as researchers in operations research realized). Moreover, even when exact solutions do exist, domain-specific decision heuristics may be more effective than domain-general logical approaches, which are often computationally infeasible (as AI found). This has led to research programs such as the study of ecological rationality by Gigerenzer, Todd, and colleagues. Their program focuses on precisely specified computational models of heuristics and how they are matched to the ecological structure of particular decision environments. It also explores the ways that learning and evolution can achieve this match in human behavior, something that has already been widely accepted for other animals in research on rules of thumb in behavioral ecology.

In the following sections, the focus is on research in psychology exploring heuristics proposed to model how people search for information and make decisions and choices. Researchers such as Payne, Bettman, and Johnson and Svenson have been concerned with psychological heuristics for preferences, but here we are mostly concerned with inference heuristics. Heuristics and shortcuts are also important in human perception and higher order reasoning processes (e.g., hypothesis testing), planning and problem solving, as well as in computer applications in these domains.

II. UNBOUNDED RATIONALITY VERSUS THE BOUNDED REALITY OF HUMAN DECISION MAKING

Both the heuristics-and-biases program and the recently emerging work on ecologically rational heuristics have been linked to Herbert Simon's notion of bounded rationality. This concept can be understood by contrasting it to the traditional decision-making approach embodied in unbounded rationality, illustrated by the following example. Imagine being faced with the decision of whether or not to marry. How can this decision be made in a rational way? Assume that you attempted to resolve this question by maximizing your subjective expected utility. To compute your personal expected utility for marrying, you would have to determine all the possible consequences that marriage could bring (e.g., children, companionship, and countless further consequences), attach quantitative probabilities to each of these consequences, estimate the subjective utility of each, multiply each utility by its associated probability, and finally add all these numbers. The same procedure would have to be repeated for the alternative "not marry." Finally, you would have to choose the alternative with the higher total expected utility.

Maximization of expected utility in this way is probably the best known realization of the prominent vision of unbounded rationality. Models of unbounded rationality have been criticized for having little or no regard for the constraints of time, knowledge, and computational capacities that real humans face. For instance, while you are deliberating about whether marrying is the right choice, considering each of the myriad conceivable consequences and assigning probabilities to each, any potential partner will probably have married someone else. To this criticism proponents of unbounded rationality generally concede that their models assume unrealistic mental abilities, but they nevertheless defend them by arguing that humans act as if they were unboundedly rational. In this interpretation, the models of unbounded rationality do not describe the process but merely the outcome of reasoning.

If the lofty ideals of human reasoning do not capture the processes of how real people make decisions in the real world, what then are those processes? In other words, what models take into account the challenging conditions under which people have to solve problems, including limited time, knowledge, and computational capacity? Herbert Simon proposed that these constraints force humans to use "approximate methods" (heuristics) to handle most tasks. These approximate methods form the basis of bounded rationality.

Simon's vision of bounded rationality has two interlocking components that act like a pair of scissors to shape human rational behavior. The two blades in this metaphor are the computational capabilities of the actor and the structure of task environments. First, the computational capability blade implies that models of human judgment and decision making should be built on what we actually know about the mind's limitations

rather than on fictitious competencies assumed in models of unbounded rationality. There are two key limitations central to bounded rationality. First, contrary to models of unbounded rationality, humans cannot search for information for all of eternity. In computationally realistic models, search must be limited because real decision makers have only a finite amount of time, knowledge, attention, or money to apply to a particular decision. Limited search requires rules to specify what information to seek and in what order (i.e., an information search rule) and a way to decide when to stop looking for information (i.e., a stopping rule).

Another key limitation of the human mind is that the pieces of information uncovered by the search process are not likely to be processed in an overly complex way. In contrast, most traditional models of inference, from linear multiple regression models to Bayesian models to neural networks, try to find some optimal integration of all information available: Every bit of information is taken into account, weighted, and combined in some more or less computationally expensive way. Models of bounded rationality instead rely on processing steps that are computationally bounded. For instance, a bounded decision or inference can be based on only one or a few pieces of information, whatever the total amount of information found during search. The simple decision rule used to process this limited knowledge need not weigh or combine pieces of information, thus eliminating the need to convert different types of information into a single common currency (e.g., utilities). Note that decision rules and information search and stopping rules are connected. For instance, when a heuristic searches for only one (discriminating) cue, this largely constrains the possible decision rules to those that do not integrate information. On the other hand, if search extends to many cues, the decision rule will be less constrained. The cues may then be weighted and integrated, or only the best of them may determine the decision.

These two key limitations, limited information search and limited processing of information, can be instantiated into models of heuristics. The limitations help explain how heuristics achieve one of their most important advantages, namely, speed. In fact, for much decision making in the real world—the stock broker who decides within seconds to keep or sell a stock, or the captain of the firefighter squad who within a few moments must predict how a fire will progress and whether or not to pull out the squad—speed is often the crucial objective.

The second blade in Simon's scissors metaphor, operating in tandem with computational capability, is environmental structure. This blade is of crucial importance in shaping bounded rationality because it can explain when and why heuristics perform well, namely, if the structure of the heuristic is adapted to the structure of the environment (i.e., if the heuristic is ecologically, rather than logically, rational). Simon's classic example concerns foraging organisms that have a single need—food. An organism living in an environment in which little heaps of food are randomly distributed can survive with a simple heuristic: Run around randomly until a heap of food is found. For this, the organism needs some capacity for movement, but it does not need a capacity for inference or learning. For an organism in an environment in which food is distributed not randomly but in patches whose locations can be inferred from cues, more sophisticated strategies are possible. For instance, it could learn the association between cues and food and store this information in memory. The general point is that in order to understand which heuristic an organism employs, and when and why the heuristic works well, one needs to examing the structure of the information in the environment.

III. HEURISTICS FOR HUMAN JUDGMENT, CHOICE, AND SEARCH BEHAVIOR

The models of heuristics for human judgment, choice, and decision making that have been proposed in psychology since 1970 can be linked to two traditions of heuristics with earlier beginnings as described previously. These two traditions have employed different levels of description. First, following the line of the heuristic methods studied by the Gestalt psychologists, one class of heuristics consists of psychological principles that are verbally described. These models are only relatively loosely specified and usually do not explicate all the processes they involve (e.g., in terms of information search, stopping, and decision rules). Second, following from Simon and Newell's computer-based models of human decision making, another class of heuristics has been formulated as process models, with explicit specification of the processes involved. Because of this explication, the latter heuristics can be both mathematically analyzed and tested with the help of computer simulations. We consider each class of heuristics in turn.

A. Heuristics for the Judgment of Probability and Frequencies: Availability, Representativeness, and Anchoring and Adjustment

The heuristics most widely studied within psychology are those that people use to make judgments or estimates of probabilities and frequencies in situations of uncertainty (i.e., in situations in which people lack exact knowledge). Most prominent among these are the availability, representativeness, and anchoring and adjustment heuristics.

The availability heuristic leads one to assess the frequency of a class or the probability of an event by the number of instances or occurrences that can be brought to mind or by how easy it seems to call up those instances. For instance, which class of words is more common: seven-letter English words of the form "_ _ _ _ _ n _" or the form "_ _ _ _ i n g"? According to the availability heuristic, to estimate the frequency of occurrences people draw a sample of the events in question from memory. Specifically, for this case they retrieve words ending in –ing (e.g., "jumping") and retrieve words with "n" in the sixth position (e.g., "raisins") and then count the number of words retrieved in some period or assess the ease with which such words could be retrieved. They then answer that the more numerous or easier class of words is more common. Because people find it easier to think of words ending with –ing than to think of words with the letter "n" in the next-to-last position, they usually estimate the class "_ _ _ _ i n g" to be more common. This judgment, however, is wrong because all words ending with –ing also have "n" in the sixth position; in addition, there are seven-letter words with "n" the sixth position that do not end in –ing.

The availability heuristic has been suggested to underlie diverse judgment errors, ranging from the tendency to overestimate how many people die from some specific causes of death (e.g., tornado) and underestimate the death toll of others causes (e.g., diabetes) to why people's answers to life satisfaction questions ("How happy are you?") may be overly influenced by events that are especially memorable.

The representativeness heuristic has been proposed as a means to assess the probability that an object A belongs to a class B (e.g., that a person described as meek is a pilot) or that an event A is generated by a process B (e.g., that the sequence HTHTHT was generated by randomly throwing a fair coin). This heuristic produces probability judgments according to the extent that object A is representative of or similar to the class or process B (e.g., meekness is not representative of pilots, so a meek person is judged as having a low probability of being a pilot). This heuristic can lead to errors because similarity or representativeness judgments are not always influenced by factors that should affect judgments of probability, such as base rates. The representativeness heuristic has also been evoked to explain numerous judgment phenomena, including "hot hand" observations in basketball (the belief that a player is more likely to score again after he or she already scored successfully than after missing a shot) and the gambler's fallacy (the belief that a successful outcome is due after a run of bad luck).

Another heuristic, anchoring and adjustment, produces estimates of quantities by starting with a particular value (the anchor) and adjusting upward or downward from it. For instance, people asked to quickly estimate the product of either $8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$ or $1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7 \times 8$ give a higher value in the former case. According to the anchoring and adjustment heuristic, this happens because the first few numbers presented are multiplied together to create a higher or lower anchor, which is then adjusted upwards in both cases, yielding a higher final estimate for the first product.

Although it has been pointed out that availability, representativeness, and anchoring and adjustment are quite useful heuristics (because they often lead to good judgments without much time or mental effort), most of the large body of evidence amassed that is consistent with the use of these heuristics comes from studies showing where they break down and lead to cognitive illusions or biases (i.e., deviations from some normative standards). This heuristics-and-biases research program has caught the attention of numerous social scientists, including economists and legal scholars. There are good reasons for this attention, since systematic biases question the empirical validity of classic rational choice models (i.e., models of unbounded rationality) and may have important economic, legal, and other implications.

However, the exclusive focus on cognitive illusions has evoked the criticism that research in the heuristics-and-biases tradition equates the notion of bounded rationality with human irrationality and portrays the human mind in an overly negative light, with some researchers even arguing that cognitive illusions are the rule rather than the exception. It has also been criticized that, to date, the cognitive heuristics posited have not been precisely formalized such that one could either simulate or mathematically analyze their

behavior, leaving them free to account for all kinds of experimental performance in a post hoc fashion. For instance, it is still an open question of how people assess similarity to make probability judgments with the representativeness heuristic or how many items (e.g., words ending with –ing) the availability heuristic retrieves before it affords a frequency estimate of a class of object (albeit theoretical progress has been made, for instance, by testing whether availability works in terms of ease of recall or number of items recalled). Moreover, the heuristics-and-biases program focuses on human computational capabilities (the first blade of Simon's scissors), largely ignoring the role of the environment by not specifying how such heuristics capitalize on information structure to make inferences. Finally, this program appears to consider heuristics as dispensable mechanisms (that would not be needed if people had the right tools of probability and logic to call on), in contrast to Simon's view of indispensable heuristics as the only available tools for solving many real-world problems.

Kahneman and Tversky have countered some of this critique by drawing a parallel between their heuristic principles and the qualitative principles of Gestalt psychology—the latter being still valuable despite not being precisely specified. Irrespective of the various criticisms, the heuristic and biases program has undoubtedly led to a tremendous amount of research into the idea that people rely on cognitive heuristics made up of simple psychological processes rather than on complex procedures to make inferences about an uncertain world. As a result, this insight has been firmly established as a central topic of psychology.

B. Fast and Frugal Choice Heuristics

More precisely specified models of heuristics have been studied by a another research program that emerged in psychology in the 1990s. This new program considers fast and frugal heuristics for making decisions as the way the human mind can take advantage of the structure of information in the environment to arrive at reasonable decisions. Thus, it focuses on how mental capabilities and structured environments together can lead to accurate and useful inferences rather than focusing on the cases in which heuristics may account for poor reasoning. Most of the fast and frugal heuristics that Gigerenzer, Todd, and colleagues have proposed model the way humans make choices rather than probability judgments (a few others deal with additional tasks, such as estimation and classification).

Many of the choices humans make involve an inference or prediction about which of two objects will score higher on a criterion: Which soccer team will win? Which of two cities has a higher crime rate or higher cost of living? Which of two applicants will do a better job? When making such choices, we may have different amounts of information available. In the most limited case, if the only information available is whether or not each option has been encountered before, the decision maker can do little better than rely on his or her own partial ignorance, for instance, choosing recognized options over unrecognized ones. This may not sound like much for a decision maker to go on, but there is often information implicit in the failure to recognize something, and this failure can be exploited.

This kind of "ignorance-based" decision making is embodied in the recognition heuristic. This heuristic states that, when choosing between two objects (according to some criterion), if one is recognized and the other is not, then select the former. For instance, if predicting whether Manchester United or Bayer Leverkusen will win the European Soccer Champions League, this heuristic would lead most of us to bet on Manchester United. Why? European soccer teams are often named after European cities (e.g., Arsenal London or AC Milano), and people who are ignorant of the quality of European soccer teams can still use city recognition as a cue for soccer team performance. Cities with successful soccer teams are likely to be large, and large cities are likely to be recognized; hence, Manchester, which is more than two times as large as Leverkusen, is also more likely to be recognized and thus chosen as the winner.

The recognition heuristic will yield good choices more often than would random choice in those decision environments in which exposure to different possibilities is positively correlated with their ranking along the decision criterion being used. Animals also behave as if they apply similar rules: Norway rats, for instance, prefer to eat things they recognize through past experience with other rats (e.g., items they have smelled on the breath of others) over novel items.

Employing the recognition heuristic can lead to a surprising phenomenon called the less-is-more effect. This is the analytical and empirical observation that an intermediate amount of (recognition) knowledge about a set of objects can yield the highest proportion of correct answers—knowing (i.e., recognizing) more than this will actually decrease the decision-making performance. A context in which this effect appears in the reasoning of people is judgments about

demographics. When American students were asked to pick the larger of two cities, they scored a median 71% correct inferences when city pairs were randomly constructed from the 22 largest U.S. cities and 73% when city pairs were from the 22 largest cities in Germany. The result is counterintuitive when viewed from the premise that more knowledge is always better: The students knew a lifetime of facts about U.S. cities that could be useful for inferring population, but they knew little or nothing beyond mere recognition about the German cities—and they did not even recognize about half of them. The latter fact, however, allowed them to employ the recognition heuristic and to pick German cities that they recognized as larger than those they did not. This heuristic could not be applied for choosing between U.S. cities, however, because the students recognized all of them and thus had to rely on additional retrievable information.

In choosing one of two options, most of the time we have more information than just a vague memory of recognition to go on. The situation of American students comparing American cities is just one example. When multiple pieces of information (or "cues") are available for guiding decisions, how can a heuristic that limits search and processing of information proceed? A decision maker following the dictums of unbounded rationality would collect all the available information, weight it appropriately, and combine it optimally before making a choice. A more frugal approach is to use a stopping rule that terminates the search for information as soon as enough has been gathered to make a decision. One such approach is to rely on "one-reason decision making": Stop looking for cues as soon as one is found that differentiates between the two options being considered. This allows the decision maker to follow a simple loop, depicted in Fig. 1: (i) Select a cue dimension; (ii) search for the corresponding cue values of each alternative; (iii) if they differ, then stop and choose the alternative for which the cue indicates a greater value on the choice criterion; (iv) if they do not differ, then return to the beginning of this loop to search for another cue dimension.

This four-step loop specifies a stopping rule (stopping after a single cue is found that enables a choice between the two options) and a decision rule (deciding on the option to which the one cue points). Depending on how cue dimensions are searched for in the first step, (i.e., depending on what kind of specific information search rule the heuristic uses), different one-reason decision-making heuristics can be formed. The Take The Best heuristic searches for cues in the order of their ecological validity (i.e., their correlation with the decision criterion). Take The Last searches for cues in the order determined by their past success so that the cue that was used for the most recent previous decision is checked first during the next decision. Finally, the Minimalist heuristic selects cues in a random order.

For example, consider the task of inferring which of two cities in the United States has a higher homelessness rate. Assume that possible cues are "rent control," "temperature," "unemployment," and "public housing," each turned into a binary value (0 or 1) according to whether the actual value is below or above the median for U.S. cities. Rent control has the highest ecological validity, temperature has the second highest, and so on. The Minimalist heuristic only needs to know in which direction a cue "points." For instance, the heuristic needs to know only whether warmer or cooler weather indicates a city with a higher rate of homelessness (in the United States, warmer weather is indeed associated more often with higher homelessness rates than with lower rates). The strategy of Minimalist is to search for cues in random order,

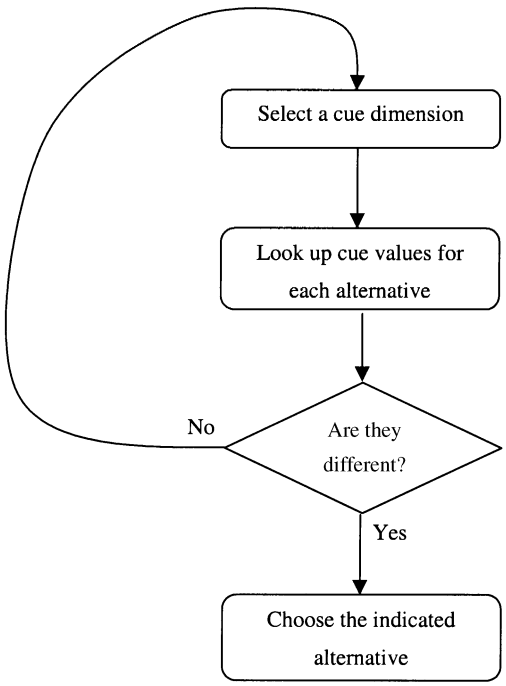


Figure 1 A flowchart of one-reason decision making. First, select a cue dimension and ascertain the corresponding cue values for each alternative; next, check whether the values for that cue discriminate between the alternatives. If so, then choose the indicated alternative; if not, select another cue dimension and repeat this process. Random choice can be used if no more cues are available.

stop cue search when a cue is found that discriminates between the two cities, and then choose the city that has the cue value 1 when the other city has cue value 0. For instance, when inferring whether New York or Los Angeles has a higher homelessness rate, the unemployment cue might be the first cue randomly selected, and the cue values are found to be 1 for both cities. Because this cue does not discriminate between the cities, search is continued, the public housing cue is randomly selected, and the cues values are 0 for New York and 1 for Los Angeles. Search is stopped at this discriminating cue and the inference is made that Los Angeles has a higher homelessness rate, as it indeed does.

The Take The Best heuristic is exactly like Minimalist except that it considers cues in order of their validity from highest to lowest. If the highest validity cue does not discriminate, the next best cue is tried, and so forth. Thus, Take The Best differs from Minimalist only in the information search rule, but it has the same stopping and decision rule. Take The Best (unlike the Minimalist, Take The Last, and recognition heuristics) is an instance of the class of lexicographic decision strategies. This term signifies that the cues are looked up in a fixed order of validity, and the first cue where choices differ is used alone to make the decision, like the alphabetic order used to arrange words in a dictionary. The Arabic number system is also lexicographic. To determine which of two numbers with equal digit length is larger, one has to start by examining the first (leftmost) digit: If this digit is larger in one of the numbers, the whole number is larger. If they are equal, one has to examine the second digit, and so on (a simple method that is not possible for Roman numbers). There is growing empirical evidence that people actually use lexicographic heuristics such as Take The Best, particularly when time is limited.

How well do these one-reason decision heuristics perform? Table I compares the performance of three fast and frugal heuristics (Minimalist, Take The Best, and Take The Last) to that of multiple regression and Dawes's and Franklin's rule. Unlike the heuristics, multiple regression is a computationally expensive linear strategy that calculates weights that reflect the covariances between predictors or cues. When the task is merely fitting the given data set, multiple regression is the most accurate strategy, by two percentage points, followed by Take The Best. However, when the task is to generalize from a training set to a test set, a simple heuristic such as Take The Best can outperform multiple regression (note that multiple regression has all the information Take The Best uses and more). The reason is that by being simple, the heuristics can avoid being too matched to any particular environment—that is, they can escape the curse of overfitting.

Table I

Performance of Three Fast and Frugal Heuristics (Take The Best, Minimalist, and Take The Last) and Three Linear Strategies (Dawes's rule, Franklin's rule, and multiple regression) Averaged across 20 Empirical Data Sets[a]

Heuristic/strategy	Frugality	Accuracy	
		Fitting (% correct)	Generalization (% correct)
Take The Best	2.4	75	71
Minimalist	2.2	69	65
Take The Last	2.1	70	65
Franklin's rule	7.7	75	71
Dawes's rule	7.7	73	69
Multiple regression	7.7	77	68

[a]The average number of predictiors available in the 20 data sets was 7.7. Frugality indicates the mean number of cues actually used by each strategy. Accuracy indicates the percentage of correct answers achieved by the heuristics and strategies when fitting data (i.e., fit a strategy to a given set of data) and when generalizing to new data (i.e., use a strategy to predict new data).

Overfitting refers to the problem of a model that is closely matched to one situation (set of data) failing to predict accurately in another similar situation (another set of data). This phenomenon can arise from assuming that every detail in a given environment is of great relevance. Consider forecasting of the U.S. presidential elections as an example. Beyond traditional variables such as incumbency and the state of the election-year economy, a plethora of additional variables have been suggested as predictors of recent U.S. presidential elections, including the voting behavior in Okanogan County (a rural stretch of north-central Washington), the rise or fall of women's hemlines, and the height of the candidates. General strategies such as multiple regression can in fact incorporate each of these and many more variables into the unlimited collection of free variables in their forecast models. As accurate as such parameter-laden forecast models may be for describing particular recent presidential elections, their accuracy in predicting other situations (e.g., earlier U.S. presidential elections or elections in other locations) may well be minimal. That is, these models can easily overfit the particular (training) data set and thereby fail to generalize to the new (testing) data set. In contrast, if a forecast model uses many

fewer parameters, for instance, just incumbency and height of the candidates (which predicted the winner of every election since World War II, except in 1976 and 2000), it is likely to avoid overfitting and thereby generalize better to new situations.

Fast and frugal heuristics (like lexicographic strategies) are noncompensatory, meaning that once they have used a single cue to make a decision no further cues in any combination can undo or compensate for that one cue's effect. When the information in the decision environment is structured in a matching noncompensatory fashion (i.e., the importance or validity of cues decreases rapidly such that each weight of a cue is larger than the sum of all weights to come, e.g., one-half, one-fourth, one-eighth, and so on), the Take The Best heuristic can exploit that structure to make correct decisions as often as compensatory rules. Take The Best also performs comparatively well when information is scarce; that is, when there are many more objects than cues to distinguish them. Further research is needed to explore what environment structures can be exploited by different fast and frugal heuristics.

C. Heuristics for Multiple Alternative Choices

Not all choices in life are presented to us as convenient pairs of alternatives, of course. Often, we must choose between several alternatives, such as which restaurant to go to, which apartment to rent, or which stocks to buy. Table II lists various decision heuristics that have been proposed in the psychological literature for choosing one out of several alternatives, where each alternative is characterized by cue (or attribute) values and where the importance of a cue is specified by its weight (or validity). This collection is not exhaustive, and it only focuses on heuristics for inference rather than preference (albeit some could be applied to preferences as well). However, the heuristics represent a wide range of different information search, stopping, and decision rules. Among the heuristics for multiple alternative choice, lexicographic (LEX), lexicographic semiorder (LEX-Semi), and elimination by aspects (EBA) are noncompensatory, whereas the rest are compensatory heuristics, integrating (at least some of) the available information and making trade-offs

Table II
Description of Various Multiple Alternative Choice Heuristics[a]

Heuristic	Description
Franklin's rule or weighted additive rule	Calculates for each alternative the sum of the cue values multiplied by the corresponding cue weights (validities) and selects the alternative with the highest score.
Dawes's rule or additive rule	Calculates for each alternative the sum of the cue values (discretized to either 1 or -1) and selects the alternative with the highest score.
Good features	Selects the alternative with the highest number of good features: a good feature is a cue value that exceeds a specified cut-off.
Weighted pros	Selects the alternative with the highest sum of weighted "pros." A cue that has a higher value for one alternative than for the others is considered a pro for this alternative. The weight of each pro is defined by the validity of the particular cue.
LEX or lexicographic	Selects the alternative with the highest cue value on the cue with the highest validity. If more than one alternative has the same highest cue value, then for these alternatives the cue with the second highest validity is considered, and so on.
LEX-Semi or lexicographic semiorder	Works like LEX, with the additional assumption of a just-noticeable difference. Pairs of alternatives with less than a just-noticeable difference between the cue values are not discriminated.
EBA or elimination by aspects	Eliminates all alternatives that do not exceed a specified value on the first cue examined. If more than one alternative remains, another cue is selected. This procedure is repeated until only one alternative is left. Each cue is selected with a probability that is proportional to its weight.
LEX-Add or lexicographic additive combination	Represents a combination of two strategies. It first uses LEX-Semi to choose two alternatives as favorites and then evaluates them by Dawes's rule and selects the one with the highest sum.

[a]Defined in terms of alternatives (options), cues (information), and weights (importance of information).

between the relevant cues to form an overall evaluation of each alternative.

Weighing and summing of all available information has been used to define rational judgment at least since the Enlightenment: The concepts of expected value and utility, Benjamin Franklin's moral algebra, and Homo economicus all rely on these two fundamental processes. The heuristics for multiple alternative choices in Table II can be seen as various shortcuts of these two processes. Dawes's rule, for instance, questions the importance of precise weighting. In the 1970s and 1980s, Robyn Dawes and colleagues showed that tallying information (cues) in terms of simple unit weights, such as $+1$ and -1, typically led to the same predictive accuracy as the "optimal weights" in multiple regression (particularly when generalizing to new data). Thus, in situations in which the task is to predict what is not yet known (rather than to fit what is already known), weighting information does not seem to matter much, as long as one gets the sign right.

On the other hand, LEX (a generalization of Take The Best), LEX-Semi, and EBA do not require summing procedures. All three heuristics use a simple form of weighting by ordering the cues, but they do not sum the cues. Gigerenzer and colleagues collected counterintuitive evidence that this simple weighting without summing (as in the Take The Best heuristic) can be as accurate and in some circumstances (e.g., generalization) even more accurate than complex decision strategies such as multiple regression.

Among the choice heuristics listed in Table II, EBA, proposed by Amos Tversky, is the most widely known elimination model in psychology. In sequential elimination choice models, one alternative is chosen from a set of possibilities by repeatedly eliminating subsets of alternatives from further consideration until only a single choice remains. One of the motivating factors in developing EBA in particular as a descriptive model of choice was that there are often many relevant cues that may be used in choosing among complex alternatives. EBA deals with this challenge by probabilistically considering successive cues (which are chosen with a probability proportional to their importance), selecting one at a time, and eliminating all the alternatives that do not possess this current cue, until a single alternative remains as the final choice. Other elimination heuristics select cues in a different manner (e.g., deterministically or based on validity) or use them to process the alternatives in other ways (e.g., in terms of thresholds rather than presence or absence).

D. Heuristics for Sequential Search

The heuristics discussed so far for choosing one option from many operate with the assumption that all the possible options (e.g., cities to choose between) are presently available to the decision maker. In many real-world choice problems, though, an agent encounters options in a sequence spread out over time. The options typically appear in random order and are drawn from a distribution with parameters that are only partially known in advance. In this case, the search for possible options, rather than just for information about those already present, becomes central.

The traditional normative approach to such problems is to search until one finds an option below a precalculated reservation price that balances the expected benefit of further search against its cost; this requires full knowledge of the search costs and the distribution of available alternatives. Heuristics that simplify the reservation price calculation (by replacing an integral with a weighted sum) can come very close to normative performance (e.g., at selecting good prices during a shopping trip to several stores). Other heuristics require less knowledge, such as "Keep searching until the total search cost exceeds 7.5% of the best price found." Herbert Simon's bounded rationality principle of satisficing suggests setting an aspiration level equal to an alternative that is good enough for the decision maker's needs (rather than optimal) and searching until that aspiration is met. Exactly how the aspiration level can be set varies with the search setting (e.g., whether it is a one-sided search such as shopping or a two-sided mutual search such as finding a mate). Finally, another type of search heuristic that people use stops search after a particular pattern of alternatives is encountered rather than after some threshold is exceeded (despite the fact that pattern should not matter from a normative perspective). For instance, the "one-bounce" and "two-bounce" rules state that one should keep searching for a low price until prices go up for the last or two last alternatives, respectively.

E. Social Decision Heuristics

Decision-making mechanisms can exploit the structure of information in the environment to arrive at better outcomes. The most important aspects of an agent's environment are often created by the other agents with which it interacts. Two of the key problems social agents face are the questions of how to (fairly)

divide up resources among one another and how to make cooperative decisions in situations in which the pursuit of self-interest by each agent would lead to a poor outcome for all. We consider each of these problems in turn.

The task of fairly dividing up resources is ubiquitous, ranging from distributing a cake among siblings to dividing an estate among heirs and splitting a fixed budget among a group of faculty members at an academic institution. Although there are a plethora of fair-division procedures, Brams and Taylor classified them according to a few dimensions, such as the number of players to which they are applicable (n=2, 3, 4, or more), the properties they satisfy (e.g., proportionality, envy-freeness, and efficiency), and whether or not the division has to be exact or only approximate. A simple but well-known decision heuristic that may be familiar to many parents is "one divides, the other chooses." Although this heuristic stipulates division of labor, it does not specify how the person who divides the resource (i.e., a cake) actually does it. If, however, the person who divides the cake understands the strategic interest of the other, the implied division rule is to divide up the resource such that one is indifferent between the two parts; the other person will then choose whatever he or she considers to be the larger piece. This way each person is assured of getting what he or she perceives to be at least half the resource, and neither party thinks that the other received a larger piece of cake. A fair-division procedure with these properties is said to be proportional and envy-free.

An example of a situation in which the pursuit of self-interest by each party leads to a poor outcome for all is that in which two industrialized regions of the world (e.g., the United States and the European Union) have established trade barriers to each other's exports. Because of the mutual advantages of free trade, both regions would be better off if these barriers were eliminated. However, if either region were to unilaterally give up the barriers, it would be faced with terms of trade that hurt its own economy. In fact, no matter what America does, the European Union (EU) is better off retaining its own trade barriers and vice versa. This strategic situation, in which the incentive to retain trade barriers for both regions produces a worse outcome than would have been possible had both decided to cooperate, is known as a prisoner's dilemma game. This game is just one among many situations that game theorists examine in order to analyze and model the strategic interactions of social agents.

If there is some likelihood that the players will encounter each other in the future, as in trade between the United States and the EU, the interaction become an iterated prisoner's dilemma game. There are a number of possible decision heuristics for this situation. A particularly simple but surprisingly successful decision heuristic is the tit-for-tat heuristic that Anatol Rapoport submitted to Axelrod's famous computer tournament. Given the possibility of cooperating with or defecting against the other player at each time step, tit-for-tat starts with a cooperative choice and thereafter does what the other player did on the previous move. In other words, tit-for-tat searches for a minimal amount of information (the counterpart's behavior in the last round) and cooperates if the last move was cooperative but defects if the last move was defective. Thus, akin to some of the heuristics described previously, tit-for-tat does not have to weigh and combine pieces of information in some more or less computationally expensive way. Many other successful heuristics, such as generous tit-for-tat and win–stay–lose–shift, have also been proposed for iterated prisoner's dilemma and other games.

IV. HOW TO MEASURE A HEURISTIC'S SUCCESS

The study of heuristics is a key approach to understanding how real minds make decisions for two main reasons. First, many of life's important problems, from choosing a mate to finding a job, cannot be solved in an optimal way because the space of possibilities that must be taken into account is often unlimited; hence, heuristic shortcuts are called for. Second, even when this space of possible solutions is limited and knowledge is complete, optimization may require unfeasible amounts of computation (as in trying to determine the best next move in chess) so that, again, heuristics will be an appropriate approach for the mind to take.

The fact that there are no optimal strategies for many real-world tasks, however, does not mean that there are no performance criteria. One set of criteria that is often used to evaluate judgments and decisions is their internal coherence, defined as accordance with the laws of probability theory and logic. For instance, if judgments are consistent (e.g., "I always think that event A is more likely than B") and transitive ("I think that A is more likely than B, B is more likely than C, and therefore that A is more likely than C"), this is taken as an indication that the underlying decision strategies are rational. If such criteria are violated, this is typically held to be a sign of irrationality on the part of the decision maker. The heuristics-and-biases

research program has focused on such relatively abstract coherence criteria to indicate when a heuristic produces reasonable or unreasonable decisions.

Alternatively, the success of a heuristic can be measured by comparing its performance with the requirements of its environment, such as accuracy, frugality, and speed. Lexicographic strategies (e.g., Take The Best) are often evaluated via correspondence criteria relating to real-world decision performance (such as how often they correctly choose the larger object in a pair). Comparing heuristics' performance to the requirements of the external world rather than to internal consistency stems from the view that the primary function of heuristics is not to be coherent. Rather, their function is to make reasonable adaptive inferences about the real social and physical world given limited time and knowledge.

The two kinds of criteria, coherence and correspondence, can sometimes be at odds with each other. For instance, in social situations, including some competitive games and predator–prey interactions, it can be advantageous to exhibit inconsistent (and hence noncoherent) behavior in order to maximize adaptive unpredictability (and hence correspondence with real-world goals) and avoid capture or loss. As another example, the Minimalist heuristic violates the coherence criterion of transitivity but nevertheless makes fairly robust and accurate inferences in particular environments. Thus, intransitivity does not necessarily imply high levels of inaccuracy, nor does transitivity guarantee high levels of accuracy: Logic and adaptive behavior are logically distinct.

Finally, it is important to measure the performance of decision mechanisms in terms of how well they make decisions when applied to new data; that is, how they generalize to new situations rather than merely how closely they can be adjusted or fit to a static set of data. In this regard, simple heuristics will often do very well, being about as accurate as complex general strategies that work with many free parameters. The reason is that simple heuristics can avoid being too matched to any particular environment; that is, they can escape the curse of overfitting mentioned earlier. As a consequence, a computationally simple heuristic that uses only some of the available information can be more robust, making more accurate predictions for new data.

V. CONCLUSION

Simplicity in models has aesthetic appeal. The mechanisms are readily understood and communicated, and they are amenable to step-by-step scrutiny. Furthermore, Popper has argued that simpler models are more falsifiable. However, the idea that humans make many decisions using simple heuristic mechanisms is important not just because the resulting simple models are transparent and easily falsifiable. More important, simple heuristics may be the only approach available for real minds making decisions in the real, uncertain, time-pressured world.

See Also the Following Articles

ARTIFICIAL INTELLIGENCE • INFORMATION PROCESSING • INTELLIGENCE • LOGIC AND REASONING • NEURAL NETWORKS

Suggested Reading

Axelrod, R. (1984). *The Evolution of Cooperation*. Basic Books, New York.

Brams, S. J., and Taylor, A. D. (1996). *Fair Division: From Cake-Cutting to Dispute Resolution*. Cambridge Univ. Press, Cambridge, UK.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *Am. Psychologist* **34,** 571–582.

Gigerenzer, G., Todd, P. M., and the ABC Research Group (1999). *Simple Heuristics That Make Us Smart*. Oxford Univer. Press, New York.

Kahneman, D., Slovic, P., and Tversky, A. (Eds.) (1982). *Judgment under Uncertainty: Heuristics and Biases*. Cambridge Univer. Press, Cambridge, UK.

Newell, A., and Simon, H. A. (1972). *Human Problem Solving*. Prentice Hall, Englewood Cliffs, NJ.

Payne, J. W., Bettman, J. R., and Johnson, E. J. (1993). *The Adaptive Decision Maker*. Cambridge Univ. Press, New York.

Pólya, G. (1945). *How to Solve It: A New Aspect of Mathematical Method*. Princeton Univ. Press, Princeton, NJ.

Simon, H. A. (1955). A behavioral model of rational choice. *Q. J. Econ.* **69,** 99–118.

Simon, H. A. (1956). Rational choice and the structure of environments. *Psychol. Rev.* **63,** 129–138.

Simon, H. A. (1990). Invariants of human behavior. *Annu. Rev. Psychol.* **41,** 1–19.

Svenson, O. (1979). Process descriptions of decision making. *Organizational Behav. Hum. Performance* **23,** 86–112.

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychol. Rev.* **79,** 281–299.

Hindbrain

JOEL C. GLOVER
University of Oslo

GLOSSARY

Hox genes Encode transcription factors bearing a DNA-binding portion called the homeodomain. Originally identified in *Drosophila* through the homoeotic effects of their mutation. In vertebrates their expression is strongly associated with the rhombomeric domains of the hindbrain.

Pax genes Encode transcription factors with a DNA-binding domain encoded by a sequence called the "paired box." Certain Pax genes are expressed in longitudinal domains within the hindbrain.

raphe Ventral seam of the brain stem, derived from the embryonic floor plate. Contains commissural axons and neuron populations termed the raphe nuclei.

reticular formation Central region of the brain stem extending from mesencephalon to medulla oblongata and containing relatively diffuse populations of neurons. Has a reticular appearance in histological sections stained to reveal nerve fibers. Regulates a wide variety of neural functions.

rhombencephalon Most caudal of the three primary brain vesicles; synonymous with the term "hindbrain." Assumes a form roughly approximating a rhombus during later development and eventually gives rise to the medulla oblongata, pons, and the greater portion of the cerebellar primordium.

rhombic lip Region of the hindbrain rimming the dorsolateral recess of the fourth ventricle.

rhombomeres Rhombencephalic segments, transiently visible as swellings or partitions during early stages of hindbrain development.

The hindbrain, or rhombencephalon, is an embryological term denoting the caudalmost of the three primary embryonic brain vesicles that form in the rostral neural tube. It gives rise to the pons and medulla oblongata as well as the greater portion of the cerebellum. Here, we restrict our view of the hindbrain to that region encompassing the pontine and medullary divisions since the cerebellum is discussed in a separate article. The pons and medulla oblongata are complex anatomical structures that subserve a wide range of vital functions. This article provides a general description of the anatomical and functional organization of the pons and medulla and then discusses the embryonic development of the hindbrain, including mechanisms involved in patterning its anatomy and function. It concludes with some brief considerations of pathology.

I. ANATOMICAL AND FUNCTIONAL ORGANIZATION

A. General Features

The pons (metencephalon or "behind-brain") and medulla oblongata (myelencephalon or "medulla-brain") are the two most caudal divisions of the brain, lying between the mesencephalon and the spinal cord. Seen from the ventral surface (Fig. 1A), the boundaries between mesencephalon and pons (pontomesencephalic sulcus) and between pons and medulla oblongata (pontomedullary sulcus) are clearly demarcated by the massive population of transversely oriented pontocerebellar fibers. In contrast, there is a smooth transition from medulla oblongata to spinal cord the only indication being the pyramidal decussation. Most of the dorsal surface of the pons and medulla oblongata

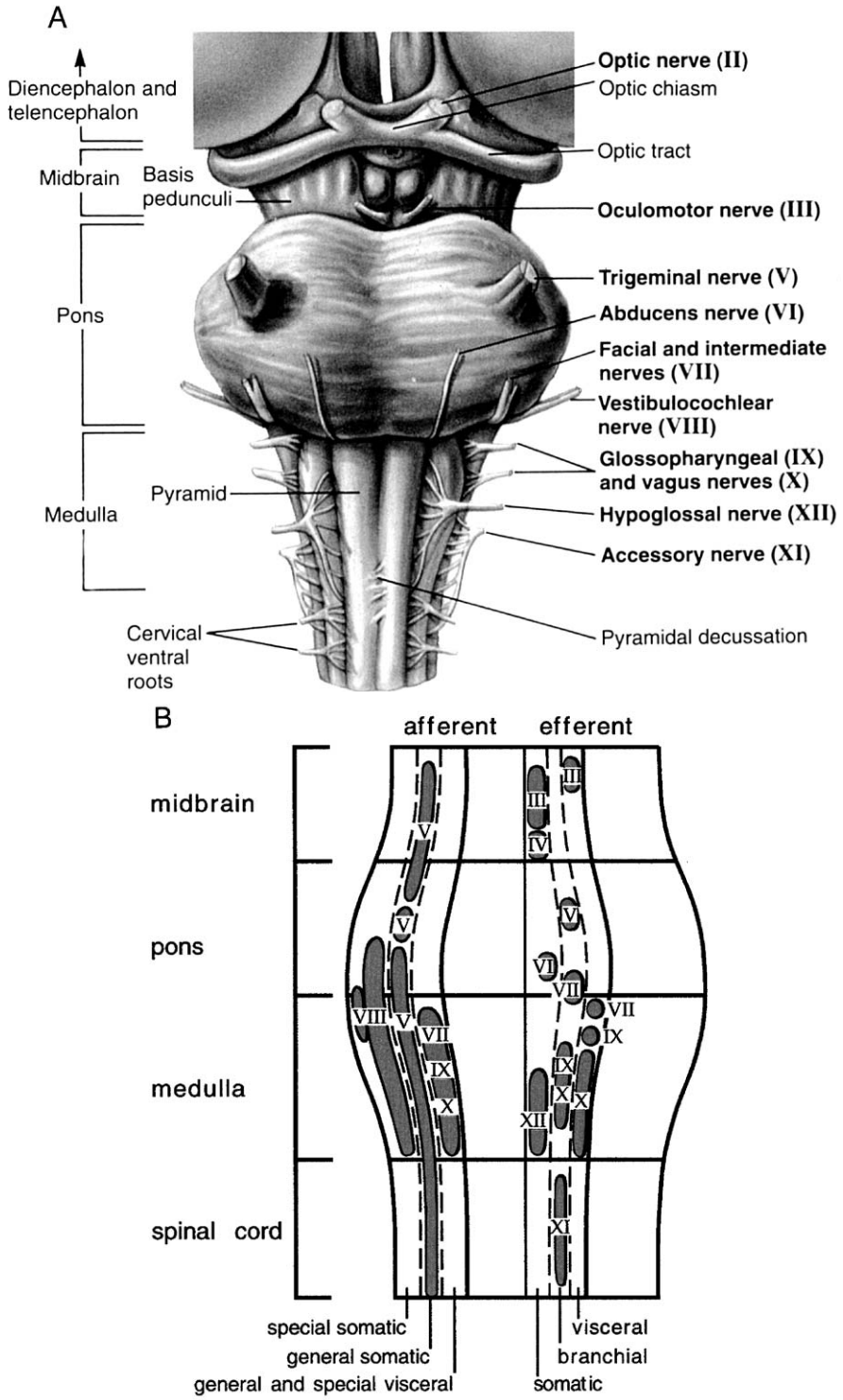


Figure 1 (A) Ventral view of the human brain stem showing the major structural features of the pons and medulla oblongata, including the cranial nerves (modified from Kandel *et al*., 1992, *Principles of Neural Science*, 3rd ed., with permission of the McGraw-Hill Companies). (B) Organization of cranial nerve-associated nuclei into afferent (sensory) and efferent (motor) columns. The locations of nuclei relative to mesencephalic (midbrain), pontine, and medullary divisions are approximate. Phylogenetic variants on this theme exist. Because of longitudinal migrations of certain nuclear groups, embryonic origins may not correspond to final locations. For example, the trochlear nucleus (IV) has been shown to derive from the pontine division in lower mammals [redrawn and modified from Martin (1989), *Neuroanatomy*, with permission of the McGraw-Hill Companies].

underlies the fourth ventricle as its floor. The floor of the fourth ventricle and the ventral surface of the medulla oblongata show a relief of underlying longitudinal fiber tracts and nuclei. The lateral recesses of the fourth ventricle are bounded by a rim of neural tissue called the rhombic lip, which is continuous with a thin velum overlying the fourth ventricle. The caudal portion of the velum differentiates into choroid plexus. In the intact brain, the fourth ventricle is hidden by the cerebellum.

Eight of the 12 cranial nerves originate from the pons or medulla oblongata, issuing from specific sites on the ventral and lateral aspects (Fig. 1A). The trigeminal nerve issues from among the pontocerebellar fibers on the lateral aspect of the pons. The abducens, facial, and vestibulocochlear nerves issue from respectively medial to lateral positions at the pontomedullary border. The glossopharyngeal, vagus, and spinal accessory nerves issue from respectively rostral to caudal sites along the ventrolateral aspect of the medulla oblongata, dorsal to the inferior olive. The hypoglossal nerve issues from the ventral aspect of the medulla oblongata between the inferior olive and the pyramid.

The nuclei associated with the cranial nerves are organized into longitudinal motor and sensory columns that are subdivided according to which peripheral structures are innervated (Fig. 1B). The organization is similar to that in the spinal cord except for the presence of special columns that innervate structures specific to the head. The motor columns lie more medially and innervate either striated muscle (somatic and branchial columns) or parasympathetic ganglia (visceral column). The sensory columns lie more laterally and transmit tactile, proprioceptive, pain, or temperature signals (general somatic and visceral columns) or other specific sensory modalities such as audition and balance (special somatic column) and taste and olfaction (special visceral column). The sulcus limitans, a shallow longitudinal groove visible in the floor of the fourth ventricle, separates the motor from the sensory columns (Fig. 2).

The pons and medulla oblongata contain many fiber tracts that relay information between the cerebrum, cerebellum, and the spinal cord. The transversely oriented pontocerebellar projection is one of the largest of these and is the dominant external feature of the pons (Fig. 1A). Within the pons, longitudinal fiber tracts course internally to or intermingled with the pontocerebellar fibers. Within the medulla oblongata, most longitudinal tracts course at or near the outer (pial) surface. Some longitudinal tracts,

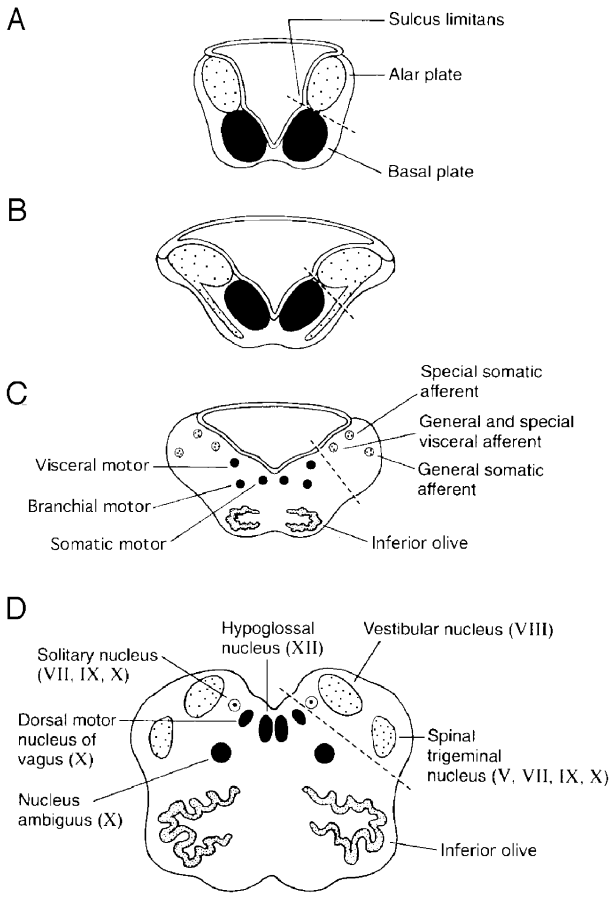


Figure 2 Series of transverse sections through the medulla oblongata at different stages of development. Different nuclei originate from basal and alar plates in a specific pattern. Motor and sensory cranial nerve-associated columns, for example, derive from the basal and alar plates, respectively. Note that inferior olive neurons derive from the alar plate but emigrate to the basal plate where they coalesce to form the nucleus [modified from Kandel *et al.*, (1992), *Principles of Neural Science*, 3rd ed., with permission from the McGraw-Hill Companies].

however, course at deeper locations near the ventral midline of the medulla oblongata. These include several important descending tracts to the spinal cord. The largest longitudinal tract, the corticospinal tract, is visible along the ventral medullary surface as the bilaterally paired pyramids, one of the dominant external features of the ventral medulla oblongata (Fig. 1A).

Like the cranial nerve nuclei, the longitudinal fiber tracts are differentially situated according to functional modality. Ascending tracts conveying sensory information from spinal and medullary centers to higher centers generally course at more dorsal and

dorsolateral locations, whereas descending tracts generally course at more ventral and ventromedial locations. Exceptions to this general rule are the rubrospinal tract, which attains a lateral position within the medulla oblongata as it descends to the spinal cord; the medial lemniscus and ventral trigeminal tract, two sensory tracts that initially have a ventromedial course; and the spinal trigeminal tract, a sensory tract that descends along the dorsolaterally located spinal trigeminal nucleus.

Several well-defined nuclei exist within the pons and medulla oblongata in addition to those associated with the cranial nerve nuclei. The largest of these is the inferior olive, which produces a prominant bulge along the ventrolateral surface of the rostral medulla (Fig. 1A), and whose characteristically convoluted appearance in transverse sections makes it an unmistakable landmark (Fig. 2). The inferior olive is the source of climbing fiber afferents to the contralateral cerebellum.

The bulk of the pontine and medullary core is filled by a more diffusely organized population of neurons that make up the reticular formation, so named because it appears highly reticulated when stained nonspecifically for nerve fibers. This appearance, along with early physiological findings that demonstrated a lack of modality-specific activity in reticular efferents, led to the early notion that much of the reticular formation functioned as a distributed network involved in general activation and arousal. Recent anatomical and physiological studies, on the other hand, have demonstrated a higher degree of anatomical and functional mosaicism within the reticular formation than was previously appreciated. Anatomical subdivisions with characteristic patterns of connections and neurotransmitter profiles exist, and a number of well-defined premotor networks have been identified that organize and integrate specific goal-directed movements.

Certain reticular neuron populations are distinct enough anatomically to be defined as nuclei, even though most of these are not sharply delimited from the rest of the reticular formation. Two of the most distinct reticular nuclei are of special interest because of their neurotransmitter phenotypes. The raphe nuclei are clusters of neurons located within the ventral raphe, the majority of which are serotonergic. The locus coeruleus is a collection of noradrenergic neurons in the pons. The raphe nuclei and the locus coeruleus have exceptionally widespread projections and terminations and exert modulatory effects on a variety of neural systems.

B. Overview of Constituent Neuron Populations

1. Efferent Neurons

These include (i) motoneurons of cranial nerves V–VII and IX–XII that innervate striated muscle in the head and neck region (trigeminal motor, abducens, facial, ambiguus, and accessory nuclei), (ii) preganglionic neurons that innervate cranial parasympathetic ganglia (superior salivatory nucleus of cranial nerve VII and inferior salivatory nucleus of cranial nerve IX) and autonomic ganglia in the trunk (dorsal motor nucleus of nerve X), and (iii) cochlear and vestibular efferents from respectively the superior olive and reticular formation that innervate hair cells in the inner ear. All of these efferent populations are cholinergic.

2. Nuclei That Receive Cranial Nerve Sensory Afferents, and Some of Their Secondary Nuclear Targets

These include (i) sensory columns of the trigeminal system which receive inputs from nerves V, VII, and X; (ii) the vestibular and cochlear nuclei, which receive inputs from nerve VIII; and (iii) the solitary nucleus, which has subdivisions that receive afferents conveying taste via nerves VII, IX, and X, afferents from receptors in cranial skin and mucous membranes, and afferents from receptors in the gut, cardiovascular system, and lungs.

3. Other Relay and Integration Centers

These include the pontine nuclei, which relay primarily cortical inputs to the contralateral cerebellum; the inferior olive, which supplies climbing fiber inputs to the contralateral cerebellum, the nucleus prepositus, which is involved in regulating eye movements; the dorsal column nuclei, which integrate afferent information from the spinal cord and relay it to the thalamus; the deep cerebellar nuclei, which relay efferent information from the cerebellum; and various subdivisions of the reticular formation.

4. Sources of Descending Inputs to the Spinal Cord

These include the vestibulospinal, reticulospinal, and raphespinal populations and the locus coeruleus.

C. Functional Centers and Networks

In addition to being clustered into discrete nuclei with specific modalities, the neuron populations of the pons

and medulla oblongata constitute interconnected networks that integrate somatic and autonomic sensory and motor functions. Many of these networks subserve reflex pathways that link the afferent inputs from certain cranial nerves with the motor output of others via interposed populations of interneurons within the reticular formation and specific relay nuclei. The complex coordination of the mouth, tongue, pharynx, and upper alimentary canal during the eating of a meal exemplifies the substantial degree of integration of different cranial nerve nuclei. The control of eye movements, the regulation of cardiovascular function, and a variety of autonomic reflexes, such as hiccuping, sneezing, and vomiting, are all examples of the interplay of different cranial nerve nuclei. The networks of interneurons that mediate these reflexes are not fully characterized, but some of them have been localized to specific sites within the reticular formation. Thus, in addition to functioning to some extent as a distributed integrative network, the reticular formation exhibits a substantial degree of functional compartmentalization.

Indeed, many of the regulatory functions exerted by the pons and medulla oblongata on other regions of the brain arise from specific, localized centers within the reticular formation. These include the regulation of pain impulses, the initiation of locomotion, and the control of cardiovascular function and respiration.

II. EMBRYONIC DEVELOPMENT

The development of the hindbrain has been the focus of intensive research during the past 10 years. Although the human hindbrain has been the subject of a few studies, most of our knowledge of the developmental programs and mechanisms that pattern the hindbrain is derived from animal studies. Since the hindbrain is an evolutionarily ancient brain structure, many of the key developmental features gleaned from other species are likely to be conserved throughout the vertebrate radiation, including primates. Nevertheless, species differences do exist, and the reader should note that some of the description that follows may differ in certain details from the developmental events occuring in the human embryo.

A. General Features

1. Morphogenesis

The brain and spinal cord originate from the embryonic ectoderm through the action of inductive signals from underlying mesoderm and within the ectoderm. During this process, a region of dorsal ectoderm called the neural plate is delineated, folds together at the midline to form the neural tube, and invaginates into the dorsal aspect of the embryo. Even before closure and invagination of the neural tube are complete, three initial subdivisions of the brain can be discerned, first by the appearance of indentations within the neural plate and subsequently by the expansion of the intervening regions into vesicles of the neural tube. These are the primary brain vesicles, the prosencephalon, mesencephalon, and rhombencephalon (Fig. 3). Within a short time (by about 4 weeks of development), further subdivisions of the neural tube, called neuromeres, can be seen. The neuromeres of the rhombencephalon are called rhombomeres (Fig. 3A).

During flexure of the developing neural tube, the rhombencephalon assumes a shape approximating a rhombus, hence the name (Fig. 4). This occurs at least in part because the mesencephalic and cervical flexures create constrictions in the neural tube, whereas the pontine flexure creates a lateral expansion. Initially, the wall of the rhombencephalic neural tube has approximately the same thickness around its entire circumference. As the flexures appear, however, morphometric changes occur that presage the mature structure (Fig. 2). The ventral portion starts proliferating extensively and eventually gives rise to the bulk of the pons and medulla oblongata. A specialized structure at the ventral midline known as the floor plate eventually gives rise to the raphe. The dorsal portion, also called the roof plate, becomes relatively much thinner and eventually gives rise to the velum of the fourth ventricle with its associated choroid plexus. No neurons differentiate within the roof plate and its derivatives. At pontine levels, the dorsal portion of the neural tube, on either side of the roof plate, develops into progressively thicker flaps of neural tissue that establish the major portion of the cerebellar primordium. A longitudinal indentation, called the sulcus limitans, appears in the floor of the fourth ventricle about midway between the floor plate and roof plate. This sulcus divides the wall of the neural tube into two longitudinal plates or columns, the basal plate and the alar plate.

Shortly after the hindbrain neural tube forms, and during the period when rhombomere and flexure formation is shaping the hindbrain, a population of progenitor cells at the dorsal aspect of the tube emigrates into the periphery. This population, a part of the cranial neural crest, contributes much of the mesenchyme of the cranium and branchial arches and

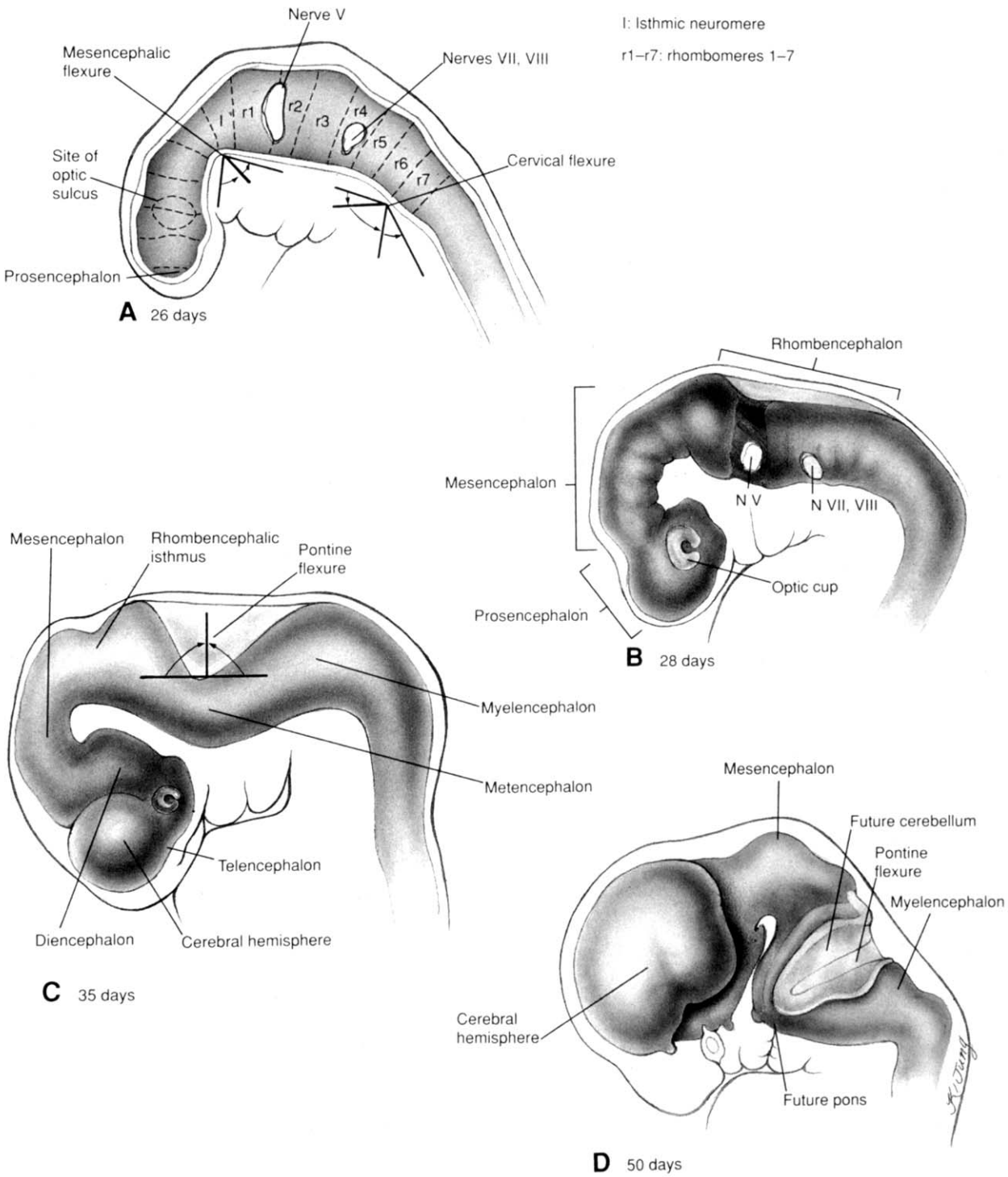


Figure 3 Important morphogenetic changes in the neural tube that shape the hindbrain [reproduced with permission from Larsen (1997), *Human Embryology*. Churchill Livingstone, New York].

of certain regions of the thorax including the developing thymus and heart. The neural crest cells proliferate and differentiate into a large number of cell types, including cartilage and other connective tissue structures, smooth muscle, melanocytes, peripheral neurons, and Schwann cells.

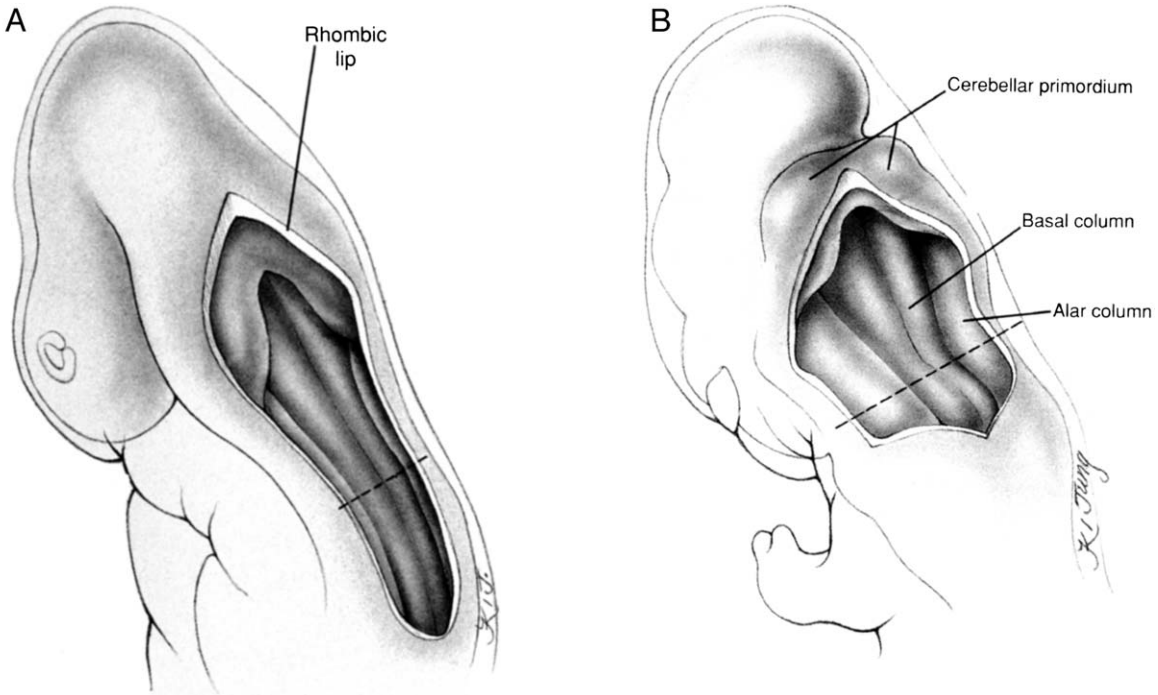


Figure 4 Dorsal views of the human embryo at approximately 4 weeks (A) and 5 weeks (B), showing the development of structural features in the floor of the fourth ventricle and the dorsolateral aspect of the hindbrain. The dotted line indicates the level of section shown in Fig. 2 [modified from Larsen (1997), *Human Embryology*. Churchill Livingstone, New York].

2. Neurogenesis and Migration

Neural progenitor cells of the neural tube proper are located and undergo mitosis near the inner (luminal) surface of the tube in a layer called the ventricular zone. As postmitotic neurons are generated, they migrate radially toward the outer (pial) surface, establishing a mantle zone where they begin to differentiate and aggregate into nuclei. Most neurons take up residence somewhere along this radial trajectory, but some turn to migrate circumferentially, either toward the floor plate or toward the roof plate, or longitudinally. Neurons generated in the alar plate, particularly the rhombic lip region, have a particular predilection for circumferential migration. Indeed, several nuclei, including the raphe nuclei and the inferior olive, are established within the basal plate through immigration from the alar plate (Fig. 2B).

By and large, the alar and basal plate domains maintain a coherent relationship to the sensory and motor divisions of the cranial nerve nuclei, respectively (Fig. 2), with the sulcus limitans persisting in the mature hindbrain as a landmark of this division. In general, motoneurons are generated within the basal plate, whereas the neurons of the sensory nuclei are generated within the alar plate. Some hindbrain cranial nerve nuclei, however, undergo circumferential migration, such as the trigeminal motoneurons, which migrate away from the floor plate, and the cochlear efferent neurons, which migrate toward the floor plate and cross the midline. The disposition of neuron groups in the mature hindbrain therefore does not necessarily accurately indicate their embryonic origins.

3. Axon Outgrowth and Tract Formation

Axon outgrowth is an early feature of neuronal differentiation and may be quite advanced even as neurons are migrating. Many axons are commissural, crossing the midline to course in axon tracts and innervate synaptic targets on the opposite side of the neuraxis. Once axons begin to project longitudinally, they tend to fasciculate into bundles that are located predominantly either near the pial surface or adjacent to the raphe. Two major axon bundles are present at early stages of hindbrain development: a medial longitudinal fascicle (mlf) and a lateral longitudinal

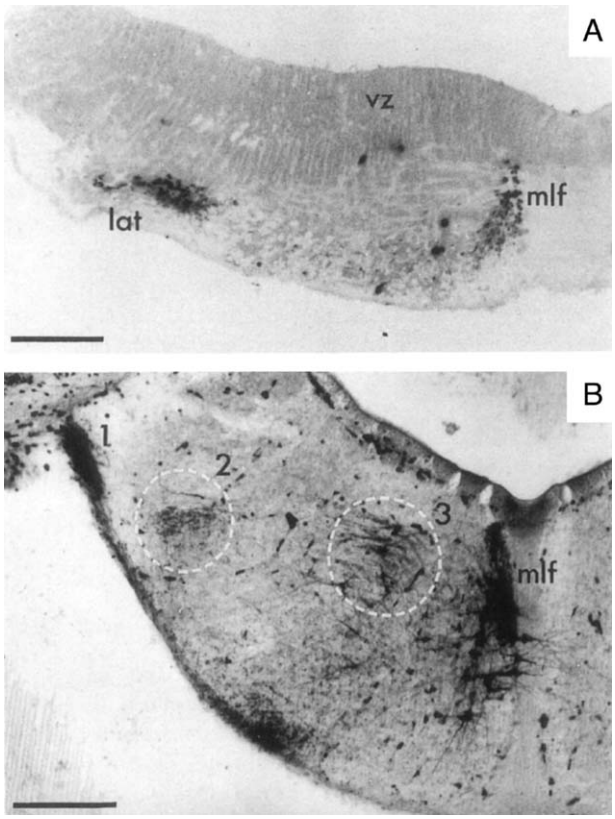


Figure 5 Transverse sections through the developing medulla oblongata of a chicken embryo illustrating the initial two longitudinal fiber bundles [A; medial longitudinal fascicle (mlf) and lateral longitudinal fascicle (lat)] and the splitting of the lateral longitudinal fascicle into separate tracts (B; 1, spinocerebellar tract; 2, spinal trigeminal tract; 3, lateral vestibulospinal tract) (modified from Glover and Petursdottir, 1991).

fascicle (llf) (Fig. 5). Each contains ascending and descending axons, which tend to be segregated into subfascicles containing axons of a given type. As the transversal area of the hindbrain expands, the mlf maintains its coherence alongside the raphe and remains as the tract of the same name in the mature brain. The llf, on the other hand, splits into several tracts, including various ascending sensory tracts, the spinal trigeminal tract, and the lateral vestibulospinal tract. This splitting is accompanied by intercalation of increasing numbers of neurons and other axons. Thus, the early formation of longitudinal axon tracts occurs on a simple scaffold that increases in complexity by a process of fission. Later formed tracts, such as the corticospinal tract and the pontocerebellar projection, are layered externally to the early formed tracts.

B. Gene Expression and Regionalization

1. Rhombomeres and Longitudinal Patterning

The rhombomeres are transient subdivisions that were described in the embryos of a number of species including the human starting in the late 1800s. Their analysis underwent a dramatic renaissance nearly 100 years later with the application of modern cellular and molecular techniques, spearheaded by the efforts of Andrew Lumsden and colleagues.

As many as eight rhombomeres have been described, and these are denoted r1, r2, r3, and so on from rostral to caudal (Fig. 3). The second through sixth rhombomeres (r2–r6) are similar in length and readily visible in most species. The first and last two rhombomeres may represent transitional neuromeres at the junctions between hindbrain and mesencephalon and hindbrain and spinal cord.

Rhombomere boundaries have several features that contribute to a physical segmentation of the hindbrain neural tube. The proliferation rate is lower at the boundaries, creating the indentations that mark their positions visibly. The pattern of intercellular communication via gap junctions is modulated at the boundaries, several extracellular proteins are preferentially expressed there, and cells in alternating rhombomeres have different cell adhesion properties such that they tend to segregate if mixed. All of this contributes to the formation of a physical barrier to cell movement over rhombomere boundaries. Indeed, in contrast to the extensive circumferential migration that occurs among certain hindbrain neuron populations, rostrocaudal migration is much more limited. This has led to the idea that the rhombomeres compartmentalize the hindbrain, confining progenitor cells and their offspring to specific rostrocaudal domains. Of course, longitudinal fibers penetrate the rhombomere boundaries, and some neuron populations breach the barriers as well, especially at sites of intersection by fiber tracts.

The rhombomeres appear during a period when many hindbrain neurons are being generated, and they fade away by the time most of the major nuclei of the hindbrain have appeared. The temporal concurrence with neurogenesis and differentiation suggests that the rhombomeres play a role in patterning that differentiation. Molecular correlates of rhombomeric compartmentalization strongly support this notion. In particular, the rhombomeric domains are correlated with the expression of a number of developmental regulatory genes that are involved in controlling the

differentiation of neurons into specific phenotypes along the rostrocaudal axis.

One class of regulatory genes that figures prominently is the Hox gene family, which encodes transcription factors whose differential expression establishes regional patterns of differentiation in a variety of organisms and organ systems. Hox genes are ordered in a specific sequence within the genome, a relationship that is reiterated in their tissue expression pattern. Within the hindbrain, Hox genes have a sequentially overlapping pattern of expression that is strongly correlated with the rhombomeres and provides each of them with a unique combinatorial address (Fig. 6). Evidence that the Hox genes are important determinants of neuronal differentiation comes from a variety of manipulations that alter the pattern of Hox gene expression. These include transgenic knockouts of specific Hox genes, ectopic expression of Hox genes by retroviral gene transfer, heterotopic transplantation of rhombomeres, treatment with retinoids, perturbations of Hox gene promoter sequences that alter the region-specific expression pattern, and manipulation of other genes that regulate Hox gene expression. In parallel with alterations in Hox gene expression, these manipulations lead to changes in the regional pattern of neuronal differentiation, such that neuronal phenotypes characteristic of a given rhombomere appear in a different rhombomere.

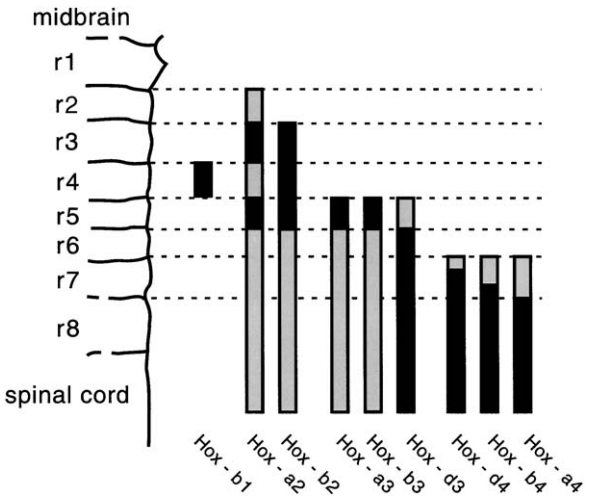


Figure 6 The relationship of rhombomeric domains to the longitudinal expression patterns of Hox genes in the mouse hindbrain. Black indicates strong expression, and gray indicates weaker expression (reproduced with permission from Keynes and Krumlauf, *Annual Review of Neuroscience*, Vol. 17, © 1994 by Annual Reviews, www.AnnualReviews.org).

Hox genes are also expressed by the emigrating cranial neural crest cells and are involved in patterning the various mesenchymal derivatives of the cranium, branchial arches, and neural crest-derived thoracic structures.

How is the normal, longitudinally sequential pattern of Hox gene expression established? This appears to be a complicated issue that involves a number of factors. Hox gene expression is known to be regulated by retinoids, acting through nuclear retinoid receptors, which are ligand-dependent transcription factors that interact directly with Hox gene promoter sequences. This regulation is concentration dependent. Since retinoid synthesis has been shown to be high in the spinal cord but very low in the hindbrain, a diffusion gradient of retinoids from the spinal cord rostrally into the hindbrain could contribute to setting up the pattern. Hox gene expression is also regulated by the action of other transcription factors via specific promoter sequences that direct expression differentially according to tissue type and region. The pattern of expression of such transcription factors is therefore pivotal in setting up the pattern. Lastly, Hox gene expression is cross-regulated and autoregulated by Hox proteins, in cooperation with other transcription factors. An important feature that underscores the complexity of Hox gene regulation is the dynamic pattern of expression in the hindbrain. Hox genes do not merely pop up in specific longitudinal domains but may be expressed in broader domains that eventually become restricted and that are modulated over time also in the transverse plane.

The Hox genes are not the only genes that exhibit rhombomere-related patterns of expression, but they are the most extensively studied so far with respect to a role in regulating neuronal differentiation. Some of the other rhombomere-related genes are likely to be downstream targets of the Hox genes because they code for membrane receptors and signaling molecules that are involved in features of differentiation, such as directed cell migration and axonal outgrowth. Others, such as the *kreisler* and *Krox-20* genes, are also transcription factor genes that participate with the Hox genes in the network of gene activity that sets up the regional patterning of the hindbrain.

2. Patterning in the Transverse Plane

In addition to the longitudinal patterning of the hindbrain exemplified by the rhombomeres, there is a

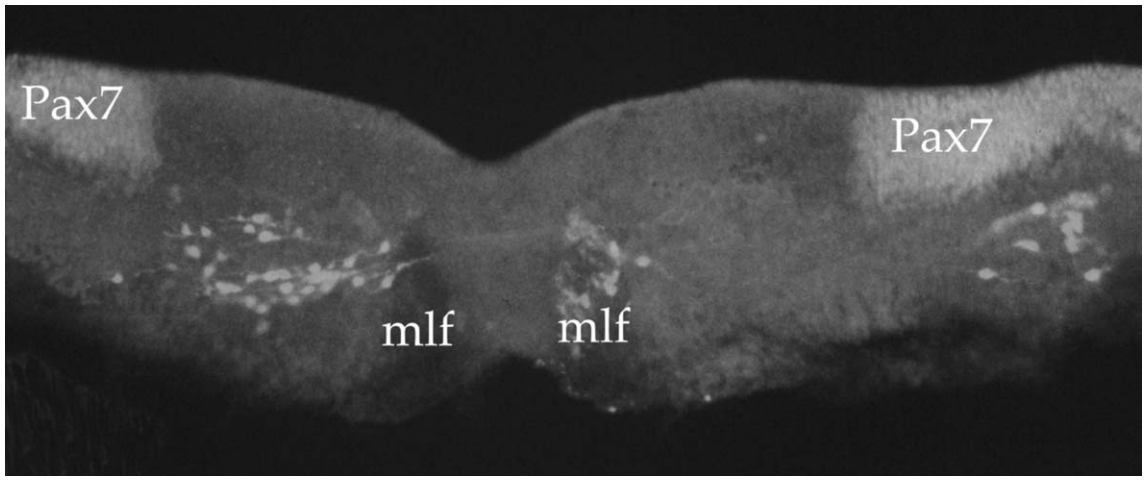


Figure 7 An example of hindbrain patterning in the transverse plane. Expression of the transcription factor *Pax7* by progenitor cells within the proliferative zone (immediately subjacent to the fourth ventricle) defines two zones with a sharp boundary at a specific site along the floor plate–roof plate axis. The hindbrain has been opened dorsally and laid flat prior to sectioning, such that the floor plate–roof plate axis runs medial to lateral on each side. Two different hindbrain neuron groups are retrogradely labeled from their axons in the right-side medial longitudinal fascicle (mlf). The group on the left side derives from the *Pax7*-negative progenitors, whereas the group on the right side derives from the *Pax7*-expressing progenitors.

systematic patterning along the floor plate–roof plate axis. This is not as obvious from the morphological standpoint because there are few related structural landmarks aside from the sulcus limitans. The organization of the cranial nerve nuclei into longitudinal columns, however, gives an immediate reflection of this transverse element of patterning. Transcription factors also figure prominently in setting up longitudinal domains within the hindbrain. Expression of Pax and Nkx genes within the progenitor cell population provides a good example. These genes, like the Hox genes, encode transcription factors known to regulate the regional differentiation of cells in a variety of tissues and species. They are expressed in overlapping longitudinal bands along much of the neuraxis, including the hindbrain, with very sharp boundaries of expression at specific levels along the floor plate–roof plate axis (Fig. 7). They can exhibit dynamic changes prior to boundary formation, and they can also exhibit modulations in expression intensity within the expression domain that becomes established. Thus, these genes provide a sequentially ordered combinatorial scheme of gene expression along the floor plate–roof plate axis in much the same way that the Hox genes pattern the longitudinal axis.

How is the normal pattern of gene expression in the transverse plane generated? This has been studied primarily in the developing spinal cord, but the emerging picture appears to be generally applicable to the hindbrain as well. Diffusible signals are released at or near the ventral and dorsal poles of the neural tube. Ventral signals derive from the notochord, a mesodermal structure that lies immediately ventral to the neural tube, as well as from the floor plate. Dorsal signals derive from the roof plate and overlying surface ectoderm. These signals establish opposing gradients that dictate which genes are expressed at different positions along the floor plate–roof plate axis. The way these signals work and the way the genes whose expression they regulate interact are complex issues that remain the subject of active research. The end result, however, is the establishment of overlapping, longitudinal bands of differential gene expression that, like the zones of differential Hox gene expression, can be correlated with the differentiation of specific neuron groups (Fig. 7).

3. From Gene Expression to Neural Networks

In combination, transcription factors encoded by these and other gene families are expressed in a kind of checkerboard pattern of intersecting rhombomeres and longitudinal bands. How does this relate to the regional pattern of neuronal differentiation? The spatial and temporal correlation of hindbrain neuron types to gene expression patterns is far from complete, but several conclusions can be made.

First, certain neuron groups are neatly delimited longitudinally by rhombomeric domains, and certain groups are neatly delimited along the floor plate–roof

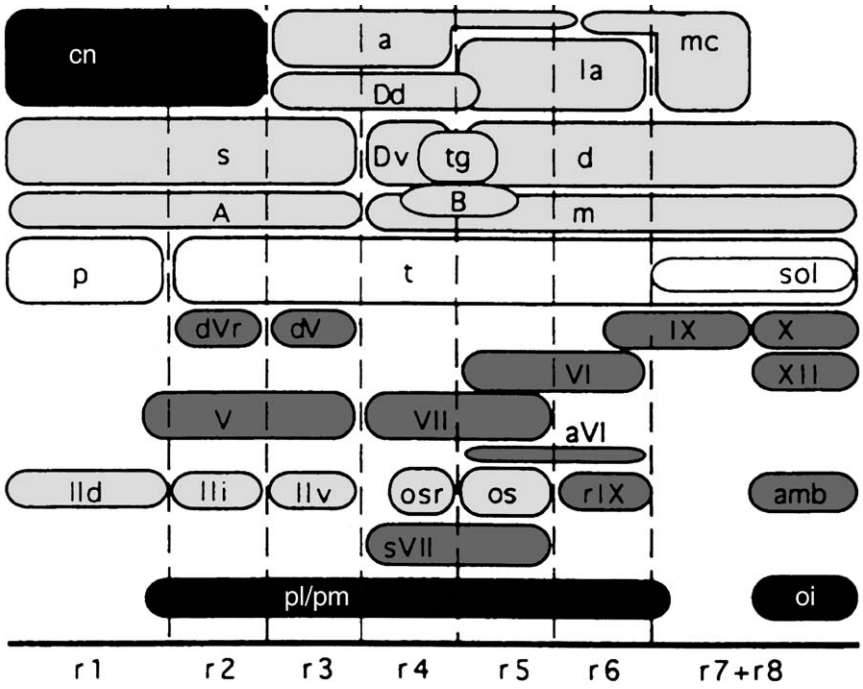


Figure 8 The relationship of hindbrain nuclei to rhombomeric domains in the chicken embryo. Nuclei are ordered roughly along the floor plate–roof plate axis from bottom to top. Different shades of grey indicate classes of related nuclei. Note that the rhombomeric pattern exhibits phylogenetic variants. cn, cerebellar nuclei; a, angularis; la, laminaris; mc, magnocellularis; Dd, Deiters dorsalis; s, superior vestibular; Dv, Deiters ventralis; tg, tangentialis; d, descending vestibular; A, vestibular cell group A; B, vestibular cell group B; m, medial vestibular; p, principal trigeminal; t, descending trigeminal; sol, nucleus of solitary tract; dV, dorsal trigeminal (r, rostral part); V, trigeminal motor; VI, abducens; aVI, accessory abducens; VII, facial; IX, glossopharyngeal; X, vagus; XII, hypoglossal; lld, nucleus of lateral lemniscus, (d, dorsal; i, intermediate; v, ventralis); os, superior olive (r, rostral part); rIX, retrofacial glossopharyngeal; amb, ambiguus; sVII, superficial facial; pl/pm, lateral and medial pontine nuclei; oi, inferior olive (modified from Marin and Puelles, 1995).

plate axis by longitudinal gene expression domains, whereas others are not (Fig. 8). There are several potential explanations for examples of noncongruity between neuron groups and these domains. For example, the pattern as described to date may be incomplete. Additional genes could subdivide the currently known expression pattern into additional domains; similarly, additional information about the phenotypic diversity of neuron groups may introduce novel group subdivisions. Alternatively, the noncongruity might result from the dynamic features of both gene expression and neuron group formation. As noted previously, gene expression patterns can change over time, and neurons can migrate, so correspondences may be evident only within very particular time windows and then disappear.

Second, the relationship between gene expression patterns and neuron groups is not necessarily applicable only to the classically defined cytoarchitectonic nuclei of the hindbrain. Rather, correlations may be stronger to neuron groupings defined by specific phenotypic characters, such as neurotransmitter profile or axon projection pattern. For example, within such populations as the reticulospinal neurons and the vestibular nuclear complex, subdivisions on the basis of axon projection pathway are more readily correlated to gene expression domains than are the classical nuclear divisions.

Third, it appears that the relationship between gene expression patterns and neuron groups can be extended to the connectivity patterns of hindbrain neurons. This feature has only been examined in a few instances, but there are compelling examples of neuron groups whose subdivision according to termination patterns onto synaptic targets can be correlated to gene expression domains. The vestibular nuclear complex provides an example. Here, the different subgroups connect in stereotyped patterns to target motoneurons, creating highly specific reflex pathways for eye and body movements. Although the action of particular genes in establishing this pattern of connectivity has not been experimentally tested, the striking

correlation of the vestibular subgroups to gene expression domains suggests a direct link. Other functional systems within the hindbrain similarly appear to be constructed through the action of regional patterns of gene expression. For example, primordial respiratory activity is generated by a neural network with definable components localized to specific rhombomeres. Genetic manipulations that perturb the rhombomeric pattern lead to specific functional deficits in the network.

To summarize, hindbrain neurons are organized into distinct cranial nerve and other nuclei, with the reticular formation as a central core. The pattern of specific functional subdivisions within these nuclei and within the reticular formation is likely to be directly linked to the highly mosaic pattern of gene expression seen at early stages of hindbrain development. Moreover, this relationship likely contributes to establishing the basic pattern of synaptic connectivity within hindbrain networks. Identifying the gene combinations responsible for specifying the various neuron types and their synaptic connections is one of the major challenges of future research on the hindbrain.

III. BRIEF CONSIDERATIONS OF PATHOLOGY

Given the many critical functions of the pons and medulla oblongata, pathology in these hindbrain derivatives can have a wide variety of effects. Specific or diverse symptoms may result, depending on whether the pathology is focal, involving a particular nucleus or pathway, or more widespread. In either case, the pathological manifestations can be devastating and potentially life-threatening since so many areas of the pons and medulla participate in vital processes. Thus, contusions, tumors, and vascular lesions are often fatal. Several well-known pathological syndromes can be distinguished on the basis of particular combinations of symptoms, as these relate directly to the location of the lesion and the level at which fiber tracts decussate. For example, medial lesions typically present symptoms such as contralateral hemiparesis (corticospinal tract) and loss of proprioception (medial lemniscus). In contrast, lateral lesions typically present symptoms such as ipsilateral loss of facial cutaneous sensation (trigeminal sensory nucleus and tract), contralateral loss of pain and temperature sensation (spinothalamic tract), and Horner syndrome (ipsilateral descending autonomic fibers). In addition, each type of lesion will potentially affect the function of specific cranial nerves.

The different medial and lateral pontine and medullary syndromes lie within the basic diagnostic repertoire of the practicing neurologist. Hindbrain pathology reulting from genetic disorders is a less common class of diagnosis but almost certainly occurs more frequently than is currently recognized. Mutations in transcription factor genes have been established as the cause of a few pathological syndromes, such as Waardenberg's syndrome, that involve developmental defects in the cranial neural crest (cristopathies). Similar mutations in genes patterning the hindbrain neural tube could lead to catastrophic consequences if they disrupt major elements of the pattern, but they could also lead to more focal defects compromising the function of specific networks, with more subtle symptoms as a result. For example, nonlethal mutation of the *Krox-20* gene in the mouse, which disrupts rhombomeric patterning, perturbs the development of respiratory networks and leads to varying degrees of apnea. There are probably a large number of idiopathic conditions affecting the hindbrain that have a genetic basis, and it is not unreasonable to expect that as genetic analysis of hindbrain patterning progresses new syndromes will be classified for which mutation of identified patterning genes is the underlying cause.

See Also the Following Articles

BRAIN ANATOMY AND NETWORKS • BRAIN DEVELOPMENT • BRAIN STEM • CEREBELLUM • FOREBRAIN

Suggested Reading

Brodal, P. (1992). *The Central Nervous System: Structure and Function* (see Chapter 4). Oxford Univ. Press, Oxford.

Champagnat, J., and Fortin, G. (1997). Primordial respiratory-like rhythm generation in the vertebrate embryo. *Trends Neurosci.* **20,** 119–124.

Ericson, J., Briscoe, J., Rashbass, P., van Heyningen, V., and Jessell, T. M. (1997). Graded sonic hedgehog signaling and the specification of cell fate in the ventral neural tube. *Cold Spring Harbor Symp. Quant. Biol.* **62,** 451–466.

Glover, J. C. (1993). The development of brain stem projections to the spinal cord in the chicken embryo. *Brain Res. Bull.* **30,** 265–271.

Glover, J. C. (2000a). Neuroepithelial "compartments" and the specification of vestibular projections. *Prog. Brain Res.* **124,** 3–21.

Glover, J. C. (2000b). The development of specific connectivity between premotor neurons and motoneurons in the brain stem and spinal cord. *Physiol. Rev.* **80,** 615–647.

Keynes, R., and Krumlauf, R. (1994). Hox genes and regionalization of the nervous system. *Annu. Rev. Neurosci.* **17,** 109–132.

Larsen, W. J. (1997). *Human Embryology*, 2nd ed. (see Chapter 13). Churchill Livingstone, New York.

Lee, K. J., and Jessell, T. M. (1999). The specification of dorsal cell fates in the vertebrate central nervous system. *Cell* **96,** 211–224.

Lumsden, A. (1990). The cellular basis of segmentation in the developing hindbrain. *Trends Neurosci.* **13,** 329–335.

Lumsden, A., and Krumlauf, R. (1996). Patterning the vertebrate neuraxis. *Science* **274,** 1109–1115.

Marín, F., and Puelles, L. (1995). Morphological fate of rhombomeres in quail/chick chimeras: A segmental analysis of hindbrain nuclei. *Eur. J. Neurosci.* **7,** 1714–1738.

HIV Infection, Neurocognitive Complications of

IGOR GRANT
University of California, San Diego, and VA San Diego Healthcare System

GLOSSARY

acquired immune deficiency syndrome A disease resulting from collapse of cell-mediated immunity, resulting in infections, cancers, and other complications that lead to death. Human immunodeficiency virus is the causal agent.

antiretroviral drugs Drugs that interfere with various stages of the reproduction of human immunodeficiency virus.

CCR5 A type of chemokine receptor found on macrophages, dendritic cells, and certain other cells.

CD4 cells A type of lymphocyte that is important in coordinating numerous immune events. Infection and depletion of CD4 lymphocytes by HIV results in the evolution of acquired immunodeficiency.

cerebrospinal fluid Fluid surrounding the brain and spinal cord and also contained within the cavities of the brain. Analysis of cerebrospinal fluid may provide clues about pathological processes in the brain.

chemokines, chemokine receptors Chemokines are a family of molecules that are produced in the course of inflammation. Docking sites for such molecules (receptor sites) may be important for HIV entry into a host cell and also in HIV-mediated neural injury.

CXCR4 A type of chemokine receptor found predominantly on CD4 lymphocytes.

dementia, cortical A pattern of neurocognitive change similar to that seen in Alzheimer's disease.

dementia, subcortical Pattern of neurocognitive changes resembling those seen in Huntington's disease, Parkinson's disease, and certain "white matter" diseases of the brain.

gp120 A molecule in the envelope coating of HIV that may be neurotoxic.

human immunodeficiency virus The virus that causes HIV disease and AIDS.

integrase An enzyme of HIV necessary for the process whereby proviral DNA is integrated into the genome of the host cell.

neurocognitive Mental processes whose disruption strongly suggests reversible or irreversible brain injury. The neurocognitive functions (abilities) include attention, perceptual motor abilities, abstracting (including problem solving, planning, and executive functions), learning, remembering, and speeded information processing.

neurocognitive complications A spectrum of disturbances of neurocognitive functions ranging from asymptomatic impairment to frank dementia.

protease An enzyme contained in HIV that is involved in the late stages of HIV maturation.

protease inhibitors Drugs whose antiviral activity derives from their ability to interfere with the HIV enzyme protease.

reverse transcriptase An enzyme carried by HIV that allows it to form viral DNA from its RNA template.

reverse transcriptase inhibitors Drugs whose antiviral properties derive from their ability to interfere with the HIV enzyme reverse transcriptase.

virion A particle of HIV.

Acquired immune deficiency syndrome (AIDS) results from infection with human immunodeficiency virus (HIV). HIV infection is associated with destruction of certain immune cells, most notably the CD4 or "helper" T lymphocytes, leading ultimately to collapse of cell-mediated immunity and consequent susceptibility to various types of "opportunistic" infections and cancers. In the course of HIV disease, virus enters the central nervous system and this can result in disturbances in neurocognitive function—that is, deficits in mental processes such as attention, learning, remembering, problem solving, speed of information processing, and various sensory and motor abnormalities. In addition to the direct effects of HIV on brain function and structure, late complications that involve infection of the brain by other pathogens or development of neoplasia or vascular disturbances can also contribute to neurocognitive complications in late-stage HIV disease.

I. HUMAN IMMUNODEFICIENCY VIRUS

HIV is a member of the lentivirus subfamily of retroviruses. Retroviruses carry their genetic information on RNA (rather than DNA) but require the formation of viral DNA as a step in reproduction. This is accomplished by harnessing the machinery of the cell that the virus infects, wherein the viral enzyme reverse transcriptase facilitates forming DNA from the two strands of viral RNA within the infected host cell. The viral DNA so formed is integrated into the cell's genetic material and directs the host cell to manufacture new viral constituents. Essentially, these consist of a viral core surrounded by a glycoprotein envelope. Steps involved in viral replication are illustrated schematically in Fig. 1.

From a genetic standpoint, there are two types of HIV: HIV-1 and HIV-2. The predominant cause of AIDS worldwide is HIV-1, whereas HIV-2 remains limited to portions of western Africa. HIV-1, in turn, is classified on the basis of genetic analysis into a major group (group M) and an outlier group (group O). Again, the vast majority of infections worldwide are with group M viruses, with subtype B of group M predominating in the industrial world. The greatest diversity of subtypes is found in Africa.

A. Transmission of HIV

In the industrialized world, transmission occurs most commonly through men having sex with men and through sharing of infected "works" by injection drug users. Heterosexual transmission is on the rise in industrialized countries, and it is the most prevalent form of transmission in Africa and India. Transmission also occurs from infected mother to fetus and through the administration of contaminated blood products. The latter is no longer an important risk in industrialized nations but poses a continuing hazard elsewhere. From the standpoint of sexual transmission, both unprotected anal and vaginal intercourse constitute the most risky circumstances, with orogenital sex representing a fairly small risk and other activities such as kissing regarded as essentially risk free.

B. Course of HIV Infection

Not all exposure to HIV, be it through unprotected sex or sharing of needles, results in infection. Whether infection occurs seems to depend on factors such as amount of inoculum, pathogenicity of virus, and host resistance. In terms of sexual transmission, the presence of sores, other venereal disease, and mucosal tears may all facilitate transmission.

Once HIV gains access to a tissue, its replication depends on its ability to enter host cells. Such entry appears to require at least two types of receptors (docking mechanisms on cell surfaces). Sections of viral envelope protein (gp120) are capable of attaching to CD4 receptor sites on host cells. However, full attachment leading to fusion of virus with cell requires a coreceptor of the chemokine type. Dozens of types of chemokine receptors are known, but the most important in terms of HIV infection appear to be CXCR4 [found predominantly on CD4 (T4) lymphocytes] and CCR5 (found on macrophages, dendritic cells, and other cells).

Once virus successfully enters host cells, replication begins. The presence of viral RNA can be found in the blood of infected individuals within 1 week of infection. Initially, there is a rapid proliferation of virus, with as many as 10 billion virions being formed every day, accompanied by massive destruction of CD4

lymphocytes of which millions must be replaced on a daily basis. Within several weeks to months, however, a "steady state" is reached wherein the host's defenses succeed in suppressing viral replication. At this point, viral load in the blood is significantly reduced or may even become undetectable. Antibody to HIV is present, however, and it is the basis for the commonly used HIV test (the enzyme-linked immunosorbent assay).

For reasons that are still incompletely understood, host responses are not ultimately effective in eradicating HIV. Chronically infected host cells (e.g., in lymph nodes) continuously seed HIV, which may gradually undergo further genetic variation (termed quasispecies formation) that may create genotypes that are increasingly more successful in eluding host defenses. Over the years, the host's capacity to mount an effective immune response diminishes, leading to a critical drop in CD4 lymphocyte concentration (normal CD4 counts are on the order of 1000 cells per cubic millimeter; persons with CD4 counts below 200 per cubic millimeter are diagnosed as having AIDS).

The collapse of cell-mediated immunity sets the stage for infection by organisms that are normally held in check by these mechanisms. These include various mycobacteria, fungi, yeasts, protozoa, as well as cancers stimulated by proliferation by other viruses (e.g., Kaposi sarcoma, which is linked to herpesvirus infection).

C. Classification of HIV Disease

AIDS represents the final stage of HIV disease. It is defined as the experience of certain major complications of HIV infection and/or the presence of CD4 cell counts below 200 (Table I). Although the course of HIV infection varies widely, it is typical for a person to experience HIV infection for 10 years before AIDS-defining events occur. Prior to that time, the individual may spend many years as an asymptomatic carrier (CDC stage A; Table I).

D. Scope of the Problem of HIV Infection

The World Health Organization estimates that approximately 33 million people are currently HIV infected worldwide. There are probably about 750,000 HIV-infected persons in the United States, and approximately 600,000 persons have died of the disease in the United States since it was first described in 1981. In some countries of southern Africa, 25% or more of the population is HIV infected.

E. Treatment and Prevention of HIV Disease

In terms of prevention, some efforts have shown singular success (e.g., the virtual elimination of medical transmission in industrialized countries by ensuring the safety of blood supplies and use of universal precautions during medical treatments). On the other hand, the modification of sexual and drug use practices has been more difficult. Some successes have been achieved in selected high-risk groups (e.g., homosexual/bisexual men and some injection drug users participating in needle exchange programs). On a worldwide level, anti-AIDS campaigns have often been limited by political, cultural, socioeconomic, and education factors.

Progress toward developing a vaccine that might prevent the establishment of infection in the first place has been slow because of the great genetic diversity of HIV and its tendency toward rapid mutation. Several vaccines are in the process of being tested, but it remains to be seen whether such vaccines, usually based on the subtype B2 of HIV-1, will be useful in areas of the world where other subtypes predominate.

Considerable progress has been made in the development of antiretroviral medications that interfere at various stages of HIV replication. The earliest example was zidovudine [azidothymidine (AZT)], which interferes with HIV's ability to utilize reverse transcriptase (RT) to form viral DNA. There are now multiple examples of RT drugs of the nucleoside type, of which AZT is an example. Other types of RT inhibitors, termed nonnucleoside RT, have also been developed. Recently, drugs that interfere with another critical step in viral assembly, the protease inhibitors, have become available. The use of combination therapies that interfere at different stages of viral replication has proved effective in lowering plasma viral load to undetectable levels in many individuals that are treated. However, these drug combinations often have very significant undesirable and toxic side effects that may limit their tolerability; additionally, HIV's ability to rapidly mutate can eventually produce quasispecies that are resistant to various drug combinations. It is uncertain whether combination therapies can be effective in controlling viral replication on an indefinite basis. For a list of currently available drugs, see Table II. The U.S. federal guidelines for use of these drugs are shown in Table III.

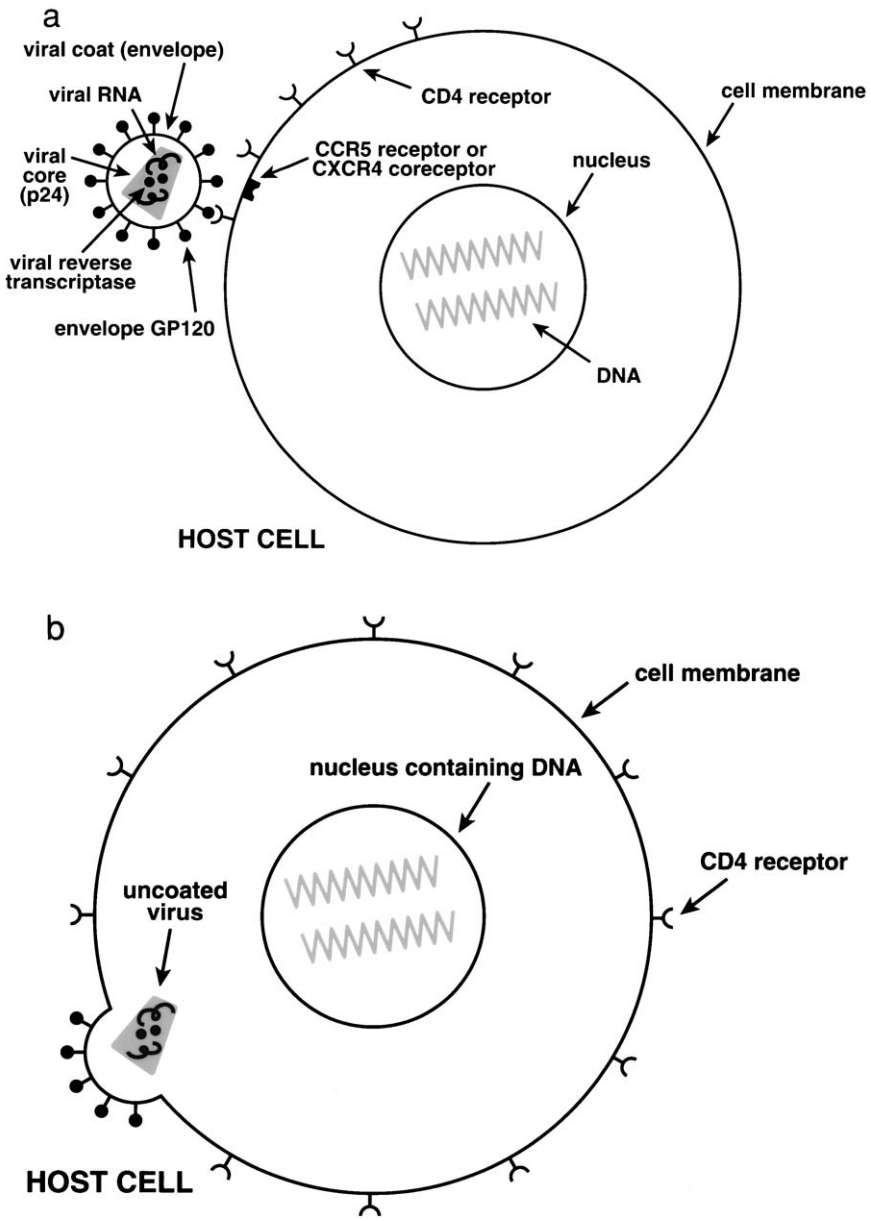


Figure 1 Stages of HIV replication. (a) HIV virion attaching to host cell. A segment of HIV envelope protein gp120 attaches to host cell CD4 receptor. Further process of attachment will involve linkage to a second host cell site, a chemokine coreceptor either of the CXCR4 (lymphocyte) or CCR5 (macrophage, dendritic cell) type. (b) After fusing with host cell wall, the uncoated HIV enters host cell cytoplasm. (c) Using its RNA template, HIV utilizes the viral enzyme reverse transcriptase to manufacture viral DNA within the host cell. This viral DNA ultimately enters the host cell nucleus and becomes integrated into host cell's DNA. (d) The viral DNA that is integrated into host cell's DNA directs formation of a viral RNA transcript, which leads to formation of viral RNA and genomic RNA. Through a translation process viral proteins are formed. (e) Viral constituents are assembled, virus particle buds off from host cell, and viral proteins are clipped into functional units by the viral protease enzyme.

II. HIV INFECTION OF THE BRAIN

There were reports of neuropsychiatric complications in AIDS even before HIV was determined to be the cause of the disease. Brain autopsies of persons dying with AIDS revealed a number of changes beyond those attributable to opportunistic infections or neoplasms. These HIV-related brain changes include

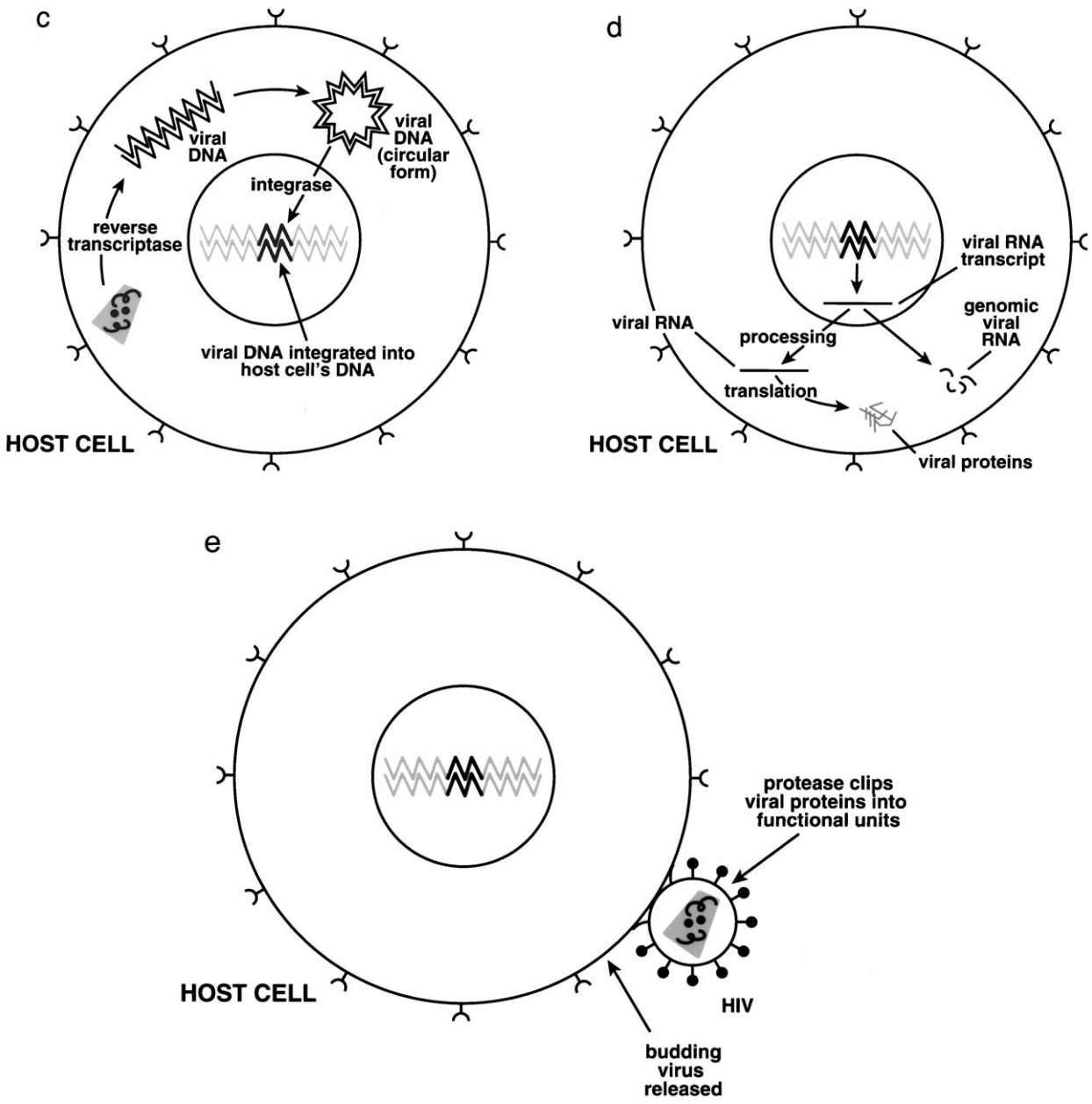


Figure 1 *(continued)*

inflammation, white matter abnormalities (vacuolar myelopathy), and nerve cell loss (poliodystrophy). The inflammatory process can be evidenced by perivascular lymphocytic infiltrates, accumulations of microglia into nodules, formation of multinucleated giant cells, as well as astrogliosis (Fig. 2).

Most persons dying with HIV infection have some detectable HIV in the brain. Although HIV can be found in any brain region, the greatest concentration of virus tends to be in subcortical gray structures (e.g., caudate nucleus) and surrounding white matter. HIV is localized within microglia and multinucleated giant cells but is not found in neurons. Despite this, the neurons of persons dying with HIV often display injury consisting of dendritic simplification and loss of synapses (Fig. 3). This has led to speculation that neural damage, and ultimately neuronal loss, may represent some combination of toxic factors and

Table I
1993 CDC Classification System for HIV Infection[a]

	Clinical categories		
	A[b]	B[c]	C[d]
CD4+ cell count categories	Asymptomatic or lymphadenopathy	Symptomatic but not A or C conditions	AIDS indicator conditions
$>500/mm^3$	A-1	B-1	C-1[e]
$200–499/mm^3$	A-2	B-2	C-2[e]
$<200/mm^3$	A-3[e]	B-3[e]	C-3[e]

[a]From CDC (1992).
[b]Category A includes acute HIV infection, asymptomatic infection, and progressive generalized lymphadenopathy.
[c]Category B includes conditions associated with HIV infection but that were not included in the CDC's 1987 case surveillance definition of conditions associated with severe immunodeficiency. Examples of "less severe" conditions include oropharyngeal candidiasis (thrush), persistent vulvovaginal candidiasis, severe cervical dysplasia or carcinoma, oral hairy leukoplakia, and recurrent herpes zoster involving more than one dermatome.
[d]Category C conditions are those associated with severe immunodeficiency identified in the CDC's 1987 surveillance definition for AIDS.
[e]CDC (1993) AIDS indicator conditions.

inflammatory mechanisms, balanced by protective mechanisms (Fig. 4).

III. MECHANISMS OF HIV NEUROPATHOGENESIS

A. Toxic Factors

It has been suggested that viral products, such as envelope protein gp120, may be neurotoxic. For example, gp120 causes cell death in neuronal cultures, and transgenic mice that constitutively overproduce gp120 show neuronal injury, with loss of dendritic spines and synapses. Lipton and colleagues suggested that gp120 may exert its effect by activating *N*–methyl-D-aspartate (NMDA) receptors, causing influx of calcium. Further evidence of this has been suggested by the fact that memantine, an NMDA antagonist, can block gp120-induced neurotoxicity. Recently, it has also been suggested that gp120 may bind to chemokine receptor sites on neurons, and this process may lead to activation of NMDA receptors and excitotoxicity.

Nonviral toxic products have also been implicated. For example, activated macrophages and astrocytes within the central nervous system are capable of producing increased quantities of quinolinic acid, an excitotoxic molecule. Previous research has found that the amount of quinolinic acid in the cerebrospinal fluid increases with stage of disease and is highest in those with HIV dementia (Fig. 5).

B. Inflammatory Factors

HIV probably enters the central nervous system within macrophages from the periphery, thereby establishing an intracerebral focus of infection in closely related microglia. Microglia (immune cells within the brain closely related to macrophages) may become infected, and these two cell populations feature prominently in two of the hallmark changes found in the brain of persons dying with HIV encephalitis—microglial nodules and multinucleated giant cells (Fig. 2). Astrocytes may also become activated; in addition to producing abnormal quantities of various molecules

Table II
Currently Available Antiretroviral Medications

Medication	Generic name	Brand name	Usual abbreviation
Nucleoside analog reverse transcriptase inhibitors	Abacavir	Ziagen	ABC
	Didanosine	Videx	ddI
	Lamivudine	Epivir	3TC
	Stavudine	Zerit	d4T
	Zalcitabine	Hivid	ddC
	AZT + 3TC	Combivir	
	AZT + 3TC + ABC	Trizivir	
Nonnucleoside analog reverse transcriptase inhibitors	Delavirdine	Rescriptor	DLV
	Efavirenz	Sustiva	EFV
	Nevirapine	Viramune	NVP
Nucleotide analog reverse transcriptase inhibitors	Tenofovir	Viread	TFV
Protease inhibitors	Amprenavir	Agenerase	APV
	Indinavir	Crixivan	IDV
	Lopinavir + RTV	Kaletra	LPV
	Nelfinavir	Viracept	NFV
	Ritonavir	Norvir	RTV
	Saquinavir	Fortovase Invirase	SQV

Table III
U.S. Federal Guidelines for Initial Treatment of HIV Infection

One from group A (highly active protease inhibitors) and one combination from group B (NARTIs):	
Group A	Group B
Indinavir	AZT + ddI
Ritonavir	d4T + ddI
Nelfinavir	AZT + ddC
Saquinavir SGC	AZT + 3TC
	d4T + 3TC
or	
A combination of ritonavir + saquinavir	

that may be neurotoxic (e.g., lymphokines and quinolinic acid), they may also alter their production of trophic factors that are necessary to sustain neurons. Examples of immune signaling molecules that can be damaging to neurons include tumor necrosis factor-alpha (TNF-α) and interleukin-6.

Masliah and colleagues suggested that different neuronal populations may be sensitive to differing types of injury. For example, pyramidal neurons may be particularly sensitive to NMDA-linked excitotoxic injury. However, interneurons, that are also damaged in HIV disease, contain proteins such as calbindin and parvalbumin that tend to protect against disruptions of calcium homeostasis. Masliah and colleagues argue that these neurons may be more vulnerable to damage by inflammatory mechanisms.

C. Protective Factors

Whether or not a neuron is injured may depend not only on the presence of inflammatory and toxic factors but also on protective mechanisms. For example, Sanders and colleagues recently reported that regions of the brain in which fibroblast growth factor (FGF) is expressed show less evidence of neural injury in the context of HIV infection. They suggested that FGF may be involved in altering intracellular signaling so that cascades leading to apoptosis (programmed cell death) are downregulated.

IV. DIAGNOSIS AND CLASSIFICATION OF NEUROCOGNITIVE COMPLICATIONS

The diagnosis of neurocognitive complications associated with HIV-1 infection requires the demonstration that there has been an acquired change in cognitive performance that has occurred since the person became HIV infected and that cannot be explained by other causes. Three levels of impairment have been described.

A. Asymptomatic Neuropsychological Impairment

This mildest form of neurocognitive complication is characterized by subtle changes in cognitive functioning that appear not to interfere in any obvious way in day-to-day functioning. Typically, a person will report feeling that he or she is not as "sharp" as he or she used to be and may complain of mild memory difficulties, difficulties in concentration, or some slight slowing down in mental functions. However, self-report alone is not sufficient to establish a diagnosis because this may be biased by mood disturbance (depression may lead to complaints that are unrelated to actual cognitive functioning) or lack of insight. The diagnosis requires demonstrating at least mild impairments in two different cognitive areas on comprehensive neuropsychological testing (Table IV).

B. Mild Neurocognitive Disorder

As described by the American Academy of Neurology, mild neurocognitive disorder, also known as minor cognitive motor disorder (MCMD), represents another mild form of cognitive abnormality that, however, is of sufficient magnitude that there is some impairment in at least one area of life functioning, such as inefficiency at work or management of domestic or financial affairs (Table V). As with asymptomatic neuropsychological impairment, comprehensive neuropsychological testing is necessary to establish that there are at least two areas of cognitive function in which performance is at least one standard deviation below that expected for persons of similar age, education, and sociodemographic background.

C. HIV-Associated Dementia

Dementia is diagnosed when there is impairment in at least two (and usually multiple) cognitive areas of sufficient severity to interfere markedly with day-to-day life (Table VI). Persons with HIV dementia are

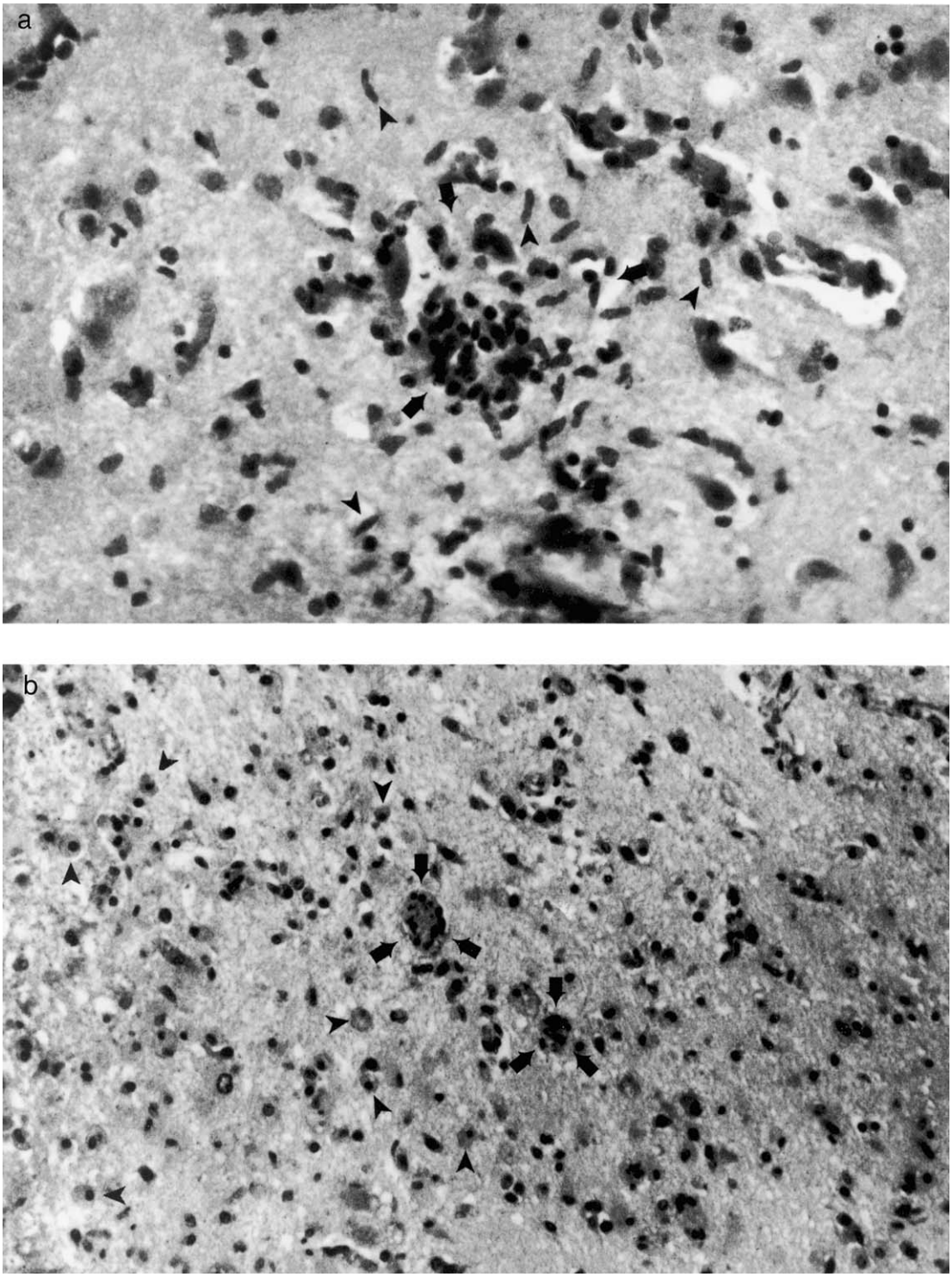


Figure 2 Microscopic anatomy of HIV brain disease. (a) Hematoxylin and eosin-stained, paraffin embedded section of AIDS brain tissue. A large microglial nodule is present in the center of the section (surrounded by three arrows). Numerous elongated microglial nuclei are present both within the nodule and in the surrounding tissue (arrowheads). Original magnification, × 600. (b) HIV p24 immunoperoxidase-stained, paraffin embedded section of AIDS brain tissue. Several large multinucleated giant cells (surrounded by arrows) contain black, amorphous immunoprecipitate of density similar to that of the oval nuclei. Numerous mononuclear macrophages (arrowheads) are distributed throughout the tissue. Original magnification, × 300 (courtesy of Clayton A. Wiley, University of Pittsburgh).

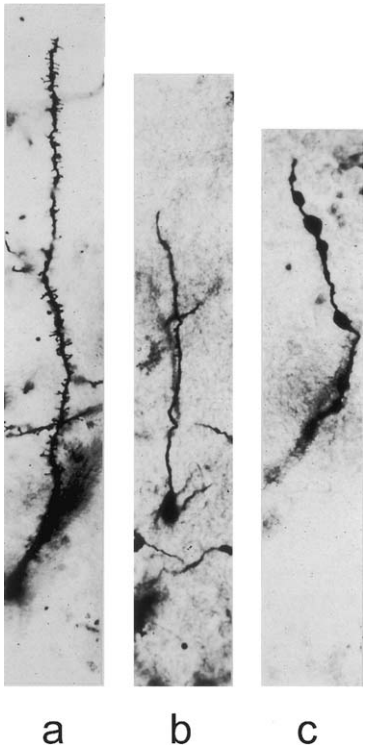


Figure 3 Dendritic pathology in HIV. Golgi-impregnated dendritic segments from control (a) and two cases with HIV dementia. There is loss of dendritic spines (b and c) and distortion and abnormal vacuolation of dendritic segment (c) (courtesy of Drs. Eliezer Masliah, University of California, San Diego, and Clayton A. Wiley, University of Pittsburgh).

typically unemployable and may have difficulty with independent living. The symptoms include significant psychomotor slowing and incoordination, marked inattention, slowness in information processing, difficulty in learning new information, and impairment in fluency. In some cases, the dementia may be accompanied by psychosis, including manic and paranoid phenomena. Some patients become markedly agitated, and others become apathetic and sometimes mute. The onset of frank dementia is a poor prognostic sign in HIV disease and is predictive of near future death in many instances.

V. EPIDEMIOLOGY OF NEUROCOGNITIVE COMPLICATIONS

Neurocognitive complications become more prevalent with progression of HIV disease. Thus, HIV dementia almost never occurs in stages A and B of HIV disease, but it may be found in 4–7% of those with frank AIDS. In earlier stage disease, neurocognitive impairment, if it is seen, is typically mild and may have a relapsing–remitting course. Observers continue to disagree on the prevalence of these complications in earlier stage disease. Data from the San Diego HIV Neurobehavioral Research Center are presented in Fig. 6. It can be seen that the rate of impairment increases with stage of disease. In the medically asymptomatic form of disease, only the mildest form of neurocognitive impairment tends to be observed (i.e., asymptomatic neuropsychological impairment). Minor cognitive motor disorder becomes more prevalent in stage B and stage C disease.

VI. ASSOCIATED FACTORS AND CORRELATES OF HIV NEUROCOGNITIVE COMPLICATIONS

It is difficult to predict who will and who will not develop neurocognitive complications. Clearly, progression of disease is one factor, as the data in Fig. 6 illustrate. However, general markers of disease progression, such as a decrease in CD4 count and an increase in beta-2 microglobulin (a marker of immune activation), are not powerful predictors of future cognitive complications. Plasma viral load is associated with neuropsychological impairment, but measuring the concentration of HIV in the cerebrospinal fluid (CSF) may provide a more specific indicator. For example, Ellis and colleagues reported that in patients with AIDS there was an association between CSF viral load, but not plasma viral load, and likelihood of neurocognitive impairment.

Markers of immune activation in the CSF have also been associated with neurocognitive impairment (e.g., an increase in CSF neopterin and CSF beta-2 microglobulin). As noted previously, concentration of the

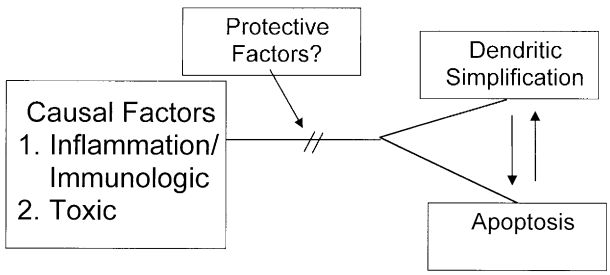


Figure 4 General model for HIV neural injury.

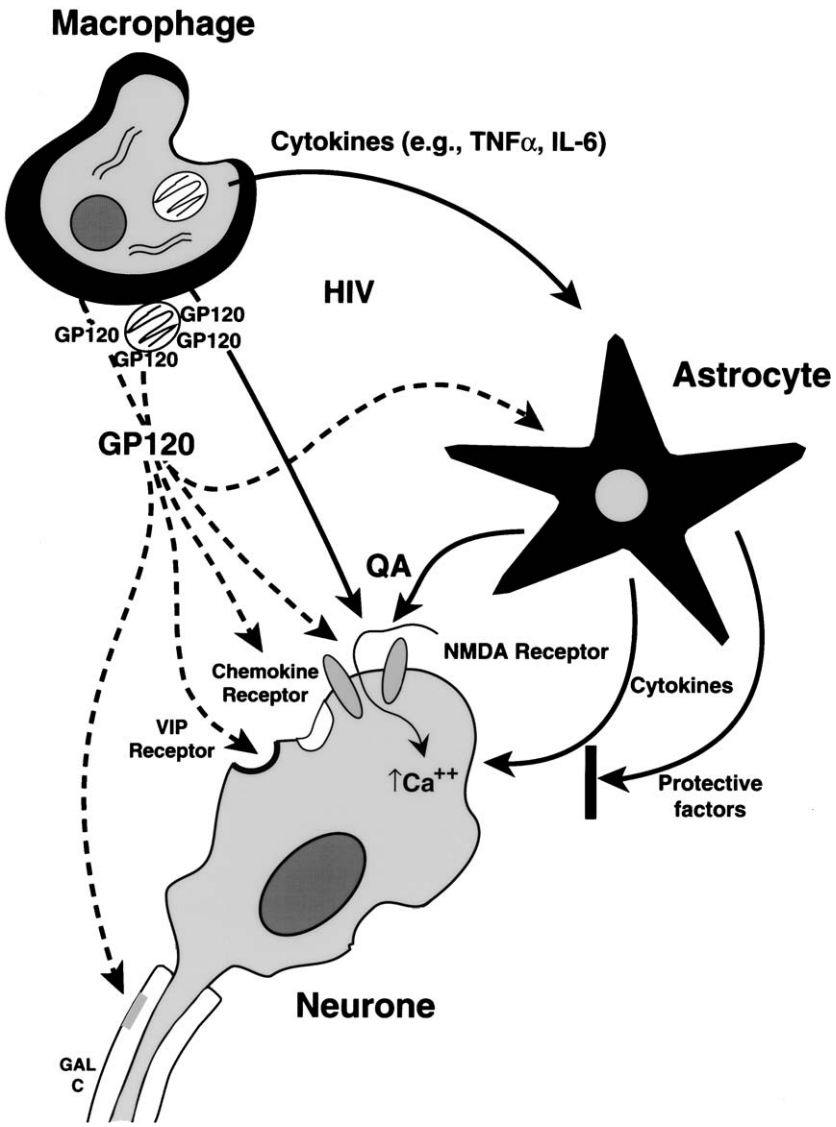


Figure 5 Some putative pathways in HIV neuropathogenesis.

excitotoxic compound quinolinic acid is also elevated in the CSF of those with HIV dementia; however, it is not clear whether quinolinic acid is another marker of immune activation or is somehow causally related to the neurocognitive impairment.

In terms of risk factors, contrary to expectation, there is no clear-cut association between injection drug use and heightened risk of neurocognitive complications. Although the rates of neuropsychological impairment tend to be higher among injection drug users generally, there does not seem to be an interaction between drug abuse and HIV status. An exception to this rule may be dependence on central stimulant drugs, especially methamphetamine. Preliminary observations suggest that history of methamphetamine dependence may enhance the likelihood of HIV-associated neurocognitive impairment perhaps because of some commonalities in mechanisms of neural injury that involve excitotoxicity.

There has been speculation that some of the subtle neuropsychological impairment found in HIV-infected persons might be due to depression, fatigue, or other nonspecific factors. This matter has received considerable exploration and the overall conclusion is that depression and medical symptoms generally do not explain the neuropsychological impairment. For

Table IV
Research Definition for HIV Neuropsychological Impairment

Performance at least 1.0 standard deviation below age–education norms in at least two different cognitive areas[a]
The impairment cannot be explained by comorbid conditions (e.g., substance abuse and medications)
The impairment does not occur solely as part of a delirium (e.g., due to CNS toxoplasmosis, lymphoma, or CMV)

[a]At least five of the following ability areas must be assessed: attention/information processing, language, abstraction/executive, complex perceptual motor, learning, recall/forgetting, motor skills, and sensory.

example, although depressive symptoms increase with frequency with disease progression, there is not a strong association between such symptoms and the likelihood of finding neurocognitive impairment. Similarly, Heaton and colleagues found that neuropsychological impairment could not be explained simply on the basis of fatigue and constitutional symptoms.

VII. COURSE OF NEUROCOGNITIVE COMPLICATIONS

HIV-related neurocognitive complications differ from those seen in degenerative disorders, such as Alzheimer's disease or Parkinson's disease. Most persons with mild impairments do not progress to develop dementia; indeed, many recover and some have a relapsing–remitting course. This is illustrated by the data in Figs. 7 and 8, derived from research at the San Diego HIV Neurobehavioral Research Center (HNRC). In Fig. 7, it can be seen that after a period of 1 year only about half of the cases judged to have MCMD or asymptomatic neuropsychological impairment remain in the same category. About one-fifth improve, and a small proportion worsen. When data over a period of 5 years were considered, it was found that about one-fourth of HIV-infected persons had a "wobbly" or relapsing–remitting course. This pattern of waxing and waning symptomatology is consistent with the presumed underlying etiology, which is thought to be linked to periodic flare-ups of viral activity or immune activation within the central nervous system.

VIII. QUALITATIVE FEATURES OF NEUROCOGNITIVE COMPLICATIONS

HIV neurocognitive complications are often described as having "subcortical" features. This means that the pattern of impairment is somewhat reminiscent of that seen in neurological diseases that affect primarily the subcortical structures or white matter and possibly pathology involving frontostriatal circuits (e.g., Huntington's disease, Parkinson's disease, and multiple sclerosis). Persons with this pattern of neuropathology tend to have difficulties in psychomotor abilities, speed of information processing, initiation, divided

Table V
Criteria for Mild Neurocognitive Disorder[a]

Acquired impairment in cognitive functioning, involving at least two ability domains, documented by performance of at least 1.0 standard deviation below the mean for age–education-appropriate norms on standardized neuropsychological tests. The neuropsychological assessment must survey at least the following abilities: verbal/language, attention/speeded processing, abstraction/executive, memory (learning and recall), complex perceptual–motor performance, and motor skills.
The cognitive impairment produces at least mild interference in daily functioning (at least one of the following):
Self-report of reduced mental acuity or inefficiency in work, homemaking, or social functioning.
Observation by knowledgeable others that the individual has undergone at least mild decline in mental acuity with resultant inefficiency in work, homemaking, or social functioning.
The cognitive impairment has been present at least 1 month.
The cognitive impairment does not meet criteria for delirium or dementia.
There is no evidence of another preexisting cause for the MND.[b]

[a]As defined by Grant and Atkinson (1995).
[b]If the individual with suspected mild neurocognitive disorder (MND) also satisfies criteria for a major depressive episode or substance dependence, the diagnosis of MND should be deferred to a subsequent examination conducted at a time when the major depression has remitted or at least 1 month has elapsed following termination of dependent-substance use.

Table VI
Criteria for HIV Dementia[a]

Marked acquired impairment in cognitive functioning, involving at least two ability domains (e.g., memory and attention): typically, the impairment is in multiple domains, especially in learning of new information, slowed information processing, and defective attention/concentration. The cognitive impairment can be ascertained by history, mental status examination, or neuropsychological testing.
The cognitive impairment produces marked interference with day-to-day functioning (work, home life, and social activities).
The marked cognitive impairment has been present for at least 1 month.
The pattern of cognitive impairment does not meet criteria for delirium (e.g., clouding of consciousness is not a prominent feature) or, if delirium is present, criteria for dementia need to have been met on a prior examination when delirium was not present.
There is no evidence of another, preexisting etiology that could explain the dementia (e.g., other CNS infection, CNS neoplasm, cerebrovascular disease, preexisting neurological disease, or severe substance abuse compatible with CNS disorder).

[a]As defined by Grant and Atkinson (1995).

attention, learning difficulties, difficulties in retrieval of information but not accelerated forgetting, and some executive dysfunction. To the extent that there are language problems, these are more in the area of fluency rather than naming. In contrast, the so-called "cortical" dementias (Alzheimer's disease and multi infarct dementia) are characterized by severe memory impairment that includes difficulty in learning new information as well as rapid forgetting, problems in naming and comprehension, and disturbances of praxis.

Thus, HIV-infected persons with asymptomatic neuropsychological impairment or MCMD tend to have mild learning difficulties, some problems with attention, difficulties with speed of information processing, some psychomotor slowing, and occasionally, difficulties with fluency. Although this may be the most typical pattern, it should be noted that since HIV-associated neurological injury can be widespread in the brain, there are some cases that have symptoms that are more cortical in nature, and others that have mixed features.

IX. SIGNIFICANCE OF NEUROCOGNITIVE COMPLICATIONS

Despite typically being mild and fluctuating in nature, HIV-associated complications can affect multiple aspects of life. For example, Heaton and colleagues demonstrated that those with impairment were twice as likely to be unemployed as persons without impairment. Also, even among the employed, those with mild impairment were performing at a level less

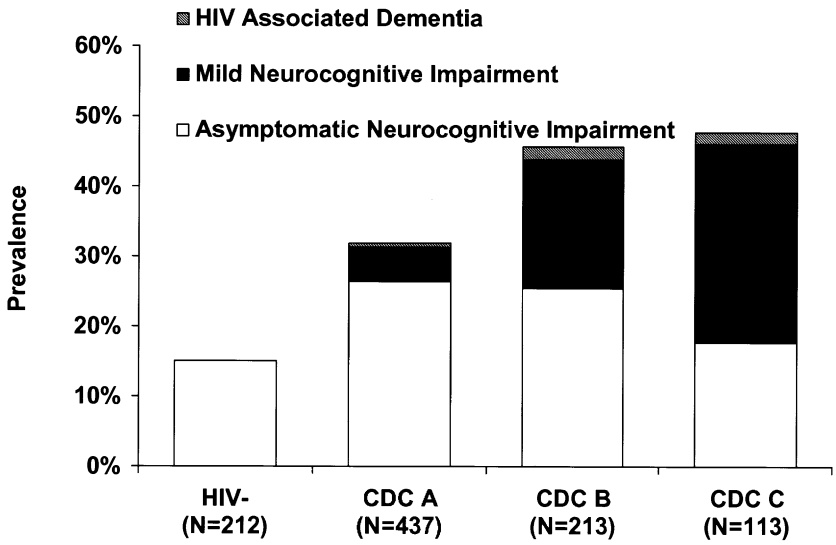


Figure 6 Prevalence of HIV neurocognitive complications at different stages of disease.

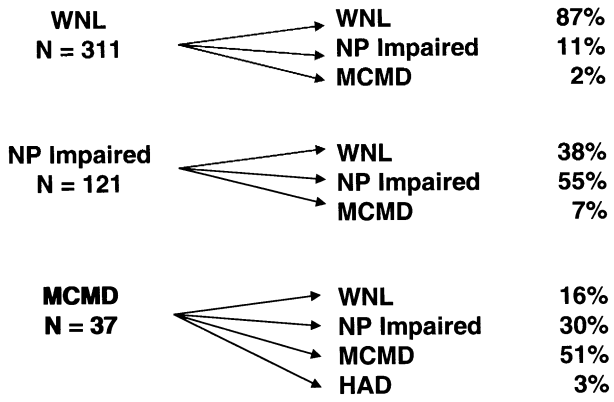


Figure 7 One-year progression of neurocognitive complications. HAD, HIV-1-associated dementia; NP, neuropschological; MCMD, minor cognitive motor disorder; WNL, within normal limits.

than expected based on uninfected comparison groups and also on infected but not impaired controls. Neuropsychological impairment associated with HIV may also affect important life activities, such as driving and medication management. For example, Marcotte and colleagues, utilizing the driving simulator, noted that those with impairment had more simulated accidents than the unimpaired. Recent data suggest that such individuals also had more actual on-road accidents and incidents. Recent observation at the HNRC also indicates that those with neuropsychological impairment may have more difficulties with managing their antiretroviral medication regimens. For example, Fig. 9 shows that more of the mildly impaired patients fail to take their medications as scheduled or as directed.

The presence of neurocognitive complications also predicts earlier death. For example, Ellis and colleagues noted that patients with MCMD had a median survival of 2.2 years, and those with asymptomatic neuropsychological impairment 3.8 years, versus 5.1 years for those who were neurocognitively normal. Adjustment of the survival analysis for CD4 count and other disease indicators revealed an independent effect for neurocognitive impairment. The mechanism for this remains unclear, but similar data were also reported from the Columbia University cohort.

X. TREATMENT AND PREVENTION OF NEUROCOGNITIVE COMPLICATIONS

The extent to which antiretroviral therapy protects against neurocognitive complications remains uncertain. Historical data indicate that the advent of the first antiretroviral (AZT) was associated with reduction in diagnosis of AIDS-associated dementia. However, not all investigators have reached the same conclusion.

The advent of potent antiretroviral combination therapies holds with it the promise of reducing viral load to unmeasurable or very low levels. Studies are under way to determine whether such viral load reductions are associated with neurocognitive improvement. Preliminary observations indicate that reduction of viral load in the CSF may be more specifically associated with neurocognitive improvement than reduction in plasma load. If correct, this would raise questions about the differential efficacy of new antiretrovirals. In other words, the possibility arises that agents may be extremely potent in lowering peripheral viral load but, because of their poor

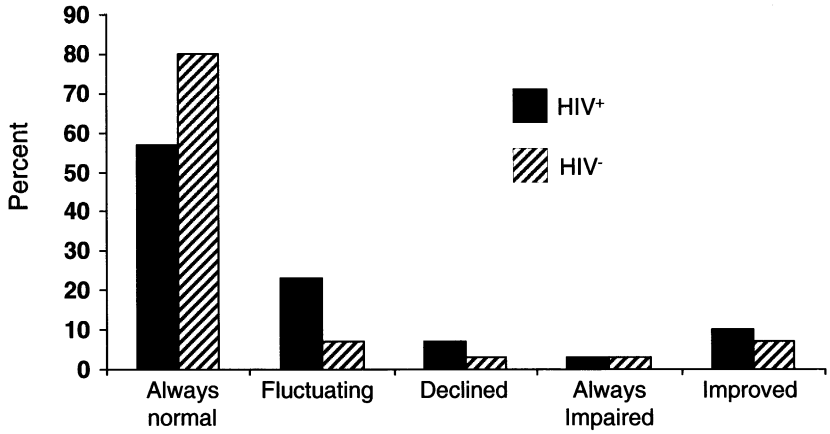


Figure 8 Neuropsychological course in participants with 5 years of annual assessments.

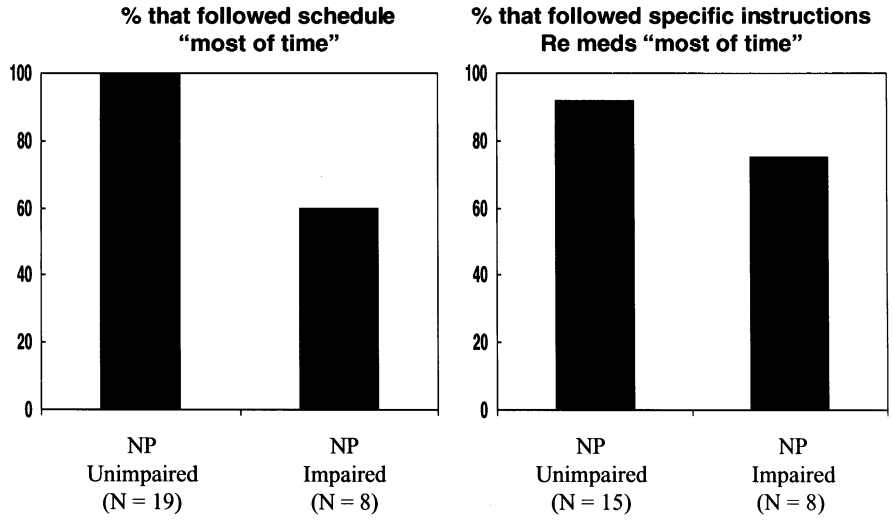


Figure 9 HIV neurocognitive impairment associated with worsened adherence to antiretroviral treatment.

penetration into the central nervous system, they may not be equally effective with regard to CNS complications.

Research is also ongoing to test drugs that may impact the putative mechanisms of neural injury. For example, a treatment trial is currently under way with the NMDA blocker memantine to determine whether reducing activity at glutamatergic receptor sites may be associated with neurocognitive improvement. A trial of peptide T, which putatively competes with viral envelope protein gp120 at neural receptor sites (e.g., receptor sites for VIP and possibly chemokine receptor sites), produced mixed results. Overall, it appeared that peptide T did not yield neurocognitive improvement; however, an analysis based on a subset of definitely impaired HIV-infected individuals did suggest some benefit. Thus, the question of peptide T's effectiveness remains unresolved. Pentoxyfylline, which blocks the action of TNF-α, was not promising in improving neurocognitive functioning in a preliminary trial.

XI. SUMMARY

Neurocognitive complications commonly occur in HIV disease, and their prevalence increases with disease progression. The impairments are typically mild in nature, often wax and wane, and affect primarily the capacity to learn new information, cognitive and psychomotor speed, and attention—a profile reminiscent of "subcortical" dementias such as those associated with Huntington's disease, Parkinson's disease, and white matter dementias. One of the fundamental substrates of the cognitive impairment is neural injury, including loss of dendritic spines and synaptic simplification. The mechanism of injury may involve the toxic effects of viral products such as gp120 as well as inflammatory mechanisms involving perhaps abnormal expression of various lymphokines. Protective factors may also be important, although the role of FGF and other trophic factors has yet to be established. Alhough they are usually mild in nature, HIV-associated neurocognitive disturbances can have substantial effects on day-to-day life, including employment, and day-to-day tasks such as medication management and driving skills. The presence of impairment is also associated with earlier mortality. Currently available antiretroviral drug combinations have increased survival of patients with HIV, but their potential to avert or improve neurocognitive complications is not conclusively established.

Acknowledgments

The work summarized in this article was performed in the context of the HIV Neurobehavioral Research Center (HNRC) (principal support by National Institute of Mental Health Grant MH45294). The San Diego HIV Neurobehavioral Research Center (HNRC) group is affiliated with the University of California, San Diego, the Naval Hospital, San Diego, and the San Diego Veterans Affairs Healthcare System and includes the following: director, Igor Grant,

codirectors, J. Hampton Atkinson, J. Allen McCutchan; center manager, Thomas D. Marcotte; Naval Hospital, San Diego, Mark R. Wallace; neuromedical component, J. Allen McCutchan, Ronald J. Ellis, Scott Letendre, and Rachel Schrier; neurobehavioral component, Robert K. Heaton, Mariana Cherner, and Julie Rippeth; imaging component, Terry Jernigan, and John Hesselink; neuropathology component, Eliezer Masliah; clinical trials component, J. Allen McCutchan, J. Hampton Atkinson, Ronald J. Ellis, and Scott Letendre; data management unit, Daniel R. Masys and Michelle Frybarger, (data systems manager); statistics unit, Ian Abramson, Reena Deutsch, and Tanya Wolfson.

See Also the Following Articles

AUTOIMMUNE DISEASES • BORNA DISEASE VIRUS • CANCER PATIENTS, COGNITIVE FUNCTION • CEREBRAL WHITE MATTER DISORDERS • COGNITIVE REHABILITATION • DEMENTIA

Suggested Reading

Centers for Disease Control (1992). 1993 revised classification system for HIV infection and expanded surveillance case definition for AIDS among adolescents and adults. *Morbidity Mortality Weeky Rep.* **41,** 1–19.

Ellis, R. J., Hsia, K., Spector, S. A., Nelson, J. A., Heaton, R. K., Wallace, M. R., Abramson, I., Atkinson, J. H., Grant, I., McCutchan, J. A., and the HIV Neurobehavioral Research Center Group (1997). Cerebrospinal fluid human immunodeficiency virus type 1 RNA levels are elevated in neurocognitively impaired individuals with acquired immunodeficiency syndrome. *Ann. Neurol.* **42,** 679–688.

Grant, I., and Atkinson, J. H. (1999). Neuropsychiatric aspects of HIV infection and AIDS. In *Kaplan and Sadock's Comprehensive Textbook of Psychiatry/VII* (B. J. Sadock and V. A. Sadock, Eds.), pp. 308–335. Williams & Wilkins, Baltimore.

Grant, I., Atkinson, J. H., Hesselink, J. R., Kennedy, C. J., Richman, D. D., Spector, S. A., and McCutchan, J. A. (1987). Evidence for early central nervous system involvement in the acquired immunodeficiency syndrome (AIDS) and other human immunodeficiency virus (HIV) infections: Studies with neuropsychologic testing and magnetic resonance imaging. *Ann. Intern. Med.* **107,** 828–836.

Grant, I., Olshen, R. A., Atkinson, J. H., Heaton, R. K., Nelson, J., McCutchan, J. A., and Weinrich, J. D. (1993). Depressed mood does not explain neuropsychological deficits in HIV-infected persons. *Neuropsychology* **7,** 53–61.

Heaton, R. K., Velin, R. A., McCutchan, J. A., Gulevich, S. J., Atkinson, J. H., Wallace, M. R., Godfrey, H. P. D., Kirson, D. A., Grant, I., and the HNRC group (1994). Neuropsychological impairment in human immunodeficiency virus-infection: Implications for employment. *Psychosom. Med.* **56,** 8–17.

Heaton, R. K., Grant, I., Butters, N., White, D. A., Kirson, D., Atkinson, J. H., McCutchan, J. A., Taylor, M. J., Kelly, M. D., Ellis, R. J., Wolfson, T., Velin, R., Marcotte, T. D., Hesselink, J. R., Jernigan, T. L., Chandler, J., Wallace, M., Abramson, I., and the HNRC group (1995). The HNRC 500—Neuropsychology of HIV infection at different disease stages. *J. Int. Neuropsychol. Soc.* **1,** 231–251.

Heaton, R. K., Marcotte, T. D., White, D. A., Ross, D., Meredith, K., Taylor, M. J., Kaplan, R., and Grant, I. (1996). Nature and vocational significance of neuropsychological impairment associated with HIV infection [Published erratum appears in *Clin. Neuropsychol.* **10,** 236, 1996]. *Clin. Neuropsychol.* **10,** 1–14.

Lipton, S. T., and Gendelman, H. E. (1995). Dementia associated with the acquired immunodeficiency syndrome. *N. Engl. J. Med.* **332,** 934–940.

Marcotte, T. D., Heaton, R. K., Wolfson, T., Taylor, M. J., Alhassoon, O., Arfaa, K., Grant, I., and the HNRC group (1999). The impact of HIV-related neuropsychological dysfunction on driving behavior. *J. Int. Neuropsychol. Soc.* **5,** 579–592.

Masliah, E., Ge, N., Achim, C. L., DeTeresa, R., and Wiley, C. A. (1996). Patterns of neurodegeneration in HIV encephalitis. *NeuroAIDS* **1,** 161–173.

Masliah, E., Ge, N., and Mucke, L. (1996). Pathogenesis of HIV-1 associated neurodegeneration. *Crit. Rev. Neurobiol.* **10,** 57–67.

Masliah, E., Heaton, R. K., Marcotte, T. D., Ellis, R. J., Wiley, C. A., Mallory, M., Achim, C. L., McCutchan, J. A., Nelson, J. A., Atkinson, J. H., Grant, I., and the HNRC group (1997). Dendritic injury is a pathological substrate for human immunodeficiency virus-related cognitive disorders. *Ann. Neurol.* **42,** 963–972.

Mayeux, R., Stern, Y., Tang, M.-X., Todak, G., Marder, K., Sano, J., Richard, M., Stein, Z., Ehrhardt, A., and Gorman, J. (1993). Mortality risks in gay men with human immunodeficiency virus infection and cognitive impairment. *Neurology* **43,** 176–182.

McArthur, J. C., and Grant, I. (1998). HIV neurocognitive disorders. In *Neurology of AIDS* (H. E. Gendelman, S. Lipton, L. Epstein, and S. Swindells, Eds.), pp. 499–523. New York, Chapman & Hall, New York.

Sanders, V., Everall, I., Johnson, R., and Masliah, E. (2000). Fibroblast growth factor modulates HIV co-receptor CXCR4 expression by neural cells. *J. Neurosci. Res.* **59,** 671–679.

Homeostatic Mechanisms

PIERRE-MARIE LLEDO
Pasteur Institute

GLOSSARY

cybernetics A term introduced by the mathematician N. Wiener (1894–1964) from the Greek *kubernetes* (''steersman''). This term, which defines a theory of feedback systems (namely, self-regulating systems), could be applicable not only to living systems but also to machines.

emotion From the earliest philosophical speculations onwards, emotion has often been seen as interfering with rationality, as a remainder of our presapiens inheritance: Emotions seem to represent unbridled human nature ''in the raw.'' Emotions are the product of an individual's own processing of occurrences on the basis of his or her own prior history and biology, and emotional response activates neural and neuroendocrine effector systems and leads to a variety of short- and long-term consequences that may or may not result in disease.

homeostasis The physiologist C. Bernard (1813–1878) introduced the concept of the constancy of the *milieu intérieur*. This internal milieu ensures the biological unity of the organism and confers a certain autonomy relative to the external milieu. The term homeostasis was coined some years later by W. B. Cannon (1871–1945). He developed the concept of homeostasis, which in modern terminology is the feedback control of servo-systems. This concept was not mathematically expressed until the 1940s, when it became the basis of cybernetics. Today, the term homeostasis refers to the adaptative response of an organism and tends to be substituted by a newer term, allostasis, which means ''stability through anticipatory change''; the long-term consequences of continued demand on the physiologic response are referred to as allostatic load.

hypothalamus A brain area that encompasses the most ventral part of the diencephalon where it forms the floor and, in part, the walls of the third ventricle. The hypothalamus consists of several nuclei that form a neuronal continuum. It plays a central role in homeostasis by controling the autonomic nervous system, the neuroendocrine system through its control of both the anterior and posterior parts of the pituitary gland, and the motivational states.

limbic system In 1878, P. Broca was the first to describe an annular ring of tissue on the medial face of the cerebral hemisphere that represents the free edge of the cerebral cortex. He named this part of the brain *le grand lobe limbique* (''the great limbic lobe''), which led to the concept of the limbic system. This system includes the hippocampal formation, entorhinal area, olfactory regions, hypothalamus, and amygdala. Functionally, the limbic system is generally thought to be concerned with visceral processes, particularly those associated with the emotional status of the organism. In fact, the interaction of all the structures in the complex, from the entorhinal area to the hypothalamus, plays a major role in the elaboration of the final actions of an organism in a particular environment and in the formation of adaptive behavior patterns.

How does the body adapt to environmental conditions? How does it organize its reactions to the world and other people? What are desire, pleasure, and pain? Going beyond the traditional dichotomies of body and soul, or reasonable brain and passionate body, we shall deal with what is called the constancy of the internal milieu for a human being brought by homeostatic mechanisms. Are we reductionist when we decide to approach the molecular mechanisms of our emotions? It must be recognized that while we announce that being is not just the sum total of the parts of the machine, this very machine shows itself to be increasingly complex as it gradually yields the secrets of its inner workings. It is indeed astonishing that we can analyze networks of billions of interconnected elements of an extraordinary

complexity and at the same time develop a single molecule capable of causing or correcting the most inextricable disorders of the mind. Therefore, the scientist is not the only one accused of reductionism: Nature provides an example of radical simplification.

Even a unicellular being has a certain degree of freedom between the information receptors on the cell surface and the effectors on the inside. The evolution of a species consists of a gradual increase in the number of intermediaries between information from the outside world and effectors responsible for actions. An animal's freedom increases in proportion to the number of these intermediaries. However, it is only because the liquid element and the substances it transports bring a solution of continuity to cell organization that this freedom is possible.

We shall therefore deal with the constancy of the internal milieu. The external milieu was a Greek invention, but the concept of internal milieu was introduced by C. Bernard (1813–1878). For Greek doctors and their disciples, man lived in harmony with nature. Temperament fixed the conditions of this harmony. However, the living being had no real identity or biological unity: The humors were nothing but a kind of reproduction, inside the animal, of the surrounding natural elements; there was no substantial difference between nutrients and living matter. The internal milieu ensures the biological unity of the animal and confers a certain autonomy relative to the external milieu. It is supposed to reconstitute around the cells the characteristics of the original marine environment.

In homeostasis, any departure from the norm draws mechanisms into play that tend to bring the trouble spot back to its initial state. Passions could thus be interpreted as a kind of neurosis of the normal, itself a fictitious immobile system of reference. In fact, behind the impassivity of the internal milieu a confused mesh of false constants is hidden, all of which are more or less dependent and variable from one species to the next, from one individual to the next, and, within each individual, from one situation to the next.

We can see the kind of safeguard that the constant agitation of the humors of the internal milieu offers the nervous system and its operational flexibility: Perhaps the brain runs the risk of falling victim (losing its soul?) to such a commotion. For its own protection, it can organize its own disorder: The brain–gland reveals itself as grand master of the humors by its multiple secretions of neurohormones. Like the brain–machine, the humoral brain simultaneously acts as the passionate victim and the orchestrator of its own passion.

I. THE INTERNAL MILIEU

Homeostasis is a widely and somewhat loosely used term for describing all kinds of responses following the principle of negative feedback control. The main concept of homeostasis is founded in the production of stability in dynamic systems by negative feedback. This concept is the basis of cybernetics, whose founding fathers were N. R. Ashby and G. Walter in the 1950s. However, the term homeostasis was coined years before cybernetics by the founders of modern physiology—C. Bernard, W. B. Cannon, and W. R. Hess. In Cannon's germinal book, *Wisdom of the Body*, the basic idea of feedback as a fundamental physiological principle is stated. In this context, constancy in the internal environment of the body is the result of a system of control mechanisms that limit the variability of body states. The internal milieu takes the form of a certain number of volumes called regulated variables. Without regulation, the changes in the external milieu and the functioning of the cells would make these volumes vary, when the very survival of the organism depends on their stability. Hence, stability is obtained as a result of a regulating system comprising several subsystems, each of which is subjected to control mechanisms and responsible for controlled variables. A regulated variable thus remains fixed within strict limits because of the intervention of controlled variables that have a much wider scope for variation. It is clear that this is an extremely important principle for almost all physiological processes as well as for the guiding of skilled behavior. Indeed, such a concept serves as a theoretical basis for the physiology of regulation. However, homeostasis, in the sense of constancy, does not adequately describe normal physiology, in which blood pressure, heart rate, endocrine output, and neural activity are continually changing—from sleeping to waking—in response to external factors and in anticipation of future events. At all times in the daily cycle, these parameters are maintained within an operating range in response to environmental challenges. The operating range, and the ability of the body to increase or decrease vital functions to a new level within that range upon challenge, particularly in anticipation of a challenge, has been defined as *allostasis* (or "stability through change"). The operating range for most physiological systems is larger in health than in disease, and it is larger in younger compared to older individuals. Exceeding this range can lead to disaster, as is the case when exertion leads to a myocardial infarction.

It is worth mentioning that the cell theory is inseparable from the concept of the internal milieu. The organism is composed of a host of cells that are scattered or grouped together in tissues. Each cell, individualized by its plasmatic membrane, plays out its fate under the genetic control of the nucleus. These cells are bathed with water-like fluid that forms the extracellular space, providing a medium for diffusion and homogenization around the cells. Bernard, comparing the weight of a mummy with that of a living human being of the same size, estimated the water content of the latter to be 90%—to be more precise, two-thirds water for one-third dry matter. This extracellular space, including the blood and lymph, is indeed a unifier of the organism. Unlike the external environment, which is subjected to uncontrollable change, the internal milieu oscillates slightly around normal values. Thus, the autonomy acquired by the organism relative to its external environment gives it an independent and free life since the constancy of the internal milieu does not mean fixity but rather a possibility to evolve.

The regulated variables define the constancy of the internal milieu. The most important are the gas content of the blood, acidity or pH, temperature, sugar content, blood pressure, and osmotic pressure. For example, we know that the more salty a solution, the higher its osmotic pressure. If for any reason an animals loses water, the salt concentration in the internal milieu (i.e., the osmotic pressure) increases. Because it is a regulated variable, osmotic pressure will be kept constant by regulating mechanisms: diminishing the outflow of water and/or increasing the intake. The outflow is reduced by slowing down the elimination process through the kidneys. Vasopressin, an antidiuretic hormone secreted by the brain, performs this function. The amount of this antidiuretic hormone circulating in the blood is a controlled variable that increases in responses to any increase in osmotic pressure. This is an example of hormonal regulation. The best way to increase intake is to drink. Beyond a certain level, the increase in osmotic pressure causes thirst and an urgent need to drink. Therefore, the regulating mechanisms can be of two kind: hormonal or behavioral. Despite the dry heat of the desert, a camel does not suffer from a much higher osmotic pressure than that of a bartender; it merely possesses more powerful regulating mechanisms that are adapted to the external milieu and that give the controlled variables, diuresis and water intake, a wider scope for action.

Nevertheless, there is a hierarchy to be respected. The most important constants must be maintained at all costs, even if this means sacrificing one of the lower orders. In case of need, a regulated variable can become a controlled variable. For example, blood pressure is constant, but if the oxygen content of the blood is endangered, because it is a hierarchical superior it will rise in order to provide a higher flow of gas and will thus temporarily become a controlled variable instead of remaining a regulated variable.

The internal milieu defined by Bernard—the blood and the fluids in which the cells are bathed—is thus the unifier of the organism. The cell draws from the extracellular fluid the nutrients its need, the fuel and oxygen that provide its energy, and the chemical factors that keep it in working order. It discharges into this milieu its waste and the produce of its activity. Here, we have Bernard's second idea, internal secretion, which is inseparable from the concept of the internal milieu. He discovered internal secretion while describing the glycogenic function of the liver. The hepatic cell draws from its reserves of glycogen the sugar that the organism needs and reintroduces it into the bloodstream. Internal secretion, which differs from excretion, demands a fluid medium that can receive the cell's outpourings. The term endocrinology, introduced by N. Pende (1909) to refer to the study of internal secretions, is now used only for the secretions of the so-called vascular glands, which are now called the endocrine glands. The function of these glands, which is much narrower than that of internal secretion, refers to a cellular secretion with no strictly metabolic function but that has a communicative role.

Although virtually all of the brain is involved in homeostasis, neurons controlling the internal environment are mainly concentrated in the hypothalamus, a neuronal structure located at the interface of the brain and peripheral functions. Here, I first focus on the anatomy of the limbic system and then on the anatomy of the hypothalamus, a small area that belongs to the diencephalon and that comprises less than 1% of the total brain volume. I then consider how the hypothalamus and other closely linked structures in the limbic system receive information from the internal environment and how they act directly to keep it constant by regulating endocrine secretion and the autonomic nervous system. Finally, other parts of the brain that may indirectly affect the internal environment by acting on the external environment through emotions and drives are described.

II. ANATOMY OF THE LIMBIC SYSTEM AND THE HYPOTHALAMUS

In 1953, J. Olds and P. Milner reported that the weak electrical stimulation of specific sites, most located in the hypothalamus or in its rostral continuation, the septum, could elicit in experimental animals an internal state of pleasure or in any case what psychologists describe as a state of reward. For the first time, this work provided a basis to the claim that the hypothalamus, and more generally a continuum of brain tissue in which the hypothalamus is central, is implicated not only in endocrine and visceral functions but also in affect and motivation. I first consider the structures and connections of the brain tissue related to the hypothalamus that belong to this neuronal continuum before examining the anatomy and numerous functions of the hypothalamus.

A. The Limbic System and Its Connections with the Hypothalamus

The neural continuum in which the hypothalamus is central is composed of a part of the brain stem and the limbic system. Part of the brain stem, the mesencephalic reticular formation, which receives inputs from spinoreticular fibers, possesses axons that ascend to the hypothalamus. There are also connections formed by axons directed upward to the hypothalamus from the nucleus of the solitary tract, a cell group in the medulla oblongata. These connections are quite revealing. The nucleus of the solitary tract is the only known case of a circumscribed secondary sensory cell group whose primary sensory input is from the visceral domain.

A second part of the continuum in which the hypothalamus is central lies rostral to it. It is largely interconnected with the phylogenetically primitive cortical tissue that surrounds the upper brain stem. A little more than a century ago, P. Broca observed an almost annular ring of tissue on the medial face of the cerebral hemisphere that represents the free edge of the cerebral cortex. This part of the brain, called *le grand lobe limbique* (the great limbic lobe) by Broca, has led to the concept of the limbic system. This "lobe" surrounds the diencephalon and the cerebral peduncles. He called it *limbique* from the Latin *limbus* because he conceived it as a threshold to the newer pallium. It is sometimes called the rhinencephalon to indicate that these regions of the brain derived during the course of evolution from structures previously associated with the sense of smell. However, the fact that the rhinencephalon is highly developed in animals such as the dolphin and man, in whom the sense of smell is nonexistent or limited, shows that it participates in other activities.

We shall retain only its key-ring structure opening upwards toward the neocortex and downwards toward the brain stem. Like the limbo of Christian mythology, the limbic system is the intermediary between the neomammalian brain heaven (represented by the neocortex) and the reptilian brain hell (including the reticular formation and the striate cortex). The limbic lobe includes the parahippocampal, the cingulate, and the subcallosal gyri. It also includes the underlying cortex of the hippocampal formation, which is composed of the hippocampus, the dentate gyrus, and the subiculum.

In the 1930s, it became evident to J. W. Papez that the limbic lobe formed a neural circuit that provides the anatomical substratum for emotions. He proposed that the hypothalamus is connected with higher cortical centers since it plays a crucial role in the expression of emotion. According to this idea, the neuronal circuit originally proposed by Papez consists of the cortex, which influences the hypothalamus through connections of the cingulate gyrus to the hippocampal formation. Information is then processed by the hippocampal formation and projected to the mammillary bodies of the hypothalamus by way of the fornix. The hypothalamus in turn provides information to the cingulate gyrus through a pathway from the mammillary bodies to the anterior thalamic nuclei and from the anterior thalamic nuclei to the cingulate gyrus (Fig. 1). P. MacLean's resynthesis of Papez's theory of emotions resurrected Broca's concepts and breathed new life into the concept of the all-pervasive limbic system. He included in the limbic system other structures anatomically and functionally related to those described by Papez, including parts of the hypothalamus, the septal area, the nucleus accumbens, neocortical areas such as the orbitofrontal cortex, and the amygdala (Fig. 1).

It is also noteworthy that all of the senses represented in the neocortex—vision, hearing, and the somatic sense—direct part of their information toward either one or both of two cortical districts: the frontal association cortex and the inferior temporal association cortex. The two are interconnected by a massive fiber bundle called the uncinate fasciculus. In turn, the inferior temporal cortex projects to the entorhinal area. The entorhinal area could be

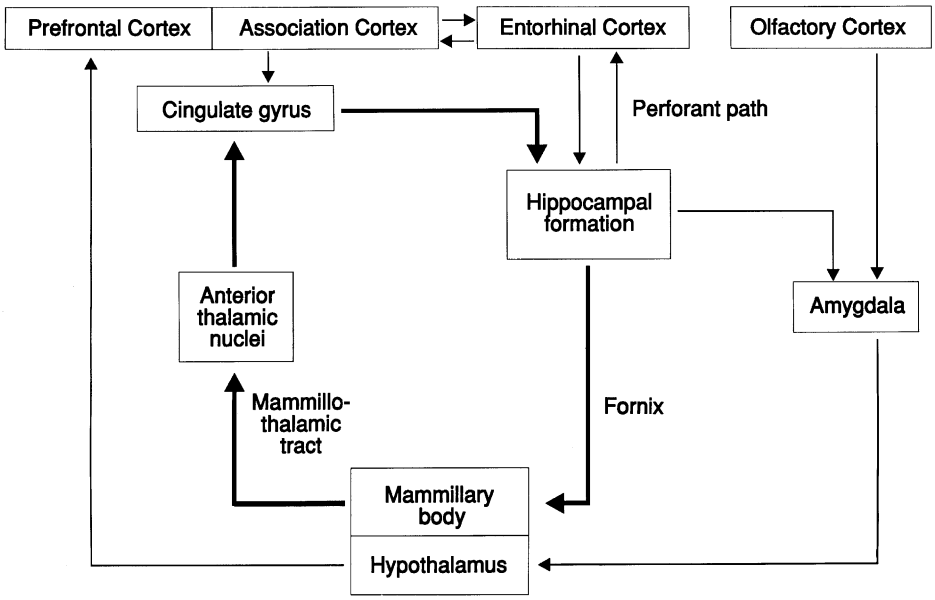


Figure 1 The neural pathways for emotion. The first circuit proposed by Papez is indicated by thick lines, whereas thin lines illustrate recently described connections.

considered as a cortical gateway for projections to the amygdala. In fact, in primates it gives the amygdala its single most important input. The projection is reciprocated; indeed, the amygdala directs its cortical projections to the inferior temporal cortex and to the frontal cortex (specifically the orbital surface of the frontal cortex). Therefore, the amygdala projects to the parts of the neocortex in which the final stages of the cascade of sensory information occur. Evidently, the amygdala also screens its neocortical input. Therefore, it has been tempting for many scientists to speculate that such a brain region could intervene in ideation and cognition. Ordinarily, one thinks of brain function as working inwards, i.e., sensory information being directed from sensory receptor organs over a sequence of synapses to the sensory cortex and from there (in what Papez called "the stream of thought") toward the limbic system. Here, we encounter the opposite: a set of connections directed outward. It is indeed as if the amygdala were participating in the brain's appreciation of the world.

The interoceptive and exteroceptive data reaching the neural continuum in which the hypothalamus is central are clearly distinguishable. The former consist of visceral sensory signals from the spinal cord and the brain stem. These data are unconditional stimuli pertinent to the maintenance of life. On the other hand, what enters the limbic system from the neocortex is fundamentally different. One might call it a repeatedly preprocessed, multisensory representation of the organism's environment. In this situation, the perception of the world is only biased by physiological needs.

It is also remarkable that among all the senses, olfaction possesses a particular link with the limbic system that was taken to be the "nose–brain". Today, it is clearly established that the primary olfactory cortex projects to the entorhinal area, which in turn contact the hippocampus. Thus, after years of fervent affirmation followed by years of fervent denial, the idea that the hippocampus receives olfactory signals was reintroduced. Indeed, the pathway that links olfaction with the limbic system is privileged. Hence, the path from the olfactory epithelium is more direct than the path from sensory surfaces such as the skin. Moreover, the primary olfactory cortex projects to the amygdala, in large part onto a particular cell group (the lateral nucleus of the amygdala), by bypassing the neocortex (Fig. 1). However, although it is clear that the main olfactory bulb (the first central relay for olfaction) projects to the amygdala in rodents, one wonders whether this connection is still present in humans. Indeed, the existence of a specialized area of the nasal mucosa called the vomeronasal organ, which sends information to a compartment of the accessory olfactory bulb, has been demonstrated in animals such as rats. The vomeronasal organ and the corresponding region of the accessory olfactory bulb are thought to

form an apparatus dedicated to the processing of sexually significant odors, but in the fully formed human body none of these structures have been identified. Finally, to emphasize the privileged link between olfaction and the limbic system, it has to be mentioned that the primary olfactory cortex also projects to the hypothalamus.

B. Structure of the Hypothalamus

The hypothalamus in the mammalian brain encompasses the most ventral part of the diencephalon, where it forms the floor and, in parts, the walls of the third ventricle. Its upper boundary is marked by a sulcus in the ventricular wall, the ventral diencephalic or hypothalamic sulcus, which separates the hypothalamus from the dorsally located thalamus (Fig. 2). Caudally, the hypothalamus merges without any clear limits with the periventricular gray and the tegmentum of the mesencephalon. However, it is customary to define the caudal boundary of the hypothalamus as represented by a plane extending from the caudal limit of the mammilary nuclei ventrally and from the posterior commissure dorsally. Rostrally, the hypothalamus is continuous with the preoptic area, which lies partly forward to and above the optic chiasm.

By means of the previously mentioned external landmarks at the ventral surface of the brain, the hypothalamus can be subdivided in the anterior–posterior direction into an anterior part that includes the preoptic area, a middle part, and a posterior part. Another subdivision in the lateral–medial direction consists of three longitudinal zones recognized as the periventricular, the medial, and the lateral zones. The

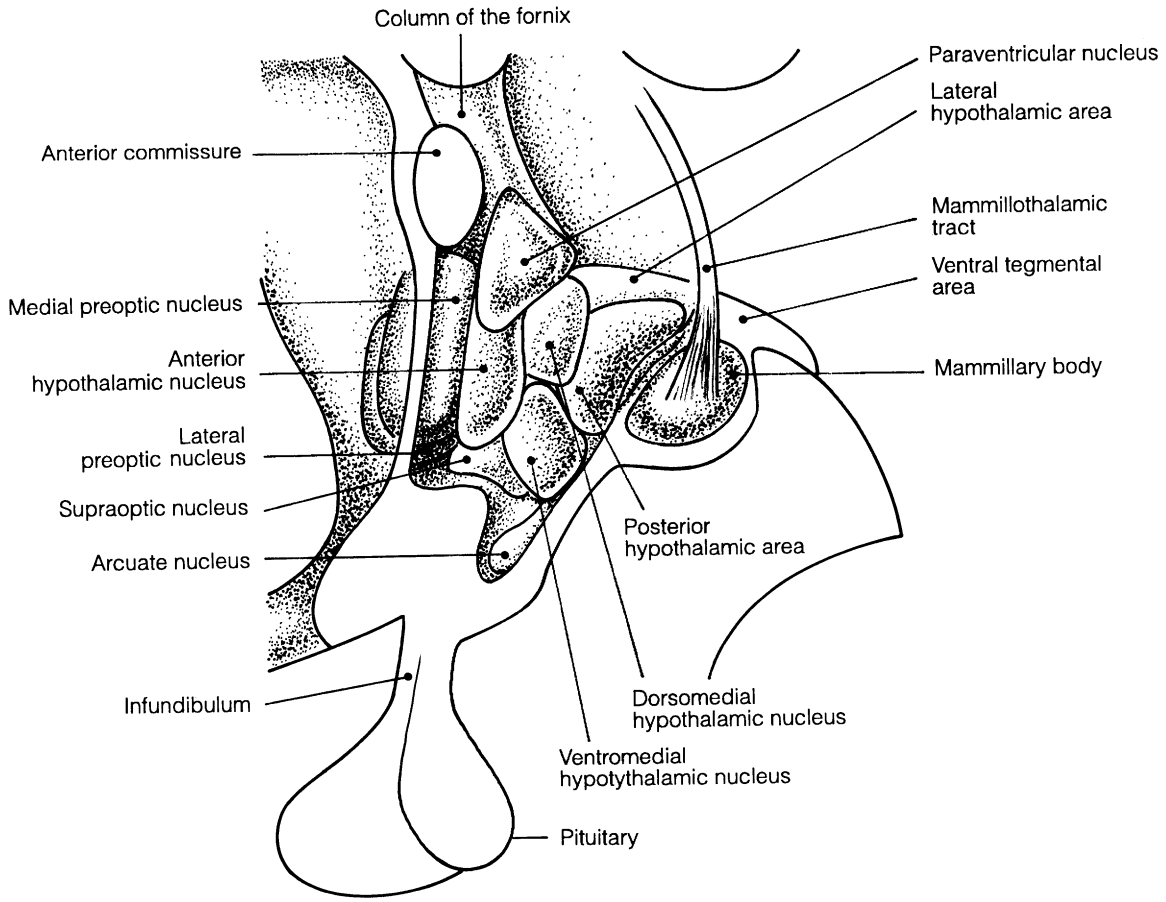


Figure 2 The location of the main hypothalamic nuclei shown in a medial view. The hypothalamus contains a large number of neuronal circuits that regulate vital functions, such as body temperature, heart rate, blood pressure, blood osmolarity, water and food intake, emotional behavior, and reproduction.

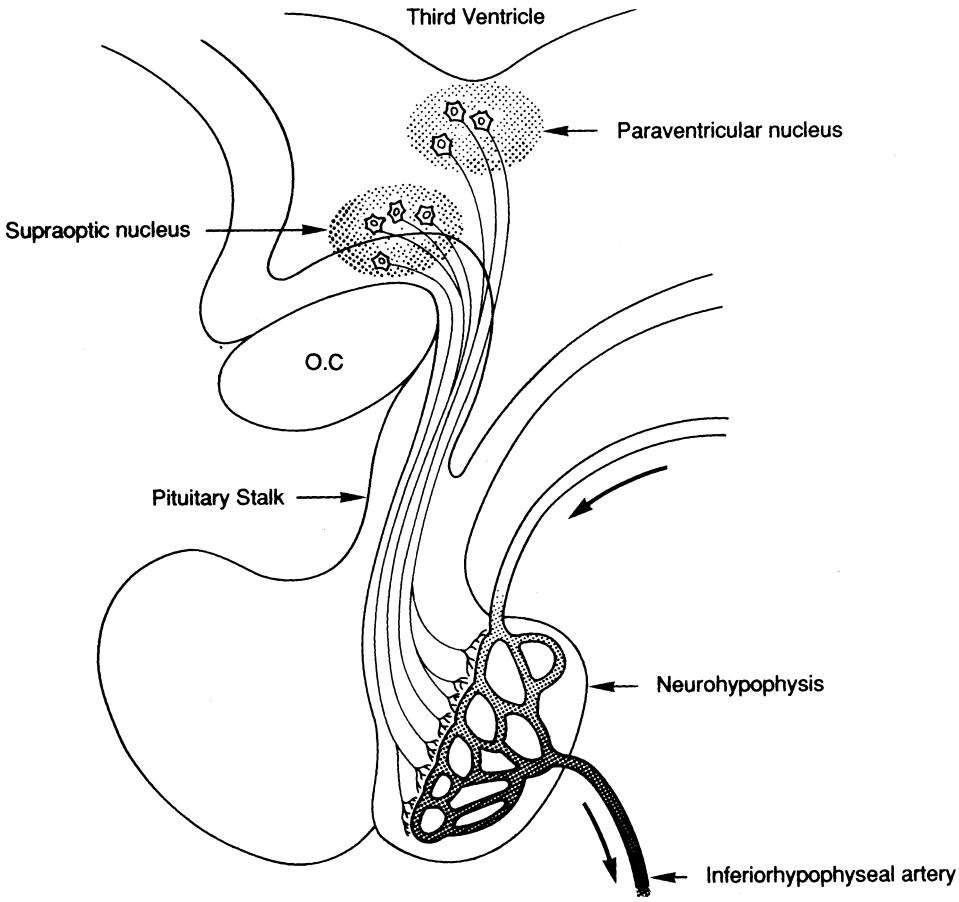


Figure 3 The posterior lobe of the pituitary gland. In the posterior lobe, axons from hypothalamic cell groups called supraoptic and paraventricular nucleus release vasopressin and oxytocin into the systemic circulation (inferior hypophyseal artery). OC, optic chiasm.

periventricular zone consists mostly of small cells that, in general, are oriented along fibers parallel with the wall of the third ventricle. The medial zone is cell rich, containing most of the well-delineated nuclei of the hypothalamus that include the preoptic and suprachiasmatic nuclei in the anterior region; the dorsomedial, ventromedial, and paraventricular nuclei in the middle region; and the posterior nucleus and mammillary bodies in the posterior region (Fig. 2). The lateral zone contains only a small number of cells interposed between the longitudinal fiber system of the medial forebrain bundle. This region possesses long fibers that project to the spinal cord and cortex as well as extensive short-fiber, multisynaptic ascending and descending pathways. The basal portion of the medial region and the periventricular region contain many of the small hypothalamic neurons that secrete the substances that control the release of anterior pituitary hormones. Most fiber systems of the hypothalamus are bidirectional. Projections to and from areas caudal to the hypothalamus are carried in the medial forebrain bundle, the mammillo-tegmental tract, and the dorsal longitudinal fasciculus. Rostral structures are interconnected with the hypothalamus by means of the mammillo-thalamic tracts, fornix, and stria terminalis. However, there are two important exceptions to the rule that fibers are bidirectional in the hypothalamus. First, the hypothalamo-hypophyseal tract contains only descending axons of paraventricular and supraoptic neurons, which terminate primarily in the posterior pituitary (Fig. 3). Second, the hypothalamus receives one-way afferent connections directly from the retina. These fibers terminate in the suprachiasmatic nucleus, which is involved in generating light–dark cycles. The role of these rhythms in the control of motivated behaviors is discussed later.

The following section describes the interrelated functions of the hypothalamus and the pituitary gland as well as some of the major functions of the limbic system.

III. PARTICIPATION OF THE HYPOTHALAMUS AND THE LIMBIC SYSTEM IN HOMEOSTASIS

The internal environment of the body, a term embracing tissue fluids and organ functions such as blood pressure, heart rate, and respiration rate, is under the control of three independent processes. The autonomic nervous system plays an important role in homeostasis. Hence, in the brain, neurons that affect the activity of the preganglionic motor neurons of the sympathetic and the parasympathetic nervous systems are concentrated in the hypothalamus. The evidence is clear: When the hypothalamus of almost any animal, including man, is suddenly destroyed, the animal dies as a consequence of severe disruption of what Bernard called the internal milieu of the body. However, controling the autonomic nervous system is not the only means by which the hypothalamus maintains homeostasis. In addition, the hypothalamus governs the neuroendocrine system through its control of both the anterior and posterior parts of the pituitary gland, which in turn play a major role in the constancy of the internal milieu. Finally, the internal environment of the body is also regulated by motivational states; therefore, we shall also consider some of the hypothalamic funtions involved in a repertoire of voluntary behavioral responses.

In analyzing the nerve mechanisms involved in homeostasis, two centers, one an inhibitor and the other an excitator, are always involved in controlling the same function. This dualistic conception is in the traditon of C. Bernard and W. Sherrington. The former taught us that the centers come into play each time the internal milieu records an abnormal value, and the latter noted that these centers are linked by the principle of reciprocal innervation. The nervous control of eating behavior illustrates this mechanism perfectly when the hunger center is inhibited by the satiety center. An entire theory of paired centers could be built on the hypothalamus. Besides the centers of hunger and satiety, these are centers of pleasure and aversion, approach and retreat, a parasympathetic region opposed to the orthosympathetic region. This anatomical and functional Manicheism goes beyond the hypothalamus to other structures, especially the limbic system and the amygdala, in which facilitating and inhibiting areas confront each other for each of the passions.

A. The Autonomic Nervous System

The autonomic nervous system is primarily an effector system that innervates smooth musculature, heart muscle, and exocrine glands. It is a visceral and largely involuntary motor system. Anatomical principles underlying the organization of both somatic motor and autonomic nervous systems are similar (Fig. 4) and the two systems function in parallel to adjust the body to environmental changes. Nevertheless, the two systems differ in several ways. Within the autonomic nervous system, two subsystems, the sympathetic and the parasympathetic, have long been distinguished by means of anatomical, chemical, and functional criteria. The sympathetic is the most extensive of the two systems. Its preganglionic motor neurons are located in the spinal cord, where they occupy a region called the latteral horn of the spinal gray matter. The preganglionic fibers employ the neurotransmitter acetylcholine, whereas the postganglionic fibers often must travel a substantial distance and employ the neurotransmitter norepinephrine. The sympathetic division of the autonomic nervous system promotes the organism's ability to expend energy (Hess called it ergotropic) and governs an endocrine gland, the adrenal medulla, considered to be a modified autonomic ganglion. It is indeed a universal mobilizing mechanism, valuable in emergencies, with postganglionic ramifications throughout the visceral realm.

In contrast, Hess described the parasympathetic nervous system as trophotropic to signify that it promotes the restitution of the organism. The parasympathetic nervous system in fact antagonizes the sympathetic's effects. The preganglionic motor neurons of the parasympathetic nervous system are in the brain stem and in a short stretch of the spinal cord near its caudal tip. Like the preganglionic motor neurons of the sympathetic nervous system, they employ acetylcholine as their transmitter, but the postganglionic transmitter of the parasympathetic is also acetylcholine. Their axons are long because the ganglia to which they project lie near the tissues of the viscera and sometimes even inside them (Fig. 4). The resilence of autonomic control is in good accordance with what is known about the conduction lines descending from the hypothalamus. Indeed, the hypothalamus emits axons

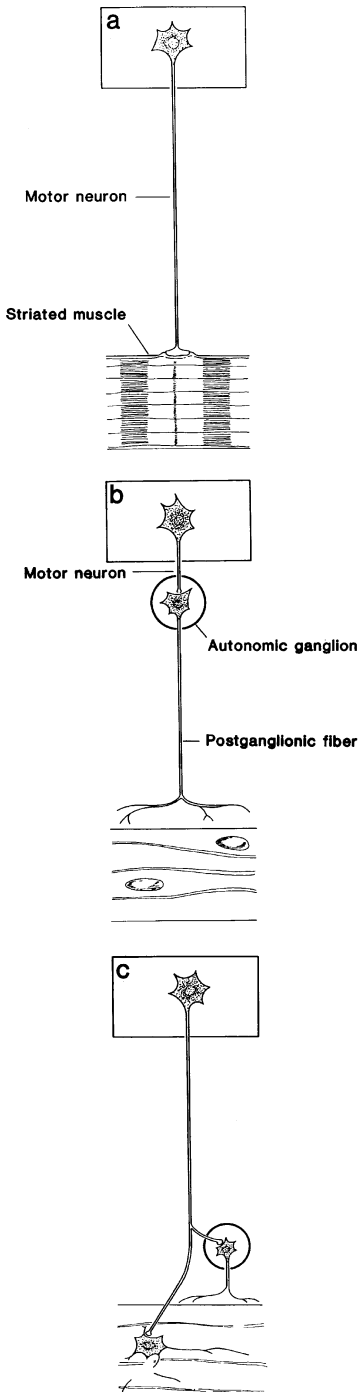


Figure 4 Three different motor innervations. (a) In the somatic motor pattern, a motor neuron from the spinal cord or in the brain stem animates striated muscles directly. In the visceral motor pattern, a two-neuron chain is required. (b) The sympathetic nervous system stations a "preganglionic" visceral motor neuron in the spinal cord. (c) The parasympathetic nervous system employs a two-neuron pathway. The first neuron is situated in the brain stem or toward the bottom of the spinal cord, and the ganglion is close or even inside the viscera.

that descend toward both sympathetic and parasympathetic preganglionic visceral motor neurons, thus regulating the viscera. Hence, the hypothalamus may function as the so-called head ganglion of the autonomic nervous system that mediates conventional reflexes involving neural inputs and outputs. Fibers passing directly from the hypothalamus to the lateral horn of the spinal cord's gray matter, where the preganglionic motor neurons of the sympathetic nervous system are situated, have recently been found. However, these fibers seem to constitute a small minority of hypothalamic efferents; the hypothalamus has nothing similar to a pyramidal tract to carry its descending outputs. Instead, it appears in large measure to project no further than the midbrain, where neurons of the reticular formation take over. It is noteworthy that pathways descending to autonomic motor neurons are interrupted at numerous levels, at which further instructions can enter the descending lines.

Most of the regions of the brain that influence the autonomic nervous system's output (e.g., the cerebral cortex, the hippocampus, the entorhinal cortex, parts of the thalamus, basal ganglia, cerebellum, and the reticular formation) produce their actions by way of the hypothalamus, which integrates the information it receives from these structures into a coherent pattern of autonomic responses. The hypothalamus controls the output of the autonomic nervous system in two ways. The first one is direct and consists of projections to nuclei in the brain stem and the spinal cord that act on preganglionic autonomic neurons to control respiration, heart rate, temperature and blood pressure. Thus, stimulation of the lateral hypothalamus leads to general sympathetic activation (increase in blood presure, piloerection, etc.). Second, the hypothalamus that governs the autonomic nervous system by controlling the endocrine system, which releases hormones that influence autonomic functions is discussed in the following section.

B. The Neuroendocrine System

The Scharrers (1940) first hypothesized that the peptides of the neurohypophysis were in fact synthetized by specialized hypothalamic neurons and transported within their axons to the neural lobe to be released into peripheral blood. In the late 1940s, Harris and collaborators proposed the hypophysial portal chemotransmitter hypothesis of anterior pituitary

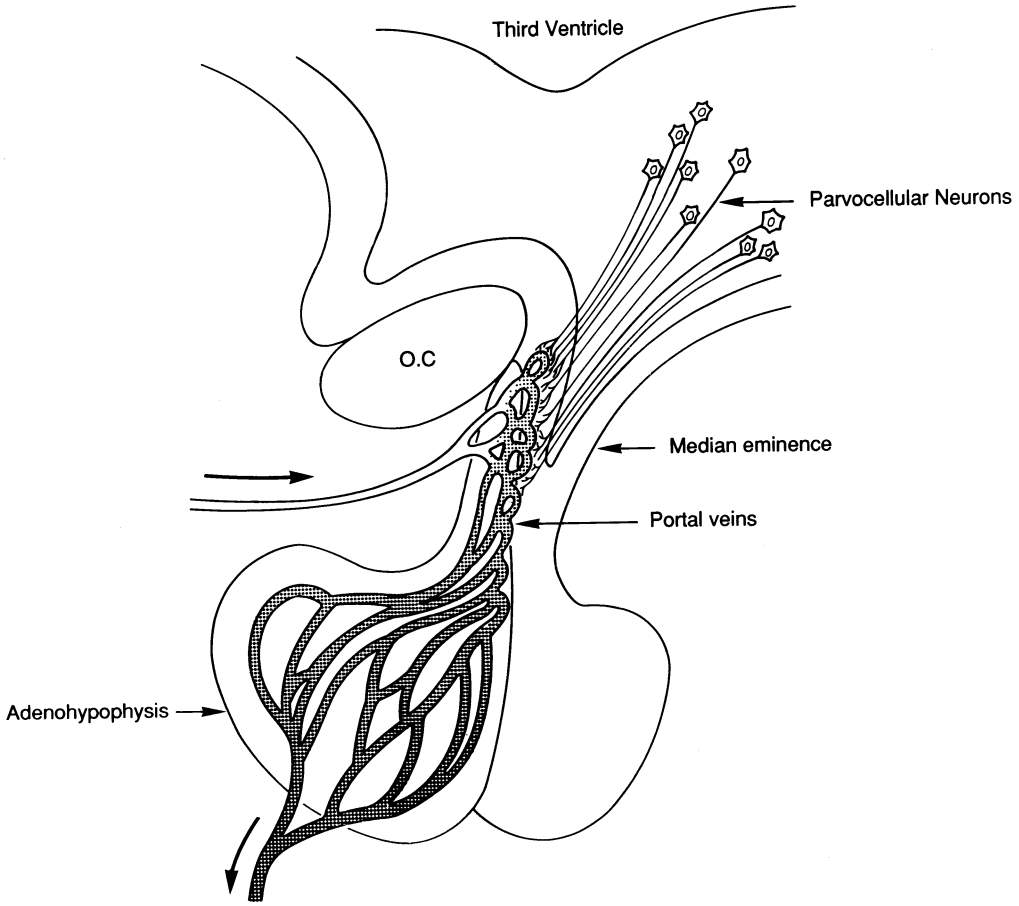


Figure 5 The anterior lobe of the pituitary gland. The anterior lobe synthetizes several hormones. Their release is induced by chemical signals, called releasing factors, that are secreted by hypothalamic neurons. These factors enter the hypothalamo-pituitary portal system, comprising first a capillary bed in the hypothalamus and then a venous drainage channel and, in the anterior lobe, a second capillary bed. OC, optic chiasm.

control, which stated that the factors regulating the anterior lobe are formed by hypothalamic neurons (later termed hypophysiotropic neurons) and transported to be released into the hypophysial portal circulation and carried to the anterior pituitary (Fig. 5), where they control the synthesis and release of anterior pituitary hormones into the general circulation. Both of these hypotheses have been confirmed and rationalized into a unified theory of neurosecretion in which the nervous system controls endocrine function. Neurosecretion is the phenomenon of synthesis and release of specific substances by neurons. Some neurosecretions are exported into the peripheral or hypophysial blood and act as true hormones; others, released in close apposition to other neurons, act as neurotransmitters or neuromodulators. Translation of neuronal signals into chemical ones has been termed neuroendocrine transduction and the cells have been called neuroendocrine transductors by R. J. Wurtman and F. Anton-Tay (1969). Two types of neurotransducer cells regulate visceral function: (i) neurosecretomotor cells, in which the neurosecretion acts directly through synapses on gland cells, and (ii) neuroendocrine cells, in which the neurosecretion passes into the blood and acts on distant targets.

Neurosecretory cells possess in common with other neurons the usual aspects of neuron functions. Most of the insight into the physiology of neurosecretory systems has been gained from studies of the hypothalamo-hypophyseal system. This system brings nervous and endocrine cells together in one anatomical entity in which the nervous system and the glandular cells of the

anterior hypophysis communicate. These two structures share common properties. They both secrete peptidergic hormones (releasing and inhibiting hypothalamic factors and the hypophyseal stimulins), and both exhibit electrical properties such as excitability, with production of action potentials. Thus, electrophysiological techniques, which were previously reserved for studies of nerve and muscle cells, can be applied to the hypothalamo-hypophyseal system in both its nervous and endocrine structures. The electrophysiological properties of these cells reveal the existence of (i) stimulus-secretion coupling, particularly at the level of neurosecretory terminals in the posterior hypophysis, the median eminance, and the endocrine cells of the anterior hypophysis, and (ii) modifications in membrane electrical properties exerted by the binding of different regulatory factors to their receptors. These observations are used to explain the modulatory mechanism of membrane properties brought into play by each factor in order to enhance or inhibit hormonal release. Thus, the electrical properties play a central role in the regulation of endocrine secretion in the anterior hypophysis. These electrical properties, common to nervous and endocrine cells, are linked to changes in membrane permeability to different ions (i.e., Ca^{2+}, Na^{+}, K^{+}, and Cl^{-}).

The pituitary gland is divided into two main functional units—the neural (or posterior) lobe and the adenohypophysis (or anterior) lobe (Figs. 3 and 5). In many mammalian species an intermediate lobe (derived embryologically from the same tissue as the anterior lobe) is present, but in humans these lobe cells are dispersed throughout the entire pituitary gland.

In the neurohypophyseal system, hypothalamic neurons transmit action potentials along their axons in a similar manner to that used by unmyelinated neurons, and each action potential triggers the release of secretory granules from nerve endings by calcium-dependent exocytosis into the general circulation. The neural lobe is an anatomical part of the neurohypophysis that is commonly viewed as consisting of three portions—the neural lobe (infundibular process or posterior pituitary), the stalk, and the infundibulum. This latter portion forms the base of the third ventricle (Figs. 3 and 5). In fact, there is a fourth intrahypothalamic component of the neurohypophysial system that consists of the cells of origin of the two principal nerve tracts that terminate in the neural lobe—supraopticohypophysial and paraventriculohypophysial.

The neurohypophysial hormones secreted by the magnocellular neurons are vasopressin (antidiuretic hormone) and oxytocin, which are synthesized within the cell bodies in association with specific proteins, the neurophysins. Like most peptidergic hormones, vasopressin and oxytocin are cleaved from a larger prohormone. These prohormones are synthesized in the cell bodies of the magnocellular neurons and are cleaved within vesicles during their transport down the axons. Vasopressin, oxytocin, and at least two forms of neurophysins are secreted into the blood circulation; they are responsive to appropriate physiological stimuli and can be altered by stressful conditions.

Although the anterior lobe does not receive any direct nerve supply, its secretions are under control exerted by the hypothalamus. This control is mediated by chemical factors (hypophysiotropic hormones) secreted by the parvocellular neuroendocrine neurons located in several hypothalamic regions: the medial basal and periventricular regions and the arcuate, tuberal, preoptic, and paraventricular nuclei. Parvocellular neurons secrete peptides in the interstitial space of the base of the third ventricle and then diffuse into the capillary plexus of the median eminence that is interposed between the peripheral arterial system and the pituitary sinusoidal circulation (Fig. 3). By this anatomical arrangement, neurohormonal mediators synthesized and released by the hypothalamus are brought into direct contact with the adenohypophysis. The hormones released by the anterior lobe of the pituitary gland are called tropic (literally "switch-on") hormones. Each is the second and final messenger in a sequence of chemical signals leading from the brain to a particular endocrine gland. All the tropic hormones of the anterior lobe are simultaneously trophic hormones, in whose absence their target glands atrophy.

A different functional link leads from the neurons of the hypothalamus to the posterior lobe of the pituitary complex. This link is more direct since it does not include part of the circulatory system. It begins in two circumscribed magnocellular nucleus. They are the first hypothalamic nuclei whose function has been identified with some precision. All, or nearly all, of the axons originating in the supraoptic nucleus, along with approximately 30% of the axons originating in the paraventricular nucleus, pass through the pituitary stalk and reach the posterior lobe of the pituitary. The remaining 70% have several destinations, of which one is especially notable: The paraventricular nucleus is a substantial contributor to the pathway descending from the hypothalamus direct to the lateral horn of the spinal cord, which contains the spinal cord's preganglionic sympathetic motor neurons.

Unlike the anterior one, the posterior lobe is a part of the brain. Nevertheless, it contains no neurons; the

terminations of the supraoptic and paraventricular axons make no synaptic contacts. Instead, they lie embedded in a tissue composed of modifed glial cells called the pituicytes and a dense plexus of capillaries. The glandular products of the supraoptic and paraventricular nuclei are synthesized in the cell body and packaged in neurosecretory vesicles in which some hormonal maturation may occur. These neurosecretory vesicles are transported down the axon to the neural terminal, where hormones are stored and released by secretion when the neuron is stimulated.

Recently, it has been demonstrated that a type of ependymal glial cell called tanycyte, which ensheaths the terminals of hypothalamic neurons, regulates the release of luteinizing hormone-releasing hormone (LHRH) from the hypothalamus and may therefore play a key role in the onset of puberty. LHRH axons that travel with the processes of tanycytes can be covered by slips of glioplasm. At the perivascular space level, the nerve terminals may be partially covered or exposed, potentially impeding or enhancing the secretion of LHRH into capillaries. Such observation, realized at a cellular level, illustrates the tremendous plasticity of the hypothalamus. Interestingly, it was found that these glial cells in the median eminence possess estrogen and epidermal growth factor receptors, whereas LHRH neurons apparently do not. Taken together, these observations provide strong evidence that, at puberty, glia is a crucial target for estrogenic action that may induce morphological changes accompanied by release of chemical signals that modulate hypothalamic neurons.

C. Behavioral Responses

When Hess succeeded in implanting electrodes in the brain and permanently fixing them to the skull of animals, he found that stimulation of different parts of the hypothalamus produced a array of behavioral responses. For example, electrical stimulation of the lateral hypothalamus in cats elicited autonomic and somatic responses characteristic of anger: increased blood pressure, raising of body hair, pupillary constriction, raising of the tail, and other characteristic emotional behaviors. Thus, the hypothalamus is not only a motor nucleus for the autonomic nervous system, as well as a neural part controlling the neuroendocrine system, but also a coordinating center that integrates various inputs to ensure a well-organized, coherent, and appropriate set of autonomic and somatic responses. In line with this view, vasopressin, oxytocin, and other regulating hormones are not the only peptides of neurobiological interest that can be found in the hypothalamus. The opioid peptides, β-endorphin, and the enkephalins can also be detected in this structure, as can angiotensin II, substance P, neurotensin, cholecystokinin, and a host of other peptides known to be involved in multiple behavioral responses. Interestingly, almost every type of peptidergic neuron previously studied, including both parvocellular and magnocellular hypothalamic neurons, has been found to contain more than one type of peptide that could act synergistically. Furthermore, peptides released by the hypothalamic magnocellular and parvocellular neurons are not unique to these cells; they have also been found in other regions of the nervous system. Such peptidergic projections are well suited for coordinating neuroendocrine and autonomic responses. For example, regulatory peptides released at brain sites other than the median eminence may modulate behavior by actions independent of the release of pituitary hormones. The behavioral effects of regulatory peptides are thematically related to the type of endocrine effects produced by the same peptide acting on the pituitary. Corticotropin-releasing hormone (CRH) is an example of such a regulatory peptide. On the one hand, it acts on the pituitary to stimulate the release of adrenocorticotropic hormone in response to stress. On the other hand, when injected intracerebroventricularly, CRH evokes many of the behavioral and autonomic reactions normally seen in response to stress.

D. Limbic Functions

In 1937, Papez suggested that the limbic lobe formed a neuronal circuit that provided the anatomical substratum for emotions. Based on experimental results suggesting that the hypothalamus plays a critical role in the expression of emotions, Papez argued that since emotions reach consciousness and thought and, conversely, higher cognitive functions affect emotions, the hypothalamus must communicate reciprocally with higher cortical centers.

The representation of the outside world and the internal milieu are superimposed in the limbic system. All the sensory information about the perceiver's environment is inscribed in the neuronal network of the limbic cortex, the hippocampus, and the amygdala. The vegetative, nervous, and humoral functions that

contribute to homeostasis are represented simultaneously in the limbic system. Moreover, the hippocampus has been described as a gatekeeper embodying the brain's ability to commit things to lasting memory. Evidence for such a role of the hippocampus is clear. For instance, the neurosurgical removal of the hippocampus on both sides of the human brain, as a treatment of otherwise intractable forms of epilepsy, leads to a central disorder called hippocampal amnesia. The patient retains the memories he or she collected well before the surgery but cannot collect new ones.

Other evidence suggests that the amygdala is not only essential for olfactory discrimination but also commands a number of adaptative responses. Lesions and electrical stimulations of the amygdala produce a variety of effects on autonomic responses, emotional behavior, and feeding. Consequently, the amygdala has been implicated in the process of learning, particularly learning those tasks that require coordination of information from different sensory modalities or the association of a stimulus and an affective response.

Finally, it has been extensively described that the interplay between the neural activity of the hypothalamus and the neural activity of higher centers results in emotional experiences that we describe as fear, anger, pleasure, or satisfaction. For example, the behavior of patients from whom a part of the limbic system (frequently the prefrontal cortex) has been removed supports this idea. Indeed, these patients are no longer bothered by chronic pain or, alternatively, when they do perceive pain and exhibit appropriate autonomic reactions the perception is no longer associated with a powerful emotional experience.

In summary, neurons from the limbic system form complex circuits that collectively play an important role in numerous behavioral responses, such as learning, memory, and emotions. Such a role played by motivational states in homeostasis is discussed in the next section.

IV. ROLE OF MOTIVATIONAL STATES IN HOMEOSTASIS

The role of the hypothalamus and the limbic system in the neuroendocrine and autonomic regulation of homeostasis was previously described. Here, I discuss the control of homeostasis by motivational states, the internal conditions that arouse and direct elementary behavior. Motivational states (also called drives) are inferred mechanisms to explain the intensity and direction of a variety of voluntary behaviors, such as temperature regulation, feeding, consumption of water, and sexual behaviors. It is the internal state that creates drives by deviations from the norm that defines the conditions of equilibrium for the milieu. A drive is not the stimulus that triggers the behavioral response but rather the internal force that underlies it. However, a stimulus may cause a drive when it has been associated in the past with a particular internal state. Because of the flexibility of internal parameters that define a pseudoequilibrium for the internal milieu of a living organism, it would be more appropriate to refer to the internal state as a fluctuating central state.

Specific motivational states possess two components: *needs* and *rewards* participating in homeostatic drives. Drives represent urges or impulses based bodily needs that lead animals into action. This concept is central to Freudian psychology, which is related to needs and experiences of satisfaction. Needs are experienced as an intolerable internal situation that must be stopped. This internal state, called motivation by psychologists, induces a drive to accomplish the act that will relieve it. For example, a temperature-regulating drive is said to control behaviors that directly affect body temperature, such as rubbing one's hands together. Therefore, physiological deprivation may lead to the satisfaction of a need and this, in turn, will lead to response reinforcement. This means that this action will become more likely in similar future situations. Such a psychological process is considered by some behaviorists to be the basis of learning.

So far, I have dealt with the role of tissue needs in generating appropriate behaviors and physiological responses to fight against a bodily deficit. However, another component linked with motivated behaviors is reward, which may lead to a profit. For example, sexual responses do not appear to be controlled by the lack of specific subtances in the body but are rather oriented toward hedonic factors. In this case, drives, by producing goal-oriented behaviors, are defined by the goal to be reached and justified by the reward obtained. One form of reward is pleasure; therefore, the duality of profit and pleasure is the major component of drive. On the other hand, drive–reward constitutes one of the rules of learning.

Finally, in describing the factors that regulate motivated behaviors, I must also be noted out the role of ecological constraints and anticipatory mechanisms. The characteristics of most behavioral responses are determined by evolutionary selection,

which retains only appropriate responses in a defined ecological surounding. In this context, one of the most determinant parameters that participate in keeping a specific behavioral response is the cost-benefit ratio. The other component that also controls motivated behaviors, namely anticipator mechanisms, gives to the homeostasis concept its temporal dimension. Sometimes, lack rather than need is able to activate drives. In the case of sexual arousal, a feeling of lack becomes a simulation of need. Accordingly, homeostatic regulation is often anticipatory and can be initiated before any physiological deficit occurs. Such a role is played by clock-like mechanisms that turn physiological behavioral responses on and off before the occurrence of any tissue deficits. The master clock mechanism that drives and coordinates many rhythms is located in the suprachiasmatic nucleus of the hypothalamus. One such common cycle is a daily rhythm called the circadian rhythm, which controls feeding, drinking, locomotor activity, and several other responses. After experiencing jet-lag due to long distances, travelers can confirm the important role played by circadian rhythms.

V. CONCLUSIONS

Homeostatic processes can be analyzed in terms of control systems or servomechanisms, comprising a set point, an error signal, controlling elements, a controlled system, and feedback detectors. This approach has provided a convenient and precise language to describe both concepts and experimental results. Moreover, it has been successfully applied to temperature regulation, feeding, and drinking. For example, in the temperature regulation system, the integrator and many controlling elements appear to be located in the hypothalamus. The normal body temperature is the set point and the feedback detector collects information about body temperature from two main sources—peripheral and central temperatures. The analysis of feeding and thirst behaviors can also be approached in terms of a control system, as for temperature regulation, although at every level of analysis the understanding is less complete than for the control of temperature.

The hypothalamus is concerned with the regulation of various behaviors directed toward homeostatic goals, such as consumption of food and water or sexual gratifications. Through its control of emotions and motivated behavior, the hypothalamus acts indirectly in maintaining homeostasis by motivating animals and human beings to act on their environment.

In regulating emotional expression, the hypothalamus functions in conjunction with higher control systems in the limbic system and neocortex. In addition to regulating specific motivated behaviors, the hypothalamus and the cerebral cortex are involved in arousal, namely the maintenance of a general state of awareness (the level of arousal varies from different degrees of excitement to coma, sleep, and drowsiness). However, because of its intimate relationship with both the autonomic and the endocrine systems, the hypothalamus appears to play a central role in regulating homeostatic behaviors. The hypothalamus contributes to these adaptative behaviors by integrating information from both external and internal stimuli that report on the homeostatic state of the animal.

See Also the Following Articles

AROUSAL • ARTIFICIAL INTELLIGENCE • BEHAVIORAL NEUROIMMUNOLOGY • BIOFEEDBACK • CIRCADIAN RHYTHMS • EMOTION • HYPOTHALAMUS • LIMBIC SYSTEM • NEUROTRANSMITTERS • PSYCHONEUROENDOCRINOLOGY • STRESS: HORMONAL AND NEURAL ASPECTS

Suggested Reading

Cannon, W. B. (1929). *Bodily Changes in Pain, Hunger, Fear and Rage*. Appleton, New York.

Damasio, A. R.(1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. Grosset/Putnam, New York.

Ekman, P., and Davidson, R. J. (1994). The nature of emotion. Oxford Univ. Press, New York.

Papez, J. W. (1937). A proposed mechanism of emotion. *Am. Med. Assoc. Arch. Neurol. Psychiatr.* **38,** 725–743.

Sacks, O. (1987). *The Man Who Mistook His Wife for a Hat and Other Clinical Tales*. HarperCollins, New York.

Wurtman, R. J., and Anton-Tay, F. (1969). The mammalian pineal as a neuroendocrine transducer. *Prog. Horm. Res.* **25,** 493–513.

Humor and Laughter

JYOTSNA VAID
Texas A & M University

GLOSSARY

Duchenne display A characteristic facial display accompanying laughter in humans involving the joint contraction of lip and eye muscles.

gelastic seizure Typically unprovoked, stereotyped, and inappropriate laughter accompanying an epileptic seizure.

The ability to perceive and produce humor, highly valued across human societies, has been the subject of much philosophical interest but little sustained, empirical study. This article summarizes the variety of theoretical perspectives and available research relevant to the study of humor, including its nature, its origins in ontogeny and phylogeny, its behavioral manifestation in laughter, its psychological and evolutionary significance, and its hypothesized neural mediation. Although responding with humor may appear to be a way of evading a serious problem, it might also turn out to be a rather effective way of counteracting the seriousness, or even danger, in a situation by deflecting its adverse effects and by offering an alternative to despair. Humor can provide a way for individuals to conceal or to reveal their vulnerabilities; it can also enable them to transcend a predicament. A person with a good sense of humor is viewed more favorably than one who is humorless. Humor is said to confer mental flexibility, openness to experience, playfulness, and maturity and, as such, is highly valued whether in a coworker, a leader, a friend, or a prospective mate. Indeed, the ability to be playful and humorous appears to be universally valued.

Although scholarly interest in the nature and functions of humor has a long history, empirical inquiry into the forms, uses, and biological bases of humor and laughter is fairly recent. This article addresses both well-researched and relatively unexplored questions pertaining to humor and laughter. The following are some of the questions that are considered here: What is humor? What is the role of laughter in humor? What are the ontogenetic and phylogenetic origins of laughter? How does humor work, cognitively? What is the functional significance of humor? What are the varieties of laughter and how are they perceived? What kinds of disorders of laughter have been documented and what do they suggest about the neural circuitry underlying laughter and humor?

I. DEFINING HUMOR

Humor normally requires at least two individuals (the humor initiator and the receiver) who cooperate in setting up what Gregory Bateson termed a play frame,

in which the parties involved tacitly agree that what is inside the frame is not to be taken seriously. A play frame may also be set up by a single individual in response to a perceived incongruity in an event or situation. Verbal humor is a particular form of skilled language use in which at least two disparate meanings are interwoven into a text by making use of ambiguity, polysemy, intertextuality, or inconsistency in such a way that the listener is led to expect one meaning but actually experiences the other. The pleasure of humor is thought to arise upon the sudden recognition of the mismatch between the expected and the experienced meaning.

II. HUMOR AND LAUGHTER

In his 1872 monograph, *The Expression of the Emotions in Man and Animals*, Charles Darwin noted that "Joy, when intense, leads to various purposeless movements—dancing about, clapping the hands, stamping, etc., and to loud laughter." Laughter is indeed commonly seen as an expression of joy, happiness, or amusement. It typically occurs in informal social situations, usually in the presence of a close friend, sibling, caregiver, or intimate.

Although laughter is normally taken to be an indicator that the laugher is in a happy emotional state, it does not exclusively or necessarily signal such a state. One can clearly find something amusing without it actually making one laugh out loud. Similarly, one can feign laughter even when one is not actually amused. Other emotional states that give rise to laughter include scorn, embarassment, and nervousness. It remains to be determined whether the laughter in these different states is morphologically different. Under certain clinical conditions, laughter can be triggered in adults without their ability to control it. In such cases it occurs for no apparent reason and usually without accompanying positive affect. The occurrence of such dissociations between the motoric act of laughter and associated affective or cognitive states raises the possibility that laughter is under the control of a variety of different neural structures and systems at different levels and that these may be selectively disrupted in pathology.

III. ONTOGENY OF LAUGHTER

Laughter and crying both appear to be innate mechanisms in humans, although laughter's onset occurs later: whereas crying emerges at birth and smiling at approximately 2 or 3 weeks of age, the characteristic expiratory movement of laughter does not appear until approximately 4–6 months. However, cases of so-called gelastic (or laughing) seizures in neonates indicate that neural and physiological structures subserving laughter are in place at birth. Laughter's innateness is further suggested by the fact that it is observed in deaf–blind children, even those who could not have learned about it by touching people's faces. Although laughter initially occurs involuntarily, whether in response to tickling or peek-a-boo games or as a reaction to a sudden change in the sensory environment, such as an unexpected noise, over the course of development laughter becomes more regulated, under voluntary control, and elicited in response to cognitive and social stimuli rather than physical stimuli per se.

IV. PHYLOGENY OF LAUGHTER

Laughter is estimated to be 7 million years old. Like other vocalizations, such as moaning, sighing, and crying, laughing is thought to have preceded speech and, like these vocalizations, it may also have a communicative function.

Although laughter is often claimed to be unique to humans, behaviors similar to human laughter have been observed in other primates and something akin to laughter has even been argued to be present in rodents. In 1997, Signe Preuschoft and Jan van Hooff examined variations in the contexts and social functions of bonding displays in various primate species, including Old World primates, macaques, baboons, great apes, and humans. They noted a silent bared-teeth display (also called a grin, grimace, or smile) associated with inhibited locomotion, evasive and protective body movements and postures, and grooming and embraces. This display was observed in all macaque and baboon species and humans. A relaxed, open-mouth display ("play face"), marked by a widely opened mouth but without pronounced baring of the teeth, was observed in all macaque and baboon species, in each of the great apes, including the chimpanzee, bonobo, orangutan, and gorilla, in humans, and in more distant species such as vervets and squirrel monkeys.

One variant of the play face noted in some species was an open mouth, bared-teeth display with associated staccato breathing and bursts of vocalization. This latter display, which Preuschoft and van Hoof

termed the "laugh face," appeared strikingly similar to human laughter. It was accompanied by boisterous body postures, brusque movements, mock biting, playful chasing, and evasive, repetitive glancing movements rather than a tense gaze. The laugh face was found only in certain macaques, mandrills, geladas, the orangutan, the bonobo, and, of course, humans, and it occurred primarily in social play contexts, although in some cases it was observed to accompany solitary play, but in these cases it occurred only when conspecifics were near. The presence of an audience thus appears to be at least facilitative, if not necessary, for releasing the laugh face display.

In 1987, Preuschoft and van Hooff proposed a functional significance of play that is relevant to the current analysis. They noted that in bonding behaviors (i.e., those involving care, sex, and affiliation), it is important to deemphasize competition and focus on shared interests. Play is one form of bonding behavior. An important element of play is the partner's unexpected performance of an expected behavioral act. Preuschoft and van Hooff suggested that the "essence of play seems to be to actively bring about incongruous and unexpected behaviors and to interpret them as nonthreatening." This, they argued, allows for affective and cognitive mastery of incongruous situations.

Whereas laughter in nonhuman primate species is found principally in play contexts and primarily occurs in the young of the species, laughter in humans occurs in a variety of social contexts throughout the life span. Moreover, there is no evidence that nonhuman primates have anything comparable to the human ability to produce, comprehend, and enjoy humor.

V. PRECONDITIONS FOR THE EMERGENCE OF HUMOR

Various factors have been proposed as preconditions for the emergence of humor perception in a species, including sociality, an exploratory drive, a system of communication (to allow for the members of the species to express and communicate beliefs and feelings), an ability to mentally represent and order the physical world, and an ability to construct possible imaginary worlds. Additional criteria proposed include a capacity to use logic and reasoning independently of affective or real-world constraints, a capacity to deceive or misrepresent the truth, a capacity for collaboration, reciprocity, compassion, and a theory of mind.

VI. WHAT MAKES SOMETHING HUMOROUS?

Speculating about the causes of mirthful laughter in human adults, Darwin noted that "[s]omething incongruous or unaccountable, exciting surprise and some sense of superiority in the laugher, who must be in a happy frame of mind, seems to be the commonest cause." In comparing laughter arising from actual tickling and that arising from the tickling of the imagination, he noted:

> *From the fact that a child can hardly tickle itself, or in a much less degree than when tickled by another person, it seems that the precise point to be touched must not be known; so with the mind, something unexpected—a novel or incongruous idea which breaks through an habitual train of thought—appears to be a strong element in the ludicrous.*

Darwin's insights anticipate many subsequent theoretical accounts of humor. With respect to its cognitive basis, an early formulation of the mechanisms underlying humor may be found in the work of a Gestalt psychologist, Norman Maier. Maier noted that a key element of humor is a sudden and unexpected restructuring of the elements of a configuration, not unlike that experienced during a flash of insight. A humorous narrative manipulates our expectations by leading us down a garden path only to present us with an altogether different conclusion than the one we were led to expect. Inasmuch as the conclusion disrupts the way in which we have been thinking about the events in the narrative, we are totally unprepared for it. After a momentary confusion of thought, we experience the newly restructured configuration with clarity; the amusement arises when we realize how we were misled. In 1992, Wyer and Collins proposed that diminishment is a key element in humor elicitation: For humor to be elicited, the new perception of the situation must in some sense be diminished in importance in comparison to the apparent reality that was first assumed.

Subsequent theorists have elaborated on these accounts. For example, cognitive approaches characterize humor perception in information processing terms as involving at least two and possibly three stages: a setup stage, a stage in which the incongruity is recognized, and a stage of incongruity resolution. Although there is debate regarding whether humor appreciation requires that the incongruity be satisfactorily resolved (with some arguing that incongruity recognition per se is sufficient to experience humor),

theorists generally agree on the importance of the juxtaposition of two or more mental representations for humor to be perceived, following Arthur Koestler, who, in *The Act of Creation*, described humor as an example of bisociative thinking, which juxtaposes and brings together two disparate matrices of thought. Recent theoretical formulations in cognitive science have used concepts such as frame shifting and conceptual integration of mental spaces to formalize the processes underlying humor perception. To date, there has been little cognitively oriented experimental research on humor comprehension and even less on the processes underlying humor generation.

It has also been acknowledged that it is not enough to postulate incongruity as a prerequisite of humor because, in certain situations, incongruity may give rise to other emotional reactions, such as fear or apprehension, rather than humor. A humorous rather than apprehensive reaction is more likely when the receiver does not have too much invested in the content of the humor—that is, when the humor does not threaten the receiver's beliefs or feelings in any profound or disturbing way. Second, a humorous reaction is more likely when the receiver was led to expect something dangerous that turns out not to be. Indeed, Ramachandran suggests that laughter may have evolved as a false alarm signal—that is, as an immediate signal to conspecifics that some potentially threatening event has trivial rather than terrifying implications. This notion is similar to the diminishment view of humor developed by Wyer and Collins.

VII. FUNCTIONAL THEORIES OF LAUGHTER AND HUMOR

Theories about the uses of laughter and humor range from psychological and social/anthropological approaches to metaphysical and evolutionary ones.

A. Psychological

Among psychological theories, that of Sigmund Freud, as described in *Jokes and Their Relation to the Unconscious* (1963), is one of the earliest. For Freud, jokes allow a temporary expression of socially undesirable impulses from the subconscious. Freud notes that tendentious humor makes possible "the satisfaction of an instinct (whether lustful or hostile) in the face of an obstacle that stands in its way." The association of humor with aggressive or sexual impulses has characterized subsequent accounts as well, although there is no current consensus as to whether hostility is inherent in humor.

Proponents of the so-called superiority theory (e.g., Aristotle, Plato, Hobbes, and Bergson) view laughter as reflecting a moral stance on the part of the laugher. For Bergson, laughter asserts the human values of spontaneity and freedom and therefore erupts whenever a person behaves rigidly, like an automaton: "Humor consists in perceiving something mechanical encrusted on something living." Hobbes described the passion of laughter as "nothing else but sudden glory arising from a sudden conception of some eminency in ourselves by comparison with the infirmity of others, or with our own formerly." Superiority is asserted not just by laughing at others' defects but also by showing that one can laugh at (and rise above) one's own imperfections.

In contrast to negative uses of humor, as in put-down or derisive humor, recent clinically based research has directed considerable attention to positive aspects of humor. Humor is increasingly viewed as a useful mechanism for coping with stress and regulating affect. In 1999, Galloway and Cropley proposed that laughter may reduce some existing mental health problems and a sense of humor may moderate the perceived intensity of negative life events.

B. Social/Anthropological

Social and ethological theorists (e.g., Martineau) have viewed laughter as a marker of group membership and solidarity (for those who share a joke) or exclusion (for those who are the "butt" of the joke) and as a means of maintaining social control and group cohesiveness.

Anthropologists have described "joking relationships" in various traditional societies; these refer to ritualized teasing between specific kin in which the receiver is required to take no offense. In an incisive analysis, anthropologist Mary Douglas suggested that joking is a subversive act in that it levels hierarchy and represents a triumph of intimacy over formality and of unofficial values over official ones. Not surprisingly, this subversive aspect of humor is particularly evident among members of minority groups who may use humor to portray reality in a way that exposes social inequities. Much feminist humor and political humor, particularly that found in Central and Eastern European countries while under authoritarian rule, is of

this type. So-called "gallows" humor also has elements of this subversive aspect of humor.

C. Metaphysical

Although laughter appears to be associated with happiness and joy, some have taken the position that laughter arises precisely from the experience of suffering. The philosopher Nietzsche, for example, argued that humor was invented precisely because of the extent to which humans suffer. This view is echoed in existentialist philosophy. For the most part, though, humor has not been taken seriously in classical philosophical thought in the Western tradition, which has tended to regard it as irrational, irresponsible, and frivolous. In contrast, Eastern traditions, such as Hindu, Buddhist, and Sufi worldviews, regard humor as an appropriate reaction in the face of the vicissitudes of death, disease, and aging; humor here is seen as an insight that all human aspirations are ultimately comical and that wisdom ensues from that insight.

D. Evolutionary

Laughter is clearly an important aspect of human nature. All societies appear to value it. It has distinct facial and vocal manifestations. It emerges spontaneously in early childhood and persists through the life span and it is an intensely pleasurable social activity. All these aspects suggest that laughter may be a reasonable candidate for a psychologically adaptive characteristic. Indeed, it has been accorded a special place in several recent theoretical accounts of human evolution. Aside from the false alarm theory mentioned earlier, at least five different evolutionary accounts of laughter and humor may be distinguished.

1. Humor as a Temporary Disabling Mechanism

When we are truly amused by something, more often than not it takes us by surprise and distracts us from whatever it was that we were in the middle of doing or thinking. If we are heartily laughing, we quite literally cannot think of, let alone do, anything else. Wallace Chafe suggests that this disruptive effect of humor may be evolutionarily significant (i.e., that humor's basic adaptive function is a disabling one). Chafe proposes that the humor state may have arisen precisely in order to keep us from doing things that might be counterproductive; that is, things that our usual form of reasoning might lead us into but that, in a larger sense, would be undesirable. This view fits with a related notion of humor as a device for pointing out and sharing counterexamples, or providing disconfirmatory evidence, which in turn is useful in reasoning and problem solving. Chafe further suggests that the fact that laughter provides an audible signal to others may be significant because laughter signals to the receiver that the laugher is in a humor state and thus cannot be taken seriously; moreover, the laughter may in turn infect others present, rendering them temporarily disabled as well.

The notion that humor is a disabling device that is ultimately adaptive is an intriguing hypothesis and one that readily lends itself to empirical test. For example, one could ask whether the effect attributed to humor depends on the type of humor, the type of problem from which the laugher has been disrupted, the relevance of the humor content to the problem content, the presence of others, etc.

2. Humor as Social Learning

According to G. Weisfeld, humor provides exposure to fitness-relevant scenarios in a nonserious, playful context, motivating the practice of social skills that will be useful in serious contexts. In this view, humor is basically a form of social–intellectual stimulation that allows for the seeking and practicing of social skills that are fitness enhancing at some later point. Laughter, according to Weisfeld, conveys appreciation and gratitude to the humorist for having provided such stimulation.

3. Humor as Status Manipulation

The notion that mirthful laughter serves to create and solidify group boundaries forms the centerpiece of an evolutionary account of humor proposed by Richard Alexander in which humor is seen as a way of favorably manipulating one's status in a group to improve one's access to resources for reproductive success. Specifically, it is proposed that humor developed as a type of ostracism, a means of manipulating one's social status in a group (and thereby one's access to critical resources) by facilitating bonds with certain members and ostracizing others, through explicit or implicit exclusion.

Alexander's theory views humor primarily as a device for establishing dominance and thereby for being more competitive for critical resources. He

suggests that humor, as a form of social scenario building, gives one an edge in learning how to negotiate in fitness-relevant domains. Humorous put-downs may be more successful than direct criticism or insults as a way of establishing dominance since they are indirect and therefore more face-saving ways of manipulating status than direct displays of hostility or strength. In Alexander's model, all humor, whether or not it has an ostensible butt or target, essentially developed in the service of status manipulation.

4. Humor as Vocal Grooming

An emphasis on vocal grooming characterizes an evolutionary view of language proposed by Robin Dunbar that has implications for humor. Dunbar speculates that human language evolved as a vocal extension of physical grooming, allowing, as it were, the simultaneous grooming across space of more than one partner and thereby the facilitating of social bonds with larger groups of animals. Language fosters social bonding by permitting the exchange of gossip, i.e., socially relevant information (specifically, who is doing what with whom), as well as the sharing of one's own experiences, actual or desired. The fact that laughter typically accompanies speech suggests that it facilitates and cements the social bonding process.

In this view, the pleasure in shared laughter is analogous to the pleasure of actual physical touch. Humor has in fact been characterized as a form of vicarious touch. The importance of touch for primate emotional development is well documented. Laughter may have served to attract like-minded others in the bonding process; as Konrad Lorenz noted, finding the same thing funny is not only a prerequisite to a real friendship but also very often the first step to its formation. Thus, just as language facilitates social bonding, laughter accompanying speech, and humorous discourse with or without laughter, could have evolved as a way of solidifying affectional bonds.

5. "The Wit to Woo": Humor as Mate Attraction

The most recent evolutionary proposal regarding humor emphasizes its importance in courtship and suggests that the creativity and unpredictability that is inherent in humor is seen as attractive to a prospective mate. In this account, proposed by Geoffrey Miller in *The Mating Mind*, creativity is seen as not just a random by-product of chaotic neural activity but also as something that evolved as an indicator of intelligence and youthfulness and as a way of playing into our attraction to novelty. Creativity not only allows one to find unpredictable solutions to problems but also provides inherent pleasure because of its protean, unpredictable nature. Humor encapsulates these elements of creativity and it is therefore not surprising that a sense of humor is highly desired in a prospective mate. In this view, humor evolved because of its importance in courtship and mate choice.

The five evolutionary hypotheses emphasize different elements of humor. Humor is hypothesized to disrupt our routine for the better, to teach and reward us in our social interactions in fitness-relevant domains, to unify a group but also create group divisions, to create and strengthen affective bonds, and to attract and sustain a partner. Although there is evidence consistent with each of these views, research designed to examine the assumptions and claims of the different accounts is clearly warranted.

In the remaining sections we review theory and research on the expression and neural mediation of normal and abnormal manifestations of laughter and humor.

VIII. NORMAL LAUGHTER

A. Visuomotoric and Physiological Aspects

Laughter is a motoric and vocal activity requiring the coordinated action of 15 different facial muscles and associated clonic contractions of the thoracic cage and the abdominal wall. Laughter produces spasmodic skeletal muscle contractions, tachycardia, changes in breathing pattern, and increases in catecholamine production. Hearty laughter increases heart rate, blood pressure and respiratory rate, and muscular activity.

Physiologically, laughter is the opposite of crying. Although the upper half of the laughing face appears indistinguishable from the crying face, the lower half and the respiratory pattern in laughter are the reverse of those of crying. In laughter a characteristic facial display known as the Duchenne display is invoked. This display involves the joint contraction of the zygomatic major muscles (i.e., pulling the lip corners back and upwards) and the orbicularis oculi muscles (i.e., raising the cheeks and causing the eyes to wrinkle). Indeed, the movement of the eye muscles may serve as a marker for distinguishing emotional from voluntary (forced) laughter since the orbicularis oculi muscles do not contract during voluntary laughter. Laughter is also accompanied by the flaring of

nostrils, mandible retraction, and brightening and sparkling of the eyes.

In terms of respiration, laughter typically begins with an abrupt exhalation followed by rhythmic expiration/inspiration cycles, which may or may not be phonated as "ha ha ha." No inspiration preceding the laugh is needed since laughter is produced at a low lung volume. The rhythmic pattern of laughter respiration is produced by contractions of muscles that are typically passive during normal expiration, i.e., the diaphragm, the abdominal (rectus abdominus), and the rib cage muscles (triangularis sterni). These muscles work together with the larynx. In crying, the saccadic contractions occur mainly with inspiration, whereas in laughing they occur with expiration; in both cases the contractions are accompanied by short, broken sounds. Autonomic correlates of laughter include arousal and muscle tension followed by relaxation of muscle tone, dilatation of cutaneous vessels of the face, neck, and hands, lacrimation, and fatigue.

In 2000, Niemitz, Loi, and Landerer examined specific visuomotoric aspects of the human laughing face and their affective interpretation by human raters. The expressive facial movements of 45 videotaped laughing adults and 13 children were shown to more than 100 adult raters. The results suggested that visual aspects of laughing faces that are judged as cheerful, winning, or generally positive were those in which (i) the laughs were fairly long, ranging from 3 to 6 sec; (ii) there were rapid eye and mouth movements during the first second of the expression of laughter; (iii) there was repetition of the mouth and eye movements; and (iv) there was almost complete closing of the eyelids, sometimes repeatedly, for 0.1–1.5 sec. There is a need for more such research quantifying how the various laughing expressions are decoded and interpreted, not just in the visual channel but also acoustically.

B. Acoustic Aspects

Willibald Ruch and Paul Ekman proposed a useful framework for describing the structure of laughter. They define a laughter bout as all the respiratory, vocal, facial, and skeletomuscular elements involved in a particular laughter event. A bout of laughter, in turn, consists of an onset (the prevocal facial component), an apex (involving vocalization or forced exhalation), and an offset (the postvocalization part, typically a smile that fades out). The apex contains laugh cycles, (i.e., repetitive laugh pulses); there are typically 4 pulses in a laugh cycle, although the number can range from 9 to 12 depending on lung volume. There are usually about 5 pulses per second. At the level of laryngeal movements, a laugh pulse can be further subdivided into the number and duration of vibratory cycles of the vocal cords. In one analysis, the duration of laughter pulses varied from 30 to 100 msec. Phonation in laughter pulses involves a series of stereotypic laryngeal adjustments that include four stages: an interpulse pause (a moment of quiet aspiration in the periods between voicing), adduction (closing) of the arytenoid cartilages, vibration of vocal chords, and abduction (opening) of the arytenoid cartilages.

Several features in the acoustic signal may serve to cue the degree of positive affect experienced by the laugher. Widening versus narrowing of the pharynx is known to affect voice quality and may signal friendly versus scornful laughter, respectively. Other cues may be provided by harmonics, melodic contour, and duration. The melodic contour, including the intonation and pitch contour, may be particularly informative of emotional state and meaning. The vocal ligaments are more likely to be tensed under conditions of arousal or anticipation, giving rise to laughter that has a rising melodic contour. The lips and the cheeks are typically elevated in joyful smiling or laughter; contraction of the muscles involved in the Duchenne display changes the form of the mouth opening, constraining the vowel sounds that can thus be produced. Indeed, one study showed that listeners can reliably infer a smiling from an unsmiling version of the same spoken message purely on the basis of the voice.

Fewer than a dozen studies have been conducted on the acoustic characteristics of laughter. Almost all have used adult participants. One study examined the acoustic characteristics of a group of 30 male and female adults from whom one laugh was elicited under each of four conditions: social, humor, tension release, and tickle. Significant differences in duration, intensity, and mean frequencies of three vowel formants above the fundamental frequency were found between the humor and social laughs and the tension release and tickle laughs. Moreover, listeners could reliably classify the laughs into the different types, suggesting a communicative aspect to the laughs. The mean fundamental frequency of laughter for males (175 Hz) was higher than that for females (160 Hz) and highest for the tickling condition. In another study, a group of 11 male college students were recorded while they viewed a videotape of comedic storyteller Bill Cosby. Analyses

of 55 bouts of laughter showed that the mean peak fundamental frequency while laughing was more than twice as high as that when the subjects were counting. Moreover, an extended range of frequencies was exhibited during laughter, with a difference of 344 Hz between the lowest and the highest range observed. In general, there was considerable variability both within and between subjects on all the measures studied.

A limitation of existing studies with adults is the lack of naturalistic social interaction contexts. Studies with children fare better in this regard. One study recorded laughter among four 3-year-olds during three sessions of spontaneous free play between mother and child. An acoustic analysis revealed that the duration of laughter syllables in children was about the same as that found in adult laughter (i.e., about 200–220 msec). The total duration of the average laugh was also similar to that of adults. The main difference between adult and child laughter was in fundamental frequency, with most children's laughter having a frequency in the upper range of adult female laughter (400–500 Hz) or, in the case of one type of laughter (squeals), even higher (nearly 2000 Hz). Four distinct laughter types were noted in the child sample: (i) comment laughter (i.e., laughter occurring in conversational contexts that lasted about one-fifth of a second), which was subdivided into dull comment and exclamatory comment laughter; (ii) chuckle laughter, which lasted half a second and tended to occur in situations provoking more excitement than comment laughter; (iii) rhythmical laughter, which lasted 1–15 sec; and (iv) squeal laughter, which lasted half a second, had a very high fundamental frequency, and usually occurred without a break. The latter two kinds were more prevalent in the context of physical stimulation, anticipation, and fear.

It appears, therefore, that different types of spontaneous laughter have different acoustic characteristics. Moreover, laughter differs in different social contexts. Additional studies are needed with different kinds of communication dyads and with different age groups to examine in more detail the relationship between acoustic distinctions in laughter production and their affective interpretation by listeners. The specific ways in which spontaneous versus contrived laughter may differ, both structurally and neurobehaviorally, also warrant study.

C. Health Aspects

An implicit belief underlying most accounts of humor is that it confers psychological benefits. As reviewed by Martin, the benefits of humor may also extend to the physiological level. For example, laughter is known to increase respiratory rate and clear mucus. It is entirely conceivable that laughter may thus be beneficial to patients with chronic respiratory conditions such as emphysema. The increased heart rate and blood pressure accompanying laughter can exercise the myocardium and improve arterial and venous circulation, allowing a greater flow of oxygen and nutrients to tissues. This in turn may facilitate the movement of immune elements useful in fighting infections. Muscle relaxation following hearty laughter may break the spasm–pain cycle in patients with neuralgias or rheumatism. A reduction in laryngeal muscle tension accompanying laughter may help patients with vocal fold pathology to produce a more relaxed voice, and it may facilitate the recovery of phonation in patients with psychogenic dysphonia, an inability to phonate during speech despite intact articulation and the absence of any identifiable pathology of the larynx. Finally, the increased catecholamine levels associated with laughter may be responsible for the beneficial effects humor is thought to exert on mental functions, such as alertness and creativity. Many of these claims of the healthful benefits of laughter and humor have yet to be subjected to systematic study.

D. Neuroanatomical Hypotheses

There has been very little direct empirical investigation of neuroanatomical correlates of normal laughter. A hypothetical neural circuit was first theorized in 1924 by Kinnier Wilson. According to Wilson's model, laughter is produced by a medullary effector center that links the seventh nerve nucleus in the pons with the 10th motor nucleus in the medulla and with phrenic nuclei in the upper cervical cord. This center is modulated by the cerebral cortex and limbic structures by means of an integrative center in the mesial thalamus, hypothalamus, and subthalamus.

Activity of the laughter center is thought to be determined by a voluntary pathway (corticobulbar fibers) and fibers extending from the orbital surface of the frontal lobes through the bulbar nuclei. Input to these fibers comes from an involuntary pathway (the basal ganglia), which appears to be inhibited by the voluntary one. No single cortical area supplies the origins for the voluntary and involuntary fibers; the input derives from diffuse cortical regions including the frontal, premotor, motor, parietal, temporal, and hippocampal regions.

The notion of hypothalamic integration of cortical and limbic control on the brain stem laughter center remains to be substantiated, although it is consistent with clinical observations.

IX. DISORDERS OF LAUGHTER

Disorders of laughter and crying are very rare. Although strong emotional reactions are not uncommon following brain injury, the most common reaction described is fear. Medical research on laughing and crying disorders consists largely of case reports, with few systematic, large-scale investigations. Abnormal laughter, defined as laughter that is involuntary and inappropriate to the situational context, is most often seen in generalized affective or cognitive disturbances such as psychoses (e.g., schizophrenia). Individual cases of hysterical laughter spells, marked by silly, unrestrained, unmotivated, and unprovoked laughter, have been reported, as has epidemic hysterical laughter.

In many cases of abnormal laughter, the motor act of laughing may be dissociated from its emotional aspect. Such a condition has been termed pathological laughter. Kinnier Wilson defined pathological laughter and crying (PLC) as "a sequel to and consequence of a recognizable cerebral lesion or lesions in which attacks of involuntary, irresistible laughing or crying, or both, have come into the foreground of the clinical picture." This definition draws attention to the impaired control and the episodic aspect of the abnormal emotional expression.

In 1994, Shaibani, Sabbagh, and Doody proposed four criteria to distinguish pathological laughter and crying from normal laughter and crying. First, PLC is inappropriate to the situation since it occurs spontaneously or in response to nonspecific stimuli or inappropriate, arbitrary stimuli (e.g., one patient with a left cerellar hematoma showed pathological laughter following left hand tremor). Second, PLC is unmotivated; that is, there is no relation between the affect and observed expression, nor is there relief or mood change afterwards. Third, PLC is involuntary; that is, it has its own pattern and occurs against the patient's will. Neither the duration nor the content of PLC can be controlled, and patients do not gradually change from smiling to laughing but have a sudden, brief outburst without any warning. Fourth, PLC differs from emotional lability, which refers to an exaggerated emotional response to a normal stimulus. The latter characterizes patients with multiple sclerosis and Alzheimer's disease. These patients are overcome by uncontrollable laughter and crying that is usually appropriate to the situation and accompanied by mood alteration.

Very few patients actually meet the criteria for PLC. However, researchers do not concur on the ideal classification system and whether PLC must involve a dissociation between the expressed affect and the mood. It is acknowledged that the degree of volitional control of PLC varies from complete absence to some control.

The most common conditions associated with PLC are summarized in the following sections, subdivided into disregulatory and excitatory conditions.

A. Disregulatory Conditions

One disregulatory condition giving rise to abnormal laughter is amyotrophic lateral sclerosis (ALS). In 25% of ALS patients there is bulbar involvement; of these, 30–50% develop PLC. Patients without bulbar involvement do not develop PLC.

Multiple sclerosis is another condition that in 7–10% of patients leads to PLC, sometimes with euphoria. Characteristics of so-called pseudobulbar palsy associated with multiple sclerosis (as with bilateral strokes) include dysarthria, dysphagia, bifacial weakness, and weak tongue movements but preserved coughing, yawning, laughing, and crying.

In one large-scale study, PLC was observed in 15% of stroke patients 1 month after the stroke, in 21% at 6 months after the stroke, and in 11% a year following the stroke. Abnormal emotional reactions were particularly associated with lesions in the left frontal and temporal regions.

A rare syndrome of PLC is associated with acute infarction—the so-called "fou rire prodromique," first described in 1903 and in only a few cases since. In these cases, laughter lasts from between 15 minutes to 24 hours and is almost always followed by death. Lesions in this syndrome typically involve the left internal capsule-thalamus, left basal ganglia, or ventral pons. This syndrome contrasts with pseudobulbar palsy since it is not recurrent and, unlike epileptic laughter, it is not associated with electroencephalograph (EEG) changes or confusion.

A host of extrapyramidal disorders, including Parkinson's and Wilson's disease, are also associated with PLC. The syndrome of Angelman, a genetic disorder characterized by mental retardation and stiff puppet-like movements in children between the ages of

2 and 6 years, is also associated with frequent bursts of laughter. A variety of toxins, such as nitrous oxide and insecticides, have also been related to pathological laughter. Finally, various malignant brain stem tumors, such as clival chordoma and pontine glioma, have been implicated in pathological laughter.

B. Excitatory Conditions

The most common excitatory condition that is associated with pathological laughter occurs in epileptics. Laughter as part of an epileptic seizure was documented by the neurologist Trousseau in 1873 and was described in Dostoevsky's novel *The Idiot*. In 1957, the term "gelastic epilepsy" was coined to refer to epileptic fits in which laughter is the only or the most common symptom. More than 160 cases of gelastic epilepsy have been reported in the literature. Gelastic epilepsy often occurs in patients with hypothalamic hamartomas and precocious puberty, in patients with complex partial seizures and temporal lobe origin seizures, and in children with infantile spasms. The seizures usually begin in infancy or childhood and are associated with cognitive decline in later years. In patients with these seizures, involuntary, mechanical giggling typically occurs as the initial ictal behavior before an alteration in consciousness. The duration of the laughter in patients with hamartomas is usually less than 30 seconds and the seizures occur several times a day.

In a recent review of published reports of seizures involving laughter, Biraben and colleagues proposed the following neural hypotheses for the genesis of laughter:

1. Laughter arising as a reactional behavior (i.e., in response to a pleasant feeling or mirth: Only a few such cases have been observed, all involving seizures with a temporal focus. In one such case, involving stimulation of the left temporobasal region (fusiform and parahippocampal gyri), the individual experienced a change in the semantic connotation of stimuli (things became funny); in another case with stimulation in the same region, the modification experienced was perceptual (things changed in a funny way). Biraben *et al.* suggest that the laughter in these cases might be a physiological response to a modified cognitive process.
2. Laughter arising as a forced action or an automatism: Seizures with a frontal focus have characteristically produced laughter described as forced and unmotivated. Arroyo and colleagues describe one patient with a cavernoma of the anterior cingulate gyrus whose laughter appeared to be an irrepressible motor behavior; resection of the lesion eliminated the laughing behavior. They suggest that the premotor mesial system acts as an interface between the limbic loop, which includes the anterior cingulate gyrus, and the motor loop, which includes the supplementary motor area. Biraben *et al.* propose that cases of gelastic seizure originating in the anterior cingulum involve a critical functional disconnection between the motor loop and the limbic loop within the mesial premotor system. The laughter arising in these cases reflects a behavioral output from a motor program separated from all motivation.

Laughter in one other epileptic patient reviewed, a nongelastic case, occurred after cortical electrical stimulation of the lateral border of the rostral part of the supplementary motor area. The laughter was natural, rather than forced, and was accompanied by a general feeling of amusement toward the environment at large. Biraben *et al.* suggest that such laughter may reflect an imbalance between the mesial and lateral premotor systems.

C. Neuroanatomical Hypotheses

Evidence for the neuroanatomy of pathological laughter and crying comes from the following sources: a limited number of autopsy reports; studies of congenitally malformed infants; case reports of patients with pathological laughter following neurological disease; EEG activity and electrical stimulation of the cortex; injection of a barbiturate to the right or left cortex, typically done in patients with epilepsy to determine the language-dominant hemisphere; and studies of humor comprehension in patients with left versus right hemisphere damage.

The autopsy data are of limited value in localizing laughter since the patients had suffered diffuse brain pathology. A review of such cases revealed no single cortical lesion as causing PLC. Together with clinical case reports, the autopsy data indicate that both unilateral and bilateral lesions that affect the descending tracts to the bulbar nuclei can cause PLC, as can lesions on either side in the anterior limb and genu of the internal capsule adjacent to the basal ganglia, thalamus, hypothalamus, and pons. Even localized brain stem lesions can cause pathological laughter. Case studies do not explain why some patients develop pathological laughter while others develop pathological crying or both. Studies of congenitally malformed

newborns with severe anencephaly but with preserved pons and medulla indicate that newborns do not smile but can cry; those with intact midbrains can both cry and smile.

The predominant neuroanatomical account of pathological laughter and crying, first proposed by Wilson, regards it as arising from a loss of direct motor cortical inhibition of a laughter and crying center located in the upper brain stem. However, this explanation does not fully account for the range of phenomena observed in PLC. A recent alternative account, developed by Parvizi and colleagues, in light of new neuropathological findings, suggests that PLC is caused by dysfunction in circuits that involve the cerebellum (as well as the cortex) and influence brain stem nuclei.

There is evidence that the cerebral hemispheres may play different roles in the control of emotional states. In normal subjects the right hemisphere appears to be specialized for the perception and expression of emotion, particularly negative emotion. Lesions in the right hemisphere have been found to impair prosodic and lexical expression of emotion. Moreover, in epileptic subjects, injection of the right hemisphere with intracarotid sodium amytal has elicited unprovoked laughter, whereas left-sided injection has tended to produce bouts of crying. Patients with crying seizures tend to show right-sided foci, whereas those with gelastic seizures show mainly left-sided foci.

Patients with right hemisphere damage show a preserved sensitivity to the surprise element of humor but a diminished ability to establish narrative coherence. Disorders of humor, such as foolish or silly euphoria (so-called *moria*), and a tendency toward making inappropriate jokes (so-called *witzelsucht*), have been reported in patients with frontal lobe disorders including neurosyphilis. In 1999, a functional neuroimaging study by Shammi and Stuss determined that deficits in humor appreciation are restricted to patients with right frontal damage, supporting the view that the right frontal lobe serves an important role in integrating cognitive and affective information.

X. CONCLUSION

As may be evident from this overview, laughter and humor offer fertile ground for future investigation undertaken from a variety of theoretical and methodological perspectives, including ethological, cognitive, linguistic, clinical, and neuroscience. In a recent review, Jaak Panksepp observed that incisive research in this area has only just begun and that substantive research remains meagre. Many basic questions remain, such as the occurrence and significance of laughter in play and other social contexts; the role of laughter and humor in courtship, mate choice, and relationship maintenance; what different kinds of laughter communicate and to whom; and the neural circuitry responsible for normal laughter and feelings of mirth and joy.

See Also the Following Articles

CREATIVITY • DEPRESSION • EMOTION

Suggested Reading

Biraben, A., Sartori, E., Taussing, D., Bernard, A., and Scarabin, J. (1999). Gelastic seizures: Video-EEG and scintigraphic analysis of a case with a frontal focus; Review of the literature and pathophysiological hypotheses. *Epileptic Disorders* **1**(4), 221–228.

Brownell, H., and Stringfellow, A. (2000). *Cognitive perspectives on humor comprehension after brain injury*. In *Neurobehavior of Language and Cognition: Studies of Normal Aging and Brain Damage* (L. Connor and L. K. Obler, Eds.), Kluwer Academic, Boston.

Galloway, G., and Cropley, A. (1999). Benefits of humor for mental health: Empirical findings and directions for further research. *Int. J. Humor Res.* **12**(3), 301–314.

Goel, V., and Dolan, R. (2001). The functional anatomy of humor: Segregating cognitive and affective components. *Nature Neurosci.* **4**(3), 237–238.

Hull, R., and Vaid, J. (2001, July). *Cognitive basis of incongruity in verbal humor: An experimental inquiry*. Poster presented at the annual meeting of the Society for Text and Discourse, Santa Barbara, CA.

Martin, R. A. (2001). Humor, laughter, and physical health: Methodological issues and research findings. *Psychological Bull.* **127**(4), 504–519.

Mendez, M., Nakawatase, T., and Brown, C. (1999). Involuntary laughter and inappropriate hilarity. *J. Neuropsychiatr. Clin. Neurosci.* **11**(2), 253–258.

Niemitz, C., Loi, M., and Landerer, S. (2000). Investigations on human laughter and its implications for the evolution of hominoid visual communication. *Homo* **51**(1), 1–18.

Panksepp, J. (2000). The riddle of laughter: Neural and psychoevolutionary underpinnings of joy. *Curr. Directions Psychol. Sci.* **9**(6), 183–186.

Parvizi, J., Anderson, S., Martin, C., Damasio, H., and Damasio, A. (2001). Pathological laughter and crying: A link to cerebellum. *Brain* **124,** 1708–1719.

Provine, R. (2000). *Laughter: A Scientific Investigation*. Viking, New York.

Ruch, W., and Ekman, P. (2001). *The expressive pattern of laughter*. In *Emotions, Qualia and Consciousness: Proceedings of the International School of Biocybernetics Casamicciola, Naples, Italy, 19–24 Oct 98*. (A. Kaszniak, Ed.). World Scientific, Tokyo.

Shammi, P., and Stuss, D. T. (1999). Humour appreciation: A role of the right frontal lobe. *Brain* **122,** 657–666.

Vaid, J. (1999). *The evolution of humor: Do those who laugh last?* In *Evolution of the Psyche* (D. Rosen and M. Luebbert, Eds.), pp. 123–138. Praeger, Westport, CT.

Vaid, J., and Kobler, J. B. (2000). Laughing matters: Toward a structural and neural account. *Brain and Cognition* **42,** 139–141.

Vaid, J., and Ramachandran, V. S. (2001). Laughter and humor. In *The Oxford Companion to the Body* (C. Blakemore and S. Jennett, Eds.), pp. 426–427. Oxford University Press, Oxford.

Hydrocephalus

CHIMA OHAEGBULAM and PETER BLACK
Children's Hospital/Brigham & Women's Hospital, Boston

GLOSSARY

basal cisterns The cerebrospinal fluid containing spaces on the undersurface of the brain, the largest of which is the cisterna magna in the angle between the cerebellum and the back of the brain stem.

cerebrospinal fluid (CSF) The fluid contained within the ventricles and surrounding the brain and spinal cord.

choroid plexus The structures within the ventricles that produce most of the CSF.

fontanelle The soft membranous gap between growing skull bones in the infant skull, the largest of which is the anterior fontanelle.

lumbar puncture The insertion of a needle between lumbar vertebrae in the midline of the back into the space around the spinal cord containing CSF.

shunts Systems, typically involving tubing and a valve, diverting CSF from the ventricles to another body site.

ventricles The system of interconnecting fluid-filled cavities within the brain.

Hydrocephalus is the abnormal accumulation of cerebrospinal fluid (CSF) within the ventricles of the brain. This always involves enlargement of ventricles, and there may or may not be increased intracranial pressure. Hydrocephalus is almost always a result of impaired CSF absorption or circulation.

I. INCIDENCE/PREVALENCE

The true incidence or prevalence of hydrocephalus is unknown. As an isolated congenital disorder, it probably occurs in about 1 in 1000 live births. There are no good data on the incidence of adult hydrocephalus.

II. HISTORY

A. Early Concepts/Myths

Early references suggest that Hippocrates was aware of the condition now described as hydrocephalus. In 1768, Whytt published "Observations on the Dropsy in the Brain," in which he cited other writers as far back as the 13th century. Numerous herbal remedies were recommended in those times, though success rates were described as very low. Other treatments included head binding, leeching or bloodletting, injection of strong iodine solution into the ventricles, and exposure to the sun.

B. First Therapies Based on Anatomy

CSF was given its name by Magendie in 1825, and he and others, including Key and Retzius, elucidated its circulatory pathways later in the 19th century. Better understanding of the location of CSF within the ventricles led to more direct surgical approaches such as ventricular puncture. In 1891, Quincke described lumbar puncture as a treatment for hydrocephalus,

and this has evolved into the important diagnostic and therapeutic tool that it is today.

C. Early Shunts

As a better understanding of CSF circulation emerged, there were attempts to circumvent "obstructions" in the ventricular system in order to relieve hydrocephalus. Initial attempts to "shunt" CSF included external drainage in the late 19th century, invariably complicated by fatal infections. Other attempts were made using glass wool, gold tubes, catgut strands, and other materials to create conduits from the ventricles to the space beneath the scalp, the dura, and other areas. In 1908, Payr attempted the use of autologous vein to drain the ventricles into the sagittal sinus initially and later the jugular veins with some success. Others could not duplicate his results.

Cushing attempted bypass from the spinal subarachnoid space into the peritoneal cavity or retroperitoneal space by passing silver cannulae through the fourth lumbar vertebra. This had some success. The introduction of vulcanized rubber led to the availability of suitable conduit for the creation of shunts from the ventricular system to other body cavities. The first major innovation with the use of rubber catheters was by Torkildsen in 1939 to divert CSF from the lateral ventricles to the cisterna magna, but operative mortality remained high. It was eventually replaced by the more modern shunting procedures.

Numerous shunting procedures were created to other body cavities, notably the ureters. The ventriculoureteral shunt, reported by Matson in the 1950s, required a nephrectomy and had several infectious and metabolic complications. It also had a number of long-term survivors, however, and was the first effective shunt system used. Later shunts were developed to the atrial system, the pleura, and the peritoneum.

D. Other Operations

Walter Dandy, with Blackfan, identified the choroid plexus as the primary source of CSF production. On physiologic principles, in 1919, Dandy introduced choroid plexectomy as a treatment for hydrocephalus, in which the choroid plexus was ablated by cautery. This was initially performed as an open operation with a very high mortality rate but later was performed endoscopically by himself, Scarff, and others with a success rate of 70–80% and a mortality rate of approximately 15%. This procedure eventually fell out of favor with the subsequent discovery of significant extrachoroidal sources of CSF as well as the introduction of simpler shunting procedures.

Dandy also introduced open procedures for creating openings in the third ventricle into the basal cisterns on the undersurface of the brain. These were again developed into endoscopic procedures with success and mortality rates similar to those for choroid plexectomy. Patient selection for these was problematic before the modern imaging era because the optimal patient was one who had an obstruction in the aqueduct of Sylvius, which was the outflow tract for the third ventricle. There is a resurgence of interest in this procedure because of more sophisticated endoscopic techniques and equipment as well as the availability of appropriate imaging to select the right candidates.

E. Modern Era Shunts/Valves

The modern era of shunts owes its origin to the development of one-way valves by Spitz and Holter as well as by Pudenz. Polyvinyl chloride was initially used, but it was soon evident that silastic was better tolerated by the body.

The most popular shunting operation of this era was the ventriculoatrial shunt, in which the distal end of the catheter was introduced into the jugular vein and then to the right atrium. Numerous other sites for drainage have been reported, including the subdural space, mastoid air cells, thoracic lymphatic duct, fallopian tube, gall bladder, salivary ducts, stomach, and small intestine. Most of these sites are rarely, if ever, used today.

Ventriculoatrial shunting remained the most popular shunting procedure until ventriculoperitoneal procedures began to gain popularity at the end of the 1960s. The ventriculoperitoneal shunt was attractive for its simplicity, the absence of a permanent foreign body in the vascular system, and the reduced need for revision with inevitable growth of the patient.

III. CSF PRODUCTION AND ABSORPTION

A. Anatomy

1. Ventricular System

The foramen of Munro allows CSF to travel from the lateral ventricles to the third ventricle and the

aqueduct of Sylvins connects the third and fourth ventricles. Much of CSF is produced by the choroid plexus of the ventricular system and leaves through the openings in the fourth ventricle (foramina of Luschka and Magendie) into the subarachnoid space that surrounds the brain and spinal cord. It then travels around the convexities up to the sagittal sinus.

2. Choroid Plexus

The choroid plexus is the source of 60% of CSF production: It is composed of villi, which secrete CSF. Hypersecretion of CSF is virtually never a source of hydrocephalus. It should be noted that 40% of CSF is probably derived from fluid from brain parenchyma.

3. Arachnoid Villi

Arachnoid villi are specialized clusters of arachnoid cells that appear to allow CSF absorption into the sinuses and perhaps along nerve roots. In dogs and some other animals, there are no villi: CSF is absorbed along nerve root sheaths. In humans, arachnoid villi seem to absorb about 60% of CSF; the rest may be absorbed by nerve root sheaths as well.

B. CSF

CSF is continuously produced at the rate of approximately 500 cc/day. The total volume of CSF is approximately 150 ml at any time. CSF is formed as an ultrafiltrate from the villi of the choroid plexus and production is partially regulated by the enzymes sodium potassium ATPase and carbonic anhydrase. One function of the CSF appears to be mechanical, serving as a kind of water jacket for the spinal cord and brain, protecting them from potentially injurious blows to the spinal column and skull and acute changes in venous pressure. CSF hydraulically balances the brain within the skull.

The CSF is also believed to serve to remove waste metabolites of cerebral metabolism and may act as a pathway for distribution of peptides and hormones within the nervous system. This possible function has never been adequately explored.

IV. CLASSIFICATION

Hydrocephalus may be classified as acute or chronic (in acute, there are symptoms for weeks; in chronic, there are symptoms for months); communicating versus noncommunicating (in communicating hydrocephalus, there is flow between the ventricular system and subarachnoid space; in noncommunicating hydrocephalus, there is a block to flow within the ventricular cistern); congenital (acquired at birth) versus acquired; and normal pressure versus high pressure, a distinction that depends on the symptom complex.

A. Congenital

Several congenital structural abnormalities can lead to obstruction of CSF flow and result in hydrocephalus.

1. Aqueductal Stenosis

Different conditions may obstruct the aqueduct, leading to CSF accumulation above the level of the aqueduct that may be congenitally stenosed, occluded by a septum, compressed by gliosis in the surrounding periaqueductal tissue, or "forked." It can also be compressed or kinked by structural abnormalities or masses.

In aqueductal stenosis, the fourth ventricle, which lies beyond the point of obstruction, is typically normal in size. The term triventricular hydrocephalus has been used for this condition, referring to the enlargement of the third and both lateral ventricles.

2. Dandy–Walker Malformation

In this abnormality, there is atresia of the foramina of Luschka and Magendie along with agenesis of the cerebellar vermis. A large posterior fossa cyst results that communicates with an enlarged fourth ventricle. Hydrocephalus results in most of these cases.

3. Neonatal Intraventricular Hemorrhage

This condition is acquired particularly in premature infants and is characterized by progressive ventricular enlargement.

B. Acquired

1. Mass Lesions/Tumors

Masses may compress CSF pathways. This is a potential cause of hydrocephalus in those portions of

the ventricular system proximal to the obstruction. Any tumor along the CSF path may be involved. Colloid cysts of the third ventricle, pineal region tumors, and fourth ventricular tumors are examples. A tumor in the midbrain, for example, can cause critical aqueductal compression, resulting in hydrocephalus similar to congenital aqueductal stenosis. A cyst around the foramen of Monro can cause trapping of the lateral ventricle on that side. A posterior fossa tumor blocks aqueductal flow as well.

Tumors associated with abnormally high CSF protein concentrations impede reabsorption; ependymomas and other tumors of the spinal canal can do this. Diffuse spreading of tumor in the subarachnoid space can lead to hydrocephalus presumably due to both a higher CSF protein concentration and blockade of the arachnoid granulations.

2. Infection

Meningitis can result in fibrosis of the basal cisterns and arachnoid granulations that can lead to an obstruction of CSF around the brain stem. This includes tuberculosis meningitis, which is notorious for producing hydrocephalus.

3. Posthemorrhagic

Subarachnoid hemorrhage (most commonly from aneurysmal bleeding) and intraventricular hemorrhage can result in hydrocephalus because of the protein obstruction of arachnoid villi as well as inflammation of the aqueduct.

4. Idiopathic Normal Pressure Hydrocephalus

This entity was described by Hakim in 1965. It comprises communicating hydrocephalus with normal intraventricular pressures. Although a cause is not identified in most cases, it is usually believed to result from one of the obstructive causes described previously. An initial increase in pressure may lead to ventricular dilatation that reaches equilibrium, with pressures then returning to the normal range.

It is an important diagnosis because it is a treatable cause of dementia and gait disorder in the elderly. It is classically characterized by the triad of progressive gait disorder, dementia, and urinary incontinence. The gait disorder is usually shuffling with a broad base. Dementia is a recent memory disorder.

C. Pseudotumor Cerebri/Idiopathic Intracranial Hypertension

This diagnosis of exclusion is characterized by elevated intracranial pressure and papilledema in the absence of intracranial tumor or other recognized cause for increased pressure or hydrocephalus. Unlike hydrocephalus, most adults with this condition have normal or even "slit" ventricles on imaging studies. Headache is the cardinal syndrome. The mechanism of increased pressure is unclear in this condition but it appears to be a result of blockage at CSF outflow.

D. "Arrested" Hydrocephalus

This term refers to a condition of ventriculomegaly with no symptoms: The ventricles may sometimes be very large. It is not clear whether it is truly asymptomatic, as there may be very subtle problems associated with it.

V. DIAGNOSIS

A. Clinical

Infants may present with such symptoms as a large head, irritability, delayed development, and vomiting. Older children and adults may describe headaches, nausea/vomiting, or visual changes (diplopia, decreased acuity, or field cuts) as high-pressure symptoms. In normal pressure hydrocephalus the symptoms are gait difficulty, slowing of action, memory loss, and incontinence.

1. Exam

a. Infants Before the cranial sutures are "closed," the size of the head will enlarge in hydrocephalus. The fontanelles may be noted to be bulging and/or firm, with palpable separation of the suture lines between cranial vault bones. The scalp veins may be engorged and the eyes may not elevate above the meridian. The head circumference is increased when compared with standard growth charts.

b. Ophthalmologic Findings Infants may present with the "setting sun sign." This results from upward gaze palsy from pressure in the suprapineal recess.

There may also be an abducens nerve palsy that manifests as a weakness of sideways gaze. The long intracranial course of the abducens nerve makes it vulnerable to stretch injury in hydrocephalus.

These symptoms may also present in adults, but they are typically later or more unusual findings. More common is the presence of papilledema that results from transmission of CSF pressure along the sheath of the optic nerve to the optic disc.

2. Other Findings

Severe cases or those of rapid onset may be associated with reflex bradycardia or, in infants, apneic spells. Sudden death may occur from rapid decompression. A variety of endocrine abnormalities have also been associated with hydrocephalus, including infantilism and precocious puberty. These have been ascribed, at least in part, to the compression of the pituitary gland by ballooning and thinning of the floor of the third ventricle.

Memory loss, spastic paraparesis with mild spastic weakness of the upper extremities, and less often a mild dysmetria of the extremities may be signs and symptoms. Agitated state is sometimes a factor.

In adults and children older than 1 year, hydrocephalus may present with either a high-pressure or normal pressure syndrome. The high-pressure syndrome includes headache, nausea, and vomiting. The normal pressure syndrome includes gait disturbance with a broad base and small steps, memory loss, urinary incontinence, and slowing of action.

B. Imaging

1. Computed Tomography

Computed tomography (CT) scanning will show ventricular enlargement and may also help determine the cause of hydrocephalus (e.g., tumors/cysts, congenital anomalies, or the presence of blood). Asymmetry of the ventricular system or the pattern of ventricular enlargement may also suggest the cause, as in aqueductal stenosis. Periventricular hypodensity may be present in some cases, suggesting **extravasation** of fluid into the periventricular space. Besides their role in diagnosis, CT scans are also essential for assessing response to treatment as well as follow-up.

2. Magnetic Resonance Imaging

Conventional magnetic resonance imaging (MRI) provides similar information as that of CT scanning about ventricular size, but MRI has a higher yield in identifying or characterizing associated or causative abnormalities. Special sequences are useful for identifying or quantifying flow through the aqueduct, for example, and making the diagnosis of aqueductal stenosis as well as determining the efficacy of certain forms of treatment, such as endoscopic ventriculostomy.

3. Ultrasound

This imaging modality is practical only in infants who have a patent fontanelle. Its usefulness lies in its relative low cost (compared to CT and MRI) as well as its ease of application at the bedside or in restless infants. It is excellent at establishing ventricular size, symmetry, and even the cause of hydrocephalus (most commonly, intraventricular blood in this age group). It is also extremely useful for follow-up after treatment until the age of approximately 6 months, when the anterior fontanelle becomes too small.

C. CSF Pressure Monitoring

Monitoring of ventricular pressure is used occasionally in the context of normal pressure hydrocephalus. Lumbar puncture is an important tool in assessing CSF pressure but should be avoided in noncommunicating hydrocephalus. Measurement over several hours may result in the diagnosis of normal pressure hydrocephalus.

VI. TREATMENT

The treatment of hydrocephalus is largely surgical. Medical therapies are either obsolete or temporizing.

A. Medical Therapy

Abandoned therapies include glycerol, isosorbide, radioactive gold, and head wrapping. Diuretics are occasionally used for neonatal hydrocephalus secondary to intraventricular hemorrhage to temporize ICP control until enough blood is reabsorbed. Acetazolamide and furosemide are the drugs typically used, but both can cause significant electrolyte imbalance.

In neonatal IVH, spinal taps also serve as a temporizing measure to keep ICP in the ''safe'' range

until CSF reabsorption resumes. In cases in which ventricular enlargement persists after the disappearance of blood and return of CSF protein to near normal levels, the patients must proceed to surgical therapy.

B. Surgical Therapy

The bulk of therapy for hydrocephalus is surgical. In an acute situation, CSF may be drained via a ventricular catheter to a sterile bag. Normally this is not done for longer than 2 weeks. Most hydrocephalus is treated by shunt placement. Shunts are permanently implantable devices for the diversion of CSF to one of several extracranial sites. They consist of a ventricular catheter, a valve, and tubing draining to a body cavity.

1. Temporary Diversion

Temporary diversion is sometimes sought when there is known infection in the CSF space precluding the implantation of permanent devices. This temporary diversion typically employs a catheter inserted into the ventricle that is tunneled underneath the scalp and connected to an external drainage bag. These systems are utilized for a few days until a permanent shunt can be placed if possible.

2. Routes of Diversion

The ventriculoperitoneal shunt is the most frequently employed route of diversion. Advantages include its relative simplicity and its complications are typically easier to manage than with other routes of diversion.

The ventricular catheter is inserted into the frontal horn of one of the lateral ventricles via frontal or occipital approaches. It is attached underneath the scalp to a valve, from which tubing passes subcutaneously down to the peritoneal cavity. This operation is often performed employing only two small incisions—the first for passing the catheter into the ventricle and the second for entry into the peritoneal cavity. The tubing is passed from one incision to the other using specially designed tunneling devices.

Ventriculoatrial shunting was the preferred route for shunting before ventriculoperitoneal shunting became popular. It is still widely employed primarily in some centers or where peritoneal shunting cannot be used.

The ventricular catheter is placed in the standard way, and the distal tubing is passed into the internal jugular vein and down to the right atrium, usually through the common facial vein, which is a small tributary of the jugular vein in the neck.

Other distal sites include the pleural cavity (a common option for diversion after the peritoneal cavity), the gallbladder, and ureter. These are rarely employed because of technical difficulty, much higher complication rates, and diversion to the ureters has typically required a nephrectomy.

Torkildsen preceded extracranial CSF shunting in its development. It involves the placement of a catheter from the lateral ventricle into the cisterna magna—the large subarachnoid space adjacent to the cerebellum and brain stem. It is useful only for aqueductal stenosis, a fourth ventricular pathology. It has been replaced by endoscopic ventricular fenestration.

There are two main valve types: differential pressure valves and variable resistance constant flow vavles. Different pressure valves provide a constant resistance and permit the flow of CSF when a certain hydrostatic pressure has been exceeded. This is the common basic valve design and includes slit valves, ball-in-cone valves, and diaphragm valves. Recently, a variable pressure valve has been created that controls the pressure of CSF release.

Variable resistance constant flow valves attempt to maintain more natural constant flow rates by varying the amount of resistance to flow in response to pressure changes and, indirectly, to posture. These valves are believed to lead to overdrainage less often than do differential pressure designs.

The use of lumboperitoneal shunts fell out of favor largely because of distal obstruction and the high incidence of kyphoscoliosis. Use of these shunts involved a laminectomy, and the tubing initially used had a propensity to cause a chemical arachnoiditis.

This technique has been repopularized with the introduction of less irritating Silastic tubing and due to the ability to place the tubing percutaneously, with specially designed needles, into the subarachnoid space without a laminectomy.

3. Endoscopic Therapy

The most common endoscopic technique in use today is the third ventriculostomy method. This is used when CSF absorption is believed to be normal but there is an obstruction to its egress from the third ventricle. In this procedure, an endoscope is introduced into the third ventricle (typically via a frontal approach) and a small hole is made in the floor of the third ventricle to permit the flow of CSF from this cavity into the subarachnoid

space, thereby bypassing any mechanical obstruction in the ventricular system. The major indication for this procedure is aqueductal stenosis. Most surgeons, however, are not specifically trained in performing this procedure, and its widespread application is further limited by patient selection.

Newer endoscopic techniques include aqueductoplasty, which involves the endoscopic repair of short-segment strictures in the aqueduct, but experience is still limited with this procedure.

4. Lesion Removal

Occasionally, there are clinical situations in which a single lesion can be removed to cure, or at least improve, hydrocephalus. Certain tumors are amenable to removal. A choroid plexus papilloma, which overproduces CSF, can be resected, thereby obviating the need for a shunt. Another typical example is a colloid cyst, which commonly occurs around the foramen of Monro, obstructing CSF flow. Resection can lead to complete resolution of the obstructive hydrocephalus.

5. Choroid Plexectomy (Obsolete)

This procedure was performed as an open operation and involved the cauterization of choroid plexus. This carried a very high mortality early on, and it fell into disfavor. Later attempts were made to perform this endoscopically, and mortality rates decreased significantly, but it has not regained widespread use because of the relative safety and ease of other operations.

VII. COMPLICATIONS

Hydrocephalus, treated or untreated, results in a number of problems, some of which are described here. Untreated, hydrocephalus can result ultimately in death. Where its onset is relatively acute, the resulting increase in intracranial pressure will compress brain stem structures to the point at which neurovegetative functions are compromised. In more chronic courses, progressive neurological decline occurs, which will affect motor function, gait, and behavior.

Blindness may result from cortical lesions or damage to the visual pathways. Cortical blindness is caused by occlusion of the posterior cerebral arteries by downward pressure on them against the tentorial notch. Papilledema caused by increased CSF pressure may lead to optic atrophy and blindness.

Neurodevelopmental impairment can be attributed to both hydrocephalus and the initial condition that resulted in it. Hydrocephalus, where untreated, appears to cause sustained stretching of neurons and associated fibers, and this probably results in many of the clinical manifestations seen in the condition.

A. Shunt Malfunction

Shunt malfunction is typically the result of an obstruction in the shunt system. This may be proximal (in or around the ventricular catheter) or distal (in or around the peritoneal or other distal end). Occasionally, the valve may be obstructed by debris. These situations will often require surgery for replacement of the obstructed portion.

Proximal obstruction of the ventricular catheter is typically caused by adherent choroid plexus, debris, blood, or adherent brain tissue. Rarely, growing tumor or the inflammation resulting from infection can also result in proximal obstruction.

In the peritoneal cavity, obstruction is typically the result of low-grade infection that eventually causes formation of a pseudocyst around the tip of the catheter. Obstruction may also be caused by debris or malposition of the catheter from the initial surgery. Progressive growth of a patient who had a catheter placed early in childhood can lead to its withdrawal from the peritoneal cavity because it becomes too short for the patient's gradual increase in length. This previously common occurrence is relatively rare today because long redundant lengths of tubing are left in the abdomen at initial placement to account for expected growth. Disconnections, breaks, kinks, and other mechanical disruptions in the distal tubing may also occur.

B. Overdrainage

It is possible to "overdrain" CSF with a shunt. In the short term, this can result in symptoms similar to those of a "spinal headache," specifically headaches, nausea, and vomiting that tend to be postural and are relieved by recumbency. These symptoms will sometimes improve with time, but they may require shunt revision with a valve of higher pressure/lower flow rate or the addition of a so-called "antisiphon" device to the system to prevent overdrainage with upright posture. With variable pressure valves the pressure can simply be raised.

Overshunting of ventricles in the presence of thinned or atrophic brain can lead to collapsing away from the skull. This creates a subdural hygroma. This can result in tearing to bridging veins in the subdural space, causing bleeding, and the accumulation of a subdural hematoma.

Slit ventricle syndrome is a poorly understood complication of ventricular shunting characterized by collapsed or slit ventricles. It appears to occur when chronic overshunting of the ventricles results in a very low compliance state. The ventricles become slits and their walls may occlude the catheter impeding CSF drainage. The low compliance of the ventricular system then causes an abnormal increase in intraventricular pressure with small accumulations of CSF and symptoms of increased ICP result.

This problem is very difficult to treat. Options include replacing the ventricular catheter, changing to a higher pressure valve, or inserting an antisiphon device. Skull decompression is sometimes employed as a last resort on the principle that the miniscule increase in intracranial volume is enough to compensate for the changed compliance of the brain in this condition.

C. Shunt Infections

Shunt infections occur in about 10% of shunt procedures in children but are much less common in adults. They may occur at any time in relation to shunt placement and have been reported more than 10 years after shunt placement. The bacteria causing such infections are typically skin flora such as *Staphylococcus epidermidis* and *Propionibacterium* spp.

Most authorities recommend removing the entire shunt system in these circumstances, using an external diversion system during treatment with appropriate intravenous antibiotics, and placement of a new shunt after an interval from several days to a couple of weeks depending on a number of factors, including the virulence of the identified organism and the sterilization of CSF on serial cultures.

Other complications with shunts include the following:

- Occasional seizures (1%) that can result from the cortical irritation from catheter placement.
- Ascites (with peritoneal shunts) or pleural effusions (with pleural shunts) from poor local absorption of CSF.
- Bowel perforation, which can occur at the time of shunt insertion or even years later from gradual erosion of the shunt tubing into the bowel lumen.
- Specific problems with atrial shunting, including bacteremia and endocarditis as well as shunt nephritis, which for poorly understood reasons appears to result from the deposition of antigen–antibody complexes in the renal glomeruli following shunt infection. Pulmonary emboli (which can be septic) can also occur, leading to pulmonary infarction.

Outcome in hydrocephalus is related to the increased ICP as well as the primary cause of the condition. Deaths from hydrocephalus in countries in which neurosurgeons and shunt are available are uncommon. One study cited an 80–90% 5-year survival rate in children with hydrocephalus. Most data on mortality are set in the context of treatment complications. Death from shunt failure is rare, with reported rates of approximately 1%.

Cognitive deficits are difficult to interpret in the context of underlying brain abnormalities. One study examining children with arrested (untreated) hydrocephalus found 25% with IQs below 50 and 45% with IQs higher than 85%. Another study examining children with treated hydrocephalus found 72% with IQs between 70 and 100 and 32% with IQs higher than 100. Only 4% had IQs below 70.

VIII. OUTCOME

Some patients have a higher likelihood of benefiting from CSF diversion than others, in terms of motor function. Up to three-fourths of patients with normal pressure hydrocephalus in which gait dysfunction is the primary symptom will improve after shunting. On the other hand, patients with dementia and no gait disorder will rarely benefit from surgery.

In children, shunt placement for high pressure hydrocephalus is likely to lead to complete reversal of high pressure symptoms.

See Also the Following Articles

BRAIN ANATOMY AND NETWORKS • VENTRICULAR SYSTEM

Suggested Reading

Scott, R. M. (Ed.) (1990). *Hydrocephalus*. Williams & Wilkins, Baltimore.

Hypothalamus

J. PATRICK CARD and LINDA RINAMAN
University of Pittsburgh

GLOSSARY

circumventricular organs Areas of the brain that lack a blood–brain barrier and are found on the midline surrounding the third and fourth ventricles. These regions are essential "windows" through which the brain exerts humoral control of peripheral systems and are also responsive to feedback humoral influences from the systems that they modulate.

fornix A well-defined myelinated fiber bundle that serves as an important histological landmark in the hypothalamus and is also the principle conduit through which neurons in the subiculum project to the mammillary bodies.

infundibular stalk The ventral evagination of the floor of the third ventricle that connects the hypothalamus to the pituitary gland. It contains the long portal vessels that transport release and inhibiting factors from hypothalamus to the anterior lobe of the pituitary gland and is also the conduit for axons of magnocellular neurons that terminate in the posterior (neural) lobe of the pituitary.

median eminence The highly vascularized portion of the tuber cinereum that is essential for hypothalamic regulation of the anterior lobe of the pituitary. The blood vessels in this region lack a blood–brain barrier and are therefore capable of transporting peptides and neurotransmitters from the hypothalamus to the anterior lobe.

stria terminalis A major fiber pathway through which the amygdala communicates with the hypothalamus.

tuber cinereum The protuberance on the ventral surface of the diencephalon that contains the median eminence and gives rise to the infundibular stalk.

The hypothalamus is a remarkable region of the central nervous system. This small subdivision of the ventral diencephalon communicates extensively with other regions of the neuraxis via classical synaptic interactions and also has profound influences on the hormonal regulation of peripheral organ systems. In essence, it is directly responsible for the regulatory control of homeostatic systems essential for survival of the parent organism. The ability of this region to exert such profound influence over behavioral state and physiology is reflected in both the properties of hypothalamic neurons and the mechanisms through which they communicate. Thus, the dynamic regulatory capabilities of the hypothalamus are defined by the unique properties (e.g., timekeeping capabilities) of its constituent neurons, the ability to influence peripheral systems by virtue of hormonal or "humoral" communication, and the responsiveness of hypothalamic neurons to feedback control by the peripheral systems that they regulate. This article reviews the basic organizational principles fundamental to hypothalamic function, focusing on well-studied hypothalamic systems that illustrate the functional parcellation of this small but influential region of the brain.

I. HYPOTHALAMIC ORGANIZATION

As its name implies, the hypothalamus is found below the thalamus in the diencephalon (Fig. 1). It is

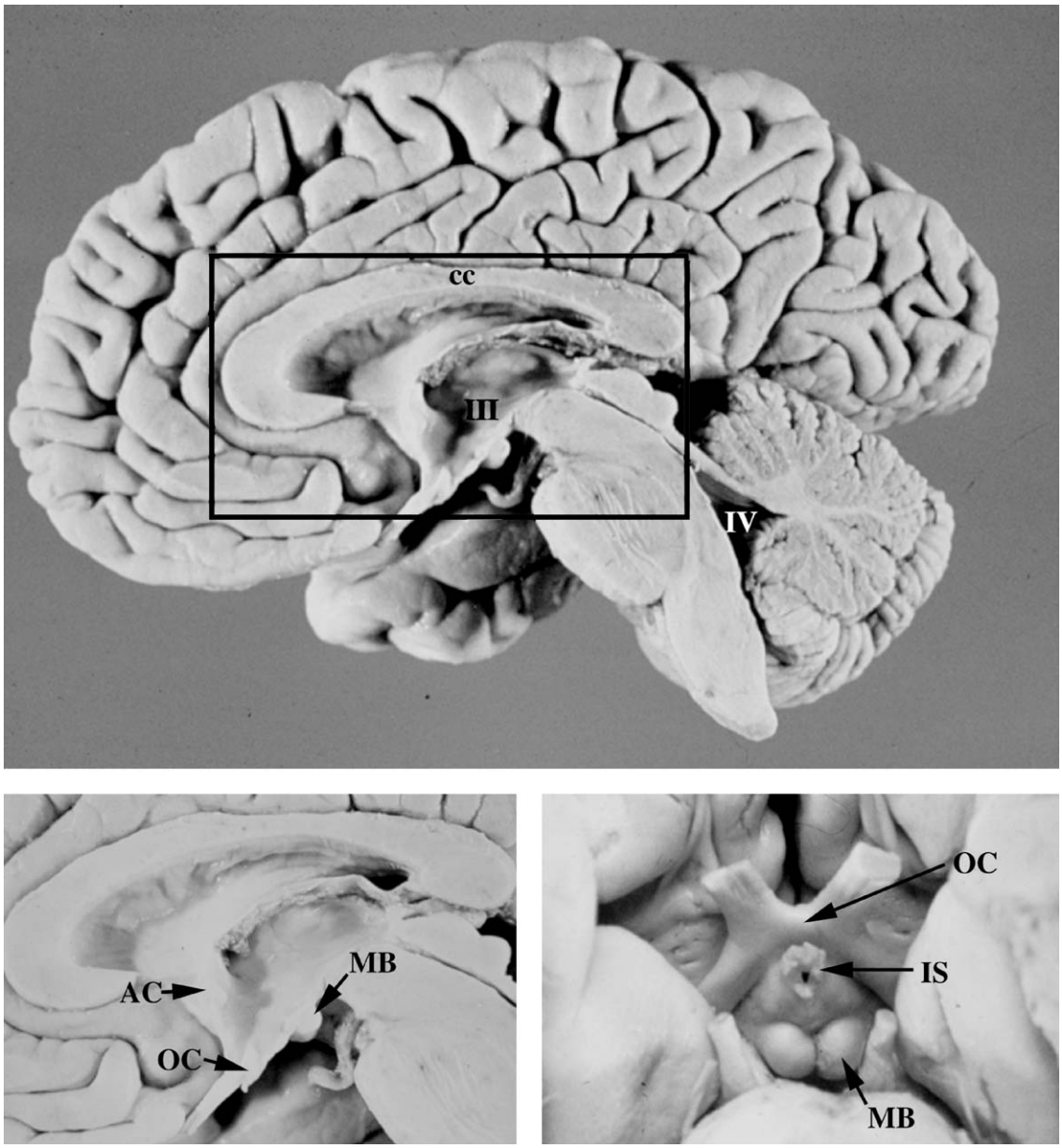


Figure 1 The disposition of the hypothalamus in relation to other regions of the central nervous system is illustrated in sagittal and ventral exposures of the human brain. In sagittal exposures of the brain the hypothalamus occupies a small region bounded by the anterior commissure (AC), optic chiasm (OC), and mammillary body. Ventral exposure of the brain reveals the three prominent landmarks that define the floor of the hypothalamus: the OC rostrally, the infundibular stalk (IS) arising from the tuber cinereum, and the paired spherical protuberances that constitute the MBs. III, third ventricle; IV, fourth ventricle.

distinguished by distinct external landmarks on the ventral surface of the brain that define its full rostrocaudal extent. These include the optic chiasm rostrally, the intermediate tuber cinereum marked by the prominent infundibular stalk, and the caudally placed mammillary bodies. Classic literature used these external landmarks to define internal subdivisions of the hypothalamus. Thus, it is still common to find references to the chiasmatic, tuberal, and mammillary subdivisions of hypothalamus. However, as our knowledge of the organization, connectivity, and function of hypothalamic cell groups has improved,

the rationale for dividing the hypothalamus into three regions defined by external landmarks has become less compelling. Nevertheless, these landmarks remain useful designations for defining the general location of hypothalamic cell groups and are commonly found in the literature.

Coronal sections through the hypothalamus reveal prominent internal landmarks that have proven useful in defining the basic organization of hypothalamic cell and fiber systems. Among the most prominent is the third ventricle that separates much of the hypothalamus into identical halves. The fluid-filled reservoir is particularly prominent in the intermediate portion of the hypothalamus demarcated by the tuber cinereum. In coronal sections through this level, the lumen of the ventricle defines the dorsoventral extent of the hypothalamus. The floor of the third ventricle is formed by a thin bridge of tissue, commonly known as the median eminence, that is continuous with a stalk connecting the ventral hypothalamus to the pituitary gland. The median eminence is among the most important interfaces through which the hypothalamus exerts regulatory control over peripheral systems in that it is essential for regulation of hormone secretion from the pituitary gland. It is similar to another midline strip of tissue that contains the organum vasculosum of the lamina terminalis (OVLT) and that forms the rostral wall of the third ventricle in that it lacks a blood–brain barrier (BBB). The absence of the BBB at these sites provides the essential means for neurohumoral communication and will be considered in greater detail later.

The internal organization of the hypothalamus is characterized best in histological preparations of coronal sections. Classic Nissl preparations reveal three longitudinal zones that extend throughout the rostrocaudal extent of the hypothalamus. The location of each zone can be defined in relation to its proximity to the third ventricle and a prominent myelinated fiber bundle, the fornix, that enters the hypothalamus rostrally and then courses caudally to terminate in the mammillary bodies. The *periventricular* zone is a densely packed group of neurons immediately adjacent to the third ventricle. Occasionally, it contains well-demarcated cell groups, such as the suprachiasmatic, arcuate, or paraventricular nuclei. However, in the majority of its extent it is composed of a thin, densely packed group of neurons immediately adjacent to the ependymal lining of the ventricle. The *medial* zone is found between the periventricular zone and a vertical plane passing through the fornix. Nuclear groups in this region (e.g., the ventromedial and dorsomedial nuclei) are among the most prominent and well-delineated cell groups in the hypothalamus. The *lateral* zone is found between the medial zone and the optic tract and internal capsule at the lateral extent of the diencephalon. Neurons in this region are more dispersed and do not form the distinct nuclear groups characteristic of the periventricular and medial zones. Nevertheless, immunohistochemical studies that have revealed distinct phenotypic parcellation of neurons in the lateral zone have contributed greatly to improving our understanding of the functional organization of this region.

II. MAJOR HYPOTHALAMIC CONNECTIONS

Connectional analyses have contributed greatly to our understanding of hypothalamic function and organization. Two well-known fiber tracts, the fornix and stria terminalis, provide major conduits through which the hypothalamus interacts with subdivisions of the limbic system. Each of these fiber tracts originates in the temporal lobe and pursues a looping trajectory over the thalamus to enter the hypothalamus dorsally and rostrally. An interesting new literature, particularly with respect to the fornix, has provided novel insights into the way in which axons traversing these pathways influence hypothalamic function. These advances have not only improved our understanding of how the hypothalamus functions within a larger ensemble of neurons but also provided a functional rationale supporting the aforementioned zonal organization of the hypothalamus.

As noted previously, early descriptions of hypothalamic organization established the intrahypothalamic course of the fornix as an important landmark separating the medial and lateral hypothalamic zones. Fibers traversing this tract divide into pre- and postcommissural bundles that pass either rostral or caudal to the anterior commissure. Individual axons of the postcommissural bundle leave the fornix along its course to terminate in hypothalamic nuclei while others continue caudally to terminate in the mammillary bodies. Axons in the precommissural fornix pass rostrally to terminate topographically within subdivisions of the lateral septum. The lateral septal neurons, in turn, project densely on the periventricular, medial, and lateral zones of the hypothalamus. Thus, the hippocampus provides a prominent topographically organized influence upon hypothalamic zones through parallel organized disynaptic projections involving the lateral septal nuclei. From the standpoint of sheer

magnitude and number of hypothalamic neurons targeted by these disynaptic projections, it would appear that the precommissural fornix has a much more substantial influence upon hypothalamic function than the axons that course through the postcommissural branch.

Prominent limbic influences on hypothalamic function also arise from the amygdaloid complex and course into the hypothalamus through either the stria terminalis or the ventral amygdalofugal pathway. In contrast to the fornix, these afferents do not segregate within hypothalamic zones. Rather, evidence indicates that they provide more directed input to hypothalamic areas and cell groups. For example, axons that course through the stria terminalis terminate more densely in the rostral rather than the caudal hypothalamus and also provide dense input to subsets of nuclei within a zone.

The medial forebrain bundle (MFB) is another prominent fiber tract associated with the hypothalamus. The trajectory of this projection system brings it through the full rostrocaudal extent of the lateral hypothalamus. This pathway is more diffusely organized than the postcommissural fornix and carries both ascending and descending fibers. Many of the axons that pass through the MFB are simply traversing the hypothalamus in transit to other forebrain targets. The prominent dopaminergic projections from the substantia nigra and ventral tegmental area fall into this category. Brain stem noradrenergic and cholinergic neurons also project through the MFB to both diencephalic and telencephalic targets. Thus, whereas the MFB is an important conduit for axons innervating the hypothalamus, it is also a major projection pathway for axons projecting to and from forebrain nuclei. Loss of function in response to lesions that interrupted these fibers of passage confounded the interpretation of results from early studies that incorrectly ascribed functions to hypothalamic nuclei that were, in fact, subserved by axons passing through the hypothalamus in the MFB. Improvements in tract tracing and lesioning methods, as well as the development of other functional probes, have clarified many of these false interpretations.

III. THE HYPOTHALAMUS AND THE TEMPORAL ORGANIZATION OF BEHAVIOR

The hypothalamus has long been recognized as an important integrative area for the regulatory control of behavioral state and the temporal organization of behavior. Early evidence from lesion and stimulation paradigms implicated regions of the hypothalamus in the control of sleep and other rhythmic aspects of behavior and physiology. However, identification of the specific circuitry through which the hypothalamus coordinates internal homeostatic function with sensory stimuli of the environment only occurred with the advent of technical advances that permitted the identification and functional dissection of populations of hypothalamic neurons. Today, our knowledge of the neural systems that exert regulatory influence over these functions is far more precise, and in some instances the molecular mechanisms that contribute to this control are beginning to be unraveled. The following sections review the functional organization of hypothalamic systems that participate in the control of rhythmic functions, sleep, and arousal.

A. A Biological Clock in Hypothalamus

It has long been clear that mammals and other organisms exhibit daily cycles in physiology and behavior that persist in the absence of sensory input from the environment. These cycles are approximately 1 day in length (hence the term circadian) and in normal circumstances are synchronized with the environment. In essence, they are rhythms of behavior and physiology that are tightly coupled to the most pervasive signal in our environment, the daily rhythm of light and dark. The adaptive significance of such an endogenous timekeeping mechanism is obvious when one considers the practical benefits of synchronizing behavior to the light-dark cycle. Certainly, the restriction of the activity of nocturnal animals to nighttime has the practical advantage of reducing the possibility of predation, and the ability to precisely measure day length permits seasonal breeders to deliver their progeny during the portion of the year when nutrients are most prevalent. Thus, it is not surprising that this endogenous timekeeping system is among the oldest and most highly conserved systems of regulatory control in the animal kingdom.

Although circadian rhythms in both plants and animals have been long been recognized, determination that a "biological clock" resides in the hypothalamus of mammals is a relatively recent event that can be traced to anatomical studies conducted in the early 1970s. Recognizing that the entraining influences of light are essential to any timekeeping system, two research groups independently utilized new methods

of defining neuronal connectivity to demonstrate that a circumscribed group of neurons overlying the optic chiasm receives dense retinal inputs. These neurons, comprising the suprachiasmatic nuclei (SCN), became the experimental focus of circadian biologists and there is now considerable evidence supporting the conclusion that a biological clock resides in the SCN of mammals. Specifically, animal studies have shown that the cells of the SCN exhibit a circadian rhythm of activity that is entrained by light, and that rhythmic aspects of physiology and behavior are abolished in animals in which the SCN have been destroyed. Importantly, it is also known that the rhythmicity of SCN neurons is genetically determined rather than the emergent property of a network, and "clock" genes that impart rhythmicity to SCN neurons have been identified. Thus, the hypothalamus contains a clock, the SCN, whose activity is synchronized to the external environment by virtue of sensory input transduced by the retina.

Elucidating the connections of the SCN has been an important component of understanding how this group of hypothalamic neurons imposes its temporal message on the physiology and behavior of the parent organism. One of the most well-characterized systems in this regard is the circuitry through which the SCN exerts regulatory control over the secretion of the hormone melatonin. Melatonin is secreted by the pineal gland in a circadian manner but is also responsive to light such that light stimulation during the dark phase of the photoperiod inhibits the normally high levels of melatonin secretion. This dynamic regulatory capacity renders the temporal profile of melatonin secretion a precise measure of day length. A large literature has established that the SCN controls both the circadian and photoperiodic aspects of melatonin secretion through multisynaptic pathways that sequentially involve the paraventricular hypothalamic nucleus, the intermediolateral cell column of the spinal cord, and neurons of the superior cervical ganglion that project to the pineal. Additionally, binding studies and localization of melatonin receptors have revealed that melatonin exerts feedback influence on the brain by binding to neurons in the SCN, paraventricular thalamic nucleus, and pars tuberalis of the infundibular stalk. Thus, the SCN not only regulates the secretion of melatonin but also is subject to feedback regulation by the hormone that it modulates. This is a common feature of the neurohumoral regulation exerted by the hypothalamus.

The SCN also imparts temporal organization to other systems and appears to do so through efferent projections that are largely confined to the hypothalamus. SCN neurons project to a relatively restricted group of nuclei in the hypothalamus that includes, but is not limited to, the region subjacent to the paraventricular nucleus (the subparaventricular zone), the preoptic area, and medial hypothalamic nuclei (e.g., arcuate and dorsomedial) involved in neuroendocrine regulation of pituitary function. These projections provide a substrate through which the SCN imparts temporal influences on a variety of systems.

B. The Hypothalamus and Sleep

Evidence that the hypothalamus is involved in the control of sleep emerged from a large literature dating to the early 1900s. Lesion studies correlating anterior hypothalamic damage with insomnia and caudal hypothalamic damage with somnolence were particularly informative. These and subsequent studies resulted in the concept of hypothalamic "sleep centers." A fascinating recent literature has demonstrated a cellular basis for hypothalamic influences on sleep and also provided insights into the means through which temporal organization is imparted on this behavior. It is now apparent that at least two distinct populations of neurons in the rostral and caudal hypothalamus are responsible for the hypothalamic effects on sleep. Using a creative experimental approach, it was demonstrated that neurons in a circumscribed region of the preoptic area [the ventrolateral preoptic nucleus (VLPO)] in rats express Fos, the protein product of the protooncogene *c-fos*, shortly following the onset of sleep. Since Fos expression reflects neuronal activation, this observation raised the possibility that VLPO neurons are involved in the initiation of sleep. A number of subsequent observations have validated this hypothesis and also revealed the larger network of hypothalamic neurons that participate in this function. Specifically, evidence now supports the conclusion that VLPO neurons inhibit arousal through projections to histaminergic neurons in the tuberomammillary (TM) nuclei of the caudal hypothalamus. In support of this conclusion, it was shown that GABAergic neurons in VLPO synapse on TM neurons and that pharmacological inhibition of TM neurons or blockade of histaminergic receptors promotes sleep. Those who recall the drowsiness that typically follows the use of early antihistamines (which act on central as well as peripheral histamine receptors) prescribed for the treatment of colds and allergies can

appreciate the major influence of the diffusely projecting neurons of the TM nuclei upon arousal.

Importantly, neurochemical lesions of VLPO and surrounding neurons have revealed greater functional parcellation of the anterior hypothalamic circuitry involved in sleep regulation. Cell-selective lesions that do not interrupt fibers of passage were shown to compromise different aspects of sleep based on the localization of the lesion. Lesions confined to the compact portion of VLPO dramatically reduce non-REM sleep and, in circumstances in which lesions are incomplete, the amount of non-REM sleep is linearly correlated with the number of Fos-expressing neurons in the portion of the VLPO that survived the lesion. Interestingly, lesions dorsal to VLPO that eliminate galanin-containing neurons that project to TM produce sleep deficits more closely associated with REM than with non-REM sleep. Collectively, these observations provide compelling evidence in support of a prominent role for the hypothalamus in sleep regulation and further indicate that there is functional parcellation in the neurons of the VLPO that participate in this control.

It is also clear that the hypothalamus plays an important role in the temporal organization of the sleep–wake cycle. Sleep is a circadian function, and although the SCN is not essential for the generation of sleep, it is responsible for consolidation of sleep within cycles that occur within a circadian framework. Thus, if the SCN are destroyed, rats will sleep approximately the same amount of time but this sleeping time will be distributed in many short bouts throughout the light–dark cycle rather than in a consolidated period. The circuitry through which the SCN imposes this circadian influence on sleep remains to be established, but it is likely that it occurs via polysynaptic connections that link the clock to nuclei involved in sleep regulation.

C. Hypothalamic Influences upon Arousal

Certainly, the aforementioned studies that have demonstrated a role for hypothalamic nuclei in sleep indicate that the hypothalamus influences arousal states. However, recent studies have demonstrated that the influence of the hypothalamus on arousal is not restricted to the TM neurons in caudal hypothalamus. In particular, a prominent group of neurons confined to the lateral hypothalamus has recently been implicated in the sleep disorder known as narcolepsy. These neurons express novel neuropeptides known as hypocretins or orexins and are differentially concentrated within the perifornical nucleus that surrounds the fornix in the tuberal hypothalamus. Mapping studies have shown that hypocretin/orexin neurons are similar to TM neurons in that they are confined to hypothalamus and give rise to extensive projections throughout the neuraxis. However, it is also clear that these neurons densely innervate areas (e.g., locus coeruleus) involved in the control of arousal, and there is good evidence that pathology of signaling pathways involving hypocretin neurons may be causal in narcolepsy. In this disorder, individuals exhibit daytime sleepiness and lapse unexpectedly into bouts of REM sleep. A dog model of the disease has identified a deletion in the gene encoding the hypocretin 2 receptor, and knockout mice lacking the hypocretin gene exhibit sleep disturbances similar to narcolepsy. Further, examination of postmortem human brains of narcoleptics has revealed substantial reductions in the number of hypocretin neurons, raising the possibility that the disease may be due to an autoimmune attack on the neurons. Thus, there is strong evidence that caudal hypothalamic neurons play an integral role in the regulation of arousal states.

IV. THE HYPOTHALAMUS AND NEUROENDOCRINE REGULATION

Neural regulation of pituitary secretion is one of the most important, well-characterized, and diverse functions of the hypothalamus. In fact, the demonstration that the hypothalamus exerts regulatory control over the pituitary is one of the landmark discoveries that tied the fields of neurobiology and endocrinology together. Recognition that hypothalamic control over the pituitary was humoral in nature resulted from structural studies that demonstrated a vascular link, or portal plexus, connecting these structures. Thus, the organizational features of the hypothalamic–pituitary axis are introduced below as a prelude to examining specific examples of hypothalamic control over pituitary secretion.

As noted earlier, the ventral diencephalon is connected to the subjacent pituitary via the infundibular stalk. Large-caliber magnocellular axons whose parent neurons reside in rostral hypothalamus traverse this stalk to terminate directly in the posterior lobe of the pituitary gland. This is the only direct neural

connection between the hypothalamus and pituitary and represents only a small portion of the regulatory capacity of the hypothalamus over pituitary function. Control of the secretory activity of the anterior pituitary is achieved through a portal vascular plexus that arises in the median eminence and then pursues a directed course along the stalk to end in the anterior lobe of this gland. An important feature of this portal system is the absence of the BBB in the capillaries of the median eminence. Fenestrations in the median eminence vessels allow peptides and neurotransmitters released from axons in their vicinity to gain access to the portal plexus, whereupon they are transported to the anterior pituitary to influence the secretory activity of cells in that portion of the gland. This architecture forms the basis for the neurohumoral regulation of pituitary function that is the foundation of neuroendocrinology (Fig. 2).

The absence of the BBB is a defining feature of a group of structures known as circumventricular organs (CVOs). These regions, which include the median eminence, are found on the midline of the brain

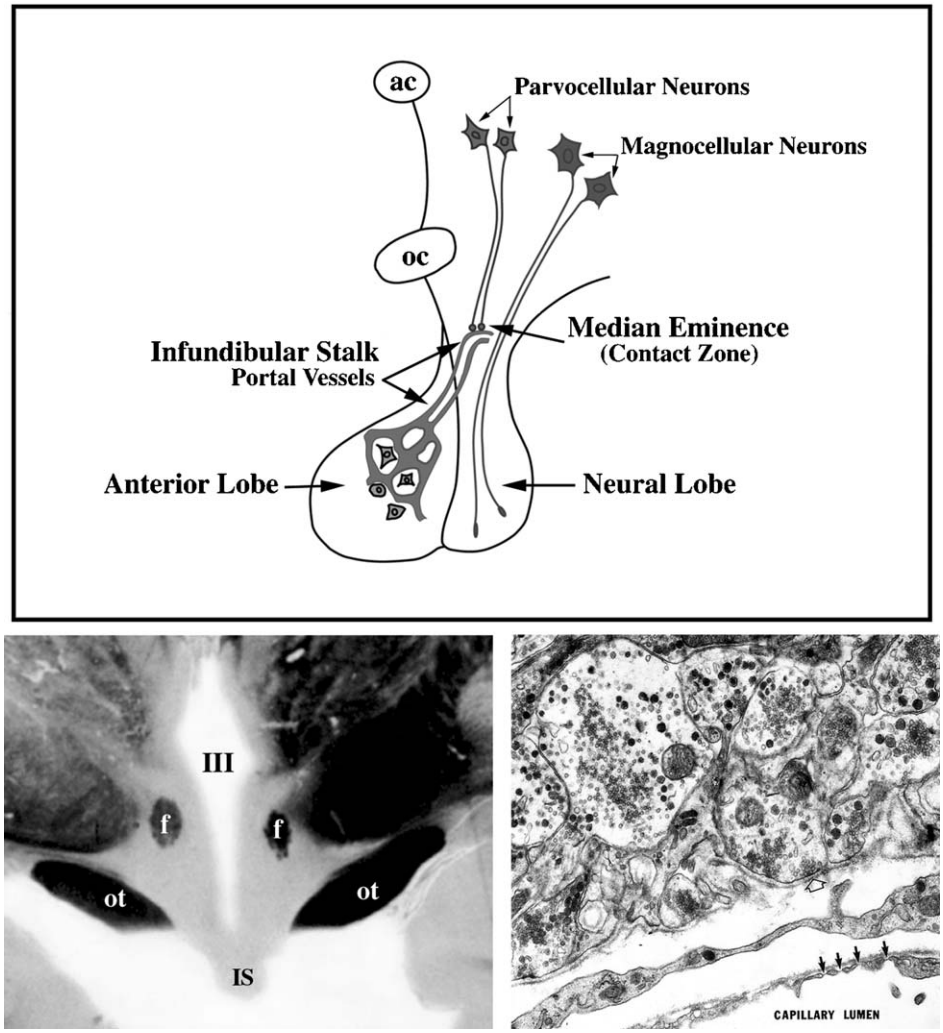


Figure 2 The basic organization of the hypothalamic–pituitary axis is illustrated in the schematic diagram. The anterior lobe of the pituitary is regulated by hypothalamic peptides and neurotransmitters that are released from parvocellular hypothalamic neurons into a vascular (portal) plexus and then travel to the anterior lobe to either stimulate or inhibit the release of hormones from cells in this portion of the gland. This is possible because the vessels at the neurohemal contact zone in the median eminence are fenestrated and large numbers of axon terminals terminate on the perivascular space adjacent to these vessels (bottom, right). Large magnocellular neurons project through the infundibular stalk (IS) to terminate in the posterior, or neural, lobe of the pituitary. ac, anterior commissure; f, fornix; OC, optic chiasm; ot, optic tract; III, third ventricle.

surrounding the third and fourth ventricles. Only two of the CVOs (the median eminence and OVLT) are found within the hypothalamus, but all are intimately associated with hypothalamic function by virtue of the connections that they maintain with various hypothalamic nuclei. For example, the absence of the BBB in the subfornical organ and area postrema allows neurons in these regions to respond to circulating cues and then modulate the activity of the hypothalamic–pituitary axis through classical synaptic connections with hypothalamic neurons. Thus, the absence of the BBB is essential not only for the ability of the hypothalamus to control the secretory activity of the pituitary but also for the feedback regulation of the hypothalamic–pituitary axis that imparts precision on endocrine regulation of peripheral systems.

The neurons that contribute to regulation of anterior pituitary secretion exhibit common features that are reflective of their function. First, they are confined to the hypothalamus and give rise to axons that terminate in the median eminence. Two hypothalamic nuclei, the paraventricular and arcuate, are particularly devoted to anterior pituitary regulation. Substantial numbers of parvocellular neurons in these nuclei give rise to dedicated projections to the external zone of the median eminence where their terminals abut on the perivascular space (contact zone) surrounding the fenestrated capillaries of the portal plexus. Neurons exhibiting the same organization are also dispersed in other areas of hypothalamus. For example, neurons involved in the regulation of growth hormone secretion are concentrated in the rostra1 periventricular and arcuate nuclei, whereas those that are important for regulation of ovulation in females are dispersed throughout the preoptic area. All these neurons project exclusively to the portal plexus in the median eminence and thereby exert their function in a neuroendocrine fashion.

A second feature of these neurons is their neurochemical diversity. Although they all have common structural features, the differing functions of these neurons are defined by their neurochemical phenotype. Many of the neurons manufacture and release small peptides that either stimulate or inhibit the secretory activity of cells in the anterior pituitary. Others utilize small molecule neurotransmitters such as dopamine toward the same end. These ''releasing'' or ''inhibiting'' factors impart another level of regulatory control over the hypothalamic–pituitary axis that is best illustrated by considering a specific example, such as growth hormone (GH) secretion. Peripheral metabolism is heavily influenced by release of GH from the anterior pituitary gland and this release, in turn, is regulated by two populations of hypothalamic neurons that produce opposite effects in the pituitary. Stimulation of GH release is under the control of neurons in the arcuate nucleus that produce growth hormone-releasing hormone, whereas inhibition of GH release results from the activity of somatostatinergic neurons in the rostral periventricular nucleus. Differential activation of these opposing systems permits precise regulation of GH secretion and emphasizes the importance of feedback regulation in activating the appropriate regulatory circuitry.

A role for hypothalamic timing mechanisms in neuroendocrine regulation is predicted by the rhythmic profiles of hormone release by the anterior pituitary and its target organs. This is clearly exemplified by the temporal profile of cortisol secretion by the adrenal gland. Release of plasma corticosteroids is under control of the hypothalamic–pituitary–adrenal (HPA) axis and exhibits a circadian profile, with peak levels occurring at the end of the dark phase in humans and the end of the light phase in rat. Although the temporal relations of these peaks differ between the two species, the temporal association of the peaks to activity is the same. Release of corticosteroids from the adrenal is under the control of corticotropin-releasing factor (CRF) neurons in the paraventricular hypothalamic nucleus. Release of CRF into the portal plexus at the median eminence elicits the synthesis and release of adrenocorticotropin hormone from the anterior pituitary, which subsequently stimulates corticosteroid release from the adrenal. Evidence suggests that the SCN modulates the activity of the HPA axis through a disynaptic pathway in which SCN neurons synapse on neurons in the dorsomedial hypothalamic nucleus that are presynaptic to CRF neurons in the paraventricular nucleus. Control of the HPA axis is also exquisitely sensitive to other sensory cues, such as feeding and stress. Thus, control of the HPA axis is a dynamic process in which the hypothalamus plays an important integrative role that defines the magnitude and temporal profile of corticosteriod release. It is probable that similar organizational principles account for the rhythmic release of other hormones.

V. HYPOTHALAMIC CONTROL OF AUTONOMIC OUTFLOW

As noted earlier, the hypothalamus plays an essential role in the maintenance of homeostasis. Neuroendocrine regulation of pituitary function constitutes one of

the primary ways through which this is achieved. However, the autonomic nervous system (ANS) is also intimately involved in homeostatic regulation. A substantial body of work clearly indicates that the hypothalamus plays a major role in controlling the activity of the ANS. Extensive descending connections to brain stem and spinal preganglionic neurons of the ANS arise from hypothalamic nuclei, and these hypothalamic nuclei are the target of feedback regulation that is both synaptic and humoral in nature. However, it is important to emphasize that these hypothalamic nuclei function within a larger set of forebrain areas that influence autonomic outflow either indirectly by virtue of their connections with hypothalamus or directly via descending projections to autonomic nuclei in the brain stem and spinal cord. These areas include the visceral cortices (prefrontal and insular cortex) and components of the limbic system such as the amygdala.

The major hypothalamic efferents to autonomic nuclei arise in the paraventricular nucleus and the lateral hypothalamic area. Neurons in both of these regions give rise to descending projections to preganglionic neurons in both the sympathetic and parasympathetic subdivisions of the spinal cord and brain stem. The paraventricular nucleus (PVN) is a particularly important integrative center for both neuroendocrine and autonomic regulation of homeostatic function. As noted earlier, parvicellular neurons in the PVN are major contributors to the neuroendocrine regulation of anterior pituitary secretion by virtue of their projections to the portal plexus in the median eminence. Additionally, magnocellular vasopressinergic and oxytocinergic neurons in the PVN and supraoptic nuclei project through the median eminence to terminate in the posterior lobe of the pituitary gland. These peptidergic systems are important for the regulation of fluid homeostasis and lactation and are distinct from the large numbers of PVN parvicellular neurons that give rise to descending projections to autonomic nuclei. The latter projections arise from phenotypically distinct PVN neurons sequestered within subfields of the nucleus that are devoted to autonomic function (dorsal parvicellular subdivision) or to both neuroendocrine and autonomic regulation (medial parvicellular subdivision). This juxtaposition of neurons in the PVN that contribute to homeostatic regulation via neuroendocrine or autonomic pathways makes efficient use of the sensory feedback signals that are relevant to both modes of regulation. Defining the means through which this sensory information is integrated in the PVN to influence endocrine and autonomic function remains an active and important area of research. The complexity of the integrative capacities of the PVN is illustrated in the following section on the neural control of feeding.

VI. HYPOTHALAMIC CONTROL OF FEEDING

Our understanding of how the hypothalamus contributes to the regulation of feeding behavior has advanced substantially during the past decade. Identification and functional dissection of phenotypically distinct populations of hypothalamic neurons that are now known to influence ingestive behavior and energy metabolism have contributed greatly to these advances. Recent identification of peripheral signaling molecules that act centrally to modulate these hypothalamic circuits has also provided tremendous insights into how the hypothalamus participates in energy homeostasis.

Results from classic lesion studies carried out in the 1940s and 1950s led to a ''dual-center'' model for the hypothalamic control of hunger and satiety that drove scientific research in this area for more than 30 years. The model was based on profound and consistent changes in ingestive behavior observed in rats after bilateral mechanical or electrolytic lesions centered in the ventromedial nucleus of the hypothalamus (VMH) or lateral hypothalamic area (LHA). Rats with VMH lesions displayed an apparently insatiable hunger and would become quite obese when food was made freely available. Conversely, rats with LHA lesions displayed adipsia and anorexia, and they required careful nursing during the postlesion period to prevent fatal dehydration and starvation. The dual-center model of hunger and satiety enjoyed strong appeal as an organizing framework for understanding how the hypothalamus contributes to the central control of ingestive behavior. However, the model ultimately was undermined by demonstrations that the method commonly used to create the VMH and LHA lesions produced effects that were not specific to hypothalamic regulation of ingestive behavior but extended to other sensorimotor and metabolic control systems.

Although the original dual-center model of hunger and satiety eventually fell out of favor, it remains generally accepted that certain subregions of the hypothalamus play a key role in energy homeostasis, at least in part by influencing ingestive behavior. Recent findings have served to renew scientific interest in the role of the LHA in controlling ingestive behavior

and energy balance. We now know that LHA neurons receive synaptic inputs from multiple sensory modalities relevant to ingestive control, and that their responses to these inputs vary as a function of nutritional status. Two distinct but partially overlapping populations of LHA neurons have been identified based on their content of unique neuropeptides (melanin-concentrating hormone and hypocretin/orexin) that appear to play important roles in the control of food intake and energy metabolism through distinct central neural pathways. LHA neurons that express these neuropeptides project to key areas of the brain stem and forebrain that mediate ingestive behavioral responses. Improved chemical methods for lesioning the diffusely scattered LHA neurons without damaging intermingled fibers of passage (including, most notably, the medial forebrain bundle) have demonstrated that the LHA is indeed necessary for normal feedback regulation of ingestive behavior.

Synaptic inputs to orexin/hypocretin- and MSH-containing neurons in the LHA arise from many regions of the central neuraxis. Perhaps most relevant to the control of ingestive behavior and energy balance is a recently discovered circuit that originates in the arcuate nucleus of the hypothalamus. Early studies indicated that neurotoxic damage to the arcuate nucleus (e.g., as produced by systemic administration of monosodium glutamate during early development) could produce syndromes of overeating and obesity in laboratory animals. New and exciting research findings have provided a possible explanation for this phenomenon. In 1994, a hormone called leptin was discovered that appears to exert tremendously potent inhibitory effects on feeding behavior. Leptin also produces significant physiological effects on thermoregulation and on the reproductive, thyroid, and adrenal axes. Leptin is released continuously from adipocytes and is present at a relatively high concentration in plasma (approximately 4 ng/ml). Leptin binds to receptors expressed in peripheral tissues and in a circumscribed set of brain regions, most notably the arcuate nucleus of the hypothalamus. Animals seem to adapt to their own unique circulating leptin levels (which are directly proportional to body adiposity), such that experimental perturbation of those levels will shift the animal's behavior and physiology toward either gaining or losing body weight in an apparent effort to restore normal leptin levels. For example, mice with genetic mutations that render them incapable of producing either leptin or leptin receptors become obese, whereas experimental elevation of circulating leptin levels causes animals to lose body weight. A wealth of recent data indicate that circulating leptin provides a physiologically important negative feedback signal to constrain body weight by affecting both ingestive behavior and metabolic activity.

There is general consensus that the arcuate nucleus plays a major role in transducing the effects of leptin on both the behavioral and physiological components of energy balance. Arcuate neurons that express functional leptin receptors project to the LHA and to other hypothalamic areas implicated in energy homeostasis, including the dorsomedial and PVN nuclei. Arcuate projections to the PVN provide a basis for documented leptin effects on parvicellular neurosecretory neurons comprising the central limbs of the HPA and hypothalamo–pituitary–thyroid axes, the activities of which are regulated by leptin. The arcuate nuclei and PVN also provide important descending projections to autonomic neurons in the brain stem and spinal cord, and may play a role in leptin-mediated increases in energy expenditure through the stimulation of brown fat thermogenesis. Leptin-sensitive arcuate neurons that project to the LHA and PVN contain neuropeptides (including neuropeptide Y, proopiomelanocortin, and agouti-related protein) that exert potent effects on food intake and other relevant neuroendocrine and autonomic aspects of energy homeostasis. Thus, the arcuate nucleus can be viewed as a nodal point in the hypothalamic regulation of energy balance.

VII. CONCLUSIONS

There can be little doubt that our understanding of hypothalamic organization and function has improved dramatically during the past 25 years. Many of these advances can be traced directly to technical advances that have allowed the dissection of hypothalamic circuits from both organizational and functional perspectives. Thus, the ability to define the connectivity and phenotype of hypothalamic neurons and to assess the activity of anatomically defined circuits with functional probes such as Fos has proven to be enormously informative in defining the function of hypothalamic cell groups. Continued application of these experimental approaches promises to further illuminate our understanding of this small subdivision of the diencephalon that has long been recognized as a central integrative center for the control of homeostasis and behavioral state.

See Also the Following Articles

AROUSAL • CHEMICAL NEUROANATOMY • CIRCADIAN RHYTHMS • HOMEOSTATIC MECHANISMS • NERVOUS SYSTEM, ORGANIZATION OF • PSYCHONEUROENDOCRINOLOGY • SLEEP DISORDERS • TIME PASSAGE, NEURAL SUBSTRATES

Suggested Reading

Card, J. P., Swanson, L. W., and Moore, R. Y. (1999). The hypothalamus: An overview of regulatory systems. In *Fundamental Neuroscience* (M. Zigmond, F. E. Bloom, S. C. Landis, J. L. Roberts, and L. R. Squire, Eds.), pp. 1013–1026. Academic Press, San Diego.

Iversen, S., Iversen, L., and Saper, C. B. (2000). The autonomic nervous system and the hypothalamus. In *Principles of Neural Science* (E. R. Kandel, J. H. Schwartz, and T. M. Jessel, Eds.), 4th ed., pp. 960–981. McGraw-Hill, New York.

Parent, A. (1997). Hypothalamus. In *Carpenter's Human Neuroanatomy* 9th ed., pp. 706–743. Willams & Wilkins, Baltimore.

Saper, C. B. (1990). Hypothalamus. In *The Human Nervous System* (G. Paxinos, Ed.), pp. 389–414. Academic Press, San Diego.

Sawchenko, P. E. (1998). Toward a new neurobiology of energy balance, appetite, and obesity. The anatomists weigh in. *J. Comp. Neurol.* **402**, 435–441.

Imaging: Brain Mapping Methods

JOHN C. MAZZIOTTA
University of California, Los Angeles School of Medicine

RICHARD S. J. FRACKOWIAK
University College, London

Disorders of the human nervous system are among the most debilitating and devastating of all human illnesses. Such disorders not only affect the physical abilities of patients but often severely compromise the quality of life, the ability to function in society, family relations, and the ability to maintain gainful employment. As such, neurological, neurosurgical, and psychiatric disorders affect not only the patients but also their families and society at large because of the tremendous economic burden that results from their prevalence. Undoubtedly, this is one of the reasons brain mapping methods have advanced so rapidly and hold such promise for improved diagnostics, monitoring of therapeutics, and providing insights into the basic mechanisms of brain disease. Never before have so many methods been available to tackle these vexing problems. This article provides an overview of the methods and their applications in human diseases.

Throughout this article, it is important for the reader to keep in mind a number of important physical and physiological factors when trying to understand the approaches of different brain mapping applications to the study of human cerebral disorders (Table I).

First, it is important to keep in mind that specific methods address only one or a few aspects of underlying cerebral physiology and pathophysiology. For example, electromagnetophysiological techniques, such as electroencephalography (EEG), event-related potentials (ERPs), and magnetoencephalography (MEG), provide information about large constellations of neurons and the net electromagnetophysiological vector that their firing produces. The electrical techniques are weighted toward surface structures, whereas the magnetic ones convey information about deeper brain structures with less distortion. This indicates another important aspect in assessing methods for the evaluation of disease states. That is, certain techniques are better suited to examination of particular sites in the brain than others.

Second, a number of techniques are devoted specifically to describing brain structure. These include X-ray computed tomography (CT), conventional magnetic resonance imaging (MRI), and blood vessel imaging using conventional angiography, magnetic resonance angiography (MRA), or helical CT. A number of techniques assess hemodynamic responses as a measure of function. These techniques include xenon-enhanced CT, functional MRI (fMRI), perfusion MRI, and cerebral blood flow or blood volume measurements using positron emission tomography (PET) or single photon emission computed tomography (SPECT). All the techniques that evaluate cerebral

This article has been reprinted from Mazziotta and Frackowiak (2000). The study of human disease with brain mapping methods. In *Brain Mapping: The Disorders* (Toga and Mazziotta, eds.) pp. 3–31. Academic Press, San Diego.

Table I

Advantages and Limitations of Current Brain Mapping Techniques of Use in the Study of Patients with Neurological, Neurosurgical, and Psychiatric Disorders

Method	Advantages	Limitations
X-ray computed tomography	Excellent bone imaging	Ionizing radiation
	~100% detection of hemorrhages	Poor contrast resolution
	Short study time	
	Can scan patients with ancillary equipment	
	Can scan patients with metal devices/ electronic devices	
Magnetic resonance imaging	High spatial resolution	Long study duration
	No ionizing radiation	Patients may be claustrophobic
	High resolution	Electronic devices contraindicated
	High gray–white contrast	Acute hemorrhages problematic
	No bone-generated artifact in posterior fossa	Relative measurements only
	Can also perform chemical, functional, and angiographic imaging	
Positron emission tomography	Can perform hemodynamic, chemical, and functional imaging	Ionizing radiation High initial costs
	Quantifiable results	Long development time for new tracers
	Absolute physiologic variables can be determined	Limited access
	Uniform spatial resolution	Low temporal resolution
Single photon emission computed tomography	Can perform hemodynamic, chemical, and functional imaging	Ionizing radiation Relative measurements only
	Widely available	Nonuniform spatial resolution
		Low temporal resolution
Xenon-enhanced computed tomography	Uses existing equipment	Ionizing radiation
		High xenon concentrations have pharmacologic effects
Helical computed tomography (CT angiography)	Provides high-resolution vascular images	Ionizing radiation
		Vascular and bony anatomy only
Electroencephalography	No ionizing radiation	Low spatial resolution
	High temporal resolution	Weighted toward surface measurements
	Widely available	
	Can identify epileptic foci	
Magnetoencephalography	No ionizing radiation	Low spatial resolution
	High temporal resolution	
	Can identify epileptic foci	
Transcranial magnetic stimulation	No ionizing radiation	Low spatial resolution
	Potential for therapy	Has produced seizures in certain patient groups
	Can be linked to other imaging methods (PET, MRI)	
Optical intrinsic signal imaging	No ionizing radiation	Complex signal source
	High temporal resolution	Invasive only (intraoperative)
	High spatial resolution	

function based on hemodynamic measurements are, by their very nature, at a physiological "distance" from the actual neuronal event. These methods assume that neuronal firing and blood flow increments or decrements are tightly coupled. In most cases, this holds true in the normal brain, but it may not always be true in pathologic states.

Third, the determination of chemical processes in the brain falls mainly in the domain of PET and magnetic resonance spectroscopy (MRS). The former can measure cerebral glucose metabolism, protein synthesis, amino acid uptake, pH, and other variables and does so in a quantitative manner reported in physiological units when appropriate rate constants and other factors are incorporated into mathematical models for their estimation. MRS provides relative measurements of chemical compounds relying primarily on hydrogen spectra but, at higher magnetic field strengths, can also estimate relative quantities of sodium fluorine-carbon, and phosphorus-containing molecules as well.

Fourth, the evaluation of receptor systems in the brain, both transmitter molecules and receptor complexes, has been an active area of research in both health and disease using PET and SPECT. Most information has been derived for the dopaminergic system but data also exist for the cholinergic, serotonergic, opioid, and benzodiazepine systems.

Last, there are interactive approaches. The most time honored, of course, is the direct observation of signs and symptoms in patients with cerebral disorders. Such information can be obtained in the traditional clinical setting or in the highly unusual circumstance of awake surgical procedures where recording from or stimulation of cerebral tissue can be correlated with behavioral states in a conscious patient. These latter methods, while in use for more than 50 years, still provide unique information about structure–function relationships in the human brain. Recently, such measurements have been augmented through the use of optical intrinsic signal imaging, in which changes in optical reflectance from the cortex are measured during surgery. Measurements can be made either in awake patients during behavior or with anesthetized patients receiving sensory stimulation or direct peripheral nerve stimulation. They provide measures of functional specificity for different brain regions in the operative field. The technique of transcranial magnetic stimulation allows the investigator to stimulate the cortex of the brain magnetically, resulting in an induced electrical discharge in the cortex and an observed or reported behavior from the subject. This technique has been used experimentally to map normal brain systems and in patients for experimental, diagnostic, and therapeutic purposes.

Each technique is capable of making measurements with a characteristic resolution in both the spatial and temporal domains. Tomographic imaging techniques produce the highest spatial resolution currently available, whereas the electromagnetophysiological methods provide the highest temporal resolution. Although knowledge of resolution is important, it must be matched to the question of interest in a particular patient population—a decision that also requires knowledge of the sampling characteristics of the technique. This latter term refers to the volume of brain tissue that can be assessed with a particular measurement. Thus, whereas fMRI may survey functional responses through out the entire brain, electrophysiological measurements from a depth electrode, despite producing data with exquisite spatial and temporal resolution, sample only a very small volume of brain tissue. As such, investigations aimed at trying to identify a functional disorder in a disease of unknown etiology (e.g., autism) would be better done with a global technique that surveys the entire cerebral landscape rather than with measurement at multiple sites with an electrophysiological method such as depth electrodes. The latter are better used to understand the local electrophysiology of a site that has a high probability of being abnormal and possibly also causative of a given disorder.

With these principles in mind, we now consider a more specific examination of the different strategies that can be employed with modern brain mapping methods to assess patients, either as individuals or as groups, with neurological, neurosurgical, or psychiatric disorders (Table II).

I. DIAGNOSTIC METHODS

A. Individual Processes

1. Structural Anatomy

a. Computed Tomography X-ray CT was the first noninvasive imaging technique that allowed for direct visulization of the brain parenchyma. It revolutionized the evaluation of patients with neurological and neurosurgical disorders because it could image bone and provided the first opportunity to see the brain directly. It has a small dynamic contrast range so that

Table II

Brain Mapping Methods of Use in the Study of Human Disease along with the Types of Measurements They Provide and Some of the Clinical Situations in which They May Be of Use

Method	Measurements provided	Disorders
X-ray computed tomography (CT)	Brain structure	Acute/chronic hemorrhages
	Blood–brain barrier integrity	Acute trauma
		General screening of anatomy
		Focal or generalized atrophy
		Hydrocephalus
Magnetic resonance imaging	Brain structure	Acute ischemia
	Brain and cervical vasculature	Neoplasms
	Relative cerebral perfusion	Demyelinating disease
	Chemical concentrations	Epileptic foci
	Fiber tracts	Degenerative disorders
	Blood–brain barrier integrity	Infections
		Preoperative mapping
Positron emission tomography	Perfusion	Ischemic states
	Metabolism	Degenerative disorders
	Substrate extraction	Epilepsy
	Protein synthesis	Movement disorders
	Neurotransmitter integrity	Affective disorders
	Receptor binding	Neoplasms
	Blood–brain barrier integrity	Addictive states
		Preoperative mapping
Single photon emission computed tomography	Perfusion	Ischemic states
	Neurotransmitter integrity	Degenerative disorders
	Receptor binding	Epilepsy
	Blood–brain barrier integrity	Movement disorders
Xenon-enhanced computed tomography	Perfusion	Ischemic states
Helical computed tomography (CT angiography)	Vascular anatomy	Vascular occulsive disease
	Bony anatomy	Aneurysms
		Arteriovenous malformations
Electroencephalography	Electrophysiology	Epilepsy
		Encephalopathics
		Degenerative disorders
		Preoperative mapping
Magnetoencephalography	Electrophysiology	Epilepsy
Transcranial magnetic stimulation	Focal brain activation	Preoperative mapping
Optical intrinsic signal imaging	Integrated measure of blood volume, metabolism, and cell swelling	Intraoperative mapping

differentiation of gray and white matter is difficult (Fig. 1A). However, it is very sensitive for identifying cerebral hemorrhage and also lesions associated with an alteration in the blood–brain barrier, by virtue of leakage of iodinated contrast material in them. These abilities make X-ray CT ideal for direct and immediate assessment of patients with cerebral hemorrhage, multiple sclerosis, brain tumors, and traumatic

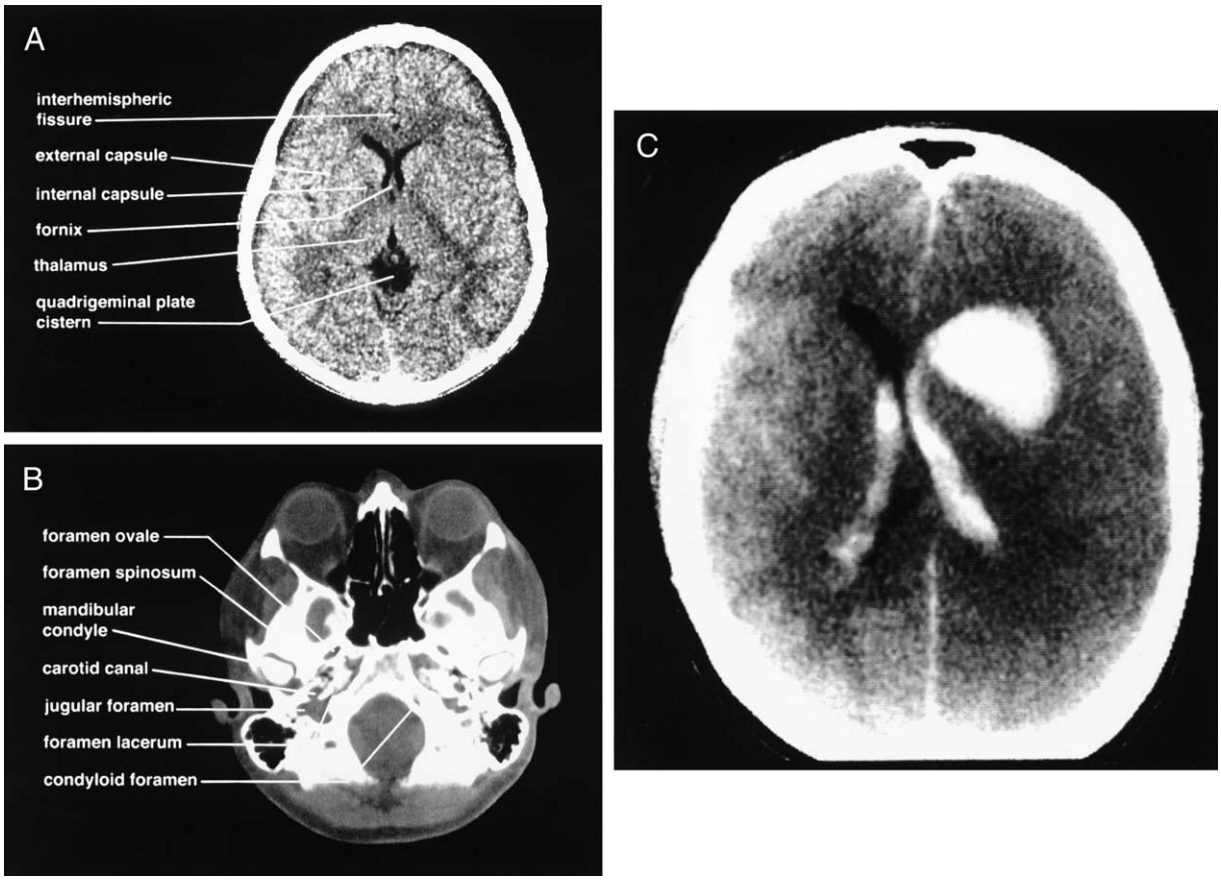


Figure 1 X-ray CT. (A) Images of the human brain from an X-ray CT device, demonstrating good anatomical detail, particularly of the skull and ventricular system as well as the subarachnoid CSF spaces. Note that there is less gray–white contrast than in MRI images (Fig. 2A). (B) X-ray CT provided very detailed images of bony structures that surround the central nervous system. This is particularly useful in evaluating pathologic states at the base of the skull, where conventional radiography is often difficult because of patient positioning and the overlap of bony structures in a two-dimensional radiograph. Furthermore, in situations in which trauma is a factor, often patients cannot be manipulated easily because of the possibility of fractures at the base of the skull or in the cervical spine. (C) Intracerebral hemorrhage demonstrated by X-ray CT. Sensitivity for detection of intracranial bleeds is effectively 100% with X-ray CT and it remains the imaging modality of choice in acute patients when identification of cerebral hemorrhage is urgent and important. This is typically the case in patients with an acute cerebral deficit when cerebral hemorrhage must be identified if thrombolytic or anticoagulant therapy is being contemplated.

injuries. Contrast sensitivity and the time needed for scanning have improved since X-ray CT was introduced, but with the advent of MRI technology many diagnostic studies that were formerly in the province of X-ray CT are now done with MRI. Nevertheless, X-ray CT continues to have an important role in resolving certain diagnostic questions and in particular patient circumstances.

X-ray CT has remained the imaging modality of choice for patients requiring urgent evaluation of suspected intracranial hemorrhage (Fig. 1C) and in patients with acute head trauma. In both these circumstances, the speed of the study and ease of patient access as well as availability of equipment are well matched to the ability of CT to evaluate such patients. X-ray CT is also the procedure of choice when evaluating abnormalities of bony structures of the head particularly the skull base (Fig. 1B). Lastly, patients who cannot tolerate MRI because of claustrophobia or implanted ferromagnetic or electronic devices or that need to be attached to ancillary equipment such as is frequently encountered in critical care situations are also best scanned by X-ray CT if structural information is needed for clinical evaluation.

b. MRI Magnetic resonance imaging is the structural imaging modality of choice in all other situations. Its superior spatial resolution and contrast range, particularly useful in differentiating gray and white matter, are but two features of MRI that make it

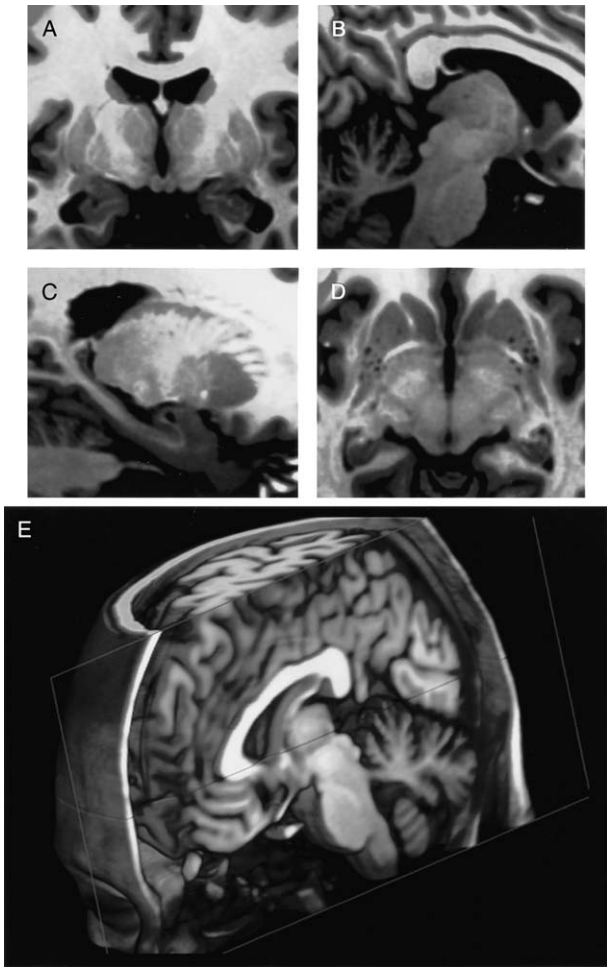


Figure 2 Magnetic resonance imaging. (A–D) Typical two-dimensional MRI images of the brain. Notice that the detailed anatomy of the brain parenchyma has better gray–white contrast than X-ray CT images (Fig. 1A). Notice also that there are none of the typical CT artifacts caused by the juxtaposition of dense bone and brain parenchymay. (A) Coronal view through the thalamus demonstrating the subnuclei of the thalamus, the mamillary bodies, and the internal, external, and extreme capsules as well as the two segments of the globus pallidus. Also note the detailed anatomy of the hippocampi. (B) Sagittal view demonstrating the colliculi of the midbrain, the midline of the thalamus, and the detailed anatomy of the midsagittal region of the cerebellum. (C) Sagittal view through the hippocampus and striatum. Note the fine bridges of gray matter between the caudate and the putamen. (D) Transverse section through the basal ganglia and upper midbrain. Note the periaqueductal gray matter, the detailed anatomy of the hippocampi, and the bilateral flow voids produced by the presence of the lenticulostriate arteries in the posterior portion of the putamen. (E) Three-dimensional reconstruction with cutaway of an MRI data set demonstrating the kind of anatomical detail that can be provided with three-dimensional MR images (courtesy of Colin Holmes and colleagues, UCLA School of Medicine, Los Angeles). [A–D from Holmes *et al.* (1998). Enhancement of magnetic resonance images using registration for signal averaging. *J. Computer Assisted Tomogr.* **22**(1), 139–152].

superior to CT for structural imaging (Fig. 2). The ability to image brain structures from any angle and avoidance of CT artifacts that result from soft tissue-dense bone boundaries in the field of view provide further arguments for the use of MRI in patients with cerebral disorders. This advantage is most notable in the posterior fossa, where artifacts from the dense petrous bones often obscure or obliterate relevant clinical information about the brain stem and cerebellum. Gadolinium and other paramagnetic contrast agents can be used with MRI to provide the ability to detect blood–brain barrier defects in a manner analogous to that used in X-ray CT with iodinated contrast agents.

2. Vascular Anatomy

a. MR Angiography The observation that protons leaving the field of view reduce the local signal in MRI studies has led to an entire field of MR angiography and associated flow-based MRI techniques. The so-called flow void occurs in the vascular system when blood that encounters a radiofrequency pulse leaves the field of view of the scanner and the resultant energy is therefore emitted outside the field of view. The result is a loss of signal within the lumina of blood vessels. When partitular pulse sequences are utilized to optimize this effect, an image of vascular anatomy results. Like most angiographic procedures, the image depicts the contents of the blood vessel (within the lumen) as opposed to the blood vessel wall and associated structures. However, unlike conventional angiography, in which arterial, capillary, and venous phases are distributed in time and images of each phase can be produced independently, MRA provides a composite image of all medium-to large-diameter vessels, including arteries and viens.

Such studies have provided important opportunities for the evaluation of intracranial and cervical, medium and large vessels for abnormalities, including arteriovenous malformations, aneurysms, and occlusive disease. Smaller caliber vessels still require conventional angiographic or helical CT evaluation for the assessment of disorders such as vasculitis.

b. Helical CT In this technique, also called spiral CT or CT angiography, conventional CT technology is modified to produce very rapid sequential images of the head by having the relationship between the patient and the X-ray tube/detector system traverse a helical course through the tissue. This process is rapid and so

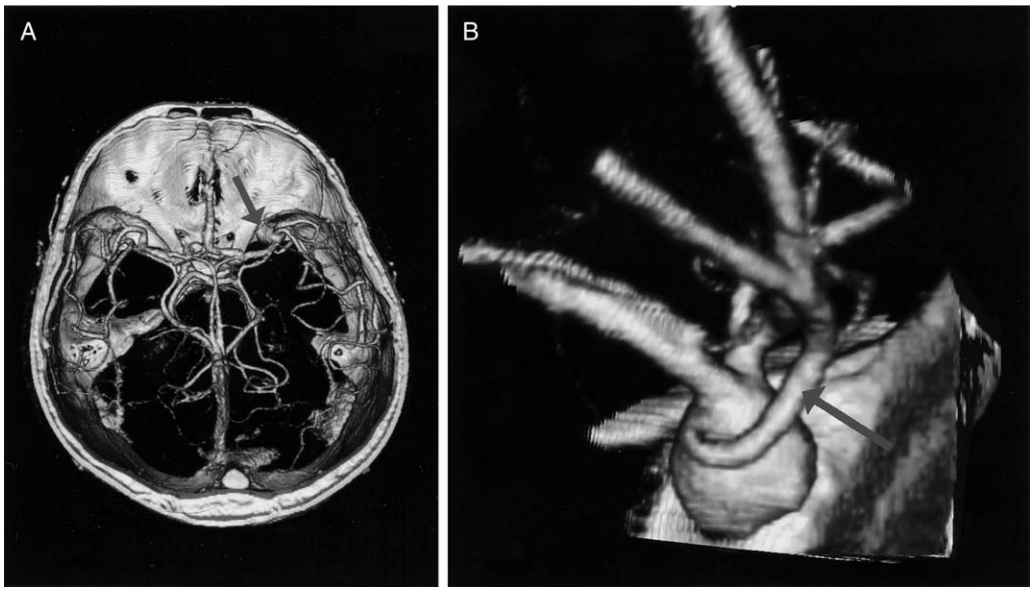


Figure 3 Helical CT angiography. (A) The cerebral vasculature is superimposed on this cutaway view of the skull seen from above. Note the line detail provided for the vascular structures, including the circle of Willis, the anterior, middle, and posterior cerebral arteries, and many of their branches. The arrow indicates an aneurysm arising from the middle cerebral artery at the anterior edge of the middle cerebral fossa. (B) Close-up view of the aneurysm demonstrated in A. Note that in this three-dimensional reconstruction, it is possible to see all of the blood vessels that contribute to or arise from the aneurysmal sack (left). Unlike conventional angiography, where the overlap of the aneurysm and the parent or daughter vessels can be obscured because of the two-dimensional projection required in this technique, full three-dimensional images are possible with helical CT angiography. By digital reconstruction and rotation of the data set from helical CT, these complex and important relationships can be evaluated prior to surgery, whereas with conventional angiograms multiple view would be required and still may not provide a sufficiently detailed view of these relationships. In addition, multiple views obtained with conventional angiography expose the patient to additional radiation and contrast risks. In this case, the angular artery (arrow) arises from the aneurysm sack. Endovascular coil placement might lead to obstruction of this important vessel, thereby making such a patient a candidate for surgical rather than endovascular treatment of this lesion (courtesy of Pablo Villablanca, UCLA School of Medicine).

arranged that it is coincident with the delivery of a bolus of iodinated contrast material into the cranial and cervical vessels from a peripheral vein. The resultant images are high-resolution depictions of the intracranial anatomy that can be reconstructed in three dimensions (Fig. 3A). Currently, limited information is available about the technique in terms of its clinical applications, but it is likely that there are circumstances in which it may be superior to conventional angiography. One of those is in the assessment of the local vascular anatomy of patients with aneurysms and arteriovenous malformations. In this case, conventional angiography, while high in spatial resolution, collapses the three-dimensional structure of such lesions into a two-dimensional projection. As such, the important relationship between aneurysm neck and a parent or daughter vessel may be difficult to ascertain or may require multiple intraarterial contrast injections and radiation exposure for the patient. Helical CT allows for a true three-dimensional reconstruction of local anatomy that can be manipulated to assess such relationships in greater detail, while subjecting the patient to only a single radiation exposure and dose of iodinated contrast material (Fig. 3B). A similar situation exists when defining feeder vessels to arteriovenous malformations.

3. Blood Flow and Perfusion

a. PET The assessment of cerebral perfusion (i.e., cerebral blood flow per volume of tissue) can be determined quantitatively with PET using ^{15}O-labeled water, ^{11}C-labeled butanol, and potentially other agents. These agents are freely diffusible and knowledge of the time–activity relationship of the tracer compound in arterial blood and the tissue concentration over time in the brain permits a calculation of cerebral perfusion with approximately $5 \times 5 \times 5$-mm spatial resolution. These methods have been used extensively to evaluate increments in perfusion associated with underlying neuronal activity such as that

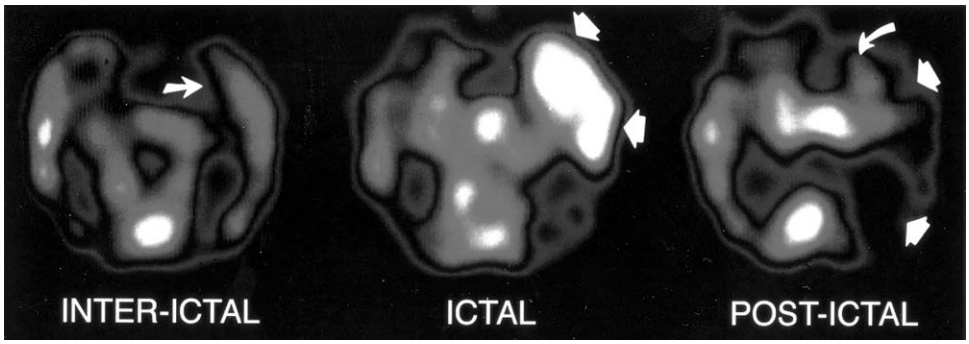


Figure 4 SPECT studies of a seizure focus in a patient with epilepsy, obtained using technetium-^{99}m-labeled HMPAO. By comparing the ictal study (hyperperfusion), obtained by injecting the tracer on a telemetry ward, with the postictal and interictal scans (hypoperfusion), it is possible, with a high degree of accuracy, to identify seizure foci responsible for focal epilepsies. Although such SPECT studies of relative perfusion may be confusing or misleading when evaluated in only one of these states, the composite information obtained from ictal, postictal, and interictal studies, particularly when the studies are aligned, registered, and subtracted, is useful in the presurgical evaluation of patients with focal or complex partial epilepsy [courtesy of Sam Berkovic, Austin Hospital, Melbourne, Victoria, Australia. From Berkovic, S. F., Newton, M. R., and Rowe, C. C. (1991). Localization of epileptic foci using SPECT. In *Epilepsy Surgery* (H. Luden, Ed.), pp. 251–256. Raven Press, New York].

associated with the performance of behavioral tasks, and as such can be used in patients to evaluate critical cortical and subcortical areas as part of the process of mapping brain regions adjacent to abnormalities that are under consideration for surgical resection.

Perfusion measurements are also of value in assessing patients with ischemic cerebrovascular disease not only to determine baseline conditions but also in the assessment of "cerebral perfusion reserve" by challenging such patients with cerebral vasodilation drugs such as acetazolamide.

b. SPECT Perfusion measurements with SPECT can be obtained in a semiquantitative way using xenon-133 or in a relative fashion using a host of technetium-99 m labeled agents (e.g., HMPAO). Studies have been used to evaluate patients with cerebrovascular disease in the assessment of patients with epilepsy. In the latter circumstance, ictal, postictal, and interictal studies are made using these agents and subtracted to identify foci of epileptic discharges that increase cerebral perfusion during seizures and reduce perfusion postictally and interictally (Fig. 4).

c. Xenon-Enhanced CT Nonradioactive xenon-133 alters the X-ray attenuation characteristics of the tissues that absorb it. Xenon-133 can be readily inhaled. Using strategies similar to those discussed previously, one can calculate changes in X-ray attenuation associated with the concentration of xenon in tissue to calculate an index of cerebral perfusion. This approach has been used to evaluate patients with cerebrovascular disease and head trauma. One confounding factor with this approach is that relatively high concentrations of xenon are needed in brain tissue to produce accurate perfusion estimates. At these concentrations, patients experience pharmacological effects, including sedation and, ultimately, anesthesia. These direct neuronal effects of xenon can, in turn, affect cerebral perfusion and contaminate physiological measurements. Nevertheless, such studies have been used to assess patients in specific diagnostic categories when alternate techniques are not available.

d. MRI Relative and semiquantitative measurements of cerebral perfusion can be obtained with MRI using a variety of techniques. The basic difference among these techniques is whether exogenous agents are infused into the patient or whether changes in endogenous signals from the vascular system are monitored. With the former, bolus contrast agents such as gadolinium are delivered to a subject intravenously. Their local concentration in the cerebral vasculature estimates local cerebral perfusion, which in turn is proportional to neuronal firing rates. The change in MR signal, induced by a local change in concentration of a contrast agent, reflects a relative change in local perfusion that can be estimated. A refinement of this approach requires information about the concentration of the contrast agent in the blood over a period of time. This can be measured directly from an arterial source, although such a method is rarely employed, or from a large-diameter

blood vessel in the scanner's field of view (e.g., arteries in the circle of Willis).

Alternatively, one can record endogenous local signal changes associated with alterations in cerebral perfusion induced by changing neuronal activity. When local neuronal firing increases, there is an increment in cerebral blood flow to an area. However, there is a proportionally smaller change in local tissue oxygen metabolism. As a result, more oxygen is delivered per unit volume of tissue (because of decreased fractional extraction) but there is little change in the amount of oxygen used. Thus, the oxygen content (i.e., the concentration of oxyhemoglobin) in venous blood increases. When compared to images in a state of ''baseline'' neuronal firing, this local change in venous oxyhemoglobin results in an alteration of the MR signal since deoxygenated blood is more paramagnetic than oxygenated blood (the difference is about 0.2 ppm). The change in signal provides an estimate of relative local cerebral perfusion changes associated with the change in underlying neuronal activity. This observation has resulted in the widespread use of the so-called blood oxygen level dependent technique in the evaluation of normal subjects performing behavioral tasks and in patients who are candidates for surgical resection of brain lesions such as tumors, vascular malformations, or epileptic foci. Such methods may ultimately supplant more invasive approaches to determining language dominance and memory function such as those employing the intraarterial injection of barbiturate compounds (e.g., Wada testing).

Diffusion imaging is also possible using MR scanning. Water, and hence protons, is freely diffusible. In the brain, however, the microscopic anatomy of the tissue compartmentalizes (with cells, along axons, etc.) this process. It is possible to obtain MRI images that are diffusion weighted (DW) and in which the signal is dependent on the ease with which protons diffuse in their local environment. Initially tested in cats, DW images were shown to demonstrate the boundaries of acute infarcts within minutes. Infarcted or ischemic areas are visible as regions of hyperintensity corresponding to local decreases in the apparent diffusion coefficient of water. The use of DWI in screening patients with early ischemia who are candidates for thrombolytic therapy is especially relevant (Fig. 5). When combined with MRI perfusion imaging, an assessment of the proportion of jeopardized and infarcted tissue can be made.

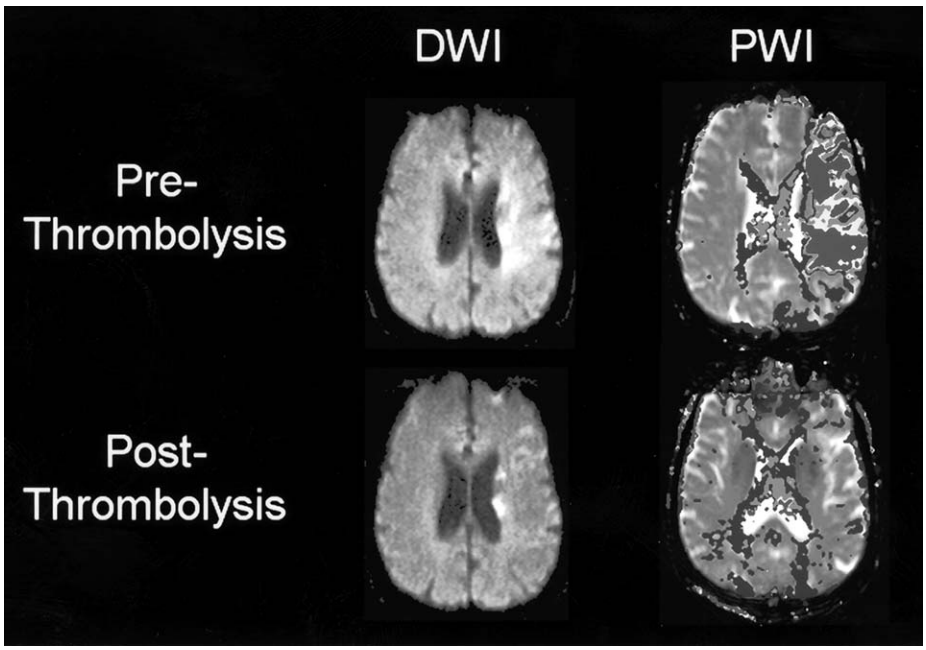


Figure 5 Diffusion-weighted (DWI) and perfusion-weighted (PWI) images from MRI are useful in selecting patients for cerebrovascular intervention therapies and monitoring their outcome. Here, a patient with a large region of hypoperfusion in the distribution of the middle cerebral artery already has some evidence of tissue ischemia as demonstrated by the white areas that border the ventricle in the prethrombolysis DW image (top left). Following thrombolysis (bottom row), note that perfusion has been reestablished to the area of the middle cerebral artery and the patient is left with only a small ischemia injury on the posterior border of the lateral ventricle (courtesy of Chelsea Kidwell, Jeffrey Saver, and Jeffrey Alger, UCLA School of Medicine).

4. Metabolism

a. PET Glucose and oxygen metabolism can be assessed using PET and ^{18}F-labeled fluorodeoxyglucose (FDG), ^{11}C-labeled glucose, and ^{15}O-labeled oxygen. A common approach has been to use FDG to evaluate glucose metabolism. Thought to be primarily an assessment of synaptic activity (i.e., deoxyglucose uptake is maximal in the neuropil rather than in regions dominated by cell bodies), scanning of patients using FDG has provided useful observations in a wide range of disorders. Hypometabolic regions are found interictally at epileptic foci and in specific cortical and subcortical regions in a wide range of degenerative and dementing processes and also appear to reflect the malignancy grade of cerebral neoplasms, particularly gliomas (Figs. 6–8). Oxygen metabolism and extraction measurements are of greatest utility in assessing patients with cerebrovascular disease, particularly that unique subset of patients for which extracranial–intracranial bypass or other surgical reperfusion approaches are contemplated.

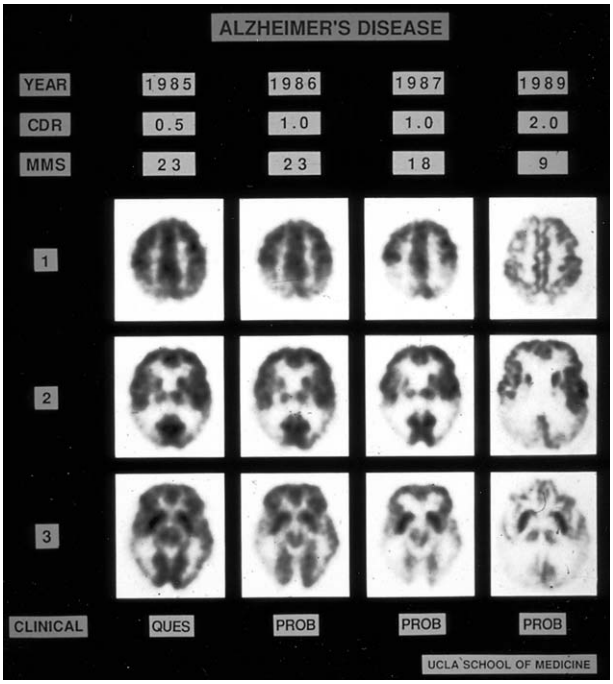


Figure 6 PET studies of cerebral glucose metabolism in a patient with Alzheimer's disease demonstrating the characteristic pattern of hypometabolism that occurs in the parietal and superior temporal cortices. As the disease progresses (from left to right), hypometabolism worsens in both spatial extent and magnitude, ultimately involving all neocortical structures.

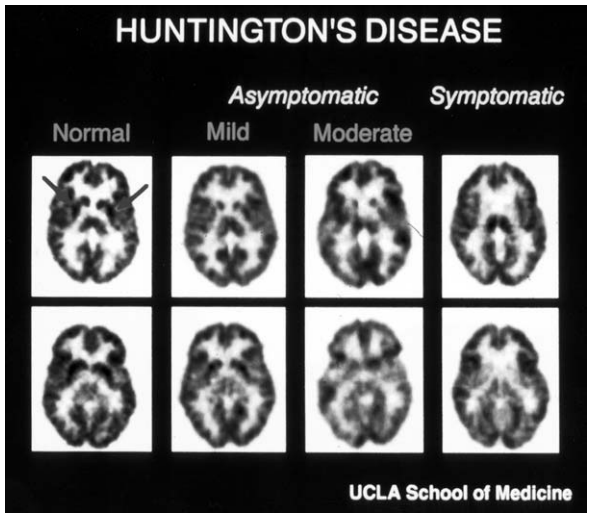


Figure 7 Glucose metabolism measured with PET in Huntington's disease. Two anatomical levels (rows) of a normal subject demonstrating glucose metabolism in the caudate and putamen as compared with three patients, two with mild or moderate asymptomatic Huntington's disease and one with frank symptoms. Note that in the symptomatic patient there is profound hypometabolism of the caudate and most of the putamen, whereas in the asymptomatic, gene-positive subjects, there is progressive loss of metabolism in the caudate. Such changes can be identified 5–7 years prior to the onset of symptoms in patients who carry an expanded triplet repeat sequence of the Huntington's disease gene. With advancing disease, hypometabolism extends throughout the striatum and can also involve the frontal cortices (courtesy of Scott Grafton and John Mazziotta, UCLA School of Medicine).

b. MR Spectroscopy Concentrations of a wide variety of compounds can be estimated using MR spectroscopy. Proton spectra provide information about lactate as a measure of carbohydrate metabolism. Other chemical species as well as spectra obtained from other isotopes can provide additional clues about various chemical pathways. In addition, "tracer" techniques can be employed, e.g., using $^{17}O_2$, ^{13}C-labeled glucose, analogous to those used with PET or SPECT.

5. Ligands and Neuroreceptor Imaging

a. PET Ligands have been developed for PET to image both pre and postsynaptic neuroreceptors. Since one or both sides of a synapse can be affected by neuropsychiatric disease, the ability to image the integrity of both components of synaptic structure and function is useful. There are a large number of neurotransmitter systems in the human brain but

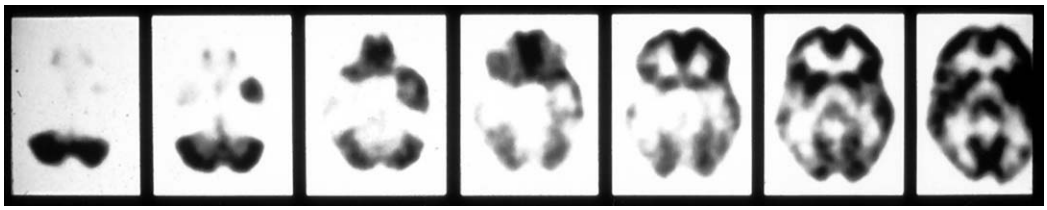


Figure 8 Interictal hypometabolism of the temporal lobe in a patient with complex partial epilepsy referable to that structure. Note the profound asymmetry of metabolism, particularly in the medical but also in the lateral portion of the temporal lobe, in six out of the seven axial images. Such studies are typically performed in the evaluation of patients who are candidates for surgical treatment of their epilepsy. In many patients, brain mapping techniques have obviated the requirement for depth electrodes and subdural grids in the presurgical evaluation of such individuals (courtesy of Jerome Engel, Jr., *et al*., UCLA School of Medicine; J. Engel, P. H. Crandall, and R. Rausch (1983). The partial epilepsies. In *Clinical Neurosciences* (R. N. Rosenberg, Ed.), pp. 1349–1380. Churchill Livingstone, New York).

PET ligands have been developed for only some of these. Fewer still have been completely validated and are in clinical use.

The most extensively studied neurochemical system is the dopaminergic network. Presynaptic imaging of dopamine synthesis and reuptake have been evaluated with fluorinated (^{18}F) ligands. The uptake, metabolism, and flux of ^{18}F-labeled L-DOPA in presynaptic dopaminergic terminals have been used to evaluate movement disorders, particularly Parkinson's disease (Fig. 9), and a number of neuropsychiatric syndromes. The same is true for the evaluation of presynaptic dopaminergic reuptake sites (e.g., with the WIN compounds). On the postsynaptic side, numerous ligands have been developed with a range of affinities for the different postsynaptic dopaminergic receptors. These tracers include raclopride, spiperone, and ethylspiperone. They have provided data for the movement disorders, schizophrenia, pituitary tumors, and other disorders of the brain. In addition, the labeling of drugs of abuse, such as ^{11}C-labeled cocaine, that bind to receptors in the dopaminergic system has

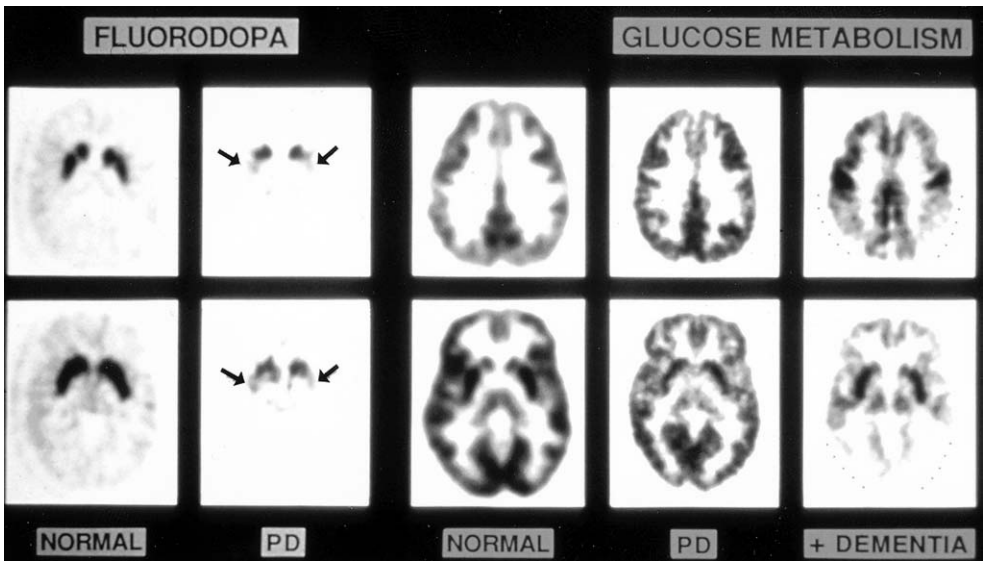


Figure 9 Evaluation of patients with Parkinson's disease using PET. Glucose metabolism in Parkinson's disease demonstrates a normal pattern in such patients (PD label) when compared to a normal control (middle column). When the dopamine-specific ligand [^{18}F]fluoro-L-DOPA is employed, however, a striking reduction in the uptake of this tracer is identified in the posterior putamen of patients with Parkinson's disease when compared to normals (columns 1 and 2). This is because the majority of the presynaptic dopaminergic terminals in Parkinson's disease are lost from the putamen at the onset of Parkinson's disease symptoms. Nevertheless, this population of synapses represents only a minority of all the synapses in this structure. As such, glucose metabolism remains normal but the uptake of fluoro-L-DOPA is dramatically reduced since this tracer images only that subpopulation of cells that have dopamine as their neurotransmitter. In patients with Parkinson's disease plus dementia, some patients also show hypometabolism of neocortex (+ DEMENTIA) similar to patients with Alzheimer's disease (see Fig. 6).

been helpful as a means of exploring the neurobiology of chemical addiction.

Cholinergic tracers are in use for the study of neurodegenerative disorders such as Alzheimer's disease and markers of the serotonin system and have been employed in the evaluation of psychiatric disorders. The central benzodiazepine system can be scanned with flumazenil labeled with carbon-11. Important findings have been made with this tracer in patients with epilepsy (Fig. 10A). Finally, opioid receptors have been studied with a host of ligands that bind with varying affinities to the many subclasses of opiate receptor. These compounds have been used in the exploration of eiplepsy, psychiatric disease, and movement disorders.

b. SPECT A number of iodinated compounds have been developed and serve as analogs of the PET tracers described previously. Used in a similar fashion, relative estimates of binding and receptor uptake can be obtained with SPECT (Fig. 10B). Although the number of compounds of this type is few compared to the inventory of PET ligands, interest in their development and use will result in an ever-increasing set of SPECT tracers.

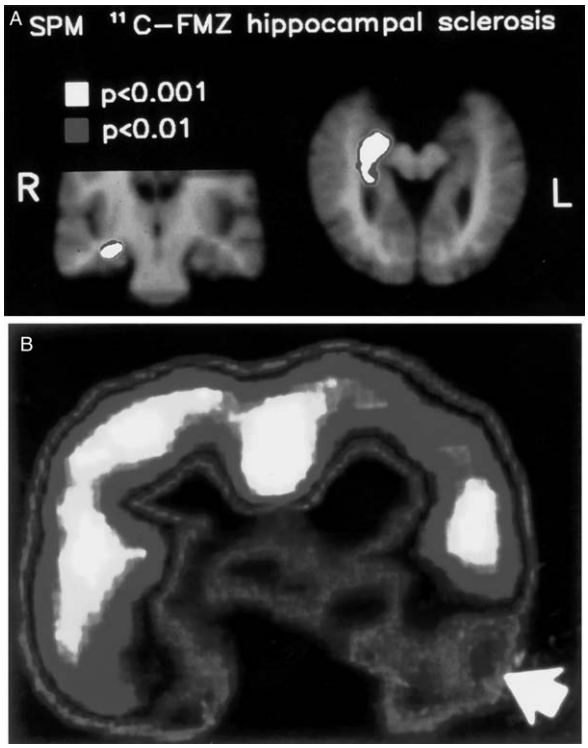


Figure 10 Benzodiazepine receptor imaging in patients with focal epilepsy. (A) PET evaluation with [^{11}C]flumazenil. In patients with focal temporal lobe epilepsy from hippocampal sclerosis producing complex partial seizures, flumazenil uptake is reduced in the medial temporal lobe. These changes are typically smaller in spatial extent than hypometabolism detected with FDG–PET. Comparison of the two studies may ultimately lead to more selective surgical resections of medial temporal lobe structures in such patients. [courtesy of John Duncan, National Hospital for Neurologic Disease, Queen Square, London. From *Brain* (1996), **119,** 1677–1687, with permission of Oxford University Press]. (B) Similar imaging can be performed using SPECT and the iodinated compound iomazenil. Although slightly lower in spatial resolution, such studies provide information comparable to that discussed for flumazenil PET imaging in patients with focal epilepsy [courtesy of A. C. van Huffelen, University of Utrecht (van Huffelen *et al.* 1990). From Berkovic *et al.* (1993). In *Surgical Treatment of the Epilepsies* (J. Engel, Jr., Ed.), 2nd ed., p. 238. Raven Press, New York].

6. Electrophysiology

a. EEG Electrophysiological techniques provide the best temporal resolution for studying neuronal activity in patients with neurosurgical, neurological, and psychiatric disorders. Although more limited in spatial resolution than the tomographic techniques, EEG data provide the investigator and clinician an opportunity to learn about the timing of events and their synchronicity. Particularly important in the investigation of patients with seizures, EEG has been the mainstay of diagnostic evaluation of such patients. The limited spatial resolution of scalp EEG can be overcome by the use of more invasive techniques when patients are surgical candidates. In that setting, subdural grid electrodes, depth electrodes, and cortical surface electrodes can be used to obtain local extracellular field potential recordings and electrocorticograms.

System-specific information can be obtained with EEG when recording is accompanied by specific sensory, motor, or cognitive stimulation, a technique known as ERP recording. The resultant ERP maps are used to identify the general location and relative timing of a cortical representation of such functions in the brain. This approach has been used extensively to evaluate interruptions of the visual, auditory, or somatosensory systems (e.g., in multiple sclerosis) noninvasively and to provide more detailed cortical maps intraoperatively, or with subcortical or depth electrodes, in the evaluation of patients with seizure foci. Analysis of power spectra from scalp EEG and ERP recordings have also provided information of clinical use in neurodegenerative disorders and psychiatric syndromes.

b. MEG The measurement of minute magnetic fields in the brain with MEG is analogous to the measurement of electrical fields with EEG. Requiring far more complex equipment, the MEG method may have greater spatial resolution and greater accuracy in identifying electrophysiological dipoles, both at the surface and in the depths of the brain. Clinically, the method has been used to identify seizure foci and in a research setting has been employed for the experimental investigation of a wide range of neuropsychiatric disorders. Data with this technique are limited in number due to the fact that only a small number of MEG installations are currently operational and evaluating patients clinically.

c. Transcranial Magnetic Stimulation The creation of an intense focal magnetic field in the human cerebral cortex results in the induction of an electrical current that discharges cells lying tangentially within that volume of tissue. The discharge can result in a pseudophysiologic response. That is, if the motor cortex is stimulated, the appropriate contralateral muscle groups will contract. If the visual cortex is stimulated, a subject or patient will see a flash or phosphene in the contralateral visual field. Transient high-frequency stimulation of the cortex by magnetic stimulation will temporarily and reversibly deactivate it.

This approach has been used to create reversible lesions and has also been used as a therapeutic maneuver in the treatment of chronic depression, in a fashion analogous to electroconvulsive shock therapy delivered focally to the frontal cortex. When linked to a tomographic functional imaging technique, TMS can be used to map functional pathways directly in patients and normal subjects. There are guidelines for the safe use of this technique, which must be used with caution in epileptic patients.

7. Optical Intrinsic Signal (OIS) Imaging

The newest brain mapping technique to be used in a clinical setting is optical intrinsic signal imaging. This method provides information about cortical blood flow, blood volume, and metabolism in an integrated fashion. The approach is straightforward. White light is shone onto the exposed cortex and the amount of light of different wavelengths reflected from the cortex is measured. The reflectance and wavelength composition of light changes as a function of the neural activity of the illuminated tissue as a function of blood volume, blood flow, cell swelling, and the oxidative state of the tissue (among other variables). An invasive technique, OIS is used in the operating room to provide functional maps in which neuronal activity is varied by stimulation of peripheral nerves or, in the awake patient, by the performance of behavioral (e.g., language) tasks. The method has the best spatial and temporal resolution of all the functional imaging techniques, approaching 50 μm in the spatial domain and 50 ms in the temporal domain.

B. Multiple Modality Imaging

There are two types of multimodal integrative imaging studies that are important in the clinical evaluation of patients. The first is the within-subject integration of information from multiple brain mapping techniques, or the serial integration of multiple imaging studies in time using the same technique in the same individual. The second approach is the averaging or integration of information from multiple subjects, a much more difficult problem due to the great anatomical and functional variability that exists between individuals. The within-subject integration problem has yielded to a variety of excellent and elegant mathematical approaches for the alignment and registration of data. The between-subject problem is a more difficult one that is being resolved through the use of warping and morphing techniques.

1. Within-Subject Registration

A composite image of a patient derived from imaging using multiple imaging modalities or serial studies over time is a critical indicator of the clinical picture from an imaging perspective-for example, in a patient with a brain tumor in whom the natural history of the enlargement of the lesion can be evaluated quantitatively and objectively. Similarly, the integrated image of functional activation in cortical regions surrounding a lesion that is to be surgically resected can predict the relative risk of functional damage due to resection of normal cortex in the process of tumor ablation (Fig. 11).

Images of interictal spikes can be obtained by combining EEG and fMRI. Such images capitalize on the excellent temporal resolution of EEG and the complementary high spatial resolution of MRI. The relationship between hypometabolism in a seizure focus, determined with PET measurements of glucose metabolism, can be compared with benzodiazepine receptor binding, thereby increasing the specificity and

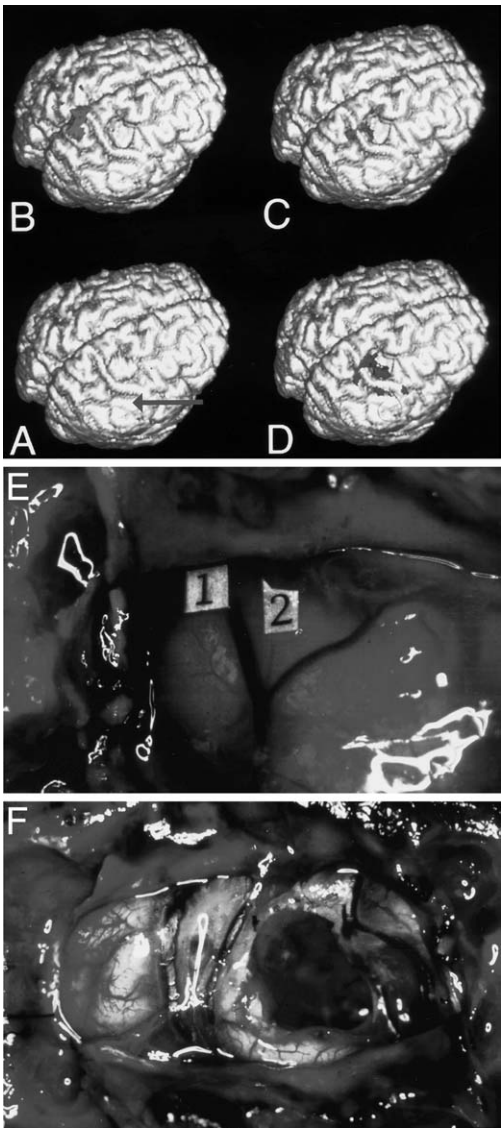


Figure 11 Within-subject registration techniques. (A) Three-dimensional reconstruction of a patient's brain with a cortical tumor in the region of the sensorimotor cortex (arrow). The structural data set, reconstructed from MRI, can then be combined with functional information about cerebral blood flow changes associated with motor tasks, derived from ^{15}O-labeled water PET studies when the patient moved the left leg (B), shoulder (C), or fingers (D). (E) Intraoperative view of the hand area of the motor cortex where the localization of sensorimotor hand function was identified intraoperatively with electrophysiologic techniques and labeled "1" and "2." (F) Operative view of site following resection of tumor. Because of the close proximity of the tumor to the activation site for finger movement, identified with preoperative PET imaging, it was predicted that this patient would have loss of fine motor control of the hand following complete resection of the lesion. This was in fact the case and is indicative of the accuracy and predictive power of preoperative mapping in patients with cortical lesions close to vital cortical structures (courtesy of Roger Woods, ULCA School of Medicine, and Scott Grafton, Emory University School of Medicine).

sensitivity of the combined result. One can also combine electrophysiologic data sets with tomographic data from PET, MRI, and MRS to provide a composite preoperative assessment of patients with epilepsy. PET measurements of cerebral blood flow, oxygen extraction, and oxygen metabolism in the same subject and comparison with diffusion-weighted MRI and MR angiography (or helical CT) provide a very complete picture of the supply–demand relationships of the brain parenchyma in patients with ischemic cerebrovascular disease. Such combined studies will undoubtedly become a clinical norm rather than an exception in the future.

Comparisons of scans between individual patients will become increasingly important. Such comparisons require that the scans be spatially normalized to account for individual differences in brain structure. The ability to normalize a scan to a standard brain space means that individual patient scans can be compared with a representative scan from a population of normal subjects that takes into account a realistic estimate of the anatomical variability in that population. In the structural domain, this ability may increase the sensitivity with which subtle heterotopias or other migrational abnormalities are identified in patients with focal epilepsy. Similarly, selected patterns of atrophy in neurodegenerative diseases should be detected in a more sensitive and specific fashion. The ability to compare representative scans across patient groups could also have importance for clinical trials where a patient group on an experimental therapy could be compared with a control group in an objective and quantifiable manner.

II. SURGICAL STRATEGIES

A. Preoperative

The preoperative investigation of patients with cerebral lesions falls into two general categories. The first is targeting areas for the purpose of stimulation or ablation. These circumstances occur in patients with movement disorders in which parts of the basal ganglia or other subcortical regions are selected for lesioning (e.g., pallidotomy) as a means of improving symptoms. Lesioning using stereotactic focal radiation or direct surgical ablation, by heating or freezing, are of interest for the treatment of cerebral neoplasms and vascular malformations. In a similar fashion, stimulating electrodes are now being employed in the treatment of Parkinson's disease and certain types of tremors.

The exact location for the placement of these electrodes requires knowledge of the structural anatomy and local electrophysiology or function to obtain maximal therapeutic benefit.

1. Targeting

Brain mapping techniques have already been employed for targetting. Currently, these approaches are limited to obtaining better definition of a patient's structural anatomy and developing better atlases to identify selected portions of the brain, given the individual variability among patients. With high-resolution structural imaging, specific locations for potential lesions can be identified anatomically and a frameless stereotactic approach can be used to direct an ablation probe or stimulation electrode. Once located, electrophysiological recordings can be used to verify the local functional environment of a given anatomical site.

Functional activation of deep brain sites (e.g., medial globus pallidus) is an important area of current investigation. Ideally, such sites should be located both structurally and functionally to identify a surgical target more accurately preoperatively. Such a facility would mean less retargeting and repositioning of electrodes and probes, thus reducing operating room time and morbidity, resulting in a higher success rate. Currently, there are no validated clinical examples of such an approach.

2. Differentiation of Normal from Abnormal Brain

Another important aspect of presurgical investigation of patients with brain mapping techniques is the identification of normal cortex or deep brain structures so that they may be avoided during surgical resection or ablation of cerebral lesions. The goal is to remove an abnormality in its entirety without removing normal brain tissue. Preoperative evaluation with PET, SPECT, fMRI, or transcranial magnetic stimulation may be employed to identify the functional anatomy of an individual's brain to determine the safety of surgery and a strategy for reaching an abnormal brain region to remove pathologic tissue.

Such functional imaging techniques may ultimately replace procedures such as the Wada test or reversible pharmacologic interruptions of brain function. The selective administration of barbiturates to brain regions that produce transient deficits has long been used to determine the relative safety associated with removal of a portion of the brain. Nevertheless, such tests are difficult to perform, particularly in younger children, and the exact distribution of the pharmacologic agent can be difficult to verify. If the same information can be obtained through presurgical use of functional imaging, it is hoped that both the accuracy and the ease of obtaining such data will be enhanced (Fig. 11).

The combined use of all of the noninvasive scanning methods, both current and experimental, resulting in integrated and composite images linked to interactive graphics stations in the operating room or the interventional neuroradiological suite, will surely become a part of such interventional procedures in the future. To be successful, these approaches need to be less costly, more accurate, and associated with lower eventual morbidity than conventional surgery without such ancillary noninvasive procedures.

B. Intraoperative Mapping

The most frequently used intraoperative brain mapping techniques are electrophysiological. These include electrocorticography for the identification of epileptic foci. Where cortical resections are indicated, evoked potential recordings with cortical electrodes are combined with peripheral nerve stimulation in anesthetized patients. Such an approach is frequently used to identify sensory and motor cortices. When more complex information is needed, particularly about language areas of the cortex, anesthesia is reversed and the patient is awakened during surgery. Direct electrical stimulation of the cortex is then used to reversibly disrupt local neuronal function while the patient performs behavioral tasks. In this setting, a patient must be psychologically able to accept the disturbing aspects of awake neurosurgery. Ideally, the examiner and the patient should have good rapport so that behavioral testing can proceed in an efficient and cooperative manner.

Optical intrinsic signal imaging, currently used in a research setting, may soon augment intraoperative electrophysiological measurements by directly providing visualization of cortical maps and functional responses seen ultimately through the operating microscope in real time. The high spatial and temporal resolution of this approach can be used to validate preoperative images of functional anatomy from individual patients. In addition, the functional brain maps can be updated as they become distorted from the preoperative state by changes in brain shape resulting from osmotic dehydration, ventricular

drainage, or local edema during the course of the operation.

As more interventional neuroradiology and neurosurgery are performed within imaging devices, the reacquisition of structural and functional information as an invasive procedure unfolds will become possible and realistic. The most likely source of these data will be from MRI devices, where either patients are moved in and out of the device to update imaging data or the interventional procedure is performed within the magnetic field directly. In either case, direct updates of structural and, potentially, functional information can be provided to the clinical team performing the procedure.

III. ATLASES

The advent of modern mathematical and computational approaches to averaging imaging data across subjects has led to the generation of population-based probabilistic atlases. Such atlases are already in existence for the normal brain at different age ranges and for other regions of the body. Disease-based atlases may be useful for the differential diagnosis of human cerebral disorders.

The basic approach to generating such atlases is to obtain images from a large number of subjects (i.e., typically hundreds or even thousands) in a mathematical framework that produces a database that is probabilistic. Such an atlas allows the user to obtain relative information that takes into account the variance in structure and function in the human population (Fig. 12). Once established, such an atlas can interact with new data sets derived from individual subjects and patients or groups of subjects or patients. Thus, a clinician or investigator who performs an MRI scan of a single patient with focal epilepsy could call on a digital probabilistic atlas of normal subjects and compare the patient with the average normal atlas. The atlas will use the normal variance information estimated from the population of normals from which it is generated to determine whether a patient's scan falls within or outside normal morphometric limits. If the atlas is constructed from sufficient subjects, a subpopulation could be selected that more closely resembles a patient's demographic profile. In such a case, one might ask for only those normal subjects from the atlas who are right-handed, of a particular racial origin, and females ages 25–30. An increasing number of variables can be included in such a prior specification depending on the size of the data set constituting the atlas and the range of demographic information collected about the contributing subjects. As a result, it would become possible to detect subtle abnormalities of diagnostic importance that would not be identified by the less sensitive conventional approach of qualitatively examining two-dimensional image sets by eye. In addition, such an atlas-based approach will give an objective and quantifiable magnitude to any detected abnormality. The scan data from any patient can be added to an atlas database, increasing its value with regard to particular patient groups.

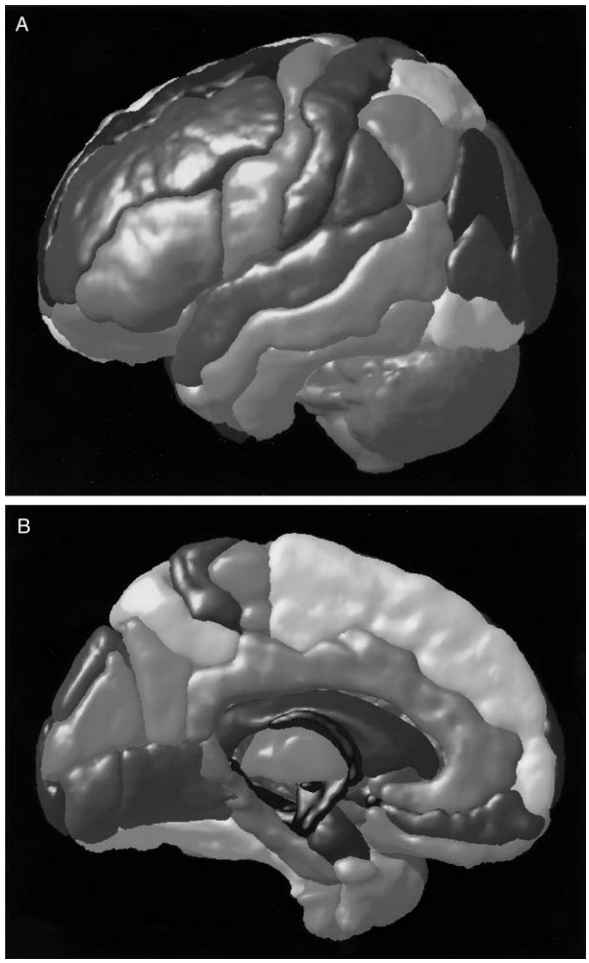


Figure 12 Probabilistic atlases. Population-based probabilistic atlas of the normal human brain derived from 67 subjects, ages 20–40 years, seen from the lateral (A) or midsagittal (B) views. The structures have been segmented to show cortical regions at a 50% confidence limit in the population [courtesy of Alan Evans and colleagues. Montreal Neurologic Institute].

Disease-based atlases, thus generated, are currently being assessed. It is possible to imagine morphometric or functional atlases for Alzheimer's (Fig. 13) and

Parkinson's disease, schizophrenia, and other disorders. Such atlases would also provide a population and disease-based opportunity to examine the natural history of morphometric or functional abnormalities as a function of disease progression, age of onset, or other variables. Such atlases could also be used to identify changes in natural course as a function of therapeutic intervention. Consider, for example, a clinical trial with a new drug for Alzheimer's disease. The Alzheimer's disease population atlas would provide estimates of morphometric changes in focal atrophy as well as, for example, alterations in crerebral glucose metabolism as a function of disease progression. A population of patients at a certain stage of the disease could be divided into two groups, one given an experimental therapy and the other given placebo. Serial imaging of both groups with the appropriate techniques would then provide longitudinal imaging data. Comparisons of morphometric and metabolic changes as a function of time between the two groups would be undertaken to detect objective and quantitative differences between the two groups. Any differences would represent a measure of the effect of the therapeutic intervention on progressive atrophy or glucose metabolism due to the natural history of the disease. It is probable, although currently unproven, that such an approach will be more sensitive in detecting differences between control and experimental groups, thereby requiring either fewer subjects or shorter time frames for therapeutic assessment, thus resulting in lower costs of clinical trials.

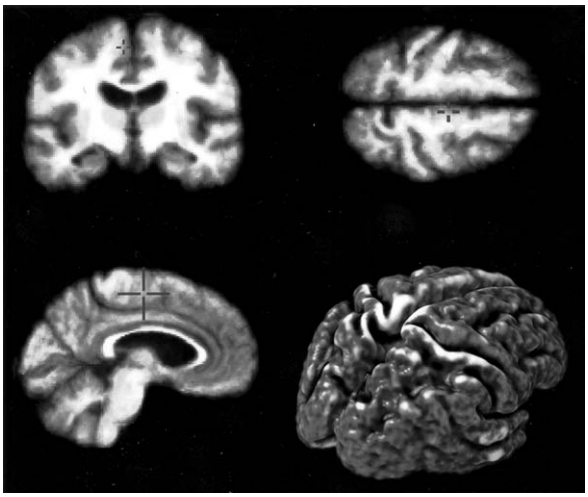


Figure 13 Probabilistic population atlas derived from nine individuals with Alzheimer's disease. This atlas is presented as a set of two-dimensional orthogonal views plus a three-dimensional rendering (bottom right) and is produced using a continuum-mechanical approach. Note the influence of atrophy on the composite image demonstrating widening of the major fissures of the brain as well as sulci in the neocortex. Such disease-based population atlases will be useful not only in tracking the natural history of cerebral disorders but also in providing objective and quantifiable information about structural and functional changes associated with experimental therapy for these disorders (courtesy of Paul Thompson and colleagues. UCLA School of Medicine).

IV. PLASTICITY

One of the most exciting and dramatic observations to come from human brain mapping with a wide range of structural and functional techniques has been the dynamic plasticity of function in both normal brains and the brains of patients with neurological and neuropsychiatric disorders. Brain maps must therefore be viewed as dynamic, changing with development, disease progression, and normal learning and in the recovery of function after acute injury. The dynamic plasticity of functional brain maps provides an exiting opportunity to study these processes. It also means that the use of brain maps must take into account such variability in the design of brain mapping studies for patients with cerebral disorders.

For example, just as structural and functional studies must be normalized for spatial variability in the population, disease-based maps must be normalized in time to account for dynamic changes that occur with progression. Thus, a comparison of patients with Alzheimer's disease or other neurodegenerative disease should be stratified by time of onset or other variables that take into account the pattern of changing functional maps. The same is true after an acute brain injury, such as trauma or cerebral infarction. The complex interaction and highly variable changes in blood flow, blood volume, water diffusion, oxygen, and glucose extraction and metabolism will all be more appropriately interpreted if they are stratified by time from onset of cerebral injury. So too will plastic changes associated with recovery and reorganization after irreversible damage. Compensatory properties of the human nervous system have been clearly demonstrated in studies of patients following stroke who recover motor function (Fig. 14). The study of drug-induced, behaviorally associated and surgically promoted plasticity will, we predict, be an important part of brain mapping in the study of patients with neurological, neurosurgical, and psychiatric disorders.

The value of imaging data depends on an appreciation of the changing landscape of functional patterns.

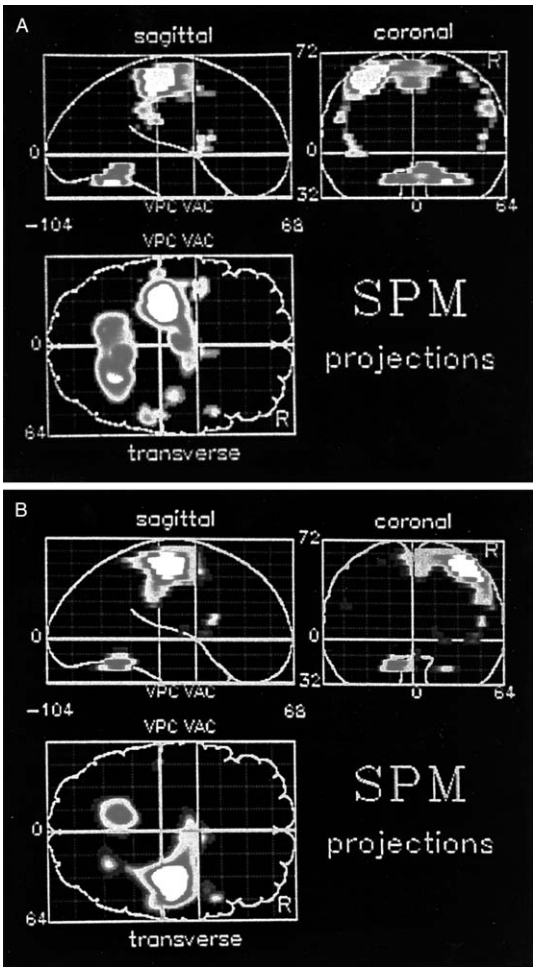


Figure 14 Compensatory reorganization of the brain after acute injury induced by cerebral infarction. (A) Normal response of a group of control subjects performing a motor task with the right upper extremity. Relative increases in cerebral blood flow derived from PET measurements demonstrating increased perfusion in the contralateral hand area of the motor cortex and the ipsilateral cerebellum. (B) The same motor task and methods were applied to a group of patients who had small subcortical cerebral infarctions associated with upper extremity paresis, all of whom recovered in the days to week following the acute ischemic injury. Note that when these subjects perform the same task, not only are there responses in the expected areas previously identified in the normal controls—that is, the contralateral hand area of the motor cortex and ipsilateral cerebellum—but also there are relative increases in cerebral blood flow in the ipsilateral hand area of the motor cortex and the contralateral cerebellum. Such studies provide useful insights into how the brain reorganizes following acute injury and can also be used to study compensation in more chronic states such as might be encountered in neurodegenerative disorders. These types of insights may be useful for designing more efficient, effective, and timely neurorehabilitation protocols employing behavioral, pharmacologic, or, potentially, surgical interventions for the restoration of function or its maintenance following acute or, chronic injury to the brain [courtesy of Francois Chollet and colleagues, Toulouse, France. From Chollet *et al.* (1991). *Ann. Neurol.* **29,** 63–71].

This is particularly true for techniques that make relative measurements. For example, relative cerebral blood flow measurements obtained with fMRI, SPECT, or PET may be misleading in patients with large cerebral infarctions. The evaluation of motor reorganization after cerebral infarction in a patient or group of patients must take into account a variable cerebral blood flow baseline in the setting of hemodynamically unstable tissue. Cerebral blood flow may be very low acutely, rise dramatically soon thereafter, and then reach some stable new level days, weeks, or months later. Currently, it is uncertain how increments or decrements in blood flow associated with neuronal firing changes that are task induced will behave under these different conditions of baseline blood flow. Is it valid to compare motor task activation in a cortical region when the "resting state" blood flow is altered from the normal value by 50–200%? Are there ceiling or floor effects in these responses? These issues need to be addressed before a proper interpretation of scans from such patients will be possible.

We predict that the ability to image plastic reorganization, both in normal and in pathologic states, will provide new insights, previously unavailable, about the constant reorganization of the human brain. Such information will be valuable in the design of behavioral, surgical, and pharmacological interventions in patients that facilitate and maximize the efficiency of the natural recovery processes. The imaging techniques should also provide a means to evaluate specific rehabilitation interventions, to determine their appropriateness, effectiveness, and timing, and to select patients for them. These abilities are currently lacking because the necessary information about the variables discussed has not been previously available.

See Also the Following Articles

CEREBRAL CORTEX • ELECTRICAL POTENTIALS • ELECTROENCEPHALOGRAPHY (EEG) • EVENT-RELATED ELECTROMAGNETIC RESPONSES • FUNCTIONAL MAGNETIC RESONANCE IMAGING (fMRI) • MAGNETIC RESONANCE IMAGING (MRI) • NEOCORTEX • NEUROIMAGING • PSYCHOPHYSIOLOGY

Acknowledgments

The authors thank contributing investigators for allowing their work to be reproduced in this article. This work was partially supported by a grant from the Human Brain Project (P01-MH52176-7) and generous support from the Brain Mapping Medical Research

Organization, The Ahmanson Foundation, the Pierson–Lovelace Foundation, the Jennifer Jones–Simon Foundation, the Tamkin Foundation, the Northstar Fund, and the Wellcome Trust. The authors thank Laurie Carr for the preparation of the manuscript and Andrew Lee for assistance with the illustrations.

Suggested Reading

Atlas, S. W. (1991). *Magnetic Resonance Imaging of the Brain and Spine*. Raven Press, New York.

Berkovic, S., Newton, M. R., and Rowe, C. C. (1992). Localization of epileptic foci using SPECT. In *Epilepsy Surgery* (H. Luders, Ed.), pp. 251–256. Raven Press, New York.

Chiappa, K. H. (1983). *Evoked Potentials in Clinical Medicine.* Raven Press, New York.

Chollet, F., DiPiero, V., Wise, R. J., Brooks, D. J., Dolan, R. J., and Frackowiak, R. S. (1991). The functional anatomy of motor recovery after stroke in humans: A study with positron emission tomography. *Ann. Neurol.* **29**(1), 63–71.

Daly, D. D., and Pedley, T. A. (1990). *Current Practice of Clinical Electroencephalography*, 2nd ed. Raven Press, New York.

Holmes, C. J., Hoge, R., Collins, L., Woods, R., Toga, A. W., and Evans, A. C. (1998). Enhancement of MR images using registration for signal averaging. *J. Computer Assisted Tomogr.* **22**(2), 324–333.

Luden, H. (Ed.) (1991). *Epilepsy Surgery*. Raven Press, New York.

Mazziotta, J. C., and Gilman, S. (1992). *Clinical Brain Imaging: Principles and Applications*, Davis, Philadelphia.

Mazziotta, J. C., Toga, A. W., Evans, A., Fox, P., and Lancaster, J. (1995). A probabilistic atlas of the human brain: Theory and rationale for its development. *NeuroImage* **2,** 89–101.

Newton, T. H., and Potts, D. G. (Eds.) (1981). *Radiology of the Skull and Brain: Technical Aspects of Computed Tomography*. Vol. 5. Mosby, St. Louis.

Shellock, F. (1997). *Pocket Guide to MR Procedures and Metallic Objects: Update 1997*. Lippincott–Raven, Philadelphia.

Toga, A., and Mazziotta, J. (1996). *Brain Mapping: The Methods*. Academic Press, San Diego.

Toga, A., and Mazziotta, J. (2000). *Brain Mapping: The Systems*. Academic Press, San Diego.

Williams, A. L., and Haughton, V. M. (Eds.) (1985). *Cranial Computed Tomography: A Comprehensive Text*. Mosby, St. Louis.

Information Processing

JOE L. MARTINEZ, Jr., STEPHEN A. K. HARVEY, ADRIA E. MARTINEZ, and EDWIN J. BAREA-RODRIGUEZ

University of Texas, San Antonio

GLOSSARY

amygdala A group of nuclei in the anterior–medial part of the brain that are involved in emotional memory.

declarative memory Memory that involves the conscious recollection of events.

DNA (deoxyribonucleic acid) The central library of an organism's genetic information.

DNA microarray Technique used to investigate thousands of genes from the same tissue sample.

engram The memory trace laid down in the brain. It is believed to be formed by changes in the synapse.

Hebbian synapse When a signal between two neurons is strengthened due to the simultaneous activation of the presynaptic and postsynaptic neurons.

hippocampus A structure located in the medial temporal lobe that is involved in spatial learning.

learning A change in behavior as a result of experience.

long-term potentiation A form of synaptic plasticity induced by brief high-frequency stimulation.

memory The ability to store and recall an experience.

nondeclarative memory Memory usually measured by performance, such as automatic motor skills.

RNA (ribonucleic acid) Short-lived molecule that contains information about the DNA.

transduction The change of physical energy into neural signals.

The neuron is the principal functional unit of the brain. Neurons both transform physical energy such as pressure and temperature into neural energy and conduct light and sound energy transformed by the eyes and ears. This transformation is the first step in information processing. Because we all have slightly different perceptual worlds and experiences, our individual identity is based on our memories of perceptions. This article presents learning and memory as an example of an information processing system. The article takes a cognitive and molecular biology approach in understanding how memories are formed.

> *The world first exists, and then the states of mind; and these gain a cognizance of the world which gets gradually more and more complete.*
>
> William James (1892)

Understanding movement of sensation from the external environment to the inside of one's head as useful information is a process that has consumed philosophers and scientists for hundreds of years. Think of the Challenger space shuttle explosion. In your mind's eye you can visualize the twin plumes of smoke streaking higher into the sky that represented the end of the mission. How did this image get into your brain and why will it be with you for the rest of your life? These are two questions this article seeks to address.

The human body exists in a sea of physical stimuli, only a tiny fraction of which we are aware. Consider

the electromagnetic spectrum, which includes X-rays able to permeate solid objects and ultra-low-frequency radio waves with which we can contact submerged submarines thousands of miles away. We see only a tiny fraction, from 400 to 760 nm. We do not see cosmic rays or television transmissions. Perhaps our tiny range of vision allows us to focus on those external objects necessary for our survival. Our senses, then, are transducers that convert physical energy to neural energy. Ears convert sound, nerve endings convert temperature and pressure, the labyrinth converts gravity, the tongue converts acids and sugars, and the nose captures molecules from the air and gives them each a unique sensation. Consider the curious life of synasthets, people who hear light or see sound when their neural energies are crossed at the level of the thalamus so that hearing neural energy goes to the visual areas and *vice versa*.

We do not all perceive the same world because physical energy is transduced by biological systems that have many states and conditions. A person who walks into a darkened movie theater does not see nearly as clearly as the person who has been in the theater for a long time and who is dark adapted. At that moment their perceptual worlds differ. A person unable to distinguish red and green, a dichromatic, sees a different world, and a blind person does not see the world at all. Because we all have slightly different perceptual worlds and experiences, our individual identity is based on our memories of perceptions. As William James said, cognizance of the world is an iterative process that depends on perception and memory; together, perception and memory allow information processing. Everyone agrees that the brain is the organ that perceives the world and stores our individual memories, and the computational unit of a brain is the neuron (Fig. 1). As a cell, the neuron exhibits continuous metabolic activity but exists in only one of two transmission states, "on" and "off." Thus, it is analogous to a computer memory address, which contains a bit of information—either 1 or 0. Although the content of a computer memory address is set by a central processor, a neuron is much more complicated, gathering information from its dendrites and deciding whether and how many on states, called action potentials, it will generate. Neurons connect through special contacts called synapses at the end of axons. As depicted in the Fig. 1, a neuron connects with many other neurons, usually on a part of the neuron called a dendrite.

Figure 2 shows that neurons are organized into units called nuclei and these nuclei perform functions and communicate with other nuclei. Neuroanatomy is the study of the organization of nuclei. Continuing our computer analogy, nuclei function like integrated circuits. Note that in Fig. 2, neurons become more complex as one moves up the phylogenetic scale from mouse to human. Because it is more complex, the human neuron has more computational power. This is the basis of superior information processing in humans. Note that rat and human brains are similar in architecture. Both have cerebral hemispheres, both have a thalamus, and both have a cerebellum. Based on this similarity, we expect humans and rats to have similar functions, and both are capable of learning mazes, for example.

Now we can ask a difficult question: How can a perception become a memory and vice versa? Figure 3 shows a most intriguing result. In this experiment, a monkey was trained to look at a fixed point in the center of the circular pattern, and this was followed by

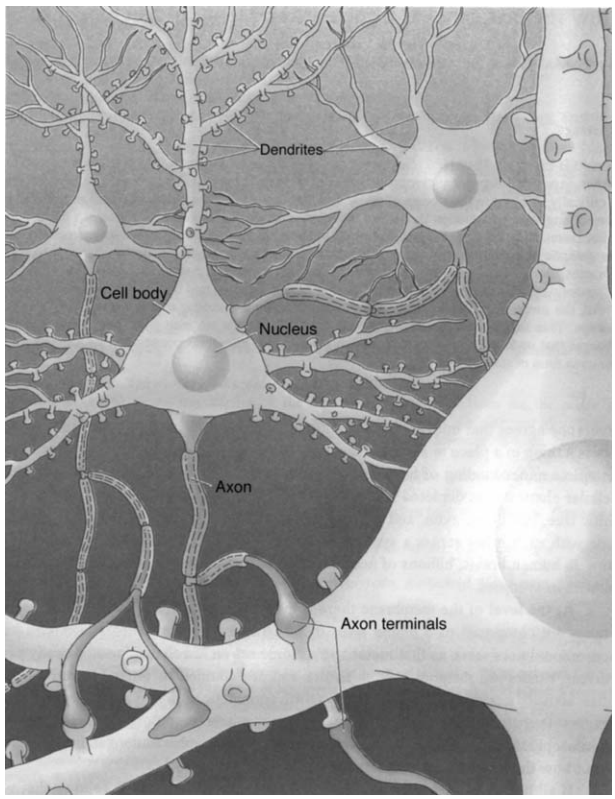


Figure 1 The neuron is the basic signaling unit of the nervous system. There are 10^{12} neurons in the brain and most have a cell body that gives rise to the axon, dendrites, and synaptic boutons or terminals [reproduced with permission from Rosenzweig, M. R., Leiman, A. L., and Breedlove, S. M. (1996). *Biological Psychology*. Sinauer Associates, Sunderland, MA].

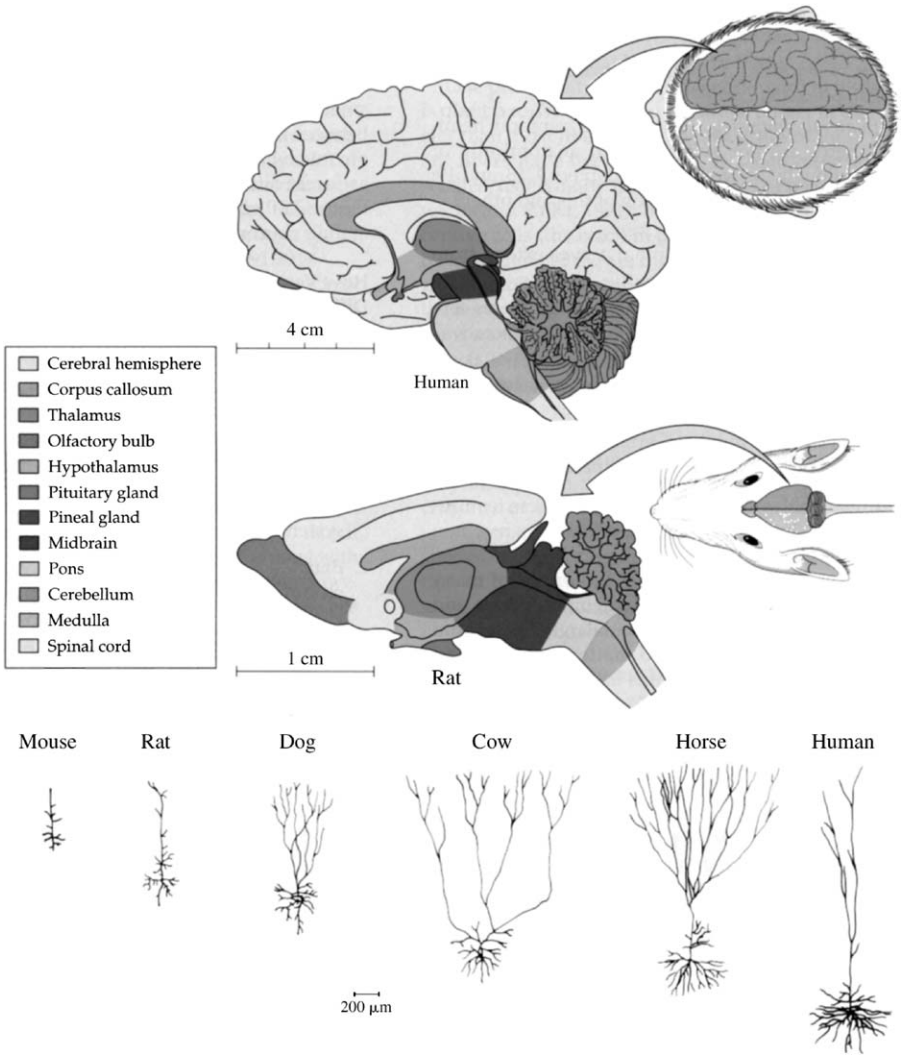


Figure 2 Midsagittal view of human and rat brains. The structures observed contain neurons and each performs different functions [reproduced with permission from Rosenzweig, M. R., Leiman, A. L., and Breedlove, S. M. (1996). *Biological Psychology*. Sinauer Associates, Sunderland, MA].

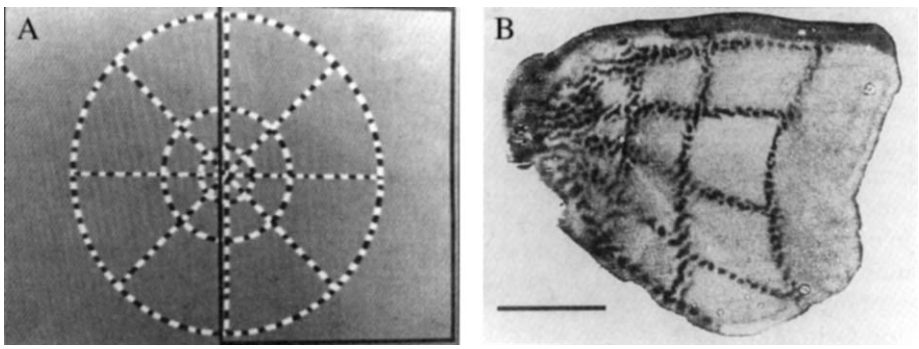


Figure 3 (A) A pattern of flickering lights shown in the visual field of a macaque monkey (B) Pattern is mapped in the striate cortex using 2-deoxyglucose (reproduced with permission from Tootell *et al.*, 1988. Deoxyglucose analysis of retinotopic organization in primate striate cortex. *Science* **218,** 902–904).

an injection with a synthetic form of glucose. Neurons that were metabolically active look darker. The circular pattern is recreated on the cortex of the monkey. A network of interconnected neurons represents the perception. The pattern can be stored in the monkey brain and made into a memory if the network that represents the circle can be made permanent. Hebb proposed this idea in 1949, but he called the network of neurons a cell assembly. In order for the monkey to re-experience the pattern, the network has to be reactivated in its brain.

Hebb proposed that the pattern could be made permanent if the connections between neurons were made stronger in a functional sense. That is, if neuron A, the first in the network representing the circle pattern, persistently fires neuron B, and so on, then some process takes place in either A or B and all the neurons that represent the pattern so that the efficiency of A firing B is increased. In other words, when neuron A fires the pattern is recreated. Today, the phenomenon by which one cell creates a larger response in the second by repeated stimulation is called long-term potentiation (LTP), and it is considered to be the best model of how the brain may store information.

Are there any nuclei in the brain whose function is to make memories? In the 1950s, a patient H.M. had his hippocampus removed on both sides of his brain to control his grand mal epilepsy. Following the surgery, H.M. could remember the past, but he could not remember anything new. He was said to have a deficit in his ability to consolidate or make permanent new memories. Interestingly, Fig. 4 shows that H.M. could learn a backward mirror drawing task, even though he did not remember the apparatus from day to day. It is thought that the hippocampus is a brain structure whose function is to lay down new declarative memories or "knowing that" (the apparatus) rather than "knowing how" (to draw backwards). In Section III.A, we take a closer look at the hippocampus and determine whether its organization or arrangement of

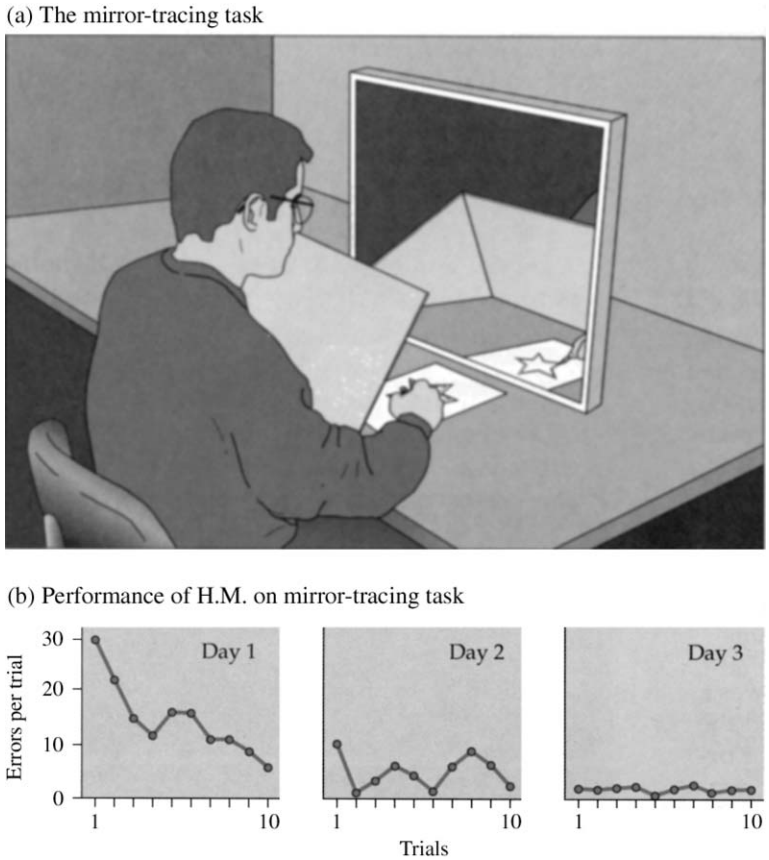


Figure 4 (a) The mirror-tracing task used to test memory in patient H.M. (b) Performance of H.M. on this task. Despite the improvement over days in this task, H.M. did not remember having performed this task in the previous days [reproduced with permission from Rosenzweig, M. R., Leiman, A. L., and Breedlove, S. M. (1996). *Biological Psychology*. Sinauer Associates, Sunderland, MA].

cells provide clues as to how it might lay down memories. In Section III.G, we examine the phenomenon of LTP in the hippocampus. If the hippocampus really is a nucleus designed to lay down memories, then LTP should abound in this structure. Finally, we examine an interesting phenomenon of memory: Many are permanent. What could possibly change in a neuron so that you remember the first time you rode a bicycle?

I. LEARNING AND MEMORY AS AN INFORMATION PROCESSING SYSTEM

In the morning an employee arrives to work and realizes that her usual parking spot is taken. She parks in another parking area, but before she leaves she inspects the area looking for landmarks that will help her find her parked car at the end of the day. At the end of the day she searches for her car in the parking lot by looking for specific landmarks. The ability to find the location of the car by searching for landmarks is possible because the person makes an association between the location of the parked car and the landmarks.

It is 12 noon, you are feeling hungry but still have work to do. You know that when the bell of the microwave oven sounds, your coworker, who is a great cook, is warming up food. Today, you hear the bell and your mouth waters. Sometime in the past an association was established between the bell and the food. In both the cases of using landmarks to one's car (declarative/spatial learning) and the sounding of the microwave oven bell causing one's mouth to water (nondeclarative/classical conditioning learning), it is said that learning occurred. Both visual and auditory stimuli were processed and associated in memory.

Using the spatial learning example (this also applies to auditory information), we can construct a model that explains how the person establishes an association between landmarks and the parked car. This model can be divided into a set of stages. The first stage involves the registration of the cues and the location of the car in the visual system. Thus, as described earlier, physical energy, light, which is used by the visual system to form the images (light pole, trees, or a sign with the lot number) that surround the car, is transformed by the sensory register into neural energy. Transduction is the first step in information processing—the transformation of physical energy into a type of energy that the brain understands. The second stage involves the processing of neural energy by the neural structures involved with establishing the association (the hippocampus). In the third and final stage, the information is retrieved as needed.

The information processing model constructed previously uses two approaches to explain the surrounding landmarks–parked car associations: cognitive psychology and computer science. The cognitive psychology approach seeks to understand the mental processes that take place during the time that a stimulus is perceived and the behavior that it elicits in an individual. The computer science approach helps us understand how information is processed so that it can be used at a later time. Interestingly, we can borrow many concepts from computer science. For instance, in order for perceived information to be used it needs to be encoded, stored, and retrieved. One major difference between the brain and a computer is that the brain uses billions of neurons in parallel fashion to process information. The computer has a main processor that allows the processing of a number of serial operations. You know that you have never seen the word "tedinu" (united backwards); you do not have to scan your memory from aardvark to zygote, as a computer does.

Humans are processors of information and human memory is an exemplar information processing system. In the past four decades we have learned a great deal about memory and the underlying mechanisms that subserve the encoding of information, and we are learning more about the mechanisms involved in the storage and retrieval of information. Next, we discuss the concept of the memory trace and current evidence that shows how a number of neural structures are involved in mediating the processing of information that leads to the formation of spatial and emotional memories.

II. IN SEARCH OF THE MEMORY TRACE (ENGRAM)

Curiosity in the understanding of the ability to remember information (memory) dates back as far as 800 BC, when memory was considered a virtue. Interestingly, even in those days teachers and philosophers recognized the importance of the sense of sight to develop strategies to remember information, also known as mnemonic techniques. The context became associated with all the information acquired in a place. Although an engram, the memory trace, was formed with a given experience, little was speculated about the localization of the engram.

III. ANATOMICAL BASIS OF THE MEMORY INFORMATION SYSTEM

Here, we focus on two types of learning presented previously—spatial (declarative) and classical conditioning (nondeclarative) learning that allow us to empirically understand the role of the brain in information processing. Interestingly, both humans and animals utilize these forms of learning. Furthermore, one type of learning may exist without the other. Before we discuss different types of learning, it is important to describe two structures believed to be involved in learning and memory—the hippocampus and the amygdala.

A. Anatomy of the Hippocampus and Amygdala

As noted earlier, the hippocampus was found to be involved in memory when H.M.'s hippocampi were removed and he exhibited anterograde amnesia. Figure 5 shows a view of the hippocampus and the amygdala using a transparent model of the human brain. The hippocampus was so named because it has a close resemblance to a sea horse ("hippo" means horse, and "kampos" is Greek for "sea monster"). The hippocampus is an elaborate infrastructure of computational systems. However, the complex basis on which the hippocampus functions can be understood in a simpler way by describing its trisynaptic structure.

The hippocampal formation is composed of the hippocampus proper and is formed by layers of neurons that form areas CA1–CA3 (cornu ammonis 1–3), the dentate gyrus, the entorhinal cortex, and the subiculum. The hippocampal formation has an internal circuit known as the trisynaptic circuit. First projection is from the entorhinal cortex to the dentate gyrus. Second, the dentate gyrus neurons give rise to the mossy fibers (axons) that project to CA3 cell fields. Third, the CA3 neurons give rise to the Schaffer collaterals (axons), which project to the CA1 cell field. In addition, the hippocampal formation receives projections from many subcortical structures (e.g., amygdala). Neuroanatomy is replete with nerve tracts that run from one physical location in the brain to another, but the hippocampus can be thought of as a loop that returns its signals to a location close to their cortical origin.

The amygdala, derived from the Greek word for "almond," is a collection of nuclei located in the temporal pole of the cerebral hemisphere rostral to the hippocampal formation. The amygdala comprises the lateral, medial, basolateral, and central nuclei. These nuclei have a number of functional roles, from receiving and processing olfactory information to sending projections to areas involved in autonomic responses. Studies suggest that the amygdala is directly involved in classical conditioning associated with fear.

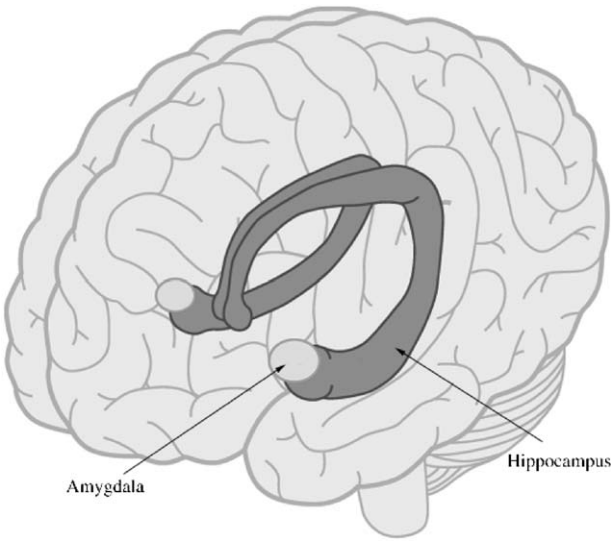


Figure 5 Studies from human and rodents show that the hippocampus mediates declarative information. Studies derived from rodents show that the amygdala is involved in classical conditioning (reproduced with permission from *www.brainconnection.com*).

B. Declarative Learning and the Hippocampus

Remember H.M.; imagine not being able to remember what you ate for breakfast, lunch, and dinner. Picture yourself at a dinner party unable to recognize any of the faces of the people whom you had met only minutes earlier. Recall that because H.M. suffered from epileptic seizures, doctors suggested that his hippocampus be removed. The result was both successful and unfortunate. H.M.'s seizures ended, but H.M. could not form any new memories. In addition to H.M.'s anterograde amnesia, he was unable to remember the 3 years of his life prior to the surgery (retrograde amnesia). Through endless studies and tests performed on H.M., many theories about learning and memory were revolutionized. Previous views focused on the concept of functional localization; this meant that specific brain regions were responsible for particular functions. This view (for some) was

modified. Research from H.M. suggested that although functional localization is legitimate, there is a more complex circuitry associated with memory than once believed. It could have been hypothesized that because H.M. was incapable of transferring short-term memories into long-term memories, he could not have learned anything new. This was not entirely true. Remember that H.M. could learn skills such as the backwards mirror-drawing task. H.M. illustrated learning by improving at the task, despite his inability to recognize the test each day. This gave much insight into the process of converting short-term into long-term memory, a process called consolidation. H.M.'s ability to learn and improve at a task but to have no recollection of tracing the lines in the star gave novel insight into the role of the hippocampus in storing declarative memories.

C. The Role of the Hippocampus

Think of the hippocampus as the place where short-term memories are processed until they turn into long-term memories. Because H.M. lacked any ability to "remember" postsurgery, doctors and researchers deduced that the hippocampus was the locale where short-term memories turned into long-term memories. Keep in mind that H.M. could recall his life 3 years prior to the removal of his hippocampus. This illustrated the fact that long-term memories were not stored in the hippocampus; they were subsequently believed to be stored in another part of the brain called the neocortex.

D. Declarative Memory: One Type of Memory

Declarative memory is what we most often think of as memory. Remembering where you were when you first rode your bicycle or where you were when the Challenger exploded are examples of declarative memory. What defines declarative memory is the ability to consciously recollect the situation in which you learned something new. Declarative memory is context based. H.M. was able to recall experiences before his retrograde amnesia. H.M. had intact declarative memory to an extent, but he was unable to form any new declarative memories because he lacked the anatomical structure to do so.

E. The Animal Model

These findings provided an understanding about the hippocampus as a whole, but structurally there was much more to learn. Because of H.M. and considerable research on amnesic patients, we know now that the hippocampus is critical in transferring new memories into stored ones. How does this system work? Researchers needed to first find a model to test that was comparable to humans. Rats have similar nervous systems to humans, but unlike humans they are not able to verbally recall (declarative) what they have learned. They do use spatial cues to find their way in the environment (spatial learning). Spatial learning is mediated by the hippocampus, and thus used as a model for declarative learning in animals. The Morris water maze tests spatial memory by placing rats in a pool of opaque water. The rats have to find a submerged platform only by using cues (a large black star on the wall or an inflated beach ball) which are placed around the pool. Rats must first make a choice about what they did before (how they found the platform) and then use visual cues. Because the time spent finding the platform decreases with each new trial, it is believed that they have learned to find the platform by using the visual cues. Research on spatial information processing in the hippocampus shows that there are cells, known as place cells, that code for specific cues in the environment. Animal studies have shown that the dorsal hippocampus is specifically required for spatial learning, and the equivalent (posterior) part of the human hippocampus shows increased metabolic activity when navigational information is being recalled. Some species actually show an increased hippocampal volume during periods when they need to bury and retrieve cached food. Is there an equivalent change in humans? A professional requirement for London taxi drivers is a detailed knowledge of the tortuous streets of that large city. Interestingly, these taxi drivers have an increase in volume of the posterior hippocampus relative to a matched control population. Moreover, the longer they have been taxi drivers, the more pronounced is this change.

F. Nondeclarative Learning and the Amygdala

Previously, we used the example of a microwave bell and mouth watering to introduce nondeclarative learning. Physiologist Ivan Pavlov was the first to investigate classical conditioning. His studies showed

that by repetitive pairing of a conditioning stimulus (bell) with an unconditional stimulus (food), dogs salivated at the sound of the bell. It was Pavlov's idea that the cerebral cortex was involved in classical conditioning.

Studies with animals show that the amygdala is involved in classical conditioning. In this case, the pairing of a tone with a foot shock elicits a freezing response when the tone is presented alone. That is, the tone predicts the presentation of the shock. Neurons in the amygdala respond to a shock-related tone only. Furthermore, damage to the amygdala eliminates the physiological changes associated with the freezing response. What is the circuit involved in this type of information processing? Studies by LeDoux and coworkers suggest that auditory information arriving at the auditory cortex travels to the basolateral nucleus and the central nucleus of the amygdala; first the auditory information arrives at the basolateral nucleus, then to the central nucleus. The central nucleus then sends information to the hypothalamus, which mediates autonomic responses (increased in heart rate); periaqueductal the gray area the brain stem, which mediates the actual emotional response (freezing); and the cerebral cortex, which mediate the emotional experience.

In the next section, we show that information processing can also be studied at the cellular and molecular level. Indeed, as suggested by Semon, Santiago Ramon y Cajal, and William James, the engram is formed by physiological changes in neurons. Neurons in both the hippocampus and the amygdala display physiological changes that are believed to mediate information storage.

G. Long-Term Potentiation Is a Form of Synaptic Plasticity That May Be Involved in Information Processing

As addressed earlier, Hebb increased our understanding of how information processing occurs in neurons by suggesting that activity in groups of neurons mediates memory formation. Because neurons communicate with each other only at synapses, changes in activity of groups of neurons likely occur at the synapses, and therefore memories are stored through LTP. The NMDA receptor is involved in forming LTP. Activation of the NMDA receptor requires both the release of the neurotransmitter glutamate and stimulation of the postsynaptic cell. That is, the NMDA receptor must be activated by glutamate and simultaneously there must be sufficient depolarization of the postsynaptic membrane to relieve a magnesium block in the NMDA-associated ion channel, which allows entry of the ion calcium (Ca^{2+}) into the postsynaptic terminal. Ca^{2+} activates a number of changes in the cells. Thus, this receptor acts as a coincidence detector, informing the neuron that it was activated in rapid succession. This is an example of a cellular mechanism that may be involved in information processing.

IV. RELATIONSHIP BETWEEN LTP AND LEARNING

The first evidence to associate LTP with learning was derived from studies that used drugs to block the activity of the NMDA receptor. The rationale behind these studies was that the blockade of the NMDA receptor would prevent changes in synaptic function. In the most comprehensive studies, the drugs were administered directly into the ventricles. This procedure allows the drugs to diffuse to the hippocampus in addition to other brain areas. The behavioral task investigated was the Morris water maze task, which is a spatial learning task. Blocking the NMDA receptor indeed affected spatial learning and also blocked LTP. Presumably, the blockade of changes in synaptic function prevented learning. As expected, LTP is also found in the amygdala, and NMDA receptor blockade in the amygdala prevents both nondeclarative learning and LTP.

A. LTP or Memory Formation Causes Changes in Gene Expression

Previously, we presented LTP as a model of information processing that leads to memory storage. We also considered the idea that information processing is associated with changes in synaptic function. However, changes in synaptic function alone cannot account for the permanence of memory storage. That is, there must be other mechanisms associated with information processing that allow for memories to remain for days to years. Many believe that the permanent storage of information involves a signal cascade that begins at the level of the synapse and includes the activation of a receptor and the activation of cellular changes known as second messenger

systems. Then, at the level of the cell body, it involves the transcription of genes that have a short-lasting function (immediate early genes) and the activation of genes that have a longer lasting function. The activation of the latter may account for a more stable memory. The molecular biological revolution has taught us that long-term changes of cell function, as must occur in long-term memory storage, are controlled by gene expression and resultant protein production. Interestingly, almost every aspect of the signal cascade has been investigated in both learning and LTP.

How might changes in the synapse and then at the nucleus be involved in LTP and learning? If one accepts the hypothesis that long-term potentiation subserves memory, then one can alter information processing either by inducing LTP or by training an animal so that it forms a memory (task acquisition). After either of these experimental perturbations, electrophysiological and morphological responses are known to occur in the brain. For example, the electrical properties of a neuron could be changed by altered production of ion channel proteins, and dendritic morphology in neurons could be altered by changes in cytoskeletal and/or surface adhesion proteins.

Changes in gene expression can be measured in a number of ways. In "fishing" for the expression of genes that are unknown, two techniques have been used successfully: subtractive hybridization and differential display. Subtractive hybridization permits a mathematical subtraction of gene expression in control tissue from expression in treated or perturbed tissue, allowing isolation of mRNA specifically from genes that have been upregulated or downregulated. Differential display involves amplification of both control and perturbed sequences and subsequent separation of the amplified products, which permits one to determine directly which sequences have been increased. However, with the completion of the Human Genome Project and with other entire mammalian genomes in reach, a technique called DNA microarray analysis has become increasingly important. In principle, DNA microarray techniques can measure expression of all known genes at one time, and it likely will become the method of choice for some investigations. DNA microarray techniques have already been applied to studies of alcoholism, Alzheimer's disease, schizophrenia, and multiple sclerosis; they have examined the changes wrought in the brain by aging, sleep, environmental enrichment chemically induced seizures, and amphetamine.

Indications so far are that these conditions cause altered expression of a small subset of genes (approximately 12). One possibility is that this altered expression is sufficient to cause profound changes in a complex system such as the brain. Another possibility, which has yet to be examined, is that DNA microarray analysis of a relatively large piece of tissue containing millions of cells misses the changes that occur in a small subset of those cells. After all, neuroanatomy tells us that changes in highly localized nuclei or tracts within the brain will have profound effects. Other molecular biological techniques have been used to examine gene expression in a single cell, and it seems likely that DNA microarray analysis will be adapted to provide the same kind of specificity.

B. Restriction or Enhancement of Gene Expression Changes LTP and/or Memory Formation

Previously, we considered perturbing brain function and measuring the consequent changes in gene expression. Conversely, one can perturb normal gene expression and determine what effect this has on the ability of the brain to support LTP or the ability of the animal to acquire a task. For instance, single genes, controlling what are hoped to be specific events within cells, can be eliminated and the resultant effect can be studied simultaneously in whole animals minus one gene (so-called knockouts) for LTP and learning.

One reason to target genes is that these genetic procedures have the potential to overcome the current limitations of pharmacology. As mentioned previously, drugs used to block the NMDA receptors affected not just the NMDA receptors located in the hippocampus but also those all over the brain. In addition, drugs used to block second messenger systems are not specific. Thus, the idea of gene manipulation is to delete genes in local areas. Homologous recombination, a process that gives rise to natural diversity, is used to create knockout animals. To use this technique *in vitro*, one must know the DNA sequence bracketing the gene in question; also, the efficiency of recombination is low. In the first studies of genes related to LTP and learning, an area of focus was proteins known as kinases. Kinases are important because they activate other proteins (e.g., receptors). This presented a problem because the kinase family is composed of a number of subtypes, which appear to have varied functions. However, we can selectively impair the

function of a specific kinase isoform by using knockout mutants.

The first kinase targetted was α-CaMKII, the type II α-calcium/calmodulin kinase. This kinase is important because it is activated when calcium enters through the NMDA receptor. Mice lacking this kinase had difficulty learning the Morris water maze. In the mutant mice, only the probability of LTP induction was altered; LTP induction was not abolished.

Protein kinase C is another kinase investigated using the knockout technology. In these mice the probability of LTP induction was reduced in the mutants much as it had been in previous studies employing knockouts. However, if the mutant mice were first treated with low-frequency stimulation, then the LTP was indistinguishable from that observed in wild-type controls. Regarding behavior, there were interesting findings in these mice. The mutant mice learned the Morris water maze at the same rate as did the normal mice; however, they had a deficit in contextual fear conditioning, a task that requires intact hippocampal function.

One potential function of kinases is to activate transcription factors, which are molecules that activate other genes and are found in the nucleus. Transcription factors are interesting because they mediate cellular and molecular mechanisms downstream of receptor activation (i.e., the cell body). One transcription factor involved in learning and memory is CREB. Previous studies indicated that better learning occurs when spaced trials are given as opposed to massed trials. One group used this finding to investigate learning in mice lacking CREB. Unique to these animals is impairment in long-term but not short-term memory. In this elegant study, it was found that in a number of behavioral tasks, including contextual conditioning, socially transmitted food preferences, and spatial learning, there was no impairment in learning when the animals were trained using spaced trials. The impairment observed with massed trials is explained by the fact that there is insufficient CREB to activate the long-term regulatory genes, whose proteins are essential for long-term memory. The use of spaced trials allows a limited amount of CREB to initiate sufficient transcription to allow long-term memory. These findings are important because they indicate the complexity that is inherent in the mechanisms involved in information processing in the brain. That is, since the deletion of CREB occurs throughout the entire animal, it is difficult to separate its primary effects in the central nervous system from effects outside the central nervous system. Since the animal's entire development occurs in the absence of the gene product, a viable adult likely indicates that compensatory changes have been made in the levels of other gene products. DNA microarray analysis provides a means of measure for these compensatory changes.

Recently, second-generation knockouts have been created in which the gene deficiency can be localized to a specific region of the brain or can be turned on or off by treatment with a specific drug. Spatial localization within the brain is performed using two distinct genetically engineered animals. In one animal, part of the target gene is replaced by an identical sequence that is flanked by short, highly specific sequences of DNA (called lox sequences); these sequences do not affect gene function. In the second animal, the gene for a bacterial enzyme called Cre is inserted under the control of a promoter that is known to be activated in the brain. For example, the enzyme calcium/calmodulin kinase is highly expressed in the forebrain; therefore, inserting the promoter sequence for calcium/calmodulin kinase in conjunction will cause a high level of Cre expression. When these two mice are mated, some progeny will possess both the Cre enzyme and the lox-bracketed gene. The normal function of the Cre enzyme is to locate two lox adjacent sequences and join them together, excising the portion of the DNA that lies between them. In so doing, Cre inactivates the target gene. This inactivation will occur only in cells that express Cre, and since the expression of Cre is different among strains of engineered mice, workers were able to obtain a strain in which expression was highly localized to the principal cells of hippocampal region CA1. Studies accomplished with the first target gene—NMDAR1 or subunit 1 of the NMDA receptor—illustrate the elegance of this system. Conventional first-generation knockouts of NMDAR1 die shortly after birth. Since normal calcium/calmodulin kinase expression (and therefore Cre expression) does not occur until later in development, the second-generation knockouts have NMDAR1 present during critical early periods and are viable adults. However, the adults are deficient in long-term potentiation in area CA1 and have a greatly reduced ability to learn a spatial task.

The genetic techniques mentioned previously are all implemented in the mouse. Although behavioral training of mice is well established, the induction of LTP *in vivo* is technically more challenging in these smaller animals. The advantage of antisense oligonucleotides is that they can be used in the rat, which is the neurophysiologist's animal of choice. There is still no satisfactory explanation as to why this technique works, but it is particularly successful in the brain.

The target in this case is not the gene but the transcribed messenger RNA (mRNA), which directs protein synthesis. An oligonucleotide sequence approximately 20–25 bases long is synthesized that is complementary to part of the target mRNA. This antisense oligonucleotide is injected stereotaxically into the brain, where it enters the cells. Within the cytosol the antisense binds specifically to the target mRNA. The resultant nucleic acid duplex is vulnerable to cellular enzymes (e.g., RNase H) that cut both strands, inactivating the target mRNA. One disadvantage is that to ''knock down'' the levels of proteins that are constitutively expressed, a number of injections have to be given, often over several days. The antisense technique is best used against proteins that are expressed transiently, such as immediate early genes (IEGs). A further disadvantage is that to ensure that any effects seen are due to the specific antisense sequence used, it is useful to use more than one control sequence.

C. Transgenic Mice

As mentioned previously, the predominant form of LTP in the brain is dependent on the NMDA receptor. Functional NMDA receptors comprise more than one subunit and usually more than one type of subunit. For example, NMDAR1 subunits seem to be absolutely required for the survival of mice and are capable of assembling to form a working receptor. However, most receptors also contain NMDAR2 subunits, of which these are four different kinds (2A–2D). The NMDAR2 subunits do not form functional receptors on their own but modify the characteristics of NMDAR1 when they associate with it; each of the four subtypes combines with NMDAR1 to produce a receptor with slightly different characteristics. In the rat, NMDAR2A and NMDAR2B are strongly expressed in the hippocampus. The hippocampus has essentially no NMDAR2C (which is predominantly located in the cerebellum) or NMDAR2D (which is found in the thalamus and hypothalamus). Mice with targeted disruptions (knockouts) of NMDAR1 or NMDAR2B show no gross anatomical abnormalities in the brain but die after birth of respiratory failure and impairment of suckling response, respectively. Of the hippocampal NMDAR2s, electrophysiological studies suggest that NMDAR2B should be more effective than NMDAR2A in causing LTP. In a recent study, a transgenic mouse overexpressing the NMDAR2B was created and evaluated; the researchers called their transgenic ''doogie'' after a precocious fictional medical student Doogie Hauser. These mice show enhanced retention of spatial memory and of both context (hippocampal-dependent) and cued (hippocampal-independent) fear conditioning, and they were dubbed ''supermice'' by the media. However, these mice show faster extinction (i.e., forgetting) of fear conditioning and enhanced novel-object exploratory responses, which are not traits likely to enhance survival in the wild. Between 20 days of age and adulthood, the amounts of NMDAR2A and NMDAR2B in the hippocampus decline slightly. A better name for this transgenic mouse might be Peter Pan since the decreased expression of the 2B subunit in adulthood is counteracted by overexpression, keeping the complement of 2B in a ''child-like'' state.

D. The Temporal Control of Gene Expression

Specialized proteins bind to the control regions of a gene (called promoters or operators) and increase or decrease the probability of transcription. In bacteria, genes to counter tetracycline antibiotics are kept switched off by the binding of a repressor protein to the appropriate operon (bacterial control sequence). Tetracyclines bind to the repressor and release it from the operon, permitting gene expression. A useful mutation of this system works in reverse: An activator of transcription does not bind the control region unless tetracycline is present so that antibiotic switches the system on. Insertion of these bacterial genes into mammals can be used to provide precise temporal control of gene expression. In one mouse, the activator protein (called rtTA) is placed under the control of calcium/calmodulin kinase, so it is expressed abundantly, but only in the forebrain. A second mouse has a transgene inserted that is under the control of tetO, the region of DNA to which rtTA binds. Progeny of these two mice will show elevated gene expression in the forebrain in response to a tetracycline-class antibiotic, such as doxycyclin. A proof-of-concept study used calcineurin as the transgene. Calcineurin is a protein phosphatase, and as a simplistic model it predicts that if protein phosphorylation by specific kinases is critical for the induction of LTP and for memory, then overexpression of phosphatases such as calcineurin should neutralize the kinase effects by dephosphorylating some of the critical proteins. This is indeed the case: The mutant progeny show normal LTP and a

normal ability to recall their spatial training until doxycyclin is added to their food. Doxycyclin treatment significantly decreases both LTP and the ability to retrieve (recall) spatial training. When doxycyclin is withdrawn and the antibiotic clears from the animals' bodies, their LTP and ability to recall training revert to normal. Doxycyclin treatment does not alter the characteristics of the parental strains, which contain the individual components of the expression control system.

V. THE STORY OF ARC: A MOLECULAR BIOLOGICAL EXPLORATION OF LTP AND MEMORY

IEGs are expressed rapidly (within 30–60 min) after a stimulus. They include transcriptional factors, which act as intracellular signals, switching the expression of other genes on or off. However, IEGs also include proteins necessary for rapid extracellular signaling or for structural changes.

As an example of how molecular biology can be used to probe the processes underlying memory, consider the discovery and characterization of an IEG called *Arc* (activity-regulated cytoskeleton-associated protein) by Lyford and associates and Link and associates. Messenger RNA was extracted from the hippocampi both of control rats and of rats subjected to electroconvulsive stimulation. Subtractive hybridization was performed to isolate genes that were upregulated by stimulation. In order to confirm that the identified Arc DNA sequence could code for a functional protein, it was subjected to *in vitro* transcription/translation, which uses a mixture of the appropriate isolated cellular components to produce protein in a cell-free system. The predicted protein sequence of Arc indicated that it had no signal sequence for export outside the cell. Nor did it have long hydrophobic sequences or glycosylation sites, which are often diagnostic of a membrane protein; long hydrophobic sequences permit a protein to span the cell membrane, and glycosylation is a common feature of the extracellular domains of proteins. Sequence comparisons showed that one-half of Arc is closely related to α-spectrin, a cytoskeletal protein. The other half interacts with calcium/calmodulin-dependent protein kinase II. Measurement of Arc mRNA using Northern blot analysis showed that the unstimulated brain was enriched in Arc relative to other tissues, and that following electroconvulsive stimulation hippocampal Arc was elevated between 30 min and 8 hr, returning to basal levels by 24 hr.

When a gene product shows increased expression, the following is one of the first questions asked: Where in the tissue of interest is this expression taking place? This question is answered by *in situ* hybridization: Tissue slices are taken and exposed to radioactive nucleic acid sequences that are complementary to the mRNA of interest and therefore bind to it specifically. Exposure to film results in an autoradiograph that localizes the mRNA to anatomic structures and even to subcellular regions. At 15 and 30 min after electroconvulsive stimulation, Arc was expressed principally in the cell bodies of the dentate gyrus granule cells. After 1 hr, Arc was also found in the dendrites (see Fig. 2).

In order to confirm that the expression of mRNA is paralleled by protein production, bacterial expression systems are used to generate large quantities of pure protein from the mRNA sequence. Immunizing animals with the protein yields antibodies that react specifically with the protein. These antibodies can be used to localize proteins in a tissue slice, a technique called immunohistochemistry. The general pattern of Arc protein localization in unstimulated tissue matched that obtained by *in situ* hybridization. Four hours after electroconvulsive stimulation the number of granule cells in the dentate gyrus that expressed Arc protein increased from 1–2 to nearly 100%. Moreover, protein was not expressed in axons or synaptic terminals but was expressed in cell bodies and also in dendrites. In these regions, Arc protein was subplasmalemmar (i.e., just inside the cell membrane), and this localization resembles that of spectrin and F-actin, important cytoskeletal components. Since Arc resembles a cytoskeletal protein in structure and localization, its ability to interact with other cytoskeletal proteins was measured. Arc did not interact with highly purified actin, but it did bind to a less purified preparation containing actin and actin-associated proteins.

The behavior of Arc, a protein rapidly synthesized in the nerve dendrites and capable of interacting with the cytoskeleton, would be interesting if it occured only in response to a pathological stimulus such as electroconvulsive stimulation. Does synthesis of Arc occur under conditions that more closely resemble normal physiology? Stimulation that leads to LTP induction *in vivo* causes changes in Arc mRNA expression similar to those seen in response to electroconvulsive stimulation. Blocking the induction of LTP in this model with an NMDA receptor antagonist also blocks the changes

in Arc expression. Further analysis demonstrates that Arc is actually localized to the dendrites that are stimulated. As dentate granule cells extend into the molecular layer, they are contacted first (proximal to the cell body) by commissural/associational neurons, then (at an intermediate distance) by the medial perforant path, and finally (distal from the cell body) by the lateral perforant path. When electroconvulsive stimulation is used to generate Arc, there is no specific localization within the granule cell dendritic tree. However, when the afferent paths are stimulated separately, Arc mRNA is increased in the appropriate region of the dendritic tree. Immunohistochemistry confirms that Arc protein is localized in a similar way.

Are the increased expression of Arc and its subsequent localization distinct processes or are they related? Also, how is Arc targeted? As mRNA alone, with translation to protein occurring only at its dendritic destination, or as a polyribosome complex, with the newly synthesized protein required for localization? When Arc expression is increased by electroconvulsive stimulation (a process that causes no specific localization in the dendritic tree), subsequent high-frequency stimulation of the medial perforant path causes appropriate localization of the Arc mRNA within the dendritic tree, suggesting that localization is a process distinct from increased expression. Moreover, localization occurs in the presence of the protein synthesis inhibitors cycloheximide and puromycin, demonstrating that the signal sequence for localization is in Arc mRNA. An important negative control for all these data is that no other IEG so far discovered shows dendritic localization: mRNAs for these IEGs are retained in the cell body. A logical speculation is that Arc is required for the postsynaptic remodeling that occurs in the dendrite following high-frequency stimulation and that is necessary for synaptic plasticity.

It is known that the survival time of cellular mRNA is markedly decreased if it encounters a sequence that is complementary (antisense). The precise mechanism for this decreased survival time is incompletely understood, although it likely involves RNase H, an enzyme that specifically cleaves double-stranded RNA. If our speculation about the function of Arc is correct, treatment with oligonucleotide sequences (ODNs) antisense to Arc should change the pattern of LTP seen as a result of high-frequency stimulation. This proves to be the case: Direct injection of Arc antisense ODNs into the dorsal hippocampus significantly decreases the induction of LTP in the perforant path to dentate gyrus in the awake, behaving rat. In this experiment animals are injected in one hippocampus with antisense ODN; as a control, a scrambled-sequence ODN is injected in the other hippocampus. Both sides show the same initial LTP, but from 4 hr up to 5 days a significant difference is apparent, with antisense-treated hippocampi showing only 30–60% of control LTP.

Given the likely correspondence between LTP and memory, it is interesting that the injection of Arc antisense ODNs also impairs the ability of rats to retain a spatial memory. In this experiment, rats receive bilateral injections of either Arc antisense ODNs or the scrambled control; three hours later, they receive training in the spatial Morris water maze for 1 hr. It is known that this spatial training, like the different forms of electrical stimulation mentioned previously, normally causes an increase in Arc expression in the hippocampus. Both sets of rats acquire the task successfully, but when they are tested 2 days later control animals can successfully distinguish the target location, whereas the rats treated with Arc antisense ODNs cannot. Since Arc-dependent consolidation of the memory likely occurs between 1 and 2 hr after training, antisense ODNs can actually be injected immediately after the spatial training and still interfere with the 2-day long-term memory; when injected 8 hr after training, the antisense has no long-term effect.

It is well-known from electrophysiological experiments that the hippocampus contains "place" cells, whose electrical activity increases when the animal is in a particular spatial environment. If Arc is involved in spatial memory, then whenever the animal is placed in a particular environment Arc expression should be increased in the corresponding place cells. Placing the animal in a different environment should activate Arc in a different set of place cells. As noted previously, microscopic localization of Arc in tissues can be achieved using *in situ* hybridization; if the visualization process uses fluorescent instead of radioactive probes, the acronym FISH (fluorescence *in situ* hybridization) is used. The method used to measure the increased expression of Arc very shortly after environmental stimulation is denoted compartmental analysis of temporal FISH (catFISH). Like all mRNAs, Arc originates in the nucleus but is rapidly exported to the cytoplasm of the cell body. Unlike other IEG mRNAs, Arc is then cleared from the cell body by targeting to the dendrites. By examining Arc's distribution relative to the nucleus, one can determine how recently a given cell was stimulated. For example, taking an animal from its home cage and placing it in a particular environment (e.g., environment A) 5 min before sacrifice yields a subpopulation of hippocampal cells

with only nuclear labeling. If the sacrifice of the animal is delayed for more than 30 min, then there is a subpopulation of cells with only cytoplasmic labeling. How do we know that these subpopulations in different animals are comparable, or even that they are truly place cells? If an animal is exposed to environment A, returned to its home cage, and then reexposed to environment A immediately before sacrifice, a single subpopulation of cells shows both nuclear and cytoplasmic labeling—that is, the *same* subpopulation has been stimulated twice, at different times in the past. In hippocampal region CA1, this subpopulation represents about 40% of all cells. A completely different pattern is seen if the animal is exposed to environment A, returned to its home cage, and then exposed to environment B. Only 16% of cells in the CA1 show both nuclear and cytoplasmic labeling. If another 40% (fraction 0.4) of CA1 cells are activated by any given environment, then the number of cells stimulated in common by two different environments will be $(0.4)^2$ or 0.16 (i.e., 16%). We predict that 24% (40%−16%) of cells will be uniquely stimulated by environment A and the same number by environment B. The cells corresponding to environment A will show exclusively cytoplasmic labeling because of the delay prior to sacrifice. The cells corresponding to environment B will show exclusively nuclear labeling because no delay ensued. The experimentally determined values are 22 and 23%, respectively, showing the model to be internally consistent.

Given that the Arc story began in the dentate gyrus, it is interesting that although hippocampal regions CA1 and CA3 and the parietal cortex all show similar results using catFISH, the dentate gyrus shows essentially no Arc response to environmental change, or that ARC is not involved in storing spatial information in the dentate gyrus. This suggests that spatial data pass through the dentate gyrus without causing a change.

In summary, studies show that cellular and molecular events are the underlying mechanisms associated with the processing of information that leads to the permanent storage of memory. Interestingly, as suggested by Hebb, these cellular and molecular events are initiated by the activity between neurons.

See Also the Following Articles

ARTIFICIAL INTELLIGENCE • ATTENTION • CREATIVITY • HEURISTICS • INTELLIGENCE • LANGUAGE AND LEXICAL PROCESSING • LOGIC AND REASONING • MEMORY, OVERVIEW • NERVE CELLS AND MEMORY • NEURAL NETWORKS • NUMBER PROCESSING AND ARITHMETIC

Suggested Reading

Finger, S. (1994). *Origins of Neuroscience: A History of Explorations into Brain Function*. Oxford Univ. Press, New York.

Hebb, D. O. (1949). *The Organization of Behavior*. Wiley, New York.

Link, W., Konietzko, U., Kauselmann, G., Krug, M., Schwanke, B., Frey, U., and Kuhl, D. (1995). Somatodendritic expression of an immediate early gene is regulated by synaptic activity. *Proc. Natl. Acad. Sci. USA* 5734–5738.

Lyford, G. L., Yamagata, K., Kaufmann, W. E., Barnes, C. A., Sanders, L. K., Copeland, N. G., Gilbert, D. J., Jenkins, N. A., Lanahan, A. A., and Worley, P. F. (1995). Arc, a growth factor and activity-regulated gene, encodes a novel cytoskeleton-associated protein that is enriched in neuronal dendrites. *Neuron* 433–445.

Martinez, J. L., Jr., Barea-Rodriguez, E. J., and Derrick, B. E. (1998). Long-term potentiation, long-term depression and learning. In *Neurobiology of Learning and Memory* (J. L. Martinez, and R. Kesner, Eds.), pp. 211–246. Academic Press, San Diego.

Squire, L. R., and Kandel, E. R. (1999). *Memory: From Mind to Molecules*. Sci. Am., New York.

Inhibition

AVISHAI HENIK
Ben Gurion University of the Negev, Israel

THOMAS H. CARR
Michigan State University

GLOSSARY

anterior cingulate cortex A region of the medial frontal lobe involved in a variety of cognitive, motor, and emotional–motivational functions, including Brodmann's area 24 and perhaps the closely associated area 25.

attention A class of cognitive processes involving prioritization or selection among stimuli to be processed, tasks to be performed, or responses to produce.

frontal eye field A region centered at the intersection of the precentral sulcus with the superior frontal sulcus, that is, Brodmann's area 6. This region is involved in generation and control of eye movements.

reaction time The time elapsed from onset of the imperative stimulus to initiation of the subject's response.

saccade Rapid eye movement that changes the point of fixation from one location to another.

stimulus onset asynchrony The time elapsed from the onset of the first stimulus (or the cue) to the onset of the imperative stimulus.

Inhibition refers to mechanisms by which the nervous system suppresses information, restricts its use, or restrains its transmission from one place in the brain to another. This article addresses hypotheses about the role of inhibition at the systems level, considering the possible functional consequences of inhibitory processes for cognition and behavior. Ultimately, one might guess that at the neuronal level, inhibitory processes would be implemented via neuron-to-neuron communication in which release of inhibitory neurotransmitters reduces the probability of action potentials in postsynaptic target neurons of brain structures whose activity needs to be curtailed. As will be shown, however, reduction or suppression of activation is not the only way in which a systems-level inhibitory outcome can be achieved in cognition or behavior, although it is the most commonly proposed mechanism.

I. THE ROLE OF INHIBITION IN COGNITION AND BEHAVIOR

Cognition and the control of overt behavior rely on real-time orchestration of component cognitive processes or "mental operations." Each operation achieves a step in the sequence of steps leading from stimulus to response, intention to action, or thought to thought. A major distinction is made between reflexive or "automatic" operations and voluntary or "controlled" operations. The more automatic an operation is, the more able it is to occur without intention, needing only the appropriate stimulus conditions or information inputs to trigger it; to occur outside of

conscious awareness, without being noticed phenomenologically; and to run in parallel with other mental operations.

Automatic operations gain these properties either from genetic hardwiring or, more frequently, from repetition under relatively unchanging conditions. Hence, they sometimes occur as reflexes, and they become quite prominent in familiar situations and well-practiced tasks. Controlled operations are the opposite. The more voluntary or controlled an operation is, the more its execution is intentional, conscious, and demanding of serial attention. Controlled processes become prominent when dealing with novel situations to which reflexes and habits are poorly adapted or when pursuing particular goals in situations that are likely to trigger reflexes and habits that would produce incorrect or inappropriate behavior if unrestrained.

To achieve flexibility in dealing with both the familiar, unchanging aspects of the world and with novel events, it is important to make automatic and controlled processes work in concert. Inhibition provides a tool for curbing or regulating automated responses in the service of controlled assessment and reaction.

Just how fundamental a role is played by inhibition of automated responses can be appreciated by thinking about the developmental trajectory of reflexes across the life span. One of the central principles of neurology is that disease processes affecting higher brain centers, especially the cerebral cortex, are revealed by reappearance of primitive reflexes. The knee jerk of a normal person is tonically inhibited, but it can become hyperactive after damage to the spinal cord that interrupts descending inhibitory pathways from the motor cortex. The sucking reflex of infants disappears after the nursing years but can reappear in a patient with Alzheimer's disease. Presumably, with the development of the nervous system, these primitive reflexes are inhibited throughout adulthood but may be disinhibited and reappear due to nervous system insult.

Other reflexes remain active throughout the life span. Among the most common are reflexive orienting reactions that involve the automatic deployment of attention to a suddenly appearing visual stimulus or a loud sound. Because these are very common occurrences, such attentional reactions are frequent and do not always occur at convenient times. When controlled deployment of attention is required by task performance, disruptive reflexive deployments may need to be inhibited.

Analogous problems can arise in controlling responses that are not reflexes but have become automated through practice. Attention deficit and hyperactivity disorder is a persistent individual difference characterized by an impulsive inability to resist engaging in prepotent or automated actions triggered by task-irrelevant stimuli in the environment or task-irrelevant thoughts in the mind. Intrusion of an unwanted or inappropriate automatic response occasionally occurs in normal children and adults as well. This can be seen in "slips of action" in which distraction of attention while carrying out a low-frequency task can result in inadvertently executing a different task that is higher in frequency given the situation and stimulus environment.

Other inhibitory functions are aimed at suppressing unwanted or incorrect perceptions and thoughts rather than controlling attention or inhibiting overt actions. These extraneous mental representations may become activated through relatively automatic processes of generalization because they are similar to correct perceptions and thoughts, or they may become activated because the environmental situation is ambiguous and admits multiple interpretations, only one of which is relevant to the task at hand. It is obvious that the wrong interpretation can be reached in an ambiguous situation, and if so it will need to be replaced. Less obvious, perhaps, is the possibility that all the interpretations of an ambiguous stimulus might be computed relatively automatically at an early stage of processing on every occasion, with one chosen for further processing at a later stage. In either case, contextually inappropriate interpretations will be active at some point in task performance and will need to be put aside or eliminated. This may involve inhibition. In this regard, Sir John Eccles wrote the following in 1977:

> *I always think that inhibition is a sculpturing process. The inhibition, as it were, chisels away at the diffuse and rather amorphous mass of excitatory action and gives a more specific form to the neuronal performance at every stage of synaptic relay.*

In this article, we present several examples representing the types of inhibitory function we have just described. Some are examples of inhibition regulating the deployment of attention. Others are examples of inhibition enabling disengagement from ongoing or automatically triggered responding so that overt behavior can be under voluntary and strategic control.

Still others are examples of inhibition tuning and sharpening the representation of the stimulus produced by perception or the interpretation of the situation produced by higher cognitive processes.

II. ATTENDING TO A SPATIAL LOCATION

Orienting of visual attention to a point of interest is commonly accompanied by overt movements of the head, eyes, or body. Attending may originate at will, such as when we decide to look at a particular location where something of interest is expected, or it may originate reflexively without intention when something captures our attention, such as when we orient to a flash of light in the dark or to a movement in the periphery of our vision. In everyday life there are constantly competing demands on attention by the outside world as well as from internally generated goals. The need for mechanisms to arbitrate between these competing demands is straightforward: This must be done so that they can be integrated, prioritized, or selected among to provide coherent and adaptive behavior.

Michael Posner developed a paradigm widely employed to study visual spatial attention. Figure 1 shows the basic features of this paradigm. In a typical experiment, the subject is first presented with three boxes on a computer screen. A trial begins with a fixation cross at the middle box. After a short interval, one of the boxes may flash briefly, and a target (an asterisk) appears in one of the peripheral boxes. The subject is asked to respond as quickly as possible, by pressing a key, to the appearance of the target. Reaction time (RT) is measured (in milliseconds) from target onset until the subject's response. It is possible to study covert attention by asking the participants not to move their eyes and by measuring keypress responses (rather than saccade latencies).

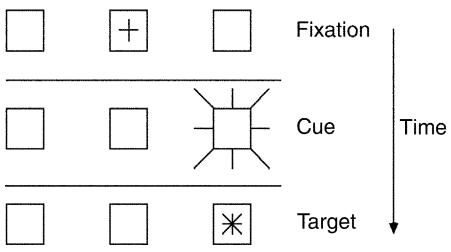


Figure 1 Basic paradigm for spatial attention, showing an exogenous cue and a valid trial.

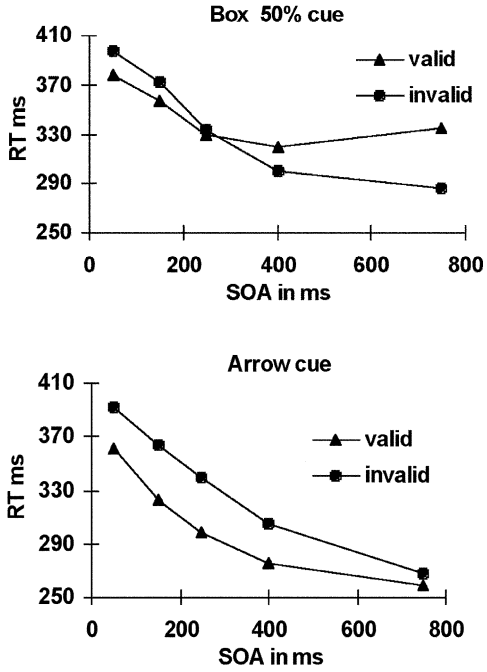


Figure 2 Typical time course for effects of a peripheral nonpredictive luminance cue (box 50% cue) and a central predictive arrow cue. The task is a simple RT key press response to the appearance of the target. Mean detection RTs are presented as a function of cue–target interval (SOA) for valid and invalid cue conditions.

The cue may summon attention to the target location, in which case it is a *valid* cue, or it may summon attention to the wrong location, in which case it is an *invalid* cue. For volitional goal-directed shifts of attention, often called *endogenous* shifts of attention, a central arrow serves as a cue and predicts where the target is likely to occur in most trials. That is, in 80% of the trials the target will appear at the valid cue location and in 20% of the trials the target will occur at the invalid cue location. For reflexive stimulus-driven shifts of attention, often called *exogenous* shifts of attention, the peripheral luminance change that serves as a cue has no predictive value with respect to where the target will occur (e.g., in 50% of the trials the target will occur at the valid cue location and in 50% of the trials the target will appear at the invalid cue location). In order to measure the effectiveness of the cue in summoning attention, researchers have manipulated the time interval between cue onset and target onset [the stimulus onset asynchrony (SOA)]. The typical effects of the two types of cues are depicted in Fig. 2. The top panel of Fig. 2 shows the time course of a

nonpredictive peripheral luminance cue on summoning reflexive exogenous attention, and the bottom of Fig. 2 shows the time course of a predictive central arrow. These results were achieved in a covert attention experiment, with no movement of the eyes. The two cues are similar in the sense that in both types of cues the facilitory effects begin at 50 msec. However, there are differences. With endogenous cueing, the facilitory effect appears to be more sustained. With peripheral cueing, the advantage at the cued location changes, after a few hundred milliseconds, into an inhibition resulting in longer RTs for the cued location. No such inhibition is seen with endogenous cueing. These features of Fig. 2 will be discussed next, each related to a different mechanism of inhibition generated by the deployment of spatial attention.

A. Noise Reduction and Suppression of Competing Stimuli

Figure 2 shows that for about 200 msec after an exogenous cue and for 600 msec or longer following an endogenous cue, RT to detect the onset of a target is shorter when a target appears at the valid cue location compared to the invalid cue location. This RT effect indicates that attention has been drawn to the location of the cue, thereby facilitating the detection of the target at the valid location compared to the invalid location.

It is generally agreed that the presence of attention due to valid cueing increases the signal-to-noise ratio or sensitivity of information processing at the cued location. Of course, this could happen either if the signal from the attended location were enhanced (a facilitatory process) or if noise that might interfere with that signal were reduced or suppressed (an inhibitory process). The available evidence supports a combination of these two effects.

First, valid cueing can increase sensitivity even when a single target appears in an otherwise blank visual field. Since there are no other stimuli around to produce any noise or distraction, it would appear that attention is enhancing the signal from the attended location. However, the effect of cueing a target in an otherwise blank field is often small and sometimes disappears. The impact of cueing is usually greater when the target is accompanied by distracting stimuli at other spatial locations. From such evidence it has been argued that the primary contribution of spatial attention, especially in real-world visual environments, is to reduce noise or cross talk from unattended stimuli that if unsuppressed would interfere with processing the attended stimulus.

Direct evidence for inhibition of cross talk-based interference from unattended stimuli comes from single-cell recording in extrastriate cortical areas of the ventral object-processing pathway of monkeys. When two stimuli are present in the receptive field of a cell in one of these areas (e.g., V4 or TEO), the cell's response is less than if only one of them was present. This reduction in activity shows that the two stimuli interfere with each another's ability to activate the cell. However, if the monkey attends to one of the stimuli for the purpose of performing an experimental task, the cell's response returns to approximately the same level as if the attended stimulus were presented alone. This restoration of responsiveness shows that the unattended stimulus is being filtered out of the cell's receptive field and hence no longer interferes with processing the other stimulus. Analogous evidence that humans have a similar attention-driven mechanism of noise reduction through suppression of competing stimuli comes from functional magnetic resonance imaging (fMRI) experiments reported by Kastner and colleagues.

B. Inhibition of Return

A second important feature of Fig. 2, (seen only in the top), is that when SOA's following an exogenous cue are longer than 2–300 msec RT to detect a target is *longer* when a target appears at the valid location compared to the invalid location. That is, facilitation is transformed into inhibition. This is a standard outcome of experiments with nonpredictive cues, in which subjects have no reason to use the cue to guide attention and would prefer to keep attention focused at fixation or spread diffusely across the display.

Michael Posner and Yoav Cohen analyzed the effect of nonpredictive cues at the longer SOA durations, now known as the inhibition of return (IOR) phenomenon. Facilitation changes to inhibition (i.e., IOR) 200–300 msec after cue onset. IOR lasts for about 3 or 4 sec; it works in environmental rather than in retinal coordinates, it is not generated by endogenous cues unless the oculomotor system has been activated, and it declines as the distance from the original cued location increases.

What is the explanation for this counterintuitive transformation of facilitation into inhibition? Exogenous cueing commonly produces a reflexive shift of attention to the cued location or object (producing the

early facilitation in Fig. 2). Even when people are asked not to pay attention to the cued location or it is in their favor to ignore the cued location, they find it difficult to avoid reacting to a peripheral luminance change and often cannot refrain from orienting to this kind of cue. As much as such efficient and rapid orienting is important for predatory and defensive behavior, voluntary control of reflexive orienting and the ability to strategically search the environment are also critical for survival. It appears that IOR is a mechanism that enables the organism to disengage from reflexive orienting and switch to the control of a more voluntary attentional system. How does it work? It seems that a location (or an object) that was recently cued or searched is tagged and IOR biases attention away from responding to events occurring at the tagged locations. Avoidance of tagged locations encourages search in new spatial locations. Accordingly, IOR seems to be a mechanism that supports efficient foraging behavior, which involves strategic search of the environment and use of knowledge about previous searched locations or objects.

Several lines of evidence point to the midbrain superior colliculus (SC) as the neural substrate for the implementation of IOR. IOR is abnormal in patients with damage to the SC. This has been shown in patients with midbrain degeneration due to progressive supranuclear palsy and in a patient with unilateral lesion to the SC. IOR is preserved in the presence of hemianopia, in which only the extrageniculate pathways are available to process visual information. It is present in newborn infants in whom the geniculostriate pathway is not yet developed. It is generated asymmetrically in the temporal and nasal visual fields. The temporal hemifield has stronger collicular representation than the nasal hemifield; accordingly, for monocular presentations of stimuli, IOR is larger for stimuli presented to the temporal than to the nasal hemifield. However, it seems that cortical structures play a role in the generation of IOR. In particular, the parietal lobe has been suggested as a structure that conveys the spatial coordinates of the tagged locations to the SC.

C. Inhibition of Reflexive Orienting

As suggested previously, other cortical structures in addition to the parietal lobes regulate activation of the SC. In 1966, James Sprague reported that occipitotemporal cortical ablation in cats produced stable hemianopia. However, visually guided behavior was restored by ablation of the dorsal midbrain contralateral to the cortical lesion. He noted,

> *This initial hemianopia is apparently due to depression of function of the colliculus ipsilateral to the cortical lesion, a depression maintained by influx of inhibition from the crossed colliculus. Thus, removal of the contralateral tectum, or splitting of the collicular commisure, abolishes this inhibition and allows the return of function in the ipsilateral colliculus, and with it the recovery from hemianopia.*

This phenomenon is called the Sprague effect. A reversed version of the Sprague effect can be found with lesions to the frontal lobe, which will be described later.

The SC is involved in triggering reflexive saccades, whereas cortical mechanisms are needed for generating voluntary saccadic eye movements under strategic guidance. Early work showed that patients with frontal cortex lesions have difficulty executing saccades in response to verbal commands. When these patients scan pictures their eye movements are controlled by external stimuli rather than by instructions. In addition, lesions of the frontal lobes produce a deficit in inhibiting ''reflexive glances.'' The patients seem unable to resist moving their eyes in the direction of a peripheral stimulus even when instructed to prevent such eye movements. It appears that the frontal eye fields (FEFs) are critical components of the frontal circuitry underlying inhibition of reflexive saccades and endogenous control of eye movements. Unilateral lesions of the FEF may shorten latencies of reflex saccades to targets in the contralesional field. Moreover, single cell recordings show that the FEF contains cells that respond in temporal correlation with purposive saccades even in the absence of an exogenous visual stimulus.

In line with these reports, Henik, Rafal, and Rhodes found that lesions to the FEF have opposite effects on endogenously activated and visually guided saccades to external stimuli. In a typical trial of an experiment examining saccades in FEF patients, subjects saw a visual cue (''get ready'') followed by a target (''go''). The cue was informative (a small arrow) or neutral (a double-headed arrow). There were two types of target go signals: an asterisk appearing at the periphery (used to elicit reflexive, exogenously triggered saccades) or a large arrowhead in the center of the display pointing left or right (used to measure voluntary, endogenously generated saccades). The efficacy of saccade

preparation was measured as facilitation in saccade latency in the informative cue condition compared to the neutral cue condition. Patients with FEF lesions presented slower endogenous saccades (in response to the central arrow) to the contralesional field, whereas exogenously triggered saccades (in response to the peripheral targets) were faster to the contralesional field (a reversed Sprague effect). These results indicate that the FEF is involved in generating endogenous saccades and therefore lesions in this region increase their latency. In addition, the FEF is involved in visually guided saccades through inhibitory connections to the SC. It seems that the FEF has the opposite effect on the extrageniculate visual system from occipital lesions. That is, whereas occipital lesions reduce activation of the ipsilesional SC (producing increased activation of the contralateral SC), FEF lesions disinhibit the ipsilesional colliculus, producing suppression of collicular function contralateral to the cortical lesion.

Taken together, this evidence indicates hierarchical control over eye movements. A relatively automatic system depending on SC is stimulus driven, responding to the appearance of new objects in the visual field and to movement, especially in the periphery. This system is also sensitive to sound via interactions between superior and inferior colliculus. Barry Stein and colleagues have shown that reflexive saccades and head turning are elicited when a movement that is subthreshold for triggering an orienting response is accompanied by a sound from the same spatial location that is also too weak by itself to cause orienting. The second level of control is cortical, depending on FEF. It exerts endogenous control rather than responding reflexively to stimulus inputs, implementing voluntary eye movements, and inhibiting the responses of the automatic system when such responses would conflict with the intended voluntary movement.

Interactions between automatic orienting and voluntary control have been studied extensively in the *antisaccade* task. This task requires subjects to move their eyes *away* from a stimulus that appears abruptly in the visual field—a stimulus that would ordinarily *attract* a reflexive saccade. Endogenous control is far from perfect in this difficult task. Errors occur in which the eyes move toward the abruptly appearing stimulus, and changes in error rate with experimental manipulations and subject characteristics can be used to diagnose the effectiveness of endogenous control. More evidence is provided by the latencies of correct responses away from the stimulus, which are slower than either reflexive saccades toward the same stimulus or voluntary saccades toward stimuli that do not elicit reflexive saccades (either because they are already present in the visual field or because they are new but their onset is ''ramped up'' slowly to eliminate visual-onset transients). In general, factors that increase the perceptual salience of the stimulus, such as brightness, tend to increase errors and slow latencies, as do factors that reduce the subject's ability to concentrate on the task, such as attentional load, and factors that reduce the subject's ability to predict when effortful control will be needed, such as randomly varying foreperiods prior to stimulus onset. Furthermore, patients with frontal lobe damage have great difficulty with the antisaccade task.

III. ''SELECTION FOR ACTION'' AND CONFLICT RESOLUTION

The antisaccade task involves more than inhibiting the reflexive response toward the abruptly appearing stimulus. The subject must successfully implement another response at the same time the reflex is being inhibited. This second response is less well learned or less natural than the one that must be inhibited. Hence, the competition created by the automatic response is powerful, and resolving the conflict in favor of the less well-learned response is difficult. The antisaccade task is one example of a common method for studying the regulatory processes of cognitive control. This method involves creating a conflict situation in which the subject has to respond to one stimulus or to one aspect of the stimulus and ignore another stimulus or another aspect of the stimulus. In these situations, the subject needs to focus on the target (a stimulus or an aspect of a stimulus) and ignore all the rest of the display. Failures in attention are commonly revealed in two ways: (i) reduction in efficiency of responding to the target when the irrelevant features of the display are present and (ii) indications for processing of the irrelevant material, especially when it clearly interferes with processing of the target. The two most widely used paradigms for studying this type of selection are *Stroop color naming* and *negative priming*.

A. Stroop Color Naming

J. Ridley Stroop, a theologist with a side interest in psychology, sought an experimental method that would enable him to measure the interference of one stimulus dimension on attempts to process another. In

1935, he published a seminal paper describing three experiments. The second experiment asked subjects to name the color of the ink of color words (e.g., the word ''red'' printed in green ink) or the color of colored squares. The first condition is called ''incongruent'' and the second ''neutral.'' Stroop found that the words interfered with naming the color. That is, RT to the incongruent condition was slower than RT to the neutral condition. Later, researchers added a congruent condition to the task (e.g., the word ''green'' printed in green ink). The congruent condition enabled one to look at facilitation, which is the difference in RT between congruent and neutral trials. It is commonly found that although facilitation in the congruent condition is small and on many occasions not significant, interference in the incongruent condition is robust and significant. Since 1935, this type of conflict has been studied extensively using variations of Stroop's color-naming task and in other Stroop-like situations with a wide variety of different stimuli and task demands. All studies converge on the conclusion that people cannot suppress the irrelevant dimension if it is heavily practiced and consequently overlearned (like reading words). Hence, the Stroop effect is considered a powerful example of automatic processing.

Nevertheless, several studies have shown that readers can modulate and partially control the impact of the word. Increasing the proportion of congruent trials relative to incongruent trials produces a larger Stroop effect. Moreover, even when the numbers of congruent and incongruent trials are kept constant while the proportion of neutral trials changes, the effect can be altered. It seems that the expectation to face a relatively large proportion of conflict trials prompts the adoption of a strategy that helps reduce interference. In addition, it seems that language competence may modulate the effect. When bilinguals are tested, under certain conditions they can reduce the effect in their first language but not in their second language. Note that they experience interference in both their first and their second language, but they are better able to control reading (i.e., reduce Stroop interference) in the language in which they are more competent.

Neuropsychological studies of brain-injured individuals suggest that the left frontal lobe is crucial to successful performance of the Stroop task. In particular, it has been reported that injury to the left dorsolateral prefrontal cortex results in enlarged Stroop interference. This result suggests that the Stroop interference presented by noninjured individuals is an underestimate of the potential interference. Stroop interference presented by the noninjured individuals is the product of automatic intrusion of the irrelevant word and their ability (admittedly not perfect) to inhibit this reading. In addition, it points to the involvement of the left dorsolateral prefrontal cortex in the control processes by which Stroop interfence is modulated. Consistent with this lesion evidence, neuroimaging studies of blood flow and changes in blood oxygenation during task performance show that the incongruent condition of the Stroop task activates left dorsolateral prefrontal cortex more than the neutral or congruent conditions. Even more noticeable in neuroimaging studies is differential activation of the anterior cingulate gyrus. Barch, Braver, Sabb, and Noll suggested that this medial–frontal structure is also active in a variety of other tasks in which selections must be made among competing stimuli, stimulus properties, and responses to them.

Of course, the crucial question in the present context is whether resolution of conflict in the Stroop task involves inhibition. Many theorists interpret the task in this way, although a well-known computational model of Stroop performance suggested by Cohen, Dunbar, and McClelland is able to account for many aspects of performance in the Stroop task by facilitation of the less automated process of color naming rather than inhibition of the more automated process of word reading. The phenomenon discussed next offers a more demanding and therefore more analytic test.

B. Negative Priming

In the incongruent or conflict condition of Stroop color naming, an additional slowing of performance is observed, over and above the usual interference, if the color name to be produced on any given trial is the same as the color word that had to be ignored on the immediately preceding trial. This effect, which has been dubbed *negative priming*, can be found in a wide variety of task situations in which two stimuli occur on each trial, one to be ignored and the other requiring a response. When a just-ignored distractor becomes the target on the next trial, responding is slowed relative to not having ignored the current target item in the recent past. Similar slowing occurs when there is just one stimulus on each trial to which subjects must produce a newly learned arbitrary response. Suppose that subjects must say ''car'' whenever they see ''bike'' or a picture of a bike, ''plane'' whenever they see ''car'' or a picture of a car, and ''boat'' whenever they see ''plane''

or a picture of a plane. Now suppose that as a prime the subject sees "car" and correctly says "plane." If the succeeding target is "bike," production of the correct response "car"—the overlearned response that had to be avoided when processing the prime—will be slower than if the prime was an unrelated stimulus that required an unrelated arbitrary response, such as "plane"–"boat."

The dominant interpretation of negative priming invokes a selective process that occurs while processing the prime. To respond appropriately, subjects must select against the distracter item that is the nonimperative component of the prime (or, in a task such as that of Shiu and Kornblum's just described, they must select against the overlearned but contextually inappropriate response that the prime tends to elicit). The act of selection leaves the distractor or the overlearned response in a state that makes it more difficult to process if it recurs as the target.

Two sorts of hypotheses have been offered about how this act of selection could be implemented. According to the distractor inhibition hypothesis, attentional operations actively inhibit either the prime's perceptually activated mental representation, in order to prevent it from competing for access to response selection operations, or the link between the prime's mental representation and the action ordinarily associated with the prime, in order to make that response unavailable. Inhibition takes time to dissipate. Therefore, if the inhibited representation or response link is needed soon thereafter for processing a target, more time and effort will be required for its activation. George Houghton and Steven Tipper constructed a simulation model embodying such processes.

According to an alternative proposal, the episodic retrieval hypothesis of W. Trammell Neill and colleagues, attentional operations mark the distractor item with a "do not respond to this stimulus" tag or the overlearned response with a "do not produce this response" tag. The tag provides an instruction that guides decision and response selection, and it remains a part of the experience of having processed the prime that is stored in episodic memory. When the target appears it acts as a cue to retrieve this memory. The tag that served the subject well when the tagged stimulus or response needed to be ignored causes confusion and interferes with performance when the tagged item becomes the target.

Note that the episodic retrieval hypothesis does not propose inhibition in the classic sense of reduction or suppression of activation. Its inhibitory process is "symbolic" rather than "analog," acting through an influence on a mental representation's informational content rather than its level of activity. Considerable effort has been expended attempting to distinguish these two underlying mechanisms by which an overtly inhibitory behavioral outcome might arise. Arguments have begun to appear that both mechanisms may be at work, and partly as a solution to the dilemma, in 1998 Milliken *et al.* made an important attempt to reinterpret negative priming in terms of the difficulty of deciding whether retrieved episodic memories of past processing should or should not be used to guide current performance rather than whether they have been inhibited or tagged with negative content. This debate illustrates a crucial point mentioned earlier. Inhibition of behavior (i.e., slowing of overt performance) does not necessarily signal inhibition of mental processing, if what one means by "inhibition of mental processing" is reduction or suppression of activation. Considerable theoretical analysis and empirical investigation are often required to make this determination.

IV. PRIMING AND RETRIEVAL FROM MEMORY

It is often possible to retrieve information from memory at will (though sometimes, of course, such intentional attempts at retrieval fail). At the same time, there are often occasions when thoughts come to mind without any apparent intention. Here again, we are dealing with two fundamental aspects of cognition: controlled and automatic processing. In the domain of retrieval of information from memory (general knowledge as well as specific episodes of experience), these two aspects of cognition have been investigated by exploiting *priming effects*. "Priming" consists of an alteration in the speed or accuracy of responding to a stimulus such as a word or object due to a previous encounter with that stimulus or with related stimuli.

A. Repetition Priming

Repetition priming refers to the change in responding to a word or an object as a result of a previous encounter with that same item, either in the same task or in a different task. Responding to words or objects is typically improved due to this previous experience (usually more so when the task as well as the stimulus item remains the same). For example, if a word appears for a second time in a task that requires reading the target aloud (naming task) or deciding whether the

target is a word or a nonword (lexical decision task), responding is faster and more accurate than for words appearing for the first time at the same level of overall practice in the task. In addition, repetition priming can influence the probability of producing a previously encountered item as a response when a task allows multiple possible answers on each trial. For example, suppose that in the study phase of a two-phase experiment subjects are asked to study a series of words. After a delay that can range from minutes to hours they are given three-letter word stems and asked to complete each stem with the first word that comes to mind (e.g., gre_ for green). In this word stem completion task, the repetition priming effect appears when subjects respond more frequently with words that had been studied earlier than with words that were not encountered in the study phase.

As can be seen from the examples, repetition priming is commonly examined using tasks that do not require conscious recollection of past experience with the stimulus item to complete. Therefore, it potentially represents a nonconscious or unintended effect of that experience. Such effects are called implicit memory in the repetition priming literature. When it can be documented that repetition priming has occurred without conscious recollection, the priming is called implicit memory, and as such it represents another example of automatic processing.

Brain imaging studies using positron emission tomography (PET) and fMRI have shown that priming is accompanied by reduced neural activity within areas that were initially activated to perform the task. Reduced neural activity was found in occipital visual cortex (Brodmann's area 19), left frontal cortex, and inferior temporal cortex. Whereas the reduced neural activity in visual cortex is specific to visually presented stimuli, the activity reductions in left frontal cortex occur regardless of cue modality (visual or auditory). The reduced activity within the left frontal cortex suggests an amodal priming effect that represents access to the meaning of the word rather than to its visual or orthographic representation. Analogous modality-specific and amodal effects have been observed with objects.

Single cell recording in monkeys provides evidence about the neural mechanisms that might mediate these repetition priming effects, at least for nonverbal stimuli. Repeated experience with the same visual stimulus leads to suppression of neuronal responses in subpopulations of visual neurons. This "repetition suppression" was found in the delayed matching-to-sample task. In these studies a monkey was presented with a sample stimulus followed by a sequence of test stimuli. The animal was rewarded for indicating which test stimulus matched the sample. For example, the monkey was presented with the sequence A … B … C … A and was supposed to respond to the final A. Under these conditions, the common type of neural response was suppressive, and it was graded by similarity. The more similar the test stimulus was to the sample, the more the neural response was suppressed. Moreover, repetition suppression was associated with item repetition, whether the repeated item was the target or a distractor. Repetition suppression was found to be stimulus specific and long-lasting. In addition to visual cortex, this effect was recorded in the inferior temporal cortex and also in some regions of the prefrontal cortex.

Robert Desimone suggested that the reduction in cortical activation in human neuroimaging studies such as those described earlier was due to a repetition suppression effect such as that documented in monkeys. This repetition suppression effect at the neuronal level would result in a decrease in the total number of activated cells (and hence in a reduced demand for oxygenated blood, producing the signal change measured in PET and fMRI). This reduced population of neurons, according to Desimone, provides a sharpened stimulus representation. The prime tunes the population of neurons so that a selective subpopulation that carries the critical features of the stimulus gives a robust response when the stimulus recurs, whereas other neurons, which are probably related to other stimuli, are suppressed. The more selective representation allows for a more efficient responding upon the next encounter with the stimulus. Desimone's argument recalls the words of Sir John Eccles quoted earlier.

B. Semantic Priming

Semantic priming arises because the brain makes use of relations among similar or related stimuli in addition to using past experiences with the same stimulus. In the basic version of the semantic priming paradigm, subjects are presented with two successive stimuli called the prime and the target. They are usually asked to respond overtly only to the target. When words are the stimuli, the task may be naming or lexical decision. Supposing that the target is the word "nurse," the prime can be a related word (e.g., "doctor"), an unrelated word (e.g., "bread"), or a neutral stimulus (e.g., a row of X's). Under these conditions, the

semantic priming or relatedness effect emerges. This effect can be described as a greater speed and accuracy of performance in the response to a target word when it is presented after a semantically related prime word than when it is presented after an unrelated prime word or after a neutral stimulus. This effect has been documented in a variety of situations. The semantic priming effect can occur when people are asked to pay attention to the prime and also when they do not pay attention to the prime, or even when they are unaware of its identity and do not phenomenologically realize that a prime has occurred. That is, semantic priming effects still exist when the target is presented very briefly and masked by visual noise presented immediately following the prime so that people believe they have seen only the visual noise and are not aware of the presence of the prime.

James Neely studied automatic and controlled processes by examining the ability of subjects to switch between semantic categories in an arbitrary fashion. He asked his subjects to think of parts of a building when the prime was the category name "body" (e.g., "body"–"door"). Subjects were able to follow the instructions when the time between the prime and the target (i.e., SOA) was long enough (e.g., 750 msec). That is, they responded faster to "door" following "body" relative to "door" following an unrelated prime (e.g., "tree") or even a neutral prime (e.g., XXXX). However, they were not able to switch from one category to another when the SOA was short (e.g., 250 msec). What was the fate of the rejected category? That is, when the prime was "body" and the subject made an effort to think of the category "building," what happened to parts of the body such as "arm"? "Arm" appearing after "body" was facilitated at short SOAs (<300 msec) and inhibited at long SOAs (between 500 and 2000 msec)—that is, RTs for related trials ("body"–"arm") were longer than RTs for neutral trials and equivalent to those for unrelated trials ("bird"–"arm").

Hence, priming can be achieved both by unintentional automatic activation, as shown by priming from masked words of which subjects are unaware, and from consciously perceptible words at short SOAs even when subjects are trying to think of unrelated words. However, priming can also be achieved by intentionally focusing on a concept and generating possibilities following some rule, even an arbitrary rule as in Neely's "switch condition," although such intentional focusing takes more time than automatic activation. Thus, these findings support the existence of two mechanisms of semantic priming. The first is automatic, nonconscious, and can work without attention. By analogy to repetition priming, one might wonder if it is mediated by reduction of neural response in structures that store semantic knowledge. The second is voluntary, conscious, and occupies attention. One might expect that this mechanism would be associated with neural structures involved in the executive control operations of working memory, and that invoking this mechanism would increase neural activation in those structures. To date, however, we do not know of any neuroimaging studies of semantic priming.

C. Inhibitory Semantic Priming and the Center–Surround Theory of Retrieval Operations

Neely's evidence for inhibition of related words in his switch condition is one of the few instances in the literature in which semantic relatedness between successive stimuli harms performance in speeded tasks, such as naming or lexical decision. Another instance was reported by Dale Dagenbach and colleagues. They found that in certain circumstances lexical decisions following semantically related primes are slower than lexical decisions following unrelated primes—an absolute inhibition effect associated with semantic relatedness. This inhibitory priming occurs as a consequence of attempting to retrieve the meaning of a perceptually presented word when the meaning is weakly activated—either because the word is masked and hence perceptual input is easily confused with other input or because the word is newly learned and hence its representation in semantic memory is weak and easily confused with or overwhelmed by other representations. In either case, the weakly activated meaning is likely to suffer interference because other representations are activated that are similar but incorrect. Dagenbach and colleagues proposed that in such circumstances, the attempt to retrieve the weakly activated semantic code is accompanied by active inhibition of the related information that is producing the interference. This inhibition reflects the operation of a center-surround attentional mechanism, which works to facilitate a semantic code on which it is focused or "centered" while inhibiting "surrounding" codes. These are codes that are similar or related to but different from the desired code and are competing with it for retrieval. The center-surround hypothesis predicts that repetition priming should be facilitatory at the same time that semantic priming is

inhibitory—a prediction that has been confirmed. Analogous findings have been reported by Steven Lehmkuhle and colleagues in the domain of spatial rather than semantic processing, suggesting that inhibition of similar or closely related representations may be a generalizable strategy for conflict resolution in the central nervous system.

V. EXECUTIVE CONTROL OF TASK PERFORMANCE: STOPPING AN ONGOING THOUGHT OR ACTION

Sometimes we start thinking about something or start to perform an action, only to realize it does not fit with our primary goal or our general plan at the moment or that our goal has changed and it is no longer appropriate. Such situations call for a change in the course of thought or action. We must stop the current thought or action in order to be able to switch to another one. An experimental method for studying this common example of executive control is the *stop signal paradigm* developed by Gordon Logan.

In this paradigm subjects are engaged in a primary task. In a relatively small proportion of trials they are signaled to stop before executing the response to the primary task. For example, participants can be asked to respond in each trial by pressing a button to an X and another button to an O. In addition, they are asked to withhold their response if a tone is presented at any point during a trial. The tone is presented in 25% of the trials at various delays after onset of the primary stimulus. Performance in this task can be successfully modeled as a race between a "go process" (perceiving and responding to the primary stimulus) and a "stop process." The stop process is conceived to be a separate sequence of mental operations that involves perceiving the stop signal and intervening in the ongoing sequence of operations involved in the go process or primary task. If participants finish the stop process before the go process, they inhibit their response to the primary stimulus. If they finish the go process before the stop process, they produce a response to the primary stimulus, failing to inhibit it. RT to the go signal can be measured directly, whereas stop-signal RT (the time needed to cancel the planned response) cannot be measured directly (since its behavioral signature is the absence of action and hence there is nothing overt to measure). However, Logan's race model provides quantitative methods for estimating the stop-signal RT.

Research applying this model shows that young adults can stop a wide variety of actions (key presses, hand movements, squeezes, and speech) very quickly, with an estimated latency from the stop signal of about 200 msec. Williams and colleagues reported that the ability to stop improves developmentally throughout childhood (i.e., stop-signal RT decreases) and then remains approximately constant across much of adulthood, falling off slightly but nonsignificantly in old age. Interestingly, speed of responding in most tasks that would be used as a primary task in this paradigm—the go process—also improves throughout childhood, but peak performance in young adulthood is followed by significant slowing beginning in middle adulthood. Thus, there appears to be a difference between the developmental trajectory of the execution of primary tasks and that of the type of inhibitory control represented by stopping. This supports Logan's hypothesis that processes governing inhibition are separate from those governing execution of speeded primary processing. Other research, however, shows that in old age the probability of stopping successfully does deteriorate, even if stop-signal RT remains fairly constant on those trials in which the stopping process succeeds. This suggests that old age may involve a loss of concentration in which inhibitory control processes are less likely to be implemented appropriately, even though they may still work effectively once deployed. Although this conclusion is consistent with the available data from the stopping paradigm, there is more to the relationship between age and inhibition. As will be discussed later, a major theory of cognitive aging proposes that most inhibitory functions decline in old age.

In addition to developmental changes, the ability to inhibit one's actions in the stopping paradigm is related to some personal characteristics and individual differences. Stop-signal RT varies with impulsivity, being longer for more impulsive individuals. Hyperactive children have trouble stopping. Their stop-signal RT is longer than that of normal controls and they fail to inhibit responding on many occasions. This pattern is not due to a failure to detect the stop signal itself but rather to a deficit in the inhibitory mechanism that implements stopping. Moreover, stimulant medication (methylphenidate) that improves behavioral symptoms of hyperactive children also improves their stopping performance.

Event-related potentials suggest two loci at which the stop process exerts its impact—a central locus in frontal cortex that acts on motor planning and execution operations and a more peripheral process that acts on descending motor commands after they have left cortex. Evidence on the frontal locus comes

from application of the stop-signal paradigm to examine gaze control in the FEF of monkeys by Jeffrey Schall and colleagues. Recording from single neurons in FEF showed that movement-related activity, which began to increase when a signal to move the eyes was presented, decreased when a signal to cancel the saccade was presented. The activity associated with this inhibition began to decrease before the stop-signal RT was over. It seems that the preparation of a movement is a controlled process composed of both execution and inhibition processes. FEF is involved in the generation of saccades, as mentioned earlier; in addition, neurons in the FEF can specify saccade cancellation—a specific and well-documented example of inhibitory function.

VI. DEVELOPMENTAL ASPECTS

A. Development of Reaching and the "A Not B Error"

Studies of reaching behavior in human infants and monkeys suggest that development of such behavior involves both the ability to plan and execute sequences of action and the ability to inhibit certain reflexive actions or dominant response tendencies.

Piaget suggested that infants have difficulty in understanding objects and their properties, including spatial relations. However, it seems that infants, even as young as 5 months old, can understand the object concept but they have difficulty demonstrating this understanding by their reaching behavior. Reaching behavior of infants has been studied by Adele Diamond and colleagues using several paradigms. In one paradigm infants were presented with a Plexiglas box with a building block inside or outside the box. Infants were able to retrieve the object when a direct line of reach was possible (the object was in front of the box touching its front wall or in the middle of the box, away from its front wall). However, when the object was placed inside the box touching its front wall infants were unsuccessful in retrieving it. To retrieve the object in the latter situation there is a need to execute a sequence of two movements in order to avoid touching the front wall of the box—one away from the object and a second in the direction of the object. Moreover, when the infants touched the front of the box they reflexively grasped it or withdrew their hands. Seven-month-old infants rarely continued their reaching toward the object. In contrast, 10-month-old infants were much less likely to show these reflexive behaviors upon touching the edge of the box. These infants were able to retrieve the object in these circumstances. It seems that the older infants developed the ability to execute a reach that requires a change of direction and to inhibit reflexive reactions of the hand.

Infants also have difficulty detouring around a barrier to retrieve an object. Here, the infant's task is to retrieve the object from a transparent box that has an open side. Infants 6.5–7 months old reach straight through the side at which they are looking at the object. If they see the object through the open side, they can retrieve it. Otherwise, they cannot retrieve the object and they do not try alternative reaching behaviors. Toward the end of the first year of life they can look through a closed side but retrieve the object from any open side of the box. In order to develop this ability infants need to inhibit the tendency to reach according to the line of sight.

In another paradigm the infant is presented with an object in one of two places, A or B. The locations are covered to hide the object from sight and the infant is then allowed to search for it by lifting the covers. After the infant retrieves the object, the object is again placed in one of the two locations and the search task continues. Suppose the object is first hidden at location A and the infant retrieves it successfully. If the object is then hidden at the second location B, the infant will often try to retrieve the object from A, even though the infant watched the experimenter hiding the object at B. This is the A not B error. Infants continue to make the A not B error from about 7.5 to 12 months of age, as long as the delay between hiding and retrieval is incremented as the infant gets older. Perhaps the most striking aspect of the phenomenon is that infants appear to know that the object is at B despite the fact that they reach toward A. Visual habituation and other visual memory tests indicate that infants remember the correct location. Sometimes, in the search task, the infant fixates B while he or she is reaching toward A. Moreover, infants show the A not B error even when the covers are transparent and the object can be seen. In order to prevent this type of error there is a need to hold details of the task and the situation briefly in short-term memory, and there is a need to inhibit the tendency to reach to A. The tendency to reach to A develops because reaching to A was reinforced earlier by successful finding of the object, and it increases with the number of times the infant has retrieved the object from location A before it is hidden at location B. The ability to meet both of these needs imposed by the task improves during infancy. Short-term memory improves as evidenced by the longer retention interval between

hiding and allowing the infant to reach for the object needed to elicit the A not B error, and inhibition of the incorrect response tendency improves as evidenced by the eventual disappearance of the A not B error.

It seems from this evidence that during the second half of the first year of life infants begin to gain control over their reaching actions. They can inhibit interfering automatic tendencies and demonstrate planned and goal-directed control over manual behavior. The gradual ascendance of planned and goal-directed behavior over reflexive or practiced stimulus-action routines has been modeled both by Diamond and colleagues and Stuart Marcovitch and Philip Zelazo as changes with age in the relative strength or dominance of competing systems, one like working memory and the other like conditioned or procedural learning. Evidence from lesions and single-cell recordings suggests that the increase in dominance of the working memory-like system is achieved, at least in part, through the maturation of several components of frontal cortex: the supplementary motor area (SMA) and the dorsolateral prefrontal cortex. Reflexive grasping is released in adult humans following lesions of the SMA, and the same is true in the case of lesioned monkeys. Lesions of the dorsolateral prefrontal cortex in monkeys produce the A not B error and difficulties inhibiting the urge to reach straight ahead to retrieve an object.

B. Development of Selection and Conflict Resolution

It is possible to view the previously mentioned studies as conflict situations in which habitual or endogenous tendencies compete for control of behavior. Throughout early childhood there is development in the ability to resolve such conflict and select among competing stimuli, stimulus properties, and responses. Several researchers have suggested that central to this achievement is the development of the ability to effectively inhibit stimuli or associations that are irrelevant to the task. Moreover, it has been suggested that such a development relies on maturation of the frontal lobes.

The Stroop task and its many variations have played a major part in the study of this development. For example, when preschool children are asked to say ''day'' to a picture of a moon and ''night'' to a picture of a sun, their accuracy is reduced relative to a neutral condition (e.g., responding ''day'' or ''night'' to a checkerboard). In addition, their accuracy decreases across the experimental session. Older children are able to maintain above-chance accuracy throughout a session. Similar trends have been found in latency of responding. Other conflict situations present similar results and document continued development of inhibitory function during childhood. For example, when Stroop color naming is examined across the school years, intrusion of the irrelevant word in place of the color name decreases in the incongruent or conflict condition, as does the impact of the irrelevant word on latency of correct responding.

Combining this evidence on conflict resolution with the evidence on stopping discussed previously suggests a general improvement in the ability to inhibit prepotent and automated responses from infancy to adulthood. This ability frees the system from stimulus-driven control, enabling strategic control and planning to play a more dominant part in behavior. Moreover, it seems that such development is dependent on high levels of cortical maturation. Other trends in development are consistent with this idea that inhibitory function depends on cortical input. One example is IOR, discussed earlier. The SC, which seems to be the generator of IOR, is already developed in infancy. However, there appears to be a need for the parietal lobes to provide this system with the spatial coordinates necessary for producing IOR. Although in some circumstances IOR can be observed even in infants only a few days old, the appearance of robust and widespread IOR in the eye movement patterns of infants awaits parietal development, rather than depending only on maturation of the SC.

C. Cognitive Aging

Lynn Hasher, Rose Zacks, and colleagues have amassed a considerable body of evidence that inhibitory functions decline in old age. As a consequence, the cognitive processes of older adults are increasingly susceptible to interference from irrelevant or unwanted perceptions, thoughts, and tendencies toward action that are more successfully ignored by younger adults. Hasher, Zacks, and May suggested that the deleterious impact of interference appears to occur in addition to the generalized slowing observed in a wide variety of speeded task performances. Older adults produce more errors and slower correct responses in antisaccade tasks. They suffer greater interference in conflict resolutions tasks such as Stroop color naming, and they find it more difficult to inhibit primary task performance in the stopping paradigm. They suffer greater proactive interference in short-term memory

tasks. They make more errors and retrieve answers more slowly in associative learning tasks when more than one response item is associated with a stimulus item (and in analogous "fan effect" fact-learning tasks in which multiple facts are learned about a single entity or topic item). They show less evidence of intentional or controlled forgetting in directed forgetting tasks, and they show less evidence of inhibiting irrelevant or inappropriate senses of ambiguous words and interpretations of ambiguous referring expressions in sentence reading and text comprehension. In many area of cognitive activity, older adults are susceptible to what Hasher, Zacks, and colleagues call "mental clutter."

Given the role played by prefrontal cortex in a number of the inhibitory functions reviewed in this article, one might wonder whether aging is particularly damaging to the integrity or efficiency of processing in frontal tissue. There is considerable evidence to support such an idea.

There are factors that moderate the impact of aging on cognitive activities with strong inhibitory components. One is practice, with task- and stimulus-specific practice being the most effective. Another is the level of circadian arousal. It is well-known that arousal levels vary systematically with time of day, and that individual differences in the time of occurrence of periods of optimal arousal create "morning people" and "evening people." These individual differences extend to task performance. Both younger and older adults perform a wide variety of tasks better during their optimal periods of arousal than during off-peak periods, and the impact of variation in circadian arousal is greater for older adults. Indeed, if testing is done in the morning when older adults are more likely to be in an optimal period and younger adults are more likely to be in an off-peak period, age differences in task performance are minimized and in some cases nearly eliminated. Testing in the evening, when older adults are likely to be in an off-peak period and younger adults in an optimal period, exaggerates age differences. Thus, circadian variation in locus coeruleus activity and right-prefrontal activity related to vigilance and preparation interact with other aging-induced changes in cortical efficacy.

Finally, there are a number of tasks in which performance does not appear to decline with age, at least in healthy adults free of nervous system insult such as stroke or Alzheimer's disease. Many of these are related to language use. Vocabulary scores continue to increase with age, although word-finding problems during real-time speech production and question answering may also increase. Sentence completion scores in fill-in-the-blank cloze tasks are not affected by age, nor is the accuracy of semantic categorization. All these language tasks that show little or no decline in old age are also affected little if at all by variation in circadian arousal. Perhaps the most surprising of the spared language functions is the ability to select one out of several possible and hence competing syntactic structures for application during real-time sentence production, which has been studied by Douglas Davidson in Zacks' laboratory. Here, the integrity of older adults' performance extends to speed as well as accuracy, violating both of the two best-established outcomes of cognitive aging—susceptibility to interference in conflict resolution situations and generalized slowing in speeded performances of many kinds. Thus, as has often been argued, language skills represent a domain-specific specialization whose operating principles seem to depart from those in many other areas of human cognition.

VII. CONCLUSIONS

Inhibitory functions are part and parcel of the cognitive processes that generate and control behaviors designed to provide specific solutions to dealing with environmental demands. In reviewing examples of inhibitory functions, we have seen that inhibition helps create coherent experience of the world along with the flexibility and efficiency required for skilled behavior.

The ability to inhibit prepotent or reflexive attentional and behavioral reactions and to stop unnecessary or inappropriate behavior develops throughout childhood. These developments free us from interference by otherwise dominant tendencies. This, in turn, enables us to exercise choice and intention over our actions. It seems that such changes are achieved through the development of various brain structures such as regions of the frontal lobes, including anterior cingulate, SMA, dorsolateral prefrontal cortex, and FEF. In doing their jobs, these control structures may interact or cooperate with regions of orbitofrontal cortex and amygdala involved in emotional regulation and sensitivity to delayed and long-term reinforcement contingencies. Note, however, that the ability of higher cortical structures to control reflexive behaviors by inhibiting them is not the only development that takes place. In some functional domains, inhibition already being produced by lower brain structures is made accessible to the influence of higher levels in the system.

An example of this type of development is IOR, which appears to occur in retinal coordinates early in development when SC alone is responsible for it, but it occurs in environmental and object-based coordinates once parietal cortex becomes sufficiently mature to contribute. Here, inhibition is already produced by certain brain structures but the development of cortical mechanisms allows this inhibition to be modulated by the higher brain mechanisms.

In addition, we presented results from memory and language processing suggesting that inhibition helps focus on an object or a concept by sharpening the activated representation. Moreover, it seems that this sharpening can sometimes proceed automatically, with no involvement of attention, and can be achieved without awareness, at least in some circumstances. In other circumstances, deployment of inhibitory tuning mechanisms appears to be under the control of intentions to process a particular kind of information.

In conclusion, inhibitory processes are ubiquitous in human cognition and vary in terms of the levels at which they operate and in terms of their relationship to various mental operations carried out in order to produce behavior.

See Also the Following Articles

ANTERIOR CINGULATE CORTEX • ATTENTION • COGNITIVE PSYCHOLOGY, OVERVIEW • CONSCIOUSNESS • EMOTION • HOMEOSTATIC MECHANISMS • NEUROFEEDBACK • NEUROTRANSMITTERS • STRESS: HORMONAL AND NEURAL ASPECTS • SUPERIOR COLLICULUS • UNCONSCIOUS, THE

Suggested Reading

Barch, D. M., Braver, T. S., Sabb, F. W., and Noll, D. C. (2000). Anterior cingulate and the monitoring of response conflict: Evidence from an fMRI study of overt verb generation. *J. Cog. Neurosci.* **12,** 298–311.

Cohen, J. D., Dunbar, K., and McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing model of the Stroop effect. *Psychol. Rev.* **97,** 332–361.

Dagenbach, D., and Carr, T. H. (Eds.) (1994). Inhibitory Processes in Attention, Memory, and Language. Academic Press, San Diego.

Desimone, R. (1996). Neural mechanisms for visual memory and their role in attention. *Proc. Natl. Acad. Sci. USA* **93,** 13494–13499.

Desimone, R., and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* **18,** 193–222.

Hasher, L., Zacks, R. T., and May, C. P. (1999). Inhibitory control, circadian arousal, and age. In *Attention and Performance XVII. Cognitive Regulation of Performance: Interaction of Theory and Applications* (D. Gopher and A. Koriat, Eds.), pp. 653–675. MIT Press, Cambridge, MA.

Henik, A., Rafal, R., and Rhodes, D. (1994). Endogenously generated and visually guided saccades after lesions of the human frontal eye fields. *J. Cog. Neurosci.* **6,** 400–411.

Kastner, S., De Weerd, P., Desimone, R., and Ungerleider, L. G. (1998). Mechanisms of directed attention in the human extrastriate cortex as revealed by functional MRI. *Science* **282,** 108–111.

Klein, R. M. (2000). Inhibition of return. *Trends Cog. Sci.* **4,** 138–147.

Marcovitch, S., and Zelazo, P. D. (1999). The A-not-B error: Results from a logistic meta-analysis. *Child Dev.* **70,** 1297–1313.

Milliken, B., Joordens, S., Merikle, P. M., and Seiffert, A. E. (1998). Selective attention: A reevaluation of the implications of negative priming. *Psychol. Rev.* **105,** 203–229.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In *Basic Processes in Reading: Visual Word Recognition* (D. Besner and G. Humphreys, Eds.), pp. 264–336. Erlbaum, Hillsdale, NJ.

West, R. L. (1996). An application of prefrontal cortex function theory to cognitive aging. *Psychol. Bull.* **120,** 272–292.

Wiggs, C. L., and Martin, A. (1998). Properties and mechanisms of perceptual priming. *Curr. Opin. Neurobiol.* **8,** 227–233.

Williams, B. R., Ponesse, J. S., Schachar, R. J., Logan, G. D., and Tannock, R. (1999). Development of inhibitory control across the life span. *Dev. Psychol.* **35,** 205–213.

Intelligence

ROBERT J. STERNBERG and JAMES C. KAUFMAN
Yale University

GLOSSARY

biological approaches to intelligence Approaches emphasizing the anatomical and physiological substrates of intelligence.

cognitive approaches to intelligence Approaches emphasizing thinking and learning processes in intelligence.

intelligence The ability purposively to adapt to, shape, and select real-world environments.

psychometric approaches to intelligence Approaches emphasizing measurement operations for intelligence.

Intelligence can be defined as the ability purposively to adapt to, shape, and select real-world environments. However, most investigators of intelligence want to go beyond simple dictionary-like definitions to a deeper understanding of the construct.

I. DEFINING INTELLIGENCE

If you ask people what intelligence is, the answer depends on whom you ask and the answer differs widely across disciplines, time, and place. We begin this article by discussing the diversity of views regarding what intelligence is because empirical studies often assume rather than explore the nature of the construct they are investigating—in this case, intelligence.

A. Western Psychological Views

How have Western psychologists conceived of intelligence? Almost none of the Western views are adequately expressed by Boring's 1923 operationistic view of intelligence as what intelligence tests test.

For example, in a 1921 symposium on experts' definitions of intelligence, researchers emphasized the importance of the ability to learn and the ability to adapt to the environment. Sixty-five years later, Sternberg and Detterman conducted a similar symposium, again asking experts their views on intelligence. Learning and adaptive abilities retained their importance, and a new emphasis crept in—metacognition or the ability to understand and control one's self. Of course, the name is new but the idea is not because long before Aristotle emphasized the importance for intelligence of knowing oneself.

B. Cross-Cultural Views

In some cases, Western notions about intelligence are not shared by other cultures. For example, at the mental level, the Western emphasis on speed of mental processing is not shared in many cultures. Other cultures may even be suspicious of the quality of work that is done very quickly. Indeed, other cultures emphasize depth rather than speed of processing. They are not alone: Some prominent Western theorists have

pointed out the importance of depth of processing for full command of material.

Yang and Sternberg reviewed Chinese philosophical conceptions of intelligence. The Confucian perspective emphasizes the characteristic of benevolence and of doing what is right. As in the Western notion, the intelligent person spends a great deal of effort in learning, enjoys learning, and persists in life-long learning with a great deal of enthusiasm. The Taoist tradition, in contrast, emphasizes the importance of humility, freedom from conventional standards of judgment, and full knowledge of oneself as well as of external conditions.

The difference between Eastern and Western conceptions of intelligence may persist even in the present day. A study of contemporary Taiwanese Chinese conceptions of intelligence found five factors underlying these conceptions: (i) a general cognitive factor, much like the *g* factor in conventional Western tests; (ii) interpersonal intelligence; (iii) intrapersonal intelligence; (iv) intellectual self-assertion; and (v) intellectual self-effacement.

The factors uncovered in both studies differ substantially from those identified in U.S. people's conceptions of intelligence—practical problem solving, verbal ability, and social competence—although in both cases, people's implicit theories of intelligence seem to go far beyond what conventional psychometric intelligence tests measure.

Another study varied only language. It explicitly compared the concepts of intelligence of Chinese graduates from Chinese-language versus English-language schools in Hong Kong. It was found that both groups considered nonverbal reasoning skills as the most relevant for measuring intelligence. Verbal reasoning and social skills were considered second in importance, followed by numerical skill. Memory was viewed as least important. The Chinese-language-schooled group, however, tended to rate verbal skills as less important than did the English-language-schooled group. Moreover, an earlier study found that Chinese students viewed memory for facts as important for intelligence, whereas Australian students viewed this skill as of only trivial importance.

A review of Eastern notions of intelligence suggested that, in Buddhist and Hindu philosophies, intelligence involves waking up, noticing, recognizing, understanding, and comprehending, but it also includes determination, mental effort, and even feelings and opinions in addition to more intellectual elements.

Differences between cultures in conceptions of intelligence have been recognized for some time. One study noted that Australian university students value academic skills and the ability to adapt to new events as critical to intelligence, whereas Malay students value practical skills as well as speed and creativity. Another study found that Malay students emphasize both social and cognitive attributes in their conceptions of intelligence.

The differences between East and West may be due to differences in the kinds of skills valued by the two kinds of cultures. Western cultures and their schools emphasize what might be called "technological intelligence"; thus, things such as artificial intelligence and so-called smart bombs are viewed, in some sense, as intelligent or smart.

Western schooling also emphasizes generalization or going beyond the information given, speed, minimal moves to a solution, and creative thinking. Moreover, silence is interpreted as a lack of knowledge. In contrast, the Wolof tribe in Africa views people of higher social class and distinction as speaking less. This difference between the Wolof and Western notions suggests the usefulness of examining African notions of intelligence as a possible contrast to U.S. notions.

Studies in Africa in fact provide another window on the substantial differences. Some psychologists have argued that in Africa conceptions of intelligence revolve largely around skills that help to facilitate and maintain harmonious and stable intergroup relations; intragroup relations are probably equally important and at times more important. For example, one study found that Chewa adults in Zambia emphasize social responsibilities, cooperativeness, and obedience as important to intelligence; intelligent children are expected to be respectful of adults. Kenyan parents also emphasize responsible participation in family and social life as important aspects of intelligence. In Zimbabwe, the word for intelligence, *ngware*, actually means to be prudent and cautious, particularly in social relationships. Among the Baoule, service to the family and community and politeness toward and respect for elders are seen as key to intelligence.

Similar emphasis on social aspects of intelligence has been found among two other African groups—the Songhay of Mali and the Samia of Kenya. The Yoruba, another African tribe, emphasize the importance of depth—of listening rather than just talking—to intelligence and of being able to see all aspects of an issue and to place the issue in its proper overall context.

The emphasis on the social aspects of intelligence is not limited to African cultures. Notions of intelligence

in many Asian cultures also emphasize the social aspect of intelligence more than does the conventional Western or IQ-based notion.

It should be noted that neither African nor Asian notions emphasize exclusively social notions of intelligence. A current project is studying conceptions of intelligence in rural Kenya. The Kenyans that have been studied have variegated conceptions of intelligence, distinguishing school intelligence (*rieko*) from other nonschool kinds of intelligence (such as *luoro*, which has more of the quality of character). Near one village (Kisumu), many and probably most of the children are at least moderately infected with a variety of parasitic infections. As a result, they experience stomachaches quite frequently. Traditional medicine suggests the usefulness of a large variety (actually, hundreds) of natural herbal medicines that can be used to treat such infections. Children who learn how to self-medicate via these natural herbal medicines are viewed as being at an adaptive advantage over those who do not have this kind of informal knowledge. Clearly, the kind of adaptive advantage that is relevant in this culture would be viewed as totally irrelevant in the West and vice versa. Children who do better on tests of adaptive knowledge of this kind actually do worse on Western tests of intelligence and in Western schooling in English and mathematics.

These conceptions of intelligence emphasize social skills much more than do conventional U.S. conceptions of intelligence, while simultaneously recognizing the importance of cognitive aspects of intelligence. However, it is important to realize that there is no one overall U.S. conception of intelligence. Indeed, one study found that different ethnic groups in San Jose, California, had different conceptions of what it means to be intelligent. For example, Latino parents of schoolchildren tended to emphasize the importance of social-competence skills in their conceptions of intelligence, whereas Asian parents tended to heavily emphasize the importance of cognitive skills. Anglo parents also emphasized cognitive skills. Teachers, representing the dominant culture, emphasized cognitive more than social-competence skills. The rank order of children of various groups' performance (including subgroups within the Latino and Asian groups) could be perfectly predicted by the extent to which their parents shared the teachers' conception of intelligence. In other words, teachers tended to reward those children who were socialized into a view of intelligence that happened to correspond to the teachers' own view. However, as we shall argue later, social aspects of intelligence, broadly defined, may be as important as or even more important than cognitive aspects of intelligence in later life. Some, however, prefer to study intelligence not in its social aspect but in its cognitive one.

II. COGNITIVE APPROACHES TO INTELLIGENCE

In 1957, Cronbach called for a merging of the two disciplines of scientific psychology—the differential and experimental approaches. Serious responses to Cronbach came in the 1970s, with cognitive approaches to intelligence attempting this merger. One team introduced the cognitive-correlates approach, whereby scores on laboratory cognitive tests were correlated with scores on psychometric intelligence tests. Another psychologist introduced the cognitive-components approach, whereby performance on complex psychometric tasks was decomposed into elementary information processing components.

In the 1990s, cognitive and biological approaches began to merge. A prototypical example is the inspection-time task. In this task, two adjacent vertical lines are presented tachistoscopically or by computer, followed by a visual mask (to destroy the image in visual iconic memory). The two lines differ in length, as does the amount of time for which the two lines are presented. The subject's task is to say which line is longer. However, instead of using raw response time as the dependent variable investigators typically use measures derived from a psychophysical function estimated after many trials. For example, the measure might be the duration of a single inspection trial at which 50% accuracy is achieved. Correlations between this task and measures of IQ appear to be about 0.4, slightly higher than is typical in psychometric tasks. There are differing theories as to why such correlations are obtained, but such theories generally attempt to relate the cognitive function of visual inspection time to some kind of biological function, such as speed of neuronal conduction. Next, we consider some of the biological functions that may underlie intelligence.

III. BIOLOGICAL APPROACHES TO INTELLIGENCE

An important approach to studying intelligence is to understand it in terms of the functioning of the brain, in particular, and of the nervous system, in general. Earlier theories relating the brain to intelligence tended

to be global in nature, although not necessarily backed by strong empirical evidence.

A. Early Biological Theories

Halstead suggested that there are four biologically based abilities, which he called the integrative field factor, the abstraction factor, the power factor, and the directional factor. Halstead attributed all four of these abilities primarily to the functioning of the cortex of the frontal lobes.

More influential than Halstead has been Hebb, who distinguished between two basic types of intelligence: intelligence A and intelligence B. Hebb's distinction is still used by some theorists today. According to Hebb, intelligence A is innate potential; intelligence B is the functioning of the brain as a result of the actual development that has occurred. These two basic types of intelligence should be distinguished from intelligence C—intelligence as measured by conventional psychometric tests of intelligence. Hebb also suggested that learning, an important basis of intelligence, is built up through cell assemblies, by which successively more complex connections among neurons are constructed as learning takes place.

A third biologically based theory is that of Luria, which has had a major impact on tests of intelligence. According to Luria, the brain comprises three main units with respect to intelligence: a unit of arousal in the brain stem and midbrain structures; a sensory-input unit in the temporal, parietal, and occipital lobes; and an organization and planning unit in the frontal cortex.

B. Modern Biological Views and Research

1. Speed of Neuronal Conduction

Recent theories have dealt with more specific aspects of brain or neural functioning. For example, one view suggested that individual differences in nerve conduction velocity are a basis for individual differences in intelligence. Conduction velocity has been measured either centrally (in the brain) or peripherally (e.g., in the arm).

Some investigators tested brain nerve conduction velocities via two medium-latency potentials, N70 and P100, which were evoked by pattern-reversal stimulation. Subjects saw a black-and-white checkerboard pattern in which the black squares would change to white and the white squares to black. Over many trials, responses to these changes were analyzed via electrodes attached to the scalp in four places. Correlations of derived latency measures with IQ were small (generally in the 0.1 to 0.2 range of absolute value) but were significant in some cases, suggesting at least a modest relation between the two kinds of measures.

Other investigators reported on two studies investigating the relation between nerve conduction velocity in the arm and IQ. In both studies, nerve conduction velocity was measured in the median nerve of the arm by attaching electrodes to the arm. In the second study, conduction velocity from the wrist to the tip of the finger was also measured. Vernon and Mori found significant correlations with IQ in the 0.4 range as well as somewhat smaller correlations (approximately -0.2) with response time measures. They interpreted their results as supporting the hypothesis of a relation between speed of information transmision in the peripheral nerves and intelligence. However, these results must be interpreted cautiously because a later study did not successfully replicate these earlier results.

2. Glucose Metabolism

Some of the most interesting recent work using the biological approach has been done by Richard Haier and colleagues. For example, their research showed that cortical glucose metabolic rates as revealed by positron emission tomography scan analysis of subjects solving Raven matrix problems were lower for more intelligent than for less intelligent subjects, suggesting that the more intelligent subjects needed to expend less effort than the less intelligent ones to solve the reasoning problems. A later study showed a similar result for more versus less practiced performers playing the computer game of Tetris. In other words, smart people or intellectually expert people do not have to work as hard as less smart or intellectually expert people at a given problem.

What remains to be shown, however, is the causal direction of this finding. One could sensibly argue that the smart people expend less glucose (as a proxy for effort) because they are smart rather than that people are smart because they expend less glucose. Also, both high IQ and low glucose metabolism may be related to a third causal variable. In other words, we cannot always assume that the biological event is a cause (in the reductionistic sense). It may be, instead, an effect.

3. Brain Size

Another approach considers brain size. Investigators correlated brain size with Wechsler Adult Intelligence

Scale (WAIS-R) IQs, controlling for body size. They found that IQ correlated 0.65 in men and 0.35 in women, with a correlation of 0.51 for both sexes combined. A follow-up analysis of the same 40 subjects suggested that, in men, a relatively larger left hemisphere better predicted WAIS-R verbal ability than it predicted nonverbal ability, whereas in women a larger left hemisphere predicted nonverbal ability better than it predicted verbal ability. These brain size correlations are suggestive, but it is currently difficult to determine what they indicate.

4. Behavior Genetics

Another approach that is at least partially biologically based is that of behavior genetics. The literature is complex, but it appears that about half the total variance in IQ scores is accounted for by genetic factors. This figure may be an underestimate because the variance includes error variance and because most studies of heritability have been performed with children, but it is known that heritability of IQ is higher for adults than for children. Also, some studies, such as the Texas Adoption Project, suggest higher estimates: 0.78 in the Texas Adoption Project, 0.75 in the Minnesota Study of Twins Reared Apart, and 0.78 in the Swedish Adoption Study of Aging.

At the same time, some researchers argue that effects of heredity and environment cannot be clearly and validly separated. Perhaps, future research should focus on determining how heredity and environment work together to produce phenotypic intelligence, concentrating especially on within-family environmental variation, which appears to be more important than between-family variation. Moreover, peers seem to have a particularly large effect on the development of various personal attributes, probably including cognitive skills. Such research requires, at the very least, very carefully prepared tests of intelligence—perhaps some of the newer tests described in the next section.

IV. PSYCHOMETRIC APPROACHES TO INTELLIGENCE

The psychometric approach to intelligence is among the oldest of approaches and dates back to Galton's 1883 psychophysical account of intelligence, which attempts to measure intelligence in terms of psychophysical abilities (such as strength of hand grip or visual acuity), and to Binet and Simon's 1916 account of intelligence as judgment, involving adaptation to the environment, direction of one's efforts, and self-criticism.

A. Theoretical Developments: Carroll's and Horn's Theories

Two of the major new theories proposed during the past decade are Carroll's and Horn's theories. The two theories are both hierarchical, suggesting more general abilities higher in the hierarchy and more specific abilities lower in the hierarchy. Carroll's theory will be described briefly as representative of these new developments.

Carroll proposed his hierarchical model of intelligence based on the factor analysis of more than 460 data sets obtained between 1927 and 1987. His analysis encompasses more than 130,000 people from diverse walks of life and even countries of origin (although non-English-speaking countries are poorly represented among his data sets). The model Carroll proposed, based on his monumental undertaking, is a hierarchy comprising three strata: stratum I, which includes many narrow, specific abilities (e.g., spelling ability and speed of reasoning); stratum II, which includes various group-factor abilities (e.g., fluid intelligence, which is involved in flexible thinking and seeing things in novel ways, and crystallized intelligence, the accumulated knowledge base); and stratum III, which is a single general intelligence, much like Spearman's 1904 general intelligence factor.

Of these strata, the most interesting is perhaps the middle stratum, which includes, in addition to fluid and crystallized abilities, learning and memory processes, visual perception, auditory perception, facile production of ideas (similar to verbal fluency), and speed (which includes both sheer speed of response and speed of accurate responding). Although Carroll does not break much new ground, in that many of the abilities in his model have been mentioned in other theories, he does masterfully integrate a large and diverse factor-analytic literature, thereby giving great authority to his model.

B. An Empirical Curiosity: The Flynn Effect

We know that the environment has powerful effects on cognitive abilities. Perhaps the simplest and most

potent demonstration of this effect is the Flynn effect, named after its discoverer, James Flynn. The basic phenomenon is that IQ has increased over successive generations throughout the world during most of the past century—at least since 1930. The effect must be environmental because, obviously, a successive stream of genetic mutations could not have taken hold and exerted such an effect over such a short period of time. The effect is powerful—about 15 points of IQ per generation for tests of fluid intelligence. Also, it occurs throughout the world. The effect has been greater for tests of fluid intelligence than for tests of crystallized intelligence. The difference, if linearly extrapolated (a hazardous procedure, obviously), suggests that a person who in 1892 was at the 90th percentile on the Raven Progressive Matrices, a test of fluid intelligence, would in 1992 score at the 5th percentile.

There have been many potential explanations of the Flynn effect, and in 1996 a conference was organized by Ulric Neisser and held at Emory University to try to explain the effect. Some of the possible explanations include increased schooling, greater educational attainment of parents, better nutrition, and less childhood disease. A particularly interesting explanation is that of more and better parental attention to children. Whatever the answer, the Flynn effect suggests we need to think carefully about the view that IQ is fixed. It probably is not fixed within individuals, and it is certainly not fixed across generations.

C. Psychometric Tests

1. Static Tests

Static tests are the conventional kind in which people are given problems to solve and they solve them without feedback. Their final score is typically the number of items answered correctly, sometimes with a penalty for guessing.

Psychometric testing of intelligence and related abilities has generally advanced evolutionarily rather than revolutionarily. Sometimes, what are touted as advances seem cosmetic or almost beside the point, as in the case of newer versions of the Scholastic Assessment Test (SAT), which are touted to have not only multiple-choice but also fill-in-the-blank math problems. Perhaps the most notable trend is a movement toward multifactorial theories, often hierarchical ones, and away from the notion that intelligence can be adequately understood only in terms of a single general or g factor. For example, the third edition of the Wechsler Intelligence Scales for Children offers scores for four factors (verbal comprehension, perceptual organization, processing speed, and freedom from distractibility), but the main scores remain the verbal, performance, and total scores that have traditionally dominated interpretation of the test. The fourth edition of the Stanford–Binet Intelligence Scale also departs from the orientation toward general ability that characterized earlier editions, yielding scores for crystallized intelligence, abstract visual reasoning, quantitative reasoning, and short-term memory.

Two new tests are also constructed on the edifice of the theory of fluid and crystallized intelligence of Cattell: the Kaufman Adolescent and Adult Intelligence Test and the Woodcock–Johnson Tests of Cognitive Ability–Revised. Although the theory is not new, the tendency to base psychometric tests closely on theories of intelligence is a welcome development.

The new Das–Naglieri Cognitive Assessment System is based not on fluid-crystallized theory but rather on the theory of Luria. It yields scores for attention, planning, simultaneous processing, and successive processing.

2. Dynamic Tests

In dynamic testing, individuals learn at the time of test. If they answer an item correctly, they are given guided feedback to help them solve the item until they either get it correct or the examiner runs out of clues to give them.

The notion of dynamic testing appears to have originated with Vygotsky and was developed independently by Feuerstein and colleagues. Dynamic assessment is generally based on the notion that cognitive abilities are modifiable, and that there is some kind of zone of proximal development, which represents the difference between actually developed ability and latent capacity. Dynamic assessments attempt to measure this zone of proximal development or an analog of it.

Dynamic assessment is cause both for celebration and for caution. On the one hand, it represents a break from conventional psychometric notions of a more or less fixed level of intelligence. On the other hand, it is more a promissory note than a realized success. The Feuerstein test, the Learning Potential Assessment Device, is of clinical use but is not psychometrically normed or validated. There is only one formally normed test available in the United States, by Swanson, which yields scores for working memory before and at various points during and after training as well

as scores for the amount of improvement with intervention and the number of hints that have been given and a subjective evaluation by the examiner of the examinee's use of strategies.

3. Typical Performance Tests

Traditionally, tests of intelligence have been maximum-performance tests, requiring examinees to work as hard as they can to maximize their scores. Ackerman recently argued that typical performance tests, which, like personality tests, do not require extensive intellectual effort, should supplement maximal-performance ones. On such tests, subjects might be asked to what extent they agree with statements such as "I prefer my life to be filled with puzzles I must solve" or "I enjoy work that requires conscientious, exacting skills." A factor analysis of such tests yielded five factors: intellectual engagement, openness, conscientiousness, directed activity, and science/technology interest.

Although the trend has been toward multifaceted views of intelligence and away from reliance on general ability, some have bucked this trend. Among those who have are Herrnstein and Murray.

D. The Bell Curve Phenomenon

A momentous event in the perception of the role of intelligence in society occured with the publication of *The Bell Curve* by Herrnstein and Murray. The impact of the book is demonstrated by the rapid publication of a number of responses. A whole issue of *The New Republic* was devoted to the book, and two edited books of responses were quickly published. Some of the responses were largely political or emotional in character, but others attacked the book on scientific grounds. A closely reasoned attack appeared a year after these collections. The American Psychological Association also sponsored a report that although not directly a response to *The Bell Curve* was largely motivated by it.

Some of the main arguments of the book are that (i) conventional IQ tests measure intelligence, at least to a good first approximation; (ii) IQ is an important predictor of many measures of success in life, including school success, but also economic success, work success, success in parenting, avoidance of criminality, and avoidance of welfare dependence; (iii) as a result of this prediction, people who are high in IQ are forming a cognitive elite, meaning that they are reaching the upper levels of society, whereas those who are low in IQ are falling toward the bottom; (iv) tests can and should be used as a gating mechanism, given their predictive success; (v) IQ is highly heritable and hence is passed on through the genes from one generation to the next, with the heritability of IQ probably in the 0.5–0.8 range; (vi) there are racial and ethnic differences in intelligence, with blacks in the United States, for example, scoring about one standard deviation below whites; (vii) it is likely, although not certain, that at least some of this difference between groups is due to genetic factors; and (viii) tests can and should be used as a gating mechanism, given their success.

Herrnstein and Murray attempted to document their claims using available literature and also their own analysis of the National Longitudinal Study of Youth data that were available to them. Although their book was written for a trade (popular) audience, it was unusual among books for such an audience in its use of fairly sophisticated statistical techniques.

It is not possible to review the full range of responses to Herrnstein and Murray. Among psychologists, there seems to be widespread agreement that the social policy recommendations of Herrnstein and Murray, which call for greater isolation of and paternalism toward those with lower IQs, do not follow from their data but rather represent a separate ideological statement. Beyond that, there is a great deal of disagreement regarding the claims made by these authors.

Our view is that it would be easy to draw much stronger inferences from the Herrnstein–Murray analysis than the data warrant and perhaps even than Herrnstein and Murray would support. First, Herrnstein and Murray acknowledge that, in the United States, IQ typically accounts only for approximately 10% of the variation, on average, in individual differences across the domains of success they survey. In other words, about 90% of the variation, and sometimes much more, remains unexplained.

Second, even the 10% figure may be inflated by the fact that U.S. society uses IQ-like tests to select, place, and, ultimately, to stratify students so that some of the outcomes that Herrnstein and Murray mention may actually be results of the use of IQ-like tests rather than results of individual differences in intelligence per se. For example, admission to selective colleges in the United States typically requires students to take either the SAT or the American College Test, both of which are similar (although not identical) to conventional tests of IQ. Admission to graduate and professional programs requires similar kinds of tests. The result is that those who do not test well may be denied access to

these programs and to the routes that would lead them to job, economic, and other socially sanctioned forms of success in our society.

It is thus not surprising that test scores would be highly correlated with, for example, job status. People who do not test well have difficulty gaining access to high-status jobs, which in turn pay better than other jobs to which they might be able to gain access. If we were to use some other index instead of test scores (e.g., social class or economic class), then different people would be selected for the access routes to societal success. In fact, we do use these alternative measures to some degree, although less so than in the past.

Finally, although group differences in IQ are acknowledged by virtually all psychologists to be real, the cause of them remains very much in dispute. What is clear is that the evidence in favor of genetic causes is weak and equivocal. We are certainly in no position to assign causes at this time. Understanding of group differences requires further analysis and probably requires examining these differences through the lens of broader theories of intelligence.

V. BROAD THEORIES OF INTELLIGENCE AND KINDS OF INTELLIGENCE

In recent years, there has been a trend toward broad theories of intelligence. We consider some of the main such theories next.

A. Multiple Intelligences

Gardner proposed that there is no single, unified intelligence but rather a set of relatively distinct, independent, and modular multiple intelligences. His theory of multiple intelligences (MI theory) originally proposed seven multiple intelligences: linguistic, as used in reading a book or writing a poem; logical–mathematical, as used in deriving a logical proof or solving a mathematical problem; spatial, as used in fitting suitcases into the trunk of a car; musical, as used in singing a song or composing a symphony; bodily–kinesthetic, as used in dancing or playing football; interpersonal, as used in understanding and interacting with other people; and intrapersonal, as used in understanding oneself.

Recently, Gardner proposed one additional intelligence as a confirmed part of his theory—naturalist intelligence, the kind shown by people who are able to discern patterns in nature. Charles Darwin would be a notable example. Gardner has also suggested that there may be two other "candidate" intelligences: spiritual intelligence and existential intelligence. Spiritual intelligence involves a concern with cosmic or existential issues and the recognition of the spiritual as the achievement of a state of being. Existential intelligence involves a concern with ultimate issues. Gardner believes the evidence for these latter two intelligences is less powerful than the evidence for the other eight intelligences. Whatever the evidence may be for the other eight, we agree that the evidence for these two new intelligences is speculative.

In the past, factor analysis served as the major criterion for identifying abilities. Gardner proposed a new set of criteria, including but not limited to factor analysis, for identifying the existence of a discrete kind of intelligence: potential isolation by brain damage, in that the destruction or sparing of a discrete area of the brain may destroy or spare a particular kind of intelligent behavior; the existence of exceptional individuals who demonstrate extraordinary ability (or deficit) in a particular kind of intelligent behavior; an identifiable core operation or set of oeprations that are essential to the performance of a particular kind of intelligent behavior; a distinctive developmental history leading from novice to master, along with disparate levels of expert performance; a distinctive evolutionary history, in which increases in intelligence may be plausibly associated with enhanced adaptation to the environment; supportive evidence from cognitive experimental research; supportive evidence from psychometric tests; and susceptibility to encoding in a symbol system.

Since the theory was first proposed, a large number of educational interventions have arisen that are based on the theory, sometimes closely and other times less so. Many of the programs are unevaluated, and evaluations of others seem to be ongoing, so it is difficult to say at this point what the results will be. In one particularly careful evaluation of a well-conceived program in a large southern city, there were no significant gains in student achievement or changes in student self-concept as a result of an intervention program based on Gardner's theory. There is no way of knowing whether these results are representative of such intervention programs, however.

B. Successful Intelligence

Sternberg suggested that we should pay less attention to conventional notions of intelligence and more to

what he terms successful intelligence—the ability to adapt to, shape, and select environments so as to accomplish one's goals and those of one's society and culture. A successfully intelligent person balances adaptation, shaping, and selection, doing each as necessary. The theory is motivated in part by repeated findings that conventional tests of intelligence and related tests do not predict meaningful criteria of success as well as they predict scores on other similar tests and school grades.

Successful intelligence involves an individual's discerning his or her pattern of strengths and weaknesses and then figuring out ways to capitalize on the strengths and at the same time compensate for or correct the weaknesses. People attain success, in part, in idiosyncratic ways that involve their finding how best to exploit their own patterns of strengths and weaknesses.

Three broad abilities are important to successful intelligence: analytical, creative, and practical abilities. Analytical abilities are required to analyze and evaluate the options available to oneself in life. They include identifying the existence of a problem, defining the nature of the problem, setting up a strategy for solving the problem, and monitoring one's solution processes.

Creative abilities are required to generate problem-solving options in the first place. Creative individuals are ones who "buy low and sell high" in the world of ideas: They are willing to generate ideas that, like stocks with low price–earnings ratios, are unpopular and perhaps even deprecated. Having convinced at least some people of the value of these ideas, they then sell high, meaning that they move on to the next unpopular idea. Research shows that these abilities are at least partially distinct from conventional IQ, and that they are moderately domain specific, meaning that creativity in one domain (such as art) does not necessarily imply creativity in another (such as writing).

Practical abilities are required to implement options and to make them work. Practical abilities are involved when intelligence is applied to real-world contexts. A key aspect of practical intelligence is the acquisition and use of tacit knowledge, which is knowledge of what one needs to know to succeed in a given environment that is not explicitly taught and that usually is not verbalized. Research shows that tacit knowledge is acquired through mindful utilization of experience, that it is relatively domain specific, that its possession is relatively independent of conventional abilities, and that it predicts criteria of job success about as well as and sometimes better than does IQ.

The separation of practical intelligence from IQ has been shown in different ways in different studies. Scribner showed that experienced assemblers in a milk processing plant used complex strategies for combining partially filled cases in a manner that minimized the number of moves require to complete an order. Although the assemblers were the least educated workers in the plant, they were able to calculate in their heads quantities expressed in different base number systems, and they routinely outperformed the more highly educated white-collar workers who substituted when the assemblers were absent. Scribner found that the order-filling performance of the assemblers was unrelated to measures of academic skills, including intelligence test scores, arithmetic test scores, and grades.

Ceci and Liker carried out a study of expert racetrack handicappers and found that expert handicappers used a highly complex algorithm for predicting post time odds that involved interactions among seven kinds of information. Use of a complex interaction term in their implicit equation was unrelated to the handicappers' IQ.

In a series of studies by Lave and colleagues, it was shown that shoppers in California grocery stores were able to choose which of several products represented the best buy for them, even though they did very poorly on the same kinds of problems when they were presented in the form of a paper-and-pencil arithmetic computation test. The same principle that applies to adults appears to apply to children as well: Carraher, Carraher, and Schliemann found that Brazilian street children who could apply sophisticated mathematical strategies in their street vending were unable to do the same in a classroom setting. In study by Grigorenko and Sternberg, practical intelligence was found to be a better predictor of everyday adaptation than was academic intelligence among adults in contemporary Russia.

One more example practical intelligence was provided by a study in which individuals were asked to play the role of city managers for the computer-simulated city of Lohhausen. Dorner and colleagues presented a variety of problems to these individuals, such as how best to raise revenue to build roads. The simulation involved more than 1000 variables. No relation was found between IQ and complexity of strategies used.

There is also evidence that practical intelligence can be taught, at least to some degree. For example, children in middle school given a program for developing their practical intelligence for school

(strategies for effective reading, writing, execution of homework, and taking of tests) improved more from pretest to posttest than did control students who received an alternative but irrelevant treatment.

None of these studies suggest that IQ is unimportant for school or job performance or other kinds of performance, and indeed, evidence suggests the contrary. The studies do suggest, however, that there are other aspects of intelligence that are relatively independent of IQ and that are important as well. A multiple-abilities prediction model of school or job performance would probably be most satisfactory.

According to the theory of successful intelligence, children's multiple abilities are underutilized in educational institutions because teaching tends to value analytical (as well as memory) abilities at the expense of creative and practical abilities. Sternberg, Ferrari, Clinkenbeard, and Grigorenko designed an experiment in order to illustrate this point. They identified 199 high school students from throughout the United States who were strong in analytical, creative, or practical abilities, all three kinds of abilities, or none of these kinds of abilities. Students were then brought to Yale University to take a college-level psychology course that was taught in a way that emphasized memory, analytical, creative, or practical abilities. Some students were matched, and others mismatched, to their own strength(s). All students were evaluated for memory-based, analytical, creative, and practical achievements.

Sternberg and colleagues found that students whose instruction matched their pattern of abilities performed significantly better than did students who were mismatched. They also found that prediction of course performance was improved by taking into account creative and practical as well as analytical abilities. In the broader test of abilities they used, confirmatory factor analysis failed to reveal a general factor. In a separate study, Sternberg, Torff, and Grigorenko found that fourth-graders and eighth-graders taught either social studies or science in a way that emphasized analytical, creative, and practical thinking performed better on tests of achievement than did children taught in a way that emphasized primarily memory, even if the children's achievement was measured by memory tests.

C. True Intelligence

Perkins proposed a theory of what he refers to as true intelligence, which he believes synthesizes classic views as well as new ones. According to Perkins, there are three basic aspects to intelligence: neural, experiential, and reflective.

Neural intelligence concerns what Perkins believes to be the fact that some people's neurological systems function better than do the neurological systems of others, running faster and with more precision. He mentions "more finely tuned voltages" and "more exquisitely adapted chemical catalysts" as well as a "better pattern of connecticity in the labyrinth of neurons," although it is not entirely clear what any of these terms mean. Perkins believes this aspect of intelligence to be largely genetically determined and unlearnable. This kind of intelligence seems to be similar to Cattell's idea of fluid intelligence.

The experiential aspect of intelligence is what has been learned from experience. It is the extent and organization of the knowledge base and thus is similar to Cattell's notion of crystallized intelligence. The reflective aspect of intelligence refers to the role of strategies in memory and problem solving, and it appears to be similar to the construct of metacognition or cognitive monitoring.

D. The Bioecological Model of Intelligence

Ceci proposed a bioecological model of intelligence, according to which multiple cognitive potentials, context, and knowledge are all essential bases of individual differences in performance. Each of the multiple cognitive potentials enables relationships to be discovered, thoughts to be monitored, and knowledge to be acquired within a given domain. Although these potentials are biologically based, their development is closely linked to environmental context; hence, it is difficult if not impossible to cleanly separate biological from environmental contributions to intelligence. Moreover, abilities may express themselves very differently in different contexts. For example, children given essentially the same task in the context of a video game and in the context of a laboratory cognitive task performed much better when the task was presented in the context of the video game. Part of this superiority may have been a result of differences in emotional response.

E. Emotional Intelligence

Emotional intelligence is the ability to perceive accurately, appraise, and express emotion; the ability to

access and/or generate feelings when they facilitate thought; the ability to understand emotion and emotional knowledge; and the ability to regulate emotions to promote emotional and intellectual growth. The concept was introduced by Salovey and Mayer and popularized and expanded upon by Goleman.

There is tentative evidence for the existence of emotional intelligence. For example, researchers found that emotional perception of characters in a variety of situations correlates with SAT scores, with empathy, and with emotional openness. Full convergent–discriminant validation of the construct, however, appears to be needed.

Some scholars still hold a relatively simple view of intelligence not much different from the view proposed by Spearman in 1904. However, with the introduction of emotional intelligence and all the other kinds of intelligences, it seems like a simple view may fail to capture intelligence in all its richness.

See Also the Following Articles

ARTIFICIAL INTELLIGENCE • BILINGUALISM • CATEGORIZATION • COGNITIVE AGING • COGNITIVE PSYCHOLOGY, OVERVIEW • CREATIVITY • EVOLUTION OF THE BRAIN • INFORMATION PROCESSING • LANGUAGE ACQUISITION • LANGUAGE AND LEXICAL PROCESSING • LANGUAGE, NEURAL BASIS OF • LOGIC AND REASONING • PROBLEM SOLVING • SPEECH

Acknowledgments

Preparation of this article was supported in part under the Javits Act Program (Grant R206R500001) as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The opinions expressed in this article do not necessarily reflect the positions or policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

Suggested Reading

Ackerman, P. L., and Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychol. Bull.* **121,** 219–245.

Ceci, S. J. (1996). *On Intelligence: A Bioecological Treatise on Intellectual Development* (expanded ed.). Harvard Univ. Press, Cambridge, MA.

Cole, M. (1996). *Cultural Psychology: A Once and Future Discipline*. Harvard Univ. Press, Cambridge, MA.

Fischer, C. S., Hout, M., Sanchez Janowski, M., Lucas, S. R., Swidler, A., and Voss, K. (1996). *Inequality by Design: Cracking the Bell Curve Myth*. Princeton Univ. Press, Princeton, NJ.

Gardner, H. (1999). Are there additional intelligences? The case for naturalist, spiritual, and existential intelligences. In *Education, Information, and Transformation* (J. Kane, Ed.), Prentice Hall, Englewood Cliffs, NJ.

Goleman, D. (1998). *Working with Emotional Intelligence*. Bantam, New York.

Grigorenko, E. L., and Sternberg, R. J. (1998). Dynamic testing. *Psychol. Bull.* **124,** 75–111.

Harris, J. R. (1998). *The Nurture Assumption*. Free Press, New York.

Jensen, A. R. (1998). *The g Factor*. Greenwood, Greenwich, CT.

Mayer, J. D., and Salovey, P. (1997). What is emotional intelligence? In *Emotional Development and Emotional Intelligence: Educational Implications*. (P. Salovey and D. Sluyter, Eds.), Basic Books, New York.

Neisser, U. (Ed.) (1998). *The Rising Curve*. American Psychological Association, Washington, DC.

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., and Urbina, S. (1996). Intelligence: Knowns and unknowns. *Am. Psychol.* **51,** 77–101.

Sternberg, R. J. (1996). *Successful Intelligence*. Simon & Schuster, New York.

Sternberg, R. J., and Grigorenko, E. L. (Eds.) (1997). *Intelligence, Heredity, and Environment*. Cambridge Univ. Press, New York.

Sternberg, R. J., and Horvath, J. (Eds.) (1999). *Tacit Knowledge in the Professions*. Erlbaum, Mahwah, NJ.

Ion Channels

B. ALEXANDER YI and LILY Y. JAN
University of California, San Francisco

GLOSSARY

complementary DNA DNA that is synthesized from mRNA and therefore contains the coding sequence of the gene with no introns.

equilibrium potential The membrane potential at which there would be no net movement of ions across the membrane.

ligand A molecule that binds a protein.

neurotransmitter A chemical substance that is stored in vesicles at the nerve terminal and released to cause a change in the postsynaptic membrane, usually a change in the membrane to ions.

protease An enzyme that cleaves proteins.

resting potential The membrane potential of the cell in its quiescent state.

second messenger A molecule that is generated by the activation of surface receptors in response to hormones or neurotransmitters that lead to changes in the functional state of the cell.

In one sense, a cell is similar to a battery. Approximately one-third of the cell's metabolic energy is stored as an electrical potential in the form of ionic gradients across the plasma membrane. This energy is released at precise moments by "holes" in the membrane that allow ions to move down their electrochemical gradients across the membrane. These holes are ion channels. In the brain, a diverse array of ion channels coordinates their actions to generate complex waveforms that are used to transmit signals across long distances or between cells.

I. ROLE OF ION CHANNELS IN PHYSIOLOGY

Ion channels are membrane proteins that catalyze the transfer of ions down their electrochemical gradients across the plasma membrane. Ion channels are necessary because the plasma membrane is hydrophobic and thus by themselves they are impermeable to ions. In this article, we review what is known about the role of ion channels in physiology and their structure and function. More is known about some ion channels than others. In the last section, we introduce some of the major ion channel families that have attracted the interest of scientists who study ion channels.

There are two types of proteins that are able to move ions across the plasma membrane: ion pumps and ion channels. There are several features that distinguish them. Ion pumps are able to perform thermodynamic work (i.e., they are able to move an ion against its electrochemical gradient). This is accomplished by consuming energy in the form of ATP hydrolysis or the concentration gradients of other ions. Ion channels cannot perform thermodynamic work and the direction of ions traveling through an open channel is solely dependent on the electrochemical gradient. Sometimes ion pumps are said to carry out active transport, whereas ion channels carry out passive transport. Another major difference is the rate at which ions move through these proteins. Ion channels are essentially pores in the plasma membrane and the throughput of an ion channel can be fast—up to 100 million ions per second. The turnover rate of ion pumps is typically orders of magnitude slower.

Ion channels are found in nearly all cells in nature and play integral roles in a cell's basic physiology. It is likely that ion channels were among the proteins found

in the earliest forms of life on this planet. Through millions of years of evolution, the number and diversity of ion channels have expanded to take on more complex functions such as those involved in learning and memory in the nervous system. Classically, ion channels are introduced via a discussion of neuronal action potentials that transmit signals down an axon. Beyond action potentials, ion channels play roles in other processes too many to enumerate. On a fundamental level, the activity of ion channels can change the membrane potential of the cell or alter the concentrations of ions inside the cell. These processes are basic to the cell's physiology; therefore, it is easily imagined that ion channels may be involved, directly or indirectly, in virtually all cellular activities. More directly, some of these activities may include setting the membrane potential, allowing the entry of ions for nutritive needs, allowing the entry of Ca^{2+}, which is used as a second messenger, and controlling cell volume.

In other cases, the physiological role of an ion channel is unknown. We have often learned about the function of ion channels by discovering instances when their activity goes awry—namely, disease states. It is becoming apparent that defective ion channels underlie the pathogenesis of many human diseases. The term channelopathy has been coined to refer to the expanding list of diseases in this class. For example, cystic fibrosis stems from a mutation in a chloride channel, cystic fibrosis transmembrane regulator (CFTR). In the lung, mutations in CFTR disrupt normal Cl^- efflux, which is necessary for the secretion of fluid that coats the airway epithelium. Consequently, patients with cystic fibrosis develop viscous mucus secretions that can obstruct the airways and are prone to acquiring life-threatening pulmonary infections. Another example is long-QT syndrome. Individuals with long-QT syndrome have abnormally prolonged action potentials in the heart and are at risk for ventricular arrhythmias that can lead to sudden death. Some individuals with long-QT syndrome have mutations in potassium channels that repolarize the cardiac muscle. The direct importance of ion channels in cardiac function is underscored by the effectiveness of antiarrhythmic drugs, many of which act on ion channels.

II. PRINCIPLES OF ION CHANNEL MECHANISMS

A. The Electrochemical Gradient

The electrochemical gradient determines the direction that ions will flow through an open ion channel and is a combination of two types of gradients: a concentration gradient and an electrical field gradient. We can consider these two gradients separately. Figure 1a shows two compartments that contain an aqueous solution of ions separated by a membrane. It is apparent that there is a concentration gradient, since the left side contains more ions than the right. Assuming the membrane is permeable to the ion, there will be a net movement of ions from the right to left side until the concentrations of ions on both sides are the same. In this case, when the concentrations on both sides equalize, the solution will have reached

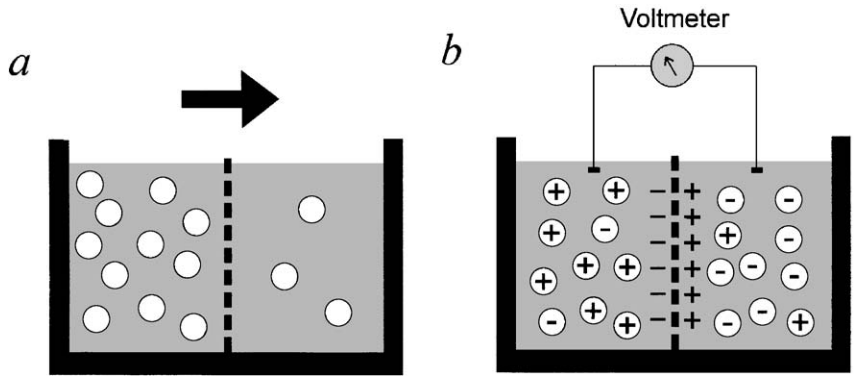


Figure 1 The electrochemical gradient is a combination of the concentration gradient and the electrical potential. The panels show two compartments that contain a solution of ions (circles). An ion-permeable membrane separates the two compartments. (a) There is a higher concentration of ions on the left side; therefore, ions will tend to diffuse from the left to the right (direction of arrow). It is also apparent that there is an osmotic gradient. Initially, water will tend to flow from the right to the left. At equilibrium, however, the two sides will be isoosmotic. (b) A voltage has been applied across the membrane. As a consequence, positively charged ions are drawn to the left and negatively charged ions to the right.

equilibrium. At equilibrium, an equal number of ions will diffuse across the membrane in both directions, and the concentrations of ions on either side will not change.

The electrical field gradient takes into account the charge on the ion. In Fig. 1b, an electrical potential has been applied so that the left side is negatively charged and the right side is positively charged. Ions that are positively charged will flow into the left compartment until it reaches a new equilibrium, in which the electrostatic forces that pull the cations into the left side are balanced by the tendency for the ions to move down its concentration gradient. Negatively charged ions will tend to flow into the right compartment. In this equilibrium, the final concentrations of ions on both sides are not equal.

The relationship between the electrical potential and the magnitude of the concentration gradient that is created is intuitive: The stronger the electrical potential, the greater the concentration gradient. The Nernst equation describes this relationship:

$$E_x = \frac{RT}{zF} \ln \frac{[X]_{\text{out}}}{[X]_{\text{in}}} \qquad (1)$$

where E_x is the electrical potential with units of millivolts, R is the gas constant, T is the temperature, z is the valence of the ion, and F is Faraday's constant. In words, the Nernst equation states that an electrical potential, E_x, will produce a concentration gradient with the ratio $[X]_{\text{out}}/[X]_{\text{in}}$ when the membrane is permeable to the ion. The converse is also true; a concentration gradient, $[X]_{\text{out}}/[X]_{\text{in}}$, will generate an electrical potential, E_x. Near room temperature (20°C), the Nernst equation simplifies to

$$E_x = \frac{58}{z} \log_{10} \frac{[X]_{\text{out}}}{[X]_{\text{in}}} \qquad (2)$$

E_x is also referred to as the equilibrium potential or the Nernst potential.

The cell is similar to the compartments in Fig. 1, only more ions need to be considered. The membrane potential is determined by the permeability of the membrane to a given ion. Figure 2 gives the concentrations of Na^+, K^+, Cl^-, and Ca^{2+} inside and outside a typical cell. Two processes work to maintain the concentration gradients of these ions. The action of ion pumps helps keep the cytoplasmic concentrations of Na^+, Cl^-, and Ca^{2+} low and the K^+ concentration high. Second, the presence of macromolecular anions inside the cell, such as proteins, tends to produce gradients of ions on their own. The redistribution of ions due to fixed charges in the cell is referred to as the Donnan effect.

At rest, the membrane is permeable to K^+; therefore, the resting potential of the cell is near E_K (in Fig. 2, approximately −84 mV). The flow of ions through the membrane is not large enough to affect changes in the ionic composition inside or outside the cell. Experimentally, currents can be elicited by changing the electrical potential across the membrane in short pulses. At membrane potentials of approximately −84 mV, there will be small K^+ currents since the membrane potential is near E_K, at which there is no net flow of K^+ across the membrane. Above E_K, there will be a net flow of K^+ outward because the electrical field gradient is not large enough to balance the concentration gradient. This results in an outward current (Fig. 3a). Below E_K, there will be a net inward current because the stronger electrical field gradient will tend to pull more K^+ into the cell. With the opening of sodium channels, the membrane becomes predominantly permeable to Na^+ rather than K^+. The low concentration of Na^+ inside the cell and the negative

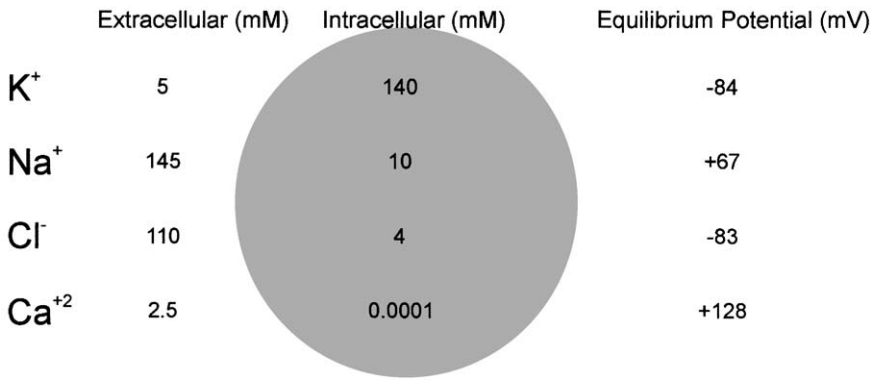


Figure 2 The concentration of ions inside and surrounding a typical mammalian cell. The equilibrium potentials were calculated using Eq. (2).

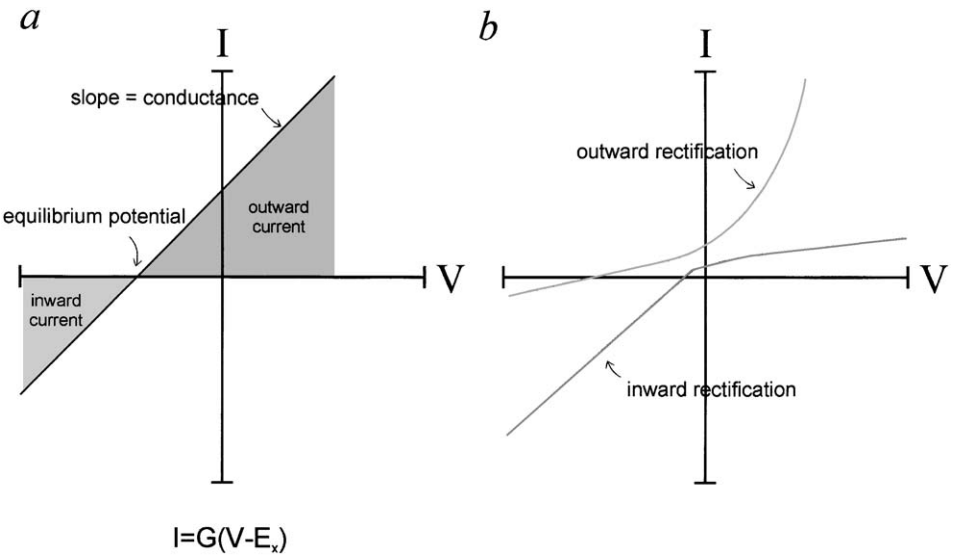


Figure 3 Current–voltage plots of ion channels. (a) An *I–V* curve of an ion channel that conducts positively charged ions. Above the equilibrium potential, there is outward current; below the equilibrium potential, there is inward current. An ohmic ion channel has a linear *I–V* relationship. (b) Rectifiers pass more current in one direction. *I–V* curves of ion channels that rectify curve upwards or downwards.

membrane potential initially create a strong driving force for Na^+ to enter the cell and can cause the membrane potential to approach E_{Na}, (e.g., at the peak of action potentials).

B. Current–Voltage Relationships

Current is the movement of charge, and in cells the flow of ions through ion channels can be measured. By expressing cloned ion channels in heterologous systems or by silencing other channels with chemical blockers, the current from a single ion channel can be isolated. In these settings, it is often useful to apply different electrical potentials to the membrane and measure the magnitude and direction of the current as a function of voltage. If current through the ion channel is linearly related to the membrane potential as in Fig. 3a, then the ion channel is said to be ohmic because it behaves like a resistor and follows Ohm's law:

$$V = IR \tag{3}$$

where V is voltage, I is current, and R is the resistance. For the purposes of studying ion channels, it is useful to modify Ohm's law into the form

$$I = G(V - E_x) \tag{4}$$

Here, V is replaced with $(V-E_x)$ because $I=0$ at the equilibrium potential, and R is replaced with its inverse, the conductance G. The conductance is the ease with which ions will flow through an ion channel, which is more intuitive since ion channels with a high conductance will conduct larger currents. $(V-E_x)$ constitutes the driving force, which together with the conductance determines the amount of current that flows through the ion channel. From Fig. 3a or Eq. (4), we can see that G represents the slope of the *I–V* curve. Ion channels with a higher conductance will have steeper *I–V* curves.

I–V curves show the sensitivity of an ion channel to voltage. One should bear in mind that all ion channels are not ohmic. Ion channels that conduct more ions in one direction are said to rectify. For example, inward rectifier potassium channels preferentially conduct more K^+ into the cell than out of the cell (Fig. 3b). Other ion channels display outwardly rectifying currents.

C. Anatomy of a Typical Ion Channel

In the early 20th century, ion channels as we know them were considered nothing more than "holes in the membrane" that allowed ions to pass through them. We now know that ion channels are more complex than simple holes. In particular, there are two properties that distinguish them from simple holes: They exhibit selectivity for certain ions, and they open and close in response to stimuli.

All ion channels display many of the same basic properties. All ion channels are integral membrane proteins that form a pore in the lipid bilayer (Fig. 4). Like other membrane proteins, ion channels contain stretches of hydrophobic amino acids called transmembrane segments that anchor the protein within the lipid bilayer. These transmembrane segments pack against each other to stabilize the basic pore structure. Part of the pore is narrow and forms a barrier beyond which impermeant ions cannot cross. This region is lined with amino acid sidechains or backbone atoms that interact with the ion briefly as it crosses the pore. A flexible region of the ion channel forms a gate that acts to open and close access to the pore. The opening and closing of the ion channel is referred to as gating. The position of the gate can be influenced by regions that are modified by enzymes such as kinases or phosphatases. Finally, many ion channels extend processes that dock onto intracellular scaffolding proteins that are part of the cellular architecture. These interactions help direct those ion channels to specific locations within the cell such as the synapse or an intracellular organelle.

D. Ion Selectivity

All ion channels conduct certain ions over others—a property referred to as ion selectivity. Some ion channels are permeable to a class of ions; cation channels are one example, though most ion channels conduct a single type of ion. The selectivity of some ion channels is extraordinary. Voltage-gated calcium channels are 1000 times more selective for Ca^{2+} over other cations and pass Ca^{2+} almost exclusively. This is remarkable considering that the concentrations of other ions such as Na^{+} can be much higher than that of Ca^{2+}. The mechanism by which ion channels pick and choose certain ions is still not fully understood.

At the molecular level, ions are little more than point charges with different valences. For ions with the same charge the only distinguishing feature is their ionic radius. Na^{+} and K^{+}, which both have a charge of $+1$, have radii of 0.95 and 1.33 Å, respectively. Another feature is the water molecules that surround and interact with the ion. In solution, ions are surrounded by multiple layers of water molecules that are continuously exchanging with other "free" waters around it. Based on the mobility of ions in solution, Na^{+} behaves as though they are larger than K^{+}. This is because its charge is concentrated in a smaller space and thus binds its waters of hydration more tightly.

The narrowest part of the pore has been termed by Bertil Hille as the ion selectivity filter. A filter with a fixed diameter is one way of explaining ion selectivity. In this view, ion channels are like sieves that allow small ions to pass but retain large ions. Although this mechanism is certainly at work in ion channels, an ion selectivity filter does not explain how an ion channel could be permeable only to large ions.

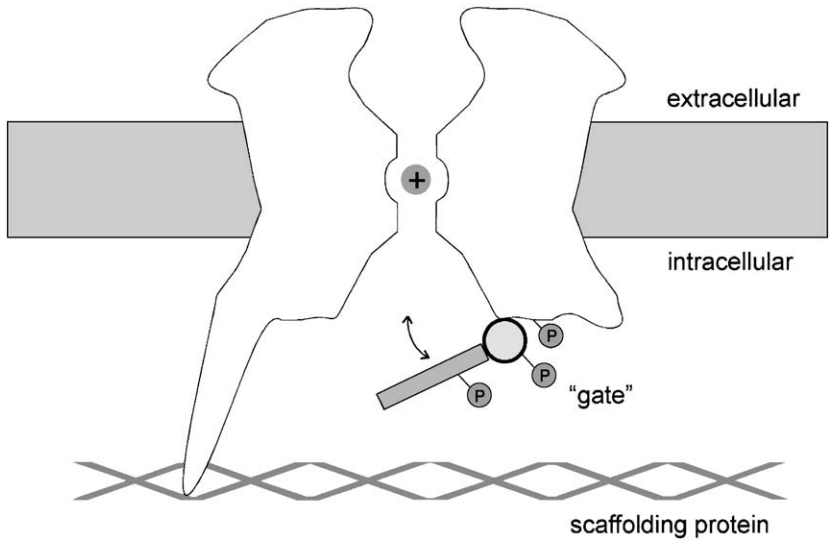


Figure 4 The "anatomy" of a typical ion channel. Ion channels are integral membrane proteins that sit in the plasma membrane. An ion is drawn sitting in a binding pocket in the ion selectivity filter (narrow part) as it passes through the ion channel. A gate (drawn as a door) swings open and closed.

It is likely that the permeant ion makes direct contact with the ion channel at some points during its passage. These interactions at specific binding sites may explain how an ion channel can be selective for large ions or discriminate between ions of similar radii. The binding sites can be depicted with the aid of an "energy landscape" diagram (Fig. 5a). Unfavorable locations are represented as regions of high energy, or peaks, whereas stable regions such as binding sites are represented by valleys. It is clear that some ion channels have more than one binding site within the channel. Multiple binding sites in close proximity may explain the high turnover rate of ion channels by setting up electrostatic repulsion between ions.

All potassium channel genes share a P region. The P region can be easily identified in potassium channel sequences since it contains the signature amino acid sequence GYG. The P region forms a part of the pore that is involved in ion selectivity since mutations in this region can alter the ion selectivity of potassium channels. The physical nature of the binding sites formed by the pore regions of most ion channels is unknown; however, it is likely that they are assembled from polar and charged amino acid sidechains or carbonyl oxygens from the protein backbone. These moieties could mimic the waters that may have been stripped away when the ion enters the pore. These interactions are likely to be transient and of low affinity given the high rate of turnover of an ion channel. One measure of the affinity can be obtained by measuring the current size at given ion concentrations (Fig. 5b). With no permeant ions present, an open ion channel will conduct zero current. As the concentration of ions is increased, the current will increase since the greater number of permeant ions will allow more ions to flow through the channel. Eventually, the current will reach a plateau or saturate. Plotted on a graph, the current can be fit with the Michaelis–Menton equation, and a constant (K_M), a measure of the affinity of the ion channel for the ion,

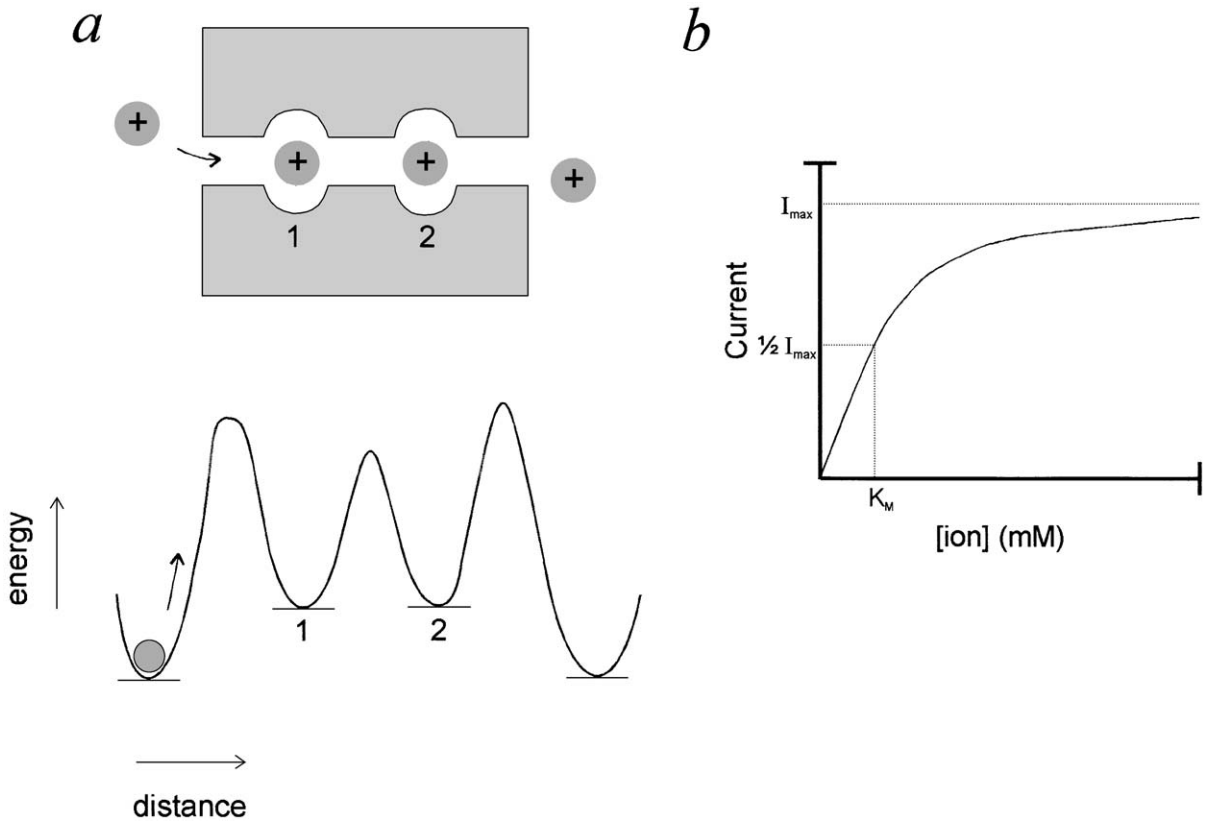


Figure 5 Ion channels contain binding sites for ions in the pore. (a) Ions move through the pore in single file from left to right. Binding sites 1 and 2 are places where the ions make interactions with the ion channel. The procession through the pore can be represented on an energy landscape diagram in which the binding sites are drawn as valleys. (b) The current amplitude eventually plateaus as the concentration of permeant ions is increased. This data can be fit using the Michaelis–Menton equation to derive the K_d, a measure of affinity.

can be obtained. Typically, ion channels show a K_M in the millimolar range.

E. Ion Channel Gating

Ion channels undergo a conversion between two types of states—closed and open—in a process referred to as gating. The gating behavior of an ion channel defines its functional role in physiology and is typically tied to the name of the ion channel. Ion channels are able to respond to a wide range of stimuli. The binding of a neurotransmitter (ligand-gated ion channels), a change in the membrane potential (voltage-gated ion channels), a physical pull on the ion channel protein (mechanosensitive ion channels), and heat are known to activate certain ion channels.

An ion channel with two states, one closed and one open, can be represented by a state diagram:

$$C \leftrightarrow O \tag{5}$$

where C stands for closed and O stands for open. The $C \rightarrow O$ transition is referred to as activation; the reverse process, $O \rightarrow C$, is called deactivation. Since ion channels are metastable and can exist in multiple states, there is no way of knowing a priori which state a single ion channel will be in at any moment in time. After observing the activity of an ion channel for some length of time, however, one can collect statistics that describe the probability that an ion channel will be in the closed or open state, the average duration of each closing and opening, and the frequency of switching between states. These parameters are part of a set of fundamental variables that uniquely describe the activity of an ion channel. A stimulus can activate the ion channel by destabilizing the closed state or stabilizing the open state and thus increase the probability that an ion channel will be in the open state. A closed state that is accessed after the ion channel opens is referred to as the inactive state (I). An ion channel with three states may have the following state diagram:

$$C \leftrightarrow O \leftrightarrow I \tag{6}$$

The $O \rightarrow I$ transition is referred to as inactivation. Often, the state diagrams of ion channels are complex and can contain multiple interconnecting closed, open, and inactive states.

The physical change in the ion channel structure that is responsible for gating remains an area of active research. Studies of several model ion channels have revealed a range of mechanisms by which ion channels may gate. For one channel, the nicotinic acetylcholine receptor (nAChR), scientists have been able to get a glimpse of the gating process with the use of the electron microscopy. Thus far, this has only been possible for the nAChR because the *Torpedo* electric ray produces abundant amounts of this protein that can form crystalline arrays in the membrane. These images revealed that this ion channel has an outer vestibule that makes it a great deal longer than the vertical height of the plasma membrane. Images taken before and after treatment with acetylcholine have been used to model the conformational changes that occurred after ligand binding. The binding site for acetylcholine lies on the outer vestibule, and the binding of ligand is transduced as a signal to the transmembrane segments that then undergo a concerted change in structure to open the pore. Unfortunately, the resolution of these images is too low to provide a detailed view of the gating process. One model postulates that the pore is lined by "kinked helices" whose vertices project into the pore in the closed state. After treatment with acetylcholine, the kinked helices appear to rotate away from the pore. This rotation may move hydrophobic residues that block the pore out of the way and replace them with polar residues.

Another well-studied gating mechanism is the inactivation gating of voltage-gated potassium channels (Fig. 6b). Inactivation can be eliminated, while other properties are left intact, by treating the inside of the channel with proteases. This and other experimental observations of the inactivation process can be explained by a "ball-and-chain" mechanism. In this model, part of the ion channel that binds the pore (ball) is tethered to the ion channel by a linker (chain) and blocks the pore once the channel has been activated. The region of the ion channel that forms the ball can be expressed by itself. As further evidence of this mechanism, when the ball is directly applied onto protease-treated ion channels, inactivation can be restored.

Inward rectifier potassium channels illustrate that the gate does not need to be an intrinsic part of the ion channel (Fig. 6c). Inward rectifying potassium channels pass larger currents into the cell than out of the cell. The channel alone, however, exhibits little or no rectification but displays rectification when it is brought near a cell, suggesting that a soluble factor is involved. In the case of inward rectifier potassium channels, it was discovered that Mg^{2+} and

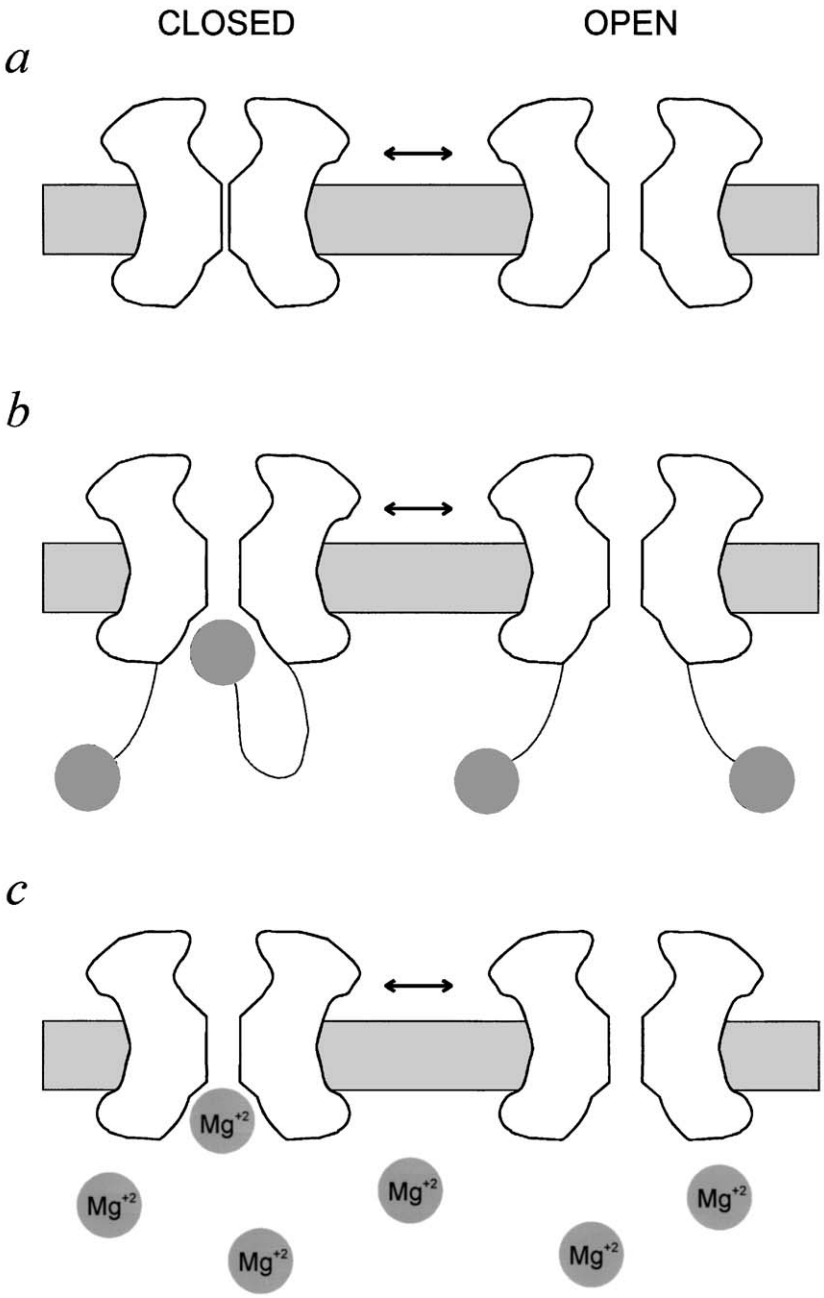


Figure 6 Three examples of ion channel gating mechanisms depicted as a cartoon. (a) Gating occurs via a generalized change in the structure in the region of the pore. The nicotinic acetylcholine receptor undergoes a conformational change of its M2 segment that closes the ion conduction pathway. (b) "Ball-and-chain" mechanism of inactivation gating in voltage-gated potassium channels. (c) Inward rectifiers are gated by extrinsic factors (e.g., Mg^{2+}) that block the pore. The block is voltage dependent; therefore, inward rectifiers conduct more current below E_K and little current above E_K.

polyamines, amino acid metabolites, produced rectification by binding to the pore of the channel at depolarized membrane potentials but not at hyperpolarized membrane potentials.

F. Ion Channel Modulation

Aside from the binding of a ligand or other stimuli that activate the ion channel, the gating of the ion channel

can be affected by other factors that affect the ease of opening the ion channel. This process is loosely referred to as modulation and can occur through the binding of a second messenger or a covalent modification of the ion channel such as phosphorylation. The role of modulation is to tweak or fine-tune the gating of an ion channel, although in some instances the line between modulation and gating can be blurred. For example, the phosphorylation of CFTR is necessary for channel openings to occur.

In general, modulation events originate at receptors that are at a distance from the ion channel. The response is usually slower in onset because a second messenger needs to be generated and then diffuse to the ion channel. The signal transduction cascade also branches off to influence other effectors that produce major changes in a cell's functioning. The modulation of ion channels can be part of long-lasting changes in the cell.

III. MOLECULAR PROPERTIES OF ION CHANNELS

A. Primary Structure of Ion Channels

What do ion channels look like? Although ion channel currents have been measured for many decades, only recently with advancements in molecular biology have their structures been determined in any detail. The goal of understanding ion channel structure is a biochemical one. Usually, the initial step is to determine the primary structure of the ion channel protein. Currently, this is achieved by first cloning the ion channel gene and then reading the protein sequence from its cDNA. The nicotinic acetylcholine receptor was the first ion channel to have its sequence identified through the work of Shosaku Numa and colleagues in 1982. They first purified the ion channel protein from the *Torpedo* electric ray and obtained small bits of protein sequence. Then they designed short oligonucleotides that corresponded to their protein sequence and hybridized them to a cDNA library that contained genes from *Torpedo*. In this way, they obtained the entire cDNA sequence and read the entire amino acid sequence of nAChR.

Ion channels have also been identified using genetic methods. *Shaker*, a voltage-gated potassium channel from *Drosophila melanogaster*, was cloned by Lily and Yuh Nung Jan and colleagues in 1987. *Shaker* derives its name from the behavior of mutant flies when exposed to ether and was long suspected to be an ion channel based on biophysical studies. *Shaker* had already been localized to a region of the X chromosome. To identify part of the gene sequence, the Jan group performed Southern blots on a series of *Shaker* mutant flies with chromosomal rearrangements. Then, they were able to obtain the entire *Shaker* sequence by probing a *Drosophila* cDNA library with their partial DNA sequence.

Once a member of an ion channel family has been identified, other members of the same family can be identified by sorting through DNA libraries for genes that have related sequences. What was in the past done with petri dishes and nitrocellulose membranes is now being done using computers. With the growing information collected from genome sequencing projects, it is increasingly common for new ion channel genes to be identified through the Internet. Once a putative ion channel sequence has been found, a relatively straightforward sequence of steps can be used to clone the gene by amplifying from a sample of genomic DNA.

A great deal of information about what the ion channel looks like can be obtained from the primary structure. For example, it is known that membrane proteins contain discrete stretches of hydrophobic amino acids that span the lipid bilayer. These transmembrane segments can be identified with computer programs that search the protein for long stretches of hydrophobic residues. Combined with other information, one can derive a model of the membrane topology of the ion channel. The models that are generated, however, can be imperfect. There are several experimental methods by which scientists can test whether a region of the protein lies on the extracellular or intracellular face of the membrane. One is to identify sites that are glycosylated since they are known to be present only on the extracellular parts of a protein. Another is to raise antibodies against specific stretches of the protein and determine whether the antibodies bind from the inside or the outside surface of the cell.

The membrane topography is a useful scheme for classifying ion channels (Fig. 7). For example, there are three classes of potassium channels. The gene for voltage-gated potassium channels has six transmembrane segments, numbered S1–S6, and another hydrophobic region between S5 and S6 called the P region, which does not cross the plasma membrane. The P region contributes to the pore of the ion channel. The inward rectifier potassium channels have two transmembrane segments with a P region between M1 and M2. A third class of potassium channels is called two-pore channels because each gene contains two P regions. Two-pore channels have four transmembrane

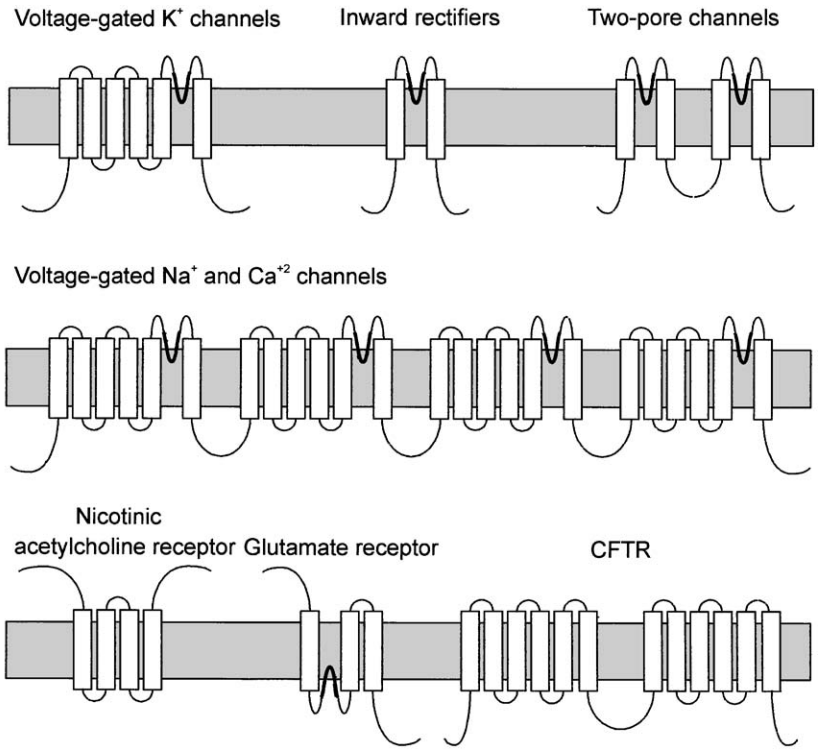


Figure 7 The membrane topology of ion channels. Voltage-gated sodium and calcium channels resemble four voltage-gated potassium channel subunits linked together. The nicotinic acetylcholine receptor and the glutamate receptor are both ligand-gated ion channels but have different membrane topologies. Two-pore channels look like two inward rectifiers fused together. Among ion channels, CFTR belongs to a unique class of membrane proteins.

segments, M1–M4. The membrane topology also gives clues about the function of different parts of the ion channel. For example, the binding site for acetylcholine in nAChR would be limited to those regions that faced the extracellular side.

Sequence analysis of different ion channel families suggests that ion channels are evolutionarily related. Two-pore channel genes resemble two inward rectifier potassium channel genes linked together, and voltage-gated sodium channels and voltage-gated calcium channels resemble four voltage-gated potassium channel genes linked in tandem.

The cloning of ion channels has revolutionized the study of ion channels by allowing scientists to make changes in the amino acid sequence and then test them. One particular strategy of making mutations—making chimeric ion channels—has been especially powerful in determining ion channel structure and function. In one well-known example, Numa, Sakmann, and colleagues used this method to conclude that M2 in the nicotinic acetylcholine receptor is important for determining the conductance of nAChR. In this strategy, one starts with two related ion channels that have different properties. In this case, it was known that nAChR expressed with the δ subunit from calf has a conductance of 65 pS, whereas the δ subunit from *Torpedo* has a conductance of 87 pS. In order to identify the residues that were responsible, Numa and colleagues made hybrid genes by splicing parts of the gene for the calf δ subunit to the *Torpedo* δ subunit and measured their conductance. With successive chimeras that contained increasingly smaller parts of the *Torpedo* gene, they were able to determine that the M2 segment is the region that determines the difference in conductance between the calf and *Torpedo* δ subunits.

This approach has been instrumental in assigning functional roles to parts of ion channels. Using this approach, it has been revealed that ion channels are modular in design (Fig. 8). One caveat, however, is that one cannot exclude the involvement of other regions of the gene since replacement of those regions would not be detected if they shared a similar sequence or function.

Our understanding of ion channels has facilitated efforts to find more ion channel genes in computer

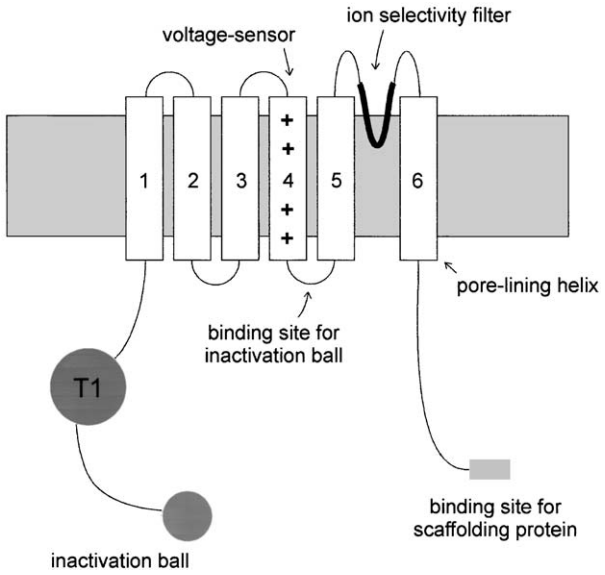


Figure 8 The modular design of voltage-gated potassium channels. Regions of the gene perform separate functions. Some of these functions can be transferred to other ion channels by transplanting that region alone.

databases. The parts of the ion channel that are functionally important are more likely to be conserved through evolution; therefore, database searches that are weighted to these residues are more likely to find homologs that may otherwise have little sequence conservation. The completion of the genomes of organisms with nervous systems, such as the fruit fly, *D. melanogaster*, or the worm, *Caenorhabditis elegans*, has ushered in a new set of more complex questions to be addressed. What are all the ion channels in a nervous system or in an animal? Given the redundancy of ion channel genes, what is the minimum set of ion channels needed for a functional nervous system?

B. Ion Channel Assembly

Ion channels are typically assembled from many subunits that form the pore-lining structure. The number of subunits that form ion channels varies from subfamily to subfamily. For example, four subunits of voltage-gated potassium channels assemble to form a single ion channel. Likewise, inward rectifier ion channels are tetramers. Most ligand-gated ion channels are pentamers, and gap junction ion channels are hexamers.

Ion channels that are members of the same subfamily can assemble with each other because they share molecular determinants that allow them to interact. For example, there are four subfamilies of voltage-gated potassium channels, Kv1–Kv4, related to *Shaker*. It is known that members of one Kv subfamily can coassemble together, but not with members of the other subfamilies. By carefully designing chimeras between the different Kv genes, the Jan group identified a stretch of amino acids in the N-terminal cytoplasmic domain before S1 that determines whether two Kv channels can interact with each other. This region, called the T1 domain, appears to form a structural domain that by itself can form a stable tetrameric structure. The T1 domain has been solved by X-ray crystallography. Analysis of the structure of the T1 domain from different Kv subfamilies shows that the interface between the T1 domain from each subfamily differs structurally. This suggests that members of different Kv subfamilies cannot coassemble because their T1 domains are incompatible.

Ion channels assembled from different subunits will exhibit different functional properties. Because ion channels can assemble from several genes of a given subfamily, the number of potential ion channels is expanded combinatorially. The reason for this potential diversity is unknown. One possibility is that the incorporation of certain subunits is used to regulate ion channel selectivity, gating, or biosynthesis. The inclusion of a subunit with certain sequence motifs has been shown to be able to target an ion channel to specific compartments within the cell or alter the stability of an ion channel on the plasma membrane.

In vivo, ion channels are often part of a complex with accessory proteins that can modify its functional properties or stability in much the same way that mixing subunits can do so. These other proteins are often referred to as β subunits, with the ion channel being the α subunit. β subunits can be soluble or integral membrane proteins. In some cases, the ion channel cannot be expressed in heterologous systems without its accessory subunits, suggesting that the β subunit is an essential part of the functional ion channel complex.

C. The Three-Dimensional Structure of Ion Channels

The final frontier is to be able to visualize the three-dimensional structure of ion channels. In 1998, Rod MacKinnon and colleagues reported the three-dimensional structure of a potassium channel and Doug Rees

and colleagues reported the structure of a mechanosensitive channel using X-ray crystallography. High-resolution structural studies of ion channels are difficult to perform because it is difficult to obtain sufficient quantities of ion channel protein for crystallization experiments and it is harder to coax membrane proteins into forming crystals than it is to coax soluble proteins. Both MacKinnon and Rees used bacterial ion channels that are easier to produce recombinantly.

Rod MacKinnon and colleagues determined the structure of the KcsA potassium channel from the bacteria *Streptomyces lividans* (Fig. 9a). KcsA has two transmembrane segments, which might place it in the same category as inward rectifier potassium channels. However, sequence analysis of KcsA suggests that it is more closely related to voltage-gated potassium channels than inward rectifiers. In support of this view, KcsA interacts with a peptide toxin, agitoxin2, that blocks *Shaker* but not inward rectifiers. One of the most satisfying aspects of the channel structure has been how well it agrees with predictions from functional studies of potassium channels. The structure is 45 Å long and the pore is narrow at the top, where the selectivity filter is predicted to be located. The helical structure resembles an "inverted teepee," and in the center of the ion channel there is a wide cavity that MacKinnon refers to as the "lake." As predicted by earlier experiments, the ion selectivity filter is formed from residues in the P region, and the carbonyl oxygens of GYG are held at a diameter that is optimal for allowing K^+ to pass. Previously we stated that ion channels catalyze the transfer of ions across an electrostatic barrier, the plasma membrane. The KcsA structure depicts two elegant mechanisms by which an ion channel overcomes the barrier. First, a water-filled lake reduces the electrostatic barrier by simply surrounding the ion in an aqueous environment. Second, the negative ends of the dipoles formed by four pore helices point toward the center of the lake, forming a point of negative electrostatic potential in the center of the membrane that is favorable for a cation.

Doug Rees and colleagues determined the structure of a mechanosensitive ion channel, MscL, from *Mycobacterium tuberculosis* (Fig. 9b). MscL is activated when lateral tension is applied to the lipid bilayer and is used by bacteria to rapidly release intracellular solutes when they are placed in a hypoosmotic environment. One motivation for studying MscL was to understand how mechanical stress could gate this ion channel. Since the open MscL channel has been shown to be able to pass a whole protein, thioredoxin, it is likely that the structure in Fig. 9b represents the closed channel. It is possible that normally lateral pressure in the membrane clamps the ion channel shut. When this pressure is released, MscL may expand into the open state.

IV. ION CHANNELS BY FAMILY AND FUNCTION

So far, we have discussed ion channels in the general sense. In the short time since the cloning of the nicotinic acetylcholine receptor, scientists have cloned and identified many ion channels, but it is clear that there are more ion channels that have yet to be discovered. An equally daunting task is to investigate the role that ion channels play in physiology. Although a complete review of the ion channel literature is beyond the scope of this articles in this section we introduce the major ion channel families and attempt to explain why some of these ion channels have attracted the interest of scientists.

A. Voltage-Gated Ion Channels

Broadly, voltage-gated ion channels are involved in the generation of electrical signals in excitable cells such as neurons (Table I). Voltage-gated sodium channels are activated when the membrane potential reaches a certain threshold potential, and they contribute to the rapid depolarization of the membrane potential. Some invertebrate species lack voltage-gated sodium channels. In these animals, voltage-gated calcium channels may partially fulfill the roles of voltage-gated sodium channels. Voltage-gated calcium channels also mediate the entry of Ca^{2+} in response to depolarization. In nerve terminals, activation of voltage-gated calcium channels by axonal action potentials is a critical step in the release of synaptic vesicles.

The voltage-gated ion channels are related in structure. Potassium channels contain one domain of six transmembrane segments per subunit, whereas sodium and calcium channels contain four domains in one large α subunit. The distinguishing feature of this group of ion channels is that they are sensitive to changes in the membrane potential. A remarkable feature of these ion channels is that they contain basic amino acids interspersed in the fourth transmembrane segment (S4). These charges could potentially sense the electric field across the membrane, and movement of S4 could transmit changes in the membrane potential to the gate. Although S4 is likely a part of the

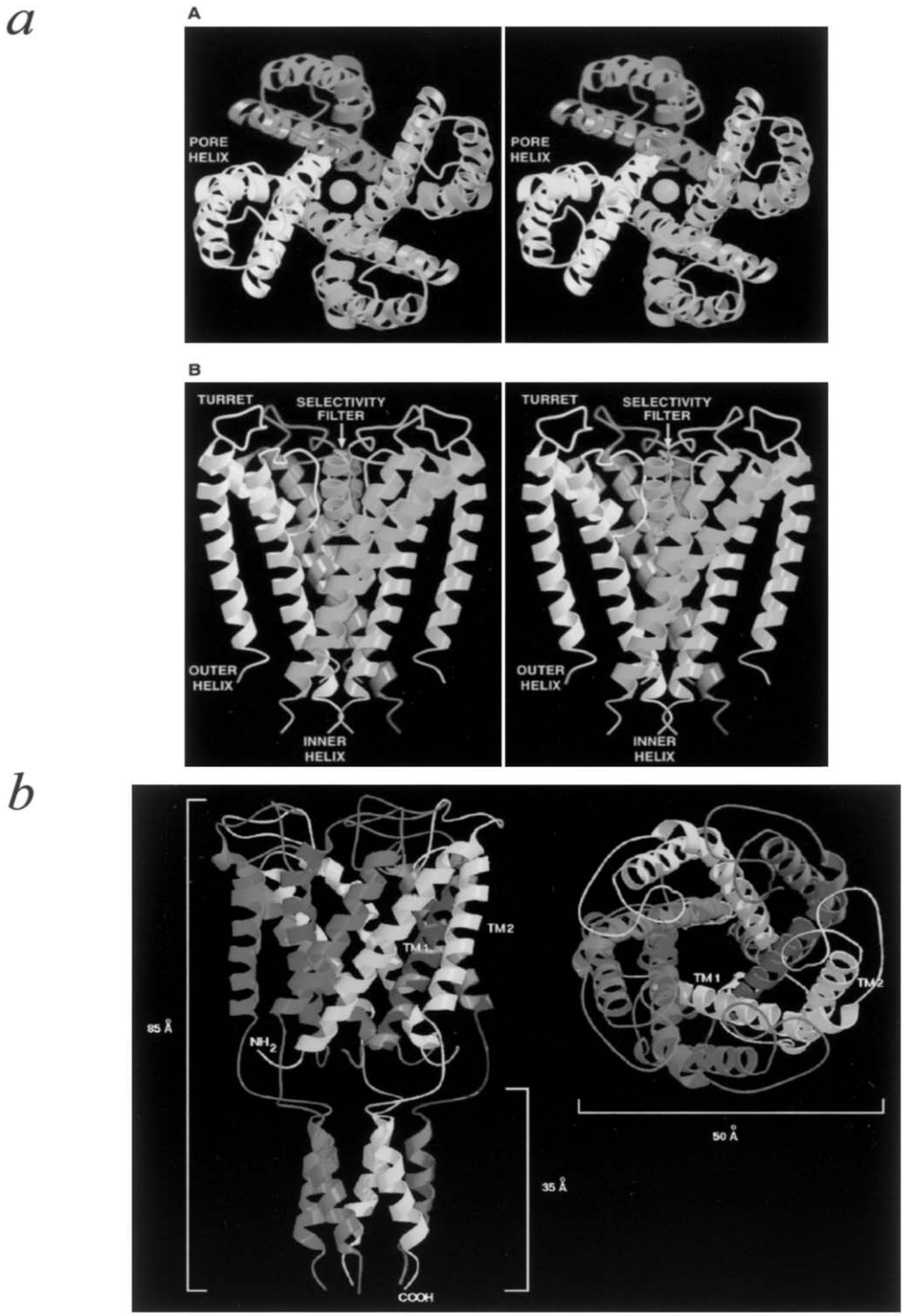


Figure 9 The three-dimensional structure of ion channels. (a) The top panels show stereoviews of the KcsA potassium channel as viewed from above the plasma membrane. KcsA is a tetramer and each of the subunits are shaded separately. The lower panel shows a side view (reprinted with permission from D.A. Doyle *et al*., The structure of the potassium channel: Molecular basis of K conduction and selectivity. *Science* **280,** 73. Copyright © 1998 American Association for the Advancement of Science). (b) The side and axial views of the MscL mechanosensitive ion channel from *M. tuberculosis*. MscL is a pentamer (reprinted with permission from G. Chang *et al*., Structure of the MscL homolog from *Mycobacterium tuberculosis*: A gated mechanosensitive ion channel. *Science* **282,** 2223. Copyright © 1998 American Association for the Advancement of Science).

Table I
Voltage-Gated Ion Channels

Ion channel	Ion selectivity	Function
Voltage-gated sodium channel	Na^+	Generates the upstroke in action potentials
Voltage-gated calcium channel	Ca^{+2}	Maintains the plateau of action potentials; allows the entry of Ca^{+2} in nerve terminals and muscle cells
Voltage-gated potassium channel	K^+	Generates the downstroke in the action potential; controls the frequency of action potentials; transports K^+ across membranes
Hyperpolarization-activated cation channel	Na^+, K^+	Mediates I_h and I_f; generates synchronous firing patterns in cells

mechanism by which ion channels sense voltage, precisely how voltage-gated ion channels respond to changes in the membrane potential and how this change is transmitted to the gate are still not completely understood.

Voltage-gated sodium channels and calcium channels are important targets of drugs used in the treatment of hypertension, cardiac arrhythmias, epilepsy, and pain. Voltage-gated sodium channels are famous for being the target of tetrodotoxin (TTX), a naturally occurring toxin most commonly associated with the pufferfish *Fugu* (it has been proposed that tetrodotoxin is actually synthesized by microorganisms living on the pufferfish). TTX blocks voltage-gated sodium channels with high affinity and thus inhibits action potentials, making it a potent neurotoxin. Marine species use TTX either to paralyze prey or to discourage natural predators. In Japan, fugu is treasured as a delicacy since low doses of TTX can produce paresthesia around the mouth when consumed (ironically, encouraging a predator). Fugu can only be prepared by licensed chefs; however, deaths from fugu poisoning are still reported.

In addition to their role in generating action potentials, voltage-gated potassium channels are involved in a wide array of other functions. This is reflected by the expression of voltage-gated potassium channels by cells outside the nervous system. The voltage-gated potassium channel family is much more diverse than the voltage-gated sodium and calcium channel families. Since the cloning of *Shaker*, the family of voltage-gated potassium channels has grown to include at least nine different subfamilies. One group, the calcium-activated potassium channels, is sensitive to changes in intracellular Ca^{2+} concentration as well as membrane potential. This feature allows them to help control the frequency of firing of action potentials. The entry of Ca^{2+} following repeated action potentials helps to activate these channels, which prolongs the duration of the undershoot of the membrane potential following the spike. Longer afterhyperpolarizations can slow the rate of action potential firing or stop it.

A new class of voltage-sensitive ion channels that mediates a hyperpolarization-activated cation current was recently identified. This current, I_h or I_f, is also referred to as the pacemaker current because it is important in establishing the rhythmic oscillatory firing of action potentials. In the heart, I_f regulates the beat-to-beat variations in the heart rate, and in the brain it helps generate synchronous oscillations in neuronal networks that can be observed in electroencephalograms. The gene that encodes the pacemaker current appears to be a cousin of voltage-gated potassium channels; however, it is not apparent how this channel is activated by hyperpolarization while other voltage-gated ion channels are activated by depolarization.

B. Ligand-Gated Ion Channels

Ligand-gated ion channels are activated upon the binding of a neurotransmitter to the ion channel and are involved in fast synaptic transmisssion in the nervous system. Many of the ion channels in Table II have a wide tissue distribution outside of the nervous system and have other functions beyond synaptic transmission. These neurotransmitters include acetylcholine, γ-aminobutyric acid (GABA), glycine, glutamate, and serotonin (5-hydroxytryptophan), and the ion channels are more commonly referred to as receptors. Several of these neurotransmitters also activate a second distinct type of receptor. This second receptor type is a heptahelical receptor that produces intracellular second messengers through the modulation of G proteins. Most commonly in the context of glutamate receptors, the ligand-gated ion channel is referred to as an ionotropic receptor, whereas the G

Table II
Ligand-Gated Ion Channels

Ion channel	Ion selectivity	Function
Nicotinic acetylcholine receptor	Na^+, K^+	Fast synaptic transmission in the nervous system and at the neuromuscular junction
$GABA_A$ receptor	Cl^-	Fast inhibitory transmission in the brain
Glycine receptor	Cl^-	Fast inhibitory transmission, predominantly in the brain stem and spinal cord
Serotonin receptor	Na^+, K^+	Fast synaptic transmission
Glutamate receptor	Na^+, K^+, (Ca^{+2})	Fast excitatory transmission in the central nervous system; mediates ischemic neuronal cell death; involved in learning and memory?
P_{2X} receptor	Na^+, Ca^{+2}	Fast synaptic transmission; sensation of pain in the trigeminal system

protein-coupled receptor is referred to as a metabotropic receptor. A familiarity with the nomenclature is helpful in determing whether a neurotransmitter receptor is ionotropic or metabotropic.

On the basis of sequence similarity, it was originally assumed that these ligand-gated ion channels formed one superfamily. The current view is that glutamate receptors have a different transmembrane topology and belong to a distinct family of ligand-gated ion channels.

The nicotinic acetylcholine receptor is a pentamer, and each subunit contains four transmembrane segments with the N and C termini outside of the membrane. Among ligand-gated ion channels, the nicotinic acetylcholine receptor has been the most thoroughly characterized. This is due to the fact that the *Torpedo* electric ray has been a source of abundant amounts of the protein and this receptor mediates synaptic transmission at the neuromuscular junction, which is a synapse that has been amenable to biophysical studies. At the neuromuscular junction, acetylcholine is released from the motoneuron and diffuses across the synaptic cleft to activate the receptor. The opening of ion channels depolarizes the membrane and generates an action potential in the muscle that initiates muscle contraction.

The nicotinic acetylcholine receptor derives part of its name from nicotine, which is an agonist of this ion channel. The nicotine in tobacco products mediates the stimulatory and addictive effects of smoking and suggests that nACh receptors in the brain play a role in cognition as well as the activation of reward pathways. Along similar lines, experimental evidence suggests that patients with Alzheimer's disease have depleted levels of acetylcholine, indicating the importance of nicotinic acetylcholine receptors in normal cognitive functioning.

GABA is the major inhibitory neurotransmitter in the brain. $GABA_A$ receptors are ligand-gated ion channels, whereas $GABA_B$ receptors are G protein-coupled receptors. In the brain stem and spinal cord, glycine is also used as a neurotransmitter in inhibitory synapses. Both $GABA_A$ and glycine receptors are chloride channels that when activated hyperpolarize the cell and make it more difficult for excitatory neurotransmitters to depolarize the membrane. Clinically, $GABA_A$ receptors are the target of benzodiazepines, which are used as sedatives, muscle relaxants, and anticonvulsants. Benzodiazepines work by a distinctive mechanism. They do not activate $GABA_A$ receptors by themselves. Instead, they facilitate the action of GABA that is released by binding to an allosteric site on the ion channel.

Glutamate is the major excitatory neurotransmitter in the brain, and many scientists believe that glutamate receptors are involved in learning and memory. Under pathological conditions, glutamate receptors also mediate the neurotoxicity associated with cerebral ischemia. By analogy to the nicotinic acetylcholine receptor, glutamate receptors were originally believed to have four transmembrane segments. This model has since been revised so that the second transmembrane segment does not traverse the plasma membrane, leaving three membrane-crossing segments—M1, M3, and M4 (the numbering of the transmembrane segments was not changed). This moved the M3–M4 loop to the outside of the cell and the C terminus to the interior of the cell.

The glutamate receptors are heterogeneous and classified based on sequence and pharmacological profile into NMDA receptors, AMPA receptors, and kainate receptors. Many of the glutamate receptors have been cloned. The distinctive features of NMDA receptors versus non-NMDA receptors fit well with

proposed mechanisms of neuronal plasticity. Near the resting potential, the majority of fast excitatory neurotransmission is mediated by AMPA receptors, which are permeable to Na^+ and K^+. NMDA receptors carry little current since they are blocked by Mg^{2+} at resting potentials. With a strong stimulus, however, enough AMPA receptors are activated to depolarize the cell and relieve the Mg^{2+} block of NMDA receptors. This is significant because NMDA receptors are highly permeable to Ca^{2+}. The entry of Ca^{2+} can initiate a signal transduction cascade that produces a long-lasting potentiation in subsequent synaptic potentials. Mechanisms of use-dependent enhancement of synaptic efficacy are often implicated in models of learning and memory.

Given the significance of Ca^{2+} permeability in glutamate receptors, the low Ca^{2+} permeability of AMPA receptors is achieved by a remarkable genetic mechanism. Sequencing of the gene and cDNAs of a particular AMPA receptor subtype (GluRB) revealed a discrepancy in the codon at a position in the second transmembrane segment. The gene contains a glutamine (Q) codon, whereas the cDNA contains an arginine (R) codon. It was later discovered that this codon is edited posttranscriptionally by the action of an enzyme in the nucleus that converts an adenine base into inosine. This single position, referred to as the Q/R site, also controls the Ca^{2+} permeability of AMPA receptors. In heterologous systems, AMPA receptors with Q in the Q/R site display high Ca^{2+} permeability, whereas AMPA receptors with R in the Q/R site display low Ca^{2+} permeability.

C. Inward Rectifiers and Two-Pore Potassium Channels

Inward rectifier potassium channels constitute another family of potassium channels distinct from voltage-gated potassium channels. Inward rectifiers are involved in maintaining the resting membrane potential near E_K or mediating the transport of K^+ across membranes (Table III). Some subfamilies of inward rectifiers respond to intracellular effectors that are produced by the activation of surface receptors. The activity of GIRKs (Kir3.0) is activated by the binding of the $G\beta\gamma$ subunits of G proteins. The activation of GIRKs is associated with the generation of slow inhibitory postsynaptic potentials in the brain and the slowing of the heart rate in response to vagal stimulation.

Inward rectifiers have the simplest structural plan of any ion channel. Like voltage-gated potassium channels, they are tetramers and contain a P region that forms part of the ion selectivity filter, but they contain only two transmembrane segments, M1 and M2. Inwardly rectifying potassium channels, as the name implies, allow more K^+ to enter the cell than to leave the cell. This property is a result of blocking of the pore from the intracellular side by polyamines or Mg^{2+} at depolarized potentials. Two acidic amino acid residues in M2 and the C terminus of inward rectifiers have been identified as part of the binding sites for the blocking particles.

Members of the Kir6.0 subfamily associate with the sulfonylurea receptor, a member of the ABC family, and form K_{ATP} ion channels that are sensitive to intracellular levels of ADP and ATP. Therefore, these ion channels link the membrane potential to the metabolic state of the cell. This property at work is best illustrated in the feedback regulation of blood glucose levels by the controlled release of insulin. In pancreatic β cells, the metabolism of glucose generates ATP, which inhibits K_{ATP}. This depolarizes the membrane potential and activates voltage-sensitive calcium channels that allow Ca^{2+} to enter the cell. The entry of Ca^{2+} then triggers the release of vesicles containing insulin, which stimulates glucose uptake by other tissues. Clinically, this pathway is utilized to help

Table III
Inwardly Rectifying Potassium Channels

Ion channel	Ion selectivity	Function
ROMK (Kir 1.0)	K^+	Salt reabsorption in the distal kidney
IRK (Kir 2.0)	K^+	Setting the resting membrane potential
GIRK (Kir 3.0)	K^+	Slowing the heart rate and slow inhibitory postsynaptic potentials in the brain
Kir 6.0	K^+	With the sulfonylurea receptor, forms K_{ATP}; regulates hormone release; cardioprotection during ischemia; regulates vascular tone

manage individuals with diabetes mellitus with oral hypoglycemics drugs that block K_{ATP} and stimulate the release of insulin.

Two-pore channels are a newly identified family of potassium channels. Their name derives from the presence of two P regions per gene. They have four transmembrane segments and resemble two inward rectifier subunits fused together. By analogy to inward rectifiers, it is believed that two-pore channels are dimers and not tetramers. Unlike inward rectifiers, two-pore channels display an outwardly rectifying current and are active at rest. Two-pore channels can be modulated by arachidonic acid or other unsaturated fatty acids, and inhibition of two-pore channels is a potential mechanism of excitatory neurotransmission. Evidence suggests that activation of two-pore channels may mediate some of the anesthetic effects of volatile gases such as chloroform or ether.

D. Chloride Channels

In addition to the $GABA_A$ and glycine receptors described previously, the chloride channels include the CLC family, CFTR, and other channel families that have yet to be identified. Chloride channels are involved in the regulation of cell volume, the transport of Cl^-, pH homeostasis, and membrane excitability. The CLC channels contain numerous hydrophobic stretches; however, the membrane topology of the CLC channels is unknown, with various models having 8–12 transmembrane segments.

The CFTR is a chloride channel that was cloned from molecular genetic studies of patients with cystic fibrosis. Its major role is in the transport of Cl^- across epithelial cells in many organs, such as the pancreas, lung, sweat glands, and kidneys. CFTR is a member of the ATP-binding cassette (ABC) family of ion transporters. ABC proteins are generally known for mediating the ATP-driven transport of substances; for example, the MDR protein pumps chemotherapeutic drugs out of the cell. Therefore, the CFTR ion channel is a unique member of this family.

Ion channels are fantastic molecular machines. Their functions are fundamental—they act as valves for ions to move into or out of the cell—and thus they play important roles in many physiological processes. The study of ion channels is challenging but rewarding since it ranges from exploring the function of ion channels on the cellular and systems level down to the atomic level to investigating how ion channels as proteins work. The first crystal structures of ion channels have ushered in a new era in ion channel research. With more structures and functional studies of ion channels and the cloning of ion channels involved in channelopathies, the next decade promises to be an exciting time for ion channel research.

See Also the Following Articles

ELECTRICAL POTENTIALS • GABA • NEUROTRANSMITTERS

Suggested Reading

Ackerman, M. J., and Clapham, D. E. (1997). Ion channels—Basic science and clinical disease. *N. Engl. J. Med.* **336,** 1575.

Aidley, D. J., and Stanfield, P. R. (1996). *Ion Channels: Molecules in Action*. Cambridge Univ. Press, Cambridge.

Ashcroft, F. M. (2000). *Ion Channels and Disease.* Academic Press, San Diego.

Dani, J. A., and Mayer, M. L. (1995). Structure and function of gluamate and nicotinic acetylcholine receptors. *Curr. Opin. Neurobiol.* **5,** 310.

Doyle, D. A., *et al.* (1998). The structure of potassium channel: Molecular basis of K conduction and selectivity. *Science* **280,** 69.

Foskett, J. K. (1998). CIC and CFTR chloride channel gating. *Annu. Rev. Physiol.* **60,** 689.

Hille, B. (1992). *Ionic Channels of Excitable Membranes*, 2nd ed. Sinauer, Sunderland, MA.

Jan, L. Y., and Jan, Y. N. (1997). Cloned potassium channels from eukaryotes and prokaryotes. *Annu. Rev. Neurosci.* **20,** 91.

Johnston, D., and Wu, S. M. (1995). *Foundations of Cellular Neurophysiology.* MIT Press, Cambridge, MA.

Nicholls, J. G., Martin, A. R., and Wallace, B. G. (1992). *From Neuron to Brain*, 3rd ed. Sinaver, Sunderland, MA.

Seeburg, P. H. (1993). The molecular biology of mammalian glutamate receptor channels. *Trends Neurosci.* **16,** 359.

Siegelbaum, S. A., and Koester, J. (2000). Ion channels. In *Principles of Neural Science* (E. R. Kandel, J. H. Schwartz, and T. M. Jessell, Eds.), 4th ed. Elsevier, New York.

Language Acquisition

HELEN TAGER-FLUSBERG
Boston University School of Medicine

GLOSSARY

binding principles According to the syntactic theory, known as government binding theory, these are the rules of our grammar that dictate the relation between words, for example, pronouns and their referents.

communicative competence Linguistic competence and knowledge of the social rules for language use.

constraints Limits or biases that children bring to the task of language acquisition. A constraint may dictate a cognitive or pragmatic strategy in the interpretation of words.

grammar The finite set of rules shared by all speakers of a language that allows all possible sentences to be generated.

intentional communication Any communicative act that a person engages in purposefully.

language A symbolic system, based on syntactic, semantic, and phonetic features, that allows mutually intelligible communication among a group of speakers.

learnability Various models of language acquisition based on several assumptions concerning the nature of children, known learning mechanisms, the structure of language, and the logical inferences that can be drawn from these assumptions.

morphology The rules that govern the use of free or bound morphemes, the minimal meaningful units of language.

overregularization errors A common error among children that involves applying regular and productive grammatical rules to words that are irregular or exceptions.

parameter A type of linguistic switch that is set after children are exposed to particular forms in their native language; it is one of a finite number of values along which languages are free to vary.

phoneme A speech sound that can signal a difference of meaning in a language.

semantics The study of the meaning system of language.

speech acts Aspects of the pragmatic systems referring to different functions of utterances, such as requesting or promising.

theory of mind Knowledge about the mind and mental states of other people that is used to interpret their actions and the intended meaning of utterances.

universal grammar Hypothetical set of restrictions governing all forms that human languages may take.

Language acquisition refers to the process of achieving the ability to speak and understand the particular language or languages to which a child has been exposed. By the time most children reach the age of 5, they are highly competent speakers of their native language. During these early years of development, children acquire the ability to perceive and produce the speech sounds of the language to which they are exposed and the phonological rules for combining them to create words. They also acquire a large and varied vocabulary and the rules for combining them into complex grammatical sentences with correct morphology to mark tense, mood, number, and so forth. Finally, they become proficient users of this linguistic system to perform a range of different speech acts appropriate to varied social contexts. These remarkable achievements in the acquisition of language occur without explicit instruction or even significant feedback from others. Language is a complex, componential system composed of an abstract phonological rule system and lexicon, syntax, morphology, pragmatic, and discourse rules. These components depend on the development of different cognitive and neural mechanisms that interact over the course of acquisition.

I. PRELINGUISTIC DEVELOPMENTS

A. Speech Perception

Infants come into the world prepared to acquire language. At birth they are able to distinguish speech from other sounds and, indeed, to perceive and discriminate speech sounds in the same way as adults. They can discriminate between syllables that differ by a single phonetic feature (e.g., ba vs pa). Studies have demonstrated that during the first months of life infants discriminate all phonetic contrasts between speech sounds in their native language as well as those occurring in other languages to which they have not been exposed.

During the second half of the first year speech perception abilities undergo significant changes and reorganizations. As a result of continued exposure to the sound properties of their native language, infants show a reduction and eventual loss in the capacity to discriminate speech sound contrasts that are present only in foreign languages. At the same time, by the age of 6 months infants become more attuned to the sound structure of their native language, preferring the prosodic patterns of their own language as opposed to those found in languages they have not heard. By the age of 9 months, infants also prefer listening to word lists in which the words follow the constraints on ordering phonetic segments in their native language over those in which the words violate those constraints.

B. Social Developments

From the start, these sophisticated speech perceptual abilities are closely tied to the infant's social experience. Studies of newborns have shown that they distinguish their own mothers' voices from those of other mothers, which is most likely related to prenatal exposure to the acoustic properties of the mothers' speech. In the visual domain, newborns also show a preference for human faces and can even imitate facial expressions. Thus, from the beginning, the social niche for language is clearly established.

During the first few months of life, rapid changes take place. Mothers and their infants begin to interact in a finely tuned way with one another. They synchronize their patterns of eye gaze, movements, and facial expressions of affect in ways that resemble turn-taking patterns in conversations. By the age of 4 months, there is a marked increase in vocal turn-taking during these rich interactions between infants and their caretakers. Toward the end of the first year of life, vocalizations as well as other nonvocal behaviors again become genuinely integrated into social interaction as infants' developing social cognitive capacities lead to the onset of intentional communication. At this point, infants becomes capable of coordinating their attention to objects or events with other people through eye gaze patterns (joint attention), gestures, and vocalizations. This developmental achievement is generally viewed as a critical step in language acquisition, with the onset of communicative intent. Infants at this stage are able to communicate a variety of meanings, including protodeclaratives, which involve pointing or other gestures to draw another person's attention to an object of interest, and protoimperatives, meaning a gesture or vocalization to express a request or demand for an object. The significance of these communicative attempts is that they suggest the infant is capable of understanding the intentions of others (the beginning of a theory of mind), at least in a rudimentary or implicit form.

C. Speech Production

During the first year of life the infant's vocal abilities are also changing. Initially, because of anatomical limitations, newborns produce mostly cries or occasional gurgling. Infants begin producing sounds at approximately the age of 2 months, with the onset of cooing—vocalizations produced in the back of the mouth. By 4 months, most infants engage in vocal play, including a broad range of different kinds of sounds such as some rudimentary consonant–vowel (CV) syllables. By 6 months, infants produce canonical babbling, consisting of systematic CV syllables with adult-like timing. During this stage, infants are sensitive to the feedback they receive from their own babbling, and there are increases in the variety of consonants produced. By 10 months, infants babble in a conversational or modulated manner, with more complex strings of sounds that have varied intonation. These stages form the backdrop against which phonological development takes off during the second year of life.

One question that has been addressed in research on babbling is whether there is a connection between the sounds that an infant selects to produce and those in the target language in the infant's environment. A

second question concerns the relationship between babbling and the onset of language. There is growing evidence suggesting that there is important continuity between prelinguistic babbling and the first words produced by infants. Not only do these stages overlap but also there is considerable overlap in sounds incorporated in babbles and early words. Furthermore, research on the development of vocal babbling in deaf infants suggests that the auditory feedback that infants receive from their own as well as others' sounds is crucial in shaping the course of development. Thus, recent studies have shown that deaf infants begin babbling later than hearing infants and in a more limited way. Even the sounds are somewhat different, presumably because deaf infants lack auditory feedback from their own sounds and have no access to the speech sounds in their environment. At the same time, deaf infants exposed to sign language during the first year of life begin manual babbling: They produce repetitive sequences of sign-language formatives that parallel the syllabic units of vocal babbling.

D. Mechanisms Underlying the Capacity for Language Acquisition

The development of the basic capacities that form the foundation of language acquisition in infancy is largely determined by the maturation of linguistic, motor (articulatory or manual), and social mechanisms that undergo developmental changes as a result of exposure to and feedback from particular language environments. Two distinct biologically based mechanisms are considered crucial for language acquisition: a domain-specific language system and a social mechanism that is dedicated to the development of understanding of other minds. Normal development of language depends on the integration of these distinct mechanisms, both of which depend on stimulation from the linguistic and social environment.

II. PHONOLOGICAL ACQUISITION

A. Stages of Development

As noted previously, it is generally agreed that there is essential continuity between prelinguistic babbling and the earliest stages of phonological development evidenced in the child's first words. The majority of sounds produced in the earliest words of children are the same as those preferred in their babbles. Initially, children's words are composed of simple CV syllable structures, using a relatively small inventory of sounds. Gradually, over time and with growth in the child's vocabulary, there is an expansion in the range of sounds produced by children. Although there is no universal order in the acquisition of phonological features, certain regularities have been found in the phonological sounds that are used across children. Mastery over vowels occurs before consonants. The main consonant classes that are used earlier in development include stops (e.g., b and d), nasals (e.g., m and n), and glides (e.g., w). Later developing consonants include fricatives (e.g., v) and liquids (e.g., l and r).

B. Systematic Error Patterns

As children begin producing more elaborated syllable structures and a wider range of sounds, they begin producing speech sound errors. The most striking feature of these errors is that although there are individual differences in the particular kinds of errors made, the errors are not random but instead fall into common patterns. In children with more severe articulation difficulties, words may be produced that involve combinations of different error patterns.

1. Omissions

Children will often omit syllables or specific sounds as they attempt to reproduce more complex adult words. Typically, unstressed syllables will be omitted—those occurring at the beginning of words (e.g., "mato" for "tomato") or in the unstressed medial position (e.g., "e'phant" for "elephant"). Consonant clusters will lead to omissions. For example, in English, one common error is to omit the /s/ in /s/ + stop consonant clusters (e.g., "top" for "stop"). Sometimes, final consonant sounds will be omitted (e.g., "go" for "gone").

2. Feature Changes

Another class of error patterns is the changing of sounds at the level of individual articulatory features. For example, voiced consonants may be changed to unvoiced consonants (e.g., bot for pot or gat for cat). Place changes also may be found in some children, with back consonants becoming more frontal (e.g., dame for game and bup for cup). These kinds of errors demonstrate the significance of features in children's phonological representations.

3. Assimilation Errors

A third class of errors illustrates how the child's representation of the target word may influence the kinds of sound substitutions that are made. Assimilation errors entail the change in one sound in the target word to make it more similar to another sound in that word. Such errors may involve assimilation in different feature classes, such as voicing (e.g., "doad" for "toad") or place ("gog" for "dog").

C. Theoretical Explanations

One of the main debates in the literature on phonological development has been between those advocating the view that very early phonological development is a discontinuous stage, which does not map onto the adult system, and those who argue that from the beginning young children's phonological systems share properties of the adult system. More evidence has accrued in favor of the latter view, called the continuity hypothesis, which suggests that the same underlying mechanisms are used in phonological acquisition as in the adult speaker.

One important piece of evidence for the continuity hypothesis is that infants are capable of perceiving speech in a mature, adult-like way before they even begin producing their first words. Furthermore, it is clear from the developmental and error patterns described here that children's attempts at producing target words are guided by abstract representations of speech sounds, including representations of syllable structures and distinctive articulatory features. At the earliest stages, the representations of syllable structures are very simple (e.g., simple CV structures referred to as minimal words), and these become more elaborate and complex over time. Similarly, children's earliest words may only include a small group of distinctive features that are considered marked (such as labials). Again, these expand over time, until by age 3 or so, the majority of children have mastered the phonology of their native language.

III. LEXICAL–SEMANTIC DEVELOPMENT

A. Stages of Development

Lexical development can be divided into three periods. The first period covers the acquisition of the initial 50 words or so, during which children are learning what words do. At this stage, some words appear to be tied to particular contexts and serve primarily social or pragmatic purposes. Word learning during this initial phase is relatively slow and uneven. A word may be equivalent to a child's holistic representation of an event. Especially in Western middle-class children, the child's vocabulary at this stage is dominated by names for objects, including animals, people, toys, and familiar household things. There will also be some social words (e.g., "hi" and bye), modifiers (e.g., "more" and "wet"), and relational terms that express success, failure, recurrence, direction and so forth.

By the middle of the second year, there is a significant increase in the rate at which children acquire new words. This new period is usually referred to as the vocabulary spurt, or naming explosion, and may be punctuated by many requests from children for adults to label things in the world around them. Words are learned very quickly, often after only a single exposure that may take place without any explicit instruction. This process of rapid word learning is referred to as "fast mapping." This phase of vocabulary growth is marked by a close relationship between lexical and grammatical development.

By the time children reach their third birthday, they begin to develop a more organized lexicon, in which the meaning relations among groups of words are discovered. For example, at this time children begin to learn words from a semantic domain, such as kinship, and they are able to organize the words according to their similarities and differences on dimensions of meanings. For nouns labeling concrete objects, children begin to organize taxonomies, also learning words at the superordinate and subordinate levels and understanding the hierarchical relations among terms such as dachshund, dog, and animal. Semantic developments at this stage will often lead to reorganizational processes as these kinds of relationships among words are realized by the child. The rate of word learning continues to be very rapid, with estimates suggesting that children acquired about 15–20 new words a day during the preschool years and beyond.

B. Developmental Processes

1. Conceptual Development

The fact that children can grasp the meaning of a word without explicit instruction, and in a variety of circumstances, suggests that there is a significant role played by preexisting or ongoing developing conceptual representations. During the early phases of word

learning, studies demonstrate how specific conceptual developments at this stage are closely related to the acquisition of particular words. For example, infants develop the ability to retrieve hidden objects within a few weeks of acquiring words such as "*gone*" that encode the concept. In some children the words were acquired before the concept, in others the reverse was found. This suggests that at this early phase, while conceptual development can influence semantic development, it is also the case that semantic development can influence conceptual change. Thus, the relationship between language and conceptual development, or more generally between language and thought, is highly complex, with each system placing constraints on the other and both dependent on the social environment for their elaboration in development.

During the toddler years, objects tend to be named at the so-called basic object level (e.g., dog or car) rather than at the subordinate (e.g., dalmatian and Mercedes) or superordinate (e.g. animal and vehicle) levels. Objects within the same category at the basic object level tend to share perceptual and functional features, and they do not overlap with related semantic categories. Thus, this level may be the most useful for children for both functional and cognitive reasons. Parents also have been shown to name objects for children at the basic object level and this too might explain why this is the preferred level for children's early words.

Once a new word is learned it is quickly generalized to new contexts. Much of the focus of research on word meanings has been on the extension of a word. At this stage, children will sometimes overextend the meaning of a word, broadening the use of a term beyond its semantic boundaries. Typical examples include calling all women "Mommy" or using "ball" to name any round object. Overextension errors may be made on the basis of functional or, more frequently, perceptual similarity, or they may involve an associative complex of features. Another kind of extension error that is not so easily noticed occurs when the child *underextends* the use of word, not using a word to label an appropriate referent. Underextension errors tend to be noted at earlier stages of lexical development, whereas overextension errors are more typical of this period, after the naming explosion.

The most widely accepted view of what guides the acquisition of word meanings at this stage is the child's conceptual representations. The initial representation may be of a particular referent to which new examples are compared. Later, this semantic representation becomes more abstract and may be composed of a composite image or set of features for a prototype or best exemplar. This theory can explain both underextension and overextension errors; however, it is a theory of lexical development that is most usefully applied to the child's acquisition of names for concrete objects but not to other kinds of word classes.

2. Contextual Influences

Even during the earliest stages of word learning, children rarely make errors about the mapping between a word and its referent. Despite the fact that children may not have words for most concepts in the world, they almost always hone in on the correct meaning of a word learned in various social contexts. One important developmental process that makes this possible is that children understand other peoples' intentions and bring this knowledge of other minds to the task of word learning. Thus, they will carefully observe what a speaker is looking at or playing with when a new word is spoken. Children monitor the speaker's line of regard and assume that a novel word refers only to the object that is in the attentional focus of the speaker, indicating that they are sensitive to subtle cues to a speaker's referential intentions.

3. Constraints on Word Meaning

A number of researchers have argued that what makes word learning possible, especially when the child is capable of fast mapping, is a set of constraints that guides the child's hypotheses about the possible meanings of words. Eve Clark proposed a very general kind of constraint, called the *principle of contrast*, which states that every two words in a language contrast in meaning. This principle operates in conjunction with the *principle of conventionality*, which states that there are conventional words that children expect to be used to express particular meanings so that if a speaker does not use the conventional word, then the child assumes that the new word must have a somewhat different meaning. A different version of the principle of contrast is called the *mutual exclusivity constraint*. This constraint leads the child to assume that each object only has a single name, and that a name can only refer to one category of objects. When children hear a new word, they will look around for a referent for which they do not currently have a label. This explains why young children are reluctant to accept superordinate labels for individual objects. Other constraints that have been proposed include the *whole-object constraint*, which states that new words

refer to whole objects rather than parts of objects (if, however, the child already knows the name of the object, then the word might be considered as labeling a part or property of the object), and the *taxonomic constraint*, which states that words refer to categories of objects.

Although some view these kinds of constraints as innate principles that are specific to lexical development, others view them as more general biases that may be an aspect of broader pragmatic or cognitive processes. Although there are still disagreements about how to characterize constraints on the child's hypotheses about the meanings of new words they encounter, most researchers agree that children use these heuristics to help them with the rapid mapping of words onto underlying meaning representations.

4. Syntactic Bootstrapping

Much of the research on semantic development has focused on the acquisition of nouns. Verbs, on the other hand, pose a different kind of problem because there is often not enough information in the context to help the child distinguish between related verbs such as "look" and "see." Children need to use syntactic information to help them figure out the meanings of verbs. The particular kinds of information that children can use include the number and kind of arguments that occur with the verb. Thus, transitive verbs take object arguments, whereas intransitive verbs do not. This kind of information is useful in helping the child interpret verb meaning. Syntactic bootstrapping is also useful for helping the child distinguish mass nouns (e.g., spaghetti) from count nouns (e.g., a potato) or common nouns from proper names, and very young children have been shown to be able to use this information when they hear new words in ambiguous contexts. As children's language progresses and they begin acquiring knowledge about the syntactic frames in which words occur, they begin to integrate syntactic and semantic information in this way. This process underscores the interrelationships that drive both semantic and syntactic development.

IV. THE ACQUISITION OF SYNTAX AND MORPHOLOGY

A. Stages of Development

Before the end of the second year, soon after the spurt in vocabulary development, children reach the next important milestone in language development: They begin to combine words together to form their first sentences. This is a crucial turning point because even the simplest two-word utterances show evidence of early grammatical development. The child's task in acquiring the grammar of her native language is complex. First, children need to segment the stream of language into morphemes (the minimal unit of language that carries meaning), phrases, and sentences. They must then discover the major word classes, such as noun, verb, and determiner, and map the appropriate lexical terms into these word classes. Children then learn how to grammatically encode tense, plurality, gender, and so forth, often using morphemes that are attached to verbs or nouns. At the same time, they acquire the major rules for organizing basic phrasal units such as noun phrase (e.g., article + adjective + noun—"*The tall man*") and verb phrase (e.g., verb + tense + prepositional phrase—"*walk-ed to the park*") as well as for organizing basic sentence structures for declaratives, questions, and negation. In the final stages, children figure out the syntactic rules for complex sentences involving coordinating and embedding multiple clauses.

B. Measuring Grammatical Development

1. Production

One of the obvious ways that children's sentences change over time is that they gradually grow longer. This fact is the basis for one of the most widely used measures of grammatical development, the mean length of utterance (MLU), which is the average length of a child's utterances as measured in morphemes. The assumption underlying this measure is that each newly acquired element of grammatical knowledge adds length to the child's utterances. Studies confirm that MLU increases gradually over time, and that it is a better predictor of the child's language level than chronological age. Nevertheless, it is only valid as a measure of development up to an average sentence length of four morphemes, and it may not be useful without significant modifications as a measure for languages other than English.

2. Comprehension

It is much more difficult to measure the child's comprehension of syntactic and morphological structures. Although in naturalistic contexts young children

give the impression they understand significantly more that they say, this may reflect the child's use of nonlinguistic context and other cues rather than knowledge of abstract linguistic structure to compute the underlying semantic relations of sentences.

Methods to assess comprehension include a variety of paradigms, each of which has both advantages and disadvantages. The oldest method is the use of diary studies, which document the conditions and contexts in which a child understands or fails to understand a particular structure. Experimental procedures may include act-out tasks, in which an experimenter asks the child to enact a sentence or phase using a set of toys and props, or direction tasks, in which the child is asked to act out an event or command. Choice selection paradigms have also been developed, including picture-choice tasks, in which the child selects from a set of pictures the one that best represents the linguistic form presented by the experimenter, and preferential-looking tasks that have successfully been used with infants. In this kind of task infants listen to a linguistic message and have the choice of two videos to observe, only one of which matches the message. Infants who look reliably longer at the matching scene are credited with understanding the linguistic structure that was presented.

C. Early Word Combinations

When children begin to combine words to form the simplest sentences, most are limited in length to two words, although a few may be as long as three or four words. These early sentences are often unique and creative, composed primarily of nouns, verbs, and adjectives. In English, function words (such as articles or prepositions) and other grammatical morphemes, such as noun (e.g., plural –s) and verb inflections (e.g., past tense -ed or present progressive –ing), are usually omitted, making the child's productive speech sound "telegraphic"; however, this is less true for children learning other languages, such as Italian or Hebrew, that are rich in inflectional morphology.

1. Semantic Relations

Cross-linguistic studies of children at this stage have shown that there is a universal small set of meanings, or semantic relations, that are expressed, including agent + action, action + object, entity + location, entity + attribute, and demonstrative + entity. Children talk a lot about objects by naming them and by discussing their locations or attributes, who owns them, and who is doing things to them. They also talk about people, their actions, their locations, their actions on objects and so forth. Objects, people, actions, and their interrelationships preoccupy young children universally.

2. Limited Scope Formulae

Initial studies of utterances produced in the two-word stage found that children used highly consistent word order. Indeed, the semantic relations approach assumed that the child uses a productive word order rule that operates on broad semantic rather than syntactic categories. This research was limited by focusing primarily on languages that make extensive use of order to mark basic relations in sentences and on a small number of children. It is now acknowledged that there is considerable individual variation among children learning different languages, and even for children learning English. Nevertheless, word order rules are used at this early stage of grammatical development, but they are more limited and more narrowly defined in semantic scope than is suggested by the semantic relations approach and therefore have been called *limited scope formulae*. For some children ordered combinations of words may even be based on specific lexical items rather than on semantic categories. Over time, these more limited rules expand to encompass broader semantic and later syntactic categories and begin to resemble the adult grammar.

3. Null Subjects

One characteristic of children's two-word sentences is that they often omit the subject. Recently, this has been interpreted from the perspective of current linguistic theory, which proposes a parameter-setting approach. Some theorists argue that all children begin with the subject parameter set in the null position (which holds for languages such as Italian or Spanish) so that children learning English must eventually switch the parameter setting to the position marked for required subjects.

Although this proposal is attractive because it connects early grammar to linguistic theory, there are several criticisms of this approach. Although English-speaking children do omit subjects, in fact they include them significantly more often than Italian-speaking children, which suggests that they know that subjects need to be expressed. Subjects are probably omitted

because young children have limited processing capacity, and for pragmatic reasons subjects are more readily omitted than objects because they are often provided by the context.

D. Development of Grammatical Morphology

1. Invariant Order of Acquisition

As children progress beyond the two-word stage, they gradually begin to fill in the inflectional morphology and function words that are omitted in their early language. The process of acquiring the major grammatical morphemes in English is gradual and lengthy and some are still not fully controlled until the child enters school. Studies have found that the order in which English morphemes are acquired (e.g., articles, past tense, prepositions, or auxiliary verbs) is strikingly similar across children. The order of acquisition is not accounted for by frequency of use by the child or mother; instead, it is related to measures of both semantic and syntactic linguistic complexity.

2. Overgeneralization and Rule Productivity

One striking error that children make in the process of acquiring grammatical morphemes is the overregularization of regular forms to irregular examples. For example, the plural *-s* is frequently added to nouns that take an irregular plural, such as "*mans*" instead of "*men*" or "*mouses*" instead of "*mice*," and the regular past tense ending *-ed* is sometimes used on verbs that are marked with an irregular form, such as "*falled*," "*goed*," or "*teached*." These errors may not be frequent, but they can persist well into the school years and are quite resistant to feedback or correction. They are taken as evidence that the child is indeed acquiring a rule-governed system rather than learning these inflections on a word-by-word basis.

Other evidence for the productive use of morphological rules comes from an elicited production task introduced by Jean Berko Gleason called the Wug test. The child is shown drawings depicting novel creatures, objects, and actions and asked to supply the appropriate description that would require the inclusion of noun or verb inflections. For example, a creature was labeled a wug, and then the child had to fill in the blank for "*there are two* ____." Preschool aged children performed well on this task, demonstrating their internalized knowledge of English morphological rules that can be applied productively.

Steven Pinker has argued that two different mechanisms are involved in acquiring regular and irregular forms. Regular forms involve a linguistic rule-governed mechanism, whereas irregular forms are retrieved directly from the lexicon and thus involve a memory storage system. This dual-mechanism hypothesis has been challenged by models developed within connectionist frameworks, in which only a single mechanism is needed to compute the correct form for regular and irregular examples, after being trained on mixed input. The debate between these camps continues.

3. Cross-Linguistic Evidence

There is a growing literature on the acquisition of morphology in other languages. Overgeneralization errors have been recorded in children learning many different languages suggesting this is a universal pattern for this aspect of grammatical development. However, the slow and gradual development of English morphology does not hold up for languages that have richer morphological systems. For example, children acquiring Turkish use suffixes on nouns that mark the noun as either the subject or object of the sentence, at even the earliest stages of language development, and children learning Italian acquire verb inflections marking person, tense, and number very rapidly and in a less piecemeal fashion than has been found for English morphology. These cross-linguistic variations seem to reflect differences among languages in the amount of inflectional morphology within a language and the degree to which inflections are optional. For example, English marks verbs only for the past tense, third person singular present tense (e.g., "*he walk-s*"), or progressive aspect (e.g., "*he is walk-ing*"), whereas Italian verbs are always marked in various ways. Children appear to be highly sensitive to these differences from the beginning stages of acquiring grammar.

E. The Acquisition of Sentence Modalities

1. Simple Declaratives

As children progress beyond the two-word stage, they begin combining words into three- and then four-word sentences. In doing so, they link together two or more basic semantic relations that were prevalent early on. For example, *agent* + *action* and *action* + *object* may be linked to form *agent* + *action* + *object*. These simple

declarative sentences include all the basic elements of adult sentences. Gradually these may become enriched with the addition of prepositional phrases, more complex noun phrases that include a variety of modifiers, and more complex verb phrases including auxiliary and modal verbs. All these additions add length to the declarative sentences of young children.

2. Negation

Although children do express negation even at the one-word stage (e.g., using the word "*no!*"), the acquisition of sentential negation is not fully acquired until much later. There are three stages in the acquisition of negation in English: (i) The negative marker is placed outside the sentence, usually preceding it (e.g., "*not go movies*" and "*no Mommy do it*"); (ii) the negative marker is sentence internal, placed adjacent to the main verb but without productive use of the auxiliary system (e.g., "*I no like it*" and "*don't go*"); and (iii) different auxiliaries are used productively and the child's negations approximate the adult forms (e.g., "*you can't have it*" and "*I'm not happy*"). Although the existence of the first stage has been questioned by some researchers, there does appear to be cross-linguistic support for an initial period when negative markers are placed outside the main sentence.

Negation is used by children to express a variety of meanings. These emerge in the following order, according to studies of children learning a wide range of languages: "nonexistence," to note the absence of something or someone (e.g., "*no cookie*,"); "rejection," used to oppose something (e.g., "*no bath*"); and "denial," to refute the truth of a statement (e.g., "*that not mine*"). Some children show consistent patterns of form–meaning relations in their negative sentences. For example, one child used external negation to express rejection while at the same stage reserved sentence internal negation forms to express denial. These patterns may have had their source in the adult input.

3. Questions

There are several different forms used to ask questions, including rising intonation on a declarative sentence; yes–no questions, which involve subject–auxiliary verb inversion; *wh-* questions, which involve *wh-* movement and inversion; and tags, which are appended to declaratives and may be marked lexically (e.g., "*we'll go shopping, okay*?") or syntactically ("*we'll go shopping, won't we*?"). Children begin at the one- or two-word stage by using rising intonation and one or two fixed *wh-* forms, such as "*what that*?" Gradually, over the next couple of years syntactic forms of questions develop with inversion rules acquired simultaneously for both yes–no and *wh-* questions. Some data suggest that for *wh-* questions, inversion rules are learned sequentially for individual *wh-* words, such as "*what*," "*where*," and "*who*," "*why*," and may be closely linked in time to the appearance of those words used as *wh-* complements. Thus, syntactic rules for question formation may be *wh-* word specific in early child language.

Several studies of English and other languages have investigated the order in which children acquire various *wh-* questions and the findings have been consistent. Children generally begin asking and understanding "*what*" and "*where*" questions, followed by "*who*," then "*how*," and finally "*when*" and "*why*" questions. One explanation for this developmental sequence is that it reflects semantic and cognitive complexity of the concepts encoded in these different types of questions. Thus, questions about objects, locations, and people (i.e., *what, where*, and *who*) involve less abstract concepts than those of manner, time, and causality (i.e., *how, when*, and *why*). The early emerging *wh-* questions are also syntactically less complex in that they involve simple noun phrase replacement, whereas the later developing questions involve prepositional phrases or full sentence complements.

4. Passives

Despite the rarity of the passive construction in everyday conversations in English, a good deal of attention has been paid to how children use and understand passive sentences. Because the order of the agent and patient is reversed, this particular construction can reveal a great deal about how children acquire word order rules that play a major role in English syntax.

Elicited production tasks have been used to study how children construct passive sentences, typically using sets of pictures that shift the focus to the patient. Younger children tend to produce primarily truncated passives (e.g., "*the window was broken*") in which no agent is specified. These truncated passives generally have inanimate subjects, whereas full passives are produced by children when animate subjects are involved, suggesting that full and truncated passives may develop separately and be unrelated for the younger child. It has been suggested

that truncated passives are really adjectival, whereas the later appearing full forms are complete verbal passives.

Numerous studies have used an act-out procedure to investigate children's comprehension of passive voice sentences. Typically, these studies compare children's comprehension of passive sentences to active sentences that are either reversible, in which either noun could plausibly be the agent (e.g., "*the boy kisses the girl*" or "*the boy is kissed by the girl*"), or semantically biased, in which one noun is more plausibly the agent than the other (e.g., "*the girl feeds the baby*" or "*the girl is fed by the baby*"). Studies find that children correctly interpret the plausible passive sentences before they do the reversible sentences. Preschoolers acquiring English tend to make errors systematically on the reversible passive sentences, suggesting the use of a processing strategy, called the word-order strategy whereby noun–verb–noun sequences are interpreted as agent–action–object. Children learning languages other than English may develop different processing strategies that closely reflect the canonical ways of organizing the basic relations in a sentence in their native language. For example, Japanese is a verb-final language that marks the agent with a suffix *-ga* rather than with a fixed word order, although there is a preference for an *agent–object–verb* order. Preschool-aged Japanese children tend to use a strategy that takes the first noun marked with *-ga* as the agent of the sentence. Thus, children's processing strategies are tailored to the kind of language they are acquiring and show that preschoolers have already worked out the primary ways that their language marks the basic grammatical relations.

Studies of the acquisition of other languages such as Sesotho, in which the passive construction is very frequent because subjects always mark sentence topic, have found that children acquire the passive much earlier and use it much more productively than do English-speaking children. Again, this suggests that children are sensitive to the typology of their language and that these factors influence the timing of development for the passive.

The semantic characteristics of the verb also influence the child's comprehension of passive sentences. Although 5-year-olds do correctly understand passive sentences that have action verbs, they find it more difficult to interpret passive sentences with nonaction verbs (e.g., "*Donald was liked by Goofy*"). Thus, the acquisition of passive voice continues into the school years as the child's knowledge becomes less constrained by semantic aspects of the verb.

F. Complex Sentence Structures

1. Coordinations

As early as 30 months of age, children begin combining sentences to express compound propositions. The simplest and most frequent method children use to combine sentences is to conjoin two propositions with "*and*". One question that has been investigated in numerous studies regards the order in which different forms of coordination develop. Both sentential (e.g., "*Mary went to school and Peter went to school*") and phrasal coordinations (e.g., "*Mary and Peter went to school*") tend to emerge at the same time in development, suggesting that these forms develop independently and are not, for young children, derived from one another. Children form phrasal coordinations by directly conjoining phrases, not via deletion rules.

Semantic factors influence the course of development of coordination. Children use coordinations first to express additive meaning, where there is no dependency relation between conjoined clauses (e.g., "*maybe you can carry this and I can carry that*"). Later, temporal relations (e.g., "*Joey is going home and take her sweater off*") and then causal relations (e.g., "*she put a Band-aid on her shoe and it maked it feel better*") are expressed, suggesting that children begin demonstrating greater semantic flexibility even while limiting themselves to the use of a single connective, "*and*."

2. Relative Clauses

Sometime after children begin using coordination, relative clauses emerge in their spontaneous speech. Initially, they are used to specify information exclusively about the object of a sentence (e.g., "*let's eat the cake what I baked*"), and often the relative pronoun is omitted or incorrect. The use of relative clauses in the spontaneous speech of young children is quite rare, perhaps because children avoid these syntactically complex constructions or because they lack the occasion to use them when the context is shared by the speaker and listener.

Elicited production techniques have been used successfully with preschoolers. These studies have also found that children find it easiest to add relative clauses to the ends of sentences rather than to embed them within the matrix clause. This suggests that some processing constraints operate on young children's productive capacities.

3. Anaphoric Reference

Children's knowledge of grammar continues to develop beyond the preschool years. One area that has received a good deal of attention from researchers is their knowledge of coreference relations within sentences, especially how anaphoric pronouns and reflexives link with referents. This research has been conducted primarily within a government-binding theoretical framework, investigating children's knowledge of the main binding principles. Spontaneous productions of pronominal forms suggest that quite young children use them correctly in their productive speech, however, the limits of their knowledge cannot be accurately assessed in naturalistic contexts.

Generally, children appear to develop knowledge of the main principles in the following order. By age 6 children know principle A, which states that reflexives are bound to referents within the same clause (e.g., "*John watched Bill wash himself*"; "*himself*" must refer to Bill, not John). Sometime later, knowledge of principle B emerges, which states that anaphoric pronouns cannot be bound to referents within the same clause (e.g, "*John asked Bill to hit him*"; "*him*" him must refer to John in the "*ask*" clause, not Bill in the "*hit*" clause). The last principle to emerge sometime during middle childhood is principle C, which states that backward coreference is only allowed if the pronoun is in a subordinate clause to the main referent (e.g., "*when he came home, John made dinner*"). Some researchers have argued that the grammatical knowledge of these principles is acquired much earlier that the research would suggest but that children's performance on tasks that tap this knowledge is limited by processing factors, pragmatic knowledge, or lexical knowledge. This debate continues in the developmental psycholinguistic literature.

G. Theoretical Explanations

1. Semantic Bootstrapping

Current theories in language acquisition attempt to address the central question of how the young child acquires the abstract and formal syntactic system of his or her language so rapidly, without formal instruction and with no feedback about whether he or she is using correct or incorrect forms. In the past two decades, one idea that has gained prominence in the literature is that children may use semantics or meaning to help break into the grammar of their language. Steven Pinker has been the main proponent to argue that children may use semantics as a bootstrap into syntax, particularly to acquire the major syntactic categories on which grammatical rules operate. Thus, children can use the correspondence that exists between names and things to map onto the syntactic category of noun, and they can use physical attributes or changes of state to map onto the category of verb. At the initial stages of development all sentence subjects tend to be semantic agents, and so children use this syntactic–semantic correspondence to begin figuring out the abstract relations for more complex sentences that require the category of subject.

2. Functionalism

A very different theoretical approach has been taken by those who view the central task of the child as gaining communicative competence. Much of the research conducted within this framework has focused on the acquisition of pragmatic aspects of language, including the functions of utterances and their use in discourse and other communicative contexts. Within research on grammatical development the functional approach does not take formal syntactic theory as its primary model. Instead, the structure of language is viewed from a functional or processing perspective. One example is the competition model of language acquisition proposed by Elizabeth Bates and Brian MacWhinney. In this model, the child begins by establishing the basic functional categories: topic–comment and agent. Different surface representations of these functional categories then compete for expression and initially the child may use a simple one form–one function mapping. Eventually, children move toward the adult system of form–function mappings.

3. Distributional Learning

At some point, all theories of acquisition need to consider how the child learns the major syntactic categories, even if they begin with a simple lexically specific, functional, or semantic approach. Thus, from the stream of words that children hear around them they must figure out which ones belong to the different major word classes, such as nouns, verbs, or adjectives. One important approach to this learning problem is the distributional learning view, according to which children not only use semantic mappings to acquire a category such as verb but also use distributional factors, such as it takes an *-ed* ending to express pastness or an *-ing* or *-s* ending in present tense

contexts, it occurs with auxiliary verbs, and so forth. In this view, children come to know that a particular morpheme is a verb because they hear that morpheme cooccurring with inflections that mark tense, for example. This kind of approach argues that children are sensitive to all kinds of distributional patterns in the linguistic input to which they are exposed.

4. Parameter-Setting Theory

Linguists working within a government-binding framework who have taken an interest in the question of how children acquire the grammar of their language claim that the central task of acquisition is to set the parameters of universal grammar in the direction appropriate for the language that is being acquired. Some argue that the parameters are initially set in one position, which may then have to be switched. An alternative view holds that parameters start off neutrally—that is, they are not set in any position. As children are exposed to their native language, they use linguistic evidence present in the environment to set the parameters accordingly.

V. PRAGMATIC DEVELOPMENT

In recent years, there has been increased interest in investigating how children acquire the ability to use language to fulfill a range of functions and in a variety of communicative contexts. This emphasis reflects the notion that to become a competent speaker requires not only knowledge of the structural forms and meanings of a language but also the ability to communicate using those forms in a competent, flexible, and appropriate manner. Some researchers have argued that language forms develop to serve new communicative functions, not *vice versa*. This aspect of language development is closely tied to the child's developing theory of mind and related social knowledge.

A. Communicative Functions

1. Speech Act Theory

What are the communicative functions expressed in children's speech? This approach to the child's language is based on the theory of speech acts proposed by Austin and Searle, among others, and has been very influential in child language research. These philosophers argue that many utterances do not simply make an assertion, but they also operate as *performatives*—that is, they perform an act (e.g., promise or refuse). Each utterance has three components: the *illocutionary intent*, or goal of the speaker; the *locutionary act*, or the actual form of the utterance; and the *perlocutionary effect*, the influence on the listener. In this way, we can account for the many different utterances that can be used, both direct and indirect, literal or metaphorical, to convey the same message. Thus, one of the questions that derives from this approach to language is how children come to use and interpret indirect and nonliteral uses of language.

2. The Development of Speech Acts

A number of researchers have focused on identifying and classifying the functions of early language, investigating the development of illocutionary intent and its relation to locutionary acts, and several systems have been developed. Children less than age 2 use language to fulfill a number of different functions, including getting people to do things, regulating their own or other's behavior, for social interaction, and as part of their imaginative play. By the time children are 3 years old, new functions emerge, including the use of language to describe objects or events and to assert an opinion, and also a range of conversational devices. At this point, children are also able to express each of these functions using a variety of different syntactic forms.

There is a more protracted period of development for indirect forms, such as indirect requests. Although 2-year-olds use terms such as "*want*" or "*need*" as a way of asking for something (e.g., "*I need new ball*"), genuine indirect requests do not emerge until approximately age 3 (e.g., "*Where is the truck?*"). By age 4, children can use polite forms, including modal verbs to make their request (e.g., "*Would you give me a cookie?*"), but hints or oblique indirect requests are not used until the early school years. Children aged 2 do not discriminate between requests for action and requests for information; however, throughout the preschool and early school years children gradually become able to understand increasingly more oblique levels of indirect speech acts.

B. Social Use of Language

A number of studies have focused on children's developing awareness of the perlocutionary effect of

their utterances and the ability to modify locutionary acts to take into account their listeners' knowledge. Although 5-year-olds know when to use definite and indefinite articles ("*the*" and "*a*"/"*an*"), depending on the listener's presuppositional knowledge about an object, children younger than age 7 or 8 do not perform well on referential communication tasks, in which they are required to describe a scene or unusual object that is hidden from the listener's view. Using more naturalistic data, such as spontaneous speech, other studies have found that even 4-year-olds change the way they speak. For example, they use simpler language if they are talking to 2-year-olds, which shows some awareness of the distinct needs of a very young conversational partner.

C. Conversational Abilities

Communicative competence also entails knowing how to engage in conversations in appropriate and informative ways. A number of studies have focused on the development of conversational abilities, especially the ability to take turns and maintain the topic of conversation. From the earliest stages children are able to take turns in conversation, following their mothers' utterance with their own, usually, although not always, in a semantically related way. This ability to maintain topic increases during the preschool years and the child is now able to respond to his or her mother by expanding on the information in her utterances. This ability to add new information correlates highly with the child's developing linguistic skills, as measured by MLU, and leads to the ability to maintain a topic over longer chains of conversational turns.

The use of language in various contexts provides the interactive, communicative framework within which children acquire knowledge of the linguistic structures available in their native language so that they can express more fully the ideas that are generated by their developing cognitive and social systems.

See Also the Following Articles

ARTIFICIAL INTELLIGENCE • BILINGUALISM • CATEGORIZATION • CREATIVITY • EVOLUTION OF THE BRAIN • INFORMATION PROCESSING • INTELLIGENCE • LANGUAGE AND LEXICAL PROCESSING • LANGUAGE DISORDERS • LANGUAGE, NEURAL BASIS OF • READING DISORDERS, DEVELOPMENTAL • NEUROPSYCHOLOGICAL ASSESSMENT, PEDIATRIC • SEMANTIC MEMORY • SPEECH

Suggested Reading

Bloom, P. (2000). *How Children Learn the Meanings of Words*. MIT Press, Cambridge, MA.

Fletcher, P., and MacWhinney, B. (Eds.) (1995). *The Handbook of Child Language*. Blackwell, Oxford.

Jusczyk, P. (1997). *The Discovery of Spoken Language*. MIT Press, Cambridge, MA.

Locke, J. (1993). *The Child's Path to Spoken Language*. Harvard University Press, Cambridge, MA.

McDaniel, D., McKee, C., and Cairns, H. (Eds.) (1996). *Methods for Assessing Children's Syntax*. MIT Press, Cambridge, MA.

Pinker, S. (1994). *The Language Instinct: How the Mind Creates Language*. Morrow, New York.

Pinker, S. (1999). *Words and Rules: The Ingredients of Language*. Basic Books, New York.

Tomasello, M. (1998). *The New Psychology of Language: Cognitive and Functional Approaches*. Erlbaum, Mahwah, NJ.

Language and Lexical Processing

RANDI C. MARTIN, MARY R. NEWSOME, and HOANG VU
Rice University

GLOSSARY

aphasia Disorder of language production and/or comprehension following brain damage.

Broca's area Area in the frontal lobe traditionally thought to contain the motor representations for words.

discourse processing Deriving meaning from conversation and text.

dissociation Preserved performance in one cognitive domain and impaired performance in another. A double dissociation occurs when one patient shows a disruption of one domain and preservation of the other, whereas another patient shows the reverse.

functional level First level of linguistic encoding in sentence production in which nonverbal conceptual information is used to select words and represent their relations.

immediacy of processing theory of comprehension Syntactic analysis and semantic interpretation are carried out on a word-by-word basis as each word is perceived.

lexicon The mental dictionary in which sound, spelling, meaning, and grammatical aspects of words are represented.

modularity hypothesis Hypothesis that different cognitive subsystems (i.e., modules) function independently.

orthography The written representations and rules for spelling of words.

parser A processing module that assigns syntactic structure to a sentence.

phoneme The basic unit of speech sounds.

phonological system The sound system of language.

positional-level Second level of linguistic encoding in sentence production in which the syntactic structure of the sentence is created and phonological representations are selected.

proposition Representation of part of a text that specifies the relations between actions and entities (people or objects) and between entities and their states or properties.

semantic system System that represents word and sentence meaning.

syntactic system System that represents grammatical information in the language.

thematic roles The roles that entities play with respect to a verb, such as agent, theme, or recipient.

Wernicke's area Area in the posterior temporal lobe traditionally thought to contain sound representations for words.

Within the realm of understanding and producing language, the brain is responsible for a wide range of tasks, from recognizing written words to understanding narrative. There is a fairly long history of behavioral studies of normal and brain-damaged populations that has delineated the organization of the component processes involved in various language tasks. Recently, lesion data have been supplemented by electrophysiological and neuroimaging data to reveal the regions of the brain that support language functions. A comprehensive view of language processing must take into account the theoretical developments from both behavioral and neuroimaging approaches.

The task of language comprehension is to transform written words or speech input into a conceptual representation, whereas that of language production is to translate a conceptual message into its spoken or written forms. Language processing can be examined

at the word level, sentence level, or at higher discourse levels (e.g., text and narrative). Various models of language processing have been proposed to represent these levels of analysis. These models vary in scope and detail, but most assume that processing at a given level is carried out by a set of systems. For example, spoken sentence comprehension involves lexical, syntactic, and semantic systems. The lexical system includes the mental lexicon (or mental dictionary), which contains all the information about individual words. In the lexical subsystem, acoustic input is matched to a sound representation of a word (termed a lexical phonological representation). A match to a phonological representation then allows access to other aspects of a word's representation, including its meaning and its grammatical features (such as noun vs verb). A syntactic system uses the grammatical information from each word to construct a representation of the relationships among words within the sentence (i.e., syntactic relationships). The semantic system uses individual word meaning, syntactic information (e.g., which noun is the subject of the verb), and general world knowledge to construct an overall interpretation of sentence meaning.

Some researchers have hypothesized that the separate systems within a domain are modular—that is, they function independently. According to a strong version of the modularity hypothesis, once a module (i.e., system) receives input from another module (e.g., the syntactic system receives grammatical information about a word from the lexical system), the module carries out its computations (e.g., determines how the word fits into the syntactic structure) without further influence from any other modules (termed discrete processing). The output from the module is then passed on to the next module. According to a weaker version of the modularity hypothesis, each module may be represented independently in the brain; however, while a module is carrying out its computations, it may receive input from other modules that influences its processing (termed interactive processing). For example, compare the sentences "Mary saw the bird with binoculars" and "Mary saw the bird with red feathers." Both have the same syntactic structure but semantic information most likely causes listeners to assume that the prepositional phrase "with binoculars" modifies "saw" in the first sentence, whereas "with red feathers" modifies "the bird" in the second. The strong version of modularity would assume that this semantic information only comes into play after a syntactic decision has been made, perhaps leading to a reinterpretation of syntactic structure. The weaker version of modularity would assume that semantic information influences the initial decisions that are made regarding syntactic structure (i.e., that the semantic and syntactic systems interact during early stages of processing). A completely nonmodular theory of language processing would assume that there is no distinction between syntactic and semantic representations—that is, both are represented in the same system, obey the same principles, and have been learned according to the same principles. As discussed later, the evidence from brain-damaged patients supports some form of modularity.

Many current models of language processing are instantiated in a network framework—that is, a network consisting of nodes that represent different types of units (e.g., letters, phonological representations, and concepts) and connections among these nodes. These network models are sometimes referred to as neural networks because the nodes and connections are argued to be analogous to neurons and their synaptic connections. The nodes of a certain type are all represented at the same level and the connections may be between levels or between units at the same level. For example, in written word recognition, there are layers of nodes representing visual features (e.g., straight line and curve), letters, and written word forms. In this type of model, processing is carried out via activation that spreads throughout the network. Stimulus presentation activates the first layer of nodes, and then activation spreads to subsequent layers. For example, if the word "bear" were presented, the visual features corresponding to each letter would be activated; activation would then spread along the connections between features and letters so that the appropriate letters would be activated. Activation of the letters would then lead to a spread of activation along connections between letters and words so that the appropriate word would be activated. In this type of model, partially consistent but incorrect letters and words would be activated, at least temporarily. For example, since "pear" shares features and letters with "bear," this word would be activated to some extent from presentation of "bear." In purely "feedforward" network models, activation flows in only one direction (in this case, from visual features to words). In "feedback" models, activation flows in both directions. In a feedback version of this model, activation from letters would activate corresponding words; however, in addition, activation at the word level would flow back to the letter level, boosting the activation of letters consistent with activated words.

I. LEXICAL PROCESSING

A. Word Recognition

In visual word recognition, a whole word may be viewed at once (provided that it is short enough), and recognition is achieved when the characteristics of the stimulus match the orthography (i.e., spelling) of an entry in the mental lexicon. Speech perception, in contrast, is a process that unfolds over time as the listener perceives subsequent portions of the word. Upon hearing the first syllable of a spoken word such as the "un" in "understand," several words may be consistent with the input (e.g., "under," "until," and "untie"). As subsequent portions are perceived the pool (or "cohort") of words will be narrowed down, until only one word remains.

Despite these differences in the temporal course of processing, there are many commonalities in spoken and written word recognition. In both cases, the goal is to go from the perceptual information to the lexical form in order to access semantic and syntactic information about the word. In visual word recognition, a letter level intervenes between visual processing and lexical access. In auditory word perception, it is often assumed that a phoneme level intervenes between the acoustic input and lexical access. Phonemes are assumed to be the basic sound units of speech perception (and production). In English there are approximately 40 different phonemes, corresponding to the consonant and vowel sounds. The phonemes of other languages overlap those of English to a large degree, although some languages may lack some of the phonemes in English or may contain phonemes that do not exist in English. For example, Chinese does not distinguish between the "l" and "r" phonemes, and some African languages include clicking sounds as phonemes.

There is general agreement that spoken and written word recognition involve access to the same semantic and syntactic representations. There has been some disagreement, though, about whether there are separate lexical representations for spoken and written words. Some researchers have argued that written words have to be transformed into a sound representation in order to access semantic and syntactic information about the word. If so, then only a phonological representation (e.g., one that indicates the sequence of constituent phonemes and the stress pattern) is needed for each word. However, considerable neuropsychological evidence suggests that there are separate phonological and orthographic representations for words, and that access to word meaning can proceed for written words without conversion to a phonological form. Nonetheless, it is the case that for normal individuals the phonological representation of a written word appears to be computed automatically (through an implicit "sounding out" or "letter–sound" conversion process) when a written word is perceived. This derived phonological information can influence the time course of lexical access, making word recognition slower for words that have an unusual letter–sound correspondence, particularly if these words appear infrequently in print (e.g., "yacht"). Despite this slowing, the correct word is typically accessed, indicating that readers cannot be relying solely on letter–sound correspondences in accessing the meaning of written words.

B. Word Production

Spoken and written word production involve the reverse of the processing steps in word perception: conceptual processing to lexical to phonological or orthographic. Motor execution stages involved in articulation and writing would complete the output process.

More research has been devoted to speech production than to writing. In the domain of speech production, a distinction has been made between a nonverbal conceptual representation of the message to be expressed and a semantic representation that is specific to words. Speech production begins with the formulation of the nonverbal conceptual representation, followed by two steps of lexical access. In the first step, a lexical–semantic representation is selected (which also contains syntactic information about the word), and in the second step the phonological form corresponding to the semantic representation is accessed. In a strict modular approach with nonoverlapping stages, only one lexical–semantic representation is selected before processing proceeds to the phonological level. Other approaches assume what is termed "cascaded processing," in which activation spreads from the lexical–semantic level to the phonological level before a single lexical–semantic representation is chosen. Suppose that the speaker wishes to communicate the concept "cat." The conceptual representation for cat would serve to activate most highly the lexical–semantic representation for "cat" but would also activate related lexical–semantic representations to some extent, such as "dog" and "lion," because of

shared conceptual features. In the cascaded model, the phonemes in "dog" and "lion" would also be activated to some extent, even though "cat" might eventually be the most activated at the lexical–semantic level. In a cascaded model with feedback, the activation of the phonemes in "cat" would cause backwards activation of words that shared phonemes with "cat," such as "mat," even though such words had no semantic relationship to "cat."

In writing, as in reading, some have argued that the orthographic forms are dependent on the phonological forms. That is, the writer is assumed to have followed similar steps in written word production as in spoken word production and has accessed a phonological form. This phonological form is then translated into a written form through a phoneme–grapheme conversion process. However, it is even clearer in writing than in reading that orthographic knowledge specific to individual words (i.e., a lexical orthographic representation) is needed for correct spelling. Even for a word with a regular correspondence between sounds and letters, there may be several alternative "regular" spellings (i.e., spellings that follow typical sound-to-letter conversion patterns in English). For example, "kat" or "cat" would be regular spellings for "cat," and "leaf," "leaph," "leef," and "leeph" would be regular spellings for "leaf." Thus, producing the correct spelling depends on having stored knowledge about the sequences of letters in words.

C. Neuropsychological Evidence on Lexical Processing

Evidence for independent modules in lexical processing can be obtained from individuals with brain damage due to a stroke or other injury who can competently produce or understand some types of linguistic information but not others. On the other hand, evidence that the same module is involved in performing two different tasks can be obtained by showing strong correlations between the factors affecting performance on each. In the domain of word recognition, double dissociations between written word and spoken word comprehension have been reported. That is, some patients who show a deficit in recognizing printed words can nonetheless recognize spoken words, whereas other patients show the reverse. For some of these patients, the deficit in spoken or written word perception cannot be attributed to difficulties with basic aspects of visual or auditory perception because the patient can recognize nonverbal materials in both modalities. Instead, the deficit is specifically in the phonological or orthographic processing systems. The existence of patients who can understand written words but who cannot understand spoken words because of disrupted phonological representations argues against the necessity of converting written words to phonological forms in order to access meaning. Evidence that the same lexical–semantic system is involved in comprehending spoken and written words comes from patients who show comprehension difficulties for certain words or certain semantic categories (such as animals or tools), and the same words are affected irrespective of whether the input is spoken or written. If there were separate semantic representations for spoken and written words, it would be highly unlikely that the same categories of words would be affected for both modalities.

Category-specific deficits have been noted for many patients and raise interesting questions concerning the nature of semantic representations in the brain. Although more specific deficits have been observed, these deficits tend to occur in the categories of animals, plants, and artifacts (i.e., man-made objects), with the most common deficit for animals or, more generally, for living things. One possible explanation of category-specific deficits is motivated by the fact that semantic properties of an object tend to be interrelated (e.g., having eyes usually occurs with having a nose) and that objects in the same superordinate category (e.g., tools or fish) share properties. If constellations of shared properties are organized together in the brain, then when damage occurs to a region in which semantic properties are stored, deficits that affect certain categories will result. The only difficulty with this account is that it does not explain why deficits tend to occur in three categories—that is, any constellation of shared properties should be subject to damage and we should observe patients with highly specific deficits such as a deficit for vehicles but not other artifacts.

Another possible explanation for category-specific deficits is that there are two separate semantic systems in the brain—one that represents sensory knowledge and another that represents functional knowledge (i.e., the functions that objects perform). Researchers have argued that knowledge of animals is mainly sensory, whereas knowledge of artifacts is mainly functional. Consequently, damage to the sensory knowledge system results in a deficit specific to animals, whereas damage to the functional system results in a deficit

specific to artifacts. However, findings from some brain-damaged patients are problematic for this sensory/functional explanation. For example, some patients have been reported who have semantic deficits for only some subsets of living things (e.g., fruits and vegetables), and others have been reported who have a disruption of knowledge of both sensory and functional attributes of animals but a preservation of both sensory and functional knowledge for artifacts.

As discussed earlier, models of word production take either a discrete stage or interactive activation approach. Some data from speech errors in normal subjects (either spontaneous or experimentally elicited) support the interactive view. For example, sound exchange errors are more likely to occur if the exchange results in two words (saying "barn door" for "darn bore") than if the exchange results in nonsense words (saying "beal dack" for "deal back"). This effect of lexical status of the resulting error can be attributed to feedback from the phoneme level to the lexical level. The word production errors of aphasic patients can also be better accounted for by an interactive approach. Such an approach provides a means of accounting for some patients' tendency to produce words phonologically related to a target word (so-called "formal errors," such as saying "mat" for "cat") and for some patients' tendency to produce a large proportion of errors that are both semantically and phonologically related to a target (saying "rat" for "cat").

Double dissociations between deficits in written and spoken word production have been observed. Again, in many of these cases, basic deficits in the motor processes involved in speaking or writing can be ruled out, indicating that the modality-specific deficit is specific to phonological or orthographic output processing. Further evidence for the separation of phonology and orthography comes from patients who make semantic errors in only one output modality—for example, producing "pillow" as the name for a picture of a bed when speaking but producing "bed" correctly in writing. Such a pattern would not be expected if it were necessary to use the phonological form to guide spelling.

Evidence from speech production deficits provides information about the representation of grammatical information. Deficits specific to certain grammatical categories have been reported because some patients have selective difficulties in the production of function words (i.e., words such as prepositions, pronouns, and auxiliary verbs that play primarily a grammatical role in a sentence). Such difficulties are remarkable given that these grammatical words are often quite short and easy to pronounce (e.g., "to" and "will") and are the most frequently occurring words in the language. Some patients have demonstrated greater difficulty in producing nouns than verbs, and others have demonstrated the reverse. As with the semantic category deficits, there is no consensus on the explanation for these grammatical class deficits. In some cases, these apparent grammatical class effects have a semantic basis. For example, better production of nouns than verbs and better production of verbs than grammatical words may be observed because the patient is better able to produce more concrete words. However, for some patients, it appears that grammatical class effects cannot be reduced to a semantic basis; consequently, these deficits suggest that at some level in the production system words are distinguished neurally with regard to the grammatical role that they play in a sentence. The separability of grammatical information from other types of lexical information is supported by other findings showing that some patients with picture-naming deficits can provide grammatical information about a word, such as its gender (in a language such as Italian or French), even though they are unable to retrieve any of the phonemes in the word.

Neuropsychological research with brain-damaged patients, more so than research with normal subjects, has addressed the issue of the relation between the phonological processing systems involved in speech perception and production and the relation between the orthographic systems involved in reading and writing. Some patients show an excellent ability to recognize and remember input phonological forms (e.g., being able to decide whether a spoken probe word rhymes with any of the words in a preceding list but have great difficulty in producing output phonological forms (e.g., naming a picture). Other patients show the reverse pattern of great difficulty in holding onto input phonological forms (performing at chance on the rhyme probe task) but showing normal speech production. Similar double dissociations have been documented for orthographic processing. Thus, input and output forms in both speech and writing appear to be represented in different brain areas. However, although the input and output forms may be different, they are linked to each other. For example, individuals can repeat nonwords, converting an input to an output form. A close coupling between input and output forms appears to be involved in the development of speech production and in the maintenance of accurate speech production throughout adulthood.

D. Localization of Word Comprehension and Production

As we have seen, patient studies can be used to posit the nature of the functional components of language comprehension and production and their connections. Studies of localization reveal where in the brain those components may lie. Models of brain areas and their functions can be traced back to the late 1800s, beginning with Lichtheim and Wernicke. Wernicke's model assumed that language was represented in the left hemisphere. Findings since that time have indicated that although the left hemisphere is dominant for language in most individuals, the right hemisphere also plays some role. In addition, the right hemisphere is dominant in some individuals. The model incorporated a concept center that received input from sensory word images and provided output to motor word images. The sensory word images were thought to be represented in Wernicke's area, found in the superior (upper), posterior (near the back), left temporal lobe, and the motor images in Broca's area, found in the left frontal lobe near the motor cortex. Damage to Wernicke's area was thought to result in an inability to recognize and understand spoken words, although speech was thought to be fluent. Broca's aphasics were thought to have an impairment in articulating speech (i.e., "essentially mute, except for the repetition of the same few utterances"), but their comprehension was intact. Wernicke's and Broca's areas, though separate, were thought to be connected through a neural fiber tract (i.e., the arcuate fasciculus). This fiber tract was thought to be involved in translating an auditory word representation into a motor word representation. Based on Wernicke's model, one would predict that patients who had intact Broca's and Wernicke's areas but damage to this fiber tract should have good language comprehension and production but have difficulty repeating words and sentences. Patients were reported who showed this pattern (termed conduction aphasics), which seemed to provide a strong confirmation of the model.

There are a number of observations that cause difficulties for the classical model. For example, Broca's aphasics are not equally impaired on all types of words, typically being better able to produce nouns than verbs. Although they have difficulty producing function words, function words occurring in the middle of sentences are omitted less frequently than those that occur in the beginning of sentences. An additional complication for the Wernicke/Lichtheim models is that although Wernicke's aphasics' speech is fluent it shows numerous aberrations, such as misordered phonemes, incorrect words, and nonsense words such as "tarripoi" in the statement spoken by one Wernicke's aphasic, "I can't mention the tarripoi." Moreover, current studies suggest that among individuals classified into any traditional syndrome category (Broca's, Wernicke's, and conduction aphasia), there are wide variations in the nature of their deficits. For example, within conduction aphasics, some patients produce fluent and appropriate speech output but make semantic substitutions or paraphrases in repeating sentences, suggesting that they have difficulty retaining phonological information on the input side. Other conduction aphasics make numerous phonemic errors in their speech output but do not make semantic substitutions in repeating, suggesting that they have difficulty constructing and maintaining phonological representations on the output side. Thus, localization of function on the basis of classical syndromes appears to be a misguided effort. However, careful examination of a series of individual cases who show similar behavioral deficits allows for the determination of the lesion site that is affected in all such individuals. On the basis of these data, one can look to behavior–lesion correspondences to draw conclusions about localization.

Recently, using methods that measure the physiological changes that occur in the brain, it has been possible to determine the brain areas activated in normal individuals while they are performing language tasks. Electrical activity occurring in the brain during various tasks [i.e., event-related potentials (ERPs)], can be measured. These ERPs show negative and positive electrical potentials that are time locked to the onset of the presentation of verbal stimuli. The brain distribution of these potentials can be plotted. The flow of blood into different regions of the brain can be detected through via positron emission tomography (PET) and functional magnetic resonance imaging (fMRI). The use of PET and fMRI for functional localization thus assumes that greater blood flow is observed in brain areas with greater neural activity. Although all these methodologies provide information about when and where activation occurs in the brain, ERPs are known for their good temporal resolution, whereas the strength of fMRI lies in its spatial resolution. The spatial resolution of PET is also substantially better than that from ERPs but worse than that of fMRI.

The traditional model of language processing predicts that processing of phonological information of heard words should occur in the posterior left

temporal lobe (Wernicke's area) and phonological activation for spoken words should occur in left frontal lobe (Broca's area). Although the brain areas implicated in the traditional model have been confirmed by imaging studies, additional areas have been found to be active. That is, although neuroimaging studies show consistent evidence of Wernicke's area activation during phonological encoding, temporoparietal and frontal activation have been found as well. Similarly, studies have found activation both in Broca's area and in the lower portion of the posterior temporal lobe (Wernicke's area is the upper portion of the posterior temporal lobe) during language production. One caveat regarding the role of Broca's area is that it has been activated when research participants are asked to make hand and tongue movements, suggesting that activation of this area reflects general motor programming rather than motor programming specific to phonological output.

If visually presented words are converted into their phonological representations, then activation should occur in areas similar to those implicated for auditory word processing. Studies have found activation during reading of visually presented words in left temporoparietal cortex, which is also active during auditory word processing. The left posterior middle temporal gyrus, in addition to the occipital–temporal and left inferior temporal/fusiform regions, has also been shown to be active during word reading.

A variety of areas have also been implicated for semantic processing. Two tasks meant to measure semantic processing, categorizing words and generating actions that can be performed with objects, have shown activation in left frontal regions of Broca's and surrounding areas. Other research has suggested that the frontal cortex is involved in retrieving and maintaining semantic information rather than being the locus of representation of semantic knowledge. Recent studies with normal and brain-damaged research participants have suggested the role of temporal areas in semantic processing, including left posterior temporoparietal, inferior temporal, and anterior temporal cortex. Researchers investigating the localization of categorical knowledge with neuroanatomical and neuroimaging studies found a wide variety of brain areas to be involved during tasks that test knowledge of different categories. Even within a category, various brain areas have been activated. For example, the left temporal lobe, the right temporal lobe, and even the frontal and inferior parietal areas have all been reported to be active during processing of animals and living things. Areas responsible for processing nonliving things also appear to be diverse. Despite the varied areas that have been reported to be involved in semantic function, one area that has shown activation across several different studies is the temporal lobe. Work with Alzheimer's patients who have semantic impairment suggests a further category distinction. Alzheimer's patients whose knowledge of the perceptual features of objects is disrupted tend to have temporal region damage, and those whose knowledge of how objects function tend to have frontoparietal impairment.

More consistent findings have been obtained regarding different localization for words of different grammatical classes. Neuropsychological and neuroimaging results suggest that nouns are represented in the temporal lobe and verbs in the frontal lobe. Some ERP studies have indicated that content words and function words elicit different timing of brain potentials and the involvement of different brain areas. For instance, one study showed early frontal activation from both word types with a similar pattern brain distribution at approximately 200–300 msec after the onset of the word. Approximately 150 msec later, however, differences for the word types were observed, with a different timing of the potential waveforms for the two word types and greater left hemisphere activation for the function words than for the content words. However, because this study measured processing of these word types in a sentence context, the different patterns may be due to the different roles that these words play in sentence processing rather than to different localization of lexical representations. For example, function words may lead to more predictions about the syntactic structure of upcoming words in a sentence.

Recent behavioral and neuroimaging data suggest that the right hemisphere may play more of a role in language processing than was previously assumed. Behavioral studies on this issue have used a priming paradigm in which word recognition (as measured by time to pronounce a word or to make a word–nonword decision) is facilitated by the prior presentation of a related word. To study hemispheric contributions, written words are presented to the left or right visual field, which engages processing in the contralateral cerebral hemisphere first. Priming studies have shown that priming occurs for only the most highly related words in the left hemisphere but for a broader range of related words in the right hemisphere. Also, for ambiguous words such as "bank" that have a dominant meaning (money) and a subordinate meaning (river), priming results indicate that immediately

following presentation of an ambiguous word both meanings are activated in both hemispheres. However, within a short time, only the dominant meaning is available in the left hemisphere, whereas both meanings are still available in the right hemisphere. Taken together, the evidence suggests that the meanings of words are more coarsely coded in the right hemisphere. Neuroimaging studies of word processing have also often found activation in the right hemisphere in areas corresponding to the traditional language areas on the left, although the activation is often less than that obtained in the left hemisphere.

II. SENTENCE PROCESSING

A. Sentence Comprehension

Sentence comprehension processes involve more than combining the meaning of individual words in a sensible fashion. Even though the sentence "the girl chased the dog" has the same content words as "the dog chased the girl" and "the girl was chased by the dog," it is clear that the first sentence means something different than the latter two. Clearly, the grammatical roles that nouns play (e.g., subject vs object) and sentence structure (e.g., active vs passive) play a role in determining sentence meaning. Most models of sentence comprehension assume that in order to understand a sentence, the constituents (e.g., noun phrase, verb phrase, and prepositional phrase) of a sentence must be identified and related to each other structurally. The "parser" is the term used to describe the module that assigns syntactic structure to a sentence. Parsing allows the determination of the grammatical roles, such as subject, direct object, and indirect object. These grammatical roles are assumed to be mapped onto "thematic roles"—that is, conceptual roles that entities play with respect to a verb such as the agent (i.e., the entity carrying out the action), theme (i.e., the entity being acted upon), recipient, or location. The mapping between grammatical roles and thematic roles varies for different verbs and may vary for the same verb depending on the sentence. For example, the thematic roles of the subjects of the phrases "the boy received," "the girl opened," and "the key opened" are "recipient," "agent," and "instrument," respectively. Because of this variation in mapping, the lexicon needs to contain verb-specific information concerning the possible mappings between grammatical and thematic roles.

It is often the case that the syntactic structure to be assigned to a sequence of words is ambiguous. For example, for a sentence beginning with "The defendant examined," "examined" could be the main verb or could be part of a reduced relative clause construction (e.g., "The defendant examined by the lawyer was innocent"). A major issue in research on sentence processing has been whether a strict modular approach applies in which the parser assigns syntactic structure purely on the basis of syntactic rules of thumb (e.g., "choose the simplest structure") or whether an interaction between semantic and syntactic information affects parsing decisions. For example, if semantic information influences parsing, then "examined" should be less likely to be taken as the main verb in a sentence beginning "The evidence examined" since "evidence" is unlikely an agent of "examine." Current evidence supports the interactive view.

During comprehension, it is necessary to integrate different constituents of a sentence, sometimes over distances spanning several words. For example, in the sentence "The truck that the car splashed was green," it is necessary to retrieve "truck" as the direct object of "splashed" and to integrate "green" with "truck" rather than "car." Consequently, researchers have assumed that comprehension makes demands on working memory resources. The concept of working memory is similar to that of short-term memory, involving the temporary maintenance of information. The term "working" is used to signify that this capacity is used in carrying out computations in addition to maintaining information. Individual differences in working memory capacity of normal individuals have been found to correlate with aspects of sentence comprehension performance. However, neuropsychological findings suggest that there are different components of working memory that play different roles in sentence comprehension.

B. Sentence Production

In sentence production, as in comprehension, the mapping between thematic and grammatical roles and the construction of syntactic structure must be included in the processing stages. The processing stages in sentence production include the message level stage, in which a conceptual representation of the sentence is formed, two stages of linguistic encoding, and an articulatory stage. Thematic role relations between actions and entities are assumed to be represented at the message level. At the first level of linguistic encoding (termed the functional level), the message level information is used to select lexical representations

(i.e., semantic–syntactic lexical representations) and to construct grammatical relations among these lexical representations based on the thematic role and other semantic information in the message. The prosodic structure of the sentence is also encoded during this stage. Prosodic factors such as word stress (the emphasis given to words) and intonation (using pitch to signify different meanings) are important methods of varying speech to facilitate communication. The second level of linguistic encoding (called the positional level) creates the syntactic structure for the sentence in terms of word order, function words, and grammatical markers (e.g., plural markers and past tense inflections). The phonological forms for the content words are also retrieved at this stage. Once the phonological representations have been retrieved, plans for articulation can be formed.

C. Neuropsychological Evidence on Sentence Processing

1. Sentence Comprehension

In the 1970s and 1980s, researchers demonstrated that some aphasic patients who understood the content words in a sentence might nonetheless show a failure of comprehension for the sentence as a whole if comprehension depended on the correct analysis of syntactic information. For example, these patients might have difficulty understanding ''reversible'' sentences—that is, sentences in which either noun could play the thematic roles of agent or theme, as in ''The dog was chased by the girl'' or ''The truck that the car splashed was green.'' This comprehension difficulty could be demonstrated by asking the patient to choose between a picture depicting the correct thematic role relations and one depicting the reverse role relations (e.g., a girl chasing a dog vs a dog chasing a girl). Although the patients might perform at chance with such picture contrasts, they would do well if one of the pictures substituted an incorrect noun or verb (e.g., a girl chasing a cat or a girl walking a dog). One early hypothesis about the nature of such a comprehension deficit was that it derived from a general failure of syntactic processing because patients who showed this comprehension problem also produced ''agrammatic speech.'' That is, their speech production was marked by simplified syntactic structure and the omission of function words and grammatical markers. However, recent findings have demonstrated that some patients may show only one side of this deficit (i.e., impaired syntactic comprehension but not agrammatic speech, or the reverse). In addition, several studies have shown that patients who show this comprehension problem on sentence–picture matching may do well on judging the grammatical acceptability of sentences. Thus, rather than a global deficit in all aspects of syntactic processing, these patients may have a more restrictive comprehension deficit, such as a deficit in mapping between the grammatical structure of the sentence and the thematic roles that entities play in the sentence. For instance, for the sentence ''The truck that the car splashed was green,'' the patient might be able to determine that ''car'' is the grammatical subject of the verb ''splashed'' but be unable to determine that the car is doing the splashing.

Other findings indicate that although agrammatic speakers may not provide the clearest evidence of a dissociation between syntactic and semantic knowledge, other patients do provide such evidence. Some patients with Alzheimer's dementia demonstrate very impaired knowledge of word meanings but show preserved grammatical knowledge. For example, they might be unable to realize that a phrase such as ''The jeeps walked'' is nonsensical but be able to detect the grammatical error in a phrase such as ''The jeeps goes.'' Other case studies of aphasic patients show the reverse pattern of preserved semantic knowledge but disrupted grammatical knowledge. Thus, the findings indicate some degree of modularity in the sentence comprehension system, with separate modules for semantic and syntactic processing. However, other evidence from patients comports with the findings from normal subjects in showing that these modules do not typically operate in isolation but instead interact. That is, patients may use the grammatical structure of sentences when there are weak semantic constraints (e.g., understanding that ''tiger'' is the agent of ''chased'' in ''The lion that the tiger chased'') but fail to use the grammatical structure when there are strong semantic constraints (e.g., mistakenly interpreting the ''woman'' as the agent of ''spanked'' in ''The woman that the child spanked''). These findings support the view that information from both semantic and syntactic sources typically combines during comprehension, but when one of these systems is weakened due to brain damage, the other system may override its influence.

As discussed previously, theories of comprehension often assume a role for a short-term or working memory system that is used to hold partial results of comprehension processes while the rest of a sentence is processed and integrated with earlier parts. Aphasic patients often have very restricted short-term memory

spans, being able to recall only one or two words from a list compared to normal subjects' ability to recall five or six words. Many of these patients appear to have a deficit specifically in the ability to retain phonological information. Although it may seem intuitively plausible that restricted short-term memory capacity would impede comprehension, a number of studies have shown that patients with very restricted memory spans may show excellent sentence comprehension even for sentences with complex syntactic structures. Such findings support immediacy of processing theories of comprehension that state that syntactic analysis and semantic interpretation are carried out on a word-by-word basis, to the extent possible. Recently, some patients have been identified whose short-term deficit appears to be due to a difficulty in retaining semantic information rather than phonological. For such patients, their restricted ability to retain semantic information does impede comprehension for certain sentence types—that is, those that put a strain on the capacity to retain individual word meanings. Specifically, these patients have difficulty comprehending sentences in which the structure of the sentences delays the integration of word meanings into larger semantic units. One sentence type causing difficulty includes sentences with several prenominal adjectives, such as "The drab old red swimsuit was taken to the beach." In this example, the meaning of "drab" cannot be integrated with the noun it modifies until two intervening words have been processed. These patients do not have difficulty comprehending similar sentences in which word meanings can be integrated immediately. For example, these patients can understand sentences in which several adjectives follow the noun (e.g., "The swimsuit was old, red, and drab, but she took it along anyway"), because these sentences allow for the immediate integration of each adjective with the preceding noun.

2. Sentence Production

Sentence production deficits have also been a focus of research in aphasia, although in this domain much of the work has originated from a syndrome-based approach, concentrating on patients showing agrammatic speech. Agrammatism occurs predominantly in patients with articulatory disturbances who are typically classed as Broca's aphasics. However, several studies have documented that some features of agrammatism may appear in patients who are fluent speakers. Moreover, other studies have shown that the sentence structure and function word difficulties of agrammatic patients may dissociate, with some patients demonstrating reduced sentence complexity but accurate production of function words and inflections and others showing the reverse. In order to accommodate the dissociation between sentence structure and function word difficulties, deficits at different levels in the production process have been postulated. Several different suggestions have been made as to what these different deficits might be. An interesting recent approach relates deficits in sentence structure to deficits in the knowledge of verb representation. The verb plays a major role in structuring the roles of nouns (such as agents, patients, and recipients) with respect to the action in the sentence. The specific verb to be used dictates what grammatical role a noun with a specific thematic role will play (e.g., the recipient will be the subject of an active sentence using the verb "receive" but the indirect object of an active sentence using "give"). A deficit in knowledge of the relations of semantic and grammatical roles entailed by verbs could lead to a reduction in sentence structure, such as the failure to produce a required indirect object.

The disruption in the production of function words and inflections might be a result of a disruption at a different stage, specifically the stage at which the syntactic structure of the utterance is specified in terms of word order and grammatical markers (i.e., the positional level). Of course, many patients might have a disruption both in representing the relations of the verbs to the nouns and in constructing a syntactic specification, resulting in prototypical agrammatic speech.

D. Localization of Sentence Processing Mechanisms

Because of the association of Broca's aphasia with agrammatism and difficulties comprehending reversible sentences, some researchers have tried to link the frontal lobe, specifically Broca's area, to syntactic parsing. However, as discussed earlier, patients who are not Broca's aphasics have been shown to have difficulties understanding reversible sentences and to make grammatical errors in language production. With regard to comprehension, a study of lesion overlap among a group of aphasic patients with syntactic comprehension difficulties revealed that the critical area appeared to be in the anterior temporal lobe rather than Broca's area. ERP studies have revealed different patterns of brain potentials in response to semantic and syntactic errors in a sentence,

supporting the independence of these processes. The semantic anomalies elicit negative-going waves approximately 400 msec after the onset of the anomalous word, whereas the syntactic anomalies elicit positive-going waves approximately 600 msec after the onset of the ungrammatical word. Both the negative semantic wave and the positive syntactic wave have central or posterior brain distributions, with a somewhat more posterior distribution for the positive syntactic wave.

There have so far been relatively few neuroimaging studies of sentence comprehension. All of the traditional left hemisphere language areas (i.e., areas surrounding the Sylvian fissure, including Broca's area and Wernicke's area) are activated; however, activation in the anterior temporal lobe is also observed in some studies. Given the many processing components involved in sentence comprehension (e.g., syntactic parsing, thematic role determination, semantic integration, and maintenance in working memory), it is unclear exactly what function is carried out in which part of this broad network of activation. Some studies have suggested a specific role for Broca's area in the comprehension of syntactically complex sentences; however, not all types of complex sentences induce activation in this area. It is possible that, instead, this area carries out a working memory function related to maintaining words that have not yet been assigned a thematic role. Only certain types of syntactically complex sentences would place a demand on such a working memory function.

With regard to sentence production, there have been few studies using lesion localization that have attempted to isolate different aspects of production. One well-established finding is that there are frontal areas involved in speech articulation. However, for more central linguistic encoding aspects of language production, the data are scarce. Some patients with frontal lesions who show difficulty comprehending sentences in which integration of meanings is delayed (i.e., sentences with several adjectives preceding a noun) also have difficulty producing similar sentence constructions. These findings suggest that the same frontal working memory area is involved in maintaining semantic representations in comprehension and production. Clearly, neuroimaging data would be useful in helping to localize aspects of language production. However, there are methodological difficulties in obtaining such data. If subjects move during PET and fMRI studies, this movement can induce the appearance of activation in brain areas that are in fact not activated. Thus, having subjects overtly produce speech could introduce movement artifacts in such studies. Having subjects speak silently to themselves would prevent the experimenter from monitoring what the subject was saying. Future research may find means to obviate these difficulties.

III. HIGHER LEVEL LANGUAGE PROCESSING

A. Discourse Comprehension

The comprehension of continuous discourse involves more than word recognition, syntactic parsing, and deriving the meaning of individual sentences. Successful comprehension requires an understanding of the relationships among the various parts of the discourse context and is dependent on the reader's or listener's general world knowledge. Many researchers have argued that story comprehension implies the derivation of propositional representations. Propositions derived from a story specify relations between actions and participants in the actions, the states and attributes of the participants, the time and locations of the actions, etc. Propositions are specified in terms of entities (typically nouns) and predicates (such as actions or states) that apply to these entities. For example, consider a story titled "The Picnic," beginning with the following: "The sky was cloudy and the weather forecast was not encouraging. There was a 70% probability of showers. However, the Brown family had no intention of renouncing their plan." Propositions derived from these sentences might include time (past), sky (cloudy), forecast [not (encouraging)], and renounce (agent, family; object; plan). It is clear, however, that comprehension of the story depends not only on deriving propositions but also on relating propositions to each other and to world knowledge. Even for the first sentence in this example, not only are "sky" and "cloudy" linked, such that "cloudy" specifies the state of the sky, but also it is likely that this information would to be related to the title and to long-term knowledge that people prefer sunny skies for picnics. Moreover, this information is reinforced by propositions derived from the following sentences. Overlap among the propositions and their associations in long-term memory makes the text more easily comprehended and remembered.

Propositions can be related to each other by various means. A person or object mentioned earlier can be referred to again using an anaphor, such as a pronoun. For example, "she" is an anaphor for "The little girl" in the passage "The little girl wanted to play ball. She got the bat out of the car." Propositions can also be

related via inferences. An inference is the drawing of a conclusion that is not explicitly mentioned in a discourse but, rather, is based on general world knowledge. In the picnic example discussed previously, someone hearing this story is likely to infer that it was the Browns who were planning the picnic mentioned in the title and that, contrary to usual expectations, they were not going to cancel because of the weather. In some cases, discourse or narrative texts (such as a mystery novel or an adventure story) may follow a familiar structure (termed a schema or script), and the presence of this structure may help the comprehender to organize the information.

B. Discourse Production

Narrative production requires that the speaker (or writer) produce a sequence of sentences that is coherent and comprehensible to the conversational partner (or the reader). The speaker has to be able to keep in mind what has already been stated and what remains to be stated and have some plan regarding the order in which information should be presented. Thus, in order to study narrative, measures beyond those of the individual phrase or sentence must be used in order to determine whether speakers are producing coherent narratives. Measures such as the overlap in propositions between utterances and the use of appropriate anaphors have been employed. Researchers may elicit target narratives by having the subjects view pictures or films and then asking the subject to tell the story of what happened. In such cases, the subjects' production can be assessed regarding whether the major elements of the story are present (such as introduction of people and situation, discussion of sequence of events and complication of the events, and resolution of the complication). Other measures can be used that assess whether the speaker uses vocabulary appropriate to the target audience and provides enough information, given the listener's knowledge, to allow for appropriate inferences.

C. Neuropsychological and Localization Evidence on Discourse Processing

As opposed to the dominance of the left hemisphere in carrying out single-word and sentence processing, the right hemisphere appears to play an important role in discourse comprehension and production. Research with right hemisphere-damaged patients demonstrates an impaired understanding of various forms of discourse, including indirect requests ("Can you reach the bowl?"), jokes, ironic comments, and metaphors. This problem with discourse materials is not limited to verbal materials, but may be observed with stories that are told through cartoons or other pictures. Left hemisphere-damaged patients, in contrast, perform better than the right hemisphere-damaged patients and may even display normal levels of performance, particularly if the material is presented in a nonverbal format.

As discussed earlier, research on single-word and sentence processing indicates that multiple meanings or senses of words are activated and maintained by the right hemisphere. This has led to the proposal that semantic coding in the right hemisphere is coarse, allowing for a broad but weak activation of semantic representations that are distantly related to the word or context being processed. Weak activation of a broad semantic field may be of little use when the literal meaning of a word is required or when the most likely inference is the correct one. However, the activation and maintenance of multiple representations may allow the comprehension system to recognize and capitalize on distant semantic relationships in the comprehension of jokes or metaphors or in instances in which the comprehender has to revise an initial interpretation.

Like discourse and text comprehension, narrative production can be impaired even when production at the single-word or sentence level is preserved. That is, a speaker may fail to have an organized plan for communicating information, going off on topics tangential to the central one. The speaker might also fail to make coherent anaphoric reference between sentences or omit information that would be critical to the listener for drawing the appropriate inferences. Patients with right hemisphere brain damage and those with frontal damage have been shown to have discourse production deficits that are disproportionate to any deficits in vocabulary and grammar. Although relatively few studies have been carried out with right hemisphere-damaged patients, their deficits include an underspecification of information and a failure to take into account the listener's point of view. A larger number of studies have examined discourse deficits in individuals with closed head-injury (e.g., from a car accident). Individuals with closed-head injury often have frontal damage, and it has been assumed that the discourse deficits displayed by these patients are due to damage to this region. These individuals also display reduced information content and a lack of cohesion among the sentences in a story.

Some caveats should be kept in mind when considering the findings on discourse deficits. Much of this research has been carried out using group studies, even though researchers have found wide variation in the discourse deficits exhibited by right hemisphere-damaged or closed-head injury patients, with some patients showing no deficits. Clearly, the right hemisphere and the frontal lobes (both left and right) are very large cortical regions, and it is likely that only specific regions within these large areas are critical for discourse processing. Moreover, there are a variety of aspects of discourse processing that could be affected in different subgroups. Some individuals might have difficulty with the working memory demands involved in maintaining information across sentences during discourse comprehension or in planning a narrative to produce. Other patients might have difficulty taking into account other individuals' points of view, which makes their discourse production lack cohesion. Still others might have difficulty inhibiting irrelevant information that becomes activated during comprehension or production. Yet others might have subtle deficits at the level of lexical retrieval that become critical when several words have to be kept active simultaneously in planning discourse production. Carefully designed case studies testing for all these possibilities are necessary to delineate which components of discourse production are functionally isolable and which brain areas provide the cortical substrate for these functions.

IV. CONCLUSIONS

A substantial body of behavioral data has been amassed concerning language and lexical processing. Studies of brain-damaged patients indicate that considerable modularity of processing components exists in these domains. Although there is some knowledge concerning the brain areas involved in various aspects of language processing, the knowledge is currently at a broad rather than fine-grained level. Some of the same brain areas are implicated in a variety of language functions. For example, the temporal lobe appears to be involved in phonological, semantic, and syntactic processing. It is possible, however, that further study will reveal that nonoverlapping areas within the temporal lobe are involved in these different aspects. Future research using series of case studies with well-identified functional deficits and neuroimaging studies with normal subjects should provide a better specification of the localization and interaction of brain areas involved in word, sentence, and discourse processing.

See Also the Following Articles

APHASIA • BILINGUALISM • BROCA'S AREA • CREATIVITY • EVOLUTION OF THE BRAIN • INFORMATION PROCESSING • LANGUAGE ACQUISITION • LANGUAGE DISORDERS • LANGUAGE, NEURAL BASIS OF • NUMBER PROCESSING AND ARITHMETIC • READING DISORDERS, DEVELOPMENTAL • SEMANTIC MEMORY • SPEECH • WERNICKE'S AREA

Suggested Reading

Ainsworth-Darnell, K., Shulman, H., and Boland, J. (1998). Dissociating brain responses to syntactic and semantic anomalies: Evidence from event-related potentials. *J. Memory Lang*. **38,** 112–130.

Berndt, R. S. (1991). Sentence processing in aphasia. In *Acquired Aphasia* (M. Sarno, Ed.), 2nd ed., pp. 223–270. Academic Press, San Diego.

Caplan, D., Alpert, N., and Waters, G. (1998). Effects of syntactic structure and propositional number on patterns of regional cerebral blood flow. *J. Cognitive Neurosci*. **10,** 541–552.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., and Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychol. Rev*. **104,** 801–838.

Eggert, G. H. (1977). *Wernicke's Works on Aphasia: A Sourcebook and Review*. Mouton, The Hague.

Joanette, Y., and Brownell, H. (1990). *Discourse Ability and Brain Damage: Theoretical and Empirical Perspectives*. Springer-Verlag, New York.

Linebarger, M., Schwartz, M., and Saffran, E. (1983). Sensitivity to grammatical structure in so-called agrammatic aphasics. *Cognition* **13,** 361–392.

Martin, R. C. (2001). Sentence comprehension deficits. In *Handbook of Cognitive Neuropsychology*. (B. Rapp, Ed.), Psychology Press, Philadelphia.

Mazoyer, B. M., Tzourio, N., Frak, V., Syrota, A., Murayama, N., Levrier, O., Salomon, G., Dehaene, S., Cohen, L., and Mehler, J. (1993). The cortical representations of speech. *J. Cognitive Neurosci*. **5,** 467–479.

Shelton, J. R., and Caramazza, A. (1999). Deficits in lexical and semantic processing: Implications for models of normal language. *Psychonomic Bull. Rev*. **6,** 5–27.

Tyler, L. (1992). *Spoken Language Comprehension: An Experimental Approach to Disordered and Normal Processing*. MIT Press, Cambridge, MA.

Language Disorders

DAVID CAPLAN
Massachusetts General Hospital

GLOSSARY

agrammatism Omission of function words, prefixes, and suffixes from speech.

agraphia Disturbance of writing.

alexia Disturbance of reading.

anomia Difficulty producing words, often nouns.

aphasia Disorders of language due to disease of the brain.

Broca's aphasia A type of aphasia with nonfluent speech and relative preserved comprehension.

paragrammatism Production of incorrect function words, prefixes and suffixes.

phonemic paraphasia Sound substitution in speech.

Wernicke's aphasia A type of aphasia with fluent error-containing speech and impaired comprehension.

Language is a code that relates different types of forms to semantic meanings. The forms of the code and their associated meanings are activated in the tasks of speaking, comprehension of spoken language, reading, and writing. Disorders of the brain can affect the ability to activate these representations in these tasks. Neurological and psychiatric disorders can also affect the use of language to accomplish tasks such as communicating ideas, storing information in long-term memory, reasoning, and solving problems. Language disorders that have been most extensively studied neurologically are those that affect language processing in the tasks of language use.

I. LANGUAGE STRUCTURE AND PROCESSING

Human language can be viewed as a code that links linguistic representations to aspects of meaning. The basic types of representations of language (or language levels) are simple words (the lexical level), words with internal structure (the morphological level), sentences, and discourse. The lexical level of language makes contact with the categorial structure of the world. Lexical items (simple words) designate concrete objects, abstract concepts, actions, properties, and logical connectives. The basic form of a simple lexical item consists of a phonological representation that specifies the segmental elements (phonemes) of the word and their organization into metrical structures (e.g., syllables). The form of a word can also be represented orthographically. Words are associated with syntactic categories (e.g., noun and verb). The morphological level of language allows the meaning associated with a simple lexical item to be used as a different syntactic category (e.g., noun formation with the suffix *-tion* allows the semantic values associated with a verb to be used as a noun, as in the word "destruction" derived from "destroy") and thus avoids the need for an enormous number of elementary lexical items in an individual's vocabulary. The sentential level of language makes use of the syntactic categories of lexical items to build hierarchically organized syntactic structures (e.g., noun phrase, verb phrase, and sentence) that define relationships between words relevant

to the propositional content of a sentence. The propositional content of a sentence expresses aspects of meaning, such as thematic roles (who did what to whom), attribution of modification (which adjectives go with which nouns), the reference of pronouns, and other referentially dependent categories. Propositional meanings make assertions that can be entered into logical and planning processes and that can serve as a means for updating an individual's knowledge of the world. The discourse level of language includes information about the general topic under discussion, the focus of a speaker's attention, the novelty of the information in a given sentence, the temporal order of events, causation, and so on. Information conveyed by the discourse level of language also serves as a basis for updating an individual's knowledge of the world and for reasoning and planning action.

The forms of language and their associated meanings are activated in the usual tasks of language use—speaking, auditory comprehension, reading, and writing. There is wide agreement among linguists, psychologists, and computer scientists that these different forms are activated by different "components" of a "language processing system." Components of the cognitive processing system are devices that accept as input a certain type of representation and operate on these inputs to activate another type of representation, where at least one of these representations is part of the language code. For instance, a component of the language processing system might accept as input the semantic representation (meaning) activated by the presentation of a picture and produce as output a representation of the sound pattern of the word that corresponds to that meaning.

The operations of the components of the language processing system are obligatory and largely unconscious. The obligatory nature of language processing can be appreciated intuitively by considering that we are generally unable to inhibit the performance of many language processing tasks once the system is engaged by an appropriate, attended input. For instance, we must perceive a spoken word as a word, not just as a nonlinguistic percept. The unconscious nature of most of language processing can be appreciated by considering that when we listen to a lecture, converse with an interlocutor, read a novel, or engage in some other language processing task, we usually have the subjective impression that we are extracting another person's meaning and producing linguistic forms appropriate to our intentions without paying attention to the details of the sounds of words, sentence structure, etc.

In general, cognitive processes that are automatic and unconscious are thought to require relatively little allocation of mental resources. However, many experimental results indicate that language processing does require the allocation of attention and/or processing resources. The efficiency of each of the components of the language processing system is thought to be a function of the resources available to that component, up to the maximum level of resource utilization of which the component is capable. Components of the system are remarkably efficient. For instance, it has been estimated on the basis of many different psycholinguistic experimental techniques that spoken words are usually recognized less than 125 msec after their onset (i.e., while they are still being uttered). Similarly, normal word production in speech requires searching through a mental word production "dictionary" of over 20,000 items, but it still occurs at the rate of about three words per second with an error rate of about one word misselected per 1 thousand and another one word mispronounced per 1 thousand. The efficiency of the language processing system as a whole reflects the efficiency of each of its components but also is achieved because of the massively parallel computational architecture of the system, in which many components of the system are simultaneously active.

Functional communication involving the language code occurs when people use these processors to undertake language-related tasks to accomplish specific goals—to inform others, to ask for information, to get things done, etc. The language code is remarkably powerful with respect to the semantic meanings it can encode and convey, and psycholinguistic processors are astonishingly fast and accurate. The ability to use this code quickly and accurately is critical to human success, both as a species and as individuals.

II. CLASSICAL APHASIC SYNDROMES

The language disorders that are best understood are those that affect the representation and processing of language in the usual tasks of language use. These disorders are known as "aphasia." By convention, the term aphasia does not refer to disturbances that affect the functions to which language processing is put. Lying (even transparent, ineffectual lying) is not considered a form of aphasia, nor is the garrulousness of old age or the incoherence of schizophrenia.

Neurologists and speech–language pathologists often describe aphasic patients as having one of a number of syndromes, which were first described in the second half of the nineteenth century. Because these syndromes are the foundation for contemporary descriptions of language impairments, and because they are still widely used by clinicians, I shall describe them here.

The paper that first led researchers to these syndromes was written by Paul Broca in 1861. Broca described a patient, Leborgne, with a severe speech output disturbance. Leborgne's speech was limited to the monosyllable "tan." In contrast, Broca described Leborgne's ability to understand spoken language and to express himself through gestures and facial expressions, as well as his understanding of nonverbal communication, as being normal. Broca claimed that Leborgne had lost "the faculty of articulate speech." Broca related this impairment to the neural tissue most badly damaged in Leborgne—the posterior portion of the inferior frontal convolution of the left hemisphere, which became known as Broca's area.

During the ensuing years, many cases of language impairments were described. In some, speech impairments were related to lesions in the left frontal lobe. Other speech impairments were associated with more posterior lesions. In 1874, Carl Wernicke, then a medical student, published a paper that appeared to reconcile many of these different findings. Wernicke described a patient with a speech disturbance, but one that was very different from that seen in Leborgne. Wernicke's patient was fluent; her speech, however, contained words with sound errors, other errors of word forms, and words that were semantically inappropriate. Also unlike Leborgne, Wernicke's patient did not understand spoken language. Wernicke related the two impairments—the one of speech production and the one of comprehension—by arguing that the patient had sustained damage to "the storehouse of auditory word forms." Under these conditions, speech would be expected to contain the types of errors that were seen in this case, and comprehension would be affected. Establishing the location of the lesion in this case was more problematic, however. Wernicke did not have the opportunity to perform an autopsy on his patient. However, he did examine the brain of a second patient, whose language had been described prior to her death by her physician in terms that made Wernicke think that she had a set of symptoms that were the same as those he had seen in his patient. The lesion in this second patient occupied the posterior portion of the first temporal gyrus, also on the left. Wernicke suggested that this region, which came to be known as Wernicke's area, was the locus of the storehouse of auditory word forms.

These two seminal papers gave rise to a model of the relation of language processing to the brain. In this model, speaking involves activating the forms of words (in Wernicke's area) from concepts and transmitting these word forms to Broca's area to plan speech; Broca's area is also directly activated by concepts. Comprehension involves activating word forms from auditory input and then activating word meanings. Repetition involves transmitting the forms of words from Wernicke's area to Broca's area. Based on this model, researchers identified a number of aphasic syndromes, summarized in Table I.

The first two of these syndromes are Broca's and Wernicke's aphasia. Broca's aphasia, which affects primarily expressive language and leads to nonfluent speech, is due to lesions in Broca's area, the center for motor speech planning adjacent to the motor strip. Wernicke's aphasia follows lesions in Wernicke's area that disturb the representations of word sounds. Pure motor speech disorders arise from lesions interrupting the motor pathways from the cortex to the brain stem nuclei that control the articulatory system. These disorders differ from Broca's aphasic because they are not linguistic; they affect articulation, not the planning of speech. Pure word deafness affects the transmission of sound input into Wernicke's area. It therefore disrupts word recognition but not speech since words are intact and accessible for speech production purposes. Transcortical motor aphasia results from the interruption of the pathway from the concept center to Broca's area. This affects speech but not repetition or comprehension. Transcortical sensory aphasia follows lesions between Wernicke's area and the concept center. Repetition of words is intact, but comprehension is affected. Finally, conduction aphasia follows from a lesion between Wernicke's area and Broca's area. Repetition is affected, but comprehension is intact. Speech is also affected, in the same way as it is affected in Wernicke's aphasia, because the sound patterns of words, though activated, are not transmitted properly to Broca's area to be used to plan speech.

These syndromes have defined the domain of aphasia as a description of performances in the usual tasks of language use—speaking, understanding spoken language, reading, and writing. In terms of the linguistic elements, the disorders affect words, sounds of words, word endings, and classes of words such as function words that are produced, recognized, or understood. In this content, they contrast with other

Table I
Classical Aphasic Syndromes

Syndrome	Clinical manifestations	Postulated deficit	Classical lesion location
Broca's aphasia	Major disturbance in speech production with sparse, halting speech, often misarticulated, and frequently missing function words and bound morphemes	Disturbances in the speech planning and production mechanisms	Posterior aspects of the third frontal convolution (Broca's area)
Wernicke's aphasia	Major disturbance in auditory comprehension; fluent speech with disturbances of the sounds and structures of words (phonemic, morphological, and semantic paraphasias)	Disturbances of the permanent representations of the sound structures of words	Posterior half of the first temporal gyrus and possibly adjacent cortex (Wernicke's area)
Pure motor speech disorder	Disturbance of articulation; apraxia of speech, dysarthria, anarthria, aphemia	Disturbance of articulatory mechanisms	Outflow tracts from motor cortex
Pure word deafness	Disturbance of spoken word comprehension	Failure to access spoken words	Input tracts from auditory system to Wernicke's area
Transcortical motor aphasia	Disturbance of spontaneous speech similar to Broca's aphasia, with relatively preserved repetition	Disconnection between conceptual representations of words and sentences and the motor speech production system	White matter tracts deep to Broca's area connecting it to parietal lobe
Transcortical sensory aphasia	Disturbance in single word comprehension, with relatively intact repetition	Disturbance in activation of word meanings despite normal recognition of auditorily presented words	White matter tracts connecting parietal lobe to temporal lobe or portions of inferior parietal lobe
Conduction aphasia	Disturbance of repetition and spontaneous speech (phonemic paraphasias)	Disconnection between the sound patterns of words and the speech production mechanism	Lesion in the arcuate fasciculus and/or corticocortical connections between Wernicke's and Broca's areas
Anomic aphasia	Disturbance in the production of single words, most marked for common nouns with variable comprehension problems	Disturbances of concepts and/or the sound patterns of words	Inferior parietal lobe or connections between parietal lobe and temporal lobe; can follow many lesions
Global aphasia	Major disturbance in all language functions	Disruption of all language processing components	Large portion of the peri-Sylvian association cortex
Isolation of the language zone	Disturbance of both spontaneous speech (similar to Broca's aphasia) and comprehension, with some preservation of repetition	Disconnection between concepts and both representations of word sounds and the speech production mechanism	Cortex just outside the peri-Sylvian association cortex

approaches to the description of aphasia. For instance, Hughlings Jackson described a patient, a carpenter, who was mute but who mustered the capacity to say "master's" in response to his son's question about where his tools were. Jackson's poignant comments convey his emphasis on the conditions that provoke speech rather than on the form of the speech:

> *The father had left work; would never return to it; was away from home; his son was on a visit, and the question was directly put to the patient. Anyone who saw the abject poverty the poor man's family lived in would admit that these tools were of immense value to them. Hence we have to consider as regards this and other occasional utterances the strength of the accompanying emotional state.*

Jackson and others sought a description of language use as a function of motivational and intellectual states and tried to describe aphasic disturbances of language in relationship to the factors that drive language production and make for depth of comprehension. This is a vital aspect of understanding language impairments. In many ways, it is more humanly

relevant than a description of language impairments in terms of which phonemes are produced in spontaneous speech or repetition. Unfortunately, it is a very intractable goal, both in terms of psychological descriptions and in terms of relating these specific motivational states to the brain. The researchers who conceived and developed the framework of the classical syndromes focused aphasiology on the description of the linguistic representations and psycholinguistic operations that are responsible for everyday language use.

III. PROBLEMS WITH THE CLASSICAL APHASIC SYNDROMES

A major limitation of the classical syndromes is that they stay at arm's length from the linguistic details of language impairments. The classical aphasic syndromes basically reflect the relative ability of patients to perform entire language tasks (speaking, comprehension, etc.), not the integrity of specific operations within the language processing system. This is not to say that there are no linguistic or qualitative descriptions of language in the characterizations of the classical aphasic syndromes, only that they are incomplete and unsystematic.

A second problem with these syndromes is that they do not classify many aphasic patients very well. In practice, most applications of the clinical taxonomy result in widespread disagreements as to a patient's classification and/or to a large number of "mixed" or "unclassifiable." The criteria for inclusion in a syndrome are often arbitrary: How bad does a patient's comprehension have to be called a Wernicke's aphasic instead of a conduction aphasic or a global aphasic instead of a Broca's aphasic? Part of the problem is that patients can only be assigned to a single syndrome instead of being thought of as having multiple deficits.

A third problem that the classical aphasic syndromes face is that they are not as well correlated with lesion sites as the theory claims they should be. The correlations do not apply to many types of lesions, such as various sorts of tumors, degenerative diseases, and others. The classical syndromes are only related to lesion sites in cases of rapidly developing lesions, such as stroke; even in these types of lesions, they are not related to acute and subacute phases of the illness. Even in the chronic phase of diseases such as stroke, as many as 40% of patients have lesions that are not predictable from their syndromes according to some studies.

IV. PSYCHOLINGUISTIC DESCRIPTIONS OF APHASIC IMPAIRMENTS

Modern psycholinguistic analyses have greatly amplified the descriptions of aphasic syndromes. This work seeks to describe the aspects of language that are affected in individual patients. It differs from the syndrome approach in that a patient can have (and usually will have) more than one deficit. This work is far too extensive to review in detail here; I shall illustrate it with a few selective descriptions of analyses undertaken in this framework.

A. Disturbances of Word Meanings

Most recent research on disturbances of word meanings in brain-damaged patients has focused on words that refer to objects. Disturbances of word meanings cause poor performance on word–picture matching and naming tasks. However, the combination of deficits in word–picture matching and naming may be due to separate input- and output-side processing disturbances that affect word recognition and production. Cooccurring deficits in naming and word–picture matching are more likely to result from a disturbance affecting concepts when (i) the patient makes many semantic errors in providing words to pictures and definitions, (ii) he or she has trouble with word–picture matching with semantic but not phonological foils, (iii) he or she fails on categorization tasks with pictures, and (iv) the same words are affected in production and comprehension tasks.

Disorders affecting processing of semantic representations for objects may be specific to certain types of inputs. Elizabeth Warrington first noted a discrepancy between comprehension of words and pictures in two dementing patients. Dan Bub and colleagues described a patient who showed very poor comprehension of written and spoken words but quite good comprehension of pictures. These impairments have been taken as reflections of disturbances of "verbal" and "visual" semantic systems, although others have disputed this conclusion.

Semantic disturbances may also be category specific. Several authors have reported a selective semantic impairment of concepts related to living things and foods compared to man-made objects. The opposite pattern has also been found. Selective preservation and disruption of abstract versus concrete concepts, and of nominal versus verbal concepts, have also been reported.

Disturbances may affect the unconscious activation of semantic meanings or their conscious use. There are patients who cannot match words to pictures or name objects but who are unconsciously influenced by a word's meaning. For instance, in a word/nonword decision task, they will respond faster to the word "doctor" when it follows "nurse" than when it follows "house," indicating they are able to appreciate the relations between words unconsciously, even when they cannot indicate understanding of word meaning in conscious, controlled tasks such as word–picture matching. Conversely, some patients who appear to understand words well have abnormalities in tasks that examine unconscious processing of the meanings of words.

B. Disturbances of Oral Word Production

Disturbances affecting the oral production of single words are extremely common in language-impaired patients. There are three basic disturbances affecting word production (other than semantic deficits). They follow the stages of word sound production: accessing the forms of words from concepts, planning the form of a word for articulation, and articulation.

A disturbance in activating word forms from concepts is manifest by an inability to produce a word from a semantic stimulus (a picture or a definition), coupled with intact processing at the semantic and phonological levels (determined by answering questions about pictures, picture categorization tests, and repetition). The form of a patient's errors is not a good guide to whether he or she has an impairment at this level of the production process since disturbances in accessing word forms may appear in a variety of ways, ranging from pauses to neologisms (complex sequences of sounds that do not form words) and semantic paraphasias (words related to the meaning of the target item). Rarely, patients show an inability to name objects presented in one modality only (e.g., visually) even though they demonstrate understanding of the concept associated with that object when it is presented in that modality (optic aphasia). Because disorders of basic sensory and motor functions can be ruled out in these patients, these modality-specific naming disorders have been taken to reflect a failure to transmit information from modality-specific semantic systems to the processor responsible for activating the forms of words.

Disturbances of a patient's ability to convert the representation of the sound of a word into a form appropriate for articulatory production are usually manifest as phonemic paraphasias (substitutions, omissions, and misorderings of phonemes). Three features of a patient's performance suggest a disturbance in word sound planning. First, some phonemic paraphasias are closely related to target words (e.g., "befenit" for "benefit"). Second, some patients make multiple attempts that come increasingly closer to the correct form of a word. Third, some patients make similar phonological errors in word repetition, word reading, and picture naming. Because the form of a word is presented to the output system in very different ways in these three tasks, the errors in such patients most likely arise in the process of planning the form of the word that is suitable for articulation.

Patients with sound planning problems tend to be more affected on longer words and on words with consonant clusters. The frequency of occurrence of a word in the language has a variable effect on the occurrence of these types of errors. Planning disturbances only rarely affect function words compared to nouns, verbs, and adjectives. Some patients have trouble planning the sounds of words only when words are inserted into sentences, making phonemic paraphasias in sentence production but not naming or repetition tasks. In these cases, the errors probably arise when words are inserted into syntactic structures.

Patients often have disturbances of articulation, as shown by abnormalities in the acoustic waveform produced by a patient and in the movement of the articulators in speech. Investigators have identified two major disturbances of articulation—dysarthria and apraxia of speech. Dysarthria is marked by hoarseness, excessive nasality, and imprecise articulation, and it has been said to not be significantly influenced by the type of linguistic material that the speaker produces or by the speech task. Apraxia of speech is marked by difficulty in initiating speech, searching for a pronunciation, better articulation for automatized speech (e.g., counting) than volitional speech, abnormal prosody, omissions of syllables in multisyllabic words, and simplification of consonant clusters (often by adding a short neutral vowel sound between consonants). Both dysarthria and apraxia of speech result in sounds that are perceived as distorted. Apraxia of speech often cooccurs with dysarthria or with the production of phonemic paraphasias, and the relations between these disorders and the empirical basis for distinguishing one from another are the subject of active research.

C. Disturbances of Recognition of Auditorily Presented Simple Words

Disturbances affecting auditory comprehension of simple words have been attributed to impairments of semantic concepts, as discussed previously, and/or to an inability to recognize spoken words. The latter disturbances have, in turn, been thought to have two possible origins: disturbances affecting the recognition of phonemes in the acoustic signal and disturbances affecting the ability to recognize words despite good acoustic–phonetic processing.

Disturbances of acoustic–phonetic processing may affect the ability to discriminate or to identify phonemes. It is unclear, however, whether these disturbances lead to problems in recognizing or understanding spoken words. Several studies suggest that they do, but other researchers have found weak correlations between comprehension capacities and phoneme discrimination capacities in language-impaired patients.

Many researchers believe that patients can have disturbances of spoken word recognition despite good acoustic–phonetic processing. Such a disturbance was originally postulated by Carl Wernicke. However, there is no clear case of a patient who has intact acoustic–phonetic processing and who cannot recognize spoken words. In most cases, single-word comprehension problems are probably multifactorial in origin and result from a complex interaction of acoustic–phonetic disturbances, disturbances in recognizing spoken words, and disturbances affecting word meanings.

D. Disorders of Repetition of Single Words

Repetition of a word can be carried out in three ways: (i) nonlexically by repeating sounds without recognizing the word (as if one were imitating a foreign language), (ii) lexically by recognizing the stimulus as a word and uttering it without understanding it, and (iii) semantically by understanding the word and reactivating its form from its meaning. Any of these routes to repetition may be disturbed. For instance, one patient could only repeat by the semantic route; this patient made many semantic paraphasias in repetition and could not repeat nonwords. Patients with relatively isolated disturbances affecting the repetition of nonwords have been described, reflecting disruption of the nonlexical route. In most cases, patients have a more complicated picture, with lexical status (whether a stimulus is a word or a nonword), word frequency, and stimulus length affecting performance differently in different patients.

E. Disturbances of Processing Morphologically Complex Words

Disturbances affecting both the comprehension and the production of morphologically complex words have been described. With respect to recognition of morphologically complex words, researchers have observed that some patients who make derivational paralexic errors (e.g., "write" → "wrote," "fish" → "fishing," and "directing" → "direction") in the oral reading of complex words have particular difficulty with the recognition and analysis of written morphologically complex words compared to morphologically simple words. A patient with a disturbance affecting the auditory processing of words with inflectional but not derivational morphology has been described.

Disturbances affecting morphological processing also appear in single-word production tasks. In one study, patients had difficulties in producing plural, possessive, and third-person singular forms of nonwords. The fact that this disorder arose with nonwords that the patients were given by the experimenters suggests that the impairment affected the ability to construct new morphological forms. Such disturbances can arise in patients who perform well on tasks that require recognition and comprehension of written morphologically complex words.

Disturbances affecting the production of morphologically complex words are most commonly seen in sentence production, where they are known as "agrammatism" and "paragrammatism." The most noticeable deficit in agrammatism is the widespread omission of function words and affixes and the better production of common nouns. This disparity is always seen in the spontaneous speech of patients termed agrammatic, and it often occurs in their repetition and writing as well. Patients in whom substitutions of these elements predominate, and whose speech is fluent, are called paragrammatic. Recent observations have emphasized the fact that these two patterns cooccur in many patients. They may result from a single underlying deficit that has different surface manifestations.

Agrammatism and paragrammatism vary considerably, with different sets of function words and bound morphemes being affected or spared in different cases. In some patients, there seems to be some systematicity to the pattern of errors. For instance,

English agrammatic patients frequently produce infinitives (e.g., "to walk") and gerunds (e.g., "walking") because these are the basic forms in the verbal system. In other cases, substitutions are closely related to the correct target. Agrammatics' errors also tend to follow the tendencies seen in normal subjects with respect to errors that "strand" affixes (e.g., "I am going to school" → "I am schooling to go") and the "sonorance hierarchy" that establishes syllabic forms as easier to produce than simple consonants. Agrammatics generally produce real words, which makes for different patterns of errors in different languages that differ with respect to whether or not they require inflections to appear on a word. The fact that in almost all cases errors do not violate the word-formation processes of the language suggests that most agrammatic and paragrammatic patients retain some knowledge of the rules of word formation.

F. Disorders of Sentence Production

Disturbances at the sentence production level are the inevitable results of disturbances affecting the production of simple or complex words. In addition, many patients have problems in the sentence planning process.

Agrammatic patients usually produce only very simple syntactic structures. In one study, virtually no syntactically well-formed syntactic constructions were found in the utterances produced by one agrammatic patient. All the agrammatic patients studied in a large contemporary cross-language study showed some impoverishment of syntactic structure in spontaneous speech. The failure to produce complex noun phrases and embedded verbs with normal frequency were the most striking features of the syntactic simplification shown by these patients.

The ability to express the thematic roles of noun phrases requires the ability to use verbs. Many agrammatic patients have problems with verbs; in one, a category-specific degradation of the meaning of verbs resulted in almost no production of verbs in speech and limited the ability to convey thematic roles of nouns. In other studies, patients' inabilities to produce verbs were only partially responsible for the shortened phrase length found in their speech. It thus appears that in some patients a disturbance affecting the ability to produce verbs affects the production of a normal range of syntactic structures, whereas in others at least some syntactic structures are built despite poor verb production. Yet other patients cannot produce normal syntactic structures despite relatively good verb production.

Several studies suggest that syntactic errors in sentence production differ in paragrammatic and agrammatic patients. Five characteristic paragrammatic patients each produced many long and complex sentences, with multiple interdependencies of constituents, but tended to produce many types of syntactic errors, including errors in tag questions, illegal noun phrases in relative clauses, and illegal use of pronouns to head relative clauses. A type of error that has often been commented on in paragrammatism is a "blend," in which the output seems to reflect a conflation of two different ways of saying the same thing (e.g., "They are not prepared to be of helpful," which is a combination of "They are not prepared to be helpful" and "They are not prepared to be of help"). Some paragrammatic patients can solve anagram tasks according to syntactic constraints (whereas agrammatic patients solved them using semantic constraints). Because of the evidence that paragrammatic patients retain some ability to use syntactic structures, some researchers have suggested that the syntactic and morphological errors in paragrammatism result from the failure of these patients to monitor their speech production processes and their output.

The various disturbances that affect sentence production usually cooccur. A complex disturbance that results from the combination of the deficits in producing syntactic forms, disturbances in accessing and planning word forms, and impairments in producing morphologically complex words is known as "jargonaphasia."

Many patients have difficulty producing prosodic aspects of speech. These disturbances may be secondary to motor output disorders or associated with other sentence production disorders. However, these disturbances may occur in isolation. These are different from the aprosodias related to emotional display described after right hemisphere disease. They reflect a primary disturbance of production of intonation in right hemisphere-damaged patients that differs as a function of lesion location in the hemisphere.

G. Disorders of Sentence Comprehension

When a subject understands a sentence, he or she combines the meanings of the words in accordance with the syntactic structure of the sentence into a propositional content. There are many reasons why a patient might fail to understand the propositional

content of a sentence. In addition to the carryover effect of disturbances affecting comprehension of simple and complex words, there are disturbances affecting the ability to understand aspects of propositional meaning despite good single-word processing.

The greatest amount of work in the area of disturbances of sentence comprehension has gone into the investigation of patients whose use of syntactic structures to assign meaning is not normal. The first researchers to show that some patients have selective impairments of this ability described patients who could match "semantically reversible" sentences such as "The apple the boy is eating is red" to one of two pictures but not "semantically irreversible" sentences such as "The girl the boy is chasing is tall." The difference between the two types of sentences resides in the fact that a listener can understand a sentence such as "The apple the boy is eating is red" via the lexico-pragmatic route, whereas understanding "The girl the boy is chasing is tall" requires assigning its syntactic structure since both boys and girls are capable of chasing one another.

Disorders of syntactic comprehension have since been examined in considerable detail. Patients may have very selective disturbances affecting the use of particular syntactic structures or elements to determine the meaning of a sentence. For instance, patients may understand sentences with reflexives ("himself") but not pronouns ("him") and vice versa. Some patients can understand very simple syntactic forms, such as active sentences ("The man hugged the woman"), but not more complex forms, such as passive sentences ("The woman was hugged by the man"). Many of these patients use strategies such as assigning the thematic role of agent to a noun immediately before a verb to understand semantically reversible sentences, leading to systematic errors in comprehension of sentences such as "The boy who pushed the girl kissed the baby." Other patients have virtually no ability to use syntactic structure. Most of these patients rely on inferences based on their knowledge of the real world and their ability to understand some words in a sentence.

Some patients can assign and interpret syntactic structures unconsciously but cannot use these structures in a conscious, controlled fashion. For instance, one patient's word-monitoring performances indicated that he was sensitive to certain syntactic anomalies but could not make judgments regarding these same anomalies at the end of a sentence. Some patients who have syntactic comprehension problems (e.g., who cannot match reversible sentences to pictures) can make judgments as to whether or not a sentence is grammatical. For instance, some patients can indicate that the utterance "The woman was watched the man" is ill formed and the utterance "The woman was watched by the man" is acceptable, despite not being able to match sentences such as "The woman was watched by the man" to one of two pictures. These results suggest that these patients can construct syntactic structures but cannot use them to determine propositional meaning (a so-called "mapping" problem). As with other areas of language functioning, it appears that patients may retain unconscious, on-line sentence comprehension processes but lose the ability to use the products of these processes in a controlled, conscious fashion in some tasks.

Some researchers maintain the view that short-term memory is used in comprehending more complex sentences and have pointed out a variety of sentence comprehension disturbances in patients with short-term memory limitations. Sentence length has been shown to affect certain comprehension tasks in some patients who do not show disturbances of syntactic comprehension. However, case studies show that patients with short-term memory impairments can have excellent syntactic comprehension abilities. Although many short-term memory patients have trouble in comprehension tasks, the relationship of these short-term memory disorders to sentence comprehension impairments remains unclear.

H. Disorders of Reading Single Words

The contemporary study of acquired dyslexias has largely focused on impairments in the ability to read single words aloud. One model of the mechanisms involved in reading a word aloud claims that there are three separate and partially independent routines in the brain for converting a written word into its spoken form: The first pathway (the semantic route) involves recognizing a word visually, gaining access to its meaning, and then activating the sound of the word from its meaning. The second pathway (the whole word route) translates the orthography of the entire word directly into a pronunciation, without first contacting the word's meaning. Finally, a third pathway (the sublexical route) decomposes the word into orthographic segments (graphemes and other spelling units) and derives a pronunciation by assigning each of them a spoken (phonemic) value. Other models of the reading-aloud process combine the second and third processes into one highly interactive system.

Brain damage can selectively affect a particular routine without impairing the function of the remaining branches of the system. Certain patients (phonological alexics) lose the operation of the sublexical route, which acts on subword units to assemble a response, leaving the semantic and whole word routes available. A patient with damage to the sublexical route is impaired in the reading aloud of nonwords, whereas legitimate words are mostly read correctly. Other reading disorders appear to be the outcome of severe damage to both the whole word route and the sublexical route. Patients with this pattern or impairment, termed deep dyslexics, are unable to pronounce written nonwords (indicating severe impairment to the sublexical route) but also make semantic paralexias (e.g., they read "chair" as "table") and are poor at reading abstract words aloud relative to concrete words. The failure to read nonwords combined with the presence of semantic errors and the influence of a semantic variable on the ability to read a written word aloud is consistent with the interpretation that the patient has lost the use of both the routines from whole words and subword units to sound and is forced to use a defective routine from the visual description of the word through meaning to pronunciation. Yet other dyslexic readers (surface dyslexics) have lost the ability to use both the semantically mediated reading route and the whole-word route but retain the translation of subword units into sound (the sublexical route). These patients can still read words aloud that obey regular correspondences between spelling and sound (e.g., "hint," "mint," and "stint"), but not those that do not (e.g., "pint"). Finally, some patients with dementia show the ability to use the whole-word reading route without the benefit of semantics. These patients can read irregularly spelled words correctly that they do not understand.

The fact that a patient has difficulties in reading words and/or nonwords aloud does not necessarily imply that he or she does not recognize or understand a printed word. Some patients have a severe disturbance of reading aloud known as "letter-by-letter reading" because they name each letter in a word before attempting (often with incomplete success) to pronounce the word. These patients take longer to read longer words, and it can take them many seconds to pronounce a printed word. Several of these patients have been tested for their abilities to recognize and comprehend words that are presented for short periods of time—far less than the time needed for them to read the words aloud. One such patient could recognize words and familiar letter strings as visual patterns, and other letter-by-letter readers can extract at least some semantic information from briefly presented words. Most of these patients denied that they were able to recognize or understand these briefly presented words. These performances suggest that some reading problems arise after words have been recognized, and also that some alexic subjects, such as the patients with disturbances affecting auditory–oral processing described previously, may retain abilities to recognize and understand words without being aware that they do so.

I. Disorders of Writing Single Words

The acquired dyslexias have their counterparts in the acquired agraphias. Patients with phonological agraphia are severely impaired in their ability to spell or write nonsense words but are capable of very good performance on legitimate words, even words that are low in frequency and contain unusual spelling patterns (e.g., "leopard"). The deficit is these cases appears to lie in the ability to convert sublexical phonological units to orthographic units (graphemes and letters). The converse impairment—an inability to access the written forms of whole words with preserved ability to convert sublexical phonological units to orthographic units, termed surface agraphia or lexical agraphia—has also been described. A third disturbance, known as "asemantic writing," consists of the inability to write spontaneously but the retained ability to write to dictation. This suggests that the contents of the visual word-form system can be addressed from spoken input but not from the meaning of a word.

These disorders of writing can be strikingly different from patients' reading performances. For instance, one patient was unable to write nonsense words to dictation but read these items aloud without difficulty; another produced numerous misspellings when attempting to write orthographically irregular or ambiguous words but accurately read the majority of legitimate words perfectly. These dissociations have led researchers to infer that the orthographic knowledge necessary for word recognition in reading is different from the orthographic knowledge necessary for correct spelling in writing.

The abstract graphemic representation of familiar and unfamiliar words is ultimately converted by the writing mechanism into a sequence of rapid movements that generate letters on the page. A specialized working memory device is needed to maintain the graphemic code in a buffer zone while the spatial

identity of each letter is chosen at the next processing stage. One patient made the same kind of errors in writing, oral spelling, and typing, a result that justifies the conclusion that the disturbance arose while these items were in the planning buffer and before the programming and execution of a particular motor act began.

Written production takes place by generating the spatial form of each letter (allographs) in the correct order. A few agraphic cases have been documented in which the impairment is plainly confined to the retrieval of elements in the allographic code. In these patients, adequate knowledge of the word's orthography can be demonstrated because oral spelling is carried out extremely well. When tested, other methods of forming a printed word that do not require a written response (typing or use of block letters) may also yield a high degree of accuracy. Writing, however, is characterized by numerous errors of omission, substitution, reversals, and insertions. The agraphia is not merely a disturbance in the production of a graphic motor pattern because patients can write single letters to dictation, and their writing of words, although flawed, is clearly legible.

Finally, the motor schema for a letter appears to distinguish between the movements denoting the shape of a letter and the parameters that govern scale factors such as magnitude and orientation. Cases of apractic agraphia reveal a loss of the motor programs necessary for producing letters. Written characters are poorly formed and may be indecipherable, although even severely affected patients maintain the distinction between cursive and printed letters and between upper- and lowercase. Evidence indicates that the disturbance need not be associated with limb apraxia. Certain patients with right hemisphere damage have no difficulty constructing written letters or words, but they exceed the correct number of strokes on letters that require repetitive movements.

V. CONCLUSION

This brief overview of the major acquired language disorders reveals their heterogeneity and specificity. It also highlights the fact that patients may have trouble with conscious, controlled use of the products of language processing, but retain the ability to compute certain language structures unconsciously on-line. Additional disorders, and further characterization of the impairments of the operating characteristics of psycholinguistic processors, continue to be documented in psycholinguistic aphasiological research. Most patients with language disorders have more than one primary language disorder and often have disorders of these processors due to other cognitive impairments as well (such as attentional deficits or problems with searching through semantic memory).

There is no simple, one-to-one relationship between impairments of elements of the language code or of psycholinguistic processors, on the one hand, and abnormalities in performing language-related tasks and accomplishing the goals of language use, on the other hand. Most patients who have disturbances of elements of the language code or psycholinguistic processors experience limitations in their functional communicative abilities. However, individuals with language processing disorders adapt to their language impairments in many ways, and some of these adaptations are remarkably effective at maintaining at least some aspects of functional communication. Time, rehabilitation, support, and a positive attitude can allow many aphasic patients to be productive and happy.

See Also the Following Articles

AGRAPHIA • ALEXIA • ANOMIA • AUTISM • BROCA'S AREA • DYSLEXIA • INFORMATION PROCESSING • LANGUAGE ACQUISITION • LANGUAGE AND LEXICAL PROCESSING • LANGUAGE, NEURAL BASIS OF • SPEECH • WERNICKE'S AREA

Suggested Reading

Caplan, D. (1992). *Language: Structure, Processing and Disorders*. MIT Press, Cambridge, MA.

Damasio, A. R. (1992). Aphasia. *N. Engl. J. Med.* **326,** 531–539.

Davis, A. (1999). *Aphasiology: Disorders and Clinical Practice*. Allyn & Bacon, New York.

Goodglass, H. (1997). *Understanding Aphasia*. Academic Press, New York.

Howard, D. (1995). Language in the human brain. In *Cognitive Neuroscience* (M. R. Rugg, Ed.), pp. 277–304. MIT Press, Cambridge, MA.

Lecours, A. R., Lhermitte, F., and Bryans, B. (1983). *Aphasiology*. Balliere Tindall, London.

Language, Neural Basis of

DAVID CAPLAN
Massachusetts General Hospital

GLOSSARY

cortex The nerve cells along the outside edge of the brain; the most advanced part of the brain.

holism Models of the neural basis of cognitive functions that maintain that these functions depend on large areas of the brain (i.e., that they are not narrowly localized).

lateralization The fact that some cognitive functions rely on one hemisphere of the brain more than another.

localization The fact that sensory, motor, and cognitive functions are supported by small areas of the brain.

Perisylvian cortex The cortex around the Sylvian fissure in the brain.

Language is a distinctly human symbol system that relates a number of different types of forms to aspects of meaning. The forms of language and their associated meanings are activated in the processes of speaking, understanding speech, reading, and writing. Several types of data provide information about the way the brain is organized to represent and process language in these tasks. These include correlations between language processing deficits and brain lesions, regional cerebral blood flow and other hemodynamic responses to language processing in normal subjects, electrophysiological and magnetoecephalographic responses to language processing in normal subjects, and the effects of electrocortical stimulation on language. These sources of data indicate that language is primarily represented and processed in perisylvian association cortex, with possible contributions from other brain areas. They indicate that different aspects of language processing are localized in different parts of the perisylvian cortex and lateralized differently in the two hemispheres. They also indicate that both localization and lateralization show some degree of variability across individuals.

I. LANGUAGE STRUCTURE

Language is a distinctly human ability that is vital to the cognitive and communicative abilities that underlie the success of humans both as individuals and as a species. Although many people think of language as a form of communication, it is more accurate to think of language as a code that can serve many functions, one of which is communication. The language code links linguistic representations to aspects of meaning. The types of representations of the language code include simple words, words with internal structure, sentences, and discourse. These types of representations are also called the lexical, word-formation, sentential, and discourse levels of language.

The lexical level of language consists of simple words. The basic form of a simple word (or lexical item) consists of a phonological representation that specifies the segmental elements (phonemes) of the word and their organization into metrical structures such as syllables. The form of a word can also be

represented orthographically. Simple words are assigned to different syntactic categories, such as nouns, verbs, adjectives, articles, and prepositions. The semantic values associated with the lexical level primarily consist of concepts and categories in the nonlinguistic world. Simple words tend to designate concrete objects, abstract concepts, actions, properties, and logical connectives.

The word-formation level of language allows words to be formed from other words. In English, word formation can take place via affixation (e.g., the word "destruction" is derived from the word "destroy") and compounding (e.g., the word "paper tray" is formed from the words "paper" and "tray"). There are two basic types of affixation in English. Derivational morphological processes allow the meaning associated with a simple lexical item to be used as a different syntactic category without coining a large number of new lexical forms that would have to be learned (e.g., "destroy" → "destruction"). Inflectional morphological processes play roles in encoding syntactic relationships (e.g., subject–verb agreement: "destroy" → "destroys").

The sentential level of language consists of syntactic structures—hierarchical sets of syntactic categories (e.g., noun phrase, verb phrase, and sentence)—into which words are inserted. The meaning of a sentence, known as its propositional content, is determined by the way the meanings of words combine in syntactic structures. Propositions convey aspects of the structure of events and states in the world. These include thematic roles (who did what to whom), attribution of modification (which adjectives go with which nouns), and the reference of pronouns and other anaphoric elements (which words in a set of sentences refer to the same items or actions). For instance, in the sentence "The big boy told the little girl to wash herself," the agent of "told" is "the big boy" and its theme is "the little girl," "big" is associated with "boy" and "little" with "girl," and "herself" refers to the same person as "girl." Sentences are a crucial level of the language code because the propositions they express make assertions about the world. These assertions can be entered into logical systems and can be used to add to an individual's knowledge of the world.

The propositional meanings conveyed by sentences are entered into higher order structures that constitute the discourse level of linguistic structure. Discourse includes information about the general topic under discussion, the focus of a speaker's attention, the novelty of the information in a given sentence, the temporal order of events, causation, and so on. Information conveyed by the discourse level of language also serves as a basis for updating an individual's knowledge of the world and for reasoning and planning action. The structure and processing of discourse involve many nonlinguistic elements and operations, such as search through semantic memory, logical inferences, and others.

II. LANGUAGE PROCESSING

Current models of language processing subdivide functions such as reading, speaking, and auditory comprehension into many different, semiindependent components, which are sometimes called modules or processors. These components of the language processing system perform highly specialized operations. For instance, the process of mapping the acoustic waveform onto phonemes and other phonological units involves a large number of highly specific operations that relate specific features of the acoustic signal to linguistically relevant units of sound. We may think of these operations as all being part of an "acoustic–phonetic" processor or module, or we may consider each of these operations as a distinct cognitive function. An analogy in the area of visual perception might be the claim that one function of the system is the identification of the three-dimensional shape of an object, which involves the identification of lines, surfaces, angles, and other geometric elements of shape; we may think of each of these more elementary perceptual operations separately or consider them as a whole with respect to their contribution to shape recognition. Different aspects of language involve different operations: The operations that have been postulated to be involved in constructing the syntactic structure of a sentence from words are different from those involved in recognizing linguistically relevant sounds. The visual system provides an analogy for this multiplication of different types of processors in that it has separate mechanisms for the perception of shape, color, texture, movement, and other visually perceptible elements. In the language system, as in the visual system, these different processing components, each composed of a variety of elementary operations, are semiindependent of the others in that each yields a particular type of representation. Each representation is finally integrated with others to achieve the overall goal of the processing system.

Information processing models of language can be expressed as flow diagrams (often called functional

architectures) that indicate the sequence of operations of the different components that perform a language-related task. These models become extremely detailed and complex when all the operations and components used in a task are specified. For present purposes, it is adequate to identify the major components of the language processing system as those processors that activate units at the lexical, word-formation, sentential, and discourse levels of the language code in the usual tasks of language use—speech, auditory comprehension, reading, and writing. This approach to defining language processing components groups together different operations that all activate a similar type of linguistic representation in a given task into a single processor.

The major components of the language processing system that can be identified at this level of detail for simple words are listed in Table I, and those for the word-formation and sentence levels are listed in Table II. Figure 1 presents a model indicating the sequence of activation of components of the lexical processing system, and Fig. 2 presents a similar model of the processing system for word formation and sentences. These tables and figures are based on the results of experimental psychological research in both normal subjects and patient populations.

Table I
Components of the Language Processing System for Simple Words

Component	Input	Operation	Output
		Auditory–oral modality	
Acoustic–phonological processing	Acoustic waveform	Matches acoustic properties to phonetic features	Phonological segments (phonemes, allophones, syllables)
Input-side lexical access	Phonological units	Activates lexical items in long-term memory on basis of sound; selects best fit to stimulus	Phonological forms of words
Input-side semantic access	Words (represented as phonological forms)	Activates semantic features of words	Word meanings
Output-side lexical access	Word meanings (“lemmas”)	Activates phonological forms of words	Phonological forms of words
Phonological output planning	Phonological forms of words (and nonwords)	Activates detailed phonetic features of words (and nonwords)	Speech
		Written modality	
Written lexical access	Abstract letter identities	Activates orthographic forms of words	Orthographic forms of words
Lexical semantic access	Orthographic forms of words	Activates semantic features of words	Word meanings
Accessing orthography from semantics	Word meanings	Activates orthographic forms of words	Orthographic forms of words
Accessing lexical orthography from lexical phonology	Phonological representations of words	Activates orthographic forms of words from their phonological forms	Orthographic forms of words
Accessing sublexical orthography from sublexical phonology	Phonological units (phonemes, other units)	Activates orthographic units corresponding to phonological units	Orthographic units in words and nonwords
Accessing lexical phonology from whole-word orthography	Orthographic forms of words	Activates phonological forms of words from their orthographic forms	Phonological forms of words
Accessing sublexical phonology from orthography	Orthographic units (graphemes, other units)	Activates phonological units corresponding to orthographic units	Phonological units in words and nonwords

Table II
Components of the Language Processing System for Derived Words and Sentences[a]

Component	Input	Operation	Output
		Processing affixed words	
Accessing morphological form	Word forms	Segments words into structural (morphological) units; activates syntactic features of words	Morphological structure; syntactic features
Morphological comprehension	Word meanings; morphological structure	Combines word roots and affixes	Meanings of morphologically complex words
Accessing affixed words from semantics	Word meanings; syntactic features	Activates forms of affixes and function words	Forms of affixes and function words
		Sentence-level processing	
Lexico inferential processing	Meanings of simple and complex words; world knowledge	Infers aspects of sentence meaning on basis of pragmatic plausibility	Aspects of propositional meaning (thematic roles, attribution of modifiers)
Syntactic comprehension	Word meanings; syntactic features	Constructs syntactic representation and combines it with word meanings	Propositional meaning
Construction of sentence form	Word forms; propositional meaning	Constructs syntactic structures; inserts word forms into structures	Sentence form (including positions of lexical items)

[a]Collapsed over auditory–oral and written modalities

Tables I and II and Figs. 1 and 2 outline the way information—in this case, sets of related linguistic representations—flows through the tasks of speaking, understanding spoken language, reading, and writing. The model depicted in these tables and figures simplifies this information flow in three ways: (i) It does not specify the nature of the operations in each of the major components of the system, (ii) it does not fully convey the extent to which the components of the system operate in parallel, and (iii) it does not convey the extent of feedback among the components of the system. Despite these simplifications, the model captures enough aspects of information processing in the language system to constitute an adequate starting place for a psycholinguistic approach to the neural basis of language.

The operations of the language processing system are regulated by a variety of control mechanisms, including both those internal to the language processor and those that are involved in other aspects of cognition. The first category, language-internal control mechanisms, probably consists of a large number of operations that schedule psycholinguistic operations on the basis of the ongoing nature of a given psycholinguistic task. The second category of control mechanisms, those that are related to cognitive processing outside the language system, determines what combinations of processors become active in order to accomplish different tasks, such as reading, repeating what one has heard, and taking notes on a lecture.

Processing components are activated in serial and in parallel to accomplish language tasks such as reading a word aloud, producing a spoken sentence, and writing a word from dictation. Different tasks require different processors. For instance, referring to Table I and Fig. 1, reading a word aloud can be accomplished in several ways. All these ''reading routes'' begin with the processor that recognizes visual patterns as letters. One route then uses these letter identities to activate phonological units, which are assembled into a pronunciation. A second route uses letter identities to activate the orthographic forms of words, which in turn activate the phonological forms of words. A third route also activates the orthographic forms of words, which in turn activate meanings of words, and these then lead to the activation of the phonological forms of words. In all cases, the resulting phonological sequence is sent to the processor labeled ''output phonological

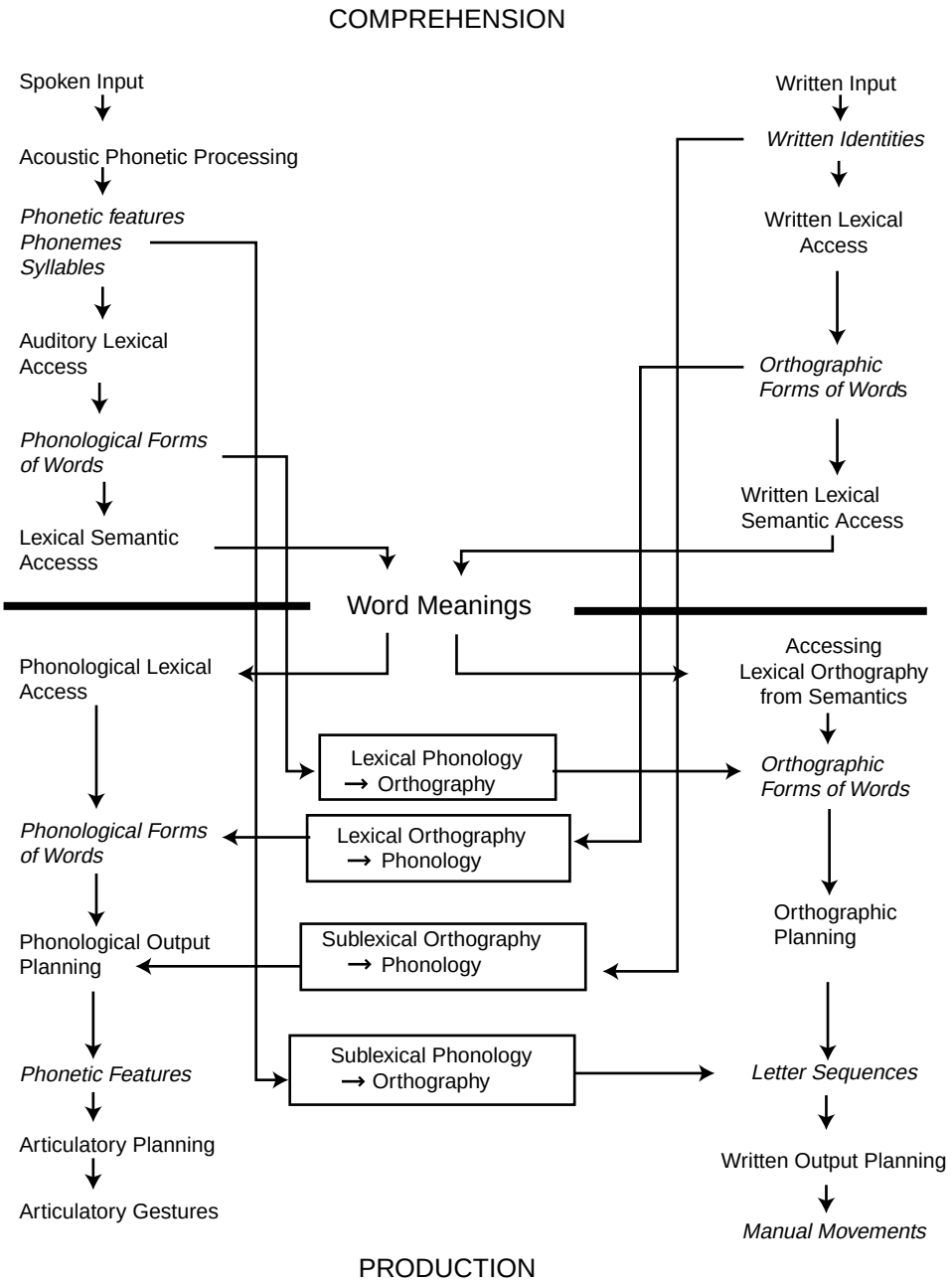


Figure 1 A model of the language processing system for single words. Processing components, indicated in boxes, activate linguistic representations in the sequences indicated by arrows. [from Caplan (1992). *Language, Structure, Processing, and Disorders*. Reproduced with permission of MIT press].

planning" to be turned into a form appropriate to activate motor speech activity. These three routes are thought to become active simultaneously. The processors used in reading aloud can be used in other language tasks. For instance, the processors involved in activating the phonological forms of words from their meanings and in output phonological planning are also used in naming pictures.

Functional communication involving the language code occurs when people use these processors to

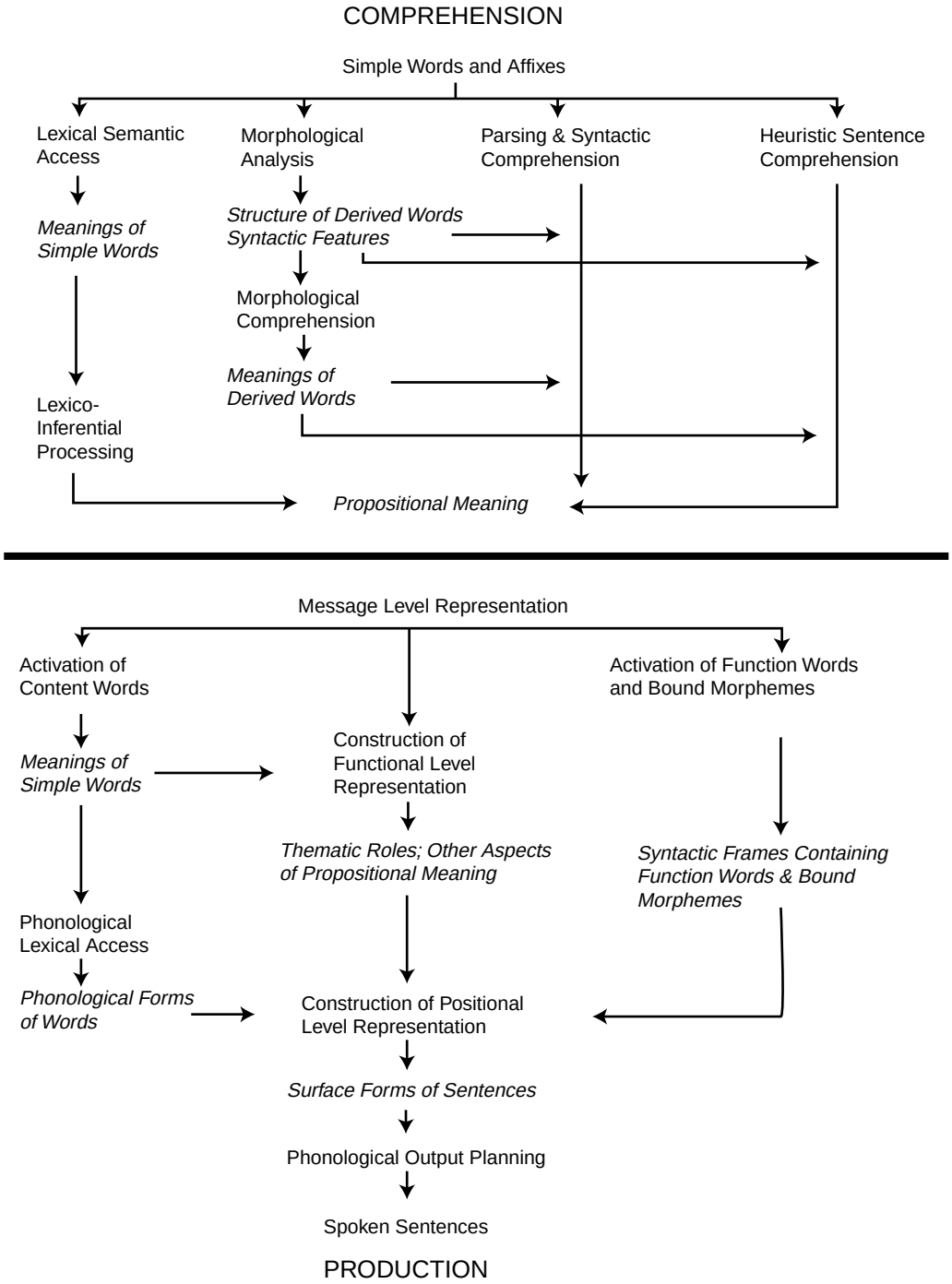


Figure 2 A model of the language processing system for derived words and sentences. Processing components, activate linguistic representations in the sequences indicated by arrows. [from Caplan (1992). *Language, Structure, Processing, and Disorders*. Reproduced with permission of MIT press].

undertake language-related tasks to accomplish specific goals—to inform others, to ask for information, to get things done, etc. It was previously noted that language is subject to control from other cognitive domains; conversely, one of the most important functions of language is to operate as a control mechanism for both intra- and interpersonal thought and action.

III. OVERVIEW OF NEURAL STRUCTURES SUPPORTING LANGUAGE

There is evidence that language processing involves the perisylvian association cortex—the pars triangularis and opercularis of the inferior frontal gyrus [Brodman's areas (BA) 45 and 44 (Broca's area)], the angular gyrus (BA39), the supramarginal gyrus (BA40), and the superior temporal gyrus (BA22: Wernicke's area) —in the dominant hemisphere. Data regarding the functional neuroanatomy of language processing were originally derived from deficit–lesion correlations and, recently, have been obtained from functional neuroimaging and electrophysiological studies in normal subjects. All these sources of data indicate that the perisylvian association cortex is involved in this function.

Patients with lesions in parts of this cortex have been described who have had long-lasting impairments of this function. Disorders affecting language processing after perisylvian lesions have been described in all languages that have been studied, in patients of all ages, with written and spoken input, and after a variety of lesion types, indicating that this cortical region is involved in syntactic processing, independent of these factors. Functional neuroimaging studies have documented increases in regional cerebral blood flow (rCBF) using positron emission tomography (PET) or blood oxygenation level-dependent (BOLD) signal using functional magnetic resonance imaging (fMRI) in tasks in associated with language processing. Event-related potentials (ERPs) whose sources are likely to be in this region have been described in relationship to a variety of language processing operations. Stimulation of this cortex by direct application electrical current during neurosurgical procedures interrupts language processing. From these data, it can be concluded that language processing is carried out in the dominant perisylvian cortex.

Regions outside the perisylvian association cortex might also support language processing. Working outwards from the perisylvian region, there is evidence that the modality of language use affects the location of the neural tissue that supports language, with written language involving cortex closer to the visual areas of the brain and sign language involving brain regions closer to those involved in movements of the hands than movements of the oral cavity and its contents. Some ERP components related to processing improbable or ill-formed language are maximal over high parietal and central scalp electrodes. This may suggest that these regions are involved in language processing, but two factors have to be considered before such a conclusion is drawn: The location in the brain of the tissue that generates an ERP wave is not easy to identify on the basis of the scalp location of that wave and may not be right below that wave, and some of these waves may reflect general processes related to detection of pragmatically implausible or unlikely events in general, not language processing per se.

Both lesion studies in stroke patients and functional neuroimaging studies suggest that the anterior temporal lobe, primarily in the dominant hemisphere, is involved in aspects of language processing. The leading candidate for such a function is the accessing of the sounds of words from their meanings in speech production. However, electrocortical stimulation studies and the effects of neurosurgical resections do not support this conclusion. On the other hand, both functional neuroimaging and electrocortical stimulation studies indicate that the inferior temporal lobe is involved in aspects of word processing, particularly the representation of meanings of nouns. Activation studies implicate the frontal lobe just in front of Broca's area in word meaning as well, although these activations may reflect switching sets rather than processing semantic representations. Injury to the supplementary motor cortex along the medial surface of the frontal lobe can lead to speech initiation disturbances; this region may be important in activating the language processing system, at least in production tasks. Activation studies have shown increased rCBF and BOLD signal in the cingulate gyrus in association with many language tasks. This activation, however, appears to be nonspecific because it occurs in many other, nonlinguistic, tasks as well. It has been suggested that it is due to increased arousal and deployment of attention associated with more complex tasks. The cerebellum also has increased rCBF in some activation studies involving both language and other cognitive functions. This may be a result of the role of this part of the brain in processes involved in timing and temporal ordering of events or because it is involved in many cognitive functions.

Subcortical structures may also be involved in language processing. Several studies report aphasic disturbances following strokes in the deep gray matter nuclei (the caudate, putamen, and parts of the thalamus), but studies of other diseases affecting the same nuclei fail to show significant language impairments. For instance, aphasias follow some caudate strokes, but language disorders are minimal in patients with Huntington's disease, even at the stage of the illness at which memory impairments are readily

documented. It has been suggested that subcortical structures involved in laying down procedural memories for motor functions, particularly the basal ganglia, are involved in "rule-based" processing in language, such as regular aspects of word formation, as opposed to the long-term maintenance of information in memory, as occurs with simple words and irregularly formed words.

Some abnormal language behaviors seen after deep gray matter lesions probably reflect the effects of disturbances in other cognitive functions on language. An example of this is the fluctuation between neologistic jargon and virtual mutism seen after some thalamic lesions. This corresponds to a more general fluctuation between states of delirium and near akinetic mutism, and it most likely reflects the effects of some thalamic lesions on arousal, alerting, and motivational functions, some of which are seen in the sphere of language. Intraoperative stimulation studies of the interference with language functions following dominant thalamic stimulation also suggest that the language impairments seen in at least some thalamic cases are due to disturbances of attentional mechanisms. Perhaps the most important consideration regarding language disorders following subcortical lesions is the question of whether they result from altered physiological activity in the overlying cortex and not from disorders of the subcortical structures. In general, subcortical lesions cause language impairments when the overlying cortex is abnormal (often, the abnormality can only be seen with metabolic scanning techniques), and the degree of language impairment is much better correlated with measures of cortical rather than subcortical hypometabolism. It may be that subcortical structures serve to activate the language processing system but do not process language.

The other major component of the subcortical region of the cerebral hemispheres is the white matter. White matter tracts transmit representations from one area to another. Lesions of white matter tracts disconnect regions of the brain from others and make the operations performed in one region unavailable to others. This can cause language disorders. The best known such disturbance is a pure alexia, in which a patient can write but not read–not even his or her own writing. This can result from a lesion that destroys the primary visual cortex in the dominant hemisphere and extends forward in the white matter to cut off visual information coming to the nondominant hemisphere from the dominant hemisphere language area. In addition to these "disconnection" syndromes, language disturbances of all sorts occur with lesions affecting many white matter tracts, whereas sparing of language functions can follow lesions in identical subcortical areas. The fact that multiple language processing disturbances occur following subcortical strokes affecting white matter is consistent with the existence of many information transfers carried out by white matter fibers, suggesting that many of the areas of cortex and/or subcortical nuclei that carry out sequential language processing operations are not contiguous.

In summary, a large number of brain regions are involved in representing and processing language. Ultimately, they all interact with one another as well as with other brain areas involved in using the products of language processing to accomplish tasks. In this sense, all these regions are part of a "neural system," but this concept should not obscure the fact that many of these regions appear to compute specific linguistic representations in particular tasks. The most important of these regions is the dominant perisylvian cortex.

IV. ORGANIZATION OF THE PERISYLVIAN CORTEX FOR LANGUAGE PROCESSING

Two general classes of theories of the relationship of portions of the perisylvian association cortex to components of the language processing system have been developed—one based on "holist" or distributed views of neural function and one based on localizationist principles. Although theories within each of these two major groupings vary, there are a number of features common to theories within each class.

The basic tenet of holist/distributed theories of the functional neuroanatomy for language is that linguistic representations are distributed widely and that language processing conmponents rely on broad areas of perisylvian association cortex. Karl Lashley identified two functional features of holist/distributed models that determine the effects of lesions on performance: equipotentiality (every portion of a particular brain region can carry out a specific function in every individual) and mass action (the larger the neuronal pool that carries out a particular function, the more efficiently that function is accomplished). The features of equipotentiality and mass action jointly entail that lesions of similar sizes anywhere in a specified brain region have equivalent effects on function, and that the magnitude of any functional deficit is directly proportional to the size of a lesion in

this specified area. Recently, models of lesions in parallel distributed processing simulations of language and other cognitive functions have provided a mathematical basis for these properties of these systems.

All the traditional theories that postulate localization of components of the language processing system maintain the view that, discounting lateralization, the localization of components of the language processing system is invariant across the normal adult population. Thus, all the traditional localizationist theories have as a corollary that lesions in particular areas of the perisylvian association cortex interrupt the same language processing components in all individuals. Many localizationist theories also maintain that the specific localization of language processing components results from a computational advantage inherent in juxtaposing particular language processing components to each other or to cortex supporting arousal, sensory, and motor processes. Modern localizationist models relax these assertions, incorporating the ideas of individual variability and some degree of specialization wthin large-scale neural nets.

Because of the plethora of specific theories within each of these two general camps, it is impossible to critically review the empirical basis of all theories that have present-day adherents. I focus on the most widely cited theories as examples of each class.

A. Holist Theories

Unlike narrow localizationist theories, there is no one holist model that has emerged as the major example of this class of theories. However, several lines of evidence are adduced as evidence for holist theories, and all holist theories suffer from similar inadequacies in accounting for certain empirical findings.

The first line of evidence supporting holist theories consists of the ubiquity of general factors in accounting for the performance of aphasic patients. For instance, factor analyses of the performances of groups of patients both on general aphasia tests and on tests of specific language abilities almost always result in first eigenvectors (usually accounting for more than half of the variance in performance) that are approximately equally weighted for most of the subtests used to test the population. Such vectors are usually taken to reflect disruption of a single factor that affects performance on all measures, such as a limited amount of mental resources available for psycholinguistic computations. The existence of such factors would be the immediate consequence of a system in which functions were disruptable by lesions in a variety of locations, and they have therefore been widely taken as evidence for a distributed basis for language functions. A second finding supporting holist theories is the frequent observation of so-called ''graceful degradation'' of performance within specific language domains after brain damage. An example of such degradation is the strong tendency of certain dyslexic patients to read irregularly spelled words according to a regularization strategy (e.g., ''pint'' is read with a short ''i''), a tendency that is inversely proportional to the frequency of the word. Graceful degradation reflects the preservation of the simplest (in many cases, the most commonly occurring) aspects of language processing after brain damage. Modern work with parallel distributed processing models, which provide formal models of holist concepts, indicates that such patterns of performance can arise following focal lesions in systems in which information is represented and processed in massively parallel, distributed forms. A third source of empirical support for holist theories comes from the finding of an effect of lesion size on the overall severity of functional impairments in several language spheres. This would follow from the principle of mass action. Therefore, these results therefore are consistent with some form of holism in the neural basis for linguistic representations and processes.

Against the complete adequacy of any holist model is the finding that multiple individual language deficits arise in patients with small perisylvian lesions, often in complementary functional spheres. For instance, studies of acquired dyslexia have documented patients who cannot read by a whole-word route (i.e., by using the entire form of a written word to gain access to the mental representation of that word) and others who cannot read by the application of spelling–sound correspondences at the letter and grapheme level. The existence of these isolated complementary deficits in different single cases indicates that at least one abnormal performance cannot result from the relative complexity of processing required by one of these tasks. Double dissociations of this sort abound in the contemporary psycholinguistic aphasiological literature. They indicate that the mode of organization of language in the brain must be one that allows focal lesions to disrupt specific aspects of psycholinguistic processing, not simply a mode of organization that produces complexity effects and degrades gracefully. Although some selective disruptions of function can occur when ''lesions'' are produced in simulated language processing systems that operate in parallel

and distributed fashion, to date no mechanism of lesioning a distributed neural system has been shown to produce the range of specific patterns of language breakdown observed in patients.

B. Localizationist Theories

Although many localizationist models exist, the connectionist model of language representation and processing in the brain, revived by Norman Geschwind and colleagues in the 1960s and 1970s, is probably the best known localizationist model of the functional neuroanatomy of language, at least in medical circles in North America. This model is based on observations of aphasic patients and the interpretation of those observations that were first made more than a century ago.

Figure 3 represents the basic connectionist model of auditory-oral language processing and its relation to areas within the dominant perisylvian cortex. This model postulates three basic "centers" for language processing, all in cerebral cortex. The first (Fig. 3A), located in Wernicke's area, stores the permanent representations for the sounds of words (what psycholinguists would now call a "phonological lexicon"). The second (Fig. 3M), located in Broca's area, houses the mechanisms responsible for planning and programming speech. The third (Fig. 3C), diffusely localized in cortex in the 19th-century models, stores the representations of concepts. A major innovation proposed by Geschwind is in the location of one aspect of the concept center. Geschwind proposed that the inferior parietal lobule—the supramarginal and angular gyri—is the location at which the fibers projecting from somesthetic, visual, and auditory association cortices all converge and that as a consequence of this convergence, associations between word sounds and the sensory properties of objects can be established in this area. Geschwind argued that these associations are critical aspects of the meanings of words and that their establishment is a prerequisite of the ability to name objects.

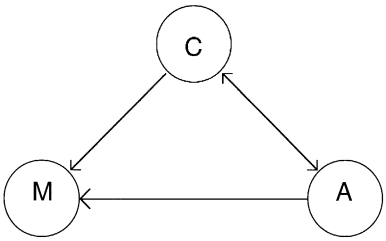


Figure 3 The classical connectionist model. A represents the auditory center for the long-term storage of word sounds. M represents the motor center for speech planning, and C represents the concept center. Information flow is indicated by arrows. The location of these centers in the brain is described in the text.

Language processing in this model involves the activation of linguistic representations in these cortical centers and the transfer of these representations from one center to another, largely via white matter tracts. For instance, in auditory comprehension, the representations of the sound patterns of words are accessed in Wernicke's area following auditory presentation of language stimuli. These auditory representations of the sounds of words in turn evoke the concepts associated with words in the "concept center." Accessing the phonological representation of words and the subsequent concepts associated with these representations constitutes the function of comprehension of auditory language. In spoken language production, concepts access the phonological representations of words in Wernicke's area, which are then transmitted to the motor programming areas for speech in Broca's area. In most versions of this model, the proper execution of the speech act also depends on Broca's area receiving input directly from the concept center. Repetition, reading, and writing are modeled as involving similar sequences of activation of centers via connections.

Recently, these aphasic syndromes have been related to the brain using a series of neuroimaging techniques—first T^{99} scanning and then computed tomography MRI, and PET. All have confirmed the relationship of the major syndromes to lesion locations. Broca's aphasia is associated with anterior lesions; Wernicke's aphasia is associated with posterior lesions, centered in the temporal–parietal juncture; pure motor deficits of speech are associated with subcortical lesions; pure word deafness is associated with lesions in the auditory association areas and surrounding white matter tracts, often bilaterally; transcortical motor and transcortical sensory aphasia are associated with watershed infarcts between the anterior and middle cerebral arteries and middle and posterior cerebral arteries; and conduction aphasia is associated with smaller lesions that often appear to affect the arcuate fasciculus. However, despite these general correlations, the classical aphasic syndromes are not as well correlated with lesion sites as the theory claims they should be. Virtually all studies exclude many types of lesions, such as various sorts of tumors, degenerative diseases, and others.

The classical syndromes are best related to lesion sites in cases of rapidly developing lesions, such as stroke. Even in these types of lesions, the syndromes are never applied to acute and subacute phases of the illness and, in the chronic phase of diseases such as stroke, between 15 and 40% of patients have lesions that are not predictable from their syndromes.

Lesion-deficit correlations have been studied in patients with more specific functional impairments than are captured by the classic aphasic syndromes (e.g., semantic memory, whole-word writing and the conversion of sounds to their corresponding orthographic units, short-term memory, and word comprehension). For the most part, these studies have involved relatively small numbers of subjects because of the difficulty in obtaining large numbers of subjects with specific deficits. For instance, one study of disorders affecting semantic memory in stroke patients could only identify three patients with a selective deficit in this function from the many patients that were screened; many other patients had problems in naming objects or in matching spoken words to pictures but not with semantic memory.

These more focused studies have provided evidence for localization of function, but the picture that emerges is complex. The localizations found have often not been consistent with the classical connectionist model. For instance, semantic memory (word meaning) appears to be disrupted after temporal, not inferior parietal, damage, and auditory–verbal short-term memory appears to be disrupted after parietal, not temporal, lesions.

An important aspect of the database relating specific language functional deficits to lesions is the finding that individual components of the language processing system can be either affected or spared following lesions in particular parts of the perisylvian association cortex. This variability in the effects of lesions at particular sites ranges across all areas of the perisylvian association cortex and is true of components of the language processing system responsible for activating any of the linguistic representations described in Section I (i.e., lexical, morphological, and sentential representations). For instance, most studies have reported phoneme discrimination deficits after both anterior and posterior lesions, in both real words and in computer-synthesized stimuli. At the same time, some studies have found that most lesions producing these impairments occur with posterior lesions. Similar individual variability has been documented in the localization of the lesions responsible for the comprehension of single words, the production of morphological forms, aspects of sentence comprehension, the production of function words in sentence production, and the production of phonemic errors in spontaneous speech, picture naming, repetition, and reading, in some cases with "central tendencies" toward deficits following lesions in specific locations. This pattern has led some rsearchers to view the entire perisylvian cortex as a neural net that supports all aspects of language processing, with some degree of specialization of parts of this net for specific functions. Models of this type cannot account for severe impairments of a single langauge operation following small lesions in different parts of this region in different individuals, however.

The finding of variability in the effects of lesions on language functions also emerges from electrocortical stimulation studies, which have studied phoneme discrimination, picture naming, sentence comprehension, and other language functions. Each of these language tasks was most likely to be interrupted with stimulation in particular regions of the perisylvian cortex across all patients, but considerable variation in sites associated with interruption of each task was also noted. For instance, phoneme discrimination was interrupted by stimulation throughout the entire central region of the perisylvian association cortex in approximately 80% of cases and in sites within the language zone that are further removed from the Sylvian fissure (such as the more dorsal regions of the inferior parietal lobule) in the remaining 20% of cases.

This variability suggests that language operations are localized in small parts of the perisylvian association cortex, but that the areas in which they are localized are different in different people. The fact that lesions restricted to specific parts of the perisylvian cortex are each associated with all levels of performance of a language processing operation, from normal performance to the worst performances seen in aphasic patients, implies that for some individuals a lesion in each of these lobar regions does not affect the operation at all, whereas for others some or all of the operation is impaired by a lesion in the same region. It is difficult to understand how this could result from anything other than individual variability in the premorbid location of these operations.

On the other hand, to the extent that specific lesions tend to be associated with certain deficits, as may be the case for some impairments, these studies also support the view that there are central tendencies in the localization of these language processing components within the perisylvian cortex. A reasonable conjecture is that the functional neuroanatomy of language in the

perisylvian association cortex has three features: (i) localization of language functions in individuals, (ii) tendencies for functions to be localized in specific portions of the perisylvian association cortex, and (iii) at least 20% of normal adults (and often many more) showing significant deviations from these central localizationist tendencies for each language processing component.

In the past 10 years, a considerable number of studies of the neural correlates of language processing have been published that employed PET, fMRI, and other "activation" techniques. As noted previously, some of these studies have led to the appreciation of possible roles for brain regions outside the perisylvian cortex in carrying out language operations. These studies also provide evidence about the organization of the perisylvian cortex for language. Many studies claim to find evidence for the localization of particular language representations, such as the phonological forms of words used in speech recognition, or language operations, such as transforming letters into their corresponding sounds, in particular parts of this cortex. Some of these localizations depend on specific tasks; for instance, the exact part of the perisylvian cortex in which blood flow changes during tasks that require comparison of sequences of phonemes has varied considerably in different studies in which different tasks were used. In other cases, there is more consensus about the areas that increase their vascular responses as a function of particular operations. Verbal rehearsal, which appears to involve Broca's area and adjacent parts of the frontal lobe, and maintenance of phonological representations in verbal short-term memory, which seems to activate the inferior parietal lobe, are examples of such functions; however, even here, different studies show some variation in the areas activated. Very few studies have examined patterns of vascular responsivity in individual subjects, which will be necessary to address the issue of variability. Much more work using these techniques can be expected, with consequent increases in the database relevant to our understanding of these issues.

V. LATERALIZATION

Most language processing occurs in one hemisphere called the "dominant" hemisphere. Which hemisphere is dominant shows considerable individual differences, which bear a systematic relationship to handedness. In approximately 98% of right-handed individuals, the left hemisphere is dominant, with the extent to which left hemisphere lesions cause language disorders influenced by the degree to which an individual is right-handed and the number of non-right-handeders in his or her family. Approximately 60–65% of the non-right-handed individuals are left hemisphere dominant, approximately 15–20% are right hemisphere dominant, and the remainder appear to use both hemispheres for language processing. The relationship of dominance for language to handedness suggests a common determination of both, probably in large part genetic (although the nature of the genetic effect remains unclear).

Although language was the first function known to be lateralized and is still the best example of a lateralized function, it is not completely lateralized. Although not as important in language functioning as the dominant hemisphere, the nondominant hemisphere is involved in many language operations. Evidence from the effects of lesions and split-brain studies, experiments using presentation of stimuli to one or the other hemisphere in normal subjects, and activation studies indicates that the nondominant hemisphere understands many words, especially concrete nouns, and suggests that it is involved in other aspects of language processing as well, such as syntactic processing. Some language operations may be carried out primarily in the right hemisphere. The best candidates for these operations are ones that pertain to processing the discourse level of language, interpreting nonliteral language such as metaphors, and appreciating the tone of a discourse as is manifest in, for instance, its being humorous. Some scientists have developed models of the sorts of processing that the right hemisphere carries out. For instance, it has been suggested that the right hemisphere codes information in a course way compared to the left hemisphere. This and other suggestions provide the bases for ongoing research programs on the nature of language processing in the right hemisphere.

Lateralization of language functions can be seen as a broad form of localization—the localization of a function in one of the two cerebral hemispheres. As noted previously, lateralization varies as a function of handedness and even within populations with similar handedness profiles. These facts suggest intriguing similarities between the phenomena of localization and lateralization of language. In both localization and lateralization, the location of a particular language processing component varies across the adult population as a whole. In both cases, however, there are

central tendencies with respect to the location of particular language processing components: There appear to be preferred sites for particular language processing functions within the perislyvian region, and there is a strong preference for language processing components to be left hemisphere based. These patterns would result from any area of either perisylvian association cortex being capable of supporting any subcomponent of the language processing system at the initial stage of language development, and from different areas of cortex assuming particular language processing roles as a function of intrinsic, genetically determined, developmental patterns, modified by other factors such as the internal organic milieu and the nature of exposure to language.

VI. CONCLUSION

Our newly acquired abilities to identify specific language processing deficits in patients, to characterize lesions with modern imaging techniques, and to use technologies such as intraoperative electrocortical stimulation, event-related potentials, and metabolic scanning to study the neural basis for language position us to investigate the neural basis for language at a level of detail previously unattainable. Research using these techniques is likely to continue to change our ideas of the way the human brain supports language functions.

See Also the Following Articles

APHASIA • BROCA'S AREA • INFORMATION PROCESSING • LANGUAGE ACQUISITION • LANGUAGE AND LEXICAL PROCESSING • LANGUAGE DISORDERS • LATERALITY • READING DISORDERS, DEVELOPMENTAL • SPEECH • WERNICKE'S AREA

Suggested Reading

Caplan, D. (1987). *Neurolinguistics and Linguistic Aphasiology*. Cambridge Univ. Press, Cambridge, UK.

Caplan, D. (1992). *Language, Structure, Processing, and Disorders*. MIT Press, Cambridge, MA.

Caplan, D. (1994). Language and the brain. In *Handbook of Psycholinguistics* (Gernsbacher, Ed.), pp. 1023–1074. Academic Press, New York.

Damasio, A. R., and Damasio, H. (1992). Brain and language. *Sci. Am.* **267,** 88–95.

Dronkers, N., Pinker, S., and Damasio, A. R. (2000). Language and the aphasias. In *Principles of Neural Science* (E. R. Kandell, J. H. Schwartz, and T. W. Jessell, Eds.), Appleton & Lange, Norwalk, CT.

Geschwind, N. (1979). Specializations of the human brain. *Sci. Am.* **170,** 940–944.

Laterality

JOSEPH B. HELLIGE
University of Southern California

GLOSSARY

corpus callosum Largest fiber tract that connects the left and right cerebral hemispheres.

handedness The tendency to prefer the use of one hand over the other and for motor performance to be better for the preferred hand.

hemispheric asymmetry Biological and functional differences between the left and right sides of the cerebral cortex.

hemisphericity Indication of the extent to which an individual is more reliant on the left or on the right cerebral hemisphere, inferred from measurement of the individual's cognitive skills and biases.

laterality Behavioral and biological manifestations of left–right brain differences or of hemispheric asymmetry.

manual praxis Purposeful, sequential actions in which spatial constraints imposed by the environment are minimal.

planum temporale Cortical area located on the superior surface of the temporal lobe posterior to the transverse auditory gyrus of Heschl.

split-brain patients Individuals whose left and right cerebral hemispheres have been surgically disconnected by severing the corpus callosum and other connecting fibers.

Sylvian fissure Deep groove on the lateral surfaces of the cerebral hemispheres that marks the boundary between the frontal and parietal lobes above and the temporal lobe below.

Laterality refers to the behavioral and biological manifesta-tions of asymmetry between the left and right cerebral hemispheres. The cerebral cortex of the human brain is divided anatomically into two hemispheres, the left and the right. Although the two hemispheres are similar in appearance and structure, they are not biologically identical and they have different information processing abilities and propensities. This article provides an overview of this brain laterality and its consequences for perception, cognition, emotion, and action.

I. LEARNING ABOUT LATERALITY: TECHNIQUES AND TOOLS

Brain laterality has been studied in a variety of populations using several research techniques and tools. In fact, the existence of so many converging techniques is a very positive feature of research in this area. I begin by reviewing the major techniques on which later conclusions rest and, for the sake of example, indicate briefly how each technique provides information about laterality for aspects of human language. The oldest technique for studying brain–behavior relationships is the observation of behavioral deficits after localized brain damage. If the brain were completely symmetric in terms of structure and function, then injury to homologous areas of the two hemispheres should have equivalent effects. However, this is not the case. For example, specific language

deficits are more common and more severe after damage to certain temporal and parietal areas of the left hemisphere than after injury to corresponding areas of the right hemisphere. The deficits for which this is true include speech production and perception, phonetic analysis of printed text, the use of syntactic cues, and access to certain types of word meaning. As discussed later, certain other deficits (e.g., processing intonation cues) are associated more with right hemisphere damage. This is important because such double dissociations rule out the possibility that the left hemisphere is simply dominant for everything. Note that identification of deficits after localized brain damage can indicate the extent to which an area within one hemisphere is necessary for performing a particular task or executing a particular process. However, such studies do not indicate the extent to which one hemisphere is sufficient for performing the task or process or that the impaired process is localized within the damaged area.

Dramatic demonstrations of brain laterality come from the study of so-called split-brain patients whose hemispheres have been surgically disconnected in order to control the spread of epileptic seizures. This is done by cutting the corpus callosum, the largest fiber tract connecting the two hemispheres, as well as other connecting fibers. By using clever techniques of stimulus presentation and response measurement it is possible to examine the positive competence of each hemisphere in split-brain patients and thereby identify tasks and processes for which each isolated hemisphere is sufficient. These techniques take advantage of the fact that human sensory projections are organized so that input from one side of the body or one side of space is transmitted directly to the contralateral cerebral hemisphere. This includes sensory input from fingers of the left and right hands, the left and right halves of each retina (so that each visual half field projects directly to the contralateral visual cortex) and, under appropriate conditions, the left and right ears. With respect to certain aspects of language, the study of split-brain patients indicates that the isolated left hemisphere is far more competent than the isolated right hemisphere. For example, a split-brain patient can say the name of a common object (e.g., pencil) when it is placed into the right hand or when its picture or printed name is projected to the right visual field (i.e., to the left hemisphere) but not when it is placed into the left hand or when its picture or printed name is projected to the left visual field (i.e., to the right hemisphere). Of course, generalizing from such patients to neurologically intact individuals is tricky, in part because brain laterality may not operate the same way in the presence of extensive interhemispheric connections.

In neurologically intact individuals, it is not possible to test the competence of each hemisphere in isolation, but it is possible to measure speed and accuracy of performance as a function of which hemisphere receives information directly and must at least initiate processing. The techniques for studying this sort of behavioral laterality in neurologically intact individuals are the same as those used to lateralize stimuli for split-brain patients. When used with appropriate methodological care, such studies provide an important converging technique for learning about functional hemispheric asymmetry in the intact brain. Although often smaller, laterality effects in intact individuals are typically in the same direction as those found in split-brain patients. For example, there is a right visual field (left hemisphere) advantage for the identification of printed words and nonwords as well as a right ear (left hemisphere) advantage for the identification of spoken words and syllables.

In recent years there have been significant advances in brain imaging techniques that permit measurement of various structures in the living brain and of the relative amounts of neural activity in different cortical regions as individuals perform experimental tasks. The techniques include computerized axial tomography scans, magnetic resonance imaging (MRI) and functional MRI, measures of regional cerebral blood flow, positron emission tomography scans, event-related potentials, and magnetoencephalography. Among other things, brain imaging has provided a great deal of converging information about brain laterality. Later, I discuss certain well-established structural asymmetries as well as individual variation in the size of the corpus callosum. With respect to language, functional imaging techniques have shown greater activation in temporoparietal areas of the left hemisphere than in the right hemisphere when normal individuals speak, identify printed words, and so forth, with the specific areas of greatest activation depending on exactly which processes are relevant to the task. Furthermore, monitoring a list of words for semantic content also produces activation of frontal areas within the left hemisphere. In addition, verbal working memory tasks produce primarily left hemisphere activation, whereas spatial working memory tasks produce primarily right hemisphere activation. Of course, increased activity in a specific brain area does not indicate exactly what that area is doing or how well it is doing it. Nor is it possible to determine whether

increased activation reflects excitatory or inhibitory neural processes. Nevertheless, functional imaging techniques have produced important results that are consistent with conclusions suggested by classical neuropsychology and clearly constitute a converging operation that will grow in importance.

II. BEHAVIORAL ASYMMETRIES IN HUMANS: LATERALITY FOR COMPONENTS OF ACTION AND COGNITION

The tools and techniques described in the preceding section have been used to identify and confirm a great many behavioral asymmetries in humans. Before summarizing a number of the most well-established asymmetries, it is useful to consider three general characteristics of laterality effects: ubiquity, subtlety, and complementarity. Ubiquity refers to the fact that the two cerebral hemispheres have different levels of ability and different processing propensities in a great many domains. In this section, I review examples in the domains of motor activity, language, perception, and emotion. Subtlety refers to the fact that it is rarely the case that one hemisphere can perform a task or accomplish a specific process quite well, whereas the other hemisphere cannot perform the task or process at all. Instead, both hemispheres typically have some ability, though they may go about a task in different ways and one hemisphere may do a better job than the other. A notable exception to this is speech production, which tends to be controlled exclusively by a single hemisphere (usually the left). Complementarity refers to the fact that in many domains, the roles for which each hemisphere is dominant can be described as complementary. Consequently, both hemispheres normally play a role in virtually all complex activities, such as understanding language or identifying faces, with their contributions fitting together like two pieces of a puzzle. Several examples of complementarity are given in the remainder of this article.

The most obvious behavioral asymmetry in humans is handedness, with approximately 92% of women and 88% of men favoring and being more proficient with the right hand for performing a variety of skilled activities, such as writing, drawing, eating, and using a needle to sew. Of the remaining non-right-handed individuals a few exhibit strong and consistent left-handedness, a few are truly ambidextrous, and others show hand preferences that vary from one skilled activity to another. In general, for both right-handed and left-handed individuals, hand differences are weaker for unskilled activities such as picking up a small object. Furthermore, for tasks that require the coordinated activity of both hands, their roles are often complementary. In general, for right-handed individuals the left hand (controlled by the right hemisphere) performs movements of relatively low spatial and temporal frequency, whereas the right hand (controlled by the left hemisphere) performs movements of relatively high spatial and temporal frequency. An example of this complimentary arrangement is handwriting, during which the left hand arranges and steadies the paper while the right hand makes more frequent and smaller movements with the writing instrument. Though handedness is the most obvious example, there are also other motoric asymmetries. For example, for right-handed individuals the left side of the body is frequently preferred for postural support. In addition, the right side of the face (controlled by the left hemisphere) is superior for making certain oral movements associated with language and other precisely sequenced activities, whereas the left side of the face (controlled by the right hemisphere) is more emotionally expressive.

Left hemisphere dominance for many aspects of language is the most obvious and cited asymmetry outside of the motor domain. From clinical neurological data as well as other sources, it is estimated that speech production is limited to the left hemisphere in approximately 95% of right-handed individuals. As noted in the preceding discussion of research tools and techniques, the left hemisphere is also dominant for many aspects of language perception and for the verbal processing of stimulus material, although in these cases left hemisphere superiority is more a matter of degree than of the all-or-none asymmetry that is characteristic of speech production. In addition, when we consider understanding language for the purpose of communication, there is growing evidence that both hemispheres make important contributions. Whereas the left hemisphere is dominant for the perception of phonetic information, for the use of syntax and for certain aspects of semantic processing, the right hemisphere is dominant for processing the sort of intonation cues and prosody that communicate such things as emotional tone (e.g., anger versus surprise). The right hemisphere is also involved in processing narrative-level linguistic information. Some of these complementary, language-related asymmetries may be related to hemispheric differences in the efficiency of processing different aspects of acoustic signals. For example, identification of many spoken phonemes requires

efficient processing of rapid changes in the acoustic signal over brief periods of time, a type of processing for which the left hemisphere is hypothesized to be superior. In contrast, identification of the emotional tone of voice requires efficient processing of much slower modulations of the acoustic signal over longer periods of time, a type of processing for which the right hemisphere is hypothesized to be superior. The two hemispheres also appear to access word meanings in complementary ways. When a word is presented, the left hemisphere restricts processing very quickly to one possible meaning, usually the dominant meaning or the meaning most consistent with the present context, whereas the right hemisphere maintains activation of multiple meanings and remotely associated words for a more extended period of time.

Neither hemisphere is uniformly superior for processing nonverbal perceptual information. Instead, both hemispheres contribute to perceptual processing and do so in ways that could be described as complementary. For example, in studies of the identification of visual patterns there is evidence of right hemisphere dominance for processing global aspects of stimuli (e.g., the outer contour of a face) and left hemisphere dominance for processing local details (e.g., small features of a face). Specifically, global and local processing are associated with the posterior superior temporal areas of the right and left hemispheres, respectively (though this may be affected by the relative sizes and perceptual clarity of the global and local levels within a stimulus). These effects may be related to hemispheric differences in the efficient use of information carried by visual channels tuned to relatively high versus relatively low spatial frequencies. Specifically, the left and right hemispheres are biased toward more efficient use of higher and lower frequencies, respectively, with at least three aspects of spatial frequency being relevant: the absolute range of frequencies contained in a stimulus, the range of frequencies that is most relevant for the task being performed, and whether the relevant frequencies are high or low relative to other frequencies contained in the stimuli used in the experiment. It has been hypothesized that analogous hemispheric differences extend to the processing of relatively high and relatively low temporal frequencies in audition and, perhaps, to movements of different temporal frequencies and spatial extent.

In the spatial domain, the brain computes at least two kinds of spatial relation representations—a categorical representation used to assign a spatial relation to a category such as ''connected to'' or ''above'' and a coordinate representation used to represent precise distances and locations. The right hemisphere makes more effective use of this latter coordinate system, whereas there is either no hemispheric difference or a left hemisphere advantage for processing categorical spatial relationships. Neural network simulations suggest that hemispheric asymmetry for making categorical versus coordinate judgments may be related to the nature of visual information that is most useful for processing categorical versus coordinate properties. For example, networks constructed to simulate relatively large overlapping receptive fields compute coordinate spatial information better than do networks constructed to simulate relatively small, nonoverlapping receptive fields. Exactly the reverse has been found for the computation of categorical spatial information. From this perspective, it is interesting that categorical spatial processing is disrupted by manipulations (blurring of stimuli) that selectively interfere with information carried by channels with small, discrete receptive fields, whereas coordinate spatial processing is disrupted by manipulations (use of a diffuse red background) that selectively interfere with information carried by channels with large, overlapping receptive fields. Additional neural network simulations have examined the possible importance of receptive field sizes for encoding information about shape. The same networks that favored coordinate spatial processing also favored coding the identify of specific shapes, whereas the same networks that favored categorical spatial processing also favored the assignment of shapes to categories. Interestingly, there is evidence of right hemisphere superiority for processing the sort of specific shape information that would be needed to distinguish among the exemplars of a single category but left hemisphere superiority for classifying the prototypes used to define different categories. Of course, it is possible to devise alternative ways of tying these different laterality effects together (e.g., in terms of left and right hemisphere attentional biases toward different ranges of spatial frequency) and more empirical as well as theoretical work is needed before we fully understand the mechanisms that underlie these complementary hemispheric specializations.

As noted earlier, the right hemisphere is superior to the left in using the intonation cues of speech to identify emotional tone of voice. The same tends to be true of identifying the emotion displayed on a face, though in both cases it is difficult to rule out interpretations in terms of the kind of auditory or visual information that is most useful for identifying

emotion. With respect to the production or experience of emotions, hemispheric differences seem to be more complementary. For example, the balance of activation between the frontal lobes of the two hemispheres is related to the valence of an experienced emotion. Specifically, positive emotions are accompanied by relatively greater left hemisphere activation and negative emotions are accompanied by relatively greater right hemisphere activation. Emotions of negative versus positive valence (e.g., sadness versus happiness) are typically characterized by states of low versus high arousal, respectively, leading some researchers to suggest that it is the difference in arousal, rather than valence per se, that accounts for the observed hemispheric asymmetries.

These representative examples of behavioral asymmetries in humans are admittedly illustrative rather than exhaustive. Nevertheless, they illustrate that laterality is a pervasive characteristic of human behavior, that computational differences between the hemispheres are often subtle, and that each hemisphere makes important contributions to most aspects of behavior.

III. BIOLOGICAL ASYMMETRIES IN THE HUMAN BRAIN

At first glance, the left and right hemispheres appear to be biologically identical, leading one to wonder why there are so many functional asymmetries. However, postmortem studies of anatomy and structural imaging studies of the living brain have documented a number of consistent physical asymmetries. For example, in the majority of human brains the frontal region is wider and extends farther forward in the right hemisphere and the occipital region is wider and extends farther rearward in the left hemisphere, giving the brain a kind of counterclockwise torque. Because the temporoparietal areas of the left hemisphere are important for language, it may not be surprising that a number of anatomical and cytoarchitectonic hemispheric differences have been found in those areas. For example, consider the Sylvian fissure, which marks the boundary between the frontal and parietal lobes, which lie above the fissure, and the temporal lobe, which lies below the fissure. This fissure tends to be longer and straighter in the left hemisphere than in the right hemisphere, in which it tends to curl upward. In addition, the planum temporale, which is an extension of Wernicke's area in the left hemisphere (known to be important for language), tends to be larger in the left hemisphere than in the right hemisphere.

There is evidence that individual variation in these structural asymmetries is related to individual variation in functional asymmetry. As noted earlier, handedness is the most obvious motoric expression of laterality. Therefore, it is interesting that the foregoing structural asymmetries are related to handedness. For example, for right-handed individuals the planum temporale is larger in the left hemisphere in approximately 65% of the cases, larger in the right hemisphere in approximately 10% of the cases, and approximately equal in 25% of the cases. The corresponding percentages for non-right-handers are approximately 25, 10, and 65%, respectively. Similar distributions characterize asymmetry of the Sylvian fissure. The ability to measure structure within the normal, living brain has also made it possible to search for relationships between structural asymmetry and a variety of other nonmotoric behavioral and perceptual asymmetries, but no clear-cut picture has emerged. Perhaps this is not surprising in view of the fact that structural characteristics revealed by contemporary imaging techniques are still relatively gross (e.g., the length of a fissure).

Despite the fact that structural characteristics such as length and size are difficult to interpret, some of the structural characteristics that have been observed are related to other aspects of individual variation. For example, the extent of leftward asymmetry of planum temporale size is greater in musicians with perfect pitch than in either nonmusicians or musicians who do not have perfect pitch. Also, although it is impossible to reach strong conclusions from the study of only one brain, there is something intriguing about the fact that neurons in areas that are known to be involved in spatial reasoning and mathematics received greater metabolic support in the brain of the great physicist Albert Einstein than in the brains of other individuals. However, it should be noted that attempts to relate structural asymmetries to other dimensions, such as sex, psychopathology, and cognitive deficit, have produced mixed results.

Though not as well established as the gross structural asymmetries, there are also indications of cytoarchitectonic and biochemical differences between the two hemispheres. For example, the extent of higher order dendritic branching seems to be greater in certain speech areas of the left hemisphere (e.g., Broca's area) than in corresponding regions of the right hemisphere. Conversely, lower order dendritic branches seem to be longer in the right hemisphere than in the left

hemisphere. It has also been hypothesized that the distribution of two important neurotransmitters is asymmetric in the human brain, with dopamine being more prevalent in the left hemisphere and norepinephrine being more prevalent in the right hemisphere.

Structural imaging has also been used to study the correlates of more or less callosal connectivity between the two hemispheres by measuring such things as the midsagittal area of the corpus callosum. Although the size of the corpus callosum increases with the size of the brain, the relative extent of callosal connectivity (e.g., corpus callosum size relative to brain size) seems to decrease with increases in the size of the brain. Furthermore, there is some indication that fewer callosal fibers connect more asymmetric regions of the two hemispheres, although the functional significance of these relationships is not clear. Studies have also begun to examine the relationship between callosal connectivity as measured from brain images and various perceptual asymmetries. One hypothesis is that as the extent of callosal connectivity increases, interhemispheric transfer of information is facilitated, thereby decreasing the size of behavioral laterality effects. Despite a certain intuitive plausibility, the experimental evidence for this hypothesis is mixed, perhaps because increases in callosal connectivity also permit more inhibition from one hemisphere to the other.

IV. UNITY OF PROCESSING FROM THE LATERALIZED BRAIN: VARIETIES OF INTERHEMISPHERIC INTERACTION

One of the most striking and consistent observations in studies that involve functional brain imaging is that many areas within both brain hemispheres are activated by even very simple tasks. This reflects the fact that the behavior of neurologically intact individuals is virtually always the result of processing in both hemispheres as well as in a variety of subcortical structures. In this section, I consider a variety of ways in which the left and right hemispheres interact and the biological mechanisms that support those interactions.

The corpus callosum, with at least 200 million nerve fibers, is the largest fiber tract that connects the two cerebral hemispheres. Comparison of split-brain patients with intact individuals provides a clear indication that the corpus callosum is critical to normal interhemispheric interaction, especially interaction that requires the transfer of information about the identify or name of a stimulus. Although there are no cortical landmarks that divide the corpus callosum, it is generally the case that different regions of the corpus callosum contain fibers originating in different cortical areas. That is, anterior portions of the corpus callosum contain primarily fibers that originate in premotor and frontal regions of the cortex, middle portions contain fibers that originate in motor and somatosensory, regions, and so forth. In addition, many callosal fibers are homotopic; that is, they connect homologous areas of the two hemispheres. An interesting hypothesis is that these fibers produce a type of homotopic inhibition at the computational level, thereby producing mirror-image patterns of activation and inhibition in the two hemispheres, which could be described as a kind of complementarity and that might contribute to the development of hemispheric asymmetry. However, the corpus callosum also contains fibers that originate in a specific region of one hemisphere and terminate in a completely different region of the opposite hemisphere, creating a mechanism whereby neural activity within one hemisphere could have more generalized excitatory or inhibitory effects on neural activity within the other hemisphere.

At a functional level, we have seen that the two hemispheres are often dominant for different task-relevant processing components. In such cases, each hemisphere is likely to take the lead for those components of processing that it handles best. In order for the two hemispheres to coordinate their activities effectively, the results of processing must be integrated across the two hemispheres. At the same time, many hemispheric asymmetries seem to involve complementary analyses of the sort that depend on incompatible neural computations, and it would seem useful to insulate those computations from each other in order for them to proceed efficiently at the same time.

With this in mind, it is not surprising that the corpus callosum has been hypothesized to play two important but very different roles: transferring information between the hemispheres and creating a kind of inhibitory barrier that minimizes maladaptive cross talk between the complementary processes for which each hemisphere is dominant.

Although the corpus callosum is certainly critical for certain forms of interhemispheric interaction, it is also clear that some types of information can be transferred subcortically. This includes information about the categories to which an object belongs, contextual information about an object, and certain aspects of information about spatial location. Subcortical structures can also play a role in producing unified

behavioral responses. Although the two hemispheres of the intact brain are capable of sharing many types of information, cooperation at all levels does not necessarily occur all the time. Studies of perceptual processing in normal individuals provide important insights about the factors that determine when it is more efficient for the two hemispheres to operate collaboratively than to operate independently. Laterality techniques described earlier have been modified to include trials that demand interhemispheric collaboration by presenting each hemisphere with only a portion of the total information needed to perform a task (the across-hemisphere condition) and comparing performance to conditions that present all the relevant information to one hemisphere (the within-hemisphere condition). One general conclusion suggested by this research is that distributing information across both hemispheres becomes more beneficial as the task becomes more demanding of attentional resources. That is, when the processing demands are minimal (such as indicating whether two letters are physically identical), there is often a within-hemisphere advantage. However, when the processing demands are increased (such as indicating whether two letters of different case have the same name or whether both are vowels), there is typically an across-hemisphere advantage. Dividing input between the two hemispheres is also beneficial when it permits them to engage in mutually inconsistent processes and at earlier compared to later stages of practice, although these may simply be additional ways of manipulating overall processing demand.

A somewhat different aspect of interhemispheric interaction has been studied in experiments that include bilateral redundant trials in which exactly the same information is presented simultaneously to both hemispheres. When normal observers attempt to identify printed consonant–vowel–consonant letter trigrams, there is a right visual field (left hemisphere) advantage, and the pattern of errors is different for the two visual fields (and hemispheres). On right hemisphere trials there are many more third-letter errors than first-letter errors, as if individual letters are processed in order one at a time. On left hemisphere trials this difference is reduced, even when the error types are normalized to compensate for the left hemisphere advantage. This may reflect the left hemisphere's ability to treat the trigram as a single pronounceable unit and thereby spread attention more evenly across the letters. When the same three-letter stimulus is presented simultaneously to both hemispheres, performance is even better than it is on left hemisphere-only trials, indicating the benefit of including both hemispheres in processing. In view of the very good level of performance on redundant bilateral trials, one might expect the error pattern to be like that obtained on left hemisphere-only trials. In fact, the error pattern on redundant bilateral trials is intermediate between the left and right hemisphere patterns (and often more similar to the right hemisphere pattern), again suggesting processing contributions from both hemispheres.

The processing strategy on redundant bilateral trials is not always a mixture of the different strategies used on unilateral trials. In some cases, the bilateral strategy has been identical to the strategy associated with one hemisphere or the other. Interestingly, when this happens it is not always the strategy of the more efficient hemisphere that emerges on bilateral trials. In fact, it is not always possible to derive the processing strategy on bilateral trials from knowledge of the strategies utilized on unilateral trials. For example, in experiments that required normal observers to make rhyming judgments, certain effects of letter font and case on bilateral trials could not be predicted at all from the complete absence of such effects on unilateral trials. This suggests that interhemispheric collaboration can also have emergent properties that are impossible to deduce from the sum of the parts provided by the two individual hemispheres. It will be important in future studies to identify the factors that determine which of these different types of functional interaction will be observed and the biological mechanisms that underlie the different types of interaction. Uncovering the mechanisms of interhemispheric interaction may also have more general implications for how it is that unified processing emerges from a brain consisting of highly specific, modular subsystems.

V. LATERALITY IN NONHUMAN SPECIES

Behavioral and biological laterality is also ubiquitous in many nonhuman species, with many instances of asymmetry being at least analogous to asymmetries found in humans. At least some may also be homologous, in the sense of sharing common structures and developmental origins. Here, I review some of the most well-established laterality effects in other species and note their relationship to human laterality.

Motor asymmetries have been discovered for a number of species, with individuals sometimes showing very strong left–right preferences. However, population-level biases have been much rarer and none

has matched the magnitude of right-handedness in the human population. Furthermore, preferences seem to depend on variables such as age and sex and on specific task demands. Despite these caveats, in several species of primates there tends to be a left-hand preference for reaching and maintaining postural control but a right-hand preference for manipulation and other high-level skilled activities. In addition to individual variation in paw preference, individual members of many species of mice and rats show a preference to turn in one direction or the other. Although an individual may exhibit a strong directional rotation bias, approximately equal numbers prefer each side. That is, there does not tend to be a population-level asymmetry. Individual rotation biases in rats have been related to asymmetries in distribution of the neurotransmitter dopamine, although the specific direction of the relationship differs for different populations.

There are also certain parallels between left hemisphere language dominance in humans and asymmetries in other species for the production and perception of vocalizations. In Japanese macaques, for example, the left hemisphere is dominant for the discrimination of species-specific vocalizations that are relevant for communication but not for the discrimination of other vocalizations. Also, in chimpanzees that have been trained to use certain visual symbols to communicate, there is evidence of left hemisphere dominance for processing those symbols but not for processing other, nonmeaningful symbols. There is even evidence that the ultrasonic calls emitted by rat pups are processed preferentially by the left hemisphere of their mother and it is well-known that there is left-brain dominance for the control of song in some species of song birds.

A number of asymmetries have been reported with respect to processing the identity and spatial characteristics of visual stimuli. In language-trained chimpanzees, for example, there is a right hemisphere advantage for processing the location of a line within a geometric figure and for identifying complex visual patterns that are not relevant for communication. In addition, rhesus monkeys have been reported to have right hemisphere superiority for recognizing monkey faces. In rats, there is evidence that the right hemisphere may be more involved than the left hemisphere in spatial exploration, although the asymmetry emerges only in rats that have been handled during the course of their early development. Pigeons and newly hatched chicks exhibit left hemisphere dominance for visual pattern discrimination. In chicks, this population-level bias occurs because light strikes only the right eye during a critical period of incubation during which the visual system is developing rapidly. Finding effects of such variables as handling and light stimulation suggests that functional hemispheric asymmetries are likely to be shaped by the complex interaction of both biological and environmental factors.

Research with rats and chicks has also demonstrated asymmetry for emotional behaviors. For example, in both handled rats and chicks the right hemisphere tends to produce emotional activity, whereas the left hemisphere tends to inhibit emotional activity. In addition to providing interesting instances of laterality, effects such as these also illustrate the importance of reciprocal activity between the left and right sides of the brain.

There are also indications that some of the biological asymmetries found in the human brain characterize the brains of certain primates, although the nonhuman asymmetries are smaller and less frequent than those of humans. For example, the brains of both humans and apes show the kind of counterclockwise torque described earlier and in chimpanzees as well as humans the Sylvian fissure tends to be longer on the left side than on the right side.

As noted previously, it is difficult to know which laterality effects in other species are truly homologous to the effects found in humans. Nevertheless, the presence of so many asymmetries in other species provides a useful range of animal models that can be used to learn about the development of laterality across the life span of an individual and across evolutionary time. Among other things, laterality in other species indicates that the emergence of language is not a prerequisite for the emergence of other behavioral and biological asymmetries.

VI. LATERALITY ACROSS THE LIFE SPAN

The various functional aspects of laterality correlate with each other weakly or not at all. This is inconsistent with the hypothesis that there is but one fundamental processing dimension along which the hemispheres differ and from which all laterality effects are derived. Consequently, laterality is unlikely to be determined by any single developmental influence. Instead, the relative independence of different manifestations of laterality indicates that there are probably several different biological factors that interact with several different environmental influences to determine an individual's pattern of laterality. Furthermore, the laterality patterns that are characteristic of mature

individuals are likely to be rooted in early stages of ontogenetic development. In this section, I outline several promising developmental possibilities.

As noted previously, the emergence of laterality in the domestic chicken illustrates how laterality can result from the interplay of biological and environmental factors. During the last 5 days of incubation, the head of the chick embryo is turned so that the left eye is pressed against the body and the right eye is turned out toward the egg shell. If the egg (and, therefore, only the right eye) is exposed to light stimulation during this critical period, the chick becomes left hemisphere dominant for making visual discriminations and the right hemisphere becomes dominant for attack and copulation behavior. Lesley Rogers and colleagues have shown that it is possible to eliminate these population-level asymmetries by incubating the egg in darkness and to reverse them by experimentally manipulating the embryo so that only the left eye receives light stimulation. It is instructive to consider whether similar developmental scenarios might exist for humans.

During the course of fetal development in humans, certain areas of the right hemisphere seem to develop earlier than homologous areas of the left hemisphere. Various possibilities have been suggested as to the manner in which certain functional asymmetries could arise from the interaction of these maturational asymmetries and changes in the nature of environmental stimulation. One promising idea is that the earlier developing right hemisphere is initially more influenced than the lagging left hemisphere by the impoverished information that the developing brain encounters. This might include nonlinguistic intrauterine noises, global properties of visual stimuli in newborns, and coarse sensorimotor feedback before and for a few weeks after birth. By being more responsive to these early environmental influences, the right hemisphere may become dominant for perceiving various nonphonetic sounds, for processing global properties of visual stimuli, and for maintaining postural control. In contrast, by being on a later developmental trajectory, the left hemisphere may be saved for complementary specializations that involve processing of more detailed or finer-grained information and sensorimotor feedback and control. In addition, other asymmetric growth spurts during childhood may provide a mechanism for the continuing unfolding of functional hemispheric asymmetry. For example, between 3 and 6 years of age frontal and occipital areas of the left hemisphere appear to develop more rapidly than homologous areas of the right hemisphere, coinciding with a developmental period that is also critical for the acquisition of several aspects of language for which the left hemisphere is dominant.

Psychologist Fred Previc has proposed that several additional prenatal asymmetries also influence later functional asymmetry. One hypothesis for which there is evidence is that certain craniofacial asymmetries that begin to appear as early as the first trimester of pregnancy lead to greater sensitivity of the right ear to sounds in the auditory frequency range that is critical for speech and that, as a result, the left hemisphere becomes more responsive to speech. It has even been suggested that this increased responsivity to speech may set the stage for the development of left hemisphere dominance for speech production. Asymmetries of fetal position during the last trimester of pregnancy have also been linked to later handedness. One ambitious hypothesis for which there is circumstantial evidence is that the typical position of the fetus during the last trimester favors enhanced development of the otolith on the left side, and this asymmetry in the organs of balance favors use of the left side for postural control and use of the right side for other more skilled activities. In view of the fact that there are neural connections between primary vestibular cortex and parietal areas, it has even been suggested that asymmetry in the organs of balance sets the stage for later right hemisphere dominance for spatial processing.

Additional research is needed to provide further tests of these hypotheses and others. Whatever the outcome, it is likely that asymmetries that are very small and subtle when they first appear can eventually have profound effects. As the developing organism encounters new and richer stimulation, the extent to which the brain's neural networks are responsive is likely to be influenced by how those networks have already been modified. It is reasonable to suppose that even a subtle difference can make one hemisphere more responsive than the other to some aspects of new sensorimotor information. In fact, neural network simulations illustrate the plausibility of such scenarios. As this process is repeated, laterality may snowball into the pattern characteristic of adults.

With the foregoing in mind, it is interesting that tasks and processes that are lateralized in adults seem to be lateralized from the time they can be accurately assessed. For example, there are indications that even in newborns there is a kind of left hemisphere dominance for speech perception and that activation of the left and right hemispheres is associated with positive and negative emotions, respectively. Also, in infants as young as a few months of age there appears

to be left and right hemisphere superiority, respectively, for processing local and global aspects of visual patterns.

At the other end of the life span, it has been hypothesized that performance of the right hemisphere declines more rapidly in old age than does performance of the left hemisphere. Despite a number of careful investigations, there is little or no evidence to support this right hemiaging hypothesis. Instead, it appears that lateralized processes remain so throughout the life span and that, on average, the performance of both hemispheres changes in the same way. However, there are indications of an age-related decline in the efficiency with which the two hemispheres can communicate and coordinate their activities.

VII. EVOLUTION OF LATERALITY

The seeds of laterality were sown in our distant evolutionary past and may even be related to asymmetry in the molecules and particles of which all living things are constructed. Here, I outline a plausible scenario for the evolutionary emergence of human laterality. Although accounts of evolution remain speculative and difficult to test, the following scenario is built on what has already been said about laterality across species and on more detailed discussions of evolutionary considerations that I and others, including Michael Corballis and John Bradshaw, have offered elsewhere.

In freely moving animals, there are strong evolutionary pressures for bilateral symmetry in the placement of sense organs and limbs and in those parts of the brain that are associated with them. This is so because such animals must be equally attentive to both sides of space and be able to move in a straight line. However, there is little pressure toward symmetry for other systems that are not involved in direct sensory registration or in direct motor responses to the environment. In fact, asymmetry may be favored in such cases as a way of packaging organs more efficiently into the body cavity or of packaging cognitive functions more efficiently into a brain whose size is limited. Once an organism's brain becomes functionally asymmetric, additional asymmetries are likely to arise via the same kind of snowball mechanism discussed earlier with respect to the development of laterality within an individual. That is, at least some of the additional evolutionary adaptations that are favored by the environment are likely to be implemented more effectively by one side rather than the other, with the result that those new adaptations also become asymmetric. Note that this would occur regardless of whether the environment favored hemispheric asymmetry for the new adaptations. As this process is repeated, the extent of functional brain asymmetry increases.

We have seen that at least some of the hemispheric asymmetries found in contemporary humans are matched by similar, although usually weaker, asymmetries in other contemporary species. By taking into account how long ago we shared a common ancestor with various contemporary species, we can use the asymmetries found in those species to suggest what sorts of laterality may have been present at different moments in our ancestral line. For example, the existence of a left-hand preference for visually guided reaching in present-day prosimians suggests that at least some type of motor asymmetry was present in our ancestral line by the time our line branched from the prosimian line, nearly 60 million years ago (mya). This left-hand advantage extends to some species of Old World monkeys. In addition, some species of Old World monkeys show right hemisphere dominance for processing monkey faces and left hemisphere dominance for processing communicatively relevant vocalizations. The similarity of these asymmetries to those of humans suggests that their precursors were present in our last common ancestors, approximately 40 mya. Contemporary chimpanzees show right-hand dominance for certain high-level tasks as well as left and right hemisphere dominance for processing communicatively relevant symbols and certain visuospatial tasks, respectively. Thus, it is possible that these asymmetries were present in the last common ancestors we shared with chimpanzees, approximately 5 mya.

Although it is admittedly speculative, the foregoing is sufficiently plausible to suggest that some forms of hemispheric asymmetry were already present in our ancestral line before the first hominids emerged approximately 4 mya. If so, then the initial appearance of hemispheric asymmetry was not triggered completely by either language or tool use, as some earlier evolutionary accounts maintained. To be sure, language and tool use are among the critical milestones that set us apart from other species, and along with other milestones, such as walking upright, they are likely to have played an important role in the continued evolution of brain laterality.

Hominids mastered the art of standing and walking upright in the period from 2 to 4 mya. This seems to have been a very important milestone for a number of

reasons. For example, an upright stance freed the hands from the need to provide postural support and minimized their use in locomotion. As a result, the hands were under less environmental pressure to be controlled in a symmetrical fashion. In fact, for efficient performance in activities such as the manufacture and use of tools, the two hands must be used in a complementary fashion; for example, one hand holds a stone while the other hand chips away at it to fashion the point of a spear or a scraper. The precursors of handedness in other primate species suggest that sufficient asymmetry already existed to produce a bias toward right-handedness for the more skilled aspects of movement (e.g., chipping a stone held by the left hand). In addition, prior left hemisphere dominance for vocalization may also have contributed to make the left hemisphere a more efficient neural substrate for the organization of purposeful, sequential actions in which spatial constraints imposed by the environment are minimal (i.e., for what has been termed manual praxis). In fact, an analysis of the flaking patterns on stone tools manufactured by *Homo habilis* suggests that the majority of the population was already right-handed as far back as 1.5–2 mya. Interestingly, study of casts taken from the inside of skulls indicates that certain structural asymmetries (e.g., longer left hemisphere Sylvian fissure) were also characteristic of the brains of *H. habilis*.

Tool making and language are both forms of praxis and involve properties such as recursion, embeddedness, and generativity. Both gestural communication and vocalization also require sequences of precisely timed movements. In view of these similarities, there may have been pressure for the same brain hemisphere to become dominant for those aspects of vocalization and language that are shared with tool making and gestural communication. Certainly, tool use and language seem to have interacted in a synergistic fashion to produce dramatic changes in human skill in the period from 10,000 to 30,000 years ago.

Additional factors are also likely to have contributed to the continued evolution of laterality. For example, we have seen that the corpus callosum may have evolved, in part, to reduce maladaptive cross talk between computations performed by homologous regions of the two hemispheres. With respect to brain asymmetry, this would add to the evolutionary advantage of segregating complementary or mutually inconsistent processes into opposite hemispheres. In addition, contemporary humans are characterized by a prolonged period of postnatal immaturity, a characteristic that is believed to have emerged approximately 2 mya. We have seen that homologous areas of the two hemispheres may mature at different rates. It seems likely that the consequences of such maturational gradients would be greater to the extent that the brain undergoes less of its development in the impoverished uterine environment and more of its development postnatally, especially because the sensorimotor capacities of newborns change dramatically during the first few days, weeks, and months after birth.

VIII. INDIVIDUAL VARIATION IN LATERALITY

Thus far, I have emphasized contemporary human laterality patterns that could be considered prototypical, as if we were all more or less identical. Although there is sufficient homogeneity to justify consideration of the prototype, there is also sufficiently reliable heterogeneity to warrant consideration of individual variation. Of particular interest has been the possible relationship of functional hemispheric asymmetry to a number of other between-subject factors, including handedness, sex, intellectual ability, and psychopathology.

Handedness is of particular interest because it is a behavioral manifestation of brain laterality for certain manual activities and it is also related to other, more cognitive aspects of laterality. Hand dominance is determined by a variety of genetic and environmental factors, both before and after birth. The direction and magnitude of hand dominance may even be determined by different factors, with the magnitude being more heritable and with the direction being more subject to environmental influence. Environmental influences include pre- and postnatal trauma, prenatal levels of testosterone and other hormones (higher prenatal levels of testosterone are associated with greater incidence of left-handedness), asymmetric positioning of the fetus *in utero*, as well as the biases of the postnatal world. As noted previously handedness is at least moderately related to other forms of laterality. On average, laterality for a group of left-handed individuals is in the same direction as that for a group of right-handed individuals, but the magnitude of the asymmetry is smaller. This group difference does not necessarily occur because individual left-handed people are uniformly less lateralized than individual right-handed people since more left-handers than right-handers show a laterality effect in a direction opposite that which is considered prototypical. Rather, there is a greater proportion of left-handed individuals who show a laterality effect in a direction

opposite that which is considered prototypical. It is noteworthy that even though the relationship between handedness and other forms of laterality is only of weak to moderate strength, it has been demonstrated by all the various techniques and tools that are used to study laterality. This suggests that if the relationship of laterality to other variables such as biological sex were as strong as the relationship of laterality to handedness, it would be relatively straightforward for these techniques to uncover.

There are many effects of fetal hormone levels on brain development in other species, and in humans there are clear relationships between biological sex and cognitive ability. For example, women tend to outscore men on tests of verbal fluency, and men tend to outscore women on tests of spatial ability. Thus, it is plausible that there are sex-related differences in at least some aspects of brain laterality. This possibility receives modest support from the fact that the incidence of left-handedness is slightly higher in men than in women, consistent with the hypothesis that higher levels of fetal testosterone promote development of the right hemisphere relative to the left hemisphere. When handedness is controlled, however, evidence for additional sex-related changes in other forms of laterality is equivocal. The most encouraging indications of such relationships have come from the observation of deficits in patients with unilateral brain damage, with some studies finding evidence of greater functional hemispheric asymmetry in men than in women. However, not all studies have shown such sex differences and there may be alternative explanations for some of them. For example, it has been suggested that there is a lower rate of aphasia after left hemisphere damage in women than in men because of sex-related differences in the organization of language within the left hemisphere rather than because of differential lateralization per se. Behavioral laterality studies using neurologically intact men and women provide, at best, weak support for the hypothesis of greater functional laterality in men because the results have been quite variable. Also, even when sex-related differences are found, there are sometimes alternative explanations in terms of the tendency for men and women to prefer different cognitive strategies that may bias performance toward different hemispheres. Perhaps most clear is that if the laterality differences between men and women were as large as the laterality differences between right-handed and left-handed individuals, they would be well-known by now.

There are indications that individual variations in brain laterality may be related to individual differences in cognitive ability. For example, relative performance on tests of verbal and visuospatial ability are related to handedness, but the relationship appears complex and is moderated by sex and by overall reasoning ability. There is also evidence that extreme intellectual precocity, especially for mathematical reasoning, is related to advanced development of the right hemisphere relative to the left, perhaps as a result of increased testosterone levels during fetal development. It also appears that some forms of dyslexia (impaired acquisition of reading) are related to subtle abnormalities within the language areas of the left hemisphere, perhaps leading to an overreliance on the less efficient mechanisms of the right hemisphere. Considerable additional work is needed, however, to substantiate these relationships and to discover the precise mechanisms that might account for them.

The existence of hemispheric asymmetry for emotion has led to consideration of the possible relationship of laterality to psychopathology. Among the more promising hypotheses is the idea that schizophrenia is related to dysfunction of an anterior region within the left hemisphere, an area that is believed to be important for language and for controlling parietal areas involved in attention. It has also been hypothesized that the corpus callosum, and therefore interhemispheric interaction, is dysfunctional in schizophrenics, but these hypotheses remain controversial and may apply to only a subset of schizophrenics. In a complementary way, depression has been linked to disturbances of the right hemisphere, but more work is needed to confirm this link and to understand its functional significance.

Popularized accounts of laterality have suggested that people can be classified as ''right-brained'' or ''left-brained'' depending on whether they prefer to use strategies and modes of cognition associated with one hemisphere or the other. Thus, left-brained people are said to be rational and analytic, whereas right-brained people are said to be intuitive, artistic, and creative. A few well-rounded individuals might even be lucky enough to be identified as ''whole-brain thinkers.'' Various paper-and-pencil schemes have been proposed to achieve this sort of classification or measure of ''hemisphericity,'' and exercises have been proposed to help more of us utilize whichever side of the brain we tend to neglect. To be sure, there is individual variation on the various dimensions of laterality, but there is no evidence that any neurologically intact individuals are functionally half-brained in the manner referred to as hemisphericity. Individuals do differ reliably in cognitive style, personality, creativity, and so forth. At the

same time, however, we have seen that hemispheric asymmetries are subtle, with no indication that aspects such as rationality and creativity are the exclusive product of one brain hemisphere. Instead, both hemispheres contribute to virtually everything we do.

IX. CONSCIOUSNESS, MIND, AND THE DUAL BRAIN

The dual nature of the human brain has led to interesting discussion of the implications for consciousness and mind. Among other things, the discovery that both of the disconnected hemispheres of split-brain patients have a good deal of competence with respect to perception and motor control has led to speculation about whether or not the surgery has produced a doubling of consciousness or resulted in people with two minds. The neurobiologist Roger Sperry, who won the Nobel prize for his work with these patients, believed that this was the case. In part, he based his conclusion on the fact that the disconnected right hemisphere has its own perceptions, cognitions, and memories of which the disconnected left hemisphere is completely unaware. Others, however, have questioned whether the right hemisphere can truly think and whether its limited abilities include the same level of awareness and consciousness that seems typical of the left hemisphere. Certainly, the disconnected right hemisphere is usually incapable of speech and other forms of verbal communication that are uniquely human and that some have argued, are essential for the concept of "mind." Of course, there is much debate among philosophers and cognitive scientists about the extent to which language is essential for thinking, conscious reflection, or mental life.

An interesting hypothesis, advanced by Joseph LeDoux and Michael Gazzaniga, is that an important function of an individual's verbal left hemisphere is the construction of an internal, subjective reality or ongoing narrative based on its observations of his or her own overt behavior. In this view, observation of one's own behavior is necessary because many processes that influence behavior are not open to conscious experience. LeDoux and Gazzaniga emphasize this with respect to behavior and bodily sensations produced by the right hemisphere, of which the left hemisphere would have no direct knowledge. However, even within the left hemisphere there appear to be a great many modules whose processes may influence behavior but that are not themselves open to conscious awareness. Thus, the need to create an ongoing personal narrative by making inferences about one's own behavior is likely to extend to covert processes performed by both hemispheres. More work is clearly needed, however, to test these and other hypotheses about the different roles played by the two cerebral hemispheres in the emergence of mind from brain.

See Also the Following Articles

BILINGUALISM • BRAIN DEVELOPMENT • CEREBRAL CORTEX • CONSCIOUSNESS • CORPUS CALLOSUM • EVOLUTION OF THE BRAIN • LANGUAGE, NEURAL BASIS OF • LEFT-HANDEDNESS • NEURAL NETWORKS

Suggested Reading

Banich, M. T., and Heller, W. (1998). Evolving perspectives on lateralization of function [Special issue]. *Curr. Directions Psychol. Sci.* **7**(1).

Beeman, M., and Chiarello, C. (Eds.) (1998). *Right Hemisphere Language Comprehension: Perspectives from Cognitive Neuroscience*. Erlbaum, Mahwah, NJ.

Christman, S. (Ed.) (1997). *Cerebral Asymmetries in Sensory and Perceptual Processing*. Elsevier, Amsterdam.

Corballis, M. C. (1997). The genetics and evolution of handedness. *Psychol. Rev.* **104,** 714–727.

Gazzaniga, M. S. (2000). Cerebral specialization and interhemispheric communication: Does the corpus callosum enable the human condition? *Brain* **123,** 1293–1326.

Halpern, D. F. (2000). *Sex differences in cognitive abilities*. Erlbaum, Mahwah, NJ.

Heilman, K. M. (1997). The neurobiology of emotional experience. *J. Neuropsychiatr. Clin. Neurosci.* **9,** 439–448.

Hellige, J. B. (1993; 2001 paperback). *Cerebral Hemisphere Asymmetry: What's Right and What's Left*. Harvard University Press, Cambridge, MA.

Hellige, J. B. (2000). Cerebral hemispheric specialization in normal individuals: Experimental assessment. In *Handbook of Neuropsychology* (F. Boller and J. Grafman, Eds.), Vol. 1**,** Part 1. Elsevier, Amsterdam.

Iacoboni, M., and Zaidel, E. (Eds.) (2002). *The Parallel Brain*. MIT Press, Cambridge, MA.

Ivry, R. B., and Robertson, L. C. (1998). *The Two Sides of Perception*. MIT Press, Cambridge, MA.

Kosslyn, S. M. (1996). *Image and Brain: The Resolution of the Imagery Debate*. MIT Press, Cambridge, MA.

Provins, K. A. (1997). Handedness and speech: A critical reappraisal of the role of genetic and environmental factors in the cerebral lateralization of function. *Psychol. Rev.* **104,** 554–571.

Rogers, L. J. (1997). Early experiential effects on laterality: Research on chicks has relevance to other species. *Laterality* **2,** 199–219.

Springer, S. P., and Deutsch, G. (1998). *Left Brain/Right Brain: Perspectives from Cognitive Neuroscience*. Freeman, New York.

Vallortigara, G. (2000). Comparative neuropsychology of the dual brain: A stroll through animals' left and right perceptual worlds. *Brain and Language* **73,** 189–219.

Left-Handedness

STANLEY COREN
University of British Columbia

GLOSSARY

Broca's area The location in the brain, typically the left hemisphere, that controls speech production.

diffuse control system This refers to control of a function or behavior that involves a number of brain centers and pathways.

genetically fixed traits Genetically determined traits that are characteristic of a particular species.

pathological handedness Handedness that results after some pathological event, such as birth stress-related events, disturbs the natural development of right- or left-handedness.

rare trait marker theory A theoretical and statistical model that explains why relatively rare, but apparently benign, traits in a species are often associated with pathological conditions.

sidedness The preferential use of the right or left hand, foot, eye, or ear.

soft sign A behavior or symptom that may suggest the presence of an underlying neurological or physiological pathology.

Wernicke's area The location in the brain, typically the left hemisphere, that controls speech comprehension.

Handedness refers to the fact that most people consistently use the same hand for tasks in which skill and dexterity are required and only one hand can be used. Thus, a person who almost always uses his or her right hand when writing, throwing a ball, cutting with a knife, or using a hammer would be defined as being right-handed. Estimates of the number of right-handers in the population are between 88 and 92%. Given such an overwhelming bias toward the right in humans, it is not surprising that researchers have asked the following questions: If, as a species, we appear to be programmed to be right-handed, then why are there any left-handers? Do left-handers differ from right-handers in any ways other than their handedness? Are there any advantages or disadvantages in being left-handed?

I. SIDEDNESS

Although everyone probably knows about the existence of handedness, most people are unaware of the fact that it is just one aspect of a group of lateral biases. In the same way that we are handed, we are also footed. One would demonstrate footedness in tasks such as habitually using the same foot to step on a bug or to kick a ball to hit a target. In footedness, humans also show a right-sided bias, with approximately 80–82% of the population being right-footed.

In addition to showing motor biases favoring one side, we also show sidedness in the use of our bilateral sense organs. We demonstrate eyedness by consistently choosing the same eye to sight down a telescope or to peep through a small hole. We would be showing earedness by usually choosing the same ear to listen to the faint ticking of a clock or to press against a door to hear noises on the other side. These manifestations of sidedness are also biased toward the right, although

not as strongly as for hands and feet. Approximately 66–71% of the population is right-eyed, and 58–60% are right-eared.

Although it is clear that as a species we are right-sided, the specific proportions may vary from one study to another, depending on the particular definition used to define any aspect of sidedness. Thus, most researchers define the dominant hand as being the hand that people prefer to use for skilled tasks. Others may use the skill or strength of the hand as a measure. In addition, some researchers require consistent right-sided preference in all tasks for an individual to be classified as right–sided for that index, whereas others only require that a majority of responses show the dextral bias. One can usually tell when this latter definition is being used since the labels given for a particular aspect of sidedness might be "right-sided" versus "non-right-sided" as opposed to "right-sided" versus "left-sided"; however, regardless of the definitions or scoring procedures used, the predominance of right–sidedness in human populations is clear.

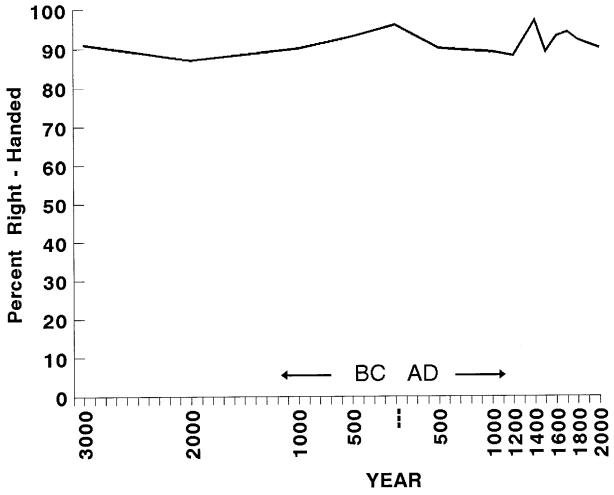


Figure 1 Analysis of handedness as shown in artworks over a period of approximately 50 centuries.

II. THE EVOLUTION OF HANDEDNESS

The nature of human handedness is unique among mammalian species. We can set up situations, analogous to handedness tests in humans, in which animals must manipulate something with only one paw. This might be a task in which a cat must reach down a relatively narrow tube to pull out a treat or in which a monkey must insert a stick into a small hole to get something to eat. When we do this, we find that most cats, rats, and monkeys are right- or left-pawed. Although individual animals show behaviors analogous to handedness, there is one major difference between these animals and humans. Whereas 9 of 10 humans are right-handed, in other species the proportion of right- and left-sided individuals is approximately 50%. In other words, there is not right-sided bias to the animal population.

It is possible to estimate the handedness of humans over history to determine if we were always right-handed. Figure 1 shows the results of a study that examined paintings and drawings, reasoning that if artists were drawing from life, then they would draw the tools and weapons held by their models in the hand that they saw the person using. Analysis of more than 50 centuries of such drawings found that the proportion of right-handedness remained at approximately 90% from the Paleolithic Era (the Old Stone Age) until the present. The date when right-handedness became dominant in our species can be pushed further back in time. Paleontologists have examined the wear patterns on stone tools and the grinding marks on devices used to grind grains. These tools and implements date back between 8000 and 35,000 BC and involve the early humanoid *Homo habilis*, one of the earliest tool makers. The wear patterns on these artifacts confirm that even at that early date, there appeared to be a consistent predominance of right-handers. Perhaps the most astounding evidence derives from more than 1.5 million years ago, involving one of our very early hominid ancestors, australopithecus. Although the australopithecines were not tool makers, they were tool users and would pick up an appropriately sized and shaped rock or stick and use it for a weapon. Examination of the skulls of baboons hunted by this early precursor to humans shows that the vast majority of these hominids were already right-handed.

III. THE GENETICS OF HANDEDNESS

Given the evidence that handedness patterns in humans extend far back into evolutionary history, it is not surprising that the vast majority of theories of handedness have included the suggestion of a genetic factor. There have been a number of studies that have examined handedness in families in order to verify this hypothesis. Unfortunately, the empirical evidence on handedness does not strongly support genetic theories.

There is a fairly consistent finding that left-handedness is more likely in children of left-handed mothers. However, when we do the fine-grain analyses needed to determine that a characteristic is inherited, the numbers relating the father's handedness to that of the offspring, the relationship between brothers and sisters, or predictions from the grandparent's handedness just do not seem to fit the patterns required by most genetic models. Some theorists are still trying to work out more sophisticated mathematical descriptions that might show that handedness is inherited; however, there is still a strong element of doubt in the data.

Perhaps the most compelling evidence against a simple, strong genetic determinant for handedness comes from the results of studies that compared the handedness of monozygotic or identical twins with those of dizygotic or fraternal twins. Twin studies are usually considered as providing the clearest indication of the presence or absence of inherited components for most behavioral traits. Monozygotic twins have an identical set of genes since they come from a single egg, fertilized by a single sperm, that later splits into two individuals in the early stages of cell division. Dizygotic twins come from two egg cells fertilized by different sperm; hence, they share only the level of genetic similarity that we would find between any pair of brothers or sisters. This means that, at a minimum, if there is a genetic component in handedness, one would at least expect that pairs of monozygotic twins would be more likely to have the same handedness than dizygotic twins. Figure 2 summarizes the results of 16 studies that compared the handedness of twins. Notice that there is no difference between monozygotic twins and dizygotic twins with regard to the likelihood that they will share the same handedness. Even more striking is that the likelihood that any form of twins will have the same handedness is no greater than if we randomly chose pairs of unrelated individuals and determined whether or not they had the same handedness.

This is not to say that genetics has no part in the determination of handedness. Certainly, genetic factors determine many human characteristics, such as the number of eyes, ears, and limbs that characterize our species. We might refer to these as genetically fixed characteristics, which are expected in all members of the species. We might contrast these characteristics to genetically variable characteristics, such as eye color, that do vary in the population, depending on the specific genes transmitted to the offspring by its parents. There is evidence that suggests that the strength or consistency of handedness is inherited from parents and, hence, is genetically variable. Children of parents with inconsistent or mixed handedness (tending toward ambidexterity) are more likely to have children with mixed handedness than are parents with consistent handedness. However, whether one is left- or right-sided seems to be a genetically fixed characteristic of species, with evolution programming humans to be right-sided. If right-handedness was meant to be a characteristic of species, then why are some people left-handed?

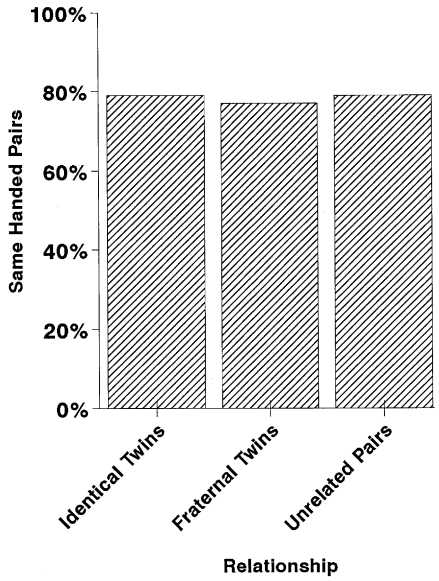


Figure 2 The number of same-handed pairs of twins does not differ among identical twins (who share identical gene sets) and fraternal twins (who are as genetically similar as any pair of siblings), and the twins are no more similar in their handedness than are unrelated pairs of individuals.

IV. THE BRAIN AND HANDEDNESS

Control of the limbs, including the hands, is contralaterally organized, meaning that the left hemisphere of the brain controls the right side of the body and the right hemisphere of the brain controls the left side of the body. In the 1860s, Paul Broca discovered that the major language production center in the brain is located in the left hemisphere of the brain. This center is now called Broca's area in his honor. There is another speech center called Wernicke's area that is associated with speech comprehension, and it is also located in the left hemisphere of the brain. The fact that

both the right-hand and primary speech functions are controlled by the same half of the brain has fuelled speculations attempting to link brain organization, language functions, and handedness. Broca started these speculations by introducing the notion of the "dominant hemisphere," by which he meant the hemisphere that directs not only the movements of writing, drawing, and other fine movements but also language and perhaps even major aspects of logical thought. He then suggested that left-handers might have a brain that is organized in a mirror image to that of right-handers, with language control on the right side of the brain.

The notion that handedness and language control are related in the brain has persisted and is still found in some theories today. The data, however, suggest that this relationship is unfounded. One important source of data on this issue has been the observations of people who are injured on one side of their brain or who acquire a pathology that damages one side of the brain. This often results in the development of aphasia, which is the loss of language abilities. Evidence from such conditions first led Broca to place the language control center in the left hemisphere. Other techniques also provide information on this matter. These involve procedures that effectively "turn off" one hemisphere of the brain. This can be done through the injection of certain drugs (such as sodium amytol) into one side of the brain. Alternatively, electric shock, such as that which is used in electroconvulsive therapy for severe psychological depression, can be used. Instead of applying the shock to the whole brain, however, it is applied to only one-half of the brain. If the drug treatment or the electric shock result in temporarily impaired language comprehension or production abilities, then we can conclude that we have turned off the hemisphere of the brain that contains the language control center. The composite results of 12 studies that attempted to locate the hemisphere containing the language control centers are shown in Fig. 3. Notice that although virtually all right-handers have language in their left hemisphere, we do not find the mirror-imaged brain that we expected in left-handers. The data show that the vast majority of left-handers have the same brain organization as right-handers, with their language control exclusively confined to the left hemisphere. Only about 2 of 10 left-handers have language exclusively in the right hemisphere, as had been predicted by Broca. There is also a small group of individuals who seem to have their language control split between both sides of the brain or duplicated on both sides of the brain.

The new technology associated with various brain scanning techniques, such as positron emission tomography, computerized axial tomography, and functional magnetic resonance imaging, has been used to explore the organization and functioning of the brains of left- and right-handers. The overwhelming number of studies have found that there is a remarkable similarity between brains, regardless of handedness, with the only major difference being that hand control regions are somewhat larger and more active on the side opposite the dominant hand. If there is any general pattern of results that seems to emerge, it is that right-handers seem to have brains that are slightly more asymmetrical, in terms of both their structure

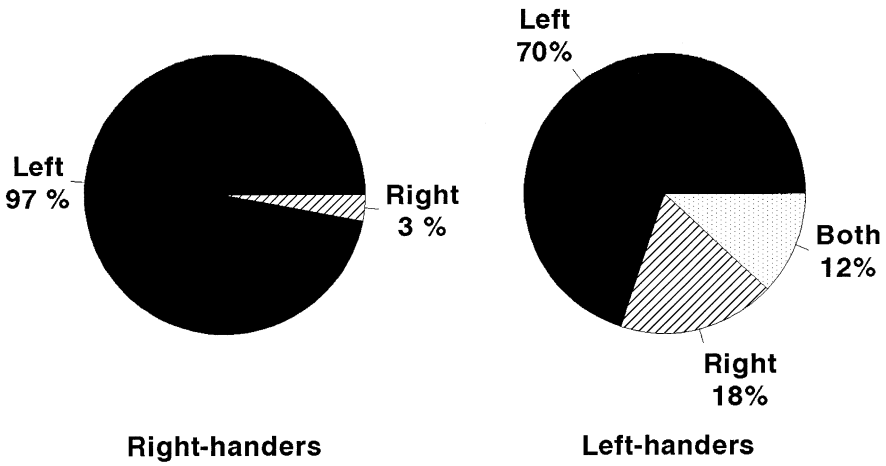


Figure 3 The side of the brain that contains the centers that control language ability for both right- and left-handers is usually the left hemisphere.

and their functional specialization. A larger proportion of left-handers seem to have more symmetrical brains and are more likely to have functions equally represented in both hemispheres. However, in most instances, a researcher presented with only data based on the structural aspects of the brain, or even the brain's function in a variety of circumstances, would be unable to determine if the brain that he or she were looking at was that of a left- or a right-hander.

V. CULTURAL AND ENVIRONMENTAL INFLUENCES ON HANDEDNESS

It should not be surprising that a technological and constructed environment created by a species in which 9 of 10 individuals are right-handed should have a bias toward use of the right hand. Most tools and equipment, furniture, traffic patterns in stores and museums, and even formal seating arrangements are structured for the convenience of the right-handed majority. Everyday implements, such as scissors, gear shifts, ice cream scoops, rulers, can openers, pencil sharpeners, and even the location of the winding stem on wrist watches, are biased toward right-handed usage. Sporting gear such as fishing reels and rifles, musical instruments such as violins, guitars, and banjos, common machinery, including voting machines, slot machines, time card punches, and also candy and soft drink dispensers are designed with a bias that facilitates right-handed operation. Even the direction in which screws are threaded favors the right-hander on the power (forward) movement, a tendency that carries the bias to other implements such as corkscrews. Given this right bias to the environment, it is apparent that left-handers are forced to learn to do many things with their right hand that a right-hander would never be expected to do with his or her left hand. If there is a learned component to handedness, this should serve to greatly reduce the number of left-handers in the world.

The presence of approximately 10% left-handers among humans is even more surprising due to direct cultural pressures to make the whole world right-handed. At some level, our society seems to intensely dislike left-handers and for proof one need go no further than our own language. The very word "left" in English comes from the Anglo-Saxon word "lyft," which means "weak" or "broken." No less an authority than the venerable Oxford English Dictionary goes on to define "left-handed" as including the meanings "crippled," "defective," "awkward," "clumsy," "inapt," "characterized by underhanded dealings," "ambiguous," "doubtful," "questionable," "ill-omened," "inauspicious," and "illegitimate."

Many common phrases in the English language demonstrate our culture's negative view of left-handedness. For instance, a left-handed compliment is actually an insult. A son from the left side of the bed is illegitimate. A left-handed marriage is no marriage at all, but refers to an unconsecrated or adulterous sexual liaison, such as in the phrase a "left-handed honeymoon with someone else's husband." Thus, a left-handed wife is actually a mistress. A left-handed diagnosis is wrong and left-handed wisdom is a collection of errors. To be about left-handed business is to be engaged in something unlawful or unsavory. Sailors speak of ships that are left-handed, meaning that they are unlucky or "wrong" in some way. It is interesting to note that there is not one positive phrase to be found in the language regarding "left" or "left-handed."

One must not suppose that speakers of the English language have a unique dislike of left-handers. The tendency appears to be universal. For example, in French the word for left is gauche, which also conveys the meanings "crooked," "ugly," "clumsy," "awkward," "uncouth," and "bashful." The German word for left-hand is links. The dictionary definition of the term linkisch is "awkward, clumsy, and maladroit." Such negative associations with left-handedness go back a long way. Even in early Latin in which the word for left was sinister, it had already taken on its alternate contemporary meaning of evil.

In many societies, use of the left hand for activities such as eating or writing is considered impolite, insulting, or the sign of ill breeding. It is therefore not surprising that perhaps between 70 and 80% of the population of left-handers report that parents or teachers made overt attempts to change them to right-handers. Some of these attempts could be quite brutal, involving punishment for using the left hand or even strapping or tying the left hand down to force right hand use.

What is most surprising about cultural pressure on handedness is that it has such a poor success rate. For females, 4 of 10 left-handers fail to change their handedness, whereas 3 of 4 males do not shift their handedness. Even for those who do change, the change appears to be only for selected actions, which society puts direct pressure on. Thus, a left-handed child may learn to eat or write with his or her right hand, but he or she will still throw a ball or brush his or her teeth with

the right hand. A good example of this can be seen in Japan, where the proportion of left-handers, measured by writing hand, is only 2 or 3%, about one-fourth of that found in North America. However, if one measures handedness by the hand used to hold a screwdriver or to use a hammer, one finds the same proportion of left-handers in Japan as in the United States.

This all leads to the conclusion that handedness is not a casual learned set of behaviors. Handedness is determined early and is quite intractable to change. If genetic factors associated with our particular species suggest that we should all be right-handed, and our culture and constructed environment bias us toward right-handedness, then why are there any left-handers?

VI. PATHOLOGICAL LEFT-HANDEDNESS

Because simple genetic explanations for the existence of left-handedness do not seem to work, an alternate theoretical position has emerged. This begins with the suggestion that, although there are some natural left-handers, right-handedness should be expected unless something unusual has occurred. Thus, although there is a genetically fixed bias toward right-handedness, and this implies some set of genes that determine handedness, in specific situations there might be partial penetrance (which simply means that the gene does not express its full set of characteristics). If the right-handed gene does not express itself, then handedness becomes a matter of chance, and a left-hander could be the result. Natural left-handedness could also occur due to nonpathological differences in the uterus during pregnancy, such as the position in which the fetus lies. For example, if a fetus comes out with the head directed to the back of the mother's right side (right occiput anterior position), the child is more likely to be left-handed than if it were born from the more common left occiput anterior position (with the head directed toward the back of the mother's left side).

Most deviations from the expected pattern of fetal development occur because something has gone wrong during pregnancy or the birth process. For this reason, researchers quickly began to focus on the relationship between handedness and birth stressors or pregnancy complications. A large body of data has accumulated that indicates that such pathological factors are associated with an increased likelihood of left-handedness. A number of particular factors have been singled out. For example, premature birth, prolonged labor, RH incompatibility, breech delivery, multiple births, and anoxia (reduced oxygen to the fetus, resulting in what is sometimes called a "blue baby") are all factors likely to result in left-handedness. Also, babies requiring resuscitation at birth, those that suffer the complications that lead to cesarian delivery, or even those that require forceps or other instruments to assist in the delivery are all more likely to become left-handed. Thus, it is not surprising that older mothers (aged 40 and older), who are prime candidates for pregnancy and birth complications, are more than twice as likely to produce left-handed children.

One of the more interesting suggestions is that certain hormonal imbalances during birth can produce an increased likelihood of left-handedness. A theory closely associated with the neuropsychologist Norman Geschwind is that while in the uterus, if the fetus is exposed to an elevated concentration of the hormone testosterone, this could also produce left-handedness by altering the relative rate at which the two hemispheres of the brain develop. Since testosterone is also the male hormone, this could also explain the observation that men are more likely to become left-handed than are women (approximately 12% left-handedness for men versus 8% for women).

The neurological reason left-handedness can arise from so many different forms of birth stress has to do with the fact that handedness may be said to have a diffuse control system. A diffuse control system simply refers to the fact that the control of handedness is neurologically complex, involving a number of brain sites and neural pathways, including at least three motor systems that originate in the cerebral cortex, several subcortical sites, several commissural systems, the pyramidal tract, the reticulospinal tract, the rubrospinal pathway, and the proprioceptive and kinaesthetic systems in the postcentral cortex. In all, at least 23 neural centers or systems have been found to be important to hand control. The notion is that anything that interferes with the normal development or functioning of any of the many sites or systems that control hand use can also alter the natural development of right-handedness. If there is any such disruption, handedness will be randomly determined, meaning that half of these birth-stressed individuals can be expected to become left-handers.

How common are pathological left-handers compared to natural left-handers? Several mathematical analyses of the distribution of handedness, birth stressors, and their interactions have led to the conclusion that left-handers are split approximately evenly between these two categories, with half having

their handedness as a result of natural, although relatively atypical, factors, and the other half obtains their left-handedness due to pathological mechanisms.

VII. LEFT-HANDEDNESS AND PHYSICAL AND PSYCHOLOGICAL FITNESS

Throughout the years there have been a large number of studies that have suggested that left-handedness may serve as a sign or a marker indicating that there may be other psychological and neurological problems present in the individual. The reasoning is as follows: To the extent that left-handedness may be caused by pathological factors, left-handedness might be a marker (at least statistically) for the possible existence of some form of neural pathology, psychological deviance, or developmental abnormality. It is suggested that, at least for the pathological left-handers, the same pathology that caused the left-handedness might have also caused some form of collateral damage that can reduce the individual's physiological or psychological fitness through direct or secondary mechanisms.

The surprising suggestion that left-handedness may be a marker for other problems is actually supported by a substantial body of research that found that left-handedness is frequently associated with a variety of symptoms or syndromes. In the vast majority of circumstances, when there are a disproportionately high number of left-handers in a group, that group is marked by some negative factor. The number of findings in which left-handedness is associated with positive outcomes in the research literature is much rarer. A review of the literature indicated at least 60 negative pathological or undesirable conditions associated with left-handedness as opposed to only 4 positive or desirable conditions. Considering associations that have been reported two or more times in the research literature, Table I provides an idea of the positive and negative conditions related to an increased proportion of left-handers.

VIII. THE RARE TRAIT MARKER MODEL

With a research literature replete with suggestions that left-handedness is predominantly associated with negative conditions, some researchers have attempted to provide theoretical models that try to explain why left-handedness may be a marker for pathology. Clinical researchers would call it a soft sign for pathology, meaning that it is not something that you can directly cut out, weigh, or subject to direct chemical or physical analysis. Handedness is merely an observable behavior that indirectly suggests a problem, rather than a hard sign, which might be actual observation of a damaged section of the nervous system. Hard signs almost always indicate pathology. Thus, a person who tests positive for HIV antibodies almost definitely has been exposed to the disease AIDS. Left-handedness is a soft sign in the sense that an individual who shows it is not definitely pathological but, rather, is more likely to have some pathological condition than a person who does not show this behavioral sign.

The linkage between handedness and these various pathological conditions occurs due to both neurological and statistical considerations. The statistical component has been mathematically explored in the form of the rare trait marker model, which indicates that a number of statistically rare traits are often found to be associated with a range of pathological conditions. Examples of rare traits include animals that are colored differently from the vast majority of their species. For instance, a "blue–marl" collie, a white dog, or an albino human often have major sensory deficits affecting their vision or hearing. These rare traits could include rare palm crease patterns, rare fingerprint patterns, or rare distributions of toe lengths, and even unusual ear shapes are often found to be associated with cognitive or physical deficits. Left-handedness, which affects only about 10% of the population, would also qualify as a rare trait.

According to theoretical considerations, if we start with a trait, such as handedness, that has a common and a rare form, we need only add the possibility that the appearance of the rare versus the common trait can be influenced by pathological factors. If we meet both of these conditions, then because of statistical considerations we have all that is needed to create a clinical soft sign for pathology.

To demonstrate the operation of the rare trait marker model, consider the situation diagrammed in Fig. 4. In this example, we suppose that we are starting with a population in which, if development proceeded naturally, 90% of the people would develop into right-handers and 10% into left-handers. Suppose that there was some disturbance in development—some aberrant condition or some birth risk factor—that caused 10% of each group to deviate from their targeted handedness. The result would be 9% of the population (10% of the 90% who would be right-handed)

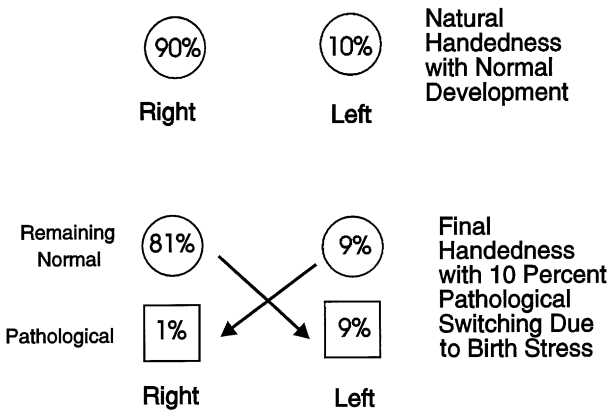


Figure 4 The rare trait marker theory is demonstrated in this diagram. The circles represent the percentages of normal individuals, whereas the squares represent the percentages of pathologically switched individuals. Notice that in the final group of left-handers, half are pathological, whereas the number of pathological right-handers is less than 2%.

"pathologically" shifting from right-handedness to left-handedness, whereas only 1% of the population (10% of the 10% originally targeted to be left-handed) shifts from left- to right-handedness. The final result is that half of the resulting population of left-handers (9% out of the total of 18% left-handers) are pathological, whereas only 1.2% (1% out of 82%) of the right-handers are pathological. This means that the relative risk of a left-hander being pathological is approximately 41% greater than that of a right-hander being pathological. How well a rare trait predicts pathology depends on its distribution in the population and the likelihood of pathological change. By chance, the approximately 10% incidence of handedness falls into the optimal predictive range of values.

The fascinating thing about the rare trait marker theory is that it really does not depend on any specific physiological, injury, or disease mechanisms. It makes no assumptions about the particular vulnerability of various parts of the nervous system. In other words, we do not have to know why or how or even when a particular injury or damage occurred. All we need to know is that there is a common and rare trait, and that some form of pathology (actually any form of pathology) can produce the rare trait.

Although the previous discussion suggests that left-handedness might be a sensitive marker for a number of pathological conditions, there is a flip side that reduces its usefulness in some ways. Given the large number of possible sites for pathology that might cause the observed left-handedness, we also must face the fact that there is little specificity. Left-handedness might be seen as a soft sign indicating increased risk for many problems; however, which specific problems an individual might suffer from, or which neural loci are involved, is not determinable. This means that given an individual about whom all one knows is that he or she is left-handed, one can list the many problems that he or she is at risk for but one cannot determine which of them this particular left-hander is most likely to have which part of the nervous system is most apt to be damaged, or what the original source of the pathology might have been. We can say with some certainty that the more diffuse the pathology that an individual has suffered, the higher the likelihood of non-right-handedness and also the higher the probability that other functions other than handedness will also be disturbed.

IX. HANDEDNESS AND LONGEVITY

The data suggest that a significant number of left-handers arrived at their left-handedness through some form of pathological event during their fetal development or associated with their birth, and that this same pathology also resulted in other physical or psychological problems. From Table I, it should be clear that a number of these problems, especially those associated with reduced immune system efficiency, might represent major health risk factors.

Since the late 1970s a number of studies have examined handedness as a function of demographic factors, such as age, sex, and race. The studies that examined handedness over the life span found something unexpected—namely, that there is a diminishing percentage of left-handers in older age groups. Although the rate of decline varies across studies, it is usually large, from about 15% left-handedness in teenagers to less than 1% in 80-year-olds. Figure 5 shows the age proportion of left-handers for various age groups based on the average values taken from seven different studies that examined this issue.

Researchers first thought that this decrease in the number of left-handers might be due to social factors since older people, educated earlier in the 20th century, are more likely to tell stories about how they were forced to use their right hands by teachers and parents than are people born later in the century. Forced switching of left-handers to right-handedness in older age groups could explain this age-related decline in the number of left-handers. However, as discussed earlier, handedness is relatively difficult to change. In addition, several analyses of the literature suggest that the proportion of adult left-handers in North American populations has not changed since the early 1900s.

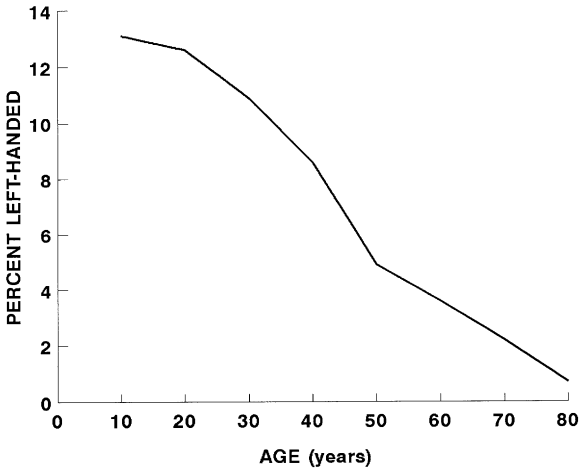


Figure 5 The percentage of left-handers diminishes with increasing age.

The alternative explanation for failing to find older left-handers might be that they are simply no longer in the population. That is, left-handers have a shorter life span, and the reason that we fail to find many older left-handers is because a large proportion of them have died early. When this hypothesis was first suggested, it evoked a lot of media attention and controversy. Recently, evidence supporting this conclusion has been obtained from several sources. For instance, a number of studies have examined the life spans of professional baseball and cricket players since records for these sports include not only birth and death dates but also information on the hand that was used in throwing. These have confirmed that left-handers seem to die younger. In other studies, researchers examined randomly selected death records and then contacted surviving next of kin to obtain handedness information. Again, the results suggest that the longevity of left-handers is less than that of right-handers. Although not all of the research that has been done on this issue has found the expected differences, whenever there is a significant relationship between handedness and life span it is always in the direction of left-handers having a poor survival rate.

Table I

Conditions That Two or More Research Reports Have Shown to Be Associated with a Disproportionately High Number of Left-Handers

	Negative conditions associated with left-handedness	
Alcoholism	Hashimoto's thyroiditis	Poor spatial ability
Allergies	Hayfever	Poor verbal ability
Asthma	High and low extremes in numerical ability	Predisposition toward aggression
Attempted suicide	Homosexuality	Psychosis
Autism	Hypopigmentation	Reading disability
Bed-wetting	Immune disorders	Reduced adult height
Brain damage	Impulsive aggression	Reduced adult weight
Chromosomal damage	Infection susceptibility	Regional ileitis
Celiac disease	Juvenile delinquency	Schizophrenia or schizotypal thinking
Crohn's disease	Juvenile-onset diabetes	School failure
Clinical depression	Language problems	Sleep difficulty
Criminality	Learning disabilities	Slow maturation
Deafness	Lower intelligence scores	Slow physical development
Depression	Manic–depressive psychosis	Strabismus
Drug abuse	Mental retardation	Stuttering
Drug hypersensitivity	Migraine headaches	Sudden heart attack death
Emotionality	Myasthenia gravis	Transsexuality
Epilepsy	Neural tube defects	Ulcerative colitis
Excessive smoking	Neuroticism	Urticaria
Eczema	Poor school performance	
	Positive conditions associated with left-handedness	
Extremely high intelligence	Good divergent thinking ability	High spatial ability
	High musical ability	

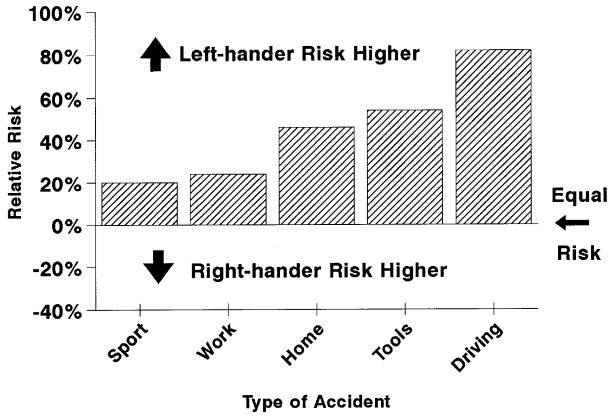


Figure 6 The relative risk of accidents is higher for left-handers in every category of accident studied.

Although many left-handers were dying of factors traceable to the problems in Table I, there was a surprise in the data: Left-handers were five times more likely to die of accident-related injuries. This brings us back to the issue that our technological world and the built environment have been designed for the safety and the convenience of the right-handed majority. There is evidence that left-handers are more likely to suffer injuries while playing sports, driving in traffic, working with tools and motorized equipment, working with kitchen and home implements and devices, and even while engaged in military activities. Figure 6 shows the relative risk of various injuries as a function of handedness based on data from four different studies. Notice that in all cases, left-handers seem to be at considerably higher risk. Obviously, these injuries are the direct result of a cultural and societal bias toward facilitating activities done with the right hand that we discussed earlier. These right-handed biases continue in today's world, even though we now know that they place the left-hander at higher risk of injury.

X. ARE LEFT-HANDERS AND RIGHT-HANDERS REALLY DIFFERENT?

When we examine all the data associated with handedness, it becomes obvious that left-handedness is not just a minor behavioral deviation shown by about 1 in every 10 humans. Handedness does not appear to be a marker for genetic differences as previously thought, neither is it necessarily a sign of difference in brain organization. The evidence indicates, however, that for about half of all left-handers, handedness does represent a sign that may indicate some level of neurological and developmental pathology. From the standpoint of behavioral medicine, we may view left-handedness as a risk factor that may well predict susceptibility to illness, physiological deficits, psychological problems, and perhaps even a shortened life span.

See Also the Following Articles

BRAIN DEVELOPMENT • BROCA'S AREA • EVOLUTION OF THE BRAIN • HAND MOVEMENTS • LATERALITY • WERNICKE'S AREA

Suggested Reading

Bishop, D. V. M. (1990). *Handedness and Developmental Disorder*. Lippincott, Philadelphia.

Corballis, M. C. (1991). *The Lopsided Ape: Evolution of the Generative Mind*. Oxford Univ. Press, New York.

Coren, S. (1993). *The Left-Hander Syndrome: The Causes and Consequences of Left-Handedness*. Random House, New York.

Coren, S. (Ed.) (1990). *Left-handedness: Behavioral implications and anomalies*. Advances in Psychology, No. 67. North-Holland, Amsterdam.

Springer, S. P., and Deutsch, G. (1995). *Left Brain, Right Brain*. Freeman, New York.

Limbic System

JOSEPH L. PRICE
Washington University School of Medicine, St. Louis

GLOSSARY

amygdala (from Latin "amygdale," almond or tonsil) A complex of nuclei embedded (like the meat of an almond) in the anteromedial temporal lobe.

cingulate gyrus (from Latin "cingulum or cingo," girdle, to surround) A long curved gyrus that surrounds the corpus callosum on the medial side of the cerebral hemisphere.

hippocampus (from Greek "hippocampos," seahorse) A very specialized region of the cerebral cortex, which is deeply folded into the lateral ventricle. In marsupials and rodents the hippocampus is largely situated in the dorsomedial cortex, dorsal to the thalamus, but in humans and other primates it has shifted into the medial temporal lobe.

limbic system (from Latin, "limbus," edge, border, or fringe) A system of interconnected brain structures situated around the medial edge of the cerebral hemisphere.

nucleus accumbens septi (from Latin "accumbens," leaning) The rostral, ventromedial portion of the corpus striatum, comparable to the caudate nucleus and putamen, that is immediately lateral to the septum and appears to "lean" against it.

nucleus basalis of Meynert Prominent nuclei of large cells in the basal forebrain, most of which use either acetylcholine or GABA as neurotransmitter.

orbital cortex (from Latin "orbita or orbis," wheel-track, circle) The cortex situated dorsal to the bony "orbit" that encases the eye.

prefrontal cortex Traditionally, the cortex in front of the "excitable" frontal cortex (i.e., the motor cortex, which was found in the 19th century to evoke body movements when excited with electrical stimulation).

The internal state of the body is largely controlled by the autonomic nuclei in the spinal cord and brain stem, which provide for reflex control of visceral functions, and by the hypothalamus, which provides for higher level coordination of autonomic and endocrine functions. As anyone who has checked his or her heart rate during an exciting action movie knows, however, visceral function is also modulated in relation to forebrain analysis of visual, olfactory, and other sensory stimuli. Furthermore, such visceral reactions are also modified in relation to previous experience, as you can tell if you watch the same action movie so many times that the surprises become expected. The modulations related to present and past sensory experience are both dependent on several forebrain structures that are linked together as the "limbic system." On the one hand, these structures have substantial connections to the hypothalamus and brain stem and provide an interface between forebrain sensory systems and visceral control areas. On the other hand, the limbic structures interact with the cerebral cortex, thalamus, and basal ganglia to relate visceral function to the ongoing sensory environment and to cognitive functions. In particular, these provide for the modulation of visceral and other functions across time, such that responses are altered in relation to previous experience. In humans and other primates, at least, this process is directly involved in conscious emotion and memory.

As defined by Broca (1878) and later by Papez (1937), the limbic system generally consists of a group

of structures around the medial edge of the cerebral hemisphere. (''Limbus'' means border or edge; e.g., to be ''in limbo'' is to be suspended between two states.) These structures include the amygdala, the hippocampus, the parahippocampal gyrus, and related structures, such as the orbital/medial prefrontal cortex, cingulate cortex, anterior and mediodorsal thalamic nuclei, the ventromedial corpus striatum (i.e., the accumbens nucleus and medial caudate nucleus), and the nucleus basalis of Meynert.

I. AMYGDALA

The amygdala is a complex of several nuclei in the rostromedial part of the temporal lobe (Figs. 1 and 2). There are several deep nuclei (lateral nucleus, basal nucleus, and accessory basal nucleus), which are substantially interconnected with the temporal, insular, and frontal cortex, the striatum, and the mediodorsal thalamus. On the surface are a number of modified cortical areas, many of which are interconnected with the olfactory system. In the dorsal part of the amygdala are the central nucleus and medial nucleus, which have connections with the hypothalamus and autonomic brain stem nuclei.

The amygdala receives input from all the sensory systems. In lower mammals, these are dominated by inputs from the olfactory system, although there are also inputs from the taste/visceral afferent system and the visual and somatosensory systems. In primates the most prominent inputs are derived from higher order sensory association cortex, especially from the visual areas in the inferior temporal cortex (i.e., the temporal visual processing stream, important for analysis of form and color and recognition of complex stimuli

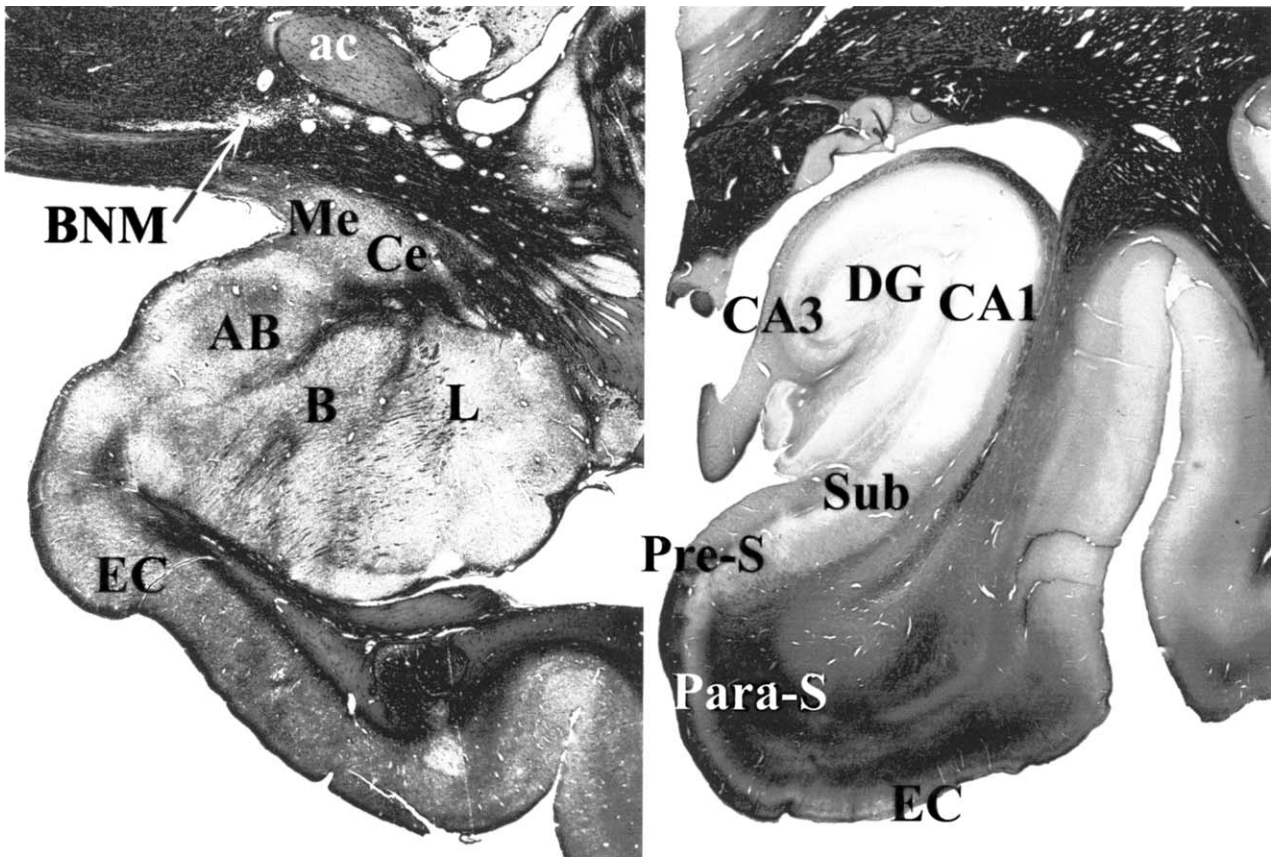


Figure 1 A photomicrograph of the amygdala, hippocampus, and parahippocampal gyrus from a human brain. BNM-basal nucleus of Meynert, ac-anterior commisure. The amygdala includes the lateral nucleus (L), basal nucleus (B), accessory basal nucleus (AB), central nucleus (Ce), and medial nucleus (Me). The hippocampus, located posterior to the amygdala in the medial temporal lobe, includes the parasubiculum (Para-S), presubiculum (Pre-S), subiculum (Sub), fields CA1 and CA3, and the dentate gyrus (DG). The entorhinal cortex (EC) occupies the parahippocampal gyrus ventral to the amygdala and hippocampus.

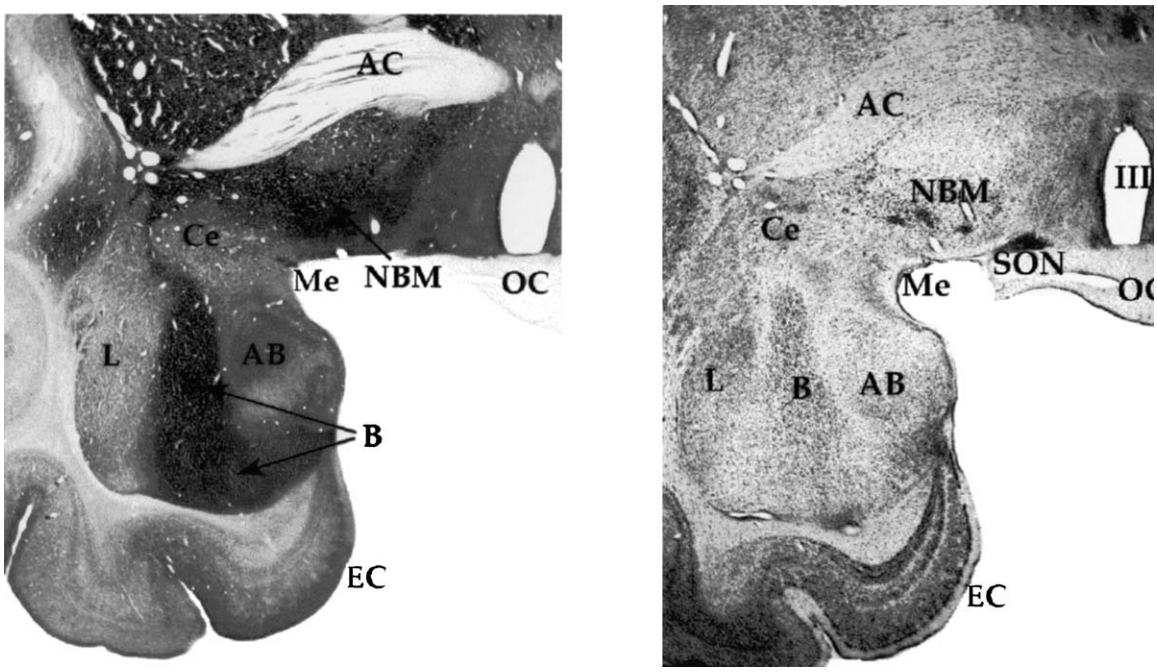


Figure 2 Photographs of the amygdala and entorhinal cortex (EC) of a macaque monkey, stained for acetylcholinesterase (left) and cells (Nissl method) (right). Note the component nuclei of the amygdala, including the basal nucleus (B), lateral nucleus (L), accessory basal nucleus (AB), central nucleus (Ce), and medial nucleus (Me). Also note the cholinergic nucleus basalis of Meynert (NBM), ventral to the anterior commissure (AC), the supraoptic nucleus (SON), the optic chiasm (OC), and the third ventricle (III).

such as faces) (Fig. 3). Recordings from the amygdala show that the neurons respond to complex sensory stimuli, including visual stimuli such as faces, as well as to other sensory modalities. In addition to the sensory aspects of the stimuli, the responses are influenced by the novelty of the stimulus or its affective sign (whether the stimulus is rewarding or aversive).

There is a complex system of intraamygdaloid connections that associate these inputs. These connect the deep amygdaloid nuclei (which interact with the cortex) with the central and medial nuclei (which connect to the hypothalamus and brain stem). There are also extensive connections to other limbic areas, including the hippocampus, parahippocampal gyrus, and the nucleus basalis of Meynert.

The outputs from the amygdala can be divided into three categories (Fig. 3):

1. Return projections back to the sensory areas that project into the amygdala: In the case of the visual system, these return projections even extend back to the primary visual cortex.

2. Descending projections to the visceral control centers of the hypothalamus and brain stem: The central amygdaloid nucleus projects to a wide variety of autonomic-related cell groups, including the lateral hypothalamus, the periaqueductal gray, the parabrachial nucleus, the nucleus of the solitary tract, the dorsal vagal nucleus, and the ventrolateral medulla. Through these projections, the amygdala can influence heart rate and blood pressure, gut and bowel function, respiratory function, bladder function, etc. For example, stimulation of the central nucleus can cause stomach ulcers as well as changes in cardiovascular function.

3. Interactions with the orbital and medial frontal cortex, both via direct amygdalocortical projections and via connections with the mediodorsal thalamic nucleus and the ventromedial parts of the basal ganglia: This circuit appears to be involved in determination of the affective "sign" of sensory stimuli (e.g., whether it is rewarding or aversive) and in setting mood.

Several additional observations may be used to illustrate the role of the amydala in the control of emotional behavior. Bilateral lesions of the amygdala and adjacent medial temporal structures in monkeys

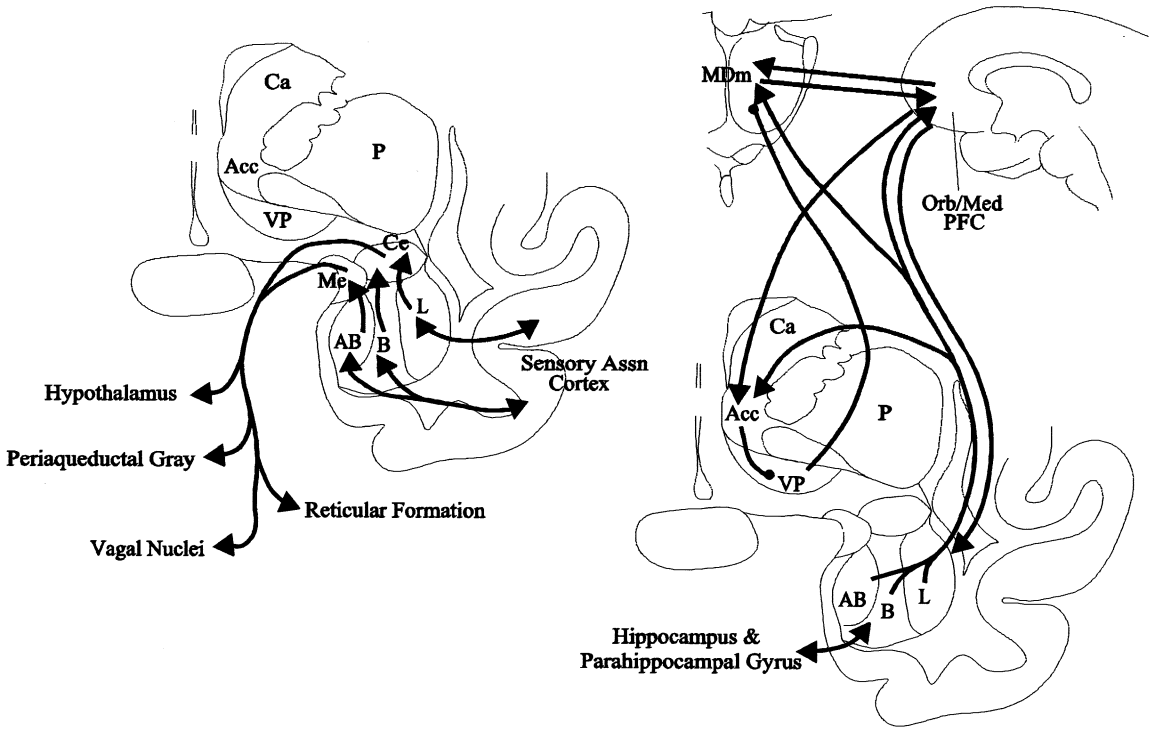


Figure 3 Schematic summary of the axonal connections of the amygdala. (Left) Inputs from sensory association cortical areas and output from the central and medial nuclei to the hypothalamus and brain stem. (Right) Connections are illustrated with the mediodorsal thalamus (MDm), the ventromedial striatum [accumbens (Acc) and caudate nuclei (Ca)], ventral pallidum (VP), and the orbital and medial prefrontal cortex (Orb/Med PFC) as well as with the hippocampal formation. The connections from Ca/Acc to VP and from VP to MDm are shown with arrowheads to indicate that they are inhibitory.

produce "psychic blindness" (Kluver–Bucy syndrome) in which the animals can see objects but are apparently unable to distinguish their significance. Such animals are not capable of appropriate social behavior, and in group settings they become isolated and solitary.

In rats, bilateral lesions of the amygdala block the acquisition of conditioned fear responses. For example, if a light is consistently coupled with a painful foot shock, the animal comes to associate the two stimuli, and the light becomes a "fearful" stimulus that can produce autonomic and behavioral responses characteristic of fear. After bilateral lesions of the amygdala, the shock produces the same response as before, but the light never becomes a fearful stimulus. Recent observations indicate that a similar deficit is present in humans who have a rare disease that results in bilateral amygdaloid destruction.

Stimulation of the amygdala in awake cats produces a defense reaction in which there are integrated behavioral and visceral changes that prepare the animal for fight or flight. Stimulation of the amgydala or hippocampus in human patients (during exploratory surgery for severe epilepsy) evokes complex sensory and experiental phenomena, often involving fear and including sensory hallucinations, feelings of deja vu, and memory-like episodes that resemble static dreams.

II. HIPPOCAMPUS

The hippocampus is a specialized part of the cerebral cortex, folded into the temporal horn of the lateral ventricle about the hippocampal fissure (Figs. 1 and 4). It is composed of several subregions, including (in order) the entorhinal cortex (in the parahippocampal gyrus), the parasubiculum, presubiculum, subiculum, fields CA1–CA3, and the dentate gyrus.

Like the amygdala, in rats the hippocampus receives strong olfactory inputs, but in primates it is dominated by multisensory inputs, mostly from higher order

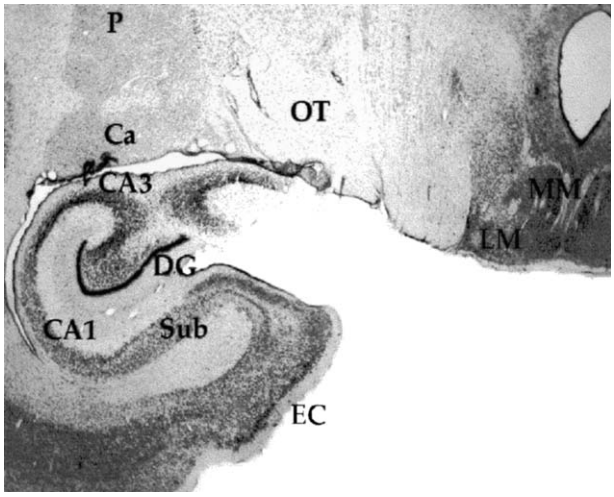


Figure 4 Photomicrograph of the hippocampal formation and the mammillary nuclei in a macaque monkey (stained with the Nissl method). Ca, caudate nucleus; EC, entorhinal cortex; DG, dentate gyrus; P, putamen; OT, optic tract; Sub, subiculum.

sensory association cortex. These inputs reach the hippocampus through a relay in the perirhinal and entorhinal cortex, which in turn receive inputs from multisensory association cortical areas. There is a relatively unidirectional set of connections through the hippocampus, beginning from the entorhinal cortex to the dentate gyrus, and then in turn to CA3, CA1, the subiculum, and back to the entorhinal cortex (Fig. 5).

Outputs to other parts of the brain mainly arise in the subiculum and entorhinal cortex. These resemble those of the amygdala:

1. Projections back to many sensory association cortical areas, including the visual and auditory areas of the inferior and superior temporal cortex.
2. Projections to the hypothalamus (but not to brain stem autonomic nuclei).
3. Connections with prefrontal and cingulate cortical areas, via direct projections and also via projections to the anterior and mediodorsal thalamic nuclei, and the mammillary nuclei, and projections to the ventral part of the basal ganglia.

The hippocampus and surrounding areas of the parahippocampal gyrus are critically involved in memory processing in general and spatial orientation in particular. Recordings in the hippocampus have demonstrated cells that fire when the animal is in a particular spatial location, as defined by characteristic sensory and other experiential stimuli. This suggests that the hippocampus provides a neural mechanism for association of different parameters that are necessary for the moment-to-moment incorporation of experience into memory. Strikingly, bilateral lesions of the hippocampus and parahippocampal cortical areas produce amnesia, an inability to form new memories (although older memories may be intact). By comparison, amygdaloid lesions that produce emotional disturbances do not produce amnesia.

III. ORBITAL AND MEDIAL PREFRONTAL CORTEX AND CINGULATE GYRUS

Cortical areas on the ventromedial surface of the frontal lobe, in the cortex dorsal to the orbit and the region rostral and ventral to the genu of the corpus callosum, are substantially interconnected with all the limbic areas discussed previously and participate in many of the same functions. These areas vary from periallocortical agranular areas with a relatively simple cortical structure in the caudal part of the region to fully developed granular cortical areas more

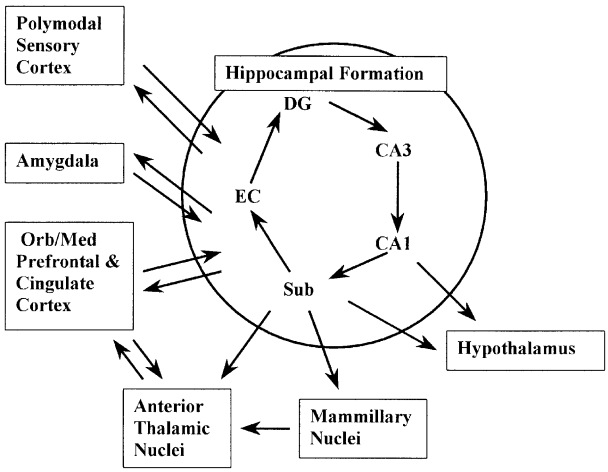


Figure 5 Schematic summary of major axonal connections of the hippocampal formation. Interactions with the polymodal sensory cortical areas occur primarily through the entorhinal cortex (EC). The amygdala, the orbital and medial prefrontal cortex, and the cingulate cortex interact with both the EC and the subiculum (Sub), whereas the Sub provides the primary output to the anterior thalamic nuclei, mammillary nuclei, and hypothalamus. Within the hippocampal formation, the principal flow of activity is from the EC to the dentate gyrus (DG), field CA3, field CA1, the Sub, and back to the EC.

rostrally. In rodents, the agranular areas dominate and the granular areas are almost nonexistent, whereas in primates and especially humans the granular areas are much larger than the agranular areas.

As a whole, the cortical areas can be divided into two relatively interconnected networks that appear to have related but distinct functions (Fig. 6). Sensory inputs from other cortical regions or the thalamus enter many of the orbital cortical areas, and these are integrated by corticocortical connections between these areas. Many of these are related to food or eating (e.g., olfaction, taste, visceral afferents, somatic sensation from the hand and mouth, and vision), and neurons in the orbital cortex respond to multisensory stimuli involving the appearance, texture, or flavor of food. In contrast, many of the areas in the medial prefrontal cortex and a few related orbital areas provide output from the cortex to visceral control centers in the hypothalamus and brain stem. Therefore, the orbital and medial prefrontal cortex appears to be adapted to evaluate feeding-related sensory information and to evoke appropriate visceral reactions.

The function of the ventromedial frontal cortex is considerably wider, however. Food is a primary reward, and many of the orbital neurons respond to rewarding or aversive aspects of stimuli beyond their sensory characteristics. In this, the cortex is closely tied to the function of the related ventromedial striatum, in which reward-related neural activity has also been found.

In addition, lesions of the ventromedial frontal lobe produce dramatic behavioral deficits, which suggests that visceral reactions evoked through this cortical

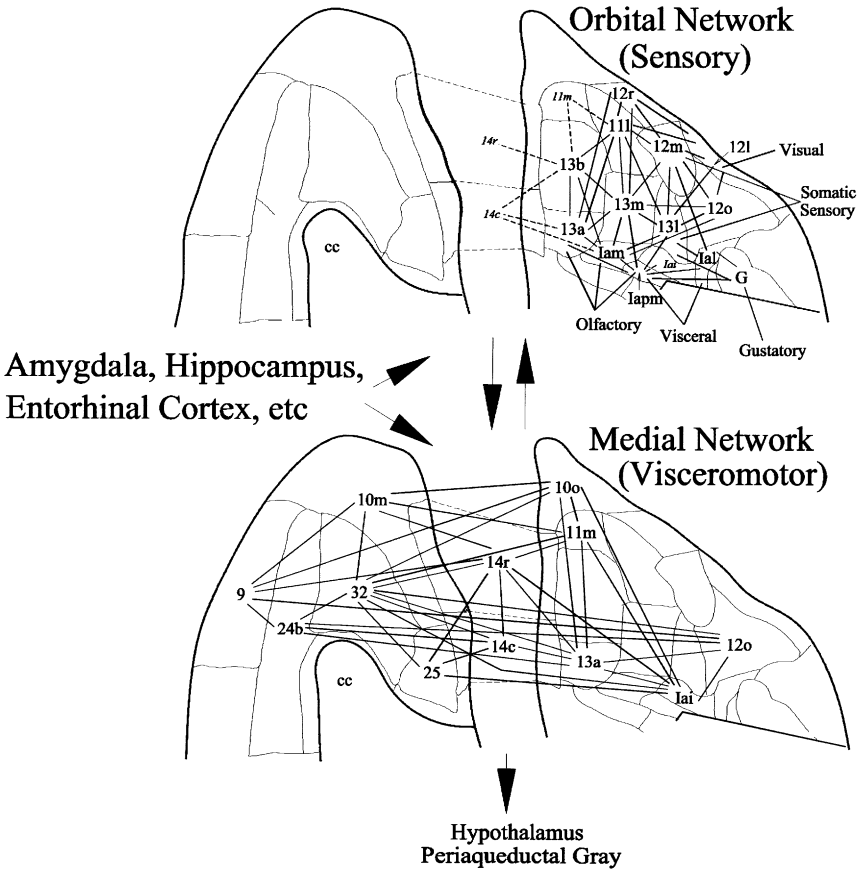


Figure 6 Diagram illustrating connections of architectonic areas in the ventromedial prefrontal cortex in macaque monkeys, drawn onto the orbital (right) and medial (left) cortical surfaces. Sensory input is particularly directed to an "orbital" network of areas (in the orbital cortex), whereas output to autonomic control centers in the hypothalamus and brain stem (especially the periaqueductal gray) arises from a "medial" network of areas (in the medial and orbital cortex). Corticocortical axonal connections between areas, which largely define the two networks, are indicated by lines connecting the areas. In addition to this sensory/visceromotor transfer, interactions with the amygdala, hippocampus, entorhinal cortex, and other limbic structure involve this cortical region in reward appreciation and affective behavior. The numbers and letters are designations of architectonic areas, modified from Brodmann; cc, corpus callosum.

area form a critical component in evaluating alternatives and making choices. As exemplified by the famous 19th-century case of Phineas Gage, individuals with damage to the ventromedial prefrontal cortex do not show deficits in motor or sensory function, or in intelligence or cognitive function, but have devastating changes in personality and choice behavior. Recent studies of such cases by Damasio and colleagues have shown that patients do not show usual visceral reactions to emotional stimuli, and this lack of response correlates closely with their difficulty in making appropriate choices. This has been explained by a somatic marker hypothesis, which postulates that visceral responses are monitored as a quick warning of choices to avoid.

In monkeys the cingulate gyrus is situated dorsal to the corpus callosum, but in humans it extends rostral and ventral to the genu of the corpus callosum and caudally around the splenium. The pre- and subgenual parts of the cingulate cortex are closely related to the medial prefrontal cortex in connections and presumably function, but more posterior cingulate regions appear to be distinct. Although the caudal pole of the cingulate gyrus is connected to the hippocampal formation, the central part of the cingulate gyrus has little relation to limbic structures.

IV. VENTROMEDIAL STRIATUM, VENTRAL PALLIDUM, AND MEDIAL THALAMUS

Although the ventromedial striatum, ventral pallidum, and medial thalamus are not part of the limbic system as traditionally defined, they have substantial connections with all the limbic structures discussed previously and are closely tied to them functionally. The amygdala, hippocampal formation, and the ventromedial prefrontal cortex all project onto the ventromedial part of the striatum, leading into a circuit that connects through the ventral pallidum to the mediodorsal thalamic nucleus (Fig. 4). Like other striato-pallidal-thalamic systems, this circuit involves two inhibitory (GABAergic) synaptic links, from the striatum to the ventral pallidum and from the ventral pallidum to the mediodorsal thalamus. In the mediodorsal nucleus, the inhibitory pallidal inputs interact with monosynaptic excitatory inputs from the amygdala, entorhinal cortex, hippocampus and other limbic structures.

Although it is usually assumed that the striatum is involved in motor functions, most of the motor-related activity is found in the dorsolateral part of the striatum, which is anatomically connected with sensorimotor areas of the cerebral cortex. In contrast, the ventromedial striatum, in keeping with its limbic associations, has been shown to be related to reward and reward-related behavior. In both cases, the role of the striatum may be to inhibit or suppress unwanted patterns of activity in order to allow other patterns to be freely expressed. Specifically, the dorsolateral striatum and related areas of the globus pallidus appear to be involved in switching between different patterns of motor behavior, whereas the ventromedial striatum and pallidum may allow changing of stimulus–reward associations when the reward status of a stimulus has changed.

V. NUCLEUS BASALIS (OF MEYNERT) AND NUCLEUS OF THE DIAGONAL BAND (OF BROCA)

These nuclei consist of scattered groups of large cells in the basal part of the cerebral hemisphere, just ventral to the anterior commissure and the globus pallidus. The nucleus basalis (Figs. 2 and 3) is most prominent at the level of the anterior commissure, but some of its cell clusters extend caudolaterally toward the amygdala and dorsally around the edges of the globus pallidus. The diagonal band nuclei extend medially and dorsally (diagonally) into the septum. Because many of the cells utilize acetylcholine as their transmitter, the nuclei stain very darkly for acetylcholinesterase. Other cells in the complex use GABA as transmitter. As a group, they project axons widely to all parts of the cerebral cortex (and the olfactory bulb and amygdala), although individual neurons within the group appear to have very restricted projections.

The principal action of acetylcholine on cortical cells is to enhance the action of other synaptic inputs. The nuclei are situated along fiber pathways that connect limbic structures such as the orbital and medial prefrontal cortex and the amygdala with the hypothalamus and brain stem, and they receive major input from all of these. The magnocellular basal forebrain nuclei are therefore well situated to modulate cortical activity in relation to limbic activity. They have been implicated in the activation (desynchronization) of the cortex that is characteristic of the waking state, and they are presumably involved in many other functions as well.

VI. THE LIMBIC SYSTEM AND PSYCHIATRIC DISORDERS

Although lesions of limbic structures do not result in apparent sensory or motor deficits, dysfunction of

these structures has been associated with a variety of psychiatric disorders, including depression, bipolar disorder, obsessive–complusive disorder, and schizophrenia. For example, structural changes have been noted in the hippocampal formation, medial thalamus, and prefrontal cortex in schizophrenic subjects. Observations from positron emission tomography indicate that the amygdala and related parts of the prefrontal cortex and medial thalamus are abnormally active in patients suffering from severe unipolar and bipolar depression. Cellular changes, especially in glial cells, have also been reported in the orbital and medial prefrontal cortex and amygdala in depressed subjects. Patients with lesions of these prefrontal areas have also been reported to have a relatively high incidence of depression. As noted previously, lesions of the ventromedial prefrontal cortex produce severe deficits in emotional reactions and in making choices.

See Also the Following Articles

CINGULATE CORTEX • ELECTRICAL POTENTIALS • HOMEOSTATIC MECHANISMS • MOOD DISORDERS • NEUROANATOMY • PREFRONTAL CORTEX

Suggested Reading

Aggleton, J. P. (Ed.) (1992). *The Amygdala: Neurobiological Aspects of Emotion, Memory, and Mental Dysfunction*. Wiley–Liss, New York.

Bjorklund, A., Hokfelt, T., and Swanson, L. W. (Eds.) (1987). Integrated systems of the CNS, part I: Hypothalamus, hippocampus, amygdala, retina. In *Handbook of Chemical Neuroanatomy*, Vol. 5**,** Elsevier, New York.

Broca, P. (1878). Anatomie comparée circonvolutions cérébrales: Le grand lobe limbique et la scissure limbique in la série des mammifères. *Rev. Anthropol. Ser. 2* **1,** 384–498.

Cavada, C., and Schultz, W. (Eds.) (2000). The mysterious orbitofrontal cortex [Special issue]. *Cerebral Cortex* **10**, 205–342.

Damasio, A. R. (1995). *Descartes' Error: Emotion, Reason, and the Human Brain*. Morrow, New York.

Fuster, J. M. (1997). *Prefrontal Cortex: Anatomy, Physiology, and Neuropsychology of the Frontal Lobe*, 3rd ed. Lippincott–Raven, Philadelphia.

LeDoux, J. E. (1994). Emotion, memory and the brain. *Sci. Am.* **270,** 50–57.

Papez, J. W. (1937). A proposed mechanism for emotion. *Arch. Neurol. Psychiatr.* **38,** 725–743.

Squire, L. R., and Zola, S. M. (1997). Amnesia, memory and brain systems. *Philos. Trans. R. Soc. London B Biol. Sci.* **352,** 1663–1673.

Vogt, B. A., and Gabriel, M. (Eds.) (1993). *Neurobiology of Cingulate Cortex and Limbic Thalamus: A Comprehensive Handbook*. Birkhauser, Boston.

Logic and Reasoning

PHILIP N. JOHNSON-LAIRD
Princeton University

GLOSSARY

deductive reasoning The process of establishing that a conclusion follows validly from premises (i.e., that it must be true given that the premises are true).

deontic reasoning Reasoning about actions that are obligatory, permissible, or impermissible.

formal rules of inference Rules that can be used to derive a conclusion from premises in a way that takes into account only the form, not the meaning, of the premises. Logical calculi rely on formal rules, and so do many psychological theories of reasoning.

implicit reasoning A fast, automatic, and largely unconscious process of making inferences in order to make sense of the world and of discourse (e.g., to select the appropriate sense of a word, or to establish the appropriate referent for a pronoun).

inductive reasoning The process of deriving plausible conclusions from premises.

logic The science of implications among sentences in a formalized language. Logical calculi are systems of proof based on formal rules of inference (proof theory); they have an accompanying semantics (or model theory).

mental models Representations of the world that are postulated to underlie human reasoning; each model represents what is true in a single possibility.

validity An inference is valid if its conclusion must be true given that its premises are true. A valid inference from true premises yields a true conclusion; a valid inference from false premises may yield a true or a false conclusion.

Logic captures the implications among sentences. A logical calculus consists of a precise definition of a language and a set of rules of inference that can be used to derive conclusions from premises. The rules are formal, that is, they operate on the form of sentences, not their meaning. The calculus, however, may have a semantics, which provides interpretations for all the sentences in the language. Modern logic lies at the heart of the development of computers and computer programming languages. However, logic is not easy to use in the evaluation of everyday inferences because no algorithm exists for translating such inferences into sentences in a logical calculus—a gap that the logician Bar-Hillel once referred to as the scandal of logic. Logic is also not a theory of how human beings reason. That topic is the province of psychology. Although psychologists studied deductive reasoning for almost the entire 20th century, they began to formulate theories of the process only in the past 25 years. Deductive reasoning is now under intensive investigation, and more is known about it than any other variety of thinking. The aim of this article is accordingly to outline the general principles of logic; to describe current theories of human reasoning, which owe much to logic; and to outline what is known about the role of the brain in reasoning.

I. LOGIC

From the founder of logic, Aristotle, onwards logicians have analyzed formal patterns of valid inference. A deduction is valid if its conclusion must be true given that its premises are true. The original aim of logic, as Leibniz remarked, was to replace rhetoric with calculation. Modern formal logic began during the last quarter of the 19th century, but nowadays logicians draw a sharp distinction between formal systems of

logic, which they refer to as proof theory, and semantic systems of logic, which they refer to as model theory. The distinction is clearest in the case of the sentential calculus. This calculus concerns implications that depend on sentential negation, as expressed by "not," and various sentential connectives, such as "if," "and," and "or," which are treated in an idealized way. The following inference is an example of a valid deduction that can be proved in the sentential calculus:

If the brakes are on and the switches are on then the engine is ready to start.
The brakes are on.
The switches are on.
Therefore, the engine is ready to start.

The inference is based on three atomic sentences (i.e., sentences that contain neither negation nor any connectives): the brakes are on, the switches are on, and the engine is ready to start. The inference is valid and has the form

If A and B then C.
A.
B.
Therefore, C.

where A, B, and C, are variables that can take as values any sentences including those that in turn contain connectives.

Logicians can set up the proof theory for a calculus in various ways. They can formalize the sentential calculus, for example, using just a single rule of inference and a set of axioms, which are assertions that are assumed to be true. However, a more intuitive method, known as natural deduction, dispenses with axioms in favor of formal rules of inference for negation and for each of the sentential connectives. Certain rules introduce connectives into a proof, such as the rule that introduces "and," using it to conjoin two premises:

A.
B.
Therefore, A and B.

Certain rules eliminate connectives from a proof, such as the well-known rule of modus ponens:

If A then B.
A.
Therefore, B.

These two rules suffice to prove the conclusion about starting the engine:

1. If the brakes are on and the switches are on then the engine is ready to start.
2. The brakes are on.
3. The switches are on.
4. Therefore, the brakes are on and the switches are on [The rule for introducing "and" applied to sentences 2 and 3]
5. Therefore, the engine is ready to start. [Modus ponens applied to sentences 1 and 4].

Table I presents a set of formal rules of inference for the sentential calculus. With such rules, you can construct a formal proof, as in the preceding example, with each step in the proof warranted by one of the rules of inference.

Your knowledge of the meaning of the connectives helps you to understand the validity of the rules in Table I. However, the rules do not rely on these meanings. They work in a formal way, allowing you to write patterns of symbols given other patterns of symbols. A proof in a formal calculus is accordingly like a computer program. A computer predicts the weather, for example, but it has no idea of what rain or sunshine is or of what it is doing. It slavishly shifts "bits," which are symbols made up from patterns of electricity, from one memory store to another, and

Table I
Formal Rules of Inference for the Sentential Calculus[a]

A ∴ Not (Not A)	Not (Not A) ∴ A
A B ∴ A and B	A and B ∴ A
A ∴ A or B, or both	A or B, or both. Not A ∴ B
Rule for conditional proof A (a supposition) … B (i.e., B can be derived from A) ∴ If A then B	Rule for modus ponens If A then B A ∴ B

[a]The rules in the left-hand column introduce negation and the sentential connectives into inferences; those in the right-hand column eliminate them from inferences.

displays symbols that meteorologists can interpret as maps of weather. Indeed, proofs and computer programs are intimately related, and certain programs can prove inferences in logical calculi. Likewise, certain programming languages, such as PROLOG, are akin to a logical calculus.

Formal proofs establish that inferences are valid, but validity is not a concept that is defined within proof theory. Its definition hinges on truth, which underlies the semantics of the calculus (i.e., its model theory). In the model theory of the sentential calculus, the truth or falsity of compound sentences depends only on the truth or falsity of their constituent sentences. Thus, an assertion of the form, A or B or both, is true if A is true, B is true, or both of them are true. Otherwise it is false. Logicians lay out these definitions in truth tables, as shown in Table II. Each row in a truth table is a "model" of a possibility and presents the truth value of the compound sentence—in this case, A or B or both—in that possibility. The first row in the table, for instance, presents the case in which A is true and B is true, and so the disjunction is true in this possibility.

One problematic connective is "if." Its everyday usage sometimes departs from its idealized logical meaning in the sentential calculus. An assertion such as

If that patient has malaria then she has a fever

is, in fact, compatible with three possibilities: The patient has malaria and a fever, she has no malaria and fever, and she has no malaria and no fever. It is false in only one case: She has malaria and no fever. The assertion is therefore equivalent to

If that patient has malaria then she has a fever, and if she does not have malaria then she either has or does not have a fever.

Logical license exists just as much as poetic license: Logicians make simplifying assumptions about the meanings of logical terms.

Table II

A Truth Table for the Disjunction *A or B or Both*, Which Shows Its Truth Value for the Four Possibilities Depending on the Truth or Falsity of A and of B

A	B	A or B or both
True	True	True
True	False	True
False	True	True
False	False	False

The validity of an inference in the sentential calculus can be established using the model theory of the calculus. Table III shows how premises can be used to eliminate possibilities from a truth table. When you have eliminated the impossible then, as Sherlock Holmes remarked, whatever remains, however improbable, must be the case. In other words, an inference is valid if the conjunction of its premises with the negation of its conclusion is inconsistent (i.e., not a single row in the resulting truth table contains the entry "true"). For instance, if you conjoin the negation of the conclusion in Table III, "The engine is not ready to start," to the premises, then it would eliminate the last remaining possibility in the truth table. It is therefore impossible for the premises to be true and for the conclusion to be false: The inference is a valid.

Any conclusion that can be proved using formal rules for the sentential calculus is also valid using truth tables and vice versa. There is also a decision procedure for the calculus; that is, the validity or invalidity of any inference can be established in a finite number of steps. Unfortunately, sentential inferences are computationally intractable. It is feasible to test the validity of inferences based on a small number of atomic sentences. However, as the number of atomic sentences in an inference increases, its evaluation in any system —no matter how large or how rapid—takes increasingly longer and depends on increasingly more memory, to the point that a decision will not emerge during the lifetime of the universe.

The sentential calculus has a decision procedure, but it is intractable. The predicate calculus includes the sentential calculus, but also deals with quantifiers—that is, with sentences containing such words as "any" and "some," as in "Any electrical circuit contains some source of current." The predicate calculus does not even have a decision procedure. Any valid inference can be proved in a finite number of steps, but no such guarantee exists for demonstrations of invalidity. Attempts to show that an inference is invalid may, in effect, get lost in the "space" of possible derivations. The principal discovery of 20th century logic, however, is Gödel's famous proof that no consistent calculus is powerful enough to yield derivations of all the valid theorems of arithmetic. Arithmetic is thus incomplete. This result drives a wedge between syntax (proof theory) and semantics (model theory). Any attempt to argue that semantics can be reduced to syntax is bound to fail. Semantics has to do with truth and validity, whereas syntax has to do with proofs and formal derivability.

Table III
The Validity of an Inference Is Shown Using a Truth Table[a]

1. If the brakes are on and the switches are on then the engine is ready to start.
2. The brakes are on.
3. The switches are on.

All that remains is the first possibility, and so it follows validly: Therefore, the engine is ready to start.

Brakes are on	Switches are on	The engine is ready to start	Possibilities that are eliminated
True	True	True	
True	True	False	Eliminated by 1
True	False	True	Eliminated by 3
True	False	False	Eliminated by 3
False	True	True	Eliminated by 2
False	True	False	Eliminated by 2
False	False	True	Eliminated by 2
False	False	False	Eliminated by 2

[a]The premises are used to eliminate possibilities.

II. DEDUCTIVE REASONING

Logic tells us about implications among sentences, but it is not a theory of human reasoning. This topic is a concern of psychology. In the last 25 years of the 20th century, psychologists proposed a variety of theories of reasoning—that it depends on a memory for previous cases, on rules that capture general knowledge, on "neural nets" representing concepts, or on specialized innate modules for matters that were important to our hunter–gatherer ancestors. However, humans have the ability to reason about matters for which they have no specific knowledge. Even if you know nothing about brakes, switches, and engines, you can grasp the validity of the earlier inference about them. This ability lies at the heart of the development of mathematics and logic. Hence, a critical question is whether it depends on syntactic or semantic principles. The following sections describe psychological theories of both sorts.

A. Formal Rule Theories

The first theories of human deductive ability postulated that the mind tacitly uses formal rules of inference like those of a system of natural deduction. Such theories continue to have many proponents, notably Daniel Osherson, Lance Rips, and the late Martin Braine and colleagues. Philosophers have proposed similar theories, and computer scientists have implemented formal systems for the computer generation of proofs. What these proposals have in common is the idea that reasoning depends on applying formal rules of inference to the premises of an inference in order to derive the conclusion in a sequence of steps akin to a proof.

Rips's PSYCOP theory was the first formal rule theory in psychology to cope with connectives and quantifiers and to be implemented in a computer program (written in PROLOG). The system is otherwise typical of formal rule theories. It postulates that reasoning depends on a single deterministic process, that it relies on natural deduction, and that it makes use of suppositions—sentences that are assumed provisionally for the sake of argument, and that have to be "discharged" if a proof is to yield a conclusion. There are two ways to discharge a supposition. First, it can be incorporated within a conditional conclusion (see the rule for conditional proof in Table I). Second, if a supposition leads to a contradiction, then it must be false given that the premises are true (according to the rule of "reductio ad absurdum," which is not shown in Table I). As an example, consider the proof for an inference of a form known as *modus tollens*. There are two premises, such as

1. If the switches were not on then the engine did not start.

2. The engine did start.

The proof starts with a supposition:

3. Suppose: the switches were not on.
4. Therefore: the engine did not start. [Rule for modus ponens applied to 1 and 3]

There is now a contradiction between a sentence in the domain of the premises (The engine did start) and a sentence in the subdomain of the supposition (The engine did not start). The rule of reductio ad absurdum uses such a contradiction to negate, and thereby discharge, the supposition that led to the contradiction:

5. Therefore, the switches were on.

Like other formal rule theories, PSYCOP does not contain a rule for modus tollens, because the inference is more difficult for logically untrained individuals than modus ponens. Hence, it depends on the chain of inferential steps just given. In contrast, an inference of the following form, which we encountered earlier,

If A and B then C.
A.
B.
∴ C.

could be derived in two steps, first conjoining A and B, and then using modus ponens to derive the conclusion. However, the inference is so easy that PSYCOP has a single formal rule for drawing the inference (a conjunctive form of modus ponens).

Formal rule theorists try to postulate psychologically plausible rules of inference and a mechanism for using them to construct mental proofs. One problem is that unless certain rules, such as the rule for introducing "and" (see Table I), are constrained, they can lead to futile derivations:

The brakes are on.
The switches are on.
∴ The brakes are on and the switches are on.
∴ The brakes are on and the brakes are on and the switches are on.
∴ The brakes are on and the brakes are on and the brakes are on and the switches are on.

and so on ad infinitum. One solution is to incorporate the effects of such rules within other rules. In computer programs, however, a rule of inference can be used in two ways: either to derive a step in a chain of inference leading forward from the premises to the conclusion or to derive a step in a backward chain leading from the conclusion to a subgoal of proving its required premises. PSYCOP allows the dangerous rules to be used only in such backward chains, and thereby prevents them from yielding futile steps. PSYCOP therefore has three sorts of rules: those that it uses only forwards, such as the conjunctive rule for modus ponens; those that it uses only backwards, such as the rule for conditional proof; and those that it uses in either direction, such as the rule for modus ponens. A corollary is that reasoners should make modus tollens inferences only when they are given the putative conclusion, or when they can guess the conclusion and then try to prove it.

Given an inference to evaluate, PSYCOP always halts after a finite number of steps either with a proof of the conclusion or else in a state in which it has unsuccessfully tried all its possible derivations. Hence, the theory implies that people infer that a conclusion is invalid only if they fail to prove it. They carry out an exhaustive search of all possible derivations, and only then do they judge that the conclusion does not follow from the premises. However, valid inferences exist that PSYCOP cannot prove. If its exhaustive search has failed to find a proof, then there are two possibilities. Either the inference is invalid, or it is valid but beyond the competence of PSYCOP to prove. A psychological corollary is that people should never know for certain that an inference is invalid.

Formal rule theories postulate that the difficulty of a deduction depends on the number of steps in its derivation and the availability and ease of use of the required rules of inference. Modus ponens is easy because it depends on a single rule; modus tollens is more difficult because it depends on a chain of inferences. Formal rule theorists have corroborated their theories in experiments using large batteries of deductions. They estimate post hoc the probability of the correct use of each rule of inference. When these empirical estimates are combined appropriately for each inference, they yield a satisfactory fit with the difficulty of the inferences in the battery.

B. The Mental Model Theory

The mind may not contain any formal rules of inference unless an individual has learned logic. Instead, inferences could be based on an understanding of the meaning of the premises. Consider the following inference:

From where I stand, the peak of the mountain is directly behind the steeple. The old oak is on the

right of the steeple, and there is a flag pole between them. Therefore, if I move to my right so that the flag pole is between me and the peak of the mountain, the steeple is to the left of my line of sight.

Reasoners might rely on axioms and formal rules to make this inference, but it seems more likely that they imagine the relevant spatial layout. This idea lies at the heart of the theory of mental models.

The theory postulates that mental models have three principal characteristics. First, each model represents a possibility. For example, the disjunction "The switches are on or the brakes are on, or both" calls for a separate model for each of the three possibilities (shown here on separate lines):

switches
brakes
switches *brakes*

where "switches" denotes a model of the switches being on, "brakes" denotes a model of the brakes being on, and the third model combines the two.

Second is the principle of truth: Mental models represent only what is true and not what is false, and in this way they place a minimal load on working memory. Hence, the preceding models do not represent the row in the truth table in which the disjunction as a whole is false (Table II). Likewise, the first model represents that the switches are on, but it does not represent explicitly that in this possibility it is false that the brakes are on. People make a mental "footnote" about what is false, but normally they soon forget it. If they retain such footnotes, however, then they may be able to flesh out their mental models to make them fully explicit. Table IV presents the mental models and the fully explicit models for sentences based on each of the main sentential connectives. Mental models are accordingly like truth tables in which there are no "false" entries.

Third, the structure of a model corresponds to the structure of the situation that the model represents. A model is accordingly like a biologist's model of a molecule. The previous notation for the models fails to capture their rich internal structure. Visual images can be derived from some models, but models are often not visualizable. Early formulations of the theory concerned only the logical terms in the language, but recently the theory has been extended to deal with various sorts of nonlogical terms, such as spatial and temporal relations, and general knowledge about causal relations.

Table IV
Models for the Sentential Connectives[a]

Connective	Mental models	Fully explicit models
A and B	A B	A B
A or else B	A	A ¬B
	B	A B
A or B or both	A	A ¬B
	B	¬A B
	A B	A B
If A then B	A B	A B
	…	¬A B
		¬A ¬B
If and only if A then B	A B	A B
	…	¬A ¬B

[a]The middle column shows the mental models postulated for human reasoners, and the right-hand column shows fully explicit models, which represent the false components in true possibilities using negations that are true: "¬" denotes negation and "…" denotes a wholly implicit model. The footnote on the mental models for "if" indicates that the implicit model represents the possibilities in which A is false, and the footnote on the mental models for "if and only if" indicates that the implicit model represents the possibilities in which both A and B are false.

Reasoners use all the information available to them to construct models—discourse, perception, general knowledge, memory, and imagination. They formulate a conclusion that holds in their models but that was not explicit in the starting information. If a conclusion holds in all the models of the premises, then it is necessary given the premises. If it holds in at least one model of the premises, then it is possible given the premises. The probability of a conclusion depends on the proportion of models in which it holds, granted that each model is equiprobable, which is an assumption that reasoners make in default of evidence to the contrary. The theory accordingly unifies reasoning about necessity, possibility, and probability. They all depend on a semantic process rather than a formal one. They all depend on a grasp of meaning, which is used to imagine the possibilities compatible with the premises.

To illustrate the theory, consider the following inference:

The switches are on or the brakes are on, or both.
The switches are not on.
∴ The brakes are on.

The disjunctive premise elicits the models:

switches
brakes
switches *brakes*

The second premise eliminates the models representing the possibilities in which the switches are on. The remaining model yields the conclusion that the brakes are on. This conclusion is valid because it holds in all the models—in this case, the single model—of the premises.

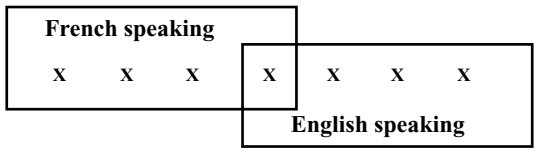


Figure 1 A counterexample used to refute an inference. Each x represents an individual: more than half of them speak French, and more than half of them speak English, but it is false that more than half speak both languages.

C. Five Empirical Phenomena

Psychological investigations have established five principal phenomena of deductive reasoning. The first phenomenon is that the more possibilities that reasoners have to envisage to draw an inference, the more difficult the inference—it takes them longer, and they are more likely to make a mistake. A simple example is that inferences based on a disjunction are more difficult when the disjunction is inclusive, as in the preceding example, than when it is exclusive and allows only two possibilities: "The switches are on or the brakes are on, but not both." The same effect of number of possibilities occurs in reasoning with other sentential connectives, in reasoning about spatial and temporal relations, and in reasoning with premises containing quantifiers, such as "all," "some," and "none."

The second phenomenon is that reasoners use counterexamples to establish invalidity. When reasoners draw conclusions for themselves, they may not consider counterexamples. However, when they reject a conclusion, they can do so by constructing a counterexample—that is, they envisage a possibility that satisfies the premises but refutes the conclusion. One experiment, for example, used problems, such as

More than half of the people in the room speak French.
More than half of the people in the room speak English.
Does it follow that more than half of the people in the room speak both French and English?

Most people responded correctly, "no," and they typically reported having envisaged a situation analogous to the one represented in Fig. 1. They drew such diagrams when they were allowed paper and pencil. They also used counterexamples when they manipulated external models—cut-out paper shapes—in order to reason with quantifiers.

The third phenomenon is that human reasoners spontaneously develop a variety of different strategies in deductive reasoning. They do not use a single deterministic strategy. For example, in reasoning based on multiple premises containing sentential connectives, some individuals develop the strategy of translating each disjunctive premise into a conditional, some base their inferences on the most informative premise, and some make use of suppositions—even when there are categorical assertions among the premises. Many distinct inferential strategies occur, but the space of possible strategies has yet to be mapped.

The fourth phenomenon is the occurrence of illusory inferences. These inferences are compelling but invalid. The following is a typical example:

Only one of the following premises is true about a particular hand of cards:
There is a king in the hand or there is an ace, or both. There is a queen in the hand or there is an ace, or both.
There is a jack in the hand or there is a 10, or both.
Is it possible that there is an ace in the hand?

Most people respond "yes." The first premise is compatible with the possibilities:

King
Ace
King *Ace*

They support the conclusion that an ace is possible. The second premise supports the same conclusion, and so reasoners are likely to respond affirmatively. However, this response overlooks the fact that when one premise is true, the others are false. Thus, if the first premise is true, the second premise is false. In which case, there cannot be an ace. Indeed, if there were an

ace in the hand, then the first two of the premises would be true, contrary to the rubric that only one of the premises is true.

The rubric "only one of the premises is true" is equivalent to an exclusive disjunction, and a compelling illusion occurs in the following inference about a particular hand of cards:

If there is a king in the hand then there is an ace in the hand, or else if there isn't a king in the hand then there is an ace in the hand.
There is a king in the hand.
What, if anything, follows?

Nearly everyone, experts and novices alike, infers that there is an ace in the hand. It follows from the possibilities that people envisage. However, given the disjunction of the two conditionals, it is an error. The disjunction implies that one or other of the conditionals could be false. If the first conditional is false, then even the presence of a king fails to guarantee that there is an ace in the hand. The fallacies arise from a failure to think about what is false. It follows that any manipulation that emphasizes falsity should alleviate them. This prediction has been corroborated experimentally.

The fifth phenomenon is that knowledge and beliefs affect both the interpretation of premises and the process of reasoning. Consider, for example, the following conditional assertion:

If she played a sport then she didn't play soccer. Conditionals are normally compatible with three possibilities (see the fully explicit models in Table IV):

sport	¬ *soccer*
¬ *sport*	¬ *soccer*
¬ *sport*	*soccer*

where ¬ denotes negation. However, the meaning of the noun soccer entails that it is a sport, and so knowledge of this meaning automatically rules out the third of these possibilities. General knowledge and knowledge of the context of an utterance can also eliminate possibilities. Individuals often know what the different possibilities are, and such knowledge modulates the interpretation of assertions. As an illustration, consider the following conditional:

If you strike a match properly then it lights.

Its interpretation includes the salient possibility:

strike	*lights*

As often happens in discourse, however, the antecedent of the conditional fails to describe in complete detail the context in which the consequent holds. There are many circumstances in which a match will not light even if you strike it properly. You know, for instance, that if it is soaking wet it will not light. In fact, you have knowledge of the following explicit possibilities:

soak	¬*lights*
¬ *soak*	¬*lights*
soak	*lights*

Now, suppose you soak a match in water and then strike it. What happens? The conditional implies that it lights. Your knowledge implies that it does not light. Your knowledge, however, takes precedence over the possibilities that the conditional asserts.

Given the following premises in a form known as a syllogism,

All the Frenchmen are wine drinkers.
Some of the wine drinkers are gourmets.

the majority of reasoners draw the plausible conclusion:

Some of the Frenchmen are gourmets.

However, with the next premises, which are identical in form,

All the Frenchmen are wine drinkers.
Some of the wine drinkers are Spanish.

few reasoners draw the conclusion

Some of the Frenchmen are Spanish.

They envisage the possibility in which the wine drinkers are of both nationalities, but they search more assiduously—and successfully—for a counterexample because this conclusion is preposterous. Hence, the main effect of beliefs on the process of reasoning is that they influence invalid inferences far more than valid inferences: People refrain from drawing unbelievable invalid conclusions.

The difficulty of coping with falsity and the effects of content come together in a well-known reasoning problem, Wason's selection task, which has been studied experimentally more than any other paradigm of reasoning. Table V presents two versions of the task, one with a neutral conditional and one with a deontic conditional concerning what is permissible. The difficulty of the version with the neutral conditional, "If a card has an 'A' on one side then it has a '2' on its other

Table V
Two Examples of Wason's Selection Task

A	B	2	3

1. The participants know that each card above has a letter on one side and a number on the other side. Their task is to select those cards that they need to turn over to discover whether the following conditional is true or false about the four cards:

If a card has an "A" on one side then it has a "2" on its other side.

Most people correctly select the "A" card, and some select the innocuous "2" card too. They fail to select the 3 card. However, if it has an A on its other side, the conditional is false.

Drinking	Not drinking	21 years	16 years

2. The participants know that each card above represents a person. One side states whether or not the person is drinking alcohol, and the other side states the age of the person. The task is to select those cards that need to be turned over to discover whether or not a person is violating the following conditional rule:

If persons are drinking then they are over the age of 20 years.

Most people correctly select the "Drinking" card and the "16 years" card.

side," arises from the participants' inability to base their selections on the possibility that falsifies the conditional:

$$A \quad \neg 2$$

They need to choose those cards that could be instances of this case, i.e., A and 3 (which is an instance of $\neg$ 2). With neutral conditionals, reasoners appear merely to select cards on the basis of their mental models of the conditional rather than its falsifying instance.

When the selection task concerns what is permissible or impermissible, such as breaking a social contract, then reasoners tend to make the correct selections (Table V, problem 2). Some psychologists argue that this version of the task maps onto mental schemas with a content that concerns such deontic matters. Evolutionary psychologists propose that social contracts mattered to our hunter–gatherer ancestors, and so an innate module evolved for reasoning about cheaters. What appears to be the case, however, is that any experimental manipulation that helps reasoners to envisage false instances of conditionals improves their performance in the selection task. Knowledge of cheating is just such a cue, but there are others. Experiments have shown, for example, that instructions to check for violations or to envisage counterexamples improve performance. The context of a conditional can also exert such effects. One study enhanced the participants' selections with a neutral conditional, "If A then 2." The participants were told that it was a rule followed by a machine that prints cards. The machine went wrong, and now the participants must check that it is printing out cards correctly.

Reasoners are sensitive to the likelihood of encountering potential counterexamples, and so some theorists have introduced probabilistic considerations into their analyses of the selection task. They defend a normative approach of this sort, arguing that participants rationally seek to maximize the expected gain in information from selecting a card. If they were testing in the real world, the following conditional:

If a creature is a raven then it is black.

it would make sense to examine creatures that are black because there are many fewer black than nonblack creatures. Hence, the argument goes, people are rational in selecting 2 rather than 3 to test the neutral conditional. In one study, however, participants were each paid 1000 pesetas (about $7) before carrying out the selection task with a neutral conditional. They were charged 250 pesetas for each card that they selected, but they were told that they could keep whatever money they did not spend provided that their evaluation of the conditional was correct. This incentive failed to improve performance. Likewise, individuals with higher SAT scores tend to do better on the selection task than those with lower scores.

Whatever "rational" is taken to mean, it seems inappropriate to apply it to those who lose money rather than gain it and to those who score lower on tests of cognitive ability.

The five sorts of phenomena reviewed in this section were all predicted by the model theory, although readers will need to consult the literature for the derivation of the predictions. Formal rule theories allow that knowledge and beliefs can affect the interpretation of premises; otherwise, the phenomena are difficult for these theories to accommodate.

III. IMPLICIT INFERENCES

Psychologists distinguish between the deliberative thinking that underlies deduction and the implicit, automatic, and largely unconscious inferences that help people to make sense of the world and its descriptions. Consider, for example, the following passage:

The pilot put the plane into a stall just before landing on the strip. He just got it out of it in time. It was a fluke.

Readers have no difficulty in understanding the passage, but every noun and verb in the first sentence is ambiguous. Also the search for the referents for the three occurrences of the pronoun "it" in the passage defeats even the most advanced computer programs for interpreting natural language. Humans have no difficulty with the passage because they are equipped with a powerful system that uses general knowledge to make implicit inferences. Readers should also have no difficulty in understanding the following passage:

Apart from her husband, a hairdresser, Eve was the only woman among 52 men on the tour. As a costumier, she filled a much needed gap, because when a company of actors is putting on a play in a different town each night, no damage to the costumes is too trivial not to be mended.

In fact, most people do not notice that the passage contains three deliberate mistakes. It implies that Eve's husband is a woman. It states that what is needed is a gap rather than Eve. It also asserts that no damage to the costumes is too trivial not be mended instead of what it surely means—no damage to the costumes is too trivial to be mended. The system of implicit inferences overrides the literal interpretation of the sentences and makes sense out of nonsense. The inferences resolve the senses of words and determine the references of pronouns and other such expressions. They enable individuals to construct a single model of the situation described in a passage, and the implicit system does not attempt to search for alternative models unless it encounters evidence for them. The process is therefore rapid, and it becomes as automatic as any other cognitive skill that calls for no more than a single mental representation at a time. For the same reason, implicit inferences lack the guarantee that their conclusions are valid. They are inductions rather than deductions. However, the implicit system is not isolated from the mechanisms of deduction. Normally, the two systems work together in tandem.

One consequence of implicit inferences is that people often jump to a conclusion, which later they have to withdraw. In logic, if a conclusion follows validly from premises, then no additional premises can invalidate it. Logic means never having to be sorry about a conclusion. As new premises are added to existing premises, then increasing numbers of logical conclusions follow (i.e., logic is "monotonic"). However, in daily life, conclusions are often withdrawn in the light of subsequent information. These inferences are "nonmonotonic". The original conclusion may have been based on an assumption made by default that turned out to be false. For instance, I tell you about my cat Hodge, and from your knowledge of cats you infer that Hodge has fur and a tail. You withdraw your conclusion, however, when you learn that Hodge is bald and tailless. Your knowledge contains various assumptions that you can make in default of information to the contrary. The whole purpose of these default assumptions is to allow you to make useful inferences that you can withdraw in the light of contrary evidence.

A more problematic sort of nonmonotonic reasoning is illustrated in the following example. You believe the following premises:

If Viv has gone shopping then she will be back in an hour.
Viv has gone shopping.

It follows, of course, that Viv will be back in an hour. However, suppose that Viv is not back in an hour. You are in a typical everyday situation in which there is a conflict between the consequences of your beliefs and the facts. At the very least, you have to withdraw your conclusion. You also have to modify your beliefs, but in what way? Should you cease to believe that Viv went

shopping or that if she went shopping she will be back in an hour, or both? Philosophers and students of artificial intelligence have made various proposals about these puzzles. Unfortunately, the understanding of nonmonotonicity in human reasoning lags behind.

Reasoning in daily life often calls for the generation of explanations and diagnoses. For example, in the case of Viv's failure to return, you do not merely modify your beliefs, you try to make diagnostic inferences about what happened:

Possibly, Viv met a friend and went for a coffee.
Possibly, Viv felt ill on the way to the shops.

One possibility leads in turn to further explanatory possibilities, for example,

Possibly, Viv couldn't get the car to start after shopping.
∴ Possibly, the car's battery is dead. Possibly, Viv left the headlights on.

You use your knowledge and any relevant evidence to generate possibilities. Human reasoners easily outperform any current computer program in envisaging putative explanations. Given two sentences selected at random from different stories, such as

Celia made her way to a shop that sold TV sets.
She had recently had her ears pierced.

they readily offer such explanations as Celia was getting reception in her ears and wanted the TV shop to investigate, or Celia had bought some new earrings and wanted to see how they looked on closed-circuit TV. This propensity to generate explanations underlies both science and superstition. The difference is that scientists test their explanations empirically.

Inferences in real life are often not deductively "closed"—that is, there is not enough information to draw a valid conclusion. Reasoners must therefore make inductions, that is, they use their knowledge to draw conclusions that go beyond the information given and that therefore may be false. There is no normative theory of induction and no comprehensive psychological theory of it, either. What does exist are a number of well-established heuristics, which were identified by two pioneers, Kahneman and Tversky. One heuristic is the availability of relevant knowledge. Most individuals, for example, judge that more people die in automobile accidents than as a result of stomach cancer. They are wrong, but the media publish more stories about auto accidents than about stomach cancer. Similarly, people rely on the representativeness of evidence. If you are told that Bill is intelligent but unimaginative and lifeless, then you are unlikely to judge that he plays jazz for a hobby, though you may find it more likely that he is an accountant who plays jazz for a hobby. If so, you have violated the principle that a conjunction (being an accountant and playing jazz) cannot be more probable than one of its components (playing jazz). The description of Bill, however, is more representative of an accountant than of a jazz musician. It has therefore led you to overlook a simple principle of probability.

IV. REASONING AND THE BRAIN

The famous Russian neuropsychologist Luria once remarked, "The cerebral organization of thinking has no history whatsoever." Fodor, the distinguished philosopher of mind, predicted that it has no future either because thinking depends on general processes rather than separate brain modules, such as those that underlie perception or motor control. Nevertheless, a start has been made in the study of the neuropsychology of reasoning. The results so far have been largely at the level of "these areas of the brain underlie reasoning," and their interpretations are at best tentative.

A. Logical Reasoning and Personal Reasoning

Clinical studies in the early 20th century often reported the loss of "abstract thinking" as a result of brain damage. Such accounts, however, suffered from two irremediable problems. On the one hand, they never succeeded in characterizing a principled difference between abstract and concrete thinking. On the other hand, they failed to pin down the particular effects of lesions in different parts of the brain. This shortcoming is understandable given that many regions of the brain are likely to underlie reasoning. Modern neuropsychological investigations suggest that the real distinction is between logical reasoning with neutral materials and personal reasoning that engages individuals' beliefs and knowledge (Table V). Some studies suggest that logical reasoning depends on the left cerebral hemisphere, whereas personal reasoning implicates the right hemisphere and bilateral ventromedial frontal cortex. Positron emission tomography scans show greater left hemisphere activity when individuals evaluate syllogisms, such as

All men have sisters.
Socrates was a man.
∴ Socrates had a sister.

or judge the plausibility of inductive inferences, such as

Socrates was a great man.
Socrates had a wife.
∴ All great men have wives.

The control task was to judge how many of the sentences had people as their subjects. The effects of brain damage also appear to support the dissociation between logical and personal reasoning. For example, left hemisphere lesions impair simple relational inferences, such as

Mary is taller than John.
John is taller than Anne
Is Mary taller than Anne?

People who live in nonliterate cultures are happy to carry out personal reasoning, but they balk at logical reasoning when the content is outside their experience. Analogous effects have been obtained using electroconvulsive therapy (ECT), which suppresses cortical activity for 30 min or more. Before ECT, the patients (depressives and schizophrenics) tended to justify their responses to deductive problems on logical grounds. They also did so more rapidly and confidently after ECT had suppressed their right hemispheres. However, after the suppression of their left hemispheres, they tended to respond on grounds of personal experience in ways similar to members of nonliterate cultures, often rejecting a logical task based on unfamiliar content as impossible because it was outside their knowledge. Similar effects of brain damage occurred in a study of the selection task with a neutral conditional (Table V). Patients with left hemisphere damage, like control subjects, tended to err in the characteristic way. Surprisingly, however, half the patients with right hemisphere damage made the correct selections.

Perhaps the right hemisphere impedes logical reasoning because it allows knowledge and probabilistic considerations to influence performance. Certainly, the right hemisphere seems to play a role in automatic implicit inferences. Given the passage,

Sally approached the movie star with pen and paper in hand. She was writing an article about famous people's views about nuclear power.

normal individuals are likely to infer that Sally wanted to ask the star about nuclear power. Patients with damage to the right hemisphere infer that Sally wanted the movie star's autograph. They are misled by the first sentence and cannot make the implicit inference from the second sentence to revise their interpretation. Patients who have had a right-hemisphere lobectomy are also poorer at reasoning from false premises than those with a left hemisphere lobectomy. In general, right hemisphere damage seems to impair the ability to "get the point" of a story, to make implicit inferences establishing coherence, and to grasp the force of indirect illocutions such as requests framed in the form of questions.

It is tempting, but erroneous, to conclude that the left hemisphere is the seat of logic, whereas the right hemisphere is the seat of personal reasoning. Damage to the right hemisphere can lead to semantic difficulties in the interpretation of words, and so it may also impair the comprehension of discourse. For instance, it impairs the deduction of converse relations, such as

John is taller than Bill.
Who is shorter?

A recent functional magnetic resonance imaging (fMRI) study confirmed the existence of dissociable networks for logical and personal reasoning, which share circuits in common in the basal ganglia, cerebellum, and left prefrontal cortex. However, the activation suggested that personal reasoning recruits the left hemisphere linguistic system, whereas logical reasoning—even in inferences of an identical form—recruits the parietal spatial system. Also, when reasoning elicits a conflict between logic and belief, right prefrontal cortex becomes active, perhaps to resolve the incongruency. Another recent fMRI study established that deductive reasoning activates right dorsolateral prefrontal cortex whereas mental arithmetic from the same premises does not. This study also showed that when an inference depends on a search for a counterexample then the right frontal pole is activated.

Frontal cortex plays a crucial role in decision making, as shown in a major series of studies carried out by Damasio and colleagues. They also investigated the selection task in testing the consequences of their somatic marker hypothesis. This hypothesis postulates that ventromedial frontal cortex underlies the typical "gut reaction" on which implicit everyday decisions rely. Considerable evidence supports this hypothesis: For example, individuals with frontal lesions tend to

go bankrupt in real life and in laboratory gambling tasks. Similarly, the investigators found that patients with lesions in ventromedial frontal cortex were unaffected by whether the selection task was based on familiar or unfamiliar neutral contents. However, patients with lesions in other areas, like normal individuals, showed the characteristic effects of content. Correct performance in the selection task depends on grasping what counts as a counterexample to the conditional assertion.

B. Imagery and Spatial Representations

Does deductive reasoning rely on visual imagery? Behavioral studies have produced little evidence to suggest this is the case. Readers might suppose that this lack of evidence counts against the model theory. This view, however, confuses models with images. The model theory distinguishes between the two: Mental models are structural analogs of the world, whereas visual images are the perceptual correlates of certain sorts of model from a particular point of view. Indeed, many mental models are incapable of supporting visual images because they represent properties or relations that are not visualizable, such as ownership, obligation, and possibility. Recent studies have sharpened the need to distinguish between the degree to which relations evoke spatial models as opposed to visual images. The studies examined three sorts of materials, as rated by an independent panel of judges:

1. Relations that are easy to envisage spatially and easy to visualize, such as above, below, in front of, and in back of
2. Relations that are not easy to envisage spatially but are easy to visualize, such as cleaner, dirtier, fatter, and thinner
3. Control relations that are neither easy to envisage spatially nor easy to visualize, such as better, worse, smarter, and dumber

The studies examined both conditional inferences and inferences about simple relations among entities. They showed that inferences were faster with contents that were easier to envisage spatially than with the control contents, which in turn were faster than contents that were easy to visualize but difficult to envisage spatially. It seems that a relation such as "dirtier", elicits a visual image, but one that is irrelevant to the construction of a mental model that allows reasoners to make the required inference. In contrast, a relation, such as "in front of" elicits a spatial model that helps individuals to draw the inference. An fMRI study has also examined spatial reasoning. Given spatial problems, such as

> *The red rectangle is in front of the green rectangle.*
> *The green rectangle is in front of the blue rectangle.*
> *Does it follow that the red rectangle is in front of the blue rectangle?*

significant activation occurred in regions of parietal cortex that are known to represent and to process spatial information. Moreover, there was no reliable difference in the degree of activation between the right and the left hemispheres. Clinical studies of how brain damage affects the use of imagery in reasoning have produced mixed results, perhaps because they have not separated the two sorts of contents—spatial and nonspatial—that are both easy to visualize.

In summary, clinical and imaging studies of the brain have yet to establish how reasoners make deductions. There is evidence for separate systems mediating logical inferences with neutral content and personal inferences with a content that engages knowledge and beliefs. Future studies may determine whether separate brain mechanisms underlie the control of different deductive strategies, the use of diagrams as opposed to verbal premises, and the construction and evaluation of multiple models.

V. CONCLUSIONS

Modern logic has developed both proof theory and model theory for systems powerful enough to cope with all the deductive inferences that human beings make. What is lacking is a systematic method for translating such inferences into formal logic. Psychologists continue to investigate deductive reasoning. Their two main theoretical accounts are based on rules of inference and on mental models, respectively—a distinction that parallels the one between proof theory and model theory in logic. Rule theorists emphasize the automatic nature of simple deductions and postulate rules corresponding to them. More complex inferences, they assume, call for sequences of simple deductions. In contrast, model theorists emphasize that reasoning is the continuation of comprehension by other means. The system for implicit inferences based on knowledge aids the process of constructing models of discourse. In deliberative reasoning,

individuals tend to focus on possibilities in which the premises are true. However, they can grasp the force of counterexamples. The evidence suggests that people have a modicum of deductive competence based on mental models. Rules of inference and mental models, however, are not incompatible. Advanced reasoners may construct formal rules for themselves—a process that ultimately leads to the discipline of logic.

See Also the Following Articles

ARTIFICIAL INTELLIGENCE • CATEGORIZATION • CREATIVITY • INFORMATION PROCESSING • INTELLIGENCE • LANGUAGE AND LEXICAL PROCESSING • PROBLEM SOLVING

Suggested Reading

Baron, J. (1994). *Thinking and Deciding*, 2nd ed. Cambridge Univ. Press, New York.

Braine, M. D. S., and O'Brien, D. P. (Eds.) (1998). *Mental Logic*. Erlbaum, Mahwah, NJ.

Brewka, G., Dix, J., and Konolige, K. (1997). *Nonmonotonic Reasoning: An Overview*. CLSI Stanford Univ. Press, Stanford, CA.

Evans, J. St. B. T., and Over, D. E. (1996). *Rationality and Reasoning*. Psychology Press, Hove, UK.

Garnham, A., and Oakhill, J. (1994). *Thinking and Reasoning*. Blackwell, Cambridge, MA.

Jeffrey, R. (1981). *Formal Logic: Its Scope and Limits*, 2nd ed. McGraw-Hill, New York.

Johnson-Laird, P. N. (2001). Mental Models and deduction. *Trends in Cognitive Scien.* **5,** 434–442.

Johnson-Laird, P. N., and Byrne, R. M. J. (1991). Deduction. Erlbaum, Hillsdale, NJ.

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M., and Caverni, J.-P. (1999). Naive probability: A mental model theory of extensional reasoning. *Psychol. Rev.* **106,** 62–88.

Kahneman, D., Slovic, P., and Tversky, A. (Eds.) (1982). *Judgment under Uncertainty: Heuristics and Biases*. Cambridge Univ. Press, New York.

Oaksford, M., and Chater, N. (1998). *Rationality in an Uncertain World: Essays on the Cognitive Science of Human Reasoning*. Psychology Press, Hove, UK.

Rips, L. J. (1994). *The Psychology of Proof*. MIT Press, Cambridge, MA.

Schaeken, W., De Vooght, G., Vandierendonck, A., and d'Ydewalle, G. (2000). *Deductive Reasoning and Strategies*. Erlbaum, Mahwah, NJ.

Stanovich, K. E. (1999). *Who Is Rational? Studies of Individual Differences in Reasoning*. Erlbaum, Mahwah, NJ.

Wharton, C. M., and Grafman, J. (1998). Deductive reasoning and the brain. *Trends Cognitive Sci.* **2,** 54–59.

Lyme Encephalopathy

RICHARD F. KAPLAN
University of Connecticut School of Medicine

GLOSSARY

fibromyalgia A chronic musculoskeletal rheumatological syndrome that is associated with muscular pain and fatigue. Symptoms include cognitive difficulties, mood change, and sleep disturbance. The etiology is unknown.

intrathecal antibody production Local production of specific antibody to *Borrelia burgdorferi* in cerebral spinal fluid.

Lyme disease A multisystem disorder of the skin, nervous system, heart, and joints caused by the tick-borne spirochete, *B. burgdorferi*.

neuropsychological assessment The use of procedures and tests that measure general intellectual abilities and specific cognitive functioning, such as memory, to aid in the diagnosis of brain dysfunction.

polymerase chain reaction A laboratory assay for *B. burgdorferi* DNA.

post-Lyme disease syndrome A condition characterized by fatigue, mood disturbance, and complaints of cognitive dysfunction that follow well-documented Lyme disease months to years after antimicrobial treatment.

single photon emission computed tomography A functional neuroimaging technique used to measure regional cerebral blood flow. Brain activity is recorded by externally placed gamma cameras, which detect a radiotracer such as t-HMPAO.

Lyme encephalopathy is a neuropsychiatric disorder beginning months to years after the onset of Lyme disease. The symptoms are non specific and typically include fatigue, sleep disturbance, slowed thinking, memory loss, naming problems, and depression. The etiology remains controversial. This article defines the nature of the syndrome and offers possible explanations from a neuropsychological perspective.

I. HISTORY AND BACKGROUND

Lyme disease, originally termed Lyme arthritis, was first recognized in 1975 following an outbreak of arthritis in and around the rural community of Lyme, Connecticut. The association of the illness with a unique skin lesion, erythema chronicum migrans (ECM), and the close geographic clustering of these early cases suggested the disease might be an infection transmitted by an arthropod. Subsequent investigations revealed that the disease was transmitted by a newly identified tick *Ixodes dammini* or related ixodid ticks. A few years later, Willy Burgdorfer and colleagues isolated the spirochete, *Borrelia burgdoferi*, from a tick, *I. dammini*, and it was determined that this was the infectious organism. A disease similar to Lyme disease had been previously described in Europe and is now known to be caused by the same spirochete. Lyme disease is now recognized as a multisystem disorder. A skin lesion, ECM, is typically the first sign of infection, although not all patients infected with Lyme disease manifest this symptom. This can be followed by systemic manifestations involving the joints, heart, and nervous system. The diagnosis of active Lyme disease is not straightforward because, unlike some bacterial infections, *B. burgdorferi* is not easily cultured. Currently, the method recommended by the

Centers for Disease Control (CDC) for diagnosing the disorder includes physician-documented ECM or the detection of antibodies in blood serum, which indicate exposure to *B. burgdorferi*, and at least one late objective manifestation of the disease, such as musculoskeletal pain, chronic arthritis, myocarditis, cranial neuritis, or radiculoneuropathy. However, not all people who are infected have the typical clinical symptoms of Lyme disease, and for this relatively small group the diagnosis is sometimes questionable. This has made the diagnosis of some disorders that have been attributed to Lyme disease, such as Lyme encephalopathy, controversial. For most patients, Lyme disease can be successfully treated with antibiotic therapy. However, a small percentage of patients develop a mild to moderate chronic encephalopathy, a peripheral neuropathy, or both, which may occur months to years after disease onset.

II. NEUROLOGICAL MANIFESTATIONS

The classic neurological symptoms of early disseminated Lyme disease are meningitis, cranial neuritis, and radiculoneurotis. These occur alone or in combination in approximately 15% of untreated patients. A unilateral or bilateral facial nerve palsy is also a relatively common symptom of early disseminated disease. Symptoms usually last for weeks to months but can become chronic. Chronic neurologic manifestations of the disorder, which include encephalopathy, polyneuropathy, and leukoencephalopathy, usually occur late in the illness. Although there have been reports of cases with severe cognitive impairment, including psychosis and dementia, and vasculitic lesions, such cases are rare. A mild chronic Lyme encephalopathy (LE) is the most common neurologic symptom in patients with late-stage disease. Months to years after disease onset, sometimes following long periods of latent infection, a small percentage of patients develop a mild to moderate encephalopathy. The symptoms tend to be diffuse and nonspecific and can include memory loss, naming problems, sleep disturbance, fatigue, and depression.

The diagnosis of LE is difficult. It is generally believed that those with immunity to *B. burgdorferi* and abnormal cerebral spinal fluid (CSF) are more likely to have a neurological basis to their illness. Even if a patient is seropositive, it is difficult to know whether active infection is causing encephalopathy because serologic testing only indicates exposure to the *B. burgdorferi*, not active infection. The standard neurological examination is usually normal. The CSF examination may show a positive polymerase chain reaction to *B. burgdorferi* DNA, intrathecal production of antibody to *B. burgdorferi*, or increased CSF protein. Traditional brain neurophysiological and neuroimaging techniques have also not been shown to be highly sensitive to the pathophysiology of LE. The routine electroencephalograph is typically normal. Magnetic resonance imaging (MRI) abnormalities have been described in Lyme patients, but these are nonspecific and relatively infrequent. When MRI abnormalities are present they are usually white matter lesions, suggesting the possibility of an inflammatory process. Single photon emission computed tomography (SPECT) imaging has shown some promise in identifying brain abnormalities in Lyme patients. Because the spatial resolution of SPECT is relatively poor, it can be used together with MRI to provide information about metabolic activity in specific brain regions (Fig. 1). SPECT has proved sensitive in identifying pathophysiologic abnormalities in other neurobehavioral disorders such as Alzheimer's disease and other dementias. In 1997, Logigian and colleagues studied a series of LE patients using a quantitative SPECT technique. In their analysis, 10 transaxial slices were imaged from each brain and divided into 4320 macrovoxels (Fig. 2). Quantitative SPECT has an advantage over visual image analysis because regional radiotracer uptake is analyzed statistically for each macrovoxel using an analysis of covariance. Patient scans are compared macrovoxel by macrovoxel to normal subjects to determine which brain regions are hyperperfused.

The Lyme patients showed patterns of multifocal hypoperfusion, most notably in the subcortical areas including the basal ganglia and white matter of the cerebral hemispheres, which was not apparent in normal controls. Patients with more objective evidence of LE, including CSF abnormalities, demonstrated significantly lower cerebral perfusion than Lyme patients with encephalopathic symptoms without objective evidence of central nervous system (CNS) involvement. Studies using visual ratings to assess the reduction in regional uptake of radiotracer have similarly reported decreased perfusion in Lyme patients. The reduction of metabolic activity in frontal and temporal lobe structures may provide a clue to the neuroanatomic basis of LE. That these same white matter regions form the large-scale neurocognitive networks involved in mediating memory and attention also suggests a possible explanation for the cognitive

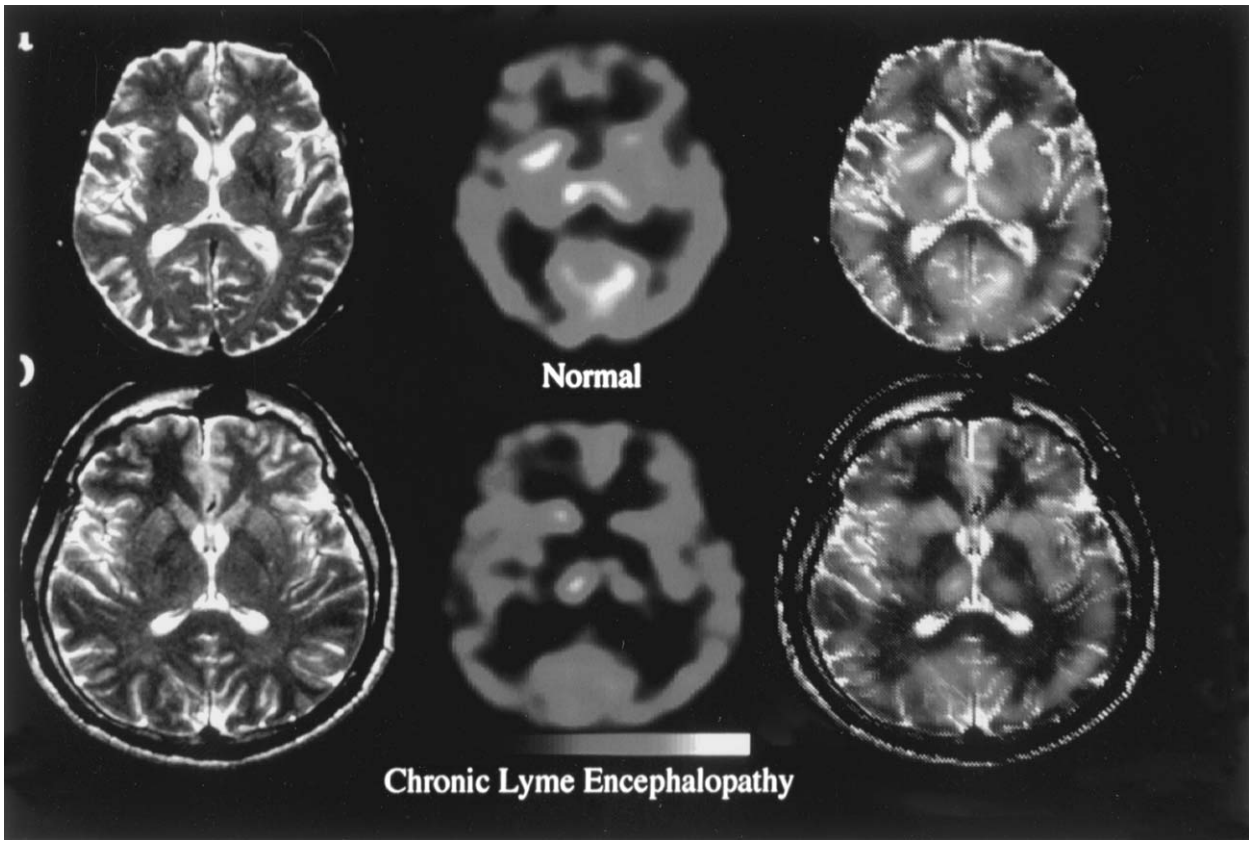


Figure 1 Representative MRI scan (left), SPECT image (center), and superimposed MRI/SPECT image (right) in a normal subject and in a patient with objective evidence of LE. The dark areas correspond to lower and the light areas to higher perfusion. (See color insert in Volume 1).

deficits described previously. Unfortunately, hypoperfusion in these brain regions is not specific to Lyme disease. Similar SPECT findings have been reported in other conditions that have symptoms similar to LE, such as depression and chronic fatigue syndrome. Although some investigators argue that the pattern of hypoperfusion in LE may differ from that seen in other disorders, current quantitative SPECT studies have not yet been done. Therefore, SPECT alone is not sufficient in diagnosing LE.

III. NEUROPSYCHOLOGICAL STUDIES

Although a number of studies that have attempted to define the relationship between positive serology, CNS infection, and cognitive dysfunction, few have used neuropsychological testing to characterize the nature of the cognitive deficits. Neuropsychological tests have an advantage over even comprehensive bedside mental status examinations in detecting cognitive dysfunction associated with LE because the deficits are often subtle. For most neuropsychological procedures, the presence or absence of functional impairment is made on a statistical basis. Ideally, performance on a given test is interpreted in comparison to normative data appropriate to the patient's age, sex, and education. Probable impairment on a specific cognitive test is inferred when performance falls below a designated cutoff score, usually two standard deviations below the normative mean. Whether or not a particular test will discriminate between normal and abnormal function for either individuals or clinical populations depends on a number of critical factors. Reliability and validity data are essential for understanding the limits of a given test. Additionally, the ability of a test to discriminate between groups depends on the purpose and population for which the test was developed. For example, tests such as the Mini-Mental State Examination that were developed to screen for dementia in elderly populations are relatively insensitive to cognitive deficits in patients with Lyme disease.

The traditional approach in neuropsychological assessment used to define cognitive deficits associated

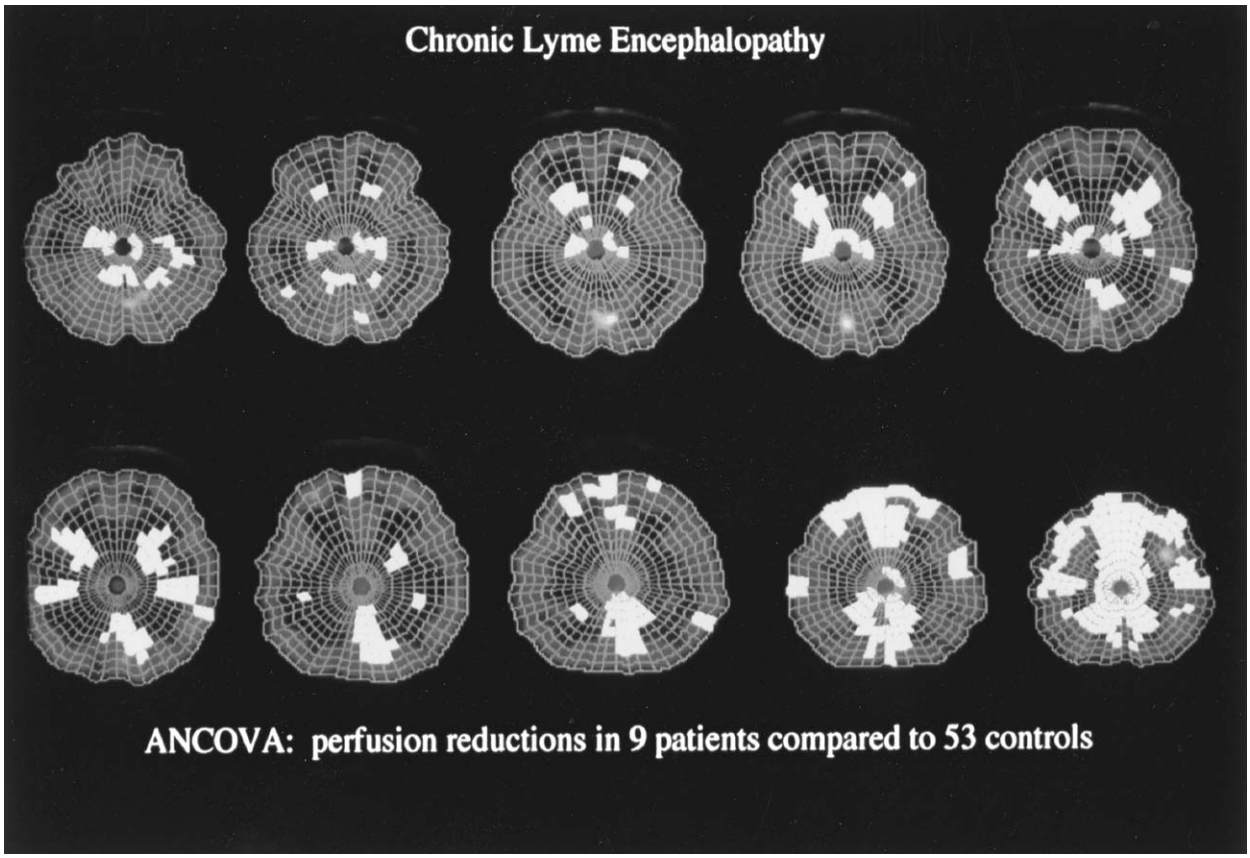


Figure 2 Quantitative SPECT analysis of nine patients with LE (mean data). Axial slices ascend from lowest (top left) to highest brain levels (bottom right). Macrovoxels whose mean perfusion is reduced at the $p < 0.001$ level below the mean of normal subjects ($n = 53$) are shown in white. The dark areas correspond to lower and the light areas to higher perfusion.

with brain disease is to administer a battery of tests. Often, differences in cognitive performance between two clinical populations, such as those with LE and those with depression alone, are best appreciated when the evaluation is based on an analysis of test score patterns. A pattern of scores not only indicates what is wrong with the patient but also what is right with the patient. The test battery usually includes both tests of general intellectual abilities, such as the Wechsler Scales, and tests developed to assess specific areas of cognitive dysfunction, such as attention, visual perception, construction, fine motor control, language, memory, and executive functioning. Test batteries also usually include measures of psychopathology in order to assess the patient's emotional state. The latter are often useful in interpreting other test results because psychological states such as anxiety and depression can impact a patient's cognitive abilities, particularly motivation and the ability to maintain and sustain attention. The importance of this cannot be overstated because although the sensitivity of neuropsychological tests in identifying brain dysfunction is quite high (between 80 and 90%), the specificity may be considerably lower. For example, a memory test that discriminates traumatic brain-injured patients from normal subjects does not indicate how important variables such as attention or affective states may impact performance. It is also incumbent upon the clinician or researcher to interpret neuropsychological test data in the context of the patients' medical and psychosocial history as well as psychiatric status. Table I provides a list of neuropsychological tests that have frequently been used in assessing cognitive deficits in Lyme disease patients.

In comparison to other more well-known brain disorders such as Alzheimer's disease, for which there are large neuropsychological databases, the number of neuropsychological studies in LE is relatively limited. The following discussion presents the neuropsychological test data in considerable detail in an attempt to

Table I
Neuropsychological Tests Frequently Used to Assess Deficits in Lyme Disease

Tested domain or function	Neuropsychological test
General cognitive abilities	Wechsler Intellignece Scales
	Shipley Institute of Living Scale
Attention/concentration	
Short-term auditory attention	Digit Span Test
Sustained attention	Stroop Test (Word and Color)
	Continuous Performance Tests
	Symbol Digit Substitution Test
Interference	Stroop Test (Color/Word)
Learning and memory	California Verbal Learning Test
	Selective Reminding Test
	Rey–Osterrieth Complex Figure Test
	Wechsler Memory Scale
Language and verbal output	
Naming	Boston Naming Test
Verbal fluency	Controlled Word Association Test
Executive functions	
Set maintenance and shifting	Trails A and B
Abstraction and set shifting	Wisconsin Card Sorting
Distractibility	Stroop Test
Problem solving	Halstead Category Test
Motor dexterity	Finger Tapping
	Grooved Peg Board
Neuropsychiatric symptoms	
Mood	Beck Depression Inventory
	Center for Epidemiologic Studies Depression Scale
	State Trait Anxiety Scale
Personality	Minnesota Multiphasic Personality Inventory
Anxiety	State Trait Anxiety Scale
Fatigue	Fatigue Severity Scale

demonstrate the nature and extent of cognitive dysfunction in LE.

In 1988, Halperin and colleagues published one of the earliest studies that included neuropsychological testing. They studied 17 Lyme patients before and after antibiotic treatment. All initially complained of memory deficits; however, only 4 had evidence of other neurological involvement and only 1 had abnormal CSF. The evaluation included tests of memory, visual spatial organization, conceptual thinking, and psychomotor and perceptual motor skills. The pretreatment results are difficult to interpret in part because these investigators failed to include a normal comparison group and also because there was no indication of the number of patients whose performance statistically fell outside the normal range. However, the authors report that there were pretreatment cognitive deficits in almost every area examined. Significant improvements following treatment were found on measures of recall memory on the California Verbal Learning Test (CVLT) as well as on tests of attention, concentration and motor speed, abstract problem solving, fine motor dexterity, and visual spatial organization. Improvement, however, was not evident on all tests. Patients' performances were unchanged on measures of tracking and sequencing, verbal fluency, auditory attention span, and recognition memory. Patients were also administered the Beck Depression Inventory (BDI), a self-administered questionnaire that measures the presence and severity of symptoms associated with depression. There was no evidence of significant depression either before or after treatment. However, the average score on the BDI, which reflects the overall severity of depression, was almost halved following treatment.

In a 1990 report, investigators at the New England Medical Center in Boston described 27 patients with chronic neurological abnormalities of at least 3 months duration following well-recognized manifestations of Lyme disease. Twenty-two of the 27 patients reported memory loss and 18 had CSF protein, evidence of intrathecal antibody to *B. burgdorferi*, or both. Each underwent a neuropsychological test battery consisting of tests of general cognitive abilities, memory, visual perception, construction, executive function, language, and fine motor dexterity. The patients were also administered the Minnesota Multiphasic Personality Inventory (MMPI) to assess emotional status. Standard scores were calculated from published, age-corrected normative data, and scores that were two standard deviations below the mean were considered as evidence of impairment on a particular test. Memory impairment was defined as two standard deviations below average on any one test or one standard deviation below average on two tests, according to a previously described system. All patients had IQ scores that were calculated or estimated to be average or above average. Twelve of the 22 patients reporting memory difficulty met the criteria for memory impairment, as did 2 patients who denied memory problems. Only 1 patient had a below average

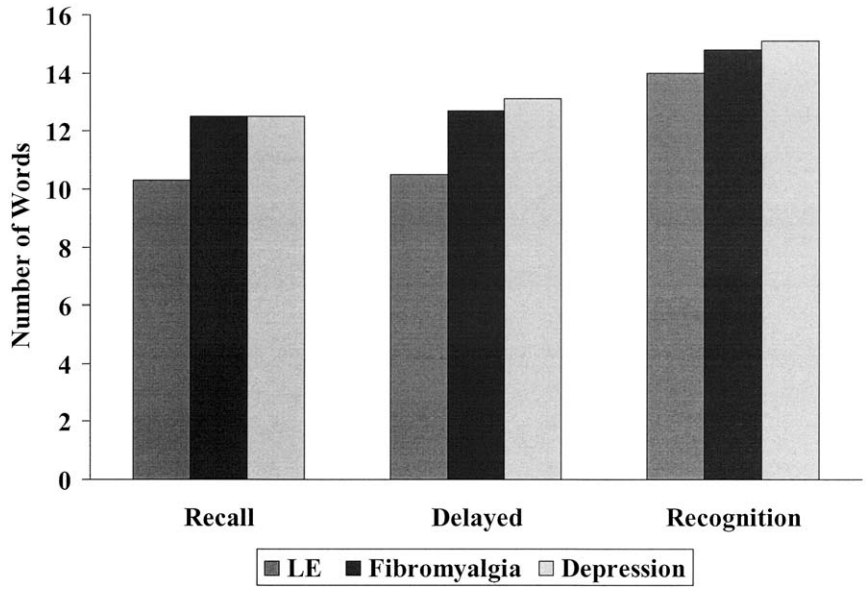


Figure 3 A comparison of memory scores for patients with LE, fibromyalgia, and depression on the CVLT.

performance on a test of abstract problem solving, the Wisconsin Card Sorting test, and 1 patient did poorly on a naming test. No patient had a below average performance on any of the other cognitive tests. Nine patients produced MMPI profiles consistent with depression. Of the 12 patients with objective evidence of memory impairment, 9 had abnormal CSF findings; however, it should be noted that 9 of 10 patients with normal memory scores also had abnormal CSF. The 4 of 24 patients with abnormal MRI scans, which were characterized as small round lesions in the periventricular white matter, all had impaired memory and 3 of the 4 had abnormal CSF.

In contrast to the earlier study by Halperin and colleagues that described a wide range of cognitive impairment, the only significant deficits in this study were in the area of memory impairment. However, unlike the patients in Halperin's study, more than half of the patients in this study had been treated with one or more courses of antibiotic treatment prior to testing. It is therefore possible that the discrepancy in results between the two studies reflects partial recovery.

Many of the symptoms of LE are similar to those reported by patients with depression. These include irritability, fatigue, emotional liability, poor concentration, impaired sleep, and memory disturbance. Depression is also common in late-stage Lyme disease. To examine the impact of psychological factors—particularly somatic concerns, depression, and anxiety—on memory impairment in LE patients, my colleagues and I compared 20 patients with well-documented Lyme disease to patients with fibromyalgia and patients with mild to moderate depression. The rationale for comparing fibromyalgia and depressed patients with LE patients was that the cognitive symptoms in these illnesses are similar. Thirteen of the 20 Lyme patients had abnormal CSF and 2 had abnormal MRI scans. Memory was assessed using the three standardized memory instruments—the CVLT, the Wechsler Memory Scale (WMS), and the Rey–Osterrieth Complex Figure. Indices of psychopathology were obtained from the MMPI and BDI. The depression and fibromyalgia groups performed significantly better than the LE group on both the CVLT (Fig. 3) and a WMS visual memory subtest, whereas recognition memory on the CVLT did not differ. The depression group also performed better than the Lyme group on the WMS Verbal Paired Associate Learning test, another verbal learning test. In contrast, both the fibromyalgia and depressed groups showed greater evidence of psychopathology. On the MMPI, the Lyme group had significantly lower scores on the scales most sensitive to depression and anxiety compared to those of the depression group and significantly lower scores on scales sensitive to the somatic concerns compared to those of the fibromyalgia group (Fig. 4). The Lyme group also had lower BDI scores, but this did not reach statistical significance. These data strongly suggest that memory impairment in LE

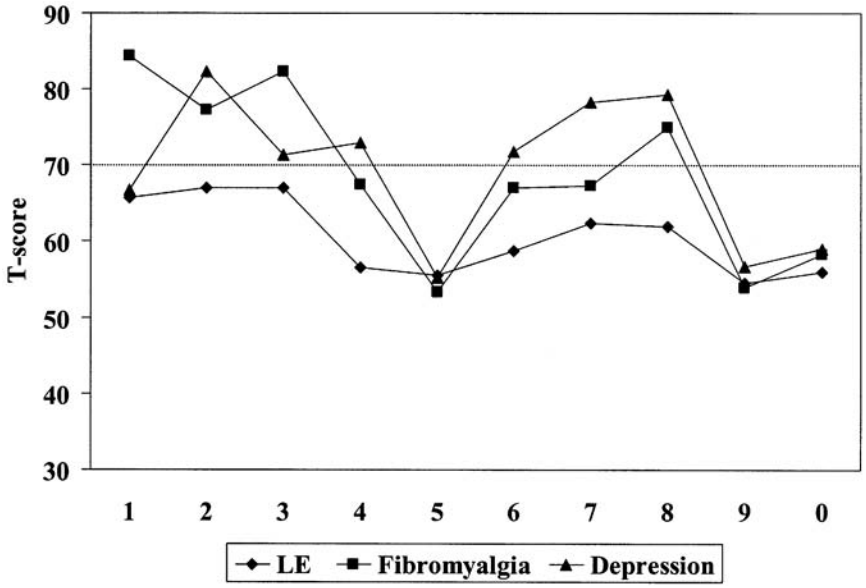


Figure 4 Composite MMPI for LE, fibromyalgia, and depression patient groups.

cannot be explained by affective symptoms alone, even though Lyme patients may report more symptoms of depression than normal controls.

If there is a pattern that emerges from the investigations outlined previously, it is that a memory disturbance appears to be the most common cognitive deficit found in LE patients. Other cognitive domains, such as attention, psychomotor performance, and executive functioning, were reported to be impaired in some but not all studies, even when observed problems in these other areas appeared to be of a lesser magnitude than the memory findings. Memory for verbal material appeared to be more affected than that for nonverbal material; however, this may be more a function of the type of test used because the verbal and nonverbal memory tests used are not of equal difficulty. The verbal list learning tests require acquisition of material over multiple trials and greater use of mnemonic strategy than the nonverbal learning tests (e.g., Rey–Osterrieth Complex Figure). It is also thought that tasks requiring sustained effort, such as the CVLT, may be more vulnerable to the effects of fatigue, and Lyme patients do report greater fatigue than controls. Furthermore, there is no anatomical reason why one would expect a material-specific memory deficit in this population. Recognition memory was spared relative to recall memory in every study. This is of interest because MRI and SPECT studies suggest that the pathophysiology of *B. burgdorferi* primarily affects CNS white matter. The pattern of memory deficits in LE parallels that of another white matter disease, multiple sclerosis (MS). On list learning tests such as the CVLT, MS patients have been shown to have lower initial acquisition rates but similar rates of learning and forgetting when compared to healthy normals. The same pattern is true for LE.

IV. POPULATION STUDIES

In the laboratory studies of the LE described previously patients were selected on the basis of neurologic complaints. Although these types of studies have been helpful in understanding the nature and severity of neurologic deficits in Lyme disease, they tell us little about the prevalence of these symptoms. Nancy Shadick and colleagues studied the prevalence of persistent neurologic symptoms in a sample of unselected patients with a history of Lyme disease in Ipswich, Massachusetts, a community endemic for Lyme disease. These investigators initially studied 38 people who met CDC criteria for previous Lyme disease and 43 people who did not. There was a slight but statistically significantly difference between groups on the CVLT, with the Lyme disease group performing more poorly. Twelve of the 38 Lyme patients scored two or more standard deviations below mean on the word list test, compared to only 5 of the 43 controls. Patients with residual symptoms, neurologic and

musculoskeletal, had longer duration of disease prior to treatment. These findings suggested that a small percentage of patients with previous Lyme disease may have permanent learning and memory deficits, albeit subtle. In a larger subsequent study on Nantucket, another community highly endemic for Lyme disease, these investigators compared 186 people with prior Lyme disease to 167 healthy controls using similar measures including the CVLT. Patients were studied an average of 6 years after infection. Although the patient group reported a higher incidence of nonspecific symptoms, including fatigue, difficulty sleeping, memory impairment, and poor concentration, there were no significant differences between groups on any of the objective tests of memory and concentration. At least two other population studies produced similar findings—namely, that the prevalence of objective measures of LE in previously infected patients who were treated for Lyme disese is very low. Thus, although patients previously infected with Lyme disease may report more neurologic symptoms than never infected controls, there is little evidence of any objective deficits. The relationship between reports of perceived memory dysfunction and performance on memory tests is discussed later.

V. LE IN CHILDREN

Although Lyme disease was first described in children, most studies of LE have involved adults. The incidence of Lyme disease is actually higher in children than in adults, but there are fewer reports of long-term problems. Even in children with Lyme arthritis, who were not treated or were treated years after the onset of infection, the incidence of encephalopathy was low. It has also been noted that children with Lyme disease almost never have the nonspecific symptoms of LE as the sole manifestation of the illness. Wayne Adams and colleagues studied 41 children, aged 6–17, with well-documented Lyme disease. All had been treated with antibiotics and the testing occurred about 2 years after disease onset. The children, 23 healthy sibling controls and 14 with subacute rheumatological diseases, were compared on a battery of standard neuropsychological tests, academic achievement tests, and parental rating scales. The neuropsychological test battery included measures of intelligence, information processing speed, fine-motor dexterity, executive functioning, and memory. The parents were also asked to complete a questionnaire rating perceived changes in their child's behavior in the areas of school, personality, family, and friends since the onset of Lyme disease. Absentee records and the results of standard academic achievement tests were obtained from the schools. No significant group differences were found on any of these measures. Moreover, among children with Lyme disease, those with early neurologic involvement did not differ on any measure from those without any neurologic symptoms. The authors concluded that although there may be individual cases of Lyme-related cognitive abnormalities, children who are properly diagnosed and properly treated are unlikely to have any long-term negative consequences from Lyme disease. Since there is evidence that LE in adults may not occur for years after the onset of the illness, these same authors reexamined 25 of the their original sample and 17 sibling controls 4 years later. Again, there was no cognitive impairment in the Lyme disease children.

VI. POST-LYME DISEASE SYNDROME

The terms post-Lyme disease syndrome and post-treatment chronic Lyme disease are used to describe patients who were previously diagnosed with Lyme disease and develop persistent encephalopathic symptoms months to years after antibiotic treatment. Not all patients with a clinical history of Lyme disease who later present with encephalopathic symptoms are seropositive or have abnormal CSF. Most patients who are seronegative do not have Lyme disease, although it is possible to be seronegative and have post-Lyme syndrome. There are also reported cases of seronegative patients with CSF evidence of Lyme infection; however, this is rare. Most patients diagnosed with post-Lyme syndrome are seropositive for Lyme disease and develop symptoms after treatment.

In 1991, Krupp and colleagues compared a group of treated Lyme disease patients who continued to report memory problems after treatment, to a group of healthy controls. All subjects received a similar neuropsychological test battery to those described previously, with the addition of a fatigue rating scale. There were no differences between patients and controls on the any of the intelligence tests, tests of tracking and sequencing, and tests of abstract problems solving. The Lyme disease group performed significantly worse on the verbal fluency and two memory measures. Patients were also more depressed than controls. Neuropsychological test scores remained lower in the Lyme group, relative to controls, even when the presence of depression was controlled

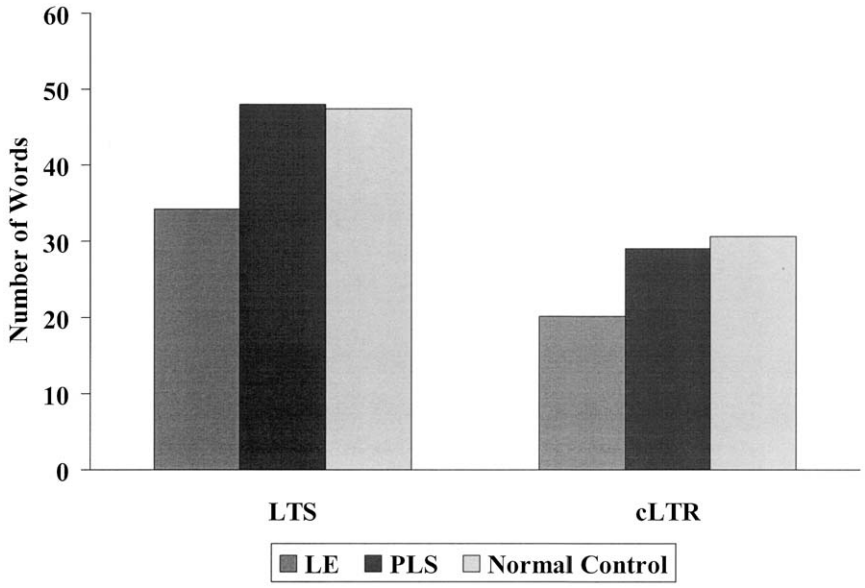


Figure 5 A comparison of memory scores for patients with LE, post-Lyme disease syndrome (PLS), and normal controls on the Selected Reminding Test. The two measures compared are total words recalled (LTS) and consistency of recall (cLTR).

statistically. Impairment was greatest for word recall on a verbal list learning test, the Selective Reminding Test (SRT). However, recognition memory did not discriminate patients and controls. The Lyme patients were also divided by the severity of neuropsychological impairment into mild, moderate, or severe categories based on the number of tests in which a patient's score fell below the published normative means (one or two standard deviations). On this basis, 9 of the 15 patients showed evidence of mild to moderate cognitive impairment, whereas 6 patients appeared normal despite complaints of memory trouble. Neither evidence of CSF infection nor MRI abnormalities correlated with the degree of cognitive impairment. However, fatigue was correlated with memory performance in the Lyme group but not the control group, with higher levels of fatigue associated with greater memory impairment. Interestingly, depressive symptoms were also correlated with memory scores. Paradoxically, however, more depressive symptoms corresponded to better memory scores.

Because it appeared that cognitive deficits exist in some Lyme patients independent of other evidence of neurological involvement, my colleagues and I studied the relationship between active infection and cognitive dysfunction. We compared 13 seropositive Lyme patients with evidence of inflammatory CSF to 20 seropositive Lyme patients without evidence of neurological disease and 14 age-matched normal controls. Most patients in both Lyme groups described memory deficits as one of their symptoms. Eleven of 13 patients with abnormal CSF and 16 of 20 with normal CSF reported memory problems, versus none of the controls. On the SRT verbal memory measure, Lyme patients with abnormal CSF recalled significantly fewer words, were less consistent in their recall, and retrieved fewer items from memory than either Lyme patients with normal CSF or normal controls, who did not differ on any of these measures (Fig. 5). On the BDI, both Lyme disease groups endorsed significantly more symptoms than the normal control group, although no group's score was sufficiently high to meet the criteria for depression. Thus, although both groups of Lyme reported memory difficulties in our study, only the group with evidence of CNS infection performed more poorly on objective memory testing. However, both Lyme groups reported more symptoms of depression and anxiety than healthy controls (Fig. 6). These data suggest that depression may be a factor in perceived memory loss.

As a group, Lyme patients report significantly more symptoms of depression and fatigue than do controls, independent of evidence of CNS disease. In several studies the average depression scores for the Lyme patients were typically higher than for healthy controls but below the suggested cutoffs for clinical depression.

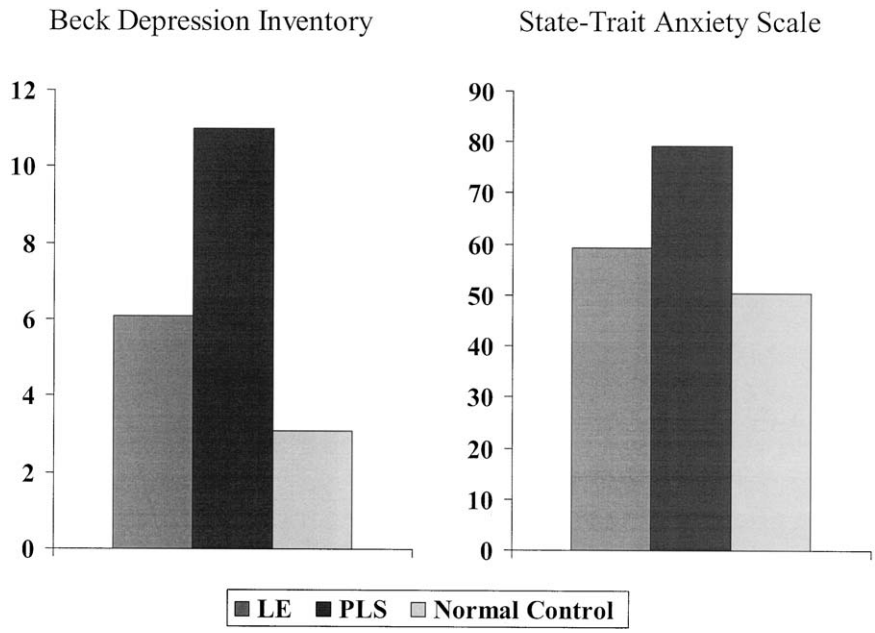


Figure 6 Depression and anxiety for LE, post-Lyme disease syndrome (PLS), and normal controls.

Moreover, although a significant proportion of Lyme patients meet the criteria for depression, depression scores and objective measures of memory loss have not been highly correlated. Fatigue, however, has been shown to be significantly correlated with poor memory in a number of studies.

In the study described previously, most Lyme patients complained of memory disturbance whether or not they had abnormal CSF. As such, the perception of memory loss was not a good predictor of objective impairment on neuropsychological testing. This finding has been replicated in several studies of Lyme disease patients. It is also not unique to Lyme disease. In most studies, self-reported poor memory is only weakly related to actual memory test performance, whereas it is strongly associated with conventional affective symptoms. Most studies have been done with elderly patients. These have shown that the relationship between subjective estimates of memory loss and objective measures of memory impairment is low. Instead, perceived memory loss tends to be related to psychological distress. As with many chronic illnesses, Lyme patients with late-stage disease often experience greater emotional distress, including depression and anxiety, than otherwise healthy people, although they do not meet the clinical criteria for psychopathology. It is therefore likely that the perceived memory disturbance in many Lyme patients may be related to the stress and affective symptoms common to many chronic diseases.

VII. CONCLUSIONS

LE was originally used to describe mild confessional state—most commonly fatigue, sleep disturbance, memory loss, and depression—in patients with active systemic Lyme disease, particularly active arthritis. However, its meaning has become more general and is often used to describe patients with similar deficits but without laboratory confirmation of active Lyme infection. This has made the diagnosis of some disorders that have been attributed to Lyme disease highly controversial. In cases in which the patient has had the characteristic clinical picture, a positive serology to *B. burgdorferi*, and abnormal CSF, the symptoms are likely due to active infection. Moreover, there is evidence, from MRI and SPECT studies, that the infection preferentially affects white matter areas of the brain, although the mechanism is not known. Neuropsychological testing in these patients typically shows mild, but statistically significant, deficits on memory testing. After adequate antimicrobial therapy, most patients show significant improvements in their cognitive functioning and return to normal. In the one quantitative SPECT study, demonstrating a

reduction of cerebral perfusion in patients with LE, there was also a partial reduction of hypoperfusion with antibiotic therapy.

The controversy regarding the diagnosis of LE stems from patients who report encephalopathic symptoms without clear evidence of an active Lyme infection, namely those with post-Lyme disease syndrome. Although some clinicians still attribute this to a subacute *B. burgdorferi* infection and recommend extended antimicrobial therapy, others question whether these symptoms are caused by active brain infection. Some neuropsychological investigations have shown these patients to perform more poorly than healthy normals, whereas others have not. Neuropsychological deficits in post-Lyme syndrome patients seem to be independent of a history of psychopathology but are correlated with concurrent fatigue. Lyme patients with independent evidence of CNS infection perform significantly worse on neuropsychological testing than post-Lyme disease syndrome patients. For patients without clear evidence of active Lyme disease, there is little efficacy to additional antimicrobial therapy. This has led some researchers to conclude that LE is probably overly diagnosed, and the lack of response to antibiotic therapy is the result of misdiagnosis. In addition to latent infection, other explanations for post-Lyme disease include the psychological consequences of chronic illness, residual deficits from past infection, and possibly an immune response. It has also been suggested that a fibromyalgia syndrome may be triggered by infection to *B. burgdorferi* after the infection is successfully treated since both disorders can result in similar cognitive complaints.

Patients with LE typically report memory problems independent of their performance on objective testing. The perception of cognitive difficulties may be common to a variety of physiological and psychological disorders. In chronic Lyme disease, patients with objective evidence of CNS infection, such as abnormal CSF, probably have a neurological basis to their reported cognitive decline. Other chronic Lyme patients are likely experiencing the stress and affective symptoms common to many chronic illnesses and similarly the perception of cognitive dysfunction.

Lastly, it is important to note that relative to the incidence of Lyme disease, the prevalence of chronic LE is quite low. Population studies have shown that only a small percentage of previously infected patients have neurological abnormalities, including objective evidence of memory impairment. Patients with residual deficits typically have had a longer duration of untreated disease prior to eventual treatment. Similarly, studies of children indicate that the chronic neurological deficits are rare when Lyme disease is adequately treated. Taken together, it is reasonable to conclude that the prognosis for returning to normal neurocognitive functioning after being infected with Lyme disease is good with adequate treatment.

See Also the Following Articles

BORNA DISEASE VIRUS • BRAIN LESIONS • CEREBRAL WHITE MATTER DISORDERS • PRION DISEASES

Suggested Reading

Adams, W. V., Rose, C. D., Eppes, S. C., and Klein, J. D. (1994). Cognitive effects of Lyme disease in children. *Pediatrics* **94,** 185–189.

Halperin, J. J., Pass, H. L., Anand, A. K., Luft, B. J., Volkman, D. J., and Dattwyler, R. J. (1988). Nervous system abnormalities in Lyme disease. *Ann. N. Y. Acad. Sci.* **539,** 24–34.

Kaplan, R. F., Jones-Woodward, L., Workman, K., Steere, A. C., Logigian, E., and Meadows, M.-E. (1999). Neuropsychological deficits in Lyme disease patients with and without other evidence of nervous system pathology. *Appl. Neuropsychol.* **8,** 3–11.

Krupp, L. B., Masur, D., Schwartz, J., Coyle, P. K., Langenbach, L. J., Fernquist, S. K., Jandorf, L., and Halperin, J. J. (1991). Cognitive functioning in late Lyme borreliosis. *Arch. Neurol.* **48,** 1125–1129.

Logigian, E. L., Kaplan, R. F., and Steere, A. C. (1990). Chronic neurologic manifestations of Lyme disease. *N. Engl. J. Med.* **323,** 1438–1444.

Shadick, N. A., Phillips, C. B., Sanga, O., Logigian, E. L., Kaplan, R. F., Wright, E. A., Fossel, A. H., Fossel, K., Berardi, V., Lew, R. A., and Liang, M. H. (1999). Musculoskeletal and neurologic outcomes in patients with previously treated Lyme disease. *Ann. Intern. Med.* **131,** 919–926.

Steere, A. C., Taylor, E., McHugh, G. L., and Logigian, E. L. (1993). The overdiagnosis of Lyme disease. *J. Am. Med. Assoc.* **269,** 1812–1816.

Magnetic Resonance Imaging (MRI)

JEFFRY R. ALGER
University of California, Los Angeles

GLOSSARY

longitudinal (T1) relaxation The process in which the nuclear spin magnetization recovers its orientation parallel to the applied magnetic field in characteristic time, T1, following a perturbation.

magnetic resonance imaging (MRI) A biomedical procedure that utilizes the magnetic resonance signal produced by the protons of tissue water to obtain vivid depictions of the internal macroscopic anatomy of soft tissues such as the brain.

nuclear spin magnetization The magnetic properties that result from the spinning behavior of a single atomic nucleus or an ensemble of atomic nuclei when placed in a magnetic field.

spin echo pulse sequence A frequently used procedure in MRI in which, for a variety of technical reasons, the appearance of the MRI signal is caused to be delayed for a defined time period after the excitation of the magnetization away from its equilibrium orientation.

transverse (T2) relaxation The process in which the nuclear spin magnetization loses its orientation perpendicular to the applied magnetic field in characteristic time, T2, following a perturbation.

time-to-echo The time delay between MRI signal excitation and the appearance of maximal signal when a spin echo pulse sequence is employed.

time-to-repeat The time between successive MRI signal excitations.

Magnetic resonance imaging (MRI) is a procedure that utilizes the magnetic resonance signal produced by the protons of tissue water to obtain vivid depictions of the internal macroscopic anatomy of soft tissues such as the brain. It has become the method of choice for nondestructive visualization of brain anatomy. The nuclear magnetic resonance (NMR) phenomenon on which MRI is based was discovered in the 1940s. NMR has become an indispensable tool in the fields of chemistry, biochemistry, and structural biology. In the 1970s, methods of forming images from the proton NMR signal produced by the water in living tissues were developed and became known as MRI. MRI is now a routinely used clinical tool and has growing utility for investigations involving brain structure. The purpose of this article is to familiarize the neuroscientifically inclined reader with key physical principles and fundamental technological aspects that underlie MRI.

I. MRI SIGNAL GENERATION

A. The Nuclear Spin

MRI is based on the fact that collections of atomic nuclei, when placed in a strong unchanging magnetic field, interact with an externally applied oscillating magnetic field when the frequency of the oscillating magnetic field meets certain specific criteria. Many texts inaccurately summarize this by saying that atomic nuclei absorb or emit electromagnetic radiation (i.e., radio waves). It is not accurate to infer an interaction between atomic nuclei and electromagnetic radiation because electromagnetic radiation requires the concurrent presence of electric and magnetic fields with specific spatial and temporal relationships. In

MRI one avoids use of oscillating electric fields because these do not interact with the atomic nuclei. Only the oscillating magnetic field interacts with the atomic nuclei. In MRI, it is the hydrogen nucleus (^{1}H), which is often referred to as "the proton," that is responsible for generating the signal from which the images are formed. Other atomic nuclei (e.g., ^{19}F, ^{31}P, and ^{23}Na) have been used in exploratory biomedical MRI studies. However, the use of nuclei other than protons is not widespread for brain MRI. The vast majority of the protons present in biological tissue are constituents of water molecules; therefore, water protons have the greatest relevance in MRI. However, the protons in the fatty acyl components of lipid molecules of fatty tissue can also be imaged. One of the reasons that MRI is a successful neuroimaging tool is that water, which is found at high concentration in soft tissues such as the brain, generates the signal from which the images are formed. MRI is distinct from conventional radiographic and computed tomographic approaches to brain imaging. In MRI, an intrinsic tissue signal is imaged, whereas X-ray techniques are based on attenuation of an X-ray beam by the tissue.

Figure 1 illustrates the fundamental properties of the proton that lead to MRI. Any particular volume element (voxel) of brain tissue contains a very large number of water molecules, each of which has two protons. A typical MRI voxel ($1 \times 1 \times 3\,mm^3$) of brain tissue contains about 10^{20} protons. Each proton behaves as a bar magnet because it acts as if it spins about an axis. The property of spin and the associated magnetism lead to the use of the terms "spin" or "nuclear spin" as synonyms for proton or hydrogen nucleus. In the foregoing, the verb "to behave" was used to reflect the fact that the magnetic and spin properties of the atomic nucleus are governed by the laws of quantum mechanics, the science that deals with the behavior of matter and energy at very small dimensions. A detailed discussion of quantum mechanical behavior of the nuclear spin is beyond the scope of this article. Accordingly, "real-world" descriptions will be used as an approximation for the behavior of the atomic nuclei in the quantum world.

B. Interaction of the Nuclear Spin with Magnetic Fields

MRI requires that the signal-generating protons be placed in a strong static magnetic field. It is conventional to denote the applied magnetic field by a vector quantity, B_0. The typical magnetic field strength used in brain imaging is 1.5 T (approximately 30,000 times stronger than the earth's magnetic field). The spatial

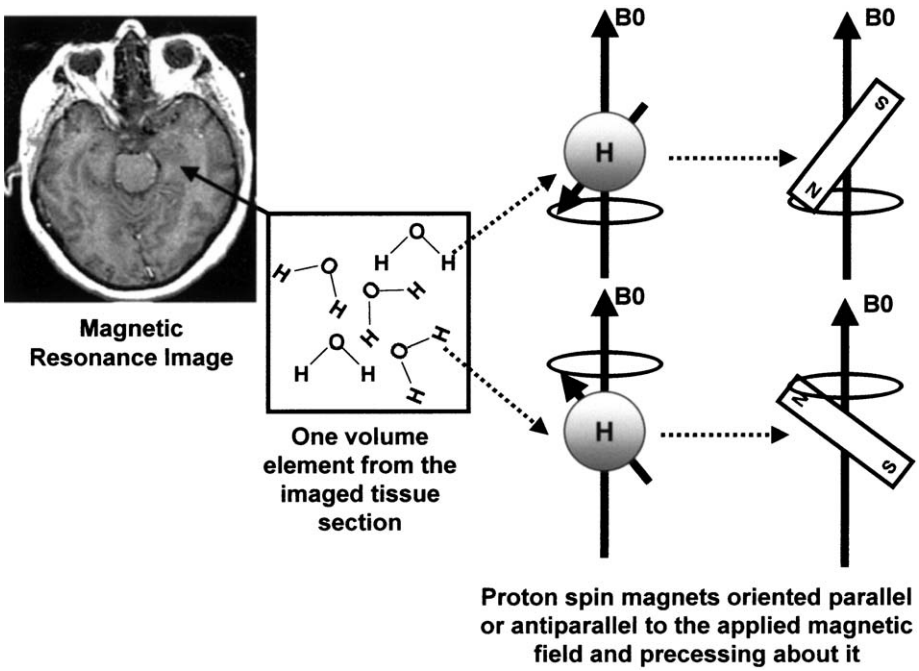


Figure 1 Key concepts related to MRI signal creation.

homogeneity of the magnetic field also plays an important role in defining feasibility of imaging. Typically, the magnetic field must vary less than 100 parts per million (ppm) over the entire brain volume and less than about 1 ppm over any particular imaging voxel. An important attribute of MRI is that the magnetic field readily penetrates bony structures, permitting MRI to "see" the brain through the bony cranium.

C. The Larmor Relationship

Figure 1 illustrates how the proton nuclear spins (equivalent to spinning bar magnets) behave in an imposed magnetic field. They tend to align parallel to (with) or antiparallel to (against) the magnetic field. Such behavior is readily appreciated when two magnets are manipulated in the macroscopic world. When a smaller magnet is placed inside a larger one, the smaller magnet's north pole tends to point toward the north pole of the larger magnet. "North to south" orientation has some stability, but a perpendicular orientation of the two magnets is highly unstable. In the quantum world, magnetic alignments are expressed in terms of their probabilities. Alignment in the parallel configuration is only slightly more probable compared to that of the antiparallel configuration. Furthermore, in the quantum world, the alignment is not perfect. The spin causes the nuclear magnet to "precess" about the applied magnetic field at a constant angle in each of the two allowed configurations. The precession frequency, v, (the number of precession cycles per second), is a linear function of the applied magnetic field

$$v = \gamma B_0$$

Precession characteristics such as these are found in everyday mechanical systems (e.g., gyroscopes). The relationship between the precession frequency and the applied magnetic field on the orbital motion of an electron was described in 1897 by Sir Joseph Larmor. Accordingly, even in the context of MRI, the expression that relates precessional frequency to magnetic field strength is known as the Larmor relationship. The Larmor relationship indicates that the precession frequency is directly proportional to the magnetic field strength, with γ (the gyromagnetic ratio) being a constant of proportionality that is a unique property of the type of nucleus. For the proton, γ is approximately 42 million cycles per second per Tesla. Therefore, the characteristic frequency at which the nuclear spins precess about the commonly used applied magnetic field strength (1.5 T) is 63.8 MHz.

D. Behavior of Large Ensembles of Nuclear Spins

A single nuclear spin would produce a signal that is far too weak to detect. Therefore, it is necessary to consider how large numbers of nuclear spins behave in a collective sense to understand how a measurable MRI signal is produced. It is the 10^{20} protons in the typical MRI voxel that produce the signal used for image formation. Some of the spins are oriented antiparallel to the applied field, whereas others are oriented parallel to the applied field (Fig. 2). Nature tends to favor the parallel configuration because this represents an energetically more stable (lower energy) state. However, the two possible configurations do not differ to a great extent in their energy stability, and there are almost equal numbers of spins in the two configurations. The magnetic field from each of the spins in the antiparallel configuration is canceled by a spin in the parallel configuration, but there are spins in the parallel configuration that not canceled by antiparallel spins. These "uncanceled" parallel spins act cohesively as an ensemble. They collectively behave as a bar magnet oriented perfectly parallel with B_0. It is common practice to refer to the magnetic properties of the ensemble as the "bulk nuclear spin magnetization" or "magnetization." The angle between the magnetization and B_0 is lost because each nuclear spin precesses about the applied field independently of the others. Nature imposes no constraints on where along

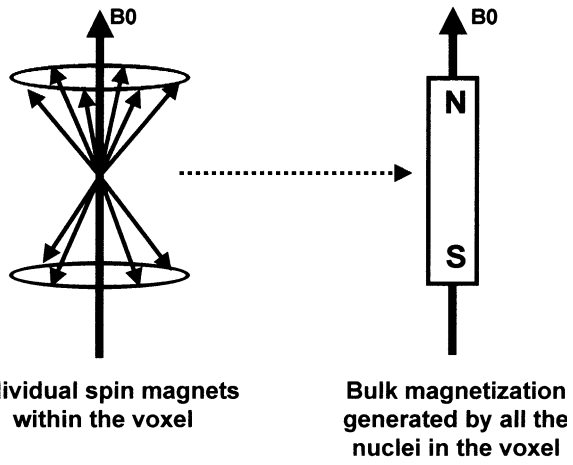


Figure 2 Ensembles of nuclear spins create a magnetization parallel to the applied magnetic field.

each precession circle each spin happens to be, and the transverse component nuclear spin magnetization perpendicular to B_0 is zero. It is important to also emphasize that the bulk magnetization produced by an ensemble of many nuclear spins behaves differently than does the spin magnetization from a single spin. The bulk magnetization from an ensemble of nuclear spins can attain any arbitrary orientation relative to B_0. Its behavior is not "quantized" as is the case for a single spin's magnetization.

E. The Magnetic Resonance Signal

In MRI, images are formed from the electrical "signals" generated by each of the volume elements in the tissue. The bulk magnetization resultant from the ensemble of spins in each of the volume elements generates the MRI signals. In order to detect the MRI signal, it is necessary to first disturb the bulk magnetization away from the equilibrium configuration shown in Figure 2. This is done using a radiofrequency (RF) transmitter pulse (Fig. 3). The working part of the RF pulse is a second magnetic field, which is usually referred to as B_1. The B_1 magnetic field differs from B_0 in that it oscillates. In order to be effective, B_1 must oscillate at a frequency that is very near to the Larmor frequency (e.g., 63.8 MHz for 1.5 T). When this is the case, there is a coupling between B_1 and the nuclear spin precessions that causes the bulk magnetization to be rotated away from its equilibrium position toward the transverse plane perpendicular to B_0. Energy from the B_1 field is used to excite the magnetization to a nonequilibrium configuration. Figure 3 shows the result of a 90° pulse in which the magnetization is twisted through an angle of 90° into the transverse plane. Coupling between the bulk magnetization and a B_1 field having the appropriate oscillatory frequency is an example of the phenomenon known in physics as resonance. The oscillatory B_1 field is created by the flow of alternating electric current through a coil of wire (i.e., the RF coil) that is placed near the tissue being imaged. The alternating electric current frequency must be very near the Larmor frequency (typically 63.8 MHz) to create the appropriate oscillation frequency for B_1. Use of RF is a central part of radio and television broadcasting; therefore, one often hears that MRI results from an interaction between "radio waves" and nuclear spins. However, as has already been stated, only oscillating magnetic fields are used in MRI, and it is not correct to state that radio waves are used in MRI.

Once the magnetization has been disturbed as shown in Fig. 3, it precesses about the transverse plane

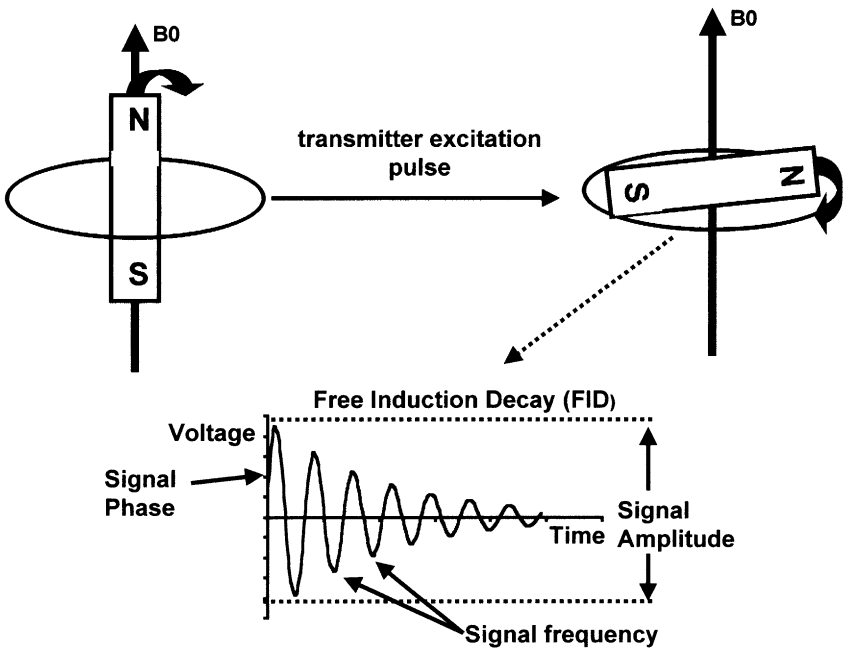


Figure 3 Perturbation of the magnetization with a radiofrequency pulse and the resultant free induction decay signal.

at the Larmor frequency. This precessional motion can be detected with a coil of wire (i.e., the RF coil) that is placed near the tissue of interest. The motion of the magnetization generates a voltage oscillating at the Larmor frequency in the coil. Therefore, one can describe the phenomenon by saying that the nuclear spins transmit a radio frequency signal on a characteristic radio "channel" that is detected by the MRI scanner. Note, however, that radio waves per se are not involved in this signal-generation process. The signal subsequently decays as illustrated in the graph shown in Fig. 3. The maximal signal amplitude (voltage) is proportional to the magnetization, which in turn is principally dependent on the number of nuclear spins in the ensemble, although other factors (described later) also play a role. The MRI signal that results immediately from a 90° pulse is known as a free induction decay (FID). It is often advantageous to cause the signal to appear sometime after the transmitter RF pulse is finished. One way of doing this is to form a spin echo. Spin echo MRI signals oscillate at the Larmor frequency just as FID signals do. However, they build up and then disappear at some defined period of time after the transmitter pulse is completed (Fig. 4). The time that elapses between the initial transmitter pulse and the formation of the echo is known as the time-to-echo (TE). The amplitude of the spin echo signal decays as the TE is progressively lengthened, as illustrated in the graph shown in Fig. 4. This will be discussed in more detail later.

Formation of a spin echo requires the use of two RF pulses given in a carefully prescribed manner. A series of pulses is referred to as a pulse sequence. In a broader sense, pulse sequences, which can be quite complicated, are frequently used in MRI to manipulate the magnetization in a defined manner so as to accentuate certain aspects of the MRI signal or for developing contrast. The spin echo pulse sequence is one of the simplest of a large and growing body of MRI pulse sequences.

II. IMAGING THE MAGNETIC RESONANCE SIGNAL

MRI visualization of structural anatomy requires measuring the MRI signal amplitude generated by each volume element within the tissue being imaged. Accomplishing this is more complicated than might be imagined. Because MRI does not use radio waves, it is not possible to focus "beams" of radio waves at suitable resolution to excite the MRI signal of only one volume element at a time and then sequentially sample all volume elements with this focused beam. Moreover, the sequential detection of the MRI signal from each volume element is highly inefficient and impractical.

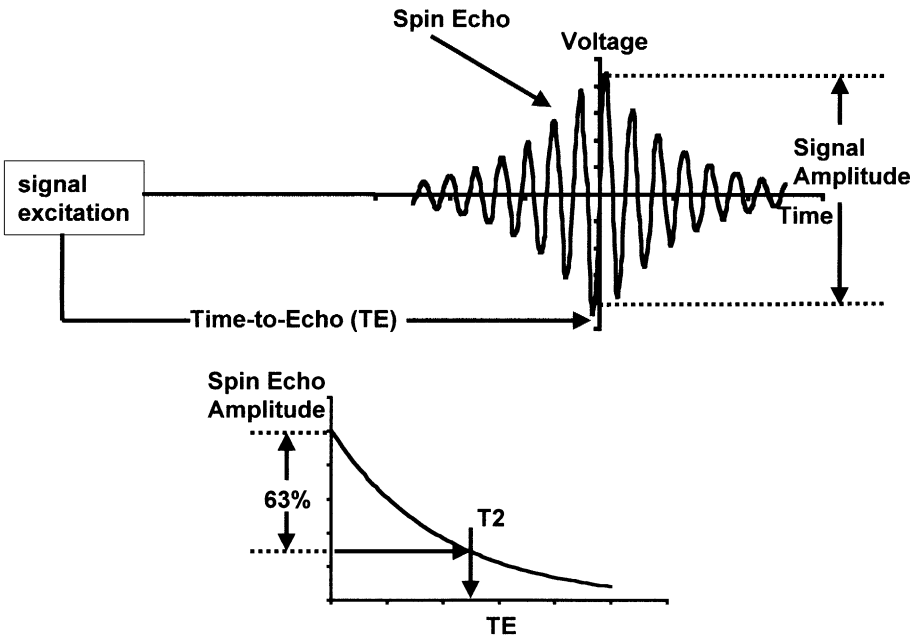


Figure 4 The spin echo signal and its dependence on TE.

Alternate ingenious methods based on the use of magnetic field gradients and precise frequency measurement are typically used for imaging entire sections or volumes of tissue.

A. Magnetic Field Gradients

A magnetic field gradient is a smooth (usually linear) variation in the static magnetic field (B_0) from one position to another position. Magnetic field gradients are purposefully applied in MRI as part of the imaging process. This is illustrated in Fig. 5. Application of a magnetic field gradient that varies smoothly in the inferior–superior direction causes B_0 in the neck to be relatively smaller compared to that in the brain. In the presence of this field gradient, the Larmor relationship ensures that there will be smooth linear dependence of the MRI signal frequency along the inferior–superior axis. Therefore, it is possible to know where along the superior–inferior axis a particular signal-generating volume element is located through precise frequency measurement. There is no strict requirement that a field gradient be oriented along any particular anatomic axis. Field gradients may be created that cause linear variation of B_0 along the left–right and the anterior–posterior axes or any arbitrary oblique axis that lies at any angle between the principal anatomic axes. Furthermore, it should be understood that magnetic field gradients can be turned on and off (i.e., switched) during the pulse sequence. This permits the application of magnetic field gradients on different axes during different parts of the pulse sequence.

B. Slice Selection

In MRI, tomographic images of tissue slices having defined orientation, thickness and location are obtained. The slice selection process is accomplished by using a frequency-selective RF pulse during the application of a magnetic field gradient. A frequency-selective RF pulse permits the highly selective excitation of MRI signals having a narrow band of Larmor frequencies. The presence of a magnetic field gradient ensures that this narrow band of Larmor frequencies will correspond with a narrow tissue section. Protons that would generate signal outside of the narrow band of Larmor frequencies (i.e., those located beyond the edges of the narrow tissue section) are unaffected and do not produce signal. Figure 5 illustrates that an axial tissue section may be selected for subsequent imaging by using a frequency-selective RF pulse in the presence of a magnetic field gradient oriented along the inferior

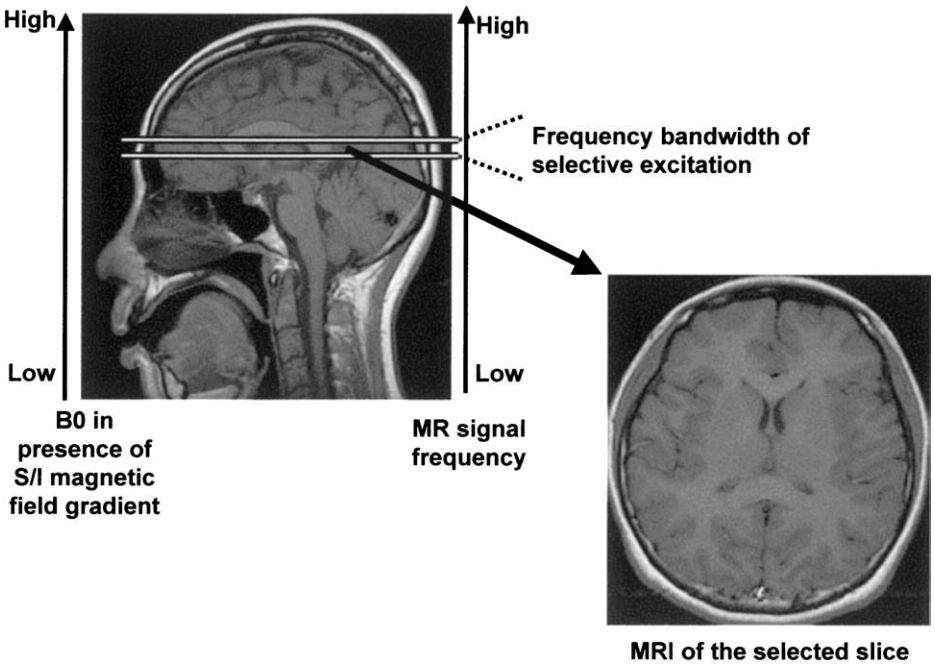


Figure 5 Slice selection in MRI.

to superior axis. Both the slice thickness and the slice location may be freely selected without physically moving the human subject or physically moving components of the scanning equipment relative to the subject. This is done by making suitable adjustments in the characteristics of magnetic field gradient or the frequency-selective RF pulse. Accordingly, the moving parts of MRI scanners tend to be limited to that needed for moving the subject into the appropriate part of the magnetic field at the beginning of the examination. Subsequently, neither the subject nor the scanning equipment moves during the image acquisition process. Furthermore, because a field gradient can be created along any arbitrary anatomic axis, it is possible to select slices in any of the three principal anatomic sectional planes (sagittal, axial, and coronal) or in any arbitrary oblique imaging plane without physically repositioning the subject or physically reorienting the imaging equipment (Fig. 6).

The obvious need to visualize anatomy in three dimensions is usually met through the use of mutlislice imaging. Figure 7 displays a series of 15 axial images that were obtained from tissue sections 3 mm thick. The conventional practice is to produce collages showing individual two-dimensional sectional images as a means of visualizing three-dimensional (3D) anatomy. If the sections can be made sufficiently thin, it is possible to create a volume rendering that displays the cortical surface anatomy. Conventional multislice MRI for human subjects is limited to a slice thickness of about 3 mm. This is relatively thick for volume rendering purposes; therefore, when volume rendering is planned, special 3D acquisition techniques are employed in which more than 100 slices having slice thickness of about 1 mm are used. Figure 8 provides an example of volume rendered images that display brain surface anatomy from a variety of perspectives. In addition to viewing surface anatomy in the manner shown, it is also possible to section the digital representation of the brain volume and produce two-dimensional images in any chosen sectional plane. This is done retrospectively using purposely designed software that sections and displays images under user control. The volumetric nature of the image data also readily lends itself to studies in which it is desired to measure the volumes of specific neuroanatomical structures that can be visualized in the images. As a result, 3D acquisition techniques are often used in research studies in which quantitative evaluation of neuroanatomic volumes are sought. The volumetric

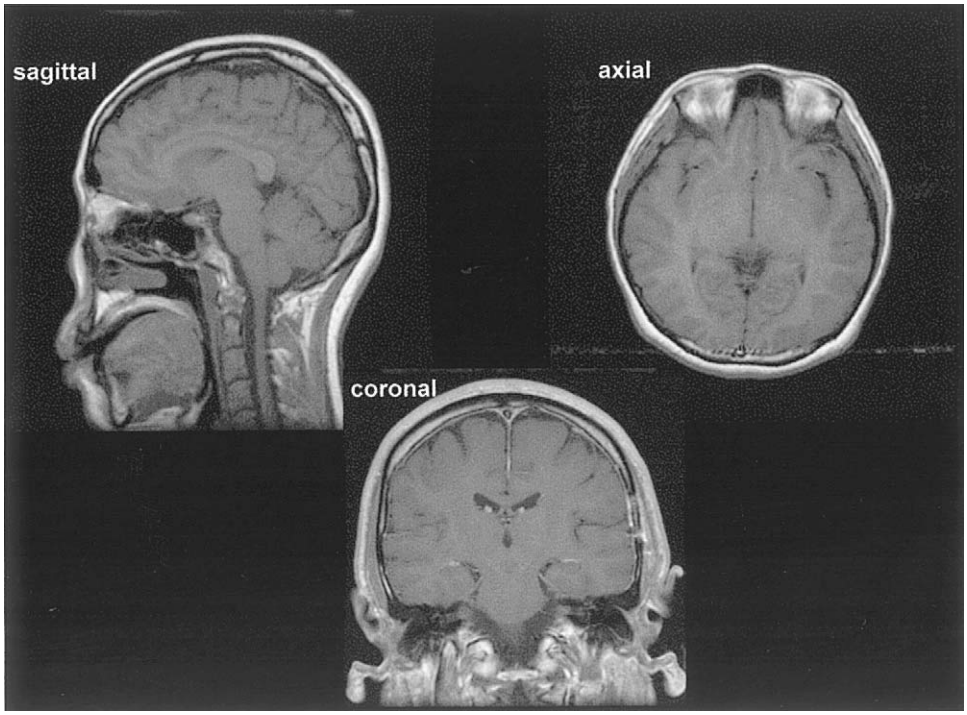


Figure 6 MRI slice selection can be in any chosen anatomic plane.

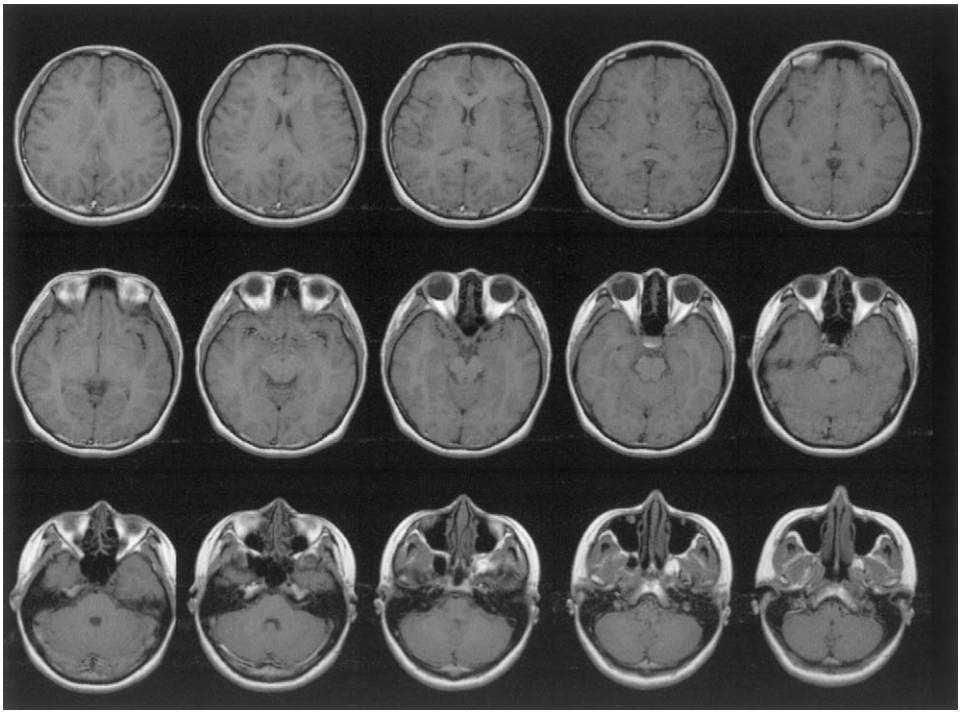


Figure 7 Multislice MRI.

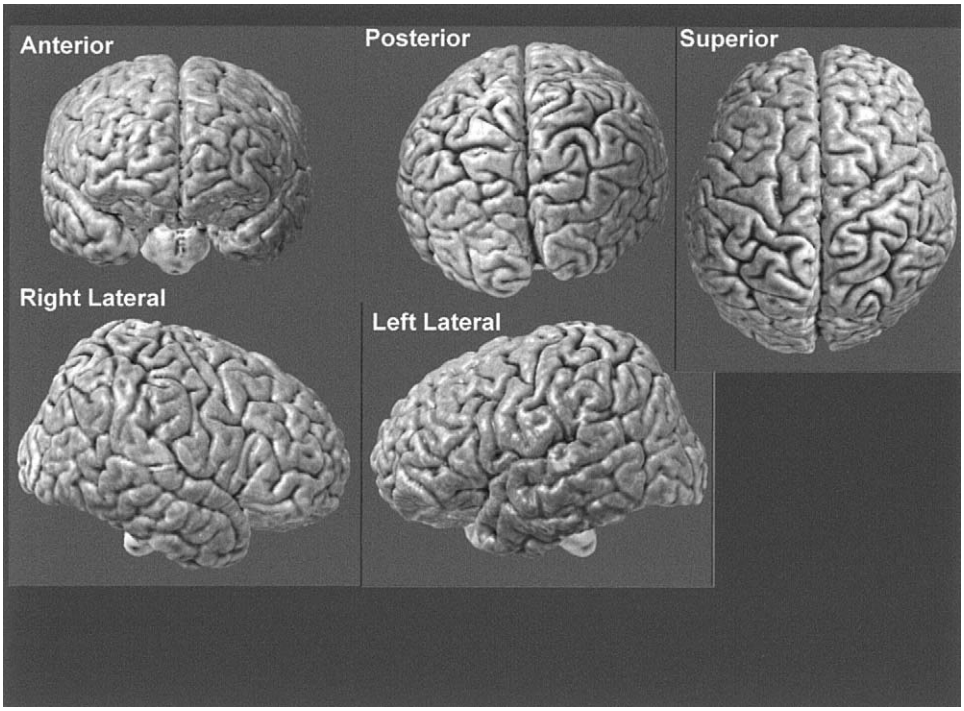


Figure 8 Volume rendering of three-dimensional MRI.

nature of the image data is also appropriate for the planning of neurosurgical procedures such as stereotactic biopsy. Accordingly, 3D volume MRI acquisition is often performed as part of the preoperative evaluation of patients for whom neurosurgical procedures are contemplated.

C. In-Plane Imaging: Frequency Encoding and Phase Encoding

Once the nuclear spins from a particular slice are excited, it is necessary to have methods of measuring how the MRI signal varies within the imaging plane. For two-dimensional imaging, two orthogonal coordinate directions are interrogated by application of magnetic field gradients that vary along these coordinate directions. Typically, frequency encoding and phase encoding gradients are used. To avoid confusing locations along the two orthogonal axes, the frequency encoding and phase encoding gradients are applied at different times during the pulse sequence. Figure 9 illustrates the general concepts. A magnetic field gradient is imposed such that the signal frequency is higher on the left compared to the right. (Note that it is customary in MRI to display the subject's left on the viewer's right.) This causes the signal frequency to depend on position along the left–right axis. A plot of signal amplitude versus frequency in the presence of this left–right frequency encoding gradient would give a one-dimensional image in which signal amplitude is projected down the columns along the anterior–posterior axis onto the left–right axis. The signal amplitude along this one-dimensional image defines the total amount of MRI signal generated by anterior–posterior columns. To obtain a two-dimensional image, a phase encoding gradient must also be used. In Fig. 9, the phase encoding gradient is applied along the anterior–posterior direction, causing the signal phase to depend on the position along the anterior–posterior direction. Phase encoding must be used because the signal frequency is already used for left–right encoding. Phase encoding is more complicated compared to frequency encoding. It is necessary to perform a number of phase-encoded signal measurements using different levels of the phase encoding to fully unravel the relationship between the position and phase. Typically, if it is desired to obtain a resolution of 256 image lines in the phase encode direction, then it is necessary to make on the order of 256 independent phase-encoded measurements.

The previous paragraph described the most commonly employed procedure for in-plane imaging. Several additional variant techniques have been

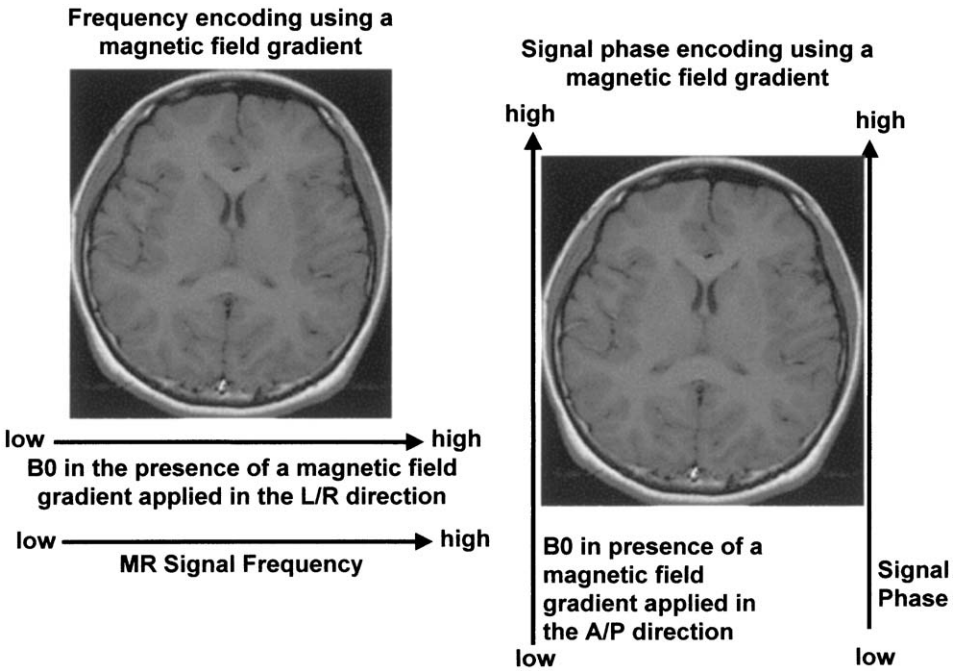


Figure 9 Frequency and phase encoding using applied magnetic field gradients with frequency measurement for in-plane imaging.

designed for specific purposes. The 3D acquisition techniques described previously use phase encoding in two dimensions and frequency encoding in one dimension without slice selection. This permits the slices to be made thinner than is typically possible when using conventional slice selection techniques. Another variant is known as echo planar imaging (EPI). In EPI, the frequency encoding and phase encoding are combined in a manner that speeds the imaging process considerably compared to conventional MRI. However, efficiency of EPI is partially offset by an increased susceptibility to artifacts arising from B_0 imperfections.

D. Resolution and Imaging Times

The typically attainable spatial resolution for brain MRI is approximately $1 \times 1 \times 3\,\text{mm}^3$. Somewhat higher resolution can be attained in special circumstances (e.g., by using magnetic field strengths greater than 1.5 T or limited field-of-view RF coils). In general, the resolution is principally dependent on the signal-to-noise ratio (SNR) of the MRI signal produced by a single voxel. MRI "noise" results from the random movement of electric currents in the detection circuitry and in the subject. The noise level can generally be assumed to be a constant and only weakly dependent on the signal detection technique. Unfortunately, the MRI signal is relatively weak compared to that detected in many other forms of spectroscopy. It is often not appreciably larger than the noise. Reducing the size of the volume element to increase spatial resolution results in a proportional decrease in the SNR and this leads to an increased level of image "graininess," complicating visualization of subtle features. SNR may be increased through signal averaging (i.e., by spending more time at image acquisition), although this provides meager returns. In general, the SNR is directly proportional to the square root of the time spent at image acquisition. Doubling the image acquisition time leads to only a $\sqrt{2}$ (about 40%) improvement in SNR. If one wishes to improve the resolution by twofold without loss of image clarity (i.e., maintain the SNR), one must increase the imaging time by fourfold.

The imaging time is also limited by the phase encoding process. A twofold increase in the resolution along the phase encoding axis requires that one spend twofold more time at phase encoding. Certain rapid imaging techniques partially circumvent this limitation. EPI and fast spin echo imaging procedures are examples. However, additional imperfections in image quality must be accepted as a consequence of more rapid imaging.

III. RELAXATION AND TISSUE CONTRAST

The process wherein the magnetization returns to its equilibrium configuration parallel to B_0 is known as nuclear spin relaxation. Typically, relaxation is conceptually separated into two separate processes: (i) relaxation of the magnetization component that is parallel to B_0 and (ii) relaxation of the magnetization component that is transverse to B_0. The former is known as longitudinal or T1 relaxation, and the latter is known as transverse or T2 relaxation.

Relaxation derives its energy from the random rotational and translational motion of the water (or lipid) molecules within the tissue. These motions are constrained by the tissue ultrastructure. It has not been possible to develop an exact theory of nuclear magnetic spin relaxation relevant to tissue due to the complex nature of the tissue ultrastructure. However, theories developed for simpler homogeneous materials can be extended to describe tissue nuclear spin relaxation. Moreover, tissue nuclear spin relaxation characteristics can be readily measured, so the absence of an exact theory is not a profound limitation. The practical utility is that different tissue types often display unique relaxation characteristics because of unique ultrastructural characteristics. Brain gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) each exhibit unique relaxation properties, and these may be used for developing image contrast between these tissue types.

A. T1 and Time-to-Repeat

T1 relaxation may be visualized as follows: At the conclusion of an excitation RF pulse, the magnetization is situated away from its equilibrium state. It takes a finite amount of time for the magnetization to recover to its equilibrium configuration. This recovery time is specified with a characteristic time constant that is known as T1. A series of image acquisitions produces signal intensity that is dependent on how rapidly the RF pulses are repeated relative to the rate at which the system is capable of relaxing. It is conventional in MRI to specify the rate of repetition with a parameter called time-to-repeat (TR). Figure 10 illustrates that repeated

acquisitions done at short TR tend to produce a relatively weak signal (compared to longer TR) because the magnetization does not have sufficient time to reattain its equilibrium configuration. A series of measurements in which signal intensity is measured as TR is increased typically leads to results illustrated in Fig. 10. Increasing signal is observed as TR is lengthened and this generally follows an exponential mathematical function. Such a mathematical function can be completely described by a single "time constant," defined as the time at which some fraction (usually 50 or 63%) of the process is complete. Therefore, the relaxation time constant, T1, is defined as the TR at which 63% of the total available signal is measured. In other words, the T1 value is determined by locating the point along the curve where 63% of the total available signal is obtained and then projecting to the TR axis.

As indicated previously, the actual value of T1 depends on the extent to which the tissue ultrastructure constrains the molecular movement. Figure 11 illustrates general features of T1 relaxation in unique brain tissues. Tissues such as CSF in which the motion of water molecules is relatively unconstrained (and therefore rapid) tend to produce relatively long T1 values. T1 values in GM or WM are smaller than in CSF because the tissue ultrastructure constrains the water movement to a greater extent. Motional constants in WM are somewhat greater than in GM due to the presence of rigid myelin structures; WM T1 values are therefore shorter compared to GM T1 values. It is indeed fortuitous that the tissue T1 values are what they are. The TR must be on the order of T1 to realize an appreciable fraction of the total available signal. Brain T1 values are such that brain MRI may be performed with TR between 500 and 2000 msec. Therefore, a typical image acquisition (128 phase encodes) requires approximately 2 min. Were the brain T1 values to be significantly greater (as is the case for distilled water) the same acquisition would take 10 times longer.

B. T1-Weighted Imaging

Figure 11 illustrates the appearance of a T1-weighted (T1w) image. By definition, a T1w image is acquired using repetitive measures (for phase encoding and signal averaging) with full excitation (90° RF pulses) and TR that is between about 300 and 1000 msec. Image acquisition at this TR maximizes signal contrast between tissue types (Fig. 11). CSF shows the lowest (nearest to black on the gray scale) signal intensity. GM shows a somewhat stronger signal intensity (closer to white on the gray scale), but this is appreciably weaker than that produced by WM. The fatty tissue (not plotted on the graph) located around the orbits and in the superficial soft tissues shows a

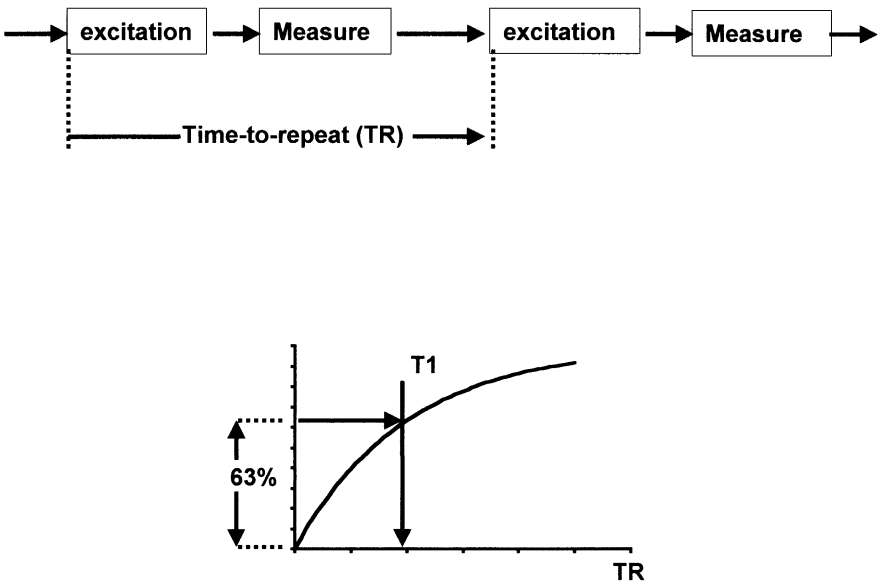


Figure 10 Longitudinal (T1) recovery and its relationship to the time-to-repeat (TR).

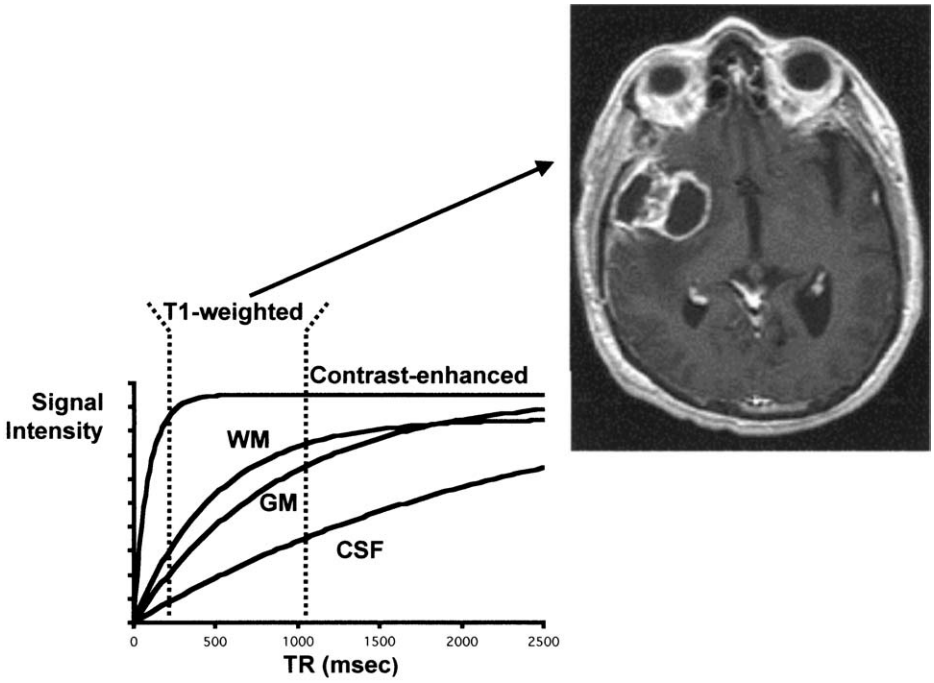


Figure 11 Contrast development through T1-weighted imaging.

very rapid T1 compared to brain tissues and this appears as intense signal in fatty tissues. The signal intensity differences in T1w imaging permit the reader to readily distinguish different tissue types and, thereby, to visualize the details of anatomic structure with great clarity.

Relaxation may be further enhanced by the use of relaxation agents, which are also called paramagnetic contrast agents. These materials have inherent magnetic properties arising from unpaired electron spins. They weakly associate with water molecules and through transient contact relax the water proton nuclear spins, making T1 very short. Intravenously infused paramagnetic contrast agents are commonly employed for clinical neuroimaging. The blood–brain barrier limits the access that such materials may have to normal brain tissue, and T1 of the normal brain tissue inside the intact blood–brain barrier is not altered. On the other hand, the agents tend to readily penetrate into tissue in which the blood–brain barrier is not intact (e.g., tumors). This leads to a relatively short T1 in tissues having a damaged blood–brain barrier and to very intense signal on T1w images. The image shown in Fig. 11 was obtained after intravenous administration of a commercially available contrast agent. It illustrates an intense double ring enhancement pattern produced by a tumor located at the right (viewer's left) frontotemporal junction. The bright "ring" is generally thought to represent highly vascular living tumor tissue that does not have an intact blood–brain barrier.

Figure 12 illustrates that T1w imaging may also be used to obtain angiographic images showing the major intracerebral vessels. This is known as magnetic resonance angiography (MRA). Blood flow tends to accentuate T1 relaxation because flow is constantly moving new water into the tissue section that is being imaged. Therefore, intravascular water tends to show full signal intensity after each excitation because it was not in the slice during previous excitations. On the other hand, the brain water needs time to relax between excitations to produce strong signal. One MRA approach is to employ very heavily T1w contrast (i.e., very short TR) so that brain signal is very weak and the intravasular water produces a much stronger signal. This is illustrated in the image shown on the left in Fig. 12. One can appreciate spots and strings of high signal intensity that depict the major vessels in this heavily T1w (flow-sensitized) image. Although single-slice views of vascular anatomy such as this are useful, display of MRA results in a manner consistent with fluoroscopic angiography is typically employed. This is illustrated in the image shown on the right in Fig. 12. To construct an angiographic projection view,

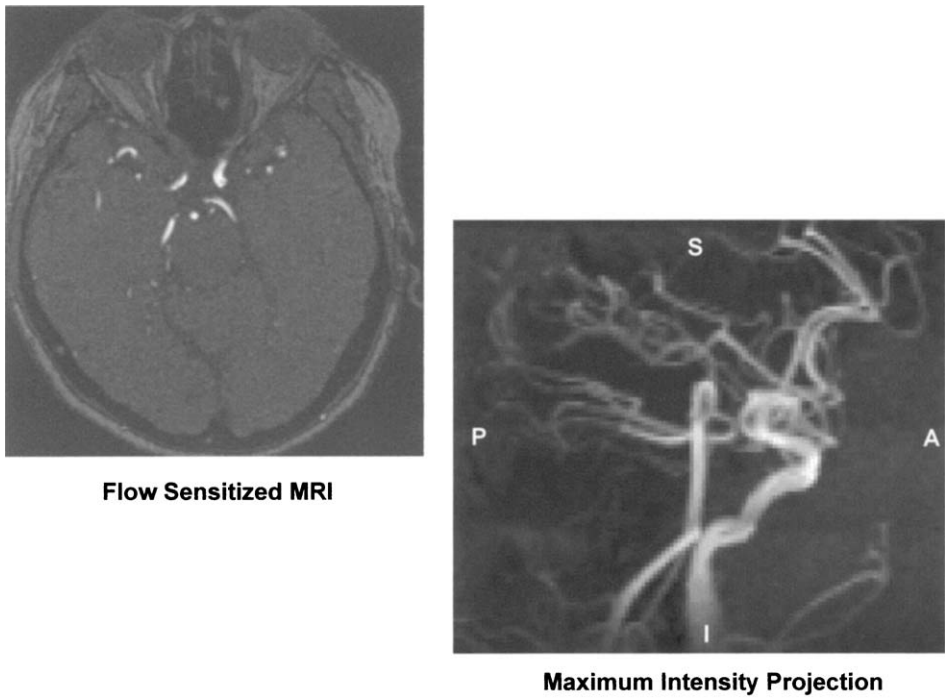


Figure 12 Magnetic resonance angiography.

flow-sensitized MRI is collected from many thin slices in a 3D format. This produces a 3D volume image in which the vessels are represented by relatively intense signal. Subsequently, the image intensities are projected onto a single viewing plane that may have any arbitrary orientation. In the example shown in Fig. 12, the angiographic information is being projected from a lateral perspective. From this perspective one can readily visualize the major feeding arteries of the corotid and basilar systems as well as many of the larger intracerebral vessels.

C. T2 and TE

Transverse (T2) relaxation is the process in which the finite components of the magnetization present in the transverse plane decay to zero (the equilibrium state). Often, T2 relaxation is discussed in terms of two separate processes. One of these is driven by molecular movements and is known as "pure" T2 relaxation. The other is the result of the presence of a magnetic field gradient within the voxel of interest and is known as T2* relaxation. To measure "pure" T2 relaxation independently of T2* relaxation, spin echo acquisitions (Fig. 4) are commonly used. The spin echo amplitude decays as the TE is lengthened, as illustrated in the graph in Fig. 4. This is generally found to be described by a decaying exponential function that has a characteristic time constant. To determine the value of the characteristic time constant for transverse relaxation (T2), one finds the point at which 63% of the signal has decayed and then projects that point onto the TE axis.

As is the case for T1, the T2 value depends on motional constraints imposed by tissue ultrastructure. However, the functional dependence of T2 relaxation differs from that for T1 relaxation. Tissues in which the water molecule motion is unhindered (e.g., CSF) tend to produce a very long T2. Increasing constraints on water movement lead to progressively shorter values for T2. WM shows a shorter T2 compared to that of GM, and both of these show a shorter T2 compared to that of CSF. The T2 relaxation characteristics of the three tissues are illustrated in the graph shown in Fig. 13. Spin echo signals produced by each of the tissues decay as TE is lengthened, with the differences becoming more pronounced as progressively longer TE values are used.

D. Proton Density-Weighted and T2-Weighted Imaging

Figure 13 illustrates that spin echo imaging can be employed to obtain two different types of tissue

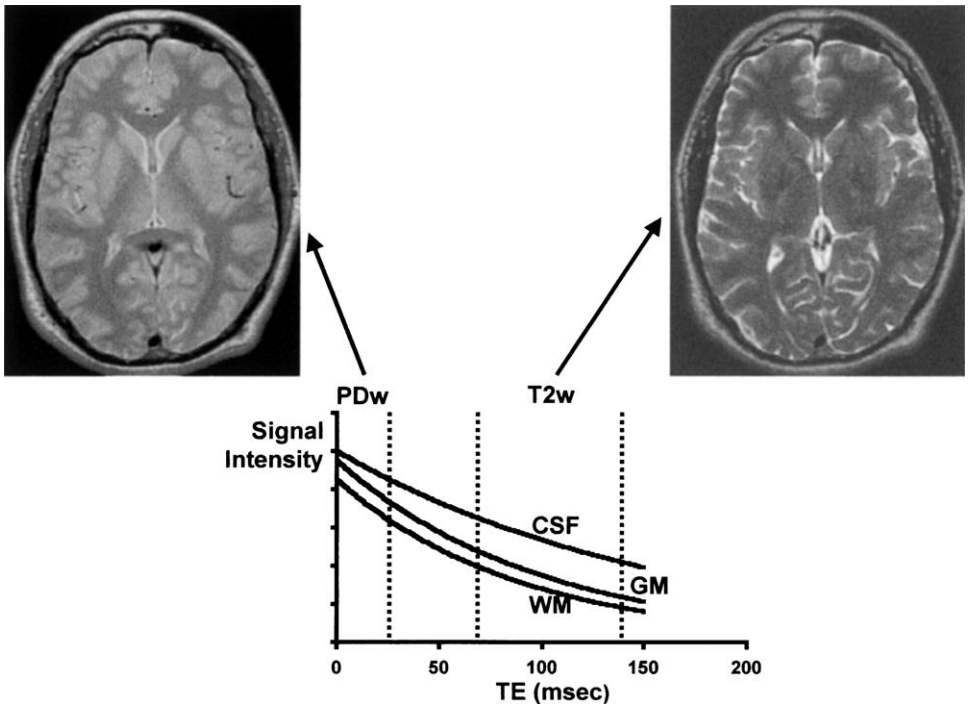


Figure 13 Contrast development through T2-weighted imaging.

contrast known as proton density contrast and T2 contrast. To obtain a proton density-weighted (PDw) image one uses a (relatively short) TE value of about 20 msec. Shorter TE values are desirable, but technical and engineering constraints usually prohibit their realization. In a PDw image the density of water protons is the predominant factor that defines what signal intensity is detected. Therefore, WM that has slightly fewer water protons per unit volume due to the presence of myelin tends to appear as lower (closer to black) signal compared to GM or CSF. CSF water density is slightly higher than in GM, causing the CSF to appear modestly more intense compared to GM. T2-weighted (T2w) images are typically obtained with TE values between 80 and 150 msec. This serves to maximize the contrast differences between the three tissue types as shown in the graph in Fig. 13. In the case of T2w images, the CSF shows the most intense signal, with GM and WM showing lower and almost equal signal intensities.

IV. MRI TECHNOLOGY

A typical MRI scanner is a complex system that utilizes advanced digital and analog electronics and expensive magnet technology. A simplified block diagram of a typical system is provided in Fig. 14. The system is centered around a computer system that interacts with a purposefully designed system controller. The system controller is responsible for coordinating each of the subsystems to produce pulse sequences and to acquire the MRI signal. The gradient system controller creates magnetic field gradient pulses by application of direct electric current to a set of gradient coils that are positioned around the subject within the magnet. The RF transmitter system is essentially a radio transmitter that is responsible for creating the RF pulses by appropriate application of alternating electric current to the RF coil. The RF receiver system is responsible for detecting the FID or spin echo signals induced in the RF coil by the movement of the magnetization.

The magnet is a relatively expensive system component. It often accounts for a large fraction of the system cost. The majority of current MRI systems use superconducting magnet technology because this provides adequate field strength, superior field homogeneity, and field stability with minimal operating costs. Despite the popularity of superconducting MRI magnets, permanent magnet technologies and conventional electromagnet technologies are also used in some commercial MRI scanners. The requirement for

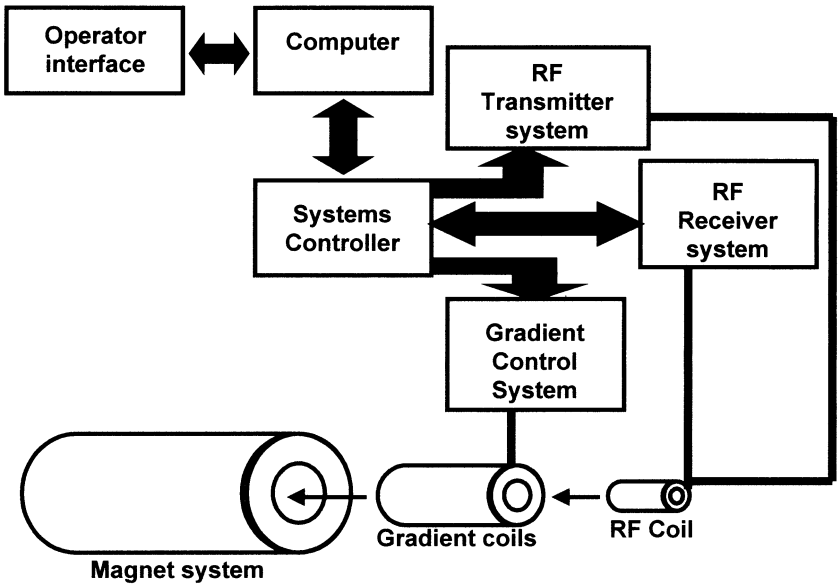


Figure 14 MRI scanner system design.

a strong magnetic field within the head necessitates that the subject be "inside" the magnet aperture (or bore). In addition, the subject's head must be placed within a RF coil, and this must be placed within a set of magnetic field gradient coils. Therefore, the RF coils and the gradient coils must also be placed within the magnet aperture (Fig. 15). In order to perform an imaging examination, the subject lies on the bed. The bed is raised to the level of the RF coil, and the RF coil is placed around the head. The subject is then moved into the magnet so the head is located approximately 1 m within the magnet bore.

V. SAFETY AND EXPOSURE

MRI is widely regarded as a safe, innocuous imaging procedure that can be repeated at virtually any desirable interval. It is generally deemed as being safe for use with normal subjects within the context of research studies. MRI is accomplished without the use of ionizing radiation (X-rays). In addition to exposing the subject to a static magnetic field of considerable strength, the imaging process also requires that the subject be exposed to magnetic fields that oscillate at a frequency used in radio and television broadcasting and also to switched magnetic field gradients. No studies have demonstrated that MRI exposure has any adverse affects on health in the short or long term. Of course, failure to detect an effect does not necessarily mean that effects do not exist. The current negative findings, however, do suggest that any possible health effects are subtle at best. Because of the possibility of unknown effects on human development, MRI is generally not used during pregnancy.

The MRI environment does present some hazards. Probably the greatest potential hazard is being struck by a magnetic object while in the MRI scanner. MRI magnets create a fringe field around them that is capable of levitating and attracting (at great speed) iron and steel objects. The majority of MRI magnets use superconducting technology and these magnets are

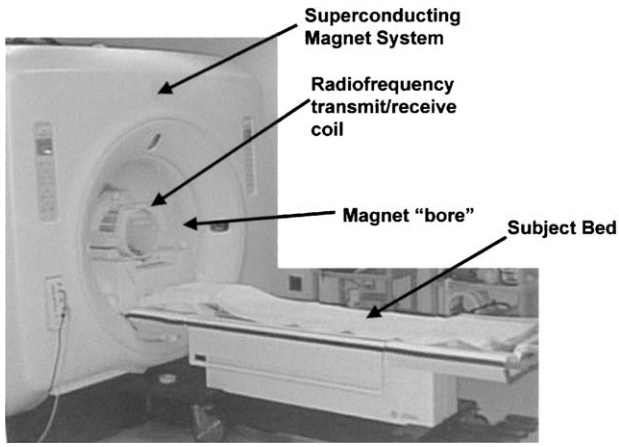


Figure 15 Photograph of a typical commercial MRI scanner.

rarely, if ever, turned off. Accordingly, the MRI scanner must be located in a controlled environment from which iron and steel objects are excluded. The magnetic field can also affect the operation of heart pacemakers and certain other electromagnetic prostheses. Individuals who rely on these devices should be excluded from the MRI environment. The magnetic field may also apply torque to certain types of aneurysm clips; therefore, subjects who have undergone vascular surgery must be carefully evaluated before being imaged with MRI. The switched magnetic field gradients used in MRI can induce electric current flow in the body and affect neuromuscular function. Accordingly, the limits of field gradient strength and gradient switching rates have been established by the appropriate regulatory authorities. Alternating RF electric currents can be induced in the body by the RF pulses used in MRI. Such currents dissipate energy in the form of heat and this can be a source of local tissue burns. This hazard becomes more probable when metallic objects (e.g., electrocardiographic electrodes) are located in or on the body in the vicinity of the RF coil. Therefore, the RF energy that can be used is regulated. With appropriate precautions, it is exceedingly rare for subjects to experience injury during MRI scanning.

Undergoing MRI is not necessarily a pleasant experience despite its innocuous health effects. MRI scanning is a relatively slow process during which the subject must remain motionless within the tight confines of the scanner hardware. Some subjects experience claustrophobia during MRI scanning. In addition, the magnetic field gradient pulses are accompanied by a high level of audio noise, which some subjects intensely dislike.

VI. SUMMARY

MRI provides exquisite images of the soft tissue structure of the brain at millimeter resolution from any viewing angle. It does this by detecting an intrinsic magnetic signal that is produced by the protons within tissue water when the body is placed in the appropriate magnetic environment. Nowhere are the uses of MRI more significant than in brain imaging. MRI is far superior to all other noninvasive brain imaging procedures because it images signal generated directly by the brain tissue and this signal is readily detected despite the presence of the bony cranium. The technology is sufficiently robust to routinely produce images at millimeter resolution with good tissue contrast. The signal detection process uses no form of ionizing radiation and is relatively innocuous. These properties permit subjects to be examined repeatedly. These attributes of MRI point to utility that extends well beyond the diagnosis and evaluation of disease. In recent years, neuroscientists have realized that MRI is a significant research tool that may be used for anatomic imaging in human and animal subjects.

See Also the Following Articles

ELECTROENCEPHALOGRAPHY (EEG) • FUNCTIONAL MAGNETIC RESONANCE IMAGING (fMRI) • IMAGING: BRAIN MAPPING METHODS

Suggested Reading

Brown, M. A., and Semelka, R. C. (1999). *MRI: Basic Principles and Applications*, 2nd ed. Wiley–Liss, New York.

Damasio, H. (1995). *Human Brain Anatomy in Computerized Images*. Oxford Univ. Press, New York.

Farrar, T. C., and Becker, E. D. (1971). *Pulse and Fourier Transform NMR; Introduction to Theory and Methods*. Academic Press, New York.

Gadian, D. G. (1995). *NMR and Its Applications to Living Systems*, 2nd ed. Oxford Univ. Press, Oxford.

Hashemi, R. H., and Bradley, W. G. Jr. (1997). *MRI: The Basics*. Williams & Wilkins, Baltimore.

Jackson, G. D., and Duncan, J. S. (1996). *MRI Neuroanatomy: A New Angle on the Brain*. Churchill-Livingstone, New York.

Lufkin, R. B. (Ed.) (1998). *The MRI Manual*, 2nd ed. Mosby, St. Louis.

Mitchell, D. G. (1999). *MRI Principles*. Saunders, Philadelphia.

Rajan, S. S. (1998). *MRI: A Conceptual Overview*. Springer, New York.

Toga, A. W., and Mazziotta, J. C. (Eds.) (1996). *Brain Mapping: The Methods*. Academic Press, San Diego.

Zimmerman, R. A., Gibby, W. A., and Carmody, R. F. (2000). *Neuroimaging: Clinical and Physical Principles*. Springer, New York.

Manic–Depressive Illness

EDWARD H. TAYLOR
University of Minnesota

GLOSSARY

cycle length The time from the beginning of one depressive or manic episode to the start of a new episode. As episodes increase, cycle length decreases. Individuals tend to establish rather constant and individualized cycling patterns. The term rapid cycling is used when a patient has four or more episodes of depression or mania, in any combination, during a 12-month period.

cognitive information processing (also known as social cognitive information processing) The ability to appropriately identify, classify, assess, and correctly act upon or disregard significant environmental and social cues. Processing of information requires the brain to accurately observe or attend to the environment, select important cues while disregarding less meaningful data, compare the selected cues with facts stored in memory, label partial and incomplete data, and develop a cognitive schema for action or determine that no action is required. All of the bipolar disorders reduce a person's information processing skills during periods of depression, mania, or hypomania.

depressive episode An extremely decreased mood that remains symptomatic across all or most of the person's daily activities. Depression removes the capacity for experiencing pleasure and often the ability to find meaning in endeavors or interpersonal relationships that were previously positive and enjoyable. Approximately two-thirds of people in a depressive episode think about suicide, and 10–15% kill themselves.

hypersomnia An excessive experience of sleepiness during day hours. The symptom occurs in some patients with depression and bipolar disorders but is more commonly part of a nonpsychiatric medical problem such as sleep apnea or narcolepsy.

manic episode An abnormally euphoric, elevated, or irritable mood that is associated with impulsive behaviors, poor judgment, concrete and grandiose thinking, reduced sleep, and impaired problem-solving skills. In the beginning individuals in a manic episode may express feelings of elation, energy, and increased insight. These positive perceptions, however, often evolve into feelings of anger, confusion, and fear. Obsessive thoughts of persecution, religiosity, sexuality, fame, economics, or other life events preoccupy individuals who are in a severe manic state.

neurobiological disorders Psychiatric disorders that relate to disruptions or pathological changes in the brain's structure and chemical neural messenger system. Brain abnormalities often found in neurobiological disorders include enlargement of the lateral and third ventricles, reduced frontal cortex volume, cerebellar atrophy, frontal cortex and medial temporal lobe metabolic changes, and changes in the size–shape structure of the temporal lobe and hippocampus.

spectrum disorder A term denoting psychiatric disorders that are thought to have a common pathophysiology and overlapping symptoms and often, but not always, can be treated with identical or similar psychotropic medications.

I. INTRODUCTION

Manic–depressive illness belongs to a spectrum of neurobiological diseases that today are most often called bipolar disorders. The terms are often used interchangeably in popular and scientific literature and refer to an interrelated group of illnesses rather than a

single disorder. The popular press and public speakers often use manic–depression and bipolar I disorder as synonyms for a single illness. This is understandable in that bipolar I disorder is one of the most severe forms of manic–depression illness. However, thinking of manic–depression as a cluster of similar mood disorders is more helpful. These illnesses, among other things, attack the brain, preventing individuals in varying degrees from willfully regulating their mood. Like all physiological and emotional problems, bipolar disorders can range in severity from mild to extremely disabling.

Definition of the disorders as simply shifts between high and low or depressed and manic moods misrepresents the affect, disabilities, and pain caused by these illnesses. People with bipolar disorders often report having moods and feelings that take control of their behaviors, motivations, thoughts, perceptions, and ultimately their lives. In a major depressive episode, feelings such as sadness, hopelessness, mental pain, and fatigue are amplified far beyond the normal human experience. Depression can create a pervasive and relentless sense of gloom, inadequacy, rumination, guilt, and worthlessness that the person is unable to dispel with logic, past experiences, or personal will. Mania swings an individual from an upward feeling of well-being, past exuberance, through a state of unexplained euphoria, and finally into a chaotic state of racing, incomprehensible, disconnected thoughts. Depressive and manic episodes are more than magnifications of everyday moods. They have the ability to decrease and control a person's cognitive information processing speed and accuracy, word retrieval, memory, motor speed and skills, concentration, abstract thinking and problem-solving skills, and perception of the social world. Whereas this article provides an overview of these disorders, readers should consult the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders*, 4th ed., for specific diagnostic and assessment criteria.

Most authorities believe that bipolar I and II and cyclothymic disorders represent differing types, symptoms, and severities of bipolar disorders. A growing number of researchers and clinicians also consider schizoaffective disorder as part of the bipolar spectrum. Bipolar I disorder is a severe form of manic–depressive illness and always includes the occurrence of at least one manic or mixed episode. Most individuals who have bipolar I disorder will at some point experience a major depressive episode, even though this is not a required criterion for the diagnosis. The illness can start with an episode of either mania or depression. Bipolar I disorder can cause acute, severe problems requiring hospitalization to control symptoms like psychotic episodes, extreme manic behavior, immobilizing depression, or suicidal thoughts. Bipolar II is diagnosed if the person has one or more major depressive and hypomanic episodes, but no history of manic or mixed episodes. Hypomania is a milder, less severe form of mania that seldom leads to hospitalization. Mental health professionals are sometimes unable to identify the hypomania episodes and misdiagnose the illness as severe depression or a personality disorder.

Cyclothymic disorder is the mildest form of manic–depressive illness. The disorder causes low-grade chronic mood abnormalities. For 2 or more years the person has numerous periods of hypomanic symptoms alternating with numerous bouts of mild to moderate depression. Individuals with this disorder find that, particularly during periods of depression, it becomes difficult but possible to maintain their normal activities. When hypomania occurs, the person experiences an increase in activity and at least a slight decrease in judgment and problem-solving skills. Irritability may occur in both the hypomanic and the depressed states. Many individuals with cyclothymia appear fully functional, maintain their employment, and never seek professional help. What is not seen is their internal turmoil, decreased problem-solving skills, and the amount of energy required to complete required daily work tasks.

Schizoaffective disorder represents the other end of the spectrum. It is an extremely severe form of illness that incorporates symptoms found in both manic–depressive illness and schizophrenia. Schizoaffective disorder can produce a state of persistent psychosis along with episodes of mania and depression. Normally the psychotic symptoms continue after the mood episodes dissipate. Some individuals appear to have a schizomanic condition, whereas others have more symptoms of schizophrenia and depression. The fact that mood-stabilizing medications help people with schizoaffective illness but generally are not helpful for patients with schizophrenia argues for classifying the illness as part of the bipolar spectrum.

II. HISTORICAL OVERVIEW

Throughout history scholars have described the symptoms and effects of manic–depressive illness. Ancient Greek, Persian, and biblical writers recorded

and attempted to explain the complexities of bipolar illness. In the second century AD Areteus wrote about patients who, in a state of euphoria, danced throughout the night, talked publicly, and acted overly self-confident then, for no apparent reason, shifted into a state of sorrow and despair. The fourth century BC Greek physicians lead by Hippocrates were perhaps the first to hypothesize that symptoms we now call bipolar disorder represented a neurological illness highlighted by major uncontrollable shifts in a person's mood. These early Greek scholars further taught that mental illness is caused by natural rather than spiritual forces, identified the brain as the major organ responsible for sanity and intellectual processes, attempted to classify major mental disorders, and developed crude medical treatments for mental disorders. Unfortunately, however, records from the ancient Egyptians, Greeks, Romans, Middle Ages, European Renaissance, and early American history indicate that this hypothesis gave way to assumptions of demoniacal possession, witchcraft, sinfulness, and other dehumanizing concepts. Nonetheless, traces of scientific and medical inquiries into bipolar disorders periodically appear throughout history.

The first person to identify the link between mania and melancholia or depression was Theophile Bonet. In 1686 Bonet described patients who cycled between high and low moods as having "manico-melancolicus." During the mid-1800s, French researchers Falret and Baillarger each independently observed that patients having manic and depressive episodes were not experiencing two different disorders, but rather two different presentations of the same illness. Falret described the disorder as "circular insanity" and listed the symptoms much as they appear in today's medical books and journals. He also (remarkably) hypothesized that the illness was hereditary and believed that through research a medication would be found for effectively treating the symptoms. The German psychiatrist Emil Kraepelin, building on Falret and Baillarger's work in the late 1800s and early 1900s, developed the definitive description and classification for manic–depressive illness that largely stands to this day. Kraepelin is credited with sensitizing past mood studies, clearly documenting that mania and depression are different symptoms of the same disorder, and with being the first researcher to assert that all mood disorders are neurologically related.

Kraepelin's basic concepts were challenged and widened by Eugen Bleuler in 1924. For Bleuler, mental disorders could not be classified into two major categories as Kraepelin claimed. Kraepelin believed that all mental illnesses fall into two basic, but separate groups. An illness was classified either as causing periodic recurring symptoms, such as manic–depression, or as a disorder characterized by ongoing neurological deterioration, such as schizophrenia. In his later work, Kraepelin did, however, clarify that it was impossible to neatly place everyone with mental illness into these two categories and that one cannot always discriminate among major disorders. Bleuler argued that manic–depressive illness and dementia praecox (schizophrenia) were not separate classifications, but rather a continuum. How a person was diagnosed and placed on the spectrum depended on the number of symptoms of schizophrenia that were found. More importantly, Bleuler broadened the manic–depression classification by identifying a number of subcategories and introducing the term affective illness. Between the early 1920s and mid-1980s, criteria independently developed by Kraepelin and Bleuler shaped most of the world's psychiatric diagnostic systems.

Between 1930 and 1940, mental health treatment providers largely abandoned the assumption that manic–depression and most other disorders, including schizophrenia and autism, developed from neurobiological abnormalities. Following World War II and until the early 1980s, mental health theory and treatment were mostly guided by psychoanalytic concepts proposed by Freud and his followers. Whereas psychoanalytic theory agreed that biological components played a role in affective disorders, practitioners insisted that early childhood parental or other environmental conflicts usually explained the onset and recurrence of manic–depressive episodes. As a result, it was thought that manic–depressive symptoms would resolve if individuals gained insight into their unconscious anger or other hidden emotional conflicts. Even though psychoanalysis and other forms of psychotherapy offered little help for most patients with severe manic–depressive problems, talk therapies nonetheless became the treatment of choice for decades. This preference continued for a number of years even after the introduction of lithium, the first drug found to successfully treat manic episodes.

Dr. John F. J. Cade, a doctor in the Mental Hygiene Department of Victoria, Australia, was dedicated to the belief that manic–depression was a biological, not an unconscious, psychological disorder. In the 1940s he was attempting to discover how urine toxicity levels from patients with various mental disorders differed. Cade wanted to inject guinea pigs with various concentrations of uric acid. However, uric acid is

insoluble in water and difficult to inject. To resolve this problem Cade mixed uric acid with lithium. To Cade's surprise, guinea pigs injected with the lithium solution had less toxicity in their urine. The scientist next injected the animals with lithium carbonate and observed that the animals remained conscious but less active and responsive to their environment. On the basis of these findings, Cade administered a lithium salt preparation to several highly agitated manic patients. Each of the patients had a remarkable reduction in symptoms. After Dr. Cade successfully treated 10 additional patients with the solution, European doctors started to quickly accept lithium as an important advancement in treating bipolar disorders. Because of safety concerns documented by cases of hypertension and deaths resulting from a consumer salt substitute containing lithium, the drug was not approved for use in the United States until 1970.

III. EPIDEMIOLOGY AND CAUSE

Bipolar disorders usually start in late adolescence and young adulthood. These illnesses, however, can appear any time between ages 5 and 50 and in rare cases beyond the age of 50. Research indicates that between 25 and 30% of the people who develop manic–depression as adults had one or more related symptoms before their 6th birthday. The more severe forms of bipolar illness (bipolar I disorder) are considered to be rare in prepubertal children. Only about 0.6% of adolescents are thought to have a bipolar I diagnosis, but estimates of teens with bipolar II disorder have reached as high as 10%.

Estimates of the number of people with a bipolar disorder vary. This occurs in part because of diagnostic difficulties and because of the fact that many people who have mild symptoms either do not seek or do not receive professional attention. At any given time, about 8% of America's population is at risk for developing a mood disorder. Most studies estimate that between 1 and 2.5% of the U.S. population has a bipolar disorder. A representative number of studies estimate that the prevalence of bipolar disorders is 3–6.5% of the U.S. population. Unlike unipolar depression, bipolar disorders are found equally in females and males. Between 5 and 20% of adult cases that are first diagnosed as unipolar depression over time will receive a reevaluated diagnosis of bipolar disorder.

Science has gained a large amount of information about manic–depressive illness but has been unable to identify specifically how the illness starts. There are most likely several causes for this syndrome of disorders. Studies of families, twins, and adoptions suggest that most cases of manic–depression are genetically inherited. There is general agreement among researchers that genetic components play a more significant role in transmitting bipolar I disorder than major depressive disorder, but this perspective continues to be debated. Family studies report that having a first-degree relative with bipolar I disorder increases the chances of developing manic–depressive illness by 8–18 times over families with no first-degree members having a bipolar disorder history. The likelihood of developing bipolar I disorder is 1.5–2.5 times greater if a first-degree family member has a major unipolar depressive disorder. Perhaps more illustrative of the genetic relationship to manic–depression is the fact that 50% of all people with a bipolar I disorder have at least one parent with a mood disorder. Additionally, a 25% probability exists of a child developing bipolar I disorder if one parent has this form of manic–depressive illness, and a 50–75% chance exists if both parents have bipolar I disorder. There is only a limited amount of information from adoption studies on bipolar disorders. The available data document that children adopted as infants from biological parents with a major mood disorder remain at an increased risk for developing bipolar disorders. The link between genetics and manic–depression has also been established through the study of twins. Monozygotic twins show a concordance rate for bipolar I disorder of 75%, whereas the concordance rate for bipolar I disorder in dizygotic twins drops to 20%. In addition, scientists have hypothesized that some cases of bipolar disorder may not stem from a genetic transmission. Researchers have found indications that bipolar disorders may be caused by *in utero* neuroviruses that attack the fetus' forming brain. There is currently growing interest in the role neuroviruses and immunologic abnormalities play in the formation and development of major mental disorders, including manic–depressive illness and schizophrenia.

IV. MAJOR SYMPTOMS

The type of bipolar disorder a person has is largely determined by identifying the severity, number, type, and duration of manic and depressive symptoms the

person has or is experiencing. Diagnosis of bipolar disorders is complicated by the fact that many symptoms of unipolar depression and manic–depression overlap. As an example, agitation and insomnia can occur in the depressed and in the manic state. Hypersomnia and psychomotor retardation, however, are observed more in bipolar than in unipolar depression.

Mania is one of the most dangerous of the abnormal mood states, but fortunately it is not present in all forms of bipolar disorder. Mania or manic behavior produces extreme and dramatic symptoms that can endanger the person's social and economic well-being and cause the individual to take life-threatening risks. The early stages of mania are experienced as pleasant and uplifting. The person feels energetic, creative, highly spirited, and capable. In the beginning of a manic episode, individuals are filled with a pleasant mood, ambitious thoughts, and self-confidence. They see great promise in their relationships, personal talents, skills, careers, and future. Goals are more clear, tasks seem less difficult, and life becomes magically filled with cosmic meaning and understanding. These exuberant and positive feelings, however, quickly pass and change into more pronounced psychiatric symptoms.

The person's cognitive information processing skills are disrupted by rapidly occurring thoughts that not only collide together but form incomplete and incongruent ideas. Additionally, the ability to screen and judge the appropriateness of one's thoughts, behaviors, productivity, and quality of work largely disappears. Furthermore, the previously elated mood filled with self-confidence, grandiosity, and positive symbolic meaning turns into unexplained, almost random anger and irritability. Whereas some individuals alternate between periods of elation and irritability, most slip into a state of mania dominated by dissatisfaction, frustration, intolerance, and an unsettling irritated mood. Some individuals also experience a constant internal rage that, with almost no provocation, can explode into verbal or physical violence. Mania also blunts and distorts learned social–cultural judgment while simultaneously stimulating a need for increased activity and excitement. In an attempt to alleviate these pressured feelings of desire, anxiety, anger, and grandiosity, the person behaves and makes decisions that are erratic and often dangerous. Without forethought or consideration of the consequences, individuals in a manic episode may run up large credit card debts, buy enormous quantities of a single and unusual item, engage in risky sexual activities with multiple unknown partners, drive recklessly, feel justified driving the wrong direction down a one-way street or disobeying traffic lights, indulge in large quantities of food and alcoholic drinks or go without eating and sleeping, drive aimlessly until the car runs out of gas, lose inhibitions and speak crudely or go nude in public, dress bizarrely, or verbally and physically lash out at others. Mania is always dangerous. As grandiosity is stimulated and judgment severely suppressed, the manic person can become extremely reckless and cause serious self-injury or death. The inability to foresee consequences and consider multiple solutions along with impulsive, agitated thinking and anger can induce rapid suicidal thoughts and behaviors. Depressed patients ruminate, plan, and deliberate the possibilities of suicide. For the depressed, death is often a means of ending mental anguish and hopelessness or stopping an unexplainable, but nonetheless constant drive and obsessive desire to die. In contrast, manic individuals kill themselves over poorly conceived, impulsive, almost momentary issues and feelings. Furthermore, severe mania often triggers physical violence and aggressive property destruction in people who have a more severe form of manic–depressive illness. People with a manic–depressive illness more often make death threats to presidents and other famous individuals than patients diagnosed with schizophrenia.

As the severity of a manic episode proceeds, thoughts can rush through a person's mind so fast that half-way through a sentence the beginning point is forgotten. This occurs when manic symptoms disrupt or block the brain's working memory. Under normal circumstances, working memory allows us to pull appropriate information from long-term memory, lock it in our mind, and manipulate the facts into interlinking complete thoughts and logical problem-solving models. In the most severe stages, mania can cause the person to experience psychotic hallucinations and delusions. Without medication, the person's manic hyperactivity and psychotic state will evolve into a stressful fatigue and loss of psychological orientation to time and place that causes the individual to appear completely confused, bewildered, and stupefied. In the past this was referred to as delirious mania. Fortunately today's modern medications and supportive treatment prevent most people with a manic–depressive disorder from reaching this level of severity.

Euphoria and hyperactivity that create difficulties but do not reach the level of severity of manic episodes is known as hypomania. This psychiatric condition

was first described in the late 1800s by the German psychiatrist Mendel. A hypomanic mood produces behavior that resembles the first phase of a full manic episode. The person has an elated mood, increased energy, rapid thinking and speaking, and reduced information processing skills. Though thoughts occur quickly, information is processed in a more narrow, concrete, and restricted manner. As an example, one's abilities to form alternative solutions, empathize, consider input from others, or perform problem-solving tasks requiring an exact sequential sequence are substantially reduced. Additionally, hypomania causes problems perceiving, organizing, and analyzing fragmented social information and interpreting social cues that have multiple meanings. As a result, hypomanic individuals make impulsive decisions, fail to consider behavioral consequences, and seldom perceive how others experience their actions. The symptoms may also include an inability to screen verbal communications. That is, the person feels an actual need or urge to voice almost every thought. When this symptom occurs, the individual interrupts others, talks incessantly, and has little concern if his or her words insult and upset the listener. Furthermore, hypomania, like the first phase of mania, can cause grandiose self-perceptions. During periods of grandiosity individuals overvalue their skills, status, or personal magnetism and may engage in behaviors like risky investments and business decisions, overspending and credit card debt; sexual experimentation and excess, and careless and reckless activities. A hypomanic person often displays seductive and addictive behaviors that may first appear as spontaneity or personality characteristics. A closer examination, however, will show that the individual's actions extend beyond the boundaries that are acceptable for most people within the same age and cultural group. Many times individuals with hypomania are labeled by families, schools, and community agencies as immature or delinquent and neither receive a referral nor seek mental health treatment.

Only a brief overview of depression symptoms is provided in this section. Readers are directed, however, to the complete article on the subject that is included in this volume. Depression is different from sadness, grief and bereavement, feelings of loneliness and isolation, or disappointment. Each of these situations creates a normal, but nonetheless unpleasant mood reaction to a real or perceived event. More importantly, one is able to shift away from the reaction, receive relief, and often block the feelings by engaging in nonrelated activities. That is, normal depressive feelings are mood reactions that seldom pervade every domain of our life for an extended period of time. Even when one's mood is lowered by a specific event, most individuals continue to experience a range of positive thoughts and feelings.

Unlike reactive sadness, depression is a downward spiraling or narrowing of feeling and emotional range across most major life domains for an extended period. Dulling and despondent feelings relentlessly occur from depression and prevent one from experiencing pleasure, accepting solace and praise, and finding emotional relief. Rumination over issues like regret, guilt, personal loss, shame, incompetency, disappointment, and hopelessness is often experienced as emotional stress and pain. Severe depression can also create a numbing emptiness and an inability to care about oneself, family, others, or the future. Additionally, during a depressive episode most people with a bipolar disorder experience not only a feeling of gloom but also restrictions and deficits in their cognition, information processing, motor, and perceptual skills. Concentration and working memory are always reduced by depression. Abstract thinking along with simple social cognitive information processing is greatly slowed. Other common symptoms include social withdrawal, insomnia or hypersomnia, weight loss or gain, fatigue, headaches, constipation, loss of sexual drive, and loss of interest and enjoyment in past pleasures or life skills. In severe depressive episodes a bipolar disordered person may also develop delusional thinking, hallucinations, catatonic states, or other forms of psychotic symptoms. Approximately 50% of all people with a bipolar disorder will exhibit psychotic symptoms. Additionally, even without entering a psychotic state, severe depression can usher in and maintain paranoid thinking for an extended period. To stop the emotional pain stemming from hopelessness, lost cognitive skills, and hurt from burdening others or feeling unloved or undeserving of love, far too many bipolar disordered patients during a depressive episode make serious and deadly suicidal attempts. Approximately 15% of individuals with a bipolar disorder make a serious attempt to take their life.

Another cluster of manic–depressive symptoms forms the mixed affective mood states or simply mixed episodes. A small subgroup of individuals with bipolar disorders will, for a week or more, concurrently experience the symptoms required for diagnosing a major depressive and a manic episode. This is a highly torturous state that simultaneously inflicts the rushing frenzy of mania and the restrictive negative sensations of depression. Even though the mood abnormality was

first described by a seventeenth century doctor, modern medicine continues to struggle with and debate the exact characteristics that define a mixed state. The prevailing symptoms can vary greatly for patients having a mixed episode. Some individuals, as an example, will feature psychotic symptoms, whereas others become highly irritable and yet others manifest more depressive behaviors. There is, however, growing evidence that a mixed state does not occur in all bipolar disorders. As a result, many experts believe that mixed episodes need to be thought of as a form of bipolar disorder or mixed mania that is separate from both depression and mania. Dysphoric mania is similar to mixed episodes, but the symptoms are less severe, often have a shorter duration, and do not qualify as full depressive and manic episodes. Patients who rapidly cycle between depressed and manic or hypomanic states can appear to be, but are not technically, in a mixed state.

The length of time between episodes of illness varies greatly among all patients with manic–depressive illness. For many people the cycle lengths shorten, causing more frequent episodes, then plateau at approximately 3–5 episodes per year, and finally shift to episodes that occur more or less annually. This appears to be the natural course of the illness. As the interval between psychiatric crises lengthens, most patients will have extended periods where their mood, and cognition, motor, and information processing skills return to normal. Unfortunately around 5–20% of bipolar patients have at least 4 manic or depressive episodes per year. This is known as rapid cycling, and episodes may take place in any combination and order.

Diagnostically, rapid cycling episodes do not differ in criteria from those that take place in non-rapid cycling. The symptoms must meet the criteria required for a manic, hypomanic, mixed, or major depressive episode. A smaller group of patients has ultrarapid cycling in which a depressive, manic, or a hypomanic episode may last for only a day or trigger multiple episodes within a 24-hr period. Unlike rapid and ultrarapid cycling, individuals with continuous cycling move through episodes without returning to their normal baseline or feeling normal for a significant period of time. With these individuals one depressive, manic, or hypomanic episode melts into another. Correct and prompt diagnosis of this illness is extremely important. Rapid cycling appears related to morbidity, is pharmacologically difficult to treat, and requires specific medication regimens. Studies show that these patients do not respond well to medications that are normally used with other bipolar disorders like lithium and antidepressants. Moreover, antipsychotic drugs may actually stimulate or exacerbate the cycling process.

V. CHILDHOOD ONSET

Even in the 1960s psychoanalytic theory hypothesized that children could not develop depression, let alone manic–depressive illness. This was founded on the assumption that depression results from unconscious superego conflicts. Because they thought children had not yet developed superegos, the assumption was that children also could not experience real depression. Today research documents that children, preteens, and adolescents can develop unipolar depression and bipolar disorders, but their course of illness and symptom patterns are often much different from these of adults. Manic–depression in children causes the general symptom categories found in adults but creates different patterns and behavioral problems. As an example, children have more chronic episodes distinguished by rapid and ultrarapid cycling. From the very beginning as infants many of these children behave differently from their peers. Mothers and nurseries report that the baby has long periods of crying, sleeplessness, and excessive activity.

As toddlers and preschoolers the children psychoanalytic theory destined for manic–depressive illness often are described as highly creative, bright, verbal, and intellectually ahead of most other children. Parallel to these positive developmental factors parents also recall that their children experienced severe separation anxiety and night terrors much longer than most youths. Additionally, many children who are developing manic–depressive illness have vivid, explicit dreams that typically focus on violent themes filled with blood and gore. This type of color-filled descriptive violence can also occur in the discourse of normal conversations. They may meet simple requests or instructive guidelines, such as gently requesting the child to remain in line, with an exaggerated physical threat. A parental ''no'' can send the manic–depressive child into a destructive rage for hours. During tantrums these children routinely cry, kick, hit, pull hair, bite, and attempt to punch holes in the wall or destroy other property. Discovery that the child has destroyed favorite personal items and damaged walls and furnishings when sent to a room for a time out punishment is a common experience among parents. Not only do many children rapidly cycle, often

switching from one mood state to another within hours, but additionally high degrees of agitation, unhappiness, anxiety, and anger can accompany both depressive and manic episodes. Manic episodes may include obsessive silly behaviors that grate on others, including peers. As the child's grandiosity increases, the youth may make unrealistic claims and brag about current and future fame. Mania can also cause the child to steal or hitchhike and travel in a totally unplanned and unprepared manner. Many times children in a manic episode will not know where they are going or why they are traveling. Still others will break into a house and take off all clothing for no apparent or logical reason. Impulsive behaviors highlight childhood mania and greatly increase the risk for suicide and self-injury.

The onset of a bipolar disorder during childhood usually starts with a depressive episode. Both unipolar and bipolar early onset depression may feature a distinct loss of pleasure in activities that had been pleasurable, tantrums or crying for almost no known reason, lethargy and oversleeping, an increase in self-consciousness, previously unseen or greatly increased phobic anxiety, increased agitation, arguing, and physical fighting. In almost all cases the child's school behavior and academic performance decrease, and many children develop or experience increased separation anxiety symptoms.

VI. TREATMENT ISSUES

When symptoms first start in children and adults, an extensive physical exam from a qualified physician is immediately required. Both depression and manic symptoms can stem from problems ranging from vitamin deficiencies or excess to major autoimmune, cardiovascular, gastrointestinal, endocrine, hematologic, neurological, and pulmonary diseases or malignancies. Additionally, a long list of medications and drug interactions can cause manic–depressive behaviors. Commonly, bipolar symptoms caused by medical problems other than a neuropsychiatric disorder will improve as the person recovers from the primary physical illness. Once alternative medical problems are ruled out and the person is diagnosed with a bipolar disorder, psychotropic medications become the first and fundamental treatment method.

Lithium along with the anticonvulsant medications carbamazepine and valproic acid are widely used as mood stabilizers. Lithium continues to be the most widely used drug for combating manic episodes and has also been found to relieve depressive symptoms for some individuals. Approximately 70% of bipolar patients experience significant benefits from taking lithium. This, however, is not without a cost. Lithium can cause numerous transient cognitive and physiological side effects and become life-threatening if the amount accumulating in the blood becomes toxic. When taking lithium, blood must be drawn routinely to monitor the blood's drug level and insure that toxicity does not occur.

Depression in bipolar disorders can be difficult to treat. Standard antidepressant medications, especially those known as tricyclics, can induce mania and rapid cycling in bipolar patients. This presents a particular hazard for individuals with bipolar II. These individuals often present as having recurring severe depression. The hypomania component can be hidden and difficult to diagnose. When this occurs, the person is at risk of being incorrectly diagnosed, prescribed an antidepressant, and sent into a state of mania or rapid cycling. Therefore, best practice guidelines recommend that bipolar depression be treated first with one of the mood-stabilizing medications mentioned earlier. Lithium is reported to be effective for 30–79% of bipolar depressed patients. The other antimania drugs are significantly less effective in treating depression. When the depressive episode is not reduced by lithium treatment, the psychiatrist will add either an additional mood-stabilizing medication or an antidepressant. Because tricyclic medications are known to increase mania, physicians most often put the patient on a monoamine oxidase inhibitor (MOAI) or one of the newer serotonin re-uptake inhibitor (SSRI) antidepressants. MOAI have generally been found to be the most reliable class of antidepressant medications for bipolar patients, but the drug tolerance and effectiveness, as with other psychoactive drugs, vary from patient to patient. Reports on best practice methods for treating bipolar depression consistently emphasize that antidepressants generally must be coadministered with a standard antimanic medication. Furthermore, unlike unipolar depression, there is no evidence that maintaining a patient on an antidepressant after the bipolar depression ends prevents the recurrence of future episodes.

Psychotherapy is always part of the comprehensive treatment of manic–depressive illness. During severe depressive or manic episodes, therapy anchors the person offering support, assurance, hope, and reflections of reality. As episodes lift and normal moods and thinking emerge, psychotherapy assists in logical and

systematic problem-solving, reinforcing positive behaviors and cognition, and understanding one's self and world. More specifically, psychotherapy can help individuals who have manic–depressive illness better understand their disorder, remain on prescribed medications, learn to predict and prevent or soften recurring episodes, learn to not be afraid when experiencing normal sadness, grief, or joy, how and when to let others take charge of their decision making, and how to better care for their significant others and friends. A combination of cognitive, behavioral, and supportive therapies can accomplish these goals with many individuals. An important part of psychotherapy is recording in a chart or journal how moods change over time, how they are affected by medication, and reactions in differing environments. Methods of this type are also used to identify internal thoughts or cognitive nonverbal self-talk. The messages we silently give ourselves can highlight strengths, fears, and oncoming depressive and manic episodes. By knowing that the patient's cognition and behaviors are changing dangerously, the treatment team can take steps for preventing or reducing the severity of an approaching illness. Prevention is accomplished by adjusting the person's medication, providing additional psychological support, and reducing stress in the person's home and work environment. Psychotherapy can help individuals make sense and meaning out of their physical and mental difficulties.

Families can provide important information concerning the patient's developmental history, illness onset, and current progress. Family members, however, also need support, understanding, and assistance from the treatment team. Living with an ill loved one is stressful and difficult. Psychotherapists or other treatment team members can offer families psychosocial education, guided problem-solving sessions, and methods for managing anxieties and organizing their thoughts. Psychosocial education provides the family information about the cause, treatment, care, and prognosis of bipolar disorders. Family members need to understand that manic–depression is a biological illness that can be treated. They also often need assistance in knowing how to relate with the ill family member and how to help with the on-going treatment. With the onset of illness, families not only must learn how to live together once again and redistribute work tasks but also concretely know how to respond to psychiatric emergencies. Family members, as an example, must know when and when not to call mental health professionals, police, clergy, and emergency rooms.

Treatment not only involves medication, psychotherapy, and education but also concrete support for the patient and the family. Comprehensive treatment must help patients and families gain real services and tangible resources. Bipolar disorders can greatly alter the patient's and the family's ability to earn money and provide transportation, housing, food, medication, and access to medical services. Family members and patients also need opportunities to separate and rest. Living daily with someone who is in the midst of a depressive or manic episode is extremely stressful. To prevent suicide or other hazardous behaviors, family members often remain with the patient continuously for an extended time. Other family members take on additional employment to pay for medical and rehabilitation expenses or spend countless hours coordinating services for the ill family member. Treatment teams need to not only understand the stressors experienced by families but substantially assist in the resolution of these problems.

VII. STRESS AND ONSET

Stress does not appear to be a major factor in explaining why individuals develop a bipolar disorder. Additionally, research does not support the idea that the number of depressive or manic episodes experienced by a patient relates to that individual's pre-onset stress level. There is also little or no scientific evidence that bipolar episodes are related to stress through the brain's kindling process. Kindling refers to the brain at the cellular level learning from repeated episodes to automatically trigger an event such as a seizure. Studies show that, after a number of electrically stimulated seizures, spontaneous epilepsy will occur without the introduction of electrical stimulation. The popular brain kindling and behavioral sensitization theory hypothesized that bipolar patients become highly aware of stress. Environmental and behavioral symbols for stress are then linked to repeated depressive and manic episodes. Eventually the brain will automatically trigger a relapse with only the slightest perception of a stressful stimulus. Whereas kindling has been shown to occur in laboratory animals, it has never been demonstrated in humans. No experimental or clinical research has shown that kindling occurs in the brain of individuals with a bipolar disorder.

There is some evidence that the onset of the first episode may partially be triggered by a stressful experience. This appears to be less true for depressive

and manic episodes occurring after the initial onset of the illness. Whereas stress may not explain why episodes occur, a strong relationship exists between the severity and frequency of symptoms and environmental stress. Reduction in interpersonal, community, and work stressors often increases the patient's level of functioning and decreases the manic–depressive symptoms. This is one of the immediate benefits some individuals gain from emergency hospitalizations and partial hospital programs. Stress management, however, is an adjunct therapy that serves as an enhancement to appropriate medication and psychotherapy treatment. Environmental manipulations cannot independently address and treat manic–depressive symptoms. Furthermore, people who are in the midst of an extremely severe episode may receive almost no relief from stress reduction and environmental enhancements.

VIII. NEUROCHEMISTRY

In the simplest terms, bipolar disorders occur from abnormalities in the brain's anatomy and chemistry. As a result, the brain is unable to orchestrate the dispatch and reception of chemical messages that direct, appropriately control, or modulate a person's mood. These are complex illnesses that cannot be explained by a single neurotransmitter or localized to one primary brain structure. Neurotransmitters are chemical messengers that carry electrical impulses between brain cells. Early research in the 1950s demonstrated that imipramine medication used for treating depression inhibited the re-uptake or increased levels of neurotransmitters known as neurogenicamines (neuroamines). Re-uptake is the process whereby chemicals that remain after transmission from one cell to another are either stored in the neuron's presynaptic terminal or eliminated by glial cells.

Norepinephrine and dopamine, considered the most important neurotransmitters in the neuroamine group, were thought to be the key for understanding bipolar disorders. This led scientists to hypothesize that depression developed principally as norepinephrine decreased and that mania occurred when the re-uptake system caused norepinephrine levels to increase excessively. In the 1960s this concept became known as the "catecholamine" hypothesis, because both norepinephrine and dopamine are classified as catecholamines. Dopaminergic pathways project throughout much of the brain but are highly concentrated in the nigrostriatal region. This area is most often associated with Parkinson's disease, stiffness, tremors, and other movement disorders. The nigrostriatal system ranges from the substantia nigra to the neostriatum and is responsible for controlling complex muscular movements and posture. As part of the dopamine pathway, the system also plays a role in the coordination of voluntary movement.

Individuals with manic–depressive illness can experience alterations in their coordination and movement in both depressive and manic episodes. It is also thought that dopamine located in the basal ganglia filters unwanted verbal expressions. This may explain why during periods of depression some individuals find word selection difficult, make distressed nonword sounds, walk with their head down, awkwardly swing their arms, distort their facial muscles, or fail to use complete sentences or speech structures. However, it is during manic episodes that people most often lose their verbal filtering skills. The most universal symptoms reported for manic patients are motor hyperactivity, rapid or pressured speech, and a decreased need for sleep.

An overabundance of dopamine appears to be related to psychotic episodes, whereas lower levels are associated with Parkinson's disease. Unfortunately, dopamine's complete role in manic–depressive illness is not yet understood. We do know that the neurotransmitter not only influences movement and verbal filtering but also is involved in arousal, attention, mood, reality testing, and cerebral blood flow. The reduced information processing, reality testing, and social judgment experienced during manic episodes may in part result from changes in dopamine levels. These symptoms are greatly influenced by the cerebral cortex, frontal and temporal lobes, and hippocampus and other limbic system areas. Dopaminergic pathways travel throughout and link each of these brain regions. Additionally, for some patients medication targeted at dopamine re-uptake provides greater relief from depression than drugs that impact serotonin neurotransmitters.

Early catecholamine research tended to focus on the actions and effects of a primary neurotransmitter rather than the interactions occurring among neurochemicals and brain structures. Manic–depression was often described as a chemical imbalance primarily caused by changes in a person's norepinephrine level. Today we know that a disruption or change within a single re-uptake system only partially explains why individuals uncontrollably cycle through states of

depression and mania. Additionally, not everyone is adversely affected by an increase or decrease in norepinephrine levels. There is strong evidence that bipolar disorders involve defects in complex interplays among multiple neurotransmitters and physical changes in the brain's structure. Chemically, moods appear to be controlled or destabilized by changes in the level and interaction of norepinephrine, serotonin, dopamine, and other neurochemical factors.

Although norepinephrine cannot stand alone as an explanation for manic–depressive illness, it nonetheless continues to be viewed as a major contributor to the illness. The neurotransmitter's circuitry illustrates why it is an important link in understanding bipolar disorders. Norepinephrine pathways originate in the pigmented locus ceruleus of the brain stem and project through the limbic system and other major brain areas. The limbic system includes the hypothalamus, amygdala, hippocampus, and septal regions. These cortical and subcortical areas play a significant role in the regulation and modulation of moods, emotions, memory, and how one responds to events in the external world. Research has consistently found that individuals with bipolar disorders have reduced norepinephrine levels. Catecholamine metabolites such as norepinephrine are excreted in urine and provide a proportional measurement to that of the primary brain source. Most, but not all, studies have found that urinary measurements of norepinephrine are substantially reduced in bipolar patients compared with individuals who have either unipolar depression or no mental disorder. The severity of manic episodes may be an influencing factor in urinary catecholamine levels. Norepinephrine levels often differ between bipolar I disordered individuals and unipolar depressed patients, but not between those with a milder bipolar II disorder and patients who experience unipolar depressive episodes. Additionally, post mortem brain studies show that an increased density of norepinephrine receptors is highly associated with a history of major depressive episodes. When synapses are deprived of transmitter molecules, the postsynaptic cells compensate by increasing the number of receptors and attempt to retrieve any additional neurochemical signals that are available. Furthermore, medications that block norepinephrine re-uptake and increase the neurochemical in synapses reduce depressive symptoms for some individuals.

In the late 1960s, researchers found that the neurotransmitters norepinephrine and serotonin not only contribute to shaping a person's mood but also impact upon each other. Serotonin-producing neurons project from the brain stem's raphe nuclei through the amygdala, hypothalamus, and cortical areas. These regions produce the major symptoms experienced during depressive and manic episodes. As an example, the amygdala controls mood and affective responses, whereas the hypothalamus regulates appetite, sleeping patterns, and sexual drive and the frontal cortex assists with social judgment, information processing, and concentration. A person with manic–depressive illness will experience a reduction or slowness in these functions during a depressive episode and a flooding or racing of uncontrolled mental activities when the manic episode starts. This cycling between lethargic depression and racing mania is thought to occur in part because the serotonin circuitry incorporates or interacts with neuron sites that secrete or control the release of norepinephrine. It is thought that serotonin reaches a point of depletion that triggers or actually permits the brain's norepinephrine level to significantly drop. Furthermore, because serotonin-producing cells are found in many diverse brain regions, for some individuals, depression may result from inadequate serotonin levels interacting with neurons and neurochemicals other than norepinephrine.

The idea that serotonin depletion triggers a change in norepinephrine or other neurochemical levels has become known as the permissive hypothesis and continues to receive attention among researchers. Though the exact roles of these neurotransmitters remain unclear, there is no doubt that they both play important roles in manic–depressive illness. Scientists, as an example, can measure the brain's serotonin level by the amount of a serotonin byproduct found in a person's cerebrospinal fluid. Studies of cerebrospinal fluid consistently find that the serotonin levels in the brain for depressed and suicidal individuals as a group are significantly below those found in populations with no history of mood disorders. Additionally, post mortem studies report that, compared to people with no history of a mood disorder, depressed and bipolar disordered individuals have an increased density of serotonin receptors (type 2). This, like the increase in norepinephrine receptors, suggests that the brain may be attempting to compensate for extended periods of reduced serotonin in the synaptic cleft.

IX. NEUROANATOMY

The brain's inability to control neurochemical functions may result in part from anatomic abnormalities. Individuals with bipolar disorders have specific areas

of their brain that structurally differ from the neuroanatomy of individuals with no history of mental illness. Furthermore, there is growing evidence that differences in brain structure are also found among individuals with manic–depressive illness and other major diagnostic categories such as major depression and schizophrenia. University of Michigan researchers using positron emission tomography (PET) found that individuals with bipolar illness have a higher density of monoamine-releasing cells than people who do not have an affective disorder. These specialized cells are responsible for controlling the discharge of norepinephrine, serotonin, and dopamine. Magnetic resonance imaging (MRI) studies show that individuals with bipolar disorders have significantly enlarged lateral ventricles, frontal and temporal lobe sulci, and Sylvian fissures. Studies of third ventricular enlargement in bipolar disorders are at best mixed and often contradictory. This appears to be one of the anatomical differences between manic–depressive illness and schizophrenia. That is, there is strong evidence in schizophrenia, but not bipolar disorders, that lateral and third ventricular enlargement may be caused by a neuropathological process that is independent of sulcal changes. Furthermore, lateral ventricular enlargement and sulcal prominence have been found in many, but not all, patients with manic–depressive illness. Therefore, we currently cannot use neuroanatomical measurements as a means of diagnostically determining bipolar or other mood disorders.

Our understanding of how neuroanatomy differs between psychiatric diagnoses was greatly advanced by MRI studies of monozygotic twins discordant for either bipolar I illness or schizophrenia. Although these are genetically identical twins, one of each pair developed either manic–depressive illness or schizophrenia whereas the other twin remained well. The MRI study found a strong trend for the ventricular enlargement among twins affected with either bipolar I or schizophrenia disorders to be similar. When measured just within each discordant group, differences between right lateral, left lateral, and third ventricles were much more statistically significant. There was a greater difference, however, between the twins affected with schizophrenia and their well sibling twins than in the twins discordant for bipolar illness. Significant differences were also found between the schizophrenia-affected twins and their well sibling twins in the right and left hippocampus, amygdala, and basal ganglia areas of the brain. Interestingly, these differences did not significantly occur in the twins discordant for bipolar illness. This is at odds with numerous MRI nontwin studies of manic–depression. The lack of agreement may exist because the middle and posterior portions of the twins' hippocampus are yet to be studied.

X. CONCLUSION

Manic–depression is a spectrum of neurobiological illnesses that involve abnormal changes in the brain's chemistry and cellular structures. The neurotransmitters serotonin, norepinephrine, and dopamine appear to play key roles in triggering manic–depressive episodes. Anatomical damage or changes have been documented in the lateral ventricles, frontal and temporal lobe sulci, and Sylvian fissures. How bipolar disorders specifically start is yet to be explained. Evidence from family histories, adoptions, and twins strongly indicates that the disorders are genetically transmitted. There is, however, growing evidence that some individuals may develop manic–depressive illness from an *in utero* neurovirus. Environmental stress does not appear to play a major role in the causation of bipolar disorder or the number of episodes experienced by a person. Brain kindling and behavioral learning theories have not been helpful in explaining the cause or course of these disorders. Once individuals are in the midst of an episode, the reduction of environmental stress does appear to help reduce or soften the symptoms. Nonetheless, medications remain the principal treatment for bipolar disorders. Psychotherapy, psychosocial education, and environmental manipulation therapies serve as important adjuncts or additive components to treatment but are never the primary or sole intervention. This is particularly true for bipolar I and II and schizoaffective disorders. Approximately 70% of patients with bipolar I disorder significantly improve after taking lithium. Psychotherapy alone seldom has a meaningful impact on severe manic–depressive episodes. The future is extremely hopeful for individuals with bipolar disorders. New medications are currently being studied clinically, and our knowledge of how the brain functions is rapidly growing. Researchers are also learning to identify early risk factors and minor symptoms that signal the onset of illness. Unfortunately, manic–depression is an illness that cannot yet be prevented and continues, for many people, to be misdiagnosed. However, as our knowledge of medications, brain functioning, and early symptoms grows, science will be better positioned to effectively prevent and treat manic and depressive episodes.

See Also the Following Articles

AUTISM • BEHAVIORAL PHARMACOLOGY • COGNITIVE PSYCHOLOGY, OVERVIEW • DEPRESSION • DOPAMINE • MOOD DISORDERS • NEUROPSYCHOLOGICAL ASSESSMENT • NEUROPSYCHOLOGICAL ASSESSMENT, PEDIATRIC • NOREPINEPHRINE • PARKINSON'S DISEASE • SCHIZOPHRENIA • STRESS

Suggested Reading

Bearden, C. E., Hoffman, K. M., and Cannon, T. D. (2001). The neuropsycholology and neuroanatomy of bipolar affective disorder: A critical review. *Bipolar Disorder* **3**(3), 106–150.

Beckham, E. E., and Leber, W. R. (1995). *Handbook of Depression*, 2nd ed. Guilford Press, New York.

Goodwin, F. K., and Jamison, K. R. (1990). *Manic–Depressive Illness*. Oxford University Press, New York.

Torrey, E. F., Bowler, A. E., Taylor, E. H., and Gottesman, I. I. (1994). *Schizophrenia and Manic–Depressive Disorder: The Biological Roots of Mental Illness as Revealed by the Landmark Study of Identical Twins*. Basic Books, New York.

Soares, J. C., and Gershon, S. (Eds.) (2000). *Bipolar Disorders Basic Mechanisms and Therapeutic Implications*. Marcel Dekker, New York.

Young, L. T., and Joffe, R. T. (Eds.) (1997). *Bipolar Disorder: Biological Models and Their Clinical Application*. Marcel Dekker, New York.

Memory Disorders, Organic

ANDREW R. MAYES
Liverpool University

GLOSSARY

amnesia Either a nonspecific term for any kind of memory disorder (usually one caused by brain damage rather than arising for psychological or motivational reasons) or the global amnesia syndrome in which an impairment of recall and recognition of postmorbid facts and episodes (anterograde amnesia) is accompanied by varying degrees of recall and recognition impairment for premorbidly acquired memories of facts and episodes (retrograde amnesia), together with preservation of intelligence and working memory.

declarative memory Memory for facts and personally experienced episodes that the subject is aware of remembering. As such, the memories can be declared either by verbal statements or by some non-verbal means such as drawing. It is contrasted with procedural or nondeclarative memory, in which subjects cannot typically indicate what they are remembering except in the broadest manner.

episodic memory Memory for particular incidents in personal life that have specific spatial and temporal contexts. It is often contrasted with semantic memory.

explicit memory A form of memory contrasted with implicit memory; the distinction between the two is very similar to that between procedural and declarative memory. Like declarative memory, explicit memory often involves intentional retrieval of factual or episodic information.

implicit memory A form of memory contrasted with explicit memory that corresponds closely with procedural or nondeclarative memory. The presence of memory is only indicated indirectly and the subject does not intentionally try to remember anything specific but tries to perform a task well, with changed performance indicating the presence of memory.

priming A form of information-specific implicit, procedural, or nondeclarative memory in which remembering is indicated by a change in the way that previously studied information is processed. In other words, studied information is processed faster, more accurately, or in some other way more efficiently, although subjects need have no awareness that they are remembering the information.

procedural memory Memory that is only accessible through performance in an indirect or implicit fashion and comprises skill memory, conditioning, priming, and simple forms of nonassociative memory. It is sometimes referred to as nondeclarative memory by those who wish to stress that it may be constituted of a heterogeneous collection of different kinds of memory, the only common feature of which is that they are not forms of declarative memory.

semantic memory Memory for facts or the kinds of information that can be stored in our mental dictionaries and encyclopedias. Unlike episodic memory, semantic memory contains no necessary reference to the personal context(s) in which it was acquired although some facts necessarily include reference to the non-personal contexts in which (for example, historical) events occurred. It is controversial, however, whether retrieving personal contextual information may nevertheless sometimes be important in remembering facts. Resolving this issue is difficult because a common but nonessential difference between semantic and episodic memory is that the former tends to be much more rehearsed than the latter.

working memory The temporary storage for a few seconds of information that is being processed in any of a range of cognitive tasks.

Damage to different parts of the brain impairs both short-term and long-term memory for different kinds of information often in very selective ways that leave basic processing intact. Lesions to association neocortex cause several different short-term memory

disorders in which memory for phonological, visuospatial, color, or other kinds of information is selectively disrupted within approximately 1 sec of presentation. These immediate memory disorders occur in the presence of normal processing and are associated with longer term memory deficits for the same kinds of information. Several different kinds of long-term memory disorders exist. Posterior neocortical lesions disrupt semantic memory, but the deficit also seems to include episodic memories and to affect older memories much more severely than newer ones. Frontal neocortical lesions disrupt the ability to plan and organize processing in a variety of ways, and this causes impairments of long-term and perhaps (less severely) short-term memory for episodes and facts. At least in monkeys, there is some evidence that ventromedial frontal cortex lesions cause organic amnesia. This syndrome, which is known to be caused by lesions to the medial temporal lobe, the midline diencephalon, the basal forebrain, or their connections, involves pre- and postmorbid impairments of memory for facts and episodes but leaves intelligence and short-term, overlearnt semantic, and several forms of procedural memory intact. Processing of factual and episodic information seems to be relatively preserved. Priming of information already familiar at study is preserved in amnesia, but whether priming of previously novel information is preserved remains controversial. Perceptual priming has been shown to be impaired after posterior neocortical lesions. Other forms of procedural memory are disrupted by subcortical lesions. Thus, motoric classical conditioning is impaired by cerebellar lesions and skill learning is particularly disrupted by basal ganglia lesions. The range of distinct memory disorders has led to the view that there are several different memory systems, each dealing with a specific kind of information, dependent on different brain regions, and perhaps also dependent on qualitatively distinct kinds of memory processes.

I. INTRODUCTION

It has been known for more than 100 years that brain damage can impair memory in a relatively selective fashion. In the past 50 years, it has also become clear that damage to different brain regions selectively disrupts some kinds of memory, but leaves other kinds of memory intact. Since the 1970s, it has become increasingly possible to scan the brains of living memory-impaired subjects to better identify the location of damage that causes specific kinds of memory deficit. In the 1990s, research exploring the effects of differently located brain lesions on memory was complemented by neuroimaging studies that have explored the brain regions that are activated when particular memory processes are engaged. These neuroimaging studies have helped clarify the normal role of brain structures that when damaged cause different kinds of memory disorder.

Memory depends on encoding (processing and representing) different kinds of information, which are then stored and later retrieved. The occurrence of selective memory disorders has several implications about the way in which the brain mediates these processes. First, the existence of selective memory disorders is surprising because it means that memory loss occurs although the information for which memory is impaired is still being processed (and hence encoded) relatively normally. This challenges a widely held assumption that information is stored in the same neural network that represents it during encoding because if the assumption is correct, then brain lesions that damage such networks should disturb not only storage but also the ability to process and represent the information. In other words, the assumption seems difficult to reconcile with the occurrence of selective memory deficits for particular kinds of information that are still processed normally at input. Because retrieval and encoding involve many overlapping processes, the assumption is still problematic even if it was postulated that a memory deficit was caused by a retrieval deficit. Therefore, care must be taken to determine the selectivity of memory disorders.

Second, the existence of different kinds of memory disorders is consistent with growing evidence that the whole central nervous system is, to varying degrees, plastic and thus capable of storing memories. The spinal cord, as well as the brain, should be regarded as a storer as well as a processor of information. That different kinds of memory should be disrupted by lesions in different brain areas should not be surprising because it is known that different kinds of information are processed and represented in different parts of the central nervous system.

Third, the existence of several different kinds of selective memory disorders caused by lesions in distinct brain regions has resulted in the view that there are several systems of memory organized in a hierarchical fashion. Memory systems differ from each other not only because they are mediated by distinct systems of neurons but also because they involve qualitatively distinct kinds of memory processes. Memory systems are organized hierarchically to the

extent that certain kinds of memory share more qualitatively similar kinds of processes with each other than they do with other kinds of memory.

What is known about the range of short- and long-term memory disorders caused by brain damage and the selectivity of these disorders to memory are outlined before briefly considering what this reveals about how the brain mediates memory.

II. WORKING MEMORY DISORDERS

A. Impairments of Phonological and Visuospatial Working Memory

It has long been believed that memory that lasts for only a few seconds without rehearsal depends on different storage processes than does memory that lasts for minutes or longer without rehearsal. Short-term storage may depend on continued patterned activity in the neurons that represent information during encoding, whereas longer term memory may depend on structural changes between the same neurons that represent the information during encoding and short-term memory. If brain lesions can disrupt short-term memory for specific information, then this view would lead one to expect that long-term memory for that information would also be disrupted. Impaired processing of the poorly remembered information might also be expected.

Cortical damage can impair short-term or working memory while leaving processing relatively intact. This pattern of impairment has been explored most extensively with phonological short-term memory, for which the ability to hold sequences of phonemes in mind for several seconds has been found to be disrupted by lesions to the left parietal lobe, particularly in the left posteroinferior parietal lobe where it conjoins with the left temporal lobe. A patient with such damage may only be able to repeat one or two spoken digits when tested immediately after their presentation compared to the normal level of approximately seven digits, and spoken nonword repetition ability is lost with similar rapidity. Even the ability to hold one digit in memory is lost pathologically fast if rehearsal is prevented. Despite this rapid loss of phonological information, whereas some patients may be dysphasic, others show no other cognitive deficits, clearly understanding speech and seemingly processing phonemes normally when memory load is minimized. For example, some patients have been shown to be unimpaired at making same–different judgments with spoken syllables. Therefore, impaired phonological working memory can occur in the presence of apparently normal processing of phonemes.

In his model of working memory, Baddeley postulated that there was a rehearsal loop as well as a short-term store for phonological information. There is evidence that the phonological short-term store and the rehearsal loop can be disrupted separately. Thus, although some patients have impaired phonological working memory despite normal rehearsal, there are other patients whose impaired phonological working memory is caused by deficient rehearsal. Interestingly, anarthria is not sufficient to cause an impairment of phonological working memory, which suggests that rehearsal of articulated information is mediated centrally and does not require overt speech.

Baddeley's model of working memory also postulated the existence of a visuospatial short-term memory store as well as a phonological one. There is evidence that working memory for visuospatial materials can be disrupted separately from phonological working memory. Thus, whereas some patients have impaired phonological short-term memory and intact visuospatial short-term memory, others have been found to show the reverse pattern of deficit sometimes following lesions in the region of the right Sylvian fissure. In other words, differently located cortical lesions can disrupt these two forms of working memory separately.

B. Other Kinds of Working Memory Disorder

Lesions also disrupt forms of working memory not postulated in the original working memory model of Baddeley. Thus, visual verbal short-term memory can be selectively disrupted, and there has also been a report of a patient with a selective short-term memory deficit for color information in the presence of normal processing of color when memory load was minimized. There is also evidence that relatively selective lexical semantic short-term memory deficits exist. Thus, a patient has been described who was more impaired than a second patient, who had a phonological short-term memory deficit, on tests dependent on lexical semantic short-term memory (e.g., word span tests) but performed better although not completely normally on tests primarily dependent on phonological short-term memory such as non-word span tests. This semantic short-term memory deficit was not caused by

a semantic processing impairment because the patient's semantic processing was usually normal when memory load was minimized. The patient's lesion involved the left posterolateral frontal lobe and adjacent parietal regions anterior to where damage probably causes phonological working memory deficits.

Future work will examine the possibility that each sensory system and the motor system has one or more short-term memory systems by exploring whether the corresponding selective short-term memory deficits exist. This work will be guided in part by neuroimaging studies of working memory, which have already supported the implications of the lesion studies. For example, visuospatial and visual object working memory tasks have been found to activate distinct cortical regions. This leads to the expectation that there should be lesions that separately disrupt working memory for these two kinds of visual information.

C. Effects of Working Memory Disorders on Long-Term Memory

Until recently, it was widely believed that long-term memory is preserved in patients with short-term memory deficits because patients with impaired phonological short-term memory show normal long-term memory for spoken verbal materials. This finding was interpreted to mean that short-term memory processes do not trigger long-term memory processes, and the two are mediated by separable groups of neurons. These conclusions do not hold, however, because the preserved long-term memory shown by patients is almost certainly for semantic information, whereas their short-term deficit is for phonological information. Patients are able to recode phonological inputs into a semantic code very rapidly so that the recoding can be achieved even in the presence of very fast loss of phonological information. If care is taken to ensure that the phonological information cannot be recoded, then long-term memory might well be impaired. This has been shown in a patient with very impaired phonological short-term memory who was also found to be completely unable to learn spoken Russian words transliterated into her native Italian. In other words, she was impaired at both short- and long-term memory for meaningless spoken words, which she had to represent as phonological sequences. In a similar way, it has been found that another patient not only had impaired visuospatial short-term memory but also had severely impaired long-term memory for spatial layouts and new faces. A third patient, who had a semantic short-term memory deficit, was also very impaired at long-term memory for lexical semantic information (e.g., word lists) but showed preserved long-term memory for any information for which he had normal short-term memory.

These results imply that short-term memory disorders are specific to particular kinds of information and can occur despite preserved processing of that information at input. Short-term memory for different kinds of information is mediated by distinct cortical systems of neurons, but it remains to be fully explored how these "working memory" neurons relate to the neurons that represent the information at initial coding. Short-term memory disorders are accompanied by long-term ones for the same information. This strongly implies that the same systems of neurons are involved in both short- and long-term storage for specific kinds of information such that damage to a specific system disrupts both short- and long-term memory for the information that it represents. The pattern of deficit, however, neither confirms nor refutes the view that short-term memory processes are necessary for *triggering* long-term ones.

III. DEFICITS IN PREVIOUSLY WELL-LEARNED SEMANTIC INFORMATION

A. The Relationship between Semantic and Episodic Memory Disorders

Memory for the kinds of fact that one might encounter in an encyclopedia or a dictionary (semantic memory) is often contrasted with episodic memory, which concerns personally experienced episodes. The two kinds of memory must interact closely, however, because the kinds of information they involve greatly overlap. Thus, memory for personally experienced episodes typically includes facts as well as perceptual information, and memory for facts may include memory for public episodes with their accompanying contextual markers (a key feature of episodic memory) and perceptual features. In comparing deficits of semantic and episodic memory, much attention needs to be paid to the information involved, and the amount of rehearsal that each kind of memory has received, because although this is not a defining feature, it tends to be greater for semantic memories. Semantic memory is certainly disrupted by brain lesions, but episodic memory may often be similarly disrupted when

allowance is made for contaminating factors such as the amount of rehearsal.

Neocortical lesions, particularly to the anterolateral temporal cortex, cause impaired long-term memory for previously well-established semantic memories. These semantic memory deficits are found in dementing conditions, such as Alzheimer's disease and the variant of frontotemporal dementia known as semantic dementia, as well as following closed-head injury and herpes simplex encephalitis. Semantic dementia patients are impaired at naming, identifying, and describing the properties of objects, at defining spoken and written words, and at identifying semantic commonalities between pictures, but they show preservation of perception, non-verbal intelligence, and syntactical and phonological abilities. It is important to determine how selectively these deficits affect factual memory, how selective the dissociations between different kinds of factual memory deficits are; whether the location of cortical damage determines the specific factual memories disrupted; and whether the deficits reflect access or storage breakdowns.

There is an incomplete double dissociation between semantic dementia and organic amnesia. Semantic dementia has usually been regarded as a selective disorder of semantic memory, and amnesia has been regarded as a selective disorder of episodic memory because memory for premorbidly overlearned facts is preserved. Therefore, this dissociation has often been interpreted as evidence that memory for facts and memory for episodes are mediated by partially distinct brain mechanisms. The two memory systems clearly interact, however, although the nature of these interactions is poorly understood. Such interactions have been used to explain the incompleteness of the double dissociation between semantic dementia and organic amnesia. Thus, it is argued that semantic dementia patients typically acquire episodic memories relatively normally but do not show completely preserved episodic memory because new episodes will often involve factual information that they have forgotten (so that semantic dementia patients will not be able to make full sense of an episode). Failure to interpret new episodes in a normal semantic fashion reduces episodic memory because people show worse episodic memory when semantic encoding is minimal. Conversely, it can be argued that amnesics fail to learn new facts normally because their episodic memory impairment prevents the rapid acquisition of new fact memories although these eventually become independent of episodic memory. The implication is that facts can initially only be acquired rapidly as components of episodic memories, which are selectively impaired in global amnesia.

There is evidence that suggests that "semantic dementia" is a misnomer. This is because the disorder does not seem to be selective to semantic memory; and it also involves an episodic memory deficit that is not secondary to the semantic memory impairment. It was found that semantic dementia patients have an inverted temporal gradient of deficit in which remote autobiographical memories (formed when semantic memory was normal) were more impaired than recent ones, which were sometimes preserved. Similarly, better recognition of famous faces and recall of information about them was found if these faces were encountered during the current time period rather than earlier.

These findings suggest that whether memories are semantic or episodic may not matter. Rather, the only thing that matters may be the memories' age. Thus, in semantic dementia, new facts and episodes seem to be learned as well as remaining semantic memory permits and such information is perhaps retained normally for a while. If memory depends minimally on encoding factual information, then in the short term it may be almost normal. For example, forced-choice recognition memory for recently studied real and chimeric animal pictures is relatively normal in semantic dementia patients. This kind of memory probably depends much more on perceptual than on factual information. However, very long-term storage even of information that can be mainly encoded in terms of perceptual features is very impaired in semantic dementia, because patients have damage to the neocortical structures, responsible for very long-term retention of facts and episodes, so that over a period of years both kinds of information are lost pathologically. This interpretation is consistent with the widely held hypothesis of medial temporal lobe (MTL) amnesia, which states that binding information for fact and episode memories is initially stored in the MTL but through processes such as rehearsal there is a gradual reorganization of storage that is transferred for very long-term maintenance to neocortical structures such as the anterolateral temporal cortex.

A few cases of very severe global amnesia have been reported in which very impaired episodic memory was accompanied by some ability to acquire postmorbid vocabulary and semantic facts, although memory for these things was also very impaired. Such findings can plausibly be regarded as evidence that there is a slow neocortical learning mechanism that gradually creates a long-term memory for semantic information over

many learning trials. A disproportionate deficit for episodic memory relative to memory for postmorbidly encountered facts may arise simply because facts, but not particular episodes, are repeatedly encountered. In other words, it may be an artifact of the number of learning trials that facts and episodes typically receive. This would be true if the slow neocortical learning mechanism has the ability to slowly create memories for personally experienced episodes as well as facts. Whether different neocortical lesions can cause dissociable deficits for previously well-established semantic and episodic memories is currently unresolved. Dissociation seems likely to the extent that episodic information differs from semantic information, provided one assumes that memories are stored where they are represented and that different information is represented in different neural structures. In this assumption, semantic memories for different kinds of information are likely to be disrupted by different cortical lesions, and the same may apply to episodic memories of different kinds of episode. The different effects of cortical lesions on memory for different kinds of fact and episode may be much more striking than any general differences between semantic and episodic memory. This issue remains to be systematically explored. However, there is evidence that different cortical lesions do dissociably disrupt different subtypes of semantic memory.

B. Subtypes of Semantic Memory Disorder

There is good evidence that semantic memory for different categories of information breaks down in a dissociable manner although how such dissociations should be interpreted remains controversial. There are examples of dissociations, including ones between impairments of word and object knowledge and between knowledge of abstract and concrete words, respectively. Some semantic memory deficits can be extraordinarily specific. For example, following a stroke, one patient was found to have an impaired ability to name pictures and objects from the categories of fruits and vegetables. This patient could name other food objects and all nonfood objects without difficulty, so his difficulty with name retrieval was specific to the semantic categories of fruits and vegetables. The most explored semantic memory deficits have been those for animate category and inanimate category knowledge. Memory deficits for animate and inanimate categories of knowledge have frequently been shown to dissociate from each other. It is still disputed by some whether this dissociation reflects uncontrolled differences between the categories tested in variables such as frequency, familiarity, and age of acquisition. Nevertheless, the dissociation might be expected if animate concepts primarily involve visual perceptual properties, whereas inanimate concepts primarily involve functional properties that are less visual in nature. The evidence is conflicting that this is the case. For example, this hypothesis has difficulty explaining why some patients with animate category knowledge deficits are impaired at identifying nonvisual as well as visual features of living things. One possibility that is supported by neuroimaging work is that visual information has to be retrieved even when nonvisual information about living things is retrieved (as revealed by activation of the left fusiform cortex), whereas this is not true about nonliving things. It remains to be seen whether this neuroimaging work can be reconciled with lesion evidence that suggests that more posterior temporal cortex lesions produce deficits in memory for man-made things, whereas more anterior temporal cortex lesions disrupt memory for animate things.

C. Evidence of Storage or Access Failure

In principle, deficits in semantic memory could occur because of degradation in storage, because of a disturbance in processes that access the stored information, or because both of these problems are present. In the 1980s, Shallice proposed five criteria for deciding whether a semantic memory deficit reflects a problem with keeping fact memories in store or with accessing them. The two most plausible of these criteria involve consistency of success or failure of retrieval and the presence or absence of priming or information-specific implicit (unaware) memory. Shallice argued that consistently unsuccessful retrieval of particular items on different occasions implies that storage has been degraded, whereas retrieval should be relatively consistently successful across occasions if storage is still intact. Conversely, if information is not recognized or recalled but is associated with normal priming, then the explicit memory deficit reflects an access problem selective to aware memory. This is because the same information is still available to unaware memory through the presumably distinct access mechanisms required for retrieving memories without awareness (priming). Minimally, Shallice's argument requires that consistent failure should not be found with intact implicit memory. The criteria

currently lack a strong theoretical base, so it remains unclear whether they cover the effects of partial storage damage or even whether, for example, consistent failure might sometimes reflect access rather than storage deficits. Nevertheless, there is evidence that, in semantic dementia, when previously retrievable memories become inaccessible, they remain so consistently, so there is a reliable pattern of semantic memory breakdown that requires interpretation.

IV. ORGANIC AMNESIA AND FRONTAL LOBE DAMAGE: EPISODIC AND SEMANTIC MEMORY

A. Organic Amnesia

The most intensively studied of the organic memory disorders is the global organic amnesia syndrome, which was first characterized in the 19th century. In this syndrome, there is impaired recall and recognition of facts and episodes encountered both postmorbidly (anterograde amnesia) and (more variably) premorbidly (retrograde amnesia). These deficits often occur even when intelligence and shortterm memory are preserved. There is also preservation of overlearned semantic memories, of various forms of motor, perceptual, and cognitive skill learning and memory, and of at least some kinds of unaware memory for specific information (priming). The deficits can be produced by lesions to structures in the MTL, midline diencephalon, basal forebrain, and possibly ventromedial frontal cortex as well as to structures that link these regions, such as the fornix.

There is growing evidence that the syndrome is functionally heterogeneous. First, it has been claimed that retrograde amnesia can be produced relatively independently of anterograde amnesia. Several studies have found poor correlations in patients between severity of anterograde amnesia and deficits in more remote premorbid memory. There have also been several reports of relatively selective or focal retrograde amnesia in which patients had a severe and enduring deficit in premorbid memory but a relatively preserved ability to acquire new memories (so that old memories could be ''relearned'' to some extent). The location of lesions that cause this condition remains to be accurately resolved, but damage often includes the anterolateral temporal neocortex, particularly on the left. There have also been reports of patients with focal retrograde amnesia who had damage to their temporopolar and frontal cortices, mainly on the left. These patients showed extensive retrograde amnesia, but their new learning abilities were relatively intact. The results suggest that old memories about autobiographical and public information may be disrupted by lesions to parts of the temporal cortex. Damage to prefrontal cortex can also disturb these memories, but it is unproved that such lesions alone can disrupt premorbid memory relatively selectively. It needs to be determined more precisely which temporal (and possibly parietal) lesions disrupt remote premorbid memories and how these relate to the damage underlying semantic dementia. There is also a particular difficulty with these focal retrograde amnesia patients in proving that they are not malingering or suffering from a psychogenic deficit. In such cases, the premorbid memory deficit is not directly caused by brain damage but may reflect the patient's conscious wish to achieve financial gain or avoid something unpleasant or the patient's unconscious avoidance of something traumatic. These cases can appear very similar to those arising from brain damage.

Second, it has been argued that amnesia is heterogeneous because lesions to different MTL structures impair memory in distinct ways. It is generally agreed that amygdala damage does not usually cause anterograde amnesia, in contrast to other MTL lesions. However, the amygdala probably plays a role in emotional memory by modulating the effectiveness with which other MTL structures store memories when emotionally evocative information is encoded. This role of the amygdala would explain why events associated with marked emotional arousal are so well remembered (flashbulb memory). The effects of amygdala lesions on memory have been very little explored in humans because of the rarity of selective amygdala lesions.

Some researchers believe that lesions of either hippocampus or perirhinal cortex within the MTL disrupt memory in the same way, but that hippocampal lesions may have a lesser effect. This view implies that free recall and item recognition are equivalently disrupted after MTL lesions. In contrast, it has been argued that hippocampal and perirhinal cortex lesions cause dissociable deficits. This view is based on animal and human evidence. First, animals with perirhinal lesions have shown item recognition impairments but intact spatial memory, whereas the reverse effect has been found with hippocampal lesions. Second, meta-analysis of human recognition data has suggested that there is a single dissociation in which patients with damage to the hippocampus or other parts of Papez circuit, such as the fornix, mammillary bodies, or

anterior thalamus, are relatively unimpaired on item recognition but as impaired as more generally lesioned global amnesics (who are also impaired at item recognition) on tests of free recall.

Several patients with apparently selective hippocampal damage were found to be clearly impaired on item recognition when an extensive battery of tests was given. However, patients with selective hippocampal sclerosis related to temporal lobe epilepsy were reported not to show impaired item recognition, and other patients with probable hippocampal damage caused by hypoxia showed a similar preservation. In addition, three patients with evidence of early selective hippocampal damage were shown to have completely selective free recall deficits because their item recognition was normal on a range of tests. Similar patterns of memory performance with relatively preserved item recognition have been found in patients with selective hippocampal damage acquired in adulthood, so it is unlikely that the pattern of selective free recall deficits reflects reorganizational processes following early brain damage.

Apparently selective hippocampal damage, sometimes causes severe item recognition deficits and sometimes leaves item recognition relatively or completely intact. Animal studies have shown that cerebral ischemia often produces damage that extends beyond the hippocampus. Because such damage may be both highly variable and difficult to identify either by structural imaging or by postmortem analysis, it is likely that those patients with more severe recognition deficits have damage extending into other brain regions important for item recognition, such as the perirhinal cortex. There is even evidence that extrahippocampal damage is exacerbated by abnormal processes in the hippocampus that are triggered by ischemia. Extrahippocampal damage is probably best detected by measuring whether blood flow is abnormal in nonhippocampal brain regions so that it can be determined whether such abnormalities are more striking in patients with severe item recognition deficits. This kind of abnormality, invisible to structural magnetic resonance imaging, has been identified using positron emission tomography in patients who suffered hypoxia following heart attacks. These patients showed reduced blood flow in regions that appeared normal in structural magnetic resonance imaging (MRI). It is possible, however, that less used and more sophisticated structural MRI procedures would identify abnormalities in these regions.

Selective damage to other parts of the Papez circuit has also been found to cause little or no disruption of item recognition. Thus, relatively selective free recall deficits have been reported after fornix lesions caused by colloid cyst surgery and also following relatively selective mammillary body lesions. Lesions that affect only the anterior thalamic nucleus within the thalamus are very rare, but one patient has been described whose thalamic damage was confined to the left anterior thalamic nucleus, although she also had damage to the head of the left caudate nucleus and the left fornix. This patient had completely normal recognition for items and a free recall deficit for verbal materials.

Some animal research supports the existence of another memory system that comprises the perirhinal cortex, the dorsomedial nucleus of the thalamus to which it is reciprocally connected, and possibly the orbitofrontal cortex to which the thalamic nucleus links. The interconnections of this system with the Papez circuit system are illustratred in Fig. 1. As indicated previously, in animals, selective perirhinal cortex lesions disrupt item recognition but not some kinds of spatial memory. They also disrupt associations between similar kinds of items. Little relevant work has been done in humans mainly because selective perirhinal cortex lesions are probably nonexistent and selective dorsomedial thalamic lesions are extremely rare.

However, total bilateral destruction of the perirhinal cortex (and other regions) in humans has been reported to show more severe impairments of recognition for complex patterns at delays of a few seconds compared to those of global amnesics without perirhinal cortex lesions. Interestingly, these patients performed normally at delays of 0 and 2 sec as well as when making judgments regarding when the stimulus was still present. This could mean that perirhinal cortex lesions selectively disrupt long-term memory because they leave short-term memory and visual perception intact. This interpretation is not undisputed because it has been argued, on the basis of animal research, that perirhinal cortex lesions disrupt the ability to represent complex conjunctions of stimulus features in perception, and that memory deficits may be secondary to these high-level perceptual deficits. If this argument is correct, then amnesia that is caused by perirhinal cortex damage does not constitute a pure memory deficit. In principle, the argument can also be extended to hippocampal lesions that might be causing disruption of the ability to represent other kinds of complex conjunctions (such as those between objects and locations or between faces and voices). Such high-level perceptual deficits would secondarily cause memory impairment, and they

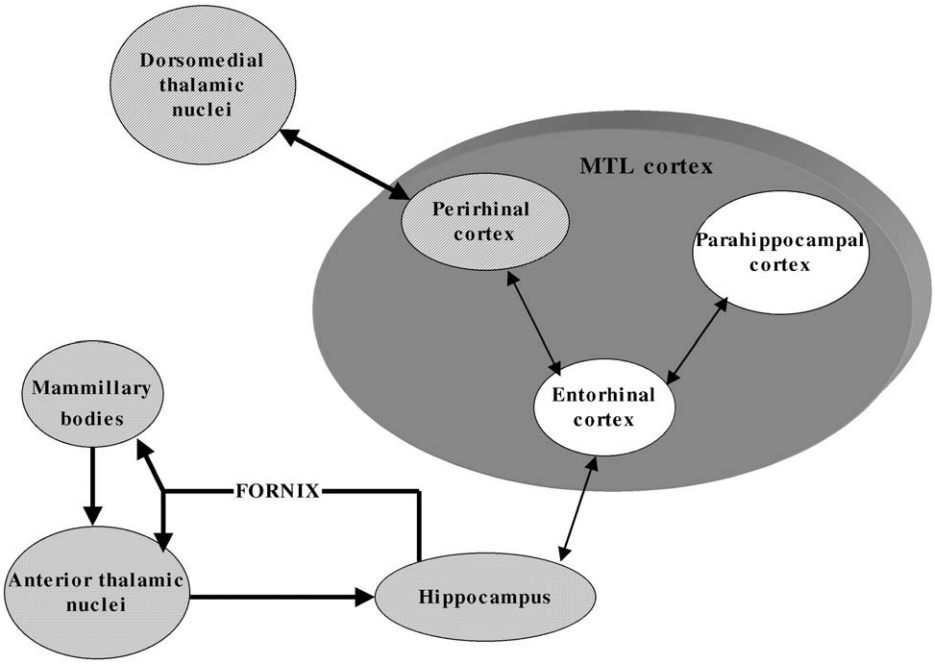


Figure 1 The connections between the structures that may constitute the two memory systems as well as interconnections between the systems and between the systems and neocortical regions. Shading with lines indicates structures that may form the perirhinal cortex–dorsomedial thalamic memory system, whereas pale gray shading indicates structures that may form part of the Papez circuit memory system. Arrows indicate that there is evidence that connections exist. Thick solid lines indicate connections that can confidently be related to the memory systems. Thin solid lines indicate connections that are known to link the two memory systems. Both systems connect to prefrontal association cortex, but it is uncertain what role this cortex plays in memory and it is also unclear whether the two memory systems interact in the prefrontal association cortex.

would be very difficult to detect because of the likelihood that affected patients would use compensatory strategies. It cannot therefore be confidently claimed that major forms of amnesia are unrelated to very specific forms of encoding failure.

The very severe long-term memory deficits that have been reported following perirhinal cortex lesions may be dependent on this damage being nearly total. Support for this possibility is provided by a subgroup of epileptic patients with selective damage that includes only part of the perirhinal cortex who show normal performance on standard tests of anterograde amnesia but accelerated forgetting over delays of weeks. By these delays, their memory is badly impaired. Partial perirhinal cortex damage may therefore cause a much less severe and much delayed item recognition deficit than does total damage to this cortex. The mnemonic effects of dorsomedial thalamic lesions may also critically depend on the extent of damage. Thus, patients with marked destruction of this nucleus show impaired item recognition as well as free recall, whereas patients with only a small portion of the nucleus damaged show little evidence of memory deficits.

Papez circuit lesions and perirhinal cortex system lesions both cause retrograde as well as anterograde amnesia. Although severe retrograde amnesia can extend back to memories acquired decades before the causative brain damage, there often appears to be a temporal gradient in which earlier acquired memories are less disrupted. Considerable evidence suggests that Papez circuit lesions cause a less severe, more steeply temporally graded retrograde amnesia than do larger lesions that involve the whole of the MTL. This would be consistent with the hypothesis that the hippocampus (and perhaps other MTL regions) only stores facts and episodes for a limited time until reorganization results in neocortical storage in the sites that presumably represent the stored information. An alternative view that challenges this claim is that there is no reorganization of the location of long-term memories as a result of processes such as rehearsal, at least for some kinds of information (such as spatial). The simplest interpretation of this view is that it should predict no temporal

gradient, although more complex interpretations are also possible that make it more difficult to distinguish between the predictions of the no change and the change views of what happens to long-term memory as it ages. Therefore, additional work is needed to determine whether memories cease to depend on the MTL as they age. Future work must also explore whether selective Papez circuit lesions minimally disrupt premorbid item recognition memory.

Exactly what processes are disrupted in patients with the amnesia syndrome? If the syndrome is heterogeneous, there must be several such processes. It is widely believed that many global amnesia patients process facts and episodes normally because their intelligence is preserved and normal kinds of information are available to them at input when memory load is minimized. As discussed previously, however, it has not been shown conclusively that subtle high-level perceptual processes may not be disrupted following MTL lesions. Evidence from transient global amnesia (TGA) indicates that it is unlikely that retrieval is impaired in global anterograde amnesia. TGA is a form of global amnesia that lasts for only a few hours and is usually caused by a reversible abnormality in the MTL, which has been revealed by showing that blood flow to the MTL (and sometimes other brain regions) is reduced during a TGA attack but usually returns to normal when tested later. Upon recovery, although all premorbid memories apart from those acquired in the few minutes or hours prior to the incident typically return, as does the ability to lay down new memories, no memories for the incident return. Because these memories do not return even when retrieval must be normal, global anterograde amnesia is probably caused by a failure to consolidate facts and episodes into long-term memory in the minutes following input.

If Papez circuit lesions do cause a selective amnesia, they will disrupt the consolidation of different kinds of information from those disrupted by perirhinal cortex lesions. The available evidence suggests that patients with selective hippocampal lesions can be relatively normal not only for recognition memory for single items (e.g., words or faces) but also for associations between the same kind of items (e.g., word–word or face–face associations). However, they are severely impaired at recognizing associations between components that would probably be processed in different cortical regions (e.g., object–location and face–voice). Similar and possibly more severe spatial memory deficits have also been reported following parahippocampal cortex lesions. The evidence is more conflicting about whether patients with selective hippocampal lesions show relatively preserved remote memory for overlearned facts. Patients are unquestionably impaired at the initial acquisition of new facts (such as vocabulary) but it remains to be shown that their impairment does not occur because their factual memory cannot be facilitated by the retrieval of associated contextual information (which involves episodic memory and is disrupted). Also, some patients may learn to compensate for their learning deficit by rehearsing factual information much more frequently than people with normal memory. Probably, hippocampal and Papez circuit lesions disrupt consolidation of both factual and episodic associations, the components of which are represented in different cortical regions.

Although lesions that include the perirhinal cortex drastically impair item recognition, little is known about the effect of selective lesions of this structure in humans. Animal studies indicate that selective damage to this cortex disrupts recognition of items and associations between similar components but not memory for some kinds of spatial information. It is uncertain in humans whether perirhinal cortex lesions disrupt memory for all the kinds of information affected by hippocampal lesions and memory for some other kinds of information as well, or whether there are some kinds of memory (such as some forms of spatial memory) that are only disrupted by hippocampal lesions.

B. The Role of the Prefrontal Cortex in Memory

Prefrontal cortex damage sometimes disrupts long-term memory for post- as well as premorbidly experienced facts and episodes. Commonly, the deficits are of free recall, with item recognition being relatively normal. The ability to remember the temporal order in which items have been presented is also often impaired, as is the ability to remember the source (who said something or whether it was encountered via TV, the radio, or newspaper) of information. Prefrontal cortex lesions have also been found to disrupt various kinds of metamemory, such as being able to predict whether one will be able to recognize information one has failed to recall when it is presented later. However, recognition deficits have also been reported in the presence of free recall that is relatively good apart from the production of a pathological level of false positives.

Such memory impairments are probably secondary to the effect of frontal cortex damage on executive processes and perhaps working memory. If sophisticated encoding processes cannot be properly

orchestrated, then memory will suffer, and the effect is likely (as is often found) to disrupt free recall, which is more dependent than item recognition on the storage of rich interitem associations. Similarly, if the organization of searching and checking operations during retrieval is disrupted, then free recall will probably be affected more than item recognition. Whether this kind of explanation accounts for the disruption of temporal and source memory in frontally damaged patients remains to be proved, but these kinds of contextual memory are likely to require considerable amounts of organization at both encoding and retrieval. Similarly, it remains to be determined whether the abnormal recognition shown by some patients with prefrontal cortex lesions is caused by the disruption of different executive processes from those affected in cases of free recall deficit or contextual memory deficit. Finally, it is believed by some that damage to the ventromedial prefrontal projections of the MTL causes a syndrome very similar, if not identical, to global amnesia.

Neuroimaging evidence indicates that the left and right prefrontal cortices are activated when verbal and difficult to verbalize information respectively are encoded into memory. This suggests that left and right prefrontal cortex lesions may respectively disrupt verbal and nonverbal memory. There is evidence that lateralized lesions do have this kind of material-specific effect, although it has been noted that laterality effects are weaker than with more posterior cortical lesions, and that verbal learning deficits are much more common after bilateral than after left prefrontal cortex lesions. Evidence for other kinds of dissociation between left and right prefrontal cortex lesions is weak. Although there may be a weak relationship between right frontal damage and retrograde amnesia for episodes, there is little evidence that right frontal cortex lesions disrupt retrieval for recently experienced episodes. However, a pathological tendency to make false alarms during recognition testing has been noted in some patients after both left- and right-sided frontal cortex damage. The manipulations that improve performance, however, are variable and include the use of a semantic orienting task at encoding and using foils that were in a different semantic category from all target items.

Long-term memory for episodes and facts involves a large network of neural structures interacting with each other. The network includes several frontal cortex regions interacting with both temporal and parietal cortices as well as subcortical structures. The precise characteristics of the memory impairments caused by lesions to specific parts of the network need further specification so as to test the functional deficit hypotheses more rigorously. Neuroimaging has provided evidence that the roles of the prefrontal and MTL cortices in memory are complementary. Thus, patients with amnesia caused by MTL lesions can show normal frontal activation patterns during encoding despite their impaired memory. If the frontal and MTL contributions to memory are complementary, then one would expect that temporal order memory deficits arise for different reasons in amnesia and following frontal cortex lesions. The impaired process is presumably different in the two cases, involving consolidation in the former and some kind of difficulty in executing effortful encoding and/or retrieval processes in the latter. It has been shown that the frontal deficit mainly occurs following intentional encoding, which indicates that the source of the problem is probably impoverished effortful encoding. If the deficit relates to poor consolidation in amnesia, amnesics should be impaired following both intentional and incidental encoding. This likely possibility remains to be tested.

V. PRIMING DEFICITS

Priming involves memory for specific information that people are typically unaware they are remembering. This kind of unaware memory is indicated by the enhanced fluency with which the remembered information is reactivated when cues that form part of the memory are encoded. Enhanced reactivation fluency of the remembered representation probably depends on storage changes at the synapses within the representing region so that the components of the memory are bound more tightly together. These changes should occur in different neocortical regions, depending not only on whether semantic or perceptual information is being implicitly remembered but also on precisely what kinds of such information are involved. Consistent with this, functional imaging has shown that whereas visual object priming produces reduced activation of visual cortex regions where the visual object information should be represented at encoding, semantic priming produces less activation in the left inferior frontal cortex where semantic information may be partially represented. This implies that perceptual and semantic kinds of priming occur in different cortical regions, although perceptual priming may not always be based in posterior cortex or semantic priming in frontal cortex regions. Lesions in these regions would be expected to disrupt the appropriate kinds of priming.

Amnesics, who do not have damage in these neocortical regions, might be expected to show preserved priming. However, if the priming involves retrieval of the kinds of association that amnesics fail to store normally, then deficits might be expected. The evidence is still incomplete. There is good evidence that amnesics show preserved priming for information that was already in memory prior to study (such as words or famous faces), but they are often impaired at priming for various kinds of information that were novel prior to study. Meta-analysis of available studies that involve priming of both novel and already familiar information reveals that although amnesic patients show completely normal priming for information that was already in memory at study, they are significantly impaired across studies at priming various kinds of novel information. One interpretation of such results is that normal people may effortfully be using explicit memory in those priming tasks at which amnesics are impaired, whereas the patients cannot do this. This interpretation, however, does not provide a principled explanation of why control subjects use or do not use explicit memory with similar priming tasks that differ merely with respect to whether studied information is novel or already familiar.

If perceptual information is processed in the posterior association cortex, then one might expect that lesions in this region will disrupt perceptual priming. This is consistent with evidence that Alzheimer patients typically show preservation of perceptual priming but impairment in more semantic kinds of priming. These patients have relative preservation of primary sensory processing regions in the posterior cortex but marked atrophy in the temporal association cortex, which together with the left frontal region plays a key role in processing semantic information. Thus, Alzheimer patients show preservation of priming for previously novel patterns but are impaired at a more semantic verbal free association priming task. There is disagreement about whether these patients are impaired at stem completion priming, which some argue may relate to this being a partially semantic priming task. A key requirement for preserved stem completion priming that may explain the inconsistent results is that patients have to explicitly process the words phonologically during study if they are to prime normally.

Alzheimer's patients are impaired not only at certain kinds of priming but also at explicit memory for facts and episodes, so they do not show evidence of impaired priming in the face of intact recognition and/or recall. However, selective impairment of perceptual priming has been shown in a patient with a right posterior neocortical lesion. This patient was impaired at certain kinds of visual repetition priming. In contrast, he was relatively normal not only at semantic priming but also at recognition of visually presented words for which visual repetition priming was impaired. A double dissociation was noted between this patient and amnesics who have disrupted recognition memory but normal visual priming of premorbidly familiar materials. A similar patient was found to be unimpaired in the familiarity component as well as in the recollection component of recognition.

Such a dissociation between perceptual priming of words and recognition of words might arise because visual word recognition usually involves familiarity as well as recollection for primarily semantic information rather than the kinds of visual information retrieved in the perceptual priming at which patients with right-sided visual cortex damage were impaired. However, it has been shown that recognition of modality and font of word presentation can also be preserved in the presence of impaired visual priming. If the perceptual information retrieved in the intact recognition and impaired priming tasks was truly matched, and the priming deficit was caused by a storage deficit, this finding would show that the perceptual fluency that underlies perceptual priming does not contribute to the familiarity component of perceptual recognition.

There has been little exploration of whether priming deficits can occur without accompanying, if subtle, processing deficits, so this remains an open issue. The priming deficits shown in the patients just described, however, are associated with some visual processing problems, so it remains possible that priming deficits are always accompanied by corresponding processing deficits. It could even be that a visual processing deficit produces the false appearance of a priming impairment, and that the memory underlying priming is preserved.

VI. DEFICITS IN OTHER KINDS OF PROCEDURAL/IMPLICIT MEMORY

Procedural memory comprises not only priming but also various forms of conditioning, various forms of skill memory, and nonassociative forms of memory such as habituation. The core feature of all these forms of memory is that remembering is not accompanied by a feeling of memory. They are probably much more functionally heterogeneous than semantic and episodic memory, but less is known about them. It is believed that they are mainly dependent on subcortical

mechanisms, but neuroimaging evidence suggests that frontal and motor cortex regions are important in the early stages of skill acquisition.

Amnesics perform normally on nonpriming kinds of procedural memory such as classical conditioning as well as skill learning and memory. For example, they show preserved delay eye blink classical conditioning. Amnesics have also been shown to acquire normally the motor skill of mirror drawing, the perceptual skill of reading mirror reversed words, and the cognitive skill of intuitively grasping the relationship between variables to achieve a target value. Furthermore, amnesic performance has been shown to be preserved in the acquisition of adaptation-level effects with weights and in the acquisition of the ability to perceive depth using random-dot stereograms. This indicates that the forebrain regions critical for episodic and semantic memory play little or no role in these forms of procedural memory.

In contrast to amnesics, patients with Huntington's disease (HD), who have neostriatal damage, have been found to show impaired acquisition and retention of skills such as reading mirror-reversed words and relatively normal performance at verbal recognition tests. Similarly, Parkinson's patients, whose substantia nigra pathology also disrupts neostriatal function, have shown impaired acquisition of cognitive skills despite being normal at some explicit memory tasks.

With respect to motor skills, and unlike amnesics, patients with HD fail to show normal adaptation-level effects with weights. This is particularly interesting because although no correlation between motor dysfunction and motor skill acquisition dysfunction has been found in these patients, there is still uncertainty about what functional deficits underlie the motor skill acquisition deficit in HD. Because the adaptation-level task is much less dependent on overt movement than motor skill acquisition, it is likely that neostriatal damage disrupts the development of motor programs vital for both normal motor skill acquisition and adaptation-level effects. The neostriatum may be particularly involved in developing motor programs for sequences of motor acts because neurodegenerative diseases affecting the neostriatum disrupt serial reaction time performance. Cerebellar atrophy disrupts performance on the serial reaction time task, which suggests that the cerebellum is also involved in developing motor programs, perhaps because it indexes the temporal order of sensorimotor events.

The involvement of the striatum in perceptual and cognitive skills is supported by several reports of deficits in the development of such skills in patients with degenerative damage in the region. However, impairments are not always found because preserved acquisition of mirror-reading skill has been found in Parkinson's disease patients. Whether cerebellar damage impairs perceptual and cognitive skill acquisition is uncertain because although deficits have been reported, there is strong evidence that patients with selective cerebellar degeneration do not show deficits in learning to read mirror-reversed text or solving the tower of Hanoi difficulty. The difficulty is that many studies have included patients with degeneration that extended beyond the cerebellum.

Cerebellar lesions have, however, have been shown convincingly to disrupt the development of delay eye blink classical conditioning while leaving explicit memory intact. However, the same patients show unimpaired autonomic or emotional conditioning. Studies with animals indicate that fear conditioning, which typically involves autonomic as well as behavioral conditioning, is disrupted both by amygdala lesions and by lesions that disrupt sensory input to the amygdala. Therefore, the neural bases of motoric and emotional classical conditioning, appear to be very different, so it needs to be shown what these two forms of conditioning share apart from the name.

Whereas delay conditioning in which the conditioned stimulus is still present when the unconditioned stimulus appears is preserved in amnesics, more complex forms of conditioning may not be. Thus, amnesics are impaired at reversal discrimination conditioning (in which a response is first conditioned to one discriminative stimulus and inhibited to the other, after which the process is reversed) perhaps because normal performance depends partly on the use of aware (explicit) memory, which is impaired in amnesics. There is conflicting evidence about whether amnesics are impaired at trace motor conditioning (in which the conditioned stimulus ends before the unconditioned stimulus appears) although a growing body of evidence suggests that patients are unpaired. This conflict probably arises because measuring conditioned responses in the trace conditioning paradigm is difficult partly because fine grained features of the response's timing may indicate whether it is an automatic conditioned response or a voluntarily produced anticipation of the unconditioned stimulus. This difficulty may also relate to the dispute about whether trace conditioning depends on explicit memory for the contingencies between conditioned and unconditioned stimuli or is automatic and independent of explicit memory, as one would expect if it is a pure form of procedural memory.

VII. CONCLUSION

Differently located brain damage disrupts memory for different kinds of information despite often having little obvious effect on the processing of the poorly remembered information. Processing may not always be unaffected (as perhaps is the case with selective priming deficits) and it is sometimes difficult to show convincingly that it is not (as with skill memory and conditioning); however, insofar as it is unaffected, the view that information is stored in the same neurons that process and represent it is challenged. At least four explanations of this apparent challenge are possible. First, the damaged region may merely modulate storage in the neural system that represents the poorly remembered information. Basal forebrain structures may play this modulatory role. Second, partial damage to the representing neural system may be sufficient to disrupt its storage abilities while having a minimal effect on its representational processing abilities. This may apply to the short-term memory disorders. Third, there may be multiple representational–storage neural systems for the same information. In other words, the same information may be represented and stored in several different neural sites. Fourth, information is not stored exactly where it is represented.

The existence of multiple memory deficits has been used as evidence to support the claim that there are different memory systems for different kinds of information, each with its own neural system that may mediate memory through the use of qualitatively distinct processes. Two comments about such memory system views are warranted. First, if memory is organized as a set of systems, then these should be arranged hierarchically. This is true of the influential taxonomy that discriminates between declarative and nondeclarative memory, where the former involves all aware forms of memory and the latter all nonaware forms. If correct, all aware forms of memory should have more in common with each other than they do with any form of nonaware memory and vice versa. This would be untrue if amnesics have impaired priming for the same novel information for which they show impaired aware memory, as the available evidence suggests, and also if they are impaired at forms of conditioning that can be shown not to depend on aware memory, as some believe to be the case. The issue is very important and currently unresolved. Either memory systems are mainly organized around whether or not they produce aware memory or the kinds of information they store is more fundamental to their organization. Resolution will involve determining the relationship between unaware and aware memory and how aware memory is produced.

Second, memory may be mediated by different brain systems, but they may work in the same way. Anatomical differences should not be regarded as equivalent to qualitative differences in processing. Although such qualitatively different memory processes may operate for memory of different kinds of information, the methods for showing this remain to be properly established. Indeed, one cannot even be sure for most organic memory disorders whether they are caused by encoding, storage, or retrieval failures. Thus, it is currently unknown whether radically different kinds of memory processing are mediated by distinct brain regions.

See Also the Following Articles

MEMORY, EXPLICIT AND IMPLICIT • MEMORY NEUROBIOLOGY • MEMORY, NEUROIMAGING • MEMORY, OVERVIEW • NERVE CELLS AND MEMORY • SEMANTIC MEMORY • SHORT-TERM MEMORY • WORKING MEMORY

Suggested Reading

Daum, I., Schugens, M. M., Ackerman, H., Lutzenberg, W., Dichgans, J., and Birbaumer, N. (1993). Classical conditioning after cerebellar lesions in humans. *Behav. Neurosci.* **105,** 748–756.

Jonides, J., and Smith, E. E. (1997). The architecture of working memory. In *Cognitive Neuroscience* (M. D. Rugg, Ed.), pp. 243–276. Psychology Press, Hove, UK.

Keane, M. M., Gabrieli, J. D. E., Mapstone, H. C., Johnston, K. A., and Corkin, S. (1995). Double dissociation of memory capacities after bilateral occipital-lobe or medial temporal-lobe lesions. *Brain* **118,** 1129–1148.

Mayes, A. R. (1988). *Human Organic Memory Disorders*. Cambridge Univ. Press, Cambridge, UK.

Mayes, A. R., and Downes, J. J. (Eds.) (1997). *Theories of Organic Amnesia*. Psychology Press, Hove, UK.

Patterson, K., and Hodges, J. R. (1995). Disorders of semantic memory. In *Handbook of Memory Disorders*. (A. D. Baddeley, B. A. Wilson, and F. N. Watts, Eds.), p. 167. Wiley, Chichester, UK

Schacter, D. L., and Tulving, E. (Eds.) (1994). *Memory Systems*. MIT Press, Cambridge, MA.

Troster, A. (Ed.) (1998). *Memory in Neurodegenerative Disease*. Cambridge Univ. Press, Cambridge, UK.

Vallar, G., and Papagno, C. (1995). Neuropsychological impairments of short-term memory. In *Handbook of Memory Disorders* (A. D. Baddeley, B. A. Wilson, and F. N. Watts, Eds.), pp. 135–166. Wiley, Chichester, UK.

Vargha-Khadem, F., Gadian, D. G., Watkins, K. E., Van Paesschen, W., and Mishkin, M. (1997). Differential effects of early hippocampal pathology on episodic and semantic memory. *Science* **277,** 376–380.

Memory Neurobiology

RAYMOND P. KESNER
University of Utah

GLOSSARY

event-based memory system This system provides for temporary representations of incoming data concerning the present, with an emphasis on data and events that are usually personal or egocentric and that occur within specific external and internal contexts.

knowledge-based memory system This system provides for more permanent representations of previously stored information in long-term memory and can be thought of as one's general knowledge of the world.

language attribute This attribute involves memory representations of phonological, lexical, morphological, syntactical, and semantic information.

response attribute This attribute involves memory representations based on feedback from motor responses (often based on kinesthetic and vestibular cues) that occur in specific situations as well as memory representations of stimulus–response associations.

reward value (affect) attribute This attribute involves memory representations of reward value, positive or negative emotional experiences, and the associations between stimuli and rewards.

rule-based memory system This system receives information from the event-based and knowledge-based systems and integrates the information by applying rules and strategies for subsequent action.

sensory-perceptual attribute This attribute involves memory representations of a set of sensory stimuli that are organized in the form of cues as part of a specific experience.

spatial (space) attribute This attribute involves memory representations of places or relationships between places.

temporal (time) attribute This attribute involves memory representations of the duration of a stimulus and the succession or temporal order of temporally separated events or stimuli.

Memory neurobiology is defined as the organization of memory systems in terms of the kind of information to be represented, the processes associated with the operation of each system, and the neurobiological substrates including neural structures and mechanisms that subserve each system.

I. INTRODUCTION

Memory neurobiology is defined as the organization of memory systems in terms of the kind of information to be represented, the processes associated with the operation of each system, and the neurobiological substrates including neural structures and mechanisms that subserve each system. Because it is assumed that memory is highly complex and distributed across many neural systems, the overall organization of memory systems can take many forms, with different emphases on the nature of memory representation and different operations associated with processing of mnemonic information.

For example, the most widely accepted view of memory is based on the idea that memory can be divided into a declarative component based on conscious recollection of facts and events and mediated by the hippocampus and interconnected neural regions, such as the entorhinal cortex and parahippocampal and perirhinal cortex, and a nondeclarative

component based on memory without conscious access for skills, habits, priming, simple classical conditioning, and nonassociative learning mediated by a variety of brain regions, including the striatum for skills and habits, the neocortex for priming, the amygdala for simple classical conditioning of emotional responses, the cerebellum for simple classical conditioning of skeletal musculature, and reflex pathways for nonassociative learning. Others have used different terms to reflect the same type of distinction, including a hippocampal-dependent explicit memory vs a nonhippocampal-dependent implicit memory and a hippocampal-dependent declarative memory based on the representation of relationships among stimuli vs a nonhippocampal-dependent procedural memory based on the representation of a single stimulus or configuration of stimuli.

According to these models, the key difference in memory representation across different brain regions is based on conscious access to the information to be processed or stored. Support for this distinction comes from studies with human amnesic patients. For example, Korsakoff patients, with presumably diencephalic damage, and patients receiving electroconvulsive shock treatments, which presumably produce major disruptive effects in the temporal lobe, can acquire and retain (for at least 3 months) a mirror reading skill as easily as normal subjects (nondeclarative memory). However, when asked to remember the words they read in this task, they were severely impaired (declarative memory). Patient H.M. with bilateral medial temporal lobe damage, including hippocampus and amygdala, can learn and remember a set of complicated skills associated with solving the Tower of Hanoi problem, but this patient cannot recall any contextual aspect of the task or the strategies involved in solving the task. Thus, it appears that amnesic patients can acquire skills necessary for correct mirror reading performance but cannot remember the specific facts or events associated with the experiment. Further support for a problem with conscious awareness associated with declarative information is based on a patient with bilateral damage to the hippocampus who was impaired in learning which stimuli were paired with an unconditioned response but learned very readily a conditioned autonomic response to the critical visual and auditory stimuli. In contrast, a different patient with a bilateral lesion of the amygdala was impaired in learning a conditioned autonomic response to visual or auditory stimuli but learned very readily which stimuli were paired with the unconditioned response. Also, amnesic subjects with damage to the hippocampus are not impaired on a variety of tests that measure implicit memory for a number of stimulus patterns using the nondeclarative memory system, but they are impaired in explicitly remembering the same stimulus patterns using the declarative memory system. Finally, amnesic subjects can learn a complex probability classification task, but they perform poorly on multiple-choice tests that attempt to measure the training experience.

Others have suggested that memory can be divided into a short-term, working or episodic memory system defined as memory for the specific, personal and temporal context of a situation and mediated by the hippocampus and a reference or semantic memory defined as memory for rules and procedures (general knowledge) of specific situations mediated by nonhippocampal brain regions, such as the neocortex. According to these models the key difference in memory representation across different brain regions is not necessarily based on conscious access to the information to be processed or stored but rather on differential processing of short-term vs long-term memory representations.

A more comprehensive view of memory organization based on multiple processes and multiple forms of memory representation is based on the neurobiology of a multiple attribute, multiple process, tripartite system model of memory that will be presented later. The tripartite attribute model of memory is organized into event-based, knowledge-based, and rule-based memory systems. Each system is composed of the same set of multiple attributes or forms of memory, characterized by a set of process-oriented operating characteristics and mapped onto multiple neural regions and interconnected neural circuits.

II. TRIPARTITE MEMORY SYSTEM

On a psychological level, the event-based memory system provides for temporary representations of incoming data concerning the present, with an emphasis on data and events that are usually personal or egocentric and that occur within specific external and internal contexts. The emphasis is on the processing of new and current information. During initial learning, great emphasis is placed on the event-based memory system, which will continue to be of importance even after initial learning in situations in which unique or novel trial information needs to be remembered.

The knowledge-based memory system provides for more permanent representations of previously stored

information in long-term memory and can be thought of as one's general knowledge of the world. The knowledge-based memory system would tend to be of greater importance after a task has been learned given that the situation is invariant and familiar. The organization of these attributes within the knowledge-based memory system can take many forms and they are organized as a set of attribute-dependent cognitive maps and their interactions that are unique for each memory.

The rule-based memory system receives information from the event-based and knowledge-based systems and integrates the information by applying rules and strategies for subsequent action. In most situations, however, one would expect a contribution of all three systems with a varying proportion of involvement of one relative to the other.

The three memory systems are composed of the same forms, domains, or attributes of memory. Even though there could be many attributes, the most important attributes include space, time, response, sensory perception, and reward value (affect). In humans, a language attribute is also added. A spatial (space) attribute within this framework involves memory representations of places or relationships between places. It is exemplified by the ability to encode and remember spatial maps and to localize stimuli in external space. Memory representations of the spatial attribute can be further subdivided into specific spatial features, including allocentric spatial distance, egocentric spatial distance, allocentric direction, egocentric head direction, and spatial location. A temporal (time) attribute within this framework involves memory representations of the duration of a stimulus and the succession or temporal order of temporally separated events or stimuli. A response attribute within this framework involves memory representations based on feedback from motor responses (often based on kinesthetic and vestibular cues) that occur in specific situations as well as memory representations of stimulus–response associations. A reward value (affect) attribute within this framework involves memory representations of reward value, positive or negative emotional experiences, and the associations between stimuli and rewards. A sensory-perceptual attribute within this framework involves memory representations of a set of sensory stimuli that are organized in the form of cues as part of a specific experience. Each sensory modality (olfaction, auditory, vision, somatosensory, and taste) has its own memory representations and can be considered to be part of the sensory-perceptual attribute component of memory. A language attribute within this framework involves memory representations of phonological, lexical, morphological, syntactical, and semantic information. The attributes within each memory system can be organized in many different ways and are likely to interact extensively with each other even though it can be demonstrated that in many cases these attributes operate independent of each other.

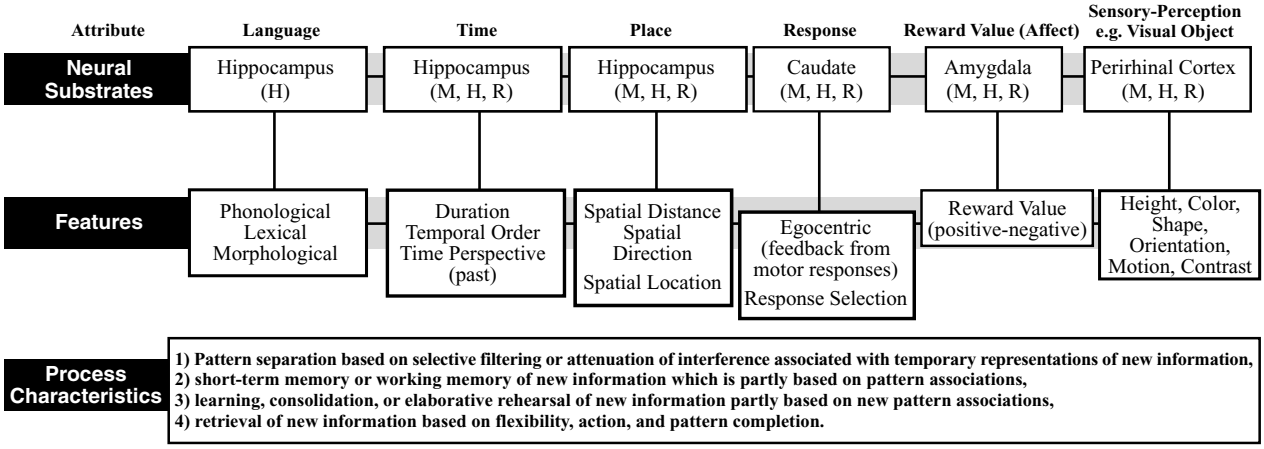
Within each system attribute information is processed in different ways based on different operational characteristics. For the event-based memory system, specific processes involve (i) selective filtering or attenuation of interference of temporary memory representations of new information, labeled pattern separation; (ii) short-term memory or working memory of new information that is partly based on pattern associations; (iii) consolidation or elaborative rehearsal of new information; and (iv) retrieval of new information based on flexibility, action, and pattern completion.

For the knowledge-based memory system, specific processes include (i) selective attention and selective filtering associated with permanent memory representations of familiar information, (ii) perceptual memory and long-term memory storage based on pattern associations, and (iii) retrieval of familiar information based on flexibility and action. For the rule-based memory system, the major process includes the selection of strategies and rules for maintaining or manipulating information for subsequent action.

III. EVENT-BASED MEMORY SYSTEM

On a neurobiological level each attribute maps onto a set of neural regions and their interconnected neural circuits. For example, within the event-based memory system it has been demonstrated that in animals and humans the hippocampus supports memory for spatial and temporal attribute information, the caudate mediates memory for response attribute information, the amygdala subserves memory for reward value (affect) attribute information, and the perirhinal and extrastriate visual cortex support memory for visual object attribute information as an example of a sensory-perceptual attribute (Fig. 1).

Evidence supportive of the previously mentioned mapping of attributes onto specific brain regions is based in part on the use of paradigms that measure the short-term or working memory process based on performance within matching or nonmatching-to-



Key: M=Monkeys, **H**=Humans, **R**=Rats

Figure 1 Representation of the neural substrates, features, and process characteristics associated with the event-based memory system for the language, time, place, response, reward value (affect), and sensory-perception attributes.

sample, delayed conditional discrimination or continuous recognition memory of single-item or lists of items tasks; the consolidation and retrieval process, based on learning and retention of a variety of behavioral tasks; and the pattern separation process based on performance within category memory or discrimination tasks.

A. Spatial (Place) Attribute

With respect to short-term or working memory, the data indicate that animals and humans with damage to the hippocampus are severely impaired in working memory for spatial information, including spatial features such as allocentric spatial distance, egocentric spatial distance, head direction, and spatial location. In contrast, rats, monkeys, or humans with damage to the hippocampus are not impaired in working memory for response, reward value (affect), or visual object attribute information. Similar patterns of results have been reported for the involvement of the hippocampus in mediating spatial but not response, reward value, or visual object information during new learning (consolidation).

In the context of the pattern separation process, it has been shown that the hippocampus responds to all sensory modalities, suggesting that a possible role for the hippocampus is to provide the opportunity for using sensory markers to demarcate a spatial location so that the hippocampus can more efficiently represent spatial information. It is thus possible that one of the main process functions of the hippocampus is to encode and separate spatial events from each other. This would ensure that new highly processed sensory information is organized within the hippocampus and enhances the possibility of remembering and temporarily storing one place as separate from another place. It is assumed that this is accomplished via pattern separation of event information so that spatial events can be separated from each other and spatial interference is reduced. This process is akin to the idea that the hippocampus is involved in orthogonalization of sensory input information, in representational differentiation, and indirectly in the utilization of relationships.

Since pattern separation paradigms are new and have been developed only recently, a more detailed description of the paradigms is presented. In this task, rats are required to remember a spatial location dependent on the distance between the study phase object and an identical object used as a foil. Specifically, during the study phase an object that covers a baited food well is randomly positioned in 1 of 15 possible spatial locations on a cheese board. Rats exit a start box and displace the object in order to receive a food reward and are then returned to the start box. On the ensuing test phase rats are allowed to choose between two objects that are identical to the study phase object. One object is baited and positioned in the previous study phase location (correct choice), and the

other (foil) is unbaited and placed in a different location (incorrect choice). Five distances (min = 15 cm; max = 105 cm) are randomly used to separate the foil from the correct object. Following the establishment of a criterion of 75% correct averaged across all separation distances, rats are given either large (dorsal and ventral) hippocampal or cortical control lesions dorsal to the dorsal hippocampus. Following recovery from surgery the rats are retested. The results indicate that whereas control rats match their presurgery performance for all spatial distances, hippocampal lesioned rats display impairments for short (15–37.5 cm) and medium (60 cm) spatial separations but perform as well as controls when the spatial separation is long (82.5–105 cm). The fact that the hippocampal lesioned group is able to perform the task well at large separations indicates that the deficits observed at the shorter separations are not the result of an inability to remember the rule. The results suggest that the hippocampus may serve to separate incoming spatial information into patterns or categories by temporarily storing one place as separate from another place. It can be shown that the ability to remember the long distances is not based on an egocentric response strategy because if the study phase is presented on one side of the cheese board and the test originates on the opposite side, the hippocampal lesioned rats still discriminate the long distances without difficulty. Furthermore, the hippocampal lesioned group has no difficulty discriminating between two short distances. It is clear that in this task it is necessary to separate one spatial location from another spatial location. Hippocampal lesioned rats cannot separate these spatial locations very well, so they can perform the task only when the spatial locations are far apart. Subsequent experiments have shown that lesions of the dorsal dentate gyrus result in the same spatial pattern separation problem, but lesions of the dorsal CA1 region do not produce an impairment, suggesting that pattern separation mechanisms might reside in the dentate gyrus. Similar deficits have been observed for new geographical information in patients with hippocampal damage due to an hypoxic episode.

Does spatial pattern separation play a role in the acquisition (consolidation) of a variety of hippocampal-dependent tasks? One example will suffice. Because rats are started in different locations in the standard water maze task, there is a great potential for interference among similar and overlapping spatial patterns. Thus, the observation that hippocampal lesioned rats are impaired in learning and subsequent consolidation of important spatial information in this task could be due to difficulty in separating spatial patterns, resulting in enhanced spatial interference. Evidence in favor of this idea comes from the observation that when fimbria–fornix lesioned rats are trained on the water maze task from only a single starting position (less spatial interference), there are very few learning deficits, whereas training from many different starting points results in learning difficulties. In a similar study it was shown that total hippocampal lesioned rats learned or consolidated rather readily that only one spatial location was correct on an eight-arm maze.

Response, reward value, and visual object pattern separation are not processed by the hippocampus but involve the caudate for response pattern separation, amygdala for reward value pattern separation, and perirhinal cortex for visual object pattern separation. With respect to associations, the hippocampus does not mediate all arbitrary associations, not even all stimulus–stimulus associations, but only associations that involve spatial or temporal information, such as object–place and odor–place associations or trace classical conditioning, but not, for example, odor–odor or odor–object associations or delayed classical conditioning.

B. Temporal (Time) Attribute

With respect to short-term or working memory, the data indicate that animals and humans with damage to the hippocampus are severely impaired in working memory for temporal information, including duration and temporal order. It has been suggested that trace conditioning requires memory for the duration of the conditioned stimulus. Thus, it is of importance to note that rabbits with hippocampal lesions and humans with hypoxia resulting in bilateral hippocampal damage are impaired in acquisition (consolidation) of trace but not delayed eye-blink conditioning.

Based on ample evidence that almost all sensory information is processed by hippocampal neurons, perhaps to provide for sensory markers for time, and that the hippocampus mediates temporal information, it is likely that one of the main process functions of the hippocampus is to encode the temporal order of events. This would ensure that newly highly processed sensory information is organized within the hippocampus and enhances the possibility of remembering and temporarily storing one event as separate from another event in time. On this basis, it has been shown

that the hippocampus is involved in temporal pattern separation for spatial, visual object, odor, or language information.

C. Language Attribute

With respect to language attribute information, it can be shown that with the use of the previously mentioned paradigms to measure short-term memory there are severe impairments for lists of words for humans with left hippocampal or bilateral hippocampal damage, suggesting that the hippocampus plays an important role in short-term memory representation of word information as an important feature of language attribute information. There is much evidence supporting the idea of lateralization of hippocampal function in humans, with the right hippocampus representing spatial information and the left hippocampus representing linguistic information. For example, patients who had left or right temporal lobectomies that included the hippocampus were tested on a task of recall for a visual location. In this task subjects made a mark on an 8-in line in order to reproduce as close as possible the exact position of the previously shown circle. Subjects with right temporal lobe lesions were impaired on this task, whereas subjects with left temporal lobe lesions were not significantly different from control subjects. In contrast, recall of a list of words resulted in an impairment for subjects with left, but not right, temporal lobe resections. Additional support for the idea that the right hippocampus mediates memory for temporal order for novel spatial information and the left hippocampus mediates temporal order for novel linguistic information comes from the demonstration that subjects with right temporal lobe lesions are impaired relative to controls for temporal order for novel spatial location information but not for the temporal order of novel linguistic information. In contrast, subjects with left temporal lobe lesions are impaired relative to control subjects for the temporal order of novel linguistic information but not the temporal order of novel spatial information. Even though hypoxic subjects or left temporal resected patients are impaired in remembering the order of presentation of words in nonmeaningful sentences requiring the processing of new event-based linguistic information, they are not impaired in remembering the order of presentation of words in syntactically or syntactically and semantically meaningful sentences requiring the processing of knowledge-based linguistic information.

D. Response Attribute

With respect to short-term or working memory, data indicate that animals and humans with damage to the caudate are severely impaired in working memory for response information, including memory representations based on feedback from motor responses (often based on kinesthetic and vestibular cues) that occur in specific situations as well as memory representations of stimulus–response associations. For rats with caudate lesions there are profound short-term memory deficits for a right or left turn response, for an egocentric distance response, for head direction, and for response-based sequential learning. For humans with caudate damage due to Huntington's disease, there are short-term memory deficits for reproducing a hand movement, remembering a list of hand motor movement responses, or learning the sequence of motor movements. Furthermore, rats, monkeys, and humans with caudate lesions have deficits in tasks such as delayed response, delayed alternation, and delayed matching to position. One salient feature of these tasks is the maintenance of spatial orientation to the baited food relative to the position of the subject's body often based on proprioceptive and vestibular feedback. These data suggest that the caudate plays an important role in short-term memory representation for the feedback from a motor response feature of response attribute information. The memory impairments following caudate lesions are specific to the response attribute because these same lesions in rats do not impair short-term memory performance for spatial location or visual object attribute information. Similar patterns of results have been reported for the involvement of the caudate in mediating response, but not spatial or reward value, information during new learning (consolidation) and response-based pattern separation. With respect to associations, the caudate mediates associations that involve the response attribute, thereby supporting primarily stimulus–response-type associations.

E. Reward Value (Affect) Attribute

With respect to short-term or working memory, data indicate that animals and humans with damage to the

amygdala are severely impaired in working memory for affect information including reward value (affect) associated with magnitude of reinforcement or for a liking response based on the mere exposure of a novel stimulus. The memory impairments following amygdala lesions are specific to the affect attribute because these same lesions in rats do not impair short-term memory performance for spatial location, visual object, or response attribute information. Similar results have been reported for the involvement of the amygdala in mediating affect, but not spatial, response or visual object information during new learning (consolidation) and to some extent reward-based pattern separation. With respect to associations, the amygdala mediates associations that involve the reward attribute, thereby supporting primarily stimulus–reward-type associations.

F. Sensory-Perceptual Attribute

With respect to sensory-perceptual attribute information, I concentrate on visual object information as an exemplar of memory representation of this attribute. In the context of short-term or working memory, data indicate that animals and humans with damage to the extrastriate or perirhinal cortex are severely impaired in working memory for visual object information. The memory impairments following perirhinal cortex lesions are specific to the visual object attribute because these same lesions in rats do not impair short-term memory performance for spatial location attribute information. Similar patterns of results have been reported for the involvement of the perirhinal cortex in mediating visual object information during new learning (consolidation) and visual object-based pattern separation. With respect to associations, the perirhinal cortex mediates associations that involve visual object information, thereby supporting primarily visual–visual-type associations.

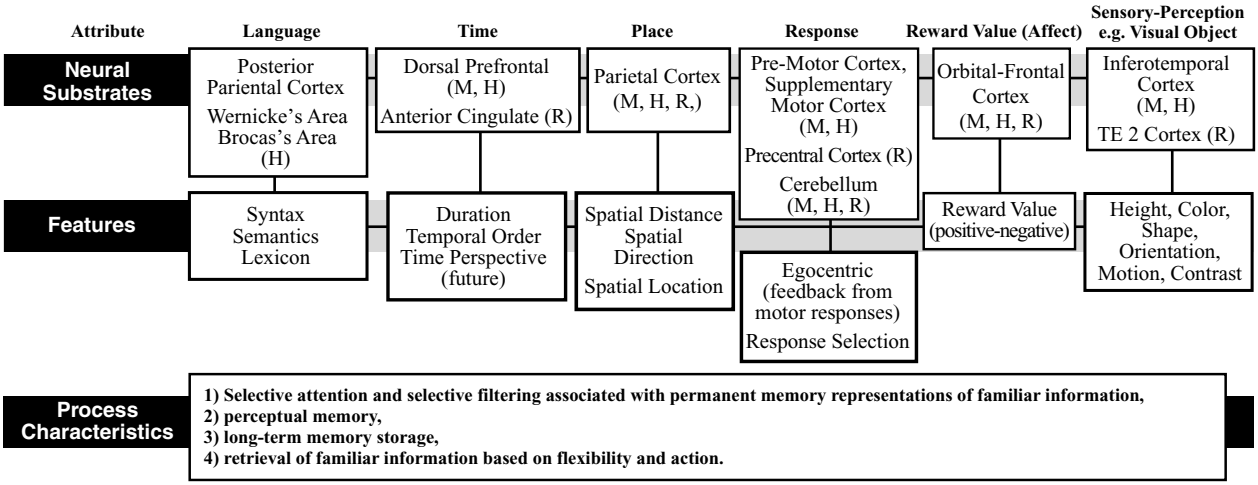
Thus, it appears that the neural systems that support different attributes within the event-based memory system can be dissociated from each other, suggesting that they can operate independent of each other. However, it is clear that many interactions can occur between the proposed brain regions (e.g., between hippocampus and amygdala, between hippocampus and perirhinal cortex, or between hippocampus and caudate nucleus). Furthermore, there are interactions that are based on convergence onto other brain regions (e.g., hippocampus and amygdala projections to the same neurons in the nucleus accumbens). Each neural system subserves a different attribute but engages the same set of processes and is therefore part of the same event-based memory system.

The event-based memory system is more akin to episodic memory; however, it also maps to some extent onto declarative memory, but without the emphasis on the need to use conscious processing and the need to require only the operation of the medial temporal cortex. Instead, the attribute model suggests that there are additional brain circuits that mediate multiple attributes or forms of memory and specifies to a much greater extent the operation of specific multiple processes. Finally, this event-based memory system incorporates rather well White's three- system model that includes the hippocampus for stimulus–stimulus associations, the caudate for stimulus–response associations, and the amygdala for stimulus–reward associations.

IV. KNOWLEDGE-BASED MEMORY SYSTEM

Within the knowledge-based memory system, it has been demonstrated that in animals and humans (i) the posterior parietal cortex supports memory for spatial attributes; (ii) the dorsal and dorsolateral prefrontal cortex and/or anterior cingulate support memory for temporal attributes; (iii) the premotor, supplementary motor, and cerebellum in monkeys and humans and precentral cortex and cerebellum in rats support memory for response attributes; (iv) the orbital prefrontal cortex supports memory for reward value (affect) attributes; and (v) the inferotemporal cortex in monkeys and humans and TE2 cortex in rats subserve memory for sensory-perceptual attributes (e.g., visual objects) (Fig. 2).

Evidence supportive of the previously mentioned mapping of attributes onto specific brain regions is based in part on the use of paradigms that measure the perceptual memory process based on performance within a repetition priming or a discrimination performance paradigm, the attention process by measuring performance within selective attention and stimulus-binding tasks, and the long-term storage and retrieval process by measuring learning and retention in a variety of behavioral tasks. Due to of space constraints, I discuss only the role of the parietal cortex in processing spatial information and that of the inferotemporal cortex in processing visual object information.



Key: M=Monkeys, **H**=Humans, **R**=Rats

Figure 2 Representation of the neural substrates, features, and process characteristics associated with the knowledge-based memory system for the language, time, place, response, reward value (affect), and sensory-perception attributes.

A. Spatial (Place) Attribute

Monkeys with lesions of the parietal cortex show deficits in place reversal, landmark reversal, distance discrimination, bent-wire route finding, pattern string finding, and maze-learning tasks. Similarly, rats with parietal cortex lesions cannot perform well in mazes. Furthermore, rats with parietal cortex lesions display deficits in both the acquisition and retention of spatial navigation tasks that are presumed to measure the operation of a spatial cognitive map within a complex environment. They also display deficits in the acquisition and retention of spatial recognition memory for a list of five spatial locations. In a complex discrimination task in which rats have to detect the change in location of an object in a scene, rats with parietal cortex lesions are profoundly impaired, but on less complex tasks involving the discrimination or short-term memory for single spatial features including spatial location and allocentric and egocentric spatial distance, there are no impairments. Similarly, there are no impairments in discriminating between visual objects in terms of either new learning or performance of a previously learned visual discrimination. When the task is more complex, involving the association of objects and places (components of a spatial cognitive map), parietal cortex plays an important role. Support for this comes from the finding that rats with parietal lesions are impaired in the acquisition and retention of a spatial location plus object discrimination (paired associate task) but show no deficits for only spatial or object discriminations.

Humans with parietal cortex lesions have difficulty in drawing maps or diagrams of familiar spatial locations. They also have problems in using information to guide them in novel or familiar routes, to discriminate near from far objects, and to solve complex mazes. There is also spatial neglect and deficits in spatial attention. There is a general loss of "topographic sense," which may involve loss of long-term geographical knowledge as well as an inability to form cognitive maps of new environments. Using positron emission tomography (PET) scan and functional magnetic resonance imaging (fMRI) data, it can be shown that complex spatial information results in activation of the parietal cortex. Thus, memory for complex spatial information appears to be processed by the parietal cortex.

The parietal cortex is probably not the only neural region that mediates long-term memory for spatial information. For example, topographical amnesia has also been reported for patients with parahippocampal lesions and spatial navigation deficits have been found following retrosplenial and entorhinal cortex lesions. Thus, other neural regions (e.g., parahippocampal cortex, entorhinal cortex, and retrosplenial cortex) may also contribute to the long-term representation of a spatial cognitive map.

B. Sensory-Perceptual Attribute

With respect to sensory-perceptual attribute information, it can be shown that lesions of the inferotemporal cortex in monkeys and humans and temporal cortex (TE2) in rats result in visual object discrimination problems, suggesting that the inferotemporal or TE2 may play an important role in mediating long-term representations of visual object information. Additional support comes from PET scan and fMRI data in humans, in whom it can be shown that visual object information results in activation of inferotemporal cortex. In a different study, it was shown that neurons within the inferotemporal cortex responded more readily after training to a complex visual stimulus that had been paired with another complex visual stimulus across a delay, suggesting the formation of long-term representations of object–object pairs within the inferotemporal cortex. Finally, it has been shown in rats that based on a repetition priming paradigm, perceptual memory for spatial location information is mediated by the parietal cortex but not the TE2 cortex, whereas perceptual memory for visual object information is mediated by TE2 cortex but not the parietal cortex, suggesting support for independent mediation of spatial location and visual object attribute information within the knowledge-based memory system.

V. INDEPENDENCE OF THE EVENT-BASED AND KNOWLEDGE-BASED MEMORY SYSTEMS

Even though the two systems are supported by neural substrates and different operating characteristics, suggesting that the systems can operate independent of each other, there are also important interactions between the two systems, especially during the consolidation of new information and retrieval of previously stored information. It is thus likely to be very difficult to separate the contribution of each system in new learning tasks since each system supports one component of the consolidation process. However, there are a few examples based on tasks in which the major consolidation processes have already taken place. Olton and Papas ran animals in a 17-arm maze with food available in 8 arms and no food available in 9 arms. In order to solve this maze, an animal should not enter unbaited arms, activating knowledge-based memory, but should enter baited arms only once utilizing event-based memory. After learning the task to criterion performance, animals were given fimbria–fornix lesions. The lesioned animals had a deficit only for the event-based memory memory component of the task. In a different study, it was shown that in an 8-arm maze parietal cortex lesions placed in rats after training on 4 unbaited and 4 baited arms resulted in a deficit in the knowledge-based but not event-based memory. If one assumes that the presentation of unbaited arms reflects the operation of the knowledge-based memory system and that the presentation of baited arms reflects the operation of the event-based memory system, then it appears that lesions of the hippocampus disrupt only the event-based memory system, whereas lesions of the parietal cortex only disrupt the knowledge-based memory system. Similarly, in humans there is evidence that patients with hippocampal damage do not have difficulty remembering knowledge-based information, whereas they have difficulty remembering event-based information. For example, hypoxic subjects or left temporal resected patients are impaired in remembering the order of presentation of words in nonmeaningful sentences requiring the processing of new event-based linguistic information, but they are not impaired in remembering the order of presentation of words in syntactically or syntactically and semantically meaningful sentences requiring the processing of knowledge-based linguistic information.

In a different series of studies a double dissociation between perceptual memory (a measure reflective of the operation of the knowledge-based memory system) and short-term memory (a measure reflective of the operation of the event-based memory system) was reported in human subjects. It was shown that patients with a right occipital cortical lesion displayed impaired performance for perceptual memory tests of visual priming for words but intact performance on short-term tests of recognition and cued recall of words. In contrast, the reverse pattern was present for amnesic subjects with hippocampal damage. Furthermore, for patients with parietal lesions resulting in spatial neglect, there was a deficit in spatial repetition priming (perceptual memory) without a loss in short-term or working memory for spatial information. In a different study, it was shown that, like humans with parietal cortex lesions, rats with such lesions are impaired in a spatial repetition priming (perceptual memory) experiment but perform without difficulty in a short-term or working spatial memory experiment, suggesting that the parietal cortex plays a role in spatial perceptual memory within the knowledge-based memory system but does not play a role in spatial memory within the event-based memory system.

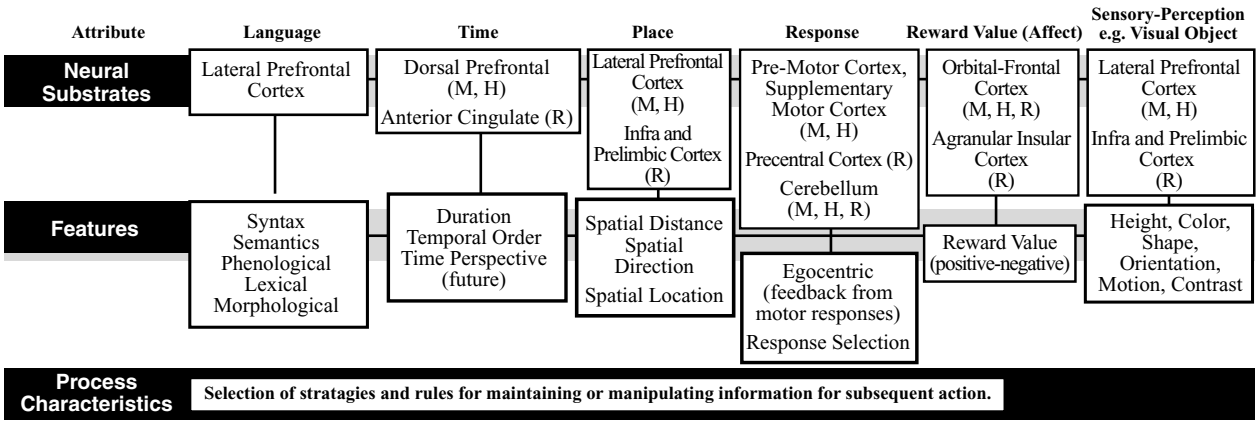
In a different set of studies it was shown that patients with temporal lobe lesions that did not involve hippocampus or perirhinal cortex had difficulty in naming familiar objects (semantic dementia) but had no difficulty in recognizing the same familiar objects, whereas amnesic subjects with hippocampal damage had no difficulty in naming objects, but were impaired in recognizing those objects. If one assumes that naming is a process that is supported by long-term memory and thus a characteristic of the knowledge-based memory system, then these data provide further support for a dissociation between the knowledge-based and event-based memory systems.

The knowledge-based memory system is akin to semantic and reference memory, but it also maps to some extent onto nondeclarative memory but without the emphasis on the need to use unconscious processing and the need to require only the operation of the medial temporal cortex. The nondeclarative memory system has a large memory representation, and compared to the tripartite attribute model it includes the knowledge-based and rule-based memory systems as well as some components of the event-based memory system. It is therefore very difficult to characterize the operations that are necessary for efficient functioning of the nondeclarative system. Instead, the tripartite attribute model suggests that for the knowledge-based memory system there are specific brain circuits that support a set of processes that differ from the brain circuits that support a different set of processes within the event-based memory system. Furthermore, it appears that the knowledge-based memory system and the event-based memory system can operate independent of each other and that information can be processed by distinct neural regions, even though there are also important interactions between the two systems, especially during the consolidation of new information and retrieval of previously stored information.

VI. RULE-BASED MEMORY SYSTEM

Within the rule-based memory system it can be shown that different subdivisions of the prefrontal cortex support different attributes. For example, the dorsolateral and ventrolateral prefrontal cortex in monkeys and humans and the infralimbic and prelimbic cortex in rats support spatial, visual object, and language attributes; the premotor and supplementary motor cortex in monkeys and humans and precentral cortex in rats support response attributes; the dorsal, dorsolateral, and mid-dorsolateral prefrontal cortex in monkeys and humans and anterior cingulate in rats mediate primarily temporal attributes; and the orbital prefrontal cortex in monkeys and humans and agranular insular cortex in rats support affect attributes. (Fig. 3).

Evidence supportive of the previously mentioned mapping of attributes onto specific brain regions is based in part on the use of paradigms that measure the use of rules within short-term or working memory based on performance within matching- or nonmatching-to-sample, delayed conditional discrimination or continuous recognition memory of single-item or lists



Key: **M**=Monkeys, **H**=Humans, **R**=Rats

Figure 3 Representation of the neural substrates, features, and process characteristics associated with the rule-based memory system for the language, time, place, response, reward value (affect), and sensory-perception attributes.

of items tasks, temporal ordering of information, and sequential learning. It is also based on paradigms that measure the use of rules in cross-modal switching, reversal learning, paired-associate, and problem-solving tasks. Here, I concentrate only on the data reported in the context of short-term memory or working memory.

A. Spatial (Place) and Sensory-Perceptual Attributes

The dorsolateral and ventrolateral prefrontal cortex in monkeys and humans and infralimbic and prelimbic cortex are involved whenever there are rules associated with working memory for spatial and visual object attribute information. Evidence for this idea comes from the finding that rats with lesions of the infralimbic and prelimbic cortex disrupt working memory for spatial information and working memory for object information. In monkeys lesions of the dorsolateral and ventrolateral regions disrupt performance on delayed response, delayed alternation, delayed occulomotor, spatial search, and visual object recognition tasks. Furthermore, in monkeys, for working memory there are delay-specific cells in the dorsolateral and ventrolateral prefrontal cortex in spatial tasks, such as delayed response, delayed alternation, and delayed occulomotor tasks, and in visual object delay tasks. In humans, based on a meta-analysis of multiple studies using neuroimaging techniques, both the dorsolateral and the ventrolateral cortex are activated during object and place working memory tasks. Thus, the data indicate that the infralimbic and prelimbic cortex in rats and dorsolateral and ventrolateral cortex in monkeys and humans play a very important role in processing visual object and spatial attribute information, suggesting a possible contribution of this region in supporting the use of working memory rules associated with object and spatial attributes. The working memory impairments following infra- and prelimbic lesions are specific to the object and place attributes because these same lesions in rats do not impair working memory performance for response or affect attribute information.

B. Temporal (Time) Attribute

The dorsal prefrontal cortex, dorsolateral prefrontal cortex, and mid-dorsolateral prefrontal cortex in monkeys and humans and anterior cingulate cortex in rats are involved whenever there are rules associated with working memory for temporal attribute information. Evidence for this idea comes from the finding that in rats, monkeys, and humans, following lesions of this region, there are deficits in working memory for temporal order information and memory for frequency information. Furthermore, using fMRI, it was shown that the dorsolateral prefrontal cortex is activated in a memory for a temporal order task. All the previously mentioned results suggest that this region plays an important role processing temporal attribute information. The working memory impairments following anterior cingulate cortex lesions in rats are specific to the temporal attribute because these lesions do not impair working memory performance for spatial, visual object, or affect attribute information.

C. Response Attribute

The premotor and supplementary motor cortex in monkeys and humans and precentral motor cortex in the rat are involved whenever there are rules associated with the processing of response attribute information. Support for this idea comes from the observation that lesions of the precentral cortex in the rat disrupt rules associated with processing of response information such as working memory for a motor (right–left turn) response. Also, in humans prefrontal cortex lesions result in a deficit for memory for a motor movement and memory for a list of motor movements, supporting a role for this region in processing response attribute information.

D. Reward Value (Affect) Attribute

The orbitofrontal cortex in monkeys and humans and agranular insular prefrontal cortex in the rat are involved whenever there are rules associated with working memory for affect attribute information, especially based on odors and tastes. Evidence for this idea comes from the findings that in both animals and humans, following lesions of this region there are deficits in working memory for taste or odor (affect attribute) information. The working memory impairments following the agranular insular prefrontal cortex lesions are specific to the affect attribute because these same lesions in rats do not impair

working memory performance for spatial attribute information.

E. Language Attribute

The lateral prefrontal cortex in humans is involved whenever there are rules associated with working memory for language attribute information. This is based on evidence demonstrating activation of the lateral prefrontal cortex in verbal learning tasks and impairments in working memory for verbal information in patients with lateral prefrontal cortex lesions.

Various subregions within the prefrontal cortex also play a role in the use of rules associated with other tasks besides working memory, such as tasks that measure the use of rules in cross-modal switching, reversal learning, paired associate, and problem solving.

It is assumed that there are important interactions between the rule-based and event-based memory systems across specific attributes. In support of this assumption, it was shown based on a disconnection experiment that there are interactions between prefrontal cortex and perirhinal cortex but not between perirhinal cortex and amygdala or hippocampus in supporting short-term memory for visual object attribute information. It should be noted that the role for the prefrontal cortex is not specified in the declarative–nondeclarative model.

VII. INTEGRATION

The overall tripartite attribute model of memory is shown in Fig. 4. Different forms of memory and its neurobiological underpinnings are represented in terms of the nature, structure, or content of information representation as a set of different attributes, including language, time, place, response, reward value (affect), and visual object as an example of sensory perception. For each attribute, information is processed in the event-based memory system through operations that involve pattern separation or orthogonalization of specific attribute information, short-term memory processing, encoding of specific pattern associations into long-term memory, and retrieval of stored information via flexibility and pattern completion. In addition, for each attribute, information is processed in the knowledge-based system through operations of long-term storage, selective attention, perceptual memory, and retrieval of pattern associations. Finally, for each attribute, information is processed in the rule-based memory system through the integration of information from the event-based and knowledge-based memory systems for the selection of strategies and rules for maintaining or

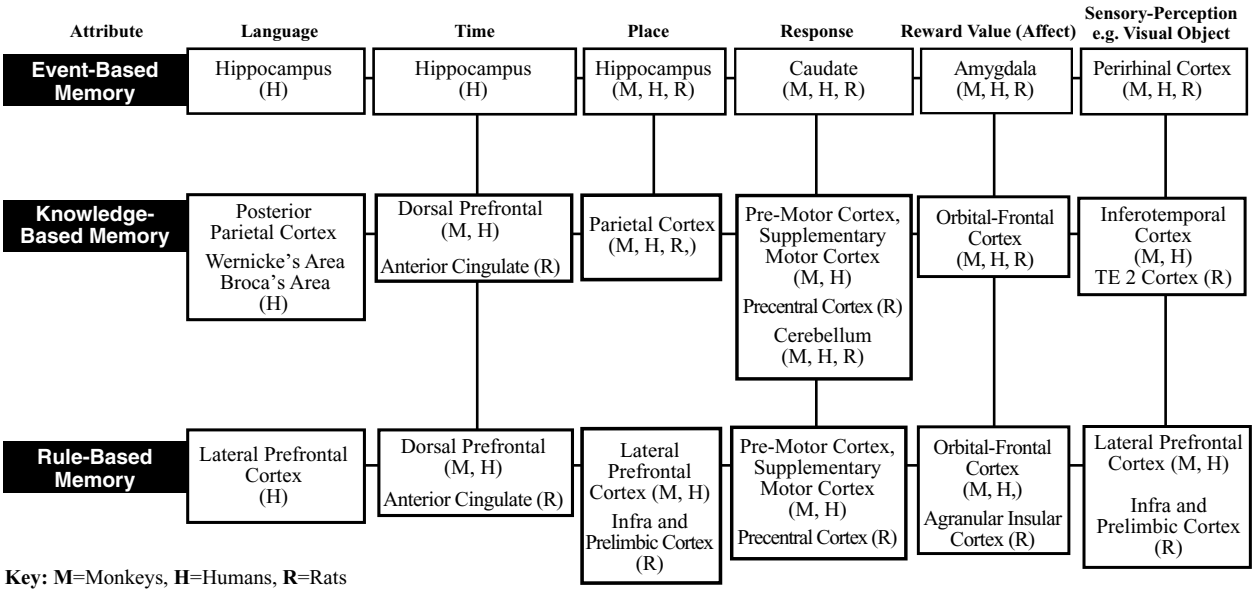


Figure 4 Representation of the neural substrates associated with the event-based, knowledge-based, and rule-based memory systems for the language, time, place, response, reward value (affect), and sensory-perception attributes.

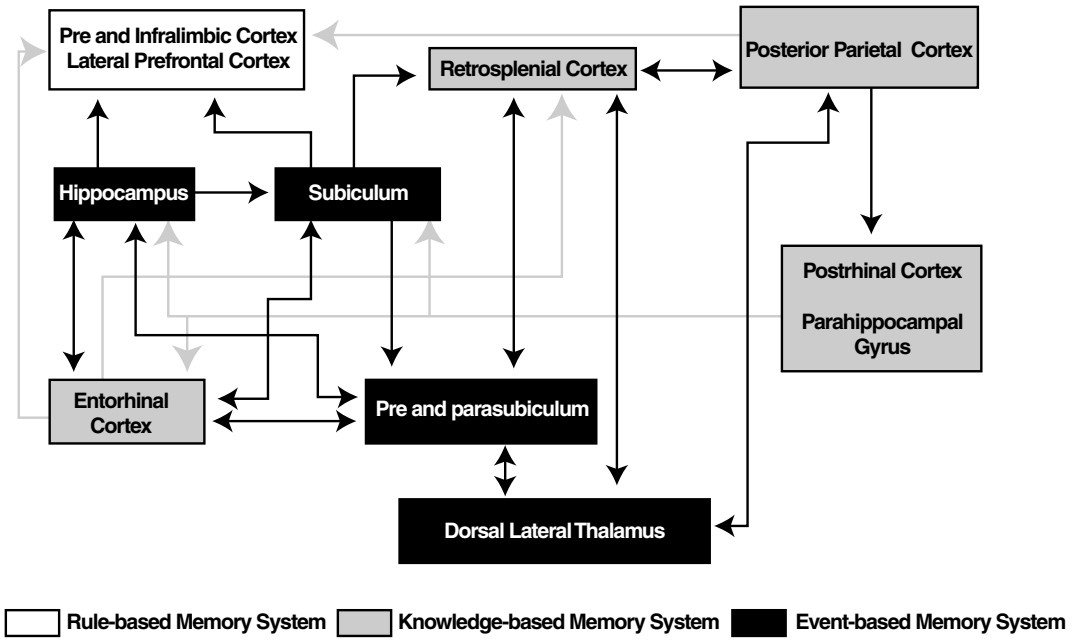


Figure 5 Representation of the spatial attribute neural circuit incorporating neural regions that mediate rule-based, knowledge-based, and event-based memory.

manipulating information for subsequent action. The neural systems that subserve specific attributes within a system can operate independent of each other, even though there are also many possibilities for interactions among the attributes. Although the event-based and knowledge-based memory systems are supported by neural substrates and different operating characteristics, suggesting that the two systems can operate independent of each other, there are also important interactions between the two systems, especially during the consolidation of new information and retrieval of previously stored information. Finally, because it is assumed that the rule-based system is influenced by the integration of event-based and knowledge-based memory information, there should be important interactions between the event-based and knowledge-based memory systems and the rule-based memory system. Thus, for each attribute there is a neural circuit that encompasses all three memory systems in representing specific attribute information. Space only allows for the presentation of one neural circuit as an example. Figure 5 depicts the neural substrates and their interconnections associated with the spatial (place) attribute across all three memory systems. Note that the dorsal lateral thalamus, pre- and parasubiculum, hippocampus, and subiculum represent neural substrates that support the event-based memory system; the entorhinal cortex, parahippocampal gyrus or postrhinal cortex, posterior parietal cortex, and retrosplenial cortex support the knowledge-based memory system; and the lateral prefrontal cortex or pre- and infralimbic cortex support the rule-based memory system. This circuit provides anatomical support for a possible independence in the operation of the hippocampus as part of the event-based memory system and posterior parietal cortex as part of the knowledge-based memory system in that spatial information that is processed via the dorsal lateral thalamus can activate both the hippocampus and the posterior parietal cortex in parallel. Also, information can reach the lateral prefrontal cortex or pre- and infralimbic cortex as part of the rule-based memory system via direct connections from the posterior parietal cortex as part of the knowledge-based memory system and hippocampus as part of the event-based memory system. Finally, spatial information can interact with other specific attributes via a series of direct connections, including an interaction with reward value attribute information via hippocampal–amygdala connections or lateral prefrontal cortex–orbital frontal cortex connections or an interaction with response attribute information via hippocampal–caudate or lateral prefrontal–premotor or supplementary motor connections. In general, the tripartite attribute memory model represents the most comprehensive memory model capable of integrating the

extant knowledge concerning the neural system representation of memory.

See Also the Following Articles

INFORMATION PROCESSING • LANGUAGE AND LEXICAL PROCESSING • MEMORY, EXPLICIT AND IMPLICIT • MEMORY DISORDERS, ORGANIC • MEMORY, NEUROIMAGING • MEMORY, OVERVIEW • SEMANTIC MEMORY • SHORT-TERM MEMORY • WORKING MEMORY

Suggested Reading

Burgess, N., Jeffery, K. J., and O'Keefe, J. (Eds.) (1999). *The Hippocampal and Parietal Foundations of Spatial Cognition*. Oxford Univ. Press, New York.

Eichenbaum, H. (1999). The hippocampus and mechanisms of declarative memory. *Behav. Brain Res*. **103,** 123–133.

Kesner, R. P. (1998). Neurobiological views of memory. In *Neurobiology of Learning and Memory*. (J. L. Martinez and R. P. Kesner, Eds.), pp. 361–416. Academic Press, New York.

LeDoux, J. E. (1995). Emotion: Clues from the brain. *Annu. Rev. Psychol*. **46,** 209–235.

Martinez, J. L., and Kesner, R. P. (Eds.) (1998). *Neurobiology of Learning and Memory*. Academic Press, San Diego.

Roberts, A. C., Robbins, T. W., and Weiskrantz, L. (Eds.) (1998). *The Prefrontal Cortex: Executive and Cognitive Functions*. Oxford Univ. Press, New York.

Schacter, D. L., and Buckner, R. L. (1998). Priming and the brain. *Neuron* **20,** 185–195.

Squire, L. R. (1994). Declarative and nondeclarative memory: Multiple brain systems supporting learning and memory. In *Memory Systems 1994* (D. L. Schacter and E. Tulving, Eds.), pp. 203–231. MIT Press, Cambridge, MA.

Memory, Explicit and Implicit

KATHLEEN B. MCDERMOTT
Washington University, St. Louis

GLOSSARY

explicit memory Intentional retrieval of the past as manifested on a test in which people are asked to recollect the past. Explicit memory is what laypeople typically mean by the term memory.

implicit memory The change in performance as a result of prior experience in the absence of intention to remember the prior event. Implicit memory is often facilitative, although it can cause interference.

generation effect The finding that requiring people to generate information (e.g., from a conceptual clue, "hot–c____") leads to a higher likelihood of remembering the generated information (e.g., "cold") on a later explicit memory test than if the to-be-remembered item is simply presented to the person (e.g., they read the word "cold" or read "hot-cold").

level-of-processing effect The finding that on most explicit memory tests, events encoded with respect to their meaning are more likely to be retrieved at a later time than those processed only to a superficial level (e.g., in terms of visual features or sound).

picture superiority effect The finding that on most explicit memory tests, items previously presented as pictures are better remembered than those previously presented as words.

priming Change in performance on a current task caused by recent prior experience in the absence of intent to use that experience on the current task. This is the measure of memory on most laboratory-based implicit memory tests.

recall A type of memory test in which the subject must produce the previously encountered items, either from minimal cues (as in free recall: "Write down as many words as you can remember seeing in the previous study list") or in response to a cue (as in cued recall: "Tell me the word previously paired with "hot").

recognition A type of memory test in which the to-be-remembered items are presented to the subject, whose task is to decide whether he or she remembers having encountered the item previously (free choice recognition) or to decide which of several items he or she remembers having encountered (forced choice recognition).

Memory researchers typically discuss memory as being a three-stage process: encoding (or the acquisition of information), storage (or the retention of information over time), and retrieval (or accessing information previously encoded). Explicit and implicit memory refer to different ways that past events can be retrieved. Specifically, one can intentionally try to retrieve the past. This is what is commonly referred to as "memory" or memory retrieval and is what experimental psychologists call explicit memory. We "search our brains" for information previously encountered. For example, remembering where you parked your car, remembering your first day of school, and remembering the last time you ate at your favorite restaurant are all examples of explicit memory. Implicit memory, in contrast, refers to the unintentional manifestation of previous experience on a current task. Psychologists generally do not use the term "remembering" when discussing implicit memory; instead, they refer to "priming." If you recently heard an unusual word such as "perspicacious," you are more likely to use that word in conversation than you otherwise would be; You are "primed" to use the word. This is an everyday example of implicit memory. As will be seen, both explicit and implicit memory are typically studied by psychologists through controlled laboratory experiments in which people are given sets of materials (e.g., pictures or words) to remember, and retention of this information is assessed by one or more types of memory test.

I. EXPLICIT MEMORY

A. Definition

Explicit memory can be thought of as intentional retrieval. That is, explicit memory is the willful process of thinking back in time for the purpose of retrieving previously encountered events. It is also sometimes referred to as *episodic memory* because explicit memory involves memory for prior episodes in one's life (as opposed to memory for general knowledge of the world, e.g., who served as the first U.S. president, which is called *semantic memory*). In psychology experiments, explicit memory is usually defined operationally in terms of test instructions. That is, if participants are asked to retrieve a previous event, then the experiment is one that taps explicit memory.

B. Measures

Explicit memory is usually measured with tests of recognition or recall. *Recognition* refers to the case in which the memory test gives an answer and the person must decide whether or not it is correct (called *free choice recognition* or yes/no recognition) or choose from among possible alternatives (*forced choice recognition*). For example, imagine that subjects in an experiment were given a list of 100 words to remember. They might then receive a free choice recognition test in which 50 of these words are presented, mixed with 50 new words (the ratio need not be 50 : 50). The job of the participant is to determine for each word whether it had been presented in the study phase. Thus, they give a yes or no answer to each word. This test is similar to true/false tests often given in school. An alternative test would be forced choice recognition, in which people would be given a set of words, with each set containing at least 1 studied word and at least 1 nonstudied word. The subject's task would be to decide which of the words had been studied. This approach is analogous to what are called multiple-choice tests in educational settings.

Recall differs from recognition in that the answer is not presented to the person. There are several types of recall as well. In *free recall* and *serial recall*, no cues are given; people are simply told to think back to the study phase (in our example here, the word list) and write down (or say) everything they remember. Free recall differs from serial recall in that order of recall does not matter in free recall; in serial recall, however, people are instructed to produce the studied items in the same order in which they were previously experienced. On a *cued recall* test, subjects are given cues to help them remember the list. For example, if the studied list contained the words "zebra," "lion," and "dog," they might be given the cue "animals" to help guide their recall. Perhaps they would be given the first few letters of the words (e.g., "ze___") to help them remember items in the list.

C. Typical Patterns of Results

As discussed previously, researchers generally conceptualize memory as having three stages (encoding, storage, and retrieval). The types of strategies that people use during the encoding (or acquisition) stage have a profound effect on what is remembered at a later time. Considered here are just a few of the classic variables shown to strongly affect performance on a later memory test.

1. Level of Processing

The *level-of-processing effect* is one of the most robust and well-known findings in the explicit memory literature. In two seminal papers on this topic in the mid-1970s, Fergus Craik, Robert Lockhart, and Endel Tulving showed that if people are encouraged to think about the meaning of words (from a list of words to be remembered later), they later recall and recognize those words with a higher probability than if they are encouraged to think about the sound (or phonology). For example, for the word "lizard", people could be asked "Is it an animal?" or "Does it rhyme with wizard?" The former question would lead to a higher probability of recall and recognition on average, across many words in the study list. Similarly, people remember words encoded with attention to sound better than words that are processed at a more superficial level (e.g., determining whether the word is in uppercase letters). This phenomenon is called *level of processing* because it was proposed that people must go through the more "shallow" levels (e.g., the letter level) to access the "deeper" levels (e.g., meaning-based processing). The level-of-processing effect is a very robust phenomenon; one of the primary principles of memory is that if one wants to remember something later, he or she will do well to think hard about its meaning and importance at the time of encoding.

2. The Generation Effect

The *generation effect* is similar to the level-of-processing effect in that it shows that the more meaningfully and effortfully items or events are processed, the better they will later be remembered on explicit memory tests (both recall and recognition). The way the effect is typically studied in the laboratory was popularized by Norman Slamecka and Peter Graf in the 1970s. People are either given antonyms to read (e.g., "hot–cold") or given the first word and asked to generate its antonym (e.g., "hot —"). Later, on recall and recognition tests, people remember "cold" better if they previously had to generate it themselves than if they simply read it. Consider the way children use flash cards to learn vocabulary words or the periodic table of the elements. Flash card techniques take advantage of the generation effect in that the information is generated by the user before looking at the answer. For example, the flash card user might see "mercury" on a card and try to generate its symbol (Hg) before checking the answer on the back of the card. Research has shown that even instances in which the generation attempt fails (e.g., the flash card user cannot successfully remember the item on the other side of the card), the generation effect occurs: Memory for the to-be-generated information can be facilitated relative to the condition of simply reading the information.

3. The Picture Superiority Effect

The *picture superiority effect* refers to the finding (made widely known through experiments by Allan Paivio in the late 1960s) that pictures are remembered better than words. This pattern occurs regardless of the type of explicit test—recall or recognition; it even occurs when the recognition test contains words (referring to pictures or words previously encoded). The source of this effect is thought to be that pictures access meaning more fully than words (and therefore are processed more "deeply" in level-of-processing terms); furthermore, pictures can often be accompanied by a verbal label. For example, if a picture of a fish is shown, people can easily think "fish" or "trout" to themselves when looking at the picture. Thus, pictures tend to access two types of codes (pictorial and verbal), whereas words tend to access only a single type of code (verbal). Of course, people could also form a mental image of a fish when given the word, in which case they would access both types of codes; indeed, when they do so, words are better remembered than when no imagery is invoked.

D. Neural Correlates

What parts of the brain contribute to our ability to remember the past? There are two primary ways of answering this question. The first and traditional way is to use unfortunate accidents of nature—naturally occurring brain lesions—to determine what cognitive processes break down when certain parts of the brain are injured due to stroke, accidents, or other insults to the brain. William Scoville and Brenda Milner described the memory impairments of a man known by his initials, H.M., who had his temporal lobes (including most of the hippocampi) surgically removed in the early 1950s in an attempt to cure intractable epilepsy (Fig. 1). The result was a profound loss of ability to remember anything that happened since the surgery (*anterograde amnesia*). However, H.M. was able to remember most things that occurred before the surgery and was able to converse somewhat normally. This pattern of results and similar outcomes exhibited by other patients with temporal lobe damage suggests that the temporal lobes (in or around the hippocampus) are necessary for the formation of new explicit memories. Interestingly, as will be discussed later, amnesic patients such as H.M. exhibit intact implicit memory (Fig. 2). The point to be taken from this study and other similar studies is that the medial temporal lobes play an important function in remembering the past; when they are removed or damaged, new explicit memories cannot be formed. One difficulty, however, lies in determining which stage(s) of the memory process (encoding, storage, or retrieval) this structure exhibits its effect. Does the information not enter the amnesic's brain properly? Is it encoded properly, with a breakdown in the storage or retrieval phase? Is it coded in one type of memory system (implicit) but not the other type (explicit)?

Recently, a new set of techniques has been developed that allows one to address these and similar questions. Specifically, *neuroimaging techniques* [e.g., *event-related potentials* (ERPs), *functional magnetic resonance imaging* (fMRI), and *positron emission tomography* (PET)] allow researchers to view the normal, living brain as it encodes and retrieves information (and while it processes information generally). By taking advantage of the fact that when specific brain regions are engaged in a task (e.g., a recognition memory test), these regions receive enhanced blood flow relative to their normal resting state (in the case of fMRI and PET) or the fact that electrical activity increases (in the case of ERP), these techniques can inform our understanding of the neural

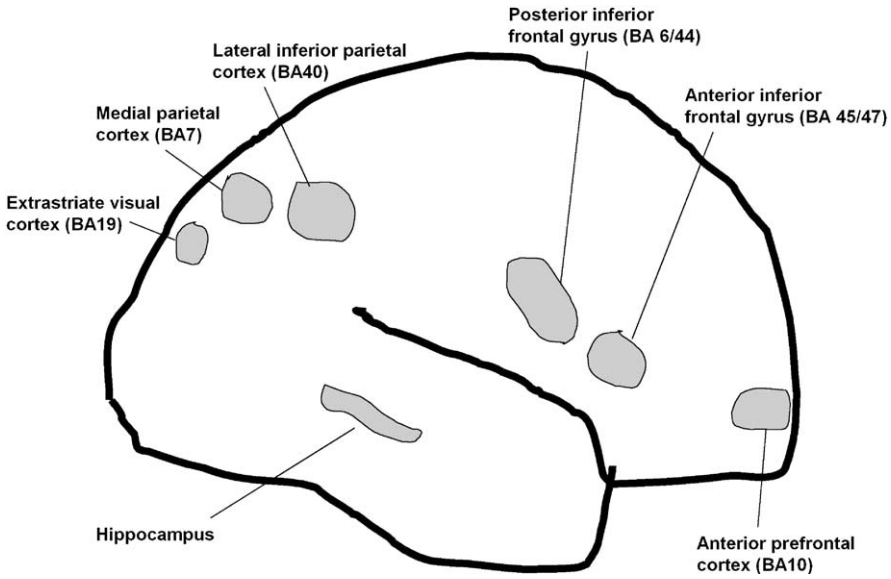


Figure 1 Schematically depicted brain regions involved in explicit and implicit human memory. The front of the brain appears on the right side of the figure. BA, approximate Brodmann's area.

underpinnings of memory. These approaches have shown that several regions are typically engaged during memory retrieval. Interestingly, although the hippocampus is sometimes shown to be active during encoding and retrieval on explicit memory tests, these techniques have pointed to other brain areas as also being important contributors to explicit memory. Some of these regions were relatively unanticipated from the literature on patients with brain lesions. Specifically, a region in the anterior portion of the frontal lobes [in or near Brodmann area (BA) 10] tends to be active when we try to think back to a previous point in time (Fig. 1). This region is sometimes active primarily on the right side and sometimes bilaterally; it is currently a topic of great interest to determine whether the two sides make different contributions and what those contributions might be. Some possibilities are that one or both sides represent the "mental set" or the fact that someone is trying to focus his or her attention to recollect an earlier point in time or that he or she performs reflective processing such as trying to recollect whether an item was seen previously as a picture or as a word. Recent findings demonstrate that these anterior prefrontal regions (in or near BA10) are more active when people successfully retrieve the past than when they try but fail to retrieve the past. A great deal of effort is currently being focused on these questions.

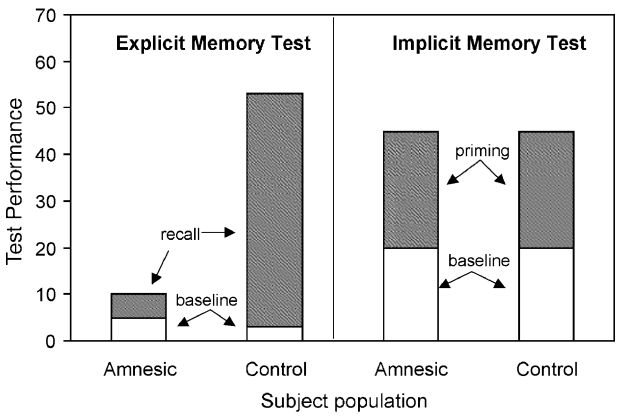


Figure 2 Explicit and implicit memory tests show different patterns of results. Data in this (and all) figures are idealized data demonstrating the general pattern observed across multiple studies. (Left) Patients with damage to the hippocampus and surrounding areas of the medial temporal lobes show profound impairments on tests of explicit memory. The baseline measure represents incorrect answers (or guessing). As shown here, amnesic patients show lower levels of recall than do normal control subjects. (Right) Patients with damage to the hippocampus and surrounding medial temporal lobes demonstrate intact priming. The baseline level represents performance in the unprimed state (e.g., completing a word fragment when the answer was not recently encountered).

A second set of regions indicated by neuroimaging lie within parietal cortex. Regions within both medial parietal cortex (BA 7) and lateral parietal cortex (BA 40) have been shown to be active during retrieval (Fig. 1). Recent studies have shown that they tend to be more active during recognition for items studied before

and correctly recognized by subjects (*hits*) than for nonstudied items correctly classified as such (*correct rejections*). These regions have thus far received less attention than the anterior frontal region, and their role in retrieval is uncertain.

Much more work needs to be done to determine the exact nature of the contribution of these memory-related regions, and such work is ongoing. The relatively new functional neuroimaging techniques will complement lesion studies in that the two techniques typically answer slightly different questions. Functional neuroimaging studies typically examine the normal human brain as it processes information. Functional neuroimaging can also be applied to patient populations, in which case patients are often compared with normal control subjects. Traditional lesion studies involve the study of people who have suffered brain damage; the goal is to determine which tasks they can no longer perform normally. Ultimately, both techniques will be important; if functional neuroimaging identifies a region implicated in performing a cognitive function, it will be important to show that this function breaks down when the region is damaged.

II. IMPLICIT MEMORY

A. Definition

As mentioned previously, explicit memory refers to intentional retrieval. In contrast, *implicit memory* can be thought of as unintentional or incidental retrieval. Implicit memory refers to the change in performance as a result of prior experience without intentionally trying to remember the prior, facilitating event. For example, if you were to read this entry a second time, you would read it faster than the first. This would happen even if you are not conscious of this difference or trying to produce it. What is important is that you are not attempting to draw on previous experience to aid you in the current task; instead, it manifests itself in the absence of your intent to use that information. As will be seen, implicit and explicit memory differ in many important ways.

B. Measures

Implicit memory is measured in terms of *priming*, or the amount of change (often facilitation) observed on an implicit memory test. In order to better understand how implicit memory works, psychologists have devised three main classes of implicit memory tests: perceptual implicit memory tests, conceptual implicit memory tests, and procedural learning.

Perceptual implicit memory tests require people to resolve a perceptually degraded object or word. For example, a word might be flashed very briefly (e.g., 30 msec) on a computer screen, and the task of the person is to try to guess the word. Accuracy in guessing the word is better if the word was recently read. Other types of tests involve completing word puzzles, such as those seen on game shows (e.g., "a _ r _ _ a r _" will be more readily recognized as "aardvark" if the intact word was recently read). Another popular test is to have people fill in the blanks to form the first word that comes to mind that begins with specified letters (e.g., "app_____"). If "apple" were previously seen, it would be primed; it would be used by subjects more often than if they had not previously seen this word. However, if "appendix" were previously seen, it would be the primed word. These tests are called word identification, word fragment completion, and word stem completion, respectively. Similar tests can be employed with pictures, unfamiliar objects, visual patterns, or sounds. For example in picture fragment identification, people are given line drawings of common objects (e.g., a lamp), but parts of the lines have been erased. Their task is to guess the name of the picture from its fragmented form.

The second type of implicit memory test, *conceptual implicit memory tests*, have received much less attention. They, too, use priming as the measure of memory, but they do so by observing how performance on a conceptual, or meaning-based task is influenced by the recent past. For example, if someone asked "What is the name of an airplane without an engine?" a person would be more likely to answer correctly with "glider" if he or she recently encountered that word. Similarly, if a person were asked to say the first word to come to mind when given "aviation", he or she would be more likely to say "glider" than he or she otherwise would if it had not recently been encountered. Peoples' thinking processes in the present are influenced by the recent past, even when there is no attempt to use that recent past to perform the task at hand.

Notice that unlike explicit memory tests, there are sometimes no correct or incorrect answers on an implicit memory test. For example, saying the first word that comes to mind when given another word (i.e., *word association*) or saying the first word that comes to mind that begins with "app____" have many "right" answers; priming, however, is measured in the enhancement of saying a word prespecified by the

experimenter and recently studied. Specifically, researchers search for enhanced probabilities of producing whatever word was recently encountered (usually called the target word or the primed word) relative to the condition in which that word was not recently encountered.

Another example of a conceptual implicit memory test is a "liking" judgment. That is, the answer to "How much do you like this?" can be affected by recent experience. A song heard several times will sometimes tend to "grow on" a person, and this, too, is a form of conceptual implicit memory.

Procedural learning is a third index of implicit memory. The previous example of reading a passage of text faster the second time than the first is an example of procedural learning. Other examples include relearning mazes and recompleting jigsaw puzzles. People can perform complex procedural tasks more quickly and efficiently if they have had recent prior experience with the same materials.

C. Typical Patterns of Results

As mentioned previously, implicit memory tests tend to show different patterns of results from those of explicit memory tests. Perhaps the most dramatic difference is that patients with amnesia (e.g., patient H.M. described previously) perform normally on these tests, despite profound impairments on explicit memory tests (Fig. 2). Indeed, this finding was the spark that produced such great interest in these tests.

With respect to independent variables, perceptual implicit memory tests show patterns markedly different from those exhibited by most explicit memory tests. Conceptual implicit tests, however, tend to exhibit patterns similar (although not always identical) to many explicit tests. The following sections discuss the three patterns of effect of encoding tasks discussed previously with respect to explicit memory (the level of processing effect, the generation effect, and the picture superiority effect), as they relate to perceptual and conceptual implicit memory.

1. Perceptual Implicit Memory Tests

Perceptual implicit memory tests tend to show no (or a very small) difference in degree of priming as a function of level of processing. The explanation is that these tests are thought to be sensitive to the degree of overlap in perceptual features between study and test, and semantic (or meaning-based) processing does not modulate performance. If one sees "aardvark," viewing the visual form of that word will facilitate later reading of the same word because the visual system has an easier time identifying the incoming information. Therefore, verbal perceptual implicit memory tests are primed by verbal stimuli; picture-based perceptual implicit memory tests are primed by picture stimuli (Fig. 3). There is little cross-form priming; that is, if one makes contact with the concept but no visual features match, little or no priming occurs (e.g., seeing a picture of a windmill does not facilitate completion of its fragmented word; similarly, seeing the word "windmill" does not facilitate identification of a fragmented picture of a windmill; Fig. 3).

Similarly, level-of-processing effects and generation effects are typically absent (or very small) in perceptual implicit memory tests. In fact, the reverse pattern is often seen with a read/generate manipulation in that generating the word from a conceptual cue leads to less priming than does reading the word (Fig. 4, left). The explanation calls on whether there is perceptual overlap from the study and test phases; this overlap exists in the read condition, but there is no overlap of visual features in the generate condition.

The picture superiority effect that is seen in explicit memory tests also differs for perceptual implicit memory tests, as alluded to previously (Fig. 3). Because these tests are sensitive to the overlap in perceptual features, whether a picture superiority

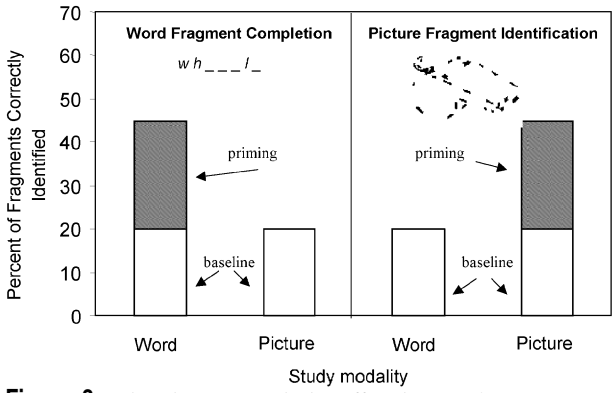


Figure 3 The picture superiority effect is not always seen on implicit memory tests. (Left) Overlap in perceptual features between the study and test phases influences priming on perceptual implicit memory tests. Robust priming occurs on a verbal perceptual implicit memory tests (e.g., word fragment completion) following visual presentation of words (e.g., seeing the word "whistle" primes solving the fragment "w h _ _ _ l _"). However, seeing a picture corresponding to the concept of interest (e.g., a picture of a whistle) does not facilitate performance on word fragment completion. (Right) The converse set of results is seen for picture fragment identification. Seeing a picture primes the identification of its later fragmented form, whereas seeing the corresponding word does not.

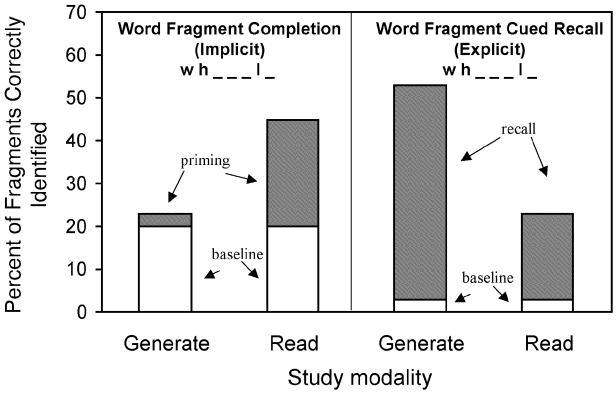


Figure 4 The retrieval intentionality criterion. The retrieval intentionality criterion is met here because the same test cues are used (word fragments) with two different sets of test instructions, and different patterns of results are observed for the implicit and explicit tests. (Left) A reverse generation effect is seen on word fragment completion. Less priming results from generating a word from a clue relative to reading the word. The explanation for this pattern is that there is transfer with respect to the visual features of the word in the read condition but not the generate condition (i.e., the word is seen in the read case, but not the generate case). (Right) The typical generation effect is seen on word fragment cued recall (an explicit test in which people are given a word fragment and asked to complete it with a word from a previously studied list).

effect is present depends on the exact type of test. If a picture-based test is given, the perceptual overlap between study and test will be enhanced for a picture study condition (relative to a word study condition). However, the opposite is true for a verbal perceptual implicit memory test: Encountering words in the study phase primes these more than pictures.

When different patterns are obtained on explicit and implicit tests, the tests are said to have been *dissociated*. The most convincing dissociation occurs when the only feature that differs between the implicit and explicit test is the instructions; when this occurs, it is said that the *retrieval intentionality criterion* is met, as described in the late 1980s by Dan Schacter, Jeffrey Bowers, and Jill Booker. For example, word stems (''whi____'') could be given either with implicit instructions (fill in the blanks with the first word that comes to mind—word stem completion) or with explicit instructions (fill in the blanks to form a word that you encountered earlier in the experiment—word stem cued recall). The only feature that differs between the two cases is instructions: One instructional set requires that people intentionally retrieve, whereas the other does not.

When different patterns of results are obtained on different memory tests, one can argue that different forms of memory underlie the different tests. For example, examine Fig. 4. People either read words during the study phase (e.g., ''whistle'') or generated the word from conceptual cues (e.g., ''blow-w____''). As discussed before, explicit tests exhibit a generation effect such that the generate condition enhances later retrieval relative to the read condition. This pattern is found for the explicit test of word fragment-cued recall in which people are given cues (e.g., ''w h _ _ _ l _'') and asked to use the cue to create a word from the study list. However, if the same cues are given with a different set of instructions, in which people are simply asked to fill in the fragment with the first word that comes to mind, the opposite pattern is observed. The read condition leads to more priming than does the generate condition. In this case, the retrieval intentionality criterion is met; the only procedural difference between the word fragment-cued recall and word fragment completion tests is instructional; Instructions for the former ask subjects to recollect the past, whereas instructions for the latter ask people simply to perform a task to the best of their ability and no mention is made of the relevance of the recent past.

2. Conceptual Implicit Memory Tests

As discussed previously, conceptual implicit memory tests show many of the same patterns exhibited by most explicit memory tests. For example, they show greater priming following meaning-based processing relative to superficial processing (a level-of-processing effect), and they also demonstrate a generation effect. One surprising finding is that a picture superiority effect is not observed on these tests, and the reasons for this unexpected finding are not well understood. In general, however, conceptual priming demonstrates the same patterns of results seen on most explicit tests.

D. Neural Correlates

As alluded to previously, damage to regions in the medial temporal lobes does not produce a general impairment on perceptual implicit memory tests. Whether general impairments occur on conceptual implicit memory tests is more controversial. The finding of intact perceptual priming in patients with damage to the hippocampus and surrounding structures within the medial temporal lobe was reported by Elizabeth Warrington and Lawrence Weiskrantz in the late 1960s. This finding spurred interest in the phenomenon of priming. Prior to this finding, it had been thought that the medial temporal lobes were globally important in memory; however, this finding demonstrated that this was not the case.

If the medial temporal lobes are not critical for producing perceptual priming, then what brain regions are? The brain mechanisms that underlie priming effects differ as a function of the type of implicit memory test. In general terms, brain regions that are critical for performing a task in the unprimed state are *less active* in the primed state. This makes sense if one thinks about priming as facilitation; the brain regions critical to performing the task have to put forth less effort in the primed (facilitated) state.

Consider first perceptual implicit memory tests. Reading a visually presented word taxes the visual system. However, regions in extrastriate visual cortex (Fig. 1) show less activity in the primed state relative to the unprimed state, consistent with the idea that less neural effort is required to perform the task; the neural pathways necessary to accomplish the goal are facilitated. Although less well studied, on the basis of this logic we would expect auditory implicit memory tests (e.g., identifying an auditory word stem) to show less activation in regions of the brain responsible for auditory processing (relative to the unprimed state).

Conceptual implicit memory tests, however, are not sensitive to the match or mismatch in perceptual features between the study and test phases; hence, the neural manifestation is not at the perceptual level. Rather, facilitation is observed at higher level regions of the brain, which are concerned with the task at hand. Consider the case of generating an associate to a presented word (e.g., given the word ''elephant,'' the person would respond with a related word, such as ''tusk''). This task calls on many brain regions, and two critically important regions lie within the left inferior frontal cortex (Fig. 1, anterior and posterior inferior frontal gyri). These regions show diminished activation in the primed condition. Again, it can be seen that regions that are important for performing the task have to put forth less effort to accomplish the task at hand in the primed condition. The brain is more efficient in the primed case.

III. IMPLICATIONS

As previously reviewed, implicit and explicit memory differ as a function of:

1. Subject populations: The finding that amnesic patients exhibit intact implicit memory despite grossly impaired explicit memory spawned a great deal of interest in characterizing implicit memory tests.

2. Independent variables: Many standard findings in explicit memory do not hold up (and often are reversed) for perceptual implicit memory tests. Explicit and implicit memory seem to be fundamentally different types of memory.

3. Neural substrates: Explicit memory tests show *increased* activity in a network of brain regions, including regions within anterior prefrontal cortex and lateral and medial parietal cortex. Implicit memory tests show *decreased* activity in regions critical for performing that task (e.g., in visual cortex as seen for word stem completion).

In the late 1980s, there was a debate over the question of whether different memory ''systems'' were responsible for performance on implicit and explicit memory. This hotly contested question was argued before the tools to observe localized brain activity relatively directly (via PET and fMRI) had become widely available. The availability of these techniques has allowed researchers to go beyond debating whether there is a system or network of regions in the brain responsible for implicit memory (or explicit memory) and instead to focus on the precise role of various specific brain regions to specific memory tasks. The general conclusion is that no single brain region is solely responsible for implicit (or explicit) memory. We do now have an emerging understanding of the network of brain regions underlying implicit memory and explicit memory tests, and there are marked differences between the two. However, differences also exist among various explicit tasks as well as among different implicit tasks. Whether one wants to refer to the networks underlying performance as brain systems is largely a matter of taste. One suggestion advocated here (in collaboration with Henry Roediger and Randy Buckner) is that it makes sense to put the question of systems aside until the individual components of the systems (i.e., individual brain regions) are better understood, with the eventual goal being to understand the entire network of brain regions contributing to implicit and explicit memory in their multiple instantiations.

IV. TERMINOLOGICAL CONSIDERATIONS

I consider here a few difficult issues involving terminology. First, implicit and explicit memory tests have been defined here according to instructions. However, consider the case in which the instructions given to subjects are for an implicit test, but people choose to ignore the instructions and decide to think back to the past in an effort to enhance performance on the test. Is

the test still implicit? Conversely, what if people decide that an explicit memory test is too difficult and therefore begin responding with whatever first comes to mind? Is the test explicit simply because the instructions asked people to think back in time? The approach advocated here is that safeguards can (and in some situations should) be built into experiments to ensure that subjects do indeed follow instructions; this is as true of explicit memory experiments as of implicit memory experiments, however. Thus, instructions to subjects define the test type to a first approximation, but it is desirable to have some behavioral evidence documenting that people did follow those instructions.

A second issue that researchers have wrestled with is the concept John Gardiner termed *involuntary conscious recollection*. This term refers to the situation in which a person vividly recollects some aspect of the past even though he or she is not trying to do so. Consider the case in which, for example, you are walking down the street when seemingly out of nowhere you recall going to the circus as a child. You did not intentionally recall that memory; it simply "popped to mind," perhaps sparked by the unconscious association to some cue in the environment. Does such an experience tend to happen on implicit memory tests—a subject recognizes a word as studied after it has been retrieved—and, if so, does it contaminate the results? Although such phenomena sometimes do occur on implicit tests, research indicates that it need not affect the implicit test results. As long as people do not alter their strategy on the implicit test (i.e., as long as they do not adopt a strategy of attempting to recollect the recent past to aid performance on the test and instead continue to follow instructions), such involuntary conscious recollection does not contaminate the results.

A third difficult issue is that implicit and explicit have been referred to here as representing test types as defined by instructions and have also been referred to as representing different manifestations of memory, which underlie performance on the two types of test. This confounding of terms is pervasive in the field but can lead to confusion. One approach (advocated here) is to apply the terms to the type of test (implicit or explicit memory test) because test type can be defined (at least approximately) by instructions, whereas "implicit memory" and "explicit memory" cannot be observed directly. However, the title of this article referred to implicit and explicit "memory," and these concepts are widely discussed in the literature; therefore, implicit and explicit are used in both senses throughout this article.

V. SUMMARY

Perceptual implicit tests of memory and traditional, explicit tests of memory demonstrate fundamentally different types of memory. The former is preserved in amnesia, sensitive to the overlap in perceptual details between the study and test phases of experiments but relatively insensitive to higher level strategies. Conversely, explicit tests of memory demonstrate profound decrements for amnesic patients. Also, they are relatively insensitive to mismatches in perceptual features between the study and test phases, but are highly sensitive to the types of study strategies invoked during the study phase. Understanding both types of memory and the brain regions contributing to them will provide a more complete understanding of the workings of human memory.

See Also the Following Articles

ALZHEIMER'S DISEASE, NEUROPSYCHOLOGY OF • CATEGORIZATION • COGNITIVE AGING • MEMORY NEUROBIOLOGY • MEMORY, NEUROIMAGING • MEMORY, OVERVIEW • NERVE CELLS AND MEMORY • PRIMING • SEMANTIC MEMORY • SHORT-TERM MEMORY • WORKING MEMORY

Suggested Reading

Gardiner, J. M., and Java, R. I. (1993). Recognising and remembering. In *Theories of Memory* (A. Collins, M. A. Conway and P. E. Morris, Eds.), pp. 163–188. Erlbaum, Hillsdale, NJ.

McDermott, K. B. (2000). Implicit memory. In *The Encyclopedia of Psychology* (A. E. Kazdin, Ed.), pp. 231–234. Oxford Univ. Press, Oxford.

Richardson-Klavehn, A., and Bjork, R. A. (1988). Measures of memory. *Annu. Rev. Psychol.* **39,** 475–543.

Roediger, H. L. (1990). Implicit memory: Retention without remembering. *Am. Psychol.* **45,** 1043–1056.

Roediger, H. L., and McDermott, K. B. (1993). Implicit memory in normal human subjects. In *Handbook of Neuropsychology* (F. Boller and J. Grafman, Eds.), Vol. 8, pp. 63–131. Elsevier, Amsterdam.

Roediger, H. L., Buckner, R. L., and McDermott, K. B. (1999). Components of processing. In *Memory: Systems, Process, or Function?* (J. K. Foster and M. Jelicic, Eds.), pp. 31–65. Oxford Univ. Press, Oxford.

Schacter, D. L. (1987). Implicit memory: History and current status. *J. Exp. Psychol. Learning Memory Cognition* **13,** 501–518.

Schacter, D. L., and Buckner, R. L. (1998). Priming and the brain. *Neuron* **20,** 185–195.

Tulving, E., and Schacter, D. L. (1990). Priming and human memory systems. *Science* **247,** 301–306.

Memory, Neuroimaging

JOHN JONIDES, TOR D. WAGER, and DAVID T. BADRE
University of Michigan

GLOSSARY

declarative memory Information stored in memory that can be retrieved explicitly.

encoding The set of processes that transform some physical event into a memory representation.

episodic information Information in memory that is tied to the time and/or place of occurrence, such as being tied to one's autobiographical past.

executive processes The set of operations that allows one to shift attention from one task to another, plan a set of operations, tie information to its context, and generally modulate the operation of other mental processes.

long-term memory The system of storage that is the large repository for information that is stored for very long periods of time.

procedural memory Information that allows one to engage in some skill, including motor and cognitive skills.

retrieval The processes responsible for extracting information from memory for some purpose. The most frequent kinds of retrieval are recall and recognition.

semantic information Information that is generic in form without being tied to some time or place of occurrence, such as knowledge of facts about the world.

storage The processes responsible for holding information in memory for some period of time.

working memory The system of storage that is responsible for small amounts of information for short periods of time.

The study of memory has been remarkably facilitated by the use of neuroimaging tools in recent years. As we now know, memory is not a unitary function or set of processes; rather there are multiple systems of memory that underlie our cognitive life. This article examines the full taxonomy of memory in humans, and reviews what we know about the brain basis of memory from neuroimaging studies. Beyond this, the purpose of the article is also to illustrate how knowledge of the brain mechanisms of memory can help inform us about a proper psychological view of how memory works functionally.

I. SOME INTRODUCTORY CONCEPTS

A. Types of Memory

One of the critical properties that makes the human mind so extraordinarily suited to understanding and dealing with the world is its ability to shift in time—to model the future and reconstruct the past. Reconstruction of the past requires memory, and memory is fundamental to nearly any cognitive skill. It is involved in complex processes such as problem-solving, and it is involved in even what seem to be the simplest skills, such as recognizing a familiar face. The role played by memory in cognition is complex enough that not just a single memory system will do. Humans and other animals have several memory systems with different characteristics and different neural implementations,

and these systems, acting in concert, contribute to the human mind's tremendous adaptability.

At the broadest level, one can distinguish between "working memory" and "long-term memory." Working memory refers to the system that stores a small amount of information for a brief span of time. Information stored in working memory is then used in the service of other cognitive tasks. For example, if we were solving an arithmetic problem such as 817 + 723 without the benefit of writing anything on paper, working memory would be used to store the problem, store the intermediate steps in the addition, and store the final solution. In addition to temporary storage, an important component of working memory is what is called "executive processing:" the set of operations that permits one to manipulate the contents of working memory. In the previous example, executive processes would be involved in switching attention from one column of addition to another and in organizing the order of steps to arrive at a final sum. Whereas there is as yet no overall agreement about a full list of executive processes, they generally can be thought of as operations that regulate the processes operating on the contents of working memory, processes such as selective attention to relevant information (more about this shortly).

In contrast to the short duration and small capacity of working memory, long-term memory is a system with very long duration memory traces and a very large storage capacity. In our previous mental arithmetic example, long-term memory would be the repository of the facts of addition that would be needed to solve the mental arithmetic problem. Of course, long-term memory stores much more than that. For example, it is the repository of all the words we know in our language, of the sensory information that we all have stored for untold numbers of events (e.g., the taste of a good chocolate), of the spatial information we have stored for navigating around our world, and so on. In addition, many pieces of information are stored that we normally do not retrieve consciously but that nevertheless guide our everyday behavior, such as the rules of language or habitual actions in which we engage every day.

Larry Squire of the University of California and Endel Tulving of the University of Toronto have proposed schemes that summarize the various forms of long-term memory. One way of synthesizing and expanding these schemes is shown in Fig. 1. The figure shows that there are two broad divisions of long-term memory: declarative and procedural. Declarative memory refers to the facts and events that we can retrieve at will, often consciously. By contrast, procedural memory refers to stored information that has an impact on our behavior but that is not willfully retrieved. Consider, for example, the concept of a bicycle. A declarative memory you might have of a bicycle is that it is blue and that it has 21 gears, mountain terrain tires, two handbrakes, and so on. These are all facts that can be willfully retrieved from memory. By contrast, you also have stored information that allows you to ride your bicycle—a task that any 6-year-old child will tell you is not easy. This information is not consciously retrievable; indeed, it is a nontrivial problem in physics and kinesiology to describe just how people are able to ride a two-wheeled bicycle without falling over. The contrast between these two sorts of memory is a contrast between declarative and procedural memory. Perhaps the most compelling evidence that procedural knowledge is different from declarative knowledge is that patients with damage to their hippocampi and surrounding medial temporal lobes can learn new procedural skills, even though they cannot encode where they learned

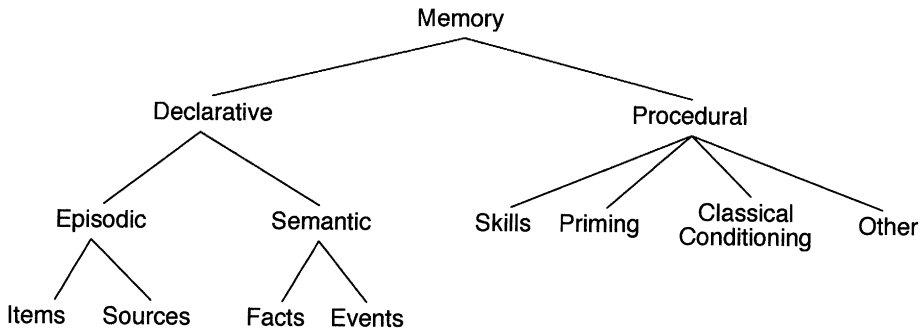


Figure 1 A taxonomy of various forms of long-term memory.

the skill or remember any details of having practiced it, even when that practice occurred very recently. Other patients with cerebellar damage can remember the practice sessions, but their skills on most motor tasks do not improve. This pattern of deficits, called a double dissociation, helps to define procedural and declarative processes as distinct types of memory.

Declarative memory itself comes in two forms. One is called episodic memory, or memory for specific events, and it consists of memory traces that are accompanied by memory for the context in which they were formed. Each piece of episodic memory has a source tag associated with it, possibly including the time and place of memory formation and other details about the context. When retrieving an episodic memory, one can retrieve either the item itself, given information about the source, or the source, given information about the item. For example, you may recall where and when you purchased your current bicycle or, given the time and place, you may recall the features of the bicycle that you purchased. The other category of declarative memory is semantic. This type of memory consists of the vast store of facts and events that you have in long-term memory, regardless of whether you can retrieve when and where you learned them. For example, you may remember the fact that bicycles can be mountain bikes, racing bikes, hybrid bikes, and so on, yet you may not be able to recall when or where you learned this semantic fact.

Procedural memories also are of various sorts. There are skills, for example, such as riding a bicycle. There are classically conditioned responses, which entail a previous pairing of an unconditioned with a conditioned stimulus to yield a conditioned response. And there are cases of priming, in which a previously learned piece of information facilitates processing of some new piece of information. Psychological measures of priming, such as decreases in response time to recognize a previously viewed word, indicate that a trace of the previously learned piece of information is affecting current cognitive processing—even if there is no conscious recollection of having seen the word before.

Another important dimension of memory, whether working or long-term, is the type of information being stored. As we shall see later, the brain circuitry involved in a memory task honors the type of information that is stored and retrieved. Perhaps the most frequently studied case of this concerns the distinction between linguistic information (such as letters, words, sentences, and stories) and visual or spatial information (such as a scene, an object, a face, or a spatial environment). By now ample evidence exists that the two hemispheres of the brain are differentially activated by these two types of information, with the left hemisphere specialized for verbal information and the right for visual or spatial information in most humans.

B. Types of Processes

Memory entails three cognitive operations: encoding, storage, and retrieval. These terms refer to the sequence in which memory processes are thought to occur. Entering information is first put into the proper internal code and a new memory trace is formed (encoding). Encoding is followed by storage of the information for some period of time. This stage may include consolidation or alteration of memory traces to make them last longer and ease retrieval. Retrieval is the process of reporting information from storage.

The nature of encoding depends on two factors: the type of material that is involved in the memory task and the task that is performed with that material. The type of material exerts a strong influence on the path of activity in the brain early in the processing sequence. The best example of this is the visual system. Spatial information about a visual stimulus is selectively routed to a dorsal stream of information processing that mainly includes the parietal lobes, whereas information about shape and other nonspatial object features of the same stimulus is processed by a ventral stream in the occipital and inferior temporal lobes. To generalize from this example, we can say that the nature of incoming information will influence the path of processing that the information takes in the brain. Beyond this, though, there is also an influence of the task with which a person is faced, as many different operations may be performed on any given type of material. For example, one can process a word by noting its meaning or by noting whether it is printed in uppercase or lowercase letters. These very different types of processing on the same stimulus yield different patterns of activation in the brain, as we shall see later.

Once encoded, information is retained for some period of time. Consistent with the fundamental distinction between working and long-term memory, the length of the retention interval in large part will determine which of these systems is most heavily involved. Retrieval after short retention intervals—up to, perhaps, intervals as long as 30 sec to 1 min—uses working memory. Retrieval of information stored for longer periods will, under most circumstances,

necessitate the involvement of long-term memory storage. Which memory system is engaged will be revealed by the circuitry that is activated. Working memory engages circuitry in frontal and parietal cortices most prominently, whereas long-term memory requires the involvement of frontal and parietal circuitry as well as hippocampal and parahippocampal mechanisms.

Just as encoding different types of material engages different mechanisms, storage of different types of material also requires different mechanisms. This has been demonstrated most handsomely in the contrast between verbal and visual material, which predominantly activate left and right hemisphere structures, respectively. This distinction has been demonstrated for both working memory and long-term memory, as we shall see later.

Once encoded and stored, information in memory can then be retrieved as needed. Suppose, for example, that we ask a person to memorize a list of words. Retrieval can be accomplished in several ways. We might simply ask the person to recall as many of the words as possible (free recall). Or we might guide recall by giving some of the words on the list as hints and asking the person to recall the others (cued recall). Or we might present the person with a longer list of words, some of which were presented on the original list and some not, and ask the person to decide which is which (recognition). Any of these procedures requires the person to access the stored information in memory and produce an explicit response that depends on that stored information. For this reason, these are often called explicit tests of memory. However, there are also implicit tests. Suppose, for example, that we presented someone with a list of words and later flashed the same words and new words, one by one, so briefly that they were difficult to identify. If the person were more accurate in identifying words that had been presented on the original list than ones that had not (which is what happens in this perceptual identification situation), then we could conclude that the original words were stored in memory even though no explicit retrieval of them was ever demanded. The process of storage and use of information without explicit memory is called priming. Evidence from positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) suggests that implicit and explicit tests of memory recruit different brain areas, as reviewed later.

With these preliminaries about memory in place, we are now in a position to review what neuroimaging evidence has contributed to understanding basic mechanisms of human memory.

II. WORKING MEMORY

The canonical model of working memory is due originally to Alan Baddeley, and it is this model that has been investigated in detail using neuroimaging methods. The model claims a fundamental distinction between short-term storage of information and the executive processes that manipulate this information. This general view is supported by the existence of patients who have intact short-term storage but deficits in executive processes; this pattern of impairments contrasts with that of other patients who have deficits in executive processing but intact short-term storage. Such a double dissociation suggests that the circuitries of storage and executive processing are separable, and imaging studies have confirmed this separability.

A. Short-Term Storage

The short-term storage of information in working memory appears to be accomplished via two mechanisms: one that retains information and another that "rehearses" that information in order to keep the memory traces active during a retention interval. This is perhaps best illustrated for verbal information. A task that has been used frequently to study the mechanisms of verbal working memory in neuroimaging experiments is the item-recognition task. In this task, participants are presented with a small number of target items, typically randomly selected letters, to store for a retention interval of several seconds. Following this interval, a single probe item is presented and participants must decide whether this item was a member of the memorized set. When participants engage in this task in PET and fMRI settings, a number of easily replicable sites of activation exist compared to a control condition that does not require memory at all or in which the memory requirement is minimal. One frequent site of activation is in the posterior parietal cortex, typically more prominently in the left hemisphere than the right. In addition, a set of activations appears in frontal areas, including the inferior frontal gyrus on the left, premotor cortex (more prominently on the left than on the right), and supplementary motor cortex. These brain regions, and all other major regions discussed throughout this article, are shown in Fig. 2.

The frontal cortical areas that are activated in this task are quite similar to those activated in a task that requires one to make judgments of rhyming, a task that

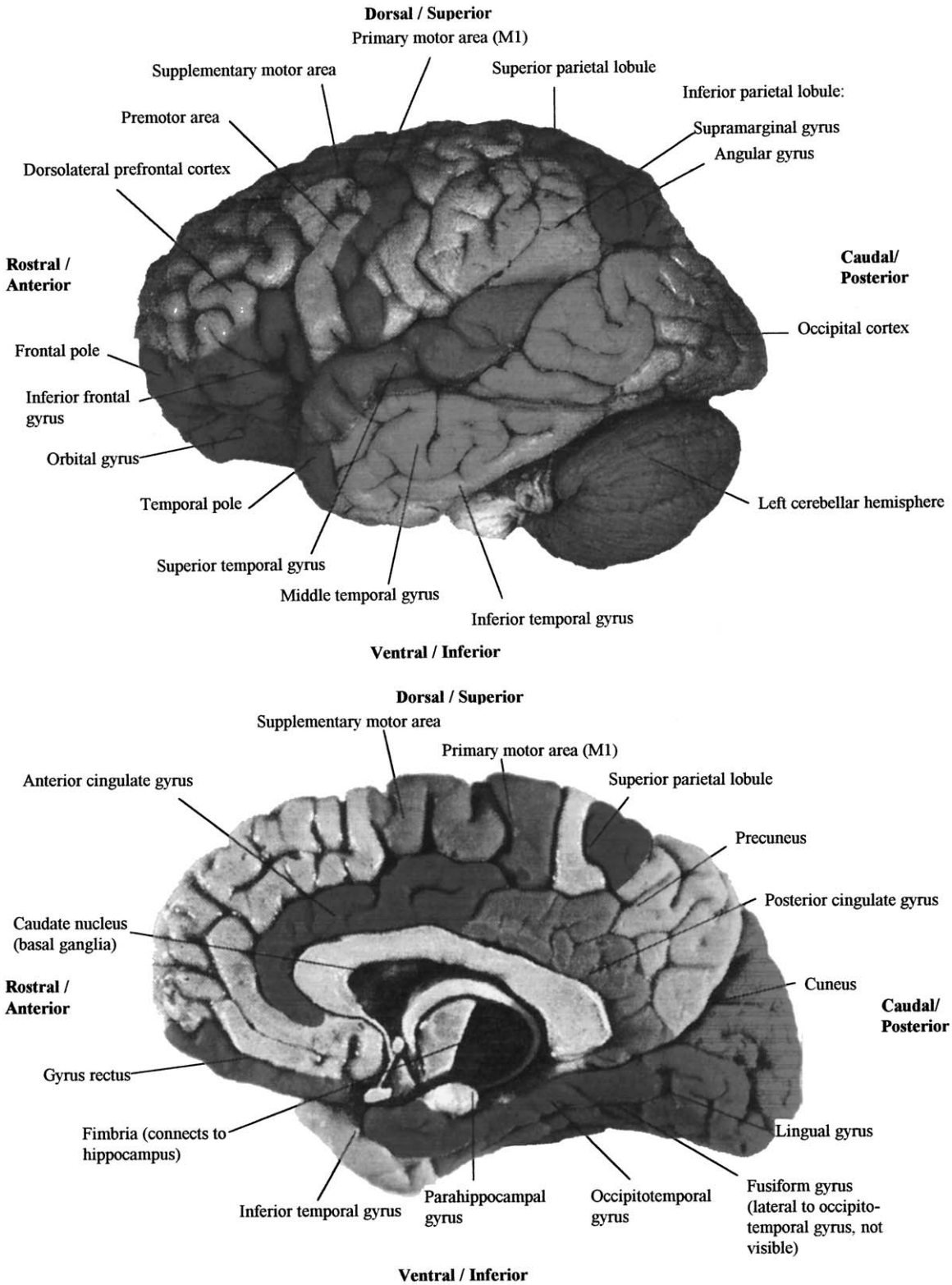


Figure 2 Brain diagrams highlighting the major structures discussed in the text. The upper figure shows a lateral view of a left hemisphere of the human brain, and the lower figure shows a view of the right hemisphere as seen from the midline of the brain. Major structures of relevance to memory are labeled.

presumably requires production of a speechlike representation. So, it is likely that these frontal areas are the ones involved in rehearsal, which involves internally generating and regenerating a speechlike code for the stored verbal material. The posterior parietal sites have been suggested as sites for the storage of verbal information as well as for switching attention between one item and another.

The purported dissociation between the frontal and parietal sites is nicely supported by a study that used a different task involving verbal working memory, the two-back task. In this task, participants see a series of letters presented at a pace of one every 2.5 sec, and they must judge whether each matches in identity the one that appeared two letters back in the series. This task clearly requires storage and rehearsal of each letter, as well as other processes that we discuss next. Compared to a task in which participants must simply judge whether each letter in the series matches a single target (say, the letter "P"), the two-back task produces activations in regions similar to those in the item-recognition task. This is as it should be if both tasks involve storage and rehearsal. Beyond this, though, the two-back task has also been compared to another condition, one in which participants had to silently rehearse letters to themselves with little storage requirement (e.g., say the letter "P" for 3 sec, followed by silently saying the letter "M," and so on). Subtraction of the activation in this rehearsal condition from that in the two-back condition revealed much lower activation in the frontal areas. The rehearsal condition is presumed to involve the explicit production of silent speech. Subtraction of the activations in this condition from those in the two-back condition reduces frontal but not parietal activation; therefore, one can conclude that the frontal activations in the two-back and other verbal working memory tasks must reflect an inner rehearsal process as part of those tasks. These same frontal regions are also activated in tasks that require a recall response, so they are not unique to the peculiarities of the item-recognition task or the two-back matching task.

Just as we can identify the frontal sites used in verbal rehearsal, we can also identify the parietal sites used in verbal storage. Evidence that the parietal sites are used in part for storage comes from a study in which subjects memorized a set of nonsense letter strings (e.g., "MAVER") and then kept these items in memory during a retention interval of some 50 sec, during which they underwent PET scanning. After the scan, they had to retrieve the items to be sure that they had been stored accurately. Scanning during just the retention interval allows one to isolate storage processes or at least to concentrate scanning on storage. One study using this procedure found posterior parietal activations, leading to the conclusion that these activations reflected storage processes and not encoding or retrieval processes.

Storage and rehearsal should not be restricted to verbal information, of course, if they are general properties of working memory as Baddeley has supposed. Indeed, many studies have investigated the storage and rehearsal circuitry used for spatial information as well. The clearest result of these studies is that the circuitry activated by spatial information in a working memory task is quite different from that activated by verbal information, even when the tasks are quite similar and only the material differs. For example, in an analog to the item-recognition task, subjects are presented with a set of dots on a screen and asked to store their locations in memory. Following a retention interval of several seconds, they are presented with a single probe dot, and their task is to decide whether it appears at the same location as one of the locations they have stored. This task has the same formal structure as the item-recognition task for letters, yet it yields activations that are quite different. In common are activations in the posterior parietal and premotor cortex, although with a tendency for greater activation in the right than the left hemisphere. However, quite different are activations in the occipital cortex, superior frontal cortex, and inferior frontal cortex, most prominently in the right hemisphere.

The common activations in parietal and premotor cortex between verbal and spatial versions of the task suggest that there are some processes in common between the tasks, possibly having to do in part with allocating attention to several items in memory. However, the differences in activations suggest that the mechanisms by which information is stored and rehearsed may be different. Indeed, there is evidence of a similarity in circuitry between processes mediating spatial working memory and those mediating shifts of attention to various locations in the visual field when stimuli are being perceived. This leads to the conclusion that spatial rehearsal may amount to a successive allocation of attention to internal representations of spatial locations, a process possibly mediated by premotor mechanisms near the frontal eye fields. This region, together with parietal cortex, may also play a role in maintaining the representations of the spatial locations as well, a conclusion that is consistent with lesion studies and electrophysiological studies of monkeys in spatial working memory tasks. So, we

can see that, although storage and rehearsal are common features of spatial and verbal working memory, they appear to be implemented in the brain in different ways.

Of course, visual information that is stored need not be spatial in nature. Features such as the shape of an object or its color are not spatial, even though they are visual. As described earlier, the brain honors this distinction in simple visual processing, and indeed neuroimaging research suggests that spatial memory and memory for other visual information are processed differently in the brain as well. One experiment that demonstrates this used pictures of three faces presented sequentially in three different spatial locations. After a retention interval, a probe picture was presented in one of the locations. When subjects were tested on their working memory for objects, they had to decide whether the probe face was the same as any of the previous three; when they were tested on spatial working memory, they had to decide whether the probe was in the same location as one of the original faces. The elegance of this design is that it involves the very same stimuli, and only the nature of the memory task changes. The results show that this change in task produces an important difference in brain activation: The object task activated regions of dorsolateral prefrontal cortex, whereas the spatial task activated a region posterior to this in the premotor cortex. Beyond this, a meta-analysis of several spatial and object working memory tasks suggests that there is also a dorsal–ventral difference in activation in the posterior cortex. Spatial working memory tasks activate more dorsal structures in the posterior cortex, whereas object working memory tasks activate more ventral structures.

B. Executive Processes

In addition to storage components, the model of working memory proposed by Baddeley includes a component due to executive processes. Although there is not yet a clear taxonomy of executive processes in hand, descriptions of them typically include the following: (a) focusing attention on relevant information and inhibiting attention from irrelevant information; (b) scheduling processes in tasks that require multiple processes; (c) planning and prioritizing a sequence of steps to meet some goal; (d) updating and checking the contents of working memory; and (e) coding internal representations for time or place of occurrence. All of these processes involve the manipulation of information that is temporarily stored in working memory. Research on executive processes using neuroimaging techniques has revealed a heavy contribution of frontal mechanisms regardless of the executive process in question.

As an example, recall the verbal item-recognition task. In that task, subjects are presented with a set of letters that they have to retain for several seconds, after which they have to decide whether a probe letter matches one of the letters in memory. Several studies have introduced an inhibitory component in this task in the following way. Trials were included in which the distractor probes (probes that did not match an item in the current memory set) were letters that did match a letter in the memory set from the previous trial. Thus, these probes were relatively familiar because they had been memorized recently. This design creates a situation in which participants have a sense of familiarity about the probe item, but they must remember that it does not match the memory set on the current trial. On such trials, subjects take longer to give a "no match" response. Both PET and fMRI studies show that there is a site in the left lateral prefrontal cortex that is activated on these trials, and the activation occurs most prominently at the time the probe is presented. Furthermore, older subjects, who show a greater interference effect on these trials, also show less activation at this left lateral site, and patients with damage to this area show a dramatically increased interference effect compared to patients with damage elsewhere in the frontal cortex. Taken together, this evidence suggests that the left lateral site is involved in resolving the conflict between familiarity and source memory that arises on these trials.

Another example of a task in which executive processes interact with storage processes is the two-back task described earlier. Recall that, in this task, single letters are presented in succession and subjects must judge whether each letter matches the one two earlier in the sequence. To succeed at this task, one not only has to store the recent stream of letters but also has to update this stored set as new letters are presented, dropping older letters and adding newer ones. This task is similar to the item-interference task in that it includes an inhibitory component, as described earlier. In addition to this executive process, the letters that are stored in memory also have to be tagged by their order of appearance so that the subject can keep in mind which one is two back, which is one back, which is three back, and so on. Thus, the two-back task must recruit an executive process responsible for temporally tagging information, a sort of

short-term episodic memory requirement. Indeed, the two-back task shows evidence of activations in the dorsolateral prefrontal cortex in addition to other sites that may well be responsible for temporal tagging. The dorsolateral prefrontal activation that arises in this task seems to be a common broad site of activation in many tasks that require manipulation of the information stored in working memory, and so this leads to the general conclusion that prefrontal mechanisms may be responsible for a wide array of executive processes.

C. Summary of Working Memory

Overall the neuroimaging research concerned with working memory has reliably revealed a set of structures that may be important for storage, rehearsal, and executive processes. Posterior parietal mechanisms have been implicated in the storage of verbal material, and prefrontal ones concerned with language processing have been implicated in the rehearsal of stored verbal material. For spatial material, the sites of storage and rehearsal are different; nonetheless, one can conclude that there are storage and rehearsal processes for nonverbal material as well, but that these may be implemented via nonlinguistic mechanisms. Finally, various sites in the prefrontal cortex, most prominently dorsolateral prefrontal areas, have been documented in the mediation of executive processes. Thus, the psychological architecture proposed by Baddeley in his model of working memory seems to be amply supported by a brain architecture that may honor the same distinctions among processes.

III. EPISODIC MEMORY

As described earlier, episodic memory can be defined as memory for information that is associated with a time and place of occurrence. Take, as an example, a semantic fact: one may know that the turn of the century French impressionist painter Claude Monet lived and worked for many years in his provincial home at Giverny. This fact is in the domain of semantic memory. However, one's memory of learning this fact in an art history course would be an episodic memory. Episodic memory is often studied in a controlled laboratory setting using recognition or recall tasks, described in the introduction to this article. These tasks require memory for a source code (e.g., time or place of occurrence) that is the essence of episodic memory. In the context of neuroimaging, the encoding and/or retrieval phases of these tasks are scanned using PET or fMRI and then compared to a control task with a diminished or absent demand on memory.

These studies have identified a set of regions underlying episodic memory. These include medial temporal structures, such as the hippocampus and parahippocampal areas, prefrontal cortex, anterior cingulate cortex, cerebellum, and parietal and superior temporal association cortices (shown in Fig. 2). Important hemispheric, regional, and functional differences exist, however, between the encoding and retrieval phases of episodic memory. In addition to exploring these differences, neuroimaging studies have also begun to examine cases in which this system performs inadequately.

A. Episodic Encoding

As discussed earlier, memory entails three important general stages: encoding, storage, and retrieval. At the encoding stage, processes must be involved that create an internal code for a piece of information and then attach a context (a place or time) to the new memory.

Several neuroimaging studies have scanned participants while they perform some task to encode a set of items. For example, participants might be asked to make a judgment about whether a word represents an animal or vegetable, an encoding task that requires access to the semantics of the word. Alternatively, a subject might simply be asked to memorize a set of items and be tested on them later. Subsequent testing of the items confirms whether subjects have effectively encoded the items. These studies show that encoding involves the left prefrontal cortex, hippocampus, parahippocampal cortex, anterior cingulate, and some superior temporal cortex. Further experimentation, including converging evidence from neuropsychology and other experimental paradigms, has begun to examine the role of each of these regions and their relationships to one another.

The hippocampus and surrounding areas have long been associated with memory. Evidence from both animal studies and studies of brain-damaged patients has shown that damage to the hippocampus can result in amnesia, one form of which is caused by damage to medial temporal structures such as the hippocampus and parahippocampal gyrus. Though amnesics typically are able to retrieve memories from their distant past, they show a profound deficit in the ability to form new memories, a phenomenon known as anterograde

amnesia. For this reason, the hippocampus is thought to be involved with the encoding and consolidation of long-term memories.

In line with this, many neuroimaging studies using the encoding paradigms described earlier have shown hippocampal activity. Neuroimaging evidence has shown, however, a selective response of the hippocampus to novelty. In one experiment, participants were shown pictures of indoor and outdoor scenes while in the MRI scanner. They were required to judge whether each scene was an indoor scene or an outdoor scene and remember the pictures for a later test. In some scans, the same two pictures were repeated many times so that participants became very familiar with them. During other scans, the scenes were entirely novel and unfamiliar. Comparison of the unfamiliar scans to the familiar scans showed activity in the parahippocampal gyrus bilaterally. Given this fact, it would seem that the medial temporal lobe is particularly responsive to novel stimuli—a finding consistent with the intuition that most episodic memory encoding occurs on the first presentation of new material.

The function of the left prefrontal cortex appears to involve processing the context (or source) in which new information is learned. An event-related fMRI experiment has studied the different functions of left prefrontal and hippocampal mechanisms in episodic memory. Event-related fMRI allows the examination of areas of the brain that are active in response to different events occurring within the context of a single cognitive task. In this experiment, participants were required to learn word pairs in which the first word served as semantic context for a second word, for example, ''athlete–boxer.'' Participants were presented with several of these word pairs during each scan. Sometimes the context for a word would change, as in ''dog–boxer.'' Other times the word would change as in ''dog–labrador.'' Both word and context could also be new or both could be old. This design permitted independent manipulation of the novelty and the context of the item to be learned during encoding. The hippocampus was active when either the context word or the related word was new, and it was most active when both were new. This corroborates the idea that the hippocampus is involved in processing novel items. The left prefrontal cortex was most active when an old context was attached to a new word or a new context attached to an old word. This finding suggests that the prefrontal cortex is involved in representing the context of the item to be remembered, a function that is critical for episodic memory.

We can test whether the effectiveness of encoding is related to the brain activations that reflect encoding by varying what is called the depth of processing participants apply to material. It is well-known that evaluation of the semantic content of material (deep encoding) leads to more elaborate processing and a longer lasting memory trace than evaluation of the physical features of material (shallow encoding). One experiment that takes advantage of this effect required subjects to judge whether a word was abstract or concrete (deep encoding) or whether it was printed in upper- or lowercase characters (shallow encoding). Deeply encoded words were remembered better than words encoded shallowly, replicating previous behavioral results. When the two conditions were compared, it was found that there was greater activity in the hippocampus and left prefrontal cortex for deep encoding, suggesting that both areas are more vigorously involved with deep than with shallow encoding.

This result by itself does not indicate that more activity in the prefrontal cortex and hippocampus produces better behavioral performance; it only indicates that depth of encoding and activation are correlated. To address the performance question, several studies have directly examined the relationship between performance on retrieving an individual item in memory and brain activation while encoding that item. After being scanned, the participants had to recognize the encoded items, and they were grouped by whether the items were retrieved correctly or incorrectly, an indication of good or poor encoding, respectively. This comparison revealed activity bilaterally in the hippocampus and in the left prefrontal cortex. Hence, it would seem that for effective encoding not only must the information be consolidated effectively by the hippocampus but the prefrontal cortex must also assist in processing the context.

To summarize, encoding recruits a set of regions that includes the left prefrontal cortex, hippocampus, parahippocampal gyrus, parietal cortex, and anterior cingulate (as shown in Fig. 2). These regions appear to be involved in transforming information into a mental code in the brain that can later be retrieved. Two processes entailed by this task are the consolidation of a novel item by the hippocampus and the processing of its context by the left prefrontal cortex. The extent to which the information being encoded can be effectively recovered at a later time is strongly dependent on the depth of encoding, which seems to have an effect on the activity of the hippocampus and left prefrontal cortex.

B. Episodic Retrieval

Retrieval of episodic memory is mediated by regions that generally are functionally and anatomically distinct from those used in encoding. Most neuroimaging studies of retrieval use a task design similar to that used in studies of encoding, in which participants must study a set of items and are subsequently tested for their memory of the items. The difference is that participants are scanned while they retrieve (recall or recognize) rather than while they encode the material. In recognition tasks, an item is shown, and it is the task of the participant to indicate whether that item was presented during the study phase. Hence, it is necessary only to access the source and not the item. In recall tests, it is necessary to generate the item as well. Neuroimaging studies of both recall and recognition typically show activity in the right prefrontal cortex, hippocampus, medial as well as inferior parietal cortex, anterior cingulate, and cerebellum. There are some important variations in this pattern, however, that are discussed next.

The hippocampus is typically considered to be involved in the consolidation of long-term memories, as discussed earlier. Although this function implies that the hippocampus should not be involved in retrieval, some studies *have* found it to be activated during retrieval tasks. To test whether the effort required for retrieval might influence activation of the hippocampus, one study varied the amount of effort required to search memory. In a ''high-recall'' condition, words were deeply encoded and, hence, less effortfully retrieved. When the recall phase was scanned, this manipulation revealed activity in the hippocampus bilaterally, supporting the view that the hippocampus is involved in effortless, conscious recall. In a ''low-recall'' condition, words were encoded more superficially and, hence, required more effortful retrieval. Scanning during this more effortful recall phase showed bilateral prefrontal but not hippocampal activation. The finding that the prefrontal cortex is involved in effortful retrieval is consistent with the view that one function of the prefrontal cortex is to implement retrieval strategies. The hippocampus, by contrast, may be involved in relatively more automatic retrieval.

Certain neuroimaging studies of episodic retrieval have found not only increased activity in the right prefrontal cortex but also decreased activity in other areas such as the left prefrontal cortex. On the basis of this effect, some have suggested that episodic retrieval is not just an active process of search and retrieval but involves the active inhibition of certain regions of the brain by other areas of the brain. By this model, the right prefrontal cortex could be actively inhibiting left frontal regions as well as inferior temporal regions, areas that sometimes show deactivations in retrieval tasks. In the case of the temporal regions, for example, this might indicate the suppression of language processes during episodic retrieval. This effect has been termed ''ensemble inhibition'' and suggests that episodic retrieval may be carried out, in part, by inhibitory processes.

Retrieval processing involves an interplay between the right prefrontal cortex and the hippocampus in the implementation of search strategies and conscious, effortless retrieval, respectively. The involvement of other areas of the brain such as the precuneus, parietal cortex, anterior cingulate, and cerebellum has yet to be fully elucidated, so much further research is required on this problem.

C. Synthesis: The HERA Model and Its Extension

Stable differences appear to exist in the activations accompanying encoding versus retrieval. The most striking pattern is in the activity of the prefrontal cortex. Most studies of encoding have shown activity in the left prefrontal cortex at a more anterior site, whereas most studies of retrieval have shown activity in the right prefrontal cortex, also at a more anterior site. This hemispheric difference in prefrontal activity in episodic memory is typically referred to as the Hemispheric Encoding–Retrieval Asymmetry model or HERA.

Other areas of the brain have also come to be included in the HERA model. For example, the left cerebellum seems to be more active during retrieval than during encoding. The cerebellum's anatomical connections are predominantly with the contralateral prefrontal cortex (via the thalamus), so that the coupling of right prefrontal and left cerebellar activations is not surprising. What function the cerebellum might be serving in the context of episodic retrieval is unclear. The cerebellum has long been associated with motor coordination and visuomotor skill learning. It is possible that the cerebellum is serving one of these general roles in effortful retrieval, but its exact role remains to be elucidated.

The association cortices have also gained some attention in regard to the HERA model. The left temporal cortex has been found to be activated in some

studies of encoding, whereas activation in the right or bilateral parietal cortex has been documented during retrieval. These findings, though not entirely uncontroversial, seem to follow the HERA pattern, with encoding being a left hemisphere function, in this case in the temporal lobe, and retrieval being a right hemisphere function, in the parietal cortex. There is a great deal of speculation as to exactly what these areas are doing. Some accounts claim that they are involved in some way in the execution of special encoding or retrieval strategies. In the case of the temporal cortex, this might be attaching some kind of mnemonic code to items to ease retrieval later on. Others suggest that the parietal cortex is involved in perceptual aspects of retrieval, such as mental imagery. Further study is necessary to fully understand the functions subserved by these regions, as well as the way that they interact with the other areas of the brain that are active during episodic encoding and retrieval.

It must be noted that there are important exceptions to the HERA model's general description of the patterns of brain activity during tasks of episodic memory. The model does not do well in predicting patterns of activation in the hippocampus. Hippocampal activation has been found unilaterally on the right and left and bilaterally in tasks of both encoding and retrieval. The pattern of activity in the hippocampus is best described as being dependent on the type of material being encoded or retrieved, not on encoding and retrieval by themselves. This is shown by systematic patterns of activation depending on whether verbal or nonverbal stimuli are used in an experiment. Most experiments using verbal information have shown predominantly left hippocampal activity. In contrast, the right hippocampus or both hippocampi may be more active in encoding visual information. For example, a study in which people retrieved information about a spatial route through a town activated bilateral hippocampi. It should be noted as well that the material specificity of activations extends to the prefrontal cortex. One area of prefrontal cortex often observed in studies of episodic memory does not follow HERA but rather depends on whether the stimulus material is verbal or visuospatial, the former producing activation on the left and the latter on the right. So, even within the prefrontal cortex, one region obeys the description given by the HERA model and another does not.

Overall, it does appear that many patterns of activity demonstrated in tasks of episodic memory follow an asymmetric hemispheric pattern in regard to encoding and retrieval. In the prefrontal cortex and in temporal and parietal association areas, activity in the left hemisphere is associated with encoding processing and activity in the right hemisphere is associated with retrieval processing. The cerebellum also shows a hemispheric asymmetry, but it is the reverse pattern with the left cerebellum engaged during retrieval. There are exceptions to this pattern, and these are seen in the hippocampus, posterior regions, and some anterior regions of prefrontal cortex as well. The patterns of activity in these regions are dependent on the modality of the information being processed—verbal information lateralized to the left hemisphere and visuospatial lateralized to the right—rather than encoding and retrieval processes.

D. False Memories

Memories of our personal experiences are extremely vulnerable because memory often is not *reconstructive* but *constructive*. When retrieving an episodic memory, we try to reproduce the event as closely as possible, constructing the most plausible approximation. Consequently, we often incorporate aspects of the event that are close to the original but not exactly it, and we can even insert information that never occurred at all. This vulnerability can even go so far as to produce elaborate situations that never actually happened, though the person might swear that they did.

Neuroimaging studies have begun to examine differences in brain activation, comparing retrieval of a true past event from false memory for an event that did not occur. Most of these studies have used a task in which participants are shown lists of words to study. After a long retention period, the participants are asked to perform a recognition task, indicating which words on a second list were present on the first list. Some of the words on the second list that were not present on the first list, called foils, are semantically related to the words on the first list. For example, the words "pajama, bed, night" might have appeared on the first list, but the word "sleep" might appear at the time of the recognition test. Semantically related foils often were falsely recognized as having appeared on the originally studied list. Furthermore, participants often rated their confidence that the words had been on the original list as highly as they rated their confidence that actual words had appeared. Thus, it appeared as if the participants had created false memories of semantically related foil words.

Neuroimaging studies have compared activations due to falsely recognized words to activations due to

correctly recognized words. One difference that emerges is activation in the frontal cortex during true recognition. This is consistent with other retrieval studies, as reviewed earlier. Another feature of activation is that words on the recognition test that had actually been presented sometimes caused activation in the primary sensory cortex of the modality in which they had appeared. For example, a word presented aurally, when tested at the time of recognition, might show activation in the superior temporal cortex. By contrast, words that did not appear showed no such sensory cortex activation. Thus, a neural signature apparently exists that permits one to distinguish actually presented words from semantically related foils, even if one does not access this signature in the recognition judgments.

IV. SEMANTIC MEMORY

Episodic memory is distinguished by the fact that it requires not only the retrieval of an item from memory but also a source or context for that item. But many times, we retrieve a fact with no knowledge of its context, as when we can identify various types of bicycles without knowing when and where we learned about the various types. Semantic memory can be defined as memory for facts about the world, naked of their source context. This kind of knowledge plays a critical role in all forms of cognition, from language to reasoning to problem-solving. Hence, semantic memory is an important topic for study. Most studies of semantic memory have focused on the retrieval of semantic information from memory because this is studied most readily and because it is more difficult, given the normal course of learning, to study encoding or storage of semantic memory.

A. Verbal Semantic Memory

Many of the concepts that make up our semantic knowledge are coded in the form of language, probably because we are such intensely linguistic creatures. These concepts come from various categories, of course, such as living things and nonliving things, distinctions that we can readily make for many concepts. Evidence from patients with focal brain injury shows that the brain seems to honor some of these categorical distinctions among concepts. For example, there are patients who appear to have lost their ability to identify living things, such an elephant or a flower, even though they are still capable of identifying nonliving things, such as tools. One interpretation of this result is that the brain's organization of semantic memory is, in part, organized by broad categories. To test whether this is so in normal adults as well as brain-injured adults, one study used PET to examine what areas of the brain are active during the retrieval of three different semantic categories. Participants were given several scans during each of which they were asked to name photographs of either famous people, animals, or tools. As predicted, different brain activations resulted from naming each kind of stimulus. Naming famous people produced activity in the most anterior part of the temporal cortex, called the temporal pole. Naming animals produced activity in a more posterior area of the temporal cortex in inferior and middle temporal gyri. Naming tools showed activity in an even more posterior portion of the inferior temporal gyrus. In general, the more specific the item that had to be named, the more anterior the activation in the temporal cortex. These findings are consistent with the notion of a visual processing stream that spreads from occipital cortex into temporal cortex, moving from general classification to more specific categorization in the anterior temporal cortex.

A related PET study of object and face naming provides corroborating evidence for these findings and further distinguishes between activations involved in the identification of specific faces versus the simple recognition of stimuli as faces. In one condition, participants were asked to make gender discrimination judgments of familiar and unfamiliar faces. Relative to a control condition, this task produced activations broadly in the extrastriate occipital cortex. Only when participants identified faces of famous people, and so had to make a specific face identification, were the temporal poles activated. Also activated were other structures in the temporal cortex (fusiform gyri, right lingual and parahippocampal gyri, and left middle temporal gyrus) as well as the orbitofrontal cortex.

Object naming in this study shared some common areas of activation with face naming, including the orbitofrontal cortex, left middle temporal gyrus, and left fusiform gyrus. Converging evidence from animal and lesion studies may shed light on the roles of these regions. Neuropsychological data suggest that the left middle temporal gyrus may be necessary for naming, but not recognizing, objects and faces. Lesions of orbitofrontal cortex have been related to visual memory impairments in animals. Finally, the fusiform

gyri appear to be involved in the recall of both faces and objects, with face recognition activating the right gyrus and object recognition the left. Interestingly, the fusiform gyrus has also been implicated in the *perception* of faces and objects, so the region responsible for semantic memory for this type of information may be similar to the region used to perceive it. This hypothesis is supported by intracortical electrode studies done on patients in preparation for possible surgery to treat epilepsy. These recordings showed that face recognition elicited electrical activity in the fusiform gyrus and that electrical stimulation in this same region resulted in an inability of the patients to name a face for the duration of stimulation.

These complementary results suggest that face perception and recognition share a common substrate and that the boundary between perception and semantic memory may be indistinct for this type of material. Studies of object naming more generally show that semantic memory about concrete objects appears to be organized, at least to some degree, in cortical modules devoted to particular types of remembered material, and these become more specific moving from posterior to anterior brain regions.

Semantic memory for words must take on more than just a simple recognition and naming function. Indeed, when faced with an object it is often more useful to know what can be done with that object rather than just its name. This type of information, as well as information about constructs that are not concrete, is within the domain of semantic memory and has been directly studied.

To examine this sort of semantic memory, a PET experiment was designed that required participants to generate an associated verb for each word in a list of nouns. For example, when shown a picture of an apple, a participant might respond "eat." This experiment revealed a preferential role of lateral and inferior frontal cortex in the generation of verbs associated with visually presented objects. It has been suggested that verbs are at the core of semantic structure, and hence activation in frontal cortical areas might be an indication of which areas are critical in mediating the generation of semantic concepts. Further study of this verb-generation situation compared a task in which participants had to name a verb for each noun presented to two control conditions—reading words and passively viewing words. Multiple subtraction conditions were used to identify areas related to the motor execution of speech (motor cortex), word reading (left insula), and verb generation (left frontal cortex, anterior cingulate, and right cerebellum). This study also revealed changes in activation with practice on this task, as reviewed later. The constellation of regions that were activated in this study probably included a complex combination of areas involved in attention, inferential reasoning, willed action, episodic encoding, and working memory, but most prominently, significant activation occurred in the left prefrontal cortex and elsewhere that could be attributed to semantic retrieval.

V. COGNITIVE SKILL LEARNING

Investigation of practice-related changes in the verb-generation task provides a convenient segue into a discussion of skill acquisition—another vital aspect of memory. After 15 min of practice on the verb-generation task, 90% of the verbs that participants generated were rote responses that had been consistently associated with the nouns. Participants no longer had to search through memory for a novel association; instead, they could quickly recall a response they gave previously. Analysis of the difference between the PET images early and late in the verb-generation task showed that, with practice, activity decreased in the anterior cingulate, left prefrontal cortex, bilateral inferior frontal gyri, left temporal cortex, and right cerebellum—the very areas that were active when verb generation was compared to word reading. In fact, after practice the PET images were indistinguishable from those of word reading. Increases in activation with practice were observed in the precuneus and cuneus on the medial wall of the posterior cerebrum and in the right superior parietal cortex. These shifts in activation could be due to decreased demand on attention and effort, decreased searching of semantic memory, or some other factor.

Without further evidence, it is hard to distinguish among these causes. Certainly evidence exists that some of these areas are involved in other cognitive processes. For example, the left prefrontal cortex is activated in verbal working memory tasks and in the encoding of long-term memories, as we reviewed earlier. The anterior cingulate is also activated in some overpracticed motor tasks, particularly when they might require attending, making inferences about a pattern, or anticipating future stimuli or feedback. One region that may have a clearer role in the practice effect seen in the verb-generation task is the insula, an area located anatomically near the region responsible for speech output and an area that showed increased activation in the verb-generation task with practice.

This activation may be an indication of increasing automaticity in producing verb associations given nouns as stimuli, a kind of stimulus–response connection that developed even over the course of relatively little practice.

Of course, one might ask whether skill in the verb-generation task is a good example of cognitive skill learning in general. The shifts in activation in this task seem to result not from an improved ability to generate *novel* verbs, but rather from the ability to call up from memory the same verb the subject gave on the last trial, a kind of automatic stimulus–response mapping. It is not yet understood how true cognitive skills, such as the ability to make inferences and manipulate abstract concepts, are learned, except that their appearance in children seems to parallel development of the frontal lobes. However, a great deal of what we normally consider to be cognitive skills, such as expertise in chess, can be explained as the formation and retrieval of ever larger and more complex sets of associations in semantic memory.

A final example of cognitive skill learning comes from a PET study of categorization. Experimental studies have shown that people learn to categorize objects in several ways: through the application of rules, learning of specific exemplars of a category, and implicit learning of an average or "prototype" of the category. Patients with medial temporal damage, who have virtually no remaining episodic memory, fail on rule- and exemplar-based categorization but learn prototype-based categorization as readily as do normal participants. In the study, people classified pictures of contrived animals based on previous practice with similar (but not identical) animals. One practice group learned to categorize animals by a rule, and the other group learned the categories by trial and error. The rule group categorized the new animals presented during scanning by applying the rule. The trial-and-error group categorized animals during PET scanning on the basis of their similarity to the animals they saw during training, an exemplar-based strategy. Only rule-based categorization activated the bilateral parietal cortex, right prefrontal cortex, and bilateral supplementary cortex, possibly reflecting greater working memory demands, attention shifting, and retrieval and application of the rule. Exemplar-based categorization activated the left extrastriate cortex and left cerebellum. The extrastriate activation may reflect greater use of a perceptually based memory trace in the exemplar-based strategy, consistent with the involvement of extrastriate visual cortex in studies of implicit memory. This study suggests that different learning regimens may have a profoundly different effect on the brain circuitry recruited to the task, at least for a task requiring categorization processes.

VI. LEARNING PROCEDURAL SKILLS

Procedural knowledge, as described earlier, consists of knowledge of how to do something. It includes all the behaviors shown as examples of procedural knowledge in Fig. 1, common to all of which is the fact that they do not require explicit retrieval of information; rather, they require the person to tap memory in the service of some other task, such as riding a bike taps memory to coordinate the muscles in various ways so that balance can be maintained. This is often called "implicit" memory. Viewed this way, it is clear that a vast number of motor skills are mediated by procedural memory. Of course, procedural memory must develop over the course of practice, and the changes that occur with practice are often highly specific, improving performance on precisely those tasks we practice. Although this may seem to limit our behavioral repertoire, in fact, once we have learned a sequence of motor movements, we can quickly learn to adapt these movements to similar situations.

Skill learning occurs when a new movement or sequence of movements is acquired, and performance becomes both faster and more accurate with training. Sometimes this learning might involve the acquisition of new sensory–motor mappings. Studies of motor skill learning have focused on either the acquisition of some new sequence of motor movements, such as a sequence of finger taps, or the acquisition of some sensory–motor mapping, such as guiding a visual cursor with joystick movements.

A note: it is often difficult to separate the brain areas involved in learning a skill from those involved in other aspects of processing that change along with skill learning. As people learn a skill, for example, they devote less attention to the task, so that neural circuitry responsible for focusing attention is less involved. At the same time, the incidence of errors decreases, so the brain activity related to error detection and correction will decrease. The brain regions responsible for mediating task performance may also change as skills become automatic. All of these changes obscure the interpretation of neuroimaging studies of the learning process by making it difficult to determine which activations are due to skill learning per se and which are due to other processes.

A. Motor Skill Learning

Execution of motor tasks such as tapping fingers in a particular sequence or maintaining contact between a target moving in a circular pattern and a hand-held marker activates large regions of cerebral cortex and subcortical structures, including sensory motor cortex (primarily the primary motor cortex), supplementary motor area (SMA), premotor area (PMA), putamen, cerebellum, and sensorimotor thalamus (see Fig. 2). Traditionally, and consistent with data on the anatomical layout of the motor system, neuroimaging studies have found activations in these areas on the side contralateral to hand movement (with ipsilateral activation of the cerebellum because of its crossed connections to the cortex). However, evidence indicates that complex movement of even one hand can activate motor areas bilaterally. For example, in one experiment, participants rotated two metal balls at a constant rate in either their right or left hand. These complex movements activated the regions mentioned previously, including significant bilateral activations in the postcentral gyrus, traditionally considered primary somatosensory cortex, and intraparietal sulcus. Notably missing from this list of activations is the basal ganglia, which appear to be preferentially activated during the performance of learned sequences of movements.

Areas involved in the learning of motor skills are largely the same areas as those used during task performance. However, some regions become active only during initial 1earning or only after performance has become automatic and requires little effort. These changes are specific to the type of task—they vary depending on whether automatic performance involves increased or decreased use of sensory cues and whether the task involves learning a new movement, a sequence of movements, or a sensory–motor association.

B. Sequence Learning

Among the first motor learning tasks studied is the sequential tapping task, in which participants must tap the fingers of their dominant hand (all studies reviewed here used right-handed subjects) in a prespecified sequence. An early study scanned participants doing sequential finger tapping at three stages in the learning process. The three levels corresponded to an initial learning phase when a skill is first being learned, the phase when a skill becomes automatic after significant practice, and skilled performance after performance level has reached its asymptote. The ipsilateral (right) cerebellum was activated in all three conditions, and the activation per movement in the cerebellum decreased as training progressed. The striatum was also activated during advanced practice, suggesting a role for the basal ganglia in the development of automaticity.

Subsequent studies of sequential finger tapping have examined the role of the cerebellum and other structures, including changes in primary motor cortex, in more depth. Unpracticed performance activated a network of areas often associated with the planning and execution of movements: contralateral primary motor cortex and putamen, bilateral PMA and SMA (with more activity on the contralateral side), and cerebellum. With practice, activity decreased in the lateral portion and deep nuclei of the cerebellum, supporting the view that the cerebellum is important in sequence learning and may be less important in the execution of highly learned sequences.

Subsequent studies examined activations during a sequential finger-tapping task in which participants had to learn the correct sequence by trial and error. The first study compared activations between new sequences, learned sequences, and a resting control. Similar to the earlier studies, the performance of learned sequences with some degree of automaticity activated the contralateral (left) primary motor cortex, PMA, SMA, putamen, bilateral cerebellar hemispheres, vermis, deep cerebellar nuclei, anterior cingulate, parietal cortex, and ventrolateral thalamus. Of course, because sequence performance was compared with a resting condition, this network includes areas responsible for processes irrelevant to sequence learning. When compared to rest, the performance of new sequences showed, in addition to these areas, increases in prefrontal cortex and more extensive activation of the cerebellum. When compared to the practiced sequence directly, learning of a new sequence produced activations in the bilateral PMA, cerebellum, anterior cingulate, prefrontal cortex, and medial thalamus. It seems likely that the requirement that the participant infer the correct sequence on the basis of feedback was responsible for the activation of the anatomically interconnected prefrontal–anterior cingulate–medial thalamus network. However, as we will see later, it is possible that the cerebellum also contributes to error detection and correction.

In another revealing manipulation, participants were asked to pay attention to their finger movements while performing a highly learned sequence. Attention

resulted in the reactivation of the anterior cingulate and prefrontal cortex. By comparing learning of a new sequence with a control condition that required a similar level of attention and similar decision and motor processes, the researchers found activation of the caudate nucleus and cerebellum. This result indicates that these two structures may be important in learning a new sequence, as opposed to other task-related processes. Together, these studies suggest that the basal ganglia and cerebellum, and possibly the PMA, are involved in skill learning, whereas anterior cingulate and prefrontal areas are involved in attention and higher level control processes.

A more controlled version of the tapping task is the Repeated Sequence Task. In this task, participants see a cue appear above one of four squares, and they must touch that square as quickly as possible. As a particular sequence appears more frequently, participants become faster in pressing the appropriate squares. The faster responses for the learned sequence indicate that implicit learning of the sequence has occurred, even though participants have no explicit, declarative memory for the sequence.

Doyon and co-workers, using this paradigm, compared PET activation among several conditions, including different amounts of learning, and included a condition in which subjects were given explicit knowledge of the sequence prior to scanning. Performance on highly learned versus random sequences resulted in increased activation in the right (ipsilateral) ventral striatum, right cerebellum, bilateral anterior cingulate, right medial parietal cortex, and right extrastriate cortex. Decreases were found in the ventrolateral frontal, frontopolar, and lateral parietal cortices, all on the right side. As with previous studies, basal ganglia activity increased when performance was highly learned. These changes are consistent with animal models suggesting a role for the basal ganglia in the performance of movement sequences.

When compared with newly learned sequences, highly learned sequences showed increased activity in the cerebellum, suggesting a role for the cerebellum in sequence performance or in the development of automaticity. This finding contrasts with previous sequence learning studies, which found that cerebellar activity decreases as performance becomes automatic. Thus, further research is necessary to investigate the role of the cerebellum. It is possible that the cerebellum has multiple roles and plays a part both in initial learning and in later retrieval of sequences. An alternative explanation for cerebellar activity in the learned–unlearned comparisons in these studies is that the cerebellum is required for speeding up movements while maintaining a criterion level of accuracy. In the early stages of training, it may be difficult to keep movements speedy, resulting in cerebellar activation that decreases as it becomes easier to perform the task at the requisite speed. This hypothesis highlights the fact that neuroimaging results may be interpreted in multiple ways.

Newly and highly learned sequences in the Repeated Sequence Test were also examined before and after participants were given explicit knowledge of the sequence. Explicit knowledge of highly learned sequences relative to implicit performance decreased activation in the right (ipsilateral) cerebellum and increased activation in the right ventrolateral frontal cortex. Explicit versus implicit knowledge of newly learned sequences resulted in increased activation of the right cerebellum and left ventrolateral frontal cortex.

The sequence learning studies reviewed here implicate the cerebellum and basal ganglia in initial learning and automatization of new skills. Three studies found significant activity in the basal ganglia with practiced, but not novel, sequences. Two other studies found basal ganglia activity in newly learned sequences, one relative to a resting control and the other relative to a free-selection tapping task. The cerebellum appears to be involved in early learning of the sequence, but it is also active during skilled performance relative to rest. Its activity may increase as participants gain skill, if that skill involves making speeded responses to visual cues.

Notably, none of these PET studies showed changes in the primary motor cortex as a function of practice. However, some researchers have shown that skill learning produces nonmonotonic changes in the strength and spread of activation within the primary motor cortex. These researchers scanned participants learning one of two similar finger-tapping sequences using fMRI. Initially, a habituation effect to sequence performance was found: activity in the primary motor cortex was lower when it was performed later in the scanning block. After 30 min of practice, activity in the primary motor cortex was higher when it was performed later in the scanning block, but only for practiced sequences, possibly reflecting fast learning processes that set up later consolidation of a new motor sequence. After 4 weeks of daily practice, more areas in the primary motor cortex were recruited during performance of the learned sequence, showing that practice resulted in a true expansion of the cortical area recruited in the primary motor cortex.

C. Learning Sensory–Motor Associations

What are strikingly missing from analyses of sequence learning, but apparent in other motor learning tasks, are changes in activation of the premotor and supplementary motor cortices. Studies of sequential finger tapping primarily require the learning of associations between movements that form a sequence. This kind of learning may be fundamentally different from learning that requires improvement in coordination or learning an altogether new motor movement. In addition, the finger-tapping task requires a series of internally generated movements rather than movements elicited by sensory cues. We have already seen that learning in the Repeated Sequence Test, which involves visuomotor associative learning, may produce a different pattern of cerebellar activity than learning of internally generated movement sequences. Perhaps examination of nonsequence motor learning, including studies of rotor pursuit, trajectory movements, joystick movements, and maze tracing, may help resolve discrepancies.

Rotor pursuit involves maintaining contact between a hand-held stylus and a target moving in a circular pattern. Several studies have compared rotor pursuit with visual tracking of the stylus, identifying changes due to skill learning of the appropriate hand movement. An initial study showed increases with practice in the primary motor cortex, SMA, and pulvinar nucleus. In a later study, PET scans over 2 days of learning revealed a similar pattern of activation on day 1, this time including practice-related changes in the ipsilateral cerebellum, cingulate, and inferior parietal cortex. On day 2, after an extensive practice period, the rotor pursuit–visual tracking subtraction revealed changes in the putamen and parietal cortex bilaterally and the left (contralateral) PMA. Improvement on the task was correlated with increased activity in the premotor, prefrontal, and cingulate areas and decreased activity in visual processing areas.

An opposite pattern of changes in the PMA and SMA were observed in a series of studies of maze tracing. In the maze-tracing task, participants practiced tracing cutout maze patterns with their eyes closed. Tracing a novel maze with the right hand in a clockwise pattern (relative to a control) resulted in increases in the right PMA and left cerebellum. Practice diminished both of these loci of activity: PET scanning after 10 min of practice produced increases in SMA and decreases in the right PMA and left cerebellum. Contrary to expectations, training affected activity in the ipsilateral cerebrum and contralateral cerebellum. A follow-up study examined maze tracing with the left hand in a counter-clockwise pattern and, strikingly, produced the same results. A third study confirmed that only the primary motor cortex and anterior cerebellum showed activation depending on which hand was used, suggesting that learning and performance in the maze-tracing task require an abstract representation of movements and patterns not directly tied to motor activity.

Contradictory effects of practice in similar tasks, such as that found in PMA and SMA in the studies reviewed earlier, may reflect something of the underlying functions of premotor brain regions. It has been suggested that the SMA is active in the internal generation of responses, and PMA is preferentially active when responses are contingent upon sensory cues. That distinction can be applied to neuroimaging studies of human skill learning. In the rotor pursuit studies, learning of a new motor skill resulted in increased activity in SMA. After the movement was automatic, PMA and basal ganglia were activated in adapting the movement to the motion of the stylus. In the maze studies, initial performance relied heavily on somatosensory feedback from the cutout maze, and so the PMA was heavily recruited. After the maze was well-learned, a coordinated, internally generated movement could be used to trace the maze successfully without reliance on sensory feedback. This type of internally generated movement appears to be the domain of the SMA. Consistent with this view, neuropsychological evidence shows that patients with PMA damage are impaired in sensory-cued motor learning.

A study in which participants had to write the letter "R" found increases in the right PMA and right parietal cortex, possibly indicating enhanced attention to sensory feedback. Again, the activation was unexpectedly ipsilateral to the hand used. When participants wrote the letter as quickly as possible, a network including the left primary motor, PMA, SMA, and right putamen was activated. These experiments also compared writing novel ideograms with the right hand to a baseline in which participants watched the figures being drawn. Novel ideograms, as compared to baseline, activated the left primary motor cortex and right cerebellum. Practiced ideograms minus baseline activated the right PMA, left SMA, and right cerebellum. Once again, cerebellar activity appeared during the initial learning and automatization phases of new visuomotor patterns, but was not apparent when writing the letter "R," a highly practiced ideogram.

These studies show increases in PMA when sensory feedback or external cues are critical for task performance. Right PMA increases appear to be modulated by task demands, e.g., asking participants to be accurate, and independent of the hand used. In the study of unilateral two-ball rotation mentioned earlier, positive correlations were found between unilateral PMA activity during movements of either hand and skill improvement for the ipsilateral hand. Skill improvement was measured during nonscanning intervals by comparing the maximum rotation speed of the balls just after the first scan and just before the last scan. The strongest correlation was between right PMA activity and skill improvement for the right hand. Correlations between left PMA and skill improvement for the left hand were also positive and significant.

The cerebellum appears to play a particular role in the learning of new visuomotor associations. In a classic study, participants made joystick movements to align a cursor with a visual target in one of three joystick–cursor mappings: normal mapping, reversed mapping in which joystick movements caused the cursor to move in the opposite direction, and random mapping with no relationship between joystick and cursor movement. The PET camera in this study was centered on the cerebellum to allow maximum sensitivity to changes in this structure. When the mapping between joystick movements and cursor movements was reversed, cerebellar activity was high during initial learning and decreased as performance improved. When the mapping between joystick movements and cursor movements was random, no learning occurred and cerebellar activity remained high. The authors concluded that the cerebellum contributes to visuomotor skill learning by participating in the detection and correction of errors. It should be noted, in closing, that the cerebellar contribution to skill learning may be primarily in sensory–motor association tasks. Patients with cerebellar atrophy can improve on some motor skill tasks, but they fail to coordinate and adapt movements to environmental contexts.

D. Learning Perceptual Skills

Research on the psychophysics of perceptual learning suggests that long-term learning of skills is not limited to motor areas. This research suggests three striking findings: learning to detect a visual stimulus is specific to a retinal location, it only occurs if the stimulus is behaviorally relevant, and effects of learning remain robust after nearly 3 years without practice.

Although there has been relatively little neuroimaging research on perceptual learning, the research that has been conducted is illuminating. Perhaps the first study was an fMRI experiment of mirror reading. Participants viewed words and matched nonwords printed backward and decided whether the stimulus was a real word. Improvement in mirror reading is still detectable after 2 months of practice, and changes in performance are evident after 1 year without practice. The researchers compared mirror reading with normal reading and practiced, mirror words with unpracticed ones. Mirror reading compared to normal reading produced increases in blood flow in a number of areas, including the medial and lateral occipital cortex, right superior parietal cortex, bilateral fusiform gyrus, pulvinar, and cerebellum, particularly on the right side. Decreases were found in nearby regions of some of the same areas, including the medial occipital and right superior parietal cortices, as well as in the precuneus and bilateral middle–superior temporal gyrus. The occipital, fusiform, and pulvinar areas form an anatomically interconnected network associated with visual processing, object recognition, and visual attention. The authors suggested that these activations were due to visual transformation of the letters in unpracticed mirror reading.

With practice, mirror reading activation increased in the precuneus, left superior parietal cortex and fusiform, and right cerebellum and superior temporal gyrus. Decreases were found in a number of areas, including left lingual gyrus, bilateral occipital cortex, right cerebellum, superior and inferior parietal, inferior temporal, and pulvinar. The deactivation of right superior parietal cortex and occipital cortex could reflect decreased involvement of attention, which is associated with right parietal activation in a number of studies, or it could result from a decreased need to rely on visuospatial transformations. Increases in the left fusiform gyrus, activated in at least one other letter recognition study and in studies of object recognition, could reflect a shift to direct letter recognition processes. Overall, the results are consistent with a shift in strategy from sensory transformation to direct recognition of mirror words.

E. Implicit Memory

All of the studies of skill learning discussed earlier involved both explicit and implicit memory for motor

activity. Although procedural in nature, in most of the tasks participants were allowed to develop conscious recognition of the sequences and movements to be learned. In one study, explicit awareness of a sequence was associated with changes in the cerebellum, ventrolateral frontal cortex, and medial frontal cortex. The remainder of the changes with skill learning, found prominently in parts of the cerebellum, primary motor cortex, SMA, PMA, and basal ganglia, presumably reflect changes in procedural motor processes: the kind you can do, but cannot describe—and, if you are amnesiac, may not remember that you ever learned.

Implicit memory is not limited to motor and perceptual learning, however. Exposure to words, pictures, faces, and other stimuli influences later processing of this information, even though subjects may not remember any of the trained items. This unconscious memory is independent of explicit recollection, meaning that the amount of explicit memory one has for a stimulus does not predict the amount of implicit memory. Also, implicit memory is not affected by depth of encoding as is explicit memory. The two forms of memory appear to be separate species that operate independently of one another. Neuroimaging results support the conclusion that explicit memory and implicit memory are mediated by separate circuits in the brain—explicit memory by the hippocampus, prefrontal cortex, and related circuitry, and implicit memory by perceptual areas similar to those discussed in the previous section.

One of the most common ways to study implicit retrieval is the word stem completion task. In the encoding phase of the task, participants are given a list of words and asked to make some judgment, semantic or otherwise, about the words. Subsequently, the participants are given a list of incomplete words, such as the first three letters, and asked to complete the fragments with whatever word comes to mind. If the fragments match words that they viewed before, participants are more likely to complete the stem with the previously viewed word, whether or not they consciously remember it.

Daniel Schacter and co-workers predicted that, when only implicit memory is involved in a task, stem completion will not activate the hippocampus, the structure critical to explicit episodic memory. They asked participants to count the number of "T" junctions in a list of words and compared activation during stem completion of words from the list with novel words. Participants did not explicitly remember the words, but their performance showed priming effects. Their PET scans revealed no hippocampal activation, but did show decreases in extrastriate (visual) cortex. In another study, Schacter *et al.* compared stem completion when participants had encoded words deeply (when they were required to make semantic judgments) or superficially (when they counted T junctions). Explicit memory and hippocampal activation during stem completion were greater for the deeply processed words. Other studies have replicated this basic finding with object naming and categorization tasks. Taken together, these results indicate a dissociation between explicit and implicit memory; explicit memory involves the activation of hippocampal structures, whereas implicit memory involves the deactivation of posterior neocortex.

Studies in other domains of semantic knowledge, such as object categorization and object naming, show similar priming-related reductions in activation. One explanation of the decrease in activation is that repeated presentations lead to more efficient perceptual processing and lower levels of activation. However, in an auditory word stem completion task studied by Schacter and colleagues, auditory presentation of words and auditory stem completion produced the same regional cerebral blood flow (rCBF) decreases in bilateral extrastriate cortex, as well as decreases in the right anterior medial prefrontal cortex, right angular gyrus, and precuneus. Thus, priming effects on the extrastriate cortex appear to be modality-independent.

VII. SOME CONCLUDING REMARKS

There has been remarkable progress in the study of memory during the last half of the twentieth century. The development of neuroimaging techniques has contributed in no small part to this progress. These techniques have provided a source of evidence about the relationship between brain structure and psychological function that complements evidence from the study of human patients with brain injury, behavioral evidence from normal human subjects, and evidence from animal models. Taken together, these sources of evidence have sketched the outlines of complex memory systems. These systems, in their interaction, form a seamless whole capable of dealing with a variety of cognitive problems. Our memory apparatus has multiple components (working memory, long-term memory), multiple representations of information in different formats (e.g., verbal versus spatial), multiple retrieval schemes (explicit versus implicit), multiple circuitries, and multiple processes (encoding, retention, and retrieval). We are far from getting our arms

around a complete description of this system, but continued use of the full array of investigative tools promises much new progress in the current millenium.

See Also the Following Articles

INFORMATION PROCESSING • MEMORY DISORDERS, ORGANIC • MEMORY, EXPLICIT AND IMPLICIT • MEMORY NEUROBIOLOGY • MEMORY, OVERVIEW • NERVE CELLS AND MEMORY • NEUROIMAGING • SEMANTIC MEMORY • SHORT-TERM MEMORY • WORKING MEMORY

Acknowledgment

Preparation of this manuscript was supported in part by a grant from the National Institute of Mental Health to the University of Michigan.

Suggested Reading

Anderson, J. R. (1995). *Learning and Memory: An Integrated Approach*. Wiley, New York.

Baddeley, A. (1990). *Human Memory Theory and Practice*. Allyn and Bacon, Boston.

Campbell, R., and Conway, M. A. (1995). *Broken Memories*. Blackwell, Cambridge.

Gordon, B. (1995). *Memory: Remembering and Forgetting in Everyday Life*. Mastermedia, New York.

Reisberg, D. (1997). *Cognition*. W. W. Norton, New York.

Schacter, D. (1996). *Searching for Memory: The Brain, the Mind, and the Past*. Basic Books, New York.

Schacter, D. (2001). *The Seven Sins of Memory: How the Mind Remembers and Forgets*. Houghton Mifflin, Boston.

Squire, L. R., and Kandel, E. R. (2000). *Memory*. Scientific American Library, New York.

Vallar, G., and Shallice, T. (1990). *Neuropsychological Impairments of Short-Term Memory*. Cambridge University Press, UK.

Memory, Overview

JOHN A. LUCAS
Mayo Clinic

GLOSSARY

Alzheimer's disease A gradually progressive degenerative dementia of insidious onset that is characterized by early, prominent memory dysfunction and loss of the ability to perform activities of daily living.

amnesia Loss of memory.

dementia A behavioral description of any acquired disorder that involves amnesia plus impairment of at least one other cognitive ability, such as language, visuospatial function, attention, or problem solving. The degree of cognitive impairment must be severe enough to interfere with social functioning, occupational functioning, or activities of daily living.

Korsakoff syndrome An amnesic disorder resulting from severe thiamine deficiency, most often seen in patients with chronic alcoholism.

neocortex In evolutionary terms, the newest, most highly evolved region of the brain, containing six neuronal layers and forming the outer surface of the brain.

Memory is the ability to maintain previously learned information within an internal storage system so that it may be accessed and used at a later time. Memory may be observed in overt behavior, such as when a student recalls information from previous readings in order to answer questions on a test, or in less readily observable events, such as neuronal development and change. This article reviews the behavioral components of memory as well as the neurobiological bases and underlying neuroanatomical substrates believed to be important to memory functioning.

I. CLASSIFICATION OF MEMORY

Memory is not a unitary construct but instead reflects a number of distinct cognitive abilities that can be categorized along a number of different dimensions. For example, one can characterize memory based on the amount of time that elapses between presentation and recall of information (e.g., short- vs long-term memory) or the nature of the information that is remembered (e.g., visual vs verbal). Memory behaviors can also be characterized by task demands (e.g., recall vs recognition) or by the cognitive processes that underlie these demands (e.g., retrieval vs retention). These and other conceptual divisions of memory are reviewed in this section.

A. Short- and Long-Term Memory

In 1890, William James distinguished between memory that endured for a very brief time and memory that lasted after the experience had been "dropped from consciousness." The former, known as short-term (or primary) memory, refers to one's ability to recall material immediately after it is presented or following uninterrupted rehearsal. The latter, known as long-term (or secondary) memory, refers to the ability to remember information at a later time without the need for intervening rehearsal.

Short-term memory is of limited capacity, holding an average of seven pieces of information at any one time. This information can be maintained for up to several minutes, but it is lost or replaced by new information if not sustained by active rehearsal. A common example of short-term memory is the act of

looking up an unknown telephone number in a directory and dialing that number. The number is held briefly in short-term memory as one looks away from the directory and dials the number. If the telephone is across the room, one can maintain the number in short-term memory by repeating it continuously until it is dialed. Soon after dialing, however, the number is usually forgotten.

In contrast, long-term memory has an extraordinarily large capacity, with the potential for holding information indefinitely without the need for continued rehearsal. For example, one can recall his or her own telephone number, as well as a myriad of other facts, experiences, and personal information, without needing to reference an external source or rehearse the information continuously.

The clinical significance of the distinction between short- and long-term memory has been exemplified in patients with amnesic disorders. One of the most famous case studies in the neuropsychological literature describes a patient (H.M.) who became amnesic following surgical treatment for epilepsy. After surgery, H.M. was unable to retain any new information that he learned, such as the names of new staff in his doctor's office or information that he read each day in the newspaper. His ability to repeat information immediately following presentation or hold onto new information with active rehearsal, however, was normal. This dissociation underscores the distinct nature of the processes required to create new short- and long-term memories.

A second notable dissociation of memory functions demonstrated by H.M. was that although he was unable to form new long-term memories, he could recall previously learned information normally. This discrepancy suggests that the cognitive processes required to retain and access stored information differ from those required to encode new information.

B. Encoding, Retention, and Retrieval

Encoding, retention, and retrieval refer, respectively, to the processes by which information is acquired and transformed into a stored mental representation, maintained over time without active rehearsal, and brought back into consciousness from storage. Successful encoding occurs when an individual demonstrates acquisition of more information than would normally be possible to hold in short-term memory alone. For example, if someone is presented a list of 16 words, he or she might be able to recall only 6 or 7 of those words after seeing the list for the first time (e.g., a normal amount of information that can be held in short-term memory). If the list is repeated, however, a neurologically intact individual will remember increasingly more words with each subsequent repetition. The ability to remember more words than could normally be held in short-term storage indicates that the information has been encoded into a more permanent, long-term storage.

Once information is encoded, it must be retained for later use and accessed when needed. Failure to remember encoded information may indicate a problem with either retention or retrieval mechanisms. These are distinct memory processes that can be distinguished by examining the relationship between free recall and recognition. Free recall of information places maximum demands on retrieval processes because an individual must actively search memory to find and access the information. Recognizing the correct answer among multiple choices, however, is a less difficult memory task because it minimizes search and retrieval demands. If information is retained successfully but cannot be retrieved, free recall will be impaired but recognition of the correct information will be disproportionately better. In contrast, if information is not retained in long-term storage, both recall and recognition of information will be equally poor. The differences in memory performances among individuals with selective deficits in encoding, retention, and retrieval are presented in Fig. 1.

C. Retroactive and Proactive Interference

Memory can be disrupted by events or information encountered at approximately the time of encoding. This disruption is known as an "interference effect."

1. Retroactive Interference

Retroactive interference (RI) refers to the disrupting effect that new learning has on the ability to recall previously learned information. For example, a student studying several text chapters for an exam will be able to recall a greater amount of information from a chapter that he or she has just finished reading than from a chapter that was read earlier in the evening. In this example, information read in the most recent chapter interferes with memory for information from previous chapters.

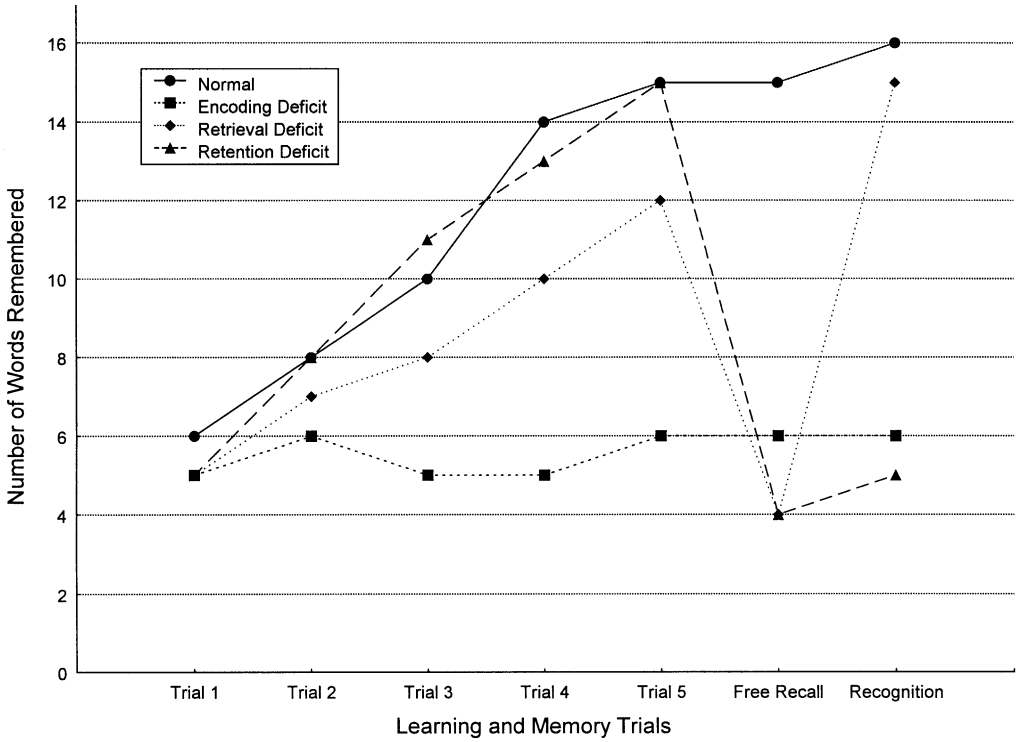


Figure 1 Memory test profiles expected in neurologically intact individuals and patients with primary deficits of encoding, retrieval, or retention. A list of 16 words is presented over five learning trials, with memory for the list assessed immediately after each trial. Memory for the word list is again assessed after a 20-min delay by means of free recall and recognition. Compared to individuals with no memory disorder, patients with encoding problems do not recall more words with repeated exposure to information. Patients with retrieval problems show poor delayed memory when assessed by free recall but normal memory when assessed by recognition. Patients with retention problems do not benefit from recognition testing.

The effect of RI is a function of the amount of new information encountered and the degree to which the interfering information and target information are similar. If a large amount of similar material is presented between initial learning and eventual recall of target information, recall will be poorer than if small amounts of dissimilar information had been presented during the intervening time.

2. Proactive Interference

Proactive interference (PI) occurs when information presented at an earlier time interferes with one's ability to learn and recall new information. For example, suppose one is asked to learn several lists of different words. After the first list is presented and memory for the words is assessed, a new list of words is presented and memory for the new words is assessed. As more lists are presented, memory for the new words declines because previously learned words produce interference.

As with RI, the effect of PI on recall also increases with the amount of interfering information and the degree of similarity to the target information. If one were presented five successive word lists, the effect of PI would be greater after the fifth list was presented than after the second list was presented. In addition, if the lists all contain similar words, such as animal names, memory for each new list will be worse than if each list contained a variety of different words.

One interesting aspect of PI is that presenting dissimilar interfering information can facilitate memory. Using the previous example, suppose a group of individuals are asked to learn four successive word lists, all of which contain different animal names. Now suppose a fifth trial is given. Half of the individuals are presented another list of animal names, whereas the other half are presented a list of words belonging to an unrelated category (e.g., clothing). Results of such an experiment are presented in Fig. 2. Individuals who are presented more animal names on trial 5 continue to demonstrate PI, whereas those presented with a

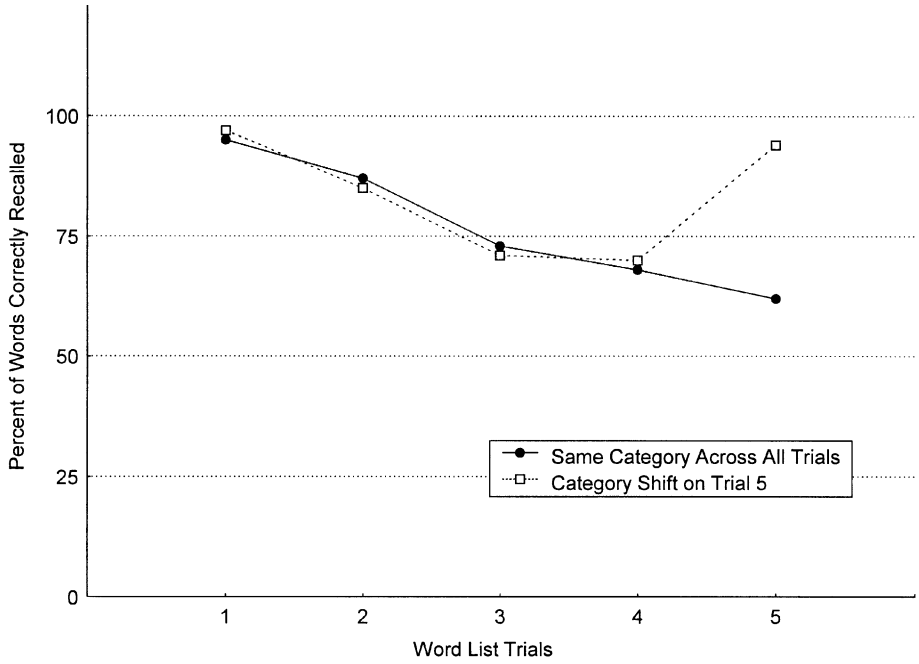


Figure 2 An illustration of proactive interference (PI). On four consecutive trials, study participants are asked to recall different lists of words from the same semantic category (e.g., animals). The effect of PI can be seen in the decline in performance over these trials. On the fifth trial, participants who are asked to recall another set of words from the same category continue to demonstrate PI, whereas those who are asked to recall words from a new semantic category (e.g., clothing) show improvement, known as "release" from PI.

dissimilar list of words demonstrate improved recall. This phenomenon is commonly known as "release" from PI.

D. Anterograde and Retrograde Memory

Anterograde memory is the ability to learn new information and to form new memories from a given moment forward. In contrast, retrograde memory is the ability to recall or recognize information or events that occurred prior to a specific moment in time.

Anterograde and retrograde amnesias are often seen in patients with brain injuries. Patients who are injured in motor vehicle accidents, for example, may not recall the events leading up to the accident (i.e., a retrograde deficit) or the events that occurred immediately following the accident (i.e., an anterograde deficit). The degree of anterograde and retrograde amnesia following a head injury is highly correlated with the severity of brain damage sustained. Anterograde and retrograde amnesias can also occur independently. It is common for patients with certain disorders to have intact memory for past events but poor ability to lay down new memories. The reverse pattern, however, is rare. Individuals who present with complete loss of past memories, personal history, and personal identity typically are found to have psychological rather than neurological disorders.

1. Recent and Remote Memory

The temporal dimension of retrograde memory is often divided into recent versus remote time frames. Recent memory typically refers to information that has been acquired within a relatively short period of time prior to an event (i.e., injury or time of evaluation), whereas remote memory refers to information about events or experiences that occurred years or decades before.

Patients with retrograde amnesia often demonstrate a temporal gradient in which memory for more recent events is disrupted to a greater extent than memory for remote events. This gradient is illustrated by the case of patient P.Z., a distinguished scientist who became amnesic secondary to alcoholic Korsakoff syndrome. Several years prior to the onset of his amnesic syndrome, P.Z. completed an autobiography. Investigators used this information to assess P.Z.'s memory for his own past life events. A temporally graded retrograde amnesic disorder is illustrated in Fig. 3, as this patient demonstrated significantly greater

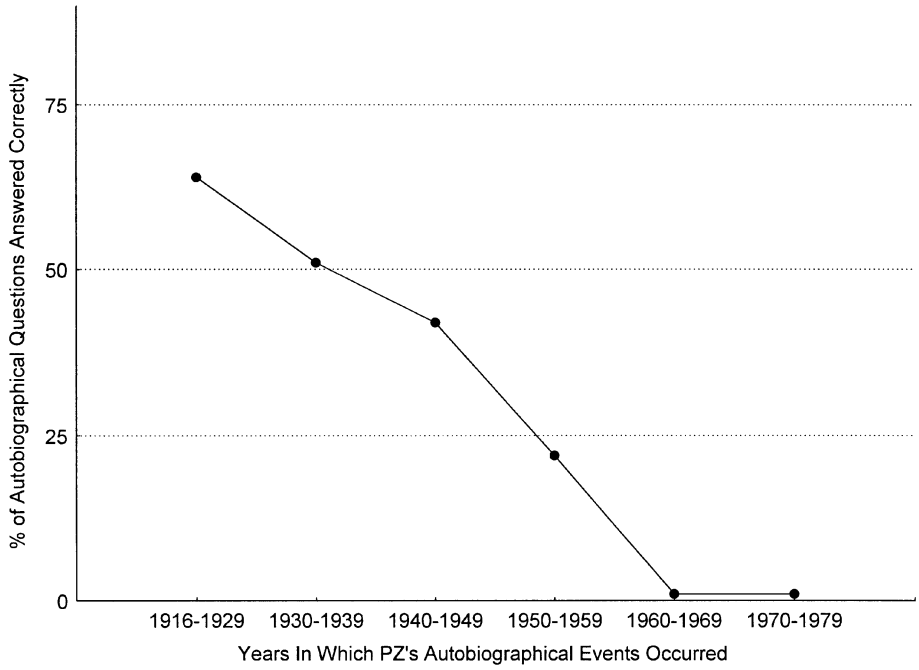


Figure 3 Temporally graded retrograde amnesia for autobiographical information as seen in patient PZ. Information from earlier decades in PZ's life was recalled better than information from recent decades.

impairment of memory for recent experiences and events compared to his memory for events from his early life.

E. Declarative versus Nondeclarative Memory

Declarative memory (also called explicit memory) refers to the acquisition of facts, experiences, and information about events. It is memory that is directly accessible to conscious awareness and thus can be "declared." In contrast, nondeclarative memory (also called implicit memory) refers to various forms of memory that are not directly accessible to consciousness. These include skill and habit learning, classical conditioning, priming, and other situations in which memory is expressed through performance or skill rather than through conscious recollection.

1. Types of Declarative Memory

a. Episodic Memory Episodic memory refers to information that is linked to a particular place and time. The ability to answer questions regarding what you ordered at a restaurant the night before or what information was presented at a meeting you attended are examples of episodic memory. That is, in order to recall the target information correctly, the individual must access information regarding the time and place the information was acquired.

b. Semantic Memory Semantic memory refers to general knowledge that is not linked to a particular temporal or spatial context. For example, defining the word "restaurant" or reciting the alphabet do not require knowledge of where or when that information was originally learned. Both episodic and semantic memories are declarative, however, in that retrieval of information is carried out explicitly, on a conscious level.

2. Types of Nondeclarative Memory

a. Procedural Memory Procedural memory is the process of retrieving information necessary to perform learned skills. These skills may be movement based, such as tying a shoe or riding a bicycle, or they may be perceptual in nature, such as learning to read mirror-reversed text. Although some aspects of skills can be declared, the skills are most often performed automatically, without conscious retrieval of information regarding the procedure. Amnesic patients such as H.M. can learn how to perform several complex tasks

and will demonstrate normal retention of these new skills despite not being able to remember having ever learned how to do the tasks.

b. Conditioning Classical conditioning is one of the most basic forms of learning. A stimulus that naturally produces a certain response is paired with a neutral stimulus. After repeated pairings, the neutral stimulus alone will elicit the response. For example, a dog will naturally begin to salivate when presented food but not when presented with the sound of a bell ringing. If, however, a bell is rung repeatedly along with presentation of food, bell ringing alone will eventually produce salivation. As with procedural memory, patients with very poor declarative memory can demonstrate normal conditioning despite being unable to recall any of the training sessions that led to the conditioned response.

c. Priming Priming is the phenomenon by which prior exposure to information influences performance on later tasks even without conscious awareness. Priming effects are demonstrated by presenting a stimulus on one occasion and measuring its influence on performance on a subsequent occasion. Depending on the level of processing, priming can be perceptual or conceptual in nature. Perceptual priming occurs when exposure to the *form* of a stimulus influences later behavior. Two examples of perceptual processing are priming for picture identification and verbal information. In picture identification tasks, an individual looks at pictures or line drawings of different objects. Later, the individual is presented a series of ''degraded'' pictures (i.e., pictures in which much of the visual information is masked or missing) and asked to identify them. Identification of degraded pictures is faster and more accurate if the individual had previously seen the complete picture than when a novel picture is presented, indicating some form of memory for the information.

Perceptual priming for verbal information is commonly demonstrated in word-stem completion paradigms. During the first phase of the paradigm, individuals perform tasks designed to expose them to different words. For example, they may be shown a series of words and asked to judge whether each word has a ''pleasant'' or ''unpleasant'' connotation. Later, during the test phase, a new task is introduced in which the first three letters of words (i.e., word stems) are presented and the individuals are asked to complete each stem with the first word that comes to mind. Study participants are more likely to complete word stems with words seen during the exposure phase of the study, even if they are less commonly used words. For example, individuals who participate in the exposure task and see the word ''motel'' will be more likely to use that word to complete the word stem ''mot__'' than individuals who did not participate in the exposure task. When not previously exposed to the word ''motel,'' the more common completion of this stem is the more frequently used word, ''mother.''

Although the previous example involves verbal stimuli, it is the perceptual aspect and not the meaning of those stimuli that is primed. When processing the meaning of a stimulus influences later behavior, *conceptual* priming has occurred. The exposure phase of conceptual priming experiments is similar to that of the word-stem completion paradigm. During the test phase, however, participants engage in word-association tasks or other activities that require the processing of word meanings. Suppose, for example, the word ''crown'' is presented during an exposure task. Individuals who were primed by exposure are more likely to respond with the word ''crown'' when asked for the first word that comes to mind in response to the word ''king,'' whereas naive (i.e., ''unprimed'') study participants more often respond with the word ''queen.''

Patients with impaired episodic memory typically perform normally on perceptual priming tasks even though they have no memory of having previously seen the words or pictures that have been primed. Conceptual priming, however, may be reduced in patients with declarative memory dysfunction.

II. NEUROBIOLOGICAL BASES OF MEMORY

Early investigators hypothesized that learning and memory might occur through growth of new neurons in the brain, much in the same way that strength is increased through growth of muscle tissue. It soon became clear, however, that the central nervous system was limited in its ability to generate new cells. Investigators eventually showed that existing neurons possessed the ability to form new processes and connections, and that learning and memory reflected modification of neuronal signal processing, interneuronal communication, and functional relationships among brain systems. There is now a rapidly growing and exciting literature suggesting that generation of new neurons may indeed be possible in adulthood. Although this avenue of exploration holds much promise, to date evidence suggests that neuronal

modification remains the primary mechanism underlying memory formation.

A. Memory as Change of Synaptic Structure and Efficacy

A complete discussion of neuronal function and synaptic transmission of information is beyond the scope of this article. Briefly, a synapse is a functional juxtaposition of two or more neurons (Fig. 4). When a neuron is stimulated to a sufficient degree, chemicals known as neurotransmitters are released from its axon terminal into the microscopic space that separates it from its neighboring neuron. The presence of neurotransmitters within this region produces characteristic changes in the membranes of neighboring neurons.

Each neuron in the brain interconnects with many other neurons in this fashion, forming a series of networks and feedback loops. In the first half of the 20th century, D. O. Hebb proposed that psychologically important events such as memory were manifestations of the flow of activity within a given network of neurons that were acting together as a single unit. He suggested that when an event or experience caused a set of neurons to be excited together, the synapses involved in that pathway became functionally connected.

Although some of the specific neuronal mechanisms that Hebb proposed were not supported by subsequent research, the basic hypothesis of memory as structural

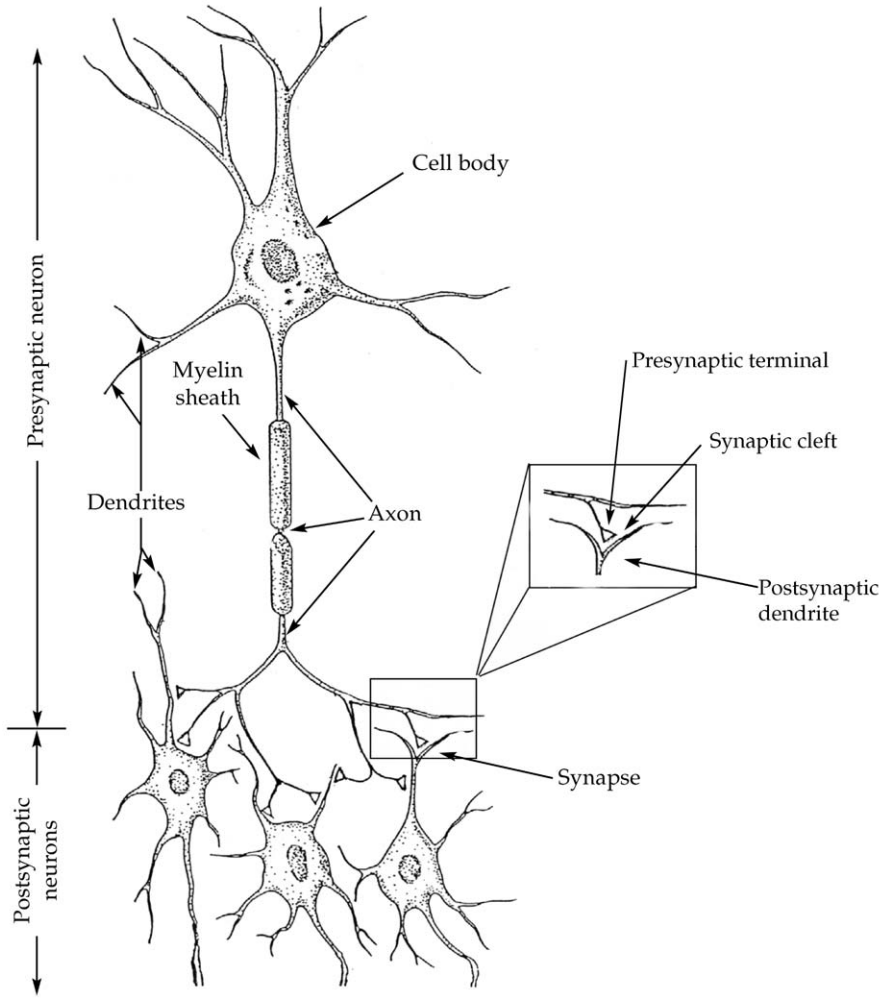


Figure 4 Illustration of the basic structure of neurons and neuronal synapses. Modified from Kandel, E. "Principles of Neural Science," 3rd ed. (1992), The McGraw-Hill Companies. Reproduced with permission of The McGraw-Hill Companies.

and functional change involving neuronal activity and synaptic transmission has endured. Research has shown that when animals are trained to perform specific tasks or are exposed to enriched environments, new synapses grow and preexisting synaptic connections become better developed. Synaptic change in response to new learning is observed in both young and adult animals and can occur subsequent to a single learning experience.

B. Neuronal Processes Underlying Short- and Long-Term Memory

A distinction can be made between the neuronal processes underlying short- versus long-term memory. Although temporary changes in the dynamics of neuronal functioning can maintain information for up to several minutes, more permanent structural changes are believed to be necessary to retain information for days or longer.

1. Neuronal Processes Underlying Short-Term Memory

Studies show that a neuron's function can be modified by intense activity. High-frequency stimulation by a presynaptic neuron, for example, tends to increase the responsiveness and efficiency of postsynaptic membranes, a phenomenon known as potentiation. Some neurons can demonstrate increased responsiveness lasting for several minutes to more than 1 hr after active stimulation has ceased. This temporary increase in synaptic effectiveness is known as posttetanic potentiation (PTP) and reflects one possible way that recently learned information can be remembered. PTP may cause neurons activated during learning to remain active for a brief time after learning has ceased, thus allowing that information to remain available.

Another possible mechanism by which neuronal activity associated with recently learned information may be maintained for a brief period of time is by means of a feedback loop. Specifically, excitatory input from exposure to information could theoretically be maintained after active input ceases if neurons within a closed loop excite each other (Fig. 5). Excitatory neuronal feedback systems such as these are known as reverberatory circuits.

The amount of time that elapses before information in short-term memory degrades is believed to be a function of PTP strength and/or the level of neuronal excitement in reverberatory circuits caused by the learning experience. More intense changes in these neuronal dynamics are typically associated with more persistent memory, whereas weaker changes are associated with more rapid forgetting. Are short- and long-term memories, then, different ends of a single graded continuum of neuronal dynamics, with long term-memory reflecting a chronic state of persisting neuronal activity?

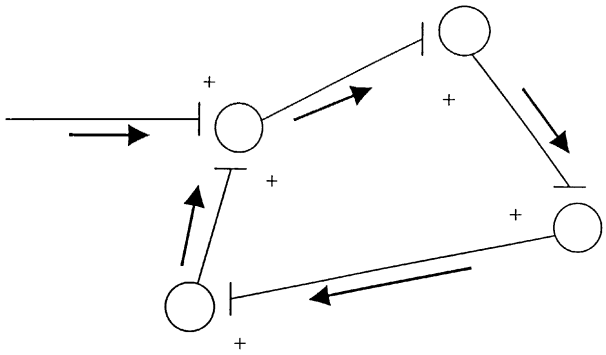


Figure 5 Schematic of a reverberatory circuit. Neuronal responses are prolonged by reexcitation of excitatory neurons.

If both types of memory relied on dynamic mechanisms such as PTP and reverberatory circuits, then interrupting those dynamics should interfere with both short- and long-term memory. Studies of patients undergoing electroconvulsive therapy (ECT), however, show that this is not the case. ECT is a treatment for medically refractory depression that works by delivering low-voltage electric current to the brain. Neuronal activity in the brain is briefly "short-circuited" by this procedure, thus interfering with all electrical activity, including PTP and activity within reverberatory circuits. Following recovery from ECT, patients demonstrate memory loss for information immediately preceding treatment; however, memories in long-term storage remain intact.

2. Neuronal Processes Underlying Long-Term Memory

The ability to encode information into a more permanent long-term storage system is believed to be a function of long-term potentiation (LTP). LTP is similar to PTP in that high-intensity stimulation increases the effectiveness of the neuronal synapse. LTP is much more powerful and longer lasting than PTP, however, and it cannot be produced by activation of only a single presynaptic neuron in a single pathway. Instead, a minimum number of inputs must be present to produce an effect.

Another difference between LTP and PTP is that LTP is associated with the activation of genes that direct the growth and structure of synapses. These neuronal changes are believed to underlie long-term memory formation and are not observed in short-term memory.

It is unlikely that every neuronal synapse is modified by each learned experience, nor is it likely that each memory corresponds to changes in a single neuron in a one-to-one fashion. So where in the brain are long-term memories formed and stored? The first neurons to be discovered that were capable of LTP were found in the hippocampus and surrounding structures, although it has since been recognized that neurons in other brain areas also demonstrate LTP. Similarly, behavioral studies of memory function and dysfunction in humans and animals indicate that a number of different brain regions are important for memory.

III. NEUROANATOMY OF MEMORY

The cumulative literature of learning and memory in humans and nonhuman animals provides a broad outline of the major neuroanatomical structures and connections believed to be important for memory functioning. In the following sections, the neuroanatomical substrates of declarative and nondeclarative memory are reviewed.

A. Neuroanatomical Correlates of Declarative Memory

Several brain regions are believed to be important to encoding and storage of long-term declarative memories. These include structures in the medial temporal lobes, medial diencephalon, basal forebrain, and prefrontal cortex. In addition, certain subcortical nuclei and white matter pathways are important for retrieving acquired information from storage. The gross anatomy of the lateral and medial surfaces of the brain is illustrated in Figs. 6 and 7, whereas some of the more important structures related to memory are illustrated in Fig. 8.

1. The Temporal Lobes

Recall from the beginning of this article the case of H.M., who suffered a profound long-term memory

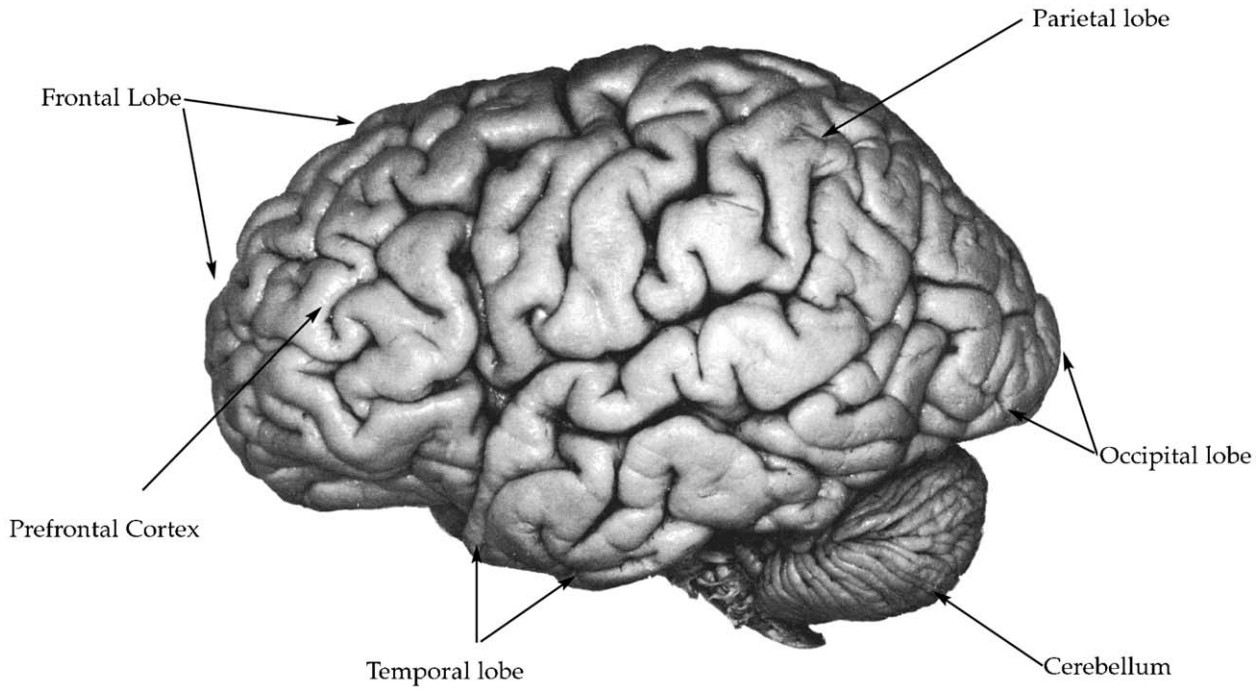


Figure 6 Lateral surface of the brain. From "Structure of the Human Brain: A Photographic Atlas," 3rd ed. by S. J. DeArmond, M. M. Fusco, and M. M. Dewey, copyright 1974, 1976, 1989 by Oxford University Press, Inc. Used by permission of Oxford University Press, Inc.

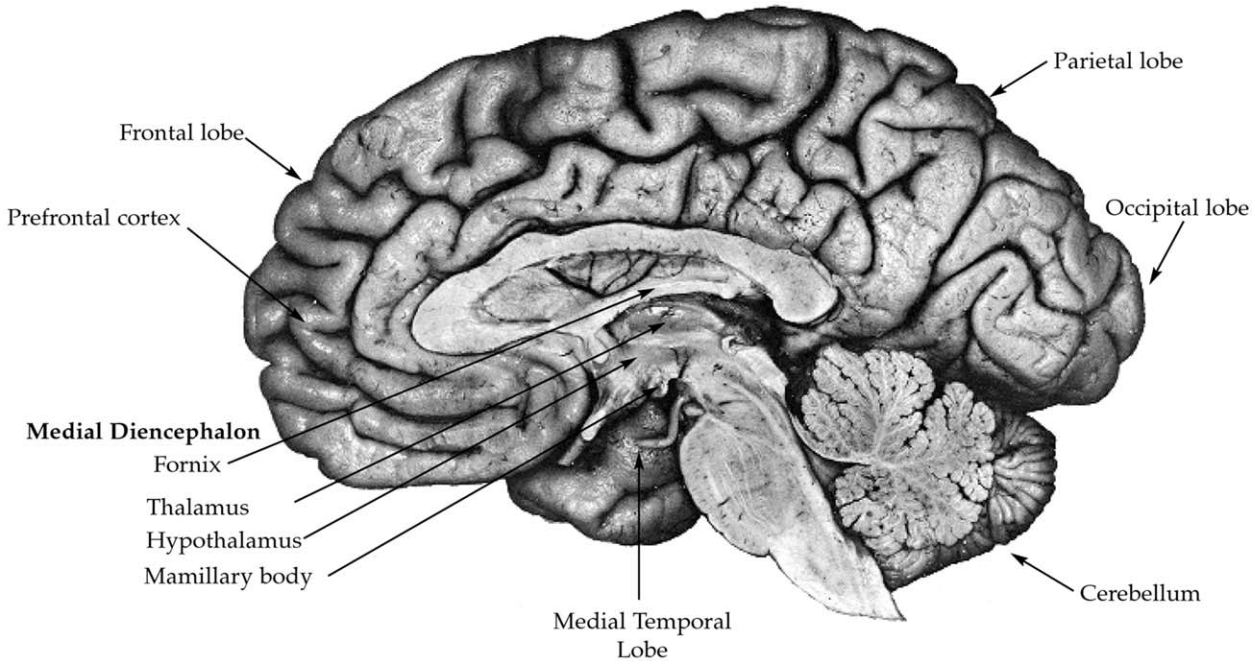


Figure 7 Medial surface of the brain. From "Structure of the Human Brain: A Photographic Atlas," 3rd ed. by S. J. DeArmond, M. M. Fusco, and M. M. Dewey, copyright 1974, 1976, 1989 by Oxford University Press, Inc. Used by permission of Oxford University Press, Inc.

disorder following brain surgery. H.M.'s surgery involved the removal of the medial portion of both temporal lobes, thus demonstrating the importance of this region to long-term memory functioning. Patients with bilateral medial temporal lobe dysfunction typically demonstrate a global anterograde memory deficit. Although short-term memory for newly presented information may be unaffected, these patients cannot encode information into long-term storage. On memory tests, they demonstrate poor recall and poor recognition of previously presented information. Patients with medial temporal lobe damage, however, demonstrate intact memory for information that had been acquired prior to their injury, semantic knowledge, and previously learned skills. They also demonstrate normal priming effects.

Patients with bilateral medial temporal lobe dysfunction demonstrate global declarative memory deficits. The most common cause of this type of brain dysfunction is Alzheimer's disease, a progressive degenerative disorder in which neuropathological changes are first observed in the hippocampus. Other brain regions are affected as dementia progresses; however, the ability to form new episodic memories tends to remain the most significant behavioral symptom.

Patients with focal brain diseases such as stroke or tumors may demonstrate unilateral temporal lobe dysfunction. In such cases, material-specific memory deficits may be observed. Patients with left temporal lobe damage typically have more difficulty learning and remembering verbal material, such as stories or word lists, than nonverbal material, such as abstract geometric patterns, faces, tonal patterns, or the spatial location of objects. Patients with right temporal lobe damage tend to demonstrate the opposite pattern of memory impairment.

The temporal lobe is a relatively large brain region with several anatomically distinct areas. Most investigators agree, however, that the structures in the medial portion of the temporal lobes—specifically the hippocampus, its surrounding cortex, and the amygdala—are most important for memory (Figs. 8–10).

a. The Hippocampus and Related Structures The human hippocampus and associated cortices are located bilaterally in the cerebral hemispheres, forming a ridge that extends along the temporal horn of each lateral ventricle. As illustrated in Fig. 9, these structures are convoluted in shape, with several distinct regions defined by differences in cellular structure and organization. Proceeding from the collateral sulcus, the parahippocampal gyrus curves dorsally, transitioning into the subiculum, which curves medially and transitions into the hippocampus

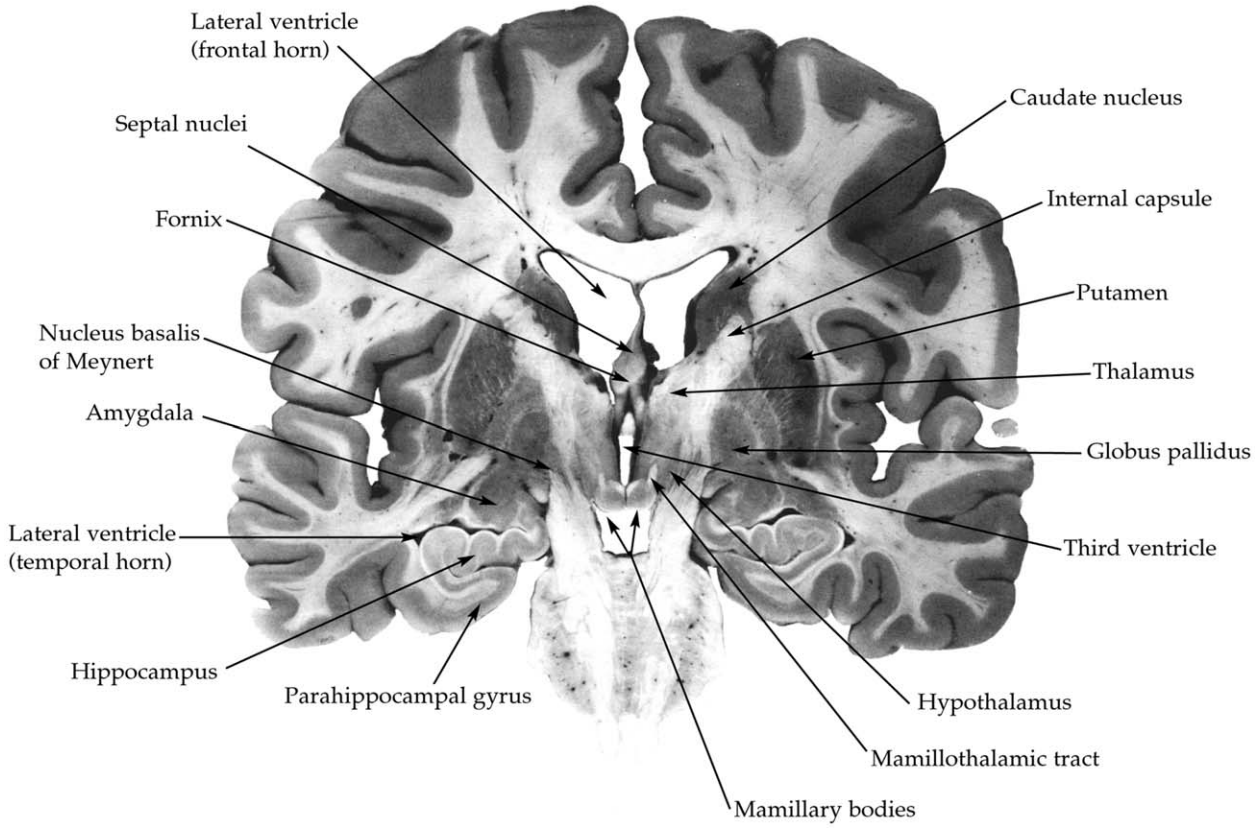


Figure 8 Coronal section of the brain at the level of the anterior portion of the temporal lobes illustrating many of the structures believed to be important to memory. From "Structure of the Human Brain: A Photographic Atlas," 3rd ed. by S. J. DeArmond, M. M. Fusco, and M. M. Dewey, copyright 1974, 1976, 1989 by Oxford University Press, Inc. Used by permission of Oxford University Press, Inc.

proper. The hippocampus curves inward again, forming the hippocampal fissure. When the hippocampal fissure is opened, a narrow layer of cortex can be observed, known as the dentate gyrus. Axons arise from neurons comprising the hippocampus proper, converge to form the fimbria, and continue into the fornix, which is the major white matter tract out of the hippocampal formation.

b. The Amygdala The amygdala is a collection of nuclei and specialized cortical areas situated in the dorsomedial portion of the temporal lobe rostral and dorsal to the tip of the temporal horn of the lateral ventricle (Fig. 8). Although early investigators believed the amygdala was primarily an olfactory structure, later studies revealed substantial inputs from brain regions responsible for processing many different types of sensory information, including visual, auditory, somatosensory, and autonomic stimuli. Outputs from the amygdala project to many brain regions believed to be important to memory, including the hippocampal formation, thalamus, hypothalamus, prefrontal cortex, basal forebrain, and corpus striatum.

The amygdala is important to emotional, autonomic, reproductive, and feeding behaviors. Electrical stimulation of the amygdala in animals typically results in a constellation of aggressive and fear-related responses. Bilateral lesions to the amygdala in humans result in Klüver–Bucy syndrome, a disorder characterized by excessive docility, lack of fear response, hypersexuality, hyperorality, and changes in dietary habits.

c. Relative Contributions of the Hippocampus and Amygdala to Memory Given the close proximity of the hippocampus to the amygdala, both structures are typically damaged in clinical cases of amnesia, thus leading to the question of which structure is more important to memory. The preponderance of case studies reported in the literature suggest that when damage is restricted to the hippocampus proper,

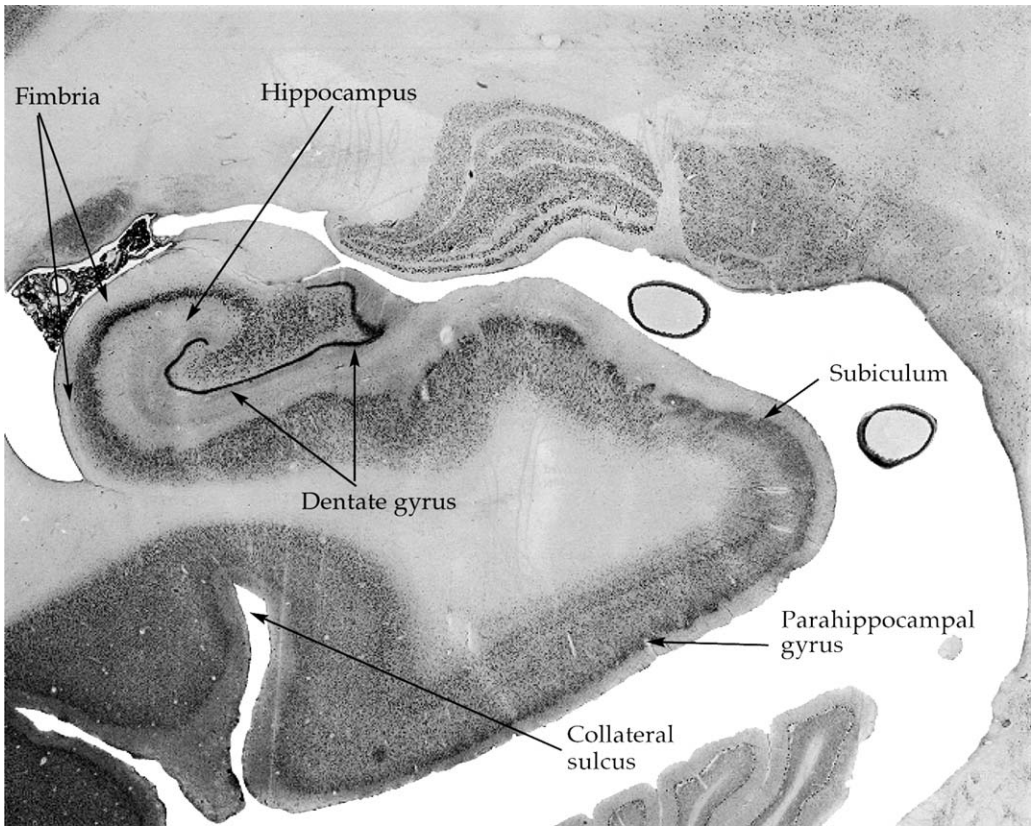


Figure 9 Illustration of a transverse section through the human hippocampal formation and parahippocampal gyrus. From "Structure of the Human Brain: A Photographic Atlas," 3rd ed. by S. J. DeArmond, M. M. Fusco, and M. M. Dewey, copyright 1974, 1976, 1989 by Oxford University Press, Inc. Used by permission of Oxford University Press, Inc.

fimbria, dentate gyrus, and subiculum, declarative memory impairment is less severe than when there is more radical loss of tissue.

Similar findings have been observed in nonhuman primates. When the hippocampal formation and its adjacent cortical region are lesioned, less severe declarative memory disturbance is observed than when the amygdala and its adjacent cortices are also involved. The most severe deficit is observed when the hippocampus, its adjacent cortical region, and the cortical region adjacent to the amygdala (i.e., the entorhinal and perirhinal cortices) are damaged. Damage to the amygdala, however, does not create any greater declarative memory deficit beyond what is caused by damage to the other three brain regions.

As discussed earlier, the amygdala is known to be important for processing sensory stimuli and information related to emotions. Studies show that the amygdala is important for remembering the emotional context of information. It is also implicated in certain types of nondeclarative memory, such as conditioning and evoking fear-related behaviors.

2. The Medial Diencephalon

The diencephalon is a region of several important nuclei located at the rostral part of the brain stem (Figs. 7 and 8). This region can be divided into four major areas: the epithalamus, thalamus, hypothalamus, and subthalamus. The divisions important to memory functioning include portions of the thalamus and hypothalamus. The thalamus is the largest subdivision of the diencephalon and is composed of several histopathologically distinct nuclei (Fig. 10). The hypothalamus lies below the thalamus and creates the floor of the third ventricle. It controls many autonomic, endocrine, and somatic functions and is composed of several distinct anatomical structures, including the mamillary bodies.

Severe amnesia results from damage to regions along the midline of the thalamus and hypothalamus.

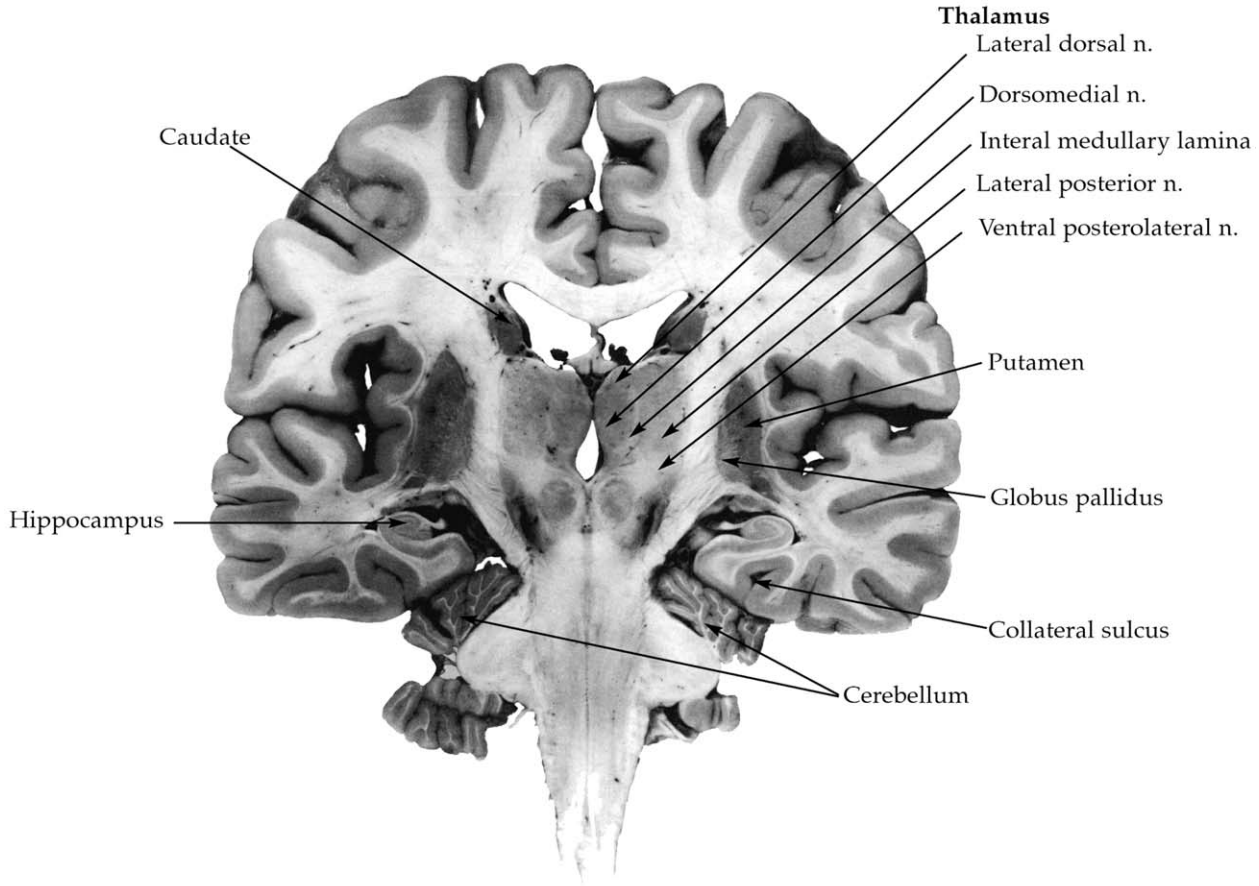


Figure 10 Coronal section of the brain at the level of the midbrain illustrating the thalamic nuclei. From "Structure of the Human Brain: A Photographic Atlas," 3rd ed. by S. J. DeArmond, M. M. Fusco, and M. M. Dewey, copyright 1974, 1976, 1989 by Oxford University Press, Inc. Used by permission of Oxford University Press, Inc.

There has been considerable controversy, however, concerning which of the many neuroanatomical structures and connections in this region must be damaged to cause memory dysfunction. The structures that have most often been implicated in diencephalic amnesia are the dorsomedial nucleus of the thalamus, the mamillary bodies of the hypothalamus, and the white matter (mamillothalamic) tract connecting these structures.

Patients who most commonly demonstrate diencephalic amnesia are individuals with Wernicke–Korsakoff syndrome. This disorder is caused by severe thiamine deficiency, often associated with chronic severe alcoholism. The first stage of this disorder involves an acute, global confusional state, oculomotor abnormalities, ataxia, and peripheral polyneuropathy. This stage is known as Wernicke's encephalopathy, during which the patient is in danger of suffering a potentially fatal midbrain hemorrhage unless treated immediately with thiamine. Approximately one-fourth of successfully treated patients regain the majority of their premorbid cognitive abilities. The remaining 75%, however, demonstrate severe, persistent anterograde and retrograde memory deficits.

a. The Mamillary Bodies and Dorsomedial Nucleus The mamillary bodies and dorsomedial nucleus of the thalamus are both frequently damaged in Korsakoff syndrome. The few case studies in the literature of patients with naturally occurring lesions suggest that damage restricted to either the dorsomedial nucleus or the mamillary bodies is capable of interfering with memory.

In monkeys, lesions involving both the mamillary bodies and the dorsomedial thalamic nucleus result in memory deficits that are substantially more severe than those caused by lesions to either structure separately. Moreover, lesions to any single

diencephalic structure appear to produce less memory impairment than lesions to the hippocampal formation.

b. The Internal Medullary Lamina A series of studies using an animal model of alcoholic Korsakoff syndrome in rats suggested that damage to the internal medullary lamina may also be critical in diencephalic amnesia. Diencephalic damage induced in rats via an experimentally controlled thiamine deficiency caused bilateral lesions in the mamillary bodies, dorsomedial nucleus of the thalamus, and internal medullary lamina of the thalamus. Rats with radio frequency-induced lesions to the internal medullary lamina demonstrated more similarities to rats with thiamine deficiency on measures of spatial memory than to rats with lesions to the mamillary bodies or midline thalamic nuclei.

3. The Basal Forebrain

Significant memory dysfunction has been associated with damage to the basal forebrain. The basal forebrain is a loose term used to describe the area of the brain superior to the optic chiasm. It includes the medial septal nuclei, nucleus accumbens, anterior hypothalamus, diagonal band of Broca, nucleus basalis of Meynert, and part of the prefrontal cortex (i.e., Brodmann's area 13). The structures within the basal forebrain project widely throughout the rest of the brain. The septal nuclei and nucleus basalis of Meynert (Fig. 8), for example, have extensive connections to and from the hippocampal formation, amygdala, and neocortex and are believed to be important to memory functioning.

Basal forebrain involvement in memory functioning is implicated primarily from the study of two patient groups: patients with ruptured aneurysm of the anterior communicating artery and patients with dementia due to Alzheimer's disease. The basal forebrain is perfused primarily by branches of the anterior communicating artery; thus, disturbances within this flow of circulation result in infarction and necrosis of basal forebrain tissue. Patients with stroke, hemorrhage, or damage to this area subsequent to aneurysm surgery often demonstrate declarative memory deficits.

Patients with Alzheimer's disease demonstrate marked degeneration within the basal forebrain as well as in the medial temporal lobes, as discussed earlier. The region of the basal forebrain most affected by Alzheimer's disease is the nucleus basalis of Meynert, a complex of neurons that produce the neurotransmitter acetylcholine. Recall that the ability to learn and remember new information reflects enhanced communication among neurons due to the development and growth of synapses. This growth occurs when neurotransmitters are in adequate supply. The primary neurotransmitter associated with the ability to encode and retain new declarative memories is acetylcholine. Thus, the depletion of acetylcholine in the brain caused by degeneration of neurons in the nucleus basalis of Meynert is believed to underlie part of the memory disorder associated with Alzheimer's disease.

In monkeys, combined lesions to the nucleus basalis of Meynert, medial septal nuclei, and diagonal band of Broca result in significant memory impairment. No significant memory impairment is noted, however, if each structure is lesioned separately. Thus, it appears that extensive damage to the basal forebrain, rather than specific damage to any given structure, is necessary to produce memory impairment. In light of these findings and the presence of extensive anatomical connections between the basal forebrain and medial temporal lobe structures, some have suggested that the basal forebrain most likely modulates medial temporal lobe memory processing but is not in and of itself a memory center.

4. The Prefrontal Cortex

Functional neuroimaging studies have found that activity in the anterior region of the prefrontal cortex (Figs. 6 and 7) increases when one attempts to retrieve previously learned information. Specifically, the prefrontal cortex is believed to play an important role in directing the attentional and organizational processes necessary for both encoding and retrieval of information. Disruption of these cognitive processes is associated with characteristic memory impairments, including source memory deficits, impaired temporal ordering, and confabulation.

Source memory and temporal ordering refer to the ability to recall the spatial and temporal contexts within which information was originally acquired. Patients with prefrontal cortical damage may recall factual information correctly, but they often have difficulty recalling where they learned the information. If they learned several pieces of information over a period of time, they may recall all the information adequately but be unable to report which information was learned first.

Another interesting behavioral consequence of prefrontal damage is the tendency to confabulate.

When patients with prefrontal damage cannot recall factual information, they often fabricate false information rather than indicate that they do not remember the correct information. This is different from lying because these patients have no desire or intent to deceive. Confabulation is believed to be a manifestation of source memory and temporal ordering deficits, coupled with poor self-monitoring abilities. When spatial and temporal contexts of information are missing it becomes very difficult to select accurate recollections from all the possible information that can be retrieved in a memory search. Patients with prefrontal damage, it is argued, simply choose one of the many alternatives retrieved from memory.

5. Subcortical Nuclei and White Matter

As illustrated in Figs. 8–10, two kinds of substances can be appreciated when the brain is cut—one light and one dark. The dark substance is known as gray matter and contains neuronal bodies, dendrites, and synapses. Gray matter covers the external surface of the brain (i.e., the cortex) and forms several discrete nuclei within the central portion of the brain (i.e., subcortical nuclei). The light substance separating the cortical and subcortical gray matter is known as white matter. White matter is made up of neuronal axons, the majority of which are covered in fatty sheaths known as myelin. It is the myelin that is responsible for the white appearance of these regions.

Information is processed and encoded in gray matter, whereas the white matter pathways provide the means by which information is transferred and communicated throughout the brain. Research suggests that the ability to search and retrieve information from long-term memory stores is a function of subcortical nuclei and white matter systems. These data derive primarily from studies of patients with basal ganglia diseases, such as Parkinson's disease and Huntington's disease, and from studies of patients with white matter diseases, such as multiple sclerosis. These patients tend to be slower when processing and learning information, and they demonstrate poor free recall. Their recognition memory, however, is usually equivalent to that of healthy controls, indicating a primary retrieval deficit.

B. A Model of Encoding and Memory Storage

Memory researchers believe that after sensory information is processed by the neocortex, it is sent along parallel pathways to the hippocampal cortices and medial diencephalon for memory processing. During encoding, associations are created among stimulus features. These associations serve as indices to the cortical sites where the information was originally processed. Input from the basal forebrain and prefrontal cortex modulates memory processing and tags the newly learned information with temporal and spatial information.

Once memories are formed, they must be stored in such a way as to be searched and accessed at a later time when the information is needed. One possibility is that long-term memories are stored in the same brain structures in which new memories are formed. Research on amnesic patients, however, suggests that this is unlikely. Patient H.M., for example, lost the ability to form new memories following bilateral removal of medial temporal lobe structures, but he was able to recall previously learned information normally. This suggests that after new memories are formed they are transferred elsewhere for long-term storage. Researchers believe that permanent memory storage develops in the neocortex; however, the exact nature of the stored information is not known. It remains unclear, for example, whether information is stored regionally or more diffusely throughout the brain and whether memories are stored as basic elemental forms or in more complex formats.

C. Neuroanatomical Correlates of Nondeclarative Memory

Nondeclarative memory is performance based and comprised of phenomena such as skill learning, conditioning, and priming. Unlike declarative memory, the formation of nondeclarative memories is not reliant on the medial temporal lobe/medial diencephalic system. Instead, memories underlying acquired skills or conditioned or primed responses are a function of the sensory and motor systems inherent in the involved behaviors.

1. Skill Learning and Memory

The ability to learn new motor skills and procedures is a function of the corticostriatal system and the cerebellum. The coroticostriatal system includes the basal ganglia and its projections from the neocortex. The basal ganglia are large subcortical nuclei that include the caudate nucleus, putamen, and globus

pallidus (Figs. 8 and 10). The basal ganglia have extensive connections with the thalamus, subthalamic nucleus, amygdala, substantia nigra, and broad regions of the neocortex.

The basal ganglia are important for motor planning and programming. Patients with diseases causing dysfunction of the basal ganglia, such as Parkinson's disease and Huntington's disease, typically demonstrate impaired motor skill learning. Learning a new skill requires development and modification of accurate motor programs. As the skill is learned, appropriate movements are performed in the correct serial order and the correct temporal pattern within that order. A feedback system detects errors and generates new, more accurate movements as a result. Dysfunction of the basal ganglia interrupts complex motor and sensory circuits and thus interferes with the ability to generate and/or modify motor programs.

Skill learning that depends on integrating visual and motor information or that requires visual feedback to develop and refine the skill is believed to be a function of the cerebellum. Lesions to the cerebellum are associated with impairments of abilities such as learning and demonstrating the skill of tracing objects when looking at them in a mirror.

Learning visual perceptual skills that do not require motor responses, such as the ability to read text that is presented in reverse mirror-image, depends on the integrity of posterior cortical regions. During learning, the right parietal region becomes activated, presumably because of the need to process the visuospatial aspects of the information. Once the skill is acquired, however, this region becomes less active while activity in the left temporooccipital region increases. The left temporooccipital region is important to normal reading, suggesting that once the skill is learned, the need to decode the visuospatial aspects of the words diminishes and skilled reading occurs.

2. Conditioning

Animal studies indicate that the cerebellum is responsible for most types of conditioned learning. Electrophysiologic changes are observed in the cerebellum during conditioned learning, whereas lesions to the cerebellum disrupt learning. Electrophysiologic changes are also noted in the hippocampus during classical conditioning. Lesions to the hippocampus, however, do not inhibit conditioned learning, suggesting that the hippocampus reflects a parallel information processing system during conditioning but does not in and of itself mediate conditioned learning. In humans, naturally occurring cerebellar lesions have been shown to interfere with conditioning, whereas lesions to the medial temporal lobes, medial diencephalon, or basal ganglia have no effect on conditioned learning.

The ability to condition a fear response by pairing a neutral stimulus, such as a light or tone, with an aversive unconditioned stimulus, such as electric shock, is a special case of classical conditioning and is a function of the amygdala. Ablation of the amygdala in a variety of mammalian species interferes with the animal's innate fear response and its ability to learn new fear responses. The site of storage for long-term fear-related memory, however, is believed to be outside the amygdala. When the amygdala is lesioned shortly after a new fear response is conditioned, the animal will not demonstrate the response; however, lesioning the amygdala several days after successful conditioning causes no disruption in the conditioned response. Brain areas that have been implicated as possible storage sites for fear memories include the insular cortex and the vermis of the cerebellum.

3. Priming

Patients with dysfunction of the medial temporal lobe, medial diencephalon, and basal ganglia demonstrate intact perceptual priming despite impaired declarative memory or skill learning. Therefore, perceptual priming does not appear to be dependent on these brain regions. Instead, perceptual priming is a function of the cortical regions necessary to process whatever information is primed. Brain imaging studies, for example, show that the right posterior cortex (i.e., the brain region responsible for processing physical features of visual stimuli) is necessary for perceptual priming of words. Patients with damage to this region do not demonstrate visual word priming despite the ability to remember the words on explicit recall tasks. Brain regions required for other forms of perceptual priming likely include cortical regions involved in processing the stimulus qualities and external demands inherent to the individual priming task. Conceptual priming, on the other hand, is believed to be a function of the left frontal cortex.

IV. SUMMARY

The past several decades have witnessed great progress in our understanding of cognitive, biological, and

neuroanatomical processes underlying memory functioning. We have learned, for example, that there are two dissociable memory systems—one that is conscious and declarative in nature and another that is unconscious and expressed through performance rather than through conscious recollection. Each of these broad categories of memory is composed of several qualitatively different functions that can be defined along a variety of dimensions.

We have also learned that encoding and memory reflect changes in neuronal function in different brain structures. The medial temporal lobe structures and medial diencephalon, for example, are critical for the formation of new declarative memories but have little impact on memory for previously learned information or nondeclarative forms of learning and memory. On the other hand, basal ganglia dysfunction can disrupt some aspects of nondeclarative memory (i.e., skill learning) while sparing declarative memory. The cerebellum is essential to most forms of classical conditioning, whereas the amygdala plays a key role in conditioned learning of fear responses and emotional modulation of memory.

Despite our advances in knowledge, we still know relatively little about where and how memories are preserved in long-term storage. We know that this information is not stored in the brain structures in which they are created and we presume instead that they are stored in the neocortex. The exact location and format in which memories are preserved over long periods of time remain unclear.

See Also the Following Articles

ALZHEIMER'S DISEASE, NEUROPSYCHOLOGY OF • DEMENTIA • INFORMATION PROCESSING • INTELLIGENCE • MEMORY DISORDERS, ORGANIC • NEOCORTEX

Suggested Reading

Butters, N., Delis, D. C., and Lucas, J. A. (1995). Clinical assessment of memory disorders in amnesia and dementia. *Annu. Rev. Psychol.* **46,** 493–523.

Cermak, L. S. (1994). *Neuropsychological Explorations of Memory and Cognition: Essays in Honor of Nelson Butters*. Plenum, New York.

Crosson, B. (1992). *Subcortical Functions in Language and Memory*. Guilford, New York.

Gabrieli, J. D. E. (1998). Cognitive neuroscience of human memory. *Annu. Rev. Psychol.* **49,** 87–116.

Kandel, E. R. (1991). Nerve cells and behavior. In *Principles of Neural Science* (E. R. Kandel, J. H. Schwartz, and T. H. Jessell, Eds.), pp. 18–32. Appleton & Lange, Norwalk, CT.

Markowitsch, H. J. (2000). The anatomical bases of memory. In *The New Cognitive Neuroscience* (M. S. Gazzaniga, Ed.), pp. 781–796. MIT Press, Cambridge, MA.

Schachter, D. L., and Buckner, R. L. (1998). On the relations among priming, conscious recollection, and intentional retrieval: Evidence from neuroimaging research. *Neurobiol. Learning Memory* **70,** 284–303.

Scoville, W. B., and Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *J. Neurol. Neurosurg. Psychiatr.* **20,** 11–21.

Squire, L. R. (1987). *Memory and Brain*. Oxford Univ. Press, New York.

Mental Retardation

SANDRA CLUETT REDDEN
University of North Carolina at Chapel Hill and Glenwood Children's Resource Center

STEPHEN R. HOOPER
University of North Carolina at Chapel Hill

MARTHA POPE
Goldsboro Middle School and Wayne County Schools

GLOSSARY

akathisia Involuntary motor restlessness with increased fidgety behaviors.

brachycephaly The physical feature of having a broad or tall head.

dyskinesia Involuntary nonrhythmic motor movements that can be slow or fast and usually involve movements of the mouth.

epicanthic folds Folds of skins at the inside corner of the eye, common in persons with Down's syndrome.

hypogonadism Decreased function of the sex glands.

idiopathic Relating to or designating a disease having no known cause.

incidence The number of new cases of disease occurring in a period of time divided by the number of people at risk during that period.

macroorchidism Enlarged testes.

nystagmus Involuntary rapid movements of the eyes.

philtrum The groove in the midline of the lips.

prevalence The total number of cases of a disorder existing within a population at a given time or in a given area.

strabismus Deviation of one or both of the eyes during forward gaze.

Mental retardation refers to substantial limitations in a person's level of present functioning. It is characterized by significantly subaverage intellectual functioning, existing concurrently with related limitations in two or more of the following applicable adaptive skill areas, communication, self-care, home living, social skills, community use, self-direction, health and safety, functional academics, leisure, and work. Mental retardation manifests before age 18.

I. INTRODUCTION

The term mental retardation represents divergent meanings to different persons. It is a general label that incorporates a heterogeneous group of persons, including those with variable cognitive aptitudes, emotional functioning and adjustment, and social development. It can be described as a concept, condition, symptom, or syndrome. The purpose of this article is to elucidate the

facets of mental retardation by focusing on its history, definitional issues, legal implications, etiology and prevention, prevalence, characteristics within specific syndromes, comorbid psychopathology, and treatment.

II. HISTORY OF MENTAL RETARDATION

There is a long history of persons with mental retardation that dates back to antiquity. In general, the view of persons with mental retardation varied based on accompanying social, political, and religious beliefs. Earliest recordings of mental retardation are found at approximately 1500 BC from ancient Egypt. These recordings refer to those whom we would distinguish as persons with mental retardation as being inferior, needing to be separate from society, and having the real possibility of being put to death because of their unusual presentation and limitations. In the Greco-Roman period (1300 BC to AD 476), persons with handicaps were not regarded as being fully human. They were placed alongside slaves and women, who could not take part in public life. They were described as the "private people" or "the idiots." Neglect, infanticide, cruelty, and abuse were common.

The Middle Ages, Renaissance, and Reformation (476 to the 16th century) began with the fall of Rome and the subsequent heightened status of Christianity. This resulted in a period of time in which the church, through foundling homes and orphanages, provided for those whom they believed were mentally deficient. These individuals came to be known as children of innocence and of God who were to be protected. Later in this period, Martin Luther proposed that "feeble-minded children" were a result of sin and the Devil, and that they and their mothers should be burned at the stake. Luther's view was one that influenced society in a pervasive manner. Throughout all the foregoing years of the Middle Ages, there was little true understanding of mental retardation and limited acknowledgment that mental retardation was separate from other disorders. Instances of formal recognition of mental retardation as a general disorder were sparse.

During the 17th and 18th centuries, the Enlightenment period, advances in medicine as well as the philosophical thought of that time played important roles in the conceptualization of mental retardation. In the medical field, many physicians wrote about mental retardation, epilepsy, hydrocephalus, and cretinism. Locke was the first to distinguish between "idiocy" and "insanity," and he and Rousseau advocated for the influences of the environment on impacting the outcome of human capacity. This cultural enlightenment served as a catalyst to the scientific study of mental retardation and the development of more humane treatment programs.

The Age of Progress (1800–1899), championed by the leadership of French physicians and educators, brought the formal emergence of the construct that today we refer to as mental retardation. It was during this era that the variability in cognitive functioning began to be appreciated, and this increased understanding contributed to the development of treatment programs. The National Institutes for Deaf-Mutes in France hired Jean-Marc Gaspard Itard in 1800 to work with Victor, a feral child found in the woods. Itard developed a broad educational program that focused on Victor's senses, intellect, and emotions. Itard believed that Victor, who by today's standards would most likely be classified as moderately or severely retarded, had not been born in his present mental state, but that his intellectual deficiencies were a consequence of the absence of sensory experiences in a socialized environment. Itard worked diligently with Victor for 5 years but considered himself a failure since Victor was not fully remediable. However, the benefits of his work, such as the thorough recounting of Victor's mental processes and the specifications of the education program, advanced the scientific study and humane treatment of persons with mental retardation.

During this time period (circa 1830) another Frenchman, Esquirol, provided the first classification scheme. He identified two levels of mental retardation that were distinguished by the extent of both cognitive deficiencies and coping impairments: idiot and imbecile. These terms are approximately synonymous with the more current terminology of moderate and severe mental retardation, respectively.

Edward Seguin, a French physician and educator, also had a significant impact on the field by his endeavoring to aggrandize treatment efforts and education programs. He was convinced all persons deserved to be educated and that this was a universal right. Seguin believed that society had an obligation to improve the circumstances of all its members, and those individuals with mental retardation were among the most needy. As such, Sequin developed gradated care based on one's level of need: full supportive care for the most severe and protective care coupled with training for the others. Seguin was later appointed head of the Pennsylvania Training School for Idiotic and Feeble-Minded Children (later renamed Pennsylvania Training School, Elwyn School, and then the Elwyn Institute). By 1857 the population at the

Pennsylvania Training School was 175 students, and less than 50 years later the Pennsylvania Training School served 1041 children and adults and had 165 staff members. By the late 19th century, public school classes for what we would call mild mental retardation were established in all major cities in the United States.

Considerable understanding in the identification of clinical conditions associated with mental retardation was gained in the 1900s. Among the new disabilities that were identified and classified were von Recklinghausen's disease, Tay–Sachs disease, Sturge–Weber syndrome, and Down syndrome (mongoloidism). J. Langston Down made a significant contribution to the field by being the first to provide a separate classification and description of mongoloidism. He proposed that idiocy could be classified into three categories: congenital (idiots), accidental (feeble-minded and idiots), and developmental (feeble-minded). In addition, the Association of Medical Officers of American Institutions for Idiotic and Feeble Minded Persons, a forerunner to the current American Association on Mental Retardation, was established.

The 1900s also brought significant changes in the care of persons with mental retardation. In the early 1900s, Binet's method of assessing intelligence was widely used in the evaluation of persons with mental retardation. Clearly, this had an immense impact on the conceptualization and operationalization of mental retardation via redefining the definitional criteria and related diagnostic procedures. However, the prevailing view in the early 1900s quickly changed to a belief that society needed to be protected from those with mental retardation. This resulted in decisions pertaining to the need for sterilization and the reluctance to accept immigrants with mental retardation. Despite this belief, a number of definitional changes occurred during this time. Edgar Doll advocated one significant change. Doll notably impacted the field with his discernment of social competence and its need to be included in the definition of mental retardation. Another notable change challenged the belief that an institutional setting was the only appropriate setting for treatment. This challenge contributed to the deinstitutionalization movement and furthered discussions regarding the need for increased community-based treatments.

III. DEFINITION OF MENTAL RETARDATION

The definition of mental retardation has been modified several times throughout the years. For the sake of brevity, the numerous reconceptualizations and recapitulations of the definition of mental retardation are not discussed in full. Despite variability in definition, the definitions have held constant with the inclusion of impairments in cognitive and adaptive abilities. One of the major changes in definition occurred in the 1950s when a limitation in adaptive behavior was included as a necessary criterion for the diagnosis. Prior to this change, having a low IQ was sufficient for a diagnosis of mental retardation. Earlier definitions included varying levels of mental retardation, mostly based on the IQ score. For example, a person with an IQ ranging from 35–40 to 50–55 was considered moderately retarded.

In 1992, the American Association on Mental Retardation (AAMR) defined mental retardation as referring to substantial limitations in present functioning. It is characterized by significantly subaverage intellectual functioning, existing concurrently with related limitations in two or more of the following applicable adaptive skill areas: communication, self-care, home living, social skills, community use, self-direction, health and safety, functional academics, leisure, and work. An age limit criterion (e.g., mental retardation manifesting before age 18) also was included in this definition.

In applying the definition, the AAMR asserted four assumptions that are essential to the application of the definition. First, a valid assessment is critical and should consider cultural, communication, behavioral, and linguistic diversity. Second, the adaptive skill limitations should occur within community environments typical of the individual's age peers (e.g., work setting and school environment) and be part of determining the individual's service needs and supports. Third, specific adaptive limitations may be present concurrently with areas of strengths in other adaptive areas. Fourth, appropriate supports provided over a sustained period will most often lead to improvements in the life functioning of a person with mental retardation. This definition was designed to reduce professionals' reliance on the IQ score for measuring the severity of the disability, and it focused on using associated adaptive skills as a means of determining appropriate services and supports necessary to maximize functioning.

When compared to the 1983 definition of mental retardation, several changes are evident. Previously, adaptive behavior had only been roughly defined; however, with the 1992 definition, 10 specific areas of adaptive skills were identified and well-defined. Second, the new definition was to be more functional in its

nature. This definition emphasized the interplay between three dimensions: one's capabilities (e.g., intelligence and adaptive behaviors), the environments in which the person functions (e.g., home, work, school, and community), and the need for varying levels of support that may change over the life span. The 1992 definition provided for a means to classify individuals in terms of level of necessary support (e.g., intermittent, limited, extensive, or pervasive) rather than only in terms of level of severity, such a mild, moderate, severe, or profound. It was hoped that the use of the categorical levels of mental retardation would be discontinued; however, to date these levels of support have not been widely acknowledged or applied. Finally, this definition allowed professionals a means to develop a profile of needed supports based on intellectual functioning and adaptive skills, psychological considerations, health factors, and environmental circumstances. Overall, the 1992 definition has been groundbreaking in its nature and theoretical underpinnings; however, it has not been without its critics, with most of the criticism focused on the discontinuation of the use of the levels of severity (e.g., intellectual levels).

In an effort to continually advance the field of mental retardation, the AAMR has prepared a new definition of mental retardation. This new definition was not formally published at the time of publication of this article; however, a proposed definition has been presented [AAMR (2001)] and is reviewed here. As a whole this proposed 2002 definition builds on the 1992 definition. This proposed definition states that mental retardation is a disability characterized by significant limitations in both intellectual functioning and conceptual, social, and practical adaptive skills. This disability originates before age 18.

There are several noted similarities between the 1992 definition and the proposed 2002 definition including: (a) the use of a functional orientation, (b) intellectual functioning, adaptive behavior, and age of the onset as diagnostic domains, and (c) a dedication to the notion that the level of needed supports for an individual should be the primary focus of a classification system.

The central differences in the proposed 2002 definition as compared to the 1992 definition include: (a) an additional standard deviation criterion within the intellectual domain, (b) the additional dimension of participation, interaction, and social roles as a means of adding to the multidimensional approach to mental retardation, (c) clarification of adaptive behavior which is encompassed in conceptual, social and practical skills that representing one's typical performance, and (d) reconfiguration of assessment of support and determination of intensity of need.

At the risk of perpetuating the earlier definition of mental retardation, this article uses the 1983 nomenclature. The reasons for the continued use of the earlier definition of mental retardation are threefold. First, the nomenclature of the 1992 definition that emphasizes level of support has not been widely used in clinical and research practices. As such, discussion of the levels of retardation is inherently found in the literature describing this population, and it continues to be the most familiar to researchers, clinicians, and other professionals. Second, the proposed levels of support do not clearly align with the accepted nosology often used in practice (e.g., *DSM-IV* and *ICD-10*). Finally, for the purposes of this article, using the more commonly accepted and understood levels of retardation will aid in the ease of interpretation and the conveying of information. Despite our choice to lean toward communicative ease, those professionals working with individuals with mental retardation should become familiar with the most recent definition and its intended clinical use.

IV. FEDERAL LAWS AND MENTAL RETARDATION

Although there are several federal and state laws that directly or indirectly address the rights of persons with mental retardation and other disabilities, three will be highlighted here. Each of these has one main focus: to protect persons with disabilities, including those with mental retardation, from discrimination due to their disability. These laws apply to different institutions that are affected by each statue; however, their overall goal is the same. Landmark legislation to be reviewed include Public Law 94-142, the Education for All Handicapped Children Act, renamed and reauthorized as the Individuals with Disabilities Education Act (IDEA) of 1997; Section 504 of the Rehabilitation Act; and the Americans with Disabilities Act.

A. Individuals with Disabilities Education Act

In 1975 Congress passed the Education for All Handicapped Children Act (PL 94-142). PL 94-142 was revised in 1990 and renamed the Individuals with Disabilities Education Act. It was reauthorized in 1997. This legislation provides for the rights of all

children (ages 3–21) to be educated, despite their disabilities, and for this education to be done in a way that addresses children's specific needs. The main thrust of IDEA is a mandate to states to provide a "free and appropriate public education" in the "least restrictive environment" while protecting a child's rights. To determine eligibility, a student must be evaluated, typically by his or her school district, to determine if he or she is eligible to receive special education and related services. If a child is deemed eligible, the next step is for an Individualized Education Plan (IEP) to be written, which in effect becomes a contract for services between the parents and the school system. This statute provided for nondiscriminatory testing and required parent participation.

In addition to mandating services for children over the age of 3, Part C of IDEA makes available services for children from birth up to age 3. Infant–toddler services within IDEA recognized the need to augment the development of children with disabilities, minimize the need for later special education via early intervention, capitalize on one's ability to live independently, and to improve the capacity of a family to meet the needs of their child.

The reauthorization in 1997 addressed several issues, including regular education teachers as IEP team members, graduation, discipline, children with attention problems, reevaluation practices, public charter schools, parentally placed children in private schools, and use of the "developmental delay" category. The latter change has specific implications as related to mental retardation. The new legislation mandated that a child could receive a diagnosis of developmentally delayed through age 9. Previously, the age limit was 5 years. This change in the legislation may serve to shift the pattern of identification of mental retardation. That is, it may be that numerous children with mental retardation may be labeled with the possibly less stigmatizing label of developmentally delayed, thus decreasing the reported prevalence of mental retardation before age 9.

B. Section 504 of the Rehabilitation Act of 1973

Section 504 of the Rehabilitation Act of 1973 applies to beneficiaries of any kind who receive federal funding, such as schools (elementary, secondary, and universities) and hospitals. Section 504 provides for prohibition of denying persons with disabilities goods, services, facilities, privileges, advantages, or accommodations that are available to those without disabilities. In addition, agencies or groups subject to Section 504 must adjust their work requirements and eligibility standards to accommodate the needs of persons with disabilities. However, Section 504 only protects an individual if the individual is deemed to be otherwise qualified and able to meet the essential eligibility requirements for the job, program, or activity. For individuals with mental retardation, Section 504 undoubtedly has created educational and vocational opportunities that would not have been possible without this legislation.

C. The Americans with Disabilities Act of 1990

The scope of the Americans with Disabilities Act of 1990 (ADA) is more expansive than either of the previously discussed landmark legislative decisions and has been called by some an "emancipation proclamation" for persons with various disabilities. The ADA borrowed heavily from the Rehabilitation Act of 1973 and built on the rights described in the earlier act. The groups of persons protected are virtually the same since both acts share the same definition of disability, but the ADA provides a broader scope of privileges because it is not limited to the purview of agencies that are funded by the federal government. The ADA mandates nondiscrimination in the areas of public entities, public accommodations, employment, architecture, transportation, and communication barriers. As with other legislative acts, the ADA likely has contributed to improving the quality of life and increased the number of opportunities for individuals with mental retardation.

V. CAUSES OF MENTAL RETARDATION AND PREVENTION EFFORTS

The etiology of mental retardation is variable and complex. In fact, there are more than 350 known disorders and conditions, both genetic and acquired, that can result in mental retardation at different developmental stages. To understand the etiology of mental retardation, it is helpful to consider biological and environmental events that may occur prenatally, perinatally, or postnatally.

Before we examine specific causative agents, it is important to note that in general the estimates of the etiology vary based on the level of severity. Mild

retardation is most commonly of idiopathic origins, whereas severe mental retardation most often results from prenatal or genetic factors such as chromosomal abnormalities. For example, for persons with mild retardation, approximately 45–60% of these cases are due to unknown reasons; approximately 10–25% are due to some prenatal, nonchromosomal cause; 4–8% are due to a chromosomal disorder; and 20% of cases can be traced to a perinatal or postnatal cause. When we consider severe mental retardation, as many as 40–70% of cases are due to chromosomal disorders or other prenatal factors. Less than 20% of cases are due to perinatal (e.g., prolonged birth anoxia) or postnatal factors, whereas 25–40% of etiologies are unknown.

A. Prenatal Etiology

Genetic factors have been shown to be one of the most common causes of mental retardation. Chromosomal abnormalities are either heritable mutations or variations of genetic material. Advances in medical technology have furthered our knowledge of the role of genetic transmission of many conditions, including mental retardation. Chromosomal abnormalities may involve either single gene disorders (autosomal-dominant syndromes or autosomal-recessive syndromes) or disorders of the sex chromosomes. Chromosomal disorders that can lead to varying degrees of mental retardation and other cognitive and behavioral characteristics include phenylketonuria (PKU), Turner's syndrome, Lesch–Nylan syndrome, Tay–Sachs disease, maple syrup urine disease, *cri du chat* syndrome, and Klinefelter syndrome.

Numerous agents can have significant deleterious effects on the fragile central nervous system of a child *in utero*. Such teratogens are nongenetic, nonchromosomal agents that are major causes of mental subnormality. These include poor nutrition, toxic substances, maternal disease or infection, blood incompatibility, drugs and alcohol exposure, and cigarettes.

For example, maternal alcohol abuse during pregnancy can result in fetal alcohol syndrome (FAS) or fetal alcohol effects (FAE). The dangers of ingestion of alcohol during pregnancy have been well-known for years. The teratogenic effects of alcohol and other illicit substances impact the developing, vulnerable brain and cause an interruption in the developmental stages of the body organs, most often the central nervous system. If a child meets all diagnostic criteria, as discussed later, he or she is said to have FAS. A child whose mother drank with some moderation during pregnancy, however, may have a typical physical appearance and milder cognitive impairments with accompanied learning and behavioral problems. Such a child would be noted to have FAE.

The prevalence of FAS is not known; however, it is clearly one of the major causes of mental retardation in the United States. It is estimated that 5000 children are born in the United States with FAS each year, with rates of 1–3 cases of FAS per 1000 births. As a whole, alcohol ingestion while a woman is pregnant has been associated with a spectrum of physical and neurodevelopmental effects on the developing child. The criteria for a diagnosis of FAS include pre- and postnatal growth retardation (e.g., low birth weight), central nervous system abnormalities (e.g., mental retardation), and craniofacial malformations (e.g., microcephaly).

Specifically, 80% of children with FAS are born low birth weight, 70% have feeding problems as infants, and most of the children are thin and short in stature. Characteristic physical features include microcephaly, widely space eyes with narrow eyelids, short and upturned noses, thin upper lips, large low-set ears, and a flattened philtrum. Vision problems, such as strabismus or nystagmus, also may be present. In addition to the noted physical and cognitive abnormalities, about two-thirds of children with FAS manifest significant behavioral and emotional problems, including oppositional and defiant behaviors, inappropriate response to social cues, social withdrawal, mood instability, and other overt (e.g., aggression) and covert (e.g., lying and stealing) psychopathology.

B. Perinatal Etiology

The perinatal stage defines the time period surrounding the birth process (e.g., time of birth ± 7 days). During this period, several obstetric complications may occur that can place a child at increased risk of having mental retardation. Perinatal factors account for approximately 20% of cases of mental retardation. These include prematurity and low birth weight. Prematurity and low birth weight are factors highly correlated with the presence of mental retardation; however, prematurity may not be the specific casual agent in and of itself. Rather, prematurity may be the outcome of several other risk factors that contribute to the manifestation of mental retardation. That is, the premature infant is born with more biological and environmental risk factors and likely will have more

risks across time. Other possible perinatal complications include anoxia, breech or other atypical presentation for delivery, birth injury, or difficulty with delivery (e.g., umbilical cord entanglement).

C. Postnatal Etiology

Overall, postnatal etiology of mental retardation is implicated less often than prenatal and perinatal conditions. Nonetheless, many conditions in these early years can lead to mental retardation. In fact, it has been estimated that between 5 and 20% of cases of mental retardation are a result of trauma or neglect. Postnatal causes include traumatic brain injury, cerebral infections (e.g., meningitis and encephalitis), child abuse (e.g., shaken baby syndrome), lead poisoning, and nutritional deficiencies. Childhood head injuries can be a result of various agents (e.g., falls, motor vehicle accidents, bicycle accidents, gunshot wounds, and sports-related injuries). Those at highest risk of a brain injury are boys between the ages of 15 and 18 years and children younger than 5 years of age.

Environmental etiologies are most often related to the presence of mild mental retardation. A large percentage of persons with mental retardation are affected by psychosocial retardation due to economic disadvantage such as poverty. This group includes those who have a diagnosis of mental retardation that did not result from a distinctive, identifiable organic or genetic anomaly.

Environmental risk factors work in combination to impact outcomes of children who do not have any identifiable biological risk but for whom early life experiences are sufficiently limiting that they impart a high probability for delayed development. Poverty has pervasive effects on a child, the family, and the environment. Poverty does not equate to defining a certain cultural group, but poverty is pervasive across cultures. However, not all characteristics described will be present in each economically disadvantaged home situation. In addition, it is important to note that many children reared in impoverished homes function within normal intellectual levels.

Many years of study have enabled researchers to identify causative factors within the impoverished environments that may directly or indirectly impact developmental outcomes. Intergenerational patterns of dysfunction can affect the cognitive development of children. In the case of poverty, the environment may be analogous to poor nutrition, health care, self-esteem, limited parenting skills due to young maternal age or lack of education of the parents, and limited educational and vocational opportunities. The home environment may be overcrowded, understimulating, unpredictable, or provide an excessive amount or inappropriate type of stimulation. In addition, disadvantaged families may not be able to provide basic survival necessities such as food or adequate shelter. Family economics often result in low parental education, low social status of the family, and low parental expectations, which all can negatively impact development resulting in cognitive and adaptive behavior limitations.

D. Prevention Efforts

In the past 25 years, efforts to eliminate or reduce cases of mental retardation have been marked. Prevention has occurred across all three levels: primary, secondary, and tertiary. In short, primary prevention involves efforts to intervene such that one eradicates or minimizes the risk factors before a disabling condition develops. Primary prevention efforts often comprise the backbone of public health programs. Secondary prevention comprises efforts toward the early detection, diagnosis, and treatment such that the duration of the disease is shortened, secondary conditions are reduced, and discomfort is minimized. To achieve this goal, professionals must recognize the precursors of a disease or condition and intervene to limit its progression. For example, neonatal intensive care units act as an agent to limit the numerous consequences of premature births. For those who have an identified disability or disease, tertiary prevention is necessary. Tertiary prevention involves promoting maximal functioning among persons with identified disorders or diseases. The intervention efforts serve to provide for functionality or normalization within the least restrictive environment. One example is the array of community-based life skills training programs for persons with mental retardation.

The goal of prevention, despite successes, remains a daunting one due to multifactorial etiologies, competition for resources, and lack of client motivation and/or empowerment for participation. Another challenge specific to the outcome of mental retardation is the long latency period between exposure to causal factors and the diagnosis of mental retardation. This latency period presents a challenge for determining the etiology of mental retardation (and other developmental disabilities). However, developmental epidemiological studies can assist in the challenge of dealing

with the latency period between exposure to causal factors and the manifestations of a developmental outcome. This can be accomplished by using linked extant databases to examine causative and protective factors and their relation to the outcome of interest. This methodology effectively creates a longitudinal file for each individual for which all data points are available without having to wait multiple years for the outcome (e.g., mental retardation) to be manifested.

To date, prevention efforts have proven successful in eliminating congenital rubella by immunization and antibody screenings; eliminating retardation due to PKU, galactosemia, and congenital hypothyroidism by newborn screenings and dietary management; and eliminating kernicterus by the use of globulin therapy. Prevention efforts have also greatly reduced morbidity from prematurity through neonatal intensive care nurseries, the presence of measles through vaccination, and Tay–Sachs disease through early screening and prenatal diagnosis. In addition, efforts have proven effective in reducing the incidence of neural tube defects by folic acid supplementation, fetal alcohol syndrome through public awareness and education, lead poisoning by environmental improvements and lead screenings, traumatic brain injury by using child restraints in automobiles, and child abuse and neglect through family education and supports. Finally, improvements have also occurred via early identification and intervention of young children at risk for developmental delays. With such efforts, various medical, allied health, and early intervention professionals have been effective in providing earlier identification and treatment for individuals with mental retardation.

1. Early Childhood Intervention

The provision of early childhood intervention services involves systematic efforts to provide at-risk children with additional educational and developmental experiences before school age. It also includes the provision of family services. Several investigations have attempted to quantify the impact of early intervention on the cognitive development of young children from disadvantaged backgrounds (e.g., Early Training Project, Perry Preschool Project, Houston Parent Child Development Center, Milwaukee Project, Carolina Abecedarian Project, and Project Care) and those with biological risk factors (e.g., low birth weight—the Infant Health and Development Project). Although these research projects defined disadvantage in slightly different ways, initiated intervention at varying ages, and included different intervention components, their collective results provide evidence that high-quality early intervention services prevent or minimize the potentially harmful effects of various environmental and biological risk factors for young disadvantaged children. These programs provide short-term benefits in the areas of cognitive functioning and long-term effects on decreased grade retention and special education placement. In addition, evidence from a few studies suggests that programs with the greatest educational intensity and duration maintain cognitive functioning benefits until adulthood. The positive cost-benefit ratio for society should not be underestimated in this regard.

In general, programs that begin earlier and continue longer afford greater cognitive functioning benefits than those that begin later and have shorter durations. Intervention studies that begin during infancy and preschool years, before at-risk children develop increasingly depressed cognitive/developmental scores, have demonstrated the largest impact on cognitive level of functioning. These intervention programs have shown that participating children continue to score in the average range on cognitive/developmental tests, whereas children not receiving the full intervention score in the low-average range. Programs that began at later ages, approximately 4 or 5, have shown that upon entry children's cognitive scores have already begun to decline. After 1 or 2 years of intervention, children's cognitive scores are higher, whereas the scores of children who did not receive the intervention have not changed. Although these types of interventions may lessen the morbidity associated with moderate and severe levels of mental retardation, they stand to impact most significantly on individuals with mild mental retardation with idiopathic etiologies.

VI. PREVALENCE OF MENTAL RETARDATION

It is estimated that people with mental retardation comprise approximately 1–3% of the population. However, these data vary based on the method of case ascertainment, definition used, and population studied. The prevalence of mental retardation has not changed appreciably for more than 60 years as a result of the stasis between advancing health care and the emergence of new cases. For example, although neonatal diagnosis of children with phenylketonuria and subsequent management has theoretically resulted in the elimination of mental retardation caused by this metabolic disorder, there has been an increase in

prenatal exposure to toxic substances (e.g., drugs and alcohol) and an increased survival rate of very low or extremely low-birth-weight, premature infants. Therefore, this balancing effect has contributed to static prevalence rates of mental retardation throughout the years.

As noted, mental retardation comprises a heterogeneous group of persons and syndromes with dissimilar prevalence, physical manifestations, cognitive abilities, and social–emotional development. Previously, all persons with mental retardation were seen as a homogeneous group of individuals. However, we now have a significant knowledge base that allows us to understand more about selective disorders. As such, we consider relevant issues within specific syndromes that are associated with mental retardation.

VII. CHARACTERISTICS OF SYNDROMES ASSOCIATED WITH MENTAL RETARDATION

A. Down Syndrome

Down syndrome (DS) is easily recognized by the lay population and perhaps thought to be synonymous with mental retardation. It is one of the most common genetic causes of mental retardation and perhaps the best known. It is estimated that DS occurs about 1 to 1.5 times every 1000 births. However, other estimates reveal DS may occur once in every 600 births. The risk of having a child with DS increases as the age of the mother increases. It is estimated that males are affected slightly more than females (1.3 to 1.0), which may be due, in part, to a higher mortality rate in females during infancy.

The most common cause of DS involves an extra chromosome existing at the 21st position, known as nondisjunction or trisomy 21. Nondisjunction results from an inappropriate separation of chromosomes during meiotic division that results in a total of 47 chromosomes. This accounts for approximately 95% of the cases of DS. However, DS can also be caused by the chromosomal abnormalities of translocation or mosacism. Translocation (responsible for 2–4% of cases of DS) occurs when part of a chromosome is attached or translocated to another location or chromosome, often to chromosome 14. Mosaicism (responsible for 1–4% of cases of DS) occurs when there is an uneven pattern of dissimilar cells with 46 or 47 chromosomes.

DS results in a cluster of physical traits that include short stature; flat, broad face with small ears and nose; brachycephaly; short, broad hands and feet; tongue protrusion and/or fissured tongue; heart defects; hypotonia; slanted and almond-shaped eyes; epicanthic folds in the inside corners of the eyes; a wide space between the first and second toe; and a single crease that transverses the palm. These physical traits can vary from individual to individual and may become less pronounced as an individual grows and develops.

Most individuals with DS often have some degree of intellectual impairment, with mild to moderate mental retardation being most prevalent. As children with DS age, there seems to be a general decline in their developmental rate and progression, with verbal short-term memory skills being most vulnerable. Some research has suggested that individuals with mosaicism have higher average cognitive scores (10–30 points higher) than individuals with trisomy 21 on standardized measures of IQ. Children with mosaicism have also demonstrated average visual-perception skills.

The accepted stereotype of persons with DS is that they are happy, social, and friendly; however, this is not scientifically based. Instead, studies have shown that children with DS have temperament profiles that are comparable to those of typically developing children. Children with DS have difficulty in the area of social competence, perhaps due to concomitant motor, cognitive, and language delays. Individuals with DS exhibit other comorbid disorders, including attention deficit/hyperactivity disorder, conduct/oppositional disorder, aggression, phobias, stereotypic behavior and self-injurious behaviors.

B. Fragile X Syndrome

Fragile X syndrome (FXS) is the most common inherited cause of mental retardation. FXS occurs in approximately 1 in 1500 male births and 1 in 1000 female births. Subsequent studies using DNA have demonstrated a lower prevalence of FXS at approximately 1 per 4000 males in the general population. It is a single gene disorder on the X chromosome that occurs in both males and females, but males are typically affected more severely. It is an X-linked disorder and results from a pinched, restricted, or "fragile" location on the long arm of the X chromosome. In 1991, the fragile X mental retardation 1 (FMR1) gene was discovered. With the discovery of FMR1, DNA testing for the presence of fragile X was available. DNA testing provides a more effective means to detect the presence of FXS and carriers of the FMR1 gene by specifically showing the genetic

mutation at FMR1. Female carriers can pass down the disorder through generations. FXS minimizes the function of the gene responsible for producing an essential protein necessary for normal brain function.

Because of the divergent impact on males and females, there is a significant degree of phenotypic variability expressed in this disorder. In fact, early in life, there is a limited physical phenotype with a distinctive appearance developing at approximately the time of puberty. That is, males with FXS may have macrocephaly, elongated face, long protruding ears, prominent jaw, macroorchidism in adulthood, large nose, and a prominent forehead. Early in life, the one physical characteristic that may be present is the large ears. Females typically have only a slightly abnormal appearance, with approximately half presenting with protruding ears and other subtle features, such as flexible joints, low muscle tone, or flat feet.

Males and females show disparate patterns of cognitive profiles. Males most often have IQ scores in the moderate to severe range of mental retardation, with some scatter to the borderline to low-average range of intellectual functioning. Behavioral manifestations also include attentional difficulties, hyperactivity, tactile defensiveness, hand flapping, seizures, stereotypic behavior, echolalia, and gaze avoidance. Females with FXS are often carriers with little or no expression of the condition they may often be diagnosed as having a learning disability rather than mental retardation.

C. The Mucopolysaccharidoses

The mucopolysaccharidoses (MPS disorders) are a group of autosomal-recessive genetic abnormalities of mucopolysaccharide metabolism that results in variable impairment across physical, sensory, cognitive, and behavioral domains. There are six main types of MPS, many of which include specific subtypes. Prevalence estimates vary based on the subtype of MPS disorders and geographic location. Prevalence estimates range from 1 in 25,000 (Sanfilippo syndrome in the United Kingdom) to 3 in 1 million persons (Morquio syndrome).

Often, MPS disorder results in impairment of the skeletal system. Other organ systems that are impacted include the cardiovascular, respiratory, and nervous systems. Skeletal impairments include widened collarbone and ribs, progressive curvature of the lower spine, short stature, shortened neck, claw-like hands, joint contractures, enlarged head, flattened bridge of the nose, protruding tongue, and thick hair coupled with excessive body hair.

Intelligence varies within subtypes of MPS from average to progressive mental retardation. Behavioral issues also vary within subtypes of the MPS disorders. Individuals with Hurler syndrome, for example, are described as being anxious, restless, and having sleep problems but are not aggressive. Children with Hunter syndrome are overactive, aggressive/destructive, fearful, defiant, and also have sleep problems. Children with Sanfilippo syndrome manifest the most noticeable behavioral symptoms. These children often wander aimlessly, are restless, mouth clothing and objects, have sleep problems, and can be aggressive even when not provoked.

D. Noonan Syndrome

Prevalence estimates of Noonan syndrome (NS) range from 1 in 1000 births to 1 in 2500 births. Males and females are equally affected, and no racial or geographical differences have been reported. Physical manifestations of NS include craniofacial abnormalities (e.g., small jaw, low-set ears that are angled toward the posterior, a deep groove in the philtrum, and a high arched palate), congenital heart disease, skeletal abnormalities (e.g., short stature, webbed neck, and abnormalities of the thorax), and genital malformations. Intellectual functioning ranges from mild mental retardation to superior intellectual levels.

Although further research is necessary regarding the social–behavioral aspects of persons with NS, it is known that social interactions may be deficient because of other concomitant disorders (e.g., low muscle tone impacts athletic participation). Overall, children with NS seem to mature more slowly than their typically developing peers, and they often prefer to play with younger children. No evidence of behavioral or psychiatric disorders comorbid with NS has been reported.

E. Klinefelter Syndrome

Klinefelter syndrome (KS) is a sex chromosome disorder and occurs due to the presence of an extra X chromosome in males (XXY). It is estimated that 1 in 700 to 1 in 900 live male births are affected by KS. Early research linked KS with psychiatric disorders, criminal behavior, and mental retardation. These studies had methodological concerns that impacted

validity; however, many still hold to these early findings. In truth, males with KS are at risk for developmental, learning, language, and behavioral problems along with psychiatric disorders, but they are not inherently criminal.

Physical manifestations of KS result in impairment of normal genital and sexual development. Although persons with KS enter puberty at the typical age, inadequate testosterone levels prevent normal pubertal progress. Hypogonadism is present with a slowed development of or lack of secondary sexual characteristics. In addition, KS leads to smaller than normal testes that often contribute to infertility. Cognitive abilities range from mild mental retardation to above-average intelligence. In general, persons with KS have deficits in verbal abilities, whereas nonverbal abilities are typically in the average range. Behaviorally, individuals with KS may be reserved, withdrawn, and immature, and they are at risk for having poor peer relationships. Self-esteem problems also have been described. In contrast, they also have been described as being easygoing, underactive, compliant, and, consequently, well liked by teachers.

F. Lesch–Nyhan Syndrome

Lesch–Nyhan syndrome (LNS) is a rare X-linked recessive disorder that typically only impacts males, whereas females act as carriers. Research has found that approximately 1 in every 380,000 births is affected, and no ethnic groups or individuals in different geographic regions are known to be at greater risk. Individuals with LNS suffer from dystonia (including hypotonia and hypertonia), spasticity of movement, speech deficits, renal disease, gout, and mental retardation. Motor impairments are severe and often restrict an individual to use of a wheelchair with necessary physical supports. In addition, motor abilities impact the ability to produce intelligible speech and perform self-care tasks independently.

Perhaps one of the most distinguishing behaviors of LNS is that nearly all individuals with LNS demonstrate severe self-injurious/self-mutilating behaviors that can begin as early as 26 months of age. These behaviors escalate with age. The self-mutilation often begins with repeated biting of the lips, hands, and fingers. The severity of self-mutilation is such that permanent tissue loss often results (e.g., self-amputation of fingers or tongue).

Cognitive assessment of individuals with LNS is challenging due to their physical limitations and self-injurious behaviors. Findings suggest that persons with LNS have varying degrees of cognitive impairments ranging from mild to severe, but it is widely held that cognitive skills may actually be higher than measured due to assessment limitations.

G. Prader–Willie Syndrome

Prader–Willie syndrome (PWS) occurs evenly across both gender and ethnic lines. Approximately 1 individual in 10,000–15,000 births is impacted. PWS is caused by a microdeletion of chromosomal matter on chromosome 15. One unique feature of PWS is that genes are expressed in a divergent manner depending on whether one inherited them from one's mother or one's father. Physical manifestations include failure to thrive (early in development), characteristic facial features (e.g., narrow nasal bridge, downturned mouth, almond-shaped eyes, and thin upper lip), hypotonia, hypogonadism, later insatiable appetite and resulting obesity, and a specific behavioral phenotype.

The behavioral phenotype includes impulsivity, aggression, obsessive behaviors, underactivity, anxiety, daytime sleepiness, and perseverative tendencies. In addition, other maladaptive behaviors and psychopathology are present in persons with PWS, including tantrums, mood changes, argumentativeness, stubbornness, oppositional behaviors, theft, and rigidity. Clearly, managing obsessive food-seeking behaviors that occur in clever and manipulative manners poses a substantial challenge for those working with and caring for these individuals. Most persons with PWS are functioning within the mild range of mental retardation, with an average IQ score estimated to be approximately 70. Individual variation has been noted, with scatter from the average level to the profound range of mental retardation. In general, cognitive abilities seem to plateau in adolescence or early adulthood. Adaptive behavior scores are generally reported as being lower than scores for their level of cognitive functioning, with steady gains in socialization and daily livings skills as these individuals grow and develop.

H. Williams Syndrome

Approximately 1 in 20,000 live births results in a child with Williams syndrome (WS), and it is equally distributed across ethnic groups. WS is a multisystem disorder that includes impairments in cognition,

congenital heart disease, dysmorphic facial features, and connective tissue problems that impact the bowel, bladder, skin, and joints. Individuals with WS also have distinctive personality characteristics. Diagnosis in infancy most often is a result of a triad of symptoms: developmental delay, severe cardiovascular diseases, and dysmorphic facial features. In early adolescence, it is often diagnosed based on a specific cognitive profile and personality characteristics.

Characteristic facial features include a flat profile, a broad brow with bitemporal narrowing, a wide mouth with full lips, full cheeks, upturned nose with a bulbous nasal tip, asymmetry in the face, and a long philtrum. Children with WS often have chronic otitis media as well as ocular and visual abnormalities. Additionally, abnormalities in the respiratory system often lead to an idiosyncratic voice pattern. This voice pattern is described as being hoarse, low pitched, and flat. Complications of the renal, gastrointestinal, musculoskeletal, and neurological systems are present as well.

Individuals with WS function across varying levels of intellectual abilities. Overall cognitive abilities range from low average to severe mental retardation, with most persons with WS categorized in the mild range of mental retardation. Children with WS typically score better on measures of verbal abilities compared to nonverbal (visual–spatial) abilities. In fact, their visual–spatial abilities often represent extreme weaknesses in their cognitive profile. The cognitive profile for persons with WS also includes auditory rote memory skills that are slightly better what would be expected based on their overall cognitive performance, and these abilities may contribute to their apparently higher verbal abilities.

Children with WS often have attentional problems, most often due to overactivity and distractibility. Thought problems and social problems have also been reported. Research on the personality characteristics of individuals with WS has shown that these individuals are intense, anxious, curious, less reserved toward strangers, overfriendly, eager to learn, and oversensitive to sound. In addition, individuals with WS have negative moods, have a lower threshold of excitement, and have difficulty sleeping.

VIII. PSYCHOPATHOLOGY AND MENTAL RETARDATION

During the past 30 years, research studies have clearly established that persons with mental retardation are more likely to experience some form of psychopathology than are persons without disabilities. Indeed, nearly all the disorders reviewed in this article have some type of unusual or distinctive social–behavioral patterns. A complex combination and diverse array of biological, psychological, and social factors interact to contribute to mental health problems in persons with mental retardation. There is unanimous agreement among professionals that persons with mental retardation are at higher risk for developing comorbid psychiatric disorders, but there is no agreement on the actual prevalence of comorbid psychiatric disorders. The prevalence estimates vary from as low as 1% to as high as 79%.

One of the major reasons for this large disparity between these prevalence estimates is due to methodological considerations. Prevalence studies diverge based on definition issues, identification, and sampling methods. That is, the studies vary in differing interpretations regarding how to define mental retardation (e.g., historically diagnosed or IQ cut-off), diagnostic procedures in establishing a formal psychiatric diagnosis (e.g., review of case files, direct assessment of psychopathology or direct care staff or other person acting as a reporter of symptomatology), type of psychopathology being examined, the criteria for establishing psychopathology (e.g., does psychopathology include ''challenging behaviors''), and population sampled (e.g., level of severity and living arrangements of sample). Despite this variability, the fact remains that the full range of psychiatric disorders can be observed within the population of individuals with mental retardation.

Another consideration relates to the heterogeneity of persons with mental retardation. Although this may be tied to the level of function issues, it could also relate to the pattern of function and/or the type of mental retardation manifested. For example, an individual with LNS manifests with a very serious level of self-injurious behaviors, whereas individuals within other types of mental retardation display very little if any self-injurious behaviors. Similarly, even within some disorders, such as FXS, the type and severity of psychopathology can vary by gender.

The third reason for disparity in prevalence estimates relates to the measurement of psychopathology in clients with mental retardation. Because of factors inherent in most individuals with mental retardation, an accurate psychiatric diagnosis may be problematic. Clearly, intellectual limitations may impact one's ability to conceptualize and effectively communicate or describe subjective feelings or objective events. In addition, limited social skills and inflexible thinking

may impact the information provided by the client and, in turn, impact diagnostic status.

In an effort to overcome these issues and thereby obtain a more accurate diagnosis of psychopathology in individuals with mental retardation, several instruments have been developed, including the Psychopathology Instrument for Mentally Retarded Adults (PIMRA), the Reiss Screen for Maladaptive Behavior (RSMB), and the Diagnostic Assessment for the Severely Handicapped (DASH). These instruments focus on the major forms of psychopathology and have been shown to be effective in diagnosing psychopathology in persons with mental retardation. The PIMRA can be administered in a self-report format or an informant version. Direct-care staff or others who are familiar with an individual complete the RSMB. The DASH involves a comprehensive structured survey answered by a direct-care staff member or other familiar individual.

IX. TREATMENT/INTERVENTION APPROACHES

The use of the term treatment can be misleading as related to mental retardation. Treatment usually infers the presence of a cure or a complete remediation of present problems. In the discussion of treatment and mental retardation, one should not interpret treatment in its literal sense but, rather, as a means of providing specific interventions that provide considerable benefit to persons with mental retardation and often target a specific deficit or problem. The main treatment approaches discussed here have received support as empirically based approaches and include counseling services, cognitive–behavioral treatments, and behavioral treatments. However, it is important for the reader to note that there are a number of other treatment approaches available (e.g., pharmacological). Furthermore, the astute reader will recognize that it is important to individualize treatment programs to a client's particular characteristics and needs, particularly with respect to developmentally appropriate practices.

A. Counseling

Clearly, the use of individual or group counseling as a treatment modality for persons with mental retardation has limitations. Persons with mental retardation often have trouble generalizing behaviors, recognizing consistent patterns of behaviors, and generating insight-oriented understandings. However, counseling that is directive and modified has been used effectively in the treatment of educational, vocational, and nonpsychiatric behavioral problems. Modifications should include the use of simplistic language, structured and directive communication, and the use of visual aids when necessary. In general, establishing realistic and attainable goals given the developmental levels and profiles of the client is the primary motivation of the counselor. Specifically, the counselor's goals are often to assist a person in understanding and labeling his or her emotional states, teach problem-solving and decision-making skills, and help the client learn that his or her behaviors result in consequences. Efficacy studies examining the use of counseling strategies for persons with mental retardation have demonstrated positive outcomes.

B. Cognitive–Behavioral Treatments

Cognitive–behavioral treatments have been effective in increasing functioning within the educational and social domains of persons with mental retardation. It has been shown that children with mental retardation are deficient in their ability to use metacognitive skills. As a whole, children with mental retardation use fewer cognitive strategies in memory and learning situations than their peers without mental retardation. In addition, they often lack the abilities to apply learned skills in a new situation. The use of metacognitive training can enhance one's memory and learning skills by teaching self-management skills. For example, a child may be taught to determine what type of math problem is presented to him or her. Subsequently, the child is taught to choose the appropriate strategy and then to evaluate the strategy choice (metacognitive skills).

Along with educational intervention, cognitive–behavioral strategies have been used in the social realm to address peer acceptance and social competence. That is, a person with mild mental retardation who is experiencing social problems can be taught to use metacognitive skills in a given situation. In doing so, one might work toward encoding properly (e.g., think about what happened), interpreting social cues (e.g., what the other person is feeling), and evaluating the possible choices for a response and its consequences.

C. Behavioral Treatments

Behavioral treatments have been shown to be effective in reducing aggression, self-injurious behaviors, and other learned maladaptive behavioral patterns in

persons with mental retardation. In addition, behavioral treatments have improved quality of life via teaching basic living skills. For many years, behavioral strategies have been the most thoroughly researched and widely used group of interventions for persons with mental retardation. These strategies rest on the belief that behavior is regulated by its consequences, and that environmental influences can promote or diminish behaviors. Consequences can increase the probability of a future occurrence of a behavior, or they can reduce the probability of a behavior occurring.

All behavioral therapies involve several key components: operationalizing target behaviors, recognizing antecedents to behaviors, identifying an appropriate array of reinforcers, and specifying when target behaviors will be rewarded or punished. With such knowledge, one can devise appropriate environments in which a person can be reinforced for desirable and adaptive behaviors and not reinforced for maladaptive behaviors.

Behavioral treatments also involve the specification and quantification of the occurrence of the maladaptive behavior in terms of its context, function, frequency, duration, and intensity. Upon determining the specific maladaptive behavior to target and its potential reinforcers, it is the goal of the professional to determine the purpose or function that one's behavior is serving. A functional analysis attempts to determine the relationship between behavior, antecedents, and consequences, which permits the choice of an appropriate treatment. In conducting a functional analysis, information can be obtained that determines whether a behavior is maintained due to antecedents specific to the environment (e.g., interaction with peers and overstimulation) or due to consequences (e.g., desired avoidance of an event).

X. FUTURE RESEARCH

There are three main areas to be addressed as we strive to improve our commitment to providing services for persons with mental retardation. First, continual work on the definition of mental retardation is necessitated by the lack of widespread use of the most recent definition. As noted previously, many clinicians and researches have not adopted the classification put forth in 1992 by the AAMR. Achieving the goal of widespread use of this contemporary language in practice is perhaps an onerous one; which is perhaps furthered by the proposed 2002 definition. Further understanding by practitioners and clinicians regarding how the contemporary nomenclature coincides with historically established taxonomy is essential.

Second, research is needed that provides insight into the relationship between genotypic and phenotypic linkages. Biomedical technology advances, such as recently witnessed via the Human Genome Project, may serve to identify the major genetic causes of mental retardation. As such, once a genetic syndrome is identified, this lends itself to study of how this disorder is expressed phenotypically. Such information allows for more accurate and earlier identification of mental retardation, which in turn allows for the most fitting service provision and supports available (e.g., provision of early intervention services).

Third, research must continue to move away from the study of persons with mental retardation as a homogeneous group. Research must move toward providing detailed descriptions of the cognitive (e.g., neuropsychological), behavioral, and social profiles, including strengths and weaknesses, of a specific group of persons with mental retardation (e.g., persons with mental retardation due to FXS). This line of research will refine clinicians' and service providers' abilities to perform their respective tasks. Clinicians will be provided with excellent tools to diagnose persons with mental retardation accurately and efficiently, and service providers will be able to target the areas of intervention more successfully.

XI. CONCLUSIONS

The field of mental retardation is quite broad, and the full range of disorders, treatment, and related diagnostic issues (e.g., assessment) cannot be realistically addressed within the confines of a single article. We have endeavored, however, to address many of the key issues inherent in this field, with a particular focus on historical foundations, definitional issues, etiological factors, and considerations such as comorbidity and treatment approaches. We also devoted significant discussion to several specific disorders, and it is believed that much of the future work in this field should address more homogeneous groupings of individuals with mental retardation. It is likely that this approach will be more informative for educators, clinicians, therapists, caseworkers, employers, and families. Even with its rich clinical and research history, this field holds great promise for future scientific advancement and improved clinical practice.

See Also the Following Articles

AUTISM • BRAIN DEVELOPMENT • CEREBRAL PALSY • INTELLIGENCE • LANGUAGE AND LEXICAL PROCESSING

Suggested Reading

American Association on Mental Retardation (2001). Request for comments on proposed new edition of *Mental Retardation: Definition, Classification, and Systems of Support*. *News and Notes* **14**(5), 9–12.

American Association of Mental Retardation (1992). *Mental Retardation: Definition, Classification, and Systems of Support*. American Association of Mental Retardation, Washington, DC.

Baroff, G. S., and Olley, J. G. (1999). *Mental Retardation: Nature, Cause, and Management*. Brunner/Mazel, Philadelphia.

Batshaw, M. L., and Shapiro, B. K. (1997). Mental retardation. In *Children with Disabilities* (M. L. Batshaw, Ed.), 4th ed. Brookes, Baltimore.

Detterman, D. K. (1999). The psychology of mental retardation. *Int. Rev. Psychiatr.* **11,** 26–39.

Goldstein, S., and Reynolds, C. R. (1999). *Handbook of Neurodevelopmental and Genetic Disorders in Children*. The Guilford, New York.

Hooper, S. R., Boyd, T. A., Hynd, G. W., and Rubin, J. (1993). Definitional issues and neurobiological foundations of selected severe neurodevelopmental disorders. *Archives Clin. Neuropsych.* **8,** 279–307.

Matson, J. L., and Mulick, J. A. (1991). *Handbook of Mental Retardation*. Pergamon Press, New York.

McLauren, J., and Bryson, S. E. (1987). Review of recent epidemiological studies in mental retardation: Prevalence, associated disorders and etiology. *Am. J. Mental Retardation* **92,** 243–254.

Pulsifer, M. B. (1995). The neuropsychology of mental retardation. *J. Int. Neuropsychol. Soc.* **2,** 159–176.

Redden, S. C., Mulvihill, B. A., Wallander, J., and Hovinga, M. (2000). Applications of developmental epidemiological data linkage methodology to examine early risk for childhood disability. *Dev. Rev.* **20,** 319–349.

Reiss, S. (2000). A mindful approach to mental retardation. *J. Social Issues* **56,** 65–80.

Reschly, D. J. (1992). Mental Retardation: Conceptual Foundations, Definitional Criteria, and Diagnostic Operations. In *Developmental Disorders: Diagnostic Criteria and Clinical Assessment*. (S. Hooper, G. W. Hynd and R. E. Mattison, Eds.) pp. 23–67. Lawrence Erlbaum & Associates, New Jersey.

Rojahn, J., and Tasse, M. J. (1996). Psychopathology in mental retardation. In *Manual of Diagnosis and Professional Practice in Mental Retardation*. (J. W. Jacobson and J. A. Mulick, Eds.). American Psychological Association, Washington, DC.

Scheerenberger, R. C. (1983). *A History of Mental Retardation*. Paul H. Brookes Publishing Company, Baltimore, Maryland.

Mental Workload

RAJA PARASURAMAN and DANIEL CAGGIANO
Catholic University of America

GLOSSARY

central executive A hypothesized control center in the brain that is involved in high-level processes, such as coordination of multiple processes, error correction, and planning.

data-limited processing Cognitive processing that is limited solely by the quality of the input to the process, whether that input is external or drawn from memory.

electroencephalography (EEG) The spontaneous electrical activity of the human brain recorded at the scalp.

endogenous driver An internal source of mental workload; reflects the individual's mental effort, task strategies, proficiency, etc.

event-related potentials (ERPs) Electrical potentials of the human brain (recorded at the scalp) evoked by stimulus and response events and present in the background EEG.

exogenous driver An external (environmental) source of mental workload.

functional magnetic resonance imaging (fMRI) A functional brain imaging technique involving a static magnetic field that is typically used to assess cerebral blood flow in humans.

mental workload A composite brain state that reflects the interaction between the environmental and task demands imposed on an individual and his or her capability to meet those demands.

positron emission tomography (PET) A functional brain imaging technique involving ionizing radiation and computed tomography that is typically used to measure cerebral blood flow and metabolism in humans.

resource-limited processing Cognitive processing whose efficiency is dependent on the amount of information-processing resources allocated to the process.

working memory A mechanism for maintaining information briefly in an active state for use in various information-processing activities.

***Mental workload* refers to a composite brain state or set of** states that mediates human performance of perceptual, cognitive, and motor tasks. Mental workload can be driven *exogenously* (or "bottom-up") by environmental sources, namely, by task load, as well as *endogenously* (or "top-down") by the voluntary application of mental effort.[1] A simple example will illustrate these two sources of mental workload. Mental multiplication of two 2-digit numbers, say 12×13, is challenging but within the capacity of most adults. If three such numbers have to be multiplied, however, say $12 \times 13 \times 14$, then mental workload rises due to the increased task load—an exogenous driver of mental workload. Another exogenous driver is time pressure, the requirement to complete the task within a specified period of time. The workload associated with mental multiplication would be relatively low if one is given unlimited time to complete the task but would be considerably increased if an answer had to be provided in only a few seconds. In addition to exogenous sources, mental workload also reflects how much effort an individual puts into the process of mental multiplication under time pressure—in lay terms, a person can work more or less hard at this task in order to meet the time deadline. More generally, and

[1]This association reveals why the term *mental effort* has been used almost synonymously with mental workload. A highly influential book by Daniel Kahneman, *Attention and Effort*, published in 1973, introduced the concept of mental effort and its physiological concomitants in human autonomic and central nervous system activity. For a variety of reasons, however, the term mental workload is currently more popular and will be used throughout this article.

in a task more complex than mental multiplication, endogenous workload will also reflect the strategies that an individual uses to solve the task. This article describes the structural and functional properties of mental workload in relation to human brain function. Neural signs of mental workload as reflected in electroencephalography (EEG), event-related potentials (ERPs), positron emission tomography (PET), and functional magnetic resonance imaging (fMRI) are described.

I. STRUCTURAL AND FUNCTIONAL CHARACTERISTICS OF MENTAL WORKLOAD

Mental workload has been variously characterized as the objective task demands imposed on an individual, the mental effort he or she exerts to meet these demands, how well the person performs the task, and his or her subjective perception of expended effort. A definition that merges these different views postulates that mental workload is an intervening construct that reflects the interaction between the environmental and task demands imposed on a person and the capabilities of that individual to meet those demands. The intensity of this interaction is reflected in brain states, whose measurement can provide for objective assessment of human mental workload.

A. Mental Work and Brain Work

The notion of mental workload as a reflection of how hard one's mind is working at any given moment is intuitively appealing. Given that the mind is a function of the brain, it follows that mental workload should be associated with brain work. How can brain work be assessed? Various candidate measures have been proposed over the years. The work of Sir Charles Sherrington, the great nineteenth century physiologist, suggested that an answer could be found in the regulation of the blood supply of the brain. Sherrington demonstrated that there is a close coupling between the electrical activity of neuronal cells, the energy demands of the associated cellular processes, and regional blood flow in the brain. His pioneering work suggested that, if mental activity results in increased neuronal response in localized regions of the brain, then in principle it should be possible to measure mental workload by assessing regional cerebral metabolism and blood flow.

Autoradiographic studies conducted in animals have confirmed Sherrington's principle for the regulation of brain blood flow and its coupling to neuronal activity and energy usage. But it would take several years before sensitive techniques were developed for measuring regional brain blood flow in humans. An early development was the invention of the xenon-133 (Xe-133) method for assessing regional cortical changes in brain blood flow and glucose metabolism. Injection of the radioactively tagged xenon gas, which passes freely across the blood–brain barrier, into human patient volunteers showed that the performance of various mental tasks (such as mental arithmetic of the type mentioned at the beginning of this article) led to increased metabolic activity in specific cortical regions.

This was the first demonstration of a link between mental work and brain work in humans. However, the Xe-133 technique was too invasive to be used routinely in normal human subjects. The development of positron emission tomography (PET) paved the way for less invasive measurement of regional cerebral metabolism and blood flow. PET is an adaptation of autoradiographic techniques originally developed for measuring blood flow in animals. Regional cerebral glucose metabolism can be determined noninvasively using PET and radioactively labeled glucose (18-fluorodeoxyglucose), whereas regional cerebral blood flow may be assessed with PET and radioactively labeled oxygen (O-15) in water. PET was also more accurate than the older methods in localizing the specific cortical regions activated by cognitive task demands. Nevertheless, the spatial resolution of PET, particularly in individual subjects, had room for improvement. Furthermore, the need for ionizing radiation, although safe when used within exposure limits, was an impediment against frequent use in studies with normal human subjects. The development of functional magnetic resonance imaging (fMRI) overcame both of these limitations. fMRI provides noninvasive, high-resolution assessment of regional cerebral blood flow. PET and fMRI studies of mental workload are discussed further later.

B. The Structure of Mental Workload: Unitary or Modular?

Brain imaging can provide for objective measurement of human mental workload. Such studies may also illuminate a fundamental issue that has been debated for several years: whether mental workload is unitary

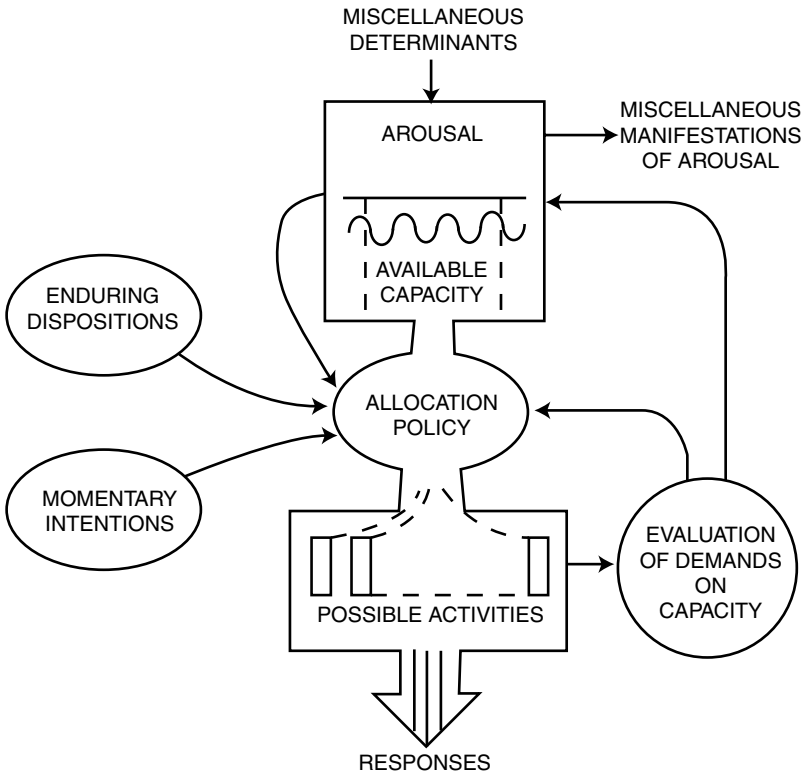


Figure 1 Unitary resource (capacity) model of Kahneman (1973). *Attention and Effort* © Daniel Kahneman. Reprinted by permission of Pearson Education Inc., Upper Saddle River, NJ.

or modular. Cognitive psychologists have addressed this question in some detail in research on human attention. Beginning with Daniel Kahneman's seminal work on attention and effort (1973), mental workload has been linked to information-processing theories of attention. These theories assume that tasks require the allocation of information-processing resources for their efficient execution (except those that can be performed automatically) and that mental workload reflects the overall demand for such resources. Kahneman's resource (or capacity) model is shown in Fig. 1. Every individual has a pool of available resources that is modulated by arousal level, with capacity being low during drowsiness and sleep and high during attentive wakefulness. The available capacity is allocated in parallel to various possible information-processing activities, according to the current behavioral goals of the person.

Some resource theories, like Kahneman's model, assume a single pool of resources that can be flexibly allocated to different processing activities. This is the unitary view. In the modular view, functionally separate resources are applied to different processing activities. Modules may be defined by such factors as sensory modality and stages of information processing, as in a model proposed by Christopher Wickens (see Fig. 2). An influential model of "working memory" by Alan Baddeley represents a hybrid view. This model proposes that modular "slave" systems for verbal and visuospatial processing are controlled by a unitary "central executive" system (see Fig. 3).

According to both the unitary and modular views, the processing resources that are applied to a task (or a component of a task) increase with the mental effort invested by the individual. Resources also increase in response to task demand. The human information-processing system, like an economic system, thus is able to adjust the supply of processing resources in response to demand. However, there is an upper bound to the amount of processing resources that can be applied: information-processing capacity is limited (see Fig. 4). Moreover, as task demand increases, the amount of "spare" capacity decreases, so that less can be allocated to other processing activities, whose efficiency may then suffer. This line of reasoning accounts for the performance decrement that can occur when two tasks are performed simultaneously and the demand of one task is increased.

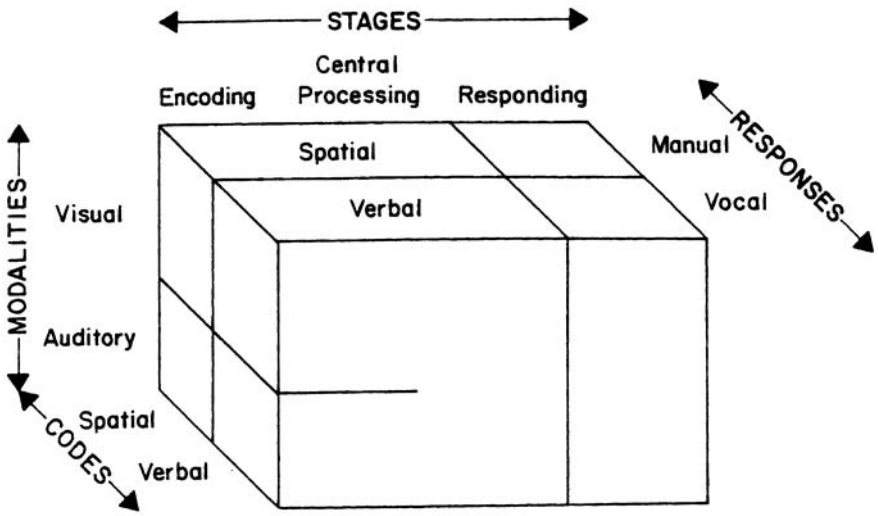


Figure 2 Multiple resource model from Wickens (1984).

Capacity limitations appear to be a fundamental property of both artificial and biological systems. Work in computer vision and robotics has shown that fully parallel processing of an entire complex visual scene is computationally inefficient in comparison to limited-capacity serial processing of parts of the scene. The large receptive fields of neurons in the higher perceptual processing areas of the primate brain are also consistent with such a computational limitation; because several objects in a visual scene fall within a neuron's receptive field, they compete for neuronal processing. Identification of an object among distracters therefore is dependent on the allocation of this limited processing resource.

Competition for limited processing resources is also a major factor in the decrement in performance that occurs when two tasks must be carried out simultaneously. According to unitary resource theories, interference between concurrent tasks occurs when

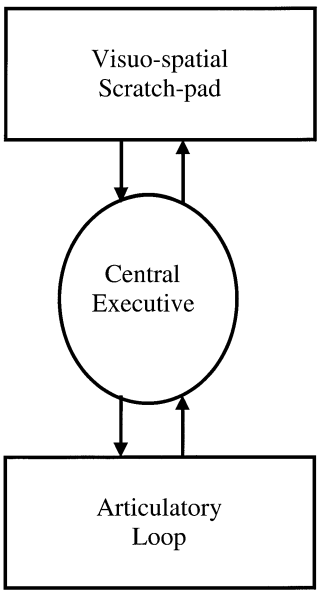


Figure 3 Baddeley (1992) model of working memory.

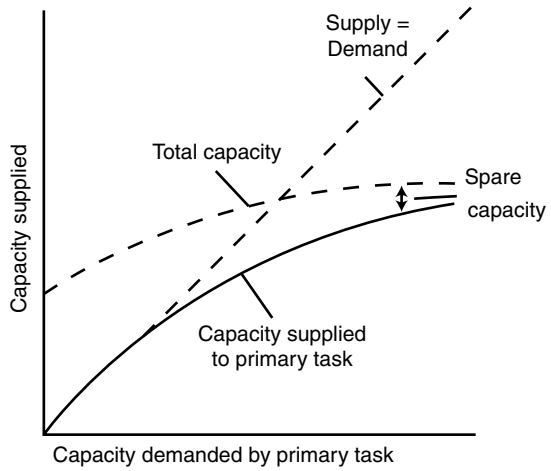


Figure 4 Relation between capacity (resources) demanded by a task and capacity supplied. From Kahneman (1973). *Attention and Effort* © Daniel Kahneman. Reprinted by permission of Pearson Education Inc., Upper Saddle River, NJ.

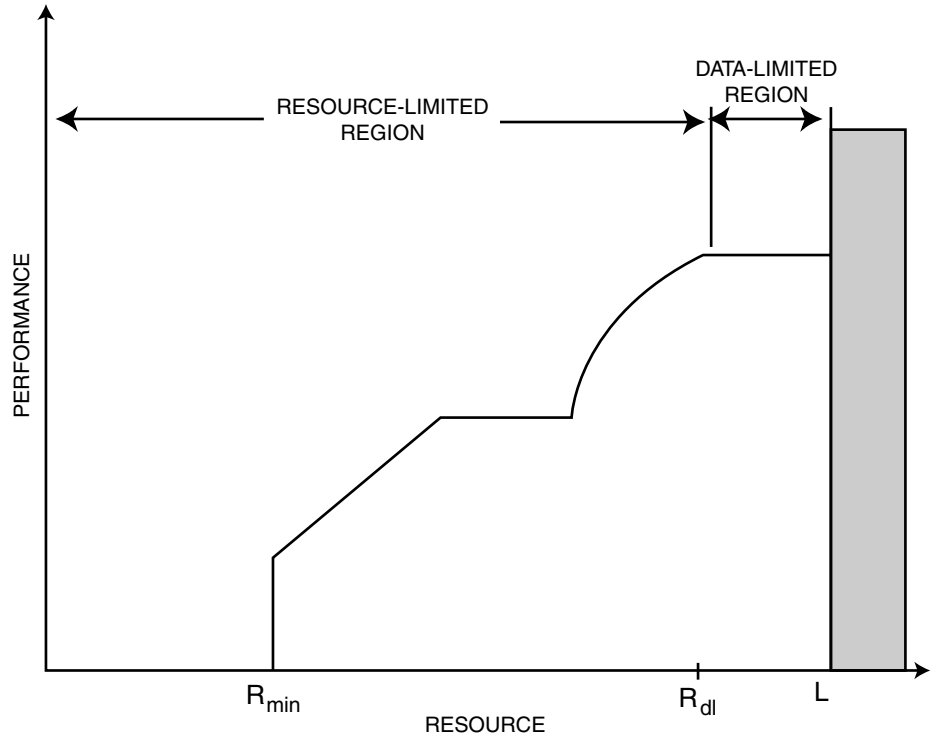


Figure 5 Relation between performance (processing efficiency) and resources for data-limited and resource-limited processes. Processing efficiency remains constant for data-limited processing but improves with further application of resources for resource-limited processing. From Norman and Bobrow (1975).

the total demand for processing resources exceeds the available supply. Modular theories predict that interference occurs only to the extent that concurrent tasks compete for the same module, e.g., two tasks that both demand visuospatial resources.

More formal resource models of mental workload distinguish between *data-limited* and *resource-limited* processes. According to this view, which assumes that resources are unitary, there are different ways in which the availability of resources can influence the functioning of simultaneously active mental processes. Once a certain minimum level of resources is allocated to activate a mental process, the application of greater resources leads to more efficient (e.g., faster, more accurate) execution of that process. Such a process is said to be resource-limited, in that the quality of its output is dependent on the allocation of resources. Subjectively, "working harder" at a task will improve its performance. As more resources are applied, however, there may come a point at which further allocation no longer improves the efficiency of the mental process. At this point, performance is said to be data-limited; efficiency is determined by the quality of the input source to the process, whether that source is sensory or drawn from memory storage (see Fig. 5). The subjective experience is that when the stimulus input is degraded, performance cannot be bettered: for example, no amount of "trying harder" will improve one's perception of a weak sound buried in loud noise.

II. CORTICAL SIGNS OF MENTAL WORKLOAD

Cognitive psychologists have deliberated on the functional structure of mental workload for a number of years. Debate has focused on the two major issues of whether resources are unitary or fractionated into modules and given the latter, how the modules are defined. There is even argument concerning the notion of resources. Some investigators question the need for this construct, arguing that many aspects of complex human behavior, e.g., concurrent performance of two or more tasks, can be explained without recourse to a resource concept. For example, it has been suggested that interference between two discrete tasks can be

explained by the postponement of the central response–selection stage of processing of one of the tasks, as in the phenomenon of the "psychological refractory period." According to this model, there is no graded sharing of resources between tasks at any stage of processing. However, this model and related structural models cannot account for all aspects of multiple-task performance, especially for continuous tasks such as tracking and reading, and other investigators have vigorously defended the need for the resource concept.

These debates have ebbed and flowed over the years but an impasse has been reached. Meanwhile, more practically oriented research on the assessment of the mental workload of human operators in real work settings has continued. Use of the traditional behavioral tools of cognitive psychology to resolve the issues may prove difficult. Like in many other fields of psychology, the advent of cognitive neuroscience has raised the hope that brain imaging studies may shed some additional light on the issues. In addition to PET and fMRI, studies using electroencephalography (EEG) and event-related potentials (ERPs) have also been reported.

A. Physiological Bases for Resources

In comparison to PET and fMRI, EEG and ERP studies of mental workload have the longer history. Of course, EEG and ERPs also assess neural activity directly (as opposed to indirectly through blood flow and metabolism with PET–fMRI), although measurement at the scalp smears and distorts the original neuroelectrical signals. Spectral power in two major frequency bands of the EEG, α (7–14 Hz) and θ (4–7 Hz) has been linked to mental workload. In fact, the reduction in the amplitude of the α rhythm during mental arithmetic was first reported in 1929 by the discoverer of human EEG, Hans Berger. Although the intracerebral generators of scalp-recorded EEG α are not known precisely, α is thought to arise in widespread cortical areas via thalamocortical interactions. Functionally, α may represent a type of cortical "idling" that is disrupted whenever attentional resources are allocated to a task. Spectral power in the α frequency band thus is inversely related to resource allocation. In contrast, more recent studies suggest that EEG θ recorded from midline sites overlying the frontal cortex increases in power with increased task difficulty and with higher memory load during working memory tasks. Frontal θ power thus varies directly with resource allocation and can be distinguished from the posterior θ waves that have traditionally been linked to drowsiness. Again, although the source of this anterior EEG rhythm is unknown, it is consistent with a generator in the medial frontal cortex, perhaps the anterior cingulate cortex (ACC).

The EEG data provide good evidence for a physiological basis for resources in the human brain. The changes in EEG α and θ may represent a relatively gross, whole-task allocation of resources, given that the associated brain electrical activity spans several seconds and is relatively invariant with the type of information processing (although hemispheric differences have been noted with different forms of processing, e.g., verbal versus spatial tasks). The processing characteristics of specific components underlying a task can be assessed by recording task-evoked ERPs, which reflect the neural activity in response to a stimulus with millisecond precision.

ERP studies of mental workload have identified two resource-sensitive components, the P300 and N100. Most of these studies have examined dual-task performance, although working memory studies have also been carried out. The amplitude of the P300 component to a secondary task of counting infrequent tones among more frequent tones (an "oddball" task) decreases when combined with a primary task such as visual discrimination or psychomotor tracking. Thus, P300 amplitude shows a dual-task decrement, i.e., it is reduced in amplitude when the eliciting task is combined with another task. Importantly, only changes in the difficulty of visual discrimination and not motor tracking affect P300 amplitude on the secondary task. Because the P300 component has been shown to be more sensitive to central stages of information processing than to the response-selection stage, this finding provides strong supporting evidence for a modular theory of resources based on stages of processing (see Fig. 2).

A strong prediction of theories that postulate sharing of scarce resources is one of *resource reciprocity*, or an inverse relationship between primary and secondary task resource allocation and performance. As resources are withdrawn from the primary task, they are simultaneously allocated to the secondary task, and vice versa. If P300 amplitude reflects the allocation of resources to a central processing stage, its amplitude should vary accordingly. Consistent with this prediction, P300 amplitude to the primary or secondary task has been found to increase or decrease appropriately as resources are applied or withdrawn. Resource allocation between two tasks can also be

manipulated endogenously (through instructions and a payoff scheme) such that resources are allocated in differing proportions between two tasks (e.g., 25–75%, 50–50%, or 75–25%). Such studies have the advantage that the stimuli and responses remain constant across conditions, as opposed to single–dual task comparisons, which confound resource allocation with stimulus and response variations. In the resource reciprocity studies, the amplitudes of both the P300 and the early-latency N100 components have been found to vary in a graded manner with resource allocation.

The significance of these results should be emphasized. The finding of a dual-task decrement in P300 or N100 amplitude is insufficient by itself to support a resource model, because factors other than resource scarcity may contribute to dual-task interference. Consequently, the demonstration that ERP components show *graded* changes between tasks as resources are dynamically traded off between each other, with invariant stimuli and response requirements, provides strong support for resource theories. The P300 data also support a modular theory based on different stages of processing.

B. Cortical Localization: Working Memory, Divided Attention, and the Central Executive

The ERP studies of mental workload have provided supporting evidence and a physiological basis for resources in relation to dual-task performance. A particularly important contribution of these studies is their support for predictions from basic cognitive theory. The ERP data also provide support for the modular view, although they do not rule out hybrid theories in which modular resources are combined with a general purpose resource that is invoked for all processing activities.

As brain electrical activity recorded from the scalp, EEG–ERP studies do not provide strong evidence for cortical localization of resources. The poor spatial resolution of EEG and ERPs can be overcome to a degree through the use of such techniques as dipole modeling and spatial deconvolution. For example, the use of these techniques has led to the view that workload-related EEG θ activity is generated in the anteromedial frontal cortex, possibly the ACC, which has also been proposed to be a high-level central executive control center on the basis of PET and lesion studies, as will be discussed further later. In general, however, brain imaging techniques offer spatial resolution superior to that of EEG–ERPs and have provided evidence on the cortical localization of resources. This line of research follows from early work on measuring cerebral blood flow and metabolism using Xe-133, as described previously.

Much of the evidence stems from PET and fMRI studies of the neural substrates of working memory. This is a type of memory involved in keeping and maintaining information "on line" so that it can be used in the service of other processing activities—in language, decision making, and problem-solving. The functional properties of working memory have been studied in the cognitive psychology literature, and a widely accepted model is that illustrated in Fig. 3. The possible neural bases of these components of working memory have been the focus of intense study. A general finding is that active maintenance of information in working memory is associated with the activation of both frontal and posterior (parietal) cortical regions, depending on the type of material encoded and the specific operation in working memory probed. For example, it is well-known that perceptual operations can be divided into object and spatial components and that these operations are mediated by cortical processing streams that activate regions in the inferior temporal (ventral stream) and parietal cortices (dorsal stream), respectively. This cortical subdivision of labor during perception has been postulated to result in similar divisions "upstream" in the frontal cortex. In support of this prediction, working memory for objects such as faces has been found to activate the lateral prefrontal cortex, whereas spatial working memory has been shown to recruit more dorsal regions of the frontal cortex in the premotor region. Brain imaging data thus support the modular resource view shown in Fig. 3 to the extent that separate frontal loci have been observed for verbal and visuospatial working memory.

PET and fMRI studies of working memory have used a variety of tasks, such as delayed match-to-sample, in which a target item (the sample) must be retained in working memory and then compared to another stimulus presented after a delay. Another task is the *n*-back task, on each trial of which the *n*th previous stimulus in a series of stimuli presented sequentially must be identified and responded to (generally $n = 1$ or 2). In such tasks, activation of the ventrolateral prefrontal cortex (VLPFC), dorsolateral prefrontal cortex (DLPFC), and posterior parietal cortex is observed. Broca's area is also activated when verbal stimuli are used, probably reflecting phonologic rehearsal of items. Such studies

have also revealed a modular distinction within the lateral prefrontal cortex. Processes associated with the maintenance of information in working memory activate VLPFC and posterior parietal cortex, whereas processes associated with the manipulation of that information activate DLPFC. DLPFC activation has also been observed for high levels of working memory load, possibly reflecting mechanisms that are needed to provide processing support when task demands exceed short-term memory capacity. It has also been suggested that the posterior parietal areas may be a more passive storage buffer of on-line information, whereas VLPFC mediates more active maintenance processes. However, other work suggests that posterior parietal areas may in fact be involved with executive processes as well. The types of manipulation processes performed in DLPFC may include monitoring, updating, scheduling, attention shifting, and inhibition. Because of the prominence of the DLPFC in manipulation and other high mental workload-related processes, it has been implicated as a possible neural basis for a central executive module.

Studies of the Stroop task have identified another frontal cortical region involved with executive control, the ACC. In this task a person is asked to name the color in which a word is written. The word itself represents the name of a color and is written in either the same color (e.g., "red" written in red: congruent condition) or a different color (e.g., "green" written in red: incongruent condition). Reaction time to color naming typically is longer in the incongruent condition: this is known as the Stroop effect. Several studies have associated the Stroop effect with the activation of both the DLPFC and the ACC. These brain regions have also been implicated specifically in response selection in divided attention tasks and more generally with executive or supervisory activity. One study showed that it is possible to functionally dissociate these two regions using fMRI. DLPFC activation was found to be associated with cognitive control or with the focusing of attention on the demands of the task. In contrast, ACC activation was related to the provision of feedback about subjects' accuracy on the task.

Divided attention (dual-task) studies have obtained results relevant to the distinction between unitary and modular theories of workload and the concept of a central executive. In these studies, brain regions activated by the concurrent execution of two tasks are compared to those activated during the execution of either task in isolation. Dual-task performance ostensibly requires the central executive because of the need for coordination (although this must be demonstrated empirically for any given pair of tasks, and not all studies have done this). Therefore, any brain region activated by dual-task but *not* by single-task performance would potentially provide evidence for a specialized central executive region, such as the DLPFC or the ACC. Two fMRI studies failed to provide support for this view. In tasks involving verbal and face working memory, no new brain area was activated for dual-task performance. Instead, activation increased with dual-task performance but in the same regions active during the performance of each task individually. Although these findings do not rule out the possibility that specific executive processes are mediated anatomically by specialized modules, they do concur with other findings and are consistent with a modular view of workload, with content-specific slave buffers but no separate central executive control "center."

C. Beyond Localization: Parametric Manipulation of and Dynamic Interactions in Working Memory

Brain imaging studies of working memory and divided attention paint a fairly consistent picture. On-line maintenance of information in support of cognitive performance, whether in specific processing domains such as verbal, object, or spatial processing or more generally as in executive processes, activates specific neural circuits in the frontal cortex. The first generation of these imaging studies was primarily concerned with localization. As such, they have value in pointing to the cortical regions putatively associated with resources and, therefore, in describing the cortical expression of mental workload. However, localization per se is only partially informative with respect to the functional aspects of the neural mechanisms underlying mental workload. Localization is only a first step in a deeper understanding of such mechanisms. More important would be evidence that shows that neural activity in specific cortical regions exhibits systematic variation with task or subject variables or, as previously discussed, with exogenous or endogenous drivers of mental workload.

An early example of such parametric manipulation of workload was reported in an fMRI study of language processing in which the computational complexity of comprehending different sentences was varied. Sentence comprehension requires working memory processing resources first to compute the comprehension operations (word integration, syntactic and semantic relations, etc.) and second to maintain

these representations actively during the period of processing, which can be considerable for a long, complex sentence. The study examined sentences that varied in structural complexity by comparing simple conjoined sentences, including subject and object clauses. Important controls were the number of words, word frequency, and the complexity of individual words, so that the sentences differed only in overall structural complexity. Left hemisphere areas associated with sentence processing were analyzed for the volume of activation in the superior temporal cortex (Wernicke's area) and inferior frontal gyrus (Broca's area). Activation volumes increased with sentence structural complexity in both cortical areas and in a graded manner. Thus, the quantitative extent of neural activity evoked by sentence processing was directly related to the associated computational demand. These findings suggest that manipulation of processing complexity unconfounded with other factors results in increased brain activation in a manner consistent with the allocation of specific processing resources.

Similar parametric studies of working memory have also been reported. Such studies have shown that, as mental workload increases with an exogeneous driver, the cortical areas associated with the type of process manipulated exhibit corresponding increases in activity. PET studies of the *n*-back task have shown monotonic increases in activity in a number of areas, including the DLPFC, VLPFC, posterior parietal cortex, and Broca's area, as the size of *n* or the amount of information retained in working memory increases. In addition, monotonic deactivation with increasing verbal working memory demands has also been reported for certain brain regions. Although no clear consensus has been reached as to the relevance of these deactivations to the functioning of working memory systems, the consistent parametric activation of process-specific "slave" modules, coupled with parametric deactivation of other brain regions, suggests a reallocation of mental resources to compensate for increased task demands.

The working memory and dual-task studies indicate first that activation of specific frontal cortical regions can be identified as cortical signs of processing resources associated with task performance. Second, activity in these regions increases, or the volume of activated tissue increases, in response to demand for greater processing resources with exogenous drivers of workload. Such increases probably also reflect endogenous sources of workload, but exogenous and endogenous effects have not yet been differentiated by brain imaging studies. An important additional question is whether such cortical signs of processing resources are *necessary* for efficient task performance. For example, is the working memory "signature" in the frontal cortex sustained across time when a target object that has to be kept in mind occurs repeatedly? Specifically, how can processing resources be deployed to sustain attention to a specific object over time in the presence of distractors? fMRI studies have shown that neural responses are changed as stimuli become more familiar or as new associations are learned. However, such alterations in activation cannot distinguish a specific object that is currently the focus of attention from equally familiar distracters that should be ignored.

The neural responses that mediate repeated target identification during working memory were examined in an fMRI study. On a given trial subjects were required to study a sample object (a face) and then to identify it repeatedly with a button press in a stream of objects containing distracters (there were a total of 13 items plus the sample in a single trial). Subjects were familiarized with every object, and a specific object was used as a target on some trials and as a distracter on other trials. Distracters were also repeated within a trial. Thus, neither familiarity nor simple repetition could be used to signal the presence of a target, but rather a representation of the target on that trial had to be kept in mind for the duration of the trial. Under these conditions, enhanced neural responses were found primarily in bilateral inferior–middle frontal and left insular cortices (see Fig. 6). These responses signaled the identification and maintenance in working memory of the currently attended target item. Targets were also responded to faster with each repetition within a trial. The neural response to targets remained constant with repetition in frontal–insular areas, whereas those to targets or distracters in posterior regions declined (see Fig. 7). Activity associated with repeated distracters in frontal regions also remained at a constant but low level throughout the trial. Thus, the enhanced frontal neural responses and the speeded response to targets signaled the maintenance of the target object in working memory, supporting a role for the active maintenance component of working memory in the selection of targets among distracters. Effective selective attention requires that the neural response to a target stimulus is enhanced and maintained during the period of time the target remains behaviorally relevant. The sustained enhancement of frontal responses across target repetitions might reflect such top-down allocation of attentional resources to target detection.

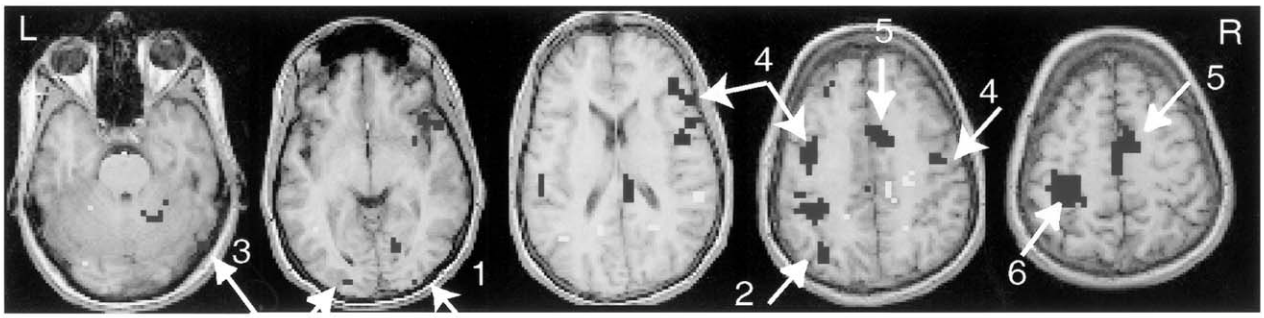
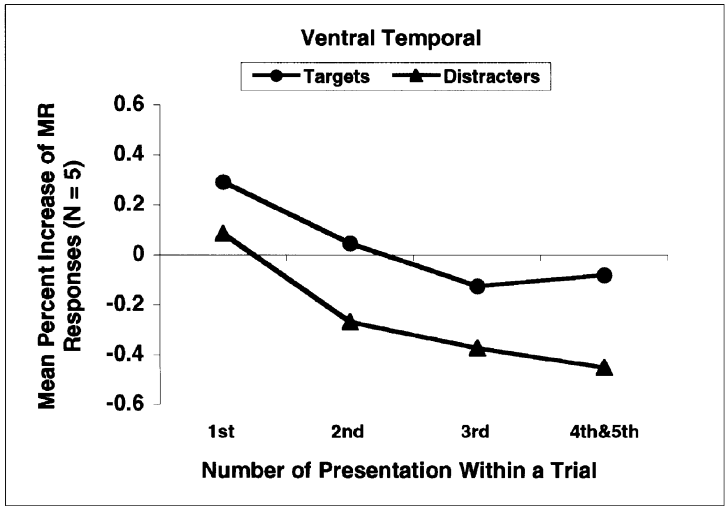


Figure 6 fMRI activation patterns showing enhanced neural responses to targets. Labels 4, 5, and 6 indicate frontal–insular areas, the supplemental motor area, and the left motor cortex, respectively. Reprinted with permission from Jiang *et al.* (2000). Complementary neural mechanisms for tracking items in human working memory. *Science* **287,** 643–646. Copyright 2001 American Association for the Advancement of Science.



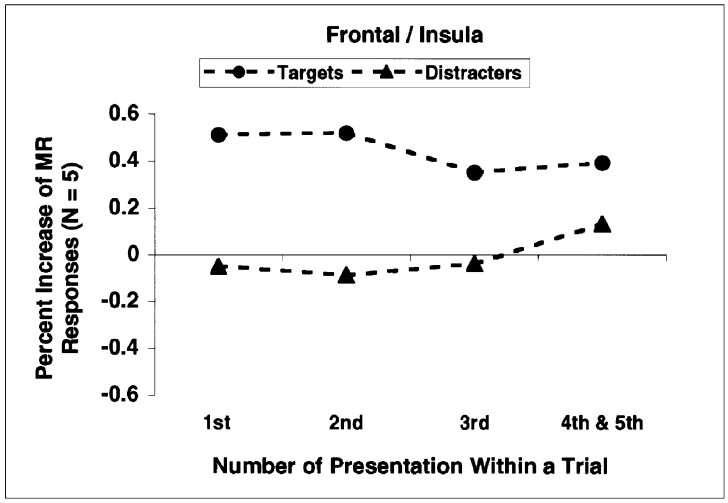


Figure 7 Mean percentage increase of fMRI responses to repeated targets and distracters in ventral temporal (A) and frontal–insular cortices (B). The neural response to targets remained constant with repetition in frontal–insular areas, whereas those to targets or distracters in posterior regions declined. Reprinted with permission from Jiang *et al.* (2000). Complementary neural mechanisms for tracking items in human working memory. *Science* **287,** 643–646. Copyright 2001 American Association for the Advancement of Science.

III. CONCLUSIONS

Mental workload represents a composite brain state that reflects the interaction between task demands and a person's capability to meet those demands by the voluntary application of mental effort. Unitary theories propose that mental workload is associated with a single pool of information-processing resources that can be flexibly allocated to different processing activities. In the modular view, functionally separate resources are applied to different processing activities.

Evidence from electrophysiological (EEG, ERP) and functional brain imaging (PET, fMRI) studies supports a physiological basis for processing resources in the human brain. In general, these indexes of human brain function support a modular view of mental workload, although they do not rule out hybrid theories in which modular resources are combined with a general purpose resource that is invoked for all processing activities. PET and fMRI studies of working memory and dual-task performance indicate that mental workload is associated with the activation of cortical networks, including the dorsolateral and ventrolateral prefrontal cortices and posterior parietal cortex. Parametric modulation of activation in these brain regions with increased workload has been reported in a few studies. Mental workload may reflect the dynamic recruitment of these anterior and posterior brain regions in support of task performance in response to increased task demands.

See Also the Following Articles

ATTENTION • CONSCIOUSNESS • INFORMATION PROCESSING • MEMORY, EXPLICIT AND IMPLICIT • MOTION PROCESSING • NUMBER PROCESSING AND ARITHMETIC • PATTERN RECOGNITION • SHORT-TERM MEMORY • VIGILANCE

Suggested Reading

Baddeley, A. D. (1992). Working memory. *Science* **255,** 556–559.

Bunge, S. A., Klingberg, T., Jacobsen, R. B., and Gabrieli, J. D. E. (2000). A resource model of the neural basis of executive working memory. *Proc. Natl. Acad. USA* **97,** 3573–3578.

Gopher, D. (1996). Attention control: Explorations of the work of an executive controller. *Cogn. Brain Res.* **5,** 23–38.

Jiang, Y., Haxby, J. V., Martin, A., Ungerleider, L. G., Parasuraman, R. (2000). Complementary neural mechanisms for tracking items in human working memory. *Science* **287,** 643–646.

Just, M. A., Carpenter, P. A., Keller, T., Eddy, W. F., and Thulborn, K. R. (1996). Brain activation modulated by sentence comprehension. *Science* **274,** 114–116.

Kahneman, D. (1973). *Attention and effort*. Prentice Hall, Englewood Cliffs, NJ.

Kramer, A. F. (1991). Physiological metrics of mental workload: A review of recent progress. In *Multiple Task Performance*. (D. Damos, Ed.), pp. 279–328. Taylor and Francis, London.

MacDonald, A. W., Cohen, J. D., Stenger, V. A., and Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science* **288,** 1835–1838.

Norman, D. A., Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cogn. Psychol.* **7,** 44–64.

Parasuraman, R. (1998). *The Attentive Brain*. MIT Press, Cambridge, MA.

Posner, M. I., and Tudela, P. (1997). Imaging resources. *Biol. Psychol.* **45,** 95–107.

Roland, P. E. (1993). *Brain Activation*. Wiley-Liss, New York.

Roy, C. S., and Sherrington, C. S. (1890). On the regulation of the blood supply of the brain. *J. Physiol. (Lond.)* **11,** 85–108.

Smith, E. E., and Jonides, J. (1997). Working memory: A view from neuroimaging. *Cogn. Psychol.* **33,** 5–42.

Wickens, C. D. (1984). Processing resources in attention. In *Varieties of Attention*. (R. Parasuraman, D. R. Davies, Eds.), pp. 63–102. Academic Press, San Diego.

Microglia

GUIDO STOLL
Heinrich Heine University, Düsseldorf, and Julius Maximilians University, Würzburg, Germany

SEBASTIAN JANDER and MICHAEL SCHROETER
Heinrich Heine University, Düsseldorf, Germany

GLOSSARY

antigen presentation Presentation of processed antigens on the surface of macrophages, microglia, or dendritic cells concomitant with major histocompatibility complex class II molecules is a mandatory step for generation and activation of antigen-specific T cells.

cytokine Soluble factor regulating interactions between immune cells.

glia Nonneuronal cells of the central nervous system encompassing microglia, astrocytes, and oligodendrocytes.

major histocompatibility complex Immunological surface molecules important in antigen-recognition processes by T cells and self/nonself discrimination.

microglia A subpopulation of glia in the central nervous system.

myelin Biological membrane around nerve fibers facilitating fast nerve conduction; produced by Schwann cells in the peripheral nervous system and oligodendrocytes in the central nervous system.

nerve regeneration Regrowth of nerve fibers after injury.

synaptic stripping Active detachment of synapses from neurons by glial processes.

Wallerian degeneration Term for the complex cellular and molecular events in the distal stump after nerve injury with degeneration of nerve fibres.

Microglia are a major glial component of the central nervous system (CNS) of presumed bone marrow origin. Only the perivascular microglia, a subtype, are regularly replaced from the bone marrow in adult animals, whereas parenchymal microglia are extremely sessile. Microglia respond to virtually any, even minor, pathological events in the CNS. This article reviews basic properties of microglia and discusses their multiple functions in the CNS during neurodegeneration, ischemia, autoimmunity, and inflammation.

I. BASIC PROPERTIES OF MICROGLIA

A. Microglia in the Normal CNS

The CNS is composed of neurons and a nonneuronal population of cells, designated glia, that includes astrocytes, oligodendrocytes, and microglia. del Rio Hortega discovered microglia in 1919 when he developed the silver carbonate methology. He identified process-bearing cells in the CNS parenchyma and named these cells microglia. Microglia contribute to about 10% of the total glial cell population in the CNS of adults. In the 1960s, the existence of ramified microglia as a distinct glial population was further confirmed by electron microscopy showing lack of intracellular intermediate filaments and intercellular gap junctions that are typical ultrastructural features of astrocytes. In contrast to oligodendrocytes, the principal myelin-forming cells of the CNS, microglia do not entertain myelin sheaths. Microglia are present in all parts of the brain and spinal cord (Fig. 1) but are

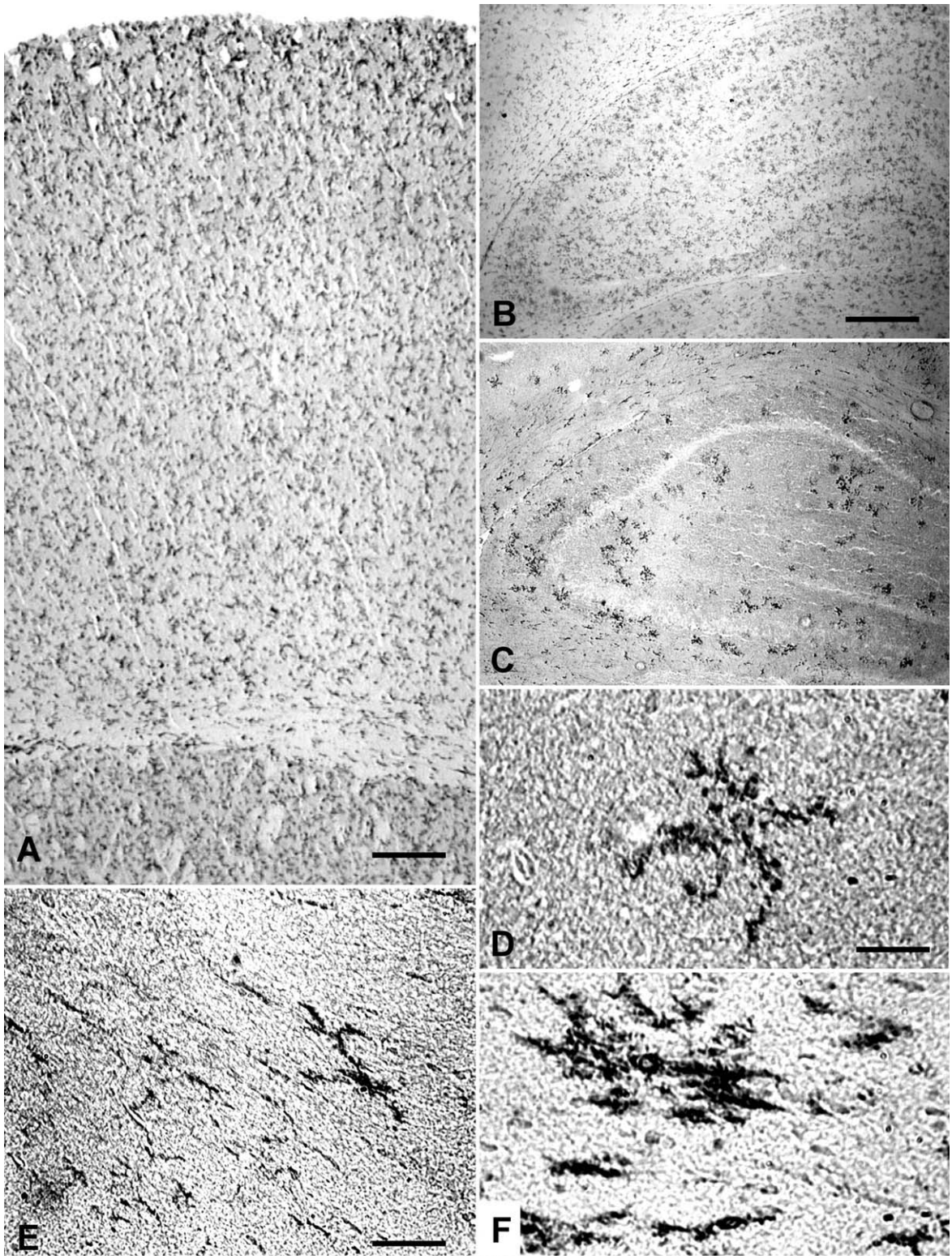


Figure 1 Parenchymal microglia in the rat brain. Sections of the sensomotor cerebral cortex (A), hippocampus (B–D), and cortical white matter (E, F) stained immunocytochemically with antibodies against the complement receptor-3 (CR-3) (A, B, E, F) and against keratan sulfate proteoglycans (KSPG) (C, D) to identify microglia. Note the widespread distribution of microglia in the cortex (A) and their typical ramified morphology (D). Microglia show a more longitidinal orientation in white matter tracts (E,F). More microglia express CR-3 (B) than KSPG (C). Scale bars = 200 μm in A, 80 μm in and B and C, 20 μm in D and F, and 40 μm in E.

not uniformly distributed. Microglia are located in close vicinity to neurons in the gray matter (Figs. 1A–1C) and between fiber tracts in the white matter of the CNS (Figs. 1E and 1F). More microglia are found in gray than in white matter. Microglia vary in morphology depending on their location. Longitudinally branched cells are found in fiber tracts (Figs. 1E and 1F). They possess long processes that are usually aligned parallel to the longitudinal axis of the nerve fibers. Radially branched cells are found throughout the gray matter (Fig. 1D). During normal aging microglia undergo morphological changes. Although in newborn brain few ramified microglia can be detected, they increase in number during postnatal development and are abundant in the middle-aged human brain. In the aged brain microglia change morphology again and show signs of activation, with enlarged longitudinally arranged processes.

Today, microglia can unequivocally be distinguished from other resident, nonneuronal cells in the CNS by immunocytochemistry. Resting microglia constitutively express the complement type-3 receptor (CR3; CD11b/CD18 complex) (Figs. 1A, 1B, and 1E), Fc receptors for binding of immunoglobulins, and galactose-containing glycoconjugates that bind isolectin B4 from Griffonia simplicifolia seeds. However, these antigens are not specific for microglia. Hematogenous macrophages that invade the CNS under certain pathological conditions also express CR3, Fc receptors, and the lectin binding site. Currently, the only antigen to our knowledge that is not shared between microglia and macrophages is a keratan sulfate proteoglycan epitope (KSPG) detected by monoclonal antibody 5D4 (Figs. 1C and 1D), which, underlies modification during immune-mediated processes in the CNS. In the gray and white matter of the normal rat CNS, a subpopulation of rat parenchymal microglia constitutively express KSPG.

Resting microglia generally express low to absent levels of major histocompatibility complex (MHC) class II molecules. These molecules are important for immune reactions. Already within a given species there are fundamental differences in the expression of microglial surface molecules. Whereas in the rat constitutive MHC class II expression of ramified microglia can be detected in Brown Norway and Dark Agouti strains, it is virtually absent in Lewis and Fischer 344 rats. Interestingly, constitutive MHC class II expression is inversely related to KSPG expression. These findings suggest a significant impact of genetic factors on the molecular differentiation of resident microglia in the normal CNS that are not yet elucidated.

Microglia respond to a variety of functional and lesional disturbances in the CNS. This response is accompanied by characteristic morphological and molecular signs of activation. Activated microglia retract processes followed by rounding of the cell body, particularly in conditions in which microglia finally act as phagocytes. This is associated with increased levels of CR3 expression on the surface. As a relatively early sign of microglial activation, MHC class I and II antigens are upregulated on the cellular surface, which enables microglia to interact directly with T cells. Except in mice, activated microglia express CD4 antigens, which are accessory molecules of T helper lymphocytes. Upon further transition into phagocytes, microglia develop intracellular phagolysosomes that can be visualized by immunocytochemistry. All these molecules indicative of microglial activation are also present on hematogenous monocytes/macrophages. This makes a distinction between the relative contribution of activated local microglia and that of invading macrophages from the blood in CNS disorders, in which both cell types are involved, difficult.

A subset of microglia, the perivascular microglia (which have also been termed perivascular cells), can be distinguished by their unique association with vessels and by immunocytochemistry using mab ED2 in the rat (Figs. 2A and 2C). These cells are spindle shaped and located around blood vessels. Perivascular microglia constitutively express the immune activation markers CD4 and MHC class II molecules (Figs. 2B, 2D, and 2E).

B. Microglia in Culture

Culture experiments have allowed the study of basic functional capacities of microglia. Most of these studies have been performed with microglia that were isolated from mixed glial cell cultures based on newborn rodent brain as starting material. In these isolated culture systems, microglia lose their characteristic ramified shape and round up to an ameboid morphology reminescent of activated microglia in areas of tissue remodeling during development and after brain injury (Fig. 3A). Compared to the normal *in vivo* situation, production of inflammatory cytokines is markedly enhanced in cultured microglia and likely further increased by the presence of serum that is routinely added to the culture medium but normally excluded from the brain parenchyma.

Treatment of cultured microglia with proinflammatory substances such as interferon-γ (IFN-γ), which is

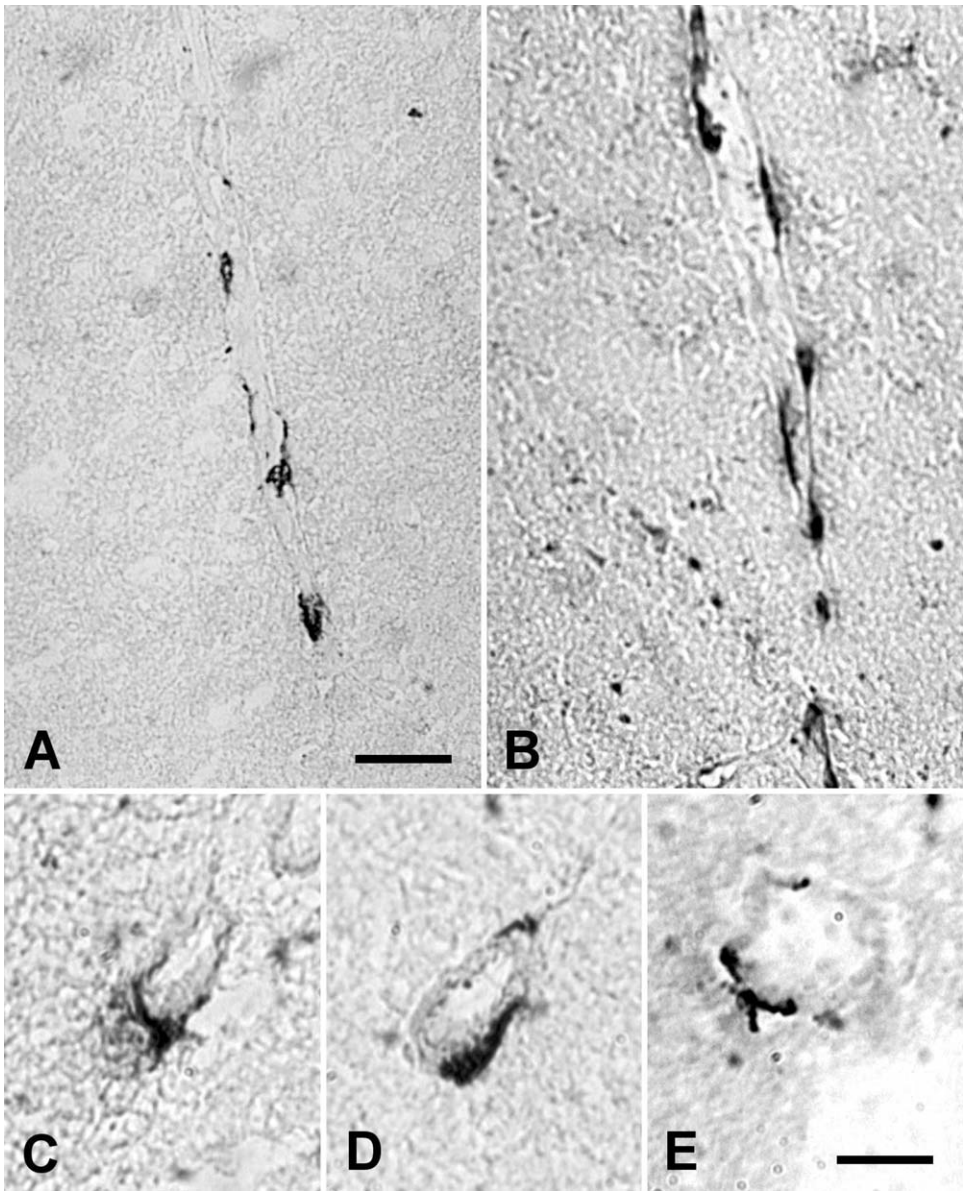


Figure 2 Perivascular microglia in the rat brain. Immunolabeling with ED2 antibody reveals a unique cell population of microglia also designated perivascular cells (A) that also express CD4 (B). Note the exclusive localization of the $ED2^+$ (A) and $CD4^+$ (B) microglia in close vicinity to cerebral blood vessels. (C–E) Perivascular microglia at higher magnification and their expression of ED2 (C), CD4 (D), and MHC class II antigens (E). Scale bars = 40 μm in A and B and 20 μm in (C–E).

secreted by T cells during inflammation, leads to induction of MHC class II antigens and a number of adhesion and costimulatory molecules, such as intercellular adhesion molecule-1 (ICAM-1), B7-1, and CD40, that are essential participants in the process of antigen-specific T cell stimulation. Accordingly, activated microglia derived from newborn rodent brain efficiently stimulated the activation and proliferation of antigen-specific T cells *in vitro*, supporting a role for microglia as the major antigen-presenting and immune-stimulatory cell type in the mammalian brain. However, this concept has been challenged by recent studies using ramified microglia isolated from the adult brain that have shown downregulatory rather than immune-stimulatory interactions of parenchymal microglia with T cells.

When microglial cells are grown on an astrocytic monolayer they develop processes similar to those of

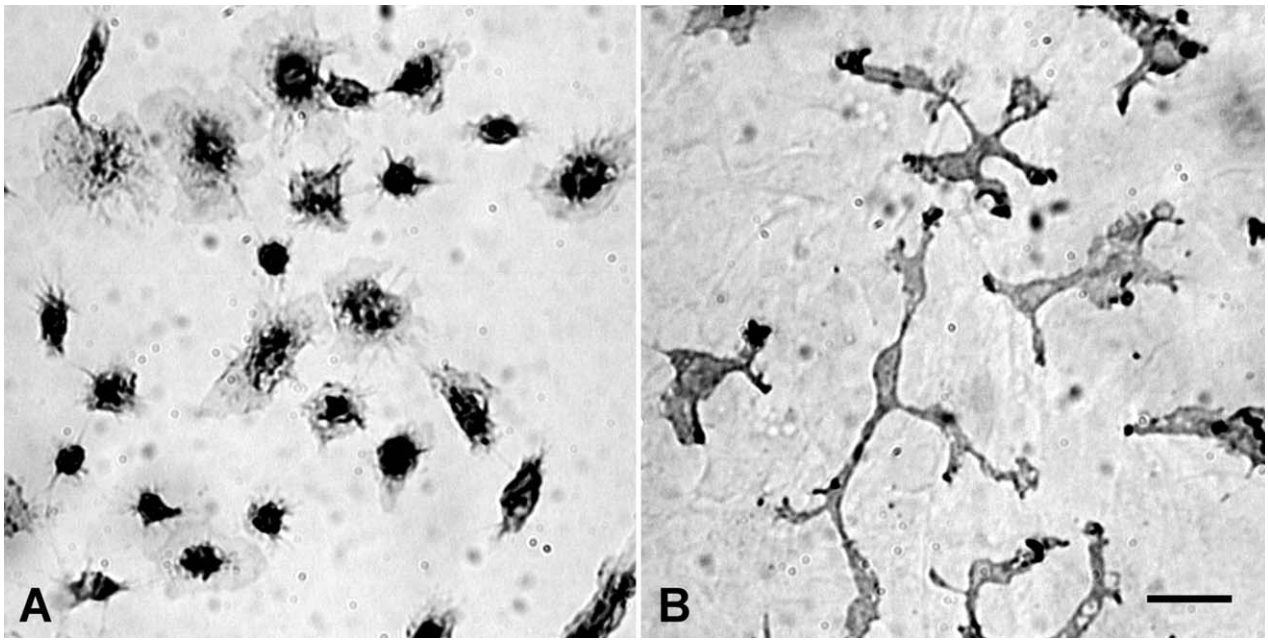


Figure 3 Rat microglia in culture dishes: immunostaining for complement receptor 3 by monoclonal antibody Ox42, a microglial marker. (A) Note the round (ameboid) morphology of isolated microglia resembling monocytes/macrophages. In the presence of astrocytes in primary mixed glial cultures (barely visible in the background), microglia develop multiple processes and acquire their distinctive morphological features (B). Scale bar = 20 μm.

ramified microglia in brain parenchyma (Fig. 3B). This points to an important influence of astroglia on microglial properties that is further substantiated by the fact that coculturing microglia with astrocytes markedly suppressed phagocytosis. Nondiffusible matrix components as well as astrocyte-derived cytokines such as the macrophage colony-stimulating factor (M-CSF) have been implicated in the induction of microglial ramification in these coculture models. In support of the proposed hematogenous origin of ramified microglia, monocytes or spleen macrophages brought into coculture with astrocytes likewise displayed a ramified morphology that was indistinguishable from that of CNS-derived microglia. Functionally, ramification was accompanied by the downregulation of immunomoleucles such as MHC class II, leukocyte function-associated antigen-1, and ICAM-1. Microglia bear a variety of receptor molecules on their surface, facilitating signaling through complement components, growth factors, cytokines, and chemokines. Antiinflammatory cytokines, such as interleukin-(IL-10) and transforming growth factor-β (TGF-β), can effectively suppress the expression of MHC class II antigens on microglia as well as neuronally produced factors such as brain-derived neurotrophic factor. *In vivo*, the cellular microenvironment of the CNS thus provides multiple autocrine as well as paracrine antiinflammatory feedback loops that tightly control the potentially harmful activation of microglia in disease states.

C. Factors Produced by Microglia

A striking feature of microglial reactivity is the ability to synthesize and secrete a large number of substances, which alone or in concert with factors derived from other brain or hematogenous cells can orchestrate immune responses and repair processes or directly cause brain damage. Factors produced by microglia include cytokines, growth factors, coagulation and complement factors, prostanoids, free radicals, extracellular matrix components, enzymes, and neurotoxins. In general, our knowledge about the circumstances in which each of these factors is produced and the regulatory factors involved is limited. This is mainly due to the complexity of possible interactions.

Microglia are an important source of the proinflammatory cytokines tumor necrosis factor-α(TNF-α), IL-1, IL-6, IL-12, and IL-18, which play an essential role in the initiation, amplification, and effector phase of T cell-mediated autoimmunity in the CNS.

Moreover, the cytokines IL1-β and TNF-α, have been inflicted in the pathophysiology of ischemic neuronal death but, on the other hand, may also exert neuroprotective effects. Interestingly, microglia also produce cytokines with antiinflammatory activities such as IL-10 and TGF-β. These cytokines provide an efficient autocrine mechanism for controlling microglial activity.

Nitric oxide (NO), whose formation is catalyzed by nitric oxide synthase (NOS), is a potent local mediator in the CNS. NO plays a major role in neural transmission, inflammation, neurotoxicity, host defense, and in the regulation of the vascular tone. Microglia derived from rodent brain are potent producers of NO upon stimulation with lipopolysaccharide (LPS) and cytokines. Contrastingly, the expression of inducible NOS and secretion of NO by human macrophages and microglia is controversial. Although low levels of NO appear to elicit protective responses, excessive NO production during autoimmunity and cerebral ischemia contributes to tissue damage. In the CNS, neurons and oligodendrocytes are most susceptible to NO cytotoxicity. Other potentially cytotoxic substances of microglia include superoxide anions, the enzymes elastase and gelatinase, excitotoxin, and quinolinic acid. However, most of the information on the cytotoxic properties of activated microglia pertains to *in vitro* observations and requires confirmation *in vivo*. On the other hand, microglia-derived neurotrophins, such as nerve growth factor, neurotrophin 3, basic fibroblast growth factor, and TGF, as well as the cellular matrix component thrombospondin likely contribute to repair processes after injury.

II. ORIGIN OF MICROGLIA

The origin of ramified microglia has been a long-standing controversial issue, although most authorities would accept that microglia are bone marrow derived and belong to the monocyte/macrophage lineage. The observation that the decline of blood-derived ameboid cells (macrophages) in the CNS during the first postnatal weeks was accompanied by a dramatic increase in the number of ramified microglia was suggestive for a transition of ameboid cells into resident ramified microglia. Based on morphological grounds, however, transitional forms between these brain macrophages and resting microglia could not be detected in the developing brain. Moreover, in an attempt to directly address the issue of transition, young mice received bone marrow transplants from transgenic mice, thereby allowing the distinction between host and donor cells in tissues. In these chimeric animals only 10% of parenchymal microglia in the CNS displayed the transgenic signal. In adult animals attempts to directly demonstrate the replacement of ramified parenchymal microglia from bone marrow-derived precursors have so far yielded inconclusive results. Ramified microglia in the adult CNS are an extremely sessile cell population exhibiting virtually no turnover from circulating monocytic precursor cells. In contrast to the parenchymal microglia, the perivascular microglia are definitely bone marrow derived and regularly replaced in the adult CNS as demonstrated by use of chimeric rats by Hickey and Kimura.

The view that parenchymal microglia are bone marrow derived has been challenged. Based on their finding that astroglial cultures initiated from newborn mouse neopallium contained bipotential progenitor cells that could give rise to both astrocytes and microglia, Fedoroff and colleagues put forward the idea that parenchymal microglia are of neuroectodermal origin, as are all other glia. This view was further supported by the observation that the majority of microglia lacked the transgenic signal after bone marrow transplantation as described previously. Despite the uncertainty about their origin, microglia share most surface molecules with bone marrow-derived monocytes/macrophages.

III. MICROGLIA AND CNS INJURY

Microglia are able to transform into large phagocytes and thereby remove debris in the CNS. Microglial responses, however, appear to be tightly controlled by environmental factors in CNS injury. This is most apparent when the phagocytic activity of microglia during Wallerian degeneration is compared to that of necrotic CNS lesions induced by stab wounds, bacterial infection, or cerebral ischemia.

A. Ineffective Phagocytic Activation of Microglia during Wallerian Degeneration

The simplest model of nerve injury is axotomy of fiber tracts. Transection or crush of a fiber tract in the nervous system leads to breakdown of axoplasm within days, loss of axonal connectivity, and to cellular and molecular changes in the distal nerve segment

referred to as Wallerian degeneration. In the peripheral nervous system (PNS) nerve fibers promptly regenerate from the proximal stump into the degenerating distal nerve segment, but no such regrowth occurs in the CNS. One of the prerequisites for axonal regeneration in the PNS is the rapid removal of growth-inhibitory myelin debris from the degenerating distal nerve segments that is facilitated by infiltration of hematogenous macrophages within the first 2 weeks. After transection of the optic nerve (a component of the CNS) and of nerve fiber tracts in the brain or spinal cord, hematogenous macrophages are largely or completely excluded, probably because the blood–brain barrier remains intact. Thus, growth-inhibitory myelin components in the distal stump persist for weeks. Only at the site of the transection where the blood–brain barrier is disrupted is a rapid macrophage infiltration observed. Although microglia could substitute hematogenous macrophages by phagocytic transformation and thereby compensate for the failure of macrophage entry into degenerating CNS fiber tracts, microglial activation is surprisingly slow and insufficient to clear myelin debris. In the PNS the complement system and lectins play a major role in myelin removal by macrophages. MAC-2 is a galactose-specific lectin mediating myelin phagocytosis by macrophages in the PNS. MAC-2 is not induced in microglia in degenerating fiber tracts in the CNS. Interestingly, the phagocytic activity of both macrophages and isolated brain-derived microglia could be enhanced upon their exposure to sciatic nerve segments but was inhibited by exposure to optic nerve segments. In conclusion, both the insufficient activation and the active suppression of microglia by unknown mechanisms in conjunction with the lack of macrophage entry may partly account for regeneration failure in the CNS. Elucidation of the different molecular signaling during Wallerian degeneration in the PNS and CNS can be expected to have important clinical implications for nerve repair.

B. The Prompt Microglial/Macrophage Response in Necrotic CNS Lesions

1. Microglial Responses to Stab Wounds and Bacterial Infection

In CNS injury leading to neuronal necrosis with breakdown of the blood–brain barrier (BBB), microglia are promptly activated and aided by infiltrating monocytes/macrophages from the circulation. As stated previously, at the stage of phagocytic transformation microglia and macrophages become indistinguishable by morphological grounds and by immunocytochemistry. Therefore, the population of phagocytes is referred to as microglia/macrophages in this article. The normal CNS appears to be relatively resistant to leukocyte diapedesis since intracerebral injections of proinflammatory cytokines such as IL-8, IL-1, and TNF-α elicit only minimal monocytic recruitment into the CNS parenchyma. However, necrotic brain lesions induce prompt inflammation. The critical role of breakdown of the BBB to macrophage recruitment into CNS lesions was demonstrated in the following experiment: Blood monocytes were prelabeled to allow follow-up of their distribution after injury. Although a mechanical CNS lesion with breakdown of the BBB led to prompt recruitment of these prelabeled peripheral monocytes, degeneration of inferior olivary neurons induced by intraperitoneal administration of 3-acetylpyridine, a neurotoxic process that respects the integrity of the BBB, led to microglial activation only.

The most simple traumatic CNS lesion with destruction of the BBB is a stab wound. A device is penetrated into the cortex. The ensuing cellular response is limited to the site of penetrating injury and the immediately surrounding tissue. Giulian and colleagues described activation of microglia at the lesion edges and infiltration of the wound site by hematogenous macrophages within hours after injury. Microglia proliferated locally with a maximum on Days 2 or 3, and mononuclear phagocytes rapidly cleared trauma-induced debris and produced IL-1, a cytokine that stimulates astrogliosis and neovascularization. Macrophages are attracted to tissues by chemokines. After penetrating mechanical CNS injury, astrocytes and endothelial cells increased mRNA levels of the chemoattractant monocyte chemoattractant protein-1 (MCP-1) and within 12 hr expressed MCP-1 protein, followed by prompt macrophage infiltration.

Microglia are involved in the formation of bacterial brain abscesses. Brain abscesses were produced experimentally in the rat by direct intracerebral injection of agarose beads laden with *Staphylococcus areus*. This approach allowed a sequential analysis of glial and inflammatory responses during different stages of abscess development. The first stage, acute cerebritis, was characterized by brain edema and diffuse infiltration of a wide rim of edematous brain parenchyma by microglia/macrophages around a necrotic center filled with granulocytes. Between Days 4 and 8 chronic cerebritis developed and microglia strongly

upregulated CD4 and MHC class II molecules. At this stage increasing numbers of hematogenous macrophages were recruited and gathered around the necrotic lesion. This finally led to capsule formation, regression of the abscess, and resolution of the edema. Interestingly, a subpopulation of macrophages recognized by the antibody ED2 in the rat preferentially infiltrated brain abscesses. Their functional properties in comparison to those of the usual population of $ED1^+/ED2^-$ macrophages in other brain lesions have not been characterized.

2. Microglia and Cerebral Ischemia

Local microglial activation and hematogenous macrophage infiltration are hallmarks of focal cerebral ischemia, which also leads to prompt breakdown of the BBB. Microglia are activated early in the border zone of infarcts, retract processes, round up, and partly transform into phagocytes (Fig. 4). In conjunction with a dramatic influx of hematogenous macrophages during the first 2 weeks after focal ischemia, microglia rapidly remove necrotic debris. Activation of microglia is accompanied by upregulation of MHC class I and II molecules. Morphological transition of microglia from normal brain tissue toward the infarct border can regularly be seen in human infarcts of recent age (Fig. 4). Macrophages are presumably attracted by chemokines released by astrocytes in the penumbra zone of infarction. The relative contribution of microglia/macrophages to the pool of phagocytes was formally assessed by depletion experiments in the model of photochemically induced focal ischemia in the rat. Macrophages can be depleted temporarily from the blood and lymphoid organs by intravenous application of dichloromethylene diphosphonate-containing liposomes, a treatment that does not affect microglia. In these macrophage-depleted rats there was no difference in the number of phagocytes on Day 3 after ischemia, but there were significantly fewer phagocytes on Day 6 compared to controls. This means that in the initial phase after focal ischemia, microglia to some extent transform into phagocytes and macrophages are recruited secondarily to aid in the removal of debris. The timing of macrophage recruitment as shown here for photothrombotic ischemic lesions may differ in other stroke models and other species.

Further phenotypical characterization revealed that besides $CD4^+$ microglia/macrophages, which are regularly seen in degenerating fiber tracts in the CNS, a novel population of microglia/macrophages appeared surrounding the infarcts predominantly between Days 3 and 6. This observation was independent of the mode of infarct induction and was a consistent finding also after permanent occlusion of the middle cerebral artery. By confocal laser microscopy and reverse transferase-polymerase chain reaction, we could show that this subpopulation of microglia/macrophages expressed the $CD8\alpha/\beta$ heterodimer, so far only described on cytotoxic/suppressor T cells. Secondarily degenerating fiber tracts and nuclei after cerebral ischemia showed the slow microglial activation and low phagocytic activity known from transected CNS fiber tracts as decribed previously and did not contain $CD8^+$ microglia/macrophages. The functional role and particular cytokine profile of this novel $CD8^+$ microglia/macrophage population in the rat CNS are unknown. Moreover, it is unclear whether this distinct microglia/macrophage population also exits in species other than the rat.

Reactive phagocytes can release neurotoxins after ischemic and traumatic injury to the CNS and thereby could aggravate tissue damage. However, neurotoxicity of microglia/macrophages in ischemic brain lesions has not been formally proven *in vivo*. The cytokine TNF-α apparently plays a key role in ischemia-induced microglial responses, although functional data are conflicting. In mice genetically deficient in TNF receptors, microglial responses were attenuated although neuronal damage increased after focal ischemia. In addition to their presumed neurotoxic actions, microglia/macrophages can also exert beneficial effects through production of neurotrophic factors.

Microglial activation is not restricted to infarcts and their immediate border zones but encompasses the entire ipsilateral, but not contralateral, hemisphere. A similar remote activation occurs in astrocytes that upregulate glial fibrillary acidic protein. This microglial and astroglial activation is transient and can partly be blocked by the NMDA receptor antagonist MK-801. This suggests that the remote glial responses are induced by periinfarct spreading depression like depolarization during the first hours after focal ischemia. The functional consequences are currently under investigation.

In contrast to the almost immediate onset of neuronal loss in permanent focal ischemia, transient global cerebral ischemia leads to delayed neuronal death in vulnerable areas. After brief periods of global ischemia (up to 15 min) damage is mainly restricted to CA1 pyramidal cells in the hippocampus. In areas of impending neuronal loss that can be delineated at 2–4

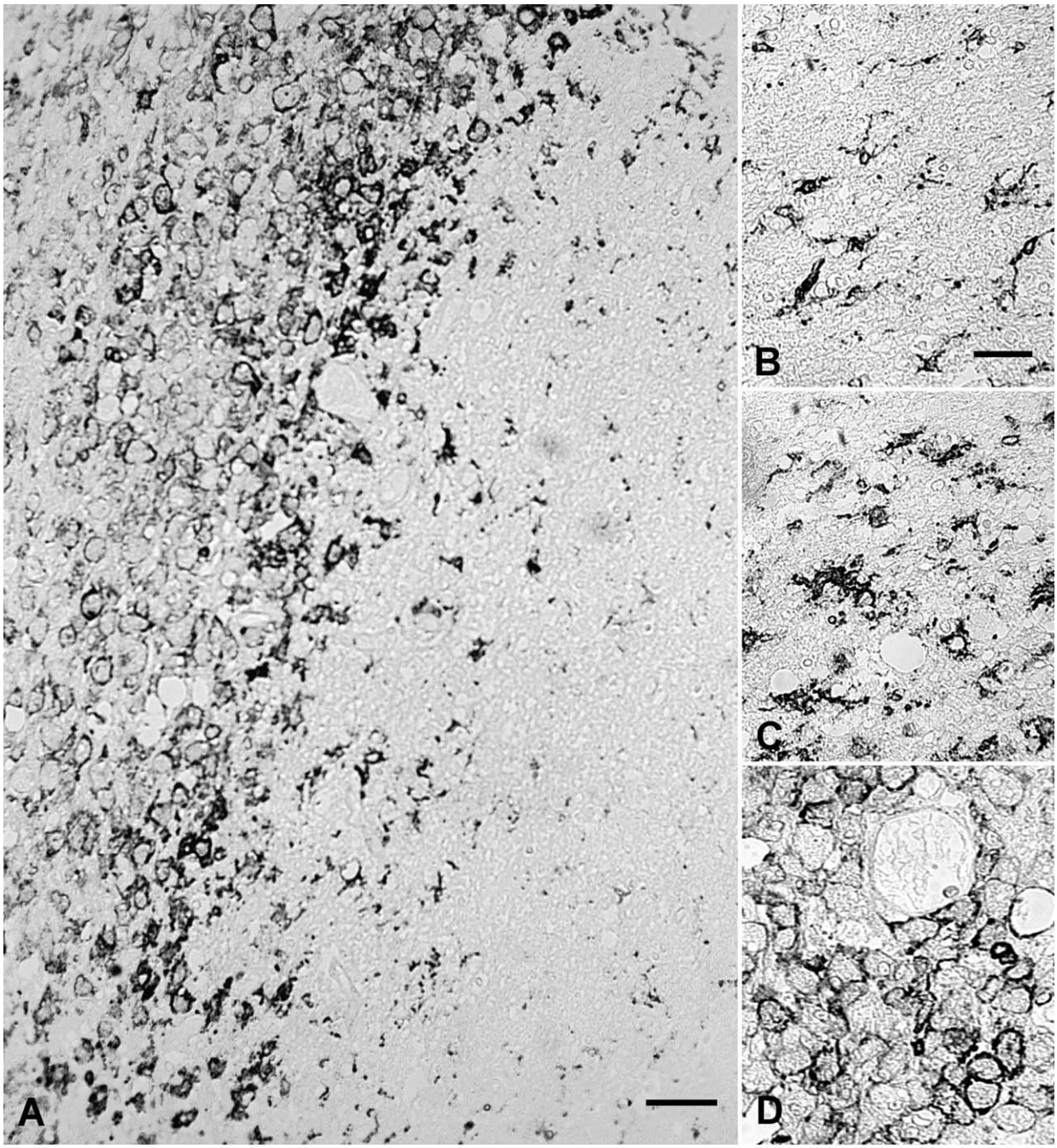


Figure 4 Microglia and macrophage responses in the human brain 5 days after cerebral ischemia (autopsy case). (A) The infarct is on the left side. Immunocytochemistry for the activation marker MHC class II reveals characteristic morphological changes in microglia from normal brain tissue (right) toward the edge of infarction (left). (B–D) Details of this morphological transformation from the typical ramified microglia (B) to ''stout'' microglia with loss of ramification (C) and finally to round phagocytes (D). At the stage of phagocytic transformation within the inner infarct border zone microglia and hematogenous macrophage are indistinguishable. Scale bars = 50 μm in A and 25 μm in B–D.

days after global ischemia, microglia show signs of activation within 24 hr after reperfusion and later cluster around degenerating neurons. In areas with a higher hypoxic resistance microglial activation is also observed, but in the absence of neuronal degeneration it occurs only transiently.

IV. MICROGLIA AND NEURODEGENERATION

A. Synaptic Stripping

Transection of the facial nerve is a useful model for studying microglial responses in the absence of infiltrating macrophages, as shown by the seminal work of Kreutzberg and colleagues. The facial nerve is cut outside the brain and the reactions of facial motoneurons and their glial environment to retrograde axonal degeneration can be studied in the brain stem, which shows no disturbance of the BBB in this lesion paradigm. After transection of the facial nerve microglia proliferated within 3 or 4 days mainly around the corresponding motoneurons and expressed long-standing activation markers (CD4, MHC class I and II, and cell adhesion molecules), amyloid precursor protein, and the cytokines TNF-α and TGF-β. TGF-β is a cytokine with a potential role in tissue repair. From Day 4 after axotomy onward, proliferating perineuronal microglia ensheathed the soma of the injured motoneurons and detached afferent synaptic terminals from the motoneuron surface, a process designated synaptic stripping. Synaptic stripping is an important process in synaptic reorganization after injury. Interestingly, microglia were activated but did not phagocytose unless the motoneurons underwent lethal injury. After selective death of motoneurons induced by retrogradly transported toxic ricin, a paradigm that also leaves the BBB intact, surrounding microglia transformed into phagocytes and removed dying neurons without the aid of hematogenous macrophages. Similarly, kainic acid-induced neuronal degeneration led to phagocytic transformation of microglia in the hippocampus.

B. Microglial Responses in Degenerative CNS Diseases

A pathogenic role of microglia has been suggested in degenerative CNS diseases, such as Alzheimer's disease (AD) and Parkinson's disease, that are characterized by selective loss of neurons in distinct regions of the brain. These conditions lack an overt inflammatory response, but microglia are activated at sites of neuronal degeneration. AD is characterized pathologically by the presence of insoluble structures in the cerebral cortex that represent either extracellular plaques consisting of amyloid-β protein (Aβ) deposits or neurofibrillary tangles within nerve cell bodies. In AD brain microglial cells accumulate in the center of plaques and in regions of perivascular deposits of Aβ. Further examination of the relation of microglial activation to different stages of plaque development revealed a preferential association of activated MHC class II-positive microglia with diffuse amyloid deposits and of enlarged phagocytic microglia with amyloid plaques. This strong association has been commonly interpretated as a beneficial microglial attempt to phagocytose and remove neurotoxic Aβ. In fact, microglia can internalize Aβ fibrills by scavenger receptors, but the rate of Aβ degradation is limited and the cells appear to be subsequently overwelmed by the amount of Aβ present. As an alternative possibility, activated microglia have been suggested to contribute to neurotoxicity. In support of this view, Aβ in culture together with the cytokine IFN-γ activated microglia to produce reactive nitrogen intermediates and TNF-α. This process led to neuronal cell injury *in vitro*. Treatment of microglia with a secreted derivate of β-amyloid precursor protein (APP) led to activation of the transcription factor NF-κB with ensuing induction of IL-1β and inducible NOS. Taken together, these findings indicate that β-APP can activate microglia and enhance their neurotoxicity.

The importance of microglial activation in degenerative CNS diseases has further been highlighted in an animal model of globoid cell leukodystrophy (GLD). GLD is a severe demyelinating disorder of the CNS. It is characterized by an increased number of MHC class II-expressing microglia/macrophages in the CNS but no T cell infiltration. Twitcher mice serve as an animal model for GLD. Mating twitcher mice with MHC class II-negative knockout mice led to a profound clinical improvement accompanied by reduced numbers of microglia/macrophages in the CNS in comparison to MHC class II-positive littermates. This points to a pathophysiological role of microglial MHC class II molecules in neurodegeneration independent of a T cell response. The underlying molecular mechanisms are unknown.

V. THE ROLE OF MICROGLIA IN CNS AUTOIMMUNITY AND INFLAMMATION

Microglia exert multiple functions in CNS autoimmunity and inflammation. Microglia can present antigens to T cells, produce cytokines and toxic radicals, and exert effector functions in CNS inflammation. Microglia are very sensitive to systemic stimuli. Intraperitoneal injections of the nonspecific immune activator

LPS led to increased IL-12 expression in parenchymal microglia and cyclooxygenase-2 activity in perivascular microglia. Moreover, microglia proliferated in response to circulating cytokines as shown for IFN-γ.

A. Antigen Presentation and Effector Functions

Experimental autoimmune encephalomyelitis (EAE) is a widely used model of autoimmune CNS disease that mimics several aspects of human multiple sclerosis (MS). It is the prototype of a T cell-mediated CNS disorder and can be induced in susceptible animals by immunization with CNS myelin components or adoptive transfer of T helper cells primed against central myelin. Clinically, the disease is characterized by weight loss, paralysis, and ataxic gait developing at the end of the second week after immunization. Depending on the immunization protocol and the species used, animals recover spontaneously and disease subsides or animals develop a relapsing–remitting course. Histopathologically, EAE is characterized by massive T cell and macrophage infiltration in the CNS and variable degrees of demyelination. The interactions between T cells and macrophages are orchestrated by cytokines. In EAE, the balance between pro- and antiinflammatory cytokines most likely determines the clinical course. In the acute phase of the disease, strong expression of proinflammatory cytokines occurs, but it subsides during the recovery phase.

Myelin-primed T cells entering the CNS in EAE have to be restimulated locally to evoke clinical disease. The fundamental role of perivascular microglia as antigen presenting cells in T cell responses in the CNS has been established in chimeric rats by Hickey and Kimura. Systemic injection of lymphocytes of the T helper phenotype ($CD4^+$) from Lewis rats directed against CNS myelin into Lewis rats elicited EAE. A transfer of the same T cells into Dark Agouti (DA) rats provoked no disease. The reason for this discrepancy lies in the fact that infiltrating T cells need to be restimulated in the CNS by local antigen presenting cells in the context of the same MHC haplotype. Lewis rats are of the MHC $RT1^1$ haplotype, and DA rats express a different $RT1^a$ MHC molecule. Hickey and Kimura created (Lewis $\times$ DA) F1 hybrid rats. Then, DA rats were lethally irradiated to remove all bone marrow-derived cells and received bone marrow from F1 hybrid rats. Two months later, 90% of lymph node cells of these bone marrow-transplanted DA rats expressed the Lewis MHC molecules ($RT1^1$), an indication that a high percentage of the immune system was repopulated with donor F1 hybrid cells. In these chimeras only bone marrow-derived cells of the F1 hybrid donor could express molecules of the Lewis MHC ($RT1^1$). When T cells of Lewis rat origin were injected into the chimeric DA animals, they developed EAE. This means that a bone marrow-derived cell of Lewis origin ($RT1^1$) must have become established in the CNS of the chimeric DA rats to restimulate the infiltrating T cells. Further immunocytochemical analysis confirmed that bone marrow-derived cells in the CNS of these chimeric DA rats corresponded in location to perivascular microglial cells previously described on morphological grounds. Moreover, these perivascular microglia expressed macrophage antigens. This important study not only clarified the origin of perivascular microglia but also established a functional role of perivascular microglia in immune responses as antigen presenting cells in the CNS. It appears that despite their extensive expression of MHC class II antigens, parenchymal microglia are not necessary for the induction of a T cell response in the CNS since these microglia were of the $RT1^a$ haplotype in Hickey and Kimura's experiments and thus they were noncompatible with the transferred Lewis rat ($RT1^1$) T cells. The concept of an immunological dichotomy of perivascular and parenychmal microglia was further supported by *in vitro* studies showing that only a minority of isolated CNS microglia/macrophages had the capacity to stimulate encephalogenic $CD4^+$ T cells, whereas the majority of ramified microglia were ineffective antigen presenting cells.

Ramified parenchymal microglia are activated and act as effector cells in the inflammatory reaction in concert with infiltrating hematogenous macrophages both in EAE and in human MS. In EAE and MS lesions, microglia/macrophages do not only respond to T cell-derived cytokines but also may undergo autocrine activation through endogenously produced IFN-γ and thereby entertain a chronic inflammatory process even at a stage in which T cells have been eliminated by apoptosis. Within and at the border of actively demyelinating MS and EAE lesions, as a sign of activation MHC class II expression on microglia is abundant. Moreover, microglia show features of transformation into phagocytes, suggesting an active role in myelin destruction. *In vitro*, TNF-α is able to destroy myelin and oligodendrocytes. Activated T cells can induce TNF-α secretion by microglia. Microglial TNF expression has been demonstrated in MS and EAE lesions. EAE could be inhibited by drugs that

selectively blocked the production of TNF-α or by soluble receptors that presumably antagonized the effects of TNF-α on the surface of responding cells. Accordingly, EAE was more severe in transgenic mice overexpressing TNF-α in the CNS. In these transgenic animals the demyelination process proceeded chronically in the absence of T cells and was accompanied by persisting microglial/macrophage activation. In TNF knockout mice, T cells and macrophages infiltrated the CNS but failed to move forward from the perivascular position into the CNS parenchyma, suggesting a pathophysiological role of TNF in the final effector phase of EAE. Experiments conducted in knockout mice, however, produced conflicting results with respect to the role of TNF. Conversely, in another study TNF knockout mice developed more severe EAE, suggesting a potent antiinflammatory action of TNF-α in autoimmune-mediated demyelination. Comparison between these studies is difficult because different animal strains, gene targeting strategies, and encephalitogenic protein antigens were used. The differential role of resident microglia versus hematogenous macrophages as sources of TNF-α in this context remains to be defined.

B. Microglia and Downregulation of Immune Responses

The CNS has a high potential for eliminating infiltrating T cells. T cell apoptosis is considered an important mechanism by which immune responses in the CNS are limited and probably accounts for spontaneous clinical recovery in EAE. Apoptosis, a form of programmed cell death, is characterized by chromatin condensation and cell shrinkage and, in contrast to necrosis, initially preserves the integrity of the cell membrane. Thus, leakage of intracellular components is avoided and inflammatory responses to dying cells are very limited. Several studies have shown that in EAE autoreactive T cells are eliminated by apoptosis. This process can be enhanced by glucocorticosteroid or high-dose antigen therapy. The precise *in vivo* mechanisms of T cell apoptosis in EAE have yet to be elucidated. There is evidence that activated ramified microglia are involved in T cell apoptosis. When tested directly *ex vivo*, purified MHC class II-positive parenchymal microglia stimulated autoreactive T cells and elicited production of the cytokines IL-2, IFN-γ and TNF-α; surprisingly, however, T cells subsequently died by apoptosis after this interaction. In contrast, perivascular microglia and meningeal macrophages led to full T cell activation and supported T cell survival. These unexpected results were confirmed by others but are in apparent contrast to results of several previous studies showing efficient stimulation of antigen-specific T cell proliferation by activated cultured microglia. The reasons for these discrepancies are unclear. Species differences or different protocols used in the generation of antigen-specific T cell lines may be responsible. The recent results raise the possibility that parenchymal microglia are more involved in downregulation than in activation of T cell responses in the CNS.

Costimulatory molecules such as B7-1 and B7-2 expressed by antigen presenting cells play an important role in T cell activation. There is evidence that B7-1 signaling shifts immune reactions toward a proinflammatory Th1 response, whereas B7-2 signaling induces an antiinflammatory Th2 phenotype on T cell populations. Interestingly, human resting microglia from brain biopsies isolated and analyzed immediately *ex vivo*, expressed the antiinflammatory costimulatory molecule B7-2, whereas the B7-1 molecule was expressed on microglia in and around active MS plaques. In the rat, *ex vivo* isolated microglia lacked significant B7-1/B7-2 expression, again indicating the importance of species differences. Currently, the molecular mechanisms underlying the presumed downregulatory role of parenchymal microglia are vague and deserve further investigation.

See Also the Following Articles

ASTROCYTES • AUTOIMMUNE DISEASES • GLIAL CELL TYPES • MULTIPLE SCLEROSIS • NERVOUS SYSTEM, ORGANIZATION OF

Acknowledgments

Work in the authors laboratory was supported by the Deutsche Forschungsgemeinschaft. Guido Stoll held a Hermann and Lilly Schilling professorship.

Suggested Reading

Del Rio Hortega, P. (1932). Microglia. In *Cytology and Cellular Pathology of the Nervous System* (W. Penfield, Ed.), Vol. 2, pp. 481–534. Hoeber, New York.

Fedoroff, S. (1995). Development of microglia. In *Neuroglia* (H. Kettenmann and B. R. Ransom, Eds.), pp. 162–181. Oxford Univ. Press, New York.

Ford, A. L., Foulcher, E., Lemckert, F. A., and Sedgwick, J. D. (1996). Microglia induce CD4 T lymphocyte final effector function and death. *J. Exp. Med.* **184,** 1737–1745.

Hickey, W. F., and Kimura, H. (1988). Perivascular microglial cells of the CNS are bone marrow-derived and present antigen in vivo. *Science* **239,** 290–292.

Kreutzberg, G. W. (1996). Microglia: A sensor for pathological events in the CNS. *Trends Neurosci.* **19,** 312–318.

Ling, E. A., and Wong, W. C. (1993). The origin and nature of ramified and amoeboid microglia: A historical review and current concepts. *Glia* **7,** 9–18.

Minghetti, L., and Levi, G. (1998). Microglia as effector cells in brain damage and repair: Focus on prostanoids and nitric oxide. *Prog. Neurobiol.* **54,** 99–125.

Perry, V. H., Brown, M. C., and Gordon, S. (1987). The macrophage response to central and peripheral nerve injury. A possible role for macrophages in regeneration. *J. Exp. Med.* **165,** 1218–1223.

Stoll, G., and Jander, S. (1999). The role of microglia and macrophages in the pathophysiology of the CNS. *Prog. Neurobiol.* **58,** 233–247.

Stoll, G., Jander, S., and Schroeter, M. (1998). Inflammation and glial responses in ischemic brain lesions. *Prog. Neurobiol.* **56,** 149–171.

Midbrain

DAVID M. WAITZMAN and DOUGLAS L. OLIVER
University of Connecticut Health Center

GLOSSARY

binaural Auditory processing that involves two ears.

binocular Visual processing that involves two eyes.

cerebral aqueduct The portion of the ventricular system in the brain connecting the third and fourth ventricles.

cuneiform nucleus Mesencephalic reticular formation.

gaze movement Combined movements of eyes and head to shift the direction of the fovea.

inferior colliculus Caudal to the superior colliculus in the tectum, involved in auditory sensory processing.

mesencephalon The midbrain. One of the five main parts of the brain.

periaqueductal gray The region of neurons that surrounds the cerebral aqueduct.

superior colliculus The most dorsal portion of the midbrain (tectum), involved in visual–motor control.

tectum The dorsal portion of the midbrain, sensory. Includes superior and inferior colliculus.

tegmentum The ventral portion of the midbrain, motor.

vergence movements Movements of the eyes toward the nose (in opposite directions).

The midbrain or mesencephalon, one of the five major subdivisions of the human brain, sits literally at the middle of the brain and forms the junction between the forebrain (thalamus and telencephalon) and the hindbrain (medulla and pons). The midbrain is divided into a dorsal *tectum* and ventral *tegmentum* by a plane that bisects the long axis of the cerebral aqueduct. The tectum forms the roof of the midbrain and includes the pretectum, superior colliculus, and inferior colliculus (Fig. 1). Together, the four colliculi (both sides of the inferior and superior colliculi) form the quadrigeminal plate, the anterior most portion of which is the superior colliculus. The tectum develops from the alar plate in the embryo. Hence, the neurons of the tectum are usually related to sensory function. In contrast, the tegmentum is the basement of the midbrain and is a basal plate derivative in the embryo. Consequently, the neurons of the tegmentum are more closely tied to motor function, and the tegmentum contains the motor nuclei of the extraocular muscles, substantia nigra, and reticular formation (Fig. 1).

This article is organized according to these major sensory and motor divisions. First, we will discuss the role of the midbrain in the auditory system where the primary focus is the inferior colliculus and sensory processing. Second, we will turn to the control of eye movements by the midbrain. This topic includes the superior colliculus as a focus of sensory–motor integration and the ventral midbrain structures used for motor control of the eyes and head. This article will conclude with discussions of pain modulation by the periaqueductal gray and the blood supply of the midbrain. A list of abbreviations is provided in Table I.

I. THE AUDITORY SYSTEM OF THE MIDBRAIN

The auditory system extends from the medulla to the telencephalon. Almost all parts of the auditory system are connected to the midbrain. The present treatment

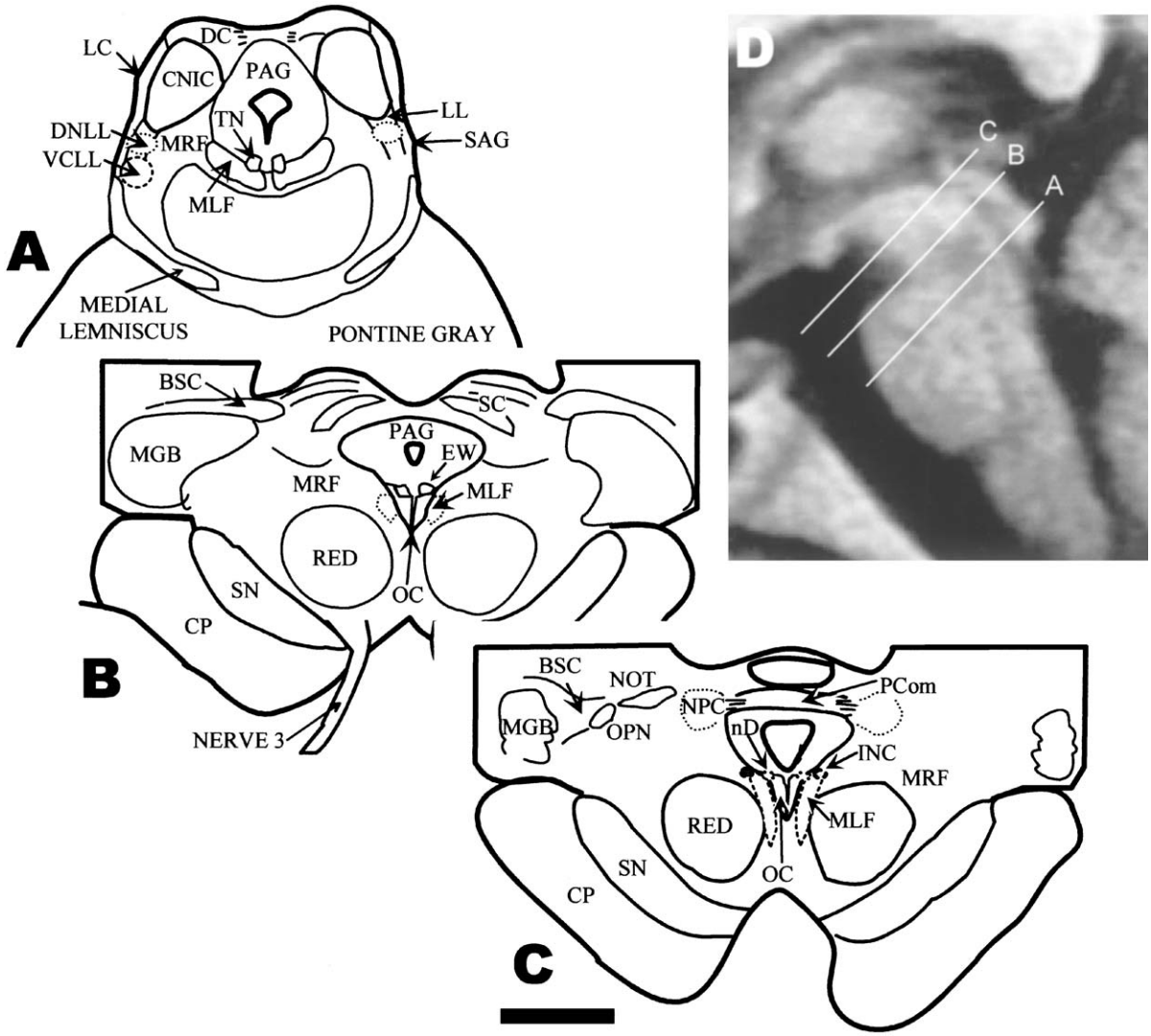


Figure 1 Anatomy of the midbrain in drawings of transverse sections. (A) Caudal level of inferior colliculus and trochear nucleus. (B) Middle level showing the superior colliculus and oculomotor nucleus. (C) Rostral level of the posterior commissure and pretectum. (D) Sagittal MRI of a normal brain at the midline showing the levels of the sections. A list of abbreviations is provided in Table I for this and all subsequent figures.

of the auditory system emphasizes the organization from the perspective of the midbrain.

A. Components of the Auditory System in the Midbrain

1. Inferior Colliculus (IC)

As the main component of the auditory system in the midbrain, the inferior colliculus represents the posterior part of the tectum (Fig. 1A). Its major role is to receive and integrate ascending input from most of the auditory system in the hindbrain and send its output toward the auditory cortex via the medial geniculate body in the thalamus (Fig. 2). It also receives major inputs from the cortex and has descending projections. The main part of the inferior colliculus is the large and prominent central nucleus. Surrounding the central nucleus, dorsally and caudally, is a cortex, and laterally and medially the paracentral nuclei are found. For each of these regions, we will describe the cellular organization, inputs, and outputs.

a. Central Nucleus of Inferior Colliculus (CNIC) The central nucleus is the major midbrain

Table I
List of Abbreviations

BA	Basilar artery
BSC	Brachium of superior colliculus
CNIC	Central nucleus of inferior colliculus
CP	Cerebral peduncle
DC	Dorsal cortex of inferior colliculus
DNLL	Dorsal nucleus of the lateral lemniscus
EW	Edinger–Westphal nuclei
INC	Interstitial nucleus of Cajal
LC	Lateral cortex
LL	Lateral lemniscus
LR	Lateral rectus muscle
MLF	Medial longitudinal fasciculus
riMLF	Rostral interstitial nucleus of the medial longitudinal fasciculus
MGB	Medial geniculate body
MR	Medial rectus muscle
MRF	Mesencephalic reticular formation (cMRF, central part of mesencephalic reticular formation)
ND	Nucleus Darkschewitsch
NOT	Nucleus of the optic tract
NPA	Anterior pretectal nucleus
NPC	Nuclei of the posterior commissure
NPP	Posterior pretectal nucleus
OC	Oculomotor complex
OPN	Pretectal olivary nucleus
OKN	Optokinetic nystagmus
P1	Proximal segment of the posterior cerebral artery
P2	Distal segment of the posterior cerebral artery
PAG	Periaqueductal gray
PC	Pretectal complex and the accessory optic system
PCA	Posterior communicating artery
Pcom	Posterior commissure
Post Com Art	Posterior communicating artery
RED	Red nucleus
SAG	Nucleus sagulum
SAI	Stratum albican intermediale
SAP	Stratum albican profunda
SC	Superior colliculus
SGI	Stratum griseum intermediale
SGP	Stratum griseum profunda
SGS	Stratum griseum superficiale
SO	Stratum opticum
SZ	Stratum zonale
SN	Substantia nigra
TN	Trochlear nucleus
VCLL	Ventral complex of the lateral lemniscus

component of the auditory pathway. In the human midbrain and most mammalian species, the central nucleus is defined by the presence of *fibrodendritic laminae*. The fiber component represents the ascending afferent axons from virtually all parts of the auditory system in the lower brain stem. Individual laminar axons form parallel sheets, several hundred micrometers wide, that extend in a ventrolateral to dorsomedial and rostrocaudal plane. These laminae represent an anatomical substrate for the *tonotopic organization* of the inferior colliculus. Acoustic stimuli at the lowest frequencies stimulate laminae in the dorsolateral part of the central nucleus, and successively higher frequencies stimulate laminae ventromedially (Fig. 2).

In the fibrodendritic laminae, the dendrites arise from the principal neurons in the central nucleus and are defined by their morphology. Disk-shaped neurons have flat, narrow, planar dendritic fields that parallel the planes of the laminar axons. The dendritic field of the large disk-shaped cell in humans is up to 1 mm long, and small disk-shaped cells have dendritic fields 600 μm long. In contrast, the width of the dendritic field is only around one-fourth the width of the axonal lamina (50 μm). This means that the disk-shaped dendrites are usually contained within a laminar plexus. A second, less common cell type is defined by a stellate or less flat dendritic morphology. Stellate cells can be small or large and have dendritic fields whose longest axis is either parallel or perpendicular to the fibrodendritic laminae.

Cell types in the central nucleus also are defined on the basis of their neurotransmitter content and by the target of their axons. A majority of the disk-shaped neurons in the inferior colliculus probably use glutamate as the neurotransmitter; however, a significant proportion of these cells use γ-aminobutyric acid (GABA) as the neurotransmitter (about 20% in the cat). Stellate cells can also synthesize GABA. Both GABAergic and non-GABAergic neurons can send their axons to other parts of the brain to form connections in the auditory pathway.

Neurons in the central nucleus are characterized by their membrane properties. At least six types of neurons can be defined using electrophysiological and pharmacological criteria in brain slice preparations in rats. The data suggest the presence of one or more unique potassium or calcium currents in each cell type, and these are likely related to unique combinations of potassium and calcium ion channels. Most cells in the central nucleus have membrane properties that tend to modify the inputs, e.g., act like filters. For

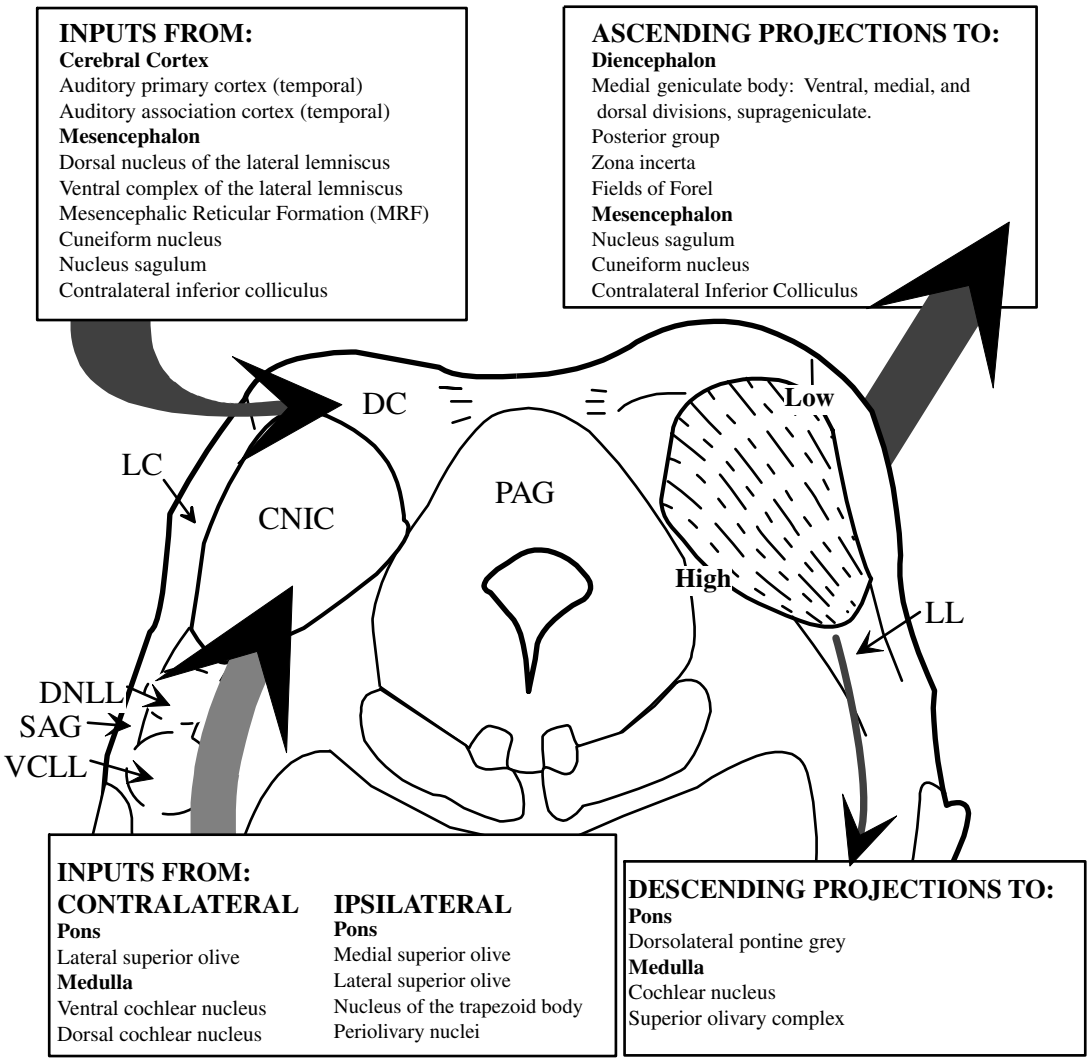


Figure 2 Inputs and outputs of the inferior colliculus. The fibrodendritic layers of the central nucleus are illustrated (right) by dashed and dotted lines. High and Low indicate high- and low-frequency laminae and the tonotopic organization. Inputs to the inferior colliculus are shown on the left side and outputs on the right side. Caudal structures are listed at the bottom, and rostral structures are listed at the top. See Table I for abbreviations.

example, in some neurons the firing rate will adapt (decrease in rate) slowly, whereas others will adapt quickly (a transient response) in response to continuous stimulation. An extreme version of the transient response is the onset response, firing once at the beginning of continuous stimulation. In contrast, some neurons in the central nucleus have responses that do not tend to modify their inputs, are not adaptive, and fire continuously in response to continuous stimulation. Thus, different membrane properties will regulate the ability of the neuron to fire and may set conditions under which it will and will not respond.

As the major midbrain component in the ascending auditory pathway, the inputs and outputs of the central nucleus include some of the main auditory pathways in the brain (Fig. 2). Ascending projections to the central nucleus arrive via the lateral lemniscus. The central nucleus receives binaural inputs from the superior olivary complex and monaural inputs from the cochlear nucleus. Of these, the largest number of synaptic inputs are excitatory inputs from the medial superior olive on the ipsilateral side. Other major excitatory inputs to the central nucleus are from the lateral superior olive, which projects to the central nucleus bilaterally, and the dorsal and ventral cochlear

nuclei on the contralateral side. The cochlear nucleus contains a variety of different cell types, but only the stellate cells from the ventral cochlear nucleus and the fusiform cells from the dorsal cochlear nucleus project to the inferior colliculus. Fine axonal inputs from the cerebral cortex to the central nucleus have been described. These are considerably fewer in number than cortical inputs to other parts of the inferior colliculus, and their influence is not well-established.

The central nucleus also receives distant projections, which are inhibitory. Much of the projection from the lateral superior olive on the ipsilateral side uses the neurotransmitter glycine and is likely to have an inhibitory influence on the cells of the central nucleus. Neurons in the dorsal nucleus of the lateral lemniscus use the neurotransmitter GABA and project primarily to the contralateral central nucleus. In contrast, the neurons in the ventral nucleus of the lateral lemniscus project to the ipsilateral central nucleus. Neurons in the ventral nucleus often contain both GABA and glycine; however, a specialized function is unclear for these colocalized neurotransmitters.

b. Cortex of Inferior Colliculus The central nucleus of the inferior colliculus is covered on its dorsal, caudal, and lateral surfaces by a cortex (Fig. 2). In most mammals including humans, a *dorsal cortex* (DC) and *caudal cortex* (CC) are easily recognized. These cortices contain stellate cells arranged in thick layers that are generally concentric and parallel to the surface of the inferior colliculus. At least four layers are found in the dorsal cortex of humans with the largest multipolar cells found in the deeper third and fourth layers. Smaller multipolar, stellate cells are found in the two superficial layers and in the caudal cortex. Disk-shaped cells are less common in cortex. Even when highly oriented dendrites are found, the neurons have local axons that are distributed widely in the dorsal cortex, unlike the disk-shaped cells of the central nucleus whose axons are confined to the same narrow lamina in which the dendrites are found. The lateral cortex of the inferior colliculus is subject to more variation between species. In humans, it is a relatively thin lateral zone. Because the lateral cortex has different connections and, possibly, a different cellular organization from the dorsal and caudal cortices, it is usually regarded as a separate part of the inferior colliculus (see later discussion).

The primary inputs to the dorsal and caudal cortices come from the telencephalon (Fig. 2). All parts of the auditory cortex project to the inferior colliculus. The heaviest inputs terminate in the dorsal cortex. Here, the axons from the telencephalon run perpendicular to the layers of the dorsal cortex, not parallel as in the central nucleus. Although all parts of the auditory cortex in the cerebrum project to the inferior colliculus, the subdivisions of the auditory cortex have different targets. Primary auditory cortex projects ventrally to the deepest layer of dorsal cortex, whereas the nonprimary auditory cortex projects to the superficial layers of the dorsal cortex. The deepest layer of the dorsal cortex, layer 4, receives inputs from both the ascending lateral lemniscus afferents and from the telencephalon. Thus, the deep dorsal cortex may have a special role to integrate ascending and descending pathways. The cortex of the inferior colliculus also receives inputs from the midbrain tegmentum and from the superior colliculus.

c. Paracentral Nuclei of Inferior Colliculus—Lateral Cortex (LC) Lateral to the central nucleus is the lateral cortex, also called external cortex (Fig. 2). The lateral cortex contains two or more cellular layers and is bounded superficially by the efferent fibers of the inferior colliculus that eventually form the brachium of the inferior colliculus. It has two parts, a superficial lateral zone (human), also called the lateral nucleus in the cat, and a *ventrolateral nucleus* at the ventral–lateral corner of the inferior colliculus. The ventrolateral nucleus may occupy a larger proportion of the lateral cortex in the rodent, whereas it is smaller and restricted to the ventrolateral corner of the colliculus in humans and cats. Despite differences in the nomenclature of this region, the structure of the lateral cortex is probably homologous in most mammalian inferior colliculi. The lateral cortex receives input from both auditory and nonauditory sources. Most of the inputs to the superficial, lateral zone are from the auditory cortex of the telencephalon and from the central nucleus. Other inputs include projections from the somatosensory system (spinal cord and dorsal column nuclei). In contrast to the lateral zone, the ventrolateral nucleus receives direct inputs from some auditory afferents that ascend to the midbrain via the lateral lemniscus. The dorsal cochlear nucleus and superior olive send important inputs to the ventrolateral nucleus.

d. Nucleus of the Rostral Pole The nucleus of the rostral pole is one of the paracentral nuclei of the inferior colliculus located rostral to the central nucleus. It contains stellate, multipolar cells organized without laminae. However, it is similar to the central nucleus in that it receives input from axons that ascend

from the lower auditory nuclei via the lateral lemniscus.

e. Commissural Nucleus The commissural nucleus of the inferior colliculus is medial to the central nucleus and dorsal cortex. It contains cells that lie among the fibers of the commissure of the inferior colliculus. Most inputs to the commissural nucleus probably originate from cells in the central nucleus and dorsal cortex of the inferior colliculus and terminate as the axons cross through the commissure.

2. Lateral Tegmental System

Ventral, medial, and rostral to the inferior colliculus are midbrain structures that have significant ascending connections with the auditory system. The lateral tegmental system includes the *superior colliculus*, *cuneiform nucleus* (see MRF), and *nucleus sagulum* (Fig. 1A). Although they do not receive direct input from the lower auditory system, they each project to the auditory part of the thalamus (the medial geniculate body). Thus, the lateral tegmental system may provide important nonauditory input to the auditory system.

3. Nuclei of the Lateral Lemniscus

The lateral lemniscus is a fiber bundle located in the lateral tegmentum of the midbrain that carries the axons from neurons in the cochlear nucleus and superior olive to the inferior colliculus. Embedded in the fibers are the nuclei of the lateral lemniscus: the dorsal nucleus and the ventral complex of nuclei.

a. Dorsal Nucleus of the Lateral Lemniscus (DNLL) The dorsal nucleus of the lateral lemniscus is situated at the top of the lateral lemniscus, just ventral to the inferior colliculus (Fig. 1A). It has many large, prominent neurons, and virtually all of the neurons contain GABA. Thus, the dorsal nucleus of the lateral lemniscus is usually thought to function as an inhibitory nucleus in the auditory pathway. The neurons in the dorsal nucleus have dendrites that are organized in horizontally oriented layers. The edges of the layers curl up in the ventral part and curl down in the dorsal part of the nucleus. Although the precise laminar arrangement has not been studied in humans, it may resemble the pattern seen in the cat rather than the more complex pattern seen in rodent species. These laminae are the substrate for a tonotopic organization with the lowest frequencies dorsally and the highest frequencies ventrally located. Most of the fibers of the lateral lemniscus penetrate the dorsal nucleus in fascicles and give off axon collaterals that contribute to the laminar organization of the dorsal nucleus.

Inputs to the dorsal nucleus arise from most of the same sources that project to the inferior colliculus (Fig. 2). These include the major input from the medial superior olivary nucleus (ipsilateral) and other projections from the lateral superior olivary nucleus (bilateral) and the ventral cochlear nucleus (contralateral). The dorsal nucleus also receives heavy inputs from the contralateral dorsal nucleus whose fibers make up the commissure of Probst. Most of these inputs are excitatory except for the input from the contralateral dorsal nucleus (GABA) and the ipsilateral projection from the lateral superior olive (glycine). Because of the major inputs from the superior olive, the dorsal nucleus of the lateral lemniscus contains neurons that are binaural and sensitive to interaural time and intensity differences.

The output of the dorsal nucleus is primarily to the inferior colliculus bilaterally and to the opposite dorsal nucleus. The projection from the dorsal nucleus to the central nucleus on the contralateral side is the major GABAergic input to the inferior colliculus. It terminates as a tonotopically organized, laminar projection whose ultrastructure and neurochemistry are characterized. Because of the large size of the dorsal nucleus neurons and axons, this inhibitory input reaches the colliculus quickly and may precede excitatory input from lower auditory structures in some cases. Inhibition in the inferior colliculus that arises from the dorsal nucleus may be important for many aspects of binaural information processing.

b. Ventral Complex of the Lateral Lemniscus (VCLL) The ventral complex of the lateral lemniscus is a collection of nuclei surrounded by the fibers of the lateral lemniscus (Fig. 1A). It extends from the rostral superior olivary complex in the pons to the ventral border of the dorsal nucleus beneath the inferior colliculus in the midbrain. The ventral complex contains at least three parts. A *ventral nucleus of the lateral lemniscus* is the main part. It is ventral and lateral and consists of many small, densely packed neurons. A majority of these neurons are immunoreactive for both GABA and glycine and are likely to release both at their axonal target. An *intermediate nucleus of the lateral lemniscus* is sometimes recognized as the region inserted between the ventral nucleus and the dorsal nucleus. It tends to have more sparsely distributed cells that are horizontally oriented.

Neurons in the intermediate nucleus are labeled with GABA and glycine less often, but they frequently have dense, glycinergic axosomatic terminals. A third region of the ventral complex is the *medial nucleus*. These are sparse neurons scattered in the fibers medial to the ventral nucleus. The ventral complex has a highly complex neuropil whose layers are suggested to resemble a helix. This complexity has delayed the comprehension of the frequency organization.

Most input to the ventral complex of the lateral lemniscus comes from the contralateral cochlear nucleus. Studies in the unanesthetized rabbit confirm that a majority of neurons of the ventral complex are monaural and respond best to sounds in the contralateral ear. However, binaural responses are seen and predominate in the medial nucleus. The binaural responses may be related to additional inputs from the ipsilateral cochlear nucleus or inputs from the superior olive.

The outputs of the ventral complex of the lateral lemniscus are primarily directed at postsynaptic targets located exclusively in the inferior colliculus on the same side of the brain. Because of the complexity of the neuropil, it had not been ascertained until relatively recently that they probably project in a topographic and tonotopic fashion. Because both GABA and glycine are found colocalized in the nucleus, it is possible that these neurons will inhibit their postsynaptic targets in the colliculus. Thus, both the dorsal nucleus and the ventral complex of the lateral lemniscus act to inhibit the activity of the inferior colliculus. Whereas the dorsal nucleus is predominantly GABAergic, binaural, and projects heavily to the contralateral colliculus, the ventral complex is predominantly glycinergic with GABA colocalization, often monaural, and projects exclusively to the ipsilateral colliculus.

B. Auditory Processing in the Midbrain

At its position at the halfway point in the auditory pathway, the inferior colliculus plays a fundamental role in the integration of information derived from the initial stages of auditory processing and transmission of this information to the auditory thalamus and, eventually, the auditory cortex. Essentially, all types of information about sound must pass through the inferior colliculus. Because the inferior colliculus is more accessible than the auditory structures of the lower brain stem, the inferior colliculus has been studied extensively with the aim of revealing the processing that occurred at previous stages of the auditory pathway as well as the processing that occurs at the midbrain. Much of the processing requires only one ear (monaural), whereas other types require the interaction of two ears (binaural).

1. Monaural Processing of Frequency, Sound Pressure Level, and Sounds with Complex Spectra or Timing

The inferior colliculus receives information about the spectrum of sound through all of its inputs, but most studies of frequency organization and spectral responses have emphasized monaural stimulation. There is only a single tonotopic map of frequency that extends across both the central nucleus and the dorsal cortex of the inferior colliculus. The characteristic frequency of single neurons in the central nucleus changes in the dorsolateral to ventromedial direction in steps of 180–200 μm, which correspond to the layers of the fibrodendritic laminae and the inputs from the auditory cortex. Characteristic frequency is defined as the frequency of sound to which the neuron responds with the lowest threshold, and it often corresponds to the frequency that drives the neuron at the highest rate at higher sound levels. There is evidence that the sharpness of tuning in the central nucleus is not uniform, and neurons with different sharpness of tuning could be mapped within the laminae.

The colliculus also participates in the processing of complex temporal signals such as speech signals, which include amplitude and frequency modulation. Neurons in the inferior colliculus have best modulation frequencies for amplitude modulation that may be mapped along the laminae of the colliculus. In some animals like the bat, there are neurons that are sensitive to the duration of a stimulus and sensitive to particular frequency modulations that are used in echo location. The role of the inferior colliculus in speech processing in the primate remains unclear. However, speech sounds are undoubtedly processed by the colliculus, and it possible that certain inputs to the colliculus are better excited by these complex signals than others.

2. Binaural Processing of Interaural Time and Intensity

The inferior colliculus receives extensive input from the superior olive that represents the initial stages for processing interaural time differences (ITD) and interaural intensity differences (IID). Both ITD and IID are necessary for the localization of sound, which is calculated neurally, unlike the localization of visual

or somatosensory stimuli, which are the result of spatially mapped receptor surfaces in the eye and skin, respectively. Inputs from the medial superior olive convey ITD information from low-frequency stimuli and are excited when coincident stimuli activate neurons. In contrast, the lateral superior olive appears to be sensitive to both IID and ITD information and covers the entire range of audible frequencies. However, the lateral superior olive tends to be inhibited when coincident stimuli arrive simultaneously. Thus, inputs from the medial and lateral superior olive have distinct properties, and the inferior colliculus appears to play an important role in preserving these properties.

3. Parallel Processing

One major question regarding the function of the inferior colliculus is the extent to which information from the different binaural and monaural pathways is transformed as opposed to relayed in an unchanged state. Both parallel processing and convergence of inputs from multiple sources appear to take place, and this conclusion is supported by both anatomical and physiological data. Parallel processing in the colliculus is supported by the continued presence of neurons with monaural responses and responses that mimic those in the medial and lateral superior olive. If all inputs were to converge onto the single frequency map in the colliculus, that preservation of response types would not occur. This principle of separation of binaural time and intensity pathways was established in avian systems (e.g., barn owl), and similar mechanisms are surely utilized by the human midbrain.

Anatomical evidence in mammals confirms that the inputs from different brain stem auditory centers do not all converge uniformly on the fibrodendritic laminae of the inferior colliculus. Instead, the laminae are probably organized into functional zones that each contain a unique set or group of laminar input axons that terminate together, in register with each other. Some inputs may terminate in one part of a lamina, whereas other inputs may terminate in another part or in a flanking lamina. The separation of parts of laminae into functional zones, called synaptic domains, may allow the basic response properties generated in the superior olive to remain intact.

4. Inhibitory Inputs to Central Nucleus Contribute to Transformations

Unlike other sensory systems, the auditory system of the midbrain contains a number of major nuclei (cell groups) that act as major inhibitory relays in the auditory network. Invariably, these inhibitory nuclei project to the central nucleus of the inferior colliculus. Thus, the functional zones of the inferior colliculus contain one or more excitatory inputs and one or more inhibitory inputs. Inhibitory inputs to the central nucleus originate primarily in the dorsal nucleus of the lateral lemniscus, the ventral complex of the lateral lemniscus, and the lateral superior olive. The binaural inputs from the dorsal nucleus of the lateral lemniscus are GABAergic and have been shown to influence or modify ITD and IID responses, especially in the contralateral central nucleus. Likewise, the binaural inputs from the lateral superior olive are glycinergic and are likely to modify binaural responses in the ipsilateral central nucleus. The frequently monaural inputs from the ventral complex are also likely to influence the ipsilateral central nucleus, but this is less well-established. Identification of the specific inputs to synaptic domains in the central nucleus of the primate is necessary before the full complexity of the auditory pathway in the midbrain can be appreciated.

C. Auditory Midbrain Outputs

1. Excitatory and Inhibitory Outputs from the Inferior Colliculus to the Thalamus

The major output of the central nucleus of the inferior colliculus is provided by a large number of neurons that project primarily to the ventral division of the medial geniculate body (ipsilateral) via the brachium of the inferior colliculus and the medial division of the medial geniculate (bilateral). These include GABAergic neurons (see later discussion). The central nucleus also has commissural projections to the contralateral inferior colliculus and descending projections to the ventromedial periolivary region of the superior olive (ipsilateral) and to the cochlear nucleus (bilateral).

The major outputs of the dorsal and lateral cortices are ascending projections to the dorsal and medial divisions of the medial geniculate body (ipsilateral). Dorsal cortex sends outputs to the dorsal and deep dorsal nuclei. Lateral cortex projections also terminate in the dorsal nucleus and in the margins surrounding the ventral division. Unlike the central nucleus, the dorsal and lateral cortices contribute few if any ascending projections to the ventral division of the medial geniculate. Thus, the dorsal and lateral cortices of the colliculus are unlikely to directly influence the projections from the thalamus to the primary auditory cortex of the telencephalon.

GABAergic projections ascend to the thalamus from the inferior colliculus in parallel with non-GABAergic projections. These parallel projections have not been confirmed in primates, but such projections would be consistent with their presence in diverse species like rats and cats. It is unusual for GABAergic cells to participate in ascending sensory projections to the thalamus, and it is not seen in the visual or somatosensory systems. Brain slice experiments in rodents have shown that short-latency, inhibitory postsynaptic potentials that use GABA-A receptors are produced in neurons of the medial geniculate after electrical stimulation of brachium axons. These short-latency potentials can precede the excitatory potentials from the central nucleus and may regulate the onset of thalamic excitation from the inferior colliculus. Many neurons in the medial geniculate receive both excitatory and inhibitory inputs from the inferior colliculus. However, some neurons in the medial geniculate receive only excitatory inputs from the midbrain.

2. Lateral Tegmental System

The superior colliculus, cuneiform nucleus, and nucleus sagulum comprise the lateral tegmental system. It may provide important nonauditory input to the medial geniculate body in the thalamus. This input to the thalamus is thought to be excitatory because few, if any, GABAergic neurons in the tegmentum or superior colliculus contribute to the thalamic projections.

The *superior colliculus* is rostral to the inferior colliculus and the other main part of the tectum (Fig. 1). It is primarily concerned with vision and the control of gaze, that is, the control of eye and head movements. However, the superior colliculus also receives inputs from the nonprimary auditory cortex and brachium of the inferior colliculus (see later discussion). Projections from the deep layers of the superior colliculus terminate in the dorsal division of the medial geniculate body, and these could provide information to the nonprimary auditory cortex about head and eye position.

The *cuneiform nucleus* (mesencephalic reticular formation, Fig. 1) is the most caudal part of the reticular formation in the mesencephalon. It is ventral to the inferior colliculus, medial to the lateral lemniscus, dorsal to the brachium conjunctivum, and lateral to the mesencephalic nucleus and tract of the trigeminal. The cuneiform nucleus receives input from the nonprimary auditory cortex among many other nonauditory sources. Outputs from the cuneiform nucleus terminate primarily in the dorsal division of the medial geniculate body, particularly in the dorsal and suprageniculate nuclei.

Nucleus sagulum is the tissue lateral to the nuclei and fibers of the lateral lemniscus (Fig. 1). It is a small-celled region of the tegmentum that is remarkably interconnected with parts of the auditory system. It receives inputs from nonprimary auditory cortex, inferior colliculus, and ventral nucleus of the lateral lemniscus and nucleus sagulum on the contralateral side. The major auditory projection of the nucleus sagulum is a projection to the dorsal division of the medial geniculate body.

3. To the Visual–Motor System: Nuclei of the Brachium of the Inferior Colliculus

Studies indicate that the brachium of the inferior colliculus may provide an important link between the auditory and the eye-movement systems of the midbrain. In mammals, there are few direct connections between the inferior and superior colliculi. Thus, the link from the auditory to visual–motor systems has not been obvious. However, such a link would be important because it would allow the location of sounds in space to be correlated with the positions of objects in space in the visual field. Sounds can initiate head and eye movements that are partially controlled by the superior colliculus. The fibers from the inferior colliculus terminate along the rostrocaudal extent of the nucleus of the brachium during their course en route to the thalamus. Neurons in the brachial nucleus project in a topographical manner to the superior colliculus, with the rostral brachium terminating rostrally and the caudal brachium terminating caudally. This rostrocaudal axis of the superior colliculus corresponds to the horizontal meridian of the visual field. Thus, the brachium projection to the superior colliculus is consistent with an auditory map of horizontal azimuth being transmitted from the auditory system to the visual system.

II. EYE MOVEMENT SYSTEMS OF THE MIDBRAIN

A. Components of Eye Movement Systems

1. Superior Colliculus (SC)

The superior colliculus forms the anterior roof of the midbrain (Fig. 1B). Its rostral border is the pretectum, and its caudal border is adjacent to the inferior

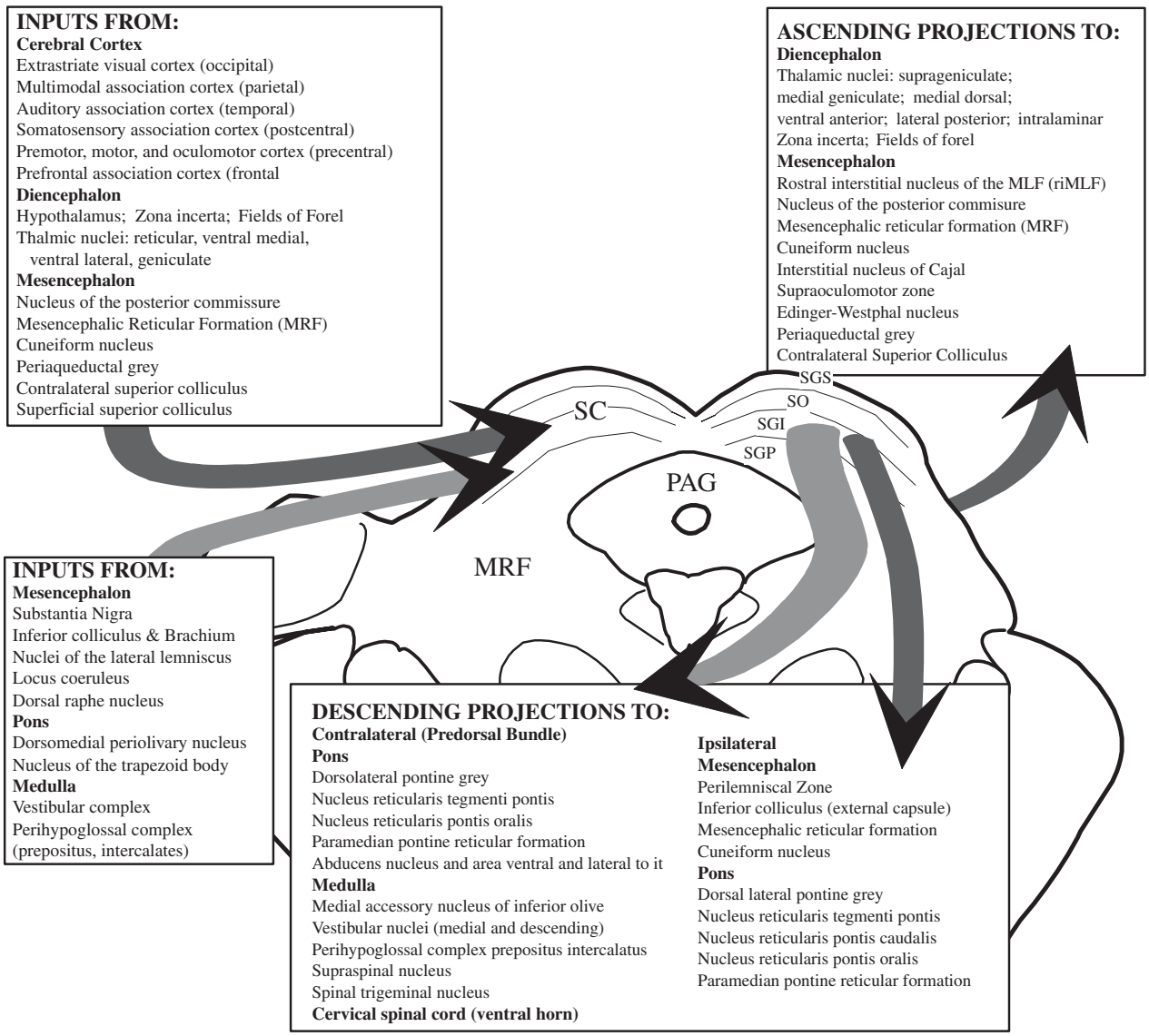


Figure 3 Inputs and outputs of the intermediate and deep layers of the superior colliculus. Layers of gray matter and optic fibers are indicated (right). Afferent connections are shown on the left side and efferent projections on the right side. Ascending projections are above the section of the superior colliculus and descending connections are below the section. See Table I for abbreviations.

colliculus. Seven alternating fiber and cellular layers make up the superior colliculus. However, the main functional divisions are a superficial visual part (three layers) and a deeper oculomotor part (four layers).

The superficial layers are composed of the stratum zonale (SZ), stratum griseum superficiale (SGS), and stratum opticum (SO) from dorsal to ventral, respectively (Fig. 3). Whereas retinal and cortical afferents innervate the SZ, SGS, and SO, the distribution of inputs to these areas is not even. Indeed, the stratum griseum superficiale can be further subdivided into three sublayers in terms of the distribution of inputs. SGS1 (25–120 μm from the surface) derives much of its input from W-type retinal ganglion cells. SGS2 occupies approximately 300 μm, and much of the input to this region is from areas 17, 18, and 19 of the visual cortex. The Y-type retinal ganglion cells synapse in SGS3. In general, the visual responses of individual superficial layer neurons are binocular, direction-selective, and have discrete retinotopically organized

receptive fields. The overall organization of retinal and visual cortical afferents to the primate superior colliculus places neurons with parafoveal receptive fields near the rostral pole, whereas neurons with peripherally located visual fields are located more caudally. The upper contralateral visual field is represented medially and the lower visual field laterally. Whereas the retinal contribution from the contralateral eye is evenly distributed across the colliculus, the input from the ipsilateral eye is densest in the central portion of the colliculus. This distribution leaves the neurons of the rostral pole of the primate colliculus driven solely by the contralateral eye (i.e., these cells are monocular). Furthermore, the contralateral retinal projection to the superficial layers is distributed to the dorsal SZ and SGS than the ipsilateral afferents that target the deeper portions of the SGS and most of the SO. Another aspect of the retinal afferents is that they tend to cluster in patches that extend across one or two of the superficial layers. Interestingly, these puffs or patches in the SZ and SGS receive predominantly contralateral, cholinergic input from the parabigeminal nucleus. Other subcortical afferents to the superficial layers arise from the ventral lateral geniculate nucleus, which primarily targets the ventral portion of the SGS and the SO. These projections are bilateral, but are predominantly ipsilateral. Last, the pretectum, especially the nucleus of the optic tract, sends a projection to the superficial layers.

The primary efferents from the superficial layers originate from the L-type neurons, small or medium-sized cells with elaborate dendritic trees. They participate in the ipsilateral descending and dorsal ascending tectofugal bundles. These cells provide synaptic input to the parent neuron through recurrent collaterals as well as more ventrally to the deeper tectal layers including the stratum griseum intermediale. The primary extratectal target of the L neurons is the parabigeminal nucleus. Other efferents arise from superficial layer neurons whose cell types have not been well-described. The stratum zonale provides efferents to the pretectal nuclei, lateral posterior pulvinar complex, and dorsal and ventral lateral geniculate nuclei. The more superficial portions of the SGS (layers 1 and 2) provide afferents to the dorsal lateral geniculate body, whereas the deeper SGS projects to the lateral posterior pulvinar.

The intermediate and deep layers of the superior colliculus (Fig. 3) are called the stratum griseum intermediale (SGI), stratum albican intermediale (SAI), stratum griseum profunda (SGP), and stratum albican profunda (SAP). These layers participate in the control of visually guided movements such as those of the eye, head, and arm. The inputs to the intermediate and deep layers include both descending cortical and subcortical contributions. Distinct from the superficial layers, the intermediate and deeper layers receive somatosensory, auditory, as well as visual inputs. The parietal (area LIP) and frontal cortices (frontal eye fields, supplementary eye fields, dorsolateral prefrontal cortex) provide much of the cortical, glutaminergic (i.e., excitatory) inputs to the intermediate and deep layers of the superior colliculus. The mesencephalic reticular formation and basal ganglia (primarily the substantia nigra pars reticulata) provide much of the GABAergic input (i.e., inhibitory) input to the superior colliculus. The afferents to the intermediate and deep layers of the superior colliculus from the pedunculopontine nucleus at the junction of the pons and midbrain are primarily acetylcholinergic (i.e., excitatory). Afferents to the intermediate and deep layers of the superior colliculus from the fastigial nucleus (deep cerebellar nuclei) and the nucleus prepositus hypoglossi are also well-described, but their transmitters have not been identified. For a listing of the afferent and efferent connections of the intermediate and deep layers of the superior colliculus, refer to Fig. 3.

Two types of efferent neurons have been distinguished in the intermediate and deep layers of the superior colliculus on the basis of their axonal branching patterns: X and T cells. Neurons of the X type are large multipolar neurons. Their cell bodies are located primarily in SGI and occasionally in the stratum opticum. They provide an axon that descends ventrally through the stratum albican profunda and then leaves the superior colliculus to reach the underlying periaqueductal gray. The axon circumscribes the periaqueductal gray and then crosses the midline ventral to the medial longitudinal fasciculus to reach the contralateral predorsal bundle. The X neurons do *not* emit any commissural collaterals to the opposite superior colliculus. In contrast, the cell bodies of the T cells are small to medium in size and reside within the ventral portion of the stratum opticum and the more dorsal portion of stratum griseum intermediale. The axons of T cells have an initial course similar to that of the X cells. Within the stratum griseum profunda, the axons of these cells ramify to provide a large number of recurrent collaterals. The primary axon then descends further to the periaqueductal gray where it bifurcates. For all T cells, one branch is directed across the collicular commissure to ramify upon neurons in the

intermediate and deep layers of the opposite superior colliculus. The destination of the second collateral can be variable. In some T cells, the second collateral branch emits collaterals around neurons in the ipsilateral mesencephalic reticular formation, whereas the main axon proceeds to circumscribe the periaqueductal gray to eventually join the contralateral predorsal bundle. The axon collaterals of other T neurons contribute to the ipsilateral ascending or descending projections of the superior colliculus. The axons of the X cells are large and probably provide much of the tectal input to cervical levels of the spinal cord. The smaller axons of the T cells probably do not reach spinal levels but end at pontine or medullary levels.

The organization of the neurons in the intermediate and deep layers of the superior colliculus generally follows that of the cells in superficial layers. The neurons in the intermediate and deep layers of the superior colliculus discharge in association with rapid eye movements. This discharge is much greater when the saccade is made in light than in dark. Two primary types of neurons are recognized. Burst neurons begin to fire 30–50 msec before the onset of a rapid eye movement that will bring the fovea within a specific portion of the oculomotor range. The collection of the movements for which the neuron will fire is called the movement field of the cell, in analogy to the visual receptive fields of neurons in the more superficial layers. Build-up neurons begin to discharge at low frequency as long as 130 msec before saccade onset and have a burst of discharges that occurs 30–50 msec before saccade onset. These neurons tend to have larger movement fields than the burst neurons, and the outside edge of their movement fields is less well-defined. With the advent of unrestrained head and eye (i.e., gaze) movement recordings, it is now recognized that neurons in the caudal portion of the intermediate and deep layers of the superior colliculus are best activated for gaze and not the separate eye or head components of the movement. As such the responses of a single neuron in the caudal SC have been described as a "gaze," not a "movement," field, i.e., the group of combined head and eye movements that cause the neurons to discharge. The neurons in the intermediate and deep layers appear to be organized in spatial register with the superficial neurons located just above. Thus, the rostral pole of the superior colliculus contains neurons that fire before small primary to secondary saccades. Cells with larger movement fields are located more caudally. Cells with upward movement fields are located medially, whereas neurons with downward movement fields are located laterally.

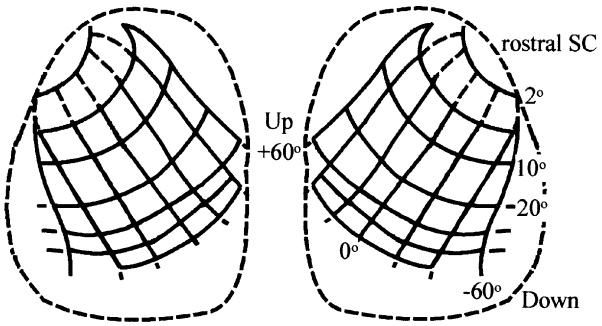


Figure 4 Physiological organization of the intermediate and deep layers of the superior colliculus based on electrical stimulation in head-fixed monkeys. Both colliculi are shown from above. The midline is between the two grids. Note that numbers along the sides of the diagrams are the amplitude of the saccade elicited in degrees. The numbers along the bottom indicate the elevation of the elicited eye movement. Positive numbers indicate upward movements found on the medial side of the colliculus, and negative numbers indicate downward movements found toward the lateral side of the colliculus. See Table I.

Electrical microstimulation in the SGI of monkeys whose head is restrained elicits saccades at short latency (<40 msec), whose maximal amplitude is primarily dependent upon the site, not the parameters of stimulation. These stimulation results have been used to generate a map of contralateral eye movement amplitudes that corresponds with the amplitude sensitivity of the cells recorded at each collicular site (Fig. 4). This will surely change in light of current experiments that are examining electrical microstimulation with the head free to move.

The possibility that the activity of the visual, superficial layers could be directly translated into the motor activity of the deeper collicular layers has driven research on the superior colliculus for longer than a century. Features of this concept have been demonstrated in many species of rodent (golden hamster and rat), but the interconnections of the superficial and deep layers have been less extensive in primates. Most likely, the reason for this species difference is dependent on the importance of the superior colliculus in visual processing. In afoveate animals (e.g., rodents), the superior colliculus is probably the primary visual and gaze processing region. Thus, direct projections from the superficial to the deep layers are advantageous. In foveate animals (i.e., monkeys and humans), the superficial layers of the superior colliculus play a much more secondary role in visual processing because of the size and complexity of the visual (occipital) cortex. The intermediate and deep layers, on the other

hand, retain considerable importance for the control of oculomotor and gaze-related behaviors, despite their duplication by other cortical bulbar pathways from the frontal eye fields and inferior parietal lobule of the parietal cortex. This is the reason that removal of the intermediate or deep layers of the superior colliculus (either by lesion or by reversible inactivation) in primates can produce an increased latency and some change in the trajectory of visually guided eye movements. Similarly, removal of the frontal eye fields alone produces minimal changes in eye movements. Simultaneous removal of both frontal eye fields and the superior colliculus produces a devastating reduction in the amplitude and speed of contralaterally directed saccadic eye movements. This deficit does not recover. In sum, the superficial layers of the superior colliculus in the primate provide visual information that is overshadowed by the amount of information provided by the occipital (visual) cortex. The intermediate and deep layers of the primate superior colliculus provide one of the parallel paths for the supranuclear control of gaze (head and eye movements.)

2. Oculomotor Complex

The oculomotor complex lies on the midline of the rostral midbrain beneath the aqueduct (Fig. 1) and contains the motoneurons whose fibers comprise the third cranial nerve innervating four of the six extraocular muscles. Medial rectus motoneurons are distributed throughout the rostral–caudal extent of the third nerve nucleus. There are three subgroups of medial rectus motoneurons. Large motoneurons of about 26 μm in diameter (cell group A) are located in the ventral portion of the rostral two-thirds of the oculomotor nucleus. Neurons of group B, again of large size averaging 30 μm in diameter, are situated in a dorsal lateral position and occupy the caudal two-thirds of the oculomotor complex. The motoneurons of groups A and B innervate the global "inner" muscle fibers situated close to the sclera of the eye. Smaller motoneurons 18 μm in diameter (cell group C) are a third population of motoneurons located in the dorsal portion of the rostral two-thirds of the oculomotor complex. Cell group C innervates the smaller "outer" fibers (i.e., the orbital muscle surface away from the globe of the eye) of the medial rectus muscle. The orbital portion of the medial rectus muscle inserts not on the globe, but on a specialized portion of the orbital tissue made up of collagen and connective tissue. This region serves as a pulley to permit shifts in the axis of rotation of the eye. Additional groups of large and medium-sized motoneurons in the oculomotor complex innervate the inferior rectus, superior rectus, and inferior oblique extraocular muscles. Interestingly, the inferior rectus motoneurons are intermingled with the medial rectus group C motoneurons in the dorsal portion of the rostral two-thirds of the oculomotor complex which participate in the generation of convergence and downward saccades. Evidence that small motoneurons of group C innervate primarily the orbital portion of the medial rectus muscle suggests that these fibers could participate in a caudal shift of the oculomotor pulley (located in the orbital tissues), thus shifting the muscle pulling directions during vergence or saccades from initial eye positions outside of primary (straight-ahead) position (i.e., secondary or tertiary positions).

A separate group of small interneurons resides within the oculomotor complex and has intracranial projections. The targets of these projections are varied but include the abducens motoneurons, reticularis tegmenti pontis, facial nucleus, cerebellum, spinal trigeminal nucleus, dorsal column nuclei, principal and dorsal accessory olivary nuclei, rostral medulla, parabrachial region, and cervical spinal cord. A role in divergence has been hypothesized for neurons projecting to the abducens nuclei, but the roles for the other projections have not been formulated.

The major inputs to the oculomotor complex include projections from the abducens nucleus and the nucleus prepositus hypoglossi (the location of the horizontal neural integrator). The terminals of the abducens internuclear neurons reach the midbrain by crossing through the contralateral abducens nucleus to the contralateral medial longitudinal fasciculus, which ascends to midbrain levels. They synapse directly on the motoneurons in the oculomotor complex, with the vast majority terminating on medial rectus motoneurons. Projections from the rostral interstitial nucleus of the medial longitudinal fasciculus (see later discussion) and the interstitial nucleus of Cajal have been demonstrated to reach the motoneurons of the primary vertical movers of the eye, including the inferior rectus, superior rectus, and inferior oblique extraocular muscles. The circuitry for the light reflex includes the crossed projections from the pretectal olivary nucleus via the posterior commissure to reach the Edinger–Westphal nucleus.

3. Medial Longitudinal Fasciculus (MLF)

The medial longitudinal fasciculus is a prominent axonal bundle running in the ventral portion of the

central gray that surrounds the aqueduct in the midbrain (Fig. 1). The medial longitudinal fasciculus connects the interneurons of the abducens nucleus with the motoneurons of the contralateral oculomotor nucleus and yokes the two eyes together. This provides a final common pathway for the generation of horizontal visually guided saccadic eye movements. The medial longitudinal fasciculus also carries fibers from the vestibular nuclei directed toward the oculomotor and trochlear nuclei as well as the interstitial nucleus of Cajal. These fibers are activated during head movements in which the vestibulo-ocular reflex is active.

4. Trochlear Nucleus

The large motoneurons of the trochlear nucleus are located within the central gray adjacent to the aqueduct of Sylvius and dorsal to the medial longitudinal fasciculus at the level of the inferior colliculus (Fig. 1). The axons of the motoneurons course caudally along the aqueduct and then ascend to the dorsal lateral aspect of the central gray. The nerves decussate completely within the anterior medullary velum (the roof of the aqueduct) to exit from the *dorsal* aspect of the brain stem just caudal to the inferior colliculus. The fourth cranial nerve then wraps around the pons lying between the superior cerebellar and posterior cerebral arteries. The nerve then ascends in the prepontine cistern along the free edge of the tentorium cerebelli before piercing the dural attachment of the tentorium to reach the cavernous sinus. The trochlear nerve enters the orbit through the superior orbital fissure to innervate a single muscle, the superior oblique. This muscle is a primary intorter of the eye in primary and abducted positions and a primary depressor of the eye in adducted (toward the nose) positions. Within the brain stem the trochlear nuclei receive direct projections from the rostral interstitial nucleus of the medial longitudinal fasciculus, which is the premotor nucleus for the coordination of vertical eye movements.

5. Mesencephalic Reticular Formation (MRF)

Anatomically, this region (Fig. 1, also referred to as nucleus cuneiformis) is linked to other oculomotor structures. It sends and receives projections from the superior colliculus. The mesencephalic reticular formation is heavily connected to the paramedian pontine reticular formation because it is reciprocally connected to the excitatory burst region and sends a heavy projection to the omnipause region (i.e., raphe interpositus) of the paramedian pontine reticular formation. The cells of the raphe interpositus pause for each saccade. The mesencephalic reticular formation provides contralateral descending projections to the nucleus reticularis tegmenti pontis (one of the major sources of oculomotor input to the cerebellum) and ipsilateral projections to the cervical spinal cord. It also provides ipsilateral ascending projections to the intramedullary lamina of the thalamus.

A number of different lines of evidence support a role for the mesencephalic reticular formation in the control of eye movements. Electrical stimulation in the mesencephalic reticular formation produces contraversive saccades at short latency and low threshold. The amplitude of these movements varies with the dorsoventral position within the mesencephalic reticular formation and not the strength of the stimulation: dorsal sites produce small contraversive horizontal saccades, whereas ventral sites produce larger, contraversive goal-directed movements. Cells in the mesencephalic reticular formation are long-lead burst neurons whose low-level discharge can begin up to 150 msec before eye movement and culminates in a high-frequency (>800 spikes/sec) burst that peaks just before the onset of spontaneous and visually guided saccades. Monkeys may develop a trimodal neglect of the contralateral sensorium (somatosensory, auditory, and visual) following electrolytic lesions of the mesencephalic reticular formation.

a. cMRF The central mesencephalic reticular formation (cMRF), located more caudally within the nucleus cuneiformis just lateral to the oculomotor nuclei (Fig. 1B), contains neurons that discharge before and during contraversive, horizontal saccades. The discharge characteristics of these cells and their anatomic connections suggest that the cMRF may participate in the local feedback pathway by providing current eye position and/or velocity to the superior colliculus or to the cerebellum via connections to the nucleus reticularis tegmenti pontis. The discharge is closely correlated with current eye displacement and/or velocity. In addition, about half of these neurons discharge in association with the appearance of visual stimuli and during the delay portion of a remembered saccade task.

b. riMLF Both clinical and experimental evidence has distinguished another subdivision of the midbrain reticular formation located in the prerubral fields near midline (Fig. 1C). The rostral interstitial

nucleus of the medial longitudinal fasciculus (riMLF) serves as the immediate premotor area for the coordination of conjugate vertical and torsional movements of the eyes. This region receives projections from other eye-movement-related areas, including the paramedian pontine reticular formation, the omnipause region (the raphe interpositus), the superior colliculus, and probably also the nuclei of the posterior commissure as well as the opposite riMLF.

The primary efferent connections of the riMLF are to the third (oculomotor) and fourth (trochlear) cranial nerve nuclei, the paramedian pontine reticular formation, the interstitial nucleus of Cajal, the vestibular nuclei, and the prepositus hypoglossi. The connections of the riMLF to the oculomotor and trochlear nuclei are ipsilateral for the downward movers (to the inferior rectus innervated by the third cranial nerve nucleus and to the superior oblique innervated by the trochlear nucleus) and bilateral to the upward movers (to the superior rectus and inferior oblique of the oculomotor nucleus). Connections from the riMLF to that of the opposite side travel across the posterior commissure to assist in the coordination of the upward movers of the eyes or travel across a commissure ventral to the aqueduct to coordinate downward movements.

6. Interstitial Nucleus of Cajal (INC)

The interstitial nucleus of Cajal is adjacent to the riMLF within the central gray (Fig. 1C). It is separated from the riMLF via the fasciculus retroflexus. As already indicated, fibers from the riMLF pass through the interstitial nucleus and provide an axon collateral. Interpretation of data from recordings, stimulation, and lesions within the interstitial nucleus and riMLF is difficult because these structures are tightly interconnected anatomically and physiologically. However, there are significant differences in the connectivity of these two structures. In contrast to the riMLF, the interstitial nucleus does not receive projections from the paramedian pontine reticular formation. In addition, the interstitial nucleus receives very strong vestibular input from the medial and superior vestibular nuclei as well as from the Y-cell group of the vestibular nuclei (i.e., it receives input from the vertical semicircular canals) and is closely connected to the opposite interstitial nucleus. Whereas earlier work suggested an input from the superior colliculus to the interstitial nucleus, more recent detailed analysis has not been able to substantiate this projection. Sparse input from the nuclei of the posterior commissure has been suggested.

The output of the interstitial nucleus is to the third, fourth, and seventh cranial nerve nuclei, as well as to the vestibular nuclei, the nucleus prepositus hypoglossi, the nucleus paramedianus, and the nucleus of Roller. The projections from the interstitial nucleus to the oculomotor and trochlear nuclei are specific for the vertical muscles and spare the medial rectus subdivisions of the oculomotor complex. In addition, the projection to the oculomotor complex is contralateral and provides a pathway for the riMLF to influence the contralateral vertical eye muscles. There are also very strong projections to the spinal cord via heavily myelinated fibers that extend to the lumbar region. The cells of the interstitial nucleus fire in the planes of head movement that activate either the anterior or the posterior semicircular canals. However, interstitial nucleus neurons that project to spinal levels are *not* responsive to vestibular stimulation.

The interstitial nucleus serves as the integrator for vertical and torsional eye movements. Unilateral electrical stimulation produces rotatory eye and head movements rather than purely vertical movements. Upward saccades were spared following bilateral interstitial nucleus lesions in cats. Impairments in downward saccades were thought to be the result of interruption of efferent fibers from the rostral riMLF. After unilateral kainic acid lesion of the interstitial nucleus in the cat, vertical saccades in all directions were preserved. Unilateral muscimol injections in the interstitial nucleus suggested that there was a significant impairment in vertical gaze holding (combined eye and head movement), producing vertical movements with torsional components (out of Listing's plane). Bilateral inactivation of the interstitial nucleus led to impairment of all vertical and torsional gaze holding. In addition, a contraversive head tilt has been documented following unilateral interstitial nucleus muscimol injections. The deficits in neck posture and vertical eye movements seen in patients with progressive supranuclear palsy have been attributed to the involvement of both the interstitial nucleus and riMLF.

7. Nuclei of the Posterior Commissure (NPC)

Five different cell groups of nuclei of the posterior commissure can be identified in the dorsal portion of the meso-diencephalic reticular formation: (1) the principal part, (2) the magnocellular part, bordering the central gray medially and within the central gray

itself, (3) rostral, (4) subcommissural, and (5) infracommissural portion below the posterior commissure. The principal and magnocellular parts are the ones typically observed. The afferents to the nucleus of the posterior commissure include a strong projection from the frontal eye fields. This cortical input is much more prominent than similar afferents destined for the paramedian pontine reticular formation or the riMLF. The superior colliculus and the dentate nucleus also provide other prominent inputs to the nucleus of the posterior commissure. The connection with the superior colliculus is reciprocal, with the nucleus of the posterior commissure providing the largest contribution of brain stem afferents to the superior colliculus.

Projections from the nucleus of the posterior commissure include one via the posterior commissure to the contralateral nucleus and a bilateral projection to the interstitial nucleus. There is some controversy about whether afferents of the nucleus of the posterior commissure reach the oculomotor nucleus. Most likely this projection is to the supraoculomotor central gray, which is concerned with lid movements. The nucleus of the posterior commissure also has a strong projection to the mesencephalic reticular formation and paramedian pontine reticular formation that may represent an alternative extracollicular pathway for cortical input for the control of horizontal saccadic eye movements (see later discussion). Smaller projections are present from the nucleus of the posterior commissure to the spinal cord and to the nucleus Darkschewitsch. Electrolytic lesions of the nucleus of the posterior commissure in monkeys and damage to this area in patients have demonstrated an upgaze paralysis and eyelid retraction (Collier's sign).

8. Nucleus Darkschewitsch (ND)

This small nucleus is located just rostral and dorsal to the interstitial nucleus of Cajal near the most rostral pole of the oculomotor complex (Fig. 1C). Its relationship to oculomotor control is unclear. It has strong projections to the dorsal cap of Kooy of the inferior olive (which projects to the flocculus of the cerebellum) and receives strong projections from the dentate nucleus of the cerebellum and from the cortex (pre- and postcentral gyrus). The afferents to the nucleus Darkschewitsch from the vestibular nuclei (via the medial longitudinal fasciculus) are small compared to the vestibular input to the interstitial nucleus of Cajal. The nucleus Darkschewitsch is closely associated with both motor (dentate nuclei and precentral gyrus) and oculomotor structures (providing floccular input), and it may participate in some aspects of hand–eye coordination (the dentate mediates control of distal upper limb segmental motoneurons).

9. Pretectal Complex and the Accessory Optic System

The optic tract fibers that do not synapse in the lateral geniculate nucleus are directed along the brachium of the superior colliculus toward the pretectum and the superior colliculus (Fig. 1C). The nuclei of the pretectal complex are situated within and below these fibers as they enter the dorsal portion of the brain stem at mesodiencephalic levels. The pretectal complex includes the nucleus of the optic tract, the olivary pretectal nucleus, the medial pretectal nucleus, and the posterior pretectal nucleus. These receive contralateral retinal efferents (Fig. 5). There is also a prominent contribution from the ipsilateral eye. There is a smaller contralateral projection from the optic tract to the medial pretectal nucleus. Other than the retina, the largest source of afferents to the pretectal nuclei is from the striate and adjacent visual association cortex, and there is also a projection from frontal eye fields and the cingulate cortex. Projections from middle temporal and middle superior temporal areas to the nucleus of the optic tract have also been demonstrated, which suggests that the nucleus may participate in smooth pursuit as well as velocity storage. Within the brain stem, the pretectal complex receives inputs from the ventral division of the lateral geniculate body, the superior colliculus, the accessory optic nuclei (medial, lateral, and dorsal terminal nuclei), and the cerebellum (dentate and interposed nuclei).

The pretectal complex provides both ascending and descending projections. The ascending efferents are primarily to the thalamic nuclei, including the intralaminar, dorsal and ventral lateral geniculate body, pulvinar, and zona incerta (subthalamus). The descending projections are clearly oculomotor and reach the interstitial nucleus of Cajal, the nucleus of the posterior commissure, the superior colliculus, the mesencephalic reticular formation, and the precerebellar nuclei in the pons and medulla (nucleus reticularis tegmenti pontis and inferior olive, respectively). The olivary pretectal nuclei have strong projections to the contralateral Edinger–Westphal nuclei via a crossing in the posterior commissure (see the section on Edinger–Westphal nuclei for functional considerations).

The accessory optic system comprises three nuclei: medial, lateral, and dorsal terminal nuclei. These

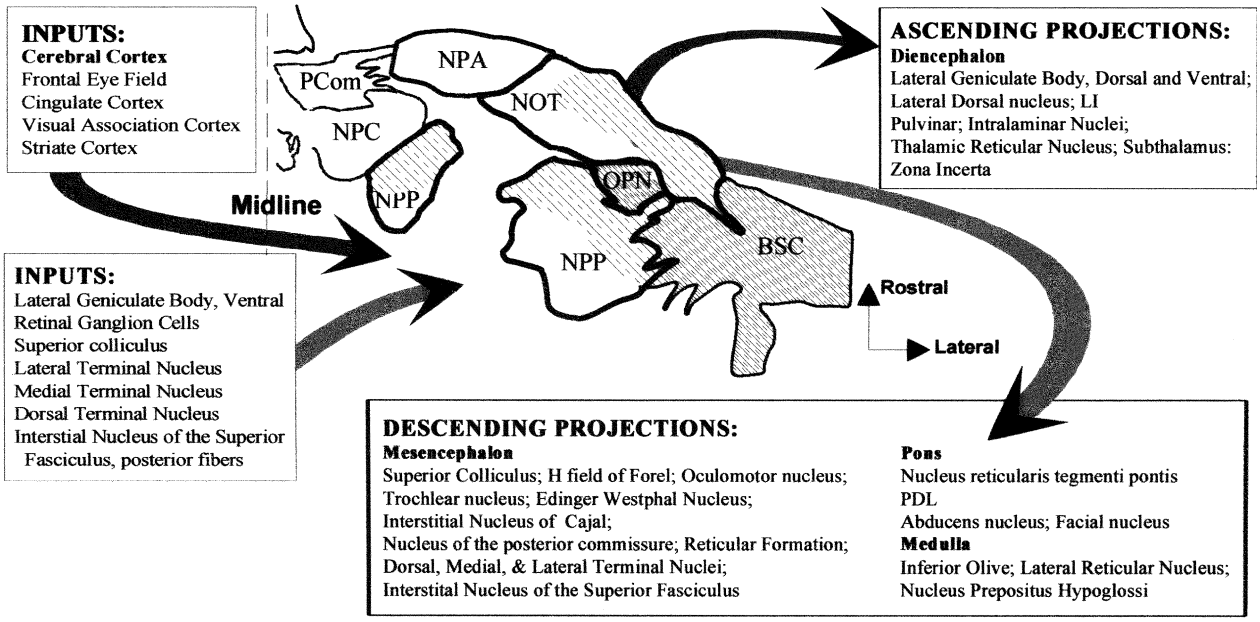


Figure 5 Inputs and outputs of the pretectal complex. The degree of shading in each portion of the horizontal section through the midbrain of the monkey illustrates the distribution of transported label over the contralateral pretectal complex following an intraocular injection of tritiated amino acids. The intensity of the hatching indicates the density of the label that was transported to the pretectum from the contralateral eye. Rostral is at the top and lateral is to the right. See Table I for abbreviations. (After Hutchins and Weber, 1985.)

nuclei are located in the rostral midbrain. The dorsal terminal nucleus is located just ventral to the nucleus of the optic tract at its most lateral extension. The lateral terminal nucleus is located ventral to the dorsal terminal nucleus, just dorsal to the substantia nigra, and lateral to the red nuclei. The medial terminal nucleus is located just lateral to the central gray and ventral to the red nucleus. These structures receive separate direct retinal input via a pathway known as the transpeduncular tract, which courses on the surface of the brain stem just anterior to the superior colliculus, travels over the brachium of the superior colliculus and then posterior to the medial geniculate body, and finally runs over the surface of the cerebral peduncle to enter the brain stem at the medial edge of the peduncle where it breaks into a number of fascicles to reach each of the terminal nuclei.

Functionally the nucleus of the optic tract and the dorsal terminal nucleus are essential in the production of the slow phases of optokinetic nystagmus (OKN) to the ipsilateral side. Electrical stimulation within the nucleus of the optic tract produces nystagmus. Lesions (both electrolytic and excitotoxic) in the nucleus of the optic tract and dorsal terminal nucleus reduce the slow rise of horizontal OKN and reduce or abolish optokinetic after-nystagmus. This suggests that the indirect path for velocity storage that produces compensatory eye-in-head and head-on-body movements via the vestibular system is impaired following lesions of the nucleus of the optic tract. This suggested that one role of the nucleus of the optic tract is to stabilize gaze in space during both passive motion and active locomotion in light. These effects are probably mediated via the strong input of the nucleus of the optic tract to the dorsal cap of Kooy (i.e., via the inferior olivary climbing fiber inputs to the flocculus) because the contribution to the medial vestibular nuclei is sparse. Evidence of cells with foveal receptive fields that are sensitive to retinal slip velocity suggested a role for the nucleus of the optic tract in smooth pursuit. This idea has been strengthened by evidence of a strong efferent projection from area MT to the nucleus of the optic tract. In the cat and rabbit, the majority of neurons in the medial and lateral terminal nuclei are sensitive to motion of the visual world along the vertical (i.e., vertical OKN). These nuclei are yet to be studied in the primate.

B. Control of Horizontal Gaze in the Midbrain

The primary focus of the current discussion is to point out particular characteristics and possible functions of midbrain structures that may participate in the

generation of horizontal saccades. The superior colliculus has received considerable attention in the control of rapid eye movements in both horizontal and vertical directions. A map of the contralateral field of movement has been demonstrated in the intermediate and deep layers of the primate superior colliculus using supramaximal electrical stimulation (Fig. 4). Movement fields (that is, the group of saccades that causes a cell to be activated during movements of particular size and direction) are characteristic of the neurons in the intermediate and deep layers of the superior colliculus. Cells in a particular portion of the colliculus have movement fields that correspond to the amplitude and direction of saccades electrically elicited from that site. Thus, activation of a particular group of neurons in the superior colliculus leads to the generation of an eye movement of particular size and direction. This has been called a *place code*. This pattern of neuronal activation is quite different from that actually required to rapidly move the eyes. The motoneurons of the abducens and oculomotor nuclei increase their discharge in relationship to the size of the upcoming movement, i.e., a *temporal code*. One of the primary problems that must be solved by supranuclear neurons in the midbrain oculomotor system is to change the place code activation of the SC into the temporal pattern of activity needed to rapidly shift the eyes. A second issue facing neurons arranged in a place code is how is this separated into horizontal and vertical signals required to move the eyes. The answers to both of these questions are not complete yet, but the following should provide some insight into how the brain stem performs these transformations.

The first step in solving these transformations is to reevaluate the connections of the superior colliculus with downstream structures involved in the generation of horizontal eye movement in the pontine reticular formation (Fig. 6). The superior colliculus provides both crossed mono- and polysynaptic inputs to the burst neurons in the paramedian pontine reticular formation (PPRF), the center for controlling horizontal gaze. A subgroup of burst neurons (the excitatory burst neurons) in the paramedian pontine reticular formation projects directly to the abducens nucleus, and its removal leads to a complete loss of rapid eye movement to the ipsilateral side. This signal is relayed by internuclear neurons, also located in the abducens nucleus, to the medial rectus subdivision of the oculomotor complex in the midbrain. The axons of these internuclear neurons travel in the medial longitudinal fasciculus to reach the midbrain (Fig. 6). At the same time, the excitatory burst neurons provide the same burst signal to the neurons of the nucleus prepositus hypoglossi. This structure is responsible for changing this pulse of activity (most closely associated with eye velocity) into a prolonged step of activity that is most closely related to eye position. This step response is then relayed back to the abducens nucleus and is responsible for holding the eyes steady at the new position. The abducens motoneurons provide an axon that reaches the lateral rectus muscle, which moves the eye toward the temporal side. Similarly, the oculomotor motoneurons provide an axon that travels in the oculomotor nerve to reach the medial rectus muscle, which moves the eye toward the nose. The combination of the PPRF, abducens nucleus, oculomotor nucleus, nucleus prepositus hypoglossi, medial rectus muscle, lateral rectus muscle, and

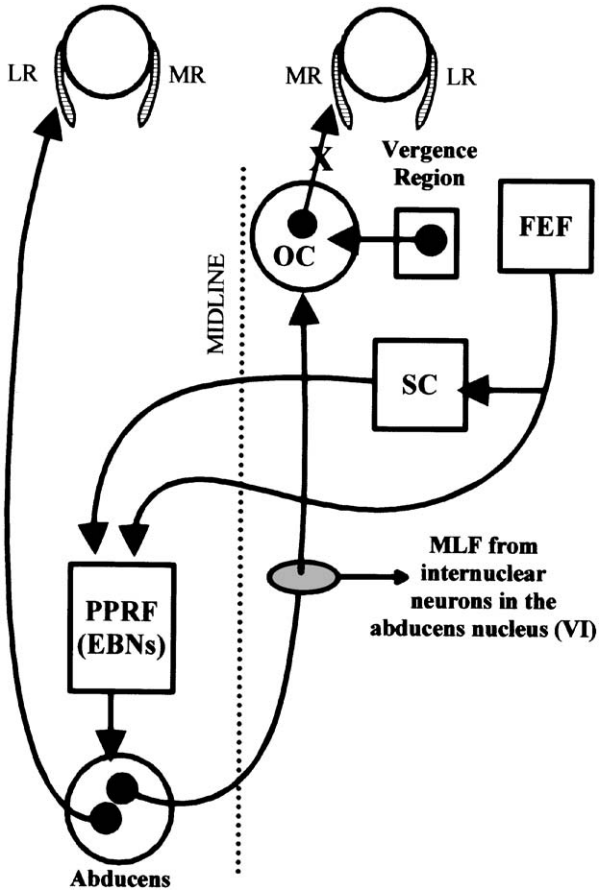


Figure 6 Schematic diagram of the final common pathway for horizontal saccades. Supranuclear inputs to the paramedian zone of the pontine reticular formation arise from the superior colliculus and frontal eye fields in parallel. Note that X marks the location of a third nerve lesion and the filled circle is a MLF lesion. A lesion in the MLF would preserve vergence, whereas a third nerve palsy would impair vergence. See Table I for other abbreviations.

medial longitudinal fasciculus is called the final common pathway responsible for the generation of horizontal saccadic eye movements.

Despite the clear connections of the final common pathway, how the superior colliculus directs eye movement control remains unclear. Work in the monkey suggests that the superior colliculus neurons provide monosynaptic connections to the long-lead burst neurons and not the excitatory burst neurons of the PPRF. In fact, most papers suggest that the input to the excitatory burst neurons of the PPRF is polysynaptic and not monosynaptic. The activity of long-lead burst neurons (LLBNs) is not uniform. Instead, a subgroup of LLBNs discharges only for particular directions (e.g., horizontal) of ipsilateral eye movements (directional LLBNs), whereas other LLBNs discharge for eye movements of very specific amplitude regardless of direction (i.e., vector long-lead burst neurons). As a result, one could generate a horizontal eye movement by projections of superior colliculus neurons to a select subgroup of vector LLBNs, which in turn activate the directional LLBNs and subsequently activate the excitatory burst neurons that project to the abducens nucleus. This scheme would explain the transformation from spatial coordinates in the superior colliculus to temporal coordinates in the excitatory burst neurons, as well as the selection of just the horizontal component of movement from a group of neurons that encodes both horizontal and vertical saccade directions.

Surprisingly, removal of the superior colliculus produces only minor eye movement deficits. An increased latency for saccades and the elimination of express saccades (i.e., saccades with latencies < 160 msec) are seen, but there are no changes in saccade accuracy. Following injections of the GABA agonist muscimol, which produces temporary inactivation of the intermediate and deep layers of the SC, saccade trajectory becomes curved and most saccades are hypometric (i.e., they fall short of the target). However, bilateral removal of both the superior colliculus and the neocortical frontal eye fields produces persistent deficits in the generation of rapid eye movements in all directions. This suggests that there are two pathways for saccade signals originating in the cerebral cortex to reach the paramedian pontine reticular formation: (1) a superior-colliculus-dependent route and (2) an extracollicular route. These experimental findings have been confirmed by clinical experience. No clinical oculomotor syndrome has been associated with damage to the superior colliculus alone. In fact, one patient examined following removal of an angioma in the superior colliculus showed increased latency for contralateral saccades, and these movements were hypometric.

On the other hand, damage (i.e., electrolytic lesion, stroke, or tumor) to the mesencephalic reticular formation is associated with distinct problems in contraversive horizontal saccades and ipsilateral smooth pursuit. One possibility is that mesencephalic reticular formation lesions interrupt the direct projections from the frontal eye fields to the paramedian pontine reticular formation. However, as pointed out earlier, these fibers are sparse. Midbrain lesions could also interrupt fibers from the nuclei of the posterior commissure to the paramedian pontine reticular formation or other parts of the reticular formation. More likely, corticobulbar projections from the frontal eye fields and parietal cortex reach the mesencephalic reticular formation and nuclei of the posterior commissure, and these structures relay them to the PPRF.

C. Control of Vertical Gaze in the Midbrain

A rostral portion of the midbrain reticular formation adjacent to the periaqueductal gray (riMLF) contains neurons whose primary on-directions are related to vertical gaze (combined head and eye movements), as do regions near the interstitial nucleus of Cajal (Fig. 1C). These burst neurons (both excitatory and inhibitory) fire during every vertical or torsional saccadic eye movement as well as during the quick phases of vertical, vestibularly induced nystagmus. Efferent projections from the riMLF synapse in, but also pass through and lateral to, the interstitial nucleus of Cajal (Fig. 7). As a result, electrolytic lesions or ischemic insult to the interstitial nucleus destroys the output of the riMLF as well as that of the neurons in the nucleus. This probably accounts for the long-held view that the interstitial nucleus of Cajal rather than the riMLF was the premotor center for coordinating vertical gaze. Instead, the cells in the interstitial nucleus produce a prolonged step of activity in response to every vertical rapid eye movement. This suggests that the interstitial nucleus is a critical structure for vertical gaze holding. In this way, it serves the same function for the vertical oculomotor system that the nucleus prepositus hypoglossus does for the horizontal eye movement system.

The projections from the riMLF and interstitial nucleus to the oculomotor nuclei account for the ability of these structures to generate vertical eye movements (Fig. 7). During downward movements, the riMLFs on both sides of the midbrain must be

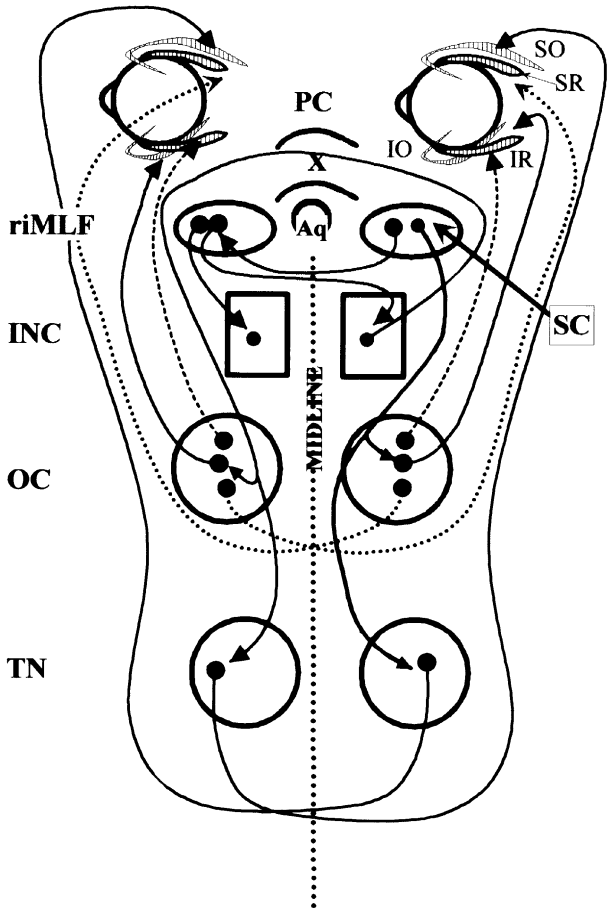


Figure 7 Schematic diagram of the final common pathway for vertical saccades. X marks the location of a cut in the posterior commissure (PC) that would produce the dorsal midbrain syndrome (see Table II). See Table I.

active. Projections from one riMLF reach the riMLF on the opposite side. There are direct projections from the riMLF to the inferior rectus and superior oblique motoneurons on the ipsilateral side. This produces a downward movement of *both* eyes (i.e., inferior rectus on the same side and the axons of the trochlear motoneurons cross the brain stem to reach the contralateral superior oblique muscle, which causes the opposite eye to be depressed and intorted). To coordinate upward movements, the riMLF provides efferent projections that synapse on the motoneurons in the superior rectus and inferior oblique subdivisions of the oculomotor complex on both sides. Similar to the efferent projections of the riMLF, the output of the interstitial nucleus is directed to the motoneurons of the oculomotor and trochlear nuclei as well as to the contralateral interstitial nucleus.

Inactivation experiments of the riMLF unilaterally demonstrated a 50% reduction in conjugate vertical eye velocity and a complete loss of one torsional direction. Clockwise (CW) rotation disappeared when the right riMLF was inactivated, whereas after inactivation of the left riMLF counterclockwise (CCW) torsion disappeared. Interestingly, changes in horizontal movements also occurred, suggesting that the short-lead burst neurons of the riMLF are divided into four groups: CW–up–left and CW–down–right in the right riMLF and CCW–up–right and CCW–down–left in the left riMLF. Bilateral inactivation led to complete abolition of both torsional and vertical saccadic eye movements. Unilateral inactivation or destruction (with kainic acid) led to similar modifications of torsional eye movements and a shift in Listing's plane. Clinical material supports the idea that the riMLF is the region equivalent to the paramedian pontine reticular formation (where horizontal movements are coordinated) for the coordination of eye movements in the vertical plane (see later discussion). On the other hand, unilateral inactivation of the interstitial nucleus of Cajal led to the loss of vertical gaze holding, skew deviation with hypertropia of the ipsilateral eye, extorsion of the contralateral eye, intorsion of the ipsilateral eye, and a contralateral head tilt (ocular tilt reaction). This confirmed the important role that the interstitial nucleus plays in maintaining vertical gaze positions. Similar to the discussion found on horizontal gaze, the separation of signals for the generation of vertical movements must occur as projections reach the riMLF. However, the same analysis of long-lead burst neurons found in the PPRF has not occurred for the vertical long-lead bursters. Finally, whereas projections from the superior colliculus, nucleus of the posterior commissure, long-lead burst neurons of the midbrain, and pontine reticular formations to the riMLF have been described, *no* direct projections from the frontal eye fields to the riMLF have been identified. This has suggested that cortical supranuclear control of vertical eye movements may be mediated through the nuclei of the posterior commissure, which receive a strong frontal eye field projection.

D. Clinical Correlation: Localization of Up and Down Gaze in the Midbrain

Experimental material and clinical cases agree that damage to the posterior commissure produces a paralysis of up gaze and a variety of associated

Table II
Features of the Dorsal Midbrain Syndrome

1. Limitation of upward eye movements:
Saccades, smooth pursuit, vestibulo-ocular reflex, Bell's phenomenon
2. Lid retraction (Collier's sign): occasionally ptosis
3. Disturbances in downward eye movements:
Downward gaze preference (setting-sun sign)
Downward saccades and smooth pursuit may be impaired, but vestibular responses are preserved
Down-beating nystagmus
4. Disturbances of vergence eye movements:
Convergence retraction nystagmus (Koerber–Salus–Eischnig syndrome)
Paralysis of convergence
Spasm of convergence
Paralysis of divergence
"A" or "V" pattern exotropia
Pseudo-abducens palsy
5. Fixation instability (square-wave jerks)
6. Skew deviation
7. Pupillary abnormalities (light–near dissociation)

findings (Table II). This condition is known by a plethora of names: Parinaud's syndrome, Koerber–Salus–Eischnig syndrome, pretectal syndrome, dorsal midbrain syndrome, and Sylvian aqueduct syndrome. Eyelid abnormalities also occur, such as Collier's tucked lid sign (i.e., lid retraction). Vergence may also be affected. Many of the features of the dorsal midbrain syndrome result from damage of the fibers crossing the posterior commissure, including those from the riMLF and interstitial nucleus destined for their namesakes on the contralateral side. Invariably, damage to other nearby structures may also occur during these lesions, the most important of which are the nuclei of the posterior commissure, the interstitial nucleus of Cajal, the fasciculus retroflexus, and the riMLF. Destruction of the riMLF bilaterally results in complete paralysis of all vertical eye movements (both up and down). Furthermore, as indicated earlier, damage to the interstitial nucleus of Cajal almost always includes damage to the fibers of passage from the riMLF. A single occlusion to the thalamoperforating artery has sometimes led to bilateral infarction of the rostral midbrain–posterior diencephalon near the midline and paralysis of vertical eye movements.

E. Autonomic Control of the Eyes in the Midbrain

1. Edinger–Westphal Nuclei (EW)

The Edinger–Westphal nuclei are located just dorsal to the third cranial nerve motoneurons and represent the visceral (parasympathetic portion) component of the oculomotor complex. The Edinger–Westphal nuclei provide preganglionic input to the ciliary ganglion mediating the pupillary light reflex and accommodation responses of the lens (i.e., the ciliary body). The nucleus of Perlia (located between the somatic nuclei of the primate) may also provide direct input into the iris and ciliary body, which bypasses the ciliary ganglion altogether. The primary input to the Edinger–Westphal nuclei is from the pretectal olivary nuclei, which receive direct crossed retinal input. Early physiologic and anatomic evidence supported direct ipsilateral and contralateral projections from the pretectal olivary nuclei to the Edinger–Westphal nucleus. Recent anatomic studies (injections of the retrograde tracer WGA–HRP into the Edinger–Westphal nuclei) have confirmed that there are no reciprocal connections between the olivary pretectal nuclei on opposite sides of the brain stem. Instead, each pretectal nucleus sends both contralateral and ipsilateral projections to each Edinger–Westphal nucleus. In addition, the Edinger–Westphal nuclei receive projections from the supraoculomotor region of the midbrain as well as from the fastigial nuclei of the cerebellum. Most neurons in the Edinger–Westphal nuclei also respond during changes in accommodation (see later discussion) used to view a near target.

2. Pupillary Light Reflex

The preceding connections permit a modern understanding of the pupillary light reflex. Light shone in one eye (e.g., the right eye) is conducted via the right optic nerve to both sides of the pretectum after crossing in the optic chiasm. The pretectal olivary nuclei on each side will then receive a neural signal related to the brightness of the light shone in one eye. This signal is relayed by each pretectal nucleus to both Edinger–Westphal nuclei. As a result there is a crossing at the chiasm and at least once in the brain stem, across the posterior commissure. A relative afferent pupillary defect (RAPD) reflects a disparity in the amount of pupillary light constriction in both eyes in response to light shone in one eye compared to light shone in the other eye. This occurs in patients who have an optic nerve lesion, and thus it localizes the defect in the

conduction of light impulses to the 2 cm of optic nerve anterior to the optic chiasm. It does not occur from a cataract or retinal disease (e.g., diabetic retinopathy).

3. Control of Vergence and Accommodation

The supraoculomotor region of the midbrain is directly involved in the control of the movement of the two eyes in opposite directions, i.e., vergence. This region is located just dorsal and lateral to the oculomotor nuclei. These neurons provide a direct, ipsilateral projection to the medial rectus subdivision of the oculomotor complex. In addition, there may also be a projection to the Edinger–Westphal nuclei on the same side, and this could account for the accommodative responses found in Edinger–Westphal neurons. The supraoculomotor region receives direct projections from the interposed nuclei of the cerebellum as well as from portions of the frontal and parietal lobes of the cerebral cortex. The cerebral regions that are known to project to this region include the prearcuate frontal cortex (area 8), area LIP, and areas MT and MST of the temporal–parietal lobes. The neurons in this supraoculomotor region are termed near-response cells because they increase their firing rate for near viewing, but not during vertical or horizontal conjugate eye movements. In addition, most of these neurons were tested in tasks that could distinguish vergence from accommodation. These neurons fired in association with changes in both vergence and accommodation. A small subset of the supraoculomotor neurons was sensitive solely to accommodation or vergence.

III. PAIN MODULATION AND OTHER MIDBRAIN SYSTEMS

A. Periaqueductal Gray (PAG)

The periaqueductal or "central" gray is a group of neurons that surrounds the aqueduct and is divided into four major zones: dorsomedial, dorsolateral, ventrolateral, and medial (Fig. 8). The medial portion immediately surrounds the aqueduct. Modulation of at least five primary behaviors has been attributed to portions of the periaqueductal gray: (1) modulation of pain pathways; (2) reproductive behavior (i.e., lordosis); (3) fear and anxiety; (4) vocalizations; and (5) autonomic regulation (i.e., control of blood pressure). The functions of the periaqueductal gray in reproductive behavior and vocalization have been demonstrated primarily in rats and cats and have not been investigated in humans. Moreover, studies have demonstrated coordinated responses between blood pressure, anxiety, and analgesia, which are mediated via competing columns of periaqueductal gray cells. As a consequence, the following discussion will be directed toward understanding three of the five primary physiologic functions: pain control, autonomic regulation (e.g., blood pressure, pulse, respiration, etc.), and the generation of fear and anxiety.

The role of the periaqueductal gray in gating pain information arising from the spinal cord has been known for at least 30 years. Electrical stimulation in the midbrain periaqueductal gray produced a level of anesthesia that was compatible with performing an exploratory laparotomy in a rat. Similar "endogenous" analgesia has been demonstrated during such activities as environmental stress, long-distance running, or sexual activity. Analgesia may also be modulated by mood and circadian rhythms. How are these functions organized?

Most of the antinociceptive effects of midbrain periaqueductal gray stimulation can be mediated via connections to the nucleus raphe magnus located in the ventral rostral medulla. In turn, the nucleus raphe magnus provides afferents to the dorsal horn of the spinal cord via the dorsolateral funiculus. Two primary pathways target the nucleus raphe magnus. An opioid-dependent pathway, located in the ventrolateral portion of the midbrain periaqueductal gray has been confirmed using a variety of experimental techniques. Injection of morphine into the ventrolateral portion of the periaqueductal gray induces analgesia. Periaqueductal gray neurons are inhibited by enkephalin, an endogenous opioid pentapeptide. The effects of stimulation in the ventrolateral portion of the periaqueductal gray (either by morphine or by electrical stimulation) can be reversed by the injection of Naloxone, an opioid antagonist. A second, nonopioid pain-modulating pathway arises from the dorsolateral portion of the periaqueductal gray. A similar level of analgesia can be generated by the injection of neurotensin or substance P, two peptide transmitters, into this region of the periaqueductal gray. The effects of these peptides and dorsolateral periaqueductal gray stimulation are not reversed by the injection of Naloxone, supporting the idea that the dorsolateral pathways are nonopioid-mediated. Some of these effects may also be modulated via *ascending* projections from the midbrain periaqueductal gray to the thalamus to cause a release of β-endorphin. The

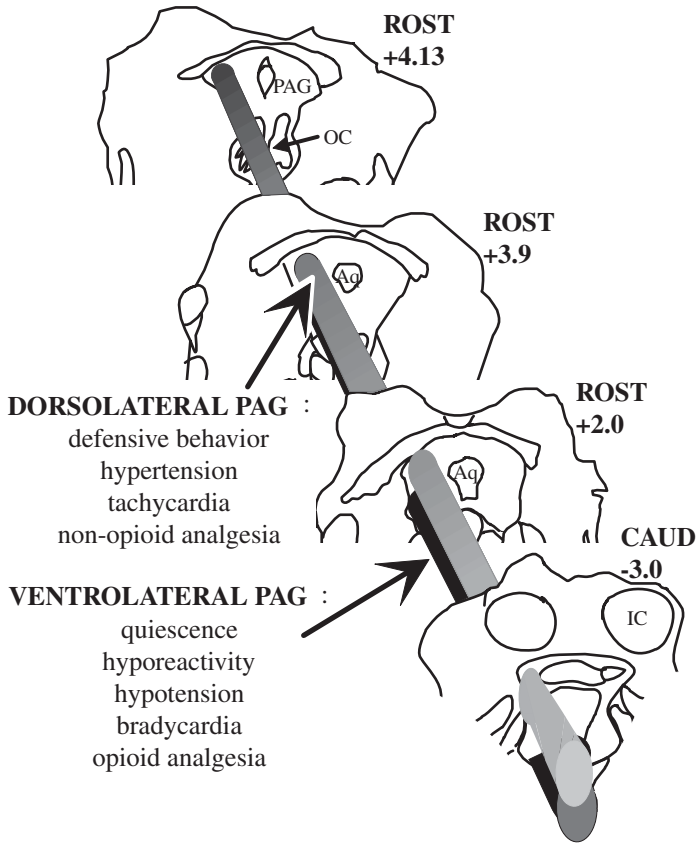


Figure 8 Schematic diagram of the dorsolateral and ventrolateral longitudinal cell columns of the periaqueductal gray controlling the fight (dorsolateral) or flight (ventrolateral) responses. See Table I.

thalamus may then reciprocally activate the midbrain periaqueductal gray. The periaqueductal gray also receives afferents from the prefrontal and insular cortices as well as the lateral and medial preoptic areas of the hypothalamus. The latter connections are much more important in the generation of lordosis behaviors than pain modulation. In sum, the primarily centrally acting, antinociceptive drugs provide much of their relief by activating opioid cells in the ventrolateral portion of the periaqueductal gray. Whereas there are a number of segmental and supraspinal descending antinociceptive systems, the periaqueductal gray provides the major descending control for both opioid and nonopioid-mediated analgesia.

Forebrain connections from the amygdala to the periaqueductal gray participate in the aversive responses of animals and in the fear and anxiety generated in humans from activation of dorsal regions of the periaqueductal gray. In addition, activation of the dorsolateral periaqueductal gray produces threat, associated with vocalization (''fight''), whereas activation of the caudal ventrolateral periaqueductal gray produces immobility or freezing effects (''flight''). These effects of stimulation (dorsolateral versus ventrolateral) are also associated with changes in blood pressure and level of intrinsic nociception. The primary networks for maintaining the blood pressure and controlling respiration and heart rate are located in the medulla. The periaqueductal gray is part of an indirect path that regulates blood pressure in response to emotional state. The periaqueductal gray is interconnected with the lateral hypothalamic nucleus, periventricular nuclei, medial preoptic nucleus, amygdala, prefrontal cortex, and insular cortex. It also has projections to all of the medullary nuclei that participate in the regulation of blood pressure. Stimulation in the dorsolateral periaqueductal gray increased blood pressure, whereas activation of the ventrolateral periaqueductal gray produced hypotension.

This has led to the concept that a constellation of physiologic responses occurs following activation of these two different regions of the periaqueductal gray (Fig. 8). The fight response following dorsolateral activation is characterized by marked hypertension,

increased blood flow to the face, decreased blood flow to the limbs and viscera, tachycardia, and nonopioid analgesia. Activation of the ventrolateral periaqueductal gray is characterized by quiescence, hyporeactivity, hypotension, bradycardia, and opioid analgesia. The blood flow increases to the limbs but decreases to the viscera and face.

The ventrolateral and dorsolateral regions are probably mutually inhibitory, mediated by the action of intrinsic GABAergic interneurons. Injection of bicuculline (a blocker of GABA-A receptors) during stimulation of the dorsal periaqueductal gray blocked the anxiolytic (antianxiety) effect of injected diazepam. Two additional neurotransmitters participate in the pharmacology of anxiety and fear: 5-hydroxytryptophan (serotonin) and CCK (cholecystokinin). In humans $5HT_{1A}$ (serotonin receptor subtype) agonists have been shown to be anxiolytic and this effect is probably mediated by inhibition of dorsal raphe neurons. Injections of CCK in human volunteers generated panic attacks. CCK also decreased the threshold to painful stimuli, whereas activation of the ventrolateral periaqueductal gray with enkephalin produced inhibition of nucleus raphe magnus projecting neurons.

Careful anatomic studies have further demonstrated that these two regions of the periaqueductal gray are organized in longitudinal rostral–caudal columns of cells with different afferent and efferent connections. The periaqueductal gray lateral, ventrolateral, and dorso*medial* to the aqueduct project to the ventromedial and ventrolateral medulla. The dorsolateral periaqueductal gray, on the other hand, projects to the cuneiform nucleus and the periabducens regions of the rostral dorsomedial pons, but not the medulla. The dorsolateral periaqueductal gray column does not stain for cytochrome oxidase, but it has intense staining for acetylcholine-esterase or NADPH–diaphorase. In addition, the labeling experiments have demonstrated that the lateral and ventrolateral portions of the periaqueductal gray have differential projections to the medulla. The ventrolateral portion of the periaqueductal gray projects to the periambigual region of the lateral medulla, which supplies vagal innervation to the heart. Ascending projections from the spinal cord and other brain stem regions have been shown to have topographic projections to the periaqueductal gray. In particular, precisely somatotopically organized afferents arising from lumbar and cervical enlargements target the dorsolateral column of the periaqueductal gray. In contrast, the ventrolateral column is targeted by lumbar and cervical afferents but without any specific somatotopic arrangement. This correlates well with the observed differential behaviors of activating the dorsolateral (fight) versus ventrolateral (flight) portions of the periaqueductal gray.

Cortical inputs to the periaqueductal gray have differential projections to the dorsolateral and ventrolateral cell columns. In the rat, descending corticofugal projections target restricted parts of the periaqueductal gray and terminate as one or two longitudinal columns along the rostrocaudal axis of the periaqueductal gray. Massive subcortical projections to the periaqueductal gray arise from the medial preoptic region and central nucleus of the amygdala and probably mediate many of the observed behavioral responses to periaqueductal gray stimulation. The medial preoptic region has been strongly implicated in neuroendocrine (gonadal steroid) regulation, sexual behavior, thermal regulation, and sleep. The projections of the medial preoptic region are primarily to the dorsomedial and dorsolateral columns of the periaqueductal gray at rostral levels. Further caudally the projection becomes distinctly bicolumnar with dense labeling in the dorsomedial and lateral cell columns. The central nucleus of the amygdala has been strongly implicated in the mediation of defensive behaviors and antinociception. Its projections at rostral levels to the periaqueductal gray are primarily confined to the medial periaqueductal gray. At the level of the oculomotor nuclei, the projection density increases markedly and forms dense dorsomedial and ventrolateral input columns separated by the dorsolateral periaqueductal gray, which receives much less input from the central nucleus of the amygdala. Caudally the central nucleus of the amygdala projects primarily to the lateral and ventrolateral but to neither the dorsomedial nor the dorsolateral periaqueductal gray column. In conclusion, there is compelling evidence for three manifestations of the columnar organization of the periaqueductal gray: (1) discrete physiologic–behavioral functions (fight or flight responses); (2) differential projections from the dorsolateral and ventrolateral portions of the periaqueductal gray to the medulla; and (3) respect of specific, longitudinal columnar boundaries for the various inputs and outputs of the periaqueductal gray.

B. Red Nucleus

The red nucleus occupies a large portion of the ventral midbrain tegmentum rostrally and contains very large

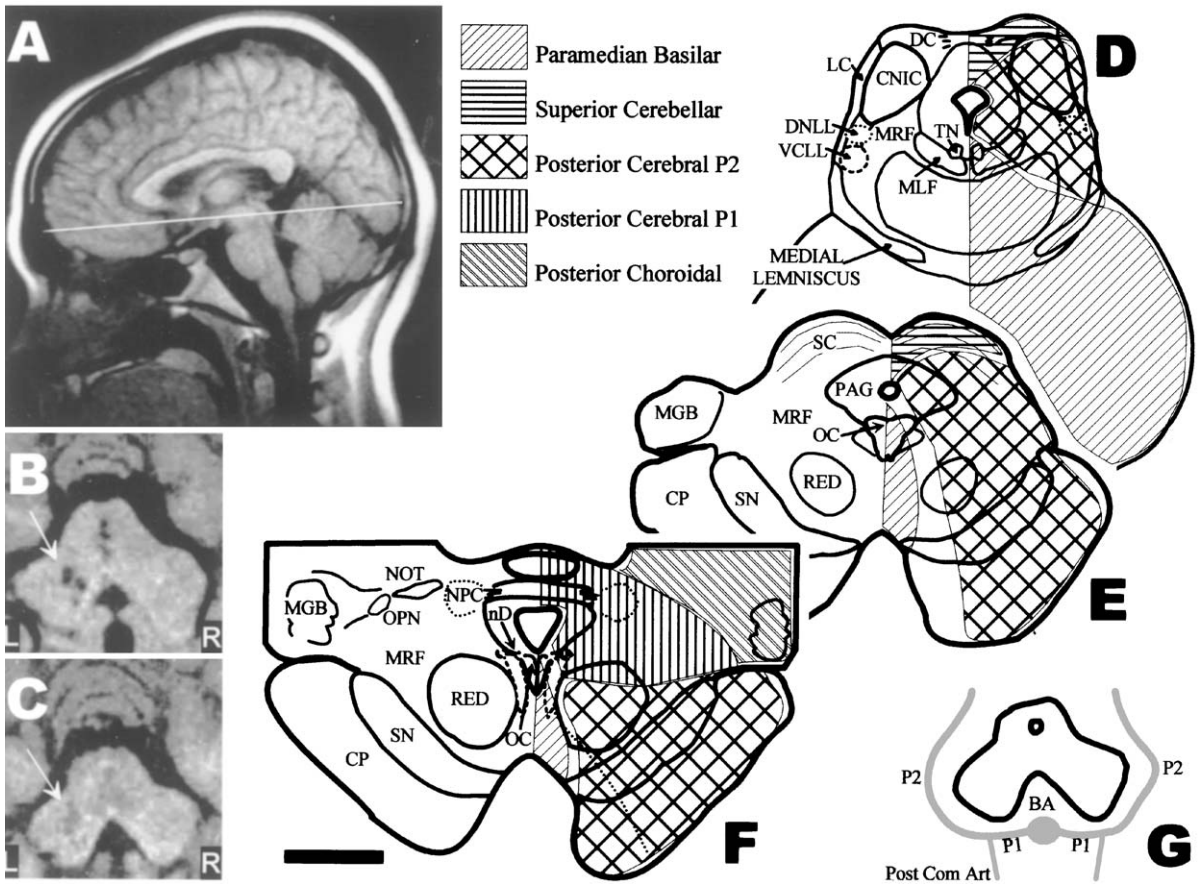


Figure 9 Blood supply of the midbrain. (A) Sagittal section of the human brain showing the horizontal plane of the brain stem sections of (B) and (C). (B) and (C) show serial sections of a patient who suffered a midbrain stroke (arrows) resulting from hemorrhage of a perforating branch of the P2 segment of the posterior cerebral artery on the left (L). Initially the patient experienced bilateral loss of eye movement, bilateral inability to open the eyes (ptosis), and mild weakness of the right upper extremity. The blood has cleared by the time of the MRI study and there is no evidence of infarction of the occipital lobes or other regions supplied by this artery. (D, E, and F) Caudal to rostral outline drawings showing the differential blood supply of the midbrain. (G) Schematic drawing of blood supply to the midbrain. Paramedian branches arise directly from the basilar artery, whereas the remaining blood supply arises from proximal (P1) and distal (P2) branches of the posterior cerebral artery. See Table I.

cells. It is primarily involved in limb control especially during reaching movements. The dentate and interposed nuclei of the cerebellum provide a large number of afferents to the red nucleus via the superior cerebellar peduncle (brachium conjuctivum). The other major input to the red nucleus comes from the motor cortex. Outputs of the red nucleus include two major descending tracts, one to the spinal cord (crossed) and one to the inferior olive (uncrossed).

Lesions within the red nucleus lead to fascicular third nerve palsies, which can involve separate divisions of the oculomotor nerve, as well as contralateral ataxia (Benedikt's syndrome). The ataxia is thought to develop from interruption of both the rubro-olivary and the dentatothalamic fibers, which travel in the adjacent brachium conjunctivum. More ventral strokes that spare the red nucleus may produce a fascicular third nerve palsy and a contralateral hemiparesis (Weber's syndrome).

IV. VASCULAR SUPPLY OF THE MIDBRAIN

The blood supply to the midbrain is via branches of the basilar artery, which bifurcates at the level of the third cranial nerves into a left and right posterior cerebral artery (PCA). The initial portion of the PCA, the P1 segment, is proximal to the posterior communicating

artery, whereas the P2 segment is distal (Fig. 9G). In as many as 10% of patients, a fetal circulation persists and the basilar communicating segment (i.e., the mesencephalic artery) is hypoplastic. In this situation the PCA is derived primarily from the internal carotid artery via a large posterior communicating artery.

In the caudal midbrain (Fig. 9), the blood supply to the ventral portion is from the basilar artery, whereas the dorsolateral midbrain is perfused via the superior cerebellar artery. At more rostral levels, the dorsal midbrain is supplied by the basilar artery and direct branches from the proximal PCAs. The vessels arising from the initial portion of the PCAs can be variable but include paramedian mesencephalic arteries and peduncular penetrating branches, which arise from the PCAs after the posterior communicating arteries join to form the P2 segment of the PCAs (see Figs. 9B and 9C).

Midbrain infarction may accompany thalamic infarction. Branches to the mesencephalon may arise either directly from the basilar artery or as a common trunk with the thalamic paramedian vessels. It is the occlusion of these paramedian thalamic–subthalamic perforating vessels that can often infarct both rostral interstitial nuclei of the medial longitudinal fasciculus and produce a complete paralysis of both upward and downward vertical and torsional eye movements. The thalamogeniculate arteries originate further posterior along the PCAs just at or after the connection with the posterior communicating artery, and they supply the posterior lateral thalamus, including the lateral geniculate, the pulvinar, portions of the hippocampal gyrus, the dentate fascia, and the dorsolateral nuclei of the thalamus. Infarctions of the midbrain are distinguished from the meso-diencephalic junction by third nerve involvement.

See Also the Following Articles

AUDITORY PERCEPTION • BRAIN ANATOMY AND NETWORKS • EYE MOVEMENTS • FOREBRAIN • HINDBRAIN • MOTOR CONTROL • MULTISENSORY INTEGRATION • NERVOUS SYSTEM, ORGANIZATION OF • PAIN • SUPERIOR COLLICULUS

Suggested Reading

Bandler, R., and Shipley, M. T. (1994). Columnar organization in the midbrain periaqueductal gray: Modules for emotional expression? *Trends Neurosci.* **17,** 379–389.

Caplan, L. R. (1996). *Posterior Circulation Disease: Clinical Findings, Diagnosis, and Management*. Blackwell Science, Cambridge, MA.

Carrive, P., Leung, P., Harris, J., and Paxinos, G. (1997). Conditioned fear to context is associated with increased Fos expression in the caudal ventrolateral region of the midbrain periaqueductal gray. *Neuroscience* **78,** 165–177.

Demer, J. L., Oh, S. Y., and Poukens, V. (2000). Evidence for active control of rectus extraocular muscle pulleys. *Invest. Ophthalmol. Vis. Sci.* **41,** 1280–1290.

Faye-Lund, H., and Osen, K. K. (1985). Anatomy of the inferior colliculus in rat. *Anat. Embryol.* **171,** 1–20.

Irvine, D. R. F. (1992). Physiology of the auditory brain stem. In *The Mammalian Auditory Pathway: Neurophysiology* (A. N. Popper, and R. R. Fay, Eds.), Springer-Verlag, New York.

Keller, E. L., McPeek, R. M., and Salz, T. (2000). Evidence against direct connections to PPRF EBNs from SC in the monkey. *J. Neurophysiol.* **84**(3), 1303–1313.

Malmierca, M. S., Blackstad, T. W., Osen, K. K., Karagüelle, T., and Molowny, R. L. (1993). The central nucleus of the inferior colliculus in rat: A Golgi and computer reconstruction study of neuronal and laminar structure. *J. Comp. Neurol.* **333,** 1–27.

Morest, D. K., and Oliver, D. L. (1984). The neuronal architecture of the inferior colliculus in the cat: Defining the functional anatomy of the auditory midbrain. *J. Comp. Neurol.* **222,** 209–236.

Moschovakis, A. K., Karabelas, A. B., and Highstein, S. M. (1988). Structure–function relationships in the primate superior colliculus. I. Morphological classification of efferent neurons. *J. Neurophysiol.* **60,** 232–262.

Oliver, D. L., and Morest, D. K. (1984). The central nucleus of the inferior colliculus in the cat. *J. Comp. Neurol.* **222,** 237–264.

Oliver, D. L. (2000). Ascending efferent projections of the superior olivary complex. *Microsc. Res. Tech.* **51,** 355–363.

Oliver, D. L., and Heurta, M. F. (1992). Inferior and superior colliculi. In *The Mammalian Auditory Pathway: Neuroanatomy* (D. B. Webster, A. N. Popper, and R. R. Fay, Eds.). Springer-Verlag, New York.

Waitzman, D. M., Silakov, V., DePalama-Bowles, S., and Ayers, A. (2000). Effects of reversible inactivation of the primate mesencephalic reticular formation. I. Hypermetric goal-directed saccades. *J. Neurophysiol.* **83,** 2260–2284.

Waitzman, D. M., Silakov, V., DePalama-Bowles, S., and Ayers, A. (2000). Effects of reversible inactivation of the primate mesencephalic reticular formation. II. Hypometric vertical saccades. *J. Neurophysiol.* **83,** 2284–2299.

Migraine

MARGARITA SANCHEZ DEL RIO and STEPHEN SILBERSTEIN
Thomas Jefferson University Hospital

GLOSSARY

familial hemiplegic migraine Hereditary subtype of migraine with aura that presents with motor paresis as part of the aura.

incidence The onset of new cases of a disease in a defined population over a given period of time.

prevalence The proportion of a given population that has a disorder over a defined period of time.

triptans Family of drugs that are widely used as migraine abortives based on their agonism on serotonin receptors of the subtype 1B, 1D, and 1F.

Migraine is a recurrent episodic headache disorder characterized by repeated attacks of pulsating or throbbing, often unilateral, headache pain of moderate to severe intensity that last 4–72 hr. Headache is typically accompanied by nausea, phonophobia, and photophobia. Migraine can be subdivided into migraine with or without aura, depending on whether focal neurological symptoms precede or accompany the headache.

I. EPIDEMIOLOGY AND COMORBIDITY

A. Migraine Prevalence

Migraine is a highly prevalent condition affecting approximately 10% of the population. Migraine prevalence is age, gender, and race dependent. Women are more affected (lifetime prevalence, 12–17%) than men (4–6%). In the American Migraine Study, the 1-year prevalence of migraine increased with age among women and men, reaching the maximum at ages 35–45 and declining thereafter. Migraine prevalence decreases in older women but never decreases to prepubertal or even male prevalence. Migraine prevalence is influenced by race and geographical region. It is highest in North America and Western Europe and more prevalent among Caucasians than African or Asian Americans. The influence of environmental and genetic factors varies. Migraine without aura is influenced by a combination of genetic and environmental factors, whereas migraine with aura has a stronger genetic influence. Behavioral, emotional, and climatologic changes may trigger migraine, modify the vulnerability to migraine, or impact on its prevalence.

Recent evidence suggests that migraine incidence is increasing. Stang and colleagues, in a population-based survey of migraine in Olmsted County, found that from 1979 to 1981 there was a striking increase in the age-adjusted incidence in those under 45 years of age. Migraine incidence increased 34% for women and 100% for men. In this study, the overall age-adjusted incidence was 137 per 100,000 person years for men and 294 per 100,000 person years for women. This was confirmed in a study conducted among schoolchildren in Finland: The prevalence of migraine increased over

an 18-year period from 1.9 to 5.7%. Neither study provided evidence for the cause of this increase.

Migraine incidence and prevalence also vary by age and sex. Migraine without aura reaches its peak at ages 14–17 in females (0.2%) and 10 or 11 in males (0.1%). Migraine with aura has maximum incidence at the age of 12 or 13 in females (0.14%) and at the age of 5 in males (0.6%). This suggests that migraine with aura in men appears very early in life. In contrast, women more often develop migraine (with or without aura) as teenagers, during or following puberty.

B. Comorbidity of Migraine

The term comorbidity refers to the greater than coincidental association of two conditions in the same individual. Migraine is comorbid with many disorders (Table I). This can alert clinicians to identify them. Comorbid illness impacts pharmacologic treatment of migraine headache. One drug may be useful for more than one disease (i.e., valproate and topiramate may be therapeutic for both migraine and epilepsy). On the other hand, some treatments may be contraindicated in certain comorbid illnesses. Beta-blockers should be avoided in patients with migraine and depression. Careful attention to a drug's effect on comorbid conditions optimizes health care use and may improve patient's quality of life.

Individuals who seek medical care are more likely to have two independent diseases than are individuals who do not seek medical care, a phenomenon referred to as Berkson's bias. This can result in associations that are not apparent in population-based studies.

1. Stroke

In a general population study conducted by the National Health and Nutrition Examination Survey I, both migraine and severe nonspecific headaches were

Table I
Comorbidity of Migraine

Comorbid disorder	Reference
Cardiovascular disorders	
Stroke	Merikangas *et al.* (1997),[a] Buring *et al.* (1995)[a]
	Yanagihara *et al.* (1992),[a] Tzourio *et al.* (1993, 1995)[a]
Hypertension	Peroutka *et al.* (1997),[a] Leviton *et al.* (1974)
Heart disease	Couch *et al.* (1989)
Cervical artery dissection	D'Anglejan-Chatillon *et al.* (1989)
Systemic lupus erythematosus	Markus and Hopkinson (1992), Sfikakis *et al.* (1998)
Antiphospholipid antibodies Sd	Silvestrinini *et al.* (1994), Levine and Brey (1996)
Raynaud's phenomenon	O'Keeffe *et al.* (1993), Planchon *et al.* (1994)
Mood and anxiety disorders	
Depression	Merikangas *et al.* (1993),[a] Breslau *et al.* (1994)[a]
Mania	Merikangas *et al.* (1993)[a]
Panic disorder	Stewart *et al.* (1994),[a] Breslau *et al.* (1994)[a]
Generalized anxiety disorder	Merikangas *et al.* (1990),[a] Breslau and Davis (1992)[a]
Phobia	Merikangas *et al.* (1990),[a] Breslau and Davis (1992)[a]
Somatoform disorders	
Neuroticism	Stewart *et al.* (1989),[a] Breslau and Andreski (1995),[a] Breslau *et al.* (1996)
Gastrointestinal disorders	Featherstone (1985)
Gastric ulcer	Featherstone (1985), Chen *et al.* (1987)
Colitis and abdominal pain	Featherstone (1985)
Epileptic syndromes	Andermann *et al.* (1987), Ottman *et al.* (1996)
Allergies and asthma	Chen *et al.* (1987),[a] Monroe *et al.* (1980), Terwindt *et al.* (2000)
Osteoarthritis	Sternfeld *et al.* (1995), Peroutka *et al.* (1997)[a]

[a]Population-based study.

associated with a significantly increased risk for schemic stroke, particularly in young women (rate of stroke was 3.7% in migraineurs versus 2.6% in nonmigraineurs). This association decreases in older patients. The association between migraine and stroke is greater for migraine with aura [odds ratio (OR) 6.2; 95% confidence interval (CI), 2.1–18.0] than migraine without aura (OR-3.0; 95% CI, 1.5–5.8), and it is greater for women than men. In women younger than age 45, migraine with and without aura is associated with a fourfold increase in the risk of stroke. The risk of stroke increases substantially in female migraineurs who are using oral contraceptives (OR-13.9; CI, 5.5–35.1) or who smoke $\geqslant$20 cigarettes a day (OR-10.2; CI, 3.5–29.9).

There appears to be a stronger association between migraine and posterior circulation stroke, a vascular territory that is otherwise infrequently affected (4%) in the general population younger than age 45. Migraine may cause stroke. Bousslavsky *et al.* reported that among patients with stroke, if the stroke occurred during a migraine attack only 9% had arterial lesion, but if the stroke was remote from a migraine attack 91% had an arterial lesion. Mitochondrial DNA mutations have been implicated as one cause of migraine-related strokes in young adults. Majamaa and colleagues found the mitochondrial encephalomyopathy, lactic acidosis, and stroke-like syndrome (MELAS) mutation in 6% of patients with juvenile occipital migraine stroke. Eighty-three percent of the migraine stroke patients carried the U mtDNA haplotype in a higher proportion than the nonmigraineur or the migraine with or without aura population, indicating that this haplotype is a risk factor for migraine stroke. Associations between migraine stroke and coagulopathies have been inconsistent. These findings suggest mechanisms other than arterial disease are responsible for migraine-related infarction.

2. Epilepsy

The 1-year prevalence of epilepsy in the population is approximately 0.5%, the prevalence of migraine in persons with epilepsy ranges from 8 to 15%, and the prevalence of epilepsy in migraine patients is approximately 5.9%. Ottman and Lipton explored the comorbidity of migraine and epilepsy based on the assumption that migraine and epilepsy share a common genetic basis. These investigators found that the incidence of migraine is 2.4 times higher in persons with epilepsy than in persons without epilepsy. The risk of migraine was not associated with the age of onset of epilepsy but was elevated in every subgroup of epilepsy defined by seizure type (higher risk for partial versus generalized onset seizures), etiology of epilepsy (higher risk for epilepsy caused by head trauma than idiopathic epilepsy), or family history of epilepsy. The risk of migraine was elevated both before and after seizure onset, indicating that migraine is not solely the cause or consequence of epilepsy. Genetic and environmental risk factors probably play an important role in the development of migraine and epilepsy, both of which share an alteration in neuronal excitability.

3. Mood, Anxiety, and Somatoform Disorders

The cooccurrence of migraine and psychiatric disorders has been studied extensively in several population-based and longitudinal surveys. Migraine is associated with both affective and anxiety disorders. Breslau and colleagues reported on the association of International Headache Society (IHS)-defined migraine with higher lifetime rates of affective disorder, anxiety disorder, illicit drug use disorder, and nicotine dependence. Migraine with aura was associated with an increased lifetime prevalence of both suicidal ideation and suicide attempts, controlling for sex, major depression, and other concurring psychiatric disorders. The relative risk for the first onset of major depression in migraineurs after the onset of migraine versus no prior migraine was 4.1 (95% CI, 2.2–7.4), whereas the relative risk for the first onset of migraine in persons with prior major depression versus no history of major depression was 3.3 (95% CI, 1.6–6.6). These data indicate that the lifetime association between migraine and major depression could result from a bidirectional influence, from migraine to subsequent onset of major depression, and from major depression to first migraine attack. Furthermore, persons with migraine have increased prevalence of bipolar disorder, panic disorder, and one or more anxiety disorders.

The mechanisms of these associations between migraine and affective/anxiety disorders are poorly understood. The simple causal model—major depression in migraineurs represents a psychological reaction to repeated disabling migraine attacks—cannot account for their comorbidity. A shared common underlying etiopathologic mechanism could be a better explanation. Several antidepressant or anxiolytic agents provide effective treatment for migraine, but the mechanism of action of these drugs seems to be independent of their antidepressant action.

4. Other Conditions

Migraine has been associated with many other conditions, including systemic lupus erythematosus, primary antiphospholipid antibody syndrome, Raynaud's phenomenon, osteoarthritis, and gastrointestinal disorders. It has been suggested but not proven that antiphospholipid antibodies may play a critical a role in the comorbidity of migraine and lupus erythematosus; however, recent studies do not support this idea. The prevalence of migraine attacks in patients suffering from lupus does not differ from that the general population. In addition, gastric ulcer disease and colitis have been associated with migraine and chronic recurrent headaches, but again a precise link is lacking. In one study, peptic ulcer occurred more often among patients with recurrent headaches (13%) than in the age- and sex-matched general population (7.5%), whereas in another study gastric ulcers were associated with migraine only among smokers. Migraine is comorbid with allergies and asthma. The risk of bronchial asthma in children whose mothers had a history of migraine was almost twice that children whose mothers were headache free (RR = 1.8; 95% CI, 1.2–2.8).

C. Quality of Life

Migraine has profound effects on daily functioning, both in the work environment and at home, as ascertained by health-related quality-of-life and functional status questionnaires. Functional status questionnaires measure how headache affects physical and emotional functioning, whereas quality-of-life questionnaires evaluate the subjective effects of the condition.

The American Migraine Study estimated that 23 million U.S. residents have severe migraine headaches. The American Migraine Study II raised this estimate to 28 million. Twenty-five percent of migraineurs have more than four severe attacks a month. Patients with migraine with or without aura are more disabled than patients who suffer from tension-type headache. In approximately 70% of adult migraine subjects, interpersonal relationships are impaired. Regular activities are limited during 78% of migraine attacks and 50% of migraineurs cancel normal activities. About 50% of subjects believed that their headaches had an effect on their families. Headache subjects tends to modify their behavior to avoid precipitating attacks. More than 75% avoided certain headache-triggering factors, such as smoke, noise, emotional stress, and certain physical activities, even if job responsibilities were affected.

Despite the high disability headaches impose on individuals, only 66% of migraineurs ever seek medical attention. Most migraineurs rely on over-the-counter medication and are at risk for analgesic abuse and rebound headache. Education of the headache population is needed to encourage them to seek appropriate medical treatment of their attacks.

D. Economic Cost

The estimates of lost work time due to migraine range from 1.570 to 4.271 days per year. The costs from lost productivity are determined not only by missed work days but also by reduced work efficiency. Approximately 80% of employed migraineurs report that their job performance has been negatively impacted by migraine. Estimates of the level of efficiency during the attack vary from 56 to 73% of full capacity. Although the estimation of efficiency is subjective, there is striking consistency in the assessment of this parameter in different studies. Extrapolating the experience of these subjects, the projected amount of lost work resulting from migraine would cost employers nationally between $5.6 and $17.2 billion annually.

Although the cost to society from lost work productivity is substantial, this may be an underestimate of migraine-related economic loss, because these measures do not take into account such things as reduced productivity resulting from sufferers who attempt to work during migraine or the cost of household productivity losses due to migraine. Appropriate and timely migraine treatment is the most cost-effective strategy. Today, there are more pharmacological, educational, and behavioral approaches to headache treatment than ever before. The ultimate success of new headache therapies may not only depend on efficacy but also may be determined by the outcome of cost-effectiveness analyses on these agents.

II. PATHOPHYSIOLOGY

With the development of minimally invasive imaging techniques, we are starting to better understand the processes that develop in the human brain during migraine attacks. The two leading theories of migraine are the vasogenic theory and the neurogenic theory.

The vasogenic theory posits that migraine aura develops as a consequence of focal ischemia secondary to vasospasm. This hypothesis predicts that decrements in blood flow precede the onset of the aura due to constriction of the blood vessels supplying the affected occipital lobe. Reactive vasodilation would explain the genesis of pain through the stimulation of the perivascular sensitive fibers. This mechanism would explain the throbbing quality of pain, its varied location, and the relief of pain with the use of vasoconstrictive agents such as ergotamine. In the 1980 s, ^{133}Xe blood flow techniques were used to investigate the hemodynamic changes during aura-like symptoms. In general, these investigations reported 17–35% reductions in cerebral blood flow in posterior regions of the brain, values arguably below ischemic thresholds. In certain cases, an anterior spread of the hypoperfusion across neurovascular boundaries was observed. It has been speculated that reported values may be underestimated due to the artifact of Compton's scatter and that the decreases in cerebral blood flow would actually be within ischemic range. Recent positron emission tomographic scan studies on the migraine aura that do not suffer from Compton's scatter clearly show spreading hypoperfusion.

In contrast, the neurogenic theory posits that neuronal dysfunction and a phenomenon similar to cortical spreading depression underlie occipital lobe dysfunction. This theory predicts that alterations in blood flow develop as a consequence of neuronal events. Thus, augmented blood flow followed by reduced perfusion during cortical spreading depression reflects a wave of neuronal and glial depolarization followed by long-lasting suppression of neural activity. Recent high-resolution, high-field functional magnetic resonance imaging studies using blood oxygenation level-dependent (Bold) and perfusion-weighted imaging during migraine with aura attacks have revealed signal changes suggestive of cortical spreading depression within the human brain. Time locked to the onset of scintillation was an initial vasodilation within extrastriate cortex that progressed contiguously over occipital cortex at 3.5 ± 1.1 mm per minute, congruent with retinotopy and visual percept. The initial vasodilation was then followed by hypoperfusion. The facts that the BOLD signal changes during aura abort at major sulci, that the light-evoked visual responses were suppressed during migraine with aura attacks and took 15 min to return to 80% of baseline, and that those areas first affected were the first to recover provide strong evidence that an electrophysiological event such as cortical spreading depression generates the migraine aura in human visual cortex.

Meningovascular or brain mechanisms have been proposed to explain pain generation. In the mid-1940s, Ray and Wolff demonstrated that the human brain is an insensate organ. Sensory nerve fibers originating from the ophthalmic division of the trigeminal nerve innervate the meningeal blood vessels and the dura mater. Together, they constitute the trigeminovascular system. These sensory fibers transmit nociceptive information centrally into the trigeminal nucleus caudalis, and peripherally may promote a sterile inflammatory response within dura mater by releasing vasodilating and permeability-promoting peptides from perivascular nerve endings (substance P, calcitonin gene-related peptide, and neurokinin A). Immunoreactive calcitonin gene-related peptide levels are elevated in external jugular venous blood during migraine attacks, and return to normal after administration of sumatriptan and amelioration of the headache. This is consistent with neuropeptide release from activated sensory nerves during the migraine attack and blockade of peptide release by sumatriptan, mediated via 5-HT1B/1D prejunctional receptors on sensory terminals. A preliminary single photon emission computerized tomography study using Tc-99 human serum albumin provided the first direct evidence of the presence of plasma protein extravasation localized to extraparenchymal regions ipsilateral to the side of pain during a spontaneous migraine attack. Neurogenic inflammation can explain the sensitization of sensory nerve fibers to previously innocuous stimuli (e.g., vessel pulsations or venous pressure changes), which manifests itself as increased intracranial mechanosensitivity and hyperalgesia worsening by coughing or sudden head movement.

Trigeminovascular activation does not account for the initiation of a migraine attack. Brain stem or neocortical structures may play an important role in the genesis or the modulation of migraines or both. It is unclear what triggers the cortical events in migraine aura, but there seem to be genetic differences that render migraineurs more hyperexcitable. These may be related to genetic factors (point mutations in genes encoding calcium channels), mitochondrial energy impairment, magnesium deficiency, or environmental factors such as stress and ovarian steroid.

Upper brain stem nuclei may participate in migraine pathogenesis. Stimulation of periaqueductal gray, locus coeruleus, and dorsal raphe nuclei either generates pain resembling migraine or suppresses pain in animals and humans. Noradrenergic and

serotoninergic nuclei participate in stress responses, anxiety, and depression. Migraineurs may exhibit central hypersensitivity to dopaminergic stimulation, which has been linked to behaviors observed during migraine such as yawning, irritability, hyperactivity, gastroparesis, nausea, and vomiting. Molecular genetic studies have provided further evidence for the involvement of the dopaminergic system. Migraine is associated with the dopaminergic hypersensitivity phenotype overexpression of the DRD2 receptor. The possibility of a "migraine generator" in the rostral brain stem was raised by a positron emission tomography blood flow study performed during spontaneous unilateral headache in nine patients without aura. Increased rCBF was found in medial brain stem predominantly contralateral to the headache, which persisted after relief of migraine pain with sumatriptan. Whether these brain stem nuclei serve as migraine generators, participate in modifying the threshold for neuronal activation, or are part of the neuronal system that terminates an attack remains to be clarified.

III. IHS CLASSIFICATION AND DIAGNOSIS OF HEADACHE

The IHS classification committee created operational diagnostic criteria and a system of classification for headache disorders in 1988 (Tables II–IV). The classification is under revision and will include diagnostic criteria for entities such as chronic daily headache. The purpose of the criteria was to standardize nomenclature and avoid ambiguous diagnosis. Headaches are broadly classified into primary and secondary disorders. In secondary headache disorders, there is an identifiable underlying cause, whereas in primary headache disorders no such underlying abnormality has been identified. Tension-type, migraine, and cluster headaches are the most common primary headache disorders.

Table II
IHS Migraine Classification

1. Migraine
1.1. Migraine without aura
1.2. Migraine with aura
1.2.1. Migraine with typical aura
1.2.2. Migraine with prolonged aura
1.2.3. Familial hemiplegic migraine
1.2.4. Basilar migraine
1.2.5. Migraine aura without headache
1.2.6. Migraine with acute onset aura
1.3. Ophthalmoplegic migraine
1.4. Retinal migraine
1.5. Childhood periodic syndromes that may be precursors to or associated with migraine
1.5.1. Benign paroxysmal vertigo of childhood
1.5.2. Alternating hemiplegia of childhood
1.6. Complications of migraine
1.6.1. Status migrainosus
1.6.1. Migrainous infarction
1.7. Migrainous disorder not fulfilling above criteria

Table III
Migraine without Aura Diagnostic Criteria

A. At least five attacks fulfilling B–D
B. Headache attacks lasting 4–72 hours
C. Headache has at least two of the following characteristics
1. Unilateral location
2. Pulsating quality
3. Moderate or severe intensity
4. Aggravation by walking stairs or similar routine physical activity
D. During headache at least one of the following
1. Nausea and/or vomiting
2. Photophobia and phonophobia
E. At least one of the following
1. History, physical, and neurological examinations do not suggest one of the disorders listed in IHS groups 5–11
2. History and/or physical and/or neurological examinations do not suggest such disorder, but it is ruled out by appropriate investigations
3. Such a disorder is present, but migraine attacks do not occur for the first time in close temporal relation to the disorder

The migraine attack can be divided into four phases: prodrome, aura, headache phase, and resolution (Fig. 1). However, not all phases can be recognized in all individuals.

A. Migraine without Aura

The headache is characterized by episodes of head pain lasting 4–72 hr and having at least two of the following characteristics: pulsatile quality, moderate to severe intensity, unilateral location, and worsening with physical activity. To fulfill the IHS criteria for migraine, headache must also have occurred on at least five occasions and have been accompanied

Table IV
Migraine with Aura Diagnostic Criteria

A. At least two attacks fulfilling B
B. At least three of the following characteristics
1. One or more fully reversible aura symptoms indicating focal cerebral cortical and/or brain stem dysfunction
2. At least one aura symptom develops gradually over more than 4 min or, two or more symptoms occur in succession
3. No single aura symptoms lasts more than 60 min. If more than one aura symptom is present, accepted duration is proportionally increased
4. Headache follows aura with a free interval of less than 60 min
C. History, physical examination, and, where appropriate, diagnostic tests exclude a secondary cause

by nausea or by photophobia and phonophobia (Table II).

B. Migraine with Aura

The migraine aura is a recurrent neurologic symptom that develops gradually (in more than 4 min) and persists for less than 1 hr. Headache, nausea, and/or photophobia usually follow within 60 min after resolution of the aura but may not necessarily develop (acephalgic migraine). Visual aura is most frequently reported (99%), followed by sensory (31%), aphasic (18%), and motor aura (6%). The stereotypical visual aura is a serrated arc of scintillating, shining, crenelated shapes that begins near the point of fixation and slowly expands, leaving a partial homonymous visual field disturbance. Symptoms referable to more than one brain region (e.g., visual and somatosensory symptoms) usually develop in sequence, and when they do occur they can prolong the total duration of the aura to more than 1 hr. Aura symptoms that persist for more than 1 hr but less than 1 week are called prolonged aura (if neuroimaging studies are normal). Symptoms that persist for longer than 7 days or are associated with ischemic infarction on neuroimaging are classified as migrainous infarction.

Table V
Diagnostic Criteria for Chronic Migraine

A. Daily or almost daily (>15 days/month) headache for >1 month
B. Average headache duration of >4 hr/day if untreated
C. At least one of the following
History of episodic migraine meeting any IHS criteria
History of increasing headache frequency with decreasing severity of migrainous features over at least 3 months
Headache at some time meets IHS criteria for migraine 1.1–1.6 other than duration
D. Does not meet criteria for new daily persistent headache or hemicrania continua
E. At least one of the following
There is no suggestion of a secondary disorder as the cause
Such a disorder is suggested, but it is ruled out by appropriate investigations
Such a disorder is present, but first migraine attacks do not occur in close temporal relation to the disorder

C. Basilar Migraine (Bickerstaff's Migraine)

Basilar migraine is characterized by aura symptoms characteristic of a brain stem disturbance (e.g., diplopia, tinnitus, ataxia, dysarthria, bilateral paresthesias or weakness, cranial nerve signs, and occasionally decreased consciousness) or from both occipital lobes. The aura often lasts less than 1 hr and is followed by headache. This syndrome affects children, adolescents, and young women, although it can occur in all age groups and in both sexes.

D. Familial Hemiplegic Migraine

Familial hemiplegic migraine is a rare autosomal-dominant migraine syndrome in which patients experience recurrent visual and somatosensory auras and prolonged hemiparesis in the context of migraine attacks. It may be associated with dysphasia, drowsiness, confusion, coma, and, in some, cerebellar symptoms, tremor and epilepsy. To establish a diagnosis, the patient must have at least one first-degree relative with identical attacks. This

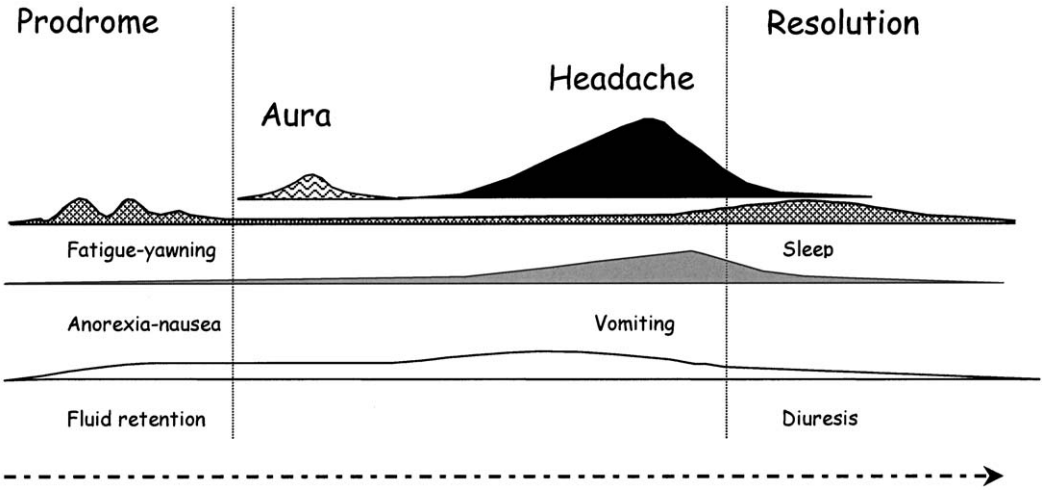


Figure 1 Phases during a migraine attack. The IHS Classification Committee states that premonitory symptoms may precede a migraine attack by hours to up to 24–48 hr. Typical symptoms consist of physical or mental hyper- or hypoactivity, depression, craving for special foods, repetitive yawning, increased bladder and bowel activity, and increased thirst. There is a growing body of biological, pharmacological, and genetic data that support a role for dopamine in the pathophysiology of the symptoms observed during the prodrome.

rare subtype has been linked to point mutations in chromosomes 19 and 1.

E. Chronic Migraine

Under the term chronic migraine, or transformed migraine, we propose to include all cases of chronic daily headache with features of a both migraine and tension-type headache that do not meet criteria for new daily persistent headache or hemicrania continua. Table V proposes criteria for chronic migraine. The typical patient is a woman with a past history of episodic migraine who develops a daily or almost daily headache that is mild to moderate in severity, with superimposed typical migraine attacks. The associated symptoms, such as phonophobia, photophobia, nausea, and vomiting, often become less severe and frequent.

Many patients with chronic migraine overuse analgesics, triptans, and ergots, leading to increased frequency of headaches, some of them in the context of "withdrawal" from acute medication—the so-called "rebound headache." Rebound headache leads to the consumption of more analgesics, creating a vicious cycle. Patients benefit from a detoxification treatment, thereby breaking the cycle, as either inpatients or outpatients. Chronic migraine may also develop from episodic mild pain without acute treatment overuse.

IV. GENETICS

Migraine genetics has experienced remarkable advances in the past 5 years. Molecular biology techniques have led to new insights into the pathogenesis of migraine.

A. Migraine with and without Aura

Migraine has a strong genetic component. In a Danish population-based survey of migraine using IHS criteria, the sex- and age-standardized risk of suffering from migraine with aura and migraine without aura among first-degree relatives was 1.9 (95% CI, 1.6–2.2) and 1.4 (95% CI, 1.0–1.8), respectively. This suggests that migraine without aura is caused by a combination of genetic and environmental factors, whereas migraine with aura is more heavily influenced by genetic factors.

Proposed modes of inheritance for migraine include autosomal-recessive inheritance for migraine with aura and sex-linked transmission, multifactorial, or autosomal-recessive inheritance for migraine without aura. Russell's epidemiology study, which included a segregation analysis of migraine with and without aura, found that both entities have multifactorial inheritance. This analysis cannot detect genetic heterogeneity and therefore cannot exclude a

mitochondrial or Mendelian pattern of inheritance. Migraine is very common. No single gene could be the cause of migraine either with or without aura; if it were, it would be more common than any other known disease-causing gene.

Migraine with aura and migraine without aura are associated with the familial hemiplegic migraine locus on chromosome 19p based on sibling pair and parametric linkage analysis of 28 unrelated migrainous families. Genetic linkage is also supported by a twin study using a polygenic, multifactorial model. In men, concordance rates were 22% in monozygotes and 4% in dizygotes. In women, concordance rates were 32% in monozygotes and 19% in dizygotes. This study estimated that 40–50% of a patient's liability to migraine is based on genetic factors; this estimate is much higher than those of previous twin studies.

B. Familial Hemiplegic Migraine

Familial hemiplegic migraine is an autosomal-dominant subtype of migraine with aura with strong penetrance. Approximately 55% of affected families can be linked to chromosome 19, 15% on chromosome 1, and 30% are still to be determined. Joutel *et al.* found that familial hemiplegic migraine was linked to chromosome 19 in two large French pedigrees. The critical area was mapped to a 30-cm region of the short arm of chromosome 19p13.1. Familial hemiplegic migraine is due to missense mutation in a pore-forming human α_{1A} subunit of neuronal P/Q-type Ca^{2+} channels (CACNA1A). More than 15 missense mutations in the CACNA1A have been reported. P/Q calcium channels are coupled to neurotransmitter release and expressed on soma and dendrites throughout the mammalian brain. How these mutations affect the functioning of the calcium channel is not well understood. Recent reports show calcium currents attributed to α_{1A} channel mutations partially diminished, and, in some studies, increased in other mutations. Tottering mice have a similar genetic defect in the α_{1A} calcium channel. These mice have resistance to the induction of cortical spreading depression. Familial hemiplegic migraine has been linked to two other gene loci on chromosome 1q21–23 and chromosome 1q31, respectively. Patients with the chromosome 19 mutation may also have cerebellar signs and essential tremor, and those with chromosome 1 mutations have associated epilepsy and febrile convulsions.

C. Mitochondrial Encephalomyopathy, Lactic Acidosis, and Stroke-like Syndrome

Migraine-like headaches are part of the clinical spectrum of MELAS, a well-characterized genetic disorder. MELAS is an inherited mitochondrial disease, resulting from different point mutations in the mitochondrial DNA. Most common is an A-to-G point mutation in the mitochondrial gene encoding for tRNA[Leu(UUR)] at nucleotide position 3243. Other point mutations occur at codons 3271 and 3252. Several studies analyzing the mitochondrial DNA in peripheral blood from migraine with and without aura patients have consistently failed to show an association between the commonly studied point mutations or deletions in mitochondrial DNA and migraine. In a Japanese study, 26% of migraine patients showed an A-to-G mutation in the coding region for the ND4 subunit of the respiratory complex I. Thus, migraine as a monosymptomatic expression of a defined mitochondrial cytopathy seems rare.

D. Cerebral Autosomal-Dominant Arteriopathy with Subcortical Infarcts and Leukoencephalopathy

Cerebral autosomal-dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL) is due to mutations in the *Notch 3* gene located on chromosome 19. The *Notch 3* gene encodes a transmembrane complex protein that functions as a cell surface receptor; one part of the protein complex resembles epidermal growth factor. The gene defects lead to either a gain or a loss of a cysteine residue in the extracellular N-terminal part of the molecule, probably causing a conformational and functional alteration. Accumulation of basophilic, PAS-positive, and osmiophilic material (electron microscopy) between degenerating smooth muscle cells in dermal arteries is a pathognomonic finding. The disorder can be diagnosed on skin biopsy.

Approximately one-third of CADASIL patients have migraine early in life with prolonged visual, sensory, motor, or aphasic aura. Recurrent strokes occur between 30 and 50 years of age secondary to a generalized arteriopathy. The arteriopathy develops slowly, resulting in destruction of smooth muscle cells and thickening and fibrosis of the walls of small and medium-sized penetrating arteries with consequent narrowing of the lumen. Multiple lacunar infarcts,

mainly in the frontal white matter and basal ganglia, lead to progressive permanent brain damage. Currently, no specific therapy is available.

Whether the appearance of aura symptoms early in the course of CADASIL is related to changes in the *Notch 3* gene or merely reflects the proximity of the *Notch 3* gene to the familial hemiplegic migraine gene defect remains to be clarified.

V. TREATMENT

Diagnosis of migraine will commence treatment. Patients want to know what is wrong with them and the cause of their headache. Treatment varies from patient to patient, based on their "expectations," headache frequency, severity, disability, comorbid disorders, and family and work environment. Recently, a multispecialty consensus has been reached to develop a unified, evidence-based approach to treating migraine by the U.S. Headache Consortium (www.AAN.com).

The folowing are general principles of treatment:

1. Use not only pharmacotherapy but also physical and psychological therapy.
2. Treat early in an attack.
3. Provide sufficient but not excessive symptomatic treatment, with backup alternatives for moderate and severe attacks, and rescue treatment when initial and backup therapy fails in order to stop suffering and avoid emergency department visits.
4. Encourage the avoidance of emergency department treatment except when all means of "at-home" therapy have proven ineffective. Do not deny patients the use of the emergency department treatment for acute, resistant headache. In case of recurring or inappropriate emergency service utilization, consider hospitalization since rebound or confounding factors are often present.
5. Provide firm limits on frequency and usage per week. Generally, the use of acute treatment should not exceed 2 days a week. Educate the patient on rebound headaches. Rebound headache results from frequent use of acute medications and is a pattern of increasing headache frequency often resulting in daily headache.
6. Treat at least two different attacks to assess effectiveness of a given acute medication. Determine that there is no interfering medication, increase the dose, and change formulation/route of administration before switching or adding another class of drug.
7. Acute treatment drug should be changed when the onset of action is too slow, response is incomplete, headache recurs, responsiveness is inconsistent, or these are important side effects or patient dissatisfaction.
8. Use oral, rectal, or nasal spray preparations initially.
9. Use nasal, rectal, or parenteral forms of medication when attacks are accompanied by significant nausea or vomiting or when there is evidence of delayed gastrointestinal (GI) absorption (gastroparesis).
10. Use adjunctive oral metoclopramide to reverse gastroparesis and improve absorption from the GI tract during severe attacks when oral symptomatic drugs are administered.
11. Administer nasal, rectal, or parenteral forms of symptomatic medication (DHE, sumatriptan, and indocin suppositories) for attacks that are not responsive to oral medication.
12. Consider hospitalization prior to the development of complications and/or addiction/dependency syndromes for patients with severe attacks who overuse treatment or whose treatments are clearly ineffective.
13. Encourage patients to participate in their own management.
14. Provide patients with a headache calendar to establish headache frequency, duration, intensity, associated symptoms, and benefits, and side effects of treatment.
15. Develop a program of symptomatic and preventive medication, often in combination, to establish the most beneficial treatment.

A. Nonpharmacological Treatment

Nonpharmacological treatments include behavioral methods and psychological treatment. Behavioral methods include biofeedback, relaxation techniques, and reinforcement of maintaining a regular schedule, exercise, healthy diet, and regular sleep. There are several modalities of biofeedback; one can monitor body temperature on electromyography. The final goal is the same: the extinction of pain behavior. Relaxation techniques include progressive muscle relaxation and imaginary- and suggestion-based relaxation. These treatments have been proven to be useful for children, pregnant women, and patients for whom stress is a major trigger.

Many patients (34–40%) consult alternative practitioners. Neck manipulation, acupuncture, hypnosis, and herbal therapy are sometimes used. There are no well-designed clinical trials to validate their efficacy.

B. Pharmacological Treatment

Depending on the severity and frequency of headache, patients may be amenable to receive acute or preventive treatment or both.

1. Acute Drug Treatment

Acute treatment is given only at the time of the migraine attack in an effort to shorten the attack duration and decrease the severity of the attack and its associated symptoms. In migraine with aura triptan, treatment during the aura phase has not proven to stop the aura progression or the subsequent headache. Therefore, in migraineurs with aura, treatment with triptans should be administered during the pain phase. Try to avoid pharmacological treatment of very mild headaches; behavioral or relaxation techniques are preferred. Stratify patients' treatment; the goal is to use information available at the initial consultation to predict treatment needs. Treatment plans are formulated based on diagnosis, severity of illness, symptom profile, treatment history, and patient's preference such as willingness to take medication and preferred route of administration. According to this, the initial acute treatment for mild to moderate pain with slow escalation consists of non-specific therapy such as NSAIDS and simple and combination analgesics. For moderate to severe pain, less severe pain and fast escalation, or poor response to NSAIDS or analgesics, go directly to specific therapy—triptans or DHE.

2. Preventive Treatment

Preventive treatment is administered on a daily basis to decrease the frequency and lessen the severity of the attacks. Indications for preventive treatment include the presence of disorders such as hemiplegic migraine, recurrent migraine attacks that significantly interfere with daily function despite acute treatment attacks that occur more than twice a week, or acute treatment that is not satisfactory is excessive, or contraindicated. Maximum benefit may take 3 months. If effective, reevaluate after 6 months and consider if medication needs to be continued. Concomitant acute treatment should be considered for acute attacks that escape preventive measures.

VI. DIFFERENTIAL DIAGNOSIS

The primary headache disorders, such as migraine, tension-type headache, and cluster headache, are those in which headache represents the primary symptom of a physiological disorder. These disorders do not have identifiable gross microscopy pathology. The secondary headache conditions are those in which the headache represents a symptom of a pathological organic process. Some conditions, such as severe hypertension, withdrawal syndromes, or cerebral venous thrombosis, may mimic benign disorders such as migraine. The importance of accurate and timely diagnosis cannot be overestimated because a delay in appropriate treatment may in certain settings result in death or permanent neurological injury. Depending on the temporal profile of presentation of headaches they can be divided into sudden, abrupt, onset headaches, which should be assumed to be the result of an acute neurological event; subacute onset headaches, in which the symptoms appear over minutes, hours, or days; and insidious onset headaches that include those illnesses that have a slow onset and a progressive course.

Note that all cases of "new onset" headache must be evaluated carefully. New onset headaches refers to headaches that have never been present before or in which the pattern or accompanying symptoms are new.

VII. CONCLUSION

Migraine is probably the most important primary headache syndrome due to its high prevalence and the disability that it causes. Recent research with both genetic and imaging techniques has led to a better understanding of the pathophysiology of migraine. However, diagnosis remains a clinical procedure and is the first step for treatment. Stratifying the treatment to meet the patient's needs' decreases disability and is a cost-effective method of managing migraine, delivering improved clinical outcomes at a small additional cost.

See Also the Following Articles

ANXIETY • CONVERSION DISORDERS AND SOMATOFORM DISORDERS • EPILEPSY • MILD HEAD INJURY • MODELING BRAIN INJURY/TRAUMA • MOOD DISORDERS • NAUSEA AND VOMITING • NEUROPHARMACOLOGY • PAIN • PAIN AND PSYCHOPATHOLOGY • STROKE

Suggested Reading

Buzzi, M. G., Di Gennaro, G., D'Onofrio, M., Ciccarelli, O., Santorelli, F. M., Fortini, D., Nappi, G., Nicoletti, F., and Casali, C. (2000). mtDNA A3243G MELAS mutation is not associated with multigenerational female migraine. *Neurology* **54,** 1005–1007.

Diener, H. C. (1999). Efficacy and safety of intravenous acetylsalicylic acid lysinate compared to subcutaneous sumatriptan and parenteral placebo in the acute treatment of migraine. A double-blind, double-dummy, randomized, multicenter, parallel group study. The ASASUMAMIG Study Group. *Cephalalgia* **19,** 581–588.

Dooley, M., and Faulds, D. (1999). Rizatriptan: A review of its efficacy in the management of migraine. *Drugs* **58,** 699–723.

Haan, J., Terwindt, G. M., Maassen, J. A., 'tHart, L. M., Frants, R. R., and Ferrari, M. D. (1999). Search for mitochondrial DNA mutations in migraine subgroups. *Cephalalgia* **19,** 20–22.

Landy, S. H., and McGinnis, J. (1999). Divalproex sodium—Review of prophylactic migraine efficacy, safety and dosage, with recommendations. *Tenn. Med.* **92,** 135–136.

Melchart, D., Linde, K., Fischer, P., White, A., Allais, G., Vickers, A., and Berman, B. (1999). Acupuncture for recurrent headaches: A systematic review of randomized controlled trials. *Cephalalgia* **19,** 779–786.

Montagna, P. (2000). Molecular genetics of migraine headaches: A review. *Cephalalgia* **20,** 3–14.

Pappagalo, M., Szabo, Z., Esposito, G., Lokesh, A., and Velez, L. (1999). Imaging neurogenic inflammation in patients with migraine headaches. *Neurology* **52,** A274–A275.

Rasmussen, B. K. (1999). Epidemiology and socio-economic impact of headache. *Cephalalgia* **19**(Suppl. 25), 20–3, 20–23.

Sanchez del Rio, M., and Moskowitz, M. A. (2000). The trigeminal system. In *The Headaches* (J. Olesen, P. Tfelt-Hansen, and K. M. Welch, Eds.), pp. 141–149. Lippincott Williams & Wilkins, Philadelphia.

Sanchez del Rio, M., Bakker, D., Hadjikhani, N., Wu, O., Cutrer, F. M., Sorensen, A. G., Rosen, B. R., and Moskowitz, M. A. (1999a). Neurovascular cortical spreading phenomenon during spontaneous visual aura. *Cephalalgia* **19**(4), 310.

Sanchez del Rio, M., Bakker, D., Wu, O., Agosti, R., Mitsikostas, D. D., Ostergaard, L., Wells, W. A., Rosen, B. R., Sorensen, A. G., Moskowitz, M. A., and Cutrer, F. M. (1999b). Perfusion weighted imaging during migraine: Spontaneous visual aura and headache. *Cephalalgia.* **19,** 1–7.

Silberstein, S. D., and Saper, J. (1993). Migraine: Diagnosis and treatment. In *Wolff's Headache and Other Head Pain*, 6th ed. (D., Dalcessio and S. D., Silberstein, Eds.), pp. 96–170. Oxford University Press, New York.

Silberstein, S., and Merriam, G. (1999). Sex hormones and headache 1999 (menstrual migraine). *Neurology* **53,** S3–S13.

Terwindt, G. M., Ophoff, R. A., Haan, J., Vergouwe, M. N., van E. R., Frants, R. R., and Ferrari, M. D. (1998). Variable clinical expression of mutations in the P/Q-type calcium channel gene in familial hemiplegic migraine. Dutch Migraine Genetics Research Group. *Neurology* **50,** 1105–1110.

Tfelt-Hansen, P., Saxena, P. R., Dahlof, C., Pascual, J., Lainez, M., Henry, P., Diener, H., Schoenen, J., Ferrari, M. D., and Goadsby, P. J. (2000). Ergotamine in the acute treatment of migraine: A review and European consensus. *Brain* **123,** 9–18.

Mild Head Injury

JEFFREY T. BARTH, JASON R. FREEMAN, and DONNA K. BROSHEK
University of Virginia School of Medicine

GLOSSARY

acceleration–deceleration injury Brain injury typically noted at the cellular level (diffuse axonal injury), caused by high-speed deceleration (e.g., automobile accidents).

cerebral autoregulation System that serves to maintain sufficient cerebral flow to balance blood supply and cellular glucose metabolism demands (facilitate homeostasis).

cerebral concussion Mild traumatic brain injury with brief or no loss of consciousness and no permanent neuropathological deficits or changes.

complicated mild head injury Trauma to the head that results in sequelae that meet the criteria for mild head injury but that reveals neuroimaging evidence of a related cerebral lesion; may also refer to mild head injury with persistent postconcussion disorder.

diffuse axonal injury Widespread histological brain trauma characterized by stretching and breaking of neuronal fibers, usually secondary to acceleration–deceleration and rotational forces of the brain inside the skull.

fluid percussion Rodent animal research model in which a saline solution is introduced under pressure to the extracerebral/subcranial space to simulate nonimpact/deceleration brain injury.

Glasgow Coma Scale A brief head injury observational scale that assesses the general level of consciousness and correlates moderately with severity of injury. This measure involves the evaluation of a patient's best oculomotor, verbal, and motor responses, yielding a severity score from 3 to 15 (3–8, severe; 9–12, moderate; and 13–15, mild).

glucose metabolism The use of glucose (sugars) to create energy for neural activation; directly related to cerebral autoregulation.

postconcussion disorder or postconcussion syndrome A constellation of symptoms including one or more of the following: headache, dizziness, nausea, vomiting, memory, learning and problem-solving impairment, attentional difficulties, problems in abstract reasoning, confusion, fatigue, frustration, depression, sleep disturbance, apathy, diplopia, tinnitus, and slowed thinking, which may persist in some mild head injuries for 6 months postinjury or longer. Postconcussion disorder is the term used in the *Diagnostic and Statistical Manual of Mental Disorders* and postconcussion syndrome is the terminology listed in the *International Classification of Diseases–9-CM*.

posttraumatic amnesia Loss of memory and confusion for events following a head injury.

second impact syndrome An often catastrophic neurological trauma resulting from two mild head injuries in close temporal proximity to each other involving a disconnection of the cerebral blood flow autoregulation system, vascular engorgement, and intracranial pressure, resulting in significant morbidity and/or mortality.

shear strain Stretching and breaking of the axonal fibers due to rotational and deceleration forces in mild head injury. Shearing trauma is usually associated with diffuse axonal injury, but it may also be focal in nature.

Mild head injury is defined as blunt trauma or acceleration–deceleration injury to the head resulting in a Glasgow Coma Scale >12, less than 20 min of unconsciousness or some brief alteration in consciousness, less than 48 hr of hospitalization (no serious extracranial injuries), and no neuroimaging evidence of brain lesion(s). Mild head injury can result in neurocognitive deficits and functional impairments, which may be temporary or persisting. There are considerable ambiguities and uncertainties associated with the study of mild head injury, and this article presents scientific facts and discusses some of the many controversies that

contribute to what has been referred to as "the silent epidemic" and those who suffer from it as the "miserable minority."

I. INTRODUCTION

Mild head injury/concussion has become the single most important disorder in neuropsychological assessment practice in the United States in the past two decades, and it will likely remain a primary focus of clinical and research endeavors for years. Its dominance in neuroscientific study is predicated on several factors: prevalence and costs, morbidity, forensic relevance, societal sensitivity, and its controversial underpinnings.

It is difficult to estimate the prevalence of head injury in the United States each year since statistical analyses must rely on information generated through emergency room records, regional and state traumatic brain injury registries, and Centers for Disease Control data collection. Most of these systems are limited to monitoring moderate to severe disorders that require active medical assessment, intervention, hospitalization, and significant rehabilitation, which are often irrelevant in mild head injury cases. Fifty-five to 75% of all emergency room-reported head traumas are considered mild in nature, and the U.S. Census Bureau estimates the incidence of mild head injury to be approximately 1.3 million per year. This figure must be considered conservative since many mild head injuries occur with little or no loss of consciousness, require no medical treatment, have rapid and complete recovery, and therefore do not come to the attention of the medical/health care community. Traumatic brain injury has long been viewed as a major public health concern. Nevertheless, as a National Institutes of Health consensus development panel suggested, the likely society burden is even greater than previously conceptualized since mild head injury is frequently underdiagnosed. Costs associated with this disorder, when considering assessment, treatment, and lost work, are estimated to be in the billions of dollars each year.

Most studies indicate that the vast majority of individuals experiencing mild head injuries suffer little if any lasting neuropsychological deficits (morbidity), but there is a substantial group of patients who demonstrate deficits in neurocognitive functioning at 1 month postinjury and a small minority with longer lasting sequelae. These latter individuals may be diagnosed with a complicated mild head injury or postconcussion disorder/syndrome. These postconcussion disorder symptoms, even though usually temporary, can be devastating to the socioeconomic, emotional, interpersonal, and occupational functioning of these individuals, who have been termed the "miserable minority."

Although the percentage of patients with mild head injury who experience more severe and longer lasting neurocognitive, medical, and emotional deficits appears to be small, the absolute number of patients in this postconcussion disorder group is likely quite large given the high overall incidence of mild head injury. Since these individuals are the exception to the rules of fast recovery and no permanent impairments, the veracity of their complaints and symptoms are often questioned, which argues for forensic considerations and solutions to personal injury compensation issues. In this context, comprehensive neurocognitive assessments and evaluation of effort and symptom validity are of critical importance. The financial implications for patients, expert medical/neuropsychological witnesses, and attorneys are substantial.

In recent years, society has become more aware and sensitive to the negative consequences of mild head injury through media exposure of celebrities, specifically athletes (particularly in professional football and hockey), who have experienced multiple concussions and whose careers have been threatened. Multiple concussive and subconcussive blows have also been linked to early development of degenerative neurologic conditions such as dementia puglisitca (boxer's encephalopathy), which has also afflicted high-profile athletes and brought sports such as boxing under greater scrutiny by safety commissions.

Finally, mild head injury has been the focus of considerable neuropsychological interest and research based on the many controversies that surround it and obfuscate our complete understanding. Controversial issues include the following:

1. Can we achieve a universally accepted definition?
2. Does mild head injury result in an identifiable neurologic injury/process?
3. What is the natural recovery curve for mild head injury and what roles do individual vulnerability and risk factors play in outcome?
4. What are the best methods for assessing mild head injury and effort/symptom validity in the forensic setting?
5. How can we best assess return to play criteria to avoid catastrophic outcome (second impact syndrome) in sports concussion?

II. DEFINITIONS

The term mild head injury has been linked to other descriptors, disorders, and controversial concepts, such as minor head injury, mild traumatic brain injury, complicated mild head injury, cerebral concussion, postconcussion disorder, posttraumatic stress disorder, somatization disorder, compensation neurosis, the silent epidemic, and the miserable minority. In the early 1980s, investigators at the University of Virginia used the following criteria to define their mild head injury research population: head injury with Glasgow Coma Scale greater than 12, loss of consciousness of less than 20 min, and hospitalization less than 48 hr. This research definition was later expanded to include the absence of neuroimaging evidence of cerebral lesion(s). Variations of this definition with minor modifications have been the mainstay for mild head injury and concussion research ever since.

In 1993, controversy arose when a special-interest group of the American Congress of Rehabilitation Medicine (ACRM) published a definition for mild traumatic brain injury. Although it was similar to the previous research definition of mild head injury, it included factors such as alterations in consciousness or mental state at the time of the accident and posttraumatic amnesia less than 24 hr. Of note, neither the ACRM nor the research definition require loss of consciousness or neuroimaging evidence of impairment. Table I presents the clinical criteria for mild traumatic brain injury. The ACRM definition goes on to state that medical, cognitive, and emotional symptoms of this mild brain injury may persist (the development of postconcussion disorder) and may produce functional disability. The symptoms of postconcussive disorder are presented in Table II.

Table I
Mild Traumatic Brain Injury: Clinical Criteria

Encompasses all aspects of mild head injury research definition and
Infers physiologic impairment and possible histologic damage at least at the time of injury
Requires some altered mental or neurologic state
Slightly more liberal research criteria (LOC < 30 min; PTA < 24 hr)
Symptoms of brain injury may or may not persist (these include physical, cognitive, behavioral, emotional, and psychosocial deficits)
Symptoms can result in temporary or permanent disability
Implies persistent postconcussion disorder
Implies possible complicated mild head injury

Table II
Postconcussion Disorder/Syndrome Symptoms

Headache	Depression/anxiety
Dizziness	Problems in abstract reasoning
Memory impairment	Impaired learning process
Impairment of attention	Confusion of mental state
Slowed mental processing/ reaction time	Disturbance in sleep
Mental and physical fatiguability	Nausea
Lowered frustration tolerance/ irritability	Vomiting
Decreased tolerance for stress and medications	Tinnitus
Apathy/poor motivation	Blurred vision

The most striking aspect of this expanded definition is not its inclusion of potential outcome statements, which is an important consideration and should be part of any clinical definition, but the fact that the general criteria for mild head injury are applied to a new term—mild traumatic brain injury. There is an obvious distinction between "head injury" and "brain injury." The research definition for mild head injury implies some temporary alteration in consciousness at the time of the accident, which also suggests the possibility of some brain impairment but not necessarily persistent functional or neurologic deficits. The ACRM definition, though it is somewhat more specific and directly addresses possible outcomes, is entirely consistent with the research definition of mild head injury. In contrast, it espouses the use of the words brain injury as more accurate terminology since an altered state of consciousness, even if momentary, is required. The use of the term mild traumatic brain injury with this definition has had a secondary, and perhaps profound, effect on society and the medical/neurological sciences, the forensic system, and the patient. Just as Jellnick did in the 1960s when he proposed the disease concept of alcoholism, the ACRM has sensitized the health care community and society to the potential "medical" seriousness of mild head injury and its possible, real underlying neuropathology. Its focus on brain injury has thus established and promoted the disease concept of mild head injury. In addition, the term brain injury to the layperson, or perhaps a jury in a personal injury case, can have an emotionally laden meaning. This meaning may include the perception that there is either a permanent lesion or functional disability, when this may or may not be the case with an appropriately

defined mild traumatic brain injury. Although no studies have directly focused on this issue, it stands to reason that patients and families may well have different reactions to being told that they have suffered a mild head injury versus a mild traumatic brain injury. This again speaks to the issue of emotionally laden terms, societal labels, and a sense of permanence and perhaps hopelessness. There is no consensus regarding the appropriate use of these terms; therefore, mild head injury and mild traumatic brain injury are often used interchangeably. Most researchers and clinicians agree, however, that these two definitions address the most important mild head injury classification criteria, even though they do not offer suggested guidelines for assessing the vague lower limits of mild head injury.

Ronald Ruff and Paul Jurica suggested a unified definition of mild traumatic brain injury that incorporates the classic research and ACRM criteria *Diagnostic and Statistical Manual of Mental Disorders (4th ed.)* as well as the suggested criteria for mild traumatic brain injury. Under their definition of postconcussion disorder, three categories of mild traumatic brain injury were created to address issues of loss of consciousness, posttraumatic amnesia, and neurological symptoms. A type I traumatic brain injury refers to individuals experiencing an altered or transient loss of consciousness, 1–60 sec of posttraumatic amnesia, and one or more neurological symptoms. A type II injury requires a definite loss of consciousness of less than 5 min or unknown duration, 1 min to 12 hr of posttraumatic amnesia, and one or more neurological symptoms. A type III injury includes loss of consciousness of 5–30 min, posttraumatic amnesia that persists more than 12 hr, and one or more neurological symptoms. This definition is broad and allows subclassification that highlights the fact that mild traumatic brain injury is multifactorial and reflects a spectrum from very mild and perhaps simple to more severe and complex.

III. HISTORY AND NEUROPSYCHOLOGICAL FINDINGS

As early as 1962, neuroscientists voiced concerns regarding the possibility of permanent effects from mild concussions, and later during that decade at least one investigator found microscopic lesions on autopsies of patients with histories of mild head injuries. These positions were reinforced in the 1970s and early 1980s by research in New Zealand and at the University of Virginia. These investigations documented neurocognitive deficits in attention, memory, new problem solving, mental flexibility, and cognitive processing speed and delayed return to work in previously employed mild head-injured patients 3 months posttrauma. These findings were in stark contrast to the prevailing medical view at the time, which suggested that mild head injury involved no neurological or cognitive morbidity and that anyone who did not make a rapid, complete recovery and return to work was likely suffering from a psychiatric/personality disturbance such as depression or hysteria.

In light of clinical research findings of neurocognitive morbidity in this population, animal studies were initiated to study neuropathological changes in primates subjected to linear acceleration–deceleration mild head injuries, similar in nature to some of the forces in motor vehicle accidents. Results revealed shear-strain trauma characterized by stretching and breaking of axonal fibers due to deceleration forces on the brain, which was particularly evident in the brain stem and the focus of the mechanical stress in these studies. These histological lesions were not apparent on gross nonmicroscopic inspection.

Although these groundbreaking human neurocognitive and animal studies alerted the medical community to the possibility of identifiable microscopic lesions and significant neurocognitive morbidity in some of the mild head injury population, the experimental designs utilized did not account for confounding factors. They also did not address issues related to differences in primate and human anatomy and physiology. In the mid-1980s and early 1990s, studies in Texas, New York, California, and Washington attempted to control some of these confounding factors, such as previous head injury, substance abuse, and litigation, by selecting participants with no history of these aforementioned risk factors. This sample of patients with uncomplicated mild head injury demonstrated neurocognitive deficits 1 month postinjury in comparison to controls, with good recovery for almost all this population after 3 months.

During this same time period, a unique approach to the study of the effects of mild head injury, preinjury neurocognitive baseline status, and extent and speed of recovery was pioneered at the University of Virginia with the development of the Sports as a Laboratory Assessment Model (SLAM). This methodology involved the study of mild head injury through a highly controlled sports "laboratory" that examined acceleration–deceleration injuries. In this investigation, more than 2300 football players at 10 universities were assessed before the football season using brief

neurocognitive tests, with the same test battery administered again postseason. Players who sustained a mild concussion, as well as a matched control student athlete, were assessed again at 24 hr, 5 days, and 10 days postinjury, in addition to postseason assessments. Statistically significant differences in neurocognitive performance were noted between the scores of concussed players and the scores of controls at 24 hr and 5 days postinjury. By 10 days postinjury, however, concussed players had made sufficient recovery to take advantage of the practice effect and closely approximated the performance of their matched control. This important research demonstrated that the recovery curve was swift in young, healthy, motivated athletes, who demonstrated very few confounding psychosocial or medical factors.

The 1990s saw many scientific studies that borrowed from all these research models. Although some results showed persistent neurocognitive and neuroimaging deficits in patients with postconcussion disorder or complicated mild head injury, other findings suggested little difference between control subjects and patients with mild head injuries at 3 months postinjury. The question remains: Does anyone really suffer in mild head injury? Although the answer may seem obvious, most investigators agree that (i) the vast majority of mild head injury patients make rapid and full recoveries; (ii) some mild head injuries can result in significant neurocognitive deficits that may persist; and (iii) outcomes are likely related to pathophysiological issues, individual risk factors or vulnerability, and recovery curves.

IV. SPORTS CONCUSSION

Perhaps the best empirical evaluation of the functional impact of mild head injury, typically referred to in the sports arena as concussion, comes from the application of the aforementioned SLAM methodology. Because the neurocognitive sequelae of concussion are often subtle and therefore difficult to detect, it is advantageous to have baseline assessments of premorbid functioning against which to compare concussed individuals. Sports involving high risk for acceleration–deceleration injury, such as football and soccer, serve as a natural laboratory in which one can obtain baseline measures of neurocognitive abilities of each athlete for later use that may be compared to his or her postconcussion functioning.

Review of the professional and amateur sports literature indicates that the incidence of mild head injury can be as high as 91% but generally ranges between 2 and 20%. Incidence varies from sport to sport, with the following rates: equestrian, 3–91%; boxing, 1–70%; rugby, 2–25%; soccer, 4–22%; and American football, 2–20%. Frequency of sports-related mild head injury extends to children and adolescents, with potential for greater morbidity in this younger population.

American football has received considerable attention due to the number of high-profile athletes sustaining concussions. Although the concussed football players in the University of Virginia study demonstrated little decline from their preseason baseline data, they failed to take advantage of the practice effects 24 hr postinjury. Many injured players achieved the same learning curve as uninjured controls by the fifth day and there were no statistical differences in performance 10 days after injury. Of note, head-injured players reported neurological symptoms, including headaches, dizziness, and memory problems, that also generally resolved by the 10th day. Other research has found that head-injured athletes performed below their baseline level on most neuropsychological tests, with the greatest decline in verbal fluency and information processing speed.

Soccer has gained increased notoriety as a source of concussion with increased popularity in the United States. Although publicity has focused on the negative effects of heading the ball, concussion most frequently results from impact collisions, such as head-to-head or head-to-goalpost contact. Some research has indicated that soccer players reported a higher frequency of head injury than boxers, with the vast majority endorsing persisting subjective deficits in memory and concentration. Other studies of active and retired soccer players revealed mild electroencephalogram abnormalities, age-inappropriate cerebral atrophy on computed tomography (CT) scans, and more than twice the rate of demonstrated neurocognitive deficits compared to age-matched controls. Given that professional soccer players average 5250 headings during their careers, recent studies have focused on the impact of heading a 12- to 14-ounce ball traveling as fast as 120 km/hr. A prominent study of Dutch professional soccer players found neuropsychological deficits in memory, planning, and visuoperceptual tasks relative to a control group of elite athletes from noncontact sports, with more severe deficits associated with increased frequency of both concussions and headings. Decreased neuropsychological functioning, such as impaired attention, concentration, mental flexibility, and general intellectual ability, was also documented

in high school, college, and professional soccer players, and it correlated with both frequency of ball heading and years of play.

SLAM not only provides a vehicle for research but also potentially yields the best protection of athletes. Optimally, a brief, 20- to 30-min preseason neuropsychological assessment provides a measure of each player's baseline abilities, including processing speed, attention/concentration, and memory. Since athletes would then serve as their own controls, neuropsychological assessment can provide a sensitive tool in identifying concussion-related impairments, even in the absence of radiographic or neurologic findings.

The danger in overlooking or minimizing a concussion is that an initial injury creates a vulnerability that can have devastating consequences if even a very mild

Table III
Virginia Concussion Guidelines[a]

Grade 1—very mild	Grade 2—mild	Grade 3—moderate	Grade 4—severe
Severity of concussion (one or more of the following symptoms in bold are essential to determine category)[b]			
"Bell ringing"—stunned, temporarily dazed/clouded	Clouded consciousness for short time	LOC less than 1 min, clouded consciousness for longer time	LOC more than 1 min. clouded consciousness for longer time
No LOC	**No LOC**	**LOC <1 min**	**LOC >1 min**
Confusion <1 min	**Confusion 1–5 min**	**Confusion 5–15 min**	**Confusion >15 min**
No neuro symptoms	**Neuro symptoms <5 min**	**Neuro symptoms 5–15 min**	**Neuro symptoms >15 min**
Acute headache	Acute headache	Persisting headache	Persisting headache
GCS = 15	GCS = 15	GCS = 13 or 14	GCS < 13
		Day of injury management	
Return to play if no symptoms after 15 min with exertion; continue close monitoring	Return to play if no symptoms after 30 min with exertion; continue close monitoring	No return to play; monitor by team physician/ATC; refer to ER/neurosurgeon for consultation if symptoms persist past game/practice	No return to play; refer to ER/neurosurgeon
		Follow-up criteria for return to play/practice	
Monitor for recurrence of symptoms with exertion. Hold from play until symptom free with exertion. Athlete takes CRI[c]—look for green light for possible return to play. Notify team physician/ATC/neuro surgeon/ neuropsychologist if symptoms develop or persist.	Monitor for recurrence of symptoms with exertion. Hold from play until symptom free with exertion. Athlete takes CRI—look for green light for possible return to play. Notify team physician/ATC/neurosurgeon/ neuropsychologist if symptoms develop or persist.	Minimum 1 week out. No contact for minimum 1 week. Must be symptom free with exertion for 1 week before considering return to play. Athlete takes CRI—look for green light for possible return to play. Consult with team physician/ATC/ neuro-surgeon/ neuropsychologist.	Minimum 3 weeks out. No contact or exertion for minimum of 1 week. Must then be symptom free for 2 weeks with exertion (no contact) and be re-evaluated at least once a week by team physician/neurosurgeon. Athlete takes CRI each week—look for green light for return to play. Consult with team physician/ ATC/neurosurgeon/ neuropsychologist.

[a]Repeated concussions with persisting postconcussion symptoms merit evaluation by neurosurgeon/neuropsychologist and consultation with ATC, physician, family, and athlete to consider ending a season. Confusion does not equate with post injury amnesia, which is defined as an inability to lay down new memories after a traumatic event. Therefore, amnesia does not mean the inability to recall the event, which is quite common. This table is intended to serve as a conservative guideline for the protection of athletes. It should enhance, but not replace, the clinical judgment of trained personnel on the field. Modifications to these guidelines may be made with appropriate consultation among the aforementioned professionals and consensus to adjust guidelines for individual circumstances.

[b]The items in bold are the essential features that are used to grade a concussion. The nonbold items are nonessential symptoms and typically accompanying features. Therefore, headache is not used to make the diagnosis in and of itself. The key symptoms are LOC, confusion, and neurological symptoms.

[c]CRI, Concussion Resolution Index (Head Minder, Inc.)—a web-based neurocognitive screening of sustained attention, processing speed, spatial memory, reaction time, and two-step problem solving.

second injury occurs in close temporal proximity. This traumatic sequence of events is called second impact syndrome (SIS). When an athlete returns to play prior to full recovery after a first mild concussion and sustains a second very mild impact to the head, a devastating process, though rare, may develop even though the athlete initially remains alert, successfully finishes the play, and walks off the field. The athlete may soon collapse, become semicomatose, and develop loss of eye movement, rapidly dilating pupils, and respiratory failure. This appears to result from disrupted autoregulation of cerebral blood flow (CBF), leading to vascular engorgement and increased intracranial pressure, followed by compression of the brain stem and possible death. Prevention of SIS is crucial, despite its relatively low prevalence rate, since it has an approximately 50% mortality rate and a morbidity rate near 100%. According to neurosurgeon Robert Cantu and the American Academy of Neurology, athletes must not be allowed to return to play or practice in contact or collision sports while still experiencing any symptoms of concussion and ideally not until at least 1 week after resolution of symptoms.

In order to allow sufficient recovery time following concussion, return–to–play criteria have been developed and implemented to specify the amount of time that a concussed athlete should be withheld from play. The most widely used criteria are those developed by Robert Cantu and the American Academy of Neurology, but a total of 18 such guidelines exist. Although these guidelines are generally formulated through professional consensus, there is little empirical basis for the decisions regarding length of time that head-injured athletes are withheld from play. In an attempt to develop criteria that combine elements of many conservative guidelines, the University of Virginia has created suggested return-to-play criteria and on-field management recommendations for use by certified athletic trainers and other sports medicine professionals on the sideline. These concussion management guidelines are presented in Table III.

The study of risk factors associated with mild traumatic brain injury via the sports arena has also focused on the investigation of the influence of genetics on boxers. Because boxers suffer repeated head injuries of varying severity, which is the primary objective of the sport, cumulative and delayed effects have been documented, including dementia pugilistica and Parkinson's-like syndromes. Additionally, recent investigations have suggested that the different expressions of the apolipoprotein (APOE) allele differentially affect the risk for morbidity in boxers who have sustained repeated concussions. Although the APOE 2 allele may serve as a protective factor, reducing the incidence of poor neurocognitive outcomes following repeated head injury, having the APOE 4 allele appears to increase the risk for developing neuropsychological deficits and degenerative neurologic conditions.

V. PATHOPHYSIOLOGY

Much of our understanding of the pathophysiology of mild traumatic brain injury (TBI) stems from animal research, which has demonstrated that physiological and metabolic disruption occurs with cerebral concussion and the previously mentioned shear strain. Although knowledge of the neurobiology of concussion is in its infancy, animal models have proven useful in delineating trauma-induced ionic flux, metabolic changes, and disruptions to CBF. Head trauma sufficient to cause concussion triggers changes in intracellular concentration of several ions, including increased potassium and calcium and decreased magnesium, also known as ionic flux. Ionic flux is thought to activate glycolysis (glucose metabolism), which reflects increased energy demands for cell membrane pumps to restore cellular ionic homeostasis. This increased glycolysis has been observed in studies of fluid percussed animals within minutes of injury, with disruption of several metabolic pathways lasting up to 10 days in mature rodent brains. Regions of the brain most prominently affected include the region of the cerebral cortex that is ipsilateral to the injury site and the hippocampus. The pathophysiology and metabolic changes associated with mild TBI are outlined in Table IV.

Ionic flux and metabolic disruption can be conceptualized as an increased "demand" for energy to restore the homeostatic functions of neuronal cells, which places these cells in a vulnerable state. Demand, however, is only part of the issue. Across animal studies of experimental TBI, CBF is reduced by as much as 50% of normal. Therefore, supply of glucose and other cellular energy nutrients required for the restoration of homeostasis can be drastically compromised even in mild head injury. This imbalance of supply and demand is referred to as "uncoupling" and reflects a disruption of the autoregulation of CBF. During changes in cerebral perfusion pressure, cerebral autoregulation serves to maintain constant CBF. Autoregulation in the normal (uninjured) brain is

Table IV
Pathophysiology of Mild Traumatic Brain Injury

Concussion-induced ionic flux		
Potassium (K^+)	**Calcium (Ca^{2+})**	**Magnesium (Mg)**
Massive increase in extracellular K^+ concentration	Massive increase in intracellular Ca^{2+} concentration	Marked decrease in intracellular Mg
Caused by sudden intense neuronal discharges	Ca^{2+} plays role in secondary cell death, which alters metabolism	Mg is important to glycolysis and oxidative phosphorylation
Extracellular K^+ surpasses physiological ceiling causing rapid release of neurotransmitters and additional K^+ fluxes		Reduced Mg may have implications for cerebral metabolic recovery
Acute metabolic changes following concussion		
Glucose metabolism (GM)	Oxidative metabolism	Lactate accumulation
Increased GM	Oxygen utilization is increased in first few minutes following trauma injury	Increased glycolysis after TBI produces decline in high-energy phosphates, which may stimulate anaerobic glycolysis
Hypothesis is that acute increase in glucose utilization following experimental TBI is due to the massive release of extracellular K^+	Likely due to disruption of mitochondrial functions	Results in accumulation of lactate and development of intracellular acidosis
Chronic metabolic changes following concussion		
Long-term effects on glucose metabolism	Long-term effects on oxidative metabolism	
Beginning 6 hr after diffuse brain injury, many brain regions enter state of hypometabolism lasting up to 10 days	Preliminary studies suggest that following concussion, TBI, or cortical ablation, oxidative metabolism is reduced for at least several days, especially within cerebral cortex and hippocampus	
Neurobehavioral deficits accompany this period of metabolic depression (diaschisis)		
Secondary cell death usually does not occur, but it may produce a level of ionic flux and metabolic derangement		
Although sublethal, this may disrupt cellular processes for several days, rendering them vulnerable to secondary insult		

responsible for maintaining the balance of cellular metabolic demands and supply of these needs via blood flow regulation. Although the mechanisms by which disrupted autoregulation occurs are not fully understood, proposed factors include vasospasm, decreased or depleted nitric oxide or nitric oxide synthetase activity, and/or the massive release of vasoconstrictive agents such as neuropeptide Y and endothelins.

Animal research has consistently demonstrated changes in CBF, metabolic function, and ionic flux immediately following experimental TBI. Although this bench research provides a process for understanding the basic pathophysiology in concussion, this process may not generalize to humans and is insufficient in terms of documenting cognitive changes associated with these injuries. Therefore, the need to move from animal models to a more specific analysis of

effects of TBI in humans is clear. Animal models do provide evidence of "windows" of change and opportunity. These changes, however, are doubleedged, indicating both a vulnerability to secondary insult and a critical period during which potential physiology-specific interventions could be effectively utilized. This will be discussed in the following review of the physiology of concussion in humans.

Impaired cerebral autoregulation following mild TBI has been observed in humans. In addition, humans also demonstrate increased glucose metabolism after TBI coupled with reduced CBF, which may represent the insufficient supply of needed energy to restore ionic homeostasis in humans. Researchers at UCLA have speculated that the duration of reduced CBF after head injury is likely to be the primary predictive factor in outcome. Additionally, the endothelial accumulation of calcium appears to have the detrimental effect of inducing vasoconstriction and limiting CBF and necessary nutrients for restoring homeostasis. Observed alterations in CBF in patients with mild TBI are an exciting new frontier for the study of mild brain injury. Frequently, these patients do not experience loss of consciousness, significant retrograde or posttraumatic amnesia, or evidence of pathology on standard neuroimaging (e.g., magnetic resonance imaging or CT), and these factors have generally been poor predictors of functional outcome. A review of the neurovascular and neurochemical effects of mild head injury in humans is warranted because such an endeavor may yield more sensitive, accurate predictors of outcome.

By examining "minor cerebral contusion" using a noninvasive technique known as Xenon inhalation, investigators have found that mild head-injured patients demonstrate significantly reduced CBF relative to controls within 10 days after insult. A significant number of these patients can demonstrate persistent CBF declines for 2 weeks, with recovery to control CBF levels appearing to take place within 4 weeks. Interestingly, those patients who show lateralized hypointensities on CT demonstrate CBF deficiencies in the ipsilateral cerebral hemisphere. Although this research is still in its infancy, patients' return to normal CBF is correlated with improvements on gross measures of cognitive functioning.

Fluorodeoxyglucose positron emission tomography has been used as a noninvasive measure of neurometabolic function in severely head-injured patients, and investigators have found regional and global hyperglycolysis. Although the pathogenesis of increased glucose metabolism is unclear, it is postulated that contributing factors include high concentrations of extracellular potassium (ionic disequilibrium), fluctuating levels of extracellular excitatory amino acids, and localized seizure activity. Of note is the fact that observed metabolic changes persist up to 2 weeks following trauma; however, using the aforementioned Xenon inhalation technique, CBF levels were not so low as to be ischemic. This human research documents metabolic destabilization (specifically heightened glucose metabolism) observed in animal models of mechanical injury. Investigators hypothesize that since animal models utilize a mechanical injury that induces a concussed state involving brief (seconds) loss of consciousness and slow recovery, this is equivalent to mild/moderate head injury in humans. Therefore, if the animal model were extended to human pathophysiology, then cerebral concussion would result in at least a brief period of hyperglycolysis concurrent with diminished supply of glucose through reduced CBF.

Recent technological advances have permitted more accurate but still noninvasive measurement of CBF in mild head injury. Using transcranial Doppler ultrasonography, investigators have compared CBF and mean arterial blood pressure in mild head-injured patients (Glasgow Coma Scale scores between 13 and 15) within 48 hr of injury with matched controls. Although head-injured patients and healthy controls demonstrate equivalent mean arterial blood pressure at rest, head-injured patients show reduced autoregulation in response to induced rapid and brief changes in arterial blood pressure. Because this reduces cerebral perfusion pressure as well as decreases CBF, it can potentially lead to ischemia. Additionally, sudden increases in mean arterial blood pressure may also contribute to secondary hemorrhage and/or edema in the head-injured patient.

In summary, disrupted autoregulation of the vascular supply creates vulnerability from two related states: reduced supply of nutrients (energy) and increased demand for those nutrients. This leaves the head-injured victim at great risk for life-threatening consequences should a second such injury ensue. Accurate assessment of the length of metabolic disruption in humans is critical because animal models likely underestimate the period of cerebral vulnerability. Unstable autoregulation in even mild TBI also has implications for decisions regarding surgery for peripheral injuries. Because blood pressure changes are common during surgical procedures, the risk for ischemia increases dramatically in patients with disrupted autoregulation and reduced cerebral perfusion pressure.

Although some consider the young brain more resilient than the mature brain, this is the subject of continued debate with regard to TBI, particularly given the occurrence of SIS. There is increasing support for a "critical period" related to cerebral maturation that affects recovery and vulnerability. Animal models yield a controlled physiological environment for studying the neurochemical and neurophysiological effects of shear strain or axonal injuries across the life span. Investigation at UCLA of fluid percussion brain injury in 17-day-old, 28-day-old, and adult rodents demonstrated increased intracranial pressure after injury in all animals, but the mortality rate for the younger rodents was greater than that for the mature ones. In the severe injury category, all immature rodents died compared to 55% of the adults. Younger animals even showed an increased duration of postinjury apnea relative to older animals at the mild injury level despite the lack of histopathological differences. These investigators postulated that physiological differences, including brain water content and skull dimensions, might account for the increased morbidity in younger rodents. Additionally, a decrease in cerebral perfusion pressure, which is essential for meeting the energy demands of the postinjury brain and restoring metabolic homeostasis, may result in vasodilation and hence increased intracranial pressure. In any case, it appears that the animal literature, at this early stage of investigation, supports the view that the immature brain is more vulnerable to injury and requires a longer time frame for physiological recovery than the mature brain.

An additional issue in the area of physiological risk and protective factors is the effect of hormones. Animal research investigating gender as a factor in TBI outcome found that estrogen has differential effects in rats undergoing experimental brain injury. It provided protective effects for males and exacerbated the injury of females. Female rats have also been noted to have higher mortality rates following fluid percussion injury. Under conditions of hyponatremia, depressed oxygen use and CBF were observed in female rodents. These animal studies suggest a pathophysiological basis for the reported poorer outcome for females following TBI.

VI. RECOVERY CURVES AND TREATMENT

The natural recovery curve for moderate to severe head injury is 18–24 months, with further restitution of function occurring at a slow rate for years after. Most investigators agree that the majority of mild head injury patients recover in less than 3 months, and only a small percentage of these patients suffer from persistent postconcussion disorder symptoms. Therefore, what are the risk factors for slow recovery and poor long-term outcome?

The first and most obvious risk factor for poor recovery is greater severity of injury and related neurological/physiological dysfunction. A complicated mild head injury defined by positive neuroimaging findings or a high probability of axonal shear strain injury (based on documented loss of consciousness, posttraumatic amnesia, and/or retrograde amnesia) places a patient in a high-risk category for slow and poor recovery. Other factors, such as pain (e.g., headache and back and neck trauma), depression, stress, sleep disturbance, poor premorbid health, and cognitive abilities, previous head injuries, psychiatric disorders, substance abuse, advanced age, poor social support systems, inadequate information about mild head injury recovery, and pending litigation, can all contribute to a patient's individual vulnerability to and risk for poor outcome. Given the complexity of these factors, and the previously mentioned neurocognitive deficits that may accompany these injuries, a comprehensive and interdisciplinary approach to assessment is clearly warranted in some cases. Neuropsychologists, neurologists, neurosurgeons, neuropsychiatrists, physiatrists, neuroradiologists, speech and language therapists, pain and sleep management specialists, and others may be called on to evaluate the mild head injury patient with incomplete recovery.

The most effective treatment approach to most mild head injuries focuses on enhancing the natural recovery curve. Unlike more severe brain injury, significant medical intervention, cognitive rehabilitation, and physical therapy are rarely indicated. Interventions that focus on the previously mentioned individual risk factors will allow the individual to take advantage of natural recovery. Specific suggestions include increasing rest and reducing stress levels, reducing or eliminating alcohol consumption, treating depression with supportive psychotherapy and medications where warranted, assessing and treating pain and sleep disturbance, and educating the patient, family, and significant others regarding the typical mild head injury symptoms and natural recovery course. Early in the intervention process, it is important for everyone involved to have appropriate reassurances and positive expectations but an appreciation for the possible development of symptoms and an understanding that although recovery is usually swift and complete, in

fact, full recovery can take weeks or months in some cases. Realistic but positive expectations can be critical for good outcome, particularly if they reduce stress and allow sufficient time for healing. Recently, researchers have begun to examine the use of pharmacological agents (e.g., psychostimulants) in the management of mild head injury. Further investigations are necessary to conclusively determine the impact of these agents on the recovery curve.

VII. CONTROVERSIES AND FUTURE DIRECTIONS

For two decades, mild head injury has been the focus of intense neuropsychological study, and significant progress has been made in our understanding of this not so silent epidemic. On the other hand, continuing controversies fire our interest in this growing area of intense medical and forensic concern. Although we have validated animal models that help to elucidate the mechanism and pathophysiology of some mild head injuries, we are still uncertain as to the prevalence and extent of real neurological (histological) impairment, the persistence of functional deficits, and possible disability. Typical recovery curves for the majority of these mild head injury patients are generally well-known; however, the risk factors (individual vulnerability) that contribute to poor recovery have not been subjected to sufficient scrutiny to enable researchers to be certain about their relative or absolute effects on outcome. In the forensic arena, we continue to face questions about the presence of underlying neuropathology, the best way to measure premorbid functioning to accurately assess the extent of neurocognitive compromise, and the relationship of compromised function to the trauma. Furthermore, we must continue to examine the veracity of subjective complaints by considering base rates for symptoms in the normal population and resolving questions regarding test-taking effort, symptom validity and exaggeration, and malingering-like behavior.

New neuroimaging techniques are evolving that enhance our capacity to identify lesions, and interfacing these techniques with physiological assessment (e.g., glucose supply and demand) and neuropsychological evaluation will certainly contribute to a better understanding of the mild head injury phenomenon. Use of formulas to estimate premorbid intellectual functioning and interviews with significant others and employers/teachers can help to establish pretrauma cognitive and functional abilities. Tests of symptom validity and effort have been developed; although controversial, they can provide important data in a comprehensive neuropsychological assessment model. Finally, in the sports arena, although return-to-play guidelines have been developed to protect athletes against possible SIS, these criteria remain controversial since they lack a sufficient empirical basis. Even with limited controlled studies, however, experts in the field have established conservative guidelines, based on medical/neuropsychological experience, for measuring severity of concussion and reduced-risk return to play.

The complexity of mild head injury issues and our lack of clear scientific data and knowledge have created discomfort in the scientific/clinical community. When faced with this complexity, there is a tendency to become reductionistic in order to limit uncertainty and increase comfort. This reductionism may lead to simple, and perhaps inaccurate, understandings and extreme positions (i.e., all mild head-injured patients are either malingerers or neurologically devastated). These simple extremes can breed scientism (extreme scientific skepticism) or charlatanism with regard to views on the previously mentioned issues in mild head injury.

One of the solutions to these problems is to develop a level of comfort with the ambiguity and lack of knowledge that we have concerning mild head injury. Recognition of the complexity of these issues will allow us to embrace this as a challenge to our clinical and scientific skills. Finally, it will be important for us to confront these controversies in our scientific inquiries and to share our concerns about issues with our patients, colleagues, and the forensic system.

See Also the Following Articles

BRAIN DAMAGE, RECOVERY FROM • BRAIN LESIONS • COGNITIVE REHABILITATION • MODELING BRAIN INJURY/TRAUMA

Suggested Reading

Alexander, M. P. (1995). Mild traumatic brain injury: Pathophysiology, natural history, and ethical management. *Neurology* **45,** 1253–1260.

Bailes, J. E., Lovell, M. R., and Maroon, J. C. (Eds.) (1999). *Sports Related Concussion*. Quality Medical, St. Louis, MO.

Cantu, R. C. (1998). Neurologic athletic head and neck injuries. *Clin. Sports Med.* **17**(1), 1–210.

Gasquoine, P. G. (1997). Post concussion symptoms. *Neuropsychol. Rev.* **7**(2), 77–86.

Kelly, J. P. (Ed.) (1998). Sports-related head injuries. *J. Head Injury Rehab.* **13**(2), 1–98.

Loring, D. W. (Ed.) (1999). *INS Dictionary of Neuropsychology*. Oxford Univ. Press, New York.

Lye, T. C., and Shores, E. A. (2000). Traumatic brain injury as a risk factor for Alzheimer's disease: A review. *Neuropsychol. Rev.* **10**(2), 115–129.

Narayan, R. K., Wilberger, J. E., and Povlishock, J. T. (Eds.) (1996). *Neurotrauma*. McGraw-Hill, New York.

NIH Consensus Panel (1999). Rehabilitation of persons with traumatic brain injury. *J. Am. Med. Assoc.* **282,** 974–983.

Prins, M. L., Lee, S. M., Cheng, C. L., Becker, D. P., and Hovda, D. A. (1996). Fluid percussion brain injury in the developing and adult rat: A comparative study of mortality, morphology, intracranial pressure and mean arterial blood pressure. *Dev. Brain Res.* **95,** 272–282.

Raymond, M. J., Bennett, T. L., Hartlage, L. C., and Cullum, C. M. (1999). *Mild Traumatic Brain Injury: A Clinician's Guide*. Pro-ed, Austin, TX.

Rizzo, M., and Tranel, D. (Ed.) (1996). *Head Injury and Post-concussive Syndrome*. Churchhill Livingstone, New York.

Rosenthal, M., and Kreutzer, J. (Eds.) (1999). *Rehabilitation of the Adult and Child with Traumatic Brain Injury*. Davis, Philadelphia.

Ruff, R. M., and Jurica, P. (1999). In search of a unified definition for mild traumatic brain injury. *Brain Injury* **13**(9), 943–952.

Varney, N. R., and Roberts, R. J. (Eds.) (1999). *The Evaluation and Treatment of Mild Traumatic Brain Injury*. Erlbaum, Mahwah, NJ.

Modeling Brain Injury/Trauma

HELMUT L. LAURER, DAVID F. MEANEY, SUSAN S. MARGULIES, and TRACY K. MCINTOSH
University of Pennsylvania

GLOSSARY

traumatic brain injury Any disturbance of the cellular integrity and/or homeostasis of the brain tissue or brain cells due to mechanical forces leading to reversible or irreversible cellular dysfunction or cell death in the brain.

An estimated 2 million cases of traumatic brain injury (TBI) occur in the United States every year, with around 500,000 sufficiently serious to require hospitalization. Additionally, a large number of cases of mild TBI remain unreported or undiagnosed each year. Therefore, TBI, with its estimated incidence of 100 per 100,000 persons, occurs more often than many of the better known diseases affecting the central nervous system tissue (e.g., Parkinson's disease, Alzheimer's disease, multiple sclerosis). Epidemiologically, TBI is most often associated with motor vehicle, bicycle, or pedestrian–vehicle accidents followed by falls and violence- and sports-related incidents, occuring primarily in people 15–24 years of age. Estimates for the lifetime cost of care for a severely injured person range from $600,000 to $1,875,000 and add up to a yearly cost of $9–10 billion for rehabilitation for new cases in the Unites States alone. Although the development of neurochemical, histopathological, and molecular techniques to study human TBI has enabled researchers to begin to elucidate the pathologic sequelae following TBI, no neuroprotective therapy is currently available and TBI remains one of the leading causes of disability and death of young adults in industrialized countries.

I. INTRODUCTION

Human head injury involves different forms of focal and diffuse injury to the brain, and the great majority of patients display more than one abnormality upon neuropathological investigation. Focal injury is characterized pathologically by the presence of contusions and lacerations, often accompanied by hematoma. In contrast, the term diffuse injury is most often used for the finding of diffuse axonal injury (DAI) observed in the direct vicinity and also remote from the injury site. Classification becomes even more complex because TBI is often classified according to different grades of injury severity, and the pathologic sequelae after TBI can be separated into primary and secondary (delayed) injury. The term primary injury encompasses the immediate damage to the central nervous system (CNS) that occurs at the moment of impact. This damage to the brain cells and tissues is nonreversible and, therefore, not curable. In contrast, secondary or delayed injury is initiated at the moment of the traumatic insult and progresses for days or months. This secondary injury to the CNS is a complex and poorly understood network of interacting cellular, structural, functional, and molecular changes, including breakdown of the blood–brain barrier, formation

of edema, impairment of energy metabolism, changes in cerebral perfusion and intracranial pressure, ionic dyshomeostasis, activation and/or release of autodestructive neurochemicals and enzymes, inflammation, and pathologic–protective genomic changes. Alone or in combination these events may lead to delayed cell death, but because many are potentially reversible a chance exists for therapeutic intervention to attenuate cellular damage directed at improving functional recovery during rehabilitation and in the chronic phase of the injury.

Experimental models of TBI have been designed to mimic closely the clinical sequelae of human TBI and play a crucial role in the process of evaluating and understanding the physiological, behavioral, and histopathologic changes associated with TBI. Because human TBI is very much a heterogeneous disease, no single animal model of TBI can mimic the whole spectrum of clinical TBI. Rather, the concurrent use of a number of distinct yet complementary models is necessary to reliably reproduce the whole range of injury severity and characteristic features observed upon clinical and post mortem examination of TBI patients. Experimental models have contributed to our insight into the posttraumatic sequelae and have prompted the development of several novel diagnostic and treatment strategies that are now either part of standard clinical practice or under intense preclinical and clinical investigation.

This article attempts to provide a broad overview of the most widely used and popular experimental models of mechanically induced TBI in whole animals. Because rodents are the species of choice in the vast majority of studies due to their obvious advantages (small size, modest cost, extensive normative data available), this article will mainly focus on results obtained in studies with rodents, except where such data are not available. Moreover, this description will not extensively review the existing literature concerning nonmechanical models to produce brain damage (thermal, chemical, or electrical), *in vitro* models, or inanimate finite element computational model characterizations.

II. IMPORTANT VARIABLES IN MODELING TRAUMATIC BRAIN INJURY

Despite differing objectives, any model designed to reliably produce experimental TBI should fulfill a number of criteria. The injury severity and the exact injury location must be defined, and the injury response must be quantifiable and reproducible not only in the same laboratory but also between different investigators. Additionally, the damage caused from the traumatic event should be part of a continuum, increasing with increasing mechanical forces applied to the head or brain. The most widely used models in the majority of studies employ standardized surgical protocols and techniques, including sham (uninjured) animals with identical surgical treatment to control for systemic variables. This experimental design controls for the possible influence of the operative procedure, anesthesia, changes in body or brain temperature, brain damage due to head restraint or the placement of intracranial probes, etc. on the posttraumatic sequelae. Additionally, the majority of the trauma devices employ computer-based measurements of the applied load, such as pressure gradients, the velocity of the impactor, or the speed of acceleration–deceleration forces to measure variations in the mechanical parameters that define inflicted injury severity. This information is used to make adjustments to the device and allows for the maintenance of a narrow range of inflicted injury severity within a particular study.

A. Injury Severity

The most commonly used clinical classification for human head injury is based on the Glasgow Coma Scale, which allows clinicians to divide admitted patients into three major categories of mild (score higher than 13), moderate (score between 9 and 12), and severe TBI (score below 9). Studies conducted to define and characterize outcomes in each of these three categories that can be used to predict spontaneous outcome and response to therapeutic intervention have proven its reliability and usefulness. To closely mimic the range of TBI severity in the clinical situation, experimental studies have modified the existing injury models to be capable of producing brain trauma over a spectrum of severity. This goal is accomplished by adjusting the main mechanical parameters of the injury device (e.g., height or mass of the free-falling weight, depth of the traumatic impact or impact velocity, height of the pressure impulse by adjusting the pendulum of the fluid-percussion device, or changes in the plane or velocity of the rotational forces). Although a scoring method for injury severity comparable to the Glasgow Coma Scale has been developed for the cat, it has yet to be defined for the

rodent. However, a number of experimental studies performed have exposed a close relationship between injury severity and the animals' posttraumatic responses and rates of recovery. As a result, a classification for the severity of experimental TBI has been developed and established that is similar to the clinical categories of mild, moderate, and severe.

B. Types of Injury

Human head injury is not a single pathophysiological entity, and the majority of patients suffering from TBI display more than one lesion upon careful diagnostic evaluation. Because the clinical situation is seldom as controlled as the experimental setting, different injury models are employed to elucidate the main characteristics of TBI, which include focal and diffuse damage. Focal abnormalities involve contusions and lacerations not always accompanied by skull fracture or hematoma formation. This type of damage occurs in the direct vicinity of the site of mechanical impact to the head and typically involves the underlying cortical and, in the case of injuries of higher severity, subcortical structures. Several experimental models have been established that mimic these aspects of focal TBI over a wide range of injury severity (weight-drop closed head injury, fluid-percussion brain injury, and rigid indentation injury). However, all models are associated with concussive events, and, if the injury severity exceeds a certain threshold, substantial displacement of the brain occurs, which adds a new, more remote component of axonal injury to the predominantly focal damage (Fig. 1). Diffuse injuries may include concussions and diffuse axonal pathology. This type of injury is sometimes more difficult to detect in the clinical setting but appears to occur more commonly than previously believed and is presumably present in the whole range from mild to severe head injury. Diffuse brain injuries are thought to occur primarily from the tissue distortion, or shear, caused by inertial forces that are present at the moment of injury. Experimental models that predominantly mimic this type of damage (e.g., models of inertial acceleration and, to a lesser extent, impact acceleration) lead to substantial diffuse injury in the absence of profound focal damage. These changes are usually observed peripheral to the vicinity of the impact and also remote to the injury site (Fig. 1).

C. Determining Injury-Induced Changes and Outcome

A number of scales have been created to evaluate spontaneous and therapeutically modified

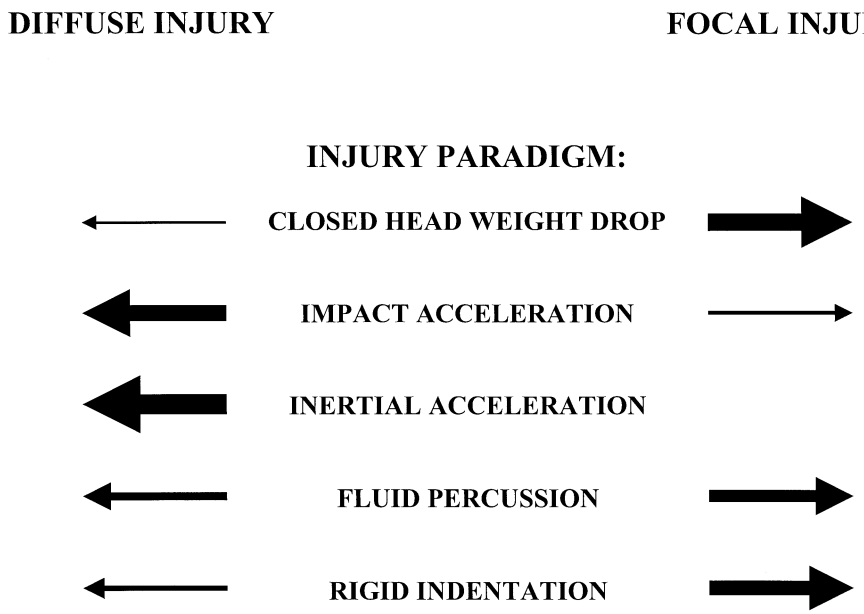


Figure 1 Injury characteristics of different experimental models to produce TBI. Each model produces a distinct pattern of focal and/or diffuse brain injury.

posttraumatic recovery, characterize permanent deficits, or estimate life quality in the chronic phase after clinical TBI. Although a satisfactory system to determine outcome of head-injured patients has yet to be described, the Glasgow Outcome Scale, which divides patients into five subgroups according to their level of recovery or persisting impairment, seems to be the most popular and widely accepted scoring method at present. Thorough evaluation of patients has made it unequivocally clear that behavioral impairments, particularly with respect to neurologic motor function and cognitive deficits, comprise the most persistent deficits after TBI, lasting for months and even years. Additionally, a number of radiological and histological studies have revealed that morphological damage to the brain tissue represents another very consistent feature of human TBI that remains unresolved and persists over time. In contrast, a broad variety of different posttraumatic events occur in the immediate and acute phase after TBI (e.g., changes in electrophysiology, blood–brain barrier dysfunction, edema formation, changes in cerebral perfusion and intracranial pressure, activation of ion channels and ion shift, genomic changes, production of free radicals, inflammation, etc.). Some of these changes are transient, but the duration of others has yet to be determined. To follow and describe injury-induced changes in the experimental setting, research tools have been developed to determine both the reversible and the persistent posttraumatic sequelae after experimental TBI in the laboratory (Table I). Although the great majority of experimental studies have been conducted with posttraumatic survival times of hours or days, a smaller number of studies have successfully identified persistent neurobehavioral impairments and histological changes in the chronic phase after experimental TBI up to 1 year postinjury. Because these persisting alterations (morphologic and behavioral changes)

Table I
Popular Methods to Detect Changes after Experimental TBI

Outcome measurement	Methods
Electrophysiological changes	EEG, somatosensory or brain stem evoked potentials
Breakdown of the blood–brain barrier	Magnetic resonance imaging, immunostaining for plasma constituents (e.g. IgG, albumin), detection of administered tracers (e.g., Evans blue, horseradish peroxidase)
Edema formation	Gravimetric detection of hemispheric swelling and increased water content, determination of specific gravity
Changes in cerebral perfusion	Laser Doppler flow measurements, arterial spin-labeled magnetic resonance imaging, microsphere or hydrogen clearance technique, autoradiography
Increase in intracranial pressure	Intracranial inserted probes
Metabolic alterations	Microdialysis and chromatography of metabolites, staining for respiratory function, determination of local cerebral glucose utilization, nuclear magnetic resonance spectroscopy, assessment of mitochondrial function
Genomic changes	*In situ* hybridization, immunohistochemistry, Western blot analysis, semiquantitative reverse transcriptase PCR
Neurotransmitter release	Microdialysis
Ionic changes	Nuclear magnetic resonance spectroscopy, ion- autoradiography, ion- selective electrode technology, microdialysis
Activation of autodestructive enzymes	Immunohistochemistry for activated enzymes or proteolysis products, electron microscopy for proteolysis products, Western blot analysis
Production of free radicals	Measurement of cyclic voltammetry, salicylate trapping method, chemiluminescence
Inflammation	Microdialysis, immunohistochemistry, *in situ* hybridization, ELISA
Cell damage and cell death	Immunohistochemistry, electron microscopy, measurement of lesion or cavity volume, measurement of cortical thickness, estimates of contusion volume or axonal damage with MRI, regional cell counts
Impairment in neurological motor function	Neurological severity score, Rotarod, rotating pole, beam walk and beam balance, composite neuroscore, spontaneous motor activity, wire grip test
Deficits in cognitive function	Morris water maze, Barnes table, water finding task, win-shift paradigm

seem to be investigated more comprehensively than any of the listed early and transient changes, it appears reasonable to use these persisting characteristics for a brief description of the most popular models to produce experimental TBI.

III. OVERVIEW OF CURRENT MODELS TO PRODUCE TRAUMATIC BRAIN INJURY

A. Models without Trephination of the Skull

1. Weight-Drop Closed Head Injury

In this model, injury is produced using the gravitational forces of a free-falling, guided weight after exposing the skull of the anesthetized animal and attaching it to the impactor or bottom end of the trauma device. The majority of investigations have been performed with the head of the animal placed unrestrained on an adjustable platform; however, in several studies the head has been restrained before the impact is delivered. Due to the height and mass of the weight, injury severity can be varied from mild to severe TBI. Because the head is freely moveable in the majority of studies and no trephination is performed, the exact injury site can be selected on the basis of the special objectives of the particular study. This aspect, together with the relatively short duration involved in the preparation of the animal for the induction of TBI (no trephination of the skull or fixing of protective plates to the skull), makes this model easy and fast to use. However, because the vertex of the skull in rodents is extremely thin, a variable and somewhat uncontrolled number of skull fractures causes variability in the results when animals are subjected to higher magnitudes of injury severity.

With low injury severity, this model is able to produce concussive-like TBI without overt contusion or focal lesion. However, in a rigorous histological evaluation of the brains of mildly injured animals, cortical cell loss directly beneath the impact site and bilateral damage to selectively vulnerable regions remote from the direct injury site have been described. If animals are submitted to an injury of higher magnitude, this model is designed to produce mostly focal damage, thus replicating contusions found in the clinical situation. This characteristic feature is confirmed by the presence of acute hemorrhagic lesions in the acute and hemosiderin-laden macrophages when traumatized brains are evaluated in the more chronic posttraumatic period. Additionally, a necrotic cavity surrounded by a rim of rarefied tissue evolving over time from the hemorrhagic contusion has been reported. Robust impairments in neurological motor function and cognitive deficits, closely correlated with injury severity, have been shown to occur after experimental TBI employing this method. However, no long-term study has evaluated behavioral dysfunction in the more chronic period after TBI using this injury paradigm.

2. Impact Acceleration Model

To overcome the risk of experimentally induced skull fracture, which is a rare finding in human TBI, a model of impact acceleration was developed that uses a stainless steel protection helmet to avoid fracture when animals are subjected to TBI of higher magnitude. This protective shield is glued to the vertex of the skull after opening the scalp and distributes the load widely over the skull, thereby minimizing the likelihood of fracture. When combined with a large blunt weight that causes acceleration of the head and minor contact phenomena, this impact predominantly leads to shear forces. To perform the trauma, the head of the anesthetized animal is placed unrestrained and in prone position on the platform, adjusted to the end of the device, and the impact is delivered via a free-falling weight. The platform has to be covered by a foam of known spring constant to allow defined movement of the head after the impact, which has been shown to profoundly determine the injury-induced changes. Because the mass and height of the weight can be varied, the posttraumatic response has been characterized over a broad range of severity.

This model was characterized to produce mainly diffuse brain injury. Massive diffuse axonal swelling with its severity related to the level of trauma was observed after injury (e.g., in cerebral peduncles, rubrospinal or corticospinal tracts, medulla oblongata, corpus callosum, and internal capsule). Moreover, structural changes in dendrites and mild subarachnoid hemorrhage without focal contusions or lesions have been reported with mild impact, whereas increasing the impact energy leads to more extensive subarachnoid and petechial hemorrhage in the acute posttraumatic period. However, neuronal damage in the supraventricular cortex directly beneath the impact site that is correlated with injury severity and cell death in the hippocampus have been described.

Evaluation of the animal's posttraumatic behavioral function showed cognitive deficits and impairments

in neurologic motor function. Although in one study the more severely injured animals did not show full recovery from behavioral impairments during the investigation period, no long-term studies have been performed to test for persisting impairments in neurologic motor and cognitive function in the more chronic period of months after TBI.

3. Inertial Acceleration Brain Injury

Rotational acceleration of the head due to unrestricted movement causes rotation and deformation of the brain if the loading conditions exceed a certain threshold. Moreover, the rotational acceleration forces necessary to induce damage increase exponentially with decreasing brain size, and no injury device exists that can meet the parameters to reliably produce experimental TBI in rodents. Because a high percentage of patients suffer from TBI not associated with direct contact forces to the head but with rotational forces leading to diffuse brain injury, a model of inertial acceleration injury has been modified from earlier studies conducted in primates to allow for further studies of this type of damage. This model uses the minipig as the species of choice because of its gyrencephalic brain structure and relatively large brain mass compared to its body weight. To produce injury, the anesthetized miniature swine is positioned prone on the injury device, the head is tightly fixed, and inertial loading is produced through a biphasic centroidal rotation for 110° within 20 msec. Additionally, a model using sheep has been developed that is considered to produce mainly diffuse TBI due to mechanical loading to the head. At present, these are the only existing models that are available to produce widely distributed traumatic axonal pathology in the deep white matter at the root of the gyri and the junction of white and gray matter. Additionally, neuronal loss in the cortex and cerebellum next to bilateral damage to hippocampal structures that are known to be selectively vulnerable to TBI is reported. Unfortunately, the lack of availability of tests to assess neurobehavioral impairment in these animals and high costs and technical demands limit the utility of these models at present. To date, only two research centers have performed a small number of histological and radiological investigations elucidating the characteristic posttraumatic sequelae after inertial loading to the brain in the experimental setting.

B. Models with Trephination of the Skull

1. Fluid-Percussion Brain Injury

The fluid-percussion model of brain injury in the rodent is the most commonly used and well-characterized experimental model of TBI. After the skull is exposed and trephination is performed, injury is produced by the impact of a rapid fluid bolus, which strikes the intact dural surface and then moves in the epidural space concentrically from the injection area leading to diffuse loading to the brain. This model subsequently has been used and modified over the past decades to produce injury in a wide range of different animal species. The fluid-percussion technique to produce TBI in the rat was first evaluated with the trephination site placed centrally in the midline. It was subsequently shifted to a lateral location and later modified for use in the mouse. Due to the height of the pendulum, which determines the force of the fluid pressure pulse that is transmitted through a saline-filled reservoir, the injury severity can be varied reproducibly, and studies using this injury model have been conducted over a range of mild to severe experimental TBI.

The fluid-percussion model reproduces contusions as the hallmark of focal (gray matter) damage with accompanying petechial or intraparenchymal and subarachnoid hemorrhages. In fact, central fluid percussion leads to a contusion in the direct vicinity of the site of the maximal fluid impulse, whereas the more pronounced contused area in the lateral fluid-percussion injury tends to be more lateral than the actual trephination and injury site. The extent of this damaged area is associated with injury severity, and a necrotic cavity evolves from the contusion over weeks, which progressively expands due to ongoing cell death up to 1 year postinjury. Injury-induced damage has also been described remote from the injury site. To date, cell damage or cell death has been demonstrated in the thalamus and hippocampus and parallels the cortical cell death up to 1 year postinjury. Although the lateral fluid-percussion model was initially considered to represent a model of unilateral damage, a number of studies have conclusively shown that contralateral damage occurs. Additionally, lateral fluid-percussion injury has been associated with diffuse white matter damage remote from the injury site, reminiscent of human traumatic white matter injury.

Transient impairments in a broad variety of tests for different aspects of neurologic motor function and cognitive deficits have been described. Moreover,

long-term investigations have reported persistent cognitive and motor impairments up to 1 year after severe lateral fluid-percussion brain injury.

2. Rigid Indentation Injury

This method to produce experimental TBI employs a rigid impactor to generate the mechanical energy to the intact dura after trephination of the exposed skull, with the head of the animal usually restrained. The most popular method to produce this type of injury employs pressurized air as the source of the mechanical energy for loading to the brain and is referred to as controlled cortical impact injury (CCI). This method was first described in the ferret and subsequently adapted for use in the rat and mouse. The advantage of closely controlling deformation parameters with pneumatically driven devices (time, velocity, and depth of impact) and the absence of risk for rebound injury make the CCI model superior to devices that are driven by the gravity of a free-falling, guided weight. However, to date, only one study has evaluated for differences in the pathologic sequelae between these two methods to produce experimental TBI.

Besides minor concussive events with low injury severity, this model is able to mimic a whole spectrum of contusions, including intraparenchymal petechial hemorrhages accompanied by epidural and subdural hematomas. Following CCI brain injury, widespread cortical damage and ablation of the gray and, to a lesser extent, the underlying white matter are frequently observed. These changes are dependent on injury severity and, in particular, the depth of indentation and impact velocity. It has been shown that the resulting cell death and hemispheric loss increase up to 1 year postinjury. At moderate and high injury severity with central trephination, axonal injury in the adjacent white matter, corpus callosum, and internal capsule has been described. However, with sensitive immunohistochemical techniques, axonal changes have been observed after mild rigid indentation over the parietal cortex. Consistent neuronal cell loss in selectively vulnerable brain structures, including the hippocampus and dentate gyrus or thalamus ipsilateral to the traumatic impact, has been reported in a number of investigations, and contralateral damage starts to occur following rigid indentation brain injury of higher magnitude. Behavioral evaluations have revealed impairment in various aspects of gross motor function and fine motor coordination following TBI employing this technique. Additionally, deficits in cognitive function (working memory, information acquisition, spatial or place learning) have been reported. The deficits in neurologic motor function seem more transient in nature but have not been comprehensively investigated in the chronic period after TBI, whereas cognitive deficits appear to persist up to 1 year postinjury.

C. Models Complicated by Secondary Events after the Traumatic Brain Injury

1. Hemorrhage–Hypotension–Hypoxia–Ischemia

A high number of patients that suffer from head injury often present with accompanying severe injury to other parts of the body, with profound hemorrhage at the site of injury. Additionally, hypotensive episodes have been found to occur in head-injured patients despite intensive care. Due to the duration and/or extent of these secondary episodes (SEs) accompanied by profoundly impaired autoregulation after TBI, these events may or may not be associated with hypoxia or ischemia to the brain, however. To test for the contribution of these SEs to the posttraumatic cascade after TBI, a number of experiments have included volume- or pressure-controlled hemorrhage, hypoxia, or pharmaceutically induced hypotensive episodes in their study design. Additionally, models of cerebral ischemia or hypoxia have been added to produce secondary ischemic or hypoxic insults to the injured brain directed to evaluate for aggravation of trauma-induced changes after TBI. To date, the models of impact acceleration, fluid-percussion brain injury, or rigid indentation combined with episodes of either hemorrhage, hypotension, ischemia, or hypoxia (or a combination of more than one of these pathologic conditions) have been used to experimentally mimic these frequent complications of clinical TBI.

Due to the broad spectrum of different investigation protocols and numerous possible combinations of the previously mentioned pathologic conditions, comprehensive conclusions are difficult to draw at present. Evidence suggests that animals showed some ability to cope with these posttraumatic SEs without further damage to CNS structures or behavioral impairment if the SE does not occur for a prolonged period. In contrast, a substantial increase in injury-induced pathologic changes has been reported if the duration of the SE or its severity exceeded a certain level or if the animal was subjected to a combination of more than one SE after the traumatic CNS insult.

To date, an increase in cell death in selectively vulnerable brain regions has been reported when trauma and SE were combined. Additionally, tests for neurologic motor function or cognitive ability revealed increased impairment if SE was added to the experimental protocol after TBI. The time frame of exposure to SEs appears to influence the magnitude of additional damage to the CNS. Longer periods of ischemia result in more pronounced damage to the cortex and subcortical hippocampal structures. Moreover, another variable that has to be considered is the time point after the traumatic impact to the brain when the SE occurs. Whereas the subcortical regions seem to be affected in the early period after TBI (immediately until 1 hr), the extent of cortical contusion increased markedly when the ischemia occurred several hours after TBI, although these reports have been contested.

Taken together, the data support a detrimental role of these secondary pathologic events in the sequelae after TBI if they exceed a certain threshold that is dependent on a number of factors. Among others, the magnitude of the single SE, the number, the temporal course and the duration appear to play a crucial role in increasing TBI-induced damage and impairment.

2. Multiple Head Injuries

Repetitive head injuries occur in a wide variety of individuals engaged in contact–collision sports. Due to reports of unfavorable outcomes and epidemiological data suggesting an increased risk for the early onset of neurodegenerative diseases for certain populations of susceptible patients suffering from repetitive head injury, experimental laboratories have begun to develop models that reproduce the effects of repetitive TBI (second impact syndrome). Different techniques described earlier in this article have been used to mimic repetitive head injury, including fluid-percussion injury and closed head injury. The majority of patients suffer from repetitive head injury of mild severity, and, accordingly, the experimental mechanical injury parameters have been designed to produce repetitive mild experimental TBI. However, the interval between the different episodes of TBI and the number of traumatic insults are important variables that have not been thoroughly assessed due to the small number of published studies. Drawing conclusions from these limited results is difficult at present, but it appears that repeated, mild head injury leads to histological damage and behavioral impairments not observed after a single mild TBI. Due to the wide popularity of contact–collision sports, these results warrant further investigation.

IV. MAJOR LIMITATIONS OF THE EXISTING EXPERIMENTAL MODELS OF TRAUMATIC BRAIN INJURY

A. Species Differences

Profound differences in brain geometry, craniospinal angle, gyral complexity, and white to gray matter ratio are well-characterized differences between the human CNS and the brains of laboratory animals. In addition to these anatomical differences, more subtle alterations in the physiologic responses to the traumatic insult, neurotransmitter receptor distribution, or genomic differences must be considered before data obtained in experimental research are translated into clinical concepts. The situation becomes even more complex because profound differences in the behavioral and histopathological responses to TBI within different rodent strains have been described and are additionally influenced by the age of the animal and its laboratory environment.

B. Modeling Injury Severity

In contrast to the clinical situation, no commonly used scoring system for injury severity based on neurological examination has been adopted for the rodent. Therefore, the classification of mild, moderate, and severe experimental TBI continues to be based on measurements of the somewhat arbitrary mechanical injury parameters. Moreover, most injury devices are custom-made and show subtle differences. As a result, scoring systems based on mechanical parameters may be specific only for a particular laboratory. Additionally, different TBI impact sites produce a distinct pattern of injury-induced changes, making experimental findings from different centers difficult to compare at best. Besides these more technical difficulties, the whole heterogeneous spectrum of human TBI is broader than can possibly be modeled in the experimental setting, and both the posttraumatic sequelae after mild TBI without overt morphological damage and the severest cases of TBI with high mortality have not been comprehensively studied.

C. Survival Times and Outcome Measurements

The great majority of studies of experimental TBI have been conducted with short survival times in the range of hours or days. Although the results obtained in these studies may help us to gain insight into the acute posttraumatic sequelae, these histological or behavioral endpoints do not represent a valid assessment of long-term outcomes. Therefore, more studies evaluating injury response and behavioral deficits in the chronic phase (weeks or months after TBI) are warranted. Additionally, the limitations linked with neurobehavioral outcome measurements (high biological variability resulting in a high number of animals required, time- and cost-intensive) have restricted the number of studies evaluating behavioral outcomes after experimental TBI. Because the main effort in patient care is directed toward the attenuation of behavioral dysfunction and chronic disability, more research designed to evaluate neurobehavioral deficits and the therapeutic efficacy of new drugs to attenuate cognitive and neurologic motor functions after TBI is sorely needed.

Although robust impairments in neurologic motor function and cognition have been reported following TBI from moderate or severe magnitude, a great majority of patients suffer from mild clinical TBI. Although the majority of these mildly head-injured patients are not clinically observed or reported, a small percentage register complaints about postconcussive symptoms weeks and months after the traumatic insult. A thorough examination employing sensitive tests directed to detect impairments in higher cognitive or emotional function have indeed shown persistent deficits after mild clinical TBI. However, no such tests are available in the laboratory setting at present. Because the high incidence of mild TBI and the related long-term impairments warrant further investigation, more sensitive tests to elucidate the posttraumatic sequelae after experimental mild TBI must be developed.

V. CONCLUSION

The number of experimental and clinical studies attempting to gain insight into the posttraumatic sequelae following TBI has expanded enormously over the past decade and has generated an abundant amount of preclinical and clinical data suggestive of a number of divergent cascades involved in the delayed damage after TBI. However, no single animal model is entirely successful in reproducing the complete spectrum of pathological changes observed after clinical TBI. Together with the existing *in vivo* models of experimental TBI, *in vitro* models and finite element characterizations have contributed to the current theories of posttraumatic sequelae. Further effort will be necessary to elucidate the acute and chronic changes occurring after TBI in a number of different preclinical models to gain greater insight into the cellular cascades. These studies must be directed at clarifying and validating the present concepts, establishing new therapeutic strategies, and extending the pharmacological armamentarium toward more effective treatment for head-injured patients. New technical breakthroughs, including new methodologies in radiological examination and the use of genomic chip arrays or transgenic animals, seem to be powerful tools in this effort. However, the living animal remains necessary to prove new concepts and make clinical trials successful and safe.

See Also the Following Articles

BRAIN DAMAGE, RECOVERY FROM • BRAIN DISEASE, ORGANIC • BRAIN LESIONS • COGNITIVE REHABILITATION • MILD HEAD INJURY • STROKE

Suggested Reading

Beaumont, A., Marmarou, A., Czigner, A., Yamamoto, M., Demetriadou, K., Shirotani, T., Marmarou, C., and Dunbar, J. (1999). The impact-acceleration model of head injury: Injury severity predicts motor and cognitive performance after trauma. *Neurol. Res.* **21,** 742–754.

Bramlett, H. M., Green, E. J., and Dietrich, W. D. (1999). Exacerbation of cortical and hippocampal CA1 damage due to posttraumatic hypoxia following moderate fluid-percussion brain injury in rats. *J. Neurosurg.* **91,** 653–659.

Carbonell, W. S., Maris, D. O., McCall, T., and Grady, M. S. (1998). Adaptation of the fluid percussion injury model to the mouse. *J. Neurotrauma* **15,** 217–229.

Chen, Y., Constantini, S., Trembovler, V., Weinstock, M., and Shohami, E. (1996). An experimental model of closed head injury in mice: Pathophysiology, histopathology, and cognitive deficits. *J. Neurotrauma* **13,** 557–568.

Crabbe, J. C., Wahlsten, D., and Dudek, B. C. (1999). Genetics of mouse behavior: Interactions with laboratory environment. *Science* **284,** 1670–1672.

Dixon, C. E., Kochanek, P. M., Yan, H. Q., Schiding, J. K., Griffith, R. G., Baum, E., Marion, D. W., and DeKosky, S. T. (1999). One-year study of spatial memory performance, brain morphology and cholinergic markers after moderate controlled cortical impact in rats. *J. Neurotrauma* **16,** 109–122.

Fox, G. B., Levasseur, R. A., and Faden, A. I. (1999). Behavioral responses of C57BL/6, FVB/N, and 129/SvEMS mouse strains to traumatic brain injury: Implications for gene targeting approaches to neurotrauma. *J. Neurotrauma* **16,** 377–389.

Graham, D. I., Raghupathi, R., Saatman, K. E., Meaney, D. F., and McIntosh, T. K. (2000). Tissue tears in the white matter after lateral fluid percussion brain injury in the rat: Relevance to human brain injury. *Acta Neuropathol.* **99,** 117–124.

Lammie, G. A., Piper, I. R., Thomson, D., and Brannan, F. (1999). Neuropathologic characterization of a rodent model of closed head injury—Addition of clinically relevant secondary insults does not significantly potentiate brain damage. *J. Neurotrauma* **16,** 603–615.

Lindner, M. D., Plone, M. A., Cain, C. K., Frydel, B., Francis, J. M., Emerich, D. F., and Sutton, R. L. (1998). Dissociable long-term cognitive deficits after frontal versus sensorimotor cortical contusions. *J. Neurotrauma* **15,** 199–216.

Lowenstein, D. H., Thomas, M. J., Smith, D. H., and McIntosh, T. K. (1992). Selective vulnerability of dentate hilar neurons following traumatic brain injury: A potential mechanistic link between head trauma and disorders of the hippocampus. *J. Neurosci.* **12,** 4846–4853.

McIntosh, T. K., Vink, R., Noble, L., Yamakami, I., Fernyak, S., Soares, H., and Faden, A. L. (1989). Traumatic brain injury in the rat: Characterization of a lateral fluid-percussion model. *Neuroscience* **28,** 233–244.

Pierce, J. E., Smith, D. H., Trojanowski, J. Q., and McIntosh, T. K. (1998). Enduring cognitive, neurobehavioral and histopathological changes persist for up to one year following severe experimental brain injury in rats. *Neuroscience* **87,** 359–369.

Smith, D. H., Chen, X. H., Nonaka, M., Trojanowski, J. Q., Lee, V. M., Saatman, K. E., Leoni, M. J., Xu, B. N., Wolf, J. A., and Meaney, D. F. (1999). Accumulation of amyloid beta and tau and the formation of neurofilament inclusions following diffuse brain injury in the pig. *J. Neuropathol. Exp. Neurol.* **58,** 982–992.

Yang, S. Y., and Cui, J. Z. (1998). Expression of the basic fibroblast growth factor gene in mild and more severe head injury in the rat. *J. Neurosurg.* **89,** 297–302.

Mood Disorders

ERAN CHEMERINSKI and ROBERT G. ROBINSON
University of Iowa College of Medicine

GLOSSARY

limbic system A group of subcortical structures (such as the hypothalamus, the hippocampus, and the amygdala) of the brain that are concerned especially with emotion and motivation.

melancholia A mental condition characterized by extreme depression, bodily complaints, and often hallucinations and delusions.

neurotransmitter A substance (as norepinephrine or acetylcholine) that transmits nerve impulses across a synapse.

proband An individual being studied in a genetic investigation.

The relationship between brain and mood disorders has been investigated based on numerous brain-related functions. This article attempts to summarize the historical development of these areas of research as well as the focus of current investigation.

I. HISTORICAL BACKGROUND

Hippocrates (460–377 BC) is credited with what is probably the earliest statement known to man relating the brain with mood disorders. He stated,

> *and men ought to know that from nothing else but hence [from the brain] comes joys, delights, laughter and sports, sorrows, grieves, despondency and lamentations … and by the same organ we become mad and delirious, and fears and terrors assail us.*

In the 18th and 19th centuries, techniques for neuropathological studies of the brain were developed, and therefore a more precise taxonomic division of mental disorders emerged. In Paris, Philippe Pinel (1745–1826), who became famous for abolishing the restraints of patients at Bicêtre and Salpétrière, divided mental disorders into melancholia and mania with and without delirium. In *Des Maladies Mentales*, Esquirol (1772–1840) created the new term "lypemania" in which the patient had fixed delusions and sadness that overwhelmed and debilitated him or her. Opposite to this was mania, defined by him as a state of generalized delusions and excited emotions.

Contemporary terms for mood disorders, however, are mostly influenced by the work of Emil Kraepelin (1856–1926). In his textbook issued between 1883 and 1914 he applied the method of systematic collection and description of facts for the clinical delineation of mental disease entities. Under the term "depressive states," he included melancholia simplex, melancholia gravis, fantastic melancholia, delirious melancholia, and stupor. Perhaps the greatest contribution of Kraepelin to modern psychiatric classification, however, was the distinction of affective disorders (for him, all forms of recurrent affective psychoses) from the more severe dementia praecox. Whereas dementia praecox exhibited a chronic deteriorating course resulting in poor long-term prognosis, manic–depressive illness had a recurrent course with full remission between psychotic episodes.

II. CLASSIFICATION OF MOOD DISORDERS

A. Background

Although modern psychiatrists seek to classify psychiatric disorders based on biochemical, genetic, and immunologic findings as well as the interaction with the environment, almost all existing psychiatric classifications use the descriptive method in which diseases are named and grouped based on similar patterns of clinical presentations across different individuals. Leonhard (1959) suggested that unipolar (UP) and bipolar (BP) disorders were different entities. After Angst (1966) and Perris (1968) reported systematic family history data supporting Leonhard's theory, these disorders were independently included in modern psychiatric classifications. In 1948, a committee was instituted by the American Psychiatric Association to create a single national system of classification that would solve the chaotic situation existing as new psychiatric diagnoses were being independently devised by different researchers since the 19th century. In 1952, the first system of classification was published, the *Diagnostic and Statistical Manual of Mental Disorders*, known as *DSM-I*. The current fourth edition (*DSM-IV*), published in 1994, uses specific diagnostic criteria (i.e., a list of individual features required for a diagnosis to be made) and a multiaxial approach to classify psychiatric disorders. Axis I refers to clinical disorders. Therefore, the presence of UP or BP disorder would be recorded in this axis. Supplemental information for the clinician is provided by axis II, which consists of personality disorders, as well as axes III and IV, which record associated physical conditions and psychosocial stressors, respectively. Using axis V, which consists of the Global Assessment of Functioning Scale, the clinician can also assess the patients' highest level of social, occupational, and psychological functioning at the time of evaluation.

B. Clinical Subtypes of Mood Disorders

1. Depressive Disorders

The *DSM-IV* classification includes three types of depressive disorders.

a. Major Depressive Disorder This disorder is characterized by a clinical course of one or more major depressive episodes without a history of manic, mixed, or hypomanic episodes. In addition, the episodes are not due to direct physiological effects of a drug or a medical condition and are not present during the course of schizophrenia, schizophreniform disorder, delusional disorder, or psychotic disorder not otherwise specified. There are two types: major depressive disorder, single episode, and major depressive disorder, recurrent.

b. Dysthymic Disorder The essential feature of this disorder is a constant depressed mood for most of the day, and for more days than not, for at least 2 years.

c. Depressive Disorders Not Otherwise Specified This category includes depressive disorders that do not meet the criteria for major depressive disorder, dysthymic disorder, adjustment disorder with depressed mood, or adjustment disorder with mixed anxiety and depressed mood such as minor depression (based on suggested research criteria).

2. Bipolar Disorders

There are four types of bipolar mood disorders.

a. Bipolar I Disorder This disorder is characterized by the occurrence of one or more manic or mixed episodes. Although it is not a requirement, most individuals also have had one or more major depressive episodes. In addition, this disorder is divided into bipolar I disorder, single manic episode, with the presence of only one manic episode and no past major depressive episodes, and other types depending on the nature of the most recent episode—hypomanic, manic, mixed, depressed, or unspecified.

b. Bipolar II Disorder This disorder is characterized by the occurrence of one or more major depressive episodes accompanied by at least one hypomanic episode.

c. Cyclothymic Disorder This disorder is characterized by the occurrence of numerous periods of hypomanic and depressive symptoms not meeting criteria for major depressive episodes for a 2-year period without symptom-free periods of more than 2 months.

d. Bipolar Disorder Not Otherwise Specified This category includes disorders with bipolar features that do not meet criteria for any specific bipolar disorder.

3. Other Mood Disorders

a. Mood Disorder due to a General Medical Condition These disorders are caused by either a depressed mood or apathy or an elevated, expanded, or irritable mood that is the consequence of a direct physiological effect of a general medical condition. These disorders have generally been called secondary mood disorders.

b. Substance-Induced Mood Disorder Depressed or elevated mood develops either after the use of medication etiologically related to the disturbance or within 1 month of substance intoxication or withdrawal.

c. Mood Disorder Not Otherwise Specified This category includes disorders with mood symptoms that do not meet the criteria for any specific mood disorder.

III. EPIDEMIOLOGY OF MOOD DISORDERS

Epidemiology is the study of the distribution of a disease or condition in a population and of the factors that influence this distribution. Due to the high prevalence and morbidity of mood disorders for which effective treatment is available, there has been growing interest in recent years in assessing the epidemiological distribution of these disorders. Results from the National Comorbidity Survey (NCS) show that mood disorders, along with simple phobia and alcohol dependence, are the most frequent psychiatric disorders. Since World War II there have been major changes in the epidemiology of mood disorders, such as an increase in the overall cumulative lifetime rates of depression and the occurrence of these disorders at earlier ages in both men and women.

In the early 1980s, approximately 20,000 community and institutionalized (e.g., in mental hospitals, nursing homes, or prison) subjects were selected according to age group proportion by the 1980 U.S. census of adults ages 18 and older as part of the Epidemiological Catchment Area (ECA) study. The overall rate for mood disorders was 5.1%, with significantly higher rates in women (6.6%) compared to men (3.5%). The highest rates were found in the age group 25–44 years (6.4%), whereas the lowest rates were found in the are group $\geqslant$65 years old. Of the sociodemographic variables, marital status accounted for the only significant difference, with a 4.1% rate of mood disorders in married subjects compared to an 11.1% rate in separated and divorced subjects. A major depressive episode occurred in 2.2% of the population during the previous year, whereas a manic episode (necessary for the bipolar I disorder diagnostic criteria) occurred in the previous year in 0.4% of the population. Dysthymia was identified on a lifetime basis in 3.3%.

The NCS assessed a probability-based sample population of subjects between 15 and 54 years of age in the United States. In order to increase the reporting sensitivity of lifetime psychiatric disorders and substance use, the interviewers encouraged respondents to recall accurately past episodes of psychiatric illnesses using a structured interview that incorporated *DSM-III-R* criteria. This survey, conducted from 1990 to 1992, surprisingly showed much higher prevalence rates than the ECA. The NCS found that the lifetime prevalence of major depressive episode was 17.1% overall (12.7% in males and 21.3% in females). For major depressive disorder, the lifetime prevalence was 14.9% overall (11.0% among males and 18.6% among females), whereas the overall prevalence for dysthymia in this age group was 6.4% (4.8% in males and 8.0% in females). The NCS also found that in those respondents who had a substance use disorder on a lifetime basis, there was a prevalence of 51% of lifetime mental disorder. The variation in the assessment instruments used in the ECA and the more recent NCS could account for the substantial difference observed in the rates of prevalence between the two surveys.

A review by Rouchell, Pounds, and Tierney of recent epidemiological studies showed the prevalence of major depression in specific medical conditions: dementia, 11%; stroke, 27%; Parkinson's disease, 28.6%; epilepsy, 55%; diabetes mellitus, 24%; coronary artery disease, 16–19%; and cancer, was 20–38%.

IV. BRAIN-RELATED MECHANISM OF MOOD DISORDERS

A. Genetic Contributions

1. Background

In the 1850s, Morel initiated the application of genetics to mental diseases. His "psychopathic degeneration" theory expressed that mental disorders have a genetic origin and worsen over generations; therefore,

neurotic symptoms in a patient would ultimately give rise to severe mental illnesses in his or her progeny. In the 1920s, Rudin, Luxenburger, and Schultz from the Munich school of medicine intensively studied the mode of inheritance of schizophrenia and refined the methodology of twin studies. However, they soon found that Mendelian modes of inheritance, which made it easy to calculate the morbid risks for different degrees of relatives of probands, could not be applied to the study of the inheritance of major psychoses.

2. Family Studies

Most studies of bipolar disorder show that this illness tends to be familial, with significantly higher risk in relatives of bipolar probands compared to the general population. Kallmann reported that the lifetime risk for parents of bipolar probands was 23.4% and that for siblings was 22.7%, whereas for second-degree relatives the rates range from 1 to 4%. Winokur found that in sibs of bipolar patients, the lifetime risk for any mood disorder was 35%, whereas in their parents the risk was 34%.

3. Twin Studies

Although monozygotic (MZ) twins are genetically identical, dizygotic (DZ) twins share only half of the genome. The twin method compares concordance rates for a trait between MZ and DZ twins. A significantly higher rate in MZ compared to DZ twins who share a similar environment suggests that genetic influences are present. Most studies using this method have found two to five times higher concordance rates for BP disorder in MZ compared to DZ twins. A Danish twin study found that when the proband had BP disorder, the concordance in MZ twins was 69% compared to 19% in DZ twins (3.5 : 1 ratio), and when the proband had UP disorder the MZ concordance was 54%, whereas the DZ concordance was 24% (2 : 1 ratio).

4. Adoption Studies

This method separates the interacting roles of heredity and environment in the etiology of diseases by comparing the rates in the adoptive parents of a proband versus the rates in his or her biological parents. Mendlewicz reported higher rates of mood disorder in biological (28%) compared to adoptive parents (12%) among parents of adoptees with BP illness. The rate of the disorder in the proband's adoptive parents was similar to that of the adoptive parents of a normal offspring control group, whereas the rate of the disorder in biological parents of adoptees probands was similar to that of parents in a nonadopted bipolar group. Cadoret also found higher rates of UP depression in adopted offspring of biological parents with mood disorders compared to adoptees with biological parents without psychiatric conditions.

5. Linkage Studies

Two traits located proximately on the same chromosome by the phenomenon of linkage have a dependent assortment during the process of meiosis. Linkage studies compare the pattern of inheritance of blocks of genes among relatives with the inheritance pattern of a trait of interest. Although this method was useful in the study of rare Mendelian disorders with known mode of inheritance and clearly specified phenotypes such as Huntington's disease, it has been less successful when applied to disorders that do not exhibit simple Mendelian patterns of inheritance, such as the major psychiatric disorders. Recent studies have demonstrated a linkage of mood disorders to chromosome 18. In a linkage study of 22 bipolar families, Berretini *et al.* found evidence for a bipolar susceptibility gene in the pericentromeric region of chromosome 18. No single gene or region has been shown to be causative of the majority of familial mood disorders.

B. Neuroendocrine Contributions

The observation that mood disorders are sometimes associated with endocrine disorders, such as Cushing's syndrome and hyper- or hypothyroidism, has led to the hypothesis that abnormal secretion of neurotransmitters, themselves implicated in the mechanism of mood disorders, also disrupts the secretion of hypothalamic neuroendocrine cells leading to an association between mood disorders and hormonal dysfunction.

1. Hypothalamic–Pituitary–Adrenal Axis

Cortisol has been recognized as a stress–response hormone because plasma levels are altered by environmental stress. Disturbances in the hypothalamic–pituitary–adrenal (HPA) axis due to mood disorders in humans were reported as early as 1943 by Pincus and Hoagland, who observed an elevation of urinary

metabolites of adrenocortical secretion related to stress among test pilots. In response to anxiety or depression, an increase in norepinephrine (NE) activates the hypothalamus to release corticotropin-releasing factor, which stimulates the secretion of adrenocorticotropic hormone. This hormone stimulates the adrenal cortex, thereby increasing plasma concentrations of cortisol. By a negative feedback mechanism, cortisol lowers the levels of NE synthesis.

The best known challenge test related to mood disorders is the Dexamethasone Suppression Test. A level of $>5\,\mu g/dl$ of cortisol at 8, 16, or 24 hr after dexamethasone administration is abnormal and indicates a lack of suppression capability by the feedback inhibition system of the HPA axis. The highest rates of abnormalities (i.e., 60–90% abnormality) are found in more severe depressive patients, usually with melancholic or delusional symptoms.

2. Hypothalamic–Pituitary–Thyroid Axis

A number of researchers have reported subnormal pituitary thyroid-stimulating hormone (TSH) levels in response to the administration of the hypothalamic thyrotropin-releasing hormone among patients with depressive disorder. Moreover, a correlation has been found between relapse of depression and a persistently low TSH response after clinical recovery.

3. Growth Hormone

The levels of growth hormone (GH) secretion by the pituitary are increased by the hypothalamic growth hormone-releasing hormone as well as by noradrenergic and dopaminergic stimulation, whereas the release of GH is inhibited by somatostatin. Several studies have reported blunted responses of GH to clonidine in depressed men and postmenopausal women, suggesting dysfunctional α_2 adrenergic receptors in depression. Also, blunted responses of GH to imipramine that persist despite effective treatment of depression have been reported, suggesting that subnormal GH response could be a trait marker in depressed patients.

C. Neurochemical Contributions

1. Norepinephrine

The norepinephrine (NE) system is composed of nuclei located in the brain stem that project to cerebral cortex, limbic system, and cerebellum. This includes monosynaptic projections to the amygdala, hippocampus, cingulate gyrus, thalamus, hypothalamus, midbrain, and cerebral cortex. As early as 1929, Cannon described the role of NE in responses to threatening stimuli. NE has been linked to visual, auditory, and somesthetic stimuli as well as pleasurable experiences. The catecholamine hypothesis articulated by Schildkraut posits that mood disorders are the result of a deficit in noradrenergic activity leading to depressive states, whereas increased noradrenergic activity leads to mania. Although the exact relationship between mood disorders and NE activity is not fully understood, a vast amount of research has examined this hypothesis, which suggests that depression is in some way related to inefficient functioning of the NE system and that greater output is required to sustain adequate functioning. Since direct measures of NE or its urinary metabolite 3-methoxy-4-hydroxyphenolglycol are difficult to interpret, physiologic markers have been developed to identify disturbances in NE functioning during depression and the mechanisms of action of antidepressant therapies. One of these markers is the orthostatic challenge test, in which depressed patients show exaggerated increases of plasma NE when rising from a supine to a standing position without altered cardiovascular responses compared to normal subjects.

2. Serotonin

Reserpine, an antihypertensive drug used in the 1950s and 1960s, depletes noradrenaline as well as serotonin and dopamine and frequently led to depression in patients who took this medication. This finding led to the hypothesis that mood disorders were caused by depletions of serotonin and dopamine. Serotonin (5-HT) is produced by the dorsal and median raphe located in the brain stem. The dorsal raphe sends projections to the frontal lobe, amygdala, hypothalamus, and basal ganglia, whereas the median raphe projects to the cingulate gyrus, septum, and hippocampus. Most serotonin is secreted from the dorsal raphe, a well-organized nucleus that coordinates excitation responses in the frontal lobe and amygdala. In 1965, Coppen suggested that a deficiency of 5-HT in the central nervous systen (CNS) could lead to depressive illness. Although it is now possible to measure 5-HT's principal metabolite, 5-hydroxy indole acetic acid (5-HIAA), several studies have reported that a high percentage of depressed patients have normal levels of 5-HIAA in spinal fluid. In addition, 5-HT has been implicated in suicidal

behavior, and lower levels are associated with more violent methods of suicide.

D. Neuroanatomical Abnormalities

1. Background

During the past 15 years, neuroimaging has provided valuable information about the structural brain abnormalities that may play a role in the etiology or mechanism of primary and secondary mood disorders. The first study using computed tomography (CT) in patients with mood disorders was reported in 1980 and the first study using magnetic resonance imaging (MRI) in mood disorders was published in 1983.

One of the major problems in using neuroimaging techniques to study primary mood disorders is that the neuropathology that is etiologically related to this disorder remains unknown (hence the term primary mood disorder). In patients with secondary mood disorders, however, neuroimaging techniques are particularly suited to identify the location and size of structural lesions associated with mood disorders. This technique, based on clinical pathological correlation with spontaneous brain lesion, was first utilized by Broca in the 1860s to identify brain regions that play an important role in specific behaviors.

Broca (1878) also first described an area of cortex that he called the limbus cortex that surrounded the midbrain. This limbus cortex included phylogenetically older areas of cortex which formed a ring around the lower brain stem structures. This term gave rise to the limbic system, which has been the presumed anatomical basis of emotion in humans. Many of these phylogenetically older areas of cortex, such as the basal temporal and inferior frontal cortex, have also been implicated in the mechanism of mood disorders.

2. Generalized Brain Abnormalities

Studies of brain structure in patients with mood disorder have searched for both generalized brain changes and specific regional changes. These studies have usually examined ventricular to brain ratios as a measure of cortical and/or subcortical brain atrophy. Although a number of studies have reported increased ventricular to brain ratios in patients with mood disorders compared to controls, this remains an area of controversy. More than half of the studies using either CT scan measurements or MRI measurements have failed to find increased lateral or third ventricular to brain ratios which would indicate subcortical atrophy in patients with either UP or BP disorders.

One study of patients with secondary depression found that patients with poststroke major depression had significantly larger lateral and third ventricular to brain ratios compared to patients with stroke who were nondepressed. This study suggested that subcortical atrophy preceding brain injury might produce a vulnerability to depression in patients with secondary depression.

Measurements of cortical sulcal width as well as overall cerebral volume have not demonstrated significant differences between patients with UP or BP mood disorders and controls.

3. Regional Brain Abnormalities

a. Cortical Changes Studies of frontal lobe abnormalities among patients with mood disorders have reported findings such as smaller mean frontal area or higher T_1 values on MRI scan in the frontal lobe in patients with UP depressive disorder but not in patients with BP disorder. However, additional studies of prefrontal cortical gray matter volume with better definition of what constitutes medial, dorsal, lateral, and orbital frontal cortex are needed to determine whether structural frontal abnormalities are associated with primary mood disorder. Based on consistent findings using functional imaging techniques of abnormal function in prefrontal cortex, it seems likely that structural abnormalities of prefrontal cortex may be identified in the future.

Studies of temporal lobe structure have reported decreased temporal lobe area in patients with mood disorders. These findings, however, have been inconsistent; in fact, one study reported an increase in left temporal volume in BP patients compared to controls. Studies of temporal lobe asymmetries have reported smaller left compared to right temporal lobe volumes in BP patients compared to controls, but one study found the reversed pattern. Overall, there is lack of agreement as to whether structural temporal abnormalities occur in either BP patients or UP patients. Further studies need to be conducted using better definitions of temporal structures and including subregions within the temporal lobe.

b. Subcortical Changes Examinations of subcortical structures have found conflicting evidence about abnormalities of the thalamus. A recent study reported increased thalamic volume in patients with BP disorder

compared to controls and decreased thalamic volume in unipolars compared to controls. Since the thalamus constitutes part of the cortical–striatal–thalamic loops that have been proposed in numerous studies to constitute the anatomical substrate of mood disorders (Fig. 1), additional studies of thalamic structures including examination of specific nuclei need to be conducted.

Several studies have examined the amygdala and hippocampus. However, there is a lack of consistent evidence demonstrating structural abnormalities in the amygdala or hippocampus in patients with mood disorders. Although the amygdala–hippocampal complex volume was reported in one study to be not significantly different than that of controls, another study reported increased hippocampal volume in bipolars, with the left hippocampus being significantly larger than the right compared with controls.

The largest number of subcortical structural studies have examined abnormalities of the basal ganglia. Several studies have reported significantly smaller putamen and caudate volumes in patients with UP disorder compared to controls. A few studies of patients with BP disorder have reported larger caudate volumes bilaterally in males compared to females.

Studies of subcortical hyperintensities on MRI scan have generally reported that the presence of basal ganglia lesions is strongly associated with mood disorder. These focal hyperintensities (thought to be predominantly vascular lesions) have been the most consistent structural abnormality found in patients with primary mood disorders. However, this appears to be predominantly a phenomenon of the elderly since the majority of studies reporting hyperintensity have been conducted in this population. Only one or two studies of nonelderly depressives found these ischemic abnormalities.

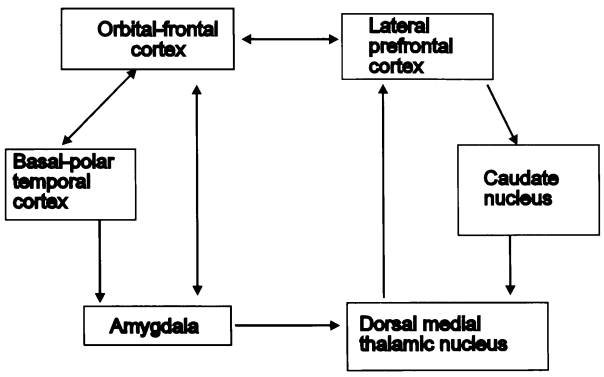


Figure 1 Schematic drawing of the relationship and direction of axonal projections for structures of the ventral–lateral limbic circuit. These interrelationships may explain why both acute lateral frontal (prefrontal) and basal ganglia (caudate) dysfunction could be associated with depression. These circuits may also explain why both metabolic and serotonin S_2 receptor changes in the temporal cortex (basal polar) may also be associated with depression.

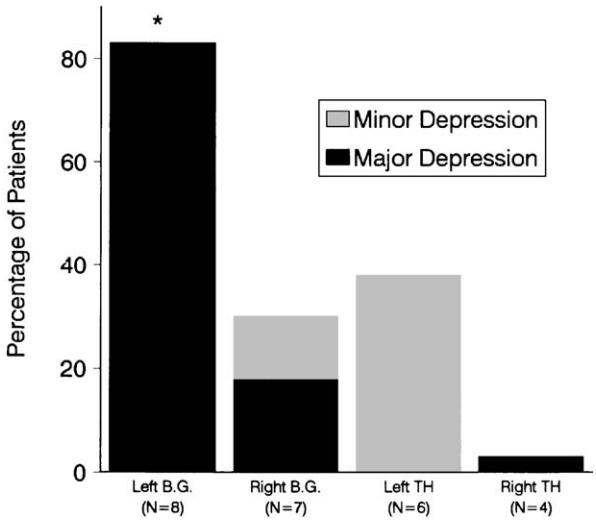


Figure 2 The percentage of patients with major or minor depression based on the localization of their subcortical lesions. Patients with left basal ganglia lesions had a significantly higher frequency of major depression than those with lesions at any other location. BG, basal ganglia; TH, thalamus (data from Starkstein *et al*., 1989, with the permission of Cambridge University Press).

Another new area of investigation in the structural brain imaging of patients with mood disorder has been the identification of silent brain infarctions. Two studies have reported an increased frequency of silent ischemic infarctions as imaged by MRI in patients with senile-onset major depression compared with presenile-onset depression. Silent cerebral infarction was found in 100% of patients with onset of depression after age 65 compared to 53.6% with onset of depression between ages 50 and 65 and 20% with onset of depression prior to age 50.

Among patients with secondary depressions, three studies have examined the role of basal ganglia lesions in depressive disorder. They have all reported a significant association between infarction in the basal ganglia and poststroke depressive disorder. In contrast, lesions of the thalamus have not been significantly associated with poststroke mood disorder (Fig. 2).

Laterality has also been examined with regard to the structural basis of mood disorders. Although

studies of primary mood disorders have not identified lateralized structural asymmetries associated with mood disorder, functional imaging studies in patients with primary mood disorders have frequently reported brain asymmetries. In addition, studies of secondary depressive disorder have frequently reported an association between laterality of brain injury and frequency of mood disorders. Five independent studies have reported that major depression was significantly more frequent among patients with acute or subacute left frontal or left basal ganglia lesions compared to those with lesions in other locations (Fig. 2). However, the association of left frontal and left basal ganglia lesions with major depressive disorder appears to be a phenomenon influenced by time since injury. Patients who are studied within the first few months following stroke show the most prominent association between the hemispheric side of the lesion and the frequency of major poststroke depression (Fig. 3). The studies that have failed to show a lateralized effect of brain lesions on depression have generally examined patients who are several months to several years poststroke.

Additional support for the suggestion that lateralized brain mechanisms may be involved in the production of mood disorders is provided by two controlled case studies and numerous anecdotal case reports of mania associated with injury to the right hemisphere. Thus, left frontal and left basal ganglia lesions have been associated with major depression, whereas right orbital frontal, basotemporal, basal ganglia, and thalamic lesions have been associated with mania.

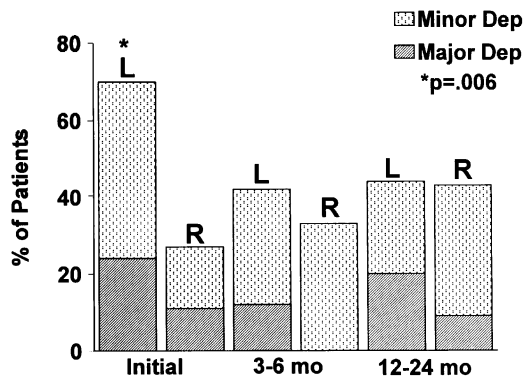


Figure 3 The frequency of major and minor depression defined by *DSM-IV* criteria associated with single lesions of the right (R) or left (L) hemisphere during the acute stroke period and at follow-up. The lateralized effect of left hemisphere lesions on both major and minor depression was found only during the acute stroke period. At short- and long-term follow-up, there were no hemispheric lesion effects on the frequency of depression. From Robinson (1998); reprinted with the permission of Cambridge University Press.

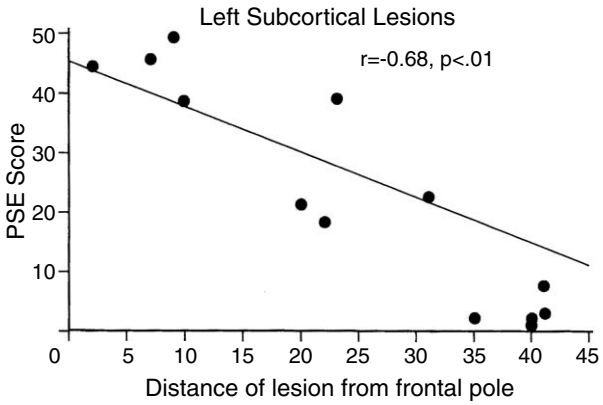


Figure 4 Scattergram showing the correlation between severity of depression based on total Present State Exam (PSE) scores and proximity of the anterior border of the lesion from the frontal pole. Patients all had subcortical lesions restricted to the basal ganglia, thalamus, or internal capsule (data from Starkstein, 1987, with the permission of Cambridge University Press).

Another phenomenon that has emerged from studies of patients with secondary mood disorders is the significant correlation between proximity of the anterior border of the lesion to the frontal pole and severity of associated depressive symptoms (Fig. 4). Eleven studies have found significant correlations between the proximity of a stroke lesion to the front of the brain and severity of depression as measured by a depression rating scale (Fig. 5). There is only one published study that has failed to identify this clinical pathological correlation, and it did so because patients with very large posterior lesions were included in the study. Some studies have found that this correlation holds for both left and right hemisphere lesions, whereas other studies have found that it is significant only among left hemisphere lesion patients or only among patients with lesions of the anterior portion of the left hemisphere. A recent study demonstrated that the clinical pathological correlation between proximity of the lesion to the frontal pole and severity of poststroke depression is also a time-related phenomenon. During the first few months following stroke, the correlation was significant only among patients with left hemisphere lesions. At 3–6 months poststroke, however, the correlation between severity of depression and proximity of the lesion to the frontal lobe was significant in both the right and left hemisphere.

Although the structural or physiological basis for this clinical–pathological correlation has not been identified, it nevertheless constitutes an intriguing phenomenon perhaps related to mechanisms of

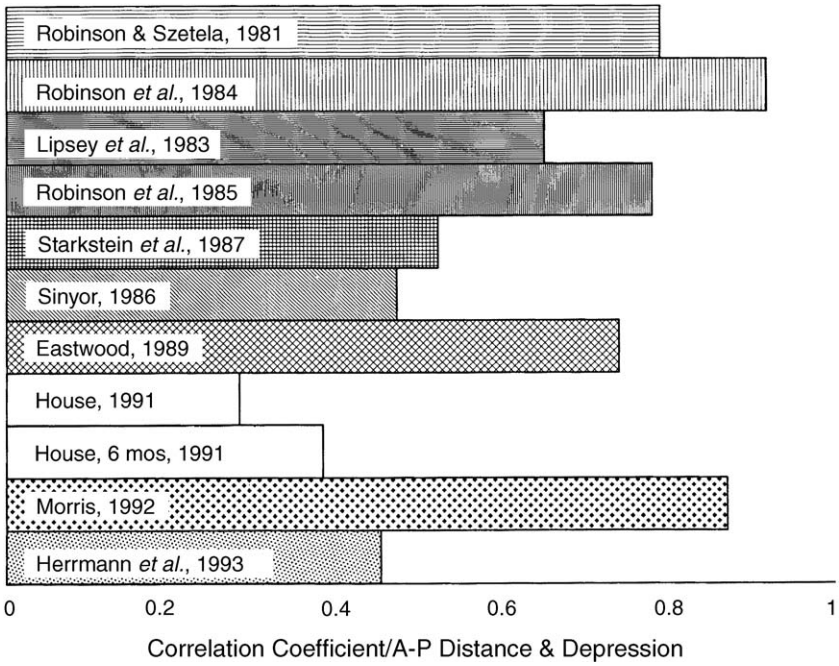


Figure 5 Magnitude of correlation coefficients between severity of depression and proximity of the anterior border of the brain lesion to the frontal pole (left hemisphere or combined left and right hemispheres) for all studies that have examined this correlation. All correlations were statistically significant and the magnitude ranged from relationships that would explain 8% to those that would explain 80% of the variance in depression severity. From Robinson (1998); reprinted with the permission of Cambridge University Press.

depressive disorder. It has been suggested that the pathophysiological basis for this linear correlation between severity of depression and proximity of the lesion to the frontal pole may be a consequence of frontal lesions producing more proximal injury to the norepinephrine- or serotonin-containing neurons and therefore greater depletions of these biogenic amine transmitters. Although other explanations might also be proposed to explain this phenomenon, the fact that multiple investigators have replicated this finding suggests that it may reflect some fundamental mechanism of mood regulation.

In summary, neuroimaging of structural changes in both primary and secondary depression has led to a number of intriguing clinical–pathological findings. Structural abnormalities of brain in primary depression and structural lesions in secondary depression have generally implicated dysfunction of the frontal cortex and basal ganglia in mood disorders. In addition, there may be a laterality effect on depression, with left hemisphere dysfunction associated with depression and right hemisphere dysfunction associated with mania. Alexander and De Long described five frontal–subcortical circuits that appear to play an important role in brain–behavior relationships: Each of these circuits involves the frontal lobe, striatum, globus pallidus, substantia nigra, and thalamus. Findings of abnormalities in primary and secondary mood disorders have suggested abnormalities in the function of one or more of these circuits (Fig. 1). These frontal brain areas associated with mood disorders also contain axonal projections from the biogenic amine-containing neurons located in the brain stem. Thus, frontal or basal ganglia lesions could disrupt function and release of noradrenergic, serotoninergic, or dopaminergic transmissions in the cortex. This combination of basal ganglia and cortical dysfunction associated with disruption of neurotransmitter mediation within the frontal and temporal cortex may provide a common mechanism for both primary and secondary depressive disorder.

E. Physiological Abnormalities

1. Background

The development of functional imaging techniques, including positron emission tomography (PET), single photon emission computed tomography, and xenon inhalation, has provided investigators with powerful tools to examine the physiological and biochemical

correlates of mood disorders. Most investigators, however, have used PET imaging with [^{15}O]-labeled water to measure regional cerebral blood flow or [^{18}F]fluorodeoxyglucose to measure regional brain metabolic abnormalities associated with mood disorders.

2. Positron Emission Tomography Studies

The most common findings of PET studies of patients with depression are caudate and prefrontal hypometabolism. One study examined brain metabolic activity in three types of depressive disorder: UP depression, BP depression, and obsessive/compulsive disorder with secondary major depression. This study demonstrated a decreased glucose metabolic activity in the anterior lateral prefrontal cortex in all three groups of patients compared to controls. There was also a significant correlation between left frontal cortical metabolic rate and Hamilton Depression Score. A study examining regional cerebral blood flow in patients with primary depression and controls found significantly reduced cerebral blood flow in the left anterior cingulate and left dorsal lateral prefrontal cortex. In addition to confirming the frontal and basal ganglia findings, another recent study reported hypometabolism in prefrontal cortex ventral to the genu of the corpus callosum in both familial BP and familial UP depression.

Studies of primary mood disorders have also identified regional brain abnormalities associated with disorders of both mental state (i.e., current clinical disorder) and trait (i.e., abnormalities that persist beyond the current clinical disorder). For example, a trait abnormality of increased blood flow in the left amygdala has been reported in patients with primary UP familial depression. On the other hand, treatment with antidepressants has been found to reverse a state of abnormality by returning blood flow in the left prefrontal cortex to control levels. A fluorodeoxyglucose study has also demonstrated a significant increase in the metabolic rate of the basal ganglia associated with treatment of depressive disorder. Another recent study of patients with UP depression demonstrated that rostral anterior cingulate hypometabolism uniquely differentiated depressed patients who did and did not respond to treatment.

Although receptor studies examining the pathophysiology of primary mood disorders have rarely been conducted, one study found an increase in mu opiate receptor binding in paralimbic areas in patients with UP depression.

Although secondary depressions have not been extensively studied using functional imaging techniques, PET studies have been conducted in patients with depression associated with stroke, Parkinson's disease, or Huntington's disease. A study of metabolic activity in patients with basal ganglia stroke demonstrated that patients with basal ganglia lesions not involving motor pathways (i.e., lesions of the head of the caudate, anterior internal capsule, or anterior or dorsal medial thalamus) had focal ipsilateral hypometabolism involving regions of frontal, temporal, or cingulate cortex. In contrast, patients with lesions of the motor pathways (putamen and posterior internal capsule) had widespread ipsilateral hypometabolism. When these patients with subcortical lesions were classified as having no mood disturbance, mania, or depression, patients with either depression or mania demonstrated hypometabolism in the temporal lobe. Patients with mania showed temporal hypometabolism in the right hemisphere, whereas those with depression showed bilateral temporal hypometabolism as well as cingulate hypometabolism.

Biochemical studies of patients with secondary depression have also been carried out using imaging of serotonin receptor binding. Serotonin S_2 receptor binding was examined in patients with unilateral stroke using ^{11}C-*N*-methylspiperone. This study found that patients with right hemisphere stroke had significantly greater ipsilateral to contralateral S_2 serotonin receptor binding in the temporal and parietal cortex than age-matched healthy controls or patients with left hemisphere stroke. However, there were no significant differences in the amount of S_2 receptor binding in the frontal cortex among patients with right or left hemisphere stroke or normal controls.

Another important finding from this study was that among patients with left hemisphere stroke, the amount of serotonin S_2 binding in the left temporal cortex was highly correlated with the Zung Depression Score (Fig. 6). This finding suggested that lower concentrations of S_2 receptor binding in the left temporal cortex were associated with higher depression scores. This only held true, however, for patients with left hemisphere stroke. There was no significant correlation between depression scores and S_2 receptor binding in the frontal, temporal, or parietal cortex following right hemisphere stroke.

This was the first study to demonstrate a lateralized biochemical response to brain injury in humans. It was also consistent with a previous finding in rats that ischemic lesions of the right but not left hemisphere produced significant depletions of norepinephrine in

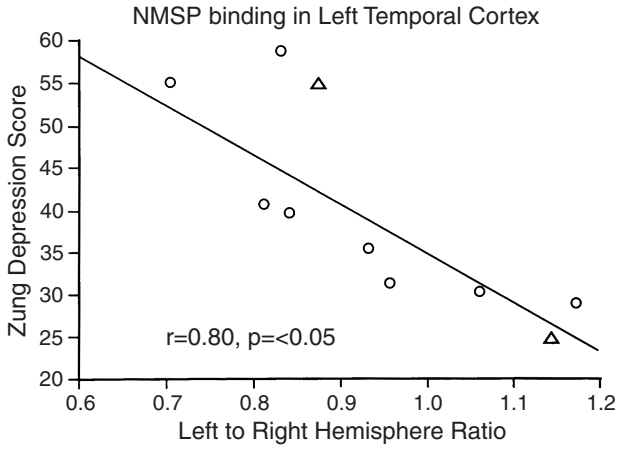


Figure 6 Relationship between Zung depression score and positron emission tomography measures of ^{11}C-*N*-methylspiperone binding to S_2 serotonin receptors in the temporal cortex following a left hemisphere stroke ($N = 8$). The binding is expressed on the horizontal axis as a ratio of binding in the left temporal cortex to binding in an identical region of the right temporal cortex. Lower numbers indicate less S_2 serotonin receptor binding. Lower numbers of S_2 receptors were associated with higher depression scores. The triangles indicate receptor changes in one patient who had spontaneous remission of a major depression. As the depression score fell from 55 to 25, the amount of serotonin S_2 receptor binding increased from 0.86 to 1.15. This patient, with a small left basal ganglia infarct, demonstrated that S_2 receptor binding may be a state marker for poststroke depression (data from Mayberg *et al.*, 1988, with permission of Cambridge University Press).

brain regions both ipsilateral and contralateral to the site of the lesion. Combining these findings from both human and animal studies, it has been hypothesized that left frontal cortical injury or left basal ganglia injury may produce mild to moderate depletions of both norepinephrine and serotonin (and perhaps dopamine). Depletions of serotonin resulting from left hemisphere injury may lead to serotonergic dysfunction in several uninjured brain regions including the temporal lobe. Temporal lobe dysfunction could then initiate changes in limbic circuitry (Fig. 1) through connections from the amygdala to the basal ganglia, thalamus, or orbital frontal cortex, and this could ultimately lead to the manifestations of depression. Right hemisphere lesions, in contrast, may initially produce even greater depletions of norepinephrine and serotonin compared with left hemisphere lesions. The combination of serotonin and norepinephrine depletion has been shown in rats to lead to greater upregulation of serotonin receptors than depletion of serotonin alone. Thus, large depletions of both norepinephrine and serotonin may lead to compensatory upregulation of S_2 receptors in the temporal and parietal cortex. This upregulation in uninjured cortex could lead to enhanced serotonergic function in the right temporal and parietal cortex and therefore block the development of depression.

In summary, functional imaging studies using patients with either primary or secondary mood disorders have, with a moderate degree of consistently, found abnormal activity in the frontal cortex (particularly in the left ventral inferior frontal cortical region) and basal ganglia. In addition, abnormalities have been reported in the cingulate, subgenual cortex, inferior temporal cortex, amygdala, and dorsal medial thalamus. These studies, as well as the structural imaging studies, are consistent with the previously stated hypothesis that frontal–striatal–substantia nigra–dorsal medial thalamus–prefrontal circuits form the anatomical substrate for mood disorders and that hypometabolic effects in the cortex produced by basal ganglia lesions disrupt the function of noradrenergic, serotonergic, and dopaminergic transmissions in the cortex, ultimately leading to either primary or secondary depression.

V. PSYCHOSOCIAL ASPECTS OF MOOD DISORDERS

Psychosocial factors have been correlated with the development and maintenance of mood disorders. Psychoanalytic theories of the etiology of depression concentrate on Freud's belief that the grievance of unresolved early losses undermines the patient's self-esteem. This theory explains that depression is an adult reaction that occurs when the patient is unable to tolerate the negative side of the ambivalence created from early lost objects.

Behavioral theories of depression are influenced by the concept of learned helplessness. Seligman showed that if an animal is repeatedly exposed to a painful stimulus from which it is impossible to escape, a passive acceptance of this stimulus occurs even when the condition is changed so that the animal can flee away from it. Therefore, patients who lack reinforcement of nondepressive behaviors, such as attention and affection, have problems responding positively to social challenges. Also, these patients learn that their depressive behaviors are rewarded with more attention from their significant others.

The cognitive model of depression maintains that biases in thinking play a central role in depression. Beck theorizes that early experiences lead to the development of schemata. These are stable negative

assumption or dysfunctional attitudes that develop during early life experiences and result in inflexible and negative views of the self, world, and future, increasing the patient's vulnerability to episodes of depression.

Although evidence suggests that depressed patients suffer an impairment of social adjustment, the impact of life events and social support on the course of depression is controversial. In their Cumberwell study, Brown and colleagues identified the following social factors that explained the more prevalent numbers of depression in working-class women compared to middle-class women: The working-class women (i) lacked an intimate relationship with a husband or cohabitant, (ii) had three or more children at home, and (iii) lost their mothers before age 11. Weissman and Paykel, using the Social Adjustment Scale, showed that depressed women had marital, parental, ocupational, social, and leisure adjustment deficits.

VI. THERAPY FOR MOOD DISORDERS

A. Antidepressant Drugs

1. Tricyclic Antidepressants

Introduced in 1959, the tricyclic antidepressants (TCAs) constituted the first line of treatment for depression until the appearance of the selective serotonin reuptake inhibitors in the 1980s. The first TCA to be introduced was imipramine, which was found to have antidepressant actions instead of sedative effects that were expected based on its structure analogous to phenothiazine. Currently, there are nine TCAs (imipramine, amitriptyline, desipramine, nortriptyline, clomipramine, trimipramine, doxepine, protriptyline, and amoxapine) and one tetracyclic (maprotiline) drug marketed in the United States. Central nervous system side effects are mostly due to their antimuscarinic action. Sedation, confusion, or seizures resulting from lowering of seizure threshold occur in about 15% of patients. Peripheral side effects include significant antimuscarinic effects on the cardiovascular system (arrhythmias and tachycardia) as well as blurred vision, dry mouth, constipation, and urinary retention.

2. Monoamine Oxidase Inhibitors

In 1952, Selikoff *et al.* reported that the antituberculosis agent iproniazid also possessed mood-elevating properties, and in 1965 Zeller explained its action as secondary to the inhibition of monoamine oxidase (MAO) enzymes. There are two isoenzymes: (i) MAO type A, which degrades mostly NE, epinephrine, and 5-HT, and (ii) MAO type B, which degrades mostly phenylethylamine, phentylamine, benzylamine, and phenylethanolamine. Tyramine and dopamine are nonselective substrates. The MAO inhibitors (MAOIs) can be classified according to the type of MAO inhibition they cause and whether the inhibition is reversible or irreversible. The available MAOIs in the United States are the hydrazine derivatives that are non-selective and have an irreversible action on the MAO enzyme. Moclobemide is a nonhydrazine MAO-A-selective and reversible compound, and deprenyl is useful in the treatment of Parkinson's disease because it increases the levels of dopamine in the CNS due to its MAO-B-selective effect in low doses. MAOIs have significant side effects and interactions. The increased availability of biogenic amines in the CNS causes an overstimulation with agitation and insomnia in some patients. Orthostatic hypotension is also a common side effect of this drug, which can precipitate falls in older people.

3. Serotonin Selective Reuptake Inhibitors

The serotonin selective reuptake inhibitors (SSRIs) are considered by most clinicians to be the first choice for the treatment of depression because of their relatively benign side effect profile. Their mechanism of action is similar to that of the TCAs but with a highly specific 5-HT reuptake inhibition action preventing the degradation of this neurotransmitter by MAO. There are currently five SSRI compounds marketed in the United States: fluoxetine, sertraline, paroxetine, fluvoxamine, and citalopram. Since these SSRIs differ in their potency, selectivity, and side effects profile, patients intolerant to one compound of this group can often be treated with another. SSRIs have a significant cytochrome P450 inhibition action that can result in a toxic level of coadministered drugs, such as alcohol, antihistamines, and anticholinergics.

B. Mood-Stabilizing Drugs

1. Lithium

After observing that lithium produced placidity in guinea pigs, in 1949 Cade suggested that this compound could possess antiexcitement properties in

humans. Although lithium was introduced as a treatment for acute mania, the antimanic effect is maximal after 7–10 days, thus, faster acting agents such as neuroleptics are often used in conjunction with lithium. Lithium has been shown to provide effective prophylaxis for BP disorder. Bipolar lithium-treated patients have a risk of relapse of 34–36% compared to more than 79% in comparable patients given placebo. Lithium has a narrow therapeutic index, however, and high levels can produce seizures, coma, and eventually death. Even when it is present in its therapeutic concentration range of 0.8–1.2 mEq/liter, patients treated with lithium can have gastrointestinal side effects, such as vomiting and/or diarrhea, endocrine side effects, such as inhibition of thyroid function, renal side effects, such as polyuria with polydipsia, and CNS side effects, such as drowsiness, memory deficits, and tremors.

2. Alternative Mood-Stabilizing Drugs

Anticonvulsants are a new choice for the stabilization of mood in BP disorder. Carbamazepine is an anticonvulsant with a tricyclic structure similar to that of imipramine that can be used in lithium-resistant or rapid-cycling BP patients.

The only placebo-controlled study with carbamazepine reported that 60% of a population of BP patients receiving this drug did not relapse compared to 22% of patients receiving placebo.

Another anticonvulsant that has been shown to be effective as a mood-stabilizing drug is valproate. The first double-blind study of valproate in acute mania found a significant effect of this drug in the treatment of acute mania as well as prophylactic effects in BP patients. Although the mechanism by which valproate produces this mood-stabilizing action is unknown, recent studies suggested that the increased CNS levels of γ-aminobutyric acid (GABA) appearing after the inhibition of its degradation by this drug might play an important role. Although it is usually well tolerated, it can produce irreversible hepatic failure that can be fatal.

C. Electroconvulsive Therapy and Rapid Transcranial Magnetic Stimulation

Since its introduction as a clinical tool by Cerletti and Bini in 1938, electroconvulsive therapy (ECT) has been the most effective treatment method for mood disorders. In a review of the literature comparing treatment for depression with real ECT or simulated ECT, placebo, TCAs, and MAOIs, Fink showed that real ECT was significantly more effective in each comparison. Studies have also shown that ECT is equally effective in depressed UP and BP patients. The mechanism of action of ECT is not completely elucidated, but animal studies have found enhancement of brain concentrations of several neurotransmitters, such as norepinephrine, serotonin, dopamine, and GABA. These studies suggest that ECT might produce similar changes in neurotransmitter receptors as those produced by antidepressant drugs. The side effects of ECT are generally less severe than those that accompany the use of antidepressants, although loss of memory is a common complaint.

Rapid transcranial magnetic stimulation (rTMS) is a new noninvasive technique that transfers energy from current in a magnetic coil to cortical neurons. The strength of the induced current is a function of the current in the stimulating coil, which develops about 1.5–2 T. The resulting electrical current is able to activate neurons up to 2 cm from the surface of the cortex. Since the energy used in rTMS is approximately 1 million times less than that used for a stimulus with ECT, this method has proven to be painless, thereby allowing its use without concomitant general anesthesia. Pascual-Leone *et al.* reported that left prefrontal stimulation has been shown to produce a significant clinical response in patients with UP medication-resistant psychotic major depression.

In summary, there are a wide variety of treatments for mood disorders that are mediated by brain mechanisms. The TCAs and MAOIs retain an important role for patients who do not respond to SSRIs. ECT remains the treatment of choice for patients who cannot take medications or who are treatment resistant. Finally, TMS is a new treatment approach with benign side effects and may play an important role in future treatment of mood disorders.

VII. SUMMARY

The relationship between brain and mood disorders has been investigated based on numerous brain-related functions, including genetics, neurotransmitters, hormonal activity, anatomical structure, physiological functions, psychosocial aspects, and treatment

response. Integrating these diverse areas of investigation into a coherent mechanism of mood disorders has not been accomplished, but significant progress is being made in all these areas of study. Since mood disorders are based on clinical and not laboratory criteria, the definition and classification of mood disorders play a central role in determining which brain changes are associated with mood disorder, and classification of mood disorders, continues to be an evolving area. The study of secondary depression, in which there is a well-defined neuropathology, has also contributed to our understanding of the brain-related mechanism of mood disorders. The exponential growth of genetic research has generated numerous investigations searching for the fundamental abnormality that leads to mood disorders. These studies continue to search for multigenic explanations for these disorders. Similarly, the development and growth of both structural and functional brain imaging techniques have led to a general consensus that the anatomical substrate of mood disorders involves dysfunction of the frontal cortical and basal ganglia–thalamic neuronal loops. In addition, neurotransmitter and neuroendocrine studies have established that changes in brain biochemistry are important elements that must be incorporated into any comprehensive explanation of the brain-related mechanism of mood disorders. Finally, one of the most successful areas of research in mood disorders is empirically based treatment. Although effective treatment suggests neurochemical mechanisms of mood disorder, the fact that controlled treatment trials have established a firm scientific basis for treatment selection has allowed clinicians to effectively manage these disorders while we await a fundamental understanding of the brain mechanisms involved.

See Also the Following Articles

DEPRESSION • LIMBIC SYSTEM • NEUROPHARMACOLOGY • NEUROPSYCHOLOGICAL ASSESSMENT • NEUROTRANSMITTERS • NOREPINEPHRINE • PSYCHONEUROENDOCRINOLOGY • SCHIZOPHRENIA

Suggested Reading

Bertelsen, A., Harvald, B., and Hauge, M. (1997). A Danish twin study of manic–depressive disorders. *Br. J. Psychiatr.* **130,** 330–351.

Checkley, S. A. (1992). Neuroendocrinology. In *Handbook of Affective Disorders* (E. S. Paykel, Ed.), 2nd ed. Guilford, New York.

Golden, R. N., and Potter, W.Z. (1986). Neurochemical and neuroendocrine dysregulation in affective disorders. In *The Psychiatric Clinics of North America* (R. M. Restak, Ed.), Vol. 9**,** Part 2. Sanders, Philadelphia.

Jarrett, R. B. (1990). Psychosocial aspects of depression and the role of psychotherapy. *J. Clin. Psychiatr.* **51**(6), 26–35.

Kessler, R. C., McGonagle, K. A., Zhao, S., *et al.* (1994). Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States: Results from the National Comorbidity Survey. *Arch. Gen. Psychiatr.* **51,** 8–19.

Mayberg, H. S. (1993). Neuroimaging studies of depression in neurological disease. In *Depression in Neurologic Disease* (S. E. Starkstein and R. G. Robinson, Eds.), pp. 186–216. Johns Hopkins Univ. Press, Baltimore.

Merikangas, K. R., Spence, A., and Kupfer, D. J. (1989). Linkage studies of bipolar disorder: Methodologic and analytic issues. *Arch. Gen. Psychiatr.* **46,** 1137–1141.

Regier, D. A., and Burke, J. D. (1995). Quantitative and experimental methods in psychiatry. In *Comprehensive Textbook of Psychiatry VI* (H. I. Kaplan and B. J. Sadock, Eds.). Williams & Wilkins, Baltimore.

Robinson, R. G. (1998). *The Clinical Neuropsychiatry of Stroke.* Cambridge Univ. Press, Cambridge, UK.

Rouchell, A. M., Pounds, R., and Tierney, J. G. (1996). Depression. In *Textbook of Consultation–Liaison Psychiatry* (J. R. Rundell and M. G. Wise, Eds.). American Psychiatric Press, Washington, DC.

Soares, J. C., and Mann, J. J. (1997). The anatomy of mood disorders—Review of structural neuroimaging studies. *Biol. Psychiatr.* **41,** 86–106.

Thase, M. E., and Howland, R. H. (1995). Biological processes in depression: An updated review and integration. In *Handbook of Depression* (E. E. Beckham and W. R. Leber, Eds.), 2nd ed. Guilford, New York.

Trimble, M. R. (1988). *Biological Psychiatry.* Wiley, New York.

Winokur, G., Clayton, P. J., and Reich, T. (1969). *Manic–Depressive Illness.* Mosby, St. Louis.

Motion Processing

PREETI VERGHESE
Smith Kettlewell Eye Research Institute

BRENT R. BEUTTER
NASA Ames Research Center

GLOSSARY

adaptation The change in sensitivity of a neuron (or of the animal as a whole) in response to prolonged viewing of a stimulus.

bandwidth A measure of the neuron's selectivity for a property such as motion direction. A large bandwidth refers to a neuron that responds to a broad range of directions, whereas a narrow bandwidth refers to a neuron that responds to a more selective range of directions.

coherent motion A type of motion in which all the elements move in the same direction and at the same speed.

contrast A measure of the deviation of local light intensity from the average intensity.

direction selectivity A preferential response to light moving in a narrow range of directions.

gabor The profile described as the product of a sine wave grating and a Gaussian that is often used to describe the spatial sensitivity of neurons in striate cortex.

gaussian A smoothing function that has a bell-shaped profile described by a normal distribution.

grating A pattern of parallel light and dark bars.

heading direction The direction in which the observer is moving.

incoherent motion A type of motion in which the elements move in different directions or with different speeds. Each element's velocity is randomly changed over time.

mask A pattern that occurs closely in space or time to a test pattern and affects the response of the neuron (or of the observer) to the test.

motion transparency The perception of transparent sheets moving over one another produced when two sets of elements move in widely different directions or speeds.

optic flow The motion on the retina produced by observers moving relative to their environment.

preferred direction The direction of motion that evokes the strongest response in a motion-sensitive neuron.

psychometric function A function relating perceptual performance to a physical stimulus variable. The measure of perceptual performance is the proportion of perceptual decisions (e.g., rightward/leftward judgments or correct/incorrect responses).

sine wave (sinusoidal) grating A pattern of light and dark bars in which the contrast across the bars varies as a sine wave.

spatial receptive field The region of the visual field in which a visual stimulus evokes a response in a neuron.

space–time orientation A property of a receptive field that makes it sensitive to a particular direction and speed. The receptive field of the motion-selective unit has a spatial response profile that changes with time.

spatial frequency A measure that is inversely related to the width of the bars of a grating. The wider these are, the lower the spatial frequency. It is measured in cycles per degree of visual angle.

temporal frequency The analog of spatial frequency in time. It refers to the rate of alternation of a stimulus and is expressed in cycles per second or hertz.

threshold A criterion performance level between chance and perfect performance that is used to compare data across different conditions.

visual angle The angle that an object subtends at the eye.

Everyday visual experience is replete with moving objects. A successful interaction with the outside world depends on the accurate perception of motion. Motion information is critical in estimating one's heading direction, in estimating the time to collision with a moving object, in providing input to the vestibular and eye movement systems, and in breaking camouflage. Motion information also helps to order surfaces according to depth, and to reveal the three-dimensional (3D) structure of objects. Although the visual system of even the fly is capable of motion processing, primates have evolved sophisticated motion processing systems that are represented in multiple cortical areas.

I. INTRODUCTION

One of the first demonstrations of a motion-selective process was the waterfall illusion. After watching a cascading waterfall for a few seconds, Robert Addams shifted his gaze to the rocks on one side of the waterfall and was surprised to find that the stationary rocks appeared to move upwards, in the direction opposite to the water flow. This phenomenon, called the motion aftereffect, has been extensively studied. Prolonged viewing of motion in one direction (the adapting stimulus) desensitizes the observer to motion in that direction, such that a stationary stimulus appears to move in the opposite direction. Recent studies by Bob Sekuler and coworkers have shown that this desensitization occurs over a narrow range of directions centered on the direction of the adapting stimulus. These results suggest that the underlying motion sensors are direction selective and respond only to a narrow range of directions.

Another motion illusion is that stationary flashing lights on movie theater marquees appear to move. This phenomenon, called apparent motion, occurs when an object (such as a flashing light) appears briefly at one location, and after a short time delay a similar object appears at another location. In the case of theater marquees, the time and distance between adjacent flashing lights are large, so we can readily distinguish this motion from a real moving object. If the space and time intervals between successive presentations are small, apparent motion is indistinguishable from continuous motion. Although both television (NTSC 60 Hz or PAL 50 Hz) and movies (24 Hz) consist of animated sequences of static images, we generally perceive them as real motion. The perceptual equivalence of the real and sampled motions suggests that motion-selective neurons are insensitive to the additional information present in the continuous motion representation. In fact, the spatial and temporal sampling rates above which smooth and sampled motion are indistinguishable correspond to the human window of visibility—the limits of spatial and temporal frequency sensitivity of the human visual system (approximately 60 Hz and 50 cycles/degree).

Figure 1 provides a brief outline of the motion processing pathways in primates. It will serve as a road map as we discuss some of the important anatomical areas involved in motion processing. The pathway for visual motion processing in primates starts with the two main cell types in the retina: the magnocellular neurons that project mainly to areas involved with motion and depth processing and the parvocellular neurons that project mainly to areas involved with form and color processing. The magnocellular neurons project to two distinct layers in the lateral geniculate nucleus (LGN) in the thalamus on their way to cortex. From the LGN, these neurons project to layer 4ca in striate cortex, which is the first visual area in the brain (V1). About one-third of V1 neurons are selective for the direction of motion, but due to their small receptive fields they respond only to "local" motion. The V1 neurons in turn project to the middle temporal area (MT), which is an extrastriate area that is specialized for motion processing in primates. About 90% of the neurons in area MT are selective for the direction of motion; furthermore, these neurons are arranged in columns, according to the preferred directions of motion. There is a smooth gradation of preferred direction across neurons in neighboring columns. Neurons in area MT can detect combinations of local motions as well as the motion of a target relative to the background. The adjoining medial superior temporal area (MST) receives projections from area MT as well as from eye movement areas, such as the superior colliculus, the lateral intraparietal area, and the frontal eye fields. Area MST has neurons capable of analyzing

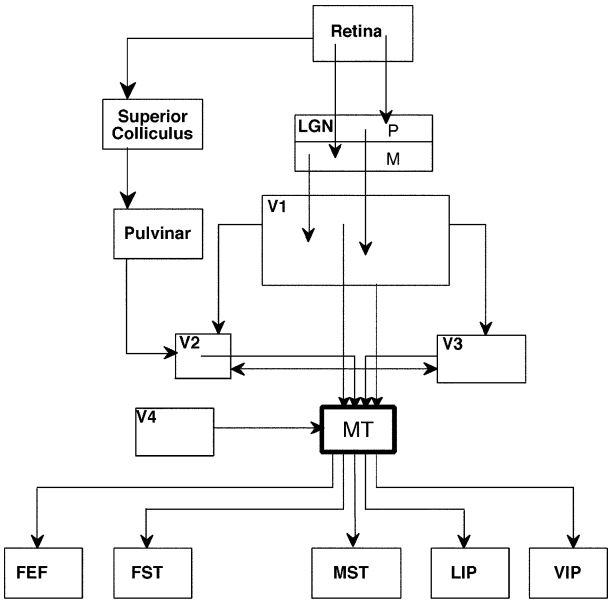


Figure 1 A broad outline of the brain pathways involved in processing visual motion information. Light is detected by the retina (top), which is connected to the lateral geniculate nucleus (LGN), which passes information to the visual cortex. Area V1 is the earliest cortical locus. From it, information is passed to the middle temporal area (MT) and the middle superior temporal area (MST), which are areas that specialize in the processing of motion information. The arrows show only the information going in the feedforward direction, but there are also feedback connections within cortical areas as well as from area V1 to the LGN. FST, fundus of the superior temporal area; VIP, ventral intraparietal area; LIP, lateral intraparietal area, FEF, frontal eye fields [modified with permission from Snowden, R. J. (1994). Motion processing in the primate cerebral cortex. In *Visual Detection of Motion* (A. T. Smith and R. J. Snowden, Eds.). Academic Press, London].

different patterns of optic flow and potentially estimating heading direction. This area also generates signals for smooth pursuit eye movements. During the course of this article, we will pose the problems that a generic motion analysis system must solve and present plausible candidate locations in the primate visual pathway. Although it is tempting to assume that a neuron with a particular property is responsible for the animal's sensitivity to that property, we believe that the animal's sensitivity is likely determined by a population of neurons sensitive to that property.

This article first discusses simple motion-related processing and then discusses more complicated motion phenomena. The organization of this article also reflects a progression from lower to higher visual areas. However, we must state that visual processing is not purely hierarchical; as evidence accumulates for certain interpretations of the visual scene, feedback from higher areas modulates the responses of lower areas, thus strengthening an emerging percept. The hardware for this interaction is certainly in place; there is abundant evidence for feedback connections from higher areas to lower areas. Also, perception might depend on more than the visual stimulus. The observer's past experience with visual stimuli can influence the perception and interpretation of a visual scene. For example, when a moving object disappears for a brief time behind an opaque occluder, there is no visual motion information during its traverse behind the occluder. However, we interpret the image correctly as an object moving smoothly behind an occluder rather than as an object that disappears for a brief time and coincidentally reappears with a similar trajectory at the other side of the occluder. This certainly has to do with our experience with the trajectories of moving objects and with occluders.

II. MEASUREMENT OF MOTION

A. Receptive Field Structure

Several biological systems need to measure motion, whether it is a frog trying to catch a fly, a prey trying to avoid a predator, or a human trying to catch a baseball. Motion is the change in position of an object over time. Therefore, one way to measure motion is to identify an object, or a complex feature of an object, measure its location at one instant in time and again at another instant in time, and use these measurements to compute both the speed and direction of motion. Human visual motion sensors do not employ this strategy of detecting high-level features and keeping track of their locations over time, but instead compute motion signals by responding directly to local changes in intensity over time. Evidence from physiology shows that although the earliest neurons sensitive to motion respond to a wide range of stimuli, they respond best to a narrow range of spatial sizes and temporal intervals. The responses of most of these neurons can be described by simple receptive fields, which are oriented in space–time and are selective for both direction and speed. Many experiments have characterized these neurons' receptive fields by determining their responses to bars of varying widths, orientations, and speeds.

As visual information is passed from the retina to higher level processing areas, the receptive fields of motion-sensitive neurons become increasingly complex. Here, we focus on how the brain achieves motion-

sensitivity in the striate cortex. The responses of motion-sensitive striate cortex cells are directionally *selective*, which means that their response to moving stimuli depends on the direction of motion. Motion-sensitive cells have large responses to motion in their preferred direction and much smaller responses to motion in the opposite direction. Direction selectivity requires receptive fields to be space–time oriented.

To appreciate how a space–time-oriented receptive field is built, we first consider the responses of nondirectionally selective neurons whose responses are independent of the direction of motion. These cells have simpler receptive fields that are space–time separable, which means that their space–time receptive field is the product of a spatial receptive field with a temporal weighting function. Readers are probably familiar with the concept of a spatial receptive field. The spatial profile of many V1 receptive fields is well characterized as a two-dimensional (2D) Gaussian, multiplied by a sinusoid as shown in Fig. 2A. This receptive field is localized in space and responds best to a particular orientation. The cell's preferred orientation and width are determined by the orientation and spatial frequency of the sinusoid. Neural responses to stimuli also depend on time. The neural response to a brief flash of light extends over a period of time (about 100 msec) and is described by a temporal weighting function (or temporal impulse response function). Space–time-separable receptive fields are the product of this spatial receptive field with a temporal weighting function. The temporal weighting function is generally modeled as a biphasic response with a positive response that is followed by a more extended negative response. Two examples of temporal response functions are shown in Fig. 2B. For space–time-separable receptive fields, the spatial receptive field polarity changes with time, but its profile does not change (i.e., it is not oriented along the time dimension). Figure 2C shows an x–t plot of a typical space–time-separable V1 receptive field that is selective for a static vertical bar. This space–time-separable receptive field is unselective for motion direction (i.e., it responds similarly to leftward and rightward motion). This can be seen by comparing such a receptive field profile to Fig. 3, which shows x–t plots of a bar that is moving rightward (Fig. 3B), moving leftward (Fig. 3C), or is static (Fig. 3D).

In contrast, a motion-sensitive cell responds strongly to motion in its preferred direction but weakly or not at all to motion in the opposite direction. Figure 4B shows four examples of motion-sensitive receptive fields that are oriented in space–time; their response at each spatial location depends on time. Such a neuron is selective for a particular velocity and direction. Achieving a spatial response that changes with time is difficult because stimulation of a particular receptive field region must produce a positive, negative, or zero response depending on the time. One way to construct a space–time-oriented receptive field is to combine space–time-separable receptive fields (Fig. 4A shows four space–time-separable receptive fields). Space–time-oriented receptive fields can be constructed by combining (either adding or subtracting) the outputs of two such cells, such that the response of the second is shifted in space and time relative to the first (Fig. 4B shows four different space–time-oriented receptive fields constructed by combining pairs of the receptive fields shown in Fig. 4A). Space–time-oriented receptive fields are selective for motion direction, but their responses are polarity specific; their response to a dark bar is inverted compared to their response to a bright bar. To detect the motion of an object reliably, the neuron's response should not depend on the polarity of the moving bar. One way of achieving this is to square the response. It appears that the visual system achieves a polarity-insensitive filter by taking the sum of the squared responses of the two filters shown at the top (or bottom) of Fig. 4B. This measure based on the squared responses of local direction- and speed-selective units is called motion energy. Ted Adelson and Jim Bergen proposed that this motion energy stage is followed by an opponent stage that takes the difference of local detectors tuned to opposite directions (e.g., the difference between rightward and leftward responses). Physiological measurements support the existence of space–time-oriented receptive fields that compute motion energy, but they do not support the existence of an opponent energy stage.

B. Direction Coding

Bob Sekuler and coworkers measured the human ability to discriminate the direction of motion of large patches of moving dots. They showed that observers could discriminate directions that were 1 or 2° apart. They estimate that the bandwidth of directionally selective mechanisms is about 60°, and that 12 such mechanisms evenly distributed about 360° can account for human direction discrimination. Humans likely achieve their sensitivity to small direction differences by comparing the patterns of activation across these broadly tuned mechanisms. This psychophysical bandwidth estimate is consistent with physiological

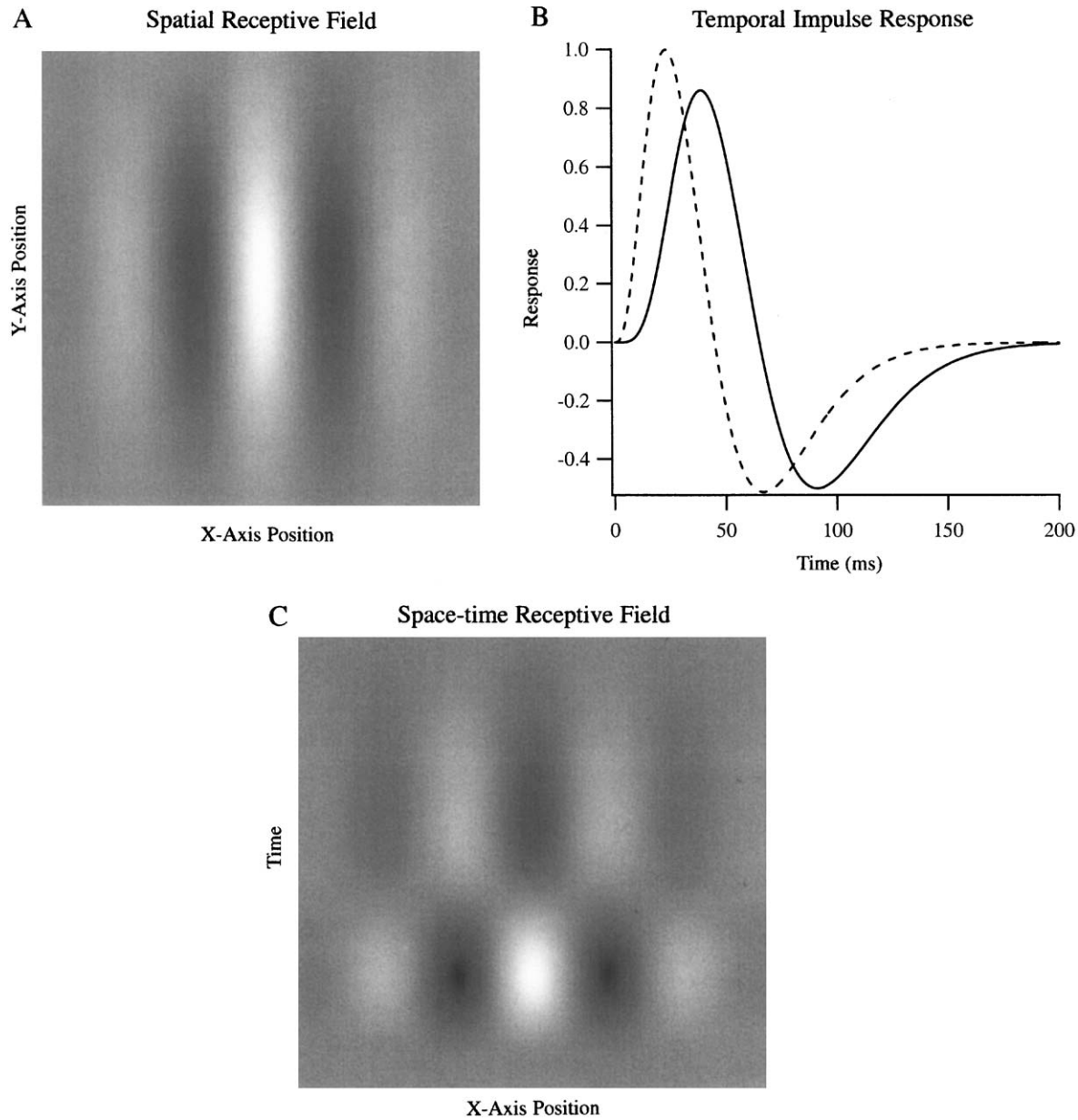


Figure 2 (A) A spatial receptive field. This is an x–y plot of a neuron's response to a spot of light at different positions. In general, the neuron's response to a spatially extended stimulus is computed by multiplying the stimulus contrast at each position by the receptive field value and summing over all receptive field locations. Bright regions indicate excitatory responses, whereas dark areas indicate inhibitory responses. The receptive field shown is a vertical spatial Gabor in cosine phase: $R(x,y) = \exp(-(x^2 + y^2)/\sigma^2) \cdot \cos(2 \cdot \pi \cdot f \cdot x)$, where σ is related to the spatial width of the Gaussian and f is the spatial frequency of the sine wave. This receptive field responds strongly to bright vertical bars with widths matched to the cell's excitatory center ($\sim 0.5/f$). (B) Two temporal weighting functions described by Adelson and Bergen. The response amplitude as a function of time to a brief pulse at time 0 is plotted. For temporally extended stimuli, the neuron's response at a given time is the sum over previous times of the weighting function multiplied by the stimulus contrast. The two temporal response functions shown have different delays: The dashed line function's response is maximal at 23 msec, whereas the solid line function's response is maximal at 39 msec. (C) An x–t plot of a space–time separable receptive field. This receptive field is unselective for direction of motion and responds best to a static bar. It was constructed by multiplying the spatial receptive field shown in A by the temporal impulse response function shown by the solid line in B and taking an x–t slice through the resulting x–y–t space.

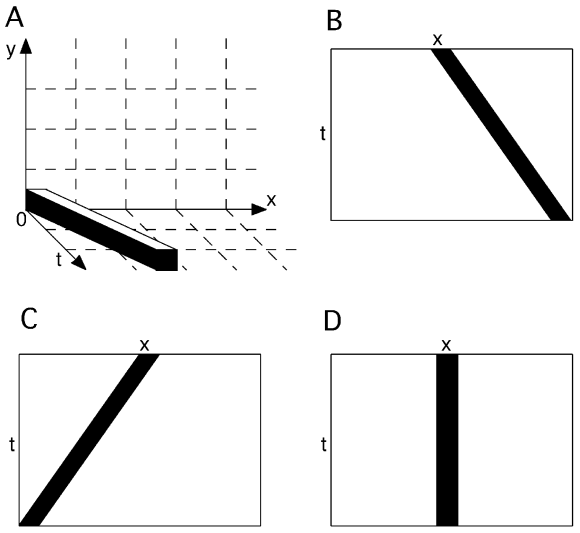


Figure 3 (A) A space–time plot of a dark rectangle translating rightward. It starts at x position = 0, and its x position (horizontal position) increases with time, whereas its y position (vertical position) remains constant. (B) An x–t plot of the rectangle's motion created by taking a slice through the space–time plot at a constant y value. Note that the x axis is horizontal and the t axis is vertical. The origin is at the midpoint of the x axis. Rightward motion starts at $x = 0$ and lies along a line that is tilted down and to right in this plot. (C) Leftward motion starts at $x = 0$ and lies along a line that is tilted down and to the left. (D) A static rectangle whose position remains fixed at $x = 0$. Because its position does not change with time, its graph is a vertical straight line.

estimates that show that primate MT neurons are arranged into direction columns according to their preferred directions of motion. There is a smooth gradation of preferred direction across neurons in neighboring columns. For similar moving dot stimuli, physiological estimates of the direction bandwidth of MT neurons range from 40 to 52°.

C. Speed Coding

Suzanne McKee and coworkers showed that humans are very sensitive to changes in speed and can discriminate a 5% speed difference despite small variations in the contrast, duration, and temporal frequency content of the stimulus. Currently, the way in which the brain processes and codes speed information is less well understood than the neural processing and coding of the direction of motion. Although the space–time-oriented receptive fields of motion energy detectors and complex cells in striate cortex of the cat do respond optimally to a unique velocity, the output of a motion energy unit also depends strongly on both the contrast and size of a moving stimulus. To extract velocity information, which is independent of stimulus contrast and size, Ted Adelson and Jim Bergen proposed a velocity computation stage that calculates the ratio of the outputs of pairs of motion-sensitive and static neurons. If the response of both the static and motion-sensitive cells depends equally on contrast and size, then this divisive normalization largely eliminates the effect of contrast and size on the output of the velocity computation stage. Nonetheless, large changes in contrast affect estimates of perceived speed, with higher contrasts associated with higher perceived speeds.

For a drifting sine wave grating, speed is the ratio of its temporal frequency to spatial frequency. Although earlier models proposed that the visual system codes speed by directly measuring temporal frequency, there is currently robust evidence that the visual system is sensitive to speed per se and not to the combinations of spatial and temporal frequency that characterize the motion. Several studies working with moving bars and dots have provided evidence for speed selectivity in areas V1 and MT. However, these studies do not distinguish between the possibility that the neurons are responding to changes in temporal frequency rather than speed. Bill Newsome and Tony Movshon have found evidence for MT neurons that are tuned to the speed of a moving grating rather than to its temporal or spatial frequency. Further experiments are needed to establish that MT is indeed the site that encodes speed.

III. INTEGRATING LOCAL MOTION SIGNALS

A. The Aperture Problem

As mentioned earlier, there is evidence for space–time-oriented units in visual area 1 (V1), the earliest cortical processing area for visual information. This area has units with small receptive fields that respond best to bars or edges that move in the cell's preferred direction within the receptive fields. Thus, these small units do not directly signal the motion of complex real-world objects but respond only to the motion of image fragments that fall within the aperture that is their receptive field. The problem that a local neuron faces is evident if one views the motion of an extended edge through an aperture. The direction of motion is ambiguous. This is the aperture problem, and it is

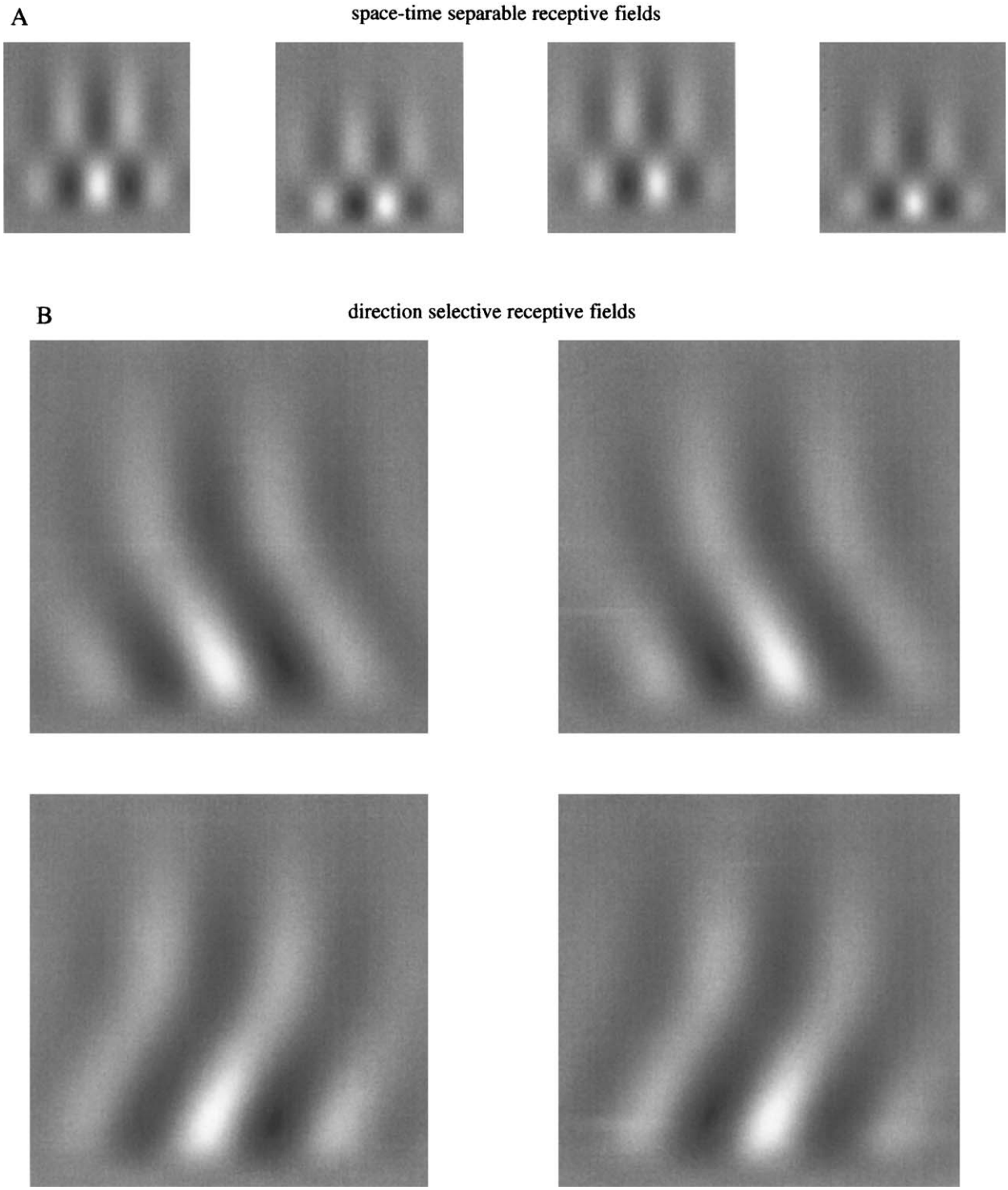


Figure 4 (A) Four space–time separable receptive fields are shown as x–t plots. These receptive fields were constructed by combining sine and cosine phase spatial receptive fields with the two temporal impulse response functions shown in Fig. 2B. All four receptive fields are not oriented in space–time and therefore are not direction selective. (B) Four direction selective receptive fields. The top two are selective for rightward motion (motion in the positive x direction), whereas the bottom two are selective for leftward motion (motion in the negative x direction). These receptive fields were constructed by taking the sum or difference of pairs of space–time separable receptive fields shown in A.

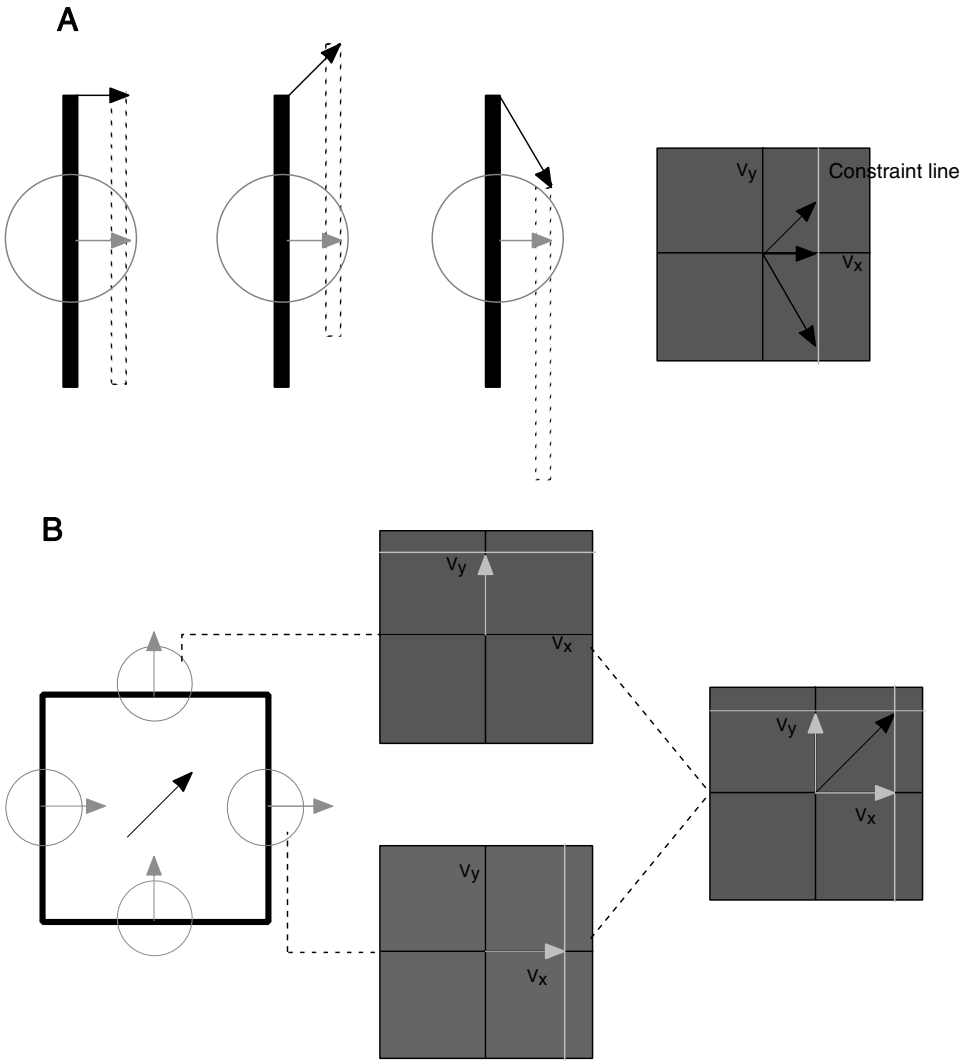


Figure 5 The aperture problem (A) Three instances of a vertical bar (thick black bar) translating in different directions, which all give rise to the same component of horizontal motion. The black arrows depict the true motion of the bar, whereas the gray arrows depict the horizontal component of the motion. The circles represent the spatial extent of a neuron's receptive field. The motion of the bar within this receptive field is identical for all three directions of bar motion. In the leftmost panel the true motion and the motion detector output are identical because the motion direction is orthogonal to the bar orientation. In the second panel the true motion is up and to the right and has a higher speed than in the first case, such that the horizontal component of motion in these two cases is the same. In the third panel, the bar is moving down and to the right but again has a horizontal component of motion equal to that of the first case. In general, an extended moving bar viewed through a small aperture provides only the component of motion perpendicular to the bar orientation. This is the aperture problem. Local measurements of motion within a restricted receptive field are thus consistent with many possible motion directions. As shown in the rightmost panel, all physical velocities with the same horizontal component lie along a constraint line that is parallel to the orientation of the bar. A motion detector tuned to rightward motion cannot distinguish between these three velocities or in fact any velocity that has an equivalent horizontal component. (B) Solving the aperture problem. The figure on the left shows a square translating diagonally upwards. As discussed for A, a detector responding to a single edge of the square can only code the component of motion perpendicular to the orientation of the edge. None of these individual detectors can signal the square's true motion, but its motion can be recovered by combining the constraints defined by the motion of at least two edges. The middle of the figure shows the constraint lines corresponding to the top and right edges of the square. The figure on the right shows that the square's true motion is determined by the intersection of these two constraint lines.

shown in Fig. 5. Figure 5A shows a moving vertical bar. The circle represents the receptive field of a motion unit that is selective for the motion of vertical stimuli. Since the preferred motion direction is perpendicular to the preferred orientation, a unit that is selective for vertical bars responds best to horizontal motion.

Figure 5A shows that this unit responds similarly to a small rightward motion as it does to a much larger diagonal (and rightward) motion. The true direction of object motion can lie anywhere along a constraint line parallel to the moving edge.

The directional ambiguity associated with local motion measurements can be resolved by combining motion estimates from more than one moving edge (Fig. 5B). Consider the case of a square that is moving diagonally upward. It has four edges at two different orientations. Local motion units that are selective for these orientations can only code the direction of motion orthogonal to these edges. Each of these measurements defines a constraint line. The intersection of these constraint lines gives the true direction of motion. The primate visual system also seems to implement such a solution. Units in V1 might detect motion of the top and bottom edges as upward and the motion of the left and right edges as rightward.

There is evidence that a proportion of units in the MT area of the brain, an area specialized for motion processing, combine these two component motions to give the true motion of the object. This corresponds to the direction specified by the intersection of constraint lines. Thus, these neurons are sensitive to the motion of 2D patterns. These pattern units respond best when the true direction of stimulus motion is in their preferred direction. Thus, they respond well to a single bar moving in their preferred direction and to a combination of bars whose intersection-of-constraints direction is their preferred direction. Furthermore, the receptive fields of cells in area MT are on average three times larger than those in corresponding retinal locations in V1, allowing them to integrate motion over a larger area. Their sensitivity to multiple directions of motion and their larger receptive fields make MT neurons well suited to solving the aperture problem. That humans are not normally subject to the aperture problem is proof that the primate visual system has implemented a solution.

B. Integration across Space

The early processing of motion is local; receptive fields of motion-sensitive neurons in area V1 are small (approximately 0.3° at an eccentricity of 1°). Receptive fields at comparable eccentricities in area MT are at least a factor of 3 larger. Although the pooling of motion signals across space reduces the accuracy of local motion estimates, it is useful for measuring large field motion in the presence of noise. Bill Newsome and colleagues demonstrated the ability of neurons in area MT to integrate local motion signals. Their stimulus was a large patch of dots, which were divided into two groups, signal and noise. Each noise dot was plotted at a new random position in each frame, but the signal dots all moved in a consistent direction. Awake behaving monkeys were trained to respond to the direction of motion of the signal dots and the response of individual neurons was recorded. The solid circles in Fig. 6A plot the proportion of correct trials as a function of the percentage of correlated dots. The solid curve through these points is the psychometric function. This function is usually summarized by the coherence threshold, which is the smallest fraction of signal dots that supports criterion performance in the animal. These researchers also showed that a neuron with a preferred direction that matched the signal had a response that increased with the proportion of coherent dots in the display. Furthermore, from the distribution of the neuron's responses to stimuli of various coherence levels, they were able to construct a neurometric function, the equivalent of a psychometric function for a neuron. The open circles and dashed line in Fig. 6A represent a neurometric function, which plots the probability that the neuron's response to stimulus motion in its preferred direction is larger than its response to stimulus motion in the opposite direction as a function of the stimulus coherence level. Thus, a neurometric coherence threshold can be measured in a manner analogous to the behavioral threshold. The resulting neurometric thresholds for the most sensitive MT cells are similar to the psychometric thresholds of awake behaving animals, and both are significantly lower than those predicted from measurements of local V1 units. This makes a good case for the role of MT in the integration of local motion signals.

Two further studies strengthen the role of MT neurons in the integration of this large field stimulus. Punctate lesions in area MT caused severe deficits in the ability of the animal to do the task when the stimulus was presented in the visual field location corresponding to the lesion site. The corresponding location in the (intact) contralateral hemifield served as a control; the monkey's response was unimpaired when the stimulus was presented at this location. Furthermore, electrical microstimulation of a neuron increased the probability that the monkey's perceptual response would be in the stimulated cell's preferred direction. Figure 6B plots psychometric functions with and without stimulation (dashed and solid lines, respectively). Note that the y axis is the proportion

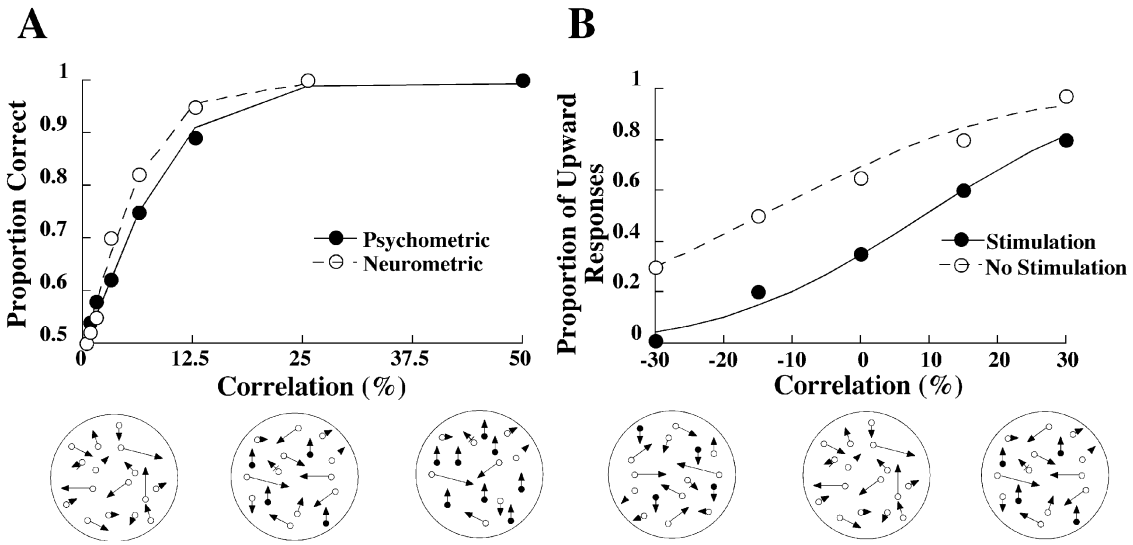


Figure 6 (A) Behavioral and neural responses to the same motion stimuli. The proportion of correct responses is plotted as a function of varying motion coherence as illustrated below the horizontal axis. The open and solid symbols represent psychophsyical and phsyiological data, respectively. (B) Psychometric functions measured with and without electrical stimulation of the neuron being recorded (solid and filled symbols, respectively). The preferred direction of the cell was upward, so the vertical axis plots the proportion of upward responses. In the absence of stimulation, an uncorrelated display (0% coherence) evoked 50% upward responses (chance performance). In the presence of stimulation, the function was shifted leftward so that an uncorrelated display evoked a significant proportion of upward responses.

of upward responses, and that a value of 0.5 represents equal upward and downward responses. In the absence of stimulation, the proportion of "upward" responses increased above 0.5 only when the stimulus had a significant number of upwardly moving signal dots. In the presence of electrical stimulation the proportion of upward responses increased above 0.5 even when the signal dots were actually moving downward (negative values of correlation). In fact, electrical stimulation shifted the psychometric function leftward so that a smaller coherence level in the presence of stimulation produces a response associated with a larger coherence level in the absence of stimulation. This effect of stimulation, along with the lesion studies and the close correspondence between behavior and neural response, provides strong evidence that MT neurons are involved in the processing of large field motion stimuli.

IV. MOTION TRAJECTORIES

Humans are very good at detecting motion along a smooth path. Vilayanur Ramachandran and Stuart Anstis showed that the direction of motion of an ambiguous motion stimulus (e.g., one that is equally likely to be perceived to move in two different directions) is resolved when it is embedded in a motion sequence that moves in a consistent direction. This phenomenon, called motion inertia, refers to the tendency to see motion along a consistent path, despite a brief motion ambiguity. Andrea van Doorn, Jan Koenderink, and coworkers showed that it is much easier to detect coherent motion embedded in incoherent motion when the stimulus is elongated in the direction of motion rather than perpendicular to it. Similarly, Scott Watamaniuk and coworkers showed that a single signal dot moving along a straight path is easily detected among noise dots moving in random directions, despite the fact that in single frames the signal dot and all the noise dots move an equal distance. A quantitative investigation of this effect showed that for extended trajectories (lasting 200 msec), human performance far exceeds the combined outputs of local motion units acting independently. This discrepancy occurs only for long trajectories. For brief trajectories (lasting 100 msec), human performance is predicted by local units with circular profiles. One explanation for this finding is that sensitivity to extended motion trajectories is due to specialized motion units that are elongated in the direction of motion. This explanation has been ruled out by experiments that show that motion signals

along any smooth motion path are easily detected, whether that path is linear, circular, or a triangular wave. It has been suggested that the detection of extended trajectories is mediated by a flexible network of local units that propagate activation among units tuned to similar directions of motion. A unit propagates activation to another unit if its preferred direction roughly points to that unit, thus facilitating the detection of motion signals along a smooth path. This interaction would have to be nonlinear to be consistent with the observed data. We have preliminary evidence that the signal that is propagated along the motion path is a measure of the probability of future motion in a particular direction, given the history of past motion.

V. SECOND-ORDER MOTION

So far, we have discussed motion for gray-scale backgrounds and objects, where the moving object is a region of luminance that changes its position over time. The moving object can also be defined as a region of color that changes positions over time. Luminance and color are regarded as first-order properties of the image. Motion that is produced directly as changes in position over time of these first-order image properties is also called Fourier motion. Our discussion of motion has so far assumed that the brightness and texture of the moving object's retinal image are constant over time. Humans accurately and reliably detect the motion of objects despite dramatic changes of luminance (motion through shadows in a forest) and rapid changes in the surroundings (animals running through waving grass). For conditions like theses, any luminance-based procedure, such as computer algorithms that detect motion by finding regions with matching luminance (or even contrast) over time, would be severely degraded. In these and other cases, there are other invariant properties that define the objects and their motion. Stimuli that have motion that is defined by higher order parameters such as changes in contrast are called second-order motion stimuli (also called non-Fourier motion). We define a second-order motion stimulus to be a stimulus that on average has equal first-order motion energy in opposite directions for all spatial and temporal frequencies. Equivalently, for second-order motion stimuli, the responses of motion energy units tuned to opposite directions are on average equal. Thus, purely first-order mechanisms would be unable to detect second-order motion. The ability of humans to perceive motion in second-order stimuli can provide important evidence about the type of motion mechanisms the brain uses.

Charles Chubb and George Sperling showed that humans perceive motion for many types of second-order motion stimuli. A single frame of a typical second-order motion stimulus is shown in Fig. 7A, and an x–t plot of its motion is shown in Fig. 7B. Humans reliably perceive this second-order motion stimulus to be moving. Chubb and Sperling showed that if the luminance values of certain second-order motion stimuli were subjected to a nonlinear transformation (such as rectification, taking the absolute value, or simple squaring), then the resulting stimulus could be detected by a standard first-order mechanism. The effect of applying a nonlinear transformation (absolute value) to the stimulus is shown in Fig. 7C. Some researchers have suggested that there are two separate motion pathways: one that processes first-order motion using motion energy detectors and a second that consists of a nonlinear stage followed by motion energy detectors. Others have proposed that a single pathway containing an early nonlinearity is sufficient to explain human perception of second-order motion. Although some experiments have found differences in the perception of first- and second-order motion stimuli, evidence is accumulating that motion perception of most second-order stimuli can be explained by a single pathway with an early nonlinearity. Many experiments have shown that there are similarities between the perception of first- and second-order motion. For instance, if a first- or second-order motion stimulus has a contrast high enough to be detected, its direction of motion can be discriminated. First- and second-order motion stimuli produce adaptation effects that are similar and are independent of which type of motion was used as the adapter and the test. They also produce similar evoked potentials. Recent work by Ethan Taub, Jonathan Victor, and Mary Conte showed that direction-discrimination and speed-discrimination thresholds of both first- and second-order motion stimuli were accurately predicted by a single pathway with an early nonlinearity. They showed that an early nonlinearity not only allows the second-order motion stimulus to be processed by first-order mechanisms but also necessarily produces other signals that behave like masks and can strongly influence motion perception. Although these effects depend strongly on the type of early nonlinearity, they likely explain why many experiments find differences between the perception of first- and second-order motion. The

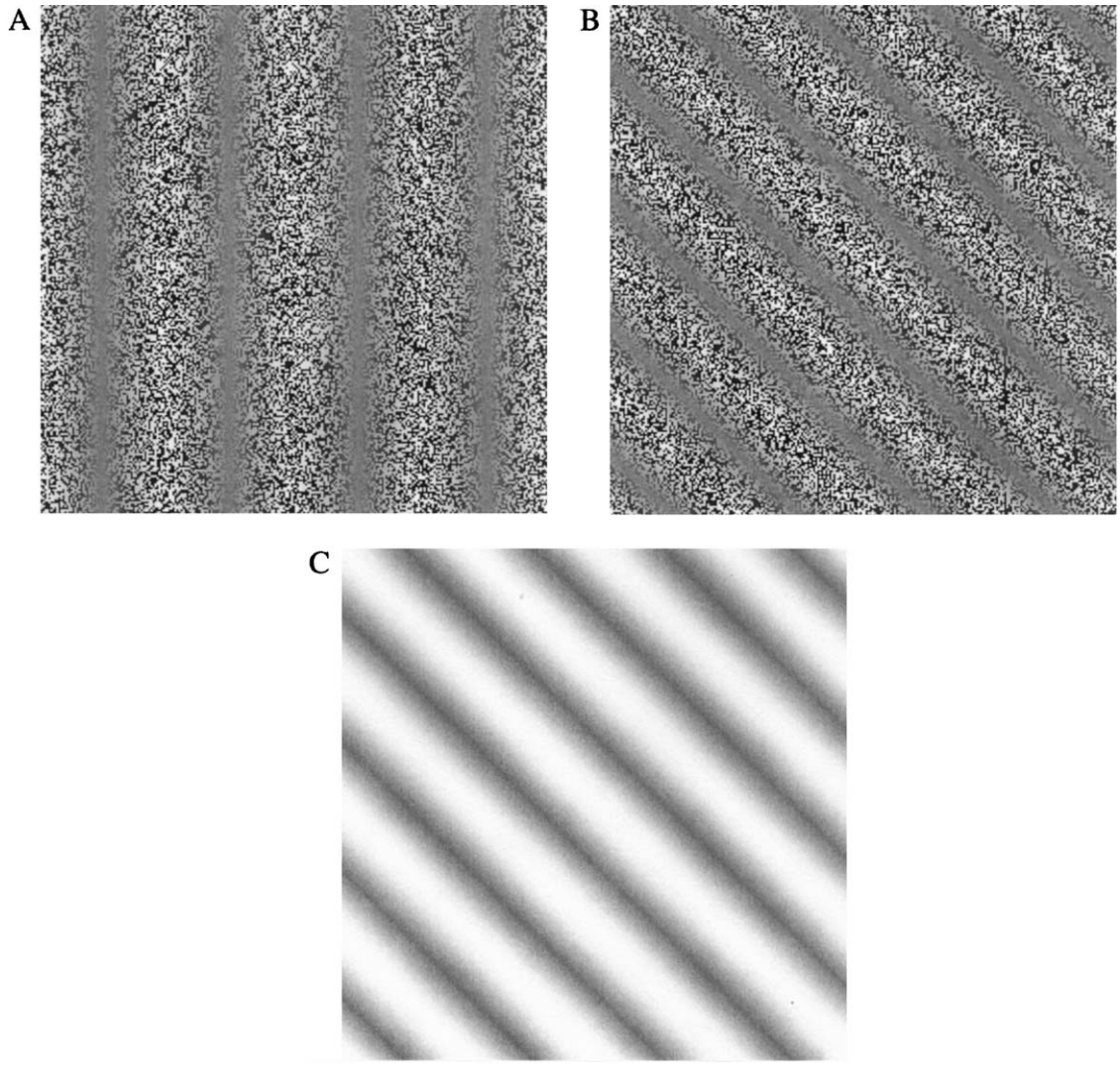


Figure 7 Non-Fourier motion. (A) An *x–y* plot of one frame of a second-order motion stimulus. This stimulus was generated by first creating a checkerboard of small squares and randomly choosing each square to be bright or dark and then multiplying each square's contrast by a sinusoid. This produces regions of very low contrast where the sinusoid is near zero (the approximately uniform gray regions) and regions of very high contrast where the sinusoid is near positive or negative one (the high-contrast speckled regions). (B) An *x–t* plot showing the second-order stimulus moving rightward. (C) The same moving second-order motion stimulus after it has been put through a nonlinearity. In this case the nonlinearity is the absolute value of the contrast. This operation ignores the sign of contrast so that bright and dark squares contrast become equally bright. This nonlinearity produces a moving rectified sinusoid.

investigation of second-order motion stimuli has thus revealed the importance of nonlinear stages in early visual processing.

VI. MOTION AND COLOR

So far, we have discussed the perception of motion defined by luminance. Motion can also be defined by another first-order property—color. Human color perception is described by three technical terms: hue, saturation, and luminance. In everyday usage the word color usually refers to the technical term hue (e.g., red, green, and blue), which is determined by the wavelength of light. Saturation measures how different the light is from white and quantifies the differences between white (zero saturation), pink (medium saturation), and pure red (high saturation). Luminance measures the total light intensity that our photorecep-

tors detect and is related to the brightness or darkness of a region. To investigate the dependence of human motion processing on color, physiologists and psychophysicists use two types of motion stimuli. In the first, luminant stimuli, the motion is created entirely by luminance differences (equal hues and saturations). In the second, isoluminant stimuli, the motion is created entirely by hue and saturation differences (equal luminances).

Many psychophysical experiments have shown that the perception of isoluminant motion is poorer than that of luminant motion. For some types of stimuli, luminant motion is readily perceived, but when the luminance differences are replaced by color differences (isoluminant motion) motion is no longer perceived. Other early experiments showed that isoluminant stimuli were perceived to move more slowly than luminant stimuli. Furthermore, basic anatomy and physiology have shown that the primate visual system has two major pathways as early as the retinal ganglion cells. The magnocellular pathway is primarily involved in motion processing and is not very sensitive to hue, whereas the parvocellular pathway is primarily involved in form and color perception and is not very sensitive to motion. Area MT receives direct input predominately from the magnocellular pathway and is most sensitive to luminant motion and less sensitive to isoluminant motion. These results led some researchers to conclude that luminant and isoluminant motion are processed by two separate pathways. Recent experiments by Brian Wandell and colleagues have directly tested this hypothesis. Extracellular recordings from area MT and functional magnetic resonate imaging results have shown that it does respond to isoluminant motion. The responses to isoluminant motion are similar to those to luminant motion but weaker. Psychophysical experiments have measured how perceived speed depends on contrast for different colors. The function relating perceived speed to contrast has the same shape for all colors but differs by an overall contrast sensitivity factor. For instance, to achieve the same perceived speed, blue stimuli require a contrast 10–20 times higher than that for yellow stimuli. This difference in the sensitivity of perceived speed to contrast is much larger than the differences in contrast detection thresholds for the blue and yellow stimuli. Thus, it is likely that area MT processes both isoluminant and luminant motion, but that MT is much less sensitive to isoluminant motion. The color-dependent differences in sensitivity can explain many of the perceptual differences between isoluminant and luminant motion, including the lower perceived speeds of isoluminant motion and the inability to perceive motion for low-contrast isoluminant stimuli.

VII. MOTION AS A CUE TO DEPTH

A. Kinetic Depth Effect

Relative motion can provide several different views of an object over time and information from these different views can reveal the 3D structure of the object. This is loosely analogous to the way stereo vision combines views from the two eyes to give depth information; since the eyes are horizontally displaced in the head, each eye provides a slightly different view of the world. These different views can be used to extract the relative depths in the image. We first consider the kinetic depth effect, where an object moves with respect to a stationary observer. Hans Wallach demonstrated this effect by projecting the image of an object rotating about a vertical (or horizontal) axis on to a screen. The object was a rigid 3D shape made up of a wire frame. The shadow of the moving object on the screen convincingly gave rise to the percept of a solid shape, showing that humans are able to integrate motion information to recover the 3D structure of objects.

B. Structure from Motion

Another compelling demonstration of how motion information gives strong 3D information is the rotating cylinder example. The observer sees a 2D projection of a pattern of dots that are painted on a rotating glass cylinder rotating about a vertical (or horizontal) axis (Fig. 8). When the cylinder is static the image on the screen looks like a random array of dots. When the cylinder rotates, the dots on the front and back surfaces move in opposite directions. Moreover, the dots in the center of the cylinder have a faster speed than the dots on the edge. Observers are able to integrate these different local motions into the percept of a transparent rotating cylinder.

How does the visual system extract depth from the 2D information generated by a moving 3D shape? Currently, there is not a clear physiological basis for this process. From geometrical constraints it has been shown that the 3D structure can be reconstructed from

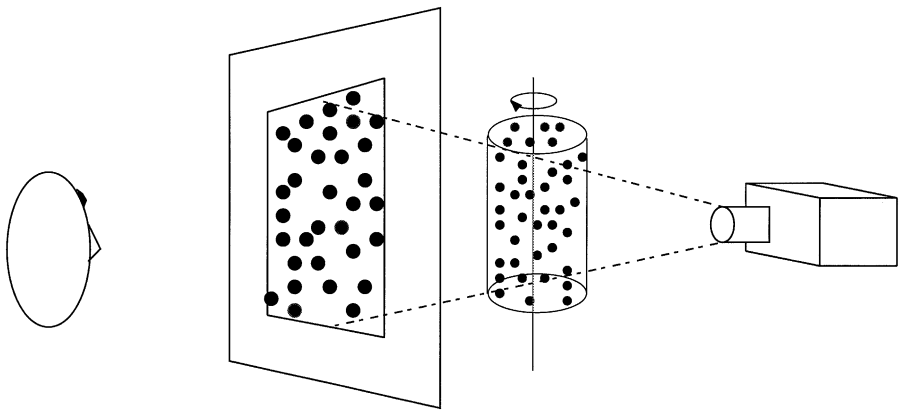


Figure 8 Structure from motion. The observer is viewing the projection of a glass cylinder rotating clockwise with dots painted on its surface. The projection falls on a 2D screen. At each instant in time the projection is a 2D pattern of dots, but as the cylinder rotates the dot pattern changes in a way that conveys the 2D shape of the cylinder. The 2D projection of the dots near the center of the cylinder appears to move at a high velocity (either leftward or rightward depending on whether the dots are on the front or back surface of the cylinder), whereas the 2D projection of the dots near the edges of the cylinder has a low velocity (because actual 3D motion of the dots is primarily toward or away from the projection screen).

just three different views of four points on a rigid object that are not in a single plane. Depth can also be mathematically reconstructed with fewer views of more points. Furthermore, the 2D pattern of moving dots on the screen is ambiguous; it is equally consistent with stationary dots affixed to a rotating cylinder and with a single plane of independent dots moving with different velocities. Humans appear to be biased to assume that the stimulus is an object that is moving rigidly. This is the rigidity constraint in the interpretation of optic flow. In fact, if we incorporate such a constraint it would have to be one of local rigidity since

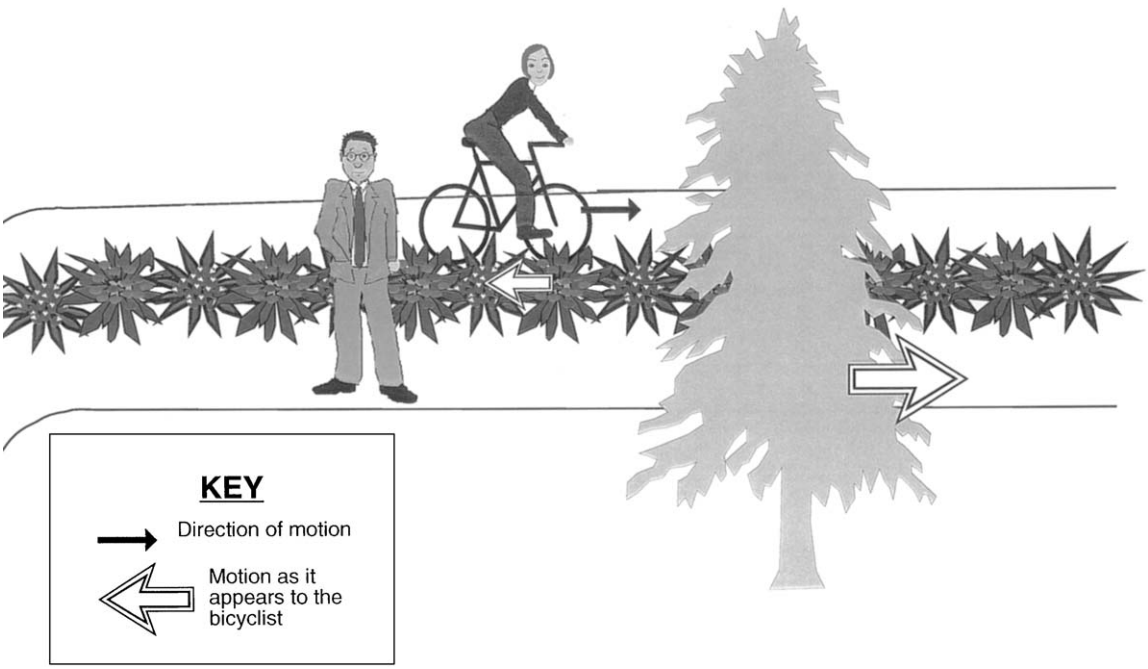


Figure 9 Motion parallax. The woman on the bicycle is looking at the man as she pedals. Objects farther than the man appear to move with her, whereas objects closer than the man appear to move in the opposite direction. Specifically, the tree appears to move in the same direction as she does, and the flower beds appear to move backward.

the movement of most animate objects is not consistent with rigid body motion. Typically, head and limbs move with respect to the torso. It is perhaps this locally rigid motion that makes the phenomenon of biological motion so compelling. Biological motion depicts only the motion of the major joints of a human or animal. Gunnar Johansenn showed that a movie of a person walking in complete darkness, except for small lights attached to major joints, was instantly recognized as a human walking.

C. Motion Parallax

Motion parallax is the relative motion of objects at different depths due to the motion of the observer. This is the kind of information that is available when we look out of the window of a moving vehicle or when we look around while on a moving bicycle. Objects farther than our point of fixation move with us; objects that are closer move in the opposite direction (Fig. 9). This is indeed a powerful monocular cue to depth. Brian Rogers and Maureen Graham demonstrated the effect of motion parallax in the laboratory. Observers viewed a flat screen with randomly positioned dots. They were asked to move their heads from side to side. This lateral motion was instantaneously measured and used to compute differential motion of the random dots for each head position. The dot motion was calculated so that it was consistent with motion parallax from a 3D surface corrugated sinusoidally in depth. This imposed motion was similar to motion that would have impinged on the retina if the observer were moving with respect to a real corrugated surface. Observers had the compelling percept of a surface modulated in depth, despite the fact that there were no stereo cues to the depth.

It is possible that neurons in MT can encode depth from motion parallax. It is known that populations of MT neurons have different preferred directions in the center and surround. A subset of these neurons have centers and surrounds with opposite preferred directions of motion. These cells could potentially encode a motion parallax signal generated by a moving observer fixating a point between two objects separated in depth. Furthermore, we also know that some cells in MT are disparity selective. A cell that combined relative motion specificity with disparity selectivity could potentially encode depth from both motion parallax and stereo information.

VIII. MOTION AS A CUE TO SEGMENTING THE VISUAL SCENE

Motion can help break camouflage; flounders on the seafloor are so well camouflaged that they are nearly impossible to detect until they move. However, when a flounder does move, we are immediately aware of its form. Such real-life examples seem to have inspired psychophysicists to use random dot patterns to study how an object defined by motion information alone segregates from the background. This is the figure-ground problem. In these experiments, the stimulus is a random dot pattern and a single patch of the stimulus moves between frames. Each frame by itself looks like an array of random dots. It is only when the frames are animated that the motion of the patch is visible. Neurons in area MT seem well suited to detecting this kind of motion. As discussed earlier, there is a class of neurons whose surrounds have direction and velocity preferences that are antagonistic to the center response. The cell has no response if the surround alone is stimulated. Some of these neurons have surrounds that are maximally suppressive when the neurons are stimulated by a large field stimulus moving in the preferred direction of the center. Other neurons have surrounds that produce suppression for field movement in any direction. These latter neurons respond optimally when stimulated with a patch that is matched in size to the center and moves in the preferred direction of the center. As the patch gets larger, the surround is also stimulated and the response of the neuron decreases. These neurons are well suited to segmenting an object that is defined by motion information alone, but they do not encode large-field motion generated by the animal's movements.

Recent studies have also shown that humans can use speed information to group areas of common motion. Mary Bravo and Scott Watamaniuk showed that dots moving at a single speed are seen as a coherent group even when their common speed is defined by different combinations of spatial and temporal displacements. Similarly, a motion stimulus composed of regions of different spatial and temporal frequencies appears to move as a single textured object if the spatial and temporal frequencies define a common speed across the object. Conversely, a pattern that differs from its surround only by a change in speed can be easily discriminated. To achieve segmentation by speed, the neurons in MT with antagonistic surrounds would have to have different preferred velocities in the center and surround.

Humans can also use speed and direction information to segment the visual scene. When overlapping sets of points move in different directions or at different speeds, the image segregates into two separate layers moving at different velocities. This phenomenon, called motion transparency, normally occurs when there are large differences in direction or speed between the overlapping motions. Ning Qian and coworkers showed that humans perceive transparency in a display with two overlapping sets of dots that move in opposite directions. Observes perceive flicker and not transparency when the dots are arranged as pairs of points that are very closely spaced and move in opposite directions. The opponent version of the motion energy model provides a potential explanation. The paired dots are at approximately the same location, so the opposing motion directions cancel each other. The unpaired dots are not closely spaced, so the two directions of motions do not cancel each other.

Andrea van Doorn and Jan Koenderink described the spatial and temporal conditions under which transparency is perceived. They used a display that was divided into horizontal strips of equal width. Alternate strips had dots that moved at two different velocities. If the strips were broad, the display appeared to be divided into separate horizontal strips moving at different velocities. If the strips were narrow, the division into horizontal strips was not perceived; instead, the percept was of "transparent sheets moving through each other." The width of the strip that supported the percept of transparency depended on the magnitude of the velocity: It ranged from 0.02 to 0.67° for velocities from .25 to 30°/sec. They also used a homogeneous display that alternated between two speeds at a variable rate. If this rate was slow, humans perceived a single sheet of dots, which changed its velocity over time. If this rate was faster than 10 Hz, then two transparent, simultaneously moving sheets of noise pattern with different velocities could be seen. Besides requiring that the alternations in space and time occurred over small distances and times, the displays appeared transparent only if speeds of the two motions differed by a factor of 4 or the directions were separated by at least 30°. At the level of MT, the spatial separation between signals that support the percept of transparency is typically small enough to fall within the receptive field of a single neuron. This suggests that the presence of two motions at adjacent locations is probably encoded within a single MT cell by subunits (V1 neurons) that have different preferred directions and speeds. In fact, Ning Qian and coworkers have shown that, on average, neurons in macaque MT respond more strongly to the unpaired dot stimuli that generate the transparent motion percept than to the paired dot stimuli that generate a flickering percept.

IX. MOTION INFORMATION GENERATED BY SELF-MOTION

Moving through complicated environments is an everyday necessity that depends critically on visual motion information. Humans, like other animals, are able to perceive and control their motion precisely through complex environments. We have two main types of sensory inputs that help us perceive self-motion: vestibular and visual. Vestibular input from the semicircular canals in the ears directly sense certain types of self-motion by detecting accelerations. Here, we focus on the ability of humans to perceive self-motion from purely visual cues, which are produced by the changing images (motion) our eyes receive as we move.

When a person moves, the environment generates a pattern of motion across the retina, which is called optic flow. As J. J. Gibson described, if the person is moving along a line without rotating (translating), the object toward which he or she is moving stays fixed on the retina but all other objects move in an outward radially expanding pattern that is centered about the direction of translation (Fig. 10A). For pure translation, this stationary point is called the focus of expansion (FOE), and the direction of self-motion is toward its location. When the person is translating slightly to the left of straight ahead, the resulting optic flow pattern is shown in Fig. 10B. Many studies have presented stationary observers with various types of translational optic flow patterns and found that humans can accurately (to within approximately 1°) determine the simulated direction of self-motion under a wide range of conditions. A different type of optic flow pattern results when the observer's position remains fixed but he or she rotates his or her eyes (or body). This rotation produces the optic flow pattern shown in Fig. 10C. If the self-motion is a combination of translation and rotation, resulting from the observer rotating his or her eyes to maintain fixation of an object while walking in a straight line in a different direction, the resulting optic flow pattern is the sum of the translational and rotational optic flow patterns

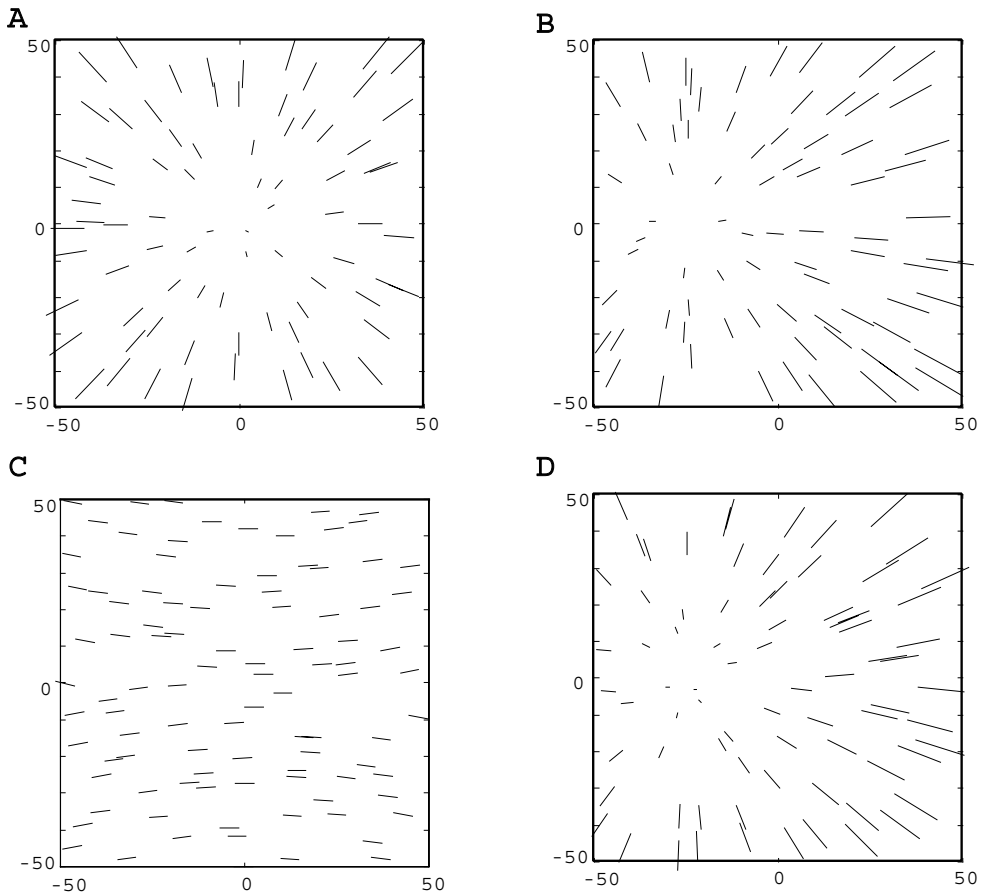


Figure 10 Optic flow (A) The flow field for an observer both looking and moving straight ahead toward a flat wall painted with dots. The lines plotted in the flow field represent the velocities of the moving dots on the retina. As the observer moves closer to the wall, the image expands and each dot moves directly away from the center of the screen. The point from which all the flow field vectors emanate is the focus of expansion (FOE). For an observer moving in a straight line in any direction, the flow field contains an FOE that is always located in the direction the observer is headed. Because the observer is moving toward the center of the wall, the FOE is in the center of the flow field. (B) The flow field produced when an observer is translating in a direction slightly to the left of straight ahead. The focus of expansion has now shifted to the left so that it remains aligned with the translation direction. (C) The flow field produced by an observer looking to the left by rotating her eyes (or head) about a vertical axis. Note that moving the eyes to the left causes all points to move predominantly to the right and that there is no FOE for rotational motion. (D) The flow field resulting from the combination of moving straight ahead, as in A, and rotating the eyes to the left, as shown in C. The resultant flow field is the sum of these two flow fields and looks very similar to the flow field in B. Although it contains a location that has no motion, this is not a true FOE and does not correspond to the direction of translation.

(Fig. 10D). The retinal motion in these optic flow patterns is more complex, and although they contain a location with no motion (a false FOE), this no longer corresponds to the direction of motion. Recent experiments have shown that humans are able to accurately perceive their self-motion in these conditions, suggesting that the brain performs a more complex computation than simply identifying the location of the FOE. Eye movements can also play a role in heading perception. Information about eye rotation has been found to improve the accuracy of heading estimation in some conditions.

The brain appears to process optic flow patterns in area MST and adjacent parietal areas such as area 7A. Complex motion-sensitive neurons are found in the dorsal portion of area MST (MSTd), which receives direct input from area MT and eye movement areas, such as the superior colliculus, the lateral intraparietal area, and the frontal eye fields. Many MSTd neurons have large receptive fields, some of which cover nearly the entire visual field. Instead of responding best to the simple motion of bars, these large-field neurons respond optimally to specific patterns of motion. Some of these cells respond best to the expansion

optic flow patterns produced by translations, other cells respond best to rotational optic flow patterns, and still others appear to respond best to specific combinations of expansion and rotation. Recently, Lee Stone and John Perrone developed a model that predicts both the response of MSTd neurons and the ability of humans to perceive their self-motion. In their model each neuron is tuned to a specific type of self-motion. The neurons have large receptive fields that receive input from MT neurons. Direct support for a role for MST in self-motion perception comes from microstimulation studies in area MST of awake behaving monkeys that were trained on a heading discrimination task. Ken Britten and coworkers showed that electrical microstimulation of MST frequently biased the monkeys' decisions about their heading, and these induced biases were often quite large. These results suggest that MST has a direct role in the perception of heading from optic flow.

X. TIME TO CONTACT

When a fly lands on a wall, a human catches a ball, or a pilot lands a plane, motion information about when the approaching object will arrive (time to contact) is necessary to interact successfully with the environment. Low-level measurements of the motion of an approaching object do not by themselves contain information about the absolute distance and velocity of the object. However, absolute knowledge of object distance and velocity is not required to determine the time to contact. David Lee showed that it is possible to calculate the time to contact with an approaching object by measuring the relative expansion rate (the ratio of the size of the object at a given instant to its rate of change of size). He referred to this ratio as τ. Humans appear to perform such a calculation and base their estimates of time to contact on the relative expansion rate, τ. Psychophysical studies show that human judgments of time to contact increase in proportion to parametric variation in τ, despite variations in initial expansion rate. Other studies suggest that certain motor actions are initiated when the time to contact reaches a critical value. For example, flies begin decelerating when the relative rate of expansion reaches a critical value; humans appear to use τ to initiate appropriate action during driving and catching or hitting a ball. However, research to date has not identified a neurophysiological correlate of τ in primates.

XI. MOTION AND PURSUIT EYE MOVEMENTS

The human visual field has a very high acuity central region, the fovea, and visual acuity declines rapidly for more peripheral regions. To position objects near the center of the fovea, humans and monkeys have developed sophisticated brain circuitry that produces two different types of eye movements: saccades and smooth pursuit. Saccadic eye movements are rapid, step-like changes in eye position, that are used to quickly look around a scene or to read. Saccades have latencies of about 200 msec, very short durations (<50 msec), and can have very high velocities (300°/sec). These extremely fast eye movements cause the world to move relative to the eye with high velocities during saccades, but we do not perceive the world to move. The high saccadic velocities and short durations generally produce world motion on the retina too fast for the visual system to detect.

Humans and monkeys use a different eye movement, smooth pursuit, to track the motion of objects (e.g., a ball rolling or a bird flying). Humans are able to pursue objects moving at speeds up to about 30°/sec with great accuracy. Pursuit eye movements are typically initiated very quickly, about 80–150 msec after the onset of motion, and accelerate rapidly to match the speed of the target in about 80 ms. During steady-state pursuit, the eye velocity approximately matches the target velocity, producing a retinal image in which the target is nearly stationary and the world is moving with a retinal velocity equal and opposite to the eye velocity. Humans nevertheless perceive the world as stationary and the target as moving, and we are able to maintain accurate smooth pursuit of the target even though its retinal velocity is close to zero. The brain computes motion signals in world coordinates by subtracting an eye velocity signal from the retinal motion signals. To obtain a very fast eye velocity signal, the brain appears to use an efference copy of the neural signal sent to the regions generating eye movements instead of direct measurement of eye velocity by proprioceptors. Evidence from several different types of monkey studies indicates that the computation of world velocity from retinal and eye velocities is performed in the lateral portion of area MST (and possibly also in the frontal eye fields). Intracellular recordings from MST neurons show that they receive retinal motion information from area MT and extraretinal eye movement signals. Lesions of MST produce deficits in both motion perception and the ability to generate smooth pursuit. Also, injecting small, carefully controlled electrical signals into localized

regions (microstimulation) of MST produces smooth pursuit and also influences the perception of motion.

Further evidence that the brain calculates an object motion signal that is shared by pursuit and perception is provided by recent experiments by Brent Beutter and Lee Stone that simultaneously measured the direction of pursuit and the perceived direction of motion. These studies examined both image segmentation and errors in perceived direction and showed that the perceptual data could be accurately predicted from measurements of pursuit. Furthermore, by using the pursuit data to predict the perceptual response to each trial, these studies showed that on a trial-by-trial basis perception and pursuit were correlated. Thus, it appears that object motion signals computed by MST are used by both perception and pursuit.

XII. CONCLUSION

Motion processing is critical to many aspects of everyday life. We discussed how the ability to extract motion signals from the environment affects diverse functions from breaking camouflage to navigating through the environment. These diverse functions are reflected in the multiple cortical areas that deal with different aspects of motion processing. Although humans can function without other aspects of visual function such as stereoscopic vision, the loss of motion processing is totally debilitating. This is not surprising considering that motion is a wealthy source of information about the environment, and that motion processing is probably the most important visual function of the brain.

See Also the Following Articles

AREA V2 • COLOR PROCESSING AND COLOR PROCESSING DISORDERS • EYE MOVEMENTS • HAND MOVEMENTS • INFORMATION PROCESSING • NEUROFEEDBACK • OBJECT PERCEPTION • RECEPTIVE FIELD • SPATIAL COGNITION • TIME PASSAGE, NEURAL SUBSTRATES • VISION: BRAIN MECHANISMS • VISUAL SYSTEM DEVELOPMENT AND NEURAL ACTIVITY

Suggested Reading

Maunsell, J. H. R., and Newsome, W. T. (1987). Visual processing in monkey extrastriate cortex. *Ann. Rev. Neurosci.* **10,** 363–401.

Nakayama, K. (1985). Biological motion processing: A review. *Vision Res.* **25,** 625–660.

Newsome, W. T., Britten, K. R., Movshon, J. A., and Shadlen, M. (1989). Single nuerons and perception of visual motion. In *Neural Mechanisms of Visual Perception* (D. M.-K. Lam and C. D. Gilbert, Eds.), Proceedings of the Retina Research Foundation, Vol. 2. Portfolio, The Woodlands, TX.

Sekuler, R., Anstis, S., Braddick, O. J., Brandt, T., Movshon, J. A., and Orban, G. (1990). The perception of motion. In *Visual Perception: The Neurophysiological Foundations* (L. Spillman and J. S. Werner, Eds.). Academic Press, San Diego.

Smith, A. T., and Snowden, R. J. (1994). *Visual Detection of Motion.* Academic Press, London.

Motor Control

STEVEN P. WISE
National Institute of Mental Health

REZA SHADMEHR
Johns Hopkins University

GLOSSARY

agonist A muscle that generates movement toward a goal.

antagonist A muscle that generates force in a direction opposite to that of an agonist.

central pattern generator Neural network in the central nervous system that produces rhythmic motor commands.

forward dynamics Computation of movements that result when muscles generate a given pattern of force.

forward kinematics Computation of limb positions for a given pattern of joint rotations.

internal model A central nervous system representation of the kinematics and dynamics of a motor task.

inverse dynamics Computation of muscle activations and forces needed to reach a goal.

inverse kinematics Computation of joint rotations or limb movements needed to reach a goal.

synergists Muscles that work together to produce a given movement.

The human motor system controls goal-directed movement by selecting the targets of action, generating a motor plan, and coordinating the forces needed to achieve those objectives. Genes encode much of the information required for both the development and the operation of the motor system—especially for actions involving locomotion, orientation and reproduction—but every individual must acquire and store motor memories during his or her lifetime. Some of that information reaches conscious awareness, but much of it does not. This article focuses on the human motor system but draws heavily on knowledge about the motor system of other mammals, especially primates.

I. WHAT CONTROLS MOVEMENT

The motor system consists of two interacting parts: peripheral and central. The peripheral motor system includes muscles and both motor and sensory nerve fibers. The central motor system has components throughout the central nervous system (CNS), including the cerebral cortex, basal ganglia, cerebellum, brain stem, and spinal cord.

A. Peripheral Motor System

1. Muscles

Skeletal muscles consist of specialized cells, which fuse during development to form fibers (technically, a syncitium). There are two types of muscle fibers:

extrafusal and intrafusal. Extrafusal fibers, which attach to tendons and then to the skeleton, produce force and movement. Intrafusal fibers, which contain muscle spindles, attach to muscles and serve a sensory function.

Force and movement depend on muscle proteins, principally myosin and actin, both of which form strands within the muscle fibers. Molecules of myosin store kinetic energy as a result of metabolizing adenosine triphosphate (ATP), and muscle activation converts this chemical energy into mechanical force and work. Muscles generate force through a cascade of electrical and biochemical events, beginning with the release of acetylcholine by motor neuron synapses at the neuromuscular junction. This excitatory neurotransmitter binds temporarily with the muscle's cholineregic receptors, leading to depolarization of the postsynaptic membrane and mobilization of intracellular calcium ions. High intracellular calcium levels expose a site on the actin filaments to which myosin can attach. Once attached, the myosin molecules then reconfigure to force the actin and myosin filaments to slide relative to each other, which either shortens the muscle or generates force.

Whether these biomechanical events cause a force leading to movement or, alternatively, force without movement depends on the interaction of those muscles with their tendons, the skeleton, and the environment. As one consequence of the properties of actin and myosin, the length of a muscle affects the force (or tension) that it generates, a property known as the length–force (or length–tension) relation. As a muscle elongates, a given amount of activity generates a proportionately larger force. In addition, the properties of actin, myosin, and other structural elements also cause muscle fibers to behave approximately like a spring (technically termed viscoelasticity). Like metal springs, muscles have varying degrees of stiffness. For example, when pulled with a given amount of force, a spring made of thick, inflexible metal will increase its length much less than one made of thin, pliable metal. The former kind of spring is called stiff, and the latter is called compliant. Stiffness is defined as the ratio of force change to length change.

a. Rigor Mortis Muscle activation increases muscle stiffness, but so does death, which leads to rigor mortis. Death causes depolarization of the muscle fibers because the eventual depletion of ATP stops the sodium–potassium pump, upon which normal cell polarization depends, as well as the hydrolysis of ATP by myosin, upon which detachment from actin depends. In both rigor mortis and movement, attachment of myosin to actin causes stiffening of the muscle fibers.

2. Motor Neurons and Motor Units

In the ventral part of the spinal cord, motor neurons are organized into segregated motor pools, which innervate particular muscles. Alpha motor neurons send their axons from the spinal gray matter to terminate on extrafusal muscle fibers. Gamma motor neurons send their axons to intrafusal muscle fibers. Motor pools extend over two to four spinal segments, with medially situated motor pools innervating axial muscles (e.g., those of the neck and spine). Laterally situated motor pools project to limb muscles, with those contacting distal muscles located most laterally.

The term motor unit applies to a motor neuron and the muscle fibers it controls. Each motor neuron branches to innervate many muscle fibers, which receive input from only one motor neuron (except very early in development and in some disease states). The number of fibers in a motor unit varies according to function and within each muscle. Motor units that contribute to fine movements, such as those of the eye or the fingers, usually have a small number of muscle fibers. For example, motor units in the eye muscles consist of three to six muscle fibers. However, gastrocnemius, which forms the belly of the calf muscle, has thousands of muscle fibers per motor unit. Large motor neurons typically innervate more muscle fibers.

There are three basic types of motor units, each categorized by the speed with which it contracts upon electrical stimulation and its fatigability upon repeated stimulation. Fast, quickly fatiguing (FF) motor units have a short contraction time and produce a high twitch tension. However, with repeated stimulation the force they generate dissipates rapidly. Fast, fatigue-resistant (FR) motor units have an intermediate contraction time and can maintain force longer, whereas slow, nonfatiguing (S) motor units have a long contraction time and show little or no loss of force with repeated stimulation. FF motor units have large motor neurons, fast conducting, large-diameter axons, and muscle fibers of relatively large diameter. S motor units have the opposite characteristics. Muscles have various proportions of motor units: For example, S motor units make up nearly the entire diaphragm (one would not want to get tired of breathing), whereas gastrocnemius has a large proportion of FR and FF units (one certainly can get tired after a few strenuous jumps).

a. Poliomyelitis: A Disorder of Motor Neurons and Motor Units The poliovirus invades motor neurons, leaving adjacent nerve cells intact. Poliovirus receptors, located at the neuromuscular junction, allow the viruses to enter the motor neuron's axon, after which they migrate to its cell body. The infected cell either overcomes the virus or dies. If it dies, that motor unit is lost, and if the entire motor pool dies permanent paralysis results. However, some motor neurons usually survive. They develop new terminal axons that sprout to reinnervate "orphaned" muscle fibers. A single motor neuron may innervate up to 10 times the normal number of muscle fibers, restoring motor function. However, in a process called remodeling, which occurs in both healthy people and polio patients, motor units continually lose old sprouts and grow new ones. Decades after the onset of poliomyelitis, the enlarged motor units begin to break down, causing renewed weakness. According to one hypothesis, intense use of the relatively few remaining motor neurons causes this cell death.

3. Muscle Afferents

Certain sensory neurons innervate muscles and provide the CNS with information about muscle length and force. These and other sensory neurons have cell bodies in a dorsal root ganglion, with one axon projecting to sensory receptors in the periphery and another terminating in the CNS.

The diameter of muscle-afferent axons determines whether they belong to group I or group II. Group I, subdivided into groups Ia and Ib, has the larger fiber diameter and therefore faster transmission rates. Group Ia and II fibers innervate muscle spindles and are therefore called muscle spindle afferents. The term spindle refers to fine intrafusal muscle fibers that taper at the end and contain a fluid-filled capsule at the center. Muscle-afferent fibers wrap around muscular elements within the capsule. Group Ia fibers are termed primary muscle spindle afferents; group II fibers are called secondary muscle spindle afferents. Group Ib fibers innervate golgi tendon organs (GTOs), which are located in the transitional region between extrafusal muscle fibers and tendons. The role of muscle afferents in reflex responses is discussed in Section II.B.

B. Central Motor System

All levels of the CNS contribute to motor control, including the spinal cord, medulla, pons, midbrain, diencephalon, and telencephalon. The following sections survey the major components of the central motor system, beginning with the spinal cord and progressing up the neural axis to the telencephalon. Figure 1 shows some of the major components and projections of the central motor system. Notwithstanding the impression that this component-by-component description might convey, the various components of the motor system work as an integrated neural network, not as isolated motor "centers."

1. Spinal Cord

In addition to motor neurons, spinal components of the motor system include sensory pathways, the proprioceptive system, and central pattern generators (CPGs). Sensory afferents bring information to the CNS from the skin, joints, and muscles, and both cells and fiber tracts in the spinal cord relay that information to structures involved in motor control. For example, primary afferent neurons terminate on the dorsal column nuclei in the medulla, which relay that information to the thalamus through a fiber pathway called the medial lemniscus. In addition, an extensive and intricate system of intrinsic spinal cord neurons underlies a group of miscellaneous functions collectively termed proprioceptive. Proprioceptors are sensory transducers in muscles, tendons, and other internal tissues. However, the concept of the proprioceptive system extends beyond this definition to include a wide variety of interneurons that relay somatosensory signals locally and between segments in the spinal cord as well as carry descending motor commands, largely within the spinal cord. CPGs are neural networks that generate patterned, rhythmic movements, such as those involved in walking and running.

2. Brain Stem

The brain stem contains motor neurons that send their axons through certain cranial nerves, primarily to muscles of the tongue, face, and eyes. Like the spinal motor pools, many of these cranial motor nuclei receive direct input from sensory neurons and less direct influences from proprioceptive interneurons. Brain stem CPGs generate rhythmic movements, such as those underlying breathing and chewing.

Some parts of the brain stem interact with spinal CPGs and other components of the spinal motor system. One such region has been termed the midbrain locomotion region, which is thought to trigger the

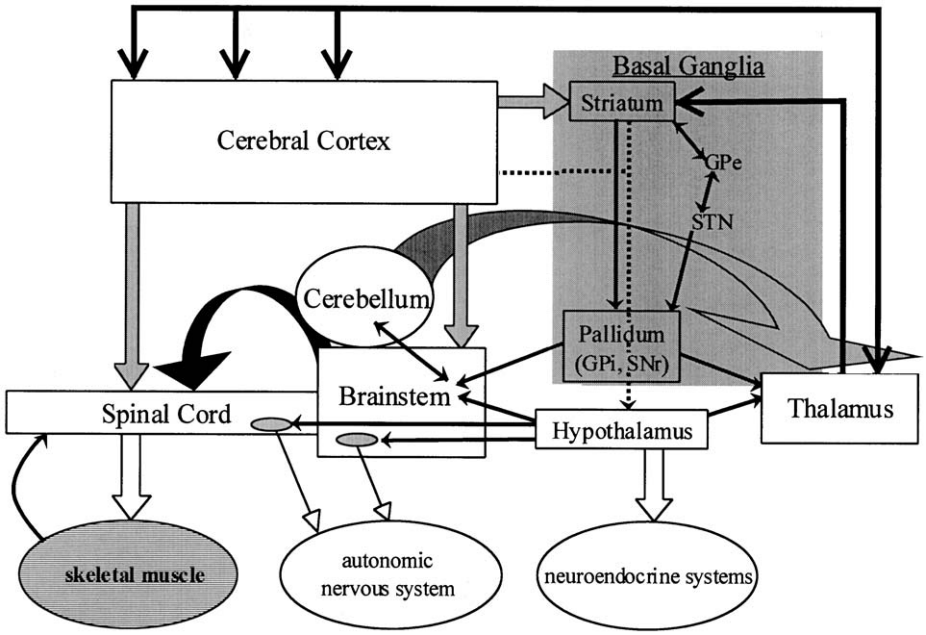


Figure 1 Major components of the motor system. Corticofugal projections are depicted as gray arrows. Projections to effectors are depicted as open arrows and arrowheads. The basal ganglia are contained within the gray box. Preganglionic autonomic motor nuclei are shown as stippled ovals. GPi, internal segment of globus pallidus; GPe, external segment of globus pallidus; STN, subthalamic nucleus; SNr, substantia nigra pars reticulata.

activity of spinal CPGs and thereby initiate locomotion. However, higher order networks, akin to CPGs but having more complex output patterns, have an important role in a number of instinctive behaviors, including aggressive posturing (a form of "body language") and inarticulate vocalization (such as crying, laughing, and screaming). Some of these networks are located in and near a midbrain structure called the periaqueductal gray.

a. Reticulospinal System Cells in the brain stem reticular formation that project to the spinal cord make up the reticulospinal system, which extends through medullary, pontine, and midbrain levels. The reticulospinal system performs a diverse set of functions, including the regulation of muscle tone, control of posture and locomotion, and integration of lower order motor signals with those emanating from the cerebellum and cerebral cortex. Different reticulospinal pathways exert influences on flexor versus extensor muscles and on proximal versus distal parts of the limb.

Part of the reticulospinal system serves as a fast transmission route to postural motor neurons and helps prevent movements from destabilizing balance. For example, when a person lifts a heavy object, the leg muscles need to stiffen before the elbow flexes. This postural adjustment prevents the object's weight from pulling the person off balance. The reticulospinal system activates leg muscles to stiffen them and help preserve balance. On the whole, while people are awake the reticulospinal system has a predominantly facilatory influence on motor pools. However, this effect changes dramatically during sleep. Then, reticulospinal neurons exert a strong inhibitory influence that, for example, prevents the performance of imagined actions during dreams.

Important influences over the reticulospinal system come from other systems, including vestibular afferents, which signal movements of the head and its orientation with respect to the earth's gravitational field, and the motor cortex, which provides information otherwise unavailable at brain stem levels. Through vestibulospinal projections, the vestibular system can contribute directly to various reflexes that adjust head position, posture, and limb movements. However, the vestibular afferents also provide inputs to the reticulospinal system. Consider the role of the reticulospinal system as a person runs through a field of obstacles. The signals conveyed by the reticulospinal

system to CPGs and spinal motor pools adjust posture and movement based primarily on vestibular and proprioceptive inputs. However, cortical and other higher order inputs supply the information needed for dynamic motor adjustments that allow people to step over and around visible obstacles.

b. Cerebellum and Red Nucleus The largest component of the brain stem motor system is the cerebellum. The medial cerebellum controls posture, whereas the lateral cerebellum participates more in voluntary movement. Accordingly, vestibular and propriospinal inputs predominate in the medial cerebellum, and inputs to the lateral cerebellum arise mainly from the cerebral cortex, relayed through mossy fibers originating in the basilar pontine nuclei. In addition, the cerebellum receives mossy fiber input from the red nucleus via the lateral reticular nucleus (which also has major spinal inputs) and from other sources. Mossy fibers terminate on the output nuclei of the cerebellum (the deep cerebellar nuclei) as well as on neurons in the cerebellar cortex. Another type of input, conveyed by climbing fibers originating in the inferior olivary complex, is thought to signal motor error or discoordination. These signals may play a central role in motor learning (see Section IV).

The output of the cerebellar cortex comes from GABAergic Purkinje cells, which inhibit neurons in the deep cerebellar nuclei and in one of the vestibular nuclei. The deep cerebellar nuclei send excitatory outputs to a variety of structures. Their largest projections terminate in the thalamus (Fig. 1), but other efferents reach the reticulospinal system, red nucleus, superior colliculus, and spinal cord. Cerebellar outputs to many of its targets are accompanied by return projections through a variety of direct and indirect pathways. One example is the cerebellar projection to the motor cortex (via the thalamus), which is returned by a cortical projection to the cerebellum (via the basilar pontine nuclei). Recurrent, excitatory circuits such as this are thought to form functional networks termed cortical–cerebellar modules.

The red nucleus plays an enigmatic role in motor control, especially in the human brain, but appears to be intimately related to cerebellar function. It receives a major projection from the deep cerebellar nuclei as well as from the motor cortex, and the largest part of the red nucleus (its parvocellular, or small-cell, component) projects predominantly to the inferior olivary complex, the source of cerebellar climbing fibers. The magnocellular (large-cell) red nucleus sends its axons directly to the spinal cord through the rubrospinal tract, which might be particularly important in stabilizing the limb by coactivating agonist and antagonist muscles. However, the magnocellular red nucleus is said to be relatively small in the human brain, which may reflect a dominant role of cortical motor control in our species.

c. Superior Colliculus The superior colliculus, although typically discussed in terms of eye-movement control, also has an important role in the control of head movements. Generally stated, its function involves the orientation of the retina and other receptors on the head, which the superior colliculus guides through its interaction with the reticulospinal system, premotor neurons in the brain stem reticular formation, and direct projections to the spinal cord (the tectospinal system).

3. Diencephalon

The hypothalamus and thalamus are the major parts of the motor system that lie within the diencephalon. The motor functions of the hypothalamus are discussed in Section II.A. The thalamus does not serve a primarily motor function when viewed as a whole. However, a major component of the thalamus receives projections from the cerebellum and basal ganglia, which play an important role in motor control. Two general regions of the thalamus receive these inputs—the ventroanterior and ventrolateral (VA/VL) nuclei. Anterior VA/VL receives basal ganglia projections; posterior VA/VL receives cerebellar input. Each part of VA/VL sends excitatory projections to the frontal cortex and receives excitatory projections from the same cortical areas. These reciprocal connections are thought to act as recurrent cortical–thalamic modules. Although the thalamic terminals from cerebellum and basal ganglia overlap very little, the thalamocortical components of these systems converge to influence most if not all motor areas jointly.

4. Telencephalon

Two large parts of the telencephalon have important roles in motor control: the cerebral cortex and the basal ganglia. Of course, the function of both structures extends beyond motor control, but this article addresses only that role.

a. Cerebral Cortex The number of functionally distinct motor cortical fields remains unknown. One

heuristically useful view of the frontal cortex divides it into three main parts: the primary motor cortex, a group of areas collectively known as nonprimary motor cortex, and the prefrontal cortex (Fig. 2). The first two components can be referred to collectively as the motor cortex, although this term is sometimes used as a synonym for primary motor cortex.

The primary motor cortex (abbreviated M1) corresponds approximately to Brodmann's area 4. It lies in the anterior bank of the central sulcus and contains a topographic representation of the musculature. Most textbooks depict this topography in the form of a homunculus (i.e., a projection of the body onto the cortical surface). Such pictures have validity only at the most superficial level. They correctly imply that the medial part of M1 contains the leg and foot representation, that a more lateral part includes the arm and hand representations, and that an even more lateral part has the face, tongue, and mouth representation. However, at any finer level of detail, the homunculus presents an inaccurate image of M1's organization. Instead, M1 consists of a mosaic of broadly overlapping muscle representations, with each part of the body represented repeatedly. The functions of M1 are discussed in Section III.

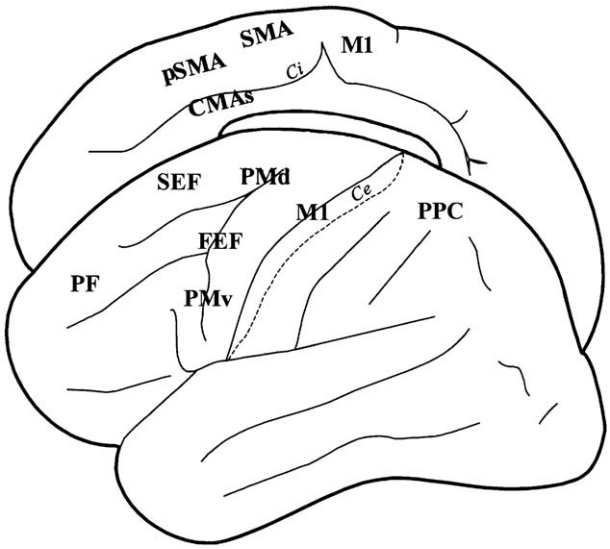


Figure 2 Motor areas of the human brain plotted onto a highly simplified sketch of the cerebral hemispheres. The lateral surface of the left hemisphere is shown below the medial surface of the right hemisphere. Rostral is to the left; dorsal is up. The dashed line indicates the fundus of the central sulcus, showing that the primary motor cortex is located mainly within the rostral (anterior) back of that sulcus. Ce, central sulcus; CMAs, cingulate motor areas; FEF, frontal eye field; M1, primary motor cortex; PF, prefrontal cortex; PMd, dorsal premotor cortex; PMv, ventral premotor cortex; PPC, posterior parietal cortex; SEF, supplementary eye field; pSMA, presupplementary motor area; SMA, supplementary motor area.

By current estimates, there are about a dozen nonprimary motor areas. Many of these fields occupy parts of Brodmann's area 6. However, parts of areas 8 and 24, the latter also known as anterior cingulate cortex, also contain nonprimary motor areas. A medial group of areas includes the well-known supplementary motor area (SMA), an area immediately anterior to it (the pre-SMA), and two or more cingulate motor areas. A lateral group of areas, often termed simply the premotor cortex (PM), can be divided into separate dorsal and ventral components, and these two divisions can be subdivided further into rostral and caudal parts. At least two eye-movement fields, the frontal eye field, and the supplementary eye field, have been identified. Finer subdivisions and combinations of these regions have been proposed and may have some validity. The functions of these areas are discussed in Sections III and V.B.

The primary and nonprimary motor areas project to the other components of the motor system, at virtually all levels of the neural axis (Fig. 1). The cortical output system, often called the corticofugal system, arises exclusively from layer 5 and includes a direct projection to the spinal cord (the corticospinal system). Some motor areas, especially M1, have monosynaptic excitatory connections with alpha motor neurons. Accordingly, the oxymoron "upper motor neuron" has been applied to M1's corticospinal neurons. However, M1 exerts its influence over genuine motor neurons in large measure through relatively indirect pathways, and its synapses on spinal interneurons vastly outnumber those on motor neurons. Corticospinal axons accumulate in large bundles traversing the internal capsule, cerebral peduncle, pyramidal tract, and corticospinal tract on the way from the telencephalon to the spinal cord. Because corticospinal axons run through the pyramidal tract, the term pyramidal motor system is sometimes applied to it. However, many fiber systems coexist within the pyramidal tract, such as neurons projecting to the dorsal column nuclei and other targets in the brain stem. Thus, the corticospinal and pyramidal motor systems do not correspond exactly. Additional outputs from the motor cortex include massive projections to the basal ganglia, the red nucleus, the cerebellum (via the basilar pontine nuclei), and the reticulospinal system.

b. Basal Ganglia Long considered part of the so-called extrapyramidal motor system, much of the

motor output from the basal ganglia depends, ultimately, on the pyramidal tract. Accordingly, the concept of an extrapyramidal motor system has been largely abandoned. The striatum is the basal ganglia's input structure, encompassing the putamen, caudate nucleus, nucleus accumbens, and other regions within the ventral forebrain. Likewise, the pallidum is the basal ganglia's output structure, including not only the globus pallidus but also the substantia nigra pars reticulata (Fig. 1) and additional parts of the ventral forebrain. The pallidum sends GABAergic inhibitory projections to the brain stem and thalamus. These thalamic nuclei connect to nearly the entire frontal lobe as well as additional areas, such as the inferior temporal contex and the posterior papietal cortex.

The major excitatory inputs to the striatum come from the cerebral cortex (including the hippocampus) and the intralaminar complex of thalamic nuclei. A major input arises from the dopaminergic cells of the midbrain, located in the substantia nigra pars compacta and the adjacent ventral tegmental area. Degeneration of the striatum results in Huntington's disease, whereas degeneration of the dopaminergic neurons causes Parkinson's disease.

Much attention has recently focused on the modular organization of cortical and basal ganglia interconnections. This trend accords with the recognition of other important modules in the motor system, such as CPGs, reflex circuits, and cortical–cerebellar modules. The cortical–basal ganglionic modules, often called "loops," consist of cortical, striatal, pallidal, and thalamic elements that form, at least in principle, a recurrent excitatory pathway. This circuit includes the so-called "direct pathway," striatal output neurons (technically, medium spiny neurons) that project directly to the pallidal projection neurons (Fig. 1), including those in the internal segment of the globus pallidus (GPi). Another part of the basal ganglia, the "indirect pathway," begins with projections from the striatum to the external segment of the globus pallidus (GPe). These neurons influence the subthalamic nucleus and pallidum, in turn (Fig. 1).

This sketch of the basal ganglia has heuristic value in understanding Parkinson's disease and other consequences of basal ganglia dysfunction. However, the reader should recognize that it represents a coarse oversimplification. Many other features of its anatomy are important to basal ganglia physiology. The subthalamic nucleus, for example, excites not just GPi but also GPe; the GPe sends inhibitory inputs "back" to the striatum; and the motor cortex sends a direct, excitatory projection to the subthalamic nucleus. Furthermore, in addition to dopaminergic inputs, serotonergic inputs arise from the raphe nucleus, and there are a large variety of intrinsic neurons using other neurotransmitters, including acetylcholine.

c. Parkinson's Disease According to one current view, dopamine acts to support activation of the direct pathway's striatal neurons but to suppress those of the indirect pathway. Accordingly, the direct pathway's inhibitory influence over GPi might wane in the absence of dopamine. This decrease in inhibition from the striatal direct-output pathway would lead to GPi neurons having greater activity and therefore greater inhibitory output to the thalamus. The indirect pathway may also contribute to greater inhibition of the thalamus. This increased inhibitory influence on recurrent cortical–thalamic modules may cause the slowing of movement (technically, bradykinesia) that is a symptom of Parkinson's disease.

II. WHAT THE MOTOR SYSTEM CONTROLS

The motor system controls innate behavior, which is encoded genetically; a wide variety of reflex responses, defined as output generated relatively directly by sensory inputs; and voluntary movements.

A. Instinctive Action

The hypothalamus plays a central role in the control of instinctive behaviors, such as those involved in locomotion, orientation, and reproduction, and in neuroendocrine function. Pools of neuroendocrine neurons in the periventricular parts of the hypothalamus secrete hormones into the vascular system. Their terminals in the neurohypophysis (posterior pituitary) release oxytocin and antidiuretic hormone (vasopressin). Although it might seem counterintuitive to consider such secretions as motor in function, they serve as a mechanism through which the CNS controls other parts of the body, just as alpha motor neurons do.

Different cells in the periventricular hypothalamus serve a less direct endrocrine control role. These neuroendocrine cells secrete higher order control hormones, often termed releasing factors, into the portal blood supply of the pituitary. The adenohypophysis (anterior pituitary) responds to these hormones,

such as corticotropin-releasing factor, which increase or decrease the secretion of pituitary hormones such as adrenocorticotropin. These hormones play crucial roles in regulating growth and in homeostatic and reproductive functions.

The hypothalamus also affects the body through its control of the autonomic nervous system (ANS). The hypothalamus sends descending projections to the autonomic motor nuclei in the spinal cord and brain stem. There, neurons send outputs to the peripheral ganglia of the two components of the ANS, the sympathetic and parasympathetic systems. Sympathetic motor neurons are located at thoracic and lumbar levels of the spinal cord. Parasympathetic motor neurons are found in the sacral spinal cord as well as in the brain stem, from which they give rise to parts of cranial nerves V (trigeminal), VII (facial), IX (glossopharyngeal), and X (vagus). Autonomic motor neurons, like alpha motor neurons, project to the periphery and release acetylcholine. In the sympathetic nervous system, most of these cholinergic influences terminate on ganglia relatively near the CNS. These sympathetic ganglia contain neurons that release noradrenaline, which induces a generalized arousal of the "fight, flight, and fright" variety. In the parasympathetic system, cholinergic motor neurons project to ganglia relatively far from the CNS, near their visceral targets. Thus, the ANS can address the targets of the parasympathetic system more specifically than it can for the sympathetic system. Most parasympathetic postganglionic synapses use acetylcholine as a transmitter, although many neuropeptides, such as vasoactive intestinal peptide, are released along with acetylcholine. The hypothalamus, through its control of the ANS, exerts an important influence over such motor functions as vasoconstriction, respiration, and heart rate.

However, the hypothalamus does not confine its motor functions to the ANS. Parts of the hypothalamus play an important role in ingestive behaviors (such as eating and drinking), defensive and agonistic behaviors (such as flight from danger and aggression), arousal and orientation, and social behaviors, including sexual and other behaviors involved in reproduction (e.g., rearing of progeny and other means by which genes are passed to future generations). Its motor roles thus include initiating complex action patterns, which include epigenetically expressed but genetically encoded motor programs often termed species-specific or species-typical behaviors. For example, aggressive displays such as snarling at or staring down an adversary might result from fear, intermale competition, or irritation, with each state and response mediated by different, partially overlapping networks in the hypothalamus. The hypothalamus effectuates many of these behaviors via projections to the thalamus and to various brain stem structures (e.g., the periaqueductal gray).

The several systems influenced by the hypothalamus—endocrine, autonomic, and instinctive—may seem dissimilar and unrelated. However, the hypothalamus (along with components of the amygdala) coordinates these aspects into a fully coordinated behavior. Imagine combat soldiers engaged in a stressful and dangerous situation. A variety of species-typical behaviors accompany such perilous situations, including heightened states of vigilance and arousal, which require high expenditures of energy. Parts of the medial hypothalamus detect inputs that signal such situations and, through their projections to the brain stem, release and intensify arousal and vigilance behaviors. Those hypothalamic regions also influence the periventricular hypothalamus to secrete corticotropin-releasing factor, which in turn induces the release and circulation of adrenocorticotropic hormone. This hormone stimulates the adrenal cortex to produce and release larger amounts of glucocorticoids. A parallel ANS control signal from the hypothalamus to the motor nucleus of the vagus promotes insulin secretion by the pancreas. Thus, the neuroendocrine system (through gluocorticoids) induces the liver to release more glucose at the same time as the ANS (through insulin) promotes the uptake of glucose into cells, especially muscle fibers. Together, the hypothalamus coordinates an adaptive response to a stressful state: the mobilization and utilization of stored nutrients to support high energy expenditure as well as the actions appropriate to that state.

B. Reflex Responses

The motor system controls a large number of reflexes, of which this article highlights only withdrawal and muscle-afferent reflexes.

1. Withdrawal Reflexes

Activation of cutaneous and deep receptors by a potentially damaging stimulus gives rise to an ipsilateral flexion reflex, accompanied by contralateral extension, called the crossed-extension reflex. A polysynaptic network in the spinal cord mediates this

response, which acts to retract the limb from the noxious stimulus (the flexion reflex) and enhance postural support, especially from the legs (the crossed-extension reflex).

2. Muscle-Afferent Reflexes

Two feedback systems, both involving muscle afferents, regulate force and muscle length through reflexes. Muscle spindle afferents transduce muscle length, whereas GTO afferents transduce muscle force.

a. Force Feedback GTOs lie in series with the extrafusal fibers and receive no motor innervation. They send force information to the spinal cord, where interneurons receive input from the brain that specifies the amount of force that a muscle should produce. If that muscle's force level exceeds this set point, the GTO inputs inhibit the alpha motor neurons innervating that muscle, which lowers the force produced unless some other mechanism cancels that signal.

b. Length Feedback In contrast to GTOs, muscle spindles lie in parallel with extrafusal fibers and have contractile elements that are activated by gamma motor neurons. Without efferent innervation, the muscle spindles would become slack when extrafusal fibers shorten. The spindle afferents would then become silent and the CNS would lose information about muscle length. Gamma motor neurons activate muscle spindles during contraction to maintain that information flow. During movement and steady posture, muscle spindle afferents sense muscle length with respect to a bias length set by their gamma motor neurons. When the gamma motor neurons have high levels of activity, the bias length is relatively short. If the muscle length exceeds this bias length, muscle spindle afferents increase their discharge rate and excite alpha motor neurons innervating the same muscle. Of course, this increase in activity tends to shorten the muscle, bringing it closer to the bias length. Generally, however, gamma motor neuron activity allows the CNS to control the sensitivity of muscle spindle afferents, which may play their most important role in regulating muscle stiffness.

Information regarding limb position does not depend entirely on muscle spindles. Other sources of input include cutaneous and joint capsule receptors, both of which contribute information about limb position and joint angle. In addition, group III and IV afferents also innervate the limbs but receive information primarily from deep receptors in the muscle and cutaneous receptors that appear to respond mainly to painful stimuli rather than force or limb position. Despite this diversity of receptors, the muscle spindles appear to be especially important for sensing muscle length, as demonstrated by the following experiment: Imagine a blindfolded person, seated with his or her elbows on a table and his or her forearm held in a vertical posture. If someone else moves one forearm, the blindfolded person can indicate the position of that arm by matching it with the other, free arm. People perform this task very accurately in normal circumstances. However, if vibration is applied to the belly of biceps, people consistently overestimate the angle of extension at the elbow. The explanation for this phenomenon involves muscle spindle afferents. Vibration provides a very powerful stimulus to muscle spindles, and the CNS wrongly interprets their increased discharge as reflecting a longer biceps muscle, which translates to increased extension at the elbow.

Muscle spindle afferents also mediate stretch reflexes, among which the monosynaptic stretch reflex is also known as the myotactic reflex or the knee-jerk response. For example, when one taps the skin surface over the knee's patellar tendon, this stretches the quadriceps muscle. Prior to the involvement of stretch reflexes, this increase in length makes the muscle generate more force through the length–tension relation. However, stretch also results in an elongation of the muscle spindles in the quadriceps, which in turn causes increased firing of the primary and secondary muscle spindle afferents. This sensory input excites alpha motor neurons and causes a knee-jerk within 15–20 msec. Particularly for the muscles of the arm, wrist, and fingers, a second pathway exists through which muscle stretch can activate motor neurons. This pathway, called the long-loop stretch reflex, also begins with muscle spindle afferents. Information from these afferents is transmitted to the thalamus and then to the somatosensory and motor cortex before returning to the spinal cord through the corticospinal projection. It takes 40–50 msec for the information to traverse this entire circuit. Although the long-loop reflex takes longer than the short-loop one, the brain can reprogram the long-loop response in a highly flexible manner. For example, if a stretch is expected, people can suppress the long-loop component of the reflex or choose to respond especially vigorously. The major role of stretch reflexes probably involves responses to an unexpected perturbation. If, during a movement, something suddenly displaces the arm, this input elicits a compensatory response from

both the short- and long-loop reflexes. These reflexes tend to change the activation levels of motor neurons in a manner that stabilizes and stiffens the limb as it moves along the desired trajectory. Reflexes may help the motor system overcome impediments that have never been experienced. If a person's goal includes producing the sound "pa," the lips must touch each other to achieve this goal. However, if an experimenter pulls on the lower lip, the motor system needs to produce more force than usual to make the lips contact each other (technically, occlusion). This response does not depend on experience with such perturbations; it occurs the first time the lip is pulled.

c. Disorders of Long-Loop Stretch Reflexes in Parkinson's Disease Normally, people can voluntarily suppress the long-loop component of the stretch reflex. In Parkinson's disease, however, patients appear to lose this ability. Regardless of the instructions given to them, stimulation of primary muscle spindle afferents produces a large response. In Huntington's disease, by contrast, there can be a complete loss of the long-loop stretch reflex. The relation between the basal ganglia and control of these reflexes remains unknown, but it probably involves its influence over the motor cortex. Recent research suggests that loss of the long-loop reflexes accounts for the jerky movements characteristic of Huntington's disease.

3. Role of Reflexes in Voluntary Movement

To what extent are reflexes used for generation of voluntary movements? Studies of patients with degeneration of large-fiber afferents (peripheral neuropathy) have led to the conclusion that reflexes play only an indirect role in volitional action. Such patients lose their stretch reflexes, limb position sense, and their ability to detect limb motion, but they can still make voluntary movements. In normal individuals, an electromyograph (EMG) pattern composed of three discrete bursts of activity characterizes the execution of a rapid, one-joint voluntary movement. First, the agonist muscle activates (AG1), followed by the antagonist muscle (ANT) and, finally, a second activation of the agonist muscle (AG2). AG1 accelerates the limb, ANT brakes the limb, and AG2 stabilizes the limb and dampens oscillations around the final position. When peripheral neuropathy patients make a rapid thumb flexion, the typical three-phase EMG pattern described previously (AG1, ANT, and AG2) appears in the normal manner. Therefore, the EMG pattern of voluntary movements appears to originate from descending motor commands to alpha motor neurons and does not depend on reflex loops. The motor command signal, however, may pass through some of the same neurons used in reflex loops, increasing the capacity of the system to integrate peripheral and central information.

C. Voluntary Movement

This article concentrates on the voluntary control of forelimb reaching movements. However, similar processes apply to other types of voluntary movements. These processes include generating torques on an articulated chain of limb segments, promoting stability through agonist–antagonist architecture, positioning end effectors, translating goals into plans for action, computing the joint rotations and patterns of force needed to implement those plans, and compensating for other forces.

1. Generating Torques on Articulated Limb Segments

A limb consists of a chain of articulated segments, with the muscles acting as the motors (technically, actuators) that control torques around those segments. Each segment of a limb can rotate with respect to the more proximal segment. The axis of rotation centers on the joint that connects the two segments, and muscles provide torques on that joint. For example, a person usually has to flex his or her elbow to lift a coffee cup and sip from it. Commands from the motor system reach the biceps muscle, activating it and producing force, which results in flexion torques on the elbow joint. As the elbow flexes, the resulting movement stretches the triceps, which in other circumstances would result in increased force output from the triceps because of the length–tension relation. This increased force would cause an extension torque on the elbow joint. For the hand to reach the mouth, flexion torques need to exceed extension torques. Thus, while sending the activating commands to the biceps to initiate the movement, the motor system usually sends an inhibitory command to the triceps' motor pool as well. This reduces some of the extension torque produced by the triceps, diminishing the resistance to the voluntary flexion that brings the cup to the mouth.

2. Producing Stability through Antagonistic Architecture

To hold a coffee cup steady requires a different approach: The motor system sends commands to activate both biceps and triceps. This coactivation results in both flexion and extension torques on the elbow joint. If the net torque, which is the sum of these two torques, equals zero, then the forearm will remain still. Why not simply shut down both muscles rather than waste the torques (and energy) to no effect? The answer has to do with limb stability. Consider what happens if an object suddenly hits a person's hand and causes the arm to flex at the elbow. During coactivation, the muscle that gets stretched because of the impact (triceps) will vigorously resist because intrinsic muscle stiffness increases with activation level. Without coactivation (i.e., with an inactive triceps), the impact would result in a much larger flexion of the elbow. Thus, coactivation promotes limb stability.

3. Positioning End Effectors

Although the final motor commands act on muscles to move limb segments, the goal of a movement often involves the positioning of an end-effector. For example, to sip from a coffee cup, most people attend to the cup and not to what the elbow is doing. Signatures provide another case in point. Everyone has a unique signature, and its distinctive character persists even if different joints and muscles perform the signing movement. This principle has been called motor equivalence. Its basis is that the pencil serves as an end effector regardless of which muscles move it. The preservation of unique elements in a signature indicates that the highest levels of the CNS represent voluntary movements not as a pattern of muscle activations but as a kinematic pattern, specifically the desired motion of an end effector.

Of course, signatures consist of complicated movement trajectories. In principle, an infinite number of possible hand trajectories can be made between two points—some straight, others curved to varying degrees. However, unless the goal includes a curved trajectory, people show a remarkable similarity in the movements that they produce in reaching from a given hand position to a target. The hand moves with a unimodal, smooth, and symmetric velocity over the time course of the action, and it takes a straight-line path. Even blind people show this feature in their arm movements. When an experimenter demonstrates the target to a blind person by moving his or her hand to the target position (later returning it to the original location), he or she makes straight and smooth reaching movements just like sighted people. This smoothness in hand trajectory contrasts sharply with the changes that occur in joint positions during the same movement. For example, consider a movement of the right arm that starts with the hand at the far left of the midline and reaches to a target at the far right. For most people, this movement would be a straight line in terms of hand position. However, examination of joint angles shows that the elbow initially flexes and then extends. Therefore, its velocity is not unimodal. Human arm movements generally appear simple when described in terms of hand positions and velocities but are complex when described in joint coordinates. This regularity remains when people perform movements with different end effectors. For example, movements remain smooth and simple when our hands hold a long stick. In this case, the end of the stick moves smoothly and in a straight line.

4. Translating Goals into Action

a. Inverse Kinematics The smooth and simple hand trajectory described previously represents a kinematic plan. Before that plan can be formulated, the motor system must estimate both current hand position and the direction and magnitude of the movement needed to reach the target. Estimation of current hand position is based on two sources: Vision and proprioception. Muscle afferents from the arm provide the information necessary for estimating the orientation of each limb segment relative to its proximal joint. If the motor system has this information, it can compute the hand position with respect to the body. The computation of limb position from a proximal, joint-coordinate-based system to a distal, hand-centered coordinate system is called forward kinematics. If someone moves a blindfolded person's hand, he or she still has a pretty good idea of that hand's location. This ability depends primarily on the computation of forward kinematics from the length sensors in the muscles. If the motor system knows the length of the limb's muscles, it knows the angles of its joints and, through forward kinematics, can compute the location of the hand. The inverse of this computation maps hand position to the joint angles that are appropriate for it. In order to move the hand to a desired position, the motor system needs information about what joint angles the muscles need to achieve in order to move the limb segments to that position. This computation is termed inverse kinematics. In other

words, if the motor system knows the desired hand position, it can compute the joint angles needed to put the hand in that position through the computation of inverse kinematics.

b. Inverse Dynamics In addition to computing the positions of the joints and the hand for a desired limb trajectory (kinematics), the motor system must estimate how much torque to produce on each joint (dynamics). Accordingly, the motor system must translate a desired motion of the end effector into a pattern of muscle activations. This does not imply that the brain calculates or represents the joint torques in absolute terms, joint by joint, but rather that the neural network must solve this problem to generate motor commands that will achieve the goal. Consider the torques needed to lift a full cup of coffee in contrast with those needed to lift an empty cup. Although the hand trajectories in the two cases may match perfectly, torques on the elbow will differ. Therefore, the motor system must take into account the weight of objects before it sends motor commands to the muscles. The computation that estimates the motion that will occur as a result of an applied force is called forward dynamics. The mass of objects held in the hand affects this computation: Activation of the biceps at a certain level will flex the elbow by a smaller amount for a full cup than for an empty cup. Forward dynamics consists of predicting the elbow angle after the biceps receives its activation command. The ability to predict the sensory consequences of motor commands relies on this computation. The inverse of this computation, called inverse dynamics, allows the motor system to transform the desired motion of the limb into the patterns of muscle activation that produce the torques required for the task.

However, in everyday life, even movements as simple as lifting a coffee cup can encounter impediments. When something disturbs arm movements (e.g., an unexpected change in the load on the hand), movements lose their smooth and regular character. However, provided that the perturbations have high predictability, with practice the movements again become straight in terms of hand trajectory. This convergence toward a straight, simple trajectory in hand coordinates (rather than joint coordinates) further supports the idea that the motor system plans movements in terms of the position of the hand and other end effectors rather than joint angles or patterns of muscle activity. In other words, the motor system plans in terms of goals rather than the components of movement. Motor learning and memory often underlies the ability to make smooth, straight movement despite external perturbations and the forces of each part of a limb acting on the others.

III. MECHANISMS OF MOTOR CONTROL

Neurophysiologists have only begun to understand the mechanisms of the motor system, and this section should be considered a highly provisional account of these mechanisms. Most of the relevant information comes from studies of neuronal activity in awake, behaving monkeys.

In generating the plan for a simple reaching movement, the initial problem involves kinematics, figuring out the current location of the end effector (often the hand), the location of the target, and perhaps the path between them. Both the premotor cortex (PM) and the posterior parietal cortex (PPC) appear to have key roles in solving this problem. The general locations of these cortical regions are depicted in Fig. 2.

A. Location of Targets and Initial Hand Position

PPC, in particular, plays an important role in determining the location of objects that could serve as the target of a reaching movement. Cells in one part of the PPC, the parietooccipital area (PO; approximately corresponding to areas V6 and V6A), respond to visual stimuli and, unlike most visual areas, their receptive fields have no bias toward the foveal representation. This characteristic suggests that the visual information PO processes relates to the control of movement rather than the analysis of an object's features. Some PPC neurons signal the presence of a target in retinal coordinates (i.e., the location of a target with respect to the fovea). These cells could be particularly important in controlling eye movements to that target, but they might also function to compute reaching, head, and eye movements within a single coordinate framework. One PPC region, the lateral intraparietal cortex, appears to be particularly important for eye movements, whereas a nearby region, the medial intraparietal cortex (MIP; also known as the partial reach region, PRR), plays a larger role in reaching movements. Other cells in PPC show an influence of "extraretinal" signals such as eye and head position. Some PO neurons, for example, appear to indicate target location with respect to the head (i.e., in head-centered coordinates). These signals mark the beginning stages of transforming visual information from a

receptor-based (retinal) coordinate frame into one more useful for reaching movements. For those movements, it is more useful to represent the target location relative to the end effector that will make the movement. Other parts of PPC play a larger role in determining the initial position of the end effector. One part of parietal area 5, for example, encodes hand position with respect to the shoulder. These and other parts of PPC operate in cooperation with the motor cortex, including both PM and M1.

Like the PPC, cells in the dorsal part of PM respond to a combination of signals relevant to voluntary reaching movement, including visuospatial and proprioceptive input, as well as inputs reflecting gaze direction, the location of objects in the environment, the orientation of spatial attention, and nonspatial visual information (such as color and form). Gaze effects show that the location of the target relative to eye position has some importance in motor control, perhaps for coordinated movements of eye and hand when people reach one place while looking at another. Attentional signals might be important when not all reaching targets can be foveated, as often is the case for a sequence of movements in a cluttered visual scene. In addition to these target-related signals, cells in dorsal PM (along with many in M1) are sensitive to the initial position of the hand, possibly through both proprioceptive and visual inputs. Similarly, cells in the ventral part of PM respond to both somatosensory and visual inputs as well as spatial acoustic signals. Whereas the PPC coordinate systems seem to reflect eye-, head-, and body-centered frames, ventral PM appears to be more specialized for the particular body-centered coordinate frame that is most useful for a given movement. In ventral PM, when a body part having a tactile receptive field moves, the visual receptive field moves in the same way: Visual receptive fields on the hand move with the hand, and those on the head move with the head. Therefore, it appears that these PM cells encode the potential targets of action relative to the body and update this map whenever the pertinent body part moves. Thus, PPC and the motor cortex derive much of the information needed for formulating a kinematic trajectory, including the starting position of the hand and the target of movement in each of several relevant coordinate frameworks.

B. Dynamics

At least two brain structures mediate the generation of force profiles that move the hand smoothly to a target: M1 and the cerebellum. The nonprimary motor cortex and basal ganglia appear to be less involved in this function.

It is likely that the network computing inverse dynamics includes the cerebellum. Damage to it results in movements that suggest an inability to compensate for the complex dynamics of multijoint reaching movements, including forces that arise due to interaction among limb segments (interaction torques). For example, the intact motor system approximates the inertia of the arm in programming activations of muscles, whereas cerebellar damage results in movements that suggest a deficit in this transformation. Normal individuals produce coordinated motion of the joints during reaching movements, whereas cerebellar patients produce movements that often consist of a sequence of single-joint motions. Neurons in the cerebellum have properties consistent with this view, including Purkinje cells that reflect movement velocity and the forces needed to achieve a certain trajectory.

M1 plays a role in limb dynamics as well, in part through its reciprocal interaction with the cerebellum. A century-long controversy has surrounded the question of whether M1 neurons encode primarily kinematics or dynamics. This problem, known informally as the muscle versus movement debate, consists of at least two parts. One part concerns how M1 addresses muscles. Some studies in the 1970s suggested that the M1 neurons address individual muscles, but refinements in research techniques have led to the understanding that individual neurons in M1 address multiple motor pools, usually of synergistic muscles. This view is consistent with the branching of corticospinal neurons to multiple motor pools and with the fact that they terminate principally on spinal interneurons rather than alpha motor neurons. These results support the ''movement'' side of the debate, but there is another issue. It involves the motor-control signal that M1 sends to the motor pools, and it yields a different answer—one more consistent with the ''muscle'' side of the debate.

A pure force signal in M1 neurons would support the view that they control muscles. Neurophysiologists have not completely settled this question, but three facts have been clearly established. First, when the hand generates a force that does not lead to movement (as happens when the hand pushes against an object so rigid that it does not move), cells in M1 have approximately the same patterns of activity as when the limb moves. Second, loads exerted on the hand can either assist a movement or oppose it. Neither kind of load changes the kinematics of movement appreciably,

but the forces involved in making the movement can differ greatly. Many cells in M1 differentiate between these two conditions, and they therefore appear to reflect limb dynamics. Third, the posture of the arm, which changes both joint angles and limb dynamics, affects the activity of M1 cells, even when the end effector path has nearly identical kinematics. The predominance of signals related to limb dynamics supports the muscle side of the debate and, therefore, a role for M1 in controlling limb dynamics. However, this view should not be taken to an extreme. There is evidence that limb kinetmatics is reflected by M1 cell activity, as well.

C. Kinematics

There is no evidence for a purely kinematic signal in M1. A purely kinematic signal would be invariant to loads on the limb or the initial posture of the arm. Such invariance has been observed in PPC (area 5), supporting the idea that it is mainly involved in movement planning in kinematic terms rather than movement execution in terms of limb dynamics. However, in M1 evidence points to a combination of kinematic and dynamic signals, as shown in the following experiment: Like people, monkeys can move their fingers upward by bending the wrist in different ways. If the palm is up, then the wrist can be flexed (i.e., moved in the palm's direction) to move the fingers upward. If the palm is down, then the wrist can be extended to achieve the same results. Some M1 neurons reflect only limb dynamics (i.e., flexion or extension), but others show an intermediate pattern of activity that takes into account both kinematics (end effector movement) and dynamics (muscle activity).

Large parts of nonprimary motor cortex, especially those on the lateral parts of the hemisphere, appear to function in the sensory guidance of movement at a kinematic level. For example, neurons in the dorsal and ventral PM have greater activity during visually guided movements than during memorized sequences, but they are not much affected by loads. The former observation points to a specialization for sensory guidance of movement, the latter to a specialization for kinematics. As noted previously, the activity of both PPC and PM neurons reflects the location of movement targets. However, this is not a purely sensory response in either PPC or PM. Instead, their activity reflects the motor significance of those signals at different levels. A term that has been used for motor significance in this sense is intention. For example, when monkeys indirectly move a spot on a video monitor by pushing or pulling on a joystick, researchers can distinguish neuronal signals related to the direction of spot movement from those reflecting the direction of limb movement. Many neurons in at least one part of PPC, MIP, signal the direction of hand motion, not spot motion. Also, they signal the location of movement-guiding spots only when the spot directs a reaching movement and not when it directs an eye movement. A population of cells in dorsal PM and the vast majority of cells in M1 do so as well. As another example, when a visual cue indicates the location of the next target of a reaching movement, but movement needs to be delayed until some future time, cells in both the PPC (area 5) and dorsal PM cortex signal the direction and amplitude of the planned movement. This "delay-period" activity begins about 100 msec after the cue appears and can continue for several seconds. PPC neurons do not distinguish between objects that will be the target of the next reaching movement and similar objects that indicate (e.g., by their color) that they will not be the subject of immediate action. Thus, PPC neurons signal potential movements or movement targets but not necessarily those that are currently planned. PPC neurons also reflect, in their activity, the expected benefit to be gained by moving toward a potential target. Neurons in dorsal PM, in contrast to PPC neurons, have delay-period activity only when the object will be the target of the upcoming movement. Thus, at the level of the nonprimary motor cortex, especially PM, neural activity appears to relate mainly to the implementation of near-term kinematic plans and other relatively high-level goals.

Cells in the basal ganglia also reflect the kinematics of reaching movements, such as direction and amplitude. However, the basal ganglia is unlikely to have a significant role in solving the inverse dynamics problem or computing the motor plan. Most basal ganglia activation or inactivation occurs too late to have a very large role in movement initiation or the planning that precedes it. Instead, the activity in basal ganglia develops at about the same time as muscle activity and continues during movement. Accordingly, it has been proposed that pallidal output plays a predominantly modulatory role. According to one hypothesis, motor-control signals depend, in part, on recurrent, mutually supporting activity in cortical–thalamic modules (or loops) that include motor cortex. Pallidal output may affect limb kinematics by facilitating or suppressing these recurrent circuits during an ongoing movement. It appears paradoxical that when the

motor parts of the basal ganglia's output projection are surgically destroyed, patients with Parkinson's disease can initiate movements more easily and move faster than before the surgery. Damage to motor-control structures usually causes rather than relieves motor dysfunction. However, the movement-modulation hypothesis resolves this paradox. Applying the brakes will slow a car, even though the brakes are not part of the movement-generation system. It is also thought that the basal ganglia function in context-dependent movement selection as well as in sequential and internally generated movements.

D. Goal Achievement

1. Limb Trajectory

M1 neurons reflect information about the direction, magnitude, and speed of the movement, in addition to postural signals. They have their greatest discharge rate for movements in one direction, with systematically less activity as the direction of movement diverges from that direction. They are therefore broadly tuned for movement direction, although this preferred direction can change with variations in starting hand position, the posture of the limb, and many other factors. On the assumption that these cells contribute to movements in their preferred direction, it is possible to compute a single vector, termed the population vector, representing the net contribution of a neuronal population. The M1 population vector anticipates the direction of limb movement for straight-line reaching movements (as well as for a variety of curved trajectories) by 30–120 msec.

Cells in the nonprimary motor cortex have properties similar to those of M1 cells but with some important differences. For example, cells in M1 are generally specific for the limb used, usually the limb contralateral to the hemisphere in which the cell is located. Cells in dorsal PM have activity broadly tuned for reaching direction like M1 neurons but have nearly the same directional preference for movements of the left hand or the right hand to the same visuospatial target. This finding shows that dorsal PM cells reflect the movement in terms of either visual targets or the trajectory of an end effector (in this case, a handle that the monkey moves). Evidence points to the former explanation. It appears that PM and M1 differ in the degree to which the visual inputs that guide a movement affect the activity of its neurons. Experiments have been done in which the monkey must follow a visual trajectory, but this visual input is projected directly to its eyes so that the relationship between vision and movement can be altered. Sometimes the visual target trajectory and the movement trajectory, perhaps an oval, are the same, but sometimes the hand trajectories must be circular to match an oval visual input or vice versa. PM populations reflect the visual target trajectory with more fidelity than the end effector trajectory, whereas M1 populations more closely reflect the movements of the end effector (the hand). This supports the idea that the motor system computes movement trajectories in terms of end effectors, with visual target trajectory predominant in PM and limb trajectory predominant in M1.

2. Compensation for External Perturbations

A part of the long-loop reflex, M1, receives somatosensory information that helps compensate for unexpected perturbations during movement. For example, when a hand movement is stopped in progress by an external force, M1 neurons that precede and accompany the movement with a burst of activity renew or increase that activity in response. This signal probably arises from the muscle spindles, which shorten in concert with the extrafusal fibers due to their input from gamma motor neurons. When the limb is stopped, both alpha and gamma motor neurons continue to discharge according to the motor plan. The extrafusual fibers build up force, but the muscle cannot shorten. The muscle spindle fibers continue to contract, which generates a signal comparable to that evoked by a muscle stretch. This signal is relayed to M1 and it causes the motor-control signal to be augmented. The resulting increase in activity serves to compensate, at least partially, for the perturbation, although it may not be adequate to overcome the impediment. This kind of mechanism could account for the achievement of goals upon the initial presentation of a particular perturbation.

IV. MOTOR MEMORY

A. Implicit and Explicit Memory Systems

The brain regions that store motor memories differ from those that store conscious memories. The former comprise an aspect of procedural memory or knowledge and the latter declarative memory or knowledge. Psychologists often refer to procedural knowledge as implicit memory and to declarative knowledge as

explicit memory. Some psychologists use the term "habit" interchangeably for procedural knowledge, but this usage should not be confused with its biological meaning, which involves instinctive behavior.

The idea that different brain structures underlie explicit versus implicit memory comes from observing the effects of brain damage. Damage to structures in and near the medial temporal lobe (MTL) results in loss of certain recently acquired information. Amnesic patients with MTL lesions can learn and retain skills such as mirror tracing, rotary pursuit, bimanual tracking, and compensation for complex forces applied to the limb during reaching movements. Despite this motor learning, the patients may not be able to recall the training episodes.

The distinction between an explicit memory system, which depends on the MTL, and an implicit motor memory system has several implications for motor control in the human brain. Voluntary actions have been defined as those that are learned, attended, and based on a comparison among alternatives. This awareness depends on the explicit memory system. Other actions, including but by no means limited to reflex movements, proceed without conscious awareness. Some subconscious movements bear obvious markings of this unawareness, such as the stretch reflex or the vestibuloocular reflex (VOR). The latter serves as a case in point. When people move their head left while looking at something, their eyes move equally fast and equally far in the opposite direction. They are probably aware of the object at the focus of attention. However, they cannot report anything about the motor memory that allows them to keep looking directly at that object, regardless of their head movements. Adjusting the VOR involves motor learning in the broadest sense but differs dramatically from that underlying voluntary movement. People cannot make VOR-like eye movements voluntarily. Movements such as the VOR and other reflexes can only be controlled implicitly.

Other movements that can be made without conscious awareness closely resemble voluntary actions. They can be guided either implicitly or explicitly. The best studied example of this phenomenon is termed blindsight. Normally, pointing to visible targets is accompanied by explicit knowledge of the action and the goal. However, some people with damage to the visual system can point to a visual stimulus while denying that they see it. Thus, some of the visuomotor networks remain functional even when the networks underlying visual perception fail. Phenomena such as blindsight have led to a distinction between CNS systems underlying vision for action (and therefore implicitly guided action) and those involved in vision for perception (which may or may not lead to explicitly guided action). The distinction between these two information processing systems does not depend on brain damage. Normal people can make finger movements that accurately match the size of objects they touch but nevertheless describe the size of those objects incorrectly due to visual illusions. People can also make accurate saccadic eye movements to fixate a visible target, although they report that the target moved in some different direction due to different kinds of illusions.

Which brain structures underlie motor memories, implicitly guided action, and procedural knowledge? To answer this question, it is useful to distinguish reflexes (such as for the VOR) from explicitly guided movements. Much of the information underlying the former is stored at the brain stem level, including the cerebellum. The cerebellum also has an important role in classical conditioning, through which sensory inputs are linked to stimuli that trigger reflex responses. Accordingly, it is clear that the cerebellum plays an important role in motor learning and memory at the reflex level. However, the situation is more complex for movements that are sometimes voluntary and under conscious control (explicitly guided action) but that also can become automatic and unattended as they become routine (implicitly guided action).

The oldest and most common idea holds that the MTL subserves explicit knowledge, whereas the basal ganglia underlies implicit knowledge. Patients with Parkinson's disease show particular deficits on tasks involving implicit estimation of event probabilities, for example, along with a wide variety of tasks involving motor skills. A newer idea holds that explicit knowledge is stored in prefrontal cortex and an associated part of the basal ganglia, along with the MTL, whereas some aspects of implicit knowledge are stored in motor cortex and their associated parts of the basal ganglia, along with the cerebellum. [The distinction between prefrontal and MTL function is that between long-term information storage of many months or years (prefrontal cortex) and intermediate-term storage of many weeks or months (MTL)]. Brain imaging studies, especially those that have explored the concept of attention to action, appear to be more consistent with the newer view. In motor learning, attention to one's actions and explicit knowledge of those actions (and outcomes) dominate the early part of training on a novel task. As a motor task becomes more automatic,

fewer attentive resources are engaged, and eventually it might be performed without awareness. Neuroimaging results have consistently shown increased blood flow, an indirect marker of synaptic activity, in the prefrontal cortex as subjects begin to perform motor tasks that involve learning a new skill. This increased activation generally declines to baseline as the task is extensively practiced and becomes automatic. Conversely, as a sequence or skill becomes more automatic, cerebellum (especially posterior parts), nonprimary motor cortex, and PPC show increases in activity. These findings support the hypothesis that prefrontal cortex–basal ganglionic modules subserve voluntary movement, whereas the motor cortex–basal ganglionic modules (along with the cerebellum and the PPC) underlie more automatic movements of the same kind.

Besides neuroimaging experiments, other evidence also implicates the cerebellum in motor learning. In a monkey trained to stabilize the wrist against an externally imposed load, for example, the cerebellum has been inactivated by cooling it. This manipulation eliminated the previously learned, predictive component of the muscle activity, which would oppose the imposed load. Furthermore, development of motor memory has been associated with an increase in the number of synapses onto the Purkinje cells in the cerebellum. Significant synaptic remodeling on Purkinje cells takes place within 1–4 hr after completion of initial training.

B. Internal Models

1. Acquisition

When a novice operator learns to control a novel mechanical system, the brain solves three types of computational problems: optimization of the task performance criteria to arrive at a kinematic plan (i.e., learning how the mechanical system should behave in order to accomplish the goals of the task), learning a model of the forward dynamics of the mechanical system (i.e., acquiring an ability to predict how the mechanical system will behave as a function of current input) and learning a model of the inverse dynamics of the controlled system (i.e., acquiring an ability to predict the inputs that should be provided to the mechanical system for a given desired change in state of that system). For example, consider the case of an operator acting on a joystick that controls the thrust produced by the motors of a remote underwater vehicle. The task involves moving the vehicle from point A to point B. The kinematic plan specifies a smooth vehicle trajectory. The forward model specifies how the vehicle will behave as the operator moves the joystick. The inverse model estimates how the operator should move the joystick so that the vehicle moves along the planned trajectory. An internal model (IM) is a blanket term used to describe the information contained in the solution to these three types of computational problems, and motor memory refers to the representation of IMs in the brain.

People use IMs in nearly every voluntary movement. For example, consider trying to lift an empty bottle of milk that has been painted white on the inside so that it appears full. The motor system will generate muscle activation patterns in the arm that provide the forces appropriate for lifting a full bottle, resulting in a flailing motion. This fact indicates that in programming the motor output to the muscles of the arm, the motor system uses certain visual characteristics of the object to predict and compensate for its mechanical dynamics. Learning IMs has been investigated in experiments in which a subject reaches to a target while holding the handle of a lightweight robotic arm. Disengagement of the robot's motors results in smooth and straight-line movements. Motor learning starts when the investigator programs a pattern of forces for the robotic arm to produce. These forces represent novel dynamics. When a person starts the training process, the computations that the brain performs in programming muscle activations do not take into account the novel forces, resulting in jerky hand movements. To compensate, initially people stiffen the entire limb through general coactivation of the muscles. This results in improved arm stability but serves only as a temporary and relatively ineffective strategy for responding to the perturbations. With training, stiffness returns to normal levels at the same time as the brain builds an IM of the novel dynamics. The motor output changes to specifically account for the additional forces. The development and use of a new IM can be demonstrated by turning off the robot's motors at the onset of movement. The resulting movement is a mirror image of that observed early in the training process. Therefore, the motor command that reaches the muscles includes a prediction by the IM of the forces required to overcome the imposed mechanical dynamics. The motor system retains this skill for months after the training session.

2. Consolidation

As one of its fundamental properties, the neural substrate of explicit memory gradually undergoes a

change, becoming more resistant to disruption. Newly acquired memories are more sensitive to new experiences and more susceptible to interference or brain injury. Posttraining treatments, including electric shocks, removal of key anatomical sites, or inhibition of protein synthesis, retard this progression and often result in loss of the recently acquired information. These interventions, however, have little effect on recall once a window of time has passed since acquisition. Consolidation refers to this time-dependent process. Does such a process occur in the storage of motor memories?

Until recently, little evidence supported the idea that time affected properties of motor memory. For example, it was found that whereas electroconvulsive shock interfered with conscious memories, it spared retention of a newly acquired skill (e.g., reading words in mirror image). Recent work, however, has found evidence for a temporal gradient in motor memory. The results suggest that within 4–6 hr after completion of training in a reaching task, functional properties of the motor memory gradually change. Subjects learned reaching movements in force pattern 1, leading to an internal model termed IM1, and then a second force pattern, termed force pattern 2, leading to an IM2. Their ability to learn IM2 depended on the time that had passed since completion of practice in force pattern 1. If only a short time had intervened (less than 4 hr), learning of IM2 was impaired compared to that of naive subjects. By the time 6 hr had elapsed, the interference caused by their training in force pattern 1 had subsided. This evidence points to the limited capacity of the system initially engaged in learning new skills. With time, the information maintained in this system fades, allowing the learning of additional skills. If this limited-capacity system serves a kind of intermediate-term memory stage for motor skills pending consolidation into a long-term memory store, then one would predict that its disruption might lead to an inability in long-term recall of motor memories. Indeed, learning IM2 within 4 hr after performing the movements in force pattern 1 results in an inability to recall IM1 in tests of long-term recall.

a. Apraxia as a Disorder of Motor Memory Apraxia occurs in numerous forms, of which at least two may represent disorders of motor memory. Ideomotor apraxia is characterized by kinematic errors made as patients copy movements or perform according to verbal instruction. Ideational apraxia is revealed by the incorrect miming of an action appropriate for a tool. In these patients, the ability to make the movements is relatively intact, but the motor system fails to produce the motor program that should be called up by the input, be it verbal or nonverbal instructions, as in ideomotor apraxia, or an object with a well-known use, as in ideational apraxia. The brain areas that fail in these situations have not been definitively established, but both PM and PPC appear to be involved. PM plays an important role in retrieval of a motor response as cued by visual or auditory stimuli, and PPC lesions cause some forms of apraxia. For example, a dentist with PPC damage suddenly complained that he "did not know" how to use a drill. This has been viewed as an example of a loss of a previously learned motor skill.

V. FLEXIBILITY IN MOTOR CONTROL

The ability to store motor memories enables the motor system to select a wide variety of movements in a highly flexible manner. People can select actions from a large repertoire of skills that have been previously learned, depending on context. In many contexts, achieving the goal of the motor system depends on movements made directly to targets, such as reaching to grasp an object. In others, more flexible relationships need to be established between objects and actions.

According to current thinking, different kinds of motor flexibility are afforded by different parts of the frontal cortex. Lateral nonprimary motor areas (such as dorsal PM) are thought to compute arbitrary mappings based on external (sensory) cues, whereas medial nonprimary motor areas (such as SMA) play an analogous role for internally generated actions, including memorized movement sequences. M1 is thought to enable a different sort of flexibility through the fractionation of motor synergies. Thus, it is commonly held that M1 functions mainly to permit motor "fractionation" (i.e., the independent control of muscle groups that usually work in concert). For example, the long muscles of the arm attach to several fingers to either flex or extend them. However, people can move their fingers one at a time.

A. Sensorimotor Mapping

Many reaching movements are made directly toward a target object. The term standard mapping applies to these kinds of movement. People often look at an object, orient attention toward it, and reach directly to

touch or grasp it. However, human behavior would be scarcely recognizable were it limited to standard mapping. People can also look in one direction while attending to a different place and reaching to a third, and everyone can use both spatial and nonspatial information to guide action.

1. Transformational Sensorimotor Mapping

In explaining the difference between standard and nonstandard mapping, prism adaptation serves as a case in point. When a diffracting prism distorts visual input, objects that lie directly in front of a person might appear to be 10° off to the right. When the person tries to reach to the object, he or she will reach ~10° too far to the right. However, the motor system can recognize this error and correct it through the process of motor adaptation, similar to that described previously for studying internal models (see Section IV.B). This behavior serves as a paradigmatic example of standard mapping: The motor system achieves the goal of directing the hand to the object, even though the prism distorts the object's location. In contrast, people can decide voluntarily to make a movement ~10° to the right of an object's location. This behavior serves as an example of nonstandard mapping. The motor system can produce an output that uses the spatial information in an object and transforms it according to some algorithm (such as 10° to the right) to produce a more flexible motor output than a system limited to standard mapping. This form of nonstandard mapping can be termed transformational mapping because the motor output depends on some transformed function of the spatial input.

2. Arbitrary Sensorimotor Mapping

Both of these spatially guided movements contrast with another form of nonstandard mapping termed arbitrary mapping. When nonspatial visual inputs, such as color, are used to determine the goals of action, an arbitrary mapping must be learned. Imagine a building with rooms of two colors: yellow rooms that require doorknobs to be twisted clockwise and blue rooms that require the opposite. Reaching toward the doorknob relies on standard mapping. Opening the doors requires arbitrary mapping, at least to do so reliably on the first attempt. Although this example is artificial, most signal- or symbol-guided behavior, including almost all language-guided behavior, depends on arbitrary mapping. Examples include stopping at a red light or at the sound of the word "stop."

B. Internal versus Sensorimotor Mapping

The idea that lateral nonprimary motor areas (such as PM) underlie external control of action, whereas medial areas (such as SMA) subserve internal control, has a link to related ideas concerning the motor functions of the basal ganglia and the cerebellum. It has been proposed that the cerebellum functions preferentially in externally guided action, whereas the basal ganglia controls mainly internally guided action. However, the functions of these structures are more complex than can be captured by such a simple dichotomy. The cerebellum participates in movements based on internal as well as external cues, and the basal ganglia plays a role in movements modulated by sensory as well as nonsensory information. Furthermore, contrary to earlier views, influences from both cerebellum and basal ganglia converge on both medial and lateral nonprimary motor areas.

How can these differing views be reconciled? Some progress can be made by recognizing functional subdivisions with both the basal ganglia and the cerebellum and by eschewing all-or-none dichotomies. Cells in caudal parts of one deep cerebellar nucleus, the dentate nucleus, have a preference for movements based on visual inputs (externally guided action) compared to kinematically similar movements generated from memory, whereas rostral parts have fewer cells with such preferences (i.e., they are relatively nonselective). Likewise, cells in the dorsal parts of the GPi have a strong preference for memorized sequences (internally guided action), whereas the ventral parts of GPi lack such selectivity. M1 and SMA receive their largest inputs (via the thalamus) from the less selective parts of the basal ganglia and cerebellum. Accordingly, M1 and SMA play a fairly general role in motor control and lack strong specializations for either internal or external control.

In contrast, PM receives its predominant inputs from the part of the dentate nucleus more selective for externally guided movements. Accordingly, PM has a large role in externally (often visually) guided action. For example, lesions to or inactivation of dorsal PM prevent monkeys from using color (or other nonspatial visual information) to choose an action.

The internally selective (dorsal) part of GPi preferentially influences the pre-SMA, which likewise appears to be specialized for internally guided action. Cells in pre-SMA have their greatest activity for internally guided movements, including movement sequences, and show other sequence- or order-specific patterns of activity. Temporary inactivation of the

pre-SMA disrupts the ability to produce a memorized (internally guided) sequence but not a visually triggered one. Brain imaging and brain lesion research also suggests the involvement of medial nonprimary motor areas in the learning of both limb- and eye-movement sequences and the "spontaneous" generation of action. In particular, pre-SMA, has been proposed to function in changing or updating motor plans based primarily on signals internal to the CNS.

Therefore, both medial and lateral parts of the motor cortex contain generalized as well as specialized components. Generalized parts include SMA (medially), and M1 (laterally). Specialized elements include the laterally situated PM, which maps sensory information onto motor outputs in a highly flexible manner, and the medially situated pre-SMA, which along with SMA maps a vast array of memorized and other nonsensory information onto motor output. A medial specialization for internal information accords with the role recently proposed for the rostral cingulate motor area in linking incentive magnitude to the selection of action, as occurs when one must choose an action based on some estimation of preferred outcome.

VI. EVOLUTION OF THE MOTOR SYSTEM

The motor system was born, not made, and many of the characteristics of the human motor system reflect its history. The CNS's chief function involves the acquisition of a behavioral repertoire, which can be stored both genetically and epigenetically, and the selection from that repertoire of the actions most likely to enhance an individual's fitness, in the inclusive sense of that term.

It seems likely that the ability to move in a goal-directed manner developed in an invertebrate ancestor of vertebrates. Many experts believe that the coordinated guidance of action represents the principal function of the CNS and comprises, in all likelihood, the crucial adaptive breakthrough made by the vertebrates and their ancestors. (Other animals made the same leap, but this article focuses on the human lineage.) The emphasis on motor control in evolution may seem paradoxical because, to us, complex and sophisticated cognition seems to be the principal characteristic of advanced animals, not motor control. However, in evolutionary history, long before our ancestors possessed the capacity for language, abstract reasoning, or highly general problem solving, they moved in relation to objects and places in their environment. It seems likely that the invertebrate ancestors of vertebrates (technically, cephalochordates) evolved from filter-feeding ancestors. They could swim in a coordinated manner, and they directed those actions with a brain and sense organs concentrated on the head, including paired eyes. The ability to do these things allowed our very distant ancestors to adopt a hunting life, one probably dominated by olfaction and vision. According to the fossil record, this revolutionary advance occurred about 500–550 million years ago. The human motor system, along with that of other vertebrates, has developed its capabilities by building on the original system. The motor system that people use to run, walk, talk, and manipulate objects evolved from the same one that originally evolved to control swimming and that was adapted to control flying, burrowing, and galloping in other species.

Thus, the motor system has a high degree of functional plasticity, at least as measured over geologic time. This plasticity of function contrasts with the conservatism of this basic morphological organization throughout vertebrate evolution. In the past, authorities have speculated that the brain evolved from caudal to rostral. That is, early comparative anatomists thought that the basal ganglia, for example, being located mainly in the telencephalon, evolved relatively recently in evolution. Anatomical names such as neostriatum testify to this notion. However, recent evidence from comparative morphology shows that the basal ganglia evolved very early in vertebrate evolution, probably at about the same time as the appearance of the first true vertebrates. Most of the remainder of the motor system also reflects our inheritance from our distant vertebrate ancestors. In addition to the basal ganglia, the superior colliculus, the red nucleus, the reticulospinal system, CPGs, and the alpha motor neurons retain their basic organizational patterns and functions from the earliest history of the vertebrate brain. The evolution of the cerebellum remains less certain, but it clearly emerged relatively early in the history of vertebrates as well.

What has changed in the past 500 million years or so? Of course, many of the basic components of the motor system have been elaborated and modified in many ways. In addition, an entirely new component has been added. In contrast to the ancient components of the motor system enumerated previously, the motor areas of neocortex have evolved much more recently. In a form that people can recognize as distinctively neocortical, a clear-cut motor cortex appears only in mammals. Thus, a separate motor cortex appears to have evolved only approximately 180 million years ago, a short time by evolutionary standards, especially

compared to the ancient lineage of the motor system's other main components. The relatively recent development of the motor cortex suggests that an important aspect of motor cortex function involves the modulation and control over subcortical motor systems, including the introduction of novel levels of flexibility in the motor system, such as transformational and arbitrary mapping, and the fractionation of relatively hardwired motor synergies.

See Also the Following Articles

APRAXIA • BIOFEEDBACK • BODY PERCEPTION DISORDERS • HAND MOVEMENTS • HYPOTHALAMUS • MEMORY, EXPLICIT AND IMPLICIT • MOTION PROCESSING • MOTOR CORTEX • MOTOR NEURON DISEASE • MOTOR SKILL • PARKINSON'S DISEASE • STRESS: HORMONAL AND NEURAL ASPECTS

Suggested Reading

Bhushan, N., and Shadmehr, R. (1999). Computational architecture of the human adaptive controller during learning of reaching movements in force fields. *Biol. Cybernetics* **81,** 39–60.

Bloedel, J. R., Ebner, T. J., and Wise, S. P. (Eds.) (1996). *The Acquisition of Motor Behavior in Vertebrates*. MIT Press, Cambridge, MA.

Goldberg, M. E., Eggers, H. M., and Gouras, P. (1999). The ocular motor system. In *Principles of Neural Science* (E. R. Kandel, J. H. Schwartz and T. M. Jessel, Eds.), 4th ed. Elsevier, New York.

Joseph, A. B., and Young, R. R. (Eds.) (1992). *Movement Disorders in Neurology and Neuropsychiatry*. Blackwell, Boston.

Leiguarda, R. C., and Marsden, C. D. (2000). Limb apraxias—Higher-order disorders of sensorimotor integration. *Brain* **123,** 860–879.

Loewy, A. D., and Spyer, K. M. (Eds.) (1990). *Central Regulation of Autonomic Functions*. Oxford Univ. Press, Oxford.

Middleton, F. A., and Strick, P. L. (2000). A revised neuroanatomy of frontal subcortical circuits. In *Frontal–Subcortical Circuits in Psychiatry and Neurology* (D. G. Lichter and J. L. Cummings, Eds.), Guilford, New York.

Milner, A. D., and Goodale, M. A. (1996). *The Visual Brain in Action*. Oxford Univ. Press, Oxford.

Passingham, R. E. (1993). *The Frontal Lobes in Voluntary Action*. Oxford Univ. Press, Oxford.

Porter, R., and Lemon, R. (1993). *Corticospinal Function and Voluntary Movement*. Clarendon, Oxford.

Stein, P. S. G., Grillner, S., Selverston, A. I., and Stuart, D. G. (Eds.) (1997). *Neurons, Networks, and Motor Behavior*. MIT Press, Cambridge, MA.

Motor Cortex

JON H. KAAS and IWONA STEPNIEWSKA
Vanderbilt University

GLOSSARY

basal ganglia A system of subcortical nuclei that receive cortical input and project to thalamic nuclear groups that project in turn to cortex. The basal ganglia thus contribute to a feedback loop for the control of movement as well as cognitive and often nonmotor functions. The globus pallidus and substantia nigra are the basal ganglia components that control motor cortical areas.

cerebellar nuclei Deep nuclei of the cerebellum that, together with the basal ganglia, control motor cortical areas via their projections to the motor thalamus.

cingulate motor cortex Motor cortical areas located in the cingulate sulcus and cingulate gyrus on the medial wall of the frontal lobe of primates.

motor cortex Cortex involved in control of movement. Often, it refers to only the primary motor field, M1, but premotor and other motor areas of frontal cortex may be included as well.

motor thalamus The part of the ventral thalamus that projects to motor cortex.

premotor cortex Nonprimary motor cortex on the dorsolateral brain surface, anterior to M1. Investigators currently describe this cortical region as having dorsal and ventral subdivisions.

supplementary motor area A motor area first described in the cortex of the medial wall of the frontal lobe of humans. It is sometimes called the second motor area, M2.

Motor cortex is the part of the neocortex of mammals that is devoted to evoking and controlling movements rather than receiving sensations and creating ideas or perceptions. Motor cortex is located in the posterior part of the frontal lobe or frontal cortex, just anterior to somatosensory cortex. Motor cortex includes a primary area, M1, where electrical stimulation of neurons evokes muscle contractions at low levels of current. M1 is characterized by large pyramidal or Betz cells and the lack of an obvious layer 4 of granular cells. Thus, M1 is referred to as agranular cortex and as area 4 of Brodmann's classical terminology. The pyramidal cells in M1 project via the pyramidal tract to motor neuron pools in the contralateral brain stem and spinal cord. A number of premotor areas are also involved in motor control. Muscle contractions can be evoked by electrically stimulating these premotor areas. Many of these areas connect to M1 and project to the spinal cord, thus directly influencing motor control.

I. HISTORY

The concept of a motor cortex, a specialized part of the neocortex of the forebrain devoted to causing movements, dates back over 130 years to the theories of the British neurologist, John Hughlings Jackson. This prominent clinician reasoned from the sequences of movements that spread from an initial site in patients with focal ("Jacksonian") epilepsy that body movements must be represented in part of neocortex in a systematic way. Jackson's theory of an organized motor cortex received substantial support in 1870 when two German investigators, Eduard Hitzig and G. Theodore Fritsch, electrically stimulated the surface of

the exposed neocortex in dogs and found that they could evoke muscle contractions on the contralateral side of the body. Soon, the British neurologist David Ferrier (1874) electrically stimulated the neocortex in monkeys and described a number of regions of cortex in which body and eye movements could be evoked. In further studies, Ferrier surgically removed the cortex where electrical stimulation evoked hand movements in monkeys, causing hand paralysis. By 1906, Charles Sherrington, a British neurophysiologist, had used electrical stimulation to map much of primary motor cortex in chimpanzees. Sherrington later received the Nobel prize. Although movements could be evoked from many regions of the brain, Sherrington stressed that movements could be evoked at the lowest stimulation thresholds from the cortex just anterior to the central sulcus in primates. This cortex was the "Betz cell" region containing large pyramidal cells first described in 1874 by the Russian anatomist Vladimir Betz, who considered the region motor in function. Later, the German neuroanatomist Korbinian Brodmann (1909) described a motor area (area 4) with giant pyramidal cells in humans and a number of other mammals. The first systematic description of the representation of movements in motor cortex of humans was by the German neurologist Otfried Foester in 1925. In the 1930s, the Canadian neurosurgeon Wilder Penfield stimulated motor cortex during surgery in patients with epilepsy and described the orderly way in which movements of body parts are represented in M1. Penfield illustrated the organization of M1 as a distorted little man, or homunculus, on the surface of cortex with the lips and the hand occupying much of M1. However, many of our current views on the organization of motor cortex were shaped by the detailed investigations of monkeys by Clinton Woolsey in the 1950s and 1960s. These studies established the basic organization of the primary motor area, M1, just anterior to somatosensory cortex. Within M1, a sequence of movements could be evoked from tail to foot, trunk, hand, face, and tongue when stimulating electrodes were moved from medial to lateral across precentral cortex.

Efforts were also made to distinguish premotor cortex from M1 and subdivide the premotor region into fields. The Betz cells of M1 were considered to be the major contributors to the pyramidal motor tract projecting to the spinal cord. Foester distinguished M1 as the cortex of the "pyramidal" system and more anterior regions of motor cortex as belonging to an "extrapyramidal" system. However, we now know that a number of areas contribute to the pyramidal corticospinal tract, and that only a minority of the axons are from the large Betz cells. The same areas also contribute to the extrapyramidal system, which indirectly projects to the spinal cord.

In approximately 1950, Penfield described a second orderly representation of movements in frontal cortex along the medial wall of the cerebral hemisphere in humans and monkeys. He referred to this motor region as the supplementary motor area (SMA). Woolsey soon described the SMA in more detail in monkeys, referring to this area as the second motor area, M2. Modern studies have greatly extended these early efforts, first by subdividing the dorsolateral premotor region into separate dorsal and ventral premotor areas and then by discovering evidence for more motor areas in cortex along the medial wall of the frontal lobe (Fig. 1).

Currently, 10–12 motor areas are thought to exist in primates. Fewer motor areas have been described in other mammals. Taxonomic groups of mammals apparently differ in the number of areas devoted to motor control. Marsupials may not have separate motor areas of the neocortex. Instead, somatosensory areas, which also contribute to motor functions in all mammals, seem to mediate the motor functions. Most mammals, however, have a M1 and at least one or more additional premotor areas. The SMA or M2 is likely to be a motor area common to most or all placental mammals.

II. PRIMARY MOTOR CORTEX

M1 occupies a broad band of cortex just anterior to somatosensory cortex (Fig. 1). In primates with a central sulcus, M1 is located just anterior to somatosensory areas 3a and 3b, on the anterior bank of the central sulcus, and extends onto the surface of most of the precentral gyrus. M1 closely corresponds to the region described by Brodmann as area 4 or the area of giant pyramidal cells (Betz cells). The term area 4 is commonly used as an alternate designation for M1. However, not all portions of M1 contain Betz cells. The lateral part of M1 is characterized by smaller size pyramidal cells. Thus, some definitions of the cortex comprising M1 may have included some of area 6. M1 is one of the thickest portions of cerebral cortex. M1 can be identified by its characteristic representation of body movements, unique histological appearance (most notably, a lack of an obvious layer 4 of granule cells and large layer 5 pyramidal cells), and consistent

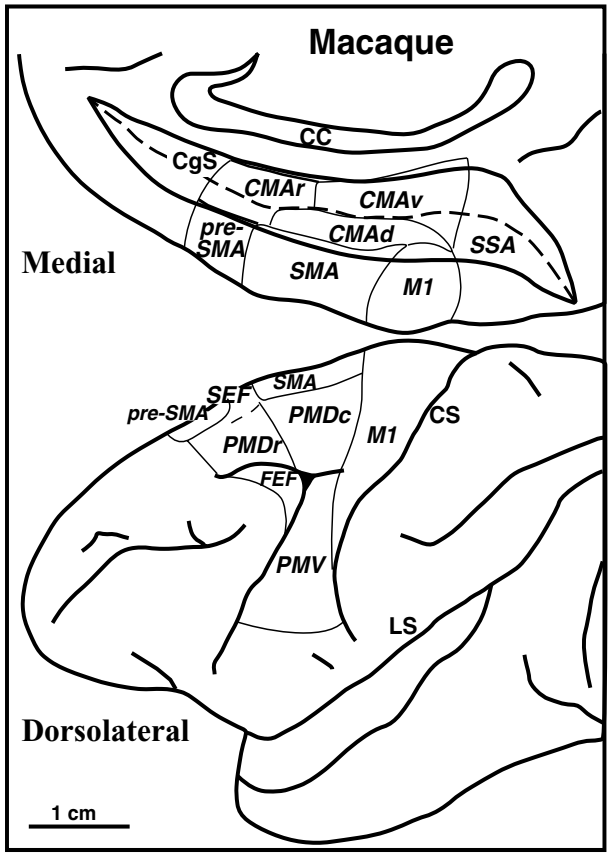


Figure 1 Motor areas of primates shown on a macaque monkey brain. (Bottom) A dorsolateral view of the frontal two-thirds of the brain. (Top) A view of cortex of the medial surface of the left cerebral hemisphere of the brain below shown rotated out so that ventral is at the top, allowing the motor areas to remain continuous. Primary motor cortex (M1) is on the precentral gyrus and it extends into the depths of the central sulcus (CS). Premotor cortex includes dorsal (PMD) and ventral (PMV) fields that appear to have rostral (PMDr) and caudal (PMDc) subdivisions (not shown for PMV). The supplementary motor area (SMA) was the first well-defined premotor area. Recently, a presupplementary motor area (pre-SMA) has been defined just rostral to SMA. On the medial surface of the cerebral hemisphere, two rostral (CMAr) and caudal (CMAc) cingulate motor areas have been defined. Some investigators divide CMAc into ventral (CMAv) and dorsal (CMAd) areas. CgS, cingulate sulcus; FEF, frontal eye field; LS, lateral sulcus; SEF, supplementary eye field; SSA, supplementary sensory area.

pattern of connections with somatosensory areas, premotor cortex, and the spinal cord.

The basic topographic organization of M1 has long been known from early studies in which many locations were electrically stimulated with electrodes placed on the surface of the cortex. Recently, the detailed pattern of how M1 represents movements has been determined with microelectrodes advanced into the cortex to allow the electrode tip to be placed close to the layer 5 pyramidal cells that project to the spinal cord. Thus, lower levels of current can be used to stimulate these output cells and cause movements, a technique known as microstimulation.

A long-standing debate has been whether M1 represents muscles or movements. Recordings from muscles, made while the cortex is microstimulated, indicate that single sites in M1 activate a number of pools of motor neurons in the spinal cord and brain stem so that each site of cortical stimulation activates several muscles that seem to be related in some form of movement. Although one muscle may be activated more strongly than others, the cortex seems to be subdivided into clusters or columns of neurons related to different types of movements or the generation of specifically directed forces. Cortical neurons typically activate several muscles that work together to cause a movement, while they inhibit outputs to muscles that produce the opposite movement. However, the specific muscles involved in making the movement are closely reflected in the activities of at least some neurons in M1.

Overall, movements of different body parts are represented in a systematic pattern across the mediolateral length of the M1 belt of cortex (Fig. 2). The most medial sites cause tail, toe, foot, and leg movements, whereas more lateral sites cause trunk movements, and a large lateral region is devoted to digit, wrist, and arm movements. The most lateral sites contribute to face and tongue movements. Because the projections of M1 via the pyramidal tract cross in the lower brain stem to terminate on motor neurons on the opposite side of the brain stem or spinal cord, these movements are all of muscles of the opposite side of the body. Parts of M1 that represent the trunk and proximal limbs of the two hemispheres are linked by connections through the corpus callosum. However, callosal connections are distinctly sparse or absent in M1 areas that represent the hands and feet. Thus, movements of hands or feet are controlled independently by a single M1. Although M1 is devoted to movements of the opposite side of the body, some coordination of movements of the two sides may occur via connections of one M1 with the other, especially for trunk and other proximal movements.

There have been various attempts to characterize the detailed local organization of M1. There does not seem to be a simple topographic pattern. Instead, the same or similar movements can be evoked from several nearby sites in the same region of M1, whereas

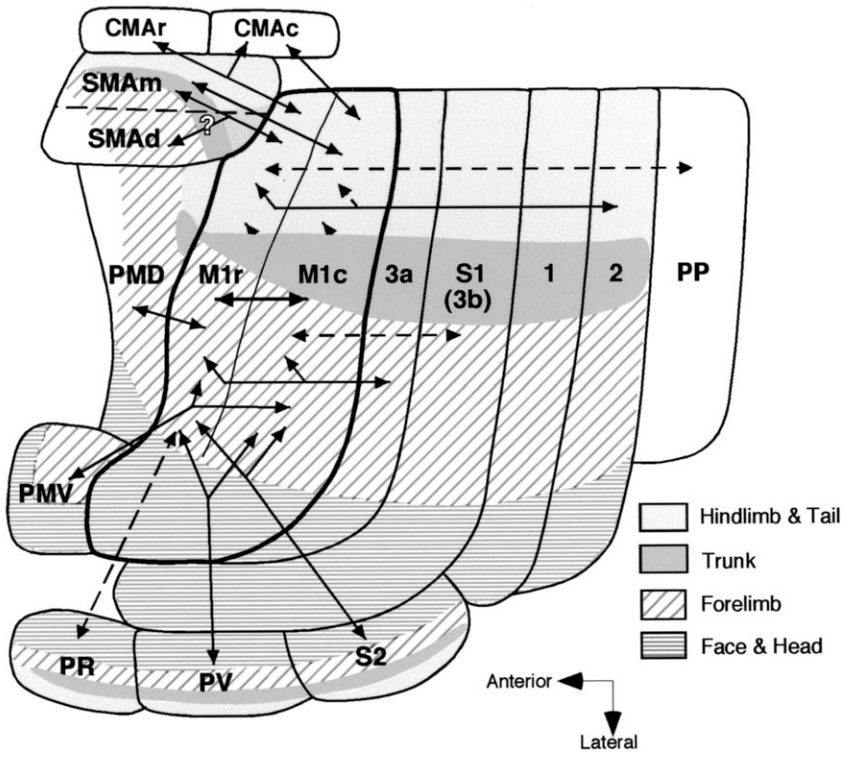


Figure 2 Some of the ipsilateral cortical connections of primary motor cortex, M1, in New World owl monkey. Most of these connections have also been reported for other species of monkeys. The organization of the sensory and motor representations in each area is indicated by shading. Major interconnections are indicated by thick arrows. Less dense connections are indicated by dashed arrows. Motor areas include M1 with rostral (M1r) and caudal (M1c) subdivisions, ventral (PMV) and dorsal (PMD) premotor cortex, supplementary motor area (SMA), presupplementary motor area (pre-SMA), and rostral (CMAr) and caudal (CMAc) divisions of cingulate motor cortex. Somatosensory areas include areas 3a, 3b, 1, and 2 of anterior parietal cortex; the second somatosensory area (S2); and the parietal ventral (PV) and the parietal rostral (PR) areas of lateral parietal cortex and posterior parietal cortex (PP).

adjoining sites evoke different but related movements. Thus, M1 appears to consist of a mosaic of small efferent zones or modules, with each module evoking a specific movement. Modules related to a given body region are grouped; within a region a specific type of module often occurs more than once. For example, modules moving different digits tend to be grouped; while bordering each other in different ways. These, in turn, border on modules for wrist, elbow, and even shoulder movements. Since neighboring modules might easily facilitate each other through local interconnections, this arrangement might allow different combinations of useful movements to be programmed in M1.

Stimulation of sites in M1 evokes small or restricted movements only when low levels of current are used. At higher levels of current, additional body parts become involved in more complex movement. For example, wrist movements may be added to digit movements. Thus, cortical sites may be involved in the control of a number of related muscles, provided that the recruitment of additional movements is not due to current spread. In addition, the local circuits in motor cortex are modifiable so that the consequences of electrical stimulation depend somewhat on ongoing sensory events or on recent activity patterns in M1. Conditioning by periods of electrical stimulation of M1, for example, can alter the precise nature of the response evoked by a subsequent stimulation. These effects are likely mediated by local connections within and between cortical modules in M1, allowing some flexibility in performance.

The details of the representations of movements in M1 can be modified by experience and training. The organization of M1 in humans can be evaluated in a noninvasive manner by stimulating motor cortex through the scalp by placing a magnetic coil over various parts of motor cortex and inducing current

flow in cortex (transcranial magnetic stimulation). Finger movements can be evoked over a larger region of cortex in highly skilled musicians who use their fingers to play instruments, suggesting that years of practice have changed the organization of their motor cortex so that more neurons in M1 are used to control finger movements. Changes in M1 with practice may also play a role in the recovery of motor control after partial lesions of motor cortex accompanying stroke. Monkeys with damage to parts of M1 that represent finger movements may reorganize so that remaining parts of M1 reinstate control over finger movements. Possibly as a result of cortical reorganization, the monkeys improve in their ability to skillfully use their fingers. Thus, the reinstatement of cortical representations may be responsible for the recovery of motor skills. Other motor areas may also be modified by experience as new motor skills are acquired.

M1 is considered to be a major area for initiating and controlling voluntary movements. Damage to M1 in humans may severely impair the ability to initiate movements, although much recovery occurs over time. Typically, lesions result in decreased use of the impaired contralateral limb or limbs and deficits in manual skill and speed of movement. Partial recovery of lost abilities may depend on the reorganization of remaining parts of M1 to devote more neurons to the needed functions as well as increased involvement of other intact parts of the motor system.

The discharge patterns of neurons in M1 reflect the important role that M1 plays in evoking movements. Neurons subserving a body part activate prior to the movement of that part, as one would expect if they function to initiate and control movements. For a particular movement, many neurons may be active, some much more than others. For a similar but slightly different movement, other neurons will be most active, but many of the same neurons will respond to a lesser degree. Neurons tend to be activated preferentially with some movements. A population of activated neurons has a different activity profile for each type of movement. The type and strength of a given movement depend on what neurons are active and on their levels of activity. In addition, some neurons become active in advance of intended movements. A sensory signal indicating that a movement should start soon initiates activity changes in such neurons.

The functions of M1 depend on its connections with other parts of the brain (Fig. 2). Motor behavior may be modified by sensory information, especially from receptors in the skin and muscles. Some of this information reaches M1 from areas of somatosensory cortex, especially areas 3a and 2 of anterior parietal cortex. These areas are activated by inputs from muscle spindle receptors. Feedback from muscle receptors during movements is very important in motor control. Other inputs come from higher order areas of somatosensory cortex such as the second somatosensory area, S2. In addition, many of the premotor areas are densely interconnected with M1. These premotor areas may influence motor behavior partly or largely through altering activity patterns in M1. Another source of sensory information is less direct. The cerebellum receives sensory inputs from muscle receptors and other receptors, integrates and adjusts this information, and sends projections to the specific region of thalamus, which in turn projects to M1 and several premotor areas. M1 sends feedback connections to the thalamus, somatosensory, and premotor cortical areas. The feedback connections from M1 to somatosensory areas of cortex alter the responses of neurons to sensory stimuli so that inputs related to self-initiated movements may be ignored. This feedback may produce corollary discharge in neurons leading to a sensation of movement. M1 also receives the broad cortical input indirectly by way of the internal segment of the globus pallidus, the portion of basal ganglia of the extrapyramidal system that projects to the motor thalamus. Thus, the processing loops in the motor network are complex. A major output of the M1 is via the pyramidal tract to motor neurons distributed in the brain stem and spinal cord. Most of the axons in the descending pyramidal tract cross from their side of origin to the other side in the lower brain stem as the pyramidal decussates. A small portion of these axons continue to descend in the ipsilateral spinal cord, where they may later cross to the opposite side or remain on the same side to influence ipsilateral motor neurons and ipsilateral body movements.

Recent anatomical and physiological data suggest that M1 has two functionally distinct subregions (Fig. 2). Cytoarchitectonic evidence for rostral (M1r) and caudal (M1c) subdivisions was found in both monkey and human brains. On the basis of differences in the cortical and subcortical connectivity of these two regions, as well as differences in the response properties of neurons, it has been proposed that these two divisions are specialized for the control of different stages of movement. M1r, which is less excitable than M1c and receives stronger projections from premotor cortex, may preferentially be involved in early stages of movements, including the postural adjustments required to maintain balance during reaching for an

object. These movements could be programmed by nonprimary motor areas and carried out with the aid of kinesthetic information reaching M1r from somatosensory areas. M1c, which is more excitable and receives diverse somatosensory inputs, may be more involved in the control of later stages of movement, such as grasping and holding objects, in which cutaneous feedback is especially important.

III. DORSOLATERAL PREMOTOR CORTEX

Premotor cortex is the cortex just anterior to M1 where electrical stimulation evokes movements, but at higher levels of current than in M1. Premotor cortex also projects much less densely to the spinal cord than M1. Premotor cortex is generally thinner than M1, but like M1 it has an indistinct or poorly developed layer 4 and thus is agranular or dysgranular in appearance. Unlike M1 (area 4), premotor cortex lacks a dense population of large pyramidal cells (Betz cells). Brodmann characterized most of this region as area 6. The frontal association cortex just rostral to area 6, generally termed prefrontal cortex, is recognized by the appearance of a distinct layer 4.

Premotor cortex is uniform in neither structure nor function, and it clearly contains several cortical fields. Most investigations distinguish at least two nonprimary motor fields: a dorsal premotor area (PMD) and a ventral premotor area (PMV) (Figs. 1–3). PMD is agranular in histological appearance, whereas PMV is dysgranular, having a thin granular layer 4. Both PMD and PMV lack a significant number of giant pyramidal cells, which are numerous in M1. The density of myelinated fibers, the density of reactivity of some metabolic enzymes, and the intensity of some types of immunoreactivity are distributed differently in PMD and PMV. Descending projections from both areas are also differently organized. These areas require different levels of electrical stimulation to evoke movements and have different patterns of movement representation.

PMD is located just anterior to M1 and lateral to the SMA. PMD extends laterally to the level of the frontal eye field. PMD is significantly less responsive to intracortical microstimualtion and has markedly higher thresholds for eliciting muscle contractions than M1. Electrical stimulation evokes hindlimb movements in medial PMD and forelimb movements in lateral PMD. Although movements of both the proximal and the distal limbs can be evoked from PMD, proximal limb movements (involving shoulder or shoulder and elbow) are most common. Neurons subserving face and eye movements may be located most laterally, just posterior to the frontal eye field.

PMD may be engaged in movements that require orienting the eyes, head, and trunk toward the target when limb movements are directed toward nearby objects and when posture is adjusted. Neurons in PMD project in such a way that those projecting to spinal cord motor pools for the leg are medial to neurons projecting to spinal cord motor pools for the arm. Anteriorly, PMD borders on prefrontal cortex, where neurons do not project to the spinal cord.

Neurons in PMD typically respond to cues that indicate that a motor response will soon be necessary. The neurons also fire during the waiting period and during the movement. As in M1, many neurons in PMD are activated during particular movements and during limb movements in a particular direction. Thus, PMD appears to have roles in the preparation and guidance of movements. Important sensory inputs to PMD originate in somatosensory and visuomotor

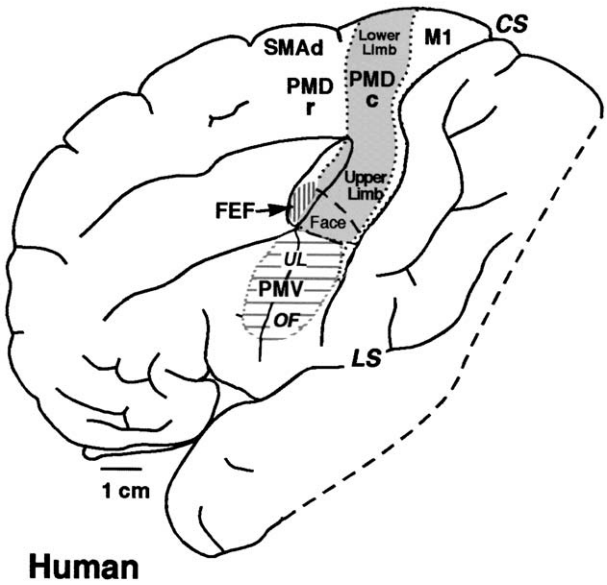


Figure 3 Proposed locations of primary motor cortex (M1) and premotor fields in the frontal lobe of humans. Rostral (r) and caudal (c) subdivisions of dorsal premotor cortex (PMD) are indicated, as are the ventral premotor area (PMV), the frontal eye field (FEF), and the dorsal division of supplementary motor area (SMAd). In PMV, the upper limb is represented medial to the face and mouth (UL and OF). Most of M1 is buried in the rostral bank of the central fissure (CS). Previous interpretations have extended M1 further onto the surface of the precentral gyrus to cortex we consider to be PMD.

areas of posterior parietal cortex. PMD is reciprocally connected with PMV, M1, the SMA, and cingulate cortex.

The rostral half of PMD is sometimes distinguished as a separate rostral field, PMDr, because current levels for evoked movements are higher and eye and face movements are sometimes evoked. Moreover, a scattering of Betz cells can be found in caudal PMD, whereas in the rostral PMD such large cells are essentially absent. Both caudal and rostral PMD regions are interconnected and the two areas may function in concert.

The PMV is located immediately anterior to the representation of face and tongue movements in M1. PMV can be distinguished from M1 and PMD by the presence of a thin but distinct internal granular layer 4. A layer 5 containing medium-sized pyramidal cells is prominent. Electrical stimulation of PMV results in forelimb, face, tongue, and eye movements. Hand movements are generally produced from cortex just dorsal to cortex subserving face movements. There is no clear evidence for a region devoted to hindlimb movements. Since the currents needed to evoke the movements can be as low as those for M1, PMV has been included within M1 by some authors.

PMV has been divided into separate rostral and caudal areas based on histological differences and different response properties of neurons or into medial and ventral parts based on differences in cytoarchitecture and connectivity. In humans, part of PMV of the left cerebral hemisphere (Fig. 3) may be specialized and is known as Broca's area, which is involved in the motor control of speech.

PMV appears to have a major role in visually guiding arm movements. Neurons in caudal PMV tend to respond to touch on the hand, arm, face, and mouth and to visual objects close to or approaching the hands, arm, face, and mouth. Neurons that respond to touch on the hand or arm have visual receptive fields that move with the hand or arm. Neurons in caudal PMV respond to appropriate sensory stimuli and during reaching and grasping movements. The more rostral part of PMV may be more involved in initiating grasping movements of objects such as bits of food and in bringing them to the mouth. Neurons in rostral PMV tend to respond while an object that is to be grasped, such as a desired food object, is observed and during reaching and grasping. Other neurons in rostral PMV, termed mirror neurons, respond during an action such as grasping a bit of food but also when another individual is observed performing the same action. These mirror neurons may have a role in learning from others by imitation or in mental rehearsal. Some investigators propose that learning by imitation plays an important role in language acquisition in humans. Thus, mirror neurons in ventral motor cortex, especially in the left hemisphere, may have language functions in humans. In monkeys, lesions of PMV disturb visually guided reaching movements. PMV's role in guiding reaching movements may depend on its inputs from visuomotor areas in posterior parietal cortex and somatosensory parts of the lateral sulcus. PMV also has connections with M1, PMD, SMA, and cingulate cortex as well as direct projections to the cervical spinal cord, in which motor neurons subserve arm and hand movements.

The dorsolateral premotor region also contains an area with visuomotor functions, the frontal eye field (FEF), which is located just rostral to the junction of PMD and PMV (Figs. 1 and 3). Electrical stimulation within the FEF evokes either saccadic or smooth-pursuit eye movements. The subregion for smooth-pursuit movements is caudal to the subregion for saccadic movements. The FEF receives inputs from a number of visuomotor areas in the temporal and parietal lobes, from adjoining regions of prefrontal cortex, and from frontal visuomotor regions such as the supplementary eye field. Important outputs are to the superior colliculus and to brain stem visuomotor nuclei.

IV. SUPPLEMENTARY MOTOR CORTEX

The SMA was first described in the cortex of the medial wall of the frontal lobe of humans more than 50 years ago. In monkeys, SMA extends onto the dorsal surface of the medial frontal lobe just rostral to M1 and medial to PMD (Fig. 1). The area is also known as the medial premotor cortex (MPC). Although SMA is usually considered to be a single area, it has been divided into medial (SMAm) and dorsal (SMAd) subdivisions in monkeys (Fig. 2). SMAm appears to be more densely connected to M1, whereas SMAd is somewhat more myelinated. More consistently, the SMA region has been divided into a SMA-proper (SMAd plus SMAm), located immediately rostral to the mesial sector of M1 representing foot, and a pre-SMA, extending toward prefrontal cortex just anterior to a SMA-proper. Both regions border the agranular cingulate cortex in the cingulate sulcus.

SMA was initially recognized when electrical stimulation of the region evoked movements in humans.

SMA has subsequently been described in monkeys and carnivores. The region known as M2 in rats is likely to be SMA. Thus, SMA may be an area that exists in many mammals. Stimulation of posterior SMA evokes movements of the contralateral leg, the middle portion is related to movements of the arm and hand, and the most anterior portion is devoted to the face. Current thresholds for evoking movements are generally higher in SMA than in M1. The internal organization of SMA is like a smaller version of M1. Like M1, similar movements may be elicited from more than one site in SMA. Matching bilateral movements are occasionally evoked from sites in SMA.

SMA is within the medial part of Brodmann's area 6. The cortex is agranular, without an obvious layer 4 of granule cells. The layer 5 pyramidal neurons are generally smaller than in adjoining M1. SMA has dense interconnections with M1 and dense projections to the motor neuron pools in the spinal cord. Inputs include those from visual, somatosensory, and auditory association and multimodal areas of the temporal and parietal lobes. SMA receives major inputs from regions of the thalamus with inputs from the internal segment of the globus pallidus, and a minor input from that part of the thalamus receiving inputs from the cerebellum.

Neurons in SMA respond to visual, auditory, and tactile stimuli when these stimuli are used as signals to start a movement or series of movements. When the signal indicates that a movement must be delayed for a short period of time, SMA neurons respond strongly during the delay period. This is partial evidence that SMA has a role in initiating and planning movements.

SMA was originally thought to have a significant role in coordinating movements of the two hands, partly because electrical stimulation of SMA neurons was once thought to commonly evoke ipsilateral movements. This has proven to be rare, however. During movements of one side of the body, only contralateral SMA shows significant activity. Although connections are strong between the SMA of each cerebral hemisphere, lesions of SMA do not produce deficits that suggest a major role in bimanual coordination. Instead, SMA appears to be important in the initiation of contralateral movements, motor programming, motor planning, and motor learning. After lesions, voluntary movements can be elicited by sensory cues and skilled movements can be executed. However, lesions impair the ability to self-initiate learned movements and sequences of movements when no sensory cue indicates the time for movement onset.

Pre-SMA is a small region of cortex just rostral to SMA. Movements are usually not evoked by electrical stimulation of pre-SMA unless much higher current levels are used and a series of current pulses of longer duration are employed. Pre-SMA lacks direct connections with M1, and it does not project or projects very weakly to motor neurons in the spinal cord. Pre-SMA has strong connections with prefrontal cortex, SMA, and cingulate motor areas. Neurons in pre-SMA are preferentially active prior to movement, and they respond less frequently to somatosensory stimuli but more frequently to visual stimuli. Pre-SMA is thought to be involved in the more cognitive aspects of motor behavior, possibly in updating motor plans and in the learning of new motor sequences.

In addition to somatomotor areas, a separate oculomotor area, the supplementary eye field (SEF), is found in the dorsomedial frontal cortex. SEF lies rostrolateral to the SMA and pre-SMA, and it is distinguishable from these two areas by its close relation to eye movements and its anatomical connections with cortical and subcortical oculomotor centers.

V. CINGULATE MOTOR CORTEX

Cingulate cortex along the medial wall of the cerebral hemisphere has traditionally been associated with the limbic system. The functionally significant subdivisions of cingulate cortex are not well established. Architectonically, the region is formed by two clearly distinct regions, areas 23 and 24 of Brodmann, with the more posterior granular area 23 having an obvious layer 4 and the more anterior agranular area 24 lacking a clear layer 4. Both areas are located ventral to the agranular cortex of pre-SMA, SMA, and M1. These two large cingulate regions have been further subdivided into smaller architectonic fields, such as area 24c and area 24d. Cingulate cortex is generally thought to be involved in emotional behavior. Recently, however, parts of cingulate cortex have been associated with the motor system.

There is good evidence for at least two motor fields in cingulate cortex (Fig. 2). Just ventral to pre-SMA along the medial surface of cortex, a rostral cingulate motor area, CMAr, occupies both the dorsal and the ventral banks of the cingulate cortex of monkeys. CMAr corresponds closely to part of area 24c of architectonic studies. Electrical stimulation of neurons in CMAr causes movements of contralateral body parts. Neurons related to hindlimb movements

are located caudally and neurons related to forelimb movements are located rostrally. The forelimb and hindlimb portions of CMAr project to the forelimb and hindlimb portions of M1 and the spinal cord. Neurons in CMAr are active before movements as well as during movements. Thus, these neurons are active during the intention to move, and they are not simply responding to the sensory consequence of movement.

A caudal cingulate motor area, CMAc occupies the dorsal and ventral banks of the cingulate sulcus just behind CMAr and ventral to SMA and M1. The cortex has a thin layer 4 and is within area 23c. Some investigators divide this region into dorsal and ventral fields (Fig. 1). CMAc projects to both M1 and the spinal cord. Electrical stimulation of CMAc evokes movements. The forelimb representation in CMAc appears to be rostroventral to the hindlimb representation. Other connections of the cingulate motor areas include dorsal and ventral premotor areas and regions of prefrontal cortex. The distribution of thalamocortical projections to CMAr and CMAc indicates that nuclei of the motor thalamus (ventral anterior nucleus and the caudal part of the mediodorsal nucleus) have projections that are greatest to area 24. Area 23 receives its principal thalamic inputs from the anterior nuclei, which are thought to participate in memory and emotion.

The functional roles of CMAr and CMAc are not well understood. The more caudal area, CMAc, with a thin layer 4 would appear to be more sensorimotor rather than purely motor in function. One proposal is that CMAr is involved in more executive functions and CMAc in more evaluative functions, such as monitoring the outcome of actions.

In addition to CMAr and CMAc, a more caudal portion of cingulate cortex projects to the spinal cord and possibly to M1. This cortex has a well-developed layer 4, and it is well within Brodmann's area 23. This portion of cingulate cortex, just ventral to somatosensory cortex, may include the region sometimes described as the supplementary sensory area (SSA; Fig. 1). The well-developed layer 4 suggests that this subdivision of cingulate cortex has important sensory functions, but the motor connections indicate a role in motor behavior as well.

In addition to CMAr and CMAc, electrical stimulation of more rostral parts of area 24 in monkeys evokes vocalizations. This region has been called a vocalization area. Vocalizations in monkeys are associated with emotional states, much like crying and laughing in humans.

VI. INFLUENCE OF THE CEREBELLUM AND THE BASAL GANGLIA ON THE MOTOR CORTICAL AREAS

Two main subcortical motor centers, the cerebellum and the basal ganglia, modulate the activity of cortical motor areas by providing feedback circuits. Their output nuclei [the cerebellar nuclei for the cerebellum and the internal segment of the globus pallidus (GPi) and the substantia nigra for the basal ganglia] receive indirect inputs from a variety of cortical somatosensory, motor, and associative areas and send their outputs, via the thalamus, back to the motor and premotor areas along with prefrontal, parietal, and other cortical areas. These motor areas receive cerebellar and pallidal inputs primarily through relays in the ventral lateral (VL) and ventral anterior (VA) regions of the thalamus. GPi sends a large number of GABA-containing fibers to the anterior region of the motor thalamus (VA nucleus) and the anterior part of the ventral lateral nucleus (VLa). Recent evidence shows that particular subsets of GPi neurons project to zones of VL that in turn project to individual cortical motor areas such as M1 or SMA, each of which contributes to the corticospinal tract. Thus, the multiple projections from GPi back to the cortex influence different cortical zones with different motor function. The deep cerebellar nuclei send excitatory input mostly to the posterior region of the motor thalamus (VL nuclei). The thalamus is not a simple relay of information from the cerebellum and globus pallidus to the cortex but, rather, a center for the integration these subcortical motor inputs with feedback from motor cortex.

The degree of overlap between the terminal fields of cerebellar and pallidal afferents in the thalamus is not well established. Although there is a broad consensus that the thalamic projections from the cerebellum and globus pallidus are mostly segregated, there is evidence for partial overlap in the motor thalamus. Cerebellar projections are concentrated posteriorly, whereas pallidal projections are concentrated anteriorly within VL nuclei. Pallidal and cerebellar afferents are highly segregated in the central core of these thalamic nuclei, but both inputs can be found within transitional zones between nuclei. In such regions the two inputs usually interdigitate, although occasionally they may overlap.

The motor cortical areas are reciprocally connected with the specific thalamic nuclear groups from which they receive their primary input, and these thalamocortical projections are critical for the functional output of the motor cortices. M1 has dense

topographically organized connections with the VL complex. The caudal part of M1 is most strongly connected with the posterior VL nucleus, a recipient of cerebellar projections, whereas the rostral part of M1 is most strongly connected with the anterior VL nucleus, a recipient of pallidal projections. Thus, both cerebellar and pallidal inputs to the thalamus are relayed to M1. Similarly, the SMA and pre-SMA areas receive both transthalamic inputs, although these inputs are largely derived from GPi. Finally, the dorsal and ventral premotor areas receive projections from anterior and posterior VL regions activated by both the pallidum and the cerebellum.

The functional contributions of the cerebellum and the basal ganglia to motor action are not clear, but both are necessary for smooth movements and posture. Degenerative diseases of the basal ganglia produce involuntary movements, paucity of movement, and abnormalities in posture. Damage to the cerebellum produces characteristic loss of coordination and accuracy of limb movements. Damage to either region may result in tremor.

VII. MOTOR FUNCTIONS OF SOMATOSENSORY CORTEX

The somatosensory areas of the parietal lobe make a substantial contribution to the pyramidal tract, and it is not surprising that movements can be evoked by electrically stimulating sites in these areas. Somatosensory cortex of humans and other primates includes four architectonic fields in anterior parietal cortex: areas 3a, 3b, 1, and 2, after Brodmann (Fig. 2). All these fields were once considered to be parts of a single primary somatosensory area, S1, but we now know that each architectonic field corresponds to a separate representation of body receptors and only area 3b corresponds to S1 of other mammals, such as rats and cats. Electrical stimulation of these fields and other higher-order fields such as the second somatosensory area, S2, of the lateral sulcus evoke movements related to the body part providing sensory inputs to the representation, but the levels of current needed to evoke movements are substantially higher than those for M1. The exception is area 3a, in which the thresholds for evoked movements are only slightly higher than those needed for M1. Area 3a may be especially involved in motor functions since it is activated by muscle spindle receptors, and these receptors provide feedback about ongoing movements. In addition, area 3a projects strongly to M1. Area 3a contains many pyramidal cells projecting to the spinal cord. Some investigators have described area 3a as having histological features of both sensory and motor cortices. S2 also projects densely to the spinal cord and to M1. Although some of the motor functions of somatosensory cortex might be mediated through these connections with M1 in intact primates, movements can still be evoked by electrically stimulating somatosensory cortex after complete lesions of M1. Also, as already noted, opossums and other marsupials do not appear to have any separate motor areas, but they have somatosensory areas that are sensorimotor in function.

VIII. CONCLUSIONS

Descriptions of motor cortex often include only the primary motor area, M1. However, humans and other primates also have a number of other cortical areas that are involved in the generation of motor behavior. Recently, evidence has been presented that at least 11 motor areas can be distinguished (9 somatomotor and 2 oculomotor). Each area is architectonically different and has a unique combination of connectivity to subcortical nuclei and cortical areas. The functional relationship of these areas involves both hierarchical and parallel components. Whereas the primary motor cortex is involved mainly in the execution of voluntary movements, nonprimary motor areas, including SMA, PMD, PMV, and cingulate areas, are engaged in higher order motor processing, such as preparation, and programming of sequences of movements and coordination of bilateral and other complex movements. These areas make up a second level of motor fields that interact with M1 and act in parallel with M1 by projecting to the motor neuron pools in the spinal cord and brain stem. By projecting to the spinal cord, each of these areas has the potential to influence the generation and control of movement independently, even when M1 is lesioned. A pre-SMA and possibly other fields represent a third level of areas with motor function. These fields interact with areas of the second level but not directly with M1 and not in parallel with M1. Areas of somatosensory cortex, especially areas 3a and S2, have motor as well as sensory functions. These cortical areas interact as part of a complex motor control network that also involves the motor thalamus, the substantia nigra, the cerebellum, and the basal ganglia. The motor thalamus receives cerebellar,

pallidal, and nigral inputs. Most motor behaviors probably involve the majority of cortical areas, and they likely have overlapping functions. However, there is evidence for specialization of fields so that different impairments emerge from lesions of different areas. There is much yet to learn about motor cortex, especially how the populations of neurons in each area contribute to specific types of motor behavior.

See Also the Following Articles

BASAL GANGLIA • CEREBELLUM • CINGULATE CORTEX • FRONTAL LOBE • MOTOR CONTROL • MOTOR NEURON DISEASE • MOTOR SKILL • NEOCORTEX

Suggested Reading

Georgopoulos, A. P. L. (1995). Current issues in directional motor control. *Trends Neurosci.* **18,** 506–510.

Kakei, S., Hoffman, D. S., and Strick, P. L. (1999). Muscle and movement representations in the primary motor cortex. *Science* **285,** 2136–2139.

Kalaska, J. F., Scott, S. H., Cisek, P., and Sergio, L. E. (1997). Cortical control of reaching movements. *Curr. Opin. Neurobiol.* **7,** 849–859.

Nudo, R. J. (1999). Recovery after damage to motor cortical areas. *Curr. Opin. Neurobiol.* **9,** 740–747.

Picard, N., and Strick, P. L. (1996). Motor areas of the medial wall: A review of their location and functional activation. *Cerebral Cortex* **6,** 342–353.

Preuss, T. M., Stepniewska, I., and Kaas, J. H. (1996). Movement representation in the dorsal and ventral premotor areas of owl monkeys: A microstimulation study. *J. Comp. Neurol.* **371,** 649–676.

Rizolatti, G., Fogassi, L., and Gallese, V. (2000). Cortical mechanisms subserving object grasping and action recognition; A new view on cortical motor function. In *The New Cognitive Neurosciences* (M. S. Gazzaniga, Ed.), p. 539. MIT Press, Cambridge, MA.

Stepniewska, I., Preuss, T. M., and Kaas, J. H. (1994). Architectonics, somatotopic organization, and ipsilateral cortical connections of the primary motor area (MI) of owl monkeys. *J. Comp. Neurol.* **330,** 238–271.

Tanji, J., and Mushiake, H. (1996). Comparison of neuronal activity in the supplementary motor area and primary motor cortex. *Cognitive Brain Res.* **3,** 143–150.

Wise, S. P., Boussaoud, D., Johnson, P. B., and Caminiti, R. (1997). Premotor and parietal cortex: Cortico-cortical connectivity and combinatorial computations. *Annu. Rev. Neurosci.* **20,** 25–42.

Wu, C., Bichot, N. P., and Kaas, J. H. (2000). Converging evidence from microstimulation, architecture and connections for multiple motor areas in the frontal and cingulate cortex of prosimian primates. *J. Comp. Neurol.* **423,** 140–177.

Motor Neuron Disease

PHILIP C. WONG, DONALD L. PRICE, and JEFFERY ROTHSTEIN
Johns Hopkins University School of Medicine

The motor neuron diseases, an etiologically heterogeneous group of disorders, are manifested by weakness (muscle atrophy) and/or spastic paralysis, reflecting the selective involvement of lower motor neurons and/or upper motor neurons, respectively. Amyotrophic lateral sclerosis, the most common adult-onset motor neuron disease, involves both lower and upper motor neurons. Spinobulbar muscular atrophy (or Kennedy's disease) and spinal muscular atrophy selectively affect lower motor neurons, whereas hereditary spastic paraplegia predominently involves upper motor neurons. During the past several years, significant progress has been made in identifying some of the genes/chromosomal loci associated with these illnesses. We briefly describe the clinical features/neuropathology of several of these diseases and then discuss recent progress in understanding the genetics and molecular/cell biology of these diseases and the status of investigations of *in vivo* and *in vitro* models relevant to some of these illnesses.

I. AMYOTROPHIC LATERAL SCLEROSIS AND FAMILIAL AMYOTROPHIC LATERAL SCLEROSIS

A. Clinical Features/Neuropathology

With a worldwide prevalence of $\sim$5 per 100,000, amyotropic lateral sclerosis (ALS) usually occurs in mid to late life and is manifest as progressive muscle weakness accompanied by hyperreflexia and spasticity associated with fibrillations, fasciculations, and giant polyphasic potentials. Muscle biopsies demonstrate denervation and muscle atrophy, and patients usually die of intercurrent illnesses. Muscle atrophy and weakness reflect selective degeneration of large motor neurons of the brain stem and spinal cord; spasticity, hyperreflexia, and extensor plantar signs are attributable to lesions of upper motor neurons.

Lower motor neurons show several abnormalities: the presence of phosphorylated neurofilament and ubiquitin immunoreactivities in inclusions within cell bodies, Lewy body-like intracytoplasmic inclusions, NF swellings of proximal axons, fragmentation of the Golgi, and attenuation of dendrites. In some cases of superoxide dismutase-1 (SOD1)-linked familial ALS (FALS), intracytoplasmic inclusions contain SOD1, ubiquitin, and neurofilament immunoreactivities, and neurofilaments accumulate in cell bodies and neuronal processes, particularly proximal axons. The roles of these inclusions are unclear. The presence of ubiquitin immunoreactivity suggests that these inclusions contain proteins destined for degradation, presumably in part via the proteosome. Whether components within these aggregates sequester essential molecules needed for the biology of motor neurons or whether the

malfolded protein can perform aberrant catalytic reactions is unknown.

Motor nerves show axonal atrophy and Wallerian degeneration. Eventually, the numbers of motor neurons in brain stem nuclei and spinal cord are reduced and there is a loss of large pyramidal neurons in motor cortex. These lesions are accompanied by degeneration of axons in peripheral motor nerves and corticospinal tracts, respectively, and denervation of target fields of these axons. Although the mechanisms of neuronal death in ALS are poorly characterized, recent studies have documented that motor neuron degeneration structurally resembles the process of apoptosis. In ALS motor cortex and spinal cord anterior horn, there is evidence of DNA fragmentation, caspase-3 activation, and redistribution of cell death proteins such as Bcl-2, Bax, and Bak.

A variety of mechanisms, including excitotoxicity, copper toxicity, oxidation/nitration-mediated damage, protein misfolding, aggregation of critical components, apoptosis, calcium-mediated processes, and alterations in the biology of neurofilaments are among the processes that have been suggested to play roles in ALS and FALS.

B. Genetics of ALS

Several gene/loci associated with ALS have been identified (Table I).

1. SOD1 Gene

Approximately 5–10% of ALS cases are familial, and in almost all cases inheritance exhibits an autosomal-dominant pattern. Approximately 15–20% of patients with autosomal-dominant FALS have missense point mutations in the gene that encodes cytosolic Cu/Zn SOD1 (Table I), the enzyme that catalyzes the conversion of $\cdot O_2$ to O_2 and H_2O_2. To date, >70 different mutations have been identified in the SOD1 gene. There is clear evidence of allelic heterogeneity, with phenotypes associated with different mutations sometimes showing differences; for example, cases of A4V SOD1 FALS have limited involvement of the corticospinal tract, whereas other SOD1 mutations are associated with a more classic ALS syndrome.

2. Other Genes/Loci Relevant to ALS

A variety of chromosomal loci have been linked to autosomal-dominant, autosomal-recessive, and X-linked-dominant forms of ALS. Notably, investigators have described a large family with autosomal-dominant juvenile ALS (JALS) linked to a locus at the 9q34 region. Manifested by the presence of slowly progressive distal atrophy, weakness, and corticospinal signs, these cases show a marked loss of spinal motor neurons and less pronounced degeneration of the corticospinal tracts. The peripheral nerves and dorsal and ventral roots show prominent axonal and

Table I
Genes and Chromosomal Loci Associated with Different Classes of Motor Neuron Disease

Motor neuron disease	Genes/chromosomal loci	Comments
Amyotrophic lateral sclerosis	NF-H	Mutations in KSP repeats; risk factor
	EAAT2	RNA editing errors; decrease in protein level
Familial amyotrophic lateral sclerosis	SOD1; Ch 21	Autosomal dominant; largely missense mutations
Juvenile amyotrophic lateral sclerosis	Ch 2q33	Autosomal recessive
	Ch 9q34	Autosomal dominant
Spinal muscular atrophy	SMN; Ch 5	Autosomal recessive; homozygous deletions
Spinobulbar muscular atrophy	Androgen receptor; Ch X	Expanded CAG repeats
Hereditary spastic paraplegia	Proteolipid protein; Ch Xq21	X-linked
	Neural cell adhesion molecule L1; Ch Xq28	X-linked
	Paraplegin; Ch 16q	Autosomal recessive; homozygous deletions
	Spastin; Ch 2p	Autosomal dominant
	Ch 2q, 8p, 12q, 14q, 15q, and 19q	Autosomal dominant
	Ch 8q	Autosomal recessive

swellings and axonal degeneration, with marked loss of both large and small myelinated fibers. These findings extend the spectrum of FALS and JALS to include a slowly progressive, autosomal-dominant, nonfatal but debilitating disease.

To date, deletion/insertion mutations in the KSP repeat motif of the neurofilament (NF)-H tail domain have been identified in 10 of 1047 patients with sporadic ALS and in 1 of 295 FALS patients. Although they are considered a risk factor, there is no direct evidence that mutations of NF genes are a primary cause of ALS. In a subset of patients with sporadic ALS, levels of EAAT2 protein were reduced compared to those of controls and individuals who died of other neurodegenerative diseases. In 17 of 28 cases of ALS, aberrant EAAT2 mRNA was found; significantly, lower levels of the EAAT2 protein were observed in these same tissue samples. Functional studies indicated that the mutant RNA species led to a dominant downregulation of EAAT2 protein synthesis. However, recent studies indicate that these aberrant RNA species were also observed in control cases. Further work is under way to clarify the importance of aberrant EAAT2 mRNA in the pathogenesis of sporadic ALS. Finally, a locus for disease has also been identified on the X chromosome in at least one ALS family.

C. Copper-Mediated Mutant SOD1 Neurotoxicity

As discussed previously, mutations in SOD1 cause up to 20% of cases of familial ALS, a fatal adult-onset motor neuron disease characterized by the selective loss of spinal and cortical motor neurons. A variety of *in vivo* and *in vitro* studies have demonstrated that the mutant enzyme causes selective neuronal degeneration through a gain of toxic property rather than a loss of SOD1 activity, consistent with FALS displaying an autosomal-dominant pattern of inheritance. Although some FALS-linked SOD1 mutants show reduced enzymatic activities, others retain nearly wild-type levels and mutant SOD1 subunits do not appear to alter the metabolism/activity of wild-type SOD1 in a dominant-negative fashion. Transgenic mice expressing different FALS-linked SOD1 mutants exhibit progressive motor neuron disease resembling that occurring in cases of ALS, and mice deficient in SOD1 do not develop a motor neuron disease-like phenotype. Taken together, these results are consistent with the view that mutant SOD1 causes disease through a gain of neurotoxic property.

Although the role of SOD1 in scavenging superoxide free radicals is well established, other functions, such as peroxidase activity or buffering against Cu toxicity, have been documented. However, the molecular mechanisms whereby mutant SOD1 causes selective motor neuron death remain uncertain. One hypothesis is that the Cu bound to mutant SOD1 plays a key role in generating the toxic property in SOD1-linked FALS; that is, mutations induce conformational changes in SOD1 to facilitate the interactions of the catalytic Cu with small molecules, such as peroxynitrite or hydrogen peroxide, to generate toxic free radicals that damage a variety of cell constituents important for the maintenance and survival of motor neurons.

Consistent with the peroxynitrite hypothesis are results showing increased levels of free nitrotyrosine in G37R mutant SOD1 transgenic mice and in both sporadic and familial ALS. However, there is no documentation of increased levels of nitrated proteins in mutant SOD1 mice compared to control mice. In addition, disease progression in G93A mutant mice lacking either neuronal or endothelial nitric oxide synthase genes showed no significant difference when compared to that of G93A mice; these results do not support a critical role for neuronal-derived nitric oxide in the pathogenesis of FALS. In support of the peroxidase-mediated mechanism, markers of oxidative damage are increased in the cortices of sporadic cases of ALS, and it has also been demonstrated that the glutamate transporter, GLT1, is inactivated by oxidative reactions initiated by hydrogen peroxide and catalyzed by FALS-linked mutant SOD1. Studies in mutant SOD1 mice, however, failed to show increased levels of hydroxyl radical when compared to control mice. Furthermore, recent studies have provided evidence against the generation of free hydroxyl radicals from the interaction of SOD1 and hydrogen peroxide. Indeed, the onset and progression of disease in G85R mutant mice are independent of the level of wild-type SOD1 activities; this is inconsistent with the peroxynitrite and the peroxidase hypotheses.

Recent studies have shown that the loss of zinc from either mutant or wild-type SOD1 is sufficient to induce apoptosis in cultured motor neurons by trophic factor withdrawal. Because the toxic effect observed required Cu to be bound to SOD1 and endogenous nitric oxide to be produced, these studies suggested that zinc-deficient SOD1 might play a role in disease of both sporadic and familial ALS through a nitric oxide-mediated mechanism. However, no *in vivo* demonstration for these aberrant Cu chemistries, proposed to be

central to the pathogenesis of FALS, has been established. Although the Cu-mediated toxicity models remain controversial, the discovery of CCS provides the opportunity to directly test whether Cu within mutant SOD1 mediates motor neuron degeneration in SOD1-linked FALS.

1. Copper Trafficking and CCS

The delivery of Cu to specific proteins is mediated through distinct intracellular pathways of copper trafficking. A family of soluble proteins termed Cu chaperones is required to deliver Cu to specific intracellular metelloproteins. Recently, the yeast Cu chaperone, termed Lys7 or its mammalian homolog CCS (copper chaperone for superoxide dismutase), was shown to be necessary to deliver copper to SOD1 in yeast. CCS is able to rescue the Lys7 null mutant and is shown to interact physically with SOD1. Biochemical and structural analysis of yeast CCS indicates that the insertion of Cu into SOD1 requires interactions among three distinct domains of the copper chaperone: an Atx1p-like amino-terminal domain responsible for Cu uptake, an SOD1-like central domain that functions in SOD1 recognition, and a carboxyl-terminal domain unique to copper chaperones that mediates Cu incorporation into SOD1. Moreover, yeast CCS is sufficient to incorporate Cu into SOD1 through a direct transfer of Cu from the metallochaperone to its target enzyme under restricted concentrations of intracellular free Cu, indicating that a pool of intracellular free Cu is not used to activate SOD1.

2. Testing the Copper Hypothesis: CCS Knockout Mice

Recent studies have shown that a common property of both wild-type and FALS-linked mutant SOD1 is that Cu incorporation is CCS dependent. Thus, with regard to mutant SOD1, aberrant Cu chemistry may mediate degeneration of motor neurons in FALS. Consistent with this view are findings that CCS physically interacts with SOD1 and that both proteins are colocalized in many cell types, including motor neurons. To determine whether CCS is necessary for incorporation of Cu into SOD1 and to test the role of Cu in the pathogenesis of FALS using a genetic approach to alter the amount of Cu bound to mutant SOD1, mice were generated with targeted disruption of *CCS* loci ($CCS^{-/-}$ mice). Although $CCS^{-/-}$ mice are viable and possess normal levels of SOD1 protein, they reveal marked reductions in SOD1 activity when compared to control littermates. Metabolic labeling with radioactive ^{64}Cu demonstrated that the reduction of SOD1 activity in $CCS^{-/-}$ mice was the direct result of impaired Cu incorporation into SOD1 and that this effect was specific because no abnormalities were observed in Cu uptake, distribution, or incorporation into other cuproenzymes. Consistent with this loss of SOD1 activity, $CCS^{-/-}$ mice showed increased sensitivity to paraquat and reduced female fertility, phenotypes that are characteristic of SOD1-deficient mice. These results demonstrate the essential role of mammalian and have important implications for the development of novel therapeutic strategies in FALS.

As discussed previously, the view that Cu bound to mutant SOD1 mediates toxicity in SOD1-linked FALS remains controversial. However, the demonstration that CCS null mice are viable and possess marked reductions in SOD1 activity offers an opportunity to directly test whether Cu bound to mutant SOD1 plays a key role in the pathogenesis of mutant SOD1-induced FALS. Cross-breeding strategies using these CCS null mice and several lines of FALS mice (G37R, G93A, and G85R) are currently under way and outcomes of these studies should be instructive in deciphering the role of Cu in SOD1-linked FALS. During the past several years, these mutant SOD1 transgenic mice have been valuable in testing other hypotheses regarding the molecular mechanisms of selective motor neuron degeneration induced by mutant SOD1, and the results of these investigations are reviewed next.

D. Mutant SOD1 Transgenic Mice

1. G93A SOD1 Mice

The G1H line of mice, which express G93A human SOD1 at approximately five times the endogenous level, develop hindlimb weakness at 3 months of age. By 5 months of age, these mice are moribund. The number of choline acetyltransferase-positive motor neurons is reduced, and the Golgi is fragmented. Vacuoles, seen in cell bodies and dendrites, are thought to represent degenerating mictochondria. Some axons are swollen, showing accumulations of neurofilaments, and other axons are reduced in caliber; fast and slow transport appear to be impaired. In late stage, there is a loss of motor neurons. In another line of mice expressing low levels of SOD1, ubiquitin immunoreactivty and Lewy body-like inclusions are present in cell bodies of motor neurons; these mice also show

axonal swellings, loss of motor neurons, and local astrogliosis.

G93A SOD1 transgenic mice have been used to test a variety of therapeutic approaches. Vitamin E and selenium modestly delay the onset and progression of disease without affecting survival; in contrast, riluzole and gabapentin do not influence onset/progression but do slightly increase survival. Although initial studies showed that oral administration of *d*-penicillamine delays the onset of disease, recent results revealed no significant effects. Overexpression of Bcl-2 extends the survival of these transgenic mice, but the presence of the gene does not change the progression of the disease. Recently, the level of Bax, a molecule that promotes apoptosis, was shown to be increased in G93A mice. However, this increase in Bax was augmented by an increase in Bcl-2/Bax heterodimers in G93A mice overexpressing Bcl-2. Interestingly, there is evidence that an apoptotic mechanism may be involved in motor neuron loss in ALS. In a small group of G93A SOD1 mice overexpressing a dominant-negative inhibitor of interleukin-1B converting enzyme, a cell death gene, there is a modest slowing of progression of disease. Recently, it was documented that oral administration of creatine significantly delays the onset of disease and extends the life span of the G93A mice. To date, it appears that agents that act in an antioxidant/antiapoptotic manner delay disease onset, whereas antiexcitotoxic agents do not affect onset of disease but do significantly prolong survival.

2. G37R SOD1 Mice

These mice accumulate the G37R SOD1 to 3 to 12 times those levels of endogenous SOD1 in the spinal cord; the mutant SOD1 retains full specific activity. Levels of the mutant transgene product determine the age of onset. At 5–7 weeks of age ($\sim$2 or 3 months before the appearance of clinical signs), SOD1 accumulates in irregular swollen intraparenchymal portions of motor axon, the axonal cytoskeleton is abnormal, and vacuoles are present in the axon. Radiolabeling studies demonstrate that both endogenous SOD1 and G37R SOD1 are transported anterogradely in axons. Deficits in slow or fast axonal transport prior to disease onset have been documented in G37R and G85R mice and in G93A mice, respectively. Thus, toxic mutant SOD1 is transported anterograde, accumulates early in the disease, and associated with early structural pathology in axons. Small vacuoles are also present in dendrites, and some vacuoles appear to originate in the space between the outer and inner mitochondrial membranes, with prominent distention of the outer mitochondrial membrane, displacement of the inner membrane, and disruption of the cristae. Both axonal and dendritic abnormalities occur months before the onset of clinical signs. By 20 weeks of age, motor axons show vacuoles, cytoskeletal alterations, and Wallerian degeneration. The cell bodies of some neurons show ubiquitin-immunoreactive inclusions and phosphorylated NF-H immunoreactivity. Recently, motor axon inclusions immunoreactive with peripherin, an intermediate filament protein that when overexpressed in transgenic mice causes motor neuron death, have been documented in presymptomatic G37R mice. The number of motor neurons was reduced. Astrogliosis is present in proximity to abnormal neurons. These findings suggest that the accumulation of mutant SOD1 is associated with subtle and early damage to organelles within axons/dendrites of motor neurons. If these early axonal and dendritic lesions occur in ALS, they are virtually impossible to detect because they occur early and are subtle; in humans, they are masked by end-stage processes, immediate antemortem events, and postmortem delays.

NFs have been implicated in the pathogenesis of ALS. To test the role of NFs in motor neuron disease caused by SOD1 mutations, transgenic mice expressing G37R SOD1 were cross-bred to transgenic mice that accumulate NF-H–β-galactosidase fusion protein (NF-H–lacZ), which crosslinks neurofilaments in neuronal perikarya and limits their export to axons, and to transgenic mice expressing human NF-H subunits. In G37R SOD1 mice expressing NF-H–lacZ, NFs are withheld from the axonal compartment, but there is no influence on the progression of disease. These results indicate that neither initiation nor progression of disease require an axonal NF cytoskeleton and that alterations in NF biology observed in some forms of motor neuron disease may be secondary responses. In contrast, the expression of wildtype human NF-H transgenes in SOD1 mutant mice significantly increases the mean life span of the G37R SOD1 mice. In contrast to the striking axonal degeneration observed in 1-year-old transgenic mice expressing G37R SOD1, NF-H transgenic mice show remarkable sparing of motor neurons. The reasons for this protection of G37R SOD1 mice by increased levels of human NF-H, particularly in cell bodies but not in axons (in which neurofilament content is reduced), are not known; it could be related to the ability of NF-H to bind calcium and therefore influence calcium-mediated damage. Consistent with

a role for calcium in disease is the observation that overexpression of the calcium binding protein calbindin D (28K) confers protection against mutant SOD1-mediated death of PC12 cells.

3. G85R SOD1 Mice

In multiple lines with different number of copies (2–15 copies), the accumulation of mutant G85R protein ranged from ~20% in endogenous mouse SOD1 (in the lowest expressing line) to endogenous levels (in the highest expressing line). The age of onset of weakness varied from 8–10 months for the highest expressing line to 12–14 months for the lowest expressing line. Two weeks after initial weakness in one hindlimb, these mice were completely paralyzed. In the preclinical period, there was a defect in axonal transport, and astrocytes contained SOD1 and ubiquitin-immunoreactive Lewy body-like inclusions; at later stages, motor neurons also exhibited SOD1 and ubiquitin-positive inclusions. At 6.5 months (coincident with the earliest pathology), there was no significant loss of axons in the motor roots, but 1 or 2 weeks after the onset of clinical signs many axons had degenerated. Similar pathology has been reported in two individuals with FALS who had a two-base pair deletion in codon 126 of SOD1, leading to a frameshift and truncation of the final 27 amino acids of SOD1. To test the role of neurofilaments in disease, NF-L null mice were mated to G85R SOD1 transgenic mice. G85R mutant mice without neurofilaments show delayed onset of disease and extension of survival. Thus, although assembled neurofilaments are not required for mutant SOD1-induced disease, the absence of neurofilaments does appear to slow the G85R SOD1-mediated disease. To further examine disease mechanisms, G85R SOD1 transgenic mice were mated to wildtype SOD1 mice or to SOD1 null mice; elevation or elimination of wildtype SOD1 did not influence the SOD1 mutant-associated disease. These studies indicate that SOD mimetic may not be useful and is inconsistent with the view that mutant SOD1-mediated toxicity derives from superoxide-associated oxidative stress.

4. G86R SOD1 Mice

Mice expressing this mutant transgene, the murine equivalent of the G85R SOD1 mutation, developed progressive paralysis at 3 months of age; some neurons exhibited cytological evidence of degeneration. At end-stage disease, there was accumulation of phosphorylated NF inclusions in surviving motor neurons, a reduction in the number of motor neurons and interneurons, and reactive gliosis. Vacuolar pathology was not observed. Calbindin-containing nerve cells appeared to be relatively spared.

As discussed previously, the observations of astroglial inclusions in ALS and in mutant SOD1 transgenic mice indicate the possibility that astrocytes might participate in the pathogenesis of ALS. To assess the involvement of astrocytes, transgenic mice expressing the G86R SOD1 restricted to astrocytes (GFAP-m SOD1 mice) were generated. Although astrocytes from GFAP-m SOD1 mice showed hypertrophy with increased GFAP immunoreactivity, these animals developed normally with no motor neuron degeneration. These observations indicate that expression of mutant SOD1 in astrocytes is not sufficient to cause motor neuron disease. However, because these mice do not express high levels of G85R SOD1, additional lines of mice expressing higher levels of the mutant protein in astrocytes may prove useful.

E. Excitotoxicity in ALS

Excitotoxicity has been suggested as one mechanism by which motor neurons are damaged in ALS. About 60–70% of sporadic ALS patients have a 30–95% reduction in the astroglial glutamate transporter EAAT2 (excitatory amino acid transporter 2), also termed GLT-1 in motor cortex and spinal cord. Loss of the major transporter could lead to an increase in extracellular concentrations of glutamate, as seen in some patients with ALS, and excitotoxic degeneration of motor neurons. Significantly, in a subset of patients with sporadic ALS (65%), multiple abnormalities of *EAAT2* mRNA, including intron retention and exon skipping, have initially been identified in tissues from the affected areas. The aberrant mRNAs were highly abundant, found only in neuropathologically affected areas of ALS patients and not in other brain regions, and detectable in the CSF of living ALS patients early in the disease. *In vitro* expression studies suggest that proteins translated from these aberrant mRNA can undergo rapid degradation and/or produce dominant-negative effects on normal EAAT2, resulting in loss of protein and activity. These findings suggest that a reduction of EAAT2 in ALS could be due to the presence of aberrant EAAT2 mRNA, presumably resulting from RNA processing errors. This aberrant RNA processing reduces the levels of this type of glutamate transport, and accumulation of glutamate could predispose to excitotoxic damage of motor

neurons. Consistent with this idea is the finding that experimental reductions in EAAT2 by antisense olgonucleotide injection *in vivo* produce progressive limb weakness and motor neuron degeneration. However, recent observations have shown that these aberrant EAATS mRNAs were also found in control cases. The significance of these aberrant mRNAs in ALS pathogenesis remains a focus of future research.

II. SPINAL AND BULBAR MUSCULAR ATROPHY

Spinal and bulbar muscular atrophy (SBMA or Kennedy's disease), an X-linked disorder, usually manifests with slowly progressive proximal weakness in the third through fifth decades. The disease, mapped to Xq11–q12, is caused by expanded CAG repeats encoding a polyglutamine tract in the androgen receptor protein. Disease-linked alleles (37–66 CAGs) change length when transmitted from parents to children, with an increased tendency to change when inherited through the father. Individuals with longer CAG repeats have early onset disease. Neuropathologically, there is degeneration of lower motor neurons in spinal cord and brain stem, accompanied by local gliosis and atrophy of skeletal muscles. Both the wild-type and the mutant androgen receptor are widely distributed, predominantly in the cytoplasm of neurons. Nuclear ubiquitin-immunoreactive inclusions containing the mutant androgen receptor are observed in spinal motor neurons of SBMA but not in other, unaffected neural tissues. Similar nuclear inclusions are seen in some nonneural tissues. The mechanism whereby the expanded trinucleotide repeats cause disease, including Huntington's disease, is the subject of exciting recent research. Although there is an indication that the expanded repeat causes some loss of transcriptional activity by the androgen receptor, the major influence of the expansion is likely related to a gain of toxic function.

Recent advances in neuronal cell culture and transgenic models have documented pathological findings that resemble those observed in individuals with Kennedy's disease. Cell culture studies have indicated that the aggregation and proteolytic processing of the androgen receptor are dependent on the length of the polyglutamine repeat. Moreover, the abnormal metabolism of the expanded repeat androgen receptor is associated with cellular toxicity. These results suggest a molecular basis for the neurotoxic gain of function associated with neuronal degeneration in SBMA. Initial transgenic mice expressing the human androgen receptor cDNAs with 45 or 66 CAG repeats driven by a variety of promoters failed to reveal any pathological phenotype. Recently, mice expressing a truncated androgen receptor with 112 CAG repeats driven by the prion protein promoter showed a remarkable neurological abnormality associated with tremor, gait problem, circling behavior, and seizures. As observed in SBMA cases, nuclear inclusions are found in motor neurons of these transgenic mice. Several lines of yeast artificial chromosome (YAC) transgenic mice carrying 45 CAG repeat expansions in the androgen receptor showed an $\sim 10\%$ rate of repeat length instability. In addition, the 45 CAG repeat tracts caused greater instability with maternal transmission and with older transmitting female. These studies indicate that human locus-specific sequences are necessary to generate trinucleotide repeat instability in mice. Further efforts to identify the *cis*-acting elements that permit CAG tract instability and the *trans*-acting factors that regulate repeat instability will be useful to clarify the molecular basis of trinucleotide repeat instability in SBMA.

III. HEREDITARY SPASTIC PARAPLEGIA

Hereditary spastic paraplegia (HSP), a genetically heterogeneous illness, is characterized by weakness, progressive and bilateral spasticity of the lower limbs, increased deep tendon reflexes, and extensor plantar responses—signs reflecting involvement of corticalspinal pathways. Axonal degeneration of the corticospinal tracts and posterior columns is a common characteristic of all types of HSP. Although this disease can be inherited as an autosomal-dominant disease (AD-HSP), an X-linked illness (X-HSP), or an autosomal-recessive disease (AR-HSP) (Table I), AD-HSP represents the majority of cases ($\sim 80\%$).

X-HSP has been linked to two regions of the X chromosome. At Xq28, mutations in the gene for neural cell adhesion molecule L1, an axonal glycoprotein involved in neuronal migration and differentiation, cause one form of HSP linked to mental retardation, whereas mutations in the proteolipid protein gene at Xq21 cause another form of HSP.

A subset of autosomal-recessive cases of HSP have been shown to be homozygous for a 9.5-kb deletion involving paraplegin, a member of the AAA protein family that is homologous to several yeast mitochondria ATPases. These ATPases exhibit both proteolytic and chaperone-like activities at the inner mitochondrial membrane. Paraplegin is localized to the inner

mitochondrial membrane, and muscle biopsies from affected individuals show impairments in oxidative phosphorylation in these tissues. The AAA family of proteins possesses diverse cellular functions, including microtubule rearrangement, protein degradation, organelle biogenesis, and vesicular transport. The membrane-bound AAA proteases have been shown to possess metal-dependent peptidase activity and serve to degrade nonassembled membrane proteins. Studies of yeast have demonstrated that Yme1, a yeast homolog of paraplegin, mediates binding of substrates through the amino-terminal region of the AAA domain to degrade unfolded membrane proteins. In addition, the purified AAA domain of Yme1 was shown to bind unfolded polypeptides and prevent their aggregation. These studies indicate that the chaperone-like activity of AAA proteins resides within the AAA domain.

Six AD-HSP loci have been mapped (Table I) and the locus at chromosome 2p accounts for ~50% of all AD-HSP pedigrees. Recently, mutations of another AAA protein, termed spastin, have been linked to a subset of AD-HSP at the chromosome 2p locus. Sequence analysis revealed various modifications of the spastin gene, including missense, nonsense, and splice site point mutations, as well as deletions and insertions. Spastin, ubiquitously expressed in both fetal and adult tissues, is thought to be an ATPase involved in the function or assembly of nuclear protein complexes. Generation of mouse models for HSP should further clarify the molecular mechanisms whereby these genetic abnormalities lead to the various types of HSP.

IV. SPINAL MUSCULAR ATROPHY

Spinal muscular atrophy (SMA), the most frequent fatal autosomal-recessive disorders in infants and one of the most important neuromuscular disorders of children, is characterized by muscle weakness and atrophy. The disease is classified as type I, II, or III on the basis of age of onset and functional ability. The incidence of type I SMA (Werdnig–Hoffmann disease) is estimated to be ~1 in 10,000 live births, with a carrier rate frequency between 1 in 50 and 1 in 80. Affected infants become weak prior to 6 months of age, ~30% of individuals show reduced fetal movements *in utero*, and these infants develop respiratory and bulbar difficulties and die before 2 years of age. Type II SMA is usually recognized before the first birthday; these children are never able to walk independently. Type III SMA begins in early childhood; affected individuals have milder degrees of disability. Affected siblings have a very similar severity of phenotype, suggesting a significant effect of allelism on phenotype or the presence of a tightly linked disease-modifying gene.

Although much remains to be learned about the neuropathology of SMA, the most characteristic feature of affected individuals is the paucity of large α motor neurons in the spinal cord and brain stem, with associated degeneration of motor roots and denervation of skeletal muscle. Some of the surviving neurons appear normal, and no characteristic neuronal inclusions have been demonstrated with the disorder. A minority of surviving motor neurons demonstrate chromatolysis, accumulations of phosphorylated NF in neuronal cell bodies. Relatively spared are cranial nerve nuclei III, IV, and VI as well as motor neurons of the phrenic nerve (which innervates the diaphragm) and Onuf's nucleus (which innervates the male genitalia and anal sphincters). Upper motor neurons do not appear to be affected.

Molecular genetic studies have mapped SMA types I–III to a single, highly complex genetic locus on chromosome 5q11.2–q13.3. The genomic organization of this region is polymorphic and complex, with numerous repeated DNA elements that occur in different orientations on chromosomes from different individuals. Two SMA candidate genes were described within this region, survival motor neuron (*SMN*) and neuronal apoptosis inhibitory protein (*NAIP*). Recent evidence is consistent with the idea that *SMN1* is the SMA–determining gene; *NAIP* may play a role in modifying the severity of the disease.

The *SMN1* gene, encoding a 294–amino acid polypeptide (32 kDa), lies in the telomeric portion of chromosome 5q11.2–q13.3, and a homologous copy of *SMN1*, termed *SMN2*, is located in a more centromeric position. Although the two ~20-kb genes encode identical proteins, they can be differentiated by single-strand conformational polymorphism analysis or reverse transcriptase-polymerase chain reaction because of differences in the transcripts at two nucleotides in exons 7 and 8. Transcripts encoded by the two genes are subjected to alternative modes of posttranscriptional processing; whereas *SMN1* produces full-length RNA, *SMN2* produces mainly transcripts lacking exon 7. Recently, it was demonstrated that exon 7 skipping is caused by a single nucleotide difference between the *SMN1* and *SMN2* genes. Thus, it has been observed that the ratio of transcripts lacking exon 7 to full-length *SMN2* transcripts correlates with the severity of SMA. The mouse

contains only one copy of *SMN*, which is also closely linked to NAIP on mouse chromosome 13. *SMN* is highly conserved between mice and humans, with 82% identity. *In situ* hybridization of adult mouse brain and spinal cord demonstrates expression of *SMN* in large motor neurons as well as most other cells. Similarly, almost all neurons in the brain express *SMN*, with highest levels seen in the hippocampus and cerebellum.

Mutations are found in nearly 100% of affected individuals, and levels of *SMN1* are different in the different types of SMA; SMA type I has the least amount of SMN in spinal motor neurons. Partial or complete deletions are detected in *SMN1* in more than 95% of well-characterized cases, with other disabling mutations found in many of the remaining cases. Homozygous deletion of *SMN1* has not been found in controls. *SMN2* has not been found to be deleted in any affected individual, although it is deleted in 4% of normal controls. This suggests that homozygous deletion of the centromeric and telomeric copies of SMN may be nonviable and likely results in an embryonic lethal event, and that levels of full-length SMN proteins govern whether the individual has SMA type I, II, or III. Thus, *SMN2* is predicted to be able to partially substitute for *SMN1*.

SMN, a member of a multiprotein complex, is involved in small nuclear ribonucleoprotein (snRNP) biogenesis and pre-mRNA splicing. In HeLa cells, SMN is present in the nucleus associated with small intranuclear structures termed ''gem bodies'' (Gemini of coiled bodies), which are closely associated with but distinct from coiled bodies. Because of the similarity between gems and coiled bodies, it is suggested that gems and SMN might play roles in the processing of small nuclear RNAs. In the cytoplasm, SMN complex is associated with snRNP Sm core proteins and is involved in spliceosomal snRNP assembly. A variety of methods have shown that SMN forms a stable heteromeric complex with a novel 32-kDa protein termed Gemin2; Gemin2 was formerly called SIP1 (SMN-interacting protein-1). Gemin2 and SMN are found in a large stable complex ($\sim$300 kDa) together with spliceosomal snRNP proteins (B, D1-3, and E) in both the nuclear and cytoplasmic fractions. SMN was shown to physically associate with Sm proteins and functional studies have revealed that Gemin2 plays an important role in the assembly of snRNPs. Studies using a dominant-negative form of SMN (SMNΔN27) showed that SMN also participates in the cytoplasmic assembly of snRNPs; overexpression of SMNΔN27 caused abnormal snRNP assembly with an enlarged gem/coiled body complex in the nucleus. Additional functional analysis revealed that SMN is required for recycling of snRNPs in pre-mRNA splicing.

Recently, other components of the SMN complex, termed Gemin3 (a DEAD box protein with putative RNA helicase activity) and Gemin4, have been identified. Gemin3 colocalizes and interacts directly with SMN in gems, and this interaction has shown to be decreased in some SMA cases in which there is deficiency in snRNP regeneration activity. Gemin4 is associated with the SMN protein complex through direct interaction with Gemin3 and it is thought that Gemin4 function as a cofactor for Gemin3. It has also been shown that Gemin4 physically interacts with several Sm core proteins and is colocalized with SMN in the cytoplasm and in gems. It is proposed that the complex composed of SMN, Gemin2, Gemin3, and Gemin4 represents an active functional unit that serves to bind substrates such as Sm proteins.

SMN has previously been documented to interact with fibrillarin, a common component of small nucleolar RNPs (snoRNPs), and hnRNP U, a member of the hnRNP protein family, which bind nuclear RNA. In yeast, fibrillarin is required for pre-rRNA processing and is essential for viability, whereas hnRNP U is believed to play a role in the transport and processing of mRNA. Recently, it was demonstrated that Gemin4 is also localized to the nucleoli. Thus, it is possible that the SMN complex may participate in snoRNP assembly and ribosomal RNA metabolism.

Taken together, these studies showed that the SMA disease gene *SMN* is linked to several fundamental biochemical pathways. It has also been demonstrated that SMN mutants found in SMA patients are defective in binding to Sm proteins because mutant SMN is incapable of forming large oligomers, which is essential for high-affinity binding to spliceosomal snRNP Sm proteins. These findings further support the view that abnormalities of spliceosomal snRNP biogenesis and metabolism are directly involved in the pathogenesis of SMA.

In an effort to further examine the *in vivo* function of SMN in SMA, mice deficient in SMN were generated. $Smn^{-/-}$ embryos died with massive cell death during the periimplantation stage corresponding to the initiation of embryonic RNA transcription, precluding analysis of any postnatal phenotype. These results are consistent with the view that SMN functions in essential cellular pathways, including the biogenesis of spliceosomal snRNPs and pre-mRNA splicing. To test whether *SMN2* can complement the embryonic lethality of $Smn^{-/-}$ embryos and to generate a mouse model for SMA, several groups generated transgenic

mice expressing human *SMN2* and cross-bred them with $Smn^{+/-}$ mice to generate *SMN2* transgenic mice lacking mouse endogenous SMN. These $Smn^{-/-}/SMN2$ mice showed neuropathological abnormalities in the spinal cord and skeletal muscles, similar to cases of SMA. The severity of the pathology in these mice correlated with the level of SMN polypeptide that retained the amino acids encoded by exon 7. These results demonstrated that *SMN2* could partially compensate for the endogenous mouse SMN and the variable phenotypes observed in the $Smn^{-/-}/SMN2$ mice recapitulate those seen in SMA types I–III. These $Smn^{-/-}/SMN2$ mice revealed clinical symptoms that can be grouped into three categories. Type 1 mice, the most severe pathological form, did not develop furry hair and died by Postnatal Day 10. Type 2 mice showed poor activity and died between 2 and 4 weeks of age. Type 3 mice survived and bred normally and had short and enlarged tails. Furthermore, the level of full-length SMN protein in these $Smn^{-/-}/SMN2$ mice correlated with the severity of the disease. These studies strongly support the view that the level of intact SMN is a determining factor for SMA severity.

Analysis of $Smn^{+/-}$ heterozygous knockout mice revealed an $\sim$50% reduction of SMN protein in the spinal cord of these mice that resulted in a progressive loss of motor neurons between birth and 6 months of age, and the phenotype of these mice resembles SMA type III. Taken together, these studies support the view that SMA is caused by insufficient level of SMN and the severity of the disease is dependent on the level of SMN generated from the *SMN2* gene. These mouse models of SMA will be useful in obtaining a better understanding of disease mechanisms and for testing therapeutic strategies.

To clarify the role NAIP played in SMA, mice with targeted deletion of *NAIP1* were recently generated. Although mice lacking *NAIP1* showed no overt phenotype, the survival of pyramidal neurons in the hippocampus after kainic acid-induced injury is greatly suppressed in these mice compared to controls. Although *NAIP1* is not required for normal development of the nervous system, it is necessary for neuronal survival under stressful situations. These studies further strengthen the view that although *NAIP* may modify the severity of the disease, *SMN1* is the SMA-determining gene.

V. MOUSE MODELS OF TAUOPATHIES

During the past several years, mutations in the tau gene linked to a heterogeneous group of neurodegenerative disorders characterized by filamentous tau inclusions have been identified. These disorders, collectively termed tauopathies, include sporadic and familial frontotemporal dementia with parkinsonism linked to chromosome 17, progressive supranuclear palsy, and Pick's disease. In humans, six tau isoforms (three isoforms with three microtubule-binding repeats and three isoform with four microtubule-binding repeats) are generated by alternative splicing of the tau gene in adult brain. In rodents, there are three four-repeat isoforms. To model neurodegenerative tauopathy, investigators have developed transgenic mice expressing high levels of wild-type tau isoforms. Overexpression of either the shortest or the longest tau human isoform in the brain and spinal cord of transgenic mice using either the Thy1 or the prion protein promoter, respectively, resulted in an age-dependent phenotype characterized by insoluble, hyperphosphorylated tau and intraneuronal inclusions composed of tau-immunoreactive filaments. These tau inclusions are found in large numbers in spinal motor neurons. These mice showed axon degeneration and decreased axonal transport. Not surprisingly, they revealed motor impairments. Although these studies showed that the overexpression of a single tau isoform can develop filamentous tau inclusions and neurodegeneration, other features of tauopathies such as filamentous tau tangles were not observed. Future efforts to generate additional tau transgenic models should provide further insight into the molecular mechanisms and pathobiology of tauopathies.

VI. CONCLUSIONS

During the past several years, substantial progress has been made in understanding the molecular events that underlie motor neuron disease. Significantly, several genes and a variety of loci have been linked to motor neuron disease. Free radical-mediated chemistries, copper toxicity, neurofilament biology, axonal transport, RNA metabolism, excitotoxicity, protein folding/aggregation, etc. appear to influence the function/viability of neurons. The mechanisms that account for "selective" neuronal degeneration are uncertain. However, the identification of specific genes/proteins that are mutated/deleted in the inherited forms of the disease has allowed investigators to create *in vivo* and *in vitro* model systems. For example, transgenic mice that recapitulate some of the features of human diseases have provided important information about the biology of these diseases. The availability of new

models will allow investigators to examine the molecular mechanisms by which mutant proteins cause selective dysfunction/death of motor neurons. Moreover, pathogenic hypotheses can be tested by experimental manipulations and by breeding mice carrying mutant genes to mice that express other transgenes or to gene-targeted mice. The results of these approaches should provide a better understanding of the pathogenic mechanisms of disease. In turn, this new knowledge should lead to the design of novel therapeutic strategies that can be tested in these animal models.

See Also the Following Articles

MOTOR CONTROL • MOTOR CORTEX • MOTOR SKILL • MULTIPLE SCLEROSIS • NEURODEGENERATIVE DISORDERS • NEURON

Acknowledgments

The authors thank Drs. Jamuna Subramaniam, Lio Tessarollo, Valeria Culotta, Jonathan Gitlin, David Borchelt, Lee Martin, Michael Lee, Sam Sisodia, Dave Cornblath, Bruce Rabin, Vassilis Koliatsos, Glen Lin, Tom Crawford, Mark Becher, Paul Hoffman, Ann Bergin, Don Cleveland, Jack Griffin, Lucie Brujin, and Carlos Pardo for their helpful discussions and contributions to some of the work mentioned in this text. Aspects of this work were supported by grants from the U.S. Public Health Service (NIH NS 37771, NS 10580, NS 20471, AG 05146, AG 07914, AG 10491, AG 10480, AG 14248 and AG 67914).

Suggested Reading

Cleveland, D. W. (1999). From charcot to SOD1: Mechanisms of selective motor neuron death in ALS. *Neuron* **24,** 515–520.

Cudkowicz, M. E., McKenna-Yasek, D., Sapp, P. E., Chin, W., Geller, B., Hayden, D. L., Schoenfeld, D. A., Hosler, B. A., Horvitz, H. R., and Brown, R. H. (1997). Epidemiology of mutations in superoxide dismutase in amyotrophic lateral sclerosis. *Ann. Neurol.* **41,** 210–221.

Culotta, V. C., Lin, S. J., Schmidt, P., Klomp, L. W., Casareno, R. L., and Gitlin, J. (1999). Intracellular pathways of copper trafficking in yeast and humans. *Adv. Exp. Med. Biol.* **448,** 247–254.

Fischbeck, K. H., Lieberman, A., Bailey, C. K., Abel, A., and Merry, D. E. (1999). Androgen receptor mutation in Kennedy's disease. *Philos. Trans. R Soc. London B Biol. Sci.* **354,** 1075.

The Hereditary Spastic Paraplegia Working Group. Hereditary spastic paraplegia: Advances in genetic research. *Neurology* **46,** 1507–1514.

Julien, J.-P. (1999). Neurofilament functions in health and disease. *Curr. Opin. Neurobiol.* **9,** 554–560.

Lee, V. M. Y., and Trojanowski, J. Q. (1999). Neurodegenerative taupathies: Human disease and transgenic mouse models. *Neuron* **24,** 507–510.

Lefebvre, S., Burglen, L., Frezal, J., Munnich, A., and Melki, J. (1998). The role of the SMN gene in proximal spinal muscular atrophy. *Hum. Mol. Genet.* **7,** 1531–1536.

Patel, S., and Latterich, M. (1998). The AAA team: Related ATPases with diverse functions. *Trends Cell Biol.* **8,** 65–71.

Price, D. L., Sisodia, S. S., and Borchelt, D. R. (1998). Genetic neurodegenerative diseases: The human illness and transgenic models. *Science* **282,** 1079–1083.

Motor Skill

ELIOT HAZELTINE
NASA—Ames Research Center

RICHARD IVRY
University of California, Berkeley

GLOSSARY

adaptation Skill learning tasks that require a remapping of motor commands to sensory feedback.

apraxia A neurological syndrome involving deficits in the production of voluntary movements without paralysis or sensory loss.

closed-loop movements Movements for which sensory feedback is used to modify the ongoing motor commands.

efference copy An internal copy of a signal to the muscles that is compared to afferent signals from the periphery.

open-loop movements Movements for which the motor commands are sent to the muscles without opportunity for modification due to sensory feedback.

sensorimotor integration Skill learning tasks that require the formation of novel associations between sensory events and motor commands.

Motor skills are learned behaviors that require patterns of activity across sets of muscles. Such a broad, encompassing description reflects the diversity of phenomena that are considered forms of motor skill learning. It is therefore not surprising that many neural systems are involved in the development of motor skills. This article focuses on the primary neural systems associated with motor learning and discusses functional hypotheses associated with these structures.

I. BACKGROUND

Our daily lives are replete with behaviors that our species had never attempted a few hundred years ago. Human activities and abilities vary dramatically across culture, occupation, and recreational taste. This diversity is the product of our brains' ability to develop new skills. Indeed, motor skill learning is a primary function of the central nervous system. Organisms see, smell, hear, and think in order to appropriately guide overt behavior.

Neuropsychology has often focused on perception, particularly vision, as a paradigmatic cognitive operation, perhaps because clinical populations have presented striking and specific deficits. In contrast, motor deficits, although diagnosed as distinct syndromes, have proven less amenable to attempts to decompose the neural systems underlying the control of action into functionally distinct neural systems. Although syndromes such as Parkinson's disease are associated with a unique set of symptoms, the difficulty in mapping the symptoms onto properties of the external world has made it difficult to develop functional models of the various neural structures involved in motor control.

Two aspects of motor control have been particularly problematic for researchers attempting to identify the neural substrates of motor learning. First, motor

behavior is thought to be governed by two distinct learning systems: declarative (or explicit) memory and procedural (or implicit) memory, a collection of memory systems including motor learning. For example, when learning to play tennis, one may initially attempt to consciously control the position of certain body parts. In this case, behavior is being shaped by the individual's declarative memories. However, as the individual becomes more practiced, the movements become fluent and less attention is required to control and coordinate the various parts of the body. In this latter case, implicit learning is taking place. Because brain lesions can lead to dramatic deficits in declarative memories while leaving many aspects of motor learning intact, two forms of learning are assumed to exist, supported by a different neural systems. However, determining the relative contributions of these two systems to a given behavior is a formidable task for researchers. In particular, given that experimental subjects may apply declarative knowledge and conscious strategies to guide their performance, it can be difficult to obtain pure measures of procedural motor learning.

Second, neural structures that are presumed to be critical for motor learning may also be involved in the production of movement. When evaluating the performance of individuals with damage in these brain regions, it can be difficult to distinguish between deficits in learning and deficits in execution. That is, the neural structures necessary to encode a motor skill may be intact, but the person may be unable to express that knowledge because the structures controlling output are not operating adequately. Similarly, impairments in performance may make movements so variable that learning is reduced even though there is no damage to the systems that normally encode new behaviors.

Despite these obstacles, dramatic progress has been made in our understanding of motor learning and its neural substrates. Much of our understanding of how the brain learns to refine motor signals in response to sensory feedback has come from studies examining neural activity in awake, behaving animals. These neurophysiological studies record changes in the firing rates of sets of neurons that accompany motor learning while interfering minimally with the brain's normal operation. Similarly, the advent of functional whole brain imaging techniques, such as positron emission tomography (PET) and functional magnetic resonance imaging (fMRI), has allowed researchers to observe changes in brain metabolism and blood flow as human subjects become more practiced with motor tasks.

These procedures have revealed learning-related changes in almost every region of the brain associated with motor control. In the cerebral cortex, these regions include the primary motor cortex (M1), the premotor cortex, the supplementary motor cortex (SMA), and both superior and inferior parietal cortex. Subcortical structures, including the basal ganglia and cerebellum, also play a critical role in motor learning in addition to their roles in motor execution.

Differentiating the roles of these various structures presents many intricate problems. Although lesions to many of these brain structures can lead to characteristic deficits in performance, understanding these impairments in terms of motor learning has been fraught with controversy. The controversy stems in part from the fact that the domain of these systems in many cases has yet to be determined. The systems may necessarily interact during the performance of most behaviors, blending their computational operations. When one system fails, the others may act in a degraded or altered fashion. Thus, determining what each of the many areas involved in motor control contributes to our diverse and multifaceted abilities has proven to be a thorny task.

II. LEVELS OF REPRESENTATION

Learning to modify actions based on sensory inputs and previous experiences is a complex computational problem. In fact, the pervasiveness of skill acquisition makes it difficult to define. For example, learning a novel, multicomponent series of movements, such as when we throw a baseball, is certainly considered motor skill learning. However, do we include other skills, such as learning to use a mouse or blink in anticipation of a puff of air? In all these cases, new relationships between particular actions and environmental events must be encoded. However, in some cases, the new behavior requires the precise timing and modulation of different muscles. In other cases, the movements are already well entrained and acquiring the new skill requires learning the appropriate goal that is associated with a particular set of environmental conditions. For example, in eye-blink conditioning, the movement is well established and learning primarily consists of determining the stimulus conditions in which this response should be emitted.

Another behavior, handwriting, serves as a useful example when considering the different processes involved in motor learning. Learning to write requires the complex coordination of multiple effectors to

produce elaborate spatiotemporal patterns of output. Handwriting is usually produced by a set of effectors including the muscles of the fingers and wrist of the dominant hand. However, it can also be produced —and with considerable skill—with other, nonoverlapping sets of effectors. For example, writing can be performed relatively effortlessly with the muscles of the right arm and shoulder, such as when writing large words on a chalkboard.

These transfer effects have led to the hypothesis that skilled movements are constrained by abstract motor programs. An abstract motor program does not specify the commands sent to the muscles or even which muscles must be activated; this information must be provided by downstream systems that implement a motor program. Exactly what a motor program represents remains controversial. However, the absence of effector-specific commands makes the abstract motor program so powerful. Once established, it can be useful in a variety of contexts, depending on the goals and circumstances confronting the organism. Given the complexity of our environments, it is advantageous since learning acquired with one set of effectors can transfer to other sets, thus allowing behavior to be both skilled and flexible.

However, transfer of knowledge to different effectors does not always appear to be complete. When right-handed individuals are asked to write with their left hands, their movements are comparatively slow and awkward, and the resulting trajectories can be considerably distorted. Such difficulty indicates that many components of the learned behavior do not transfer well to this new set of effectors. Even so, many features of an individual's unique style are present in their production with the nondominant hand. That is, the shape of the letters appears to match, although with less precision, the same template used to generate those produced by the dominant hand.

The partial transfer of handwriting to novel effectors suggests that multiple representations can subserve a single learned behavior. Representations about the desired trajectory of the movement appear to be encoded in spatial coordinates that can be accessed by various effectors. However, such representations are not sufficient to rapidly produce fluent performance. In addition to requiring a representation of the desired goal of the movement, individuals must also learn how to accomplish the goal with a specific set of muscles.

This example illustrates the distinction between the two classes of knowledge required to perform motor acts, often referred to as "knowing what to do" and "knowing how to do." Although it is tempting to associate the former with declarative knowledge and the latter with procedural knowledge, several lines of evidence suggest that "knowing what to do" can be encoded without awareness. That is, in many cases, decisions about what to do appear to be facilitated by implicit knowledge.

Understanding how these multiple levels of representation are supported by the neural structures associated with motor learning is critical for a comprehensive theory of motor skill learning. In this article, we review some experimental tasks that have been used to identify the neural loci of motor learning and develop functional hypotheses regarding the computations performed by these structures.

III. THE CEREBELLUM

The brain structure most frequently associated with motor learning is the cerebellum. Tucked beneath the visual cortex in the posterior cranial fossa, the cerebellum is the largest subcortical structure, constituting approximately 15% of the brain's mass and containing more than half of the entire neurons in the brain. The structure is divided into three lobes, which are composed of 10 lobules. Given its tremendous size, it is likely that the cerebellum contributes to many behaviors, and much recent speculation has focused on nonmotor functions of the cerebellum. Nonetheless, as emphasized in the classic observations of the British neurologist Gordon Holmes, the hallmark of cerebellar dysfunction in humans is the breakdown of skilled movement.

Some remarkable features of the cerebellum serve as a useful background for considering how this structure might perform essential computations for motor learning. Perhaps the most striking feature of the cerebellum is its uniformity. Inputs to the cerebellar cortex come from two distinct pathways, the mossy fibers and the climbing fibers. Mossy fibers originate in various brain stem nuclei and synapse onto the cerebellar granule cells. The axons of the granule cells, or parallel fibers, extend for as long as a few centimeters, synapsing with various cell types within the cerebellar cortex. Principal among these are the Purkinje cells, the sole output of the cerebellar cortex, providing an inhibitory signal on the deep cerebellar and vestibular nuclei. The parallel fibers can also influence the Purkinje cells indirectly via inhibitory interneurons called basket cells and Golgi cells. A given Purkinje cell may receive input from 200,000 parallel fibers, although each parallel fiber will only

synapse once on a particular Purkinje cell. Climbing fibers originate in the inferior olive. In stark contrast to the extreme divergence of information deriving from the mossy fibers, each Purkinje cell receives input from a single climbing fiber.

A. Eye-Blink Conditioning

Eye-blink conditioning is perhaps the best studied model system for examining the neural basis of sensorimotor learning in mammals. In this form of Pavlovian conditioning, a neutral stimulus such as a tone is repeatedly paired with an aversive stimulus such as an airpuff to the eye. The airpuff is an unconditioned stimulus (US): It automatically elicits an unconditioned response (UR), either a blink in humans or the extension of a protective membrane in rabbits. Initially, the tone, or conditioned stimulus (CS), causes no response. However, after training, the animal produces a conditioned response (CR) following the tone. The CR is timed to occur just prior to the onset of the US, and in this way it is adaptive as it attenuates the aversive consequences of the US. The order and timing of the CS–US events are critical. The tone must precede the airpuff (i.e., be predictive of the US). Also, the two events must be separated by a fixed interval or small set of fixed intervals. Training is poor or absent if the interval between the CS and US is random.

Despite the fact that eye-blink conditioning requires only extremely simple movements, the task contains some basic ingredients of motor learning: First, the appropriate response, the blink, must have no previous associations with the critical stimulus (e.g., the tone). Second, the characteristics of the response are critical for evaluating performance. If the blink is not timed correctly, then it will be of no benefit. For example, the blink is not adaptive if it occurs immediately upon presentation of the tone or if the air puff will not occur for another 500 msec. Thus, the behavior must be refined with practice so that the motor commands are precisely coordinated with environmental events. Finally, it appears that learning for this task is largely implicit. Individuals with anterograde amnesia are able to learn during eye-blink conditioning, at least under certain conditions.

Richard Thompson and colleagues have established the involvement of the cerebellum in eye-blink conditioning. Lesions of the cerebellum can abolish the CR in animals that have undergone conditioning and prevent the acquisition of the CR in a naive animal. Importantly, the UR remains largely intact, indicating that the deficit is not one of motor production but specifically related to learning. Although most work has been done on the rabbit, the relationship between cerebellar pathology and eye-blink conditioning has been observed in a range of species including humans.

Lesion and physiological methods have been used to identify the component pathways supporting eye-blink conditioning within the cerebellum. When the tone is replaced by stimulation to the mossy fibers, eye-blink conditioning proceeds normally, indicating that parallel fibers represent the CS. Similarly, when the US is replaced by stimulation to the climbing fibers, conditioning is also normal. Thus, the two input pathways to the cerebellum convey representations of the CS and US. The locus at which an association is formed between the CS and US (i.e., the site of learning) has been a subject of considerable debate. One possibility is the primary site of convergence of these inputs—the Purkinje cells of the cerebellar cortex. Another possibility is the deep cerebellar nuclei given that both the mossy and climbing fibers send collaterals to the nuclei. Indeed, lesions of the nuclei completely abolish the CR, whereas lesions of the cerebellar cortex disrupt the timing of the CR. It thus appears that the association of the CS and US can occur at multiple levels. For the CR to be acquired, the cerebellar nuclei must be intact; for the response to be appropriately timed and thus adaptive, the cerebellar cortex is required.

B. The Vestibular–Ocular Reflex

The vestibular–ocular reflex (VOR) is a second form of motor learning whose neural substrate has been largely identified within the cerebellum. The VOR causes movements of the eyes that are intended to compensate for movements of the head and body so that visual images remain stable on the retina. When we move our heads, our eyes rotate effortlessly so that the world does not appear to move in the opposite direction. If magnifying spectacles are worn, head movements lead to larger displacements of the visual scene, and the VOR adapts so that the eye movements have the appropriate magnitude.

Evidence from single-unit recordings and ablation studies shows that this adaptation requires the floccular complex of the cerebellum. Animal studies of the VOR are typically performed in the dark, with the animal placed on a turntable so that visual signals and body movements do not provide a potential source of

the neural and behavioral responses. Rotations of the head cause changes in the rate of simple spike discharges for the Purkinje cells within the floccular complex. When vision is provided and experimental manipulations cause the VOR to incorrectly compensate for the movement, the magnitude of these changes is modulated in a manner that is consistent with the appropriate alteration of the VOR gain.

In some respects it is surprising that similar circuits should be involved in adaptation-like tasks, such as adjusting the VOR, and eye-blink conditioning. Eye-blink conditioning involves discrete movements whose magnitude and trajectory require little on-line adjustment. In contrast, the VOR is a continuous behavior that involves the coordination of multiple muscles whose degree of activation must be precisely modulated. However, there is considerable overlap in the computational requirements for the two tasks. For either behavior to be successful, the organism must learn to precisely time specific muscle commands with sensory events to minimize an unwanted stimulus. For eye-blink conditioning, the stimulus is the puff of air on the animal's eye; for the VOR, the stimulus is movement of the image on the retina.

In the VOR, the initial motor reflex and vestibular inputs are represented along the parallel fibers and the Purkinje cells learn to produce the pattern of activation to tune the motor command so that the eye movement correctly compensates for the change in the head's position. Climbing fibers provide signals about the slippage of the image on the retina. Thus, there is another parallel between VOR adaptation (and normal function) and eye-blink conditioning. The simple spikes provided by the mossy fibers provide a representation of the current environmental context—a representation that likely includes an efference copy of the animal's own movements but also sensory information relevant for these actions. The climbing fibers provide an error signal, a representation of the unexpected and aversive air puff or of a mismatch between expected and actual eye position.

C. Adaptation

The VOR belongs to a broader class of motor learning termed adaptation. During adaptation the sensory consequences of movements are altered so that subjects must learn new relationships between their motor commands and the environment. In the case of the VOR, compensatory eye movements must be remapped to vestibular signals so that the same degree of head movement results in a new amount of eye movement.

The eye movements learned during VOR adaptation differ critically from many other forms of learned movements, such as reaching movements. Unlike the dynamics of the eye, the dynamics of a multijoint system, such as the arm and hand, has many degrees of freedom and is highly nonlinear. For most multijoint systems, including that composed of the wrist, arm, and shoulder, determining what muscles to activate given the desired trajectory is an ill-posed problem. Ill-posed problems are those that are insufficiently constrained so that there are multiple possible solutions. For example, there are many different sets of muscle commands that will cause the hand to move to a particular location. This situation complicates motor programming considerably because the control system must determine which solution to pursue as it attempts to optimize behavior. Therefore, attributing errors to the appropriate muscle commands becomes a complex procedure.

A common form of adaptation is prism adaptation. In this experimental paradigm, the subjects wear goggles fitted with prisms that laterally displace the visual image on the retina. Thus, if the goggles displace the image to the right, information from the left side of the visual field will fall on the fovea as the subject looks straight ahead. This apparatus alters the relationship between movements of the body and visual information. For example, in order to reach for an object that appears on the fovea when the eyes and head are pointing directly ahead, one must move the arm to a position that is displaced to the left or right of the body. This new relationship between vision and action contradicts a lifetime of experience, but particular movements can be successfully relearned in a matter of minutes.

When first worn, the prism goggles cause pointing movements to be systematically displaced from their targets, with the magnitude of the distortion approximately equal to the extent that the goggles displace the visual signal. However, performance improves quickly with practice. Within approximately 20 trials, the systematic error is eliminated and performance matches that observed before the goggles were introduced. This improvement can stem from distinct forms of behavioral adjustment that are unrelated to motor learning per se. Conscious strategies can compensate for the shifted visual image if one intentionally directs movements a few degrees to the left or right of the target to adjust for the displacement. To obtain measurements of adaptation that are uncontaminated

from compensatory strategies, data are collected after adaptation has occurred and the goggles are removed. Under these conditions, subjects initially exhibit negative aftereffects of similar magnitude but in the opposite direction of the displacements observed when they first wore the goggles. Presumably, after removing the goggles subjects abandon any conscious compensatory strategies. Thus, the negative aftereffect indicates that the adaptation reflects a learned automatic process remapping visual information and motor commands.

To better understand the computational processes that are altered during adaptation, it is useful to distinguish between open-loop and closed-loop movements. Open-loop movements are programmed without reference to any sensory feedback. Thus, error signals from the visual and proprioceptive systems are unable to alter the motor programs. In contrast, closed-loop movements allow for sensory feedback to indicate adjustments in the ongoing motor command. When movements are made relatively slowly, in a closed-loop manner, the effects of the prism goggles may not be obvious. However, when movements are made ballistically, in an open-loop manner, they will terminate in a location that is displaced from the desired endpoint.

Importantly, adaptation based on interactions between visual and proprioceptive feedback does not transfer to unpracticed movements or effectors. That is, if an individual learns to make accurate rapid pointing movements exclusively with the right arm, performance with the left arm will show little benefit. This finding indicates that learning does not entail a generic remapping of visual coordinates to an internal model of space. Instead, the remapping involves specific motor commands to coordinates in egocentric space. However, although learning appears to be specific to particular sets of movements, it does generalize to other portions of the visual field. In summary, learning appears to be better characterized as motoric rather than perceptual.

Individuals with cerebellar damage demonstrate deficits in their ability to adapt to the prism goggles. Although the patients do exhibit some improvement in performance when wearing the goggles, there are no aftereffects once the goggles are removed, suggesting that the improvement was based on conscious strategy rather than true adaptation. This finding is consistent with the proposal that the cerebellum plays a critical role in coordinating motor commands with sensory information, and that such learning occurs at a procedural level.

Adaptation can also occur without manipulating visual inputs. Peter Gilbert and Thomas Thach trained monkeys to move a manipulandum with the muscles of the wrist to a specific target location. The manipulandum was fitted with a torque motor that could apply forces to the handle for random intervals so that the monkey had to change the amount of force exerted on the handle to return it to the target location. Once the monkeys learned to maintain the handle's location with two levels of force (one requiring activation of extensor muscles and the other requiring activation of flexor muscles), a new level of force was introduced, replacing one of the learned levels. The task is considered to measure adaptation because after training with the novel level, effects are observed on the unchanged force level.

As the monkeys performed this task, single-unit recordings were made from Purkinje cells in the cerebellar cortex. The frequency of complex spikes, driven by climbing fiber activity, increased when the novel force was introduced and returned to baseline rates as the animals adapted. The frequency of simple spikes, driven by activity along the parallel fibers, decreased as the complex spike frequency increased and remained depressed as the complex spike frequency returned to baseline. These findings indicate that the climbing fiber activity is affected by learning and that the climbing fibers may serve to modulate the strength of the Purkinje cell–parallel fiber synapses.

D. General Cerebellar Function

The uniformity of the cerebellum's architecture and its well-established role in motor control, including the control of posture and multijoint movements, have motivated theorists to seek a general, unified account of cerebellar function and how it might contribute to motor control and learning. Indeed, given recent anatomical neuroimaging and neuropsychological findings associating the cerebellum and nonmotor functions, these theories have been extended to consider how the cerebellum might provide a computation that is common to a variety of motor and cognitive tasks.

Valentino Braitenberg proposed that the long parallel fibers, through the sequential synapses that they form across a series of Purkinje cells, might constitute a set of delay lines. These delay lines would allow for muscular patterns to be coordinated across synergistic muscle groups recruited in the course of complex actions. Although it appears unlikely that

delay lines based on the length and transmission velocity of the parallel fibers can produce a sufficient range of delays, the general idea that the cerebellum is essential for regulating the timing of movements is viable. As noted previously, the cerebellum, especially the cerebellar cortex, is essential for learning when the task requires the precise representation of the temporal relationship between successive events. Indeed, the cerebellum may perform similar timing computations for nonmotor tasks that require this type of representation, providing a general characterization of this system as an internal timing system. Recent computational models indicate that the representation of temporal information may emerge from the complex physiological interactions between the various cells of the cerebellar cortex that shape the output of the Purkinje cells.

Turning to a more mechanistic account of the cerebellum and motor learning, computational models have emphasized that the interactions of the parallel and climbing fibers within the cerebellar cortex can function as a supervised pattern-recognition device. The climbing fibers provide a training signal that modifies the synaptic strength of the connections between the parallel fibers and Purkinje cells. In this manner, the Purkinje cells can shape the topography of movements based on the previous reinforcement history associated with the input patterns provided by the parallel fibers.

Such a mechanism provides an excellent means for learning arbitrary input–output associations but is less obviously useful for computing the temporal relationships necessary for the fine-tuning of motor commands. Researchers, including James Houk and Mitsuo Kawato, have developed this basic model to allow for the dynamic control of the motor system. These models share the properties that critical information about the present and desired state of the effectors, or body parts, is represented along the parallel fibers and that training signals are propagated through the climbing fibers. The initial motor commands originate upstream from the cerebellum. The inhibitory output of the Purkinje cells transforms or sculpts this signal into the appropriate pattern activity to control the actual effector.

Within this framework, it has been proposed that the cerebellum serves as an inverse model of the controlled effector, a forward model, or a combination of the two. A forward model directly simulates the controlled effector, taking as input the motor commands issued to the effector and producing as output the predicted sensory feedback. One might question the use of a forward model in learning since it merely approximates the actual behavior of the controlled effector. However, sensory feedback is quite slow in relation to the speed at which movements are performed and thus, at least during ballistic movements, may be of little help during execution. A model of the effector that can anticipate the behavior of the body in response to motor commands provides one way in which feedback can be rapidly used to contribute to on-line control. The forward model can be used to anticipate the sensory consequences of motor commands, and the difference between the expected and actual sensory feedback can be used as a sort of error signal.

An inverse model, as its name suggests, performs the inverse of the computation performed by the forward model. It takes as input the present and the desired state of the controlled effector and produces as output the motor command required to achieve the desired state. Such a computational device is of obvious use: With an accurate inverse model, a system need only compute the desired trajectory for the effector and allow the model to determine the necessary motor command.

Learning to encode a forward model is straightforward: The system takes as input the present state of the effector and the motor command and learns to produce as output the sensory feedback. Learning to encode an inverse model is much more difficult because the appropriate motor command is not available to the system (if it were, then there would be little need to learn an inverse model). For an inverse model to be correctly trained, observable sensory errors must be converted into motor errors. Therefore, inverse models are comparatively computationally expensive.

Mitsuo Kawato and colleagues proposed that the cerebellum houses the neural circuitry supporting both forward and inverse models. Evidence for the presence of each type of model has come from analyses of the patterns of neural activity within the cerebellum. Such evidence is indirect, however, and researchers continue to strive to characterize the computation performed by the structure.

IV. INTERNAL MODELS AND SENSORIMOTOR INTEGRATION

Adaptation can be thought of as belonging to an even broader class of motor learning phenomena termed sensorimotor integration. Although adaptation is restricted to a remapping of motor signals to sensory

feedback, learning to track an object or move a mouse requires mapping particular sensory signals to sets of motor commands in a goal-dependent manner. These latter tasks differ from adaptation paradigms because these relationships involve unfamiliar sets of stimuli and responses.

Prominent among sensorimotor tasks has been mirror drawing. In this task, subjects are asked to trace geometric shapes with a pencil. However, they are not permitted to view their hands or the shapes directly but instead must view them indirectly via a mirror. With this apparatus, the visual feedback is opposite what one expects from his or her movements. For example, as the hand moves to the left, the feedback displays a movement toward the right. Initially, mirror tracing is quite difficult. However, after several trials of practice, subjects are able to learn to trace the shapes rapidly and smoothly, eventually approaching levels of performance that match those achieved when the hand and shape can be view directly.

The mirror-tracing task has played an important role in studies of the motor learning capabilities of patients with severe amnesia. Despite their severe memory deficits, these patients will show improvements across separate sessions of mirror drawing that are equivalent to the improvements observed in age-matched control subjects. Such behavioral evidence of learning is especially striking given that the same patients may be unable to recall having ever performed the task. The dissociation provides a powerful demonstration that motor learning is subserved by an independent set of neural structures from those associated with explicit memory.

This is not to deny that people with intact explicit memory cannot use this system to improve their performance on the mirror-drawing task. For example, an individual can learn to make movements in the reverse direction without ever performing the task. This explicit strategy can also lead to a dramatic facilitation of performance, but it is not generally referred to by the term "motor learning." In neurologically healthy individuals, it is likely that learning is occurring in parallel within both declarative and procedural memory systems.

Daniel Wolpert and colleagues used sensorimotor integration tasks that alter the visual feedback of arm movements. In these experiments, individuals make reaching movements to targets that are projected onto a mirror-lined horizontal surface. The participant's arm is below this surface, hidden from view. The phenomenal experience is that of reaching to positions along the same surface as that on which the hand is located. Feedback of arm position can be provided by the projection of a cursor onto the mirrored surface and can be either veridical or artificially distorted by the experimenter. A basic finding in the motor control literature is that the hand trajectory generally follows a straight path during arm movements, a result that mandates a more complicated trajectory of each limb segment in joint space. However, when the feedback is altered so that such movements appear curved, participants will automatically adjust the trajectory so that the perceived path is straight. To do so, they must produce curved hand trajectories, thus implying an adjustment in the transformations associated with the internal model. This finding emphasizes the importance of visual information in providing feedback for motor learning. Studies using this technique have explored whether such modifications are applied in a general manner or whether they are restricted to particular regions of space. The results of such studies show that the tuning of the internal model is specific to particular regions of space. That is, the distortion diminishes as the distance from the practice location increases.

A second common experimental procedure used to measure sensorimotor integration is rotary pursuit tracking. Like mirror tracing, individuals with dense amnesia are able to improve by practicing rotary pursuit despite being unable to remember having performed in the earlier sessions. This task requires subjects to maintain contact with the moving object with a stylus as it follows a predictable circular path. In some versions of the task, participants simply follow the target with their eyes. Accurate performance of this task requires subjects not only to make smooth, well-timed movements but also to anticipate the trajectory of the target. Unlike many of the tasks considered so far, the desired movements form a fixed sequence that can be learned to improve performance.

This task is frequently used to assess the motoric consequences of brain damage. Although levels of ability can vary significantly across different clinical populations, one can determine whether individuals are able to improve with repeated exposure to the predictably moving target. We are unaware of any studies testing participants with cerebellar damage on this task, perhaps because these individuals do not easily make the smooth tracking movements that are required during performance. In general, this problem can be addressed by slowing the movements of the target and evaluating individuals' ability to improve over time. Using such techniques, several researchers

have reported that damage to the basal ganglia impairs learning during rotary pursuit. We examine the role of the basal ganglia during the performance of predictable movement sequence in our discussion of the serial reaction time task.

The distortion and rotor-pursuit tasks demonstrate that individuals can modify internal models based solely on visual information. Other researchers have created experimental settings that emphasize proprioceptive feedback. Reza Shadmehr investigated the ability of subjects to learn to counteract predictable forces exerted on a manipulandum. In this task, subjects must grip a handle that is attached to a robot arm and attempt to move the handle to a specified position in space. As the handle is moved from its starting position, torque motors in the robot arm exert a force that is dependent on the velocity of the movement. When first performed, the added force dramatically distorts the movement trajectory so that the handle moves in a curved path toward the target location. During practice, the movements adapt to the force field and eventually form a straight path.

There are two possible ways that the brain might learn to produce straight movements in the force field produced by the robot arm: The muscles of the arm might increase their stiffness or the trajectory of the arm might be altered to compensate for the field. Because movements show an aftereffect when the force field is removed, the latter explanation appears to be correct. Thus, experimental participants must learn an internal model of the force field that is used to reprogram motor commands and adapt to the task demands.

The formation of the internal model appears to go through a consolidation process that occurs hours after exposure to the force field, even if the time is not spent performing the task. To demonstrate this, Shadmehr and colleagues used an interference paradigm in which, after a delay period, participants were trained on a second, orthogonal force field. When the interval between the two sessions was separated by more than 5 hr, internal models for both force fields appeared to be learned. However, if the interval was less than 5 hr, learning of the new field was slower and interference was observed when the participants were retested in the original force field.

A series of PET studies were conducted to identify the neural regions associated with the formation of the modified internal models as well as those involved in the consolidation process that allows for the long-term retention of these new models. During training, few systematic increases were observed as the participants became more adept at making movements in the force field. Neural activity decreased in right frontal cortex and the left motor cortex, likely reflecting the reduced error as the trajectories became more linear. Interestingly, when the participants returned to the same force field after a 5.5-hr delay, increases were observed in the left posterior parietal and premotor cortices and the right anterior cerebellum. Control participants, who performed the task in a novel force field after the delay, did not show these increases. These findings suggest that the identified regions play a critical role in the representation of the internal model, but only after consolidation has occurred.

The activity in the cerebellum is of particular interest because it indicates that the structure, along with the others, plays a role in the retention of the internal model. This hypothesis has received further support from the computational and imaging studies of a group led by Mitsuo Kawato. In their model, the cerebellum stores a large repertoire of internal models. These models not only define how we produce gestures such as reaching movements in an efficient manner but also underlie the transformations required when we interact with tools. They assume that the error signals provided by the climbing fibers are distributed widely during the course of a movement. As the error is reduced, the overall activity of the cerebellum will decrease. However, in the restricted region involved in representing a task-specific internal model, activity should increase. This general pattern—an overall decrease in cerebellar activity during learning with foci of increased activation—has been confirmed in an fMRI study. The fact that the changes are bidirectional also sheds light on the discrepancies found in the imaging literature concerning the role of the cerebellum in motor learning, with some studies emphasizing an increase in activity and others a decrease.

Imaging studies have also provided insight into the neural bases of the interference effects. Shadmehr measured brain activity during performance under a novel force field for two groups of participants, with the training occurring 10 min after the training with the original force field for one group and 5.5 hr after training with the original force field for the other group. The 10-min group showed marked increases in activity in the brain stem and ventrolateral prefrontal cortex compared to the 5.5-hr group. These regions did not overlap with those associated with learning the first field. Therefore, the researchers proposed that the prefrontal regions were suppressing the expression of the learned internal model for performance in the novel field. Such suppression was not required for the

5.5-hr group, presumably because the inverse model associated with the original force field had been consolidated.

In summary these sensorimotor tasks indicate that neural structures throughout the brain are recruited to encode novel motor behaviors. In addition to the cerebellum, the basal ganglia and cortical structures appear to play a role in the acquisition of new behaviors. Next, we examine some other motor skill learning tasks that have been shown to recruit multiple brain structures.

A. The Cerebellum and Sequence Learning

Skilled behavior rarely involves tasks that require a single discrete movement or the involvement of a single joint. Many of our skills are more complex, involving the successful integration of a series of movements into a coordinated action. One popular task for studying sequence learning is the serial reaction time (SRT) task. In this task, participants perform a sequential choice reaction time task. The stimuli can occur in a repeating sequence or in a random order. In the typical experiment, random probe blocks are interspersed with learning blocks, and after an extended training period on sequence blocks, the difference between the reaction times on the two types of blocks provides a measure of learning. Reaction times are faster on sequence blocks even when participants report being unaware of the presence of the sequence and perform poorly on tests designed to assess their explicit knowledge of the sequence.

Before discussing to the numerous neuropsychological and neuroimaging studies of the SRT task, it is useful to consider some behavioral findings that highlight important computational issues. First, though it is frequently used as a measure of motor skill learning, there is debate as to what information is actually encoded by participants. This controversy stems in part from the finding that sequence knowledge acquired during the SRT task transfers to different effector systems. That is, participants initially trained on the task with finger movements will exhibit substantial transfer when tested with another effector system, such as arm movements or vocal responses. This aspect of the learning can be contrasted with many forms of adaptation in which the learning is effector specific.

However, these findings do not imply that sequence learning is largely perceptual. There is also strong evidence that sequence knowledge transfers to novel sets of stimuli or—when more than one stimulus is used to indicate a response—new orders of stimuli, as long as the sequence of responses remains constant. Thus, it is widely assumed that learning in SRT task involves encoding some abstract representation of the response sequence, although what information is included in this representation remains to be determined.

It should also be noted that the timescale of the movement sequence acquired during the SRT task is considerably different from that of the movements optimized in adaptation and sensorimotor integration tasks. In these latter cases, the learned behavior consists of a continuous movement and typically lasts less than 1 sec, whereas the SRT studies involve linking together series of discrete movements that can be separated by 1 sec or longer. This relates to a second property of the SRT task that distinguishes it from the others. Responses made in the SRT task are typically composed of simple and distinct actions; the precise timing of movements and the coarticulation of multiple effectors are not required. That is, the keypresses used in the SRT task tend to be single finger flexions that do not require the coordination of other muscle groups. It is true that these simple movements must be linked to others in the sequence, but the timing of these actions need not be precisely controlled for successful performance of the task.

Finally, it is generally believed that, as with the mirror-tracing task, learning on the SRT task may occur implicitly (i.e., outside awareness). Of course, under certain conditions sequence knowledge may become explicit. Although the critical boundary conditions for these two forms of learning remain controversial, there is evidence that they reflect the operation of separate neural systems. Although both forms of sequence encoding are highly interesting, we focus on implicit learning because its properties more closely match what is generally meant by the term "motor skill learning."

With these considerations in mind, we discuss the question of what neural systems support learning during the SRT task and what computations these systems perform. In contrast to behaviors such as the adaptation of the VOR, much of our understanding of the brain structures supporting learning during the SRT task comes from whole brain imaging techniques such as PET and fMRI. These procedures have an advantage over neurophysiological techniques in that they allow researchers to observed changes in activity across the entire brain.

Changes in activation within the cerebellum have been reported during SRT learning, but the picture

painted by these findings remains somewhat unclear. There have been considerable differences in the locations of the identified foci within the cerebellum, which may stem from differences in the experimental procedures employed by the researchers. Interestingly, there is the question of what sorts of changes one might predict within the cerebellum as learning progresses. If one supposes that the cerebellum plays a greater role in producing the appropriate motor commands as participants learn the sequence, then an increase in activity should be observed in this structure; if one supposes that the cerebellum is fine-tuning motor representations that are housed elsewhere in the brain, then decreases should be associated with learning. Of course, both types of changes could occur, similar to what has been reported in studies of tool use. In fact, both increases and decreases in cerebellar activity have been observed as learning in the SRT task progresses. However, the dominant finding is that sequence learning is associated with decreases in cerebellar activation.

A more consistent picture emerges in neuropsychological studies in which patients with cerebellar lesions are tested on the SRT task. Across a number of studies, these patients consistently fail to exhibit evidence of learning, except when the sequence is composed of just a few elements. Interestingly, even when the patients were explicitly informed of the sequence in advance of the testing in one study, they still failed to exhibit faster reaction times on the sequence blocks compared to the random blocks. The fact that the patients' reaction times fail to decrease over either condition raises the possibility that a performance deficit may make it difficult to observe learning, were it to occur. However, that learning is observed with short sequences undercuts this explanation to some degree.

Whatever reasons emerge to explain the complex patterns of activity, it appears likely that the cerebellum may play multiple roles during the acquisition and production of complex behaviors. Based on the distribution of projections from the cerebellum to cortex and spinal tract, lateral portions of the cerebellum are generally assumed to be involved in motor planning, whereas medial portions are assumed to be involved with motor execution. Motor learning likely involves both these sets of processes.

B. The Basal Ganglia

Perhaps the most frequently identified structure in neuroimaging studies of the SRT task is the subcortical collection of nuclei called basal ganglia. Increases in the basal ganglia's activity have been observed during both implicit and explicit sequence learning, although researchers have tended to emphasize automatic processing when characterizing its computational role. They have done so in part because several lines of evidence indicate that the basal ganglia are a key component of a habit learning system, especially in relationship to the production of sequential movements.

Neurological disorders have also emphasized the role of the basal ganglia in motor control, particularly in the production of sequential movements. Two prominent degenerative disorders, Parkinson's disease and Huntington's disease, involve a loss of tissue in the basal ganglia. For the former, the disorder primarily involves a loss of dopaminergic neurons arising in the substantia nigra. For the latter, the genetic abnormalities result in extensive atrophy of the striatum, at least in the initial stages of the disease. Patients with either Parkinson's or Huntington's disease exhibit a learning deficit on the SRT task. In both disorders, the coordination problems are most pronounced during the production of sequential actions.

Kent Berridge and colleagues have taken a neuroethological approach to the study of sequential behavior, exploring the neural basis of grooming in the rat. This behavior is highly stereotyped, consisting of three distinct arm strokes arranged in an essentially fixed order. Although lesions of the cerebral cortex and cerebellum produce only transient disruptions in this behavior, extensive striatal lesions can produce a chronic impairment in grooming. Interestingly, the deficit is not manifest as a loss of particular component elements, nor are the elements produced in a shuffled order. Rather, following the lesions, the animals frequently fail to complete the production of the complete sequence.

Because the grooming sequence performed by the rat is thought to be innate, the basal ganglia's function in this example is likely unrelated to learning per se. Thus, although many researchers have emphasized that the basal ganglia are necessary for learning new associations between stimuli and responses (S–R associations), it appears that the structure is critical for performing behaviors that are well established in the organism's repertoire. Given that the component elements are still performed after the basal ganglia damage, one possible contribution of the basal ganglia may be scheduling to elements in the appropriate order. That the structure is critical for both the acquisition of new sequences and the performance of

old ones provides an intriguing clue to its computational role.

Finally, neurophysiological studies have revealed that neurons within the monkey striatum are sensitive not only to which response in a sequence is being performed but also to the context of that response. In other words, these neurons show increases in activity that are dependent on the particular response occurring in a particular sequence. For example, a monkey can be trained to perform two movement sequences, one consisting of response sequence ABC and the other consisting of response sequence CBA. A context-sensitive neuron may show increases in activity before response B in the sequence ABC but not in the sequence CBA. This pattern is consistent with the hypothesis that the basal ganglia have chunked a series of discrete responses into a sequence, a process that is essential for the automatization of skills. However, similar context dependency is also observed in the supplementary motor area (SMA), a cortical area that is reciprocally connected with the basal ganglia. These results suggest a variety of computational interpretations. One hypothesis is that as skill develops, the basal ganglia form a representation of the sequence, enabling a shift in the locus of control from the cortex to the subcortex. In contrast, it may be that the representation of the sequence remains cortical and that the basal ganglia provide a mechanism to rapidly progress through the series of gestures as the sequence unfolds. The latter hypothesis would be consistent with the grooming studies showing that the striatal lesions result in a failure to complete the sequence.

C. Comparing the Basal Ganglia and the Cerebellum

In evaluating the role of the basal ganglia and cerebellum in skill learning, it is useful to compare these two prominent subcortical structures in terms of both anatomy and physiology. Although the neural circuits of each are unique, there are correspondences between the features of the cerebellum and basal ganglia that warrant a brief description. First, both form loop-like circuits in which inputs from the cortex are processed and then relayed back to the cortex via the thalamus. Second, both have inhibitory projections to their output targets: The internal capsule of the globus pallidus inhibits the thalamic nuclei and the Purkinje cells of the cerebellar cortex inhibit the cerebellar nuclei. Finally, both the cerebellum and the basal ganglia appear to use a divergent–convergent architecture in which input signals are distributed across a vast range of neural networks before recombining into a more compact, topographic organization. In the basal ganglia, this occurs as small cortical regions project to many loci within the striatum, which in turn project reconverging output to focal regions with the globus pallidus. In the cerebellum, inputs from mossy fibers are distributed to and reintegrated by thousands of Purkinje cells. One possible function of this design is to allow for inputs to form associations with information from many sources.

Given these shared principles, it is important to consider some differences between the two structures. First, unlike the cerebellum, the basal ganglia do not have any direct efferent or afferent connections to the spinal cord and comparatively few connections to the brain stem. In this manner, the basal ganglia are positioned to modulate the cortical selection and instantiation of actions, whereas the cerebellum is likely involved in both the planning and execution of movements.

Second, the two subcortical systems seem to use feedback in a differential manner. As described previously, the climbing fiber inputs to the cerebellum are generally considered to be the source of an error signal that is minimized as learning progresses. In contrast to the prevalence of error signals in theories of cerebellar learning, models of basal ganglia have emphasized reward systems based on the dopaminergic pathways. Thus, whereas the cerebellum is assumed to learn based on negative feedback from error signals, the basal ganglia appear to learn using positive feedback from reward signals.

In trying to understand how the computational roles of the two structures may be distinguished, theorists have emphasized the role of cerebellum in movements that require precise coordination and timing: Eyeblink conditioning, VOR, and multijoint movements all necessitate that motor commands that are exactly timed in relation to environmental events or changes in the positions of other body parts. In contrast, the basal ganglia are generally associated with learning tasks in which a particular action or series of actions must be performed in a novel context. For example, in the SRT task, the simple finger press responses are easily made and benefits are likely accrued when the correct one can be anticipated.

Along these lines, Daniel Willingham has argued that the basal ganglia perform motor learning computations that are distinct from those tapped by sensorimotor integration tasks. Willingham evaluated

performance of individuals with Huntington's disease across a variety of motor learning tasks and found that the learning impairments appear limited to a subset of the tasks, including rotary pursuit, the SRT task, and a tracking task in which the target followed a fixed pattern. The patients showed normal learning on mirror tracing and on a tracking task in which the target moved randomly. Noting that the Huntington's patients' deficits appeared to be restricted cases in which the movements were predictable, Willingham proposed that the basal ganglia are critical for learning sequences of open-loop responses.

In this sense, the basal ganglia can be thought of as operating on higher level representations that involve selecting a particular goal. In contrast, the cerebellum has been shown to be required when fine-tuning is required in the motor programs responsible for accomplishing a set goal. These generalizations suggest that the basal ganglia are critical for SRT learning, whereas the cerebellum is involved in learning tasks that require precise movements and the on-line coordination of effectors with sensory input. That is, the basal ganglia support "knowing what to do" and the cerebellum underlies "knowing how to do" motor tasks. Computational models have generally been consistent with this view. The successive inhibitory processing stages of the basal ganglia, from striatum to globus pallidus and then from pallidus to thalamus, have been shown to offer the unique feature of operating as a winner-take-all process, a mechanism for selecting the appropriate output given a particular context. In contrast, the inhibitory output from the cerebellar cortex is well suited for appropriately tuning the output from the cerebellar nuclei.

Features of this proposed distinction are likely to prove useful, but in some cases it is almost certainly an oversimplification. For instance, this framework cannot easily account for the fact that patients with cerebellar lesions are most impaired on SRT learning. We would expect that the patients could learn the pattern, but that they would have difficulty in producing the coordinated sequence of actions. Perhaps the cerebellum does not distinguish between error signals associated with poor movements and those associated with meeting the task requirements (e.g., pressing the correct button).

Contrary to earlier views of brain function, the functional domains of both the cerebellum and basal ganglia are no longer restricted to motor control and motor learning. Both structures project to various association areas of the cerebral cortex, suggesting a role in higher level cognition. Moreover, imaging studies consistently reveal metabolic changes in these areas that cannot be accounted for in terms of the motor requirements of the tasks, and neuropsychological studies have found that lesions in both the basal ganglia and cerebellum can disrupt performance on a wide range of cognitive tasks. The manner in which these subcortical structures contribute to motor skill acquisition may prove useful in understanding how they influence learning more generally.

V. THE CEREBRAL CORTEX

Thus far, we have focused on the subcortical structures, the cerebellum and the basal ganglia, and their roles in motor learning. Although it is clear that the cerebral cortex is intimately involved in many aspects of motor and nonmotor learning, there is a comparative dearth of rigorous computation models regarding the role of cortical structures in skill acquisition.

There is abundant evidence that the cortical structures play a critical role in encoding new motor behaviors. For example, imaging studies of motor learning tasks, including the SRT task, have identified many activation foci in the frontal and parietal lobes. These regions include the primary motor cortex, the premotor cortex, the supplementary motor cortex (SMA), and anterior parietal cortex, all of which possess neurons whose activity is closely related to motor activity.

A. Motor Cortex

The primary motor cortex, Brodmann's area 4, is an obvious starting point. It sends direct projections to the spinal cord and is unique among cortical areas in that it possesses reciprocal connections with all of the premotor regions that project to the spinal cord. Stimulation of this cortical area produces overt movement and the neural activity is more closely related to movement parameters than other cortical motor structures. For these reasons, the motor cortex has traditionally been considered an area associated with motor output rather than performing the higher level operations associated with motor learning.

However, there is strong evidence that the physiology of motor cortex changes as new motor skills are developed. Across a variety of sequence learning tasks, motor cortex activity has been shown to increase during the course of training. Interestingly, these

changes appear to be strongest when the sequence knowledge is implicit; under conditions that favor the development of explicit knowledge, PET studies have failed to show learning-related changes in motor cortex.

One potential consequence of skill learning is that the motor fields of task-related neural ensembles are enhanced over the course of practice. Indeed, anatomical studies in the rat have shown a functional reorganization of the topographic map within motor cortex as a function of practice. The cortical region associated with the involved effectors becomes larger, paralleling learning-related changes that have been observed in various sensory cortices. The motor "receptive" field can also be measured in humans with transcranial magnetic stimulation (TMS). With this method, a large magnetic field is generated in a coil applied to the scalp and the underlying neural tissue is excited. When applied over motor cortex, discrete finger movements can be elicited. During SRT training, the region over which finger movements could be generated increased. Interestingly, as soon as the participants became aware of the sequence, the motor fields returned to their original size.

The previous results suggest an involvement of motor cortex during implicit sequence acquisition. However, these studies have all involved training that was restricted to a single session. Avi Karni and colleagues conducted a sequence learning study that spanned a much longer period. Participants were required to practice a series of finger movements over a period of several months. fMRI was used to measure activity within the motor cortex at various points during the training regimen, with the comparison made between epochs in which the participants produced the learned sequence and epochs in which the participants produced an alternative sequence. During the scanning sessions, the movements were paced so that the number of actual responses was equated for the two conditions. Nonetheless, after 3 weeks of practice (but not before), the trained sequence produced greater signals in primary motor cortex than an untrained control sequence. This increase was apparent despite the fact that the participants had full awareness of both the trained and control sequences.

There are at least two ways to reconcile the discrepancy between the results of this study and the SRT results. First, after many weeks of training, performance on the highly trained sequence may no longer depend on explicit representational systems. Indeed, at this point in training, the participants would show minimal cost on the sequencing task if required to concurrently perform a second, attention, demanding task. By this hypothesis, increases in motor cortex activity are again restricted to implicit sequence performance. Second, the increased signal may be a consequence of the pacing manipulation during the scanning sessions. By slowing down the production rate during the scanning session, motor cortex neurons may have remained active in anticipation of the next response. This latter hypothesis underscores a potential limitation with imaging studies of sequence learning. Neural activity may increase in an area not because there is a representational change within the area but because the input signal from an upstream neural region has become stronger.

Neurophysiological studies have also provided evidence for learning-related changes in motor cortex. Several groups of researchers have reported that the response properties of motor cortex neurons change as the animal learns to perform new tasks. These neurons may become more responsive to a given stimulus as the animal learns to associate it with a particular response, or they may alter their responsiveness to reflect the operation of a particular set of S–R mapping rules, regardless of the actual stimuli. Thus, it is unlikely that this region's computation is restricted to simple output properties.

B. The Premotor Cortex and SMA

Lying immediately rostral to the primary motor cortex is Brodmann's area 6. This secondary motor area is composed of two primary regions, the premotor cortex and SMA. Premotor cortex includes the lateral aspect of area 6, whereas SMA spans the medial aspect. Recent work has emphasized that there are likely many subareas within these areas; for example, four distinct motor areas can be identified within SMA. Premotor cortex and SMA are well positioned to modulate motor output given their projections to motor cortex as well as their reciprocal connections with prefrontal, parietal, and SMA regions.

Functional hypotheses concerning the role of premotor cortex and SMA have primarily been derived from physiological studies with primates. The relevant experiments have generally investigated neural activity during two types of tasks: learning to form novel associations between stimuli and responses and learning to perform novel sequences of actions. For the tasks involving the formation of S–R associations, the single-unit recordings have predominantly examined

the premotor cortex. However, for the tasks involving response sequences, the recordings have focused on the SMA.

Steven Wise and colleagues performed several experiments demonstrating that neurons in premotor cortex change their responses as monkeys learn to produce particular responses to specific stimuli. That is, these neurons become more responsive to particular stimuli as they become associated with particular responses, suggesting that they may contribute to the encoding of the S–R mapping. Other studies have recorded from neurons in motor, premotor, and parietal cortex, examining if the activity is associated with the response, the stimulus, or both. Neurons showing each of these types of response properties have been observed in all three regions. Nonetheless, there are differences in the proportions of types across the three regions, with the emphasis changing from stimulus linked to response linked between parietal and motor cortex. Thus, the functional differences within the cortex may be more quantitative than qualitative.

The emergence of a new motor skill results from the combination of basic motor elements into meaningful, sequential gestures. To type, we learn to chain together a set of simple finger and wrist movements. Moreover, to type a frequent word such as ''the,'' we encode a specific pattern of such movements. The gesture used to produce the ''e'' in this context is distinct from that associated with typing the word ''hose.'' Neurophysiologists have examined the context dependency of motor neurons to identify those that are involved in sequence learning and sequence performance. Jun Tanji and collaborators provided extensive evidence that SMA neurons exhibit such context dependency. For example, animals might be trained to press three horizontally aligned response keys in one of two sequences, left–right–middle or middle–right–left. Some SMA neurons will fire prior to the response on the right key in the first pattern but not the second; the opposite pattern will be observed in other neurons. Thus, the neural activity is associated with a specific movement (press right key) only when that movement occurs within a particular context or as part of the actions associated with a specific goal.

The association of SMA with sequential movements is one way in which this area appears to differ from motor cortex. Motor neurons show much less context dependency. Their response is much more closely tied to the individual movements that form the sequence rather than features of the overall sequence. A motor cortex neuron that responds when the right key is pressed will generally fire similarly within any sequence that contains this response.

An intriguing demonstration of the contribution of SMA to the organization of sequential behavior in humans comes from TMS study performed by Christian Gerloff and colleagues. With this technique, a coil capable of producing a strong magnetic field is placed on the skull. When engaged, the field induces widespread firing of the neurons in the cortical region lying under the coil, causing a disruption of the endogenous activity of this area. In a sense, TMS causes transient ''virtual'' lesions. In the study, TMS was targeted to disrupt activity in either motor cortex or SMA while people performed a series of finger movements. Motor cortex stimulation led to errors in the production of the ongoing movement: The selected finger might freeze or another finger might be activated and press an incorrect key. In contrast, SMA stimulation led to errors in subsequent movements. The participants reported losing track of their place in the sequence, of momentarily forgetting the series of associations. This dissociation is in accord with the hypothesis that the representation of SMA is not linked to individual movements but, rather, the sequence of movements that constitute the skill.

Neurons within premotor cortex also exhibit context dependency, responding during a particular movement, but only when that movement is performed within a specific context. However, there is one important difference between the sequential properties of neurons in premotor cortex and those in SMA/pre-SMA. The majority of premotor neurons fire when the movements are cued by external signals such as the illumination of a response key. In contrast, neurons in SMA fire most vigorously when the movements are internally generated. Thus, within either area, one will fire neurons that respond prior to the right keypress during the sequence left–right–middle. However, the typical premotor neuron will fire when the movements are signaled by an external cue, whereas an SMA neuron will fire when the sequence is being produced from memory.

The distinction between external and internal modulation of activity may offer clues to the functional specialization within secondary motor areas in terms of their contribution to motor learning. As noted previously, premotor cortex is critical for learning arbitrary associations between stimuli and responses. Externally cued movements by definition involve transforming stimulus information into the appropriate action goals (S–R translation), making such behaviors the purview of the premotor cortex. In

contrast, producing a complex sequential action likely requires that the successive movements be organized in relation to internal states, thus requiring the operation of the SMA.

From this perspective, it would be reasonable to expect both premotor cortex and SMA to be involved in the performance of skilled behaviors. Throwing a dart requires the sequential and exquisitely timed engagement of a specific set of muscles as well as the use of visual feedback to fine-tune the movements from one toss to the next. Indeed, many neuroimaging studies have observed activity in both regions over the course of learning. Although the pattern of activation in premotor cortex and SMA is frequently similar, dissociations have also been reported. Harri Jenkins and coworkers used PET to observe the neural correlates of sequential movements at different stages of learning. Unlike the serial reaction time task, these studies used a trial-and-error method for sequence learning. In each trial, the participant presses one of four keys. Following the response, a tone is generated indicating whether or not the selected response was correct. Learning progresses rapidly so that after about 10 min of practice, the sequence can be repeatedly produced without any errors. Paralleling these performance changes, the focus of cortical activity shifts from premotor cortex to SMA. If a new sequence is then introduced, premotor cortex is again engaged. As with the primate studies, these results emphasize that premotor cortex is most prominent when behavior is dependent on external sources of information; when performance can be guided by internal cues, SMA appears to become more important.

C. The Parietal Cortex

The final region to be considered in this article is the parietal cortex. Both neuropsychological and neurophysiological evidence suggests that the function of this region may be closely related to that of the premotor cortex. Extensive reciprocal connections exist between parietal and premotor cortex, and the two structures may work together to transform information about the environment into representations used for motor programming.

The classic neuropsychological syndrome involving the loss of skilled movement is apraxia, first characterized by Hugo Liepmann. The term likely includes a variety of syndromes, although a taxonomy of apraxic subtypes remains controversial. By definition, apraxia involves deficits in the production of movements, when the deficits cannot be attributed to problems in motor execution (e.g., paralysis or weakness), perception (e.g., agnosia), or comprehension (e.g., aphasia). Many investigators have limited the definition of apraxia to deficits in learned motor acts rather than novel behaviors. Apraxia can be summarized as a failure to translate goals into movements.

Apraxia can be observed following lesions of either cerebral hemisphere. In general, the deficits are most severe following left hemisphere lesions, with these patients exhibiting the disorder when using either the contra- or ipsilesional hand. Patients with apraxia resulting from right hemisphere lesions usually exhibit deficits only when using the contralesional side of the body. Within either hemisphere, apraxia has been associated with lesions of either the frontal or parietal lobes. Considerable neuropsychological research has attempted to dissociate apraxia stemming from parietal lesions from apraxia stemming from frontal lesions.

Kenneth Heilman and colleagues tested apraxic individuals with either premotor or parietal damage. They determined that although both groups demonstrated deficits in the production of learned motor behaviors, the group with parietal damage was also significantly impaired at discriminating well-performed and poorly performed actions in a task involving the perception of skilled actions. For example, the patients might not be able to judge which of three video clips depicted a person pantomiming how to use a key to open a lock. Thus, the more anterior lesions appear to disconnect the motor programs from the structures that are necessary for performance, whereas the more posterior lesions appear to disrupt the representations of the motor programs.

Neurophysiological studies have also compared how movements are represented in the frontal and parietal lobes. Much of this research has focused on "mirror neurons"—cells that become active both when a monkey performs a particular act and when the animal observes that same act being performed by another monkey (or person). The fact that the cells are responsive during both observation and performance suggests that they are coding abstract properties of the actions rather than particular motor parameters. Although the initial work on mirror neurons focused on premotor cortex, Vittorio Gallese recently reported that neurons in this inferior parietal lobe also respond during both performance and observation of goal-directed actions. It appears that the inferior parietal neurons may represent the motor acts at an even more abstract level than the premotor neurons in that the

former may be insensitive to the actual effectors used to achieve the goal. Thus, the parietal lobe may serve as a critical interface between perception and action. As the animal views the environment and identifies possible actions (e.g., obtain a banana), an abstract representation of the goal may be translated into an abstract action designed to achieve the goal, with the projections to premotor and motor areas serving to specify a particular reaching gesture.

To date, there has been little neurophysiological work addressing how the responses of neurons within the inferior anterior parietal cortex change during the course of skill acquisition. However, activation in this region is frequently observed during the encoding of new motor skills. For example, increases in area 40 are often reported during SRT learning, particularly in the left hemisphere. Several lines of evidence are consistent with the proposal that the activity in this region is associated with an abstract level of program. First, Scott Grafton and colleagues trained participants in a version of the SRT task in which the sequential key presses were made on a large keyboard that required whole arm movements. After training, the participants were transferred to a smaller apparatus on which the responses were made with finger movements. The behavioral data demonstrated that the sequence knowledge transferred from the arm movements to the finger movements. The PET data showed that whereas activity in the motor cortex was contingent on the effectors used to produce the sequence, activation of the inferior parietal lobe in the left hemisphere was contingent on the presence of the sequence regardless of how the movements were made.

VI. SUMMARY

In this article, we have focused on the neural regions involved in the performance of learned motor skills. The article is admittedly selective. For example, the cortical section did not discuss the contribution of prefrontal cortex despite extensive imaging evidence showing consistent activation in this region during motor learning and motor performance. However, the role of prefrontal cortex is likely not specific to motor learning. The prefrontal cortex is often characterized as performing operations relating to executive control rather than encoding motor skill for long-term storage. Thus, this structure may be most critical during the initial performance of novel tasks, playing a smaller role as learning progresses.

This article included a description of several cortical regions. Nonetheless, the emphasis was on the basal ganglia and cerebellum. The cerebellum and basal ganglia do not project directly to one another, and the cortex, particularly sites within the frontal lobe, is most likely the locus of interaction between the two structures. We characterized the cerebellum and basal ganglia as performing very different functions, with the cerebellum tuning motor commands in relation to sensory feedback and the basal ganglia scheduling action goals for motor processes to achieve.

Our focus on the subcortex reflects a bias in the literature: It is striking how many computational theories have been developed concerning the function of the cerebellum and basal ganglia, and that there is a dearth of such theories for cortical function, at least in terms of motor learning. Does this dominance of subcortical structures reflect limitations in our current understanding or does it indicate that the cerebellum and basal ganglia play a predominant role in motor learning? Although these possibilities are not exclusive, one factor to bear in mind is that the comparatively homogeneous architecture of the subcortical structures may make them excellent systems to model. That is, the connections within these neural structures follow a mostly linear arrangement, with neural signals moving along well-defined pathways. In contrast, the organization of the cortex appears much more complex, and the flow of information is much less easily determined. In short, the state of computational models may stem from general differences in the neural architecture of cortical and subcortical regions.

See Also the Following Articles

APRAXIA • BASAL GANGLIA • CEREBELLUM • CEREBRAL CORTEX • MOTOR CONTROL • MOTOR CORTEX • MOTOR NEURON DISEASE • MULTISENSORY INTEGRATION • NEURODEGENERATIVE DISORDERS • PARKINSON'S DISEASE

Suggested Reading

Heilman, K. M., Watson, R. T., and Rothi, L. J. (2000). Disorders of skilled movements. In *Patient-Based Approaches to Cognitive Neuroscience* (M. J. Farah and T. E. Feinberg, Eds.), pp. 335–343. MIT Press, Cambridge, MA.

Houk, J. C., Davis, J. L., and Beiser, D. G. (1995). *Models of Information Processing in the Basal Ganglia*. MIT Press, Cambridge, MA.

Imamizu, H., Miyauchi, S., Tamada, T., Sasaki, Y., Takino, R., Putz, B., Yoshioka, T., and Kawato, M. (2000). Human cerebellar activity reflecting an acquired internal model of a new tool. *Nature* **403,** 192–195.

Ivry, R. (1996). Representational issues in motor learning: Phenomena and theory. In *Handbook of Perception and Action, Vol. 2: Motor Skills* (H. Heuer, S. W. Keele *et al*., Eds.), pp. 263–330. Academic Press, London.

Karni, A., Meyer, G., Rey-Hipolito, C., Jezzard, P., Adams, M. M., Turner, R., and Ungerleider, L.G. (1998). The acquisition of skilled motor performance: Fast and slow experience-driven changes in primary motor cortex. *Proc. Natl. Acad. Sci. USA* **95,** 861–868.

Kawato, M., and Wolpert, D. (1998). Internal models for motor control. *Novartis Foundation Symp*. **218,** 291–304.

Keele, S. W. (1986). Motor control. In *Handbook of Perception and Human Performance, Vol. 2: Cognitive Processes and Performance* (K. R. Boff and L. Kaufman, Eds.), pp. 1–60. Wiley, New York.

Kim, J. J., and Thompson, R. F. (1997). Cerebellar circuits and synaptic mechanisms involved in classical eyeblink conditioning. *Trends Neurosci*. **20,** 177–181.

Mushiake, H., Inase, M., and Tanji, J. (1991). Neuronal activity in the primate premotor, supplementary, and precentral motor cortex during visually guided and internally determined sequential movements. *J. Neurophysiol*. **66,** 705–718.

Raymond, J. L., Lisberger, S. G., and Mauk, M.D. (1996). The cerebellum: A neuronal learning machine? *Science* **272,** 1126–1131.

Rosenbaum, D. A. (1991). *Human Motor Control*. Academic Press, San Diego.

Shadmehr, R., and Holcomb, H. H. (1997). Neural correlates of motor memory consolidation. *Science* **277,** 821–825.

Thach, W. T. (1996). On the specific role of the cerebellum in motor learning and cognition: Clues from PET activation and lesion studies in man. *Behav. Brain Sci.* **19,** 411–431, 503–527.

Willingham, D. B. (1998). A neuropsychological theory of motor skill learning. *Psychol. Rev*. **105,** 558–584.

Wise, S. P., and Murray, E. A. (2000). Arbitrary associations between antecedents and actions. *Trends Neurosci*. **23,** 271–276.

Movement Regulation

JOHN F. SOECHTING and MARTHA FLANDERS
University of Minnesota

GLOSSARY

feedback A signal derived from afferents used to control the movement and correct for errors in an ongoing movement.

feedforward A signal derived from efferent commands used to predict the movement that will ensue.

kinematics Dealing with movement; includes parameters such as position, velocity, and acceleration.

kinetics Dealing with the forces that generate a movement.

physical plant The mechanical properties of muscles and of the skeletal system.

Movement regulation is thought to involve a combination of feedback and feedforward mechanisms. The nature of the feedforward and the feedback control depends crucially on the characteristics of the motor plan and the properties of the physical plant that is being regulated. This article reviews current concepts of motor planning, the properties of the musculoskeletal system, and control systems models for movement regulation. The focus is on the regulation of limb movements, although many of the concepts apply equally to the control of eye movements.

I. OVERVIEW

At the beginning of the 20th century, Charles Sherrington put into place a conceptual framework concerning movement regulation and initiation that held sway for a considerable amount of time. The central component of this framework was the reflex—sensory input leading to the initiation of the movement and also providing for its ongoing control by means of feedback. The most studied of these reflexes was perhaps the stretch reflex, but a large variety of other postural reflexes involving somatic, vestibular, and even visual sensory afferents were also elaborated over the course of time. In this viewpoint, reflexes were not obligatory; instead, they could be gated and regulated by higher centers in the nervous system, thus providing for the idea of a hierarchical organization. Another important component of this framework was the concept of synergy—a group of muscles acting together in a common purpose in a particular movement and being controlled together as a unit.

In this conceptualization, limb motion from one posture to another involved two opposing groups of muscles, agonists and antagonists, each activated as a unit. Another important discovery, developed by Elwood Hennemann in the middle of the 20th century, provided an explanation for the orderly recruitment of individual motor units within the agonist and antagonist muscles. Specifically, according to Hennemann's size principle, motor units are recruited in a stereotypic fashion, according to their size. The pattern of muscle activation producing a limb movement was thought to be quite simple. There is an initial burst of agonist activity, arising from descending commands and providing for the movement's initiation. This is

followed by a burst of antagonist activity and a second (smaller) burst of agonist activity, providing for the movement's arrest and stabilization and arising at least in part from the stretch reflex.

The primacy of reflexes in controlling movement was not acknowledged universally, however. For example, although some investigators initially thought that locomotion was the result of a chain of reflexes, it was ultimately demonstrated that the locomotor pattern of alternating, rhythmic activation of flexors and extensors of each of the limbs was the result of a central pattern generator. These investigations also showed that this central pattern generator for locomotion was located in the spinal cord and that rhythmic motion could be elicited from deafferented preparations (i.e., in the absence of sensory feedback).

Thus, there were two competing viewpoints. According to one, movements were regulated essentially and importantly by feedback mechanisms dependent on sensory information generated by the movement. A diametrically opposite position held that movements were accomplished open loop in a feedforward fashion, and that the delays in feedback loops were too long for afferent information generated during the movement to modify the ongoing trajectory.

The conceptual framework from which we view movement regulation has changed dramatically in the past 20 years, but feedback and feedforward regulation remain central to the issue. One major change is the recognition that movements cannot be controlled by feedback alone and that they also cannot be performed accurately using simply feedforward. Thus, movement regulation involves a hybrid of feedforward and feedback controllers.

Furthermore, our concept of what needs to be regulated by the central nervous system has become more sophisticated. Thus, it is now clear that the physical properties of the plant—the muscles and the skeletal system—introduce appreciable complexities. Similarly, movement regulation usually involves the synthesis of sensory information derived from different modalities, such as vision, somatosensation, and the vestibular system. Sensory information from these different modalities is combined into a common frame of reference, introducing additional computational burdens. Finally, there is a growing appreciation that the control system regulating movements is not "hard-wired" but, rather, that it is subject to modification by experience and learning. We discuss each of these points in the following sections.

II. THE PHYSICAL PLANT

A. Kinetics and Kinematics

Kinetics refers to the forces (or torques) that generate a movement, whereas kinematics refers to a description of the motion—position, velocity, and acceleration. Kinematics and kinetics are related by Newton's second laws of motion. When the motion is restricted to a single joint, this relationship is simple:

$$F = ma \text{ or } T = I\alpha$$

where F is the force, m is the mass, and a is the object's acceleration. Similarly, for rotational motions, T is the torque, I is the moment of inertia, and α is the angular acceleration. Now consider the problem of predicting the movement that a particular force will generate. The solution to this problem is simple because position (or angular displacement) can be found by integrating force (or torque) twice. It is also easy to compute the force needed to produce a given movement by differentiating the desired linear displacement $p(t)$ [or the desired angular displacement $\theta(t)$] twice.

Since it is known that neural circuits can perform the mathematical equivalents of integration and differentiation, controlling simple movements does not appear to be particularly challenging. Furthermore, it was thought that more complicated movements could be generated by simply combining several simple movements. Thus, the distinction between kinematics (movements) and kinetics (forces) becomes academic; accelerations and forces are proportional to each other.

This is not the case when a movement involves rotation at more than one joint—for example, an arm movement that involves the shoulder, the elbow, and the wrist. Consider a simple example in which the arm is in the horizontal plane and is supported against gravity. Also imagine that a single shoulder muscle is activated, generating a torque about the shoulder joint. This torque will produce an angular acceleration at the shoulder. However, if the elbow and wrist joints are free to rotate, the acceleration at the shoulder will also impart a rotation at the elbow and wrist joints. Similarly, contraction of a pure elbow or a pure wrist muscle will affect the shoulder joint, changing the motion of the humerus, the limb segment proximal to the elbow. Torque is proportional to the angular acceleration at each of the joints in a skeletal linkage; conversely, the angular acceleration at one joint is proportional to the torques generated by muscles acting at each of the joints.

The equations relating kinetics and kinematics for multijoint motion also include terms known as Coriolis and centripetal accelerations. These terms are related to the angular velocities of the joint motions. Furthermore, the inertia of a linkage such as the arm depends on the angles of the joints in the linkage. Thus, torque is generally related to angular acceleration, angular velocity, and angular position. This relation is illustrated in Fig. 1. The relation between angular acceleration and torque is shown to depend on several parameters: the inertia of the limb, its angular velocity, and its posture. Posture enters into the equation in two ways: It affects limb inertia and it also determines the gravitational loads that need to be opposed.

The crucial point that emerges from this picture is the following: Inertia, posture, and velocity all vary during a movement. Therefore, if the motion that ensues from a torque profile is to be predicted, it is necessary either to measure each of these parameters accurately or to be able to predict them accurately. (Note that in principle, a single-joint movement could be performed accurately knowing only the amplitude of the required movement because the inertia of one limb segment does not change with time.) Sensing the position and velocity of a movement involves time delays. Consequently, parameters entered into the equation of motion will be in error because they are based on the state of the system in the past. Similarly, estimating position and velocity from the torque commands will also introduce errors arising, for example, from uncertainties in estimating the starting posture of the limb.

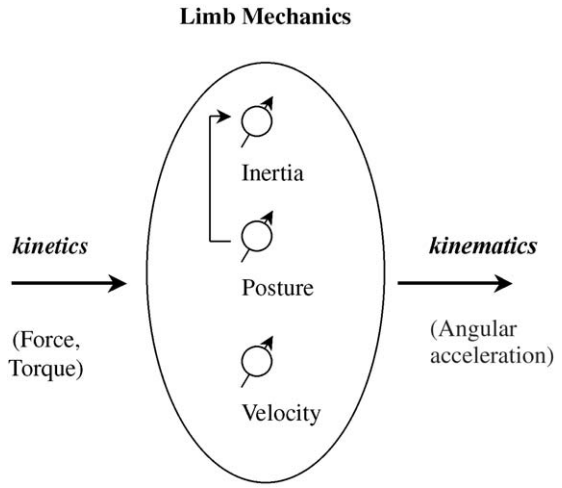


Figure 1 Relationship between kinetics (forces or torques) and kinematics (acceleration) as determined by limb mechanics. For a multijointed limb, the equations relating kinematics to kinetics include terms that depend on positional and velocity parameters, which change over time. The inertia of the limb also depends on posture and consequently also changes as the movement progresses.

Gravity is ubiquitous and gravitational torques also depend on the posture of the limb but not on the speed with which the limb moves. Thus, the control of a limb movement would need to account for the effects of gravity as well. In the mid-1980s, John Hollerbach made an important observation that provides for a potential simplification of the control of movement in a gravitational environment. He showed that despite all the apparent complexity of the relationship between torque and angular acceleration, the equations relating these parameters scale with the speed of the movement. In other words, movements of different speeds can be generated by taking one template of torques and scaling it in amplitude and in time. Since the gravitational torques do not depend on the speed, movement control can be simplified by separating the controller into two components: a postural component that does not depend on speed and a movement component that is speed dependent. Subsequent investigations have provided experimental support for this supposition.

The laws of motion relating kinematics and kinetics have other implications. First, the direction of force and the direction of movement generally do not coincide. Consider the following example. Ask a subject to exert an isometric force with the hand in some particular direction against a resistance, and then release the resistance. The hand will begin to move in a direction that need not coincide with the direction in which the force was exerted. Second, because of the interactions between the motions of the various limb segments, the sign of the torque at a particular joint need not be the same as the sign of the angular acceleration at that joint. For example, elbow extension may require elbow flexor torque. Consequently, eccentric contractions (i.e., activation of muscles that lengthen during a movement) are not uncommon. Movements depart in the correct direction, with the activation of the appropriate muscles predicted on the basis of mechanics. Accordingly, the central nervous system (CNS) must take these factors into account in specifying which muscles should be activated.

B. Muscles

The mechanical properties of muscles also need to be taken into account by the CNS. Since the work of A. V.

Hill in the 1930s, it has been appreciated that muscle force depends on muscle length (the length–tension relation) and on the rate at which muscle length changes (the force–velocity curve). Tendon compliance also affects the dynamical relation between muscle activation and force generation, and it is an important factor in mechanical models of muscle, such as those developed by Felix Zajac. It is also well recognized that the moment arm of a muscle can change with changes in the limb configuration. Since muscle torque is equal to muscle force times moment arm, changes in a muscle's moment arm will also affect the amount of torque generated by a specific level of muscle activation.

Most muscles generate torque about more than one axis. For example, biceps brachii acts as shoulder flexor and as an elbow flexor and supinator. Similarly, anterior deltoid acts as a shoulder flexor and adductor. One can always define a torque axis, a vector defining the mechanical action of that muscle. The question arises, do these actions depend on the posture of the limb? In other words, does the torque axis remain fixed as limb posture changes? If the answer is the affirmative, one can imagine two possibilities: (i) The torque axis remains fixed relative to the orientation of the distal segment, and (ii) it remains fixed relative to the proximal segment. For example, considering deltoid, one could imagine that the mechanical action was fixed relative to the trunk (proximal segment) or that it was fixed relative to the humerus (the distal segment). In fact, the answer is neither, and torque axes depend on limb posture in a manner that is intermediate to the two possibilities just outlined. The axis moves along with the distal limb segment, but not by as much as the limb segment moves. The postural dependence of the mechanical actions of muscle reinforces the conclusion drawn in the previous section: The CNS needs to take into account posture and rate of change of posture in planning and regulating movement.

Although the postural dependence of the mechanical actions of muscles is not negligible, it need not be overly complicated. For example, for shoulder muscles, simple linear models relating the directions of the torque axes to the orientation of the humerus could adequately account for the experimental data. Muscles are customarily modeled by assuming that their forces are exerted in a straight line from an effective origin to an effective insertion. (The effective origins and insertions need not coincide with the anatomical origins and insertions.) These "rubber band models" are indeed adequate to account for the actions of shoulder muscles, even those that have widely distributed origins such as the heads of deltoid. Thus, it appears that fairly simple models can be used to account for the mechanical actions of muscles and for their postural dependence.

As noted previously, the initial conception of movement regulation was that for any particular movement, there were two distinct groups of muscles —agonists and antagonists. Muscles within these groups would act as synergists (i.e., they would be activated synchronously). Initially, it was also thought that this grouping would hold for all movements [e.g., a classification of muscles as flexors and extensors (or gravity vs antigravity)], and that synergies would be hardwired. This viewpoint was subsequently modified to account for the fact that a pair of muscles could be synergists for one movement and antagonists for another. Furthermore, the need for synchronization was also abandoned and it was posited that a group of muscles belonging to a synergy could be activated in a fixed, sequential order. In this concept of "flexible synergies," the relative timing of the activation of different muscles would be fixed, at least for a broad class of movements.

A detailed investigation of the patterns of muscle activation for arm movements in different directions has shown that this is not the case. Although the basic "three-burst" pattern of electromyographic activity does hold for multijoint movements as well as for single-joint movements, the timing of the bursts in individual muscles is asynchronous. Furthermore, the relative timing of muscle bursts, among the various shoulder and elbow muscles, varies with the direction of movement. For any given muscle, there is an orderly progression of the timing of muscle activation with movement direction, but this pattern is different from muscle to muscle. Accordingly, it is not possible to define even pairs of muscles that are synergists, in the strict sense of this word. However, these studies did provide support for the idea introduced by Hollerbach in the sense that it was possible to identify for each muscle a "tonic" pattern of activity related to static postures and counteracting gravity and a "dynamic" component that scaled with the speed of the movement.

Thus, in general, it does not appear that control is simplified to two groups of muscles, acting as agonists or as antagonists for a given movement. Is it possible to conceive of movement regulation on a muscle-by-muscle level? It has long been known that some muscles are compartmentalized, with muscle fibers having mechanical actions that differ from compartment to compartment. Nevertheless, it was supposed

that, at least within each compartment, motor units were recruited in an orderly fashion, according to the size principle, and that except for this proviso motor units would be recruited similarly and simultaneously. Recent studies do not support such a simplifying assumption. Instead, they indicate that the control and activation of motor units is effected in a distributed fashion.

This assertion is based on results of studies of the recruitment of single motor units of elbow and shoulder muscles, primarily biceps and deltoid, under isometric conditions (Fig. 2). Under isometric conditions (as well as during reaching movements), muscles generally exhibit "cosine tuning." When force is exerted in a particular direction, muscle activation is greatest. When the same amount of force is exerted in a different direction, the amount of muscle activity decreases in a manner that is approximated by a cosine function. Thus, muscles can have a "best direction." (Sometimes, the tuning curve shows two peaks and thus there are two best directions.) As shown in Fig. 2A, this tuning property is also obeyed by single motor units. However, the best directions of individual motor units in one muscle do not coincide but are dispersed over a wide range of directions (Fig. 2B). Furthermore, there is no evidence for a clustering of these best directions; thus, the results cannot be explained by assuming muscle compartmentalization.

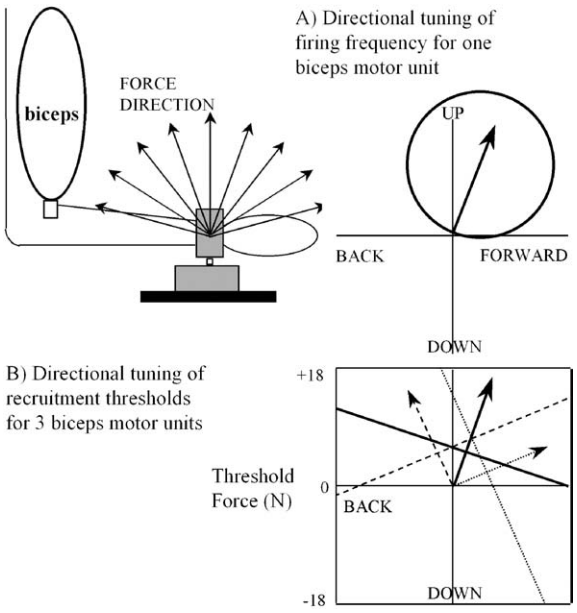


Figure 2 Directional tuning of motor units. (Top left) A typical experiment used to define directional tuning of motor units. The arm is maintained isometrically and force is exerted in a variety of directions. For each direction of force, firing frequency and the threshold of recruitment of single motor units are determined. (A) The typical tuning curve, with the distance from the origin to the edge of the closed curve representing firing frequency when the same amplitude of force is exerted in different directions. The arrow denotes the best direction for this motor unit. (B) The thresholds for recruitment of three motor units. Note that the threshold for each unit is defined by a straight line (equivalent to cosine directional tuning in A), but that the slope of the line is different for each unit. The arrows show the best direction for each unit. Note that the recruitment order of these three units depends on the direction in which force is exerted.

The threshold for recruitment of a motor unit is lowest when force is being exerted in that motor unit's best direction. If the best directions of motor units vary over a fairly wide range, as was found for the muscles studied so far, then there is no fixed recruitment order of motor units within a given muscle because there will be a different recruitment order for every force direction (see the threshold levels for the three motor units in Fig. 2B).

In summary, the experiments discussed in this section lead to two main conclusions. First, in order to predict the forces and torques developed by activation of any muscle, the motion and posture at each of the joints must be sensed (or predicted accurately) throughout the movement. Second, neural control of the musculature is not exerted over groups of muscles (synergies) or even at the level of individual muscles. Rather, this control is distributed, with individual motor units within a muscle having individual tuning characteristics.

III. MOVEMENT PLANNING

A. Robotic Models

Computationally, the control of a robotic arm by means of a computer and the control of a biological arm by the brain share many similarities. Accordingly, investigators studying the neural control of limb movements have often attempted to use the algorithms that have been used successfully by roboticists as a template for understanding biological motion. In fact, robotic models of movement planning have had considerable influence on models for neural control of movement.

Computer algorithms for movement planning evolve in a serial fashion, with one stage devoted to movement kinematics and a second stage devoted to movement kinetics. Consider the kinematic aspects of movement planning first. In this module, the first step

is to define the location of the target of the motion. This specification is said to be in extrinsic (or task) coordinates; for example, the x, y, and z coordinates with reference to some origin. Generally, the extrinsic coordinates do not define a unique configuration of robot or human arms because these arms have more than three degrees of freedom. For example, the human arm has 4 degrees of freedom, counting only the shoulder and elbow, and 7 degrees of freedom if the wrist is included. Therefore, there is not a unique solution; more than one posture is compatible with the target location. A unique solution is imposed by introducing some constraints. For example, it could be desired that the limb posture be as far away as possible from the limits of motion at each of the joints, or it could be desired to limit the amount of motion at each joint.

Two other things also need to be specified in such robotically inspired models before the movement can be constructed: the spatial and the temporal evolution of the movement. Thus, it has been suggested that movements are planned to follow straight lines with a velocity profile that is bell shaped. According to this suggestion, movement planning would involve the following sequence of steps, organized and executed serially: localization of a point in extrinsic space, specification of a rectilinear hand trajectory, and the computation of the joint angles (in intrinsic space) that correspond to each point in extrinsic space. As stated previously, there is not a unique set of joint angles for any given hand position, and a unique solution is obtained by invoking kinematic constraints.

The last step in these models of movement planning involves kinetics. Specifically, necessary joint torques are derived from the desired kinematics (joint angles). This step is commonly referred to as inverse dynamics (Fig. 3) because the forces are derived from motion rather than the opposite, as would be implied by cause and effect. For a biological system, there would be one last step—namely, the partitioning of the torque across the redundant actuators (the motor units), taking into account the musculoskeletal properties described in the previous section.

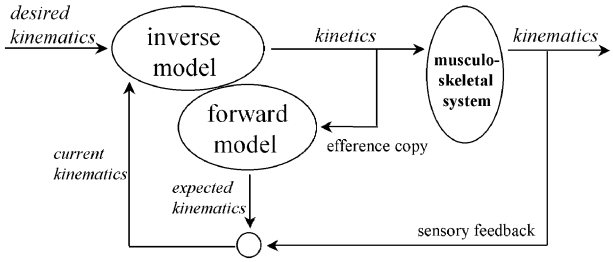


Figure 3 Hybrid model of feedforward and feedback control. The *desired kinematics* of a planned movement are transformed into motor commands (*kinetics*) through the inverse model. The motor commands are transformed by the musculoskeletal system into the actual movement *kinematics*. These are sensed by afferents to provide sensory feedback. An efference copy of the motor commands is transformed into the *expected kinematics* via a forward model. The difference between the expected kinematics and the actual kinematics (provided by sensory feedback) provides an error signal used to update the motor commands.

B. Trajectory Planning of Human Limb Motion

The robotic model has been influential in guiding experimentation and it has also received considerable support. For example, it is generally found that for arm movements from one point in space to another, the wrist follows a path that is close to straight and the wrist's speed is well approximated by a bell-shaped profile.

Furthermore, the results of psychophysical studies in humans and electrophysiological studies in nonhuman primates have been interpreted from this perspective. For example, there has been considerable work dealing with errors in pointing to targets in three-dimensional space. In these experimental paradigms, a target is presented visually or proprioceptively (by moving the arm into a given posture), and the subject is instructed to move the arm to the target in the absence of visual guidance. Both constant and variable errors in this type of task are usually very reproducible, and these errors have been interpreted as arising from errors in kinematic coordinate transformations. Specifically, it has been proposed that errors arise because of approximations in transforming target location in extrinsic, eye-centered coordinates into intrinsic, shoulder- or hand-centered coordinates.

Also, single unit recordings from parietal and frontal cortical areas have been related to kinematic parameters defined in different frames of reference. In general, cortical neurons whose activity is related to limb movement are tuned to the direction of movement (a kinematic parameter). Moreover, in some areas, direction is defined in eye-centered coordinates (i.e., a direction relative to the foveal direction of gaze). In other instances, direction appears to be specified in arm-centered coordinates (i.e., a direction relative to the orientation of the arm).

Thus, considerable evidence supports the first stage of the robotically inspired models—a transformation of a specification of target location in extrinsic space (or, more precisely, in retinotopic coordinates) into a specification of target location in joint coordinates. The question then arises: Is there also a kinematic plan of the trajectory? Evidence has been offered in support of such a supposition. Essentially, the velocity profiles of movements are highly repeatable and predictable. Specifically, if movements are relatively straight, bell-shaped (unimodal) velocity profiles are observed. If a curved trajectory is produced intentionally during drawing movements or when a curved path is necessary to avoid an obstacle, there is a precise relation between speed and the trajectory's curvature: The smaller the radius of curvature, the slower the speed. Taken at face value, these observations imply a precise plan of the moment-to-moment evolution of a movement's kinematics.

However, such a kinematic plan of movement does not require that kinematics be determined first without regard to the kinetic requirements. In fact, extensive modeling studies undertaken by Mitsuo Kawato and colleagues led to the opposite conclusion. They began with the observation that wrist trajectories for pointing movements are usually gently curved, and they were able to predict this curvature by assuming that the hand followed a path according to a "minimum torque change" criterion. Thus, kinetic constraints determined the kinematic plan of movement. Our own studies on limb posture at the end of pointing movements also support this interpretation. When subjects made arm movements from different starting locations, the arm's posture at the target depended on the starting location. This variation in the final posture could be predicted by invoking a kinetic criterion, one related to minimizing energy expenditure during the movement.

Thus, it seems clear that movement trajectory is regulated, even when it does not need to be (such as during pointing movements in the absence of obstacles). However, it is not clear that the observed regularities occur because a particular trajectory is planned explicitly, since they could also occur implicitly from regulation of the movement's kinetics.

IV. FEEDBACK REGULATION

As mentioned previously, feedback regulation of movement was a central component of the framework developed by Sherrington and followers. As is well-known, the stretch reflex involves monosynaptic excitation from primary muscle spindle afferents (as well as disynaptic inhibition arising from Golgi tendon organ afferents) onto the homonymous muscle. It has long been recognized that this arrangement provides for a regulation of muscle length through spindle feedback and for a regulation of force by means of tendon organ afferent feedback. Furthermore, since stiffness is the ratio of length to tension, the combination of length and force feedback could provide a means for stiffness regulation. Thus, as has long been recognized, this reflex potentially provides for feedback regulation of posture. In the 1950s, it was postulated that it could also provide a mechanism for a feedback control of movement. The fusimotor innervation of muscle spindles (from γ motoneurons) can change the spindles' rest length. Accordingly, it was hypothesized that the gamma innervation provided a reference length (the desired kinematics) and that spindle afferents provided an error signal proportional to a deviation from the desired kinematics.

Thus, in this classical view, the stretch reflex has the possibility of affording movement control on a muscle-by-muscle basis. Whether or not it does so effectively depends on the reflex gain. This value has been notoriously difficult to estimate because of the inherent nonlinearities in muscle mechanics. However, the best available estimates assign a relatively modest value to the reflex gain. Furthermore, there is good experimental evidence that feedback regulation is not on a muscle-by-muscle basis. As mentioned previously, the control of motion in a limb with two or more joints often involves eccentric contractions of muscles that are agonists for the movement. Similarly, when perturbations evoke rotations at more than one joint, reflex excitation of muscles that shorten consequent to the perturbation has been observed, contrary to the classical view. Spindle afferents are known to project not only to motoneurons of homonymous and synergistic muscles but also, either directly or indirectly, to motoneurons of muscles spanning distant joints. Thus, the reflex actions observed when a perturbation affects more than one joint could be purely spinal. However, there is also experimental support for "long-loop reflexes" involving supraspinal structures, including the cerebral cortex and the cerebellum.

In the classical view, feedback regulation via the stretch reflex would compensate primarily for external perturbations, but it is equally conceivable that feedback could compensate for inaccuracies in the motor

commands sent to the muscles. In fact, this supposition has experimental support. The effective moment of inertia of the arm varies with the direction of arm movement, and Claude Ghez and colleagues suggested that the initial descending motor commands do not take into account this inertial anisotropy. For movements of equal amplitude, the initial acceleration of the arm is largest for movements in directions for which the effective inertia is smallest. Nevertheless, the final amplitude of the movement is equally accurate for all directions. These investigators suggested that this was achieved by means of feedback compensation, a conclusion based on the additional observation that movement amplitude in deafferented patients varies in proportion to the variation in the initial acceleration of the arm. In other words, these patients do not exhibit any compensation for inertial anisotropies.

A final point should be made concerning feedback regulation of movements. The previous discussion considered the effects of kinesthetic feedback on the control of the evolving movement (i.e., an on-line regulation of the movement). There is also evidence that afferent inflow of information during one movement alters the execution of a subsequent movement. A simple, everyday example illustrates this point. Consider lifting a suitcase of unknown weight. If the suitcase is much heavier than it appears, the movement is likely to be hypometric, or it may even be aborted. However, the next attempt at lifting the suitcase will most likely be successful, implying that the motor commands have been modified based on previous experience. Visual feedback can also provide both on-line and trial-to-trial regulation.

V. CONTROL SIGNALS FOR MOVEMENT REGULATION

A. Feedback and Feedforward

There is consensus that neither purely feedforward, open-loop control nor purely feedback control is adequate for the regulation of movement in a biological system. Feedforward control alone cannot work because it requires that all the parameters of the system (its inertia, the starting posture, muscle mechanics, etc.) be known with a precision that appears to be beyond the realm of biological sensors. Similarly, pure feedback control is inadequate because the feedback gain is low. If the feedback gain were higher, the inherent time delays would lead to instability.

Thus, movement regulation appears to involve a hybrid of feedforward and feedback. In this conceptualization (Fig. 3), the feedforward pathway involves a model of the inverse dynamics of the limb that is a prediction of the forces that would be required to generate a desired movement. Thus, the inverse model transforms desired kinematics into motor commands (kinetics) acting on the musculoskeletal system to generate movement (kinematics). If the inverse model is accurate, the actual kinematics will match the desired kinematics. As stated previously, the inverse model will never be completely accurate, and accuracy is achieved by means of feedback.

In Fig. 3, the sensory feedback signals the actual kinematics resulting from the motor commands. These are compared to the expected kinematics, obtained using an efference copy of the motor commands. This comparison requires a transformation from the kinetic signal carried by the efference copy to kinematics. Because the transformation is in the forward, causal direction from kinetics to kinematics, it is called a forward model. The feedback regulation would then provide for a comparison of the expected motion, derived from the forward model, with the actual motion that is sensed by various afferents. This error signal could be used as an input to the inverse model to update the motor commands.

In this scheme, a mismatch between the predicted motion and the actual motion would have two effects, occurring on two different timescales. First, the motor commands would be modified, through the inverse dynamic model, to correct the ongoing motion. Second, if errors persist over many trials, the inverse dynamic and the forward dynamic models would be modified so as to bring their predictions into better accord with the actual performance.

Note that this hybrid scheme bears some resemblance to the robotically inspired scheme for movement planning outlined previously. Recall that this scheme involved kinematic stages in which ultimately the desired trajectory was specified in terms of the motion of each of the joints. A second stage converted this desired trajectory into the torques at each of the joints (and forces at each of the muscles) required to produce the requested motion. In terms of the hybrid model just described, this stage corresponds to the feedforward, inverse dynamics model part of the process. What has now been added is a feedback controller, correcting for errors inherent in the feedforward stage.

B. Anatomical Correlates and Distributed Processing

The hybrid scheme described in the previous section appears to be an attractive hypothesis, but is there any evidence that it represents the manner in which the brain regulates movements? It is a truism that the brain contains an inverse dynamic model. Consider a pointing movement to a target. Clearly, the input to this movement is information on the retina (in kinematic variables) and the output is activation of motoneuron pools generating forces (kinetics). Somewhere in between is a transformation from kinematics to kinetics.

Can one be more precise than that? Can the inverse dynamics model be localized to one particular brain structure? There is no good answer to this question, partly because there is no clear agreement concerning the parameters that are encoded by neural activity in major structures of the motor system. Consider, for example, the primary motor cortex. Almost since the time that the role of the motor cortex in generating movement was first appreciated, a debate has raged: Does motor cortex control movements or muscles? In other words, does motor cortical neuronal activity encode kinematic parameters (movements) or does it encode kinetic parameters (muscle forces)? Depending on the answer, the inverse dynamic model would be located either before or after the motor cortex, or it could in fact be within the motor cortex.

Unfortunately, there is no clear answer to this question. For limb movements in three-dimensional space, motor cortical neurons are tuned to the direction of the motion, as are neurons in parietal cortical areas and the cerebellum. At first glance, this might be taken to suppose that these structures encode kinematic parameters. However, kinetic parameters, such as torque or muscle activation patterns, also show directional tuning, and therefore this observation *per se* does not resolve the question. One way to resolve the question is to record neuronal activity for the same movement in the presence of different loads (i.e., the same kinematics but different kinetics). Such experiments have found a continuum of responses (even in one brain area): Some neurons are related only to kinematics, whereas others depend strongly on kinetic parameters as well.

An alternative approach to resolve the question takes advantage of the finding that the mechanical actions of individual muscles change in a predictable manner with changes in limb posture. For example, the directional tuning of wrist muscles for flexion/extension and abduction/adduction depends on the extent of forearm pronation/supination. When the directional tuning of motor cortical neurons in the wrist area was examined, it was found that some changed in parallel with the directional tuning of the muscles (suggesting a coding of kinetics), others changed in parallel with the changes in posture (suggesting a kinematic code), and yet others showed changes that were intermediate between these two.

Thus, it appears that the robotic models, with transformations from kinematics to kinetics and vise versa, do not show any simple isomorphisms to neural activity in different brain areas because a variety of parameters are found to be encoded in any one area. Nevertheless, it is possible that the conceptual scheme is reasonable but that the processing stages are distributed throughout and over many brain areas. However, it is equally possible that the models that have been proposed are inappropriate. Certainly, in the few instances in which behaviors are clearly understood in terms of neural processing, it has been possible to define distinct stages of transformations from sensory to motor domains and to associate neural structures with each of these stages. Some of the examples that can be cited are the processing of binaural sounds for sound localization, the jamming avoidance response in electric fish, and saccadic control of eye movements. Each of these examples involves processing in subcortical structures; processing of information in cortex may be inherently more distributed.

C. Learning and Modularity

No matter what the answer to the question posed in the preceding section turns out to be, it seems certain that the neural regulation of movement involves inverse dynamic and forward dynamic models, be they distributed or implemented in well-defined anatomical loci. Furthermore, it seems certain that these models are shaped by experience and by learning. This question has received considerable attention in recent years, and it has been shown that humans and nonhuman primates have a remarkable capacity to adapt to altered environments. For example, subjects can readily adapt to manipulations of visual input imposed by displacing prisms and they can also adapt to experimentally induced changes in kinematic or kinetic transformations. Specifically, subjects have been shown to gradually adapt when they are asked

to make directional movements against a variety of force fields.

Adaptation is usually restricted to the particular movement on which subjects were trained or to movements that are very similar. This observation has led to the conclusion that movement regulation is achieved in a modular fashion. Taken to the extreme, this viewpoint holds that there is one internal model (with inverse and forward components) for each movement. Such an arrangement would make learning feasible because only the parameters of one module would need to be modified. There are several reasons to be skeptical of this hypothesis. Although it would simplify learning, the complexities are shifted to the process of selecting the appropriate module, interactions among them, and the amount of memory required to store them all. Furthermore, there does not seem to be any compelling evidence for such an arrangement.

Nevertheless, learning does appear to be modular. For example, subjects trained to make a movement in one direction in a novel force field will adapt so that movements in that direction and in neighboring directions are correct but movements in directions far from the one in which they train will be unaffected. Similarly, the vestibuloocular reflex can be adapted in a wide variety of ways, even to generate vertical eye movements in response to head rotations in the horizontal plane. However, this adaptation is frequency specific, being greatest at the frequency at which the subject oscillated.

The apparent modularity of learning is not inconsistent with a system in which information processing is widely distributed. The apparent contradiction can be resolved by considering the tuning characteristics of individual neurons. For example, as mentioned previously, motor cortical neurons are tuned to the direction of a movement. If one supposes that adaptation is restricted to circuits only involving neurons whose best direction is similar to the direction at which subjects are being adapted, then one can readily imagine a scenario of modular adaptation.

In conclusion, although the general form of the scheme by which the nervous system regulates and controls movement is known (Fig. 3), many of the details of the implementation of this scheme are not.

See Also the Following Articles

APRAXIA • BASAL GANGLIA • BIOFEEDBACK • CEREBRAL PALSY • EPILEPSY • HAND MOVEMENTS • MOTION PROCESSING

Suggested Reading

Buneo, C. A., Soechting, J. F., and Flanders, M. (1997). Postural dependence of muscle actions: Implications for neural control. *J. Neurosci.* **17,** 2128–2142.

Flanders, M., and Herrmann, U. (1992). Two components of muscle activation: Scaling with the speed of arm movement. *J. Neurophysiol.* **67,** 931–943.

Flanders, M., Helms Tillery, S. I., and Soechting, J. F. (1992). Early stages in a sensorimotor transformation. *Behav. Brain Sci.* **15,** 309–362.

Georgopoulos, A. P. (1991). Higher order motor control. *Annu. Rev. Neurosci.* **14,** 361–377.

Ghez, C., Gordon, J., Ghilardi, M. F., Christakos, C. N., and Cooper, S. E. (1990). Roles of proprioceptive input in the programming of arm trajectories. *Cold Spring Harbor Symp. Quant. Biol.* **55,** 837–847.

Hasan, Z. (1991). Biomechanics and the study of multijoint movements. In *Motor Control: Concepts and Issues* (D. R. Humphrey and H.-J. Freund, Eds.), pp. 75–84. Wiley, Chichester, UK.

Herrmann, U., and Flanders, M. (1998). Directional tuning of single motor units. *J. Neurosci.* **18,** 8402–8416.

Kakei, S., Hoffman, D. S., and Strick, P. L. (1999). Muscle and movement representations in the primary motor cortex. *Science* **285,** 2136–2139.

Kawato, M. (1999). Internal models for motor control and trajectory planning. *Curr. Opin. Neurobiol.* **9,** 718–727.

Macpherson, J. M. (1991). How flexible are muscle synergies? In *Motor Control: Concepts and Issues* (D. R. Humphrey and H.-J. Freund, Eds.), pp. 33–48. Wiley, Chichester, UK.

Mussa-Ivaldi, F. A. (1999). Modular features of motor control and learning. *Curr. Opin. Neurobiol.* **9,** 713–717.

Soechting, J. F. (1989). Elements of coordinated arm movements in three-dimensional space. In *Perspectives on the Coordination of Movement* (S. A. Wallace, Ed.), pp. 47–83. North-Holland, New York.

Soechting, J. F., and Flanders, M. (1992). Moving in three-dimensional space: frames of reference, vectors and coordinate systems. *Annu. Rev. Neurosci.* **15,** 167–191.

Soechting, J. F., Buneo, C. A., Herrmann, U., and Flanders, M. (1995). Moving effortlessly in three dimensions: Does Donders' law apply to arm movement? *J. Neurosci.* **15,** 6271–6280.

Wolpert, D. M., and Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks* **11,** 1317–1329.

Multiple Sclerosis

LORRI J. LOBECK
Medical College of Wisconsin

GLOSSARY

Lhermitte's phenomenon Tingling or electrical-type sensation experienced as a result of neck flexion, felt due to mechanosensitivity of demyelinated axons.

major histocompatibility complex An area on chromosome 6 that encodes immune response genes; possible area of genetic susceptibility in multiple sclerosis.

myelopathy Disease of the spinal cord, typically with weakness and/or sensory disturbance below the level of demyelination.

plaque Lesions of multiple sclerosis characterized pathologically by light gray to pink discoloration in white matter regions of the central nervous system.

relapse Neurologic signs or symptoms occurring in individuals with multiple sclerosis, persist for days to weeks, followed by partial or complete recovery.

Multiple sclerosis (MS) is an inflammatory, demyelinating disease of the central nervous system that results in destruction of both myelin and axons. The cause of the disease is unknown; however, it is believed to be an immune-mediated process. The disease is characterized by relapses and remissions of neurologic signs and symptoms early in the course, followed in some cases by progressive decline of neurologic function. Although most individuals are between 15 and 60 years of age at the time of diagnosis, the disease affects all age groups and causes varying levels of disability. Women are twice as likely to develop the disease as men.

I. DESCRIPTION OF MULTIPLE SCLEROSIS

An important feature of multiple sclerosis is the marked variability in neurologic symptoms and clinical course. Four different types of disease course have been described. Most patients present with relapsing–remitting disease. An individual develops neurologic symptoms and signs over hours to days and typically recovers in 6–8 weeks. This is followed by various lengths of disease-free periods. Some individuals have few relapses and no residual symptoms, whereas others have frequent relapses and accumulating disability. Approximately one-half of the patients with relapsing–remitting disease go on to develop secondary progressive disease. In this case, features of relapsing–remitting disease are followed by a gradual decline in function with or without superimposed relapses. Primary progressive disease, which affects 10–15% of individuals with MS, is characterized by a gradual decline in neurologic condition from onset, without relapses. This form of disease progression is typically seen in older individuals, more often males, as a progressive myelopathy. A fourth type, progressive–

relapsing MS, is similar to primary progressive MS but with superimposed relapses. The term benign MS is used carefully. Current estimates of disease that truly remains benign are low—5–10% of all patients.

Based on epidemiologic and genetic data, MS is believed to be caused by some environmental agent that triggers the disease in susceptible individuals. The trigger is unknown but likely infectious. Advances in biotechnology and better trial design have led to new therapies that are at least partially effective in treating the disease.

II. EPIDEMIOLOGY

There are an estimated 400,000 individuals in the United States who have MS and 1 million worldwide. The prevalence of the disease in individuals of Northern European descent is 1 in 1000, whereas in Asians the prevalance is 2 per 100,000. The highest prevalence rate is 250 cases per 100,000 individuals in the Orkney Islands, north of Scotland. Hispanic, Asian, and African Americans are less likely than Caucasians to have the disease. MS is usually sporadic, and often there is no documented family history. However, some families have shown an increased frequency of the disease and thus attention has focused on the genetic versus environmental features of the disease. The prevalence of MS varies along a north–south gradient worldwide, increasing as one moves away from the equator, even in countries with relatively homogenous racial populations. This gradient of disease prevalence suggests an environmental cause. There are exceptions to this pattern, such as Sicily and Malta, two areas in close proximity, which have a 10-fold difference in prevalence rate. Additionally, there are ethnic groups that appear resistant to the disease in high-risk areas, such as the Hutterites of western Canada and Native Americans in North America, reinforcing the genetic features of the disease. Furthermore, the north–south gradient may be explained by the migration pattern of ethnic groups, as they tended to migrate to similar climates. Studies of individuals moving from high-risk to low-risk areas after puberty show that these individuals carry the risk of the first location for developing MS, suggesting that an environmental factor may trigger the disease early in life. Clusters of cases of MS have been described, but many have not withstood further scrutiny to provide evidence that epidemics of MS occur. In summary, the data indicate that the geographical distribution of MS cannot be due to a single environmental factor, nor can genetic factors alone explain the distribution.

III. ETIOLOGY

A. Genetic

Epidemiologic studies provided early, albeit indirect, evidence of a genetic susceptibility toward developing MS. Studies of monozygotic twins illustrated a 25–30% lifetime risk of developing the disease if one sibling is effected. Dizygotic twins have a risk similar to that of their nontwin siblings (2–5%), which is higher than that for the normal population. Additional studies have shown that the risk is higher in full siblings rather than half-siblings. Siblings of individuals with MS raised in separate homes retain a higher risk of developing the disease. First-degree, adoptive, nonbiologic relatives do not have an increased risk compared to the general population; thus, genetic rather than environmental factors are more likely to explain familial clustering of disease. In the early 1970s, an association between the risk of developing MS and the major histocompatibility complex (MHC) on chromosome 6, which encodes immune response genes, was recognized. The MHC class I and class II molecules are glycoproteins located on a cell surface that present antigen to antigen-specific T cells. An individual's ability to react to an antigen is determined by the MHC region. $CD8^+$ T cells interact with MHC class I molecules on the surface of all nucleated cells in humans; those molecules are known as human leukocyte antigens (HLA)-A, -B, or -C. Class II molecules are present on only some cells, including monocytes, macrophages, endothelial cells, and microglial cells. These antigens are described as HLA-DP, -DQ, and-DR. Although there may be a weak association with HLA-A3, multiple studies have shown an increased occurrence of HLA-DR2 haplotypes associated with MS. Most individuals with MS do not have this HLA-DR haplotype; thus, it is not essential or sufficient to cause the disease. It is also possible that an unidentified gene close to the HLA gene is the susceptibility locus. Recent genetic studies indicate that multiple unlinked genes may influence the risk for developing MS. Each gene may add to the risk. Different genes may influence susceptibility in some individuals and not others. A screen of the entire genome for regions related to risk of MS with anonymous genetic markers identified a susceptibility loci in the MHC region.

One model for MS in animals is experimental allergic encephalomyelitis (EAE). In this animal model, the susceptibility to developing demyelination is based on genetic risk. Specific chromosomes and segments have been identified in the animal model that

increase the risk. The model suggests that initial events that occur in the disease may be mediated by different genetic factors than those that cause later progression of the disease.

B. Infections

The search for an infectious cause of MS has been pursued for many years. Bacteria, spirochetes, and atypical bacteria have be considered, although not confirmed. Viruses have been the focus of the search based on epidemiologic, genetic, pathologic, and serologic studies. The nonrandom North–South gradient of disease occurrence, described previously, is atypical for other autoimmune diseases and suggests the role of an environmental trigger. The low rate of concordance for monozygotic twins is evidence against MS being a purely genetic diseases. Because twins share many childhood diseases one would expect the concordance rate to be much higher if MS was a result of a nonspecific response to multiple agents. More likely, a specific, poorly transmissible agent leads to activation of the disease. Pathologic studies show that the changes seen in the central nervous system are similar in MS and viral illnesses. Both can produce inflammatory demyelinating lesions in the brain. Persistent viral infections can cause chronic demyelination.

Immunologic studies have noted the cerebrospinal fluid is similar in viral infections and MS. In viral infections, however, the oligoclonal bands can be identified as due to specific viral antigens. In MS, the oligoclonal bands are not known to react with a specific antigen. Antibodies are elevated in the cerebrospinal fluid and serum in patients with MS to such viruses as measles, Epstein–Barr virus, rubella, mumps, and herpes simplex, but this is more likely due to a nonspecific increase due to genetic or immunologic factors rather than a specific response to a virus.

There are animal models of viral demyelinating diseases, that resemble MS, including canine distemper virus, Theiler's murine encephalomyelitis, murine coronavirus, and visna. Furthermore, viruses are known to result in demyelinating diseases in humans. For example, postinfectious encephalomyelitis or acute disseminated encephalomyelitis (ADEM) have occurred following infection with measles, varicella, or Epstein–Barr virus. Persistent infections with papovavirus cause progressive multifocal leukoencephalopathy (PML) in immunocompromised individuals.

Proposals to explain how viruses may cause MS include an immune system response to a chronic or transient virus, reactivation of a persistent infection, or viral infection of immunocompetent cells including lymphocytes. "Molecular mimicry" may explain the immune system response to virus, or other infectious agents. This model suggests that a viral peptide is similar to a component of myelin. The immune system recognizes the foreign viral peptide but cross-reacts with a myelin, resulting in activation of the immune system. Further damage results with exposure of additional myelin antigens. Exposure of antigens may result in epitope shifting, in which the initial inciting antigen no longer remains the sole perpetrator in the immune process. Some viruses (e.g., Epstein–Barr virus, cytomegalovirus, and measles) are known to have peptides that are similar to antigens identified on myelin components. Furthermore, the antigen may not be a component of myelin but an enzyme or regulatory protein, which may affect the immune function in other ways (e.g., permeability of the blood–brain barrier or myelin metabolism). Superantigens, bacterial or viral envelope components that activate large populations of T cells or multiple clones of T cells, may be another mechanism by which a clone of auto reactive T cells may be activated. Others argue that a latent viral infection results in disease. Periodic relapses would be explained by reactivation of the virus. Progression may be due to failure of the host to inactivate the viruses. Several viruses are known to exist in the human central nervous system and to reactivate. For example, herpes zoster causes chicken pox in childhood and then reactives in the adult as shingles, a disease with rash in a dermatomal pattern. In other cases, a virus remains as a persistent infection. The measles virus causes subacute sclerosing panencephalitis. HIV infections result in various neurologic complications, including dementia. HTLV1 causes a progressive myelopathy. Whether the virus causes direct injury to neural tissue or causes bystander inflammation and demyelination is unknown. Finally, it is possible that viral infection of lymphocytes or other immune cells leads to changes in function of these cells and ultimately altered immunity.

Many viruses have been suggested as inciting agents for MS. Recently, there has been interest in human herpesvirus 6 (HHV-6). Most, if not all, types of herpes are latent in the nervous system. They are transported within axons and are known to cause demyelination. HHV-6 causes roseola or sixth disease in children, and

like other herpesviruses it may reactivate in the adult nervous system. The presence of this virus has been confirmed in the brains of individuals with MS and others. Determination of the role of this virus in MS requires additional testing.

IV. IMMUNOLOGY

Most authorities believe the inflammatory, demyelinating disease process is immunemediated, even though the cause remains unknown. The target of inflammation may be a component of myelin. The process most likely begins with the T cell. Antimyelin T cells are recognized in the normal adult peripheral blood, but in MS these cells migrate to the central nervous system. Humeral immune responses are also likely, based on the intrathecal presence of immunoglobulins. The antigen recognized by these antibodies has not been determined. Myelin basic protein (MBP) and proteolipid protein make up 30–50% of myelin protein. Myelin oligodendroglia protein (MOG) makes up 50% of central nervous system myelin. This protein is not present in the peripheral nervous system and may be the antigen in this exclusively central nervous system disease. MOG is located on the oligodendroglia surface of the outer lamella of the myelin sheath and is thus accessible to attack. Other possible targets include myelin-associated glycoprotein, $\alpha\beta$ crystalline, 2–3-cyclic nucleotide, 3′-phosphodiesterase, and viral antigens.

The marmoset model of relapsing experimental allergic encephalomyelitis, which is histologically similar to MS, has demonstrated that T cells recognize MBP and MOG. These T cells induce inflammation, but demyelination does not occur unless the MOG antibody is present. Although MOG may be the initial antigen and inciting factor, exposure as a result of inflammation induces new antigens and epitope shifting. The antigen for this disease is difficult to identify possibly because there are multiple antigens.

Figure 1 illustrates the proposed mechanisms for the immune-mediated injury in MS. Genetic or environmental factors (viruses, bacteria, or superantigens) cause activation of circulating T cells. Adhesion molecules on the circulating T cell surface, very late

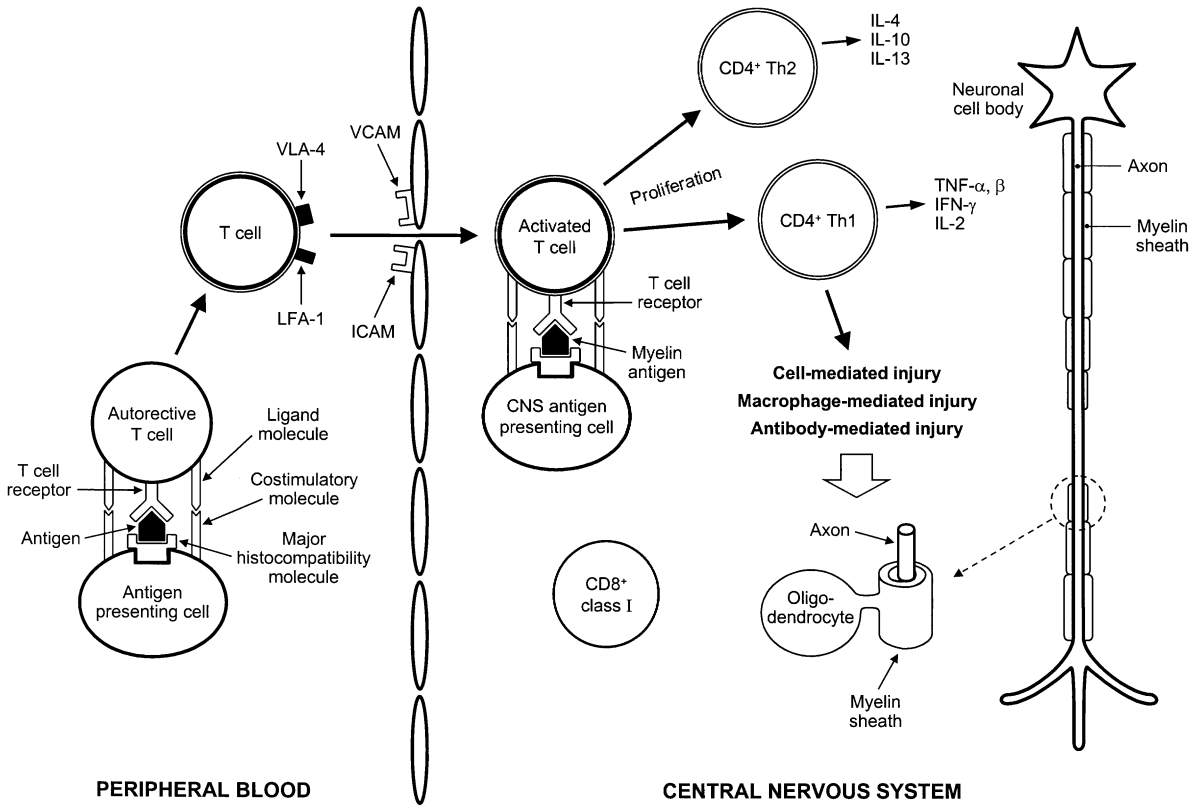


Figure 1 Proposed mechanisms for immune-mediated injury in MS.

antigen (VLA-4) and lymphocyte function associated antigen (LFA-1), bind complementary receptors on the endothelium, intracellular cell adhesion molecule-1, and vascular cell adhesion molecule-1. Once in the central nervous system, the activated T cell secretes cytokines, including tumor necrosis factor-β (TNF-β) and interferon-γ (IFN-γ), which in turn activate antigen presenting cells (APCs) including astrocytes, macrophages, and microglia. Class II MCH molecules on APCs and T cells interact in the presence of central nervous system antigen. Costimulating molecules on the APC and T lymphocytes are necessary for activation of T cells. $CD4^+$ cells differentiate into proinflammatory cells ($CD4^+$Th1) or antiinflammatory ($CD4^+$Th2) cells. Proinflammatory cells (Th1) secrete cytokines such as TNF-α or -β or IFN-γ. The result is activation of various process, that cause injury to myelin or oligodendrocytes. Antibody-mediated injury may be cell dependent or occur via complement activation. Cell-mediated injury induces damage by further release of proinflammatory cytokines. Binding between FAS–ligand and FAS or binding of $\alpha\beta$ crystallin will result in apoptosis or programmed cell death. Macrophages induce myelin injury by phagocytosis and secretion of toxic substances, such as proteases, nitric oxide, oxygen radicals, and proinflammatory cytokines. Apoptosis may occur via the FAS–ligand interaction.

In some cases, cytotoxic $CD8^+$ autoreactive T cells may be the primary cause of injury by binding class I MHC antigens on oligodendrocytes. They may also bind via FAS and induce apoptosis. In addition, these cells may release perforins, which create membrane pores and kill cells. Finally, other factors may lead to injury of the oligodendroglia, resulting in "dying back oligodendrogliopathy." The result is varying degrees of demyelination and axonal destruction.

Even as some areas are damaged, others are in various stages of repair. An antiinflammatory process activated via the CD4 Th2 cells, which secrete cytokines such as IL-4, -10, and -13, decreases inflammation and downregulates the immune response. This process occurs simultaneous to the inflammatory response.

Thinly myelinated axons in pathologic specimens indicate that remyelination occurs in patients with MS. Repair likely begins with resolution of inflammation and edema. New sodium channels may develop in a demyelinated axon to allow for propagation of active potential. It is not known if these sodium channels develop from the neuron or surrounding glial cells. Oligodendroglia progenitors capable of proliferation have been identified but the specific role of these cells is not clear. It is known that remyelination can occur if inflammation is controlled.

The final result of this inflammatory process, irreversible axonal injury, is known to occur early in the disease. These findings suggest that early intervention of the disease course, which prevents or decreases inflammation, may prevent axonal injury and allow for remyelination. Multiple potential sites for therapeutic interventions can be considered including blockage of adhesion or co stimulating molecules, cytokine therapy, or modulation of the T cells via vaccination.

V. PATHOLOGY

Pathology sections of the brain in individuals with MS reveal well-demarcated lesions or plaques that appear pink or gray. Microscopically, the lesions are characterized by inflammation, demyelination, or gliosis. Inflammatory lesions have both perivascular and leptomeningeal infiltration of lymphocytes, monocytes, and macrophages. Histopathologic studies show that macrophages and $CD4^+$ T cells are the most prominent cell types within lesions. There are fewer B cells and plasma cells. Lesions can be further characterized as active (acute), chronic active, or chronic inactive. In the acute lesion, perivascular and parenchymal infiltration with mononuclear cells, such as T cells and macrophages, exists. Cells are mostly large, round, and lipid-laden macrophages. There are variable numbers of CD4 and CD8 cells present. MHC class II positive cells are abundant. Similar findings are seen in ADEM and animal models that have T cell-mediated immune responses. In these cases, the demyelination that occurs with MS-type plaques is not present. This implies that there must be another factor present for demyelination to occur. Chronic active plaques have MHC class II cells and lipid-laden macrophages at the border of the lesion but not centrally. Chronic inactive lesions have few MHC class II-positive cells and are hypocellular. Recently evidence indicates that the pathologic process may vary between individuals. Lucchinetti and Lassmann described different types of pathological appearances based on immunohistochemical studies. The first type shows areas of demyelination with no oligodendroglia loss. There is local deposition of demyelinating antibodies and complement activation. The preserved oligodendroglia cells favor remyelination and recovery. A second type has demyelination; however, there is destruction and loss of oligodendroglia cells with the

presence of T cells and macrophages in these areas. Local precipitation of antibody and complement is not evident. A third type involves primary demyelination and a gradient loss of oligodendroglia cells toward the center of an inactive plaque. In inactive plaques there is profound oligodendroglia loss. A fourth type describes primary oligodendroglia destruction in white matter near the plaque and resultant secondary demyelination. A fifth type describes demyelination with profound loss of axons, oligodendrocytes, and astrocytes. This is not likely to be due to different stages of the disease since individuals tend to have similar patterns of pathologic changes throughout the brain. Rather, the types of injuries may be the result of different immunologic mechanisms that occur between patients and not within a patient.

The irreversible disability of MS has been attributed to axonal injury, which is evident in autopsy specimens. Charcot recognized degeneration of axons in MS as early as 1877. Recently, axonal transection has been demonstrated even in the early stages of the disease. These findings suggest that accumulating axonal injury from early on may result in progressive disability. Attempts to prevent axonal damage may prevent long-term disability.

VI. NEUROPHYSIOLOGY

Myelin acts as an electrical insulator in the brain and spinal cord. This lipid sheath is produced by oligodendrocytes in the brain and spinal cord. One oligodendrocyte insulates 1–100 axons. The myelin, produced in layers or lamellae, is interrupted by nodes of Ranvier. At the node of Ranvier there is a high density of sodium channels, which are responsible for sodium ion shifts for depolarization. These areas allow for production and propagation of action potentials. Conduction is a saltatory (not a continuous) process. The internodal areas (areas between the nodes of Ranvier under the myelin sheath) have a low proportion of sodium channels and a high proportion of potassium channels, especially in the perinodal region. These potassium channels oppose depolarization. Depolarization produces an electrical charge, which is conducted through the myelinated segment to the next node for depolarization. Myelin increases the speed of conduction and efficiency. With demyelination, slowing of action potentials and conduction block occur. Conduction block prevents propagation of any axon action potential. Sodium channels develop along the demyelinated internodal area as a way to allow for propagation of an action potential, although this is slower. This reorganization of the demyelinated axon results in a higher than normal density of sodium in demyelinated regions. Other factors may interfere with conduction, such as circulating immune factors or edema. In addition to reorganization of sodium channels, resolution of edema or clearance of these factors may restore conduction. Synaptic changes and conduction through normal unmyelinated axons may also restore function.

It is known that small changes in temperature and electrolytes can produce conduction block in demyelinated axons. Patients' symptoms may recur when they have an elevated body temperature as a result of illness or increased environmental temperature. Uthoff's phenomenon is one of the best known of these phenomena. With elevation of body temperature, such as during exercise, individuals develop decreased visual acuity in an eye previously affected by optic neuritis. Symptoms resolve when body temperature returns to normal. Trains of impulses can also block propagation of action potentials by prolongation of the refractory period—the time during which a neuron cannot depolarize. Clinically, this occurs as patients complain of fatigability of muscle strength with repetitive use of any extremity, such as occurs during walking. It is possible that modification of the sodium and potassium channels in demyelinated neurons may result in improvement in clinical symptoms.4-Aminopyridine, a potassium channel blocker, is a drug that improves symptoms for patients with MS, especially those who are heat sensitive. Currently, it is not a Federal Drug Administration (FDA)-approved drug, but further research into modification of sodium and potassium channels may lead to therapies, that provide improved conduction in demyelinated axons.

Paroxysmal symptoms of MS are a unique feature of the disease that have their basis in the physiology of demyelinated axons. These symptoms include episodes of recurrent hemibody sensory symptoms, paroxysmal dysarthria, and paroxysmal pain such as trigeminal neuralgia. These phenomena may be a result of ephaptic transmission, which occurs when adjacent demyelinated axons communicate in a lateral fashion and result in recurrent abnormal stimuli. Lhermitte's phenomena is a transient electric shock-like sensation in the spine and/or extremities believed to be due to increased mechanosensitivity of demyelinated axons in the spinal cord and it results when neck flexion causes mechanical changes in the spinal cord.

VII. CLINICAL FEATURES

The clinical features of MS are based on the location of lesions rather than the severity of these lesions. Lesions must occur in eloquent areas to produce symptoms. Many lesions occur in areas that are not clinically recognized as symptomatic, including periventricular white matter lesions identified near the lateral ventricles. A relapse of MS is defined as an exacerbation or clinical worsening of neurologic signs and symptoms that persists for weeks to months. This is followed by a period of remission. Symptoms must be present for at least 24 hr to be considered a relapse, but generally they are present longer. Most common are sensory symptoms. They are often the earliest symptoms and occasional are peculiar in distribution. In addition, sensory symptoms may not have objective findings; thus, they are ascribed to other causes such as stress. Patients often complain of "numbness," but they are often experiencing positive sensory phenomena such as tingling or dysesthesias. The ability to detect vibration in the extremities is often decreased and identified by neurologic examinations early in the disease. Often, the sensory symptoms of MS are due to lesions of the posterior columns rather than the spinothalamic tracts; however, decreased pinprick and temperature sensation can occur. Lhermitte's phenomenon is a sensory symptom in which patients experience sudden, brief electrical sensations along the spine or in the extremities. Less commonly, a radicular pattern of pain may occur and is likely due to demyelination of the sensory root entry zone. Lancinating neuralgias or dysesthesias are not uncommon. A "useless hand" develops when there is sensory disruption in the sensory spinal tracts that affects vibration, two-point discrimination, and proprioception. Often, this occurs as a result of a high cervical spinal cord lesion of the posterior columns or in the medial lemniscus system of the brain stem. Although strength is preserved, the hand is difficult to manipulate for tasks because of the sensory disruption. This may occur in other spinal disorders, but when it occurs in young individuals it is most likely due to MS.

Disorders of the visual system are common at the onset of MS. Optic neuritis results in acute or subacute monocular loss of vision. Patients complain of blurred vision or dimming of vision, often associated with photophobia and pain that increases with eye movement. Examination often discloses decreased visual acuity and central scotoma of visual loss. The optic nerve head may appear swollen, with hemorrhages or exudate (papillitis), or may be normal appearing (retrobulbar neuritis) by funduscopic examination. Color vision is often affected and can easily be tested by asking a patient view a red object with each eye separately and compare the difference of appearance. An afferent pupillary defect, elicited by shining a bright light on each pupil, may demonstrate delay of pupillary response or dilatation when the light is brought to the affected eye. Often, visual acuity returns to normal within several weeks. Later, optic atrophy may be detected as a pale optic nerve head evident by funduscopic examination. Other types of visual field defects have been described, including bilateral homonymous hemianopia and quadrantanopias, indicative of lesions more posterior in the visual system. Although optic neuritis is commonly associated with MS, other conditions including systemic lupus erythematosus or sarcoidosis must be considered.

Extraocular movement disorders are also relatively common. Examination will often show horizontal nystagmus, which is asymptomatic. Rotary, upbeating and downbeating nystagmus has also been described. A bilateral internuclear ophthalmoplegia (INO) is a result of a lesion in the brain stem in the medial longitudinal fasciculus. Patients develop failure of the ipsilateral eye to adduct, whereas the contralateral eye develops abducting nystagmus. Patients may be asymptomatic with central gaze, but with horizontal gaze they develop diplopia or oscillopsia, a sense of movement of images. The INO may be bilateral or unilateral. Incomplete INO is frequent, detected only by slowed adduction or mild abducting nystagmus. A skew deviation occurs when one eye is slightly elevated compared to the other. Sixth nerve palsies resulting in horizontal diplopia occur, although third and fourth cranial neuropathies are rare.

Vertigo may occur with demyelinating lesions in the brain stem. It is often difficult to determine if vertigo is due to a peripheral or central disorder. Other features of brain stem injury may be helpful in defining the cause of vertigo but are not necessary to conclude that vertigo is a result of MS. Although infrequent, hearing loss is known to occur. Dysfunction of the facial nerve develops from a lesion of the seventh nerve in the brain stem or with a subcortical lesion, causing a central or upper motor neuron type of weakness confined to the lower aspect of the face. Myokymia, which is an involuntary wave-like fasciculating movement in the face, also results from demyelination.

Speech abnormalities include scanning dysarthria, in which each word or syllable is given equal emphasis. Other types of dysarthria, including pseudobulbar or nasal speech, are less common. Symptoms due to

cortical dysfunctions, such as aphasias or agnosias, are not typical of MS. Seizures are rare and tend to occur later in the disease. Cerebellar findings are not common early in the disease, but with progression ataxia, dysmetria, intention tremor, and dysdiadochokinesias may all occur.

Cognitive dysfunction is estimated to occur in: 15–50% of all patients with evidence of neuropsychological dysfunction. This aspect of the disease can be very mild to extremely severe with profound dementia. The most common features of cognitive dysfunction include difficulties with recent memory, sustained attention, verbal fluency, conceptual reasoning, and visual–spatial perception as determined by neuropsychological testing. These difficulties are not easily defined by simple procedures in the clinic, such as the Mini-Mental status examination. The severity of physical disease does not correlate well with cognitive dysfunction. Cognitive difficulties are significantly disabling and secondary to fatigue, which is the most likely cause of patients discontinuing to work. Depression is very common and may interfere with cognitive function. Euphoria, when present, is associated with significant cognitive difficulties or subcortical forebrain lesions.

Bladder dysfunction, although unusual at onset, is common with progressive disease. Symptoms include frequency, urgency, and incontinence. These symptoms are a result of a failure to store or a failure to empty the bladder. With failure to store, the bladder wall is sensitive and expels small amounts of urine unexpectedly due to bladder wall contractions. With failure to store, the detrusor muscle is weak and urine difficult to expel. Incontinence occurs with overflow of a full bladder. Often, patients have a combination of these problems, referred to as detrusor dyssynergia. Bladder wall contraction occurs simultaneously to closure of the external urethral sphincter and prevents emptying of the bladder. With long-term retention, vesicular ureteral reflux may occur along with hydronephrosis and renal failure. Bowel dysfunction includes constipation, bowel urgency, or incontinence. In an attempt to control bladder symptoms, many patients decrease fluid intake, which worsens constipation. Sexual dysfunction is common and is likely a result of both physiologic and psychosocial factors. Male sexual dysfunction, which has been studied more extensively, includes erectile dysfunction, decreased sensation, and decreased libido resulting in an inability to have erections or ejaculate. Female patients experience decreased sensation, decreased or absent orgasm, decreased arousal, and vaginal dryness.

One of the early manifestations of MS may be transverse myelitis. Patients develop acute or subacute motor and sensory symptoms in the lower extremities extending into the thorax with a definable sensory level. Transverse myelitis may represent the first episode of MS or remain as a monosymptomatic illness, similar to optic neuritis. Para-, hemi-, or quadriplegia may develop during exacerbations. With progressive disease, patients develop a myelopathy with progressive weakness in lower extremities, later involving upper extremities. Associated with the weakness is an increase in muscle stretch reflexes and spasticity. Patients experience stiffness or cramping and intermittent spasms. Discomfort or pain may result.

Finally, paroxysmal symptoms—the stereotypical, repetitive symptoms and signs that occur in clusters, are a unique feature of MS. Patients may develop "tonic seizures" in which dystonic posturing of the hand or arm lasts from seconds to minutes. The individual is unable to control the involuntary flexion or stiffness. Other paroxysmal symptoms include transient dysarthria and sensory phenomena. One of the most common paroxysmal symptoms is trigeminal neuralgia, in which patients experience severe lancinating pain in the face on a repetitive basis. Trigeminal neuralgia may occur in individuals without MS, but it tends to occur at a younger age and is more frequently bilateral in individuals with MS.

VIII. DIAGNOSIS

The diagnosis of MS is based on the clinical features of the illness. There is no one definitive test for MS. Signs and symptoms define lesions that develop over time (at least 1 month apart) and space (different areas of the central nervous system). Most symptoms tend to develop over several days and then persist for several weeks to months, with improvement and in many cases return to normal. The neurologic examination generally provides objective evidence of the patient's symptoms, however, sensory symptoms may be entirely subjective. Paraclinical data, which include magnetic resonance imaging (MRI), cerebrospinal fluid (CSF), and evoked potentials (EPs), are used to confirm or dispute the clinical diagnosis. The most recent criteria were developed by an International panel of experts in the field MS (Table I), These criteria incorporate MRI as a tool of assessing subclinical disease activity and providing for a diagnosis before a second clinical event. The criteria includes visual

Table I
MS Diagnostic Criteria[a]

Clinical attacks	Objective lesions	Additional requirements to make diagnosis
2 or more	2 or more	None
2 or more	1	Dissemination in space by MRI or positive CSF and 2 or more MRI lesions consistent with MS or further clinical attack involving different site
1	2 or more	Dissemination in time by MRI or second clinical attack
1	1	Dissemination in space by MRI or positive CSF and 2 or more MRI lesions consistent with MS
(clinically isolated syndrome)		AND
		Dissemination in time by MRI or second clinical attack
		Positive CSF
		AND
0	1	Dissemination in space by MRI evidence of 9 or more T2 brain lesions
(progression from onset)		or 2 or more cord lesions or 4–8 brain and 1 cord lesion
		or abnormal VEP with 4–8 MRI lesions
		or abnormal VEP with less than 4 brain lesions plus 1 cord lesion
		AND
		Dissemination in time by MRI or combined progression for 1 year

[a]Modified from McDonald, *et al.* (2001). Recommended Diagnosic Criteria for MS. *Ann. Neurol.* **50,** 121–127.

evoked potentials as the only form of evoked potentials to support the diagnosis of MS. Although unusual, the paraclinical data can be normal in some patients with the clinical diagnosis of MS.

IX. CEREBROSPINAL AND BODY FLUIDS

CSF has been evaluated from diagnostic and research standpoints. Typical CSF findings in patients with MS include less than 20 white blood cells per cubic milliliter. These cells are mostly lymphocytes and occasionally macrophages. Total protein and glucose levels are usually normal; however, protein levels may be slightly elevated but rarely above 100 mg/dl. The abnormalities in CSF are primarily those of immunoglobulin production. Oligoclonal bands (O bands) represent multiple monoclonal globulins that are mostly of the immunoglobulin-G (IgG) type. O bands are determined by agarose electrophoresis or isoelectric focusing and indicate two or more IgG bands in the gamma region. Agarose electrophoresis is less sensitive but currently more commonly performed. Once O bands are present, they persist indefinitely. There is a low false-positive rate of approximately 4 or 5%. The presence of O bands may be a predictor of the chances of developing clinically definite MS in those individuals with monosymptomatic disease. Patients with O bands in the CSF are more likely to have a confirmed diagnosis of clinically definite MS. Other measures of CSF IgG production include the IgG synthesis rate and the IgG index, which are based on comparison of albumin and immunoglobulin production in the CSF to serum. These measures are considered less specific and sensitive than oligoclonal bands, but they are useful in indicating immune abnormalities and confirming the clinical diagnosis.

Because CSF is more difficult to obtain, other markers for disease activity have been pursued. Urine is an easily accessible fluid, although volume, collection timing, and infections may influence results. Myelin basic protein-like (MBP-L) is smaller than MBP found in the CSF. It has a cryptic epitope that is not exposed in central nervous system myelin—possibly a small peptide. MBP-L fluctuates independent of acute relapses; however, a sustained increase is present in patients with secondary progressive MS. Levels were also elevated in those individuals with relapsing–remitting MS who went on to develop secondary progressive MS. There are currently no markers in the blood that are used for diagnosis of MS. Markers that have been considered include circulating adhesion molecules, antibodies, cell subpopulation, cytokines, and cytokine receptors.

X. EVOKED POTENTIALS

Evoked potentials were more frequently used in the diagnosis of MS prior to the advent of MRI. The data are generated by presenting repeated stimuli and recording a wave form or EP from scalp surface electrodes. An alternating checkerboard pattern is used as a visual stimuli for visual EP, repeated auditory clicks are used for brain stem auditory EP, and repeated electrical sensory stimulation is used for tibial and somatosensory EP. As a result of repeated stimulation, the EP has a defined latency and amplitude. Based on standards established in each laboratory, a prolongation of the latency of these responses suggests that demyelination has occurred. A loss of amplitude of those responses is abnormal but not necessarily due to demyelination. These studies are nonspecific and are most helpful in confirming a suspicious clinical finding. For example, a distant history of unusual symptoms suggestive of optic neuritis may be confirmed by a prolonged visual evoked response recorded from stimulation of one eye. Confirmed clinical findings do not need to be further assessed with EPs since additional information is not gathered. Currently, EPs are not used to define prognosis or as surrogate markers of disease activity in clinical trials.

XI. MAGNETIC RESONANCE IMAGING

Magnetic resonance imaging was first used in 1981 to study patients with MS. The technique was rapidly accepted in the clinical realm. Images are generated as a result of energy release from protons (positively charged hydrogen nuclei) that have been aligned in the axis of a strong magnet. Water, lipids, and other molecules contain hydrogen atoms, which spin and precess around the main axis. This axis is referred to a magnetization vector. Radio frequency pulses cause the protons to rotate and spin away from the main axis. Once the radio frequency pulse is removed, a signal is induced as the protons return to their original rotation. The signal decays over time and is referred to a longitudinal decay (T1) and transverse relaxation (T2) time. T1 and T2 times are the result of the environment in which these proteins exist (e.g., CSF protons require a longer time to return to equilibrium). The radio frequency pulses can be given at different times and for different lengths of time to generate different types of images that are "weighted" toward T1 or T2. This weighting is based on repetition time (TR) or time between pulses and echo time (TE), the time to generate the echo. T1 images have both short TE and TR, whereas T2 images have both long TE and TR. CSF is white in T2-weighted images, as are demyelinating lesions of MS (Fig. 2). On T1-weighted images some chronic lesions appear as dark areas or "black holes," which indicate tissue destruction and probable axonal loss. Newer techniques, such as fluid-attenuated inversion recovery imaging, cancel the signal from CSF, and as a result demyelinating periventricular lesions are more easily detected (Fig. 3). Gadolinium, a rare element with paramagnetic substance, is given intravenously to provide contrast enhancement of areas of breakdown of the blood–brain barrier. Because gadolinium shortens the T1 and T2 relaxation time, images appear bright where the blood–brain barrier is disrupted. This enhancement, which is the earliest standard MRI evidence of a newly developing MS lesion, indicates acute inflammation and often persists 2–6 weeks in areas of demyelination.

In general, the lesions of MS are ovoid or round and located adjacent to the body or temporal horn of the ventricles, in the corpus callosum or infratentorial or cortical–subcortical areas. Size ranges from a few millimeters to larger confluent areas. Lesions change in size over time and may disappear, however, often an abnormality will persist on T2-weighted images. The white matter may appear diffusely abnormal (i.e., "dirty white matter"). Cortical lesions are present pathologically but more difficult to detect with MRI. New lesions or enlargement of old lesions occur while other lesions are shrinking. Spinal cord lesions are detected most often in the cervical region. The spinal cord may appear swollen during the acute phase and may be enhanced with use of gadolinium. Optic nerve abnormalities are visualized with special techniques to suppress signal from orbital fat.

Because of the frequency of nonspecific white matter changes in MRI, various criteria have been suggested to define when brain MRI abnormalities are consistent with the clinical diagnosis of MS. One frequently used set of criteria defines significant findings as three or more lesions, that are equal to or greater than 3 mm, with at least one located in periventricular or infratentorial areas. At least 90% of patients with clinically definite MS have MRI abnormalities. Other explanations for white matter abnormalities exist and include patchy white matter abnormalities related to hypertension or diabetes mellitus. Lacunar lesions due to strokes appear isointense to CSF because of complete tissue destruction, which is not typical of MS lesions. Other individuals with risk for cerebrovascular disease

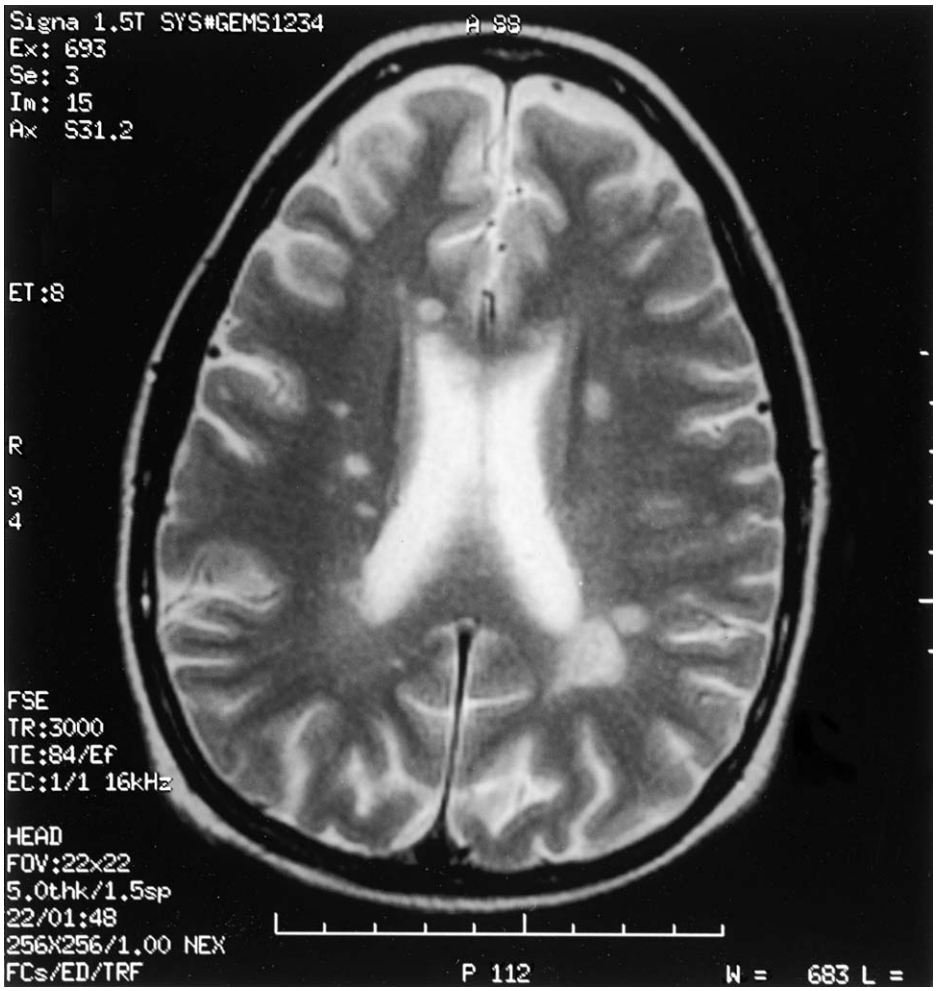


Figure 2 T2 weighted MRI demonstrating white matter lesions consistent with a diagnosis of MS.

may have incidental lesions scattered in the deep white matter. Lesions seen with systemic lupus erythematosus are also located in the deep white matter but are not typically periventricular. Neurosarcoidosis may cause white matter abnormalities but frequently has associated meningeal enhancement due to leptomeningitis. ADEM may appear identical to MS. Lesions tend to resolve and do not recur. Gliomas may appear as solitary lesions, as may ''pseudotumor'' lesions of MS. Follow-up MRI may be definitive, although biopsy may be necessary to distinguish the two. Finally, inherited disorders of myelin, the leukodystrophies, demonstrate symmetric confluent rather than patchy, discrete lesions. A clinical history and examination are most helpful in providing alternate explanations for these abnormalities.

MRI has provided much information about the underlying pathophysiology of MS. Serial imaging has demonstrated that white matter lesions are present even when individuals are not clinically aware of disease activity. This is not a consistent process and varies from month to month, however, the frequency of subclinical lesions may be 5–10 times greater than clinical disease. Clinically evident lesions tend to be present in the brain stem and spinal cord. Subclinical lesions in the cerebrum tend to increase with clinical events. The extent of lesions present on T2 scan does not correlate well with the clinical examination, other than a possible correlation with cognitive dysfunction. However, an increase in the extent of lesions over time correlates with increasing physical disability. The number and area of contrast-enhancing lesions does not correlate with disability. In primary progressive MS, there are often very few lesions present on T2-weighted images, compared to relapsing–remitting or secondary progressive MS in which multiple new and

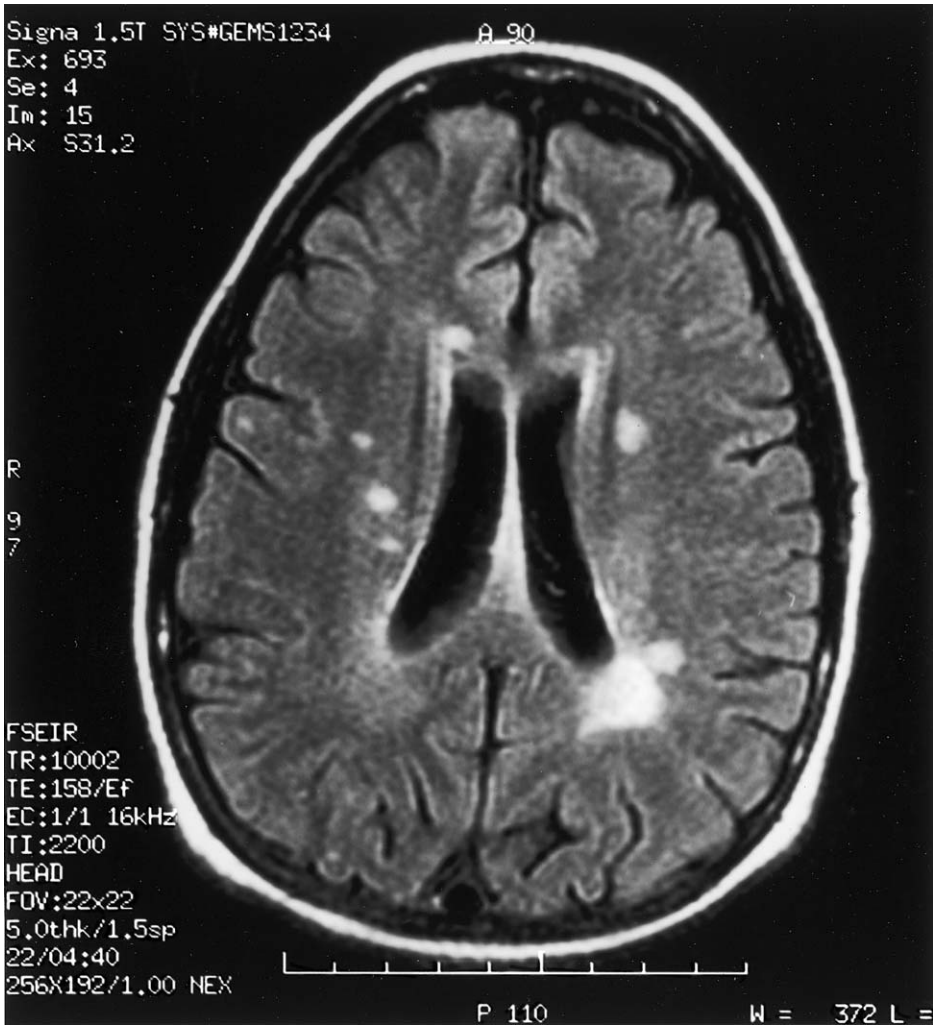


Figure 3 Fluid attenuated inversion recovery sequence demonstrating white matter lesions as seen in MS.

enhancing lesions tend to develop over time. This may indicate a pathophysiologic difference in disease type. MRI also has a predictive value for the diagnosis of MS in individuals who have had monosymptomatic disease, such as transverse myelitis or optic neuritis. The risk for an eventual diagnosis of clinically definite MS is higher in those individuals with abnormal brain MRI. MRI has been used as a tool in the clinical trial realm, allowing for a shorter clinical trial because it is a surrogate marker for disease activity. This becomes increasingly important as the partially effective therapies available will need to be incorporated into clinical trials of new agents, thus requiring more patients in order to prove efficacy. Problematic in the use of MRI has been clinical trials showing an effect on the clinical or MRI aspect but not vice versa. This apparent mismatch of data will need to be clarified.

Newer techniques add to our understanding of why T2-weighted images may not correlate well with clinical disability. Magnetization transfer imaging (MTI) relies on a different relaxation time of protons bound to macromolecules versus that of those protons that are freely moving in water. The normal appearing white matter of T2 weighted scans is abnormal when assessed with MTI, implying a more diffuse process than demonstrated by conventional T2 images. Magnetic resonance spectroscopy (MRS) uses MRI to generate a spectra of hydrogen or high energy phosphorus-containing metabolites. The spectra is used to define pathology of lesions, including tumors and demyelination. The spectra reveal major resonances from choline, creatine and phosphocreatine, and *N*-acetylaspartate (NAA). Choline is present in membranes and also present with increased myelin

breakdown products. Creatine and phosphocreatine tend to be stable in MS other than acute lesion. NAA is present only in neurons and neuronal processes and can be used as a specific axonal marker in white matter. Decreased NAA is indicative of axonal pathology. MRS has shown a decreased NAA concentration throughout the brain of individuals with MS, implying diffuse injury beyond the site of lesions detected by routine imaging techniques.

Brain atrophy documented by MRI shows volume loss around the third and lateral ventricles and decreased corpus callosum area and brain width. Atrophy has been documented early in the course of the disease, before significant clinical disability, and is likely another indicator of underlying axonal damage.

Thus, standard and newer MRI techniques have shown that MS can be a subclinical process resulting in increasing MRI burden of disease. Findings also support the hypothesis that axonal injury occurs early in the course of the disease and may be more diffuse than previously thought.

XII. DIFFERENTIAL DIAGNOSIS/DISEASE VARIANTS

Prototypic MS is one of a spectrum of inflammatory demyelinating diseases. This includes acute MS (Marburg variant), Balo's concentric sclerosis, neuromyelitis optica (Devic's disease), and ADEM. Acute MS presents as a fulminant disease that causes severe disability shortly after onset and often death within 1 year. Lesions are more destructive than those of typical MS with axonal destruction and neurosis. Age of lesions is mixed, suggesting that disease is present prior to the clinical diagnosis. Balo's concentric sclerosis is another acute illness that has bands of demyelination alternating with preserved myelin areas. The cerebellum, brain stem, optic chiasm, and spinal cord are often spared. Devic's disease or neuromyelitis optica is defined by optic neuritis and myelopathy occurring together or within a short period of time. Pathologically, necrosis has been evident in both optic nerves and spinal cord, probably as a result of ischemia from swollen tissues.

ADEM is a polysymptomatic disease that occurs after upper respiratory infection or other viral infection or following immunization. The disease occurs more commonly in children than in adults and includes demyelination in multiple areas, including brain stem, spinal cord, optic nerves, cerebrum, and cerebellum. Patients may have associated febrile illness and can become comatose in rare cases. The percentage of patients that ultimately have a diagnosis of MS is low (i.e., 5%), but initial presentations, especially if not fulminant, can be confused with the first event of MS. Pathologically, there is infiltration of mononuclear cells with limited periventricular demyelination.

Other illnesses that present as a relapsing–remitting type of illness include Behcet's disease, characterized by oral or genital ulcerations, iridocyclitis, meningoencephalitis, and thrombophlebitis. Lyme disease typically has radicular or peripheral nerve involvement but can be mistaken for MS because of its relapsing–remitting course and occasionally by its appearance on MRI. Sjogren's syndrome, in which individuals have vasculitis of the skin and peripheral nervous system associated with dry eyes and dry mouth, has also been confused with MS, in part because of MRI abnormalities that may mimic MS. Neurosarcoidosis may be a relapsing–remitting type of illness, but it generally has multiple cranial mononeuropathies. MRI shows leptomeningeal enhancement associated with intracranial disease. Cerebrovascular disease may be confused with MS, especially when interpreting MRI scans. Primary central nervous system vasculitis or systemic lupus erythematosus (SLE) may mimic the disease because of relapsing–remitting-type illnesses and white matter lesions present on MRI. The lesions seen with SLE tend to be subcortical and not periventricular. Vitamin B_{12} deficiency remains in the differential of MS; however, it tends to be a progressive-type disease with symmetric findings consistent with myelopathy and peripheral neuropathy. Myasthenia gravis may fluctuate in severity and could be confused with MS, especially when presenting with diplopia. Other bulbar features are uncommon in early MS, and lack of upper motor neuron findings can be helpful in distinguishing this illness. The hereditary dysmyelinating disorders are usually not confused with MS, although with adult onset of progressive myelopathy one must consider illnesses such as adrenolmyeloneuropathy.

XIII. PROGNOSTIC FEATURES

Much has been written about the natural history of MS. Because effective therapies will alter the natural history of the disease, this information is invaluable. Many patients expect to have a progressive disease when they are told a diagnosis of MS. In reality, 50% of patients will be walking without assistance 15 years after the diagnosis. As mentioned previously, the majority of patients present with relapsing–remitting disease. Approximately one-half of relapsing–remitting

patients experience progression. Many features have been assessed as prognostic markers for future outcome, including the extent of disability at 5 years, age, sex, the extent of initial symptoms, complete or partial remission, type of symptoms at onset (e.g., optic neuritis, sensory symptoms, and cerebellar findings), and attack frequency in the first 2 and 5 years. Favorable prognostic factors are outlined in Table II. Because of the unpredictable nature of the disease, no firm predictions for an individual's course can be determined early in the disease. Of all predictors, the extent of disability at 5 years is most reliable, but patients who have a very mild early course may later develop significant disability. Death as a result of a relapse is extremely rare, although it may occur as a complication of the disease, such as pneumonia or suicide.

XIV. TREATMENT

Advances in biotechnology have had a significant impact on the treatment of MS. For many years, treatment relied on the use of broad-spectrum immunosuppressive agents that had numerous toxicities. Early trials reported success with agents that did not withstand larger or more detailed investigations. Historically, the disease has been difficult to study because of the wide variability in presentation and clinical course. Large numbers of patients are necessary to show a consistent reproducible response from a drug. The gold standard for measuring disease activity has been the Kurtzke Expanded Disability Scale (EDSS). This 10- point scale, with half-step gradients, measures physical disability. Unfortunately, it is not a linear scale and is heavily weighted toward gait, especially in the middle range of the scale. Newer measures of disability include the Functional Composite, which evaluates gait (25-ft timed walk), upper extremity function (9-hole peg test), and cognitive function (Paced Auditory Serial Addition Test). MRI is used as a surrogate marker for disease activity.

Table II
Prognostic Features of Multiple Sclerosis

Favorable	Less favorable
Onset of disease before age 40	Onset after age 40
Monosymptomatic onset	Polysymptomatic onset
Complete recovery of symptoms	Lack of recovery from the first attack
Female sex	Male sex
Optic neuritis or sensory symptoms at onset	Early cerebellar or motor symptoms
Little disability at 5 years	

Genetic engineering has led to the manufacture of compounds such as interferon-β (IFN-β), which is where produced by *Escherichia coli* or mammalian ovarian cells after gene insertion. Copaxone (glatiramer acetate) is a synthetic compound made of four amino acids. These agents allow for selective immunomodulation and thus lower toxicity. There is conclusive evidence from large-scale trials that these agents are partially effective in the treatment of MS. IFN-β-1b (Betaseron) was the first of these agents to be assessed. The drug was shown to effectively decrease the risk of exacerbation of MS and had the effect of slowing accumulation of new lesions on MRI. Although the drug did not show an effect on slowing progression of disease, as measured by the EDSS, the effect on MRI led many to believe that the long-term outcome of a patient treated with this will be better than the expected natural history. IFN-β-1a (Avonex) was the second agent to be approved by the FDA. This drug has a similar effect on decreasing the rate of relapse. The primary end point of the clinical trial was efficacy in slowing disability as measured by the EDSS, and this drug has the unique feature of FDA approval for that indication. Copaxone was initially investigated as a potential agent to induce EAE. Conversely to what was expected in the animal model, there appeared to be a protective effect, which led to additional clinical trials. Copaxone decreases the risk of a relapse and has favorable effects on MRI. Currently, it is difficult to argue that one of these agents is superior to the others for treatment of relapsing–remitting MS. Fortunately, it allows for patients to have an option. Worldwide use and access are limited by the high cost ($12,000 per year) and administration (intramuscular or subcutaneous injections). The National Multiple Sclerosis Society has issued a recommendation advocating the use of these drugs early in the course of the disease based on increasing evidence that early treatment is likely beneficial in decreasing the inflammatory phase of the disease and preventing axonal injury. Recently, trials of secondary progressive MS have found mixed results with use of beta-interferons. Mitoxantrone (Novantrone) has received FDA approval in the United States for treatment of secondary progressive MS. This immunosuppressive agent was shown to be effective in European studies in slowing progression of

individuals with secondary progressive disease or relapsing–remitting disease unresponsive to other therapies.

Future clinical trials will likely lack a placebo group because it is unethical to leave a patient untreated when partially effective therapies are available. As a result, larger and lengthier clinical trials will be necessary to show an effect of newer agents. Surrogate markers for disease activity, including MRI, will be used to assess the utility of drugs prior to large phase 3 trials needed to show clinical efficacy. As our understanding of the pathological processes of the disease increases, it is possible that different therapies will be indicated for different types of MS. For example, in individuals with a notable inflammatory component as determined by MRI or an other tool, use of immunomodulator therapies that have a specific effect on inflammatory aspect of the disease may be most beneficial. In other cases in which little inflammatory disease appears evident, growth factors or other means to stimulate regenesis of cell life may be more appropriate. In cases in which drugs are not shown to be efficacious for MS, the understanding of the pathophysiology of MS increases and thus the trials cannot be considered failures. Finally, until the cause of the disease is found, a cure is not likely. Fortunately, advances in understanding the disease process have been exponential in recent years, allowing hope for control, if not cure, for this disease in the near future.

See Also the Following Articles

BRAIN LESIONS • CEREBRAL WHITE MATTER DISORDERS • GLIAL CELL TYPES • MOTOR NEURON DISEASE

Suggested Reading

Fazelas, F., Burkhot, F., *et al.* (1999). The contribution of magnetic resonance imaging to the diagnosis of multiple sclerosis. *Neurology* **53,** 448–456.

Lucchinetti, C. F., Bruck, W., Rodrigues, M., and Lassmann, H. (1996). Distinct patterns of multiple sclerosis pathology indicate heterogeneity on pathogenesis. *Brain Pathol.* **6,** 259–274.

McDonald, W. I., Compston, A., Edan, G., *et al.* (2001). Recommended Diagnostic Criteria for Multiple Sclerosis: Guidelines from the International Panel on the Diagnosis of Multiple Sclerosis. *Ann. Neurol.* **50,** 121–127.

Noseworthy, J. (1999). Progress in determining the causes and treatment of multiple sclerosis. *Nature* **399**(Suppl.), A40–A46.

Paty, D. W., Noseworthy, J. H., and Ebers, G. C. (1998). *Diagnosis of Multiple Sclerosis* (D. W. Paty and G. C. Ebers, Eds.), pp. 48–134. Davis, Philadelphia.

Waxman, S. G. (1998). Demyelinating diseases—New pathological insights, new therapeutic targets. *N. Engl. J. Med.* **338**(5), 323–325.

Weinshenker, B. G. (1998). The natural history of multiple sclerosis: Update 1998. *Sem. Neurol.* **18,** 301–307.

Multisensory Integration

BARRY E. STEIN, PAUL J. LAURIENTI, MARK T. WALLACE, and TERRENCE R. STANFORD
Wake Forest University School of Medicine

GLOSSARY

cross-modal From two or more different sensory modalities. Used here to refer to (i) combinations of stimuli from different sensory modalities (e.g., a combination of light and sound) that normally evoke different subjective experiences, (ii) the spatial register among the different receptive fields of a multisensory neuron, and (iii) the spatial register among different sensory maps.

modality specific From a single sensory modality (i.e., unimodal). Used here in two forms: (i) to categorize a neuron based on the stimuli to which it can respond (e.g., a neuron responsive only to light would be a modality-specific neuron) and (ii) to categorize a particular neuronal response regardless of the type of neuron from which it is evoked (e.g., a response to light, even in a multisensory neuron, is a modality-specific response).

multisensory Refers to neurons that are capable of responding to stimuli from more than a single sensory modality and to the neural processes associated with these responses (e.g., multisensory integration).

multisensory integration A statistically significant difference between the number of impulses evoked from a neuron by a combination of cross-modal stimuli compared to the responses to the most effective of these stimuli individually.

qualia (singular quale) Bits of experience; used here to refer to the particular subjective impressions that are associated with a given sensory modality (e.g., hue, pitch, tickle, and itch).

receptive field The area in space (or on the body) in which a stimulus can evoke a response from a particular neuron.

topographic Used here to define a systematic arrangement of the components of a sensory or motor representation in the brain (e.g., the arrangement of the visual receptive fields into a map of visual space).

Multisensory integration refers to the brain's ability to synthesize the information that it derives from two or more senses. There are a number of other terms in use, such as cross-modal integration and intersensory integration, but all refer to the same phenomenon. The longstanding interest in the underlying neural mechanisms by which multisensory integration is accomplished is due, in large part, to its remarkably robust perceptual and behavioral products, products which cannot be understood by dealing with the senses as independent of one another. The present discussion will use the individual multisensory neuron in the midbrain as a model, and will focus on how its responses are altered in predictable ways by converging inputs from the visual, auditory and somatosensory modalities.

I. CROSS-MODAL PERCEPTION

One of the driving forces in evolution has been the creation of multiple means by which animals can sample their environments. This has involved the development of numerous sensory systems that are responsive to very different forms of physical energy, because a primary imperative of survival is knowledge

of environmental circumstances, information that is as important for the location and capture of prey as it is for the avoidance of danger. Extant animals maintain a rich variety of such systems, thereby increasing the probability of their survival and the circumstances in which they can flourish. Being able to detect many of the same events with different sensors allows some to substitute for others when the circumstances require it. For example, touch and hearing can substitute for vision in the darkness. The combination of inputs from sensors that can function simultaneously also provides an animal with multiple "views" of a single event. As a result, the inherent ambiguities that may be present when evaluating an event along one sensory dimension (e.g., how it looks) can be obviated by the additional information provided along another sensory dimension (e.g., how it sounds).

Our sensory systems function so well in these regards that we have come to have great faith in the accuracy with which they reflect the properties of the physical world. Consequently, we are sometimes amused and sometimes distressed to learn that our sensory judgments are not absolute and can vary considerably in different situations. However, it appears that our reactions to this knowledge differ in different contexts, and for most people it is far less surprising when errors are made in estimating the intensity of a sound because of the presence of background noise than when a sound is mislocated because of the confounding influence of a visual stimulus. Perhaps this is because we are more familiar with the relativity of perceptual judgments within a given sensory modality than across different sensory modalities. We readily accept the fact that visual judgments are influenced by context in the visual arts—painting, photography, and film—and we are not at all disturbed by the use of perspective and shading to alter our judgments of dimensionality, relative distance, or even the stature of a vertically challenged actor. However, finding that a sensory judgment can be substantially altered by seemingly irrelevant cues from another sensory modality is somehow more surprising and less intuitive. Nevertheless, cross-modal influences on perception are both common and potent. The presence of a neutral auditory cue can make a dim light appear substantially brighter, the vibration of the muscles on one side of the neck can make a stationary light in a darkened room appear to be displaced or even appear to be moving toward the contralateral side, rotation of the body can make a horizontal line appear oblique, and changing gravitational forces can seriously disrupt our localization of visual cues.

Of course, cross-modal influences are not restricted to altering visual judgments; judgments in all sensory modalities are subject to such influences. A common example of the sensitivity of the vestibular system to the effect of visual cues is evident in modern wide-screen cinematography, in which films are made from the perspective of a pilot flying much too low and much too fast over steep mountain peaks and into deep canyons or from the front seat of a roller coaster. Inevitably, the viewer experiences the same vestibular and gastrointestinal sensations that make clear air turbulence and high seas so amusing.

The general impression that different senses function in entirely separate realms and are thus unlikely to interact with each other may also be partly a consequence of the unique subjective impressions associated with each of them. These impressions, called "qualia," are modality specific. For example, the perception of hue or color is specific to the visual system. Although we may speak of the smell or taste of green, this is really a consequence of associations that are developed through experience because there is simply no nonvisual equivalent to the experience of "green." This becomes quite obvious when trying to describe color to a congenitally blind person. The same sort of problem would be encountered in trying to describe music to someone who is congenitally deaf because pitch is unique to the auditory system. Similarly, the sensations of tickle and itch are peculiar to somatosensation, and there is no nontaste equivalent to "salty," nor a nonolfactory equivalent to "burnt almond." The unique nature of these qualia has been the subject of discussion for much of the history of neuroscience and led Johannes Muller to propose a theory of "specific nerve energies" in 1826. According to Muller, each subjective impression is attributable to the activation of modality-specific nerve fibers and their target neurons in the brain, a concept that has been supported by studies in which electrical stimulation of modality-specific cortical regions in patients (which is helpful in avoiding sensory areas during brain surgery) produced the appropriate modality-specific sensations.

These observations might lead one to expect that the brain must avoid cross talk between the modalities in order to maintain their unique qualia. However, this is not the case. Not only do the senses regularly affect each other, as indicated by the cross-modal influences on perception noted previously, but when cross-modal

stimuli are slightly discordant, this cross talk can often result in any of a host of unexpected and intriguing illusions. One of the best known of these is the ventriloquism effect, and few people have not delighted in the illusion that a wooden doll is actually speaking. The name of the effect is derived from the world of entertainment, but the phenomenon properly refers to a broad class of events in which a visual cue produces the apparent translocation of an auditory cue. Although this illusion need not involve speech, the impact of television and movies has rendered speech translocation its most common form. Each voice in a movie or television program seems to come from the appropriate character regardless of his/her location on the screen. However, in reality all sounds are derived from the same location: speakers at the sides of the television cabinet or the movie screen. This underscores the fact that despite the apparent skills of some entertainers, the ventriloquist's trick is less a reflection of his ability to ''throw his voice'' than the susceptibility of the brain to use the visual system to localize external events. Thus, the lip movements and gestures associated with speech produce the compelling illusion that the corresponding sound must be coming from the source of those movements, even if it is a dummy.

Among speech professionals, the so-called McGurk effect reigns supreme. In this illusion, the meaning rather than the location of the auditory signal is altered by lip movements. Its popularity is due to its ability to dramatically demonstrate how important the integration of visual and auditory cues is to speech perception and the significance of this cross-modal integration when, for example, trying to understand a spoken message in a noisy room. The integration of visual and auditory signals has also recently been shown to play a significant role in nonhuman primate communication. Originally, the McGurk effect was shown by having the lips of the demonstrator form a syllable, for example, ''ga,'' but the sound associated with the lip movements (prerecorded and played through a speaker in synchrony with them) was ''ba.'' The listener perceived ''da,'' a percept that represents neither of the original cues but rather the product of their synthesis. The McGurk effect has also been shown to be readily induced by using strings of spoken or visible (lip movements) sounds, each of which has no meaning by itself but when combined they result in an intelligible sentence (Fig. 1). It has even been shown to work with words that already have meaning, but their combination results in the perception of an entirely different word (e.g., hearing ''bows'' and seeing ''goes'' results in the percept ''doze'').

There are many cross-modal illusions, some of which seem like curiosities and others that must be fully appreciated because they can confound judgments in life-threatening circumstances. A recently described auditory–tactile illusion that is particularly compelling, but that is unlikely to produce significant problems in daily life, is the so-called ''parchment-skin'' illusion. In this case, a subject is fitted with

Figure 1 An example of the McGurk effect. Here, the pairing of a nonsense phrase delivered from the speaker (''my bab pop me poo brive'') combined with the visual cues of a person whose lips are forming a different nonsense phrase (''my gag kok me koo grive'') results in the meaningful percept: ''My dad taught me to drive'' (adapted and reprinted with permission from Stork and Massaro, *American Scientist* **86,** 236–244, copyright 1998).

headphones and a microphone, which is routed through the headset, is placed near one of his hands. Then, the subject rubs his fingers together and the sound of the rubbing is amplified and played through the headphones. The result is that the subject perceives his skin to feel rough and dry, and the degree of roughness perception changes with variations of the composition of the sound frequency. On the other hand, visual–vestibular illusions are of particular importance to aircraft pilots, who must correctly judge the speed with which the nose of the aircraft is rising despite the potent, and often confusing, influences of vestibular cues produced by the strong gravitational forces of the high-speed takeoff. Errors in judgment in this circumstance can have severe consequences for both pilot and plane.

The few cross-modal illusions described here should be sufficient to emphasize the point that changing the normal relationships among cross-modal stimuli can have substantial disruptive effects on perception. Sensory systems have evolved to function in concert, and the neural integration of their inputs, as well as the perceptual products that result from this integration, reflect the normal mode of brain function. Some of the direct behavioral consequences of the integration of cross-modal signals involve both an increase in the probability of detecting an external event and a substantially more rapid response to it. Often, these changes in detection and reaction speed exceed statistical models based on reactions to each of the individual modality-specific stimuli alone.

One of the key features that can be used to predict whether cross-modal stimuli will produce enhanced neural, perceptual, and/or behavioral responses is whether or not these stimuli are derived from the same event because such cues are generally coincident in both space and time. The relative timing of the different stimuli is critical because if cross-modal stimuli are sufficiently disparate in time, the brain does not integrate them, but treats them as separate events. However, from the perspective of neural function, the temporal window during which this information can be integrated is very long, lasting hundreds of milliseconds. This makes it possible for the nervous system to integrate, for example, visual, auditory, and somatosensory cues that are derived from the same event despite the fact that these inputs arrive at a multisensory neuron at very different times. The difference in arrival time is due to the nature of the stimuli and the nature of the nervous system. Somatosensory cues do not travel in space; they are delivered to the skin of the observer and have a reasonably short conduction time to the brain. Sound is also rapidly transmitted from the ear to the brain, but it travels very slowly in space and takes an appreciable time to reach the ear. Although for all practical biological purposes light transmission is instantaneous, it requires a good deal of processing time in the retina before it can be sent to the brain.

The spatial relationships among cross-modal stimuli are also critical, not only for determining whether a multisensory interaction will take place in the brain but also for determining the kind of interaction that will occur. Spatially coincident stimuli generally result in enhanced brain signals and enhanced perception and behavior; spatially disparate stimuli either produce no interaction and are treated as separate events or they inhibit one another and degrade brain signals as well as their perceptual and behavioral products. It is when these spatial and temporal relationships are neither so close that they produce "normal" products nor so discrepant that they are treated as separate events that aberrant neural, perceptual, and behavioral consequences result. The illusions discussed previously are good examples of some of these "anomalous" products.

The integration of cross-modal inputs (multisensory integration) is also responsible for providing unity to our perceptions. This observation was obvious to Aristotle and was discussed in *De Anima*. Aristotle was one of the first to grapple with the concept of multisensory integration by noting that although each of our five senses provides us with different information, the resultant perception is of a single world. He concluded that there must be a mechanism by which the disparate information that is obtained from different sense organs is brought together into a unified whole, a view that anticipated modern discussions of what has become known as the "binding" problem. If we state this problem from a modern neurological perspective, it might be posed as follows: How can we perceive objects as unitary entities when their individual features are processed separately in different populations of neurons in different regions of the nervous system? This issue is as germane to within-modal issues as to cross-modal issues, because individual components of a modality-specific stimulus (e.g., its motion, direction of movement, and color) are dealt with by different populations of neurons in different regions of the brain, just as different modalities engage different populations of neurons in different regions of the brain (e.g., their different modality-specific pathways). Although there is no widely accepted theory explaining how the brain solves this "problem," there

are a number of reasonable possibilities that are currently being considered.

II. MULTISENSORY INTEGRATION: EVOLUTIONARY CONSIDERATIONS

We are far from ignorant about some of the neural mechanisms that underlie the integration of information from different sensory modalities. The study of single neurons that receive inputs from more than a single sensory modality has produced a growing body of information about the sites in the nervous system in which cross-modal convergence takes place as well as some of the mechanisms by which information from different sensory modalities is brought together and integrated. The manner in which these multisensory neurons deal with their converging inputs and the consequences of their action are among the subjects of this article. However, before dealing with this issue, it is important to note that although we often tend to frame questions about the phenomenon of multisensory integration in terms of human experience (especially when considering its perceptual consequences), its appearance did not await the evolution of *Homo sapiens*. Rather, it is an ancient scheme of information processing that is likely to have been present in the presumptive primordial unicellular organism. The caveat "presumptive" is used here because no one can say with certainty what that unicellular organism was like. Nevertheless, it is reasonable to suppose that it, like its modern counterparts, had different sensory receptors tuned to the transduction of chemical, mechanical, and/or photic energy, and that each of these sensory transducers used the same signaling mechanism (an electrical current produced by the movement of ions across the cell's membrane). Consequently, the primordial unicellular organism, like its extant cousins, was a multisensory organism. Because the various currents produced by activating its different sensory receptors can readily affect one another, unicellular organisms are obligatory multisensory integrators. Simple multicellular organisms are also unable to sequester specific sensory signals because of the nature of their intercellular connections, and the segregation of at least some modality-specific information probably did not occur until the appearance of advanced invertebrates. Thus, multisensory integration, rather than sensory segregation, is likely to have been the initial form of sensory information processing.

The capacity for multisensory integration was retained in some form throughout the evolution of multicellular organisms, and there is no known beast, vertebrate or invertebrate, in which there exists a complete segregation of sensory signals on a sense-by-sense basis. Presumably, the ability to derive information from the combined action of different senses that would be unavailable from their individual action had considerable survival value and thereby engendered the maintenance and elaboration of systems capable of multisensory synthesis. However, during the evolution of complex species, an interesting duality was formed;

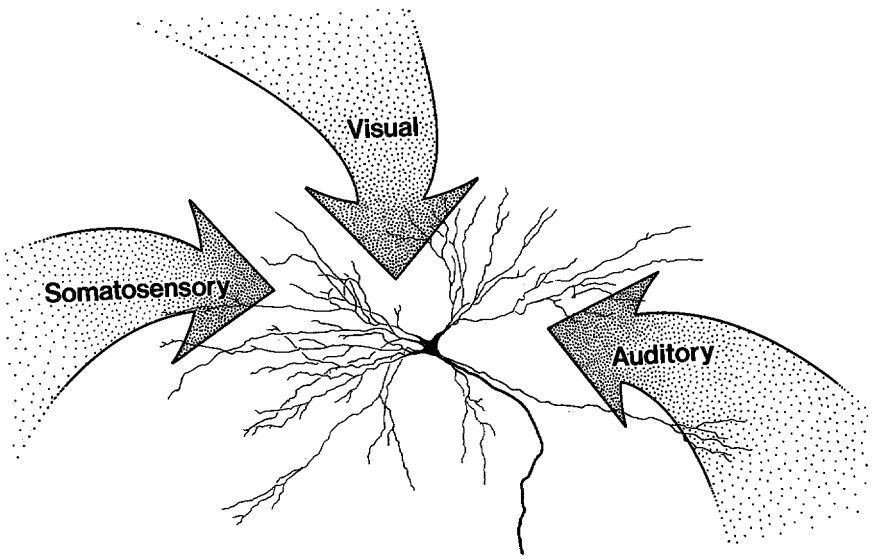


Figure 2 Inputs from multiple sensory modalities converge on an individual neuron in the superior colliculus (reprinted from Stein and Meredith, "The merging of the senses," 1993, with permission from The MIT Press).

some regions of the nervous system became specialized for processing information based on its modality and others for processing information regardless of the modality from which it was obtained. Presumably, it is the action of the former systems that accounts for the sensory qualia referred to previously. The preeminent example of such a system is the primary projection pathway (but see Section VII for evidence that cross-modal influences are evident even in the primary projection systems). In the vertebrate visual system, for example, this involves the projections from the retina to the primary thalamic relay station (the lateral geniculate nucleus) and from there to primary visual cortex. Multisensory areas exist outside the primary projection pathways and have been found at every level of the neuraxis.

In light of the importance of multisensory information processing for normal perception and behavior, it is surprising to note that we are only beginning to understand the neural bases by which the brain accomplishes this feat. When compared to our understanding of how modality-specific sensory cues are processed, our knowledge of multisensory integration seems rudimentary. However, we do know that whenever different modality-specific inputs converge at the level of the single neuron (Fig. 2) there is the opportunity to synthesize them.

III. THE SUPERIOR COLLICULUS: A MODEL FOR UNDERSTANDING THE NEURAL BASES OF MULTISENSORY INTEGRATION

Perhaps the brain's best studied multisensory neurons are those in the superior colliculus (SC). This structure, which is located on the surface of the midbrain (Fig. 3) plays its primary role in attentive and orientation behaviors. It uses visual, auditory, and somatosensory information individually or in concert to direct the peripheral sensory organs toward an object or event of interest. Its best known role is in the initiation and control of eye and head movements (collectively referred to as gaze) in order to fixate a target; however, in species with highly mobile ears, it has also been shown to be involved in generating coordinated and directed ear movements. Recent evidence has also implicated the SC in the control of some limb movements. The importance of the SC in the current context is that it has been used as a general model for understanding the neural principles of multisensory integration and relating them to overt behavior.

The SC receives its inputs from many sources, both sensory and motor. Most of the sensory structures that project to the SC contain neurons responsive only to a single sensory modality—visual, auditory, or somatosensory. Thus, rather than reflecting multisensory information that is preprocessed elsewhere, the SC is a primary site of cross-modal convergence. This is underscored by the finding that although association cortex, one of the major sources of input to the SC, contains both modality-specific and multisensory neurons, only its modality-specific neurons project to the SC. Despite the counterintuitive nature of this finding (it would seem fitting to have multisensory neurons communicate directly with other multisensory neurons), the independence of this cortical circuit has been an experimental asset. It has allowed investigators to test whether the principles of multisensory integration found in the SC also apply to this independent multisensory circuit and, by extension, to other multisensory regions of the brain.

A. Single-Neuron Studies

1. Topographic Register of Sensory and Motor Representations

Regardless of their origin, the visual, auditory, and somatosensory inputs that reach the SC are arranged systematically to form topographic, or map-like, representations. These sensory maps, like those found in other areas of the brain, are based on the organization of receptive fields (Fig. 3). A neuron's receptive field defines that region of space from which it can be activated by a sensory stimulus, and the arrangement of neurons in the SC is such that adjacent neurons have their receptive fields in adjacent regions of space. When one speaks of a topographic or map-like representation, a frame of reference is always implied, if not stated explicitly. For example, in primary visual structures, the position of a neuron's receptive field is indexed relative to the fovea, which defines the center point of the axes of "visual space." Similarly, the receptive field of a neuron in a primary auditory structure would be referred to the center of the head, and in the case of a somatosensory neuron the referent would be the body surface (e.g., face, hand, and leg). One could write volumes about how a topographic representation is an efficient way to encode information about the nature of objects as defined by the spatial relationships of their components (e.g., size, shape,

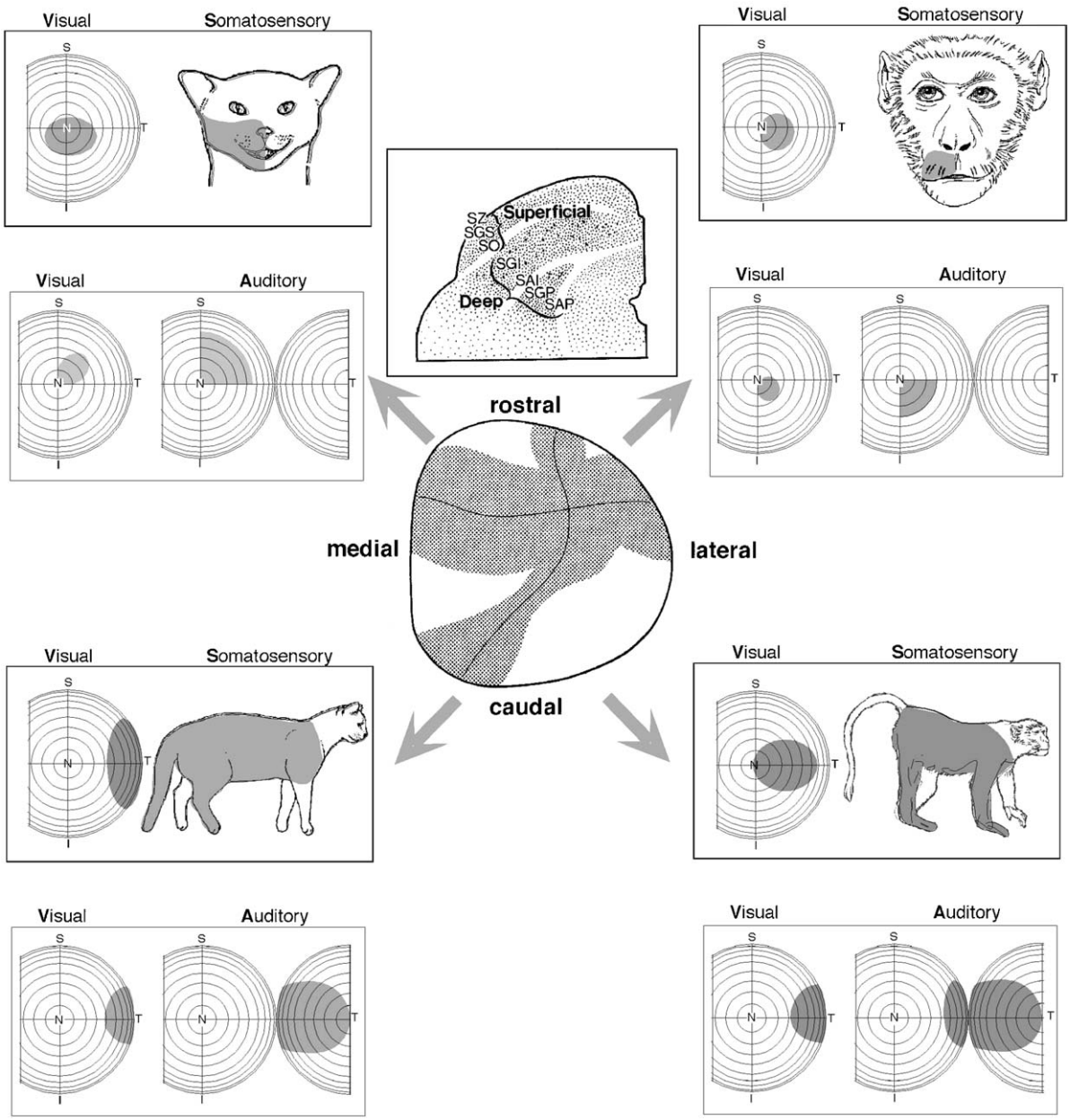


Figure 3 Visual, auditory, and somatosensory receptive fields are organized in map-like fashion in the superior colliculus (SC) and all share the same axes. The overlapping nature of these representations is particularly evident in individual multisensory neurons. Shown here at the center is a schematic of the left SC as viewed from the top of the brain. Rostral aspects of the SC contain receptive fields that are found in frontal space, caudal locations contain receptive fields in temporal or caudal space, etc. (see text for further detail). The curved line running from the rostral to the caudal SC is the horizontal meridian, which divides the representation of superior from the representation of inferior sensory space. Similarly, the curved line running from medial to lateral represents the vertical meridian, which divides ipsilateral and contralateral sensory space. Representative receptive fields (shading) for multisensory neurons from cat (left) and monkey (right) SC are drawn from rostral (top) and caudal (bottom) sites in the SC. Each concentric circle in the schematics of visual and auditory space represents 10°. In the representation of auditory space, the caudal half of contralateral space is shown by the split hemisphere that has been folded forward. Note the register among the receptive fields for visual–somatosensory and visual–auditory neurons. Visual receptive fields in nasal (frontal) space on the right are linked to somatosensory and auditory receptive fields in frontal space on the right. Similarly, temporal (or caudal) receptive fields in one modality are linked to similar locations in other modalities. Multisensory neurons are found in the deep layers of the SC. These layers are illustrated on a schematic of a coronal (vertical) section through the structure. S, superior; I, inferior; T, temporal; N, nasal; SZ, stratum zonale; SGS, stratum griseum superficiale; SO, stratum opticum; SGI, stratum griseum intermediale; SAI, stratum album intermediale; SGP, stratum griseum profundum; SAP, stratum album profundum.

and texture). However, the ''sensory'' topographies found in the SC do not serve this purpose and are thus quite distinct from those found in primary sensory structures. The principal function of the SC is to generate motor commands for the purpose of orienting to a sensory stimulus. Thus, whether one is dealing with a visual, auditory, or somatosensory stimulus, the problem is the same: to determine where the stimulus is with respect to the part of the body that must be oriented to it. For example, knowing where a visual stimulus is with respect to the fovea is not sufficient for programming a movement to reach out and grasp the object; for this, one needs to know where the stimulus is with respect to the hand. Similarly, knowing the location of an auditory stimulus with respect to the head is not sufficient to program an eye movement (i.e., gaze shift) to look at the stimulus; one needs to know where the stimulus is with respect to the current position of gaze. In other words, the topographic referent must be based on the structure or body part that is to be oriented toward the stimulus.

The SC is known to be involved in the generation of many types of orienting movements. The best established of these motor representations is the one that controls gaze. Therefore, the following description of the motor topography is based on what is known about gaze control, although in theory similar schemes could be constructed with ear or limb movement maps. Within the SC there is a gaze topography so that the site of activity codes for the distance and direction of a gaze shift. The components of this ''motor'' map are neurons that each have a ''movement field,'' the motor analog of a sensory receptive field. Neurons with movement fields discharge in association with gaze shifts within a particular range of amplitude and direction and are arranged systematically according to this movement range; neurons that represent small contralateral movements are located in rostral SC; neurons representing progressively larger contralateral movements are located in progressively more caudal aspects of the structure. From medial to lateral in the SC the gradient is from neurons representing movements with upward directions to those representing movements with downward directions. The relationship between the SC sensory and motor representations is summarized as follows: The locus of sensory-evoked activity represents the location of a sensory stimulus with respect to the current gaze position; the location of motor-related activity represents the amplitude and direction of the movement required to shift gaze from the current position toward the location of the stimulus.

Tying the sensory topographies to the position of gaze ensures that the activity evoked by a sensory stimulus activates a region of the SC in which motor-related activity would be appropriate for shifting gaze to the stimulus. It also means that visual, auditory, and somatosensory stimuli share the same reference frame and will activate the same region of the SC if they originate from the same location in space. As such, the maps of visual, auditory, and somatosensory modalities will always be in register, regardless of the relative position of the eyes, head, and body to each other. The general alignment of these representations has been demonstrated in studies carried out in anesthetized animals with eyes, ears, head, and body facing forward so that the axes of visual, auditory, and tactile space are in approximate alignment with the direction of gaze. For each sensory modality, receptive fields shift systematically across space as one samples neurons in any given direction across the SC. For example, visual or auditory neurons in the rostral aspect of the SC have their receptive fields in central contralateral space (consistent with sites in the motor map that produce small contralateral gaze shifts), whereas those located progressively further rearward (caudal) in the structure have their receptive fields shifted progressively more eccentric into the peripheral aspects of contralateral space (in register with sites that produce larger contralateral gaze shifts). In short, when the animal is facing forward, neurons in the front of the structure represent the space in front of the animal and those in the rear of the structure represent space in the periphery. The somatosensory representation corresponds to this organization in that neurons with receptive fields on the face are located rostral in the SC and those with receptive fields progressively further back on the body (toward the rump) are located progressively more caudal. Neurons found medial in the structure have receptive fields in upper space (or on the upper body) and laterally located neurons have receptive fields progressively lower in space (or lower on the body).

Because many SC neurons are responsive to stimuli from more than one sensory modality, they contribute to the formation of multiple sensory representations in the SC. There are several varieties of SC multisensory neurons, and the order of incidence of bimodal neurons is as follows: visual–auditory, visual–somatosensory, and auditory–somatosensory. Although trimodal (visual–auditory–somatosensory) neurons are sometimes encountered, their incidence is

comparatively low. However, regardless of its specific modality convergence pattern, each multisensory neuron has multiple receptive fields (Fig. 3), one for each of the sensory modalities to which it responds. As would be expected based on the topographic register described previously, the different receptive fields of each multisensory neuron correspond with each other (Fig. 3). One of the consequences of this receptive field correspondence is that any of the different sensory cues that are derived from the same event can activate the same neurons and thereby access the same circuitry to evoke a given orientation response. As discussed later, this cross-modal register of receptive fields is crucial for proper integration of the information derived from different sensory modalities.

2. Response Enhancement and Response Depression in Multisensory Neurons

Multisensory neurons do more than simply respond to a variety of different sensory stimuli: They transform the signals arriving from multiple sensory channels into an integrated multisensory product. In operational terms, multisensory integration is defined as a statistically significant difference between the number of impulses evoked by a cross-modal combination of stimuli (e.g., a visual and an auditory stimulus) and that evoked by the most effective of these stimuli alone. The integration results in either a response enhancement that can exceed the arithmetic sum of the individual modality-specific responses (Fig. 4), or a response depression that can eliminate responses altogether. The specific multisensory response that is achieved depends on the spatial relationships among the cross-modal stimuli. When two cross-modal stimuli are derived from the same event, they originate from the same location in space. Because the different receptive fields of the same multisensory neuron overlap one another in space, such an event can stimulate both excitatory receptive fields. These two inputs interact synergistically, and the degree to which the response is enhanced (above that evoked by the best modality-specific stimulus) depends on the relative effectiveness of these stimuli; combinations of weakly effective modality-specific stimuli usually result in the largest multisensory enhancements. Thus, multisensory enhancements are particularly beneficial when modality-specific stimuli are weak. If, on the other hand, the same cross-modal stimuli are spatially disparate, as would happen if they were derived from separate events, and one of them falls outside its receptive field, either no interaction results (the extrareceptive field stimulus produces no input to the neuron) or response depression is produced. The latter effect is generated when the extrareceptive field stimulus falls into an inhibitory region that borders the excitatory receptive field of some SC neurons, thereby generating an inhibitory input powerful enough to degrade or suppress the excitation produced by a stimulus within its receptive field.

B. Behavioral Studies

Behavioral experiments have been conducted to determine whether the cross-modal stimulus configurations that enhance multisensory responses at the single-neuron level also enhance multisensory responses at the behavioral level and whether the cross-modal stimulus configurations that reduce multisensory responses at the single-neuron level also degrade multisensory responses at the behavioral level. In these experiments groups of cats were trained in a simple perimetry device (Fig. 4), conceptually similar to the device used by ophthalmologists to map a patient's visual field. Each animal learned to fixate on a point directly ahead. If a light came on anywhere in that perimetry, the animal's task was to break fixation, look at the light, and immediately approach it. The reward for correct behavior was a food treat. Some of these animals were taught to respond in the same way to a brief sound (group 1). Others were trained to ignore the sound (group 2) by never being rewarded for responding to it, and still others had no auditory training (group 3). To examine the possibility that response enhancement would be induced with spatially coincident cross-modal stimulus pairs, it was necessary to keep performance to the individual stimuli below 100%. Thus, when testing, the intensities of the stimuli were lowered so that they were very difficult to detect, and correct performance for either one alone was 50% or less. When the light and the sound were presented simultaneously and in spatial register, correct responding was enhanced dramatically, far more than predicted based on performance in response to either cue individually. On the other hand, even when the intensity of the visual stimulus was increased so that it could be detected on every trial, when the auditory cue was spatially disparate to the visual cue there was a dramatic drop in performance. This was true not only for animals that learned to ignore the sound during training but also for animals that were never exposed to it during training (only groups 2 and 3 were tested for response depression). These results indicate that

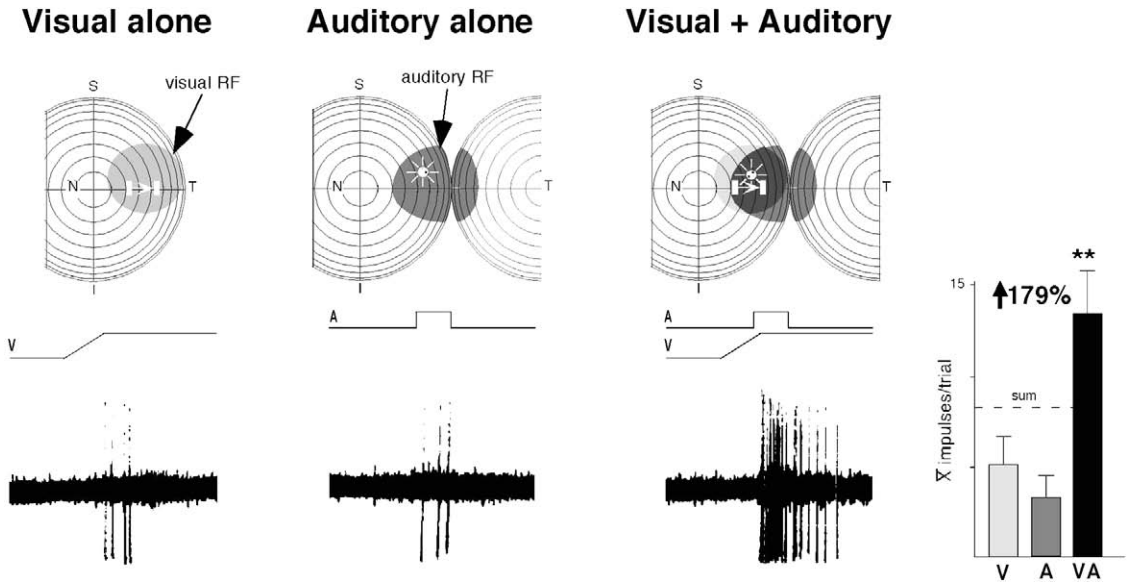


Figure 4 Multisensory response enhancement. Observations are taken from studies of a visual–auditory neuron. Receptive fields (shading) and the locations of the visual and auditory stimuli (icons) within these receptive fields are shown in the schematics at the top (see Fig. 3 for conventions). The visual stimulus (V) was a moving bar of light, and the auditory stimulus (A) was a noise burst. Below the receptive fields are shown the neuronal responses to the visual stimulus alone (left), the auditory stimulus alone (center), and the visual–auditory combination (right). The ramps above the oscillograms (V) represent the movement of the visual stimulus across its receptive field, and the square wave (A) represents the onset, plateau, and offset of the auditory stimulus. The oscillographic traces at the bottom show the neuronal impulses that were evoked by these stimuli on a single trial. The summary bar graph on the right shows the averaged response for each of the conditions as well as the predicted sum based on the addition of the two modality-specific responses. Note that the magnitude of the multisensory response enhancement was greater than predicted by the sum of the modality-specific responses. ** Statistical significance ($p < 0.01$) (reprinted from Wallace *et al.*, *J. Neurophysiol.* **80,** 1006–1010, copyright 1998, with permission from The American Physiological Society).

the same principles that govern multisensory responses at the level of the single SC neuron are operative at the level of overt orientation behavior.

IV. THE ROLE OF ASSOCIATION CORTEX IN MEDIATING MULTISENSORY INTEGRATION IN THE SUPERIOR COLLICULUS

A. Single-Neuron Studies

It might reasonably be assumed that whenever different sensory inputs converge on a single neuron, that neuron is immediately rendered capable of integrating those inputs. In this way, any of the various cross-modal convergence patterns that are possible among the more than 40 subcortical and multiple cortical inputs to the SC would be effective in creating the substrate for multisensory integration. However, this assumption proved to be incorrect. In cat, one of the primary models used for understanding the neural bases of multisensory integration, it has been estimated that 20% of the multisensory neurons in the SC are incapable of synthesizing their cross-modal inputs (i.e., "nonintegrative"). These neurons respond quite well to different modality-specific inputs and come in all varieties (visual–auditory, visual–somatosensory, auditory–somatosensory, and trimodal), but their responses to cross-modal combinations of cues are no greater or less than their responses to the most effective of these stimuli alone. A key component of the SC circuitry underlying multisensory integration appears to be the inputs from association cortex.

Inputs from two regions of association cortex have been found to be critical in this regard: the anterior ectosylvian sulcus (AES) and the rostral portion of the lateral suprasylvian sulcus (rLS). This conclusion is based on studies in which the modality-specific and multisensory responses of cat SC neurons were studied before, during, and after reversibly blocking (via cryogenic deactivation) the neuronal activity in these areas. The profound effect of blocking these cortical inputs is illustrated in Fig. 4. In this example, blocking AES activity eliminated the multisensory response enhancement in the SC. However, as was typically

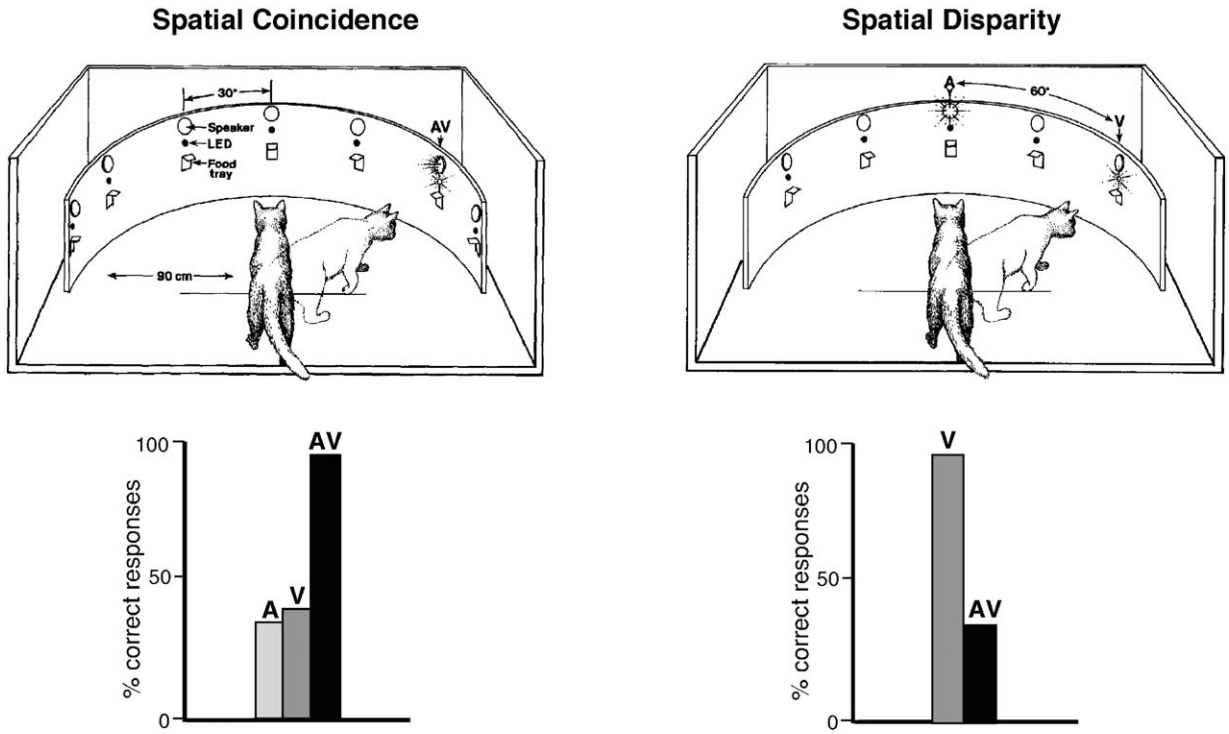


Figure 5 Orientation behaviors are altered by multisensory stimuli such that cross-modal stimulus combinations at the same location enhance performance while those at different locations impair performance. Cats were trained in a perimetry device (top) to attend straight ahead and then to orient to and approach a flashed light-emitting diode (LED) or a brief noise burst (see text for details). When the stimuli were of low intensity, they were difficult to notice and the animal responded correctly to them in less than 50% of the trials (left). However, when they were presented simultaneously and at the same location (AV) performance was enhanced more than predicted by their sum. Another animal was trained to respond only to the visual stimulus (it was never presented with the auditory stimulus during training). Its responses to the LED were markedly impaired when the neutral auditory stimulus was presented 60° medial to the LED, and this was most evident when correct responses were high (right). In this training paradigm performance could also be enhanced when the LED was made dimmer and the two stimuli were presented at the same location (not shown here) (reprinted from Stein *et al.*, *J. Cogn. Neurosci.* **1,** 12–24, copyright 1989, with permission from The MIT Press).

found, the influences from the AES (or rLS) were quite selective; they were critical only for multisensory integration. The neuron's responses to modality-specific stimuli were not significantly altered.

Some SC neurons depend only on influences from AES to exhibit multisensory integration, and others depend only on influences from rLS. However, many receive combined influences from these two cortical areas, and both cortices influence the multisensory integration of such SC neurons. In most of these cases, it is necessary to maintain the influences from both cortices; if either one is removed, the SC neuron loses its capacity for multisensory integration. It is not clear why these different patterns of dependence on AES and rLS among multisensory SC neurons have emerged, but it does raise the possibility that one cortex could compensate for early damage to the other.

B. Behavioral Studies

Whatever the reason that two cortical areas have assumed shared control over multisensory integration in SC neurons, it is interesting to note that these "association" cortices achieve at least some of their associative functions via a very distant subcortical structure. This allows them to influence, quite directly, SC-mediated behaviors. One might expect, then, that the same perturbations that compromise multisensory integration in single SC neurons would also compromise SC-mediated multisensory behaviors.

This possibility was tested experimentally using a paradigm similar to that described earlier. Although cats were trained to respond as described previously, their performance was assessed when association cortex was intact or temporarily deactivated. Cortical

deactivation was accomplished either by locally anesthetizing an area via lidocaine injections through chronically implanted cannulas or by cryogenic blockade of the activity in that area via implanted cooling probes. Deactivating association cortex had no effect on orientation to an individual modality-specific cue, but the spatially coincident combination of visual and auditory stimuli no longer could enhance performance and spatially disparate cues could no longer degrade performance. These effects were not seen when other visual or auditory cortical regions were deactivated. That these effects were dependent on disrupting SC multisensory integration by depriving its neurons of critical cortical inputs, and were not due to the loss of association cortex *per se*, was evident from the results of experiments in which the functional integrity of the SC was destroyed with a neurotoxin. In these cases, the enhanced multisensory behaviors were lost in the absence of the SC despite the functional integrity of association cortex. However, modality-specific behaviors were unaffected. The results of behavioral and electrophysiological experiments underscore the dependence of SC multisensory integration on influences from association cortex.

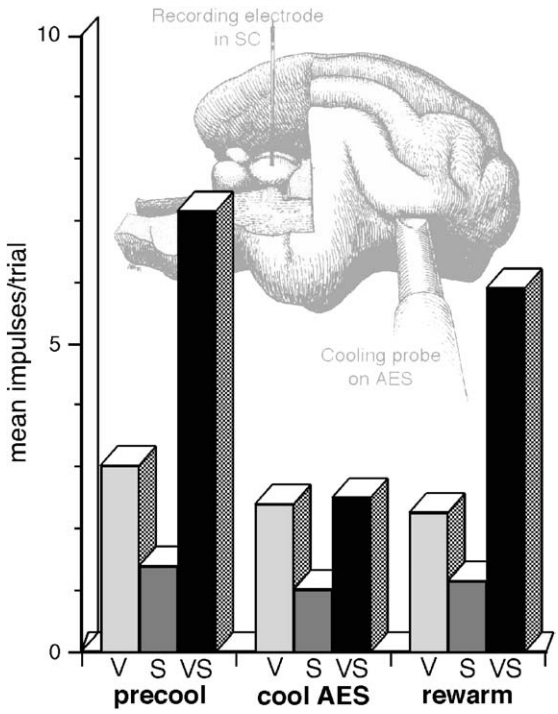


Figure 6 Deactivation of association cortex disrupts multisensory integration in the SC. The shadowed image shows the electrophysiological paradigm. While recording from a multisensory neuron in the SC, the anterior ectosylvian sulcus (AES) was reversibly deactivated by cooling. Representative data are shown for a single SC visual–somatosensory neuron. Prior to AES deactivation (precool), this SC neuron showed a modest response to the visual and somatosensory stimuli when presented individually and a large response enhancement when they were paired. When the AES was deactivated, the neuron's modality-specific visual and somatosensory responses remained intact, but its multisensory response enhancement was abolished. Multisensory response enhancement was reinstated after rewarming the AES.

V. DEVELOPMENT OF MULTISENSORY INTEGRATION IN THE SUPERIOR COLLICULUS

Scientists have argued for some time about whether the newborn brain is already capable of engaging in multisensory integration or if it requires postnatal sensory experience to do so. In the perceptual literature this issue has sometimes revolved around a syndrome called synesthesia. The literal meaning of synesthesia is "joining of the senses" and, in synesthetes, strong modality-specific stimuli initiate not only the "proper" sensations but also sensations that are appropriate for other modalities. For example, a sound is not only heard but also may trigger a specific taste or visual image. These cross-modal effects have recently been related to specific changes in brain activity.

Some investigators hold that synesthesia is the normal condition in neonates—that sensory impressions form a "primitive unity" at birth and that only with age and experience are they differentiated from one another. Others believe quite the opposite—that the senses are differentiated at birth and one must learn to associate among them. Studies of SC neurons in animals seem more consistent with the latter view. In newborn cat there are no multisensory neurons in the SC. Initially, the only sensory-responsive neurons are those that respond to somatosensory cues, presumably to help the neonate in finding the nipple. Auditory responsiveness develops in SC neurons at approximately Postnatal Day 5, and visual responses in the multisensory regions of the structure do not appear until about 3 weeks. Multisensory neurons appear in the second week of life (the first are auditory–somatosensory) and gradually increase in number. Although these neonatal multisensory neurons can respond to cues from different sensory modalities, they cannot integrate them to produce the response enhancement or response depression that is characteristic of the mature SC. In this regard, neonatal multisensory neurons appear very much like the nonintegrative multisensory neurons in adult animals described earlier. It is only after many weeks

that adult-like multisensory integration can take place. The initiation of this event in any individual neuron is linked to the maturation of functional inputs to it from association cortex. Thus, early neonatal multisensory neurons, mature nonintegrative neurons, and mature integrative neurons that have been deprived of their inputs from association cortex all respond in the same way to cross-modal cues.

At birth, the rhesus monkey is far more mature than is the cat. It already has a complement of multisensory SC neurons, although one that is not equal to that found in the adult monkey. Nevertheless, like multisensory neurons in the neonatal cat, these neurons are incapable of synthesizing cross-modal cues to produce response enhancement. Presumably, these neurons also await the development of cortical influences, but this remains to be determined. Nevertheless, it is safe to say that, based on experiments in cats and monkeys, the newborn SC is incapable of multisensory integration. Whether this is also true of neurons in other polysensory areas of the brain is not known. Most important for perceptual theorists would be testing whether multisensory integration is possible in polysensory cortex, and one excellent structure in which to examine this question is association cortex, specifically AES. As noted earlier, AES contains multisensory neurons that form a circuit independent of the SC, and baseline studies of the multisensory properties of adult AES neurons have already been performed.

VI. COMMONALITIES AMONG STRUCTURES: SUPERIOR COLLICULUS VERSUS MULTISENSORY CORTEX

Using the information obtained from the SC as a guide, experiments were conducted with multisensory neurons in AES to determine if they are governed by the same principles of multisensory integration. This cortex consists of three largely modality-specific subregions (visual, auditory, and somatosensory), with multisensory neurons clustered most heavily at the borders between these subregions. AES multisensory neurons proved to have overlapping receptive fields, much like their counterparts in the SC. This was particularly surprising given that no overall topographic organization appears to be present in the visual and auditory subregions of AES. Only the somatosensory representation is topographic. This suggested that cross-modal receptive field register would also be a key feature of multisensory integration in these neurons.

Indeed, just as in SC neurons, the spatial correspondence of cross-modal cues was a critical determinant of the nature of an integrated multisensory response: spatially coincident stimuli produced multisensory response enhancement in AES neurons, whereas combinations of spatially disparate cross-modal cues either failed to produce an interaction or resulted in response depression (albeit overall the response depression appeared to be less severe than in SC neurons). Furthermore, the greatest proportionate response enhancements were produced by combining weakly effective modality-specific cues, just as in the SC, and the temporal window within which multisensory integration was possible was similar to that found in the SC. Unfortunately, it is not known if it is possible to selectively depress multisensory integration in AES neurons by deactivating a specific input without eliminating their modality-specific responses.

Similar principles of multisensory integration have also been found in polysensory cortical regions in rodents, and the overlapping nature of cross-modal receptive fields has proved to be present in nearly every polysensory area of primate cortex, suggesting that here, too, the spatial principles of multisensory integration will be operative. Such a constancy in the principles of multisensory integration would be a parsimonious way of uniformly increasing (or decreasing) the salience of the same stimulus complex throughout the brain and would also be an effective way of appropriately matching the intensity of immediate behaviors with the higher order perceptual, cognitive, and emotive processes that provide conscious dimension to sensory experience.

VII. STUDIES OF MULTISENSORY INTEGRATION IN HUMAN CORTEX

In the past few years there have been a number of studies exploring the neural bases of multisensory integration in human cortex. Initially, such studies utilized event-related potential (ERP) recording techniques in which the averaged responses of thousands of neurons are recorded from surface electrodes on the scalp. The temporal resolution of ERP recordings is excellent (events can be measured in milliseconds) and can easily be combined with conventional behavioral or perceptual measures in order to examine the covariance between changes in behavior/perception and neural activity. One of the surprising results of these studies is the finding that activity in primary sensory cortex (e.g., auditory) can be modulated by

"nonappropriate" sensory stimuli (e.g., visual), an observation that questions the accepted dogma that each primary sensory cortex is responsive only to a single sensory modality. It also confirms some early observations (that were largely ignored) from single-neuron studies in animals suggesting that some of the neurons in primary sensory cortex are not modality specific and can, in fact, be activated by stimuli from other sensory modalities.

ERP studies have also identified presumptive polysensory cortices in humans. The changes in human brain activity that are induced by multisensory stimuli have also been associated with an increased response accuracy and reaction speed to these stimuli. Unfortunately, the spatial resolution of the ERP technique is not sufficient to allow researchers to identify the specific sulci and gyri that are involved in these processes.

It is in specifying the locus of evoked activity in the human brain that functional imaging techniques have had their greatest impact, because they have centimeter or subcentimeter resolution. Two imaging techniques, both of which utilize changes in cerebral blood flow to assess neural activity, have become quite popular in studies of multisensory integration: positron emission tomography (PET) and functional magnetic resonance imaging (fMRI). These techniques have confirmed the ERP observation that the activity in primary visual and auditory cortices can be modulated by other sensory stimuli, and they have also identified specific regions of cortex that are devoted to processing multisensory stimuli, including right insulaclaustrum, left basal posterior temporal lobe, and the medial inferior parietal lobe. Recently, fMRI techniques have strongly suggested that multisensory integration in human cortex is subject to some of the same rules of multisensory integration that have been demonstrated at the single-neuron level in the midbrain and cortex of animals. Thus, superadditive response enhancement (Fig. 7) was obtained in the left superior temporal sulcus when the visual signal (lip movements) matched the auditory signal (heard speech), whereas responses became subadditive when the lips and the heard signal were mismatched.

Surprisingly, these imaging studies have demonstrated that primary sensory cortex is subject to a remarkable degree of developmental plasticity. Blind or deaf individuals have their denervated cortices "taken over" by other sensory modalities. Thus, for example, disrupting the activity of visual cortex in a blind subject via transcranial magnetic stimulation interferes with the subject's ability to read Braille. Apparently, the absence of visual input allowed visual cortex to capture some of the normal functions of somatosensory cortex, thereby expanding the brain tissue devoted to processing information from the body. Similar changes in the brain are likely to accompany damage to each of the senses. Nevertheless, whether blind or deaf people are actually better at using the information they process in their remaining senses than are normal individuals remains a controversial issue.

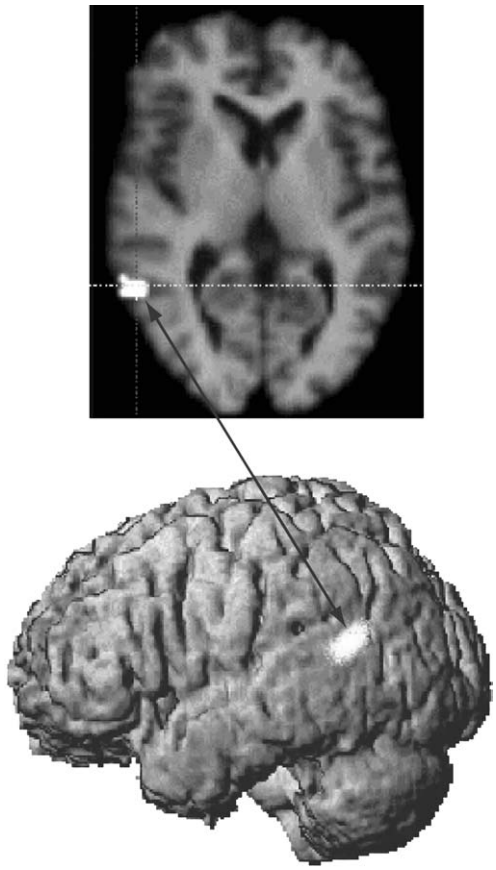


Figure 7 (Top) A magnetic resonance image (MRI) of multisensory enhancement in the human brain. An illustration of a horizontal section through the brain is shown. The white area (double–headed arrow) is a region of activation as measured by functional MRI (i.e., increased blood flow). The region is located in the ventral bank of the left superior temporal sulcus and is known to be involved in speech perception. Its response to a cross-modal (visual–auditory) stimulus combination exceeded that predicted by the sum of the individual modality–specific responses. The area is also illustrated on an image of the whole brain. (reprinted from *Curr. Biol.* **10,** Calvert *et al.*, 649–657, copyright 2000, with permission from Elsevier Science).

VIII. CONCLUSIONS

The observations detailed previously suggest that at least some of the fundamental principles of multisensory integration are generalized across species and brain structures. Undoubtedly, specialized multisensory properties will be found that are best adapted to the specialized needs of different species and different brain regions, and this is one of the challenges of future research. Nevertheless, it is likely that such specialized properties will prove to be overlaid on a fundamental core of principles that have common evolutionary origin and widespread utility. Despite a remarkable degree of diversity among species and their ecological circumstances, the presence of striking commonalities in individual sensory systems argues for similar commonalities in the principles that interlink them. From the perspective of the individual, a constancy in the fundamental principles of multisensory integration at different levels of the neuraxis ensures that the salience of a given event is approximately equivalent in the many different areas of the brain that contribute to the same behavior.

See Also the Following Articles

AUDITORY PERCEPTION • HAND MOVEMENTS • MIDBRAIN • MOTOR SKILL • NEURAL NETWORKS • RECEPTIVE FIELD • SPEECH • SUPERIOR COLLICULUS • VISION: BRAIN MECHANISMS

Acknowledgments

The authors' research is supported by NIH Grants NS22543 and NS36926, and editorial assistance was provided by Nancy London.

Suggested Reading

Baron-Cohen, S., and Harrison, J. E. (1997). *Synaesthesia*. Blackwell, Oxford.

Cytowic, R. E. (1989). *Synesthesia: A Union of the Senses*. Springer-Verlag, New York.

Edwards, S. B., Ginsburgh, C. L., Henkel, C. K., and Stein, B. E. (1979). Sources of subcortical projections to the superior colliculus in the cat. *J. Comp. Neurol.* **184,** 309–330.

Finger, S. (1994). *Origins of Neuroscience*. Oxford Univ. Press, Oxford.

Gibson, J. J. (1966). *The Senses Considered as Perceptual Systems*. Houghton-Mifflin, Boston.

Huerta, M. F., and Harting, J. K. (1984). The mammalian superior colliculus: Studies of its morphology and connections. In *Comparative Neurology of the Optic Tectum* (H. Vanegas, Ed.), pp. 687–773. Plenum, New York.

Kujala, R., Alho, K., and Naatanen, R. (2000). Cross-modal reorganization of human cortical functions. *Trends Neurosci.* **23,** 115–120.

Lewkowicz, D. J., and Lickliter, R. (1994). *The Development of Intersensory Perception*. Erlbaum, Hillsdale, NJ.

Massaro, D. W., and Stork, D. G. (1998). Sensory integration and speechreading by humans and machines. *Am. Sci.* **86,** 236–244.

Maurer, D., and Maurer, C. (1988). *The World of the Newborn*. Basic Books, New York.

Sparks, D. L., and Nelson, J. S. (1987). Sensory and motor maps in the mammalian superior colliculus. *Trends Neurosci.* **10,** 312–317.

Stein, B. E., and Meredith, M. A. (1991). Functional organization of the superior colliculus. In *The Neural Bases of Visual Function* (A. G. Leventhal, Ed.), pp. 85–110. Macmillan, Hampshire, UK.

Stein, B. E., and Meredith, M. A. (1993). *The Merging of the Senses*. MIT Press, Cambridge, MA.

Stein, B. E., Meredith, M. A., Huneycutt, W. S., and McDade, L. (1989). Behavioral indices of multisensory integration: Orientation to visual cues is affected by auditory stimuli. *J. Cognition Neurosci.* **1,** 12–24.

Wallace, M. T., Meredith, M. A., and Stein, B. E. (1998). Multisensory integration in the superior colliculus of the alert cat. *J. Neurophysiol.* **80,** 1006–1010.

Music and the Brain

DAVID W. PERRY
University of California, San Francisco

GLOSSARY

absolute pitch A rare ability possessed by some individuals who can name the pitch of a musical tone, or sing a named pitch on demand, without referring to any other sounds.

amusias Acquired disorders of music perception, performance, reading, or writing that are due to brain damage and are not attributable simply to the disruption of basic perceptual, motoric, or cognitive functions.

consonance A subjective quality of the pleasantness of combinations of tones.

contour The pattern of changes in frequency direction within a tone sequence, regardless of the magnitude of those changes.

harmony The subjective effect produced by the simultaneous sounding of musical tones or tone sequences.

melody A sequence of tonal pitches, usually varying in duration, organized to express a musical idea.

meter A repeated pattern of stress against which other temporal variations are ordered.

pitch A subjective quality of the relative highness or lowness of sounds.

rhythm The musical organization of the temporal durations of sounds.

scale The basic set of pitches used in a particular musical composition (identical in every octave), arranged in ascending or descending order.

song A short musical composition integrating words and music.

timbre A subjective quality imparted to harmonic complex sounds by the relative distribution and time course of energy at the different harmonic partials but distinct from their pitch.

tonal melody The pattern of intertone intervals formed by a tone sequence, including their specific size and order, that may be transposed in frequency.

tonal working memory The ability to hold sequences of tones in mind, often in order for more complex processing to be carried out.

tone A sound of distinct pitch and duration, such as a musical note.

The study of how the human brain produces musical behaviors, both receptive and expressive, is at least as complex as the similar study of language. As in the study of language, the neural substrates of these highly evolved and complex human activities, such as listening to (or performing in) an orchestra with chorus, are best approached by breaking them down into component processes, each of which depends on distributed networks of neural processors.

I. INTRODUCTION

The following fundamental questions occupy most of the energies of those engaged in attempting to unravel the brain's role in musical perception and production:

1. How does the brain *receive* musical sound?
2. How does the brain *process* and organize musical sounds and extract the "information" they convey?
3. How does the brain *store* and *retrieve* musical sounds and information and the motor programs required to reproduce them?
4. How does the brain control the *production* of musical sounds (whether audibly by singing, instru-

ment playing, or whistling or inaudibly, for the "mind's ear" alone, through mental imagery)?

5. How does the brain control the *writing* and *reading* of graphic notations used to record and transmit musical ideas or "road maps" for musical production?

6. How does the brain mediate *emotional* responses to music and emotional expression through music?

7. How does the brain mediate the integrated perception and production of *music and language* as in song?

8. How is the brain *changed* by musical experiences, not just in the sense of "engrams" or "circuits" underlying memory for specific musical ideas or actions, but in more generalized ways?

Most of what is known about how the human brain perceives and produces music comes from the study of two very different sources of information: (i) neurophysiological activity measured noninvasively during the exercise of musical perceptual, motoric, and/or cognitive functions and (ii) measurement of the behavioral effects of focal brain lesions on these same musical functions. The study of how the brain produces musical behaviors is in fact a microcosm of the field of cognitive neuroscience in general, with different aspects of music touching on most of the fundamental questions of that domain. This article can be viewed as a framework for studying the neural organization of the range of normal musical functions as well as for studying the range of musical disorders that may follow the disruption of normal neural functioning.

II. MUSICAL PERCEPTION

Musical perception may be considered in terms of the reception of those perceptual–acoustic characteristics that define the basic elements of musical sounds, such as frequency, pitch, consonance, timbre, duration, intensity, and spatial location. This level of inquiry has received by far the most attention by researchers and is amenable to investigation in both human and animal brains.

A. Perceptual Acoustic

Pitch can be considered the most essentially musical of these acoustic characteristics, being a central organizing principle for virtually any imaginable musical culture.

1. Pitch

Pitch is a subjective quality, the exact physical correlates of which have been controversial since the time of Heinrich von Helmholtz in the late 19th century. Pitch is traditionally understood to be derived by the auditory nervous system through analysis of the spectral frequencies in sounds. In fact, the physical attribute of frequency (for pure tones) can be mapped point for point from the inner ear, in the topography of the basilar membrane (apex, high frequency; base, low frequency), through portions of the brain stem nuclei, culminating in the ventral division of the medial geniculate nucleus and projecting to tonotopically organized cortical fields, the most well studied of which is the primary auditory cortex.

Pitch can be dissociated from frequency by studying responses to stimuli with virtual pitch, or pitch experienced at a "missing" fundamental frequency that is derivable from the pattern of upper harmonics, as the root of a corresponding harmonic series. Several pivotal experiments have taken advantage of such stimuli. In 1976, Henry Heffner and I. C. Whitfield demonstrated that cats could be trained to discriminate periodicity pitch as well as frequency. In 1980, Whitfield showed that following bilateral lesions of primary auditory cortex, the cats lost the ability to discriminate virtual pitch but could still discriminate absolute frequency. This suggests that auditory cortex may be essential for the discrimination of pitch but not frequency.

In 1988, Robert Zatorre tested patients whose temporal lobe surgeries for the control of seizures either included or excluded Heschl's gyrus, the anatomical landmark for human primary auditory cortex, based on surgeons' reports. Patients with right temporal lobe removals including Heschl's were the only group impaired in discriminating virtual pitch.

Christo Pantev and colleagues used a noninvasive imaging technique, magnetoencephalography (MEG), to examine the magnetic fields evoked by tones with virtual pitch and pure tones and found that the source of the primary component of the evoked magnetic field (the M100) was identical for a tone composed of the fourth through the seventh harmonics and for a pure tone at the frequency of the missing fundamental. However, corroborating evidence is scant, and controversy exists regarding the interpretation of MEG demonstrations of the topographic mapping of pitch, and even of frequency.

Recalling some of the patients mentioned previously, Ingrid Johnsrude, Virginia Penhune, and Robert

Zatorre determined from high-resolution magnetic resonance imaging (MRI) that only excisions extending into right Heschl's gyrus disrupted thresholds for the judgment of the direction of a change in frequency. No excisions impaired the ability to discriminate the presence or absence of a frequency change.

Taken together, these results suggest that certain higher level aspects of pitch perception, such as the perception of virtual pitch and the *direction* of frequency change, may depend on the integrity of primary auditory cortex and, in man, particularly on that in the right hemisphere.

As mentioned earlier, the exact nature of the physical parameters corresponding to the subjective experience of pitch has been controversial for more than 100 years because the neural excitation patterns produced by the cochlea contain not only spectral peaks, from which the pitch of harmonic sounds may be derived, but also repetition rates that correspond to the pitch of the sound. For virtual pitches formed by upper harmonics (i.e., above the eighth in the series), spectral peaks are no longer resolvable and yet pitch is still experienced. Sounds ("iterated rippled noises") can be synthesized that are spectrally identical to one another by passing a random noise through a cascade of delay-and-add networks and yet produce different subjective pitches, with a pitch corresponding to the reciprocal of the time delay.

Using positron emission tomography (PET), Timothy Griffiths and colleagues presented rapid sequences of tones differing in pitch that varied progressively in the number of iterations from 0 to 16 (and hence in the strength of the pitch percept). They sought brain areas whose activation covaried with the number of iterations. The only such areas found were near the primary auditory cortex bilaterally, probably just lateral and/or inferior to it, and were approximately symmetrical. These results indicate that the temporal integration allowing this time-based code to be transformed into a spatial one occurs at or before the primary auditory region. Animal studies by Gerald Langner and Christoph Schreiner indicate that this integration occurs in the cat by the level of the inferior colliculus and results in a mapping of pitch that is orthogonal to that for frequency. Results suggestive of a similar arrangement within human auditory cortex have been obtained by Langner, Mikko Sams, and colleagues using MEG but remain to be fully confirmed. Furthermore, cortical neurons do not exhibit sufficient temporal sensitivity to encode pitch in the temporal patterns of their firing rates. Nevertheless, Griffith's PET study suggests the possibility that at the cortical level the temporal analysis of pitch within the auditory nervous system may be more bilaterally distributed than the spectrally based analysis. However, this study differs in other ways from most studies demonstrating right auditory cortical superiority, e.g., in the rapidity of the tone sequences (96 events within 1 sec).

Although interpretation of far-field neurophysiological measures such as MEG is controversial, near-field measurements have provided more direct evidence for a tonotopic gradient within human auditory cortex. Catherine Liegeouis-Chauvel and colleagues, using intracranially measured auditory evoked potentials (from depth electrodes placed for seizure localization), demonstrated a medial (high frequency) to lateral (low frequency) gradient within Heschl's gyrus, more sharply tuned in frequency within the right hemisphere. In addition, they measured evoked potentials (EPs) to voiced (/ba/) and unvoiced (/pa/) speech syllables and found that only EPs from sites in left Heschl's gyrus were differentially sensitive and showed evidence of encoding the timing of voice onset.

Zatorre proposed that greater spectral resolution within the right primary auditory cortex is fundamental to the development of complementary hemispheric specialization. Specifically, the inherent trade-off between spectral and temporal resolution resulted in the development of complementary specializations for their processing and was probably driven first by the demands on temporal processing imposed by perceiving speech. To test this hypothesis, he and Pascal Belin examined the correlation of cerebral blood flow (CBF), as measured with PET, with two parametrically varying characteristics of pure tone sequences: increasing frequency discrimination (with rate held constant) and increasing temporal rate (with frequency difference a constant). An area in the anterior temporal lobe showed increasing CBF with rate in the left hemisphere and with frequency discrimination in the right.

Regardless of whether right or left auditory cortex specialization came first in human evolution (and thus possibly music and language in human history), there is ample evidence supporting a preferential role for the right auditory cortex in the processing of spectral pitch.

2. Consonance

Consonance is a subjective quality of the pleasantness of combinations of tones. Its physical counterpart is found in the phenomenon of roughness, or changes in the amplitude of components of a complex signal that

can result in the sensation of beats. When the discharges of single fibers in the auditory nerve, which projects from the cochleae to the first brain stem relay (the cochlear nucleus), are measured, these beating patterns can be discerned. Mark Tramo, Peter Cariani, and Bertrand Delgutte measured such responses in the cat to pairs of pure tones, the ratios of whose frequency components constituted a subset of the musical intervals central to many musical systems: the major fifth (1:2), major fourth (2:3), minor second (16:l5), and tritone (32:45). They also examined the responses to pairs of complex tones (composed of six harmonics) whose fundamental frequencies formed the same musically relevant ratios. For pure tone pairs, only the most discordant minor second produced beating patterns. For complex tone pairs, beating patterns resulted from the most closely spaced component frequencies, and there was a high positive correlation between the total number of auditory nerve fibers showing beating patterns and the perceived dissonance of the musical interval, maximal for the minor second and intermediate for the tritone.

Tramo recorded the responses of single cells in auditory cortex to these same stimuli and found evidence for a dual time code. Local periodicities in the submillisecond range appear to encode the harmonic relationship among partials, whereas local periodicities in the range of tens of milliseconds encode the total roughness among partials.

These results suggest that the perceptual attribute of dissonance results from the integration of temporal information across populations of auditory nerve fibers. It is not clear how or where this integration takes place. Furthermore, as Tramo pointed out, measures of roughness alone do not distinguish between the perfect fourth and fifth, suggesting that the perception of harmony may derive more from higher leve1 analysis of pitch relationships than from physiological roughness.

3. Timbre

Timbre refers to another subjective quality imparted to harmonic complex sounds by the relative distribution of energy at the different harmonic partials. In a dynamic sense, the time course of this spectral distribution contributes to our ability to distinguish different musical instruments (e.g., the differences in attack between a trumpet and an oboe). In 1962, Brenda Milner demonstrated deficits on the timbre discrimination subtest of the Seashore Tests of Musical Talent after right but not left temporal lobectomy. Severine Samson and Robert Zatorre confirmed and extended this observation by finding deficits in both spectral and temporal aspects of timbre recognition.

Recognition of musical instruments is also often severely disturbed in cases of generalized auditory agnosia, or a loss of the ability to recognize auditory "objects" following bilateral auditory cortical damage and, in some cases, unilateral right hemisphere damage. However, even in cases in which such loss of recognition is restricted to musical sounds, such as those studied extensively by Isabelle Peretz and colleagues, musical instrument recognition may be selectively spared (e.g., patients CN and GL both display profound deficits in the recognition of previously familiar tunes, but musical instrument recognition was disturbed for CN but not for GL). At least for recognition, bilateral cortical contributions are suggested, but with a right temporal lobe predominance. More investigation is needed before further conclusions can be drawn regarding the critical stages in the auditory pathway for processing musical timbre.

4. Intensity

The detection of changes in sound intensity is of great musical relevance, forming the basis of musical "dynamics" or expressive changes in sound intensity, both abrupt and gradual. Cortical contributions to the perception of intensity are suggested by studies of the response properties of neurons in animals. In particular, some cells in the primary auditory region of cats and monkeys show sharp tuning of response to particular intensities, especially, as demonstrated by Gregg Recanzone and colleagues in awake behaving macaque monkeys, within the tonotopically organized field just anterior to the primary area (field R).

Pascal Belin and associates measured CBF changes using PET while human subjects detected progressively finer intensity differences between tones. Similar increases were observed within the right posterior temporal gyrus regardless of the subjects' intensity discrimination performance. These results are partially corroborated by Brenda Milner's finding in 1962 that right but not left anterior temporal lobectomy resulted in a decrement on the intensity subtest of the Seashore Tests of Musical Talent. Increases were also observed that varied with individual performance level in the right inferior frontal gyrus and, linearly, in right parietal cortex. This frontal–parietal network is very similar to that observed by Robert Zatorre and Todd Mondor during auditory attention to either frequency or spatial location, and it may be involved in the

allocation of attention to various acoustic attributes and even to particular sensory modalities. It thus appears that simply discriminating between the intensity of two tones involves not only the encoding of basic acoustic parameters within auditory cortex, but more widespread cortical networks.

It remains to be investigated whether higher level judgments of intensity change, such as the direction of change (i.e., louder vs softer), show dependence on cortical levels of auditory processing, as do analogous judgments of frequency direction. The processing of more complex intensity changes (e.g., decrescendo and crescendo) extended in time also awaits investigation.

5. Duration

Studies of the behavioral performance of patients with temporal lobe lesions have shown deficits in thresholds for temporal discrimination following damage to the left hemisphere. A recent study of patients with medial temporal lobe epilepsy by Severine Samson and associates examined thresholds for the detection of deviations within otherwise isochronous tone sequences. For detection of interonset intervals (IOI) of 80 msec, left temporal lobe patients showed increased thresholds when compared to right temporal patients and normal controls. At IOIs of 300 msec and higher, thresholds for anisochrony detection were normal. Detection of temporal perturbations within familiar melodies (i.e., lengthening one interval by 25%) was measured following right vs left temporal lobectomy. Only left temporal lobectomy resulted in a decrement in the detection of temporal deviations within a melodic context.

Since mesial temporal lobe atrophy is accompanied by decreased metabolism in the lateral temporal lobe, these results support the hypothesis that the left temporal lobe is critical for the fine-grained perception of temporal variations.

6. Sound Localization

Whereas *frequency* is a physical attribute that is known to be "mapped" directly in the spatial organization of cortical neurons throughout the main ascending auditory pathway, the *spatial locations* of sounds are computed primarily from differences between the inputs to the two ears (but also from monaural spectral cues resulting from the shape of the pinnae). Recent evidence from animal neurophysiology favors an interpretation of the cortical encoding of acoustic space within the population structure of neuronal responses.

Although not the most musically relevant of the basic perceptual–acoustic attributes of musical sounds, auditory spatial localization can aid in the segregation of particular instruments or musical groupings from competing sounds. Auditory scene analysis, or the extraction of individual auditory "objects" from a complex mixture of sounds, is a higher level function dependent on the integration of spectral, temporal, and/or spatial attributes of auditory stimuli and is discussed in the following section. Human and animal studies indicate the involvement of both primary and secondary auditory cortices and portions of the posterior parietal lobe in simple localization of sounds in space.

B. Perceptual–Organizational

Musical perception can also be studied at the level of the perceptual organization of auditory events along these basic perceptual–acoustic axes, both when the events occur simultaneously or overlap in time, and when they are extended across time sequentially. Although the previous section touched on this level, the following discussion focuses on the perceptual organization of multiple auditory events.

1. Pitches

As discussed in the previous section, there is considerable evidence supporting a critical role for the right auditory cortex in the processing of pitch, as distinguished from frequency, particularly when judgments must be made. In this section, we consider what is known about the perceptual processing of multiple pitched auditory events, occurring either simultaneously (as in musical chords of two or more notes) or sequentially (as in tone sequences or melodies).

a. Harmonic (Simultaneous) Few studies have systematically examined the localization of component processes involved in the perception of harmony or the subjective effect of particular combinations of musical pitches or pitch sequences. Mark Tramo, Jamshed Bharucha, and Frank Musiek asked a patient with bilateral temporal lobe strokes to make judgments of mistuning within a three-note chord that was preceded by a different three-note chord. Normal subjects showed facilitation of mistuning judgments when the preceding chord was harmonically related. Even though the patient was impaired in tuning judgments, with a bias toward dissonance, his judgments were still normally "primed" by the harmonic

relatedness of the preceding chord. This result suggests that simple judgments of consonance, which may be dependent on primary auditory cortex, can be dissociated from higher level effects of harmonic relatedness and expected sequence, which may be mediated by secondary auditory cortical areas.

b. Tonal (Sequential) The processing of sequences of auditory events possessing pitch is obviously central to music perception. In its simplest form, each of these events consists of a harmonic complex tone with a relatively constant pitch. The neural basis for processing sequences of such tones has been explored using several different methods. Sequences of tones differing in pitch can be encoded relatively crudely in terms of the pattern of changes in frequency direction, regardless of the magnitude of change. This type of pattern is called contour. Alternatively, the specific pitches (or the *size* of the intervals they form) can be encoded, and the resulting melodic contour or tonal melody may be transposed in frequency.

Only a few neurophysiological studies have examined single cell recordings in response to the individual tones contained in such stimuli. Norman Weinberger and T. McKenna studied the responses of neurons to five-tone sequences in the cat's auditory cortex. Strong responses were obtained only to the first tone of each sequence under anesthesia, so measurements were carried out while awake, resulting in strong responses to each tone. Discharges to the same tone varied depending on the contour of the particular sequence in which they were embedded, within both primary and secondary auditory cortical areas.

In a related study using three-tone sequences, I. Espinosa and George Gerstein recorded from primary auditory cortex in the cat, testing all permutations of the three tones. In a groundbreaking experiment, they recorded up to 24 neurons simultaneously, permitting analyses of the functional connectivity between pairs of neurons. These patterns of connectivity, elicited in response to the same tone as it occurred in different sequences, were varied, suggesting that contour and/or tonal melody are encoded in the responses of *populations* of neurons.

Techniques for recording from many neurons or neuronal clusters simultaneously have undergone rapid advances recently, thus permitting further investigation of these phenomena. Only then will we be able to proceed beyond these initial suggestive inquiries toward a more complete understanding of how sequences of pitches are encoded within the tonotopic auditory nervous system. Also, only then will we be able to judge the prescience of William James, who speculated in 1890 that memories might be retained as "'paths' … in the finest recesses of the brain's tissue," recalled by "the functional excitement of the tracts and paths in question."

Using PET, Zatorre and colleagues measured increases in CBF in response to standard Western diatonic melodies, in contrast to amplitude envelope-matched noise bursts. In a recent (2000) reanalysis of the original 1994 report, the localization of the focus within the right superior temporal gyrus was further refined to its ventrolateral aspect, anterior to Heschl's gyrus, and the smaller increase in the left superior temporal gyrus reached significance (Fig. 1). Although

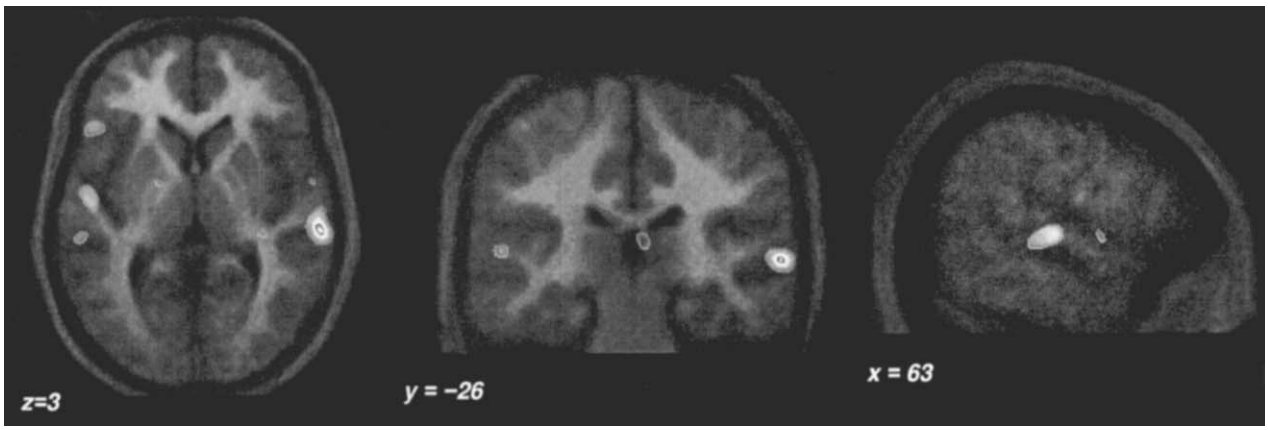


Figure 1 Melody perception vs matched noise bursts. Horizontal, coronal, and saggital views through the peak of CBF increases in the superior temporal gyrus, as measured with PET, superimposed on the group-averaged MRI volume. The peak is located on the lateral surface of the right superior temporal gyrus, posterior to Heschl's gyrus (original data from Zatorre *et al.*, 1994; reanalysis and figure reproduced with permission from Zatorre and Binder, 2000).

the contrast between melodies and noise sequences is complex, these results suggest the involvement of secondary auditory cortical areas in the processing of melodies, with a right hemisphere predominance that is nevertheless relative.

A recent study by Anniruddh Patel and Evan Balaban measured the magnetoencephalogram while subjects listened to tone sequences varying parametrically in structure from random to music-like. Patterns of temporal synchronization between brain regions were greatest for the most structured, melody-like sequences and suggested a high degree of inter-hemispheric connectivity.

Although a number of studies of the effects of brain damage on the processing of tonal melody indicate a relative right hemisphere predominance, many also indicate a critical involvement of processing in the left hemisphere as well, particularly for more complex sequences and tasks. Ensuring that the melodies do not have associated lyrics, that they are novel, and that the task does not involve naming is not sufficient to eliminate potentially critical left hemisphere contributions to tone sequence perception.

For example, Peretz found that groups of patients with right or left hemisphere lesions were both impaired on a task of making same–different judgments to pairs of novel melodies, relative to normal control subjects. The right hemisphere group was more impaired for trials in which the contour was violated. However, both groups were equivalently impaired when contour was preserved, and comparison was thus based solely on pitch interval information. Peretz hypothesized that the right hemisphere is predominant for the encoding of contour and the left hemisphere for the encoding of specific pitch intervals. She further proposed that the encoding of specific pitch intervals is dependent on the formation of a representation of contour, resulting in the observed pattern of deficits. However, studies by Zatorre and colleagues found normal effects of these cues on melody recognition regardless of which hemisphere was lesioned and overall found greater and more consistent effects of right hemisphere excisions, particularly those that encroach on the primary auditory region.

Differences in tasks and in the types and extents of lesions are no doubt critical, but clearly cortical areas in the left hemisphere also contribute to melody perception.

2. Duration

The musical organization of durations is called *rhythm*, and it often reflects an underlying *meter* or repeated pattern of stress against which other temporal variations are ordered. These aspects of music perception have received less study than tonal pitch organization, and there are fewer hypotheses regarding their neural substrates.

a. Metric Peretz compared rhythm and meter discrimination by groups of right and left hemisphere-damaged patients. Although rhythm discrimination was impaired for both groups, neither group was impaired in meter discrimination. This evidence that meter discrimination is more resistant to cortical damage may implicate subcortical and/or spared cortical processing.

b. Rhythmic A recent study using PET carried out by Virginia Penhune, Robert Zatorre, and colleagues found right-sided activity in the superior temporal gyrus when the perception and manual reproduction of isochronous sequences were contrasted to perception alone. These results suggest the involvement of secondary auditory cortical areas in the right hemisphere for the simplest form of rhythmic perception and/or reproduction. When reproduction of more complex rhythmic sequences was contrasted to reproduction of a similar repeated, well-learned sequence, increases were seen in the cerebellar vermis and hemispheres for both auditory and visual stimulus presentation. This cerebellar activation may result from its contribution to the flexible production of precisely timed sequential movements and/or to the perception of complex, changing rhythms, but in either case it is not unique to auditory rhythms. The perception of visual rhythm may itself be considered an important musical function, at least for members of conductor-led ensembles.

3. Intensity

Although the intensity patterns underlying musical phrase structure have been studied extensively by cognitive scientists such as Henry Shaffer, Eric Clarke, and Caroline Palmer, the neural basis of their perceptual processing remains to be explored.

4. Integration of Tonal, Rhythmic, and Intensity Organizations

In actual musical performance, and thus in its product, musical sound, the tonal structure, rhythmic structure, and intensity pattern are usually highly interrelated. Their integrated product, the musical phrase at one

level of local detail, is perhaps one of the most essential and defining elements of music. However, a neural understanding of the mechanisms supporting this integration must follow further study of the component processes outlined previously.

Finally, more complex levels of perceptual organization can be considered, such as the perception of an auditory "scene" and the integration of auditory and visual perception across time.

5. Auditory Scene Analysis

Spatial location is only one potential attribute of musical sounds that may be used dynamically, along with temporal and spatial cues, in order to extract auditory objects or events from complex sound mixtures. The physical separation of players in an orchestra can facilitate their extraction from the sound of the orchestra as a whole, finding an extreme in the off-stage ensembles sometimes employed in complex orchestrations. However, temporal and spectral cues also play prominent roles in the segregation of auditory objects. Sufficient differences on at least two of these three axes are needed in order to segregate an auditory object from its surround. In an experiment examining the combination of spectral and spatial differences, David Perry and Pierre Divenyi found lower thresholds for spatial separation within the left auditory hemifield in normal controls and disturbances following cortical damage in the contralateral temporoparietal junction. The normal asymmetry is suggestive of a right hemisphere preference for some aspect of auditory scene analysis based on temporal/spectral integration and could be explained simply by the demand for spectral processing.

A recent study of professional conductors found evidence from EPs of enhanced auditory localization when compared to that of nonmusicians or pianists. Although the ability to accurately localize the sources of sounds is critical for a symphony conductor, auditory spatial perception can also enhance the perception of music for casual listeners, whether real, as in live performance, or virtual, as in studio-produced recordings for two or more speakers. Specifically musical investigations must in large part follow further basic neuroscientific investigations of auditory scene analysis.

6. Auditory–Visual Integration

The integration of auditory and visual information is an active focus of basic neuroscience research that has relevance for understanding music perception. Although music can be enjoyed as a purely auditory phenomenon, without correlated visual input, live or videotaped performances of music present the listener with the correlated movements of players, singers, and conductors. Choreography set to music, live or videotaped dance, movie musicals, music videos, cartoons, and other music-video compositions present the listener/viewer with artificially assembled auditory–visual coincidences. Concomitant visual inputs can influence music perception, just as they can influence speech perception. Advances in basic research on low-level auditory–visual integration (e.g., a tone paired with a spatial grating) and on visual influences on speech perception will facilitate investigations of the neural mechanisms supporting auditory–visual integration in specifically musical contexts.

7. Absolute ("Perfect") Pitch

Some individuals, who are said to possess absolute or perfect pitch (AP), can name the pitch of a musical tone or sing a named pitch on demand without referring to any other sounds. Gottfried Schlaug measured the planum temporale or secondary auditory cortex on the superior temporal plane posterior to Heschl's gyrus and found that AP possessors exhibited an exaggerated leftward asymmetry.

Robert Zatorre, David Perry, and Christine Beckett used PET to measure CBF increases while musicians with and without AP performed a series of simple tasks involving pairs of complex tones. The musicians either simply listened to the tone pairs or decided whether they formed major or minor intervals. For listening in contrast to noise bursts, both groups showed bilateral activation of auditory cortex, but only the AP group showed an additional focus in the left posterior dorsolateral frontal cortex. However, when making interval judgments both groups showed activation in this region (Fig. 2). Studies in nonhuman primates by Michael Petrides have implicated this region as critical for conditional associative learning. AP can be conceptualized as a form of conditional associative learning, in which a stimulus attribute (tonal pitch) is arbitrarily associated with a verbal label (note name). Therefore, the activation of left posterior dorsolateral frontal cortex when simply listening to the associated stimulus (and while note names are obligatorily retrieved) may be explained by the retrieval of these associations. Its activation when musicians labeled intervals as "major" or "minor" may similarly be explained by the association between an arbitrary label

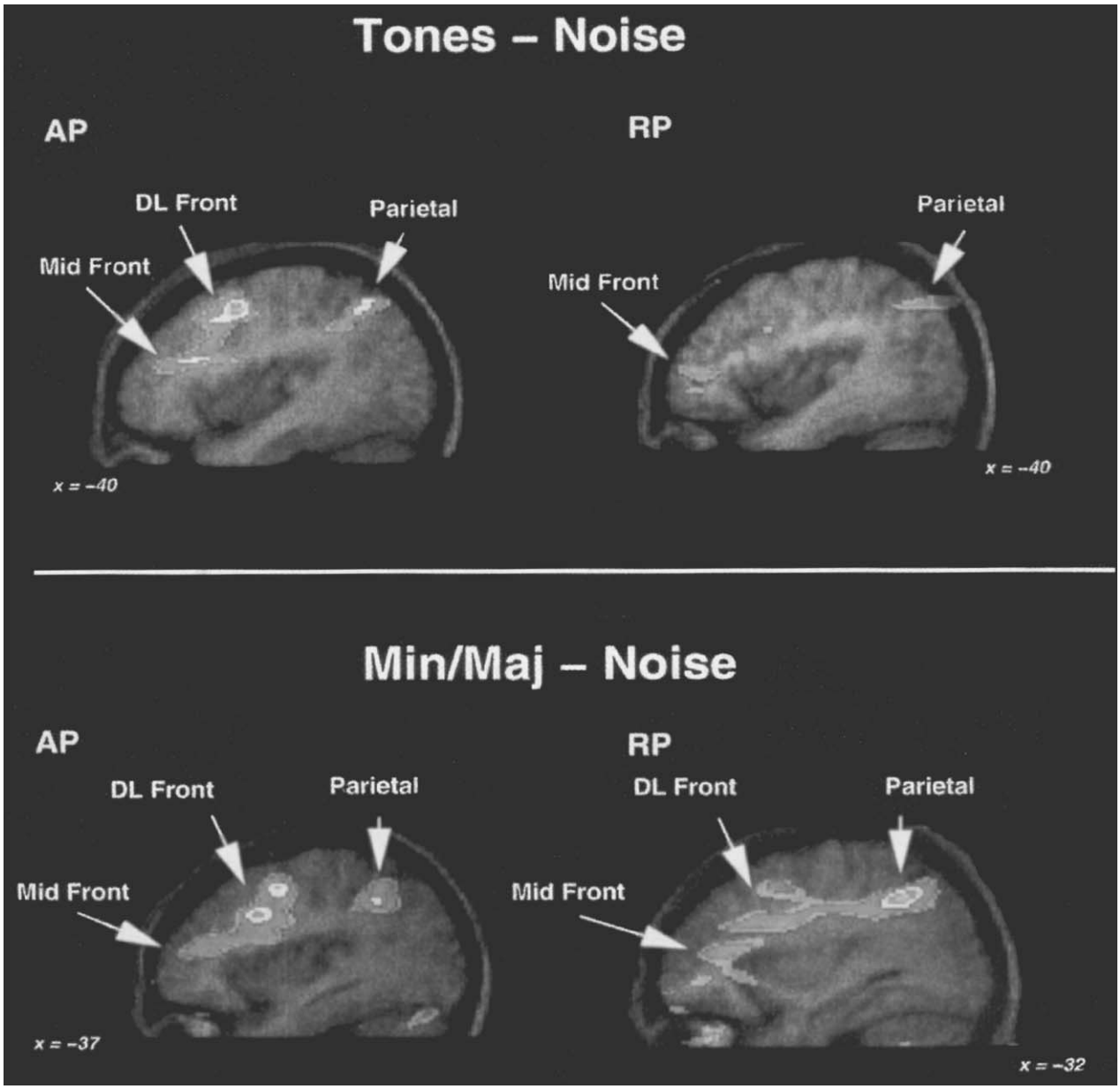


Figure 2 Musicians with absolute vs relative pitch ability. Tones-Noise: CBF increases while listening to pairs of sequential tones that formed either major or minor intervals vs amplitude envelope-matched noise bursts. An increase within the left posterior dorsolateral (DL) frontal cortex was observed for musicians with absolute pitch (AP) but not for those with only relative pitch perception (RP). Min/Maj-Noise: CBF increases while listening to the same tone pairs, but with the instruction to press a button if they formed a minor interval. Increases were seen within posterior DL frontal cortex in both groups, consistent with the hypothesis that these activations are related to retrieving conditional associations between particular auditory stimuli [pitches (AP) or the ratio of two pitches (AP or RP)] and verbal labels (note or interval names) (reproduced with permission from Zatorre *et al.*, 1998).

and a *relative* pitch stimulus attribute (the ratio of the fundamental frequencies of the tones making up a pair), an association learned by both groups as part of their conservatory training.

III. MUSICAL MEMORY

As in other cognitive domains, memory functions of several distinct types support musical processing and

behavior. Since music is inherently extended across time, music perception demands an ability to retain sounds temporarily in memory, either passively, as in short term or "echoic" auditory memory, or actively as the contents of auditory working memory. Other cognitive operations can then be applied to these contents. If sufficiently encoded, they may become part of long-term episodic memory, consciously recallable along with the original context ("They're playing our song"). Procedural memories for musical information, and for the motor movements to produce them, may also be formed even without conscious awareness ("What's that tune I'm whistling?").

A. Sensory Short-Term Retention

Zatorre and Samson presented a task to temporal lobectomy patients in which they compared the pitch of two tones across a silent interval. Neither right- nor left-sided excisions resulted in impairment on this task. Although more posterior excisions or longer retention intervals might result in deficits in short-term retention of pitch, these results indicate that unilateral lesions involving the primary auditory region did not impair brief sensory retention.

B. Auditory Working Memory/Imagery

In a second condition of the same experiment, Zatorre and Samson presented patients with the same task, but with the retention interval filled by interfering tones. The right temporal lobe patients were impaired in this version of the task, regardless of whether Heschl's gyrus was involved. A group of patients with right frontal lobectomies were similarly impaired. Although the laterality of the frontal effect could not be assessed, the right temporal lobe effect was clear-cut and indicated a critical role in auditory working memory for secondary auditory cortex in the right hemisphere.

David Perry carried out a study of ear-of-presentation asymmetry (a small behavioral effect presumed to result from predominance of the contralateral ascending auditory pathway) for the subsequent recall of melodies by experienced pianists who replayed them on a computer-interfaced keyboard. The interval between the end of the melody and the beginning of attempts at keyboard reproduction was recorded. During this interval, most subjects reported mentally resinging and rehearsing their memory of the melody before beginning to play. A right ear asymmetry was observed in sequence recall accuracy. Furthermore, an asymmetry was observed in this delay interval that was positively correlated with that in accuracy, such that the delay between stimulus and response (lasting several seconds on average) was longer after presentation to the preferred ear.

This asymmetry in an interval associated with internal rehearsal of just heard information by "inner singing" suggested the possibility of an asymmetry in the neural substrate for *auditory–tonal working memory*, conceived of by analogy to Alan Baddeley's "phonological loop." Just as inner speech can serve to refresh the contents of a specialized auditory–verbal store (as when remembering a just heard phone number), so too inner singing (based on vocal fundamental frequency control rather than articulatory control processes) can refresh the contents of a specialized tonal store through the functioning of the "tonal loop" (Fig. 3). It is possible that for instrumentalists "inner playing" might also be utilized as an internal rehearsal strategy, although whether mental

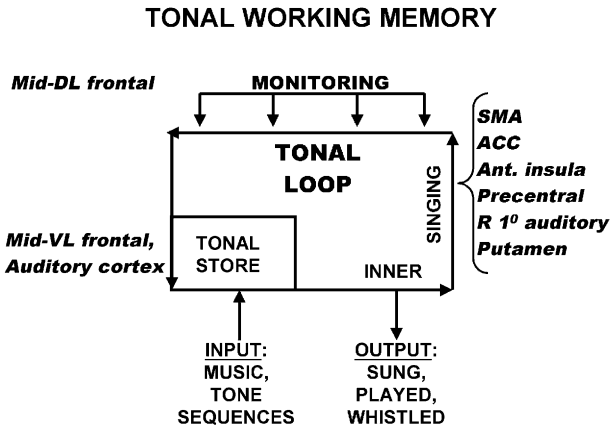


Figure 3 Tonal working memory. Tonal loop model of auditory–tonal working memory based on the phonological loop model of Alan Baddeley for auditory–verbal working memory. Music or nonmusical tonal information is first held in a time-limited tonal store, whose proposed neural substrate includes auditory cortex and mid-ventrolateral (VL) frontal cortex. The contents of this store can be refreshed and held on-line through inner rehearsal (e.g., "inner singing") based on vocal fundamental frequency control processes (rather than on articulatory control processes as proposed for inner speech by Baddeley). Inner or imagined singing is proposed to depend on most of the areas active during actual singing. The supplementary motor area (SMA) is particularly strongly associated with imagined singing. Finally, the contents of working memory may be consciously monitored, e.g., to update an ongoing record of events or actions, an executive working memory function for which the mid-dorsolateral frontal cortex is critical (adapted with permission from Marin and Perry, 1999).

motoric rehearsal and its associated auditory image can be fully dissociated from inner singing remains an empirical question.

Based on their studies of anatomical projections in the macaque brain between the frontal lobe and posterior sensory association cortex, Deepak Pandya and associates hypothesized that retention of *auditory* information might involve specific temporal–frontal projections, just as Patricia Goldman and colleagues had proposed for *visuospatial* retention and parietal–frontal projections.

Based on a long series of behavioral lesion analyses in nonhuman primates and further analyses of posterior association cortex–frontal projections, one of Pandya's associates, Michael Petrides, has articulated a hierarchical theory of frontal contributions to mnemonic processing. Petrides proposes that each major sensory modality represented in the posterior association cortices (auditory, visual–spatia1, visual–object, and somatosensory) projects, with some degree of topographical specificity, to each of several frontal lobe cortical regions critical for distinct executive functions. As mentioned in the discussion of absolute pitch, the posterior dorsolateral frontal cortex is proposed to play a role in conditional associative learning, one higher order frontal lobe memory function.

First in the hierarchy of proposed frontal lobe contributions, ventrolateral frontal cortex is hypothesized to be critical for the repetition, selection, comparison, and judgment of stimuli held in working memory. Auditory cortex may be sufficient for the passive retention of tonal information, but ventrolateral frontal cortex may be required for any form of more active maintenance. Furthermore, as suggested by the patient study of Zatorre and Samson and the asymmetry in the aforementioned rehearsal intervals, there may be a complementary right hemispheric asymmetry in frontal as well as temporal cortical contributions to pitch processing.

In a PET experiment inspired in part by the lesion analysis of pitch retention described previously, Zatorre presented eight-note melodies and instructed subjects to simply listen or to compare the pitch of either the first and the second notes or the first and final notes. The melody perception condition was discussed previously. When the two-note condition was contrasted to melody perception, frontal activation was observed in right frontopolar cortex. For comparison of the first–last pitch vs perception, additional increases were observed bilaterally in both ventrolateral and mid-dorsolateral frontal cortex, and there was also an increase in the right middle temporal gyrus.

Thus, only in the first–last condition was activation seen in the frontal regions predicted by Petrides' model of frontal lobe contributions to working memory. The two-note comparison, despite the fact that it requires judgment of pitch direction, was designed to make minimal demands on active maintenance of pitch information, which may not have been sufficient, particularly in contrast to melody perception, to result in measurable ventrolateral frontal activation. However, the first–last condition requires maintenance of the first pitch across interfering tones and then comparison to the last pitch. Since all the tones in the melody are related, it may not be possible to hold one tone in mind and simply ignore the rest. If subjects hold the entire sequence briefly in mind and compare the first and last tones mentally, then they are monitoring the contents of their working memory store. In a sense, this task thus resembles the *n*-back tasks, in which subjects are asked to compare the current item to one *n* items back, which have been demonstrated to activate mid-dorsolateral frontal cortex just as the self-ordered tasks do. Therefore, activation of both stages of Petrides' model by the first–last eight-note melody task is consistent with the task demands.

In a PET study designed to isolate the "monitoring" component of auditory–tonal working memory, Perry, Petrides, and Zatorre presented subjects with two tasks requiring them to monitor the contents of working memory. In a self-ordered condition, subjects were asked to sing six-tone sequences composed of two pitches by choosing each at random until a sequence with an equal number of both pitches was completed. In an externally ordered condition, they listened to five-tone sequences and sang the final pitch such that the sequences again had an equivalent number of high and low pitches. Both conditions were contrasted to a sensorimotor control involving singing a single pitch at the same rate. Increases were observed in both mid-dorsolateral and ventrolateral frontal cortex (Fig. 4), consistent with Petrides' two-level hypothesis of frontal lobe contributions to working memory that emphasizes the role of projections between mid-dorsolateral and mid-dorsolateral frontal cortex in monitoring. The activations within mid-dorsolateral frontal cortex were both greater in the right hemisphere, whereas no consistent asymmetry was observed within ventrolateral frontal cortex.

The portion of ventrolateral frontal cortex activated was actually deep within the upper bank of the horizontal ramus of the Sylvian fissure, part of cortical area 45 according to recent cytoarchitectonic analyses of this region of human frontal cortex by Petrides and

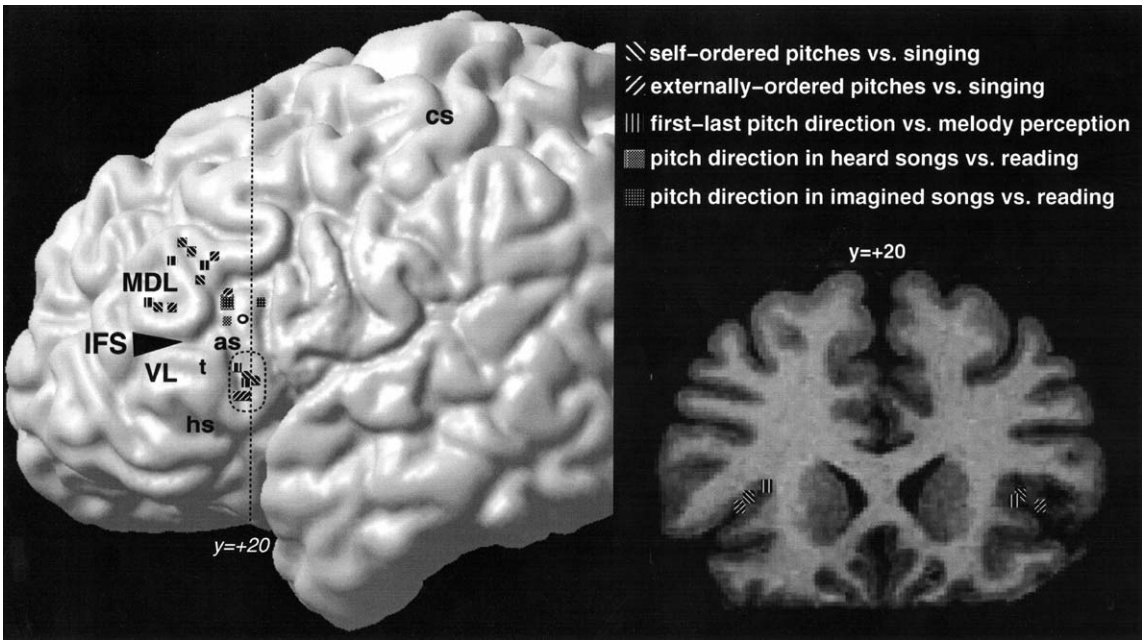


Figure 4 Mid-dorsolateral and mid-ventrolateral frontal activation during musical tonal working memory tasks. The coordinates of peaks of activation from three PET studies of rCBF changes during tasks involving working memory for musical tones. All data are group averaged and transformed into a common proportional stereotaxic space (i.e., Taillarach and Tournoux). (Left) The coordinates are plotted onto one hemisphere of a surface reconstruction of a similarly transformed normal brain chosen at random from the Montreal Neurological Institute database. All foci are displayed on the surface of the left hemisphere by projecting a line perpendicular to the ac–pc line until it meets the cortical surface. The black arrowhead marks the inferior frontal sulcus (IFS) that separates mid-dorsolateral frontal cortex (MDL), or cytoarchitectonic areas 46 and 9/46 in the nomenclature of Petrides and Pandya, from ventrolateral frontal cortex (VL), or areas 45 (pars triangularis; t), 47/12, and 44 (pars opercularis; o). The ascending sulcus (as) that forms the border between the pars triangularis and opercularis and the horizontal sulcus (hs), which branches from the end of the Sylvian fissure, are also indicated, as is the central sulcus (cs). Note that the self-ordered and externally ordered "monitoring" tasks and the first–last note judgment of pitch direction within melodies all resulted in highly overlapping sets of foci within MDL frontal cortex. All require maintaining and examining a mental record of recent events (i.e., tones). In contrast, the tasks involving pitch direction judgments within familiar songs, either heard or imagined, resulted in more ventral foci that fell in or just below the IFS within the pars opercularis or area 44. Finally, the monitoring tasks also resulted in a highly circumscribed and overlapping area of activation within the depth of the upper bank and fundus of the horizontal sulcus, part of area 45. The dotted circle around these foci serves to emphasize the fact that these foci are located deep to the cortical surface. The dotted line indicates the plane of the adjacent MRI slice. (Right) Coronal slice through the portion of sulcal area 45 with the coordinates of regional peaks of CBF increase marked by crosses. All foci fell within ± 2 mm along the y axis from the slice depicted [Perry *et al*., 1993; Zatorre *et al.*, 1994, 1996].

Pandya. Their localization accords precisely with the frontal opercular peaks observed by Zatorre in the first–last condition vs melody perception (Fig. 4) and in the first–last vs two-note in a later reanalysis.

Further evidence comes from another study by Perry, Zatorre, and Petrides designed to measure CBF during the simple operation of the tonal loop with no monitoring requirements. Subjects listened to two-note sequences drawn from a major scale, followed by a silent interval twice as long as the stimulus. In one condition they listened with no task imposed. In contrast to a silent baseline, this condition resulted in activation of right ventrolateral cortex as well as bilateral auditory cortex, more extensively in the right posterior temporal plane. The portion of ventrolateral cortex activated was also within the upper bank of the horizontal sulcus but near the lateral surface. Petrides and colleagues noted a similar distinction within the same region associated with auditory–verbal memory tasks (i.e., more lateral foci for simple verbal repetition vs rest, and more medial foci for free recall vs repetition). This simple and basic musical task thus activated the basic substrate proposed for the first stage of frontal executive processes in auditory–tonal working memory with a clear-cut right hemisphere asymmetry in both temporal and ventrolateral activations.

In order to measure activation during inner singing, or the covert rehearsal component of the tonal loop,

subjects were asked to immediately repeat the two tones once either was "in their head." The imaged condition, when contrasted to listening, resulted in activation of motor cortical areas, including the supplementary motor area and the right putamen —areas proposed to form part of the network supporting vocal fundamental frequency control and thus the rehearsal component of the tonal loop.

Robert Zatorre, Andrea Halpern, and David Perry examined CBF increases during a more complex task involving imagery for familiar songs with lyrics. Subjects were given two words and asked to judge whether the pitch associated with the second was higher or lower than the first. They either replayed the songs in their "mind's ear" until able to make the comparison or listened to a recorded rendition. In order to perform the imaged task, subjects had to play out the song in their minds in real time, as Andrea Halpern demonstrated by varying the distance between the two words in the song and examining reaction times. Both tasks activated very similar regions bilaterally in the temporal and frontal lobes, in the supplementary motor area (pre-SMA), and in the left supramarginal gyrus (Figs. 5.I–5.III). Their bilateral nature is in keeping with the integration of lyrics and tune in song representations, and the specific regions activated are consistent with the task's demands (i.e., the superior temporal gyri and lateral frontal cortex due to the involvement of both auditory–tonal and auditory–verbal working memory).

The foci of frontal activations fell in lateral frontal cortex just below or within the inferior frontal sulcus (Fig. 4) and thus in ventrolateral rather than mid-dorsolateral frontal cortex as did those during the similar first–last pitch comparison within novel melodies. Because the songs were very familiar, no record of events, mental or actual, needed to be kept. Ventrolateral frontal cortex activation is clearly predicted by Petrides' model due to the requirement to make a judgment on stimuli held in working memory (i.e., of pitch direction).

Activation of the superior temporal gyrus was of course much greater in the song perception condition but was nevertheless observed, exclusively within auditory association cortex, during song imagery without any auditory stimulation.

The common activation of SMA by both the perceptual and imaged task variants may be due to the fact that subjects were still at least partially "singing along" mentally with the presented song in order to disentangle the normally integrated representation of lyrics and melody in familiar songs.

Activation unique to imagery was seen in frontopolar cortex bilaterally, in subcallosal cingulate cortex, and in the right thalamus, possibly due to the retrieval demands of the imagery task (Fig. 5.IV).

In a recent study, Halpern and Zatorre measured CBF changes during tasks based on imagining well-known lyricless tunes, cued by the first few notes. When cued tune imagery was contrasted to hearing a random permutation of the five-note cue, activations were seen in the right superior and inferior temporal gyri, left sulcal 45, bilateral mid-dorsolateral frontal cortex, bilateral precentral gyri, anterior cingulate and the SMA. Of these areas, only the SMA and precentral gyri were activated during imaging of the random control sequence vs hearing it, without cued melody retrieval.

The activation of mid-dorsolateral and -ventrolateral frontal cortex by cued recall is consistent with the results of studies in other informational domains. The lack of ventrolateral frontal cortex activation during recall of the random sequence may be due to the fact that the contrasted condition (listening to the same sequences) also activated ventrolateral cortex, just as listening to two-note musical sequences did in contrast to silence.

All three studies of musical imagery demonstrated activation of the SMA, a motor area also active during singing out loud. This accords well with the proposed inner singing component of the tonal loop. It remains to be demonstrated whether auditory imagery for less clearly vocalizable sounds will also activate SMA. However, for imagery of melody, song, or musical intervals, inner singing and hence activation of SMA and other motor areas associated with vocal motor control may be obligatory. Thus, a remaining question concerns the dissociability of auditory imagery from inner vocalization.

C. Long-Term or Episodic Memory

After being thoroughly encoded, and perhaps after many repetitions, musical information may be stored in long-term memory so that it can be consciously retrieved even after it has been erased from the contents of working memory. For example, a melody may be consciously retrieved ("What is the theme to 'The Lone Ranger' or the 'William Tell Overture'?"), as may memorable instances or episodes of hearing it.

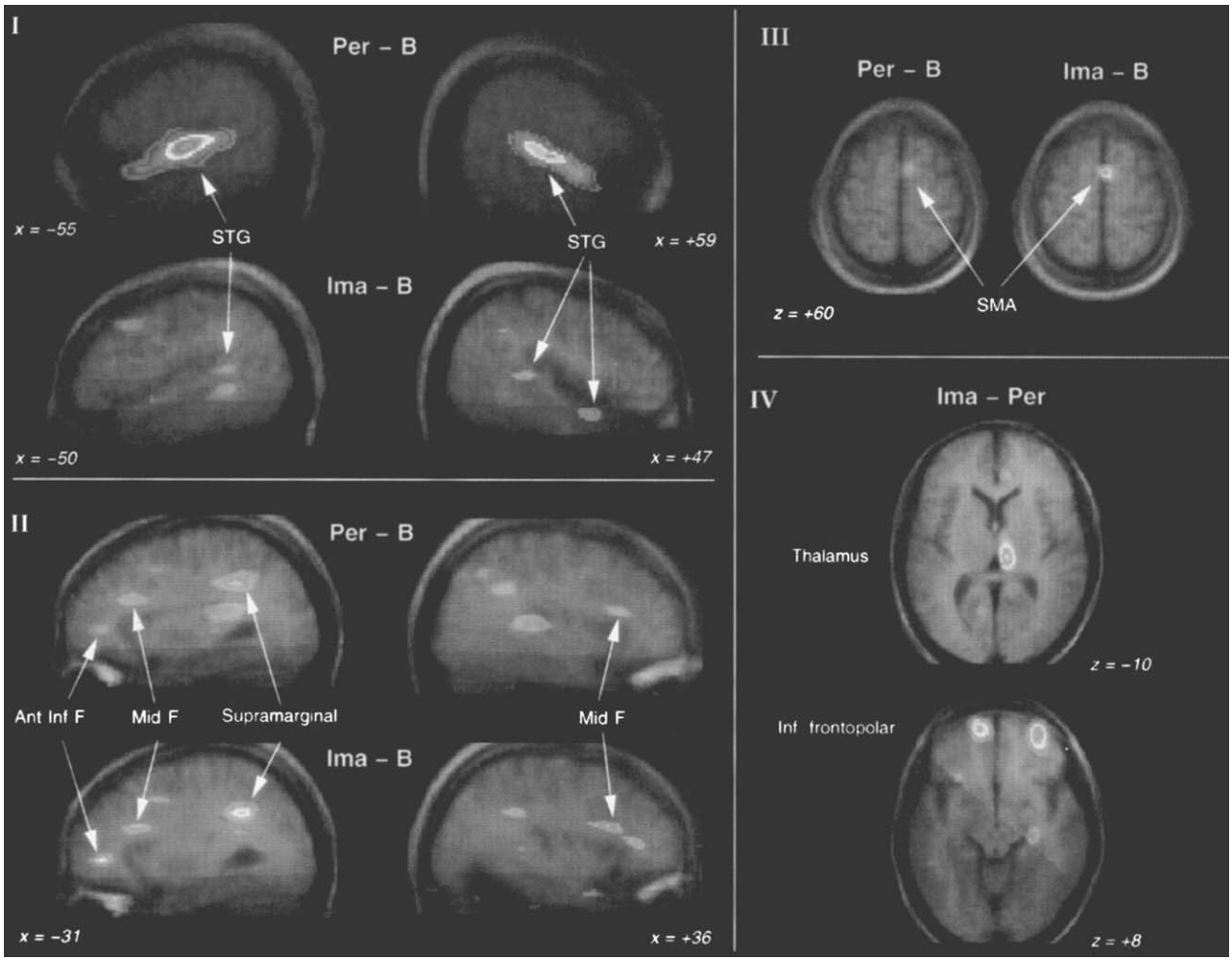


Figure 5 Pitch direction judgments within perceived or imaged songs. CBF increases during pitch comparisons between two notes within familiar songs, cued by the presentation of two words from their lyrics. Both conditions were contrasted to a visual baseline consisting of reading similar words. In the imagery (Ima) condition, subjects had to replay the song mentally, whereas in the perception (Per) condition they heard a recorded rendition. (I) Left and right saggital views through peaks in the superior temporal gyri. Although much less robust, activation similar to actual perception was observed during song imagery in secondary auditory cortices, with no auditory stimulation. (II) Left and right saggital sections through frontal lobe foci. The overall patterns of activation are strikingly similar for perception and imagery and include activation in mid-dorsolateral frontal cortex, consistent with the monitoring demands of this task. (III) CBF increases within the supplementary motor area (SMA), believed to be associated with inner singing, are more robust during imagery. (IV) Increases during imagery in contrast to perception. Foci are in the right thalamus and bilateral frontopolar cortex (reproduced with permission from Zatorre *et al.*, 1996).

The formation of long-term memories has been extensively studied, and general mechanisms, such as the critical importance of entorhinal cortex in the formation of new long-term memories, must be common to music. However, little work of specific musical relevance has been carried out. Important observations from the cognitive study of memory for melody, such as the greater importance of exact interval information in long-term memory for melody as opposed to the predominance of contour in short-term recall, will no doubt inform the design of such experiments.

D. Procedural Memory

Procedural or implicit memories can also be formed without conscious awareness. Their effects can be detected in facilitated responses to primed stimuli. The chord priming paradigm developed by Jamshed

Bharhucha has been used to test an auditory agnosic, and the results suggest that this sort of priming depends on secondary rather than primary auditory cortex. Procedural learning can also be demonstrated in melodic paradigms, in which previous exposure to a tone sequence can prime choices for its completion later without conscious awareness of having heard it before. The neural bases underlying the formation and retrieval of such procedural memories have not been explored in specifically musical contexts. Procedural learning is ubiquitous in music perception and in music production. A full understanding of how the human brain processes and produces music will depend in part on further studies in this area.

E. Lexical–Semantic or Cultural Memory

Musical stimuli are often associated with linguistic labels. In the PET study of musicians with absolute pitch, the association of verbal labels (note names) with musical pitches was hypothesized to depend in part on left dorsolateral frontal cortex, as was the association of the names of musical intervals (e.g., major/minor third) with pairs of tones whose frequencies formed the same ratio. Many other types of verbal labels are associated with musical sounds: the names of melodies, songs, or pieces; composers; styles; and specific musical elements or functions (e.g., eighth note and decrescendo). Amusias following focal damage or dementia may sometimes be confined to the loss of the ability to retrieve such verbal associations, particularly following left hemisphere damage.

IV. MUSICAL EXPRESSION

In order to perceive music, someone must first produce it. Although computer-generated music can be highly abstract, most music is played on an instrument, sung, or possibly whistled. These are fundamentally motoric functions. What makes them uniquely musical, besides the high level of skill embodied in virtuoso performance, is the fact that the end result is musical sound, and an exquisite degree of sensorimotor integration (and attention to auditory feedback) is required.

Two levels of motor nervous system functioning relevant to musical expression are discussed next: motor productive, or the basic aspects of motor control necessary to produce musical sounds, and motor programmative, or the cortical control required to produce these sounds and the movements that engender them in sequences of varying length and complexity, from single notes to complex pieces 1 hr or more in length.

A. Motor Productive

Probably the most fundamental musical motor activity is singing. Perry, Zatorre, Petrides, and associates measured increased CBF during simple singing using PET. Repetitive singing of a single pitch was contrasted with listening to complex tones at the same pitch and rate. The set of regions activated overlapped those previously observed during speech (SMA, anterior cingulate, insula/frontal operculum, and precentral gyrus) (Fig. 6). The main differences were in the direction of hemisphere asymmetry within a subset of these regions. First, the CBF increase was much greater in the right primary auditory region, a result that Perry and Zatorre later replicated for singing a single pitch continuously on each breath in contrast to listening to playback of that singing. They hypothesized that this asymmetry may be related to deriving the fundamental frequency of one's own voice for feedback guidance of vocal motor production (given a right auditory cortex preference for spectral analysis).

This hypothesis received support from a subsequent analysis. The fundamental frequency of subjects' vocalizations was measured. When the total amount of pitch excursion within each continuously sung note was quantified and covaried against CBF in the whole brain, a region of positive covariation was observed in the right primary auditory region.

Second, though less striking, an asymmetry favoring the right hemisphere was also observed within the right ventral precentral gyrus or the orofacial region. The close correspondence between the regions activated by singing and speaking suggests that both may have evolved from a complex system for the voluntary control of vocalization. Their divergences suggest the later evolution of complementary hemispheric specializations for both the perception and production of singing and speech.

An incredible variety of musical instruments have been developed by human cultures, our knowledge of which begins with 5000-year-old Neanderthal flutes made from bone. The proliferation of wind, string, percussive, and other instruments for the production of musical sound is ongoing, and the basic aspects of motor control required vary from purely labial and respiratory (e.g., bugle) to labial, respiratory, and manual (e.g., trumpet) or purely manual (e.g., violin and piano). Although peripheral aspects of this motor

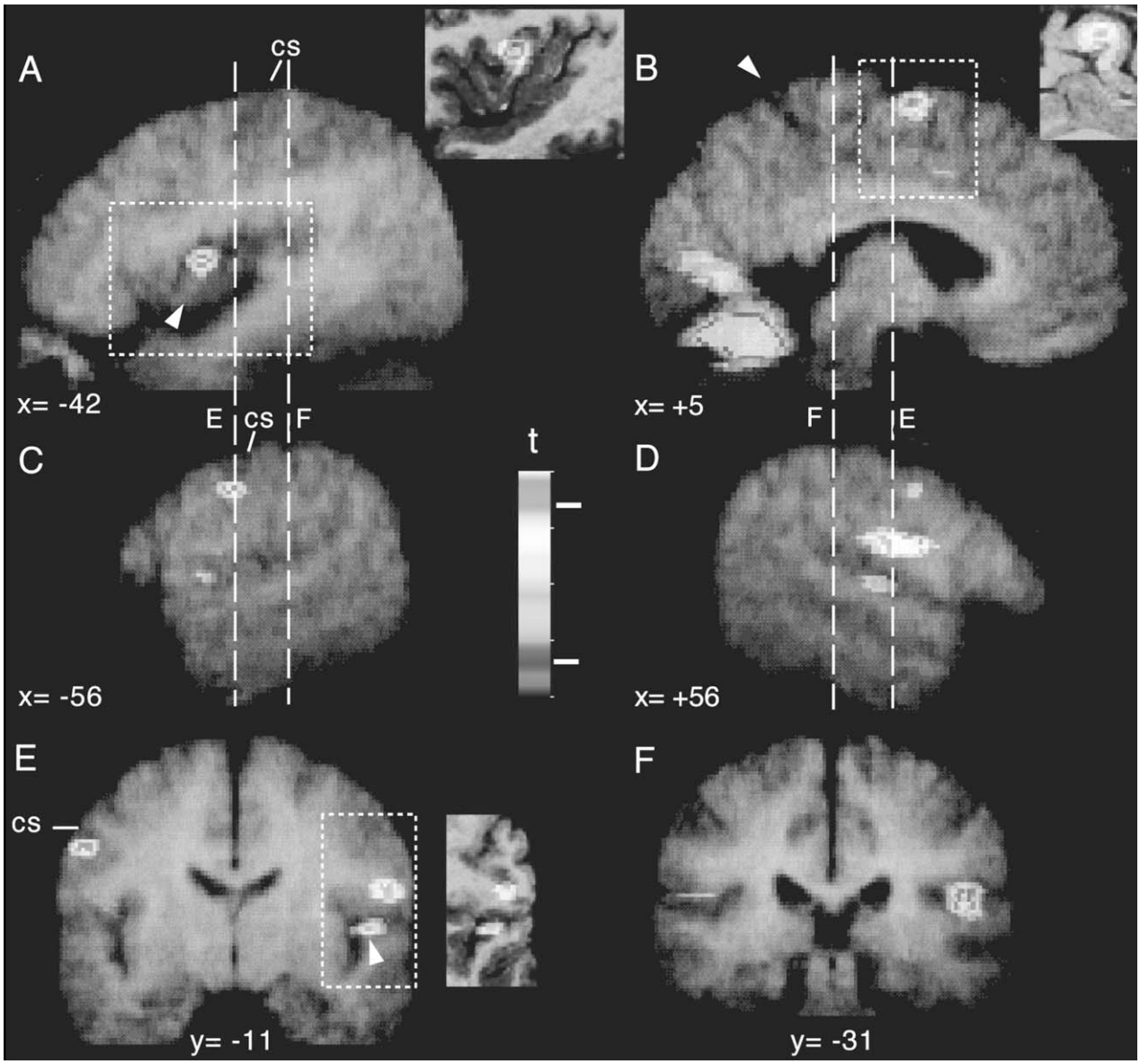


Figure 6 Localization of cerebral activity during singing. Peaks of CBF during repetitive singing of a single pitch on the vowel /a/ at a rate of about 1250 msec as measured with PET. Saggital (A–D) and coronal (E and F) slices from the mean CBF change t-statistic volume superimposed on the averaged MRI in Taillarach coordinates from a group of 13 nonmusician subjects. In critical regions, the same rCBF data are shown superimposed on an individual MRI in order to view their relationship to sulci not clearly visible in the group average. The dotted rectangles in slices A, B, and E correspond to the regions depicted in the individual MRI offset to the right. Peaks were observed as follows: (A) in the first long gyrus (precentral) of the insula, also critical for motor speech; (B) (from top to bottom) the supplementary motor area, anterior cingulate sulcus (note the marginal ramus of the cingulate sulcus; arrowhead), visual cortex activation that is part of a right lingual peak, and cerebellar activation coextensive with more lateral peaks; (C) relatively dorsal left central sulcus (note the position of the central sulcus at the dorsal edge of the slice and in the more medial slice A) (also visible is the lateral extension of the insular peak seen in A); (D) dorsal and ventral right central sulcus—the activated voxels just below the Sylvian fissure extend from the more medial peak in Heschl's gyrus; (E) right Heschl's gyrus (position indicated by the arrowhead and more clearly visible on the individual MRI), dorsal left central sulcus (seen also in C), and ventral right central sulcus (also seen in D); (F) right superior temporal plane (STP) and the overlying right parietal operculum and the left parietal operculum (reproduced with permission from Perry *et al.*, 1999).

control have been studied for many years (e.g., the work of Otto Ortmann on piano performance), study of the cortical control required in specifically musical contexts is sparse. A few studies have utilized computer-interfaced piano keyboards to study piano performance by musicians, the cortical control of sequential finger movements in nonmusicians, or the production of rhythms using a single computer key as if it were a piano key. The simplest sequences studied were repetitive, isochronous presses of a single key. In contrast to the vocal studies mentioned previously, in several of these studies no sound was produced. Although the behaviors are thus not fully musical, they allow separation of motor control from the auditory–motor integrative aspects.

Penhune and Zatorre measured CBF increases during the reproduction of an isochronous sequence (cued either visually or auditorily) on a single computer key using the right hand, in contrast to perception of the cue alone. Common areas of activation were seen in the contralateral somatomotor cortex, basal ganglia, and ipsilateral cerebellum. Activation in the SMA was also seen, but only for the auditorily cued condition. In a similar auditorily cued task, Stephen Rao and colleagues used functional MRI to examine CBF as subjects continued tapping a key at a fixed rate with the right hand. Increases were seen in the same areas: contralateral somatomotor cortex, basal ganglia, ipsilateral cerebellum, and SMA.

These studies illustrate the cortical areas involved in the control of manual motor output in a quasi-musical context, consistent with what would be expected for timed manual movements in general. These areas are similar to those seen during isochronous, isofrequency vocal motor output, except that processing of auditory feedback is excluded and somatosensory feedback is emphasized.

B. Motor Learning

Even the simple acts of motor production described previously (e.g., singing an isochronous series of a single pitch or tapping an isochronous sequence on a single key) require the execution of a learned motor program or an integrated sequence of movements. More complex sequences that follow a precise temporal plan and involve multiple vocal or instrumental pitches and variable timing and intensity are integral to musical expression. Such sequential movements performed as a unit require advance programming prior to their execution.

In the experiment discussed previously (see Section III.B), Perry, Zatorre, and Petrides asked their subjects to listen to a two-note sequence drawn from a major scale and either imagine it in their heads, or hum it out loud. When they hummed out loud, in contrast to just listening to the sequences, activation was seen not only in motor cortex, SMA, and the right putamen as seen during imaging but also in the left putamen, cerebellum, right primary auditory region, and more extensively in motor cortex bilaterally. Thus, this simple act of musical motor programming activated the same areas thought to be involved in the execution of movement sequences generally (i.e., primary and premotor cortex, SMA, basal ganglia, and the cerebellum).

The SMA activation fell anterior to that seen during simple, repetitive singing of a single pitch. Rather than falling in "SMA proper," as defined by Nathalie Picard and Peter Strick in their cross-species neuroanatomical analysis, it fell in pre-SMA, a cortical area thought to be important for producing organized sequences of movements. Thus, more complex sequential singing resulted in additional activation of premotor cortex, pre-SMA, and the basal ganglia (putamen).

Justine Sergent and associates carried out a study of right-handed piano performance using PET. When playing a scale was contrasted to listening to scales, CBF increases were seen in contralateral primary motor cortex, SMA proper, and the ipsilateral cerebellum.When sight-reading an unfamiliar Bach melody was contrasted to reading a score of it while listening to a recorded performance, additional activation was seen in contralateral premotor cortex and in pre-SMA. Although the increase in motor programming complexity between playing scales and sight-reading a melody would seem to be much greater than that between singing a single pitch and singing two-note sequences, the results are highly similar. Both comparisons emphasize the activation of premotor cortex and pre-SMA during the expression of more complex and variable motor programs.

Although these two studies measured actual musical expression, few others have done so. However, others have studied quasi-musical tasks helpful for isolating components of musical expression.

In the study of rhythm production described in Section IV.A, Penhune and Zatorre asked subjects to reproduce sequences with elements of variable durations, either short (250 msec) or long (750 msec), again cued either visually or auditorily. In one condition, subjects reproduced a previously learned sequence

repetitively. When contrasted to the isochronous condition, additional activation during sequence reproduction was seen in the cerebellum and in the contralateral primary motor cortex only for the visually cued condition. In another condition, they reproduced novel sequences. In contrast to the repeated sequence condition, additional activation was seen across tasks in the cerebellar vermis and bilaterally in the cerebellar hemispheres. Additional activation was seen in pre-SMA and the basal ganglia for the auditory condition only.

At least for the auditorily cued condition, these results for rhythm production parallel those described previously for singing and piano performance: activation of SMA proper for simple, predictable movements and of pre-SMA for more complex and less predictable movement sequences. The activation of the basal ganglia by the task placing more demands on motor programming and execution is also consistent for both the vocal pitch and manual rhythm reproduction tasks.

Christian Gerloff and associates used repetitive trains of transcranial magnetic stimulation (TMS) that temporarily disrupt neural activity to examine its effects on the performance of finger sebquences on a (silent) piano keyboard. The sequences differed in complexity, and TMS was applied over the contralateral primary motor cortex or over the SMA. Subjects thoroughly learned the 16-item sequences first. The simplest sequence consisted of 16 repeated index finger presses of a single key; a scale-like sequence consisted of alternating 4-item ascending and descending subsequences; and the most complex was a variable 16-item sequence. Stimulation over the primary motor cortex induced errors in both the complex and scale-like sequences and, with sufficient intensity, in the simple repetitive sequence. Stimulation over the SMA disrupted only the complex sequence.

These TMS results are consistent with activation studies that demonstrated increased activation of primary motor cortex and pre-SMA with increased demands for motor sequence planning. Although the role of primary motor cortex was once thought to be restricted to the execution of simple voluntary movements by individual muscles, increasing evidence indicates its involvement in the learning and execution of more complex movement sequences.

Norihiro Sadato and colleagues measured CBF increases during the execution of sequential finger movements (touching fingers to thumb) of varying complexity using PET. Each sequence (4–16 items) was thoroughly learned before the PET scan, and the rate was determined by following a metronome. Activation in one set of motor-related areas did not vary with the complexity of the sequences: bilateral primary sensorimotor cortex, contralateral ventral premotor cortex, SMA proper, contralateral putamen, and ipsilateral cerebellum. Four areas showed a linear increase in CBF with increasing sequence length: right dorsal premotor cortex, right superior parietal cortex, left thalamus, and the cerebellar vermis. The association of sequence complexity with premotor cortex activation agrees with results of subtractive studies mentioned previously. However, the basal ganglia were activated equivalently by all tasks and activation of pre-SMA was not demonstrated.

Subtle differences in task and experimental design are well-known to affect the outcome of imaging studies, and major differences exist between all these tasks. For example, Sadato's study required synchronization with an auditory metronome, perhaps resulting in activation of the basal ganglia across tasks. Future studies systematically varying such task parameters will further our understanding of the roles of each of these areas in movement sequencing.

Precise programs for movement sequences must first be learned. Then they can be retrieved and executed as one unit. Some studies have attempted to minimize learning by having subjects thoroughly learn the sequences beforehand (e.g., Gerloff, Sadato). Obviously, study of the brain mechanisms supporting learning as well as retrieval and execution of complex motor sequences that produce musical sound is central to a biological understanding of musical behaviors.

One study used TMS in a very different manner to investigate cortical changes associated with learning a motor sequence on a piano keyboard with auditory feedback. Alvaro Pascual-Leone and colleagues repeatedly mapped the contralateral motor cortex over a 5-day period during which subjects practiced a simple five-finger exercise on a computer-interfaced piano keyboard. They learned to play a 10-item scale-like sequence consisting of 5 ascending and 5 descending elements, with the goal of playing it evenly at a specified rate. None had any musical instrument-playing experience. Thresholds for the production of motor evoked potentials (MEPs) in response to single TMS bursts were measured from the contralateral hand (flexor and extensor muscles). Contour maps of the probability of inducing MEPs demonstrated increases in the cortical motor areas targeting the practiced hand and decreases in threshold. No such changes were seen for the untrained hand, and only modest gains were seen for subjects who simply played the piano at will for the same length of time (Fig. 7).

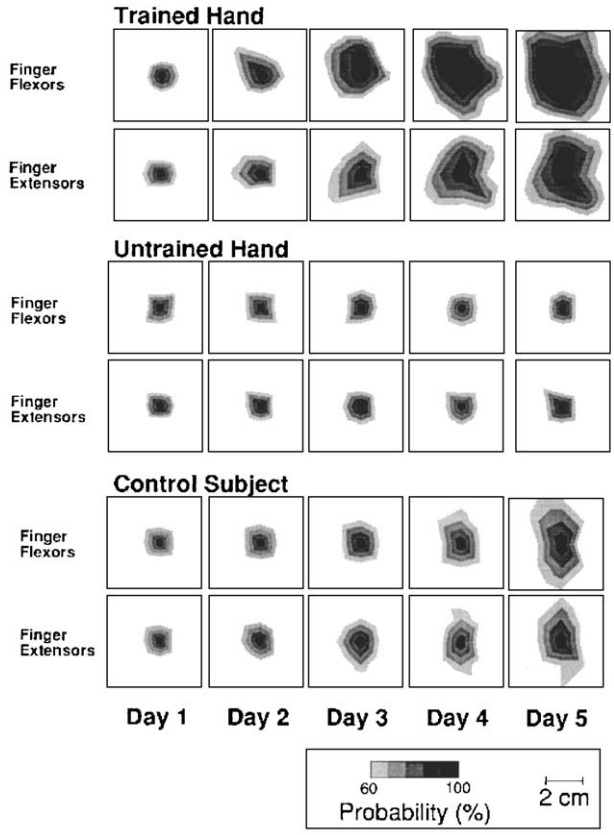


Figure 7 Piano exercise learning, trained vs untrained hand. Representative examples of cortical motor output maps for the long finger flexor and extensor muscles on Days 1–5 from a subject who practiced 2 hr/day (trained right and untrained left hand) and from one control subject who played random sequences for the same amount of time (trained right hand) (reproduced with permission from Pascual-Leone *et al.*, 1995).

Other functional imaging studies of motor sequence learning are less directly related to musical expression. Nevertheless, they are relevant for understanding the motor sequence learning component, isolated from sound production. Though complex, the results suggest an expansion of the representation of the involved fingers within primary motor cortex. A. Karni and colleagues asked subjects to learn a particular sequence of finger-to-thumb opposition movements (e.g., 4–1–3–2–4) for 10–12 min a day over the course of 5 weeks. Once a week, cortical activity was measured during performance of the practiced sequence and during performance of an unpracticed sequence (e.g., 4–2–3–1–4) with fMRI.

After 3 weeks of practice, the extent of primary motor cortex activated by the learned sequence was greater than that activated by the unpracticed one. When retested several months later, the skill and the increased extent of motor cortex activation persisted. Because the same fingers were used in both sequences, this increase is not in the cortical representation of particular fingers but in the representation of a particular sequential combination of finger movements.

William James' speculations regarding the brain mechanism underlying memory as "paths" within the brain's tissue can be applied not only to melodic memory but also to memory for the movement sequences used to produce them (i.e., as paths within somatotopically ordered cortical areas).

Other studies have shown activation of cerebellum, basal ganglia, thalamus, and dorsolateral frontal cortex during skill learning, decreasing with increasing acquisition. Plateaus in the increased activation within motor cortex during skill learning have been demonstrated, as have decreases after complete acquisition. The precise contributions of cortical and subcortical areas, including the basal ganglia, to the execution, learning, and retrieval of movement sequences remain controversial and a subject of active research.

Movement sequence learning in the service of musical expression adds further complexity by introducing the production of musical sound as the ultimate goal. It also introduces the possibility of sampling across wide variations in the degree of skill acquisition.

Extending his previous research, Pascual-Leone used single pulses of TMS and measurement of the consequent MEPs to map changes in cortical output maps during implicit and explicit sequence learning. Subjects rested four digits over four computer keys and pressed the keys sequentially in response to a visual cue. Unbeknownst to them, these cues corresponded to particular sequences of varying length. Even before becoming aware of this sequence, reaction times progressively declined (implicit learning) and rebounded when the order was changed. During this phase, the cortical motor output map to the involved contralateral muscles increased. Uninvolved muscles (i.e., the thumb) showed no change. After subjects became aware of the sequence, learning could proceed by both implicit and explicit means, and the maps tended to plateau. Finally, once the sequence was fully, explicitly learned and response times plateaued, there was a rapid reduction in the cortical output map toward baseline.

Using a different technique, trains of TMS were used to disrupt learning in the serial reaction time task during 60-sec blocks over dorsolateral frontal cortex

or SMA. The effects of stimulation were inferred from task performance and were also tested during a final random block. Stimulation over the contralateral dorsolateral frontal cortex resulted in a complete disruption of implicit learning, indicating that it is not only activated during such learning but also critical to its success.

Motor sequence learning in musical context is always explicit, at least for the intended auditory result. Nevertheless, simultaneous implicit learning must play a crucial role, particularly for the actual sequence of fingers. Thus, these results are directly relevant to hypotheses for the motor learning underlying musical expression.

C. Motoric Working Memory/Imagery

As many musicians have discovered, musical instrument playing can be rehearsed mentally. For example, in the piano keyboard melody recall task described previously, the pianists reported mentally rehearsing their keyboard recall attempts, not only through inner singing but also through mental simulation of keyboard fingering, prior to actually playing. To investigate the effects of mental practice, Pascual-Leone, in the piano keyboard sequence learning experiment described previously, also studied changes in motor cortical output maps and in performance following 5 days of *mental* practice. Subsequent performance improved steadily across the 5 days, although not as much as after an equivalent amount of physical practice (Fig. 8A). However, the increase in the extent of the cortical motor output maps was the same (Fig. 8B).

Brain imaging studies have demonstrated the activation of many of the same areas during the mental simulation of motor acts as during their overt physical performance (e.g., SMA, premotor cortex, basal ganglia, and sensorimotor cortex). In fact, all these areas were activated when subjects mentally sang musical intervals in the experiment described in Section III.B. However, for *musical* motor tasks, imagined sound output is also involved and in fact constitutes the explicit focus of attention. Activation was also seen in the ventrolateral frontal cortex and the posterior superior temporal sulcus.

In fact, this study, one of the simplest examples of musical expression or imagery, indicates the complexity involved in their study. Both fundamentally involve integrated motoric and auditory representations and the integrated functioning of motoric and auditory

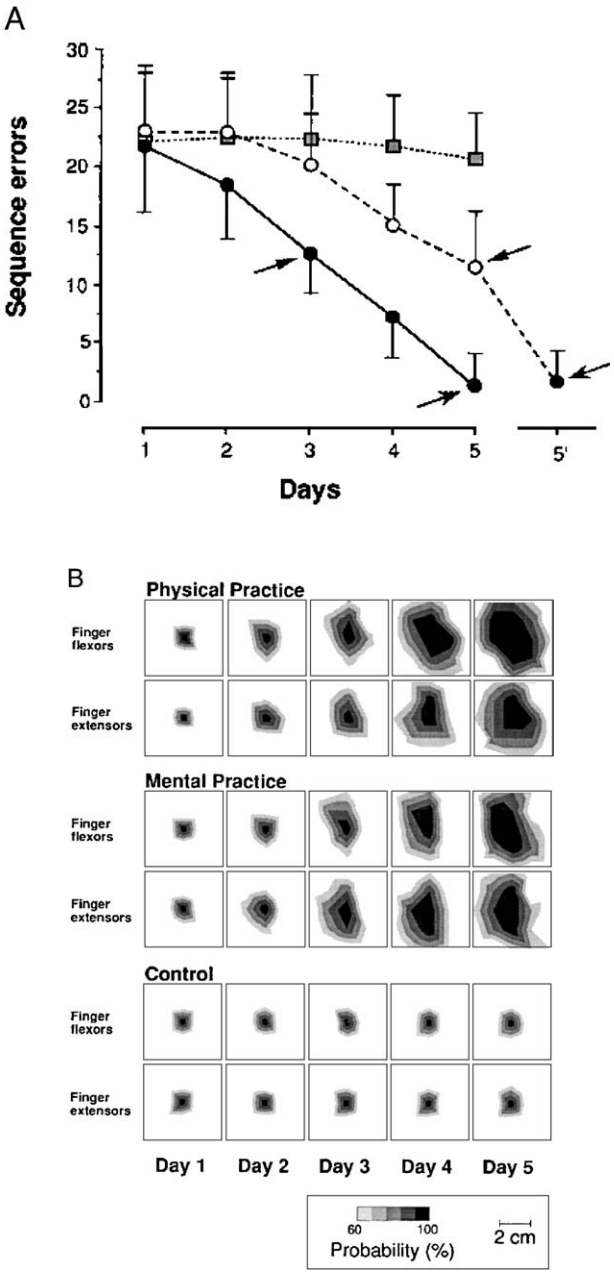


Figure 8 Piano exercise learning, physical vs mental practice. (A) Sequence errors: The number of errors in the order of notes on a posttest of 20 repetitions of the practiced exercise. ●, physical practice group; ○, mental practice group; ■, control group. 5′, in one physical practice session, the mental practice group attained the same performance level as that of the physical practice group. (B) Cortical motor output maps: Representative examples of cortical motor output maps for the long finger flexor and extensor muscles on Days 1–5 from a subject who practiced 2 hr/day, another who practiced mentally (i.e., sat in front of the piano for 2 hr and visualized the fingers performing the exercise while imaging the sound), and a control subject who played random sequences (all for the trained right hand). (reproduced with permission from Pascual-Leone *et al.*, 1995).

working memory systems. Thus, the song and imagery tasks described previously involve not only auditory imagery but also, at least implicitly, motoric simulation. Thus, all three tasks (imaging two-tone musical sequences, melodies, or songs) activated pre-SMA, in contrast to simple singing of a single note or imaging of an overlearned melody, which activated SMA proper.

The musical interval and melody imagery tasks also activated premotor cortex and the superior temporal gyrus—further evidence of simultaneous activity in the motoric and auditory working memory/imagery systems. Further work is needed in order to understand the functioning of each system in isolation as well as the mechanisms that allow their integration in melodic working memory and even in long-term melodic representations.

D. Sensorimotor Integration

One of the most defining characteristics of musical expression is that specific motoric actions produce specific musical sounds (i.e., that motor production and auditory feedback are tightly and, particularly for musical instruments, artificially integrated). Furthermore, somatosensory feedback must also be integrated, guiding motoric actions in order to produce the intended sounds. Finally, at least for playing musical instruments with eyes open, visuospatial feedback may also be closely integrated.

As discussed previously, simple singing of a single pitch or of musical intervals resulted in asymmetric activation of the right primary auditory region—activation that increased with the extent of pitch excursion. Thus, activation of primary auditory cortex in the right hemisphere appears to be particularly associated with auditory feedback from the pitch of one's own singing voice. Future studies, by manipulating feedback, could help to elucidate the neural mechanisms supporting auditory–motor integration in singing and instrument playing. Similar strategies may also shed light on the mechanisms of their further integration with visuospatial and somatosensory feedback.

V. MUSIC READING AND WRITING

Although not essential to musical behavior, many musical cultures have developed systems for graphically specifying musical sounds, intended to be read and studied by potential performers. Thus, musical reading and writing are learned behaviors that can be added to the musical functions described previously (much as linguistic reading and writing can be added to speaking). Like linguistic reading, writing, and speaking, these functions can be differentially affected by focal lesions.

A. Musical Lexical

One brain imaging study has addressed sight-reading. Justine Sargent and associates measured CBF during silent score reading of a Bach chorale and subtracted it from activation during perception of dots (and manual responses to their location). Activation during silent score reading was greater bilaterally in extrastriate visual cortex and in the left occipitoparietal junction. This suggests activity in the dorsal visual pathway, which is involved in spatial processing, as opposed to the ventral visual pathway, important for words. When score reading alone was subtracted from score reading while listening to a recorded rendition, activation was observed not only in auditory cortex bilaterally but also in the left supramarginal gyrus. However, the portion of the supramarginal gyrus activated was superior to that thought to be critical for verbal print-to-sound mapping.

Following brain damage, associations between musical and verbal alexia (or loss of reading ability) have frequently been reported, such as the cases of Bouillaud in 1865 and the French composer Maurice Ravel toward the end of his degenerative brain disorder. However, many cases of dissociation have been reported, including Ravel *earlier* in the course of his illness. Although some cases of selective musical alexia involve the right hemisphere, others involve the left hemisphere. Dissociations and associations may occur following left hemisphere damage, depending on whether adjacent but nonoverlapping critical areas are damaged together or separately.

B. Musical Graphic

Musical activation studies have not dealt with the problem of unraveling the neural substrates of music writing for obvious reasons, given the complexities involved (e.g., complex movement sequences, visual–spatial organization, and translation from auditory imagery working memory to notation). Our only other source of information is the sequelae of focal brain

damage, which again may result in dissociations between musical and verbal writing. Several musicians have become verbally agraphic but continued to write music. One, described by Signoret and colleagues, was an organist and composer who, following a left MCA stroke, became alexic and agraphic in Braille but continued to sight-read and compose music (in Braille). Thus, verbal and musical reading can be dissociated not only in their usual visual forms but also within the tactual modality.

VI. MUSICAL EMOTION

The neural substrates permitting the perception and expression of emotions through music are a relatively unexplored dimension of musical behaviour despite their obvious centrality to the enjoyment and importance of music. However, investigations into the perception of affect are under way.

Anne Blood, Robert Zatorre, and colleagues used PET to measure the correlation between CBF changes and the degree of dissonance or consonance (and unpleasantness or pleasantness) in musical excerpts constructed by altering the harmonic accompaniment of a novel tonal melody (Fig. 9A). Activation in the right parahippocampal gyrus and right inferior parietal lobe increased with increasing dissonance; that in portions of the orbitofrontal and subcallosal cingulate cortex increased with increasing consonance (Fig. 9B). Volunteers also rated the emotional quality of the excerpts along adjective pairs (e.g., pleasant vs unpleasant), two of which were known not to correlate with dissonance (happy vs sad). The fact that dissonance was associated with particular emotions (e.g., tense, unpleasant, irritated, and angry) and not others (Fig. 9A) suggests that the areas identified are involved in the perception of certain emotions, but not all.

The parahippocampal region is strongly and reciprocally connected with the amygdala, possibly involved in processing these emotions. CBF increases in the parahippocampal gyrus are also associated with unpleasant emotions in response to pictorial stimuli. Orbitofrontal and subcallosal cortex have been associated with emotional processing (e.g., the former with emotional disinhibition in monkeys following selective lesions, and deficits in emotional identification following subcallosal lesions in humans).

Blood and Zatorre measured the neural correlates of the experience of thrill during music perception. Though studied from a cognitive point of view by John Sloboda, this is probably the first study of the cortical and subcortical correlates of musical thrill. Subjects heard self-selected musical excerpts and made postscan ratings of subjective emotional intensity. Increased activity was associated with increasing intensity of thrill in the nucleus accumbens, midbrain, insula, thalamus, SMA, anterior cingulate, and orbitofrontal cortex. Decreasing activity was observed bilaterally in the amygdala and ventrolateral frontal cortex. These complex findings suggest the involvement of neural systems known to be associated with reward and arousal in the generation of musical thrill. Many of these areas have also been implicated in reward responses to other euphoria-inducing stimuli, both naturally occurring (e.g., food, sex) and artificial (e.g., cocaine, heroin). Thus music listening appears to serve some important biological function.

Peretz and colleagues noted that one of their auditory agnosia patients who no longer recognized any previously familiar melodies (''tune'' agnosia) and could not detect pitch or rhythmic variations within novel melodies, utilize melodic contour or interval information, or sing, nevertheless reported that she still enjoyed music. Tests of her ability to discriminate the happy vs sad emotional valence of musical excerpts revealed that this ability was entirely intact. When the stimuli were presented in computer-synthesized versions rather than as recordings of live performances, emotional discrimination was unaffected, indicating that at least for these stimuli, expressive variations in timing and intensity were not significantly contributing to their emotional processing. The judgments were demonstrated to rely primarily on the mode (major/minor) and tempo of the selection. This patient suffered sequential strokes affecting the left and right MCAs, which left extensive damage to almost the entirety of the left superior temporal gyrus (STG), the anterior third of the right STG, the middle and inferior temporal gyri, the left insula and frontal operculum, the right inferior and middle temporal gyri, and the left anterior parietal area. Though many hypotheses could be generated about the musical functions that might be impaired by such damage (e.g., singing and auditory–tonal working memory), it is interesting to note that most of the regions implicated by the PET studies of musical emotion described previously (with the exception of the left insula and left anterior parietal lobe) are spared. On the other hand, the emotional valence studied (happy vs sad) was not addressed by either of the activation studies. Clearly, much work remains to be done before the neural networks critical for the perception of specific categories of musically expressible emotions can be sorted out.

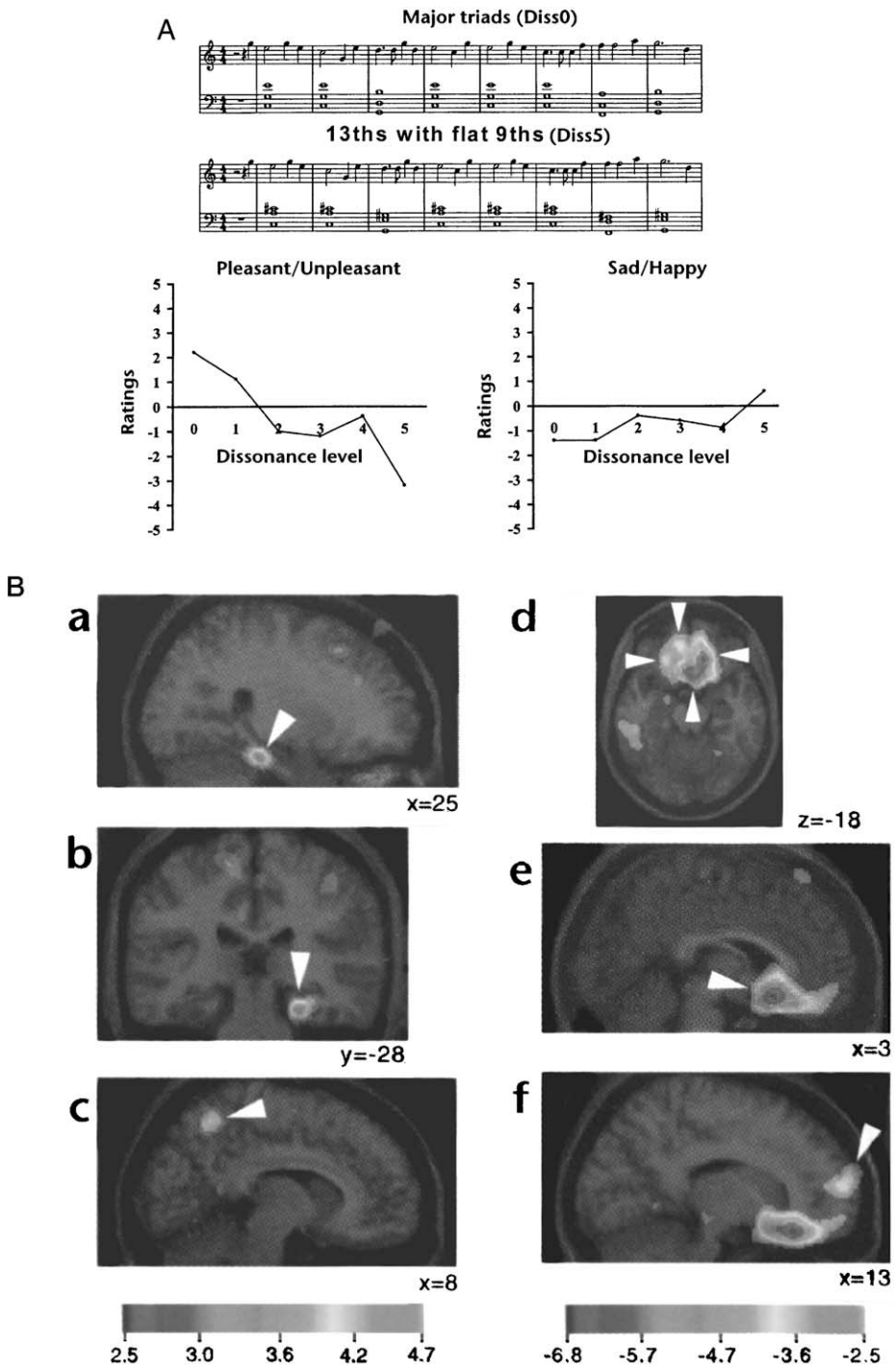


Figure 9 Consonant vs dissonant music perception. (A) (Top) Excerpts from the most consonant version (major triads; Diss0) and the most dissonant version (13ths with flat 9ths; Diss5) of the musical stimuli presented during PET scanning. (Bottom) Line graphs demonstrating averaged subject ratings following scans for each of the six versions Diss0–Diss5. Note that dissonance was related to judgments of pleasantness/unpleasantness but not happiness/sadness. (B) Regions demonstrating significant rCBF correlations with dissonance level, parametrically varied from consonant to most dissonant in five steps by altering the harmonic accompaniment to a novel tonal melody. Correlations are shown as *t*-statistic images superimposed on corresponding averaged MRI scans. (a–c) Positive correlations in rCBF with increasing dissonance in (a) right parahippocampal gyrus in saggital section and (b) coronal section and (c) right precuneus. (d–f) Negative correlations with increasing dissonance (equivalent to positive correlations with increasing consonance) in (d) bilateral orbitofrontal cortex, (e) medial subcallosal cingulate in sagittal section, and (f) right frontopolar cortex (reproduced with permission from Blood *et al.*, 1999).

Although in the context of Peretz's task of emotional classification, expressive microvariations in timing, intensity, or pitch did not play a significant role, clearly in other contexts they are manipulated by skillful musicians to effectively communicate musical emotion. Examples of masters in this musical domain include musicians such as Mahalia Jackson, Maria Callas, Nusrat Fateh Ali Khan, and Claudio Arrau. Future studies may help to determine the neural processing consequences of both the sensory characteristics used to convey particular emotions (e.g., consonance/dissonance, tempo, harmonic structure, and intensity) and the particular emotions (e.g., happiness/sadness, pleasantness/unpleasantness, and thrill).

Finally, the neural substrates supporting the *expression* of musical emotion have yet to be explored (i.e., the motor control of the actions producing these expressive variations in musical sound and the perceptual processing of the emotional content of one's own performance). Though complex, to the degree that these expressive parameters can be quantified, neural correlates of their motor control may be sought.

VII. SUNG LANGUAGE

A significant portion of the music we listen to contains words set to music (i.e., language in which the fundamental frequency pattern has become predominant, rather than providing an emotional background and semantic support as in the prosodic component of spoken language). The melodic component may be so predominant that it masks or distorts perception of the linguistic component.

A. Integrated Perception of Music and Language

The prototypical or simplest form of sung language is probably the song. In fact, the degree to which the linguistic context is attended varies from song to song, performance to performance, and listener to listener. Though complex, listening to and singing songs are clearly among the most basic and universal of musical behaviors.

The effects of unilateral cerebral excisions (for the relief of epilepsy) on a song memory task have been investigated by Samson and Zatorre. Subjects first listened to novel melodies sung to lyrics and then were asked if they recognized the original tune with new lyrics, the original lyrics with new tunes, or the original songs. Patients with left temporal lobectomies were impaired in recognition of the lyrics, and both groups were impaired in recognizing the original tune with new lyrics. In another experiment, they presented the same melodies and lyrics separately. Right temporal lobectomy impaired melody recognition, but left temporal lobectomy did not. It appears that the integration of tune with lyrics, once formed, may not be disassembled later without cost.

B. Integrated Production of Music and Language

In the act of singing (with words) there is a smooth integration of the neural systems producing speech and singing. We know that in terms of activation, these systems seem to exhibit a significant degree of overlap but also some opposite hemispheric asymmetries. These asymmetries may reflect complementary specializations that arose specifically to serve the simultaneous demands inherent in producing speech (words with prosodic intonation) or song (melody with lyrics). Regardless of its evolutionary origins, song production, like song perception, calls upon a combination of functions, some of whose components appear to exhibit opposite hemisphere asymmetries.

Numerous cases have been reported in the neurological literature of patients who were profoundly aphasic but could still sing songs with normal intonation and sometimes with perfect lyrics. The examples of combined loss of speaking and singing suggest damage at a shared level of the vocal motor control system, whereas dissociations suggest that higher levels are distinct, possibly due to opposite hemispheric specializations.

Although no brain imaging studies have examined the activity associated with song production, Zatorre, Halpern, and Perry investigated the perception and mental imagery of songs (see Section III.B). As with other nonverbal imagery tasks involving melodies or musical intervals, SMA was active during song imagery (Fig. 5.III). Since both speech and singing also activate SMA, its activation by song imagery is clearly expected.

Temporal lobe areas were activated bilaterally during song imagery, consistent with the integrated nature of melody and lyrics in song representations. During actual singing, temporal lobe activation would be predominantly driven by the processing of auditory feedback. It remains to be demonstrated whether song singing will result in strongly right hemisphere asymmetric activation, as does simple singing without words, or in significant activation of left temporal lobe areas associated with language comprehension.

VIII. MUSIC AND NEURAL PLASTICITY

In previous sections we discussed changes in the organization and responses of neurons in response to both auditory sequence and motor sequence learning, particularly within tonotopically and somatotopically organizaed cortical areas. For such learning, the concomitant neural changes formed for any given sequence may be to some extent unique. Here, we discuss changes in the brain that are more general and result from long-term exposure to musical perception or production. Considerable interest has been generated in recent years regarding the possibility of more generalized benefits of both music listening and training in musical performance.

A. Plasticity in Response to Music Production

One approach to investigating how long-term experience in music performance may result in lasting changes in the brain involves measuring the size of particular structures from high-resolution MRI images. Differences between musicians and individuals who do not engage in music performance may be the direct results of musical experience, though one can usually not rule out the alternative explanation that these morphometric differences result from innate differences in ability.

Gottfried Schlaug and associates measured the midsaggital areas of the corpus callosum, the bundle of interhemispheric corticocortical fibers. The anterior half of the corpus callosum was significantly larger in musicians, with most of the effect coming from the subgroup of musicians who began training before age 7. Since the fibers connecting the sensorimotor cortices in the two hemispheres are contained in this portion of the corpus callosum, the increase in interhemispheric connections may be related to sensorimotor cortical development in response to musical practice, especially during the first decade of life.

They also measured the intrasulcal length of the posterior bank of the precentral gyrus as an index of primary motor cortex. They found a significantly longer sulcus in the right hemisphere of musicians in comparison to nonmusicians. Furthermore, both right and left sulcal length were positively correlated with the age at which music training commenced.

Finally, Schlaug and associates used MRI morphometry to investigate the specialized ability of absolute pitch perception. The surface area of the planum temporale, or the superior temporal plane posterior to Heschl's gyrus, showed an exaggeration of the normal leftward asymmetry in musicians with perfect pitch. Zatorre and colleagues measured the cortical volume of the same region and found that it was slightly larger on the left for absolute pitch possessors than for musicians without absolute pitch. Furthermore, pitch naming performance was correlated with size of the left planum temporale, such that larger volume was associated with lower error scores. Although studies suggest a critical period early in life for the development of absolute pitch, it is not clear if these asymmetries are innate or result from musical training. In the monkey the majority of projections from the superior temporal gyrus to the frontal area analogous to that activated in absolute pitch musicians while listening to musical tones (the posterior dorsolateral frontal cortex) originate in the planum temporale.

Pantev and associates used MEG to measure the responses of large ensembles of neurons in auditory cortex to tones. Yoshihiro Hirata, Shinya Kuriki, and Pantev contrasted such responses from a group of musicians with absolute pitch to a group of nonmusicians. The best fitting single equivalent current dipoles, representing the sources of these auditory-evoked magnetic responses, were located significantly more posteriorly within the left hemisphere for the absolute pitch possessors. These results are consistent with morphometric data indicating a larger planum temporale.

Pantev and associates also measured the magnetic fields evoked by piano tones and by pure tones matched for hearing level and fundamental frequency. The strength of the primary auditory evoked response (dipole moment of the M100) in the left hemisphere to piano tones was about 25% greater than that to matched pure tones for musicians but not significantly different for nonmusicians. Moreover, the strength of cortical activation correlated linearly with the age at which musicians began playing (Fig. 10) and did not vary with absolute pitch ability. It was maximal for musicians who began to play before the age of 9. Thus, it is plausible that this increase results from cortical reorganization induced by the sensory input coincident with musical production—reorganization that results in the recruitment of a larger ensemble of auditory cortical neurons and/or an increased synchrony of neural firing.

Approximately half of the musicians in the previous study were pianists and half were string or wind instrument players, although almost all reported the piano as at least a secondary instrument. Therefore, it is not clear whether reorganization may have been

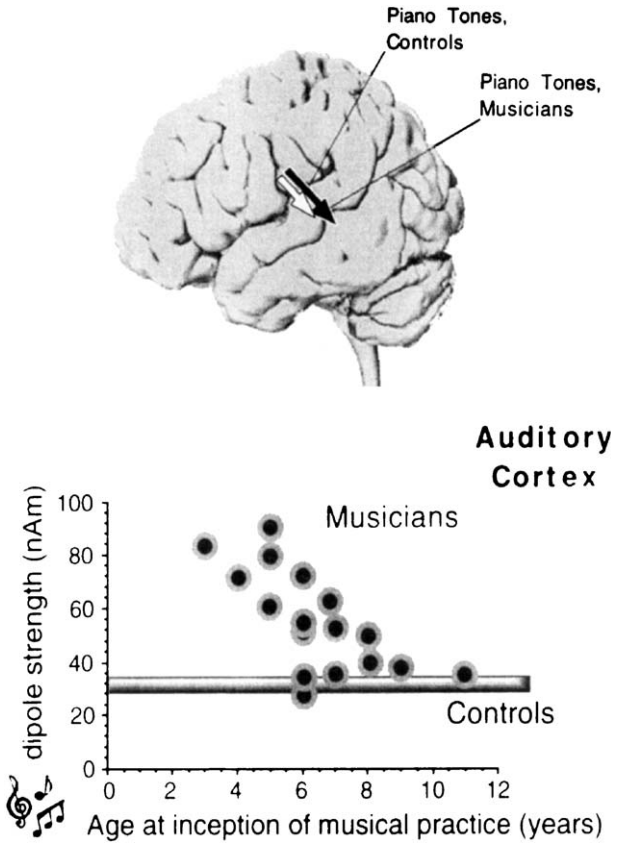


Figure 10 Enhanced auditory representation of piano tones in musicians. (Top) Mean values for the strength of the magnetic field evoked in response to piano tones (single equivalent dipole moment of the M100), plotted as arrow length, from musicians (black arrow) and nonmuscician controls (white arrow). (Bottom) ●, average dipole strength across all piano tones tested from each musician, plotted as a function of the age at which musical training began. The line indicates the average dipole strength across all piano tones and control subjects. Note that dipole strengths tend to increase with earlier training, but only for musicians who began playing before the age of 9 (reproduced with permission from Pantev *et al.*, 2001a; modified from Pantev *et al.*, 1998).

specific to the sounds produced by one's own instrument or to spectrally rich harmonic tones in general. To investigate the instrument specificity of this effect, Pantev and colleagues compared the auditory responses evoked by trumpet and violin notes (of the same fundamental frequency) in groups of trumpeters and violinists (none of whom played both instruments). The amplitude of the M100 was greater for a musician's own instrument and was greater overall for the trumpeters (Fig. 11). Although these results do not rule out generalized enhancement of neuronal responses for spectrally rich harmonic tones among musicians, they do indicate at least an additional enhancement in response to one's own instrument. The instrument-specific response enhancement may result from long-term exposure to feedback from one's own instrumental performance.

Plasticity in somatosensory cortex is also suggested by a study carried out by Thomas Elbert, Pantev, and colleagues, who used MEG to measure and compare the cortical representation of the digits of string players to those of nonmusicians. They found that

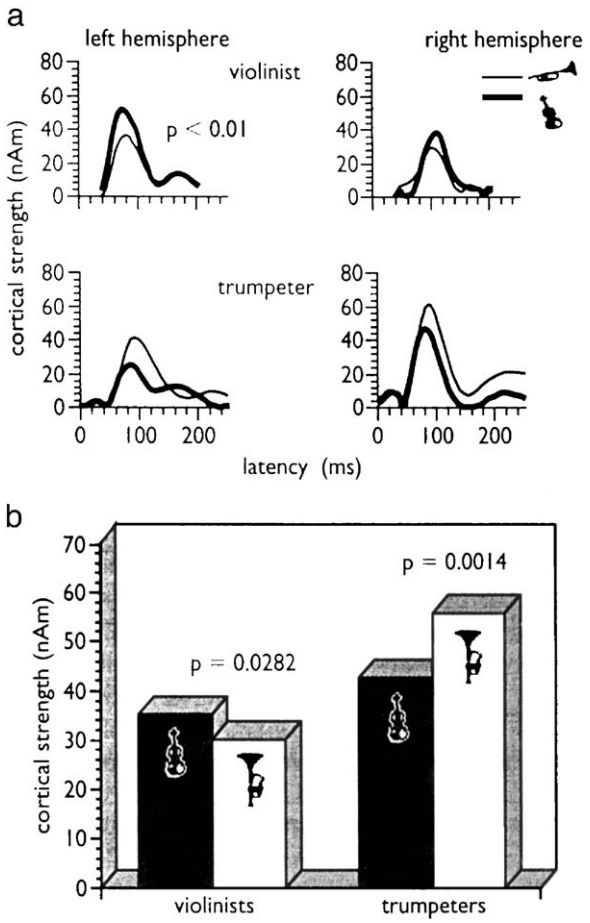


Figure 11 Enhanced auditory representation of tones played on one's own instrument (violin or trumpet). (a) Time courses of the strength of the magnetic field (as dipole moment) by hemisphere from one violinist and one trumpeter, evoked in response to trumpet tones (fine lines) and violin tones (heavy lines). Note that peak responses at ~100 msec (M100) are greater for one's own instrument. (b) Average dipole strengths across both hemispheres for all violinists and all trumpeters in response to violin tones (black bars) and trumpet tones (white bars). The interaction of stimulus type with instrument group was significant, and response strength was reliably greater for the timbre of the instrument of training within both groups (reproduced with permission from Pantev *et al.*, 2001b).

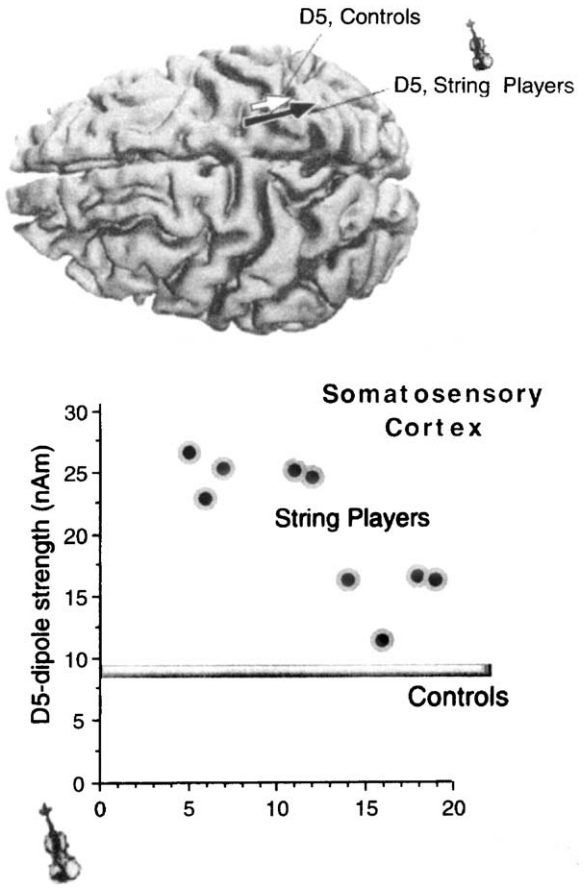


Figure 12 Enhanced somatosensory representation of left-hand digits in string players. (Top) Equivalent current dipoles (ECDs) elicited by stimulation of the little finger (D5) superimposed onto an MRI reconstruction of the cortical surface from one control subject. The arrows indicate the location, orientation, and moment (length) of the modeled dipoles averaged across sting players (black arrow) and nonmusician controls (white arrow) (Bottom) ●, magnitude of the dipole moment of D5 somatosensory-evoked responses from each string player, plotted as a function of the age at which musical training began. The line is the mean of the same response across all nonmusician subjects. Note that dipole strengths are particularly enhanced for those who began training at ≤ 12 years (reproduced with permission from Pantev *et al.*, 2001a; modified from Elbert *et al.*, 1995).

the cortical representation of the digits of the left hand was greater for string players (smallest for the thumb) but equivalent for the right hand. The magnitude of the effect correlated negatively with the age at which training began (Fig. 12). These results suggest cortical reorganization in response to somatosensory feedback from one's own instrumental performance since violinists typically use the four digits of the left hand to manipulate the strings. The results are consistent with direct mapping of somatosensory cortex in monkeys before and after tactile discrimination training. Greg Recanzone, Michael Merzenich, and colleagues found that the area of somatosensory cortex representing the portion of the skin of a trained digit increased up to threefold.

Focal hand dystonia, the loss of motor control of one or more digits, is a serious occupational hazard for those engaged in rapid, finely coordinated hand movements such as pianists. Robert Schumann may have suffered from focal hand dystonia beginning in his twenties. Its cause is controversial, but some theories suggest it is caused by maladaptive cortical plasticity, particularly an overlap of receptive fields within somatosensory cortex. Using MEG, Elbert, Pantev, and colleagues found that the single equivalent current dipoles representing somatosensory-evoked responses to individual digits were closer together for the affected hand of dystonic musicians than for nonmusician controls. These results are consistent with the hypothesis of maladaptive, overlapping somatosensory receptive fields.

Jesus Pujol, Pascual-Leone, and colleagues used functional MRI to measure brain activity while guitarists played a simple finger exercise (arpeggios) on a computer-interfaced guitar. When not experiencing dystonia, the pattern of motor cortical activation within contralateral sensorimotor cortex and SMA was the same for dystonic and nondystonic musicians. When dystonic symptoms were provoked by guitar playing, activity in sensorimotor cortex became elevated and motor cortex activity declined.

Further brain imaging studies during the active expression of dystonic symptoms will help to reconcile findings of disordered and hyperactive somatosensory functioning associated with focal hand dystonia, and it is hoped that they will lead to successful treatments for this all-too-common affliction.

Finally, we consider the most generalized claims of changes in the brain as a result of exposure to music production. M. Gardiner studied the effects of music training during childhood on development of other cognitive abilities. He found that children who participated in an intensive, active music training program (Kodaly method) also showed large improvements in math and reading. However, a causal link has not been established, nor have the mechanisms by which musical training might result in changes that could enhance other cognitive functions been determined. Continuing developmental research may help to provide new insights into very old controversies concerning the place of music within general education.

B. Plasticity in Response to Music Perception

In the previous section, I discussed changes that may occur in auditory cortex as a result of long-term exposure to auditory feedback during music production. Generalized changes within the brain may also result simply from exposure to music perception. The duration of exposure varies significantly across studies. In an influential study, Greg Recanzone, Christoph Schreiner, and Michael Merzenich trained monkeys to make pitch discriminations within particular frequency ranges over several weeks. After training, a frequency map of the primary auditory region was derived from multi-unit recordings. The cortical representation of the trained frequency band and its area correlated positively with performance.

Hans Menning, Larry Roberts, and Pantev carried out a related experiment in humans using MEG. Volunteers were trained to make increasingly fine pitch discriminations referenced to 1000-Hz pure tones over a period of 3 weeks. The M100 in response to the frequencies trained increased in amplitude during training. After 3 weeks without training, it decreased. Increased amplitude of the M100 could indicate that more neurons are firing, and it is consistent with the increased area of cortical representation seen in the analogous monkey experiments.

Pantev and colleagues asked volunteers to listen to music from which a narrow band of frequencies centered at 1 kHz had been removed by a "notch" filter. Magnetic fields evoked by a bandpassed noise corresponding to the notch filter and a similar control stimulus centered at 0.5 kHz were recorded before and after 3 hr of listening on 3 consecutive days. The amplitude of the response to the 1-kHz centered stimulus decreased after listening to the notched music,

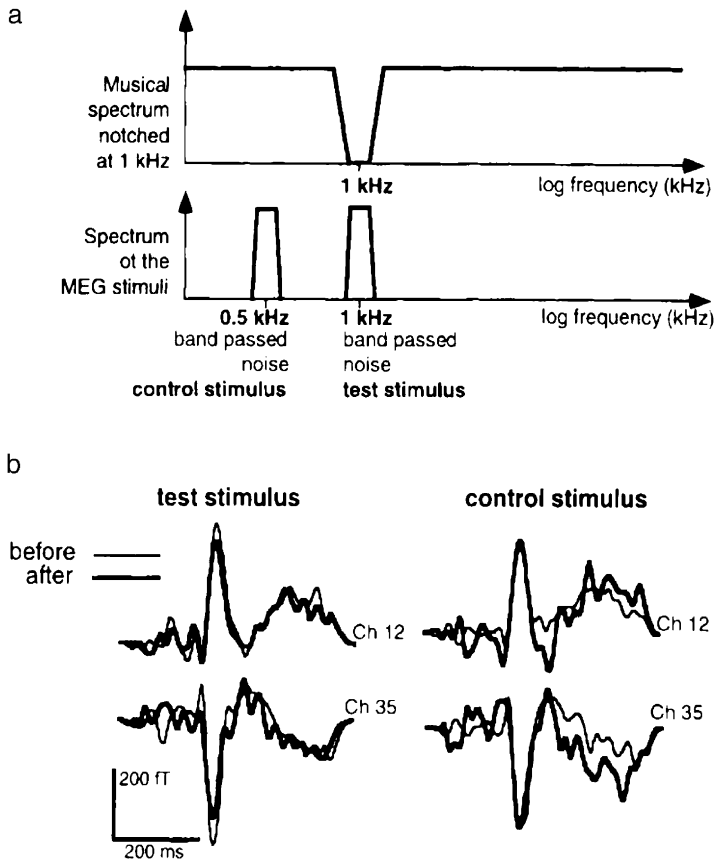


Figure 13 Short-term reduction of neural response to a specific frequency range after listening to "notched" music. (a) Spectrum of the notch filter centered at 1 kHz that was applied to self-selected music. Spectrum of the two stimuli, band-passed noise bursts centered at 1 and 0.5 kHz, used to measure the evoked magnetic response. (b) Auditory evoked fields from one representative subject in response to the two bandpassed noise bursts immediately before and after 3 hr of attentive listening to notched music on 3 consecutive days. Note that response magnitude is diminished afterwards only for the test stimulus (1 kHz), corresponding to the removed frequencies. This effect was not present 24 hr later (i.e., at the beginning of the next session), suggesting that auditory cortical neurons reverted to their initial frequency tuning (reproduced with permission from Pantev *et al.*, 2001a, modified from Pantev *et al.*, 1999).

but the response to the control stimulus was unaffected (Fig. 13). The 1-kHz response rebounded by the beginning of the next session. These findings point to the existence of more rapid and transient changes that may take place in response to changes in one's acoustic environment.

Finally, I consider claims regarding more generalized effects of music perception on the brain. In particular, much attention has been drawn recently to the so-called "Mozart" effect, or modest gains in cognitive test performance following 10 min of listening to a Mozart sonata (primarily in tests requiring visual–spatial manipulation). First, the effect is not specific to Mozart, and it has been reported by Glenn Schellenberg in response to Schubert, or even listening to a passage from a story. For a given individual it appears to be maximal for the preferred stimulus. Its persistence across time has not been demonstrated.

The Mozart effect appeals to the popular imagination in that it appears to promise a passive route to cognitive enhancement (much as the passive exercise machines of the 1950s did for physical improvement). Though considerable controversy still exists, perhaps the moral is that effort and time are required before long-lasting changes can be induced in the brain. Regardless, human beings will continue to make and listen to music, primarily for noncognitive reasons and without any goals for self-improvement.

As this article makes clear, music, even when only perceived, engages and activates the human brain at many levels: auditory, cognitive, emotional, and, when working memory or imagery are involved, motoric. Across a lifetime of exposure, musical experience, both active and passive, has the potential to leave lasting traces in our brains, not only through the formation of memories for specific musical sounds and gestures (a musical "lexicon") but also in more general ways.

Like language, music is an accumulated product of humankind that is continually being passed on and added to. Questions regarding how (and why) the human brain produces and responds to music will continue to occupy researchers and philosophers for many years to come and touch upon many of the brain's most complex and integrative abilities.

See Also the Following Articles

AUDITORY CORTEX • AUDITORY PERCEPTION • CREATIVITY • HEARING • LANGUAGE AND LEXICAL PROCESSING • NEUROPLASTICITY, DEVELOPMENTAL • SEMANTIC MEMORY

Suggested Reading

Blood, A. J., Zatorre, R. J., Bermudez, P., and Evans, A. C. (1999). Emotional responses to pleasant and unpleasant music correlate with activity in paralimbic brain regions. *Nature Neurosci.* **2,** 382–387.

Blood, A. J., and Zatorre, R. J. (2001). Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion. *Proc. Natl. Acad. Sci. USA* **98,** 11818–11823.

Elbert, T., Pantev, C., Wienbruch, C., Rockstroh, B., and Taub, E. (1995). Increased cortical representation of the fingers of the left hand in string players. *Science* **270,** 305–307.

Marin, O. S. M., and Perry, D. W. (1999). Neurological aspects of music perception and performance. In *Psychology of Music* (D. Deutsch, Ed.), 2nd ed., pp. 653–724. Academic Press, New York.

Pantev, C., Oostenveld, R., Engelien, A., Ross, B., Roberts, L. E., and Hoke, M. (1998). Increased auditory cortical representation in musicians. *Nature* **392,** 811–814.

Pantev, C., Wollbrink, A., Roberts, L. E., Engelien, A., and Lütkenhöner, B. (1999). Short-term plasticity of the human auditory cortex. *Brain Res.* **842,** 192–199.

Pantev, C. A., Engelien, A., Candia, V., and Elbert, T. (2001a). Representational cortex in musicians: Plastic alterations in response to music practice. In *The Biological Foundations of Music* (R. J. Zatorre and I. Peretz, Eds.). *Annals New York Acad. Sci.* **930,** 300–314.

Pantev, C., Roberts, L. E., Schulz, M., Engelien, A., and Ross, B. (2001b). Timbre-specific enhancement of auditory cortical representations in musicians. *NeuroReport* **12,** 169–174.

Pascual-Leone, A., Dang, N., Cohen, L. G., Brasil-Neto, J. P., Cammarota, A., and Hallett, M. (1995). Modulation of muscle responses evoked by transcranial magnetic stimulation during the acquisition of new fine motor skills. *J. Neurophysiol.* **74,** 1037–1045.

Peretz, I. (2000). Musical perception and recognition. In *The Handbook of Cognitive Neuropsychology* (B. Rapp, Ed.). Psychology Press, Philadelphia.

Perry, D. W., Zatorre, R. J., Petrides, M., Alivisatos, B., Meyer, E., and Evans, A. C. (1999). Localization of cerebral activity during simple singing. *NeuroReport* **10,** 3453–3458.

Weinberger, N. (1999). Music and the auditory system. In *The Psychology of Music* (D. Deutsch, Ed.). Academic Press, San Diego.

Zatorre, R. J., and Binder, J. R. (2000). Functional and structural imaging of the human auditory system. In *Brain Mapping: The Systems* (A. Toga and J. Mazziotta, Eds.), pp. 365–402. Academic Press, San Diego.

Zatorre, R. J., and Peretz, I. (Eds.) (2001). The biological foundations of music. *Ann. N. Y. Acad. Sci.* **930**.

Zatorre, R. J., Evans, A. C., and Mayer, E. (1994). Neural mechanisms underlying melodic perception and memory for pitch. *J. Neurosci.* **14,** 1908–1919.

Zatorre, R. J., Halpern, A., Perry, D. W., Meyer, E., and Evans, A. C. (1996). Hearing in the mind's ear: A PET investigation of musical imagery and perception. *J. Cognitive Neurosci.* **8,** 29–46.

Zatorre, R. J., Perry, D. W., Beckett, C. A., Westbury, C. F., and Evans, A. C. (1998). Functional anatomy of musical processing in listeners with absolute and relative pitch. *Proc. Natl. Acad. Sci. USA* **95,** 3172–3177.

Nausea and Vomiting

ROBERT J. NAYLOR
University of Bradford

GLOSSARY

area prostrema Located on the ventral surface of the brain stem and contains chemoreceptors; synonymous with the chemoreceptor trigger zone (CTZ).

emetic stimuli Arise from the stimulation of mechanoreceptors or chemoreceptors in the gastrointestinal tract, other body tissues, and the brain or from pain, psychogenic stimulation, or vestibular–visual conflict.

nausea A distinctive and unpleasant feeling of malaise and drowsiness, accompanied by autonomic changes.

nucleus tractus solitarius Lies close to the area postrema and receives emetic stimuli and autonomic connections; a crucial structure in the emetic reflex.

retching and vomiting (or emesis) Rhythmic contractions of the abdominal muscles and contraction of the diaphragm compressing the stomach to expel its contents through the mouth.

vestibular nuclei Receive information concerning body and head movements.

vomiting center A useful concept to explain the still undetermined coordination of the emetic reflex.

The act of vomiting has been considered one of the most primitive functions in animals and humans. There are pictorial representations of vomiting from Roman times, with the first recorded experimental work associating changes in stomach function with emesis being performed in the seventeenth century. By the late nineteenth century, a ''vomiting center'' in the brain had been hypothesized, the importance of the vagus nerve was established, and certain hindbrain areas had been shown to be more relevant than others to mediate drug-induced emesis. Little further progress was made until the 1950s when Borison and colleagues, using brain lesion and electrical stimulation techniques, substantiated and clarified earlier findings to demonstrate a functional and anatomical distinction between a ''chemoreceptor trigger zone'' and a ''vomiting center.'' This was again followed by little serious medical or research interest: nausea and vomiting were rarely a medical problem, more a difficulty for the patient, family, or nursing staff. Yet the introduction of aggressive and severely emetogenic chemotherapy in the 1980s stimulated medical concern to discover effective antiemetic treatments for the cancer patient. This was realized with the discovery of the antiemetic action of the 5-HT_3 receptor antagonists. These advances complemented the continuing efforts of the military and other researchers to investigate the mechanisms involved in the sickness of motion and space travel and the problems inherent in their treatment. This article indicates how an understanding of the stimuli, pathways, and structures within the brain and gut has begun to allow control of the emetic reflex and the therapeutic successes that have revolutionized treatments and caused a renaissance in emesis research.

I. THE PHYSIOLOGY OF NAUSEA AND VOMITING

Nausea is a highly subjective and peculiarly unpleasant sensation, quite distinct from other feelings. It is ''felt''

in the throat or stomach or as a ''sinking'' sensation in the epigastrium; the symptoms may include malaise, drowsiness, anxiety, and reduced vigilance. It is invariably accompanied by autonomic changes including vasomotor disturbances, causing vasoconstriction–pallor and pupil dilation, tachycardia, salivation and sweating, and a relaxation of the lower part of esophagus and abdominal muscles. The latter tends to increase tension of the gastric and esophageal muscles, which may directly contribute to the sensation of nausea. In the acute stage it is perceived as an unpleasant and temporary effect that precludes other mental and physical activity. It frequently precedes vomiting (although either may occur alone) and usually is relieved by an emetic episode, which the subject or patient may welcome. The presence of chronic nausea severely reduces the quality of life.

Following the relaxation of the lower part of the esophagus and abdominal muscles, a contraction of the upper small intestine occurs closely followed by contraction of the pyloric sphincter and the pyloric portion of the stomach. These changes will empty the contents of the upper jejunum, doudenum, and pyloric portion of the stomach into the fundus and body of the stomach, which are relaxed. Also relaxed are the cardiac sphincter, esophagus, and esophageal sphincter. In this manner the system is prepared for retching and vomiting, which are reflex in origin and serve to remove the contents of the upper gastrointestinal tract.

This involves a series of highly coordinated changes in gastrointestinal motility, respiratory movements, and posture. Emesis, which is initiated by a deep and sharp inspiration, is immediately followed by reflex closure of the glottis and raising of the soft palate, which prevent the passage of vomitus into the lungs and nasal cavity. The abdominal muscles then contract in the rhythmic manner of ''retching'' movements, which compress the stomach between the contracted diaphragm and abdominal organs. The inevitable increase in intragastric pressure causes evacuation of the stomach contents through the relaxed esophagus; definite antiperistalsis in the stomach itself is rarely observed. These events and pressure changes are shown in Fig. 1. This profile of activities can vary somewhat between species and also in the human infant. In the latter, the abdominal muscles or diaphragm apparently does not play a role in, for example, the regurgitation of an oversized meal; the reverse peristalsis occurs by contraction of the stomach muscle alone.

The feelings of nausea followed by retching or vomiting are the expression of a precisely controlled reflex, albeit of an extraordinarily complex nature. Almost irrespective of the causes of nausea and emesis, the autonomic and somatic patterns of motor activities are the same. Also, because the presence of the emetic reflex occurs in so many species, it would have been predicted that the actual function of nausea and

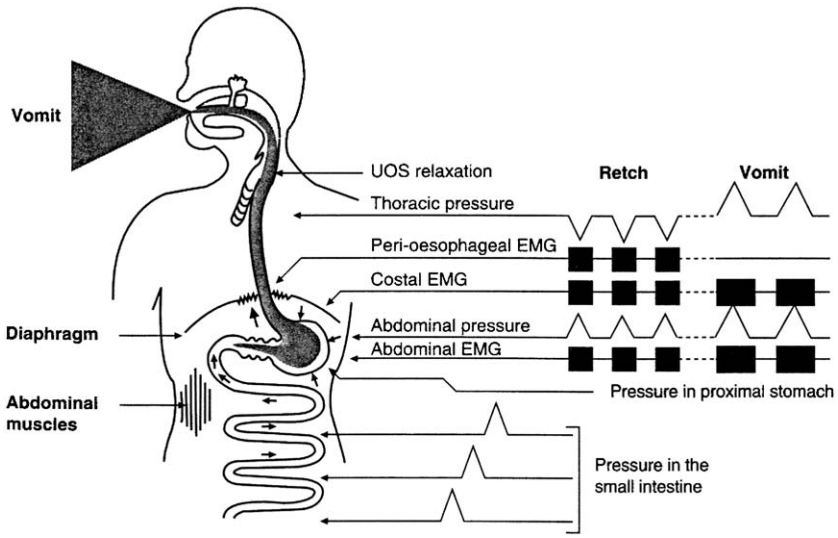


Figure 1 The major mechanomotor components of retching and vomiting. Abbreviations: EMG, electromyogram; UOS, upper esophageal sphincter. Andrews, P. L. R., and Davis, C. J. (1995). The physiology of emesis induced by anticancer therapy. In *Serotonin and the Scientific Basis of Antiemetic Therapy* (D. J. M. Reynolds, P. L. R. Andrews, and C. J. Davis, Eds.), pp. 25–49. Oxford Clinical Communications, Oxford. Reproduced by permission of Oxford Clinical Communications.

vomiting should be well-established. But this has not occurred. The physiological "value" of nausea and vomiting is usually described as a "defense mechanism" to protect the host against a toxic challenge. Thus, the activation of smell or taste receptors causing nausea could warn the organism of a noxious or poisonous substance in order to avoid its ingestion. Similarly, if toxins are actually ingested, the chemoreceptors in the stomach would warn of the presence of danger at an early stage and emesis would promptly expel the threat. However, such explanations remain unconvincing.

First, there are numerous toxins that do not cause nausea or emesis and simply kill the host; the defense mechanism is inadequate at best. More important, the defense theory fails to provide any reasonable explanation for the induction of nausea or emesis caused by nontoxins and, therefore, all other emetic stimuli, e.g., changes in blood pressure, irritation of the throat, pain, psychogenic stimuli, increased intracranial pressure, and motion. It is concluded that we have no satisfactory explanation for the phylogenetic persistence of the emetic reflex: its value appears extraordinarily limited. The absence of an emetic reflex in rats, mice, and other species confirms its nonessential role.

II. THE STIMULI GIVING RISE TO NAUSEA AND VOMITING

A classification of the stimuli giving rise to nausea or emesis is the essential precursor to an understanding of the mechanisms and systems mediating nausea and emesis. However, the reductionist approach has had an inevitable effect of "individualizing" stimuli to a particular "sickness." For example, the stimulus of motion sickness has generally been restricted to motion, travel, or space sickness and its potential importance to medically related conditions has been generally ignored. This limitation should be carefully considered in the delineation of the stimuli causing nausea and emesis.

A. Stimuli Arising from the Alimentary Canal

For many years it was the opinion of physiologists that the autonomic nervous system was composed mainly of motor (efferent) fibers. However, more recent observations in both animals and human, based on morphological and electrophysiological techniques, have clearly shown that sensory fibers are much more numerous than motor fibers in the physiological regulation of gut homeostasis, visceral motility, and alimentery behavior. Their normal function is related to events accompanying digestion. They signal events related to the mechanical state of the digestive tract (i.e., mechanoreceptors measuring contraction or distension) or the physicochemical properties of the gut contents (i.e., chemoreceptors measuring chemical composition, concentration of nutrients, osmotic pressure, acid and alkaline sensitivity, general chemosensitivity). Other polymodal receptors may respond to a broad range of stimuli, and digestive nerve endings appear to exhibit a general sensitivity to neurotransmitters, hormones, and other agents. A disturbance of this control can trigger changes in peristaltic contractions, the coordination of gastrointestinal activity, gastric emptying, and food intake.

These receptors can be stimulated by a diverse range of foreign stimuli of varying intensity. For example, simple overeating or serious obstruction can both cause distension and induce nausea and vomiting by mechanoreceptor stimulation. Similarly, the oral administration of a moderately innocuous solution of sodium chloride or a severe gut infection with staphylotoxin enterotoxin will both stimulate chemoreceptors to induce nausea and vomiting (see Fig. 2). Stimulation of receptors within the alimentary canal is the single most important cause of the induction of nausea and emesis in the normal population (Fig. 3). The vagus nerve is particularly important in sensory viscerotopic organization. Its afferent components contain general somatic, special visceral (taste), and general visceral sensory fibers, which carry both mechano– and chemoreceptive information from the gut to terminate in the nucleus tractus solitarius. The splanchnic nerves also relay sensory information from the gut. The efferent components of the vagus nerve contain both general and special visceral motor fibers arising in the dorsal motor nucleus of the vagus (DMV) and the nucleus ambiguous, respectively. These medullary systems comprise a final common pathway from the brain to control the activity of the gastrointestinal tract.

B. Stimuli Arising from the Area Postrema–Related Areas

It was established nearly 100 years ago that the application of chemicals to the dorsal surface of the brainstem or the ventricular system could induce emesis. In these early studies it was assumed that this

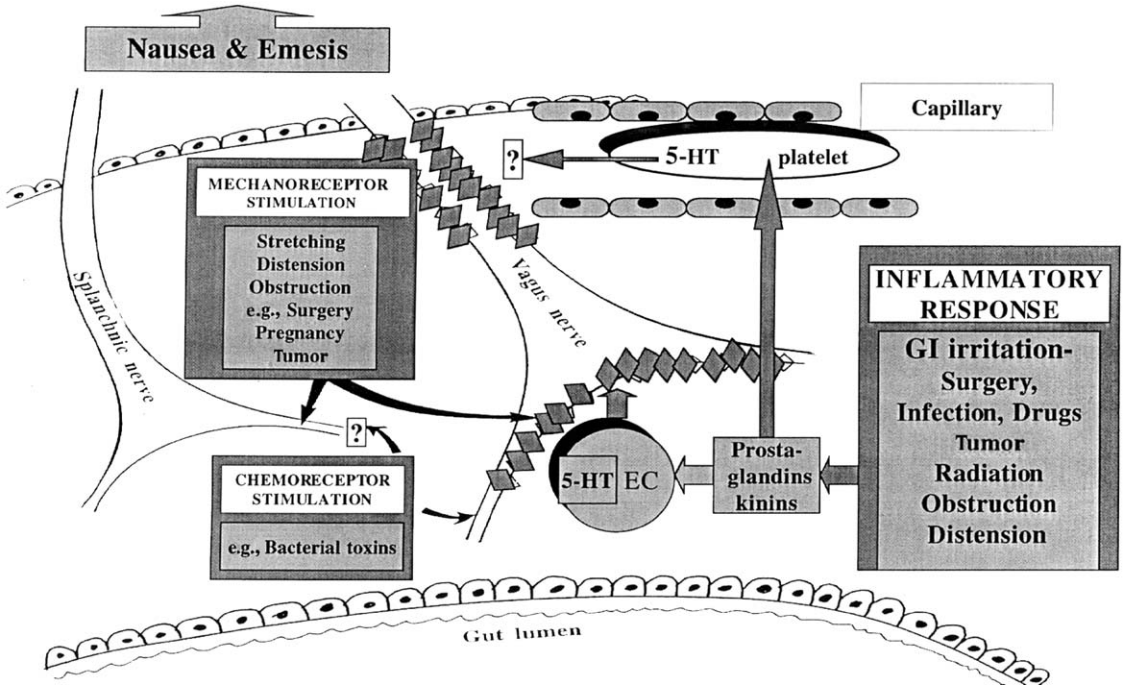


Figure 2 The stimuli causing chemo- and mechanoreceptors for stimulation of the vagus and splanchnic nerves in the gut wall to induce nausea and emesis. An inflammatory response may damage cellular elements leading to the release of prostaglandins and kinins which may stimulate the enterochromaoffin cells (EC) or possibly platelets. This will cause the release of serotonin, which may activate 5-HT_3 receptors located on the vagus nerve to trigger the emetic reflex.

occurred through a "vomiting center" in the brain stem. Again, it was the studies of Borison and colleagues that showed that a number of stimuli were actually detected by cells in the caudal part of the floor of the fourth ventricle, the area postrema, which was termed a "chemoreceptor trigger zone" for emesis. The area postrema is a circumventricular organ of the brain outside the blood–brain barrier, with the capillaries within the area lacking endothelial occluding junctions. Also, the ventricular ependymal cells lack occluding junctions, making the structure permeable to molecules in the blood or ventricles to provide an efficient chemoreceptor sensor for numerous chemicals and endogenous substances (Fig. 4).

C. Chemicals as Emetogenic Stimuli

Very few drugs or chemicals are administered to humans to purposefully induce emesis. However, chemoreceptor stimulation caused by concentrated saline solutions taken orally have a nauseous taste and may be given in an attempt to cause emesis and remove toxic substances from the stomach. (The insertion of a finger down the throat has a similar effect, irritating the mechanoreceptors in the pharynx.) The alkaloid ipecacuanha has well-known emetic effects and acts by stimulating chemoreceptors in the area postrema. It has also been used in subemetic doses in some paracetamol tablets to prevent self-poisoning; a person taking an excessive number of tablets would finally receive an emetic dose of ipecacuanha, which would expel the paracetamol from the stomach. A different use was found for disulfiram in the treatment of alcoholism. If taken with alcohol, it leads to an accumulation of acetaldehyde, which induces nausea and sickness. It was used as an aversion therapy for drug abuse; apomorphine was used as an alternative treatment. With these exceptions, drug-induced emesis is unwanted but occurs with many different types of medical treatments. It can vary from a transient effect of little consequence to a persistent and chronic effect to severely reduce the quality of life.

The nausea and emesis induced by cancer chemotherapy (and radiation) are correctly perceived as the most severe emetogenic stimulus induced by any drug treatment. However, it is emphasized that the potential emetogenicity of chemotherapy is rated as absent, low,

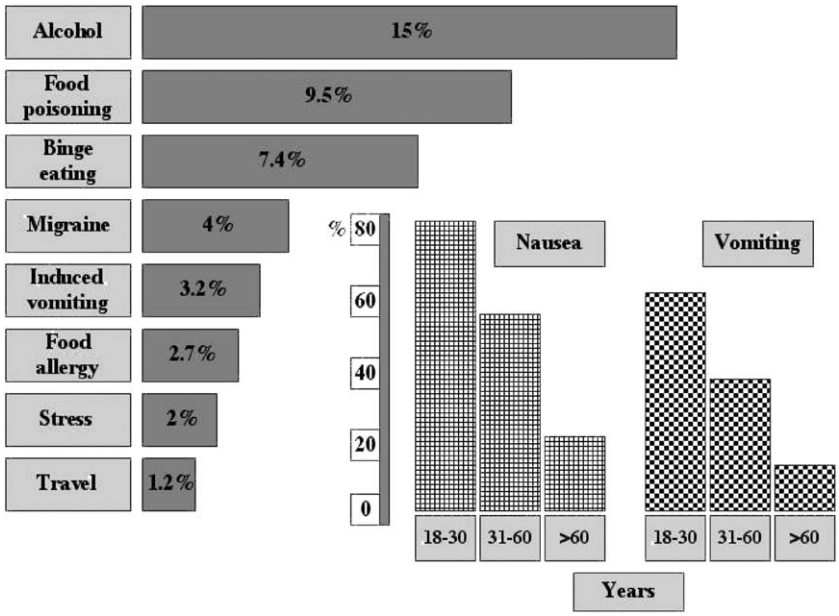


Figure 3 The causes and incidence of nausea and vomiting from an otherwise healthy population in the United Kingdom. The percentage of respondents in each group who reported nausea and vomiting at least once during the previous 12-month period.

moderate, or severe, depending on the drug used, regime of drug administration, dose, and patient variables. For example, all patients receiving cisplatin would be predicted to be very sick, whereas methotrexate causes little if any effect. Nausea and vomiting induced by radiation are also related to the dose used and to the area and extent of the body irradiated. Both the cytotoxic therapy and radiation treatment frequently cause tissue disruption to the gastrointestinal tract. This damage may cause the release of toxic products from the tissue destruction and may induce a local inflammatory response influencing vagal afferent nerve endings within the gut to trigger the emetic reflex. This will depolarize the vagus nerve, and this chemically induced response mimics the effects of electrical stimulation of the vagus nerve, which is known to cause vomiting. The release of 5-hydroxytryptamine from damaged enterochromaffin cells within the gut provides just one important example whereby an endogenous substance may stimulate the 5-HT_3 receptor subtype located on the vagal nerve (see Fig. 2). Additionally, substances such as 5-HT or other agents released as a consequence of cell damage may be transported in the blood system (with cytotoxic agents themselves and also the products of tumor breakdown) to directly stimulate the central components mediating the emetic reflex within the area postrema and perhaps the nucleus tractus solitarius.

A second group of therapeutic agents that inevitably cause nausea as the dose is increased is the dopamine agonist anti-Parkinson drugs, apomorphine, L-dopa, and the ergot derivatives. A major effect is to directly stimulate the dopamine receptors located in the central chemoreceptor mechanisms, with a lesser action to induce gastric stasis.

A third group of drugs that invariably induce nausea or emesis is the opiate analgesics. The acute administration of morphine and related drugs to opioid-naive patients frequently induces nausea and sometimes vomiting. However, tolerance rapidly develops to such effects, and indeed the first treatment may antagonize the emetic effects to a second opioid injection or other emetogens. The emetic effect may be mediated in the chemoreceptor trigger zone, whereas the antiemetic effect may be mediated closer to the vomiting center. The antiemetic effect may relate to an endogenous inhibitory tone exerted by opioids from the enkephalin or related series. The ability of narcotic antagonists such as naloxone to precipitate nausea or vomiting would support this hypothesis.

A fourth group of drugs is those enhancing 5-HT function, for example, 5-hydroxytryptophan, the precursor of 5-HT, or 5-HT re-uptake inhibitors, e.g., fluoxetine and paroxetine, which have been reported to induce nausea and occasionally vomiting as a side

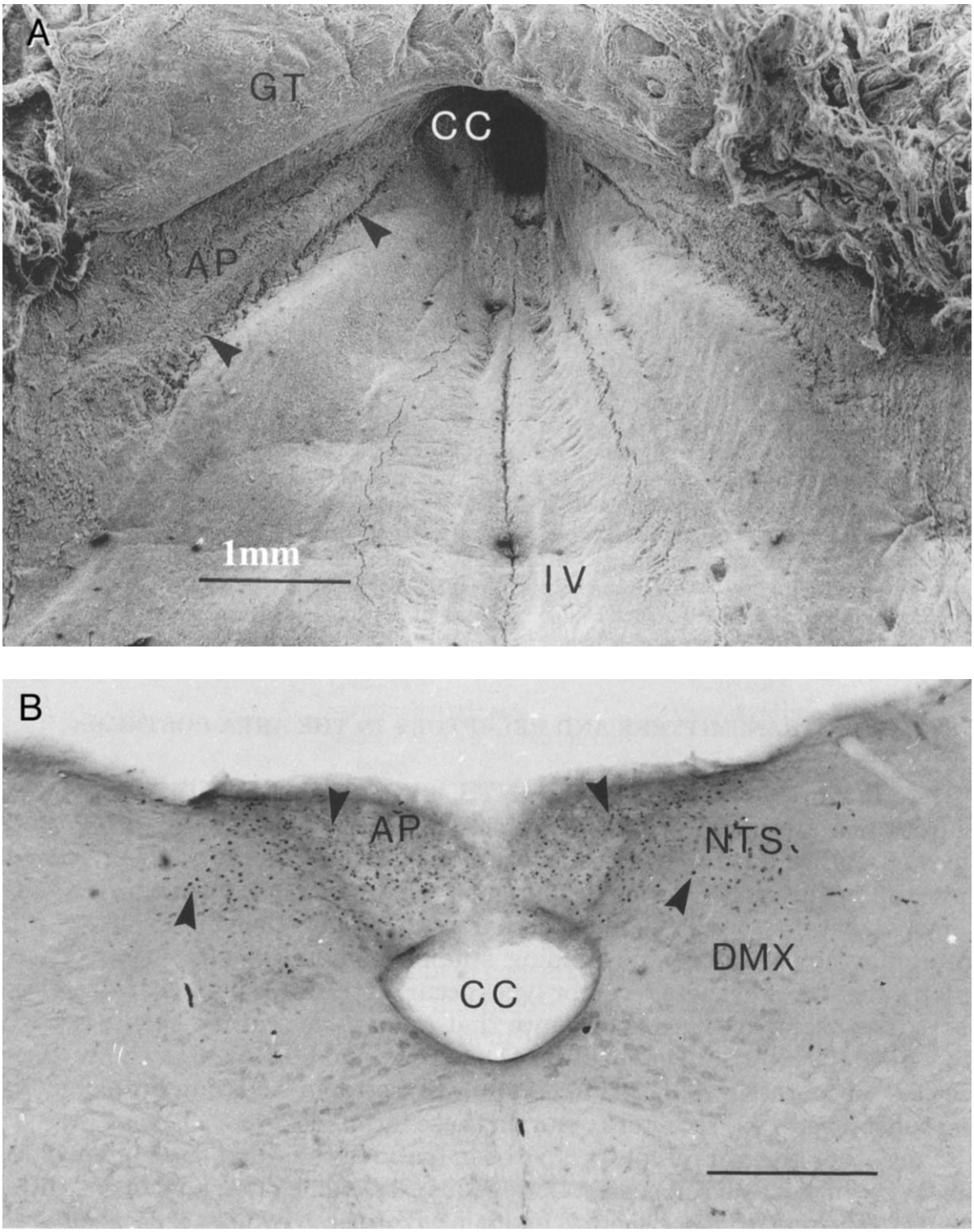


Figure 4 (A) A scanning electron micrograph showing the human area postrema (AP) on the caudal floor of the fourth ventricle (IV) looking toward the central canal (CC; GT, gracile tubercle). The arrowheads indicate the abrupt transition between the area postrema and underlying cells. (B) Section through the medulla oblongata of a ferret that had been given cisplatin to reveal an increase in cellular activity using Fos-like immunoreactivity. This appears as dark dots (indicated by the arrowheads) is visible throughout the area postrema, and extends into the nucleus tractus solitarius (NTS; DMX, dorsal motor nucleus of the vagus nerve).

effect. This may relate to an increased 5-HT activity in both central and gut tissues.

Numerous other drugs also induce nausea and emesis. For example, cardiac glycosides such as digoxin, at doses only slightly greater than the therapeutic dose, induced abdominal pains, nausea, and vomiting. This probably relates to a central action at the chemoreceptor trigger zone and an irritant action within the gastrointestinal tract, which may be exacerbated by cardiac dysrhythmia. Also, the gastrointestinal irritation caused by antibiotic treatment or nonsteroidal anti-inflammatory agents frequently causes gastrointestinal distress, nausea, or vomiting, triggering the emetic reflex via the vagal and splanchnic nerves.

D. Stimuli Arising from the Vestibular Labyrinth and Visual System

All animals that move must remain oriented to their surroundings, and for most movements this requires maintenance of balance and posture, measurement of the changes of bodily position, and the adjustment of eye position when the head is moving. This is also important for the maintenance of homeostasis upon a change of posture; compensation must be rapid to maximize respiratory requirements and direct blood flow to the areas of need. However, unlike most sensations, orientation, movement sensation, and associated autonomic changes never enter the awareness; we are not conscious of this level of control. However, unusual disturbing conditions of orientation or movement, of either the body or environment, exist that certainly enter the conscious domain, which can induce reliable and intense feelings of dizziness, vertigo, nausea, and vomiting. This occurs when the orienting response that has been configured to normal terrestrial movement for each species has been subject to an alien stimulation. In humans, the major orienting detectors are the vestibular organs and the eyes.

1. Vestibular Organs

These comprise the saccule, utricle, and semicircular canals, which in humans are designed to measure changes in gravity, linear movements, and rotary acceleration. Hair cells present in the fluid-filled membranous sacs of the saccule and utricle are bent by the free moving otoliths (calcium carbonate crystals) depending on the extent and direction of linear displacement, i.e., linear acceleration or changes in the rate of motion. The saccule and utricle are positioned to respond to vertical and horizontal movements, respectively; both respond to gravity. The three semicircular canals are fluid-filled and widen at the base to form the ampulla, which contains sensory hair cells fixed to the cupula and the base called the crista. Each of the three canals lies at approximately right angles to each other and to a major plane of the body; the cupula is bent by the pressure of fluid movement in the canal to stimulate the hair cells, which relays the nature of rotary acceleration to the head.

The orienting system has evolved in each species to meet the movements that are normally required. In humans, this is usually active movements and brief passive (imposed) movements and measures best acceleration or deceleration rather than constant velocity. For example, on a plane or train having attained constant speed with the constant passive movement of fluid, otoliths and receptors soon move at the same speed and there is no relative movement between the various components of the vestibular mechanism and, therefore, no stimulation.

2. Vestibular Stimulation and Occular Function

In response to rotation of the head, the stimulation of the semicircular canals transmits information to many body systems, including the occular muscles, which creates reflexive eye movement. The eyes move slowly opposite, the movement of rotation and then quickly return. The reflex that coordinates the eye and head positions is known as vestibular nystagmus, which facilitates the perception of a stable visual environment. In other words, the visual image of a fixed object remains stabilized on the retina.

3. Incidence and Susceptibility to Disorienting Stimuli

All individuals or animals with a normal vestibular system are susceptible to the sickness induced by an unusual movement or visual scene, and this is best known with respect to motion sickness. It is observed in numerous species, e.g., cows, sheep, dogs, birds, and monkeys, in addition to humans. It is not reliably demonstrated in children less than 2 years old; susceptibility is highest between 3 and the early teenage years and usually decreases thereafter. Parents who suffer from motion sickness report a greater incidence of susceptible children. Women are also generally believed to be more susceptible than men.

People may feel ill or nauseous and 1 or 2% may actually vomit during normal travelling by boat, coach, car, train, or airplane, which can all precipitate motion sickness. The problem has been studied intensively for pilots with increasingly serious implications for private license, commercial airline, and military pilots. The problem can be most intense and disabling during space travel, with some 75% of all astronauts being affected for 3–4 days. The occurrence of drowsiness (the Sopite syndrome) can also be significant.

A number of tests have been developed to assess individual susceptibility to motion sickness and to assess the mechanisms involved. One of the most common is the Coriolis Sickness Susceptibility Index, in which subjects are required to make a series of head movement in the pitch (sagittal) and roll (frontal) planes while undergoing body rotation in the yaw (horizontal) plane.

4. The Sensory Conflict Theory

Disorientation-induced sickness is generally considered to result from a sensory conflict between the data presented from the vestibular and visual systems against those that are expected on the basis of past experience. For example, reading a book in a moving car or train can be provocative; the constant visual image is at variance with the movement recorded by the vestibular system. In contrast, looking out of the car or train window with a fixed horizon presents a moving image that is in harmony with the input from the vestibular nuclei.

The importance of visual perception is shown dramatically by the sensory conflict imposed by the presence of a changing visual image but static vestibular input. For example, a subject sits quietly on a chair with no movement but is made to watch motion pictures shot from a moving vehicle or is placed inside an optokinetic drum or rotating room. In brief, the subject is stationary while the environment is moving. An alternative would be for the subject to operate a vehicle training device (a stimulator sickness) equipped with a moving visual display. In both cases, the ''visual sickness'' created by the visual record of movement in the absence of body movement creates a conflict sufficient to induce sickness. A most impressive demonstration of the power of visual input to influence vestibular function is shown by postural sway experiments. In these experiments, a stationary, standing observer is placed in a special room whose walls and ceiling can move backward and forward. The subject, who is unaware of the swaying of the room, begins to sway in synchrony with the swaying environment. The postural (vestibular) adjustment occurs only in response to the changing visual scene. These examples provide convincing evidence that visually induced sensations of motion appear equivalent to those produced by actual body motion.

E. Psychogenic Stimuli

Fear or disgust at the sight, smell, or taste of something intensely disagreeable can precipitate nausea or even vomiting. The stimuli arising in the visual, olfactory, or gustatory receptors are finally relayed to the cortical and limbic systems, which interpret and give meaning to the experience. In addition, anxiety itself is well-known to predispose one to nausea and vomiting. Also, higher functions in the brain can learn or be conditioned to respond with emesis to specific triggers. This involuntary or Pavlovian response, seen in cancer patients whose previous nausea and emesis to cytotoxic treatments had been poorly controlled, can be precipitated by even the sight or smell of the hospital where or the medical or nursing staff from whom they had received treatment.

F. Pain

The specific stimulus of intense and acute pain, caused by distension of intestinal, bile, or urethral ducts, or severe somatic or cardiac pain can induce nausea and vomiting.

III. THE CENTRAL AND PERIPHERAL SYSTEMS MEDIATING NAUSEA AND EMESIS

The stimuli influencing the emetic reflex provide important clues as to the brain circuitry mediating nausea and vomiting. Concepts based on results obtained mainly from animals using central and peripheral nerve lesions, from intracerebral injection of drugs into and electrophysiological stimulation from discrete brain regions, and more recently from human brain imaging have begun to establish key structures and pathways. Pharmacology, in discovering new antiemetic treatments, is establishing the relevance of specific neurotransmitters within these pathways.

A. The Area Postrema: The "Chemoreceptor Trigger Zone"

An important structure is the "chemoreceptor trigger zone" (CTZ) in the area postrema, which has numerous afferent and efferent connections with underlying structures, the subnucleus gelatinosus and nucleus tractus solitarius, and the vagus nerve.

B. The Nucleus of the Solitary Tract

The nucleus of the solitary tract consists of medial and lateral divisions, which are innervated in a highly selective manner by the terminals of the vagal sensory neurons; this provides the basis for a central sensory viscerotopy. The solitary nucleus is a complex integration center subserving many functions and also receives information from the parvocellular vasopressin containing neurons in the hypothalamus; increased levels of vasopressin are associated with nausea. Further, the amygdala innervates the solitary nucleus and has been implicated in the integration of autonomic and somatic components of emotional behavior. Its influence on cardiovascular and respiratory functions, together with cortical input, may contribute to the nausea and emesis induced by psychogenic stimuli. The aortic and carotid sinus afferents also relay chemo– and baroreceptor information to the solitary nucleus, and hypotension is a known factor to induce nausea and emesis (Fig. 5).

The nucleus tractus solitarius may also serve a chemoreceptor function. For example, the injection of the ∂_2 adrenoceptor agonist clonidine into the solitary complex induces bradycardia and hypotension; the reduction in blood pressure in humans may be related to the actions of clonidine to induce an intensity of nausea and dizziness that will cause at least 10% of patients to discontinue the treatment.

It is possible that psychogenic stimuli mediated via the amygdala on the solitary nucleus may modify cardiovascular and respiratory functions to reduce blood pressure; this may then trigger a lowering of the threshold for nausea and vomiting. It is possible that projections from the area postrema to the solitary nucleus may also interact with the visceral and other afferent terminals to modulate interoceptive information relevant to nausea and vomiting.

The solitary tract is a substantial structure and located close to the area postrema and vestibular nucleus. The solitary nucleus sends a major projection to the medullary reticular formation, has direct projections to the limbic and other forebrain regions,

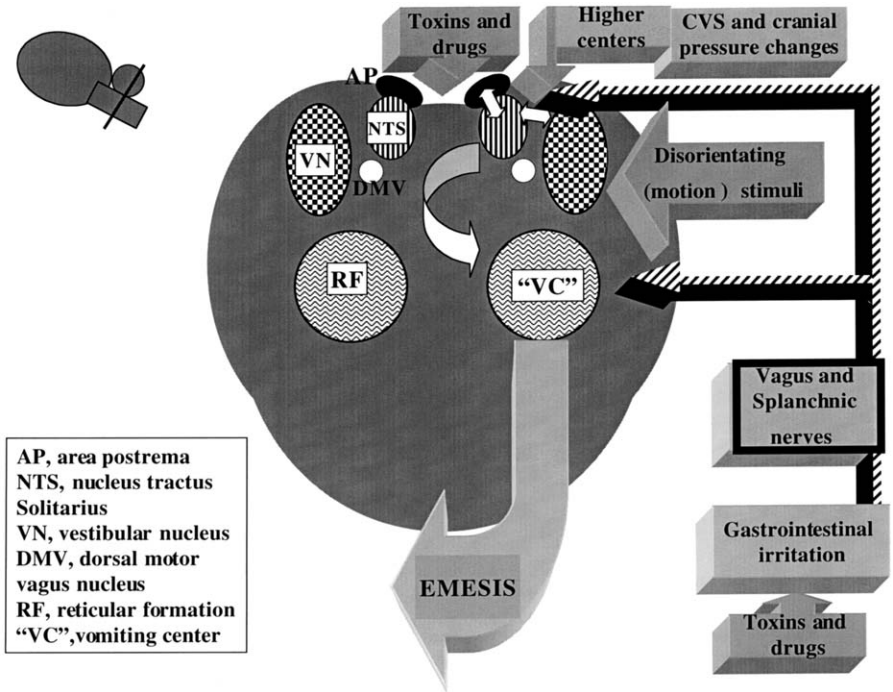


Figure 5 The major medullary structures and stimuli that contribute to the induction of emesis. CVS, cardiovascular.

and has indirect connections to the cortex. This may allow the interoceptive information generated in the solitary nucleus to influence both higher and lower brain centers relevant to nausea and emesis.

C. The Vestibular System

The vestibular pathways that begin in the hair cells of the christae and maculae have their cells bodies in the vestibular ganglion and their axon terminals in the vestibular nuclei of the medulla and cerebellum. It is well-established that impulses are then relayed to a number of sites, including cranial nerves III, IV, and VI to control head movements with eye movements and motor neurons to control posture and balance.

However, more recently, like the vestibulo–ocular and vestibular spinal circuits, pathways have been established that appear to integrate vestibular and autonomic information. These may be relevant to the autonomic manifestations of vestibular dysfunction, gastrointestinal discomfort, nausea, and vomiting and complement the cerebellar complexes that influence the somatic motor system. The pathways are shown in Fig. 6. Both physiological and anatomical evidence indicates a convergence of descending vestibular and visceral information on neurons in the solitary tract and the rostroventral medullary reticular formation. An ascending projection to the parabrachial nucleus may integrate vestibular and autonomic influences via limbic (affective), cortical (cognitive), and hypothalamic (neuroendocrine) pathways. This may provide a substrate for the interactions between affective (e.g., anxiety) and cognitive (e.g., learned behavior) factors to predispose one to motion sickness and psychiatric disorders. The model provides a useful working hypothesis to further establish the autonomic changes that can be shown to occur following vestibular stimulation.

D. The "Vomiting Center"

The vomiting center was an area originally discovered by electrically stimulating brain stem medullary structures to induce emesis. However, it is now more

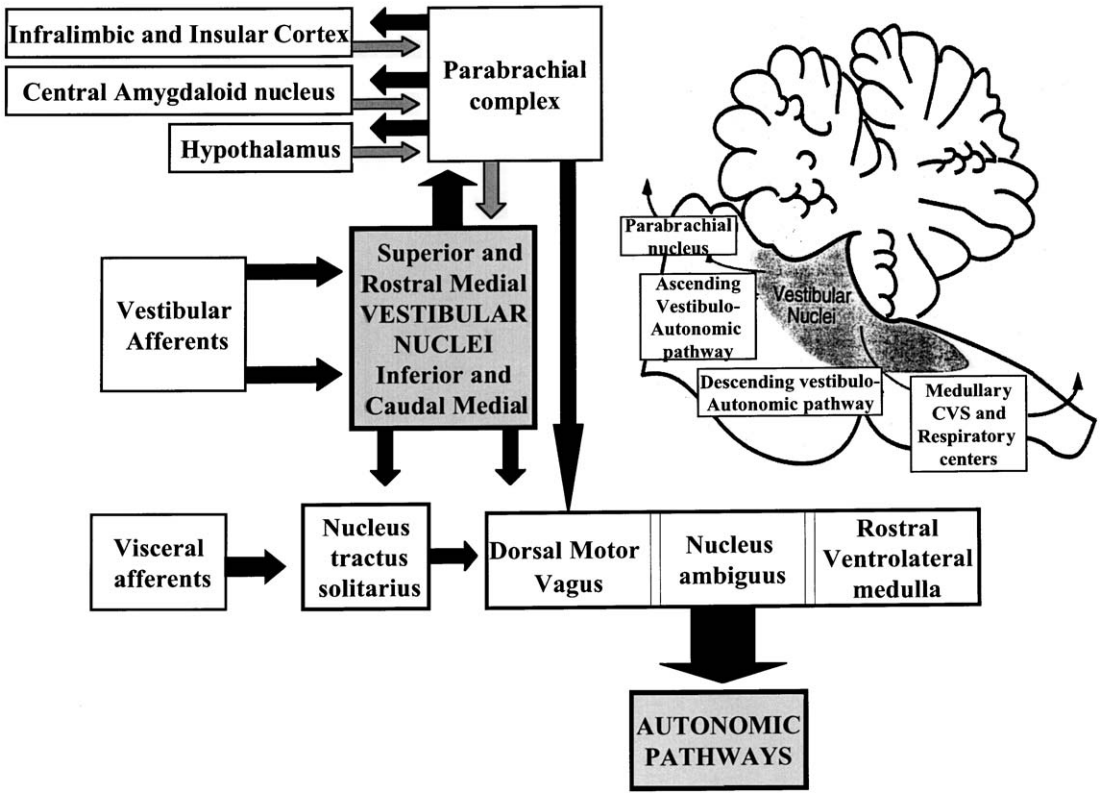


Figure 6 Diagram of the vestibuloautonomic pathways and the interactions with higher brain function relevant to emesis. Adapted from Babalan and Porter (1998) and Furman *et al.*(1998). Clinical evidence that the vestibular system participates in autonomic control. *J. Vestibular Res.* **8**(1), 27–34. Reproduced by permission of Elsevier Science.

accurately described as a collection of effector nuclei rather than a discrete brain area. The vomiting center receives major inputs from the chemoreceptor trigger zone, the nucleus tractus solitarius, the vestibular nucleus, a vagal and sympathetic input from the gut, and input from the cardiovascular system and a number of limbic brain nuclei, e.g., the olfactory tubercle, amygdala, hypothalamus, and ventral thalamic nucleus. Electrical stimulation of all of these structures can induce emesis. The latter nuclei may be involved in olfactory, emotional–anticipatory, hormonal–stress, and pain–induced vomiting, respectively.

E. Brain Mechanisms and Nausea

Finally, the brain regions that produce the sensation of nausea have been difficult to assess; nausea is a subjective human experience, and reliable animal models are not available. However, noninvasive magnetic source imaging in humans has revealed that neuronal activation occurs in the cortex in the inferior frontal gyrus in volunteers made nauseous with ipecacuanha or vestibular stimulation: the activation caused by ipecacuanha (but not vestibular) was antagonized by ondansetron. This brain area may be important in the perception or sensation of nausea (Fig. 7).

IV. THE PATHOLOGY OF NAUSEA AND VOMITING

A. Gastrointestinal Irritation

A summary of the major stimuli of the chemoreceptors and mechanoreceptors in the gut have been provided in Fig. 2. Gastrointestinal trauma following the accidental or deliberate injestion of toxic materials such as alcohol is the single most common cause of nausea and emesis in Certain chemotherapeutic agents such as cisplatin or radiation used in the treatment of

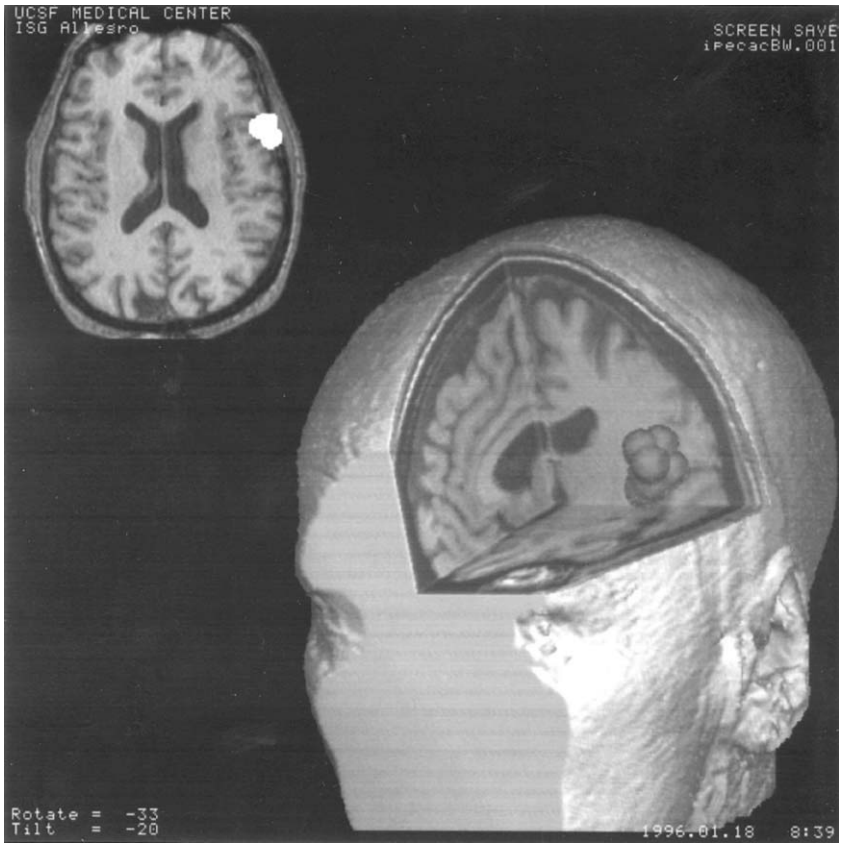


Figure 7 Location of current dipoles, indicative of neuronal activation, in the inferior frontal gyrus in a subject made nauseous by the oral ingestion of ipecacuahna syrup. The dipoles reflect magnetic source imaging of only one side of the head during a recording session.

cancer cause considerable damage to gut tissue, including the enterochromaffin cells. These contain 5-HT and it is hypothesized that damage causes the release of 5-HT to stimulate 5-HT_3 receptor located on the vagus afferent nerve fibers. This will depolarize the vagus nerve and trigger the emetic reflex.

B. Vestibular Pathology

The importance of the labyrinth to motion sickness is shown in labyrinthectomized monkeys and labyrinthine-defective human subjects; they are immune to motion sickness and probably visually induced sickness. Interestingly, chronic labyrinthectomy in dogs also impairs their ability to vomit in response certain poisons. However, in the presence of a defective vestibular function as is found in Ménières disease, the patient experiences sudden episodes of violent vestibular activity causing nystagmic eye movements, extreme vertigo, severe nausea, and distortions of hearing.

C. Intracranial Pathology

A sudden bout of vomiting associated with a severe headache may be the consequence of raised intracranial pressure caused by intracranial hemorrhage or a severe inflammatory response. A severe headache is also frequently observed following lumber puncture, which alters pressure within the cerebral ventricular system. This may be due to a direct influence on the central structures coordinating the vomiting reflex, with additional input from nociceptors present in the vascular smooth muscle. Migraine provides a less dramatic pathology and is the most common neurological disorder, presenting as the classical unilateral pulsating headache with gastrointestinal effects and nausea and vomiting. There is a little understood link between the susceptibility to migraine attacks and motion sickness.

D. Cyclic Vomiting Syndrome

The cyclic vomiting syndrome is an uncommon and unexplained disorder in children (and some adults), with an onset usually between 3 and 7 years and persisting for months to decades, with episodes occurring several times per month or year. It is characterized by recurrent attacks of severe and relentless nausea, vomiting (5–10 times an hour at peak effect) with prostration for hours to days; gastrointestinal disorders and intense abdominal pain, diarrhea, and fever also occur. The dehydration and electrolyte imbalance cause intense thirst and may be life-threatening. The child generally is well between attacks. Frequently, there is a familial history of migraine.

E. Hyperemesis Gravidarum

Nausea occurs in over 80% and vomiting in over 50% of pregnant women during the first trimester. The "morning sickness," which can actually occur at any time, then declines. Only a small number, possibly 1 in 1000, show the continued vomiting of hyperemesis gravidarum. The cause is not known but may involve changes in hormonal function, intraabdominal pressure, gastric emptying, metabolism, and pyschogenic influences.

F. Metabolic Disorders

A wide variety of metabolic disorders may induce nausea and vomiting, e.g., hypoglycemia due to over dosage with insulin and uremia following kidney damage.

G. Psychogenic Nausea and Retching or Vomiting

A raised level of anxiety predisposes one to nausea and emesis:

1. Chronic "psychogenic vomiting" (or usually retching) usually occurs upon getting up or just after breakfast; it may persist for years.
2. "Nervous dyspepsia" is used to describe symptoms of a feeling of satiety, abdominal discomfort, nausea, and vomiting associated with psychoneurotic features such as anxiety, irritability, and loss of weight.
3. Anorexia nervosa and bulimia have a well-established psychopathology in which retching–vomiting is a symptom of serious psychiatric illness.

H. Post-Operative Nausea and Vomiting (PONV)

General anesthetics were first administered in 1846 and were immediately associated with a very high incidence of postoperative nausea and vomiting. After 100 years the introduction of halothane caused a fall in the

incidence of PONV, but this remained at some 25–30%. Whereas PONV is generally is self-limiting and rarely is of medical or surgical importance, for many patients it remains the most distressing part of the operation.

PONV is multifactorial and may be triggered by any or all of the following factors:

1. Inhaled agents are variably associated with PONV, nitrous oxide being noteworthy. However, intravenous anesthetics and spinal anesthesia are also potentially emetogenic.
2. Stimulation of the oropharynx.
3. The nature of the surgery, e.g., gynecological and pediatric strabismus surgery have a high incidence of nausea and vomiting. Also, abdominal surgery may cause stretching, distention, or tissue damage, i.e., gastrointestinal irritation.
4. Pain resulting from surgery or disease.
5. Hypoxia, hypotension, and carbon dioxide retention.
6. Clumsy movement of the patient in the recovery room or ward or following day-case surgery, causing labyrinthine disturbance.
7. The nature of any pre- or postoperative drug treatments, e.g., opioid analgesics.
8. Psychogenic factors such as anxiety.
9. An increase in body weight increases the prevalence of PONV.
10. Adult females have three times more risk than adult males of developing PONV, and children are twice as susceptible as adults.

V. NEUROCHEMICAL COMMUNICATION WITHIN THE EMETIC REFLEX

An understanding of the neurotransmitters within the pathways of the emetic reflex is essential to the design of rational pharmacological antiemetic treatments. Neurochemical and immunocytochemical analyses of medullary nuclei have revealed a wealth of neuroactive substances that may contribute to a neurotransmitter, neurohumoral, and neuromodulatory role: e.g., serotonin, noradrenaline, γ-aminobutyric acid, acetylcholine, substance P, enkephalin, oxytocin, vasopressin, neurotensin, and thyrotropin releasing hormone. The identification of receptor binding sites within the medullary nuclei has further identified the potential loci of drug action. However, it is not yet possible to characterize the precise neurotransmitter pathways that exist between the nuclei of the emetic reflex. Also, the presence of a neurotransmitter or its receptors provides no guidance as to its functional relevance. This can only be undertaken by local and discrete injections of transmitters and receptor agonists and antagonists into the medullary nuclei. Such studies, relevant to emesis, remain to be performed.

VI. THE PREVENTION OR TREATMENT OF NAUSEA AND VOMITING

A. Overview

Nausea and vomiting are symptoms of a cause that may require investigation and treatment. A transient bout of emesis induced by a single psychogenic stimulus is unpredictable, by definition untreatable, and only of interest to the sufferer. Similarly, a brief bout of nausea and vomiting during travelling is essentially a social inconvenience of little consequence to anyone other than the afflicted and fellow passengers. However, when motion sickness interferes with the awareness, emotional, and cognitive competence of a professional or military pilot, the consequences are profound and will require careful investigation. The devastating nausea and vomiting induced by highly emetogenic chemotherapy or the relentless nausea and vomiting observed in the cyclical vomiting syndrome are of serious medical concern and require treatment. Finally, retching, vomiting, or regurgitation occasionally constitutes a medical or surgical emergency, which requires immediate treatment when it is induced

1. by major intracranial pathology or intestinal obstruction.
2. by in infants, where fluid loss may cause dehydration.
3. by postoperatively where the force of retching or vomiting may tear esophageal tissue or damage delicate surgical repair.
4. by in an emergency, when the patient has recently had a meal and therefore has a high risk of developing aspiration pneumonia.
5. by hyperemesis gravidarum.

It remains clear that, to a patient, either a brief or a persistent period of nausea and vomiting is always of concern. Persistent nausea will incapacitate the patient, and persistent vomiting may result in the loss of fluid and hydrochloric acid, leading to dehydration, alkalosis, reduced food intake, malnutrition, and serious debilitation, severely reducing the quality of life.

B. The Preferred Treatment of Nausea and Vomiting Is Removal of the Cause

Persistent nausea or vomiting may be indicative of gastrointestinal, neurologic, or metabolic disorders that require direct treatment, and it may be desirable to withhold antiemetic therapy until a diagnosis has been made. However, it remains clear that the treatment of a cardiovascular pathology such as migraine with sumatriptan will relieve the neurologic manifestations, headache, gastrointestinal effects, nausea, and vomiting. Yet sumatriptan has no direct effect on the emetic reflex. However, for most causes of nausea and emesis, e.g., motion sickness, cancer chemotherapy, and antiparkinsonian medication, removal of the cause is not an option. Also, there is a marked variation between individuals in their response to nausea–emetic stimuli, although previous sensitivity to a nausea–emetic stimulus is indicative of future sensitivity. For example, patients who have had nausea and vomiting following previous surgery are likely to experience it again. Also, women who experience pregnancy sickness are much more likely to develop nausea and vomiting in response to any hormonal disturbance (e.g., the contraceptive pill) or travelling and with a migraine. People who have an emetic response to a first drug challenge are likely to show a similar response to a subsequent challenge, i.e., an individual tends to have a consistent response. A simple enquiry therefore may identify people at greater risk, i.e., those who have a lower emetic threshold. These people are particularly likely to develop nausea and vomiting under emotional strain as occasioned by illness.

It should also be considered that many presentations of nausea and vomiting may be induced by a combination of different stimuli. This has remained the single most unexplored territory in the treatment of nausea and emesis.

There are now at least five major groups of compounds with defined pharmacological profiles to control nausea and emesis. These include the following:

- Dopamine receptor antagonists
- Muscarinic receptor antagonists
- Histamine H_1 receptor antagonists
- 5-HT_3 receptor antagonists
- NK_1 receptor antagonists
- Sedatives and hypnotics
- Phenothiazines

Procedures and treatments that alleviate retching and vomiting generally prevent nausea.

1. Dopamine Receptor Antagonists

Apomorphine induces intense nausea and vomiting and has a high affinity for the dopamine receptor. Dopamine receptors are found in high concentrations in the area postrema, the dorsal motor nucleus of the vagus nerve, and the nucleus tractus solitarius. The traditional view has been that apomorphine, levodopa (via dopamine), lergotrile, bromocriptine, and other dopamine agonists used in the treatment of Parkinson's disease induce nausea and vomiting by stimulating dopamine receptors in the CTZ. Highly potent dopamine receptor antagonists such as haloperidol block such receptors and thereby prevent nausea and emesis. However, the use of these drugs is limited in two ways:

- First, they generally do not inhibit nausea and emesis induced by stimuli other than dopamine agonists (although droperidol is used in PONV).
- Second, they have major adverse effects of motor impairment, severe akinesia and muscle rigidity, and dystonias (muscle spasm) caused by striatal dopamine receptor blockade, particularly in young people.

The adverse effects are less with the dopamine receptor antagonists domperidone and sulpiride because they are less able to penetrate the blood–brain barrier in the striatal areas but can easily access structures such as the area postrema, which lacks a blood–brain barrier.

Metroclopramide is also a dopamine receptor antagonist that has been widely used as an antinauseant–antiemetic for gastrointestinal disorders and migraines and, in much higher doses, for cancer chemotherapy and radiation-induced sickness. Its unusual and established ability to facilitate gastric emptying and intestinal activity may contribute directly to its antinauseant–antiemetic actions in gastrointestinal disorders and migraines. However, such actions reflect not a dopamine receptor blockade but more probably an agonist action at the 5-HT_4 receptor. Furthermore, its ability to prevent chemotherapy–radiation-induced emesis, when used in exceptionally high doses, is more readily attributed to its low-potency 5-HT_3 receptor antagonism.

2. Muscarinic Receptor Antagonists

Scopalamine is the most effective remedy for motion sickness of all types, although nausea and vomiting induced by extreme changes in motion and space travel

remain an intractable problem and scopalamine fails to control all types of movement sickness (e.g., due to severe vestibular disturbance). Scopalamine has a greater central depressant effect than atropine, and its antiemetic action is attributed to a blockade of muscarinic cholinergic receptors in the area postrema or associated nuclei of the dorsal vagal complex. Scopalamine is also a potent inhibitor of gastrointestinal movements and relaxes the gastrointestinal tract. These effects may make a modest contribution to the antiemetic action.

Adverse effects of scopalamine include sedation and predictable autonomic effects with increasing dose (e.g., blurred vision, urinary retention, decreased salivation). The sedation and blurred vision preclude its use in airline pilots, train, bus, and car drivers, and those operating machinery.

3. Histamine H_1 Receptor Antagonists

Antihistamines used for motion sickness include buclizine, cyclizine, dimenhydrinate, meclizine, and promethazine. Their antiemetic actions are attributed to a central blockade of H_1 receptors in the area postrema and possibly underlying structures. However, most antihistamines are also potent muscarinic receptor antagonists, and this behavior may make an important contribution to their antiemetic actions. Indeed, chlorpheniramine, which is an H_1 receptor antagonist that fails to block centrally mediated cholinergic effects, does not inhibit motion sickness. Whatever the contribution of an H_1 and ACh receptor blockade to the antiemetic potential, a sedative potential may also contribute to the effect, although this is less or absent for cinnarizine. However, some antihistamines that cause sedation are not antiemetic. The H_1 receptor antagonists may play some role in the treatment of PONV and pregnancy sickness.

4. 5-HT_3 Receptor Antagonists

Ondansetron, granisetron, and tropisetron were introduced 10 years ago as antiemetic agents. Their ability to antagonize chemotherapy- and radiation-induced emesis was first established in animal models in 1986. Their efficacy relates to the blockade of 5-HT_3 receptors, which are located in high density in the area postrema and nucleus tractus solitarius and on vagal afferent nerve endings in the gut.

It is hypothesized that severely emetogenic regimens such as cisplatin cause gastrointestinal tissue disruption, which initiates the release of 5-HT from the enterochromaffin cells within the mucosa to trigger vagus nerve firing to initiate the emetic reflex. The 5-HT_3 receptors within the area postrema–nucleus tractus solitarius lie on the vagus nerve terminals, because they disappear if the vagus nerve is cut.

Ondansetron and the other 5-HT_3 receptor antagonists control nausea, retching, and vomiting in patients with cancer receiving emetogenic treatments, especially during the acute phase (i.e., day 1 of treatment): the symptoms are completely controlled in 70% of patients and there is reduced vomiting in the others. However, on the second and subsequent days (i.e., a delayed phase) the antimetic effects are less pronounced and the addition of a glucocorticosteroid is needed to produce a maximally efficacious antiemetic regimen. Dexamethasone or methylprednisolone is now used frequently in combination with ondansetron or other 5-HT_3 receptor antagonists to secure optimal control, even on the first day of treatment. This is vitally important for the patient with cancer, because the development of nausea and vomiting during treatment can lead to learned anticipatory nausea and vomiting. Such anticipatory nausea–vomiting is resistant to all drug treatments, including the 5-HT_3 receptor antagonists.

The 5-HT_3 receptor antagonists have no effect on motion sickness or apomorphine-induced vomiting. They will, however, block the nausea and vomiting induced by ipecacuanha, which has been used as a tool in animals and humans to determine the dose regimens of 5-HT_3 receptor antagonists that will be required to block nausea and emesis in the cancer patient. Ondansetron has also been shown to antagonize the nausea induced by morphine and the nausea and adverse gastrointestinal effects of SSRIs. The lack of interaction between 5-HT_3 receptor antagonists and other drugs also prompted the use of ondansetron in the treatment of PONV, for which it is as efficacious as any of the existing remedies (e.g., cyclizine, droperidol) but does not produce their adverse effects. Its use in the treatment of other causes of gastrointestinal irritation and pregnancy sickness is being investigated.

5. Sedatives and Hypnotics

Vomiting is blocked by anesthesia and lesser degrees of CNS depression, which may attenuate a low emetic threshold that normally would trigger intractable nausea and vomiting. Barbiturates and chloral hydrate have been used, but benzodiazepines are a more contemporary treatment and have the advantage of causing amnesia to prevent recall of a highly

distressing event. However, caution is needed as the potential for vomiting or aspiration in the presence of reduced consciousness is a serious hazzard.

6. Phenothiazines

Agents such as chlorpromazine, promethazine, prochlorperazine, and trimeprazine have a mixed pharmacology as antagonists at muscarinic, histaminergic, dopaminergic, adrenergic, and serotonergic receptors. Nausea and vomiting that fail to respond to specific pharmacologic treatments may sometimes respond to this multireceptor blockade (e.g., pregnancy sickness, PONV). Because the mechanisms mediating most types of nausea and vomiting are not clear but there may be many different contributory stimuli, the use of a mixed pharmacologic approach is logical. This may account for the extensive use of trimeprazine in palliative care.

7. NK_1 Receptor Antagonists

Antagonists acting at the substance P NK_1 receptor subtype have the most interesting antiemetic spectrum of action recorded so far. In many animal species they have consistent actions to inhibit emesis induced by many different drugs and also motion sickness. They appear to be broad spectrum antiemetic agents, where the breadth of antiemetic activities may be revealing of an inhibitory effect on the final pathways of the emetic reflex, although this remains to be proven. The only action against which they have not yet been tested is psychogenic induced or anticipatory emesis.

In humans the NK_1 receptor antagonists have been tested in the cancer patient and show considerable potential to antagonie chemotherapy-induced emesis and nausea, although this remains to be proven. In any event, these agents undoubtedly will receive intensive clinical testing to clearly establish their full potential and limitations.

8. Ancillary Treatments of Nausea and Vomiting

Movement disturbances and visual, olfactory, and emotional or psychogenic factors are potent but variable emetic stimuli and can be expected to contribute to the development of nausea and vomiting in many, and perhaps most, patients. For example, the sight, sound, or smell of just one traveller or patient who is vomiting may be sufficient to trigger the same response in others. Thus, it could be questioned whether the residual vomiting of patients with cancer placed on an open ward after 5-HT_3 receptor antagonism could be partly psychogenic rather than exclusively chemotherapy-induced. Incautious movement of a patient postoperatively, during pregnancy, or during other nausea- and emesis-inducing procedures may also contribute to observed symptoms.

Nonpharmacological remedies for nausea and emesis that are available for all individuals, and not specifically motion sickness, include restricting head movements, lying in a supine position, and closing the eyes. There are reports that acupuncture at P6 is effective to relieve nausea and vomiting in the cancer patient and during pregnancy. Electrical stimulation of P6 is also being investigated. Specialized procedures of vestibular desensitization and biofeedback prior to flight and space travel aid susceptible fliers in overcoming sickness; the success of many of these programs is excellent.

9. Future Drug Treatments

Dexamethasone has antiemetic activity in its own right and is being investigated in motion sickness. The anticonvulsants carbamazepine and phenytoin, the calcium channel blocker flunarizine, the herbal product ginger and the 5-HT_{1A} receptor agonists are all being investigated in animal models and in the clinic. Their value remains to be determined.

10. Summary of Treatment for Nausea and Emesis

1. The subject or patient should, as far as possible, be provided with an environment that will minimize noxious visual, vestibular, olfactory, and emotional precipitants of nausea and vomiting.
2. Psychological procedures to offer relaxation and reassurance and relieve anxiety are important.
3. Advice should be given about eating habits.
4. Medical, surgical, acupuncture, psychological, or psychiatric procedures are used to alleviate the cause or symptoms.

See Also the Following Articles

CHEMICAL NEUROANATOMY • MIGRAINE • MILD HEAD INJURY

Suggested Reading

Andrews, P. L. R., and Davis, C. J. (1995). The physiology of emesis induced by anticancer therapy. In *Serotonin and the Scientific*

Basis of Antiemetic Therapy (D. J. M. Reynolds, P. L. R. Andrews, and C. J. Davis, Eds.), pp. 25–49. Oxford Clinical Communications, Oxford.

Davis, C. J. (1995). Emesis research: A concise history of the critical concepts and experiments. In *Serotonin and Scientific Basis of Antiemetic Therapy* (D. J. Reynolds, P. L. R. Andrews and C. J. Davis, Eds.), pp. 9–24. Oxford Clinical Communications, Oxford.

Gralla, R. J. (1998). The evolution of antiemetic treatment. In *Medical Management of Cancer Treatment Induced Emesis*. (M. A. Dicato, Ed.), pp. 1–8. Martin Dunnitz, London.

Millar, A. D. (1991). Motion induced nausea and vomiting. In *Nausea and Vomiting : Recent Research and Clinical Advances* (J. Kucharczyk, D. J. Stewart, and A. D. Millar, Eds.), pp. 13–41. CRC Press, Boca Raton, FL.

Millar, A. D., Rowley, H. A., Roberts, T. P. L., and Kucharczyk, J. (1996). Human cortical activity during vestibular- and drug-induced nausea detected using MSI. *Ann. NY Acad. Sci*. **781,** 670–672.

Morrow, G. R., and Roscoe, J. A. (1998). Anticipatory nausea and vomiting: Models, mechanisms and management. In *Medical Management of Cancer Treatment Induced Emesis* (M. A. Dicato, Ed.), pp. 149–166. Martin Dunnitz, London.

Tache, Y., and Wingate, D. (1991). *Brain Gut Interactions*. CRC Press. Boca Raton, FL.

Watson, J. W., Nagahisa, A., Lucot, J. B., and Andrews, A. L. R. (1995). The tachykinins and emesis: Toward a complete control? In *Serotonin and the Scientific Basis of Antiemetic Therapy* (D. J. M. Reynolds, P. L. R. Andrews, and C. J. Davis, Eds.), pp. 233–239. Oxford Clinical Communications, Oxford.

Yates, B. J., Sklare, D. A., and Frey, M. A. B. (1998). Vestibular autonomic regulation: Overview and conclusions of a recent workshop at the University of Pittsburgh. *J. Vestibular Res*. **8,** 1–5.

Neocortex

JON H. KAAS
Vanderbilt University

GLOSSARY

area A major subdivision of neocortex. Each area performs a specific set of functions. classically, areas have been called the organs of the brain.

column or module Subdivisions of areas that mediate a function or functions that are repeated many times within other modules of the same type within the area. Areas may have two or more types of intermixed modules.

cortical magnification The greater representation of important parts of sensory surfaces in cortical areas.

representation Areas are said to represent a sensory surface, such as the retina, skin, or cochlea, when stimulation in different parts of the sensory surface activates neurons in different parts of the area in a matching or isomorphic pattern. Muscles and movements are also represented in areas of motor cortex. Some areas may have more abstract, higher order representations. In each case, the representation is based on a spatial pattern of neural activity.

Neocortex is the part of the outer shell or "bark" of the forebrain that was thought to be new with the evolution of mammals. The cerebral cortex or pallidum covers the deeper parts of the forebrain or telencephalon, and it is generally divided into three parts: the lateral paleocortex or olfactory cortex, the medial archicortex or hippocampus and subiculum, and the neocortex lying in between. The archicortex and paleocortex are easily recognized in reptiles, and therefore they have terms that reflect the early belief that they are phylogenically old parts of the brain. Whereas all mammals have an obvious neocortex, nothing quite like neocortex exists in reptiles. Hence, early investigators coined the term "neocortex" to refer to this seemingly new part of the brain. Nevertheless, neocortex is not really new with the advent of mammals. Rather, neocortex is homologous with the dorsal cortex of reptiles, a rather small and unimpressively thin sheet of tissue with hardly more than a single row of neurons, a marked contrast to the thick neocortex with several layers of neurons. Because neocortex really did not originate with mammals, some investigators prefer to call the structure isocortex, using an early term that refers to the relatively uniform structure or appearance of neocortex throughout. However, most investigators continue to use the term neocortex, and it is accurate to say that its thick, laminated form is new with mammals. No reptile has a dorsal cortex that looks like neocortex, and no mammal has a neocortex that looks like dorsal cortex. Thus, a simple, stable, and functionally limited structure that has been retained from early to present-day reptiles has been transformed into a more complex, highly variable, and remarkably flexible structure in mammals. No living vertebrates have a cortex that is intermediate between dorsal cortex and neocortex. Thus, mammals are characterized by neocortex as much as by mammary glands. Whereas

neocortex has a somewhat uniform appearance, being easily recognizable as neocortex throughout and across species, its variability in size and organization is what allows mammals to be so different in behavior and abilities. Neocortex allows us to be human and rats to be rats. To understand how this is possible, we need to identify the common features of neocortex, as well as how it varies from mammal to mammal.

I. LAMINAR ORGANIZATION

The basic design of neocortex is similar in all mammals. Neocortex is a thick (compared to dorsal cortex) sheet of tissue of varying size but usually 1–3 mm in depth. The thickness of cortex results from a proliferation of neurons, such that the cell bodies of over 100 would be encountered by a pin stuck through its surface to the underlying axon fibers that shuttle information to and from the cortex. This contrasts to the one or two neurons that would be encountered by a pin stuck through the dorsal cortex of reptiles. Whereas the thin row of neurons in dorsal cortex both receives input from other parts of the brain and provides output, the thicker neocortex provides a more complex and fundamentally different type of processing. First, the neurons across the thickness of cortex are of different morphological types and they have different connections and functions. Second, neurons of similar appearance and connections are grouped according to depth in the cortex so that groups of cells form a stack of layers. Early investigators numbered and named the layers in various ways, but a scheme of six basic layers, stemming from the publications of Korbinian Brodmann around 100 years ago, has become a nearly universal standard (Fig. 1). Some of the layers of other early investigators are now considered to be sublayers within the standard scheme.

Layer 4, a middle layer of neurons, is the main receiving layer of neocortex (Fig. 2). Neurons in this layer receive most of the activating input from other structures, including nuclei of the dorsal thalamus and other subdivisions of the cortex. The connections of neurons in layer 4 are predominantly local within layer 4 and with neurons immediately above or below in adjoining layers. There, the information that is received in layer 4 rapidly spreads in a narrow focus across the layers. These and similar vertical connections of neurons in other layers form vertical columns of coactive neurons in neocortex. This is a highly characteristic feature of cortical organization.

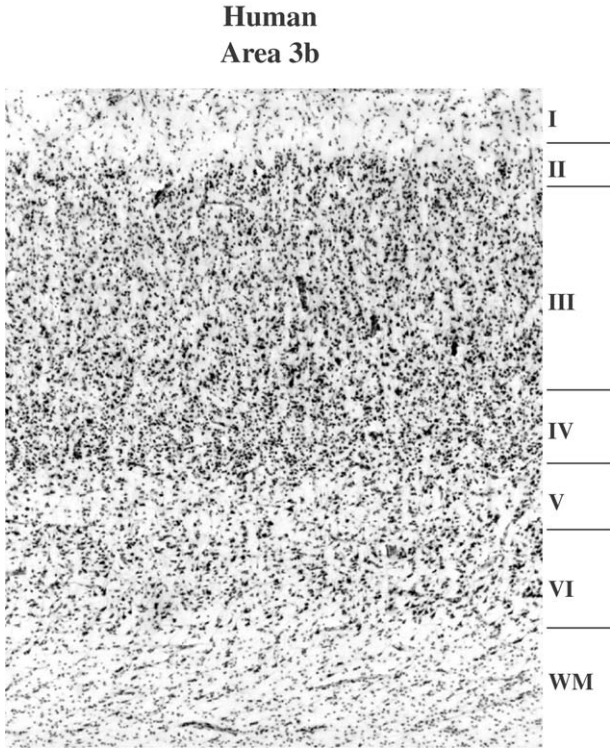


Figure 1 The laminar organization of the neocortex. Part of the neocortex was cut into thin slices across its thickness, and the sections were stained to reveal cell bodies (dark ovals) but not axons and dendrites. The cell bodies vary in size, shape, and packing density across the thickness of neocortex in a laminar pattern. It is now usual to divide the neocortex into six main layers (I–VI; for historical reasons, Roman numerals are usually used to number layers) from the surface to the underlying connecting axons or white matter (WM). Sublayers are also commonly distinguished. Layers play different functional roles in the processing that goes on in the neocortex (see text). Different areas of the neocortex are specialized for different functional tasks, and such specialization may be reflected in variations in the appearance and relative thickness of the layers. The section of cortex shown here is from the primary somatosensory cortex of a human, commonly referred to as area 3b after a numbering scheme devised for cortical areas by Korbinian Brodmann over 100 years ago. This cortex is just over 2 mm thick. Area 3b and other sensory areas of neocortex are characterized by a relatively thick layer IV that is densely packed with neurons having small cell bodies. Layer III is especially thick in the large brains of humans.

For activation, layer 4 neurons depend on focused synapses from only a few axons from another structure, and they spread this information to only a few nearby neurons that are immediately deeper or more superficial in the cortex. Thus, layer 4 neurons are rather small, with radial dendrites spreading around the cell body and a restricted and local axon arbor contacting adjacent neurons in layer 4 and neurons

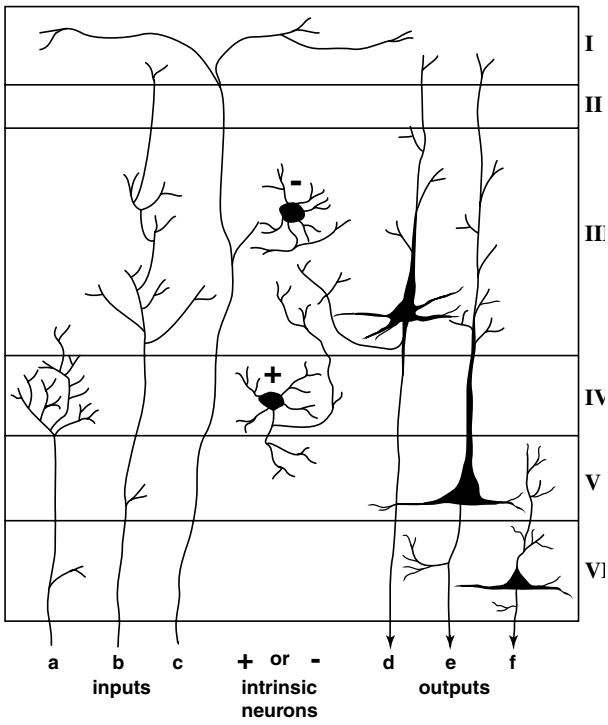


Figure 2 Major input axons, output neurons, and intrinsic neurons of the neocortex. Input to cortex includes (a) the major activating input from the thalamus or other areas of cortex that terminates largely in layer IV; (b) the feedback activating connections from other areas of cortex that terminate more diffusely above and below layer IV; and (c) the diffusely distributed modulating input that terminates above layer IV. Intrinsic neurons include excitatory (+) stellate or granular cells in layer IV with local dendrites and a local axon arbor and an inhibitory (−), nonspiny class of stellate cells or local circuit neurons of various types in all layers. Output neurons include (d) layer III pyramidal cells that project to other areas of the cortex and to subcortical targets; (e) larger layer V pyramidal cells that project mainly to subcortical targets including the spinal cord; and (f) pyramidal-like neurons of several types in layer VI that provide feedback connections to the

above and below layer 4. The small neurons in layer 4 are typically called stellate cells because their short dendrites protrude in all directions like the rays of a star. Because layer 4 cells are often so small as to appear to be no more than grains of sand, they are also called granule cells. Layer 4 is sometimes called a granular layer.

Layer 5 neurons, just under layer 4, are activated by layer 4 and other neurons in the superficial layers. The major neuron type is the large pyramidal neuron (Fig. 2). The cell body is triangular in shape with the apex pointing outward as it gives rise to a long apical dendrite that ascends toward or to the surface of layer 1. The base of the triangular cell body also gives rise to basilar dendrites that are much shorter and spread out horizontally away from the vertical cell column and ascending apical dendrite. The dendrites are covered with synaptic contacts from other neurons. The arrangement of the dendrites indicates that layer 5 pyramidal cells are designed to gather information somewhat horizontally in layer 5 via the basilar dendrites, and more locally from the more superficial layers via the long ascending apical dendrite, which has short radiating branches along its ascending stem. Thus, layer 5 pyramidal neurons have access to input from a vertical array of neurons, including those of a vertical row along the apical dendrite and those in adjacent rows contacted by the basilar dendrites or the branches of the apical dendrite. The layer 5 neuron sums all this information, and it provides the major output of the neocortex. In addition to forming a local axonal arbor, most layer 5 pyramidal neurons project over quite long axons to nuclei in the dorsal thalamus, structures in the subcortical basal ganglia, the midbrain, and other brain stem structures, and even the spinal cord. Thus, the output of neocortex that most directly affects behavior and performance stems from layer 5 pyramidal neurons. Some layer 5 neurons also send information to more distant locations in the neocortex, either in the same cerebral hemisphere or in the opposite cerebral hemisphere via the corpus callosum. Thus, layer 5 neurons help to inform other areas of neocortex about ongoing computations. In brief, a major function of layer 5 neurons is to send information to other parts of the brain. Because these targets can be relatively far from the neocortex and distance is time for the nervous system, layer 5 neurons have long, thick axons. Because thicker axons have faster conduction rates, the transfer of information is speeded up. Neurons with long, thick axons and long, branching dendrites need large cell bodies to maintain these structures, and layer 5 pyramidal cells are large. They also vary in size according to axon length and thickness, so that the large Betz cells of the motor cortex project all the way to the lower spinal cord, whereas the large Meynert cells of the visual cortex project to the opposite hemisphere, to a distant visual area in the neocortex, and to the brain stem.

The other main output layer of the neocortex is layer 3, and it also contains pyramidal cells with basilar and apical dendrites (Fig. 2). However, layer 3 pyramidal cells generally are not as large as layer 5 pyramidal cells, because their apical dendrites are shorter and, more importantly, their axons are shorter, traveling to other areas of cortex rather than to more distant subcortical structures. Thus, neurons in any part of the

neocortex send information to other parts of the neocortex, and this is done largely by layer 3 pyramidal neurons. In a manner comparable to layer 5 pyramidal neurons, layer 3 pyramidal neurons receive local input along the vertical apical dendrite extending and branching throughout the depth of layer 3 and the basilar dendrites extending horizontally, sometimes to adjacent cell columns. In general, deeper layer 3 pyramidal cells are larger than more superficial layer 3 pyramidal cells. Layer 3 often has obvious sublayers. Layer 3 neurons send information to layer 4 neurons of other regions of the cortex or directly to other layer 3 neurons. The great computational power of the neocortex stems from these layer 3 pyramidal cells because they transmit information from one region of neocortex to another. Each local group or column of neurons performs a similar and somewhat simple operation or computation, but it is the consequence of a series of simple computations that gives the cortex its great computational power, a power beyond that of the input–output operations of the dorsal cortex of reptiles.

Other layers add to the functional complexity of neocortex. Feedback about computational outcomes is important in information processing systems, and neocortex is designed to provide feedback both to the subcortical thalamic nuclei that provide all sensory information to the neocortex and to other cortical areas that provide information for further processing. Layer 6 is the main feedback layer (Fig. 2). Neurons in layer 6 typically provide feedback projections to the same structure that provides the major activating input to a cortical area or region. If, for example, the major activating input for a subdivision of visual cortex is from a nucleus in the visual thalamus, then layer 6 neurons project back to that same nucleus. If layer 3 neurons in one area of cortex provide the major activating input to layer 4 of another area of cortex, then layer 6 neurons from the target area typically project back to the activating region. Other neurons in other layers may also participate in providing some feedback. The layer 6 neurons receive rather direct information from upper layers, especially layer 4, and they even receive a small amount of direct input from branches or collaterals of axons terminating in layer 4. Thus, input from the visual thalamus to the visual cortex terminates largely in layer 4, but it also provides a few direct branches to layer 6. Layer 6 neurons in turn project back to the sending neurons in the visual thalamus to inform these neurons about the state of the cortex and to influence the next burst of information being sent to the cortex. Layer 6 neurons, by concentrating on responding to local layer 6 connections and input from more superficial neurons, have basilar dendrites and a variable but often long apical dendrite. Thus, they are considered to be pyramidal cells or modified pyramidal cells, although layer 6 is sometimes referred to as the internal granular layer, with layer 4 being the external granular layer. However, layer 6 neurons differ from layer 4 neurons by projecting to distant structures, and thus layer 6 neurons generally are larger. Besides having projections to thalamic nuclei and other cortical areas, layer 6 neurons project to a subcortical structure, the claustrum, which in turn projects back to layer 6. This reciprocal pattern of connections functions to modify the response properties of layer 6 neurons.

The other two layers of neocortex are layers 1 and 2. Layer 2 is a thin layer of densely packed small neurons with local connections and a role in modifying local processing via contacts on apical dendrites of deeper pyramidal cells. Layer 1 is a fiber layer with few neurons that largely consists of the ends of apical dendrites and axons that course along the brain surface and contact these dendrites. Thus, layer 1 is reminiscent of the outer fiber layer of the dorsal cortex of reptiles where input contacts dendrites. Some of the input to layer 1 is modulating input from brain stem neurons. Other input is feedback connections from other cortical areas.

II. LOCAL CIRCUIT FUNCTIONS

Whereas the basic structure of neocortex involves five layers of neurons and one layer of fibers that differ in functional role, neuron types, and connections, the functioning of the cortex cannot be fully understood from this framework alone. It is also important to consider several additional features of cortical organization. First, all layers of cortex contain populations of two fundamentally different types of neurons. The ones discussed so far excite each other and use the excitatory neurotransmitters. The release of an excitatory neurotransmitter at each synaptic contact tends to depolarize the contacted neuron and increase the probability of an action potential and the transmission of information. These excitatory pyramidal and spiny stellate cells constitute roughly 80% of the total number of neurons. The other 20% of neurons, distributed across the layers, are nonspiny stellate cells that release the inhibitory neurotransmitter, GABA, which hyperpolarizes the contacted neuron and

decreases the probability of an action potential. The GABAergic or inhibitory neurons in the cortex all have short axons that distribute locally, and thus they are considered to be interneurons or local circuit neurons (Fig. 2). Inhibitory and excitatory neurons interact to produce important outcomes within the circuitry of neocortex. First, circuits tend to respond briefly to changes in input. Input to layer 4 excites both excitatory and inhibitory neurons. The excitatory neurons transmit excitation to other excitatory and inhibitory neurons. The inhibitory neurons inhibit excitatory neurons and other inhibitory neurons. The immediate result is to limit the duration times of a cortical response to any input. The initial excitation gets through to activate the neuron, but later arriving inhibition dampens subsequent activity. Thus, the system selectively responds to what is new and changing in time. Response to temporal changes in the environment is usually what is significant for adaptive behavior and survival. Second, circuits compare and contrast adjacent inputs. The synaptic contacts of inhibitory neurons are local, and they tend to distribute horizontally to adjacent neurons. Thus, when adjacent vertical arrays of neurons respond to the same stimuli, they tend to inhibit each other and overall excitation decreases. In contrast, if input to adjacent neurons is activated by different stimuli so they do not respond at the same time, less lateral inhibition occurs to dampen local excitation. The circuits of neurons are most excited by local differences in activating input that is often related to spatial contrast in the stimuli falling on the retina, skin, or cochlea. Spatial antagonism between a core of neurons and a surround of neurons provides a neural mechanism for emphasizing spatial differences over uniformity, again something that is very useful biologically. A fly on a sheet of paper, for example, activates cortical neurons more than any comparable location on the uniformly white sheet of paper, and we easily detect the fly. By repeating neural computations in a series of successive circuits, the temporal and spatial contrast detection features of circuits can progress from local to global features and from mediating simple and "mindless" abilities to the complex and astonishing.

What else is important about neural circuits in the neocortex? Local circuits in vertical arrays in neocortex interact with each other within each area over short, lateral, or horizontal connections. Excitatory neurons, especially in layer 3, often have horizontal connections extending for several millimeters within the layer. The existence of these connections, which are relatively sparse, has long been known, but their significance is only now beginning to be understood. Basically, it appears that these sparse and relatively weak connections correlate the ongoing activities of separate groups of neurons. If the neurons are already responding to something, it does not take much in the way of interconnections to shift the times of neuronal action potentials or spikes so that the spikes occur in the separated neurons at the same time. If some of these neurons project to the same group of neurons in another structure, the arrival of these spikes from different groups of coactive neurons at the same time will have a powerful activating effect on the target neurons. There may be many uses for such correlated firing of neurons. One suggestion is that such correlation mediates "perceptual binding." All of the neurons firing at once, being activated by different aspects of the same stimulus object, signal by their correlation that they are responding to the same object. The object is thereby seen as one whole, rather than as fragments.

Another feature of neocortex that makes it so biologically useful is that the circuits are adjustable, rather than fixed. They are adjusted by "experience" in the way they respond to stimuli and other neurons so that they function in more useful and productive ways. This is largely because neurons have two types of membrane receptors for the excitatory neurotransmitter, glutamate. One type of receptor, the non-NMDA receptor functions as an excitatory channel that tends to generate neuronal spikes. The other receptor, the *N*-methyl-D-aspartate or NMDA receptor, is often blocked by a magnesium ion at resting membrane potentials so that glutamate release has no effect and the NMDA receptor fails to participate in the neuron's response. However, when the neuron has been depolarized by the activation of the non-NMDA receptors, the magnesium ion is released and the NMDA receptors can be activated and pass calcium ions by glutamate release, but only for a small fraction of a second. In some sense, the receptors function as "coincidence" detectors, because they are active when two sources of excitation are closely timed. The activation of NMDA receptors then provokes internal alterations in the neuron that strengthen the effectiveness of the synapses that were just active. Thus, some connections in the circuit are made more powerful. This change is often called long-term potentiation or LTP because synapses are potentiated or strengthened for up to weeks at a time. Other activity patterns weaken synapses. This change is called long-term depression or LTD. Thus, local circuits of neurons are always being altered in activity-dependent ways. These alterations are important in shaping brain circuits

during development and throughout life. They adjust the performance of local circuits and are necessary responsible for learning from sensory experience.

Neocortex has another way of adjusting its performance. Besides excitatory and inhibitory connections input exists that releases more long-lasting neurotransmitters that alter the probability of neurons producing spikes. Neural circuits in the neocortex do not always function at the same level of excitability. When we are drowsy or sleeping, the circuits may be harder to activate. When we are alert and motivated, the circuits may respond more readily. When we focus our attention, some groups of circuits become more responsive and others less responsive. These adjustments need not be spatially nor temporally precise, and they are mediated by neuromodulatory systems that project from the brain stem to the cortex in a widespread but regionally variable manner. When they are activated, these fiber systems release such transmitters as norepinephrine, dopamine, serotonin, or acetylcholine. Such releases generally alter the responsiveness of cortical neurons to create greater outputs. These increases in neural activity patterns also promote longer lasting alterations of neural circuits. Acetylcholine release, for example, is associated with arousal, and it promotes LTP and allows circuits to be modified more easily, thus improving learning during arousal.

In addition to these brain stem modulating projections that distribute broadly especially to the outer layers of cortex, some of the thalamic input also appears to be of the modulating type in that it avoids layer 4 neurons and instead terminates in the outer cortical layers, including layer 1. Such terminations, often on the distal dendrites of neurons, have only minor excitatory effects that add to the major sources of activation. The functions of such modulation are not well-understood, but they certainly add to the functional flexibility of the cortex.

III. AREAS OF NEOCORTEX

The basic organization of the neocortex helps explain how it can be such a powerful and flexible computing machine, but it does not explain the great variability in cortical function that occurs in different species of mammals, making one mammal a mouse in performance and capability and another a man. Whether you are a ''man or a mouse'' largely depends on your neocortex. A major difference in the neocortex of humans and mice is in the size of the cortical sheet. Mice have about 2 cm^2 of neocortical surface per cerebral hemisphere, whereas humans have 800 cm^2 or roughly 400 times as much. Human cortex is also 2–3 times as thick. Simply put, we expect a computer that is about 800 times larger to be better, and it is. A mouse will never speak or perform higher mathematics.

But as we know, size is not everything. Otherwise, mammals with brains of the same size would perform in similar manners. Instead, the neocortex can be quite different in internal organization from species to species. In addition, acquired differences in cortical organization exist in each individual. The cortex of a great tennis player is organized differently from that of the great scholar, largely as a consequence of experience-related modifications in the neocortex.

How does neocortex differ from species to species, other than in size? The most conspicuous ways are in numbers and types of cortical areas. Early investigators such as Korbinian Brodmann proposed that neocortex is subdivided into a number of functionally distinct regions or areas. For Brodmann, the cortical areas were the ''organs of the brain,'' each with a set of specific functions. The existence of such organs was suggested by the impairments produced by damage to different regions of the neocortex. As a well-known example, in 1861 Pierre Broca proposed that a speech area, now known as Broca's area, is located in the left frontal lobe after he noted that speech production problems followed lesions in the ventral part of the left frontal lobe. But working approximately 100 years ago, Brodmann had only one way to delimit and define areas of the cortex: he stained thinly cut sections from the brain for the cell bodies of neurons and used the slight differences in appearance that he noted from one region of cortex to another to define areas. The major or primary sensory areas, for example, have a thicker layer 4 that is more densely packed with small neurons (Fig. 1). This feature reflects the fact that layer 4 is the major receiving layer, and sensory areas are specialized for receiving input from the thalamus. In contrast, primary motor cortex is specialized for sending motor commands to the brain stem and spinal cord, and the major output layer 5 is especially thick, with large pyramidal neurons. Because the sensory role of the motor cortex is reduced, it does not have an obvious layer 4, also known as the external granular cell layer. Thus, primary motor cortex is sometimes called agranular cortex. Brodmann both named and numbered his cortical areas. Thus, primary visual cortex was named the calcarine type for its location in the calcarine sulcus of the occipital lobe and given a number, 17, as the seventeenth area described in an

arbitrary order. Area 1, just on the crest of the postcentral gyrus in the middle of the cerebral hemisphere, is a subdivision of somatosensory cortex, whereas the primary motor cortex was labeled area 4 by Brodmann.

Brodmann's approach was to study brain sections stained for cell bodies and look for distinctions that would suggest or indicate the boundaries between functional subdivisions. Boundaries between areas were usually or often assumed to be sharp, but Brodmann allowed for the possibility of gradual change between some areas. The method of study is known as cytoarchitectonics or the study of tissue structure from brain sections stained for cells. Soon, others used stains for myelin, the fatty substance around axons. The method of using regional differences in the myelination pattern to divide the neocortex into areas is called myeloarchitectonics. Currently, there are many ways of processing brain tissue for proteins and other substances in neurons, so that a field of chemoarchitectonics has emerged. Many new methods have proven to be very useful in defining cortical areas. Yet, the process has been and remains difficult. Often, architectonic results are difficult to interpret, and different investigators come to different conclusions.

Brodmann boldly defined over 50 areas in human neocortex and found fewer in other mammals such as monkeys and small-brained hedgehogs. Several of his general conclusions remain valid today. We now have considerable evidence that small-brained mammals such as hedgehogs have few areas, perhaps 15–20, monkeys have more, and humans have still more, probably many more than the 50 or so proposed by Brodmann. Moreover, the brains of all mammals share a few areas, as these areas emerged with the evolution of mammals and have been retained in most or all lines of evolution. In addition, as some mammals evolved larger brains, they added cortical areas. How cortical areas were added is unknown, but Brodmann believed that they emerged by existing areas gradually becoming differentiated into two or more areas.

Because only a few areas of the brain are so specialized that they could be easily seen in stained brain sections, Brodmann only correctly identified a few brain areas. Other investigators at that time came up with other proposals of how brains are divided into areas, but only Brodmann's numbering scheme remains in wide use. Unfortunately, the numbered areas largely mark locations in cortex rather than functionally distinct areas. Thus, many investigators are currently trying to determine how the cortex is divided into areas in humans and other mammals. The evidence that neocortex is subdivided into a patchwork of cortical areas in all mammals and that different species of mammals differ in number of areas is very strong, but the task of accurately locating and defining most of the areas has proven to be very difficult. Investigators now consider task-related neuronal activity patterns in the cortex and the ways in which cortical regions are connected with each other and other structures as additional sources of information. Of course, from the very beginning, investigators such as Pierre Broca used the relationship of different behavioral impairments to lesions of regions of cortex as a way of surmising the location and function of cortical areas.

Despite the lack of complete knowledge about cortical organization, it seems certain that a major reason that the human brain functions so well is that it has so many cortical areas. Studies of cortical organization in monkeys have given us some clear evidence for how the neocortex is organized in humans, and this evidence has provided some surprises. Most notably, many scientists once thought that most neocortex in humans and monkeys would be devoted to large regions of "association cortex" for associating the information from the different senses. Such large areas would deal with abstractions, so that they would be activated by visual, auditory, or somatosensory events. Thus, most neocortex was considered to be responsive to several senses or multimodal. We now believe that a great deal of the neocortex is devoted to processing information within a sense, and the creation of accurate, modality specific perceptions is not an easy task for the brain. Consider the inaccurate perceptions of a bird attacking its own reflection in a glass window, and the impact of limited processing becomes apparent. Monkeys have over 30 cortical areas for processing visual information, at least 15 for processing somatosensory information, and 15 or so for auditory information. Humans may have more. Why so many?

The reason for so many processing areas seems to be that each area performs a few rather simple computations. However, if the results are fed into another area for a second level of processing, roughly the same types of simple computations can add a level of complexity and sophistication. Complex abilities depend on a chain of processing events, simple at each step, but complex in the end. Of course, processing is not just in a series or chain, because feedback always (or nearly always) occurs between interconnected areas, cross-talk occurs between areas of roughly the same level, and much processing is over parallel chains of circuits.

Nevertheless, small-brained mammals are behaviorally limited compared to large-brained mammals, not only because they have less neocortex but also because they have fewer cortical areas.

All mammals appear to have primary visual, somatosensory, and auditory areas, as well as one or two associated secondary areas for each sense, and most mammals have one motor area or more (Fig. 3). Visual areas are located caudally in the occipital region or lobe, and they extend ventrally into the temporal lobe in mammals with more visual areas. Auditory areas are in temporal cortex or the temporal lobe, somatosensory areas are in parietal cortex, and motor areas are in the frontal lobe. In mammals with more areas, the visual, auditory, and somatosensory areas remain grouped. The grouping of areas for each sense in the same region of neocortex allows for more rapid processing over shorter interconnections.

IV. SENSORY REPRESENTATIONS (MAPS)

In all mammals, much of the neocortex consists of orderly representations or maps of receptor surfaces (see Fig. 3). Visual areas of each cerebral hemisphere represent the contralateral half of the visual field via the nasal hemiretina of the contralateral eye and the temporal hemiretina (actually a smaller and species variable portion that is less than half) of the ipsilateral eye. Somatosensory areas represent the receptors of the contralateral body surface. Auditory areas represent tones from high to low frequencies as they activate successive locations along a row of cochlear hair cell receptors in the contralateral and ipsilateral ears. Other cortical areas relate to taste and vestibular input. Smell is relayed to piriform cortex from the olfactory bulb. The cortical representations are activated from the thalamus via orderly, topographic projections from the receptor sheets to subcortical nuclei that relay to the thalamus, or to the thalamus directly, and then to cortex. The maps closely, but not precisely, reflect the order of the receptors on the receptor sheet. The maps might have disruptions so that receptors next to each other are represented in nonadjacent parts of the map, and the maps might have modular repetitions related to receptor class or eye or ear of origin (see later discussion). In addition, the receptive fields, the portion of the receptive sheet that activates a neuron, become larger for neurons in each successive level of a series of connected nuclei and areas, so that the maps in initial subcortical structures are the most detailed, the maps in primary sensory areas are somewhat less detailed, and secondary and higher order sensory areas contain progressively less detailed maps. These changes in receptive field size and the detail of map topography are compatible with the changing functions of neural circuits in representations that are early or late in a processing sequence. The number of topographic maps of a sensory surface in cortex varies with the number of sensory areas, with the early areas in sequences being most topographic and later areas often reflecting little of the order of the receptor sheet, but possibly representing in a systematic manner some product or feature that has been derived from other sensory areas. Topographic maps place neurons that most interact within the area close together and provide a suitable substrate for the common spatiotemporal computations of neural circuits.

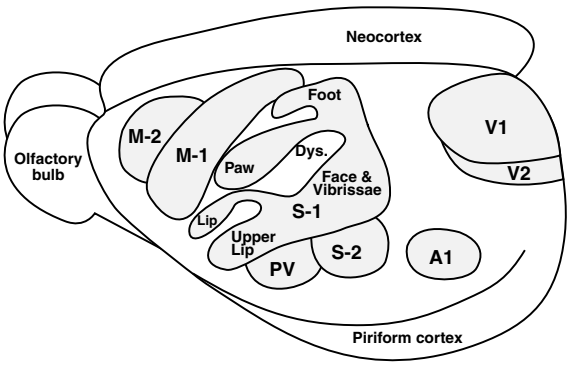


Figure 3 Some of the currently proposed areas of neocortex in rats. Like most other mammals, rats have a primary somatosensory area, S1, and a secondary somatosensory area, S2, a primary visual area, V1, and a secondary visual area, V2, a primary auditory area A1, and a primary motor area, M1. There is also evidence for a secondary or premotor area, M2, a dysgranular (Dys) somatosensory area with a less developed or "dysgranular" layer IV, and a parietal ventral (PV) somatosensory area. S1, S2, and PV each represent the receptors of the skin in an orderly manner. The locations of activating input relayed from major body parts are indicated for S1. Other areas of neocortex in rats have been proposed, but they are not indicated. Mammals, such as monkeys and humans, with larger brains and more neocortex have the same areas as those proposed here for rats, at least a few additional areas that they share with rats, and a number of additional areas that are not found in rats. Thus, species differ in numbers of cortical areas, as well as the proportional extent of neocortex. Piriform cortex and the olfactory bulb are indicated for reference.

V. MOTOR REPRESENTATIONS

Areas of neocortex also represent muscles and movements. Long before it was technically possible to

record the small changes in the electrical activity of neurons in neocortex, it was possible to generate and deliver an electrical current to the surface of neocortex and observe the effects. From quite early studies, it became apparent that the electrical stimulation of locations in the frontal lobe was especially effective in evoking movements of various parts of the body and that different sites in the frontal cortex related to different movements. Parts of the frontal cortex were seen to systematically represent body movements, and a primary motor area, M1, was defined first, followed by a supplementary motor area, SMA. M1 is a striplike area that extends mediolaterally along and in the central sulcus of the human brain just rostral to the somatosensory cortex, and it represents the body from foot, trunk, hand, face, and tongue in a mediolateral sequence. M1 largely, but not precisely, corresponds to cortex defined earlier by Brodmann as area 4, but many modern depictions make M1 coextensive with a redefined area 4. SMA is rostral to the more medial part of M1. The second motor area, M2, of rats (Fig. 3) might correspond to SMA, which is best known from studies in primates.

Detailed studies of the organization of M1 have used penetrating microelectrodes to more precisely activate small numbers of the large pyramidal cells of layer 5 at low levels of current, thereby evoking movements in only a few muscles and allowing the motor map to be seen in detail. As a result of microstimulation studies, it is now apparent that M1 and probably other motor representations are organized as a patchwork or mosaic of small regions of cortex that are devoted to specific movements, that of a single finger for example. Each type of patch of cortex typically occurs more than once but next to various other patches of motor cortex for other related movements. Such an organization may give the motor cortex great flexibility in the control of various combinations of movements.

Humans and other primates have additional motor areas in the frontal lobe that are collectively called premotor areas. They include the dorsal and ventral premotor areas just rostral to medial and lateral portions of M1, respectively, the supplementary motor area rostral to M1 on the medial wall of the cerebral hemisphere, and other motor areas on the medial wall and dorsolateral surface of the frontal lobe. Movements can be evoked by electrical stimulation from all of these areas, and each area represents the body in a sequence across the field. To evoke movements in these premotor areas takes higher levels of current than in M1, and the movements evoked are more complex. Movements also can be evoked by stimulating areas of the somatosensory cortex, but at high levels of current. Thus, a number of areas of cortex have motor functions. Motor and premotor areas also have sensory input, and they may also contain sensory maps.

VI. OTHER CORTICAL AREAS

Neocortex also has several types of areas that are not strictly sensory or perceptual in nature and not closely tied to motor behavior, although neurons in these areas may respond to sensory stimuli and the areas may be closely involved in behavior. One such region of neocortex in primates is dorsolateral prefrontal cortex, where neurons appear to briefly retain information and be involved in short-term or working memory. Areas of dorsolateral prefrontal cortex have not been defined completely, but several areas probably exist. The more ventral cortex of the frontal lobe has areas that appear to provide evaluative functions and emotional tones to experienced events. Cingulate cortex along the medial wall of each cerebral hemisphere has rostral divisions that are more motor in function and caudal divisions that are more sensory in function. Areas of the cingulate cortex appear to be involved in instinctive behavior and motivation. In humans, other areas, usually in the left cerebral hemisphere, are involved in processing language. Broca's area in the left frontal lobe, as a probable specialization of a ventral premotor area, is critically important for the production of speech. More caudal areas in the temporal lobe are used to extract meaning from speech and provide meaningful structure to speech. Part of the superior surface of the left temporal lobe is known as Wernicke's area after the description of Carl Wernicke in 1874 of defects in speech that result from lesions in this region. Such individuals easily produce words and phrases, but the speech has little meaning. Wernicke's area may actually contain several processing areas, as the organization of this cortex is not well-understood. Other areas appear to be more like the "association" cortex of early proposals and multimodal or polysensory in nature. Such a multimodal region, the superior temporal polysensory area (or region, because it probably contains several areas), has been described in the temporal lobe of monkeys. Whereas the functions of the superior temporal polysensory region are still largely unknown, neurons in the region respond to visual, auditory, and somatic

stimuli, and this cortex could be a site for integrating information across sensory modalities.

VII. SPECIES DIFFERENCES IN CORTICAL ORGANIZATION

As mentioned, species of mammals vary in amount of neocortex and in number of cortical areas. These differences obviously are important in mediating functional differences. In addition, neocortex and areas of neocortex differ from species to species in several other ways. These organizational differences are responsible for mediating many important variations in behavior and ability.

A. Cortical Magnification

Sensory areas, especially those early in the cortical processing sequence, are said to represent the receptor sheets. Thus, stimuli falling on different parts of the receptor sheet activate neurons in different parts of sensory areas in a pattern that is isomorphic with the receptor sheet. However, not all parts of the receptor sheet are equally important. For example, we only see well within a few degrees of central vision corresponding to the specialized center of the retina, the fovea, and the immediate surround, where receptor cells are very densely packed. In humans and monkeys, much of the visual cortex is devoted to processing information from the fovea. In part this is no more than a reflection of the greater receptor density of the fovea. However, the amount of cortex devoted to the fovea exceeds the proportional relationship to retinal receptors. Thus, there is a distortion within sensory representations in the neocortex that follows the variable density of receptors across the sensory surface (receptor-based cortical magnification) and a distortion that exceeds that dictated by receptor density (functionally based cortical magnification). Other mammals may distort sensory maps in other ways. For example, ground squirrels benefit from accurately seeing predators coming from the side as well as from the front. Their retinas have a horizontal strip of high receptor density, and their visual cortex is largely devoted to processing information from this strip of receptors. Rats devote most of their somatosensory cortex to processing information from the long whiskers on the sides of the face that they whisk forward to contact objects, whereas humans devote much of their somatosensory cortex to information coming from our sensitive finger tips. Thus, mammals vary in sensory and perceptual abilities, not only from specialization and alteration of the receptor surfaces but also by changes in the proportions of cortical areas that are devoted to different parts of the receptor sheet.

B. Laminar Differentiation

Neocortex varies in another way that also adds flexibility to its function. The laminar organization of neocortex has already been mentioned, as have the slight variations in the appearance and thickness of layers from cortical area to cortical area that reflect functional specialization. The same area can also vary across species in differentiation, so that the primary visual cortex in highly visual mammals, such as monkeys and humans, is more distinctively laminated than in poorly visual rats. Apparently, something is to be gained by tightly grouping cells in layers and sublayers of the same functional types and developing cells of different shapes to allow specialized functions, but there must also be some cost or loss for such differentiation or rats would have a highly differentiated visual cortex. Perhaps neurons in the primary visual cortex of rats, with poor vision and few visual areas, need to retain broad, general functions and thereby do not specialize. However, one of the sources of flexibility in the neocortex is in the degree to which specialized neurons and groups of neurons are tightly confined within layers and sublayers.

C. Modules or Columns

Another source of variability and specialization in the neocortex is in types of modular organization. Cortical areas are at least sometimes, and possibly always or often, divided into smaller, functionally distinct regions called columns or modules. The concept of the cortical column was introduced by the neurophysiologist, Vernon Mountcastle, who noted during recordings from the somatosensory cortex that neurons with similar response properties were grouped into territories or columns across the thickness of cortex that were less than one 1 mm in width. Investigators soon described "ocular dominance" columns in the visual cortex of cats and monkeys, where inputs to the primary visual cortex related to one eye or the other eye were found to be segregated into alternating bands of tissue in layer 4 of something close to 0.5 mm in width.

One set of bands was activated by input related to the ipsilateral eye and the other set by the contralateral eye. Neurons in other layers were activated less exclusively by one eye or the other, but they were said to be "dominated" by one eye. Whereas cats and ferrets also have ocular dominance columns, rats and apparently many other mammals do not.

Subsequent to the discovery of ocular dominance columns, the primary visual cortex of monkeys was found to have a dotlike pattern of clusters or "blobs" of neurons of high metabolic activity that appeared to be more related to processing color information, each surrounded by tissue devoted to other functions. Next, the second visual area, V2, of monkeys was found to contain repeating sequences of three types of bands crossing the field, with each type of band having input from functionally different types of neurons in V1 and each type of band having output to different targets. Soon thereafter, a visual area in the upper temporal lobe, the middle temporal visual area or MT, was found to contain a patchwork of small regions with neurons more sensitive to either global motion or local motion. In the somatosensory cortex of monkeys, area 3b or the primary somatosensory cortex was found to contain narrow regions of neurons responding to either slowly or rapidly adapting classes of cutaneous afferents. These and related findings suggest that many or even most cortical areas are subdivided into two or more types of smaller regions, the cortical columns or modules.

It was once thought that columns would all be of roughly the same shape, ovals of about 0.5 mm in diameter and extending columnlike from the brain surface to the fibers underlying cortex. Now it is apparent that columns come in a range of shapes, from a patchwork of small dotlike centers with large surrounds to bands. The shapes may depend in part on the ratio of neurons of one type or the other. Sizes also vary, but not by much, from widths of less than 0.5 mm to well over 1 mm. The differences in the shapes of the groups of neurons are more compatible with the term "module" than "column," although both terms are in use and are interchangeable. Modules allow areas to perform more specialized functions, with each area acting as two or three. By grouping neurons into processing units, the lengths of connections between neurons are reduced. Mammals not only differ in numbers of cortical areas and in how areas are differentiated into cell types and layers but also in how areas are divided into modules. Thus, areas can be highly specialized for the needs of a particular species or line of evolution.

D. Connections

Input and output obviously are important in mediating the functions of an area, and the same area in different species can have somewhat different connections. This aspect of variability has not been studied extensively, but clear examples of variability are known. One example might be the projections from the primary visual cortex in primates to the middle temporal visual area (MT) in the temporal lobe. All primates appear to have this direct projection, but MT has not been identified in any nonprimate mammal. Thus, the projections to MT appear to be a unique feature of the primary visual cortex in primates. As a related example, evidence exists that the primary visual cortex projects to the frontal lobe in some mammals, such as rats, but not in others such as monkeys. Across species, cortical areas vary as to with which other areas and subcortical structures they are connected, the density of these connections, and the distribution of connections of specific types within an area. Most primates have few connections between the primary visual areas in the opposite cerebral hemispheres, but prosimian primates and several other mammals have many such connections.

VIII. DEVELOPMENT AND EVOLUTION

The functional flexibility of the neocortex is related to its capacity to be altered in evolution and development. Neocortex develops from a thin sheet of germinal cells along the margin of the cerebral ventricle. The germinal cells undergo mitosis and divide to form more germinal cells, some of which ultimately differentiate into neurons or glial cells. A marginal zone of glial processes soon forms over the germinal cells, and neuron precursors migrate from the deeper germinal layer toward the surface to form a cortical plate and then cortical layers. Neurons reach the cortex largely by migrating along glial processes that extend radially from the inner ventricular surface to the outer margin of the developing hemisphere. Surprisingly, the neurons that arrive first form the deeper layers because later arriving neurons pass through the earlier arrivals to reach the surface. Most neurons pass from a specific location along the ventricle to a specific location in the cortex by migrating along the process of possibly a single radial glial cell. A row of neurons across the thickness of the cortex may largely originate from a single germinal cell

that has divided over and over again to produce a series of migrating offspring. Thus, the thickness of the cortex, constituting surface to white matter rows of roughly 100 or so neurons, may depend on roughly 100 cell divisions. Of course the number varies across cortical areas and the same cortical area across species, but not by that much. One of the reasons for the relative stability of the thickness of the cortex and the number of neurons across this thickness is the developmental process of repeatedly dividing the parent cell to produce offspring in series. Thus, a large change in the final number of neurons in a row would depend on a major change in the number of germinal cell divisions that produce neuronal precursors. Many of the inhibitory neurons of the neocortex do migrate in from a lateral location, but the major point to row migration pattern appears to provide some stability in one feature of cortical organization, the thickness. It may also be important that neurons migrate to specific regions of the cortex, because this might initiate the process of forming cortical areas.

In another way, the generation of cortex is a flexible process that easily allows increases in the size of the neocortical sheet. As neurobiologist Pasko Rakic has stressed, each symmetrical division of germinal cells doubles the number of germinal cells. When each germinal cell finally changes its role to produce a row of cortical neurons, then the number of rows and the size of the cortical sheet will be doubled by each of the previous doublings of the population of germinal cells. Whereas cortical thickness may be hard to change in development, increases in cortical size appear to depend on only a few more cell divisions in the early stages of development. Small genetic changes in the control of the number of cell divisions to form more germinal cells would have major consequences. Thus, we see great variability in the size of the neocortex, but little in the thickness.

Another important factor in the evolution of variability in the cortex is that developing cortical neurons are very sensitive to environmental influences. Developing neurons use information in neural activity patterns to form functional groups. As is sometimes stated, ''neurons that fire together wire together.'' Correlated activity in interconnected neurons maintains connections, whereas discorrelations remove those connections. This allows developing cortical neurons to respond in an adaptive way to variable features in the environment. The somatosensory cortex is normally subdivided so that groups of neurons process information from each finger or other subdivisions of the body surface. If a change in development alters the number of fingers or the number of sensory whiskers on the face, the areas of the cortex develop a matching number of modular subdivisions. In other words, brain regions develop to match each other in modular subdivisions that correspond to input from segregated groups of receptors. Each match is based on the transfer of information from one structure to another, rather than on the gradual selection for a long sequence of matches across nuclei and areas by the accumulation of changes in genetics. By using information from the environment, the rapid evolution of major alterations in the functional organization of the neocortex is possible. The monotreme mammal, the duck-billed platypus, has evolved a novel class of receptors on its bill, electroreceptors that are sensitive to weak electrical currents in water produced by the contracting muscles of swimming prey. The brain and neocortex adjusted to this new receptor by subdividing the primary sensory area, S1, into one set of modules related to tactile input and another set of modules related to electroreception. The developmental plasticity of the cortex allows useful adjustments for such changes in input, and the somatosensory system of monotremes was able to acquire new functions. The neocortex is designed not only to readily accommodate new or changed input originating from the receptor surfaces but also to accommodate input from functionally new classes of neurons that are created in the cortex by cortical circuits.

The ability of the neocortex to accommodate alterations in the nature of activating input by forming new circuits and modules may also relate to the evolution of new cortical areas. How the number of cortical areas increased in some lines of evolution is uncertain, but there are several logical possibilities. First, body parts sometimes duplicate in evolution so that suddenly extra parts emerge. A mutation in genes for some control mechanism could lead to the duplication of a cortical area. The two daughter areas could then be gradually differentiated by subsequent changes in genetic control, so that the two daughter areas come to mediate different functions. Another possibility is that new functions emerge as classes of modules are formed within areas. Gradually, over many generations, modules of each class could merge, reducing the lengths of the interconnections between modules. A full merger with complete separation of the two classes of modules would result in two areas emerging from one.

Whatever the mechanisms of cortical evolution, they work by modifying development. In order to

better understand the evolution of the neocortex, it will be productive to further investigate the great variability in neocortical organization that exists and the nature of the developmental changes that allow this variability.

Given the great potential for the neocortex to vary in size, accommodate changes in input, and subdivide into areas and modules specialized for functionally adaptive tasks, one might wonder why all mammals do not have expanded sheets of neocortex with many subdivisions. The major reasons for having less neocortex seem to be that the neocortex takes a long time to develop and mature and that the neocortex is metabolically costly. Other constraints undoubtedly also exist. For example, the weight of a large brain would be difficult to manage for bats as flying mammals, for whom all weight is costly. Mammals must be able to compensate for the metabolic and developmental costs of having more neocortex by having the gain in computational ability lead to getting better and more food and longer life spans, so that successful reproduction is improved or at least maintained. Humans, with exceptionally large sheets of neocortex, have used these sheets in ways that have allowed considerable reproductive success, with our species now exceeding 6 billion individuals, but most of this reproductive success has been exceedingly recent, over the last few hundred years. Great apes, as our closest relatives, have a large amount of neocortex, although much less than us. However, they have not done well and border on the verge of extinction. In contrast, small-brained rats and mice thrive over most parts of the earth. The rats and mice are not very bright and they make many mistakes, but they are able to replace themselves in large numbers very rapidly. However, mammals with little neocortex often specialized their small sheets of neocortex for particular functions. Thus, echolocating bats devote much of their small neocortex to auditory input and the sounds used for echolocating objects and prey. The large, costly sheets of neocortex that characterize the brains of higher primates remain unusual because they require long periods of safe development, long reproduction cycles, and consistently available foods of high metabolic value.

IX. SUMMARY

Although the neocortex is homologous with the dorsal cortex of reptiles, the neocortex is new with mammals as a thick, layered structure. The neocortex is an especially important part of the brain that varies greatly in structure and function across species. The neocortex can be changed in structure and function with training and experience so that members of the same species can become individuals. Several consistent features of the neocortex provide a framework with great structural and functional flexibility. These include the laminar organization of the neocortex with six basic layers of neurons with differing functional roles and patterns of connections, the existence of classes of neurons including small receptive neurons, large pyramidal output neurons, and local circuit inhibitory neurons, and the pronounced vertical connectivity of neurons across the depth of cortex. Species vary in amount of neocortex, types, numbers and sizes of areas, how areas disproportionately represent movements or sensory surfaces, modular organization within areas, laminar differentiation and specialization within areas, and patterns of connections within and across areas and with subcortical structures. The neocortex is a structure with great information-processing and -storing capacity. In humans and possibly other mammals, the neocortex mediates consciousness. Most importantly, the neural circuits in the neocortex are modifiable in development and throughout life, so that species become specialized for relevant perceptual and behavioral abilities, and individuals acquire individual skills, abilities, personalities, and memories.

See Also the Following Articles

BROCA'S AREA • CEREBRAL CORTEX • CINGULATE CORTEX • EVOLUTION OF THE BRAIN • FOREBRAIN • GABA • NEUROANATOMY • NEUROTRANSMITTERS

Suggested Reading

Creutzfeldt, O. D. (1993). *Cortex Cerebri, Performance, Structural and Functional Organization of Cortex*. Hubert and Co., Gottingen, Germany.

Kaas, J. H. (1987). The organization of neocortex in mammals: Implications for theories of brain function. *Ann. Rev. Psychol.* **38,** 129.

Krubitzer, L. (1995). The organization of neocortex in mammals: Are species differences really so different? *Trends Neurosci.* **8,** 408.

Mountcastle, V. P. (1998). *Perceptual Neuroscience: The Cerebral Cortex*. Harvard Univ. Press, Cambridge, MA.

Northcutt, R. G., and Kaas, J. H. (1995). The emergence and evolution of mammalian neocortex. *Trends Neurosci.* **18,** 373.

Peters, A., and Jones, E. G. (Eds.) (1984). *Cerebral Cortex, Volume 1: Cellular Components of the Cerebral Cortex*. Plenum Press, New York.

Rakic, P. (1995). A small step for the cell—a giant leap for mankind: A hypothesis of neocortical expansion during evolution. *Trends Neurosci.* **18,** 383.

Nerve Cells and Memory

PETER S. ERIKSSON
Sahlgrenska University Hospital

GLOSSARY

face agnosia The inability to recognize and remember faces.

neuronal stem cell A cell with the potential to give rise to all cell types found in the brain.

progenitor cell A cell with restricted potential to give rise to some cell types of the brain, e.g., astrocytes and oligodendrocytes.

prosopagnosia The inability to recognize and remember faces.

During the past century, neuroscientific research has generated a tremendous amount of knowledge on the function of the brain, but despite a substantial effort we still lack knowledge in such important areas as storage mechanisms for long-term memory, mechanisms responsible for higher intellectual functions such as reasoning, and decision-making processes based on judgment. There are, however, some areas in which it has been possible to localize behavior to small groups of neurons or even single nerve cells. Only a few mechanisms rely on spatially closely located neurons in the brain. Instead, many systems, such as the storage of long-term memories, seem to rely on diffuse, spatially nonlocalized circuitry throughout the brain. This article will focus on some examples for which the relationship between nerve cells and behavior has been partially unraveled.

I. THE BRAIN AND BEHAVIOR

The brain relays an enormous number of incoming and outgoing impulses every second. Input may stem from any of our senses, such as hearing, seeing, or feeling. More profound, and considered by the brain to be more important, information may result in the perceptions being stored as memories. These memories may be retrieved and perhaps used while processing subsequent incoming impulses. The result of this processing may be either rather simplified and stereotypical actions, such as reflexes (the knee jerk reflex), or more complex output, such as decisions based on judgment. The more complex output can also be termed behavioral output, or behavior. Although we consider these two examples of elicited behavior to be fundamentally different, they have many features in common. It is widely recognized that the lowest common denominator in all processing, including both simple and extremely complex behavior, is the nerve cell or the neuron. Behavior is thought of as the concerted signaling of an enormous number of neurons. In fact, it has been estimated that the human brain contains about 10^{11} neurons, classified into as many as 10,000 different types. One can easily understand the complexity within a system consisting of this number of signaling units, especially considering the fact that each neuron can have connections with perhaps thousands of other neurons. Another intriguing feature of the brain is the presence of glial cells. Cell number estimations suggest that the human brain may

contain as many as 10 times more glial cells than neurons.

II. THE BRAIN CONSISTS OF TWO MAJOR CLASSES OF CELLS

As indicated earlier, two primary cell types exist in the brain: nerve cells (neurons) and glial cells. Nerve cells, which have been the focus of neuroscientific research during the past century, have a number of basic common features. A typical neuron consists of a cell body containing the nucleus with the genetic material; neurons generally are highly metabolically active and thus contain relatively large amounts of energy producing mitochondria. Finally, the neuron usually has two types of processes: dendrites and an axon. The dendrites branch out in a treelike fashion and are generally considered to be the receiving components for input from other cells. The axon, on the other hand, is considered to be the output apparatus for the neuron, conveying output signals to other neurons. The axon's diameter can range from 0.2 to 20 μm, and it can be up to 1 m long, as are, for example, the motor neurons located in the cerebral cortex whose axons end in the spinal cord. Perhaps their most important feature is the neurons' ability to initiate and propagate an all-or-nothing transient electrical signal. This feature is called the action potential and is normally generated from the cell soma either as an independently initiated signal or as a result of incoming impulses (excitation). The action potential is initiated once a critical threshold (membrane potential) is reached.

The function of glial cells, whose name is derived from the Greek word glia, meaning glue, is less well-known. Whereas glial cells traditionally have been considered mainly as passive supporting elements providing structure to the brain, they may also provide metabolic support to surrounding neurons. There are two major classes of glial cells: oligodendrocytes and astrocytes. The oligodendrocytes produce the insulation that covers most large axons within the central nervous system. The astrocytes, on the other hand, provide metabolic support, act as scavengers, and perhaps more importantly buffer K^+ ions in the extracellular space. Another very important feature of the astroglial cells is taking up and metabolizing transmitters released by neurons during synaptic transmission. Glial cells also participate in forming the so-called blood–brain barrier. Finally, some classes of glial cells guide the migration of neurons and direct the outgrowth of axons.

Glial cells may play a very important role in scar formation after injury to the brain. Gliosis, or proliferation of glial cells, is an almost universal response to brain injury. It has even been proposed that gliosis after brain injury actually inhibits the reparative processes in the brain, in which neurons try to reestablish connections with their normal target regions. It is possible though that, by so doing, astrocytes help seal off damaged brain tissue after injury, thus inhibiting abhorrent reorganization within a damaged brain region.

It has been established that astroglial cells provide an extra neural signaling network via so-called gap junctions. Gap junctions are pores between individual glial cells that provide communication between glial cells for small molecules, up to 1000 kDa. Calcium signaling spreading within the astroglial network has been described in various model systems, including preparations of the intact brain. Although very little is known about the function of glial signaling, one can easily imagine the potential importance of a signaling network consisting of 10 times more cells than there are neurons within the brain. At present, the neurons are generally considered the most important components in the processing of sensory information, eliciting motor and emotional responses, and in learning and memory function. Although a large number of different neural types can be described on the basis of the transmitters the neurons release, what receptors they have, etc., researchers have been forced to reduce the degree of complexity by viewing neurons as more simple units, with either stimulatory or inhibitory output. By so doing, and by focusing instead on the interconnection between nerve cells, researchers have begun to emulate theoretically some of the processing capabilities of the brain. The deployment of several neural clusters or pathways to convey the same information is called parallel processing. This phenomenon has a number of advantages, including increased processing speed, increased reliability, and, of course, increased processing capacity. Scientists attempting to construct computational models of brain function have adopted this particular aspect of neural connectivity. Computer scientists initially used a serial-processing model to simulate higher levels of cognitive functions in the brain, such as the formation of memory, motor performance, and pattern recognition. It was soon realized that parallel processing allowed for the performance of these and other complex and demanding tasks. Thus, it is generally

believed that complex processing does not rely on the contribution of individual components but instead consists of the results of the interconnection of many elements. In this way the wiring between neurons provides the cellular basis for higher brain functions, such as face recognition, learning, and memory. In the following sections, we will give some examples of complex behavior resulting from activity in a small group of nerve cells.

III. NERVE CELLS AND PATTERN RECOGNITION

Face agnosia, or prosopagnosia, is a selective deficit in recognizing familiar faces. There are numerous case reports on patients suffering from this condition. The patients are aware they are looking at a face, but are unable to identify or even sense any familiarity in the faces of family members, co-workers, or other previously well-known individuals. Typically, these patients have difficulties remembering new faces. Patients with this condition normally rely on recognizing individuals by their clothing or their voices. Patients suffering from isolated prosopagnosia have raised the question of whether this condition is in fact due to selective damage to dedicated neural circuitry. There are patterns of deficits in processing face material in prosopagnosia that suggest that a number of dissociable cognitive operations are involved in processing face information. For instance, some patients retain their ability to judge the age, gender, and emotional expressions of faces whose identity remains completely unrecognized. It has been suggested that this is due to an inability to connect the obviously intact perception of face information to the stored biographical memories associated with a given facial structure necessary for identification. Prosopagnosia has been associated with damage to the right posterior cortex, and therefore face recognition was thought to be associated exclusively with neural circuits within the right posterior hemisphere. Studies on autopsy material have also suggested that bilateral damage may be necessary to induce prosopagnosia. Still, the identity and location of neurons responsible for face recognition in humans have yet to be firmly established.

Primate models of prosopagnosia suggest that the superior temporal sulcus contains a great concentration of ''face-selective'' neurons. This area is, however, well-separated from the region one would suppose to be homologous to the site of damage revealed in humans with prosopagnosia. Furthermore, bilateral ablation of the superior temporal sulcus in monkeys clearly fails to produce more than a minor decrease in face recognition. Face-selective neurons located in more ventral portions of the temporal lobe in ablation experiments have been shown to play an important role in face recognition. Removal of this area of the cortex causes severe impairments in face perception. Furthermore, it also perturbs the perception and recognition of a large number of other classes of objects.

One possible explanation for the seemingly different ways in which monkeys and humans process face recognition is that the neural substrates for recognizing faces and other classes of stimuli are not segregated as completely in monkeys as in humans. Functional neuroimaging studies suggest that the ventral occipitotemporal cortex of the right hemisphere is primarily responsible for the perceptual analysis of faces. More anterior temporal regions are then activated for the association of face information with stored information about individuals. The frontal cortex appears to contribute both to the analysis of emotional facial expression and to the maintenance of faces within the short-term memory store. Interestingly, electrophysiological experiments on patients going through presurgical epilepsy investigation have shown that large potentials are generated by faces and not by other categories of stimuli at very small sites within the ventral occipitotemporal cortex. These small groups of ''face-activated'' neurons are thought to be the human counterparts to the face-selective neurons found in monkeys. The analyses of the brain's mechanisms for face recognition thus suggest that single or small groups of neurons may in fact be responsible for distinct aspects of face recognition. These neurons therefore might be considered discrete units, possibly providing singular perceptions of complex stimuli at the level of individual nerve cells.

IV. LEARNING AND MEMORY

One structure that has been associated with memory formation is the hippocampus. The hippocampus is an area of the allocortex hidden within the temporal lobe, where it resides in the medial portion. The hippocampus is connected to a number of adjacent cortical areas. The areas affiliated with the hippocampus include the perirhinal, entorhinal, and parahippocampal cortex. Hippocampal injury in humans results in deficits in episodic memory. Memories dependent on the hippocampus also include memories for spatial material. There are case reports on patients with hippocampal

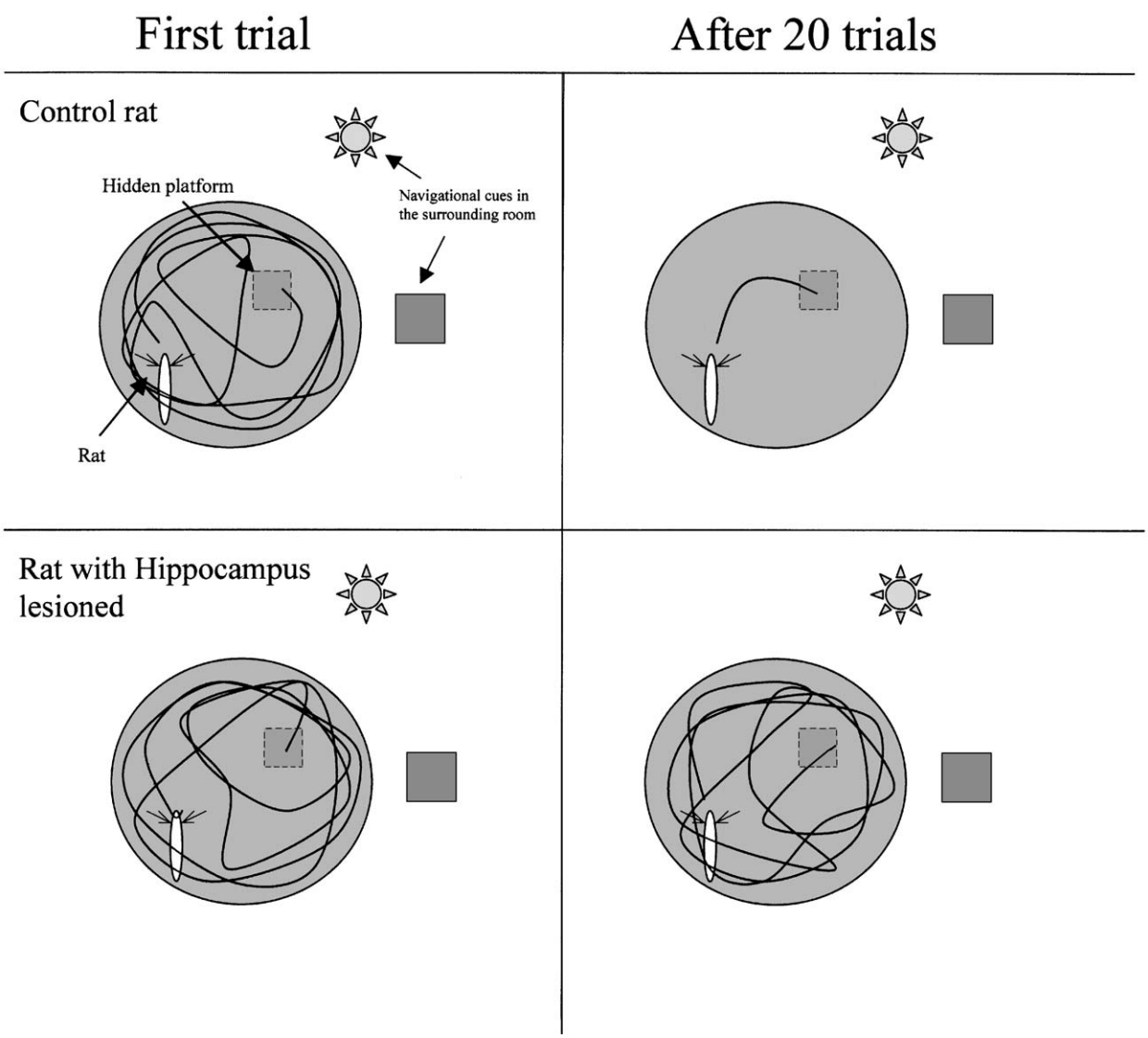


Figure 1 Spatial memory testing in rats. Rats are placed in a circular water pool filled with cloudy water. A small platform is placed at a fixed location just under the surface of the water. The pool has no any specific features but the surrounding environment usually contains a number of navigational cues (lamps, windows, doors, etc.). As the rats search for the platform, their swim path (and speed) is monitored by a video camera (indicated by the traces in the figure). After a number of trials normal rats swim directly to the platform. Rats with impaired spatial memory due to a lesion on the hippocampus fail to learn where to find the hidden platform.

injuries due to temporal lobe surgery, who suffer from an inability to learn how to find their way in unfamiliar neighborhoods (Fig. 1).

V. NERVE CELLS AND MEMORY

As mentioned earlier in this article, most aspects of memory are dissociable processes mediated by distinct and separate brain systems. Although we are still far from resolving all aspects of memory formation, converging evidence from psychology and neuroscience point to at least five major systems in humans. These are episodic memory, working memory, semantic memory, the perceptual representation system, and procedural memory (Fig. 2).

Episodic memory is the recollection of specific incidents that occurred at a particular time and place during a person's past. Researchers study episodic memory formation by asking questions about specific

Human memory formation

System:	Neuronal circuitry:
Episodic memory	Hippocampal formation Prefrontal cortex
Working memory	Hippocampal formation (possibly left temporal lobe)
Perceptual representation system	Extrastriate occipital cortex Temporal lobe and frontal cortex
Semantic memory	Lateral regions of the temporal lobe (left)
Procedural memory	Corticostriatal systems and possibly also the cerebellum

Figure 2 Converging evidence from psychology and neuroscience point to at least five major memory systems in humans with somewhat different anatomical locations.

information acquired at a certain time and place in the past. In this manner, it has been established that damage to the medial parts of the temporal lobes, especially the hippocampus, greatly impairs the acquisition of new episodic memories. Patients with acquired damage to the hippocampal formation invariably suffer impairment of episodic memory, only the degree of damage changes. This means that they have difficulty remembering events during their daily life. Bilateral hippocampal damage may also affect the ability to acquire new memories. Episodic memory also appears to rely on the prefrontal cortex, which seems to provide temporal information on episodic memories. Patients with damage to the prefrontal cortex may have severe difficulty in remembering when and even where recent events occurred. Frontal lobe damage also often produces what is known as source amnesia, which means that patients readily acquire new facts but fail to recollect where or when they learned them. Damage to the frontal lobes, which can, for instance, result from severe alcohol abuse, can also produce distortions of episodic memory in which patients may even claim to remember events that never occurred. Episodic memory studies have been performed while monitoring subjects with positron emission tomography (PET scanning) in order to measure regional cerebral blood flow. These studies confirm that the frontal lobe is activated during episodic memory tasks. Studies using PET scanning have also shown that the temporal lobe and the hippocampus are activated during both the encoding and retrieval of memories.

Working memory is a memory system used for short-term retention, operating over periods of seconds. Working memory is used during reasoning, problem solving, and comprehending. The concept of working memory emerged among other things from studies of patients suffering from amnesia due to bilateral hippocampal damage. Such patients may for instance be able to immediately remember small numbers of digits. Studies on patients with hippocampal damage therefore have revealed that a separate memory system for short-term memory exists. Case reports on patients with selective deficits in working memory have suggested that this memory system may rely on a small group of specific neurons situated in the supramarginal gyrus of the left temporal lobe.

The third memory system, semantic memory, refers to general knowledge of facts and concepts without any apparent connection to time or place. Semantic memory is important for knowing that Hawaii is in the Pacific Ocean, whereas episodic memory is used for remembering a certain visit to Hawaii. Semantic memory also relies on the function of the medial temporal lobes. The acquisition of new memories, both episodic and semantic, depends on the integrity of the medial temporal lobe and is collectively referred to as the declarative memory. The semantic and episodic elements composing declarative memories are thought of as dissociated from each other. Case reports indicate that patients with damage to the anterior part of the temporal lobe have relatively little difficulty remembering specific past episodes but more pronounced difficulties with knowledge of historic events. On the other hand, the study of patients with damage to the more lateral regions of the temporal lobe, particularly in the left hemisphere, suggests that these regions play important roles in the semantic memory system, thus implying that limited neural circuitry may in fact be responsible for the integrity of the declarative memory function.

Another memory system is generally called the perceptual representation system. This memory system plays an important role in the identification of words and objects based on their structure and form. The perceptual representation system operates at a presemantic level and is not involved in semantic memory. PET scanning studies have suggested that specific regions within the extrastriate occipital cortex

are involved in processing and representing the visual form of words, whereas regions within the temporal lobe and the frontal cortex process the meaning of words. The perceptual representation system plays an important role in the phenomenon called priming. In priming experiments, the subject studies a word or a picture and subsequently is presented with only a part of the same item. Priming means that the subject's ability to identify the partial items improves following recent exposure to the entire item. This phenomenon appears to operate unconsciously, because it can be observed in subjects who were primed under conditions resulting in their lack of explicit memory of having studied the object or the word. Amnesic patients (patients with bilateral temporal lobe damage) exhibit intact priming across a wide variety of paradigms. This suggests that the perceptual representation system does not rely on temporal structures. Functional imaging has suggested that posterior cortical regions may be responsible.

Procedural memory refers to the acquisition of skills, characterized by the ability to perform a specific task or habit. Such memories include, for instance, knowing how to ride a bike or more explicit knowledge of sequences or grammatical rules. Studies of amnesic patients have revealed that, although unable to explicitly remember past experiences, these patients can gradually acquire new perceptual, motor, or cognitive skills. These studies suggest that the acquisition of procedural knowledge is independent of the medial temporal lobe. Instead, procedural memory seems to depend heavily on corticostriatal systems, e.g., patients with Huntington's disease or Parkinson's disease may have difficulty in acquiring new motor skills without having any deficits in explicit memory. Functional imaging studies have suggested that both the motor cortex and the basal ganglia are involved in procedural learning. Interestingly, the cerebellum also appears to be involved in procedural memories, as demonstrated by the fact that patients with cerebellar damage may have difficulty learning to execute sequences of movements.

VI. MEMORY FUNCTION IN NORMAL AGING AND DEMENTIA

It has traditionally been accepted that moderate neural death occurs as an inevitable consequence of normal aging. Significant age-related neural loss has been reported in almost every region examined. Studies involving unbiased quantitative techniques for estimating cell numbers in tissue samples have led to a significant revision of traditional views on age-related neural loss. This is due to the fact that most investigators had been focusing their attention on cell density, measured as the number of neurons in a fixed volume of tissue within a brain region. The counting was typically accomplished using standard histological staining procedures to visualize and count cells microscopically. If a strategy like this is employed without taking into account the fact that gliosis within a brain region may actually alter the size of a brain structure, it can lead to the possibly erroneous conclusion that the total number of neurons is decreased. Instead, studies taking into account volumetric differences between young and aged brains have led to a reevaluation of the traditional view on age-related neural loss. Modern stereological tools have been widely used to investigate neural numbers in the aged hippocampus. Earlier studies had suggested that the hippocampus is especially susceptible to age-related cell death and that this may be the explanation for decreases in hippocampal-dependent learning and memory within aged subjects. Instead, investigations have revealed that the number of granule cells of the dentate gyrus and pyramidal neurons within fields CA2, CA3, and CA1 remain unaffected in the aged hippocampus (Fig. 3). Studies also suggest that the number of hippocampal neurons remains normal even among aged individuals with pronounced learning and memory deficits, which are normally associated with hippocampal dysfunction. Other studies have even suggested that granule cells within the dentagyrus are formed continuously throughout life. Instead, age-related neural loss associated with memory deficits preferentially targets subcortical brain systems. Aging does result in substantial subcortical cell loss, especially among neurons with ascending projections to cortical regions. The loss of cholinergic neurons in the basal forebrain has been studied extensively because of its assumed role in Alzheimer's disease. A hallmark of this degenerative condition is an accelerated degeneration of acetylcholine-containing neurons affecting cell groups that project to the hippocampus, the amygdala, and the neocortex. The loss of cholinergic cells might disrupt the circuitry responsible for information processing in these target regions. The number of remaining cholinergic neurons correlates with a magnitude of behavioral impairments in dementia patients and aged individuals. Although the loss of cholinergic neurons may not be directly responsible for the cognitive deficits seen in these patients, it probably indirectly affects other neurochemically specific

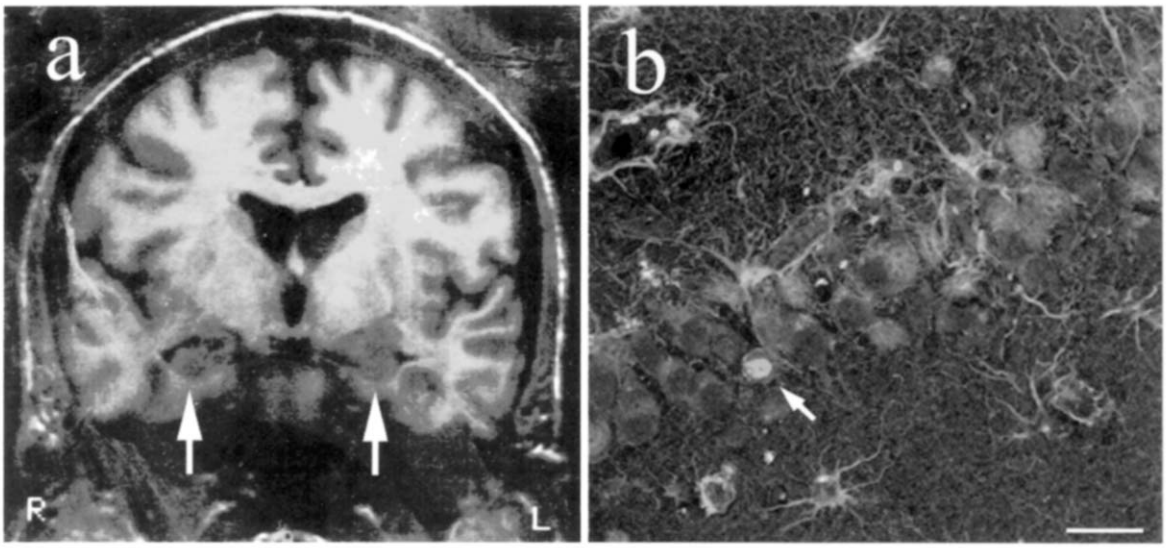


Figure 3 (a) MRI image of the human brain showing the location of the hippocampi (arrows) within the medial portions of the temporal lobe. (b) Confocal image of granule cells within the human dentate gyrus (hippocampus). The arrows show the nucleus of a newborn, NeuN immunoreactive, neuron within the adult human brain. Astroglial cells are immunostained for glial fibrillary protein. From Eriksson, P. S., *et al.* (1998). Neurogenesis in the adult human hippocampus. *Nature Med.* **4,** 1313–1317. (See color insert in Volume 1).

projection systems responsible for memory and learning. Thus, cholinergic neurons within the basal forebrain represent an example of a limited number of neurons that have a pronounced impact on behavior.

VII. MODELS FOR THE ANALYSIS OF NERVE CELLS AND SPATIAL MEMORY

In animal models using rodents, neural electrical activity can be monitored in the hippocampus while the test animal is running at arm radial mazes and performing tasks that require locomotion within a workplace. Experiments using this paradigm have revealed a remarkable degree of spatial selectivity. A surprisingly high number of neurons throughout the hippocampus have place fields. This means that a specific neuron within the hippocampus will fire most strongly while the rat is in a certain area of the workspace. In this way, it has been shown that different neurons within the hippocampus cover the workspace made up of the testing area. There are also some hippocampal neurons that are sensitive not only to the rat's location but also to the direction in which its head is pointing. This feature is most pronounced within the postsubiculum, an area closely linked to the hippocampus. Experiments have shown that these postsubicular neurons fire when the rat's head is pointing in the neuron's preferred direction and conversely decrease the firing rate as the orientation of the rat's head deviates from that direction.

Another example of a contribution from specific neural circuitry is new granule neurons formed during adulthood. Neuronal stem cells or progenitor cells residing within the dentate gyrus in the hippocampus continuously generate granule neurons throughout life. This phenomenon has been was shown to be under dynamic regulation. Increased hippocampal volume has been reported in birds and animals that engage in behavior requiring spatial memory, such as food storing. Another study demonstrated navigation-related structural changes in the hippocampus of taxi drivers, with an increased volume of the posterior hippocampus correlating to the time spent as a taxi driver. Laboratory animals housed in an enriched environment have an increased rate of neurogenesis within this region as compared to rats living in a standard environment. Furthermore, animals living in an enriched environment perform better on spatial memory tests. Running also induces this phenomenon and selectively enhances dentate gyrus long-term potentiation (LTP). Newborn granule cells within the adult rat hippocampus have been shown to project axons through the mossy fibers to their natural targets

within the CA3 region. It has, therefore, been speculated that newborn granular cells may contribute to the improvement in spatial memory test performance. The direct link between neurogenesis within the dentate gyrus and spatial memory performance, although compelling, still needs further investigation.

See Also the Following Articles

AGNOSIA • BEHAVIORAL NEUROGENETICS • COGNITIVE AGING • DEMENTIA • MEMORY, EXPLICIT AND IMPLICIT • MEMORY NEUROBIOLOGY • NEUROGLIA, OVERVIEW • PATTERN RECOGNITION • SEMANTIC MEMORY • SPATIAL COGNITION • WORKING MEMORY

Suggested Reading

Gallagher, M., and Colombo, P. (1995). Aging: The colinergic hypothesis of cognitive decline. *Curr. Opin. Neurobiol.* **5,** 161–168.

Kandel, E. R., Schwartz, J. H., and Jessel, T. M. (Eds.) (1991). *Principles of Neural Science*. Elsevier Science, New York.

Purves, D., Augustine, G. J., Fitzpatrick, D., Katz, L. C., LaMantia, A.-S., and McNamara, J. O. (Eds.) (1997). *Neuroscience*. Sinauer Ass. Inc., Sunderland, MA.

Tovee, M. J. (1998). Is face processing special? *Neuron* **6,** 1239–1242.

Wilsson, M. A., and Tonegawa, S. (1997). Synaptic plasticity, place cells and spatial memory: Study with second generation knockouts. *Trends Neurosci.* **20,** 102–106.

Zigmond, M. J., Bloom, F. E., Landis, S. C., Roberts, J. L., and Squire, L. R. (Eds.) (1999). *Fundamental Neuroscience*. Academic Press, San Diego.

Nervous System, Organization of

JAY B. ANGEVINE, Jr.
The University of Arizona, Tucson

GLOSSARY

autonomic nervous system (ANS) Efferent part of the nervous system supplying motor innervation to the viscera. Has three divisions: sympathetic (thoracolumbar), parasympathetic (craniosacral), and enteric.

axon Usually a single, long, thin, and often branched process arising from the cell body or major dendrite of a neuron, conducting nerve impulses (action potentials) to other neurons, muscles, or glands.

cell body (soma) Plump, spherical part of the nerve cell; includes the nucleus and surrounding organelles. Indispensable, irreplaceable trophic (nutritive, sustaining) center, houses genetic history of the neuron, maintains integrity of the cell and all its far-reaching processes.

central nervous system (CNS) The brain and spinal cord; the neuraxis, consisting of seven bilaterally organized, largely symmetrical, functionally specialized regions: the cerebral hemispheres, diencephalon, midbrain, cerebellum, pons, medulla oblongata, and spinal cord.

cortex Extra gray matter external to white matter in the cerebral hemispheres, roof of the midbrain, and cerebellum. Nerve cell bodies are arrayed in distinct layers; nerve fibers (afferent, intrinsic, and efferent) are arranged in orderly, geometric ways that facilitate integrative, analytic, and comparative functions.

dendrites Short, highly branched, tapered, often spiny processes extending from the nerve cell body or, in sensory neurons, from the distal end of an interposed axon; chief receptive surface of the neuron.

ganglion Spherical or fusiform cluster of sensory nerve cell bodies in the PNS or of small motor nerve cell bodies in the ANS. Has connective tissue investments like those of adjoining nerves.

gray matter Inner core of CNS, containing nerve cell bodies (somata), dendrites, the proximal extents of axons, which distally enter white matter, neuroglial cells, and blood vessels (especially capillaries).

interneurons Originally defined as all neurons lying between sensory and motor neurons, comprising most neurons in the CNS; tiny or huge, linking neurons receiving stimuli from sense organs to those innervating muscle. Now used much more restrictively for neurons intrinsic to specific CNS regions.

motor neurons The ultimate neurons or "final common path" (Sherrington) over which neural activity passes to muscles and glands; the large multipolar anterior horn cells in spinal gray and similar cells in the brain stem (hypoglossal nucleus). Paltry in number (about 2 million), profound in importance.

nerve In the PNS, a bundle or bundles of myelinated and unmyelinated fibers with supportive, protective, and nutritive connective tissue investments (endoneurium, perineurium, epineurium). Like a CNS tract, but serving a wider range of functions (sensory and motor, somatic and visceral, etc.).

neuroglia Singular noun meaning "nerve glue." One of two types of cells intrinsic to the CNS, the other being neurons. The neuroglia (astrocytes, oligodendrocytes, ependymal cells, microglial cells) outnumbers neurons 10–50 times and makes up half the volume of the CNS. More than offering structural support, neuroglial cells surround, protect, nurture, and assist neurons in important ways as they perform their integrative and communicative functions.

neuron Nerve cell, highly diversified in size, shape, and chemical properties. The human nervous system has over 100 billion, perhaps 1 trillion neurons (25 billion in the cerebral cortex alone).

nucleus Functional cluster of nerve cell bodies in CNS gray matter (not the nucleus of a cell). Their axons and dendrites frequently

Dedicated to the memory of David Bodian and Walle J. H. Nauta.

extend beyond the arbitrary nuclear boundary and their component cells exhibit differences in size and other structural features, as well as in the functions served.

peripheral nervous system (PNS) The peripheral nerves (cranial and spinal) with their sensory and motor roots, sensory ganglia, and complicated nerve plexuses.

sensory (or primary sensory) neurons Initial neurons (about 20 million) in sequential sensory data processing. Derived chiefly from neural crest, their bipolar or pseudounipolar cell bodies lie outside the CNS in the sensory ganglia of craniospinal nerves and in the olfactory mucous membrane.

synapses Intimately apposed regions (10- to 20-nm gap; 10^{14} or more) of contact between neurons or between neurons and muscle fibers, which permit functional transactions and, by allowing one-way transmission only, help determine the forward impulse traffic in the nervous system as a whole.

tract Cablelike bundle of myelinated or unmyelinated axons in CNS white matter, running from place to place, often over long distances (up to 1 m in humans) and representing a functional group of nerve fibers. Named functionally (optic tract), descriptively (medial forebrain bundle), binomially as to origin and termination (spinocerebellar tract), or analogously (lemniscus, L. for ribbon).

white matter Outer zone of CNS, surrounding the ever-present gray core; contains vast numbers of axons of various caliber and myelination, derived from nerve cell bodies in subjacent gray or PNS sensory ganglia, neuroglial cells, capillaries (fewer than in gray matter), and larger blood vessels.

The human nervous system is a hierarchy, culminating in the brain, of 100 billion or more neurons of 10,000 types, 1–10 trillion neuroglial cells, 100 trillion chemical synapses, 160,000 km of neuronal processes, thousands of neuronal clusters and fiber tracts, hundreds of functional regions, dozens of functional subsystems, 7 central regions, and 3 main divisions. All of these parts form a coherent, bodily pervasive, diversified, complex *epithelium* with interdependent connectivity of neurons, mostly neither sensory nor motor but anatomically and functionally intermediate. The key organizing principles of the system are *centralization* and *integration*. The nervous system performs two roles: *regulation* and *initiation*. In the first, it *counteracts*: responsively and homeostatically, gathering stimuli from outside and inside the body (including the brain), assessing their short-term and long-range significance, generating activity from faster breathing to stock trading, even to functional plasticity in learning or after brain damage. In the other, it *acts*: endogenously, not so homeostatically, replacing one state of neural activity with another, generating activity from doing nothing at all to creative thinking and extraordinary achievement, even taking steps toward understanding how itself, the nervous system, works. Although the divisions and regions of the nervous system are identical in all normally developed humans, their genetic specification and personal history are unique, as are the permutations and combinations of their unified function. Each human nervous system is unprecedented. The work of each (expressed or hidden) is unpredictable, ever-different, surprising, startling, at times horrifying, but not infrequently magnificent.

I. PRINCIPAL DIVISIONS OF THE NERVOUS SYSTEM

A. Central Nervous System

The central nervous system (CNS) comprises the brain and spinal cord (Fig. 1). Derived from five vesicles in the cranial part of the embryonic neural tube, the brain has five bilaterally represented, largely symmetrical, functionally specialized regions: cerebral hemispheres, diencephalon, midbrain, pons, and medulla oblongata (Fig. 2, Table I). Each has a central hollow, large or small. All are interconnected as a four-chambered ventricular system (Fig. 3) with a connecting aqueduct. In its size and specialization, the cerebellum, inseparably a part of the pons, deserves status as a sixth region. Figure 4 illustrates the embryonic and fetal development of all these regions.

The spinal cord is a tubular structure (Fig. 1). Its central canal, a remnant of the lumen of the neural tube, is no longer patent (Fig. 18). Continuous with the medulla oblongata, it has 31 segments: 8 cervical, 12 thoracic, 5 lumbar, 5 sacral, and 1 coccygeal. Segments are spinal regions served by a pair of spinal nerves. Reflecting the segmental development of the body, they are functionally and clinically important. No other demarcation of segments exists: an epigraph for the spinal cord is "continuity." Extending 43–45 cm from the base of the skull to the lumbar spine, the spinal cord mediates all sensations and movements of all parts of the human body, except those served by cranial nerves.

B. Peripheral Nervous System

The peripheral nervous system (PNS) comprises the cranial and spinal nerves (Fig. 1), with their associated roots and ganglia. Through these nerves, sensory impulses come to the CNS and motor impulses go to muscles and glands. Like the cranial structures they

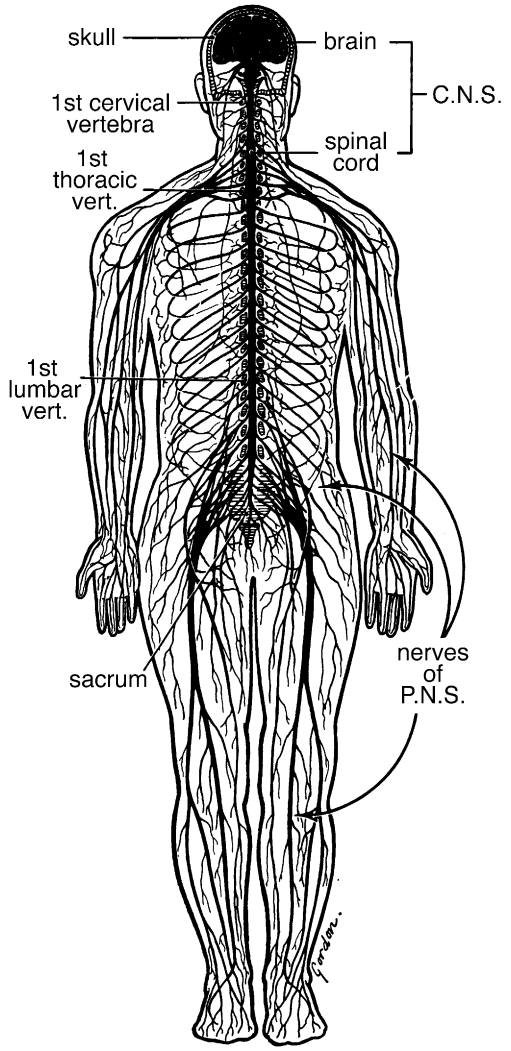


Figure 1 Basic subdivisions of adult nervous system: central nervous system (CNS) and peripheral nervous system (PNS). From Han, A. W., *Histology*, 5th ed., J. B. Lippincott, 1961 (illustration by Louise Miller).

innervate, most of the 12 cranial nerves are specialized. Some, like the optic nerve, are purely sensory, whereas others, like the abducens, are purely motor. Still others, like the vagus, are mixed sensory and motor (Fig. 5, Table II).

By contrast, the spinal nerves are much alike, providing a basic segmental pattern of sensory and motor innervation for the rest of the body, including the limbs. In the cervical, brachial, lumbar, and sacral regions, nerve plexuses interpose between the spinal cord and the nerves themselves. In these complex webs, nerve fibers leaving the spinal cord segmentally sort out and recombine into specific nerves. This involves a process of diverging and converging roots, trunks, divisions, cords, and branches, eventually combining in nerves, such as the phrenic, radial, femoral, and sciatic, to cite prominent nerves derived from each plexus in turn. These and other nerves continue to ramify.

C. Autonomic Nervous System

The autonomic nervous system (ANS) is an involuntary division for maintaining homeostasis (Fig. 6). It sends motor fibers to the viscera, blood vessels, sweat glands, arrector pili, pupillary smooth muscles, etc. and regulates heart rate. It features two small visceral motor neurons in tandem: a preganglionic neuron with its cell body in the CNS and a postganglionic neuron in an autonomic ganglion. Both have meager dendritic trees and thin, lightly myelinated, slow-conducting axons that travel in peripheral or autonomic nerves to smooth muscle in a viscus, blood vessel, or gland.

The ANS has three subdivisions: *sympathetic* (thoracolumbar), *parasympathetic* (craniosacral), and *enteric*. In the first, preganglionic nerve cell bodies lie in thoracic and upper lumbar segments of the spinal cord and postganglionic ones in the paravertebral sympathetic chain ganglia or the prevertebral ganglia (celiac, superior mesenteric, and inferior mesenteric). In the second subdivision, preganglionic nerve cell bodies lie in the brain stem (in cranial nerve nuclei III, VII, IX, and X; see Table II) and in sacral spinal cord segments 2–4, with postganglionic ones in or near the organs innervated. In the enteric component, postganglionic neurons lie in the alimentary wall in ganglia and intramural plexuses. These nets contain 100 million neurons. They function almost autonomously, subject to control and override by preganglionic parasympathetic and sympathetic neurons.

The CNS receives visceral sensory axons not included in the ANS as originally defined. Visceral afferents, from mechanoreceptors, chemoreceptors, and nociceptors, are poorly understood but important to homeostasis and behavior. The sympathetic part of the ANS forms a distinct pair of cords in the PNS: the sympathetic chain ganglia alongside the spinal cord. The parasympathetic part, by contrast, is less obvious as its fibers are components of cranial and sacral nerves.

The sympathetic subsystem is the "fight or flight" component. When energy must be burned, it acts

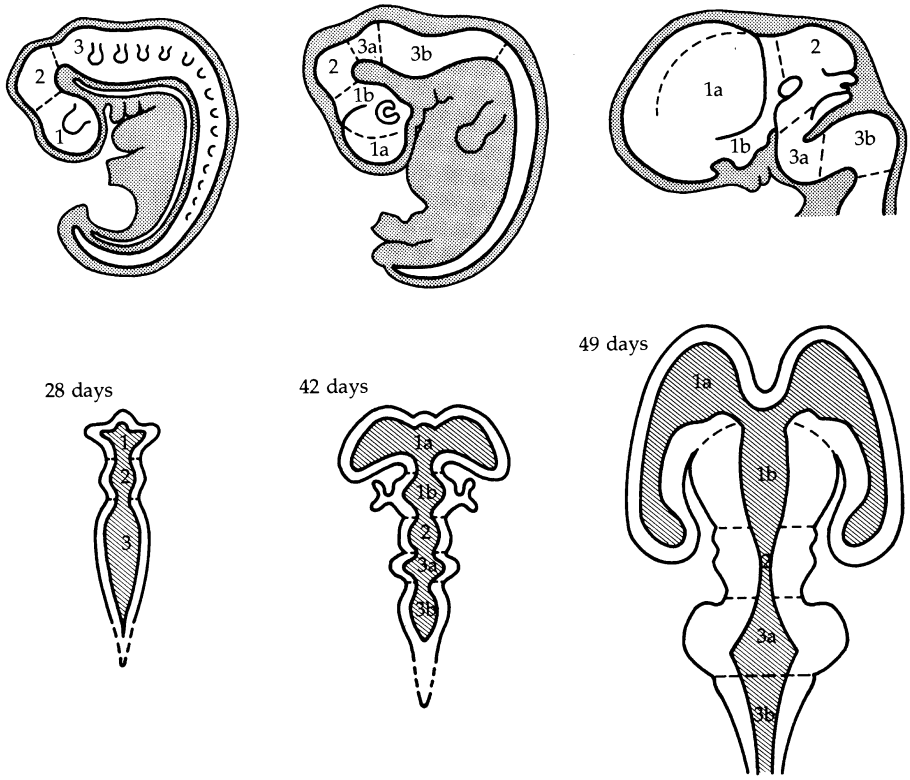


Figure 2 Basic subdivisions of embryonic nervous system: at 28 days, prosencephalon (1), mesencephalon (2), rhombencephalon (3). At 42 and 49 days, telencephalon (1a), diencephalon (1b), mesencephalon (2), metencephalon (3a), myelencephalon (3b). From *Principles of Neuroanatomy* by J. B. Angevine, Jr., and C. W. Cotman, copyright 1981 by Oxford University Press, Inc. Used by permission of Oxford University Press, Inc. (illustration by Steven J. Harrison).

almost instantly in a concerted manner and over some duration. Heart rate goes up, blood vessels in skeletal muscles dilate, as do respiratory bronchioles and pupils, suprarenal glands kick in massive secretions of epinephrine and norepinephrine, hepatic glycogenolysis provides glucose for energy, hairs stand on end, and the mouth is suddenly very dry. The parasympathetic subsystem functions less dramatically. It is the

Table I
The Main Subdivisions of the Embryonic CNS and Their Adult Fates

Three-vesicle stage	Five-vesicle stage	Adult derivatives
1. Prosencephalon (forebrain)	1a. Telencephalon (endbrain)	Cerebral hemispheres, lateral ventricles, basal ganglia, corpus callosum
	1b. Diencephalon (tweenbrain)	Thalamus, hypothalamus, third ventricle, optic nerves and tracts, retinae, pineal gland
2. Mesencephalon (midbrain)	2. Mesencephalon (midbrain)	Superior and inferior colliculi, cerebral aqueduct, cerebral peduncles, midbrain tegmentum
3. Rhombencephalon (hindbrain)	3a. Metencephalon (afterbrain)	Cerebellum, rostral part of fourth ventricle, pons, pontine tegmentum
	3b. Myelencephalon (cordbrain)	Medulla oblongata, caudal part of fourth ventricle, medullary tegmentum
4. Remaining caudal part of neural tube	4. Remaining caudal part of neural tube	Spinal cord (myel, from Greek, "marrow") and central canal

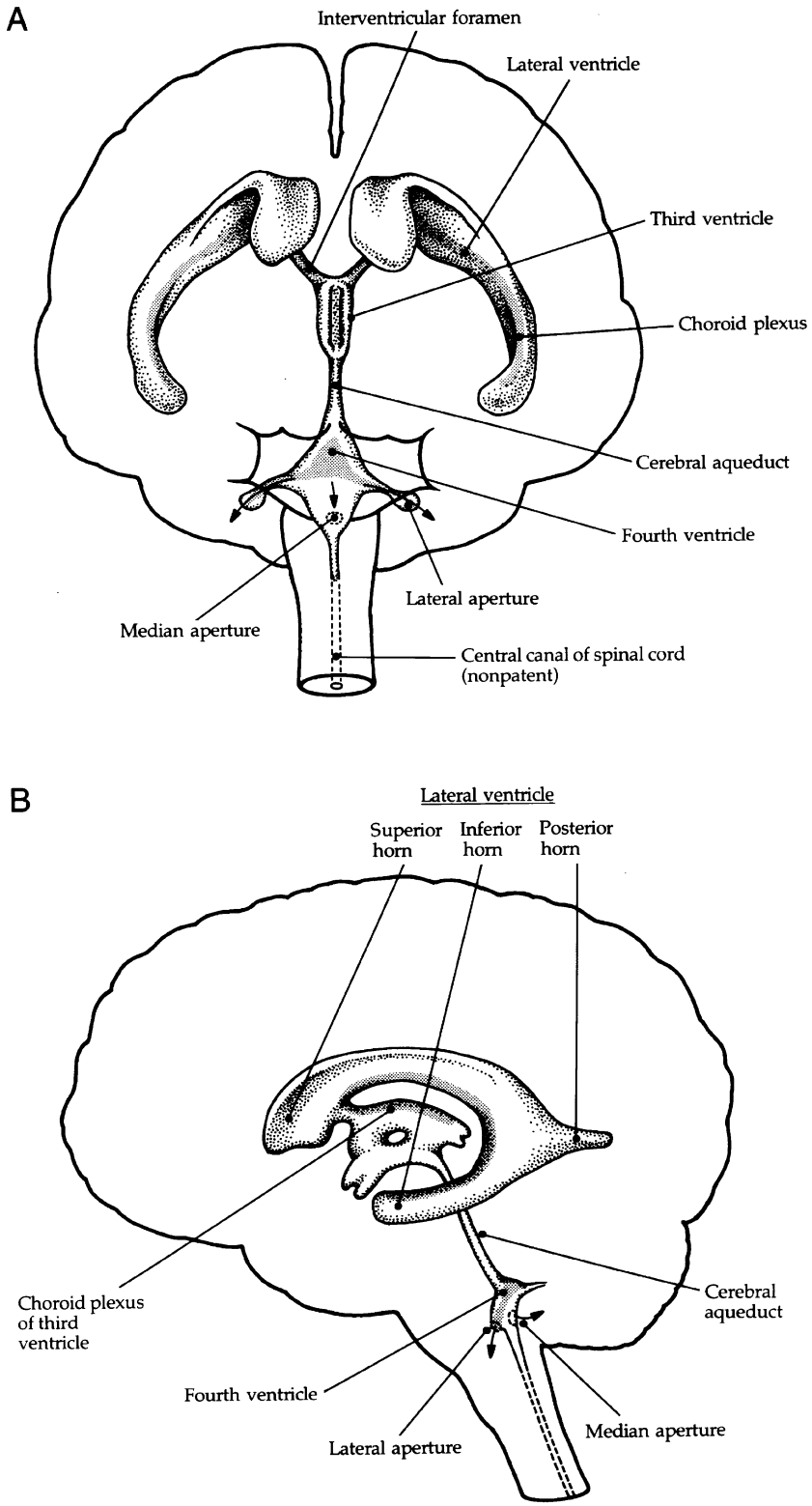


Figure 3 Brain ventricular system: the cerebral hemispheres, diencephalon, midbrain, pons, and medulla oblongata have central cavities, interconnected as a four-chambered ventricular system with a connecting aqueduct. Three apertures in the fourth ventricle allow cerebrospinal fluid (CSF) to exit the system into the subarachnoid space. The lumen of the embryonic spinal cord is no longer patent. From *Principles of Neuroanatomy* by J. B. Angevine, Jr., and C. W. Cotman, copyright 1981 by Oxford University Press, Inc. Used by permission of Oxford University Press, Inc. (illustration by Maureen Killackey).

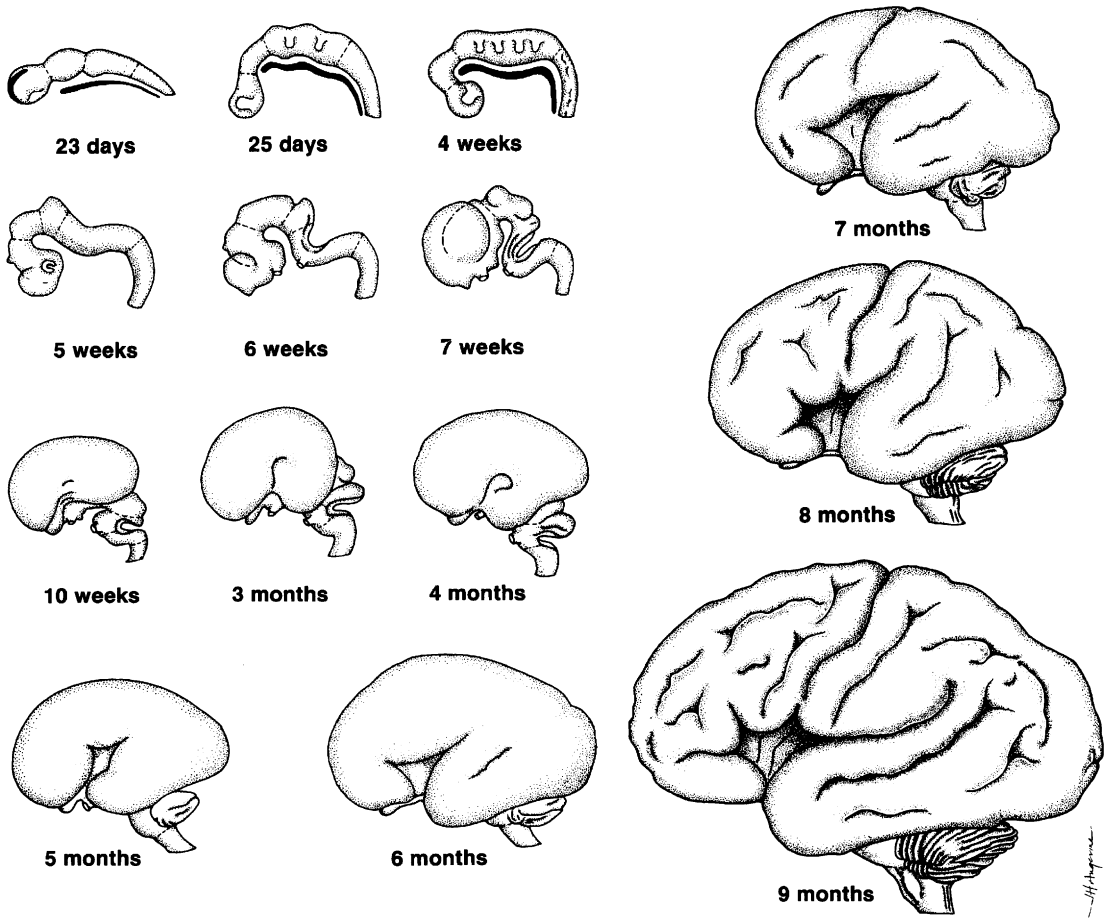


Figure 4 Embryonic and fetal brain development: flexures and differential expansion of its chambers, along with cerebral fissuration, bring about the definitive appearance of the brain. From J. B. Angevine, Jr., Morphogenesis of the Central Nervous System, *BNI Quarterly*, Vol. 5, No. 4, 1989 (illustration by Janice H. Angevine, modified from Carol Donner).

"rest and replenishment" component. When activity brings fatigue and energy must be restored, it goes about its business slowly, quietly, selectively. Heart rate goes down, blood vessels stay the same, bronchioles and pupils constrict to normal size, and digestive, urinary, and reproductive systems proceed with normal function.

Usually, sympathetic and parasympathetic divisions act separately and at times cooperatively. Sweat glands and blood vessels in the limbs have only sympathetic innervation; the pupil and bladder are dominated by parasympathetic fibers. Certain disparate domains see both divisions working together: in never-ending modulation (with neurohormonal assistance) of cardiac rhythmicity and intermittent assistance in male sexual function (erection is mediated mainly by parasympathetic fibers but ejaculation by sympathetic fibers). Visceral and somatic functions may also be performed cooperatively. This requires the interplay of many parts of the CNS: spinal cord, brain stem, limbic system, hypothalamus, basal ganglia, and others. Performances of respiratory, digestive, and sexual functions offer eloquent illustrations of concerted viscerosomatic activity.

Visceral sensations may intrude deeply upon thoughts and feelings. Often difficult to describe and localize, they are usually unpleasant and at times uncompromising. As biofeedback studies show, the ANS can be classically conditioned. Once activated, it plays a pivotal role in human decision making. It now receives major attention in studies of normal, abnormal, and criminal behavior, with benefit to informed family counseling, crisis mediation, law enforcement, corrections, and society at large.

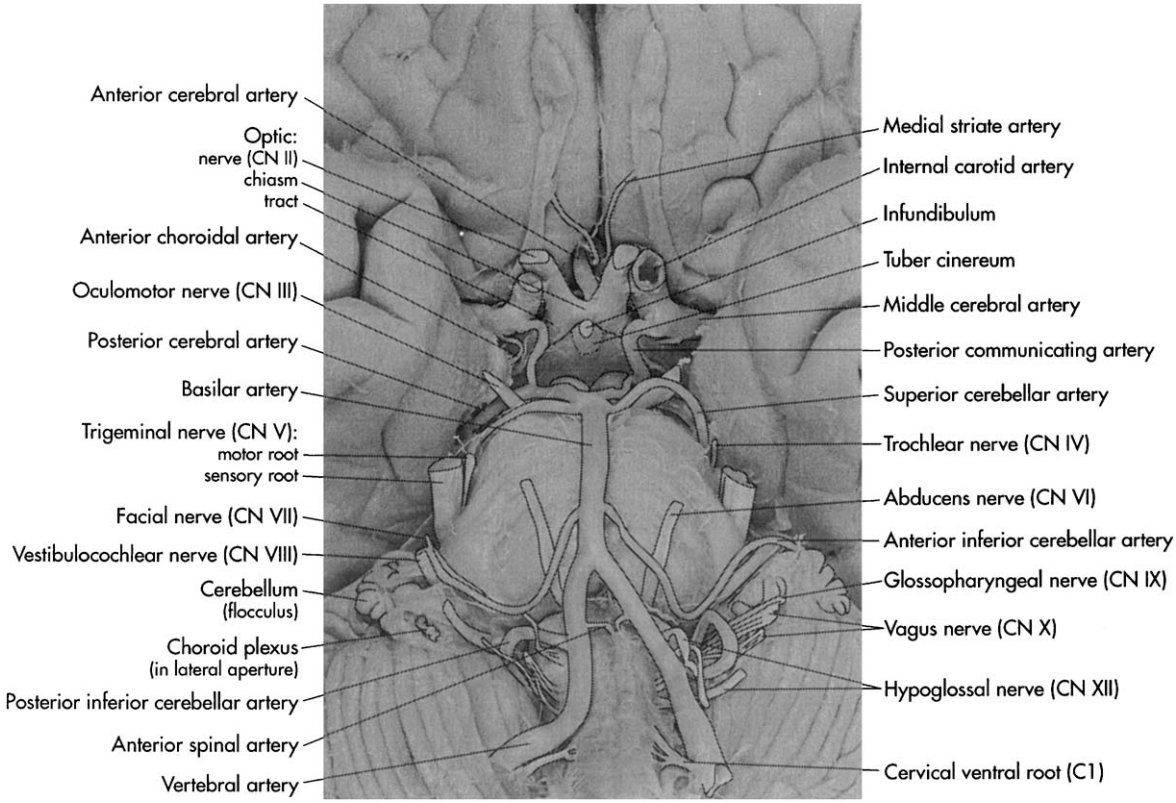


Figure 5 Base of forebrain and inferior view of brain stem showing cranial nerves and major arteries; olfactory bulbs and tracts are visible anterior to optic nerves. From J. Nolte and J. B. Angevine, Jr., *The Human Brain. In Photography and Diagrams*, 2nd ed., Mosby, St. Louis, 2000 (photograph by Biomedical Communications, The University of Arizona College of Medicine).

II. MAJOR REGIONS OF THE CENTRAL NERVOUS SYSTEM

A. The Cerebral Hemispheres

The largest, most striking structure in the adult human nervous system is the *cerebrum*. It is almost completely divided into left and right cerebral hemispheres (Fig. 7), each containing a core region of *gray matter*, the *basal ganglia*, a domelike central mass of *white matter*, and an intricately and deeply infolded thin outer sheet (1.5–4.5 mm deep by 2200–2400 cm^2 in area) of additional gray matter comprising the *cerebral cortex*. It has six lobes: frontal, parietal, temporal, occipital, limbic, and central (hidden by opercular folds of the first three; see Fig. 4, 9 months).

1. The Cerebral Cortex

The term cortex (Latin: "bark") means an external layer, in this case of the endbrain. The cerebral cortex features six distinct *layers* of neurons (estimated at 25 billion) and intrinsic axonal plexuses (Fig. 8). Nerve cell bodies are neatly regimented in these horizontal planes as to size, type, and connectivity (3×10^{14} synapses). Nerve fibers (afferent, intrinsic, and efferent) are also arranged in orderly, geometric ways. Nerve cells are of two main types. *Pyramidal* or projection neurons, about 75–80% of the total, display dendritic spines and are generally glutamatergic and excitatory. *Nonpyramidal* or local-circuit neurons lack spines; they take many forms but are usually GABAergic and inhibitory. They dominate some areas (e.g., the visual), but overall are outnumbered by the former.

Superimposed on the laminar plan is a *columnar* mode of organization, featuring about 500 million small, vertical slabs or columns (50–500 μm wide), each containing some 100–300 neurons profusely interconnected in the vertical axis. The cellular make-up of a column is similar throughout the cortex, but the input and output vary according to function and

Table II
Names, Components, Constituent Neurons, and Functions of Cranial Nerves

Nerve	Name	Components[a]	Location of cell bodies[b]	Functions
I	Olfactory	SVA	Bipolar cells in nasal mucosa	Olfaction
II	Optic	SSA	Ganglion cells of the retina	Vision
III	Oculomotor	GSE	Principal oculomotor nucleus	Elevation, depression, adduction, extorsion of eye; elevation of lid
		GVE	Nucleus of Edinger–Westphal	Pupillary constriction, accommodation of lens
IV	Trochlear	GSE	Trochlear nucleus	Intorsion (also abduction, depression) of eye
V	Trigeminal	SVE	Motor nucleus of nerve V	Chewing, control of tensor tympani muscle
		GSA	Trigeminal ganglion	Anterior cranial sensation (face, nose, mouth, dura mater)
		GSA	Mesencephalic nucleus of V (primary sensory neurons)	Stretch input from chewing muscles, pressure from teeth
VI	Abducens	GSE	Abducens nucleus	Abduction of eye
VII	Facial	SVE	Facial nucleus	Facial expression, tensing stapedius muscle
		GVE	Superior salivatory nucleus	Lacrimation, salivation (glands of oral floor)
		GVA	Geniculate ganglion	Nasal and palatal sensation
		SVA	Geniculate ganglion	Taste (anterior two-thirds of tongue)
		GSA	Geniculate ganglion	Sensation from external ear
VIII	Cochlear	SSA	Spiral ganglion	Hearing
	Vestibular	SSA	Vestibular (Scarpa's) ganglion	Sense of stability, oculocephalogyric control
IX	Glossopha-ryngeal	SVE	Nucleus ambiguus	Swallowing movements
		GVE	Inferior salivatory nucleus	Salivation (parotid gland)
		GVA	Petrosal ganglion	Pharyngeal sensation
		SVA	Petrosal ganglion	Taste (posterior one-third of tongue)
		GSA	Superior ganglion	Sensation of skin behind ear
X	Vagus	SVE	Nucleus ambiguus	Swallowing and phonation
		GVE	Dorsal motor nucleus of X	Parasympathetic innervation of viscera
		GVA	Nodosal ganglion	General visceral sensation
		SVA	Nodosal ganglion	Taste (epiglottis)
		GSA	Jugular ganglion	Sensation of skin behind ear
		GVE	Nucleus ambiguus	Parasympathetic innervation of heart
XI	Spinal accessory	SVE	Spinal accessory nucleus	Head, neck, and shoulder movements
XII	Hypoglossal	GSE	Hypoglossal nucleus	Tongue movements

[a]GSA, general somatic afferent; GSE, general somatic efferent; GVA, general visceral afferent; GVE, general visceral efferent; SSA, special somatic afferent; SVA, special visceral afferent; SVE, special visceral efferent.

[b]Note: Input from stretch receptors in the extraoccular and facial muscles may be mediated by cells in the trigeminal ganglion and similar input from lingual muscles by inconstant ganglion cells along the hypoglossal rootlets.

location in the sheet. As to function, each column is stimulus-specific. All of the neurons in a given column respond preferentially to some stimulus parameter: sensory modality, stimulus orientation, ocular dominance, etc. Like the many floors and floorplans of a tall office building, the orderly laminar and columnar features of cortex facilitate integrative, progressive analytic, and comparative functions.

Beneath the cortex, in the white matter, lie millions of myelinated axons of varying caliber, so closely packed as to seem a solid mass of myelin. All of the axons are associated with the cortex. Some go to it,

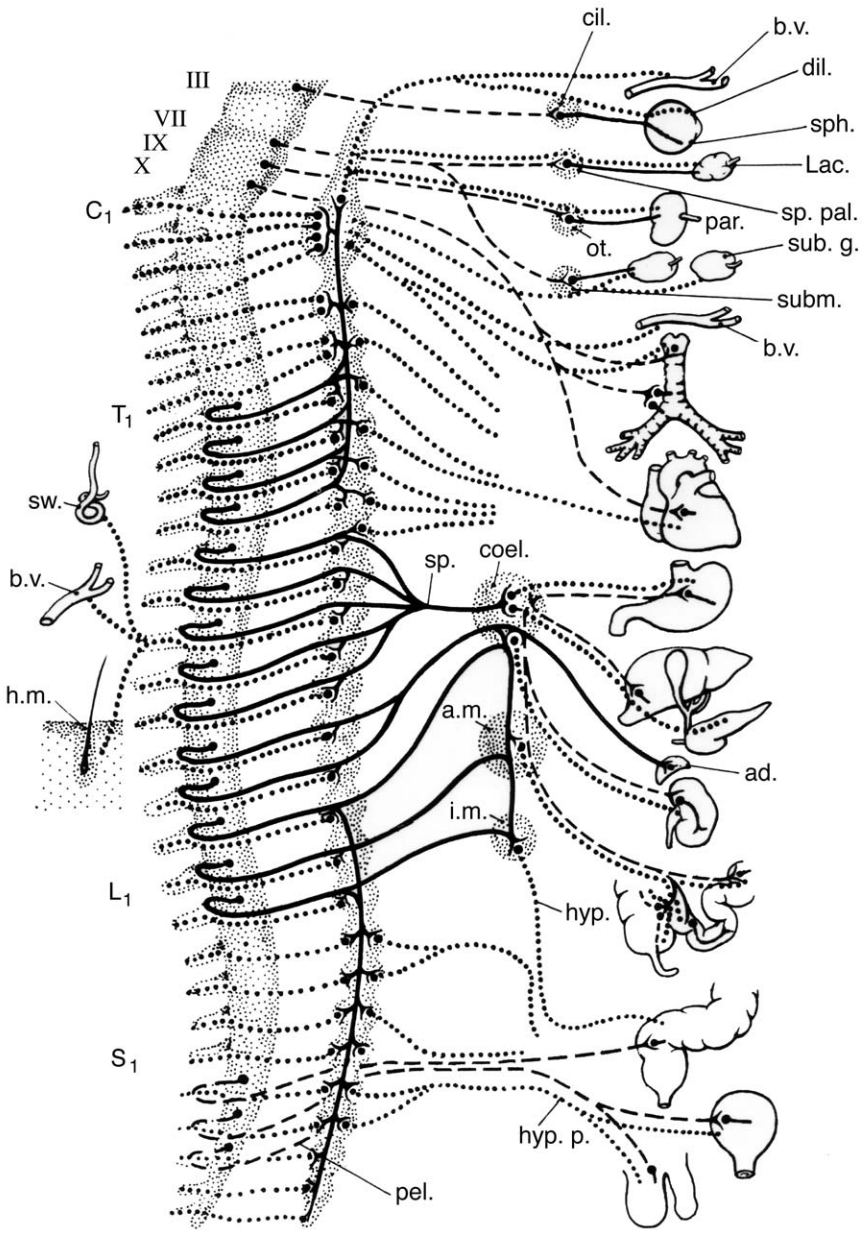


Figure 6 Autonomic nervous system (ANS) showing target organs: preganglionic sympathetic fibers (solid lines), postganglionic (dotted lines); preganglionic parasympathetic fibers (dashed lines), postganglionic (solid lines). From *The Life of Mammals*, copyright 1957 by Oxford University Press, Inc. Used by permission of Oxford University Press, Inc. (original painting by Frank Netter, modified by Jane deVere).

whereas others depart from it. The total length of axons involved in cortical connectivity is estimated at more than 100,000 km. These fibers are crucial to cortical function. They allow parts of the two cortical sheets to influence one another and parts of the underlying brain. Three types of fibers are found (see later discussion). For rapid communication, most are large, well-myelinated axons.

Association fibers interconnect cortical areas in one cerebral hemisphere. They are short, looping from one gyrus to the next, or long, extending the entire frontooccipital length of the hemisphere. They may pass from one cortical area to another in linear sequence or bypass areas, diverge, or converge in a selective manner. These cascades provide profound connectivity in the cortical sheet. Outward or

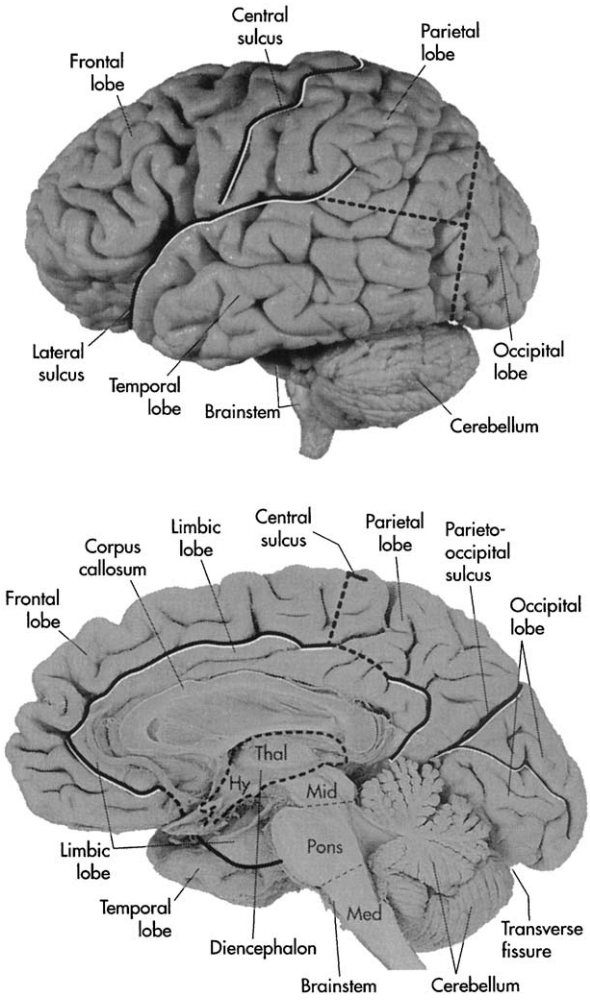


Figure 7 Lateral (top) and medial (bottom) aspects of brain, showing major fissures and lobes (except central lobe; see Fig. 4, 9 months) and principal regions of the brain stem. From J. Nolte and J. B. Angevine, Jr., *The Human Brain. In Photographs and Diagrams*, 2nd ed., Mosby, St. Louis, 2000 (photograph by Biomedical Communications, The University of Arizona College of Medicine).

feedforward connections originate and terminate in specified cortical layers. Backward or feedback connections begin and end in other specific cortical "landing strips." Wherever signals enter the network, they eventually filter through to the orbitofrontal cortex and limbic system, where emotional colorations and affect may be factored in. This network makes for unified cortical function: in analysis and integration, wherein multiple data are processed, coincidences detected, and sensory and motor data reduced and combined in schemes of reference, such as our world and ourselves within it.

Commissural fibers interconnect cortical areas across the midline. There they form a huge curvilinear bridge of white matter, the *corpus callosum* ("hard body"), the largest assembly of nerve fibers in the human brain at 700 mm long and 50–100 mm thick, comprising over 300 million well-myelinated axons (Fig. 7). Most of these make mirror-image connections between the two hemispheres, but some cross to areas different from the ones at which they arise. Much of the temporal lobe interconnects in the *anterior commissure* (Fig. 12; small oval structure just beneath the interventricular foramen).

The traditional concept is that the corpus callosum allows information sharing and teamplay between the two hemispheres. This idea is being reevaluated. Callosal connections involve a time lag (about 30 msec at 6.5 m/sec to travel the roughly 175 mm between origin and termination) that could interfere with cooperative feature analysis. Not excluding hemispheric cooperation, new ideas posit callosal connections permitting hemispheric competition. Callosal connections may effect largely inhibitory, not excitatory, influences and express true hemispheric dominance, albeit moment to moment.

Projection fibers lead to, or come from, subcortical structures. By analogy, they "cross state lines." They convey impulses between structures in different principal regions of the CNS (corticospinal, spinocerebellar, cerebellothalamic, etc.). The neocortex of mammals (as contrasted with the olfactory cortex of all vertebrates) is unique in its *direct lines* to every level of the neuraxis. Association and commissural fibers are direct lines too, from one part of the cortex to another, near or far. No region of the CNS is beyond the reach of the cerebral cortex (Table III). Cortical efferents modulate activity in all parts of sensory and motor neurons, visceral and somatic. Most cortical projections go to the thalamus, which reciprocates them. The work of the cortex would be impossible but for the thalamus. A concept of thalamocortical resonance may offer a better understanding of mechanisms of attention and perseverative thought processes.

Except for minor differences in sulcal pattern, the hemispheres look much alike, but show *functional asymmetry*. Each makes the same functional contributions, but more pronounced than those of the other side. For example, one hemisphere (usually the left) is better in speech and calculation and has a stronger link with consciousness and deliberate, analytical thought processes. The other (usually the right) is better at spatial relations, nonverbal ideation, and holistic

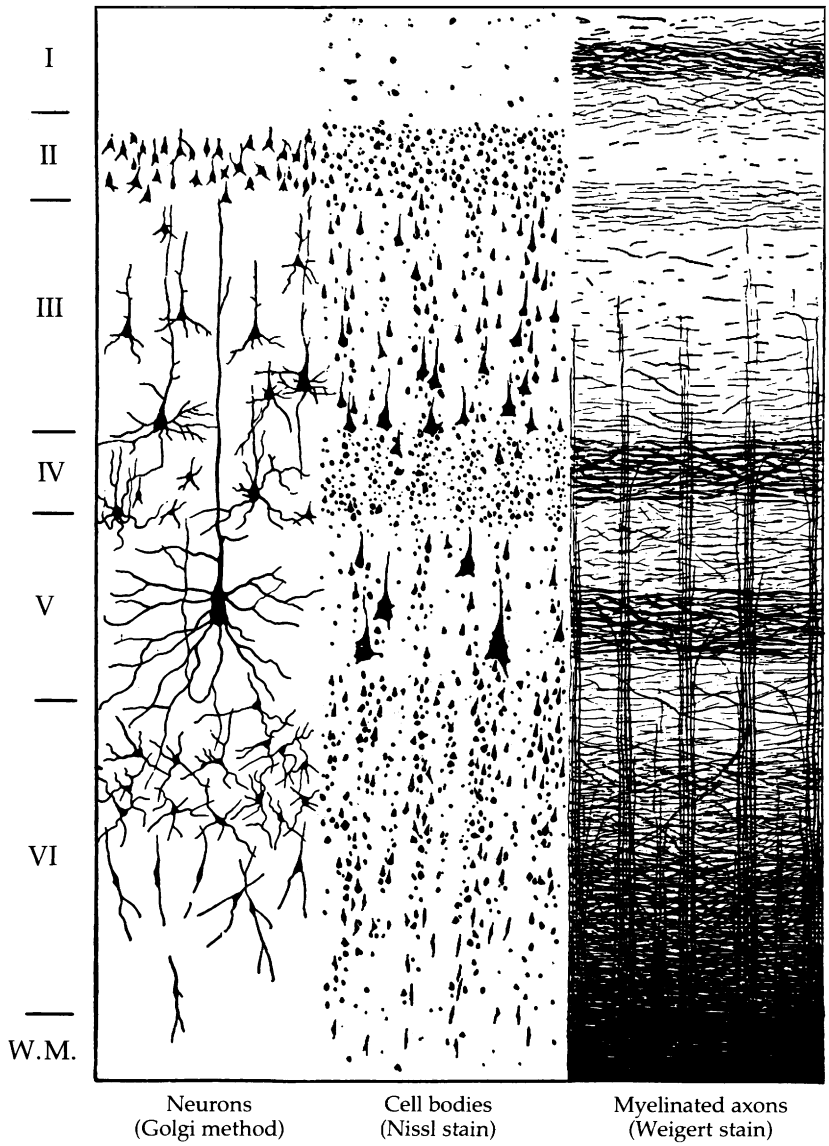


Figure 8 Cerebral cortex showing six principal layers, constituent neurons (pyramidal, stellate, multiform), and myelinated axons. Horizontal lamination is evident, but a vertical or columnar plan of organization is also prominent (see text). From *Principles of Neuroanatomy* by J. B. Angevine, Jr., and C. W. Cotman, copyright 1981 by Oxford University Press, Inc. Used by permission of Oxford University Press, Inc. (illustration by Konstantin Brodmann).

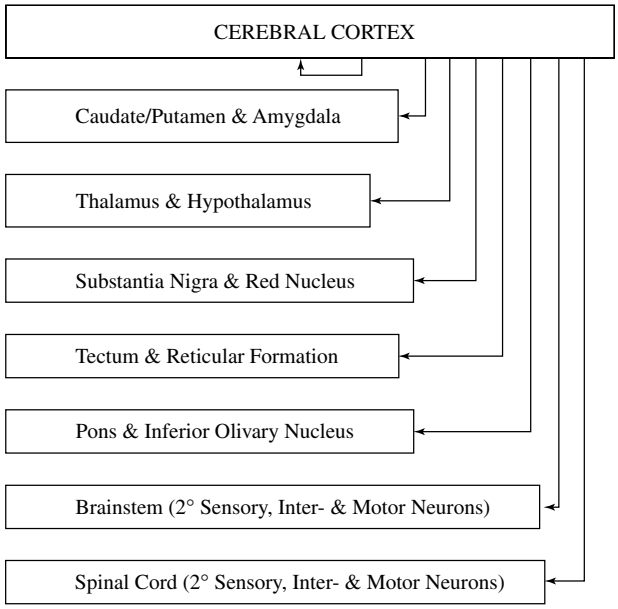
thinking. But in both hemispheres, all abilities are well-represented, and neural integration and pattern synthesis are equally profound.

2. The Basal Ganglia

The basal ganglia are large nuclei within each hemisphere, joined here and separated there, by cortical efferents. They contribute to motor control, cognition, motivation, selection and initiation of behavior, emotion, and perhaps other higher functions (Fig. 9). Only relatively recently have we begun to understand them.

A huge mass of gray matter, the striatum ("striped" by myelinated inputs), has four parts. A lateral, bulky part, the *putamen* ("a shell"), is separated by projection fibers from an inner, curving one, the *caudate nucleus* (with head, body, and tail). Below, in the

Table III



temporal lobe, is the *amygdala* (''almond''), now part of the limbic system. It mediates learning, expression, and the cognitive experience of emotion. A fourth part is the *nucleus accumbens* (not shown), also annexed by the limbic system.

Medial to the putamen is the *globus pallidus*, a major output center to the thalamus and midbrain. Two smaller centers are now considered basal ganglia: the *subthalamic nucleus* and the mesencephalic *substantia nigra*. The former provides an excitatory sidearm between the outer and inner segments of the globus pallidus (all other circuits of the basal ganglia are inhibitory). The lower *reticular part* of the nigra serves, like the inner pallidum, as an output center. Its upper *compact part* synthesizes and ships dopamine up to the caudate and putamen for use as a neurotransmitter in their activities.

The contributions of the basal ganglia have long eluded understanding. The results of their injury, however, are dramatic, expressing something terribly wrong with the teamwork of motor control. Examples include Parkinson's disease and Huntington's chorea. Great strides are now being made regarding the functions of the basal ganglia, the connections that underlie their interplay with the cortex, and the chemistry and molecular biology of their complex synaptic relationships.

The basal ganglia receive widespread cortical inputs and project back, via the thalamus, to motor cortex (Fig. 10). The circuit details are daunting, but the advantage is clear. Almost all of the cortex is involved in movements in the making. The same is true of the cerebellum. Projections from the basal ganglia and cerebellum to the motor cortex are integrated in the rostral thalamus en route.

3. The Limbic System

The limbic system (Fig. 11) is the most controversial subsystem of the human brain. It does not represent one sensory modality, nor one effector mode. Many disagree as to its components. Its functional contributions are not as defined as those of other subsystems or of new ones postulated. The system comprises closely interconnected structures with components in all brain regions, not just the forebrain as conceived. It has multimodal inputs: sensory and associative, and dual outputs: neural and endocrine. It is postulated to mediate memories, drives, and rewards underlying motivation, regulate visceral function and emotional expression, and influence and emotionally color high functions essential to perception and perhaps thoughts.

Nineteenth century neurosurgeon Paul Broca's limbic lobe (the cingulate and parahippocampal gyri, bordering the sulcus of the corpus callosum) gave the system its name. The hippocampus, fornix, mammillary body, and anterior thalamic nuclei projecting to the cingulate gyrus were key links in neuroanatomist James Papez's 1937 sulcal circuit. Then other structures were added: the septal area, temporoamygdaloid complex, habenula and its somatovisceral connections; preoptic area and hypothalamus with visceromotor and endocrine outputs; the striatal nucleus accumbens and ventral pallidum; and the ventral tegmental area and other brain stem nuclei serving visceral afferents energizing the system. Some include the orbitofrontal cortex and the thalamic medial dorsal nucleus projecting to it. This cortex has direct lines to most of the preceding places for cortical oversight and regulation.

B. The Diencephalon

1. The Thalamus

Analysis of detail and synthesis of effector patterns are carried out in the cerebral cortex to a degree unequalled by other parts of the CNS. A close competitor is the *thalamus*, a diencephalic structure between the

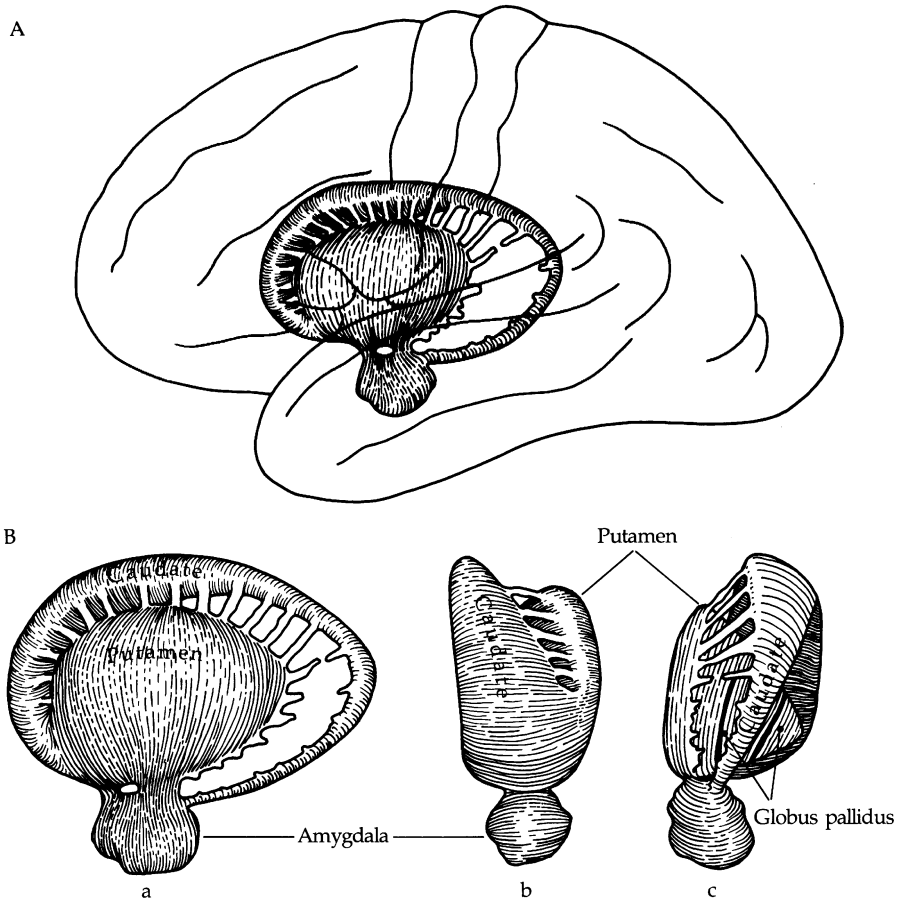


Figure 9 Basal ganglia: the massive putamen is continuous medially in many places with the fishlike caudate nucleus and inferiorly with the amygdala. Medial to the putamen and visible only in view c is the cone-shaped globus pallidus with its outer and inner divisions. From *Principles of Neuroanatomy* by J. B. Angevine, Jr., and C. W. Cotman, copyright 1981 by Oxford University Press Inc. Used by permission of Oxford University Press, Inc. (illustration by Emeline M. Angevine).

cerebral hemispheres, surrounded by them, and during development bound to them by thalamocortical fibers (Fig. 12). The thalamus thus is physically coherent with the cerebrum. It is a bilateral ovoid gray mass with a prominent pole (the pulvinar) facing posteriorly on each side. During development, its halves bulge to form a small interthalamic adhesion. In aging, this adhesion may atrophy and disappear. (Descriptively, "thalamus" often refers to one half-of the thalamus, "thalamic function" to both halves collectively. Such dual usage also applies to the hypothalamus.)

The intimate association with the optic tract is a metaphor of thalamic function. It is the portal of the cerebral cortex. All of the sensory tracts, and many nonsensory tracts, converge on the thalamus. It integrates this information, then passes it to the cortex for additional integration, correlation, and comparison. Olfaction is an exception. It is first processed in the olfactory bulb, a region of old cortex. This sensory information eventually reaches the thalamus after passing through other way stations.

Messages to the thalamus concern *discriminative* aspects of sensation (location, quality, intensity) and *affective* aspects (pleasant, unpleasant). Such inputs dominate the caudal part of the thalamus. Other messages, however, have but an indirect relation to sensation; they come from the basal ganglia, cerebellum, and limbic system. As noted, the basal ganglia and cerebellum act as consultants to the cortex in synthesizing movement patterns, the former offering complex programming of movements and the latter regulating their direction and force. Results of these consultations go to the rostral part of the thalamus (Fig. 10, VA, VL) and are projected to motor cortex,

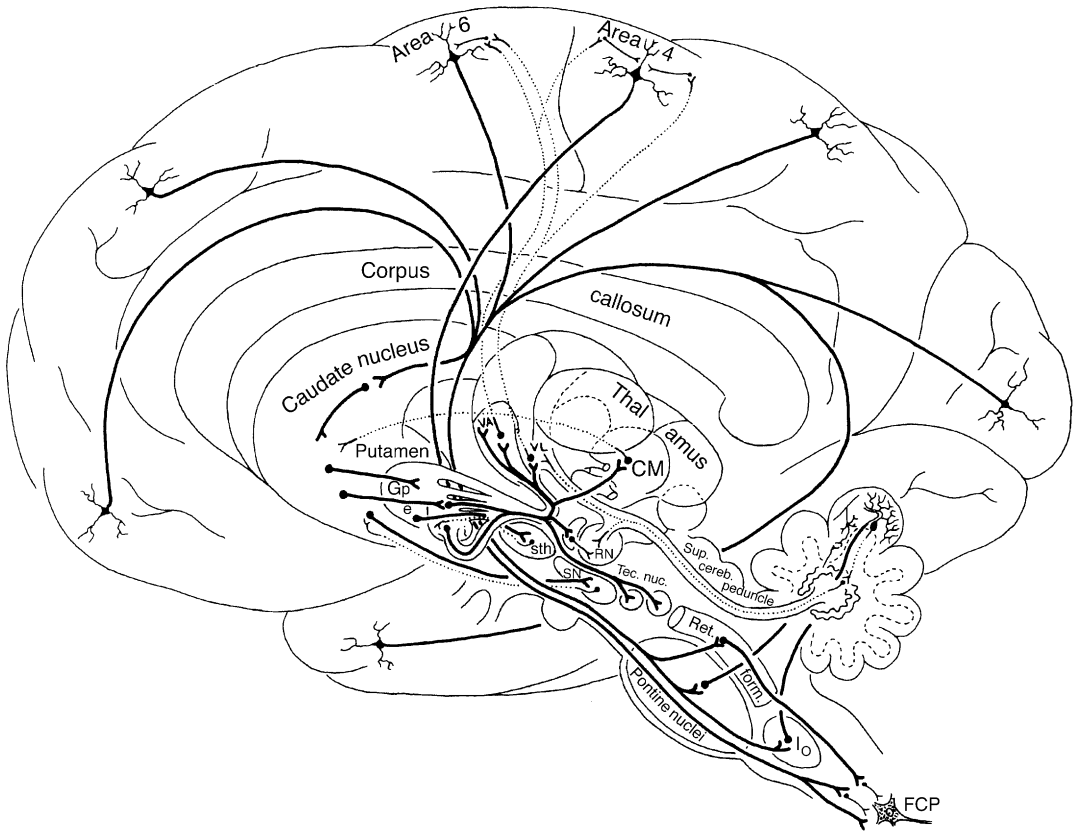


Figure 10 Motor system: an important feature of this complex circuitry is that the basal ganglia lie between the entire cerebral cortex and the motor cortex, by way of pallidothalamic feedback. The same is true for the cerebellum. Although the motor system includes many structures and connections, what emerges is refined motor control. From *Principles of Neuroanatomy* by J. B. Angevine, Jr., and C. W. Cotman, copyright 1981 by Oxford University Press, Inc. Used by permission of Oxford University Press, Inc. (illustration by Jay B. Angevine, Jr.).

where movement patterns are synthesized. Limbic input comes from the hypothalamus and orbitofrontal cortex: visceral afferent data from the former, highly refined integration of the widest range of function imaginable from the latter. As before, thalamic projections to specific cortical areas permit the necessary cortical interplay.

The human thalamus comprises two egg-shaped halves containing some 15 major nuclei and many small cell clusters. Space constraints preclude their coverage, but the structural and functional unity of the thalamus is more important. Over the last 50 years, neurophysiology has shown five functional attributes of thalamocortical connections that are especially illuminating.

Thalamic *overlap* and *fusion* blend modalities and submodalities of sensation, e.g., all sensations of the index finger are processed by neurons in a particular region of the thalamic somesthetic nucleus. *Orderly representations* of sensory or other functional domains form topographic maps of the retina, cochlear duct, and cutaneous surface, in tracts leading to the thalamus, and in certain thalamic nuclei. *Multiple representations* of venues are exemplified by the lateral geniculate nucleus (Fig. 12). It receives the fibers of the optic tract. Its six cell layers provide six tightly registered retinal maps of the contralateral half of the binocular visual field, the map in each layer representing the ipsilateral or contralateral hemiretina serving the half visual field in alternation. *Reciprocity* of function is evident in the equally numerous and precise thalamocortical and corticothalamic projections. These inhibit or facilitate thalamic neurons, perhaps for selective attention or further integration. Finally, *specificity* of thalamic neurons: David Hubel and Torsten Wiesel, Nobel laureates of 1981, showed that

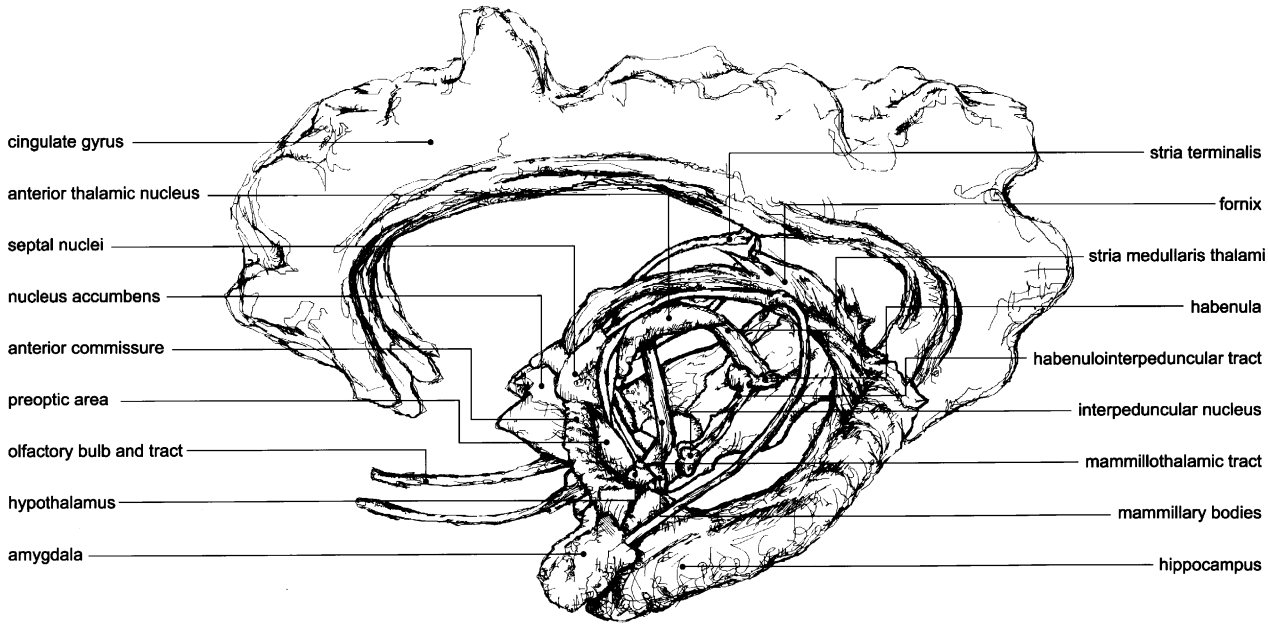


Figure 11 The limbic system: major limbic centers and connections, based on a computer reconstruction of limbic structures traced and digitized from whole brain serial sections. From Cheryl A. Cotman, Jay B. Angevine, Jr., and Kevin Head (illustration by Cheryl A. Cotman; commissioned for this article by the author). (See color insert in Volume 1).

nerve cells in the visual cortex respond to particular features of sensory information and only that feature.

2. The Hypothalamus

The hypothalamus (Fig. 13) lies at the heart of the limbic system. With dual outputs, it plays key roles in short-term and long-range homeostasis. Its neural output is effected by descending tracts in the core of the neuraxis and its endocrine output by pituitary hormones. It regulates vegetative functions: control of body temperature, caloric input, water–osmolar balance, and sleep–wakefulness. It selects and integrates autonomic responses. It controls the release of adenohypophyseal and neurohypophyseal hormones that exert far-ranging and, in some cases (growth hormone), enduring effects on the body.

The hypothalamus lies just below the thalamus. It is minute: about one-three hundredth to one four-hundredth of total brain weight (4 out of 1400 g). Yet it has over a dozen nuclei and receives, emits, or gives passage to as many tracts and to and from as many brain regions, both upstream and downstream. Neuroanatomist Walle Nauta saw it as the instrument panel or the "dashboard" of the brain. It monitors and regulates (not on dials or by switches but from receptors and via diverse connections) visceral and endocrine activities for homeostasis, offensive–defensive or trophic responses, and conative pursuits of long-range goals.

C. The Midbrain

The cylindrical midbrain (Fig. 7) separates the diencephalon and pons. Hardly more than 20 mm in diameter, its layout of structures reflects its embryonic tubular plan. Massive tracts run through it, some ascending from the spinal cord and cerebellum and others descending from the cerebral cortex. The cerebral aqueduct in its upper part is a vital, yet narrow, vulnerable conduit of cerebrospinal fluid (CSF) and a useful landmark. Above it, in the *tectum* (Latin: "roof"), lie two small hills of gray matter: the superior and inferior colliculi, which are visual and auditory way stations. They receive 10% of the fibers of the optic nerve and 100% of the auditory projection, respectively. They show cortical features of organization, perform multimodal integration, and mediate complex responses to sights and sounds.

Beneath the aqueduct is the *tegmentum* (Latin: "covering"). It covers *basal* parts of the midbrain, the cerebral peduncles, and substantia nigra. It houses

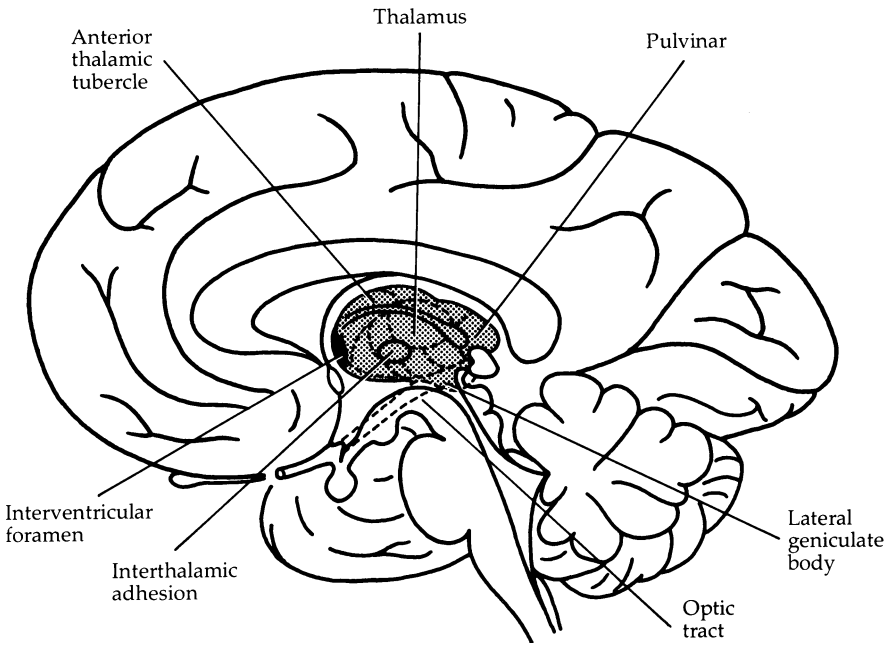


Figure 12 The thalamus, gateway to the cerebral cortex: a large mass of gray matter, comprising over a dozen recognizable nuclei, in the upper diencephalon, separated by the third ventricle into right and left halves. Its strategic central position permits a broad range of upstream input and cortical interconnections. From *Principles of Neuroanatomy* by J. B. Angevine, Jr., and C. W. Cotman, copyright 1981 by Oxford University Press, Inc. Used by permission of Oxford University Press, Inc. (illustration by Jay B. Angevine, Jr.).

cranial nerve nuclei mediating eye movements and pupillary constriction and provides the rostral part of the brain stem reticular formation, a fabric of gray and white matter forming a continuous core of the midbrain, pons, and medulla oblongata. It exerts strong facilitatory and inhibitory influences on virtually all activities up and down the neuraxis. In the reticular formation on each side is a large round mass of neurons: the highly vascularized *red nucleus*, a motor center distributing cerebellar output to nuclei in the brain stem and thalamus.

Basal components are corticospinal and corticopontine fiber bundles surmounted by a gray cushion, the *substantia nigra*. Its compact upper part, where some neurons contain melanin pigment, makes dopamine and delivers it by axonal transport to the caudate nucleus and putamen for use in their complex neuronal interactions related to motor control. The lower layer of the nigra, as stated, serves as an output center of the basal ganglia, along with the inner segment of the globus pallidus.

Injuries to the midbrain have devastating results: almost total loss of sensation, paralysis, severe extensor spasms from the interruption of key connections of the motor system, and, from damage to the reticular formation, lasting coma if not death.

D. The Cerebellum

In humans, the cerebellum is half-hidden by the overlying occipitotemporal regions of the cerebral hemispheres (Fig. 7). It is bilobed, with a narrower median part, the vermis, connecting two large, ovoid hemispheres. These are gracefully fissurated in a curving way, a feature resulting in closely apposed, almost sinusoidal folia coated by cortical gray matter. The geometrical organization of this cortex, the self-evident design of its elegant neurons, and the swift, reliable contributions to motor control make it a natural wonder. The cerebellum is found in all vertebrate brains, with tremendous variations in size, shape, numbers of neurons, and neuronal idiosyncrasies. But the usual neuronal types are always there and the basic plan of cortical connectivity is instantly recognizable in all.

The cerebellum, like a computer, regulates the rate, range, and force of movements and contributes to muscle tone and posture. Like the basal ganglia, it works in concert with the cerebral cortex. It has ties to pontomedullary vestibular centers and also input from the spinal cord. Unlike the cerebrum, it does not play a major role in perception of sensation or initiation of volitional movement. With damage to the cerebellum,

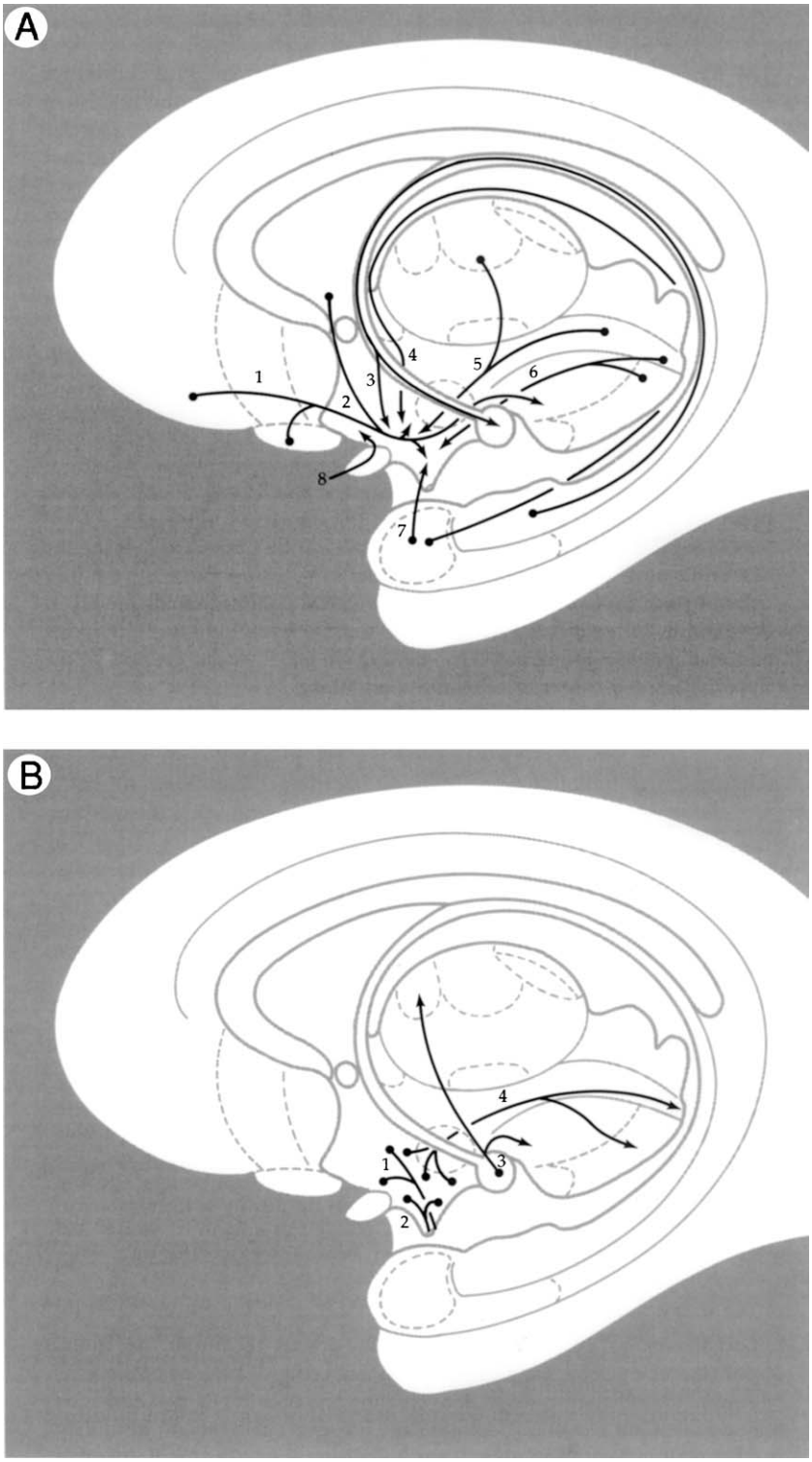


Figure 13 Hypothalamus: afferents (upper view) are corticohypothalamic fibers (1), medial forebrain bundle (2), fornix (3), stria terminalis (4) periventricular fiber system (5), mammillary peduncle (6), ventral amygdalofugal pathway (7), and retinohypothalamic fibers (8). Efferents (lower view) are the hypothalamohypophysial tract (1), tuberoinfundibular tract (2), principal mammillary fasciculus (including the mammillothalamic tract) (3), and dorsolateral fasciculus (4). These connections are largely reciprocal and form integral parts of the limbic and neuroendocrine systems of the brain. From *Principles of Neuroanatomy* by J. B. Angevine, Jr., and C. W. Cotman, copyright 1981 by Oxford University Press, Inc. Used by permission of Oxford University Press, Inc. (illustration by W. J. H. Nauta and V. B. Domesick).

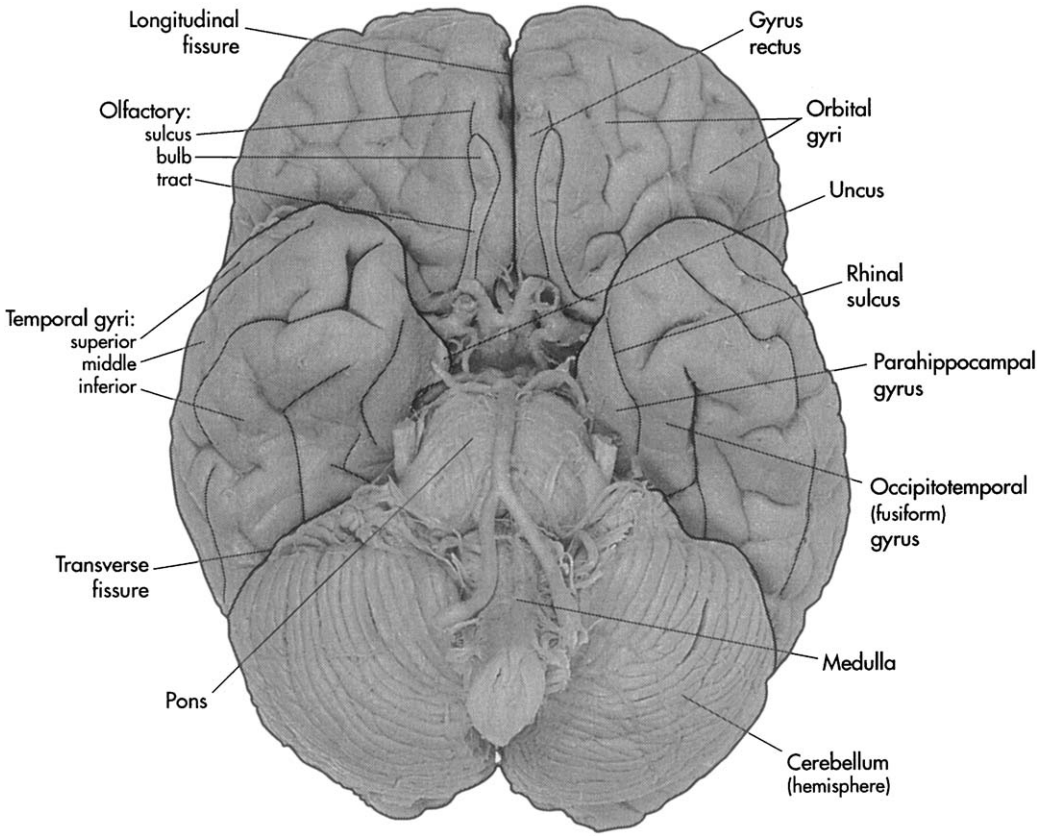


Figure 14 Base of the brain: low power view of same brain shown in Fig. 5 to show the inferior surface of the cerebellum, pons, and medulla oblongata, together with other structures. From J. Nolte and J. B. Angevine, Jr., *The Human Brain. In Photographs and Diagrams*, 2nd ed., Mosby, St. Louis, 2000 (photograph by Biomedical Communications, The University of Arizona College of Medicine).

conscious sensation is largely spared, but dexterity and smooth execution of movement (especially skilled movements of the upper extremities) are impaired, muscle tone and strength are diminished, and posture and equilibrium are deficient. The degree of deficits depends on the location, severity, and duration of the insult. In time, cerebellar circuitry often brings improvement, but damage to its outflow path, the superior cerebellar peduncle, usually effects lasting dysfunction.

E. The Pons

The pons (Fig. 14) is a thick bridge of nerve fibers, crossing the midline anteriorly. With a massive arm reaching into the cerebellum on either side, it binds that organ (its derivative) to the brain stem like a backpack strapped around one's waist. But it does more than hold the cerebellum in place. It provides a key link between the cerebral cortex and the cerebellum. The brachium pontis, or middle cerebellar peduncle, is a huge cable by which the contralateral cerebral cortex, through the liaison of 20 million pontine neurons, accesses the cerebellar movement computer. This cross-link in cerebral–cerebellar function enables movement to be performed in a smooth, coordinated, directed manner.

Caudal to the middle cerebellar peduncle is the inferior peduncle, in which sensory information from spinal cord and messages from the cortex sent through the inferior olivary nucleus of the medulla reach the cerebellum. Rostral to it is the superior peduncle. It distributes cerebellar output, via a thalamic nucleus, back to the contralateral motor cortex (thus completing a neural "double-cross"), as well as to various motor centers in the brain stem (red nucleus, reticular formation, and others).

F. The Medulla Oblongata

Protruding behind the pons (Fig. 14), the medulla oblongata tapers smoothly into the spinal cord. Like many regions of the CNS (see locus ceruleus), it is important out of proportion to its size. No bigger around than the last part of a little finger, it is vital. Even slight injury to it can swiftly lead to devastating or fatal consequences: hemianesthesia, hemi- or diplegia, and respiratory arrest. Through the medulla oblongata run all of the long ascending and descending tracts that connect the brain with the spinal cord and permit their interactions. In it are neuronal clusters of the lowermost five cranial nerves (Table II) and other nuclei that regulate breathing, heart rate, swallowing, caliber of small blood vessels, and indirectly blood pressure, waking, sleeping, and other vital functions. Some medullary neurons, like those of the brain stem serotonergic system, elicit profound effects by their neurotransmitters and through pervasive connections on general levels of activity elsewhere in the neuraxis. Consciousness and alertness depend on their upward influences, and spinal reflexes of posture, locomotion, and visceration rely on their downward effects.

G. The Spinal Cord

The caudal part of the CNS is the spinal cord (Fig. 1). Sensory impulses enter through its 31 paired spinal nerves, and sensory data processing begins. Messages flow up over modality-specific tracts, fast and direct or slow and indirect, to higher cord regions and the brain: to the medulla, pons, cerebellum, midbrain, and thalamus. The thalamus integrates this information with other input and in reciprocal association with the cerebral cortex. In due course, or on its own, commands from the cerebral cortex and certain brain stem regions flash down to spinal motor neurons and nearby local-circuit neurons. These small cells regulate the motor neurons and integrate their activity from one level to another and across the midline. Although its intrinsic spinal activity is monitored by higher centers, the spinal cord draws up swift and complex reflexes on its own, playing out movement programs (like walking or running) as if from a computer–numeric–control tape.

III. BASIC ORGANIZING PRINCIPLES OF THE NERVOUS SYSTEM

Breakthroughs have occurred in our understanding of these principles. Visual sensory processing, the task of the visual system, is a good example. Like other sensory analyses, it is far more complex than once surmised. At their first way station in the brain, the lateral geniculate nucleus (LGN), optic tract fibers end on neurons in one of two neuronal populations: a large-celled group in the lower two layers of the LGN and a more numerous small-celled group making up the upper four layers. When thalamic projections of these two types of cells reach the primary visual cortex in the occipital lobe, visual processing diverges into two distinct, only partly connected streams. The parvocellular system streams into inferior temporal cortex for analysis of color and form and the magnocellular system into posterior parietal cortex for analysis of location and movement. In brief, these ventral and dorsal streams are the "What?" and "Where?" systems. Taken together, they comprise (at last count) 32 distinctive cortical areas interconnected by association fibers in a complex, often nonlinear, reciprocal manner, affording divergence and convergence modes of data analysis.

Visual analysis draws upon other sensory systems and systems distant from visual processing: the motor system, core integrative systems (reticular formation, thalamus), the affect and drive-related limbic system, and the decision-making and memory-related frontal cortex. Perceptual meaning emerges and responses drafted as impulses diverge and converge in the network in domains of alertness, attention, comparison, affect, cognition, and conation. By feature analysis, integration, pattern synthesis, and data storage and retrieval, the nervous system may oversee activity inside and outside it, maintain homeostasis, and generate and monitor responses, all in a timely way, but not necessarily. We now consider eight attributes of the nervous system. They provide a context for its design and details.

A. Ubiquity

With 100,000 miles of nerve fibers (Fig. 1), the nervous system rivals the vascular system. Both pervade the body and function in harmony. By nerve impulses or circulating red and white blood cells, glucose, hormones, and immune principles, they integrate body activity, protect the body, enhance its performance to meet stress or demand, promote its growth and nutrition, and maintain its tone and vigor. The trunk and branches of both systems reflect body form. If either system and no other part of a person were visible, he or she would be recognizable. Density of

innervation varies as the value of parts to sensory discrimination or motor control. In well-innervated areas (lips, fingertips), stimuli are sharply discriminated as to modality, intensity, and location, but in sparsely innervated areas (flanks, legs) these are less defined. Similarly, muscles vary in the ratio of motor neurons to muscle fibers. The higher the ratio, the more precise the control of the muscle and the movements it serves (a motor neuron may excite 2000 muscle fibers in a limb muscle or as few as 5 in extrinsic ocular muscles).

B. Unity

As an epithelium, all parts of the nervous system are physically coherent and functionally linked by nerves, tracts, and specified cell-to-cell contacts. Potentially, each part communicates with all others. Some connections are direct (a two-neuron, monosynaptic reflex), whereas others involve myriad interposed neurons. Though complex, neural circuits offer total connectivity: fast, body-wide communication. Nerve impulses may originate in sensory nerve endings in any part of the body or anywhere in the system itself. *Responsive* activity complements *endogenous* activity, which is always evident in the human nervous system with its startling capacity to generate patterns of behavior and initiate events on its own. Sensory impulses, triggered by PNS primary sensory neurons, race over its nerves to the CNS, there diverging to clusters of secondary sensory neurons. Analysis begins. New impulses pass to central neurons on which related messages converge, which is a recombinant process providing *integration*. Other messages on stimulus modality, intensity, location, affective quality, body position and movement, visceral activity, fatigue, experience, and expectations are all integrated. Huge numbers of impulses are generated; untold numbers of synapses are activated. Almost instantly, nerve impulses that will elicit bodily responses stream out of the CNS to muscles and glands.

C. Centralization

The key feature of the nervous system is centralization. It offers few circuits for local interactions of body parts. The CNS is almost always involved even if the distance, as from thumb to index finger, is slight. Intercession of the brain and spinal cord ensures integrated and coordinated activity.

Exceptions are instructive. The local cutaneous response to irritating stimuli (raking a blunt probe over the skin) has three components: local reddening (vasodilation from injury), wheal formation (transient edema from tissue fluid extrusion), and ensuing widespread vasodilation (flare) with lowered threshold and increased sensitivity to pain (pinprick). The flare and hyperalgesia represent an *axon reflex*. Nociceptive (pain) nerve endings are activated by substances released by injured tissue cells, and nerve impulses are conducted a short way centrally along nociceptive axons and then distally over branches of these axons to nearby arterioles, causing them to dilate. Advanced or primitive (it is sluggish, starting in about 20 sec and developing fully in around 3 min), this reflex involves local nerve fibers only, not the CNS.

The "triple response" illustrates three concepts. Pain receptors sense chemical, as well as mechanical and thermal, stimuli. Their sensitivity is increased by substances accumulating in the damaged area. Their response includes a *neuroeffector* component. They release substances (peptides) that initiate further events, providing further protection and favoring local tissue repair.

Studies in invertebrate neural systems show extensive local control of visceral function. Exceptions to central control are also found in the mammalian ANS. Near-normal interaction of bowel segments persists in the absence of CNS innervation. Sensory fibers from the gut exert feedback in intramural autonomic ganglia on visceral motor neurons regulating smooth muscle in the intestinal wall. The nervous system has *pattern generators*, both central and peripheral: systems with cellular, synaptic, and network properties (cyclic firing rhythms, reciprocal inhibition of cell pairs, leader and follower cells) that provide automated mechanisms for generating rhythmic movements (breathing, walking) or periodic activities (sleeping, waking). Regulated by neural (sensory feedback, volitional override) or neuroendocrine influences, pattern generators are pithy examples of neural endogenous activity.

D. Specialization

Reflecting its diverse tasks, the nervous system is specialized, from the single neuron to each brain region. Specialized subsystems analyze sensations. They differ in some ways, but data processing is progressive and networked in all. Neurons and the neuroglia have special shapes and roles, but both enjoy

all criteria for cells and work in concert. Less obvious but equally specialized are subsystems for other functions: sleep–wakefulness, alertness, attention, affect, collating pages of a report, reading out loud from a book, self-awareness, brain damage control, and so on ad infinitum.

Ubiquitous specializations include those for high nerve conduction velocity (large axon diameter, thick myelin sheath), space-saving bundling (small axon diameter, thin myelin sheath, shared sheaths), short latency response (monosynaptic reflex), staggered, persistent latencies (parallel side chaining of long-axoned neurons), dependability (neuron redundancy), feature analysis (parallel processing), effect monitoring (feedback circuits), and force multiplication (feed-forward circuits). The neurons performing such tasks and the neuroglia backing them up are as specialized as these many diversified services. For neurons and the neuroglia, form indeed reflects function.

E. The Purposefulness of Neural Components

Every part of the nervous system has at least one function, often many more. Small parts of the CNS may play crucial roles, as in the extensive distribution and profound influence of axons from inconspicuous brain centers. The *locus ceruleus* (''blue spot'') on each side of the fourth ventricle contains about 12,000 large, melanin-pigmented neurons. These synthesize *norepinephrine* and release it in the cerebral cortex, cerebellum, and almost every other part of the CNS. Electrically, they are almost silent in sleep, hypoactive in wakefulness, and hyperactive in watchful or startling situations. They serve vigilance and attention to novel stimuli. They contribute, indirectly but no less crucially, to perceptual and cognitive functions. By contrast, immense structures make large but expensive contributions, as in the cognitive and motor abilities afforded us by the billions of neurons in our cerebral and cerebellar cortices.

F. Uniformity with Versatility

The vertebrate nervous system is accurately and reproducibly assembled. In animals of like genus and species it appears almost identical, although this is not absolute when genetic histories differ. Minor variations in the size of components and arrangements of cells are seen between species, striking ones between classes, orders, and families. Yet basic regions and properties, cells and circuits, and overall organization are sufficiently alike to permit instant recognition of the basic brain plan and insights as to what these parts and cells contribute to function. Humans show huge increases in brain size and regional elaboration, numbers of neurons and prominence of certain connections, variations in cerebral sulcation, hemispheric asymmetry, and long projections.

G. Plasticity

Highly reliable in a healthy person, the human nervous system has inherent modifiability, though in adulthood this attribute cannot approach that in invertebrates (moths and snails) or certain other vertebrates (teleosts and amphibians). In mammalian development, neural plasticity is striking. It continues postnatally. Abnormal visual experience at certain sensitive periods profoundly affects ocular dominance and orientation columns in the visual cortex. If an eye is closed at birth, ocular dominance columns for the other eye enlarge at the expense of adjacent blind eye columns, with thalamic fibers arriving in the cortex expanding terminal fields into them. If, shortly after birth, visual stimuli are restricted for a few weeks or even days to stripes of one orientation, cortical cells develop a response preference to lines of that orientation.

In humans, PET imaging studies of cortical blood flow show that tasks requiring tactile discrimination activate visual cortex in people blind at birth or having lost sight in childhood. This suggests that cortical connections reorganize after blindness: that afferent fibers to nearby cortical areas serving polymodal sensory integration usurp the bereft visual cortex. Such plasticity may explain the well-known tactile acuity of the blind.

In later development, neural plasticity operates on many levels, as in fine-tuning circuits to changing body dimensions. Depth perception is recalibrated as the skull enlarges and interpupillary distance increases. Even in adulthood, plasticity persists. Vilayanur Ramachandran has shown that a stroke with a cottonswab on the cheek of a young man who had accidentally lost his left arm led him to feel touch on his missing left hand. Later, the whole hand could be mapped on his face.The findings suggest that the deprived somatosensory cortical region for the hand becomes innervated by fibers from the adjacent face areas and that secondary input to a cortical neuron's broad receptive field becomes functional when primary input is lost.

After injury to the CNS, intact neurons form new terminals, by *axon sprouting*, to replace those of other neurons lost to trauma and thus reoccupy vacated synapses. Such *reactive synaptogenesis*, the clinically proven effectiveness of long-range regrowth of PNS axons, and the evident potential for axon regeneration in the CNS (as in teleosts and amphibia) hold promise for circuit restablishment. But in mammals, these factors are thwarted by myelin debris, glial scarring, usurpation of sprouts, unresponsive injured neurons, and complex central connections. Developmental neuroscience now focuses on the cerebral cortex. The human nervous system appears to learn very rapidly by using preconstructed circuits and by locking neurons into specific types and functions after cell origin.

H. Chemical Message Coding

The basic function of the nervous system, from which all others derive, is *communication*, performed (with unsung neuroglial support) by neurons. It depends on special electrical, structural, and chemical properties of these diversified cells with their long processes, on their exploitation and refinement of two basic protoplasmic properties, *irritability* and *conductivity*, on their external and internal neuronal morphology featuring multipolar shape and integrative design, almost infinite modes of dendritic and axonal branching, widespread, diversified connections, and specialized organelles, and on their use of chemical substances to encode, deliver, and decipher messages of their own and other neurons.

Neural circuits are chemically coded. Neuroanatomy encompasses interneuronal connections and also chemical mediators and transmitters. Neuroactive substances comprise neurotransmitters, neuromodulators, and neurohormones. Their definition in contexts other than site of action, postsynaptic neuronal activity, and corelease of one or more additional neuroactive substances can be misleading. Neurotransmittters are small molecules acting swiftly, locally, and briefly on target cells. Neuromodulators are very small (peptides), regulating but not effecting transmission, and neurohormones are also small, with intrinsic activity mediated by neuronal and other cells, exerting slow, widespread, and enduring influence via the extracellular fluid or bloodstream.

Neurons releasing hormones are quasi-endocrine cells, liberating secretory products from axonal endings into the perivascular space to be conveyed to blood vessels and thence to target organs. The provincial concerns of neurophysiology and endocrinology have fused into neuroendocrinology, as psychoneuroimmunology has united psychobiology, molecular neurobiology, and immunology.

IV. STRUCTURAL APPROACHES TO STUDY OF THE NERVOUS SYSTEM

Neurohistology is the study of nervous tissue per se, and neuroanatomy is the study of nervous tissue as a *communications system*. The concerns and methods of the two fields are different, but overlapping and interdependent. Both approaches shed light on the organization of the human nervous system.

A. Neurohistology

Like the microscopic anatomy of other tissues, this approach focuses on nerve cells: their types, internal and external features, structural, functional, and chemical properties, and interrelationships. But the great size, extent, and diversity of neurons, their interdependence, and their complex spatial relationships can be daunting to the histologist.

Let us examine a motor neuron. Figure 15 is a schematic illustration showing a motor neuron innervating a muscle just above the letter B. We begin with the cell body, noting its size, organelles, and inclusions, and then trace its dendrites (100 terminal ones), spines, and relationships to neuroglial processes and capillaries. We identify the axon, note its features (axon hillock, initial segment, side branches if any), and follow its course through the gray matter into the white matter (the two are not delineated), where it becomes myelinated by oligodendrocytes (OI) for rapid conduction and protection.

We follow the axon into the PNS. Here it is myelinated by Schwann cells (S_1), for the same reasons. Outside the myelin sheath, three connective tissue sheaths (endoneurium, perineurium, epineurium; not shown) provide vascular, isolating, and protective support for the delicate nerve fibers. We trace the axon for a meter or so to a skeletal muscle and inspect its preterminal branches and axon terminals on muscle fibers at the neuromuscular junction: the nerve–muscle synapse.

We cannot examine one motor neuron in all respects. For the total picture, we piece together these parts from many motor neurons. The hardest thing to grasp is the relative size of a neuron and its parts.

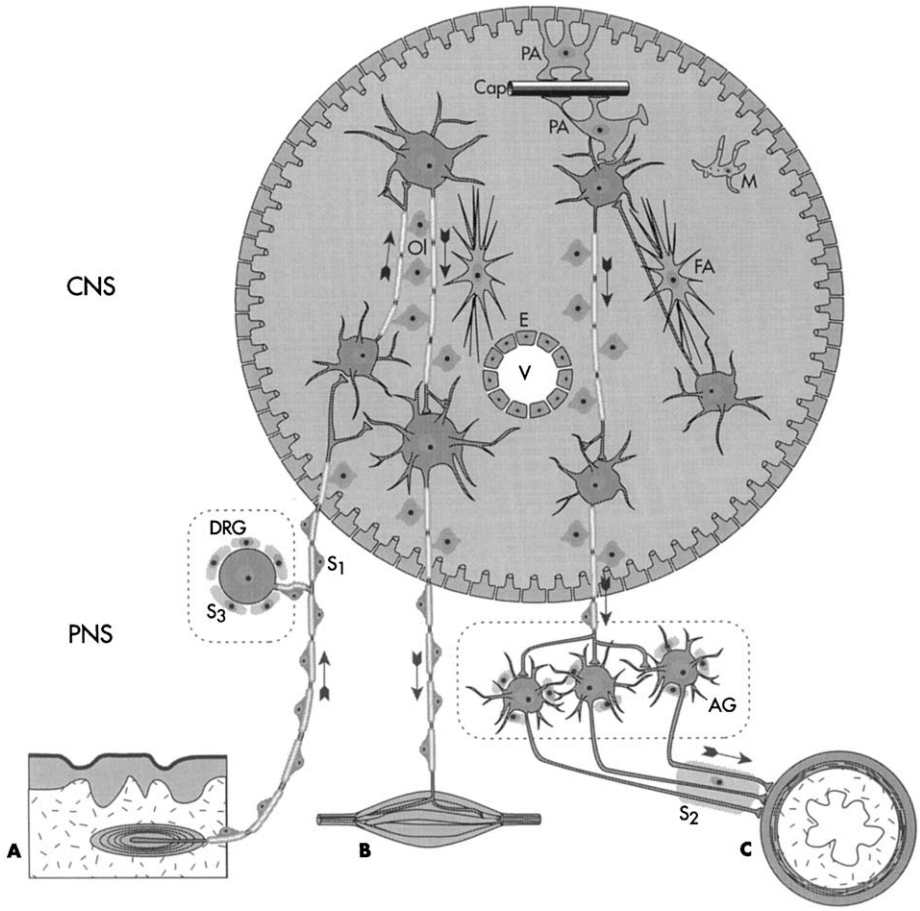


Figure 15 Neuronal and neuroglial cell types in the nervous system: the end feet of protoplasmic astrocytes (PA) form a leaky membrane around the CNS and others in the gray matter about either neurons (the dark cells) or capillaries (Cap). Fibrous astrocytes (FA) lie among axons in the white matter, many of which are myelinated by oligodendrocytes (Ol). Microglia (M) scavenge debris after injury, and ependymal cells (E) line the ventricles (V). Schwann cells are the glial cells of the PNS, myelinating peripheral nerve fibers (S_1), surrounding unmyelinated axons (S_2), and satelliting dorsal root and autonomic ganglion cells (S_3). Direction of impulses in various axons is indicated by arrows. Other abbreviations: skin (A), skeletal muscle (B), gastrointestinal tract, containing intramural smooth muscle and glands (C), autonomic ganglion (AG). See also text. From J. Nolte, *The Human Brain. An Introduction to Its Functional Anatomy*, 4th ed., Mosby, St. Louis, 2000 (illustration by Prof. Dr. Radivoj Krstic).

Neuroanatomist Jack Nolte offers this perspective: If the cell body of a motor neuron were the size of a tennis ball, its dendrites would fill a room and its axon would extend, like a 0.5-in. garden hose, nearly 0.5 mile.

B. Neuroanatomy

Here the emphasis is different. We look at a motor neuron as a part of the *motor subsystem* of the nervous system. From the number of axon terminals on its receptive surface (around 50,000, each signifying a synapse), we see that this large cell, final in the chain of communication from sensory periphery to muscles, has a vast range of inputs. This includes PNS primary afferents, spinal and brain stem connections, feedback and feedforward circuits from other motor neurons, self-directed inputs from recurrent axon collaterals of its own and neighboring local interneurons, and direct, express lines from the cerebral cortex 1 m upstream from this motor neuron. We recall that each input releases a chemical messenger eliciting excitation, inhibition, or modulation of the target cell, depending on the nature of receptor molecules and second messenger systems in it and other complex factors. We know that each input has a specified place on the

soma, dendrites, or axonal initial segment of the motor neuron and that this orderly display favors input identification and efficacy in control of this harried cell.

Moving peripherally, e.g., in the sciatic nerve, we find that the axon terminals in the muscular system end on muscle fibers in the gastrocnemius. With methods for tracing nerve fibers, we find that sensory axons supplying stretch receptors (*muscle spindles*) in this muscle lead back into the CNS over the dorsal roots of spinal nerves to synapse on our motor neuron and all its homonymous neighbors, thereby closing the external arc of a spinal neuromuscular control loop: the monosynaptic *stretch reflex*. We now see the place and role of the motor neuron in the human nervous system. Its inputs include large, well-myelinated, fast-conducting axons that excite it monosynaptically when the gastrocnemius is stretched. Its output to gastrocnemius fibers enable it, with the help of other motor neurons, to maintain our erect posture.

Our virtual excursion teaches us that neuroanatomy is less interested in cytological detail and more interested in the identity, interconnections, and interactions of neurons, in "wiring diagrams," in chemical colorations, and in the functional roles of neurons in networks.

C. Complementarity of Neurohistology and Neuroanatomy

The two approaches contrast with, yet complement each other. The study of neural circuits is enhanced by familiarity with neurons and the neuroglia, just as their study is enriched by knowledge of the uses to which these cells are put. Whatever the design and extent of its circuits, the work of the nervous system derives from neuronal activity, singly or in modular groups, within these circuits. Indeed, one neuron can be as complex as an integrated circuit, because it responds (when it does) partly in a *fractionated* manner as well as in the familiar "*all-or-none*" nerve impulse.

V. PROPERTIES OF NERVOUS TISSUE CRUCIAL TO THE NERVOUS SYSTEM

Six properties of nervous tissue are key to its communications functions. Among these, the first is cardinal, especially for the CNS. We revisit it after overviewing all six.

A. Epithelial Nature of Nervous Tissue

The nervous system is an *epithelium*, with cells close by and little space between. Far more complex than other epithelia, with flattened, cuboidal, or columnar cells in single or stratified sheets covering surfaces or lining spaces, the nervous system is a superepithelium: a huge, three-dimensional jigsaw puzzle with billions of pieces of varying size (1 to over 150 μm), length (1 μm to 1 m) and, in some cases, an extraordinarily complex and beautiful configuration. Electron micrographs show that its pieces fit together, precisely and intimately, with little space left over, an evident benefit for cell communication (Fig. 16).

B. Integrative Design of Cells and Processes

Neuronal or neuroglial, the pieces of the nervous system are of an integrative design, shaped to contact many other pieces at many places. Their integrative roles are expressed in their shapes as seen in electron micrographs: curling dendrites and thorny spines of neurons, sprawling processes of astrocytes, and fenced-in clusters of axons clearly delimit discrete territories in the puzzle (Fig. 16).

C. Interdependent Ordered Connections

Connections in the nervous system are highly ordered. In tracts and nuclei, maplike representations of functional domains are found: a sensory surface, a pattern of muscles or movements calling up particular muscles, and some part of the body, including the brain itself, that requires service. Sometimes, many maps of a venue are found in many places along many routes, allowing parallel processing of details (quality, intensity, and location of stimuli; steady-state, change of state, or affective nature).

D. Organized Display of Inputs on Neurons

Each neuron has its view of the world, a spatial arrangement on its receptive surface of the inputs (many or few) impinging on it. The location of each input (on the dendritic tree, cell body, or initial region of the axon) indicates its source and helps to determine its efficacy in bringing about excitation, inhibition, or modulation of target cell activity.

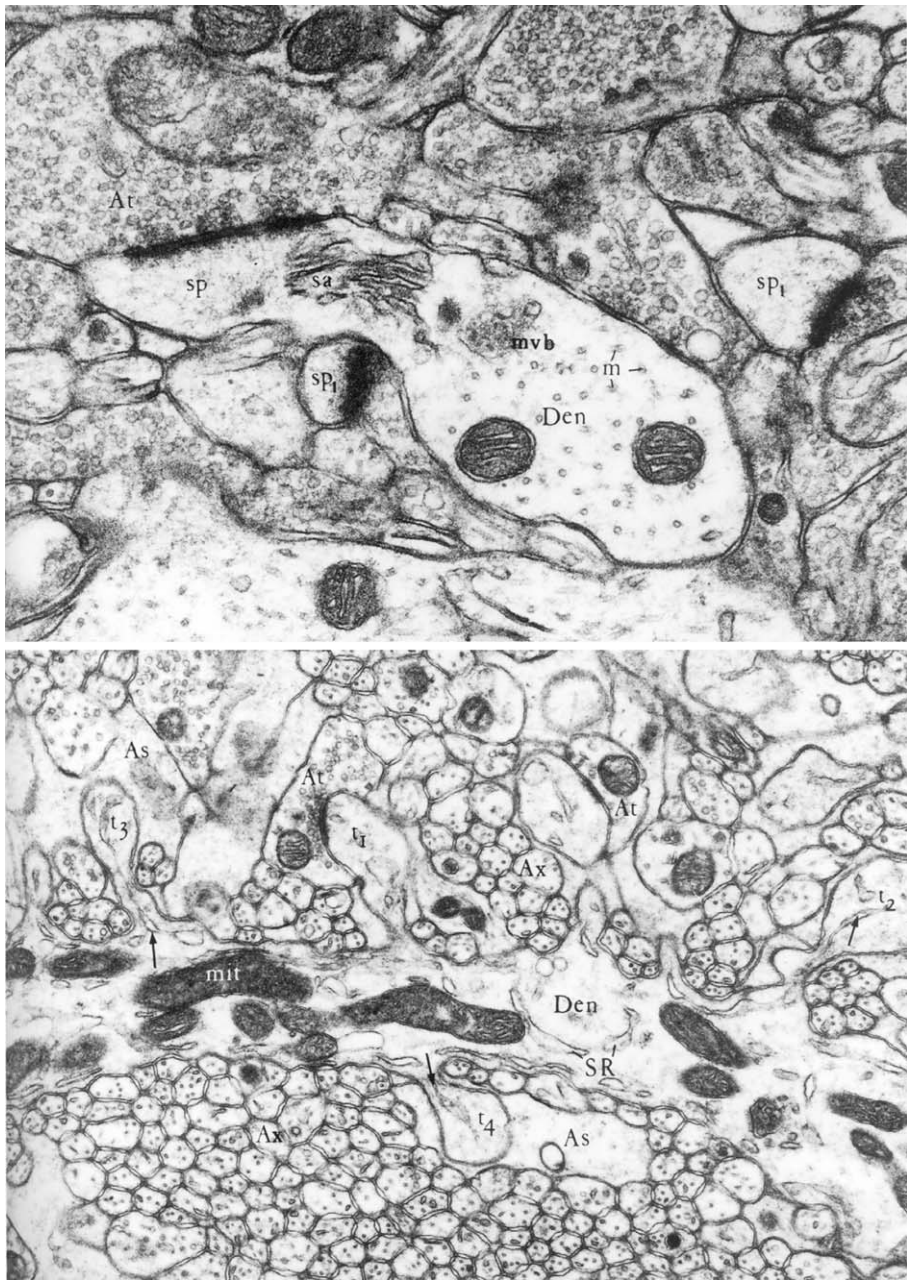


Figure 16 Dendrites: electron micrographs (EMs) show that the CNS is a hypercomplex epithelium, with billions of pieces fitting together precisely, intimately, and with evident benefit for cell-to-cell communication. Upper EM: dendrite (Den) of pyramidal cell from cerebral cortex with dendritic spine (sp) in synaptic contact with axon terminal (At) containing spherical synaptic vesicles. Lower EM: spiny branchlet of a Purkinje cell dendrite (Den) with spines (thorns; t1–t4) synapsing with axon terminals (At) of parallel fibers (Ax). See also text. From *The Fine Structure of the Nervous System: Neurons and Their Supporting Cells*, 3rd ed., by Alan Peters, Sanford L. Palay, and Henry deF. Webster, copyright 1990 by Alan Peters. Used by permission of Oxford University Press, Inc. (upper EM by Katharine Harriman, lower EM from authors).

E. Graded and All-or-None Neuronal Responses

Nervous tissue shows fractionated, graded responses peculiar to the traditional receptive regions of neurons (mainly dendrites and somata) and all-or-nothing responses, which are properties of the impulse-triggering surface (initial axon segment and downstream axonal membrane at nodes of Ranvier).

F. Chemical Messengers

At a synapse, neurons use chemicals to deliver and receive impulses. Presynaptic substances are exocytosed into the synaptic cleft and received and acted upon by postsynaptic receptors. These rapidly open membrane ion channels or act via second messenger systems (the transmitter is the first messenger) to elicit changes in postynaptic activity. Such effects, with the many diverse mediators (biogenic amines, peptides, growth hormones, gases), fluctuating activity of neurons, and multiple modes of prolonging or stopping messages, permit many combinations of factors to be in play at any time. The synapse is thus a multifactorial, extremely versatile information transfer system. This type of system is physically difficult for the investigator to study and clearly more complex postsynaptically than described in many accounts.

VI. CELLULAR ELEMENTS OF THE NERVOUS SYSTEM

A. Neurons and the Neuron Doctrine

Although the neuroglia has far greater importance than we accord it, nerve cells are regarded as the fundamental genetic, anatomical, functional, and trophic (nutritive, sustaining) units of the nervous system. These attributes are the four tenets of the neuron doctrine, a once-controversial, ever-authoritative, and still-evolving restatement of the cell theory first applied to nervous tissue in 1891. Fifth and sixth tenets were quick to follow.

In 1890, the histologist Santiago Ramón y Cajal, considering neurons as separate cells and tirelessly delineating their connections, inferred that impulse traffic in a neuron flows one way: from dendrites and soma to axon and axon terminals or from receptor parts to effector parts. In 1897, the physiologist Charles Sherrington proposed the term *synapse* for the contact between nerve cells and suggested that the one-way sensorimotor traffic in spinal reflexes rested on a one-way valvelike action of the synapse. So a fifth tenet, Cajal's law of *dynamic polarization* (like neuron doctrine, strong words), was added and then a sixth, his law of *connectional specificity*: synapses are not random, but determined.

B. The Neuron Doctrine Revisited

Electron microscopy conclusively shows that nerve cells are separate, but the neuron doctrine needs careful revision. Indeed, this has gone forward. Four issues spurred its reevaluation.

1. Gap Junctions, Cell-to-Cell Influences, and Electrical Synapses

Small, roughly circular plaques (0.1 μm to many micrometers across) of intimate apposition (2–3 nm) unite the membranes of epithelial cells. Here, tiny membrane particles (connexons) with minute pores (<2 nm) in each membrane and in register permit the flow of ions and small molecules (amino acids, cyclic AMP, tracers) between cells. Such a contact is a *gap junction*, nexus, or communicating junction. The last term is definitive: they permit cell-to-cell percolation of nucleotides and peptides that may coordinate cellular activity. They also offer low-resistance electrical coupling of adjacent cells, as in embryonic tissues. In neural development, cell-to-cell communication by gap junctions in epithelial tissues plays a role in major events, as in gastrulation and neural plate closure.

For a time, gap junctions challenged the second tenet of the neuron doctrine. They allowed chemical commerce between neurons and united them electrically. But controversy subsided. Though present in a few neurons in a few places in the adult and more widely encountered in the early CNS, such areas of interaction between separate cells do not undermine the unitary nature of neurons.

2. Versatility of Neuronal Processes and Complexity of the Neuropil

An old problem in studying neurons (especially invertebrate ones) is identifying their extensions as dendrites or axons. Some processes play both roles. Such versatility violates the law of dynamic polarization tacked on to the neuron doctrine: that dendrites and cell body receive impulses and the axon and its terminals conduct them away and deliver them elsewhere. But the parts of neurons are adaptable. Dynamic polarization was based on a neuron considered to be typical: the motor neuron.

Today, we know that the neuronal surface is a *mosaic* of receptor and effector parts and that local circuits with different polarizations need no spikes but only graded potentials to function. Cajal's law must be viewed in perspective. It was a great step in

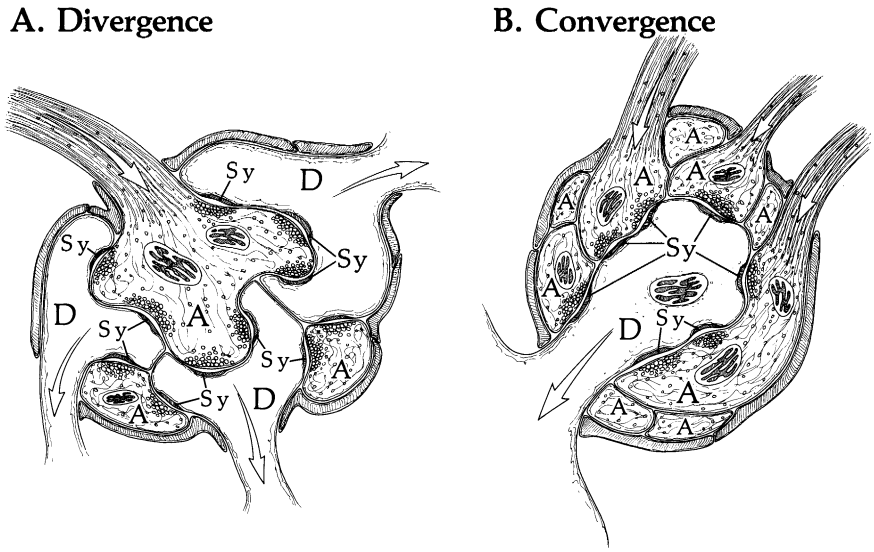


Figure 17 Glomeruli, resembling renal capillary glomeruli and long known in the cerebellum and olfactory bulb; many more complex tangles of intricately intertwined and synapsing neuronal processes of multiple origin are now recognized elsewhere. The interactions of processes in many of these knots are not known, but as shown in this schematic drawing, some provide for divergence or convergence of information. At left, a glomerulus in the cerebellar cortex centered on a single mossy fiber axon terminal; at right, a glomerulus in the thalamic pulvinar centered on a single departing dendrite. From Jay B. Angevine, Jr., (1988) Dendrites, axons and synapses, *BNI Quarterly* **4**(2), 9–19. after T. Bullock, R. D. Orkand, and A. D. Grinell (illustration by Steven J. Harrison) by permission of Barrow Neurological Institute.

understanding neuronal activity, but it cannot account for the almost limitless interactions between neurons.

Related problems are *glomeruli* (Fig. 17), which are knots of neuronal processes of diverse morphology and origin in which group synapses (rather than the familiar paradigm of one-on-one) are the rule and complex transactions of many cells, near and afar, are afoot. Here the independence of neurons is diminished, but not abrogated, by the complexity of the *neuropil*, the feltwork of neuronal and glial processes in which most synapses are found and most CNS business is transacted. Awareness of the epithelial nature of nervous tissue helps to explain these regions of multineuronal interaction.

3. Neuronal Teamwork and Distributed Systems

A critique of the functional tenet of the neuron doctrine comes from the growing awareness that the nervous system employs groups of neurons, not necessarily all in one place, to perform its tasks, and not single cells themselves. Striking examples are the progressive feature analyses by sensory systems of the CNS (notably the visual), the cooperative interplay of neurons at all levels of the motor system, and the communal organization and stepwise connections of the cerebral cortex. A fitting term for this team concept of neuronal function is *distributed system*. As a key principle that is crucial to understanding cognitive functions and disorders, it does not impugn the individuality of neurons in any respect. However, it lessens their individual importance as determinative functional or gnostic units.

4. Transneuronal Degeneration

Neuronal responses to axon interruption include degenerative changes in the severed axon and its investments, as well as reparative changes in the cell body. Undamaged neurons nearby do not react, nor does the degeneration usually involve the postsynaptic cell, with certain major exceptions: atrophy of skeletal muscle is a familiar consequence of denervation. These results illustrate tenets one and four: the anatomical individuality of neurons and the trophic importance of their cell bodies. But with a century gone by and better methods for evaluating and following up effects of injury upon neurons, the concept of a single neuron as *the* trophic unit is hardly tenable.

Transneuronal degeneration is now well-recognized. It was first noted in pathways in which a given neuron largely depends on another for its input. In the visual system, cells of the thalamic lateral geniculate nucleus degenerate after lesions to the retina, where the

ganglion cells that project to them lie. Transneuronal effects are now known in many neuronal sequences in the CNS and PNS. Even more remarkable is that transynaptic effects of injury or disuse may extend in either direction (retrograde and orthograde) and involve neurons one or more than one synapse removed (primary, secondary, tertiary, etc.). Neurons in circuits seem to depend on one another in ways that go beyond receipt and delivery of impulses. Their metabolic equilibrium may derive, in part and varied measure, from their interactions. These possibilities are relevant to interpreting experimental lesions or treating neurologic diseases. Insights on the factors underlying these trophisms may explain how the nervous system maintains itself and what it could do (with assistance) to compensate for injury or cell loss.

C. The Neuron Doctrine Today

As the *genetic unit*, the neuronal nucleus contains the chromosomal DNA, the human genome of 30–40,000 genes. A high percentage probably concern the nervous system, whereas the degree of alternate processing and the number of proteins expressed, though unknown, may be far greater.

As the *anatomical unit*, the picture proof by Sanford Palay and George Palade in 1954 that at the synapse the apposed neuronal membranes are separated by a 10 to 20-nm cleft was a triumph for electron microscopy. But for gap junctions (see previous discussion) and fine filaments seen in the EM crossing the gap between pre- and postsynaptic cytoplasm, doubts about the separateness of nerve cells were laid to rest.

As the *functional unit*, modification of the doctrine is mandatory: groups of cells are the working units of the nervous system, not single cells themselves.

As the *trophic unit*, evidence has accumulated over the last century (in neurology, neuropathology, and experimental biology) to show that the cell body of a neuron is vital to its *own* life. If its processes are severed or injured but the soma is left intact or little harmed, new ones may be formed, or at least the damage controlled. But if the soma is severely damaged, the nerve cell dies.

Still (to paraphrase neuroanatomist Alf Brodal), "The concept of a neuron as a trophic unit is scarcely acceptable as originally formulated. A single neuron is dependent on other neurons for its viability." In this regard, neurons and their processes exert nurturing and supporting effects on other tissues, but such trophic interactions have not been examined as fully in the nervous system itself.

Studies of *neurotrophisms* are numerous, and they continue. But trophic functions of neurons have not been investigated as intensively as their communicative and endocrine abilities. Less is known about them. Their clinical significance is profound, as in managing patients with spinal cord injury in which muscular atrophy and decubitus ulcers are secondary to even more debilitating problems.

Finally as to trophisms, the multitudinous, ubiquitous neuroglial cells deserve attention. The neuron doctrine was formulated a century ago for neurons and for good reasons. But, as it was for the laws of dynamic polarization and specified connections, a corollary for the neuroglia could be added. Its production, by astrocytes, of numerous *growth factors* supporting vital attributes of certain neurons merits mention in relation to the trophic tenet of the revised neuron doctrine.

D. Epithelial Design of Nervous Tissue

As stated, nervous tissue is a specialized epithelial tissue. Epithelia are sheets of cells that cover surfaces or line cavities of the body and in places grow into subjacent connective tissue to form hair follicles, glands, or other derivatives. Features of epithelia include close aggregation of functionally related cells, small amounts of intercellular substance, a basement membrane, a free surface, and utter avascularity. Nervous tissue, wherever encountered clearly shows the first two attributes and the free surface and basement membrane in the CNS as well. The presence of blood vessels might seem to depart from epithelial design, yet they are not part of the nervous tissue but only pass through it, ensheathed by basal lamina.

Nervous tissue meets all structural criteria of epithelial tissue. It arises from epithelium: the embryonic neural plate. As epithelial duties include sensory reception, protection of the body, and secretion of hormones, nervous tissue performs these functions to a fault, offering general body responsiveness, integration, and chemical transmission. It is thus properly and usefully classified as epithelial tissue. Where the resemblance of nervous tissue to epithelium ends is in the configurations of its cells. They have far more complex shapes than the squamous, cuboidal, columnar, and domelike cells of other epithelia. With its many wavy processes, a neuron looks like a medusa.

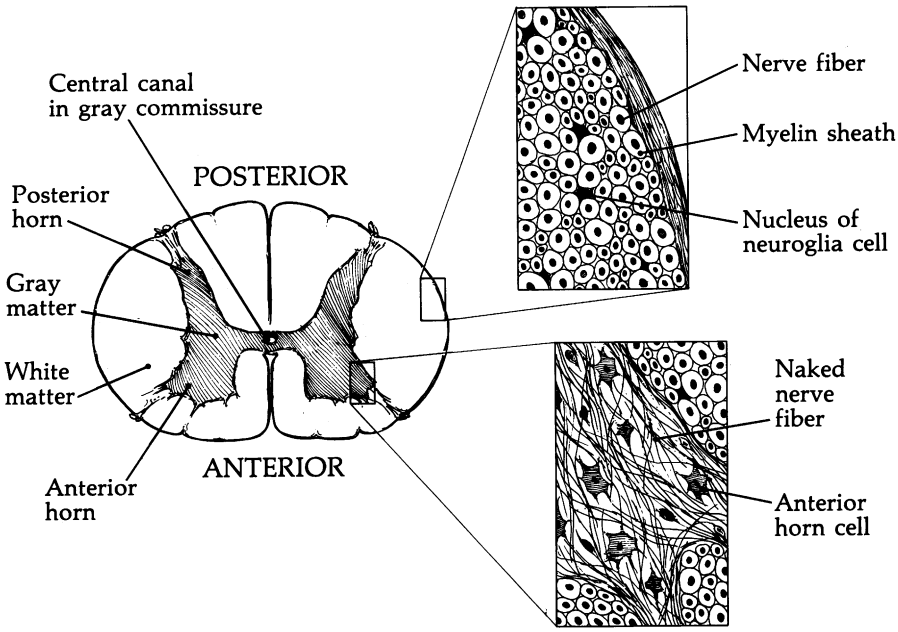


Figure 18 Gray and white matter: light micrographs (LMs) of the CNS fail to show its epithelial nature. Myelinated axons in the white matter resemble closely packed strands of spaghetti; nerve cell bodies in the gray seem to float in a watery matrix. At first glance in hematoxylin and eosin preparations, the CNS looks much like loose areolar connective tissue. From A. W. Ham, *Histology*, 6th ed., J. B. Lippincott, 1969 (illustration by Steven J. Harrison, from Dorothy Chubb).

Light micrographs of gray and white matter in the CNS (Fig. 18) convey a misleading impression, failing to show the epithelial design of the nervous system. Nerve fibers in the white matter resemble strands of spaghetti, and nerve cell bodies in the gray float here and there amid processes in a watery substance, like the fibroblasts and collagen fibers in the fluid extracellular matrix of connective tissue.

Yet nervous tissue is epithelial throughout. Little intercellular matrix exists, mainly cells and their processes. A schematic drawing (Fig. 19) shows this epithelial character. Two of the three meninges covering the CNS, the arachnoid and pia mater, separated by the subarachnoid space, lie externally. Beneath them, the CNS is bounded by a basement membrane (basal lamina). Such membranes separate each of the four basic tissues of the body, preventing contact and unregulated commerce between them and thereby serving tissue homeostasis. An excellent example is the basal lamina between axon terminals and muscle fibers at the intimate nerve–muscle synapse.

Internal to the basement membrane are three elements seen in CNS tissue: neurons, neuroglia, and blood vessels. As noted, epithelia are avascular. For oxygen and nutrients, they depend on vessels in nearby connective tissue. But in the CNS superepithelium, the volume of tissue and the never-ending demand for oxygen and glucose mandate their proximity. The blood vessels entering and exiting the CNS and their interposed capillary networks (so dense that neurons lie no more than 100 μm away) are *extrinsic* as stated, separated from nervous tissue by the basal lamina of the CNS and that of the vessels. Substances from the blood must traverse both laminae to enter the CNS and then an interposed astrocytic process before reaching a neuron.

The inner free surface of the CNS epithelium comprises a single layer of ciliated ependymal cells lining a brain ventricle. Nearby, in attenuated regions of the brain where the pia and ependyma are apposed, modified ependymal cells form the choroid epithelium, which produces the cerebrospinal fluid (CSF). Its cells are specialized for secretion, with basal infoldings, free surface microvilli, many mitochondria, and tight junctions to limit the passage of substances larger than ions.

Epithelial design is striking in electron micrographs of the CNS. Figure 20 shows the soma of a spinal motor neuron, replete with nucleus, nucleolus, and surrounding organelles. Proximal parts of three dendrites are visible, as well as other dendritic, axonal, and glial processes around the cell and two capillaries. The

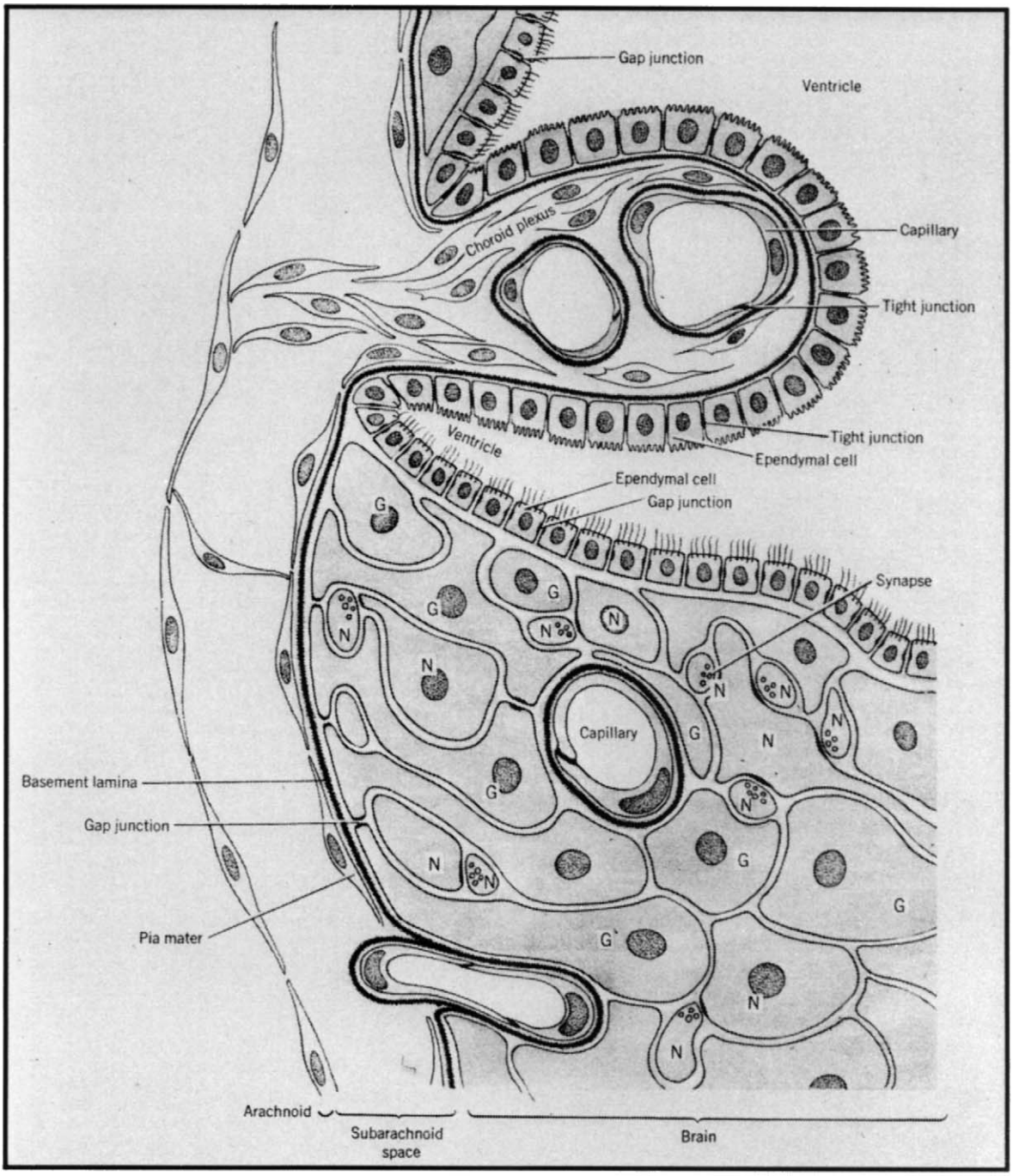


Figure 19 Epithelial nature of the CNS: neurons (N) and glial cells (G) lie within a basement lamina (basal lamina), like the diverse cells of conventional epithelia. Unlike other epithelia, however, is the presence of blood vessels; the volume of tissue and unceasing demand for oxygen and glucose mandate them, but they are separated from CNS parenchyma by their own basal lamina and that of the CNS. Blood-borne gases and substances must traverse both laminae to enter or exit the CNS. From C. R. Noback and R. J. Demarest, *The Human Nervous System. Basic Principles of Neurobiology*, 2nd ed., McGraw-Hill, New York, 1975 (illustration by Robert J. Demarest).

key point: everywhere the tissue is epithelial. Every nook and cranny is occupied, with the 2–4 nm of extracellular space invisible at this low magnification. The capillary basal lamina is barely apparent. The darkly encircled fibers in the field are small, myelinated axons of arriving input.

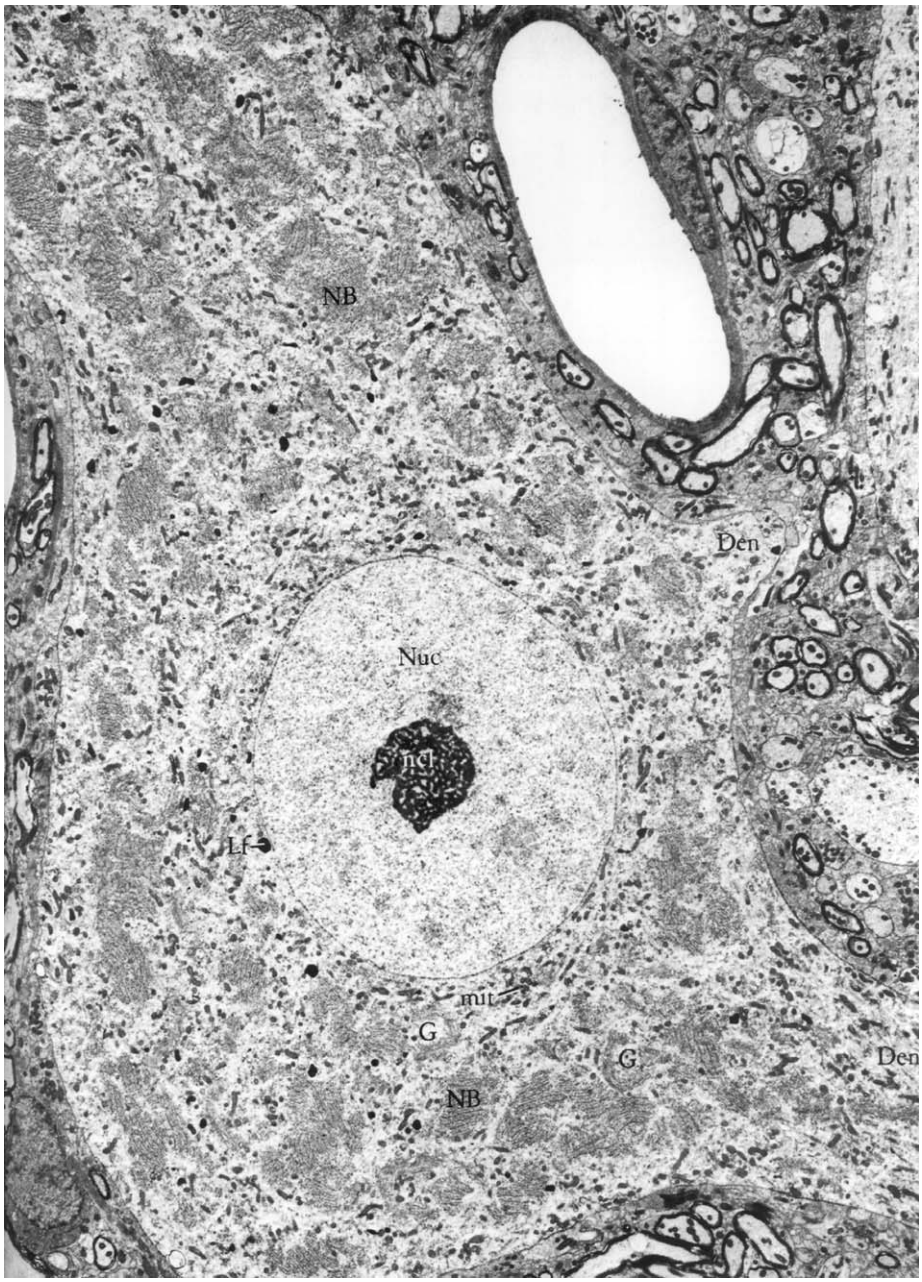


Figure 20 Neuronal cell body: low-magnification EM of a motor neuron. Nucleus (Nuc) is pale and contains a nucleolus (ncl). Perikaryon shows Nissl bodies (NB), Golgi apparatus (G), mitochondria (mit), and a few lipofuscin granules (Lf). Extending from the perikaryon are three dendrites. Other dendrites, as well as axonal and glial processes and two capillaries, lie nearby. See also text. From *The Fine Structure of the Nervous System: Neurons and Their Supporting Cells*, 3rd ed., by Alan Peters, Sanford L. Palay, and Henry de F. Webster, copyright 1990 by Alan Peters. Used by permission of Oxford University Press, Inc. (EM from collections of authors).

E. Complexity and Individuality of Neurons

Neurons typically emit many highly branched, intermingling processes. This complexity, along with their sheer size, is a problem in studying them and a reason why the neuron doctrine was controversial. Many studies, experiments, and years of bitter debate took place before the idea that nerve cells were separate was

generally accepted. Even then, some doubted it. Only with the demonstration by electron microscopists in the mid-1950s of the minute cleft between neurons at the synapse was the physical individuality of neurons established.

F. Parts and Functional Dependencies of Neurons

1. The Cell Body

A neuron (Figs. 21E–21I) invariably has a *cell body* (soma) containing the cell *nucleus* and a more or less spherical mass of ambient cytoplasm, the *perikaryon*, filled with cell *organelles* and *inclusions*. During development, the soma is the first part of the neuron to arise, and throughout the life of the cell (with few exceptions, postmitotic and incapable of cell division) it is the irreplaceable trophic part that maintains the structural integrity and health of the neuron with its far-reaching processes.

2. Dendrites

Extending from the soma in most neurons are the *dendrites* (Figs. 21A–21B). These relatively short, initially tapering processes branch obliquely, like the limbs and branches of a tree. In varying proportions, they contain the organelles seen in the soma. In a few neurons, they arise at the end of an axon. A dorsal root ganglion cell (Fig. 15, DRG) has dendritic tufts or a specialized nerve ending (a free or an encapsulated axonal tip) in such an axon in the dermis, wall of a viscus, or other peripheral area.

Dendrites may branch elaborately (Fig. 27). They are the chief receptive processes of a neuron, even though the soma and certain proximal or distal axonal regions may also receive inputs. In their ramification, they signify the relationships of the neuron to other neurons and, hence, their integrative role. They offer points of contact (synapses) for axon terminals and for axons grazing them as they pass by (*fibers of passage*). They are the prime integrative components of neurons: receiving, combining, summing, or otherwise regulating their input.

In the multipolar neurons of the CNS (Fig. 15, the six dark cells), dendrites originate from the soma as one or more primary branches. These continue to ramify, becoming finer and finer, eventually forming branchlets or twigs. The surface area and volume of the dendrites may exceed those of the soma by many times. Incoming axons end on the dendritic shaft or on *spines* (Fig. 16, upper and lower plates), which are minute thorns, that if present in large numbers give dendrites a studded appearance.

The role of spines is not understood, but (as once surmised) they do not serve to increase the receptive surface area; intervening shaft regions may have but few synapses. Spine synapses are the major sites of excitation; each spine usually has one input, but may have several. One function spines seem to play is isolating such afferents. But it is premature to characterize them. They have electrical properties that may be modified by slight changes in the form of the spine neck, perhaps during learning or in aging, when they show deterioration early on. They seem to be specialized ports of entry for currents into the dendritic shaft. They may be bounded by *astrocytic processes* (Fig. 16, lower plate). By "spongelike" action, astrocytes serve vital roles in neural communication: taking up excess potassium ions that might interfere with local neuronal activity, localizing glutamate release for optimal elicitation of excitation in target cells, and preventing glutamate-induced toxicity, which is potentially life-threatening to neurons.

3. The Axon and Axonal Collaterals

Somewhere about the soma of most neurons emerges a single, cablelike process called the *axon* or axis cylinder. It arises from a small, conical elevation (axon hillock; Fig. 21A) or from the stem of a major dendrite (Fig. 21B). Its departure may be virtually straight, presaging a clearly directed flight to a distant location (via a tract or nerve), or meandering, as if lost in the neuronal shrubbery near its origin (Fig. 22). Such direct or rambling courses provide the long-distance trunk lines and myriad local circuits over which neuronal signaling and communication take place.

The conductile axon conveys impulses to other neurons, a muscle, or a gland. In primary sensory neurons (Fig. 15, DRG) innervating the skin, impulses arise in axonal endings: naked, with minute expanded tips, or encapsulated (A), as shown. The action potentials travel to the CNS via a cranial or spinal nerve.

Compared to dendrites, the axon is relatively long (2–3 μm to 1 m), slender (0.1–20 μm), and untapering. Like dendrites, it may branch, as *axon collaterals*, but these typically depart at right (not acute) angles. Collaterals distribute neuronal output to several or many places, depending on circuit design and how many collaterals the axon has. Axonal collateralization is as expressive of neuronal interrelationships as

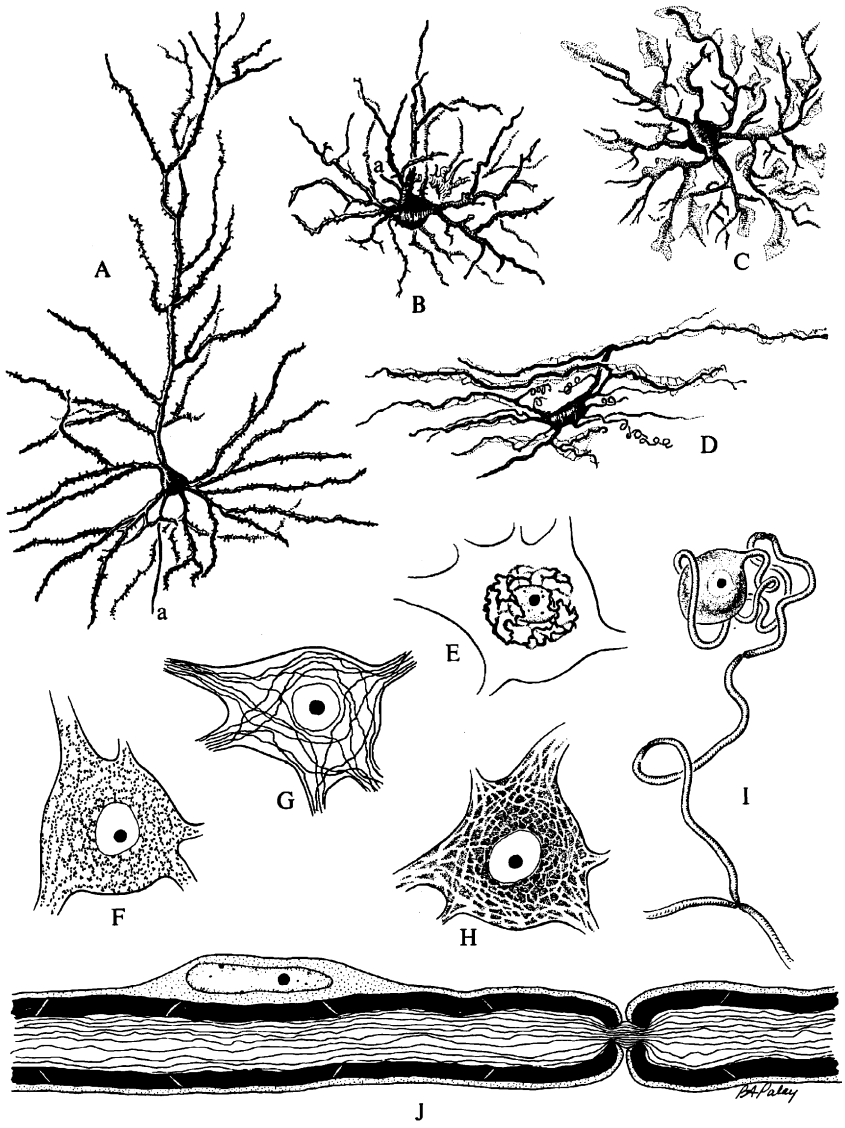


Figure 21 LM potpourri: (A) Small pyramidal cell in visual cortex, axon (a), Golgi method. (B) Small neuron in dentate nucleus of cerebellum, axon (a), Golgi method. (C) Protoplasmic (velate) astrocyte in gray matter, Golgi method. (D) Oligodendrocyte in white matter, Golgi method. (E) Spinal motor neuron showing Golgi apparatus, osmium tetroxide impregnation. (F) Motor neuron in abducens nucleus showing mitochondria, Altmann–Kull method. (G) Spinal motor neuron showing neurofibrils, Cajal's silver stain. (H) Motor neuron in abducens nucleus showing Nissl bodies, thionin stain. (I) Dorsal root ganglion cell, artist's rendition. (J) Myelinated peripheral nerve fiber, showing node of Ranvier, Schmidt–Lanterman clefts, Schwann cell nucleus, and neurofibrils, artist's rendition. From *The Fine Structure of the Nervous System: Neurons and Their Supporting Cells*, 3rd ed., by Alan Peters, Sanford L. Palay, and Henry de F. Webster, copyright 1990 by Alan Peters. Used by permission of Oxford University Press, Inc. (illustration by Betsy A. Palay).

dendritic branching. Dendrites signify the integrative power of a neuron and axons and their collaterals its distributive power, as well as the routes and addresses of impulse dissemination.

The axon is usually single, but can be double, as in bipolar neurons, or missing, as in *amacrine* cells (having no long processes), which intercommunicate through their dendrites. It is thinner and much longer than any dendrite, but most CNS neurons have a greater total length of dendrites than axons.

Single or double, short or long, branched or not, if an axon is present, it is the functional axis of the neuron, the hub along which nerve impulses travel. Over this main line speed the "all-or-none" action

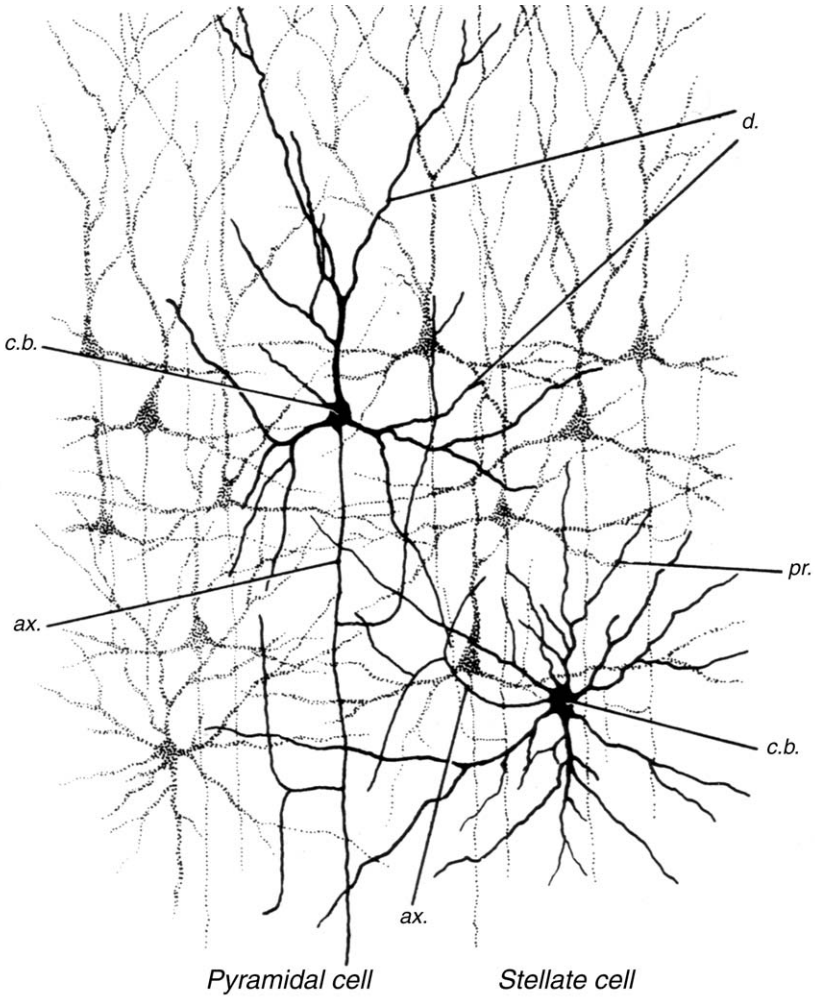


Figure 22 Direct and wandering axons: the pyramidal cell and stellate cell of the cerebral cortex shown here typify the two major modes of axonal departure: in the one, virtually straight, presaging flight to a distant location (e.g., the spinal cord), and in the other, meandering among the dendrites of the neuron of origin and those of others nearby. The former course characterizes the long-distance trunk lines of the CNS and the latter its myriad local circuits. Abbreviations: dendrites (d and pr.), cell body (c.b.), axons (ax.). From J. Z. Young, *The Life of Mammals*, 1957, copyright 1957 by Oxford University Press, Inc. Used by permission of Oxford University Press, Inc. (preparation by D. A. Sholl).

potentials (spikes), encoding output in spike trains or bursts superimposed on a discharge frequency characteristic of the neuron under given conditions. In motor neurons and other large nerve cells, these spike potentials originate in the axon hillock and axon initial segment, the "spike trigger zone" (Fig. 23).

4. Axon Terminals

Specialized terminations of axons or of their preterminal branches, these have many shapes and names: bulbous, buttonlike, claw-shaped, and so forth. Similar beaded structures may lie along the course of an axon. All are sites where frequency-coded messages are transmitted chemically to some part of the target neuron or the muscle fiber across a synapse or neuromuscular junction and where trophic effects may pass in both directions. Axon terminals are the effector parts of neurons, whereas dendrites, the cell body, and axons are the receptive, trophic, and conductile parts, respectively.

5. Flexibility of the Parts of Neurons

The parts of neurons are adaptable. There are many exceptions to these conventional definitions, and

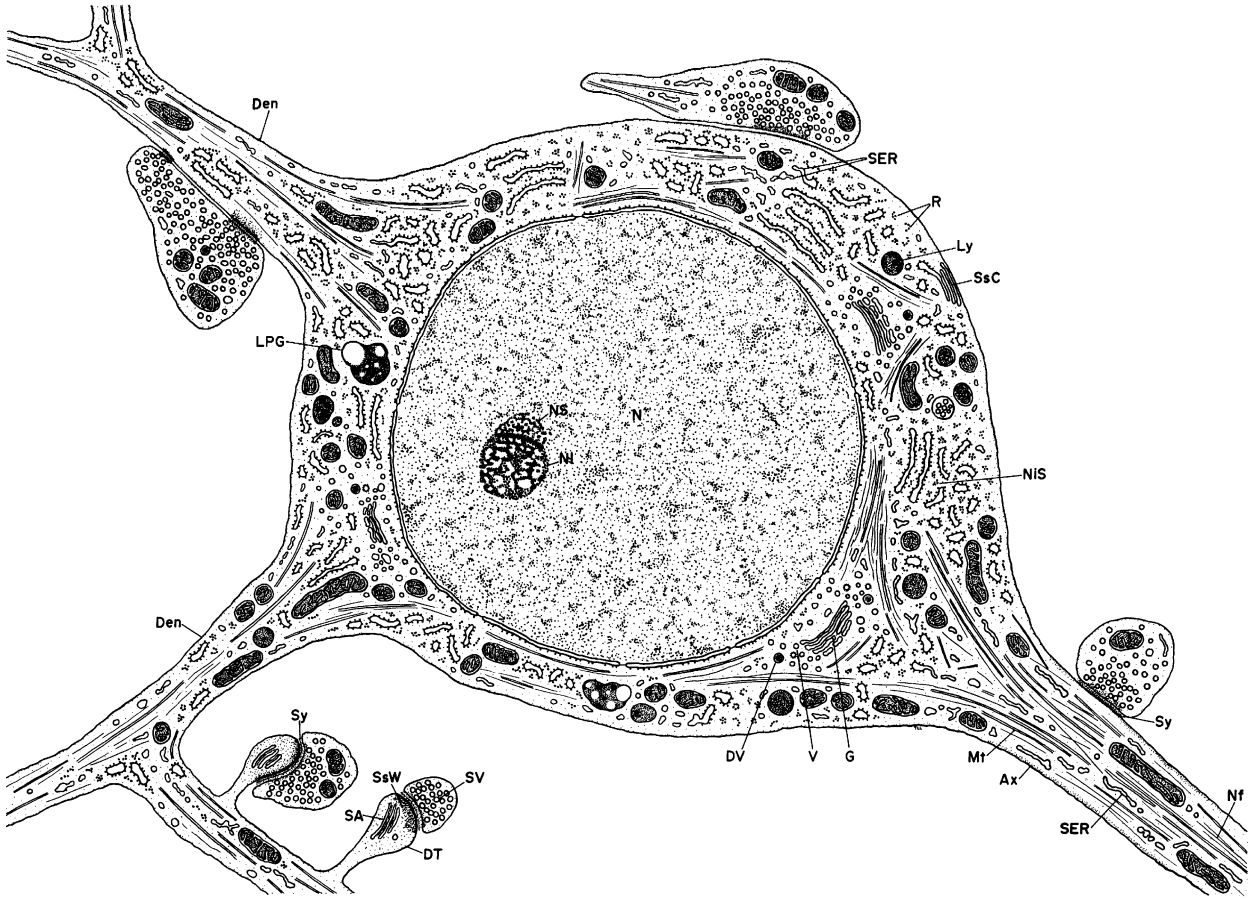


Figure 23 Schematic EM of neuron, its organelles, and inclusions: dendrites (Den), axon (Ax), nucleus (N), nucleolus (Nl), nucleolar satellite (NS), Nissl substance (Nis), rough-surfaced endoplasmic reticulum (not labeled), ribosomes (R), smooth-surfaced endoplasmic reticulum (SER), subsurface cisternae (SsC), Golgi apparatus (G), dense core vesicles (DV), lysosomes (Ly), lipofuscin pigment granules (LPG), microtubules (Mt), neurofilaments (Nf), dendritic thorns or spines (DT), spine apparatus (SA), synapses (Sy), synaptic vesicles (SV). From T. L. Lentz, *Cell Fine Structure. An Atlas of Drawings of Whole-Cell Structure*. W. B. Saunders Co., Philadelphia, 1971.

awareness of them enhances understanding of neural function and neurologic disease. Certain regions of the axon may serve receptive functions: with axoaxonic input at the axon hillock or on its presynaptic end bulbs. Dendrites may conduct impulses swiftly in an all-or-none manner like axons (as in the towering hippocampal pyramidal cells). They may also act as effectors (like axon terminals) and transmit activity fractionately, as in dendrodendritic synapses. Thus, the chemical membrane and subjacent organelles of a neuron constitute a *functional mosaic*.

It is helpful to regard neuronal design as flexible. Almost any part can perform any communication function, if circumstances or circuit design make it advantageous. The only invariant component of a neuron is the cell body. It, and only it, is the trophic part.

Flexibility is striking in relationships between neurons. Neuroscientists used to think in terms of serial progressions of neurons through axodendritic and axosomatic synapses and of similar sequences through axonal collateral channels. Now other arrangements are known (Fig. 25): *parallel* coupling (axoaxonic and somatosomatic synapses), *reciprocal* coupling (dendrodendritic synapses), *parallel–serial* coupling (axoaxodendritic synapses), and *atypical* coupling (somatoaxonic synapses). These more complex modes of cell-to-cell interplay, along with gap junctions and the unpredictable results of synaptic activity, illustrate the versatility of neural communication. More than

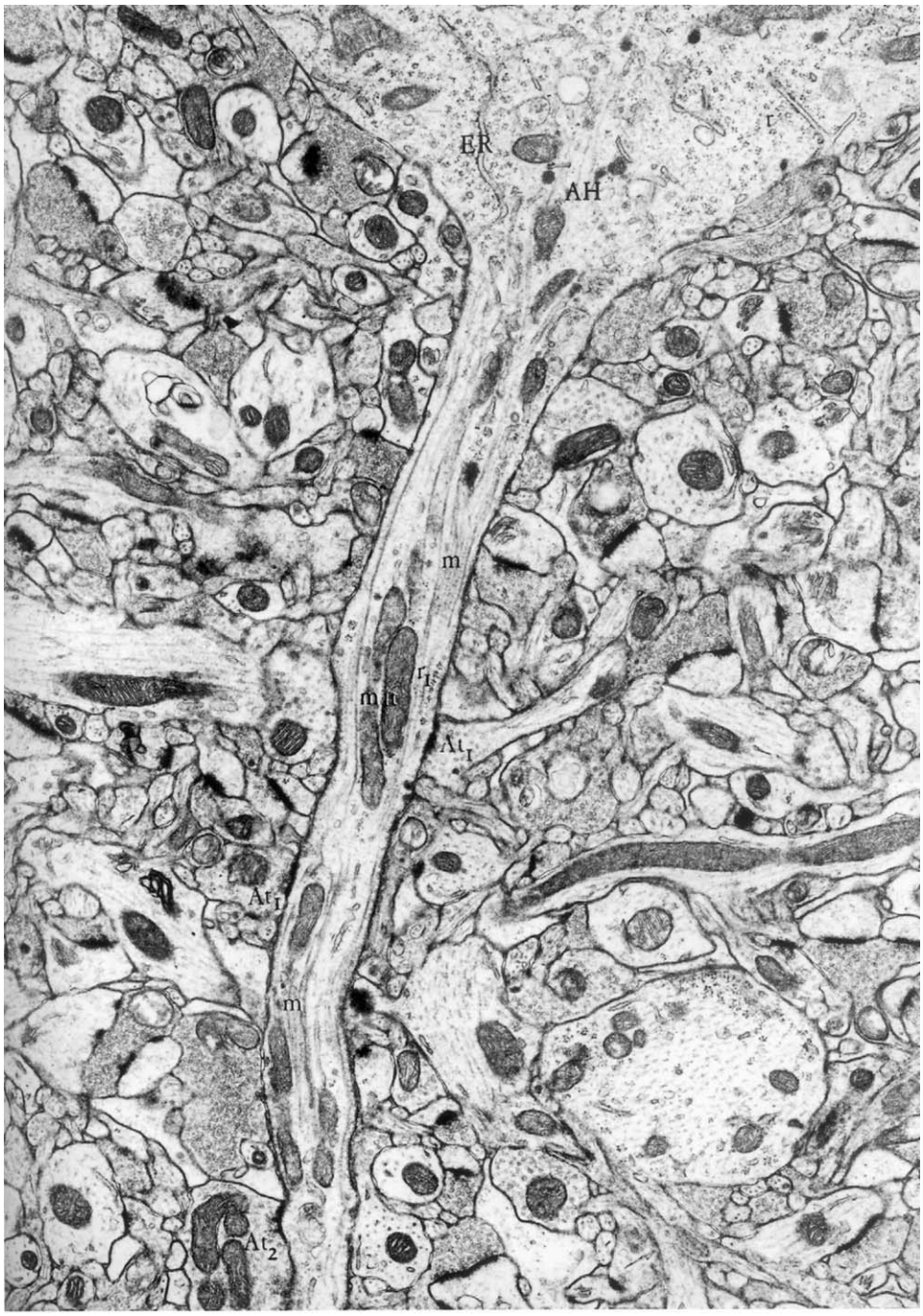


Figure 24 Axon hillock and initial segment: the axon is shown leaving the perikaryon of a pyramidal neuron at the axon hillock (AH), wherein ribosomes (r) and rough endoplasmic reticulum (ER) are seen, and descending through richly varied and epithelially organized neuropil. In the initial segment, ER is absent and ribosomes (r1) decrease in number. Mitochondria (mit) and neurofilaments enter the axon freely; microtubules (m) also enter but become bundled. Such bundling and the presence of an undercoat to the axonal plasma membrane characterize the initial segment. Three axon terminals are seen synapsing with the initial segment: two proximal (AT1) and one distal (AT2). From *The Fine Structure of the Nervous System: Neurons and Their Supporting Cells*, 3rd ed., by Alan Peters, Sanford L. Palay, and Henry de F. Webster, copyright 1990 by Alan Peters. Used by permission of Oxford University Press, Inc. (EM from collections of authors).

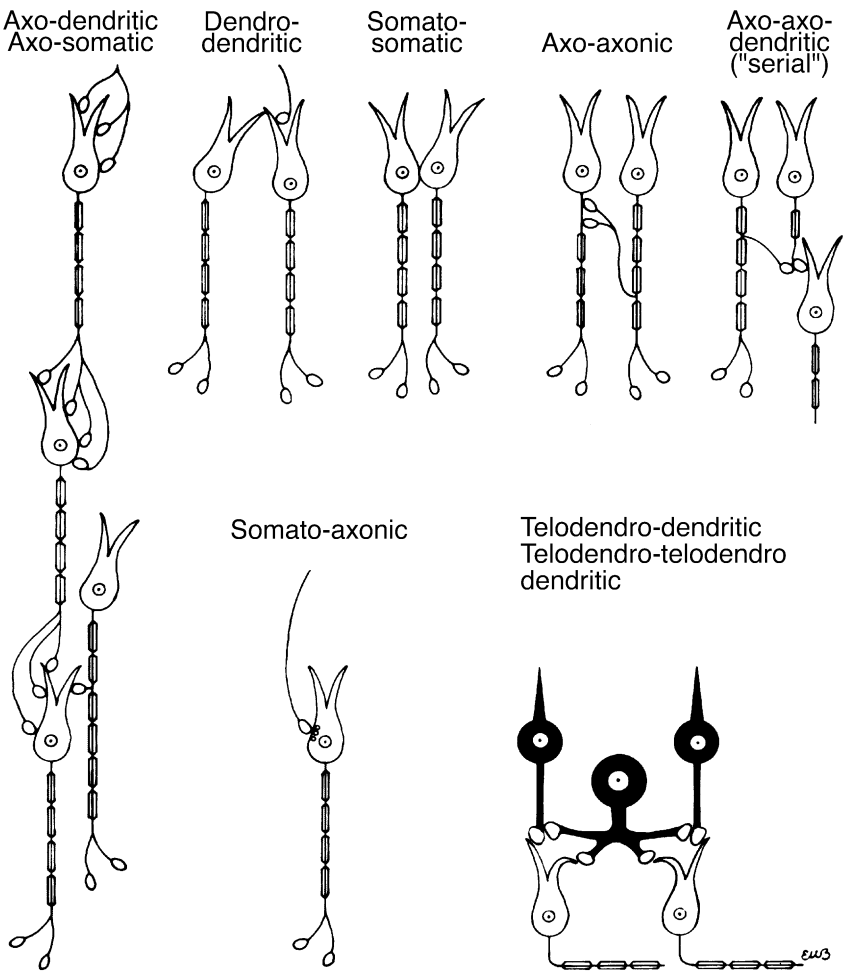


Figure 25 Neuronal coupling paradigms: for a century, neurons were considered to be coupled linearly by axodendritic and axosomatic synapses. But by 1972, as shown schematically here by David Bodian, it was clear that many paradigms of parallel coupling existed. These illustrate the adaptability of the parts of neurons, the flexibility of neuronal design, and, as Dr. Bodian observed, "the diversity of life and its processes." See also text. From Neuron junctions: a revolutionary decade, D. Bodian, *The Anatomical Record*, Copyright © 1972. Reprinted by permission of Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc. (illustration by Elinor W. Bodian).

that, they show, as neuroanatomist David Bodian observed, "the diversity of life and its processes."

6. Polarity of Neurons

In terms of *polarity*, (number of cytoplasmic processes), three kinds of neurons are recognized: unipolar, bipolar, and multipolar (Fig. 26). The probable location and functional significance of each may be inferred by inspection, but an explanation is necessary.

Unipolar *neurons* are not found in vertebrate nervous systems, although young neurons have but one process at certain developmental stages and look like them. But in invertebrates, they represent the dominant population and, hence, the largest number of nerve cells on earth.

Bipolar neurons are simple modifications, with fusiform cell bodies, of the columnar epithelial cells from which they evolved. In humans, they form primary sensory neurons in the olfactory epithelium, retina, and vestibulocochlear ganglia. Terminal ramifications in the periphery (e.g., in the organ of Corti) respond fractionately to stimuli and may be considered distant dendrites. By contrast, the long process leading to the soma, like the other process departing, is by all criteria axonal, even having a myelin sheath for increased

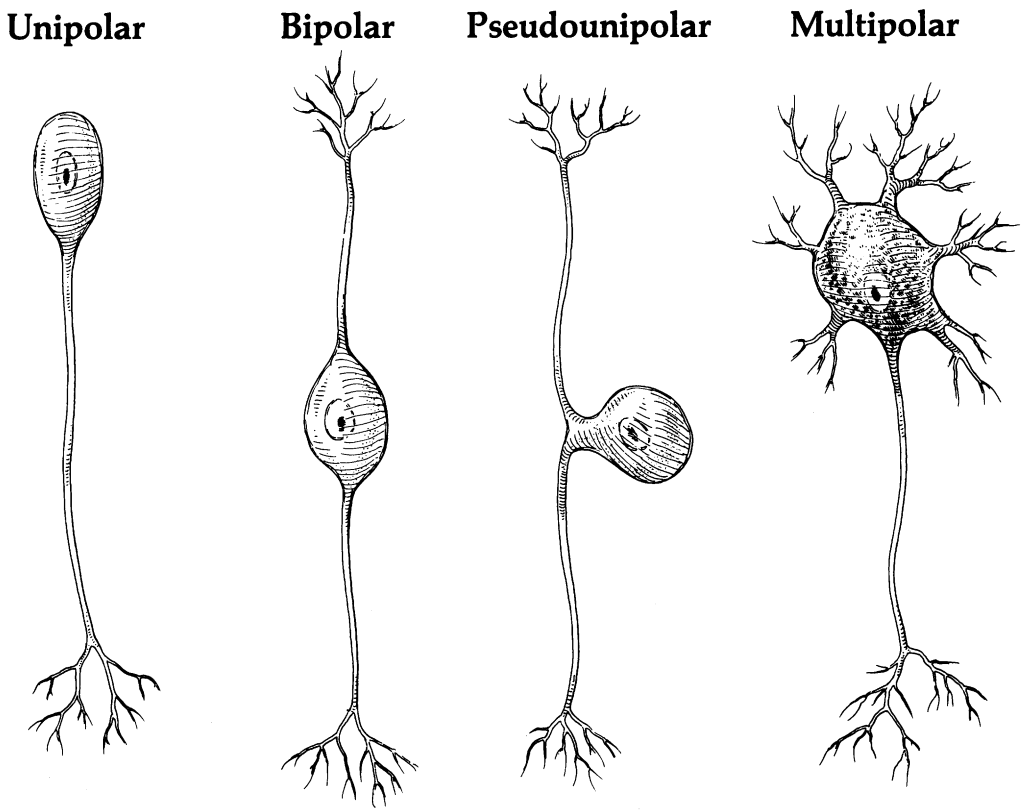


Figure 26 Polarity of neurons. In number of cytoplasmic processes, three general kinds of neurons are recognized: unipolar, bipolar, and multipolar. True unipolar neurons are not found in the adult vertebrate nervous system. Bipolar neurons and a variant, pseudounipolar neurons, make up all the primary sensory neurons of the PNS. Multipolar neurons have many variably branched processes extending in many directions; as the most common type of vertebrate neuron, they are the hallmark of the human CNS. See also text. From *Principles of Neuroanatomy*, by Jay B. Angevine, Jr., and Carl W. Cotman, copyright 1981 by Oxford University Press, Inc. Used by permission of Oxford University Press, Inc. (illustration by Steven J. Harrison).

speed of impulse conduction. By the interposition of an axonal cable between the receptive region and the cell body, spikes triggered peripherally pass swiftly to the soma, over which they flow to the other process and on to the CNS, wherein their central ramifications distribute impulses to secondary sensory neurons.

Pseudounipolar *neurons* are modified bipolar cells. The opposing processes shift around the soma during development and combine into one, at least proximally. This process takes a short, often convoluted course and then branches like the letter T: one branch going to the periphery and the other to the CNS (Fig. 21I), as in bipolar neurons. Nerve impulses in these cells pass from one branch to the other, subsequently "backfiring" the single process and soma. Pseudounipolar and bipolar cells make up all primary sensory neurons in the PNS. Both have limited integrative domains in distant dendritic tufts and no input on their somata, which are limited to trophic and housekeeping duties. Their central terminals, however, receive presynaptic endings, providing efferent modulation of their transmission to secondary sensory neurons. This arrangement is important in suppressing the receipt of painful stimuli. It is an enkephalinergic component of a complex brain stem analgesia system.

Multipolar neurons have many variably branched processes projecting in many directions. The most common type of vertebrate neuron, they comprise virtually all neurons of the human CNS. With their diversified input (from 10^4 to 3×10^5 apiece), they have tremendous integrative capacity. In the shape, size, and position of their somata, in the number, length, and branching pattern of axons and dendrites, and in neuroactive substances synthesized and released, a panoply of multipolar neurons is found in the CNS (Fig. 27). This variety is so extraordinary that neurobiologist Pasko Rakic has said, "I believe it is safe to estimate that there are more forms of cells in the

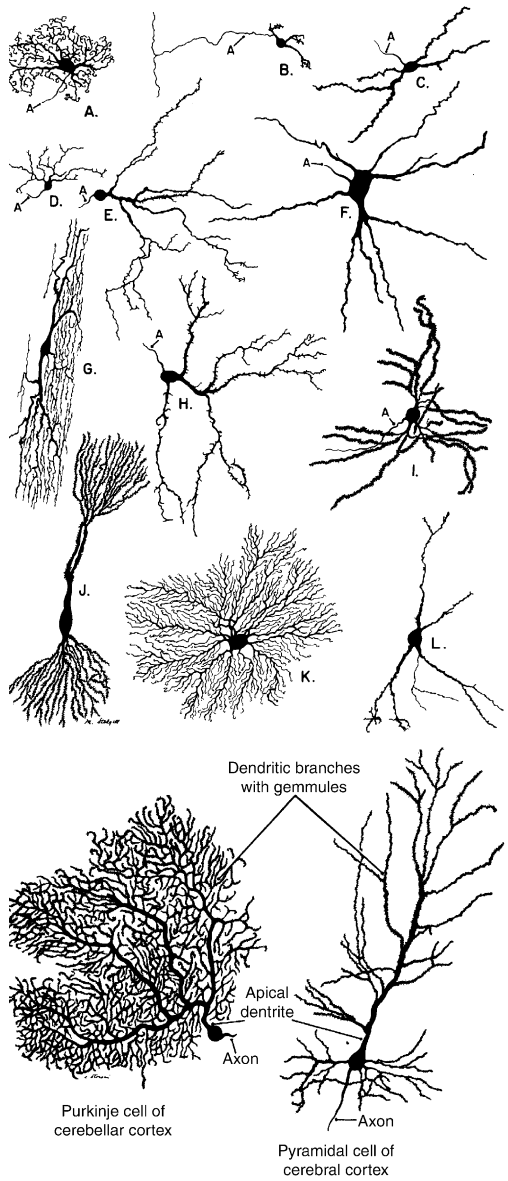


Figure 27 A panoply of neurons: a huge variety of multipolar neurons is found in the CNS. The dendrites express the integrative power of a neuron, the axon with its collaterals (or lack of same), its sphere of influence, the cell body with its minute to huge size and varied shape, its area available for axosomatic synapses, the length of its axon, and its amenability to high or low packing density. (A) Neuron of inferior olive. (B) Granule cell of cerebellar cortex. (C) Small cell of reticular formation. (D) Small gelatinosa cell of spinal trigeminal nucleus. (E) Ovoid cell of nucleus tractus solitarius. (F) Large cell of reticular formation. (G) Spindle-shaped cell, substantia gelatinosa of spinal cord. (H) Large cell of spinal trigeminal nucleus. (I) Neuron of putamen. (J) Pyramidal cell of hippocampus. (K) Cell from thalamic nucleus. (L) Cell from globus pallidus. At bottom, scaled drawings of a Purkinje cell of the cerebellar cortex and a pyramidal cell of the cerebral cortex for comparison. From Malcolm B. Carpenter and Jerome Sutin, *Human Neuroanatomy*, 8th ed., J. B. Lippincott, 1983 (drawings of Golgi-impregnated material by M. Stogdell for Dr. Clement Fox).

brain of a mammal than in all the rest of the entire body." Functionally, their vast number (100 billion, maybe 1 trillion) falls into two groups: a small number of motor neurons (about 2 million), sending axons to muscles and glands, and the remainder a host of interneurons.

As stated, the dendrites of multipolar neurons are the most intriguing variables, expressing their integrative power. They vary from few to profuse, from straight and smooth to curving and spiny, and from meagre shrubbery to magnificent arboreal extravaganzas. The axon comes next, being short or long, unbranched or rectilinearly, almost obsessively collateralized, and ruler-straight and departing the locale or sinuous and wandering through the local neuropil. Then the soma: ovoid, spherical, pyriform, fusiform, fat, skinny, big, small: miniscule (4 μm, half the size of a red blood cell) in dwarf neurons to huge (150 μm) in giants. Teased out in a tissue preparation, the Betz cells in the human cerebral cortex can be seen with the naked eye.

7. Specialization of Neurons

The variety of cell structure in the human nervous system is far more complex than in other tissues. Whereas cells elsewhere (e.g., erythrocytes, hepatocytes) are highly redundant, neurons are highly individualized. Indispensable, decision-making *command neurons* are known in invertebrate nervous systems (e.g., in the ganglia of certain arthropods). But in vertebrates, a great redundancy of neurons seems the rule: to iron out individual peculiarities of cells and for safety's sake.

The human nervous system has many neuronal subpopulations. Within each, a neuron shows its individuality by form, function (speed of response, endurance, or fatigability), position in a circuit, chemical coding of impulses, and sheer size. Neuronal specialization gives our nervous system speed, flexibility, fidelity, and staggering integrative power.

8. Schemes of Classifying Neurons

To understand such complex cells as neurons, it helps to have some means of categorization. Many classifications exist, each with advantages and limitations. Neurons are grouped as to size, shape, polarity, dendritic pattern, axonal length, long-distance or local-circuit service, presence and nature of neurosecretory activity, neuroactive substances synthesized and liberated, and identifiability as a circuit

component from one nervous system to another. But perhaps the most useful scheme is the position of a neuron in the fundamental circuit plan: a functional classification into sensory, motor, and intermediate neurons.

a. Sensory Neurons *Primary sensory neurons* are the initial neurons in sensory data processing. Derived from neural crest cells near the embryonic neural tube, their cell bodies lie outside the CNS: rod cells of the olfactory membrane, bipolar ganglion cells of the vestibulocochlear nerve (cranial nerve VIII), and pseudounipolar ganglion cells of spinal nerves and cranial nerves (CNs) V, VII, IX, and X. They are sentinals posted outside the CNS, reporting news (good and bad) from the periphery. Their number is small, perhaps 20 million.

Secondary sensory neurons are the second echelon in sensory data processing: cells in the CNS on which axons of first-order neurons terminate. Clusters of them lie in the dorsal horn of the spinal gray and brain stem nuclei (e.g., the cochlear nuclei). Whereas their afferent domain is weighted by sensory input, they are interneurons, between the first and last neurons in the plan of the nervous system. They receive input from sources beside primary sensory neurons (even from the cerebral cortex) and are multipolar, not bipolar or pseudounipolar, thereby meeting all criteria for interneurons.

b. Motor Neurons Multipolar in design, the cell bodies of *somatic motor neurons* are located entirely within the CNS. Whereas sensory neurons are the first cells in neural data processing, motor neurons are the last. They send impulses to the effector organs: muscles and glands. Like sensory neurons, their number is modest: about 2 million. Yet their role is profound. Derived from the mantle layer of the neural tube, they include the large α and small γ motor neurons in the spinal cord and brainstem that innervate *extrafusal* and *intrafusal* skeletal muscle fibers: those that do work and those that report muscle stretch, respectively. They also include the *visceral motor neurons* of the ANS, the dual arrangement of a preganglionic motor neuron in the CNS and a postganglionic one in an autonomic ganglion.

c. Interneurons Once called internuncial or "go-between" neurons, *interneurons* as originally defined are neurons interposed between sensory and motor neurons. The CNS is their domain and neuroanatomy is largely the study of them and their connections. The number of interneurons is extremely large: 10^{11} is the figure currently cited, but because large populations of minute cerebellar and hippocampal neurons may be greatly underestimated, some see it as closer to 10^{12}. They comprise almost all of our neurons. These countless separate, yet richly interconnected cells are more than responsive to stimuli, though they serve homeostasis with computer speed and reliability and in ways nineteenth century opponents of the neuron doctrine could hardly have comprehended. Interneurons are the source of endogenous neural activity, the cells that initiate new programs of behavior and, if need be, abandon old ones.

Interneuron is a useful term in invertebrate neuroscience. But it loses meaning for the staggering numbers of such cells in vertebrates, especially in the human CNS, where as noted they represent 99.9997% of the total. It is now used for neurons confined to particular regions of vertebrate nervous systems. A newer scheme of classification offers an alternative. It recognizes two major classes of neurons: *projection neurons* and *local-circuit neurons* (LCNs). Projection neurons are interneurons in the original sense, with axons running between regions of the CNS but also including sensory and motor neurons, with axons extending from the CNS to receptors and effector organs, respectively. LCNs are also interneurons, but they have axons restricted in their sphere of influence to other neurons nearby. Projection and local-circuit neurons are not classified by cell size, length of axon, or type of contact. It is only their service in long-distance or local communication that separates them.

d. Projection Neurons Most large neurons are of this class. They fit the classic mold of a neuron: multipolar, with an ample cell body, many dendrites emanating from it, and a long axon over which impulses pass to one or more distant regions, with the amino acid glutamate as the primary transmitter (usually eliciting excitation of the target cells). Pyramidal cells in the cerebral cortex and Purkinje cells in the cerebellar cortex (Fig. 27) are familiar examples. Motor neurons in the brain stem and spinal cord are others. They perform long-range signaling in the CNS (the corticospinal tract) or from CNS to muscles (the sciatic nerve). Central projection fibers (region to region), association fibers (area to area in a region), and commissural fibers (association fibers that cross the midline) arise from such cells. During neural development, projection neurons originate earlier than LCNs, as if sketching in the basic circuit plan.

e. Local-Circuit Neurons As noted, LCNs usually have a short axon or sometimes none (as in the amacrine cells of the olfactory bulb and retina). They are involved in local activity within a group of cells, not transactions between distant groups. They are far more numerous than projection neurons in the human CNS and arise later and longer in neural development. The dentate gyrus of the hippocampal formation and the cerebellar cortex have huge numbers of them. In the cerebral cortex, they are also numerous but comprise only 20–25% of the grand total; this region is noted for its long-distance, direct output (Table III). Some of these large subpopulations take a long time to produce and put in place. Indeed, millions of LCNs arise after birth and even in adulthood in some mammals, including humans.

LCNs are of great interest. With the growing appreciation that regions of the CNS are specialized for different functions, their roles in circuits peculiar to these regions are now high-priority issues. Their general structure and connections have been known for years, but until relatively recently their functions were elusive due to the difficulty of studying them. Some LCNs elicit inhibitory effects by the transmitter γ-aminobutyric acid (GABA), whereas others elicit excitatory effects by glutamate. Many LCNs are recognized in the cerebral cortex, differing in location, size and branching of processes, patterns of connectivity, places of termination on target neurons, and other features, especially the presence and corelease of other transmitters. A few LCNs, however, have been understood for years: the GABAergic basket cell in the cerebellar cortex exerts *surround inhibition* on rows of Purkinje cells, thus providing *enhanced contrast* between them and the excited Purkinje cells in nearby rows.

VII. THE FINE STRUCTURE OF THE NEURON

Despite their enhanced capacity to communicate, neurons are cells, showing familiar membranous, filamentous, and granular formed components. Yet the amount, arrangement, and function of these are distinctive, and a new component, the *synaptic vesicle*, unique to neurons is present.

A neuron has two major compartments: the *nucleus* and the surrounding *cytoplasm*. Within each, based on assumptions as to function, lie two categories of formed components: *organelles*, found in all cells and considered minute, metabolically active structures serving specific vital functions, and *inclusions*, varying from cell to cell and considered metabolically inert accumulations of cell products (pigments) or metabolites (lipids, carbohydrates). This categorization is still useful but was made when little was known about neurons' fine structure and specific function. Today, the list of organelles is increasing and that of inclusions is debatable or decreasing as highly organized internal structure and enzymatic activities of them are uncovered. Although the organization of the human nervous system is hypercomplex, simplifying principles are at hand.

Neurons, as noted, are cells of a superepithelium and, as such, are structurally and functionally polarized to play secretory and absorptive roles and to regulate traffic in solutes and ions between the outside and inside of the body. This polarity is seen in specializations for increased surface area, in the apical location of secretory products and parent Golgi complex, in basolateral adhesive and communicative junctions, and in basal membrane peculiarities serving substrate contact.

An "epithelial blueprint" of a neuron specifies its dendritic aspect as equivalent to the basolateral surface of an epithelial cell and its axonal aspect to the apical surface. This homology requires inclusion of the axon between the supranuclear Golgi complex on the one hand and the axon terminals, where secretory (synaptic) vesicles dock and discharge their contents into the synaptic cleft, on the other. An overview of neuronal fine structure follows.

A. The Cell Membrane

The external membrane or *plasmalemma* of a neuron (Fig. 23) has the general composition of all biological membranes: a bimolecular layer of opposing phospholipid molecules, with hydrophilic phosphorylated heads at the external and internal membrane surfaces and hydrophobic chains in the water-free interior. Cholesterol, proteins, glycoproteins, and glycolipids are interspersed in the phospholipid bilayer. The lipids form a barrier to prevent ions and water-soluble molecules from entering and to retain the cytoplasm. They also provide a fluid matrix in which protein molecules lie in a fluid mosaic manner. The proteins permit selective passage of certain materials through pores, as carrier enzymes, antigens, and receptor sites for bonding neuroactive substances. The apparent uniformity of the plasmalemma is misleading. Regional peculiarities in the distribution of membrane proteins and lipids are known. The apical, lateral,

basolateral, and basal domains of the columnar epithelial cell noted previously are good examples.

B. The Nucleus

Large and either spherical, or ovoid, the nucleus lies centrally located in the soma in most healthy neurons, (Figs. 20 and 21E-21I). The nucleoplasm in large, synthetically active neurons is euchromatic (lacking dense chromatin particles), amorphous, and pale. The chromosomes in these postmitotic cells no longer replicate DNA and double themselves but engage only in gene expression as partly uncoiled, indistinct, sprawling threads. Such a configuration makes the genome easily available for transcription and reflects the constant interplay between the nucleus and cytoplasm. In small neurons, the nuclear contents are concentrated and heterochromatic clumps are seen. In humans, the *sex chromatin* (condensed, genetically inactive X chromosome of the female) lies near the inner surface of the bilaminar *nuclear envelope*. In some mammals (mouse, rat, cat), it is satellited to the nucleolus (Fig. 23). In some neurons, the nucleus is shaped irregularly, with deep infoldings of the nuclear envelope.

1. Nuclear Envelope

The *nuclear envelope* displays numerous *nuclear pores* (Fig. 23). Seemingly closed by a diaphragm, each pore is a complex of eight radial spokes and a central granule that provides a 10-nm channel. The pores permit two-way traffic of molecules of less than 10 nm by passive diffusion. Because outgoing ribonucleoproteins and incoming newly synthesized nuclear proteins require a 20-nm opening, such molecules are thought to have a signal sequence of amino acids that binds to the nuclear envelope and somehow triggers opening of the channel.

2. Nucleolus

The *nucleolus* in neuronal nuclei is prominent (Figs. 21E–21I), and there may be more than one. As in most cells, it is a rounded structure, deeply stained by basic dyes, but unreactive to the Feulgen reaction for DNA. It transcribes ribosomal RNA. Newly synthesized ribosomal subunits, together with messenger RNA (mRNA) assembled in the nucleus, then pass to the cytoplasm through the nuclear pores, traveling along actin filaments within the nucleus and microtubules once outside it.

C. The Cell Body

The neuronal cytoplasm is crowded by filamentous, membranous, and granular organelles arranged concentrically about the nucleus.

1. Neurofibrils

Neurofibrils are best seen in large neurons, but are present in almost all (Fig. 21G). With metallic-impregnation, they are thin, interlacing, silver-loving threads (up to 2 μm in diameter) running through the cytoplasm and extending into dendrites and axon. With electron microscopy (EM), three kinds of filamentous structures are seen in neurons: *microfilaments* (diameter of 3–5 nm), *neurofilaments* (about 10 nm), and larger *microtubules* (20–28 nm).

Microfilaments resemble actin filaments in other cells, e.g., muscle fibers: they are polar polymers of globular (G) actin wound into two helically arranged strands. Like microtubules, they contribute to the cytoskeleton and provide rails along which organelles and proteins are driven by molecular motors, such as myosins. Neuronal myosins are also important to the production of neuronal growth cones, the retraction of axonal processes by a neuron and extension of others, and the generation of microdomains on the neuronal surface. Neurofilaments (Fig. 23) consist of three types of polypeptide threads twisted helically in a complex fashion to look like minute tubules in cross section. Microtubules (Fig. 23) are like those in most cells (e.g., mitotic spindle fibers), with differences and spatial distributions of microtubule-associated proteins (MAPs) that regulate their stability and facilitate their assembly.

The neurofibrils of classical cytology are probably bundles of neurofilaments, which have a high affinity for silver nitrate (microtubules may also be mixed in, but are not impregnated). They form roadways in the neuron, curving around and between other organelles and inclusions. Their functions are not fully understood, but as the *cytoskeleton* they support the organelles and change the shape of the cell as a whole. In the axon, neurofilaments run parallel to the fiber and may help to maintain the gelated state of axoplasm. If extruded, axoplasm is a firm gel, but it soon becomes a sol if exposed to calcium ions, triggering proteolytic neurofilament degeneration.

Microtubules, as noted, serve cell division by forming the mitotic spindle in neuronal precursors, but not in neurons, which are largely postmitotic. Microtubules play a vital role in the *intracellular transport* of vesicles and organelles that move along their surface through the cell body and along the axon (see later discussion). They turn over constantly. Their effects on cell shape are expressed not by contraction but by length changes due to polymerization or depolymerization or by self-assembly of new microtubules of different orientation. Neurofilaments are less dynamic. In degenerative diseases like Alzheimer's, their protein structure seems modified, forming characteristic lesions known as neurofibrillary tangles.

2. Nissl Substance

Nissl substance, named after Franz Nissl, a neurologist in Heidelberg in the late nineteenth and early twentieth centuries, this organelle is also termed chromophilic substance, as in other cells. It takes the form of clumps of cytoplasmic material more or less large and conspicuous, which stain deeply with basic aniline dyes (Fig. 21H). In living neurons, they are visible in phase contrast microscopy. The Nissl bodies represent the protein-synthetic machinery of the cell. Key constituents are ribosomes; they disappear in toluidine blue preparations after treatment with RNAase. The EM (Fig. 20) shows that Nissl substance is the granular or rough-surfaced endoplasmic reticulum (rER) found in other cells, especially those that synthesize protein in large amounts. Like the acinar cells of the pancreas, which contain a spiral nebula of rER about the nucleus, neurons show orderly, distinctive arrangements of it in parallel, broad cisternae stacked one atop another like empty pillowcases (Fig. 23). Anastomoses of cisternae and fenestrations of cisternal membranes are frequent. Nissl substance is a reticulum of channels within neurons. Its external membranes are studded with rows, loops, and spirals of ribosomes, although some regions are smooth.

Ribosomes also lie free in the cytoplasm as polysomal rosettes between cisternae. Rough ER is evident in the soma and proximal reaches of dendrites, wherein it mainly appears as branching and anastomosing tubules and short cisternae. It is not seen in the axon hillock, nor in the axon itself, which contains only agranular reticulum (Fig. 23). The soma is the major site of protein synthesis.

Neurons show a broad range of protein-synthetic functions, expressing far more genetic information than other cells. Like gland cells, neurons synthesize protein substances for export. But much of that produced is for maintaining and renewing its own protoplasm: of the axon and axon terminals, of the cytoskeleton, and of the watery cytosol in which all of the organelles lie suspended.

In the early stages of synthesis of neuronal proteins, mRNA molecules in the cytosol associate with free ribosomes and link them into clusters as *polysomes*. If the products are secretory, membrane, and vacuolar system proteins (ER, Golgi complex, etc.), the polysomes attach to the outer surface of the rER, continuing to assemble amino acids into polypeptides in a mRNA-determined sequence. As the amino acids are added, the polypeptide chains elongate from the larger ribosomal subunit into the lumen of the ER. When the chains attain full length, they are released from their ribosomes. Now free within the ER, these products are eventually exported as transport vesicles either to the Golgi complex or to the cell surface. All other neuronal proteins (cytoskeletal, cytosol, mitochondrial, peroxisomal, nuclear, and other vacuolar types) are synthesized on free ribosomes.

The visibility, size, form, and distribution of Nissl substance vary greatly in neurons. These variations are important in neuropathology. Nissl bodies appear as large chunks of basophilic material in pyramidal cells and motor neurons, cells with long axons and distant terminals to maintain. By contrast, they are dispersed in powdered form in sensory ganglion cells, which also have long axons. Normally, Nissl bodies are evenly distributed in the soma and proximal dendritic regions, but in a few healthy neurons they are concentrated near the plasmalemma. Such outward dispersion usually indicates a neuronal reaction to injury, anoxia, or disease and is aptly termed central chromatolysis.

Neuronal protein-synthetic machinery is eloquently expressive of cell activity (rest, work, fatigue) and sensitive in reflecting damage to the neuron of any sort. For example, when the axon of a neuron is severed (as in penetrating wounds), the Nissl substance disappears (retrograde chromatolysis), with its deeply basophilic material becoming diffused in the cytoplasm as it labors to repair that great loss.

3. Golgi Complex

The Golgi apparatus resembles a stack of inverted saucers, but it is a smooth-surfaced network of channels and spaces. Now called the *Golgi complex*, it is always present in neurons. Impregnated with silver even more deeply than the neurofibrils, it forms a rete of wavy strands encircling or apically surmounting the

nucleus (Fig. 21E). EM resolves these strands as closely apposed, flattened cisternae in stacks, as in other cells, with many small vesicles nearby (Fig. 23). The whole lies close to granular ER, from which it receives polypeptide secretory products in the form of transport vesicles. The cisternae frequently interconnect by tubules and also have pores in register between their curved aspects.

The Golgi complex is polarized with three compartments: an internal *cis* or *forming face* near the rER, where vesicles derived from it arrive and are initially sorted and incorporated; an *intermediate compartment*, comprising the bulk of the cisternae, through which delivered proteins pass stepwise, some by continuous channels but mainly by vesicular budding and fusion; and an external *trans* or *releasing face*, from which large vesicles, *secretory granules*, are liberated. A branched web of small tubules, the *trans-Golgi network* (TGN), borders the terminal cisternae. Here, small vesicles of various types and destinations bud off and depart, following the secretory pathway leading to the plasmalemma, wherein they fuse by exocytosis.

Many modifications to secretory proteins occur in transit through the Golgi complex and TGN. These include changes aimed at concentrating and packaging them, enhancing their effectiveness, binding to other macromolecules, resisting degradation, and addressing them to destinations. The last was thought to occur in the TGN, but other evidence suggests that this may occur earlier in the complex.

Reflecting its roles in refining and distributing secretory, plasmalemmal, and vacuolar proteins in the neuron, the Golgi complex is even more sensitive to changes in neuronal activity or to neuronal injury or disease than the rER. Under such circumstances, it migrates from its normal perinuclear position to the periphery of the cell body (retispersion), dissolves (retisolution), and then reconstitutes itself during repair after injury (e.g., severance of the neuron's axon).

4. Mitochondria and Peroxisomes

These organelles are considered symbiotic invaders of eukaryotic cells early in evolution. They stand apart from the vacuolar or tubulovesicular system of cells. Mitochondria self-replicate, have a short life span, and reproduce by division similar to the binary fission of bacteria. They have their own DNA and ribosomal, transfer, and messenger RNAs, but are not self-sufficient. They rely on their hosts to code and synthesize most of their proteins. Peroxisomes are less understood: mature forms may be recruited from a precursor reticulum from which they arise by budding.

Mitochondria are minute capsules of respiratory, phosphorylative, and other enzymes; they are the "power plants" of cells, which the unresting neuron, with a great need for oxygen and glucose and low energy reserves, requires in profusion. In oxidative phosphorylation, mitochondria produce large amounts of adenosine triphosphate (ATP), the chief molecule upon which the energy requirements of electrically active neurons depend. Neuronal mitochondria are rodlike or filamentous and usually more slender than those in other cells, ranging from 0.1 to 0.8 μm in diameter. They may associate closely with Nissl bodies, but generally are ubiquitous: in the dendrites and soma (Figs. 16 and 21F), along the axon, and in profusion at axon terminals (Fig. 23). A peculiarity of neuronal mitochondria is that their cristae often lie parallel (not transverse) to their long axis: cross sections of such forms appear as concentric light and dark rings. Another feature is the paucity of electron-dense granules in their matrix. Microcinematography of cultured neurons shows that mitochondria are constantly moving: along microtubules, within the soma, between it and the dendrites, and along axons. Mitochondria have intrinsic circular DNA, but most of their protein is synthesized in the nucleocytoplasmic system.

Peroxisomes are tiny (0.2–1.0 μm), spherical, membrane-limited bodies, smaller than lysosomes and aptly (if not helpfully) called microbodies when first described. They lack hydrolases but contain oxidases, which generate hydrogen peroxide, and catalase, which detoxifies it to water and oxygen, as well as over 40 other enzymes. They produce energy by oxidizing substrates, but unlike mitochondria they cannot store it as ATP for subsequent cellular use. With few exceptions in vertebrates, peroxisomes have a homogeneous structure. Clinical interest in them derives from several inherited diseases in which there is deficiency in one or more of their enzymes or defects in their assembly.

5. The Centrosome and Cilium

The centrosome, containing a pair of centrioles, is conspicuous in preneuronal proliferative cells during early neural development. It is seldom noted in light micrographs of adult neurons but sometimes seen in electron micrographs (Fig. 28, right), usually associated with a cilium. Because adult neurons, with a few

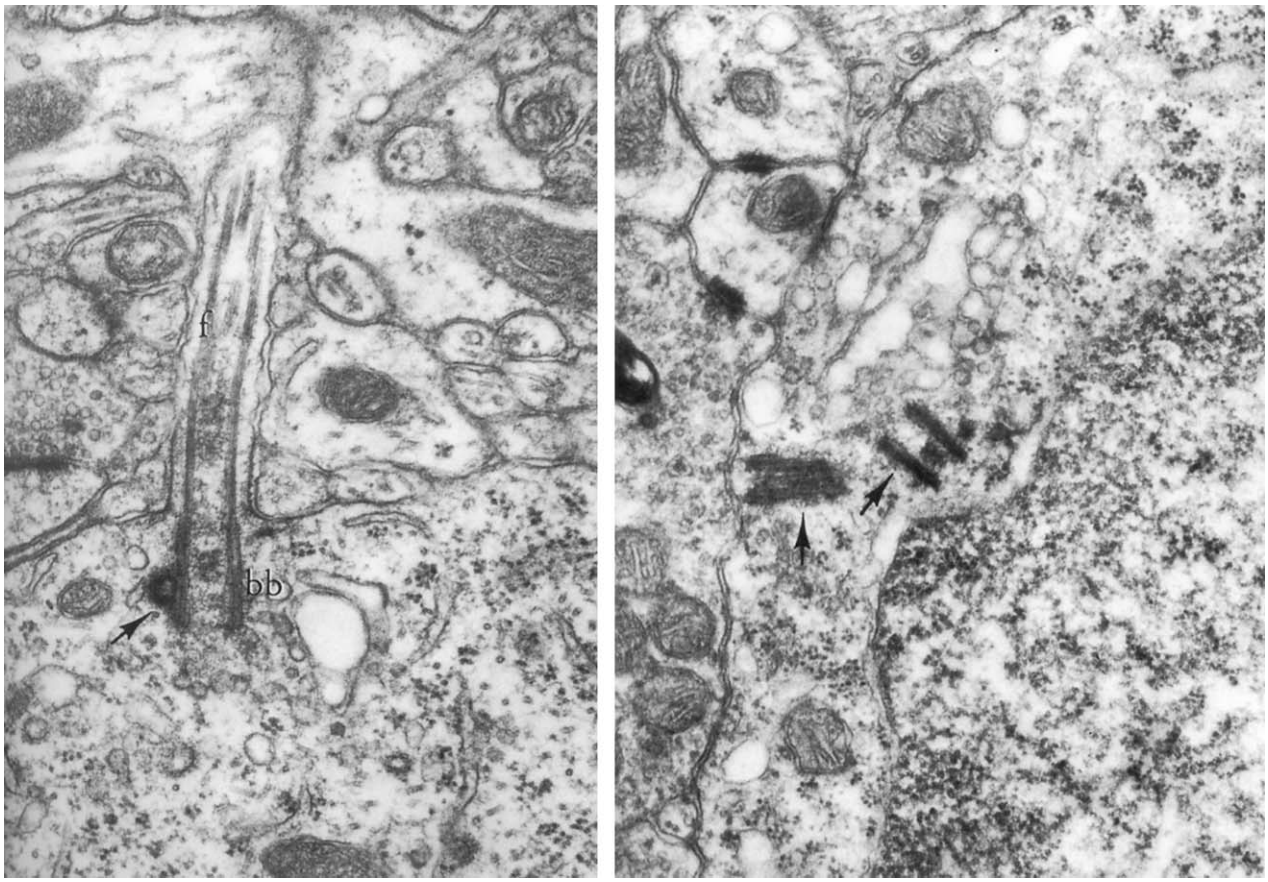


Figure 28 Cilium and centrosome: left EM shows a longitudinal section of a cilium arising from a neuron. At its base is a basal body (bb) with attached parabasal body. The microtubules in the ciliary shaft have a 9 + O configuration and continue into the basal body. Right EM shows part of the perikaryon of a cerebellar granule cell containing two longitudinally sectioned centrioles (arrows). The wall of each is formed by nine triplets of short tubules oriented parallel to its long axis. From *The Fine Structure of the Nervous System: Neurons and Their Supporting Cells*, 3rd ed., by Alan Peters, Sanford L. Palay, and Henry de F. Webster, copyright 1990 by Alan Peters. Used by permission of Oxford Unviersity Press, Inc. (EMs from collections of authors).

exceptions that may reflect the residual presence of antecedent cell types, do not divide, the role of the centrosome is puzzling: possibly it is a nondisposable relic of the mitotic apparatus of parental cells. But the recognition of the centrosome in neurons and in other cells as the *microtubule-organizing center* gives a new perspective. A cloud of proteins surrounds the centrioles. Microtubules attach to it at their minus ends and grow from it by addition to their plus ends.

The cilium (Fig. 28, left) has nine radial pairs of microtubules, distally dwindling to eight, with short connections but no arms or spokes between them as in other cilia. Distally a central pair appears. In kinocilia it is thought to switch the other pairs from power stroke to recovery stroke. Here it seems nonmotile.

6. Neuronal Inclusions

Golden brown *lipofuscin* (lipochrome or fatty pigment) granules are found in neurons of the CNS. They are harmless byproducts of lysosomal activity, a "wear-and-tear" pigment increasing with aging. Lipofuscin may eventually displace the nucleus and organelles severely to one side of the cell body. Black *melanin* pigment is seen in certain clusters of specialized neurons. The substantia nigra of the midbrain and the locus ceruleus of the pons are the prime examples. The presence of lipofuscin seems related to the synthesis of certain catecholamines, notably dopamine. *Glycogen* is found in embryonic cells of the CNS but not adult neurons. The consequences of a lack of

energy stores in neurons are obvious. *Iron* particles are seen in some neurons of older persons. In some instances, they may represent the uptake of blood pigments in hemorrhagic areas.

D. Dendrites

These extensions of the soma contain Nissl substance, mitochondria, parallel microtubules, and a less conspicuous population of neurofilaments (Fig. 23). Components of the Golgi complex may also be seen in first-order dendrites of some neurons, especially in those with a small soma. In a given neuron they may appear in some primary dendrites and not others (some speculate that the Golgi complex slips in and out in response to unknown commands). Most of the other organelles dwindle out toward the finer branches of the dendritic tree, but the number of mitochondria remains constant or even increases, reflecting the high degree of activity in these receptive cell processes.

E. The Axon

At its origin in the axon hillock, the arrangement of structures changes between that of the soma and the axon proper (Fig. 23). The microtubules funnel into bundles, the number of ribosomes wanes, and the rER almost vanishes, in keeping with the lack of Nissl substance seen in the light microscope (Fig. 24). Beyond the initial segment, the axon contains longitudinally oriented smooth agranular reticulum, long slender mitochondria, numerous vesicles, modest numbers of microtubules (not profuse as in dendrites), many neurofilaments (but fewer than in dendrites), and microfilaments.

1. Axonal Transport

Impulse conduction is not the sole function of the axon. Macromolecules essential to structure flow from the soma down the axon. Other substances, e.g., tired axon terminal membrane requiring degradation and recycling, move up. Rates of *axonal transport* of these commodities vary strikingly; some, including newly synthesized protein for neuronal repair, move slowly, at 0.2–5 mm/day, whereas others related to transmitter synthesis travel 100–400 mm/day. Two modes of axonal transport are known: *fast*, *bidirectional* (anterograde and retrograde) and *slow*, *unidirectional* (anterograde only).

Mechanisms of axonal transport have been studied intensively. Microtubules play a critical role in fast transport. Axonal microtubules are polarized: their "plus" ends face the axon terminal and their "minus" ends the cell body. Microtubules constantly turn over; tubulin dimers are added at the plus ends and depolymerized at the minus ends. Neuronal products are shipped down the axon as small vesicles. Kinesin, a microtubule-associated protein, is the motor responsible for this transport. One end of a kinesin molecule attaches to a vesicle while a binding site on the other end interacts in a cyclical manner with the wall of a microtubule, resulting in movement of the vesicle toward the plus end at about 3 μm/s.

In *fast anterograde* axonal transport, outbound materials are membrane-bound organelles from the rER and Golgi complex, synaptic vesicle precursor membranes, large dense-core vesicles, parts of the smooth ER, and mitochondria. This transport travels along microtubules, stationary raillines and fast tracks along which shipments are moved by molecular motors (kinesin and related proteins) at high speed (100–400 mm/day). It meets the critical, unceasing needs of neurotransmission.

In *fast retrograde* axonal transport, the incoming materials are endosomes (micropinocytotic vesicles) from nerve terminals, mitochondria, and ER elements returning to the soma from axon terminals. This transport travels on microtubules as in the preceding paragraph, but at half to two-thirds the speed. The motor is a microtubule-associated ATPase similar to dynein in cilia and flagella. It provides for degradation and further recycling in cell body lysosomes of worn-out membrane components, which may have already been recycled a few times in the terminals. It also implies processing and repackaging events there, much as one must evaluate and rewrap damaged goods prior to returning to the sender.

Slow *anterograde-only* axonal transport is for cytosol proteins and cytoskeletal elements. It has two rates: slow and slower. *Slow* (rate b: 2.5–5 mm/day) conveys hundreds of polypeptides from cytoskeletal proteins (like actin) to soluble metabolic (e.g., glycolytic) enzymes. *Slower* (rate a: 0.2–2.5 mm/day) is mainly (75% of the total moved) neurofibrillary protein: neurofilament and microtubule subunits. Motors are uncertain; dynein may be one. Neurofilaments seem unable to move on their own but may hitchhike on microtubules, which can extend themselves (see previous discussion). Slow axonal transport maintains the structural and functional integrity of remote regions of the neuron: perhaps 10,000 times the diameter of the cell body away.

The mechanisms of axonal transport have been explained to a degree that is a triumph for molecular

neurobiology. Study of the perpetual movement of materials along the axon clarifies other aspects of neuronal activity, especially the regeneration of PNS axons, which takes place at about 1–3 mm/day, and the daily shipment and delivery of neuroactive substances and their precursors.

Axonal transport is vital to the life of the neuron and its communicative and neurotrophic functions. It permits rapid replacement of catabolized protein in the axon and its terminals. Most constituents of such a protein cannot be synthesized in the axoplasm devoid of rER. It permits the transport of enzymes synthesized in the soma down to the axon terminals; such enzymes are necessary for transmitter synthesis there. It permits the movement of macromolecules and other dynamic cellular components within the soma itself and out into the dendrites, which have modest protein-synthetic capability.

In retrospect, "axonal transport" exemplifies tunnel vision. "Intraneuronal transport" might have been a better way to designate mechanisms of moving materials within neurons had we known then what we know now. It permits feedback from the periphery (axon terminals and dendrites) and regulation of the metabolic activity of the parent cell body: adjusting rates of protein and enzyme synthesis or reprocessing worn-out membrane components in accord with peripheral use and demand.

Experimental benefits and some risks accrue from axonal transport. A powerful neuroanatomical method for tracing axons back to their cells of origin rests upon the uptake by axon terminals of tiny enzyme particles (the glycoprotein horseradish peroxidase) and retrograde axonal transport to the somata. There, these markers can be demonstrated histochemically as a colored, insoluble polymer. Unfortunately, retrograde transport also may convey pathogens (tetanus bacterial toxin, herpes simplex, and rabies viruses) directly to nerve cell bodies, once the microorganisms are endocytosed at nerve endings. A technique exploiting anterograde axonal transport is the autoradiographic method. Radioactively labeled amino acid is injected near a nerve cell body, taken up by the cell, incorporated into newly synthesized protein, and then transported down the axon to its terminals. There, the radioactivity can be detected by autoradiography (a process similar to photography).

2. The Axonal Myelin Sheath

Axons in CNS tracts and peripheral nerves are, to varying degrees, ensheathed by sleeves of spirally wound cell membrane derived from nearby supportive cells and compacted into a substance known as *myelin* (Fig. 18). Chemically, myelin consists of a 4:1 ratio of various lipids to proteins. In the unfixed state it is glistening white. It acts as an insulator, increasing nerve conduction velocity from perhaps less than 1 m/sec in slender, unmyelinated axons to 120 m/sec in large, well-myelinated fibers. Peripheral and central *myelin sheaths* are similar but made by different cells and in different ways. In the PNS and ANS (but not the CNS, which has no intrinsic connective tissue), the myelin sheath is additionally invested by protective, supportive, and nutritive connective tissue layers.

a. Peripheral Myelin Just beyond their exit from the CNS to near their distal endings, peripheral axons are enveloped by serially arranged, flattened neurilemmal or *Schwann cells* (Fig. 15, S_1). Newly regenerated or thin axons lie in shallow indentations of the Schwann cell surface (S_2). Such fibers are "unmyelinated." The term is misleading: they are not uncovered, but minimally myelinated. The plasmalemma of the slender bay of cytoplasm enclosing them is the same membrane, as discovered by pathologist Betty Ben Geren in 1954, that when rolled into a tight spiral of many turns constitutes the myelin sheath. Such inward spiraling of Schwann cell membrane may involve only a few turns to up to 50 or more about the axon. As the spiral is compacted, cytoplasm within it is displaced outward in the spiral toward the cell body (as toothpaste is displaced upward when the tube is rolled from the bottom) and longitudinally to the edges of the sheet, where it appears as a bead of material in paranodal loops.

The larger the axon, the more myelin layers about it. In electron micrographs, myelin has alternating dark and light lines with 12-nm periodicity. The cytoplasmic surfaces of the Schwann cell appose to form the *major dense line*, and the external surfaces form a less dense *intraperiod line* with 2-nm *intraperiod gap* separation, a cleft allowing the passage of tracer molecules. Tight junctions block the passage of larger molecules and impart strong adhesion to facing cell surfaces (like Velcro). The myelin spiral starts with a tongue of Schwann cell about the axon and continues a few more turns, when compaction begins. Yet 46 years after its discovery, how the myelin spiral forms remains an unsolved mystery.

Cordons of Schwann cells along an axon segmentally ensheath it to prevent current from leaking out. *Nodes of Ranvier*, periodic gaps of myelin (250–1000 μm apart in human nerves) between Schwann

cells, expose the axonal membrane, which has a high concentration of voltage-gated Na^+ channels. Action potentials spread electrotonically and virtually instantaneously along the insulated *internodal segments*. At the nodes (Fig. 21J), they are regenerated rapidly, but with sufficient delay to suggest that they "leap" from node to node.

b. Central Myelin Central myelin is formed similarly to peripheral myelin, resembles it in almost every way, and serves equally to speed nerve impulse conduction. In its abundance along the axons of central tracts, the darkly stained myelin sheaths seen in whole brain sections strikingly demonstrate the long-distance routes of communication in the human nervous system. Key differences are described next.

Central myelin is made by *oligodendrocytes*. By producing (by unknown means) many helical turns of plasmalemma at the tips of its processes, an oligodendrocyte makes 15–40 internodal segments on several to many axons (Fig. 21D). But whereas a Schwann cell may harbor many unmyelinated or regenerating axons around its circumference, it constructs only one internode on one axon.

Other differences in central myelin are a paucity of associated cytoplasm, periodic thickenings of the axolemma at points of contact with paranodal myelin loops, a longitudinal ridge of cytoplasm outside the spiral (instead of an enveloping ring), and the absence of basal lamina around the oligodendrocyte (a cell proper to the neuroepithelium, not needing delimitation). Connective tissue investments do not surround the myelin sheath as in peripheral nerves.

As cells that myelinate CNS axons, oligodendrocytes line up along axons in tracts or diffuse white matter as *interfascicular glia*. In human fetal and postnatal development, tracts are myelinated and thus become fully functional at different times. Motor nerve roots are well-myelinated by birth, but optic nerves and sensory roots lag behind until the 3rd and 4th months after birth, respectively. The corticospinal tracts require 1 year to achieve full myelination, whereas cerebral commissural fibers requuire up to 10 years. This staggered ensheathment has important functional correlates for the developing individual: when she or he can move, see, begin to walk, perform directed movements, and so on. One intracortical plexus is not completely myelinated until middle age: the stripe of Kaes–Bechterew in layer IIIa (Fig. 8, right panel). Neuropathologist Paul Yakovlev called it "the wisdom stripe."

c. Functions of Myelin The preeminent role of this phospholipid sheath is to greatly increase the speed of nerve impulse conduction in select nerve fibers or central pathways. It acts as a high-resistance, low-capacitance insulator, confining the ionic current set up within the axon by the action potential generated at a node of Ranvier and forcing it on to the next node, where membrane depolarization regenerates the potential. But in some neural pathways, such as in the ANS and certain central pain tracts, this high velocity may be neither necessary nor desirable. Reliable delivery and persistent arrival of impulses may be more important. Thin myelin sheaths and shared sheaths of axons by oligodendroglia and Schwann cells, respectively, meet these requirements and save precious space as well.

A putative, unsubstantiated nutritive role for myelin rests on observations that tracers pass from the outer collar of Schwann cell cytoplasm inward along tunnels of residual cytoplasm in the myelin spiral (termed "incisures" from their apparent interruption of the myelin sheath as seen in light microscopic preparations; Fig. 21J) and that these molecules later show up in the axoplasm. It is unknown whether metabolites follow this path. There are also speculations about metabolic interactions between the myelin sheath and the axon and how these might change with neuronal activity.

Other functions may exist. Myelin is not confined to axons. Somata and dendrites may be loosely wrapped in it. Myelin may be protective, assuring continued connectivity. In demyelinating diseases like multiple sclerosis, demyelinated axons conduct impulses, but less rapidly and efficiently. During remission, the improvement in conduction may be greater than the remyelination that has occurred.

VIII. SYNAPTIC ORGANIZATION OF THE NERVOUS SYSTEM

This subject lies at the heart of the organization of the human or any other nervous system. A synapse, by definition, includes presynaptic and postsynaptic membranes separated by a narrow cleft about 12–20 nm across. As noted, EM demonstration of this cleft provided definitive proof of the anatomical tenet of the neuron doctrine: nerve cells are separate. The cells that face each other across this cleft may be two neurons, a neuron and a skeletal or smooth muscle fiber, or a neuron and a gland cell (usually a smooth muscle fiber in that gland).

A. Significance of the Synapse and Dual Modes of Synapses

The synapse is a region of specialized contact between neurons or between a neuron and a muscle cell for cell-to-cell communication. The estimated number of synapses in the human nervous system is 10^{14} (100 trillion), and in light of newly appreciated synaptic diversity it may be even higher. Two modes of synapses have been identified: electrical and chemical. The first has been touched on in the discussion of gap junctions and the neuron doctrine. The second is by far the more common.

1. Electrical Synapses

Electrical synapses are gap junctions. When present between neurons, they are very different from chemical synapses where the separateness of the cells is not in question. They allow the direct spread of current from one cell to another, without delay or need for receptor and decoding systems. But the individuality of the coupled cells is partly lost, and hence their utility is diminished for large nervous systems with labeled lines like those of mammals. Electrical synapses are common in invertebrate and nonmammalian nervous systems but infrequent in mammals except between neuroglial cells, where they offer the chief mode of communication. Yet they have been found between mammalian neurons and shown to transmit in a few cases. In the embryonic CNS, they are seen in many places, even in the cerebral cortex, but decline in number as chemical synapses develop. In the adult, they are usually found in cell clusters that fire action potentials synchronously, as in the lateral vestibular nucleus, which effects a rapid increase in ipsilateral extensor tone for postural maintenance, or clusters that spread influences widely, like the horizontal cells of the retina. Studies show that electrical synapses can be modulated, that they may have mechanisms favoring unidirectional conduction, and that electrical and chemical synapses have important reciprocal influences.

2. Chemical Synapses

Chemical synapses are the standard mode of neuronal communication in mammals. Many studies have been made on their structural, functional, and neuropharmacological properties. They mediate the transfer of neuronal activity by releasing a neuroactive substance of some sort. Along with the law (usually enforced) of dynamic polarization of neurons, they help to establish the one-way traffic of the nervous system. Synapses act as turnstiles that permit activity to pass in one direction only (two-way synapses exist and are now well-known, but these are a pair of oppositely polarized synapses juxtaposed). This unidirectional feature of the chemical synapse is of cardinal importance: the other parts of neural circuits (dendrites, cell bodies, and axons) can and do, in some instances, conduct both ways. Even in the chemical synapse, important effects are exerted by the postsynaptic neuron upon its presynaptic partner (see discussion to follow). Still, the polarity of the chemical synapse largely determines the forward traffic flow seen in most routes and regions of the CNS.

B. Varieties of Synapses and General Principles of Synaptic Activity

1. Synaptic Mechanisms and Second Messenger Systems

Molecular neurobiology has bared submicroscopic details of the organization of the axon terminal and clarified the complex mechanisms of synaptic transmission. Two kinds of transmission exist: rapid and slow. *Rapid transmission* depends on transmitter-gated ion channels. *Slow transmission* relies on postsynaptic receptors (membrane-spanning proteins) linked to G (guanine-nucleotide-binding) proteins. Most postsynaptic responses are depolarizing or hyperpolarizing changes in membrane electrical potential (excitatory and inhibitory postsynaptic potentials or EPSPs and IPSPs), but some are electrically silent metabolic or membrane alterations.

In rapid synaptic transmission, the synthesis of neurotransmitter, packaging and transport in vesicles, clustering, cytoskeletal anchoring, subsequent dissociation, docking at the presynaptic membrane, fusion, release of contents upon arrival of the action potential, membrane re-uptake, and recycling are stages, fleeting or time-consuming, in a continuous and widespread process inside a nerve cell. We take a closer look at rapid transmission in a subsequent section.

In slow transmission, EPSPs and IPSPs are also produced by alterations in current flow through membrane ion channels, but a staged process occurs in which transmitter binding to a G-protein-coupled receptor leads to dissociation of a subunit of it, which then moves laterally in the membrane and exerts effects ranging from ion-channel binding and

conductance alteration to enzyme activation altering concentrations of second messengers (cyclic AMP, inositol triphosphate, arachidonic acid metabolites) to alteration of gene expression.

2. Functional Specialization of Synapses

The structural diversity of presynaptic components is great: one finds club endings, large end bulbs, small end bulbs, spiral endings that wrap around the axon hillock and initial segment of the target neuron, fibers of passage merely brushing the dendrites as they pass by, claw-shaped endings that seem to seize the cell body of a neuron or grasp a dendritic spine, extended zones of apposition of an axon and a dendrite, as in the case of the cerebellar climbing fiber, where numerous synapses are made, and other types. All of these structural variations reflect functional specialization. A synapse *en passant* allows an axon (like that of a cerebellar granule cell) to distribute its influences to multiple target cells along its course, whereas the obsessive climbing fiber ending on a Purkinje cell provides the focused protracted connectivity underlying the most powerful excitatory synapse in the human brain.

3. Spatial Distribution of Synapses

The various inputs to a neuron are distributed in a precise way on the postsynaptic cell. Virtually all neurons in the human CNS show the same location specificity: pyramidal cells, Purkinje cells, mitral cells, and motor neurons. The giant pyramidal cells in the hippocampus have perhaps the most striking *lamination of afferents*, as this input distribution is called. David Bodian, a pioneer in the study of synapses, had another term for it: *mosaic segregation*. The location of the end bulb or terminal bouton indicates the source of input, efficacy of action, and local biochemical diversity of the postsynaptic membrane.

C. Fine Structure, Elements, and Stabilization of Chemical Synapses

1. Fine Structure of the Chemical Synapse

As in the neuromuscular junction, the axon terminal typically contains a few neurofilaments, many mitochondria, and an organelle unique to neurons: the synaptic vesicle (Fig. 29). The vesicles release their contents into the synaptic cleft upon the arrival of nerve impulses in the axon terminal. The substance liberated is then recovered and recycled after exerting effects on receptor sites in the postsynaptic neuron.

2. Elements of Chemical Synapses

When the two synaptic elements are both parts of neurons (dendrites, somata, axons), several combinations are possible, with terms to match. Thus, axodendritic, axosomatic, and axoaxonic synapses are recognized. These three arrangements are the "conventional" ones dominating the field until the late 1950s.

"Unconventional" synapses (Fig. 30) were then uncovered: dendrodendritic synapses are the best known, but dendroaxonic, somatosomatic, somatoaxonic, and axoaxodendritic (serial junctions) are described. Yet the most arresting ones are *glomeruli*, which are clusters of axons and dendrites (Fig. 17). These neuronal microcosms illustrate Charles Sherrington's key principles of divergence and convergence.

3. Adhesion of Neurons and Stabilization of Synapse

An unsung type of intercellular junction in the CNS is the *punctum adherens*: a small region of close contact and adhesion between adjacent cells. It corresponds to the zonula adherens on the lateral surfaces of cells in various epithelia (e.g., the simple columnar epithelium of the small intestine). In epithelia, the role of zonulae adherentes is obvious. They bind cells together and, with nearby occluding junctions, prevent the extracellular passage of materials that otherwise would leak past them. In the CNS, puncta adherentia hold nerve cells together and stabilize their points of synaptic contact, just as soldered circuits provide such stability. Without them and with sudden movements of the head, our central circuits might come unglued, neuroglia or not.

D. Mechanism of Rapid Synaptic Transmission

Arrival of an action potential at an axon terminal causes voltage-gated ion channels to open, thereby allowing Ca^{2+} ions to enter the terminal. Within 100 μsec, their presence triggers quantal release of neurotransmitter into the synaptic cleft by exocytosis

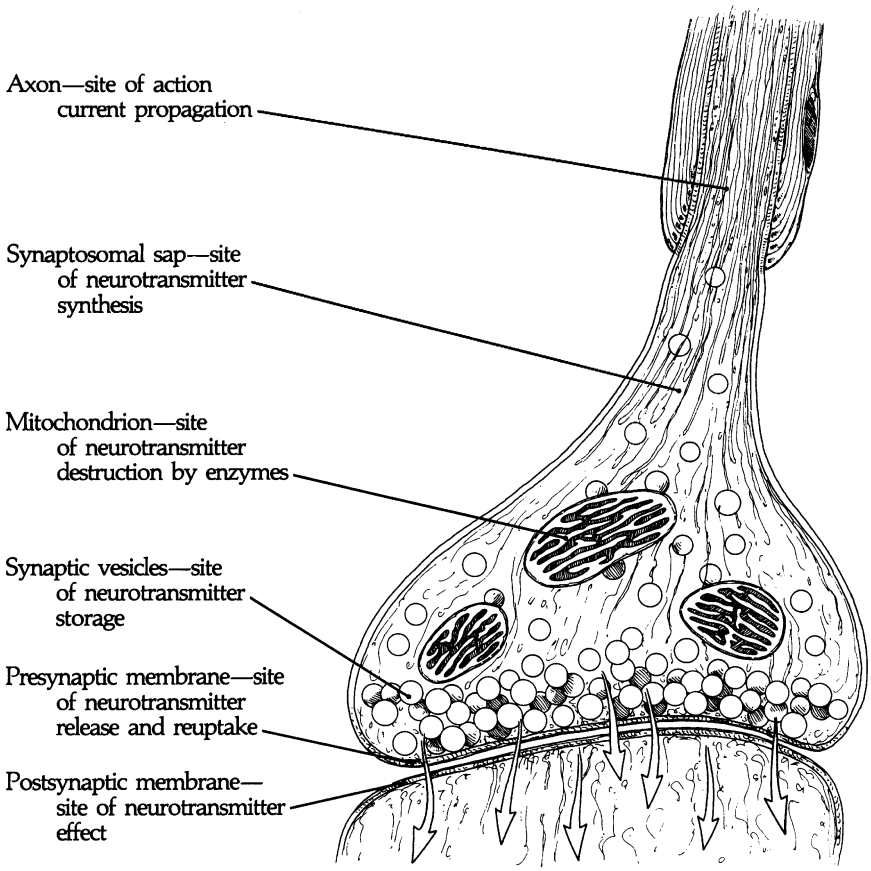


Figure 29 Chemical synapse: an axon terminal typically shows a few neurofilaments, many mitochondria, and numerous synaptic vesicles. These spherical organelles, unique to neurons, are found attached to the actin cytoskeleton in relatively large numbers or in smaller numbers released to positions of readiness or docked for immediate use at the presynaptic membrane. They release their contents into the synaptic cleft (12–20 nm wide) by coalescing with the plasma membrane upon the arrival of nerve impulses in the axon terminal. Following release, transmitter diffuses across the cleft and binds to specific receptors in the postsynaptic membrane. These receptors undergo immediate conformational change that leads to the opening of channels so as to alter the permeability of the membrane to certain ions, thus changing the membrane potential. After release, presynaptic membrane is taken up by endocytosis and synaptic vesicles are reloaded with locally synthesized transmitter. From J. B. Angevine, Jr., Dendrites, axons, and synapses, *BNI Quarterly*, Vol. 4, No. 2, 1988 (illustration by Steven J. Harrison).

from synaptic vesicles docked in the active zone of the synapse. This exocytosis is far more rapid than anywhere else. It is virtually instantaneous, probably because a subset of vesicles is already docked at the *active zone* (where the presynaptic membrane appears thickened) and evanescent fusion pores form between the vesicles and the presynaptic membrane. The transmitter diffuses directly across the synaptic cleft and binds to specific receptors in the postsynaptic membrane. The receptors undergo conformational changes that open ion channels, resulting in depolarization and excitation or hyperpolarization and inhibition of the target cell. By this time, all of 200 μsec has elapsed.

Within the axon terminal are two pools of synaptic vesicles: a relatively small releasable pool from which vesicles fuse with the presynaptic membrane when the action potential arrives and a larger reserve pool in which vesicles are bound to the actin cytoskeleton and mobilized as required. After transmitter release, membrane is taken up by endocytosis and vesicles are reloaded with transmitter synthesized locally from precursors transported to the terminal by kinesin-mediated axonal transport. As for the transmitter just released, its molecules must be removed quickly after receptor binding to prepare the postsynaptic membrane for new releases. Removal may be performed in one or more ways, acting in concert or in combinations

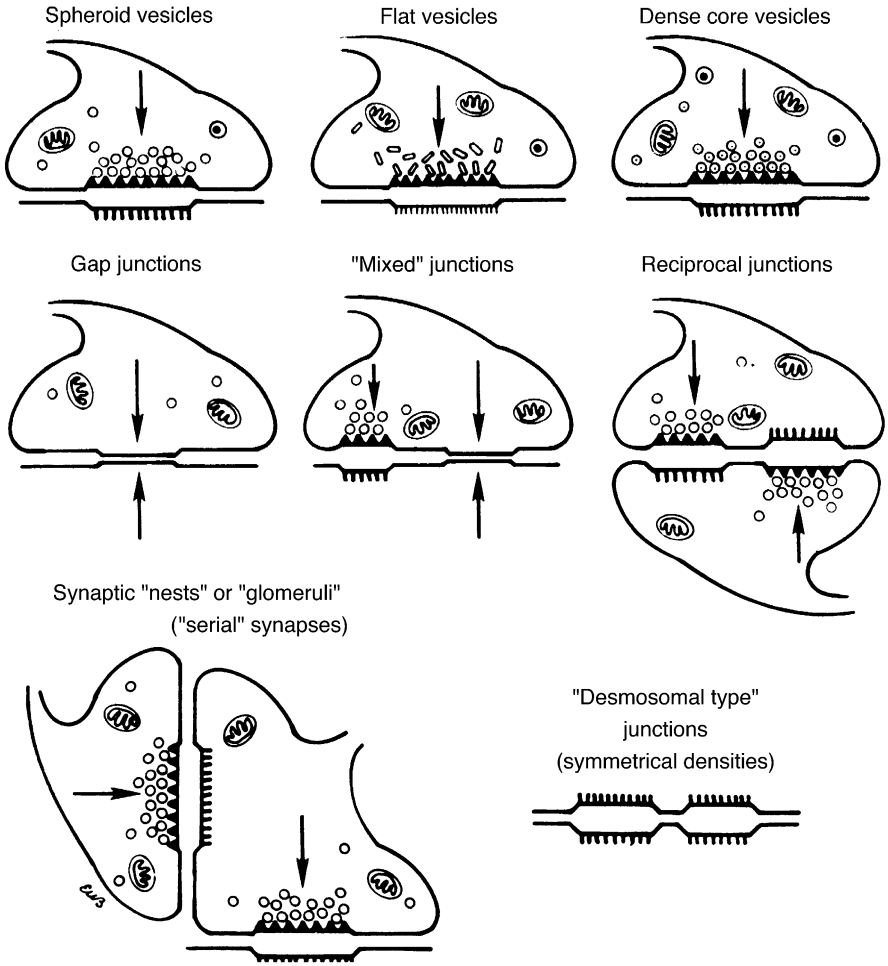


Figure 30 Synaptic variety: until the late 1950s, chemical synapses of axodendritic, axosomatic, and (later) axoaxonic types dominated the field of neurobiology. These familiar synapses were considered "conventional." But by 1972, many "unconventional" junctions had been recognized. That year, David Bodian reviewed these "unconventional" coupling paradigms. He included the intriguing glomeruli: many of these are problematic, but some exemplify, in a reductionist manner, Sherrington's key principles of divergence and convergence in the organization of brain circuitry. The desmosomal type junction shown is a punctum adherens (plural: puncta adherentia), a small region of contact and adhesion between neurons corresponding to the zonula adherens found on the lateral surfaces of cells in various nonneural epithelia (e.g., simple columnar epithelium of the small intestine). See also text. From Neuron junctions: A revolutionary decade, D. Bodian, *The Anatomical Record*, copyright 1972. Reprinted by permission of Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc. (illustration by Elinor W. Bodian).

peculiar to a given synapse: enzymatic inactivation, reuptake by the axon terminal, uptake by a nearby astrocytic process, uptake by the postsynaptic region of the target neuron, and simple diffusion away from the synaptic cleft into the extracellular space, which although narrow in the CNS (2–4 nm) would allow some movement of transmitter molecules away from the cleft. It is too slow, however, to be an effective mechanism by itself.

E. Overview of Synaptic Transmission

Synapses vary in their mode of action (electrical or chemical), form and ultrastructure, number and specified laminar organization on postsynaptic neurons, types and rapidly increasing numbers of chemical messengers (now including gases and growth hormones), functional effects, and affiliations with the neuroglia.

We have noted puncta adherentia: sticky spots that hold neurons together and make fast their synapses. They are widely distributed in the nervous system, prominent between astrocytes, and evident *everywhere*: between dendrites, cell bodies, dendrites, and axons, dendrites and cell bodies, axon terminals and axon initial segments (at axoaxonic synapses), axon membranes and their sheaths, and neuronal processes and astrocytes. The larger *zonula adherens* is seen between the lateral surfaces of ependymal cells, cells of the choroid plexus, and the subventricular tanycytes. Although these surface specializations play purely adhesive roles, they are often seen close beside synapses, chemical or electrical, and at nonspecific sites. Some see the *active zone* of a synapse, the site of transmitter release, as a modified zonula or punctum adherens. Unlike these symmetrical adhesive junctions, a chemical synapse has become structurally and functionally *asymmetrical* and, hence, polarized. If this is true, the site of transmission and permanence of place of neuronal communication evolved from simple adhesive junctions found universally in epithelial tissues.

Three key points on neuroactive substances: First, *multiple* transmitters and the *corelease* of transmitters seem now to be the rule for neurons. Second, a given transmitter may elicit *different* functional effects in different situations. Third, the effect of a neuroactive substance on its target cell depends *exclusively* on the postsynaptic receptors to which it binds.

IX. THE NEUROGLIA: BACKUP ELEMENTS OF THE NERVOUS SYSTEM

The peculiar branched, diversified cells of the neuroglia outnumber the 100 billion neurons 10–50 times, making up half the volume of the CNS. Though largely held together by means common to epithelia (puncta adherentia), neurons lie in a sea of glia. These chiefly (maybe entirely) ectodermal derivatives are not just supportive cells, though they offer mechanical support during neurogenesis and in adulthood. They are metabolically active cells that assist neurons in the performance of their communicative and integrative functions.

A. Categories of the Neuroglia

Five major categories of the neuroglia are generally recognized (Fig. 15).

1. Astrocytes

Although two major types exist, astrocytes show extreme variations in size, shape, and relationships with other structures. The Bergmann astrocyte or candelabra cell of Golgi is extreme in all three respects. *Protoplasmic astrocytes*, found mainly in gray matter, have a stellate cell body with many branching processes (Fig. 15, PA). Some of these terminate on blood vessels as end feet. Sometimes the soma itself lies against the vessel. Astrocytes near the surface of the neuraxis send similar processes to the pia mater, beneath which their end feet are apposed, fastened by puncta adherentia and displaying gap junctions. Taken together with the overlying pia, this barrier is the *pia–glial membrane*.

Protoplasmic astrocytes, as their name implies, have abundant cytoplasm with a generous number of organelles and a large, pale nucleus that helps to distinguish them from other glial cells and nerve cells in dye-stained preparations. Smaller examples lie close to neuronal cell bodies as satellite cells. *Velate astrocytes* (Fig. 21C) extend thin veils between neurons and their processes and in the case of the Bergmann astrocytes, partition input to neurons.

Fibrous *astrocytes* are found mainly in white matter, but some regions of the periventricular gray matter also show them (Fig. 15, FA). They have the familiar euchromatic nucleus and a soma with relatively few organelles: long, thin, smooth processes that branch infrequently and extend to blood vessels and the pia with end feet like the protoplasmic forms. As noted, CNS blood vessels and the subpial border of the neuraxis are surrounded by basal lamina. Protoplasmic and fibrous astrocytes may be a single cell type, varying in form in different locales, or two distinct subtypes; more study is necessary. The former expand into whatever space is available, and this may account for their irregular outline.

The shapes of astrocytes are shown, not in dye-stained material but by special metallic impregnation techniques, along with cytoplasmic fibrils, which are numerous in the fibrous type, less so in the protoplasmic type, and highly variable in number within the cell and regionally. In the EM, glial fibrils are resolved as bundles of slender intermediate filaments. Unlike neurofibrils, they are thinner (8 nm) and packed more closely. Chemically they are distinct, consisting of *glial fibrillar acidic protein* of 51,000 Da molecular weight. GFAP is specific to astrocytes, and antibodies to GFAP selectively stain and identify these cells in tissue preparations and cell cultures. The high content

of astrocytic fibrils accords with their supportive functions.

2. Oligodendrocytes

Oligodendrocytes resemble astrocytes and may be related to them, but they have fewer processes, which branch infrequently (Figs. 15, Ol and 21D). The soma is small and the nucleus distinctive. It is much smaller than the large, pale nucleus in astrocytes: round, heterochromatic, and deeply staining. The dense cytoplasm also stains darkly because as it is rich in rER and free ribosomes, with a conspicuous Golgi complex and many mitochondria. This dusky quality is perhaps the most characteristic feature, as striking in the EM as in dye-stained sections. Oligodendrocytes are usually the darkest cells around.

Interfascicular *oligodendrocytes* make and maintain central myelin (see preceding discussion). In white matter, they lie in rows between bundles of axons (Fig. 15, Ol). In gray matter, *satellite oligodendrocytes* closely associate with nerve cell bodies. The relationship is unclear. EM shows a smooth zone of apposition between glial cell and neuron, with nothing to suggest what transpires between them. Oligodendrocytes in culture show shallow rhythmic pulsations. The significance of such activity is not known.

3. Microglia

The *microglia* comprise tiny cells throughout the CNS (Fig. 15, M). Resembling oligodendrocytes, they are even smaller and darker, with dense oval, elongate, or triangular nuclei, meagre somata, and short tortuous processes with minute pointed thorns (also found on the soma).

The embryologic origin of the microglia is disputed. Most believe they are of mesodermal origin, invading the young CNS with the blood vessels. In context, the mesoderm in question derives from the neural crest: cells at the crests of the neural folds that break free to enter the mesodermal compartment at the time of neural tube closure. These cells undergo ectomesenchymal transformation. Accordingly, many embryologists characterize them as ectomesoderm. Almost taken up in the CNS, at the final moment they depart, differentiate, and rapidly migrate away to form a host of derivatives, including dorsal root ganglion cells, autonomic ganglion cells, and Schwann cells, i.e., peripheral neurons and glia. Thus, it may be that the microglia are not invading the young CNS, but coming home.

Others favor hematopoietic origin of the microglia, citing blood-borne mononuclear cells derived from bone marrow. Most neuropathologists recognize this added source of macrophages after brain insult. If vessels are damaged (they usually are), both the resident microglia and monocytes from the blood participate in repairs. Microglial cells proliferate, enlarge, and become puffy and phagocytic, clearing away cellular debris and ingesting products of degenerating myelin. Blood-borne macrophages do the same. The term *microglia* has been suggested for the minimally active resident population and *phagocytes* for cells responding to insults. The latter is appropriately noncommittal. Phagocytes may derive from the microglia, circulating monocytes, capillary pericytes, and outlying connective tissue. A third origin that has been postulated is that microglial cells arise from the same neuroepithelial stem cells that give rise to the macroglia: the astrocytes and oligodendrocytes.

4. Ependymal Cells

Ependymal cells line the brain ventricles and central canal of the spinal cord (Figs. 15, E, and 19). They arise from the pseudostratified neuroepithelium from which neurons and neuroglial cells originate. That surface is ciliated here and there, and some cilia are seen in the ependyma. They form an apparent simple cuboidal or columnar epithelium with microvilli and occasional cilia, but in fact the bases of certain cells taper into long, slender, outward processes.

In the embryo, some processes reach the surface of the neuraxis, establishing end feet there or on nearby capillaries. Later, they shorten and ultimately disappear, ending somewhere in the neuropil. Where the CNS wall is thin, as near a choroid plexus, they extend the whole way between central canal and pia, expanding into end feet that collectively make up a thin, smooth, external limiting membrane under the pial surface. Shorter processes entangle with those of astrocytes in a dense subependymal layer, the *internal limiting membrane*, close beneath the ventricular surface.

Modified ependymal cells are found in the choroid plexuses and third ventricle. In the former, they form a simple cuboidal secretory epithelium, with moderately extensive rER, many mitochondria, and irregularly oriented microvilli at their free surfaces, some with bulbous expansions at their tips. In the latter, *tanycytes* have processes extending into the hypothalamus to end near neurosecretory cells and the capillary plexus of its portal circulation. Their function is uncertain. They

may transport hormones in the CSF to these neurons to regulate the release of adenohypophyseal hormones into the portal system or from these neurons into the CSF.

5. Satellite Cells and Schwann Cells

The Schwann cell arises from the neural crest cell population that leaves the closing neural tube. As noted, it invests peripheral axons, forming and maintaining their myelin sheaths (Fig. 21J). These cells are crucial to nerve regeneration. With assistance from macrophages, they clean up debris from degenerating axons and myelin and, in cordons, guide axonal sprouts peripherally. Schwann cells, along with endoneurial and epineurial collagen and elastic fibers, provide structural support and limited elasticity to axons, which are subject to stretch during limb movements. Other functions include collagen synthesis during development and presentation of endogenous antigens to lymphocytes in autoimmune peripheral nerve disease.

A satellite cell (Fig. 15, S_3) is a type of Schwann cell that is more rounded than its counterpart and apposed to the soma of a sensory or autonomic ganglion cell. Like the satellite oligodendrocyte in the CNS, it seems to play sustaining functions, but little is known about either. In the ganglia of the VIIIth cranial nerve, satellite cells myelinate the bipolar cell bodies as well as the axons. The number required is proportional to the surface area of the soma. The ratio is fairly constant, whatever the neuronal size. One satellite cell may partially encircle the axon as it exits, so that the same cell may cover part of the cell body and part of the axon. Whether one cell type or two, Schwann and satellite cells are considered the neuroglia of the PNS.

B. Functions of the Neuroglia

1. Physical Support

We have noted the physical support afforded PNS nerve fibers by Schwann cells. In the adult CNS, astrocytes form an elaborate framework in which the neurons and their processes are deployed in an intricate, regionally distinctive way. Astrocytes have a robust cytoskeleton dominated by GFAP intermediate filaments, with many actin filaments and microtubules. They are ideal candidates for structural support. Similarly, radial glia in the embryonic neural tube support and maintain the patency of its lumen. Fluid pressure therein is crucial to normal forebrain development. Astrocytes derived from the radial glia may serve a similar function for the ventricles of the adult brain.

2. Parcellation of Input

In this astrocytic labyrinth, inputs are often individually enveloped and isolated from one another. In electron micrographs, wherever a neuron or a neuronal process is not engaged in a synapse, it is frequently enveloped by the somata or processes of glial cells. Yet exceptions are easily found. Small unmyelinated CNS axons are virtually bare, brushing by other neuronal and neuroglial elements. The extent of coverage of neurons varies greatly too, from heavily draped motor neurons and Purkinje cells, where only specifically directed axons effecting synapses attain surface contact, to almost bare neurons in the cerebral cortex bounded by few or no astrocytic processes (Fig. 16).

The distribution of astrocytic processes appears neither random nor to provide only mechanical support and nutrition to neurons. Cajal held that they are always disposed to prevent the contact of neuronal processes at points other than those suited to specific circuits. His belief is supported by electron microscopy. Sanford Palay and Alan Peters suggest that astrocytes ensure that terminals act in a discrete and localized manner. Each neuron has a distinctive pattern of glial investment suited to its pattern of synaptic connections. Velate astrocytes may play this role in the neuropil. A multineuronal, polysynaptic glomerulus is always heavily guarded by astrocytic sheets in instances forming capsules several layers thick. Thus, by isolating the diverse input converging on a neuron, astrocytes seem to play a key role in the communication functions of the nervous system.

3. Occupation of Injured Areas and Vacated Synaptic Sites

Astrocytes comprise a major part of the limited resources available to the CNS for responding to damage or disease and subsequent repair. Astrocytes or their precursors multiply in a mysterious manner. Mitotic figures are never seen; hence, the process of division is said to be *amitotic*. As a result, an increase in astrocyte number occurs along with an enlargement of cell size. Collectively, this hyperplasia and hypertrophy constitute *astrocytosis*. The hypertrophy involves both cell bodies and processes and is usually accompanied by *astrogliosis*: enlargement and excessive

production of astrocytic fibrils. Astrogliosis can also occur without hypertrophy, especially in chronic diseases, but here we focus on the destructive injury of brain tissue. Hence, the astrocyte responses are secondary to the degeneration of neurons and myelin, however caused.

4. Presynaptic Glial–Neuronal Interaction

The rapidly emerging chemical and molecular complexities of synaptic transmission are beyond the scope of this entry. But the processes of neuroglial cells, especially astrocytes, are usually on site at synapses, in some cases almost everywhere and in others less so. The intimate participation of these cells, not just in supportive, shielding, and inactivating roles but in chemical interplay with interacting neurons, is striking. Presynaptic interaction is seen in many of the synapses currently described.

5. Regulation of Electrolytes and pH in CNS Extracellular Space

As noted, astrocytes are connected by numerous gap junctions, permitting electronic coupling and passage of small molecules, especially ions. This feature enhances their spongelike ability to take up unwanted elements from the limited extracellular space in the CNS. Regulation of potassium ions therein is critical, because potassium ions could otherwise reach values incompatible with normal synaptic activity. Other significant astrocytic regulation involves extracellular pH by removal of CO_2.

6. Transmitter Reabsorption

We have already emphasized the necessity of rapid removal of neurotransmitter molecules after neurotransmission and described five ways in which this imperative could, and may in some cases, be done. One of these is uptake by astrocytic processes. The presence of these around synapses, in places close and in others not distant (see preceding discussion), facilitates the regulation of transmitter uptake and inactivation of synaptic activity. The entities removed by these processes are transmitters (notably glutamate, but including GABA and serotonin), ions (notably K^+), and toxins (e.g., metals).

7. Phagocytosis

As described, the cells of the microglia are the major in-house scavengers of the CNS. They may be reinforced by blood-borne macrophages. Astrocytes also seem to be capable of phagocytosis in situations, such as in degenerating optic nerves, where scar tissue prevents access by other phagocytotic cells. In the PNS, the ubiquitous Schwann and satellite cells play a major role in ingesting axonal and myelin fragments, supplemented by nearby connective tissue histiocytes (phagocytes) and hematogenous monocytes.

8. Participation in Neuroendocrine Regulation

Many functions are postulated for ependymal cells, as well as for intraventricular nerve endings lying between them and at the free surface. Ependymocytes seem to play roles in relation to the CSF: stir it by ciliary action, remove debris, microorganisms, and alien cells by adhesion to a cilium-borne lectin, produce the subependymal layer of the adult, which serves as a continuing source of neurons and neuroglial cells, possibly support the ventricular wall and maintain its patency, and provide sensory and secretory functions. As noted, the tanycytes in the basal third ventricle may play one or more of several neuroendocrine roles: transport hormones in the CSF to hypothalamic neurons to regulate the release of hormones into the portal system or carry them from these neurons into the CSF.

9. Myelination of Axons

Ensheathment of axons, by oligodendrocytes centrally and Schwann cells peripherally, to enhance nerve impulse conduction velocity in some or save space in bundling others was reviewed earlier.

10. Induction of Barrier Properties in Capillary Walls

Blood–brain barrier properties are encouraged by cells in the developing brain. Fenestrations and loose association of endothelial cells are obtained when the CNS is vascularized, but these properties are unsuited to CNS homeostasis. Thus, the windows shortly disappear and tight junctions begin to form. In tissue culture, these refinements are shown to require the presence and influence of brain tissue. The cells exerting these influences could be neurons, oligodendrocytes, or astrocytes, perhaps two of these, or all of the above. Through their processes and end feet upon the outer surface of blood vessels, astrocytes in some way appear to induce, along with other cells of the CNS, capillary endothelial barrier properties.

Together with cAMP agonists, they increase the interdigitation and total area of endothelial cell tight junctions.

11. Developmental Guidance of Neuronal Migration

The neuroglia are crucial to CNS histogenesis. Primitive glial cells, the radial glia, provide a structural scaffolding for migratory young neurons. Their migration involves translocation of cell bodies along a pre-existing neuronal process: *outward*, as for neurons leaving the ventricular zone for the cerebral cortex, or *inward*, as for cells of the transient external granular layer of the cerebellar cortex. These descend along inward processes past the Purkinje cell layer into the internal granular layer. In the developing cerebral hemisphere, the distances involved may be great (3000 μm in the case of the neocortex in primates). Hence, glial guidance and assistance are necessary, as follows.

Migrating young neurons shifting their somata outward in the intermediate zone of the hemisphere are apposed to radial glial fibers. At its leading and trailing processes and at its bipolar cell body, a neuron may partly encircle the vertical glial shaft (like someone climbing a rope, wrapping arms and legs around it and hugging it tightly). At such points, interstitial junctions are seen: widening of the space between, with filamentous material in it. Arrangements for descending cerebellar granule cells are similar: neuroglial contact guidance for neurons. This assistance is much enhanced by *astrotactin*, a glycoprotein neural receptor offering control of cell position along the radial fiber system.

12. Release of Neuronal Growth Factors

As noted, astrocytes produce growth factors, which may act alone or in combination to regulate the structure, proliferation, differentiation, and longevity of neurons and the development of astrocytes and oligodendrocytes and their functions as well. The organizing principle emerging is all too familiar: unified function of the cellular elements of the human nervous system.

13. The Dark Side of the Neuroglia

The neuroglia gives rise to mitotically dysfunctional forms. It has ghastly potential for abnormal growth. Glial cells (especially astrocytes, but also oligodendrocytes and ependymocytes) can wreak havoc in the CNS. Excepting metastatic cancer, these intrinsic elements of the brain comprise the largest source (50–60%) of primary brain tumors.

The neuroglia is a central topic of neuropathology. Glial tumors are now confronted by remarkable advances in treatment, stemming from molecular biology, genetics, immunology, and other fields. Reactive changes of the various glial cells, especially increased numbers of astrocytes and their fibers and the swelling of oligodendrocytes, follow almost every type of brain disease or injury and, in addition to playing phagocytic and reparative role as described, provide key early indications of abnormality. These warning signs have mainly been noted postmortem. But they can be detected or inferred from modern brain imaging and during neurosurgical procedures on the brain. Someday, we may have some high-technology methods of detecting these subtle reactions before it is too late.

X. CONCLUSIONS

The human nervous system is similar in cellular elements, structure, function, and basic plan to the nervous systems of all vertebrates, especially those of its fellow mammals. This article has highlighted many organizing principles of the human nervous system: anatomical pervasiveness, physical and functional coherence, centralized organization, structural specialization, use-designed components, phyletic uniformity with versatile adaptability, inherent plasticity, and recourse to chemical messengers (diversified neuroactive substances that perform short- and long-range tasks). These generalizations afford perspective and context for more complex principles and bewildering details.

In documenting these, this article has overviewed the three major divisions of the human nervous system, its seven principal CNS regions, its major neuronal and neuroglial subpopulations, and its diversified cellular, subcellular, and molecular features.

As is apparent, the structure and organization of the human CNS are multifaceted and complex. Those concerned with functional circuitry seek to discern the fine details of the pathways serving specific functions and unite these with the neuroanatomy described earlier. Certainly, the major pathways are defined at an anatomical level, but what remains is the integration of anatomy with function. In this task,

we must consider models of brain organization and bring these together with real structures and real connections.

First, the *linear* model. Stimuli are processed in a *pathway-specific* manner: like a road map. This is classical neuroanatomy. Whatever model pre-vails, this one will always be valuable, in the clinic and in the grand history of neuroscience before anyone recognized this supreme discipline or called it that.

Second, the multiple pathway concept, the *lens model*, of neural processing, converging on a single endpoint. Particular stimuli can select many pathways and still arrive at common endpoints. These features embody the concept of *focused output*, where the pathways are several but the output convergence points obligatory, e.g., as in recovery of function after brain damage and probably in several neurodegenerative diseases.

Third is the *network* model. It is the most versatile, addressing functional neuroanatomy in terms of probabilities and multiple cross-communicating pathways that develop a *consensus output* but nonetheless activate many outputs and options. A particular stimulus may be processed over many pathways and arrive at several outcomes, depending on variables not well-understood.

All of these models, and perhaps others, are valid. All apply, both to discrete stimuli and to output transformations. Neuroscientist Carl Cotman observes, ''Functional neuroanatomy posits we know the basic circuits. From these we elucidate pathways serving function, covering a range from normal sensory activity to the neuronal plasticity operational in brain damage and disease. It is unlikely that one model will be involved in or satisfy all computational transactions of the human nervous system, nor adequately explain them.''

From what we know of the seamless interplay of neurons, neuroglia, the central and peripheral regions, functional subsystems, chemical substances, and non-neural components of the human nervous system (notably its autoregulatory vasculature), it may be that many models of the widest sort will come together, if and when we have an explanation of the human mind.

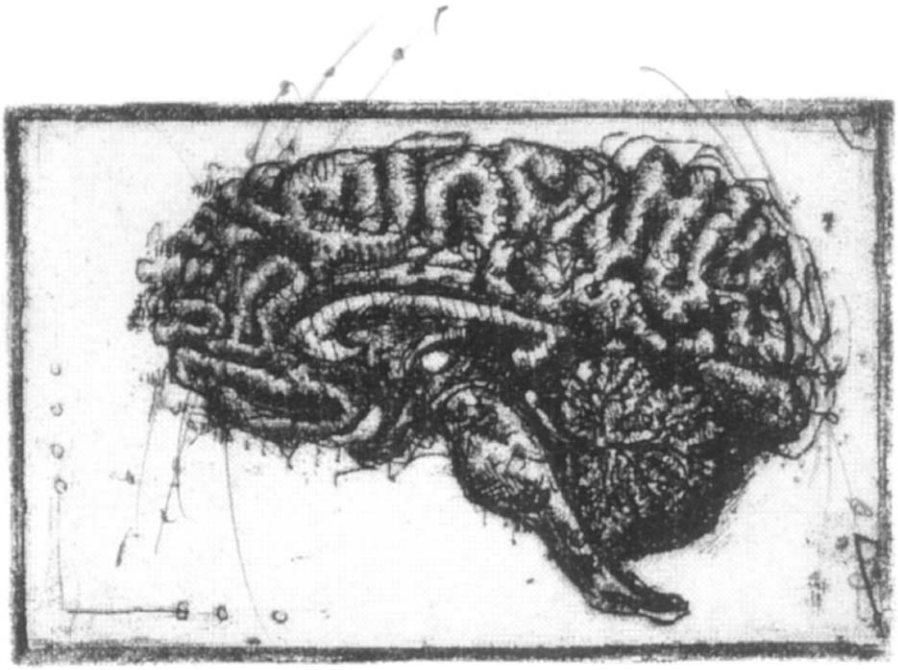


Illustration by Cheryl A. Cotman

Light breaks on secret lots,
On tips of thought where thoughts smell in the rain . . .

— "Light Breaks Where No Sun Shines"
Dylan Thomas, 1933, published 1934

See Also the Following Articles

NEUROANATOMY • NEUROGLIA, OVERVIEW • NEURON • PERIPHERAL NERVOUS SYSTEM

Suggested Reading

Angevine, J. B., Jr. (1994). The Nervous Tissue. In *Bloom and Fawcett, A Textbook of Histology* (D. W. Fawcett, Ed.), 12th ed., pp. 309–367. Chapman & Hall, New York, London.

Angevine, J. B., Jr., and Cotman, C. W. (1981). *Principles of Neuroanatomy*. Oxford University Press, New York, Oxford.

Bodian, D. (1972). Neuron junctions: A revolutionary decade. *Anat. Rec.* **174,** 73–82.

Brodal, A. (1981). *Neurological Anatomy: In Relation to Clinical Medicine*, 3rd ed. Oxford University Press, New York, Oxford.

Brodal, P. (1998). *The Central Nervous System: Structure and Function*, 2nd ed. Oxford University Press, New York, Oxford.

Bullock, T. H., Orkland, R., and Grinnell, A. (1977). *Introduction to Nervous Systems*. W. H. Freeman and Company, San Francisco.

Cooper, J. R., Bloom, F. E., and Roth, R. H. (1996). *The Biochemical Basis of Neuropharmacology*, 7th ed. Oxford University Press, New York, London.

Gazzaniga, M. S., Ivry, R. B., and Mangun, G. R. (1998). *Cognitive Neuroscience: The Biology of the Mind*. W. W. Norton & Company, New York, London.

Kandel, E. R., Schwartz, J. H., and Jessell, T. M. (Eds.). (2000). *Principles of Neural Science*, 4th ed. McGraw-Hill, New York.

Nauta, W. J. H., and Feirtag, M. (1986). *Fundamental Neuroanatomy*. W. H. Freeman and Company, New York.

Nolte, J. (1999). *The Human Brain: An Introduction to Its Functional Anatomy*, 4th ed. Mosby, St. Louis.

Nolte, J., and Angevine, J. B., Jr. (2000). *The Human Brain in Photographs and Diagrams*, 2nd ed. Mosby, St. Louis.

Peters, A., Palay, S. L., and Webster, H. DeF. (1991). *The Fine Structure of the Nervous System: Neurons and Their Supporting Cells*, 3rd ed. Oxford University Press, New York, Oxford.

Shepherd, G. M. (1994). *Neurobiology*, 3rd ed. Oxford University Press, New York, Oxford.

Zigmond, M. J., Bloom, F. E., Landis, S. C., Roberts, J. L., and Squire, L. R. (Eds.). (1999). *Fundamental Neuroscience*. Academic Press, San Diego, London.

Neural Networks

PAUL SAJDA
Columbia University

GLOSSARY

action potential Also called a "spike," the all-or-none response elicited by a neuron, the frequency of which depends upon the total input to the neuron. A sequence of action potentials is called a "spike train."

binding problem The problem of how activities of neurons, which are distributed across the brain, are linked to represent the same object.

compartmental model A biophysical model of a neuron based on a series of circuit elements representing passive and active electrical properties of cell membranes.

credit-assignment problem The problem of determining the contribution of each neuron and connection weight to the output of the network.

emergent computation Computation resulting from the interactions of many neurons within a network.

objective function The function that is optimized in training artificial neural networks and usually includes a term related to the error of the network.

supervised training A training procedure that is based on a set of exemplars composed of sample input data and the correct or desired output—"truth data." The error between the actual and desired output is directly incorporated into the objective function.

synapse From the Greek word "to clasp," the location of electrochemical communication between connected neurons. In artificial neural networks synapses are represented by connection weights.

unsupervised training A training procedure that does not rely on labeled "truth" data. Often used in data clusters and the construction of associative memories.

Neural networks are interconnections of simple units, called neurons, capable of computing complex functions. Hallmarks of a neural network are nonlinear response properties of single neurons, a high degree of interconnectivity between neurons, and adaptation and learning of connectivity and associated parameters. Neural networks are often divided into two general classes, biological and artificial, though this is less of a dichotomy than a spectrum. In the strictest sense, a biological neural network is a network comprising real neurons (i.e., constructed from living neural tissue) and found within a biological nervous system. Models of biological neural networks have been developed that abstract and simplify, at some level, biological realism in order to study a specific aspect of the network or reduce it to the hypothesized key elements underlying its behavior. These models are termed biologically based neural network models, though they do not capture all the complexity of the real biological system. As the level of abstraction increases, the tendency is toward less of a direct link with neurobiology. Artificial neural networks lie at the end of the spectrum at which mathematical description and compactness become more important than biological plausibility. Nonetheless, much of the research in the world of artificial neural networks has, at its origin, the study of biological neural networks. Computer models of both biological and artificial neural networks have been employed in the study of brain function, as well as in the development of "brainlike" systems

for prediction, recognition, and control in real-world environments.

I. THE NEURON AS A BASIC COMPUTATIONAL ELEMENT

In some sense, neurons in a neural network play the same basic computational role that transistors play in a digital computer, though they are far from isomorphic, e.g., the method of computation and the representation of information are quite different. The concept of the neuron as a basic unit in the nervous system has at its origins the "neuron doctrine," which was formulated by Wilhelm Waldeyer in the 1890s and is largely based on the seminal neuroanatomical work of Santiago Ramon y Cajal and the nerve cell stain discovered by Camillo Golgi, both of whom shared the Noble prize in 1906. The neuron doctrine states that the nervous system is composed of discrete units or cells, called neurons, which are both structurally and functionally discrete, having their own cell membranes and functioning as a fundamental signaling unit. Further, the doctrine states that the connection between these discrete units is highly specific. At the time, this was in contrast to competing theories that hypothesized the nervous system as a syncytium or as an amorphous collection of cell bodies that essentially share a common cell membrane, implying no real connection specificity. The neuron doctrine has been important in establishing the currently accepted definition of a neural network, namely, that discrete units can be connected in highly specific ways to enable complex behaviors in biological and artificial nervous systems.

There are several hundred types of excitable nerve cells in the cerebral cortex, some of which have differences in their biophysics of computation. Most classes, however, have similarities related to the nature of information flow within and between neurons, their input–output characteristics or transfer function, and the coding and representation of information.

The biophysical basis of information flow within a neuron rests in the dynamics of ionic concentration gradients. Biological neurons are described as having a set of neural processes, which include their dendrites

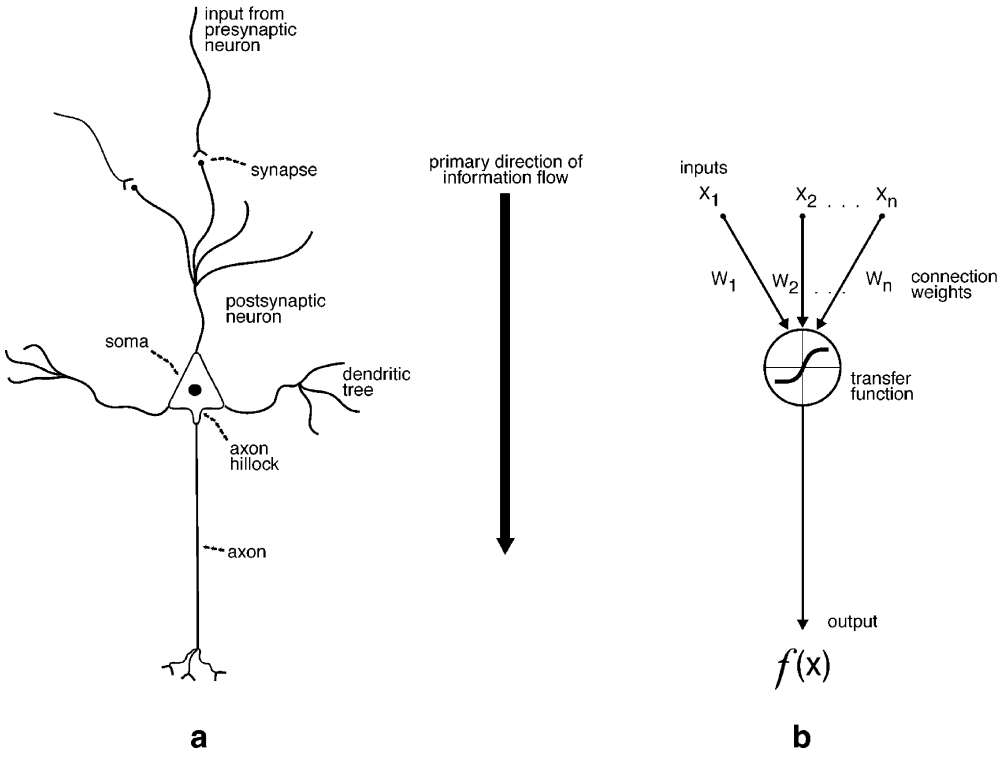


Figure 1 Modeling a neuron: (a) biological neuron, including neural processes, (b) artificial neuron.

and axon, as shown in Fig. 1a. Ionic gradients across the cell membrane produce voltage changes that propagate down the neural processes. The voltage changes are typically graded until they reach the soma where, at the axon hillock, they sum and generate, with some probability, an all-or-none response, termed a spike or an *action potential*. Action potentials travel down the axon until they reach *synapses*, where they induce the release of a neurotransmitter. There are many types of neurotransmitters, however, two of the most prominent in the brain are glutamate and gamma amino butyric acid (GABA). Neurotransmitters are often characterized by whether they excite or inhibit the postsynaptic neuron, with glutamate being excitatory and GABA being inhibitory. Neurotransmitters diffuse across the synaptic cleft and bind to the cell membrane of a neighboring neuron. Binding of neurotransmitter changes the conductance of ionic channels so that a voltage change is induced in the postsynaptic cell. At the synapse, the transduction of the electrical signal (voltage) to a chemical signal (neurotransmitter) and back to an electrical signal can amplify or attenuate the signal received postsynaptically, depending on the strength of the synapse. This ''synaptic weighting'' plays an important role in the input–output characteristics of a neuron as well as being important for learning and adaptation. The postsynaptic voltage travels down the dendrites of the second neuron and, in this way, information is communicated between the two neurons.

Information flow is thought to be primarily in the direction from the dendrites to the soma (cell body), down the axon and finally across synapses to other neurons. The concept of unidirectional information flow in a single neuron has been an important element of many artificial neural network models, although more recent experimental evidence has shown that the neurobiology is more complex. For example, there is evidence that action potentials propagate back through the dendrites. This backpropagating signal can be important for different types of adaptation and learning, including long-term potentiation (LTP) and Hebbian learning, as will be discussed later.

Neurons are interesting computational elements because of the nonlinear way in which they transform their inputs. From a purely computational point of view, complex behaviors and functions can only be computed if there is a source of nonlinearity in the information processing flow, e.g., to implement multiplication requires a nonlinearity. Warren McCulloch and Walter Pitts in 1943 formulated a nonlinear model that is a binary device, with the output of the neuron being either 1 (''on'' state) or 0 (''off'' state). The state of the McCulloch-Pitts neuron, as it became called, is determined by the synaptic input mediated by both excitatory and inhibitory synapses. The neuron transitions to the ''on'' state if the sum of its excitatory synaptic input is greater than a threshold value and no inhibitory synapses were active, otherwise it transitions to an ''off'' state. The work of McCulloch and Pitts was one of the first attempts to understand the computational properties of the nervous system through consideration of the nonlinear properties of single neurons.

The McCulloch–Pitts neuron, though having some interesting computational properties, was far removed from the neurobiology and biophysics of real neurons. In biological systems, nonlinearities are the norm, so it is important to consider which nonlinearities in real biological neurons are utilized for computation. In the 1950s, Alan Hodgkin and Andrew Huxley's Nobel prize winning work describing the properties of excitable cells began to shed light on this question. By using the giant squid axon as their subject, Hodgkin and Huxley demonstrated how action potentials are generated by the nonlinear properties of voltage-dependent ion channels in the cell membrane. They modeled this complex physiological behavior by using a set of coupled differential equations, fitting the parameters of their model to experimental data. The Hodgkin–Huxley model is still widely used for modeling biologically based spiking neurons.

Though there are other sources of nonlinearity in biological neurons, the Hodgkin–Huxley model is important for establishing the nonlinearity of action potential generation. The relationship between a neuron's postsynaptic potential (i.e., its input) and the generation of action potentials is often termed the neuron's transfer function. The transfer function determines how a neuron will map its input to an output. To simplify things mathematically and to focus more on the computation of entire networks rather than individual neurons, the biophysically based model of Hodgkin and Huxley has been abstracted in various ways to yield a variety of transfer functions. One such abstraction is the ''integrate-and-fire'' neuron. The basic biophysical mechanism governing the behavior of an integrate-and-fire neuron is the change in membrane voltage due to injection of a current, for example, at a synapse. In the integrate-and-fire neuron, the membrane essentially acts like a capacitor. If enough current is injected into the cell, the voltage will increase until it reaches a threshold, at which time an action potential is generated, the

membrane potential resets, and all of the charge is dissipated. The integrate-and-fire neuron can be made more realistic by adding a resistance into the membrane equation, allowing for the leakage of current that is observed in real neurons. These models are termed leaky integrate-and-fire neurons.

Even simpler models ignore the finer temporal information of individual spikes and condense a sequence of spikes—the spike train—into a single number called the firing rate, which represents the number of spikes generated by the neuron over a given time interval. The advantage of this abstraction is that the transfer function, which captures the relationship between input and firing rate, can be represented accurately by classes of functions with nice mathematical properties: such as, they are continuously differentiable. One such class of function is the sigmoid. For these model neurons, shown in Fig. 1b and often used in artificial neural network models, the firing rate encodes the information represented by the neuron and all information related to spike-timing is lost, e.g., when a spike occurred relative to some other spike. In biological neural networks, firing rates of a population of neurons are believed to be used for encoding movement direction. The population vector response in the primate motor cortex, first observed in the monkey by Apostolos Georgeopoulos, has each neuron encoding a movement direction. The monkey's intended movement can be predicted from the sum of the neurons' direction vectors, each weighted by their relative firing rate.

In more recent years there has been considerable debate on whether firing rate or spike timing is the more optimal coding–representation strategy. For example, humans can recognize familiar objects in 150 ms, which corresponds to <25 ms per processing stage between retina and cortical recognition areas. The biology thus dictates that each processing stage must be capable of integrating and responding to the initial wave of arriving spikes without requiring additional processing iterations. The biology also indicates that the computation cannot depend upon traditional rate coding (40 Hz firing rate=25 msec between spikes). One mechanism that allows sufficiently fast analog computation for recognition is "space–rate coding." In space–rate coding, stimulus information is encoded by the fraction of neurons in a population that are active within a short time window (e.g., 5 msec). Because the fraction active can change on a millisecond time scale, space–rate coding allows a rapid and high-resolution readout of network computations.

II. BIOLOGICALLY BASED NEURAL NETWORKS

The complex behavior of neural systems perhaps has less to do with the specifics of the individual units and more with the connection of units to form networks, leading to *emergent computation*. Computational neuroscience focuses on the modeling of biologically based neural networks. Computational neuroscience developed into its own field in the 1980s largely due to advances in neurophysiological recording and the continued development of low-cost digital computers. More recent advances in noninvasive neuroimaging, for example, the development of functional magnetic resonance imaging (fMRI), high-density electroencephalography (EEG), and magnetoencephalography (MEG), have led to a new set of tools and data for supporting the construction and validation of biologically based neural network models.

In terms of neural modeling, the hippocampus is a brain structure that has received much attention. The hippocampus plays an important role in memory, seemingly acting like a buffer for the storage of short-term memories, as well as being involved in spatial navigation tasks. Roger Traub and colleagues have constructed a neural network model of the mammalian hippocampus, which consists of roughly 10,000 neurons, each neuron built using realistic models of several types of ionic channels. Consistent with anatomical data, the neurons in the network are sparsely connected (less than 5% connectivity). Network simulations produce a population-based rhythmic activity of 4–8 Hz known as the θ rhythm, which is observed in normal hippocampus. The activity is population-based because individual neurons do not fire regularly at the θ rhythm. Only by considering the population or network response does one see this emergent behavior. The model reproduces a variety of responses observed *in vivo*; specifically, the network can generate "seizures" similar to those seen in humans and monkeys. Because the hippocampus is involved in a variety of neurological diseases, including epilepsy, Alzheimer's disease, and Down's syndrome, one promising area of research is the development of computational models for evaluating the efficacy of various types of treatment.

Other neural network models of the hippocampus, including those by John Lisman and colleagues, have exploited two types of rhythms observed *in vivo*, the θ rhythm and the γ rhythm (40–60 Hz). In Lisman's model, a short-term memory is stored in each cycle of a γ oscillation and all short-term memories are packaged in the θ rhythms. The model predicts that roughly

seven short-term memories can be stored at any given time, consistent with well-known psychological data.

Neural networks have been used to model oscillatory and rhythmic phenomena in areas other than the hippocampus. For example, several network models have been developed that hypothesize a role for oscillations and temporal patterns of spikes for solving the so-called *binding problem*. The binding problem in audition is the well-known "cocktail party problem." When attending a noisy cocktail party with multiple speakers and noise sources, you will notice it is relatively easy to separate out the voice of a single speaker and understand what he or she is saying. This is in spite of the fact that the signals from the sound sources in the room are mixed together, often in very complicated ways due to reverberation and attenuation. The brain ultimately must deal with this mixture, separating the neural signals that belong to the target speaker from those that belong to the background noise. In vision, the problem is compounded by the fact that the first thing the visual system appears to do is decompose objects into their associated color, motion, depth, position, etc. In fact, this decomposition occurs in two different streams, termed the "what" and "where" streams by Robert Desimone and Leslie Ungerleider. These streams are in different parts of the cortex—temporal and parietal lobes—and no region has been identified in the brain in which all of the information converges, implying that the representation remains distributed.

Charles Gray and Wolf Singer have found evidence that temporal patterns may play a role in binding a distributed neural representation. On the basis of their neurophysiological recordings, they propose that neurons representing the same object fire in synchrony. In a network having excitatory and inhibitory neurons, this synchronous firing could also lead to oscillations. Several neural network models have been developed that use synchrony and/or oscillatory activity as a coding scheme for representing and binding objects. John Hopfield and Carlos Brody have developed a neural network model constructed from integrate-and-fire neurons that is selective for specific spatiotemporal patterns in the input stimulus. In their model, the presence of a stimulus is signaled via transient synchrony among a population of neurons in the network. One important element of the model is that the network easily and rapidly desynchronizes when no recognized spatiotemporal patterns are present. Desynchronization is just as important as synchronization if the network is to use these temporal patterns as a coding mechanism for objects. The recognition event, that being transient synchrony, is detected by a neuron that acts as a coincidence detector.

Whether the brain explicitly solves the binding problem in order to recognize objects is still under much debate. A neural network model developed by Maximilian Reisenhuber and Tomaso Poggio illustrates that impressive visual object recognition performance is achievable using a purely feedforward model, which does not explicitly necessitate binding or segmentation. The model is based on the neurophysiological experiments of Keiji Tanaka and colleagues and their recordings from the anterior portion of the inferior temporal cortex (IT) in the monkey. IT is believed to be an important visual area for object recognition, and recordings in IT have shown cells having high specificity for individual objects, such as faces. Tanaka and his group developed a set of complex visual stimuli (e.g., starbursts, junctions of lines, bull's-eyes), which are quite different from the bars and edges traditionally used as stimuli by neurophysiologists. They found that individual IT neurons respond preferentially to these features—one neuron only fires for bull's-eyes, another only for T junctions. One hypothesis is that they function as a large set of complex feature detectors. Reisenhuber and Poggio constructed a neural network that builds up a large set of complex feature detectors using a hierarchy of linear and nonlinear combination rules. In their model, neurons at higher levels in the hierarchy have response properties, which are very similar to those observed by Tanaka *in vivo*. Reisenhuber and Poggio argue that these responses are a good basis for representing visual objects because individual objects will tend to uniquely cluster in small regions of this high-dimensional feature space. By feeding these representations into an artificial neural network, they have shown that the representation captured the necessary information for robust, rotation invariant object recognition—a difficult visual recognition problem.

Further evidence that argues against oscillatory behavior of neural networks being a necessity for solving the binding problem comes from models of the piriform cortex by Matt Wilson and James Bower. Piriform cortex is the primary cortical area for olfaction (e.g., sense of smell). It is well-known that the EEG of the piriform cortex in rats exhibits γ oscillations when the rat sniffs an odor. Wilson and Bower constructed a *compartmental model*, consisting over 4000 neurons, to investigate the nature of these oscillations. Their findings showed that even when random inputs were given to the model, γ oscillations

continued. This implies that the oscillations are not involved in olfactory computation and odor recognition, rather they are an epiphenomenon of the network architecture. The debate over the computational role of oscillations in the cortex and subcortical areas thus is not resolved.

In addition to neurophysiology, anatomical data has been important in the development of biologically based neural networks. Many regions of the neocortex are organized topographically, with precise connectivity mapping the sensory world into the brain. For example, much of the visual system is retinotopically mapped, meaning that adjacent regions in the visual cortex are stimulated by adjacent positions in the visual environment, as projected onto the retina. Other prominent areas having topographic representations are the motor cortex and somatosensory cortex. One of the unique aspects of many of these topographic maps is that they have been shown to be adaptive. Leif Finkel, Gerald Edelman, and John Pearson developed a network model consisting of approximately 1500 neural units of somatosensory cortex that exhibited this adaptive behavior, often termed plasticity. Stimulation of the network model by repeated tactile input on one of the simulated fingers results in an increase in the representation of that finger within the network model. Simulated transection of afferent nerves, eliminating input from that finger, causes the corresponding region in the model to shrink, being taken over by regions receiving active tactile input. The phenomenon of a phantom limb, where an amputee patient experiences the sensation of their limb shrinking into their body and disappearing, is the analog of the dynamic remapping observed in this model.

The learning rule governing plasticity in many biologically based neural network models is based on the classic Hebb rule. In 1949, Donald Hebb formulated a learning rule that stated how the strengths of synapses are modified on the basis of pre- and postsynaptic activities. The basic notion was that, if neurons' activities are correlated or coincident, then the synaptic strength between them should be increased. In mathematical terms, one of the simplest forms of the rule is

$$\Delta w_{ij} = \eta \cdot o_i \cdot o_j$$

where Δw_{ij} is the change in synaptic strength, η is the learning rate, and o_i and o_j are the outputs or firing rates of the presynaptic and postsynaptic neurons, respectively. In this case, the synaptic strength is increased if the presynaptic cell and postsynaptic cell fire at the same time. A symmetric rule allows the synapse to be weakened if the cells fire at different times. Computationally, the Hebb rule strengthens synapses between neurons that have correlated activity and weakens those synapses between neurons having uncorrelated firing patterns. One of the issues with the classic Hebb rule is that it is a local rule based on the generation of action potentials by the pre- and postsynaptic neurons. Because this adaptation occurs directly at the synapse, which can be far removed from the axon hillock and axon, it is unclear how information about whether the postsynaptic neuron generated action potentials would be communicated back to the synapse. Several researchers have reformulated the Hebb rule so that the change in synaptic strength is based on pre- and postsynaptic potentials (local voltages), not firing rates. However, as mentioned earlier, more recent neurophysiological evidence has shown that action potentials can quickly back-propagate from the soma to the synapse, serving as a means to communicate the firing of the postsynaptic cell to the local synapse. Regardless of which formulation is used, the Hebb rule has provided modelers with a biologically plausible activity-dependent mechanism for constructing highly specific connection patterns, leading to emergent computation.

The specificity of the connections in the visual cortex has led to the development of several models hypothesizing a computational role for these connections. Leif Finkel, Shih-Cheng Yen, and Elliot Menschik, for example, have developed a biologically based network model, which reproduces several interesting quantitative psychophysical results of contour perception. The computation in the model is largely mediated by the specificity of long-range horizontal connections in layers 2 and 3 of the visual cortex. Anatomically, these connections extend over several millimeters and subtend over 10° of visual field. In their model, horizontal connections mediate facilitation of neurons representing collinear/cocircular contour elements, as shown in Fig. 2.

Biologically based network modeling has as its goal understanding how the biophysics, neurophysiology and neuroantomy give rise to complex behavior and computation within the brain. One of the challenges has been that a mathematical analysis of biologically based networks is difficult due to the complexity of the detailed neurobiology. This has also limited the size of the models, in terms of the number of neurons in the network, and therefore puts limitations on their evaluation with more complex and realistic input. As a result, some researchers have developed neural network models that trade off biological realism in

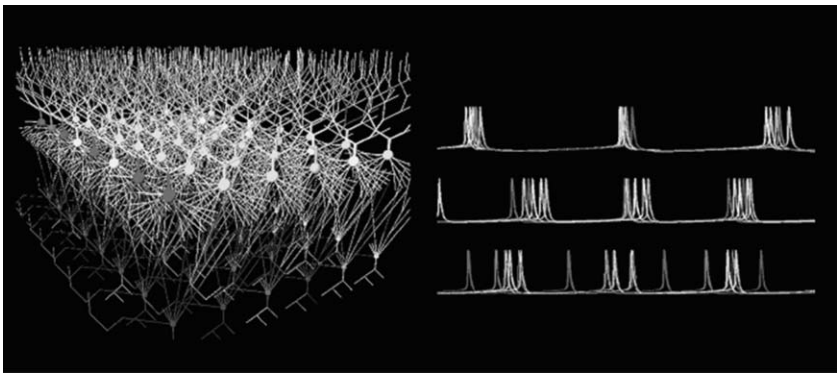


Figure 2 Cortical simulations of horizontal connections between columns of neurons in the visual cortex. Spike traces show the degree of synchronization in response to a six-element contour. Each column of neurons "sees" just one contour element; the spacing between elements determines the salience of the contour relative to background clutter. As the salience increases (top trace most salient, bottom trace least salient), the degree of synchronization decreases. Six hypercolumns are simulated with eight pyramidal cells and eight interneurons per orientation column. Each pyramidal cell is a 64-compartment model, and interneurons are 51-compartment models (reprinted with permission from L. Finkel).

favor of larger network size and whose overall behavior, through mathematical analysis, is tractable.

III. ARTIFICIAL NEURAL NETWORKS

A network whose architecture is not strongly linked to neurobiology is often termed an artificial neural network (ANN). ANNs are built by using simplified models of a neuron, for example, sigmoidal units mentioned earlier. The pattern of connections between the neural units is most often established via learning rules and gives the network its ability to compute complex functions and develop emergent behavior. The importance of connectivity in ANN models has resulted in the coining of the term "connectionism" for describing this research area.

ANN research has its origins in the development of the "perceptron" by Frank Rosenblatt in 1958. ANNs are often characterized by the number of layers of modifiable connections or synapses. The perceptron is a network consisting of a single layer of connections with output units having a nonlinear transfer function, as shown in Fig. 3a. The output, $f(\mathbf{x})$, of the perceptron is given by

$$f(\mathbf{x}) = \mathrm{sgn}\left(\sum_{i=0}^{N} w_i x_i\right)$$

where w_i are the connection weights and x_i are the inputs, with x_0 often chosen to be a constant equal to 1, called the bias. The sgn() function is equal to 1 if its argument is greater than or equal to 0 and -1 if it is less than 0. The perceptron is a linear network in that the output "decision" is linearly dependent on the input. The perceptron classifies its input, by outputting either a 1 or a -1, on the basis of how it projects onto its connection weights. Therefore, determination of the values for the weights is important for giving the perceptron its ability to classify. The connection weights are usually learned via *supervised training*.

Training ANNs involves formulating and minimizing an *objective function*. For the perceptron, the objective function, H, is given by

$$H(w) = \sum_{x_i \in S} (-w_i x_i)$$

where S is the set of examples that are misclassified, given the connection weights. One can see that $\sum_i w_i x_i \leq 0$ if x_i is misclassified. Thus, the minimum of H is 0.

Minimization of the objective function for the training data requires taking the gradient of H with respect to the connection weights and then updating the weights by moving in the direction opposite to the gradient (e.g., gradient descent). This is simply

$$\frac{dH(w)}{dw} = \sum_{x_i \in S} (-x_i),$$

and the update rule for the perceptron from time step t to $t+1$ becomes

$$w_i(t+1) = w_i(t) + \eta \frac{dH(w)}{dw} = w_i(t) + \sum_{x_i \in S} x_i.$$

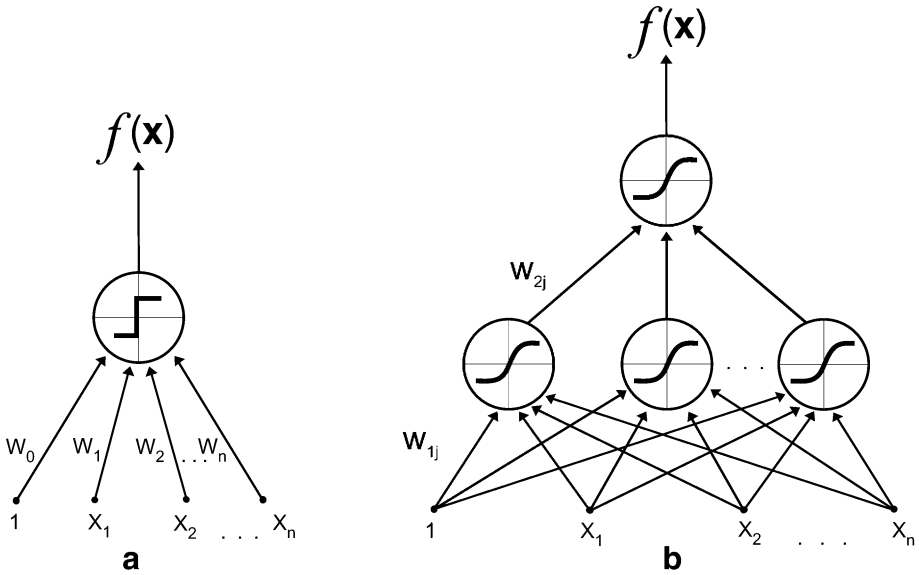


Figure 3 Artificial neural network architectures: (a) perceptron, (b) multilayer preceptron.

After learning on the training data, new data are input to the perceptron network with the goal being that it gives correct results on the new data, a property called "generalization." The perceptron generated much excitement over its ability to perform "brainlike" computation. This excitement, however, was squelched in 1969 when Marvin Minsky and Seymour Papert published their book *Perceptrons*, which clearly outlined the computational limitations of the perceptron architecture. The single-layer architecture made it impossible for the network to solve anything but linearly separable problems. A classic example in which the perceptron had difficulty was the exclusive-OR (XOR) problem shown in Fig. 4. Minsky and

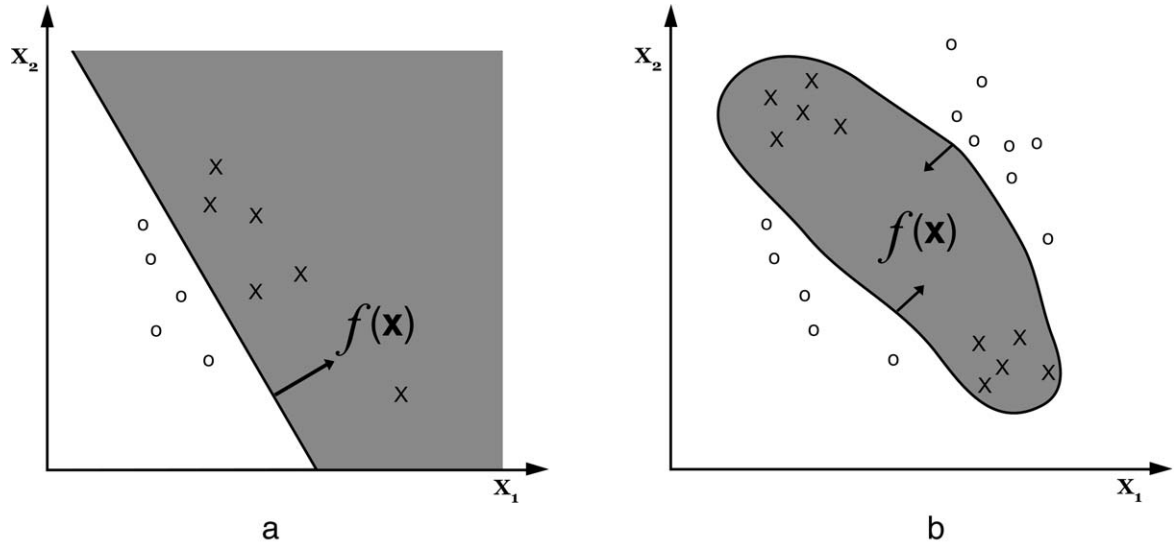


Figure 4 Linear versus nonlinear separable classification. (a) Linearly separable classification problem in x_1, x_2 feature space. A perceptron can learn the function $f(\mathbf{x})$ for solving linearly separable problems. (b) The exclusive-OR (XOR) problem is not linearly separable. In the XOR problem, data in x_1, x_2 feature space are distributed in much the same way as the exclusive-OR function in Boolean logic. A perceptron is not capable of learning a function $f(\mathbf{x})$ that classifies all the X's and O's correctly. An MLP, however, can learn $f(\mathbf{x})$ because the output unit combines the decision surfaces generated by all of the hidden units, thereby "piecing together" a complex decision boundary separating all X's from O's.

Papert's seminal work was a blow to the field, which lay largely dormant until the 1980s, when the development of a learning rule for training multilayer perceptrons (MLPs) was formulated.

MLPs, shown in Fig. 3b, consist of several layers of modifiable connection weights. Neural units in layers that are not directly connected to the output are called "hidden units." Connections are usually feedforward, with units in layer N connected to all units in layer $N-1$. The multilayer architecture enables the MLP to solve problems that are not linearly separable. In fact, the eminent Russian mathematician Andrei Kolmogrov proved that any continuous function can be implemented in an MLP with a sufficient number of hidden units. An important development for the field was the publication in 1986 by David Rumelhart, Geoffrey Hinton, and Ronald Williams of the back-propagation learning rule for training MLPs. The back-propagation learning rule was able to solve the classic *credit-assignment problem* for MLPs. The formulation of the back-propagation learning rule follows the same basic structure as the perceptron learning rule, namely, defining an objective function based on the training error, differentiating the objective function with respect to the connection weights, and then generating a weight update equation based on gradient descent. One of the important differences is that, because for an MLP the outputs are not directly connected to the inputs, the back-propagation learning rule requires the use of the chain rule of calculus to compute the required derivatives. The publication of the back-propagation algorithm started a flurry of development in the field, with ANNs being built that could learn to recognize printed characters to networks that could be trained to "talk." For example, the NETtalk network, consisting of 203 input units, 80 hidden units, 26 output units, and a total of 18,629 connections, learned to translate written text into phonemes that could be pronounced by a speech synthesizer. Today, MLP architectures and the back-propagation learning rule continue to evolve with new variants having better convergence properties, leading to more powerful networks.

ANNs are not limited to feedforward connection architectures, particularly considering how feedback connectivity can give rise to interesting dynamic properties in biological neural networks. Recurrent networks have architecture similar to feedforward MLPs, except that there are feedback connections between layers. John Hopfield, motivated by concepts in statistical physics, developed one particularly interesting class of recurrent networks. In a Hopfield network, all neurons are connected to one another—the network is termed "fully connected." Neuron responses are represented as binary states (1/0 or 1/-1), analogous to the spin of a particle ($+\frac{1}{2}$ or $-\frac{1}{2}$) in physics. Interactions between neurons, via their connectivity, influence their individual states much as interactions between particles influence their collective spins. As in the physical system, a network architecture, with a particular set of connections, will have an "energetically favorable" equilibrium state, which can be viewed as an attractor in a dynamic system. Training a Hopfield network involves learning the set of connections that makes a particular binary state of neurons an attractor. Input of a similar pattern, or the original pattern corrupted by noise, results in the network activity eventually converging to the closest attractor, with the output being the stored pattern. Hopfield networks, therefore, implement "associative memories" with properties similar to what has been observed in the hippocampus.

Many subtle issues are related to the construction, training, and evaluation of ANNs. For instance, there are issues on how to objectively select the best set of features for input to an ANN and even on how to select the best ANN model itself, e.g., what is the optimal number of layers, neural units, and connection weight given a particular data set and problem. A related issue is the "bias–variance trade-off," which can be thought of as the trade-off between the expected error of the network, d (the bias), and the variation of the error (the variance) for different subsets of training data. Ideally, one would like to minimize both the bias and variance; however, these two terms tend to vary diametrically. An overly simple network, with few parameters, will tend to have a large error across training subsets (high bias); however, the value of this error will not vary considerably across training subset (low variance). An overly complex network will tend to estimate the training data well (low bias); however, this estimate will likely vary considerably across different training subsets (high variance). The challenge is to find the optimal network that minimizes the combination of bias and variance. Readers interested in more detail are referred to Bishop (1995) and Duda, Hart, and Stork (2001).

Cognitive neuroscientists have used ANNs as tools for studying brain function, particularly for modeling high-level cognitive processing. Martha Farah and Jim McClelland have built a connectionist model composed of fewer than 200 neurons to study the organization of semantic memory and its role in agnosia, which is a failure in object recognition. They

use their model to test two hypotheses, the first being that semantic memory is organized into categories and the alternative being that the organization is based on object properties. For instance, a category-based organization might be based on living versus nonliving things, whereas a property-based system would organize on the basis of the visual and functional attributes of objects. By using a network model based on object-property organization, Farah and McClelland simulated "lesioning" their ANN by deactivating a percentage of the units in the semantic layer. These simulated lesion studies showed that loss of function in the visual and functional semantic units results in a specific categorical agnosia, with lesions to the visual units causing a loss in memory of living things and lesions to functional units resulting in a loss in memory of nonliving things. These results, consistent with neuropsychological literature, demonstrate that category-specific deficits are an emergent property of a network that is organized based on object properties.

ANNs have been used to solve very complex signal processing problems. The well-known cocktail party problem described earlier is the problem of unmixing individual speakers and other acoustic sources from a set of microphone signals. If little or no information is available about the sources or the environment in which the recordings took place, then the problem is often termed *blind source separation* (BSS). One technique for BSS that has proven very successful is independent component analysis (ICA). The goal of ICA is to find the mixture components that are most statistically independent. In 1995, Tony Bell and Terry Sejnowski of the Salk Institute showed that ANNs are very good at implementing ICA and, thus, could be used for BSS. By using a single-layer neural network with sigmoidal output units and trained with an entropy-based *unsupervised learning* rule, they demonstrated impressive results in separating acoustic sources. Since Bell and Sejnowski's original work, many researchers have developed neural network models for BSS applications. In neuroimaging, for example, electroencephalography (EEG) and magnetoencephalography (MEG) are being used to record millisecond resolution electromagnetic signals related to brain activity. High-density EEG and MEG systems, having more than 100 sensors recording activity across the brain, pick up a mixture of electromagnetic signals originating from different populations of neurons. To analyze these signals, it is often useful to separate them into independent sources, such as separating signals from the auditory cortex from signals originating in the visual cortex. ANNs are being used to recover independent sources in EEG and MEG signals for improved brain activity analysis.

IV. NEURAL NETWORK SIMULATION

Computer simulation plays an important role in neural network research. It was not until fast and inexpensive digital computers were available that it became possible to study the behaviors of biologically detailed neural network models or large connectionist ANN networks. A variety of software-based simulators have been developed that enable researchers and students to more readily construct biological and artificial neural network models and evaluate their behavior and performance using different stimuli. Three simulation packages of note are GENESIS, NEXUS, and the Neural Network Toolbox from MATLAB, which span the spectrum of neural network modeling. Each simulator typically includes a language for specifying a network architecture, numerical and optimization tools for training and simulation, and graphical tools for visualizing results.

GENESIS (GEneralized NEural SImulation System) was developed by James Bower and David Beeman and represents one of the most comprehensive efforts to provide a simulation system to the research and academic communities for modeling biologically realistic neural networks. At the core of GENESIS is an object-oriented programming language for constructing various biophysical elements, such as compartmental elements, voltage-dependent ion channels, and spike generators. A hierarchical representation is used to represent relationships between the different elements, including the connection of individual neurons into complex networks. A message passing scheme connects objects in the hierarchy. Several biologically realistic models have been developed using GENESIS, including the models of the piriform cortex and contour perception in the visual cortex described earlier.

NEXUS (Network EXperiments Using Simulation) was developed by Paul Sajda and Leif Finkel and is designed for building and simulating highly interconnected computational maps. NEXUS provides a language for describing complex connectivity patterns between maps of neural units. Though neural units can be programmed to have arbitrary complexity, units are typically abstracted as integrate-and-fire or sigmoidal activation neurons. Researchers interested in modeling the biophysical properties of single neurons and

small networks should consider using GENESIS. Many models developed using NEXUS have focused on the visual cortex, because it is well-designed for constructing retinotopic connectivity. Specific models have included textured discrimination and segmentation in the visual cortex, perceptual grouping in areas V1 and V2, and color constancy and color contrast in area V4.

The Neural Network Toolbox for MATLAB, developed by Mathworks, is a simulator for building artificial neural networks. The toolbox runs under MATLAB, a linear algebra based mathematical simulation package. The Neural Network Toolbox supports a variety of predefined neural network architectures, including MLPs, self-organizing networks, and recurrent networks. Many classes of transfer functions are supported, and a variety of learning rules can be implemented.

In addition to software simulation, hardware implementations of neural networks have proven quite useful. Often termed "neuromorphic engineering," these neural hardware systems are based on integrated circuit technology that enable real-time simulation of highly interconnected neural networks, as well as processing that is closely coupled with sensors, e.g., visual processing coupled with a silicon retina. Carver Mead and colleagues were the early developers of neuromorphic hardware systems, which have included arrays of spiking neurons, retinomorphic chips, a silicon cochlea, and reconfigurable neural network circuits. Neuromorphic systems enable researchers to analyze the information processing in neural networks at very fast time scales, as well as providing hardware platforms for accelerating their software implementations.

V. CONCLUSION

Research into the computational capabilities and limitations of neural networks is fundamental to understanding information processing in the brain. Biologically based neural network models have proven important for integrating biophysical, physiological, and anatomical information for performing experiments (simulations) that currently are not feasible for experimentalists. Artificial neural networks have proven utility in their own right, providing tools for nonlinear signal processing and pattern classification, as well as providing a mathematical framework for gathering insight into a more abstract level of brainlike computation. The challenge for the future will be to develop a tighter integration between the computational framework and neurobiology, while at the same time incorporating molecular and genomic data for discovering the basic principles governing brain function.

See Also the Following Articles

ARTIFICIAL INTELLIGENCE • BRAIN ANATOMY AND NETWORKS • HEURISTICS • INFORMATION PROCESSING • NERVOUS SYSTEM, ORGANIZATION OF • SYNAPSES AND SYNAPTIC TRANSMISSION AND INTEGRATION

Suggested Reading

Arbib, M. (Ed.). (1995). *Handbook of Brain Theory and Neural Networks*. MIT Press, Cambridge.

Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, UK.

Bower, J. M., and Beeman, D. (1998). *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System,* 2nd ed. Springer-Verlag/TELOS, New York.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*, 2nd ed. John Wiley and Sons, Inc, New York.

Jennings, C., and Aamodt, S. (Eds.). (2000). Computational approaches to brain function, Supplement to *Nature Neurosci.* **3**.

Kandel, E., Schwartz, J., and Jessell, T. (2000). *Principles of Neural Science,* 4th ed. McGraw Hill, New York.

Koch, C., and Segev, I. (Eds.) (1998). *Methods in Neuronal Modeling: From Ions to Networks*, 2nd ed. MIT Press, Cambridge, MA.

Proceedings of Advances in Neural Information Processing Systems, 1991–present, Vol. 3 etc. MIT Press, Cambridge, MA.

Neural Transplantation

DOUGLAS KONDZIOLKA, ELIZABETH TYLER-KABARA, and CRISTIAN ACHIM
University of Pittsburgh

GLOSSARY

Huntington's disease A genetic brain disorder characterized by abnormal motor and behavioral effects.

neurons The functional cellular units of the brain.

Parkinson's disease A degenerative brain diseased characterized by rigidity, tremor, and slow movements due to lack of brain dopamine.

stem cell A primitive body cell that holds the potential to become one or more mature cells in distinct body organs.

stereotactic surgery Precise, image-guided surgery of the brain using a localization frame.

transplantation Surgical transfer of cells or tissue from one patient source (donor) to another patient (host).

This article reviews the basic principles of neurotransplantation research, discusses appropriate neurodegenerative disorders and describes current science in this field.

I. PRINCIPLES OF NEUROTRANSPLANTATION

Modern restorative neurosurgery began over 25 years ago when neurosurgeons and neurobiologists envisioned the possibility of replacing human neurons in neurodegenerative diseases like Parkinson's and Huntington's diseases. Early clinical trials were first based on a direct approach targeting the replacement of missing specific neurotransmitters rather than regenerating the damaged neuronal circuitry. More recently, with the advent of therapeutic strategies developed from experimental work with stem and progenitor cells, there is hope that the final goal of reconstructing neuronal pathways may be achieved. We can summarize the goals of this field as replacement, release, and regeneration. Dead neurons have to be replaced, the grafts have to be able to release neurotransmitters, and circuits have to be rebuilt. Of course, all of these goals can be fulfilled only if our understanding of molecular mechanisms of disease will keep up the pace with the development of new bioengineering strategies. Since the pioneering work of Björklund *et al.* almost three decades ago, much progress has been made. Rather than reviewing these advances in detail, a task already performed very well by other authors, we will outline some of the current challenges and discuss potential areas of further development. For a more indepth analysis of the clinical outcome of many ongoing clinical studies, the interested reader is strongly encouraged to read one of the most comprehensive analyses to date, the work edited by Freeman and Widner (1998), a true textbook of neurotransplantation.

The field of neural transplantation for the treatment of neurological diseases first became a potential therapeutic modality in 1979 when Björklund and Perlow showed that transplantation of dopamine containing neurons in rat striatum improved functional deficits induced by damage to the nigrostriatal pathway. Since that time, advances in neural transplantation have moved from the animal model to the human model with varying degrees of success. Animal models encompass a wide variety of disease states from degenerative diseases to trauma and stroke models. Tissue used for transplantation ranges from fetal tissue to tumor lines to stem cells. In some models, implants provide a source of neurotrophic factors. Successes in animal models have led to transplant trials in the human population. Patient trials include transplantation for Parkinson's disease, Huntington's disease, spinal cord injury, and stroke. As research in animal models progresses, transplant trials may be initiated for the treatment of multiple sclerosis, traumatic brain injury, cerebral palsy, amyotrophic lateral sclerosis (ALS), Alzheimer's disease, hereditary ataxias, and other disorders.

II. NEURAL TRANSPLANTATION: CLINICAL TRIALS AND EXPERIMENTAL MODELS

A. Background

The concept of using dissociated central nervous system (CNS) tissues for transplantation was promoted by Schmidt *et al.*, who argued that this protocol may be used for intraparenchymal implantation. The same group of investigators later showed, including detailed methodology, that mesencephalic, dopaminergic rich cell suspensions can be injected in various regions of lesioned rat brains to promote regeneration. Since these early attempts, a whole field has grown largely on the idea of using human fetal tissues to replace human degenerating neurons. Parkinson's disease (PD) has become the frontier in neurotransplantation.

The hypotheses tested and lessons learned from transplantation studies in PD have been extended to other problems, like Huntington's disease (HD) and spinal cord regeneration. Some of the therapeutic principles are similar, and the obstacles to be conquered to secure graft survival and functional recovery are identical. For this reason, our article will emphasize PD studies as the benchmark for any new experimental approaches in CNS cell transplantation, although promising advances have been reported in stroke. However, the reader must be aware that many of the scientific principles pertinent to Parkinson's disease are not the same for other disorders and that our understanding of many neurodegenerative disorders is far from complete. Among the challenges to any transplantation protocol are the source and preparation of donor tissue, surgical protocol, posttransplantation hypoxia, local toxic factors like free radicals and excitatory amino acids, deprivation of neurotrophic factors (NTF), necrosis, and neuronal apoptosis. Any factors alleviating these limitations will have a significant impact on the cell transplantation protocol in any neurodegenerative disease.

B. Clinical Trials in Parkinson's Disease: The Prototype for Cell Transplantation

Parkinson's disease is characterized primarily by the degeneration of dopaminergic neurons projecting from the substantia nigra to the corpus striatum. Parkinsonian clinical symptoms like rigidity, tremor, and postural imbalance, accompanied by the histopathologic findings of nigral distrophy and Lewy bodies, are well-described and correlate with the loss of dopamine production and release in the brain. The importance of cerebral dopamine became apparent in the late 1950s, and since then our knowledge about the neuroanatomy, physiology, and especially pharmacology of this neurotranmsitter has increased immensely. Readers interested in the topic may start by consulting the chapter in the popular neuropharmacology textbook by Cooper, Bloom, and Roth. Logically, the first therapeutic approach in diseases associated with parkinsonian symptoms, for a long time equated primarily with rigidity, was based on replacing the natural neurotransmitter with precursors like levodopa. This has proved to be a highly effective treatment and is still the first choice in PD. Nevertheless, some of its limitations derived from extended use and coupled with the significant progress in developmental neurobiology have led to clinical trials using human fetal tissues as potential replacements for the degenerating adult dopaminergic neurons. For more than two decades, experimental and clinical data have accumulated to suggest that this can be a beneficial strategy for PD patients (especially those with poor response to traditional medication) and that dopamine replacement can be achieved through cell transplantation.

An important issue to be considered in brain cell transplantation is the availability and suitability of tissues from abortions. Besides the obvious ethical concerns, the basic science and clinical studies are still in progress and many questions have yet to be answered. For example, in an extensive study of 5 tissue banks funded by the National Institutes of Health (NIH), out of approximately 1500 embryonic donors, only 7 were considered to be suitable for transplantation in PD patients. Of course, this is based on the older concept that only first trimester mesencephalic tissues from multiple donors are to be used for transplantation in one patient. Fortunately, more recent studies have shown that the use of multiple donors or first trimester tissues is not an absolute requirement to generate long-term surviving human brain grafts. Furthermore, of great promise is the current work with stem cells, which are neuronal precursors and cell lines that may soon replace the use of primary human brain cells as the main donors (Fig. 1). Finally, as will be discussed in more detail later in the article, the use of neurotrophic treatments along with viral vector technology is another exciting area in this field and may become an important approach to promote brain regeneration in PD.

C. Fetal Mesencephalic Transplants

The majority of current clinical trials are based on using human fetal mesencephalic tissues from first trimester abortion specimens. The use of tissue fragments is still popular. The clinical outcome of some of the studies discussed later depends more on the surgical technique and underlying host condition rather than the quality of the donor tissue. The autopsy studies are few, and the *in vivo* assessment of the grafts is still limited by the accuracy of positron emission tomography (PET). Nevertheless, clinical symptoms do appear to improve in some patients, often for extended periods of time, and the overall results are encouraging. The investigators agree on the need for standardized tissue processing and

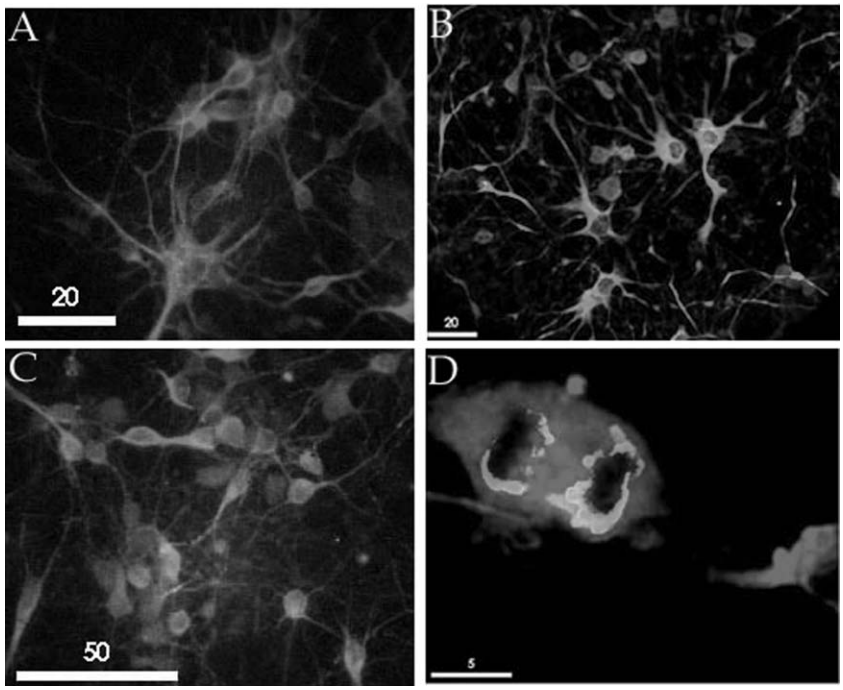


Figure 1 *In vitro*, grown in conditioned serum-free medium, fetal brain cultures differentiate into all major human brain cell populations at a physiologic ratio between various phenotypes. In these mixed cultures, (A) differentiated astrocytes (GFAP) and neurons (MAP-2) are abundant. (B) Staining for TH is abundant by day 14 *in vitro* on cells that are not GFAP negative cells, presumably neurons. (C) A subpopulation of MAP2-positive neurons express the high-affinity receptor for BDNF, TrkB. (D) Following treatment with BDNF, some of the MAP-2-positive neurons labeled with BrdU can proliferate.

implantation protocols as well as the need for more sensitive imaging techniques.

Current clinical trials explore some fundamental concepts like the need for immunosuppression and the relative contribution of the graft versus the host response in promoting dopaminergic regeneration. Since 1991, the idea that immunosuppression (discussed further later) may not be necessary has been tested in a landmark study by Henderson *et al.* In 12 advanced PD patients with grafts implanted in the caudate nucleus, long-term clinical benefits could be seen in the absence of immunosuppression. Furthermore, the authors challenged the idea of using only first trimester embryonic donors and used second trimester fetal tissues. Current discussion also centers on the number and topography of the implantation sites and the ongoing debate about the use of immunosuppressants.

In a representative study, Spencer *et al.* implanted cryopreserved first trimester tissues (one donor per host) in the caudate of advanced PD patients with 6 months of cyclosporine immunosuppression. Three out of four patients showed motor improvement and greater benefit from medication use. Similar results were reported by Freed *et al.*, who studied a larger series of patients for a longer time. Some of these patients received bilateral putaminal grafts. Also, some of the patients did not receive immunosuppresion. Interestingly, whereas the clinical outcome was largely positive, no significant differences existed between the patient groups. These studies showed that fetal transplantation was feasible and that further research should be encouraged.

The difficulty of using clinical improvement (measured on various standardized scales) or PET imaging in establishing the *direct* benefit from graft survival was overcome by Kordower *et al.* who in a detailed autopsy report showed that, in a PD patient who died of unrelated causes at 18 months postsurgery (bilateral grafts), the fetal mesencephalic tissue was surviving, was relatively well-integrated, and contained a significant number of dopaminergic neurons. Surprisingly, no host dopaminergic neuronal sprouting occurred. This appeared to be in contrast to findings from adrenal medullary grafts, which did not survive well but were accompanied by host dopaminergic sprouting. Nevertheless, the same group of investigators reported more recently that the use of abundant donor tissue (from seven donors) has resulted in significant clinical and PET improvements in one patient, even in the absence of host dopaminergic sprouting in the graft. At autopsy, the majority of the host putamen receiving the graft was positive for tyrosine hydroxylase (TH), and this may explain the postsurgery benefits.

Unilateral implantation is still used by many investigators who report significant clinical benefits when grafts are placed in the putamen with or without additional seeding in the caudate. At the same time, bilateral implantation in the putamen is gaining ground and seems to show more consistent clinical benefits. In a review of the long-term benefits of bilateral implantation, Hauser *et al.* concluded that clinical improvement appeared to be related to long-term survival and host integration of the graft. Nevertheless, the advantages of a sequential bilateral second graft have yet to be determined. It may be that ultimately the survival and long-term benefit of a graft, in addition to the quality of the initial donor cells, depend on the quality of the stereotactic surgical procedure.

From various clinical trials using fetal mesencephalic tissues, one of the most interesting observations is that the grafts appear to survive for long periods of time, even if the underlying disease affecting the host dopaminergic neurons does not change its course. These results are encouraging, to say the least. When we can overcome problems in generating reliable, high-quality, thoroughly assessed grafting material, cell transplantation could become a standard adjuvant in pharmacotherpies for PD or other novel interventions like deep-brain stimulation. PET imaging is of utmost importance in assessing these therapeutic interventions. Traditionally, dopamine production was measured by imaging of fluorodopa, but newer radiotracers, like CFT (a dopamine transporter ligand), SCH (D1 receptor), and raclopride (D2 receptor), may give a more accurate picture of the disease progress and also the survival and function of the grafted cells (Fig. 2).

D. Adrenal and Carotid Body Transplants in Parkinson's Disease

After early observations that adrenal chromaffin cells transplanted in rats can survive in the host brain, clinical trials were initiated in humans. The early results, encouraging to some extent, were questioned later, especially due to poor graft survival. One of the proposed explanations was the contamination of adrenal grafts with nonchromaffin cells. When this was corrected, improved survival and graft function

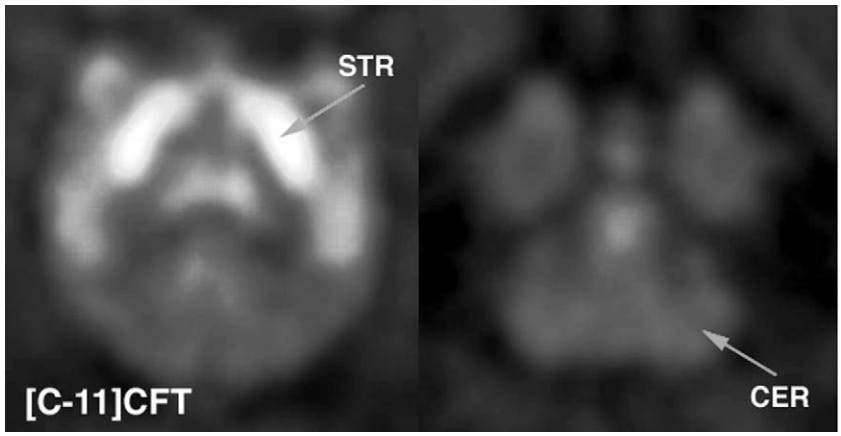


Figure 2 PET imaging of a baboon using the [C-11]CFT DAT tracer. Prominent striatal (STR) tracer activity is shown, whereas the cerebellum (CER) had no significant DAT binding.

were noted. Purified chromaffin cells can be grown and differentiated *in vitro* preimplantation or manipulated to express growth factors that may further benefit the graft *in vivo*. An intriguing but promising alternative is the use of carotid body cell aggregates that have been shown to be of significant benefit in parkinsonian rats and more recently in pilot studies in monkeys.

E. Xenografts in Parkinson's Disease

Human xenografts have been used in animal models to study principles of neurotransplantation or mechanisms of brain degeneration. More often, xenotransplantation has been tested with the goal to explore alternative sources of mesencephalic fetal tissues for grafting into the degenerating brain. Studies using porcine fetal cells in rats showed that the grafts survived and integrated successfully in the host. Currently, there are several ongoing clinical trials to assess the feasibility of porcine xenografts in PD and HD patients, and the preliminary data suggest that this may be a safe procedure. The efficacy still has to be determined, although initial reports note measurable improvement in some patients. We believe that xenotransplants will not play a significant role in neural restoration if human cells and tissues can be obtained.

F. Other Neurodegenerative Diseases

As mentioned earlier, the majority of the principles tested in PD eventually will be applied to other neurodegenerative diseseases. As another example of chronic brain degeneration, HD could potentially benefit from cell transplantation. Another instance in which graft cells could make a dramatic difference is the acute brain degeneration associated with stroke.

III. HUNTINGTON'S DISEASE

The rationale for intrastriatal grafting in HD patients in many ways is similar to that for, PD but the methodological challenges could be greater due to the more complex neurophysiological substrate. Nevertheless, extensive efforts are under way to establish a series of large-scale clinical trials, and the preliminary data are currently under evaluation. More promising are the studies analyzing the benefits of cell grafts in primate models of HD. Like in PD, much hope is invested in developing neuronal cells that can be manipulated *in vitro* to promote their *in vivo* functional integration into the degenerating striatum of HD patients.

Based on successful animal models, human trials of striatal transplantation for the treatment of Huntington's disease have been initiated. The first human transplant for Huntington's disease was performed in 1990 in Mexico using an open surgical procedure. Follow-up of this patient and a second treated by the same group showed slowed progression of the disease. Clinical studies have also been initiated in the United States. Twelve-month follow-up showed increased striatal tissue volume by MRI and improved measures of mobility and cognition. Hopefully, as human trials for the treatment of Huntington's disease progress,

large, organized, and controlled studies will emerge that will effectively evaluate this treatment modality.

A. Other Disease Models

Animal models of other neurodegenerative diseases have been developed and transplantation studies are being initiated. Huntington's disease has been modeled by injections of excitotoxic agents into the rodent striatum. Several groups have investigated transplants of fetal striatal tissue in this model with good results. Studies conducted in the nonhuman primate model also showed improvement of motor function. In these studies, transplanted neurons formed synaptic connections leading to the restoration of function. Because Huntington's disease can be identified genetically before the onset of symptoms or evidence of striatal degeneration, investigators have explored the potential of nerve growth factor to prevent striatal degeneration. Studies in the excitotoxic model and the mitochondrial dysfunction model show that grafts of fibroblasts genetically modified to secrete nerve growth factor protect striatal cells from degeneration. A murine model with Purkinje cell degeneration has been used to study hereditary ataxias. After transplantation of fetal cerebellar cells in these mice, transplanted cells migrated to the molecular layer and formed synaptic connections.

In addition to models of degenerative diseases, animal models of trauma and ischemia have been developed. Animal transplant studies in both traumatic brain and traumatic spine injuries have been initiated. Transplantation has also been evaluated in animal models of cortical and lacunar infarcts.

IV. STROKE

Transplantation of human neuronal cells is a new approach for ameliorating functional deficits caused by central nervous system (CNS) disease or injury. Several investigators have evaluated the effects of transplanted fetal tissue, rat striatum, or cellular implants into small animal stroke models. One of the best studied models of brain ischemia is the murine hippocampal stroke that results in well-defined lesions, especially in the CA1 region. The standardization of this model is invaluable to the reliable testing of various experimental protective and regenerative therapies. Among them, cell transplantation of fetal hippocampal neurons has shown that they can survive and integrate in the ischemic brain. Methodological issues are still to be resolved because subsequent studies questioned the capacity of rat fetal neocortical tissues, implanted in an infarcted area, to integrate in the surrounding host tissue. However, it has been shown that the chronic ischemic region can support graft tissue.

Because the widespread clinical use of primary human tissue is likely to be extremely limited due to the ethical and logistical difficulties inherent in obtaining large quantities of fetal neurons, much effort has been devoted to developing alternate sources of human neurons for use in transplantation. One alternate source is the Ntera 2/cl.D1 (NT2) human embryonic carcinoma-derived cell line. These cells proliferate in culture and differentiate into pure, postmitotic human neuronal cells (LBS-neurons) upon treatment with retinoic acid (RetA). Thus, NT2 precursor cells appear to function as CNS progenitor cells with the capacity to develop diverse mature neuronal phenotypes. When transplanted, these neuronal cells survive, extend processes, express neurotransmitters, form functional synapses, and integrate with the host. The final product is a $>95\%$ pure population of human neuronal cells that appear virtually indistinguishable from terminally differentiated postmitotic neurons. The cells are capable of differentiation to express different neuronal markers characteristic of mature neurons, including all three neurofilament proteins (NFL, NFM, and NFH), microtubule-associated protein 2 (MAP2), the somal–dendritic protein, and τ, the axonal protein. Their neuronal phenotype makes them a promising candidate for replacement in CNS disorders as a virtually unlimited supply of pure, postmitotic, terminally differentiated, human neuronal cells.

In support of different mechanisms for efficacy, animal transplantation studies of LBS-neurons revealed graft survival, a mature neuronal phenotype, and integration into the host brain *in vivo*.

In patients disabled by stroke, the concept of restoring function by transplanting human neuronal cells into the brain is innovative. Research in a rat model of transient focal cerebral ischemia demonstrated that transplantation of fetal tissue restored both cognitive and motor functions. Sanberg, Borlongan, and colleagues were the first to show that transplants of LBS-neurons could also reverse the deficits caused by stroke. The preclinical studies of LBS-neurons were carried out in a model of transient focal, rather than global, ischemia in order to maximize the chances of

functional recovery. Animals that received transplants of LBS-neurons (and cyclosporine treatment) showed amelioration of ischemia-induced behavioral deficits throughout the 6-month observation period. They demonstrated complete recovery in the passive avoidance test, as well as normalization of motor function in the elevated body swing test. In comparison, control groups receiving transplants of rat fetal cerebellar cells, medium alone, or cyclosporine failed to show significant behavioral improvement. A second study that evaluated response in comparison to the number of cells transplanted, confirmed the efficacy of transplanted LBS-neurons in reversing the behavioral deficits resulting from transient ischemia in an MCA occlusion rat model.

The initial objectives of the first clinical study performed at the University of Pittsburgh were to demonstrate the safety and feasibility of the neuronal-cell implantation procedure. These goals were met, in that no adverse events related to the implantation have occurred in at least 36 months of follow-up in 12 patients (Fig. 3). The adverse events that did occur in these patients were thought to be unrelated to the implantation of the neuronal cells and can be considered typical of a population with known cardiovascular disease and advanced age. This study was also intended to provide some information on the efficacy of neuronal-cell implantation in improving stroke-related neurologic deficits. In both treatment groups, mean NIHSS (National Institutes of Health Stroke Scale) total scores decreased and mean ESS (European Stroke Scale) total scores increased; both changes indicated improvement. For the ESS, the increases tended to be larger in the group of four patients receiving 6 million cells, both in the total scores and in the composite motor subscale scores. Both the Barthel Index and the SF-36 scores decreased in the group receiving 2 million cells and increased in the group receiving 6 million cells. All outcomes measurements were consistent in identifying a trend toward improved scores in the group of patients who received 6 million neuronal cells. The PET scan results also provide a suggestion of efficacy, in that increased activity at the area of the stroke was seen in 6 patients.

The neuronal cells could improve neurologic function through a number of different mechanisms. These include the provision of neurotrophic support (acting as local pumps to support cell function), provision of neurotransmitters, reestablishment of local interneuronal connections, cell differentiation and integration, and improvement of regional oxygen tension. Transplanted cells also may act to limit the reactive glial response and to limit retrograde degeneration, although this may be less feasible in a chronic injury. We believe that axonal reconnections through the grafted cells (serving as a ''bridge'') over large distances are less likely, although this phenomenon has been observed in spinal cord injury models. Phase 2 dose–response trials in patients are ongoing to evaluate further the role of neurotransplantation for patients with chronic motor deficits caused by basal ganglia region infarction.

V. ANIMAL MODELS FOR TRANSPLANTATION

A. Animal Models of Brain Degeneration in Parkinson's Disease

Among the chronic brain degenerative diseases, PD has the best characterized and most consistent models. Although the underlying mechanism of disease in these models may be significantly different from the human condition, they reproduce rather closely the pathology in the final stages of dopaminergic degeneration in PD and constitute a crucial basis to test novel therapies. Due to its well-defined neuroanatomical distribution, the nigrostriatal dopaminergic pathway can be lesioned or severed at various locations and the benefit of various cell grafting methods can be studied. Because this approach is more labor intensive and sometimes postsurgical complications may interfere with the experimental results, most investigators prefer to use chemical lesioning. The best characterized and most extensively used models are the rat 6-OHDA and the primate and murine (1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine) (MPTP) models. Both are based on the induction of a cytotoxic lesion in the dopaminergic neuronal population of the substantia nigra, resulting in massive or complete loss of striatal projections. The neurologic deficits can then be measured in various functional assays and now through PET imaging for dopamine connectivity. The autopsy findings are remarkable for many similarities with the human PD brain.

B. The Rat 6-OHDA Model for Transplantation

The rat 6-OHDA model is characterized by massive loss of dopaminergic neurons accompanied by obvious motor dysfunction that can be exacerbated by amphetamines. The effects of various treatments on the motor asymmetry can be measured over time and correlated with autopsy findings. For example, in a

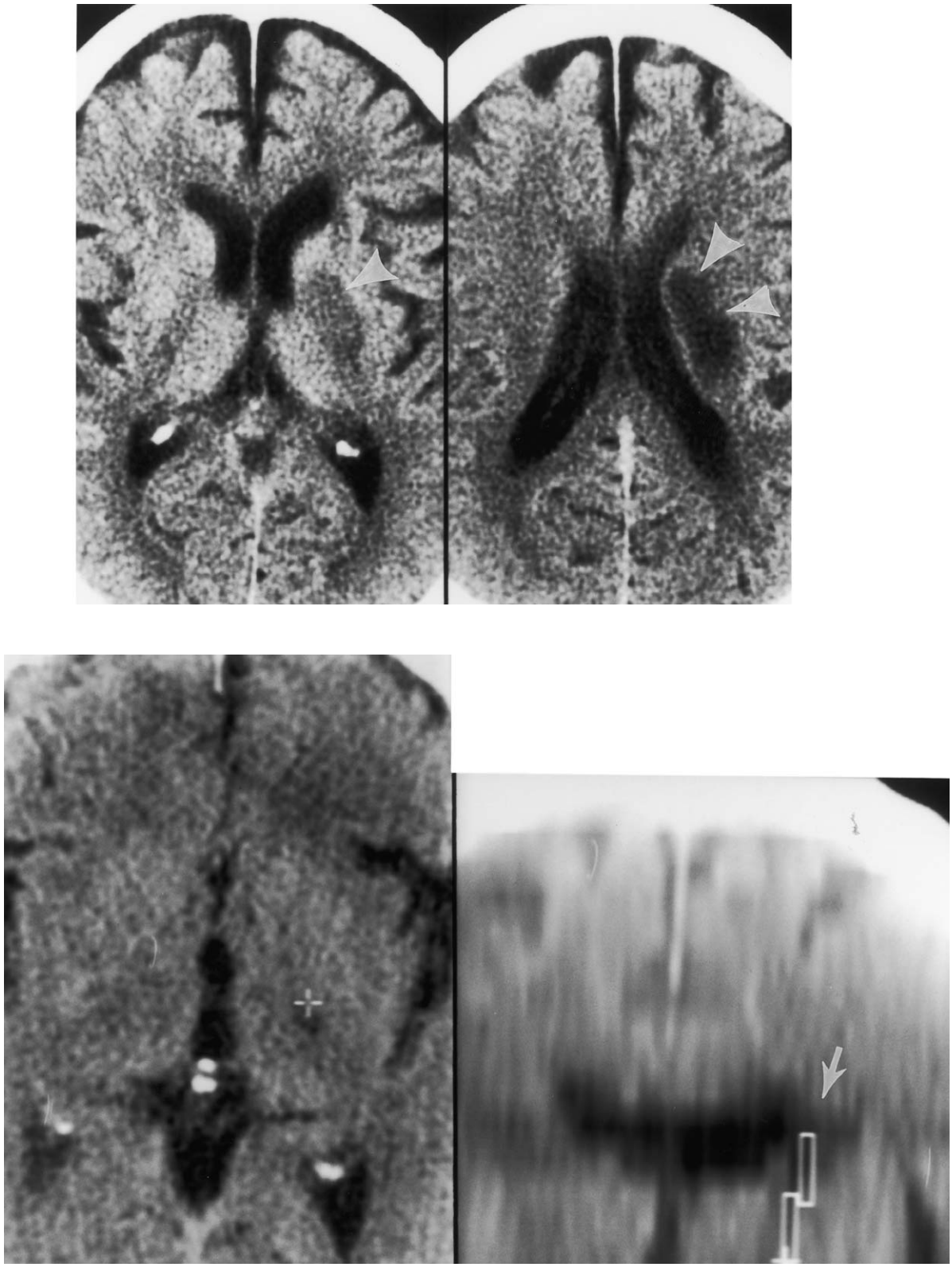


Figure 3 CT scans showing a cerebral infarction of the left basal ganglia region (top). The left scan shows the inferior target site for cell placement, and the right image shows a coronal view of the planned neuronal cell deposits (bottom).

detailed study of human fetal dopamine neurons implanted in parkinsonian rats, Clarke *et al.* showed that the functional effects of the graft were evident only when, as confirmed by serial autopsy histologic studies, they developed dopaminergic synaptic contacts with the host. This occurred at a rather late time points (4–5 months postgrafting), emphasizing again the fact that long-term survival is not sufficient if it is not accompanied by functional connectivity. These results were confirmed by other groups, who extended the studies for up to 2 years posttransplantation. In general, the grafts continue to differentiate and are characterized by abundant TH-positive neurites with many synaptic contacts. The host astroglial reactivity surrounding the graft can be detected and does not appear to be affected by host treatment.

To improve graft survival and integration, Nikkhah *et al.* developed a microtransplantation technique consisting of multiple small deposits of fewer cells rather than a single large-volume graft. This approach resulted in an increased functional benefit as measured by the significant reduction of rotational asymmetry. Another improvement in the grafting protocol in parkinsonian rats was made by Constantini *et al.*, who showed that cotransplantation of striatal tissues enhanced the benefit of the nigral dopaminergic implants, probably by offering some trophic support. This is an interesting concept and will be further discussed in this article (under trophic factors). Finally, Winkler *et al.* showed that the efficacy of nigral–striatal cotransplants can be further increased when the striatal grafts are implanted in the host nigra, with the mechanism proposed being a donor-derived, GABA-mediated, inhibitory activity on the host projection neurons.

C. The MPTP Model

The story of the MPTP model based on a series of intriguing observations in several heroin addicts with parkinsonian symptoms is well-known to all readers interested in PD. The clinical progress of these patients was followed and reviewed numerous times over the years. Burns *et al.* have reported a primate model of selective destruction of dopaminergic neurons in the pars compacta of substantia nigra (the same region affected in PD patients) following the administration of MPTP, the toxin that caused the disease in the drug users. Since then, numerous attempts have been made to develop similar models in other species, including a mouse version that shows promise, if attention is paid to the variability in the strain-related host susceptibility.

Due to its potential to mimic the human disease for testing early diagnostic methods (e.g., PET imaging), pharmacologic interventions, surgical procedures including cell grafting, and neuropathologic analysis, the primate MPTP model is one of the most powerful tools available to the PD investigators. Still, there is much to be done in terms of consistency of effects following the MPTP lesion. The diversity of protocols used by various groups, the individual response of the animals even within the same species and age group, the often subtle clinical changes, and sometimes the spontaneous recovery continue to limit our ability to measure accurately the changes made in the course of the disease following various treatments. Fortunately, increased sensitivity of neuroradiologic diagnosis to dopamine loss and more sophisticated histologic studies at autopsy could soon make this model the standard for preclinical trials in PD.

A systematic effort to improve the MPTP model was made by Bankiewicz and colleagues who demonstrated that unilateral injection in the internal carotid artery induced specific destruction of the dopaminergic nigrostriatal pathway, with sparing of other dopaminergic neurons and minimal or no injury on the contralateral side. The specific neuroanatomical injury was mirrored by limited, unilateral neurologic deficits that did not impair animal grooming and survival. This experimental development opened the door for long-term preclinical studies in nonhuman parkinsonian primates. For example, the transient efficacy of adrenal medullary transplants in humans was similar to observations in hemiparkinsonian monkeys. In similar fashion, the fundamental principles of fetal mesencephalic grafting can be tested in this model. For example, an interesting observation was that in hemiparkinsonian monkeys, the functional benefit postsurgery may be due to the host dopaminergic sprouting rather than to the donor tissue, raising the intriguing possibility that the graft itself may not need to contain dopaminergic neurons to promote a functional recovery in MPTP lesions.

D. Graft Survival: Transplantation Strategies

Significant progress has been made in the transplantation protocol itself, but we have much to learn about the intrinsic potential of various donor brain tissue

preparations and the *in vivo* cues to determine the fate of the neuroglial graft. In general, it is still believed that first trimester human fetal mesencephalic tissue is the best source for harvesting and grafting of dopaminergic neurons. This ''window'' can be somewhat extended (comparable numbers of surviving TH-positive grafted neurons) if cell suspensions are used instead of solid grafts. This finding was confirmed by different groups in various animal models, and most data suggest that, if survival of TH-positive cells is used as the main criterion of assessing the grafting efficacy, the younger embryonic tissues are superior to older fetal donors. Nevertheless, as discussed previously, several reports challenge this concept.

Although it may sound mostly of academic interest because all PD patients are adults, the influence of host age on graft survival merits discussion for identifying potential *in vivo* cues that are responsible for the integration of the implanted cells. One of the first observations made was that, in the neocortical transplants in older rats, both types of grafts, with high (early embryonic) or low (late embryonic) growth potential, survived and integrated to similar extents versus clear *in vivo* differences when younger host animals were used. Further studies have confirmed that migration of donor cells and integration of the grafts decrease in older hosts. On the other hand, studies analyzing host dopaminergic sprouting in the graft, in general very limited, find that it is not significantly different in young versus adult rats. Furthermore, these results were confirmed by functional studies showing that, even if the cytoarchitecture of the grafts in hosts of various ages may be different, the benefits are comparable. It is interesting to speculate on the clinical implications of these observations because several authors agree that the host responses are most critical to the functional benefits of fetal grafting.

VI. *IN VITRO* PROCESSING OF FUTURE GRAFTS

As mentioned earlier, the first limiting factor in the success of fetal grafting is the availability of suitable tissues. Banking of cryopreserved tissues is still employed and considered to be feasible by several authors. Most investigators now agree that, ideally, cryopreservation could be avoided, and optimized *in vitro* processing of fresh tissues for long-term survival and growth should become the standard. The University of Pittsburgh stroke transplantation trial was the first to use a cryopreserved cell line in patients. Another approach is to maintain fetal cells as aggregates in long-term suspension cultures that may even allow their expansion. However, this approach may not be feasible for processed neuronal cell lines.

There are benefits to *in vitro* processing of brain cells before implantation. First, processing may select for a viable, better characterized neuronal donor population. Second, it can help identify the donor glial–neuronal interactions that are critical *in vivo*, potentially protecting the graft from detrimental host reactions. Third, it can offer a setting for trophic and genetic manipulations (discussed later) that will enhance *in vivo* survival and integration. Finally, neuronal apoptosis in fetal grafts is well-described and appears to be significant during the first 10–15 days postimplantation. The mechanisms of neuronal death are predominantly mediated by caspases, and inhibitors of this process may have important benefits in improving the viability of the grafts.

A. Implantation Techniques

The majority of the TH-positive neurons (around 90%) within fetal grafts appear to die within the first week after implantation. Besides the pharmacologic trophic and protective treatments discussed later, changes in implantation protocols may have an important impact on the survival of the graft. One approach is to include a biologic adjuvant at the time of implantation, often as a cograft of peripheral nerve, that may offer continuous trophic support, promoting the survival and differentiation of the donor dopaminergic cells. Over the past decade, this approach was tested several times and the results, though mixed, are generally positive. Another bioengineering protocol, proposed by Brecknell *et al.*, is based on providing a ''bridge'' for the implanted dopaminergic neurons to grow into the striatum by using cells transformed to secrete growth factors. At present, stereotactic techniques are used to deposit cells at distances such that the ''sphere of influence'' of each graft might determine the distance between individual graft locations.

The most direct strategies to increase transplant integration in the host dopaminergic pathways is the use of cografts of mesencephalic and striatal donor cells or simultaneous mesencephalic grafts in the nigra and striatum of the host. Both of these interesting approaches are currently being explored in clinical trials in PD patients. Finally, the idea of using the nigra

(versus striatum) as a site of dopaminergic fetal cell grafting has gained momentum again, and studies propose it as a feasible alternative in PD. However, intranigral grafts alone fail to innervate the striatum without some additional therapy.

B. Host Immune Response

Another potentially serious limiting factor in graft survival, despite optimal *in vitro* tissue processing and efficient implantation protocol, is host rejection of the graft. Whereas the brain is still considered to be an organ with a limited immue response (historically called "immunoprivileged"), rejections of grafted tissues can occur through the classic cell-mediated immune response. The immune reactions to neural grafts have been studied for a long time, but the necessity of immune suppression is still debated. A review by Widner emphasizes the complexity of the immune response in the brain and also questions the impact of immunosupressive therapies on allo- or even xenograft survival. Furthermore, it should be noted that another host reaction, the degeneration of the graft mediated by brain macrophages that does not involve the presence of lymphocytes, can also be considered a form of rejection. The role played by reactive microglia in graft survival is the object of intense dispute. Finally, following an interesting observation in rats implanted with fetal dopamine cells, Hudson *et al*. suggested that host immunocompetence to the allogeneic graft is necessary but not sufficient to cause rejection.

Early studies in rats showed that immunosupression with cyclosporin A (CsA) was critical for graft survival and function of dopaminergic human xenografts. The benefit of CsA was confirmed in other animal xenograft models, but graft degeneration in nontreated animals was not evident. Furthermore, the lack of a detectable systemic immune response in primates with embryonic brain grafts raised the possibility that aggressive immune suppression may not be necessary and that local immune modulation or evasion strategies can be developed. If one assumes that the immune response is exclusively detrimental, which is not clear yet, methods proposed to modulate it include the cografting of Sertoli cells that seem to have both a trophic effect and a surprising immunosuppressive activity on the host response. Another proposed approach is to use specific anti-T-cell antibodies capable of inducing immune tolerance or, as suggested in an interesting report, a short course of treatment with an anti-IL-2 receptor antibody (CD25), which may be as effective as long-term CsA.

C. Vascularization and Blood–Brain Barrier Development

The next challenge in the survival of a healthy graft is the development of functional vascularization and a blood–brain barrier (BBB). The origin of angiogenesis in fetal brain grafts is still disputed. There is evidence that donor-derived vasculature already contained in the implanted tissue fragments can survive and rapidly develop connectivity with the host. When cell suspensions were used for allo- and xenografting, the BBB also appeared to form relatively soon and did not seem to be affected by concomitant treatments with CsA. These results were further confirmed, and additional findings suggest that early vascularization is determined intrinsically by the graft, possibly through locally produced growth factors (discussed later in the article).

Contrary to these reports, Geny *et al*. found that angiogenesis in the human fetal grafts is a much slower process, including a delayed redifferentiation of the endothelial cells present in the graft at the time of implantation, a phenomenon possibly dependent on the maturation of the fetal neuroglial tissue. Adding another twist to this debate, Pennel and Streit proposed that their studies with rat fetal allografts demonstrated early vascularization, but this was directly dependent on graft colonization by donor-derived microglia. In another similar study, Rostaing-Rigattieri *et al*. showed that not only microglia but also host-reactive astroglia and endothelial cells infiltrated the graft and participated in the reorganization of its cytoarchitecture. Finally, pilot studies in our lab showed that, in human fetal xenografts, the early development of a BBB, facilitated by *in vitro* manipulations before implantation, is crucial for long-term donor neuronal survival and differentiation.

D. Intraparenchymal Reaction to the Graft

It has become increasingly clear that the host parenchymal reaction to the fetal graft is possibly the decisive factor in determining its survival, integration, and functional benefit. Since the early studies with

murine models of neurotransplantation, the investigators have suspected that part of the functional benefit may be due to the host reaction to the graft or just to the various surgical procedures (e.g., cavitation). Not surprisingly, prime candidates as mediators of these effects are growth and trophic factors or other molecules produced by the host-reactive neuroglial millieu. The effects of the host-guided migration of donor-derived neurons to selective targets have been convincingly demonstrated. Nonetheless, the opposite may be as important clinically: host sprouting in a dopaminergic graft may be specific and extensive. The most intriguing observations were made in the MPTP model of PD. Bankiewicz and colleagues have demonstrated, in a long-term study of hemiparkinsonian monkeys, that cavitation alone leads to clinical improvement similar to that of autografting dopamine-secreting adrenal medullary chromaffin cells. Furthermore, similar results were obtained following implantation of fetal amnion. Both studies suggest strongly that the regenerative capacity of the host brain plays a primary role in functional recovery, at least in this primate model of PD.

Among the most prominent but least understood host reactions to the graft is the microglial response. It is well-known that reactive microglia are important immune mediators as antigen presenting cells infiltrating the grafted tissues. Still, several reports suggest that their presence and less defined activities are critical for the survival of the graft and host brain regeneration. Increasing evidence suggests that microglia are associated not only with pathological but also with normal functions in the brain. Microglia react to injury, usually by responding to and producing NTF. Furthermore, other growth factors produced by microglia, like HGF, were shown to have neurotrophic activities in the CNS. Finally, in a direct experiment, activated macrophages transplanted into a CNS injury site were crucial in triggering a regenerative cascade. *In vitro*, microglia promote the survival and differentiation of TH-positive neurons, probably through production of NTF or factors like plasminogen that enhance neurite growth. *In vivo*, nigrostriatal dopaminergic neurons develop neurites around the wound or graft, and this process appears to correlate with the neurotrophic factor production by activated macrophages and microglia.

From this brief summary of the supportive and guidance role played by the surrounding host neuroglial environment, it has become apparent that neurotrophic support is crucial in promoting graft survival and differentiation in the host.

VII. GRAFTING SUPPORTIVE TREATMENTS

In our opinion, the most important factor deciding the fate of the future graft is the trophic support offered by the intrinsic milieu, the host environment, or administered *in vivo* postgrafting. Neurotrophic factors have been the focus of intense studies discussing them as therapeutic agents in neurodegenerative diseases. Future testing in primate models will be necessary. In the field of neurotransplantation, the studies proposing NTF as *in vitro* and *in vivo* adjuvants to the graft have been significantly more successful. Overwhelming evidence from developmental neuroscience experiments demonstrated that growth factors are crucial in the differentiation of neural progenitor cells or, in general, less differentiated cells that are abundant in the fetal grafts. Among these factors, epidermal and fibroblast growth factors were most important in promoting the mitogenic capacity of neuroglial progenitors. Whereas work by Gage and colleagues established FGF as a principal mitogenic factor in the developing neuronal population, other growth factors may participate in the expansion and differentiation of the neuronal precursor cell population, as demonstrated by *in vitro* and *in vivo* experiments with insulin-like growth factor and hepatocyte growth factor. Interestingly, the latter seems to function as a primer for other trophic signals.

Hepatocyte growth factor is only one of the newer trophins for the CNS. Today, the family of growth and trophic factors proposed to affect the survival and development of neuroprogenitor cells is probably the largest in this ever expanding field. Factors like cytokines, once considered to be exclusively neurotoxic (e.g., TNF), are now studied for neurotrophic properties. Among them, leukemia inhibitory factor (LIF) and ciliary neurotrophic factor (CNTF), in addition to more traditional growth factors like platelet-derived growth factor (PDGF), are considered to be potent promoters of neuroprogenitor cell proliferation and eventually differentiation. *In vivo*, CNTF produced by reactive astrocytes can prevent the degeneration of dopaminergic neurons in adult rats. Astrocytes and endothelial cells surrounding or infiltrating the transplant are susceptible to the effects of PDGF, which may control the survival of graft through neovascularization.

In addition to the proliferative support offered to neuroprecursor cells, NTFs are key players in their differentiation to the mature phenotype. This function may be critical for the graft integration in the host environment. It was shown that, at least when cell lines

were used (e.g., PC12), *in vitro* pretreatment with NGF is crucial for the *in vivo* phenotype of the graft: differentiated (NTF-treated cells) versus nondifferentiated, tumorlike (nontreated cells). The *in vivo* effects of NGF become even more obvious when fetal grafts are continuously treated through *in situ* injection or when regeneration is promoted through gene delivery using transformed cells.

A. Brain-Derived Neurotrophic Factor

Brain-derived neurotrophic factor (BDNF), another member of the neurotrophin family (which includes NGF, NT-3, and NT-4/5), has been shown to have great potency in modulating the growth and survival of dopaminergic cells and their precursors. It is now widely accepted that the pluripotent BDNF and its high-affinity receptor trkB are widely distributed in both the developing and mature nervous system. Our increasing understanding of neurotrophin binding to their receptors, signal transduction following trk, and p75 dimerization and activation has lead to a series of exciting developments in designing experimental models to test novel trophic treatments. In addition to dopaminergic cells, BDNF was found to be potent on cholinergic and glutamatergic motor and sensory neurons, in both the central and peripheral nervous systems.

Among the most exciting new discoveries in the field of neurotrophic factor research is the ability of BDNF to be transported in an anterograde fashion. For example, it was shown that, in development, BDNF produced by dorsal interneurons stimulates the proliferation and differentiation of motor neuron progenitors after anterograde transport. Another piece of evidence comes from the studies by Kokaia *et al.*, who have shown that BDNF levels increase significantly in a rat model of focal ischemia but then decrease rapidly, suggesting anterograde transport. Finally, additional supporting data for the anterograde transport of BDNF in the adult CNS were reported by Conner *et al.*, who demonstrated that its distribution parallels axonal flow, including storage in terminals within the target.

One of the most important implications of anterograde transport of BDNF is probably its participation in synaptic transmission. This theory is supported by data suggesting that enhancement of long-term transmission is manifested through synapse consolidation rather than neuronal growth. Also, in an *in vitro* explant model using hippocampal slice cultures, Frerking *et al.*, showed that BDNF may enhance transmission in CA1 neurons by decreasing the postsynaptic inhibition through a presynaptic mechanism. There is evidence that, at the presynaptic level, BDNF potentiation is mediated by cAMP. In addition, BDNF is reported to also be able to mediate agrin-induced postsynaptic differentiation. All of these functions obviously can be critical steps in the establishment of a functional cell graft.

In spinal cord injuries, neurotrophin treatments are proposed to present significant clinical benefits. For example, BDNF infused at the site of spinal cord injury in rats showed a positive but transient effect on local reflexes. The most dramatic impact of BDNF occurred in fully transected spinal cords. When these chronic infusions were stopped, the behavioral effects disappeared. BDNF was also shown to stimulate sprouting of cholinergic fibers at the injury site, but did not affect serotonergic fibers or total axon density. In an effort to promote directional regeneration, cells transformed to secrete BDNF were grafted in trails in the transected spinal cord, and the results showed a significant positive effect on axons from trkB expressing neurons. In another study by Broude *et al.*, it was shown that in spinal cord transplants the addition of BDNF increased axonal outgrowth of axotomized neurons.

BDNF has been proposed to have an autocrine effect on dopaminergic neurons that express abundant trkB, and if these results are confirmed in humans, new hypotheses may be formulated about the mechanisms of disease in parkinsonism. Supporting this theory, several studies showed reduced BDNF protein in the substantia nigra of PD patients. In this context, BDNF treatments (similar to GDNF, discussed later) pre- and posttransplantation may serve a dual purpose: stimulate the growth of a functional dopaminergic population within the graft and promote the regeneration of dopaminergic pathways in PD.

Interestingly, some reports suggested that BDNF enhanced the function rather than survival of the grafts enriched with dopamine cells. *In vitro*, BDNF can protect dopaminergic neurons from hydroxydopamine toxicity. *In vivo*, similar protective functions were observed in rats with BDNF producing grafts, after being challenged with the active metabolite of the dopaminergic toxin MPTP. The regenerative capacity of BDNF on dopaminergic projections was shown to be both direct and indirect, mediated through improved fetal grafts. Still, only a subpopulation of nigral dopaminergic cells may be susceptible to these effects, depending on their capacity to express the high-affinity receptor trkB. Furthermore, there are some concerns

about the effects of long-term *in situ* delivery of BDNF *in vivo*, at least in the rat striatum. Currently, it still appears that the most consistently positive use of BDNF in cell transplantation is to promote dopaminergic differentiation preimplantation. When rat and human nigral fetal cell aggregates were treated with BDNF, the number of TH-positive neurons increased significantly. These effects were further enhanced when GDNF was used in combination with BDNF.

B. Glial-Derived Neurotrophic Factor

In general, GDNF has similar or even enhanced trophic functions on dopaminergic neurons and their precursors. *In vivo*, GDNF was also shown to be relatively potent in animal models of dopaminergic protection or regeneration, but human clinical trials have not been so encouraging. Nonetheless, its potential use in PD has made GDNF the prime candidate for NTF treatments in association with fetal mesencephalic transplants of dopaminergic neurons.

GDNF, a related member of the TGF-β family, was first identified and characterized in 1993 by Lin *et al.* From the beginning, it was evident that GDNF has a potent but specific–selective activity on dopaminergic neurons, inducing their differentiation in the absence of overt neuroglial proliferation. Shortly after, *in vivo* studies demonstrated that GDNF injected into the substantia nigra produced a significant decrease in the motor deficits associated with 6-OHDA lesioning in rats. When GDNF was injected in developing or mature mesencephalic grafts, *in oculo*, its primary effect was to promote dopaminergic neuritic growth rather than survival of TH-positive cells. Interestingly, administration of GDNF in a murine MPTP parkinsonian model had both protective and regenerative effects. The *in vivo* protective effects were also demonstrated in a rat nigral axotomy model, suggesting again that GDNF has a high specificity for dopaminergic neurons.

Further *in vitro* studies showed that a mixture of slow-release GDNF, fibrin glue, and fetal mesencephalic neuroglial cells resulted in a significant increase in the number of TH-positive cells and neuritic density. Another *in vitro* study indicated that GDNF may offer protection from continuous cell death after the removal of toxins and even stimulate dopaminergic fiber regrowth. When another neurotoxin, quinolinic acid, was used, GDNF showed selective protection of dopaminergic neurons against excitotoxicity.

In vivo studies using fetal mesencephalic grafts showed that injections of GDNF in the vicinity of the rat brain implant resulted in significantly increased survival and growth of TH-positive cells accompanied by marked functional improvement. The improved survival and differentiation of dopaminergic fetal grafts treated with GDNF pre- or postimplantation have been independently confirmed in numerous studies focused on the therapeutic benefits in parkinsonian models. A more recently identified member of this family of growth factors, neurturin, was found to be similarly potent in preventing dopaminergic cell death but lacked the support for TH-positive neuritic growth associated with GDNF treatments. Finally, in clinical studies, two patients with PD who received fetal dopaminergic implants pretreated *in vitro* with GDNF showed increased graft survival. The benefits of GDNF treatments in PD patients appear to be restricted to association with fetal grafts. When GDNF was injected into the cerebroventricular system of a PD patient that did not receive a graft, the results appeared detrimental. Further research is continuing on this approach.

C. Other Protective and Trophic Factors

In addition to the beneficial effects of neurotrophic factor treatments, the ganglioside GM1 has been shown to enhance effects even at minimal concentrations of BDNF. The lazaroids are another intriguing class of compounds that demonstrated a strong trophic effect on promoting the survival of embryonic mesencephalic tissues and their development *in vivo*. Interestingly, the lazaroids promoted *in vivo* survival of dopaminergic neurons but did not increase target striatal innervation. These agents have now been incorporated into clinical trials.

Mechanisms of grafted cell death include excitotoxicity and apoptosis. Among the excitotoxic inhibitors, the calcium channel–NMDA receptor antagonist MK-801 is one of the most studied. *In vivo*, this compound was not able to enhance dopaminergic neuronal survival in the graft, suggesting that cell death in the grafts may not be due to excitotoxicity. The other mechanism of cell death in transplants, apoptosis, is under intense scrutiny because caspase inhibitors seem to reduce neuronal death in the grafts. Furthermore, the combination of pretreatment with a caspase inhibitor and a lazaroid may have a significantly higher positive effect on transplanted dopaminergic neurons.

Finally, debates in neural cell grafting gravitate around the potentially beneficial effects of new generation immunosupressive drugs that have been shown to be less toxic to the brain. A representative member of this family is the FK-506 drug. This compound, together with CsA and rapamycin (approved for clinical trials), binds to receptors called immunophilins. Many of the immunophilin ligands have be shown to possess neurotrophic activities. This field is under intense scrutiny, especially since immunophilins have been proven to be abundant in the brain. Among the newer members of the immunophilin ligand family, the non-immunosuppressive drug GPI-1046 shows much promise in promoting regenerative neuritic growth from surviving neurons in various CNS lesions. Furthermore, the immunophilin ligand V-10,367 has been shown to specifically increase the growth of dopaminergic neurons (mostly neurite branching) and protect against MPTP lesioning, in a manner superior to FK-506.

D. Nonneuronal Primary Human Cells

On the basis of the observations discussed previously regarding the importance of the trophic host response in promoting regeneration posttransplantation, often independent of graft survival or function, some investigators have proposed alternate sources of donor cells that will primarily serve as promoters of host regeneration. Among the nonneuronal brain cells, astrocytes are prime candidates. A study of neuroglial cell grafting in a murine model of lesion-induced memory deficits showed that astrocytic grafts induced significantly more improvement compared to the cholinergic neuronal grafts. Another interesting method is based on transplanting activated macrophages to promote CNS regeneration. One of the rationales is that grafted macrophages may compensate for the failure of host macrophages to provide trophic support or potentially overcome their inhibitory activity. Finally, one of the most intriguing sources of donor cells is the bone marrow stroma. Azizi and colleagues have shown that human marrow stromal cells grafted in the rat brain can survive, spread, and differentiate into a neuroglial phenotype in the absence of a host immune response or any signs of rejection. An important comment is that the donor cell mixture may include a significant population of nonhematopoietic tissue precursors. If any of these methods proves to be consistent among various laboratories and disease models, they could become a powerful strategy for promoting "natural" regeneration or an important platform for delivering gene therapy.

VIII. STEM CELLS AND NEURONAL PRECURSORS

The advances in our ability to manipulate embryonic stem cells have opened a whole new spectrum of potential therapies in tissue regeneration in general and the brain in particular. It has also reignited interest in brain stem cells and neuroprecursors that show many analogies with the considerably better characterized hematopoietic system. The definition of brain stem cells versus neural precursors is increasingly more precise and tends to better delineate the distinction between totipotent, multipotent, and lineage committed. Excellent reviews about the immense potential of using stem cells in regeneration have been published. The reports that stem cells are present not only in the developing brain but can also be identified in adults have further contributed to the excitement about new possibilities to promote CNS regeneration.

Many of the early studies have often reported successful harvesting of progenitor cells from various regions of the brain (the subventricular zone being the most popular but not singular) that can be expanded *in vitro* and then differentiate *in vivo* into functional neurons. The challenge is to still better identify the neural progenitor cells and understand the mechanisms of growth and differentiation from embryonic stem cells. An important characteristic of neuroepithelial stem cells, described more than a decade ago and still widely used, is expression of the intermediate filament protein, nestin. More recently, new markers of neuronal progenitor cells, conserved in their evolution and useful for early lineage selection, include Musashi-1 and an epitope recognized by the 2F7 monoclonal antibody. The mechanisms of progenitor cell maturation appear to be similar in various brain regions and seem to depend primarily on a cascade of signals mediated by specific combinations of growth factors and molecules (like noggin) that can modulate their activity.

An understanding of the dynamics of gene expression and specific protein production in the development of the neural lineage from stem cells is critical for designing new isolation and purification methods to be used *in vitro* for preimplantation neuronal enrichment of future grafts. In addition to the traditional marker nestin, Li and colleagues reported that the transcription factors Sox1 and Sox2 could identify progenitor

cells restricted to the neural lineage with higher specificity. Furthermore, expression of a neuron-specific promoter (like Hu and TuJ1/β-III tubulin) offers another opportunity to intervene through gene engineering in the development of neural progenitor cells and select for this cell phenotype.

Regardless of their origin, fetal or adult, neural precursor cells show much promise in brain repair. They can survive and differentiate in the host lesioned brain, although it seems that their predominant differentiation *in vivo*, posttranplantation is along the astrocytic lineage. Not surprisingly, the migration and differentiation of the grafted precursor cells are significantly influenced by local cues that may even overcome the *in vitro* manipulations. Nonetheless, the potential for *in vivo* survival of grafted neural precursor cells can be fully exploited when they are used as platforms for gene delivery or engineered to modulate the neurotrophic factor environment in the host brain.

A. Neuronal Cell Lines

The use of "primitive" neuroglial cell lines that preserve their mitotic potential *in vitro* but will differentiate *in vivo* was proposed for a long time to be a feasible alternative to primary cell cultures. Whereas the PC12 line is well-established as the workhorse of a plethora of *in vitro* experiments studying neurotoxic or neurotrophic mechanisms in the CNS, its capacity to differentiate and integrate *in vivo* in the host brain is limited at best. Because the usefulness of this cell line in humans may be seriously hampered by its uncontrolled capacity to proliferate in a nondifferentiated state, significant efforts were made to identify new human neuronal cell lines that will overcome this obstacle. One of the candidates that, under certain conditions, can differentiate into a "mature neuronal phenotype" is the human cortical neuronal cell line HCN-1, which has many of the characteristics of immature neuroepithelial cells but, interestingly, does not respond to traditional growth factors for CNS cells, like βFGF. Another example is the human neuroblastoma cell line SH-SY5Y, which can synthesize dopamine *in vitro* and differentiate *in vivo*. Still, a puzzling observation seen, as discussed earlier, with primary cell grafts too is that in parkinsonian animal models the functional benefit postgrafting may not be due to dopamine release by the implanted cells. The use of immortalized primary cell cultures has become a popular approach. Experiments with SV40 tumor antigen immortalized dopaminergic cells showed that they can be used for *in vitro* analysis (e.g., the SN4741 line) of the effects of BDNF on mesencephalic cultures. *In vivo*, similar immortalized dopaminergic cells (the 1RB3AN27 clone) survived for more than 1 year in grafts and were associated with clinical improvement but, again, did not show significant differentiation and integration in the host brain.

One of the most studied neuronal cells is the NT2 line derived from a human teratocarcinoma that can differentiate *in vitro*, following treatment with retinoic acid, into cells with a mature neuronal phenotype known as "hNT" (or the older designation "NT2N") cells. *In vivo*, grafted hNT cells aquired a mature phenotype (including growth of neuritic processes and synaptic contacts) and could survive for more than a year in the mouse CNS. Interestingly, later studies showed that *in vitro* treatments of NT2 cells are not critical to their *in vivo* differentiation. One of the most convincing applications of using these cells in transplantation comes from experiments using a rat CNS ischemic model, in which hNT cells induced a more robust recovery than fetal rat striatal grafts. In the rat 6-OHDA parkinsonian model, hNT cells grafted into the striatum and substantia nigra were shown to survive in the host parenchyma and generate TH immunoreactivity when pretreated with LiCl. When made for human use, these cells are called LBS-neurons.

B. Polymer Encapsulation

As discussed earlier, one of the major challenges to graft survival in the CNS is modulation of the host immune response. Most neuroglial cells (including *in vitro* adapted lines like PC12) could, at least theoretically, benefit from protection against the host immune response and rejection by using an encapsulating physical barrier. Though many grafted cells can survive, the connectivity between the host and implanted neurons is limited and a significant number of donor cells die. In an early study by Jaeger *et al.*, dopaminergic cell lines survived within a semipermeable membrane. Shortly after, Winn *et al.* proposed that polymer-encapsulated cells may provide a means of neurotransmitter and growth factor delivery *in vivo*. This approach was proposed early on as an alternative treatment in PD, and its feasibility was demonstrated in animal models. In MPTP monkeys that showed clinical improvement after grafting, encapsulated cell implants were shown to be able to uptake and metabolize dopamine.

Another promising application of this approach is the delivery of trophic factors in the regenerating brain. The main problem with using genetically engineered cells for *in vivo* delivery of growth factors is their uncontrolled growth in the host brain. This can be readily solved when the donor cells, modified to secrete trophic factors, are encapsulated in a semipermeable membrane. Intrastriatal implants of polymer-encapsulated cells producing NGF induced a significant increase in the cholinergic neuronal immunoreactivity at the site of implantation. The long-term production of NTF by encapsulated grafts is physiologically relevant and associated with a positive host response in the absence of any detectable side effects. In parkinsonian animal models, encapsulated grafts engineered to produce GDNF added a significant functional benefit to the dopaminergic cografts.

IX. VIRAL VECTORS

The traditional strategy for viral vector mediated gene transfer into the CNS was based on a herpes simplex virus construct. Some of the limitations associated with this approach, like limited long-term robust gene expression in the CNS, have been overcome by new adenoviral vectors. MPTP lesioned mice receiving GDNF via an adenoviral vector showed significantly higher levels of dopamine in the striatum. The use of adenoviral vectors is accompanied by only a limited inflammatory host response. As mentioned earlier, primary neuroglial cells (e.g., astrocytes) are excellent candidates for gene delivering *in vivo*. In a study by Ridet *et al*., astrocytes were infected with an adenoviral construct containing the human *TH* gene. The cells transformed produced significant amounts of TH and released L-dopa. A similar approach was proposed to be used with human neural progenitor cells.

A newer category of vectors increasingly used in CNS applications is represented by adenoassociated viruses (AAVs). In the experimental MPTP model, AAVs expressing TH were introduced in monkey brains, and expression of the donor gene was restricted to neurons in the host striatum. In a similar approach, AAV–TH kindey cells infected with an AAV packaging the TH and aromatic amino acid decarboxylase genes showed increased production of dopamine *in vitro*. *In vivo*, the same viral vector construct used in dopamine-depleted MPTP monkeys induced sustained high levels of TH immunoreactivity. When the *BDNF* gene was incorporated in to an AAV vector also containing a neuronal specific promoter, no significant effects could be seen in a rat parkinsonian model.

Another class of viral vectors tested for CNS delivery is represented by retroviruses. Astrocytes transformed to produce dopa by infection with a TH containing retroviral vector induced a significant functional improvement postgrafting in parkinsonian rats. Nevertheless, the apparently shorter time course of transgene expression in this system may limit its use in brain regeneration. Finally, perhaps the most interesting and potentially useful strategies are based on using lentiviral vectors to deliver therapeutic genes to the CNS. Lentiviruses are characterized by the capacity to infect postmitotic cells (e.g., neurons). If one assumes that all of the theoretical concerns regarding the *in vivo* replication and potential lethal effects of this family of viruses, which includes HIV, are addressed, they may become ideal tools for long-term, stable, transgene expression *in vivo*. In a study of lentiviral gene transfer to the monkey brain, Kordower *et al*. reported that striatal and nigral cells infected with a vector containing β-Gal showed robust, long-term transgene expression in the absence of significant inflammatory host response.

X. SUMMARY

Tremendous achievements in neuroscience over the past three decades have provided a solid foundation for basic and clinical research in neurotransplantation. Restorative neurosurgical procedures will develop from different directions, and it is likely that a combination of approaches will be necessary to maximize patient outcomes.

See Also the Following Articles

BRAIN DAMAGE, RECOVERY FROM • GABA • MODELING BRAIN INJURY/TRAUMA • NEURODEGENERATIVE DISORDERS • NEURON • PARKINSON'S DISEASE • STROKE

Suggested Reading

Bjorklund, A., and Stenevi, U. (1971). Growth of central catecholamine neurons into smooth muscle grafts in the rat mesencephalon. *Brain Res.* **31**(1), 1–20.

Borlongan, C. V., Tajima, Y., Trojanowski, J. Q., Lee, V. M.-Y., and Sanberg, P. R. (1998). Transplantation of cryopreserved human embryonal carcinoma-derived neurons (NT2N) promotes functional recovery in ischemic rats. *Exp. Neurol.* **149,** 310–321.

Emgard, M., *et al*. (1999). Patterns of cell death and dopaminergic neuron survival in intrastriatal nigral grafts. *Exp. Neurol.* **160**(1), 279–288.

Freeman, T. B., and Widner, H.(1998). Cell transplantation for neurological disorders: Toward reconstruction of the human central nervous system. *Contemporary Neuroscience*, Vol. xviii, p. 350. Humana Press, Totowa, NJ.

Gage, F. H., *et al*. (1995a). Survival and differentiation of adult neuronal progenitor cells transplanted to the adult brain. *Proc. Natl. Acad. Sci. USA* **92**(25), 11879–11883.

Gage, F. H., Ray, J., and Fisher, L. J. (1995b). Isolation, characterization, and use of stem cells from the CNS. *Annu. Rev. Neurosci.* **18,** 159–192.

Henderson, B. T., *et al*. (1991). Implantation of human fetal ventral mesencephalon to the right caudate nucleus in advanced Parkinson's disease. *Arch. Neurol.* **48**(8), 822–827.

Isacson, O., Dunnett, S. B., and Björklund, A. (1986). Graft-induced behavioral recovery in an animal model of Huntington's disease. *Proc. Natl. Acad. Sci. USA* **83,** 27–32.

Kendall, A. L., *et al*. (1998). Functional integration of striatal allografts in a primate model of Huntington's disease [see comments]. *Nat. Med.* **4**(6), 727–729.

Kleppner, S. R., *et al*. (1995). Transplanted human neurons derived from a teratocarcinoma cell line (ntera-2) mature, integrate, and survive for over 1 year in the nude-mouse brain. *J. Comp. Neurol.* **357**(4), 618–632.

Kondziolka, D., Wechsler, L., Goldstein, S., Meltzer, C., Thulborn, K., Gebel, J., Jannetta, P., DeCesare, S., Elder, E., McGrogan, M., Reitman, M., and Bynum, L. (2000). Transplantation of cultured human neuronal cells for patients with stroke. *Neurology* **55,** 565–569

Kordower, J. H., *et al*. (1995). Neuropathological evidence of graft survival and striatal reinnervation after the transplantation of fetal mesencephalic tissue in a patient with Parkinson's disease [see comments]. *N. Engl. J. Med.* **332**(17), 1118–1124.

Mendez, I., *et al*. (2000). Enhancement of survival of stored dopaminergic cells and promotion of graft survival by derived neurotrophic factor in patients with Parkinson's disease. Report of two cases and technical considerations. *J. Neurosurg*. **92**(5), 863–869.

Snyder, E. Y., *et al*. (1997). Multipotent neural precursors can differentiate toward replacement of neurons undergoing targeted apoptotic degeneration in adult mouse neocortex. *Proc. Natl. Acad. Sci. USA* **94**(21), 11663–11668.

Thompson, T., Lunsford, L. D., and Kondziolka, D. (1999). Restorative neurosurgery: Opportunities for restoration of function in acquired, degenerative, and idiopathic neurological diseases. *Neurosurgery* **45,** 741–752.

Neuroanatomy

CHRISTOPHER M. FILLEY
University of Colorado School of Medicine and Denver Veterans Affairs Medical Center

GLOSSARY

brain stem The most caudal part of the brain, linking the cerebrum with the spinal cord and consisting of the midbrain, pons, and medulla.

cerebellum The brain structure posterior to the brain stem, the major divisions of which are the laterally placed cerebellar hemispheres and the midline vermis.

cerebrum The paired cerebral hemispheres and diencephalon, which are located rostral to the brain stem and cerebellum.

cortex A thin sheet of neurons and synapses that constitutes the outermost layer of the cerebral hemispheres.

gray matter One of the major types of brain tissue, consisting of neuronal cell bodies and synapses found in either the cortex or the subcortical nuclei.

neuron An excitable cell that serves as the basic functional unit of the nervous system because of its capacity to transmit information electrically.

synapse The point of contact between neurons at which a chemical signal from one neuron influences the excitability of another.

ventricles A series of cavities within the brain—the paired lateral ventricles, the third ventricle, and the fourth ventricle—filled with cerebrospinal fluid.

white matter The term used to describe tracts of myelinated axons responsible for providing structural and functional links between gray matter areas.

The human nervous system is a highly complex assembly of nervous tissue that is responsible for a wide range of homeostatic, motor, sensory, cognitive, and emotional functions. Neuroanatomy is the discipline devoted to the structure of the nervous system. Broadly viewed, the nervous system can be divided into several major components (Table I). The first such division is between the central nervous system (CNS), which consists of the brain and spinal cord, and the peripheral nervous system (PNS), made up of numerous spinal and cranial nerves that transmit information to and from the CNS. In this article, we consider the anatomy of the brain, which stands literally and figuratively at the top of the nervous system as the highest integrative organ in the entire body. The cranial nerves, while strictly speaking part of the PNS, will also be included in this account as they are intimately related to brain structure and cannot be omitted in describing its anatomy. This review will necessarily be confined by space limitations to salient features of the brain's anatomy and further details can be found in several comprehensive textbooks. An attempt will be made herein, however, to emphasize the aspects of brain anatomy most relevant to the phenomena of human behavior.

I. OVERVIEW OF THE BRAIN

Encased securely in the skull, or cranium, the brain is a roughly symmetric, soft, and delicate organ weighing about 3 lb (approximately 1400 g) in the adult human. At the lower end of the brain, the medulla oblongata of the brain stem is continuous with the spinal cord, which descends into the spinal canal below an opening in the base of the skull called the foramen magnum; thus, the CNS is a distinct and connected whole that is protected throughout by the skeletal system. The brain is also surrounded by three layers of covering tissues: the thick dura mater just below the skull, the weblike arachnoid attached to the inner surface of the dura, and the thin pia mater directly adjacent to the surface of the brain.

The brain is customarily divided into three gross anatomic segments: the cerebrum, cerebellum, and brain stem (Fig. 1). The cerebrum consists of the paired cerebral hemispheres, which are joined by a large white matter tract called the corpus callosum, and the diencephalon, which includes the thalamus and hypothalamus. The cerebellum is a relatively large structure situated posterior to the brain stem. The brain stem itself is made up of the midbrain, pons, and medulla (short for medulla oblongata). Also within the brain are four cavities known as ventricles: the two lateral ventricles in the cerebral hemispheres, the third ventricle between the two thalami, and the fourth ventricle between the brain stem and the cerebellum. The ventricles are filled with a clear, colorless liquid called cerebrospinal fluid (CSF). The blood supply of the brain comes from four major arteries in the neck that provide a rich source of blood to meet the high oxygen demand of the brain. There are two carotid arteries and two vertebral arteries that link together in an anastomotic structure called the circle of Willis at the base of the brain, where major arteries irrigating specific cerebral areas originate. Venous drainage is accomplished by a complex network of venous sinuses that ultimately terminate in the paired internal jugular veins in the neck.

A source of confusion to many in studying the gross anatomy of the brain is the use of the singular descriptor in discussions of what are evidently paired structures. For example, it is commonplace to refer to the "cerebrum" when in fact both cerebra are implied.

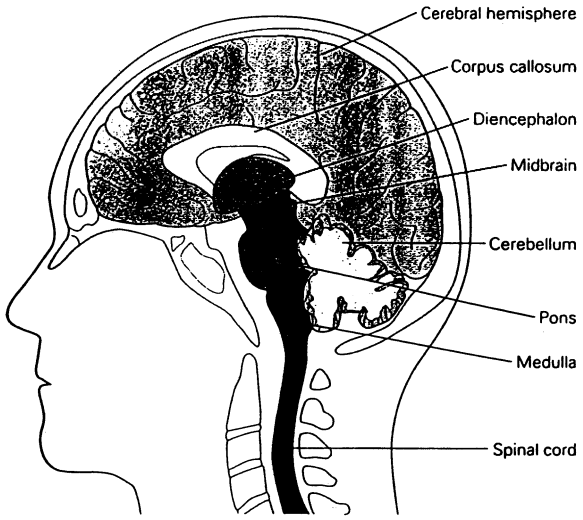


Figure 1 The major structures of the brain, illustrated from a midline view between the two hemispheres. Reprinted with permission from Kandel, E. R., Schwartz, J. H., and Jessell, T. M., Eds. (2000). *Principles of Neural Science*, 4th ed., p. 9. McGraw-Hill, New York.

Table I
Major Neuroanatomic Divisions

Nervous system
Central
Peripheral
Central nervous system
Brain
Spinal Cord
Brain
Brain stem
Midbrain
Pons
Medulla
Cerebellum
Cerebrum
Cerebral hemispheres
Gray matter
Cortex
Basal ganglia
White matter
Diencephalon
Thalamus
Hypothalamus
Cerebral cortex
Frontal lobe
Temporal lobe–limbic system
Parietal lobe
Occipital lobe

This convention, well-entrenched in the scientific literature, is maintained for simplicity because the singular serves as a convenient shorthand. In this article, use of the singular descriptor for gross anatomic regions will imply that a pair of these regions exists unless otherwise specified.

A vast number of cells are found within the brain itself. The cellular composition of the brain includes neurons, or nerve cells, which are the basic functional units of the nervous system, and glial gells, or glia, which have a variety of supporting roles. Estimates vary, but it is possible that the brain contains up to 100 billion neurons and perhaps 10 times as many glial cells. Neurons are excitable cells that function to integrate signals they receive and transmit impulses to other neurons, and each one is connected to many others at points of juncture called synapses. Thus, the brain can be considered to be a densely interconnected electrical organ. The ongoing neuronal activity that results permits an impressive range of integrative functions that the brain maintains over all bodily and mental processes.

A final anatomic distinction to be made in the brain is between the gray matter and white matter. The cell bodies of neurons are concentrated in the gray matter, which largely consists of neuronal somata and their extensive synaptic connections. In gross anatomic terms, gray matter includes both the cortex of the brain, its outer layer, and a number of structures below the cortex known as nuclei. In contrast, the white matter consists of the axons of neurons, which travel throughout the brain and allow gray matter areas to communicate with each other. The white matter is characterized by myelin, a fatty substance that imparts a glistening white appearance to the cut brain. Myelin serves to insulate axons and in so doing greatly increases the efficiency of interneuronal communication by increasing the speed of neuronal conduction. White matter tracts course within and between the hemispheres to connect gray matter areas in functionally unified networks subserving many functions.

II. MICROSCOPIC ANATOMY

As indicated earlier, the brain contains extraordinary numbers of neurons and glial cells. The remarkable capacities of the brain result in part from the wealth of neurons and their connections, and the glia provide important support for the optimal function of the neuronal population. In this section, we consider microscopic aspects of the major cell types in the brain.

The neuron is the fundamental functional unit of the nervous system. Neurons are anatomically specialized to transmit information, which may take the form of a sensory stimulus received from the periphery or a motor signal destined to produce movement in a muscle. In the brain, neurons have many additional functions that can be generally termed "information processing", and presumably this category includes all of the cognitive and emotional operations traditionally associated with mental activity and behavior. Most brain neurons, in fact, are classified as interneurons because they are interposed between sensory input and motor output and mediate these other, "higher" functions. To accomplish all of these various tasks, neurons have a typical arrangement that includes a cell body, a variable number of dendrites, and an axon (Fig. 2). The cell body, also known as the soma or perikaryon, contains the cell nucleus and other organelles that maintain the metabolic status of the neuron and synthesize macromolecules essential for its function. Dendrites are relatively short processes extending from the cell body that receive input from adjacent neurons via synaptic contacts (see later discussion).

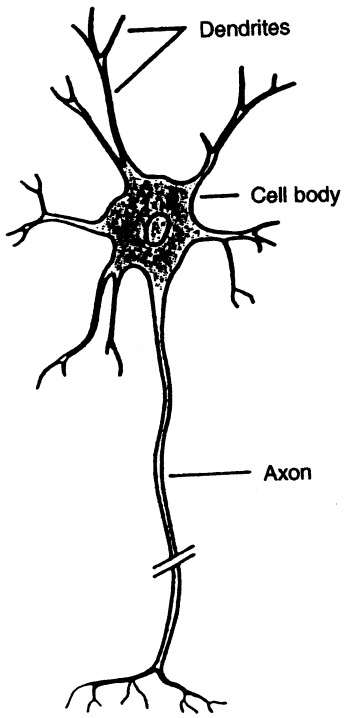


Figure 2 Drawing of a typical neuron, with its characteristic cell body, dendrites, and axon. Reprinted with permission from Kandel, E. R., Schwartz, J. H., and Jessell, T. M., Eds. (2000). *Principles of Neural Science*, 4th ed., p. 24. McGraw-Hill, New York.

The axon is a long, cylindrical process that provides for the output of the neuron, again via synaptic contact with adjacent neurons.

The transfer of information in the nervous system is both electrical and chemical in nature (Fig. 3). Within a single neuron, the signal is electrical and takes the form of an action potential. This is an electrical impulse (often referred to as a spike) propagated along the axon by virtue of a sudden influx of sodium ions that transiently reverses the polarity of the axonal membrane. The resting membrane potential is quickly restored after the action potential passes, and after a short refractory period another spike can be propagated. The action potential is an all-or-none phenomenon that depends on the balance between excitatory and inhibitory input received by the neuronal dendrites.

Whereas the neuron itself conducts information electrically, the junction between neurons, known as the synapse, operates by means of chemical transmission (Fig. 3). A synapse is a specialized region at which a chemical messenger known as a neurotransmitter from one neuron (presynaptic) diffuses across a narrow synaptic cleft to activate another neuron (postsynaptic). When the neurotransmitter binds with the receptor on the postsynaptic membrane, it acts to produce either a depolarization (excitatory stimulus) or hyperpolarization (inhibitory stimulus), and the summation of these many postsynaptic potentials determines whether an action potential is generated in the postsynaptic neuron. Neurotransmitters are generally small amines, amino acids, and neuropeptides, and their pharmacology promises many avenues for the successful manipulation of abnormal physical and mental states.

Far more abundant even than neurons in the brain are glial cells. Glia of the brain and spinal cord are classified into four types: astrocytes, oligodendrocytes, microglia, and ependymal cells. Astrocytes are star-shaped cells found in both gray and white matter that have a role in the mechanical support of neurons, contribute to metabolic regulation of the microenvironment of the brain, and participate in its response to injury. Oligodendrocytes are confined mainly to white matter, where they are responsible for the myelination of brain axons, just as Schwann cells perform this function in the PNS. Microglia are small cells found in gray and white matter that serve as the phagocytes of the brain, migrating as necessary to damaged areas where they consume pathogens and neuronal debris.

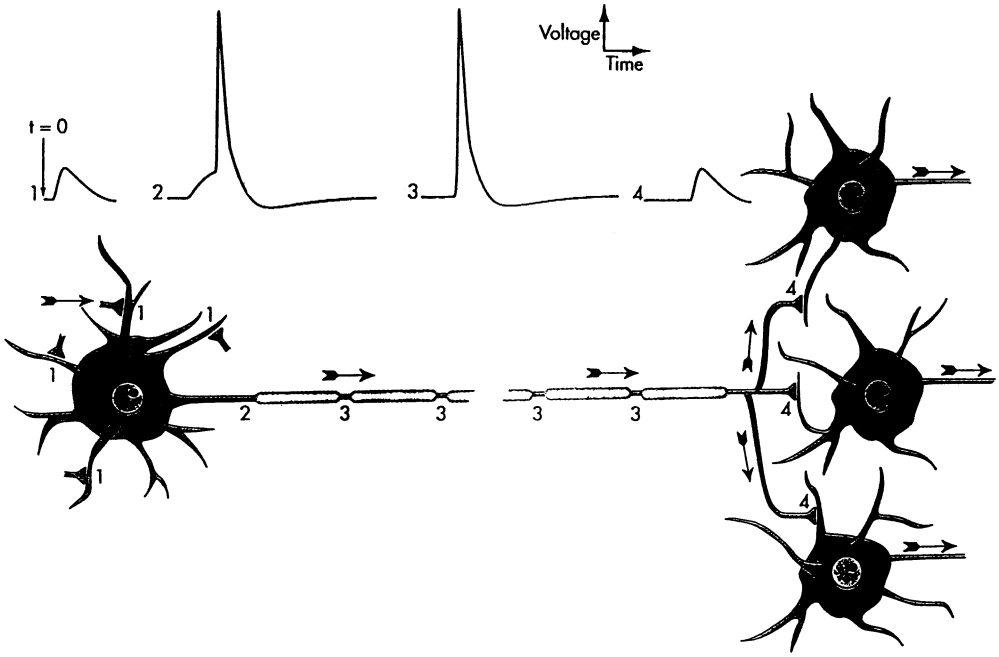


Figure 3 General depiction of information transfer between neurons. Chemical signaling by neurotransmitter release onto a neuron (1) causes the generation of an action potential (2), which is an electrical signal propagated down the axon (3). The action potential then causes further neurotransmitter release onto subsequent neurons (4). Reprinted with permission from Nolte, J. (1999). *The Human Brain*, 4th ed., p. 145. Mosby, St. Louis.

Ependymal cells line the ventricles of the brain, and at a specialized structure called the choroid plexus, one of which is found in each ventricle, ependymal cells form a secretory epithelium that produces the CSF that fills the ventricles and bathes the entire CNS.

III. BLOOD SUPPLY

The brain is extremely vulnerable to even a momentary reduction in or loss of blood flow, and an uninterrupted supply of well-oxygenated blood is critical. A complex system of arteries and veins convey blood to and from the brain, and interruption of the vascular system may result in stroke, one of the most common and devastating neurologic disorders.

The arterial supply of the brain comes entirely from four major vessels that originate in the neck (Fig. 4). The right and left common carotid arteries arise from the right subclavian artery and ascending aorta, respectively, and within a few centimeters each bifurcates into an external branch, which supplies extracranial structures, and an internal carotid artery (ICA), which irrigates major portions of the brain. The paired vertebral arteries are somewhat smaller and arise from the subclavian arteries; these ascend in parallel to a level just below the pons, where they join to form the single basilar artery.

The two internal carotid arteries and basilar artery then contribute to a vascular structure at the base of the brain called the circle of Willis. This is a continuous vascular loop comprising paired posterior cerebral arteries (PCAs) that bifurcate from the top of the basilar artery and go on to supply posterior parts of the brain, paired posterior communicating arteries (PCoAs) that connect the PCAs with the ICAs, paired anterior cerebral arteries (ACAs) that arise from the ICAs and go on to irrigate anterior regions of the brain, and a single anterior communicating artery (ACoA) that joins the two ACAs. Another important artery arising from the circle of Willis is the middle cerebral artery (MCA), one of which ascends to each hemisphere and nourishes the lateral aspects of the brain. Clinically, the MCA, ACA, and PCA are the most important cerebral arteries in that the majority of strokes occur in these distributions and lead to major neurologic deficits because of the large areas of the cerebrum that may be damaged (Fig. 5).

The venous drainage of the brain is accomplished by a complex system of superficial and deep veins, all of which are richly anastomosed with each other. The superficial veins drain into the superior sagittal sinus and the deep veins into the straight sinus. The final destinations of venous blood from the brain are the paired internal jugular veins in the neck, by which the blood is conveyed ultimately to the heart. Disorders of the venous system are far less common than those affecting the arterial system of the brain.

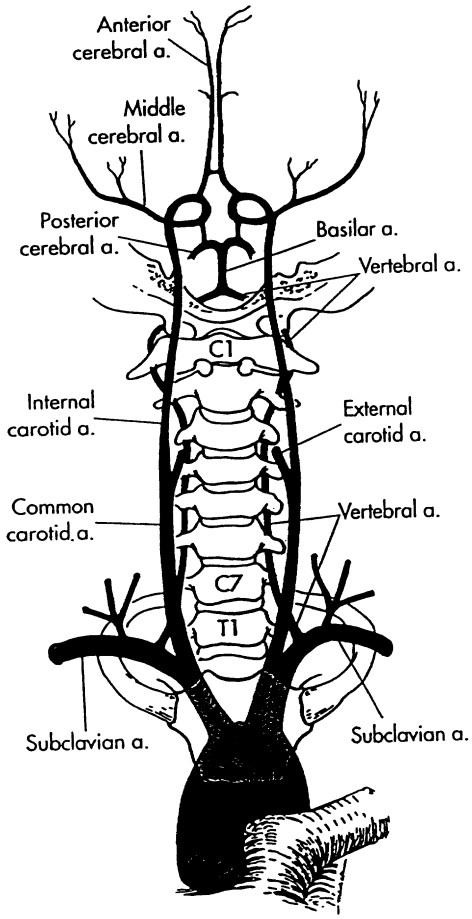


Figure 4 The blood supply of the brain, originating from four major arteries that arise from the aorta. Reprinted with permission from Nolte, J. (1999). *The Human Brain*, 4th ed., p. 145. Mosby, St. Louis.

A final point regarding the blood supply of the brain is the existence of a blood–brain barrier. This term refers to an anatomic–physiologic complex at the level of the capillary endothelium that prevents the entry of many substances into the brain from the blood stream. The blood–brain barrier offers clear advantages, as in the case of barring the intrusion of blood-borne pathogens, but it also carries the liability that antibiotics and other drugs may have difficulty reaching the sites in the brain where they are needed.

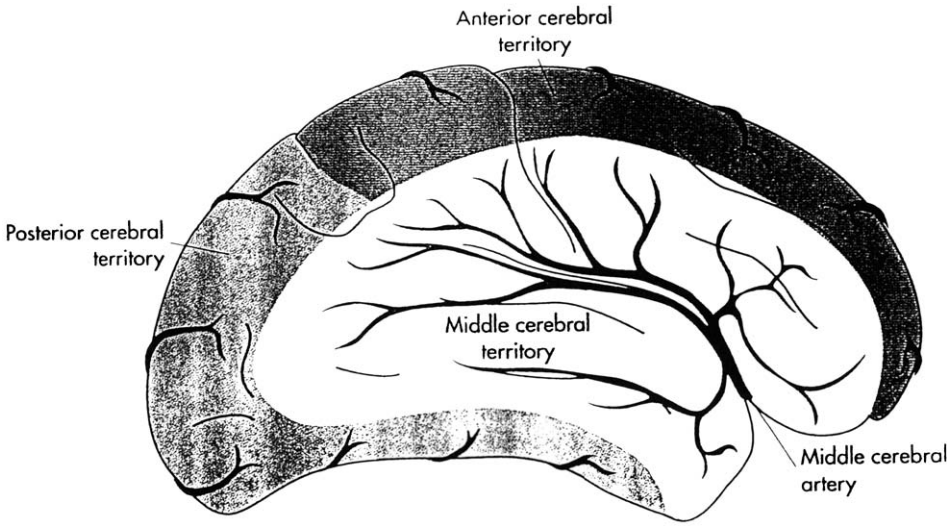


Figure 5 Distributions of the middle cerebral, anterior cerebral, and posterior cerebral arteries on the lateral surface of the brain. Reprinted with permission from Nolte, J. (1999). *The Human Brain*, 4th ed., p. 122. Mosby, St. Louis.

IV. VENTRICLES AND CEREBROSPINAL FLUID

The brain contains four internal cavities called ventricles (Fig. 6). These cavities are filled with CSF, which is produced in the ventricles and serves to bathe the entire CNS. An understanding of the ventricular system is important for an appreciation of fluid and pressure dynamics in the brain, as well as for its anatomy.

The two largest ventricles are the lateral ventricles, one in each hemisphere, which underlie the frontal, parietal, occipital, and temporal lobes. These communicate with a single third ventricle, narrow and situated between the two thalami, via an opening in each lateral ventricle called the foramen of Monro. Finally, the tent-shaped fourth ventricle, just dorsal to the brain stem, is connected with the third ventricle by a small conduit in the midbrain called the cerebral aqueduct. The fourth ventricle, in turn, empties into a region of subarachnoid space called the cisterna magna through three apertures, the midline foramen of Magendie and the two lateral foramina of Luschka. More caudally, the CSF flowing from the ventricles travels down around the spinal cord to the lower end of the spinal canal and also circulates rostrally to the convexities of the brain, where it is eventually absorbed into the cerebral venous sinuses through structures known as arachnoid villi.

CSF is produced by the choroid plexus in all four ventricles, and the entire volume of the CSF in and around the CNS is about 140 ml. The volume of the ventricular system is actually rather small, about 25 ml. Because CSF is produced at the rate of about 450 ml/day, the entire CSF volume turns over more than three times daily. The CSF has an obvious supportive role in that it provides a buoyancy that keeps the brain from settling down upon rigid, bony surfaces. The CSF also takes part in regulating the chemical environment of brain neurons.

In clinical terms, the ventricular system and the CSF have many implications. For example, CSF analysis

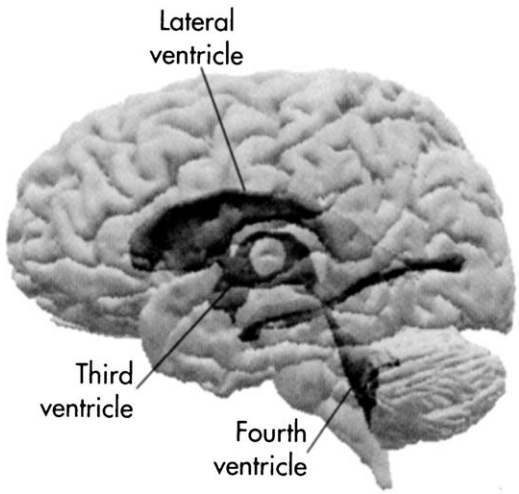


Figure 6 The position of the four ventricles—the two lateral ventricles, the third, and the fourth—within the brain. Reprinted with permission from Nolte, J. (1999). *The Human Brain*, 4th ed., p. 65. Mosby, New York.

after lumbar puncture is essential in the diagnosis of many neurologic disorders, such as meningitis and encephalitis. Moreover, enlargement of the ventricles due to an excess of CSF, as is seen in hydrocephalus or with certain mass lesions, can have major neurologic consequences.

V. CRANIAL NERVES

The 12 cranial nerves are important for providing motor and sensory innervation of the head and neck, and their anatomy is inextricably linked with that of the brain. Although the cranial nerves are generally regarded as belonging to the PNS, the second cranial nerve, the optic nerve, is in fact considered a part of the CNS. All cranial nerves exist in pairs, and their crossed and uncrossed connections with central structures are important for an appreciation of the neuroanatomy of the brain.

A. Cranial Nerve I

Olfaction is the sense of smell. Although lower animals have a more highly developed olfactory system, this chemical sense persists in humans. The olfactory system originates in the roof of the nasal cavity with the olfactory epithelium, which is a collection of olfactory receptor cells. The axons of these cells are known collectively as the olfactory nerve. Ascending through the cribriform plate of the ethmoid bone, the olfactory nerve terminates in the olfactory bulb at the base of the frontal lobe. From there the olfactory tract projects to the olfactory cortex in the medial temporal lobe.

B. Cranial Nerve II

The visual system is extremely important in human life, and a substantial percentage of CNS neurons is devoted to its function. Incoming visual stimuli are first processed in the eye, where photoreceptor cells in the retina, called rods and cones, transduce patterns of light into electrical signals that travel to the brain. The optic nerve begins at the back of the eye, where it is visible with an ophthalmoscope as the optic disk. Visual information is first transmitted via the optic nerve to the lateral geniculate nucleus of the thalamus, and then additional relays through the temporal and parietal lobes project to the primary visual cortex in the occipital lobe (Fig. 7). As it travels to the thalamus, the optic nerve splits into two components, one remaining on the same side as the eye from which it came and the other passing over to the other side. The importance of this anatomic feature is that each hemisphere receives visual input from the contralateral visual field; thus, the right hemisphere interprets visual input from the left side of vision and vice versa. This aspect of crossed function is typical of the organization of many brain systems and has important clinical and behavioral implications.

C. Cranial Nerves III, IV, and VI

These three nerves are grouped together because they all are exclusively involved with eye movements, and routine clinical testing of these movements permits a quick survey of the function of each nerve. The oculomotor nerve (CN III) originates from the oculomotor nucleus in the midbrain and innervates the following muscles: the levator palpebrae, which raises the eyelid; the superior rectus, inferior rectus, medial rectus, and inferior oblique, which mediate most voluntary directions of gaze; and the pupillary sphincter, which constricts the pupil of the eye when light stimulation is provided. The trochlear nerve (CN IV) also begins in the midbrain but innervates only the superior oblique muscle; this muscle moves the eye downward and laterally. The abducens nerve (CN VI) comes from the pons and supplies the lateral rectus muscle, which allows lateral deviation (abduction) of the eye.

D. Cranial Nerve V

This nerve, known as the trigeminal nerve, carries out important sensory and motor functions for the face. The trigeminal nerve is the general sensory nerve for the head; it mediates somatic sensation in the face by means of three subdivisions—the ophthalmic (V1), maxillary (V2), and mandibular (V3) branches—which converge in the trigeminal ganglion outside the brain stem and then enter the pons. A secondary relay to the brain then sends this facial somatosensory information to the ventral posterior medial nucleus of the thalamus, where it is processed and brought to consciousness. Somatic sensation for the head posterior to the face is mediated by the second and third

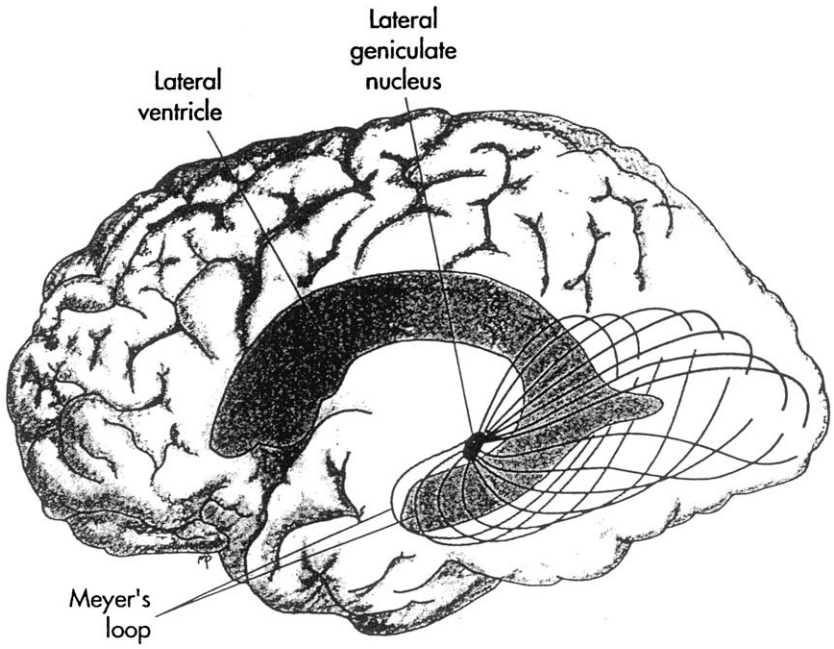


Figure 7 Central connections of the visual system. Fibers from the eyes reach the lateral geniculate nucleus, whence relays proceed to the occipital cortex. The relay in the temporal lobe is known as Meyer's loop. Reprinted with permission from Nolte, J. (1999). *The Human Brain*, 4th ed., p. 419. Mosby, New York.

cervical nerves that enter the spinal cord before traveling to the thalamus. The trigeminal nerve also has a motor division that originates in the pons and supplies the muscles of mastication (chewing).

E. Cranial Nerve VII

This is the facial nerve, which has a primary motor function. The facial nerve originates from the facial nucleus in the pons, where it receives input from the corticobulbar tract, the source of motor innervation from the precentral gyrus of the cerebral cortex. After its axons leave the pons, the facial nerve innervates the ipsilateral muscles of facial expression. Damage to the facial nerve or its nucleus thus produces weakness of one entire side of the face. However, damage to the corticobulbar fibers arriving at the facial nucleus in the brain stem produces a somewhat different pattern of weakness. In this instance, typically following a contralateral stroke in the brain, only the muscles of the lower face are weak because the upper facial muscles are supplied by corticobulbar fibers from both hemispheres and the intact side can compensate for the damaged side. This anatomic feature permits the crucial clinical distinction between a stroke and the less ominous situation of weakness from seventh nerve involvement. The facial nerve also has one notable sensory function, which is to convey taste from the anterior two-thirds of the tongue, via a branch called the chorda tympani, to the solitary tract in the pons and medulla.

F. Cranial Nerve VIII

The eighth cranial nerve is known as the vestibulocochlear nerve because it is made up of two special sensory components called the vestibular division and the auditory division. Correspondingly, this nerve has a dual role: it participates in the vestibular or balance system and also in the sense of audition or hearing. Each division makes use of mechanoreception; cells of the vestibular division are sensitive to positional movements of the head, and those of the cochlear division are sensitive to sound stimuli. The receptors for both divisions are found in the inner ear, well-protected deep within the temporal bone of the skull. By means of complex processes of tranduction, mechanical and auditory stimuli are sent to vestibular and cochlear nuclei, respectively, in the lower pons. From these sites, vestibular input is extensively processed in the brain stem and cerebellum, whereas auditory input is sent rostrally up the brain stem to the

medial geniculate nucleus of the thalamus and finally to the primary auditory cortices in the temporal lobes (Heschl's gyri). One of the central functions of hearing in humans is that it serves as a necessary precursor to language.

G. Cranial Nerve IX

The glossopharyngeal nerve participates in motor, sensory, and autonomic functions. Motor fibers of CN IX arise from the nucleus ambiguus in the medulla and innervate the stylopharyngeus muscle of the pharynx. Sensory fibers from the tongue, nasopharynx, and middle and outer ear travel to the spinal trigeminal nucleus, and taste fibers from the posterior one-third of the tongue terminate in the solitary tract. Finally, there is parasympathetic innervation of the parotid gland that originates from the inferior salivatory nucleus.

H. Cranial Nerve X

This nerve, the vagus, is the most widely distributed cranial nerve, as it provides parasympathetic innervation to many thoracic and abdominal visceral organs from neurons situated in the dorsal motor nucleus of the vagus in the medulla. The vagus nerve also shares many similarities with the glossopharyngeal nerve. Motor fibers of CN X arise from the nucleus ambiguus and innervate most of the muscles of the larynx and pharynx, taste fibers from the epiglottis enter the solitary tract, and somatosensory fibers from the outer ear enter the spinal trigeminal tract. In clinical practice, the gag reflex allows for the simultaneous testing of both the ninth and tenth cranial nerves, as the afferent and efferent limbs of this reflex are mediated by the glossopharyngeal and the vagus nerves, respectively.

I. Cranial Nerve XI

The XIth nerve is also called the accessory nerve, and it contains only motor fibers. These arise from the lower medulla and upper spinal cord and innervate the ipsilateral sternocleidomastoid and trapezius muscles.

J. Cranial Nerve XII

Like the accessory nerve, the hypoglossal nerve (CN XII) is purely motor. It arises from neurons in the hypoglossal nucleus of the medulla and enters the tongue from below to supply its musculature.

VI. BRAIN STEM

The brain stem is the most caudal portion of the brain, and structurally it serves to connect the cerebrum with the spinal cord as well as to anchor the cerebellum behind it. Its three major segments are the midbrain, which lies just below the diencephalon and is continuous with the thalamus, the pons, which is below the midbrain and anterior to the fourth ventricle, and the medulla, which is continuous with the spinal cord below. In addition to harboring many cranial nerve nuclei and tracts, the brain stem contains long ascending and descending tracts to and from higher structures, as well as an important integrative structure called the reticular formation.

As reviewed previously, cranial nerves III–XII all have their central termini in the brain stem. Thus, the brain stem serves as a general relay station conveying sensory, motor, and autonomic information between the CNS and the tissues and organs of the face and body. Damage to the brain stem can thus have a wide range of effects on cranial nerve function that can often be detected clinically, as is well-described by the observations of classic neurology.

The long tracts that are found within the brain stem are all continuations of tracts that begin above or below. Four major tracts deserve mention. First is the corticospinal tract, which begins in the precentral gyrus of the frontal lobe and descends to the spinal cord, where it provides supraspinal input to motor neurons that directly innervate voluntary muscles. In the brain stem, this tract occupies the ventral portion of the midbrain, pons, and medulla, and near the bottom of the medulla it crosses (decussates) so that most corticospinal fibers travel to the opposite side of the spinal cord. Thus, as in the visual system, there is a crossing of neural systems that renders one side of the brain responsible for nervous activity on the other side of the body. The second major tract is the corticobulbar tract, which performs a role similar to that of the corticospinal tract, but terminates in various brain stem motor nuclei. The other two long tracts of the brain stem are sensory. The first is the medial lemniscus, which is a continuation of the dorsal column system in the spinal cord that conveys information about vibration and position sensation to the ventral posterior lateral (VPL) nucleus of the

thalamus and thence to the somatosensory cortex of the parietal lobe. The other is the spinothalamic tract, a similar fiber system that transmits pain and temperature sensation from the spinal cord again to the VPL nucleus of the thalamus and ultimately the parietal cortex.

The integrative function of the brain stem is carried out by a collection of diffusely organized nuclei and tracts in the brain stem core called the reticular formation. This area is defined more by its physiological characteristics than by discrete anatomic boundaries, and it includes some important neurotransmitters, among them norepinephrine from the locus ceruleus and serotonin from the dorsal raphe nuclei. The reticular formation participates in sensation, movement, and autonomic function, but perhaps its most important role is in the maintenance of consciousness. A portion of the reticular formation in the upper pons and midbrain, called the ascending reticular activating system (ARAS), sends projections to the intralaminar nuclei of the thalamus, which in turn project to the entire cerebrum (Fig. 8). By virtue of the general cerebral activation enabled by this system, the ARAS has a crucial role in wakefulness and is largely responsible for the sleep and wakefulness cycle of normal humans. Damage to the ARAS, as from a brain stem stroke or traumatic brain injury, may produce a loss of arousal and result in the clinical state of coma. Thus, the ARAS is necessary for the level of consciousness; conversely, the content of consciousness, as will be seen, is elaborated by more rostral structures in the brain.

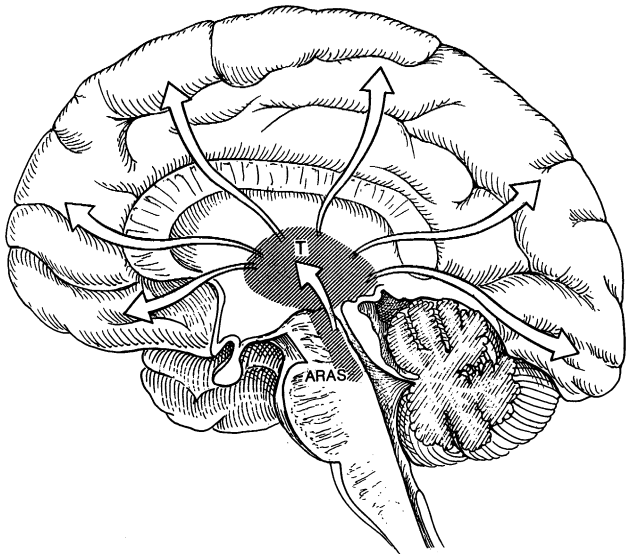


Figure 8 Drawing of the ascending reticular activating system (ARAS), which is found in the upper pons and midbrain and sends projections rostrally to the thalamus. Reprinted with permission from Filley, C. M. (2001). *Neurobehavioral Anatomy*, 2nd ed., p. 42. University Press of Colorado, Boulder, CO.

VII. CEREBELLUM

The cerebellum is a large, semidetached region of the brain lying behind the brain stem, to which it is extensively attached. Even though the cerebellum receives substantial sensory input, it is considered a part of the motor system because of its primary involvement with coordination, postural control, equilibrium, and muscle tone.

In gross anatomic terms, the cerebellum can be divided into the body of the cerebellum (corpus cerebelli) and the much smaller flocculonodular lobe. A more useful functional distinction, however, can be made between the two lateral cerebellar hemispheres and the centrally located vermis (Fig. 9). This division is important because the cerebellar hemispheres are concerned with coordination of the limbs, and the vermis is devoted to postural adjustment. Thus, damage to these regions causes, respectively, limb ataxia and postural instability (also known as trunkal ataxia).

Like the cerebrum, the cerebellum contains both gray and white matter. The gray matter includes billions of cerebellar cortical neurons, which are arranged in three layers: a superficial molecular layer, an intermediate Purkinje cell layer, and a deeper granular cell layer. In addition, there are four cell collections deep within the cerebellum called the dentate, globose, emboliform, and fastigial nuclei (Fig. 9). The white matter consists in part of axons traveling to and from the cerebellar cortex. There are also three prominent cerebellar peduncles—inferior, middle, and superior—that connect the cerebellum with the medulla, pons, and midbrain, respectively (Fig. 9).

The major role that the cerebellum plays in the motor system results from its intermediate position between multiple sensory inputs and its output to motor areas in the cerebral hemispheres. Through projections in the inferior and middle cerebellar peduncles, the cerebellum receives a variety of vestibular, spinal, and cerebral cortical inputs that reach the cerebellar cortex where extensive processing takes place. From this point, cerebellar output occurs mainly by relays in the four deep nuclei, which send axons through the superior cerebellar peduncle to the

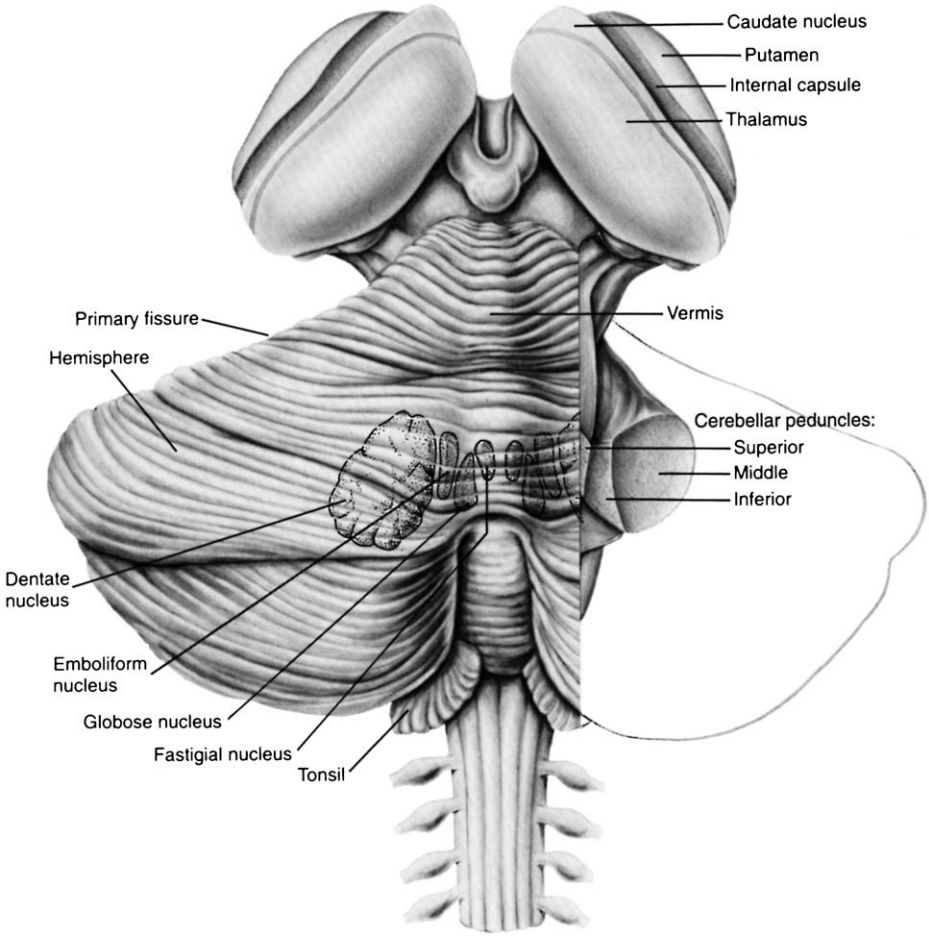


Figure 9 The cerebellum, as depicted schematically from behind. The major divisions are the two hemispheres, situated laterally, and the single midline vermis. Reprinted with permission from Kandel, E. R., Schwartz, J. H., and Jessell, T. M., Eds. (1991). *Principles of Neural Science*, 3rd ed., p. 628. Elsevier, New York.

midbrain and eventually the ventral anterior and ventral lateral nuclei of the thalamus. These thalamic connections allow the cerebellum to exert a powerful influence on motor regions of the cerebral cortex, providing fine-tuning of movements of the limbs and trunk. An important clinical point is that ataxia on one side of the body is ipsilateral to the side of the cerebellar lesion. Because the cerebellar output crosses to the opposite thalamus and then the subsequent corticospinal system crosses again, the bodily deficit is on the same side as the cerebellar damage.

Cerebellar lesions are most often associated with the clinical findings of ataxia, which may affect the limbs, trunk, or even speech (producing a specific type of dysarthria known as scanning speech), dysequilibrium as manifested by a wide-based gait, and muscular hypotonia. Findings have also suggested that the cerebellum has a contribution to higher motor and even cognitive functions. The acquisition of a motor skill such as playing the piano, an example of procedural learning, seems to depend in part on the cerebellum, and there are increasingly frequent reports of individuals with cerebellar lesions who develop cognitive deficits.

VIII. DIENCEPHALON

The diencephalon consists of four structures located deep in the cerebral hemispheres just rostral to the midbrain and surrounding the third ventricle: the thalamus, hypothalamus, subthalamus, and epithalamus. Despite its small size, the diencephalon has major importance in brain function, particularly through the

activities of the thalamus and hypothalamus. The subthalamus is a very small region inferior to the thalamus that contains the subthalamic nucleus and zona incerta; these areas have connections to the basal ganglia and cerebral cortex, but their functions are largely unknown. The epithalamus is found superior and caudal to the thalamus and contains the pineal gland and the habenular nuclei. The pineal gland is an unpaired structure that was once thought to harbor the seat of the soul; today, it is known to secrete a hormone called melatonin, hypothesized to have a role in sleep and gonadal function.

The thalamus is an egg-shaped collection of nuclei that makes up about 80% of the diencephalon (Fig. 10). It consists of a collection of nuclei that participate in sensation, movement, cognition, emotion, and arousal. The most characteristic thalamic function is to serve as a sensory relay station for stimuli that will eventually reach the cerebral cortex. All sensory systems with the exception of olfaction pass through the thalamus en route to their respective cortical areas. Thus, somatosensory information from the body and face reach the VPL and ventral posterior medial (VPM) nuclei, respectively, and taste fibers also project to the VPM nucleus. Visual projections from the optic nerve synapse in the lateral geniculate nucleus and auditory fibers in the medial geniculate nucleus. The ventral anterior and ventral lateral nuclei receive fibers from the cerebellum and in turn send projections to the basal ganglia, so that they contribute to the motor system. The dorsal medial nucleus and the pulvinar are the major association nuclei, connected with the frontal and the parietal–temporal–occipital cortices, respectively. The anterior nucleus has a role in the limbic system (see following discussion). Finally, the intralaminar nuclei, the two largest of which are the centromedian and the parafascicular, receive input from the ARAS in the brain stem and then relay this general information rostrally to activate the cerebrum as a whole.

The hypothalamus is considerably smaller than the thalamus, but it exerts powerful effects on autonomic, endocrine, and emotional functions. Located inferior to the thalamus and superior to the pituitary gland, the hypothalamus is a collection of many tiny nuclei that

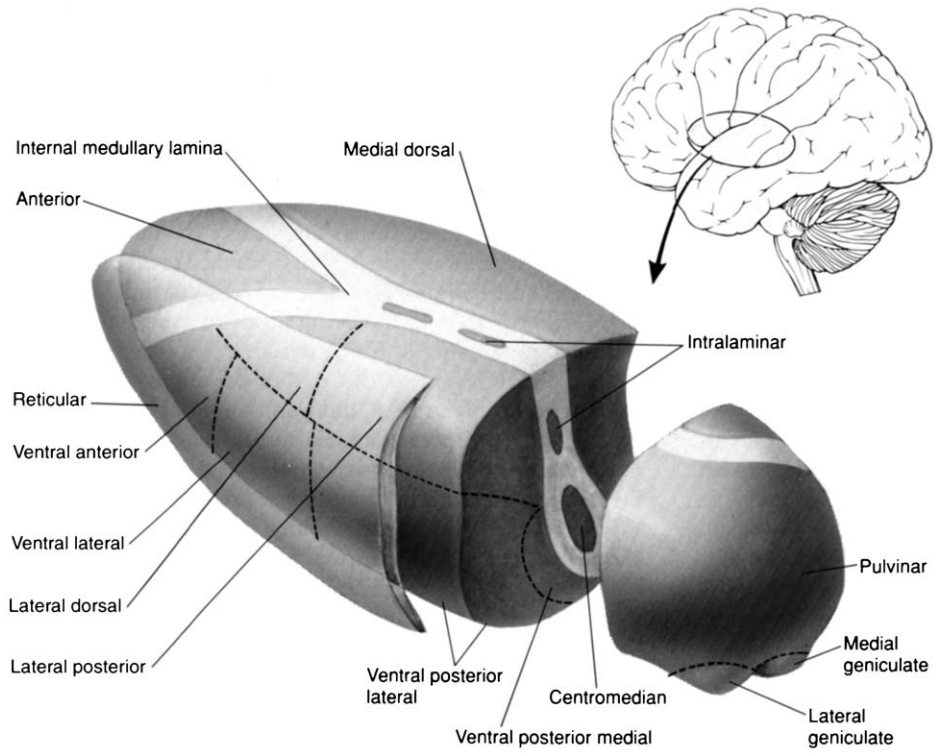


Figure 10 The thalamus, with its many nuclei depicted. These nuclei act as relay stations for information traveling to and from the cortex. Reprinted with permission from Kandel, E. R., Schwartz, J. H., and Jessell, T. M., Eds., (1991). *Principles of Neural Science*, 3rd ed., p. 291. Elsevier, New York.

contribute to the regulation of the body's homeostasis. The hypothalamus is the control center of the autonomic nervous system, an involuntary portion of the nervous system that regulates aspects of body temperature, digestion, circulation, water balance, and sexual function. The autonomic nervous system is divided into the parasympathetic branch, which is generally associated with the anterior regions of the hypothalamus, and the sympathetic branch, which is associated with posterior sites. Endocrine function is also governed by the hypothalamus by virtue of its strong neural and vascular connections with the two lobes of the pituitary gland. Finally, the hypothalamus shares the mammillary bodies in common with the limbic system and, in part because of this linkage, plays a role in emotional function. Hypothalamic lesions in experimental animals and in humans, for example, have sometimes been noted to produce rage behaviors.

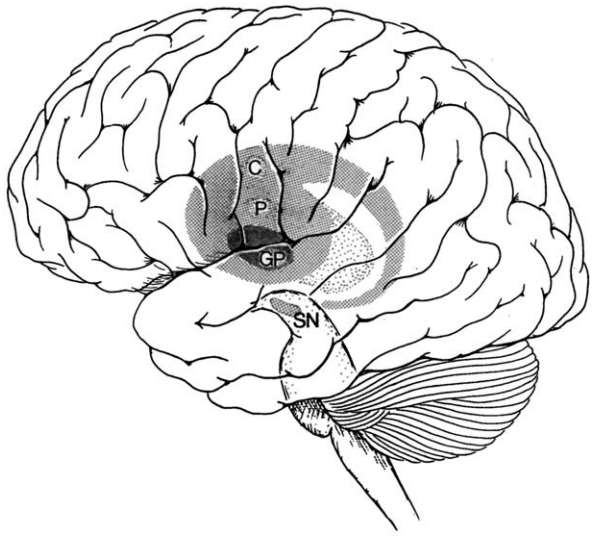


Figure 11 The positions of the basal ganglia and the substantia nigra within the hemispheres. Abbreviations: C, caudate; P, putamen; GP, globus pallidus; SN, substantia nigra. Reprinted with permission from Filley, C. M. (2001). *Neurobehavioral Anatomy*, 2nd ed., p. 181. University Press of Colorado, Boulder, CO.

IX. BASAL GANGLIA

The basal ganglia are a group of large nuclei located deep in the cerebral hemispheres. They are important because of their exclusively motor affiliation and because of the many disorders of movement that result from damage to these structures.

There is no uniformity of opinion about which structures should be included among the basal ganglia, but there is general agreement that the caudate nucleus, putamen, and globus pallidus should be listed under this heading. Many authors also include the midbrain substantia nigra and the subthalamic nucleus of the diencephalon as basal ganglia as well. In any case, the caudate, putamen, and globus pallidus are all generally situated at the base of the hemispheres, lateral to the thalamus and medial to the cerebral cortex (Fig. 11). The caudate is separated from the putamen and the globus pallidus by a thick band of white matter called the internal capsule, among the functions of which is the conveyance of motor impulses from the cortex to the motor neurons of the face and body below. Combinations of the basal ganglia also have frequently used names; hence, the caudate and putamen are known as the striatum, and the putamen and globus pallidus are referred to as the lenticular nucleus.

The principal function of the basal ganglia as an integrated unit is to modulate the activity of the motor cortex in the cerebrum as it organizes voluntary movement of the bodily musculature. Consistent with this role, the basal ganglia are connected with the cortex via a series of parallel loops that permit extensive involvement in cortical motor output. The primary such loop consists of the following: multiple cortical inputs reach the striatum by way of the internal capsule and another white matter tract called the external capsule; from there, connections proceed first to the globus pallidus and to then the ventral anterior and ventral lateral thalamic nuclei; the final link involves connections that return back to the motor cortex, again via the internal capsule. Thus, the basal ganglia join the cerebellum as regions strongly connected to the voluntary motor system by relays in specific thalamic nuclei. This anatomic linkage provides powerful modulatory input to the motor system, integrating two other brain areas into the organization of movement and emphasizing the general importance of motor action in human life. As reviewed previously, the cerebellum has a prominent role in coordination; the basal ganglia, in contrast, can be thought of as contributing to the initiation and timing of movements, although the exact contribution of these nuclei is still being clarified.

A final aspect of basal ganglia anatomy to be considered is its neurochemical input that arises from the substantia nigra of the midbrain (Fig. 11). Pigmented cells of the substantia nigra send axons rostrally, where they deliver the neurotransmitter

dopamine to the striatum. Dopamine serves as an activator of the basal ganglia and the motor system as a whole, and its deficiency or absence results in dramatic alterations in motor function. Individuals with Parkinson's disease, characterized by a loss of dopaminergic cells in the substantia nigra, display the classic clinical features of bradykinesia (slowness of movement), rigidity, and tremor. This is the most important movement disorder because of its high prevalence and because of the fact that the provision of dopamine or similar agents can significantly alleviate the clinical symptoms of the disease.

X. LIMBIC SYSTEM

The limbic system is a loosely defined collection of structures at the junction of the diencephalon and the cerebral hemisphere. Some authors consider the limbic system a discrete lobe of the brain, but because of its extensive thalamic and hypothalamic connections, we will regard the limbic system as a transitional region between the diencephalon and the hemisphere. Its two most important structures—the amygdala and the hippocampus—are found within the temporal lobe. The name "limbic" derives from the Latin *limbus*, meaning "border," and the limbic system was originally named by the French neurologist Paul Broca in 1878 to reflect the location of a number of cortical areas at the margin of the hemisphere. These cortical regions—the cingulate gyrus, the parahippocampal gyrus, and the hippocampus—are typified by phylogenetically primitive cortex, and it was long thought that they were devoted to the sense of smell because of their connections with the olfactory system at the base of the frontal lobe (Fig. 12). However, later studies provided ample evidence that, whereas the limbic system in animals is indeed strongly linked with olfaction, the limbic system in humans is dedicated to other, more important functions. It is now widely accepted that the human limbic system has a prominent role in both emotion and memory.

In 1937, an influential paper by James Papez proposed that an interconnected series of structures —the cingulate gyrus, the parahippocampal gyrus, the hippocampus, the fornix, the mammillary bodies, the mammillothalamic tract, and the anterior nucleus of the thalamus—comprised the cerebral basis of emotion. This group of limbic structures came to be known as the Papez circuit (Fig. 13), and despite decades of debate, this region endures as a central concept in the neuroanatomy of emotion. Studies in the past decade have also demonstrated that the amygdala, located adjacent to the hippocampus (Fig. 13), has an important role in emotional learning, and this nucleus has joined the Papez circuit as a core constituent of the limbic system. Sensory input to the limbic system undergoes processing in the amygdala, and the emotional significance of a stimulus is thereby determined. In this way, the experience of emotion is mediated by limbic structures. There is also an effector component of the limbic system; connections with the hypothalamus account for the important autonomic and endocrine reactions to powerful emotions that have been referred to as the "fight or flight" response.

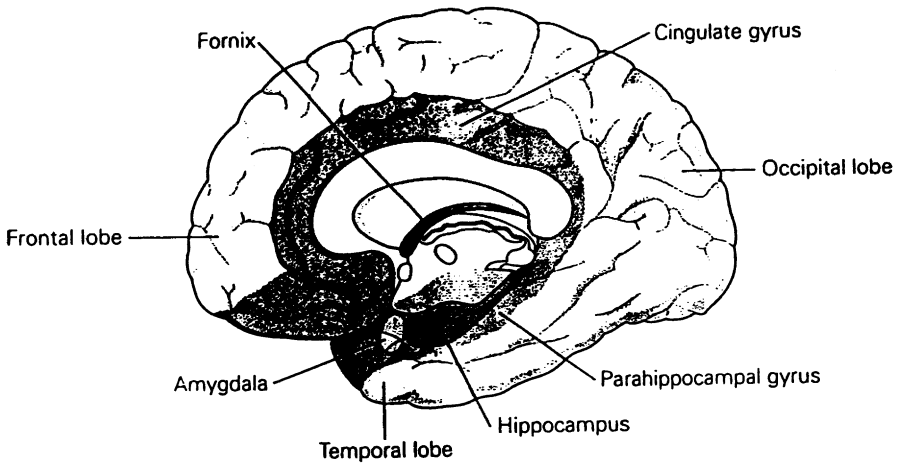


Figure 12 Illustration of the relationship of the limbic system to the cerebral hemisphere. Reprinted with permission from Kandel, E. R., Schwartz, J. H., and Jessell, T. M., Eds. (2000). *Principles of Neural Science*, 4th ed., p. 987. McGraw-Hill, New York.

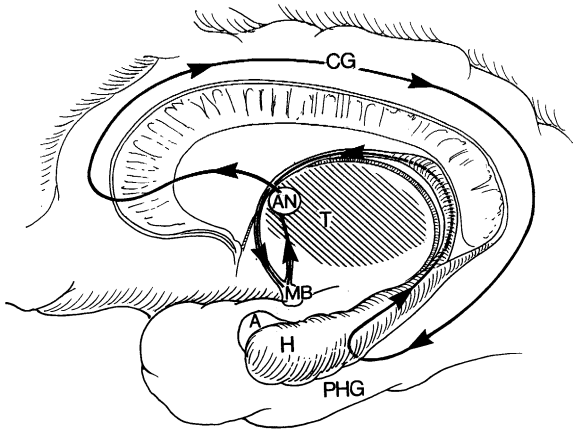


Figure 13 The Papez circuit shown schematically. The arrows indicate connections between components of the circuit. Abbreviations: A, amygdala; H, hippocampus; PHG, parahippocampal gyrus; CG, cingulate gyrus; AN, anterior nucleus of the thalamus; MB, mammillary body; T, thalamus. Reprinted with permission from Filley, C. M. (2001). *Neurobehavioral Anatomy*, 2nd ed., p. 27. University Press of Colorado, Boulder, CO.

In a parallel development, the hippocampus came to be understood as a key area for the processes of memory formation. The acquisition of declarative memory, which refers to facts and events as opposed to the skills that are associated with procedural memory, of course is vital for successful human existence. This capacity is apparently dependent on the hippocampus, a three-layered cortical region tucked deep within the medial temporal lobe and connected with other regions also concerned with recent memory, including the dorsal medial nucleus of the thalamus and the basal forebrain (Fig. 14). Whereas lesions in these latter regions have also been associated with memory dysfunction, the most dramatic demonstrations of recent memory disturbance have been observed in individuals with bilateral destruction of the hippocampus. The overlap of brain systems mediating emotions and memory is an intriguing neuroanatomic feature and suggests that those events with the most emotional significance are most likely to be encoded in memory; this aspect of mental life generally can be confirmed by common experience.

XI. WHITE MATTER

The white matter of the brain constitutes nearly one-half of its volume and serves the important general function of linking cortical and subcortical gray matter regions with each other. White matter consists of collections of axons ensheathed with myelin that are most often called tracts, but that may also be termed fasciculi, bundles, lemnisci, funiculi, and peduncles. Traveling extensively throughout the brain to link widely dispersed gray matter areas, these groups of fibers integrate cortical and subcortical regions into functionally unified neural networks. These networks subserve the many unique functions of the brain, from basic sensory and motor activities to complex cognition and emotion. By virtue of the dramatic increase in axonal conduction velocity that is afforded by myelin, rapid and efficient transfer of information across white matter tracts occurs, which enables the highest functions in the cerebral hemispheres. The many neurologic and neurobehavioral deficits sustained by individuals with the white matter disease multiple sclerosis (MS) testify to the importance of white matter in brain function.

White matter tracts in the brain can be considered under three general categories: projection fibers, commissural fibers, and association fibers. Projection fibers are those that ascend to the cortex from lower structures (corticopetally) or descend from the cortex to lower regions (corticofugally). Major corticopetal tracts are the thalamic radiations, relaying somatosensory information from the thalamus to the parietal cortex, and the optic radiations, projecting from the

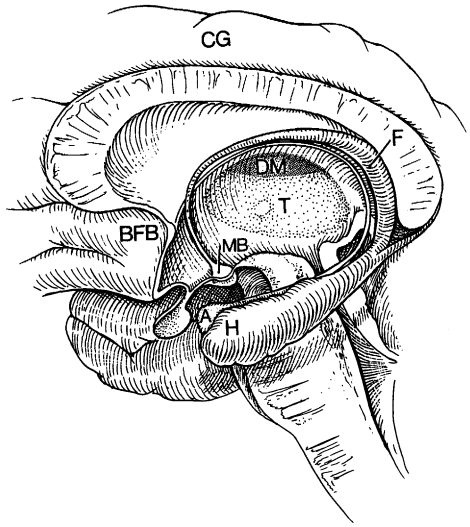


Figure 14 Cutaway illustration of the hippocampus, which lies deep in the medial temporal lobe. Abbreviations: H, hippocampus; A, amygdala; F, fornix; MB, mammillary body; DM, dorsal medial nucleus of the thalamus; BFB, basal forebrain; CG, cingulate gyrus. Reprinted with permission from Filley, C. M. (2001). *Neurobehavioral Anatomy*, 2nd ed., p. 59. University Press of Colorado, Boulder, CO.

lateral geniculate body of the thalamus to the occipital cortex. The major corticofugal tract is the corticospinal tract, which projects from the motor cortex to lower motor neurons in the spinal cord. A similar role is played by the corticobulbar tract, which also originates in the motor cortex but terminates on lower motor neurons in the brain stem. Both of these tracts first travel through the large internal capsule; corticobulbar fibers then cross to join various brain stem motor nuclei, whereas corticospinal fibers continue rostrally to the spinal cord, with most first decussating in the lower medulla. Projection fibers are therefore involved solely with elemental sensory and motor functions.

More important for behavior are the commissural and association fiber systems (Fig. 15). Commissural fibers are those that travel between the hemispheres in the cerebral commissures. The largest of these by far is the corpus callosum, a massive white matter tract that connects the four lobes of the brain on each side with their counterparts on the other. Much smaller commissural fiber systems include the anterior commissure and the hippocampal commissure. The association fibers join gray matter regions within each hemisphere. Among these, anatomists have distinguished two types: short and long association fibers. Short association fibers, also called arcuate or U fibers, connect adjacent cortical gyri and are found throughout the cerebrum. Long association fibers primarily link the lobes of the brain: these are the superior occipitofrontal fasciculus, the inferior occipitofrontal fasciculus, the arcuate fasciculus, the uncinate fasciculus, and the cingulum. All of these tracts have one termination in the frontal lobe, whereas their other terminus is variably in more posterior regions.

Two other white matter tracts deserve mention. The fornix is a prominent curved structure in the limbic system that connects the hippocampus and the mammillary bodies. The medial forebrain bundle connects the hypothalamus with both caudal and rostral brain regions and participates in hypothalamic control of the autonomic nervous system.

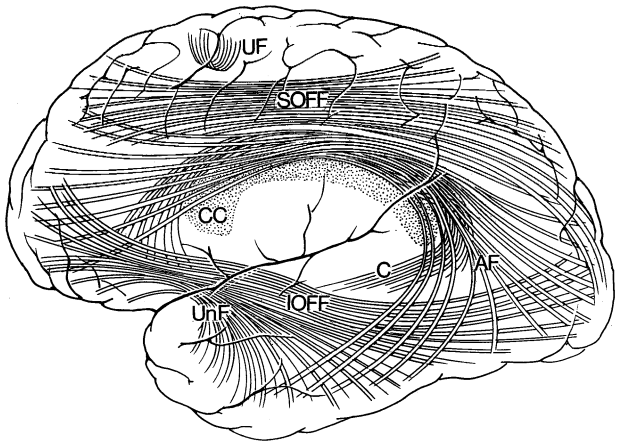


Figure 15 Diagram of white matter commissural and association tracts in the cerebrum. Abbreviations: CC, corpus callosum; UF, U fibers; SOFF, superior occipitofrontal fasciculus; IOFF, inferior occipitofrontal fasciculus; AF, arcuate fasciculus; UnF, uncinate fasciculus; C, cingulum. Reprinted with permission from Filley, C. M. (2001). *Neurobehavioral Anatomy*, 2nd ed., p. 185. University Press of Colorado, Boulder, CO.

XII. CEREBRAL CORTEX

Cortex is the Latin word for "bark," and the cerebral cortex is the outermost layer of the cerebrum. The surface of the brain is virtually entirely composed of cortex, which consists of a thin sheet of neurons averaging 3 mm in thickness. The number of cortical neurons in the brain is estimated at 25 billion, and the number of synapses in the cortex may be an astonishing 300 trillion. The computational power made possible by this extraordinary number of connections renders the cortex the site of the most advanced functions of the human brain. For the neuroscientific study of the mind and all of the capacities implied by this concept, an understanding of neocortical structure and function is indispensable.

At the microscopic level, the cortex is made up mainly of pyramidal cells, neurons named for the shape of their cell body, which each have a long axon projecting to other cortical areas or to subcortical sites. More than 90% of the cortex is a six-layered structure called neocortex, a term that refers to the relatively recent arrival of this brain structure in the course of evolution. Thus, the neocortex has a horizontally laminated pattern in which can be found, in sequence from the outermost layer inward, the molecular layer, the external granular cell layer, the external pyramidal cell layer, the internal granular cell layer, the internal pyramidal cell layer, and the multiform layer. There is also a vertical organization to the neocortex, such that cells arranged in a column perpendicular to the cortical surface are linked so as to respond as a unit to a given stimulus. Hundreds of millions of these vertical modules exist, and they are extensively linked to other modules by the axons of pyramidal cells.

The remaining 10% of the cortex is known as allocortex, which is made up in turn of paleocortex and archicortex, cortical types of ancient origin that are

more prominent in lower animals. Much of the allocortex is concerned with the olfactory system, which has limited importance in humans. However, as has been discussed, an important allocortical region is the hippocampus, a three-layered cortical region that, as part of the limbic system, plays a crucial role in memory and emotion. The three layers of the hippocampus are the molecular layer, the pyramidal cell layer, and the polymorphic layer.

Many attempts have been made by anatomists to divide the neocortex into discrete zones. The best known of these cortical parcellations is that of Korbinian Brodmann, a German anatomist who described about 50 areas of the cortex, each with distinct histological features (Fig. 16). Although it is not clear in many cases whether these anatomically distinct zones reflect functional distinctions, the cortical map of Brodmann has endured for almost a century, and reference to his numbered zones is customary in discussions of neocortical anatomy and function. Many of Brodmann's areas will be mentioned in the discussion that follows.

The neocortex forms the most superficial portion of all four lobes of the brain, and these lobes have much clinical utility in considering the functional affiliations of cortical regions and their underlying connections. Thus, the cortical surface can be divided into the frontal, temporal, parietal, and occipital lobes (Figs. 17 and 18). Each of these lobes has important functional affiliations (Table II), which will be discussed later. The frontal lobe is the most rostral of the four, lying anterior to the Rolandic fissure and superior to the Sylvian fissure. The temporal lobe is positioned inferior to the Sylvian fissure, and its posterior boundary is determined by the junction of two lines: one from the parietal–occipital junction to the preoccipital notch and the other running posteriorly from the end of the Sylvian fissure. The parietal lobe lies posterior to the Rolandic fissure, and its inferior margin is also defined by the two lines that form the posterior extent of the temporal lobe. The occipital lobe is situated posterior to both the temporal and parietal lobes. Another small neocortical region that is not visible on the surface of the brain is the insula, or island of Reil, which is concealed under the Sylvian fissure by portions of the frontal, temporal, and parietal lobes.

The frontal lobe is the largest of the four lobes, accounting for more than one-third of the entire cortical surface. As the most recently acquired cortical region in the course of evolution, the frontal lobe houses a variety of motor, cognitive, and emotional functions. First, the frontal lobe contains two areas important for voluntary movement: the precentral gyrus (Brodmann area 4), from which originate the corticospinal and corticobulbar tracts that descend to the spinal cord and brain stem, and the adjacent supplementary motor area (area 6). The remainder of the frontal lobe consists of the prefrontal cortex, a confusing term that simply means frontal association cortex not devoted to motor function. The prefrontal regions mediate many of the highest functions of the human brain. Among these advanced functions are the motor aspects of language: the fluency of language, for example, is mediated by Broca's area on the left (areas 44 and 45), whereas the expression of prosody, the emotional content of language, is provided by the homologous region of Broca's area in the right frontal lobe. Working memory, the capacity to hold information in mind long enough to apply it to the solution of a problem, is another important role of the frontal lobe, specifically the dorsolateral prefrontal cortex (areas 9 and 46). Working memory is thus an aspect of the general phenomenon of attention, which is also represented in the right hemisphere (see later discussion). In addition, working memory contributes to the concept of executive functions, a term meaning the ability to plan, carry out, and monitor a sequence of actions that culminates in the attainment of a goal. Executive functions are thought to be mediated by areas 8, 9, 10, 46, and 47. Another critical ability of the frontal lobe is in the realm of comportment, by which is implied the control of limbic impulses so that a socially appropriate behavioral repertoire is maintained. The orbitofrontal cortex (areas 11, 12, and 25) is extensively connected with the limbic system, and the normal function of this region prevents the dramatic behavioral disturbances that may occur with disinhibition. Finally, the medial frontal cortex (areas 24, 32, and 33), which corresponds essentially to the anterior cingulate gyrus, seems to have a role in motivation and the initiation of voluntary action.

The temporal lobe has a prominent role in hearing, language, memory, and emotion. Audition is represented at the cortical level by the primary auditory cortex (Heschl's gyrus, areas 41 and 42). From there, auditory input is relayed to an adjacent temporal region, the posterior part of area 22, for further processing. On the left, this is known as Wernicke's area, and it is here that the comprehension of spoken language is organized. The homologous region in the right temporal lobe mediates the comprehension of the emotional content of language, which has been called sensory prosody. The hippocampus, folded into the

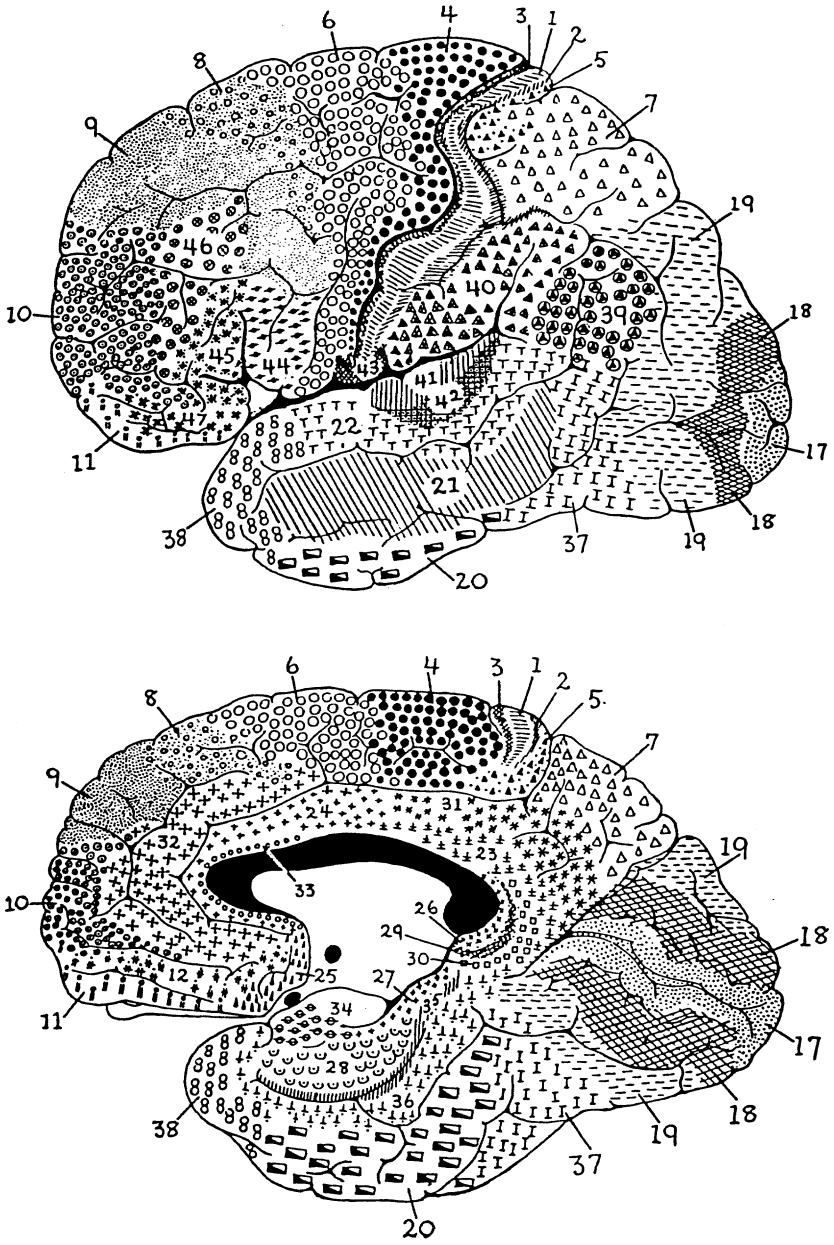


Figure 16 The cortical map of Brodmann, which distinguishes regions of the cortex on the basis of their microscopic features. Reprinted with permission from Filley, C. M. (2001). *Neurobehavioral Anatomy*, 2nd ed., p. 10. University Press of Colorado, Boulder, CO.

medial temporal lobe, is essential for the learning of new declarative information. The amygdala, immediately adjacent to the hippocampus, is a key nuclear structure in the cerebral organization of emotion. In a manner recalling hippocampal function, the amygdala is responsible for emotional learning; one powerful emotion, for example, is fear, and it is through amygdalar function that an individual learns to avoid dangerous environmental stimuli.

The parietal lobe has both sensory and cognitive roles in brain function. Immediately posterior to the motor cortex lies the primary sensory cortex (areas 3, 1, and 2), which receives sensory information from the thalamus below. Higher order sensory association cortex is located posterior to this region, in areas 5 and 7, and here further processing occurs. This arrangement permits the reception and interpretation of tactile stimuli from the opposite side of the body and face. On

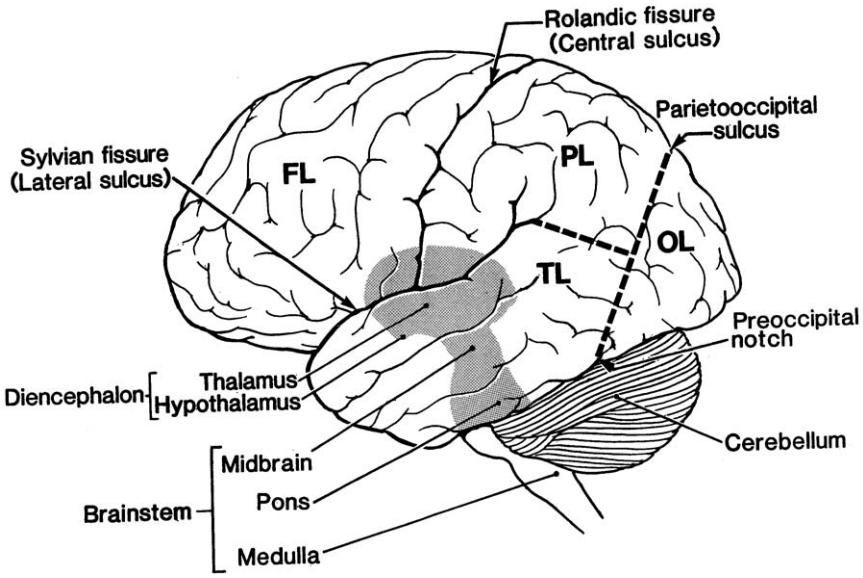


Figure 17 Lateral view of the brain showing the location of the four lobes of the hemispheres. Abbreviations: FL, frontal lobe; TL, temporal lobe; PL, parietal lobe; OL, occipital lobe. Reprinted with permission from Filley, C. M. (2001). *Neurobehavioral Anatomy*, 2nd ed., p. 6. University Press of Colorado, Boulder, CO.

the right, the parietal lobe is primarily involved in the capacity known as visuospatial function, which is the ability to orient within three-dimensional space and respond appropriately to it. The right parietal lobe is also central to the processes of attention, the ability to survey external space and focus on relevant stimuli while excluding those that are irrelevant; this lobe is the most prominent of a series of linked right cerebral

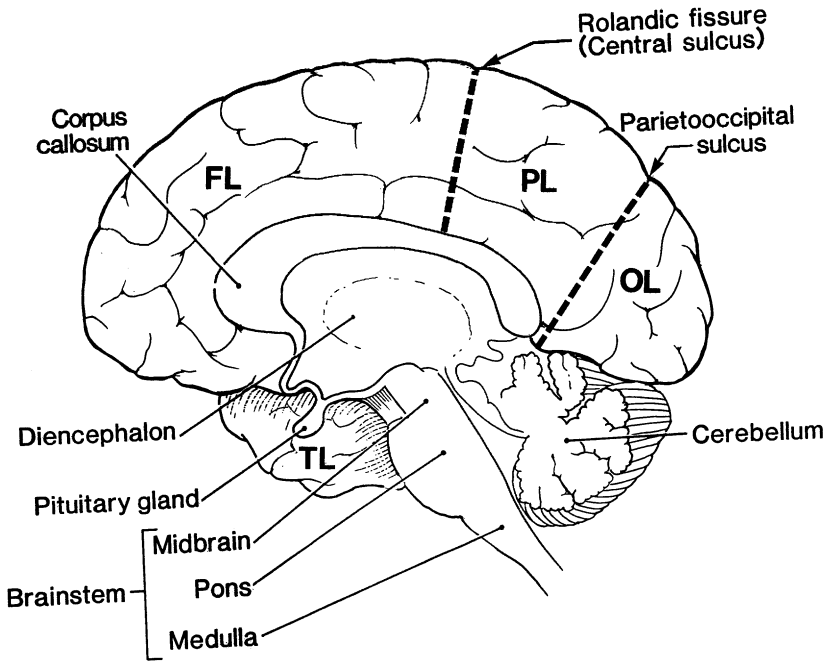


Figure 18 Medial view of the brain depicting the four lobes on the medial surface of the hemispheres. Abbreviations: FL, frontal lobe; TL, temporal lobe; PL, parietal lobe; OL, occipital lobe. Reprinted with permission from Filley, C. M. (2001). *Neurobehavioral Anatomy*, 2nd ed., p. 7. University Press of Colorado, Boulder, CO.

Table II
Functional Affilations of Cortical Regions

Frontal lobe
Voluntary movement
Language fluency (left)
Motor prosody (right)
Working memory
Executive function
Comportment
Motivation
Temporal lobe
Audition
Language comprehension (left)
Sensory prosody (right)
Memory
Emotion
Parietal lobe
Tactile sensation
Visuospatial function (right)
Attention (right)
Reading (left)
Calculation (left)
Occipital lobe
Vision
Visual perception

regions that renders the right hemisphere dominant for attention. On the left, important areas of the parietal lobe are the angular gyrus (area 39) and the supramarginal gyrus (area 40), which are involved in reading, calculation, and other cognitive domains.

Th occipital lobe is the smallest and most posterior lobe of the brain, but its importance lies in its singular affiliation with the visual system. Primary visual input from the thalamus is received here in the calcarine cortex of the medial occipital lobe (area 17), and then further processing occurs in the higher order visual association cortices adjacent to the calcarine cortex (areas 18 and 19). Reception of visual information is therefore as dependent on the visual cortex as it is on the eyes themselves. In addition, perception of visual aspects such as color, shape, form, motion, and location is made possible by the participation of the visual association areas.

A final point to be discussed in considering cortical structure and function is the specialization of the cerebral hemispheres. It is true that the two halves of the cerebrum are roughly symmetrical in a gross anatomic sense, and for elemental sensory and motor functions the two hemispheres can be considered comparable. However, there are significant interhemispheric differences in the organization of the higher functions. One of the most apparent of these is in the realm of language, as it has long been known that the left hemisphere is specialized for language in most individuals. This asymmetry of language function is linked to handedness, in that people who are right-handed (about 90% of the population) nearly always have left hemisphere dominance for language. However, it is important to recognize that the term ''dominance'' applies not just to language but to other cognitive functions as well. For example, the right hemisphere is dominant for visuospatial function and for attention, both of which are equally critical for successful human existence. Thus, the assumption that dominance only applies to language is misleading.

The two hemispheres, as is true of all areas of the brain, each contribute in their unique fashion to the extraordinary range of human capacities. These functions in general are becoming increasingly well-understood, and neuroanatomy continues to provide the foundation for understanding their neurobiological organization. From this basis, contemporary neuroscience is achieving an ever expanding knowledge of brain function, and, as further advances in neuroanatomy and related disciplines are made, a richer and more complete appreciation of the wide variety of human behavior can be expected.

See Also the Following Articles

BASAL GANGLIA • BRAIN STEM • CEREBELLUM • CEREBRAL CIRCULATION • CEREBRAL CORTEX • CEREBRAL WHITE MATTER DISORDERS • CRANIAL NERVES • LIMBIC SYSTEM

Suggested Reading

Brodal, P. (1981). *Neurological Anatomy*, 3rd ed. Oxford University Press, New York.

DeArmond, S. J., Fusco, M. M., and Dewey, M. M. (1989). *Structure of the Human Brain: A Photographic Atlas*, 3rd ed. Oxford University Press, New York.

Filley, C. M. (2001). *Neurobehavioral Anatomy*, 2nd ed. University Press of Colorado, Boulder, CO.

Kandel, E. R., Schwartz, J. H., and Jessell, T. M. (Eds.) (2000). *Principles of Neural Science*, 4th ed. McGraw-Hill, New York.

Nolte, J. (1999). *The Human Brain*, 4th ed. Mosby, New York.

Papez, J. W. (1937). A proposed mechanism of emotion. *Arch. Neurol. Psychiatry* **38,** 725–743.

Paxinos, G. Ed. (1990). *The Human Nervous System*, 3rd ed. Academic Press, San Diego.

Neurobehavioral Toxicology

MARCIA HILLARY RATNER,* ROBERT G. FELDMAN,*,†,§ and ROBERTA F. WHITE*,†,#

*Boston University Schools of Medicine and †Public Health and §Harvard Medical School, and #Boston Veterans Administration Hospital

GLOSSARY

cognitive domain Various modalities of behavior as expressed by the terms attention, executive function, memory (short-term and remote), visuospatial ability, motor function, and mood and affect.

critical effect The earliest discernible or measurable effect of exposure to a toxic substance.

critical organ The initial tissue to be affected by an exposure to a toxic substance.

encephalopathy A state of impaired functioning of the brain, which may include motor, sensory, and cognitive abnormalities and changes in mood and affect, as well as disturbances in attentiveness and consciousness.

executive function The behavioral process by which a person utilizes current sensory input and previously stored information and the complementary interactions of several cognitive domains to process and carry out complex independent, purposive, and problem-solving behaviors.

The adverse effects of exposure to neurotoxic chemicals on the brain manifest themselves behaviorally as changes in mood and in cognitive function. Neuropsychological testing has emerged as a means of documenting and measuring these changes. This methodology has several advantages for use in both the clinical diagnosis of chemically induced disorders and the epidemiologic investigation of neurotoxicant exposure–outcome relationships. These advantages include the documented validity and reliability of neuropsychological tests and their known sensitivity for detecting the effects of cerebral pathology and for localizing the probable specific anatomical sites associated with many neuropathological and neurodegenerative processes. In addition, methods of neuropsychological testing used to estimate premorbid intellectual abilities allow investigators to uncover acquired changes in cognitive functioning associated with brain insults, including the effects of exposure to neurotoxicants. The pathological effects following brain insults that may produce measurable deficits on formal neuropsychological testing may or may not be detectable by conventional magnetic resonance imaging (MRI), computed tomography (CT), and electroencephalography studies. Although the newer imaging technologies [i.e., positron emission tomography (PET), single photon emission computed tomography (SPECT), functional MRI (fMRI), and magnetic resonance spectroscopy (MRS)] can provide additional objective evidence of an organic basis of the neurobehavioral manifestations, formal neuropsychological assessment nevertheless remains the principal means for documenting impaired function among patients exposed to neurotoxicants. This article discusses the neurobehavioral syndromes related to neurotoxicant exposure and presents a review of the clinical diagnostic approaches most utilized in their evaluation.

I. DEFINITIONS AND ACTIONS OF NEUROTOXICANTS

A. Functional and Structural Effects

Many naturally occurring (neurotoxins) and synthetic chemicals (neurotoxicants) can cause damage to the

nervous system. Disruptions of neuronal cell homeostasis occur when natural protective mechanisms fail to detoxify and eliminate a potentially hazardous chemical substance before it causes tissue damage. A neurotoxic chemical is defined as a substance that is directly and/or indirectly capable of the following: (1) altering the integrity of nerve cell membranes, thereby affecting neuronal excitability, neurotransmitter release, and synaptic activity of neurons; (2) disturbing the flow of axoplasm, thereby interfering with the transport of neurotransmitters and nutrient substances along the axon to and from the cell body; (3) disrupting cellular respiration processes; (4) disrupting protein synthesis; (5) affecting neuronal functions indirectly by damaging Schwann cells and peripheral myelin, oligodendrocytes, and central myelin and/or disrupting the normal functioning of astrocytes and microglia; and (6) disturbing extracellular fluid volume and flow by damaging capillary endothelium, thereby resulting in disruption of the integrity of the blood–brain barrier (BBB) or blood–nerve barrier (BNB). Alterations in the normal functioning of the various affected target cellular elements, such as neurons, glial cells, myelin sheaths, or blood vessels, result in neurobehavioral manifestations.

Symptoms of exposure appear when tissue concentrations of a neurotoxic chemical reach a critical threshold level, above which intracellular processes such as oxidative respiration and axonal transport become impaired. Initially, reversible functional alterations occur; often these early effects are subclinical and may only be detectable by electroencephalography, evoked potential studies, or peripheral nerve conduction velocity testing. Irreversible damage to neural systems permanently interferes with function, resulting in impaired or disabled performance of ordinary activities of daily living.

B. Behavioral Manifestations

The severity of the behavioral manifestations associated with exposure depends on the potency and the dose of the particular neurotoxicant or neurotoxin. The total dose is dependent on the intensity and duration of the exposure. Acute exposures to higher concentrations of a neurotoxic chemical require less time than do more prolonged, lower levels of exposure to reach a threshold at which there is a clinical effect. The reversibility of a neurobehavioral change also depends upon the dose and type of neurotoxicant or neurotoxin. Chronic exposures are often associated with more gradual development of behavioral changes, which may be reversible or permanent.

Neurotoxicants and neurotoxins affect brain structures that mediate motor, sensory, and/or cognitive functioning. Many neurotoxic chemicals affect overall neurological and cognitive performance, and behavior to some degree, irrespective of their predilection for producing focal effects. Thus, diverse patterns of symptoms, signs, and neuropsychological performance deficits are not uncommon. Patients exposed to neurotoxic chemicals often complain of symptoms of encephalopathy, including changes in mood and affect, and of attention and memory problems. Overt signs of neurotoxicant exposure-induced motor system dysfunction may include spasticity, paralysis, bradykinesia, dyskinesia, dystonia, tremor, and incoordination. Subtle dysfunction of the motor system may not produce overt signs but may, nevertheless, affect performance on tests of motor function such as the Santa Ana Formboard, Purdue Pegboard, or Wechsler Digit Symbol Test. Neurotoxic chemicals affecting structures of the extrapyramidal motor system including the basal ganglia, such as manganese, produce overt signs of parkinsonism (e.g., tremor, dystonia, dyskinesia, and bradykinesia). Damage to the basal ganglia can directly and indirectly affect performance on neuropsychological tests of motor function. Cerebellar structures are sensitive to the effects of neurotoxic chemicals. These various patterns of clinical manifestations and neuropsychological performance deficits reflect an underlying neuropathology, which is dependent on the mechanism of action of the particular neurotoxicant, the exposure dose, and the type of exposure (acute or chronic) (see Table I).

II. DIAGNOSIS OF NEUROBEHAVIORAL DISORDERS

A person exposed to neurotoxic chemicals may be completely unaware of his or her own changes in behavior. This unrecognized impairment carries a risk for work-related accidents and injuries. Co-workers and/or family members often are the first to recognize changes in the patient's attention, memory, and mood and affect. If a source of chemical exposure is not immediately suspected or if the possible neurotoxic effects of a particular chemical are not well-known, the

Table I
Categories of Toxic Encephalopathies[a]

Disorder	Symptoms	Symptom duration	Pathophysiology
Acute Organic Mental Disorders			
Acute intoxication	Depression of CNS, attention and psychomotor deficits;	Minutes to hours	Reversible disruption of neutrotransmission
	No permanent sequelae		
Acute toxic encephalopathy	Confusion, seizures, coma; Cognitive deficits may persist	Hours to days	Hypoxia, disruption of BBB, cerebral edema
Chronic Organic Mental Disorders			
Organic affective syndrome	Disturbances of mood and affect including depression, irritability, anxiety, and fatigue;	Days to weeks	Reversible disruption of neurotransmission
	No permanent sequelae		
Mild chronic toxic encephalopathy	Disturbances of mood and affect plus cognitive deficits;	Weeks to years	Reversible disruption of neurotransmission
	Improvement with cessation of exposure but mild cognitive deficits may persist indefinitely		Limited neuronal/glial cell loss but no frank neuropathology
Severe chronic toxic encephalopathy	Disturbances of mood and affect plus severe cognitive deficits (including memory impairments) that interfere with activities of daily living;	Years	Cortical atrophy and/or loss of white matter in CNS; focal lesions may also be seen
	Deficits persist indefinitely		

[a]From White, R. F., *et al*. (1992).

patient's behavioral changes may be attributed to other possible neurological conditions. Risk of further exposure before removal from the source(s) of exposure is thus increased. The possibility that a neurotoxic illness may underlie the patient's presenting complaints should be fully investigated using the patient's medical history, laboratory findings, and occupational and environmental exposure histories.

The observations made by the clinician during a neurological examination can be used to infer the probable anatomical site(s) of nervous system dysfunction and to describe the patient's functional status. Abnormal neurologic symptoms and signs are expressions of impaired function or damage to particular neural structures, regardless of the specific etiology of the lesion. Thus, neurologic findings arising from the effects of exposures to neurotoxicants may resemble those found in primary or non-neurotoxic neurologic illness. The diagnostic process integrates the clinician's observations of the patient and the results of tests on physiological, anatomical, and behavioral functions, along with his or her acumen and judgment accumulated from experience with similar cases, and reference to a background of information contained in previously published literature.

A. Assessment of Behavior

The neurological examination begins with an assessment of the patient's ability to comprehend speech, follow simple instructions, perform complicated cognitive tasks, and perceive and identify sensory stimuli. The patient's spontaneous remarks and movements as well as those made in response to commands are evaluated for any digressions from expected norms of behavior, noting indications of the patient's orientation to person, place, time, and circumstances. Disturbances in mood and affect emerge during conversation. Engaging the patient in a conversation often reveals difficulties with attention if he or she is unable to focus well enough to follow the verbal exchange. Language comprehension is assessed by having the patient repeat simple phrases and follow simple commands. Expressive language is assessed by

listening to the patient's free speech for dysarthria, paraphasias, and neologisms. A clinical test of attention and mental control is to ask the patient to spell a word backward or to recite lists of words or numbers backward. Memory can be assessed through the details a patient is able to recall concerning medical history or through brief clinical assessment. Visuospatial and constructional ability can be assessed by having the patient draw the face of a clock with the hands set at a particular time.

Although these brief tests often reveal gross cognitive deficits associated with severe acute intoxication, acute toxic encephalopathy, and/or severe chronic toxic encephalopathy, patients with less severe or mild chronic toxic encephalopathy may present with subtle performance deficits that require formal neuropsychological assessment to detect and document. All patients exposed to neurotoxic chemicals who present with complaints of central nervous dysfunction should have formal neuropsychological testing to ascertain whether they are suffering from toxic encephalopathy.

B. Assessment of Motor and Sensory Functions

Neurological examination of motor functioning assesses the functional integrity of neurons in the cerebral cortex as well as their connections with subcortical, brain stem, cerebellar, and spinal cord pathways and the effector muscles that produce observable actions. Dysfunction in the motor cortex (upper motor neurons) results in weakness and spasticity of the limbs that are contralateral (i.e., on the opposite side of the body) to the site of the lesion. Basal ganglia dysfunction alters muscle tone and speed of response, causing bradykinesia (i.e., slowness of movement). Midbrain and brain stem structures control the coordination of cranial nerve functions such as conjugate eye movement, articulation of speech, and swallowing. Impaired cerebellar functioning results in ataxia of gait and tremulous trunk, head, and outstretched extremities. Tremors resulting from cerebellar and vestibular dysfunctions appear during actions and increase upon intention, whereas the tremor associated with Parkinson's disease appears during rest and disappears during action.

Spinal cord function is assessed by observing and recording the patient's gait, posture, muscle tone, fine motor control, and coordination. These features must be differentiated from those arising from damage to the brain. Reflexes are evaluated and muscle tone is noted. Spinal cord lesions produce motor dysfunction on the same side (ipsilateral) as the lesion. Weakness in the arm and leg indicates disturbance in the spinal pathways on the same side if a lesion exists below the foramen magnum. In such instances, muscle atrophy occurs as well. Loss of muscle tone and total paralysis without spasticity reflects dysfunction in the lower motor neurons.

Dysfunction in the structures of the brain responsible for processing sensory information can result in syndromes of contralateral neglect, visuospatial disturbances, or cortical blindness. Sensory deficits associated with exposure to neurotoxic chemicals can be detected by studies of brain stem auditory evoked responses and visual evoked responses. (see Section III.A). Impairments in sensory function indicate the probable location of disturbed anatomy in the spinal cord, thalamus, or sensory cortex. Sensory disturbances may be ipsilateral or contralateral to the site of the spinal cord lesion, depending on the specific sensory modality affected. Clinical assessment of sensory function is readily accomplished at bedside with a wisp of cotton, a new safety pin, a tuning fork, and a reflex hammer. Fibers carrying information about light touch, vibration, and position sensation ascend without crossing in the cord (these fibers cross in the medial lemniscus); thus, sensory dysfunction is found ipsilateral to the site of the cord lesion. Conversely, fibers carrying pain and temperature sensation cross immediately after entering the cord; thus, spinal cord lesions affecting these fibers produce sensory impairments on the side of the body contralateral to the site of the lesion. The loss of sensation to pain and temperature suggests dysfunction in the ventral spinal thalamic tracts (anterolateral system) of the spinal cord, whereas position and light touch represent anatomical structures in the dorsal columns of the spinal cord.

III. DOCUMENTATION BY FORMAL DIAGNOSTIC TESTS

A. Neuropsychological Testing

1. History Taking and Clinical Interview

The questions included in the interview should address information about the patient's current symptoms as well as his or her past medical history. This process can be facilitated by having the patient complete a standardized questionnaire such as the "Boston

Occupational and Environmental Neurology Questionnaire.'' Information about the patient's performance in school and work should be obtained. It should be determined whether the patient has had any previous psychiatric, neurological, or developmental–academic problems that could affect his or her performance on neuropsychological tests. The patient should also be questioned about the history of his or her immediate family members, including details about education levels and occupations as well as medical and psychiatric histories. Past and present recreational use of drugs, alcohol, and prescribed medications that may affect performance on neuropsychological tests should be noted. The particulars of the circumstances surrounding the alleged exposure to neurotoxicants must be defined and documented. Material safety data sheets and documentation of current and past exposure levels should be obtained whenever possible.

2. Methods of Neurobehavioral Assessment

The clinical approach to the neuropsychological assessment of cognitive deficits and behavioral changes attributable to exposure to neurotoxicants is essentially the same as that applied to any neuropsychological assessment situation. The clinician must be familiar with the expected or likely behavioral effects of the various neurotoxic chemicals found in the workplace and environment to which the patient may have been exposed. An experienced neuropsychologist will be able to carry out the differential diagnostic process of determining the most likely etiology of any deficits that may be revealed by testing.

The clinical neuropsychological evaluation of patients with possible toxic encephalopathy necessitates that the clinician perform a careful and thorough examination of each patient. Many areas of cognitive function must be assessed so that exposure-related effects can be detected and other possible diagnoses comprehensively evaluated and ruled in or out. Because there is overlap between the behavioral effects of exposure to certain neurotoxic chemicals and those associated with developmental disorders (e.g., learning disabilities, attention deficit disorder), psychiatric conditions (e.g., posttraumatic stress disorder, bipolar disorder), neurological diseases (e.g., multiple sclerosis, cerebrovascular disease, primary progressive dementia, parkinsonism), and the exposure to ethanol, medications, and illegal drugs, the test battery must allow for consideration of these alternative or contributing conditions.

The literature on behavioral toxicology includes the results of epidemiologic studies of exposed populations as well as clinical case reports. A number of batteries have been proposed for the assessment of neurotoxic effects of industrial chemicals. It is important to note that, whereas these batteries may be well-suited to the *epidemiological* investigation of neurotoxic effects, they are too brief and lack some tests essential to the process of making a clinical differential diagnosis. In addition, some of these batteries contain tasks with no available norms upon which a clinical diagnosis could be based. The results from epidemiological studies using these test batteries are nevertheless important to the neuropsychologist performing clinical case assessments. They can provide insight into the domains of function expected to be affected and unaffected by exposure to specific neurotoxicants and aid in the selection of appropriate tests expected to be sensitive to the effects of a particular toxic chemical.

There are fundamental differences between epidemiologic and individual clinical testing endeavors. In the epidemiologic setting, tests are designed or selected to ascertain dose–response relationships between the degree of exposure and neurobehavioral outcomes among members of a group of exposed persons. In the clinical setting, one is looking for deficits relative to the expected premorbid level of performance for an individual. In the epidemiologic setting, neuropsychological test scores within the ''normal'' range nevertheless may support an adverse effect of exposure based on dose–response outcomes. In the clinical setting, however, scores must be at least 1–2 standard deviations below normative expectation for the subject in order to be considered indicative of a deficit. Thus, clinical examinations may require a greater degree of dysfunction in order to conclude that an exposure is having (or has had) an effect than might be seen in the research setting.

The field of neuropsychology has a number of assessment approaches, including the Halstead–Reitan method, which typically employs a set battery of tasks, and the behavioral neurology or Process Approach, which employs a flexible battery in evaluating patients with specific kinds of referral issues or who show certain types of processing deficits during the evaluation itself. We have developed a battery of tests sensitive to the effects of neurotoxic chemical (Table II).

The neuropsychologist typically employs tests with which he or she is familiar and uses frequently to examine the toxicant-exposed patient. These tests are generally classified according to the cognitive domains

Table II
Boston Extended Neurotoxicologic Battery: Clinical Version[a]

Behavioral test	Cognitive domain assessed
WAIS; WAIS-R: WAIS-III	
Information	Basic academic verbal skills
Vocabulary	Verbal concept formation
Comprehension	Verbal concept formation
Similarities	Verbal concept formation
Digit Spans	Attention
Arithmetic	Attention and calculation
Picture Arrangement	Sequencing and visual spatial
Block Design	Visual spatial
Object Assembly	Visual spatial
Digit Symbol	Psychomotor speed
WMS-R	
Information	
Orientation	
Mental control	Attention and cognitive tracking
Digit Spans	Attention (verbal)
Visual Spans	Attention (visual)
Logical memories	Verbal memory
Verbal Paired Associates	Verbal memory
Visual Paired Associates	Visual memory
Visual Reproductions	Visual memory
Figural Memory	Visual memory
Continuous Performance Test	Attention (vigilance) and reaction time
Trail-Making Test (A and B)	Attention, cognitive tracking and sequencing
Wisconsin Card Sort Test	Set formation and set shifting
FAS-verbal fluency	Language
Boston Naming Test	Language
Boston Diagnostic aphasia Exam - reading comprehension subtest	Language
Wide Range Achievement Test	Basic academic skills
Boston Visuospatial Quantitative Battery	Visuospatial
Santa Ana Form Board	Psychomotor speed
Milner Facial Recognition Test	Visuospatial and visual memory
Benton Visual Rention Test	Visual memory
Difficult Paired Associates	Verbal memory
Albert's Famous Faces Test	Retrograde memory
Profile of Mood States	Affect
Minnesota Multiphasic Personality Inventory	Personality and affect

[a]From White, R. F., and Proctor, S. (1995). In "Neurotoxicology: Approaches and Methods" (L. W. Chang and W. Slikker, eds.), Ch. 46, Academic Press.

they tap. Though no test is so pure that it taps only one type of cognitive processing skill or functional domain, many load heavily on one area of function or another. The most commonly assessed functional domains in neuropsychology include attention, executive function, fine manual motor skills, visuospatial abilities, language and verbal skills, anterograde (short-term) and retrograde memory, and affect and personality. Regardless of the battery of tests used, it is essential that the clinician determine an estimate of premorbid ability patterns for the patient. Any deficits uncovered on neuropsychological testing should be related to this estimate of baseline function. Neuropsychologists are typically interested in academic skills such as reading and arithmetic (which are often helpful in determining premorbid ability patterns) and in motivation to perform well on testing which can be influenced by aspects of secondary gains such as monetary compensations. The reader is referred elsewhere for descriptions of the tasks assessing these domains. The various domains mentioned earlier as well as the expected changes in performance associated with exposure to neurotoxicants are described next to provide additional insight into the neurobehavioral features of toxic encephalopathy.

a. Language–Verbal Function Language and verbal functioning are typically preserved in adults exposed to neurotoxicants. This aspect of cognitive function is relatively resistant to the effects of neurotoxic exposure-induced brain damage compared with dynamic cognitive processes, such as encoding of new memories. This is in stark contrast with the effects of stroke and other focal lesions, which may have profound impacts on language function. However, patients exposed to certain neurotoxic chemicals (e.g., carbon monoxide) may show deficits on tests requiring the application of verbal and language skills. Motor aspects of writing may be affected in those patients with movement disorders (e.g., tremor) resulting from neurotoxic exposure, whereas the grammatical aspects of writing remain intact. Exposure to neurotoxic chemicals is more likely to produce language deficits in children than in adults because disruption of encoding processes occurs during development and, thus, can lead to problems with language acquisition. The severity of the deficits seen depends upon the age of the child at the time of exposure; younger children are more vulnerable.

b. Attention and Executive Function Deficits in *attention* and *executive function* may be found on

formal testing of patients with exposures to neurotoxicants. Tests of attention measure the following: (1) *simple (immediate) attention*, i.e., how much information can be grasped at once; (2) *divided attention*, which is the ability of an individual to attend to more than one task simultaneously; and (3) *vigilance or sustained attention*, which measures the ability of the subject to remain focused on a single task for long durations of time. Attention deficits impair the patient's ability to selectively focus on relevant stimuli and, therefore, may have a direct effect on concentration and an indirect effect on memory function.

Executive function is a higher order behavioral process involving the ability of the subject to appreciate and respond to complex changes in neurobehavioral task demands, including recognizing, maintaining, and shifting set as necessary to carry out such tasks. *Cognitive tracking* is an aspect of executive function often found to be impaired in those patients exposed to neurotoxic chemicals. Trail making (Trails A and B) is a cognitive tracking task that has been shown to be sensitive to problems associated with exposure to neurotoxic chemicals. *Cognitive flexibility* is another aspect of executive function that is affected in toxic encephalopathy. Difficulties in cognitive flexibility may be revealed by tasks such as the Wisconsin Card Sorting Test. Although some patients show deficits on tests of both cognitive tracking and cognitive flexibility, other patients exhibit deficits on one but not the other. It is impossible to predict which type of deficit will be seen in a patient with toxic encephalopathy, and it does not appear to be related to the type or severity of the exposure. Patients with severe deficits in executive function may have problems with their activities of daily living.

c. Memory Function Memory can be affected at several levels, including encoding of new information, retrieval of encoded information, ability to inhibit interference during learning and retrieval, and retention of encoded information. Many patients exposed to neurotoxic chemicals have relative deficits on tests of short-term or anterograde memory function compared with retrograde memory or long-term memory function. This dichotomy reflects the sensitivity to neurotoxic exposure-induced brain damage of the complex dynamic processes involved in short-term memory function, particularly encoding processes, and the relative resilience of the previously stored information tapped by tests of retrograde memory function, which is considerably more dependent upon retrieval mechanisms. Because of the divergent patterns of memory impairment possible following toxic exposure, a rather extensive memory battery is used to assess these patients.

d. Motor Skills Motor function is affected in some patients exposed to neurotoxic chemicals. Patients may show performance deficits on tests sensitive to motor deficits, including Finger Tapping and the Santa Ana Formboard. Motor function deficits can interfere with the patient's performance on many neurobehavioral tests (e.g., Digit Symbol). For example, a patient with bradykinesia may perform poorly on timed tasks measuring performance in domains other than motor function because they simply move more slowly. The neuropsychologist using the behavioral neurology or Process Approach can take into consideration the effects of motor deficits on other neurobehavioral tasks and utilize this information in his or her assessment and interpretation of the patient's performance.

e. Visuospatial Abilities Visuospatial deficits are seen following exposure to certain neurotoxicants (e.g., mercury). Performance on tests such as Block Design and the Rey–Osterreith Complex Figure (copy trial) may be below expectation in a patient exposed to neurotoxicants.

3. Strengths of Neuropsychological Tests

Neuropsychological test procedures have some inherent strengths and weaknesses with regard to making clinical diagnosis.

A. They are reliable because they have been standardized with regard to scoring and administration. For this reason, they can be given in the same manner by different clinicians.

B. Normative values are available for these tests so that performance can be judged with regard to level of performance. These norms allow the clinician to compare the patient's performance to that evidenced by persons of the same age and, often, of the same gender and educational achievement. For some tests, norms specific to countries outside of the United States are available, as are norms for specific groups within the U.S. population (e.g., Hispanics).

C. The tests have been well-validated, and a great deal is known about how performance on one test relates to performance on similar tests. In addition, extensive information is available concerning the brain structures that participate most directly in the completion of tasks (brain–behavior relationships revealed

by the tasks) and the patterns of performance on the tests among patients with specific developmental, neurological, motivational, and psychiatric disorders.

4. Weaknesses of Neuropsychological Tests

A. In persons with low premorbid intellectual abilities, it can be difficult to identify subtle deficits in function because they are performing at the floor of the tests already (many tasks have high floors).

B. When there is a long delay between exposure and neuropsychological testing, physiological recovery and/or use of compensatory strategies by the patient may occur, obscuring changes in function that may have been evident at the time of exposure. Among patients whom we have seen longitudinally from the time of exposure until several years after exposure ceased, it is clear that for some cases we would not have been able to diagnose the toxicant-related effects or determine the minor residuum of those effects had we not seen the patient early.

C. For some subgroups of patients within the U.S. population, particularly immigrants who do not speak English as a first language, normative values may be unavailable or available only for a limited number of tests.

D. In patients for whom sick-role playing or embellishment is an issue, inconsistencies in test performance or exaggeration of performance deficits may obscure mild or subtle brain dysfunction associated with exposure to toxic chemicals.

E. Conditions with overlapping pathologies can make differential diagnosis difficult. For example, patients with white matter lesions may have multiple sclerosis, a toxicant-induced condition, or both. Similar problems hold for parkinsonism affecting basal ganglia function, which can occur secondary to toxicant exposure, infections, infarcts, or a combination of these brain insults. Cerebrovascular disease is also a problem because it presents with a frontal–subcortical picture on neuropsychological testing similar to that of many toxic chemicals. Furthermore, exposure to certain neurotoxicants (e.g., mercury) that may be associated with hypertension can lead to brain dysfunction.

B. Neurophysiological Testing

Neurophysiological tests of central nervous system functioning, such as electroencephalography and evoked potentials, are obtained to provide additional objective evidence of an organic basis for the clinical behavioral findings. These tests can be used to detect possible diseases other than the disorder attributable to exposure to neurotoxicants. Abnormalities are not always revealed by these tests, but, when present, they may lend further support to a diagnosis of toxic encephalopathy.

1. Electroencephalography

The electroencephalogram (EEG) seen in a healthy adult has a relatively consistent pattern under normal waking, resting, and sleeping states. Disruptions of normal brain electrical activity are associated with changes in the symmetry, amplitude, and/or frequency of the EEG patterns. For example, epileptic activity is associated with a paroxysmal quality to the waveforms and sharp, spiked discharges. A focal concentration of sharp, slow, and/or paroxysmal waves in a particular area may indicate an underlying structural lesion, such as a neoplasm, stroke, or traumatic brain injury. Marked asymmetry of the EEG pattern suggests lateralized pathology. In contrast, encephalopathy due to endogenous or exogenous toxic disturbances is associated with diffuse bilateral slowing of the background rhythm and the disappearance of normal resting frequencies. As with all laboratory tests, the significance of the EEG report depends on correlation with other clinical information. In the differential diagnosis of neurotoxic syndromes and non-neurotoxic, neurological disease, the EEG is most helpful when an abnormality is seen during or in close chronological proximity to neurotoxicant exposure because the EEG pattern normalizes in reversible acute encephalopathy when the patient is removed from the source of exposure. The EEG tracing may not return to normal after the patient is removed from exposure to certain neurotoxicants (e.g., trimethyltin) and, therefore, is useful in documenting the presence of organic changes attributable to exposure.

2. Evoked Potentials

Evoked potentials (EPs) are used to assess the integrity and function of sensory pathways and are recorded after stimulation of peripheral sensory (afferent) nerve fibers in the extremities and/or by direct recording over the dorsal columns of the spinal cord. The EPs consist of visual evoked potentials, brain stem auditory evoked potentials, and somatosensory evoked potentials.

Visual evoked potentials (VEPS) are used to assess the integrity of the pathways of the optic system, including the optic nerve and chiasm, the optic tract to the geniculate nuclei, and the calcarine cortex. Two types of VEPs, flash visual evoked potientals (FVEPs) and pattern shift visual evoked potentials (PSVEPs), have been used to study the effects of exposure to neurotoxic chemicals.

Brain stem auditory evoked potentials (BAEPs) arise following stimulation of the auditory nerve. The auditory pathway generates a complex waveform that is recorded through electrodes attached to the patient and projected on the screen of an oscilloscope. Wave I reflects activation of the auditory nerve, whereas waves II and III reflect the activation of structures in the pontomedullary region. The sources of waves IV and V are less clearly defined but appear to be related to functions of the upper pons and lower midbrain. The absolute latencies of each wave are recorded, but interpeak latencies of waves I–III, III–V, and I–V are more consistent and reproducible and, therefore, are utilized in clinical testing. Brain stem auditory evoked potentials can be used to detect insults caused by ototoxic substances.

Somatosensory evoked potentials (SEPs) are recorded from electrodes placed over the sensory cortex after activation of a peripheral sensory or mixed nerve. The stimulus is conveyed centrally in the spinal cord and is then projected to the contralateral cerebral cortex. Technique is important in obtaining reliable measures. SEPs are usually tested in both upper and lower extremities, and interpeak latencies are more consistent and can help to localize pathology along the path of the peripheral nerve through the spinal cord, brain stem, and thalamus to the cortex.

Because many neurotoxic chemicals affect peripheral nerves, it is commonly the most distal sites of sensory conduction that are slowed and affect the cortically evoked SEPs in patients exposed to neurotoxic chemicals. Therefore, it is uncertain whether SEPs offer any advantage over standard nerve conduction velocities, except for studies of conduction through the spinal cord posterior columns and in instances where proximal nerve blocks or asymmetrical problems are being considered in the differential diagnosis.

C. Neuroimaging Studies

Highly sophisticated, computerized radiological methodologies can delineate images of the cerebral hemispheres and ventricular systems, as well as differentiate between cerebral cortex and white matter structures. The sensitivity of neuroimaging techniques in the detection of neurotoxic effects is limited. Nonetheless, there have been reports of clinical correlations between images obtained with magnetic resonance imaging (MRI), single photon emission computerized tomography (SPECT), and positron emission tomography (PET) in the presence of exposure to various neurotoxicants.

1. Magnetic Resonance Imaging

Magnetic resonance imaging studies are invaluable in the differential diagnosis of various encephalopathies. These studies often demonstrate structural abnormalities due to strokes and neoplasms; MRI studies of severe acute and chronic toxic encephalopathy may reveal white matter changes. However, the MRI studies of patients with less severe chronic toxic encephalopathy may appear normal, despite clinical neurological and neuropsychological evidence of behavioral abnormalities and performance deficits.

In cases of hypoxia or carbon monoxide poisoning, the MRI findings show signal intensity changes in the basal ganglia (globus pallidus) that may be bilateral or unilateral. Changes in cerebral white matter are seen upon brain MRI following chronic exposure to toluene; these findings have been correlated with changes in neuropsychological performance. MRI changes are also associated with exposure to organolead compounds and inorganic mercury.

2. Magnetic Resonance Spectroscopy

Magnetic resonance spectroscopy (MRS) has emerged as a possible sensitive measure of the structural and functional abnormalities associated with central nervous system dysfunction. It is an advanced level of the technology used in conventional MRI, which can detect chemical characteristics in addition to image data. Three main peaks reflecting the concentrations of *N*-acetyl-L-aspartate (NAA), creatine–phosphocreatine (Cr), and choline-containing compounds (Cho) are recorded from selected areas of interest. NAA is present within all neurons, and its concentration is elevated in several degenerative neurological conditions including amyotrophic lateral sclerosis. The Cr peak seen on MRS reflects levels of creatine and phosphocreatine, which serve as a reserve for high-energy phosphates in the cytosol of neurons. The Cho peak represents choline-containing compounds.

Choline is a precursor for the neurotransmitter acetylcholine and for the membrane constituent phosphatidylcholine. Additional chemicals of interest detectable with MRS include lactate, glutamate, glutamine, myoinositol, and γ-aminobutyric acid (GABA). Lactate is an end product of anerobic respiration and may be elevated following exposure to hypoxia-inducing neurotoxicants such as carbon monoxide and hydrogen sulfide.

The value of MRS in the diagnosis of exposure to neurotoxicants has not been fully elucidated, but published reports indicate that this technology holds considerable promise as a diagnostic tool. For example, MRS studies of a 10-year-old boy who had documented elevated blood lead levels of 51 μg/dl at 3 years of age revealed spectra that deviated from the expected pattern in all metabolite ratios analyzed, with a reduction in the NAA : creatine ratio for both gray and white matter. The conventional MRI of his brain done at the same time was normal. Formal neuropsychological assessment of the child was also performed at this time and revealed deficits in attention and mental control. The child also had reading, writing, and linguistic performance deficits. In contrast, his scores on tests of general knowledge were within normal limits. These neuropsychological and MRS findings were in stark contrast with those obtained for his cousin, a 9-year-old boy who was not exposed to lead and who served as a control. The neuropsychological assessment of the cousin was within normal limits and no abnormalities were seen on his MRI or MRS studies. Furthermore, the MRS findings in the patient's unexposed cousin were entirely consistent with the spectral pattern reported in previous studies of healthy individuals.

IV. BEHAVIORAL MANIFESTATIONS FOLLOWING EXPOSURE TO SELECTED CHEMICAL NEUROTOXICANTS

A. Lead

The acute symptoms of exposure to lead include abdominal colic, constipation, anorexia, vomiting, headaches, lightheadedness, dizziness, forgetfulness, anxiety, depression, irritability, excessive sweating, and muscle and joint pain. Acute symptoms subside following cessation of exposure and the reduction of blood lead levels. Chronic lead exposure results in more persistent neurological manifestations, including peripheral neuropathy and encephalopathy. Encephalopathy is the most serious consequence of acute and chronic lead poisoning in adults and children. Seizures and coma are seen in both children and adults with inorganic lead encephalopathy, and death occurs in the most severe cases. Early recognition of the symptoms and signs of lead poisoning can minimize the neurotoxic effects of lead exposure and prevent permanent brain damage or death.

The neuropsychological testing of lead exposed children reveals impaired performance on tests of memory, visuospatial abilities, and concept formation. Meta-analysis of the published data on the effects of lead on the brain suggests that children's IQ scores are inversely related to their lead body burden. Residual cognitive deficits and behavioral disturbances have been reported in middle-aged adult survivors of childhood lead poisoning. Language function is typically spared in lead-exposed adults but is often impaired in those individuals exposed to inorganic lead as children.

We previously reported the results of comprehensive neuropsychological assessment of 18 adults (7 male, 11 female) who had been exposed to lead as children 50 years earlier. All of the exposed subjects were under the age of 4 years at the time of exposure; the mean age of the exposed subjects at the time of neuropsychological testing was 54.4 years. Eighteen age- and sex-matched subjects with no history of lead exposure served as controls. All of the exposed subjects exhibited lead lines (metaphyseal bands) upon the X ray of at least one long bone and had presented during childhood with symptoms consistent with overt lead poisoning, indicating that blood lead levels in all cases had exceeded 60 μg/100 ml. The neuropsychological test battery used included four subtests from the *Wechsler Adult Intelligence Scale-Revised* (*WAIS-R*), including Similarities, Vocabulary, Picture Completion, and Block Design, the *Wechsler Memory Scale* (*WMS*), Trail Making Tests A and B (attention and visuomotor functioning); controlled word association test (verbal fluency), *Raven Progressive Matrices* (nonverbal reasoning), finger tapping (motor speed), and the *Profile of Mood States* (*POMS*).

Results of neuropsychological testing revealed significant deficits on the Picture Completion and Logical Memories subtests of the *WAIS-R* and *WMS*, respectively. In addition, the lead-exposed subjects showed impaired performance on the *Raven Progressive Matrices* and Trail B. These findings are consistent with persistent impairment of function in the cognitive domains of attention, executive function, concept formation, and short-term memory. Furthermore,

these findings indicate performance deficits on complex cognitive tasks that may impede the exposed individual's ability to learn new information. Whereas such subtle cognitive deficits may not be apparent during most activities of daily living, they can interfere with the exposed individual's academic and occupational achievements and advancement. Subtle cognitive deficits such as those reported in this study are particularly significant among those individuals whose premorbid IQ was below average.

Neuropsychological testing can also be used to detect and document the subclinical and clinical central nervous system effects of inorganic lead poisoning in persons exposed to lead as adults. Neuropsychological assessment of asymptomatic workers with blood lead levels greater than 50 μg/100 ml reveals cognitive deficits on tests of psychomotor and memory functioning. Although comparisons of responses of workers on the *Profile of Mood States* (*POMS*) before and after reductions in levels of lead exposure indicate that tension, anger, depression, fatigue, and confusion are decreased by reductions in exposure, neurobehavioral performance deficits may not be entirely ameliorated by a similar reduction in lead exposure levels.

We examined a 31-year-old woman who had personally undertaken the job of deleading her recently purchased home. She developed symptoms indicative of inorganic lead encephalopathy after approximately 9 months of exposure. Her blood lead level when she first presented was 146 μg/100 ml of whole blood. Neuropsychological testing revealed poor attention and impaired memory for new material, especially visual information. Specific tests revealing impairments of central nervous system function included Digit Span, Digit Symbol, and Block Design. We reported on a similar pattern of performance in a 19-year-old man who had worked as a professional deleader for approximately 10 months. His lead levels ranged from 70 to 100 μg/100 ml of whole blood. He sought medical attention because of problems with attention and memory. Neuropsychological testing revealed impaired performance on tests of attention, short-term memory, and visuospatial and visuomotor functioning. Specific tests revealing impaired functioning included Digit Span, Digit Symbol, Block Design, and Picture Arrangement. Performance on tests of language and vocabulary were within expectation.

The trialkyl metabolites of the organolead antiknock agents tetramethyllead and tetraethyllead (i.e., trimethyllead and triethyllead) are potent neurotoxicants. Very few studies have specifically evaluated the persistence and/or severity of the cognitive effects associated with chronic exposure to organic lead in humans. The severity of the encephalopathy associated with organic lead poisoning has been related to the patient's body burden of inorganic lead as reflected by measurement of blood lead levels.

The findings on formal neuropsychological testing of patients with organic lead encephalopathy may be correlated with the findings on EEG studies. Both of these markers of neurotoxicant effect may show improvement after cessation of exposure, and these findings may be further improved with chelation therapy. Cognitive functioning was assessed in a 41-year-old woman who presented with symptoms of organic lead poisoning (blood lead level = 110 μg/100 ml) and an 8-month history of sniffing leaded gasoline for its euphoric effects. Neuropsychological testing revealed her to be well-oriented with intact remote memory, but her attention span and short-term memory functioning were severely impaired. The patient was unable to perform simple oral arithmetic problems, indicating deficits in attention and working memory. In addition, her performance on delayed recall tests of short-term memory indicated that she had severe deficits in this cognitive domain as well. Her attention and memory as well as other symptoms of lead poisoning were improved after chelation therapy.

Neuropsychological functioning has been assessed in a group of 39 organic lead manufacturing plant workers with a mean exposure duration of 14.7 years. The mean lifetime blood lead level among these workers was 26.1 μg/100 ml. Eighteen (46.2%) of them had neuropsychological performance deficits on tests of attention, memory, and psychomotor function. These 18 workers all underwent additional testing to rule out the existence of metabolic, infectious, or structural etiologies for their performance deficits. No alternative explanation could be found for the neuropsychological deficits seen in any of these workers.

The associations between neurobehavioral functioning and tibial bone lead and chelatable lead levels in former organic lead workers have also been investigated. Higher peak tibial lead levels were significantly associated with poorer performance on the *Wechsler Adult Intelligence Scale-Revised* vocabulary subtest, serial digit learning test, *Rey Auditory–Verbal Learning Test*, Trails B, finger tapping, Purdue pegboard, and the *Stroop Test*. Chelatable lead concentrations were significantly associated with choice reaction times. These findings suggest that past exposure to organic lead may be associated with persistent neurobehavioral performance deficits,

particularly in the domains of manual dexterity, executive function, verbal intelligence, and verbal memory, and that the severity of these deficits is related to the peak tibial lead levels that occurred during the exposure period.

B. Mercury

Symptoms of acute exposure to elemental mercury vapor include respiratory irritation, headache, fever, chills, chest pain, general malaise, nausea, and vomiting. These symptoms are often referred to as "metal fume fever syndrome." Emotional lability, depression, social withdrawal, tremors, delirium, and coma develop within 24 hr after exposure to elemental mercury. The respiratory symptoms typically resolve within days to weeks after cessation of exposure, but the CNS disturbances persist.

The tremor associated with mercury poisoning begins in the fingers and hands, then progresses to affect the eyelids and face, and eventually affects the head, neck, and torso. The tremor is rapid, may be quite severe, and is accentuated by activity and emotional excitement, increasing in amplitude of excursion upon activation. The tremor is not a resting tremor and thus its frequency is faster and differs from the characteristic resting pill-rolling tremor seen in Parkinson's disease. Continued exposure to mercury vapor results in worsening tremor, gingivitis, mood changes, withdrawal from social interactions, greater memory loss. Abnormalities also include deficits in attention, executive functioning, short-term memory, and visuospatial ability. Language, basic academic skills, and retrograde memory are usually unaffected.

Neurobehavioral effects of acute exposure to elemental mercury vapor include depression, anxiety, and social withdrawal. Assessment of cognitive and emotional functioning conducted after cessation of exposure in workers exposed elemental mercury vapor may reveal persistent cognitive deficits on tests of motor coordination, processing speed with and without a motor component, cognitive flexibility, verbal fluency, verbal memory, and visual problem-solving and conceptualization.

Adults chronically exposed to elemental mercury may demonstrate cognitive impairments, including deficits in attention, executive function, short-term memory, visuospatial ability, and motor function. Language function and long-term memory are typically spared in adults but may be impaired in persons exposed *in utero* or during childhood. A case study of a 19-year-old man with a history of chronic exposure to mercury vapors during childhood (ages 4–9 years) revealed persistent tremor and behavioral abnormalities indicative of developmental toxin-induced encephalopathy secondary to mercury exposure, including deficits on tests of language, executive function, visuospatial skills, and fine motor control. The *POMS* revealed irritability and depressive affect.

The neuropsychological effects of mercury exposure can be correlated with current and cumulative exposure doses. For example, Digit Span has been used in research settings to document a correlation between current urine mercury levels and impairments of short-term memory function among exposed persons. The Bender–Gestalt test has also revealed impairments of visuospatial skills in persons with elevated tissue mercury levels. Impaired short-term memory functioning has been correlated with the duration of exposure to mercury. The serial neuropsychological assessment of construction workers exposed to elemental mercury revealed acute and persistent CNS effects. Initial testing revealed impaired performance on tests of attention and executive functioning (Trails A and B, *Stroop Test*) and motor skills (finger tapping, grooved pegboard). Performance was correlated with cumulative excretion of mercury. Performance on Trails A and B was improved following chelation therapy. However, performance on finger tapping and the *Stroop Test* remained unchanged and performance on the grooved pegboard worsened after cessation of exposure. Thus, our experience and a review of the literature suggest that elemental mercury exposure can be expected to affect performance on tests of attention, executive functioning, and motor functioning.

Persistent impairments in short-term memory, coordination, and simple reaction time may be seen up to 10 years after removal from exposure to elemental mercury. Research studies comparing the neuropsychological functioning of former chloralkali workers with that of age-matched unexposed controls with similar educational levels reveal that performance on certain tests, including the grooved pegboard and *Benton Visual Retention Test*, is worse among those previously exposed to mercury. Formerly exposed subjects may also perform worse on other tests sensitive to attention and psychomotor function, such as Trails A and B and Digit Symbol.

The clinical picture of methylmercury encephalopathy differs somewhat from that associated with

exposure to elemental mercury. Exposure to methylmercury during critical developmental periods can lead to severe mental retardation. Neuropsychological development was reported to be so severely delayed in an individual exposed to methylmercury *in utero* during the Minimata outbreak that she never learned to speak. Adult exposure to methylmercury is associated with lesions in the occipital cortex, and performance deficits may be seen on tests of visuospatial ability. Deficits have also been reported on tests of manual dexterity in adult women exposed to methylmercury.

C. Manganese

Occupational exposure to manganese (Mn) occurs among miners and welders. Frequently encountered organic Mn compounds include methylcyclopentadienylmanganese tricarbonyl (MMT), which is used as an antiknock additive in gasoline. Encephalopathy and basal ganglia dysfunction with parkinsonian signs including rigidity, gait abnormalities, dysarthria, hypomimia, and bradykinesia have been reported in patients with increased brain Mn levels due to chronic liver failure. The similarities between the clinical manifestations associated with Mn intoxication and those seen in Parkinson's disease suggest that occupational and environmental exposures to this and other neurotoxic chemicals may be involved in the pathogenesis and/or alter the prognosis of certain neurodegenerative diseases. It has not been established whether exposure to Mn precipitates the occurrence of progressive idiopathic Parkinson's disease, but research suggests that the prognosis may be poorer among those persons who are predisposed to develop idiopathic Parkinson's disease who also have a history of exposure to Mn.

The initial clinical manifestations of manganese poisoning often include behavioral changes referred to collectively as "manganese psychosis." The clinical symptoms include mood changes, emotional lability, uncontrolled laughter, and hallucinations. Performance on formal neuropsychological tests may be impaired at this time. Motor disturbances, characterized by tremor, dysarthria, gait disturbance, slowness and clumsiness of movement, and postural instability, emerge with continued exposure to Mn. If exposure continues, the psychosis typically subsides, and dystonia and an awkward high-stepping gait emerge. The gait disturbance associated with Mn poisoning is easily distinguishable from the shuffling gait of idiopathic Parkinson's disease. In some cases, the extrapyramidal symptoms of Mn poisoning may progress following cessation of exposure.

Formal neuropsychological testing can be used to document the behavioral and cognitive effects of Mn exposure in clinical and research settings. Tests of psychomotor functioning such as the Santa Ana Formboard, Digit Symbol, and finger tapping are sensitive to the neurotoxic effects of Mn. Attention and memory impairments induced by Mn exposure are revealed by tests sensitive to functioning in these two cognitive domains. Language and verbal functioning are typically spared in Mn-exposed adults, but children exposed to Mn may develop persistent impairments of language function.

Research studies investigating neuropsychological performance among Mn-exposed persons have revealed deficits at concentrations ranging from 1 to 28 mg/m^3. Tests in research settings used to document deficits have included finger tapping, reaction time, and digit spans. Visual spans and tests of visual spatial ability have also been used successfully in research settings to document the effects of Mn.

The neuropsychological test performance of Mn-exposed workers with parkinsonism has been compared with those of normal controls and patients with idiopathic Parkinson's disease. Mn-exposed workers with parkinsonism perform significantly worse than do unexposed controls on the *WAIS-R*, *Milner Facial Recognition Test*, Purdue pegboard, and the Continuous Performance Test. Mn-exposed workers with parkinsonism typically perform better on the Purdue pegboard than do patients with Parkinson's disease.

Neurobehavioral testing has been performed on a group of workers from a ferromanganese alloy plant. The workers had been exposed for 1–28 years, and the mean duration of exposure was 13 years. They were divided into three groups on the basis of their Mn exposure histories. The *low-exposure* group comprised foremen, clerks, and laboratory technicians exposed to Mn concentrations of only 0.009–0.15 mg/m^3 (mean blood Mn = 6.0 μg/liter, mean urine Mn = 1.7 μg/liter). The *medium-exposure* group consisted of maintenance workers with exposure levels ranging from 0.072 to 0.76 mg/m^3 (mean blood Mn = 8.6 μg/liter, mean urine Mn = 2.3 μg/liter). The *high-exposure* group consisted of those workers with the highest levels of exposure of up to 2.6 mg/m^3 (mean blood Mn = 11.9 μg/liter, mean urine Mn = 2.8 μg/liter). Cumulative exposure indices (CEI) were determined for each subject. Cognitive domains assessed included

attention (simple and complex reaction times), executive function (arithmetic), psychomotor function (Digit Symbol and finger tapping), short-term memory (Digit Span), and verbal understanding (vocabulary). The results of tests of psychomotor function (finger tapping and Digit Symbol) and short-term memory (Digit Span) were correlated with environmental exposure levels, as represented by individual CEI scores.

D. Trichloroethylene

Trichloroethylene (TCE) is commonly used as a degreasing solvent in industrial settings. The neurotoxic effects of TCE are associated with exposure to the parent molecule and to its environmental degradation product, dichloroacetylene (DCA). Because DCA is derived from TCE, individuals working with TCE or encountering it in nonoccupational settings are at risk for the neurotoxic effects of both chemicals.

Acute exposure to TCE induces narcosis characterized by subtle behavioral changes and subjective symptoms of drowsiness and an inability to concentrate. At higher levels of exposure, nausea, vomiting, headache, dizziness, confusion, stupor, and loss of consciousness occur. Acute exposure to very high levels of TCE has been associated with persistent neurological symptoms and death. The salient persistent effects of a severe acute exposure to TCE include facial anesthesia, reduced perception of taste, dysarthria, flattening of the nasolabial folds, ptosis, reduced pupillary response, constricted visual field, and sensorimotor neuropathy.

The 25-year serial follow-up examination of a 26-year-old man who experienced toxic encephalopathy after a single severe exposure to TCE was reported by us. Initial neuropsychological assessment of this individual revealed difficulty with sequential problem-solving and short-term memory tasks. Although daily functioning improved over the course of several years, the patient felt chronically depressed and apathetic and continued to experience problems with attention and short-term memory. These problems, which were severe enough to interfere with his ability to perform his daily activities, were documented by serial neuropsychological assessments. At a follow-up examination 16 years after the exposure incident, his performance IQ (PIQ) was 19 points lower than his verbal IQ (VIQ). Significant visuospatial deficits accounted for a large part of the discrepancy between the patient's PIQ and VIQ. Tests of short-term memory showed deficits in both immediate and delayed recall of verbal and visual information. Tests revealing deficits included the *Benton Visual Retention Test*, Logical Memories, Paired Associates, and Visual Reproductions. In addition, the *Minnesota Multiphasic Personality Inventory* (*MMPI*) indicated that the patient was depressed.

Toxic encephalopathy was also seen in a 62-year-old machinist who experienced a brief acute exposure to TCE. A follow-up neuropsychological assessment of this patient performed 5.5 years after the incident revealed performance deficits on the Digit Symbol and Object Assembly sections of the *Wechsler Adult Intelligence Scale* (*WAIS*). Furthermore, the patient exhibited overt impairment of executive function, lack of insight, and low motivation, all of which were associated with his past exposure to TCE.

Symptoms of chronic exposure to TCE develop insidiously whether intake occurs via a pulmonary or oral route. Memory and affect are affected by chronic exposure. For example, subjective symptoms among residents of Woburn, MA, who were exposed to TCE-contaminated well water, included headache, dizziness, fatigue, irritability, insomnia, memory and concentration impairments, and paresthesias. Effects associated with chronic occupational exposure to TCE vapors include forgetfulness, dizziness, headache, sleep disturbances, fatigue, irritability, anorexia, trigeminal nerve symptoms, sexual problems, and peripheral neuropathy. Speech and hearing disorders have been reported among children exposed to TCE through the use of contaminated drinking and bathing water. The cognitive domains most frequently affected by TCE include attention, executive functioning, short-term memory, and visuospatial ability; language and verbal skills are typically spared in adults. Children exposed to TCE may develop persistent impairments of verbal functioning, which interfere with learning later in life.

Neuropsychological testing documented mild-to-moderate encephalopathy in 24 of 28 individuals who were chronically exposed to TCE-contaminated drinking and bathing water. Impaired cognitive performance was seen on the following tests: Visual Reproductions, Logical Memories, Word Triads, and the *Benton Visual Retention Test*. Significant memory impairments were seen in 24/28 cases. Attention and executive function deficits were seen in 19/28 persons, whereas visuospatial deficits and manual motor function deficits were seen in 17/28. Language and verbal functioning among the adults were almost

always within expectation. However, the neuropsychological evaluations of the children in this group indicated that the developmental stage at the time of exposure is related to the type of neurobehavioral deficits seen postexposure. Children exposed before age 18 years were shown to have deficits in a greater number of neuropsychological domains than individuals exposed as adults. In addition, these children showed a decrease in performance on the *Boston Naming Test* that was not seen in their parents. These findings suggest that children exposed to TCE suffer more diffuse damage to the brain and are more likely to develop learning disabilities.

E. Organophosphorus Compounds

Organophosphorus compounds (OPCs) are used primarily as pesticides, but many (e.g., Sarin and VX) are also used as chemical warfare agents. These compounds inhibit the activity of acetylcholinesterase, resulting in the accumulation of acetylcholine at receptors on neurons in the central and peripheral nervous systems.

Exposure to OPCs is associated with three clinically distinct syndromes: (1) acute cholinergic crisis; (2) intermediate syndrome; and (3) organophosphate-induced delayed peripheral neuropathy (OPIDN). The clinical presentation of these three syndromes reflects the different pathophysiological mechanisms that underlie each. The *acute cholinergic crisis* develops within hours of exposure to OPCs and may last as long as 96 hr. The signs and symptoms seen during the acute cholinergic crisis are due to the inhibition of acetylcholinesterase. The clinical manifestations reflect excessive stimulation of nicotinic and muscarinic receptors. Clinical manifestations include weakness, muscle fasciculations, tachycardia, miosis, lacrimation, excessive salivation, seizures, and coma.

The *intermediate syndrome* follows the acute cholinergic crisis. Symptoms begin to emerge within 24–96 hr after removal from exposure. Symptoms are due to adverse effects on muscle cells, and can persist for up to 6 weeks and include weakness of the proximal muscles of the limbs and neck. Tendon reflexes are also reduced. The cranial nerves and respiratory muscles may also be involved. Weakness of the diaphragm may necessitate intubation and respiratory support. Death may occur in the most severe cases. Prognosis is good if the patient survives, and clinical recovery is typically complete.

Organophosphate-induced delayed neuropathy (OPIDN) is characterized by clinical and electrophysiological signs and symptoms of neuropathy that emerge 1–5 weeks after exposure. The clinical features of OPIDN include flaccid paralysis of the distal muscles of the lower and upper extremities. This clinical picture is in contrast with that of the intermediate syndrome, which involves the more proximal limb muscles. Deep tendon reflexes are reduced or absent in those patients with OPIDN. Patients report sensory symptoms of numbness and paresthesias that are aggravated by exercise. The prognosis for patients with OPIDN is variable, with clinical recovery requiring at least several months; recovery of function may be incomplete in the legs and feet of the most severe cases. Signs and symptoms indicative of CNS pathology are also noted in those patients with OPIDN. Patients with greater CNS involvement show increased tendon reflexes and the Babinski sign may be present.

Acute and chronic behavioral effects have been associated with exposure to OPCs. Changes in mood and affect including irritability, nervousness, and depression are seen during the acute cholinergic crisis. Cognitive deficits associated with the acute cholinergic crisis range from overt confusion to mild forgetfulness. These neurobehavioral manifestations may subside within days or may persist indefinitely. Formal neuropsychological testing reveals deficits in attention, executive functioning, psychomotor skills, and short-term memory. Significant neuropsychological performance deficits have been found in these patients up to several years after cessation of exposure, indicating that toxic encephalopathy can occur in some individuals exposed to OPCs.

Neuropsychological functioning has been assessed in workers with chronic exposure to OPCs who have not experienced any episodes of overt acute poisoning. The exposed workers show significantly slower reaction times when compared with unexposed controls suggesting that psychomotor processing is slowed.

Serial neuropsychological testing was completed in a 44-year-old EPA field inspector who was soaked with a large quantity of phosmet released from a crop duster airplane that was flying overhead. His acute symptoms included nausea, excessive sweating, salivation, blurred vision, and headache. He remained symptomatic over the next several days and reported experiencing anxiety, irritability, weakness, photophobia, and insomnia. He soon recovered from the acute cholinergic symptoms such as excessive sweating and salivating, but noted that he was now experiencing attention deficits and dizziness that he had never before experienced. A neurological examination and a brief neuropsychological assessment were performed 6 months

after the exposue incidant. The patient's cognitive performance was within the normal ranges on Digit Span, the *Benton Visual Retention Test*, Block Designs, *Wisconsin Card Sort Test*, and the Purdue pegboard. Despite this reassurance that he was cognitively "normal," the patient continued to experience difficulty with his activities of daily living.

A more detailed neuropsychological assessment was performed approximately 2 years after the exposure incident. At testing the patient's verbal IQ was 126 and was significantly greater than his performance IQ (106). Performance deficits were noted on Trails B (attention and executive function) and Digit Symbol (psychomotor speed). Memory deficits were noted on delayed recall of both visual and verbal material; immediate recall was within expectation. His answers tended to be concrete and he missed the gestalt on tests of complex verbal reasoning. The *Profile of Mood States* (*POMS*) revealed depression, anger, and fatigue.

A follow-up neuropsychological assessment was performed 13 months later at 40 months after the OPC exposure incident. Overall improvement was noted on tests of executive function, visuomotor speed, and retrieval of information on delayed recall. However, the patient continued to show deficits on tests of attention and had difficulty with recall of verbal paired associates. His major problem on this testing continued to be difficulty with complex verbal reasoning tasks. The *POMS* revealed only irritability. It was concluded that the overall improvement in cognitive functioning with some residual impairments seen in this patient is consistent with recovery from an acute toxic encephalopathy that could have resulted from OP exposure. Of interest in this case is that the patient went on to complete law school, and his only residual complaints reported 10 years after the OP exposure were that he occasionally experienced unexplained symptoms of nausea, blurred vision, and anxiety. Such an instance was when he entered a local grocery store for a brief period of time and became nauseous and had blurred vision. He went home and contacted the store manager by telephone, inquiring about the possibility that he might have been exposed to a chemical used in the store. The manager informed him that the store had been spayed with an insecticide earlier that same day. In addition, he has recognized that he feels anxious and has nonspecific discomfort at those times when he can detect aromatic substances in the air, such as perfumes, petroleum products, or auto exhaust. He has adjusted to his apparent multiple chemical sensitivity (MCS) by simply avoiding such circumstances.

The number of individuals who experience the preceding constellation of symptoms is sufficiently large that a clinician should recognize it as a behavioral syndrome. The exact pathogenesis of MCS syndrome is still unclear. Components of autonomic nervous system reactivity, olfactory triggers of emotional responses, and PTSD and features of depression make it difficult to identify a specific therapeutic intervention. Much further work is needed on this condition before it can simply be dismissed as psychogenic in origin.

V. CONCLUSION

Behavioral syndromes following exposure to neurotoxic chemicals must be differentiated from other neurological disorders. This is only accomplished with careful analysis of the time line relationship between reported exposure conditions and the emergence of symptoms. Markers of exposure such as air or water samples and biological specimens also should be obtained whenever possible to support the diagnosis. Clinical findings and/or corraborative diagnostic test data are of value when these results are interpreted in the context of developmental, academic, and social history, medical and psychiatric history, information gleaned from interviewing the patient and/or significant others, the patient's behavior in the test situation, and qualitative findings observed in the test material. Follow-up testing is useful in documenting prognosis, confirming the effect of exposure to neurotoxic chemicals, and that a neurodegenerative process is not responsible for the individual's symptoms and signs of neurologic disfunction.

See Also the Following Articles

BEHAVIORAL NEUROIMMUNOLOGY • BEHAVIORAL PHARMACOLOGY • CHEMICAL NEUROANATOMY • MOTOR SKILL • NEUROPHARMACOLOGY • NEUROPSYCHOLOGICAL ASSESSMENT

Suggested Reading

Baker, E. L., White, R. F., Pothier, L. J., *et al.* (1985). Occupational lead neurotoxicity: Improvement in behavioral effects after reduction of exposure. *Brit. J. Ind. Med.* **42,** 507–516.

Diamond, R., White, R. F., Gerr, F., and Feldman, R. G. (1995). A case of developmental exposure to inorganic mercury. *Child Neuropsychol.* **1,** 1–11.

Eto, K., Oyanhei, S., Itai, Y., *et al*. (1992). A fetal type of Minamata disease. An autopsy case report with special reference to the nervous system. *Mol. Chem. Neuropathol.* **16,** 171–186.

Feldman, R. G. (1999). *Occupational and Environmental Neurotoxicology*. Lippincott-Raven Publishers, Philadelphia.

Fiedler, N., Kipen, H., Kelly-McNeil, K., and Fenske, R. (1997). Long-term use of organophosphates and neuropsychological performance. *Am. J. Ind. Med.* **32,** 487–496.

Law, W. R., and Nelson, E. R. (1968). Gasoline-sniffing by an adult. Report of a case with the unusual complication of lead encephalopathy. *JAMA* **204,** 144–146.

Lezak, M. D. (1995). *Neuropsychological Assessment*. Oxford University Press, New York.

Mathiesen, T., Ellingsen, D. G., and Kjuus, H. (1999). Neuropsychological effects associated with exposure to mercury vapor among former chloralkali workers. *Scand. J. Work Environ. Health* **25,** 342–350.

Mitchell, C. S., Shear, M. S., Bolla, K. I., and Schwartz, B. S. (1996). Clinical evaluation of 58 organolead manufacturing workers. *J. Occup. Environ. Med.* **38,** 372–378.

Needleman, H. L., and Gatsonis, C. A. (1990). Low-level lead exposure and the IQ of children. A meta-analysis of modern studies. *J. Am. Med. Assoc.* **263,** 673–678.

Smith, P. J., Langolf, G. D., and Goldberg, J. (1983). Effects of occupational exposure to elemental mercury on short-term memory. *Br. J. Ind. Med.* **40,** 413–419.

Stewart, W. F., Schwartz, B. S., Simon, D., *et al*. (1999). Neurobehavioral function and tibial and chelatable lead levels in 543 former organolead workers. *Neurology* **52,** 1610–1617.

Trope, I., Lopez-Villegas, D., and Lenkinski, R. E. (1998). Magnetic resonance imaging and spectroscopy of regional brain structure in a 10-year-old boy with elevated blood lead levels. *Pediatrics* **101,** E7.

White, R. F., Feldman, R. G., and Proctor, S. P. (1992). Neurobehavioral effects of toxic exposure. In *Clinical Syndromes in Adult Neuropsychology: The Practitioner's Handbook*, (R. F. White, Ed.), Chapter 1, pp. 1–51. Elsevier Science Publishers, Amsterdam.

White, R. F., Diamond, R., Proctor, S., *et al*. (1993). Residual cognitive deficits 50 years after lead poisoning during childhood. *Brit. J. Ind. Med.* **50,** 613–622.

Neurodegenerative Disorders

LEE J. MARTIN
Johns Hopkins University School of Medicine

GLOSSARY

antioxidant An enzyme or chemical that inactivates reactive oxygen species.

apoptosis A form of programmed cell death that has a characteristic structural appearance and occurs through the activation of intrinsic cell death pathways.

apoptosis–necrosis continuum The concept that neuronal cell death can occur as typical apoptosis, typical necrosis, or as intermediates or hybrids of cell death with varying, overlapping contributions of apoptosis and necrosis.

astroglia Glial cells in the central nervous system that have long radial processes that ensheathe neurons and synaptic complexes. Astroglia regulate the extracellular chemical and ionic environment and secrete peptides, growth factors, cytokines, and chemokines.

axotomy Experimental injury to the neuronal axon (or bundles of axons) by transection, avulsion, or trauma.

central nervous system The brain and spinal cord.

death protein A protein that regulates programmed cell death.

excitotoxicity A neurotoxic process that is mediated by excessive activation of excitatory glutamate receptors.

experimental neuropathology The study of neurodegeneration in animal or cell culture model systems.

glial cell Nonneuronal cell in the nervous system (e.g., astrocyte, oligodendrocyte, microglial cell, Schwann cell).

glutamate receptor A family of cell membrane proteins located on neurons and glia that bind glutamate or related chemicals, causing depolarization by ionic conductance or activation of enzymes resulting in the production or release of intracellular second messengers.

microglia The resident small phagocytic cells of the brain and spinal cord that are related to the mononuclear phagocyte lineage and function as immune accessory cells that secrete cytokines and chemokines.

mutation Change in the DNA nucleotide sequence of a gene.

necrosis A form of cell death that has a characteristic structural appearance and occurs through the failure of homeostatic mechanisms (e.g., energy production, cell volume).

neurotransmitter Signaling molecule secreted by the presynaptic terminal of a neuron at chemical synapses to relay a signal to a postsynaptic neuron (e.g., glutamate).

NMDA receptor A type of ion channel glutamate receptor that controls neuronal excitation and is very important for synaptic plasticity. It regulates intracellular Ca^{2+}. Overactivation of NMDA receptors causes excitotoxic neuronal death.

oligodendroglia Glial cells that provide myelin sheaths for axons within the CNS and secrete peptide growth factors.

programmed cell death A form of cell death that is brought about by intrinsic cellular pathways involving specific death proteins.

reactive oxygen species An oxygen molecule containing an odd number of electrons rendering it chemically reactive due to an open bond (e.g., superoxide, hydroxyl radical, nitric oxide).

stem cell Relatively undifferentiated cell that can divide into daughter cells that can undergo terminal differentiation into particular cell types (e.g., neurons).

target deprivation The removal or ablation of a brain region or peripheral tissue that is the site with which a group of neurons connects.

Neurodegeneration is the pathology of neurons instigated by an acute insult or a chronic perturbation in cell function. The process of neurodegeneration can be rapid or it can be slow and progressive, and it can result in neuronal loss and neuronal dysfunction with neurological consequences. Acute neurodegeneration is caused by cerebral ischemia (stroke and cardiac arrest), toxins, and trauma. Chronic neurodegeneration is caused by diseases such as Alzheimer's disease, amyotrophic lateral sclerosis, Huntington's disease, and Parkinson's disease. In different neurological disorders, different groups of neurons are selectively vulnerable to the molecular pathology of the degenerative process. The basis for this selective vulnerability may be intrinsic to the different groups of neurons. Models of experimental neuropathology are essential for discovering the mechanisms for selective neuronal vulnerability to injury and neurodegenerative disease.

I. INTRODUCTION

A variety of neurodegenerative disorders affect the brain or spinal cord of humans (Tables I–III). The degeneration of nervous tissue can result from acute injury and from chronic disease. Acute neuropathology in adults, children, and infants can arise from head or spinal cord trauma, infection, toxicity, liver failure, and cerebral ischemia resulting from stroke, cardiac arrest, or asphyxiation. Chronic, progressive neuropathology occurs in adult disorders such as Alzheimer's disease, amyotrophic lateral sclerosis (Lou Gehrig's disease), Huntington's disease, multiple sclerosis, and Parkinson's disease and in infants and children with spinal muscular atrophy. These disease processes involve the degeneration of neurons and glial cells. Different populations of neurons in different nervous system regions have differential susceptibilities to disease and injury (Table I). This differential susceptibility of groups of neurons provides the basis for the yet to be understood concept of selective vulnerability (Table I). The underlying mechanisms for selective vulnerability are possibly related to specific properties of neurons, including size, axonal length, connections, metabolism, and gene mutations. This neuronal degeneration causes tragic neurological and behavioral disabilities ranging from memory loss to paralysis. For example, neuronal degeneration is responsible for the memory disturbances that occur in individuals with Alzheimer's disease (AD) and in people that have experienced cardiac arrest, asphyxiation, or strokes. Nerve cell loss is also responsible for the profound abnormalities in movement that occur in individuals with amyotrophic lateral sclerosis (ALS), Parkinson's disease, or Huntington's disease and in children with cerebral palsy or spinal muscular atrophy.

The impact of neurodegenerative disorders on our society is revealed by epidemiological studies. AD affects approximately 4 million adults (most are >65 years of age) and is the 4th-leading cause of death in the United States. AD accounts for $>100{,}000$ deaths annually. ALS affects approximately 30,000 Americans (4–6 people in 100,000), whereas Huntington's disease annually affects 4–7 people per 100,000. Estimates of the incidence of Parkinson's disease vary depending on the populations of individuals studied, with the annual incidence ranging from 7 to 19 affected people per 100,000. More than 100,000 Americans under the age of 18 years have some degree of neurological disability attributable to cerebral palsy.

Table I
Selective Vulnerability of the Human Central Nervous System in Neurodegenerative Conditions

Neurodegenerative disorder	Vulnerable neurons–brain regions
Alzheimer's disease	Neocortex, hippocampus, amygdala, basal forebrain cholinergic neurons, brain stem monoaminergic neurons
Amyotrophic lateral sclerosis	Lower motor neurons (spinal cord and brain stem) and upper motor neurons (motor cortex)
Cerebral ischemia (cardiac arrest and stroke)	Neocortex, hippocampus, striatum, cerebellum
Huntington's disease	Striatum
Multiple sclerosis	White matter
Parkinson's disease	Substantia nigra
Spinal muscular atrophy	Lower motor neurons (spinal cord and brain stem)

No drugs are available yet that can prevent the degeneration of neurons in the brain and spinal cord in people with these neurodegenerative disorders (Table I).

II. ALZHEIMER'S DISEASE IS A PROGRESSIVE NEURODEGENERATIVE DISEASE OF ELDERLY INDIVIDUALS THAT CAUSES SYNAPSE DEFECTS AND DEMENTIA

Alzheimer's disease is the most common cause of dementia occurring in middle and late life. About 70% of all cases of dementia are due to AD. It affects 7–10% of individuals >65 years of age and up to 40% of people >80 years of age. The prevalence of AD is increasing proportionally to increased life expectancy. AD thus will continue to be a major health concern because estimates predict that $\sim 25\%$ of the population will be >65 years of age in the year 2050. AD now affects >4 million people in the United States. Most cases of AD have unknown etiology and are called sporadic; however, some cases of AD, particularly those with early onset, are familial and are inherited as an autosomal dominant disorder linked to mutations in the gene that encodes amyloid precursor protein or in the genes that encode for presenilin proteins (Table II). For late-onset sporadic cases, a variety of risk factors have been identified in addition to age. The apolipoprotein E (apoE) allele is a susceptibility locus with the apoE4 type showing dose-dependent contributions. Cardiovascular disease and head trauma are additional risk factors for AD.

The mechanisms that cause the profound brain atrophy (Fig. 1), neuronal degeneration and progressive impairments in memory and intellect that occur with AD are not understood. Atrophy of the cerebral cortex (i.e., loss of cerebral cortex volume) correlates strongly with cognitive decline. Normal cognition and memory, as well as neuronal survival, depend on synapses (Fig. 2A). Regulated release of neurotransmitter-containing vesicles is necessary for normal synaptic function. Some proteins that control synaptic operation are abnormal in AD (Fig. 2B). These proteins function in the presynaptic terminal by controlling the regulated exocytosis of neurotransmitter packets. Particular proteins are reduced in the hippocampus of individuals with AD who have moderate to severe abnormalities in memory (Figs. 2 and 3). Another important discovery is that the synaptic proteins that control neurotransmitter vesicle translocation and priming at the release site of the presynaptic nerve terminal are more vulnerable than proteins involved in vesicle exocytosis at the cell membrane. It appears that these proteins begin to be lost in individuals in the early stages of AD who do not yet have detectable cognitive impairments, but as the synaptic defects progressively become more severe, individuals manifest more severe memory impairment. Abnormalities in the presynaptic proteins that regulate neurotransmitter release may be an early pathological process in the development of AD (Fig. 2).

A variety of brain lesions occur in people with AD. The major lesions are called neurofibrillary tangles and senile plaques (Fig. 4). These brain lesions are formed abundantly in individuals with AD and patients with Down's syndrome and less frequently in people aging

Table II
Gene Mutations Associated with Some Neurodegenerative Diseases

Neurodegenerative disease	Gene mutation–deletion	Prevalence
Familial amyotrophic lateral sclerosis[a]	Copper–zinc superoxide dismutase	10% of all FALS cases
Familial Alzheimer's disease[b]	Amyloid precursor protein	$<10\%$ of all familial AD cases
Familial Alzheimer's disease	Presenilin-1	$<10\%$ of all familial AD cases
Familial Alzheimer's disease	Presenilin-2	$<10\%$ of all familial AD cases
Huntington's disease	Huntingtin	100% of cases
Parkinson's disease	α-Synuclein	Very rare
Spinal muscular atrophy	Survival motor neuron, neuronal apoptosis inhibitory protein	Majority of cases

[a]ALS occurs as a familial or sporadic disease. Approximately 5–10% of all ALS cases are familial. The majority of cases are sporadic, with no yet identified mutations.
[b]AD occurs as a familial or sporadic disease. Approximately 10% of all AD cases are familial. The majority of AD cases are sporadic.

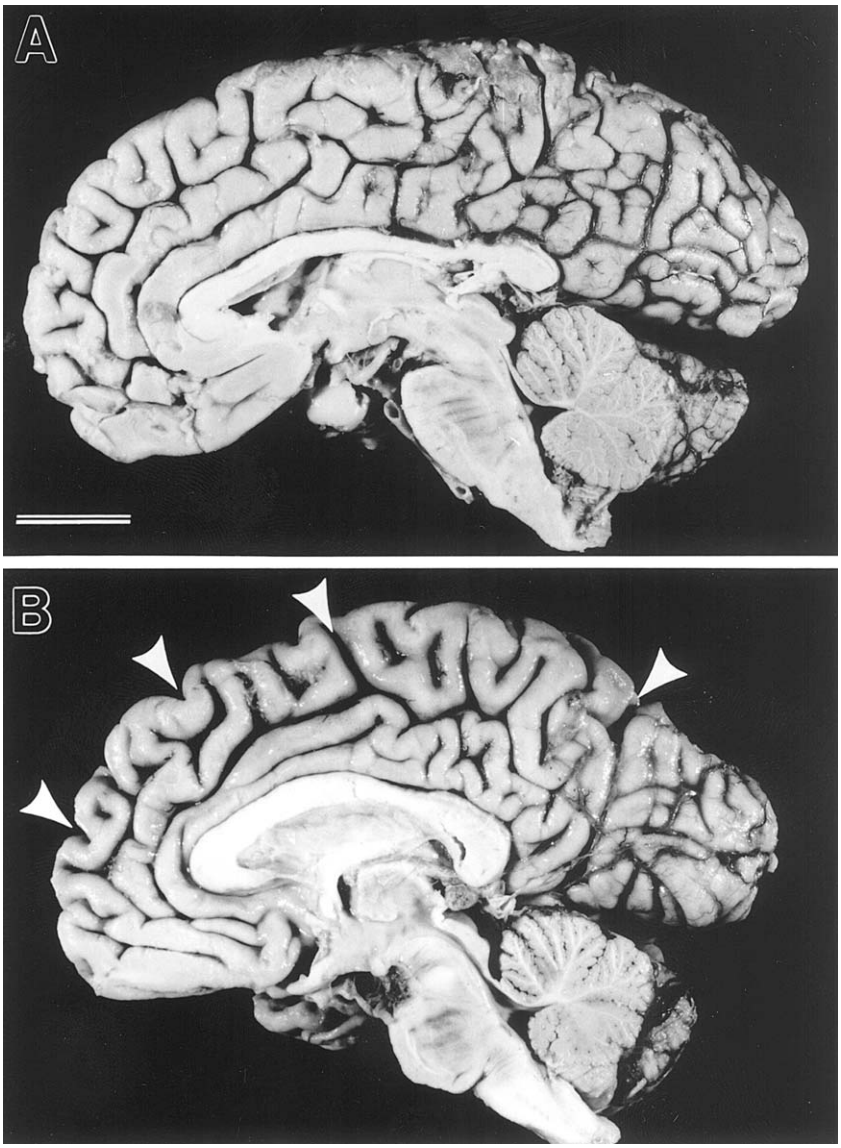


Figure 1 Severe atrophy of the brain occurs in Alzheimer's disease. (A) Midsagittal view of the brain from an 86-year-old normal, individual. The cerebral cortex is normal, with broad gyri and narrow sulci. Scale bar = 23 mm (same for B). (B) Midsagittal view of the brain from an 85-year-old individual with Alzheimer's disease. The cerebral cortex is atrophic, as indicated by widening of the sulci and narrowing of the gyri (white arrowheads).

normally. Pyramidal neurons in the neocortex and hippocampus (Fig. 4A, Table I) are highly vulnerable to the formation of neurofibrillary tangles. Neurofibrillary tangles are abnormal bundles of protein filaments that occur within neurons (Fig. 4B). These tangled masses consist of paired helical filaments containing τ protein (Fig. 4C). Senile plaques are formed throughout the brain parenchyma, and in AD they occur in large numbers (Fig. 4D). The composition of senile plaques is very complex (Figs. 4D and 5), consisting of dystrophic neurites (damaged and swollen dendrites or axon terminals), activated astrocytes and microglia, and extracellular deposits of insoluble amyloid. Amyloid occurs as fibrils composed of a small peptide (Aβ) consisting of 40–42 amino acid residues. This abnormal protein fragment is derived proteolytically from the amyloid precursor protein (APP), a cell surface protein. APP may participate directly in the pathogenesis of AD because mutations have been identified in the APP gene that are linked to early–onset AD in some families (Table II).

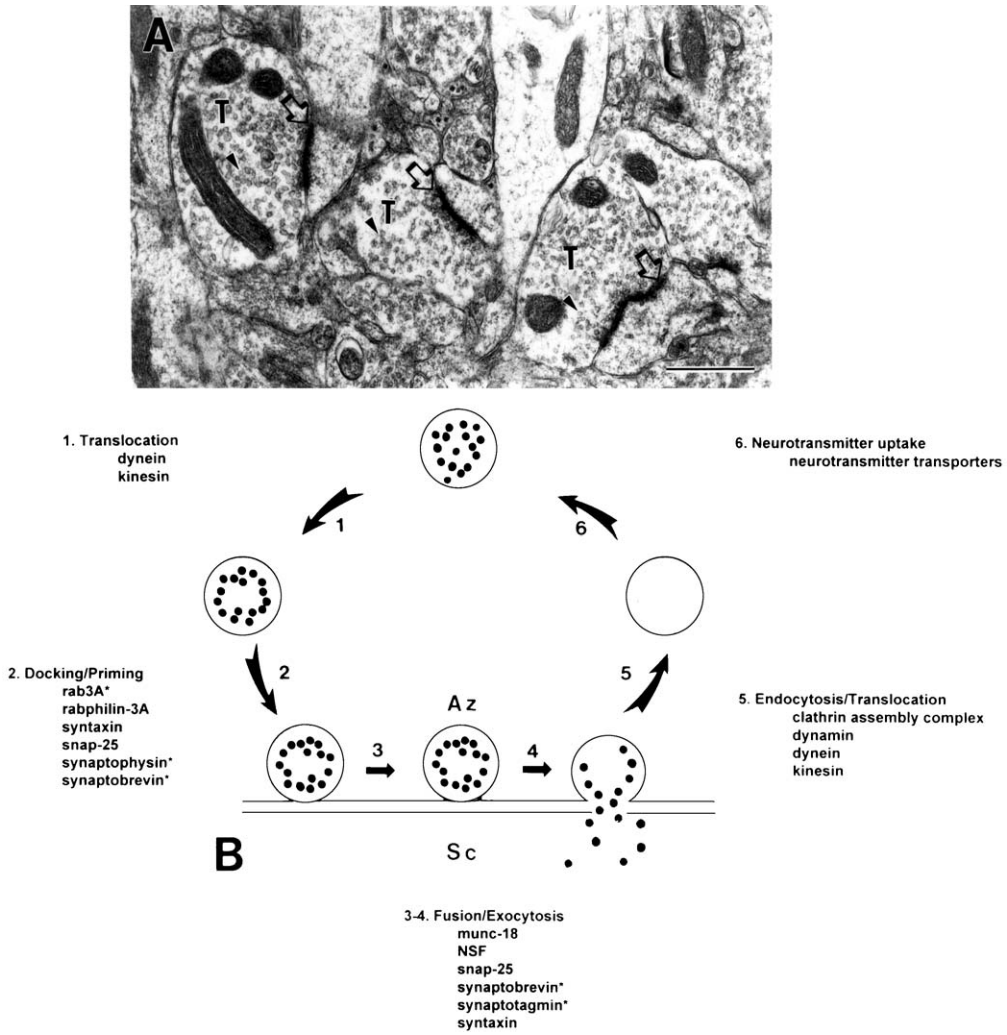


Figure 2 The synaptic vesicle cycle may be abnormal in Alzheimer's disease. (A) Electron micrographs of synapses in the cerebral cortex of a rhesus monkey. Axon terminals (T), containing many small clear synaptic vesicles (arrowheads), from synapses (open arrow) with postsynaptic structures. Synaptic vesicles cluster at the active zone of the presynaptic membrane. The narrow space between the presynaptic and postsynaptic components is the synaptic cleft. Scale bar = 1.0 μm. (B) Diagram of the synaptic vesicle cycle. This cycle functions in the regulated release of neurotransmitter-containing vesicles (filled circles) from the presynaptic terminal. Abbreviations: Az, synaptic active zone; Sc, synaptic cleft. The synaptic vesicle cycle can be divided into six major components, as indicated. Some of the specific proteins that function at each phase of the cycle are indicated. Some of these proteins (indicated by the asterisk) appear to be selectively vulnerable in individuals with Alzheimer's disease.

The functions of APP are still not well-defined, although it appears that APP functions at synapses. APP is an abundant and ubiquitous protein within CNS and other tissues. APP has structural features similar to some cell surface receptors and may be a G-protein-coupled receptor. Secreted and nonsecreted forms of APP exist, with different APP derivatives having neurotrophic or neurotoxic actions. APP is incorporated into the extracellular matrix and, thus, may have roles in cell–cell and cell–substrate adhesion. Furthermore, APP may function in the regulation of neurite outgrowth, perhaps by modulating the effects of neurotrophins and cytokines in the responses of neurons and glia to brain injury.

In cell culture, APP normally undergoes constitutive proteolytic cleavage by an α-secretase. This enzyme cleaves APP within the Aβ region at or near the plasma membrane, thereby generating secreted forms of APP and precluding the formation of full-length Aβ peptide fragments. APP is also metabolized by an endosomal–lysosomal pathway that, unlike the α-secretase pathway, yields amyloidogenic fragments of Aβ that are deposited in senile plaques (Fig. 4D). Aβ can be formed normally *in vivo* and *in vitro*, and studies of

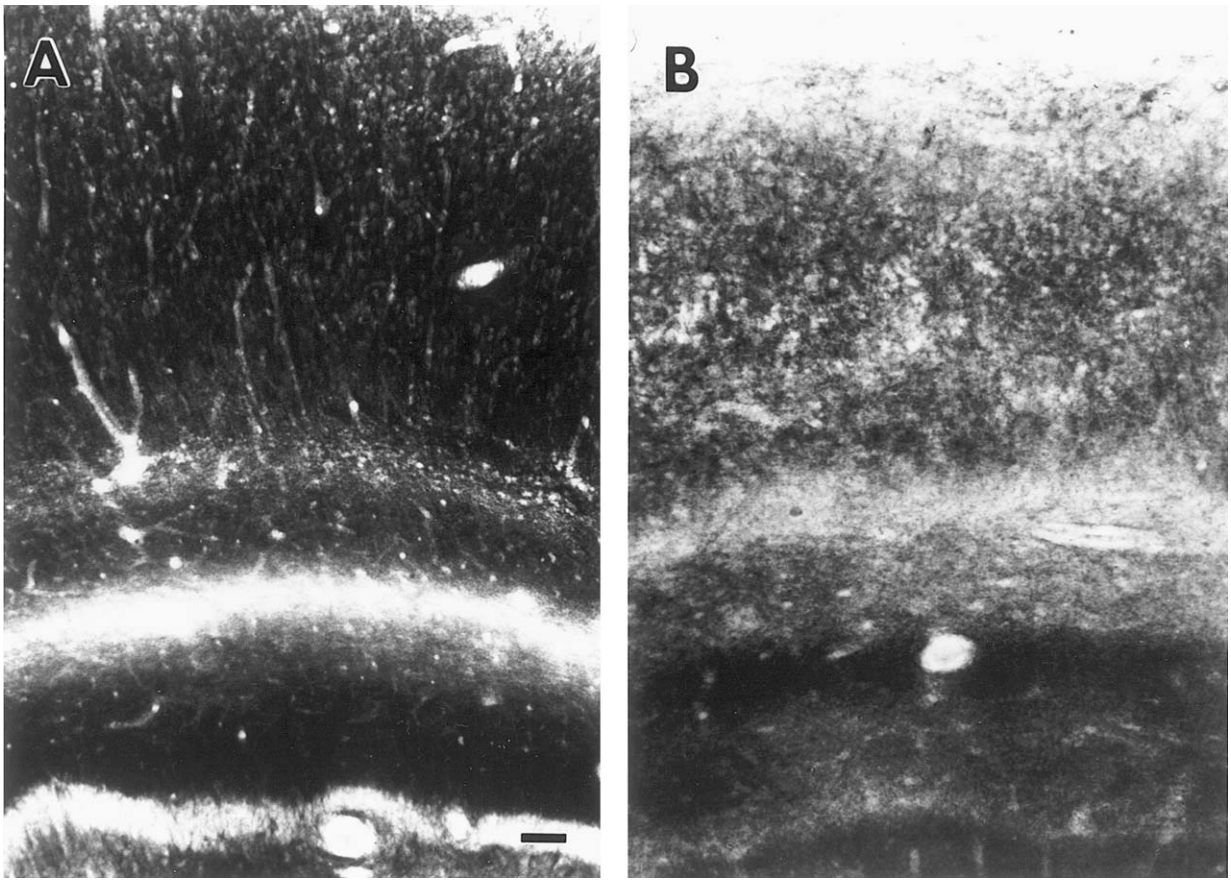


Figure 3 Synapses in the hippocampus of individuals with Alzheimer's disease are abnormal. (A) The hippocampus of normal, aged control humans has very high levels of synapses, as revealed by the synaptic vesicle protein synaptophysin (the intensity of the black staining reflects the amount of synaptophysin immunoreactivity). Scale bar = 42 μm (same for B). (B) The hippocampus of an individual with Alzheimer's disease shows a marked depletion of synaptophysin immunoreactivity (as reflected by the less intense black staining).

cultured human cells and aged non-human primates show that it is generated intracellularly.

Although Aβ has been shown to be neurotoxic in cell culture, a causal role for Aβ in neuronal degeneration within the brain remains speculative. A β-secretase cleaves APP at the N-terminus of Aβ, and a γ-secretase cleaves APP at the C-terminus of Aβ, causing the formation of Aβ that is either 40 or 42 amino acids long. This pathway for APP metabolism is found within the endoplasmic reticulum and Golgi apparatus of neurons. Mutant presenilins (Table II) promote Aβ42 generation. Presenilins, which are present at relatively low levels in the brain, localize to the endoplasmic reticulum and Golgi apparatus. Mutant presenilin is processed differently from normal presenilin, and fragments that are normally subject to endoproteolytic cleavage tend to accumulate. Thus, metabolism of APP through the β- and γ-secretase pathways may be promoted by *presenilin-1* and *presenilin-2* gene mutations linked to early-onset familial AD (Table II).

We and others have shown that APP is present in essentially all neurons and in some astroglia, microglia, and vascular endothelial cells. The most prominent neuronal localization of APP is within cell bodies and dendrites and is particularly enriched postsynaptically at some synapses. The expression of APP in nonneuronal cells in the brain is low in comparison to the dominant expression of APP within neurons and their processes. It appears that astroglia and microglia constitutively express APP at low levels in the resting state. However, the relative enrichment of APP within these neuroglial cells changes in response to brain injury and synaptic abnormalities. This idea is supported by our finding that APP is expressed prominently by activated astroglia and microglia within senile plaques of aged nonhuman primates and by other reports showing that APP is localized to

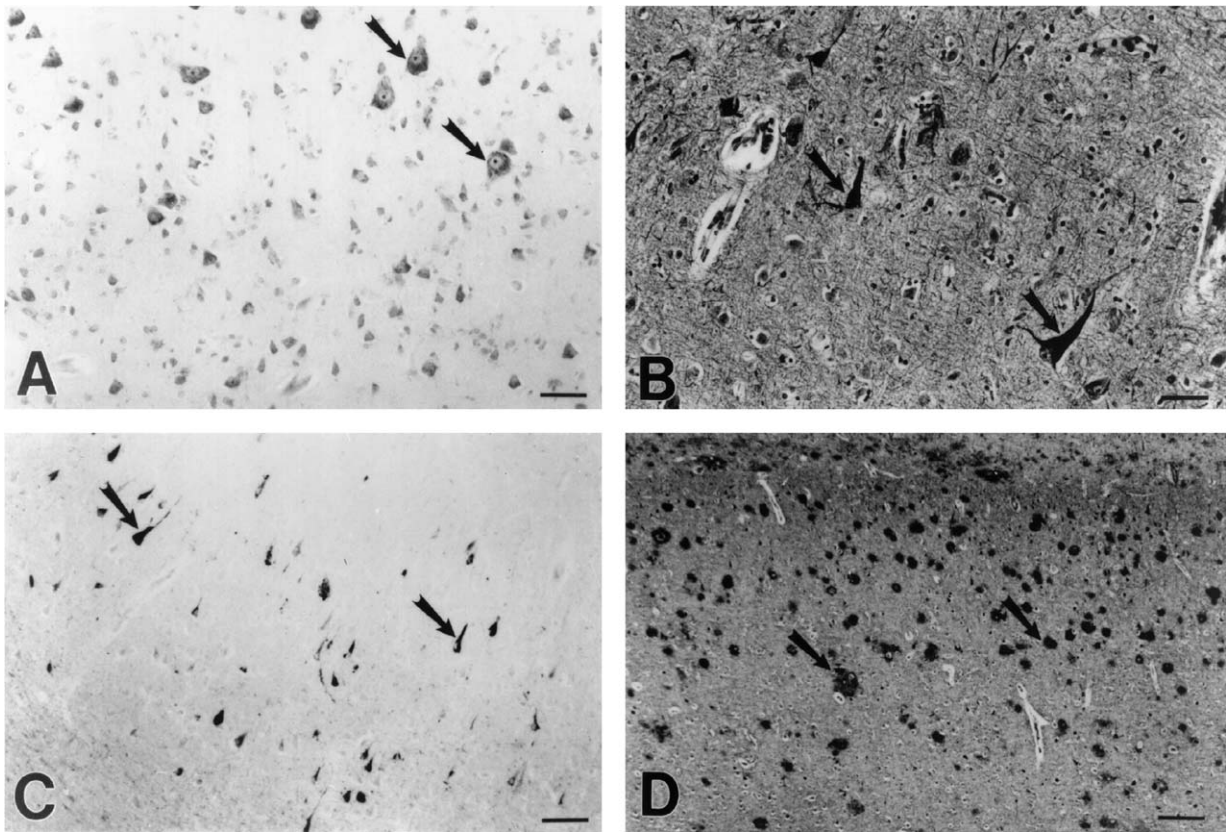


Figure 4 Neurofibrillary tangles and amyloid deposits are brain lesions that are formed in patients with Alzheimer's disease. (A) Pyramidal neurons in the neocortex (arrows) are vulnerable in individuals with Alzheimer's disease. Scale bar = 50 μm. (B) Neurofibrillary tangles, which are abnormal intracellular aggregates of protein (arrows), are formed in pyramidal neurons in patients with Alzheimer's disease. Scale bar = 50 μm. (C) Neurofibrillary tangles are composed of τ proteins (arrows). Scale bar = 100 μm. (D) Individuals with Alzheimer's disease form numerous abnormal extracellular deposits of Aβ amyloid protein in the brain (arrows). Scale bar = 200 μm.

astrocytes in senile plaques in cases of AD. Other studies have shown that some forms of APP are expressed in reactive astrocytes in the early stages of brain damage. Because levels of APP in some neurons and nonneuronal cells are increased by the cytokine interleukin-1, it is likely that the expression of APP is inducible in glia when these cells are activated in response to neuronal injury.

Several theories for the formation of senile plaques and Aβ deposits have been presented. The genesis of senile plaques may begin with the abnormal processing of APP via β-secretase and the formation of extracellular Aβ before the degeneration of cellular elements within these brain lesions. Alternatively, Aβ may be derived from degenerating axonal nerve terminals or dendrites containing APP that evolve into neurite-rich foci that form Aβ at the cell plasma membrane by aberrant processing of APP within neurons. In addition, invading reactive microglia and astroglia as well as capillaries may actively produce Aβ from APP.

We have found that senile plaques are dynamic brain lesions that evolve from early defects in synapses within the neuropil to mature plaques and extracellular deposits of Aβ (Figs. 4D and 5). The staging of these lesions is thought to be the degeneration of neuritic structures, followed by the attraction of reactive glia and the subsequent deposition of extracellular Aβ derived from microglia or astrocytes. Studies demonstrate that structural and biochemical perturbations within neuronal and non-neuronal cells, importantly glia, occur before the deposition of extracellular Aβ fibrils. Furthermore, these results suggest that focal abnormalities in synaptic contacts within the neuropil (synaptic disjunction) may instigate this complex series of events resulting in the formation of diffuse senile plaques and deposits of Aβ (Figs. 4D and 5). In

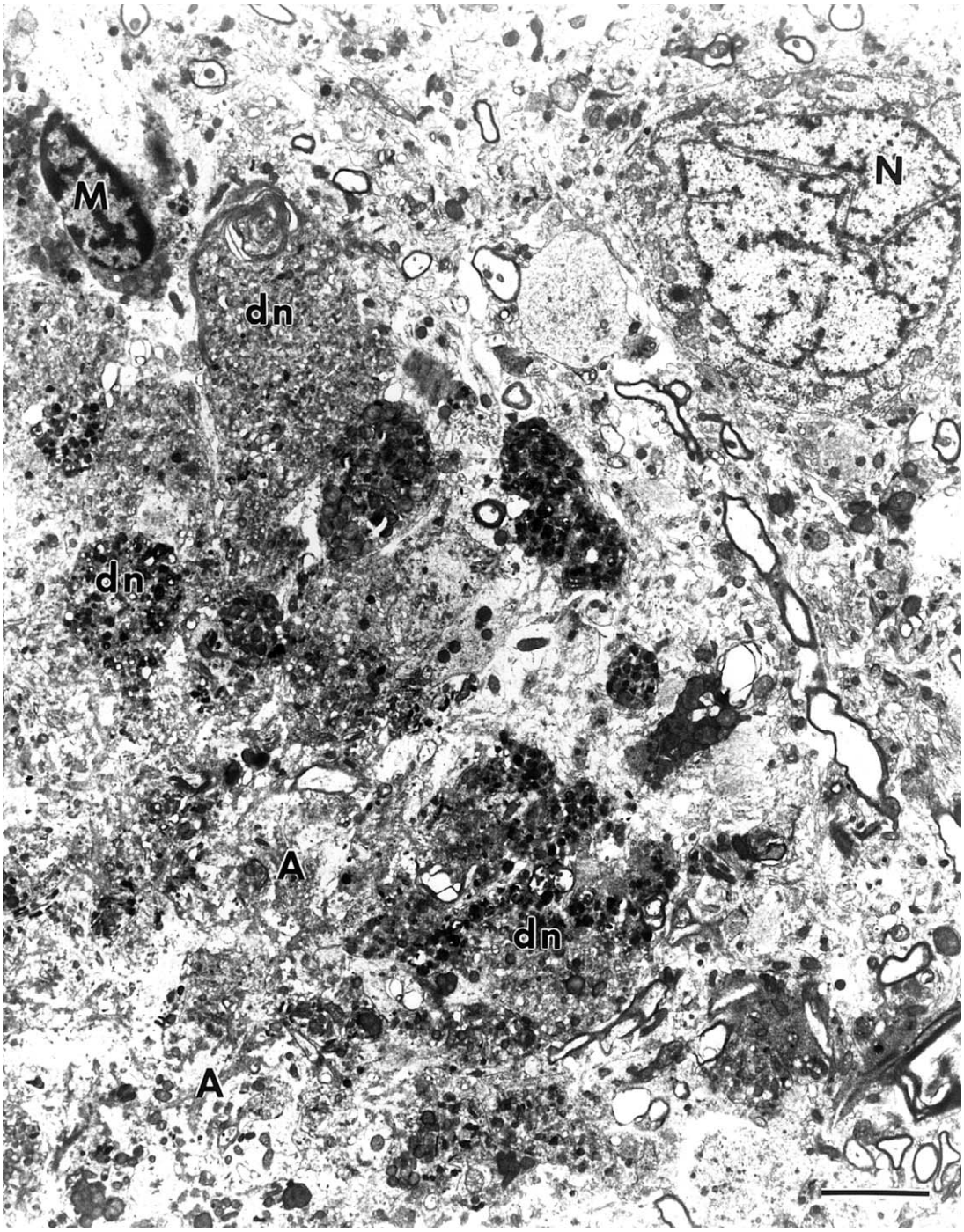


Figure 5 Electron microscopy reveals the complexity of the senile plaques that are formed in the cerebral cortex. Extracellular amyloid (A) is surrounded by numerous dystrophic, swollen neurites (dn), which are filled with abnormal membranous organelles. Microglia (m) infiltrate into the plaques. A nearby neuron (N) appears structurally normal Scale bar = 4 μm.

response to synaptic disjunction in the aged brain, astroglia and microglia produce Aβ. The molecular pathology we and others have identified at the synaptic level in humans with AD (Figs. 2 and 3) may be related to senile plaque lesions and Aβ deposits (Figs. 4D and 5). Defects in the synaptic vesicle cycle (Fig. 2) and synaptic disjunction in the neuropil may lead to abnormal APP processing

within neuroglia and could be early events in the formation of senile plaques and Aβ lesions.

III. ALS IS A DISEASE OF MOTOR NEURONS

ALS is a fatal neurological disease that causes a movement disorder (Table III) characterized by progressive muscle weakness, muscle atrophy, and eventual paralysis. Individuals affected with ALS die within 3–5 years of clinical onset. This disease is neuropathologically characterized by progressive degeneration of the upper and lower motor neurons in the brain and spinal cord (Table I, Fig. 6). The neuropathology of ALS is primary degeneration of the upper (motor cortical) and lower (brain stem and spinal) motor neurons. The amyotrophy refers to the neurogenic atrophy of affected muscle groups, and the lateral sclerosis refers to the hardening of the lateral white matter funiculus in the spinal cord (corresponding to degeneration of the corticospinal tract) found at autopsy. Because the mechanisms for the motor neuron degeneration in ALS are not understood, this disease has no precisely known causes and no effective treatments. Two major forms of ALS exist: idiopathic (sporadic) and heritable (familial). The vast majority of ALS cases are sporadic with no known genetic component. The familial forms of ALS (FALS) are autosomal dominant and make up about 10–20% of all ALS cases. In a subset of familial ALS cases (about 5–10%), missense mutations have been identified in the gene for superoxide dismutase 1 (SOD1), also called copper–zinc superoxide dismutase (Table II).

A variety of theories have been proposed for the possible causes of neurodegeneration in ALS (Table VI). A major center of attention is the mutant forms of SOD1 found in FALS. Because SOD1 is widely expressed in cells throughout the body and in CNS tissue the expression is very ubiquitous, the basis for the selective vulnerability of motor neurons in the presence of SOD1 mutations is still not clear. Initial experiments hinted that mutations in SOD1 could lead to motor neuron degeneration by decreasing enzymatic activity, resulting in neurotoxicity of reactive oxygen species, notably superoxide radicals that are inefficiently scavenged by mutant SOD1. However, it was found that FALS-linked mutations in SOD1 generally do not impair enzymatic activity but instead decrease protein stability. It was then proposed that mutant SOD1 acquires a neurotoxic gain in function. Mutations in SOD1 may convert this enzyme from a protein with antioxidant–antiapoptotic functions to a protein with apoptosis-promoting effects. In addition to the dismutation of superoxide, SOD1 also has peroxidase activity, and this peroxidase activity is enhanced in mutant SOD1 compared to normal SOD1. This gain of function could lead to the enhanced production of reactive oxygen species that could damage motor neurons. In mice with forced expression of mutant forms of the gene encoding for SOD1, motor neuron degeneration does occur. Unfortunately, however, this degeneration in transgenic mice overexpressing mutant forms of SOD1 is neuropathologically different from the degeneration of motor neurons in people with sporadic and familial ALS.

Studies have identified that the degeneration of motor neurons in ALS is a form of apoptotic cell death that appears to occur by a programmed cell death (PCD) mechanism (Fig. 7). PCD is a type of cell death that is triggered by intrinsic cellular pathways involving specific death proteins. This PCD of motor neurons in ALS could be due to a gain in function of the tumor suppressor protein p53. p53 is a DNA-binding phosphoprotein that functions in genome surveillance, DNA repair, and gene transcription. p53 commits to death cells that have sustained DNA damage from genotoxic agents and reactive oxygen species. DNA lesions have been found in individuals with ALS possibly because of free radical damage and defective DNA repair. Thus, p53 may participate in the mechanisms for motor neuron death in ALS in response to DNA damage. The gene expression of some cell death proteins is promoted by p53. For example, the levels of a protein called Bax (see Table IV) are regulated by p53. Bax is critical for neuronal

Table III
Classification of Some Movement Disorders in Humans

Motor neuron diseases	Akinetic disorders	Hyperkinetic disorders	Ataxic disorders
Amyotrophic lateral sclerosis	Parkinson's disease	Huntington's disease	Spinocerebellar degeneration
Spinal muscular atrophy	Progressive supranuclear palsy	Dystonia	Ataxia–telangiectasia

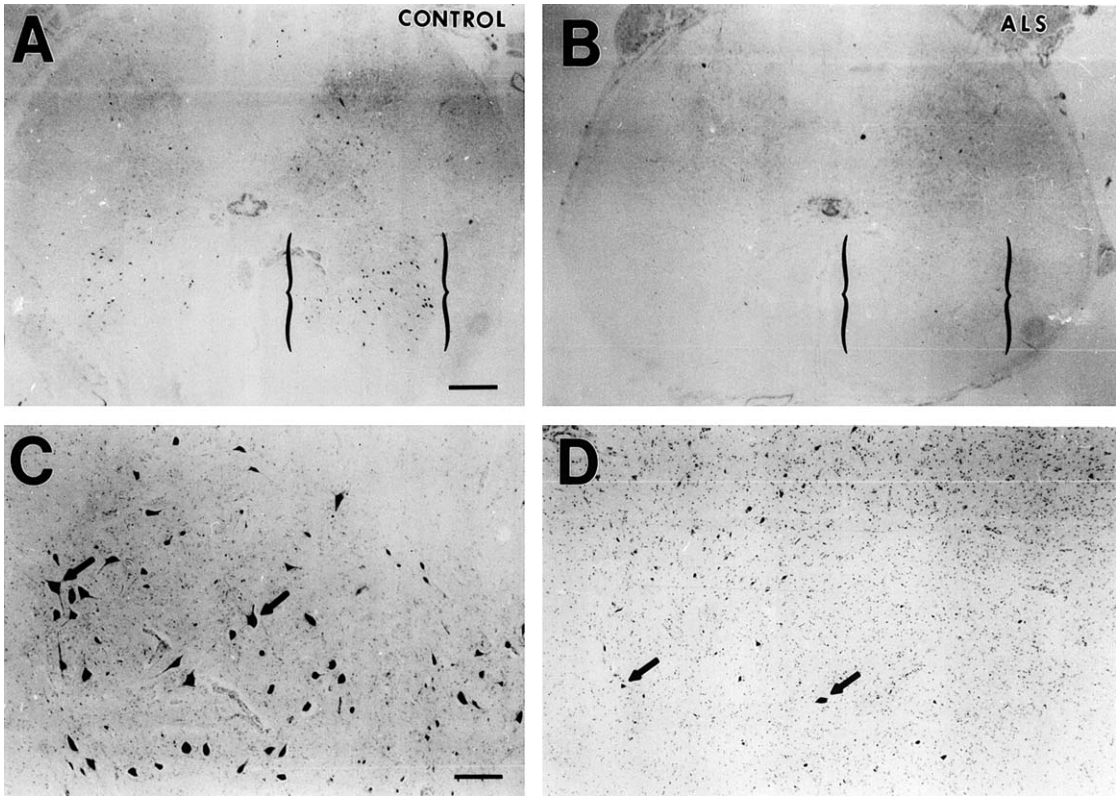


Figure 6 Motor neurons degenerate in patients with ALS. (A) In normal control individuals, the anterior horn of the spinal cord (brackets) is populated with many neurons (black dots within brackets). Scale bar = 0.7 mm (same for B). (B) In ALS, the anterior horn is depleted of neurons (brackets). (C) The anterior horn of the spinal cord in normal subjects is populated with large, multipolar motor neurons (arrows). Scale bar = 200 μm (same for D). (D) In ALS, the anterior horn contains many shrunken neurons (arrows) instead of large multipolar cells.

apoptosis. In ALS, we have found that p53 levels are elevated and that this p53 has competent DNA binding activity; moreover, Bax levels are elevated in individuals with ALS. This information is novel and is conceptually very important for further understanding the pathobiology of motor neuron death in ALS. It suggests that critical molecular mechanisms for regulating human cancer are overactive in vulnerable CNS regions in a human-age-related neurodegenerative disease. This theory may advance the

Table IV
Leading Theories on the Possible Causes of Motor Neuron Degeneration in ALS

Mechanism	Comment
SOD1 mutation	Found in some FALS cases. Resulting in a toxic gain in function or modified stability of SOD1.
Excitotoxicity	Resulting from abnormal glutamate receptor activation and defects in glutamate transport.
Neurotrophin withdrawal	Resulting from insufficient muscle cell- or glial-cell-derived trophic support or defective neurotrophin receptor signaling.
DNA damage–repair defects	Resulting from oxidative stress or inefficient DNA repair enzyme function. May involve both mitochondrial and nuclear DNA damage.
Autoimmunity	Resulting from autoantibodies to motor neuron antigens.
Aberrantly occurring programmed cell death	May be triggered by all of the precedings mechanisms.

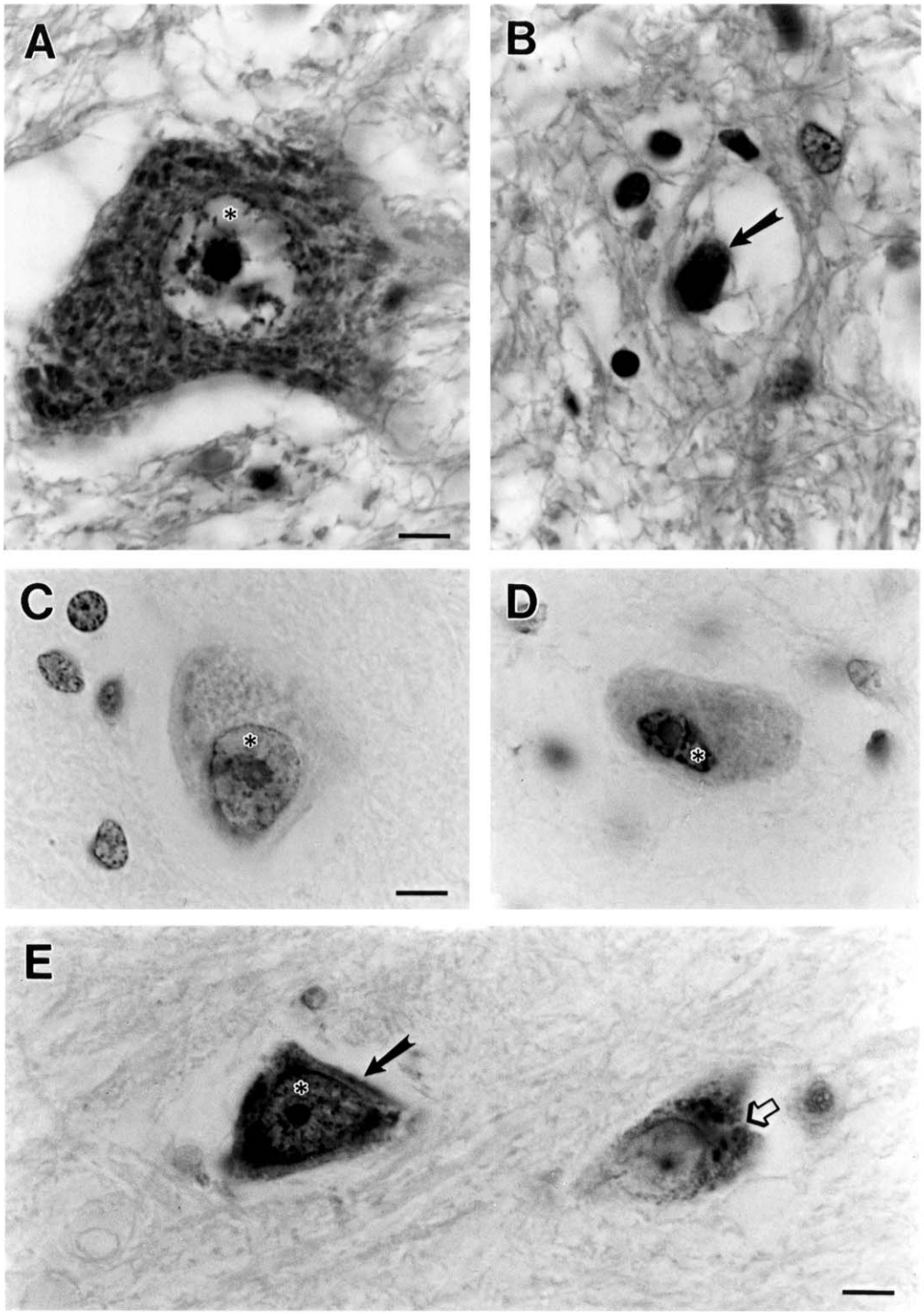


Figure 7 Motor neuron degeneration in ALS is a form of apoptosis that may be mediated by a p53-dependent mechanism. (A) Normal appearing spinal motor neuron with a large, multipolar cell body and a large nucleus (asterisk) containing a reticular network of chromatin and a large nucleolus. Scale bar = 7 μm (same for B). (B) Near end stage apoptotic motor neuron in ALS (arrow). The cell has shrunken to about 10% of normal size and has become highly condensed. (C and D) Nuclear DNA fragmentation (asterisk) occurs in motor neurons in patients with ALS as the nucleus condenses (asterisks) and the cell shrinks. Scale bar = 7 μm. (E) In individuals with ALS, p53 accumulates in the nuclei (asterisk) of motor neurons (arrow). The nearby neuron (open arrow) has an unlabeled nucleus for comparison. Scale bar = 10 μm. (See color insert in Volume 1).

understanding of motor neuron degeneration in ALS and perhaps other age-related neurodegenerative disorders in which chronic and progressive DNA damage may occur.

IV. NEURODEGENERATIVE DISEASES OF THE BASAL GANGLIA CAUSE MOVEMENT DISORDERS

The basal ganglia are subcortical brain structures that function in sensory–motor integration and in the planning and initiation of skeletal muscle movements. The striatum (caudate nucleus and putamen), globus pallidus (external and internal divisions), substantia nigra (pars compacta and reticulata), and subthalamic nucleus are the major components of the basal ganglia (Fig. 8). The basal ganglia do not control movement through direct connections with lower motor neurons, but rather the functions of the basal ganglia are executed by the frontal cortex through corticobulbar and corticospinal projections to brain stem and spinal motor neurons, respectively. The operation of the basal ganglia involves forebrain–diencephalon–midbrain circuitry loops. The different circuits within the basal ganglia utilize different neurotransmitters (Table V). The primary input to the basal ganglia originates from the neocortex and is directed to the striatum (Fig. 8). This corticostriatal projection uses the excitatory neurotransmitter glutamate. The primary output center of the basal ganglia is the globus pallidus, which conveys signals to back to the neocortex through the thalamus (Fig. 8). The pallidothalamic projection is inhibitory; in contrast, the thalamocortical projection is excitatory (Table V). Thus, the globus pallidus functions to inhibit the excitatory thalamic drive of neocortex. Somatic movements occur when thalamic neurons are released from tonic inhibition. This release occurs when corticostriatal projections excite striatal neurons that can phasically inhibit the neurons in the

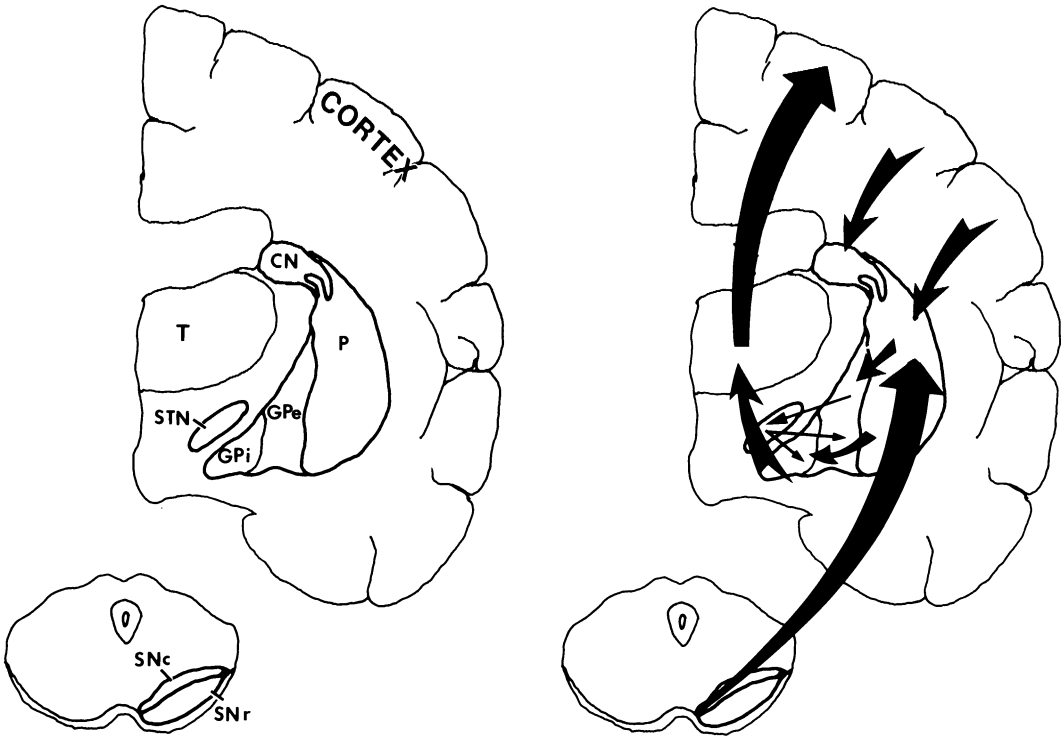


Figure 8 Basal ganglia circuits control movements. The basal ganglia comprise (left panel) the caudate nucleus (CN), putamen (P), globus pallidus external (GPe) and internal (GPi) divisions, the subthalamic nucleus (STN), and the substantia nigra compact (SNc) and reticular (SNr) divisions. The cerebral cortex and thalamus (T), although not part of the basal ganglia, participate in the connectivity loops (right panel). See Table V for connections. The major excitatory input to the striatum (the caudate nucleus and putamen) is from the cerebral cortex (right panel). The striatum in turn projects to the globus pallidus and the substantia nigra reticular division. Striatal activity is modulated by extensive dopaminergic input from the substantia nigra compacta. The major output of the basal ganglia is directed toward the thalamus, originating from GPi and SNr (not shown). The thalamic projection to the cerebral cortex (premotor and supplementary motor areas) drives the activity of the motor cortex, which executes somatic movements.

Table V
Primary Connections of the Basal Ganglia and Their Major Neurotransmitters[a]

Projection	Neurotransmitter
Neocortex to striatum (corticostriatal)	glutamate
Striatum to globus pallidus (striatopallidal)	GABA[b]–neuropeptides[c]
Globus pallidus externa to subthalamic nucleus (pallidosubthalamic)	GABA
Subthalamic nucleus to globus pallidus externa–interna (subthalamopallidal)	Glutamate
Subthalamic nucleus to substantia nigra reticulata (subthalamonigral)	Glutamate
Substantia nigra compacta to striatum (nigrostriatal)	Dopamine
Substantia nigra reticulata to thalamus (nigrothalamic)	GABA
Globus pallidus interna to thalamus (pallidothalamic)	GABA

[a]See Fig. 8 for illustration of connections.
[b]γ-Aminobutyric acid.
[c]Principal striatal neurons coexpress a variety of neuropeptide transmitters, including enkephalins, tachykinins, and dynorphin (see the article on Neuropeptides and Hormones in the Brain and Spinal Cord).

globus pallidus that inhibit thalamic neurons. The resulting activation of thalamocortical projections excites the premotor and supplementary motor areas of frontal neocortex that activate the motor cortex, thereby facilitating movement.

There are two major circuits within the basal ganglia, designated as the direct and the indirect pathways. These two pathways counterbalance each another. The direct pathway is the striatal projection to the globus pallidus (internal segment) and the substantia nigra reticulata, which in turn project to the thalamus (Fig. 8). Activation of the direct pathway results in the activation of thalamocortical projections and the facilitation of movement. The substantia nigra dopaminergic neurons activate striatal neurons of the direct pathway, thus functioning to facilitate movement. The indirect pathway, in contrast, involves a loop with the subthalamic nucleus (Fig. 8), which changes the physiological outcome in the thalamus. In the indirect pathway, the striatum inhibits the globus pallidus (external segment), which functions to inhibit the subthalamic nucleus. The subthalamic nucleus projects back to the globus pallidus (both segments) and to the substantia nigra reticulata to activate these regions; the internal segment of the globus pallidus and the substantia nigra reticulata then inhibit the thalamus. Thus, activation of the indirect pathway decreases movement. The substantia nigra dopaminergic neurons inhibit striatal neurons of the indirect direct pathway, thus opposing the activation of the indirect pathway and functioning to facilitate movement.

Diseases of the basal ganglia can cause involuntary movements, lack of movement, or slowness of movement (Table III). In Parkinson's disease, degeneration of the dopaminergic neurons in the substantia nigra results in the loss of dopamine in the striatum. Parkinson's disease causes involuntary movements (tremors), lack of movement (akinesia), and slowness of movement (bradykinesia). Some of the neurological features of Parkinson's disease are thought to arise from the loss of dopamine inhibition of striatal neurons in the indirect pathway (Fig. 8), resulting in increased inhibition of the globus pallidus external segment and greater excitatory drive of the subthalamic nucleus on the globus pallidus interna and substantia nigra reticulata. Moreover, the loss of dopamine results in decreased activity of striatal neurons in the direct pathway, leading to increased pallidothalamic inhibition. In Huntington's disease, degeneration of subsets of the principal neurons in the striatum may lead to decreased activity of the globus pallidus internal segment via the indirect pathway (Fig. 8). This failure to suppress thalamocortical activity is thought to cause slow writhing movements of the extremities (athetosis) and abrupt movements of the limbs and facial muscles (chorea).

V. NEURONAL DEATH OCCURS IN DIFFERENT FORMS

Neurons can die in different ways (Fig. 9). The death of cells has been classified generally as two distinct types: apoptosis and necrosis. These two forms of cellular degeneration are classified differently because they are believed to differ structurally and biochemically. Apoptosis is generally regarded as physiological cell death and is considered to be an organized PCD that is mediated by active, intrinsic mechanisms through which certain molecular pathways are activated to initiate apoptosis (Table VI). In contrast, necrosis is cell death resulting from a failure to sustain homeostasis due to extrinsic insults to the cell (e.g., osmotic, thermal, toxic, traumatic). The process of cellular necrosis involves damage to the structural and functional integrity of the cell plasma membrane, a rapid influx of ions and H_2O, and, subsequently, dissolution

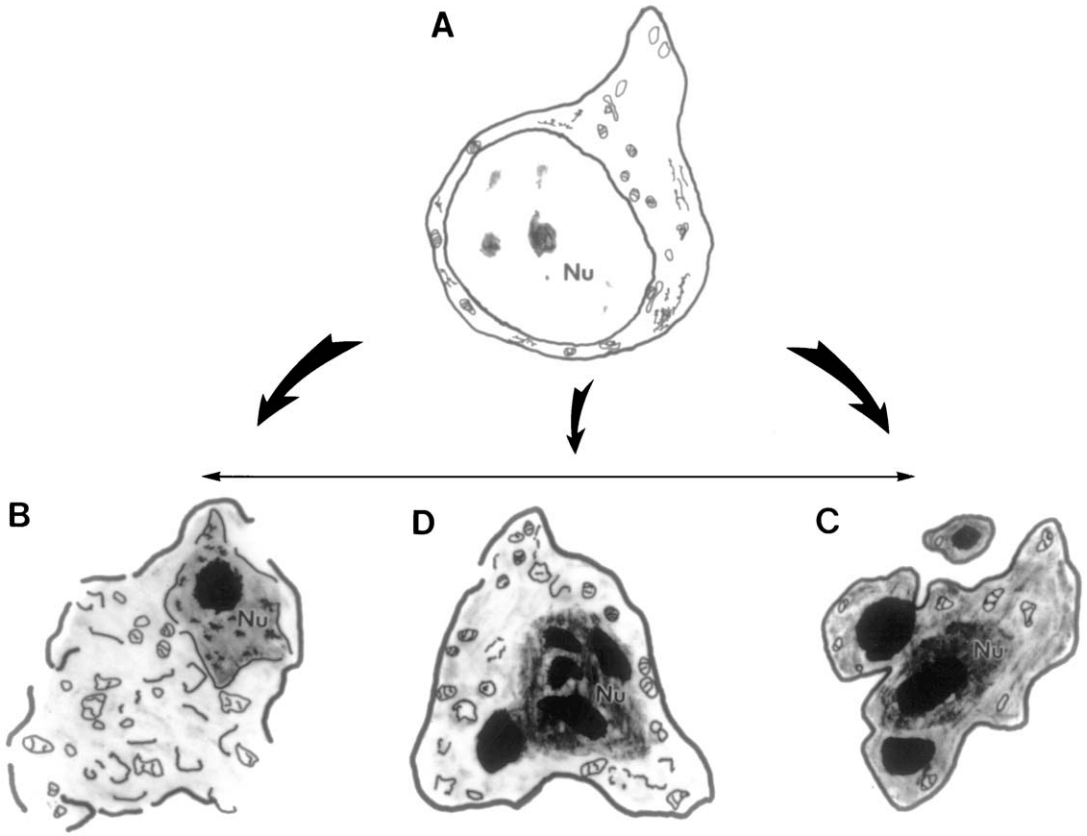


Figure 9 Neuronal cell death occurs as an apoptosis–necrosis continuum. According to the traditional binary scheme for cell death, a neuron (A) can die by either necrosis (B) or apoptosis (C). These forms of cell death were thought to be structurally distinct (representative images of cells are shown) and mutually exclusive. With cellular necrosis (B), which occurs as groups of cells, massive damage to organelles and the plasma membrane occurs with the release of cellular constituents. The nucleus (Nu) undergoes fragmentation, with the condensation of nuclear material being irregular and distinct from that occurring in apoptosis. In necrotic neurons the nucleolus can remain intact. In contrast, apoptosis (C) is an organized form of cell death that occurs generally as isolated cells within groups of cells. The nucleus containing the DNA is packaged into uniformly condensed masses (shown by the round or elliptical black structures), and the surrounding cytoplasm becomes shrunken and condensed (shown by the gray) with the organelles generally preserved until end stage apoptosis. Small fragments of cell cytoplasm surrounding packaged chromatin bud from the dying cell and are engulfed by glial cells. The concept of the apoptosis–necrosis continuum is based on the observation that neurons can die with a structure that is a hybrid of apoptosis and necrosis (D). The nuclear (Nu) and cytoplasmic changes are intermediate between those occurring in apoptosis and necrosis. The packaging of the chromatin occurs as large irregular masses in the nucleus. In the cytoplasm, some mitochondria remain intact whereas others are swollen.

of the cell. Thus, cellular necrosis is induced not by an intrinsic program within the cell per se (as in PCD) but by abrupt or slow homeostatic perturbations and departures from physiological conditions. It has been realized that an abnormal activation of PCD in brain and spinal cord neurons may also play a role in the disease process in humans with neurodegenerative disorders; therefore, deciphering of the contributions of the different types of cell death in degenerative diseases of the human CNS could help to develop treatments for these diseases. These treatments could possibly be drugs that inhibit the actions of key enzymes, ion channels in cell membranes, or numerous other proteins, as well as drugs (e.g., antioxidants) that block or inactivate the production of toxic chemicals (e.g., free oxygen radicals) that are generated during the process of neuronal death.

VI. PROGRAMMED CELL DEATH OCCURS NORMALLY DURING NERVOUS SYSTEM DEVELOPMENT

Naturally occurring apoptotic degeneration of neurons occurs normally in the developing nervous system (Fig. 10). Animals are born with excess numbers of

Table VI
Molecular Regulation of Programmed Cell Death

Bcl-2 family				
Antiapoptotic proteins	**Proapoptotic proteins**	**Caspase family**	**IAP family**	**Tumor suppressor family**
Bcl-2[a]	Bax[a]	Apoptosis "initiators:" caspase-2, -8, -9, -10	NAIP	p53[a]
Bcl-x_L[b]	Bak[a]	Apoptosis "executioners:" caspase-3[a], -6, -7	IAP1	p63
Boo	Bcl-x_S	Cytokine processors: caspase-1, -4, -5, -11, -12, -14	IAP2	p73
	Bad			
	Bid		XIAP	
	Bik			

[a]Proteins that have been shown to be abnormal in individuals with ALS.
[b]Proteins that have been shown to be unchanged in ALS.

neurons. The elimination of these supernumerary neurons in the developing nervous system occurs by normal programmed neurodegeneration at specific times. This normal neurodegeneration is thought to be important for matching the size of neuronal groups to the size of their targets (other groups of neurons to which they are connected by axons and nerve terminal synapses) as well as to their own synaptic inputs from other regions. This developmental neuronal death (particularly of spinal and sympathetic ganglion neurons and motor neurons) is thought to be partially controlled by the supply of sustaining neurotrophic factors that are synthesized by the target axon-associated glial cells or by the input regions. Thus, an insufficient supply of neurotrophic molecules triggers an apoptotic process within interconnected groups of neurons by PCD.

Apoptosis is regulated by specific molecules within cells (Table VI). Several genes that regulate apoptosis were originally identified in a nematode worm. Homologous genes have been identified in mammalian cells. The molecular mechanisms for apoptosis involve the participation of at least three groups of proteins that are made from three different gene families. One set of proteins is the caspase family of cysteine-containing, aspartate-specific proteases (14 members have been identified to date). A second group includes the death-promoting and death-suppressing proteins in the *Bcl-2* family (e.g., *Bcl-2*, *Bcl-x_L*, *Bax*, *Bak*, *Bad*, *Boo*). A third group of proteins is designated as the inhibitor of apoptosis protein (IAP) family.

Caspases exist as dormant proenzymes in healthy cells and are activated through regulated proteolysis. These proteins function in the execution phase of apoptosis with "initiator" caspases activating "effector" caspases, which subsequently cleave a variety of proteins, thereby causing the molecular and structural changes of apoptosis. Once activated, caspases act on nuclear proteins, cytoskeletal proteins, or cytosolic proteins. Two different caspase cascades mediate PCD. One pathway involves the regulated release of cytochrome c from mitochondria that promotes the activation of caspase-9 through Apaf-1 and then caspase-3 activation. Another pathway is initiated by the activation of cell-surface death receptors, including Fas and tumor necrosis factor receptor, leading to caspase-8 activation, which in turn cleaves and activates downstream caspases such as caspase-3, -6, and -7.

Apoptosis regulation by the *bcl-2* protooncogene family is a very complex and exciting process (Table VI). Of these genes, *bcl-2* and *bcl-x_L* are antiapoptotic (death-suppressing), whereas *bax*, *bcl-x_S*, *bad*, *bak*, and *bik* are proapoptotic (death-promoting). Membership into the family of Bcl-2-related proteins is defined by homology domains within their amino acid sequences. These domains function in the interactions (i.e., binding) between members. Bcl-2 family members exist as monomers (single proteins) that form homodimers (two of the same proteins bound together), heterodimers (two different proteins bound together), and higher order multimers (more than two interacting proteins). For example, Bax forms homodimers or forms heterodimers with either Bcl-2 or Bcl-x_L. When Bax is present in excess, it antagonizes the antiapoptotic activity of Bcl-2. The formation of Bax homodimers promotes apoptosis, whereas Bax heterodimerization with either Bcl-2 or Bcl-x_L blocks apoptosis. Thus, the complex steady-state array of protein–protein interactions among members of the

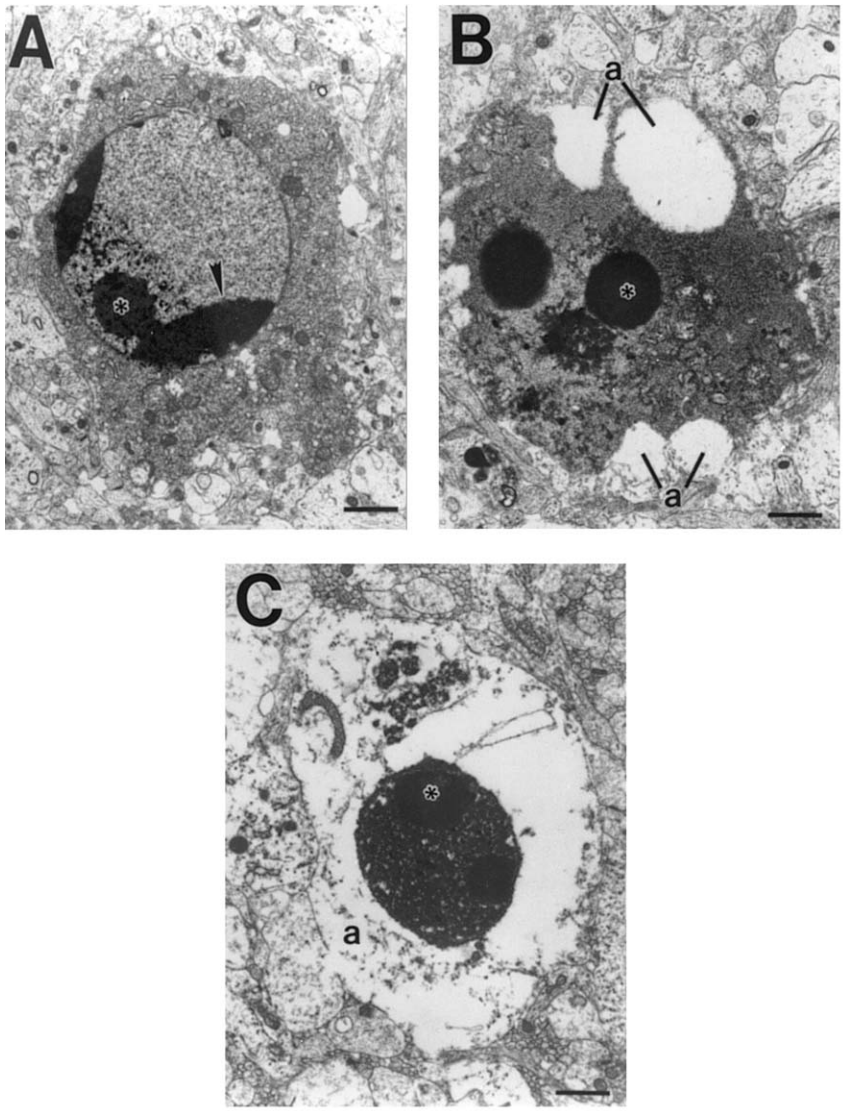


Figure 10 Neurodegeneration in the form of apoptosis is important for the normal development of the nervous system. By electron microscopy this naturally occurring programmed cell death has very characteristic features. (A) In the developing rat striatum, degenerating neurons in the early stages of apoptosis are characterized by chromatin condensation into crescentic caps (arrowhead) abutting the nuclear envelope and into round aggregates (asterisk). The surrounding cytoplasm condenses (as indicated by the uniformly dark staining), although most of the mitochondria remain intact. Scale bar = 2.5 μm. (B) As apoptosis progresses, major changes occur in the nucleus and cytoplasm. The chromatin packaging into dense round clumps (asterisk) becomes more advanced, and the integrity of the nuclear envelope is lost. The cytoplasm becomes more condensed and the mitochondria become damaged. Astrocytic processes (a) begin to surround the degenerating cell. Scale bar = 3.0 μm. (C) At end stage apoptosis, fragments of cells consist of round packages of chromatin (asterisk) surrounded by condensed cytoplasm. These apoptotic fragments are engulfed by astrocytic processes (a). Scale bar = 4.0 μm.

Bcl-2 family functions in dictating whether a cell lives or dies by apoptosis.

Cell death is also regulated by the IAP (inhibitor of apoptosis protein) family (Table VI). This family includes X-chromosome-linked IAP, IAP1, IAP2, and NAIP (neuronal apoptosis inhibitory protein). Survival motor neuron is another apoptosis inhibitory protein. The primary mechanism identified by which IAPs suppress apoptosis is the prevention of proteolytic processing of specific caspases. It appears that procaspase-9 is the major target of IAPs. However, IAPs do not prevent caspase-8-induced proteolytic activation of procaspase-3. IAPs can also block apoptosis by reciprocal interactions with the nuclear

transcription factor NFκB. NAIP is abnormal in infants and children with spinal muscular atrophy (Table II).

The regulation of PCD in the mammalian nervous system by caspases and *bcl-2* family members has been substantiated by using transgenic mouse technology and neuronal culture systems. In mice, deficiencies in the genes for caspase-3 and caspase-9 result in perinatal death and cerebral malformations, possibly caused by reduced PCD during brain development. Inhibition of caspase-1 and caspase-2 blocks the apoptosis of cultured dorsal root and sympathetic ganglion neurons when deprived of the neurotrophin nerve growth factor; furthermore, inhibition of caspase-1 arrests the PCD of motor neurons in cell culture resulting from neurotrophic factor deprivation and in the nervous system during the period of naturally occurring cell death. Mice that overexpress the *bcl-2* gene or have the *bax* gene eliminated fail to exhibit normal PCD of neurons in some nervous system regions, whereas *bcl-2*-deficient mice show progressive neurodegeneration after the period of PCD.

VII. GLUTAMATE RECEPTOR EXCITOTOXICITY KILLS NEURONS

Neurons communicate by neurotransmission at synapses (Fig. 2A). In the CNS, the amino acid glutamate is the major excitatory neurotransmitter that is packaged into small clear synaptic vesicles (Fig. 2A). Glutamate is released from nerve terminals into the synaptic cleft by regulated exocytosis of synaptic vesicles (Fig. 2B). Concentrations of glutamate at the synaptic cleft have been estimated to be approximately 1 m*M*, whereas the concentration of interstitial glutamate is about 1 μM. Glutamate can bind and activate several types of glutamate receptors (GluRs) on neurons (Table VII). These GluRs are classified broadly as either ion channel or metabotropic G-protein-coupled receptors. These classes of GluRs have distinct molecular compositions and distinct signal transduction mechanisms.

The ion channel GluRs are the *N*-methyl-D-aspartate (NMDA) receptors and the non-NMDA receptors. The non-NMDA GluRs are further divided into the α-amino-3-hydroxy-5-methyl-4-isoxazole propionate (AMPA) and kainate (KA) receptors (Table VII). The ion channel GluRs all form monovalent cation (Na^+, K^+) conducting channels, but they have differences in their permeabilities to divalent cations (Ca^{2+}). The activation of ion channel GluRs directly changes the conductance of specific ions through the receptor–ion channel complex, thereby inducing membrane depolarization. Fast, short-lived (1–10 msec) excitatory postsynaptic currents in most neurons in the CNS are mediated by these receptors. These receptors are oligomers, most likely pentameric heterooligomers, of homologous subunits encoded by distinct genes. The NMDA receptor subunits are NR1, NR2A–NR2D, and NR3, the AMPA receptor subunits are GluR1–GluR4 (or GluRA–GluRD), and the kainate receptor subunits are GluR5–GluR7 and KA1–KA2.

The metabotropic GluRs (mGluRs) are G-protein-coupled receptors that are single proteins encoded by

Table VII
Molecular Classification of Glutamate Receptors

Ion channel (ionotropic) receptors			G protein-coupled (metabotropic) receptors		
NMDA	Non-NMDA		Group I	Group II	Group III
	AMPA	Kainate			
	Receptor subunits				
NR1	GluR1	GluR5	mGluR1	mGluR2	mGluR4
NR2A	GluR2	GluR6	mGluR5	mGluR3	mGluR6
NR2B	GluR3	GluR7			mGluR7
NR2C	GluR4	KA1			mGluR8
NR2D		KA2			
NR3					

single genes. The mGluRs do not form ion channels but are instead linked to signal transduction molecules within the plasma membrane. mGluRs have slower electrophysiological characteristics (latencies >100 msec) than ion channel GluRs. Group I mGluRs (mGluR1 and mGluR5) operate through activation of phospholipase C (PLC) by G_q proteins, phosphoinositide hydrolysis and generation of inositol 1,4,5-triphosphate and diacylglycerol, and subsequent mobilization of Ca^{2+} from nonmitochondrial intracellular stores. Group II mGluRs (mGluR2 and 3) and group III mGluRs (mGluR4 and 6–8) function by G_I- or G_o-protein-mediated inhibition of adenylyl cyclase and modulation of ion channel activity.

Although glutamate and GluR activation are critical for normal nervous system function, glutamate is toxic to neurons at abnormally high concentrations if the GluRs on neurons are excessively activated. This process is called excitotoxicity. The excessive stimulation of GluRs by glutamate or chemical analogs of glutamate produces abnormalities in intracellular ions, pH, protein phosphorylation, energy levels, and reactive oxygen species. Acute excitotoxicity causes degeneration in neuronal cultures of animal brain and spinal cord and after intracerebral delivery of GluR activators into the CNS of experimental animals. In addition, excitotoxicity participates in the mechanisms for neuronal degeneration in animal models of cerebral ischemia, as well as brain and spinal cord trauma, and in the neurotoxicity in humans resulting from consumption of mussels contaminated with the KA receptor activator domoic acid. Excitotoxicity is also suspected as a culprit in the nerve cell loss associated with AD, ALS, Huntington's disease, and Parkinson's disease.

The precise mechanisms for GluR-mediated excitotoxic degeneration of neurons are not understood. Both neuronal culture and animal model data are discordant with regard to whether excitotoxic neuronal death is apoptosis or necrosis. Activation of neuronal GluRs kills neurons by pathways that may involve alterations in cytosolic free Ca^{2+} homeostasis and activation of Ca^{2+}-sensitive proteases, protein kinases, endonucleases, lipases, and phospholipases. Excitotoxicity results in an activation of endonucleases (DNA-cleaving enzymes) and internucleosomal digestion of genomic DNA into 180–200 base pair fragments 12–48 hr after intracerebral injections of excitotoxins in rats. Internucleosomal fragmentation of DNA also occurs in cultures of cortical neurons, although others have not found internucleosomal DNA fragmentation in cell culture.

The structural changes that occur in neurons in the adult rat brain after an excitotoxic insult include swelling and vacuolation of the cell body and dendrites, fragmentation of the nucleus into irregular clumps of chromatin, and damage to membranous organelles including the Golgi apparatus, endoplasmic reticulum, and mitochondria. This damage is thought to be typical of cellular necrosis (Fig. 9). However, in the immature brain, excitotoxicity can cause neuronal death very similar to apoptosis.

VIII. NEURONAL DEGENERATION CAN OCCUR AS AN APOPTOSIS–NECROSIS CONTINUUM

We have developed the concept that neuronal degeneration is influenced by brain maturity and GluR subtype. We tested the hypothesis that GluR-mediated excitotoxicity in the brain induces neuronal death with characteristics that vary depending on the maturity of the brain at the time of the insult and the GluR subtype that is activated. In the newborn rat brain, excitotoxic activation of NMDA and non-NMDA GluRs (Table VII) causes neuronal death with phenotypes ranging from apoptosis to necrosis (Fig. 9). Three structurally different forms of dying neurons were identified initially: a classic apoptotic form, a vacuolated form, and a classic necrotic form. When the progression of excitotoxin-induced neuronal death in the newborn brain was evaluated, it was found that the vacuolated form is a precursor stage of apoptosis, which has many similarities to the PCD that occurs naturally in the developing brain. Thus, some neurons die as a hybrid of apoptosis and necrosis. In contrast, when the adult rat brain is exposed to excitotoxins, the degeneration of neurons caused by NMDA receptor activation is morphologically necrotic; however, the neuronal death produced by non-NMDA receptor activation (Table VII) is distinct from that caused by NMDA receptor stimulation. Non-NMDA receptor-mediated neuronal death in the adult brain has some cytoplasmic and nuclear features reminiscent of neuronal apoptosis in the excitotoxically injured newborn brain, although non-NMDA receptor excitotoxic neurodegeneration in the adult brain and naturally occurring apoptosis in the developing brain are very different structurally. Surprisingly, both NMDA and non-NMDA receptor-mediated excitotoxic neurodegeneration occur in the presence of apoptotic-like internucleosomal DNA fragmentation.

Our experiments using this animal model of excitotoxic degeneration of neurons led to the novel concept of an apoptosis–necrosis continuum for neuronal death in the CNS. We concluded that the excitotoxic death of neurons does not have to be strictly apoptotic or necrotic, according to a traditional binary classification of cell death (Fig. 9), but it can also occur as intermediate or hybrid forms of cell death with coexisting characteristics that lie along a structural continuum with apoptosis and necrosis at the extremes. This continuum is influenced by the subtype of GluR that is activated (Table VII); hence, excitotoxic neuronal death may not be identical in every neuron, possibly because of the high diversity in the expression, localization, and function of GluR subtypes and second messenger systems in the CNS. We also concluded that the structure of neuronal death is influenced by CNS maturity, because excitotoxic degeneration of adult neurons does not occur with apoptotic structural features that closely resemble those seen during naturally occurring cell death in the developing nervous system.

This new concept may be important for understanding how neuronal degeneration occurs in neurological disorders that affect the human brain and spinal cord (Table I) and, thus, may be important for future studies aimed at the prevention of neuronal loss in human neurodegenerative diseases. The clarification of the relationships between mechanisms of neuronal death (active or passive) and the resulting structure of dying neurons in human neurodegenerative disease is important, particularly when addressing hypotheses as to whether PCD and apoptosis are equivalent and whether apoptosis and necrosis are mutually exclusive forms of neuronal cell death. Furthermore, if brain maturity dictates how neurons die, then, in humans, neuronal degeneration in adults may be fundamentally different from neuronal degeneration in newborns or children. For example, mature neurons appear to be less capable than immature neurons of displaying an apoptotic structure after an excitotoxic insult. An injury that produces a hybrid of apoptosis and necrosis in the adult CNS is more likely to elicit primarily apoptosis in the immature CNS. We speculate that PCD mechanisms may be activated more readily after an injury to the immature brain than to the mature brain, because immature neurons are closer than mature neurons to the period of naturally occurring developmental PCD.

IX. APOPTOSIS MAY HAVE IMPORTANT CONTRIBUTIONS TO NEURODEGENERATIVE DISORDERS IN HUMANS

It has been discovered that the degeneration of neurons in ALS is a form of apoptosis (Fig. 7). Vulnerable brain and spinal cord regions in ALS (Table I) have abnormalities in the balance of Bcl-2, Bax, and Bak proteins and abnormalities in their interactions (Table IV). However, the initial molecular pathology and upstream signals for motor neurons to engage PCD mechanisms are not known, although we suspect DNA damage. Furthermore, it is still unknown whether the neuronal degeneration in other age-related neurological disorders such as AD, Parkinson's disease, and Huntington's disease is related causally to an abnormal activation of PCD pathways in selectively vulnerable neurons.

Much uncertainty also centers around the possible role of apoptosis in the nerve cell degeneration resulting from cerebral ischemia caused by heart failure, asphyxiation, and stroke. Historically, cellular degeneration resulting from these abnormalities has been considered a form of cellular necrosis, but it has been suggested that postischemic neurodegeneration is apoptosis possibly mediated by a PCD mechanism. However, the contribution of apoptosis to the selective degeneration of neurons after ischemia is not yet resolved.

X. ANIMAL MODELS OF NEURODEGENERATION ARE NECESSARY TO UNDERSTAND HOW NEURONS DIE

Animal models of neuronal degeneration are crucial for improving our understanding of the mechanisms and progressive stages of neuronal death. These models provide an experimental system to identify how nerve cells die in paradigms that mirror certain neuropathological and clinical features of a neurological disorder that occurs in humans (Tables I and VIII). With animal models, the process of nerve cell death can be studied at the structural, biochemical, and molecular levels, and then subsequently the model can be used to test new therapies to prevent the degeneration of neurons in a biologically relevant system.

A. Neuronal Degeneration in Models of Axotomy and Target Deprivation

Animal models of axotomy (axon cutting or transection) and target deprivation (target removal) provide

insight into the mechanisms of progressive neuronal degeneration and are relevant to acute and slow, chronic degenerative disorders that affect the human brain or spinal cord (Table VIII). The progression of axotomy–target deprivation-induced neuronal degeneration and the likelihood of subsequent neuronal death or survival are influenced by several variables, including whether the cell body of an axotomized neuron resides within the peripheral nervous system (PNS) or CNS, the age of the animal at the time of injury, the location of axonal trauma in relation to the cell body, and the animal species. In the immature brain and spinal cord, axotomized neurons often die rapidly. Axotomy-induced degeneration of motor neurons in the immature CNS appears to be apoptosis on the basis of structural evidence in mouse and chick and the finding that overexpression of the *bcl-2* gene reduces motor neuron death in newborn mice in response to facial nerve transection or sciatic nerve transection. In contrast, in the adult nervous system, axotomized neurons can recover or persist in some altered form or they can undergo apoptosis. The outcome depends upon the type of model. Thus, in general, neuronal apoptosis is induced more easily by axotomy in the immature or newborn CNS than in the mature or adult CNS. Differences in the fate of target-deprived neurons in the developing and adult CNS may be related to the extent of target deprivation and collateral connections of deprived neurons with other brain regions and related to the differential ability to activate a PCD pathway.

Interruption of specific axonal pathways by transection in rodents and nonhuman primates can be used as a model for some forms of neuronal degeneration and behavioral deficits found in humans with certain neurological disorders (Table VIII). The hippocampal formation and septum form a neural system in the brain that functions in learning and memory. This system is vulnerable in AD (Table I). These brain regions are interconnected by an axon pathway called the fimbria–fornix. Transection of the fimbria–fornix is a model of target deprivation–axotomy that results in degeneration of neurons in the basal forebrain cholinergic complex and partial removal of glutamate-utilizing synaptic inputs (deafferentation) to neurons in the lateral septal nucleus. By using this model in rats, we have shown that these neurons that are deprived of their target or of inputs undergo chronic atrophy rather than death. They shrink but survive in a sickly state. Thus, long-term damage occurs within the cell bodies and dendrites (the major synapse–receiving areas) of these neurons, including the formation of vacuolar pathology, but they do not die. This is encouraging because, if the process is reversible, therapeutic interventions could restore these neurons to a normal healthy condition.

This neuronal injury within the septal complex of rat brain after fimbria–fornix transection has some pathological similarities to neuronal damage caused by excitotoxic GluR activation (Table VII), suggesting a mechanistic overlap. This idea has been confirmed by experiments showing that a drug that blocks NMDA receptors reduces the neuronal damage in the septum after fimbria–fornix transection. This work shows that, following interruption of the fimbria–fornix, the neuronal damage that occurs within the septum is

Table VIII
Human Neurodegenerative Disorders and Animal Models

Neurological condition	Age of onset	Major vulnerable CNS regions	Animal model
Alzheimer's disease	Adult (mid to late life)	Neocortex, hippocampus, amygdala, basal forebrain	Axotomy/target deprivation, excitotoxicity, aging nonhuman primates, transgenic mice
Amyotrophic lateral sclerosis	Adult (midlife)	Motor neurons in spinal cord and brain stem, motor cortex	Axotomy–target deprivation, excitotoxicity, transgenic mice
Parkinson's disease	Adult (midlife)	Substantia nigra dopaminergic neurons	Target deprivation, excitotoxicity, MPTP poisoning, transgenic mice
Huntington's disease	Adult (midlife)	Striatum (caudate nucleus)	Excitotoxicity, metabolic toxins, transgenic mice
Spinal muscular atrophy	Infancy–childhood	Motor neurons in spinal cord and brain stem	Axotomy–target deprivation, transgenic mice
Cerebral ischemia	Any age	Cerebral cortex, striatum, cerebellum, thalamus	Global cerebral ischemia, focal cerebral ischemia (stroke), excitotoxicity, axotomy–target deprivation
CNS trauma	Any age	Cerebral cortex, striatum, cerebellum, thalamus, spinal cord	Contusion–compression injury, excitotoxicity, axotomy–target deprivation

partly caused by a sublethal excitotoxic process involving NMDA receptors. Although GluR activation has been implicated in a variety of neurodegenerative processes within the CNS, these experiments have identified important links between degenerative processes in neurons caused by axon transection, target deprivation, and deafferentation (removal of inputs) and those caused by excitotoxic mechanisms that may be relevant to the neurodegeneration in AD.

It was discovered that the neuronal degeneration in ALS has structural features of apoptosis and molecular characteristics of PCD. Because ALS is an adult-onset neurodegenerative disease, and because of the newly developed concept of the apoptosis–necrosis continuum, it is necessary to study the mechanisms of neuronal apoptosis in the CNS of adult animals. Therefore, models of *bona fide* neuronal apoptosis in the adult brain and spinal cord have been developed. In one model, the neocortex is damaged by unilaterally ablating the visual cortex in the rat. This model axotomizes and deprives thalamic relay neurons in the dorsal lateral geniculate nucleus of their neocortical target. In another model, the sciatic nerve is removed by avulsion. This model axotomizes and deprives motor neurons in the lumbar spinal cord of vital muscle- and Schwann-cell-derived survival factors.

By using these models of retrograde neuronal degeneration, it has been discovered that neuronal apoptosis occurs in association with the accumulation of functionally active mitochondria and oxidative (free radical) damage to the DNA in the cell nucleus. An important new theory of neuronal degeneration has been developed on the basis of these animal models of definitive neuronal apoptosis. Mitochondrial accumulation within the neuronal cell body and increased cytochrome c oxidase activity may result in overwhelming generation of reactive oxygen species within the vicinity of the neuronal nucleus, depletion of mitochondrial and cytosolic antioxidant mechanisms, subsequent oxidative damage to proteins and nucleic acids, and failure of DNA repair mechanisms. The mitochondria become progressively damaged, as evidenced by their swelling and lysis of the inner membrane that manifest concurrently with incipient condensation of the neuronal cytoplasm, supporting the hypothesis that apoptosis-initiating factors are sequestered within the mitochondrial intermembrane space, which upon their release activate an apoptotic cell death cascade. Damage to DNA then activates a p53-dependent form of PCD in injured neurons. At present, it is still not known whether this process truly mimics the condition in humans with a neurodegenerative disease, but this idea is being actively studied.

The process of axotomy-induced neuronal apoptosis is associated with hydroxyl radical damage to DNA. Hydroxyl radicals are products of the transition metal (e.g., iron) catalyzed, Haber–Weiss- and Fenton-type reactions that use superoxide and hydrogen peroxide as substrates, respectively. Among the reactive oxygen species, hydroxyl radicals are highly reactive and are thought to be genotoxic by interacting with DNA and producing DNA strand breaks and base modifications. Our experiments are very important because they show, for the first time, the formation of hydroxyl-radical-modified DNA during the progression of neuronal apoptosis in animal models that are relevant to both AD and ALS.

B. Neuronal Degeneration in Models of Cerebral Ischemia

An interruption in the normal supply of oxygen to the brain resulting from abnormal blood flow or low amounts of oxygen in the blood causes cerebral hypoxia–ischemia. If the entire brain is affected, as in heart failure or asphyxiation, the abnormality is called global cerebral hypoxia–ischemia. If only specific regions are affected as in stroke, the perturbation is called focal cerebral ischemia. The mechanisms by which cerebral ischemia produces irreversible neuronal death are still not fully understood. Interestingly, only certain regions of the brain are selectively damaged by global cerebral hypoxia–ischemia (Table I).

We have studied neuronal degeneration by using models of global ischemia in experimental animals. The degeneration of pyramidal neurons in the neocortex and hippocampus, Purkinje cells in the cerebellum, and medium spiny neurons in the striatum after global ischemia (Table I) pathologically is a form of cellular necrosis rather than apoptosis. We have observed patterns of DNA fragmentation, with a progression that is very consistent with acute damage by reactive oxygen species and subsequent neuronal necrosis. The degeneration of selectively vulnerable neurons after ischemia (Table I) evolves in association with damage to organelles that function in protein synthesis, posttranslational modification, and secretion. Disaggregation of polyribosomes and fragmentation or vesiculation of the endoplasmic reticulum and Golgi apparatus are prominent examples of this subcellular pathology that would likely alter protein

synthesis–processing capabilities. This degeneration of neurons after ischemia is structurally similar in the different brain regions, regardless of the severity of the ischemia and whether the degeneration occurs acutely ($\leq$ 24 hr) or is delayed (3–7 days). Thus, the acute and delayed neuronal deaths after cerebral ischemia are structurally, and perhaps mechanistically, identical. In several models of global cerebral ischemia, neuronal degeneration in selectively vulnerable regions is indistinguishable structurally from the neuronal death caused by excitotoxicity, and, specifically, it closely resembles the neuronal necrosis caused by NMDA GluR activation in the adult brain.

In the same animal models, apoptotic cell death occurs in some neuronal groups that are not typically regarded as selectively vulnerable to ischemia. For example, subsets of granule cells in the hippocampal dentate gyrus and cerebellar cortex and subsets of neurons in the thalamus undergo apoptosis. In addition, prominent apoptotic death of white matter oligodendrocytes can occur. It has been theorized that apoptosis in these groups of cells after global ischemia is a form of PCD caused by target deprivation and axonal degeneration in response to necrotic degeneration of selectively vulnerable populations of neurons (Table I). Thus, necrosis and apoptosis, occurring as secondary degeneration in remote or distant brain areas resulting from target deprivation, both contribute to the neurodegeneration after cerebral ischemia.

Interestingly, cell death in the different settings of naturally occurring and pathological neuronal degeneration may be mechanistically distinct, or some of the underlying mechanisms in these different settings of neuronal degeneration may be shared to varying degrees, but they differ in the rate at which the primary mechanisms of injury occur (e.g., how fast an oxidative stress occurs). For example, oxidative stress and free radical damage in neuronal death caused by target deprivation–axotomy evolve slowly, and the corresponding neuronal death is apoptosis; in contrast, if oxidative stress evolves rapidly, as in the death of striatal neurons after ischemia, then the neuronal death is necrosis.

An important, unresolved question regarding the pathobiology of neuronal degeneration is whether the absence of the classic apoptotic structure in vulnerable populations of neurons is sufficient evidence to exclude the possibility that PCD participates in the pathogenesis of neuronal degeneration. All forms of PCD may not occur via apoptosis. For example, the death of T cells during negative selection in mouse thymus and the death of intersegmental muscles during metamorphosis in the moth *Manduca sexta* both occur by PCD; however, these cell populations die with distinct structural and biochemical features. Thus, some cells can die by a PCD mechanism that is not associated with the structure of classic apoptosis. Clearly, much more work using animal and cell models is needed to answer this complex question as it may relate to neuronal degeneration in disorders of the human brain and spinal cord.

XI. THE FUTURE IS PROMISING FOR UNDERSTANDING THE CAUSES OF NEURODEGENERATIVE DISORDERS IN HUMANS AND IDENTIFYING TREATMENTS

Unfortunately, no effective treatments are available to prevent or cure any of these neurological disorders that affect humans. Molecular genetics studies are likely to identify additional gene mutations or deletions that are associated with progressive neurodegenerative disorders (Table II). Animal models of experimental neuropathology will remain critical for studying how nerve cells die in the brain and spinal cord. Neurons within the CNS can be damaged by neurotoxin exposure, by cutting nerves and white matter pathways, by reducing blood flow to the brain producing oxygen deprivation, by forcing brain cells to express mutant genes, and by deleting genes. These animal models of neuronal degeneration are relevant to several neurological disorders that affect humans, including Alzheimer's disease, ALS, and cerebral ischemia (Table VIII). A better understanding of the pathogenesis of neuronal degeneration in human diseases and in animal models that mirror this degeneration is critical for the future development of effective therapies for patients with Alzheimer's disease, ALS, and other conditions. A more complete understanding of the characteristics and mechanisms of neuronal death within the CNS, provided by animal models of injury and transgenic gene expression–deletion, is important for the subsequent development of effective pharmacological and biological treatments to prevent or limit neurodegeneration in human neuropathological conditions.

It is still generally believed that, once neurons die in the mature CNS of mammals, these neurons cannot be replaced by nearby neurons (unlike the liver for example) because the remaining neurons are postmitotic (i.e., after neurons have achieved their mature

state, they cannot enter back into the cell cycle and undergo cell division). However, nervous tissue grafting or stem cell implantation as a means for neuronal replacement is an experimental approach that offers some hope, though much work needs to be done to evaluate the feasibility and efficacy of this approach. Importantly, it remains to be shown whether the immature or adult CNS provides a permissive environment for the appropriate integration of exogenous or grafted cells into a functional neural system.

See Also the Following Articles

ALZHEIMER'S DISEASE, NEUROPSYCHOLOGY OF • BASAL GANGLIA • GLIAL CELL TYPES • MOTOR CONTROL • MOTOR NEURON DISEASE • MOTOR SKILL • MULTIPLE SCLEROSIS • NEURAL TRANSPLANTATION • NEURON • NEUROTRANSMITTERS • PARKINSON'S DISEASE • STROKE

Acknowledgments

Dr. Martin is supported by grants from the U.S. Public Health Service, National Institutes of Health, National Institute of Neurological Disorders and Stroke (NS 34100), National Institute on Aging (AG16282), and the U.S. Army Medical Research and Material Command (DAMD17-99-1-9553). The author is grateful for the expert assistance of Ann Price, Frank Barksdale, and Adeel Kaiser, and he thanks Drs. Carlos Portera-Cailliau, Stephen Ginsberg, Chun-I Sze, Akiko Furuta, Nael Al-Abdulla, Frances Northington, JoAnne Natale, and Ansgar Brambrink for their contributions to his laboratory. This chapter is dedicated to my wonderful twin daughters, Gabrielle and Isabella.

Suggested Reading

Martin, L. J., Pardo, C. A., Cork, L. C., and Price, D. L. (1994). Synaptic pathology and glial responses to neuronal injury precede the formation of senile plaques and amyloid deposits in the aging cerebral cortex. *Am. J. Pathol.* **145,** 1358–1381.

Martin, L. J., Brambrink, A. M., Lehmann, C., Portera-Cailliau, C., Koehler, R., Rothstein, J., and Traystman, R. J. (1997). Hypoxia–ischemia causes abnormalities in glutamate transporters and death of astroglia and neurons in newborn striatum. *Ann. Neurol.* **42,** 335–348.

Martin, L. J., Al-Abdulla, N. A., Brambrink, A. M., Kirsch, J. R., Sieber, F. E., and Portera-Cailliau, C. (1998a). Neurodegeneration in excitotoxicity, global cerebral ischemia, and target deprivation: A perspective on the contributions of apoptosis and necrosis. *Brain Res. Bull.* **46,** 281–309.

Martin, L. J., Portera-Cailliau, C., Ginsberg, S. D., and Al-Abdulla, N. A. (1998b). Animal models and degenerative disorders of the human brain. *Lab Animal* **27,** 18–25.

Martin, L. J. (1999). Neuronal death in amyotrophic lateral sclerosis is apoptosis: Possible contribution of a programmed cell death mechanism. *J. Neuropathol. Exp. Neurol.* **58,** 459–471.

Martin, L. J., Price, A. C., Kaiser, A., Shaikh, A. Y., and Liu, Z. (2000a). Mechanisms for neuronal degeneration in amyotrophic lateral sclerosis and in models of motor neuron death. *Int. J. Mol. Med.* **5,** 3–13.

Martin, L. J., Sieber, F. E., and Traystman, R. J. (2000b). Apoptosis and necrosis occur in separate neuronal populations in hippocampus and cerebellum after ischemia and are associated with alterations in metabotropic glutamate receptor signaling pathways. *J. Cereb. Blood Flow Metab.* **20,** 153–167.

Martin, L. J., Brambrink, A. M., Price, A. C., Kaiser, A., Agnew, D. M., Ichord, R. N., and Traystman, R. J. (2000c). Neuronal death in newborn striatum after hypoxia–ischemia is necrosis and evolves with oxidative stress. *Neurobiol. Disease* **7,** 169–191.

Mouton, P. R., Martin, L. J., Calhoun, M. E., Dal Forno, G., and Price, D. L. (1998). Cognitive decline strongly correlates with cortical atrophy in Alzheimer's disease. *Neurobiol. Aging* **19,** 371–377.

Portera-Cailliau, C., Price, D. L., and Martin, L. J. (1997). Excitotoxic neuronal death in the immature brain is an apoptosis–necrosis morphological continuum. *J. Comp. Neurol.* **378,** 70–87.

Sze, C.-I., Troncoso, J. C., Kawas, C., Mouton, P., Price, D. L., and Martin, L. J. (1997). Loss of the presynaptic vesicle protein synaptophysin in hippocampus correlates with cognitive decline in Alzheimer's disease. *J. Neuropathol. Exp. Neurol.* **56,** 933–994.

Sze, C.-I., Bi, H., Kleinschmidt-DeMasters, B. K., Filley, C. M., and Martin, L. J. (2000). Selective regional loss of exocytotic presynaptic vesicle proteins in Alzheimer's disease brains. *J. Neurol. Sci.* **175,** 81–90.

Neurofeedback

JAMES R. EVANS
University of South Carolina

GLOSSARY

amplitude Height of an EEG wave from baseline, reflecting the magnitude of voltage in the EEG.

artifact Contamination of an EEG wave by factors other than brain electrical activity, e.g., eye or other body movements, muscle electrical activity.

canned protocol A very specific neurofeedback treatment plan often applied in a rigid manner with clients.

electrode A small metallic conductor. These are placed on the scalp at points where EEG activity is to be detected and conducted to other processing equipment.

empirical Derived from experience and/or scientific experiment.

entrain To cause the interaction of one organismic rhythm with another, resulting in identical or related rhythms (frequencies) between the interactants.

filtered Having certain frequencies removed from the full EEG frequency spectrum. For example, frequencies above 32 Hz often are removed prior to further EEG analysis.

Fourier analysis A mathematical process, named after the nineteenth century French physicist, J. B. J. Fourier, enabling analysis of a complex waveform into simple sine waves. This permits the calculation of power (voltage) in a specific EEG frequency band(s) of interest.

frequency band A specified portion of a total frequency range, e.g., the 8.0–12.0 Hz portion of an entire 1.0 to 32.0 Hz EEG record.

Hz Abbreviation for hertz. A frequency unit of 1 cycle per second, named after the nineteenth century physicist, Heinrich Hertz.

international 10–20 electrode system A standardized system for locating EEG electrodes on the scalp, which corrects for individual differences in skull size.

phase A stage in anything that is cyclic in nature, e.g., an EEG wave. Two waves are synchronized if a given stage occurs simultaneously in each.

placebo effect Changes following a treatment that are due to the suggestion or expectancy of change.

Neurofeedback is a specific case of biofeedback in which the physiological process involved is the electrical activity of the brain. This article presents an overview of the field of neurofeedback.

I. INTRODUCTION

As is evident from other articles of this text, the twentieth century saw major advances in knowledge of the functioning of the human brain. Many such advances were associated with technological developments in computer science and brain imaging techniques, such as functional magnetic resonance imaging (fMRI), and with developments in psychopharmacological treatments in psychiatry. Similarly, twentieth century advances in the cognitive sciences have been associated with the development of effective cognitive and behavioral therapies for mental disorders and cognitive rehabilitation techniques for acquired brain damage. However, despite the many advances, treatment of psychiatric type disorders still primarily involves temporary control of the symptoms rather than permanent change in the underlying neurophysiology or actual cure. Pharmacological treatments have unwanted and often dangerous side effects,

whereas behavioral changes resulting from behavior modification procedures often fail to generalize to a sufficiently wide range of life situations, and cognitive rehabilitation has had limited success in many cases of traumatic brain injury. Furthermore, some neurophysiological diseases such as Alzheimer's disease have responded to treatment only marginally. Obviously, the need remains for new and/or improved treatment procedures for disorders now known to have neurophysiological bases (or correlates). Many alternative treatments have been proposed, with advocates of each claiming positive effects on neurophysiological function, e.g., biofeedback, chiropractic care, music therapy, meditation. This article is about one of the more recent ones: neurofeedback (NF) (sometimes referred to as neurotherapy or EEG biofeedback). Biofeedback may be defined generally as a research or clinical treatment technique in which electronic equipment presents a person with information about ongoing physiological processes that otherwise could not be monitored readily, thus allowing the person to learn through operant conditioning to self-regulate those processes. NF is a specific case of biofeedback in which the physiological process involved is the electrical activity of the brain.

Due in part to the newness of NF as a treatment modality, there is considerable controversy regarding which topics are of most importance to the field. Topics selected for inclusion in this article and many of the ideas presented are based on the author's 30 years of following developments in EEG biofeedback and several years of experience with NF in a part-time private practice in clinical psychology. The content also is based in part on the author's experiences in co-editing, with Andrew Abarbanel, the text *Introduction to Quantitative EEG and Neurofeedback*, and in part on the results of a survey questionnaire concerning the current status of NF returned to the author by 10 of today's leading NF practitioners. Hopefully, this article presents an accurate overview of the field of NF as it exists at the beginning of the twenty-first century. The author, however, assumes sole responsibility for the content, which may be perceived by some as neglecting certain important topics and misrepresenting or overemphasizing others.

A. Procedures and Premises

In the practice of clinical NF, a client requesting treatment first is evaluated to determine whether NF is an appropriate primary or adjunctive treatment modality. This involves considerations such as specific symptoms, likelihood of motivation and ability to persist in attempting to modify his or her EEG over the usual 20–50 sessions required for optimal results, use of prescribed (or illegal) pharmacological agents, and co-existing and contraindicative psychosocial, environmental, psychiatric, and/or general medical problems (e.g., certain types of seizures) for which a referral for mental health and/or medical treatment is needed. To obtain this information and other data that may be used for pre- and posttreatment comparisons in evaluating progress and outcome, practitioners often use structured interviews, behavior rating scales, and various standardized psychological tests such as continuous performance tests. Many NF practitioners also require an EEG assessment prior to beginning treatment. This most often involves a quantified (quantitative) EEG (QEEG) evaluation and comparison of obtained measures of wave frequency, amplitude, phase relationships, and coherence between scalp electrode sites to a lifespan database to determine which, if any, are abnormal and, hence, candidates for attempting normalization through NF.

Actual treatment begins once a client is considered acceptable for this type treatment and the EEG parameters to be modified are decided upon. An electrode (or more than one electrode in some applications) is placed at an appropriate scalp location, the EEG activity emanating from that location is filtered and amplified, and a feedback signal is presented to the client reflecting the current status of a selected feature (or features) of the ongoing EEG. The EEG activity actually eliciting the feedback varies with equipment. After filtration and amplification, some equipment provides a direct analogue or digital representation of the raw EEG in the feedback. Most often, however, the raw EEG signal is averaged over short periods of time (epochs) and a Fourier analysis is made for each epoch. It is these averaged and transformed data that are reflected by the feedback.

NF clinicians vary in their rationale for selecting a specific scalp electrode site(s) for training. Some train at sites from the international 10–20 electrode system where QEEG abnormalities were greatest for the client, whereas others regularly use specific sites identified through their clinical experience (or the experiences of others) to be the most effective for specific conditions, e.g., Cz for attention problems. Feedback usually is auditory (tones varying in pitch and/or volume) or visual (e.g., moving colored bars, puzzles, or animations are presented on a computer monitor). With most modern equipment, feedback

possibilities are nearly endless considering all the possible pitch, volume, and visual stimulus combinations. Various combinations usually are tried during early sessions until one considered acceptable to the client is agreed upon. Most often, feedback is given for desired changes in the amplitude of specific EEG frequency bands (or duration of specified frequency amplitudes) over the course of a training session, e.g., for changes in commonly used EEG frequencies such as δ (0.5–4 Hz), θ (4–8 Hz), α (8–12 Hz), sensorimotor rhythm (12–15 Hz) and/or β (>12 Hz bands). In some treatment protocols, feedback is provided to facilitate increased amplitude in one frequency band and simultaneous decreases in another, e.g., increase β–decrease θ. Although wave amplitude is the EEG characteristic presently involved most often in NF, degree of coherence, i.e., correlation and similarity of waveform morphology, phase lag, i.e., transmission time in milliseconds, between two (or more) scalp electrode sites, and various other EEG measures increasingly are being used. As with electrode site selection, there is considerable variability in the rationale for selecting a specific EEG feature for feedback. Some practitioners base selection on QEEG findings on the individual client, whereas others regularly use ''canned'' protocols based on clinical lore regarding what is effective for a given condition. Modern NF equipment also monitors EEG artifacts such as those caused by eye or other body movements or muscle tension (electromyogram or EMG) and inhibits feedback when such artifacts are above a level preset by the neurotherapist. Thus, feedback is provided only when the client's target EEG feature is changing above (or below) a desired threshold and artifacts are below a preset inhibit level. Typically, a cumulative score is kept of the percentage of time that the client received feedback during a session. Neurotherapists generally agree that, for optimal learning, thresholds and artifact inhibit levels should be set so that the client is able to meet all criteria and, thus, obtain feedback 70–80% of the time. As learning occurs, thresholds and inhibit levels may be made increasingly challenging. There is general agreement that training sessions should occur 2–3 times per week, especially during the earlier and middle portions of the total treatment.

The basic premises of NF, in many respects, are quite simple:

1. Functioning of the central nervous system is electrochemical in nature. It is commonly accepted that changes in neurochemistry can result in changes in behavior, perception, emotions, and cognition. It is logical to assume that changes in electrophysiology (i.e., EEG activity) also can result in such changes.
2. If an appropriately motivated person is provided immediate, accurate feedback regarding aspects of his or her ongoing brain electrical activity (EEG), he or she can learn to regulate it, i.e., operant conditioning of the EEG activity can occur.
3. If such operant learning is repeated a sufficient number of times, the electrophysiological changes will become quite stable.
4. Appropriate stable changes in brain electrophysiology can result in relatively permanent positive changes in perceptual, cognitive, and behavioral functioning across various settings and activities of daily life.

II. HISTORICAL NOTES

A. Early Beginnings

Until about 40 years ago, it was generally believed that brain electrophysiology was not under conscious control and could not be modified through learning. In fact, even today despite much evidence that EEG can be modified rather easily, there are some who assert that this occurs only through mediation by voluntary control of a skeletal muscle response, e.g., deep breathing. The birth of the field of neurofeedback (NF) commonly is reported to have occurred with a serendipitous discovery in the early 1960s. A University of Chicago physiological psychologist, Joe Kamiya, noted that some persons could consciously control bursts of α frequency EEG if asked to do so by simply observing their ongoing EEG tracings. This was quite a revolutionary idea at the time and led to a large amount of research and clinical work on EEG biofeedback through the 1970s.

Researchers and clinicians in both medicine and psychology claimed to have demonstrated that most persons not only can learn to self-regulate aspects of their EEG but that certain modifications (especially α-frequency amplitude and duration at certain scalp sites) are associated with major changes in consciousness, including perceived decreases in anxiety. For example, Tom Budzynski and Johan Stoyva published results of successful use of biofeedback to increase α levels in conjunction with systematic desensitization therapy for a phobic type condition. Barry Sterman

reported decreases in epileptic seizure activity, and Joel Lubar and M. N. Shouse published evidence of positive behavioral changes in a ''hyperkinetic'' child, with both of these being associated with training of the sensorimotor rhythm (SMR) and inhibiting θ activity at central scalp sites.

Under the leadership of Barbara Brown, a biofeedback research society was formed in 1969, and several books on biofeedback were published soon thereafter. However, during the 1970s there was increasing criticism of EEG biofeedback. For many it became associated with ''consciousness raising'' and other ''New Age'' movements of the time, and the claims of its advocates often were strongly attacked by researchers publishing in a few prestigious journals. For example, one especially influential article reported that biofeedback did not enable persons to increase α frequency amplitudes beyond those normally occurring when simply sitting in a dark room with eyes closed.

By the late 1970s the earlier promise of EEG biofeedback had faded. Although the larger field of biofeedback (e.g., temperature, heart rate, muscle tension feedback) continued to prosper to some degree, proponents and practitioners of EEG biofeedback were few. Joel and Judith Lubar at the University of Tennessee and Southeastern Biofeedback Institute pursued research and treatment with children having learning and attentional disorders, Lester Fehmi continued to use multi-EEG channel synchrony training to develop flexible attention, and Margaret Ayres reported major success using EEG biofeedback equipment of her own design with victims of head injury and stroke. Dale Walters and Alyce and Elmer Green continued to use and teach these procedures at the Menninger Institute in Kansas. During the early 1980s, Charles Stroebel and Adam Crane continued to perfect equipment for providing feedback of complex patterns of EEG activity.

B. Revival

A major turning point for NF came in the late 1980s largely due to several nearly simultaneous events: publications by Joel and Judith Lubar detailing successful treatment of attention deficit disorder with EEG biofeedback, a research publication by Eugene Peniston and Paul Kulkosky, and the development of high-quality, relatively low priced equipment for both QEEG evaluation and EEG biofeedback by companies such as Lexicor Medical Technology of Boulder, CO. The 1989 Peniston and Kulkosky research combined conventional treatment and relaxation and temperature biofeedback with NF in treating 10 male alcoholics and compared the results to those of a conventional treatment only control group. The finding of much better and more enduring results with the experimental group aroused strong interest in the potential of NF for treating this notoriously treatment-resistant group. During the 1990s, there was what some describe as an explosion of interest in NF that continues today. A few highlights will be mentioned briefly in the next few paragraphs.

Perhaps to shed an earlier negative image, the terms neurofeedback and neurotherapy largely replaced the term EEG biofeedback by the early 1990s, with the latter now referring primarily to research rather than clinical applications. Whereas scientific research activity in this area is growing but remains modest in extent, the clinical use of NF is spreading very rapidly, both nationally and internationally. It is estimated that approximately 1500 clinics and practitioners now use it as one of their treatment modalities. Some of the pioneers continue to expand their NF practices (e.g., Joel and Judith Lubar at Southeastern Biofeedback Institute and Margaret Ayers at Neuropathways), and there are many relative newcomers with large practices. Notable among the latter are Seigfried and Susan Othmer, who have trained many of today's practitioners and help new trainees establish practices throughout the country.

An early professional group, the Biofeedback Society of America, and its descendant and largest biofeedback association, the Association for Applied Psychophysiology and Biofeedback (AAPB), have always had members with interests in EEG biofeedback. However, until about the mid-1990s, many of these members apparently felt treated as ''second class citizens,'' perhaps related to the negative image EEG biofeedback had acquired during the 1970–1990 period and mainstream biofeedback's desire to avoid it. Partially as a result, a separate group was formed in 1992 and presently is known as the Society for Neuronal Regulation (SNR). Consisting of 450 members in 2001, it sponsors annual national conferences at which NF research and related developments are presented. At approximately the same time, AAPB changed its structure to allow for specialty interests, and in 1993 an EEG interest group was formed. This EEG Biofeedback Division within AAPB also has grown rapidly and now is well-accepted, as evidenced by the election of the NF pioneer Joel Lubar as the president of AAPB in 1996. In 1995 a journal devoted

entirely to NF-related topics, the *Journal of Neurotherapy*, was first published. It now has a circulation of 450 and is abstracted through the American Psychological Association's "Psychlit" service. The first text in the field, *Introduction to Quantitative EEG and Neurofeedback*, edited by this author and Andrew Abarbanel, was published by Academic Press in 1999. Several NF websites are operative, including that of SNR: www.snr-jnt.org. Although proper training and credentials continue to be controversial topics, more states are including biofeedback (and thus neurofeedback) as practice activities in licensing laws for professionals such as psychologists and counselors. Several groups provide intensive training seminars, and NF certification is available through the Biofeedback Certification Institute of America (BCIA).

III. SPECIFIC CLINICAL APPLICATIONS

As noted earlier, training of the amplitude of specific EEG frequencies presently is the most common NF application. This often involves simultaneous training for increasing the amplitude of one frequency and decreasing the amplitude of another. That is, some protocols train for changing amplitude ratios between two frequencies, e.g., lowering the θ:β ratio. Within the general area of frequency amplitude training, there are two often independent groups of NF practitioners: (1) those who emphasize training to increase the amplitude of higher (β) frequencies (and, in some protocols, decrease lower frequencies) and (2) those who train increases in the amplitude of lower frequencies (α, θ). The first group, typified by those who treat attention deficit disorder or seek to develop "peak performance," is interested in developing higher levels of arousal in presumably underaroused (or at least not optimally aroused) cortical areas. This is based on the fact that higher levels of brain activation are reflected in higher concentrations of high-frequency EEG. The second group typically is interested in "quieting" an overactive central nervous system and/or inducing an altered state of consciousness for therapeutic purposes. For example, training to increase the amplitude of the α frequency may be part of a relaxation training program, whereas the increasing θ (or α–θ) amplitude may be aimed at helping the client enter a state in which early childhood trauma may be reexperienced and resolved, with concomitant psychological reintegration.

A. Attention Deficit Disorder

Among NF practitioners there seems to be nearly universal agreement that this can be a highly effective treatment for many cases of attention deficit hyperactivity disorder (ADHD) in both children and adults. This may be especially true for cases of ADHD that involve cortical underarousal, and a growing body of research using brain imaging techniques such as functional magnetic resonance imaging (fMRI) and QEEG suggests that these may constitute the majority of cases. Relatedly, among the first and still most commonly used NF protocols for attention disorders is one designed to increase cortical activation (β or SMR amplitude) at frontal and/or central (e.g., Cz) scalp sites (often with simultaneous training of decreases in θ or α and θ amplitude). Joel and Judith Lubar pioneered this protocol, which in its original or slightly modified form continues to be considered highly effective. Some practitioners believe that NF can be effective as the sole treatment procedure, but most appear to advocate its use as a supplement to stimulant medication and/or psychological techniques such as family therapy, education about ADHD, and cognitive behavioral modification procedures. Indeed, some report past experiences in which a chaotic family situation seemed to prevent NF progress and refuse to treat children from highly dysfunctional families until that situation is addressed either before or during NF.

Whereas concurrent use of stimulant medication is often recommended and sometimes considered essential to enable the patient to attend sufficiently to the feedback stimuli during early sessions, there are frequent reports that less medication is needed as treatment progresses. In fact, some practitioners have reported that failure to adjust medications during treatment interferes with NF progress. In some cases the client reportedly has been able to cease the use of medication while remaining symptom-free.

The author's survey of highly experienced NF practitioners mentioned earlier found 42 NF treatment sessions (range = 25–60) to be the average number reported to be necessary for optimal results with attention deficit disorder without hyperactivity (slightly more when hyperactivity also was present). Many practitioners recommend having the client engaged in "challenge tasks" such as reading, writing, and listening during NF training, especially during later sessions. This is believed to facilitate the generalization of desired EEG changes to relevant real life situations. Ten-year (and more) follow-up of successfully treated clients has indicated that they remain symptom-free,

suggesting a permanent "cure." However, it also commonly is recommended that some clients return a few times annually for "tune-up" sessions to help reinforce the ability to appropriately self-regulate the EEG. All practitioners have encountered situations, in which the usual protocols for ADHD were ineffective. When this occurs, common recourses are to reevaluate the QEEG findings to determine whether other protocols might be more appropriate (e.g., training coherence between certain scalp sites, decreasing rather than increasing activation at specific sites) and checking further into the client's background to determine whether factors such as lack of motivation, use of illegal drugs, or family, psychiatric, or other medical problems are precluding progress (or are the true cause of the ADHD symptoms).

Although there are some vocal critics whose views range from considering NF with ADHD a promising but experimental procedure in need of more research to considering it a passing fad capitalizing on placebo effects, many case reports and a few small-scale studies involving control groups have supported the efficacy of NF for treatment of ADHD. A large multisite research study on NF and ADHD currently is in progress, and results should lend greater credibility to NF as a useful treatment. However, the definitive large-scale study incorporating double-blind procedures demanded by some NF critics remains to be done. Although ADHD is the disorder for which NF currently most often is used, several others will be discussed briefly in the following paragraphs.

B. Traumatic Brain Injury

Since the 1980s, Margaret Ayers has reported several studies successfully using NF to bring patients out of long-standing coma states and, in some instances, to help those with stroke and traumatic brain injury (TBI) return to premorbid levels of function. Several NF clinicians such as Daniel Hoffman and Steven Stockdale have published evidence that the majority of TBI clients undergoing NF treatment report improvement in cognitive functioning, headaches, and/or ability to relax along with significant normalization of EEG. Whereas these reports often have been considered incredible by mainline medicine, they seem less surprising in view of scientific evidence for nerve regeneration or neural reorganization and recovery of neurological function even in older adults. Given the fact that strokes, TBI, and other neurological disorders (depending on age, location of injury, etc.) can result in an almost infinite variety of EEG disturbances and behavioral disorders, it is not surprising that there is little consensus among NF clinicians regarding best treatment protocols. Some note the finding that excessive EEG slowing (as evidenced by abnormal increases in δ, θ, and/or low α frequency amplitude) occurs at sites of brain injury and, accordingly, train for increased amplitude of higher (SMR or β) frequencies and/or decreases in lower frequency amplitudes at those sites. Others note the intra- and interhemispheric disconnections of neural communication that often occur in brain injury from shear and strain forces on short- and long-distance nerve fibers (and other factors), resulting in abnormal EEG coherence among certain sites. These clinicians tend to emphasize training to normalize QEEG coherence measures. Certainly it would seem that the wide variety of EEG abnormalities resulting from TBI and neurological disease requires QEEG or traditional EEG assessment prior to selecting a NF protocol for those type disorders. Whereas some practitioners have reported obvious and significant improvements in TBI clients after as few as 2–5 sessions (even years posttrauma), the experienced therapists surveyed reported needing an average of 53 treatment sessions (range = 30–80) for optimal success. Whereas much more quality research obviously is needed, the many clinical reports of NF successes with certain clients point toward its promise as a major form of treatment for TBI and other neurological disorders.

C. Depression

Some interest in using NF for treating depression followed reports by Davidson in 1995 that depressed persons, and even those in remission, show greater α amplitude at left frontal regions than homologous right frontal areas. This was interpreted as evidence that decreased left frontal activation is correlated with negative affect and has led to attempts to use NF to decrease left frontal α amplitude and thus decrease abnormal differences in frontal EEG asymmetry (e.g., the research of Elsa and Rufus Baehr). Whereas the need for such treatment may seem questionable given the effectiveness of present antidepressant medications, it may contribute to greater understanding of the causes of mood disorders and appears to have none of the side effects associated with medications.

Interestingly, there are anecdotal reports that depressed persons successfully treated with NF have remained symptom-free after ceasing to take

antidepressant medications. In fact, there also are reports that some developed severe flulike symptoms in reaction to their medications during the course of NF treatment. NF researchers and clinicians working in this area note that it is not effective in all cases and emphasize the need for research to determine which mood disorders are responsive, whether successful treatment is specific to the use of the asymmetry reduction protocol (as compared to other potential NF protocols), and the degree to which NF alone may be effective with mood disorders or needs to be part of a more comprehensive treatment protocol involving psychotherapy and/or medication.

D. Dissociative Disorders

Dissociative disorders occur when there is a disruption in the usually unified aspects of consciousness such as memory, affect, perception, and sense of identity. One of the most severe of these disorders is dissociative identity disorder (DID), formerly referred to as multiple personality disorder. Some of the earliest work involving NF and dissociative disorders was done with clients suffering from posttraumatic stress disorder for whom one of the symptoms was dissociation. More recently, some researchers and clinicians have reported successful treatment of DID using NF. Researchers and clinicians who have published reports of their work in this area commonly use other techniques to supplement NF. For example, Carol Manchester required subjects to learn to control body temperature through temperature biofeedback prior to NF and to use various visualization techniques. Along with other NF practitioners who work with DID clients, Manchester trained subjects to increase amplitudes of lower frequency EEG. A rationale for this is that a lower state of arousal is facilitated, which enables one to access repressed memories of traumatic events (often related to childhood abuse). With sufficient NF practice, persons are believed to be able to more readily "connect and disconnect" to psychobiological responses they had to previous psychological trauma and eventually resolve them, with accompanying reintegration of their separate personalities. Some have speculated that the increase in amplitudes of lower (low α–θ) frequencies during NF sessions results in an EEG pattern characteristic of childhood when the trauma occurred and, therefore, enables recall of the trauma (i.e., state-dependent memory).

Thomas Brownback and Linda Mason use similar procedures in their protocol for the treatment of dissociation but have added other components. For example, they train to increase θ amplitude at 12 different scalp electrode sites, incorporate the training of frequencies other than θ at points in their protocol, and use imagery, breathing techniques, psychotherapy, and other methods to achieve the goal of increased attentional flexibility in their clients. The use of NF in the treatment of DID reportedly cuts the time necessary for reintegration of separate personalities by about one-half. Because the usual time required for such reintegration ranges from 3 to 5 years, the value of NF in the treatment of dissociative disorders is worthy of further investigation. However, research is needed to help determine the relative importance of the various components of these treatment "packages."

E. Miscellaneous Applications

As mentioned in an earlier section, it was a report of the successful use of a treatment protocol involving NF with a group of alcoholic clients that stimulated revival of interest in NF during the late 1980s. There continue to be reports of the successful use of that protocol and others involving NF with alcohol and drug addiction. However, to the author's knowledge, there have been no published reports of controlled research on this topic. As with NF treatment of dissociative disorders, most of the treatment protocols have included other procedures, and the relative contribution of NF to treatment success is unknown.

A rather large number of NF practitioners are claiming significant improvement in clients diagnosed with chronic fatigue syndrome and in many diagnosed with fibromyalgia. Such reports have gained considerable attention because these disorders have proven difficult to treat through traditional medical procedures. Several years ago it was often reported that chronic fatigue syndrome was characterized by abnormally high amplitude EEG slow waves, especially at frontal sites, and NF treatment was designed to increase activation (raise β amplitude) at those sites. More recently, however, there seems to be disagreement among even experienced practitioners, who report that other abnormalities are common in chronic fatigue syndrome, including generalized EEG slowing, generally low EEG amplitudes, and unusual variability of brain electrical activity.

Another area in which NF reportedly has been successful is specific learning disability. Most often the anecdotal reports and papers presented at NF

conferences involve the successful treatment of dyslexia (specific reading disability). Unlike the protocols for the disorders mentioned earlier, which primarily involve amplitude training of specific frequencies, protocols often mentioned in the successful treatment of reading disorders have involved phase and/or coherence training. Usually the training is designed to help the client learn to modify phase and/or coherence relationships between left posterior scalp sites or between those sites and left frontal sites, specifically between sites overlying Broca's and Wernicke's areas. Modification of such relationships involves modification of neural timing relationships between the sites involved rather than changing amplitudes at those sites. Because considerable research suggests that neural timing abnormalities may be the basic causes of dyslexia, it seems reasonable that NF of this type would have the potential to change timing relationships in a favorable manner and, thus, facilitate reading readiness.

At SNR conferences there have been reports of the effectiveness of NF with various sleep disorders. In fact, 6 of the 10 experienced researchers surveyed prior to the preparation of this article reported using NF to treat sleeping disorders. Interestingly, they reported fewer sessions needed for positive results than with any of the other disorders mentioned in the survey (range = 10–30 sessions).

A few practitioners have reported the successful treatment of violent individuals, including some diagnosed with intermittent explosive disorder. QEEG findings among persons convicted of violent acts often show excessive slow wave activity and other abnormalities at right frontal and/or right anterior temporal sites. NF training targeting those sites (and other sites when appropriate) appears to have some promise in the rehabilitation of violent offenders and perhaps in the prevention of violence in those showing such EEG patterns but who have not yet displayed episodes of violent behavior.

No discussion of the present status of NF would be complete without mention of its use in facilitating optimal or peak performance in nonclinical populations. Such performance enhancement training using NF is a rapidly growing branch of the field, with its own certification requirements and treatment protocols. Treatment is aimed at improving self-discipline, attentional focus, flexibility, sustained alertness, visualization skills, and various other functions. Actual treatment procedures vary widely, but many practitioners concentrate on increasing the amplitude of β or α frequency and increasing phase synchrony among multiple sites or between homologous right and left hemisphere sites. Consumers of these services range from professional athletes seeking a competitive edge to amateur golfers aspiring to achieve respectable scores to persons from all walks of life who simply seek to improve efficiency.

IV. MODELS OF EFFECTIVENESS

Within the field there are various opinions regarding the mechanisms through which NF treatment leads to remission of symptoms, improvement in performance, etc. Whereas some feel that at this point in time the field needs to demonstrate definitively that it truly is an effective treatment rather than being concerned with why, nevertheless, considerable effort has been put into speculation about dynamics. Greater mainstream acceptance is likely if a credible model of efficacy can be formulated to help account for the often incredible results reported. Some of the more popular models will be mentioned here.

A. Site–Frequency Specificity

One model assumes that specific perceptions, cognitive events, emotions, and behaviors are enabled, facilitated, hindered, or precluded by certain brain rhythms at or between specific brain sites. Abnormal EEG activity at or between sites is seen as an underlying factor in disordered behaviors that can be normalized with NF, resulting in positive behavioral change. One of the more common examples of this type reasoning is that excessive power in the θ and/or lower α frequency ranges at central or frontal cortical sites hinders sustained attention, and learning to decrease power at those frequencies while increasing power in higher frequencies will enable the development of attentional skills. Similarly, such site–frequency specific effects often are implied in NF training to increase power in α or θ frequency bands at posterior sites. For example, it generally is assumed (with some empirical support) that high-amplitude EEG in the α frequency range facilitates a subjective state of relaxed attention in many persons. Hence, NF training to increase α power is likely to be recommended for a client with generalized anxiety who also has abnormally low EEG α (or abnormally high β) amplitudes at posterior sites. This is the same type of reasoning mentioned in the previous section where high θ amplitude is believed to enable memories of childhood events and therefore is the goal of NF training in some cases of dissociative disorder. Advocates of this model naturally tend to place much

emphasis on prior QEEG or EEG findings in order to more accurately "pinpoint" the underlying abnormalities of brain rhythm needing to be addressed. They expect to see increasing normalization of the EEG accompanying decreases in behavioral symptoms. Those who support this model see the development of QEEG procedures involving much larger numbers of electrode channels than the 19 commonly used now as promising to provide a degree of resolution that will more accurately delineate the site–frequency abnormalities to be modified through NF.

B. The Self-Regulating Brain

A second model centers around the view of the central nervous system as a self-regulating system and postulates much more generalized, systemic effects of NF training. This seems to be the model implied when practitioners make statements such as, "NF lets the brain see what is possible and then the brain proceeds to optimize its own function." Seigfried and Susan Othmer and associates have developed this theme in some detail. If one proceeds from a view that rhythmicity regulates and facilitates communication within the central nervous system, it is reasonable to assume that EEG abnormalities or dysrhythmias are indicative of central nervous system disregulation. A dysrhythmic, disregulated brain may be considered a disorganized, unstable brain, giving rise to disorganized, unstable thoughts, moods, levels of arousal, and behaviors. As viewed by Othmer and associates, NF challenges the self-regulating function of the brain. That is, by challenging the brain out of its present homeostatic state (which may be an unhealthy state), a new and improved homeostatic state and greater stability of the brain's regulatory function are attained. With such attainment, not only will specific symptoms such as those for which a client may have sought help be attenuated but there will be generalized improvement in functioning because appropriate neuroregulation is basic to all behaviors. Those who emphasize this model tend to be less concerned about matching NF training to specific site–frequency abnormalities on the basis of the assumption that normalization of any one of them may have the same effect, i.e., assisting the self-regulating brain to improve its neuroregulatory function.

C. Other Models

Outside the field, the positive results reported for NF training often tend to be attributed to other factors not directly related to EEG changes. For example, it sometimes is noted that being provided with reinforcement one or more times per week for attending closely to relatively boring stimuli for increasingly long periods of time could explain results. In other words, operant conditioning is occurring but quite independently from the specific EEG features concurrently being trained. This view is disputed by some who cite evidence that the specific EEG changes being trained are strongly correlated with expected changes in behavior. Nevertheless, NF practitioners often report cases in which behavioral improvements have occurred without changes in the targeted EEG features. It also is common to dismiss NF results as being due to placebo effects. In fact, it has been said that NF approaches the "ultimate placebo" in that it involves expensive, sophisticated electronic equipment and esoteric jargon and purports to make direct changes in brain function. As with any treatment technique, placebo effects can be expected. However, NF practitioners strongly refute the notion that this is the only or primary reason for results. They defend their position by citing cases in which positive effects went beyond anticipated behavioral changes, cases in which there were initial, temporary negative effects of the treatment, and the commonly observed course of behavioral change in which initial positive changes are followed by a plateau or increase in symptoms prior to a gradual decrease in symptoms.

Finally, there is the view that it is a sense of self-efficacy and self-confidence that comes with successfully mastering one's brain wave activity that leads to increased self-regulation of behaviors. This, however, seems unlikely to be a major factor inasmuch as children with ADHD very often enjoy and master computer games with no obvious improvement in symptoms.

V. CURRENT ISSUES

Considering its rapidly growing popularity as a therapeutic procedure for a wide range of disorders, it is not surprising that there are many controversial issues surrounding NF. Some of the more frequently encountered ones are discussed next.

A. Need for Prior Diagnosis

As noted earlier, there is controversy within the field concerning the need for a thorough diagnostic

assessment prior to beginning NF. This relates especially to the need for a prior QEEG evaluation. Those who believe that NF's success depends on the training of specific sites and frequencies at which abnormalities exist insist upon prior QEEG data (usually in conjunction with a normative database). This is commensurate with the common medical practice of seeking specific underlying causes for specific disease conditions. In this model, the diagnostic data not only guide details of treatment but also can be used to evaluate treatment progress periodically in terms of normalization of the presumed underlying causes of the disorder (i.e., abnormal brain electrical activity). Advocates of the view that NF has more generalized effects, with positive results being mediated by facilitation of the brain's self-organizing ability, generally have little or no concern about prior QEEG data. Some attention may be given to site–frequency, but more in relation to ''canned protocols,'' which have been observed clinically to be the most effective for a given condition or syndrome. Proponents of the need for prior QEEG evaluation generally see their position as more defensible from a scientific point of view and fear that failure to require a thorough diagnostic workup will result in overlooking specific EEG differences underlying abnormal behaviors in many clients and failing to detect conditions that need referral for medical treatment. They emphasize that, even though QEEG assessments are expensive, all NF practitioners need these data and should either purchase and become trained in the correct use of QEEG equipment or contract with others who can provide these data that they consider essential. The ''canned protocol'' and general effects advocates often counter this with claims that they obtain equally positive results with no more side effects and at much less cost to their clients. However, they have not published scientific evidence for this. Certainly both sides should agree that it would be unethical and dangerous to initiate NF without some type of evaluation to help determine whether conditions exist that are known to be effectively treated with medication or other treatment modalities. Whether prior QEEG data is essential in this regard apparently remains to be determined.

B. Decreasing Treatment Time

NF practitioners are well aware of a need to shorten treatment time. The average 40–50 sessions now commonly employed limit the number of clients a practitioner can treat and the number who can afford treatment. It also brings criticism from cost-conscious insurance companies and from NF critics who may see it as a ''fleecing of the public.'' However, exactly how to shorten treatment time is controversial.

Some NF practitioners claim that they have found the ''answer'' to decreased treatment time for some given disorder in a specific site–frequency combination. Some have even obtained patents on their specific protocols. Some claim that factors such as speed of feedback (time between desired change in a specific EEG frequency and presentation of feedback) is critical and that providing an optimal time interval will speed treatment. Others emphasize differences in how accurately a feedback signal actually reflects a desired EEG change. For example, NF equipment that processes EEG signals digitally and in real time rather than by use of averaging techniques and the fast Fourier transform is said by some to allow increased accuracy of the match between raw EEG activity and feedback and, hence, facilitate faster learning and decreased treatment time.

Another frequently advocated way to decrease treatment time involves supplemental use of auditory–visual stimulation (AVS). It is well-known that exposure to a flashing light can entrain EEG activity to conform to the frequency of the flashes. This has led some NF practitioners to attempt to entrain EEG activity in a desired direction (e.g., increased β frequency) by presenting lights flashing at the desired frequency to their clients in conjunction with or prior to NF training. In some cases this has included simultaneously presented auditory stimulation of the same frequency and is referred to generally as AVS. An assumption often made is that such exposure may serve as a sort of prime to facilitate a client's learning to achieve voluntary control of a desired EEG frequency, i.e., elicitation of the particular frequency via AVS may enable the client to develop a ''sense'' of the state associated with the frequency and, thus, gain enhanced ability to achieve it again during NF. Some have tried variations such as using AVS that varies in frequency during the stimulation period, e.g., beginning at 13 Hz and gradually progressing to 20 Hz, or starting at 13 Hz, progressing to 20 Hz, and then returning to 13 Hz. Still others have used EEG-driven AVS in which equipment detects EEG peak frequency at a given electrode site and then continually adjusts the AVS stimulation up (or down) by some standard increment to entrain gradually higher or lower frequencies. Some of these latter procedures have been patented. There also have been attempts to

add rhythmic tactile stimulation to AVS, resulting in even broader multiple simultaneous sensory stimulation.

Although some research supports the contention that AVS can result in temporary EEG changes in targeted directions and several NF practitioners report being able to decrease the number of treatment sessions significantly by the use of AVS, there is controversy about its use. There have been isolated reports of seizures being induced by this method. However, this apparently is extremely unlikely in clients with no prior history of seizures, and its use has been approved for research by human subjects ethics committees in several major universities. A concern heard more often is that, like many medications, its effects are temporary, whereas NF results in enduring changes because the client is learning voluntary control of EEG rather than having changes imposed on it. Some critics of AVS see it as, at best, serving as an adjunct to NF in difficult cases in which initial learning of EEG control does not occur and/or during times of prolonged learning plateaus.

AVS has promise as a means of expediting NF progress. However, as with NF, there are a large number of variables needing to be investigated, e.g., the importance of color and intensity of the visual stimuli, sound wave features and intensity of the auditory stimuli, and the nature of presentation of stimuli [e.g., alternating between right and left eyes (and/or ears) or simultaneous].

C. Training Requirements

Few states regulate the practice of NF, yet it is being used with clients with a wide range of medical and psychological conditions. Many both inside and outside the field express alarm about potential dangers to the public when practitioners with minimal or no training in neurology or neuropsychology, or even general medicine or psychology, are treating persons with ADHD, depression, brain injury, and other conditions.

Sometimes this is without supervision from or collaboration with trained professionals who normally would treat such clients. Although BCIA certification requires considerable training and supervised experience, some consider it to be too little and advocate the establishment of one or more new certifying agencies with higher standards. Some suggest that two or three levels of certification should be established, with the PhD or MD degree and extensive neurofeedback training and experience required for the highest level. Opponents of higher standards sometimes note that undesirable side effects of NF are rare and that, although it may seem incredible (and threatening to established health care professionals), training beyond the bachelor's degree level is not necessary to practice the art of NF and higher standards would be unfair to lesser educated practitioners and to the many persons in need of NF treatment. Although most, if not all, groups currently providing NF training limit enrollment to those with master's level or higher degrees, the field has an unusually diverse group of professionals, including physicians, professional counselors, psychologists, physical therapists, chiropractors, social workers, and nurses. Whereas NF may be a very safe procedure and it may not be essential that its practitioners be knowledgeable about details of EEG dynamics, neuroanatomy, or psychiatric diagnoses, having rigorous certification standards likely would aid in gaining greater acceptance of the field by insurance carriers, other professions, and the general public.

D. Miscellaneous Issues

There is some controversy among NF practitioners about which site–frequency combinations should be used for specific syndromes. For example, even in the case of ADHD in which the increase β (or SMR)–decrease θ band amplitude protocol is used extensively, there are some who argue that decreasing θ is sufficient or that increasing only SMR (or β) or some narrower band within one or the other of those frequency bands is most efficient and speeds treatment. To the author's knowledge, however, these claims have not been supported by scientific research. An area of some disagreement among practitioners is whether the time of day of NF treatment needs to be matched to the individual client to conform to his or her unique biorhythms. This concern also has been voiced by some in regard to traditional medical treatments and seems worthy of further consideration. The importance of personality characteristics of the NF practitioner also is debated. It seems essential that he or she be empathic, patient, and highly skilled at motivating the client to put adequate effort into the task, e.g., knowing when to adjust EEG target and artifact thresholds to maintain an optimal reinforcement rate. These characteristics seem especially critical when working with children.

Criteria for when to terminate NF also are controversial. Some have advocated the cessation of treatment after only a few sessions if no desired changes in targeted EEG features are observed, whereas others recommend continuing through 50 or more sessions, trying various site–frequency combinations (or phase or coherence training) even if no EEG or behavioral changes are noted. The latter may cite cases in which a positive response occurred quite suddenly only after a very large number of sessions. There have been some attempts to provide specific guidelines for termination. Joel and Judith Lubar, for example, suggest termination when a client's learning curve plateaus for at least 2 weeks (or after a minimum of 25 sessions), there is approximation of normative age-based EEG criteria, and maintenance of a 50–70% reward level and 30–50% inhibit activity.

VI. FUTURE DIRECTIONS

The history of self-regulation of brain waves (neurofeedback) has had some wavelike characteristics itself. As noted earlier, the field rose rapidly in popularity and peaked in the early 1970s, followed by the trough of the late 1970s to mid-1980s. Now its popularity is rapidly rising, with no sign of cresting. This section includes speculation on whether NF's popularity will continue to rise until cresting and persisting at a level reflecting its true value or whether it will again crash (temporarily or permanently).

The scenario of steady growth and continuing popularity is likely if there are continuing reports of successful clinical use backed by credible research findings, favorable media coverage, and increasing acceptance by insurance carriers and managed care groups. A "crest and crash" scenario is likely if controlled research findings unfavorable to NF are published in prestigious journals and publicized in popular media outlets. Related to this, there is considerable concern among NF practitioners that laboratory research on NF will be so controlled as to be "sterile" in nature and preclude the positive results commonly reported by clinicians. Practitioners often state that it is the total treatment situation that yields positive results, i.e., client–therapist rapport, ongoing adjustment of electrode placement, as well as the EEG feature being trained. Many also fear that there will be bias in the large-scale, federally funded research being called for by some NF critics because researchers may be major stakeholders in the present health care system (e.g., drug companies, physicians) who stand to profit from continuation of the status quo. Relatedly, NF advocates often report a reflexive, negative response when the topic is mentioned to members of some other professions. Responses appear to range from skepticism to derision and often include automatic dismissal of claims of NF efficacy. Whether this situation stems from perceived threats to the health care status quo, perception of NF claims as incredible, lingering memories of early 1970s association of biofeedback with "flower children," and/or perceived lack of data to support claims, it impedes professional development of the field. Such attitudes by leading professionals in related fields and in university training programs need to be modified if the field of NF is to attract bright, young persons who could help insure its future development.

EEG biofeedback as a treatment method seemingly has "risen from the ashes" under the new title NF and is rapidly growing in popularity as a sole or adjunct treatment for an increasingly large number of disorders, many of which have been resistant to more traditional therapies. Whereas there are reports of near-miraculous results with minimal or no side effects, NF practitioners often express dismay that their results are dismissed by mainstream medicine, psychology, and education rather than embraced as evidence that treating abnormalities of the electrical aspects of the electrochemical functioning of the brain can be at least as effective as treating chemical imbalances. Considering the rapidly increasing scientific evidence for neural plasticity, central nervous system influence on immune system function, and a rhythmic basis for much transfer of information within the brain (time-binding theory), it does not seem unreasonable to expect that modification of aspects of central nervous system rhythms could have profound effects on health and learning rates. However, before becoming overly eager about the unique value of NF, advocates should consider that several other alternative medicine fields (e.g., chiropractic, nutrition) claim major and near-miraculous treatment successes for an equally wide range of disorders. Some of these, such as music therapy, also may be modifying basic central nervous system rhythms through entrainment or other means. If NF is to maintain its current popularity and grow in acceptance by other professions, it must be demonstrated through well-designed, nonbiased research to be not only an efficient treatment but a unique one that offers advantages beyond those of other available treatments.

Acknowledgment

The author thanks Joel Lubar of the University of Tennessee for his helpful comments during preparation of this article.

See Also the Following Articles

ATTENTION • BIOFEEDBACK • DEPRESSION • ELECTRICAL POTENTIALS • ELECTROENCEPHALOGRAPHY (EEG) • EVENT-RELATED ELECTROMAGNETIC RESPONSES • NEUROIMAGING

Suggested Reading

Evans, J. R. (1998). Reflections on neurotherapy: Past, present and future. *J. Neurotherapy* **2**(4), i–vi.

Evans, J. R., and Abarbanel, A. (Eds.). (1999). *Introduction to Quantitative EEG and Neurofeedback*. Academic Press, San Diego, CA.

Kaiser, D. A. (2000). QEEG: State of the art or state of confusion? *J. Neurotherapy* **4**(2), 57–75.

Lubar, J. S. (1995). Neurofeedback for the management of ADHD disorders. In *Biofeedback: A Practitioner's Guide* (M. S. Schwartz, Ed.), 2nd ed. Guilford, New York.

Moore, N. C. (2000). Special issue on neurotherapy. *Clin. Electroencephalography* **31**(1).

Robbins, J. (2000). *A Symphony in the Brain: The Evolution of the New Brain Wave Biofeedback*. Atlantic Monthly Press, New York.

Sears, W., and Thompson, L. (1999). *The ADD Book*. Little-Brown and Co., Waltham, MA.

Striefel, S. S. (2000). The role of aspirational ethics and licensing laws in the practice of neurofeedback. *J. Neurotherapy* **4**(1), 43–55.

Thatcher, R. W. (2000). EEG operant conditioning (biofeedback) and traumatic brain injury. *Clin. Electroencephalography* **31**(1), 38–44.

Neuroglia, Overview

ANGUS M. BROWN and BRUCE R. RANSOM
University of Washington School of Medicine

GLOSSARY

astrocytes Neuroglial cells of diverse morphology and function located within the central nervous system.

enteric glia Neuroglial nonneuronal cells within the enteric nervous system.

microglia Neuroglial macrophage cells within the central nervous system.

neuroglia Nonneuronal cells located within the nervous system.

oligodendrocytes Neuroglial cells that myelinate axons within the central nervous system.

Schwann cells Neuroglial cells that myelinate axons within the peripheral nervous system.

This article provides an overview of glial cells, highlighting the distinct types of glial cells and describing their individual properties and functions, with a slant toward ongoing research in the authors' laboratory.

I. INTRODUCTION

Virchow initially described glial cells in 1846 as cellular elements that filled in the spaces between neurons but were distinct from neurons. Dieters further defined the description of glial cells in 1865 as cells that did not possess an axon. The concept that glial cells were a heterogeneous population of cells was realized by Golgi, who developed a technique of staining cells that allowed details of cellular morphology to be closely scrutinized. Golgi's studies indicated that glial cells had variation in their morphologies and led to subsequent subclassification of glial cells. Glial cells in fact outnumber neurons in a ratio of 10:1 and constitute 50% of the cellular volume of the central nervous system (CNS; not more than 50% because of their relatively smaller size compared to neurons).

The implication that glial cells act as a form of inert Styrofoam packing between neurons became an entrenched idea from these initial studies, and it is only within the last two decades that the true range and diversity of glial cell function have been fully realized. The impression of glial cells as inert entities with a purely supportive role in part may have arisen during the pioneering electrophysiological work in the 1930s to 1950s, when all interest was focused on cells that were electrically excitable and could fire action potentials. Because glia are incapable of firing action potentials, there was relatively little interest in them. It was not until the 1960s that the studies of Steven Kuffler suggested that glial cells, although incapable of firing action potentials, did indeed have specific roles in the nervous system. It is now realized that neurons cannot function as they should without normally functioning glial cells in close proximity. Under normal conditions there is a steady flow of information between neurons and glia. Glial cells are equally important under pathological conditions in the CNS.

They can contribute to the pathology, for example, by adding the excitotoxin glutamate to brain extracellular space (ECS) or potentially limit injury by minimizing ionic changes in ECS or providing energy substrate to neurons, which do not possess any internal energy stores. Thus, to understand pathological processes it is necessary to understand the glial reactions to the pathology. Current research on glial cells indicates that new properties and functions are still being assigned to them.

II. CLASSIFICATION

The classification of glial cells originally was based entirely on their morphological characteristics and dates back almost 150 years. Since then, additional techniques such as cell-specific proteins, electrophysiology, ultrastructural features, immunocytochemistry, and fluorescent imaging have been used to refine the classification of glia. Andreizen in 1893 first described two types of glial cells, fibrous and protoplasmic, a classification endorsed by Cajal in 1913. The initial characterization of glial cells was carried out by using light microscopy, but the advent of improved staining techniques at about the turn of the century accelerated the distinction of different cell types. It is now widely recognized that glial cells fall into the following categories: astrocytes, oligodendrocytes, and microglia in the CNS, Schwann cells in the PNS, and enteric glia in the gut. Astrocytes and oligodendrocytes are derived from neuroectoderm and are sometimes considered together as macroglia. The main functions of these cells can be categorized as follows: myelination of axons (oligodendrocytes in the CNS, Schwann cells in the PNS), maintenance of ionic and neurotransmitter stability of the brain's extracellular microenvironment (astrocytes), nurturing neurons (astrocytes), and acting as the CNS macrophages (microglia). Each cell type will be described in detail later in this article, however, a few brief words about the common features of glial cells are needed. One feature in which glial cells differ from neurons is their intercommunication. Neurons generally communicate at chemical synapses, where individual neurons use neurotransmitter molecules released into the synaptic cleft between the neurons as mediators of communication. Glial cells, especially astrocytes and oligodendrocytes, directly communicate with each other via gap junctions. Gap junctions are formed from protein molecules called connexins. These aggregate into transmembrane hexamers called connexons and fuse with connexons on the membranes of adjoining cells, forming large pores that allow the passage of substances up to 1 kDa in weight. In this way, glial cells can communicate directly and allow rapid intercellular transport.

III. ASTROCYTES

A. Anatomy

Astrocytes were originally described as filling the gaps between neurons. However, Golgi described in detail the morphological distinction between cell types in the CNS. Astrocytes can be divided into three groups: fibrous and protoplasmic astrocytes and radial glia. The latter cells project from the ventricular surface to the pial surface and are mainly prominent during early development. Fibrous astrocytes are located almost exclusively in white matter and have long, radially distributed processes that often terminate as end feet on capillaries. Protoplasmic astrocytes are located in gray matter and are characterized by short bushy processes; these processes also form vascular end feet. The three different types of astrocyte share common properties, including the synthesis of glial fibrillary acidic protein (GFAP). This protein is easily recognized in electron micrographs because it forms paired helical filaments, so-called intermediate filaments.

Astrocytes form connections with other astrocytes via gap junctions, and this allows intercellular communication that directly links the cytoplasm of the astrocytes. Neuronal communication, by comparison, is primarily mediated by chemical synapses. Gap junctions have a pore about 2 nm in diameter and couple astrocytes together into a syncitium, which allows the passage of all inorganic ions and small organic molecules up to 1 kDa. The use of dye injection into astrocytes allows an estimation of the extent to which the cells are coupled. The channels comprising gap junctions are large pores that extend from one cell across the extracellular space (ECS) into an adjacent cell. Each connexon hemichannel is composed of six symmetrical subunits (connexins). There are several different types of connexin, but the most common type in astrocytes is connexin 43. A gap junction is formed when a connexon from one cell aligns with a connexon from another cell. Gap junctions open and close abruptly, and the conductance varies between 50 and 150 pS depending on the type of connexin. Gap

junction openings are affected by transjunctional voltage and are decreased by increases in intracellular pH, Ca^{2+}, or octanol.

A striking feature of astrocyte morphology is the end feet surrounding capillaries. This morphology initially suggests that all substances in the blood must pass through astrocytes before being passed on to neural elements via the ECS. The astrocytic end feet do not represent a tight barrier, however, and substances that have passed the capillary endothelium (which is a tight barrier) can diffuse directly into brain extracellular space without passing through astrocytes. This notwithstanding, there is growing evidence that glucose and perhaps other blood-borne substances are taken up by astrocytes. Astrocytes possess glucose transporters, which would allow the movement of glucose from blood vessels directly into the capillaries. In the extreme, this would require that all neuronal nutrition be mediated via astrocytes (i.e., all glucose, the principal fuel of the brain, would go first to astrocytes). In actuality, some glucose undoubtedly bypasses the astrocytic end feet and diffuses directly to neurons, but astrocytes do provide fuel to neurons, probably in the form of lactate (see later discussion).

Astrocytes also seem to surround synapses and have been proposed as the main site of neurotransmitter uptake after synaptic activity. Astrocytes contain a variety of neurotransmitter transport mechanisms that are not found on neurons, in addition to enzyme systems that are capable of metabolizing neurotransmitters to more manageable inert compounds for transport back to neurons, where they are recycled back into the active neurotransmitter.

B. Function

Astrocytes share a common ECS with neurons, and virtually every neuron in the brain has a large percentage of its membrane immediately adjacent to astrocyte membrane. This situation lends itself to a role in regulating the ionic and/or chemical stability of extracellular fluid, which is necessary for normal neuronal function. In this capacity, it is true that astrocytes function to maintain constant ionic concentrations in the extracellular space by buffering K^+ and H^+, as well as by removing neurotransmitters. The membrane of astrocytes contains many transport and exchange mechanisms that allow the movement of ions into astrocytes. Astrocytes are the ideal cell type to buffer K^+ ions. K^+ accumulates extracellularly after neuronal activity, with the degree of activity determining the degree of K^+ elevation. It is thought that a single action potential can raise K^+ by as much as 1 m*M* in localized areas. After intense activity, $[K^+]_o$ can rise from a baseline level of 3 m*M* to about 12 m*M*, the so-called ceiling level. This ceiling level for $[K^+]_o$ is only breached with spreading depression and under pathological conditions such as trauma or ischemia, where concentrations of 60 m*M* or more may be recorded. There is some controversy as to how extracellular K^+ is removed or "buffered" during neural activity, but likely candidates include K^+ channels, Na^+ K^+ ATPase, and anion transporters such as the $Na^+K^+2Cl^-$ cotransporter. Once K^+ enters astrocytes, it can be dispersed via gap junctions to distant astrocytes and then released back into the ECS as levels return to normal (ultimately, of course, the neurons that released the K^+ in the first place must reaccumulate what was lost; see Fig. 1). Extracellular pH also changes with neural activity, and astrocytes have a unique transporter, the $Na^+HCO_3^-$ cotransporter, that helps to regulate this important ion level.

Astrocytic membranes contain neurotransmitter transporters. During neuronal activity synaptic neurotransmitters would rapidly build up if they were not sequestered intracellularly or metabolized

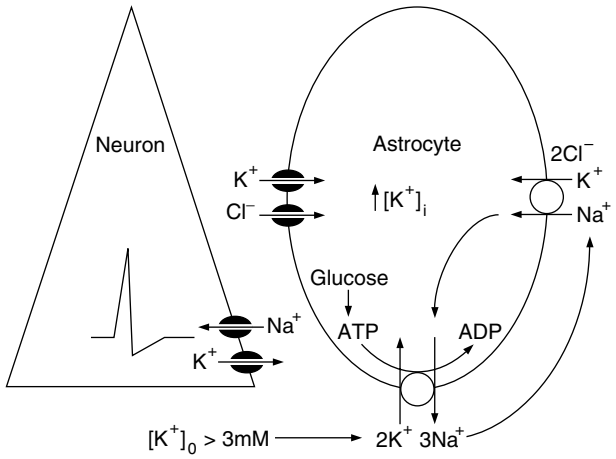


Figure 1 Schematic representation of mechanisms of K^+ uptake in astrocytes. K^+ released by active neurons is actively accumulated by astrocytes in three ways. The sodium pump and an anion transporter both take up K^+. The sodium pump relies directly on the availability of ATP, whereas the anion transporter is indirectly powered by the energy stored in the transmembrane Na^+ gradient. The presence of channels for Cl^- and K^+ allow Donnan forces to produce KCl influx. These mechanisms, along with K^+ spatial buffering (see text), prevent $[K^+]_o$ from exceeding $\sim$12 m*M*. Increases in $[K^+]_i$ are seen during neural activity as $[K^+]_o$ increases.

extracellularly. The most common excitatory CNS neurotransmitter is glutamate. Astrocytes take up ~90% of the glutamate released at synapses via two specific glutamate transporters called GLT1 and GLAST. The energy for this transport is provided by the transmembrane Na^+ gradient (see Fig. 2). Once inside the astrocyte the glutamate is converted to glutamine by the astrocyte-specific enzyme glutamine synthetase, resulting in the dephosphorylation of one molecule of ATP. The glutamine is shuttled out of the astrocyte and into the neuron by the glutamine transport protein, where it is converted back into glutamate by the enzyme glutaminase. Similar mechanisms exist for the uptake of the major inhibitory CNS neurotransmitter, GABA.

A major function of astrocytes is as an energy store. The full importance of this function is just being realized. Astrocytes are the only cells in the CNS that store glycogen, and as such astrocytes are the only cells in the CNS with energy stores. Glycogen is formed from glucose within astrocytes via the intermediary compounds glucose 6-phosphate, glucose 1-phosphate, and UDP-glucose in a reaction requiring ATP. Once glucose enters astrocytes it is immediately phosphorylated to glucose 6-phosphate, but the lack of the enzyme glucose-6-phosphatase, which converts glucose 6-phosphate to glucose, determines that glucose is not transported out of astrocytes. Glycogen is broken down to glucose by action of the enzyme glycogen phosphorylase. Thus, there is equilibrium between glucose and glycogen. In periods of high glucose the equilibrium favors the formation of glycogen (glycogenesis), and during periods of low glucose the equilibrium favors the breakdown of glycogen to glucose (glycogenolysis). It has been proposed that astrocytic glycogen can act as an energy source for neural elements in both gray and white matter during periods of energy deprivation.

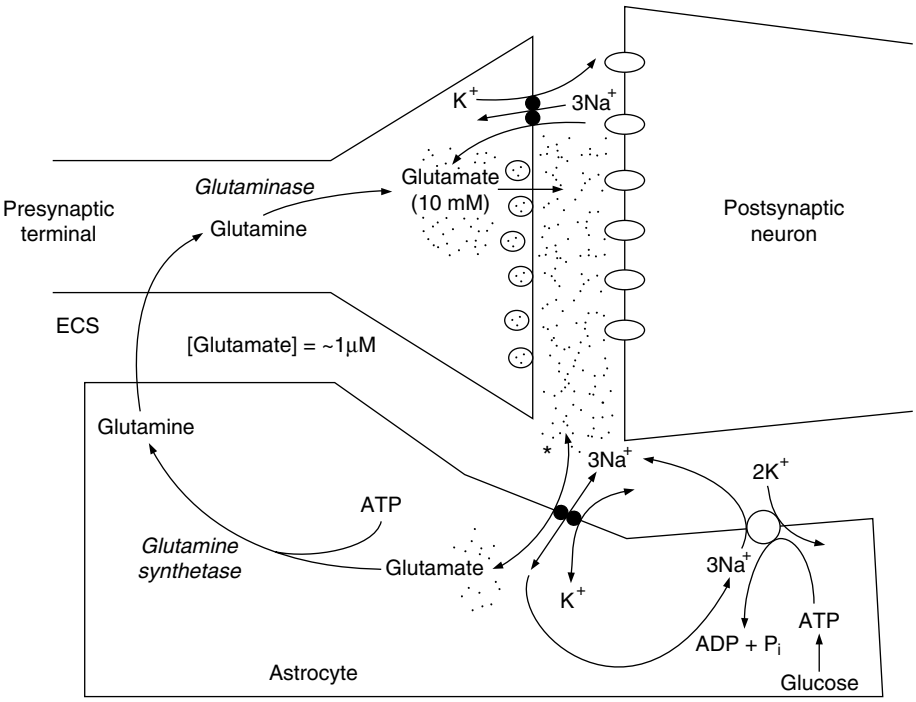


Figure 2 Scheme showing how astrocytes are involved in glutamate metabolism and uptake. Only astrocytes contain the enzyme glutamine synthetase, which converts glutamate to glutamine in an ATP-requiring reaction. Glutamine is transported to nearby presynaptic terminals, where it is converted to glutamate for synaptic release. Finally, the released glutamate is recaptured by astrocytes via a high-affinity glutamate uptake system. Although glutamate transporters are present in neurons, astrocytes are the most active in removing glutamate (see text). In the absence of the normal transmembrane Na^+ gradient maintained by the ATP-dependent Na^+ pump, the glutamate transporter ceases to remove glutamate and can run in reverse so that it pumps glutamate into the ECS.

Astrocytic glycogen is broken down to lactate, which is easily shuttled out of the astrocyte via the monocarboxylate transporter MCT1 and into the neuron–axon via the monocarboxylate transporter MCT2, where it is converted to pyruvate and incorporated into the citric acid cycle to yield ATP (see Fig. 3). Neurons in gray matter and axons in white matter are sustained by lactate as well as by glucose.

C. Lineage

There appear to be two separate pathways that result in the formation of astrocytes. The first pathway suggests that astrocytes are produced by immature cells of the ventricular zone but must temporarily pass via a radial glial cell intermediary. The other pathway proposes that astrocytes are direct descendants of immature cells in the germinal zones and do not pass via a radial cell intermediary stage. The first pathway results in response to epidermal growth factors and is separate from oligodendrocyte lineage. The second pathway shares a common progenitor with oligodendrocytes. The nature of astrocyte maturation seems to indicate that two separate pathways can produce the same cell, and although there are regional differences between astrocyte morphology, it is not known whether this is as a result of the maturation phase.

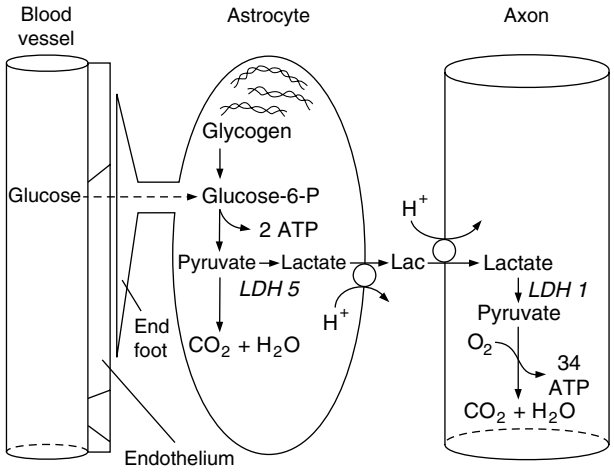


Figure 3 Schematic illustration of how astrocytic glycogen appears to fuel axons in the absence of glucose. Blood glucose first encounters astrocytic end feet as it is transported into the brain. In the absence of glucose, astrocytic glycogen is broken down into lactate, which is transported to the extracellular space via a monocarboxylate transporter (MCT). It is then taken up by a MCT in axons and is oxidatively metabolized to produce the energy needed to sustain excitability. LDH5 preferentially reduces pyruvate to lactate, whereas LDH1 preferentially oxidizes lactate to pyruvate. This scheme recognizes that astrocytes can subsist, at least transiently, on glycolytic energy metabolism, whereas axons require oxidative metabolism.

D. Physiology

The membrane of astrocytes contains a variety of neurotransmitter transporters, ion channels, and gap junctions. The neurotransmitter transporters act to clear the ECS of neurotransmitters, which accumulate during electrical activity. The continued presence of neurotransmitters in the ECS could lead to cell death by increased intracellular Ca^{2+}, e.g., glutamate. Alternatively, the sustained presence of GABA may lead to prolonged inhibition of cells. Astrocytes also contain a variety of ion channels. Although Na^+ channels are present, they are not present in sufficient density relative to K^+ channels to cause action potentials. Multiple K^+ channels are present and because astrocytes are almost exclusively permeable to K^+ at rest, they act as K^+ electrodes and the resting membrane potential is dictated almost exclusively by the transmembrane concentration of K^+ ions. As such, astrocytes are exquisitely sensitive to changes in extracellular K^+ concentration, and changes in this concentration due to increased neuronal activity are immediately sensed by astrocytes. Astrocytes also contain Ca^{2+} channels, but there has been controversy about whether Ca^{2+} channels are present *in vivo*. In cultured astrocytes and in astrocytes studied after having undergone trauma, Ca^{2+} channels have been shown to be present. However, one study could find no evidence of Ca^{2+} channels present *in vitro*, suggesting instead that increased intracellular Ca^{2+} occurred due to the activation of metabotropic glutamate receptors. These data suggested that the presence of Ca^{2+} channels in previous studies was due to up-regulation of Ca^{2+} channel expression induced by culture conditions or injury. However, technical limitations have made answering this question difficult. Patch clamp recordings suffer from the effects of space clamp induced by the gap junctional dissipation of current. However, a study using electrophysiological techniques and antibodies directed at Ca^{2+} channel subunits indicated that nifedipine-sensitive L-type Ca^{2+} channels are present on both astrocytes and axons in the adult rat optic nerve.

IV. SCHWANN CELLS

A. Anatomy

Schwann cells surround all axons in the peripheral nervous system either by forming channels that are contiguous with the axonal membrane or by wrapping myelin around the nerves. Peripheral nerves are composed of an outer epineurium, which is composed of blood vessels, connective tissue, and fibroblasts, a perineurium, which is composed of flattened cells surrounding the nerves, and the intrafascicular endoneurium, which contains supportive cells and the ECM that surrounds bundles of nerve fibers. The supporting cells include the Schwann cells that ensheath the axons, fibroblasts, and vascular elements. The perineurial cells form a diffusion barrier that isolates the endoneurium from the connective tissue of the surrounding epineurium.

B. Function

The function of Schwann cells is to provide myelin and a suitable environment for peripheral axons. Schwann cells can be divided into two classes: the ensheathing Schwann cell and the myelinating Schwann cell. It is the axons that determine whether a Schwann cell will become ensheathing or myelinating. This was shown by studies in which unmyelinated nerves were incubated with ensheathing Schwann cells, which subsequently differentiated into myelinating Schwann cells. Both types of Schwann cell originate from a common precursor, the neural crest cell, and they can be identified by expression of the calcium-binding protein S-100. The majority of ensheathing Schwann cells are associated with unmyelinated axons of DRG neurons and the myelinated efferent sympathetic axons of autonomic ganglia. There is a basal lamina surrounding each axon–Schwann cell unit, which provides both the structure and flexibility required for axonal function. In addition to this function, Schwann cells also form a myelin sheath around selected axons. This process occurs by separation of the axon from its neighbors, and a one-to-one relationship between axon and Schwann cell occurs. The Schwann cell envelops the axons during myelination, forming a very tight wrapping around the axon. The diameter of the myelin sheath is proportional to the diameter of the axon it myelinates, resulting in the optimal conduction velocity of the axon. The length of myelin provided by one Schwann cell is between 250 and 1000 μm and is separated from the adjacent segment by the node of Ranvier. The nodes are interspersed to give optimal conduction velocity. The node has a high density of Na^+ channels, which produces the upstroke of the action potential. The repolarizing K^+ channels are spread more evenly along the axons and can be found underneath the myelin. Schwann cells have a small cell body located at the midpoint of the length of the cell.

1. Regulation of Nerve Development

Schwann cells have been shown to regulate nerve development. In Erbb3– mice the survival of neuronal populations, including DRG neurons, depends on trophic factors released from Schwann cell progenitors and immature Schwann cells. Candidates for this function are glial cell line derived neurotrophic factor (GDNF) and a related compound, neurturin. GDNF mRNA is found in Schwann cells and it is essential for motorneuron survival. Neurotroph 3 (NT3) is a possible candidate. It is formed by Schwann cells at birth and promotes survival and differentiation of early sensory neurons. Thus, NT3 seems to support DRG neurons. Additionally, DHH may help form fibroblasts, and Schwann cell survival and MAC may support axonal architecture.

2. Repair of Axons

Schwann cells have been implicated in the repair of axons after injury. One of the major differences between peripheral and central axons is that the peripheral nerves will regenerate to a large degree whereas central axons seem unable to do so. This accounts for the devastating injury seen in patients whose central axons have been damaged, e.g., spinal cord trauma and multiple sclerosis. In the peripheral nervous system, cut axons can sprout and these sprouts elongate and reconnect with the correct targets. In the CNS, axons sprout but cannot advance to their original targets; thus, regeneration fails. In the peripheral nervous system the neurotrophins NGF, NT3, and BDNF show beneficial effects on the survival of axons, and these are also released from Schwann cells. More recent studies have indicated that peripheral nerves can be used in the regeneration of central axons. In the spinal cord, injection of Schwann cells into the site of injury is a promising strategy to aid regeneration.

C. Lineage

The majority of Schwann cells in the peripheral nervous system originate from neural crest cells, as do satellite cells, the other main type of glial cell in the PNS. The neural crest comprises a multipotent population of cells that arises from dorsal areas of the neural tube during development. The cells disperse through distant pathways in the mesenchyme and differentiate to form a variety of cell types. It has been shown that the progenitor neural crest cells can give rise to more than one cell type and that at some point in their differentiation the cells "decide" which way to differentiate. In order to understand this process of lineage choice, it is important to understand the effect of extrinsic signals on lineage outcome. In order for developing Schwann cells to survive, they must associate themselves with a developing peripheral nerve. This is likely to happen in the anterior parts of neural tube where both neural crest cells and peripheral axons are directed, and therefore they end up in the same place at the same time. These immature cells invade and ensheath bundles of developing axons, a process called radial sorting. They then differentiate into myelinating and nonmyelinating Schwann cells. It is not known whether the ultimate lineage of the cells is determined by the axons or whether the crest cells' lineage has already been decided prior to the meeting of cells and axons. There is evidence that the neural crest cell lineage is predetermined prior to meeting with the axons. Neural crest cells express the gene for the peripheral myelin protein P_o in small quantities prior to meeting the axon. However, once the cone is in contact with the axons, the expression of P_o is unregulated. Thus, expression of P_o may indicate cells that have entered Schwann cell lineage.

When do Schwann cell precursors give rise to Schwann cells? In rat sciatic nerve the phenotype of cells is determined between day E14 and birth. Cells from days E14 and E15 do not survive more than 20 hr in routine culture medium when removed from the nerve, but cells removed from nerves E17 and older survive. This suggests that at E14–E15 the nerve contains precursors, whereas at day E17 immature Schwann cells are present. To support this theory, nerves at day E16 contain both cell types.

Thus, it appears that Schwann cell development occurs by the following process. Neural crest cells form two main intermediates: the Schwann cell precursor found in rat nerves aged E14–E15 and the immature Schwann cells, which are present from E17 to birth. At birth the cells start to differentiate and form mature Schwann cells, of which there are two types: myelinating and nonmyelinating. It appears that the embryonic phase of development is controlled by a single signal, NRG-b. This substance is found in the membrane of axons, and it is likely that survival of Schwann cells is controlled by release of this substance from axons associated with the Schwann cells. High concentrations of NRG mRNA are found in the motor neurons of the ventral horn of the spinal cord and in DRG neurons, the two major sources of axons in the embryonic peripheral nerves.

Schwann cells, however, also provide support for peripheral neurons. Mice deficient in the NGR receptor ErbB3 have about 80% cell death by E18. This indicates mutual dependence between developing neurons and Schwann cells, in which the survival of the Schwann cell depends on NGF released from the axon and neuronal survival depends on an as yet unidentified precursor-derived signal.

D. Physiology

Until the advent of the patch clamp technique in the early 1980s, Schwann cells were thought to be electrically passive. This advance in technology revealed the possible presence of voltage-gated ion channels on Schwann cells. It has since been shown that Schwann cells from a variety of preparations express a wide range of ion channels, including Na^+ channels, K^+ channels, Cl^- channels, Ca^{2+}-dependent K^+ channels, and Ca^{2+} channels. The role of these channels in Schwann cells remains unclear, as the conventional role for Na^+ and K^+ channels, i.e., action potential conduction, obviously does not occur. It is thought that the role of the channels is more likely to be the maintenance of normal function of both the Schwann cell and the extracellular ionic milieu. K^+ channels may play a role in cell proliferation, as inhibition of K^+ current leads to a decrease in Schwann cell proliferation. Specifically, it is the TEA, 4-AP-sensitive, type 1 K^+ channel that appears to be involved in proliferation, as inhibition of the type 2 K^+ channel with dendrotoxin had no effect on proliferation. Ca^{2+} channels have been located on Schwann cells but their function remains unknown. To date it appears that only L- and T-types have been located in Schwann cells, but it is difficult to categorize these channels types without the use of antibodies directed at specific α_1-subunits. It appears, however, that Ca^{2+} channels are unregulated by neighboring DRG cells. Neuronal

activity in DRG cells results in the up-regulation of Ca^{2+} channels in cultured Schwann cells; presumably the DRG neurons release a diffusible factor that acts on Schwann cells.

Schwann cells have been implicated in Charcot–Marie–Tooth disease, a common inherited heterogeneous group of peripheral neuropathies. This disease is usually inherited as an autosomal dominant disorder, although recessively inherited forms do occur less commonly. Charcot–Marie–Tooth (CMT) disease is associated with severe demyelination of peripheral nerves, which results in greatly slowed conduction velocities in the nerves and results in muscle dysfunction and atrophy. It is a slowly progressive disease, and patients may ultimately become unable to walk. The most frequent form of the disease CMT1, designated CMT1A, is caused by abnormalities of one of several genes expressed in Schwann cells. The majority of cases have been shown to be associated with duplication in the p11–p12 region of chromosome 17, which contains the peripheral myelin protein 22 (PMP22) gene, which encodes one of the major PNS myelination proteins. A less common form of CMT1 (CMT1B) is caused by mutation in the protein zero gene (P0), which encodes the major PNS myelin structural protein. Point mutation in the early growth response gene 2 (EGR-2) of Knox 20, which is expressed in Schwann cells, has been implicated in the autosomal dominant form of CMT1. An X-linked form of CMT (CMTX) is caused by mutations in the connexin 32 gene, which is expressed in myelinating Schwann cells. CMT leads to axonal degeneration, which results in the clinical phenotype of the disease. This is analogous to the axonal dysfunction seen in the CNS in multiple sclerosis pathology.

V. OLIGODENDROCYTES

A. Anatomy

Oligodendrocytes are the cells in the CNS that manufacture and maintain myelin, and as such they play an analogous role to the Schwann cells in the PNS. Hortega characterized oligodendrocytes in the 1920s using silver carbonate impregnation techniques. Not only did Hortega offer the first detailed morphological description of these cells, he also implied their role in myelination. Whereas oligodendrocytes are classified as a single cellular entity, they display a great degree of polymorphism. They are distributed throughout the entire CNS but are most prominent in white matter areas. They tend to be lineated in one of three ways: (1) aligned in rows along nerve fascicles, (2) juxtaposed against neuronal somata, and (3) abutting blood vessels. On the basis of these lineations, Hortega classified oligodendrocytes as interfascicular, perineuronal, and perivascular. In addition to categorizing oligodendrocytes on the basis of lineation, Hortega classified oligodendrocytes into four groups on the basis of morphology. Type 1 oligodendrocytes have spherical somata from which numerous processes project toward nerve fibers. Type 2 oligodendrocytes are located exclusively in white matter areas and have a cuboid cell body shape, with fewer and thicker processes associated with nerve fibers. Type 3 oligodendrocytes have only three or four processes emerging from the somata extending toward nerve fibers. Type 4 oligodendrocytes occur near the entrance of nerves into the CNS and adhere directly to nerve fibers. Improved staining technology using intracellularly injected dyes has increased morphological resolution and provides a more detailed picture of oligodendrocytes. In addition to improved intracellular staining techniques, specific markers for oligodendrocytes have been developed, of which antimyelin basic protein is one of the most commonly used.

B. Function

The function of oligodendrocytes is to ensheath axons with myelin and to both support myelin production and ensure that it myelinates axons in the optimal fashion. Myelin is essential for action potential conduction and acts to increase the speed of action potential firing. Immature axons are premyelinated and separated into groups by primitive sheetlike glial processes from immature undifferentiated glial cells. This gives way to more organized wrapping of the axons with myelin by the oligodendrocytes. The myelin sheet extends from oligodendrocytes and is connected to the cell body via only a very thin process. During the process of myelination, the oligodendrocyte cell body remains stationary and the myelin sheet wraps around the axon. The myelin sheet is assembled at the local level as it sits close to the axon during the ensheathing process. There are several points to bear in mind about myelination. Oligodendrocytes myelinate more than one axon and can myelinate axons of different diameters. Despite a common cell body, myelin sheaths that myelinate large-diameter axons

are thicker than myelin sheaths myelinating small-diameter axons. At the time of myelination it appears that axons secrete molecules at the site of myelination that control the behavior of myelinating oligodendrocytes. The process is not fully understood, but it appears that there is an increase in the intramembranous particles (IMPs) on the protoplasmic side of the axon membrane. This occurs during the initiation of myelination and has been hypothesized to be the promoter of myelination.

Multiple sclerosis is a condition that occurs due to demyelination of central axons. The white matter lesions that define this disease are visible by T2-weighted MRI. The condition is due to an autoimmune inflammatory response that results in focal lesions in the CNS. The condition starts with an increase in the permeability of the blood–brain barrier in association with inflammation. The area of abnormality increases over a period of several weeks but then diminishes, leaving a scar that may be gliotic or may exhibit axon degeneration. Function decreases in tandem with the increasing lesion size, but then returns completely as the lesion diminishes. However, over time the attacks recur and function diminishes. Lesion sites have been shown to contain expanded extracellular space due to the loss of axons. This loss of axons results in dysfunction that can lead to limb paralysis and the eventual inability to walk. The potential for treating multiple sclerosis is determined by two main factors: remyelination of axons and replacement of dead axons. The first area is a subject of intense research, with stem cell transplants appearing to offer the best chance of success. Injection of immature oligodendrocytes into rat brain has indicated that they are capable of remyelination, but direction of this process remains a problem. Axon death can be countered by devising neuroprotective strategies for axons during acute attacks. In this it is similar to neuroprotective strategies during ischemic stroke. The advantage in the case of multiple sclerosis is that the attacks when axons die are clinically apparent, and thus therapy could be timed to coincide with attacks.

C. Lineage

Oligodendrocytes develop from migratory, proliferating progenitor cells called O-2A progenitors. These cells develop from neuroepithelial cells in the walls of the embryonic neural tube. Two markers are used to identify oligodendrocyte precursor cells: the chondroitin sulfate proteoglycan NG2 and platelet-derived growth factor α receptor (PDGFαR). PDGFαR-positive cells show a distribution similar to that of NG2+ cells. NG2+ cells are first detected in the developing ventral spinal cord at E14 and are found in the hindbrain and within the basal forebrain by E16. At birth NG2 cells are distributed widely in the CNS, indicating a rapid increase in cell distribution. From E17 to birth, all cells positive for NG2 also express PDGFαR, strongly suggesting that NG2+ cells are markers for oligodendrial precursors. *In vivo* NG2+ cells form oligodendrocytes, but in tissue culture these cells can also form astrocytes as well as oligodendrocytes. However, there is no evidence that this occurs *in vivo* and it may be an artifact of the culture conditions. It has also been shown that NG2+ cells persist in the mature CNS, indicating that not all of the NG2+ cells form oligodendrocytes but may have some additional function. However, the question remains whether these NG2+ cells are a separate cellular entity than oligodendrocytes. NG2+ cells in the mature CNS have a complex morphology, which is different from that of oligodendrocytes. They have highly irregular cell bodies from which fine branches radiate. In gray matter NG2+ cells lie closely apposed to neuronal cell bodies, whereas in white matter NG2+ cells are tightly packed between myelinated axons and have elongated cell bodies. These NG2+ cells are slowly dividing and have been implicated as having a role to play during remyelination. There are increased numbers of NG2+ cells at lesion sites, and there is limited evidence that these cells form new remyelinating oligodendrocytes. There is also evidence that separate progenitor cells may give rise to oligodendrocytes. These cells are characterized by the expression of plp/dm-20, and during development evidence exists that plp/dm-20 expressing cells differentiate into oligodendrocytes and that these plp/dm-20 precursors do not depend on PDGFαR to proliferate, as they do not appear to be precursors of PDGFαR progenitors.

D. Physiology

Despite having a common lineage with astrocytes, oligodendrocytes exhibit different ion channel distribution, most likely due to the different functional properties of each cell type. The mature oligodendrocyte expresses an inwardly rectifying K^+ channel, which constitutes the predominant ion channel in differentiated oligodendrocytes *in vitro*. This channel

has a conductance of 29 pS. Mature oligodendrocytes never express voltage-activated Na^+ channels, which are lost at an early stage of maturation. However, mature oligodendrocytes do express a tetrodotoxin-sensitive, voltage-independent Na^+ channel that is activated by acidification. Stepping of pH from 7.9 to 6.7 induces a large, slowly inactivating current that is potentiated by reducing ambient $[Ca^{2+}]_o$. A functional role for this channel has not been established.

VI. MICROGLIA

A. Anatomy

Microglia can be present in three states: resting, activated, and phagocytotic. The morphology of microglia differs from that of other glial cells in that it changes in response to the CNS perturbations. Thus, under normal conditions microglia are in the resting state, but they will become activated or phagocytotic in response to CNS injury and will invade the CNS. These changes in morphology are intimately linked with microglial function, as will be discussed later. The first selective stain for microglia was silver carbonate developed by Rio Hortega and, although unreliable, was the standard method for 50 years. The most reliable current histochemical staining method for identifying resting microglia uses the enzymes thiamine pyrophosphatase and nucleoside diphosphatase. Activated microglia can be distinguished from resting microglia in that they exhibit increased activity of enzymes that are absent from resting microglial cells. These enzymes include acid phosphatase, 5′-nucleotidase, oxidoreductase, nitric oxide synthase, lysosomal proteinases, plasminogen activator, lysozyme, purine nucleoside phosphorlyase, and elastase. However, these enzymes are not specific markers for activated microglia because they are found in other glial cells. Resting microglia can be distinguished both by their unique morphology and by their phenotype. Morphologically they have highly branched processes, a small amount of perinuclear cytoplasm, and a small, dense, and heterochromatic nucleus, and they can be positively distinguished as the only cells that express the CR3 complement receptor. At the ultrastructural level they are recognizable as true parenchymal constituents of the nervous system, i.e., they are located outside of the vascular basement membrane. Microglia are not a homogeneous cell population throughout the brain because their morphology varies with brain area. Specifically, in gray matter areas of the brain, resting microglia tend to be profusely ramified with processes extending in all directions, whereas resting microglia in white matter align to their cytoplasmic extensions in parallel but also at right angles to nerve fiber bundles.

B. Function

Microglia are the brain macrophages that respond to a variety of CNS injuries. There are three states of microglia: (1) resting, which are highly ramified cells, (2) activated microglia, which are cells responding to injury with morphological (enlarged cell bodies, contraction of processes) and immunophenotypic changes as well as proliferation, and (3) phagocytic microglia, which are full-blown brain macrophages with amoeboid morphology and expression of a number of immunomolecules. Activated microglia retract their processes followed by a rounding of the cell body. This is accompanied by increased expression of complement receptor 3. MHC class I and II antigens are upregulated on the microglial cell surface, which enables the microglia to interact with immunocompetent cells such as T-cells. Once activated, phagocytic microglia are able to remove debris. The phagocytic microglia invade the area of damage and appear over a period of days, depending on the nature, severity, and location of the CNS damage. However, the process of debris removal can take from days to many weeks. This is in contrast to damage in the PNS, where macrophages have cleared debris within 2 weeks. As well as having a role in acute CNS injury, microglia may also play a role in neurodegenerative diseases such as Alzheimer's disease and Parkinson's disease. These conditions are characterized by the selective loss of neurons in distinct areas of the brain, areas in which microglia are activated. In Alzheimer plaques containing amyloid-β protein, microglia are present in the center of these plaques. It has been shown that amyloid-β protein precursor can activate microglia, which may act to enhance their toxicity.

Microglia are also involved in CNS autoimmunity and are activated by inflammatory signals such as LPS and interferon. Microglia are involved in HIV encephalitis, and although it is not clear how HIV interacts with microglia, it is thought that the HIV virus enters the CNS parenchyma hidden in infected monocytes and may subsequently spread to microglia.

C. Lineage

The origin of microglial precursors is a controversial issue with two camps claiming distinct origins.

1. Microglial cells are of mesodermal origin.
2. Microglial cells originate from neuroepithelial cells.

The first view is supported by the majority of researchers, who believe that microglia derive either from monocytes that leave the blood and colonize the nervous parenchyma or from primitive hemipoietic cells (stem cells) that differentiate into microglia within the CNS. The evidence for each group can be summarized as follows. Evidence that microglial cells are of mesodermal origin is based on the fact that microglial cells and cells of monocytic lineage share common features, including the enzymes nucleoside diphosphatase, nonspecific esterase, and acid phosphatase. Additionally, both cell types contain vault particles and are labeled by several lectins. Antibodies have been developed that recognize both cell types in a variety of species. Ink- or carbon-labeled monocytes injected into the blood stream of newborn rats are incorporated into ramified microglia in adult animals, suggesting that microglia originate from monocytes that enter the nervous parenchyma via the blood stream. These findings together with the similarity of function of microglia and monocytic cells suggest a common origin. Less convincing evidence is that cells with morphological properties and patterns of membrane currents similar to those of microglia are derived from monocytes.

However, some authors maintain that at least some microglial cells are of neuroectodermal origin. Autoradiographic studies indicate that microglia are derived from glioblasts that also produce astrocytes, indicating that proliferating glioblasts form microglia. Microglial cells are found within the matrix cell layer during development, but these cells may be merely crossing the neuroepithelial layer after entering the nervous parenchyma and may not be derived from that area. Antibodies labeled against the protein lipocortin-1 label both microglia and neuroepithelial cells, and some of the antibodies that recognize microglial cells also label cells of neuroectodermal origin. Additionally, microglial cells can be produced in mouse neuroepithelial cell cultures thought to be devoid of mesenchymal precursors.

More recent data have shown that both microglia and astrocytes can derive from the same progenitor cell, suggesting either that some microglia are of neuroectodermal origin or that some astrocytes can be derived from mesenchymal cells, challenging the long-held view that astrocytes are of purely neuroectodermal origin. As such, the derivation of microglial cells has not been conclusively shown, but it is fair to say that

1. Microglial cells derive either directly from blood cells or are derived from cells of blood cell lineage.
2. During development microglial precursors invade the CNS.
3. Microglial precursor cells migrate within the CNS parenchyma to their final location.
4. Microglial cells resemble amoeboid microglia during migration but differentiate and become mature, ramified microglia once they reach their final location.

D. Physiology

In order for microglial cells to perform their task of responding to changes in CNS microenvironment, they must be able to sense changes in the extracellular microenvironment. This function can be achieved through the activation of ion channels. To date microglial cells have been shown to possess six types of K^+ channel, H^+ channel, Na^+ channels, Ca^{2+} channels, Ca^{2+}-release-activated Ca^{2+} channels, and voltage-dependent and voltage-independent Cl^- channels. However, the level of channel expression depends upon the state of the microglia, with activated microglia exhibiting different levels of channel expression than resting microglia. The inward rectifier K^+ channels are inhibited by G-protein activation and appear to be involved in regulating membrane potential, but so far no role has been suggested for delayed rectifier channels, although they may be involved in membrane repolarization and cell volume regulation . Both inward and outward rectifier channels are inhibited by increased intracellular Ca^{2+}, whereas, as expected, an increase in Ca^{2+} to 1 μM activates the Ca^{2+}-activated K^+ channel. Voltage-gated H^+ channels are expressed in phagocytic microglia. Their function is thought to be H^+ extrusion after phagocytic production of H^+. Increases in intracellular H^+ (decreased pH) and cell swelling activate this current, suggesting that it acts as a negative feedback mechanism to protect microglia from cytotoxic intracellular acidification and acidosis-induced swelling.

Na^+ channels are expressed in ramified microglia and have been proposed to play a role in membrane depolarization and regulation of morphological changes, although there is no firm evidence for this. L-type Ca^{2+} channels are present on microglia and are thought to be the pathway of increased intracellular Ca^{2+} by PrP and β-amyloid fragments. Capacitative Ca^{2+} currents were induced in rat microglia being perfused with IP3. However, although their existence has been shown, too little is known about them to compare them to known CRAC channels. However, it has been proposed that they play a part in microglial superoxide production. It is thought that functional Cl^- channels are required for changes in membrane potential and for the induction of microglial ramification, but not for maintenance of the shape. Microglia also contain a host of surface receptors, including AMPA receptors and thrombin receptors, although future research will no doubt uncover more receptor expression on microglia.

VII. ENTERIC GLIA

A. Anatomy

Enteric glia are different from other glial cells in that they are not located in the CNS but rather are located in the gut as part of the enteric nervous system (ENS). As such they are obviously distinct from glial cells in the central and peripheral nervous systems, but they do share common features that justify their inclusion in the family of neuroglia. The enteric nervous system is functionally different from the peripheral nervous system in that it is capable of mediating reflex activity without the input of the PNS or CNS. This has given rise to the concept that the gut is controlled by the activity of local isolated microcircuits located within the ENS. The structure of the ENS can be considered to resemble the CNS in certain respects. Nerve cell bodies and axons in the ENS are not surrounded by connective tissue as occurs in the PNS. This resemblance to the CNS suggests that support for ENS neurons comes from glial-like cells—the enteric glia. Morphologically, enteric glia differ from Schwann cells in that enteric glial cells are more irregularly shaped. In the myenteric plexus, the cell body projects long processes that end in small boutons that abut the basal lamina surrounding ganglia and interganglionic connectives. This enteric glial sheath is not complete and does not totally separate the myenteric plexus neurons from the extraganglionic connective tissue, thus allowing neurons to contact the periganglionic basal lamina directly.

B. Function

Enteric glial cells establish the perimeter of the myenteric plexus and partition myenteric ganglia into compartments. The surfaces of enteric glia are devoid of laminin and are only in contact with laminin where their end feet are in contact with the basal lamina. The enteric glial cells are packed with 10-nm intermediate filaments, and as such enteric glia are rich in glial fibrillary acidic protein (GFAP). GFAP-positive cells cannot be identified exclusively as enteric glial cells because PNS Schwann cells also express GFAP.

C. Lineage

As is the case with microglial cells, there are two possible origins for enteric glial cells; however, each theory has enteric glia originating in the neural crest. The first theory is that enteric glial cells migrate to the bowel within the original population of vagal and sacral crest-derived cells that also form enteric neurons. The second theory is that enteric glial cells enter the gut later in development. In order to investigate the first theory, portions of fetal gut were isolated before innervation by extrinsic nerves occurred. When these portions were grown *in vitro*, it was found that GFAP-positive cells developed, indicating that enteric glial precursors are present in the gut at the time when the explants were removed. Thus, cells capable of producing glial cells are present in the gut before extrinsic innervation arrives and proves that glial cells can be derived from the original wave of neural crest cells that colonize the gut. A similar type of study was carried out in which explants of the cells of presumptive ganglion cells were removed from ls/ls mice. In these mice, the terminal colon is aganglionic because it is not infiltrated by enteric glial precursors from the neural crest. However, this region is not denervated as extrinsic nerves grow into the region from the ganglia in more proximal regions. These nerves exhibit GFAP immunoreactivity. The mutation in ls/ls mice is due to the inability of the tissue to support inward migration of neural crest-derived neuroglial cells. Explants of ls/ls bowel give rise to cultures that contain neither neurons nor glia as these segments of gut are devoid of precursor cells. It appears that the supporting cells in

this area are derived from Schwann cells. Thus, glial components may be derived from separate lineages of cells, of which one is the crest-derived precursors that colonize the gut early and the other is the Schwann cell population that reaches the bowel later in development.

D. Physiology

Enteric glial cells possess a variety of surface receptors that can be used to distinguish them from nonmyelinating Schwann cells, with which they can be easily confused. These include reacting to the marker RAN-1, although it is only seen in cultured enteric glial cells, not *in situ*. The enzyme plasmalemmal 5′-nucleotidease is found on the surface of enteric glial cells but not on Schwann cells.

See Also the Following Articles

ASTROCYTES • MICROGLIA • NEURON

Suggested Reading

Dawson, M. R., Levine, J. M., and Reynolds, R. (2000). NG2-expressing cells in the central nervous system: Are they oligodendroglial progenitors? *J. Neurosci. Res.* **61,** 471–479.

Eder, C. (1998). Ion channels in microglia (brain macrophages). *Am. J. Physiol.* **275,** C327–342.

Fedoroff, S., Zhai, R., and Novak, J. P. (1997). Microglia and astroglia have a common progenitor cell. *J. Neurosci. Res.* **50,** 477–486.

Gershon, M. D. (1998). Genes, lineages, and tissue interactions in the development of the enteric nervous system. *Am. J. Physiol.* **275,** G869–G873.

Gershon, M. D., and Rothman, T. P. (1991). Enteric glia. *Glia* **4,** 195–204.

Hansen, A. J. (1985). Effect of anoxia on ion distribution in the brain. *Physiol. Rev.* **65,** 101–148.

Izumi, Y., Benz, A. M., Katsuki, H., and Zorumski, C. F. (1997). Endogenous monocarboxylates sustain hippocampal synaptic function and morphological integrity during energy deprivation. *J. Neurosci.* **17,** 9448–9457.

Kamholz, J., Menichella, D., Jani, A., Garbern, J., Lewis, R. A., Krajewski, K. M., Lilien, J., Scherer, S. S., and Shy, M. E. (2000). Charcot–Marie–Tooth disease type 1: Molecular pathogenesis to gene therapy. *Brain* **123,** 222–233.

Mirsky, R., and Jessen, K.R. (1999). The neurobiology of Schwann cells. *Brain Pathol.* **9,** 293–311.

Raisman, G. (1997). Use of Schwann cells to induce repair of adult CNS tracts. *Rev. Neurol. (Paris)* **153,** 521–525.

Rothman, T. P., Tennyson, V. M., and Gershon, M. D. (1986). Colonization of the bowel by the precursors of enteric glia: studies of normal and congenitally aganglionic mutant mice. *J. Comp. Neurol.* **252,** 493–506.

Sontheimer, H., Trotter, J., Schachner, M., and Kettenmann, H. (1989). Channel expression correlates with differentiation stage during the development of oligodendrocytes from their precursor cells in culture. *Neuron* **2,** 1135–1145.

Stoll, G., and Jander, S. (1999). The role of microglia and macrophages in the pathophysiology of the CNS. *Prog. Neurobiol.* **58,** 233–247.

Wiesinger, H., Hamprecht, B., and Dringen, R. (1997). Metabolic pathways for glucose in astrocytes. *Glia* **21,** 22–34.

Wender, R., Brown, A. M., Fern, R., Swanson, R. A., Farrell, K., and Ransom, B. R. (2000). Astrocytic glycogen influences axon function and survival during glucose deprivation in central white matter. *J. Neurosci.* **20,** 6804–6810.

Neuroimaging

GIORGIO GANIS and STEPHEN M. KOSSLYN
Harvard University

GLOSSARY

brain activation Fast, transient biophysical and biochemical variations in neurons and, to some extent, in the glial cells that support them.

cerebral cortex Continuous sheet of tissue covering each hemisphere of the brain composed of billions of neurons and glial cells.

Gauss Unit of magnetic field strength. The preferred International System unit of field strength is the tesla (1 T = 10,000 G).

gray matter The gray tissue of the nervous system that contains a high proportion of nerve cell bodies.

hemodynamic response Variation of a vascular property, such as regional cerebral blood flow, in response to a change in neural activity.

sampling rate Rate at which physiological signals are sampled for subsequent analysis.

spatial resolution Precision with which physiological variables can be localized in space by a given neuroimaging technique.

voxel "Volume element" in a three-dimensional image.

white matter The white tissue of the nervous system consisting mainly of axons connecting nerve cells.

The ultimate goal of human neuroimaging is to monitor the living brain at all levels of structure and function, from neurotransmitter and receptor molecules to large networks of brain cells, with time scales ranging from milliseconds to minutes. Achievement of this goal would provide psychologists and neuroscientists with biological building blocks roughly similar to those that the human genome project provides geneticists. However, this analogy is misleading because the task of the neuroimager is probably more complicated than that of the geneticist. For instance, whereas the human genome contains an estimated 3×10^9 base pairs within at most approximately 10^5 genes, the human brain contains perhaps 10^{14} synapses on about 10^{11} neurons. In both the brain and genetic realms, it is important to realize that even perfect knowledge of biological building blocks would not automatically lead to knowledge of how the mechanisms function. Although knowledge of the structure of the brain tells us something about how it operates (just as knowledge about the structure of a lens tells one something about why people sometimes wear them in front of their eyes), further research will be required that specifically seeks to characterize how mental processes arise from neural tissue.

I. INTRODUCTION

The classical and most direct method of measuring neuronal activity in individual cells *in vivo* is by means of invasive electrical recordings in animals. In this method, tiny electrodes are placed into or near neurons, and the effects of different activities on the firing rate of the neurons are recorded. Although this method and its variants have provided a wealth of knowledge about how neurons operate, the invasive nature of intracranial electrical recording methods strictly limits their application to only special cases in humans (e.g., patients undergoing brain surgery). A host of noninvasive neuroimaging methods for

measuring brain activity in human beings has evolved. Herein, a brief overview is offered of the major human neuroimaging methods currently in use, emphasizing the strengths and limitations of each. Currently, there are two broad classes of noninvasive neuroimaging techniques to measure brain activity, which are the focus of this article: (1) direct methods that monitor electrical or magnetic fields *directly* linked to neural activity and (2) indirect methods that monitor metabolic and vascular changes associated with neural activity.

Note that in the literature the first category of methods often is not referred to as "neuroimaging." This is misleading not only because images can be produced on the basis of data obtained with electromagnetic measurements but also because research often integrates multiple methods, resulting in combined maps with high spatial and temporal resolution. In addition to neuroimaging techniques to monitor human brain activity, there are also techniques that provide structural information about the brain; these will be addressed only very briefly. Finally, techniques based on brain stimulation, such as transcranial magnetic stimulation, although important for cognitive neuroscientists, will not be discussed here because they are not neuroimaging techniques.

II. ELECTROENCEPHALOGRAPHY (EEG) AND MAGNETOENCEPHALOGRAPHY (MEG)

Noninvasive human electrophysiology was first accomplished in a systematic way with the electroencephalogram (EEG) in 1929 by the German psychiatrist Hans Berger, who recorded the α rhythm (10-Hz oscillations) from two electrodes placed on the front and back of his son's scalp. A variant of EEG is the event-related potential (ERP), in which changes in electrical activity immediately following the presentation of a stimulus or decision are recorded; ERPs are thus "time-locked," with changes being described relative to a particular event (hence the name, "event-related"). The first magnetic recordings of brain activity, using magnetoencephalography (MEG), were produced only much later in 1971.

Methods for recording EEG, ERP, and MEG provide information about the summed electrical events produced by individual brain cells. EEG and ERPs are recorded from electrodes placed on the scalp (although, very rarely, they can be recorded from within the brain in patients being prepared for surgical treatment of epilepsy). MEG and its time-locked derivative, event-related fields (ERFs, which are the magnetic analogs of ERPs), are recorded from arrays of superconducting quantum interference devices (SQUIDs). These detectors are so cold that tiny fluctuations in magnetic fields can induce slight currents, which can then be amplified and recorded.

A. Nature of the Signals

Signals that are recorded outside the head are very small, especially in the case of MEG. Consider that the magnetic field of the earth is about 0.5 G and that of the heart is 10^{-6} G. The α wave, a relatively large signal, is only 10^{-9} G, and most experimental effects are one-tenth of this or less. Indeed, a major expense of setting up a MEG, and to a lesser extent an EEG, system is the construction of an appropriate recording chamber to attenuate external interference.

At a cellular level, both electrical and magnetic activity originates from ionic current flowing across neuronal (and glial) membranes.

For a small brain volume, a single so-called "equivalent dipole" can be defined that approximates the locations where net local current flows either out from the volume (sources) or back into the volume (sinks). In general, an externally observable electrical or magnetic signal is produced if and only if (a) sources and sinks within the brain volume of interest are distributed in a non-radially symmetric manner, (b) the neurons are aligned somewhat systematically (so-called "open fields,") and (c) the neurons are activated synchronously. Thus, fields generated by brain regions in which the neurons have no consistent orientation (e.g., in the basal ganglia or thalamus) or are roughly radially symmetric ("closed fields," which is the case for many brain nuclei) generally contribute little to the EEG and MEG recorded from the scalp.

The gray matter of the neocortex is the main brain structure that meets all three constraints. The neocortex is a large, folded sheet a few millimeters thick composed of 70% pyramidal neurons oriented orthogonally to the cortical surface, with apical dendrites extending from the soma (cell body) to the sheet surface. When apical dendrites are activated near the soma by thalamic inputs, current enters the somatic region and exits the cell at sites near the cortical surface. This creates a source–sink configuration that is approximately dipolar and is oriented perpendicularly to the cortical sheet; a dipolar field of the opposite orientation is produced when apical dendrites near the cortical surface are activated. The sum of

post-synaptic activity in the apical dendrites of the hundreds of thousands of pyramidal neurons within a neocortical patch is considered to be the primary source of scalp-recorded EEG and MEG.

EEG and MEG provide differing but *complementary* views of underlying brain activity, and both therefore should be recorded simultaneously for the most complete picture of neural processing, timing, and localization. The complementary differences between EEG and MEG follow from differences in how electrical and magnetic signals depend on structural factors, such as head shape and the location and orientation of dipoles. Specifically, one difference stems from the fact that EEG recordings detect electrical current conducted through brain tissue and cerebrospinal fluid as differences in potential at the scalp. Electrical currents are greatly impeded and redirected by tissue boundaries (e.g., skull, skin, and cerebrospinal fluid) that differ substantially in conductivity. In contrast, MEG is virtually unaffected by these factors.

Two other differences must be understood within the context of the folded nature of the human neocortex. A crucial difference between EEG and MEG is that dipoles in any orientation (i.e., radial and tangential to the scalp) can modulate electrical potentials, whereas tangential (but not radial) components of the dipole are the principal contributors to the magnetic field outside the head. Another difference is that MEG detects primarily superficial sources, whereas EEG can sense deeper sources as well. This follows from the fact that magnetic field strength dissipates faster than electrical potentials with dipole depth; thus, SQUIDs can detect magnetic fields generated in large assemblies of neurons that are primarily near the outer surface of the brain.

In sum, MEG detects less brain activity than EEG, and all signals observable with MEG are also seen in EEG, but not vice versa. EEG provides a broader view of neocortical activity that includes both superficial and deep sources at many orientations relative to the scalp. In contrast, MEG reveals a more limited view of brain activity (perhaps half the generators in the human brain), being primarily selective to superficial and tangentially oriented dipoles, and, hence, detecting primarily activity originating in the sulci near the surface. Radial sources are essentially invisible to MEG. However, the electrical currents detected by EEG travel both over the cortex and over the scalp, and they are distorted by their passage through the skull (which has varying thickness at different places). In contrast, MEG is affected by none of these factors. Thus, at least in principle, MEG affords better spatial localization than EEG.

At first glance, it is counterintuitive that the lack of sensitivity of MEG to radial sources can be an advantage for the technique. However, given the extreme complexity of the recorded signals, resulting from the superposition of a large number of equivalent dipoles, the ability to rule out a subset of sources turns out to be very helpful. For example, MEG was able to resolve a localization controversy surrounding the early somatosensory potential N20 observed in EEG (the "N" means that it is a negative potential, and the "20" means that it occurs about 20 msec after a stimulus is presented). Because a magnetic N20 was observed, it could be inferred that the signal was generated by a tangential source in somatosensory cortex rather than by thalamic or radial neocortical sources alone because radial sources would have been absent from the MEG.

Although the main strength of EEG and MEG methods is to define the time course of neural information processing with high temporal resolution, it is important to localize the neural sources of EEG and MEG to specify the order in which each brain area contributes to the performance of a particular task. By present methods, in realistic situations EEG or MEG *alone* can often localize single dipolar sources to within 8–10 mm. The complementary nature of EEG and MEG sensitivity suggests that the most accurate localization should be achieved by taking into account both types of information simultaneously.

B. Analysis of EEG and MEG Data

In general, there are two major ways of interpreting EEG and MEG data; this dual approach can be understood to some extent by analogy to the wave and particle views in modern physics. One way is to consider EEG and MEG recordings as composed of discrete, relatively short-lasting neural events ("particles") triggered by specific events (plus "noise" not related to any event of interest). To enhance such neural events, the EEG or MEG can be time-locked to particular experimental events and averaged to yield ERPs or ERFs, respectively. Inferences from ERPs and ERFs are made primarily from their changes over time. In addition, the location of such changes often also provides useful data. The classical approach has been to define "components" in terms of their polarity [ERPs only: positive (P) and negative (N)], latency (milliseconds), scalp distribution, source location, and

function (i.e., modulation with specific experimental manipulations). For example, the P150 (or ERF M180) is a broadly distributed peak that is positive (hence the P) at the vertex around 150 msec. This event has been found to index early perceptual categorization of well-learned visual stimuli (i.e., words and faces), which seems to reflect activity in specific parts of the brain (the posterior fusiform gyrus and occipitotemporal sulcal regions).

A common assumption is that each ERP or ERF peak reflects one or more perceptual or cognitive processes and, therefore, is a neurophysiological index of those processes. To determine whether two experimental conditions engage the same or different processes, researchers evaluate whether both conditions evoke the same ERP or ERF component, and the degree of activation can be inferred from peak amplitude or area. More important, the time course of activation of these neural processes can then be inferred from the peak latency or the time when the conditions diverge.

Another way to analyze EEG and MEG recordings is to consider them as long stretches of data produced by a large ensemble of neuronal oscillators (''waves''). This view has led to analyses of the power (amplitude) of different frequencies of oscillations, which shift over time and when people are in different mental states. The first major success of this approach was to reveal that sleep involves different stages, as indicated by shifts in the power of the oscillations at different frequencies.

Hybrid approaches integrate the particle and wave methods by using ''wavelet filters'' to extract frequency information from the EEG while retaining temporal information. These approaches allow the detection of frequency changes occurring immediately after an event and have been used to discover EEG correlates of specific cognitive processes (e.g., linguistic processes).

The statistical analysis of EEG and ERPs is not trivial because of the large quantity of data acquired in most studies: for a typical experiment samples are recorded at thousands of time points from 100 or more spatial locations. Most investigators compare conditions of interest by performing tests based on the general linear model on data from a subset or all recording sites within various time windows. Similar to other neuroimaging techniques, the probability of finding an effect just by chance (so-called type I error) increases with the number of comparisons. The probability of a type I error is influenced by several factors, including the spatial and temporal correlations in the data. Although a few methods have been proposed to keep the overall probability of a type I error within acceptable limits, currently no universal solution exists.

C. Spatial Localization of EEG and MEG

The problem of determining the spatial location of the EEG and MEG sources, their orientation, and their strength from scalp recordings is often referred to as the ''inverse problem.'' The inverse problem is ill-posed (i.e., there is no unique solution): infinitely many distributions of dipoles are consistent with any set of EEG or MEG measurements. To resolve this ambiguity, additional constraints on the solution must be placed on quantitative models, including those specified by functional neuroanatomical data from other neuroimaging methods. A widely used approach assumes that EEG and MEG signals originate from a few focal, temporally uncorrelated sources and models the generators as some fixed number of discrete equivalent dipoles, where each represents activity within a specific brain volume. Although suitable for early sensory responses that are expected to be focal and recruit few brain areas, these assumptions are less appropriately applied to higher level cognitive activity or seizure-related processes. Cognitive functions often seem to involve widespread networks of somewhat synchronously active regions, and the spikes in partial epilepsy spread quickly across multiple regions.

Alternatively, the inverse problem can be reduced to an (linear) estimation of dipole strength everywhere across the entire folded cortical sheet (i.e., as a continuous dipole distribution) based on assumptions about the neocortical pyramidal cell origin of electromagnetic signals. Because the gray matter is the primary signal source, information about the geometry of the cortical surface, obtained by magnetic resonance imaging (discussed later), can be applied to greatly reduce the solution space of the inverse problem. The number of ''patches'' and corresponding equivalent dipoles that are needed to represent the cortical surface with sufficient precision, however, is much greater than the number of EEG or MEG sensors that can be placed as a practical matter. To circumvent this, a unique solution can be achieved by the further constraint of choosing the ''minimum-norm solution,'' which is the solution with the smallest overall activity (in absolute value); however, additional constraints from methods with high spatial resolution, such as functional magnetic resonance imaging,

are necessary to increase the likelihood that the unique solution is indeed also the correct one. Present methods for combining MEG with functional magnetic resonance imaging and structural magnetic resonance imaging can yield high-resolution spatiotemporal maps of human brain activity that are timed to within milliseconds of execution and localized to within millimeters. The combination of methods is a significant step in the direction of the ultimate goal of neuroimaging, as noted earlier.

D. Summary of Advantages and Disadvantages of EEG and MEG

To summarize, the chief strengths of EEG and MEG are the following. First, they are direct measures of the physiological activity of neural systems. Second, they are noninvasive. Third, electrical and magnetic monitoring of brain activity can reveal changes over very small delays (1 msec or less), which are crucially within the range of interneuronal communication time. EEG has an additional advantage of being one of the most inexpensive methods of neuroimaging; cost is actually a disadvantage of MEG, which is very expensive (at least 10 times more than EEG) in large part because of the cost of acquiring and maintaining the SQUIDs. The main disadvantage of both EEG and MEG is that they cannot indicate the location of activity on the surface of the brain very well (on the order of 10 mm) and are even worse at indicating activity in the neural structures located deep beneath the surface.

We now turn to methods that have properties complementary to EEG and MEG: good spatial resolution, but poor temporal resolution.

III. POSITRON EMISSION TOMOGRAPHY (PET), FUNCTIONAL MAGNETIC RESONANCE IMAGING (FMRI), AND OPTICAL IMAGING

The second broad class of neuroimaging methods is based on the measurement of metabolic and vascular changes that are associated with neural activity; these measures include oxygen and glucose consumption, cerebral blood flow, and cerebral blood oxygenation. The brain requires a continuous supply of glucose via the blood to meet its metabolic requirements; this is because only a little glucose is stored by the brain itself. Neural activity at a particular brain location leads to localized metabolic changes (such as those in glucose and oxygen consumption) and to vascular changes (such as those in regional cerebral blood flow, rCBF). These metabolic and vascular changes are referred to as "secondary changes," the challenge is to infer the nature of neural activity on the basis of such changes. This task is not trivial in part because multiple aspects of neural activity are reflected by single quantities; for instance, both synaptic inhibition and excitation produce similar increases in rCBF.

A. Nature of the Measured Changes

Numerous studies have shown that local glucose consumption and rCBF in the rat brain are highly correlated and that both in turn are highly correlated with underlying neural activity. Studies in humans using positron emission tomography (PET) have also shown close coupling between local glucose consumption and rCBF. Metabolic and vascular changes associated with neural activity appear to arise from synaptic processes rather than from action potentials in the cell bodies. Indeed, most of the increase in metabolism (deoxyglucose uptake) appears to occur not in the cell bodies but in the region occupied by synapses between axons and dendrites. Thus, of the three potential neural sites that contribute to metabolism, soma, axon, and synapses, the latter is the one that is the most highly correlated with the metabolic requirements of neurons. Furthermore, studies have shown that the increase in glucose metabolism is brought about by presynaptic, rather than postsynaptic activity (that is, activity of the "sending" neurons, not the "receiving" ones). This is an important observation because the source of EEG and MEG signals is thought to be primarily postsynaptic.

Work conducted to understand the energy-consuming processes taking place in axon terminals has shown that glucose consumption is correlated with the activity of the Na^+ pump, suggesting that this is the crucial event in the coupling between neural activity and energy consumption. If increased metabolic activity reflects presynaptic activity, then one would predict that both excitatory and inhibitory synaptic activity results in similar increases in energy needs. Consistent with this prediction, stimulation of excitatory and inhibitory afferents of the lateral superior olive within the brain stem auditory system of the cat brings about similar increases in deoxyglucose uptake. Note that, although increased excitatory and inhibitory presynaptic activity may produce identical local increases in deoxyglucose uptake, they will have opposite effects on the postsynaptic neurons being

driven. Thus, rCBF maps reflect a mixture of neurophysiological processes that require energy; such processes include excitatory and inhibitory synaptic activity as well as additional processes needed to regulate the extracellular environment. This observation should remind us that what is generally referred to as "brain activity" in the neuroimaging literature potentially reflects a mixture of many different energy-consuming phenomena.

B. PET

Although the exact mechanism by which synaptic processes cause metabolic and vascular changes is not well-understood, the empirical relationship between these parameters is very reliable. PET is one of the techniques that has exploited this empirical relationship to image neural activity of the human brain *in vivo*. In PET, participants are injected with (or in some cases breathe) a radioactively labeled compound (e.g., one containing ^{11}C, ^{15}O, ^{18}F, or ^{13}N). These compounds reach the brain about 20 sec after administration. When the isotope decays, it emits positrons that travel a short distance (a few millimeters on average, depending on the compound) until they encounter an electron; at this point they annihilate each other, producing two high-energy photons traveling 180° apart along a single line. The PET detectors (rings of "scintillation detectors" arrayed in parallel planes) record the near-simultaneous arrival of photons along a single line and employ the difference in arrival time to compute exactly where along the line the photons were produced. A three-dimensional image is then created on the basis of this information.

The relationship between local neural activity and regional glucose metabolism has been used in PET studies employing ^{18}F-fluorodeoxyglucose (^{18}F-DG), which is metabolized only partially (unlike glucose 6-phosphate), and more of it accumulates intracellularly when the local metabolic rate is greater. By employing ^{18}F-DG, it is possible to visualize brain regions containing neural populations that were metabolically active during the period between tracer injection and detection of the radioactive decay. Because of the long half-life of ^{18}F-DG (about 2 hr; recall that a half-life is the time required for one-half the radiation to decay), this technique has been employed to study phenomena with very slow time courses, such as sleep. To study faster phenomena, such as perception and attention, better temporal resolution is required. $H_2^{15}O$, with its half-life of about 2 min, became the most widely used tracer for these types of studies. The short half-life allows the administration of multiple experimental conditions in the same session. Because $H_2^{15}O$ accumulates locally, with linear increases relative to rCBF, it is possible to measure rCBF (which is tightly coupled with synaptic activity) without the need to measure the time course of radioactivity by arterial blood sampling.

1. Analysis of PET Data

Typical preprocessing stages in the analysis of PET data include the removal of global fluctuations in rCBF (i.e., changes in rCBF that occur over the entire brain), spatial smoothing (i.e., averaging of information from adjacent voxels to reduce the noise and increase the signal), and spatial normalization to correct for brain shape differences between individuals, after which statistical tests are applied, typically on a voxel-by-voxel basis. The simplest method consists of normalizing and subtracting two images acquired during the administration of two conditions and then looking for values significantly different from zero. What counts as "active," thus, depends on the statistical threshold used. Furthermore, the same issues regarding multiple comparisons described for the EEG and MEG techniques apply here. A commonly adopted approach consists of creating a "statistical parametric map" (SPM) from the single-voxel tests and computing corrected probability values for single voxels and clusters of voxels, such that the overall probability of obtaining significant differences by chance alone remains within acceptable limits. The resulting maps indicate the loci where one condition evoked more activation than another; these maps are sometimes overlaid onto high-resolution images of a standard brain obtained with magnetic resonance imaging.

2. Summary of Advantages and Disadvantages of PET

Currently, a major strength of PET for cognitive studies is that it can provide absolute measures of rCBF, which are much more difficult to obtain with other techniques. Such measures are particularly useful in individual differences studies that focus on the correlation between performance and brain activation. Another major strength of PET that is particularly important in the clinical domain is the possibility of mapping the distribution of particular receptors in the brain by using appropriate

"radio-ligands." These radioactive chemicals mimic various neurotransmitters and are lodged in receptors that typically accept such neurotransmitters. Thus, it is possible to map the distribution of receptors by noting where these chemicals come to reside after they are administered. The major disadvantages of PET are that it is very expensive (partly because the radioactive tracers need to be manufactured on the spot), is invasive, and has relatively poor temporal resolution (it requires at least 40 sec to obtain an image).

C. MRI

Advances in magnetic resonance imaging have made it possible to image with high resolution not only the structure of the human brain *in vivo* (including details of the white matter connecting various brain areas by means of diffusion-weighted MRI) but also *functional* changes. For a long time it was believed that local glucose consumption, local oxygen consumption, and rCBF were routinely coupled. The reasoning was that increased neural activity requires more glucose, more glucose requires more oxygen to be utilized, and more glucose and oxygen are delivered by more blood flow. These parameters are indeed strongly correlated *at rest*. However, rCBF and local glucose uptake increase much more than local oxygen consumption during physiological increases in neural activity, suggesting true uncoupling. At first glance this uncoupling between local oxygen consumption and rCBF might seem like a potential problem, but in fact it is another factor that can be used as an index of neural activity. Such uncoupling results in an increase in the concentration of oxygenated hemoglobin (HbO or oxy-Hb) and an apparent decrease in deoxygenated hemoglobin (Hb or deoxy-Hb) in neurally active areas relative to inactive areas. The most popular types of fMRI and optical imaging exploit the magnetic and optical properties of Hb and HbO.

1. Physical Foundations of MRI

Because fMRI is the neuroimaging technique most widely used today, we will describe some of its physical foundations in slightly more detail. According to quantum theory, electrons, protons, and neutrons possess a fundamental property referred to as "spin." We will consider a small group of protons, H^+. The spins of these protons can be thought of as magnetic moment vectors, which cause the protons to behave like tiny magnets with a north (n) and a south (s) pole. When the protons are placed in an external magnetic field with a north (N) and a south (S) pole, their spin vectors align themselves with the external field, just as a magnet would. For each proton there is a low-energy state in which the poles are aligned N-s-n-S and a high-energy state N-n-s-S (opposites attract and like signs repel; thus, more energy is required to keep the s-S and n-N poles aligned than when opposite poles are aligned). At room temperature, the number of spins in the lower energy level, $N+$, slightly outnumbers the number in the upper energy level, $N-$, depending on the molecular makeup of the substance. The difference between $N+$ and $N-$ is referred to as the "net magnetization" of the substance. The net magnetization vector is typically decomposed into two components: the *longitudinal magnetization*, in the direction of the external field, and the *transverse magnetization*, orthogonal to it. At equilibrium, the net magnetization vector lies along the axis of the imposed magnetic field and there is no transverse magnetization.

A crucial idea for magnetic resonance imaging is that the net magnetization vector can be moved by exposing the spin system to energy of a frequency equal to the energy difference between the spin states (e.g., by a radio frequency pulse). If enough energy is delivered to the system, it is possible to make the net magnetization vector orthogonal to that of the external magnetic field. Upon removal of the external energy source, the longitudinal magnetization returns to its equilibrium state, with a time constant labeled T_1. T_1 depends on the particular substance; for example, the T_1 of brain white matter is about 500 mse. The difference in T_1 between different types of tissues (such as gray versus white matter in the brain) is one of the main parameters used to construct structural MRI; these MRI images picture the physical structure of the brain, not its functioning.

While returning to equilibrium, the net magnetization vector will start to rotate about the axis of the external field. This is often depicted as a slow "wobble" or, more technically, *precession*. The precession of the net magnetization vector generates an electromagnetic signal that can be detected by an appropriate receiving coil. The precession occurs at a rate (the Larmor frequency) that depends on the properties of the material and is directly proportional to the strength of the applied field. Thus, if the strength of the applied magnetic field has a linear gradient of strength, then the precession rates at different spatial locations along the gradient will be different in predictable ways: the frequency of the electromagnetic signal received encodes the spatial location of its source.

To summarize so far: the key idea is that the net magnetization vector can be moved from the equilibrium position along the axis of the external field by applying a radio frequency pulse at the proper frequency, thus giving rise to a transverse magnetization component rotating around the axis of the external magnetic field. Such a component is the result of all of the spins precessing synchronously, thanks to the radio frequency pulse.

As soon as the pulse is turned off, the various spins begin to fall out of phase and the MRI signal begins to decay. This decay occurs for two reasons: intrinsic and extrinsic. The intrinsic factor is that the random configuration of the spins itself creates small inhomogeneities in the local magnetic field. The magnetic field is slightly stronger in regions where many spins line up with the field and slightly weaker in regions where many spins are in the opposite direction. These slight changes in the local magnetic field result in slight variations in the precession frequency, causing dephasing. The time constant that describes the return to equilibrium of the transverse magnetization is called T_2. T_2 depends on the molecular environment of the spins (it is about 70 msec for brain white matter). The extrinsic factor for dephasing is the slight inhomogeneities in the external magnetic field. Substances that possess paramagnetic properties (such as Hb) will distort the field, causing faster dephasing (the time constant in this case is referred to as T_2^*).

2. MRI Scanners

MRI scanners exploit the principles just described to create images of the human brain. In practice, an MRI scanner is composed of a magnet, usually a horizontal tube (referred to as the "bore") in which a person is placed, a radio transmitter, and a radio receiver. The radiofrequency pulses described earlier are transmitted by electrical coils located inside the magnet. Often, the same coils also receive the radiofrequency signals generated after the excitation radiofrequency pulse is turned off. The signals generated during this phase induce voltage changes in the receiver coil, which are then sent to suitable amplifiers. The image is reconstructed on the basis of these signals. The intensity of the MRI signal increases with field strength, which explains the trend toward building scanners with higher and higher field strengths: although most existing scanners have a field strength of 1.5 T, the number of 3- and 4-T scanners is rapidly increasing, with a handful of sites in the process of building 7-T scanners.

3. Blood Oxygenation Level-Dependent Contrast (BOLD)

The most popular type of fMRI, blood oxygenation level-dependent (BOLD) contrast, is based on the fact that Hb is paramagnetic whereas HbO is not. As described earlier, rCBF increases brought about by increased synaptic activity are not matched by corresponding increases in oxygen extraction. This causes an apparent decrease in Hb, which results in less rapid dephasing of protons (longer T_2^*). The final outcome is that a local increase in synaptic activity will result in an increased MRI signal. This phenomenon was first demonstrated in humans in 1992.

Although BOLD contrast is the predominant technique for functional mapping, other techniques have been developed. Among them, techniques to measure rCBF noninvasively by using arterial spin labeling (ASL) hold much promise. In essence, ASL techniques "label" arterial blood by changing the magnetic state of arterial water protons as they are carried up through the brain and then following the effect of labeled blood on the amplitude of the MRI signal. The advantages of ASL over BOLD are that, in theory, it should provide better localization of functional activation because of reduced sensitivity to intravenous signals, and that ASL signals are quantitatively related to CBF, which in turn is generally assumed to be tightly coupled with local neural activity. The main disadvantages of ASL relative to BOLD are that the ASL signal is 2–4 times smaller than the BOLD signal, and that ASL is limited in its maximum rate of image acquisition relative to BOLD imaging because of the time required for the tagged blood to flow into the portion of the brain being imaged.

The relationship between the BOLD signal, and the underlying activity is more complex than the well-established one between rCBF and neural activity. Thus, the issue of the extent to which the BOLD signal actually reflects neural events arises. Studies comparing the BOLD signal with rCBF (measured with PET $H_2^{15}O$) have found a linear relationship between PET and fMRI across sets of distributed regions. Animal studies have also provided evidence that BOLD, rCBF, and evoked potentials in response to visual stimuli at various frequencies are highly correlated, which is consistent with the hypothesis that the BOLD signal reflects neural events in a quantitative manner. However, exceptions to this relationship have also been found: in some cases, fMRI seemed more sensitive to activation in a number of cortical areas,

whereas PET seemed more sensitive to activation in deep nuclei.

Even without considering the spatial limitations of the imaging devices, the locus of the metabolic and especially of the vascular responses may not coincide with that of brain cell activity. The theoretical spatial resolution of the BOLD signal is limited by the size of the smallest vascular unit that can be modulated independently in response to neural activity. If synaptic activity only produced changes in the BOLD signal in large blood vessels, then the spatial resolution of BOLD would be very low. In theory, the smallest of such vascular units is the capillary. Based on the fact that capillary walls possess some contractile elements and that there are precapillary sphincters, the capillary recruitment hypothesis states that a major mechanism of blood flow regulation in the brain relies on the complete opening and closing of capillaries. This hypothesis is problematic, however, because under resting conditions all cerebral capillaries seem to be continuously perfused with plasma and most of them also contain moving red blood cells. From these considerations, we can infer that the spatial resolution of the hemodynamic response is at least 1 mm^3, the approximate size of arterioles feeding capillaries. Empirical work, noted shortly, suggests that in fact the spatial resolution of fMRI using BOLD may be a bit better than this.

One method researchers have used to assess empirically the spatial resolution of fMRI, and particularly that of the BOLD signal, has been to map anatomical structures with known functional properties. Activation in the lateral geniculate nucleus of the thalamus (LGN) during photic stimulation has been demonstrated by using BOLD contrast in a 4-T magnet. The LGN is a very small structure and is located near other structures (e.g., the hippocampus) that did not exhibit activation during this task. Although the BOLD signal in the LGN was smaller than that in the primary visual cortex (area V1), it had a similar time course. Furthermore, task-related activation in the pulvinar nucleus of the thalamus was also detected, and its activated location could be discriminated from that of the LGN. More recent work has also shown visual topography in the LGN. Furthermore, a study used high spatial resolution fMRI to map the activation in the ocular dominance columns (less than 1 mm thick) of area V1 during alternate visual stimulation of the left and right eyes. (Note that the success of this study indicates that the spatial resolution is better than 1 mm.)

Animal studies have also provided evidence regarding the spatial resolution of BOLD contrast. For example, stimulation of a rat's whiskers elicits BOLD signals with a spatial distribution that overlaps that of electrical activity in the rat whisker "barrel cortex." These results show empirically that the site of the BOLD activation is spatially very close to the sites of expected neural activity. Despite this convergent evidence, some findings also suggest that caution is warranted in the interpretation of the spatial location of fMRI foci based on BOLD contrast. Other work has compared BOLD signals with rCBF measured by PET during a visually cued sequential finger movement task. The results showed a general similarity in the pattern of activation, but in some cases they also showed a discrepancy between the precise location of foci detected with PET and fMRI (almost 1 cm average), perhaps due to the higher sensitivity of the BOLD contrast to signals coming from draining veins.

We have focused so far on the spatial resolution of fMRI, but the technique is also useful for its temporal resolution—especially in comparison with PET (which requires about 40 sec to obtain an image). The temporal resolution of fMRI depends not only on the sampling rates possible with current MRI hardware but also on the temporal characteristics intrinsic to the hemodynamic response (relative to that of the neural events that generate it). MRI sampling rates depend on the pulse sequences used. In practice, a whole-brain image can be obtained on the order of 1–2 sec with the commonly used gradient echo echoplanar imaging (EPI). Note that the MRI signal becomes smaller and smaller for sampling rates faster than about 3 sec because the magnetic spins do not have enough time to return to equilibrium. A clever method to get around this problem is to vary the onset of the trials relative to the sampling times. For example, if a voxel is sampled every 3 sec, half of the trials can be initiated in synchrony with the excitation pulses whereas the other half can be initiated 1.5 sec later. By combining the two trial types, one can achieve a temporal resolution that is twice the actual sampling rate. The temporal characteristics of the BOLD response have been well-characterized. The onset of detectable differences in the BOLD response, relative to the putative onset of neural activity, is about 2 sec; the BOLD response peaks between 6 and 9 sec and returns to baseline slightly more slowly after the cessation of neural activation. An initial undershoot, reaching a maximum around 1 sec after stimulus onset, has been observed in some studies, although the reliability and nature of this phenomenon are still under investigation. A more reliable post-undershoot in the BOLD

signal, peaking a few seconds after stimulus offset and lasting for about 20 sec, has also been observed.

The effective temporal resolution is affected by the variability of the parameters of the hemodynamic response (such as rise time) rather than by the absolute value of such parameters. For example, if the onset time is very large (e.g., several seconds) but the variance is very small (e.g., 100 msec), one can still determine relative latency shifts between conditions with accuracy. A few studies have addressed the issue of latency estimation variability within a region of interest. These studies measured the variability of a number of parameters such as the onset time and the time to peak of the hemodynamic response within specific regions of interest elicited by a brief sensory stimulus. Typically, the stability of the onset time and the time to peak in a given region are very high, especially within the same subject.

Several studies have shown large differences in the onset latency of the hemodynamic response in different brain regions. Variations in the delay of the BOLD response between 4 and 8 sec in visual and motor cortical areas are common. Such variations are too large to be accounted for by variations in neuronal activity latencies. Even longer delays, between 8 and 14 sec, can be observed in large vessels. Given the large spread of hemodynamic response latencies over space (several seconds), in general it is not possible to infer the temporal order of activation of two arbitrary regions (less than 1 sec for most cognitive tasks) from the absolute hemodynamic onset latencies. It is, however, possible to determine the *relative* timing of neural activation stages within a region of interest in response to two or more experimental manipulations. Although more problematic, it is also possible to determine the relative timing of neural activation stages between two regions of interest if these have comparable hemodynamic properties (e.g., striate cortex representation of the left and right hemifields). A few studies have shown that, within these limits, subsecond temporal resolution can be achieved.

4. Analyses of fMRI Data

A wide spectrum of methods for the analysis of fMRI data is currently in use. Not surprisingly, the range of techniques has expanded over the last few years. It is interesting to note that the analysis methods used most widely depend, to a large extent, on the availability of software packages. Often the software package that is the least difficult to use and the least expensive becomes the one used most widely, to a large extent independent of the soundness of the implemented algorithms. Although some preprocessing and artifact removal issues are specific to fMRI, the general statistical problems are similar to those encountered in the analysis of EEG, MEG, and PET data but are more difficult to solve because of the size of the data sets. A typical fMRI session produces tens of thousands of time series, one for each voxel. Typically $64 \times 64 \times 64 = 262{,}144$ voxels are monitored. Thus, the same kinds of type I error issues described for the previous techniques arise with fMRI. Many analysis techniques take advantage of the fact that the probability of obtaining clusters of active voxels decreases with cluster size (the actual probability depends on the detailed assumptions made).

The first fMRI paradigms were based on the same activation paradigms employed with PET, often referred to as blocked paradigms. These paradigms consist of administering the various conditions for several tens of seconds in a blocked manner. Analyses typically involve generating statistical maps of contrasts between images collected during the various conditions. Developments in the analysis of fMRI data have expanded the range of the paradigms available with fMRI. Such paradigms originated from researchers familiar with the ERP methodology and are thus referred to as event-related paradigms. Event-related paradigms essentially allow the presentation of trials in the various conditions in a mixed manner, thus eliminating several confounds inherent in the blocked paradigms, such as the effect of expectancy. The analytical methods used to analyze event-related paradigms usually rely on the assumption that the hemodynamic response is linear and time-invariant. That is, the methods assume that the response to stimuli A and B presented in close temporal succession can be estimated by summing the responses to stimuli A and B presented independently of each other. Results from several studies are consistent with the assumption that activation-induced hemodynamic responses behave in a roughly linear manner. For example, the hemodynamic response in the visual cortex is approximately linear, although the response to brief stimuli is greater than that predicted by a linear, time-invariant system. Despite such documented deviations from linearity, several studies have shown that useful results can be obtained by using the linearity assumption. For example, responses to visual stimuli delivered as rapidly as 1 per second can be resolved. Furthermore, reliable activation maps can be obtained with stimuli presented at an average rate of 2 per second, provided that the interval between stimuli

varies randomly. The possibility of using rapid presentation paradigms with fMRI is important in part because it has allowed the use of exactly the same paradigms typically employed with electromagnetic techniques, which has promoted the direct comparison of fMRI results with those obtained from EEG and MEG.

5. Summary of Advantages and Disadvantages of fMRI

To summarize, fMRI is the most common form of functional neuroimaging today for good reasons: in addition to its relatively good spatial and temporal resolution, it is also widely available (most hospitals have an MRI machine) and is relatively inexpensive, especially compared to PET. However, the technique is not ideal. First, it is very noisy; the shifting magnetic fields cause physical movements of the magnets, which, like a conventional loudspeaker, displace air and thus produce sound. Second, the hole in the bore is relatively narrow but very deep, which some people find uncomfortable. Third, the technique is very sensitive to motion. Although statistical methods have been designed to correct for the distortions produced by motion, this is still a major problem. Fourth, gradient echo EPI, the method most widely used for fast imaging, does not work well for areas adjacent to tissue–air interfaces.

D. Optical Imaging

Another way to obtain information about the functional state of neural tissue is to measure how light interacts with it. The most basic such method analyzes the light reflected by the exposed cortex during various functional states, a technique that has been mainly used in animal studies. This technique has provided *in vivo* high-resolution information, for example, of the functional organization of the visual cortex in cats and monkeys and of the whisker "barrel cortex" in rats. Because this technique requires drilling a hole in the skull and exposing the cortex, it is not generally suitable for human neuroimaging.

1. Near-Infrared Spectroscopy (NIRS)

The technique most widely used for noninvasive human optical imaging is near-infrared spectroscopy (NIRS). NIRS is possible because visible light can travel a significant distance through human tissue and still be detected. At wavelengths in the near-infrared region of the spectrum (700–900 nm), the optical penetration depth is several centimeters in biological tissues. Moreover, light absorption in this region is primarily due to HbO and Hb. Because different frequencies of light are absorbed differently by HbO and Hb, the concentrations of each can, in principle, be determined by measuring tissue absorption at multiple wavelengths. Thus, in NIRS, weak lasers shine light with the appropriate frequency from the scalp downward, and a set of photodetectors surrounds the location of the laser on the scalp. The photodetectors detect the amount of light transmitted at the different frequencies, and the amounts of HbO, Hb, and total hemoglobin ($HbT = HbO + Hb$) are calculated on this basis.

The average path photons travel from the source to the detectors is shaped roughly like a semicircle. The average depth reached by the photons is between 50 and 100% of the distance between the source and the detector. Thus, with a distance between source and detectors of about 4 cm, the average photon path encompasses gray matter in addition to other tissues. This observation illustrates two of the main limitations of optical imaging. The first is the maximum depth than can be probed with this technique, which is on the order of 2–3 cm. This means that subcortical nuclei and cortex lying deep in the sulci cannot be reached with optical imaging. The second limitation, related to spatial resolution, is caused by the various tissues interposed between the light sources, the gray matter, and the detectors. Of these layers, the most important is probably the cerebrospinal fluid, which produces "tunneling" phenomena—where light travels at a different rate through the fluid than though brain tissue, which complicates the interpretation of the optical signals.

a. Analysis of NIRS Data In general, the concentration of a chromophore (i.e., a light-absorbing molecule) in a volume of tissue can be determined by measuring the amount of light absorbed by the volume. All else being equal, an increase in the concentration of the chromophore results in an increase in the amount of light that is absorbed. The amount of absorption is quantified by the Lambert–Beer law, which assumes an infinitesimally small concentration of the chromophore and no scattering in the medium. Because biological tissues such as bone and white matter are highly scattering, a modified Lambert–Beer law has been used instead. This modified law introduces a factor to account for

scattering-related attenuation; furthermore, the distance between the source of the light and the detector is multiplied by an experimentally determined mean differential path length factor (DPF) that accounts for the longer path length of the photons in scattering media. It must be noted that this modified Lambert–Beer law works very well when changes in the concentration of the chromophores occur globally in the image volume, under the assumption of constant absorption along the photon pathlength. This assumption is not generally satisfied when changes in chromophore concentration occur only in a small volume within a larger sample, which is the typical situation encountered during focal brain activation. Once the time series of relative concentrations of Hb and HbO are calculated from the raw data and preprocessed (for example, by removing artifacts produced by the heart cycle and breathing), analysis essentially proceeds as it does for ERP data.

2. Diffuse Optical Tomography (DOT)

An important advance in optical imaging allows the computation of spatial variations in Hb and HbO by using multiple lasers and photodetectors. This approach produces an image of the pattern of activation over most of the cerebral cortex; this approach is referred to as diffuse optical tomography (DOT), and it affords continuous, real-time measurements of the spatial distribution of HbO, Hb, and HbT (total hemoglobin). Given that DOT can assess the same variable measured by the fMRI BOLD technique, in addition to other variables not measured by fMRI, the combination of these two techniques should be particularly productive. So far, only a handful of studies have compared the results obtained with fMRI and optical imaging techniques. These studies have shown reasonable consistency in the spatial localization of cortical activity between the two methods, although the spatial resolution of these optical imaging techniques remains to be systematically tested. The devices used to perform DOT are more complex than those required to perform NIRS with a single light source–detector pair because it is necessary to identify each source of light that reaches any detector. Thus, the sources must either be pulsed individually at distinctive frequencies or the light must be encoded in such a way that each source is specified. Furthermore, to resolve Hb and HbO within the sample volume, the volume must be probed with multiple wavelengths of light and the DOT system must be able to distinguish these signals at each wavelength. Even these DOT systems, however, are at least 10 times less expensive than an MRI or a PET scanner.

3. Analysis of DOT Data

With many sources and detectors arranged over the head, an image can be constructed by using ''back-projection algorithms.'' In these algorithms, each source–detector pair provides a ''projection'' of the underlying tissue in separate semicircle-shaped traces. These projections can then be combined to form an image of the spatial distribution of the tissue absorption and scattering coefficients. As for data from EEG and MEG, this is an ill-posed inverse problem in which many solutions are possible given any set of measurements. A number of current algorithms for image reconstruction are based on the ''photon diffusion equation,'' which is only accurate for highly scattering media. The nonscattering properties of the cerebrospinal fluid are known to introduce perturbations in photon migration. More recent algorithms are based on the more accurate radiative transport equation. By taking into account the anatomical and functional information provided by MRI, the accuracy of the inverse solution in the optical domain is likely to improve, just as it does in the case of EEG and MEG.

4. Fast Optical Signals

In addition to monitoring vascular changes, optical imaging also has the potential to monitor some aspects of brain function that are directly related to neural activity. *In vitro* studies have shown that light scattering in neural tissue changes with activation, probably because of changes in the refraction index of neural membranes. An increase in scattering causes photons to bounce around more times, thereby increasing the time they take to travel from the source to the detector (*time-of-flight* or *delay*). Note, however, that photon delay is also affected by tissue absorption, so that mathematical models are required to tease apart the scattering and absorption components. Studies have shown that fast task-related signals can be detected by measuring photon delay. The time course of these short-latency optical signals is comparable to that obtained with electrical event-related potentials. Although these findings await replication by other laboratories, they suggest that optical imaging is a very promising technique for noninvasive neuroimaging.

5. Summary of Advantages and Disadvantages of Optical Imaging

In sum, optical imaging techniques have the advantages of being relatively inexpensive, potentially having excellent temporal resolution, and having moderately good spatial resolution (probably comparable to PET). The major disadvantages are that they can record signals only from the cortex and only from regions that are not buried too deeply in sulci.

IV. FUTURE DIRECTIONS

Although ongoing incremental improvement of existing techniques will take us closer to the ultimate goal of neuroimaging, as described in the introduction, the most promising direction for functional neuroimaging is probably the combination of imaging techniques with complementary strengths that can, in principle, provide better spatial or temporal resolution than any one technique by itself. Combining techniques, however, is very challenging. On the one hand, different techniques often measure different aspects or consequences of neural activity. Thus, the combination of techniques must deal with the problem that each technique may register information not registered by the other and vice versa. In addition, the spatial localization of the same sources may be systematically different with different techniques. For example, as discussed earlier, MEG is thought to reflect postsynaptic activity, whereas fMRI reflects the indirect consequences of presynaptic activity; thus, it is conceivable that the locations of ''activation'' detected with the two techniques might not always coincide. On the other hand, the ability to obtain simultaneous measurements with two or more neuroimaging techniques sometimes is technically challenging or even impossible. For example, it is difficult to obtain good simultaneous EEG and fMRI measurements, and it is not currently possible to obtain simultaneous MEG and fMRI measurements.

Given the enormous advances in neuroimaging techniques over the past 15 years, we have every reason to be optimistic that current limitations will be surmounted as available techniques are refined and new techniques introduced. The end of the twentieth century was a watershed period for the study of the human brain, in large part because of the flowering of this technology.

See Also the Following Articles

ELECTROENCEPHALOGRAPHY (EEG) • FUNCTIONAL MAGNETIC RESONANCE IMAGING (fMRI) • IMAGING: BRAIN MAPPING METHODS • MAGNETIC RESONANCE IMAGING (MRI)

Suggested Reading

Damasio, H. (1995). *Human Brain Anatomy in Computerized Images*. Oxford University Press, New York.

Frackowiak, R. S. J., Friston, K. J., Frith, C., Dolan, R., and Mazziotta, J. C. (Eds.). (1997). *Human Brain Function*. Academic Press, New York.

Moonen, C. T. W., and Bandettini, P. A. (Eds.). (1999). *Functional MRI*. Springer-Verlag, Berlin.

Nunez, P. L. (1981). *Electric Fields of the Brain*. Oxford University Press, New York.

Roland, P. E. (1994). *Brain Activation.* Wiley-Liss, New York.

Toga, A. W., and Mazziotta, J. C. (Eds.). (1996). *Brain Mapping: The Methods.* Academic Press, San Diego, CA.

Toga, A. W., Mazziotta, J. C., and Frackowiak, R. S. J. (Eds.). (2000). *Brain Mapping: The Systems*. Academic Press, San Diego, CA.

Villringer, A., and Dirnagl, U. (Eds.). (1997). *Optical Imaging of Brain Function and Metabolism: Physiological Basis and Comparison to Other Functional Neuroimaging Methods*. Kluwer Academic–Plenum Publishers, New York.

Wong, P. K. H. (1991). *Introduction to Brain Topography*. Kluwer Academic–Plenum Publishers, New York.

Neuron

MARYANN E. MARTONE and MARK H. ELLISMAN
University of California, San Diego School of Medicine

GLOSSARY

action potential Term applied to the electrochemical signal caused by a rapid and reversible fluctuation in membrane potential used by the nerve cell for intracellular communication. Also known as a nerve impulse.

axon A cylindrical process extending from the nerve cell body specialized for conducting electrochemical signals away from neurons to their targets.

dendrite Name given to one or more tapering, typically branching processes extending from the nerve cell body that are specialized for receiving signals from other nerve cells and the environment.

neurotransmitter Any one of a number of small molecules that are released by a nerve cell at the synapse to pass signals between nerve cells and their targets.

synapse A term applied to the specialized junction between a nerve cell and its target where signals are passed from one cell to the next, usually in the form of chemical signals.

This article describes the major features of the neuron, the class of cell most closely identified with the functions of the nervous system. The major emphasis is on the structure of nerve cells and their basic physiology, along with their biochemical specializations.

I. INTRODUCTION

The neuron is a member of the class of cells most closely identified with the functions of the nervous system, namely, the transduction, processing, storage, and translation of environmental information into meaningful behaviors. The term neuron was coined by the German neuroanatomist Wilhelm Waldeyer in 1891. Neurons are unique among cells in their large size, complexity, and the diversity of their forms, which are closely related to the functions they subserve. The neuron is the fundamental anatomical building block of the neuronal circuits underlying behavior, but it is not the only cell type in the nervous system nor indeed the most numerous. That distinction belongs to a second class of cells called the neuroglia or simply glia. Although the focus of this article is the neuron, a neuron does not exist in isolation nor can it properly be described in the singular, because its interactions and associations with other neurons, glia, and target cells are integral to its function.

II. HISTORY OF THE NEURON

Most types of cell do not have a history.
Peters, Palay, and Webster (1991)

The challenge to neuroscientists imposed by the size and complexity of neuronal form is illustrated by the colorful and somewhat contentious history of the neuron concept. Once the concept of the cell as the unit of structure and function in all living things was established by Schleiden and Schwann in the early nineteenth century, scientists rapidly accepted that various tissues were made up of individual cells. Not so with nervous tissue, where nearly another 100 years would pass before the fact that neurons were distinct cellular entities, also known as the Neuron Doctrine,

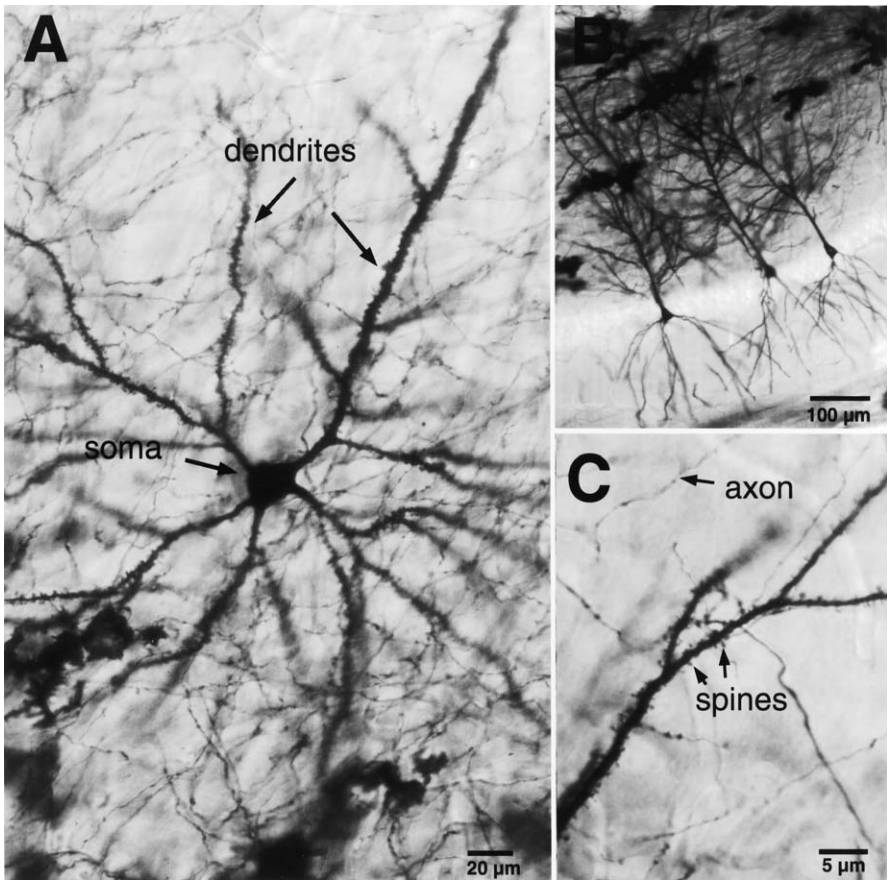


Figure 1 Neurons stained by Golgi impregnation. (A) Pyramidal neuron from the rat cerebral cortex. The plexus of fine fibers visible in the background is likely axon collaterals. (B) Three pyramidal neurons from the rat hippocampus. (C) Higher magnification view of a spiny dendrite showing the spiny protuberances from the dendritic shaft and varicose axons in the background.

would become universally accepted. Competing with the adherents of the Neuron Doctrine were the reticularists, who maintained that the nervous system comprises a syncitia of anastomosing filamentous processes in protoplasmic continuity with one another, much like a network of blood vessels. A leader of the reticularist school was the Italian neuroscientist Camillo Golgi, also famous for introducing the eponymous stain that was instrumental in elucidating the structure of the neuron, the Golgi reaction. Until the development of the Golgi reaction, most stains revealed only a complex jumble of fibrillar structures intertwined with corpuscular globules. The Golgi stain had the advantage of staining only a small number of neurons but revealing their entire form in glorious detail, including their far-reaching processes (Fig. 1). Ironically, the man most instrumental in putting forth the Neuron Doctrine, Santiago Ramon y Cajal, based most of his observations on material stained with the Golgi method developed by his competitor. The two scientists shared the Nobel prize in 1906.

With the development of biological electron microscopy in the 1950s, the limiting membrane of nerve cells and the separation between cells at points of contact were clearly visualized. However, whereas these early images from electron microscopy failed to reveal evidence for cytoplasmic connections between neurons at these arborized sites, more recently developed methods have revealed filamentous structures bridging the synaptic gap. This system, dubbed the transcellular filament system, may represent the constituents stained by the Golgi method, which Golgi interpreted as a network. Indeed, in the end, Golgi may have been partially right! Clearly it was correct to extend the cell doctrine to the nervous system and define the neuron. It appears that it was also correct to posit that a continuum of filamentous structures provides a reticular fabric for the nervous system and

that this is transcellular, especially at the synapses, which were at the heart of this controversy.

The establishment of the neuron doctrine did not mark the end of the evolving conception of the nerve cell. As our understanding of nerve cell physiology, biochemistry, and molecular biology has progressed, neuroscientists have had to create and modify models of information processing in the nervous system. For example, simple views of nerve cell processes as cables that either actively or passively propagated information in the form of electrical signals have given way to an increased emphasis on chemical signaling pathways and compartmentalization. The emphasis on molecular interactions has been fueled by advances in molecular genetics and methods for identifying and localizing the chemical constituents of nerve cells by using techniques such as immunocytochemistry. Thus, any review of the neuron represents a snapshot in time with an emphasis that is highly dependent upon the prevailing winds of scientific inquiry. However, regardless of the prevailing wind, any discussion of the neuron necessarily begins with its morphology.

III. STRUCTURE AND FUNCTION OF NERVE CELLS

A. Neurons Are Excitable Cells

Neurons are specialized to sense, transduce, process, store, and transmit information. Information is carried both within and between neurons in the form of electrical and chemical signals. Neurons are known as excitable cells, in that they are capable of altering their membrane potential to serve as a signaling mechanism. The membrane potential arises because the plasma membrane is not freely permeable to ions present in the intracellular and extracellular environment. The major ions used by nerve cells to carry signals are Na^+, K^+, Ca^{2+} and Cl^-. Ion channels specific for these ions are present in the plasma membrane. When they open, ions may pass into or out of the cell depending upon their concentration gradients, changing the membrane potential and also initiating second messenger cascades, activating enzymes, modifying the cytoskeleton, and regulating gene transcription. Ionic gradients are maintained by the action of electrogenic pumps, which actively pump ions into and out of the cell against their concentration gradients. These pumps are energy-dependent, that is, they require the hydrolysis of adenosine triphosphate (ATP) and are thus referred to as ATPases.

B. Neuronal Development

Embryologically, the nervous system derives from the thickening and infolding of the ectoderm, which also gives rise to the skin. Common progenitor cells give rise to both neurons and glia. Nerve cells retain many characteristics of epithelia, e.g., a polarized structure, little intracellular space, and the presence of numerous contacts between cells. The bulk of neurons in mammals are generated before birth, and once differentiated neurons are almost exclusively postmitotic. However, as with most well-established truths about the nervous system, this concept has had to be revised in light of evidence that nerve cells in some brain regions continue to be produced throughout the life of the organism, notably in the olfactory bulb and the dentate gyrus of the hippocampus. Work has also established that pluripotent stem cells are found in the brains and spinal cords of adult animals, which are capable of transforming into neurons and glia when given the proper signals.

C. Neurons Are Highly Polarized Cells

The structure of the nerve cell is well-suited for intracellular and intercellular information transfer. Neurons are highly polarized cells, with long processes that extend from the cell body to interact with other nerve cells, muscles, or glands (Figs. 1 and 2). Neurons are often classified according to the number of processes extending from their cell bodies as multipolar, bipolar, and monopolar. Neurons are the most morphologically diverse population of cells in an organism, varying widely in their size, form, and number of processes. The following will serve as a general description of the most common attributes of neurons, with a focus on vertebrate neurons. However, as with almost every aspect of the nervous system, exceptions to nearly every general statement can be found.

Two types of processes are distinguished in the typical nerve cell: dendrites and axons. These processes are sometimes referred to generically as neurites, particularly in developing neurons grown in culture and in invertebrates where many of the distinctions described in the following sections are less clear. Dendrites can be considered as extensions of the cell body specialized for receiving information from other nerve cells or the environment. The number of primary dendrites emanating from the cell body can vary from zero to many and is characteristic of a particular cell

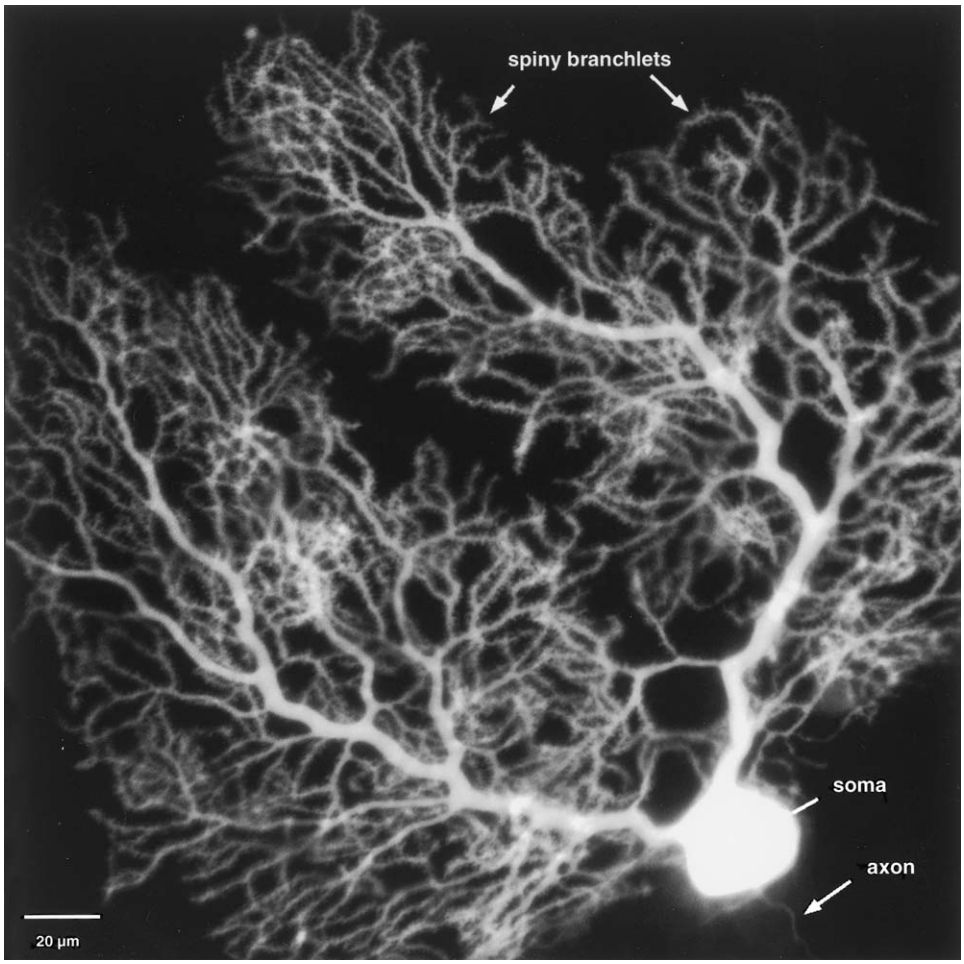


Figure 2 Purkinje neuron from the rat cerebellum stained by intracellular injection of the fluorescent dye lucifer yellow. The magnificent dendritic tree is revealed in its entirety. The fuzzy appearance of the finer caliber dendritic branches is due to the high density of dendritic spines (spiny branchlets). A single axon is visible emerging from the cell soma, although it is stained very faintly compared to the dendritic tree due to its small caliber.

type. The form of the dendritic tree can also be very distinctive, ranging from the simple morphologies of sensory cells to the magnificent tree of the Purkinje neuron (Fig. 2). The axon carries information away from the cell soma and transmits it to other neurons, muscles, and glands at a specialized cellular junction called the synapse. Information is conducted down the axon in the form of electrochemical signals known as action potentials. Transmission of signals to the next cell in the network usually involves the release of chemical neurotransmitters at the synaptic terminal, which interact with receptors in the target cell. The view that dendrites and axons are specialized for reception and transmission, respectively, was first proposed by Ramon y Cajal as the Law of Dynamic Polarization. However, examples where dendrites pass information between them (dendrodendritic transmission), signals are passed from the dendrite to the axon, and cells have no axons at all are numerous enough to suggest that this concept is only useful as a starting point for understanding neural circuits.

Since the time of Ramon y Cajal, neuroanatomists have spent considerable effort describing the morphology of nerve cells, from their overall form to their internal organization. The overall form has been studied by using techniques like the Golgi reaction (Fig. 1) and, more recently, by the injection of intracellular dyes, which reveal the pattern of the dendritic and axonal arbors of individual nerve cells (Fig. 2). The technique of intracellular dye injection is very powerful because it can be combined with intracellular recording in living tissue. By using this

method, neuroscientists have produced elegant studies correlating the form, chemical identity, and physiology of individual neurons. The form of a neuron is important in several ways in determining its functional properties. For example, the distribution and extent of the dendritic tree and axon will determine how a neuron fits into neural circuits. A neuron like the pyramidal cell of the cortex has dendrites that extend hundreds of microns and will receive different types of information compared to a cell such as the stellate cell, which has a much more limited dendritic field. Simple properties such as the diameter of the axon will affect the speed of neurotransmission, with larger caliber axons conducting faster than smaller caliber axons. Similarly, the diameter and branching pattern of dendrites along with the distribution of ion channels and other signal transduction molecules can affect the spread of a signal in the dendritic tree.

To develop and maintain their characteristic forms, neurons have well-developed cytoskeletons. The cytoskeleton is important not only for establishing and maintaining neuronal form but also for the transport of proteins and organelles throughout the neuron. The axons and dendrites of nerve cells can extend for hundreds or thousands of microns, necessitating an efficient mode of intracellular transport. The three major types of cytoskeletal elements found in all cell types, including neurons, are (1) microfilaments, e.g., actin, $\sim 7\,\mu m$ in diameter, (2) intermediate filaments, including neurofilaments, one of the neuron-specific classes of intermediate filaments, $\sim 10\,nm$ in diameter, and (3) microtubules, $\sim 25\,nm$ in diameter.

Although the morphology of a neuron is a good starting point for the discussion of neuronal diversity and function, the distinction between different types of neurons goes well beyond their unique morphology. Neurons are as heterogeneous in their biochemistry and physiology as they are in their morphology. Indeed, the nervous system by far displays the most molecular diversity of any tissue. The advent of immunocytochemical techniques to localize proteins and other small molecules and *in situ* hybridization to localize unique nucleic acid sequences has shown that different populations of neurons can be distinguished by their complement of neurochemical constituents. Such constituents include neurotransmitters, neurotransmitter receptors, ion channels, calcium-binding proteins, and neuropeptides. Even morphologically identical populations of cells can sometimes be divided into subclasses on the basis of their neurochemical signatures. Neurons are also distinguishable by their physiological properties, such as whether they excite or inhibit their targets, and by their pattern of firing. A major goal of modern neuroscience is to create an integrated view of nerve cells whereby the physiological properties of neurons can be understood in terms of their structural and biochemical specializations.

IV. PARTS OF THE NERVE CELL

Nerve cells are typically divided into several distinct functional domains (Fig. 3), each with its own unique set of structural and biochemical characteristics.

A. The Cell Body

The cell body of the neuron, also referred to as the soma or perikaryon, can range in size from 5 to $100\,\mu m$. The neuron cell body is like that of any other cell with the usual complement of organelles: a nucleus, usually with a prominent nucleolus, which contains the DNA of the cell, rough and smooth endoplasmic reticulum (RER and SER), free polyribosomes, a well-developed Golgi apparatus, mitochondria, and lysosomes (Fig. 4). Neurons are particularly rich in mitochondria, which produce cellular energy in the form of ATP, in part because the maintenance of the ionic gradients described earlier is an energy-dependent process. Cytoplasmic proteins are synthesized by free ribosomes in the cytoplasm, whereas those destined for insertion into membranes or for secretion are synthesized by ribosomes on the RER and further processed by the Golgi apparatus. In many neurons, the RER is arranged into parallel stacks close to the nucleus. These stacks stain prominently with a class of basic stains called Nissl stains, e.g., cresyl violet and toluidine blue, and for that reason are sometimes referred to as Nissl bodies (Fig. 4). An interesting feature of neurons is that they do not contain stores of glycogen, as do many types of cells. Thus, they are reliant on a constant supply of blood to meet their energy needs. However, astrocytes, the most numerous type of glia, do contain glycogen stores and are intimately associated with both neurons and blood vessels, suggesting that they provide some metabolic support to the neuron.

The cell body is the trophic center of the cell. It performs the bulk of protein synthesis for the neuron and contains the nucleus. *In situ* hybridization studies show that most mRNA species are confined to the cell soma. Dendrites and axons separated from their soma cannot survive and will eventually degenerate.

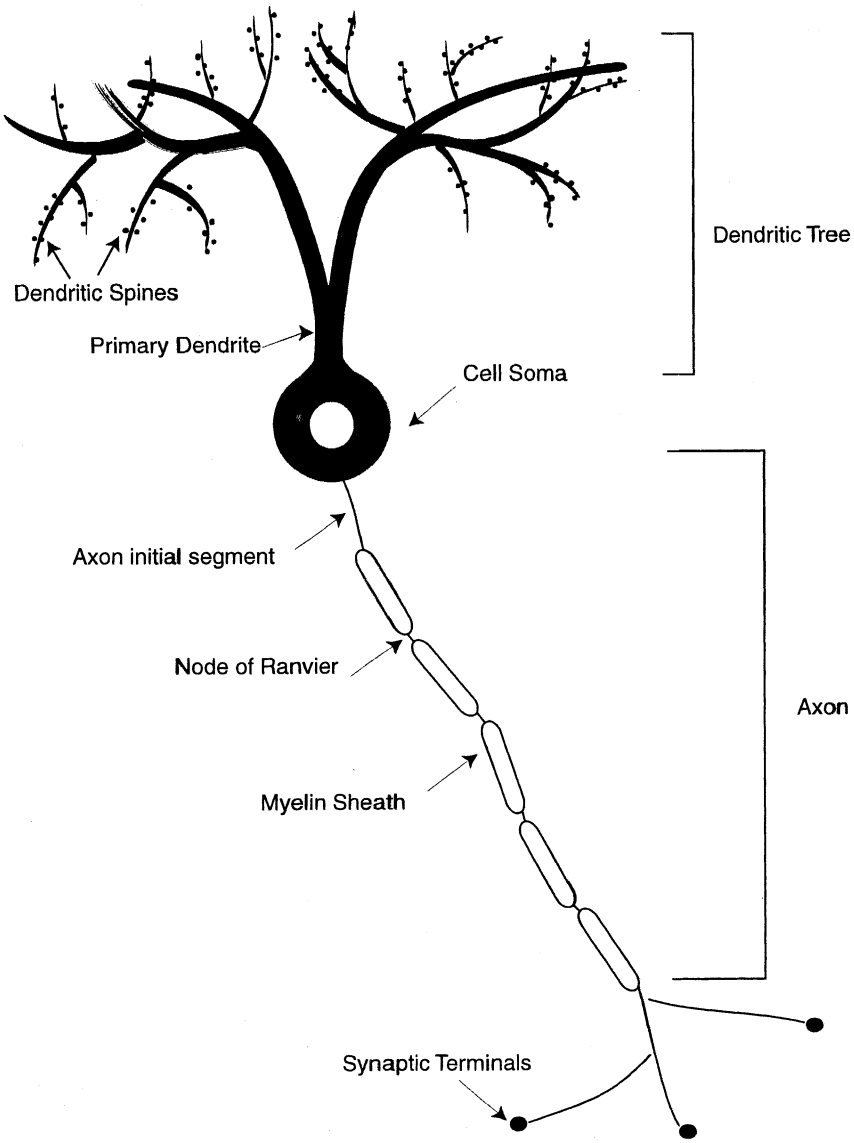


Figure 3 Schematic drawing of a neuron showing the major compartments and features.

However, neurons are also dependent for survival upon certain molecules, called growth factors, obtained from their targets via their axons and transported back to the cell soma. Especially during development, a neuron that fails to establish contacts with an appropriate target will degenerate.

When brain tissue is stained with a technique that identifies the cell bodies, e.g., a Nissl stain, the density of neuronal somata is very low compared to other tissues like muscle or skin. The bulk of nervous tissue is not composed of cell bodies in close apposition or extracellular space, but rather is composed of the intermingling of dendrites, axons, dendritic spines, and glial elements, which is called neuropil. Under the electron microscope, the neuropil appears as a dense conglomerate of small profiles (Figs. 4, 5, and 10).

B. Dendrites

One or more dendrites (Gr: "tree" or "branch") extend from the cell soma with a gradual tapering. Dendrites can extend for tens to hundreds of microns in length and can branch extensively, usually at acute angles and

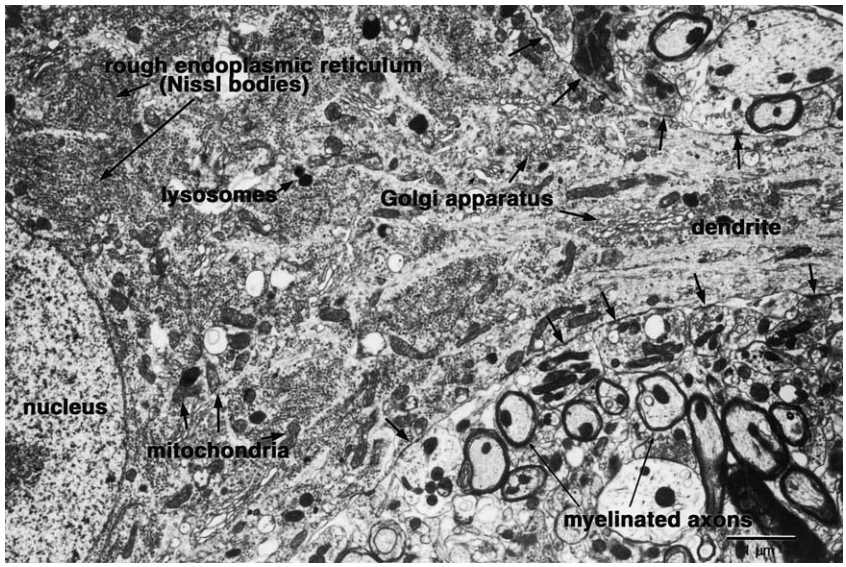


Figure 4 Electron micrograph of a motor neuron in the rat spinal cord showing the transition between the soma and a primary dendrite. The soma is the area surrounding the nucleus which gradually tapers into a dendrite. Numerous synaptic contacts onto the plasma membrane are visible (arrows). A group of myelinated axons is visible next to the soma.

with subsequent diminishment in dendritic diameter. There is often no clear morphological distinction between proximal dendrites and the soma; the same organelles, with the exception of the nucleus, are found in cell bodies and proximal dendrites, although the amounts of rough endoplasmic reticulum and Golgi apparatus diminish and eventually disappear from more distal dendrites (Fig. 4). In more distal dendrites, the only organelles found with regularity are mitochondria, the smooth endoplasmic reticulum, and multivesicular bodies (Fig. 5). The smooth endoplasmic reticulum is able to uptake, sequester, and release calcium, and dendrites express high levels of SER proteins involved in intracellular calcium regulation (Fig. 6). In some neurons, an expansion of the SER into cisterns closely apposed to the plasma membrane is seen. This expansion is very characteristic of Purkinje neurons and spinal motor neurons and is called the hypolemmal cisternae.

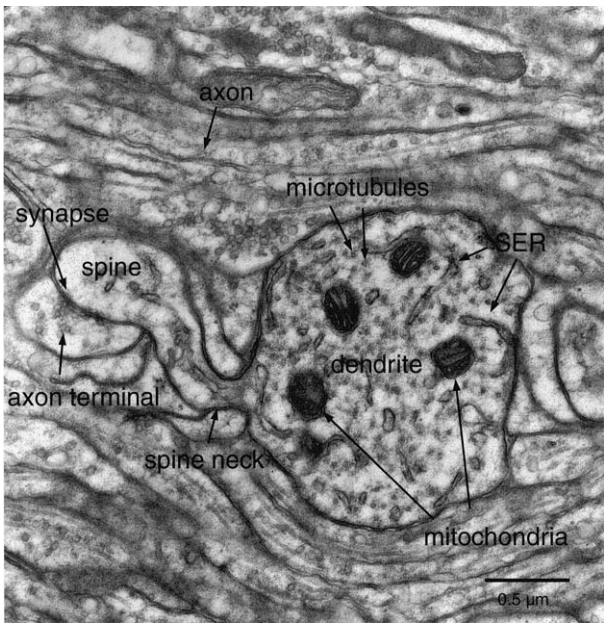


Figure 5 Electron micrograph of a cross section of a Purkinje cell dendrite in the chicken cerebellum with a single spine emerging from the dendritic shaft via a thin spiny neck. Small-caliber axons are visible coursing on either side of the dendrite.

Microtubules are the most prominent and abundant cytoskeletal element in dendrites and are aligned along the long axis of the dendrite, whereas intermediate filaments are less common. Although microtubules are found in the soma, axon, and dendrites, microtubules in each region are distinguished by the proteins associated with them. For example, dendrites are distinguished by high levels of MAP2 (microtubule-associated protein, type 2) compared to the soma and to axons. These microtubule-associated proteins appear to be important in determining the stability and arrangement of microtubules. Staining for MAP2 is often used as a method for identifying a neuronal process as a dendrite.

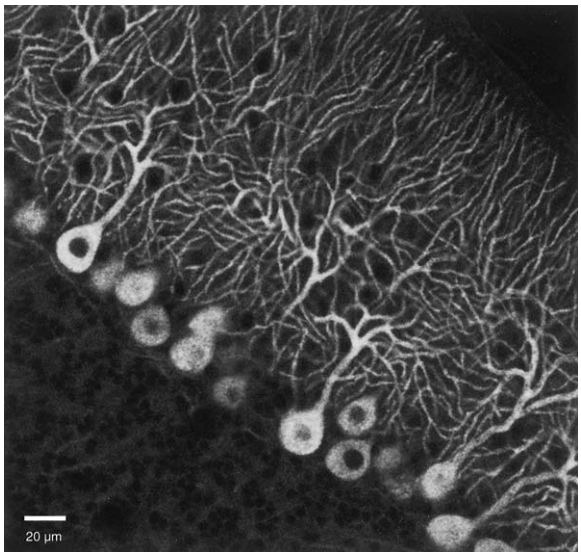


Figure 6 Purkinje neurons in the chicken cerebellum stained for the ryanodine receptor, a calcium channel located in the membranes of the smooth enodplasmic reticulum. When activated, the channel opens and releases calcium from stores inside the SER. Many other neurons are present in this brain region, but they do not stain for this particular protein.

Proteins and other constituents required for dendritic maintenance and functioning are transported into the dendrite from the soma by an active mechanism at the rate of about $\sim$20 mm/day. Whereas identifiable RER and the Golgi apparatus are almost never seen in distal dendrites, ultrastructural studies have shown that polyribosomes, some in close association with membranous cisterns, are found throughout the dendritic tree. Many of the polyribosomes are found near synaptic contacts. With the development of high-resolution *in situ* hybridization techniques, researchers discovered that a small number of mRNA species are also found localized in dendrites. These mRNAs are specifically targeted to dendrites by signal sequences contained within their 3′-untranslated regions and often encode proteins important for signal transduction. The most well-studied dendritic mRNA is that encoding the α form of calcium/calmodulin-dependent protein kinase II (CamKII), which is found in high levels in the dendrites of pyramidal neurons in the hippocampus. CaMKII is a multifunctional enzyme important in signal transduction at the synapse. Neurons grown in culture have been demonstrated to support protein synthesis in their dendrites, and some evidence exists that this also occurs *in vivo*. Researchers have hypothesized that these targeted mRNAs support protein synthesis that is activated in response to local activation of synapses and, thus, may be related to synaptic plasticity or the modulation of synaptic efficacy.

C. Dendritic Spines

Certain dendrites possess fingerlike extensions, 1–2 μm in length, called dendritic spines. In low magnification images, a high density of dendritic spines gives dendrites a fuzzy appearance (Figs. 1 and 2). Spiny protrusions are also found on the cell soma and on the axon initial segment, but the dendritic spines are more numerous and have received the most study. Neurons that have large numbers of spines on their dendrites are referred to as spiny neurons, whereas those with smooth dendrites or few dendritic spines are termed aspiny. Dendritic spines are one of the most intensively studied structures in the central nervous system because they receive the bulk of excitatory synaptic input and are thought to be key sites of synaptic plasticity, the process by which neurons are able to modify their properties in response to activity.

Dendritic spines vary widely in their size and morphology between classes of neurons and even within the same neuron (Fig. 7). The prototypical spine, seen in the hippocampus, neostriatum, cerebral cortex, and cerebellar cortex, consists of a thin stalk or neck, $\sim$100 nm in diameter, which arises from the main dendritic shaft, and a larger, rounder head. Dendritic spines can be distinguished from the main dendrite by the lack of mitochondria and microtubules, although some of the larger spines found in the olfactory bulb and hippocampal area CA3 contain both of these structures. Dendritic spines have very high concentrations of microfilaments, in particular filamentous actin. The majority of spines receive only a single synaptic contact onto the spine head, which usually uses the neurotransmitter glutamate. In the cerebral cortex and the neostriatum, about 20% of the spines receive a second synapse onto the spine neck, which is inhibitory or modulatory in action. The only other organelle seen with any regularity in the spine is the smooth endoplasmic reticulum, although coated and uncoated vesicles, multivesicular bodies, and single ribosomes have also been reported. Larger spines in the hippocampus, cerebral cortex, and neostriatum also contain a curious structure called the spine apparatus, consisting of lamellar plates of endoplasmic reticulum alternating with dense, filamentous material. The spine apparatus has

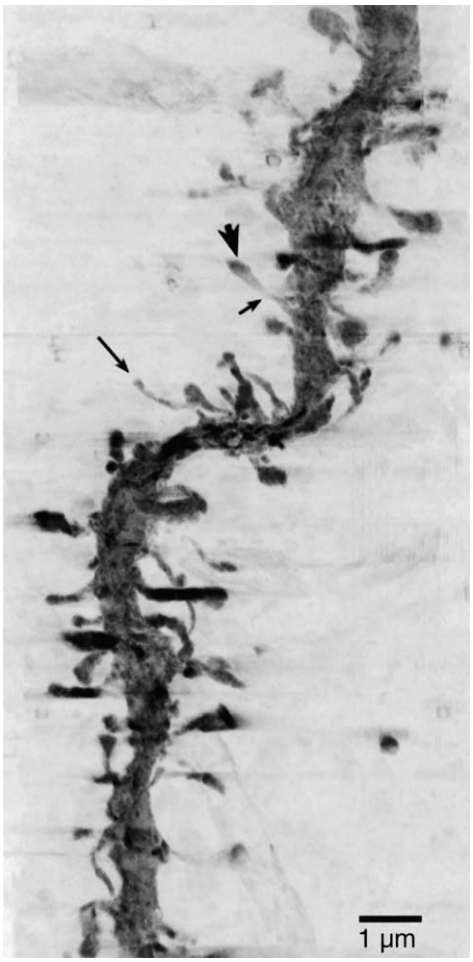


Figure 7 High-voltage electron micrograph of a spiny dendrite from the rat neostriatum showing the variation in size and shape of spines, even on a single dendrite. Spines can exhibit many morphologies, ranging from short stumpy spines to long thin spines without an obvious spine head (long thin arrow) to larger spines with an obvious head (short thick arrow) and neck (short thin arrow). The high-voltage electron microscope allows the microscopist to use much thicker sections than is possible in a conventional electron microscope. The thicker sections permit the viewing of dendritic spines in their entirety. The spiny dendrite was first injected with a fluorescent dye, which was then converted into a label that is visible under the electron microscope.

filamentous links with the postsynaptic site and may be involved in the shuttling of proteins to the synapse. As is the case with the SER in the dendritic shaft, proteins involved in the the uptake, storage, and release of intracellular calcium have been localized to the spine SER and spine apparatus. Release of calcium from these intracellular stores is thought to be an important signal transduction mechanism for synaptic transmission and may be involved in spine dynamics.

The most densely spiny cell in the CNS is the Purkinje neuron, which has been estimated to have upward of 150,000 spines per neuron. Precise counts of dendritic spines are difficult to obtain because they are too small to be resolved sufficiently in the light microscope, but they are too large to be contained completely within the thin sections typically employed for electron microscopy. To gain accurate spine counts, it is necessary to perform serial section reconstruction at the electron microscopic level or use high-voltage electron microscopy (Fig. 7). However, electron microscopic analysis is typically performed over a small portion of the entire dendritic tree, and the density of dendritic spines tends to vary over the length of the dendrite. Thus, a review of the literature can result in wildly disparate estimates of the total spine number on a given neuron type. Quantification of spine density is important because it gives an estimate of the total synaptic input to a given neuron.

Dendritic spines are labile structures in that neurons can change the size, shape, and number of their dendritic spines in response to developmental, environmental, pathological, and experimental influences. For example, researchers have shown that spine numbers are very sensitive to reproductive hormones such as estrogen. During the estrous cycle of the rat, the number of spines on a hippocampal pyramidal neuron can fluctuate by as much as 20%. Spine dynamics can be visualized directly by observing living neurons in culture or in brain slices. In these preparations, individual spines were observed to change their shape within seconds after a drug or electrical stimulus was applied to the culture dish. By using time lapse microscopy, researchers have observed new spines growing or established spines disappearing over the course of minutes.

The function of dendritic spines has been a matter of debate since their first description by Cajal. Their function clearly goes beyond simply increasing the surface area available for synaptic contact, because few synapses are established on the dendritic shafts of spiny neurons and only $\sim$10% of the total spine surface is taken up by a synaptic contact. Many theories have focused on their possible role in synaptic plasticity. One method by which neurons are thought to store information long term is by structural or biochemical modification of certain synapses. Because each spine usually receives only a single synapse, dendritic spines serve to isolate individual synapses. Rapid changes in spine shape could alter the signaling properties of a particular synapse, and the spine itself may serve as a local compartment whereby

modification of a given synapse can occur in isolation. Dendritic spines may also serve to protect the main dendrite from any negative effects associated with synaptic stimulation. As a consequence of synaptic activation, levels of calcium are raised inside the cell. These high concentrations of intracellular calcium are necessary for signal transduction but are also deleterious to the cell. Dendritic spines may allow local increases in intracellular calcium in the vicinity of the synapse while restricting its spread to other parts of the neuron.

D. The Axon

Axon is short for axis cylinder. Nerve cells may possess many dendrites but usually have only a single axon, which can be impressively long. Nerve cells projecting between the spinal cord and the brain and from the brain stem and spinal cord to muscles and glands can have axons over 1 m in length. At the other extreme, some small neurons seen in sensory systems have no axons, and these are termed amacrine cells. The main axon may give rise to many collateral branches, which typically branch off at obtuse angles. In some cases, a collateral can ramify within the dendritic field of the parent cell and nearby cells, which is known as a recurrent collateral, and examples of neurons synapsing upon themselves (autapses) have been described. Unlike dendrites, the diameter of the axon usually does not diminish with increasing distance from the cell soma. In a given brain region, neurons are typically classified into projection and intrinsic neurons, according to the trajectory of their axonal ramification. Projection neurons, also called principal, relay, or Golgi type I cells, have long axons and project outside of the region in which the parent cell resides. Intrinsic neurons, also called interneurons or Golgi type II cells, have shorter axons and ramify locally within a brain region. Interneurons tend to have inhibitory influences on their targets and are very important for shaping the spatial and temporal responses of principal neurons.

Axons can arise from either the cell body or primary dendrite, from a coned-shaped region termed the axon hillock. Unlike dendrites, where there is often a gradual changeover in cell constituents from the soma, there is an identifiable transition from soma to axon. The most proximal portion of the axon, which can be recognized in the electron microscope by its characteristic morphology, is referred to as the initial segment. Action potentials arise from the initial segment, which has an extremely high density of Na^+ channels. No rough endoplasmic reticulum, ribosomes, or Golgi apparatus extend into the axon much beyond the initial segment. Consistent with the lack of protein synthetic machinery, no protein synthesis is seen in axons. Initial segments can also be recognized in the electron microscope by bundles of microtubules aligned along the axon cylinder and by the presence of a fine, electron-dense fuzz lining the plasma membrane. As with dendrites, the smooth endoplasmic reticulum extends throughout the axon, where it is sometimes called the axoplasmic reticulum. Microtubules are also found throughout the axon, although in larger caliber axons, neurofilaments predominate. The microtubule-associated protein τ is concentrated in axons and can be used to identify a neurite as an axon in cell cultures.

Axons are almost totally reliant on the cell body for trophic support, and so axons separated from their somas will obligatorily undergo degeneration. This fact was exploited by early neuroanatomists, particularly in the 1950s and 1960s, for tracing axonal projections. A lesion was placed in an area containing cell bodies and then the brain was stained with silver-based stains specific for degenerating axons. By following the paths of degenerating axons, anatomists could infer the projection patterns of the lesioned neurons. Of course, any axons passing through the lesioned region damaged along with the cell bodies would also degenerate. In the modern anatomical age, degeneration-based methods have been largely replaced by the use of tracers that are injected into the brain, taken up by the cell body, and transported down the axon to their targets.

1. Axonal Transport

Axonal transport, the process by which protein complexes and membranous organelles are transported within axons, has been studied extensively. Axons are capable of bidirectional transport. Transport from the soma to the distal axon is known as anterograde transport, whereas transport from distal regions back to the soma is known as retrograde transport. Axonal transport is an energy-dependent process that involves microtubules and the microtubule-based motor proteins, the dyneins and kinesins. Several distinct components have been identified in axonal transport, which differ in their cargoes and the rate of transport. Small membrane-bound organelles such as clear-lumened vesicles and vesiculotubular structures are transported out of the cell soma in the fast component, capable of moving up to 100 mm/day.

Larger membrane-bound structures such as multivesicular bodies carry materials back to the cell body and are also transported by a fast mechanism. Mitochondria move independently of the anterograde and retrograde flows, and the axoplasmic reticulum is stationary relative to fast transport. Cytoskeletal and cytoplasmic proteins are transported much more slowly, on the order of 1–10 mm/day.

2. The Action Potential

Axons are able to transmit signals to their targets by the conduction of action potentials along their length. The action potential involves a sudden membrane depolarization followed by a rapid reversal. The ionic basis of the action potential was elucidated by Alan Hodgkin and Andrew Huxley in the 1940s and 1950s in pioneering studies using the giant axon of the squid. The main currents involved in generating the action potential are carried by Na^+ and K^+ ions. Na^+ is much more concentrated in the extracellular environment, whereas K^+ tends to be more concentrated inside the axon. The plasma membrane of axons contains ion channels permeable to Na^+ and K^+. These ion channels are known as voltage-gated channels in that the internal potential of the cell will determine whether the channel is open or closed. If the potential reaches a certain threshold value, the ion channel will undergo a conformational change that causes it to either open or close. The normal resting potential of a neuron is on the order of -60 to -70 mV, which is largely the result of a high concentration of K^+ ions inside the axon.

An action potential is initiated in the axon hillock when the synaptic signals received by the dendrites and soma are sufficient to raise the intracellular potential to the threshold potential of -55 mV. When this potential is reached, the Na^+ channels present in the axon initial segment will open. Na^+ ions rush into the cell, causing rapid reversal of the membrane potential from -90 to $+40$ mV. When the membrane potential reaches $+40$ mV, the Na^+ channels close and the voltage-gated K^+ channels open. K^+ ions move out of the axon, thereby repolarizing the membrane. As a result of this coordination of Na^+ and K^+ channels, the action potential is circumscribed in time and restricted in space to a local patch of membrane. However, some current spreads passively down the axon, driving the membrane potential toward threshold in an adjacent patch of membrane and the process is repeated. In this way, the action potential is propagated down the axon in an anterograde direction without a loss of amplitude. The action potential is prevented from traveling back toward the soma due to factors that make the preceding patch of membrane refractory to stimulation. Following the cessation of the action potential, Na^+/K^+ ATPases in the membrane remove Na^+ from the interior of the cell and uptake K^+ from the extracellular space, restoring the proper ionic balance. The speed of conduction in different axons varies tremendously from slow conduction (~ 0.5 m/sec) to very fast (80–120 m/sec). The speed of conduction of a given axon is directly proportional to the axon diameter and the density of ion channels in the membrane. Another major determinant of the speed of axon conduction is the presence or absence of the myelin sheath.

3. The Myelin Sheath

Fast-conducting axons in the vertebrate nervous system are wrapped in a thick, fatty coating called myelin, which is formed by glial cells. Myelin in the peripheral nervous system is formed by a type of glial cell known as the Schwann cell, whereas in the central nervous system, myelin is formed by the class of glia known as oligodendroglia or oligodendrocytes. Under the electron microscope, myelin appears as a series of concentric layers enveloping the axon (Fig. 8). These layers are actually the membrane of a single myelinating glial cell spiraled around the axon cylinder with the cytoplasm of the glial cell squeezed out. The dark staining is due to the fixation of the nerve cell with the heavy metal osmium tetroxide, which reacts primarily

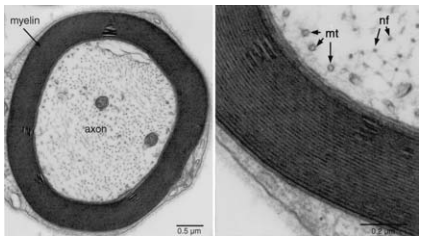


Figure 8 Electron micrograph of a myelinated nerve fiber in a rat peripheral nerve. (Left) Cross section of a myelinated axon showing the characteristic dense staining of the myelin sheath. Part of the cytoplasm of the Schwann cell forming the myelin sheath is also visible. (Right) Higher magnification view of the myelin sheath showing the concentric layers of the myelin sheath formed by the plasma membrane of the Schwann cell spiraling around the axon with the cytoplasm squeezed out. Some of the Schwann cell cytoplasm is visible around the outer layer of the myelin sheath. Microtubules (mt) and neurofilaments (nf) cut in cross section are clearly visible.

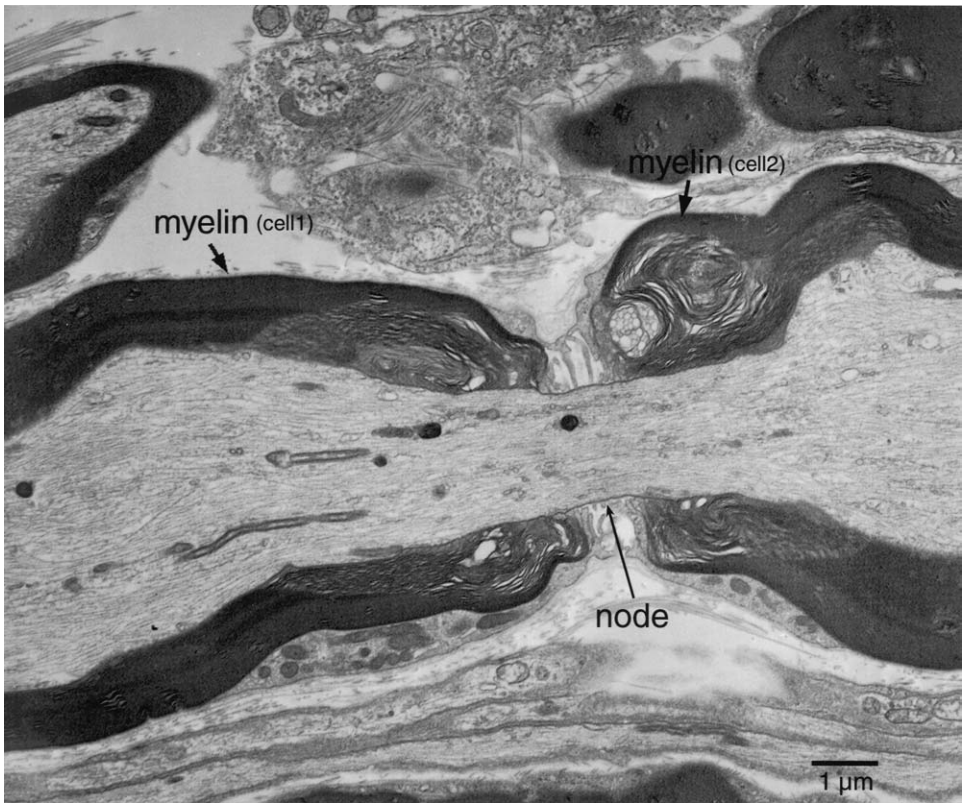


Figure 9 Electron micrograph of a node of Ranvier from a rat peripheral nerve. The nodal gap between the myelin formed by two adjacent Schwann cells is clearly visible.

with lipids and adds significantly more electron-scattering stain to these sites. In unstained tissue, the presence of myelin is what gives bundles of axons their macroscopic white, shiny appearance.

The presence of the myelin sheath tremendously increases the speed and efficiency of nerve conduction by improving the capacitive and resistive properties of the axon cable. The myelin sheath of these insulated axons is not continuous but is interrupted at regular intervals, 1–2 mm apart. The interruptions occur where the ensheathment of one glial cell ends and another begins (Fig. 9). The gaps in the myelin sheath were first described in the 1870s by the French neuroanatomist Louis-Antoine Ranvier, who also described myelin, and have been named nodes of Ranvier in his honor. Nodes of Ranvier have high concentrations of Na^+ and K^+ channels in the plasma membrane. The myelin sheath serves as an insulator to the axon, preventing the leakage of ions across the membrane in the internodal regions. Action potentials are generated only at the node, rather than continuously down the length of the axon. This form of axonal conduction is called saltatory conduction because the action potential jumps from node to node. Myelination is usually observed in axons $>1\,\mu m$ in diameter. Smaller axons in the peripheral nervous system are still ensheathed by Schwann cells, but no myelin is formed. In the central nervous system, small axons usually have no glial covering. The importance of the myelin sheath for normal nerve function is underscored by the devastating effects of demyelinating diseases, such as multiple sclerosis, which result in significant neurological impairments.

E. The Synapse

The term synapse (Gr: "fasten together") was first coined by the physiologist Charles Sherrington in the late nineteenth century. Synapses are specialized cellular junctions that are the site of intercellular communication between neurons and their targets. Two types of synapses are distinguished, chemical and electrical. Electrical synapses are also known as gap

junctions and are formed by proteins called connexins, which provide a low-resistance conduit between cells where ions and other small molecules can pass directly from cell to cell. Transmission through electrical synapses is very rapid and can provide a mechanism for synchronizing the firing of groups of connected neurons. Electrical synapses are more common between neurons in invertebrates but are found in vertebrates as well. Gap junctions are not unique to nerve cells but are also formed between glial cells and are found in other types of tissue. The most common form and best-studied synapse in the vertebrate nervous system is the chemical synapse, where signals are passed from one cell to the next through the agency of chemical neurotransmitters (Fig. 10). The neurotransmitter is released from small vesicles contained in the axon terminals of one cell (designated the presynaptic cell), diffuses across the extracellular space, and interacts with receptor proteins located in the postsynaptic cell. The postsynaptic cell is most commonly a neuron, but it can also be a muscle, gland, or organ. In fact, most of what is known about synaptic transmission is derived from studies of the neuromuscular junction, the synapse between motor neurons and the skeletal musculature, because this synapse is large and relatively accessible compared to synapses in the central nervous system.

Chemical synapses provide polarized, point-to-point zones of intercellular communication. Anatomically, synapses are difficult to identify conclusively at the light microscopic level without some sort of special stain, although Ramon y Cajal correctly identified the expansions at the axon terminal closely apposed to the dendrites and somata of other nerve cells as the point of intercellular communication. The anatomical correlate of the synapse was described definitively by Sanford Palay, George Palade, and others with the advent of biological electron microscopy starting in the 1950s (Fig. 10). Synapses can occur at the terminal end of axons, in which case they are called synaptic terminals or synaptic boutons, or they can occur along the length of the axon at local swellings. This type of synapse is very common in the mammalian nervous system and is called an *en passant* synapse. In myelinated axons, the *en passant* synapse occurs at the nodes of Ranvier. In the case of the terminal boutons, axons lose their myelin sheath prior to the terminal expansion.

The synapse and its associated specializations are most clearly visualized under the electron microscope (Fig. 10). Presynaptically, the synapse is characterized by the presence of a concentration of clear vesicles,

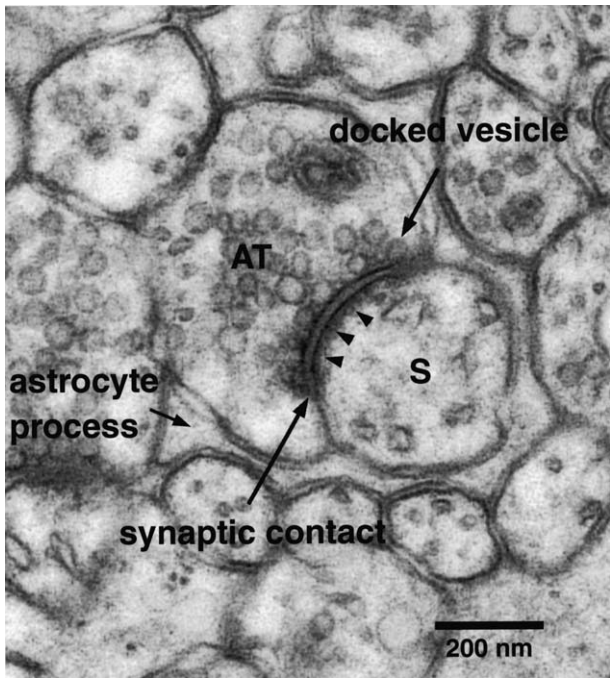


Figure 10 Electron micrograph of an asymmetrical synaptic contact onto a dendritic spine (S) in the rat cerebellum. The characteristic features of a synaptic contact are clearly visible, including the concentration of vesicles in the presynaptic axon terminal (AT), the widening of the extracellular cleft at the site of contact, and the presence of a prominent postsynaptic density (arrowheads). This synaptic complex is surrounded by an astrocytic process.

referred to as synaptic vesicles, $\sim$40 nm in diameter. These vesicles contain neurotransmitter molecules, which are released upon activation (see later discussion). Vesicles are clustered around presynaptic membrane specializations, consisting of a series of regularly spaced, electron-dense projections termed the active zone, believed to aid in the docking of vesicles for the release of neurotransmitter. More elaborate synaptic specializations can be seen in some synapses, such as the synaptic body of the vestibular hair cell and the ribbon class synapses found in several cell types. These specializations are thought to serve as a mechanism for conveying vesicles to docking sites. Other elements present in the presynaptic terminal include dense-core vesicles, mitochondria, microtubules, neurofilaments, and smooth endoplasmic reticulum.

The synaptic cleft, the intercellular gap across which the neurotransmitter diffuses, is also distinct. It tends to be wider than the average nonsynaptic distance between cells, on the order of $\sim$20–25 nm, and

contains electron-dense material. Fine filaments have been described spanning the cleft. Numerous cell adhesion molecules such as the NCAMs (neural cell adhesion molecules) and cadherins are found at the synapse, which serve to anchor the pre- and postsynaptic membranes. The adherence at the synapse is very strong and has been exploited in biochemical studies where intact synaptic complexes, or synaptosomes, can be recovered using cell fractionation techniques.

On the postsynaptic side, the major feature is the presence of a postsynaptic specialization consisting of electron-dense, filamentous material apposed to the postsynaptic membrane directly across the cleft from the active zone (Fig. 10). The electron dense material in the postsynaptic cell is called the postsynaptic density. In some types of synapses, especially those occurring in dendritic spines, the postsynaptic density is quite prominent, giving an asymmetrical appearance to the synapse. These synapses are often referred to as asymmetrical synapses and tend to be associated with excitatory synaptic transmission. Other synapses have much less prominent postsynaptic densities and are usually referred to as symmetrical synapses. Often, these synapses are associated with inhibitory synaptic transmission. These distinctions are not absolute and many synapses of intermediate morphology have been described. Asymmetrical and symmetrical synapses were first described by Edward Gray in the 1950s and are sometimes referred to as Gray's type I and type II synapses, respectively.

Much progress has been made to elucidate the composition of the postsynaptic membrane and density. In addition to neurotransmitter receptors, the postsynaptic membrane also contains proteins called cadherins as well as other adhesion molecules responsible for the tight adherence of pre- and postsynaptic membranes. Cytoskeletal proteins such as actin, tubulin, and fodrin are present in high abundance. Proteins involved in the anchoring of neurotransmitter receptors and ion channels to the cytoskeleton have also been identified. These proteins, also known as PDZ domain containing proteins, serve both to anchor neurotransmitter receptors and ion channels in the postsynaptic membrane and to link neurotransmitter receptors to downstream signaling proteins forming large macromolecular complexes. Other multifunctional signal transduction proteins such as CaM kinase II, DARP32, and protein kinase C may also be present in the postsynaptic density. Thus, the postsynaptic density serves to concentrate and organize signal transduction machinery at the synapse.

Synaptic complexes do not usually occur in isolation in the vertebrate nervous system but are closely associated with glial cells. In the central nervous system, the processes of astrocytes envelop or partially surround the presynaptic terminal and sometimes wrap around both the pre- and postsynaptic elements (Fig. 10). Astrocytes are the most numerous type of glial cell in the central nervous system. The extent of glial investiture of synapses varies across brain regions and even from synapse to synapse in a given brain region. Astrocytic processes are known to be able to take up certain neurotransmitters like glutamate from the extracellular environment and also to regulate the extracellular concentration of K^+ ions. Studies have suggested that glial cells are important for regulating synaptic transmission at glutamatergic synapses and for protecting against neurotoxic concentrations of glutamate in the extracellular space.

1. Synaptic Transmission

Chemical neurotransmission is a carefully regulated process. Synaptic transmission at the chemical synapse involves multiple steps and thus is slower than transmission through electrical synapses, but it provides the advantage of directionality and amplification of signal transmission. The major steps involved in the most rapid form of chemical synaptic transmission can be summarized as follows. Action potentials invade the synaptic terminal and open voltage-sensitive calcium channels present in the plasma membrane. Calcium triggers a series of molecular interactions ultimately causing vesicles in the vicinity of the vesicle docking area or active zone to fuse with the presynaptic membrane and release their contents into the synaptic cleft. This process is very fast, occurring on a submillisecond time scale. The neurotransmitter diffuses across the synaptic cleft and interacts with neurotransmitter receptors anchored to the postsynaptic membrane by cytoskeletal proteins in the postsynaptic density. Excess neurotransmitter either is broken down by enzymes located in the cleft, diffuses away from the synaptic site, or is taken up by the presynaptic cell through the action of transporter molecules present in the presynaptic membrane or surrounding glial cells. Lipids and specific proteins contained in the synaptic vesicle are then recycled from the plasma membrane using a clathrin-mediated mechanism, and in most systems the recycled vesicles are refilled with neurotransmitter and otherwise readied for another release cycle. A large number of proteins involved in vesicle docking, transmitter uptake and release, signal

transduction, and vesicle recycling have been identified. Unraveling of the molecular players in synaptic transmission is critical not only because synaptic transmission is fundamental to the functioning of the nervous system but also because many toxins and neuroactive drugs target proteins involved in this process.

The binding of a neurotransmitter to its receptor can result in several postsynaptic events occuring simultaneously. In some cases, the neurotransmitter receptor is an ion channel that allows Na^+ or Ca^{2+} into the cell (called ionotropic receptors). Ionotropic responses tend to be fast, occurring in the millisecond range. In other cases, the neurotransmitter activates a second messenger system like cyclic AMP or phosphoinositol breakdown (metabotropic receptors). The actions initiated via these metabotropic receptors are much slower than ionotropic response, occurring over seconds or even minutes. The result of neurotransmission can be either excitatory or inhibitory. Excitatory neurotransmission drives the postsynaptic neuron closer to firing an action potential, whereas negative neurotransmission drives the neuron farther from firing an action potential. Excitatory responses tend to involve intracellular increases of Na^+ or K^+ ions, resulting in membrane depolarization, whereas inhibitory responses usually involve increases of Cl^- ions, resulting in hyperpolarization. Beyond the voltage changes observed in the postsynaptic cell, a host of other signal transduction cascades are activated, usually in response to a rise in intracellular calcium. Calcium can activate various enzyme complexes such as protein kinases and proteases. Signals can be propagated to the nucleus to activate gene transcription. Several immediate early genes such as *c-fos* are turned on as a result of synaptic activation.

A variety of small molecules are now known to act as neurotransmitters. The prototypical fast excitatory neurotransmitter in the vertebrate nervous system is the amino acid glutamate, whereas the most common inhibitory neurotransmitter is γ-aminobutyric acid, usually referred to as GABA. Although these molecules are referred to as excitatory or inhibitory, in fact a neurotransmitter is neither excitatory or inhibitory, because its action depends upon the type of receptor present and the state of the postsynaptic cell. Other well-characterized neurotransmitters include acetylcholine, glycine, and the biogenic amine transmitters serotonin, dopamine, and norepinephrine. These molecules have all been identified as neurotransmitters on the basis of well-established criteria, such as their presence in presynaptic terminals and their interaction with specific receptors. Many other neuroactive substances are also released from the synapse along with recognized neurotransmitters, such as the fatty acid arachadonic acid, ATP, and even gases such as nitric oxide. Many of these act as neuromodulators, in that they modify the response of the cell to a neurotransmitter. These substances are not necessarily contained in synaptic vesicles but may be released directly from the terminal.

The brain contains a large number of neuroactive peptides, short chains of amino acids, such as the enkephalins, substance P, and neurotensin, which also function as neurotransmitters and neuromodulators. Neuropeptides are contained in vesicles in the terminal, but in the large dense-core vesicles and not the small clear vesicles seen in Fig. 10. Unlike other neurotransmitters that are synthesized locally in the axon terminal via specific enzymes, peptides are synthesized in the cell body, packaged in the vesicles via the Golgi apparatus, and transmitted to the terminal. Peptides interact with specific receptors that activate second messenger molecules and, thus, tend to be associated with slow synaptic transmission. They are often colocalized and coreleased from the synaptic terminal with one of the better characterized transmitters. However, exocitosis from dense-core vesicles does not occur at the presynaptic active zone but can occur anywhere in the nerve terminal or along the axon. Release from dense-core vesicles requires a higher concentration of Ca^{2+} than does release from clear vesicles and generally occurs when the nerve is firing at high frequency. Thus, although both fast and slow neurotransmitters may be contained in the same nerve terminal, they may be released under different circumstances.

2. Synaptic Arrangements and Integration

The most common site of synaptic contact on a neuron is on the dendrites or dendritic spines. These types of synapses are referred to as axodendritic or axospinous. In cells that possess dendritic spines, the vast majority of synapses occur onto the spine heads, whereas few synapses are observed onto the dendritic shaft. In fact, it has been estimated that over 90% of excitatory synaptic inputs in the brain occur onto dendritic spines. Synapses also occur onto the cell soma and axon initial segment. In general, those synapses that occur onto the cell soma, initial segment, and proximal dendrites tend to be inhibitory in nature. For example, in the cerebellum, hippocampus, and cortex, an inhibitory interneuron termed the basket cell forms

pericellular baskets surrounding the somata of projection neurons and establishing numerous inhibitory synaptic contacts. Inputs more distally and onto dendritic spines tend to be excitatory. In some areas of the brain, most notably the olfactory bulb, dendrites are capable of forming synapses onto other dendrites. These special arrangements are called dendrodendritic synapses because a dendrite is serving as both the pre- and the postsynaptic element. Reciprocal synapses, where each dendrite serves as both the pre- and the postsynaptic cell, have also been described.

Whether a neuron will initiate an action potential in response to synaptic activation will depend upon the spatial and temporal integration of excitatory and inhibitory activation over the whole cell. Excitatory and inhibitory influences are summed over the entire dendritic tree and cell body and integrated in the soma. If threshold is reached, an action potential is initiated in the axon hillock, the spike initiation zone. Previously, it was thought that ionic spread from dendrites to soma occurred exclusively through passive diffusion. However, several ion channels have been localized in the dendritic plasmalemma and some dendrites are known to possess the ability to fire action potentials.

Individual neurons vary widely in the number of synapses they receive, averaging approximately 1000 per neuron. Generally, if a neuron receives a large number of synapses, each synapse is relatively weak and the simultaneous activity of multiple synapses is necessary to excite the postsynaptic cell. An example of a cell that receives large numbers of weak synapses is the Purkinje cell found in the cerebellum, which receives up to 150,000 synapses onto its dendritic spines from the cerebellar granule cell. Each granule cell exerts relatively little influence on the Purkinje cell, and so multiple granule cells must be activated to cause the Purkinje cell to fire an action potential. In contrast, other types of synapses are very powerful, such as the synapse formed at the neuromuscular junction where every action potential in the nerve fiber results in a corresponding action potential in the muscle cell. Individual neurons can receive a mixture of strong and weak synapses. The strength of an individual synapse will be determined principally by the amount and duration of transmitter released upon stimulation and also its location on the postsynaptic cell. Those synapses occurring closer to the cell soma tend to exert a stronger postsynaptic effect than those occurring more distally.

The large number of synapses received by an individual nerve cell indicates that there can be a tremendous convergence of information at the cellular level. Divergence of information also occurs, because a single axon can synapse with many cells either through the *en passant* type along the course of the axon or by ramification of synaptic terminals. In several regions of the brain, a specialized arrangement called a glomerulus (L: "small ball") is seen. The glomerulus consists of a very large presynaptic bouton, which contacts multiple postsynaptic dendrites. In the cerebellar mossy fiber glomerulus, the presynaptic terminal can measure 10–20 μm in length and can contact up to 50 dendrites. Glomerular arrangements are also seen in the thalamus, olfactory bulb, and several other brain regions.

V. SUMMARY

Neurons are more than globes filled with talented molecules…

Harris and Kater (1994)

The morphology and general characteristics of neurons have been established through pain-staking investigation over the last 100 years. Our understanding of neurons has evolved over the same period from ill-defined corpuscular globules consisting of a lump of protoplasm containing a nucleus to the exquisitely detailed, highly compartmentalized, information processing cellular machines of today. The major challenge facing neurocytologists is the specification of the molecular machinery and the organization of functional domains used by the neuron to accomplish its daunting task. Modern neuroscientists have the advantage of access to powerful computers and algorithms for modeling the behavior of individual neurons and neuron ensembles. As our understanding of neurons, glia, and their interactions increases, further breakthroughs in understanding the functioning of the nervous system in health and disease can be expected.

See Also the Following Articles

ACTION POTENTIAL • AXON • NERVOUS SYSTEM, ORGANIZATION OF • NEUROTRANSMITTERS • SYNAPSES AND SYNAPTIC TRANSMISSION AND INTEGRATION

Suggested Reading

Ellisman, M. H. (1987). Transcellular filament system. In *Encyclopedia of Neuroscience* (G. Adelman, Ed.), Vol. II, pp. 1232–1233. Birkhauser, Boston, MA.

Harris, K. M. (1999). Structure, development, and plasticity of dendritic spines. *Curre. Opin. Neurobiol.* **9,** 343–348.

Harris, K. M., and Kater, S. B. (1994). Dendritic spines: Cellular specializations imparting both stability and flexibility to synaptic function. *Annu. Rev. Neurosci.* **17,** 341–371.

Kandel, E. R., Schwartz, J. H., and Jessell, T. M. (2000). *Principles of Neural Science*, 4th ed. McGraw-Hill, Health Professions Division, New York.

Kennedy, M. B. (1997). The postsynaptic density at glutamatergic synapses. *Trends Neurosci.* **20,** 264–268.

Kennedy, M. B. (1998). Signal transduction molecules at the glutamatergic postsynaptic membrane. *Brain Res. Rev.* **26,** 243–257.

Levitan, I. R., and Kaczmarek, L. K. (1997). *The Neuron: Cell and Molecular Biology*, 2nd ed. Oxford University Press, New York.

Peters, A., Palay, S. L., and Webster, H. D. (1991). *The Fine Structure of the Nervous System : Neurons and Their Supporting Cells*, 3rd ed. Oxford University Press, New York.

Shepherd, G. M. (1998). *The Synaptic Organization of the Brain*, 4th ed. Oxford University Press, New York.

Shepherd, G. M. (1991). *Foundations of the Neuron Doctrine*. Oxford University Press, New York.

Neuropeptides and Islet Function

BO AHRÉN
Lund University

GLOSSARY

exocytosis Extrusion of secretory vesicle from cell to extracellular space.

islet The endocrine pancreas, consisting of cell islets scattered throughout the pancreas and comprising several different types of endocrine cells.

neuropeptides Peptides localized to nerve terminals, released when the nerves are activated and functioning as neurotransmitters.

secretory granules Cellular vesicles storing the secretory product of the cell.

This article summarizes the present day knowledge of the islet neuropeptides, which are of relevance for the regulation of islet function. Normal islet function is of vital importance for glucose and lipid homeostasis, because a main function of the islets is to secrete an optimal amount of insulin and glucagon to regulate metabolism. Important factors for the regulation of islet function therefore are glucose and free fatty acids. Circulating hormones, mainly gut hormones, also are of importance for normal islet function, notably after food intake, as are the various peptide hormones produced within the pancreatic islets, which affect the secretory rate of the other islet hormones. In addition, the autonomic islet nerves are involved in the regulation of islet function, and the central nervous system therefore is able to adjust the secretory rates of the islet hormones for the optimization of metabolism under various conditions. Effectors for this action are the neurotransmitters in the islet nerves, which besides acetylcholine and norepinephrine also include several different neuropeptides.

I. ISLET ANATOMY

The pancreatic islets are scattered throughout the entire pancreas, and each islet mainly consists of four different endocrine cells synthesizing and secreting the four islet hormones: insulin (β cells), glucagon (α cells), somatostatin (δ cells), and pancreatic polypeptide (PP, F cells) (See Table I for a list of abbreviations used throughout the text.) Microanatomical studies have shown that the β cells usually form the central portion of the islet structure, whereas the other cells form a mantle zone surrounding the β cells. The afferent vessels to the islets first reach the centrally located β cells. Blood vessels then pass through the central portion of the islets to reach the mantle zone, from which they leave the islets in efferent vessels. In addition, the islets are densely innervated, and parasympathetic, sympathetic, sensory, and other

Table I
Abbreviation List

AA	Arachidonic acid
ADP	Adenosine diphosphate
ATP	Adenosine triphosphate
CALCA	Calcitonin complex gene A
CALCB	Calcitonin complex gene B
CALCP	Calcitonin complex pseudo-gene
cAMP	Cyclic adenosine monophosphate
CCK	Cholecystokinin
CCK_A	CCK receptor A
CCK_B	CCK receptor B
CGRP	Calcitonin gene-related peptide
CPON	C-terminal flanking peptide of NPY
CRLR	Calcitonin receptor-like receptor
CTR	Calcitonin receptor
DAG	Diacylglycerol
GalR1	Galanin receptor subtype 1
GalR2	Galanin receptor subtype 2
GalR3	Galanin receptor subtype 3
GIP	Gastric inhibitory peptide
GLP-1	Glucagon-like peptide-1
GMAP	Galanin message-associated polypeptide
GRP	Gastrin releasing peptide
IAPP	Islet amyloid polypeptide
IP_3	Inositol 1,4,5-triphosphate
NMB	Neuromedin B
NPY	Neuropeptide Y
PAC_1	PACAP receptor subtype 1
PACAP	Pituitary adenylate cyclase activating polypeptide
PHM	Peptide histidine methionine
PKA	Protein kinase A
PKC	Protein kinase C
PLA_2	Phospholipase A_2
PLC	Phospholipase C
PLD	Phospholipase D
PP	Pancreatic polypeptide
PRP	PACAP-related peptide
PYY	Peptide YY
RAMP	Receptor-activity-modifying protein
VIP	Vasoactive intestinal peptide
$VPAC_1$	VIP PACAP receptor subtype 1
$VPAC_2$	VIP PACAP receptor subtype 2

autonomic nerves innervate the islet endocrine cells. These nerves harbor the classical neurotransmitters, acetylcholine and norepinephrine, and a number of neuropeptides.

II. ISLET FUNCTION

A main function of the pancreatic islets is to secrete a sufficient amount of insulin and glucagon to keep the circulating glucose level within its normal range. The regulation between the islet hormones and circulating glucose is a closed circuit in which glucose stimulates insulin secretion but inhibits glucagon secretion, whereas insulin reduces circulating glucose but glucagon increases circulating glucose (Fig. 1). This allows insulin to be responsible for protection from hyperglycemia, which is of primary importance after food intake when glucose levels increase due to nutritional ingestion and glucose absorption. This effect of insulin is achieved by augmenting the uptake of glucose in a variety of tissues, mainly skeletal muscle cells and adipocytes. Conversely, reduction of the glucose concentration to hypoglycemic levels elicits a stimulation of glucagon secretion, allowing glucagon to increase glucose levels through a stimulated hepatic glycogenolysis, which eventually restores normal glucose concentrations. Therefore, glucagon protects from hypoglycemia. The islet nerves are involved in the optimization of this tight control of circulating glucose, allowing the brain to react to observed or anticipated changes in circulating glucose and eliciting islet hormone secretion to be inhibited or stimulated to adjust glucose concentrations (Fig. 1). Malfunction of this optimization of islet function may lead to disturbed control of circulating glucose, which may be manifested as glucose intolerance after food intake or inadequate glucose recovery after hypoglycemia. The islet nerves and their neuropeptides therefore might be involved both in the normal control of islet function and glucose homeostasis and in the development of malregulation of circulating glucose.

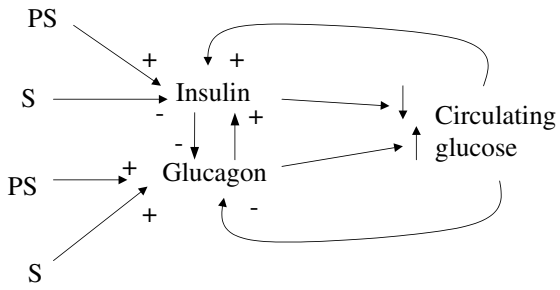


Figure 1 Schematic illustration of the influences of parasympathetic (PS) and sympathetic (S) nerves and glucose on insulin and glucagon secretion and influences of insulin and glucagon on the secretion of each other and on circulating glucose.

III. CELL BIOLOGY OF INSULIN SECRETION

The islet nerves and their neurotransmitters alter islet hormone secretion through surface-located receptors, which affect the cell biology of exocytosis of the islet hormones via signaling systems. Several transduction mechanisms are operative in β cells, transmitting signals for the stimulation or inhibition of insulin secretion. A primary mechanism underlying insulin secretion involves β cell metabolism of glucose, which increases the cytosolic ratio of ATP/ADP. This depolarizes the plasma membrane through the closure of ATP-regulated K^+ channels, which causes the opening of voltage-sensitive Ca^{2+} channels and the uptake of extracellular Ca^{2+} to raise the cytosolic concentration of Ca^{2+}. Another signaling pathway is initiated by phosphoinositide hydrolysis using phospholipase C (PLC) and phospholipase D (PLD), which produces diacylglycerol (DAG) and activates protein kinase C (PKC). PLC also induces the formation of inositol 1,4,5-triphosphate (IP_3), which stimulates the liberation of Ca^{2+} from intracellular storage sites. A third pathway involves the activation of adenylate cyclase, which produces cyclic AMP and activates protein kinase A (PKA). Finally, a fourth pathway involves the activation of phospholipase A_2 (PLA_2), which generates arachidonic acid (AA). These second messengers (cytosolic Ca^{2+}, PKC, PKA, and AA) then activate the exocytotic machinery, which releases insulin. The neurotransmitters affecting insulin secretion may execute their actions either by affecting any of these signaling mechanisms or by directly affecting the final exocytosis of insulin-storing granules. The mechanism of action of the neurotransmitters is different for each receptor activated.

IV. RELEVANCE OF THE AUTONOMIC INNERVATION OF THE PANCREATIC ISLET

The brain controls islet function through the autonomic nerves. Both parasympathetic and sympathetic nerves form a dense innervation of each islet, enabling release of their neurotransmitters in close apposition to the individual islet endocrine cells. A dense parasympathetic innervation has been verified by light and electron microscopy after staining sections with cholinesterase. Furthermore, histochemical data have shown a 10-fold higher concentration of choline acetyltransferase in islet tissue than in pancreatic tissue as a whole, which illustrates the high density of islet innervation. The parasympathetic nerve fibers innervating the islets originate in the intrapancreatic ganglia. These ganglia are in turn controlled by the preganglionic fibers, which emanate mainly from the dorsal motor nucleus of the vagus. The preganglionic fibers traverse through the vagus, both as part of the bulbar outflow tract and the hepatic and gastric branches of the vagus. They enter the pancreas along the vessels and terminate at intrapancreatic ganglia, from which the postganglionic nerves originate. These in turn penetrate the islets to terminate in close proximity to the endocrine cells. Activation of the parasympathetic nerves has been undertaken experimentally by electrical activation of the vagus, which stimulates the secretion of both insulin and glucagon. Therefore, the brain exhibits a potent autonomic system to stimulate the secretion of the two main glucoregulatory hormones from the islets.

Parasympathetic activation of islet hormone secretion is of importance in two conditions. First, it is involved in the cephalic phase of insulin secretion during food intake. This is the rapid and early increase in insulin secretion that is seen during the first 3–4 min after the initiation of food intake before any nutrient has reached the circulation to affect islet function. This phase is triggered by sensory mechanisms involving afferent pathways activated by olfactory, visual, gustatory, and oropharyngeal mechanical receptors. The afferent pathways are then integrated centrally, eliciting stimulation of the efferent pathways activating the secretion of insulin. A cephalic phase regulation of insulin secretion has been verified by demonstrations of insulin secretion (1) during sham feeding, i.e., feeding without food actually entering the gastrointestinal tract, (2) during imaginary food ingestion under hypnosis, (3) after the sight, smell, and expectation of food, and (4) after the ingestion of nonmetabolizable food. Furthermore, this early insulin response during food intake is also apparent when blood sampling is rapidly undertaken after food ingestion when an increase in circulating insulin is already evident before any significant increase in postabsorptive glycemia. Experimental studies have shown that the cephalic phase of insulin secretion is activated through the stimulation of oral taste receptors. It has also been demonstrated that the central integrative circuit is localized to the ventromedial hypothalamus and to the dorsal motor nucleus of the vagus and that the effector pathway is mediated by parasympathetic neurons within the vagus nerves. Consequently, the cephalic phase of insulin secretion is

abolished by vagotomy and by ganglionic blockade. Even though this cephalic phase of insulin secretion during food intake is short-lived and in magnitude corresponds to only 1–3% of the total insulin secretion after intake of a meal, it is of major importance for glucose tolerance after feeding. Thus, prevention or inhibition of the cephalic phase of insulin secretion results in glucose intolerance, and a brief insulin administration during inhibition of the cephalic phase restores glucose tolerance.

Another important function of the parasympathetic nerves is to stimulate glucagon secretion during hypoglycemia, which provides a mechanism for the recovery of circulating glucose. The protection against hypoglycemia is also achieved by other mechanisms, for example, the secretion of adrenaline and cortisol from the adrenals as well as direct stimulation of the release of glucose from the liver, but the primary involvement of glucagon in this respect has been established. The glucagon response to hypoglycemia is due to direct stimulation of glucagon secretion by the low glucose level and by the reduced intraislet concentration of insulin, and several studies have shown that the autonomic nerves also contribute to a major degree. Thus, ganglionic blockade is associated with more than 75% inhibition of the glucagon response to hypoglycemia in conjunction with lowered responses of markers for parasympathetic activation, as demonstrated both in humans and in experimental animals. In summary, the parasympathetic nerves are known to innervate the pancreatic islets, to stimulate insulin and glucagon secretion, and to be of importance for the cephalic phase of early insulin secretion during food intake and for the stimulated glucagon response to hypoglycemia.

The pancreatic islets are also innervated extensively by the sympathetic nerves, as demonstrated by fluorescence microscopy, electron microscopic studies, and immunocytochemistry. The islet sympathetic nerves are postganglionic with most nerve cell bodies located in the celiac ganglion or in the paravertebral sympathetic ganglia. The preganglionic nerve fibers originate from nerve cell bodies in the hypothalamus and leave the spinal cord at the level of C8–L3, whereafter the fibers traverse the lesser and greater splanchnic nerves to reach the paravertebral or celiac ganglia. From the ganglia, the nerves pass within the mixed autonomic nerves, which enter the pancreas along its vessels, although preganglionic sympathetic nerve fibers may also directly enter the pancreas. Activation of the sympathetic nerves has been undertaken by electrical activation of either the splanchnic nerves or the mixed autonomic nerves entering the pancreas after atropinization. This results in the inhibition of insulin secretion and stimulation of glucagon secretion. This activation is of marked importance during various forms of stresses or physical exercise, when hyperglycemia is achieved by the sympathetically induced inhibition of insulin secretion and stimulation of glucagon secretion. Therefore, the hyperglycemia and increased glucose turnover associated with these conditions are executed to a large degree by the islet sympathetic nerves.

In addition to innervation by parasympathetic and sympathetic nerves, the islets are also innervated by sensory nerves whose nerve terminals harbor the sensory neuropeptides, mainly calcitonin gene-related peptide (CGRP). These sensory islet nerves have been shown to innervate in particular the peripheral portions of the islets, although occasional nerve fibers are also found in the central part. The fibers leave the pancreas along the sympathetic fibers within the splanchnic nerves to reach the spinal cord. The role of the islet sensory nerves is not known. Rodents treated with capsaicin, which results in deafferentation of small unmyelinated C-fibers and leads to sensory denervation in conjunction with a substantial reduction in the number of islet CGRP nerves, have been shown to exhibit increased insulin secretion and glucose elimination. This suggests that the sensory nerves inhibit insulin secretion. This is in line with several studies under a variety of experimental conditions, demonstrating that exogenous administration of CGRP inhibits insulin secretion. Hence, it is possible that islet sensory nerves have an efferent function through the local liberation of CGRP, which inhibits insulin secretion, although the physiological role of this effect is not known.

V. EVIDENCE FOR NONCHOLINERGIC, NONADRENERGIC NEURAL ISLET EFFECTS

Because evidence for parasympathetic innervation of the islets is based on the localization of cholinesterase and because exogenous administration of acetylcholine or other muscarinic agonists mimics the stimulatory effects of vagus nerve activation on insulin and glucagon secretion, it may be anticipated that the parasympathetic effects on islet function are mediated by cholinergic mechanisms. However, several studies have shown that noncholinergic mechanisms also contribute to these effects, because it has been

demonstrated that vagally induced islet hormone secretion is not always sensitive to inhibition by muscarinic antagonism by atropine. For example, vagally induced insulin secretion in the pig pancreas is resistant to atropine, although it is inhibited by ganglionic blockade, and, furthermore, glucagon secretion during vagal nerve activation in the dog is resistant to atropine. In regard to sympathetically induced actions on insulin secretion, evidence exists that these are mediated by the classical neurotransmitter, norepinephrine. Thus, norepinephrine is both localized to islet sympathetic nerve terminals and released from the pancreas during electrical activation of these nerves. Furthermore, upon exogenous administration, norepinephrine inhibits glucose-stimulated insulin secretion. However, evidence also exists for the contribution of nonadrenergic mechanisms for sympathetically induced islet effects. Thus, combined and complete α_2- and β-adrenoceptor blockade by yohimbine and propranolol does not counteract the inhibition of basal insulin secretion induced by electrical sympathetic activation, as demonstrated in dogs. Furthermore, exogenous administration of norepinephrine does not mimic the inhibition of sympathetic nerve activation of basal insulin secretion. Moreover, immunoneutralization of one of the neuropeptides in the sympathetic nerves, galanin, prevents exercise stress from inhibiting glucose-stimulated insulin secretion, as demonstrated in mice. Therefore, these studies suggest that both noncholinergic and nonadrenergic mechanisms contribute to the islet actions of parasympathetic and sympathetic nerves. These actions, as the action of the sensory nerves, may be mediated by the neuropeptides localized to the islet nerve terminals.

VI. ISLET LOCALIZATION OF NEUROPEPTIDES

Immunocytochemistry has revealed three neuropeptides to presumably parasympathetic nerve terminals in pancreatic ganglia and in islets: gastrin releasing peptide (GRP), vasoactive intestinal peptide (VIP), and pituitary adenylate cyclase activating polypeptide (PACAP). These neuropeptides are released from the pancreas upon electrical vagal activation and they stimulate both insulin and glucagon secretion, as demonstrated in a number of studies *in vivo* as well as *in vitro*. From both a morphological and a functional point of view, they are therefore candidates to contribute to islet hormone secretion induced by vagal nerve activation. Similarly, two neuropeptides may mediate the nonadrenergic contribution to the inhibitory action of sympathetic nerve stimulation on insulin secretion: galanin and neuropeptide Y (NPY). These two neuropeptides are localized to the islet sympathetic nerve terminals and to nerve cell bodies in the celiac ganglion. Furthermore, galanin and NPY are also released from the pancreas upon electrical activation of the sympathetic nerves and inhibit insulin secretion, as demonstrated both *in vivo* and *in vitro*. In addition, CGRP has been shown to be localized to islet nerves that are sensitive to capsaicin and, therefore, presumably sensory nerve terminals. Finally, cholecystokinin (CCK), which was originally described as a gut hormone, also is localized to islet nerves. The nature of these latter nerves has, however, not yet been identified, and therefore they are referred to as "other" nerves. Hence, at least seven neuropeptides seem to be involved in the local regulation of islet function. Table II summarizes these neuropeptides, their presumed

Table II

Islet Neuropeptides, Islet Neuropeptide Receptors, Their Main Signaling Pathways, and Their Effects on Insulin and Glucagon Secretion

Neuropeptide	Receptor	Main signaling pathway	Effect on insulin secretion	Effect on glucagon secretion
GRP	GRPR	PLC–PLD	Stimulation	Stimulation
VIP	$VPAC_2$	cAMP	Stimulation	Stimulation
PACAP	PAC_1	cAMP	Stimulation	Stimulation
	$VPAC_2$	cAMP		
Galanin	GalR1	K^+–Ca^{2+}	Inhibition	Stimulation
NPY	Y1	K^+–Ca^{2+}, cAMP	Inhibition	Stimulation
CGRP	CRLR–RAMP	K^+–Ca^{2+}, cAMP	Inhibition	Stimulation
CCK	CCK_A	PLC–PLA_2	Stimulation	Stimulation

receptors and signaling pathways when affecting islet hormone secretion, and their islet effects.

VII. ISLET NEUROPEPTIDES

A. Gastrin Releasing Peptide (GRP)

GRP is the mammalian homolog of the amphibian peptide bombesin. Bombesin is a 14-amino acid peptide that was initially isolated from the skin of the frog *Bombina bombina* in 1971 by Anastasi and collaborators. Three different families of bombesin-like peptides exist in amphibians (bombesins, ranatensins, and phyllolitorins). These peptides are localized to myoepithelial poison glands that are under adrenergic control and therefore likely function in defence against predators. During the 1970s it became evident that bombesin-like peptides also exist in mammals, and in 1979 a mammalian homolog to bombesin was isolated from the porcine gastrointestinal tract by Dr. McDonald in the laboratory of Professor Mutt in Stockholm. Because this peptide was found to stimulate gastrin release from a porcine stomach model, it was called GRP. A few years later, a second bombesin-like peptide was isolated in mammals and called neuromedin B (NMB). These two peptides show a resemblance to different bombesin families. Thus, whereas GRP is structurally related to bombesin, NMB is structurally related to the ranatensins. GRP consists of a 27-amino acid residue that is α-amidated at its C-terminal methionine. The peptide was highly conserved during evolution, and the human and porcine forms of the peptide differ in only two residues (Fig. 2). The gene for GRP is located on chromosome 18q21 and consists of three exons (Fig. 2). The first exon encodes for the signal peptide and for the first 23 N-terminally located amino acids of GRP, whereas the second exon encodes for the remaining amino acids in GRP and for the first 74 amino acids in a C-terminal extension peptide of GRP. Finally, the third exon encodes for the remaining 21 amino acid residues in this long extension peptide. Due to alternative donor and acceptor sites in the second intron, three different GRP mRNAs are formed by differential splicing. These three GRP mRNAs therefore arise from the same gene and they code for three different proGRPs; however, all consist of 125 residues and differ only in their C-terminal parts. The proform of GRP contains, besides GRP (=proGRP$_{1–27}$), the C-terminal extension peptide of GRP (=proGRP$_{31–125}$). This extension peptide contains a 95-amino acid residue, but the

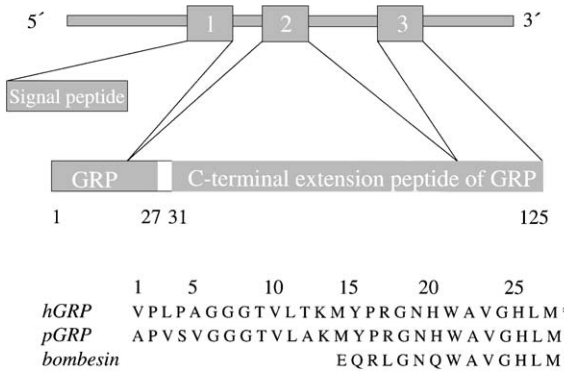


Figure 2 Schematic representation of the GRP gene and proGRP. The first exon encodes for the signal peptide and the N-terminal portion of the sequence of GRP, exon 2 encodes for the remaining portion of the GRP sequence and the major, N-terminal portion of the C-terminal extension peptide of GRP, and exon 3 encodes for the remaining sequence of this extension peptide. At the bottom of figure are the amino acid sequences of human and porcine GRP and bombesin. * indicates a C-terminal NH_2 group. The amino acids in this and the other figures are abbreviated according to the one-letter system: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; Y, Tyr.

amino acids are different depending on which mRNA is translated. This is in contrast to the GRP sequence, which is identical in all three GRP gene-derived mRNAs. proGRP$_{31–125}$ is posttranslationally modified to several different smaller peptides. GRP itself is also posttranslationally modified, and the major molecular forms in various tissues are GRP14, i.e., GRP27$_{1–14}$, and GRP27. The NMB gene is located on chromosome 15 and encodes for a 76-residue proNMB, which is processed to NMB (10 and 22 residues) and a 17-residue C-terminal extension peptide of NMB. It is thought that the GRP and NMB genes developed during evolution by duplication of the bombesin gene.

GRP is widely distributed in mammalian tissues, with the highest concentrations in the lung, central nervous system, and gut. In the peripheral nervous system, GRP is involved in the regulation of a variety of physiological processes, such as exocrine and endocrine secretions and smooth muscle contractions, and GRP is a powerful trophic agent as well. GRP is also localized to the brain with particular density in the hypothalamus, and centrally the peptide has been shown to be involved in the control of food intake and behavior. In the pancreas, GRP is localized to nerves with particular density in the ganglia, and GRP has also been shown to be released from the isolated pig pancreas when the attached vagal nerve is activated.

Furthermore, when GRP is exogenously added to different experimental models, both *in vivo* and *in vitro*, and to both the intact pancreas and isolated islets, the peptide efficiently stimulates both insulin and glucagon secretion. Because the pancreatic density of GRP localization is highest in the ganglia, it is possible that the peptide exerts actions both locally in the islets and also through ganglionic activation. Experimental support for such a notion was presented because ganglionic blockade inhibited the insulin response to exogenously added GRP in mice. Therefore, morphological and functional evidence suggests that GRP is a pancreatic parasympathetic neurotransmitter, the effect of which may contribute to islet actions of parasympathetic nerve stimulation, and this may be achieved both by a direct islet action and by an indirect ganglionic action.

The bombesin-like peptides, like GRP, are known to bind to G-protein-coupled receptors that are of the seven transmembrane domain type. Three receptors for the bombesin-like peptides have been cloned in mammals: the GRP receptor, the NMB receptor, and the bombesin receptor subtype 3. It is also possible, although not yet demonstrated, that bombesin receptor subtypes 2 and 4 might exist in mammals. However, in contrast to the GRP and NMB receptors, no endogenous ligand for these receptors has been found in mammals. The genes for both the GRP receptor and the bombesin receptor subtype 3 are localized to the X chromosome, whereas the gene for the NMB receptor is localized to chromosome 6. The chromosomal site for the GRP receptor is chromosome Xp21.2-p21.3, and this gene contains three exons. The receptor consists of 384 amino acids and shows a high degree of conservation during evolution, with homology between species of approximately 90%. Interestingly, the GRP receptor shows 60% homology with the NMB receptor. It has not yet been established whether the GRP receptors are expressed in the islet endocrine cells, although it is known that the islets are equipped with specific GRP binding sites. In islets, activation by GRP has been shown to be coupled mainly to PLC and PLD, and the islet action of GRP therefore is related to changes in cytoplasmic Ca^{2+}, formation of DAG, and activation of PKC. Studies have also demonstrated that GRP induces typical patterns of oscillation in cytosolic Ca^{2+}, which is due to both dynamic release of Ca^{2+} from intracellular stores and passage of Ca^{2+} through plasma membrane Ca^{2+} channels.

The relative importance of GRP for islet function has not yet been established. One study has used a specific GRP receptor antagonist, *N*-acetyl-$GRP_{20–26}$-amide, which is a fragment of GRP in which methionine has been deleted from position 27, to examine the potential involvement of GRP in islet function in mice. This antagonist has been used to inhibit the actions of GRP in a number of experimental systems and has also been shown to inhibit insulin secretion stimulated by exogenously added GRP in mice. However, the antagonist did not reduce islet hormone secretion induced by autonomic nerve activation, which suggests either that GRP does not contribute to islet function during autonomic nerve activation or that this antagonist is not a suitable tool for examining this contribution *in vivo*.

Another approach to the study of the physiological impact of GRP and GRP-related peptides is the use of receptor-deficient mice. Dr. Wada and collaborators in Japan have used genetic targeting to develop mice with deletion of either the bombesin receptor subtype 3 or the GRP receptor. Studies presented so far indicate that the mice with bombesin receptor subtype 3 deleted develop obesity. However, until the endogenous ligand for this receptor in mammals is found, the physiological relevance of this finding is not known. Furthermore, the GRP-receptor-deficient mice show impaired insulin response to gastric administration of glucose and to autonomic nerve activation, which suggests that GRP is involved in the parasympathetic regulation of islet function through activation of the GRP receptors. This would imply that GRP physiologically contributes to parasympathetically mediated insulin secretion during food intake, which supports a role for this neuropeptide in the regulation of islet function through the activation of GRP receptors.

B. Vasoactive Intestinal Peptide (VIP)

VIP was originally isolated from the porcine small intestine in the early 1970s by Dr. Said in the laboratory of Professor Mutt. The peptide has been shown to consist of a 28-amino acid residue with a C-terminal α-amidation. It shows structural similarities to the so-called glucagon family of peptides, i.e., glucagon, gastric inhibitory peptide (GIP), glucagon-like peptide-1 (GLP-1), and secretin. There is a high degree of identity in the VIP sequence between different species. For example, human, porcine, and rat VIPs are identical, differing from guinea pig VIP in only two amino acids. The human gene coding for VIP is located on chromosome 6q24 and contains seven

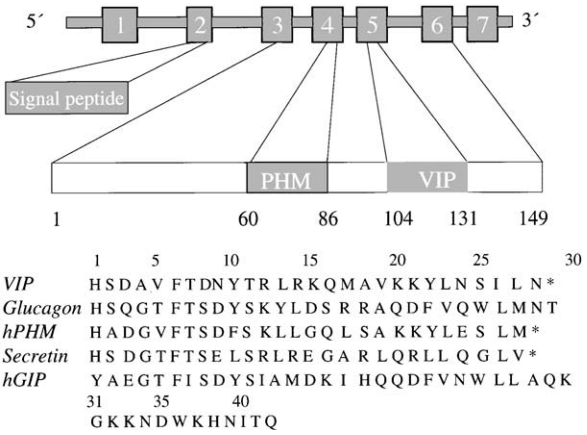


Figure 3 Schematic representation of the VIP gene and proVIP. The first exon is non-encoding, exon 2 encodes the signal peptide, exons 3–6 encode the proVIP sequence, and exon 7 is non-encoding. At the bottom of the figure are the amino acid sequences of VIP, glucagon, human PHM, secretin, and human GIP. * indicates a C-terminal NH_2 group.

exons. It encodes, besides for VIP, for peptide histidine methionine (PHM). This is a 27-residue peptide showing a high degree of homology to VIP because 15 of the amino acids are identical. VIP is encoded in exon 5, whereas PHM is encoded in exon 4 of the gene. Exon 2 encodes the signal peptide, which contains 21 residues, and proVIP is encoded in exons 3–6, which contains 149 residues, of which PHM is equivalent to $proVIP_{60–86}$ and VIP is equivalent to $proVIP_{104–131}$ (Fig. 3).

VIP is a neuropeptide exhibiting widespread distribution in the body. Thus, VIP nerves are localized to both the central nervous system and the respiratory, gastrointestinal, and genitourinary systems. In these systems, a variety of actions of VIP have been reported, such as relaxation of smooth muscle cells and stimulation of exocrine and endocrine secretions. A notably powerful action of VIP is its relaxation of smooth muscles in vessels, leading to vasodilatation. In the 1970s, it was also demonstrated that VIP occurs in the pancreas. Initially, the peptide was assumed to be an endocrine peptide in the islets, but it is now established that VIP is exclusively a neuropeptide localized to nerve terminals in the islets, ganglia, and exocrine pancreatic tissue. A neural VIP network surrounding the islets has also been demonstrated in both humans and experimental animals, suggesting important contributions of VIP to islet function. VIP may be involved in the local regulation of islet blood flow, because the peptide is a powerful vasodilatory agent and also has been demonstrated to increase islet blood flow upon exogenous administration. VIP may also be involved in the local regulation of islet hormone secretion. Thus, VIP is released from the pancreas when the vagal nerves are activated, and, furthermore, VIP is a powerful stimulator of both insulin and glucagon secretion under a number of conditions.

Two different VIP binding receptors have been cloned. They both show affinity for PACAP and are called the $VPAC_1$ and $VPAC_2$ receptors, respectively. They are both of the seven transmembranous domain type and they are G-protein-coupled. The $VPAC_1$ receptor consists of 460 residues and the $VPAC_2$ receptor consists of 438 residues. $VPAC_1$ is encoded by chromosome 3p22, whereas $VPAC_2$ is encoded by chromosome 7q36.3. Both of these genes consist of 13 exons. Activation of the VIP receptors is followed by activation of adenylate cyclase with the formation of cAMP, as has been demonstrated in islet cells. Cyclic AMP in turn activates PKA, which enhances exocytosis. Hence, the stimulation of insulin and glucagon secretion by VIP seems to be a classical cAMP–PKA-mediated process.

The relative contribution of VIP to islet function has not been established. Changing the VIP structure by substituting methionine in position 17 for leucine and inserting a 4-chloro substitution on phenylalanine in position 6 results in a VIP antagonist, [4-Cl-D-Phe^6, Leu^{17}]VIP, which inhibits islet hormone secretion after exogenous administration of VIP. However, this antagonist was without effect on the insulin or glucagon response to autonomic nerve activation in an *in vivo* study of the mouse. Therefore, further studies are required to examine the role of VIP in the regulation of islet function.

The VIP gene product, PHM, may also be of potential importance in the regulation of islet function, because this peptide, like VIP, stimulates insulin and glucagon secretion. However, much less is known about this peptide. For example, the receptors mediating its actions and its signaling pathways have not been established.

C. Pituitary Adenylate Cyclase Activating Polypeptide (PACAP)

PACAP was originally extracted from the pituitary gland in 1989 by Dr. Arimura and collaborators. Two forms of PACAP exist consisting of 27 and 38 amino acids, respectively, and PACAP, like VIP, belongs to the glucagon superfamily of peptides. In fact, PACAP

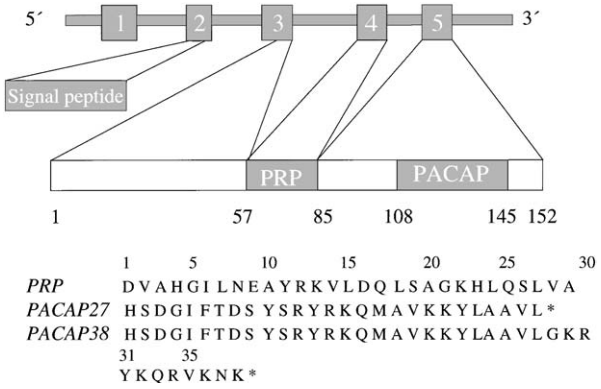


Figure 4 Schematic representation of the PACAP gene and proPACAP. The first exon is non-encoding, exon 2 encodes the signal peptide, and exons 3–5 encode the proPACAP sequence. At the bottom of the figure are the amino acid sequences of PACAP27, PACAP38, and PRP (PACAP–related peptide). * indicates a C-terminal NH_2 group.

shows 68% identity to VIP. The gene for PACAP is located on chromosome 18p11 and consists of five exons. The entire PACAP sequence is included in the fifth exon. Transcription and translation of the second exon yield a 24-residue signal peptide, whereas the fifth exon encodes the PACAP precursor, which is 152 residues long. PACAP38 is equivalent to proPACAP$_{108–145}$ (Fig. 4). PACAP27 is then cleaved from PACAP38, consisting of its 27 N-terminal amino acids (PACAP38$_{1–27}$). In various tissues, including the pancreas, the main PACAP form is PACAP38. In addition, a PACAP-related peptide (PRP) is formed from proPACAP$_{57–85}$. Although this peptide, like PACAP, is distributed in a variety of organs and its structure shows a 44% resemblance to PHM, its function is still not known.

Like VIP, PACAP is also distributed ubiquitously and occurs mainly in nerves in the central nervous system, the lungs, and the gastrointestinal tract. PACAP has been demonstrated to exert a multitude of effects in these organs, such as relaxation of smooth muscle cells and stimulation of secretion from the pituitary and adrenal glands. In the central nervous system, a neurotrophic action of the peptide has also been reported. The neuropeptide is located in islet nerve terminals and in pancreatic ganglia, and because it is released from the pancreas during electrical activation of the vagal nerves, it is thought to be mainly a parasympathetic neuropeptide. It has, however, also been reported that PACAP is a neurotransmitter in sensory nerves. Finally, in one study PACAP has also been reported to be expressed not only in pancreatic nerves but also in islet β cells, although this has not been confirmed in other studies.

Like parasympathetic nerve activation, PACAP potently stimulates the secretion of both insulin and glucagon, which has been demonstrated both *in vivo* and *in vitro* in a number of experimental conditions. Like VIP, PACAP stimulates insulin secretion in a glucose-dependent manner accompanied by increased action of adenylate cyclase with increased formation of cAMP and activation of PKA. PACAP also exerts, however, other signaling actions in insulin producing tissues, like an increase in the cytoplasmic concentrations of both Ca^{2+} and Na^{+} and a distal effect on the exocytosis machinery. It is possible that the potency of PACAP to stimulate insulin secretion, which exceeds that of VIP, is due to the diversity of signaling actions induced by PACAP in islet β cells.

The close relation between PACAP and VIP is evidenced by the structural similarity of the peptides and by the affinity of both peptides to the $VPAC_1$ and $VPAC_2$ receptors. However, the findings that PACAP is more potent in stimulating insulin secretion than VIP and that signaling pathways other than the cAMP–PKA pathway also seem to be activated by PACAP suggest the existence of a third PACAP receptor, which is specific for PACAP. Such a receptor has also been cloned and named the PAC_1 receptor, and it is, like the two VPAC receptors, of the seven transmembranous domain type and G-protein-coupled. The PAC_1 receptor gene is located on chromosome 7p15-p14, and the receptor consists of 525 amino acids. The PAC_1 receptor is further divided into eight subtypes, all resulting from alternative splicing. The function of this splicing is not known, although different affinities for various PACAP forms have been documented for the different splice variants. In pancreatic islets, the occurrence of PAC_1 and $VPAC_2$ receptors has been verified.

The functional relevance of the PACAP nerves for islet function has started to emerge. One study used a PACAP antagonist, the 22 C-terminally located residues in PACAP27, i.e., PACAP$_{6–27}$, and administered it to mice. It was first demonstrated that the antagonist indeed is a PACAP antagonist, because insulin secretion stimulated by exogenously administered PACAP27 was inhibited by the antagonist. It was also found that the antagonist was without effect on the insulin response to exogenous VIP, suggesting that the antagonist is specific for PAC_1 receptors. When this antagonist was administered to mice challenged with a gastric administration of glucose, the resulting insulin response was impaired. This suggests that PACAP is

involved in the neural mediation of the insulin response to food intake. Similar results were also shown in a mouse model from which the PAC_1 receptor had been deleted. The model was developed by Dr. Brabet and collaborators in Montpellier by genetic targeting, and a gastric challenge with glucose was found to display a reduced insulin response in these mice compared to wild-type controls. Furthermore, studies undertaken in isolated islets have demonstrated that specific PACAP antisera not only reduce the insulin response to PACAP but also reduce the insulin response to glucose alone. Because this was not observed in islets that had been cultured for 48 hr and therefore were devoid of any remaining PACAP in surviving islet nerve terminals, it is likely that the insulinotropic response to glucose is dependent on release of PACAP and intact PACAP receptors. Similar results with reduced glucose-stimulated insulin secretion were also reported in the PAC_1-receptor-deleted mouse model. Hence, PACAP seems to be an islet neuropeptide of importance for islet function, notably after activation by glucose and postprandially.

D. Galanin

Galanin was isolated from the porcine gastrointestinal tract by Dr. Tatemoto in the laboratory of Professor Mutt in 1983. It was given its name by its glycine C-terminus and alanine N-terminus. It is a highly conserved peptide consisting of 29 amino acids in most species and C-terminally amidated. However, the human form of galanin consists of a 30-amino acid residue that is not C-terminally amidated. Whereas the C-terminal amino acid of galanin in the pig is alanine, it is threonine in rat galanin and serine in human galanin. Nevertheless, the name galanin has been retained for the peptide in rats and humans. The human galanin gene is located on chromosome 11q13.3-q13.5 and consists of six exons. The first exon is noncoding, whereas the coding region for progalanin consists of five exons (Fig. 5). Progalanin is a 123-residue peptide, which includes the signal peptide (23 amino acids), galanin (30 amino acids), and galanin message-associated peptide (GMAP, 61 amino acids). Posttranslational modifications yield the 91-amino acid progalanin consisting of galanin and GMAP. The N-terminal amino acids 1–13 in the galanin molecule are encoded in exon 3 and the remaining 17 amino acids in exon 4, which is of interest because it is the N-terminal end that is biologically active.

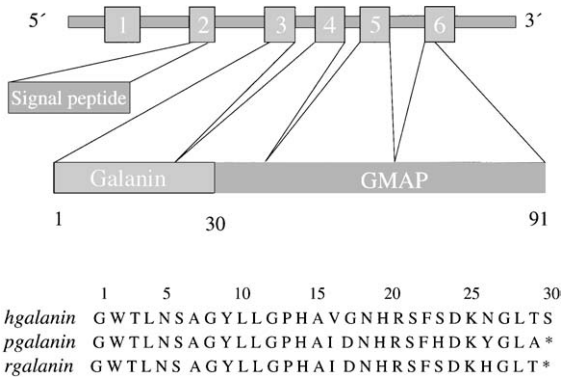


Figure 5 Schematic representation of the galanin gene and progalanin. The first exon is non-encoding, exon 2 encodes the signal peptide, exons 3–5 encode for the sequence corresponding to galanin, and exons 5 and 6 encode for the sequence corresponding to GMAP (galanin message-associated peptide). At the bottom of the figure are the amino acid sequences of human (h), porcine (p), and rat (r) galanin. * indicates a C-terminal NH_2 group.

Galanin is widely distributed in the peripheral nervous system, in both the sympathetic and parasympathetic nerves. It exerts a wide spectrum of actions, such as stimulatory or inhibitory influences on smooth muscle cells depending on location and inhibition of secretion of gastrointestinal hormones and gastric acid. Galanin is also widely distributed in the central nervous system with major localization to the nuclei of the septum–basal forebrain complex, the hypothalamus, and the dorsal raphe nucleus. Centrally, a main function of galanin is its stimulation of feeding, as has been demonstrated in rats. In the pancreas, a dense galanin innervation of the islets was first demonstrated in dogs. These nerves were found to be sympathetic, because galanin is colocalized with tyrosine hydroxylase in nerve terminals and because galanin occurs in nerve cell bodies in the celiac ganglion. In other species, islets are innervated by galanin nerves, although with a lesser density than in the dog. Galanin has been demonstrated to potently inhibit insulin secretion and to stimulate glucagon secretion in a number of experimental systems both *in vivo* and *in vitro*. The potential effect of GMAP on insulin secretion has also been examined in one study in isolated islets. However, the peptide had no influence over a wide dose range and at different glucose concentrations.

Three galanin receptors have been cloned and called GalR1, GalR2, and GalR3, respectively. They are all encoded by different genes located on different chromosomes (GalR1 on 18q23, GalR2 on 17q25.3, and

GalR3 on 22q13.1), and they are all G-protein-coupled receptors. The structural organization of the genes encoding for GalR2 and GalR3 are conserved during evolution, suggesting a common evolutionary origin, whereas the structure of the GalR1 gene is unique among the G-protein-coupled receptors, the relevance of which is not yet known. It is this galanin receptor subtype, GalR1, that has been shown to be expressed in insulin producing cells. Whereas the structure of the GalR1 gene is different between species, the structure of the receptor itself shows high conservation during evolution, and human (349 amino acids) and mouse (348 amino acids) GalR1 displays 93% identity. The powerful inhibitory influence of galanin in rodent islets has been shown to be accompanied by a complex signaling mechanism involving hyperpolarization due to the opening of K^+ channels and a concomitant reduction in the cytosolic concentration of Ca^{2+}, although reduced formation of cAMP and inhibition of the exocytotic mechanism by a direct effect on the exocytosis machinery may also contribute.

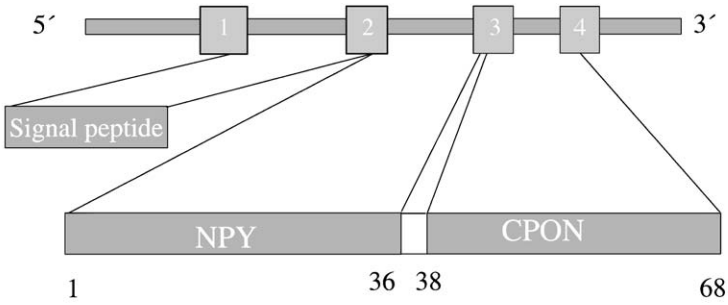
Functional studies have indicated that galanin contributes to the sympathetically induced inhibition of insulin secretion. Thus, it has been demonstrated that galanin is released from the dog pancreas during sympathetic nerve activation and that the amount of galanin released under these conditions is sufficient to mimic the inhibition of insulin secretion induced by sympathetic nerve stimulation. The physiology of galanin has also been studied in a physiological model of swimming mice, in which swimming for 2 min is accompanied by a 50% inhibition of glucose-stimulated insulin secretion as a sign of the stress associated with the physical exercise. When galanin was immunoneutralized in these mice by pretreatment with a high-titer galanin antiserum, the impairment of glucose-stimulated insulin secretion during the swimming was abolished. This suggests that galanin released from the sympathetic nerves during the swimming contributes to the inhibition of insulin secretion. Studies to conclusively establish this remain to be performed, however. One approach might be the use of selective GalR1 antagonists, a few of which have been reported, like galantide. Another approach would be the use of galanin- or GalR1-deficient mice. It must be emphasized, however, that species differences seem to be of particular relevance regarding the role of galanin in islet function. For example, whereas it is established that extensive galanin innervation exists in dog islets, there is only scanty innervation in rat or human islets. Furthermore, whereas potent inhibition by galanin of insulin secretion has been reported in dogs, no such effect is evident in humans. In addition, in the pig galanin has been shown to stimulate, not inhibit, insulin secretion. Therefore, the involvement of galanin in islet physiology remains to be established.

E. Neuropeptide Y (NPY)

NPY was isolated from the porcine brain by using an assay developed by Dr. Tatemoto and Professor Mutt in 1982 to detect C-terminally α-amidated peptides. Because the isolated peptide was found to consist of tyrosine in both its N- and C-terminal ends, it was named NPY. The peptide consists of a 36-amino acid residue and shows a high structural homology to peptides YY (PYY) and PP. The human NPY gene is located on chromosome 7p15.11, consists of four exons, and encodes for the 29-residue signal peptide and the 68-residue proNPY. proNPY is further processed to NPY at its N-terminal end and to the 30-residue peptide, CPON (C-terminal flanking peptide of NPY), at its C-terminal end (Fig. 6). Both NPY and CPON are highly conserved between species. Thus, NPY in humans and rats has an identical sequence, differing from porcine NPY in only one amino acid (position 17 is a lysine residue instead of methionine). Furthermore, human and rat CPON shows only two amino acid differences (position 19 being arginine and position 28 being serine in rat CPON).

NPY is a widely distributed neuropeptide both in the brain and in peripheral tissues, with primary localization to adrenergic neurons. In the central nervous system, localization to both the cerebral cortex and the forebrain has been reported, and a particularly high content of NPY has been observed in specific nuclei in the hypothalamus, mainly in the paraventricular nucleus and the arcuate nucleus. In these locations, the peptide seems to be of major importance for the regulation of feeding behavior and energy balance, and it is of great importance that its expression is inhibited by leptin. In the peripheral nervous system, NPY exists in almost all organs, particularly in perivascular adrenergic nerve fibers, and a major function of the neuropeptide is its involvement in the local regulation of blood flow. NPY also affects smooth muscle contractions in other locations and may be involved in the regulation of the gastrointestinal tract and the genitourinary system.

In the pancreas, NPY containing neurons are richly distributed around vessels in the exocrine portion of



	1 5 10 15 20 25 30 35
hNPY	Y P S K P D N P G E D A P A E D M A R Y Y S A L R H Y I N L I T R Q R Y*
hPP	A P L E P V Y P G D N A T P E Q M A Q Y A A D L R R Y I N M L T R P R Y*
hPYY	Y P I K P E A P G E D A S P E E L N R Y Y A S L R H Y L N L V T R Q R Y*
hCPON	S S P E T L I S D L L M R E S T E N V P R T R L E D P A M W

Figure 6 Schematic representation of the NPY gene and proNPY. Exons 1 and 2 encode the signal peptide and exons 2–4 encode the proNPY. At the bottom of the figure are the amino acid sequences of human VIP, human PP, human PYY, and human CPON. * indicates a C-terminal NH_2 group.

the islets, and NPY nerves also innervate the islets. Because NPY is colocalized with tyrosine hydroxylase both in islet nerves as well as in nerve cell bodies in the celiac ganglion, it is supposed that NPY is an islet sympathetic neuropeptide. This is supported by findings that chemical sympathectomy by 6-hydroxydopamine substantially reduces the pancreatic NPY innervation in islets. However, NPY has also been shown to be colocalized with VIP in islet nerves in the pig, suggesting that nonadrenergic NPY nerves might also exist. In addition, under some conditions, as in insulin resistance, islet β cells also seem to express NPY. NPY therefore might function both as a sympathetic and as a parasympathetic neuropeptide and as a local autocrine islet peptide. A main function is, however, the role of NPY as a sympathetic neurotransmitter.

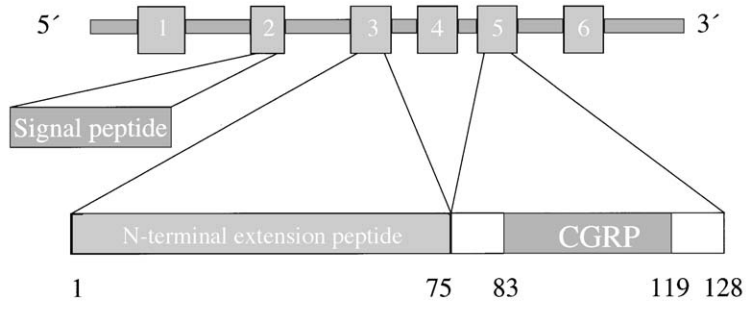
NPY has been shown to inhibit glucose-stimulated insulin secretion, i.e., to exert a sympathetic-like effect, in several studies both *in vivo* and *in vitro*. Six different NPY receptors have been cloned (Y1, Y2, Y3, Y4, Y5, and Y6 receptors). Some of these receptors show affinity for PYY and/or PP. These receptors are all of the seven transmembranous domain, G-protein-coupled type but show differences in localization and characteristics with regard to the spectra of affinity for the peptides in the NPY–PP–PYY family. In insulin producing cells, most evidence shows that it is the Y1 receptor subtype that mediates islet actions of NPY. Thus, analogs of PYY that are specific for Y1 receptors mimic the effects of NPY, and a specific Y1 receptor antagonist prevents NPY from inhibiting insulin secretion. The gene for the human Y1 receptor is located at chromosome 4q31.3-32, and the full-length receptor consists of a 384-amino acid residue. The human and mouse forms of the receptor show 93% identity, suggesting a high degree of conservation of the Y1 receptor during evolution. Studies have indicated that two different isoforms of the Y1 receptor exist, called the Y1α and Y1β receptors, respectively. They show a high degree of identity, differing in only seven amino acids, and they are encoded by the same gene and are generated by alternative splicing. The Y1β receptor seems, however, to be expressed mainly during embryogenesis and therefore might not be of relevance after birth.

The mechanism underlying the inhibition by NPY of glucose-stimulated insulin secretion is not fully established but seems to involve inhibition of adenylate cyclase with reduced formation of cAMP. However, distal mechanisms in relation to the formation of cAMP also seem to be involved because insulin secretion stimulated by a cAMP analog is inhibited by NPY, as demonstrated in a perfused rat pancreas model. The physiological importance of NPY for islet function remains to be established. One study has demonstrated that immunoneutralization of NPY in isolated islets leads to increased insulin secretion, which would suggest a physiological inhibitory effect of islet NPY on insulin secretion. However, it is not clear whether this is exerted through its neural

localization or through its expression in islet β cells, which occurs under some conditions. Furthermore, results from studies using a Y1-receptor-specific antagonist or NPY-deficient mice have found no effect on glucose and insulin levels, and the NPY-deficient mice do not develop diabetes or other abnormalities in glucose homeostasis. However, the issue of whether NPY contributes to sympthetically induced regulation of islet function has not been examined in these mice. Therefore, the role of NPY in the regulation of islet function remains to be established.

F. Calcitonin Gene-Related Polypeptide (CGRP)

The localization of the 37-amino acid peptide CGRP to islet nerves was first reported by Rosenfeld and collaborators in 1983 and later confirmed in other laboratories. CGRP was originally described in a medullary thyroid carcinoma cell line, where it was found to be encoded in the same gene as calcitonin. This gene is called the calcitonin complex gene A (CALCA), is located on chromosome 11p15.2-p15.1, and consists of six exons. This gene encodes for two different mRNAs, one being translated to preprocalcitonin and the other to preproCGRP. The N-terminal portions of procalcitonin and proCGRP, consisting of 75 amino acids, are identical in the two prohormones and encoded in exon 3. The remaining part of the 141-amino acid preprocalcitonin is encoded in exon 4, whereas the rest of the 128-amino acid preproCGRP is encoded in exon 5 (Fig. 7). Hence, by alternative splicing, CALCA will be transcribed and translated to either procalcitonin or proCGRP. This alternative splicing has turned out to be tissue-specific. Later, an additional gene that also encodes for CGRP was described and named calcitonin complex gene B (CALCB). This gene is located at chromosome 11p12-14.2 and encodes for a form of CGRP, which is called β-CGRP and shows a high degree of similarity to the α-CGRP that is encoded in CALCA. In addition, two other genes have been grouped together with CALCA and CALCB to a calcitonin gene family. These other genes are the calcitonin complex pseudogene (CALCP), the function of which is yet unknown, and the islet amyloid polypeptide (IAPP) gene, located at chromosome 12p13.2-12.1 and encoding for IAPP. CGRP consists of 37 amino acids with a high degree of identity between species. The two forms of CGRP, α and β, differ in three amino acids in humans and in one amino acid in rats. CGRP is also structurally related to IAPP, calcitonin, and adrenomedullin. CGRP has been demonstrated to be a ubiquitously distributed neuropeptide with localization in both the central nervous system and peripherally. It seems to be mainly localized to sensory nerves. A multitude of biological



	1 5 10 15 20 25 30 35
hαCGRP	A C D T A T C V T H R L A G L L S R S G G V V K N N F V P T N V G S K A F *
hβCGRP	A C N T A T C V T H R L A G L L S R S G G M V K S N F V P T N V G S K A F *
hIAPP	K C N T A T C A T Q R L A N F L V H S S N N F G A I L S S T N V G S N T Y *
hcalcitonin	C G N L S T C M L G T Y T Q D F N K F H T F P Q T A I G V G A P *

Figure 7 Schematic representation of the calcitonin complex gene A (CALCA) and proCGRP. The first exon is non-encoding, exon 2 encodes the signal peptide, exon 3 encodes the N-terminal extension peptide of proCGRP, which is identical to the corresponding N-terminal extension peptide of procalcitonin, exon 4 encodes for the remaining portion of procalcitonin, exon 5 encodes the remaining portion of proCGRP, and exon 6 is noncoding. At the bottom of the figure are the amino acid sequences of human α- and β-CGRP, human IAPP, and human calcitonin. * indicates a C-terminal NH_2 group.

actions of CGRP have been demonstrated in a variety of experimental conditions, and these include the induction of vasodilatation, increase in tissue blood flow, activation of adrenergic nerve outflow, inhibition of gastric acid and thyroid hormone secretion, and stimulation of exocrine pancreatic secretion.

CGRP nerves are also localized to the pancreas, where they are scattered throughout the parenchyma with particular density along small blood vessels and within the islets. Exogenous administration of the peptide under various experimental conditions leads to the inhibition of insulin secretion and stimulation of glucagon secretion. The physiological relevance of the CGRP nerves has been studied in animals treated with capsaicin, which destroys C-fibers and is followed by marked reduction of the CGRP nerves in the pancreatic islets. In capsaicin-treated mice, the glucagon response to neuroglycopenia induced by the glucose analog, 2-deoxyglucose (2-DG) is inhibited and the glucose recovery from insulin-induced hypoglycemia is impaired, suggesting a stimulatory role of the sensory fibers on glucagon secretion. Furthermore, insulin secretion is augmented after capsaicin, suggesting a tonic inhibitory influence of the sensory nerves on the β cells. These studies therefore suggest a physiological importance for CGRP nerves in the regulation of both insulin and glucagon secretion.

CGRP acts through the calcitonin receptor-like receptor (CRLR) or the calcitonin receptor (CTR). Both of these receptors show affinity for IAPP, calcitonin, and adrenomedullin. Three different receptor-activity-modifying proteins (RAMPs) seem to govern the specificity of the receptors for these ligands. For example, elegant studies by Sexton and collaborators in Melbourne have shown that the RAMP1-transported CRLR is a CGRP receptor, whereas RAMP2- or RAMP3-transported CRLR is an adrenomedullin receptor. They also demonstrated that IAPP binding is more related to RAMP-transported CTR. These findings suggest an interesting mode of regulation of cellular activity. Of particular interest for islet physiology is the relation between CGRP and IAPP, both in terms of sequence similarity and in terms of similarities in receptors activated by the peptides. IAPP is a β-cell peptide with a potential role in diabetes pathophysiology because it both inhibits insulin secretion and forms amyloid fibrils, which are common in islets of subjects with type 2 diabetes. It should also be emphasized that, in some species, CGRP seems to be expressed in the islet endocrine cells. Thus, in rat islets, CGRP is found in the D cells. The role of islet cellular CGRP is not known.

G. Cholecystokinin (CCK)

It has been demonstrated that nerve fibers that are not strictly related to the sympathetic, parasympathetic, or sensory nervous system innervate the endocrine pancreas. These nerve fibers have been shown to harbor different neurotransmitters from the other types of autonomic nerve terminals. For example, nitric oxide synthase and cholecystokinin (CCK) have been localized to these nerve terminals, suggesting that nitric oxide and CCK are islet neurotransmitters. CCK was initially demonstrated in 1928 by Ivy and Oldberg as a gut hormone released into the circulation, stimulating gall bladder contraction. It is known today that the peptide is produced in the intestinal I cells, which are situated mainly along the mucosa of the duodenum and the proximal jejunum. Besides its main actions to stimulate gall bladder contraction and exocrine pancreatic secretion, CCK also inhibits gastric emptying and intestinal motility. During the 1970s and 1980s it was demonstrated that CCK is a neuropeptide with localization to both central and peripheral neurons. An interesting central action of CCK is inhibition of food intake. In 1980 it was reported by Dr. Rehfeld and collaborators in Copenhagen that CCK is a neuropeptide in pancreatic islets. The human gene for CCK is located on chromosome 3q12-3pter and consists of three exons. It encodes for a proCCK with a 115-amino acid residue including the 20-amino acid signal peptide, which is posttranslationally modified to generate several different forms of CCK (Fig. 8). The

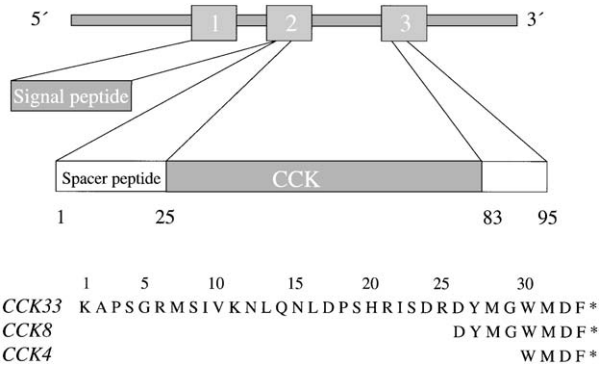


Figure 8 Schematic representation of the CCK gene. The first exon and the 5′-end of the second exon encode the signal peptide. Exon 2 also encodes a 20-amino acid spacer peptide in proCCK and, together with the 5′-end of exon 3, the remaining portion of proCCK. At the bottom of the figure are the amino acid sequences of CCK33 ($CCK58_{26–58}$), CCK8 ($CCK58_{51–58}$), and CCK4 ($CCK58_{55–58}$). * indicates a C-terminal NH_2 group. Observe that the Y residue in position 27 (in both CCK33 and CCK8) is sulfated.

initial translational product is CCK83, which is further processed by the removal of a 25-amino acid spacer peptide at its N-terminal end, yielding CCK58, the initial stable product of the gene. CCK58 is further processed by N-terminal truncation to CCK variants with 39, 33, 25, 22, 18, 12, 8, 7, 5, and 4 residues, respectively. The processing mechanisms and the function of all of these intermediates or CCK forms are still not established. However, CCK33 and CCK8 appear to be the main forms of CCK in circulation, whereas in the neurons most evidence favors the smaller CCK forms, like CCK4, as the main products that therefore probably function as neurotransmitters. The processing of proCCK in the nerves has not been examined in such detail as in the gut, although it has been revealed that it is the same CCK gene that is expressed in nerve cell bodies as in the I cells.

In the pancreatic islets, CCK has been shown to potently stimulate insulin secretion, and this has been demonstrated under a variety of experimental conditions. This action may be of physiological importance after food intake, when the circulating levels of CCK are increased. It is unlikely, however, that CCK is of physiological importance after food intake, because the circulating levels of CCK achieved after food intake are lower than those required for stimulation of insulin secretion and because inhibition of CCK release or action does not affect meal-induced insulin secretion in humans. Therefore, the main function of CCK in relation to the pancreatic islets is probably as a neurotransmitter. However, the role of these "other" nerves in the regulation of insulin secretion, particularly in relation to the influence of the three main branches, is not known. CCK has also been shown to stimulate glucagon secretion, although this has not been studied in such great detail as its effects on insulin secretion.

Two different types of CCK receptors have been described, both of which are of the G-protein-coupled, seven transmembranous domain type linked to the activation of PLC. The CCK_A receptor gene is located on chromosome 4p15.1-p15.2, whereas the CCK_B receptor gene is located on chromosome 11p15.4-p15.5. The CCK_A receptor type consists of 428 amino acids and is expressed in the gall bladder smooth muscle cells, pancreatic acinar cells, gastrointestinal muscular cells, and various parts of the brain. This receptor type shows an almost 1000-fold higher affinity for CCK than for gastrin. The CCK_B receptor type consists of 447 amino acids and is expressed in the cerebral cortex and other brain areas, gastric parietal cells, ECL cells, and gastrointestinal muscular cells. This receptor type shows equal affinity for CCK and gastrin. In the islets, most studies favor the CCK_A receptor subtype as the one that mediates the actions of CCK.

The insulinotropic activation of CCK is mainly mediated by PLC, which stimulates the hydrolysis of phosphoinositides. Phosphoinositide hydrolysis in turn yields the formation of IP_3, releasing Ca^{2+} from intracellular Ca^{2+} stores to increase the cytosolic concentration of Ca^{2+}. Consequently, CCK increases the cytosolic concentration of Ca^{2+} independently from the uptake of extracellular Ca^{2+}, which is accompanied by rapid stimulation of exocytosis and insulin secretion. CCK has also been shown to stimulate the generation of arachidonic acid (AA) through the activation of PLA_2. AA in turn stimulates the exocytosis of insulin through several mechanisms. In contrast, CCK does not seem to increase the islet formation of cAMP. The combined action by CCK to release Ca^{2+} from intracellular stores to increase the cytosolic concentration of Ca^{2+} and to stimulate the formation of AA and activate PKC seems to contribute to the insulinotropic action after activation of the CCK_A receptors.

VIII. CONCLUSION

It is known that the pancreatic islets are innervated by the autonomic nerves. The net effects of activation of the nerves on islet hormone secretion, the presumed neurotransmitters mediating these effects, and the putative receptors and their signaling pathways have also been established. In particular, several neuropeptides have been localized to islet nerve terminals, and their actions on islet hormone secretion have been documented under a variety of experimental conditions. Studies are now required to establish the physiology and pathophysiology of the islet nerves and their neurotransmitters, particularly concerning the possible involvement of the nerves in the pathophysiology of glucose intolerance, type 2 diabetes, and obesity.

See Also the Following Articles

HOMEOSTATIC MECHANISMS • NEUROTRANSMITTERS • NOREPINEPHRINE • PEPTIDES, HORMONES, AND THE BRAIN AND SPINAL CORD

Suggested Reading

Ahrén, B. (2000). Autonomic regulation of islet hormone secretion. Implications for health and disease. *Diabetologia*, **43,** 393–410.

Ahrén, B., and Lindskog, S. (1992). Galanin and the regulation of islet hormone secretion. *Int. J. Pancreatol.* **11,** 147–160.

Ahrén, B., and Pettersson, M. (1990). Calcitonin gene-related peptide (CGRP) and amylin and the endocrine pancreas. *Int. J. Pancreatol.* **6,** 1–5.

Ahrén, B., Taborsky, G. J., Jr., and Porte, D., Jr., (1986). Neuropeptidergic versus cholinergic and adrenergic regulation of islet hormone secretion. *Diabetologia* **29,** 827–836.

Arimura, A., and Shioda, S. (1995). Pituitary adenylate cyclase activating polypeptide (PACAP) and its receptors: Neuroendocrine and endocrine interaction. *Front. Neuroendocrinol.* **16,** 53–88.

Colmers, W. F., and Wahlestedt, C. (1993). *The Biology of Neuropeptide Y and Related Peptides*. Humana Press, Totowa, NJ.

Filipsson, K., Kvist-Reimer, M. and Ahrén, B. (2001). The neuropeptide pituitary adenylate cyclase-activating poly-peptide and islet function. *Diabetes* **50,** 1959–1969.

Karlsson, S., and Ahrén, B. (1992). Cholecystokinin and the regulation of insulin secretion. *Scand. J. Gastroenterol.* **27,** 161–165.

Mutt, V.(1988). Vasoactive intestinal polypeptide and related peptides. Isolation and chemistry. *Ann. NY Acad. Sci.* **527,** 1–19.

Sundler, F., and Böttcher, G. (1991). Islet innervation, with special reference to neuropeptides. In: *The Endocrine Pancreas* (E. Samols, Ed.), pp. 29–52. Raven Press, New York.

Woods, S. C., and Porte, D., Jr. (1974). Neural control of the endocrine pancreas. *Physiol. Rev.* **54,** 596–619.

Neuropharmacology

EDWARD A. WORKMAN

Lakeshore Mental Health Institute and University of Tennessee Medical Center, Knoxville

GLOSSARY

anti-convulsant An agent designed to reduce nervous system irritability and reactivity by changing sodium or calcium channel activity.

antidepressant An agent designed to treat depressed mood by altering metabolism and the transmission of neurotransmitters such as serotonin or norepinephrine.

anti-psychotic An agent designed to treat psychotic thought processes (e.g., cognitive disorganization, paranoid thinking) by altering the metabolism of dopamine, among other neurotransmitters.

bipolar disorder A major psychiatric disorder presumptively caused by a neurotransmitter imbalance, resulting in impairment of judgment and dramatic swings in mood from highly energized irritability to despair.

dementia Any one of the usually progressive, organic brain disorders characterized by disorganized thinking, severe memory dysfunction, and eventual functional debility.

dependence The process wherein an individual persistently uses a deleterious substance in order to avoid withdrawal effects (which are aversive) in the face of tolerance, which requires the use of progressively larger amounts of the substance to achieve the desired pleasurable effect.

depression A psychiatric disorder caused by aberrant metabolism of one or more neurotransmitters, resulting in persistent blue mood, low energy, slowing of cognitive tempo, and difficulty maintaining functional status.

drug Any externally administered agent that induces some biological response.

dyssomnia A disorder characterized by difficulty initiating or maintaining a normal pattern of sleep and waking.

neuroleptic An old term referring to pharmacologic agents designed to treat psychosis by inhibiting dopamine neurotransmission.

neuron A highly specialized type of tissue that has the capability to conduct electrical impulses via changes in sodium and calcium channels, thus allowing for communication from one site to another.

neurotransmitter A biochemical substance, usually a protein or a protein containing compound, that transmits information from one part of the nervous system to another.

schizophrenia The brain disease presumptively caused by aberrant dopamine neurotransmission with an onset usually in early young adulthood, characterized by cognitive disorganization, severe deficits in social functioning, and, often, delusions and hallucinations.

seizure A paroxysmal burst of electrical activity in the brain that lacks organization and focus, usually resulting in involuntary movements.

Neuropharmacology is the medical science field that examines the effects of pharmacologic agents (i.e., drugs) on the nervous system. Although the field is quite broad, encompassing both animal and human studies and both basic ''bench'' science and clinical science, the focus of this article is on the latter. Our perspective throughout this article is on those pharmacologic agents that are used to treat disease processes in the human nervous system. This group of diseases is also rather broad and falls under the purview of multiple medical fields, including psychiatry, neuropsychiatry, and neurology. The emphasis here, thus, is on clinical neuropharmacology or what

could be termed ''neuropsychopharmacology.'' The organization of the article is designed to reflect the emphasis on clinical neuropharmacology. By following a concise and simplified but, hopefully, reasonably inclusive discussion of the basic principles of neuroscience that underlie neuropharmacology, the reader will be oriented to the basic pathological processes encountered in psychiatry and neurology. The section on psychiatric disorders will review mood and thought disorders, the neurotransmitter imbalances thought to produce them, and the drug therapies that contemporary psychiatrists use to treat them. Neuropsychiatric disorders within our coverage include sleep disorders, cognitive deficit disorders, traumatic brain injury, and chemical dependence disorders. Finally, we will review the pharmacologic basis and treatment of neurological disorders, including seizure disorders, movement disorders, and headaches. This organization rather clearly reflects the clinical emphasis of the article. There is, of course, an enormous body of knowledge and literature on the basic science aspects of neuropharmacology. It is, however, beyond the scope of this article, and the reader who is interested in exploring that field is referred to the Suggested Reading section at the end of the article.

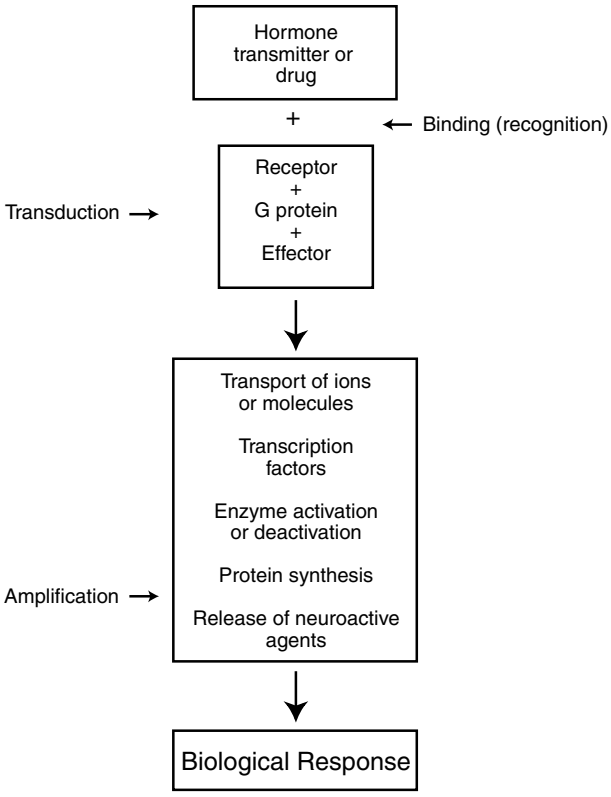


Figure 1 Model for receptor–ligand interaction.

I. BASIC PRINCIPLES AND THEORETICAL MODELS IN NEUROPHARMACOLOGY

Neuropharmacology depends upon many fields for its underlying structural basis. Cellular biology, molecular biology, biochemistry, and branches of neuroscience from neuroanatomy to neurophysiology all have made, and continue to make, substantial contributions to this field. For the sake of simplicity and conceptual economy, we will focus on that small set of models that provides the basic concepts for understanding how diseases develop in the nervous system and how drugs can treat these diseases.

Perhaps the most fundamental process in neuropharmacology is shown in Fig. 1. This figure portrays a neuroactive agent interacting with a receptor, eventually resulting in a specific biological response. In the natural state, neuroactive agents can include hormones, enzymes, or neurotransmitters. In the clinical setting, the neuroactive agent is a drug, a concept that is ubiquitous in its usage but frequently confused in its meaning. In this article we will use the term drug to include any externally administered neuroactive agent that results in a biological response (of any kind).

As can be seen in Fig. 1, the neuroactive agent interacts with (e.g., binds with) a receptor, which is a protein structure that is produced in the cell body of the neuron (nerve cell) and then embedded into its membrane surface. The binding of the neuroactive agent (NA) and receptor involves recognition of one for the other due to structural compatabilities. In other words, due to their specific molecular structures, a given NA will bind with some types of receptors but not others. This is the concept of transmitter–receptor specificity and forms a fundamental structural basis of neuropharmacology. In other words, a specific type of NA (e.g., the neurotransmitter, serotonin) naturally interacts with some specific types of receptors (e.g., serotonin type 2 and 3 receptors) but not others (e.g., adrenergic β1 and 2 receptors).

Once bound, the NA–receptor unit experiences changes in its inherent protein structure, allowing it to then bind to an intracellular G protein (e.g., adenylyl cyclase, phospholipase C). The NA–receptor–G-protein complex, once created, spontaneously experiences further structural changes (called changes

in conformation) that trigger interaction of the complex with an intracellular enzyme. This process is called transduction, because chemical energy is transduced from the structure of the NA to the NA–receptor–G-protein–enzyme complex, and results in the manufacture of what is called a second messenger molecule. Second messenger molecules include cyclic adenosine monophosphate (cAMP) and phosphatidylinositol (PI); first messengers are the neuroactive agents themselves. These second messengers comprise an effector system, which can yield effects in other parts of the neuron.

The second messenger system is extremely complex and remains under intense investigation at present. However, we do know that the second messenger molecules interact with additional intracellular enzymes to accomplish various biological effects. The processes involved include the following: (1) opening ion channels (e.g., Ca^{2+} ion channels) at the cell membrane, resulting in changes in intracellular concentrations of various agents and, thus, changes in the reactivity of the neuron itself; (2) inducing the neuron to produce various biochemical agents (e.g., neurotransmitters, hormones, and enzymes) by interacting with mRNA and altering the neuron's DNA instructions; and (3) up- or down-regulating the number and density of various receptors in the neuronal membrane via production or destruction of receptors and their structural components. Through these processes, naturally occurring NAs have their biological effects. By understanding how these processes take place, drugs can be designed that mimic the effects of natural NAs, and therein lies the basis of neuropharmacology. Drugs are developed or discovered for the purposes of mimicking or modulating a naturally occurring process that, for some reason, no longer occurs or does so in an inadequate manner (i.e., a disease or disorder).

II. NEUROTRANSMITTERS

The known number of neurotransmitters occurring naturally in the human nervous system appears to grow progressively as our investigational tools become more sensitive and sophisticated. Currently, Table I represents a reasonable articulation of the known human neurotransmitters; the reader should be aware that, in another decade, this table will be, perhaps substantially, larger and will likely include a number of surprises as our knowledge base grows.

Literally, a book chapter could be written about each of the categories of neurotransmitters in Table I. However, the clinical focus of this article dictates that we restrict our discussion to those neurotransmitters that currently have utility in clinical medicine. Although intense research effort is being applied to evaluating the clinical significance of many, if not all, of the neurotransmitters in the table, our focus here will be on the amine category.

A. Serotonin

Serotonin, or 5-hydroxytryptamine (5HT), is perhaps the most widely known neurotransmitter in a popular sense. The reason for this is the massive press given to the drug Prozac, a NA that results in an increase in the amount of serotonin available around neurons, a feat accomplished by inhibiting the destruction of serotonin. Serotonin or 5HT was originally discovered in the gastrointestinal mucosa in the 1940s. Its chemical structure is very similar to the amino acid tryptophan, which is, in fact, the dietary precursor of 5HT itself. 5HT is found in virtually all vertebrate organisms, as well as in many invertebrates such as wasps, scorpions, and ocean crustaceans. It is also found in abundance in several plant species, including bananas and pineapples. Although a neurotransmitter of major importance, 5HT is distributed throughout the body. Interestingly, approximately 90% of the total 5HT in the human body is in the gastrointestinal tract, 8% is in blood platelets (where it is involved in the clotting process), and only 2% is in the central nervous system.

5HT is synthesized in the terminal end of the neuron and transported to synaptic vesicles (cystlike structures that exist at the neuronal cell membrane), from which it is released into the synapse for interaction with receptors in adjacent neurons. Once released freely into the synaptic space, molecules of 5HT will either stimulate receptors in other neurons or they will interact with autoreceptors from their original neuron (the neuron that produced them). Interaction with adjacent neurons results in neurotransmission of a nature specific to the type of adjacent neuron. Interaction with autoreceptors results in an important process called re-uptake. Neurotransmitter re-uptake results in a decrease in the amount of the neurotransmitter (in this case 5HT) that is available in the synaptic space. Through re-uptake, the 5HT neuron can exert modulatory control over the amount of serotonergic neurotransmission generated from that neuron: net transmission can be reduced through the

Table I
Neurotransmitters in the Brain

Amines	Serotonin (5HT)
	Dopamine (DA)
	Norepinephrine (NE)
	Epinephrine (EPI)
	Acetylcholine (ACH)
Pituitary peptides	Corticotropin (ACTH)
	Growth hormone (GH)
	Lipotropin
	Oxytocin
	Vasopressin
	Melanocyte stimulating hormone
Circulating hormones	Angiotensin
	Calcitonin
	Glucagon
	Insulin
Amino acids	γ-Aminobutyric acid (GABA)
	Glycine
	Glutamic acid
	Aspartic acid
Gastrointestinal hormones	Cholecystokinin
	Gastrin
	Motilin
	Secretin
	Substance P
	Vasoactive intestinal peptide
Opioid peptides	Dynorphin
	β-Endorphin
	Met-enkephalin
	Leu-enkephalin
	Kyotorphin

production of more autoreceptors, which, in turn, trap and destroy free 5HT molecules in the synapse. By increasing the density of autoreceptors, the neuron reduces the amount of 5HT that can exert its effect in the synapse. On the other hand, the nerve cell can increase serotonergic transmission by reducing production of autoreceptors, thereby destroying less 5HT and thus increasing the net amount of serotonergic neurotransmission.

5HT is synthesized from the amino acid tryptophan. Tryptophan is hydroxylated to L-5-hydroxytryptophan, which is, in turn, decarboxylated to the serotonin molecule itself. In addition to the preceding mechanisms at the neuron's disposal for regulating serotonergic neurotransmission, it can also modulate the production of 5HT via the regulation of the activation state of the enzyme, tryptophan hydroxylase. Once synthesized, 5HT is involved in a wide variety of behavioral processes, including hunger and food intake, mood regulation, aggressive behavior and its control, sleep architecture, and anxiety and its regulation.

Two decades of research have yielded a listing of the clinical conditions that have been shown to be influenced by altered serotonergic neurotransmission. Although surely not exhaustive, the list underscores the importance of serotonergic neurotransmission processes and documents the wide-ranging distribution of 5HT systems. The disorders include mood disorders, anxiety disorders, eating disorders, migraine headaches, neuro-degenerative processes and aging, obsessive compulsive disorder, substance abuse, pain sensitivity (chronic pain), posttraumatic stress disorder, schizophrenia, sexual dysfunction, and sleep disorders.

B. The Catecholamines: Dopamine, Norepinephrine, and Epinephrine

The catecholamines, including dopamine (DA), norepinephrine (NE), and epinephrine (EPI), belong to a group of neurotransmitters called monoamines. These molecules contain a single amine ($-NH_2$) group, a catechol nucleus (consisting of a benzene ring with two hydroxyl groups), and a side chain consisting of an ethylamine or a closely related derivative. EPI was originally called adrenaline and NE was called noradrenaline, as both were initially found in the adrenal gland. Retaining these original names, neurotransmission via EPI and NE is still termed adrenergic and noradrenergic, respectively.

Catecholamine neurotransmitters were first discovered by Walter Cannon and his associates in the 1920s. In their classic experiments involving nervous system arousal, the researchers discovered that these substances (then called sympathins) were released whenever sympathetic nerves were stimulated. Thus, they recognized that catecholamines were involved in nervous system arousal and stimulatory states.

All three of the preceding catecholamines are synthesized from the amino acid tyrosine as derived dietary proteins. Tyrosine is hydroxlated (via tyrosine hydroxylase) to DOPA, which is then decarboxylated to dopamine. Dopamine can then be transported to the synaptic space and released for use in dopaminergic neurotransmission, or it can be further metabolized within the neuron. In the latter case, dopamine (DA) is β-hyroxylated to norepinephrine (NE), which can, likewise, be transported and released for neurotransmission or further metabolized. If further metabolized, NE is methylated to epinephrine (EPI). Thus, DA can be used by the neuron as a neurotransmitter or as a precursor substrate for the production of two other neurotransmitters, NE and EPI.

As was the case with the 5HT neuron, the DA neuron can control the amount of neurotransmission via its production of autoreceptors, thereby controlling re-uptake of the catecholamines. Additionally, the DA neuron has modulatory control over DA, NE, and EPI neurotransmission via its control over the rate-limiting enzyme, tyrosine hydroxylase. Production of NE and EPI is modulated by regulation of the intraneuronal enzymes dopamine β-hydroxylase and phenylethanolamine *N*-methyltransferase, respectively. Further, the catecholamines (as well as 5HT) can be destroyed in the synaptic space by an enzyme called monoamine oxidase (MAO); thus, neuronal production and release of MAO into the synaptic space are yet other mechanisms whereby neurotransmission can be modulated by the neuron.

Once synthesized and released into the synaptic space, the catecholamines serve multiple functions throughout the human nervous system, particularly the brain itself. Although it is usually an arousal-inducing neurotransmitter, DA neurotransmission varies in function depending upon the neuroanatomic system involved. In the nigrostriatal system (with fiber tracts running from the substantia nigra to and from the basal ganglia), for example, DA transmission controls motor behavior and its fine control. In the mesolimbic system (where fiber tracts traverse the brain stem and the nucleus accumbens), DA transmission is involved in cognitive and affective processing and emotional regulation. When involved in the mesocortical system (with fibers traversing the tegmentum and the frontal lobes), dopaminergic neurotransmission plays a role in attention, arousal, and motivation. In the tuberoinfundibular pathway (with fibers running to and from the hypothalamus and the tegmentum), DA is involved in the regulation of prolaction secretion (and, relatedly, bone mineral metabolism), fertility, and sexual function. In the median forebrain bundle and neuronal fibers connecting to the hypothalamus, DA appears to be intimately involved in processes of reward, pleasure, and motivation.

NE is involved in the regulation of arousal throughout the nervous system. EPI is involved with a relatively small number of neuron systems and primarily has very rapid short-lived and localized effects. Thus, NE has more substantive neuropharmacologic significance and will be the focus of this section. Centrally, adrenergic neurotransmission (via NE and, to a lesser extent, EPI) is involved in myriad processes including activation, arousal, attention–concentration, hunger, and feeding behavior. Peripherally, adrenergic transmission is involved in cardiac function, blood pressure regulation, and muscle tone, among other processes.

C. Acetylcholine

The neuroactivity of acetylcholine (ACh) has been characterized since the early 1920s, and the molecule was first discovered in the mid-1800s. It was, thus, the first neurotransmitter to be identified and has been the focus of a massive amount of research over the past century.

ACh is synthesized in cholinergic neurons from the precursors acetyl coenzyme A and dietary choline via the catalyst choline acetyltransferase. Several mechanisms regulate the production of ACh, including (1) feedback inhibition wherein increasing amounts of ACh in the nerve terminal ''feed back'' to the producing neuron to inhibit the activity of choline acetyltransferase, thus reducing further ACh production, (2) the availability of dietary choline, and (3) neuronal activity wherein ACh is depleted in the nerve terminal, thus stimulating the accelerated production of ACh. In addition to these regulatory mechanisms, ACh activity is also modulated by the presence of enzymes that hydrolytically break down ACh. These enzymes are called acetylcholinesterases and, as will be seen later, are the targets of various pharmacologic agents.

Once synthesized, ACh is found to be widely dispersed throughout the body, particularly in the brain and spinal motor neurons. In the central nervous system, ACh is heavily involved in the striatal complex (e.g., the caudate–putamen region and the nucleus accumbens) where it is, perhaps, the most important

motor system neurotransmitter. In the substantia nigra region, ACh is a cofactor in dopaminergic neurotransmission. The basal forebrain region represents a major ACh neurotransmission area, with cholinergic fiber projections traversing the area from the medial septal nucleus to the limbic cortex and hippocampus with involvement in memory, cognitive processing, and emotional functions. In the diencephalon, cholinergic neurons and their projections are involved in a variety of neuroendocrine functions, whereas the cholinergic fibers in the medulla innervate the cerebellar cortex and motor components of some cranial nerves. In the peripheral nervous system in spinal and cranial motor neurons, ACh is involved in the movement of limbs, head, trunk, eyes, face, tongue, and jaw. In the parasympathetic nervous system, ACh functions as a neurotransmitter in connections between ganglionic fibers and smooth muscle tissue in virtually all systems. In the sympathetic nervous system, ACh is involved in neurotransmission to sweat glands.

ACh is, thus, widely dispersed and nearly ubiquitously involved as a neurotransmitter. It has been demonstrated to function as a regulator of aggressive behavior, sensory processes, learning and memory, sleep and arousal, eating behavior, digestion, and sexual behavior. In the 1970s, several laboratories demonstrated that ACh neurotransmission tended to decline with age and that cholinergic neurotransmission activity was particularly deficient in the brains of patients with Alzheimer's dementia. As will be seen in subsequent sections, these findings have become the focus of an important area of research in neuropharmacology.

D. Amino Acid Neurotransmitters: GABA, Glutamate, Glycine, and Aspartate

Although the catecholamines ACh and serotonin have extremely important neurotransmitter roles in brain function, the amino acid neurotransmitters are more common by far and exist in substantially higher nervous system concentrations. Despite their ubiquitous distribution in mammals, these neurotransmitters have proven more difficult to study than any others. Nevertheless, there is a growing body of literature regarding their characteristics and roles.

γ-Aminobutyric acid (GABA) was originally synthesized in the 1880s and for years was thought to be exclusively a product of plant metabolism. In the 1950s, however, it was found to be a constituent of mammalian tissue and was determined to exist in large quantities in the brain and spinal cord. Over the past several decades, much evidence points to GABA's role as an inhibitory neurotransmitter. GABA is synthesized in GABAergic neurons by the conversion of L-glutamate via the catalytic enzyme, glutamate decarboxylase (GAD). GABA synthesis is regulated primarily by the presence and concentration of GAD and its cofactor, pyridoxal phosphate (a form of vitamin B_6), within the neuron. Following its synthesis and release into the synaptic space, GABA interacts with postsynaptic receptors to exert inhibitory effects; GABA thus is a neurotransmitter that serves to dampen or depress the effects of other neuron systems. Although there is more active research than conclusive evidence about the inhibitory effects of GABA, it is thought to be involved in the regulation of anxiety and relaxation states and overall motor tone. A growing body of research is also implicating GABA in seizure disorder development. The inhibitory effects of GABA appear to involve one of two processes, both of which result in making the affected target neuron more difficult to stimulate: hyperpolarization and depolarization. Hyperpolarization appears to be more the more common GABAergic mechanism in central cortical neurons, whereas depolarization predominates in spinal cord neurons.

Another inhibitory amino acid neurotransmitter is glycine, the amino acid with the simplest structure. Found in all mammalian proteins and tissues, glycine is synthesized primarily via hydroxymethylation from serine. Once synthesized, glycine functions as an inhibitory neurotransmitter in the spinal cord and, to a lesser extent, in the brain. In the brain, glycine's role primarily appears to involve the inhibition of neuronal activity in the striatum, substantia nigra, and cerebellum. Pharmacologically, research is still ongoing regarding the drugs that can affect this neurotransmitter, but it is well-characterized in many of its actions due to its inhibition by the poison strychnine. Observation of strychnine effects has led to an understanding of the various regulatory functions of glycine, including relaxation of muscles of facial expression and mastication, limb and trunk movement, relaxation of respiratory and cardiac muscles, as well as vascular wall musculature, and regulation of various visual, auditory, cutaneous, and vestibular functions.

Two amino acids function as excitatory neurotransmitters: glutamate and aspartate. Both are nonessential amino acids, can, thus, be synthesized in the body, and do not require a dietary source. Glutamate is

synthesized primarily as a byproduct of glucose metabolism (the Kreb's cycle) via a transamination reaction. Once synthesized, glutamate can, interestingly, serve as the precursor for GABA, or it can function as an excitatory neurotransmitter in the spinal cord and, literally, throughout the brain in virtually every region. It appears to be involved in learning, memory, and the regulation of neuronal oxygenation states and is, thus, purported to be important in the regulation of neuronal damage following strokes and heart attacks. Although aspartate is known to be excitatory, it is less well-characterized as a neurotransmitter, and most of the research on this agent is in the early investigatory stages.

E. Histamine

The final neurotransmitter upon which we will focus here is histamine. Like serotonin and the catecholamines, histamine is a biologically active amine derived from amino acid precursors. It is formed by the decarboxylation of the amino acid histidine and is degraded via methylation to type B monoamine oxidase. Histamine, although heavily involved in the regulation of allergic and inflammatory reactions, has been found to be a neurotransmitter that is involved in a wide variety of behavioral processes. It is involved in the regulation of ingestive behavior (eating and drinking), sleep and arousal, sexual behavior, pain tolerance, learning and memory, and blood pressure.

III. NEURO-RECEPTORS: THE OTHER SIDE OF NEUROTRANSMISSION

Receptors in the nervous system are essentially glycoprotein-based recognition sites for neurotransmitters. When interacting with a given receptor, a neurotransmitter can be a full or partial agonist, which stimulates maximal or submaximal effect, respectively, or an antagonist, which blocks the usual effect of stimulating the receptor. A neurotransmitter can also be an inverse agonist at a receptor, stimulating the receptor to have the opposite effect that it usually has when stimulated by an agonist (inverse agonists do not block receptors, they stimulate them). Chronic interaction between an agonist neurotransmitter and its receptor typically results in down-regulation (decreased production and distribution) of the receptors; chronic antagonism results in up-regulation (increased production and distribution) of receptors. In this manner, transmitter–receptor interactions serve as a negative feedback loop based, self-regulatory system for the control of net neurotransmission processes.

Table II
Neurotransmitter Families

Adrenergic receptors	α 1a, 1b, 1c, 1d, 2a, 2b, 2c, 2d
	β 1, 2, 3
Dopaminergic receptors	D1, D2, D3, D4, D5
GABA receptors	GABA-A, GABA-B1A, B1γ B2, GABA-C
Glutaminergic receptors	NMDA, AMPA, kainate, MGLUr-1–7
Histamine receptors	H1, 2, 3
Cholinergic receptors	
	Muscarinic: M1–5
	Nicotinic: muscle, neuronal
Opioid receptors	μ, γ, κ
Serotonergic receptors	5HT-1A, 1B, 1D, 1E, 1F, 5HT-2A, 2B, 2C, 5HT-3, 4, 5, 6, 7

Table II lists the currently known neurotransmitter receptors. As can be seen, most receptors have multiple subtypes, a fact that became apparent during the past decade and has added to the complexity of neuropharmacologic processes.

Adrenergic receptors are G-protein-coupled receptors for catecholamines. Two primary subtypes exist, α and β. α adrenergic receptors are located in blood vessels, smooth muscle sites, and throughout the nervous system. α-1 receptors are postsynaptic receptors and are characterized by high sensitivity to the agonists phenylephrine and methoxamine and to the antagonists prazosin and phenoxybenzamine. α-2 receptors are primarily presynaptic and are activated by clonidine and α-methylnorepinephrine and antagonized by yohimbine, rauwolscine, and idazoxan. Subtypes of both α-1 and α-2 receptors have been found and characterized by their differential responsiveness to various agonists and antagonists. β adrenergic receptors are found in the heart, blood vessels, and other smooth muscle sites, as well as in nervous system tissue. β-1 and β-2 receptors differ in their reactivity to EPI (NE and EPI are equally potent in β-1 activation, whereas EPI is far more active than NE in β-2) and localization. β-3 receptors are atypical and have been found to exist in adipose tissue as well as in several types of gastrointestinal cells. Stimulation of these receptors results in lipolysis and thermogenesis.

Dopaminergic receptors, like adrenergic receptors, are G-protein-coupled receptors. The first group of DA receptors, found in the 1970s, consisted of D-1 and D-2. Currently five subtypes of DA receptors have been found and characterized, D1–5. It is likely that many more exist. Although receptors of the D-1 family exist in wide distribution in mammalian nervous tissue, they are particularly densely located in the mesostriatal regions of the brain. They are, thus, heavily involved in neuromotor functions. Interestingly, the D-5 receptor is considered to be a part of the D-1 receptor family and is found primarily in the thalamohippocampal regions of the brain.

The D-2 family of dopaminergic receptors was an early focus of research because of its affinity for the antipsychotic drugs, such as thorazine (formerly called neuroleptics). D-2 receptors are found in a variety of locations in addition to the mesostriatal region, including the mesolimbic and mesocortical areas. D-2 receptors are also located in a chemical trigger zone of the medulla, where stimulation with apomorphine induced nausea and vomiting. The highest density of D-2 receptors appears to be in the caudate–putamen, nucleus accumbens, and substantia nigra regions. Classic antipsychotic agents, such as thorazine, mesoridazine, and haloperidol, are antagonists for the D-2 and, to a lesser extent, the D-1 receptor systems. Thus, the effects of these agents are not restricted to emotional and cognitive processes but also affect neuromuscular function. D-3 and D-4 receptors differ from the D-1 and D-2 families on the basis of their different localizations and affinities for agonists and antagonists. D-3 receptors appear to be particularly dense in the limbic regions of the brain and are thought to represent an additional site of action of antipsychotic agents. D-4 receptors, first characterized and cloned in the early 1990s, are thought to be sites of action of the atypical anti-psychotic, clozapine.

Multiple serotonin (5HT) receptors have been discovered since the mid-1980s. 5HT-1 receptors were the identified first, and currently five subtypes of this receptor are known, designated 1A, 1B, 1D, 1E, and 1F. The 5HT-1A subtype is the best characterized and is activated by the agonists buspirone and gepirone (both azapirone derivatives). These receptors are quite dense in the hippocampus, raphe nucleus, amygdala, and cortical regions and have been implicated in the regulation of feeding, temperature, and anxiety–arousal. Perhaps the most important role of the 5HT-1 receptor involves its function as a postsynaptic auto receptor, wherein it detects the presence of 5HT in the synapse and blocks further release of 5HT. In this manner, it functions as the regulator of 5HT synthesis in neurons. Additional 5HT-1 subtypes have been characterized on the basis of their differential interactions with agonists and antagonists, but their functions remain elusive.

The 5HT-2 receptor family has three currently known subtypes, A, B, and C. 5HT-2A receptors are densely distributed in the hippocampus and cortical and neocortical regions of the brain. The B subtype is found in the cortex as well as the gastrointestinal tract, accounting for the side effects (see later discussion) of the serotonergic antidepressants. The C subtype is distributed primarily in the hippocampus and medulla. The 5 HT-2 receptor is stimulated to action by the agonists α-methyl-5HT (a 5HT analog), the phenylalkylamines, and lysergic acid diethylamide (LSD). Methylchlorophenylpiperzine (m-CPP), a metabolite of the antidepressant trazodone, is a partial agonist at the 5HT-2 sites. Antagonists include ketanserin, ritanserin, mesulergine, and quinolone. Ordinarily, chronic administration of these antagonists would be expected to result in an up-regulation of 5HT-2 receptors; however, the 5HT-2 system exhibits a paradoxical characteristic wherein chronic antagonism results in down-regulation. As such, it is possible that what were originally thought to be antagonists are actually inverse agonists. These findings underscore the fluid and evolving nature of the knowledge base of neuropharmacology.

5HT-3 receptors have been found to be densely located in the brain stem and the area postrema. At both locations they appear to be involved in the regulation of nausea and vomiting. Unlike most other receptor families, they are not G-protein-coupled but are operated through a gated channel in the neuronal membrane.

The 5HT-3 receptor is structurally and functionally more similar to the nicotinic cholinergic receptor family than the other 5HT receptors. This receptor is activated by methyl-5HT and antagonized by ondansetron, granisetron, metoclopromide, and MDL72222, a derivative of cocaine. It is thought that 5HT-3 antagonists exert an antiemetic effect by blocking the effect of 5HT on the gastrointestinal tract and by blocking the release of 5HT from the vagus nerve (which triggers emesis by the brain stem's vomiting control center).

ACh receptors are of two subtypes, muscarinic and nicotinic. Nicotinic ACh receptors are always excitatory and transmit nerve impulses very rapidly, on the order of milliseconds. They are blocked by tubocurarine, the active agent in the paralytic poison curare,

hexamethonium, mecamylamine, and gallamine. Nicotinic receptor agonists include nicotine itself, decamethonium, and succinylcholine. Nicotinic ACh receptors are distributed throughout muscle and nerve tissue and are particularly important in striated muscle and sympathetic and parasympathetic neurons.

Muscarinic ACh receptors can be either excitatory or inhibitory and have a long latency of action (on the order of 100 msec). They are antagonized by atropine or scopolamine and activated by muscarine, pilocarpine, and oxotremorine. Muscarinic receptors are also widely distributed, particularly in parasympathetic neurons that subserve cardiac and smooth muscle at sites throughout the mammalian system. However, most of the central nervous system cholinergic activity takes place via muscarinic receptors in the neocortex, hippocampus, striatum, and thalamus. Four subtypes have been found, M1–4, that vary in their localization and, to some extent, the agents by which they are activated or antagonized. For example, M1 receptors are heavily distributed in the striatum on dopaminergic terminals of the nigrostriatal system; stimulation of these receptors by ACh results in the release of dopamine. In the septal–hippocampal pathways, M1 receptors, upon simulation, release ACh, which in turn inhibits the release of the excitatory neurotransmitters, glutamate and aspartate. ACh receptors are, thus, involved in diverse regulatory functions involving other families of neurotransmitters, in addition to having their own roles in neuromuscular and cortical transmission.

Now that we have explored the basic principles and models that underlie neuropharmacology, we will turn our attention to the clinical significance of this information. The remainder of this article will focus on the disorders of brain chemistry to which neuropharmacology applies its knowledge base.

IV. PSYCHIATRIC DISORDERS

A. Mood Disorders

Contemporary biological psychiatry strongly emphasizes the neurobiology of mood and thought disorders as the area upon which it focuses its diagnostic and treatment armamentarium. Although psychiatry deals with a wide variety of disorders, most psychiatrists spend the majority of their time diagnosing and treating major depression, bipolar disorder, schizophrenia, and anxiety spectrum disorders. Although the exact neurobiological mechanisms for these disorders are yet to be fully elucidated, reasonable theoretical models for their development have been articulated and are well-accepted in the medical community.

Major depression is a severe mood disorder characterized by blue mood, extreme sadness, loss of interests, low energy, and cognitive slowing, among other symptoms lasting for at least several weeks (DSM-IV). Unlike simple "bad moods," major depression is a painful, debilitating, and sometimes, in the case of suicidally depressed patients, life-threatening. Approximately 15% of the population can expect to suffer from severe major depressive illness at some point in their lives, and 1 in 50 of these people will require hospitalization. Even those not requiring inpatient treatment will experience an increase in their cardiac mortality and morbidity rate, increased likelihood of depressive recurrences, and some degree of functional debility during their depressive episode.

In the "monoamine hypothesis of depression," depressive symptoms are thought to be due, at least in part, to a deficiency in neurotranmission involving one or more of the neurotransmitters serotonin, norepinephrine, or dopamine. This deficiency can be the result of genetics or the individual's interactions with a highly stressful environment, as is the case with most psychiatric disorders.

The first suggestion of a biochemical basis for major depression came from early work in the 1950s, when a drug, iproniazid, that affected monoamine re-uptake (as a side effect) was used to treat tuberculosis and found to improve the mood of the patients. This led to the development of a class of drugs called monoamine oxidase inhibitors (MAOIs). These drugs exerted antidepressant effects by inhibiting the MAOs in the neuronal synapses that destroyed the monoamines 5HT, NE, and DA, thus increasing the synaptic concentration of these neurotransmitters. MAOIs include two groups: the hydrazines, iproniazid, isocarboxazid, and phenelzine, and the nonhydrazine tranylcypromine. These agents result in the nonselective increase in synaptic NE, EPI, 5HT, and DA and their effects are irreversible. Although highly efficacious in the treatment of major depression, these agents fell into disuse due, in part, to their requiring a special tyramine-free diet to avoid a malignant hypertensive crisis (tyramines are found in common foods such as cheeses and red wine and, when ingested along with MAOIs, they produce the release of supranormal amounts of NE, causing a rise in blood pressure). Additionally, MAOIs interact with a variety of other pharmacologic agents, including alcohol, opiates, barbiturates, and aspirin, altering the body's ability

to metabolize them and resulting in a prolongation and enhancement of their effects.

Two subtypes of MAO have been identified: MAO-A and MAO-B. The former appears to be most active in deaminating 5HT, NE, EPI, and DA, whereas the latter deaminates DA, tyramine, and phenylethylamine. Agents that are selective for MAO-A are under investigation as antidepressants and include moclobemide, clorgyline, and cimoxatone. Theoretically, such agents should not exhibit a side effect profile as problematic as that of the original MAOIs. Nevertheless, the MAOIs are still in use, but only as third and fourth line agents with the most treatment-resistant cases of major depression.

Following development of the MAOIs, apparently safer antidepressants were developed that specifically blocked the re-uptake pumps for 5HT, NE, or DA, thus increasing the amount of neurotransmitter available in the synapse. These were the tricyclic antidepressants (TCAs), and they were the first efficacious, widely used and marketed agents for the treatment of major depression. These agents had a three-ring structure, hence their name. The most well-known of these include imipramine, amitriptyline, desipramine, and nortriptyline. Desipramine almost exclusively blocked the re-uptake pump (autoreceptor) for NE; the others had mixed effects, blocking the re-uptake autoreceptors for both 5HT and NE. The TCAs were essentially of equal effectiveness as antidepressants but differed in their side effect profiles. Unfortunately, all of the TCAs were highly nonspecific in their target effects. In addition to blocking re-uptake pumps in their targeted synapses, they all had additional untoward effects including the blockade of ACh and histamine receptors. The former resulted in dry mucous membranes (eyes and mouth), changes in vision, confusion (particularly in the elderly), constipation, and urinary difficulties. The latter resulted in increased appetite (especially for carbohydrates), weight gain, and sedation. Additionally, TCAs also have antagonistic effects on the α-1 receptor, resulting in blood pressure decrements (particularly upon standing), and possess the ability to prolong the QT interval on EKG due to their sometimes altering a susceptible patient's cardiac rhythm. The latter requires clearance with a normal EKG prior to starting TCA therapy in older patients. Although they were highly effective in the treatment of major depression, TCA side effects frequently resulted in poor compliance and lack of full patient satisfaction.

In order to provide an alternative to TCAs, a new class of antidepressants was developed in the 1980s. These were the selective serotonin re-uptake inhibitors (SSRIs). The first of these was fluoxetine (Prozac), an SSRI that was shown to be as efficacious as the TCAs but without the troublesome side effects of ACh and histamine receptor blockade. SSRIs exert their antidepressant effect by inhibiting the serotonergic re-uptake pump in the presynaptic 5HT neuron, thus blocking the destruction of 5HT. This, in turn, results in an increase in 5HT in the synapse, an increase in serotonergic neurotransmission, and down-regulation of the 5HT-1A receptors. These processes occur in multiple cortical pathways that are thought to be deficient in serotonergic neurotransmission in major depression: the frontal cortex, basal ganglia, hippocampus, and raphe nucleus.

Since the development and success of Prozac (fluoxetine), four additional SSRIs have been developed and marketed: sertraline (Zoloft), paroxetine (Paxil), fluvoxamine (Luvox, which was actually developed for obsessive compulsive disorder), and citalopram (Celexa). All have the same mechanism of action, and all have been used successfully in both depressive and anxiety spectrum disorders (e.g., panic disorder, obsessive compulsive disorder, and posttraumatic stress disorder). Their use in both anxiety and depressive disorders has called into question the differences (or lack thereof) in the underlying neurobiological mechanisms of these disorders, and this represents an intense area of investigation at present.

Although the various SSRIs all inhibit the presynaptic re-uptake pump, they also affect other serotonergic receptors to varying degrees. When they stimulate (agonize) 5HT-2 receptors, the result can be agitation, anxiety, sexual dysfunction (particularly anorgasmia), and insomnia. They also can stimulate 5HT-3 receptors; in some patients, as would be predicted by the foregoing discussion, this results in nausea, dyspepsia, and diarrhea. 5HT-3 receptor agonist effects have also been used to explain the side effect of headaches. The SSRIs, in general, also interact with a liver enzyme system called the cytochrome P450 system and its isoenzymes 1A2, 2D6, 2C19, and 3A4. This mitochondrial system functions to metabolize a wide variety of drugs (e.g., carbamazepine, theophyllin), and its inhibition can result in an increase in the serum level (and, sometimes, the toxicity) of these drugs. The various SSRIs differ in the degree to which they inhibit a given p450 isoenzyme (e.g., fluoxetine has a high impact on 2D6 and 2C19, but less on 1A2; paroxetine has a high impact on 2D6, but a low impact on 1A2), so care must be taken

for the patient to avoid problematic drug interactions with the SSRIs (e.g., opiate analgesics).

Since the development of SSRIs, several atypical antidepressants have been developed and marketed for major depression. Venlafaxine (Effexor) is a dual-channel agent that inhibits the re-uptake of 5HT, NE, and DA. Interestingly, venlafaxine only inhibits the re-uptake of 5HT at a low dose; at moderate doses both 5HT and NE re-uptake is blocked. At high doses, re-uptake of 5HT, NE, and DA is blocked. At low doses, as would be predicted, it has the same side effects as SSRIs; at medium to high doses, it can, in some individuals, cause hypertension, insomnia, agitation, and headache. It should be noted that venlafaxine has some, albeit low level, interactivity with all of the known P450 isoenzymes.

Trazodone is an antidepressant that was actually developed in the early 1980s. It is unique in its structure and mechanism: it antagonizes 5HT-2 receptors and functions as a 5HT re-uptake inhibitor simultaneously. It also blocks α-1 and histamine receptors, resulting in some side effects reminiscent of TCAs. More recently, the trazodone cogener nefazodone (Serzone) was developed. Nefazodone, like trazodone, antagonizes 5HT-2 receptors and blocks 5HT re-uptake, but, in addition, it blocks re-uptake of NE. This agent has been particularly useful in major depression associated with anxiety, agitation, or insomnia and appears to have fewer sexual side effects than SSRIs. Unfortunately, both trazodone and nefazodone produce a metabolite, mCPP, that results in dizziness and lightheadness in a relatively small percentage of patients.

Bupropion (Wellbutrin) is an atypical antidepressant that blocks the re-uptake of NE and DA. It is used widely in those cases of major depression where the level of deenergization is a major problem. The enhancement of noradrenergic and dopaminergic neurotransmission likely results in increased feelings of energy. Also, it is virtually without sexual or gastrointestinal side effects and is often used in cases where these have been problematic.

Finally, mirtazapine (Remeron) is an atypical antidepressant with an extremely unusual receptor profile. It blocks α-2 receptors on noradrenergic and serotonergic neurons, resulting in an increase in available NE and 5HT. Simultaneously, it blocks 5HT-2 and 5HT-3 receptors, minimizing the side effects (mentioned earlier) associated with stimulation of these receptors and, thus, avoiding sexual and GI side effects. It also blocks the histamine receptor, which likely counteracts the noradrenergic propensity for increased anxiety. Unfortunately, this action can also sedate and cause weight gain.

Currently, a wide variety of antidepressant agents are available to the clinician. A great deal of work is currently ongoing that attempts to match agent(s) to the specific subtype of major depression (e.g., agitated–anxious, deenergized, comorbid with pain) to be treated. However, selection is usually based upon the fact that the various agents vary somewhat in their side effects.

Another type of mood disorder often encountered by psychiatrists is bipolar affective disorder (BPAD) (DSM-IV). This is a serious mood disorder that occurs in about 1% of the population and is characterized by severe mood swings—from agitated, uncontrollable excitement (or, alternatively, hostility and irritability) to abject depression and back, often accompanied by psychotic delusions (usually of grandeur, wherein the patient believes himself or herself to be a celebrity or some other prominent person, or paranoia, wherein he or she believes that others are trying to harm him or her in some way). Various subtypes of BPAD include the classic syndrome of mood cycling from manic excitement to depression, cycling from mania to euthymia, and rapid cycling from one state to another in a matter of a few days. Ordinarily, the mood swings of BPAD patients last for many days or weeks, and their behavior during these intervals can cause serious problems in their relationships, careers, and financial well-being. Etiologically, the specific mechanism of BPAD is not elucidated, but there is clearly a genetic component (the disorder tends to be heritable) and manic episodes tend to be stimulated by stress. It is thought that aberrations in the noradrenergic and serotonergic systems are responsible for manic excitement, and it has been shown that antidepressants, particularly TCAs, can stimulate a manic episode. Thus, a complex balance of noradrenergic and serotonergic neurotransmission systems may be dysfunctional in BPAD. Further, it is thought that the psychosis that sometimes accompanies acute BPAD episodes is related to excess dopaminergic transmission. The responsiveness of BPAD to specific medication therapies underscores these theoretical assertions.

The first pharmacotherapy for BPAD was lithium carbonate. A simple salt of the element lithium, which was used in the nineteenth century for gout and arthritis, this agent's therapeutic potential was discovered in the late 1940s, but was not approved for use in the United States until the 1970s due to its toxicity in overdose. Almost immediately upon approval it

became the most widely used treatment for BPAD and was shown to be highly effective in the stabilization and long-term treatment of manic illness. Although its precise mechanism of action is not known, lithium has been shown to (1) block the neuron's ability to recycle inositol phosphate and thus synthesize inositol from glucose (this reduces the firing rate of affected neurons and dampens their overall sensitivity) and (2) modulate both serotonergic and noradrenergic neuronal systems. Lithium enhances the effect of 5HT itself and elevates the brain concentrations of tryptophan (a 5HT precursor) and 5HT. Regarding its noradrenergic effects, lithium increases the re-uptake of NE and also reduces its release. Both of these actions result in a net decrease in noradrenergic neurotransmission activity, which likely accounts for its calming effect on patients with manic excitement. Furthermore, these effects, in the face of the clear efficacy of lithium as a treatment for BPAD, suggest that the etiological basis of BPAD involves neuronal hyperexcitability and inadequate serotonergic activity combined with excessive noradrenergic activity.

Despite the efficacy of lithium therapy in BPAD, the agent is not without problems and complications. First, lithium, unlike the vast majority of other psychotropic agents, is metabolized by the kidneys rather than the liver. This means that it can interact with a variety of other renally metabolized agents, including the ubiquitous nonsteroidal anti-inflammatories, most diuretics, some antibiotics, and the popular angiotensin converting enzyme (ACE) inhibitor antihypertensives. Lithium effectiveness, furthermore, requires that the patient attain a serum lithium level between 0.8 and 1.2 mmol/liter; steady state (attained after 4 days of dosing) serum levels under these limits are usually ineffectual, whereas levels over 1.5 mmol/liter result in progressive toxicity. Both lithium's interaction profile and its tight therapeutic range require that the serum level be periodically monitored. Close clinical monitoring of patients in lithium therapy is necessitated by an array of potential side effects on the thyroid, gastrointestinal, cardiac, and dermatologic systems.

Due to the obvious complexities in the clinical use of lithium, other pharmacologic agents have become increasingly more popular in the treatment of BPAD. These belong to the general group of anticonvulsants, including carbamazepine (Tegretol), valproic acid (Depakote), and, more recently, lamotrigine (Lamictal) and gabapentin (Neurontin). Originally developed for the treatment of various seizure disorders, these agents serve as mood stabilizers at least in some subgroups of BPAD patients. Although, as with lithium, the precise mechanisms of action of anticonvulsants on BPAD are not clear, a partial picture is emerging. Valproic acid exerts an effect on the GABA system, thus stimulating inhibitory brain pathways. It also appears to modulate dopaminergic neurotransmission. Carbamazepine modulates the activity of several neurotransmitters, including 5HT, NE, and DA, via sodium channel dampening of each neuron system's excitability. Less is known about the mechanisms of lamotrigine and gabapentin, but both apparently act by effecting a decrease in neuronal excitability, presumptively in the 5HT, NE, and DA systems.

B. Schizophrenia

Schizophrenia is, without doubt, the most severe and devastating of all the psychiatric disorders. It occurs in approximately 1% of the population, has its onset in the early adult or late teen years, and has a clear heritable component (it occurs more often in individuals with affected family members). The most often encountered subtypes of schizophrenia are paranoid schizophrenia and disorganized schizophrenia. In the former, the individual is plagued by delusions of being persecuted by others, sometimes one's family or even by strangers. Sometimes the delusions are accompanied by hallucinations of voices that threaten or give commands to the affected patient. In disorganized schizophrenia, the patient may or may not have delusions and hallucinations, but most prominently suffers from loose associations, inability to process and follow a logical train of thought, and frequently, extreme social withdrawal and lack of appropriate social interaction skills. The etiology of schizophrenia is complex and multifactorial, but it clearly involves biochemical brain aberrations including excessive dopaminergic activity in the mesolimbic and mesocortical brain pathways. The former appears to produce the positive (or outwardly manifested) symptoms of schizophrenia, such as delusions and hallucinations. It is thought that the dopaminergic aberrations in the latter pathways produce the negative (or the absence of normality) symptoms of the disease, such as disorganization and social withdrawal.

Early drug treatment of schizophrenia involved the use of so-called neuroleptics, including phenothiazines (chlorpromazine or Thorazine, thioridazine or Mellaril, and trifluoperazine or Stelazine). Later more potent and presumptively less sedating antipsychotics

were developed, including the butyrophenones Haloperidol and Droperidol. All of these agents block D2 receptors. In the mesolimbic and mesocortical systems, this is desirable and efficacious. However, blockade of D2 receptors in the nigrostriatal and tuberoinfundibular regions results in side effects such as motor dyscontrol (e.g., extrapyramidal symptoms, tardive dyskinesia) and hyper-prolactinemia (resulting in breast secretions and sexual dysfunction).

In order to avoid the side effects associated with mass D2 blockade at multiple indiscriminate sites, research efforts have resulted in the development of a new class of antipsychotic agents called serotonin dopamine antagonists (SDAs) or atypical antipsychotics. These agents include Clozapine (the first of its class), Olanzapine, Risperidone, and Quetiapine. Ongoing research suggests that these agents (1) have more specificity for the mesolimbic and mesocortial D2 receptors, (2) have less specificity for striatal and infundibular D2 receptors, and (3) antagonize 5HT-2A receptors, which is thought to increase dopamine availability at striatal and other sites. Continuing clinical use of these agents does, in fact, suggest that they have less propensity for producing the side effects associated with typical antipsychotics or neuroleptics, while being at least equally as effective in treating positive symptoms of schizophrenia. Perhaps of equal importance is the growing body of evidence that these agents may be more effective than the typicals agents for the negative symptoms of schizophrenia.

V. NEUROPSYCHIATRIC DISORDERS

In this section, we will explore the pharmacologic basis and treatment of several disorders that span the realms of psychiatry and neurology. The disorders that we will address include cognitive deficit syndromes (e.g., dementias), sleep disorders, and chemical dependence disorders.

A. Dementias

Dementias involve cognitive deterioration (e.g., memory decrements and deficits in attention and concentration) and confusion that progress as a function of age. The most common and most widely studied dementia is Alzheimer's disease (AD). This condition remains a focus of extensive contemporary research, and its etiology is far from clear. However, research has indicated several likely etiologic factors, including the abnormal deposition of amyloid plaques around neurons and neuron bundles (neuritic plaques), which form neurofibrillary tangles, and the depletion of ACh particularly in the cortex and hippocampus. Also, it has been shown that M2 and nicotinic ACh receptors are reduced in patients with AD. The latter body of research has led to pharmacologic treatments designed to treat the cholinergic deficits in AD. Perhaps the most widely used agent of this genre is tetrahyroaminoacridine (Tacrine). This is a high-affinity, noncovalent inhibitor of acetylcholinesterase that, by reduction of the enzyme that inhibits ACh activity, increases net cholinergic activity in the brain. Other acetylcholinesterase inhibitors in clinical use are velnacrine, donepezil, and metrifonate. All have the same basic mechanism but differ in their affinities for acetylcholinesterase and metabolism and, therefore, their side effects. A summary of the research has shown that, although not all studies demonstrate effectiveness, the bulk of the evidence indicates that these agents can improve the behavioral and cognitive functions of AD patients to varying degrees.

B. Sleep Disorders

Sleep disorders are rather common in the general population and include (1) dyssomnias, (2) parasomnias, and (3) sleep dysfunction secondary to other neurological, general medical, or psychiatric disorders. Dyssomnias are by far the most common sleep disorders and represent conditions wherein the patient has difficulty either initiating or maintaining sleep; some have both syndromes. Both are characterized by chronic irritability, feelings of tiredness, excessive daytime sleepiness, poor concentration, and, in some cases, significant memory impairment. Etiological factors are widely variable and include recent significant stressors, poor sleep hygiene, and disturbances in circadian rhythm. The former is primarily an excessive arousal reaction to stress, which can be treated pharmacologically with arousal modulating agents. Trazodone has long been used to reduce nocturnal arousal and normalize sleep architecture. More recently, Nefazodone has been used similarly and shown to improve sleep architecture. The benzodiazepines (e.g., Lorazepam and Valium) have long been used to induce and maintain sleep but have fallen into disfavor of late due to the potential for producing dependence and dysfunction in memory and rapid eye movement (REM) sleep with prolonged use. The non-benzodiazepine hypnotic, Zolpidem, has seen heavy clinical use

in the past decade and has been shown to induce and maintain sleep without dependence or changes in cognition and sleep staging.

Poor sleep hygiene, although frequently treated pharmacologically with agents such as those listed previously, is best treated by changing the patient's sleep habits. For example, avoiding the intake of caffeine in the 4–5 hr prior to sleep and going to bed only when sleepy (or for sex) (to avoid associating the bed and bedroom with insomnia) have both been shown to be effective in some patients. Circadian rhythm disturbances are common in specific occupational groups such as factory shift workers, nurses, and others with periodic changes from day to night shift work. Pharmacologic induction of sleep is often used on a temporary basis with such patients, but attaining a stable and consistent work schedule has been shown to be preferable "treatment."

Parasomnias are relatively uncommon sleep disorders and involve unusual behaviors, movements, and sensations during sleep. Often, the patient will experience jerking or myoclonus of the lower extremities, sometimes called restless legs syndrome. Repetitive, rhythmic movements of the head, neck, and limb affect another small subgroup of patients; some patients will find themselves awakened by their walking about while asleep. Although detailed sleep studies are usually necessary to design an individualized treatment program for parasomnia patients, many are helped pharmacologically by the administration of sedating antidepressants such as trazodone and nefazodone. Those with problematic nocturnal muscle movements are responsive to the high-potency, highly sedating, and skeletal muscle relaxing benzodiazepine, Clonazepam. This agent, which is also classified as an anticonvulsant, appears to act by very significantly reducing cognitive arousal as well as by reducing nocturnal skeletal muscle tone.

Sleep dysfunction due to other general medical or psychiatric conditions is most often due to pharmacologic side effects. For example, the SSRIs, venlafaxine, and bupropion often cause sufficient arousal to result in difficulty initiating sleep and, thus, require very slow titration from low doses as the patient becomes accustomed to their effects. Prednisone and theophylline, both widely used for inflammation and respiratory dysfunction, respectively, are well-known for their tendency to disrupt sleep and inhibit its onset. Medical and psychiatric conditions that commonly cause sleep dysfunction include hyperthyroidism, chronic obstructive pulmonary disease (COPD), Parkinson's disease, degenerative arthritis, depressive spectrum disorders, bipolar affective disorder (BPAD), and various anxiety disorders (e.g., panic disorder, generalized anxiety disorder). In the case of general medical disorder induced sleep disorders, pharmacologic treatment with sedation antidepressants or with Zolpidem is often effective. With psychiatric disorders, effective treatment of the disorder itself usually results in the resolution of the secondary sleep disorder.

C. Chemical Dependence Disorders

Chemical dependencies are characterized in DSM-IV and other sources by intense and consuming substance preoccupation, intoxication, tolerance, withdrawal, and, most importantly, significant impairment of social and occupational functioning due to the substance's use. One of the major problems in developing empirically based treatment for these disorders is that, despite the present consensus criteria for their diagnosis, there is rather astounding conceptual confusion around the concepts of addiction, dependence, and abuse of various substances. Related to this conceptual confusion is the difficulty in making a firm and reliable chemical dependence diagnosis, even when done by well-trained professionals.

Regardless of the conceptual confusion around issues of what constitutes abuse vs dependence and despite the difficulties in making a diagnosis in the complex behavioral arena of chemical dependence, psychiatrists frequently encounter patients with substance use that seriously impairs their ability to function. The most common substance dependencies seen by psychiatrists include opiate, stimulant (e.g., cocaine and amphetamine), and alcohol dependence.

Opiates are narcotic analgesics whose primary action is pain reduction. Secondary actions involve the production of an intense state of euphoria and relaxation, the basis of their dependence induction. For hundreds if not thousands of years, the sole opiate ingested was opium, from which the term opiate was derived. In the early 1800s, the active ingredient, morphine, was synthesized, and in 1874, the morphine molecule was altered to synthesize the highly dependence inducing agent, heroin. Morphine and its derivatives cause their CNS effects via interaction (as agonists) with the opiate receptors: μ, γ and κ (σ receptors were once thought to be opiate receptors, but this theory is no longer considered tenable). μ receptors are heavily distributed in the cerebral cortex, thalamus,

hippocampus, locus coeruleus, amygdala, caudate, putamen, and the dorsal portion of the spinal cord. This receptor is involved in the production of both analgesic and euphoric states. γ receptors are distributed similarly to μ receptors, but more restrictively. Their primary locations are in the neocortex, striatum, and substantia nigra, with the expected involvement in cognitive processes and motor integration. The κ receptor binds ketocyclazocine, an opiate analog molecule that produces hallucinations and dysphoric mood states; thus, it is thought to have a possible role in psychiatric disorders, but presently this is far from clear. Opiate interactions with κ receptors are involved in GI motility, food and water intake, thermoregulation, and pain perception.

Natural opiate-like neurotransmitters, called endogenous opioids or endorphins, exist throughout the nervous system and function as agonists with the preceding opiate receptors. The currently known and characterized endogenous opioids include the enkephalins met-enkephalin and leu-enkephalin, β-endorphin, and dynorphin. These agents have multiple precursors that may also have the ability to interact with opiate receptors.

Opioid effects in the nervous system are highly complex and involve interactions with other neurotransmitter systems. For example, opioid interactions with μ receptors in the brain stem result in the stimulation of GABA activity, with resultant inhibition of sensory processes in the spinal cord. Enkephalins in the caudate are localized with and probably modulate the activity of the neurotransmitter glutamate. It behaves similarly in the raphe nucleus and area postrema with 5HT. Dynorphin in the arcuate and supraoptic nuclear regions interacts with and modulates tyrosine hyroxylase and thereby influences the catecholamines. Research is ongoing in attempts to more fully characterize how opiates interact with their own receptors and with other neurotransmitter systems to produce analgesia, euphoria, and dependence states.

Despite the at best partial picture of the actions of opiates, pharmacologic methods have been developed to treat opiate dependence. The earliest treatment protocols involved an attempt to substitute agents that do not have the problematic effects of opiates. For example, methadone therapy involves the controlled administration of methadone, a synthetic narcotic analgesic similar in effect to morphine. The theory was that methadone would reduce narcotic craving, making the patient more amenable to rehabilitation. In reality, methadone is cross-dependent with morphine and heroin, and one chemical dependence is basically substituted for another, albeit a more "controlled" and clinically administered dependence in the case of methadone.

Detoxification under clinical supervision is another pharmacologic treatment protocol for opiate dependence. In this model, the symptoms of opiate withdrawal are modulated and minimized via pharmacotherapy. Clonidine, for example, is an α-2 agonist that results in the inhibition of excess arousal generated by withdrawal. It thus decreases the syndrome of sweats, myospasm, and stomach cramping. Benzodiazepines, in systematic tapered regimens, also have found usefulness in minimizing the arousal of withdrawal.

Finally, a more recent treatment option involves the use of narcotic antagonists following detoxification. These agents include naloxone, naltrexone, and cyclazocine and function to block the euphoric effects of opiates. Naltrexone is a μ receptor specific antagonist, has a relatively long duration, and has few side effects; thus, it is the most commonly used agent. When one takes naltrexone, the ingestion of heroin or other opiates has no euphoric effect, thus reducing the motivation for their ingestion. Unfortunately, the opiate-dependent patient can simply stop taking the antagonist agents for several days to re-instate the euphorigenic potential of opiates.

Alcohol is the most widely used substance that results in chemical dependence and appears to interact with GABA receptors at multiple sites. This results in reduced skeletal muscle tone, decreased arousal and anxiety, and impaired cognitive and sensory processes. Chronic alcohol use also appears to result in down-regulation of the dopaminergic system and a reduction in NE synthesis and level, possibly resulting in another avenue of decreased arousal and anxiety. Although the withdrawal symptoms of alcohol dependence are effectively treated with benzodiazepine in a taper down protocol, alcohol dependence treatment itself has not generated a literature showing clear efficacy. Disulfiram (Antabuse) is sometimes used to prevent alcoholics from consuming alcohol due to the fact that it produces toxicity when combined with consumed alcohol. This agent irreversibly inhibits the action of aldehyde dehyrogenase, an enzyme that converts acetaldehyde (a major metabolite of alcohol) to acetate (which is readily excreted). Thus, consumption of alcohol while on disulfiram results in a buildup of acetaldehyde, which results in a highly unpleasant state including tachycardia, flushing of the face, nausea and vomiting, dizziness, and shortness of

breath. This provides a form of "aversion" therapy wherein alcohol use, instead of resulting in a pleasant relaxed state, results in a highly noxious state. Unfortunately, many if not most alcohol patients learn quickly that all they need do is stop taking disulfiram so that they can drink alcohol with "impunity." Another pharmacologic treatment of alcoholism is naltrexone, the μ antagonist used to treat opiate dependence. Naltrexone is used on the theory that its blockade of μ receptors may block "brain reward" centers that are stimulated with alcohol and, thus, reduce the alcoholic's propensity for alcohol consumption. Unfortunately, there is no unequivocal evidence for its efficacy, particularly in the face of the ability of the alcoholic patient to terminate its use to enhance the pleasure producing effects of alcohol, as is the case with disulfiram.

Stimulants are drugs that produce increased alertness, arousal, and excitement, as well as euphoria. The most commonly abused stimulants are amphetamine and cocaine. Amphetamine blocks the re-uptake of DA and NE (and possibly 5HT) throughout cortical regions and is also a potent stimulator of DA release from neurons. It, thus, increases the amount of DA available and then blocks the neuron's ability to regulate DA along with NE and possibly 5HT. Cocaine is an alkaloid stimulant that exerts its stimulatory action by blocking the re-uptake of DA, NE, and 5HT. Both amphetamine and cocaine can produce chemical dependence syndromes. Stimulant dependence is often treated pharmacologically with dopamine agonists and antidepressants.

Bromocriptine is a D2 receptor agonist that has been used on the theory that stimulant "craving" and withdrawal are due to DA depletion; a D2 agonist would, thus, replete synaptic DA levels. Unfortunately, this treatment has not been shown to be clearly efficacious. Fluoxetine has also been used on the assumption that its arousal inducing properties (although serotonergic rather than dopaminergic in nature) would replace the "need" for the stimulant. Buproprion, an antidepressant that has both noradrenergic and dopaminergic properties, has been used in the treatment of amphetamine and cocaine dependence, but there is no consistent evidence of its efficacy.

While on the subject of chemical dependence, the reader is likely aware that there is currently much discussion of dependence upon the mild stimulants, nicotine and caffeine. Although there is clear evidence that smoking is a health risk factor in a variety of illnesses and excess caffeine consumption may be similarly implicated in some gastrointestinal illnesses, their status as dependence inducing agents is far less clear than the media and public health authorities would suggest. Some researchers have argued strongly against the use of a dependence model to explain nicotine use, citing evidence that, despite allusions to the similarity of heroin and nicotine in the press and public health statements, there is, in fact, little similarity when one examines the impact the use of the respective substances has on immediate functional status. Other researchers have demonstrated that cocaine and amphetamine result in a dramatic ($>50\%$) lowering of the threshold for brain reward stimulation in the dopaminergic reward centers of the brain, providing a marker for their dependence inducting capabilities; agents such as caffeine and nicotine result in substantially smaller and similar (to each other) effects ($<30\%$ brain reward stimulation threshold lowering), strongly suggesting that the two sets of pharmacologic agents are quite different in their effect on brain pleasure centers and, thus, possibly their dependence inducing ability.

Despite the conceptual and definitional problems in the literature, many people want to stop smoking and there have been neuropharmacologic treatments designed to assist. The most common involve nicotine gum and the nicotine transdermal patch. The former provides oral nicotine via chewing, whereas the latter provides nicotine secreted through the skin in a progressive taper down dosage form. Both are based on the assumption that people benefit from the ingestion of nicotine (nicotine stimulates ACh receptors and this can release DA into the synapse, likely resulting in an alteration of arousal state), and replacement in an oral or transdermal form replaces that obtained from smoking. Unfortunately, only about 10% of patients report that these methods help them quit smoking, and several studies have shown that the presence or absence of nicotine in delivery devices has little, if any, effect on reports of withdrawal symptoms. The low success rate of nicotine replacement clearly indicates that smoking involves factors in addition to attaining the pharmacological effects of nicotine.

Another pharmacologic strategy for smoking cessation involves the attempt to replace a state of high arousal and evidence that many smokers become depressed when they cease the use of nicotine and may, thus, be self-treating depressive symptoms with nicotine. This treatment involves the antidepressant bupropion, marketed for smoking cessation as Zyban. As the reader recalls, bupropion has both noradrenergic and dopaminergic properties, and it is one of the most

arousing of the antidepressants available. Unfortunately, there is no conclusive data on its efficacy in smoking cessation.

VI. NEUROLOGICAL DISORDERS

The final section of this article focuses on several disorders that have traditionally fallen under the purview of neurology. These include seizure disorders, movement disorders, and headaches.

A. Seizure Disorders

Seizure disorders involve episodic abnormal discharges of groups of neurons in the central nervous system. Their etiology lies in the paroxysmal stimulation of neuronal firing, and this is modulated by sodium channel current states, GABAergic activity in affected neurons, and calcium channel current states. The classification of these disorders is nearly a field unto itself, and a full discussion of this issue is beyond the scope of this article. In general, seizures can be viewed as falling into three categories based upon clinical manifestations and localization vs generalization of symptoms: tonic–clonic generalized, tonic–clonic partial, and absence.

Tonic–clonic seizures can involve generalized convulsive activity (involving the entire body) or partial convulsive activity, wherein one arm or one leg, for example, exhibits convulsive activity. Most anticonvulsants are used for both types of seizures, but the drugs of choice vary according to type. In generalized tonic–clonic seizures, the current drugs of choice are valproic acid (discussed in the section of BPAD) and topiramate, despite their pending FDA approval status for these purposes. Carbamazepine, phenytoin, phenobarbital, primidone, and lamotrigine are also used with generalized seizures. The drugs of choice for partial tonic–clonic seizures are carbamazepine, phenytoin, and valproic acid. Also used are phenobarbital, primidone, gabapentin, lamotrigine, felbamate, topiramate, tiagabine, and vigabatrin.

Absence seizures involve the absence of typical convulsive activity and instead, are characterized by episodic periods of the absence of normal awareness and arousal. The patient reports an episode where he or she "lost touch" or "just went away." Observers typically only notice that the individual appears not to pay attention to what was going on in the environment. When an EEG study is made during such an episode, it shows typical seizure-like brain wave activity despite the lack of muscle system involvement. The drugs of choice for the treatment of absence seizures include ethosuximide and valproic acid. Other drugs include clonazepan (a benzodiazepine), and lamotrigine.

The definitive picture of the mechanism of anticonvulsants, like antidepressants and antipsychotics, is not fully clear. However, it is known that most, if not all, serve to modulate neuronal activity by altering sodium channel activity, calcium channel activity, and GABAergic activity in neuron groups that are associated with a given type of seizure. Some agents, such as phenytoin and carbamazepine, primarily have their effect on sodium channels where they block activity by binding to and inactivating the channel. Phenobarbitol interacts primarily with the GABA system and facilitates the opening of chloride channels in response to GABA, thus serving to inhibit affected neurons. Phenobarbitol also interacts with glutamate by interfering with its excitatory actions on relevant neuron groups. Topiramate blocks sodium channels, enhances GABA inhibitory effects, and antagonizes glutamate. Ethosuximide has its effect by inhibiting calcium channel activity in the thalamic regions, where the synchronous firing of neuron groups appears to summate into absence seizures. Two of the newest anticonvulsants are vigabatrin and tiagabine. Both of these agents have their effects by enhancing GABA activity, which, as the reader knows, is inhibitory. Vigabatrin binds to and inactivates GABA transaminase, the enzyme responsible for degrading GABA. Tiagabine blocks the re-uptake of GABA, much as the antidepressants do with 5HT and NE, thus increasing the availability, concentration, and duration of the inhibitory actions of GABA.

B. Movement Disorders

Movement disorders involve abnormal, insufficient, or excessive motor activity. Most of the movement disorders involve stiffness, rigidity, bradykinesia (slow movements), tremors, tics, gait difficulties, and loss of control of movements. The most common movement disorder in neurology is Parkinson's disease, which is by far the most studied and well-understood. Other conditions include Tourette's syndrome, tremors, and myoclonus. Parkinson's disease is characterized by resting tremor (a tremor that worsens at rest, compared to during activity), cogwheel rigidity (rigidity that, in the arms, can be felt with movement showing stiffness, then giving way, then restiffening), slowing of

movements, and, finally, the absence of movement (for example, the classic loss of facial expression due to the inability to move the facial muscles). The disease usually starts on one side and remains most severe on that side. Also, its final stage involves a severe dementia that is similar in presentation to that of Alzheimer's disease described earlier.

Parkinson's disease involves etiologically obscure degeneration of the nigrostriatal pathway, a dopaminergic neuron system that originates in the substantia nigra and terminates in the striatum. The nigrostriatal pathway synthesizes and stores a large amount of DA. As the reader recalls, DA is synthesized from tyrosine, which is converted to L-dihyroxyphenylalanine (L-DOPA) by the enzyme tyrosine hyroxylase. L-DOPA is then converted to DA by decarboxylation via dopa decarboxylase. DA actions are regulated by its re-uptake and degradation by the enzymes MAO and catechol-*O*-methyltransferase (COMT). Although Parkinson's clearly involves the dopaminergic system, the ACh, 5HT, and GABA systems have also been implicated in its etiology.

The treatment of Parkinson's disease has focused upon symptomatic relief in reducing the motor symptoms of the disease by targeting various points in the synthesis of DA described earlier. The most common strategy has been to replace DA by giving the patient L-DOPA, the DA precursor. Another strategy has been to treat with either carbidopa or benserazide, which are both peripheral dopa decarboxylase inhibitors. Most often, the two agents are administered together for maximal effect. Both agents obviously focus on directly combating the loss of DA with the degeneration of the nigrostriatal system in the disease. Another strategy has involved the administration of DA receptor agonists, such as bromocriptine, pergolide, and lisuride, in hopes of enhancing dopaminergic activity. Selegeline, an inhibitor of the B subtype of MAO, which thus blocks the destruction of DA, has been used with some success. More recently, two new agents, entacapone and tolcapone, have also been used with some success. Both agents inhibit COMT at multiple metabolic sites, thus dramatically increasing the availability of DA. Unfortunately, none of the anti-Parkinson agents have done any more than delay the worsening of the disease, because the primary lesion involves the large and important dopaminergic motor system in the brain and its progressive degeneration cannot be stopped by simply replacing DA or slowing its degradation. As with many if not most of the diseases in this article, a real cure awaits much more research in the neurobiology of the diseases for which neuropharmacology offers therapeutic protocols.

C. Headaches

The next disorder we will focus upon involves headaches. Headache is ubiquitous and is always one of the top five reasons why people visit doctors. Headaches are classified by various schemes, but, most simply, they can be viewed as involving migraine headache, tension headache, cluster headache, and trigeminal neuralgia. Migraine headache in its various manifestations can involve unilateral or bilateral headache that is usually associated with a prodrome (preceding the headache) or codrome (occurring with the headache) of nausea, vomiting, photosensitivity–photophobia, and alterations in mood. Its etiology is not fully known, but it is now thought to involve excessive serotonergic activity due to histamine-induced release of 5HT from mast cells in the vertebrobasilar arterial system. In this model, the excessive serotonergic activity results in alternating constrictions and relaxation of vascular structures, which, with repetition, result in vasospasm. The vasospasm is thought to be the primary mechanism of the headaches themselves.

Historically, a wide variety of agents have been used to treat migraines, but ergotamine and dihydroergotomine have been, until recently, the mainstays of migraine treatment. Both agents inhibit NE re-uptake at sympathetic nerve endings, thereby causing vasoconstriction. The treatment rationale was that "forced" vasoconstriction would stop the process of sequential vasorelaxation and vasoconstriction that results in vasospasm. Other agents used in the treatment of migraine, especially in the chronic preventive rather than acute abortive treatment, have included tricyclic antidepressants, SSRIs, neuroleptic antipsychotics, β-adrenergic antagonists, anticonvulsants, and nonsteroidal anti-inflammatories (NSAIDS). Aside from NSAIDS, where the rationale was simply to stop the vascular inflammatory process thought to be involved in the initiation and maintenance of migraine, the rationale for the use of many of these agents was sometimes obscure and based more on clinical lore and clinician observation of efficacy than on empirical investigation.

Methysergide, a serotonin antagonist, was shown to be efficacious in the 1970s but was plagued by a variety of side effects, including a uncommon but severe syndrome of posterior mediastinal and cardiac fibrosis. As the serotonergic basis of migraine became more

substantiated in the 1990s, a new class of antimigraine abortive agent emerged, the 5HT-1 receptor agonist. The 5HT-1 receptor is a presynaptic autoreceptor that functions as a re-uptake pump. A 5HT-1 agonist thus stimulates the 5HT re-uptake pump, making less 5HT available. In the case of migraine, this is thought to stop the serotonergically mediated vasospasm process. The currently available 5HT-1 agonists include sumatriptan, naratriptan, and zolmitriptan. Sumatriptan is a 5HT-1 agonist that is structurally similar to 5HT itself and interacts with the 5HT-1 site without the usual effects of 5HT. Naratriptan is essentially identical in function to sumatriptan. Zolmitriptan is a selective agonist of the 5HT-1B and 5HT-1D receptor family and is thought to stimulate re-uptake without the side effects associated with blocking the 5HT-1 site in a more generalized manner.

Tension headache is characterized by a band of muscle tension in the frontal, temporal, and occipital regions of the scalp, locations where most people experience tension due to stress and anxiety. Pharmacologic treatment of these common headaches is usually reserved for those with extremely severe and intractable conditions. NSAIDS are commonly used for their straightforward analgesic effects. Additionally, tricyclic antidepressants as well as SSRIs have become more commonly used with severe tension headache on the assumption of an underlying mood or anxiety disorder. In some cases, antispasmodic agents such as isometheptene (combined with the anticonstrictives dihloralphenazone and acetaminophen with the trade names Midrin and Isopap) and baclofen have been used with some success.

Cluster headaches are cyclical headaches that literally occur in clusters (of daily headaches for several days), often in cycles varying with seasonal changes. They usually awaken the patient after 1–2 hr of sleep. The etiology is unknown, but several neuropharmacologic agents have been used in their treatment, including lithium carbonate (on the theory that the cyclical headaches are akin to mood cycling) and methysergide (on the theory that cluster headaches are a migraine variant with serotonergic etiology).

Trigeminal neuralgia is a syndrome of head pain due to blood vessel or inflammatory impingement upon the trigeminal nerve root. The syndrome is viewed as a type of seizure-like involvement of that specific nerve root due to excessive mechanical stimulation via impingement. The mainstay of treatment is and remains carbamazepine, as this agent modulates the sodium channel activity of the trigeminal nerve root and decreases its firing rate, thus reducing painful sensations. The other agent used in the treatment of trigeminal neuralgia is phenytoin, with a mechanism essentially identical to that of carbamazepine.

D. Chronic Pain Syndromes

Related to the subject of headache is a vast array of disorders subsumed under the concept of chronic pain syndrome (CPS). CPS can involve any severe pain in virtually any area of the body that has lasted for 6 months or longer. The pain can stem from muscle, joint, or skeletal tissue, called musculoskeletal pain, or it can be generated by damaged nerve tissue, called neuropathic pain.

Musculoskeletal CPS can involve a single joint, i.e., arthritis, a single muscle or set of muscles, i.e., myofascial pain, or multiple sites, such as in the syndrome fibromyalgia, wherein at least 11 specific body sites have pain and tenderness. The mainstay of musculoskeletal CPS treatment involves NSAIDs, as described previously. The basic theory is that musculoskeletal pain, at least partially, involves inflammation of the target tissue. NSAIDs, thus, attack inflammation as a pain generator. A second set of pharmacologic agents used widely in musculoskeletal CPS includes the antispasmodics. Two such agents, cyclobenzaprine and chlorzoxazone, both widely used, inhibit the activity of polysynaptic reflexes at the neuromuscular junction. As such, they reduce skeletal muscle tone. Another second line agent is baclofen, a much more powerful antispasmodic. Baclofen acts by stimulating the GABA receptor in neuromuscular regions and, therefore, increases the inhibition of muscle activity. There is also some evidence that baclofen has an inhibitory effect on substance P, a major transmitter of painful sensory processes.

Neuropathic CPS involves pain stemming from damage to nerve tissue. The pain generator can be a single large nerve or set of nerves (as in polyneuropathy), or it can involve networks of small nerve structures, as in what was formerly called reflex sympathetic dystrophy (now called complex regional pain syndrome or CRPS). The etiologic agent in neuropathic pain syndrome can vary widely, ranging from diabetes (high blood glucose can damage nerve fibers), to impingement of nerve fibers by other structures (tumors, ruptured vertebral disks), to direct trauma to a nerve or nerve networks (projective wounds, crush injuries, or displaced fractures that tear nerve tissue).

Despite the varied etiology and nature of neuropathic CPS, one commonality is its treatment. Anticonvulsants, as described earlier, are the mainstay of its treatment. The theory is that anticonvulsants reduce nerve fiber reactivity and, thus, reduce the transmission of pain impulses. Nerve blocks with anesthetic agents such as Lidocaine (which reduces nerve impulse activity by modulating sodium channels) are also often used in trying to modulate neuropathic pain.

Regardless of etiology or type, CPS is often a challenge to treat pharmacologically. Often physicians must resort to the use of opiate narcotic analgesics to control pain. Although chemical dependence among CPS patients is not nearly as common as once thought, care must be taken to avoid side effects, abuse, or dependence upon narcotics in the CPS population. The complexity of pain management is such that a new field of medicine has emerged, pain medicine, dedicated to diagnosing and appropriately treating pain in all its forms. This author, who is a member of the American Academy of Pain Medicine, believes that this new field will surely generate a plethora of new neuropharmacological approaches to pain control in the future.

VII. SUMMARY

This article has explored neuropharmacology from the standpoint of the most important clinical conditions in the fields of psychiatry, neuropsychiatry, and neurology. Neuropharmacology contributes to these fields by bringing to bear the models of receptor–ligand interactions as one basis for the diseases upon which they focus. These models yield neuropharmacologic agents that can alter aberrant neurophysiological and neurochemical processes. Although many of the diseases upon which this article focused are not fully elucidated in terms of mechanism, ongoing neuropharmacologic research, by developing and testing more receptor-selective drugs, is slowly providing a means for piecing together their etiologies. Through a cyclical process, whereby neuropharmacologic agents are suggested by receptor models of disease, the receptor models are refined on the basis of the effects of the agents, and new agents are, thus, developed on the basis of receptor model refinements. Many of the diseases about which we know little in the year 2002 may be more treatable and possibly even "cured" in the years to come.

See Also the Following Articles

BEHAVIORAL PHARMACOLOGY • CATECHOLAMINES • CHEMICAL NEUROANATOMY • COGNITIVE PSYCHOLOGY, OVERVIEW • DEMENTIA • DEPRESSION • MANIC–DEPRESSIVE ILLNESS • MOOD DISORDERS • NEUROBEHAVIORAL TOXICOLOGY • NEUROPSYCHOLOGICAL ASSESSMENT • PSYCHOACTIVE DRUGS • SCHIZOPHRENIA

Suggested Reading

Bozarth, M., Pudick, C., and KouLee, R. (1998). Effect of chronic nicotine on brain stimulation reward. *Behav. Brain Res.* **96,** 185–194.

Cooper, J., Bloom, F., and Roth, R. (1996). *The Biochemical Basis of Neuropharmacology*. 7th ed. Oxford University Press, New York.

Gori, G. (1996). Failings of the disease model of addiction. *Human Psychopharmacol.* **11,** s33–s38.

Kinon, B., and Lieberman, J. (1996). Mechanisms of action of atypical antipsychotic drugs: A critical analysis. *Psychopharmacology* **124,** 2–34.

Linsen, S., Zitman, F., and Breteler, M. (1995). Defining benzodiazepine dependence: The confusion persists. *Eur. Psychiatry* **10,** 306–311.

Luik, J. (1996). "I can't help myself": Addiction as ideology. *Human Psychopharmacol.* **11,** s21–s32.

Macdonald, R. (1992). Seizure disorders and epilepsy. In *Principles of Drug Therapy in Neurology*. (Johnson, Macdonald, and Young, Eds.). F. Davis Company, Philadelphia.

Penney, J., and Young, A. (1992). Movement disorders. In *Principles of Drug Therapy in Neurology*. (Johnson, Macdonald, and Young, Eds.). F. Davis Company, Philadelphia.

Scheuer, M., and Cott, A. (1998). Antiepileptic drugs. In *Current Neurologic Drugs*, 2nd ed. (L. Rowland, Ed.). Williams and Wilkins, New York.

Shaffer, H. (1997). The most important unresolved issue in the addictions: Conceptual chaos. *Substance Use Misuse* **32,** 1573–1580.

Silberstein, S. (1998). Agents for migraine and other headaches. In *Current Neurologic Drugs*, 2nd ed. (L. Rowland, Ed.). Williams and Wilkins, New York.

Tallman, J. (1999). Neuropsychopharmacology at the new millennium: New industry directions. *Neuropsychopharmacology* **20,** 99–105.

Tsuang, M., and Faraone, S. (1990). *The Genetics of Mood Disorders*. Johns Hopkins University Press, Baltimore.

Waldman, S. (1996). Headache and facial pain. In *Pain Medicine: A Comprehensive Review* (P. Raj, Ed.). Mosby–Year Book, New York.

Warburton, D. (1989). Is nicotine use an addiction? *The Psychologist: Bull. Br. Psychol. Soc.* **4,** 166–170.

Neuroplasticity, Developmental

DAPHNE BAVELIER
University of Rochester

HELEN NEVILLE
University of Oregon

GLOSSARY

altered experience Experience that is atypical for the organism, such as deprivation of input or exposure to stimulation outside the range normally encountered by the organism.

compensatory plasticity Changes in behavior and brain organization resulting from altered experience.

sensitive–critical period Developmental time period during which experience can significantly alter the organism's behavioral performance and related aspects of brain structure and/or function.

synaptogenesis–neurogenesis Generation of new synapses and new neurons.

The term neuroplasticity is generally used to refer to the capacity of the nervous system to modify its organization. Such changes may occur as a consequence of many different events, including normal development and maturation of the organism, the acquisition of new skills (learning), damage to the nervous system, and sensory deprivation. Studies to date of the molecular and cellular events underlying neural plasticity in such different conditions have revealed a limited set of mechanisms available to induce changes in the organization of the neural networks of the brain. Such reports raise the hypothesis and the hope that the diverse phenomena referred as neuroplasticity will be elucidated in the not-too-distant future. However, whereas there is evidence to suggest that there is considerable overlap in the cellular and molecular mechanisms that mediate neuroplasticity, we are still left with the challenge of understanding why a second language can be difficult for adults to learn, why children born with only one cerebral hemisphere fair so well compared to individuals who lose a hemisphere in adulthood, or why some skills are easier to learn than others. Studies of changes in behavior as a result of development or learning suggest more specificity than might have been predicted from the reports of a limited repertoire of cellular or molecular changes. Our goals are to qualify the specificity of plastic changes, characterize the parameters that determine this specificity, and elucidate its neural mechanisms. Although there is no agreed upon way to divide up the field of neural plasticity, the little data currently available suggest a greater potential for plastic changes during development than adulthood. Even if some of the mechanisms of neuroplasticity are similar or even identical during developmental and adult plasticity, the fact that these mechanisms operate on nervous systems that are structurally and physiologically different is likely to result in quantitative and/or qualitative differences in neuroplasticity in immature and mature organisms. For example, the number of synapses is 50% greater in the immature human brain than in the adult brain. This redundant connectivity

exists at different times in different brain regions and almost certainly constrains the nature and extent of modifications that can occur. Accordingly, studies of animals indicate considerable variability and specificity in the types of plastic changes that can occur from system to system as a function of age. In view of these differences between developmental and adult plasticity, we focus on developmental plasticity in this article. This article starts by reviewing neural events during human brain development. It then turns to within-modality plasticity and reviews the effects of altered visual experience during development on the cerebral organization for vision. Between-modality plasticity is discussed next by considering the evidence for reorganization across modalities in humans, including tactile and auditory skills in the blind and visual abilities in the deaf. Finally, plasticity within the language system is considered.

I. HUMAN BRAIN DEVELOPMENT: SIMILARITIES AND DIFFERENCES BETWEEN HUMAN AND NONHUMAN PRIMATES

The observation that the programs that govern development are remarkably conserved throughout phylogeny has led to the wide acceptance that human brain development goes through stages of development similar to those described for nonhuman primates. Whereas the earliest stages of development show limited room for plastic changes, plasticity naturally mediates the later stages of development as early experience in part controls the modulation of the connectivity between neurons. In most species, including humans, the pattern of connectivity that emerges as a result of early prenatal stages of development is only a rough sketch of the final wiring. Postnatally, the pattern of connectivity is refined through progressive and regressive events during which axons, dendrites, synapses, and possibly neurons show exuberant growth and major loss, leading to remodeling of the neural circuitry. The extensive changes in connectivity observed during that stage of development are believed to be limited to certain time periods. Reshaping of the connectivity during this period of time appears to be mediated by mechanisms similar to those observed at other ages, particularly activity-dependent mechanisms that are modulated by experience. However, the amplitude of the changes may be considerably greater in the immature system.

A. Role of Experience in Development

The role of sensory experience during sensitive periods has been documented in numerous animal studies. Experience in part controls the selection of axons, dendrites, synapses, and neurons that will form the functional neural circuits. For example, during the sensitive period for ocular dominance, visual deprivation induced by monocular eyelid suture results in shrinkage of the ocular dominance columns serving the closed eye. Outside the sensitive period, visual deprivation has little effect on the pattern of ocular dominance. There is good agreement in the literature that experience affects the organization of local circuits rather than major pathways because the main topographical and columnar organization of the cortex has already been achieved by the time most sensitive periods have been reported to occur. Experience has also been implicated in the onset of sensitive periods. In cats, rearing in the dark results in delayed onset of the sensitive period for ocular dominance formation. Similar observations of the role of experience have been made in the auditory systems of songbirds and humans. For example, the maturation of an early auditory evoked response displays an extended time course of development after cochlear implantation in congenitally deaf children. In this case, the number of years of auditory experience, rather than chronological age per se, was predictive of the maturational time course, even in individuals implanted during adulthood. Whereas this result indicates that this early auditory response retains its capacity to mature throughout life, most systems typically exhibit limits on the period of time when experience leads to normal maturation. Kittens raised with both eyes sutured at the age of 1.5–16 months show significantly reduced contrast sensitivity, human babies with bilateral cataracts show little recovery if operated on after 2 years of age, and children exposed to a natural language for the first time around puberty fail to master that language.

Evidence indicates that some sensitive periods are mediated, at least in part, by changes in the sensitivity of the *N*-methyl-D-aspartate (NMDA) receptors; in particular, these receptors require stronger input for learning to proceed as a sensitive period unfolds. This change in sensitivity occurs over time but also appears to be modulated by experience. For example, the visual cortex of older, binocularly deprived cats exhibits the same large NMDA component in their response to light as younger kittens

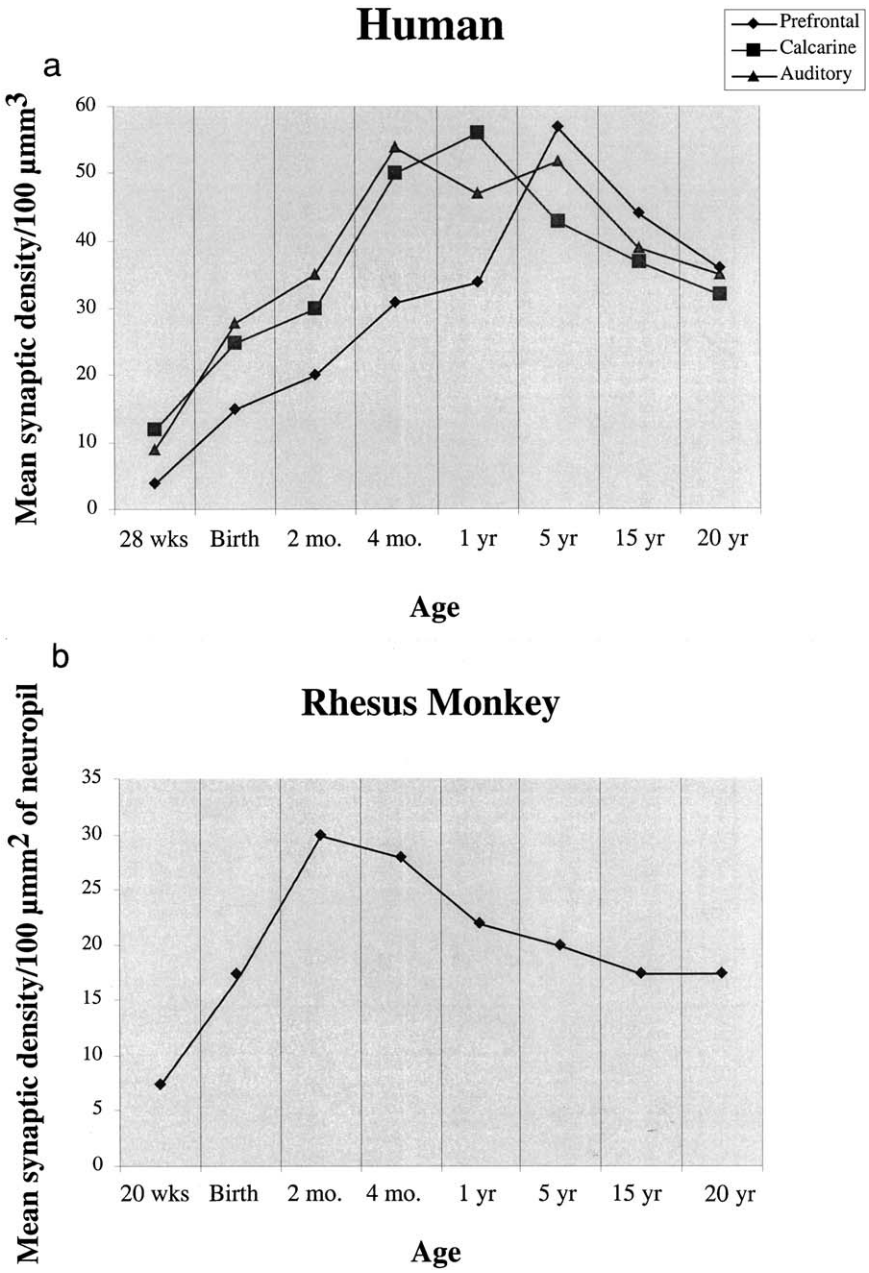


Figure 1 (a) Mean synaptic density in synapses/100 μm^3 in visual, auditory, and prefrontal cortices at various ages during human development. The data suggest regional differences in synaptogenesis [adapted from Huttenlocher and Dabholkar (1997)]. (b) In contrast, the mean synaptic density in synapses/100 μm^2 of neuropil was found to be comparable across visual, somatosensory, motor, and prefrontal cortices during development in monkeys [adapted from Rakic, Bourgeois (1986)].

that have had a similar amount of exposure to light, indicating that visual experience affects this change. Thus, the maturational status of the brain is affected by both chronological age and experience, illustrating the inextricable roles of nature and nurture during the course of development.

B. Mechanisms of Plastic Changes during Development and Learning

Although little is known about the factors that control the duration and timing of sensitive periods,

it is commonly believed that they are closely related to synaptogenesis, i.e., the phase of overproduction of synapses observed during cortical development. Studies of nonhuman primate development have indicated that about 35% of axons–neurons are lost from the peak of development to adulthood and that the onset of this loss co-occurs with the period of synaptogenesis. Similar studies conducted on human brain tissue also document periods of overproduction of synapses. Researchers have used electron microscopy to map out the synaptic remodeling that occurs during human development. They have compared synaptogenesis and synapse elimination within several different brain areas. In primary visual cortex, a burst in synaptogenesis occurs at about 3–4 months of age with maximum density reached at 4 months. In contrast, synaptogenesis in the middle prefrontal cortex takes longer, reaching maximum synaptic density at about 3.5 years of age. Furthermore, the same researchers have shown that the time course for synapse elimination occurs significantly later in the middle frontal gyrus (until 20 years of age) than in the primary visual cortex (converged on adult levels by age 4 years; see Fig. 1). These findings suggest that brain areas differ in their rate of development. The anatomical measures of synapse proliferation in the human brain describe developmental time courses similar to those observed in physiological studies using positron emission tomography (PET) with fluorodeoxyglucose (FDG), a technique that traces glucose metabolism. These studies show a rapid rise in cerebral metabolism during infancy, perhaps reflecting the burst of synaptogenesis described in the structural studies. This is followed by a decrease in brain glucose metabolism later in childhood, with a timing sequence similar to that observed for the loss of synapses. In the metabolic studies, increased glucose metabolism is observed in primary sensory and motor cortices, the hippocampal region, and the cingulate cortex before other cortical regions; one of the latest structures to show increased glucose metabolism is the prefrontal cortex. These structural and physiological findings support the view of different maturational timetables for distinct brain structures, with primary cortices developing before higher association cortices in humans.

The proposal that cortical maturation occurs earliest in sensory–motor cortices, then in brain areas mediating language, and last in frontal areas mediating complex social or planning behavior is tantalizing, because this sequence mimics in part the time course of sensory, motor, and cognitive development in the child. There is, however, much debate about the link between synaptogenesis and behavioral stages. First, whereas the findings in humans suggest a different time course of synaptogenesis for different brain systems, detailed studies by researchers have reported the concurrent onset of synaptogenesis across different brain areas in macaque monkeys. Although this may indicate different constraints during development across species, the source of this difference may be due to the technical difficulties encountered when working with human tissue and/or to the compressed rate of maturation in nonhuman primates. Furthermore, and more importantly, it is unclear whether synaptogenesis should be considered as a reliable marker of cortical maturation. The few studies of cortical development that consider the development of the connectivity of separate cortical layers within a region establish that synaptogenesis can only be part of the story. For example, detailed studies of the establishment of local connectivity in human V1 indicate that the development of intralaminar connections within intermediate layers of V1 overlaps with the period during which the overall synaptic density of V1 has been described to decrease. Clearly, an undifferentiated measure of synaptic density cannot be used as the sole marker of cortical maturation. Only by characterizing the development of different subsets of projections can one hope to link the emergence of cortical circuits with the maturation of behavioral functions.

So far, we have considered plasticity mediated by changes in connectivity as axons, dendrites, or synapses form or disappear; however, plastic changes may also be mediated by the generation of new neurons. The preeminent view has been that neurogenesis, or the period of time during which new neurons are generated, is restricted to a short period of development between embryonic days 42 and 120 in humans. This view has been challenged on two grounds. Reanalysis of the longitudinal data set of human brains collected by Conel has led some to propose that there is a 2-fold increase in the number of neurons between 15 months and 6 years of age. Although methodological considerations prevent a firm conclusion on these grounds, there is now agreed-upon evidence that neurogenesis can occur in adult primates, including humans. Neurogenesis has been documented in the human dentate gyrus as well as in the association cortices of nonhuman primates, opening new perspectives on neuroplasticity.

C. Summary

There is common agreement that experience plays a critical role in shaping the organization of the brain. We have seen that a number of brain systems show sensitivity to experience only during limited time periods, called sensitive periods. This is a time period during which experience may profoundly and durably alter the organization of that system. In contrast, other systems display life-long sensitivity to experience. There is ample evidence that similar molecular and cellular mechanisms are at play to mediate these different kinds of plastic changes. Why then does plasticity during sensitive periods appear to be more pervasive than at other stages of life? Though there is no clear answer to that question, a few hints are available. First, there appear to be molecular and cellular differences between the immature and later stages of life. For example, the NMDA receptor that mediates learning in adults is also found in young animals but in a form that facilitates learning and plastic changes. The NMDA receptor in young animals is triggered more easily and leads to larger neuronal responses than in adults; these properties suggest that learning is likely to happen more readily in young animals. It has also been proposed that these properties may be especially suited to maintain synapses in regions that do not receive well-correlated inputs, in effect keeping the multipotentiality of the young cortex. This view is supported by computational models of development, which illustrate that the redundancy of the connectivity at the onset of sensitive periods may allow for easier learning and adaptation than what can be achieved once the architecture of the brain has committed to a given pattern of organization. Thus, even if some of the mechanisms are similar or even identical between developmental and adult plasticity, the fact that these mechanisms operate on nervous systems that are structurally and physiologically different may account for the observation that plastic changes in development often become irreversible, whereas plastic changes in adults often are reversible provided adequate training.

II. WITHIN-MODALITY PLASTICITY

This section will focus on early changes in visual experience, such as strabismus, and how they affect the development of visual functions and, ultimately, their organization in the adult brain. We review evidence that suggests that, because different visual functions have different developmental timetables, the period of time during which optimal plasticity is observed is specific to each visual function. The relationship between developmental events and plastic changes is first considered by reviewing fast-developing visual functions such as orientation selectivity or stereopsis. Functions with a more protracted period of development are then considered. A major consequence of these differences in developmental time course is that the timing and nature of the experience that leads to plastic changes differ for different visual functions.

A. Development and Plastic Changes in Orientation, Motion, and Stereopsis

There is abundant evidence that the development of most visual functions in humans proceeds at a fast pace during the first 2 years of life, with different maturational schedules for different functions, such as orientation, motion, and stereopsis.

Although the human newborn is able to perform gross discriminations of orientation, it is commonly accepted that orientation selectivity appears around 3 weeks of age in human babies. Most human babies can distinguish gratings of different orientation at around 6 weeks of age, and this skill appears to be mature by 12 weeks of age. Orientation selectivity is largely mediated by the organization of thalamic projections in layer 4 of the primary visual cortex. In humans, as in other primates, these projections are among the first part of the visual cortical pathway to develop, showing a largely mature pattern of organization just before birth. In accordance with the principle that early developing thalamocortical pathways show little sensitivity to experience, it has proven very difficult to alter orientation columns at the level of layer 4 in animal studies. In one study, kittens reared in a striped environment showed an over-representation of the rearing orientation as measured by optical imaging; however, orientations that had never been shown were still represented, suggesting that visual experience can only partially alter orientation maps. Thus, orientation selectivity, which develops very early, can only be partially affected by experience.

Less is known about the development and plasticity of the motion pathways. The connectivity in V1 layer 4B is believed to control the maturation of motion perception because this layer projects heavily to MT–MST, the cortical area specialized for motion processing. This connectivity appears to be mature at 8 weeks in human babies. Whereas this provides adequate

timing for the onset of motion direction discrimination in babies, little is known about the developmental timetable of the motion complex MT–MST, which mediates motion discrimination in adults. Studies that probe the developmental time course of motion pathways have often relied on the analysis of motion tracking. Newborns can easily track a slowly moving target. However, at that age, pursuit is not entirely smooth but, rather, saccadic. It is only at 2 months of age that babies exhibit the smooth pursuit typical of adults. Research shows that this ability is impaired by early altered experience such as strabismus. This may result from alterations in the development of the cortical MT–MST complex or of the subcortical pathways involved in the control of eye movements. An animal model of motion processing after early strabismus indicates modifications in the visual pathway as early as V1. Further research is needed, however, to map out the plasticity of the different brain mechanisms that mediate motion perception.

The onset of stereopsis has been reported around 10 weeks in humans; most babies display stereo depth perception capabilities around 21 weeks and a rather mature pattern of stereovision at 26 weeks. This sequence of events is rather well-matched with what is known about the maturation of the brain architecture that mediates stereopsis. The development of stereopsis is one of the most studied models of visual plasticity. Computation of stereo information requires the preservation of eye of origin information, and it is widely accepted that this cannot be achieved before the initially overlapping connections from the two eyes become anatomically segregated to form ocular dominance columns (ODC). In a seminal study, researchers showed that visual experience plays a key role in ODC formation by establishing that monocular eyelid suture results in the shrinkage of the ODC serving the closed eye. Work on cats suggests two distinct mechanisms at play for plastic changes in ODC: (i) a competitive activity-dependent mechanism controlling the competition between left and right eye input that is affected by unilateral deprivation and (ii) an absolute activity-dependent process that controls the sensitivity of the visual pathway to visual information and that is triggered as soon as the eye receives a patterned input. Accordingly, the abnormal pattern of ODC after binocular deprivation indicates that visual experience is important for normal development, and the greater abnormalities noted after monocular deprivation stress the central role of interocular competition in the formation of ODC.

Early medical conditions in humans such as strabismus or cataracts result in deficits similar to those noted in animal studies. In particular, greater abnormalities in visual functions are noted after monocular congenital cataracts than after bilateral ones. This effect is so potent that it has become normal practice to recommend the patching of the good eye for 50–90% of the time in patients operated for unilateral cataracts to encourage the use of the previously deprived eye. Additionally, visual deprivation has dramatic effects on stereopsis. Children treated for congenital cataracts display little measurable stereopsis, a finding consistent with observations in animal models of a loss of neurons tuned to binocular disparity and a reduction in cells that respond to both eyes after deprivation by lid suture. Interestingly, these individuals (who still have strabismus in adulthood) report deficits of binocular integration and complain of visual suppression or interference between contours in the image. Again, this finding is in line with the animal literature. The strong binocular facilitation observed when moving bars of similar orientation are presented simultaneously to corresponding locations in the two eyes is significantly reduced or even turns into suppression in cats with strabismus, providing a possible neural basis for the visual suppression reported in strabismic humans.

In accordance with the developmental timetable of stereopsis, the percentage of human infants demonstrating stereopsis in a control group and in a group with early strabismus was comparable until about 4 months of age. During the following months, the proportion of normal infants demonstrating stereopsis increased whereas that of infants with strabismus decreased. These data suggest that the neural substrate that mediates stereopsis is present and functional at the onset of the sensitive period for stereopsis but is subsequently lost because of abnormal visual experience.

The finding of a restricted period of time during which stereopsis can be altered or corrected in humans is consistent with animal studies showing that the effect of experience on ODC is restricted to a sensitive period. Little effect of experience is noted on the pattern of ODC when deprivation occurs after that period, although some deprivation-induced changes can be shown in cortical cell physiology. In humans, normal ODCs have been reported in an individual who developed a convergent squint after 2 years of age, suggesting that the sensitive period for ODC may be over by that time. Similarly, individuals who developed cataracts late (after 6 years) show evidence of

binocular functions when tested with displays that appropriately correct for their low acuity and the presence of strabismus. Thus, as in animal studies, the time course of the ODC development in humans appears to be tightly coupled to the sensitive period for stereopsis. Although it was first thought that the end of the sensitive period for stereopsis corresponded to the time of full development of the ODC, it is now clear that the period of development of ODC corresponds to the most sensitive part of the sensitive period, but that the sensitive period extends beyond the full formation of ODC. This may explain some of the variability reported in humans for the oldest age at which corrective surgery for cataracts and/or strabismus is successful.

B. Protracted Visual Functions: Opening New Doors to Plastic Changes

Though so far we have concentrated on visual functions that have a short and well-defined maturational time course, other visual functions, such as visual acuity, long-range orientation integration, or orientation toward peripheral space, seem to have a protracted time course of development and a much longer period of sensitivity to altered visual experience.

Grating acuity or the capacity to discriminate sinusoidal gratings from noise is extremely low at birth; it develops rapidly during the first 6 months of life, but then continues to improve very slowly until about 6 years of age. Although adultlike performance is achieved around that age, it is not until about 10 years of age that visual acuity is no longer sensitive to altered visual experience. Early visual deprivation such as cataracts or strabismus results in enduring loss of visual acuity. Researchers have shown that tests of visual acuity just after corrective surgery indicate acuity within the range of newborns, even when deprivation lasted up to 9 months. Within an hour, the acuity of the treated eye was shown to improve dramatically to the level of a normal 6-month-old infant. These results highlight the triggering role of patterned visual input in the development of visual acuity. The timing of that patterned input is critical, however. Although the grating acuity of children treated within their first year for cataracts or strabismus continues to improve with age, it does not keep pace with normal development and falls below normal limits as development proceeds. Whereas this effect is quite similar for unilateral and bilateral deprivation, the effects of early unilateral deprivation appear to be more drastic later in life than those of bilateral deprivation. For example, acuity as measured by the ability to recognize or match letters (e.g., Snellen chart) is much worse after unilateral cataracts than bilateral cataracts. Additionally, the outcome after bilateral cataracts shows little sensitivity to the duration of deprivation. In contrast, whereas there is no effect of deprivation duration after unilateral cataracts if operated early (before 6 weeks of age), deprivation duration affects the outcome of recovery in cases of unilateral cataracts treated after 6 weeks. By this age, uneven competition between the two eyes appears to limit recovery. Accordingly, the patching strategy adopted after surgery has been found to be a major determinant of treatment outcome. These findings support the view that plastic changes in visual acuity are mediated by at least two separate mechanisms with different time courses. One mechanism is driven by the absolute amount of patterned input to which the visual system has access; this mechanism is present at the onset of development and probably throughout development. The other mechanism, which emerges only later in life, appears to be driven by the competition between the two eyes and quickly becomes the main predictor of recovery.

Another skill with a protracted time course of development is the capacity to integrate long-range spatial information. For example, the detection of contours that are defined solely on the basis of long-range orientation domain correlation has been found to improve throughout childhood, reaching adult levels around 14 years of age (Fig. 2). This skill requires the integration of visual information across space and is believed to rely in part on the integrity of long-range horizontal connections in layers 2 and 3 of the visual cortex. This skill is also jeopardized by alterations of visual experience such as in amblyopia. Little is known at this time about the time frame during which visual deprivation affects that skill, but its protracted period of development suggests that it may be influenced by altered visual experience throughout childhood.

A similarly protracted visual function is the ability to orient to peripheral information. Whereas the ability to detect large, bright peripheral events is present in the newborn, the automatic ability of adults to detect pinpoints of light at the far edge of the visual field develops slowly, reaching adult levels by 12–14 years of age. In accordance with its slow developmental time course, this skill is altered when deprivation occurs not only early in life but also at 6 years of age.

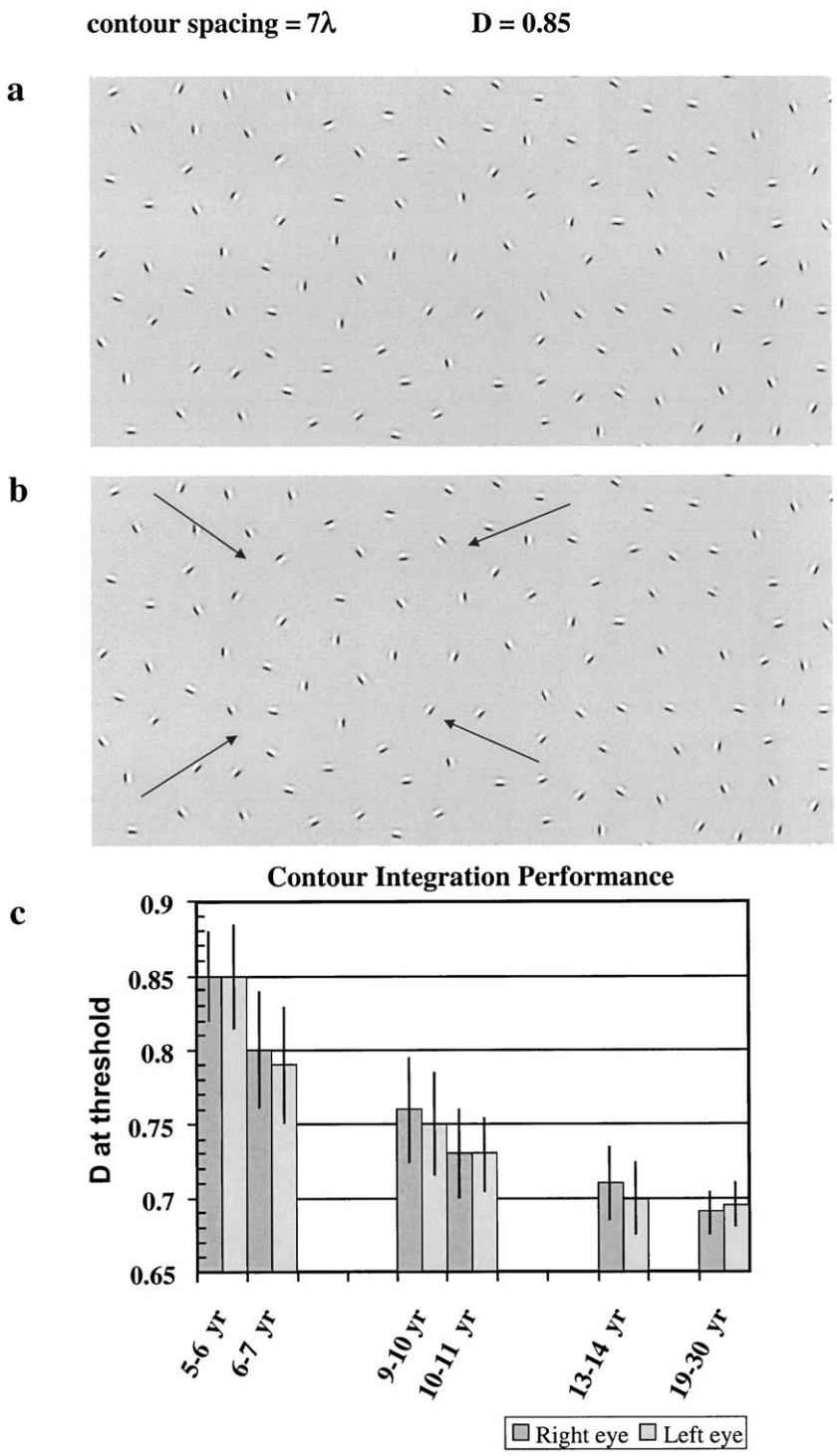


Figure 2 (a) Examples of stimuli used to study visual–spatial integration of long-range orientation. Subjects were presented with displays like the one in (a) and asked to delineate the elements that form a contour [answer given in (b)]. The difficulty of the task was manipulated by varying the spacing and numbers of elements used. (b) The contour present in (a) is indicated by arrows. (c) The developmental time course of this skill appears to be protracted, with adult performance only seen at 13–14 years of age. Reprinted from Kovacs, Kozma *et al.* (1999), Late maturation of visual spatial integration in humans. *Proc. Natl. Acad. Sci. USA* **96**(21), 12204–12209, copyright 1999, National Academy of Sciences, USA.

C. Summary

Studies of visual plasticity in humans suggest constraints in human developmental plasticity similar to those unveiled by animal studies. (i) Visual experience during sensitive periods plays a central role. As a result, functions that have a protracted sensitive period have more opportunity to be altered by experience. (ii) There is a close correspondence between the onset of the sensitive period for a skill and the development of the neural circuitry that mediates that skill. (iii) There are different developmental time windows for different visual functions. As a result, altered visual experience affects various visual functions and systems differently, preventing gross generalization from system to system and indicating the need for separate studies of each of these systems.

III. BETWEEN-MODALITY PLASTICITY

Whereas anecdotal evidence of better audition after early blindness or better vision after early deafness exists, the available data are quite mixed. The long-held belief that multisensory integration is a necessary step in optimal development led investigators to focus initially on the disabilities caused by early blindness or deafness. For example, a number of studies, mostly from the 1970s and the early 1980s, document deficient spatial abilities in the blind and deficient visual perception in the deaf. The realization of the adaptability of the brain led investigators to carefully review this issue. It is now evident that when the etiology and characteristics of the population tested are carefully controlled and the task is appropriately chosen not to rely on encoding strategies that are not available to the deprived subjects, convincing evidence of compensatory plasticity can be established.

The mechanism at play in compensatory plasticity is still a question of debate. Whereas within-sensory modality changes appear to be mediated by local changes within a limited set of cortical areas, between-modality reorganization is often thought to imply more drastic modifications of connectivity across areas. Most animal models of cross-modal plasticity have surgically forced the input of one modality to be rerouted to the primary cortex of another modality. For example, researchers have shown that, when retinal inputs are rewired to innervate the auditory thalamus and, therefore, to project to the auditory cortex in ferrets, cells in the primary auditory cortex exhibited sensitivity to visual stimulation and a substantial degree of orientation selectivity and can be used to mediate visual orientation behavior. Studies of cross-modal plasticity in animal models suggest that, at least early in development, cortical areas can change their functional specificity depending on which inputs they receive. Can such cross-modal rewiring be observed in humans? There is very little evidence to date on the question of whether compensatory plasticity, in the absence of surgically induced rewiring, can occur in primary cortices in humans or animals. However, there is converging evidence that secondary and association cortices can compensate for the loss of one modality. This rewiring is believed to involve the stabilization of early multisensory corticocortical connections that exist in the newborn but are eliminated in the presence of normal input. For example, researchers have shown that the associative cortical area 19, which is predominantly visual in normal animals, becomes more sensitive to tactile information in monkeys deprived of vision early in life. In the absence of competition from visual inputs, early transient connections may remain and recruit the visual cortex for tactile and/or auditory processing. Evidence supporting this view in humans is reviewed next.

A. Effects of Congenital Blindness on Touch

A number of studies have confirmed the functional participation of visual areas during somatosensory tasks in early blind individuals. By using PET, researchers compared tactile discrimination in early blind Braille readers and control subjects. Blind subjects revealed the activation of visual cortical areas, whereas these regions were deactivated in controls. The functional relevance of visual areas in tactile discrimination was further established in a transcranial magnetic stimulation experiment. Transient stimulation of the occipital cortex induced errors on a tactile task in early blind subjects but had no effect on the sighted controls. It is worth noting that not all aspects of somatosensory processing recruit visual areas in blind subjects. For example, simple tactile stimuli that did not require discrimination produced little activation in the visual areas of blind subjects. This finding is in agreement with the hypothesis that different neurocognitive systems and subsystems exhibit different sensitivities to altered experiences.

B. Effects of Congenital Blindness on Audition

The few studies that have addressed the accuracy of sound localization in the blind report a mixed pattern of results. Whereas impaired performance in blind individuals has been reported, other studies report no change or enhancement in the blind. A closer look suggests that, whereas distance perception may be hindered in congenitally blind subjects, sound localization is either similar or enhanced in blind subjects. In particular, pointing accuracy was comparable in the blind and sighted for binaural targets and enhanced in the early blind for the localization of monaural cues. Unlike sighted subjects, which systematically failed to correctly localize the monaural sounds on the side of the obstructed ear, half of the blind subjects were able to do so quite accurately, suggesting that early blind individuals may use monaural cues more efficiently than sighted subjects. This point is further supported by the work of researchers who have studied auditory localization abilities in blind humans. Event-related potentials (ERPs) were recorded as congenitally blind adults and sighted controls attended to either central or peripheral sound sources in order to detect a rare noise burst at either the 0° or the 90° loudspeaker (on different blocks). Behavioral data revealed a higher spatial resolution in the blind when attending to the periphery. Gradients of ERP amplitudes suggested a sharper auditory spatial attention focus in the blind compared to the sighted when attending to the periphery. This result suggests that the representation of peripheral space may be more altered by early sensory experience than that of central space.

C. Effects of Congenital Deafness on Vision

Studies of visual functions after early genetic deafness indicate enhanced visual processing, at least for visual motion, visual attention, and peripheral vision. For example, deaf adults are better than hearing controls at detecting the onset or the direction of motion of a peripheral stimulus. They are also faster at switching visual attention toward a near-periphery target in the presence of distractors located at the fixation point. Electrophysiological recordings while subjects monitored moving stimuli have indicated larger visually evoked responses for deaf than hearing adults over occipital and temporal sites. These group differences were especially marked for peripheral stimuli. In an fMRI study, the effects of visual attention on motion processing were compared in deaf and hearing individuals. When participants monitored the peripheral visual field, greater recruitment of the motion-selective area MT was observed in deaf than in hearing participants, whereas the two groups were comparable when attending to the central visual field (Fig. 3). Further analysis suggested that changes in peripheral attention in the deaf are mediated through the modulation of the connections between earlier sensory areas and the posterior parietal cortex, which is one of the main centers of attention.

The functions altered in the deaf, i.e., motion, visual attention, and peripheral processing, share the property of being mediated predominantly by the dorsal visual pathway that projects from V1 to the motion area (MT–MST) and the parietal cortex. To test the specificity of dorsal pathway enhancement after early deafness, motion processing was compared to color processing. It is commonly accepted that, whereas motion processing is primarily mediated by the dorsal pathway, color is primarily mediated by the ventral pathway projecting from V1 to IT. Electrophysiological recordings while subjects monitored either high-spatial-frequency colored gratings (color) or low-spatial-frequency, gray scale, moving gratings (motion) were compared in deaf and hearing persons. Several specific group differences occurred in the amplitude and distribution of early sensory responses recorded over anterior and temporal regions. Deaf subjects displayed significantly greater amplitudes than hearing subjects, but this effect occurred only for moving stimuli not for color stimuli (Fig. 4). Further, whereas in hearing subjects, color stimuli elicited larger responses than did motion stimuli, in deaf subjects responses to motion stimuli were as large as those to color stimuli. These data suggest that larger changes in motion than color processing occur after early deafness. Taken together, the available data suggest that there is considerable specificity in the aspects of visual processing that are altered in congenitally deaf individuals.

Because many genetically deaf individuals learn American Sign Language (ASL) as a first language, some of the changes reported could have been due to deafness or to the acquisition of a signed language. Signing has been shown to affect performance on tasks that require visuospatial transformations and are likely to recruit the dorsal pathway. Deaf native signers are both faster and more accurate than controls on tasks of mental rotation or when identifying objects presented from a noncanonical viewpoint. This effect is not specific to auditory deprivation; hearing native

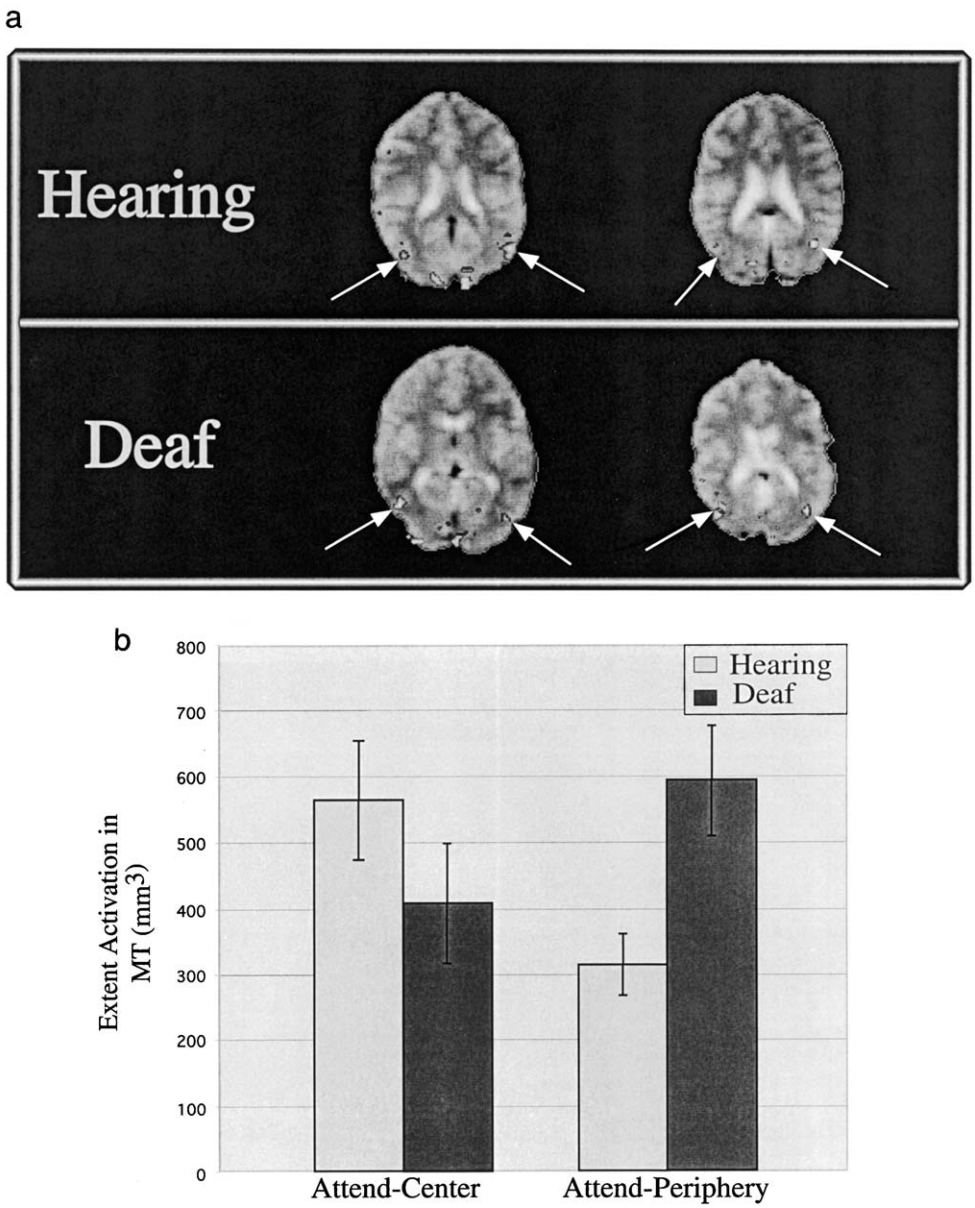


Figure 3 Extent of activation in MT–MST in deaf and hearing individuals as they monitored moving stimuli for luminance changes in either the center or the near periphery of the visual field. Enhanced recruitment of MT is observed for the peripheral condition in the deaf.

signers also exhibit better mental rotation performance than nonsigners. Researchers have also shown that deaf and hearing signers are faster at generating mental images than hearing nonsigners. This work establishes that familiarity with ASL results in behavioral enhancement in a number of visuospatial tasks that are likely to recruit structures in the dorsal pathway. Acquisition of ASL has also been linked with brain reorganization for motion processing. The few studies on that topic have reported a left hemisphere advantage for motion processing in native signers, whereas hearing nonsigners displayed a tendency for a right hemisphere advantage. Interestingly, the studies available indicate that the lateralization pattern for motion processing is guided by sign language acquisition, whereas the enhancement of motion processing in the visual periphery is specific to deafness. This research illustrates the specificity of the plastic changes as a function of the nature of altered experience.

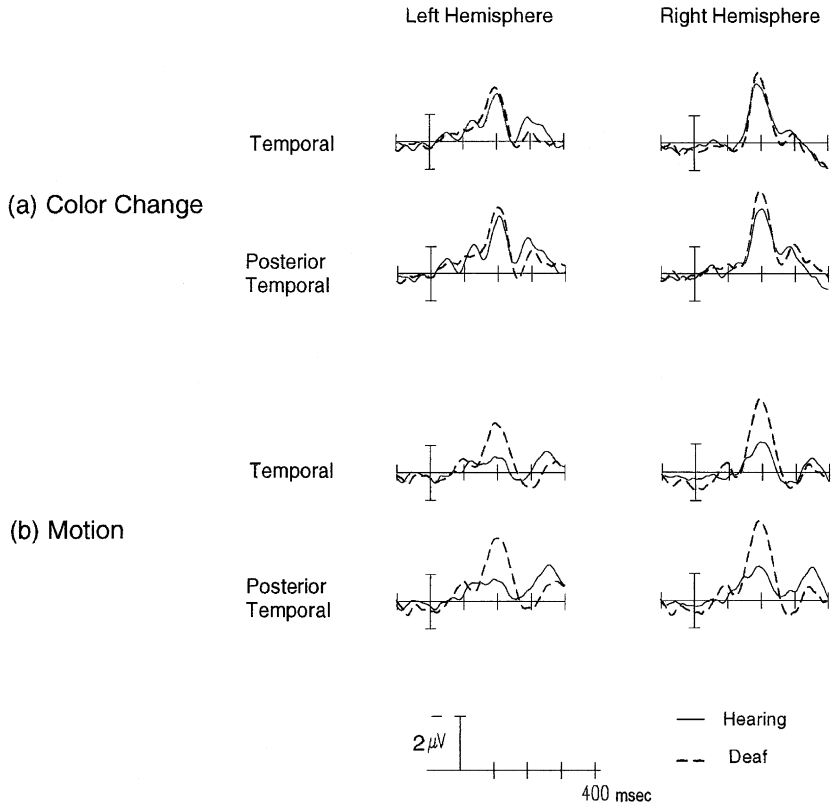


Figure 4 ERPs elicited by (a) color change and (b) motion in normally hearing and congenitally deaf adults. Recordings are from temporal and posterior temporal regions of the left and right hemispheres. Reprinted from Neville and Bavelier (1999), In "The New Cognitive Neurosciences," 2nd ed., M.S. Gazzaniga, (Ed.), pp. 83–98, with permission of MIT Press.

D. Summary

The existing literature suggests that, across auditory and visual modalities, the representation of peripheral space is more altered by early sensory experience than is the representation of central space. Close examination of the behavioral data for blind cats indicates a similar effect, i.e., a larger advantage in sound localization for blind cats at peripheral locations. Importantly, the available studies indicate that not all aspects of the remaining senses are altered after early blindness or deafness. Rather, different neurocognitive systems and subsystems exhibit different sensitivities to altered experience.

IV. PLASTICITY WITHIN THE LANGUAGE SYSTEM

It is reasonable to assume that the rules and principles that govern the development of the sensory systems also guide the development of language-relevant brain systems. Next, we briefly review facts about the normal development of language and the capacity for plasticity in the language system after altered experience and brain lesions.

A. Facts about the Normal Development of Language

Developmental studies indicate several milestones during language acquisition. By 12 weeks, most infants produce vowellike sounds called cooing. By 20 weeks, vocalizations begin to include more consonant sounds, a stage termed babbling. Whereas initially these vocalizations are very different from the sounds of the language environment, they come to resemble the syllables of the surrounding language by 8 months of age. Isolated words are produced around 1 year; these usually include common nouns that describe everyday objects or frequent social words such as hello. Sometime during the second year, vocabulary size

begins to grow at a dramatic pace and short, 2-word sentences appear. These sentences, although very brief, already display considerable structural information. Thus, young learners of English typically say "Daddy eat" or "eat pizza," showing sensitivity to the subject–object distinction. Between 3 and 5 years of age, patterns of grammatical development occur, including syntactic and morphological developments such as over-regularization (goed vs went), question formation, negation, and passive. The finding that these milestones are shared in all learners all over the world supports the proposal that language learning has a significant biological basis.

Indeed, several reports indicate that the milestones of language development are extremely robust in the presence of significant variations in experience. For example, the same stages of language development are observed across variations in the mother's speech. Children that are carried on their mother's backs and seldom talked to show the same milestones of language development as children raised in a more nurturing environment. Similarly, deaf children acquiring a visuospatial language also exhibit the same milestones. This is not to say that language experience does not guide language development. It is clear that the language environment determines which kind of language is learned. But a striking feature of language development (like visual development) is that the effects of experience are constrained by the age of the learner. For example, learners exposed to language late in life (whether second-language learners or individuals deprived of language exposure early in life) do not display the same level of proficiency as early learners. Thus, as in other domains, this research indicates that the effect of experiential influences during language development are constrained by the maturational status of the learner.

A few studies have begun to chart the changes in brain organization as children acquire their primary language and to ask whether different maturational constraints may exist for different types of processing as has been reported earlier for most sensory domains. One of the best-documented dichotomies in language processing is that between lexical–semantic information and grammatical information. For example, nouns and verbs (e.g., table, run) that convey semantic information elicit a different pattern of brain activity (as measured by ERPs) than do function words that convey grammatical relationships (e.g., within, out, the) in normal, right-handed, monolingual adults. Similarly, sentences that are semantically nonsensical (but grammatically intact) elicit a different pattern of ERPs than do sentences that contain a violation of syntactic structure (but that leave the meaning intact). These results are consistent with several other types of evidence that imply two separate systems within language, one for lexical–semantic and one for grammatical processing. Specifically, the studies available indicate a greater role for more posterior temporal–parietal systems in the left hemisphere for lexical–semantic processing; frontal–temporal systems within the left hemisphere are implicated for grammatical processing. How does this functional specialization arise during development? ERPs for open- and closed-class words have been compared in infants and young children from 20 to 42 months of age (Fig. 5). All children understood and produced both the open- and closed-class words presented. At age 20 months, ERPs in response to open- and closed-class words did not differ. However, both types of words elicited ERPs that differed from those elicited by unknown and backward words. These data suggest that in the earliest stages of language development, when children are typically speaking in single-word utterances or beginning to put two words together, open- and closed-class words elicit similar patterns of brain activity. At 28–30 months of age, when children typically begin to speak in short phrases, ERPs for open- and closed-class words elicited different patterns of brain activity. However, the more mature left hemisphere asymmetry for closed-class words was still not observed. By 3 years of age, most children speak in sentences and use closed-class words appropriately to specify grammatical relations, so that like adults, ERPs from 3-year-olds displayed a left hemisphere asymmetry to closed-class words. The results across the three groups are consistent with the hypotheses that, initially, open- and closed-class words are processed by similar brain systems and that these systems become progressively more specialized with increasing language experience.

B. Plasticity in the Language System after Altered Language Experience

1. Delayed Exposure to Language

How is language learning affected by delayed exposure to a language? Many investigators have studied this question by comparing individuals who learned a second language at different times in development. English–Korean speakers who came to the United States and were immersed in English at varying ages have been studied. Their data clearly establish a steady decrease in performance as a function of age at first

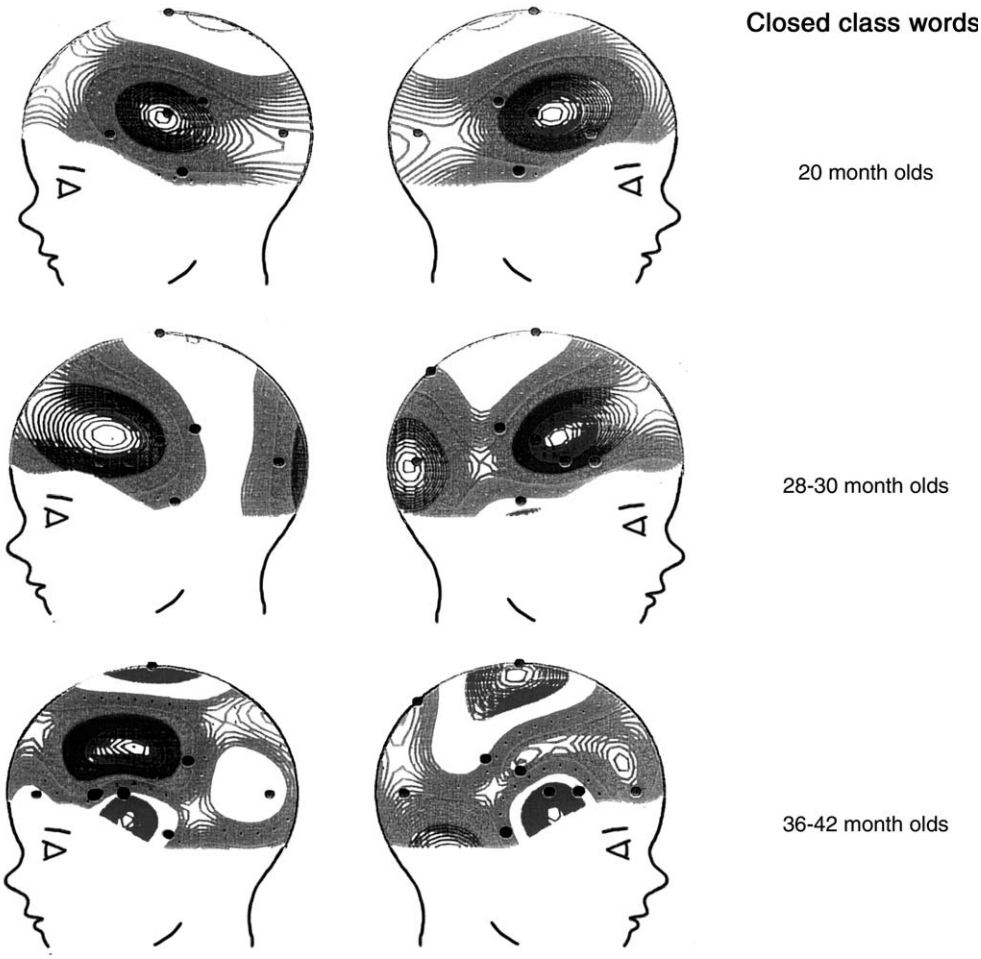


Figure 5 Current source density (CSD) analyses of neural activity in response to closed-class words at 200 msec. The CSDs illustrate sinks, i.e., activity flowing into the head (purple), and sources, i.e., activity flowing out of the head (orange), at three age groups. Top: At 20 months the CSD shows sinks over both the left and right hemispheres. Middle: At 28–30 months the CSD show sinks that are bilateral but slightly more prominent over the right than the left hemisphere. Bottom: At 36–42 months the CSD shows a sink over left anterior regions. Reprinted from Neville and Bavelier (1999), In "The New Cognitive Neurosciences," 2nd ed., M.S. Gazzaniga, (Ed.), pp. 83–98, with permission of MIT Press.

exposure (Fig. 6). The effect of age of exposure was especially striking for those aspects of language that rely on grammatical processing. Interestingly, duration of exposure cannot compensate for this effect; thus, a 14-year-old who has been immersed in English for only 4 years is more likely to show a good command of English intricacies than a 50-year-old who has been immersed in English for the last 25 years. It has been argued that the difference in late learners of a second language is due to competition from their extensive knowledge of a first language rather than their late start in learning. The observation of parallel results in deaf adults exposed at varying ages to their first and unique language (American Sign Language) suggests, however, that age of exposure is the key factor in this pattern of results. As in second language learners, late learners of American Sign Language (ASL) had particular problems with the ASL equivalent of morphemes and syntactically complex sentences.

The available results suggest that aspects of semantic and grammatical processing differ markedly in the degree to which they depend upon language input. Specifically, grammatical processing appears to be more vulnerable to delays in language experience. In Chinese–English bilingual speakers, delays of as long as 16 years in exposure to English had very little effect on the organization of the brain systems active in lexical–semantic processing. In contrast, delays of only 4 years had significant effects on those aspects of brain organization linked to grammatical processing. Further evidence on this point was provided by ERP

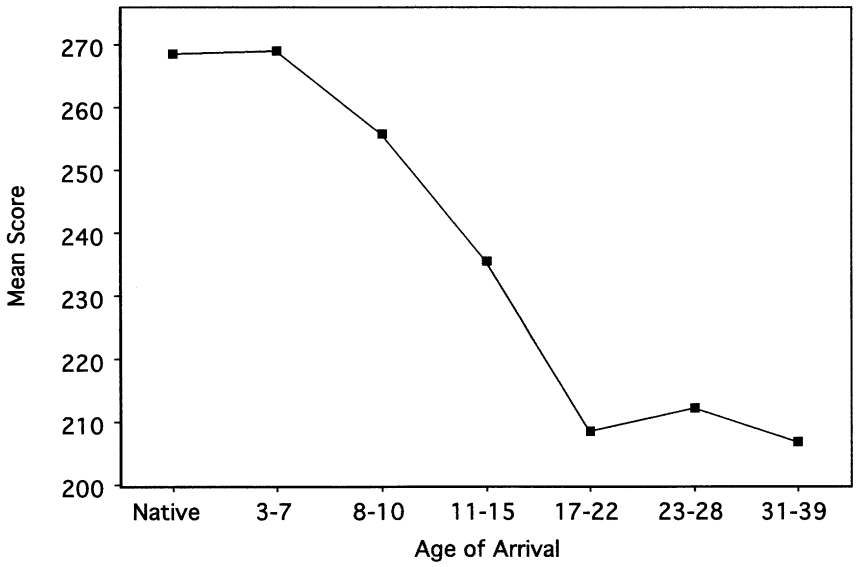


Figure 6 Total score on a test of English grammar in relation to age of arrival in the United States. These data show that the participants' performance declines as the age of exposure to the language is delayed. Reprinted from Johnson and Newport (1989), *Cog. Psych.* **21,** 60–99, with permission of Academic Press and Erlbaum Inc.

studies of English sentence processing by congenitally deaf individuals who learned English late and as a second language (ASL was the first language of these subjects). Deaf subjects displayed ERP responses to nouns and semantically anomalous sentences in written English that were indistinguishable from those of normal hearing subjects who learned English as a first language. These data are consistent with the hypothesis that some aspects of lexical–semantic processing are largely unaffected by the many differences in language experience between normally hearing and congenitally deaf individuals. By contrast, deaf subjects displayed aberrant ERP responses to grammatical information like that presented in function words in English. Specifically, they did not display the specialization of the anterior regions of the left hemisphere characteristic of native, hearing–speaking learners (Fig. 7). These data suggest that the systems that mediate the processing of grammatical information are more modifiable and more vulnerable in response to altered language experience than are those associated with lexical–semantic processing.

2. Visuospatial Languages

We have employed ERP and fMRI techniques to further pursue the preceding hypothesis and also to obtain evidence on the question of whether the strongly biased role of the left hemisphere in language occurs independently of the structure and modality of the language acquired first. ERPs recorded from native signers in response to open- and closed-class signs in ASL sentences displayed timing and anterior–posterior distributions similar to those observed in native speakers processing English. However, whereas in native speakers of English responses to closed-class English words were greatest over anterior regions of the left hemisphere, in native signers closed–class ASL signs elicited activity that was bilateral and that extended posteriorly to include parietal regions of both the left and right hemispheres. These results imply that the acquisition of a language that relies on spatial contrasts and the perception of motion may result in increased recruitment of right hemisphere regions into the language system. Both hearing and deaf native signers displayed this effect. However, hearing people who acquired ASL in their late teens did not show this effect, suggesting there may be a limited time (sensitive) period when this type of organization for grammatical processing can develop. By contrast the response to semantic information was not affected by age of acquisition of ASL, in keeping with the results from studies of English that suggest these different subsystems within language display different degrees of sensitivity to altered experience.

In an fMRI study comparing sentence processing in English and ASL, researchers observed evidence for biological constraints and effects of experience on the

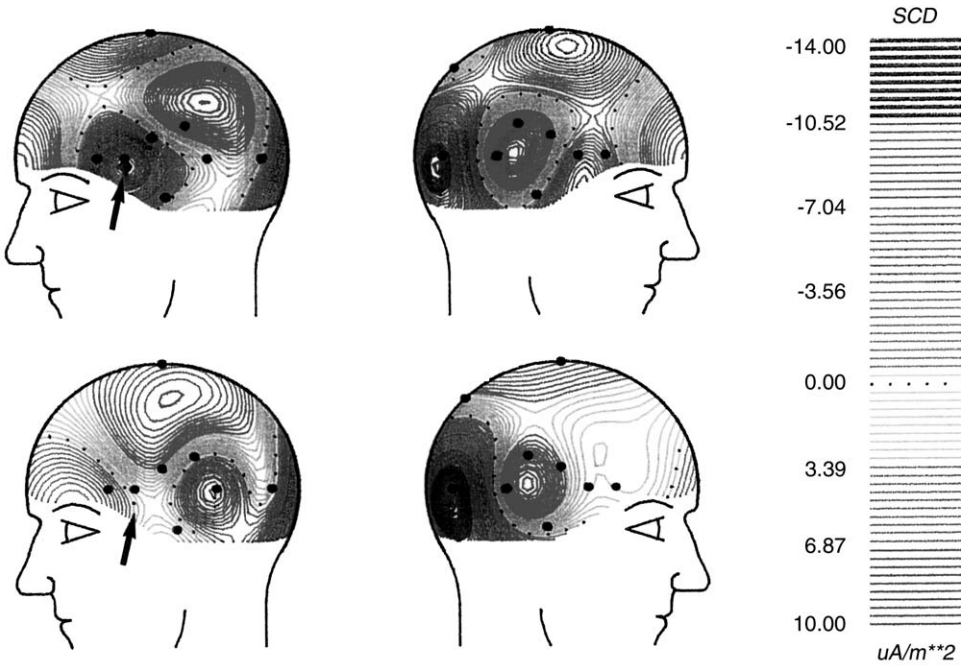


Figure 7 Distribution of current flow for the N280 peak elicited by closed-class words in English. Maps in the top row show the prominent response in the left hemisphere of hearing subjects (blue, marked by arrow). Bottom row maps display results from congenitally deaf individuals who lack the response (arrow). Reprinted from Neville, Mills, and Lawson (1992), *Cerebral Cortex* **2,** 244–258, with permission of Oxford University Press.

mature organization of the language systems of the brain. When hearing adults read English (their first language), there is robust activation within the left (but not the right) hemisphere and in particular within the inferior frontal (Broca's) region. When deaf people read English (their second language, learned late and imperfectly), there is no activation of these regions observed within the left hemisphere. Is the lack of left hemisphere activation in the deaf linked to the lack of auditory experience with language or to incomplete acquisition of the grammar of the language? ASL is not sound-based but displays each of the characteristics of all formal languages, including complex grammar (that makes extensive use of spatial location and hand motion). Studies of the same deaf subjects when viewing sentences in their native ASL clearly showed activation within the same inferior frontal region of the left hemisphere that is active when native speakers of English process English. These data suggest that there is a strong biological bias for these neural systems to mediate grammatical language regardless of the structure and modality of the language acquired. However, if the language is not acquired within the appropriate time window, this strong bias is not expressed. Biological constraints and language experience interact epigenetically as has been described for the other systems mentioned earlier.

The fMRI data also indicate a robust role for the right hemisphere in processing ASL. These results suggest that the nature of the language input shapes the organization of the language systems of the brain. Further research is necessary to specify the different times in human development when particular types of input are required for optimal development of the many systems and subsystems important in language processing.

C. Plasticity in the Language System after Brain Lesions

Studies of cerebral organization for language in the adult imply a greater role for perisylvian regions within the left hemisphere in language processing. This overall pattern appears ubiquitous in adults, and many investigators have suggested that the central role of the left hemisphere in language processing is strongly genetically determined. Certainly the fact that most individuals, regardless of the language they learn, display left hemisphere dominance for language in-

dicates that this aspect of neural development is strongly biased. Nonetheless, it is likely that language-relevant aspects of cerebral organization are dependent on and modified by language experience.

Some investigators have studied this question by assessing the language skills of children with left or right hemisphere lesions. Unlike what is observed in adults after lesions, children with left hemisphere damage (LHD) do not necessarily show poor verbal IQ and those with right hemisphere damage (RHD) poor pictorial IQ. Upon close examination, some language deficits can be observed after LHD. Most common problems include weak fluency, poor scores in formulating sentences giving oral directions or recalling sentences, and difficulties in written language particularly with spelling and production. They also show delays in language development, such as delays in prelinguistic development like babbling or in vocabulary acquisition. However, these children do not display the typical symptoms of aphasia reported in LHD adults. Although there are sizeable individual differences in the verbal IQs of children with LHD, overall they show receptive skills within normal limits and no major impairment in clinical evaluations; most are able to read and can create connected discourse. The interpretation of these findings is, however, complicated by the large interindividual variability in language performance found in that population, by the fact that most of these children function in the range of mental retardation (allowing confounding between intellectual functioning and language functioning), and by the mixing of children whose lesions occurred at different times.

A more recent study that included only children with early focal brain injuries indicates that language-relevant aspects of cerebral organization are dependent on and modified by early experience. First, this study confirms that LHD in children does not lead to the pattern of aphasia observed in adults. In fact, the early stages of language development (10–17 months) were more likely to be delayed by RHD than by LHD. Second, frontal damage was associated with greater delays in the emergence of expressive grammar that occurs around 2 years of age; however, unlike in adults, this effect was as severe after RHD as after LHD. Finally, damage to the left temporal lobe (but not the right one) led to delays in word production and the emergence of grammar. Although this disadvantage could be observed until 6 years of age, it was not detectable after 6 years of age, suggesting that significant compensatory plasticity had occurred by that age (although it is unclear whether differences would still be observed with finer tests of language abilities). These data are consistent with the proposal that the early stages of language development are represented more bilaterally and that the left hemisphere specialization for language is expressed only slowly over the course of development.

The striking result of lesion studies, however, is that the outcomes of a left hemisphere lesion are much better in infancy than in adulthood. The finding that plasticity for language decreases between birth and adulthood is consistent with the findings described earlier showing that the age of the learner affects language acquisition. It is likely that, as in other systems, nature and nurture continue to conspire during development to shape the functional organization of the brain for language. A series of studies on the development of object recognition in monkeys provides new clues for the understanding of the relative roles of nature and nurture in the development of slowly maturing functions like language. In the adult monkey, object recognition is mediated by the inferior temporal cortex including, area TE. After bilateral lesions of TE, adult monkeys are unable to perform tests of object recognition. However, bilateral removal of TE in infant monkeys leads to performance that is only slightly below that of age-matched controls. Thus, as is the case for language, the outcome of a lesion is dramatically different in adults and in infants. By lesioning other areas in TE-lesioned infants, the researchers showed that object recognition in these individuals is mediated by the recruitment of a large network of other cortical areas, which normally mediate other visual tasks or multisensory information. Further studies suggest that this reorganization is possible in infants because of the existence of early diffuse anatomical connections that disappear during development under normal circumstances, but that in the lesioned case are maintained and recruit the cortical areas they contact for their functional purpose. This research illustrates how the biological endowment of the infant, such as the pattern of redundant connections, constrains development. But it also shows how experience shapes the path of development by controlling the selective loss of some and not other connections. From what we know of biology, it is very likely that, throughout development, the genetic material provides the learner with a brain architecture that constrains the possible path of development, but experience, by altering the existing architecture, in its turn influences the next set of possible genetic instructions.

V. SUMMARY AND CONCLUSIONS

The results from the language studies taken as a whole point to different developmental time courses and developmental vulnerabilities of aspects of grammatical and semantic–lexical processing. Thus, they provide support for conceptions of language that distinguish these subprocesses within language. Similarly, following auditory deprivation, processes associated with the dorsal visual pathway were altered more than functions associated with the ventral pathway, providing support for conceptions of visual system organization that distinguish functions along these lines. A general hypothesis that may account for the different patterns of plasticity within both vision and language is that systems employing fundamentally different learning mechanisms display different patterns of developmental plasticity. It may be that systems that display experience-dependent changes throughout life, including the topography of sensory maps, lexical acquisition (i.e., object–word associations), and the establishment of form, face, and object representations (i.e., ventral pathway functions), rely upon very general, associative learning mechanisms that permit learning and adaptation. By contrast, systems that are important for computing dynamically shifting relations between locations, objects, and events (including the dorsal visual pathway and the systems of the brain that mediate grammar) appear to be dependent on and modifiable by experience primarily during more limited periods in development. This could account for both the greater developmental deficits and enhancements of dorsal pathway function following various developmental anomalies and the greater effects of early altered language experience on grammatical functions. Further research is necessary to characterize systems that become constrained in this way and those that can be modified throughout life. This type of developmental evidence can contribute to fundamental descriptions of the architecture of different cognitive systems. Additionally, in the long run, they may contribute to the design of educational and habilitative programs for both normally and abnormally developing children.

See Also the Following Articles

AUDITORY PERCEPTION • BRAIN DAMAGE, RECOVERY FROM • BRAIN DEVELOPMENT • NEURAL NETWORKS • SENSORY DEPRIVATION • SYNAPTOGENESIS • VISUAL SYSTEM DEVELOPMENT AND NEURAL ACTIVITY

Acknowledgments

We are grateful to Dick Aslin and Elissa Newport for stimulating discussions. We thank Daphne Maurer and an anonymous reviewer for sharing their views on the development of the visual system. This work was supported by a Charles A. Dana Foundation grant to D.B., and NIH Grant DC044184 to D.B. and DC00481 to H.J.N.

Suggested Reading

Bavelier, D., Tomann, A., Hutton, C., Mitchell, T., Corina, D., Liu, G., and Neville, H. J. (2000). Visual attention is enhanced in congenitally deaf individuals. *J. Neurosci.* **20,** RC93 1–6.

Bates, E. (2000). Plasticity, localization and language development. In *The Changing Nervous System: Neurobehavioral Consequences of Early Brain Disorders* (S. H. Broman and J. M. Fletcher, Eds.). Oxford University Press, New York.

Emmorey, K. (1998). The impact of sign language use on visuospatial cognition. In *Psychological Perspectives on Deafness* (M. Marschark and M. D. Clark, Eds.), Vol. 2**,** pp. 19–52. Lawrence Erlbaum Associates, Inc., Mahwah, NJ.

Huttenlocher, P. R., and Dabholkar, A. S. (1997). Regional differences in synaptogenesis in human cerebral cortex. *J. Comp. Neurol.* **387,** 167–178.

Hyvarinen, J., Carlson, S., and Hyvarinen, L. (1981). Early visual deprivation alters the modality of neuronal responses in area 19 of monkey cortex. *Neurosci. Lett.* **26,** 239–243.

Maurer, D., Lewis, T. L., Brent, H. P., and Levin, A. V. (1999). Rapid improvement in the acuity of infants after visual input. *Science* 108–110.

Nelson, C. A., and Luciana, M. (2001). *Handbook of Developmental Cognitive Neuroscience*. MIT Press, Boston.

Neville, H. J., and Bavelier, D. (1998). Neural organization and plasticity of language. *Curr. Opin. Neurobiol.* **8,** 254–258.

Newport, E. (1990). Maturational constraints on language learning. *Cogn. Sci.* **14,** 11–28.

Rakic, P., Bourgeois, J. P., Eckenhoff, N., Zecevic, N., and Goldman-Rakic, P. S. (1986). Concurrent overproduction of synapses in diverse region of the primate cerebral cortex. *Science* **232,** 232–235.

Röder, B., Teder-Salejarvi, W., Sterr, A., Rösler, F., Hillyard, S. A., and Neville, H. J. (1999). Improved auditory spatial tuning in blind humans. *Nature* **400**(6740), 162–166.

Sadato, N., Pascual-Leone, A., Grafman, J., Ibanez, V., Deiber, M.-P., Dold, G., and Hallet, M. (1996). Activation of the primary visual cortex by Braille reading in blind subjects. *Nature* **380,** 526–528.

Simons, K. (1993). *Infant Vision: Basic and Clinical Research*. Committee on Vision, Commission on Behavior, and Oxford University Press, New York.

von Melchner, L., Pallas, S. L., and Sur, M. (2000). Visual behaviour mediated by retinal projections directed to the auditory pathway. *Nature* **404**(6780), 871–876.

Webster, M. J., Ungerleider, L. G., and Bachevalier, J. (1995). Development and plasticity of the neural circuitry underlying visual recognition memory. *Can. J. Physiol. Pharmacol.* **73,** 1364–1371.

Neuropsychological Assessment

AARON P. NELSON* and MEGHAN M. SEARL†
*Harvard Medical School * and Boston University †*

GLOSSARY

baseline examination The first in a series of neuropsychological assessments, the results of which are compared with those of subsequent assessments in order to detect changes in neuropsychological status over time.

battery A selection of specific neuropsychological tests intended for use in the context of an assessment.

diagnosis An identification or label denoting that an individual suffers from a specific illness or condition; the art or act of identifying a disease from its signs and symptoms.

differential diagnosis A list of possible conditions in order of likelihood that could potentially account for a patient's symptoms, findings, and results of examination.

dominant hemisphere The hemisphere of the brain that subserves language functions in an individual.

lesion An area of abnormal brain tissue resulting from injury, disease, or anomalous development.

lesion method The process of inferring a causal brain–behavior relationship based on cognitive or behavioral dysfunction following a specific brain lesion.

limbic system A network of structures deep in the brain important for memory and emotion. Major components relevant to memory functioning include the hippocampus and amygdala.

localization Identification of a specific brain area or network subserving a particular cognitive or behavioral function.

neuropsychological test A standardized procedure designed to elicit a behavioral sample, usually related to a particular cognitive function, that can be used to guide inferences about the integrity of brain function.

process approach A hypothesis-driven method of neuropsychological assessment that tailors the selection of tests to the particular patient, uses test behavior and performance during the examination to inform the subsequent course of the assessment, and incorporates qualitative aspects of testing performance into the process of analysis and diagnostic formulation.

prognosis The anticipated future course of a patient's illness or condition.

psychometric method A means of scientifically quantifying human cognitive abilities, originally designed for the purpose of investigating individual differences.

psychophysics The science of measuring basic perceptual functions; a branch of psychology concerned with the effect of physical processes on mental processes.

reliability The consistency with which a test yields the same results on repeated occasions.

validity The accuracy with which an assessment instrument measures the function it was intended to measure.

Neuropsychology is the science of human behavior based on the function of the brain. Clinical neuropsychology is an applied discipline that uses principles of brain–behavior relationships to evaluate and diagnose abnormality in the realm of behavior and cognition. This process of evaluation and diagnosis, termed "neuropsychological assessment," involves obtaining information about an individual's history and current functioning as well as administering a variety of behavioral measures designed to probe specific cognitive functions. A clinical neuropsychologist analyzes and integrates information collected during the assessment process with the aim of providing diagnostic clarification or helping to guide treatment planning.

I. HISTORICAL DEVELOPMENT

From the dawn of self-reflective thinking, humans have wondered about the source of mental activity. The solution to this puzzle was largely relegated to superstition, faith, and philosophy. The 137th Psalm (sixth century B.C.E.) may constitute one of the earliest recorded descriptions of a neurobehavioral syndrome: "If I forget thee, O Jerusalem, let my right hand forget her cunning. If I do not remember thee, let my tongue cleave to the roof of my mouth; if I prefer not Jerusalem above my chief joy." The author appears to be depicting the major sequelae of a left hemisphere stroke: right hemiparesis (weakness) and aphasia (inability to speak). We can glimpse the first stirrings of a science in the writings of Hippocrates, who posited that the brain was the organ of intellect. However, it was not until the time of Flourens in the 1800s that the first studies of brain function were conducted.

Clinical neuropsychology traces its roots to the confluence of three major tributaries of psychological science: psychophysics, the psychometric method, and the lesion method in behavioral neuroanatomy. New methodologies for viewing the brain, both structurally and functionally, continue to illuminate the everyday miracles of thought, memory, and emotion. The integration of clinical work with experimental discoveries has reached a new zenith in cross-fertilization of both endeavors. The phenomenal growth of clinical neuropsychology over the past three decades owes much to contemporaneous work in clinical neuroscience.

Although it is not possible to pinpoint the date of its establishment as a discipline, Arthur Benton points out that clinical neuropsychology lacked the usual trappings of a coherent specialty prior to 1960. There were no professional organizations, journals, or training programs. In its earliest form neuropsychology was practiced in medical settings in association with departments of neurology and neurosurgery. Psychologists with expertise in psychological measurement and a special interest in the behavioral effects of brain injury designed tests to assess various abilities in their patients.

Although theories of brain function can be traced back to the writings of the ancients, the term *neuropsychology* came into use only relatively recently. The ancient Egyptians originally localized the seat of human thought to the heart. This notion remained largely unchallenged until some 3000 years later when a Greek student of Pythagoras proposed that the brain was responsible for sensation and thought and advanced the notion that specific aspects of mental function were represented in specific regions of the brain. This approach to thinking about the brain was later termed localization and became the center of great controversy 2000 years later. Hippocrates, writing in the fourth century BC, claimed the brain as the organ of intellect, sensation, and emotion. Furthermore, he posited that mental illness was rooted in abnormal brain function. Along this line, he advanced the hypothesis that epilepsy was not the result of demonic possession, but rather an organic ailment.

Beginning in the middle of the nineteenth century, psychophysics comprised a refined approach to the precise measurement of various human attributes. Galton's studies of individual differences in sensory and psychomotor responses epitomized this approach, as did the early work of Ebbinghaus on "individual memory curves of retention." At the beginning of the twentieth century, Binet and others introduced the revolutionary concept of measurable intelligence, a notion that continues to have a profound and controversial impact on human society. The psychometric method was employed to design and construct measurement methods possessing validity and reliability, which are crucial underpinnings of all psychological and neuropsychological tests. The lesion method has its origins in early case reports of brain–behavior effects in injury and illness. The famous case of Phineas Gage (1848) described a railroad worker who survived a catastrophic injury in which an iron spike was blasted through his head, essentially disconnecting the anterior portion of his brain. The resulting pattern of spared and affected functions in Gage gave rise to an early understanding of the functional role of the frontal lobes in human behavior. The pioneering work of Paul Broca, a French neurologist who described his patient "Tan" (1861), was instrumental in unveiling the role of the dominant hemispheres in human language function. More recently, in the 1950s, Scoville and Milner's studies of patient HM shed new light on the role of the limbic system in anterograde memory function.

The advent of World War I saw the first period in history when modern American medicine had to cope with the tragedy of "industrial" warfare. Legions of combatants returned from Europe with life-altering brain injuries. Assessing the cognitive consequences of such injuries spurred the growth of early assessment methodologies. The aftermath of World War II marked the establishment of modern clinical psychology, with an emphasis on the special skill of psychological testing. Measurement of intelligence was the

venue in which early testing applications were most significant.

The earliest codification of neuropsychological examination methods can be attributed to Dr. Ralph Reitan. A pioneering neuropsychologist, Reitan and his colleagues assembled a battery of test measures in which specific patterns of scores were linked to the dysfunction of associated brain regions. Later approaches diverged from reliance on a set battery of tests and instead used a more dynamic method, the "process" approach to neuropsychological assessment, developed by Dr. Edith Kaplan in her work at the Boston V.A. Hospital. This technique entails ongoing *in situ* hypothesis testing in which particular attention is paid to the qualitative aspects of a patient's test response. Although quantitative data are also important, the patient's route to solving a particular cognitive problem often reveals an underlying process that can be reflective of spared and affected brain regions. In contemporary practice, both of these methods rely on a vast body of knowledge about characteristic syndromes that are known to be associated with underlying disease states.

The current practice of clinical neuropsychology concerns itself with the assessment of cognitive functions, through the use of test instruments, for the purpose of understanding the functional integrity of the brain. By relying on knowledge of brain–behavior relationships and characteristic syndromes, the results of assessment are useful for the diagnosis and treatment of brain-related injuries and illness.

II. FUNDAMENTAL ASSUMPTIONS

The study of clinical neuropsychology is firmly rooted in the larger field of neuroscience and, therefore, rests on the assumption that the nervous system impacts behavior and cognition. Conversely, inferences can be made about the integrity of the brain based on observable behavior. The ability to make accurate and meaningful inferences is predicated on a thorough understanding of two streams of knowledge: (a) the neural infrastructure underlying normal human cognition and behavior and (b) the characteristic profiles of neurocognitive and neurobehavioral syndromes.

Observable behavior is frequently the most sensitive manifestation of brain pathology. Such behavior can range from a person's manner of dress to his or her performance on a specific neuropsychological test. A competent neuropsychologist will sample multiple domains of behavior. Observable behavior, including "test behavior," reflects an interaction between the domains of person and environment. Variables from each domain must be considered in order to arrive at an understanding of the clinical significance of a given behavior.

Individual tests used in neuropsychological practice focus on measurement of particular cognitive processes. In order to be useful, each test must be constructed according to sound psychometric principles and possess adequate validity and reliability. In addition, there must be appropriate normative data with which to compare a single patient's performance. It is the responsibility of the neuropsychologist to have a comprehensive understanding of test construction and proper usage so that it may be applied appropriately in relevant clinical circumstances. This includes choosing tests with normative data derived from subjects of similar age, education, and nationality to the patient currently being tested.

The interpretation of test results depends on an understanding of the component processes involved in any given test response. For example, the ability to name an object depends on multiple processes that are mediated by different brain systems. A person must first orient and attend to the stimulus. Second, the stimulus must be accurately registered at the level of visual perception. Third, the neural pathway that links the visual percept to meaningful recognition must be intact. Fourth, the ability to assign a phonemic/lexical label to the object must be intact. Finally, the person must be able to convey a response through speech. Because a complex cognitive process, such as naming a simple object, can be undermined by perturbation at any point along the route between these processes, a neuropsychologist must understand the component processes involved in each behavior. By examining the patient's function in multiple domains with multiple measures, the neuropsychologist can determine which processes and associated neural circuits are functioning abnormally.

Performance on a single test is not sufficient to make a diagnostic inference. A common misconception is that a poor score on a particular test denotes an impairment in the domain that the test is nominally designed to assess. For example, if a patient performs poorly on a memory test, this does not necessarily signify an impairment in memory, but could indicate difficulties with attention or language. A neuropsychologist will evaluate a profile of test results in a dynamic, interactive fashion in order to arrive at a diagnostic formulation.

Neuropsychological testing is simply one means of obtaining a sample of behavior. A neuropsychologist

must proceed with caution in using test data to predict behavior. The testing environment is, by necessity, contrived to promote a standard approach to test administration. This contrivance constitutes a challenge in understanding how test performance corresponds to "real-life" behavior. For example, a patient who complains of difficulty with concentration and memory at work may perform quite normally in the context of the quiet, distraction-free consultation room. Discrepancy between test behavior and real-life behavior is the source of ongoing challenge for the design of an ecologically valid assessment environment.

III. CLINICAL APPLICATIONS

In its infancy, neuropsychological testing was used primarily as a means of localizing the site of a lesion in the brain. Developments in the field of neuroimaging have obviated the need to depend exclusively on neuropsychology in this role. In contemporary clinical practice, neuropsychology is primarily devoted to diagnostic formulation, entailing the identification and description of cognitive and behavioral syndromes. Hand in hand with advanced imaging techniques and other diagnostic methodologies, the evolution of a true behavioral neuroscience has been ongoing.

Clinicians often refer patients for neuropsychological assessment when there is a question about diagnosis. Neuropsychological assessment can be useful in describing the nature of cognitive or behavioral deterioration and can help clarify the pathology underlying such a process. For example, neuropsychologists often see elderly individuals who have been referred for the evaluation of a possible neurodegenerative illness, such as Alzheimer's disease. Through an analysis of a variety of behaviors, neuropsychologists can often clarify the diagnosis and, in so doing, help guide decisions about appropriate treatment and provide information concerning prognosis. It is often difficult to arrive at a definitive diagnosis; in these cases, a neuropsychologist can assist in narrowing the *differential diagnosis*. Serial follow-up examinations can further clarify the diagnosis by tracking the course of cognitive symptoms across time.

Neuropsychological test data can be useful for establishing a baseline of cognitive functioning against which to calibrate the degree of change in patients with a known history of neurological disease, injury, or developmental abnormality affecting cerebral functions. This involves both tracking decline in the context of illness and also measuring treatment gains. With rich quantitative and qualitative data, neuropsychologists are in a unique position to evaluate the efficacy of treatment interventions (e.g., pharmacotherapy, electroconvulsive therapy, neurosurgery) by establishing a patient's baseline performance and reevaluating following initiation of therapy.

Neuropsychological assessment can delineate the potential impact of brain pathology on venues of real-life functioning, ranging from the most rudimentary human activities (bathing and dressing) to work life and the complexities of intimate relationships. Neuropsychologists can provide an explanatory context for a patient's cognitive, social, or behavioral difficulties. In addition, in the case of progressive disease, neuropsychologists can provide education about how the patient's changing condition will impact day-to-day functioning differently over time. It is not unusual for an otherwise sensitive spouse to blame an affected husband or wife for a particular behavior, which had been assumed to be under voluntary control. Learning that the behavior is part and parcel of a disease process dramatically alters this attribution and allows both patient and spouse to work together on the problem.

Because of its knowledge base in behavioral neuroscience, neuropsychology has an important role in treatment planning, Neuropsychologists make specific and detailed recommendations for rehabilitation in cases of brain injury, tumor, and stroke and are central to the planning process for patients and families confronting dementing illnesses.

Neuropsychological consultation is frequently critical in cases involving the adjudication of criminal or civil matters. Neuropsychologists with special training and experience in forensics play an important role in the determination of competency in matters ranging from criminal responsibility to guardianship.

IV. APPROACH TO NEUROPSYCHOLOGICAL ASSESSMENT

Two major approaches to neuropsychological assessment have emerged during the last century.

A. Fixed Battery Approach

The fixed battery approach to neuropsychological assessment was developed primarily by Ralph Reitan and his colleagues. As the name suggests, practitioners of this approach typically utilize a standard set of tests in the context of assessment and do not vary this battery, if possible. The battery is designed to assess each major cognitive domain and therefore ensures a

comprehensive assessment of every patient. A clear advantage of this approach is the ability to detect deficits that are unsuspected or not otherwise evident. Another advantage is that practitioners of this approach become very familiar with the subtleties of the battery and develop a greater sensitivity to slight variations in the performance of component tasks. Disadvantages of this approach include inefficiency (i.e., a time-consuming battery is administered in all cases, including those for whom more focused assessment would be sufficient) and lack of flexibility. Some have also criticized the fixed battery approach for insufficient assessment of key functions, including memory.

B. Flexible Battery or Process Approach

The other major approach to neuropsychological assessment is the Boston Process approach, primarily developed by Edith Kaplan and her colleagues. This type of approach proceeds in a hypothesis-driven manner and uses the patient's test performance to guide and inform an evolving and dynamic assessment strategy. The process-oriented evaluation is individually tailored to each patient and uses specific tests for the purpose of answering particular questions. As answers to these questions emerge during the evaluation, the neuropsychologist moves through a series of decision points, probing particular cognitive functions as needed. Tests are administered in a standardized fashion but can also be modified to test the limits of cognitive function and to produce richer qualitative data. As with the administration of tests, the data can be viewed from both a qualitative and a quantitative perspective. Task performance is analyzed across multiple measures to parse out component processes and identify specific cognitive defects.

V. CLINICAL METHOD

Clinical neuropsychological assessment comprises a series of procedures during which information is elicited and analyzed in the service of diagnosis and treatment planning. This process is illustrated in Fig. 1. While test administration is the most time-consuming component of the process, it is often the case that some of the most valuable information emerges while discussing a patient's history and current life situation. The information obtained during the initial interview is important for differential diagnosis and frequently guides the remainder of the assessment.

A. Referral Question and Chief Complaint

The first task of a neuropsychological assessment is to clarify the reason for referral. Patients are typically referred for neuropsychological assessment by other clinicians involved in their care, often for assistance in diagnosis and treatment planning. After the referral question is ascertained, the patient is seen and a chief complaint is elicited, including a clear description of the onset and course of the complaint (e.g., symptoms, concerns) as well as information regarding the medical and social context in which the problem(s) emerged. The patient's overall understanding of the rationale for the consultation and appreciation of his/her current circumstances is also sought.

B. History

Information is obtained from a variety of sources, including the patient's self-report, observations of family members or close friends, medical records, and prior evaluations from academic or work situations. Information is obtained regarding (1) past medical history, including illnesses, injuries, surgeries, medications, hospitalizations, substance abuse, and relevant family medical history, (2) past psychiatric history, including hospitalizations, medications, and outpatient treatment, (3) developmental background, including circumstances of gestation, birth and delivery, acquisition of developmental milestones, and early socialization skills, (4) social development, including major autobiographical events and relationships (a three-generation genogram is highly useful in gaining relevant family information), (5) educational background, including early school experiences and academic performance during high school, college, postgraduate study, and other educational and technical training (the presence of a learning disability should be assessed) (6) vocational history, including work performance, work satisfaction, and relationships with supervisors and co-workers, and (7) recreational interests and hobbies.

C. Behavioral Observations

Physical appearance is inspected, including symmetry of anatomical features, facial expression, manner of dress, and attention to personal hygiene. The patient is asked specific questions regarding unusual sensory or motor symptoms. Affect and mood are assessed with respect to range and modulation of felt and expressed

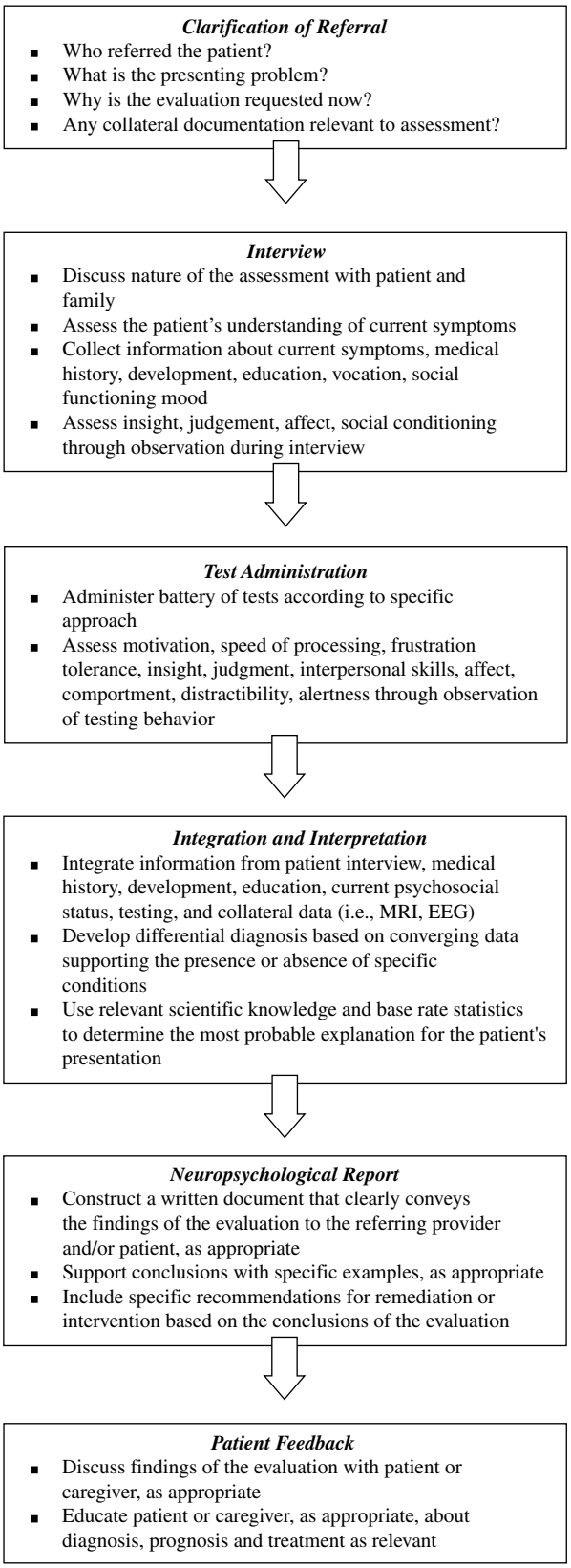


Figure 1 Schematic of neuropsychological evaluation.

emotions and their congruence with concurrent ideation and the contemporaneous situation. Interpersonal comportment is assessed in the context of the interview. Specifically, attention is paid to whether the patient's behavior reflects a normal awareness of self and other in the interaction. Whether the patient is motivated and complies with examination requests, instructions, and test procedures. An appreciation of the patient's level of motivation to participate in the evaluation and comply with examination instructions is crucial to assessing the validity of the test data. General level of *arousal* or *alertness* is determined by observing the patient's degree of drowsiness, tendency to yawn or fall asleep during the interview, level of interpersonal engagement, and speed of response in conversation. Environmental and diurnal factors can modify arousal, and an attempt should be made to assess whether this is relevant in the case of a particular patient by inquiring about the consistency of arousal level and any fluctuations that the patient or caregiver has observed.

D. Examination

A sufficiently broad range of neuropsychological functions is evaluated using tests and other assessment techniques. The major domains to be surveyed include general intellectual ability, attention, executive functioning and comportment, memory, language, visuospatial abilities, motor functioning, and mood/personality. As a prelude to test administration, it is imperative to establish the integrity of sensation and perception because impairments in these areas can invalidate the results of examination. For example, it would be incorrect to conclude that a patient has a receptive language impairment when, in fact, there is a primary hearing deficit.

Significant impairment of sensory function (auditory, visual, kinesthetic) is usually obvious and may indicate a need for specialized assessment procedures. However, when there is no obvious impairment, it is the responsibility of a neuropsychologist to inquire about all sensory modalities, particularly with regard to visual, auditory, and kinesthetic function. Vision can also be examined with tests of acuity, tracking, scanning, depth perception, color perception, and attention–neglect for visual field quadrants. Simple auditory function can be assessed by finger rub stimuli to each ear. Unusual or abnormal gustatory and olfactory experiences should be sought through direct questioning. Kinesthetic perception is assessed with tests of graphesthesia and stereognosis. Double simultaneous stimulation can be

used in auditory, visual, and kinesthetic modalities to determine whether hemiextinction occurs. Once it is clear that basic sensory and perceptual functions are intact, neuropsychological tests can be administered in a standard manner.

E. Diagnostic Formulation

Data from the history, observation, and testing of the patient are analyzed collectively to produce a concise understanding of the patient's symptoms and neuropsychological diagnosis. A configuration of abilities and limitations is developed and used both diagnostically and as a framework through which to address the goals of treatment. When possible, the diagnostic formulation should identify the neuropathological factors that give rise to the patient's clinical presentation, including underlying anatomy and disease process.

F. Recommendations and Feedback

Consultation concludes with the process of feedback, through which the findings of the evaluation and treatment recommendations are reviewed with relevant individuals (i.e., referring physician, the patient, family, treatment team members). A variety of treatment plans may be advised, including neurological evaluation, psychiatric consultation, psychotherapy, cognitive–behavioral remediation, and vocational guidance. Recommendations should be pragmatic and individually tailored to each patient's specific needs. Strategies for optimizing performance in personal, educational, occupational, and relational spheres are identified and discussed in lay language that the patient and family member can understand. When possible, the clinician outlines concrete strategies to facilitate the remediation of identified problems. Further clinical evaluations and other neurodiagnostic procedures are suggested, when appropriate, in order to provide more information relevant to differential diagnosis, response to treatment, and functional status over time. Appropriate neuropsychological follow-up is also arranged, when indicated.

VI. DOMAINS OF NEUROPSYCHOLOGICAL FUNCTION

The fractionation of neuropsychological functions into specific domains is a somewhat arbitrary organizational contrivance. In reality, there is considerable overlap within and between cognitive domains. For example, working memory shares much common ground with aspects of attention and language.

A. General Intellectual Ability

Determination of an individual's level of intellectual functioning is a fundamental component of the neuropsychological assessment. Once established, general level of intelligence serves as a point of reference from which to evaluate performance in other domains. Intelligence encompasses a broad range of capacities, many of which are not directly assessed in the traditional clinical setting. The estimate of general intellectual ability is based on both formal assessment methods and a survey of demographic factors and life accomplishments. Formal measures of general intellectual function typically assess a broad range of functions through multiple subtests (e.g., *WAIS-III*, *WASI*, *WISC-III*), and yield an "intelligence quotient" as well as other derived index scores. Other test instruments obtain high correlations with IQ measures and have been used to estimate overall intellectual ability (e.g., Ravens Progressive Matrices, Shipley Institute of Living Scale). In cases of known or suspected impairment, premorbid ability can also be surmised from performance on measures presumed less sensitive to cerebral dysfunction (e.g., vocabulary). Single-word reading tests (e.g., *AmNART*) have been used to estimate baseline verbal intelligence in patients with early degenerative conditions. In addition, so-called "best performance methods" can be used. This method assumes that the patient's highest level of performance can be used to set a reference point for optimal baseline ability (Table I).

Demographic variables, including level of education and professional achievement, avocational interests and pursuits, and socioeconomic status, can also be used to gauge an individual's current or premorbid general intellectual ability. Relationships among these variables are complex, and particular care must be exercised in the evaluation of patients from diverse sociocultural backgrounds. It is important to consider real-world accomplishments in the unique context of the individual, with attention to environmental factors that influence opportunity and achievement.

B. Attention and Executive Function

1. Attention

Many cognitive capacities are inherently predicated on a fundamental ability to attend to the surrounding

Table I
General Intellectual Ability

Test name	Description
Wechsler Intelligence Scales	The *Wechsler Adult Intelligence Scale-III* (*WAIS-III*) is composed of 13 individual subtests. Administration of all subtests generates three Intellectual Quotients (IQs): Full-Scale IQ, Verbal IQ and Performance IQ, and four different performance indices: Verbal Comprehension Index, Perceptual Organization Index, Working Memory Index, and Processing Speed Index.
	The *Wechsler Abbreviated Scale of Intelligence* (*WASI*) is designed to be a short and reliable measure of intelligence that produces VIQ, PIQ, and FSIQ scores that are similar to those obtained with the *WAIS-III*.
	The *Wechsler Intelligence Scale for Children* (*WISC-III*) is used for testing children and adolescents ranging in age from 6 to 17.
	The *Wechsler Preschool and Primary Scale of Intelligence* (*WPPSI*) is used for testing children ranging in age from 4 to 6.5 years.
American National Adult Reading Test (*AmNART*)	The *AmNART* is a measure of recognition vocabulary requiring oral reading of 45 phonetically "irregular" words. It is used to estimate premorbid baseline intellectual ability in patients with known or suspected dementia. Similar measures include the *NART*, *NART-Revised*, and *ANART*. Errors consist of word mispronunciations.
Shipley Institute of Living Scale	The *Shipley ILS* consists of a multiple-choice measure of vocabulary (knowledge of synonyms) and a measure of verbal reasoning (providing the missing member of an established logical pattern). The two components are scored together to estimate a *WAIS-R FSIQ* score. Errors consist of incorrect responses to the task involving choosing a synonym for the target word and providing no response or responding with an incorrect continuation of the logical pattern.
Raven's Standard and Colored Progressive Matrices	These are standardized measures of nonverbal analogical reasoning widely used both within and outside the U.S. as a "culture fair" measure of general intellectual ability. Both tests require the patient to demonstrate an understanding of the logic underlying visual patterns by selecting the missing component of the pattern from a series of choices. The *Standard Matrices* contain 60 black and white items ranging from simple to extremely difficult, while the *Colored Matrices* consist of 36 colored items that span a limited range of complexity. Errors consist of incorrect identification of the missing component of the visual pattern.

environment. For example, an individual cannot effectively name an object if the object is not first attended to and visually processed. Because attention is a prerequisite for other aspects of cognitive function, disruption of attention can generally skew the results of a neuropsychological assessment. Early assessment of attention is vital for informing the scope of the examination and the analysis of test data (Table II).

Attention is a general term that encompasses a number of different component processes. *Attention span* refers to the number of unrelated "bits" of information that can be held on line at a given moment in time. Assessment of attention span is typically accomplished through the recall of progressively longer series of information bits, such as numbers (digit span) or spatial locations (spatial span). *Sustained attention*, also called *vigilance*, refers to the capacity to maintain active attention over time. The most common method of assessing vigilance utilizes a target detection paradigm. Here the patient is instructed to respond to an infrequently occurring target stimulus. For example, on a measure of auditory vigilance, the patient hears a series of letters of the alphabet and must signal by pressing a response key each time a particular target is read (see *CPT*, *PASAT*). *Selective attention* is similar to sustained attention but requires a response only to a particular class of stimuli, but not other stimuli. *Set-shifting* refers to the capacity to relinquish an existing procedural strategy in favor of a new response, based on recognition of a change in reward contingencies. It is typically measured with tasks requiring the patient to shift focus among stimulus features of test display (see *Trailmaking A and B, Wisconsin Card Sorting Test, Luria Graphomotor Sequences*). *Resistance to interference*, also called *response inhibition*, refers to the ability to sustain a given response, even in the face of a salient distraction designed to undermine the target response.

This is assessed with tasks requiring the patient to inhibit overlearned responses or other distractions that could undermine a desired response (see *Stroop Interference Test*, *Trailmaking Test*).

2. Executive Functioning

Executive functions require the capacity to process information in a planful, organized, and contextually appropriate manner. Formal tests of executive function assess a number of different capacities, including some functions mentioned earlier (*set-shifting*, *overcoming interference*, *response inhibition*), and also *planning*, *perseverence*, *initiation*, *reasoning*, and *abstraction*. *Planning* involves thinking several steps ahead of one's current circumstances for the purpose of informing and altering a course of action. Tests measuring planning ability often require a person to determine the correct series of steps needed to successfully reach a particular goal (see *Tower of Hanoi*, *Tower of London*). *Perseverence* is the ability to sustain a particular course of action, even in the absence of an external prompt. Measurement of perseverance often begins with both examiner and subject performing the same task, but involves the subject continuing the task even after the examiner has stopped (see *Luria Motor Sequences*, *Verbal and Design Fluency*). *Initiation* refers to the ability to spontaneously commence an action in the absence of a direct prompt from the external environment. This function is measured with the presentation of a task followed by a period of time during which the subject is expected to respond independently (see *Verbal Fluency*, *Go-No-Go*). *Reasoning* involves using a system of logic to solve a particular problem or task. This can be measured in a variety of ways, including the use of visual puzzles and verbal analogies (see *Raven's Progressive Matrices*, *Shipley Institute of Living Scale-Abstraction*, *WAIS-III Comprehension*). *Abstraction* is the ability to articulate shared attributes of dissimilar objects (see *WAIS-III Similarities*, *Wisconsin Card Sorting Test*) (Table II).

C. Comportment, Insight, and Judgment

Although few tests probe these functions in a formal manner, they are important components of neuropsychological function and are frequently disturbed in cases of neurological and neuropsychiatric illness. *Comportment* refers to the ability to behave in a contextually appropriate manner. Disturbances often manifest as socially inappropriate behaviors that suggest an insensitivity to accepted cultural norms. Examples might include making offensive comments, crossing interpersonal boundaries, interrupting during conversation, or failing to attend to personal hygiene. In the context of neuropsychological assessment, *insight* involves an accurate perception of one's mental and physical condition as well as appreciation of the impact of one's behavior on others. Cognitively impaired individuals often lack one or both of these components of insight. *Judgment* involves the capacity to perceive and assess one's environment accurately and to make decisions that reflect a sensitivity to preserving the safety and integrity of oneself, one's resources, and one's environment.

A neuropsychologist can assess these functions informally through naturalistic observation, reports from individuals familiar with the patient, and by inquiring about any accidents or legal infractions involving the patient. Formal measures of these functions exist primarily in the form of questionnaires that quantify the degree to which the patient manifests disturbances in these general areas. The *Frontal Lobe Personality Scale* (*FLOPS*) comprises 46 descriptions of various behaviors that are characteristic of patients with frontal lobe damage. Each behavior is assigned a severity rating by the patient and by a family member, and the overall severity rating is thought to reflect the degree of behavioral disturbance present. The *Neuropsychiatric Inventory* (*NPI*) is another self-report rating scale that is completed by a family member or caregiver. Both the extent of the patient's behavior and the degree of subsequent familial distress are rated.

D. Learning and Memory

The assessment of learning and memory function is perhaps the most complex endeavor of the neuropsychological examination. *Working memory* involves holding a stimulus or set of stimuli in mind in order to either produce it after a delay (i.e., looking up a telephone number and remembering it until it is successfully dialed) or use it in a mental procedure involving manipulation of information (i.e., carrying out mental arithmetic). The simpler aspect of working memory, also called *maintenance* of information, can be tested by requiring a subject to hold information in mind and reproduce it after a short delay (see *Digit Span Forwards*). The more complex components of working memory can be tested in a number of ways, all of which involve on line maintenance *and*

Table II
Attention and Executive Functioning

Function	Test	Description
Attention span	*Digit Span*	The examiner reads increasingly long strings of numbers aloud. The examinee must repeat the numbers aloud in the same order, first forward and then in the reverse order.
	Spatial Span	The examiner taps a series of blocks in fixed locations. The examinee must repeat this series in the same order and then in the reverse order.
Sustained attention and vigilance	*Auditory and Visual Continuous Performance Test* (*CPT*)	Basic auditory vigilance is tested by having the patient listen to a series of letters read serially and responding to a single target letter or series of letter configurations. Visual vigilance can be tested by showing a series of single numbers on a computer screen and requires responding to a target stimulus, but not to non-target stimuli.
	Paced Auditory Serial Addition Test (*PASAT*)	The patient hears a tape recorded voice reading numbers at various rates, ranging from every 2.4 seconds to every 1.2 seconds. The objective of the task is to sum the last two numbers heard and voice the sum aloud (e.g., if the numbers from the tape were "5, 2, 8, 4," the patients responses would be the following (in italics): "5, 2, *7*, 8, *10*, 4, *12*," etc.). The process of voicing the sum aloud serves as interference, which must be overcome in order to attend to the following number. In total, 60 numbers are read in each of the four trials and every subsequent trial is faster than the one preceding it.
Set-shifting	*Trail Making Test A and B*	This measures visual scanning, visuomotor tracking, and response set flexibility. Trails A involves connecting consecutively numbered circles, from 1 to 25. Trails B requires the patient to continually shift set, alternating between letters and numbers (i.e., 1, A, 2, B, etc.). Both tasks must be performed as quickly as possible and without lifting the writing utensil from the paper.
	Wisconsin Card Sorting Test	A measure of nonverbal concept formation, response set flexibility, sustained attention, and ability to integrate corrective feedback. The patient must sort cards according to underlying principles (color, form, number), which must be deduced and which are shifted at set intervals.
	Luria Graphomotor Sequencing	Involves using a pencil to copy a series of patterns (i.e., m-n-m-n, peaks and plateaus, and multiple loops) without lifting the pencil. Errors consist of failing to alternate (i.e., m-n-m-m-m-m).
	Luria Motor Sequences	Involves performing repeated sequences of hand movements. Errors consist of failure to maintain the order of movements within each sequence.
Response inhibition	*Stroop Color-Word Interference Test*	Composed of three parts that require (1) reading a series of black and white color words (red, blue, green), (2) identifying the color of red, blue, and green "Xs," and (3) identifying the ink color of incongruent color words (i.e., the correct response to the word "red" printed in blue ink would be "blue"). In all conditions the patient is asked to perform as quickly as possible.
	Motor Go-No-Go	Involves responding to a target signal (one loud knock) by lifting a finger, but not to a second signal (two loud knocks). The tendency to respond to the second signal must be inhibited. The examiner produces a series of one or two knocks and observes the patient's responses. Each hand is tested separately.
Planning	*Tower of London*	Involves ordering three or four colored beads within a set of constraints. Only one bead may be moved at a time, and beads moved from their initial placement may not be returned. Similar tests include *Tower of Hanoi* and *Tower of Toronto*.
Perseverance	*Verbal and Design Fluency*	Letters: the examinee must recite as many words as possible that start with a particular letter, with the exception of proper nouns, numbers, and more than one iteration of the same root word. This is repeated with three different letters in total. Categories: the examinee must recite as many words as possible from a particular category. This is repeated with three different categories in total. Design: the examinee must create as many unique designs as possible by connecting lines between dots laid out in a grid.

manipulation of information (see *Tests of Mental Control*, *WAIS-III Letter–Number Sequencing*, *WAIS-III Arithmetic*, *PASAT*). Memory is assessed with respect to modality of presentation (auditory vs visual), material (linguistic vs figural), and locus of reference (personal vs nonpersonal). Also important is the time of initial exposure, namely, whether information was learned before the onset of brain damage (retrograde memory, see *Boston Retrograde Memory Test, Transient Events Test*) or after (anterograde memory, see *WMS-III*, *RAVLT*, *Bushke SRT*, *Three Words–Three Shapes*, *Warrington RMT*). The evaluation of memory should include measures that allow the dissociation of the component processes entailed in the acquisition and later recall of information, namely, encoding, consolidation, and retrieval. To this end, measures are used to assess performance with respect to length of the interval between exposure to information and demand for recall (immediate vs short vs long delay) and extent of facilitation required to demonstrate retention (free recall vs cued recall vs recognition). The assessment of retrograde memory function poses a special problem insofar as it is difficult to know with certainty what information was previously registered in the remote memory of a particular patient. Although there are a number of formal tests that can be used for this purpose (e.g., *Boston Remote Memory Battery*, *Transient Events Test*), we also assess this aspect of memory function by asking for personal information that presumably is or had been well-known at one time by the patient (i.e., names of family members, places of prior employment). In these instances it is helpful to obtain confirmation of the accuracy of this information from family members or friends, if possible. (See Table III).

E. Language

Language is the medium through which much of the neuropsychological examination is accomplished. Language function is assessed both opportunistically, as during the interview, and via formal test instruments (Table IV). Conversational speech is observed with respect to fluency, articulation, and prosody. The patient's capacity to respond to interview questions and test instructions provides an informal index of receptive language ability or comprehension. Visual confrontation naming is carefully assessed so that word-finding problems and paraphasic errors may be elicited. Repetition is measured with phrases of varying length and phonemic complexity. Auditory comprehension is evaluated by asking the patient questions that range in length and grammatical complexity. Reading measures include identification of individual letters, common words, irregularly spelled words, and nonwords, as well as measures of reading speed and comprehension. Spelling can be assessed in both visual and auditory modalities. A narrative handwriting sample can be obtained by instructing the patient to describe a standard stimulus scene.

F. Visuospatial Functions

After the integrity of basic visual acuity is established, the spatial distribution of visual attention is evaluated. The presence of visual neglect is assessed through the use of tasks that require scanning across all quadrants of visual space. Assessment of left–right orientation involves directing patients to point to specific body parts, either on themselves or the examiner. Topographical orientation can be tested by instructing the patient to indicate well-known locales on a blank map. Graphic reproduction of designs and assembly of patterns using sticks, blocks, or other media are used to assess visual organization and constructional abilities. Facial recognition represents a special component perceptual process and can be measured using Benton's *Facial Recognition Test*. The *Judgement of Line Orientation Test* assesses perceptual accuracy in judging the angular displacement of lines. Warrington's *Visual Object Space Perception Battery* is an example of a collection of measures, designed to assess various aspects of perceptual function. (See Table V).

G. Motor Functions

Naturalistic observations of the patient's gait and upper and lower extremity coordination are an important part of the motor examination. Hand preference should be assessed either through direct inquiry or a formal handedness questionnaire. (Table VI) Motor speed, dexterity, and programming are tested with timed tasks, some of which involve the repetition of a specific motor act (e.g., finger tapping, peg placement) and others involve more complex motor movements (e.g., finger sequencing, sequential hand positions). Manual grasp strength can be assessed with a hand dynamometer.

H. Affect, Mood, and Psychological Functioning

Standardized measures of mood, personality, and psychopathology can be used to assess the contribution

Table III
Learning and Memory

Function	Test	Description
	Tests of Mental Control	The examinee is asked to recite familiar sequences such as the alphabet, days of the week, months of the year, and numbers from 1 to 20 as quickly as possible. The examinee then must recite all sequences, except the alphabet, in reverse order. Serial subtractions involve subtracting a particular number until a predetermined point is reached.
Working memory[a]	*WAIS-III Letter–Number Sequencing*	The patient hears a series of alternating numbers and letters and is required to reconfigure them so that all of the numbers are recited first, in ascending order, and then letters in alphabetical order.
	WAIS-III Arithmetic	Involves mental calculation of aurally presented arithmetic problems of increasing difficulty.
	Boston Retrograde Memory Test	Involves showing a series of black and white photos of famous persons from the 1920s to the 1980s. If the patient cannot spontaneously generate the correct name, the examiner can give a semantic cue (i.e., "he was a singer in the 1920s"), and then a phonemic cue (i.e., first name). This test assumes that the patient has been exposed to the information in the first place.
Retrograde memory	*Transient Events Test* (*TET*)	This is a measure of memory for popular news events from the 1950s through the 1990s. Items were selected by way of the *New York Times* index according to the criteria that they were mentioned at least 250 times during a particular year and less than five times over the subsequent two years. Hence, all items were of transient notoriety thereby minimizing confounding effects of overexposure. Free recall and recognition are tested.
	Wechsler Memory Scale (*WMS-III*)	This is a composite battery of tests assessing orientation, attention, learning, and memory for verbal and visual information across immediate and delayed intervals. It yields a series of index scores.
	Rey Auditory Verbal Learning Test	This measure of verbal encoding, learning, and retention involves drilling the examinee on a series of 15 unrelated words over five successive trials. Learning is followed by an interference trial, immediate recall, and 30-minute delayed recall and recognition. Various comparisons yield information regarding sensitivity to proactive and retroactive interference and rate of forgetting.
Anterograde memory	*Bushke Selective Reminding Test*	This is a special type of list-learning test that is most helpful in cases where encoding is intact, but where there is a question of impairment at the level of consolidation or storage. A list of 12 words is read and the patient must repeat as many words as possible. However, different from the previous list-learning tasks, the examiner then reads only the words that the patient did not recall. This continues across six trials and each time, the patient must try to recite as many words as possible, but only hears the words not recalled on the preceding trial. There are immediate and delayed recall trials followed by visual multiple choice recognition paradigms.
	Three Words Three Shapes Memory Test	The patient is instructed to copy three words and three shapes after which incidental recall is tested. The patient is drilled on the words and shapes until criterion is reached and recall is tested after intervals of 5, 15, and 30 minutes. Recognition is tested using distracter shapes and words.
	Warrington Recognition Memory Test	Involves the visual presentation of single words and faces at the rate of one every three seconds. The patient is instructed to read each word silently and make and report a judgment regarding his association (pleasant or unpleasant) to it. Immediately afterward, the patient is shown a pair of words, each containing a target word and a distracter, with the instruction to identify the one presented previously. Memory for faces is tested in the same way.

[a] See also *Digit Span* forward and *PASAT* tests in Table II.

Table IV
Language Function

Test name	Description
Boston Diagnostic Aphasia Examination (BDAE)	This comprises measures that assess all aspects of expressive and receptive language function, including naming, comprehension, repetition, reading, writing, praxis, and prosody.
Boston Naming Test	One component of the *BDAE*, this measure of confrontation naming is often administered independently. It consists of a series of 60 black and white line drawings of objects. Naming difficulty increases as the objects progress from high-frequency to low-frequency words. Stimulus cues are provided in the event of perceptual difficulty. Phonemic cues are used to distinguish between retrieval difficulties and lack of knowledge of a particular object name.
Nelson Denny Reading Test	This test contains two multiple choice measures that assess vocabulary and reading comprehension. Reading speed is also computed. Reading comprehension can be scored on the basis of both a standard and an extended length of time.
Wide Range Achievement Test-III (*WRAT-3*)	This standardized battery of acquired scholastic skills includes measures of spelling, written arithmetic, and single-word reading.
Woodcock–Johnson Tests of Achievement Revised	Selected subjects such as dictation, writing fluency, writing sample, proofing, word attack, and reading comprehension provide measures of reading encoding, comprehension, and written language abilities. This is another useful measure in the assessment of language-based learning disabilities.

Table V
Visuospatial Functioning

Test name	Description
Benton Facial Recognition Test	This task is composed of two parts. The first involves matching a target face with one of six faces. The target stimulus is always identical to the correct answer. The second part of the task involves choosing the three photographs, out of the array of six, that contain the same face as the target photograph. Increasing use of camera angle and shadow contribute to the progressive difficulty of the task.
Benton Line Orientation Test	This task involves judging the spatial orientation of sets of line segments by comparing them to a grid composed of 11 radii. It is sensitive to visuoperceptual deficits associated with posterior right hemisphere lesions.
Visual Object Space Perception Battery	This battery contains eight individual tests that each probe a specific component of object or space perception. Individual subtests are untimed and can be given in isolation or within the context of the full battery. Normative data are based on healthy control subjects as well as patients with right and left-hemisphere lesions.
Visual Cancellation	The objective of this task is to circle each instance of a target letter or symbol from among a field of similar stimuli. There are a total of 60 targets evenly distributed among the four quadrants of the 8-1/2 $\times$ 11 inch page. Errors of omission involve failing to respond to the target stimulus. Errors of comission involve responding to stimuli other than the target stimulus.
Hooper Visual Organization Test	This task involves examining line drawings of objects that have been broken into fragments and rotated. The objective is to mentally reorganize each set of fragments and subsequently identify the corresponding coherent whole.
Complex Figure Drawing	This task involves copying a complex line drawing, usually the Rey-Osterreith complex figure or Taylor complex figure. Ability to reproduce the gestalt as well as the internal details of the design facilitate the detection of various perceptual deficits. Significant distortion of or failure to copy one side of the figure can indicate the presence of hemispatial neglect.

Table VI
Motor Functioning

Test name	Description
Finger oscillation test	Finger tapping speed is measured by having the patient tap a key as quickly as possible over a period of ten seconds, using the index finger. Each hand is tested a number of times and trial totals are averaged. Poor performance consists of slow tapping speed. Unilateral motor weakness can be assessed by comparing tapping speeds of each hand. Bilateral weakness is assessed through comparison with age-matched norms.
Hand dynamometer	Grip strength is measured in each hand by having the patient squeeze a pressure-calibrated instrument. Unilateral motor weakness can be assessed by comparing performance with each hand. Bilateral weakness is assessed through comparison with age-matched norms.
Grooved pegboard	Measures of fine motor speed and dexterity, entailing placement of pegs in a pegboard, are obtained with each hand separately. Poor performance consists of difficulty grasping and manipulating the pegs, resulting in slowed performance.
Reitan-Klove Sensory-Perceptual Examination	Collection of measures of tactile, auditory, and visual perception using unilateral and double simultaneous stimulation. Finger tip number writing, visual fields, and tactile finger recognition are tested.

of psychiatric illness to the patient's presentation and diagnosis (Table VII). However, it is important to understand that many neurological illnesses can cause disorders of affect and mood. In addition, certain neurological illnesses can produce symptoms that overlap with particular psychiatric disorders. Therefore, neuropsychologists must have a thorough understanding of the psychiatric profiles associated with various neurobehavioral syndromes and proceed with caution when evaluating psychiatric symptoms in the presence of neurological or medical illness.

I. Dementia Screening Tools

Because detection of age-related cognitive impairments and dementia plays a significant role in neuropsychological evaluation, a number of specific screening tools have been developed (Table VIII). Many of these measures are designed to quickly assess gross level of functioning in major cognitive domains and can be administered in a matter of minutes (see *Blessed Dementia Scale*, *MMSE*). The *Mattis Dementia Rating Scale* is a more comprehensive set of items designed to stage severity of impairment in patients with known dementia.

VII. CONCLUSIONS

Although clinical neuropsychology is a relatively new discipline, it traces its roots back to the ancients who wondered about the "seat of the soul" and the source of human thought. Drawing from the earliest studies of brain–behavior relationships in patients with naturally acquired lesions and the accumulating body of

Table VII
Psychological Functioning and Mood

Test name	Description
Beck Depression Inventory	This instrument is used to assess depression severity based on self-reported ratings of a number of different relevant symptoms. Higher scores indicate greater severity of symptoms.
Beck Anxiety Inventory	This instrument is used to assess anxiety severity based on self-reported ratings of a variety of somatic, cognitive, and psychological symptoms of anxiety. Higher scores indicate greater severity of symptoms.
Minnesota Multiphasic Personality Inventory-2 (*MMPI-2*)	This series of 537 true/false questions load on to a number of different subscales that correspond to various personality traits or types of psychopathology. Scores on each subscale are standardized. Combinations of high and low scores on individual subscales correspond differentially to the presence or absence of various psychopathologies. Careful interpretation of subscale scores is crucial to the accurate use of this measure.

Table VIII
Dementia Screening Tools

Test name	Description
Blessed Dementia Scale	Consists of two parts: (1) a rating scale of activities of daily living to be completed by a caregiver or other independent rater and (2) a brief screening measure of orientation, concentration, and memory. Higher scores indicate greater severity of dementia.
Mini Mental Status Exam (MMSE)	This is a set of brief tasks that can be administered at the bedside and used to screen for obvious cognitive impairment. Includes items that assess attention, orientation, language, memory, and construction. Lower scores indicate greater severity of dementia.
Mattis Dementia Rating Scale (DRS)	This scale assesses a wide range of neuropsychological domains, including attention, initiation and perseverance, construction, conceptualization, and memory. It is used for grading and tracking the overall degree of dementia. Lower scores indicate greater severity of dementia.

literature in cognitive neuroscience, neuropsychologists have created a diverse collection of test instruments and other assessment methods which permit the precise measurement of specific components of human thinking. Together with a comprehensive knowledge of neuropathologic syndromes, the results of neuropsychological examination provide both descriptive and diagnostic information regarding the condition of the brain. As in all venues of medicine, diagnostic precision is a prerequisite for sound therapy. Because of its relative newness as a clinical discipline, several applied paradigms currently exist and there is some debate regarding preferred methods. Neuropsychologists in the 21st century face the challenge of establishing and promulgating an approach to assessment and treatment that unifies clinical practice and embraces the need for progressively refined normative information suited to diverse patient populations.

See Also the Following Articles

BEHAVIORAL NEUROGENETICS • BRAIN LESIONS • COGNITIVE PSYCHOLOGY, OVERVIEW • INTELLIGENCE • NEUROPSYCHOLOGICAL ASSESSMENT, PEDIATRIC • PHINEAS GAGE • PSYCHONEUROENDOCRINOLOGY • PSYCHONEUROIMMUNOLOGY • PSYCHOPHYSIOLOGY

Suggested Reading

Feinberg, T., and Farah, M. (1997). *Behavioral Neurology and Neuropsychology*. McGraw-Hill, New York.

Grant, I., and Adams, K. M. (1998). *Neuropsychological Assessment of Neuropsychiatric Disorders*, 2nd ed. Oxford University Press, New York.

Heilman, K., and Valenstein, E. (1993). *Clinical Neuropsychology*, 3rd ed. Oxford University Press, New York.

Jarvis, P. E., and Barth, J. T. (1994). *The Halstead–Reitan Neuropsychological Battery: A Guide to Interpretation and Clinical Applications.* PAR, Odessa, FL.

Kaplan, E. (1983). Process and achievement revisited. In *Toward a Holistic Developmental Psychology* (S. Wapner and B. Kaplan, Eds.). Erlbaum, Hillside, NJ.

Kaplan, E. (1988). A process approach to neuropsychological assessment. In *Clinical Neuropsychology and Brain Function: Research, Measurement, and Practice* (M. Dennis, E. Kaplan, M. Posner, *et al.*, Eds.), American Psychological Association, Washington, DC.

Kolb, B., and Wishaw, I. Q. (1995). *Fundamentals of Human Neuropsychology*, 5th ed. W. H. Freeman, New York, NY.

Lezak, M. (1995). *Neuropsychological Assessment*, 3rd ed. Oxford University Press, New York, NY.

Mesulam, M. M. (2000). *Principles of Behavioral and Cognitive Neurology,* 2nd ed. Oxford University Press, Philadelphia, PA.

Mitrushina, M. N., Boone, K. B., and D'Elia, L. F. (1999). *Handbook of Normative Data for Neuropsychological Assessment*. Oxford University Press, New York, NY.

Reitan, R. M., and Wolfson, D. (1993). *The Halstead–Reitan Neuropsychological Test Battery: Theory and Clinical Interpretation*, 2nd ed. Neuropsychology Press, Tucson, AZ.

Spreen, O., and Strauss, E. (1998). *A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary*, 2nd ed. Oxford University Press, New York, NY.

Vanderploeg, R. D. (2000). *Clinician's Guide to Neuropsychological Assessment*, 2nd ed. Erlbaum, Mahwah, NJ.

Walsh, K. W., and Darby, D. (1999). *Neuropsychology: A Clinical Approach*, 4th ed. Churchill–Livingstone, Edinburgh, UK.

Walsh, K. W. (1992). Some gnomes worth knowing. *Clin. Neuropsychol.* **6,** 119–133.

Neuropsychological Assessment, Pediatric

LAWRENCE C. HARTLAGE
Augusta Neuropsychology Center, Georgia

GLOSSARY

attention deficit hyperactivity disorder Condition presumably reflecting delayed maturation of a given brain area, characterized by poor attention, concentration, and possibly impulse control relative to those expected for the child's age.

cortex Area on the lateral surface of the brain involved with higher mental functions.

dyslexia Generic name for poor reading ability. Specific dyslexia is related to dysfunction in the area of the supramarginal angular gyrus of the dominant cerebral hemisphere. Dyslexia may also represent a manifestation of generalized language disorder, reflecting dysfunction of broader areas of the dominant cerebral hemisphere.

frontal lobes Involved with impulse control, planning, and executive function. Maturation rate is quite slow, typically not reaching maturity until the teenage years in girls and later in boys.

neurobehavioral Behavioral functions (e.g., monitoring behavior, judgment, impulse control) dependent on or subserved by the (central) nervous system.

neurocognitive Cognitive functions (e.g., memory, intelligence) dependent on or subserved by the (central) nervous system.

psychometric Measurements of psychological functions with objective, calibrations, such as intelligence quotient or learning levels, corrected for age.

Pediatric neuropsychological assessment deals with the evaluation of brain–behavior relationships in a pediatric population. Such assessment measures the developmental status and functional integrity of given brain areas and systems by utilizing tests and measurement procedures based on psychometric principles of psychological science.

I. BACKGROUND

Neuropsychological assessment grew from infancy to maturity in the half century between 1950 and 2000. Originally developed with a primary focus on the assessment of neurocognitive correlates of neurological disorders in adults, by the 1970s neuropsychological assessment had expanded to encompass neurodevelopmental phenomena among children. Pediatric neuropsychological assessment in its early years was characterized by the use of measurement concepts and instruments found to be of value with adults, and early practitioners were often individuals sophisticated in the physiological bases of adult cognitive dysfunction with little clinical experience with or knowledge of a pediatric population.

Growing awareness and interest in brain–behavior relationships in children, typified by such issues as minimal brain dysfunction or attention deficit hypotheses to account for comparatively poorer performance by some children in academic and social milieus, generated government support during the Great Society era of a number of training programs designed to incorporate pediatric neuropsychological assessment into both health care and academic settings. The increasing involvement of individuals with a background in child development and the problems unique to a pediatric population heightened recognition that

conceptual models typical of brain–behavior relationships in adults were not necessarily applicable to children. This recognition has led to three decades of research and development involving issues of unique relevance to pediatric neuropsychological assessment.

II. CONCEPTUAL ISSUES IN PEDIATRIC NEUROPSYCHOLOGICAL ASSESSMENT

Neuropsychological assessment of a pediatric population differs in a number of respects from such assessment of an adult population. Whereas a primary feature of adult neuropsychological assessment is that it is based on the loss or impairment of a previously acquired function, child neuropsychological assessment is concerned with a comparable function that may be slowed in development, limited in its extent of development, or totally precluded from development. A critical feature in this respect involves the age at which insult to the central nervous system is sustained and, to a lesser but interactive extent, the brain area or functional system sustaining insult. For more than 100 years it has been documented that, in an adult, damage to a specific brain area (i.e., in the dominant cerebral hemisphere foot of third frontal convolution) may produce loss of the ability to speak (i.e., Broca's type aphasia). However, damage to the same area in a very young child may produce no apparent problems in the development of expressive speech due to the recruitment of intact areas of the nondominant cerebral hemisphere to subserve this function. Neuropsychological assessment of such children, however, may reveal impairment of the functions typically subserved by cortical areas that have been recruited to accommodate the mechanisms of expressive speech. Interestingly, there appears to be an upper age level, somewhere around 8 years of age, after when such contralateral recruitment no longer occurs or occurs to a substantially attenuated extent.

Obviously the diagnosis of an ability (e.g., expressive speech) that has been acquired and then lost is much easier than that of the impaired emergence or slowed rate of development of a skill not within the individual's prior repertoire. Accordingly, neuropsychological assessment of pediatric age populations requires considerable knowledge of the expected ages at which neurodevelopmental phenomena affecting cognitive and behavioral adjustment may be expected to occur, as well as the normal sequence of stages through which such development occurs.

This age specificity constitutes perhaps the most important difference between neuropsychological assessment in pediatric as opposed to adult populations. In the adult population the span of a decade—say between 30 and 40 years of age—produces relatively little change in neurological maturation or organization, and level of performance on most diagnostic measures can, for the most part, be interpreted in the same way for a younger vs an older individual in this age range, with little differentiation between genders except on measures of strength. Conversely, the difference in neuropsychological repertoire between a 2 year-old and a 7-year-old is at once remarkable and striking, reflecting the phenomenal changes in neurological organization occurring during this 5-year period. Accordingly, pediatric neuropsychological assessment needs to consider and take into account the emerging states of central nervous system structures and functions subserving given abilities measured by assessment and be prepared to incorporate such considerations into a diagnostic profile.

In a typical 2-year-old, receptive language is usually reflected in the ability to recognize a few words of command and to match pictures of a few familiar objects to the spoken word (e.g., cat, cup, chair) associated with such objects. By age 7 years, brain maturation has occurred to an extent sufficient to permit the recognition of abstract symbols representing letters of the alphabet and words composed of such symbols. Conceptually, this age may also include understanding of bigger–smaller, pretty–ugly, and alike–different at varying levels generally corresponding to younger vs earlier stages of this 5-year range. Obviously, pediatric neuropsychological assessment needs to be based on an understanding and appreciation of such development to permit understanding of whether a given child's performance on given tests suggests dysfunction in the central nervous system. An example of the interaction between developmental level and specific neurocognitive function may help to provide an illustration of the interdependence of development and central nervous system, normally vs abnormally.

Consider a specific function mediated by the central nervous system, expressive language. Expressive language at a 5-year-old level in an 8-year-old child of normal intellectual ability may be indicative of brain dysfunction involving frontal areas of the dominant cerebral hemisphere, whereas such expressive language functions in an 8-year-old of borderline intellectual ability may represent nothing more remarkable than a manifestation of intellectual function.

More complex functions mediated by the central nervous system in children are encountered with considerably greater frequency than single problems such as expressive language, which may ultimately depend on a single location in the brain (i.e., foot of the second frontal convolution). A diagnostic question frequently presenting for pediatric neuropsychological assessment involves that of reading difficulty. Obviously, before the possible central nervous system components of such a problem can be assessed, the pediatric neuropsychologist must rule out or at least consider the potential effects of such factors as poor educational instruction, mismatch between educational instructional procedure and the child's learning style, possible language barriers, sensory problems involving sight or hearing, impaired motivation, genetic factors, emotional problems, or health factors including medication use that might be operative. Next, the evaluation of the child's reading problem needs to proceed along anatomical parameters. Is there a specific word blindness that might implicate brain areas in the supramarginal–angular gyrus area of the dominant cerebral hemisphere? Does the problem involve associational deficits corresponding with frontal lobe functions? Can the child actually perceive and recognize words but have difficulty translating this into an appropriate response? Is there dysfunction among functions mediated by the arcuate fasciculus?

III. DEVELOPMENTAL ASYMMETRY AND ASYNCHRONICITY

In right-handed individuals and approximately 94% of left-handed individuals, the left cerebral cortex is specialized for language functions and sequential information processing, whereas the right hemisphere is specialized for spatial functions and wholistic or intuitive information processing. Just as children demonstrate growth spurts or uneven patterns of physical growth, the evolutionary functions of given brain areas tend to show asymmetric, nonlinear growth patterns in what some child development specialists have referred to as brain growth periodization. A child may show delayed language production with accelerated motor skills and then demonstrate accelerated language growth corresponding to dominant hemisphere growth acceleration. The importance of this asymmetry and asynchronicity in pediatric neuropsychological assessment is obvious: over-reaction to such false positive indices of abnormality leads to over-diagnosis of abnormality, whereas failure to account for these phenomena may miss important clues to emerging problems and the opportunity for timely intervention. Over-diagnosis of the presentation of normal variations in central nervous system developmental symmetry and synchronicity can lead to unnecessary parental concern, initiation of costly and unnecessary interventions and treatment, and even precipitate spurious lawsuits for damages when none have occurred. Conversely, under-diagnosis can produce problems resulting from lack of timely intervention, whether from early stages of arteriovenous malformations or neoplasms or from language delays wherein critical periods for intervention may be missed. Issues of asymmetry and asynchronicity tend to be more pronounced during younger ages when brain growth is most rapid and tend to become less critical as children approach the end of the pediatric period and segue into young adulthood. However, at this stage the evolution of anterior cortical brain areas—heavily influenced by gender differences as previously noted—begins to represent an increasingly important facet of pediatric neuropsychological assessment. Factors such as impulse control, consideration of alternatives and consequences, monitoring and self-evaluation of responses, and executive functions become progressively important foci of pediatric neuropsychological assessment. Adolescents with delayed frontal maturation may present on examination with findings of abilities spuriously depressed due to impulsive responding or failure to consider attenuators. Knowledgeable and sophisticated examiners recognize the possible effects of such determinants and will structure the examination to detect such possible confounds and correct for their impact on findings.

A. Clinical Manifestations of Developmental Asymmetries

For many years, it has been documented that depressed dominant (almost always left) cerebral hemisphere development will lower verbal IQ and depressed nondominant hemisphere development will lower performance IQ. Similarly, developmental tasks such as acquisition of reading skills or expressive language are primarily dependent upon dominant hemisphere maturation (in the posterior and anterior portions, respectively). Thus, the measurement of developmental phenomena relevant to the acquisition of such academic pursuits as learning to read and spell or to recognize and reproduce images of given objects

depends on the developmental status of cortically mediated functions, which in turn are dependent on the development of given cortical areas. Neuropsychological assessment of pediatric age patient is an important adjunct and contributes to the understanding of skill development (or lack of development) of a number of functions central to the acquisition of educational mastery at given age levels.

IV. DIAGNOSIS OF NEUROPSYCHOLOGICAL DISORDERS

Although a wide range of classification procedures have been developed for neuropsychological disorders in a pediatric age population, a relevant dichotomization may segregate such disorders into those representing conditions whose treatment is typically relegated to medical specialists vs those conditions whose alleviation or remediation is more commonly within the domain of educational–behavioral interventions.

A. Disorders Involving Medical Intervention

1. Neoplasms

Perhaps primary among such disorders would be neoplasms or tumors, because early detection and intervention can have crucial implications. Whereas primary brain tumors do not typically occur with high frequency in a pediatric population, neuropsychological assessment can detect them at an early stage, frequently before the child would be referred for computerized tomography or other radiological procedures. This is due to the fact that neuropsychological assessment measures *functional* processes, such as interaction of functional systems of the brain, whereas X-ray procedures do not reflect abnormalities until some structural alteration of the brain is manifested. It is not uncommon for changes in rate of development to become manifested before structural changes (e.g., tumor growth) in the brain are suspected. Whereas surgical intervention would not be based solely on neuropsychological findings, such findings can and often do serve as a first documentation that neoplastic growth is occurring. For more than a quarter of a century, it has been recognized that neuropsychological assessment can provide accurate information concerning location, size, and growth velocity of neoplasms, information that can be applied to make inferences concerning the specific nature of the tumor. Other conditions, such as arteriovenous malformations or epileptogenic seizure foci, are similarly amenable to neuropsychological evaluation.

2. Traumatic Brain Injury

A very frequent occurrence among children of all ages involves traumatic brain injury. Such injuries can occur at various levels, classified as severe (e.g., extended coma, multiple potential residua), moderate (briefer coma, typically fewer physical developmental residua), or mild (brief or no loss of consciousness). Most common causes involve automotive injuries, falls, and sports injuries. The large majority of traumatic brain injuries in the pediatric age population are typically mild concussive incidents with little or only brief loss of consciousness. Diagnosis of such injuries by standard imaging (CAT or MRI scan) or electroencephalographic (EEG) procedures are often not productive, because such procedures are not particularly sensitive to such injuries. That such injuries are important to diagnose and can have serious consequences is, unfortunately, not widely recognized. Neuropsychological examination is consensually recognized as the most definitive approach for determining whether neurocognitive, neurobehavioral, or both types of problems have resulted. Whereas some 80–85% of mild concussive disorders tend to show recovery without treatment, a second mild blow to the head shortly following an initial mild concussion can produce coma and death. Further, among 15–20% of children who do not make spontaneous recovery, impairments may range from social inappropriateness, depression, irritability, learning difficulty, and intellectual slowing to increased risk of delinquency. Even among many of those children who appear to have made a spontaneous recovery, long-term follow-up has shown mild residual neuropsychological deficits, especially as a result of automotive injuries or repeated very mild blows to the head such as are experienced among children involved in football, soccer, and boxing. Research on both former soccer players and former boxers has found the degree of neuropsychological deficit to be directly related to the amount of time involved in participation in the sport, with a similar but somewhat weaker correlation found between such deficits and time involved in participation for football players.

Neuropsychological assessment of these children is important for a number of reasons. Primarily, it can identify those children at risk for "second impact"

injuries, and aid in the prevention of such potentially serious or even fatal injuries. It can also provide the basis for understanding changes in personality or school performance that might ordinarily be referred to a mental health professional or educational therapist for treatment. It can provide guidelines for parents concerning substrates of learning or behavioral difficulties, with guidelines for management and longer term planning. In cases of liability, such as in automotive injuries, neuropsychological assessment can provide definitive documentation concerning the nature and extent of brain injuries related to or resulting from the given incident, and the likely effect(s), if any, of that injury on the child's longer term educational and personal–social adjustment and eventual vocational potential.

3. Chemical Exposure

Although not widely recognized, neurotoxic chemical exposure is not uncommon among children and exerts considerably more adverse and deleterious effects than among adults. Particularly among younger children, whose brains are growing rapidly, the metabolism of neurotoxic substances like lead takes place faster than with adults. Further, the comparatively smaller size of children makes them especially vulnerable: a serving of mercury-contaminated fish may realistically be expected to have a greater effect on a 25-lb child than on a 150-lb adult. Further, whereas the adult brain typically has established patterns of speech, language, memory, and problem-solving, insult to these systems by neurotoxic exposure represents a subtractive process that is not likely to obliterate an entire system. Conversely, among children, especially younger ones who do not yet have fully developed systems of receptive or expressive language, social behavior, or judgment, insult to the central nervous can interfere with or even prevent the development of such functional systems in the brain. Neuropsychological assessment often provides the first documentation that an interruption of central nervous system development has occurred, thus leading to specific definitive testing for blood (urine, hair, bone) samples to determine the specific neurotoxins to which the child has been exposed. Once such a determination has been made, neuropsychological assessment can ascertain how much damage has occurred to specific functions, such as short-term memory, attention, and concentration, and establish educational, behavioral, and related goals and expectancies on the basis of remaining intact functional systems. Finally, neuropsychological assessment at intervals over an extended time can monitor response to treatment and identify emerging manifestations of neuropsychological deficits relative to the given age expectancies.

This listing by no means exhausts the range of neuropsychological assessment applications to medically determinable phenomena, but it does represent illustrative elaborations of how such assessment may relate to three types of medical problems.

B. Disorders Involving Educational–Behavioral Intervention

It is noteworthy that the medical phenomena just described, as well as a wide range of other medical phenomena with neuropsychological substrates (e.g., HIV–AIDS, sickle cell anemia, hypoxic episodes, and degenerative disorders, genetic and metabolic disorders), may profit from educational modifications to accommodate neuropsychological anomalies associated with the disorder, as well behavioral intervention in the form of counseling for educators and parents concerning reasonable expectations and the need for behavioral support in form of behavioral modification or supportive therapy intervention.

1. Attention Deficit Hyperactivity Disorder

Attention deficit hyperactivity disorder (ADHD) represents a manifestation of neuropsychological disorders estimated to occur in approximately 5% (or more, depending on source) of school-aged children. Boys are referred approximately 3–4 times more frequently than girls. Neuropsychological assessment is indicated for determining the exact nature of the disorder: (1) classic (hyperactive, restless, distractible); (2) inattentive (absent-mindedness without hyperactivity); and (3) temporal lobe (aggressive, memory problems). Whereas SPECT imaging can document the physiological basis of these types of ADHD, neuropsychological assessment is method of choice for documenting the presence and type of disorder and recommended intervention approaches.

2. Learning Disabilities

Learning disabilities are found in some 5% (or more, depending on source) of school-aged children. In most states, a learning disability is documented by a 15-point or greater (i.e., 1 standard deviation or more), difference between ability as measured by individual

IQ test and school achievement, as measured by standardized individual achievement testing measuring basic skills such as reading, arithmetic, and spelling. Although arithmetic disorders are not particularly rare, language-related learning disabilities exert by far the greater impediment to learning and are the most frequent learning disorders noted. Language learning disorders can be of two generic types.

a. Language Learning Disabilities Language learning disability, reflecting generalized neurodevelopmental asymmetry involving the dominant cerebral hemisphere, is the more common type of learning disability seen for neuropsychological evaluation. The learning problem may involve recognition–decoding problems related to more posterior brain structures (e.g., angular gyrus), associative problems involving the accurate fasiculus–temporal lobe areas, expressive functions more related to frontal cortical areas, or any combination of these problems. Typically, problems are manifested to some extent in all three functional systems.

b. Specific Development Dyslexia Specific developmental dyslexia was first described in the late nineteenth century as congenital word blindness and is usually related to dysgenesis in the area of the supramarginal and angular gyri of dominant cerebral hemisphere. Anatomic differences in brain morphology have been demonstrated upon autopsy of individuals with this type of dyslexia. Tending to occur approximately 4:1 in boys over girls, there appears to be a genetic pattern. This problem may occur in individuals at any intellectual level and, in fact, is often not detected until around the 4th grade level due to ingenious subterfuges and compensatory mechanisms employed by individuals with this specific reading disorder. Whereas some cases of this disorder occasionally may be detected by school psychometric testing, the definitive diagnosis is dependent on neuropsychological assessment.

V. CONCLUSION

Neuropsychological assessment of pediatric age patients is an important component of documentation of the presence of a disorder subserved by the central nervous system, of a nature that may present as a psychiatric disorder, stubbornness, failure to apply oneself to learning, or related behavioral or cognitive problems. By relating the maladaptive or dysfunctional behavior (or misbehavior) to its cortical substrates, neuropsychological assessment provides diagnostic explanatory information that can represent a major source of understanding of the disorder, as well as provides rationale for the development of appropriate treatment and intervention.

The precise nature of the measurements involved, when compared with the more clinical measures used in neurological or neurosurgical evaluation, provide highly quantifiable and reliable data that can be compared to normative expectancies, used to measure responsiveness to a given drug or dosage of a given drug affecting the central nervous system, tracked over time to assess improvement or decline, or used to predict ecological phenomena such as potential to perform certain tasks. In this regard, neuropsychological assessment can be used for determining whether a child may be safely cleared to return to a sport activity, expected to perform at a given expectancy level for an academic task, or to expect success in a given curriculum focus or pedagogic procedure.

See Also the Following Articles

ADOLESCENT BRAIN MATURATION • BRAIN DEVELOPMENT • COGNITIVE PSYCHOLOGY, OVERVIEW • DYSLEXIA • LANGUAGE DISORDERS • MENTAL RETARDATION • NEUROPSYCHOLOGICAL ASSESSMENT

Suggested Reading

Barkley, R. A. (1998). *Attention Deficit Hyperactivity Disorder*. Guilford, New York.

Brown, R. T., and Sawyer, M. G. (1998). *Medications for School Age Children*. Guilford, New York.

Fisher, K. W., and Rose, S. P. (1994). Dynamic development of coordination of components in brain and behavior. In *Human Behavior and the Developing Brain* (G. Dawson, and K. W. Fischer, Eds.), Guilford, New York.

Hartlage, L. C. (1998). Assessment of mental development in children. *Behavior and Medicine*, 3rd ed. C. V. Mosby, New York.

Hartlage, L. C., and Lucas, D. B. (1973). *Mental Development Evaluation of the Pediatric Patient*. Charles C. Thomas Co., Springfield, IL.

Hartlage, L. C., and Williams, B. L. (1997). Pediatric neuropsychology. In *The Neuropsychology Handbook* (A. M. Horton, D. Wedding, and J. Webster, Eds.), 2nd ed., Chapter 7, pp. 211–236. Springer, New York.

Templer, D. I., Hartlage, L. C., and Cannon, C. W. (1991). *Preventable Brain Damage: Brain Vulnerability and Brain Health*. Springer, New York.

Yeates, K. O., Ris, M. D., and Taylor, H. G. (Eds.) (2000). *Pediatric Neuropsychology*. Guilford, New York.

Neurotransmitters

JAMES H. SCHWARTZ
Columbia University

GLOSSARY

biosynthetic pathway A series of enzymatic reactions by which a substance is made.

cholinergic Denotes a neuron that uses acetylcholine as transmitter.

cholinoceptive Denotes a target cell with a receptor for acetylcholine.

endoplasmic reticulum Polyprotein precursors of neuropeptides (and other secreted proteins) are translated from messenger RNA on polyribosomes attached to the cytoplasmic surface of the endoplasmic reticulum, a series of membranous tubules. The nascent polypeptide chain is extruded into the lumen of these tubules, processed, and then sent to the Golgi apparatus in transport vesicles.

Golgi apparatus A complex structure consisting of a series of tubules arrayed in stacks. Its primary function is to process newly synthesized proteins and package them for transport to the external cell membrane.

receptors for neurotransmitters As a rule are transmembrane proteins located on the postsynaptic surface of a target cell. Ionotropic receptors form a channel for the passage of ions when they bind the transmitter. Metabotropic receptors interact with second messenger systems when the transmitter is bound, often through a G-protein. The receptors for some messengers (e.g., NO and arachidonic acid) are enzymes or factors within the target cell.

Chemical synaptic transmission is mediated by the release of one or more substances from a nerve terminal. About 100 neurotransmitters have been identified. These belong to two categories: neuropeptides and small molecules. Neurotransmitters produce their action by binding to receptors in target cells.

I. INTRODUCTION

A neuron's key function is to communicate with other neurons or with target cells (gland or muscle). Communication is mediated either electrically or chemically. Electrical synapses are rather stereotyped. In contrast, chemical transmission, which involves the slow to rapid release of one or more of over 100 possible transmitter substances, is quite complex. This complexity, which also includes the time and amount of transmitter released, is the basis of synaptic plasticity. Synaptic plasticity, in turn, is the basis of higher mental function, for example, learning and memory. Electrical transmission essentially is instantaneous because the necessary cytoplasmic connection between the neuron and its target cell is continuous. In contrast, in chemical transmission, the target cell is situated at some distance from the signaling neuron, causing synaptic delay. For a transmitter released from the signaling neuron to act successfully as a chemical messenger, it must bind to an appropriate receptor in the target cell. Typically the separation is not much greater than the distance between two adjacent

neurons (the synaptic cleft is about 200 μm across), but occasionally the separation can be much greater.

Chemical transmission is a specialized form of secretion. In simple organisms and during embryogenesis, communication between cells is mediated only by peptide growth factors. These are hormones that are secreted by one cell and diffuse to relatively distant targets. Neurotransmitters can be categorized into two groups: neuropeptides and small molecules (classical transmitters). Each group uses a distinct set of molecular mechanisms that are familiar from general cell biology: neuropeptides are synthesized, processed, and packaged within vesicles of the *secretory* pathway. In contrast, small-molecule neurotransmitters are made in the cytoplasm and pumped into vesicles derived from the *lysosomal* pathway.

It is important to realize that neurons are constructed according to the general polarized plan of epithelial cells. Toward the end of the nineteenth century, Santiago Ramón y Cajal discussed the building plan of neurons in terms of functional polarity. He pointed out that a typical nerve cell has a receiving end (dendrites), a cell body (soma), and a transmitting end (axon). This corresponds to the basal-to-apical blueprint of a typical epithelial cell. In neurons, most of the Golgi apparatus is situated between the cell's nucleus and the axon hillock. Membranous vesicles filled with newly synthesized secretory products are transported from the *trans* face of the Golgi apparatus to the axon and from there to nerve terminals for release.

II. FOUR CRITERIA DEFINE A NEUROTRANSMITTER

The ideas defining a neurotransmitter have changed over the past century as a consequence of growing knowledge of cell biology, pharmacology, and electrophysiology. As late as 1940, there was debate about whether chemical transmission plays any role in synaptic transmission. Some investigators, John C. Eccles among them, believed that all synaptic signals were electrical. By midcentury, however, everyone agreed that there are two forms of transmission, chemical and electrical. Eccles had changed his mind to become one of the foremost investigators of the physiology of synapses.

Most influential was a simple experiment performed by Otto Loewi in 1921. It was known that the heartbeat is slowed when the vagus nerve is stimulated. In order to prove that a diffusable substance is released upon stimulation, Loewi bathed two frog hearts in a physiological saline solution. Both hearts beat at a normal rate. Then he stimulated the vagus nerve to one of the hearts. That heart slowed, as expected. But the heart with the unstimulated vagus also slowed, indicating that a substance, which Loewi called *Vagusstoffe*, was released into the bath to slow the other heart. Soon afterward, *Vagusstoffe* was identified as acetylcholine (ACh), a substance previously found in high concentrations in nervous systems. Long afterward, Loewi told the story about how this influential experiment occurred to him in dreams on two successive nights. In an autobiographical sketch, he tells how the design for the experiment came to him on the first night but was forgotten when he awoke. On the next night, he made sure he had pencil and paper ready and successfully noted the protocol of the dream experiment.

Two of the four criteria used today for deciding whether a substance is in fact a neurotransmitter are suggested by Loewi's experiment: (1) the substance must be synthesized by the neuron, and (2) the substance must be concentrated in nerve terminals and released when the neuron is stimulated. The amount of transmitter released should, of course, be sufficient to affect the postsynaptic neuron or target cell. The third criterion is pharmacological: (3) the putative transmitter must mimic the action of the substance released from the neuron when applied to the target cell in reasonable concentrations.

The last requirement reflects an important aspect of synaptic transmission: (4) mechanisms must exist for removing the putative transmitter from its target. Removal of the released transmitter is a requirement for transmitting a meaningful message. Failure to remove the transmitter would result either in a persistently activated postsynaptic element or, more commonly, a block of transmission because of postsynaptic receptor desensitization.

III. THERE ARE MORE THAN 100 DIFFERENT NEUROTRANSMITTERS

The classification of transmitters into neuropeptides and small molecules is based on how the transmitter is made. Typically neuropeptides are synthesized on polyribosomes that attach to the endoplasmic reticulum during translation. Characteristically, a polyprotein, while it is translated from messenger RNA, is extruded into the lumen of the endoplasmic reticulum and is then transported to the Golgi apparatus. During

its passage through the major membrane systems of the cell, the polyprotein is cut into peptides, which ultimately are packaged into secretory vesicles. Some of these peptides had been identified earlier as hormones in other tissues, notably the gut (for example, gastrin), or as neurosecretory products (for example, oxytocin, vasopressin, somatostatin, and thyrotropin-releasing hormone). A partial list of neuropeptides is presented in Table I.

In contrast, small-molecule neurotransmitters are synthesized in the neuron's cytoplasm, usually in short enzymatic pathways much like intermediates of metabolism (Table II). There are three types of small-molecule neurotransmitters. The first is amine transmitters, which are amines or molecules derived from amines in a few enzymatic steps. This group includes (1) ACh, (2) biogenic amines, and (3) various amino acids. Because these molecules were discovered first historically, they are often referred to as classical neurotransmitters. Loewi's experiment with acetylcholine (1921) has already been mentioned; the pharmacology of norepinephrine (NE) and related amines was being studied contemporaneously by Henry H. Dale. By the middle of the twentieth century, ideas about neurotransmitters were based almost entirely on experiments with these two small-molecule transmitters.

Table I
Some Families of Peptide Neurotransmitters

Family	Neurotransmitter
Opioid	Opiocortins, enkephalins, dynorphin, FMRFamide
Secretin	Secretin, glucagon, vasoactive intestinal peptide, gastric inhibitory peptide, growth hormone releasing factor, peptide histidine isoleucineamide
Tachykinins	Substance P, substance K (neurokinin A), neurokinin B
Neurohypophyseal	Vasopressin, oxytocin, neurophysins
Pancreatic polypeptide related	Pancreatic polypeptide, NPY
Gastrin	Gastrin, cholecytokinin
Others	
Calcitonin gene-related peptide(CGRP)	
Calcitonin	
Galanin	
Cholecytokinin (CCK)	

Table II
Small-Molecule Neurotransmitters and Their Neuron-Specific Enzymes

Transmitter	Enzyme
Amine	
Acetylcholine	Cholineacetyl transferase
Biogenic amine	
Dopamine	Tyrosine hydroxylase
Norepinephrine	Tyrosine hydroxylase
	Dopamine β-hydroxylase
Epinephrine	Tyrosine hydroxylase
	Dopamine β-hydroxylase
	Phentolamine-*N* methyltransferase
Indoleamine	
Serotonin	Tryptophan hydroxylase
Amino acid	
Glutamate	
GABA	Glutamic acid decarboxylase
Glycine	
Membrane soluble	
Nitric oxide	Nitric oxide synthase
Arachidonic acid	

Another group of small-molecule transmitters, the purines, is derived from adenosine triphosphate (ATP). Like the amine transmitters, ATP is not made within the tubules of the endoplasmic reticulum or in the stacks of the Golgi apparatus, but in mitochondria. How then do the amine transmitters and ATP get into vesicles, typically in rather high concentrations (0.01–0.5 *M*)? The answer is that membranes of vesicles contain special pumps that effectively transport small-molecule transmitters into the lumen of the vesicle. Because all transmitter vesicles contain ATP, when a transmitter is released, ATP and its catabolites, adenosine and AMP, are also released. Finally, the most recent substances to be recognized as transmitters are molecules that are membrane-soluble. Thus, nitric oxide (NO) and arachidonic acid pass from one neuron to another.

IV. NEUROPEPTIDE TRANSMITTERS

Neuroactive peptides, typically 3–15 amino acid residues in length, are processed from precursor

polyproteins that contain about 200 residues. In many precursor polyproteins there are multiple copies of a single neuropeptide sequence. These can be cut from the polyprotein so that several copies of the same peptide are produced, thereby amplifying the concentration of the peptide transmitter released. A polyprotein also can contain the amino acid sequences of several different neuropeptides. These, when cut from the precursor, will all reside in the same vesicle. As a consequence, a set of neuropeptides, usually with synergistic action, is released at the same synapse from a single vesicle.

Differential processing of a polyprotein precursor can result in the production of different neuropeptides. Thus, for example, bulky glycosylation, which takes place during the passage of the nascent polypeptide through the endoplasmic reticulum and Golgi, can block proteolytic cleavage at specific sites in the precursor and result in different sets of neuropeptides. Differential processing occurs with pro-opiomelanocortin (POMC), a precursor of one of the three distinct opioid genes. In the hypothalamus, β-endorphin and α-melanocyte-stimulating hormone (α-MSH) are formed; in the anterior pituitary, POMC yields corticotropin.

The proteases responsible for processing polyproteins into neuropeptides are present within the endoplasmic reticulum and Golgi apparatus. These enzymes are related to the well-known digestive proteases, trypsin, chymotrypsin, and pepsin, which are synthesized in the endoplasmic reticulum and packaged in secretory granules of the pancreas. Trypsin and chymotrypsin are called serine proteases because the hydroxyl group of a serine residue in the enzyme's catalytic site participates in the proteolytic reaction. The other major families of proteases are related to pepsin, a thiol peptidase, and carboxyprotease, proteases that contain a zinc ion at their catalytic center. Common sites of cleavage in the polyprotein are between an amino acid residue X and a pair of dibasic residues (for example, X-Lys-Lys or X-Arg-Lys) or between X and a single basic residue.

The great diversity of neuropeptides prevents a comprehensive description of each of the approximately 100 peptides known or even of particular peptide families. The opiate family is representative, however. There are three distinct genes that encode three different polypeptide precursors: opiocortin, enkephalin, and dynorphin. Proteolytic processing of each polyprotein precursor gives rise to peptides with narcotic properties, all of which contain the amino acid sequence, Tyr-Gly-Gly-Phe. Because the three genes are similar, it is thought that they all stem from a single ancestral gene. Even so, in humans the three opioid genes are each situated on different chromosomes.

The synthesis and processing of peptide neurotransmitters take place exclusively in the cell body of neurons. Secretory vesicles containing neuropeptides, which are similar to secretory granules in other cells, bud off from the Golgi apparatus. These large vesicles (60–175 nm in diameter) appear dense-cored in the electron microscope because their contents stain with osmium salts and, therefore, are electron-dense. The vesicles are moved out of the cell body and along the axon to nerve terminals by the process of fast axoplasmic transport. Once they bud off the Golgi apparatus, the vesicles cannot be refilled, nor is there any mechanism for new synthesis of the peptide. Consequently, once the contents of a vesicle are released, the vesicle membrane cannot be reused: the release of neuropeptides thus is a one-time-only situation.

Release of neuropeptides occurs by exocytosis, which, like secretion, is triggered by a rise in intracellular Ca^{2+} ion concentration. The Ca^{2+} ion concentration, which is increased when the neuron is stimulated, must rise throughout the nerve terminal. Typically, the neuron must be stimulated for a relatively long period of time to elevate the concentration of Ca^{2+}. There are no specialized sites at which these large vesicles undergo exocytosis. Because exocytosis can occur throughout the nerve terminal, neuropeptides are released over a large target area.

Thus, transmission with neuropeptides is slow in onset, covers a broad target area, and is long-lasting. The action of neuropeptides is long-lasting first because the binding constants of postsynaptic receptors tend to be quite low (therefore, a low concentration of the peptide is effective) and, second, because the chief mechanism for removing the peptides is diffusion, which tends to be quite slow with molecules the size of these peptides (see later discussion).

Another important reason why the actions of neuropeptides are slow and long-lasting is because the postsynaptic effects of the peptides are produced by second messenger cascades. Unlike the opening or closing of ionotropic receptors, which results from rapid allosteric changes in the conformation of a channel protein, actions involving second messengers typically require a series of discrete molecular steps to bring about a physiological change. Almost all postsynaptic receptors for neuropeptides are metabotropic seven-transmembrane-spanning proteins that are coupled to G-proteins. Each neuropeptide usually has

more than one receptor. Thus, there are three for opioids (δ, κ, and μ), all cloned. Morphine and related drugs produce narcosis (relief of pain), sleepiness, euphoria, and constipation.

An important technique for demonstrating the second criterion for a neurotransmitter—its presence in a neuron's terminals—is immunocytochemistry. Neuropeptides are excellent antigens for raising antibodies. Specific antibodies are useful for locating immunoreactivity in tissue sections of nervous tissue. For example, an antipeptide antibody can be raised by immunizing a rabbit. That antibody can be applied to a tissue section, where it binds strongly to the components in the cell containing the antigen. Once bound, the antibody can be located by applying a second antibody against the immunoglobulin of the rabbit antipeptide antibody, say, goat antirabbit antibody. These secondary antibodies are available commercially labeled with fluorescent dyes (e.g., fluoroscein and rhodamine), which permits the localization of neuropeptides by fluorescence (light) microscopy. Immunocytochemistry with the electron microscope is made possible by using antibodies linked to colloidal gold particles, which are electron-dense. These particles can be made with different diameters. Therefore, it is possible to detect two different antibodies simultaneously on the same tissue section, with each antibody being tagged with gold particles of different size. In the light microscope, neuropeptide immunoreactivity typically is intense over the cell bodies and terminals of neurons. In the electron microscope, it is easy to show that vesicles are the subcellular organelles that are most heavily labeled.

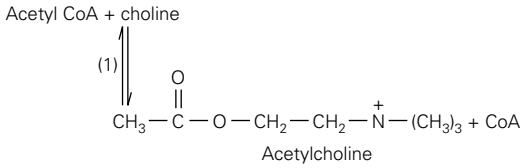
V. SMALL-MOLECULE NEUROTRANSMITTERS

Unlike neuropeptides, small-molecule neurotransmitters are synthesized in cytoplasm throughout the neuron. Thus, they can be made in nerve terminals as well as in the cell body. Small-molecule neurotransmitters are not synthesized within the major membrane system of the cell: the endoplasmic reticulum, the Golgi apparatus, or any of the vesicles that are derived from the Golgi apparatus. Rather, they are produced in short enzymatic pathways that branch off of the major pathway of intermediary metabolism. Typically, one enzyme in the biosynthetic pathway is expressed only by neurons that make use of that particular neurotransmitter (Table II). The presence of that enzyme in a neuron can be used as a biochemical criterion for that particular transmitter. In effect, it is evidence for the first neurotransmitter criterion, that the substance is synthesized by the neuron.

A. Amines

1. Acetylcholine (ACh)

ACh is formed in a single enzymatic step. The enzyme, choline acetyltransferase, catalyzes the esterification of choline by acetyl-CoA.



The transferase is specific to cholinergic neurons and is not expressed in any other cell type. (The term cholinergic is used to denote a cell that releases ACh as a neurotransmitter. Similarly, glutaminergic, dopaminergic, and serotonergic indicate that a neuron releases glutamate, dopamine, or serotonin, respectively. If a cell responds to ACh, that cell is called cholinoceptive, a term used infrequently for the other neurotransmitters; e.g., "dopaminoceptive" is unusual.) The formation of ACh is limited by the supply of choline. Choline is not made in nervous tissue, but must be obtained through the cerebrospinal fluid from dietary sources or recaptured from the synaptic cleft from the ACh released and hydrolyzed by the enzyme acetylcholinesterase (see later discussion).

There are two general classes of acetylcholine receptors (AChR): nicotinic, responding to the alkaloid nicotine, and muscarinic, responding to the mushroom poison, muscarine. ACh is excitatory at the neuromuscular junction, where it binds to postsynaptic nicotinic AChRs. As we saw with Loewi's experiment, it is an inhibitory (parasympathetic) transmitter to the heart through muscarinic AChRs. In the periphery, ACh is also the transmitter for all preganglionic neurons of the autonomic nervous system. In the brain, there are many cholinergic systems, for example, cholinergic neurons in the nucleus basalis have widespread projections to the cerebral cortex.

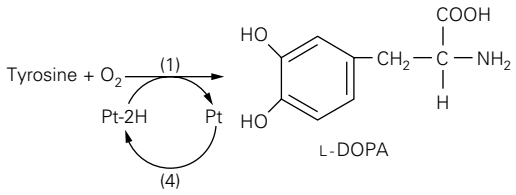
Nicotinic AChRs are *ionotropic*, meaning that, when they bind ACh, they open up to pass ions from the extracellular space into the postsynaptic neuron. Muscarinic AChRs are *metabotropic*. These receptors activate various second messenger pathways to

produce biochemical changes within the postsynaptic neuron. Thus, as with other neurotransmitters, ACh can excite or inhibit depending on the postsynaptic receptor.

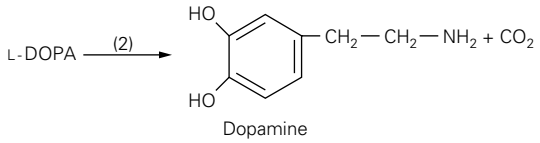
2. Biogenic Amines

The biogenic amine transmitters include the catecholamines [dopamine, norepinephrine (NE), and epinephrine], the indoleamine serotonin (5-hydroxytryptamine, 5-HT), and the imidazole histamine. Synthesis of the catecholamines and 5-HT is catalyzed by a set of enzymes that are similar in structure and mechanism. Because of amino acid sequence similarity, these enzymes are believed to come from a common ancestral molecule. The first enzymes in the synthesis of catecholamines and 5-HT are oxidases that require a pterin (tetrahydrobiopterin, THB) cofactor. These enzymes are specific and rate-limiting.

a. Dopamine For the synthesis of catecholamines, the essential amino acid L-tyrosine is converted to L-dihydrophenylalanine (L-DOPA).

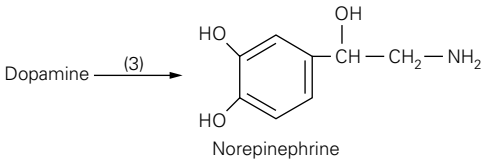


Tyrosine must be supplied in the diet and taken up by neurons. Next, L-DOPA is decarboxylated by amino acid decarboxylase to the transmitter, dopamine.



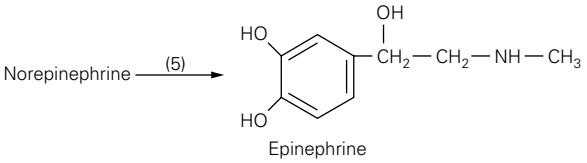
There are five types of dopamine receptors, all of which are metabotropic. Dopaminergic neurons are abundant in the substantia nigra and the ventral tegmental area; dopamine cells project to the basal ganglia, hypothalamus, and the limbic system to regulate voluntary movement, thinking, and planning. Dopamine plays an important role in the reward system and in addiction.

b. Norepinephrine NE is produced by the introduction of a hydroxyl group into the carbon chain of dopamine by dopamine β-hydroxylase.



NE is present in neurons of the locus ceruleus, which have widespread projections to the cerebral cortex, cerebellum, and spinal cord. NE is the transmitter used by the postganglionic neurons of the sympathetic nervous system.

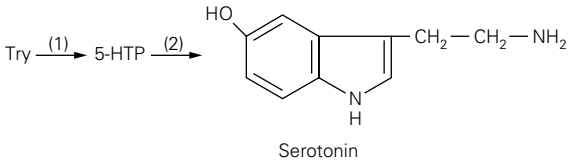
c. Epinephrine In the adrenal medulla, NE is converted to epinephrine by methylation. This reaction is catalyzed by phentolamine-*N*-methyltransferase and requires *S*-adenosylmethionine as a methyl donor.



Only a few neurons in the central nervous system use epinephrine. These are located together with noradrenergic cells in the medulla oblongata, solitary tract nucleus, and medial longitudinal fascicle.

As we have seen, the synthesis of epinephrine is carried out in four enzymatic steps, the first two being common to dopamine, NE, and epinephrine, the first three being common to NE, and epinephrine, and the last needed only for the synthesis of epinephrine. Why are NE and epinephrine not made in dopaminergic neurons? The particular neurotransmitter produced is determined by the absence of the next enzyme. Thus, no products past dopamine are produced in dopaminergic neurons because dopamine β-hydroxylase is absent. Similarly, epinephrine is not produced in neurons that lack the methyltransferase.

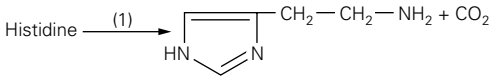
d. Serotonin (5-HT), an Indoleamine The essential amino acid L-tryptophan is oxidized to 5-hydroxytryptophan (5-HTP) by tryptophan hydroxylase, an enzyme homologous to tyrosine hydroxylase. The amino acid decarboxylase that operates in the catecholamine pathway also decarboxylates 5-HTP to serotonin.



Serotonin

About 13 5-HT receptors have been characterized. Only one is ionotropic. Transmission with 5-HT is critical for sleep, mood, sexual behavior, and aggressive behavior.

e. Histamine Histamine is formed by decarboxylation of the amino acid L-histamine catalyzed by histidine decarboxylase.



Histamine

There are only about 100,000 neurons in the brain that are histaminergic; they are situated in the posterior hypothalamus and have widespread projections. Three types of histamine receptors have been identified, all metabotropic. Central histaminergic neurons control appetite and regulate the secretion of pituitary hormones. In the periphery, histamine is familiar as a hormone that is important in the inflammatory reaction and in the activity of exocrine glands, for example, in the secretion of acidic gastric juice.

B. Amino Acids

1. L-Glutamate

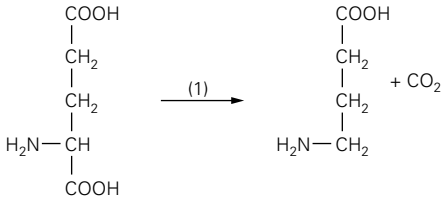
This common amino acid is synthesized from α-ketoglutarate by transamination or by the hydrolysis of glutamine. At least half of all of the synapses in the brain are glutaminergic, and most of these are excitatory. Five types of excitatory glutamate receptors have been identified. These are all ionotropic and have been classified by their sensitivity to various glutamate agonists. An important example is the NMDA (*N*-methyl-D-aspartate) receptor. This receptor permits the passage of Ca^{2+} as well as Na^{+} and K^{+} ions only when the neuron is significantly depolarized. The NMDA receptor is thought to play a key role in learning and memory. Other receptors for glutamate are metabotrophic and mediate modulatory synaptic actions.

Under certain pathological conditions such as stroke or injury, large amounts of glutamate are released into the brain, which activate these glutamate receptors and cause neuronal death. This process is called excitotoxicity and is a significant way in which brain tissue is damaged during a stroke. Many neurons that are excited by glutamate are also excited by L-aspartate when it is applied. L-Aspartate fulfills all of the criteria of a neurotransmitter except one: it is not released from presynaptic neurons.

Released glutamate is removed from the synaptic cleft by a specific uptake mechanism in the membrane of glutaminergic neurons and surrounding glial cells. These transporters pump the transmitter molecules in against a concentration gradient and, therefore, require the energy of ATP. In glial cells, the glutamate is converted to glutamine by the enzyme glutamine synthase. The glutamine is then transported into neurons to be hydrolyzed back to glutamate.

2. GABA

γ-Amino butyric acid (GABA) is the chief inhibitory neurotransmitter in the brain. It is formed from glutamate by decarboxylation with glutamic acid decarboxylase (GAD) using pyridoxal phosphate as a cofactor.



Glutamate GABA

Antibodies against GAD are used to identify GABAergic neurons.

3. Glycine

Glycine is synthesized from serine. It is the major inhibitory neurotransmitter in the spinal cord and retina. Receptors for GABA and glycine allow the influx of Cl^{-} ions, which hyperpolarizes the postsynaptic cells.

C. Purines

ATP is present in high concentrations in transmitter vesicles, both for neuropeptides and for small-molecule transmitters. ATP is synthesized in mitochondria

and is abundant in cytoplasm: it gets pumped into vesicles. Because it is present in vesicles, it is routinely coreleased with neurotransmitters. In some postsynaptic cells, there are receptors that respond to ATP, adenosine, or AMP where they modulate the transmission of other neurotransmitters. Thus, they can inhibit by activating autoreceptors. Receptors are present in the vas deferens, gut, smooth muscle, and neurons of the dorsal horn of the spinal cord. There are two classes of purine receptors, P1 and P2. P1 receptors are sensitive to adenosine and AMP and are all metabotropic. P2 receptors respond to ATP and AMP; a few of these are ionotropic.

Purinergic transmission illustrates an important aspect of chemical synaptic transmission. The abundant release of a substance is necessary, but not sufficient evidence that the substance is a transmitter, even though that substance has been shown to act as a transmitter elsewhere in the body. A postsynaptic receptor must also be present. A substance is not a neurotransmitter if it is released into an area that lacks a receptor for it. It may seem improbable that this occurs frequently, but ATP is actually released at all synapses.

VI. THE SYNAPTIC MESSAGE IS ENCODED IN THE POSTSYNAPTIC RECEPTOR

It is important to appreciate the nature of the information being sent during synaptic transmission. Release of a neurotransmitter is triggered when a neuron is depolarized with the consequent influx of Ca^{2+} ion. The particular transmitter substance released does not determine the type of action in the target cell. That is determined entirely by the postsynaptic receptor. Thus, whereas the release of transmitter is permissive, the receptor is determining. All transmitters have more than one type of receptor. As we have seen, ACh activates a variety of ionotropic receptors, all of which are excitatory. Its metabotropic receptors can be inhibitory, like those that Loewi studied in the heart, or excitatory, as those in the gut, bladder, bronchi, and iris. Thus, the actions of ACh can be excitatory or inhibitory. Another example is glutamate. In the hippocampus, glutamate activates two types of ionotropic receptors, the NMDA and the non-NMDA [AMPA (α-amino-3-hydroxy-5-methylisoxazoleproprionic acid), quisqualate, and kainate] receptors, and a metabotropic receptor. At normal resting potentials, the dendritic spines of a hippocampal neuron are briefly depolarized by a pulse of glutamate, because non-NMDA receptors respond but NMDA receptors do not. When the dendrites are depolarized for a longer time, the NMDA receptors are activated, causing prolonged Ca^{2+} entry and inducing a series of molecular changes within the dendrite. These changes enhance the responsiveness of the synapse, which can persist for many hours. This phenomenon is called long-term potentiation (LTP).

VII. SYNAPTIC VESICLES

Our thinking about neurotransmission was strongly influenced by two important discoveries made at midcentury. By using electrophysiological techniques to study nicotinic cholinergic transmission (primarily at the neuromuscular junction), Bernard Katz and his co-worker discovered that transmission is quantized. The postsynaptic responses (excitatory postsynaptic potentials, EPSPs) occur in discrete units of strength, presumably because the presynaptic component of the synapse releases ACh in specific units of about 3000 molecules. Thus, depending on the intensity of stimulation, it was estimated that 3000, 6000, 9000, etc. molecules of the neurotransmitter are released. At the same time, Eduardo De Robertis, using the electron microscope, examined the fine structure of synapses. He observed that nerve endings typically have a cluster of small (0.50 nm in diameter) vesicles crowded over a thickened membrane specialization. This observation prompted the idea that guided subsequent research on synaptic transmission. Katz proposed that his quantum is the amount of ACh contained in a single small vesicle when it is emptied into the synaptic cleft. Depending on the intensity with which the neuron is stimulated, one or more vesicles would participate in transmission.

Unlike the larger secretory granules that carry neuropeptides, the small vesicles appear clear or electron-lucent in electron micrographs. Examination of presynaptic elements that release small-molecule neurotransmitters other than ACh reveals a similar morphology. Synaptic vesicles can be isolated intact using differential ultracentrifugation. One convenient preparation, developed independently by both Victor Whittaker and De Robertis, is called synaptosomes. Nervous tissue is gently ground in a glass–glass tissue grinder. During this procedure, a large proportion of nerve endings are twisted off their axons to become closed spheres. These spheres contain the presynaptic components of synapses (with frequent contamination of some of the postsynaptic elements). After these

synaptosomes are isolated, they can be broken by osmotic shock to release the small vesicles. By using this technique, the content of ACh in one synaptic vesicle was by biochemical assay found to match the amount of neurotransmitter present in a quantum estimated electrophysiologically.

Unlike the neuropeptides, small-molecule transmitters are not synthesized and processed within membranes of the endoplasmic reticulum. Rather, they are made in the cytoplasm of nerve terminals (as well as in the cell body, axon, and dendrites). Their concentration in vesicles (0.1–0.5 *M*) is much greater than that in cytoplasm. Therefore, these transmitters must be taken up against considerable concentration gradients. For this purpose, there are at least four different types of active transport systems in vesicle membranes that pump in potential transmitters. These systems are ATPase pumps that use the energy of ATP to create proton gradients (the contents of the vesicles are acidic) to trap the positively charged amine molecules within the vesicle.

The membranes of some large, dense-cored vesicles also contain pumps for small-molecule transmitters. Therefore, these vesicles can contain both neuropeptides and small-molecule neurotransmitters. In addition some neurons contain both large vesicles and small vesicles. As a result, both types of transmitter can be released from the same neuron. The discovery of this phenomenon (called coexistence or cotransmission) caused considerable astonishment when it was first discovered in the 1970s. Neurobiologists were surprised because the discovery of cotransmission appeared to disobey a law called Dale's principle. In 1935, Dale suggested that all parts of a neuron would make the same neurotransmitter (because of the "unity of biochemistry"). Later this idea became much more stringent. Dale's principle became "one neuron, one neurotransmitter substance." Because we now understand how peptides and small-molecule transmitters are packaged into vesicles, we should not be troubled that a small-molecule transmitter can be packaged in both small synaptic vesicles and in large, dense-cored vesicles but that neuropeptides exist only in the large vesicles.

The two types of neurotransmitters can be released differentially. As discussed earlier, large dense-cored vesicles undergo exocytosis when the concentration of Ca^{2+} ion is elevated throughout the terminal. This condition occurs during sustained stimulation. If the large vesicles contain a small-molecule transmitter as well as a neuropeptide, both transmitters will be coreleased. As might be expected from evolutionary considerations, two coreleased transmitters usually have synergistic actions on target cells. As an example, motor neurons release calcitonin gene-related peptide (CGRP) together with ACh. ACh causes the ionotropic AChRs to open and produce muscle contraction, and CGRP stimulates adenylyl cyclase to initiate cAMP-dependent protein phosphorylation within the muscle fiber that both enhances the force of contraction and activates energy metabolism.

Small synaptic vesicles, which are derived from early endosomes, exclusively package small-molecule transmitter substances. Their membranes contain special proteins (synaptotagmin, synaptophysin, synaptobrevin) that are required for the rapid membrane fusion that these vesicles undergo during exocytosis. These constituents belong to families of proteins that mediate the interaction of membranes in all cells, such as the budding off of transport vesicles from the endoplasmic reticulum, their fusion with the Golgi apparatus, and vesicles shuttling between the Golgi stacks and between the Golgi and the plasma membrane. Characteristic of the small synaptic vesicles is their ability to fuse with the synaptic plate membrane specialization to open and release their contents into the synaptic cleft. Exocytosis takes place at specialized docking sites, which are arrayed regularly along the synaptic plate. As first described by Rodolfo Llínas, just adjacent to each docking site is a specialized Ca^{2+} ion channel. Depolarization produced by brief stimulation will result in the rapid influx of Ca^{2+} ions near the docking site, rapidly raising the local concentration high enough to cause membrane fusion.

In summary, the cell biology of the two vesicle types to a large extent determines the physiology of the synaptic transmissions that they mediate. Small synaptic vesicles discharge rapidly after the neuron has been briefly depolarized. Only small-molecule transmitters are released. Because the synthesis of these transmitters can take place in nerve terminals, synaptic vesicles can be used many times and new transmitter can be pumped into the recycled vesicle. This results in rapid and repeated transmission. In contrast, the large vesicles require relatively prolonged stimulation to be released. There are no specialized proteins in their membrane, so exocytosis occurs anywhere in the nerve terminal. Release is sustained, because the built-up Ca^{2+} ion concentration remains elevated for relatively long periods of time throughout the terminal. Large vesicles can release both neuropeptides and small-molecule transmitters. Once used, however, these vesicles cannot be used again because there is no mechanism for refilling with neuropeptides at nerve terminals. Thus, transmission is slow and long-lasting.

After the work of Katz and De Robertis it became almost axiomatic that a substance must be packaged in vesicles and released at nerve terminals in order to be a true neurotransmitter. Whereas it is prudent to keep this axiom in mind, more recent discoveries indicate that this traditional view must be expanded. Exocytosis is not the only mechanism by which small-molecule neurotransmitters can be released. Thus, dopamine (in the substantia nigra) and GABA (in the retina) can be released from nerve cell bodies. Release occurs by reversal of transporter pumps. If this nonexocytotic release affects a nearby target cell, the process must be called neurotransmission.

VIII. MEMBRANE-SOLUBLE MESSENGERS

Another type of nonvesicular neurotransmission is mediated by small molecules that are membrane-soluble. The best studied are the gas nitric oxide (NO) and the fatty acid arachidonate and its metabolites. These molecules diffuse through membranes and do not need vesicles to be released. Both NO and arachidonic acid have long been known to play important physiological roles in tissues other than the nervous system: NO primarily as a regulator of contractility of blood vessels and arachidonate in inflammation. NO is formed by the oxidation of arginine catalyzed by the enzyme nitric oxide synthase and an electron donor (FAD, NADPH, or THB) as cofactor. The aminoacid citrulline is produced.

Arachidonic acid is released from membrane phospholipids by the receptor-mediated activation of phopholipase A_2 through a G-protein. The fatty acid released is then metabolized through two general pathways: (1) lipoxygenase, of which there are several (5-, 8-, 12-, 15-), and (2) cyclo-oxygenase. Metabolites from both of these pathways have been shown to be active in other tissues. Thus far, only 5-, 8-, and 12-lipoxygenase products have been implicated in synaptic transmission.

It has been proposed that both NO and arachidonic acid act as retrograde messengers. Thus, in LTP, glutamate released from the presynaptic neuron would, in addition to depolarizing the postsynaptic cell, also activate the synthesis of membrane-soluble messengers, which then would diffuse back to the presynaptic element to enhance further release of glutamates.

IX. REMOVAL OF TRANSMITTERS

There are three ways by which transmitters are removed from the synaptic cleft: (1) enzymatic degradation, (2) specific uptake by neurons and glia, and (3) diffusion.

A. Enzymatic

Removal of ACh is primarily enzymatic. Acetylcholinesterase is anchored extracellularly to the postsynaptic membrane near the AChRs and hydrolyzes the transmitter to acetate and choline. The enzyme is so effective that on average an ACh molecule binds only once to a receptor. Thus, hydrolysis of ACh serves to punctuate the message; it also is a conservation mechanism (see later discussion).

Both small-molecule and peptide transmitters are degraded by enzymes, but these do not play an important role in synaptic transmission. A clinically important example is monoamine oxidase (MAO), which is located in the outer membrane of mitochondria. MAO catalyzes the oxidative deamination of catecholamines to the corresponding aldehyde. Another nonspecific degradative enzyme is catechol-*O*-methyltransferase (COMT), which is present in the blood. The transferase inactivates catechols by methylating one of the ring hydroxyl groups. Unlike acetylcholinesterase, these enzymes are not located strategically at synapses, and they operate over a long time relative to synaptic transmission. Nevertheless, physiologically they are important because they ultimately contribute to the regulation of the amount of transmitters in the brain. Regulation of transmitter concentration is thought to be important in some psychiatric conditions, for example, aggression and depression.

B. Specific Uptake

As described earlier, choline is not synthesized in nervous tissue. The choline produced by the hydrolysis of ACh by acetylcholinesterase is taken back up into the presynaptic neuron by a transport mechanism specific to cholinergic neurons and is used for the resynthesis of transmitter. As also described previously for glutamate, the small-molecule neurotransmitters are taken up by high-affinity transporter mechanisms. Uptake by means of specific transport is the most

important way classical transmitter molecules are removed from the synaptic cleft. Neuropeptides are not taken back up into the neuron and are removed by diffusion and by cleavage by extracellular peptidases.

C. Diffusion

Some transmitter is removed by diffusion. With small molecules this process is relatively fast; with peptides, on the other hand, the slow speed of diffusion contributes to the long-lasting quality of peptidergic transmission.

See Also the Following Articles

CHEMICAL NEUROANATOMY • DOPAMINE • ENDORPHINS AND THEIR RECEPTORS • GABA • NEURAL NETWORKS • NEURON • NEUROPEPTIDES AND ISLET FUNCTION • NOREPINEPHRINE • PEPTIDES, HORMONES, AND THE BRAIN AND SPINAL CORD • PSYCHONEUROENDOCRINOLOGY • SYNAPSES AND SYNAPTIC TRANSMISSION AND INTEGRATION

Suggested Reading

Cooper, J. R., Bloom, F. E., Roth, R. H. (2002). *The Biochemical Basis of Neuropharmacology*, 8th ed. Oxford University Press, New York.

Iversen, L. L. (1995). Neuropeptides: Promise unfulfilled. *Trends Neurosci.* **61,** 839–849.

Kandel, E. R., Schwartz, J. H., and Jessell, T. M. (2000). *Principles of Neural Science*, 4th ed. McGraw-Hill, New York.

Katz, B. (1969). *The Release of Neural Transmitter Substances*. Liverpool University Press, Liverpool, UK.

Kupfermann, I. (1991). Functional studies of cotransmission. *Physiol. Rev.* **71,** 683–732.

Llínas, R. R. (1982). Calcium in synaptic transmission. *Sci. Am.* **247,** 56–65.

Loewi, O. (1960). An autobiographical sketch. *Perspectives Biol. Med.* **4,** 3–25.

Siegel, G. T., Agranoff, B. W., Albers, R. W., Fisher, S. K., and Uhler, M. D. (1999). *Basic Neurochemistry: Molecular, Cellular, and Medical Aspects*, 6th ed. Lippincott-Raven, New York.

Nociceptors

LORNE M. MENDELL
State University of New York, Stony Brook

GLOSSARY

Aδ-fiber A sensory neuron whose axon in the peripheral nerve is lightly myelinated. These axons have the lowest conduction velocity among myelinated axons.

allodynia The pain elicited by a normally nonpainful stimulus after inflammation or damage to peripheral or central neurons.

C-fiber A sensory neuron whose axon in the peripheral nerve is unmyelinated. These axons are the slowest conducting in the peripheral nerves.

dorsal horn The region of spinal cord gray matter in which nociceptive neurons terminate. This region is generally divided into laminae (laminae I–V). Lamina I is also known as the marginal zone, lamina II as the substantia gelatinosa, and laminae III–V, as nucleus proprius.

dorsal root ganglion The location of cell bodies of sensory neurons innervating the body. A corresponding ganglion for the face is the trigeminal or gasserian ganglion.

dorsal root ganglion cell A sensory neuron consisting of a cell body in the dorsal root ganglion, a peripheral process innervating the peripheral tissues (skin, muscle, etc.), and a central process projecting into the spinal cord to synapse on postsynaptic neurons.

hyperalgesia The increase in perceived intensity of a normally painful stimulus after injury.

nociceptor A class of sensory neuron responding selectively to tissue damage in its receptive field.

sensitization A process that decreases the stimulus intensity required for a nociceptive response or that increases the magnitude of the nociceptive response elicited by a given nociceptive stimulus.

Nociceptors belong to a specialized class of sensory fibers that transmit information concerning stimuli that are potentially damaging to the organism; they generally elicit a response of pain. This article discusses the properties of this important class of neuron, specifically its specialized physiology, anatomy, and biochemistry. In addition to acutely signaling the presence of a potentially damaging stimulus, nociceptors can exhibit significant plasticity in response to injury in their receptive field, and this will be discussed in the context of persistent or chronic pain.

I. INTRODUCTION

Nociceptors are sensory fibers that respond to stimuli that are potentially damaging to the organism. In practice this can mean a variety of stimuli, ranging from intense pressure, extremes of temperature, to inflammation. Impulse activity in nociceptors activates central circuits that can lead to what is subjectively referred to as *pain*. It is necessary to specify "can lead" because neural systems originating in the midbrain and medulla and projecting to the spinal cord can suppress this input at the level of the spinal cord or medulla where impulses from nociceptors first

activate neurons in the central nervous system. Other more complex mechanisms at the thalamic and cortical levels also can affect the perception of the input from nociceptors. Thus, a stimulus that elicits pain under some circumstances does not always do so. Conversely, as will be discussed later, stimuli that usually do not elicit a report of pain can do so under certain conditions.

Effective nociceptive inputs instigate removal of the affected area from the source of the potentially damaging stimulus and help to protect the injured area from further damage, i.e., to promote healing of the tissue. Removal of the affected area can occur via a combination of segmental reflex mechanisms (e.g., via a flexor reflex) as well as more conscious behavior. However, in many cases the precipitating event can also result in persistent changes in the nociceptors that outlast the initiating event. Remarkably, changes can also occur in sensory neurons normally responding only to nonnociceptive stimuli, such that they can elicit nociceptive reactions. In some cases these changes are adaptive, e.g., keeping the skin cool after sunburn or immobilizing a limb after muscle damage. However, in other cases the changes can last well beyond the time when it is useful, and this chronic pain can be extremely maladaptive, e.g., in postherpetic neuralgia.

It could be anticipated that afferent fibers transmitting nociceptive information from the periphery would be among the smallest because of early findings that the stimulation of peripheral nerves in humans elicited reports of pain only when the electrical stimuli were sufficiently intense to activate small myelinated (Aδ) or unmyelinated (C) fibers. However, early work using either direct or indirect methods to monitor the electrophysiological responses of small-diameter fibers yielded only scattered reports of small fibers activated selectively by intense stimuli. This raised the possibility that nociceptive stimuli were coded primarily by increased frequency of firing in primary afferent fibers also responding to innocuous stimuli, in effect the possibility that discharge *pattern* in afferent fibers is the prime neural representation of noxious vs nonnoxious stimuli. The development of techniques to record reliably from small-diameter fibers in peripheral nerves by Perl and his associates led to a clear demonstration of several classes of small myelinated and unmyelinated fibers innervating receptors that respond exclusively to stimuli that are potentially damaging to the organism. Nociceptors have been found in skin, muscle, joints, and viscera (see Section III). It is activation of these fibers rather than specific impulse discharge patterns in fibers also responding to innocuous stimuli that signals the occurrence of a stimulus that is nociceptive. This is true under normal conditions; however, after certain injuries or under certain pathological conditions, fibers normally responding to gentle stimulation can elicit pain (allodynia; see Section VIII).

Although nociceptors are restricted largely to sensory neurons with small axons, not all sensory neurons with small axons are nociceptors. A group of sensory neurons with Aδ-axons innervates sense organs in the skin known as Down hairs (D-hairs) that are exquisitely sensitive to innocuous mechanical stimulation. Unmyelinated C-fibers can also innervate receptors responding to innocuous mechanical or thermal stimulation. Conversely, cells with Aβ-axons innervating high-threshold mechanoreceptors have been reported. These Aβ-nociceptors tend to be among the more slowly conducting Aβ-fibers.

II. UNIQUE ANATOMICAL AND BIOCHEMICAL PROPERTIES OF NOCICEPTORS

Although the cell bodies of nociceptors and low-threshold mechanoreceptors are located in the sensory ganglia [the dorsal root ganglion (DRG)] or similar structures for cranial nerves (e.g., the gasserian ganglion of the trigeminal nerve), they exhibit numerous differences in their projections and in their chemistry that determine their properties as nociceptors (Fig. 1). All of these contribute to the basic characteristics of nociceptors and nociceptive pathways, including their high threshold and their ability to be sensitized after certain types of stimuli.

From the early functional studies indicating that axons of nociceptors are among the smallest in peripheral nerves, it was believed that the cell bodies of nociceptive afferents would be among the smallest, and this has been confirmed with correlative electrophysiological and anatomical studies of individual sensory neurons. However, there is considerable scatter in soma size for axons of a particular group (e.g., Aδ-axons), making soma size somewhat less reliable as an indicator of nociceptive vs nonnociceptive neurons than axon size. Nonetheless, many workers have considered a soma diameter of 30 μm to be a dividing line below which the population is highly enriched in nociceptors.

Various markers have been found to correlate with cell size, and thus putatively to correspond to cells with nociceptive and nonnociceptive function. Many investigators have divided the dorsal root ganglion cell

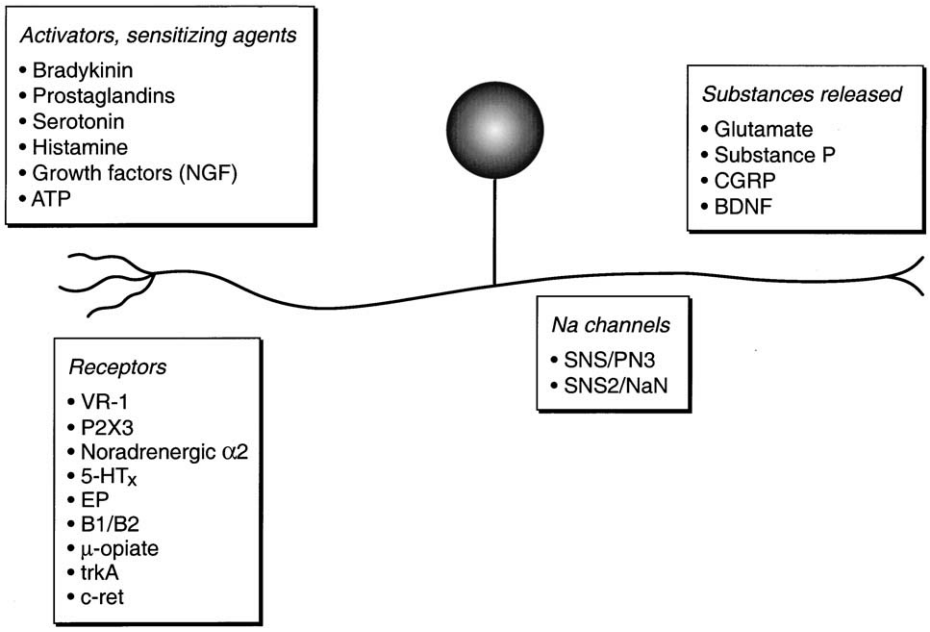


Figure 1 Chemical specialization of nociceptive sensory neurons that distinguishes them from nonnociceptive afferents. These specializations are divided into activators and sensitizing agents, receptors, Na channels, and released substances. In the latter group, glutamate is not unique to nociceptive afferents, but as discussed in the text it activates specialized glutamate receptors that result in the unique central action of nociceptors.

population into small dark and large light cells on the basis of their appearance in Nissl-stained sections. The large light cells stain with RT-97, an antibody against the phosphorylated form of the 200-kDa neurofilament subunit, whereas the small dark ones do not. Cells staining with RT-97 are believed to have myelinated axons ranging over the entire size spectrum (Aβ and Aδ).

A number of biochemical labels are present only in small somata and, thus, presumptively those that are nociceptors. These markers include the lectin IB4, the enzyme fluoride-resistant acid phosphatase (FRAP), trkA (the high-affinity receptor for NGF), c-ret (the receptor for glial-derived neurotrophic factor, GDNF), and various peptides such as substance P (SP), calcitonin gene-related peptide (CGRP), and somatostatin. However, none of these markers is present in the entire population of small cells, and some such as CGRP are also present to some degree in larger cells. Thus, the population of small DRG neurons is quite varied in its expression of these and other markers. Loss of specific functional capacity when a specific gene is eliminated, i.e., in "knockout" mice, suggests that the expression of some of these markers corresponds to specific psychophysical modalities, e.g., substance P expression relates to the ability to sense very intense thermal stimuli.

Although small DRG cells are highly varied in their expression of FRAP, substance P, etc., some of these markers tend to be frequently coexpressed, whereas others are coexpressed very infrequently. For example, c-ret, IB-4, and FRAP tend to be coexpressed and these cells do not express peptides. Another group of cells expresses trkA as well as peptides such as CGRP and/or substance P. The population of cells coexpressing markers in these different groups is relatively small. The trkA/peptide-expressing cells include some with small myelinated and some with unmyelinated axons, whereas the IB4/c-ret/FRAP-expressing neurons have largely, if not exclusively, unmyelinated axons.

Nociceptive neurons are also unique in their projection into the central nervous system. Numerous evaluations of their termination sites, including the location of peptide-expressing terminals as well as the transport of various materials such as horseradish peroxidase (HRP) in physiologically identified nociceptive afferents impaled at the dorsal root entry zone, have demonstrated that they terminate largely though not exclusively in the most superficial laminae of the dorsal horn. Small myelinated nociceptive afferents terminate in lamina I (also known as the marginal zone) and lamina V (the deepest part of the nucleus proprius), whereas unmyelinated afferents terminate in lamina II.

Many workers have subdivided lamina II into inner and outer portions (lamina II_o and lamina II_i). The IB4/c-ret/FRAP-expressing neurons terminate largely in lamina II_i, whereas the peptide-containing neurons terminate in laminae II_o and I. Physiological evidence suggests that cells in lamina II_i respond almost exclusively to nonnoxious stimulation. This would indicate that the afferent population projecting into this zone responds to nonnociceptive stimuli. However, many IB4-expressing neurons express the receptor for capsaicin (VR1), which would suggest that they are responsive to noxious heat and chemical stimulation, specifically protons. This apparent discrepancy remains to be resolved, but it is important to recall that, in this as in other studies, the recordings from central neurons were made in reduced preparations where their properties may be distorted.

III. PHYSIOLOGICAL PROPERTIES OF NOCICEPTORS IN DIFFERENT TISSUES

A. Skin

Nociceptors innervating the skin can be differentiated into two major groups on the basis of the conduction velocity of the axon transmitting impulse discharge from the skin to the spinal cord. These are the small myelinated Aδ-fibers and unmyelinated C-fibers. Two groups of Aδ-nociceptive afferents have been described. Both discharge in response to high-intensity, potentially noxious mechanical stimuli but can be distinguished on the basis of their response to noxious heat. One group with a high thermal threshold ($>53^{\circ}$C) exhibits a long response latency to heat stimuli and tends to sensitize in response to successive thermal stimuli. These are known as type I AMH (A mechanoheat) receptors, but because of their high thermal threshold this response is often ignored; thus, they are denoted as high-threshold mechanoreceptors (HTMs). The other group, known as type II AMH fibers, has a lower heat threshold (46°C), tends to respond quickly, and desensitizes in response to repeated stimuli. Some of these fibers exhibit sensitivity to chemical stimuli. Nociceptors innervated by C-fibers also exhibit responses to noxious mechanical and thermal stimuli (threshold averaging about 46°C) and often to chemical stimuli such as lowered pH and bradykinin. These sensory neurons are often referred to as polymodal nociceptors.

Many investigators have identified a population of small-diameter afferent fibers whose terminations in the skin can be identified by electrical stimulation but that exhibit no response to peripheral stimuli of any known modality. These "silent" nociceptors (presumably nociceptors whose sensitivity can be brought to a detectable level by sensitizing stimuli) represent an extreme of the spectrum of nociceptive thresholds in normal skin.

B. Muscle

Muscle afferent fibers innervating endings activated by potentially damaging (nociceptive) stimuli also transmit their impulses to the spinal cord via small myelinated and unmyelinated afferents. These fibers are generally identified by the Roman numeral system in keeping with D.P.C. Lloyd's original classification of muscle afferents (I–IV). Thus, the small myelinated afferents are known as group III (equivalent in fiber diameter to Aδ-fibers) and unmyelinated afferents are known as group IV (equivalent to C-fibers). These small fibers innervate receptors in muscle whose adequate stimulus is best described as squeezing the muscle (normally quite painful). Those with group III axons tend to be activated at lower stimulus intensity than those with group IV axons. Many of these receptors can also be excited by chemical stimuli such as bradykinin or 5-HT. These substances are released when the muscle is under stress due to ischemia, lactic acid buildup (lowered pH), etc. Some of these receptors also exhibit responses to noxious temperature ($>43^{\circ}$C). Thus, nociceptors in muscle resemble those in skin in responding to one or more modalities of stimulation.

C. Visceral Afferents

Many visceral organs (heart, veins, lungs, colon, ureter, bladder, etc.) are innervated by small-diameter afferent fibers that selectively respond to intense mechanical stimuli and also to irritant chemical stimulation such as bradykinin, capsaicin, and serotonin.

D. Joint Receptors

Nociceptors are innervated by group III and IV afferents (as described in the section on muscle afferents) and are defined as being activated only when the joint moves outside its normal physiological range. Most of these afferents can also be activated by

various chemical agents such as bradykinin, prostaglandins, and serotonin.

E. Cornea

The cornea occupies an important place in studies of nociception because the early histology indicated exclusively free nerve endings, whereas contemporaneous psychophysical studies indicated that gentle stimulation elicited nonnociceptive responses. This challenged the dogma originating at the end of the nineteenth century that free nerve endings are associated exclusively with nociceptors and pain. Studies of the physiological properties of fibers innervating the cornea indicate the presence of polymodal nociceptive afferents with Aδ- or C-fibers responding to mechanical, thermal, and chemical stimuli. The mechanical threshold of polymodal nociceptors is slightly lower than that of mechanosensory units. Thermal threshold is about 39°C. Chemical sensitivity can be demonstrated for protons, bradykinin, prostaglandin, and serotonin. A third type of unit known as a mechanoheat unit that exhibits a higher mechanical threshold than the others has also been described. Such units can develop chemical sensitivity after repeated thermal stimulation. Recent psychophysical studies of corneal sensation suggest that gentle stimulation elicits unpleasant if not frankly painful sensations, in line with the adequate stimulus of the afferent fibers innervating this tissue.

F. General Comments

Although nociceptive afferents are not a uniform or homogeneous population, certain general conclusions emerge. First, they transmit via small-diameter, slowly conducting axons. Second, they tend to transmit information about more than a single modality. Third, the spectrum of stimuli that activate these nociceptive afferents is remarkably similar from tissue to tissue. Finally, as was only touched upon in the foregoing brief description, the properties of these receptors are not static; they can be influenced by the state of their termination site and by preceding activity.

IV. THE CONTRIBUTION OF NOCICEPTORS TO THE SUBJECTIVE PERCEPTION OF PAIN

Evidence in favor of the role of physiologically described nociceptors in mediating pain sensation has been obtained for thermal nociceptive responses using several different approaches. An important contribution has come from comparison of the physiological responses of nociceptors in primates with the results of psychophysical experiments in human subjects. There is general correspondence between changes in the intensity of the subjective response and changes in the firing rate of the nociceptive afferents. In the case of a long-lasting thermal stimulus, the response of nociceptors indicates that C-fiber nociceptors mediate the initial component of the subjective response, whereas Aδ-fiber nociceptors (type I AMH) mediate the maintained response. The finding of severely reduced nociceptive responses in humans congenitally lacking unmyelinated fibers supports a crucial role for unmyelinated fibers in thermal nociception. Furthermore, pain is reported by subjects with conduction block of myelinated fibers in the nerve supplying the tested area.

One of the important implications of the nociceptor hypothesis, i.e., that pain is the response to activation of a special group of peripheral afferent fibers responding exclusively to damage in the periphery, is that selective activation of such fibers in peripheral nerves should elicit pain referred to the receptive field of the fiber. In human subjects, stimulation through a microelectrode recording activity from a single or a small number of C-fiber nociceptors results in a report of pain from a projected field (i.e., the region of skin to which the pain produced by the electrical stimulus is referred) close to the receptive field of the recorded fiber(s). The differences are within the uncertainty in the ability to localize nociceptive stimuli. It is difficult to be certain that only a single C-fiber is being activated under these conditions. However, the agreement of stimulus modality and receptive field location of the recorded fiber(s) with the projected field from the stimulation suggests that the electrical nerve stimulus activates the same fiber(s) is/are driven by the nociceptive stimulus used to test the adequate stimulus of the recorded fiber(s).

It should not be concluded from the findings under the experimental conditions described earlier that impulses in nociceptive fibers always elicit the same central effects. In the 1960s, the *gate theory* of Melzack and Wall described a physiological mechanism by which the spinal actions elicited by impulses in small-diameter afferent fibers depended on the balance of ongoing activity in small and large fibers. Specifically, segmental inputs in large-diameter afferents were proposed to attenuate the synaptic effects evoked by small-diameter afferents. More recently, descending

systems have been identified that also elicit such inhibitory effects. These fibers activate spinal interneurons that release the opiate agonist enkephalin. This transmitter activates opiate receptors on terminals of sensory neurons and this reduces Ca entry into the terminals and, thus, transmitter release. Because opiate receptors are expressed primarily on terminals of small-diameter nociceptive afferents, this is a specific action on nociceptive afferent fibers resulting in presynaptic inhibition of transmitter release.

V. MEMBRANE PROPERTIES

Nociceptors exhibit specialized electrical properties that set them apart from low-threshold mechanoreceptors, whose cell bodies also are located in sensory ganglia. At a descriptive level, the action potentials in nociceptors recorded intrasomatically have consistently different shapes, suggestive of the contribution of different ion channels (Fig. 2). Specifically, the spike in nociceptors tends to be larger and broader than spikes in low-threshold mechanoreceptors. A shoulder is often observed on the descending limb of the spike. Spikes in these cells also exhibit a longer lasting after-hyperpolarization. It has been suggested that the differences in spike properties between nociceptors and nonnociceptors are caused by a reduced delayed rectifier, leading to a larger upstroke and a larger Ca channel conductance. This would account for both the shoulder and the longer postspike after-hyperpolarization (AHP), the latter mediated by an enhanced Ca-mediated K conductance. These differences in spike configuration have been useful in studies carried out in dissociated cell culture where other criteria to classify nociceptors (e.g., response threshold) are unavailable.

Nociceptors also have a unique complement of Na channels that distinguishes them from nonnociceptors. Na channels are highly differentiated and can be classified in part on the basis of their sensitivity to the neurotoxin tetrodotoxin (TTX). Some Na channels can be blocked by TTX concentrations in the nanomolar range, whereas others are insensitive to TTX concentrations as high as 10 μM. Several sensory-neuron-specific Na channels have been identified, including the TTX-sensitive channel called PN1 and the TTX-insensitive channels called SNS/PN3 and SNS2/NaN. These Na channels have been cloned, and the expression of SNS/PN3 and SNS/NaN has been found to be restricted to small somata, i.e., putative nociceptors. In agreement with these studies, the somatic spike of nociceptors recorded *in vivo* is resistant to TTX, in contrast to the somatic impulses in low-threshold afferents that are blocked by TTX. However, axonal spikes in both classes of afferents can be blocked by TTX. Thus, the membrane of individual DRG cells is not necessarily homogeneous with respect to Na channel expression. Significantly, there is also evidence for TTX-resistant Na channels in the peripheral terminals of nociceptive neurons. Thus, any physiological role for TTX-resistant channels obtained from studies of the cell body can be applied to generation of the impulse in the sensory terminals (see Section VIII.A on peripheral sensitization).

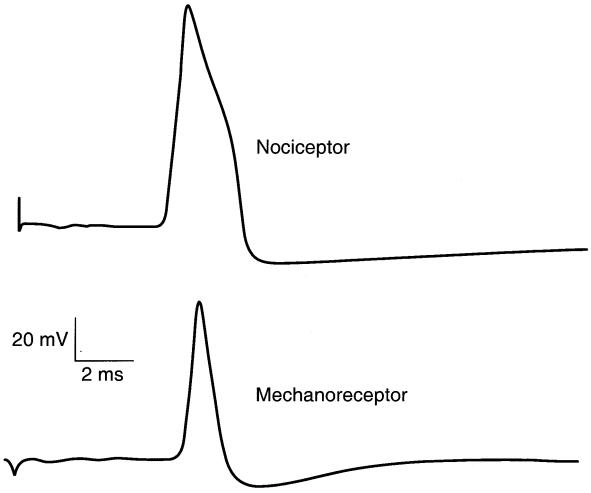


Figure 2 Examples of action potentials recorded intracellularly from the soma of a nociceptor and a low-threshold mechanoreceptor. Note the difference in spike configuration. Adapted from Traub, R. J. and Mendell, L. M. (1988). *J. Neurophysiol.* **59,** 41–55, with permission. Further details are given in the text.

VI. CENTRAL TRANSMITTER ACTION

Glutamate is the major transmitter released onto spinal neurons by nociceptive afferents. It activates AMPA–kainate receptors on central neurons to produce rapid EPSPs that excite them. This system is responsible for the rapid, nonpersisting pain acting as a signal to the organism to remove the stimulated tissue from the potentially damaging stimulus.

Cells in the superficial dorsal horn also express another glutamate receptor known as the *N*-methyl-D-aspartate (NMDA) receptor, whose activation can lead to long-lasting activity and extensive metabolic changes in these cells, resulting in persistent pain.

Activation of these receptors is facilitated by corelease of peptides, particularly substance P, from terminals of the afferent fibers in response to intense activity. Substance P elicits a long, slow depolarization of its target cells via activation of tachykinin (NK1) receptors. This depolarization acts to remove the block of postsynaptic NMDA receptors that occurs in the presence of extracellular Mg at normal resting potentials. The details of pre- and postsynaptic changes in central neurons as a consequence of activation of NMDA receptors as well as the interaction between these neural elements in eliciting these long-term changes are presently not completely characterized. Not all nociceptive afferents may be able to elicit long-term changes in this manner because many IB4/c-ret/FRAP-expressing nociceptive afferents are not peptidergic.

VII. TRANSDUCTION MECHANISMS

The availability of molecules eliciting pain has provided important evidence concerning the transduction of noxious stimuli. An extract from hot peppers known as capsaicin is a powerful stimulant of nociceptors and, when administered to human subjects, it evokes the sensation of noxious heat. The receptor for this molecule, VR1, a member of the TRP receptor family, has been cloned and has been found to be expressed in small sensory neurons coexpressing other markers characteristic of nociceptors. When VR1 is expressed in oocytes, they exhibit sensitivity to noxious heat and other stimuli that activate nociceptors, such as protons (low pH). The response of VR1 to heat can be enhanced when the pH of the surrounding medium is reduced, corresponding to the reduced threshold to heat observed behaviorally during inflammatory conditions when pH naturally falls. Thus, this receptor is a logical candidate to be an important participant in the transduction of these nociceptive stimuli. No corresponding molecule for high-threshold mechanoreceptors has been definitively identified.

ATP and its associated receptor family (the P2X family) have also been implicated in peripheral nociceptive events. Damaged cells release ATP, and the fraction of the cytosol that elicits pain when applied to blisters is rich in ATP. In rats, a member of the P2X receptor family, P2X3, is expressed exclusively in sensory neurons, and evidence suggests that its expression is confined to nociceptors. Together these findings suggest that ATP could be an important agent for evoking pain when tissue is damaged.

VIII. SENSITIZATION

Nociceptive pathways are subject to sensitization as a consequence of their previous activation. Elucidation of the mechanisms underlying these changes is exceedingly important in achieving an understanding of nociceptive physiology and pathology. Two important terms used in describing these changes are *hyperalgesia* and *allodynia* (Fig. 3). Hyperalgesia refers to the increase in perceived intensity of a normally painful stimulus after damage to the stimulated region. Allodynia refers to the pain elicited by a normally innocuous stimulus as a consequence of injury. In terms of sensory pathways, hyperalgesia can be thought of as an increased activation of normally nociceptive pathways, whereas allodynia can be considered as a nociceptive response to the activation of normally nonnociceptive pathways. It is now clear that these sensitization processes involve both peripheral and central mechanisms.

A. Peripheral Sensitization

Peripheral sensitization of nociceptive afferents is measured electrophysiologically as a reduction in response threshold to a nociceptive stimulus as well as an increased discharge in response to a given

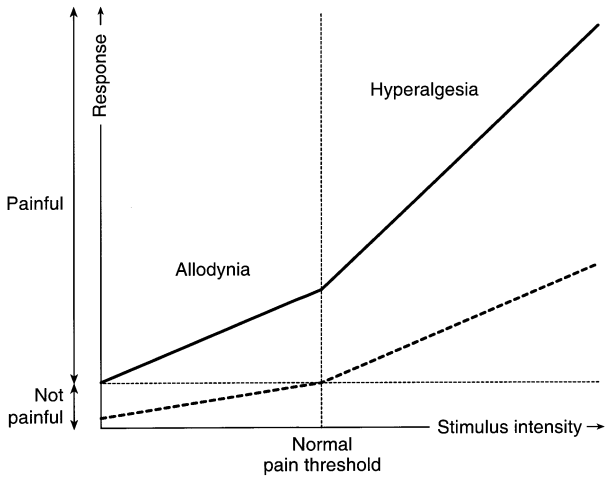


Figure 3 Graphical representation of two components of sensitization. Stimulus–response function before sensitization is shown with the dark dashed line; after sensitization the response increases so that in this example all stimuli elicit the report of pain. Allodynia refers to the painful response to a previously nonpainful stimulus; hyperalgesia refers to the exaggerated response to a normally painful stimulus. The dashed horizontal and vertical lines are provided to give thresholds.

suprathreshold stimulus. Damage to peripheral tissue generally results in sensitization of peripheral nociceptors, although a given stimulus does not sensitize all types of nociceptive responses (e.g., to noxious heat or to high-intensity mechanical stimulation) equally. Similarly, sensitization is not elicited equally well by noxious mechanical and heat stimulation. There are other complications in classifying sensitization responses, e.g., whether it is *primary hyperalgesia* (sensitization in a region that overlaps the region of injury) or *secondary hyperalgesia* (sensitization in a region outside the region of primary injury). Primary heat hyperalgesia after a burn injury involves sensitization of the response of nociceptors to noxious thermal stimuli. Primary mechanical hyperalgesia after a burn injury does not involve peripheral sensitization to mechanical stimuli, and so central changes (i.e., central sensitization; see the following section) are believed to be responsible. However, sensitization of the response of mechanical nociceptors does occur and is believed to account for the mechanical hyperalgesia after chemical sensitization. In summary, there is substantial evidence that both thermal and mechanical responses of nociceptors can be sensitized by natural stimuli.

It is now clear that peripheral sensitization after inflammation involves very complex mechanisms in the periphery resulting from the release of numerous substances in the periphery. There are two major sources of these substances. The first of these are peptides such as substance P and/or CGRP released from the nerve terminals themselves due to a local axon reflex and/or impulses elicited in the spinal terminals of the sensory axons and conducted antidromically, the dorsal root reflex. These peptides evoke several effects in the periphery, notably vasodilatation and neurogenic extravasation (leakage of large molecules from capillaries into the skin). In parallel with these changes is the release into the skin of numerous substances, which have been demonstrated to induce activation and/or sensitization of nocieceptive afferents. These include bradykinin, prostaglandin E2 (PGE2), serotonin, histamine, and nerve growth factor (NGF), each of which is a consequence of a different component of the inflammatory response. Together, this mixture of substances has been referred to as the "inflammatory soup." Numerous cell types in the skin such as macrophages, mast cells, and Schwann cells have been shown to be sources of these substances. It has also been demonstrated that certain cytokines such as TNFα and IL-1β are important precursors to the up-regulation of these substances. Expression of so many substances leading to hyperalgesia suggests that it is a highly adaptive response to inflammatory injury, whereby an inflamed limb is protected to promote the healing process.

The mechanism by which these substances sensitize nociceptive afferents is not presently well-defined. In the case of sensitization to noxious heat, there is evidence that some sensitizing agents, e.g., NGF and PGE2, can acutely enhance the response to capsaicin, suggesting a direct effect on the transduction process mediated by the VR1 receptor believed to mediate the response to noxious heat, at least in part. Attention has been focused on second messenger signaling pathways that elicit sensitization or more definitively on antagonists of these pathways that block the sensitization elicited by endogenously released agents (Fig. 4). For example, activation of the adenylyl cyclase/cyclic AMP/PKA intracellular signaling pathway by forskolin, a specific activator of this pathway, can result in acute sensitization of the response to PGE2. However, the situation is undoubtedly more complex because activators and inhibitors of another signaling pathway involving the molecule PKC can also affect peripheral sensitization to noxious heat.

TTX-resistant sodium channels are also believed to participate in the peripheral sensitization produced by inflammation. These channels are expressed preferentially in nociceptive afferents and their terminals (see previous discussion), and molecules such as PGE2, bradykinin, serotonin, and NGF that elicit inflammatory hyperalgesia selectively enhance the current produced by these channels via intracellular messengers such as PKA and PKC (Fig. 4). This would be expected to decrease the threshold for activation of nociceptive afferents and increase the discharge in response to a given level of depolarization, because TTX-resistant Na channels have a more rapid recovery from inactivation, rendering them more appropriate for steady discharge than TTX-sensitive channels.

In summary, these studies indicate that peripheral sensitization can involve the action of the sensitizing agents on the transduction process (e.g., via VR1), the cell's excitability, and possibly the spike-encoding process (via the TTX-insensitive Na channel).

B. Central Sensitization

It is apparent that not all sensitization in response to injury can be accounted for by changes in the threshold of peripheral nociceptors. Furthermore, under some conditions, preventing afferent impulses from reaching

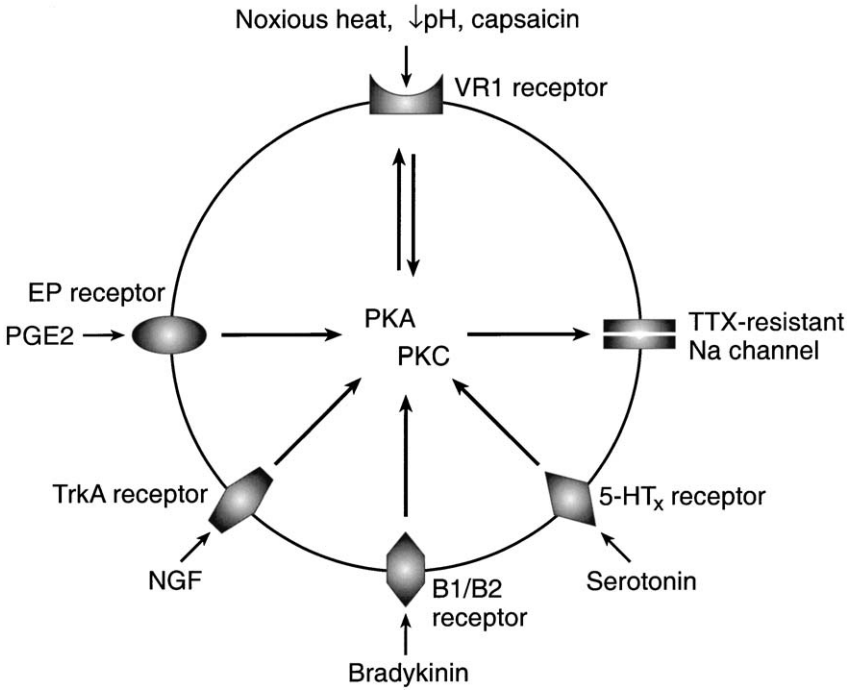


Figure 4 A sensory terminal illustrating the receptors and channels as well as some elements of signaling pathways (PKA and PKC) that may be involved in sensitization. Sensitizing molecules PGE2, NGF, bradykinin, and serotonin act on their receptors to affect the VR1 receptor and the TTX-insensitive Na channel via intracellular signaling. Sensitization of VR1 should result in an increased response to noxious heat and protons as well as capsaicin. Sensitization of TTX-resistant Na channel results in a reduced threshold for steady impulse discharge.

the spinal cord can abolish sensitization. This suggests that some of the changes underlying sensitization must take place central to the peripheral nerve. One mechanism may involve the dorsal root reflex conducted antidromically in sensory nerve fibers from their terminals in the spinal cord, but there is evidence that exclusively central mechanisms also contribute.

It has been recognized for many years that small-diameter afferent fibers differ qualitatively from large ones in the central effects that they exert. Specifically, successive volleys in large-diameter fibers elicit constant central effects, whereas the response to volleys in small-diameter afferent fibers, particularly unmyelinated fibers, can increase in frequency and duration if the peripheral stimuli occur at a high enough rate (at least once per 4 sec in the cat). This progressive increase in the central discharge of cells in the spinal cord has been referred to as *windup* and represents one component of the more comprehensive phenomenon known as central sensitization. Other components are much longer lasting and are thought to require metabolic changes in postsynaptic cells that are induced by the activation of NMDA and metabotropic glutamate receptors. They also lead to changes in gene expression in postsynaptic cells, e.g., *c-fos*, and additionally to a wider zone of postsynaptic cells being activated due to structural changes. The possibility exists that these changes also lead retrogradely to changes in transmitter release from presynaptic terminals via signaling molecules such as nitric oxide.

IX. DEVELOPMENT OF NOCICEPTORS

Sensory neurons are known to be overproduced during development, and in the immediate prenatal period many DRG cells undergo programmed cell death (apoptosis). There is considerable evidence that sensory neurons compete for a trophic substance in the periphery, and those that fail to obtain adequate levels of such a substance are the ones that do not survive. In the case of nociceptive neurons the trophic substance is nerve growth factor (NGF), and small-diameter neurons express the high-affinity receptor trkA for NGF. If NGF is experimentally depleted, fewer than normal nociceptors survive. In the rat the dependence of nociceptors on NGF for survival lasts until about postnatal day 2.

Nociceptors continue to require NGF beyond postnatal day 2 for at least some functions. Their response to noxious heat is substantially reduced if NGF levels are decreased by chronic treatment with an NGF antibody or with an immunoadhesin molecule, trkA–IgG, that binds endogenous NGF. Another indication of trophic support of nociceptors by NGF emerges from the restoration of function when it is provided to axotomized nociceptors. For example, the TTX-resistant Na current and the SNS/PN3 channel both decline in nociceptor somata after peripheral axotomy, and these can be partially restored by the administration of NGF to the cut proximal nerve stump. Similar findings have been made with regard to SP and CGRP expression in these nociceptors.

The dependence of certain nociceptive functions on NGF in postnatal animals implies that nociceptors continue to express the trkA receptor. However, it has been demonstrated in rats that as the animal develops into a young adult, some nociceptors cease expressing trkA and begin to express the receptor c-ret, for which glial-derived neurotrophic factor (GDNF), a member of the TGF-β superfamily, is a ligand.

In the prenatal rat nociceptive afferents, at least those associated with unmyelinated fibers, exhibit responses to skin stimulation similar to those in the adult even before their central terminals penetrate into the spinal cord gray matter. These afferents send their collaterals into the spinal cord at embryonic day 18–19 (E18–E19), which is somewhat later than cells that will become the large nonnociceptive afferents (E15–E16). The biochemical markers associated with nociceptive afferents begin to be expressed during this prenatal period. The spinal terminals of small-diameter afferent fibers are restricted to the superficial laminae of the dorsal horn from the time of their arrival into the cord, as they are in the adult. However, in the first postnatal week, C-fiber volleys are unable to elicit discharges in dorsal horn neurons. After postnatal day 10, activation of C-fibers elicits the same responses as in adult cord, specifically the windup indicative of the rapid component of central sensitization, and activation of immediate early genes, i.e., *c-fos*. Thus, in the rat the nociceptive system activated by C-fibers is not an effective sensory system until the second postnatal week.

X. PLASTICITY OF NOCICEPTORS

One of the hallmarks of nociceptors and the nociceptive pathway is their susceptibility to change in response to various alterations in their environment (Fig. 5). Inflammation in the periphery can result in a diminished threshold to noxious stimuli. In part this reflects the action of various substances in the "inflammatory soup" (see Section VIII.A). However, there are also changes in properties of the nociceptor itself, such that it responds differently to components of the inflammatory soup. For example, the response of individual nociceptive afferents to capsaicin can be enhanced if NGF or PGE2, known to sensitize individual nociceptors, is applied to the neuron. Thus, sensitization at the level of the periphery is thought to represent a change in the sensitivity of the receptor itself, in this case quite possibly a change in the phosphorylation state of the capsaicin receptor (i.e., VR1).

A second locus at which nociceptors exhibit plasticity is the cell body. Inflammation in the periphery results in up-regulation of substance P and CGRP as well as the neurotrophin brain-derived neurotrophic factor (BDNF) in nociceptors. The enhanced level of these substances is believed to result in a more intense driving of the postsynaptic cells on which these nociceptors terminate. The peptides act to prolong the depolarization, thereby relieving the Mg block of postsynaptic NMDA receptors and enabling these cells to be more intensely activated. BDNF directly increases the NMDA-receptor-mediated responses and thereby enhances the response of nociceptive neurons of the dorsal horn that are known to express NMDA receptors.

After peripheral inflammation, some large-diameter afferents that are activated by innocuous stimuli can undergo a phenotypic switch and display characteristics of nociceptors, specifically the expression of substance P in their somata. Release of this peptide could result in intensification of the discharge of some central neurons that they activate. Activation of cells in lamina II by large-diameter afferents can be enhanced under these conditions. These changes could contribute to the allodynia that is often reported after inflammatory injury.

Nociceptive afferents also display considerable plasticity when their axons are cut in the periphery. The level of some peptides (CGRP, SP) decreases, whereas the level of others (e.g., neuropeptide Y and GAP43) increases. This is thought to reflect, at least in part, the changing metabolic priorities of these neurons from a transmission mode to a regeneration mode. Similar changes occur in nociceptors of aged animals, and this has been shown to reflect diminished neurotrophic support from the periphery, e.g., a

Nociceptor Plasticity in Different States

Inflammation	Axotomy	Aging
Increase in NGF/trkA signaling (and in other peripheral sensitizing agents, e.g., PGE2)	Decrease in trkA expression	Decrease in NGF/trkA signaling
Increased levels of SP, CGRP in soma	Reduced levels of SP and CGRP in soma	Reduced levels of SP and CGRP in soma
Increased levels of BDNF in soma	No change in BDNF level in soma	Increase in GDNF/c-ret signaling
	Accumulation of Na channels at neuroma	
	Upregulation of adrenergic receptors	

Figure 5 Nociceptor plasticity in different states. Note the difference between the effects of inflammation and axotomy and the similarity between the effects of axotomy and aging.

diminished level of NGF. However, peripheral levels of GDNF are higher in aged animals, indicating increased trophic support of certain nociceptive functions, i.e., those associated with the IB4/FRAP population of nociceptors.

Afferents that have undergone damage to their axon exhibit other important changes that contribute to very painful conditions after peripheral nerve injury, e.g., causalgia and reflex sympathetic dystrophy. The pain associated with these syndromes is related at least partially to alterations in the expression of channels in the nociceptor membrane. The proximal stump of a cut nerve forms a tangle of nerve fibers referred to as a neuroma. This region becomes very sensitive to mechanical stimuli. This is believed to result from the increased level and duration of firing produced in response to mechanical stimulation by the presence of an increased density of Na channels in the membrane of the growing tip and in preterminal demyelinated segments.

It has long been recognized that the sympathetic nervous system can interact with nociceptive afferents and that certain pain states can be exacerbated by increased sympathetic outflow or diminished by a reduction in sympathetic outflow. These effects are particularly evident in conditions where peripheral nerve damage has taken place, e.g., conditions such as causalgia or reflex sympathetic dystrophy. An increase in responsiveness to adrenergic transmitters is also observed in the terminals of injured nerve fibers with α-adrenergic transmission being the major contributor to this action. For example, after partial lesions of mixed peripheral nerves, intact nociceptive afferents running in the same nerve developed increased expression of α_2-adrenergic receptors and an increased sensitivity to agonists of this receptor.

Together these studies demonstrate that nociceptive afferents are not fixed in their properties. Conditions known to result in changes in nociceptive processing generally are accompanied by changes in the properties of the afferents that contribute to the observed changes. A complete understanding of these changes is required to devise strategies to reduce the undesirable persistence of pain after certain types of injury.

XI. CONCLUSIONS

The wealth of data that is now available on the properties of nociceptors indicates that these afferent fibers have unique properties beyond their fundamental one that distinguishes them from other sensory afferents as nociceptors, namely, the ability to respond only to stimuli that are potentially damaging to the organism. They express a unique blend of receptors and transmitter molecules, resulting in an ability to undergo sensitization peripherally and to induce central sensitization. They follow a different developmental timetable than nonnociceptive afferents, and they exhibit greater plasticity than nonnociceptive afferents. They activate unique populations of neurons in the central nervous system, and the inputs that they provide to the CNS are subject to modification by impulses in fibers descending from the medulla. These

properties provide a significant targets for therapeutic intervention, a fortunate opportunity given the debilitating consequences of persistent pain. Although central targets certainly exist for a therapeutic approach, the nociceptor itself and neurons to which it projects would appear to be a particularly suitable focus for intervention because of their relative isolation as well as the chemical uniqueness referred to earlier.

Some important new approaches are already in progress. For example, it is known that the action of substance P released by nociceptive afferents is terminated by re-uptake into spinal neurons. This knowledge has been used to design a cytotoxin, substance P–saporin, whose uptake kills the cell. This treatment that kills certain populations of spinal neurons has been found to diminish inflammatory and neuropathic thermal hyperalgesia and mechanical allodynia with no loss of response to acute mild nociceptive stimuli. It is thus clear that continued exploration of the detailed molecular properties of nociceptors and their peripheral and central projections should be highly profitable from both the scientific and therapeutic points of view.

See Also the Following Articles

MIGRAINE • OPIATES • PAIN • PERIPHERAL NERVOUS SYSTEM

Suggested Reading

Baron, R., Levine, J. D., and Fields, H. L. (1999). Causalgia and reflex sympathetic dystrophy: Does the sympathetic nervous system contribute to the generation of pain? *Muscle Nerve* **22,** 678–695.

Basbaum, A. I., and Woolf, C. J. (1999). Pain. *Curr. Biol.* **9,** R429–431.

Belmonte, C., and Cervero, F. (Eds.) (1996). *The Neurobiology of Nociceptors*. Oxford University Press, New York.

Dubner, R., and Gold, M. (Eds.) (1999). Colloquium on the Neurobiology of Pain. *Proc. Natl. Acad. Sci. USA* **96,** 7627–7755.

Fields, H. L., Rowbotham, M., and Baron, R. (1998). Postherpetic neuralgia: Irritable nociceptors and deafferentation. *Neurobiol. Dis.* **5,** 209–227.

Julius, D., and Basbaum, A. I. (2001). Molecular mechanisms of nociception. *Nature* **413,** 203–210.

Kumuzawa, T., Kruger, L., and Mizumura, K. (Eds.) (1996). The polymodal receptor—A gateway to pathological pain. *Prog. Brain Res.* **113**.

Levine, J. D., Fields, H. L., and Basbaum, A. I. (1993). Peptides and the primary afferent nociceptor. *J. Neurosci.* **13,** 2273–2286.

Lewin, G. R., and Mendell, L. M. (1993). Nerve growth factor and nociception. *Trends Neurosci.* **16,** 353–359.

Melzack, R., and Wall, P. D. (ed.). (1999). *Textbook of Pain*. Churchill-Livingstone, New York.

Mendell, L. M., Albers, K. M., and Davis, B. M. (1999). Neurotrophins, nociceptors, and pain. *Microsc. Res. Tech.* **45,** 252–261.

Scott, S. A. (Ed.) (1992). *Sensory Neurons: Diversity, Development and Plasticity*. Oxford University Press, New York.

Snider, W. D., and McMahon, S. B. (1998). Tackling pain at the source: New ideas about nociceptors. *Neuron* **20,** 629–632.

Wood, J. N., and Perl, E. R. (1999). Pain. *Curr. Opin. Genet. Dev.* **9,** 328–332.

Norepinephrine

CANDICE DROUIN and JEAN-POL TASSIN
Collège de France, Paris

GLOSSARY

attention Condition of selective awareness governing the extent and quality of the One's communication with one's environment, although not necessarily held under voluntary control. The psychologist William James held that attention made humans perceive, conceive, distinguish, and remember.

electroencephalogram (EEG) The electrical changes taking place within the healthy human brain can be recorded from electrodes attached to the scalp. Such recording enables observation of the minute patterns of voltage fluctuation that take place as the brain cells process information and relay messages.

event-related potentials (ERPs) Consistent patterns that accompany the perception and evaluation of sensory information and that can be recorded on an electroencephalogram. These changes extend over the period of 0.5 sec or so immediately following the onset of the signal concerned. ERPs are composed of a relatively consistent pattern of positive and negative electrical peaks that vary systematically when the properties of the signal that elicits them change. The succession of the different ERP components constitutes a convenient and meaningful indicator of the various aspects of information processing being carried out on the signal. For example, a novel sound produces a prominent negative component about 100 msec (P100) after the sound is heard. A modulation can, however, occur. Indeed, if an individual who is hearing different voices speaking simultaneously in each ear is told to listen for a particular word spoken by one voice only, all words spoken by the "attended" voice elicit a larger P100 than those spoken by the other voice. Only the designated word elicits a later prominent, electrically positive component, occurring about 300 msec (P300) after it is spoken.

habituation Basic mechanism whereby response to novelty wanes with repeated and regular presentation of the same signal, which may be described as a progressive loss of behavioral responsiveness to a stimulus as its lack of adaptive significance, is discovered. Naturally, if the signals have special significance for the individual, they will continue to be attended and responded to, even though they may be, repetitive. For loud or bright stimuli, habituation may take place, but only very slowly.

locus ceruleus Group of neurons situated in the brain stem of vertebrates containing the largest number of noradrenergic cell bodies in the central nervous system of mammals. The term ceruleus (blue) was attributed because the high concentration of the pigment, neuromelanin, in noradrenergic neurons made this nucleus appear as a blue spot under the fourth ventricle in human brain sections.

neuromodulator Refers to neurotransmitters that do not directly convey neuronal information by classical synaptic transmission but rather exert a diffuse modulatory action on information processing. Monoamines (especially serotonin, norepinephrine, and dopamine) exerting their effects via divergent projections through metabotropic rather than ionotropic receptors are considered to be neuromodulators, as opposed to GABA, glycine, and excitatory amino acids.

stress Concept that arose from observations of the endocrinologist Hans Seyle that a large number of damaging or potentially damaging stimuli led to similar physiological reactions. Sympathoadrenal activation is an essential component of the physiological response to stress. Various stimuli, aversive or considered as a challenge to the organism, increase heart rate and plasma epinephrine, norepinephrine, and corticosteroid levels.

vigilance Sustained attention, referring to the state in which attention must be maintained over time. Often this is to be found in some form of "watch-keeping" activity when an observer or listener continuously has to monitor a situation in which significant, but usually infrequent and unpredictable, events may occur. Among the factors that influence vigilance performance, the most important is the frequency with which task-relevant events occur.

Norepinephrine (NE), one of the first neurotransmitters discovered, is commonly associated with the rapid behavioral and physiological response to stress, the *fight or flight* response, related to the secretion of corticosteroids by the adrenal cortex and the release of epinephrine and NE by the adrenal medulla. Present in the brain, NE plays the role of a neurotransmitter in the sympathetic ganglia and that of a hormone following its secretion by the adrenal medulla. In the central nervous system, NE is considered as a neuromodulator. There is now general agreement about the concept of specific neurons that modulate information processing rather than convey sensory or motor signals. In any representation of the primary pathways responsible for the processing of sensory stimuli or motor outputs, it is notable that these pathways include neurons releasing γ-aminobutyric acid (GABA) and excitatory amino acids, possibly acetylcholine, whereas monoaminergic (i.e., noradrenergic, dopaminergic, or serotonergic) cells do not appear to be involved. NE therefore is one neuromodulator among others that, due to its multiple modes of action, is implicated in a wide variety of functions, such as the regulation of blood pressure, sleep–wake cycle, anxiety, pain, or cognitive processes. In this article, we will focus on the actions of NE in the central nervous system; however, general information concerning NE synthesis or adrenergic receptors also applies to the periphery.

I. DISCOVERY OF NOREPINEPHRINE

NE was discovered together with epinephrine in the peripheral nervous system. Initially, a substance that increased the heart rate and was released in the blood following the stimulation of the sympathetic ganglion was characterized by a German physiologist, Otto Loewi. This substance, named *Accelerans* by Otto Loewi and later *sympathine* by Walter Cannon, was finally identified as epinephrine. NE was first considered as an inactive precursor of epinephrine, but Ulf von Euler demonstrated in 1940 that NE was also involved in the sympathetic response in mammals. In 1954, Marthe Vogt further extended the role of NE as a neurotransmitter in the central nervous system.

II. ANATOMY OF NORADRENERGIC NEURONS

Taking advantage of the fluorescent properties of amines after condensation with aldehydes, histochemistry was developed by Erankö in 1955 and applied to brain structure by Falck and Hillarp in 1962. Thanks to this technique, noradrenergic pathways are probably the best characterized in the peripheral and central nervous systems. The noradrenergic cells in the central nervous system are localized in the most posterior part of the brain in two sets of neurons (Fig. 1). First, the locus ceruleus (LC, cell group A6) and adjacent nuclei (A5 and A7) are in the upper pons. Second, the medullary nuclei (cell groups Al and A2) innervate forebrain areas as the hypothalamus, amygdala, septum, and piriform cortex via the ventral NE bundle and play a role in the control of vegetative functions and endocrine regulations. Because this article's main subject is the brain, we will focus on the LC, the major source of central noradrenergic innervation, with the medullary nuclei functions being mainly peripheral. The anatomical, as well as physiological, characteristics of the LC cells have been studied most in the rat, cat, and primate. The most obvious species difference is that LC neurons are essentially noradrenergic in rats and primates and that, in contrast, NE-containing neurons are interspersed with non-noradrenergic neurons in the LC of other species including cat.

A. Locus Ceruleus Efferent Fibers Innervate Many Target Structures

A striking feature of the LC is the extreme divergence of its projections, partly due to an intense collateralization of the axons. Indeed, a few neurons (20,000 in humans and 1,500 in rats) provide extensive innervation of the brain from the olfactory tubercle to the spinal cord (Fig. 1).

Projections of the LC are organized in two ascending fiber systems: first, the dorsal NE bundle innervating the amygdala, olfactory tubercle, septum, bed nucleus of the stria terminals, hippocampus, entire neocortex, habenula, ventral and anterior thalamus, and hypothalamus, and second, the rostral limb of the dorsal periventricular pathway. Projections also include descending fibers to the cerebellum, medulla oblongata, and spinal cord. A certain topographical organization is observed because neurons from the ventral part of the LC preferentially project to the spinal cord, cerebellum, and hypothalamus, whereas neurons from the dorsal part project to the cortex and hippocampus. A functional organization was also observed concerning the projections to somatosensory structures. Studies employing double-fluorescent

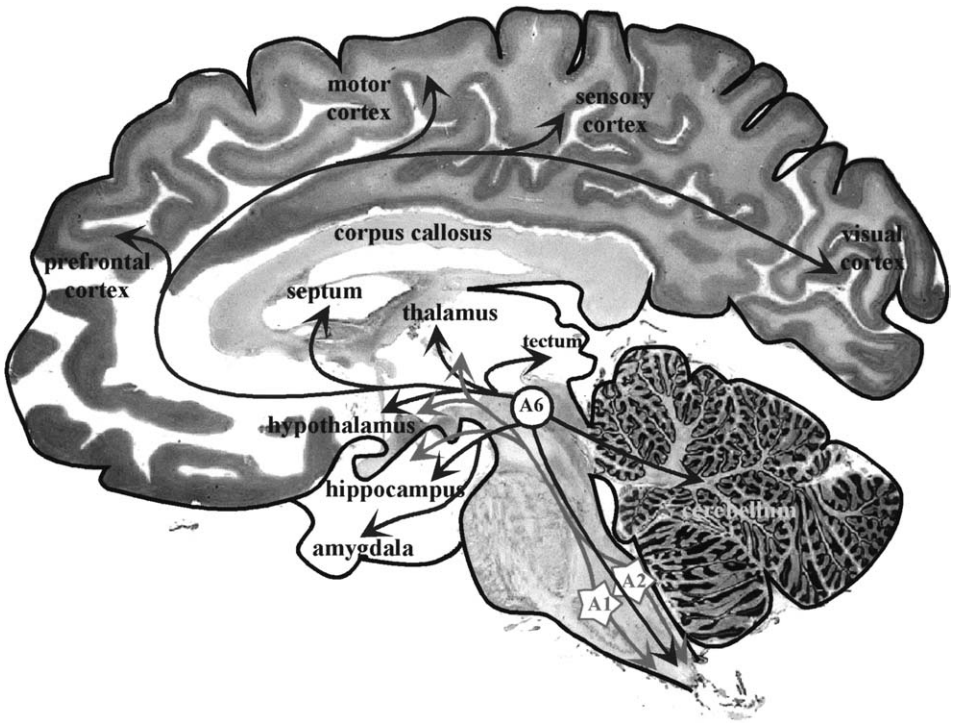


Figure 1 Noradrenergic pathways in the human brain. Sagittal human brain section containing the main projection sites of noradrenergic neurons. A6: Noradrenergic nucleus corresponding to the locus ceruleus. It widely projects to the anterior part of the brain and also to the cerebellum and spinal cord. A1–A2: Medullary noradrenergic nuclei, projecting essentially to the hippocampus, thalamus, and hypothalamus.

retrograde labeling strategies revealed that single LC neurons preferentially collateralize to different sites that process the same sensory information.

B. Coexisting Neurotransmitters in Central Noradrenergic Neurons

Some distinction between noradrenergic neurons is also possible in the function of the neuropeptides they coexpress. Up to 14 different neuropeptides are expressed in the human LC, but the most abundant are galanine and neuropeptide Y (NPY). Galanine is expressed in 80% of LC neurons projecting to the hypothalamus, medial thalamus, cerebral cortex, hippocampus, and spinal cord. NPY is present in approximately 20–40% of the LC neurons, very few of which project to the hypothalamus, instead projecting to the thalamus, cortical areas, ventral hippocampus, and spinal cord. Whereas corelease of NE and NPY was observed in the peripheral nervous system, it was never demonstrated in the brain and the function of such a release is not clearly established. Galanine inhibits LC neuron firing and NPY depresses the inhibitory postsynaptic potential induced by NE in pontine slices, which suggests that NPY and galanine exert inhibitory control on noradrenergic activity.

III. BIOCHEMISTRY OF NE

NE synthesis starts in the cytosol and ends in synaptic vesicles, where it is stored at extremely high concentrations (100 m*M* in vesicles versus a few micromolar in the cytosol). The release of NE from synaptic vesicles is triggered by the arrival of an action potential. Its removal from the extracellular space is due to very efficient re-uptake (about 80% of the NE released is re-uptaken in the synaptic terminals). NE essentially is finally metabolized by a monoamine oxidase (MAO) into dihydroxyphenylglycol (DHPG). A small amount of NE is re-uptaken in nonneuronal cells and metabolized by catechol-*O*-methyltransferase (COMT) (Fig. 2).

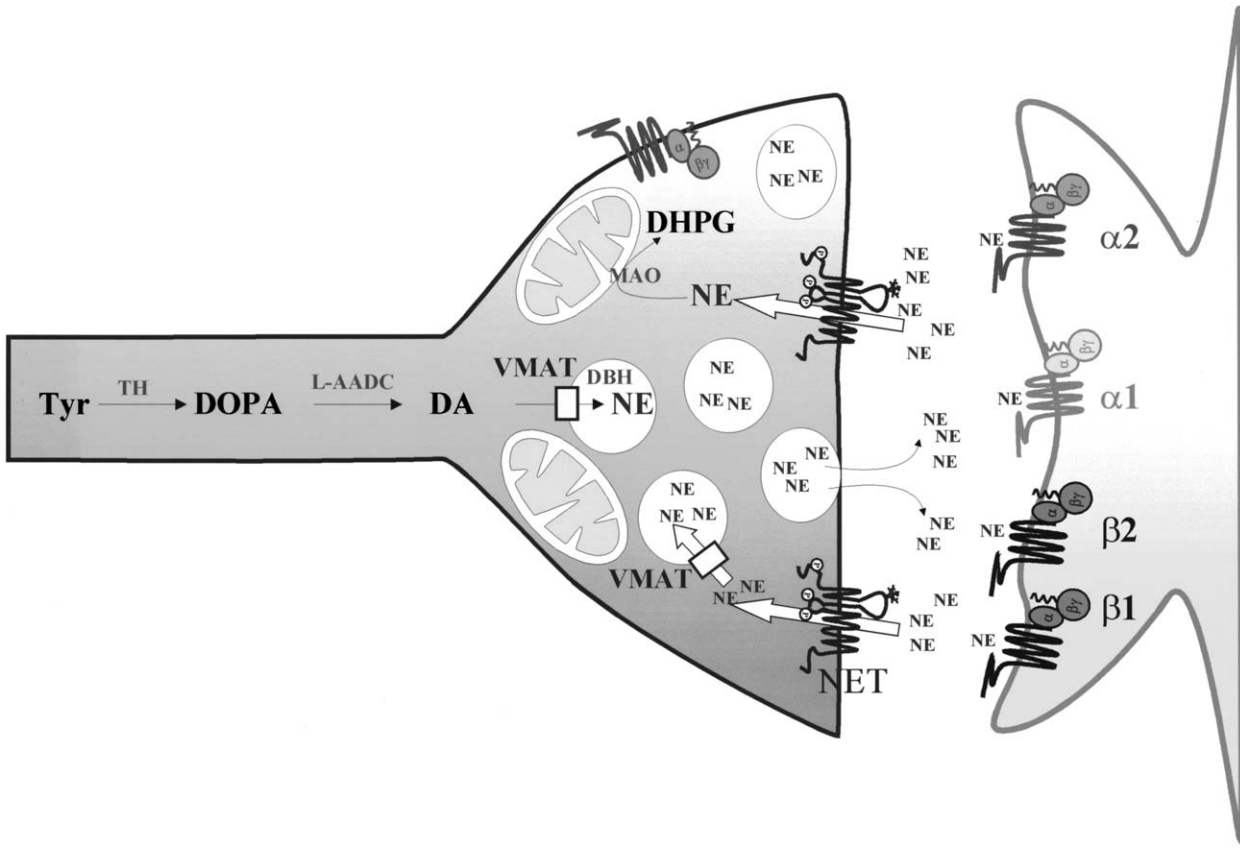


Figure 2 Noradrenergic synapse. NE is synthetized from the amino acid tyrosine (Tyr) via three enzymes, tyrosine hydroxylase (TH), L-aromatic amino acid decarboxylase (L-AADC), and dopamine-β-hydroxylase (DBH). NE and dopamine (DA) are transported in vesicles through the vesicular monoamine transporter (VMAT). After release from synaptic vesicles, NE is re-uptaken by the NE transporter (NET) and further metabolized by the monoamine oxidase (MAO) into dihydrophenylglycol (DHPG). DOPA: dihydroxyphenylalanine.

A. NE Synthesis and Storage

NE is synthesized from the amino acid tyrosine. Four enzymes are necessary (Fig. 3). The first, tyrosine hydroxylase (TH), converts tyrosine to L-dihydroxyphenylalanine (L-DOPA). This enzyme is rate-limiting for the synthesis of both NE and dopamine. It requires a reduced pteridine (Pt-2H) cofactor, regenerated from pteridine (Pt) by another enzyme, pteridine reductase, which is not specific to neurons. TH is subject to multiple controls. Short-term regulation includes activation of TH in response to increased nerve traffic and negative feedback control through end-product (i.e., dopamine and NE) inhibition. Short-term activation of TH involves phosphorylation at four sites corresponding to serine residues 8, 19, 31, and 40 in the rat sequence. In particular, the phosphorylation of serine 40 increases the dissociation rate between TH and its endogenous inhibitor. Long-term regulation involving the production of new TH enzyme occurs at the level of transcription and translation in response to various stimuli via three main second messengers: cyclic AMP, diacylglycerol, and calcium. Nerve growth factor and cell–cell contacts are important during development to promote and maintain the catecholaminergic phenotype, and neurotransmitters and glucocorticoids mediate the activation of TH in response to environmental changes such as stress and exposure to certain drugs such as reserpine, nicotine, or cocaine.

L-DOPA is decarboxylated by a nonspecific L-aromatic amino acid decarboxylase to give dopamine and carbon dioxide. Dopamine is then transported into vesicles by the vesicular monoamine transporter (VMAT) using the proton electrochemical gradient across the vesicle membrane.

Finally, dopamine β-hydroxylase (DBH), which is localized in a membrane-bound (77-kDa) or soluble

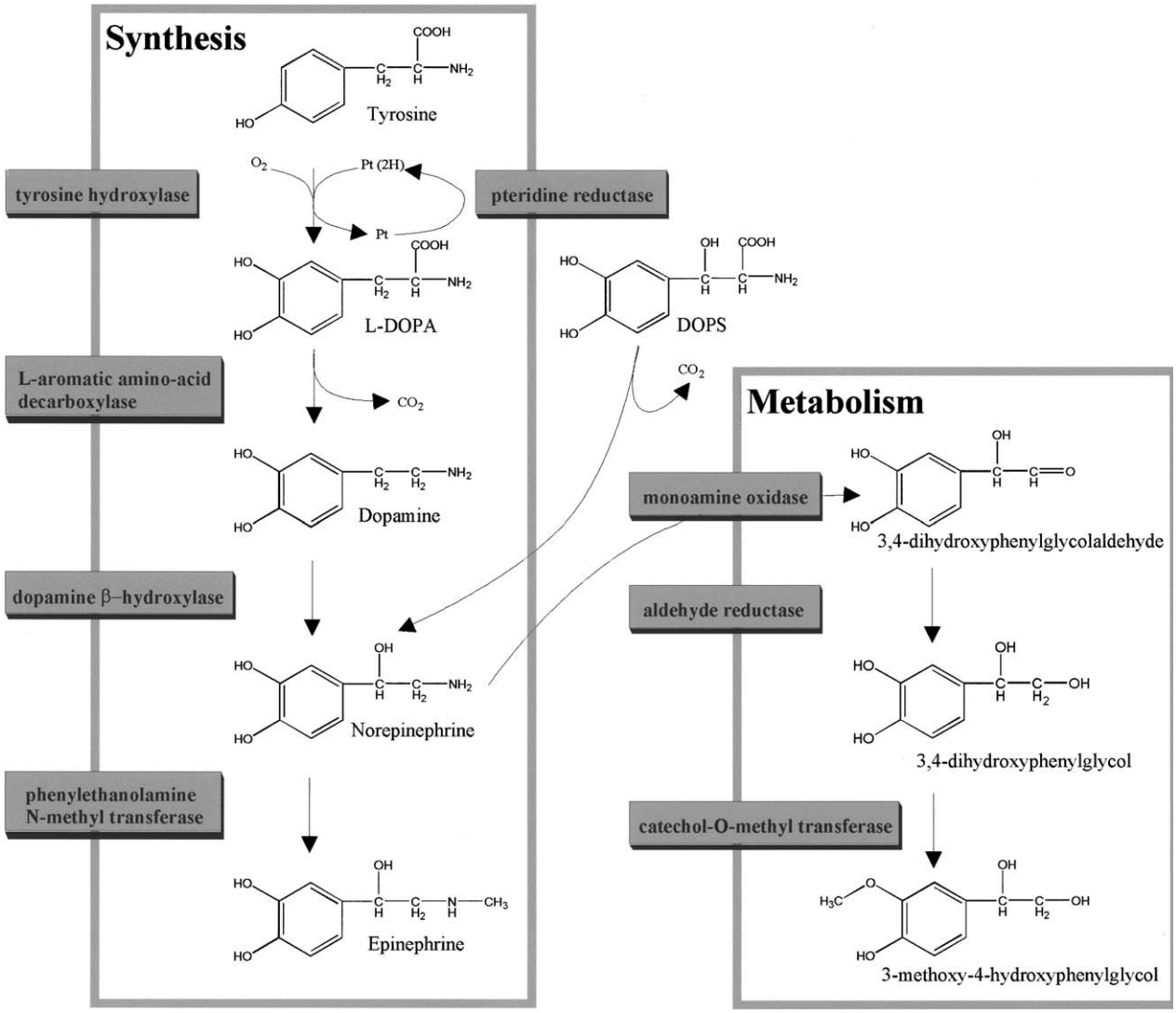


Figure 3 NE synthesis and metabolism. NE is synthetized from the amino acid tyrosine. Tyrosine is transformed into L-dihydroxyphenylalanine (L-DOPA), together with the oxidation of a cofactor, reduced pteridine (Pt-2H). Pteridine is regenerated by pteridine reductase. L-DOPA is decarboxylated by the L-aromatic amino acid decarboxylase to form dopamine. Dopamine is converted to NE by dopamine-β-hydroxylase. In adrenergic cells, phenylethanolamine-*N*-methyltransferase methylates NE to form epinephrine. In the case of dopamine-β-hydroxylase defiency, NE can be synthesized from dihydroxyphenylserine (DOPS), which is decarboxylated by the L-aromatic amino acid decarboxylase to form NE. NE is metabolized by monoamine oxidase into dihydroxyphenylglycolaldehyde, which is immediately reduced to dihydroxyphenylglycol (DHPG). Dihydroxyphenylglycol is methylated by catechol-*O*-methyl transferase to form methoxyhydroxyphenylglycol (MHPG).

form (73-kDa) in the synaptic vesicles, converts dopamine to NE. Both forms of DBH arise from a single translational product. The 77-kDa form has an uncleaved signal sequence. DBH gene expression is activated by a subset of conditions that elevate TH gene expression, but DBH gene expression often requires more severe and more prolonged treatment than TH. In the adrenal medulla and also in sparse neurons located in the central nervous system, a fifth enzyme, phenylethanolamine-*N*-methyltransferase, methylates NE to form epinephrine.

B. Noradrenaline Re-uptake and Metabolism

Re-uptake of released NE into presynaptic nerve terminals is responsible for the rapid termination of

neurotransmission in noradrenergic synapses. Transport of NE by the neuronal NE transporter (NET) is dependent on extracellular Na^+ and Cl^-. cDNAs of a series of neurotransmitter transporters have been cloned, and the NET has been shown to be a member of the superfamily of structurally related Na^+- and Cl^--dependent transporters for monoamines (dopamine, serotonin, and NE) and certain amino acids such as GABA and glycine. Transporters of this family are structurally characterized by 12 transmembrane domains, intracellular amino and carboxy termini, and a large second extracellular loop (Fig. 4). Three consensus sequences for phosphorylation by protein kinase C are found: one in the second intracellular loop and two in the carboxy-terminal end. Phosphorylation of NET by protein kinase C results in down-regulation of NE transport, presumably corresponding to a diminution of the number of transporters expressed to the membrane. NET is one of the different pharmacological targets of several psychotropic substances, such as tricyclic antidepressants and psychostimulants (amphetamine and cocaine) (Fig. 8).

Following re-uptake, monoamine oxidase (MAO) is essential in the neuronal degradation of NE (Fig. 3). Two isoforms of MAO were described, MAO-A and MAO-B. MAO-A is predominant in noradrenergic neurons, whereas MAO-B is found in serotonergic cells. MAO is located on the outer membrane of mitochondria where it uses the FAD cofactor to convert NE into dihydroxyphenylglycolaldehyde, which is reduced into dihydroxyphenylglycol (DHPG). DHPG is further metabolized into methoxyhydroxyphenylglycol (MHPG) outside neurons.

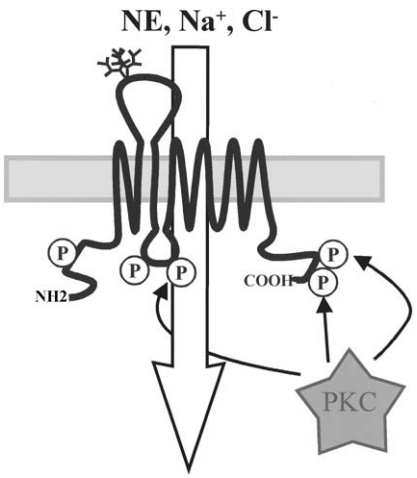


Figure 4 NE transporter. NE transporter is a 12-transmembrane-domain protein with intracellular amino and carboxy termini, a large glycosylated second extracellular loop, and five sites of phosphorylation (P), three of which are consensus sites for protein kinase C (PKC). NE is cotransported with sodium (Na^+) and chloride (Cl^-).

IV. CONTROL OF NORADRENERGIC NEURON FIRING

A. Organization of Afferents to the LC

The organization of afferents to LC has been a subject of controversy for a long time. Early retrograde tracing studies (using the horseradish peroxidase–diaminobenzidine technique) led to the conclusion that LC receives afferents from more than 30 regions (including the dorsal prefrontal cortex, amygdala, ventral tegmental area, raphe nucleus, contralateral LC and deep cerebellar nuclei, and dorsal horn of the spinal cord). However, more recent experiments using new tracers allowing more focal injections restricted the source of afferents to only two structures: the nucleus paragigantocellularis (PGi) and the nucleus prepositus hypoglossi (PrH) in the medulla oblongata. Anterograde analysis further confirmed that the ventral tegmental area, the dorsal horn of the spinal cord, the rostral solitary tractus, and the prefrontal cortex do not project to the LC but to adjacent structures.

Electrophysiological studies did not confirm, however, that PGi and PrH are the sole sources of afferents to the LC. First, the destruction of both PGi and PrH does not prevent LC responses to sensory stimuli. Second, it remains that stimulation of the ventral tegmental area, prefrontal cortex, and raphe nucleus alters LC firing, a finding suggesting that their projections to structures adjacent to the LC might be implicated.

The answer to this paradox came from light and electronic microscopic analyses, which revealed that many LC neurons extend their dendrites to pericerulear regions (rostromedial and caudal juxtaependymal regions). These pericerulear regions are sometimes denoted as the shell of the LC and distinguished from the core of the LC that represents the LC proper. These shell dendrites are selectively targeted by a number of extrinsic afferent inputs that do not project appreciably into the LC core.

Although it is not the primary LC afferent, PGi remains an important afferent to the LC and is a key sympathoexcitatory region in the brain. There are

widespread afferents to the PGi from diverse brain areas. The PGi may be an integrative pathway for activating the LC by a variety of mechanisms because it is involved in the control of arterial blood pressure, cardiopulmonary reflexes, and sympathetic function. It has been suggested that the peripheral sympathetic nervous system is activated in parallel with the LC by projections to both areas from the PGi.

B. Neurochemical Regulation of LC Neuron Firing

Immunohistochemical methods have shown that many neurotransmitters may contribute to the regulation of LC–NE neuron activity. Most of them are found in neurons of the two major LC afferents, the PGi and the PrH. These include epinephrine, excitatory amino acids, enkephalin, corticotropin releasing factor (CRF), substance P, serotonin (5-HT), and γ-aminobutyric acid (GABA). Adrenergic innervation of the LC derives from the PGi and may mediate part of the postactivation inhibition in LC responses to footshocks. GABA innervation arises from the PrH, the PGi, and other nuclei. The LC receives a rich enkephalinergic innervation, a large percentage of which arises from the PGi and PrH, and LC neurons are particularly invested with μ-opiate receptors. CRF innervation of the LC is noteworthy because the LC–NE system is sensitive to most stressors (see Section IV.C). CRF is presumed to play a crucial role in eliciting the coordinated set of endocrine, autonomic, and behavioral events that constitute the stress response. This peptide was initially characterized as the hypothalamic hormone inducing adrenocorticotropin release, but CRF is now considered as a neurotransmitter that, among others, stimulates the LC in response to hypotensive stress. CRF seems to originate from the PGi and the hypothalamus.

LC neurons also carry α2-adrenergic autoreceptors, allowing inhibitory feedback via recurrent collaterals. Intracellular studies have demonstrated that the stimulation of α2-adrenergic receptors inhibits firing by opening an inwardly rectifying potassium channel via a G-protein. The injection of clonidine, an α2-adrenergic agonist that suppresses LC neuron activity, is often used as a tool to characterize NE–LC neurons. These cells also contain many μ-opiate receptors, whose similarly activation results in an inhibition of the firing by a mechanism identical to that of α2-adrenergic receptors. In particular, μ-opiate and α2-adrenergic receptors share the same potassium channels.

Several inputs to the pericerulear region exert an excitatory influence on LC discharge, suggesting that these projections may be glutamatergic. Other transmitter receptors were identified on these distal dendrites, including CRF receptor, met-enkephalin, and μ-opiate receptors.

C. Electrotonic Coupling in LC Neurons

In some cases, LC neurons can exhibit synchronized activities. This property does not require classical synaptic transmission, but rather requires gap junctions that are located on pericerulear dendrites allowing communication between neurons with distant cell bodies. Electrical and dye couplings between LC neurons are often found in rats younger than 10 days old, whereas such a coupling is difficult to demonstrate in adults.

D. *In Vivo* Responsiveness of LC Neurons

1. Spontaneous Activity

The spontaneous discharge of LC neurons varies with the stage of sleep–wake cycle, 2 Hz during waking, lower during slow-wave sleep (from $\sim$1.5 Hz in stage 2 to $\sim$0.2 Hz in stage 4), and virtually absent during paradoxical sleep ($\sim$0.02 Hz). Rate of discharge also varies during different kinds of waking behavior, suggesting that LC activity is correlated with the state of arousal even during waking. Typically, LC neurons exhibit relatively little activity during low-vigilance behavior, such as grooming and food consumption, whereas their activity increases during exploratory behavior.

2. Responsiveness to Sensory Stimuli

LC neurons also exhibit phasic responses of short latencies (15–50 msec) to sensory stimuli in every modality (auditory, visual, somatosensory, and olfactory). These responses consist of a brief excitation followed by an inhibition lasting for several hundred milliseconds.

The responsiveness of LC varies between species. In rats, any of a variety of intense stimuli evokes LC responses in a majority of sensory trials. However, the response habituates with repetitive stimulus

presentation. In monkeys, LC is influenced less strongly by intense stimuli, but again, when responses occur, they fade after the first few trials. Intense conspicuous stimuli that interrupt behavior and elicit an orientating response reliably induce a LC response. Similarly, in complex behavioral tasks such as the *oddball*, where target cues are presented in a semirandom fashion with nontarget cues, target stimuli, even nonintense, are sufficient to induce a LC response (Fig. 5).

In cats, when the animal is well-habituated to its environment and is in a quiet waking state, LC neurons often display very little response to previously presented neutral stimuli. The presentation of novel or noxious stimuli will, in contrast, elicit a burst of unit activity (see the following section).

More generally, it can be considered that *tonic activity* of noradrenergic LC neurons is highly correlated with changing degree of alertness, whereas LC neurons respond *phasically* to behaviorally significant stimuli.

3. Responsiveness to Stressful Stimuli

Stressful stimuli such as confrontation with a dog (for cats), footshock, cold environment, loud noise, and immobilization reliably stimulate LC activity, an excitation usually accompanied by a sympathetic activation. This has led to the proposal of the LC as a central analog of peripheral sympathetic ganglia.

Physiological stress (hypotensive challenge, bladder distension, colon distension, and sciatic nerve stimulation) also reliably stimulates LC neurons even in nonnoxious magnitudes. Some of these effects are mediated via CRF release within the LC (for example, hypotensive stress as previously mentioned), but this is not the rule and different stresses act through different mechanisms.

Whereas chronic presentation of a nonnoxious stimulus results in the diminution of the LC response, chronic stress often results in sensitized noradrenergic activation by a subsequent stress. This noradrenergic

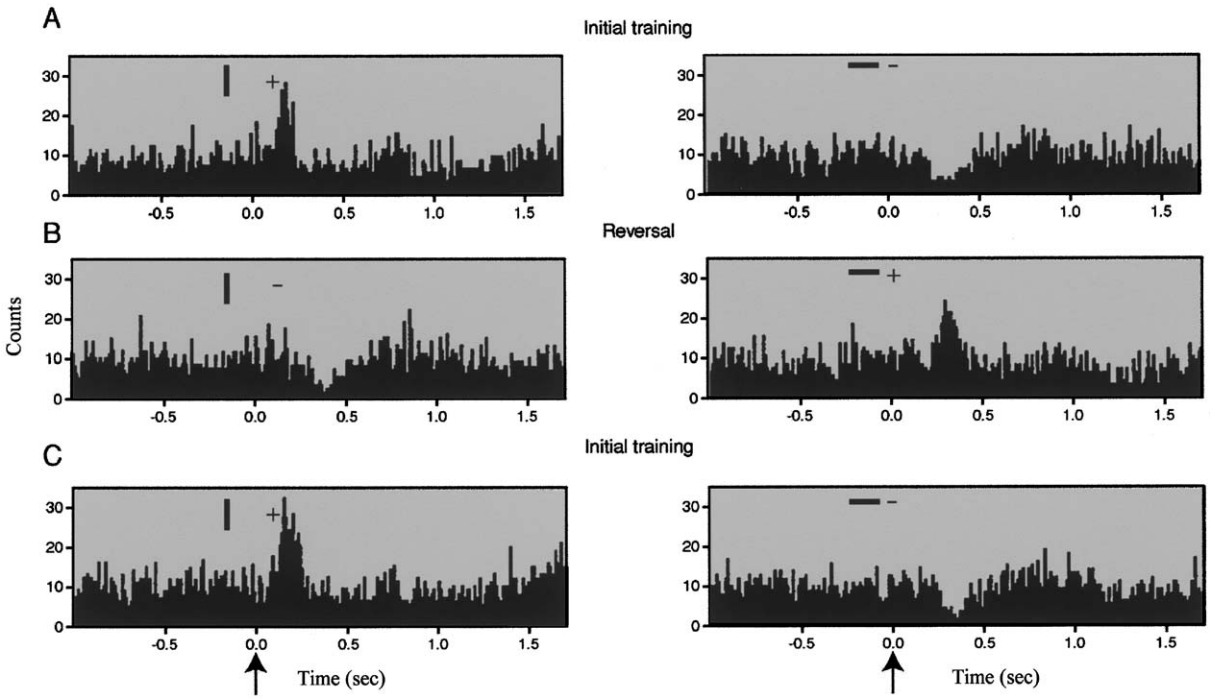


Figure 5 Noradrenergic neuron responses during the *odd ball* task. In the *odd ball* task, either vertical or horizontal bars are presented to monkeys. The activity of their noradrenergic neurons is shown here in a peristimulus histogram, which means that all recorded responses were added, with time zero corresponding to stimulus presentation (arrows). In A, vertical bars are presented in 10% of the trials and the animal response is rewarded (+). Horizontal bars are presented in 90% of the trials and are not rewarded (−). In B, the rule is reversed. Horizontal bars are presented 10% of the time and rewarded (+), whereas vertical bars are presented 90% of the time and not rewarded (−). In C, the initial rule is back. LC neuron firing is specifically increased by the presentation of the rewarding stimulus. From Aston-Jones, G., *et al.* (1999). Attention. In *Fundamental Neuroscience*, Academic Press, San Diego.

sensitization parallels an increased behavioral response.

V. IMPACT OF NE ON TARGET STRUCTURES

A. The Variety of Noradrenergic Receptors

Adrenoreceptors are membrane-bound, G-protein-coupled receptors located throughout the body on neuronal and nonneuronal tissues, where they mediate a diverse range of responses to NE and epinephrine. They share the common features of an extracellular amino terminus, seven transmembrane domains, and an intracellular carboxy terminus.

1. Adrenoreceptor Classification

The adrenoreceptor family was initially divided into two subtypes, the α- and β-adrenoreceptors, as first proposed in 1948 by Raymond Ahlquist, an American pharmacologist. A quarter of a century later, the α-adrenoreceptors were further subdivided by Langer according to their anatomical location, with α-adrenoreceptors located on sympathetic nerve terminals designated as α2-adrenoreceptors and those located postsynaptically designated as α1-adrenoreceptors. This anatomical classification rapidly gave way to the identification of pharmacological differences between α-adrenoreceptors, notably, the ability of yohimbine and rauwolscine to act as α2-adrenoreceptor antagonists. Subsequent studies using pharmacological and molecular biological techniques have further subdivided the α–adrenoreceptor family; three subtypes within each group have now been cloned and pharmacologically characterized. The αl-adrenoreceptor subtypes have been classified as the α1A-, α1B-, and α1D-adrenoreceptors, and the α2-adrenoreceptors have been classified as α2A-, α2B-, and α2C-adrenoreceptors (see Fig. 6).

a. α1-Adrenergic Receptors Subdivision of the αl-adrenoreceptors has been facilitated by both pharmacological and molecular biological techniques. The initial classification of α1-adrenoreceptors as α1A and α1B subtypes was determined from differences in the binding characteristics of the competitive antagonist WB 4101 and a site-directed alkylating agent, chloroethylclonidine (CEC). In addition, three different cDNAs, which coded for α1 subtypes, were isolated. These have since been characterized and are believed to code for three functional α1-adrenoreceptors: the α1A and α1B subtypes as described earlier and a third subtype, the α1D. A putative α1L-adrenoreceptor shows characteristics similar to those of the α1A- and α1D-adrenoreceptors, but the gene has not been identified.

The α1-adrenoreceptors mediate their response via the Gq/l1 mechanism. All of the subtypes are coupled to phospholipase C, and activation of the receptor results in the production of inositol triphosphate (IP_3) and diacylglycerol (DAG). The production of these second messengers results in the activation of both

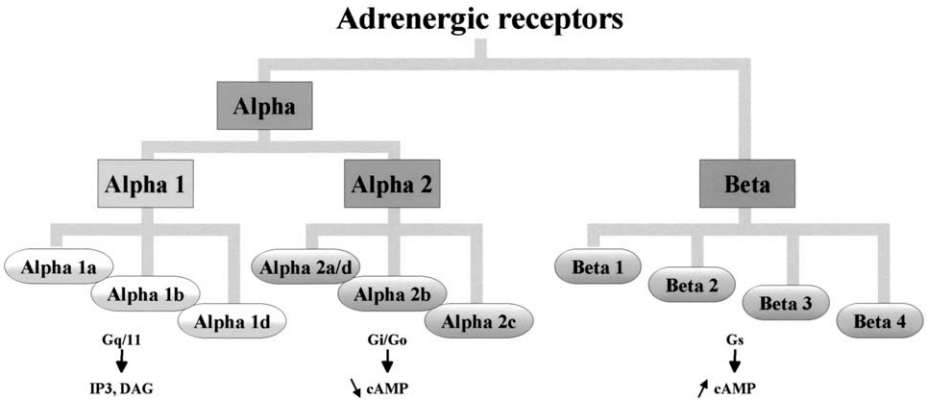


Figure 6 Adrenergic receptor classification. Adrenergic receptors are divided into α1-, α2-, and β-adrenoreceptors. These categories are further subdivided into subtypes, for which genes have been identified. α1-Adrenorecptors are coupled to the G-protein Gq/11, which activates the phospholipase C, producing inositol triphosphate (IP3) and diacylglycerol (DAG). α2-Adrenoreceptors are coupled to the G-protein G_i/G_o Which inhibits cyclic adenosine monophosphate (cAMP) production by adenylyl cyclase. β-Adrenoreceptors are coupled to the G-protein G_s, which stimulates adenylyl cyclase production of cAMP.

voltage-dependent and -independent calcium channels, as well as protein kinase C and phospholipase A2 and D stimulation, arachidonic acid release, and cyclic AMP formation.

The α1-adrenoreceptors are located in both the central and peripheral nervous systems. In the CNS, they are predominantly located postsynaptically where they mediate an excitatory response and are particularly concentrated in the anterior cerebral cortex and thalamus. Peripheral α1-adrenoreceptors are located on both vascular and nonvascular smooth muscle, where their activation results in contraction. On the vascular smooth muscles, the α1-adrenoreceptors are located intrasynaptically where they mediate the response to endogenous neurotransmitter release. They are also located in the heart, where they mediate a positive ionotropic effect, and in the liver, where they activate glycogen phosphorylation.

b. α2-Adrenergic Receptors The α2-adrenoreceptors are located on both pre- and postsynaptic neurons, where they mediate an inhibitory role on the central and peripheral nervous systems. The heterogeneous nature of the α2-adrenoreceptor was first determined from the different pharmacological profiles of the receptor between species, and subsequent studies have revealed the presence of different subtypes within the same tissue. Thus, on the basis of radioligand binding profiles, amino acid sequence, and chromosomal location, four distinct subtypes of α2-adrenoreceptor have been characterized. These α2-adrenoreceptor subtypes, α2A, α2B, α2C, and α2D, are found in a variety of species and tissues and have been characterized by using tissue and cell lines expressing only one subtype. The α2D subtype exhibits a distinct pharmacological profile, but from the sequence homology it is believed to be a species variation of the α2A subtype and is not recognized as separate.

α2-Adrenoreceptors mediate their functions through a variety of G-proteins including G_i and G_o. All of the subtypes have been shown to be negatively coupled to adenylyl cyclase and to mediate an inhibitory effect through the inhibition of cyclic AMP production. In addition, evidence now links the α2-adrenoreceptors to the stimulation of Ca^{2+} influx and also the activation of K^+ channels, phospholipase A2, and Na^+–H^+ exchange.

α2-Adrenoreceptors are found in both the central and peripheral nervous systems, located either pre- or postsynaptically. In the CNS, this receptor can regulate neurotransmitter release as an autoreceptor when located on noradrenergic nerve terminals or as a heteroreceptor when bound on nonnoradrenergic nerve terminals. This role in regulating the release of both NE and serotonin has stimulated the investigation and development of α2-antagonists such as idazoxan in order to cure mental depression.

c. β-Adrenergic Receptors The β-adrenoreceptors were first subdivided into β1- and β2-adrenoreceptors following comparison of the rank order of potency of various adrenergic agonists. The β1-adrenoreceptor is predominant in the heart and adipose tissue and displays equal affinity for epinephrine and NE. In contrast, the β2-adrenoreceptor is predominant on vascular, uterine, and airway smooth muscle and exhibits a higher selectivity for NE than epinephrine.

The classification of β-adrenoreceptors is not limited to β1- and β2-adrenoreceptors. Characterization of β-adrenoreceptor-mediated responses resulted in evidence for further atypical subtypes, β3 and β4, which are insensitive to typical β-adrenergic antagonists. The β4-adrenoreceptor is localized in the cardiac tissue.

The β1-adrenoreceptor is positively coupled to the adenylyl cyclase via activation of G_s. β2- and β3-adrenoreceptors are also coupled to G_s, but their activation can result in either the stimulation or inhibition of adenylyl cyclase. Activation of the β4-adrenergic receptor results in increased cyclic AMP and the stimulation of cyclic-AMP-pendent protein kinase. There is also evidence to suggest that β-adrenoreceptors are linked to voltage-gated Ca^{2+} channels.

The β1- and β2-adrenorecetors have distinct patterns of distribution in the CNS, as determined by using *in situ* hybridization. In the rat, β1-adrenoreceptors are found in high densities in the striatum, although there is an almost complete absence of noradrenergic innervation in that structure, whereas the highest brain concentration of β2-adrenoreceptors is found in the cerebellum.

2. Adrenoreceptor Desensitization and Resensitization

A regulatory feature shared by many of the members of the superfamily of G-protein-coupled receptors is that of desensitization. In response to prolonged or repeated agonist exposure, dampening of the signal transduction process is observed. Desensitization represents the summation of several different processes, including receptor phosphorylation, receptor

sequestration, enhanced degradation of intracellular messengers, and degradation of receptor protein.

Rapid receptor desensitization appears to be mediated by uncoupling of the receptor from its respective G-protein, a consequence of receptor phosphorylation. This mechanism has been studied particularly for the β2-adrenoreceptor, which is phosphorylated on serine and threonine sites by PKA or PKC (Fig. 7A).

β2-Adrenoreceptors are also phosphorylated by β-adrenergic receptor kinase (β-ARK), a member of the G-protein-coupled receptor kinase (GRK) family (Fig. 7B). GRKs are kinases specialized in receptor desensitization. After phosphorylation by GRKs, receptors are recognized by cytosolic proteins from the arrestin family, and binding of arrestin proteins blocks the activation of the G-protein. Once uncoupled from the G-protein, the receptor function can only be restored by receptor dephosphorylation. The extent and duration of receptor desensitization depend not only on the activity of GRKs but also on the activity of the G-protein-coupled receptor phosphatases (GRP). β2-Adrenergic receptor uncoupling is rapidly followed by sequestration into an intracellular compartment distinct from the plasma membrane. The low pH in these endocytosis vesicles induces a conformational change in the receptor, allowing dephosphorylation by GRP. Once dephosphorylated, the receptors are recycled back to the plasma membrane.

Although it has been studied particularly for β-adrenergic receptors, desensitization also occurs for α-adrenergic receptors; β1-adrenergic and α2-adrenergic receptor desensitization has been proposed as a mechanism of action of chronic antidepressant treatment.

B. Several Cellular Targets

The hormonal function of NE raised the question of whether NE had a classical synaptic transmission in the brain. Early studies on the noradrenergic nerve terminals concluded that they do not make synaptic contacts. But more recent data obtained from the cerebral cortex, diencephalon, and cerebellum indicated that noradrenergic nerve terminals exhibit conventional synapses. Both synaptic and paracrine modes of transmission are likely to coexist. Additionally, NE also participates in neuroastrocytic transmission. Cultured glial cells express both α- and β-adrenergic receptors. Brain β2-adrenergic receptors are bound on astrocyte membranes rather than on neurons, and contacts between noradrenergic nerve terminals and astrocytic processes bearing β-adrenergic receptors have been observed. Stimulation of these receptors was shown to increase glycogenolysis and induce NGF release. α1-Adrenergic receptors are observed more frequently on neuronal cells (Fig. 8) but are also present on astrocytes, and their stimulation was found to reduce dye coupling in striatal astrocytes.

C. Postsynaptic Effect of NE and LC Stimulation

1. Effect on Neuronal Discharge

The postsynaptic effects of NE have been studied in several target structures of the LC, such as the neocortex, hippocampus, and thalamus. The impact of LC on the electrophysiological activity of its target neurons was assessed by *in vivo* or *in vitro* application of noradrenergic agonists or antagonists and by *in vivo* LC activation.

Numerous studies have been conducted on the spontaneous and sensory-evoked responses of the neocortical neurons, indicating that NE application or LC stimulation is particularly efficient to enhance the responsiveness of these neurons to sensory stimulation. Indeed, NE does not change the amplitude of evoked potentials but reduces the spontaneous activity of neocortical target cells via both excitatory and inhibitory inputs. Altogether, NE seems to enhance the signal-to-noise ratio of sensory-evoked responses, thus making salient new features of the environment. More recently, noradrenergic activation was found to "gate" inputs to target neurons, which means that subthreshold synaptic inputs become suprathreshold in the presence of NE.

Similarly, facilitory NE effects were demonstrated on various types of synaptically elicited activity in the hippocampus. In addition to these effects, there have been numerous demonstrations of NE enhancement of long-term potentiation in CA3 and the dentate gyrus.

2. Effects on Electroencephalograms (EEG)

Activation of LC neurons by local application of cholinergic agonists induces a desynchronization of cortical EEG, characterized by a shift to high-frequency, low-amplitude activity. This desynchronization is generally considered as an index of sensory

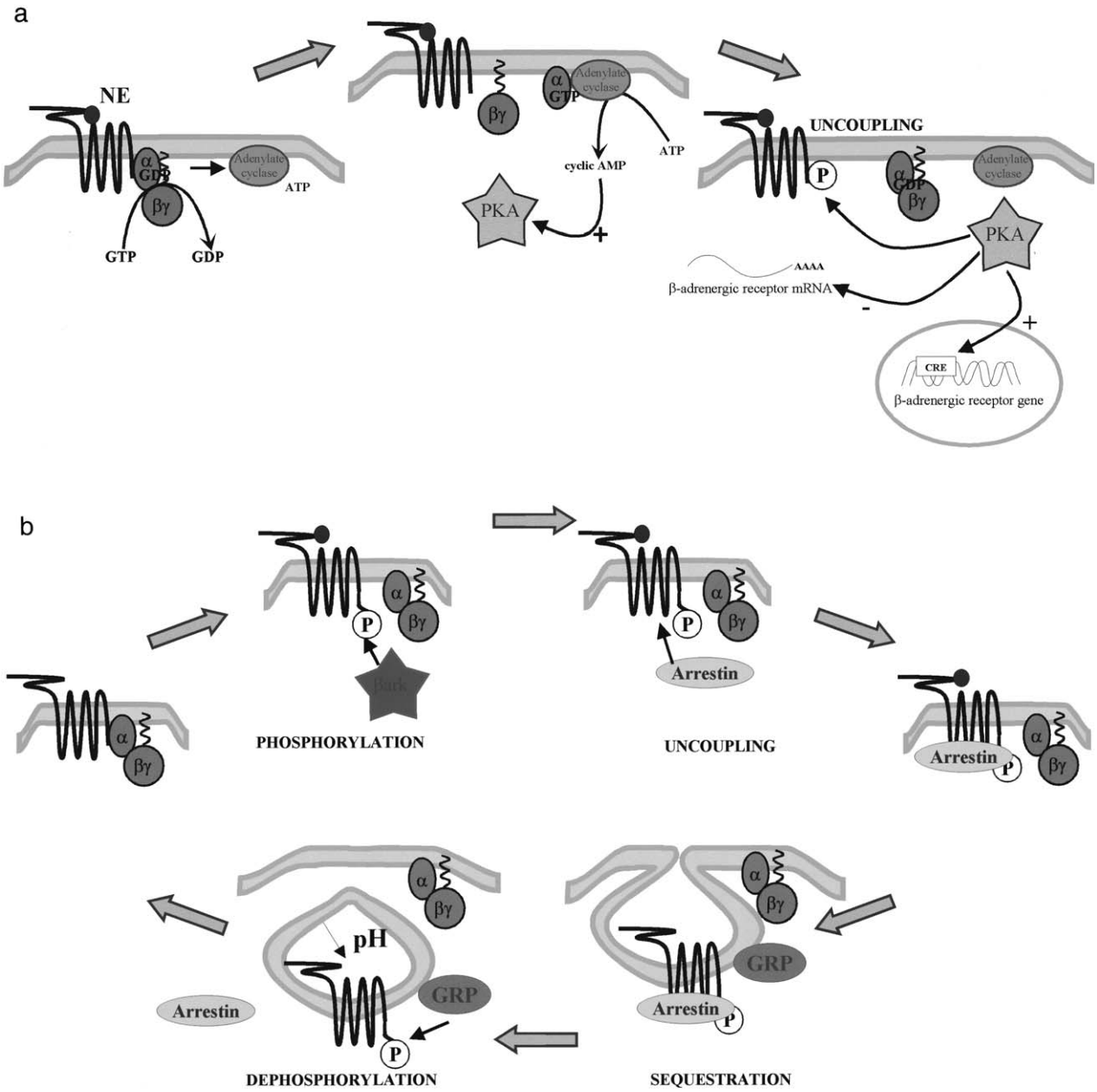


Figure 7 β-Adrenergic receptor desensitization. (A) β-Adrenergic stimulation by NE results in the activation of adenylyl cyclase and the transformation of adenosine triphosphate (ATP) into cyclic adenosine monophosphate (cAMP). cAMP stimulates protein kinase A (PKA). Activated PKA phosphorylates the β-adrenergic receptor, which results in the uncoupling of the receptor from the G-protein. PKA also destabilizes the receptor mRNA and activates the transcription of its gene via the promoter cAMP-responsive element (CRE). Abbreviations: GTP, guanosine triphosphate; GDP, guanosine diphosphate. (B) Stimulated β-adrenergic receptor can also be phosphorylated by the β-adrenergic receptor kinase (β-ARK). A cytosolic protein, arrestin, binds to the phosphorylated receptor, provoking the uncoupling of the receptor from the G-protein and its sequestration in a clathrin-coated vesicle. Following its internalization, the receptor conformation adapts to the low pH maintained in the vesicle. Arrestin separates from the receptor, allowing its dephosphorylation by the G-protein-coupled receptor phosphatase (GRP) present in the membrane and its reexpression to the cell surface.

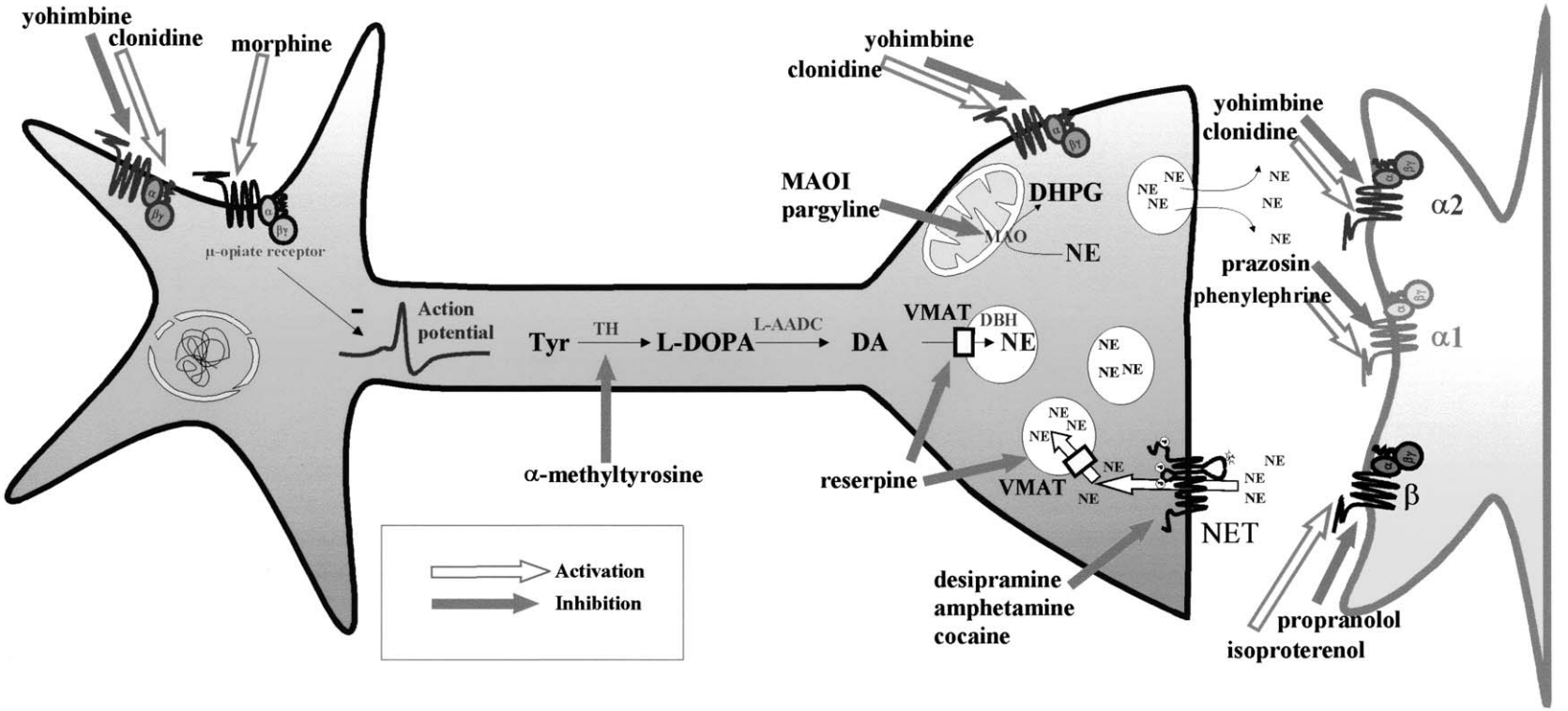


Figure 8 NE pharmacology. Several pharmacological agents are able to modify the noradrenergic transmission. α-Methyltyrosine inhibits tyrosine hydroxylase (TH). Reserpine provokes the release of NE from the synaptic vesicles. Desipramine and cocaine inhibit NE re-uptake by the NE transporter (NET). Amphetamine promotes reverse transport through the NET. Pargyline and many antidepressants (called MAOI for monoamine oxidase inhibitors) block the degradation of NE by monoamine oxidase (MAO). Prazosin, yohimbine, and propranolol (among others) inhibit the action of NE on α1-, α2-, and β-adrenergic receptors, respectively. Phenylephrine, clonidine, and isoproterenol stimulate α1-, α2-, and β-adrenergic receptors, respectively. The action of clonidine or yohimbine on α2-adrenergic autoreceptors also results in the inhibition or activation, respectively, of noradrenergic neuron firing and NE release. Finally, morphine inhibits noradrenergic neuron firing via somatodendritic μ-opiate receptors.

perception. Conversely, LC inactivation by local or peripheral clonidine administration induces a shift in neocortical EEG to low-frequency, large-amplitude activity. Whereas LC innervates the cerebral cortex, the influence of LC on cortical EEG at least partly results from the stimulation of β-adrenergic receptors in the medial septum–diagonal band of Broca, a region that receives dense noradrenergic innervation from the LC.

Long-latency components of event-related potentials include the positive potentials recorded approximately 300 msec (P300) after the occurrence of novel or attention-eliciting events. These P300-like potentials are often considered as electrophysiological correlates of human cognition. Although the electrolytic destruction of LC cells disrupts P300 potentials, this effect is not observed following specific 6-hydroxydopamine chemical destruction of noradrenergic ascending fibers. It is, therefore, possible that non-noradrenergic cells located in or passing through the LC may act in synergy with noradrenergic neurons.

VI. NE IMPACT ON PHYSIOLOGY AND BEHAVIOR

The impact of NE on physiology and behavior was studied in animals by several approaches, including the administration of noradrenergic agonists and antagonists, the stimulation or lesion of noradrenergic neurons, and, more recently, the disruption of genes implicated in noradrenergic transmission. As the widespread distribution of LC innervation through the brain may indicate, NE is implicated in a great variety of physiological and behavioral responses. Most of these responses, including increased heart rate and blood pressure, increased vigilance, hyperarousal, and enhancement of memory storage, are elicited when animals are submitted to unusual or stressful situations.

A. Pharmacological Studies

The effects of noradrenergic agonists and antagonists have been known for several decades, and many are now used clinically. Their major effects are observed at the periphery on the cardiovascular system, interfering with the NE released from sympathetic neurons. Due to their multiple sites of action, systemic administration of these agents sometimes led to difficulties of interpretation, and drugs had to be administered locally in brain structures.

1. α2-Adrenoreceptors

The effects of systemic administration of an α2-adrenergic agonist, such as clonidine, or antagonist, such as yohimbine or idazoxan, are mainly due to their action on autoreceptors, leading to inhibition or activation of noradrenergic neuronal activity, respectively. These effects also implicate the action of NE on postsynaptic adrenergic receptors.

Moderate inhibition of α2-adrenoreceptor by low systemic doses of yohimbine or idazoxan, which increase noradrenergic neuron firing, stimulates locomotion and exploratory behavior toward novel and unexpected objects and exerts a variety of positive effects on learning, such as an increase in food-rewarded operant responding, a reduction in the delay to learn a task when rules are changed during learning, and alleviation of forgetting after a 4-week posttraining delay. At low and moderate doses, idazoxan and yohimbine dose dependently potentiate the amphetamine-induced hyperlocomotion. At high doses, they inhibit locomotion, exploratory behavior, and learning and increase neophobia, suggesting that excessive stimulation of LC neurons may have negative effects. Indeed, exploratory behavior, measured by the time an animal spends investigating objects in a novel environment, has been shown to be sensitive to prior exposure to stressors. At high doses, idazoxan and yohimbine elicit stresslike decreases in exploratory behavior. The α2-adrenergic agonist clonidine antagonizes the restraint-induced decrease in exploratory behavior and the reduction in exploratory head dipping and locomotor activity produced by high doses of yohimbine. Clonidine totally eliminates the preference for novel objects and reduces amphetamine-induced hyperlocomotion.

Feeding behavior is stimulated by low levels of clonidine and decreased by further production of NE. The anorectic action of amphetamine is reversed by centrally administered clonidine.

Clonidine and dexmedetomidine (a highly specific α2-adrenergic agonist) treatments also cause sedation mainly by increasing the time spent in non-REM sleep. These hypnotic effects are not mediated solely by changes in noradrenergic transmission, but are strongly associated with a decrease in serotonergic neurotransmission and diminished by stimulation of 5-HT2 receptors. Similarly, yohimbine produces a dose-dependent anticonflict effect by increasing

serotonergic transmission. Noradrenergic effects on sedation and aggressive behavior seem to be mediated via α2-adrenergic heteroreceptors located on serotonergic neurons.

Finally, phentolamine, a non-subtype-specific α-adrenergic antagonist, inhibits NE facilitation of substantia nigral self-stimulation behavior, but propranolol, a β-adrenergic antagonist, does not.

2. α1-Adrenoreceptors

Intraventricular administration of terazosin, a nonspecific α1-adrenergic antagonist, produces a dose-dependent suppression of locomotor activity, whereas β1- and β2-adrenergic antagonists have a smaller or no inhibitory effect. This activity-reducing effect is not due to sedation or motor coordination impairment and may have a motivational or sensory gating component. On the contrary, administration of NPY or phenylephrine, an α1-adrenergic agonist, into the prefrontal cortex increases locomotion and exploratory behavior. NPY effects are blocked by prazosin, an α1-adrenergic antagonist analog to terazosin, suggesting that NE modulation of locomotor and exploratory behaviors involves the α1-adrenoreceptor located in the prefrontal cortex. Alternatively, electrophysiological studies have established that α1-adrenergic receptors in the dorsal raphe nucleus exert a tonic excitatory action on serotonergic neurons, which may also be implicated.

Prazosin and corynanthine, another α1-adrenergic antagonist, dose dependently attenuate amphetamine-induced hyperlocomotion. In contrast, the selective α1-adrenergic agonist, ST 587, potentiates the locomotor effect of amphetamine. It was found that blockade of α1-adrenergic receptors in the prefrontal cortex inhibits not only the amphetamine-induced locomotor hyperactivity but also, as measured by microdialysis, the functional release of dopamine in subcortical structures. Such a finding is one piece of evidence of the existence of functional coupling between cortical noradrenergic and subcortical dopaminergic systems.

Intraventricular terazosin, an α1-adrenergic antagonist, also produces hypothermia and a reduced respiratory rate suggestive of reduced sympathetic outflow.

3. β-Adrenoreceptors

Evidence indicates that NE modulates memory storage through the activation of β-adrenoreceptors in the amygdala and hippocampus. In newborn rats, propranolol, a β-adrenergic antagonist, injected immediately following training in a classical conditioning test impairs memory in a dose-dependent manner. Isoproterenol, β-adrenergic agonist, injected immediately after training also impairs memory performance. In adults, LC stimulation enhances memory retention in a food-rewarded maze task. This enhancement is blocked by propranolol. β-Adrenergic antagonists increase working memory errors induced by NMDA receptor antagonist administration in the hippocampus, whereas α-adrenergic antagonist administration in the hippocampus has no effect on working memory. However, the enhancing effect on memory retention of amygdala β-adrenergic activation by clenbuterol is attenuated by coadministration of prazosin, an α1-adrenergic antagonist.

β-Adrenergic receptors also participate in maternal behavior. Activation of β-adrenergic receptors in the olfactory bulb of the sheep is required to establish olfactory recognition of the lamb. Propranolol administration in the olfactory bulb before parturition prevents ewes from recognizing their lambs. Propranolol treatment increases maternal onset latencies in new mother rats with brief maternal experience, whereas isoproterenol reduces these latencies.

Like α-adrenergic receptors, β-adrenergic receptors are implicated in novelty-elicited behaviors. Propranolol totally eliminates the preference for holes with novel objects while having no effect on total time spent investigating the holes. Propranolol administration in the basolateral nucleus of the amygdala reduces neophobia for food in animals highly responsive to novelty.

Systemic propranolol prevents restraint-induced changes in behavior in mice. Intraventricular isoproterenol administration produces an increase in defensive withdrawal, which is inhibited by propranolol. Finally, blockade of β1-adrenergic receptor in the CNS by betaxolol mimicks the inhibitory effect of stress on swimming and locomotion and increases grooming behavior.

B. LC Stimulation

Electrical stimulation of the LC produces a series of behavioral responses similar to those observed in naturally occurring or experimentally induced fear. Such behaviors are also elicited by the administration of α2-adrenergic antagonists that stimulate LC firing via their action on autoreceptors.

LC activation can also produce potent antinociception by reducing the response of neurons of the dorsal horn of the spinal cord through the stimulation of α2-adrenergic receptors.

C. Lesions of NE Neurons

Lesions of NE neurons can be achieved by different methods, including excitotoxic lesion of the locus ceruleus, local administration of 6-hydroxydopamine in the main LC projection pathway, or systemic administration of DSP-4. Although easy to use because of the possibility of a systemic injection, DSP-4 is not a suitable tool because it only induces a 70% loss in tissue NE and makes interpretations difficult when no behavioral deficit is observed.

In the rat, nearly complete depletion of NE (>99%) in the forebrain can be achieved by injecting 6-hydroxydopamine, a neurotoxin selective of the catecholaminergic neurons locally into the trajectory of the LC axons, i.e., the dorsal noradrenergic bundle (DNAB). Such a lesion has no effect on eating, drinking, or spontaneous locomotor behavior but affects responses to novelty and the acquisition of conditioned behavior. DNAB lesions impair appetitive and aversive conditioning whenever the rat has to learn a behavior in response to the discriminative stimulus, but only when the lesion occurred before the acquisition of the task. Tasks in which a simple behavioral response is required, such as conditioned taste aversion, are resistant to DNAB lesion. DNAB lesion also spares behavioral tasks in which the animal has to refer to distal cues, as in contextual aversive conditioning or spatial learning in a Morris maze. Taken together, these results suggest that DNAB-lesioned animals are only impaired in tasks requiring a high level of arousal and focused attention to local cues. Additionally, DNAB lesions have been shown to exert anxiolytic effects in anxiogenic behavioral test situations.

D. Knockout and Transgenic Mice

Several genes coding for adrenergic receptors or proteins implicated in the synthesis or metabolism of NE have been deleted from mice with the knockout technique. Most of the initial studies have focused on the cardiovascular functions of these knockout mice, and only few behavioral experiments performed on these animals have yet been described.

1. Dopamine β-Hydroxylase (DBH) Knockout Mice

Most DBH−/− mice die *in utero*. However, survival can be enhanced by perinatal administration of dihydroxyphenylserine (DOPS), a direct precursor of NE (see Fig. 3). DBH−/− mice have impaired adaptation to cold and have elevated basal metabolic rates without abnormalities in thyroid hormone levels. A deficit in maternal behavior in DBH−/− females may also exist. The results of cross-fostering between DBH−/− females and DBH+/− females suggest that an important interaction mediated by NE occurs between the dam and the neonate during the first 24 hr for establishing maternal behavior.

The results of behavioral tests indicate that NE plays an important role in motor learning and performance and in the retention of several behaviors. Restoration of NE with DOPS eliminates the motor deficits and improves fertility in males, indicating that these differences are due to the physiological absence of NE rather than lasting developmental defects secondary to NE deficiency. More recently, it was found that DBH−/− mice exhibit altered ethanol-induced behavioral and physiological responses. DBH−/− mice exhibit a reduced ethanol preference in a two-bottle choice paradigm and delayed extinction in the ethanol-conditioned test aversion. They are hypersensitive to the sedative and hypothermic effects of systemic alcohol administration. Interestingly, NPY−/− mice show increased consumption and less sensitivity to the hypnotic effects of ethanol, whereas NPY-overexpressing mice decrease their consumption of ethanol. This suggests that NPY, which, as previously mentioned, is colocalized with NE in LC cells (see Section II.B), counterbalances NE effect.

2. NE Transporter Knockout Mice

Although such animals have been already obtained and seem to survive at normal Mendelian ratios, data concerning their phenotypes are scarce. There is no doubt that mice lacking the NE transporter will be an interesting tool to use to approach the different roles of NE, especially if one considers that NET is a major target site for most antidepressants and psychostimulants. Recent data indeed indicate that mice lacking NET are hypersensitive to behavioral effects of psychostimulants.

3. α2-Adrenergic Receptor Transgenic Mice

Mice with the deletion of α2A-, α2B-, or α2C-adrenergic receptor genes, as well as point mutation

of the α2A gene and a 3-fold overexpression of the α2C gene, have been generated and facilitated the assignment of different functions of α2-adrenergic receptors to specific subtypes because subtype-specific agonists and antagonists were lacking. Studies with these mice indicate that most of the classical physiological functions mediated by the α2-adrenergic receptors, such as hypotension, sedation, analgesia, and hypothermia, are mediated by the α2A subtype. The α2B subtype, the principal mediator of the hypertensive response to α2-agonists, appears to play a role in salt-induced hypertension and may be important in developmental processes.

The α2C subtype appears to be involved in many central nervous system processes, such as the startle reflex, stress response, and locomotion. In particular, α2C-knockout mice show an increased locomotor response to amphetamine, whereas mice overexpressing the α2C gene show decreased activity in response to the drug. In an isolation-induced aggression paradigm, α2C knockout mice present decreased attack latency, whereas mice overexpressing the α2C gene increase attack latency. α2C overexpression increases and lack of α2C-adrenergic receptor decreases the immobility of mice in the forced swimming test, i.e., α2C-adrenergic receptor expression appears to promote the development of behavioral despair. α2C-adrenergic receptor is also involved in spatial learning. In the Morris water maze, α2C-adrenoreceptor overexpressing mice develop an ineffective thigmotaxic search pattern characterized by swimming close to the pool walls during both spatial and nonspatial water maze training, whereas their swimming pattern is normal when no cognitive component is required. In the T-maze delayed alternation task, α2C-knockout mice present lower performance than wild-type mice due to a higher number of perseverative errors.

Both α2A and α2C subtypes are important in the presynaptic inhibition of NE release and appear to have distinct regulatory roles. However, simultaneous disruption of both α2A and α2C genes leads to far more severe deficits than the single disruption of either of these genes.

4. α1-Adrenergic Receptor Knockout Mice

Only data on α1B-adrenergic receptor knockout mice were published at that time. The mutants have no apparent phenotype changes except a decreased phenylephrine-induced blood pressure response. However, interestingly, experiments have indicated that these knockout mice exhibit locomotor responses to psychostimulants, such as cocaine or amphetamine, that are dramatically decreased when compared to those obtained with wild-type congeners. Because it has been claimed that psychostimulants exert their behavioral effects through an increase in subcortical dopaminergic transmission, such data suggest, as previously mentioned, the presence of coupling between noradrenergic and subcortical dopaminergic systems. In other words, it cannot be excluded that psychostimulants exert some of their behavioral effects through the noradrenergic stimulation of α1B-adrenergic receptors.

5. β-Adrenergic Receptor Knockout Mice

Ninety percent of β1-adrenoreceptor (−/−) mice die *in utero* between days 10.5 and 18.5 of gestation. Surviving mice appear normal and are fertile. β2-adrenoreceptor(−/−) knockout mice are normal and fertile. Physiological responses (blood pressure and heart rate) of the mice to β-adrenergic agonists and antagonists were tested and confirmed their absence of effects. Behavioral experiments have not been reported yet.

VII. CLINICAL IMPLICATIONS OF CENTRAL NORADRENERGIC SYSTEMS

Several agonists and antagonists of noradrenergic receptors are currently used in the pharmacopea to interact with different diseases related to NE in the peripheral nervous system. For example, inhalation of β2-adrenergic receptor-selective compounds has long been established as an effective therapy for asthma and other bronchospastic condition, and β-blockers are used in the treatment of angina pectoris and cardiac arrythmias. They are both used as a treatment for acute congestive heart failure or for long-term management of patients who survive myocardial infarction. In addition, β-adrenoreceptor antagonists have been used as effective antihypertensive drugs for several decades. α-Adrenoreceptor ligands, such as α1-adrenoreceptor antagonists and α2-agonists, are also widely employed as antihypertensive agents.

It must be recalled, however, that brain noradrenergic neurons are protected from peripheral NE by the blood–brain barrier and that several peripheral noradrenergic agents have been developed in such a way that they do not pass through this barrier. Consequences of central or peripheral noradrenergic

system dysfunctions therefore have to be considered separately. For instance, because of the widespread projections of the central noradrenergic system and the numerous functions NE plays in the brain, it can be anticipated that any dysfunction of central noradrenergic neurons may induce a series of different neurological and psychiatric diseases.

A. Parkinson's Disease

Besides the cardinal signs of tremor, bradykinesia, rigidity, and postural instability, patients suffering from Parkinson's disease (PD) may also present many other neurological symptoms and signs, including cognitive, sensory, and autonomic disturbances. The implication of NE in PD is based on the observation that PD is associated with the degeneration of not only dopaminergic neurons located in the substantia nigra but also NE–LC cells. This LC alteration results in a drastic diminution of NE levels in brain areas such as the prefrontal cortex and also in the peripheral system. As for dopaminergic neurons, noradrenergic neurons contain neuromelanin, a pigment that results from oxidative catabolism of catecholamines and that may sensitize noradrenergic cells to oxidative stress. It should be noted, however, that noradrenergic medullary nuclei remain unchanged.

The role of NE in PD is not as clear as it is for dopamine. One possibility is that the depletion of noradrenergic cells is responsible for secondary symptoms of PD such as depression, dementia, or deregulation of endocrine function. This hypothesis is supported by the fact that NE loss is significantly higher for PD patients that also suffer from dementia. Another possibility has been raised by experiments on MPTP-treated monkeys. MPTP is a neurotoxin whose injection results in the degeneration of dopaminergic cells and the development of Parkinson-like symptoms, which may be followed by a partial recovery. Bilateral LC lesions, which do not induce any particular motor deficits in normal monkeys, strongly aggravate the symptoms in MPTP-treated monkeys and delay their recovery.

B. Psychiatry

Studies about the role of NE in psychiatric disorders often rely on the measurement of plasma or urinary levels of NE and its neuronal metabolites (MHPG and DHPG), which represents the most direct method available to measure noradrenergic activity in living human subjects. However, the variations in these levels more probably correspond to NE released by the sympathetic system rather than to NE released in the brain. Alternative methods have been considered (MAO activity of platelets, platelet density of α2-adrenergic receptors, growth hormone release in response to clonidine), but results are still contaminated by noradrenergic activity at the periphery. The limitation of such approaches has always led to debate. Central noradrenergic activity can easily be measured in animals, but animal models of psychiatric disorders are subject to controversy. The most convincing arguments often arise from the pharmacological properties of psychotropic drugs and, conversely, the psychotropic effects of noradrenergic agents.

1. Anxiety: Panic Disorders and Posttraumatic Stress Disorder (PTSD)

Anxiety is phenomenologically similar to states of fear, with the difference that fear is related to real threat, whereas anxiety is an excessive response when little or no real danger is present. There is considerable evidence for a relationship between noradrenergic brain systems and behaviors associated with stress and anxiety. In animals, behaviors that are characteristically observed in situations of stress and fear are associated with an increase in activation of LC–NE systems. Similarly, in healthy human subjects, many studies found significant correlations between states of anxiety and plasma or urinary levels of MHPG. Conversely, stimulating NE transmission induces fear reactions. Infusion of NE into the hypothalamus of cats results in defensive–aggressive behaviors such as hissing, growling, and ear retraction. Stimulation of LC in monkeys induces behavior seen in the wild when the animal is threatened. Panic attacks are induced by yohimbine, an α2-adrenergic antagonist that increases LC firing, in approximately 60–70% of patients with panic disorders.

In animals, chronic stress is associated with a sensitization of both behavioral and NE responses after reexposure to a subsequent stress. This stress sensitization might be relevant to the neurobiology of disorders such as panic disorders or PTSD, in which patients have had a history of previous exposure to stress. This suggests that panic disorders and PTSD may result from abnormally high LC activity. Indeed, some studies have demonstrated increased urinary NE and blunted growth hormone response to clonidine in

both patients with panic disorders and those with PTSD. Studies on other anxiety disorders, such as generalized anxiety disorder and obsessive–compulsive disorder, do not support an important role for NE brain systems.

Finally, the therapeutic efficacy of tricyclic drugs or monoamine oxidase inhibitors, usually considered as antidepressants, may result from their actions on the noradrenergic system. Benzodiazepines that are highly efficient in reducing panic disorders also reduce LC firing.

2. Depression

In 1965, Joseph Schildkraut initially proposed the hypothesis that noradrenergic systems would play a role in depression and be the major site of action of antidepressant drugs. Although more recent data clearly indicate that other neurotransmissions, such as serotonergic and dopaminergic ones, are also implicated, the role of noradrenergic systems nevertheless remains prominent. For example, in bipolar patients, urinary MHPG levels are lower during the depressed phase and higher during the manic phase than during periods of euthymia. Furthermore, unipolar and bipolar depressive patients demonstrate greater increases in plasma NE after moving to an upright position than do controls. Finally, post mortem studies have reported higher β-receptor densities in the brains of suicide victims than in controls.

Since the demonstration that most antidepressants decrease β1-adrenergic receptor transduction in the cerebral cortex of the rat, this parameter has generally been considered as a biochemical correlate of therapeutic activity. This observation is noteworthy not only because it seems to be a common consequence of chronic treatments with most antidepressants whatever their mechanism but also because its development parallels clinical improvements. Indeed, in animal studies, down-regulation of βl-adrenergic receptors only appears following 10–20 days of chronic treatment, a delay that corresponds to clinical observations. Not all antidepressants, however, induce desensitization of β-adrenergic receptors. A reactivation of serotonin transmission, which can be obtained by some antidepressants such as SSRIs (specific serotonin re-uptake inhibitors), can hamper the development of β-adrenergic receptor desensitization even when noradrenergic transmission is also reactivated. The lack of a response by β-adrenergic receptors has been proposed to be due to past receptor events at the level of phosphorylations mediated by PKA and PKC.

β-Adrenergic receptors are not the only receptors desensitized by antidepressants. A serotonin receptor subtype, 5-HT2, is also frequently affected. This should not imply, however, that an up-regulation of β1-adrenergic or 5-HT2 receptors is responsible for the disease and that, consequently, down-regulation of these receptors is the goal to achieve. These observations rather suggest that both types of receptors, β1-adrenergic and 5-HT2, are very sensitive to any modification of their respective neurotransmissions and that an activation of the latter occurs following chronic treatment with antidepressants. According to that view, a deactivation of serotonergic and/or noradrenergic neurons is probably the main biochemical characteristic of depression.

It can be added that tricyclic antidepressants are generally considered to have better clinical therapeutic efficacy than SSRIs in major depression. Both groups share the property of inhibiting the re-uptake of NE and serotonin, but tricyclic antidepressants are, in contrast to SSRIs, potent cholinergic and αl-adrenergic receptor antagonists. Although α1-adrenergic antagonism may represent an adverse side effect because of its hypotensive action, it cannot be excluded that the blockade of central α1-adrenergic receptors at least partly explains the better clinical efficacy of tricyclics.

3. Schizophrenia

Since the early 1970s, the dopamine hypothesis has been at the forefront of explanations for the pathogenesis of schizophrenia. This was founded on the fact that classical neuroleptics are potent dopaminergic antagonists. However, estimation of NE and its metabolite MHPG post mortem and in cerebrospinal fluid has produced more consistent findings than similar studies on dopaminergic systems.

Moreover, psychophysiological abnormalities observed in schizophrenic patients, such as dysfunction in smooth pursuit eye movement (smooth pursuit being replaced by stepwise pursuit or spiky pursuit) or in skin conductance responses to novel sensory stimuli (absence of response or failure of its habituation), which may reflect under- or overarousal, are also observed as a consequence of noradrenergic system sub- or supersensitivity.

The interest in noradrenergic systems was renewed with the apparition of atypical antipsychotic agents such as clozapine. Clozapine is one of the first agents demonstrating superior efficacy compared with classical neuroleptics (particularly on negative symptoms)

and, perhaps because it has a relatively low affinity for dopaminergic receptors of the D2 subtype, is devoid of extrapyramidal side effects. It has also been found that chronic clozapine treatment increases plasma DHPG levels in a manner that is correlated with clinical improvement. This study is in agreement with experiments in rats showing that chronic clozapine increases the firing rate of LC neurons. Like other atypical neuroleptics, clozapine is a potent αl-adrenergic receptor antagonist. If we assume, as previously mentioned, that the coupling between noradrenergic and dopaminergic systems is due to the stimulation of α1-adrenergic receptors, it may explain that blockade of the latter can moderate the increased dopaminergic subcortical activity generally considered as one of the main biochemical features of schizophrenia. Finally, clonidine, an α2-adrenergic agonist that decreases LC neuron firing, shows a strong therapeutic effect on positive symptoms in acute treatment but is not used because tolerance develops rapidly with chronic medication and provokes a rebound exacerbation of the symptoms after withdrawal.

4. Attention Deficit Hyperactivity Disorder

Attention deficit hyperactivity disorder (ADHD) is a childhood psychiatric disorder characterized by inattention, impulsivity, and overactivity. Children with ADHD have difficulty completing tasks, they make careless mistakes, do not listen, lose things, and avoid tasks that require concentration, and they are forgetful and disorganized. The behavioral deficits ADHD are estimated to be present in 3–5% of all school-aged children and are believed to arise in early childhood. The neurotransmitter systems most commonly implicated in the pathophysiology of ADHD are the catecholamines, dopamine and NE. Virtually all medications that are effective in the treatment of ADHD, in particular psychostimulants such as methylphenidate, affect catecholamine transmission, and medications that do not interact with catecholamine transmission are rarely effective in the treatment of ADHD, thus indicating the possibility of a biological etiology. A relatively early model has postulated that a noradrenergic dysfunction in the LC produces the deficits in vigilance and sustained attention observed in children with ADHD. However, although studies employing peripheral measures to assess catecholaminergic function in ADHD are plentiful, they are highly inconsistent in their findings. Urinary MHPG levels, for example, were either higher, not changed, or lower in children with ADHD than in normal controls.

The prefrontal cortex and the basal ganglia play prominent roles in a complex neural system that serves to regulate motor function and behavior via working memory. There are different indications suggesting that working memory in the prefrontal cortex is under the control of the stimulation of D1-dopaminergic receptor subtype, whereas activation of cortical α1-adrenergic receptors inhibits the effects of D1-receptor stimulation. Complex interactions between prefrontocortical Dl, D2, and α1-adrenergic receptors have been demonstrated, and it cannot be excluded that ADHD corresponds to some dysfunction of these interactions, which, in turn, are responsible for the expression of subcortical functions.

5. Pharmacodependence

Although multiple lines of research have implicated the mesolimbic dopaminergic system in drug reward, the role of noradrenergic systems in pharmacodependence processes should not, however, be overlooked. Indeed, virtually all classes of abused drugs affect LC discharge characteristics at doses that are in the range of those abused by humans. These include hallucinogens, psychostimulants, opiates, alcohol, nicotine, and benzodiazepines. Much of the work on the LC concerning substance abuse has focused on physical dependence and withdrawal symptoms. This is due to the fact that α2-adrenergic agonists such as clonidine, which suppresses LC activity, partly alleviate withdrawal symptoms for several dependence-producing substances. Indeed, the cessation of drug use in chronic opiate abusers produces a severe withdrawal syndrome that is highly aversive. Although increased NE in the brain has long been implicated in opiate withdrawal, it was not clear whether noradrenergic systems were involved until studies, performed in the rat, were completed to indicate that the noradrenergic inputs to the bed nucleus of the stria terminalis arising from noradrenergic cell groups of the caudal medulla are critically involved in the aversiveness of opiate withdrawal.

As mentioned earlier, LC receives a prominent enkephalinergic input associated with a high density of opioid receptors in that region. Opioid antagonists have no effect on LC activity by themselves, but produce a dramatic long-lasting excitation in rats that have chronically received opiates. It is this effect that has been the basis for rationalizing the use of clonidine, which inhibits LC discharge, in the treatment of opiate withdrawal.

The case of nicotine is also interesting because it produces a potent activation of LC neurons when administered systemically. Surprisingly, evidence indicates that this effect of nicotine is not mediated in the LC, or even initiated in the brain, but results from nicotine activation of primary sensory C-fiber afferents. In addition to increasing LC discharge, nicotine induces an increase in the frequency of burst activity, suggesting that the net effect of nicotine is to elicit more NE release in targets and produce short-lasting periods of enhanced arousal.

Dopaminergic systems are, however, more generally considered as the main targets of drugs of abuse than noradrenergic ones because it is believed that psychostimulants, such as amphetamine and cocaine, or opiates, such as morphine and heroin, cause addiction in humans and induce locomotor hyperactivity in rodents through increased release of dopamine in a subcortical structure, the nucleus accumbens. Nevertheless, as noted earlier, experiments performed on rats have indicated that prazosin, an α1-adrenergic antagonist, could hamper the locomotor hyperactivity induced by D-amphetamine. Similarly, we have previously mentioned that mice lacking the α1B-adrenergic receptor were considerably less sensitive to the locomotor effects of amphetamine and cocaine than their corresponding controls. This suggests that, in addition to its role in drug withdrawal, NE, via the stimulation of α1B-adrenergic receptors, may enhance the release of subcortical dopamine and, therefore, amplify the rewarding effects of drugs of abuse. Such a role for NE may provide some new insight into the problem concerning the great variability in the sensitivity to drugs of abuse observed in humans and animals. Indeed, because LC cells are extremely sensitive to environmental stimuli, small genetic or epigenetic variations in the reactivity of noradrenergic neurons to environmental cues may affect the activation of mesencephalic dopaminergic neurons and, more generally, the sensitivity to drugs of abuse.

VIII. CONCLUSION

Several theories have been proposed to describe LC functions, ranging from notions of reinforcement and arousal to the mediation of anxiety or the control of selective attention. Because there is a good chance that most of these theories, although different, originate from the same observations, it seems wise to restrict this conclusion to a brief description of experimental data that have reached general agreement. On an anatomical point, LC cells send widely ramifying axons, which innervate diffuse forebrain structures such as the cerebral cortex and the hippocampus. The electrophysiological data and also neurochemical indices indicate that LC neurons stop firing during REM, are especially active during high states of arousal, and exhibit phasic responses to novel or intense stimuli. Finally, the release of NE in terminal regions enhances the signal-to-noise ratios of event-related action potentials.

From these data, it follows that LC neurons improve the processing of external events, whether these events are novel or salient because of previous conditioning, making them even more salient. Although this article was devoted to NE in the central nervous system, it must be noted that interactions with the periphery do exist, and, as proposed by Gary Aston-Jones, whereas the "activation of the peripheral sympathetics system prepares the animal physically for adaptive phasic responses to urgent stimuli, … parallel activation of LC increases attention and vigilance, preparing the animal cognitively for adaptive responses to such stimuli." Indeed, LC activity participates in the mechanism, which effectively focuses attention on salient stimuli in threatening or demanding situations. It can be considered, therefore, that LC cells do not mediate anxiety or stress per se, but rather a state of arousal leading to attentional and cognitive change. Any over- or underactivation of these LC cells will induce exaggerated or attenuated responses, leading to diseases such as panic attacks or depression. Finally, it must be recalled that noradrenergic systems interact with other neuromodulatory neurons, especially serotonergic and dopaminergic, and that there is little doubt that malfunction of the former will have consequences on the functions controlled by the latter.

See Also the Following Articles

ANXIETY • ATTENTION • CATECHOLAMINES • DEPRESSION • DOPAMINE • EVENT-RELATED ELECTROMAGNETIC RESPONSES • GABA • MANIC–DEPRESSIVE ILLNESS • NEUROTRANSMITTERS • PARKINSON'S DISEASE • SCHIZOPHRENIA • VIGILANCE

Suggested Reading

Aston-Jones, G., Chiang, C., and Alexinsky, T. (1991). Discharge of noradrenergic locus ceruleus neurons in behaving rats and monkeys suggests a role in vigilance. In *Progress in Brain*

Research (C. D. Barnes and O. Pomeiano, Eds.), pp. 501–519. Elsevier, Amsterdam.

Bremner, J. D., Krystal, J. H., Southwick, S. M., and Charney, D. S. (1996). Noradrenergic mechanisms in stress and anxiety: I. Preclinical studies and II. Clinical studies. *Synapse* **23,** 28–51.

Goldstein, D. S., Eisenhofer, G., and McCarty, R. (1999). Catecholamines: Bridging basic science with clinical medicine. *Adv. Pharmacol*, Vol. 42.

Rasmussen, K., Morilak, D. A., and Jacobs, B. L. (1986). Single unit activity of locus ceruleus neurons in the freely moving cat. I. During naturalistic behaviors and in response to simple and complex stimuli. *Brain Res*. **371,** 324–334.

Usher, M., Cohen, J. D., Servan-Schreiber, D., Rajkowski, J., and Aston-Jones, G. (1999). The role of locus ceruleus in the regulation of cognitive performance. *Science* **283,** 549–554.

Number Processing and Arithmetic

KAREN WYNN
Yale University

GLOSSARY

acalculia Severe deficit in arithmetical calculation abilities, often resulting from a brain lesion or degenerative neurological disorder.

cardinality The total number of items in a set.

counting Determining the number of items in a set by engaging in a serial process that satisfies the counting principles.

enumeration Determining the number of items in a set by any means including, but not limited to, counting.

individuation To determine the boundaries of a specific individual and maintain a representation of that entity as a distinct unit.

subitization A process of rapidly determining how many items are in a display; applies over small numbers only.

The formal system of mathematics is arguably one of the greatest and most beautiful of human achievements. Mathematics captures and describes many aspects of the physical world, yet it is an abstract logical system that is unconstrained by physical reality. How is it that the human mind is able to grasp an understanding of such abstract knowledge? How are we able to determine the numbers of things in our world? How does the human mind represent number concepts, and how does it operate on these representations to perform basic numerical computations, including arithmetic? These questions form the central focus of this article. Empirical findings suggest that some understanding of number is part of the inherent structure of the human mind. Evidence from comparative psychology and cognitive neuroscience suggests that we may be born with an evolved mental mechanism for number, which we share with a number, of animal species; specific brain structures are responsible for our sense of magnitude and our ability to engage in basic numerical thought.

I. SOME ABILITIES OF NORMAL HUMAN ADULTS

A. Linguistic Counting

One obvious means we have for determining the numerosity of a set is to purposefully count the members of a set, to apply our list of number words to items in the set. Psychologists Rochel Gelman and Randy Gallistel, in their classic book *The Child's Understanding of Number*, outlined the functional principles required of any and all counting procedures. The Stable-Order Principle specifies that the number labels must be used in a consistent order across all counts. The One-to-One Correspondence Principle specifies that one and only one number label must be applied to each item to be counted, and the Cardinality Principle specifies that the number label applied to the final item in a count serves to represent the total number, or cardinality, of the set of items counted. Furthermore, the Order-Irrelevance Principle states that the same answer will be obtained regardless of the order in which the items in the set are counted. Finally, the Abstraction Principle states that any items can be

counted; items need not be homogeneous to be grouped into a single count.

To engage in this procedure clearly requires having learned both the counting routine and the number words of one's language. But there is evidence that our ability to determine number is not entirely language-dependent. Human adults possess other cognitive processes that will also yield numerical representations.

B. Subitization and Estimation

1. Subitization

It has been known for a long time that human adults can identify small numbers of objects precisely without having to consciously count them. If someone places a small handful of up to three or four coins onto a table, one would have the experience of immediately being able to tell how many there were, without counting. This ability is known as subitization and has been studied experimentally since the 1940s. There are two experimental findings that typify the subitization phenomenon: (1) If varying numbers of items are visually presented very briefly (say, 200 msec) to adult subjects whose task is to indicate the number of the items, subjects tend to show virtually perfect performance for numbers up to three or four items, with performance error increasing linearly with numbers beyond this range. (2) If visual displays containing varying numbers of items are presented to subjects whose task is to respond *as accurately as possible* with the number of items, subjects show very little increase in reaction time as the number of items in the display increases from zero to three or four, and then show a much steeper linear increase with additional items (see Fig. 1). Note that the first measure in essence holds reaction time constant (by presenting all stimuli equally briefly) and measures the effect of number on accuracy, whereas the second measure holds accuracy constant (by asking subjects to be as accurate as possible) and measures the effect of number on reaction time. Thus, there is a trade-off between accuracy and reaction time for numbers above the subitization range, typically at about four items and higher, but little trade-off between the two for numbers within the subitizing range.

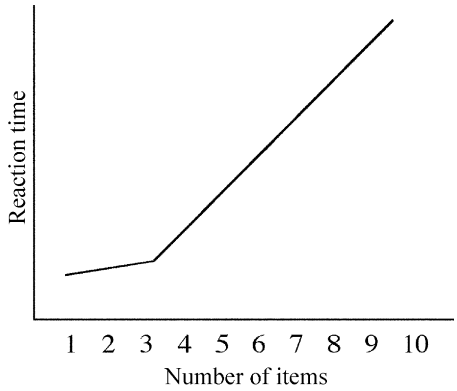


Figure 1 Typical reaction-time function for subitization.

2. Estimation

Adults' ability to determine number without conscious counting is not limited to small set sizes. We can estimate the number of items in a display containing many elements, for example, the number of black dots on a page filled with dots. In such situations, whereas the exact number estimated may be quite off the actual number, adults quite reliably give much higher estimates when presented with, say, 2000 dots than when presented with, say, 500 dots. Moreover, adults are good at representing both the absolute and relative frequencies of many different kinds of items: visual entities, sounds, events, number of words in a list that begin with the letter r, and so on. This is true even when subjects have no knowledge, when presented with the stimuli, that the task will involve subsequently giving a numerical judgment.

3. Theories of the Cognitive Processes Underlying Subitization and Estimation

What is the nature of each of these abilities to determine and represent number? There have been different proposals as to the underlying nature of the subitization process. One early proposal was that subitization is a visual pattern-recognition process, in which subjects are not actually determining the number of items present in a display but are recognizing distinct patterns that typify displays of a given number. For example, two points define a line and three points in random configuration typically define a triangle. Subjects could very quickly recognize such patterns and simply associate each pattern with its corresponding numerosity. This would explain the relatively flat reaction time slope for numbers in the one to three range; the different patterns (point, line, triangle) associated with these numbers are all recognizable in about the same amount of time. However, as

the number of items increases to four or more, the number of possible configurations for a given number increases exponentially, and the variety of configurations for displays of different numbers begin to overlap each other much more heavily, so that pattern recognition would break down at about this number, precisely where the subitization reaction-time data show a sharp "elbow." In this proposal, subitization (which ends at about three items) and estimation (which begins at around four and more items) are inherently distinct processes. However, this theory has largely fallen out of favor, primarily because data show that subitization effects are obtained even when subjects could not be identifying each number with a unique perceptual pattern (for example, when all points are arranged collinearly or when objects are not simple dots but complex household items each with a unique contour, making the overall contour of even a one- or two-item display complex and variable rather than linear).

A very different possibility is that subitization is a rapid and automatic serial enumeration process: subjects are determining the number of items per se in the displays through some form of unconscious counting. One specific model of such a counting process is the accumulator mechanism. This model will be described in greater detail later. Briefly, in the accumulator model, numerosities are represented by magnitudes, so that the counting process is akin to filling a bucket with a burst of water, one burst for each item to be counted. The final fullness level of the bucket represents the total number of the items counted. Evidence favoring the view that subitization is some form of serial counting process is that the reaction time curve is not flat for the numbers one to three but shows a shallow slope even across this range indicating some "cost" to each additional item in terms of processing time—a hallmark of serial processing.

How does this theory account for the subitization "elbow" in reaction times and for the decrease in accuracy with larger numbers? In this theory, there is variability in the counting process, an error term that is additive with each item counted (the bursts of water are not all identical in amount). Thus, with larger counts there will be more variance in the final fullness level. This means that larger numbers will be more difficult to discriminate than smaller numbers, leading to a decrease in accuracy as number increases. Whereas the mean fullness levels for, say, five and six will be as far apart as the mean fullness levels for two and three, the normal distributions of the fullness levels for the former will overlap more than those for the latter. In order for subjects to maintain near-perfect accuracy in subitization tasks for numbers large enough that their fullness values are not reliably discriminable, subjects will have to resort to other measures, such as consciously counting the items. When consciously counting, one must serially attend to each item in the display, such that there is a constant increase in processing time required for each added item to be counted, thus explaining the linear increase in reaction time (RT) for numbers four and higher. In this proposal, both subitization and estimation reflect the *same* enumeration process at work: because of the additive nature of the variance, this automatic enumeration process enables highly reliable discrimination and fast reaction time for very small numbers of items (i.e., subitization) but enables only approximate, "ballpark" identification of larger numbers (estimation).

Further evidence suggests that there may be additional processes involved in subitization as well. Lana Trick and Zenon Pylyshyn have suggested that subitization may be based in part on certain encapsulated processes within visual cognition. Specifically, they propose that the visual system comes equipped with a small number (three or four) of pointers or "fingers," which pick out selected objects in the visual field. Each pointer is assigned to an object the visual system wishes to track and indexes the current location of the object, updating the location as required (as the object moves, for example, or over brief occlusions and successive saccades). These fingers are assigned in parallel during the early preattentive stages of visual processing. In order to determine the numerosity of a group of items, one must first individuate the items, keeping track of the items to be counted and mentally keeping separate the already-counted items from the to-be-counted items. After individuating each item, one must operate over each item serially, for example, assigning a number word (if one is engaged in verbal counting) for each item or incrementing an accumulator with one increment for each item (the accumulator model). If there are specialized processes within the visual field for rapidly and in parallel individuating three to four items in a display, then this will enable the serial enumeration operation to commence significantly earlier, considerably speeding up—and rendering more error-free—the process for displays containing at most three or four items. In displays containing larger numbers of items, some of the items will be individuated "for free" by these preattentive pointers within the visual system, but the remainder of the items will need to be individuated serially themselves,

lengthening the time course of the process and also increasing the likelihood of error.

Evidence for such a component within the visual system comes from numerous sources. First, when subjects are presented with displays in which serial attentional processes are required to individuate the items to be enumerated (for example, displays containing items with overlapping contours such as concentric circles or containing distractor items from which the target items do not "pop out" but must be searched for serially), the subitization function disappears: there is a linear increase in reaction time for each added number of items, even in going from one to two items. Moreover, neuropsychological data obtained by Stanislas Dehaene and Laurent Cohen showed a clear dissociation between enumeration within the subitizing range and enumeration beyond this range. Their subjects were brain-lesioned human adults who were severely simultagnosic, that is, who showed significant deficits in serial visual exploration but whose parallel preattentive visual processes were largely intact. By hypothesis, these patients should not be hindered in subitizing, if subitizing is dependent on parallel, preattentive visual processes. However, they should show severe deficits in their ability to engage in serial counting processes, those hypothesized to be required for numerosities of four or more. All patients showed an intact ability to accurately quantify smaller numbers, one, two, and sometimes three, but showed significantly impaired ability to enumerate larger sets of items, showing that subitization can be preserved even when counting is impaired.

C. Calculation

Humans are able to do much more than simply determine numbers of entities; we are able to determine, for example, whether two values are equal or whether one is greater or less than the other, and we can perform numerical computations, determining the sum, difference, or product of two numbers, for example. Some of these computations may be supported by an accumulator-type representation. One finding in studies of adults' calculation performance is that, when subjects have to state, as rapidly as possible, which of two presented numbers is larger, reaction time decreases as the proportionate difference between the two numbers increases; this is known as the Distance Effect. For example, subjects are much faster to respond that 5 is greater than 4 than they are to respond that 15 is greater than 14. But they respond about equally quickly to the comparison between 4 and 6 and that between 12 and 18. This effect fits well with predictions of the accumulator mechanism. In the accumulator, numbers with a smaller proportionate difference will have more confusable magnitude representations and, thus, comparison will be slowed, and when the proportionate difference becomes so small that the magnitudes cannot be reliably distinguished, other processes will have to be recruited to complete the comparison (for example, rapidly mentally counting up the number word list or mentally scanning a number line to determine which number comes first).

However, there are many aspects of our arithmetical competence that cannot be attributed to an accumulator-like magnitude representation of number. For example, our ability to compute the correct answers to precise arithmetical problems involving large numbers clearly cannot rest (at least not solely) on the representations generated by something like an accumulator mechanism, because such representations are much too approximate to support the precise results of our computations with numbers above 3 or 4. Human adults can quite rapidly, and with relatively low error rates, state the answer to the question, "What is the sum of 32 and 24?" These abilities clearly rest on additional cognitive structures and involve a different kind of numerical representation.

There is abundant evidence that numerical calculation is a specific ability unto itself, dissociable from language abilities, general memory function, and general reasoning ability. For example, neuropsychologist Brian Butterworth presented the case of Monsieur Van, who, due to the onset of Alzheimer's disease, had very poor general reasoning abilities, showing an inability to engage in even relatively basic problem-solving strategies. Moreover, he failed on Piaget's typical number-conservation task, in which two rows containing equal numbers of elements are arranged in one-to-one correspondence before the subject. Once the subject agrees that the two rows are equal in number, the experimenter spreads out one of the rows and asks if they still have the same number. Even 6-year-old children will typically answer correctly, unlike Monsieur Van. However, when presented with a wide range of calculation, estimation, and comparison problems, he performed quite well. Contrast this case with that of Butterworth's patient Signora Gaddi, who had very good general reasoning abilities but was unable to reason about or perform even the simplest computations on numbers higher

than four. Butterworth has further case studies of (1) amnesics able to perform normally on number computation tasks, even in the face of drastically impaired short-term memory spans, suggesting that special memory systems may exist for arithmetical facts, and (2) patients with certain kinds of brain damage, who, while retaining their normal long-term memories and short-term memory capacity, lose all of their stored number facts.

There is some evidence that the brain stores well-learned numerical facts (such as those in the addition and multiplication tables, for example) separately from procedures for performing numerical operations such as addition, multiplication, and division and even that distinct numerical operation procedures are stored separately from each other. That is, arithmetical calculation appears to be not just one, but rather several distinct (albeit related) systems. Signore Tiziano, a patient studied by Lisa Girelli and Margarete Delazer, suffered a stroke that affected his left parietal lobe; subsequently, he became unable to perform subtraction procedures correctly, while retaining his proficiency with other arithmetical operations. Or consider the patient of Alphonse Caramazza and Michael McCloskey, who became unable to follow the standard procedure for multiplication but could still perform long division correctly. Such patients, and others like them, were still able to retrieve memorized number facts (such as the addition and multiplication tables). Their problems were at the *procedural* level; they were unable to identify the appropriate sequence of operations and often unable to correctly relate the different numbers in the problem to each other. Conversely, patients like Dr. Constable exist, studied by Elizabeth Warrington, who show the opposite pattern of abilities: Dr. Constable, following a stroke, was unable to recall even basic number facts such as $5+7$ (he had to work out the problems, often by counting up or down) but was well able to instantiate the correct procedures and sequence of operations for carrying out all of the arithmetical operations. Moreover, he could quickly give very good *approximate* answers to questions (such as "$5+7$ is approximately 13") before working them out to obtain the precise answers.

II. A NEUROPSYCHOLOGICAL MODEL OF NUMBER PROCESSING

A. Description of the Model

The abilities, performance patterns, and dissociations described in Section I, together with other findings relating to how human adults represent and process number, led the French cognitive neuroscientist Stanislas Dehaene to develop a cognitive neuropsychological model of number processing that is the dominant model today. In this model, our number knowledge consists of three distinct representational systems. Each of these systems represents numbers in a very different format from the others and so supports different aspects of our numerical knowledge and abilities. Each is localized in a different brain region (see Fig. 2).

A *verbal word* representational system is located in the classical language-specific areas within the left hemisphere only (such as the left inferior frontal and superior and middle temporal gyri) and is responsible for the recognition and processing of spoken or written number words. This system represents numbers as syntactically organized sequences of words (employing

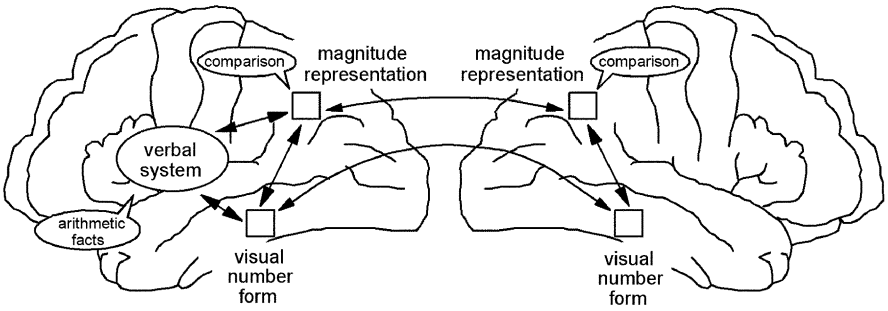


Figure 2 Stanislas Dehaene's triple-code model of number processing (courtesy of *Mathematical Cognition*).

the syntactic rules governing the number naming system of the language) and operates over phonological representations of the number words. This system also stores well-learned arithmetical facts, such as the memorized facts from addition and multiplication tables (subtraction and division facts are not memorized by rote and so are *not* hypothesized to be stored within this system). Finally, this system is heavily implicated in precise mental arithmetic such as multidigit calculations, which involve both retrieval of stored verbal facts and visuospatial representations of numbers.

A *visual Arabic numeral* representational system, located in the left occipitotemporal region of both hemispheres, underlies the recognition of Arabic digits and numerals. This system represents numbers in their Arabic numeral format as strings of digits.

Finally, an *analogical magnitude* representational system, located in the inferior parietal cortex regions of both hemispheres, employs representations that inherently embody the magnitude of a given numerical quantity, representations such as those given by the accumulator mechanism, for example. It is only within this third system that information about what we think of as the meaning or sense of a number—its quantity, magnitude, or size—is represented. Thus, comparison and relational information, such as that one number is smaller or larger than another, is only available within this system. This system also supports approximate, as opposed to precise, arithmetic calculations; such calculations depend on a sense of the magnitudes the values being operated on and the outcome value.

Connections exist (a) between the verbal, Arabic numeral, and magnitude systems within the left hemisphere, (b) between the visual and magnitude systems within the right hemisphere (recall that the verbal system exists solely in the left hemisphere), and (c) across the two hemispheric components of the visual system and of the magnitude system. These connections enable access to multiple kinds of information about a number that has been accessed through one of the systems. For example, the connections between the Arabic numeral code and the magnitude code mean that, when we see an Arabic numeral, say seven, we can then access its magnitude and so perform operations over the magnitudes and/or make magnitude judgments, such as which of two numerals represents the larger or smaller number.

There is considerable neuropsychological evidence, from case studies of brain-damaged patients as well as from studies of normal adults, that these systems are both anatomically and functionally distinct.

B. Evidence from Brain-Damaged Patients

First, consider patients with major left-hemisphere lesions. As one would expect given that language faculties reside in the left hemisphere, such damage is highly associated with severe problems in understanding and speaking language. But patients with this kind of damage can also be acalculic, showing profound deficits in their ability to deal with numbers: an inability to understand number words or relate them to each other and an inability to perform relatively simple arithmetical calculations such as adding $9+7$. However, in the face of this extreme impairment, nonverbal digit recognition, magnitude judgments, and some arithmetical operation abilities may still be functioning relatively normally. For example, one patient, NAU, who had suffered a major lesion of the posterior left hemisphere, had great difficulty reading number words and moreover was unable to solve even such simple arithmetic problems as $2+2$! However, when presented with digits, he could indicate where they should fall on the number line by pointing to roughly the correct location, showing that his digit recognition was preserved and that he was able to access the approximate magnitude of the number represented by the digit. Moreover, and most interestingly, whereas he was unable to give the correct answer to a problem such as $2+2$ and unable to state accurately the truth or falsity of arithmetical statements such as $2+2=5$ or $2+2=3$, he *was* able to state the falsity of statements such as $2+2=9$! Moreover, when presented with two digits, he could immediately say which of the two was larger. He had also apparently retained the approximate, but not the exact, magnitude values of numbers, stating (for example) that there are about 6 or 10 eggs in a dozen, about 350 days in a year, and about 50 min in 1 hr. In sum, NAU was able to identify and operate over the approximate magnitudes of the digits and obtain the approximate magnitudes of the solutions to simple arithmetic problems. In contrast to such effects of damage to the left hemisphere, damage to the right hemisphere has little or no effect on number processing.

Consider now lesions to the inferior parietal cortex. According to the model, such patients can be expected to have a preserved ability to call up memorized arithmetic facts and to correctly read and recognize number words and Arabic digits, but to have no sense of the meanings of these numbers and no understanding of numerical magnitudes and magnitudinal relationships. Again, numerous cases fit this description.

Mr. M., who had a lesion to the inferior parietal cortex, was unable to subtract (saying, for example, that $3-2=2$) and had great difficulty in saying which of two digits was larger. He was also unable to generate the solution to what number lies midway between two given numbers; not only were his responses wrong, they were frequently ridiculous, for example, he replied that the number that lies midway between 10 and 20 is 30 or 25. In essense, Mr. M. had lost all sense of the meanings of numbers. However, he was well able to read number words and digits and to generate answers to addition and multiplication problems learned by rote in school, such as that $3 \times 9=27$.

Finally, split-brain patients present an interesting test of the model. Because all three numerical codes are present in the left hemisphere, numerical tasks presented to the left hemisphere should show normal performance. However, because the right hemisphere lacks the verbal number system, numerical tasks presented selectively to the right hemisphere should show the same pattern of results as that for patients with left-hemisphere lesions. Indeed, this is the case: larger–smaller comparisons are performed by split-brain patients equally well (and on a par with the performance of normal controls) whether presented to the left or right hemisphere. In contrast, performance on mental calculation tasks equals performance by control subjects when the tasks are presented to the left hemisphere but is severely impaired when the tasks are presented to the right hemisphere.

C. Evidence from Normal Adults

In studies of normal adults, researchers Elizabeth Spelke and Stanislas Dehaene and their colleagues have found evidence that precise arithmetic facts are stored in a language-specific format, whereas approximate number facts and knowledge appear to be language-independent. In several experiments, bilingual subjects were trained on precise number facts (e.g., they had to choose which of two answers was the correct sum to an addition problem when given the correct answer versus a distractor answer very close in magnitude to the correct answer) and on approximate number facts (e.g., they had to choose the correct sum when given the choice between the correct answer and an answer that was much farther off from the correct sum, such that an idea of the approximate result would be sufficient to rule out the latter). After training, subjects were tested on these facts in the language in which the facts were trained and in the subjects' other language. The acquisition of the exact number facts did not transfer to the untrained language, whereas acquisition of the approximate number facts did: performance was equally good in both the trained and untrained language for the approximate facts. Thus, precise number facts appear to be stored in the same language in which they are learned, whereas approximate number facts, which engage subjects' magnitude sense of number, are represented in a language-independent format.

Brain-imaging experiments on normal adults also show that these two systems engage distinct brain regions. Several fMRI studies examined the brain structures recruited while subjects were engaged in precise number calculations and compared them with those structures activated during the processing of approximate number calculations. In accordance with the model, both left and right parietal lobes (along with some additional areas associated with visuospatial and analogical mental transformations) showed greater activation for approximate than for exact calculation. In contrast, the left inferior frontal lobe showed greater activation for exact calculation than for approximate.

The preceding sources of evidence all suggest that number processing rests on a variety of different mechanisms: some visuospatial, some highly language-dependent, and some involving a sense of numerical magnitude. Clearly, the visual and verbal representational systems for number are acquired—we are not born knowing the number words of our language or the Arabic numeral system. But where does our magnitude sense of number come from? If it is truly independent of language, we might see evidence of this component of numerical understanding in prelinguistic and nonlinguistic subjects, such as human infants and nonhuman animals. Before turning to empirical evidence addressing this question, however, let us look at this component of number knowledge in closer detail, in particular, the mental structures that underlie our sense of numerical magnitude.

III. MAGNITUDE REPRESENTATIONS OF NUMBER: THE ACCUMULATOR MODEL

A. Description of the Model

The accumulator model, briefly described in Section I, was originally developed by comparative

psychologists Warren Meck and Russell Church to account for a variety of numerical abilities that they found in rats. Meck and Church noticed a number of similarities in rats' ability to determine number and their ability to measure temporal duration. To account for these similarities, they proposed that a single mechanism underlies both abilities and expanded an existing model for measurement of temporal intervals so as to incorporate a counting component. Their proposed accumulator mechanism works as follows: a *pacemaker* puts out pulses of energy at a constant rate, which can be passed into an *accumulator* by the closing of a *mode switch*. In its timing mode, the switch closes at the beginning of the temporal interval being timed and remains closed for its duration, passing energy into the accumulator continuously at a constant rate. Thus, the amount of energy in the accumulator varies in direct proportion to the length of the timed duration. The fullness of the accumulator after timing of some duration can be compared with fullness values previously stored in memory to determine whether the just-timed duration is longer, shorter, or the same as a duration associated with some event. In its counting mode, when an entity is experienced that is to be counted, the switch closes for a brief, fixed interval and then opens again. Thus, when counting, the accumulator fills up in equal-sized increments, one increment for each entity counted, and its final fullness value varies in direct proportion to the number of entities so counted (see Fig. 3). Note that there will be differences in the exact fullness of the accumulator on different counts of the same number of items, resulting from inherent variability in the rate at which the pacemaker is generating pulses and in the amount of time the switch closes for each increment. This variability is normally distributed around a mean fullness value for each number, and the variance increases in proportion to the numerosity represented. This mechanism contains numerous accumulators and switches, so that the animal can count different sets of events and measure several durations simultaneously.

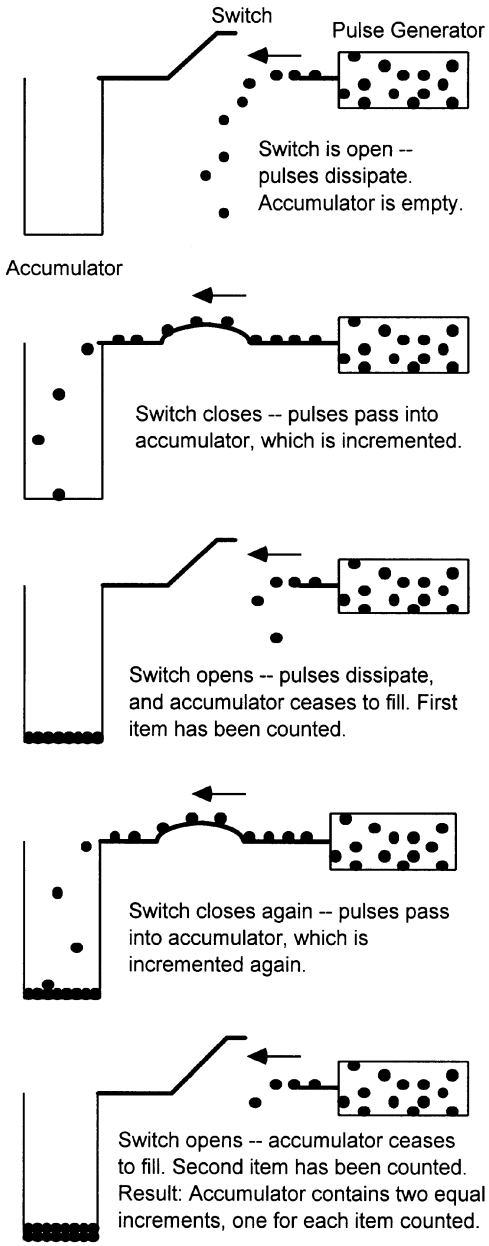


Figure 3 Schematic diagram of the states of the Meck and Church accumulator mechanism as it enumerates two items. The resulting fullness level of the accumulator is the mental representation for two (courtesy of *Current Directions in Psychological Science*).

B. Evidence for Similarity between Counting and Timing Processes in Animals

Evidence for functional similarity between rats' timing processes and their counting processes comes from several experiments. First of all, methamphetamine increases rats' perceptions of duration and numerosity by exactly the same factor, suggesting that the same mechanism is affected in both cases. This effect could be explained in the model by the drug causing an increase in the rate of pulse generation by the pacemaker, leading to a proportionate increase in the final value of the accumulator regardless of the mode in which it was operating. Second, both numerical and duration discriminations transfer to novel stimuli equally strongly when rats trained on auditory stimuli

were then tested on mixed auditory and cutaneous stimuli. Finally, an experiment tested the following prediction: If the animal's decision is based on a comparison of the final value of the accumulator with a previously stored value of the accumulator, then one might expect there to be transfer from making an evaluation on the basis of the output of the timing process to making an evaluation on the basis of the output of the counting process, so long as the final output value of the accumulator in the two cases was identical. For example, a count that yielded the same final fullness value in the accumulator as for a previously trained duration might be responded to as if it were that duration. This prediction was confirmed: when rats were trained to respond to a specific duration of continuous sound, they immediately generalized their response when presented with a certain number of 1-sec sound segments that had been calculated by the experimenters to fill up the accumulator to the same level as that for the duration on which the rats had been initially trained. Transfer was equally strong from number to duration. Meck and Church concluded that the same mechanism underlies both counting and timing processes in rats.

C. Predictions of the Model

A number of consequences fall out as predictions of the accumulator by virtue of the inherent structure and functioning of the model.

1. It predicts that animals (including human infants) should be able to discriminate different numbers of entities on the basis of number per se, not nonnumerical properties of arrays such as light–dark contrast, configuration, contour complexity, physical magnitude (size, volume, height, etc.), and so on. The accumulator is a mechanism that determines numbers of individual entities; it takes as input discrete *individuals*, not perceptual information. Moreover, because there is nothing inherent in the structure of the accumulator to restrict the kinds of idividuals it can count, we would expect infants and animals to be able to determine numbers of different kinds of entities. From an evolutionary standpoint, there are many kinds of entities it would be advantageous to be able to enumerate; thus, we might expect that a mechanism that evolved specifically for the task of determining the discrete number of things and that has no structural limitations on the kinds of input it can take, beyond the requirement of *discrete inividuals*, should be able to count different kinds of things.

2. Because of the fact that variance in the fullness of the accumulator increases with numerosity, with larger numbers representation of these numbers becomes more rough and approximate. As a result, discriminability for a given numerical difference should decrease as number increases. Because the variance is additive, the discriminability of two values should rest more on their proportionate difference than on their absolute difference; thus, animals and infants should be able to discriminate larger numbers, provided their proportionate difference is sufficient.

3. Because the representations have a structure in which the magnitudinal relations between different numbers are inherently embodied, we might expect animals and infants to possess procedures for extracting such information, that is, we would expect them to be able to use these representations in numerically meaningful ways.

The following section examines how these predictions bear out in empirical studies of numerical abilities in infants and animals. To preview, a review of the findings shows considerable evidence that this aspect of numerical understanding may form part of the inherent structure of the mind: a sense of numerical magnitude appears to be present early on in infancy and is evident in a wide range of animal species.

IV. NUMERICAL ABILITIES IN INFANTS AND ANIMALS

A. Sensitivity to Number per se

Animals can discriminate numerosities of simultaneously presented objects. For example, a raccoon was successfully taught, when presented with an array of three to five Plexiglas boxes each containing from one to five items, to choose the box containing three items; nonnumerical cues such as size, stimulus density, odor, and location of target box were controlled. Chimpanzees have been trained to pick out the correct Arabic number symbol when presented with arrays of zero to six physical objects. In all of these studies, results generalized to novel items without additional training.

In a particularly compelling study, Irene Pepperberg trained Alex, an African gray parrot, to speak the appropriate number word when presented with up to five objects. Most interestingly, because Alex had also been taught to "name" different kinds of objects, he could be asked to identify the number of a subset of items of the display. For example, when presented with

a tray with two corks and three keys randomly scattered over it, Alex could successfully answer all three of the questions, how many (correct answer: five), how many cork (two), and how many key (three). Here, he had to selectively attend to different aspects of the same display based on the question asked. Again, results generalized to items not previously presented to the parrot for numerical evaluation. The flexibility of this controlled focus of selective attention suggests a general concept of number accessible to higher level central processes.

In addition, many different species of animals are able to keep track of a number of sequentially occurring events. In a popular experimental paradigm, rats are trained to press a certain number of times on a particular lever before pressing a single time on a second lever for a reward; penalties of varying degrees are imposed upon the rats for pressing the second lever too early. Rats can learn this task with at least as many as 24 presses required. That their response is based on number of presses rather than on elapsed time is clear: when rats are trained to press for a certain amount of time, they continue to press for a certain extra *proportion* of the trained time on each trial in order to ensure that they have satisfied the criterion for reward, but when trained to press a certain number of times, they press a certain extra *constant number* of presses independent of the required number. Similar abilities have been shown in pigeons with a somewhat different task.

Rats have also been trained to turn down the third, fourth, or fifth left-hand tunnel in a maze and, once trained, did so even when the distance between the tunnels was varied from trial to trial and a corner had to be turned before the rewarded tunnel was reached. The rats could not simply have been running for a fixed length of time before turning left or responding to some level of fatigue. They had to have encoded the numerosity of the tunnels on the left in order to succeed at the task. Birds show similar abilities. In one experiment, canaries were successfully trained to select an object on the basis of its ordinal position in an array. Ten cubicles were spaced along a runway, and the canaries had to walk along the runway and choose the cubicle containing, say, the fifth aspirin in the series. Which cubicle contained the relevant aspirin was varied from trial to trial, ruling out any regularity of distance from the starting point. To control for the possibility that the birds might use rhythm as the basis of their judgments, the number of aspirins per cubicle ranged from zero to two, and the distance between cubicles varied from trial to trial. Given this, the birds must have been succeeding on the basis of the ordinal position of the aspirin.

The preceding review shows that a wide range of animal species are sensitive to number; they can discriminate different numbers of entities even when nonnumerical cues such as size, color, odor, length of run (and therefore amount of fatigue), density, and spacing, are controlled. They are also able to determine numbers of many different kinds of entities: physical objects, self-generated actions, sounds, and cutaneous stimuli.

Although the experimental evidence with human infants is less extensive than that with animals, the evidence does suggest that infants have similar abilities. Many habituation studies have found that infants (from a few days old to over 1 year of age, depending on the study) can discriminate number: when repeatedly presented with (''habituated to'') displays containing a given number of items, infants' interest will decrease and their looking time at each display will drop. They will subsequently look longer at a display containing a new number of items than at a new display containing the same, habituated number of items. In a study by Prentice Starkey, Elizabeth Spelke, and Rochel Gelman, for example, 7-month-old infants were habituated to visual displays of either two or three randomly arranged objects. The displays were constructed of photographs of various household objects that were different in each picture, for example, one picture might consist of an orange and a glove and the next might include a key chain and a banana. Following habituation, the infants were shown test pictures of two and three items containing new objects not seen in the habituation pictures. Infants looked significantly longer at test pictures containing the number of items that differed from what they had been habituated to, indicating that they discriminated between the two numbers.

In an intriguing series of experiments by Erik van Loosbroek and Ad Smitsman, infants of 5, 8, and 13 months were habituated to displays of two, three, or four moving ''objects'' displayed on a video screen. Each ''object'' was constructed by randomly filling in a certain proportion of the rectangles defined by a 16-by-16 grid. Stimuli were constructed at three different levels of density by specifying a higher or lower proportion of the grid to be filled in. The objects' paths of motion occasionally intersected so that one object overlapped another. Thus, a static view of the display would not guarantee correct number information: the number of objects could only be determined by watching the display in motion over time.

Nonetheless, infants of all ages tested discriminated two from three and even three from four, and the older two groups of infants also distinguished four from five.

Some of these results have come under question due to research findings by Melissa Clearfield and Kelly Mix. In several experiments, they found that infants were sensitive to the *contour length* of a display—the sum total of the contours of the individual items in a display. When infants were habituated to displays of a single number and contour length and then in testing presented with new displays containing alternately (a) the habituated contour length but a new number of items and (b) a new contour length but the habituated number of items, infants looked longer at the second kind of test trial, suggesting that the new contour length was of greater interest than the new number. This raises the possibility that, in previous experiments in which infants discriminated numbers of objects, they might have been doing so on the basis of contour length and not number per se, as the contour length of a display tends to covary with the number of items in that display.

However, there are reasons to doubt that all of infants' putative number-discrimination abilities can be accounted for by this explanation. There are now several studies in which infants have successfully discriminated number even when perceptual factors such as contour length and visual area were strictly controlled. In a study conducted by Paul Bloom, Wen-Chi Chiang, and the author, we found that 5-month-old infants could determine how many *groups* of objects were present in a display, independent of the number of individual objects themselves. Half the infants were habituated to two groups of three objects each moving on a computer screen, and the remaining infants were habituated to four groups of three objects each. Infants subsequently discriminated between two groups each containing four objects and four groups each containing two objects; in both habituation groups, infants preferred the test display containing the new number of *groups* of objects. Note that both test displays contained the same total number of objects—eight—and thus were equated on many perceptual dimensions such as density, area covered, and contour length of the display. Similarly, experiments examining infants' ability to discriminate relatively large numbers of objects have also controlled strictly for contour length, area, density, etc. of displays and have found successful numerical discrimination in infants under these conditions.

Like many nonhuman animal species, human infants are able to enumerate different kinds of entities. Not only can they identify the number of visual objects in displays, they can also determine numbers of physical actions, such as the jumps of a puppet. This has been shown in several studies conducted by Tanya Sharon and the author. A sequence of actions differs from a display of visual objects in several respects. Objects exist continuously through time and are separated from each other in two-dimensional (in the case of photographs or pictures) or three-dimensional space. Sequential actions, on the other hand, have a momentary existence, are separated from each other in one-dimensional time, and may or may not occur at precisely the same location in space. Perhaps most importantly, one has perceptual access to the entire collection of objects at once, whereas one has perceptual access to just one element at a time in a collection of sequential actions and cannot anticipate the final element.

In one set of experiments, 6-month-old infants were habituated to a puppet jumping; for half of the infants it jumped two times and for the other half three times. Upon completion of the jump sequence, infants' looking time at the now-motionless puppet was measured. Following habituation, infants received test trials in which the puppet jumped two times and trials in which it jumped three times. The jump sequences were structured to control strictly for the tempo and overall duration of jump sequences. We found that infants looked significantly longer at the puppet following a new number of jumps than following the old number of jumps. Thus, infants were able to identify the individual jumps in the sequence and to enumerate them. Subsequent experiments showed that infants could also enumerate heterogeneous action sequences composed of several distinct kinds of actions.

Additional studies show that infants are able to determine the number of other kinds of entities as well. One study (described earlier) showed that 5-month-olds can enumerate collective entities; they attended to how many groups of items were present in a display. Finally, some studies have found that infants can determine numbers of sounds; however, other researchers have obtained conflicting data, so that further research is required to clarify this issue.

B. Discriminability and Set Size

The finding that, as numerosity increases, discriminability of a give numerical difference decreases is a

robust effect, found within animals, human infants, and human adults. In an experiment in which rats were required to discriminate two from three and three from four sounds, all three subjects successfully discriminated two from three, but only two learned to discriminate three from four. Similarly, in an experiment requiring squirrel monkeys to pick out the smaller of two simultaneously presented arrays of objects, subjects were able to distinguish six from seven and seven from eight, but only one subject was able to distinguish eight from nine. Discrimination of adjacent numbers by human infants is also stronger with smaller numerosities. Two studies obtained results in which infants discriminated two from three but not four from six; in another study, infants discriminated two objects from three reliably, discriminated three from four only in certain situations, and did not discriminate four from five. In yet another study, three different ages of infants discriminated two objects from three and three from four, but only the oldest age groups (8- and 13-month-olds) discriminated four from five. These results, though obtained with a very different methodology, parallel the finding that human adults also show an effect of set size, even within the subitization range, when asked to rapidly identify the numerosity of an array; the identification of numbers is both faster and more error-free the smaller the number.

However, both animals and human infants are able to represent approximate values of larger numbers and to discriminate two numbers when the *proportion* by which they differ is sufficiently large. When rats are trained to press a lever a minimum required number of times in order to obtain a reward, the variance of their response increases with the required number, but they are nonetheless able to perform this task even with numbers as high as 50 (the largest tested). Human infants can discriminate 8 items from 16 and 16 from 32, with nonnumerical factors such as area, density, and contour length controlled.

C. Computational Capacities

There is more to numerical knowledge than the ability to distinguish different numbers. The ability to distinguish numbers does not entail an ability to reason about those numbers, to determine, for example, that five is larger than three or that two is composed of one and one. To determine such relationships, the animal or infant must not only be able to construct mental representations of the relevant numbers but be able to manipulate these representations in numerically meaningful ways. There is empirical evidence that infants and animals can determine the results of certain numerical operations on small numbers of physical objects.

Studies in the author's lab have investigated human infants' numerical reasoning capacities. In one experiment, 5-month-old infants were divided into two groups. Those in the "$1+1$" group were shown a single item being placed into an empty display area. Then a small screen rotated up, hiding the item from view, and the experimenter brought a second identical item into the display area in clear view of the infant. The experimenter then placed the second item out of the infant's sight behind the screen (this sequence of events is shown in the top portion of Fig. 4). Thus, infants could see the nature of the arithmetical operation being performed but could not see the result of the operation. The screen was then dropped to reveal an outcome of either one (the impossible outcome) or two (the possible outcome) objects. Infants in the "$2-1$" group were similarly presented with a sequence of events depicting a subtraction of one item from two items (shown in the bottom portion of Fig. 4). Again, after this sequence of events was concluded, the screen rotated downward to reveal either one (now the possible outcome) or two (impossible outcome) items in the display case.

Infants' looking time at the display was recorded when the screen dropped. The prediction was that infants would be surprised by an apparently impossible result. Thus, the two groups should show significantly different looking patterns; infants in the $1+1$ group should look longer when the result is one than when it is two, in comparison to the $2-1$ group, which should show the reverse pattern. (A pair of pretest trials, in which infants were simply shown displays of one and two items, revealed that infants in the two groups did not differ in their baseline looking patterns at one and two items.) This was, in fact, the pattern of results obtained; infants in the two groups differed significantly in their patterns of looking in the test trials. Infants in the $1+1$ group looked longer when the addition appeared to result in a single item than when it resulted in two items, whereas infants in the $2-1$ group looked longer when the subtraction appeared to result in two items than when it resulted in a single item. To ensure that infants were determining the exact result of the operation (that is, that they were expecting two objects in the $1+1$ situation and one object in the $2-1$ situation) rather than simply

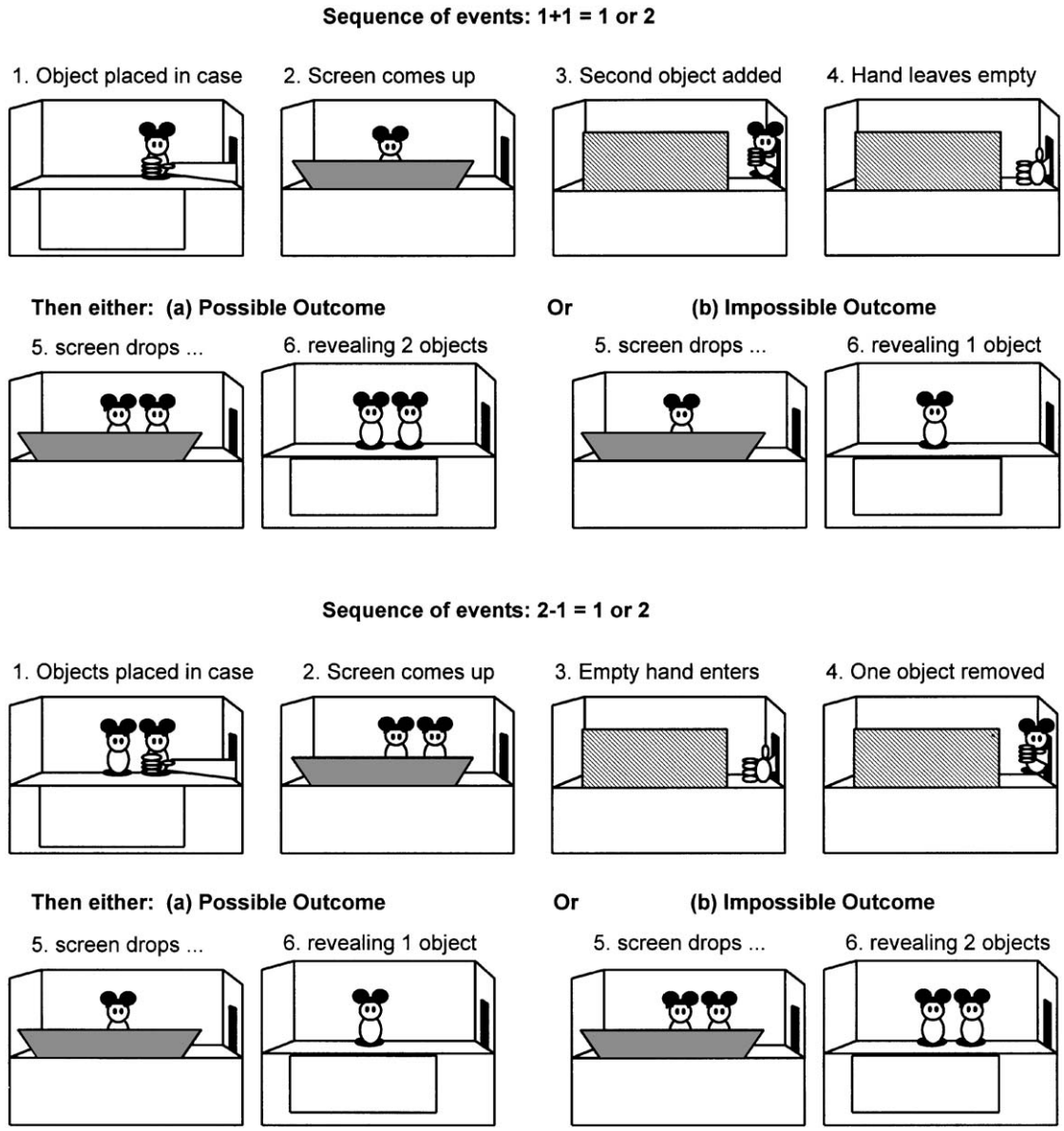


Figure 4 Sequence of events shown to infants in Wynn (1992) (courtesy of *Nature*).

expecting the number to have been changed in some way as a result of the operation without having expectations as to precisely what the result should be, another experiment was conducted. Here, infants were shown an addition of 1 + 1, with outcomes of two and three objects. In this case, both outcomes are different from the initial number of objects (one) placed in the display. Thus, if infants only expect *some change* to obtain as a result of the addition, they will not be surprised by either outcome. However, if they are computing the precise nature of the numerical change, they will be expecting two objects behind the screen and will look longer at the incorrect result of three objects. (A pretest condition showed that infants looked equally long at two and three objects.) This was the pattern of results obtained: infants looked longer when the addition appeared to result in three items than in two items.

The studies reviewed previously and others like them show that infants as young as 5 months of age are sensitive to the numerical relationships between small numbers and are able to determine the results of simple numerical operations. Similar abilities have also been shown in somewhat older infants by a task that

required them to perform motor actions based on their numerical expectations. In this study 18- to 35-month-olds saw from one to five identically colored ping-pong balls placed into an opaque box and then saw an experimenter either add or remove a small number of balls. They then were allowed to reach into the box to retrieve the objects. The box was constructed so that infants could not see into the box as they were reaching into it, so that each reach into it allowed contact with, and removal of, only one object at a time. Thus, the number of reaches into the box that the infants made indicated how many items they believed the box to contain. Even the 18-month-olds showed a knowledge of how many objects the box contained when small numbers were involved.

Some nonhuman species are also able to compute results of certain numerical operations. The most conclusive evidence comes from studies conducted by Sarah Boysen and Gary Berntson, who taught a female chimpanzee to associate the Arabic numerals 0–4 with their respective numerosities. Without further training, she was able to choose the numeral representing the sum of oranges hidden in any two of the three possible hiding places in her pen. Most impressive of all, when the sets of oranges in the hiding places were replaced with Arabic numerals (one card with an Arabic numeral printed on it hidden in each of any two hiding places), she was immediately able to choose the Arabic numeral representing the sum of the two found numerals. That is, without training, she was able to operate over two symbols representing numerosities in such a way as to arrive at the symbol representing their sum. In another study, free-range wild rhesus monkeys were given a version of the $1+1$ addition situation described earlier for human infants. When shown one eggplant and then another eggplant placed in a box out of sight, the moneys looked significantly longer when only one eggplant was revealed than when two were, despite showing no preference in a control situation for looking at one eggplant over two.

Suggestive findings have been obtained with other species as well. Rats appear to anticipate when they are approaching the required number of presses when they must press a required minimum number of times on a lever to obtain a reward. Rats will frequently check for the reward before they have given the required number of presses in situations when there is no penalty for checking for the reward too early and, upon finding no reward, will return to the lever to increase their number of presses. Interestingly, the greater their number of presses before pausing to check for the reward, the smaller their number of additional presses upon returning to the lever. That is, they appear to know how close they are to the needed number, not only whether they have or have not reached that number yet. Finally, in one study, rats were trained to press a lever on the left when presented with either two sounds or two light flashes and a lever on the right when presented with four sounds or four light flashes. Following training, rats were presented with two sound–light flash pairings (a sound accompanied by a simultaneous light flash, followed by another sound accompanied by a light flash). In this situation, rats pressed the right-hand lever, showing that they had computed that there were four stimuli altogether. They were either adding the number of sounds to the number of light flashes or enumerating the stimuli independently of their kind.

D. Summary

As shown in the preceding review, both human infants and nonhuman animals are sensitive to number. They can determine numbers of fundamentally different kinds of items—objects, actions, collections—and can do so for items presented simultaneously as well as those presented sequentially. Moreover, animals and infants have procedures for computing the numerical relationships that hold between different numbers, which entails that their representations for these numbers have a structure that allows for the extraction of such information. These findings suggest the existence of an unlearned capacity for numerical representation and reasoning that spans these different species and is operational quite early on in life.

V. CONCLUSIONS

Evidence from cognitive neuropsychology suggests that our number processing and arithmetical capacities comprise a set of dissociable systems: verbal knowledge of number and stored precise number facts, procedural knowledge for numerical computations, visual knowledge of Arabic numerals, and a magnitude sense of numbers. Evidence from comparative and developmental psychology suggests that both human infants and nonhuman animals share one of these component systems: a magnitude sense of number that supports enumeration, numerical comparison, and numerical computation operations. Language and learning appear to play key roles in the

formation of the other components; in particular, evidence suggests that we may store learned numerical facts in the specific natural human language in which they are learned. Much more remains to be learned about (a) how these systems interrelate, (b) precisely how we build onto the initial sense of numerical magnitude, and (c) how we obtain genuinely new mathematical knowledge.

See Also the Following Articles

ARTIFICIAL INTELLIGENCE • COGNITIVE REHABILITATION • CREATIVITY • INFORMATION PROCESSING • INTELLIGENCE • LANGUAGE AND LEXICAL PROCESSING • LOGIC AND REASONING • PROBLEM SOLVING • VISION: BRAIN MECHANISMS

Suggested Reading

Ashcraft, M. H. (1995). Cognitive psychology and simple arithmetic: A review and summary of new directions. *Mathemat. Cogn.* **1,** 3–34.

Boysen, S. T., and Capaldi, E. J. (Eds.). (1993). *The Development of Numerical Competence: Animal and Human Models*. LEA, Hillsdale, NJ.

Butterworth, B. (1999). *The Mathematical Brain*. Macmillan, London.

Campbell, J. D. (Ed.). (1992). *The Nature and Origins of Mathematical Skills*. Elsevier, Amsterdam.

Davis, H., and Perusse, R. (1988). Numerical competence in animals: Definitional issues, current evidence and a new research agenda. *Behav. Brain Sci.* **11,** 561–615.

Dehaene, S. (1997). *The Number Sense: How the Mind Creates Mathematics*. Oxford University Press, Oxford, UK.

Dehaene, S., and Cohen, L. (1995). Towards an anatomical and functional model of number processing. *Mathemat. Cogn.* **1,** 83–120.

Donlan, C. (Ed.). (1998). *The Development of Mathematical Skills*. Psychology Press, East Sussex, UK.

Gallistel, C. R. (1990). Number. In *The Organization of Learning*, Chapter 10, pp. 317–350. MIT Press, Cambridge, MA.

Wynn, K. (1998). Psychological foundations of number: Numerical competence in human infants. *Trends Cogn. Sci.* **2,** 296–303.

Wynn, K. (1998). An evolved capacity for number. In *The Evolution of Mind* (D. Cummins and C. Allen, Eds.). Oxford University Press, Oxford, UK.

Object Perception

PEPPER WILLIAMS
University of Massachusetts, Boston

GLOSSARY

primitives The fundamental elements, or basic building blocks, of a representation. These can range from simple spots of light of varying intensity to complex three-dimensional volumes with varying aspect ratios, degrees of curvature, cross-section shapes, etc.

representation A thing that stands for another thing. In other words, a representation can be used in place of its representee to carry the same meaning. For example, France is represented by a flag with red, white, and blue horizontal stripes; the United Nations can fly this flag in front of its headquarters to indicate that this country is a member of the institution.

structural description A form of representation that uses relatively complex primitives, usually corresponding to object parts, and specifies the relationships between parts explicitly.

view A form of representation that uses relatively simple primitives and does not specify relations between them explicitly.

Object perception refers to the process of identifying an external stimulus from the image cast by the object onto the retina. This ability is remarkable because of the fact that an object's image can change radically when it is viewed under different circumstances, for example, from different viewpoints and distances. Furthermore, before perception of a single object can even begin, the object of interest must be separated from other objects in the background and foreground of the scene. Some theories propose that the visual system constructs a *structural description* of to-be-perceived objects, in which the shapes of and relations between the object's parts are explicitly represented. Other models posit simpler representations in which parts and relations are specified only implicitly in a view of the object; these models then specify processes for matching perceived to encoded views. Neither type of theory is universally accepted, but the theoretical debate has inspired, and in turn been constrained by, a great deal of interesting neurophysiological and behavioral research.

I. OBJECT CONSTANCY

Consider the objects in Figs. 1A and 1B. Call them SOGI and DUVA, respectively. Although you have never seen these objects before, you should have little difficulty perceiving that the images in Figs. 1C–1I are of the SOGI, not the DUVA. The ability to treat different two-dimensional (2D) images as the same three-dimensional (3D) shape is known as *object constancy*, and the acheivement of object constancy is the fundamental goal of object perception. As is the case with most fundamental cognitive processes, we generally take this ability for granted, but also as with many other cognitive processes, the mechanisms that allow us to perceive objects are far from simple.

One way to appreciate the difficulties inherent in achieving object constancy is to consider what we would have to do to progam a computer to recognize objects. To a computer, the images in Fig. 1 are nothing more than collections of numbers that specify the intensity of each small dot, or *pixel*, in each image. To teach the computer to associate the collection of pixels in Fig. 1A with the label SOGI and the pixels in Fig. 1B with the label DUVA is a simple matter of

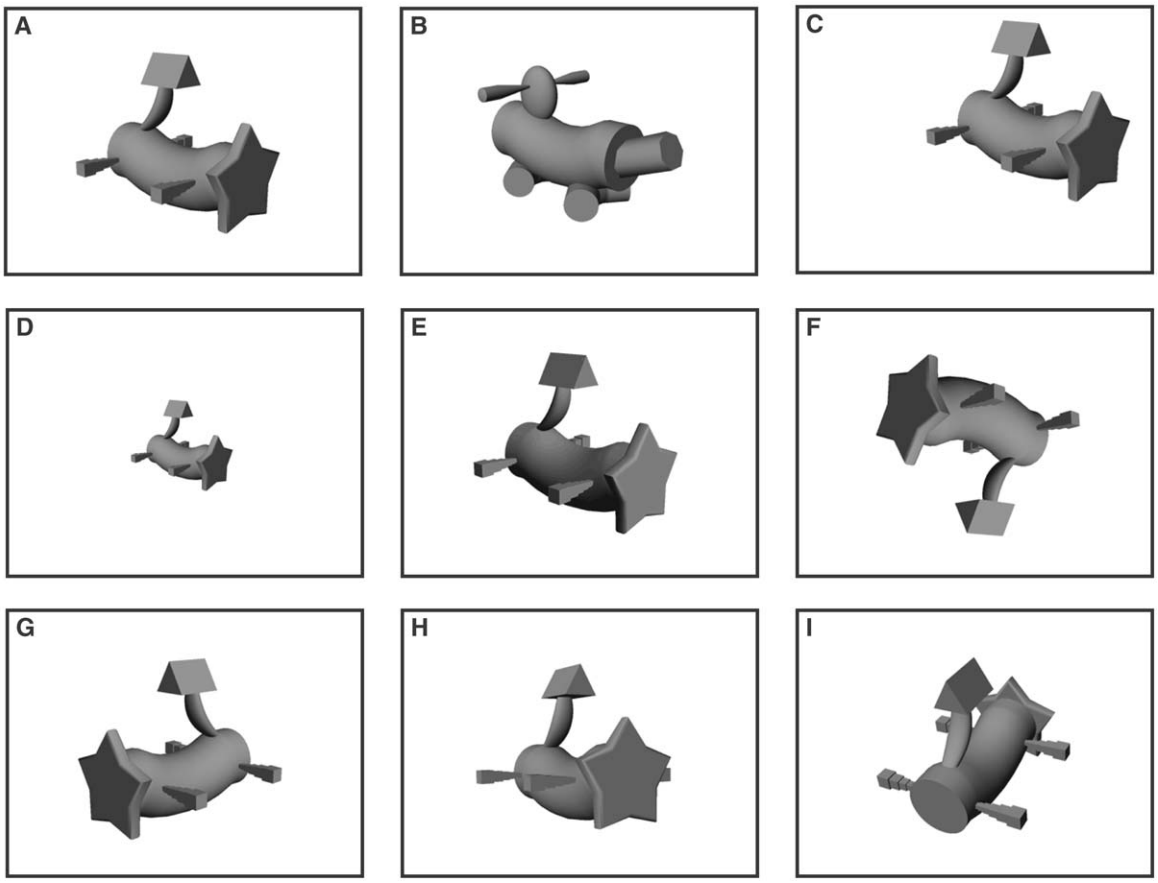


Figure 1 (A) and (B) show two novel objects under similar viewing conditions. (C)–(I) show how the former object would look under different viewing conditions. More specifically, the images illustrate (C) a translation (caused by the observer shifting his or her gaze or the object moving in the picture plane), (D) a size change (caused by the observer moving backward or the object moving away from the observer in depth), (E) a shift in lighting direction, (F) a shift in picture plane orientation (caused by the object turning upside down or the observer standing on his or her head), (G) a mirror reflection, and (H–I) viewpoint shifts (caused by the object rotating in depth or the observer moving around the object). Each of these changes to viewing conditions introduces significant alterations to the image delivered to the retina, and the human object perception system must somehow compensate for these alterations in order to successfully identify the object.

storing the pixel values and labels of both images in memory. When one or the other image is encountered again, the new pixel values can be compared to those of the previously seen images. If the new values match those of Fig. 1A, the SOGI label is retrieved, whereas if the values match those of Fig. 1B, the DUVA label is retrieved. But what will happen when the images in Figs. 1C–1I are presented to the computer? The collections of pixels in these images are identical to neither the learned SOGI nor the learned DUVA image, so our simple pixel comparison process will fail to produce a perfect match with either remembered image.

The human visual system faces the same conundrum: the raw input for vision is a set of firing rates of receptor cells (rods and cones) in the retina, roughly analogous to the set of pixel values that serve as input to the computer. If an object was always seen under the same viewing conditions, its image would always produce the same retinal firing rates and object perception would be straightforward. But Figs. 1C–1I illustrate several different ways in which alterations in viewing conditions produce different retinal input from the same object. Somehow, the visual system must find a way to compensate for these changes to the retinal image so that the perception of the object remains constant. Let us consider each of these forms of object constancy in turn.

In Fig. 1C, the image of the SOGI has been moved, or *translated*, relative to the bounding rectangle. If one's gaze is fixed at the center of the rectangle, the

pattern of retinal activity will be almost completely different for this image than for Fig. 1A. The visual system usually elicits eye movements that bring the image of a to-be-recognized object onto the fovea, a process that effectively compensates for image translations. However, common experience and psychophysical experiments indicate that objects can be recognized before being foveated: even when pictures of objects are shown experimentally for less than 0.3 sec, precluding eye movements, translations produce a minimal, if any, decrease in recognition accuracy or time. Therefore, some other compensation mechanism besides eye movements must also be at work.

In Fig. 1D, the image of the SOGI has been shrunk to one half its size in Fig. 1A. Although actual, physical size is an important characteristic of every object, the size of an image on the retina is determined jointly by the size of the object and the viewing distance (when viewed from certain distances, a tennis ball, a basketball, and the moon may all produce the same retinal size despite wide variations in physical size). To use object size as a cue for identity, then, depth cues would have to be consulted to determine object distance. Whereas this strategy may sometimes be utilized, the images in Figs. 1A and 1D contain no depth cues, yet we have a strong predilection to perceive them as the same object (and thus the same size), demonstrating that the visual system often somehow ignores retinal size when computing object identity. As with translations, most psychophysical studies that have investigated the issue have found little if any effect of object size on recognition fluency.

Fig. 1E shows what would happen if the light source were moved from above and to the left of the SOGI, as it is in Fig. 1A, to a position above and to the right of the object. Although one might not even notice this change if it was not explicitly pointed out, almost all of the pixels in the interior of the objects are different in the two images. For example, in Fig. 1A, the side of the SOGI's prism-shaped "head" is brighter (i.e., the pixel values here are more intense, and retinal receptor cells will fire at higher response rates to this surface) than the front of the star-shaped part, whereas in Fig. 1E this relationship is reversed. Because humans and other diurnal animals evolved under conditions in which the angle of their only light source (the sun) was constantly changing throughout every day, compensation for lighting direction may have been an important factor in the evolution of object perception.

Fig. 1F shows what the SOGI would look like if the observer were to stand on his or her head or if the SOGI were to be flipped upside down: the image's orientation has been shifted 180° in the picture plane relative to Fig. 1A. Because humans usually stand on their feet and many objects have a normal upright position (the trunks of trees and wheels of cars almost always touch the ground), orientation often provides a diagnostic cue for recognition. Psychophysics studies discussed later show that, although we can identify upturned trees and cars with high accuracy, the time taken to recognize such misoriented objects varies with the amount of misorientation, indicating that the visual system may take advantage of this orientation diagnosticity by representing objects in their canonical positions.

Fig. 1G is a mirror image of Fig. 1A. Although the amount of pixel shifting is about the same between mirror reflections as between upright and upside down images, reflections cause very small (if any) decreases in performance in psychophysical experiments, whereas, as noted earlier, turning of an image upside down typically leads to a substantial impairment in recognition time. Explanations for this combination of results feature prominently in many theories of object recognition.

Fig. 1H shows what the SOGI would look like if the observer moved slightly to the left relative to the vantage point from which the object was seen in Fig. 1A, or if the object rotated to its right. Fig. 1I shows the result of a larger change in viewpoint. Note that, unlike all of the other image transformations discussed here, changes in viewpoint may cause different parts and surfaces of the object to become visible or invisible and, at the least, lead to changes in the projection of the 3D surfaces onto the 2D retinal image. For example, the front surface of the star-shaped part of the SOGI covers a larger amount of the image (relative to other surfaces in the object) in Fig. 1H compared to Fig. 1A and then becomes completely invisible in Fig. 1I. In this way, the image changes introduced by viewpoint shifts are more radical than those caused by the other transformations discussed earlier. Thus, compensation for viewpoint shifts seems to be a computationally more demanding problem, and the most crucial measure of any object perception theory is usually taken to be how well it can account for humans' ability to achieve viewpoint constancy.

II. PRECURSORS TO OBJECT PERCEPTION

The pictures in Fig. 1 each show a single object isolated against a uniform white background. Object perception

is a difficult enough computational problem in this context, but in the real world we face the additional challenge of perceiving an object of interest amid a cluttered foreground and background of other objects, as illustrated in Fig. 2. In order to deal effectively with the enormous amount of information entering through the eyes, the visual system employs a divide-and-conquer strategy. First, the information coded initially as raw receptor firing rates is recoded in terms of various types of elementary features. The features are then analyzed to determine which combinations form single objects, and collections of features are pieced together to form representations of various aspects of the scene, including the object we want to identify. This section briefly describes some of what we know about the breakdown of visual information into features and the later regrouping of these features.

A. Breaking Down the Image

The process of breaking the visual field down into constituent features begins before visual information has even left the eye. Whereas receptor cells simply fire action potentials more rapidly the brighter the light that hits them, ganglion cells, the neurons whose axons leave the eye through the optic nerve, respond to particular *patterns* of light over a small patch of the retina. More specifically, most ganglion cells exhibit *center-surround* receptive fields, firing most rapidly either to a small spot of light surrounded by darkness (an on-center receptive field) or to a small spot of darkness surrounded by light (an off-center receptive field).

By the time visual information has reached the primary visual cortex (also known as striate cortex and area V1), it has been recoded again. Here, neurons called simple cells respond most strongly to stationary bars (rather than spots) of light surrounded by darkness (or to dark bars surrounded by light). Furthermore, each of these neurons responds best to a bar in a particular orientation. A second type of neuron in V1, the complex cell, also responds to oriented bars but is tolerant of small shifts in location of the bars. In addition to bars, simple and complex cells also respond to the borders between regions of light and darkness. One prominent place where such borders occur is at the edges of objects, so that simple and complex cells can be considered edge detectors (see Figs. 4A and 4B). Oriented edges are thought to be some of the most important features coded by early visual processes.

Other types of elementary features thought to be coded in the primary and secondary visual cortex include color, line size, line curvature, edge corners and intersections, motion direction, and distance in 3D space. Each feature can be thought of as being coded in a topologically organized feature map, which holds information about which parts of the visual field contain the feature. Early visual cortical areas contain a large number of such feature maps, e.g., one map for right-tilting lines, one for horizontal lines, and one each for the colors blue, green, and red. For example, a pencil rolling across a desktop might engender activity on the yellow color, vertical line orientation, and rightward motion feature maps. By splitting up the information in this way, the visual system is able to process each type of elementary feature separately, but at the same time.

Figure 2 Unlike the situation shown in Fig. 1, real-world object perception almost always requires that the to-be-perceived object be separated from other elements in the visual field, as illustrated here. For example, the coffee mug is partially occluded by the plate and salad bowl and itself occludes parts of the tabletop, fruit bowl, and an apple. To deal with this complexity, the visual system breaks the image into constituent features such as edges and color patches, decides which features belong together and which belong in separate groups, and then recombines feature subsets to form representations of single objects.

B. Building Up Objects

Once the visual image has been parsed into its constituent features, the visual system must put different features together to form representations of individual objects. Scientists working in the Gestaltist tradition, a school of psychology that began in mid-twentieth century Germany, have proposed a number of principles that appear to guide this process of

perceptual organization. For example, the principle of good continuation states that, other things being equal, lines or edges that form straight or smoothly curving contours should be grouped together. Thus, in Fig. 2, the lines forming the back edge of the table are combined into a group even though they are separated by the occluding salt and pepper shakers. Other Gestalt principles include proximity (elements close to each other should be grouped together), common fate (elements moving in the same direction should be grouped together), and connectedness (elements connected by other elements should be grouped together).

Whereas Gestalt psychologists have worked to provide intuitive heuristics for *which* elements should be grouped together, another group of cognitive scientists have worked on the problem of *how* elements are combined. The key concept in this tradition is that of visual attention, and the metaphor of a spotlight is often used to describe the role of attention in feature combination. The attentional spotlight illuminates a circumscribed portion of the visual field and combines the various features present in that portion of the image into a multidimensional representation, which can then be identified by the object perception processes discussed later. Because attentional capacity is limited, we are only able to analyze in detail a relatively small portion of an image at a time. Therefore, the attentional spotlight must move from place to place in a serial fashion in order to completely analyze an image, even though all individual features are detected in parallel.

Object perception theories generally make the simplifying assumption that all of the features of a to-be-perceived object are eventually (possibly after several shifts of the attentional spotlight) combined together into a single representation, which can then be compared to representations in memory to determine the object's identity. However, research on a phenomenon termed "change blindness" indicates that this assumption may not be valid. Consider the image in Fig. 3. Is the scene depicted here exactly the same as the one in Fig. 2 (examine the two figures before reading on)? Not quite: the handle of the pepper shaker is missing from the second figure. Behavioral studies using a number of different paradigms have converged on the conclusion that subtle changes such as this one often go unnoticed, especially when visual attention is not focused on the region of the visual field where the change occurs. Even changes to a single, isolated object that is clearly the focus of attention can be missed if the object is obscured from view for a short period of time. These findings imply that our initial representations of objects may not be as complete as we intuitively believe them to be. Object perception theories may need to be altered to take this imprecision into account.

Figure 3 This image is similar, but not exactly the same, as the one in Fig. 2 (see the text for a description of the change). The fact that experimental participants are poor at detecting such changes indicates that our representations of perceived objects are not as complete as we intuitively believe them to be.

III. THEORIES OF OBJECT PERCEPTION

Object perception begins with retinal receptors firing in response to light rays that have bounced off the to-be-perceived object and into the eye. The pattern of firing rates over the millions of receptor cells *represents* the object, in that this particular pattern of firing will result whenever the object is seen from this viewpoint and distance, under these lighting conditions, etc. However, as described in Section I, this representation changes drastically when the object rotates, the lighting direction changes, the observer moves toward or away from the object, etc.

Theories of object perception seek to explain how the visual system compensates for such changes to viewing circumstances, transforming the ever-changing retinal representation into a stable representation of an object. The low-level visual processes described in Section II, which organize the retinal output into collections of elementary features, begin this transformation, but theorists disagree on how these feature assemblies are used to ultimately produce object constancy. No single model enjoys universal support, so that rather than catalog the numerous theories that have been proposed, more general distinctions between classes of theories are described next.

A. Recovery of 3D Structure

The goal of object perception is to treat different 2D views of a 3D object similarly. One intuitive way to do this is to recover the 3D structure of the object from the 2D image. If the visual system can construct an accurate 3D model of the to-be-recognized object, the problem of object constancy is automatically solved, because such a model would constitute a *viewpoint-invariant* representation—the same model would be constructed for a given object regardless of viewing conditions, so that all retinal images would result in an identical representation.

David Marr, whose theoretical work has driven much of the research in object perception, felt that the derivation of a precise, viewpoint-invariant 3D model is a fundamental step in the achievement of object constancy, and most vision scientists have agreed with this ideal. But for most objects, it is not feasible for the visual system to construct a complete 3D model from any one view of the object. For example, the structure of the back of a house is unknowable when the house is viewed from the front. Similarly, it is impossible to know whether or not the SOGI has a fourth "arm" from the view in Fig. 1A (Fig. 1I indicates that the fourth arm does in fact exist).

A more realistic goal is to represent the 3D structure of the portion of the object that can be seen from a given viewpoint. In such a system, small shifts in viewpoint (such as that from Figs. 1A to 1H) would be likely to lead to the same *structural description* and thus would cause no trouble for the visual system. Large viewpoint shifts may result in a completely different structural description, and when these shifts occur, object constancy is expected to be violated (i.e., object perception should be error-prone and/or slow under such conditions).

Despite the elegance of the structure-recovery approach, it is important to understand that limited object constancy is theoretically attainable without referring to a 3D structural description. A representation of the 2D image of an object—such a representation is referred to as *view-based*—can serve as the basis for object perception as long as perceived views can be accurately matched to encoded views. The latter part of this statement is crucial: the success of view-based models depends on the ability to specify processes that select the proper encoded representation when presented with a novel view of an object. Thus, whereas the focus of structural description theorists has been on how to construct an informationally rich representation of an object, the focus of view-based theorists has been on specifying computational procedures that achieve object constancy using less sophisticated representations but more complex matching processes.

B. Specification of Primitives and Relations

A second way in which various object perception models differ is in the elementary units, or *primitives*, that make up object representations. Again, Marr set the agenda by proposing a different set of primitives for each of the four stages in his object perception theory. In this framework, the image initially is represented by pixel intensity values, then the contours of the object are represented by oriented line segments, then the "2.5D" structure of the object is represented by surface patches, and finally the 3D structure is represented by a collection of volumetric primitives called generalized cylinders (which include such shapes as bricks, pyramids, and tubes).

Representations made up of simple primitives such as pixel intensity values have the advantage of describing an object with great precision. For example, in Fig. 4A, we see a close-up of the image pixels in a portion of Fig. 1A. A great deal of information is derivable from this simple representation, e.g., the various surfaces making up the object's "head" and "neck," the amount of light reflected by the surfaces, the precise curve of the neck, and the fact that the head is above the neck. Though not explicitly stated, these details nevertheless are implicit in the collection of pixel intensity values.

If we transform this representation into one using more sophisticated primitives, such as the oriented line segments illustrated in Fig. 4B, we lose some information but gain the advantage of explicitly representing certain properties of the object. Here, the oriented lines explicitly specify the boundary between the object and the background but do not code anything about surface lightness either explicitly or implicitly. In Fig. 4C, we see a further transformation from contours to a structural description with high-level primitives. Here, we explicitly represent the facts that there are two distinct parts and that one has the shape of a curved tube, whereas the other is wedge-shaped. But now we have lost the information about exactly how much the neck is curved.

Along with the set of primitives, an object perception model must also specify how the relationships between primitives are coded in a representation. Structural descriptions generally use high-level primitives (e.g., generalized cylinders) and specify relations

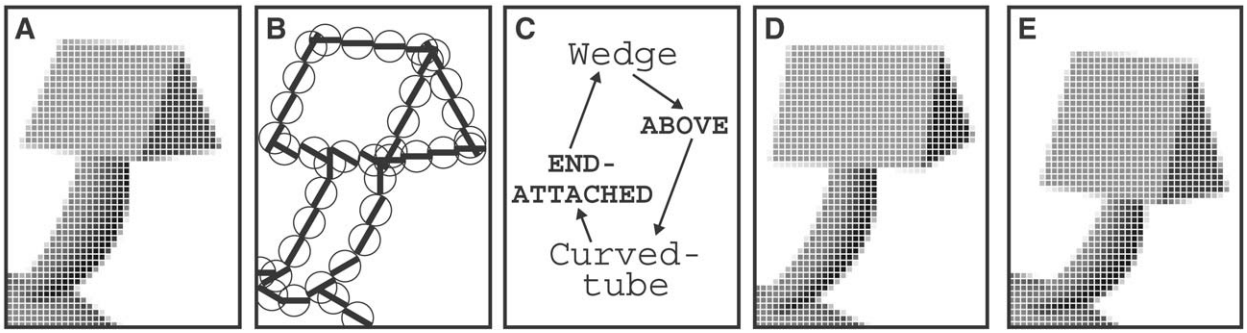


Figure 4 (A)–(C) illustrate three representations of a portion of the SOGI object using increasingly complex primitives. (A) simply includes the pixel intensity values of the image, (B) codes the orientations of the object's edges, and (C) is a structural description that explicitly denotes the shapes and relations between the two parts of the object. The latter representation is stable across the viewpoint shift that produces the image in (D) but is not sensitive enough to represent the difference between the objects depicted in (A) and (E).

explicitly—for example, the representation in Fig. 4C makes explicit the fact that the wedge is ABOVE the curved tube. Indeed, explicit specification of the relations between primitives is, for many theorists, the defining property of a structural description. View-based theories, on the other hand, generally use low-level primitives and represent relations between them only implicitly. That is, Fig. 4A includes the information that the head is above the neck only because the gray pixels of the head part have X-coordinates that are smaller than the coordinates of the pixels composing the neck part.

C. Structural Descriptions vs Views

By representing an object in terms of a limited number of high-level primitives and by explicitly specifying the qualitative relations between parts, structural descriptions are *stable* across many changes in viewing circumstances. For example, Fig. 4D shows the same object as Fig. 4A from a new viewpoint. Whereas the image (and thus any view-based representation) has changed, the same structural description should be generated. As noted earlier, such stable representations automatically solve the object constancy problem. To understand the downside of structural descriptions, consider Fig. 4E. Here, we see an image that is clearly of a different object but that also corresponds to the structural description in Fig. 4C. This example illustrates that structural descriptions may not be *sensitive* to distinctions in object structure that define the difference between two objects.

Another way to understand this trade-off is in terms of the information preserved and discarded in structural descriptions and view-based representations. As noted earlier, some information about an object that is implicitly captured in a view, e.g., the precise angle of the apex of the wedge-shaped part in Fig. 4A, is discarded in the process of forming a structural description. Throwing away this information boosts the stability of the representation, because when the object is rotated in depth, the angle may appear to change (as it does in Fig. 4D). The problem is that sometimes the discarded information is necessary for a representation to be sensitive enough to distinguish two different objects (Figs. 4A and 4E).

Marr recognized the difficulty of simultaneously achieving stability and sensitivity and proposed as a solution a hierarchical series of structural descriptions, each one more sensitive but less stable than the last. For example, in a famous illustration he showed a human form represented at the highest level by a collection of six tubes of various sizes and aspect ratios (one for the head, one for the trunk, and two each for the arms and legs). At a lower level, an arm is represented as two tubes (one for the upper arm and one for the lower arm) connected at an obtuse angle. At a still lower level the lower arm includes two parts, one for the hand and another for the portion above the wrist, and at the lowest level the shape of the fingers presumably is represented in enough detail to distinguish between the hands of different people. When an object simply needs to be classified as human, the highest level representation can be used, providing stability across different human forms, whereas when a particular person must be identified, lower level representations come into play.

As is the case for other aspects of Marr's theory, hierarchical structural descriptions are intuitively

elegant but implementationally impractical. As an alternative, many theorists have proposed that the visual system computes and encodes both a structural description and a view-based representation for each object encountered. As we will see later, both neurophysiological and psychophysical studies indicate that object perception sometimes appears to be regulated by views and sometimes by structural descriptions, consistent with the dual-system approach. Nevertheless, stalwart theorists of both persuasions continue to develop models that attempt to explain all of these results using a single form of representation. Whereas they may someday be forced into accepting dichotomous representations, single-system theories tend to generate more testable predictions, so it would be imprudent to abandon these efforts at this time.

IV. NEUROPHYSIOLOGY OF OBJECT PERCEPTION

One way to study the nature of the processes and representations used in object perception is to examine the neural tissue of the visual system directly. Several means are available to do this: scientists may record from single neurons or selectively lesion brain areas in laboratory animals; they may work with human patients who have sustained damage to selective brain regions; or they may use functional brain imaging techniques to isolate the portions of normally functioning human visual systems that are engaged while performing various tasks. This section describes some of what we have learned about object perception via these methods.

A. Feature Analysis in Inferior Temporal Cortex

As noted earlier, the pattern of receptor firing rates on the retina constitutes the initial representation of an object. In a like manner, the center-surround ganglion cells and simple and complex cells in V1 can be considered increasingly complex object representations, with each cell coding for a particular feature (e.g., an edge tilted at a particular orientation) of the object. How are objects represented in later visual cortical areas?

Axons from neurons in V1 project to several other cortical areas, including (logically enough) areas V2 and V3. From here, visual processing separates into two streams. One, commonly referred to as the ''where'' pathway, flows dorsally, ending in the posterior parietal lobe, and determines the spatial locations and/or functional aspects of visual stimuli. The second stream, which flows ventrally toward the inferior temporal lobe, is responsible for determining object identity. The present discussion is confined to this latter ''what'' pathway, because we have defined object perception as the process of recognizing the identity of an object.

Relatively little is known about the function of areas V2 and V3, in part because, whereas V1 is easily exposed by surgically removing portions of the skulls of laboratory animals, V2 and V3 are hidden in folds of the occipital cortex. It is possible that these areas, as well as area V4, to which neurons from V2 and V3 project, provide the medium for further stages of representation in the ventral processing stream. We know more about the functional characteristics of cells in the inferior temporal lobe (area IT), which receives input from V3 and V4 and is the last purely visual processing area in the ventral stream. Researchers recording from single cells in various parts of IT have found neurons that respond best to complicated patterns of light. For example, one study found a neuron that responded with a rapid firing rate to a plastic model of a pineapple (Fig. 5A). Further investigation revealed that the neuron responded equally well to the leaves of the pineapple isolated from the rest of the object (Fig. 5B), and the researchers eventually established that the neuron's most rapid firing rate occurred in response to an eight-pointed star (Fig. 5C). This type of neuron was dubbed the ''elaborate cell,'' suggesting a more complicated building block for representations of whole objects, itself constructed from sets of complex cells in earlier visual areas.

Other studies have found neurons in IT that respond best to whole objects, most notably faces. Some neurons are remarkably selective, consistently responding at their fastest rate only when a particular familiar face is presented (e.g., the face of one of the experimenters). These findings seem, on the surface, to support the concept of ''grandmother cells,'' single neurons that are each responsible for signalling the presence of a single object, for example, one's grandmother. However, several considerations indicate that this notion is almost certainly misguided. Whereas a particular neuron might respond best to a particular face, it will also respond fairly well to other faces. Furthermore, most face-selective cells in IT (as well as most elaborate cells) respond best when the image of the face is of a particular pose (e.g., profile or full frontal), is in a particular orientation (most cells

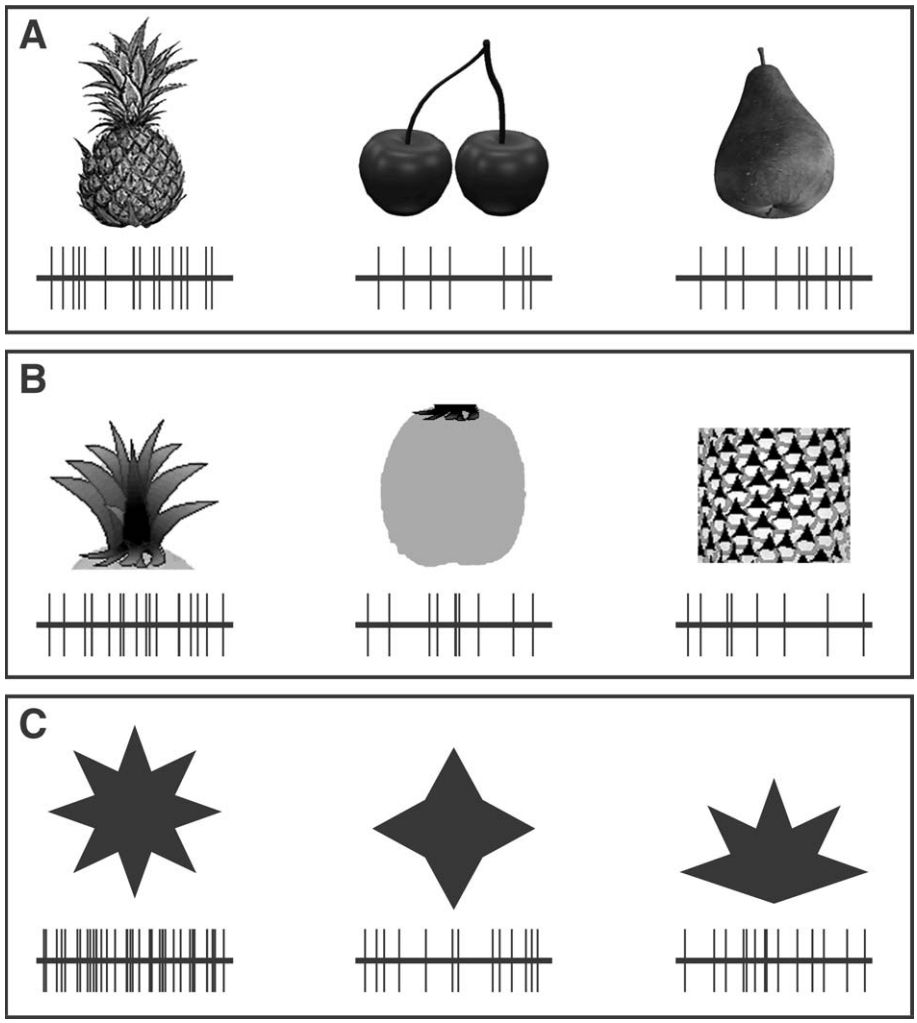


Figure 5 Schematic results from single-cell recording studies of neurons in cortical area IT. The vertical lines below each image show action potentials generated by the neuron while the image is shown to the animal subject (this figure illustrates the results of a study using macaque monkeys as subjects). Faster firing rates presumably indicate a greater preference by the cell for a given image. Results shown in (A) indicate that the cell responds better to an image of a pineapple than to other fruits. (B) indicates that the cell appears to be responding to the leafy part of the pineapple, not to the shape or texture of the bottom portion. Finally, (C) shows that the cell responds best of all to an eight-sided star shape.

respond best to upright faces, although some respond better to misoriented views), and takes up a particular retinal size (corresponding to a certain viewing distance). Thus, a neuron that responds better to grandmother's than auntie's face when they are presented from the front, tilted 30°, and viewed from 2 m, might reverse its preference if the faces were shown in profile, upright, and viewed from 10 m. The pattern of responses over a large number of IT neurons, rather than single cells, must be considered for object identity to be determined.

Because the majority of IT cells that have been found to respond best to a particular object also respond best when the object is seen from a particular orientation, view, and distance, the single-cell recording literature would seem, on the surface, to support view-based over structural description theories. That is, if area IT is the site of high-level object representations and neurons in IT show viewpoint-dependent responses, one might conclude that high-level representations are view-based. Again, though, patterns of responses over a population of neurons, rather than single neurons alone, represent objects, and it is possible that an appropriately selected population of IT neurons might respond to objects in a viewpoint-invariant manner. Furthermore, some single neurons

do respond to objects across a wide range of views, and the response properties of other neurons are completely unknown (indeed, in most studies only a relatively small proportion of neurons are found to selectively respond to any one object). At present, we can conclude that single-cell recording studies are more consistent with view-based than with structural description theories, but it is entirely possible that future studies may reveal evidence for structural descriptions (or some sort of currently unimagined hybrid representation) in area IT.

Before we leave the topic of single-cell recording, another feature of the receptive fields of IT cells should be noted: the receptive field size (the area of the retina to which a neuron responds) is greatly expanded for these cells compared to simple and complex cells in V1. V1 neurons, in turn, have larger receptive fields than retinal ganglion cells; more generally, there is a trend for receptive field sizes to increase as information gets farther along in the ventral processing stream. A natural consequence of this trend is the effective solution to translation constancy. Whereas a particular simple cell "looking" for a 45° oriented edge will only respond if SOGI's head falls within a small area, an IT cell that responds best to the SOGI as a whole will fire regardless of where in the visual field the object is located. However, it is important to note that, whereas the large receptive fields of IT cells indicate that translation constancy has been achieved by this point in processing, we do not yet know exactly *how* these cells come to respond to objects in any part of the visual field.

B. Neuropsychology of Object and Face Perception

In addition to using animal models, as described in the last subsection, it is also possible to examine the neurophysiological underpinnings of object perception with the help of humans who have sustained damage to visual areas of their brains. *Agnosia* is the general term for patients who have lost the ability to visually recognize objects. Such patients retain normal visual acuity and memory—they can see objects and remember object labels (as well as their structural and functional properties)—but they cannot retrieve an object's name from its image. Even more remarkably, patients with *associative* forms of agnosia may be able to produce excellent drawings of objects from memory, but fail to name their own drawings when shown them later.

Associative agnosic patients show myriad patterns of spared and impaired object perception abilities. Patients with *alexia* are unable to recognize letters, but recognition of many other objects appears normal. *Prosopagnosia* is a selective inability to identify individual humans from their faces with, again, many other types of objects remaining recognizable. Other patients have category-specific agnosias, only showing difficulty naming fruits and vegetables, animals, or some other class of objects. Still other patients show even more idiosyncratic deficits, for example, one well-known patient is excellent (by some tests, better than normal) at recognizing people by their faces but is severely agnosic for all other object classes.

The dissociation evidenced in the agnosia literature between the perception of faces and other objects (a number of patients showing impaired face recognition abilities with relatively intact object recognition, along with at least one patient showing the opposite pattern of abilities and disabilities) is especially interesting to object perception theorists. All faces have the same basic parts in similar relations to each other, so that it would be difficult to craft a structural description that could distinguish different faces from each other. Thus, a number of researchers have drawn the inference that faces are perceived by a visual subsystem that uses view-based representations and is damaged in prosopagnosic patients, whereas other objects are perceived by a structural-description-based subsystem that is relatively spared in prosopagnosia.

Several neuroimaging laboratories using techniques such as functional magnetic resonance imaging (fMRI) have showed that greater cellular activity is generated in a particular portion of IT cortex (more specifically, the anterior fusiform gyrus) when humans view pictures of faces than when pictures of other objects are viewed. The tantalizing conclusion is that this is the area of the brain where view-based representations of faces are processed and/or encoded and that selective injury to this area results in prosopagnosia. This hypothesis will be evaluated in greater detail later after we have considered behavioral evidence concerning the different forms of processing inherent in the perception of faces and other types of objects.

V. BEHAVIORAL STUDIES OF OBJECT PERCEPTION

The previous section described research by physiological psychologists and neuroscientists on how

different neurons and different parts of the brain contribute to the process of object perception. Other cognitive scientists have studied object perception through behavioral experiments on human subjects with normally functioning visual systems. In these studies, subjects are asked to perform a task on a selected set of stimuli, for example, naming pictures of common objects. Some aspect of the stimuli (e.g., the viewpoint from which each stimulus is pictured) or the task (e.g., the duration of time that the stimulus remains on the screen) is varied, and subjects' responses are recorded. By examining the relationship between the manipulated variable and responses, we can indirectly learn something about how the brain is processing the stimuli in order to perform the task.

A. Effects of Image Transformations

One important set of behavioral studies involves psychophysical experiments, where psychological responses to variations in a physical property of object stimuli are recorded. The basic logic of these experiments is that, if a certain stimulus dimension is encoded in the representations used for object perception, then transformations to stimuli along this dimension should cause impairments of performance (i.e., increased response times and/or increased error rates) in object perception tasks.

To take a well-known and oft-replicated example, many researchers have shown subjects pictures of common objects in different orientations, ranging from the object's canonical upright position to an upside down orientation. Example stimuli and the typical pattern of results from these experiments are shown in Fig. 6. In the initial block of trials, subjects usually require more time to recognize an object the farther it is rotated from its upright orientation. However, there is often a small dip in the response time function at 180° (the upside down position), as shown in the graph. Another important feature of these studies' results is that, if the experimental trials are repeated, subjects take considerably less time to recognize the objects in misoriented orientations—the slope of the response time function decreases with practice.

These findings indicate that something about objects' canonical orientations is encoded in the representations used for object perception. This conclusion helps to constrain both view-based and structural description theories. The fact that response time varies with degree of misorientation is explained in many

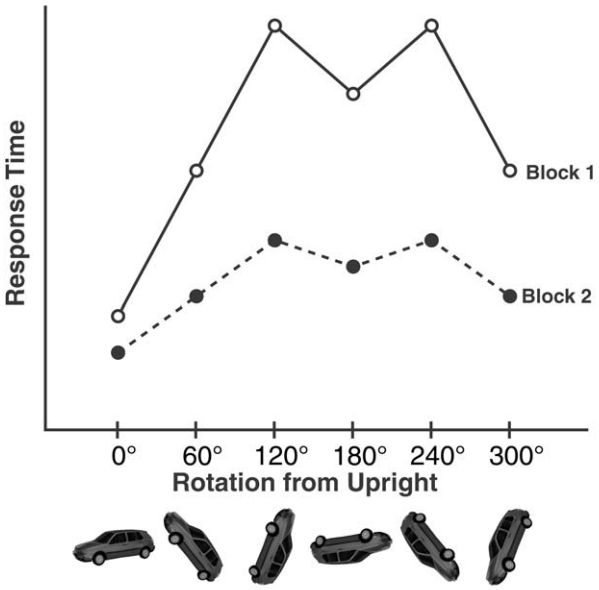


Figure 6 Schematic results from psychophysical experiments in which common objects, such as the car shown at bottom, are presented in various orientations and must be named as quickly as possible by subjects. Three results are evident in the graph: (1) naming response time generally increases as the object is rotated away (either clockwise or counterclockwise) from its canonical upright orientation; (2) there is a small dip in the response time function at 180°; and (3) the slope of this function is reduced if the stimuli are seen again in a second block of trials.

view-based theories as a consequence of a trasformation process necessary to bring a misoriented view of an object into register with the encoded, canonical view. The larger the degree of misorientation, the longer this transformation process (which is likened to mental rotation in some theories) takes, and the longer it will take subjects to recognize the object. To explain the decrease in misorientation effects over test blocks, these theories propose that multiple views of an object can be stored in memory. Once an object is seen lying on its side in the first block of test trials, a representation is encoded of the object in that orientation. When this stimulus is encountered in subsequent blocks of trials, the transformation process is no longer necessary because the test object can be directly compared to the newly encoded representation of the object on its side.

In structural description theories, orientation effects are explained as a function of shifting categorical relations between parts. The structural description of a car will include the proposition that the windshield is ABOVE the front wheel. When the car is rotated 90° in

the picture plane though, the wheel will be BESIDE the windshield, causing, as in view-based theories, an imperfect match between perceived and encoded representations, which leads to an increase in response times. The same type of mechanism also provides an interesting account for the dip in the response time function at 180°. Parts that had BESIDE relationships in upright objects (e.g., the front and rear wheels of the car, assuming the car is viewed from the side) shift to ABOVE–BELOW relations when the object is rotated 90°, but return to BESIDE relations when the object is fully upside down. Thus, in some cases the structural description of a 180° rotated object matches the upright object better than does the 90° rotated version.

Note that this account implies that structural descriptions do not include information as to whether one part is to the left or to the right of another—both LEFT and RIGHT relations are coded as BESIDE. If this is the case, then two images that are mirror reflections of each other should share the same structural description, and image reflections should have little, if any, effect on recognition accuracy or response time. As noted in Section I, this pattern of reflection invariance has, in fact, been empirically demonstrated.

Other experiments indicate that size transformations and object translations also generally have minimal effects on object perception responses. Effects of lighting direction are only starting to be studied systematically (before the ready availablility of 3D modeling software, it was difficult to generate stimuli for such experiments), with early results suggesting that lighting changes also cause only small, though systematic, disruptions to object perception processes. Object perception across rotations in the depth plane (i.e., changes in viewpoint) are also the focus of increasing scrutiny since the development of inexpensive computer modeling programs. Consistent with the research on picture plane rotations, most studies have found systematic and substantial effects of viewpoint shifts on object perception response time and error rates.

The relative invariance of object perception over mirror reflection, size, position, and lighting changes appears to favor structural description over view-based theories because, as noted in Section I, these transformations all lead to large changes in images (which are the presumed representations in the simplest view-based theories). However, structural description theories also predict at least limited viewpoint invariance, because structural descriptions should only be altered by a viewpoint shift when parts become occluded or uncovered or when categorical relations are altered. Careful investigation has shown that, even when part accretion and deletion are minimized, viewpoint costs to object perception performance are still incurred. Small picture plane rotations, which should also have no effect on categorical relations, can also have measurable effects on recognition time. Thus, view-based models appear to provide a better account than structural description theories for findings of orientation and viewpoint dependence.

The description of psychophysical research given here is, by necessity, oversimplified. There are numerous exceptions to these generalizations (e.g., object perception has been shown to be orientation-invariant in some circumstances and size-dependent on others), and many investigators have seized upon this ambiguity to argue that view-based representations are used in some object perception contexts and structural descriptions are used in others. Perhaps the most pervasive proposition is that recognition in contexts where all candidate objects are easily discriminable from each other (e.g., distinguishing between a car, a bicycle, and a jogger) relies on structural descriptions, whereas object perception in more demanding circumstances (identifying a sparrow in an area where warblers and finches also live) requires view-based representations. As intuitive as it appears to be, experiments designed to directly test this particular dual-system hypothesis have provided mixed evidence; some studies support it but others do not. Thus, the more exciting possibility remains that either a single view-based or structural description theory, or a model employing some type of hybrid representation, will eventually be able to explain all of the observed patterns of dependencies and invariances across image transformations.

B. Subordinate-Level Classification and Face Perception

In addition to the psychophysical experiments reviewed earlier, a second collection of behavioral experiments has examined the relative ease of object perception when stimuli must be identified at different classification levels. The *entry* level for any given object is operationally defined as the first name that comes to mind for that object (e.g., dog); *subordinatate*-level labels are more specific (beagle), and *superordinate* labels are more general (animal). Numerous studies have demonstrated that common objects are easiest to

classify at the entry level. These experiments typically use a label verification task, in which a word is followed by a picture of an object on each trial and the subject decides whether the word and picture match. Response times are faster and error rates lower for entry-level than for suborinate- and superordinate-level labels.

It is generally held that this entry-level advantage is at least partially perceptual in nature, that is, object perception processes are specialized to produce representations that are most efficient for identifying objects at the entry level. Neuroimaging studies have supported this assumption by showing that the extra time taken to identify objects at the subordinate level is due to additional processing by the visual system, presumably supplementing the object perception processes that allow for entry-level identification.

The most subordinate level of identification is the individual level (Snoopy), and, as mentioned already, a large literature is devoted to individual-level recognition of one particular class of objects: faces. Much of this literature is dedicated to demonstrating that face perception involves qualitatively different processes than the perception of other objects. The most well-known finding in this regard is the *inversion effect*. As already discussed, almost all objects are more difficult to recognize when inverted (turned upside down) than when upright, but the increase in difficulty for faces is especially pronounced. Faces are also more difficult to recognize than other objects when the top and bottom halves of images are misaligned (Fig. 7 demonstrates both the inversion and misalignment effects).

Although there are a number of hypotheses concerning exactly how faces are processed, most agree on the basic premise that a precise specification of the configuration of face parts is crucial for face perception. In other words, a face representation must include not just the information that the nose is ABOVE the mouth but must specify the exact distance by which these parts are separated. Such precision is not necessary to discriminate between members of most other classes of objects, such as the chairs in Fig. 7.

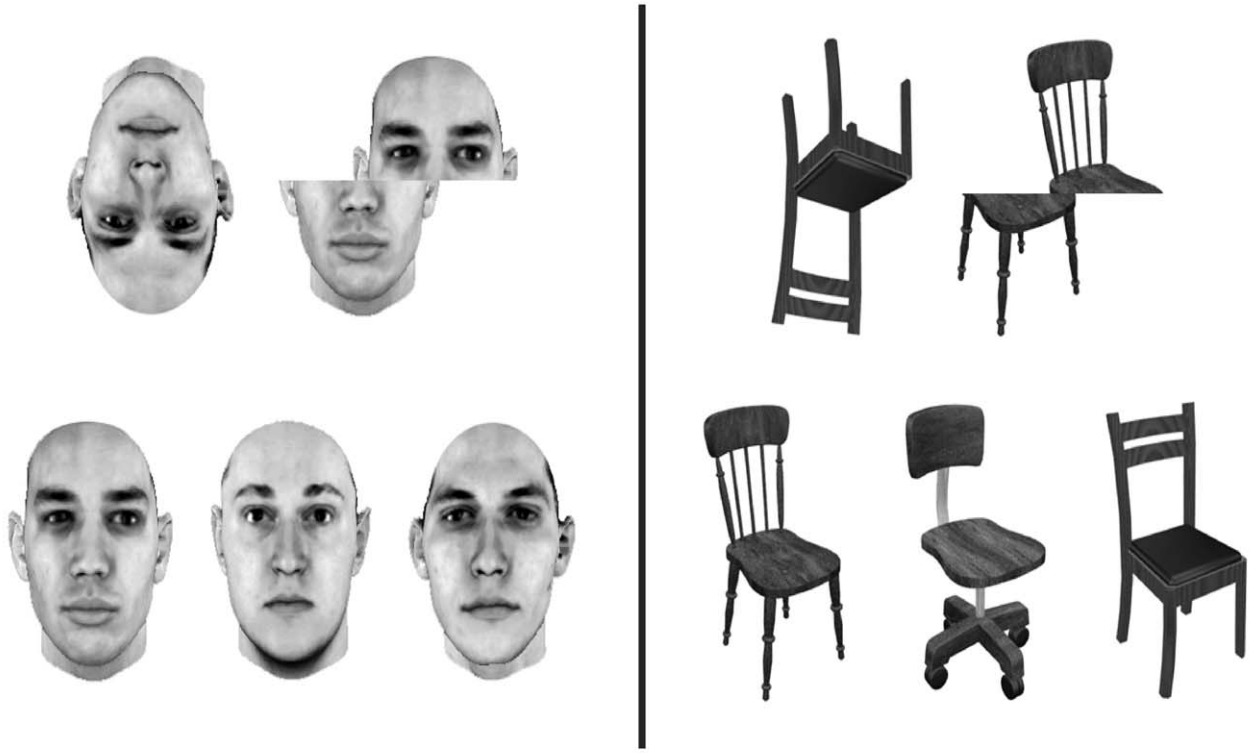


Figure 7 Face perception can be dissociated from the perception of other classes of objects by a number of behavioral effects, two of which are demonstrated here. It is quite difficult to pick out the upright faces in the bottom left triad of images that match the inverted and misaligned faces in the upper left of the figure. By contrast, matching upright with inverted and misaligned chairs is relatively easy, as shown at right.

Many researchers believe that configural processing is carried out in the anterior fusiform face area identified in the neuroimaging studies described earlier. A lively debate is currently underway in the literature concerning the nature of the processing going on in this area, most crucially whether it is a truly face-specific processing system or whether it instead can contribute to the recognition of other objects at subordinate levels. Prosopagnosic patients can recognize some types of objects at subordinate levels (one patient was even able to learn to recognize individuals from his flock of pet sheep), but careful testing shows that they are generally poorer at subordinate- than at entry-level identification.

Other research points to expertise as a crucial factor in face-specific behavioral effects. Dog show judges and bird watchers show inversion and other face-specific effects for their domains of expertise, and exciting research has shown the anterior fusiform responding preferentially to a class of novel objects only after subjects were trained to be experts at recognizing these objects at the individual level. All normally developing humans are face perception experts, so it is possible that the anterior fusiform is specialized for processing any class of objects that is both recognized at a highly subordinate level and highly overlearned. Whether this processing involves representations that are qualitatively or only quantitatively different from those involved in entry-level object perception remains a subject for future research.

See Also the Following Articles

ARTIFICIAL INTELLIGENCE • ATTENTION • CONSCIOUSNESS • INFORMATION PROCESSING • MEMORY, NEUROIMAGING • MOTION PROCESSING • MULTISENSORY INTEGRATION • PATTERN RECOGNITION • PROSOPAGNOSIA • SALIENCE • SPATIAL COGNITION

Suggested Reading

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychol. Rev*. **94,** 115–147.

Farah, M. J. (1990). *Visual agnosia: Disorders of Object Recognition and What They Tell Us about Normal Vision*. MIT Press, Cambridge, MA.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Freeman, San Fransisco.

Palmer, S. E. (1999). *Vision Science: Photons to Phenomenology*. MIT Press, Cambridge, MA.

Tarr, M. J., and Pinker, S. (1989). Mental rotation and orientation dependence in shape recognition. *Cogn. Psychol*. **21,** 233–282.

Ullman, S. (1996). *High Level Vision*. MIT Press, Cambridge, MA.

Occipital Lobe

EDGAR A. DEYOE
Medical College of Wisconsin

GLOSSARY

cerebral achromatopsia The inability to perceive and discriminate colors due to pathology within the cerebral cortex. It is to be distinguished from the more common form of color blindness caused by anomalies in the photoreceptors of the eye.

cerebral akinetopsia A rare deficit in the ability to perceive and discriminate movement thought to arise from lesions of cortical visual areas involved in the processing of visual motion information.

functional magnetic resonance imaging (fMRI) Several neuroimaging techniques capable of showing areas of the brain that are activated during sensory, motor, or cognitive tasks. The most common technique highlights areas of the brain in which blood flow and oxygenation are altered by the experimental task. fMRI images are typically displayed as an overlay on conventional structural MRI images depicting the anatomical structure of the brain.

positron emission tomography (PET) Like functional MRI, PET imaging is capable of showing areas of the brain that are active during sensory, motor, or cognitive tasks. PET is based on the injection and measurement of radioactive tracers whose concentrations are altered at brain sites where activity is experimentally altered. Depending on the tracer, this technique can be sensitive to changes in oxygen or glucose utilization or other chemical factors.

prosopagnosia Deficit in the ability to recognize previously familiar faces. It is thought to be associated with lesions of ventral occipitotemporal cortex.

receptive field That region of the retina (one in each eye), or equivalently the visual field, within which a stimulus is able to alter the response of a visual system neuron.

retinotopy The representation of retinal topography within the connections, nuclei, and cerebral cortex of the visual system. A retinotopic map is a collection of neurons whose receptive fields, in the aggregate, preserve the two-dimensional topography of the photoreceptors that provide their input, albeit to a variable degree of precision and completeness depending on the brain site. A retinotopic map may be distorted, typically due to expansion of the area devoted to the visual field near the center of gaze.

stereopsis The phenomenon of simultaneous vision with two eyes in which there is a vivid perception of the distances of objects from the viewer; it is present because the two eyes view objects in space from two points, so that the retinal image patterns of the same object are slightly different in the two eyes.

V1, V2, V3, VP, V4, hMT+, V8, FFA, LO, and KO Abbreviated names of human cortical visual areas. V1 is the primary visual cortex, also known as striate cortex. Other visual areas are collectively referred to as extrastriate visual cortex. Some extrastriate visual areas are simply numbered (e.g., V2, V3, V4, and V8); others have more descriptive names. VP, ventral posterior visual area; hMT+, middle temporal[1] visual area, also known as V5; FFA, fusiform face area; LO, lateral occipital visual complex; KO, kinetic occipital region.

visual field That portion of the world visible to the observer at any given instant. The foveal visual field refers to the portion of the visual field near the center of gaze (strictly that portion imaged on the fovea of the retina).

The occipital lobe encompasses the most posterior portion of the human cerebral cortex and is primarily responsible for vision. Direct electrical stimulation of the occipital lobe produces overtly visual sensations. Damage to the occipital lobe typically results in either complete or partial blindness or visual agnosias, depending on the location and severity of the lesion. This article will summarize the structure and function

[1]Although hMT+ derives its name from the homologous monkey visual area, its location is near the junction of the occipital, parietal, and temporal lobes on the lateral aspect of the brain, not in the middle of the temporal lobe as its name would seem to imply.

of the human occipital lobe, stressing its role in vision. Although the discussion is primarily based on studies of normal human subjects, clinical reports of brain-damaged patients and animal studies are discussed when they directly illuminate particular issues.

I. INTRODUCTION

The surface area of the human occipital lobe encompasses approximately 12% of the total surface area of the neocortex of the human brain. Its primary role is to provide the sense of vision. Vision begins with the spatial, temporal, and chromatic pattern of light falling onto photoreceptors of the retina and culminates in the perception of the properties of the objects and surfaces within the world around us. Fundamentally, then, vision is the ability to infer the attributes of objects in the visual scene from an analysis of the light patterns imaged in the eyes. The occipital lobe contains the bulk of the neural "machinery" responsible for this. However, visual perception can be affected by the context of previous experience as well as by current goals and expectations. To permit these more cognitive functions, the occipital lobe is heavily interconnected with other lobes of the brain, especially the parietal and temporal lobes, as well as with an array of subcortical structures. Through these connections, the results of visual processing in the occipital lobe enter into and can be influenced by more general processes involved in goal-directed, motivated behavior and "thinking." With these considerations in mind, the following discussion begins by describing the structure and function of the occipital lobe as a separate entity but ends with a consideration of its potential role in more cognitive functions whose neural basis may include other brain regions.

II. ANATOMICAL ORGANIZATION

Anatomy provides a description of the physical structure of the occipital lobe but can also provide a basis for identifying functionally distinct subdivisions. Although anatomy is not infallible in this respect, the discovery of new anatomical distinctions can trigger the discovery of previously unsuspected functional differences. Accordingly, the following discussion begins with descriptive anatomy but then introduces anatomical features and principles that have been useful as indicators of functional differentiation.

A. Descriptive Anatomy

1. Gross Anatomy of the Occipital Lobe

Figure 1 illustrates the location and extent of the human occipital lobe. Viewed from the medial surface of the cerebral hemisphere, it is bounded anteriorly by the parieto-occipital sulcus (POS). Ventrally and laterally there are no major anatomical features that clearly and consistently demarcate the anterior extent of the occipital lobe. For practical purposes, an imaginary line running laterally from the dorsal tip of the POS to the preoccipital notch is considered the effective boundary separating the occipital lobe from the parietal and temporal lobes.

Medially, the most prominent feature of the occipital lobe is the calcarine sulcus, which is flanked above by the cuneus and below by the lingual gyrus (cf. Fig. 1, top). The lingual gyrus is separated from the more laterally placed fusiform gyrus by the collateral sulcus. The fusiform gyrus is then bounded laterally by the occipitotemporal sulcus (OTS), although this sulcus tends to be variable and interrupted as it extends posteriorly toward the occipital pole. The lateral surface of the occipital lobe is especially variable from individual to individual. Consequently, the lateral occipital gyrus can be irregular and can be split by the lateral occipital sulcus. This latter fissure has several important variants, sometimes appearing as one, two, or even three small sulci running approximately in the anterior–posterior direction. Most dorsally on the lateral surface is the transverse occipital sulcus, which often forms the most posterior end of the intraparietal sulcus. A thorough treatment of individual variants of the occipital anatomy can be found in the atlas of Ono, Kubik, and Abernathy.

2. Cytoarchitecture, Myelination, and Histochemistry

Traditionally, occipital cortex has been subdivided into anatomically distinct regions based on cytoarchitecture. As for all neocortex, the occipital gray matter is clearly laminated, though the criteria for identifying and naming different layers have varied considerably. Figure 2A illustrates one of the more popular laminar numbering schemes. Primary visual cortex is the most cytoarchitecturally distinct region of the occipital lobe. It is also known as striate cortex due to the stria of Gennari, which can be identified even in a cross section of freshly cut tissue. This band of myelinated axons running horizontally in layer 4B demarcates the extent

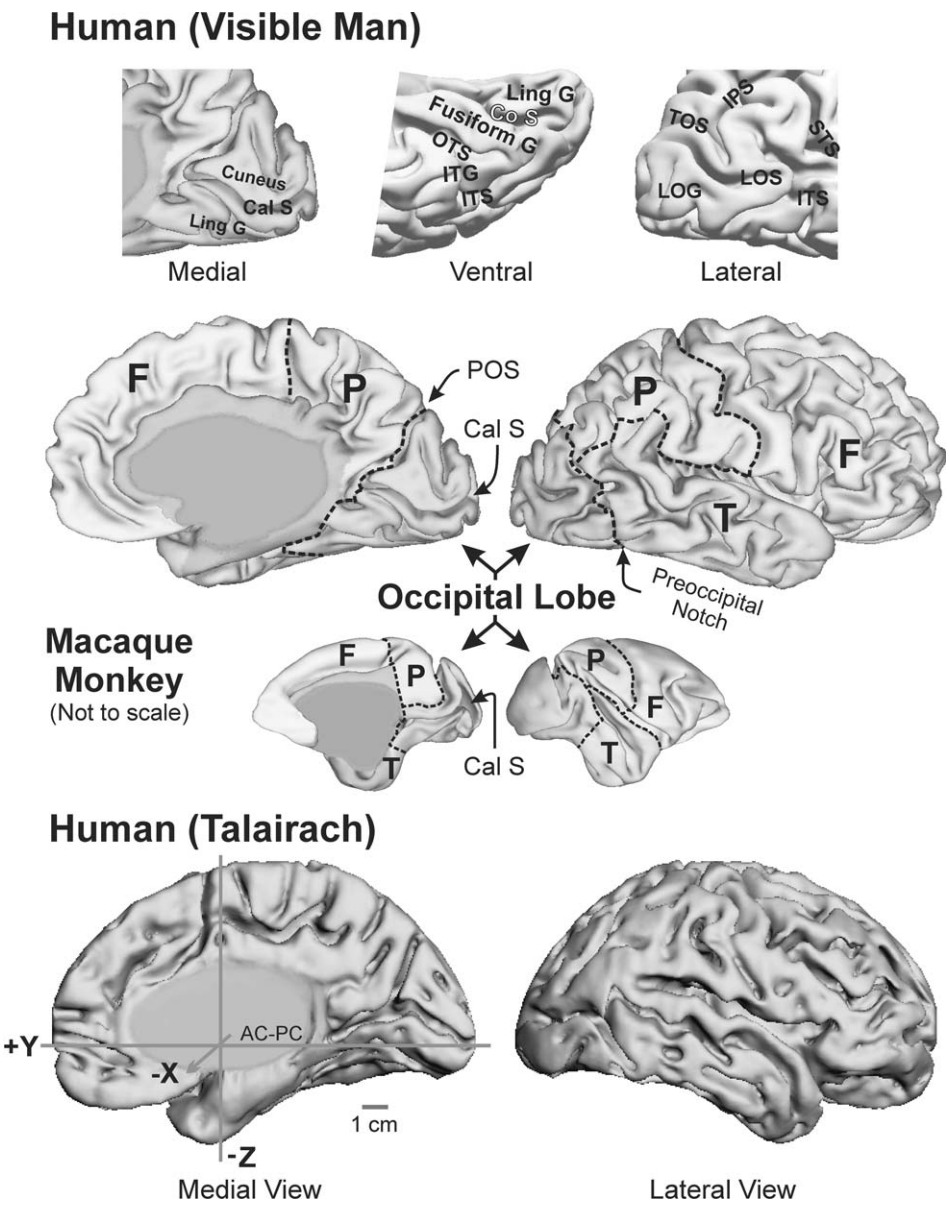


Figure 1 Gross anatomy of the occipital lobe in humans and macaque monkeys. Top: Medial, ventral, and lateral views of the occipital lobe with medial and lateral views of the whole cerebral cortex (below) from the Visible Man database. Middle: Identification of cortical lobes in the macaque monkey. Color code same as for top. Bottom: Computer brain model reconstructed from the Talairach and Tournoux *Co-Planar Stereotaxic Atlas of the Human Brain* showing the standard Talairach coordinate system, with the origin located at the anterior commissure (not shown) and the *y* axis passing through both the anterior and posterior commissures (not shown). Polarity of axes indicates conventions used in this article. Abbreviations: AC–PC, anterior commissure–posterior commissure reference line; Cal S, calcarine sulcus; POS, parieto-occipital sulcus; F, frontal lobe; P, parietal lobe; T, temporal lobe. (Figures of Visible Man and Macaque monkey brains courtesy of David Van Essen.)

of striate cortex within, and adjacent to, the calcarine sulcus (Fig. 2B). Cytoarchitectural differences among visual areas outside the striate cortex (known as extrastriate cortex) tend to be less obvious and more inconsistent, perhaps accounting for significant differences in the accounts of cytoarchitectonic parcellation of occipital cortex by early investigators.

One of the most widely used cytoarchitectonic schemes for subdividing cerebral cortex has been that of Brodmann (Fig. 2D). His scheme was primarily

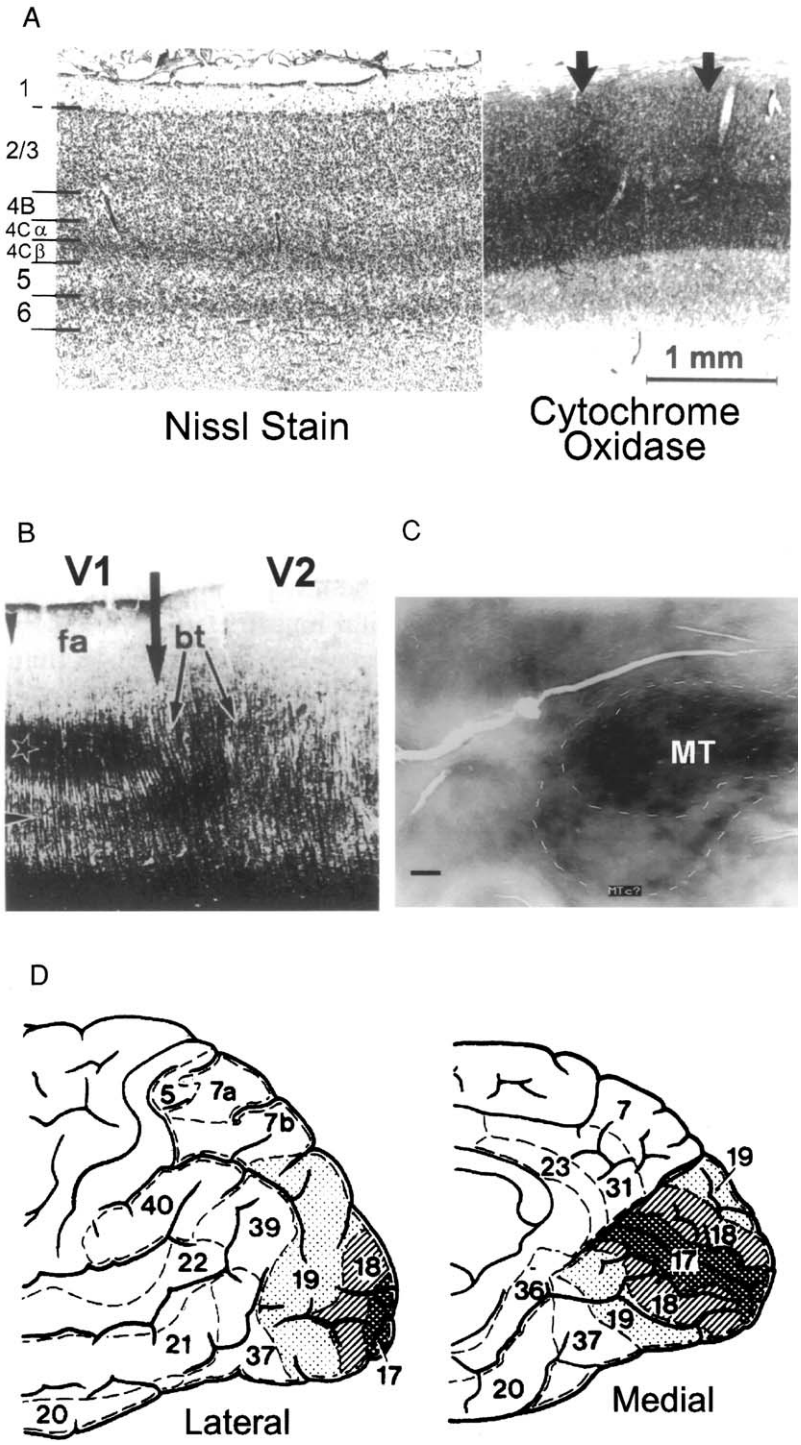


Figure 2 (A) Example of the laminar architecture of primary visual cortex visible with common Nissl stain (left) and with histochemical stain for cytochrome oxidase activity (right). Arrows indicate cytochrome oxidase dense puffs in layers 2–3. Note that different stains reveal different features. (Adapted from Horton, 1992.) (B) Example of myeloarchitecture near the border of areas V1 and V2 marked by a large vertical arrow. Abbreviations: bt, border tuft; fa, fringe area. A star marks the stria of Gennari. Small arrowhead on the left marks the inner band of Baillarger. (Adapted from Zilles and Clarke, 1997.) (C) Illustration of dense cytochrome oxidase histochemistry in human MT visual area. (From Tootell and Taylor, 1995.) (D) Brodmann's parcellation (numbers) of the human occipital lobe and adjacent cortex based on cytoarchitectonic features. (From Zilles and Clarke, 1997.)

based on differences in cell morphology (e.g., stellate–pyramidal), density, and distribution, as well as the overall thickness, proportion, and development of different layers within the cortical mantle. According to this plan, striate cortex is designated area 17 and is immediately surrounded by area 18, which in turn is bounded by area 19. Portions of Brodmann's area 37 may also be included in what we have defined here as occipital cortex (though Brodmann considered it part of the temporal cortex). Except for visual areas 17 and 18, Brodmann's areas often correspond poorly with functional distinctions identified by more modern techniques such as neuroimaging. This has resulted in diminished use of Brodmann's nomenclature as a basis for describing the functional organization of visual cortex.

Although some of Brodmann's original criteria for differentiating occipital cortex have not been functionally diagnostic, other cytoarchitectural features have been found to fare better in some cases. Indeed, researchers have argued in favor of a renewed interest in architectonic data as a useful correlate of function. Histological stains that reveal the distribution of myelinated fibers within the cortical gray matter have been successful in delineating some functionally defined visual areas. This is certainly the case for the stria of Gennari, which distinguishes the primary visual cortex. Dense myelin staining also characterizes the middle temporal visual area, hMT+, located in lateral occipital cortex near the confluence of the occipital, temporal, and parietal lobes (Fig. 2C). In macaque monkeys, the third visual area, V3, can also be identified by heavy myelin staining, though it is not clear whether this is also true in humans.

One alternative to the study of cytoarchitecture that can provide functionally relevant information is to examine the distribution of chemically distinct molecules within the cortex. This has been accomplished through a wide variety of histochemical and immunological techniques, though, historically, enzyme visualization has been especially useful in this respect. The discovery of circumscribed zones of high cytochrome oxidase activity in striate cortex (puffs, blobs) and extrastriate cortex (V2 "stripes", V4 "patches") triggered the identification and characterization of functionally distinct "modules" within occipital visual areas that had previously been thought to be functionally homogeneous. These discoveries led to the concept that individual cortical visual areas can contain multiple, distinct processing pathways or streams, thereby extending (though not necessarily reiterating) the organizational principle of multiple processing pathways found in the retina and LGN (outlined later). To a large extent, the features of cortical organization revealed by these studies are on a spatial scale that, until very recently, had been beyond the resolution of human neuroimaging and electrophysiological techniques, and so are outside the scope of this article.

B. Connectivity

Our knowledge of the anatomical connectivity of human occipital cortex largely arises from inferences based on animal experiments. In animals, differences in connectivity within the occipital lobe provide an important key to identifying different functional subdivisions and their interrelationships. To the extent that connectivity information derived from animals accurately reflects connectivity in the human brain, it provides a valuable basis for interpreting the results of neuroimaging studies and other types of human data.

1. Visual Input to the Occipital Lobe

Figures 3A and 3B summarize the major thalamic projections to occipital visual cortex. Visual input from the retina is relayed through the lateral geniculate nucleus (LGN) of the thalamus to terminate in the primary visual cortex (striate cortex, area 17, V1). A second major pathway arises from retinal projections that bypass the LGN and project to the superior colliculus. The superior colliculus then projects to the thalamic pulvinar nucleus, which in turn distributes widely to the cortex of the occipital lobe.

The geniculocortical pathways are subdivided into three main components associated with different neuronal subclasses in the retina and LGN. One pathway originates from small P-cells in the retina and is relayed through the parvocelluar layers of the LGN to terminate in layer 4Cβ of V1 and more sparsely in layers 4A and 6 (not illustrated in Fig. 3B). A second pathway to striate cortex begins with large M-type retinal ganglion cells and is relayed through the magnocellular layers of the LGN to terminate in layer 4Cα accompanied by a light projection to layer 6 (not shown in Fig. 3B). A third source of projections to V1 comes from a class of small cells within the koniocellular (K) or intercalated layers of the LGN. These neurons typically receive input from both the retina and the superior colliculus and terminate in the supragranular layers (above layer 4) of striate cortex. These afferents tend to cluster into "pufflike" regions of layers 2–3 that are also unique for their high levels of the metabolic enzyme cytochrome oxidase (co).

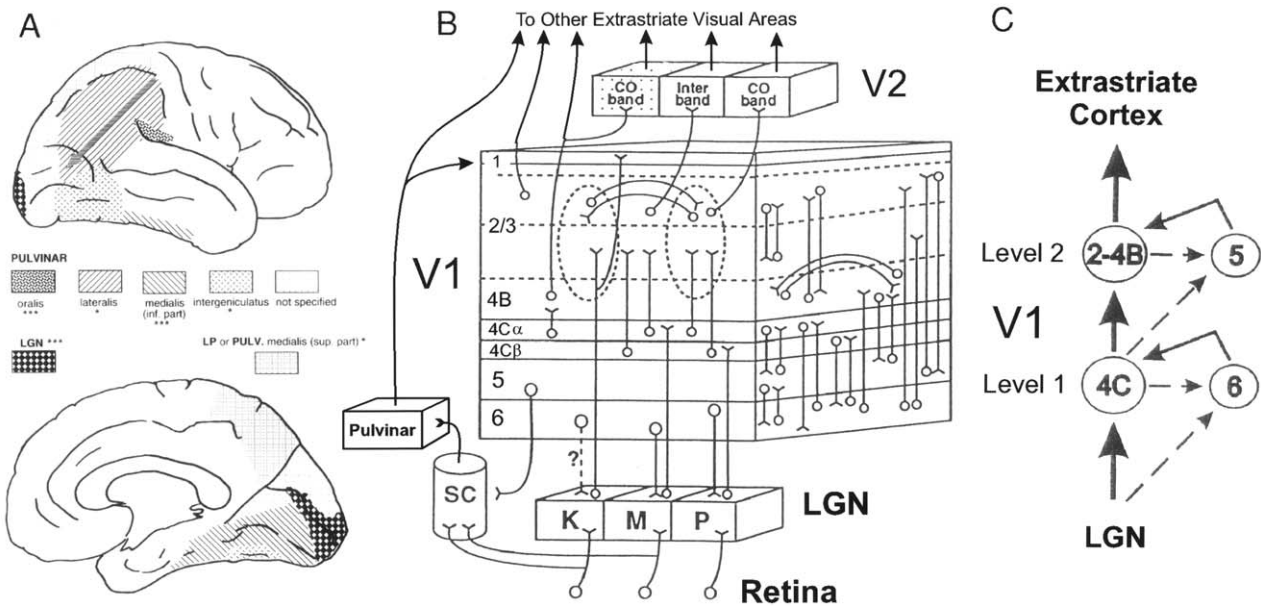


Figure 3 (A) Distribution of thalamic inputs to the occipital lobe and nearby portions of the parietal and temporal lobes. (From Zilles and Clarke, 1997.) (B) Concurrent input to V1 from K, M, and P retinogeniculate pathways with a schematic of intrinsic circuitry and cytochrome oxidase defined puffs (dashed ellipses). V2 receives segregated input from different sets of output neurons in V1 and distributes projections to different sets of extrastriate visual areas via distinct populations of output neurons in the different cytochrome oxidase defined compartments. Such circuitry forms the anatomical basis of multiple concurrent processing pathways within and among occipital visual areas. (Adapted from Casagrande and Kaas, 1994.) (C) Functional interpretation of V1 circuitry. (Adapted from Callaway, 1998.) Abbreviations: LGN, lateral geniculate nucleus; sc, superior colliculus; K, koniocellular; M, magnocellular; P, parvocellular.

Afferents to striate cortex via the alternate tectopulvinar pathway terminate most heavily in laminae 1 and upper 2–3.

Neurons of the P, M, and K, pathways differ in their visual response properties. Neurons of the parvocellular path tend to be relatively more numerous in the central retina, have smaller receptive fields with sustained responses, and are often color opponent. This makes them well-suited for conveying information about fine spatial detail and color. Magnocellular neurons tend to have larger receptive fields, prefer low spatial frequencies, and have more transient responses. They do not discriminate wavelength differences well but tend to have the greatest sensitivity to luminance contrast, especially at low spatial frequencies and low luminance levels. They are optimized for processing rapid temporal changes such as flicker and movement. The function of K-pathway neurons has not been studied as extensively as that of the M and P systems, but they appear to be a major determinant of co–puff cell properties and may play a key role in the modulation of responses evoked by the M and P pathways.

Although these three pathways are distinct, their functional capabilities can overlap significantly. It is primarily at the extremes of spatial and temporal frequencies that the M- and P-cell capabilities are most different. Both M and P pathways may contribute to processing intermediate spatial and temporal frequencies, but P-cells will tend to dominate for high spatial frequencies that are slowly varying. Conversely, magnocellular neurons will respond better than P-cells at low spatial frequencies that are rapidly varying. Chromatically, P-cell color opponency (activation by some wavelengths and suppression by others) appears to be critical for hue discrimination, but M-cells can respond more strongly to some wavelengths than others, though they are not opponent and their tuning is typically broader than that of P-cells. The significance of this functional overlap is that, under natural viewing conditions, all three inputs to the visual cortex are likely to be concurrently active. Under extreme visual conditions, the specialized capabilities of one or another of the pathways may take over and thereby extend the range of useful sight.

2. Intrinsic Circuitry

The retinal information relayed through the LGN is processed further by the intrinsic circuitry of V1 and

then distributed to other cortical areas via output neurons in layers 2–3 and 4B (see Fig. 3B). Subcortical projections to the superior colliculus and feedback projections to the LGN originate from layers 5 and 6, respectively. It is through this selective distribution of visual information combined with the characteristics of local processing that different visual areas achieve their functional uniqueness. Although a detailed review of the local circuitry of the occipital cortex is outside the focus of this article, Figs. 3B and 3C attempt to summarize the basic plan for V1. The intrinsic processing of V1 can be divided into two major levels. Input from the LGN terminates at the first level in the different subdivisions of layer 4C. Information is then distributed primarily to level 2 output neurons within layers 2–3 and 4B. The latter output signals are modified by neurons in layers 5 and 6 that combine information about the inputs and outputs of each processing level and feed it back onto neurons at the same level.

The laminar organization of V1 circuitry is complemented by horizontal connections that tend to connect functionally similar zones distributed within specific laminae. In layers 2–3, long horizontal projections preferentially interconnect pufflike zones containing cells that have similar visual response properties and that are rich in the metabolic enzyme cytochrome oxidase. Likewise, interpuff zones tend to be interconnected most strongly with other interpuff zones. In V2, a similar pattern of specific horizontal connections link co-defined compartments termed the thick-, thin-, and interstripe zones.

Together, the laminar specificity of vertical intrinsic connections and the functional specificity of horizontal connections provide an anatomical substrate for creating different subsets of output neurons whose visual response properties reflect different combinations of the P, M, and K afferent pathways, as well as modulatory effects of cortical feedback pathways and other subcortical inputs. Output neurons in layer 4B of V1 are dominated by M-pathway characteristics, as are the responses of the extrastriate visual areas, such as V2 thick stripes and area MT, to which the 4B cells project. In contrast, output neurons in the interpuff regions of V1 tend to be strongly biased by P-cell characteristics, which are subsequently imparted to downstream visual areas such as V2 (thin- and interstripe compartments), VP, and V4. Output neurons in co–puff subdivisions tend to exhibit characteristics that appear to reflect a mixture of influences. As a result, later processing stages, such as V4, also can be shown to exhibit a mixture of influences. Overall, then, the intrinsic circuitry of V1 acts to process, combine, and redistribute the visual information available in the different afferent pathways, thereby creating a new set of output signals that reflects the afferent inputs but that also encodes more complex visual properties or features. Intrinsic circuitry within subsequent cortical processing stages, though less well-studied, presumably functions in a similar manner.

3. Corticocortical Connectivity

The output neurons of layers 2–3 and 4B of V1 selectively distribute different types of visual information to subsequent processing stages in extrastriate visual cortex. Each extrastriate visual area itself has unique feedforward and feedback connections within and, often, outside the occipital lobe. These patterns of selective connectivity are a primary determinant of the functional specificity of each visual area. Finally, interhemispheric projections through the corpus callosum connect the left and right halves of each visual area into a functionally integrated whole.

In monkeys, connections among cortical visual areas tend to follow a consistent pattern of laminar distribution that differentiates forward, backward, and lateral types of connectivity (Fig. 4). Generally, forward connections arise from neurons in layers 2–3 of the lower area and distribute to layer 4, tapering off into the lower reaches of layer 3 in the higher visual area. In some visual areas, however, forward projections can also originate in subgranular layers (below layer 4), though the terminal distribution remains targeted at layer 4 of the recipient area. In contrast, backward-type projections tend to originate in layer 6 of the higher area and distribute outside layer 4 of the lower area, though some feedback projections can have a bilaminar origin. Finally, some visual areas have interconnections whose terminal fields engage nearly all laminae. This distribution pattern is thought to characterize lateral connections between visual areas at approximately the same processing level. On the basis of these patterns of connectivity, it is possible to arrange the various occipital visual areas into a processing hierarchy. This is illustrated in Fig. 5. The extent to which this plan applies to human visual cortex is not known, though the overall scheme is likely to be similar.[2]

[2]New techniques in neuroimaging, including the analysis of functional connectivity (correlation strengths among visual areas) and diffusion–tensor imaging, hold promise for directly charting connectivity in the human brain.

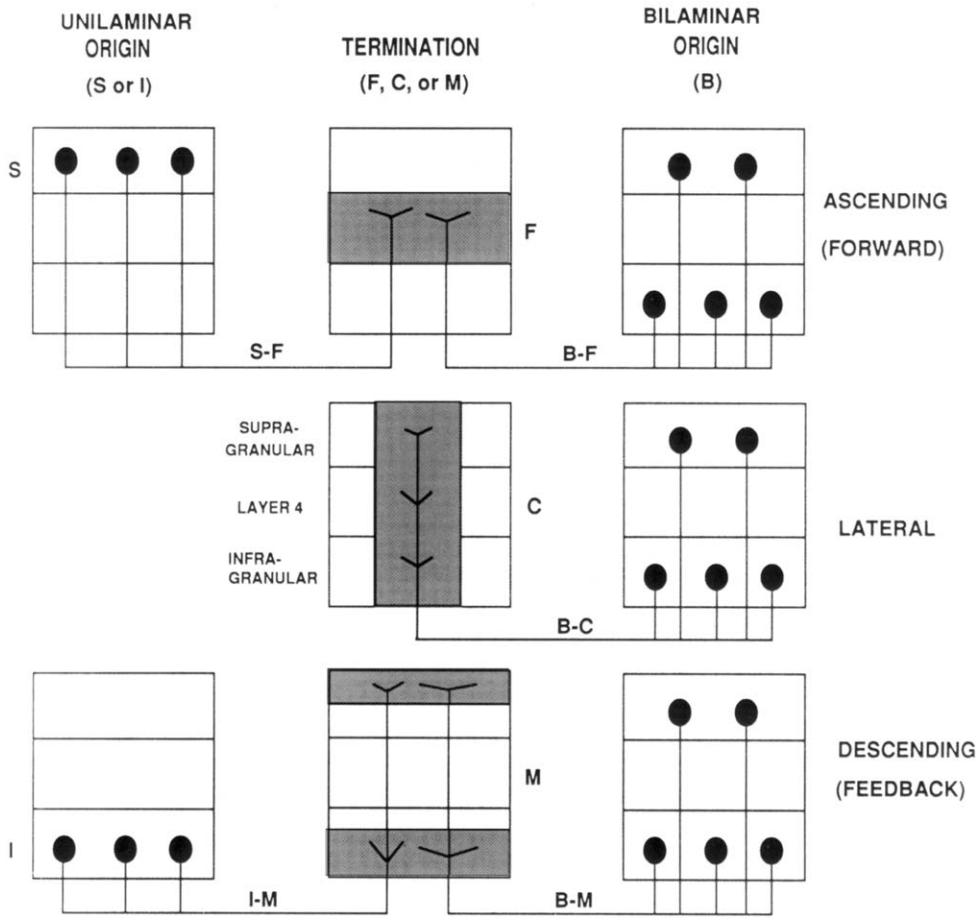


Figure 4 Schematic diagram of the laminar patterns of corticocortical connections among visual areas in macaque monkeys used to define ascending, descending, and lateral types of connectivity. The cell bodies of projection cells are found in one of two common patterns, unilaminar (left) or bilaminar (B) (right). Unilaminar projections can originate from supragranular layers (S) or infragranular layers (I). Ascending terminations tend to be focal in layer 4 (F). Lateral terminations tend to be columnar (C). Descending terminations tend to be multilayered (M). (From Felleman and Van Essen, 1991.)

This connectional hierarchy incorporates several important organizational features. First, it is generally consistent with the concept that successively more complex visual response properties are represented at successive processing stages, though detailed comparisons of the response properties of efferent neurons at one stage with the properties of recipient neurons at the next stage are generally lacking. Another key aspect of this corticocortical network is that nearly every forward connection is matched by a corresponding backward connection, thereby establishing reciprocity between visual areas. Alternate connectivity schemes based on the strength of response correlation among visual areas rather than anatomical connectivity have stressed this reciprocity and suggest that a less hierarchical, more dynamic, processing network may also provide a useful model of occipital connectivity. A third important principle illustrated in Fig. 5 is that multiple, parallel pathways exist within, and between, each processing level. There are no hierarchical levels that are interconnected by only a single pathway. [In Fig. 5, the multiple connections between entorhinal cortex (ER) and hippocampal cortex (HC) are represented only schematically.] The arrangement of multiple parallel pathways within the cortex thereby provides a basis for concurrent processing of different aspects of the incoming visual information. Thus, the M, P, and K pathways that provide concurrent

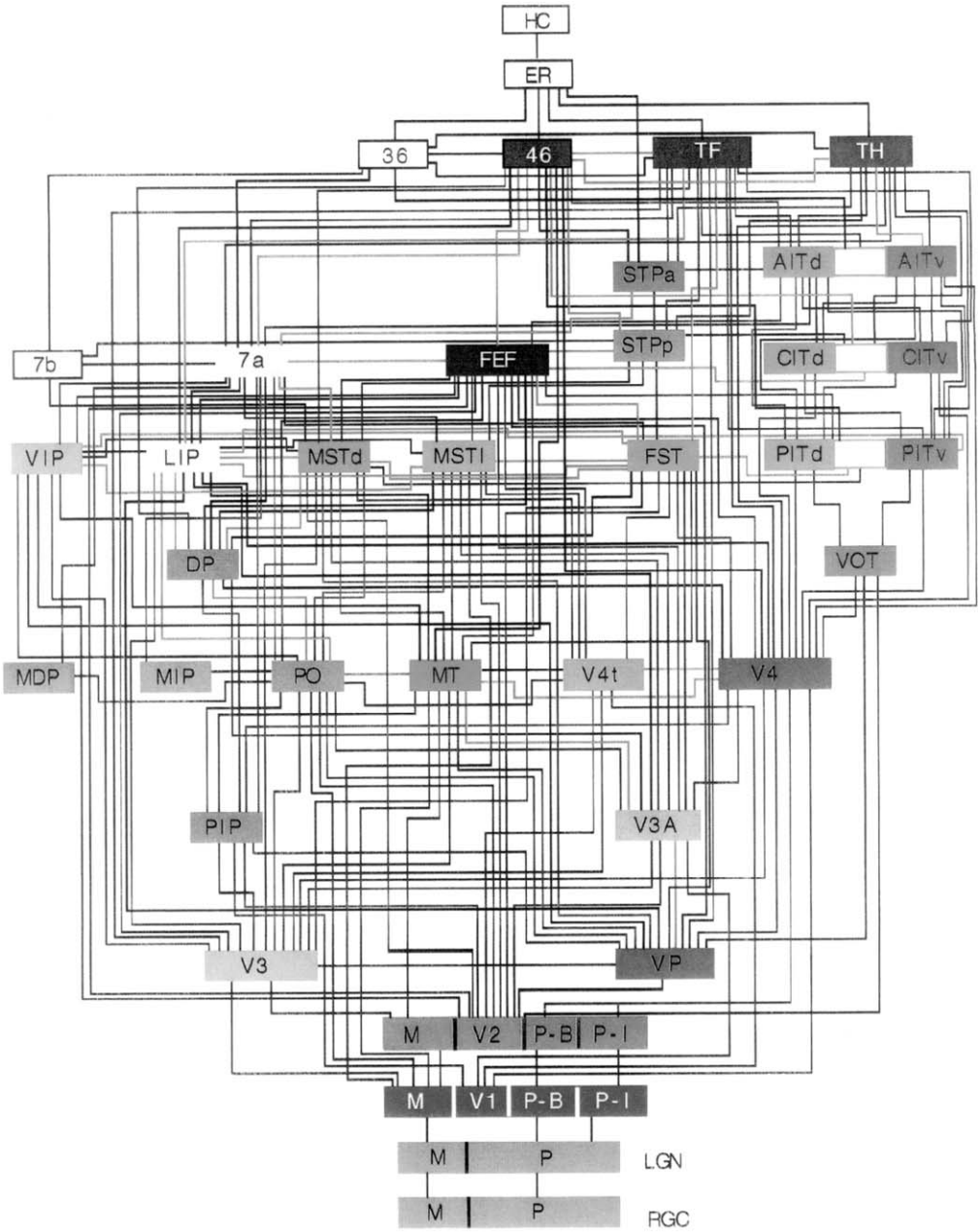


Figure 5 Connectivity and hierarchy of vision-related areas in the macaque cerebral cortex. Visual areas at a given hierarchical (vertical) position typically receive forward projections from areas below and backward projections from areas above. Most connections between areas are reciprocal (one forward, one backward). Areas at the same hierarchical level tend to have lateral connections. Forward, backward, and lateral types of connectivity are defined in Fig. 4. Visual areas V1, V2, V3, VP, V4, and V3A are completely or partially within the macaque occipital lobe. (From Felleman and Van Essen, 1991.)

processing within the retina and LGN are further modified in V1 to yield a new set of signals. These are then passed on to subsequent stages of processing via distinct sets of corticocortical output neurons. This same process is then reiterated at each successive processing stage.

Altogether, the pattern of afferent connections, the intrinsic circuitry, and the selective distribution of

multiple output signals at each stage are ultimately responsible for the unique functional properties of neurons in each cortical visual area. We turn now to a more detailed account of the topography of these areas and their functions.

III. FUNCTIONAL ORGANIZATION

Our understanding of the functional organization of the human brain has blossomed as never before thanks to the development of neuroimaging technologies that allow direct observation of the patterns of brain activity in human subjects actively engaged in sensory, motor, or cognitive tasks. In conjunction with more traditional electrophysiological, clinical, and psychophysical approaches, we now are developing a detailed picture of the topography of different visual areas and are beginning to outline their functional contributions. These new developments have taken place within the historical context of extensive earlier, and ongoing, studies of visual system organization in animals, especially nonhuman primates. Accordingly, the next section outlines this context before addressing more recent data.

A. Subdivisions of Visual Cortex

The organization of visual areas delineated in nonhuman primates provides a working hypothesis about the organization of visual cortex in humans. Figure 6 summarizes the topography of visual areas of several simian and prosimian species. The different visual areas identified in this figure have typically been defined by a convergence of several lines of evidence, including anatomical connectivity, neurophysiology, and lesion-related characteristics. Not all areas are equally well-defined, and some (possibly many) have significant internal heterogeneity. The degree to which these schemes apply to the human is not well-established. A few visual areas such as V1, V2, and MT are almost certainly homologous, whereas areas such as V3 (DM)–VP, V3A, and V4 are likely to have homologs but their identity is not yet established. For some areas, there may be no homolog. Accounts of visual area topography in humans have been heavily influenced by the macaque monkey scheme shown in Figs. 6F and 6G, though it is important to note that several alternative schemes for the macaque have been published. Due to such uncertainties, identifying homologies may not necessarily provide a definitive characterization of human visual areas. In the future, our understanding of the relationships among human and animal visual areas is likely to evolve as more data become available.

B. Visual Field Topography

Traditionally, visual field topography has been a primary source of information used to identify and map different visual areas in animals. Early work showed that a number of cortical visual areas contain a complete representation of the visual field, though each representation is split along the vertical meridian so that half of the field is represented in each hemisphere. Generally, then, the topographic arrangement of photoreceptors in the retina is maintained in the central connections. This results in an orderly arrangement of cells responsive to different locations in the field of view. In the cortex, neighboring positions in the visual field tend to be represented by groups of neurons that are adjacent but laterally displaced within the cortical gray matter. Neurons representing the vertical midline of the visual field are represented in both hemispheres and are functionally linked by interhemispheric, callosal connections. The end result is that the topography of the visual field is preserved in the visual cortex. [Because the visual field is directly imaged on the retina by the optics of the eye, visual field topography translates into retinal topography (retinotopy) but is upside down and backward due to the optical properties of the cornea and lens of the eye. Also, as will be described later, visual space is systematically distorted in the cortical representation.]

1. Charting Human Visual Areas by Retinotopy

Functional neuroimaging has been used to chart the retinotopic organization of visual areas in the human occipital lobe. Figures 7A–7C illustrate the results of such an experiment that used functional MRI. Subjects viewed a checkered hemifield that flickered at 8 Hz and slowly rotated about the center of gaze. During an fMRI scan series lasting approximately 4 min, the hemifield made five complete rotations, thereby sweeping through each angular position five times. Each location in the visual cortex went through five cycles of alternate activation and quiescence. However, cortical locations representing different angular positions in the visual field were stimulated at different times in the rotational cycle. Measurement of the temporal phase of the cyclic activation thereby

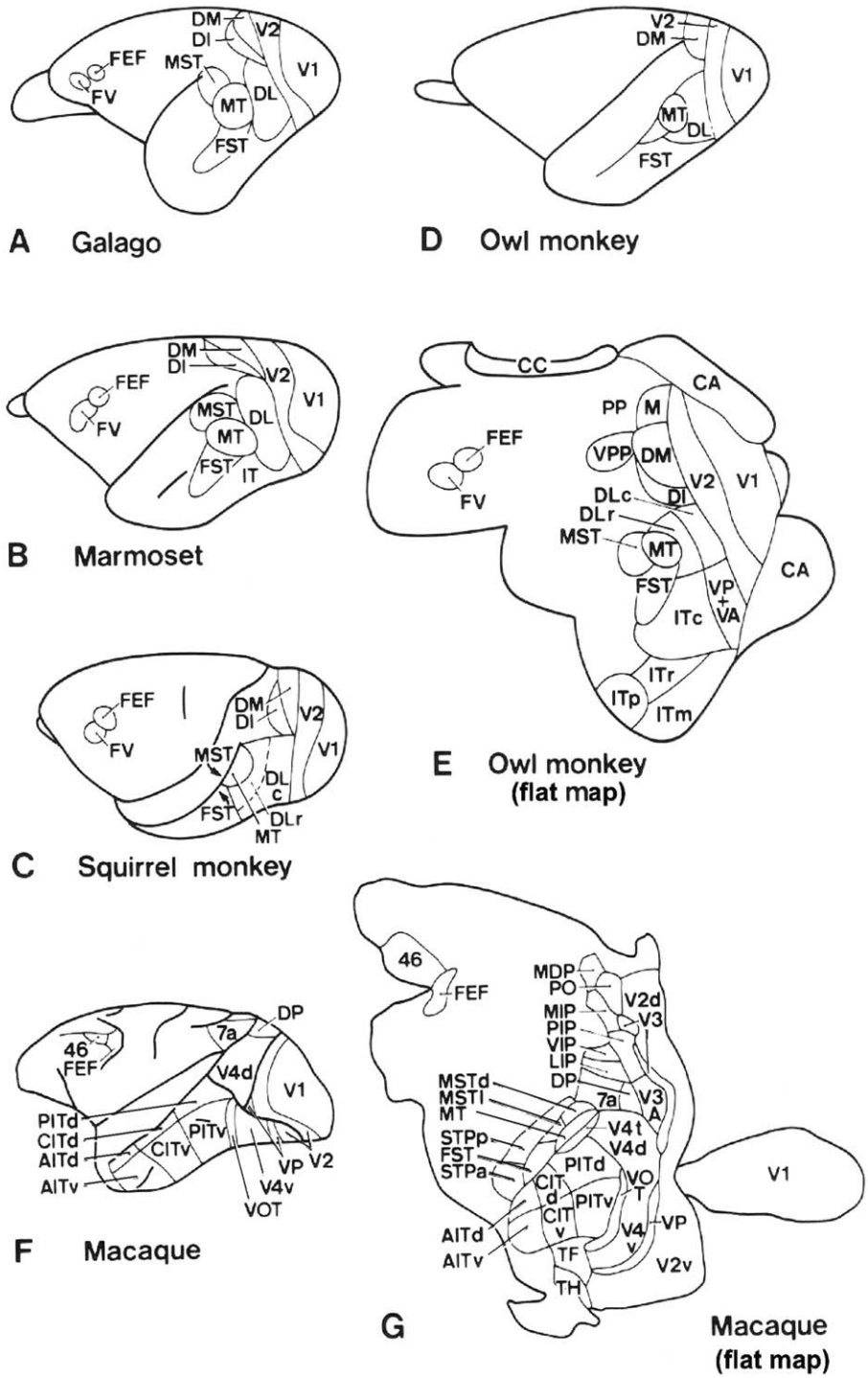


Figure 6 Summary of visual area topography for a variety of simian and prosimian species. Areas with the same name in different species are generally thought to be homologous. Other areas with apparent topographic similarity may or may not be functionally identical. Area DL is likely to be at least partially similar to macaque V4. DM may be homologous to macaque V3 and/or V3A. Note that this figure depicts left hemispheres unlike other figures in this article. (Adapted from Zilles and Clarke, 1997.)

identified the angular position represented at that cortical location. (See upper legend in Fig. 7.)

The polar angle representations depicted in Figs. 7A–7C are displayed on 3-dimensional models of each subject's occipital lobe and on corresponding maps of the unfolded and flattened cortical surface (Figs. 7A′–7C′). To facilitate the identification of cortical areas representing different visual field quadrants, the maps are pseudo-colored with a coarse scale to show brain regions that were maximally responsive when the midpoint of the checkered hemifield was within 45° of the horizontal meridian (white), the superior vertical meridian (dark gray), or the inferior vertical meridian (light gray). To produce these flat maps, the 3-dimensional cortical surface model was slit along the lateral occipital sulcus and was then unfolded using a computer algorithm. The resulting flat maps show bands of activation corresponding to alternating representations of quadrants at the horizontal and vertical meridia. On the medial wall of the hemisphere, these bands are oriented roughly parallel to the calcarine sulcus (cf. Figs. 7B and 7B′). Comparison across the three cases illustrated in Figs. 7A–7C provides an indication of the overall consistency across subjects as well as the degree of variation within each subject.

Visual field mapping can also be accomplished in a complementary manner using an expanding checkered annulus to provide information about the representation of visual field eccentricity (distance from the center of gaze). Figures 7D–7F illustrate the resulting patterns of activation.

The right side of Fig. 7 depicts the extraction of contours representing isoeccentricity and isopolar angle bands and their compilation into a final composite map showing the layout of visual space in the occipital lobe. Note that there are actually several reiterated representations of visual space, as indicated by the alternation of vertical and horizontal meridian representations (VM vs HM) labeled along the right margins of the meridia and composite maps.

From maps of the meridian representations, it is possible to estimate the locations of the borders between different visual areas. To do this, it is assumed that the borders are found at the representations of the horizontal and vertical meridia, as they are in subhuman primates. On the basis of this assumption, the layout of cortical visual areas in the human occipital cortex is illustrated on the large cortical flat map in the center of Fig. 8. To create this map, the 3-dimensional brain model (top) was slit (yellow dashed line) along the calcarine sulcus rather than along the lateral occipital surface. As a result, the vertical meridian representations (marked by black arrowheads) are oriented approximately vertically in Fig. 8 rather than horizontally as in Fig. 7.

Consistent with previous human studies, area V1 occupies the calcarine fissure but can extend onto the surface of the cuneus and lingual gyrus. The representation of the horizontal meridian contained within the calcarine fissure separates the upper and lower fields of V1 (split along the yellow dotted line in the large flat map of Fig. 8). It does not mark a border between different visual areas. As in the macaque, the borders between V1 and the two halves of V2 (V2d, V2v) are marked by vertical meridian representations for the inferior and superior halves of the visual field (VMi, VMs). The inferior vertical meridian representation is located along the upper lip of the calcarine sulcus, whereas the superior vertical meridian is located along the bottom lip. V2d and V2v together provide a second complete representation of the visual hemifield, but it is split in half with the inferior field represented dorsally and the superior field represented ventrally.

In the dorsal cortex of the cuneus, the border between V2d and V3 is marked by another representation of the horizontal meridian (HM). In complementary fashion, the border between V2v and VP is also delineated by a ventrally placed horizontal meridian representation (HM). Areas V3 and VP each contain a map of a quadrant of the visual field. Together these partial maps make up a third complete hemifield representation. In the macaque, these two quarter-field maps contain neurons with somewhat different visual response properties and anatomical connections. This suggests that they are functionally distinct, and, as a result, they have been given different names by some investigators. Others consider the formation of a complete hemifield representation as a more parsimonious way to summarize the data, even if some internal heterogeneity must be accepted. In this framework, the two quarter-field representations are termed V3d and V3v.

The anterior border of V3 is an inferior vertical meridian representation (VMi). In the macaque, this would form the border with area V3A, at least for visual field eccentricities within 10° of the fovea. Accordingly, this area has been termed V3A in humans. Some investigators have argued that parts or all of macaque V3 and V3A should be considered a single dorsomedial area (DM), such as that seen in owl monkeys (see Figs. 6D and 6E). V3A in macaque monkeys contains both an inferior and a superior quadrant representation. In one study, human V3A

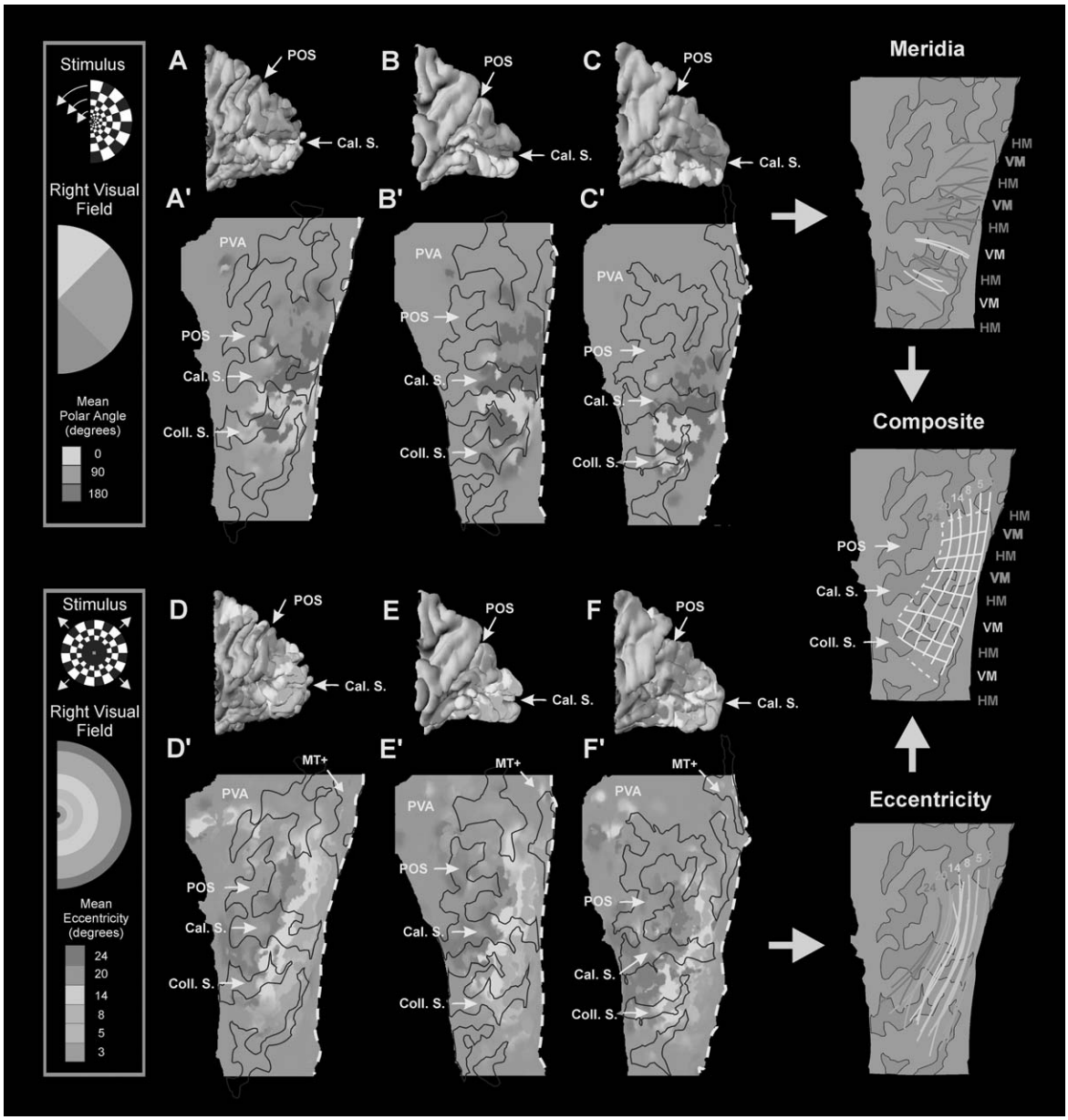
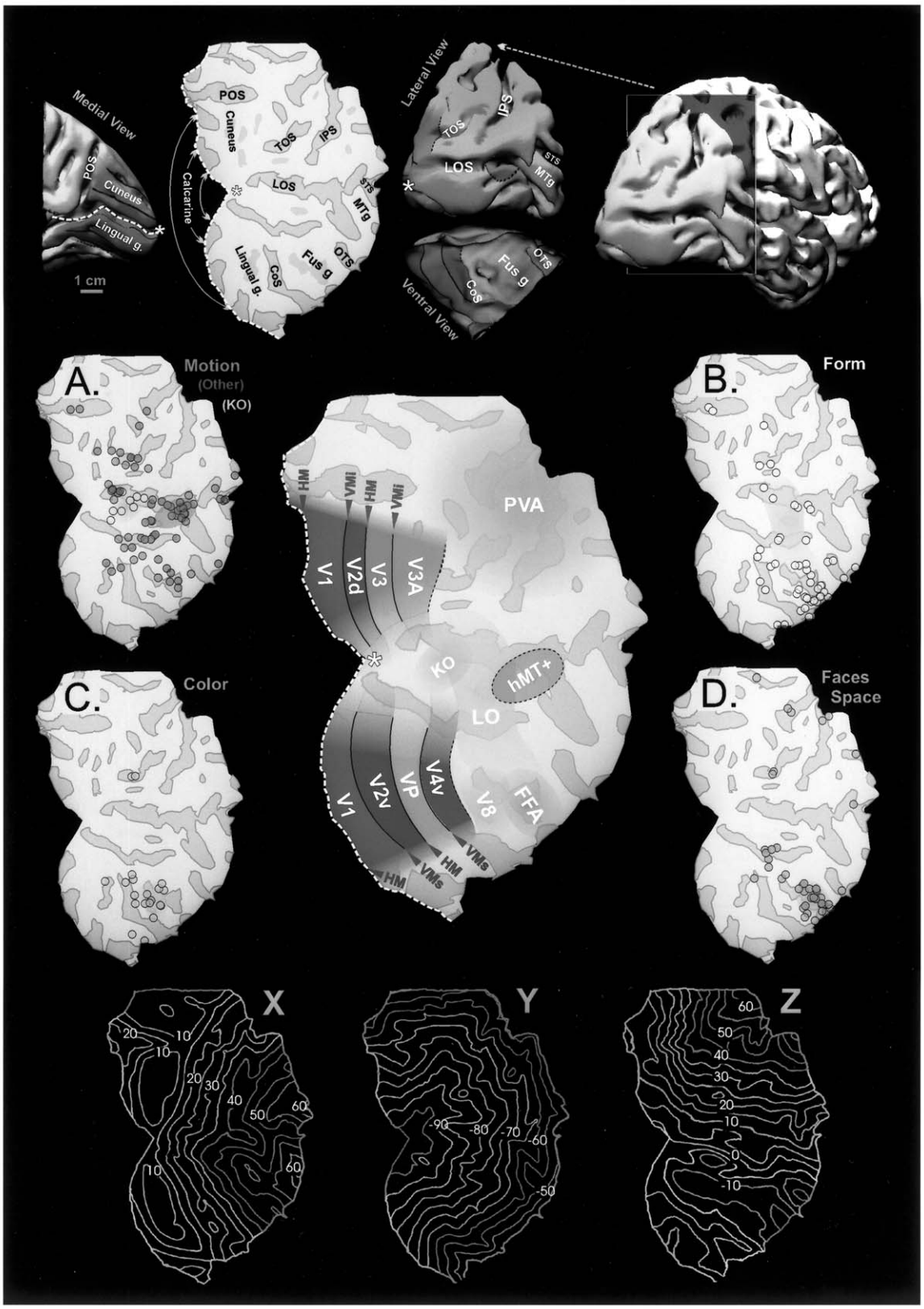


Figure 7 Visual field mapping in human occipital cortex. (A–C, A′–C′): The cortical representation of polar angle in the visual field in three subjects. The stimulus depicted at left was a counterphase flickering (8 Hz), checkered hemifield rotated slowly about the fixation point. For display purposes, the color code shows brain sites that were maximally activated when the center of the hemifield was within 45° of the horizontal meridian (white), the superior vertical meridian (dark gray), or the inferior vertical meridian (light gray). (A–C) 3-dimensional surface models of the occipital lobe plus adjacent parietal cortex, medial view, anterior to left. (A′–C′): Cortical flat maps showing the same data. White dashed lines at the edges of flat maps show margins of slit running from the occipital pole along the lateral occipital sulcus. Upper and lower halves of the dashed lines would be juxtaposed along the hidden lateral aspect of the 3-dimensional models shown in A–C. (D–F, D′–F′): Cortical representation of visual field eccentricity for the same three subjects. Stimulus: Flickering, checkered annulus that expanded slowly from 1.4° to 24°. Color code (gray to white) represents activation by six groups of three successive annuli. Inclusive eccentricities (degrees): 1.4–4.5, 2.3–7.6, 3.7–12.4, 6.3–20.8, 10.5–30, and 17.1–30. The coloring of the hemifield in the legend is only approximate; true mean eccentricities are shown at bottom. Right: Composite retinotopic grid (middle) combining data across subjects for vertical and horizontal meridian representations (right top) plus eccentricity (right bottom) displayed on a flat map of subject A. Lines in meridia and eccentricity maps pass through centers of corresponding meridian and isoeccentricity domains in A′–F′. Note multiple representations of meridia indicated at right of the composite and meridia maps. Abbreviations: HM, horizontal meridian representations; VM, vertical meridian representations (orange, lower field; yellow, upper field); Cal S, calcarine sulcus; Coll S, collateral sulcus; MT+, human middle temporal visual area and neighbors; POS, parieto-occipital sulcus; PVA, parietal visual areas. (Adapted from DeYoe *et al.*, 1996, *Proc. Natl. Acad. Sci. U.S.A.* **93,** 2382–2386.)


Medial View
POS
Cuneus
Lingual g.
1 cm
Calcarine
POS
Cuneus
TOS
IPS
LOS
STS
MTg
Lingual g.
CoS
Fus g
OTS
Lateral View
IPS
TOS
LOS
STS
MTg
Ventral View
Fus g
OTS
CoS
A.
Motion
(Other)
(KO)
B.
Form
C.
Color
D.
Faces
Space
PVA
V1
V2d
V3
V3A
KO
hMT+
LO
V1
V2v
VP
V4v
V8
FFA
HM
VMs
X
Y
Z
20
10
10
20
30
40
50
60
10
60
-90
-80
-70
-60
-50
60
50
40
30
20
10
0
-10

has also been shown to contain both upper and lower field representations, though in other respects this area may be functionally different from V3A in the macaque. Therefore, the designation V3A should be considered provisional for humans.

The anterolateral border of area VP (V3v) is formed by a second representation of the superior vertical meridian (VMs). In the macaque, this forms the border between area VP and the ventral half of V4. Correspondingly, the human area anterolateral to VP has also been designated as V4v. In both humans and macaques, V4v contains a representation of the superior field quadrant. In macaques, however, this is paired with an inferior quadrant representation located dorsally in the prelunate gyrus (cf Figs. 6F and 6G). This dorsal division, V4d, borders area V3A along an inferior vertical meridian representation. Human imaging studies have not identified a consistent focus dorsally that would correspond topographically and functionally to macaque V4d.

Anterior and lateral to V4v, near the occipitotemporal junction, lies an area, like V3A, that contains a complete representation of both the upper and lower visual fields. Topographically, this area is most equivalent to a region in the macaque known alternately as the posterior inferotemporal visual area (PIT, cf. Figs. 6F and 6G) or the temporal–occipital visual area (TEO). This area has been termed V8, but it also has been proposed as an alternative candidate for human V4 due to its selectivity for colored stimuli. The retinotopic organization of this area in both humans and monkeys is less distinct than in more posterior visual areas due to the presence of visually responsive neurons with large receptive fields (at least in the macaque).

Dorsal and lateral to V4v at the junction of the parietal, occipital, and temporal lobes is another visually responsive region that has a representation emphasizing the peripheral visual field ($>10^{\circ}$). It is strongly responsive to motion stimuli. This region is likely to be at least partly homologous to the MT–MST complex in macaque monkeys. In the macaque, visual area MT is adjacent to several small, motion-responsive areas (see Fig. 6G). The latter include V4t (a small transition area between V4 and MT), area MST (the medial superior temporal area), and area FST located in the fundus of the superior temporal sulcus. Due to the complexity of this region in the macaque, we prefer to call this region in humans the hMT+ complex pending further clarification.

The topography of human visual areas described previously is consistent with earlier anatomical studies of human post mortem material. The latter studies charted the pattern of interhemispheric callosal connections between visual areas in each hemisphere. In monkeys, axons interconnecting visual areas in the left and right hemispheres tend to terminate along the visual field representations of the vertical meridian. Accordingly, they provide an independent source of information about the boundaries of some visual areas, such as V1–V2, VP–V4, and V3–V3A. This is most convincing for the strongly retinotopic areas of visual cortex but becomes increasingly less precise for areas in which the retinotopy breaks down due to the

Figure 8 Topography and function of identified visual areas in human occipital lobe and neighboring cortex. Top: Anatomical features and topography of visual areas as displayed on a computer graphics model of the Talairach brain. Whole brain at right shows the plane used to create the separate occipital lobe model. Dashed yellow lines on the model and flat map show a cut along the depths of the calcarine sulcus to permit low-distortion unfolding of the cortical surface. A yellow asterisk marks the tip of the occipital pole. Dark gray outlines on all flat maps represent cortex buried within sulci of the 3-dimensional brain. Scale bar at the far left applies only to the 3-dimensional models. Center: Enlarged flat map of occipital cortex showing the topography of visual areas. Borders between visual areas drawn with solid black lines are relatively well-defined though the exact position of borders will vary from individual to individual. Black arrowheads indicate area borders that correspond with representations of visual field meridia: superior vertical meridian (VMs), inferior vertical meridian (VMi), and horizontal meridian (HM). Visual areas whose borders are shown as dashed lines or shading indicate greater uncertainty at the time of this writing. Center left: Flat maps showing sites (colored dots) reported in the neuroimaging literature to be selectively responsive to some aspect of visual motion (A) or color (C). Each dot is positioned at the point on the surface model that is closest to the Talairach coordinates reported in the original studies. Red-orange dots and swath in the motion map represent the approximate location of hMT+. Purple dots show other motion-responsive sites. Light-green dots and swath represent the location of the kinetic occipital area, KO. Light-blue swath in the color map represents the approximate location of area V8. Center right: Flat maps of form (B) and face or space (D) selective sites. Yellow swath in the form map represents the approximate location of the LO (lateral occipital) visual complex. Dark-green swath in the faces map represents the approximate location of the fusiform face area (FFA), Bottom: Flat maps showing Talairach coordinate isocontours (values in millimeters). Approximate Talairach coordinates of dots in the functional maps can be determined by tracing dots onto a transparent outline of the map and placing it successively over the X, Y, and Z isocontour maps. (See text for a note about accuracy.) Abbreviations: CoS, collateral sulcus; Fusg, fusiform gyrus; IPS, intraparietal sulcus; KO, kinetic occipital visual area; LO, lateral occipital visual complex; LOS, lateral occipital sulcus; MTg, middle temporal gyrus; OTS, occipitotemporal sulcus; POS, parietal–occipital sulcus; STS, superior temporal sulcus; TOS, transverse occipital sulcus. (See color insert in Volume 1).

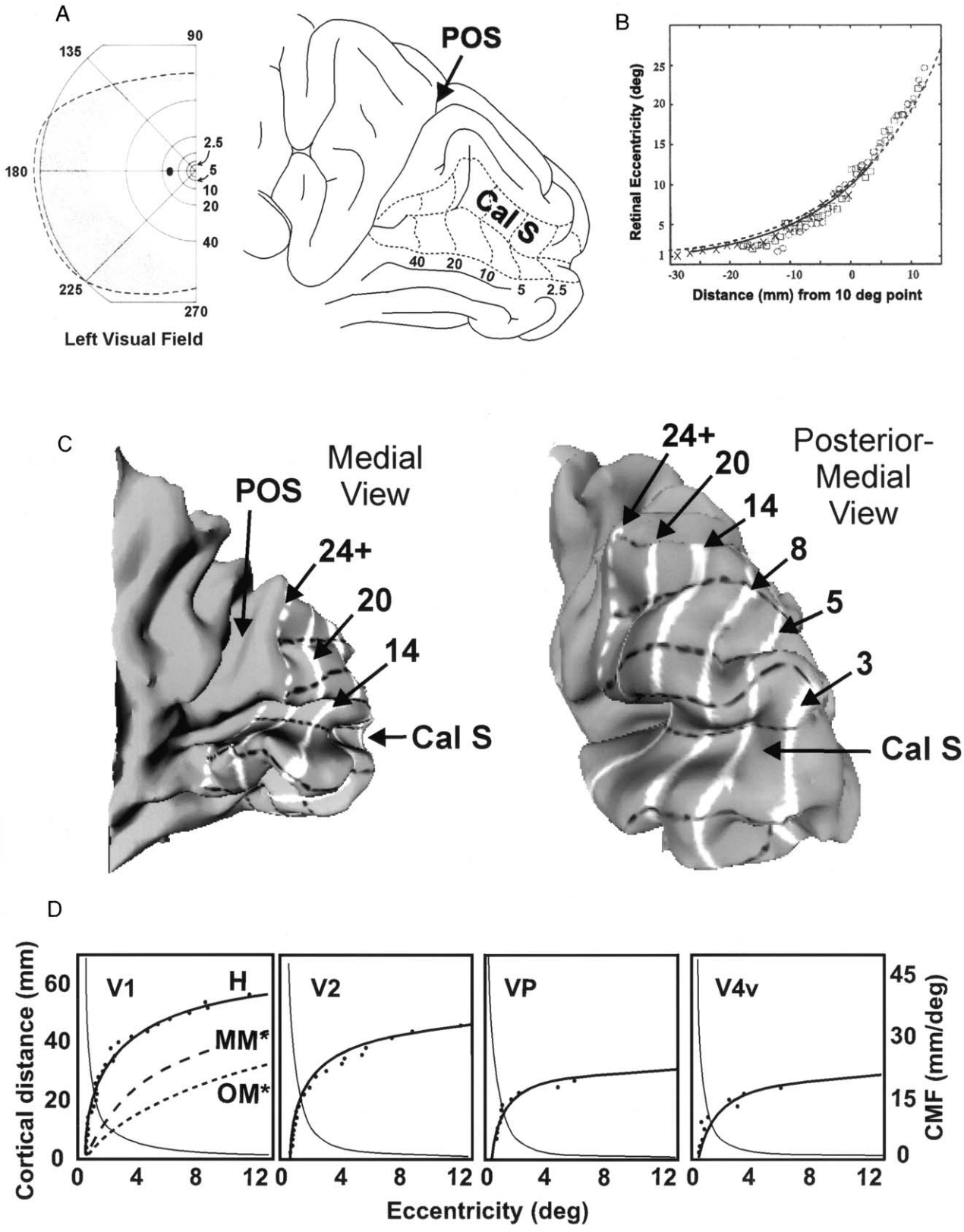
increasing size of single-cell receptive fields. In these latter areas, the callosal connections become less focused along the vertical meridian and are less useful for marking borders between visual areas. Nevertheless, the results from the human anatomical studies are in reasonably good agreement with the neuroimaging evidence for strongly retinotopic visual areas along the medial and ventromedial wall of the occipital lobe. For visual areas extending onto the dorsolateral and ventrolateral surfaces, the agreement is less certain, in part due to variability in both the imaging and anatomical data within these regions.

Several other visually responsive regions in occipital or nearby parietal and temporal cortex have been proposed as distinct visual areas. Some of these, such as the kinetic occipital area (KO), the lateral occipital area (LO), and the fusiform face area (FFA), are indicated in Fig. 8. Identification of the borders of these areas is more difficult than for the strongly retinotopic areas, and so they are intentionally shown as indistinct in Fig. 8. Other proposed visual areas, such as the parieto-occipital visual area PO–V6 and area V7, are not shown in Fig. 8 because their existence and characteristics are not yet firmly established.

2. Quantitative Retinotopy

Besides the qualitative aspects of visual field topography used to define visual area boundaries, neuroimaging has produced new data concerning quantitative features of human retinotopy. Although the qualitative topography of the visual field is preserved in cortical retinotopic maps, the metric of visual space is warped. That is, the representation of the fovea is greatly expanded relative to the representation of the periphery. (In macaques the warping is also known to be different for different visual areas.) Figure 9A shows the representation of visual field eccentricity in striate cortex as described by the researchers Horton and Hoyt in 1991. Their figure was based on a consideration of the extent of blindness in the visual field produced by partial lesions of calcarine occipital cortex. The foveal representation is located posteriorly at the occipital pole, with more eccentric visual field positions (farther from the center of gaze) located successively more anteriorly along the calcarine fissure. The representation of eccentricities close to the center of gaze is expanded relative to more peripheral eccentricities. The representation of the central 10° covers nearly half of the calcarine cortex, with the remainder of the field to beyond 40° covering the other half of cortex. At the time this article was published, little was known of the retinotopic organization of extrastriate cortex in humans. Figures 9B–9D illustrate more recent data from several neuroimaging experiments. As can be seen in Fig. 9C (and Figs. 7D–7F), the representation of visual field eccentricity within V1 of the calcarine sulcus extends both dorsally and ventrally across the entire medial surface of the occipital lobe and onto the dorsolateral and ventral surfaces. As described earlier, this region of cortex contains several different extrastriate visual areas. The results shown in Fig. 9C suggest that the representation of visual field eccentricity in these areas is topographically concordant with the representation in V1. But different visual areas vary in total size, so that the extent of cortex representing 1° of visual angle—the cortical magnification factor—will vary from area to area. Cortical mapping functions are illustrated in Figs. 9B and 9D. One study compared

Figure 9 Quantitative retinotopic organization of human visual cortex. (A) Right: Medial view of the occipital lobe with lips of the calcarine sulcus opened to show retinotopy of V1 as estimated from visual field defects in patients with occipital lobe lesions. Numbers along the lip of the calcarine sulcus indicate degrees of visual angle from the center of gaze, as indicated in the adjacent visual field schematic (A, left). (Adapted from Horton and Hoyt, 1991.) (B) Human cortical mapping function according to Engel *et al.*, 1997, *Cereb. Cortex* **7,** 181–192. Open symbols are measurements from two observers. The solid curve shows the best fitting exponential (least-squares) to the four hemispheres measured in the study. The dotted line shows an estimate derived from scotomata in human stroke patients and electrophysiological data from nonhuman primates. The x's are fMRI measurements by another researcher (see D). (C) Retinotopy of striate and extrastriate visual cortex as determined by functional MRI displayed on 3-dimensional models of the surface that would correspond to cortical layer 4 (i.e., with supragranular layers removed). The surface model has been smoothed and slightly unfolded to provide a better view of the topography within the calcarine sulcus. White lines represent the loci of constant visual field eccentricity (3, 5, 8, 14, 20, and 24+ degrees of eccentricity). Dashed black lines show representations of horizontal and vertical meridia. Dashed line within the fundus of the calcarine sulcus is the horizontal meridian representation running through the middle of V1. (Based on data from DeYoe *et al.*, 1996, *PNAS* **93,** 2382–2386.) (D) Cortical mapping functions (scale on left axis) and magnification functions (scale on right axis) for upper field striate cortex and three extrastriate visual areas in the human. Also shown are the mapping functions for V1 in the macaque monkey and owl monkey (MM* and OM*, respectively, in the first panel). The cortical mapping functions (heavy lines) show the relationship between distance along the cortex and the representation of visual field eccentricity (Note the difference in graph axis labeling compared to B.) Cortical magnification is the extent of cortex representing 1° visual angle. (Adapted from Sereno *et al.*, 1995, *Science* **268,** 889–893.)


A
90
135
180
225
270
2.5
5
10
20
40
Left Visual Field
POS
Cal S
40
20
10
5
2.5
B
Retinal Eccentricity (deg)
Distance (mm) from 10 deg point
C
Medial View
POS
24+
20
14
Cal S
Posterior-Medial View
24+
20
14
8
5
3
Cal S
D
Cortical distance (mm)
V1
H
MM*
OM*
V2
VP
V4v
CMF (mm/deg)
Eccentricity (deg)

data across visual areas and, for V1, across species. They suggested that there may be an even greater foveal expansion in humans than in monkeys, but this has been challenged. There is some disagreement among studies concerning the degree of expansion of the most central 2–3° in humans. Researchers addressed this issue by comparing their own data, with those of others and with a best fitting exponential function (solid line in Fig. 9B). These studies confirm the contention that the human foveal representation in V1 is larger than had been suggested by earlier studies, and, together, they provide the best quantitative description of visual space mapping in the human visual cortex so far available.

3. Summary of Topography

Overall, the topography of occipital visual areas in the human brain is generally consistent with the topography in macaque monkeys, but with some potential departures related to the expansion of the foveal representation and uncertainties concerning areas such as V4 and posterior–lateral occipital cortex. Compared to the macaque monkey, the placement of visual areas relative to anatomical structures in the two species is also different. Human visual areas in lateral cortex appear to be displaced posteriorly toward the occipital pole or onto the medial occipital surface. For example, the hMT+ complex is located posterior to the superior temporal sulcus (STS) in humans but is buried deep within the STS in macaques. Most of the human retinotopic visual areas (V1, V2, V3–VP, V4, V3A) are located entirely or partially on the medial and ventral aspects of the occipital lobe. In macaques, they are located partly or completely on the lateral surface of the brain. This displacement of human visual areas places the foveal representation of the visual field at the occipital pole. In macaques it is situated laterally near the ends of the lunate and inferior occipital sulci. Although the positioning of visual areas relative to anatomical features in humans and macaques is distinctly different, the positioning of visual areas relative to each other may be more consistent across species. This suggests that topographic maps of the visual cortex that are abstracted from gross anatomy (e.g., as seen in cortical flat maps) may provide a more functionally consistent view of cortical organization.

Maps of visual field topography provide one line of evidence for defining and differentiating visual areas. Such maps also provide a good organizational framework for understanding the contributions of different areas to specific visual functions such as color or movement perception. The following section uses this approach to review data concerning the functional specialization of the human visual cortex.

C. Analysis of Visual Information

1. A Conceptual Framework

To discuss the functional specialization of the human occipital lobe, it is helpful to have a conceptual framework for organizing and interpreting the relevant experimental data. One fruitful approach in this respect is to outline the computational task of vision. In other words, "What must the visual cortex accomplish?" A complete and detailed answer to this question is still lacking. Nevertheless, a broad outline such as that illustrated in Fig. 10 can be of use.

Visual information available to the brain arises from the patterns of light on the retina of each eye. These patterns vary spatially, temporally, and in wavelength composition as objects in the visual scene change and as the observer moves and explores. In a sense, the brain knows nothing about the immediate state of the external world but, rather, must infer its attributes from the incoming retinal information (potentially with help from prior experience). Our resulting visual perception is amazingly rich and includes the appreciation of object form and rigidity as well as surface properties such as color, lightness, texture, and specular reflectance. We effortlessly discern the spatial arrangement of things and surfaces as well as their movements relative to each other and relative to ourselves. Finally, we can distinguish lighting properties such as shadows and the direction of ambient illumination. On the whole, the brain does a surprisingly good job of providing a veridical impression of the attributes of the world around us, despite occasional exceptions such as those that occur in visual illusions.

The true complexity of the visual computations becomes more evident when considering how retinal image information is used to infer specific object attributes. Initially, early stages of visual processing must extract and isolate different types of information that will be useful for inferring each of the various attributes. Work in computational vision and psychophysics has helped to identify such visual cues and to outline how they contribute to the computation of specific attributes such as distance and form. For instance, each eye has a slightly different direction of

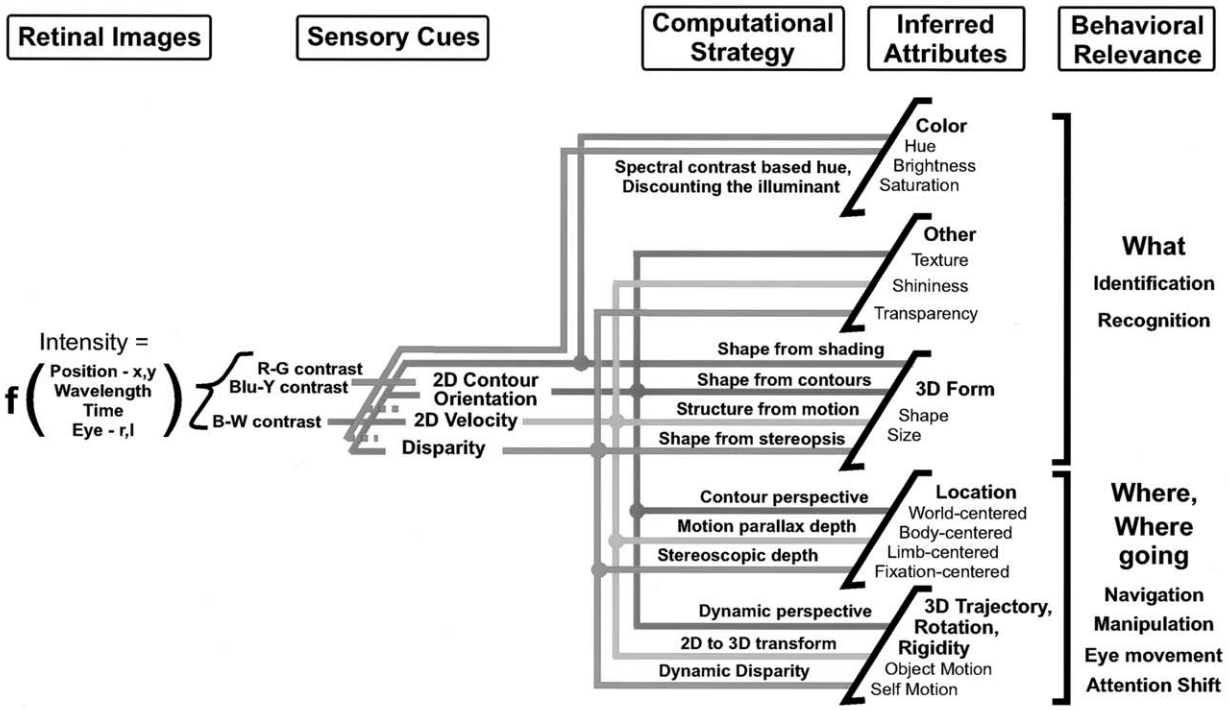


Figure 10 Conceptual framework of the information processing task of vision. Information about the characteristics of the visual scene is provided by the patterns of light imaged on the retinae of the two eyes over time. These patterns can be described physically as functions of retinal position, wavelength composition, time, and eye (left or right). From this information, the brain must infer the attributes of objects and surfaces within the visual scene. This may be accomplished through a variety of computational strategies, some of which are noted in the figure. Inferences based on any particular computational strategy depend strongly on specific "sensory cues," meaning a particular subset of the information available in the retinal images. For example, inferences about hue depend heavily on chromatic contrasts but little on motion-related information. Sensory cues noted in the diagram provide only a few rudimentary examples that are readily identified with the visual response properties of neurons at early and intermediate stages of visual processing. Red dotted lines for chromatic contrast cues indicate potential contributions of chromatic information to form and motion (see text section on color). The diagram expresses the idea that each low-level sensory cue may contribute to inferences about more than one world attribute. Moreover, each world attribute often may be inferred from more than one cue. This manifold relationship among sensory cues and inferred attributes provides a useful rationale for the multiplicity of visual areas and concurrent processing pathways within the visual cortex. In the brain, a particular path from a sensory cue to an inferred attribute may consist of many synaptic levels containing further divergent or convergent branching as well as the computation of intermediate cues. The diagram also expresses the idea that there is not usually a one-to-one relationship between a low-level sensory cue and a single inferred attribute. Missing from this outline are the effects of previous experience, expectation, and attention on the attributes inferred at any given moment. (Adapted from DeYoe and Van Essen, 1988, and Van Essen and DeYoe, 1994.)

gaze with respect to a viewed object. As a result, the images of that object in the two eyes are also slightly different. The resulting disparity in the retinal images provides a cue that can lead to a vivid impression of distance and 3-dimensional structure. Similarly, computer-generated displays of moving dots representing the vertices of an invisible, rotating cube can yield a strong impression of its 3-dimensional form. In many viewing situations, however, both types of cues—retinal disparity and motion parallax—act in concert to jointly specify an object's form. In general, different visual cues often redundantly specify a particular object attribute, thereby helping to eliminate any ambiguity in the interpretation of the retinal image data. It is also true that single visual cues can contribute to the computation of more than one attribute. Motion cues can specify both the 3-dimensional form of an object as well as the trajectory of the object through space.

As can be seen from the previous examples, the relationships among visual cues and inferred object attributes can be complex. Fig. 10 outlines some of the relationships among different visual cues and a variety of world attributes. Several important concepts are

illustrated by this figure. First, it is clear that the goal of the visual system is not to act like a camera and simply preserve the patterns of light falling on the retinae. Rather, the brain must actively convert the retinal information into neural representations of the attributes of objects and surfaces in the visual scene. This is further complicated by the need to eliminate the effects of irrelevant changes in the retinal images caused by factors such as shifts of viewing position or changes in ambient illumination. Second, the notion that the retinal images yield a number of low-level visual cues that, in turn, ultimately contribute to the inference of object attributes suggests an analysis that, overall, is inherently hierarchical in nature. At the same time, the fact that different cues can be derived from the images and can contribute simultaneously to different attributes suggests that concurrent processing will also occur. Third, the relationships between visual cues and inferred attributes are manifold. That is, multiple cues can specify a single attribute, and, conversely, a single cue can specify multiple attributes. Thus, a particular visual cue may be represented in different ways by several concurrent neural systems related to different attributes. Fourth, this leads to the more general notion that functionally specialized visual areas and networks will not be organized as isolated centers devoted exclusively to the processing of a single visual cue. This also means that we should not usually expect to find isolated processing "conduits" extending from the retina to high levels of the visual system that are dedicated along their entire length to processing a single visual cue–attribute combination. This does not preclude the possibility that some processing pathways may, at points, be confined to restricted anatomical loci, thereby allowing focal lesions to cause specific perceptual loss. Finally, it may be helpful to conceive of the visual computations as occurring in stages, even though it may be difficult to precisely identify or define such stages in a real neural network that has extensive forward and backward interconnections. At each successive stage, different types of information made explicit in the previous stage are mixed and combined in different ways in order to achieve the next level of representation. Particular attributes may become explicit at different stages. For example, the orientation of individual texture elements may be explicitly represented in V1, whereas the relatively invariant colors of a Mondrian painting may not be explicitly represented until a later stage.

With these concepts in mind, we turn now to a consideration of specific functional systems within the human occipital lobe.

2. Functional Subsystems

a. A Map of Functional Specialization[3] The middle section of Fig. 8 summarizes data from 56 neuroimaging studies concerning the functional specialization of human visual cortex. Each colored dot on the small cortical flat maps (A–C) represents a focus of activation identified in one or more of these studies. The location of each dot represents the closest point on the cortical surface model of the Talairach brain (Fig. 1) that matched the published Talairach coordinates of the site.[4]

[3]To provide a systematic frame of reference for describing the locations of functionally distinct regions of the occipital lobe, we will use the standardized coordinate system of Talairach and Tournoux as illustrated in Fig. 1. In this stereotaxic system, left–right coordinates (x axis) are measured in millimeters relative to the midline (here positive values = right). Anterior-posterior (y axis) coordinates are expressed relative to the anterior commissure (here positive numbers = anterior to AC). Superior–inferior (z) coordinates are measured relative to an axis passing through the anterior and posterior commissures (here positive values = superior). A brain normalized to this system is spatially transformed so as to match the overall size and orientation of the standard brain depicted in the Talairach and Tournoux atlas. Generally, this transformation tends to reduce the variability of coordinates for major anatomical and functional features. However, it may or may not improve the correspondence of local features, especially for structures located far from the coordinate reference points at the anterior and posterior commissures. This can be the case for occipital lobe structures and can limit the ability to make fine distinctions among functional or anatomical features in different subjects. Other, more reliable coordinate systems have been developed and are likely to become popular, but, at this juncture, the Talairach system is the most widely used.

[4]Although the creation of a composite map based on Talairach coordinates is advantageous for summarizing the results of a large number of studies, it should be stressed that there are several sources of potential inaccuracy in the resulting map. First, the process of normalizing a subjects' brain to the Talairach standardized coordinate system can be imprecise. Moreover, the published coordinates of cortical activation sites typically represent the mean centers of mass or peak responses across multiple subjects. They do not provide a measure of the extent of the activated area. Second, markers (dots) positioned on the brain model will be located at the coordinate on the surface that is closest to the actual 3-dimensional coordinate in the brain volume. If the original coordinates were not located within the cortical gray matter, the displayed location will only be an approximation. Finally, there are inevitable, though modest, distortions associated with flattening the folded cortical surface onto a plane. Due to these various inaccuracies, issues concerning differences in function of neighboring visual areas can rarely be settled on the basis of Talairach coordinates or summary maps derived from such coordinates. Nevertheless, a summary map does provide a useful starting point for discussing the functional differences that are obviously discernble in the data despite mapping inaccuracies.

It is important to recognize that this summary map represents a necessary simplification of complex neuroimaging data. To say that an area within the human brain is functionally specialized typically means that it responds differentially to two or more experimental test conditions that, ideally, are perfectly matched in all factors other than the factor of interest. We will use the term "selective" to refer to activation sites that respond in this differential manner. Selectivity at a site does not imply that it responds exclusively to changes in one particular visual cue or attribute. Rather, it means that the response is modulated most strongly by one cue, attribute, or task factor relative to the others tested. In contrast, an area that responds nonselectively will have roughly equivalent responses to different test conditions.

b. Early Stages of Cortical Visual Processing The earliest stages of visual processing involve the extraction of visual cues and rudimentary visual attributes from the incoming retinal information. Virtually all visual processing tasks activate primary visual cortex, V1, and neighboring visual areas on the medial wall of the occipital lobe. This is consistent with the known connectivity of the visual pathways in monkeys, whereby the incoming visual information is relayed from the retina through the lateral geniculate nucleus to V1 and then to V2 and other extrastriate visual areas (Fig. 5). (The role of alternate subcortical routes to the cortex has not been explored extensively in human neuroimaging studies.) At hierarchical levels beyond V1 and V2, part of the visual information goes to areas V3 and VP[5]. Although V3 and VP in the macaque are closely connected with V1 and V2, it is not clear whether they are as ubiquitously involved in visual processing as V1 and V2. Some single-unit studies of color and motion processing in monkeys suggest that, on the whole, these two areas contain cells with somewhat different visual response properties and presumably different functional roles. This may be in accord with occasional reports from neuroimaging studies of selective activation in the cuneus and lingual gyrus close to the presumed human counterparts of V3 and VP.

Early stages of cortical processing tend to be activated by virtually all visual tasks. If two tasks are similar in many respects, then subtraction of functional brain images from the two conditions (e.g., experimental vs control) can effectively "cancel out" activity at early processing stages (e.g., V1–V2), while highly specific and differential activation of later cortical visual areas remains. In the following discussion, the common activation of early cortical stages will not always be mentioned, though it should be assumed to be present unless otherwise indicated. Also, it is important to note that functional heterogeneities such as cytochrome oxidase puffs and stripes within V1 and V2 are sufficiently small that the different compartments tend to be blurred together by the relatively large size of the imaging samples (voxels) used in most human neuroimaging studies. As a result, differential patterns of activity in the microcircuitry of early visual areas may appear equivalent in fMRI or PET.

c. Motion Flat map A of Fig. 8 summarizes the results of nearly two dozen neuroimaging studies that have been concerned with the processing of visual motion information. Many of these studies involved comparison of a coherently moving field of dots or small squares with a similar but motionless field. The subject typically viewed the stimulus without performing a visual discrimination task (passive view), though in some experiments a judgment about the stimulus was required. Most of these studies found activation in visual areas V1 and V2 concordant with the presumed flow of visual information through these early processing stages to higher extrastriate areas. In addition, many of these studies reported activation within a separate area of lateral occipital cortex near the junction of the temporal and parietal lobes (red dots in Fig. 8A). The mean Talairach coordinates computed from 15 of these studies are as follows: right hemisphere, lateral 42, posterior 69, dorsal 2; left hemisphere, lateral 42, posterior 69, dorsal 3. Across the various studies, these coordinates varied by 1 cm or more in all directions. The area encompassing the average coordinates as well as a large cluster of individual sites has been shaded in red in Fig. 8. The relationship between this motion-responsive region and the local gyral–sulcal pattern in individual subjects can vary to some extent. This not only reflects true variation in the activation foci across studies but also

[5]The notion of a hierarchy is used here to facilitate discussion of a complex system that processes the visual input concurrently through multiple sequences of neuronal connections. In such a system, it may be difficult to delineate specific functional stages and pathways due to complex forward, backward, and lateral interconnections. Nevertheless, there is an overall sense in which the retinal representation is successively transformed through connections that are farther and farther removed from the input. In such a case, it is useful to talk about "higher" or "later" stages of processing, even though it may not be possible to assign a precise rank to every connection or cortical area.

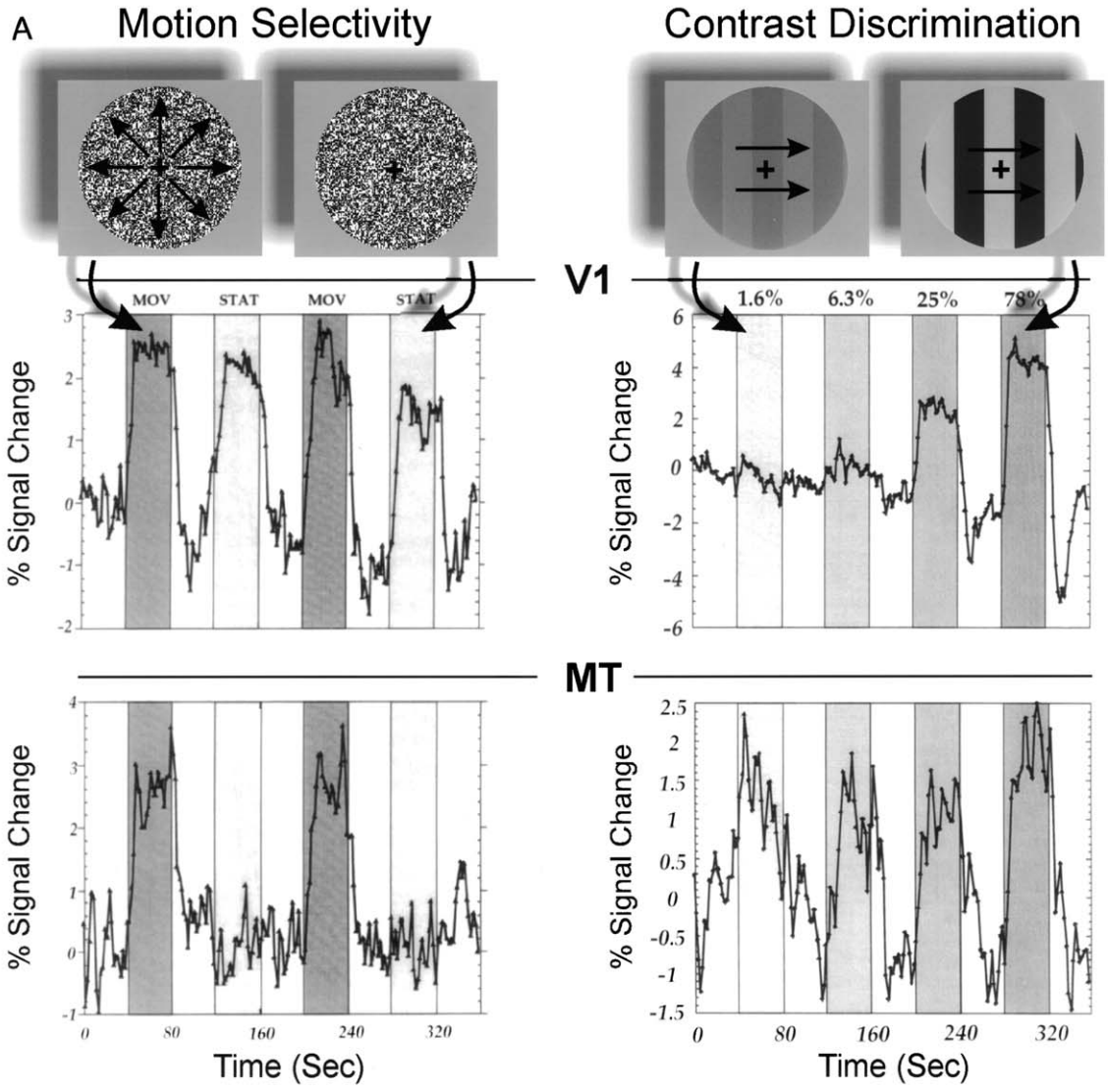
reflects the irregularity of each individual's anatomy in this region.

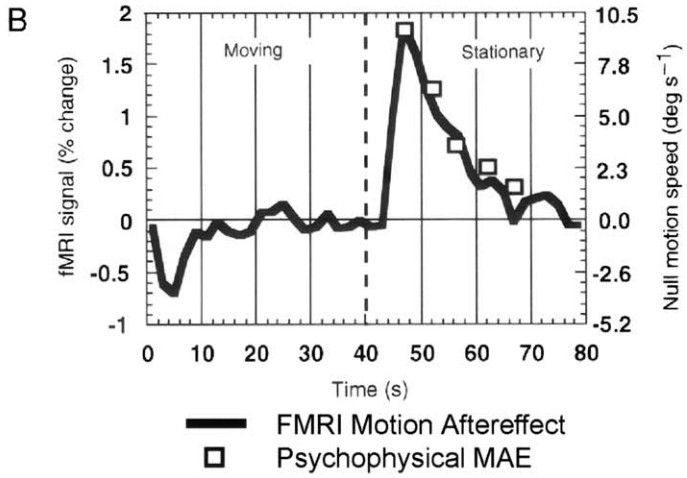
Converging evidence suggests that at least part of the lateral occipital motion focus is homologous to the middle temporal visual area (MT) in monkeys (also referred to as V5 by some investigators). Figure 11A illustrates how human fMRI responses in this area are concordant with the known motion-selective properties of single neurons in monkey MT. These effects include selective activation by moving stimuli, high contrast sensitivity, and reduced responses to equiluminant colored gratings. Anatomical and histological features in this region are also consistent with characteristics of macaque area MT. These features include dense myelination, high cytochrome oxidase activity, and staining by the antibody CAT-301. It remains to be determined whether these anatomical features are exactly coextensive with the activation sites observed in the imaging studies. In fact, the lateral occipital area activated in these studies may also include motion-selective areas other than MT. In the macaque monkey, area MT projects to neighboring area MST (of which there are three subdivisions: MSTd, MSTl, MSTf) and to FST (fundus of the superior temporal sulcus). In monkeys, these areas juxtapose MT and contain cells that are also motion-responsive. It is quite possible, then, that the area identified previously as human MT may include these areas as well. Here, we will use the term hMT+ to indicate the complex of MT and its motion-responsive neighbors.

Human hMT+ responds selectively to moving vs stationary stimuli, but also responds to other aspects of motion stimuli in ways that appear to parallel reported single-unit response properties or perceptual effects. Comparisons between globally coherent (dots move together) vs incoherent (dots move independently) motion produce a significant change in activation within hMT+, yet can result in little or no change in activation in V1–V2. This is consistent with the understanding that, on the whole, neurons in monkey V1–V2 do not readily distinguish global coherent motion from incoherent motion but that later stages of processing, such as area MT, do. hMT+ responses are also directly related to the relative coherence of a moving random dot pattern, concordant with the corresponding motion percept. Neuroimaging tests of motion opponency as well as tests contrasting first- and second-order motion (contrast- or flicker-modulated sinusoids) also differentiate hMT+ from V1–V2, in agreement with single-unit studies. As shown in Fig. 11B, motion stimuli that evoke a strong perceptual aftereffect (waterfall illusion) produce fMRI signals in human hMT+ that persist after the inducing stimulus is turned off, thereby tracking the time course of the illusory percept. Altogether, these observations provide strong evidence that motion perception may arise in hMT+ or is likely to be derived from hMT+ signals. This inference is further supported by human clinical studies in which akinetopsia (the failure to perceive motion) or more subtle motion deficits tend to be associated with lesions of the occipital lobe in the vicinity of hMT+. Together, these findings further indicate that there is a good case for homology between human and monkey MT.

In addition to hMT+, a variety of other cortical sites are activated by motion stimuli or tasks. These sites are shown by the purple and light-green dots on flat map A of Fig. 8. The complex and widespread distribution of these foci reflects the fact that motion cues contribute to a variety of perceptual phenomena, such as self-motion, "biological motion," figure–ground segmentation, structure from motion, and the control of eye movements. This complexity is consistent with the concept outlined in Fig. 10 that there are multiple motion processing pathways related to the manifold relationships between low-level motion cues and various inferred attributes of the visual scene. Attempts to identify homologies and functions for these various sites are outlined next.

Figure 11 Comparison of visual response properties of human hMT+ and V1. (A) Left: Selective fMRI activation of hMT+ by moving vs static stimuli consisting of random dots, which in the moving condition (MOV) expanded or contracted radially but in the static condition (STAT) were motionless. In the intervening periods, no dots were presented. Graphs show the time course of the fMRI signal in V1 (top) and hMT+ (bottom). (A) Right: V1 discriminates different contrast levels of grating stimulus but hMT+ does not. Graphs show the time course of fMRI signal averaged over voxels in V1 or hMT+. Human MT+ signal tends to be nosier due to smaller number of voxels contributing to the average. (From Tootell *et al.*, 1995, *J. Neurosci.* **15,** 3215–3230.) (B) fMRI-based estimate of the time course of motion aftereffect (MAE) and its relation to the psychophysical estimate of the perceptual MAE. MAE was induced by having the subject view a grating consisting of concentric rings drifting continuously outward or inward for 40 sec. This was followed by a period in which the rings were stationary. The graph at right of the dotted line shows the time course of residual fMRI signal in the presence of the stationary grating. Squares indicate the strength of the perceived motion aftereffect measured with a nulling technique. Note the correspondence of the fMRI time course and psychophysical effects. (From Tootell *et al.*, 1995, *Nature* **375,** 139–141.)


A
Motion Selectivity
Contrast Discrimination
V1
MOV
STAT
MOV
STAT
1.6%
6.3%
25%
78%
% Signal Change
MT
Time (Sec)


B
Moving
Stationary
fMRI signal (% change)
Null motion speed (deg s^{-1})
Time (s)
FMRI Motion Aftereffect
Psychophysical MAE

Coherent motion stimuli have been shown to activate sites in occipital cortex other than V1–V2 and hMT+. One possible function of at least some of these sites is the processing of optical flow information. Flow fields are generated when the viewer moves through space and they provide a key source of visual information on self-motion. (The flow can be noticed by staring fixedly out a car window as it speeds by a building or fence.) Flow fields often contain large areas of coherent motion with salient expanding or contracting components, depending on the viewer's direction of gaze relative to the direction of movement. Cells in monkey MST, but less so MT, are selectively responsive to these components and to coherent rotational motion as well. Because of this, complex flow patterns might be expected to identify the human homolog of macaque MST. In humans, stimuli that evoke a strong perception of illusory rotational motion activate foci that are partially or completely distinct from sites activated by linear coherent motion. On the basis of such data, it has been proposed that MST (also known as V5A) might be located very close to area MT, as it is in the monkey (perhaps included in the zone designated here as hMT+). A different study, employing shifts of visual attention, identified a motion-related focus in the inferior parietal lobule that might be an alternate site for human MST. In other tests, this region was activated by visually guided, smooth-pursuit tasks, concordant with the properties of neurons in macaque MST but not MT. However, such a correlation is not conclusive because other sites in ventral cortex have also been found to be responsive during smooth-pursuit tasks. At this time, it is unclear which human site, if any, should be considered strictly homologous to monkey MST.

Several studies have identified motion-selective sites that were thought to be potentially homologous to macaque visual area V3. In monkeys, cells in V3 respond better to moving than to stationary stimuli and respond better to second-order motion stimuli than to first-order. This appears to be the case for the human sites as well. However, some of these studies concluded that the activated sites were part of V3 solely on the basis of Talairach coordinates, leaving open the possibility that they were really located in adjacent visual areas such as V3A. This is important because evidence from experiments combining retinotopy with motion tests suggests that, in humans, the area corresponding retinotopically to macaque V3A may be more motion-selective than retinotopically defined human V3. The human V3A candidate contains representations of both upper and lower visual fields and possibly a separate foveal representation, consistent with V3A in macaques. Contrary to this, single-cell recordings in monkeys suggest that V3, rather than V3A, responds more selectively to motion stimuli. Whether this represents a true functional difference between monkeys and humans is unclear. All in all, differences in technique between monkey and human experiments combined with the uncertainty of Talairach coordinates make a definitive statement about differences in motion processing between human V3 and V3A premature, though clearly one (or both) of these areas is involved in motion processing.

Another potentially important function of motion processing systems is in visual tracking. Human hMT+, plus some frontal and parietal sites, are activated by moving targets that are being tracked either overtly by eye movements or covertly by shifts of attention. The human hMT+ site appears to be equivalently activated by both passively viewed target motion and active attentional tracking, whereas later processing stages are markedly more responsive during active tracking. This is consistent with the notion that motion analyses in hMT+ provide target-related signals used at later stages to control eye movements and/or shifts of attention. Indeed, visual areas in monkey parietal cortex including VIP and LIP are involved in integrating information about visual motion and eye movements, and these areas receive projections from MT–MST.

Another intriguing use of motion cues is in the perception of "biological motion," that is, the movement of people and animals. Small lights attached to a subject's joints can allow a viewer in an otherwise dark room to identify the nature of the movement (walking, dancing, jumping) and even the gender of the subject. In such instances, motion information is being used to identify and recognize the subject and his or her actions in addition to conveying information about trajectory. A few neuroimaging studies have investigated this topic and have shown that a region anterior to hMT+ in the superior temporal sulcus (STS) responds selectively to biological motion sequences of the whole body. Although located well into the temporal lobe, such a site may represent an important branch of the occipital motion processing systems, perhaps feeding into object recognition systems of the temporal lobe rather than into dorsal motion pathways concerned with trajectory or optic flow. Biological motion involving grasping movements of the hand (signaled by lights attached to the hand) activates a region of the intraparietal sulcus as well as the STS site activated by whole-body movements. Single-unit studies have

shown that specific subregions of the intraparietal sulcus in macaque monkeys contain neurons that appear to represent reaching movements in body-centered coordinates. The existence of multiple sites responsive to biological motion suggests that this type of motion information may be analyzed in different ways depending on the coordinate frame or class of motion involved.

Though retinal image motion often indicates the movement of objects (or self), motion cues can also be used to identify the boundaries and 3-dimensional shape of objects rather than their trajectories. As will be discussed later, the ventral occipitotemporal cortex appears to be heavily involved in the analysis and identification of objects. Brain regions that utilize motion for such purposes might be expected to include ventral cortical areas rather than, or in conjunction with, dorsal motion sites such as hMT+. There are sites in the ventrolateral occipital cortex that are consistently activated by motion-defined shapes rather than by uniform motion itself (light-green dots in Fig. 8A). Discussion of such sites is covered later in the section on form processing.

d. Color A number of imaging studies have investigated color processing in human visual cortex. Recordings from both implanted and scalp electrodes have also been useful in this respect. Fig. 8C shows the distribution of color-selective responses reported for ventral occipitotemporal cortex. The most recent studies have identified several distinct foci within this region. Human visual areas VP and V4v are activated somewhat more strongly by chromatic than by achromatic stimuli, but greater selectivity and task specificity are found anterior to V4v. Such selectivity has been shown to be associated with visual area V8 and, in some tasks, with a more anterior site, as shown in Fig. 12. V8 is retinotopically distinct from V4v and contains a representation of both superior and inferior visual fields. Color-selective activation sites in dorsal occipital cortex have also been observed on occasion, but their location and identity remain poorly defined.

The presence of occipitotemporal color-related sites is consistent with human lesion studies in which damage to ventral cortex results in diminished or absent color discrimination (achromatopsia). Often, these lesion studies report color perception losses in both upper and lower visual fields, even though the lesions appear to be confined to ventral cortex. Because retinotopically defined visual areas V4v and VP contain only upper field representations, they are unlikely to be the sites responsible for this cerebral achromatopsia. In contrast, area V8, which is located within the lesion-sensitive zone, does have both upper and lower field representations, thereby making it a prime candidate for the achromatopsia site. Moreover, V8 is located adjacent to the fusiform face area described in the next section. This juxtaposition of V8 and the FFA is likely to account for the fact that prosopagnosia (inability to recognize faces) is commonly associated with cerebral achromatopsia.

Although there is fairly good agreement among experts concerning the cortical locations of color- and achromatopsia-related sites in humans, there is significant controversy concerning the relationship between those sites and their counterparts in other primates. It has been proposed that the more anterior color-selective sites in human occipitotemporal cortex (labeled here as V8) should be considered the true homolog of macaque V4. Supporting evidence is based on the observation that macaque V4 (see Fig. 6) contains a significant number of cells with color-selective properties, consistent with a role in color vision. Yet the uniqueness of macaque V4 in this respect has been challenged, and there are important reservations concerning the role of simian V4 in the perception of color. Lesion studies of monkey V4 have typically failed to demonstrate severe disruption of color discrimination. Sometimes more subtle effects are observed such as the loss of color constancy under different lighting conditions, whereas hue discrimination is only mildly affected. In contrast, lesions in the posterolateral aspect of the macaque temporal lobe anterior to V4 (i.e., in area PIT) produce deficits that closely resemble those of a human achromatopsia patient tested in an identical manner. These results are consistent with the idea that, in humans, cerebral achromatopsia does not arise from lesions of V4v but from more anterior lesions encompassing V8 and/or additional neighboring areas. Despite these findings, the total number and precise topography of visual areas within posterior inferotemporal cortex in humans are still poorly defined, as are the roles of each individual area in color perception. In sum, the known properties of V8 make it a likely candidate for the achromatopsia site. This does not necessarily imply that V8 or any single visual area is solely responsible for color vision or for all color-related deficits. An achromatopsia site may simply be an especially susceptible nexus in a more extensive system of color-related processing.

Finally, it is important to note that neurons in other visual areas can be chromatically selective, that is, respond better to some colors than others, yet fail to

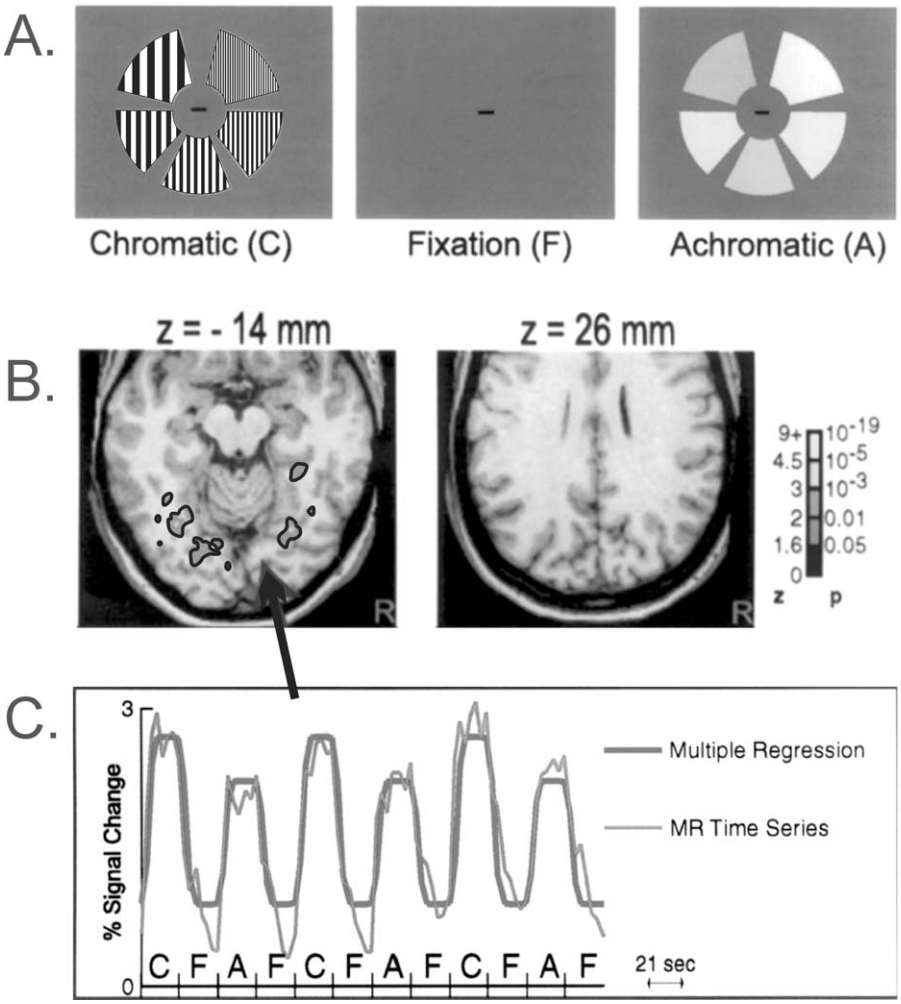


Figure 12 fMRI activation of color-selective sites in ventral occipitotemporal cortex. (A) Stimuli presented during alternating periods of the fMRI scan. (An ordered sequence of hues is represented here by different texture patterns.) During presentation of the chromatic stimuli, the subject indicated whether the hues in the color wheel were arranged in order or were random. For the achromatic stimulus, subjects made an identical judgment about the different gray levels in the wheel. (B) Brain slices through ventral occipitotemporal cortex (left) and dorsal occipitoparietal cortex (right) showing foci at which responses during the chromatic task were stronger than during the achromatic task. Note the lack of color-selective sites in dorsal cortex. (C) Time course of an average fMRI waveform for chromatically selective sites shown in B. (Adapted from Beauchamp *et al.*, 1999, *Cereb. Cortex* **9,** 257–263.)

contribute to the perception of hue. For instance, differences in chromatic contrast can help segregate an object from its background or can produce an unambiguous percept of motion in an otherwise ambiguous display. In fact, moving stimuli consisting of pure chromatic contrast can readily activate human hMT+. The amplitude of the resulting signal correlates with the perceived speed of the moving stimulus, even though the latter may not accurately reflect the true speed of the stimulus. If differences in sensitivity for chromatic vs luminance contrast are taken into account, human hMT+ (and presumably other components of the dorsal motion system) can be driven by color stimuli, though the resulting signals do not contribute to the perception of hue. These findings support the concept outlined in Fig. 10 that low-level visual cues, in this case chromatic cues, can support the perception of attributes such as movement in ways that may be unexpected.

e. Form Human imaging experiments have addressed a number of aspects of form processing ranging from the representation of simple contours to the identification of complex objects such as faces. There is general agreement that a major portion of the form processing system is located in the ventral

occipitotemporal cortex. This is illustrated in flat maps B and D of Fig. 8, which show cortical sites activated by a variety of tasks involving form perception, including recognition of faces.

i. Contours, Edges, and Contrast Passive viewing of luminance-based contours and edges readily activates retinotopic visual areas of the medial and ventral occipital cortex in humans. Generally, the higher the contrast of contours or grating patterns, the stronger the fMRI response. In one notable study, the researcher found that the size of a detectable increment in the contrast of a plaid grating was quite precisely tracked by a proportional increment in the fMRI response of visual areas V1, V2d, V3, and V3A. In area hMT+, however, the fMRI response tends to saturate above a few percent contrast, consistent with the high contrast sensitivity of MT neurons studied in animals. This suggests that accurate representation of the contrast of contours and edges is not a primary function of dorsal occipital visual areas such as hMT+, though a complete survey of contrast coding especially in the ventral occipitotemporal cortex (e.g., VP, V4, V8) has not been performed at the time of this writing. Consequently, it is not known whether the correlation between the fMRI signal and perceived contrast is best in a few key visual areas or is equivalently good throughout a wide variety of areas.

An alternate way to probe early stages of form analysis is through the use of "illusory contours," such as those illustrated in Fig. 13. Such contours—the apparently continuous outlines of the circle, square, and triangle in Fig. 13—do not exist in the physical stimulus but are nevertheless inferred by the brain mechanisms used to extract contour information from a variety of visual cues (luminance, color, and texture differences, occlusion boundaries, etc.). Single-cell recordings in animals have shown that neurons in V2, but less so in V1, are capable of responding selectively to these contours, thereby suggesting a hierarchical processing mechanism. Neuroimaging studies with human subjects have shown that illusory contours tend to activate the same neural systems that are activated by real luminance-based contours, though some types of illusory contours (Kaniza type) may preferentially activate later processing stages, including V3A, V4v, V7, V8, and LO.

ii. Motion-Defined Form In contrast to illusory contours, edges formed by motion differences along a boundary (Fig. 13) can activate early processing stages more strongly than later stages, though the effects for individual visual areas depend on the spatial density of the contours and stimulus size. It is important to note that selective responses to motion-defined edges (rather than to movement itself) are not known to be a hallmark of visual areas such as hMT+, which commonly are associated with motion processing. Rather, the most selective responses to motion-defined contours and shapes have been associated with a region termed the kinetic–occipital area (KO, indicated by light-green dots in Fig. 8A). It is located approximately 1.5 cm posterior to hMT+ in lateral occipital cortex. Whether KO is a distinct visual area or a portion of another visual area such as V4 or LO is not entirely clear, though the available evidence supports a separate designation. A more important issue is whether KO constitutes a separate, concurrent processing pathway for motion information. In one study combining evoked potential analysis with fMRI, no difference in latency of response was observed between hMT+ and ventrolateral areas (presumably

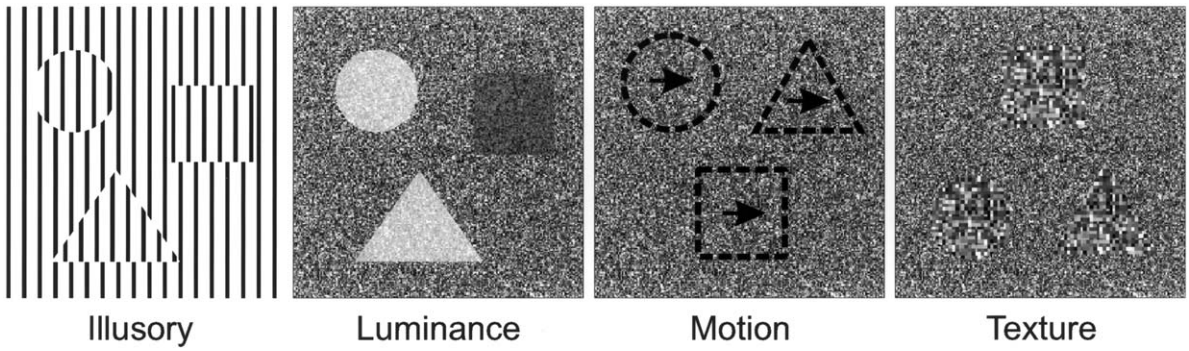


Figure 13 Examples of shapes delineated by illusory contours or by differences in luminance, motion, or texture. In the motion example, arrows indicate that dots within the dashed areas drift relative to dots in the surrounding area. Dashed lines are not present in actual stimuli. A visual area that supports a cue-invariant representation of shape would be expected to respond equivalently to a shape defined by any cue (given equivalent salience). Cue invariance is thought to be a characteristic of an abstract neural representation of shape.

KO) responding to motion-defined forms. This is consistent with the idea that low-level motion cues are processed concurrently in multiple pathways and contribute to the computation of other attributes besides trajectory.

iii. Cue Invariance From a theoretical standpoint, it would be satisfying if contour and edge properties, such as orientation, had a common neural representation regardless of the cues used to define the contour (see Fig. 13). Single-unit studies in animals indicate that at least some neurons in posterior inferotemporal cortex (PIT) are orientation-tuned for edges defined by more than one cue. Such response properties are said to be "cue-invariant" and may constitute an abstracted representation of the attribute of contour orientation. In humans, several neuroimaging studies have tested for cue-invariant responses to contours, edges, or outline figures defined by luminance, color, texture, binocular disparity, or motion. In the most comprehensive study, similar responses were observed for luminance-, texture-, and motion-defined figures in lateral occipital cortex posterior to hMT+ (labeled LO in Fig. 8) and in, or near, area V3A. In the context of human neuroimaging experiments, cue invariance simply means that the same amplitude of fMRI response to each cue condition was obtained from individual imaging voxels. This does not necessarily mean that within each voxel the individual neurons themselves are cue-invariant, but the results are consistent with that possibility. If true, the cue-invariant representation of contour information in LO and V3A may represent an important stage of abstraction in the inference of object attributes from the available retinal information.

iv. Figure–Ground Segmentation Closely related to the computation of contours and edges is the problem of segmenting the visual scene into regions associated with distinct objects or surfaces. It is thought that such figure–ground segmentation is a necessary prelude (or concomitant) to inferring more detailed attributes of objects, such as their shape and true surface color. In monkeys, it has been shown that the activity of single neurons at cortical stages as early as V1 can accurately distinguish the difference between the interior and exterior of a figure. Whether this is also true for V1 in humans is unclear, though likely. Lateral occipital cortex in humans has been shown to respond better to objects or object silhouettes than to random textures or "scrambled" objects that are not easy to segment into figure vs ground. It has been proposed that a swath of cortex located posterior to area hMT+ forms a complex specialized for the segmentation and processing of object information. This area, labeled LO in Fig. 8, is more strongly activated by discrete objects than by gratings, textures, or motion fields. Intriguingly, image manipulations that degrade the perception of a discrete object (e.g., addition of noise) also degrade the response. Conversely, manipulations that enhance the perception of an object (e.g., low-pass filtering of a coarse pixel image) also enhance the response in LO. Image manipulations that do not affect object perception, such as scaling and translation, have less of an effect on LO responses. This region does not appear to be strongly retinotopic, thereby suggesting that it contain neurons with receptive fields that encompass large portions of the visual field. This is thought to be a necessary characteristic of neurons that must represent the more global attributes of objects. Whether LO involves the representation of complete 3-dimensional objects or is primarily involved with initial stages of object representation such as figure–ground segmentation remains to be determined.

v. Complex Forms Discrimination of a change in the shape of simple rectangular targets activates a region of cortex in the collateral sulcus and adjacent lingual gyrus that is similar to retinotopically defined V4v and/or VP. These simple form tasks can also activate more lateral cortex in the fusiform gyrus or adjacent inferior temporal gyrus, sometimes extending into area LO. Complex forms or form discriminations often produce activation in ventral occipitotemporal cortex anterior to the strongly retinotopic visual areas. Together, these sites encompass a swath of cortex that extends from the occipital lobe into ventral temporal cortex. This swath is not functionally homogeneous and may be composed of subregions that are specialized for processing different types of form information. This is perhaps most consistently apparent for the analysis and recognition of faces.

Faces, Furniture, and Expert Knowledge: Identification of faces may be one of the more complex and subtle problems of form analysis. Accordingly, face stimuli and face discrimination have been studied extensively. They have been the focus of numerous animal single-unit and human neuroimaging studies. As illustrated in Fig. 14, the results from the human studies have been quite consistent and show that the fusiform gyrus from the occipital junction forward to midtemporal lobe is activated by tasks requiring face recognition. This area is selectively activated even when compared with a task requiring only classification of the face as male or female. The posterior extent

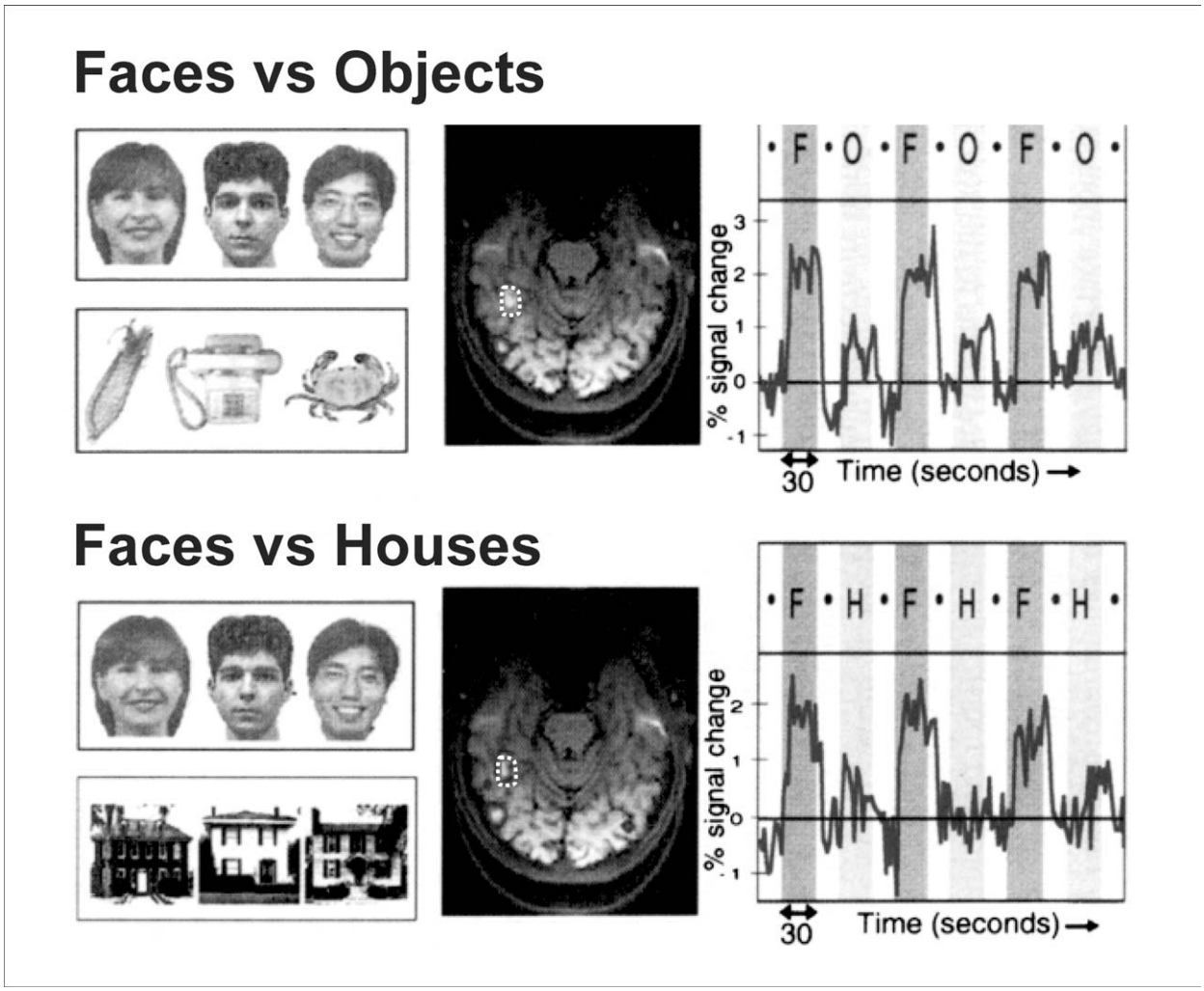


Figure 14 Selective activation of the fusiform face area by faces vs objects and houses. (Top) Left: Examples of stimuli viewed in alternating epochs during fMRI scan. Middle: Axial slice through the ventral occipital lobe and adjacent temporal cortex showing the site of the fusiform face area (dotted outline). Right: Time course of the fMRI signal during epochs in which the subject viewed fixation alone (epochs marked with dots), faces (marked with F), and objects (marked with O). (Bottom) Same as for the top but for alternation of faces and houses. (Adapted from Kanwisher *et al*., 1997, *J. Neurosci.* **17,** 4302–4311.)

of this region is lateral to the area activated during color processing (V8) or during the presentation of ''scrambled'' faces and has been termed the fusiform face area (FFA, Fig. 8).

The potential existence of a localized region of ventral occipitotemporal cortex that is critical for face recognition is consistent with the clinical syndrome of prosopagnosia, the inability to recognize previously familiar faces. Lesions that cause propopagnosia tend to have common involvement of the fusiform gyrus within the general vicinity of the FFA. Yet other evidence suggests that, if adequately tested, patients suffering from prosopagnosia can also exhibit deficits for objects other than faces. Such findings have prompted contrasting proposals that the form system either contains specialized ''modules'' for the analysis of different object categories or consists of a more universal network for analyzing all types of objects (or some combination of the two). In support of the latter concept, it has been shown that different types of objects—faces, buildings, furniture—all activate a large swath of ventral occipital cortex, but within that region there are spatially distinct ''hot spots'' whose configuration changes with object class.

Another alternative is that face recognition requires a uniquely high level of detailed, subtle analysis. Faces and their expressions convey unique social information, which is of critical importance for the individual

in terms of survival and well-being. In this sense, we are all accomplished ''experts'' at recognizing and interpreting faces. Some investigations have explored the possibility that the FFA actually constitutes an area specialized for acquiring such expert knowledge. For example, hobbyists who are particularly adept at discriminating different cars or birds will have activation in the vicinity of the FFA when performing those tasks as well as when performing face recognition tasks. This suggests the possibility that the form system may consist of both a generalized network that can deal with a vast variety of objects and specializations for fine discrimination within object categories that are especially important to the individual. These latter specializations may be malleable, thus allowing for the acquisition of new areas of expertise.

The existence of face-related modules or hot spots in human visual cortex was anticipated by studies in nonhuman primates that revealed neurons whose visual response properties were remarkably selective for face stimuli. It would seem simple to posit homology between face processing areas in the human and sites containing face-selective neurons in the macaque. In humans, face-selective responses are located ventrally, adjacent to color-selective regions in and near the collateral sulcus. In macaques, face-selective neurons have been found predominantly in anterior portions of the STS and lateral inferotemporal cortex, apparently separated from visual areas such as V4 that were traditionally thought to support color processing. This apparent discrepancy across species may be explained by more recent findings. As discussed earlier, cortex that, when lesioned, destroys hue discrimination in the macaque is located within posterior inferotemporal cortex rather than in V4. If true, color- and face-related sites in monkeys are located closer to each other than previously thought, and so are more consistent with the topographical organization in humans.

Semantic Aspects of Form Processing: Perception of the form of an object typically evokes associated semantic information, such as the class of items to which the object belongs, the utility of the object, or its ownership (*my* car). In reality, recognition of an object may be inseparably linked to at least some semantic associations. A theoretically satisfying notion in this respect is that early stages of visual processing located in the occipital lobe are most heavily involved in the direct perception of the form of an object but that later stages, especially in temporal cortex, are increasingly related to semantic associations. For instance, one researcher asked subjects to categorize pictures of objects as living or natural vs nonliving or man-made. In a control task, the subjects were asked to classify sinusoidal gratings as being oriented vertically or horizontally. Subtraction of images from the two tasks revealed activation sites in the fusiform and inferior temporal gyri that may be related to the classification of objects into semantic categories such as living vs nonliving. These results are consistent with clinical reports of patients with selective deficits in the identification of either animals (live objects) or tools (man-made objects). More recently, researchers have found that activation patterns within occipitotemporal visual areas even as early as V1–V2 can vary depending upon the semantic task involved (naming tools vs naming animals). These latter effects probably do not reflect the representation of semantic information itself in early visual cortex, but rather reflect the necessity of additional form processing when the task requires subtle distinctions between similar members of the same object class. Although such notions need to be confirmed by further study, they emphasize the fact that visual form analysis potentially involves dynamic interactions within, and across, multiple stages of cortical processing. The existence of localized brain sites that respond selectively to faces (or color or motion) does not preclude the capacity of these sites to function dynamically as part of a larger network whose operation can be modified to meet the informational needs of the observer.

Written Words and Dyslexia: Another potentially unique category of form analysis is the perception and interpretation of written words. Because this function is most highly developed in humans, it may distinguish us from other primates. Some neuroimaging studies have implicated ventral occipitotemporal cortex near, but distinct from, the FFA in the analysis of visual word form, though other studies have obtained different results. So far, then, it is unclear whether there are unique neural systems for the analysis of character and word shapes within the ventral form system.

Several studies have found an association between specific visual pathways and developmental dyslexia, an impairment of visual word processing and reading. Specifically, dyslexics appear to have impaired function of magnocellular-dominated pathways extending into the dorsal occipital cortex, especially through hMT+ to the parietal cortex. Neuroimaging studies of dyslexics have found reduced hMT+ responsiveness to both visual motion and drifting sinusoidal gratings presented at low average luminance. Such conditions stress the function of the magnocellular-dominated

dorsal stream. It has been proposed that such deficits may be related to the role of dorsal occipitoparietal systems in rapidly redirecting attention to specific targets in optically crowded environments such as printed text. If true, this suggests that dyslexia might reflect an abnormality of attentional control in addition to an inability to correctly interpret the visual form of the characters or words.

f. Depth and the Analysis of Space Compared to the concentrated work on motion and form processing, fewer human imaging studies have focused on the analysis and representation of space. The analysis of space involves inferences about a variety of spatial attributes. These include the assignment of distances of objects and surfaces relative to the self and to each other, the representation of spatial maps for navigation, the extraction of locations and orientations of objects relative to the limbs for purposes of manipulation, and, finally, the computation of object positions relative to the center of gaze or relative to the current focus of attention. (Studies concerning spatial attention, rather than the perception of space per se, are discussed in more detail in the section on attention.) Like the analysis of form and movement, the analysis of space can involve complex relationships between low-level visual cues derived from the retinal images and the various inferred attributes (see Fig. 10). In turn, this implies that there will be complex relationships in the neural pathways and cortical areas performing these analyses, including the likelihood of multiple concurrent pathways and computational hierarchies. The available evidence indicates that the cortical analysis of space heavily involves parietal and medial temporal lobe structures that receive direct or indirect projections from occipital lobe visual areas. A few such sites identified in neuroimaging studies are indicated in Fig. 8D. The following discussion mainly concerns the potential role of the occipital lobe in space-related computations.

It is clear from electrophysiological studies in animals that one important function of occipital visual areas is the extraction of relative depth–distance relationships from stereoscopic cues (retinal disparities). Although retinal disparity is not always sufficient to account for all perceived spatial attributes, nevertheless it is a major cue about distance relationships in the environment near the observer and, in combination with motion-parallax and perspective cues, typically provides for an unambiguous percept of spatial layout. Many monkey visual areas including V1, V2, V3, V3A VP, MT, and MST contain neurons that respond selectively to disparity differences. In humans, stimuli consisting of random dots alternating between zero and nonzero disparity elicit fMRI responses in the comparable human visual areas. Of these areas, V3A appears to be the most sensitive to changes in the disparity range, though all of the activated areas show a correlation between response magnitude and disparity size. One study also examined cortical responses to stimuli whose movements evoked the percept of either a 2-dimensional or 3-dimensional surface. Under passive viewing, the 3-dimensional stimuli evoked stronger fMRI activation in hMT+ and in several other foci, which (as nearly as can be determined from the report) included LO, V3A, and several parietal foci that are likely to receive projections from hMT+ or V3A. Although these studies identified visual areas that are responsive to depth cues and that may even represent the magnitude of perceived depth, it remains unclear which specific computational steps are associated with each area and which, if any, of these areas contains a comprehensive depth map of the visual scene.

Another approach to studying spatial relationships is exemplified by several imaging studies in which subjects matched the positions occupied by objects in a 2-dimensional array with a concurrently or previously presented positional cue. These tasks were typically compared to tasks in which subjects matched the identities of the objects regardless of position. In another study, subjects memorized the locations of objects in an array and then indicated whether their positions matched those of the same objects in a subsequently presented display. Most of these studies observed selective activation at one or more sites in dorsal occipital or parietal cortex (flat map D of Fig. 8). Some of these foci, especially in the superior parietal cortex, may be involved in the control of spatial attention, though it is quite possible that such sites may concurrently mediate the perception of spatial location itself. The visual areas associated with the activated sites in dorsal occipital cortex (two overlapping orange dots in the transverse occipital sulcus in Fig. 8D) were not determined, though their positions in Fig. 8D are consistent with V3A or a nearby visual area. Saccade-related activity has been found in similar locations in other studies, thereby supporting the possibility that this region may be eye-movement-related. This does not exclude a role in space perception because the constant shifting of the retinal images due to head and eye movements must be factored out to extract the invariant positions of objects relative to each other or to the self.

Accordingly, visual areas involved in the extraction of spatial information also may be involved in the analysis of eye movements. This is supported by the observation that parietal visual areas such as VIP and LIP in macaques contain cells that are active during different types of eye movements (smooth vs saccadic) and may be capable of compensating for such movements.

Other aspects of spatial processing such as the representation of mental maps of space appear to involve visual areas in the parietal, frontal, and medial temporal cortices outside the occipital lobe. Moreover, there are likely to be multiple representations of space using different coordinate frames, such as body- vs world-centered representations, rather than the retinotopic representations that are characteristic of many occipital visual areas. Given that there are a number of functionally distinct areas in macaque parietal cortex (e.g., VIP, LIP, MIP, PRR), it is likely that a similar multiplicity related, in part, to spatial processing will be described for human parietal cortex in the near future.

D. Summary: Principles of Functional Organization

From the preceding review of occipital lobe function, we can extract several general organizational principles. (1) Occipital visual cortex is composed of a number of distinct visual areas that, in many respects, resemble those of nonhuman primates such as the macaque monkey. At present, at least eight different human visual areas have been tentatively identified: V1, V2, V3–VP, V4v, V3A, hMT+ (MT plus neighboring areas such as V4t, MST, and FST), and V8. The topographic relationships among these areas are similar to those in macaques, though their positions relative to anatomical landmarks differ. Other functionally distinct zones that may or may not constitute additional visual areas include LO and KO. Also, there are additional zones of visually responsive cortex within the occipitotemporal and occipitoparietal cortices whose characteristics have not been fully explored (e.g., proposed V7). These zones may include additional homologs such as the ventral intraparietal area (VIP) or the parieto-occipital visual area (PO, also known as V6), but additional studies are needed for a comprehensive accounting. (2) Several stages of cortical visual processing (V1, V2, V3–VP, V4v, V3A, and V8) are organized retinotopically, though some areas may lack a complete representation of the visual field or may have somewhat disparate visual properties in upper vs lower field subdivisions (e.g., V4v, V3 vs VP). As in the macaque, retinotopic organization tends to become less distinct or absent in later stages of processing extending into the parietal and temporal lobes. (3) At each processing stage in the visual system there are multiple pathways that operate concurrently. By extension from animal data, these distinct pathways are undoubtedly interleaved on a fine scale within visual areas at early stages of processing (V1–V2) so that lesions there tend to produce virtually complete blindness, at least within the retinotopic confines of the damage. (4) There is a general tendency for spatial and motion-related tasks to activate dorsal sites extending into parietal cortex, whereas color and form tasks tend to activate ventral occipitotemporal sites. This is generally consistent with the principle of dorsal and ventral processing streams popularly known as "where" vs "what" pathways. However, there are examples that seem to defy this principle, such as the activation of ventral sites by motion stimuli. One key to understanding these seeming discrepancies is to clarify the distinction between the processing of low-level cues (e.g., local motion) and that of high-level attributes (e.g., motion-defined form). (5) All visual tasks produce activation throughout extended networks of visual areas, including medial occipital areas (V1–V2), plus a variable complement of extrastriate areas depending on the nature of the stimuli and the visual task. Neuroimaging data exist that are consistent with the concept that the different visual areas are arranged into a rough processing hierarchy with early stages tending to extract information about local visual cues and attributes, whereas later stages construct representations of increasingly global attributes. (6) Some specific cortical sites are preferentially activated by restricted classes of stimuli or tasks (V8–color, FFA–faces, hMT+–motion), but in all cases these areas can be activated to some extent by other classes of stimuli or tasks. The existence of functionally selective cortical sites that are critical for accurate perception is consistent with, and may explain, the fact that restricted lesions of extrastriate cortex can result in selective deficits in the discrimination of motion, color, faces, and possibly other attributes. Even so, such sites are components of extended processing networks, so that it is inaccurate to conceive of them as isolated centers or modules. Rather, such sites likely represent unique processing stages at which particular world attributes become explicitly represented in a localized population of neurons, thereby becoming vulnerable to the effects of localized cortical trauma.

IV. ADDITIONAL COGNITIVE ROLES

The preceding description of occipital lobe function was largely concerned with the neural systems involved in extracting the attributes of objects and surfaces from the incoming retinal information. However, analysis of the incoming retinal information in isolation cannot completely account for the more cognitive characteristics of vision. For example, studies of "biological motion" have demonstrated that the movement of a few lights attached to the joints of a person walking in an otherwise dark room can yield a vivid impression of the person and even allow identification of their gender. Such demonstrations suggest that the "bottom-up" analysis of the incoming visual information can be strongly affected by context and prior experience. At any given moment, we appear to be consciously aware of only a fraction of the visual information available in the retinal images. If we do not attend to a feature or event for at least a split second, we appear to be unaware of it (or at least unable to report or remember it). Together, these examples highlight the role of the observer as an active agent who is able to direct and interpret visual processing depending on the needs of the task at hand. We are still in the early stages of uncovering the neural mechanisms involved in these more complex aspects of vision, so the following discussion will only touch upon a few key cognitive functions in which the occipital lobe may participate.

A. Visual Imagery

One issue that has prompted a variety of experiments is the role of occipital visual cortex in mental imagery. In a review of experimental evidence concerning mental imagery, Farah concluded that there is "converging evidence that supports the modality-specific nature of mental images, and suggests that at least some of their neural substrates have a spatial representational format: Damage to visual areas representing such specialized stimulus properties as color, location, and form results in the loss of these properties in mental imagery." Particularly intriguing is the observation that the surgical removal of one occipital lobe appears to alter the size of the imagined visual field in a manner commensurate with the loss of a hemifield. If true, then this finding strongly suggests that occipital networks are essential for visual imagery.

A more specific question is whether mental images arise only from activity at high levels of cortical processing where, presumably, whole objects and scenes are represented. Alternatively, mental images might arise from the activation of all stages of visual processing in a way that more or less duplicates the cortical activity evoked by a real visual scene. As illustrated in Fig. 15, several studies have claimed to show activation of occipital visual areas during mental imagery, and a few have even implicated primary visual cortex. In one of these studies, subjects mentally compared the lengths of imagined stripes. This produced clear activation of V1. If transcranial magnetic stimulation (thought to directly stimulate cortical neurons) was applied to medial occipital cortex before the mental imagery task, then the subject's ability to perform the task was impaired relative to a sham control condition. However, other studies have failed to obtain clear activation of primary visual cortex during some types of imagery, thereby leading to the proposal that the neural mechanisms involved will vary depending on the specific imagery task. This suggests the possibility that a visual area may be activated if the mental image consists of a feature or attribute that is explicitly coded by the constituent neurons regardless of the area's hierarchical processing level. Whether this conjecture is true awaits further study. It is also unclear whether the activation evoked in visual cortex by mental imagery represents the imagined percept itself or rather the effects of a correlated process such as directed visual attention. Attention directed to a particular visual field location will modulate retinotopically corresponding sites in a number of occipital visual areas such as V1, V2, and V4–V8 (see later discussion). Presumably, this also occurs if attention is directed toward a mental image in order to note its features. In such a case, there is potential confusion between cortical activation representing a mental image and cortical activation associated with the processes used to explore and report on the imagined image.

B. Attention

Visual attention refers to the ability to prepare for, select, and maintain awareness of specific locations, objects, or attributes of the visual scene (or an imagined scene). The focus of visual attention can be redirected to a new target either reflexively or through the purposeful effort of the observer. The center of gaze typically follows the focus of attention, but the observer can intentionally dissociate the two, thereby demonstrating that the neural systems controlling

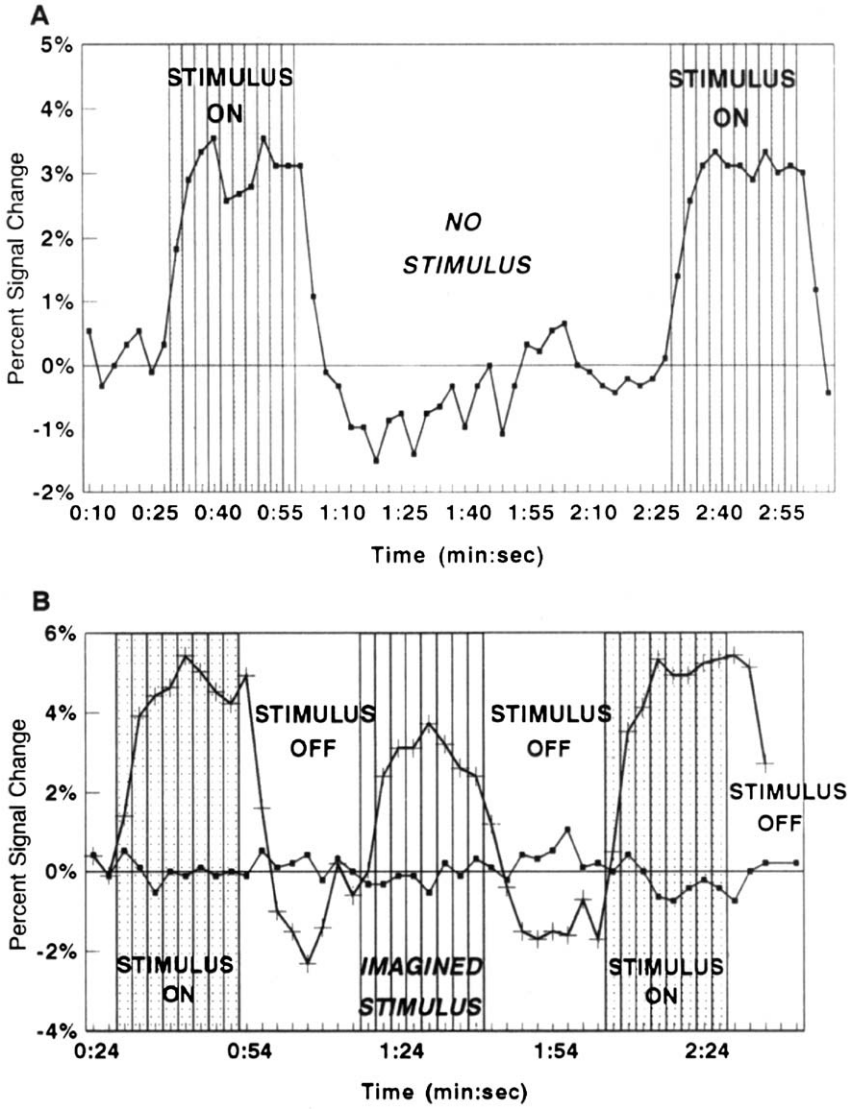


Figure 15 fMRI activation of striate cortex during mental imagery. Plots of relative fMRI signal intensity changes vs time. In A, the subject was resting between presentations of stimuli consisting of light emitting diodes arranged to form two squares. In B, the subject was asked to mentally recall an image of the stimulus in the period between the actual stimulus presentations. In this panel only, + = striate cortex and square symbols = nonstriate cortex. (From Le Bihan *et al.*, 1993.)

eye movements and attention are at least partially distinct.

The specific purpose and mode of action of attention are topics of significant debate. Direction of visual attention to a target can enhance the detection of subtle changes in brightness, color, or virtually any other attribute of interest. Changes in the visual scene not associated with the target are less readily detected unless they are sufficiently salient to reflexively redirect attention. Some visual attributes require focused attention to be perceived correctly. For example, detection of a conjunction of color and form (e.g., a red X in a field of red and blue X's and O's) requires the observer to attend to each letter in turn until the correct match is found. In a crowded visual environment, directed attention may be required if different attributes of the same object are to be correctly associated. There have been demonstrations that directed visual attention also can alter the perceived motion of certain stimuli, such as an ambiguously rotating "pinwheel" grating. In fact, quite substantial changes in the visual scene can go completely unnoticed if, for some reason,

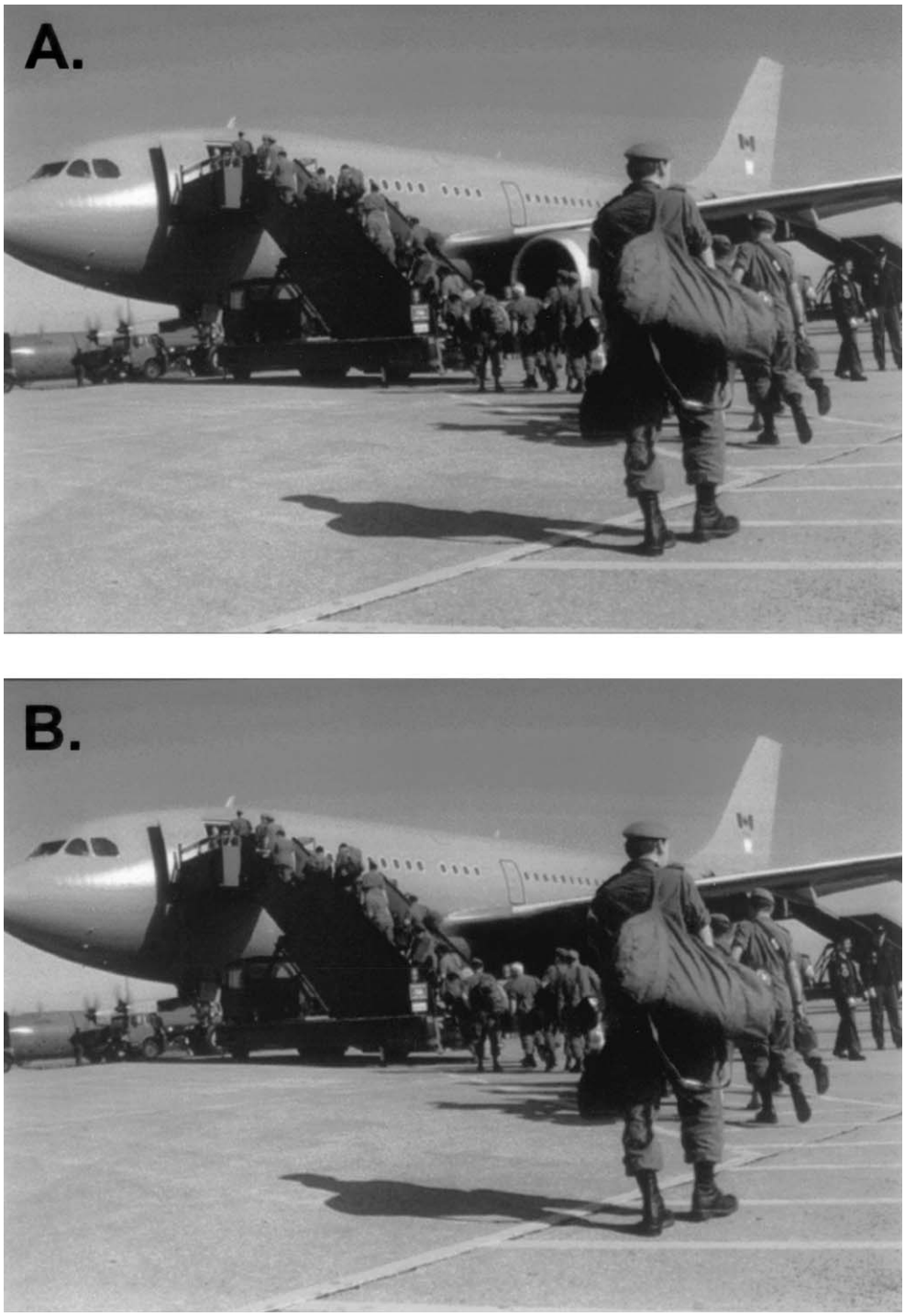


Figure 16 Stimuli used to demonstrate change blindness. The subject views A, and then the whole image flashes white at the moment that B replaces A. The two images are identical except for the missing jet engine. The subject will be unable to report the disappearance of the engine unless he or she is allowed to search the picture for an apparent discrepancy. A variety of other transient manipulations occurring at the moment of change can also block awareness of the change. It is theorized that the flash blocks the normal reflexive capture of attention by the change in the image. If the change does not capture attention, then there is no awareness. (Courtesy of R. Rensink.)

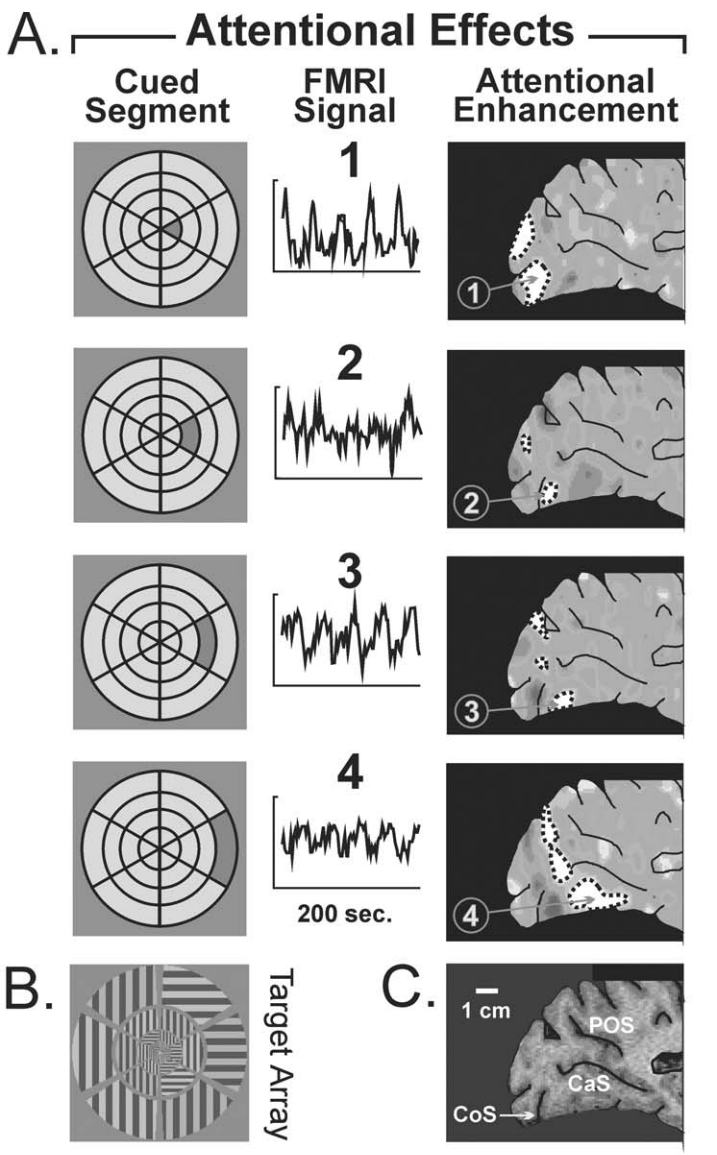
the reflexive systems that redirect attention are disabled at the moment the change occurs. This can be shown dramatically using a psychophysical procedure to induce "change blindness." By using the images shown in Fig. 16, it is easy to show that a flash occurring at the moment the images are switched can prevent a subject from detecting the disappearance of the jet engine from the wing of the airliner. It is thought that the flash reflexively redirects attention at the moment when the disappearance of the jet engine normally would have captured attention, thereby resulting in loss of awareness. (If the subject is allowed to peruse the scene after the switch, he or she can eventually infer the loss of the jet engine, though this can take considerable time.)

How and where attention exerts its effects over the processing of incoming visual information are becoming clearer. Ample evidence from animals and humans now shows that attention can modulate the responses of neurons in occipital visual cortex. Shifts of attention can alter the response amplitude or discriminability of cells in macaque visual areas such as V1, V2, MT, and V4. Attention also can selectively alter the responsiveness of neurons in V4 to one of two or more targets falling within its receptive field. In humans, electrophysiological studies have shown that evoked potentials associated with occipital visual cortex are enhanced if the subject actively attends to the location at which the stimulus is flashed. Moreover, the enhancement is localized to the occipital lobe contralateral to the attended site. Neuroimaging studies, such as the one illustrated in Fig. 17, have shown that a shift in attention from one visual target location to another results in focal enhancement of cortical activation at retinotopically corresponding sites throughout retinotopically organized occipital visual areas, including V1, V2, V3–VP, V4, and possibly V8 or neighboring cortex. Within these areas, the retinotopic locus of attentional enhancement precisely matches the retinotopic locus of activation evoked by the attended target when presented in isolation. This strongly suggests that attention exerts its spatially selective effects within occipital cortex, perhaps as early as V1 (though the latter may be due to a delayed effect relayed back from extrastriate cortex). Surprisingly, attention directed to the site of an expected stimulus will activate visual cortex before the target actually appears. This is consistent with the notion that attention is modifying responsiveness at the cortical site representing the cued location in preparation for the anticipated appearance of a target.

Although we tend to be quite familiar with directing our attention to a particular location, it is also possible to direct attention to a particular visual feature (e.g., noticing the shape vs the color of an object). By using PET imaging, researchers produced a seminal study demonstrating the featural specificity of cortical attentional modulation in humans. Attention directed to different attributes of an array of small, drifting rectangles of different shapes and colors resulted in different distributions of enhanced activation in occipital cortex, depending on the feature of interest (color, speed, shape). The patterns of activation tended to be functionally appropriate for the attended feature (e.g., attention to speed resulted in enhanced activation of hMT+). Subsequent studies have been able to show that specific visual areas such as hMT+, V4–V8, and FFA can be modulated by attention to visual attributes processed by the corresponding area (speed, hMT+; color, V4–V8; faces, FFA). Moreover, it has been demonstrated that attention to a given feature and attention to a given location can each have modulatory effects that are at least partially summative.

The preceding results suggest that the selection and enhancement of restricted aspects of the incoming visual information by visual attention occur, at least in

Figure 17 Retinotopic mapping of attention effects in the human occipital cortex. Shift in the focus of spatial attention without moving the eyes produces retinotopically corresponding foci of activation in the occipital cortex. (A) Left: Schematic representation of attended target locations. Subject fixated on the center of the target array (B) and shifted attention from the most central segment to successively more peripheral segments on the right (1–4) according to prearranged auditory cues presented briefly every 10 seconds. Middle: fMRI signals showing attention-related activation recorded at sites indicated by corresponding numbers on the right. Right: Parasagittal slice through the occipital lobe showing the foci of attentional activation (dotted outlines) that were displaced anteriorly (1–4) as the subject shifted attention to more peripheral target locations. (B) Target array. Orientation and colors of target segments changed randomly every 2 sec. No unique visual features marked the attended target segments. (C) Anatomical image. Abbreviations: CaS, calcarine sulcus; CoS, collateral sulcus; POS, parieto-occipital sulcus. (D) Quantitative comparison of visual field eccentricity (coded by temporal phase of fMRI response) for foci of attentional enhancement (y axis) vs foci activated by attended target segments presented alone (x axis). Each circle in the graph represents a single responsive voxel. Data are pooled across subjects. Left: Medial occipital cortex consisting primarily of V1 and V2. Right: Ventral occipitotemporal cortex within and surrounding the collateral sulcus. R = correlation coefficient. Dashed line indicates the locus of perfect correlation. An exceptionally high degree of correlation shows that the topography of the attentional enhancement within the visual cortex is precisely retinotopic. (Adapted from Brefczynski and DeYoe, 1999, *Nat. Neurosci.* **2,** 370–374.)


A.
Attentional Effects
Cued Segment
FMRI Signal
Attentional Enhancement
1
2
3
4
200 sec.
B.
Target Array
C.
1 cm
POS
CaS
CoS
D.
Attentional Retinotopy
Medial Occipital
Ventral Occipital
R= 0.94
R = 0.96
Attentional Phase
Single Seg. Phase
0
2

part, within occipital visual cortex. How these spatially and functionally restricted patterns of attentional enhancement are set up and impressed upon occipital cortex is not yet fully known, though the mechanism seems to involve interactions of occipital cortex with parietal and frontal cortices as well as with the pulvinar. This is an area of intense investigation, so that the underlying neuronal mechanisms are likely to be further clarified in the near future.

C. Visual Awareness

As is apparent from the foregoing discussion, visual attention is closely linked to the issue of visual awareness (though they are not synonymous). Given the extensive evidence for attentional effects in occipital cortex, one is easily led to ask whether occipital cortex directly participates in visual awareness. In other words, is our subjective perceptual experience a direct correlate of the activity of occipital neurons? Or, is occipital cortex merely extracting information from the retinal images and then passing it on to other brain regions that are responsible for awareness?

The answers to these questions are far from evident, but some relevant observations are beginning to accumulate. Several studies have used binocular rivalry in conjunction with single-unit recordings or neuroimaging to try to establish which neurons or visual areas show activity that correlates closely with perception. If an upward-drifting grating is presented to one eye but a downward-drifting grating is presented to the other eye, humans typically report the percept of one or the other and more rarely a combination of the two. In single-unit studies of monkeys performing this task, some direction-selective neurons in the superior temporal sulcus, including MT, responded in close correspondence with the alternating percept of the gratings, even though the visual stimuli themselves were unchanging. In humans, a similar rivalry experiment in which a picture of a face was presented to one eye while a house was present to the other eye resulted in alternate fMRI activation of the fusiform face area (FFA) vs the parahippocampal place area (a region in the medial temporal cortex shown to respond selectively to spatial structures) as the subject's percept alternated between face and house.

Despite these demonstrations of a close correspondence between awareness and activity in higher level visual areas, psychophysical studies indicate that at least some of the neural activity in occipital visual cortex is not accessible to consciousness. Orientation-specific adaptation to a visual grating patch can occur even when subjects cannot consciously report the orientation of the grating due to nearby distracters. Because the orientation-specific adaptation undeniably occurs (and is generally accepted to be a function of occipital neurons), it follows that the cortical activity representing the orientation-specific information itself is not accessible to conscious awareness. Furthermore, some neurons in macaque area MT apparently can discriminate the direction of motion of a noisy random dot display better than is indicated by the behavioral response. This, too, suggests that not all neural activity in the occipital cortex is capable of generating a conscious percept. Indeed, it has been proposed that consciousness may only arise when occipital activity is transferred to frontal cortex. In such a case, attention may act as a simple gating mechanism that controls the transference of sensory information to "awareness systems" in the frontal cortex. However, the available evidence does not yet support the conclusion that conscious awareness arises exclusively outside the occipital lobe or that occipital circuits never directly participate in awareness. An alternative possibility is that, at any given moment, assemblies of neurons in various locations are dynamically linked together to create a substrate for awareness. In such a scheme, attentional control mechanisms (in parietal cortex and the pulvinar?) dynamically link specific occipital circuits with frontal neurons to form an "active" configuration that generates awareness. In this fashion, neurons in a variety of cortical areas, including the occipital lobe, could transiently participate in awareness as mental events shift from one thought to another over time. At present, the validity of any of these concepts has yet to be firmly established.

V. CONCLUSION

Thanks in large part to the advent of sophisticated neuroimaging techniques, there has been a rapid increase in our understanding of the functional organization of the human occipital lobe. This progress has been built on decades of previous work in humans and animals and now permits observations from different primate species to be compared and interrelated as never before. So far, what has emerged is the realization that, in many respects, occipital lobe organization in humans is similar to that in monkeys, though we are now beginning to see indications of

potential differences. The extent of these differences and their ultimate functional importance are not yet clear, but they hold the key to a more specific understanding of the extent to which humans do or do not differ from other species in brain organization. Due to the unprecedented ability to probe brain function in humans as they perform complex perceptual and cognitive functions, it is now possible to identify and characterize the specific neural substrates of our more complex and fascinating visual abilities. Practically, this gives us new ways to understand how sensation and perception operate and to understand how they are disturbed by brain pathology. The advances in basic research have already prompted exploratory applications of brain mapping techniques in presurgical planning and in the diagnosis of visual dysfunction. Altogether, the progress of recent years has yielded a more comprehensive, predictive, and clinically relevant understanding of human occipital lobe function and, by extension, a better understanding of the relationship between brain and mind.

See Also the Following Articles

AREA V2 • MOTION PROCESSING • NEOCORTEX • PROSOPAGNOSIA • RECEPTIVE FIELD • VISUAL CORTEX • VISUAL DISORDERS

Acknowledgments

The author thanks the many colleagues who submitted materials for this article and expresses regret that only a few of the submitted illustrations could be used. The author expresses his particular appreciation to Jon Wieser for extensive technical assistance and to Dr. David Van Essen for illustrations from the Visible Man database and for use of the Caret and FlatMorph software packages. This work was supported by NIH Grants EY10244 and MH51358 to E.A.D. and by General Clinical Research Center Grant M01 RR00058 and Vision Core Grant P30 EY01931.

Suggested Reading

Callaway, E. M. (1998). Local circuits in primary visual cortex of the macaque monkey. *Ann. Rev. Neurosci.* **21,** 47–74.

Casagrande, V. A., and Kaas, J. H. (1994). The afferent, intrinsic, and efferent connections of primary visual cortex in primates. In *Cerebral Cortex* (A. Peters and K. Rockland, Eds.), Vol. 10, pp. 201–259. Plenum Press, New York.

DeYoe, E. A., and Van Essen, D. C. (1988). Concurrent processing streams in monkey visual cortex. *Trends Neurosci.* **11**(5), 219–226.

Felleman, D. J., and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex* **1,** 1–47.

Horton, J. C. (1992). The central visual pathways. *Adler's Physiology of the Eye*, pp. 728–772. Mosby–Year Book, St. Louis.

Kaas, J. (1997). The organization of visual cortex in primates: Problems, conclusions, and the use of comparative studies in understanding the human brain. In *Cerebral Cortex* (K. Rockland, Ed.), Vol. 12, Plenum Press, New York.

Zeki, S. (1990). A century of cerebral achromatopsia. *Brain* **113,** 1721–1777.

Zilles, K., and Clarke, S. (1997). Architecture, connectivity, and transmitter receptors of human extrastriate visual cortex: Comparison with nonhuman primates. In *Cerebral Cortex* (K. Rockland, Ed.), Vol. 12, pp. 673–742, Plenum Press, New York.

Olfaction

RICHARD L. DOTY
University of Pennsylvania

GLOSSARY

anosmia Inability to perceive odors.

dysosmia (parosmia) Distorted smell sensations.

hyposmia (microsmia) Decreased ability to perceive odors.

olfactory bulb Multilayered structure at the base of the brain where incoming olfactory information is first processed.

olfactory fila Collections of olfactory nerve cell axons formed by ensheathing glial cells that pass from the nasal to the brain cavity via the foramina of the cribriform plate.

olfactory nerve Collectively the ~6,000,000 odorant-sensitive receptor cells located within the olfactory neuroepithelium that send axons to the olfactory bulb.

olfactory neuroepithelium The epithelium in the upper recesses of the nose that contains, among other cells, the olfactory receptor cells where airborne environmental chemicals activate neural responses.

olfactory receptors Seven-transmembrane proteins specialized to bind odorants and transduce information that leads to neural firing and ultimately the perception of odors.

olfactory tract Fiber bundle at the base of the brain through which afferent and efferent connections pass between the olfactory bulb and higher brain structures.

Olfaction largely determines the flavor of foods and beverages and serves as an early warning system for the detection of environmental hazards, including spoiled foods, leaking natural gas, smoke, and various pollutants. This primary sensory system contributes to our quality of life, allowing us to more fully appreciate flowers, perfumes, culinary creations, the sea shore, the mountains, and the seasons of the year. From a medical perspective, smell dysfunction can be the first clinical sign of such neurodegenerative disorders as Alzheimer's disease and idiopathic Parkinson's disease, making smell testing of particular value in neurological diagnosis. This article provides basic information on the anatomy, physiology, and function of the olfactory system, emphasizing quantitative measurement and clinical applications.

I. ANATOMY

A. Intranasal Neural Systems

Three specialized neural systems are present within the left and right sides of the human nose: (i) the main olfactory system (Cranial Nerve I or CN I); (ii) the trigeminal somatosensory system (CN V); and (iii) the nervus terminalis or terminal nerve (CN 0). CN I mediates what we commonly term odor sensations (e.g., chocolate, strawberry, apple, etc.), whereas CN V mediates, via both chemical and nonchemical stimuli, somatosensory sensations, including irritation, tickling, cooling, and burning. The cool sensations of menthol are mediated by CN V, as are the sharp burning sensations of ammonia. CN 0, whose function is largely unknown, was discovered after the other

cranial nerves had been named and has remarkably constant anatomy across all vertebrates, including humans. Its peripheral component is a loose neural plexus distinguished by ganglia at nodal points. CN 0, notable for its high gonadotropin-releasing hormone (GnRH) content, ramifies throughout the nasal epithelium before coalescing into bundles that eventually traverse the cribriform plate to enter the forebrain.

Despite the fact that nearly all adult humans possess, in the lower recesses of each nasal chamber, a tube-like rudimentary vomeronasal organ (VNO) and a VNO duct approximately 15–20 mm from the posterior aspect of the external naris, they lack an accessory olfactory bulb and their vomeronasal system apparently is not functional. Attempts to trace neural connections from this organ to the brain have been uniformly unsuccessful in humans, despite the fact that local electrophysiological responses have been recorded within the VNO lumen. In contrast to the receptor-rich medial and receptor-free (but vascular-rich) lateral elements of the VNOs of species with a functioning vomeronasal organ system, the human VNO has comparatively homogeneous epithelium along both its medial and lateral aspects. Adult human VNOs contain cilia and short microvillae, unlike the elongated microvillae typical of functional VNOs.

B. The Olfactory Neuroepithelium

The olfactory epithelium, which contains the sensory receptors of the main olfactory system, lines the upper recesses of the nasal chambers, including the cribriform plate, superior turbinate, superior septum, and sectors of the middle turbinate. This neuroepithelium loses its general homogeneity postnatally as islands of respiratory-like epithelia appear a few weeks after birth, presumably as a result of insults from environmetal toxins, bacteria, and viruses. Such islands accumulate across the lifespan. Surprisingly, the size of the epithelium in humans is still not well-established, and there is some evidence that it may extend farther onto the middle turbinate than previously believed.

The mature olfactory epithelium comprises at least six biochemically and morphologically distinct cell types (Fig. 1). The first cell type—the *bipolar sensory receptor neuron*—contains the cilia that bind the odorants as they permeate the olfactory mucus. Collectively, these cells constitute CN I. In aggregate, the surface area of the cilia of these cells is quite large, exceeding 22 cm^2 in humans. The number of olfactory receptor cells ($\sim$6,000,000 in the adult) exceeds that of any other sensory system except vision. The second cell type is the *microvillar cell*, whose cell body is located near the epithelial surface. These cells, which occur in $\sim$1:10 ratio with the bipolar receptor cells, have microvillae at their apical surfaces and resemble the "brush" cells of the upper and lower airways of many species. Whether microvillar cells play any role in chemosensation is not clear, although preliminary *in vitro* patch clamp studies of dissociated microvillar cells suggest that they are not responsive to odorants. The third cell type is the *supporting* or *sustentacular cell*. Like the microvillar cells, these cells project microvillae into the mucus. Supporting cells appear to (i) insulate the receptor cells from one another, (ii) regulate the local ionic composition of the mucus, (iii) deactivate odorants, and (iv) help protect the epithelium from damage from foreign agents. These cells contain xenobiotic-metabolizing enzymes (e.g., cytochrome P-450), a feature shared with the acinar and duct cells of Bowman's glands, the primary source of mucus in the region of the olfactory epithelium. The fourth cell type is the *cell that lines the Bowman's glands and ducts*, whereas the fifth and six cell types are the *globose* (*light*) *basal cell* and *horizontal* (*dark*) *basal cell*—cells located near the basement membrane from which most of the other cell types arise. The same type of basal cell, most likely a globose cell, can give rise to neurons and nonneural cells when the olfactory epithelium is markedly damaged, expressing a multipotency rarely observed in stem cells.

The cilia of the bipolar receptor cells differ from the cilia of the cells making up the respiratory epithelium in that they are much longer and lack dynein arms (hence, intrinsic motility). It is these cilia that contain the seven-domain transmembrane receptors that interact with incoming odorants. Movement of odorants through the mucus to the cilia is aided by transporting molecules termed "odorant-binding proteins." Approximately 1000 putative odorant receptors are believed to exist, reflecting the expression of the largest known vertebrate gene family. In fact, this gene family accounts for $\sim$1% of all expressed genes. *In situ* hybridization studies suggest that the olfactory receptors of rodents are topographically organized into four strip-like zones that roughly parallel the dorsal–ventral axis of the cribriform plate. Receptors of a given type are largely confined to one of these zones. Whether the situation is analogous in humans is not yet known. Work employing a scanning electron microscope has shown a possible morphological correlate to these zones in rats—by embryonic day 16, the posterior

Figure 1 Low-power electron photomicrograph (× 670) of a longitudinal section through a biopsy specimen of human olfactory mucosa taken from the nasal septum. Four cell types are indicated: ciliated olfactory receptors (c), microvillar cells (m), supporting cells (s), and basal cells (b). The arrows point to ciliated olfactory knobs of the bipolar receptor cells. Abbreviations: d, degenerating cells; bs, base of the supporting cells; lp, lamina propria; n, nerve bundle; bg, Bowman's gland. Photo courtesy of Dr. David T. Moran.

regions (roughly corresponding to zones 1 and 2) have much higher receptor cell knob densities than the more anterior regions (corresponding to zones 3 and 4). The supporting cell microvilli appear to be longer in region 1 than in region 2, and the apical ends of the cells adjacent to the receptor cells are flatter in regions 1 and 2 than in regions 3 and 4. Regions 3 and 4 also have glandular openings and scattered microvillous cells similar in appearance to the hair cells of the inner ear.

C. The Olfactory Bulb

In humans, the axons of the bipolar receptor cells coalesce into 30–40 fascicles, termed the olfactory fila,

formed by ensheathing glia. The fila traverse the cribriform plate and pia mater. The incoming axons make up the first layer of the olfactory bulb and synapse with dendrites of the second-order neurons of the bulb in spherical structures termed glomeruli. The major discernible concentric layers of the olfactory bulb are the olfactory nerve layer, the glomerular layer, the external plexiform layer, the mitral cell layers, the internal plexiform layer, and the granule cell layer. In the rat and mouse, neurons expressing a given receptor type typically project their axons to one or, at most, two glomeruli. Thus, a given odorant activates a spatially defined or restricted set of glomeruli. Hence, the olfactory code is reflected, at this early stage, as different patterns not only across the mucosa but across the glomeruli as well. Glutamate appears to be the main, if not sole, neurotransmitter of the receptor cells.

The activity of the mitral and tufted cells is modulated by other cells within the bulb, including periglomerular cells and more deeply located granule cells. The latter cells are the most numerous cells of the olfactory bulb, and centrifugal fibers from higher brain centers serve to alter the activity of these cells. It is generally believed that both intrinsic and extrinsic feedback processes sharpen or alter the neural information transmitted from the bulb to more central olfaction-related structures. The primary second-order neurons of the bulb, most notably the mitral and tufted cells, project directly to the primary olfactory cortex via the olfactory tract without synapsing with the thalamus. Although the olfactory tract, which is relatively flat posteriorly and becomes the olfactory trigon just rostral to the anterior perforated substance, is commonly divided into ''lateral'' and ''medial'' aspects in textbooks of anatomy, there is, in fact, no medial tract in primates.

D. The Olfactory Cortex

The olfactory cortex comprises (i) the anterior olfactory nucleus (AON), (ii) the olfactory tubercle (poorly developed in humans), (iii) the prepiriform cortex, (iv) the lateral entorhinal cortex, (v) the periamygdaloid cortex (a region contiguous with the underlying amygdala), and (vi) the cortical nucleus of the amygdala. Major connections between the primary olfactory cortex and the secondary olfactory cortex in the orbitofrontal region occur via the mediodorsal nucleus of the thalamus, as well as via direct cortico-cortical projections from prorhinal cortex to the posterolateral orbitofrontal region.

The AON is located in a rostral segment of the olfactory bulb, as well as in a segment of the olfactory peduncle near the anterior perforated substance. The dendrites of the pyramidal cells of the AON receive synapses not only from the olfactory bulb but also from both ipsilateral and contralateral brain structures, including the contralateral AON and elements of the olfactory cortex via the anterior segment of the anterior commissure. The anterior commissure represents the first CNS structure through which olfactory information crosses contralaterally.

Extensive interactions occur among cells comprising the superficial and deeper laminae of each component of the olfactory cortex, as well as among the components themselves. In addition to receiving direct projections from the olfactory bulb, the entorhinal cortex also receives input from both the prepiriform and the periamygdaloid cortices. The entorhinal cortex sends projections to a number of cortical structures, including the hippocampus. However, no direct pathways exist between the hippocampus and either the olfactory bulb or the AON, and anatomic, physiologic, and ontogenic evidence suggests that the hippocampus is not a main component of the olfactory system, even though it does play a role in odor memory.

II. TRANSDUCTION MECHANISMS

A. Olfactory Receptors

During the past 10 years considerable progress regarding the initial events of olfactory transduction has occurred, beginning with the aforementioned identification of the gene family that encodes olfactory receptors. Despite the fact that a given receptor cell seems to express only one type of receptor derived from a single allele, each cell is electrophysiologically responsive to a wide, but circumscribed, set of stimuli. This suggests that a single receptor accepts a range of molecular entities and that a complex cross-fiber patterning of responses provides the neural code. Odorant binding leads to an inwardly depolarizing current within the cilia of the bipolar receptor cells, which leads to the triggering of the action potentials that collectively provide the information that is forwarded to higher brain centers.

Two approaches for functionally characterizing odorant receptors have been employed. In one, an adenovirus-mediated gene transfer procedure was used to increase the expression of a specific receptor

gene in rat olfactory receptors, demonstrating ligand-specific increases in the amplitude of summated electrical activity at the surface of the epithelium (termed the electro-olfactogram or EOG). In the other, an olfactory receptor library was generated using a polymerase chain reaction from which cloned receptors were screened for odorant-induced responsiveness to a panel of odorants (as measured by an assay of intracellular Ca^{2+} changes). Several receptor types with ligand specificity were found, including one differentially sensitive to the stereoisomers of citronellal.

B. Transduction Pathways

In general, olfactory receptor proteins are linked to the stimulatory guanine nucleotide-binding protein G_{olf}. When stimulated, they activate the enzyme adenyl cyclase to produce the second messenger adenosine monophosphate (cAMP). G_{olf}-induced cAMP activates cellular depolarization via the opening of cyclic-nucleotide-gated ionic channels and Ca^{2+}-dependent Cl^- or K^+ channels. The amount of adenylate cyclase activity produced by various odorants in a frog ciliary preparation is positively correlated with the magnitude of the frog's EOG, as well as with the perceived intensity of these same odorants to humans. Cyclic guanosine monophosphate (cGMP) is also activated by some odorants. cGMP is believed to play a role in the modulation of the sensitivity of olfactory receptor neurons, such as during adaptation. Although olfactory receptor cells express G proteins other than G_{olf} (e.g., G_{i2} and G_o), they probably are not involved in early transduction events. More likely they assist such processes as axonal signal propagation, axon sorting, and target innervation.

Before the discovery of G_{olf}, another G protein, $G_{s\alpha}$, was believed to play a major role in the initial phases of olfactory transduction in humans. Evidence for such a role came from findings of variably decreased olfactory ability in type Ia pseudohypoparathyroidism (PHP), a disease associated with generalized hormone resistance and a deficiency of $G_{s\alpha}$, as measured in erthrocytes. However, this disorder has other features that could cause or contribute to olfactory dysfunction, including Albright hereditary osteodystrophy (AHO), an unusual constellation of skeletal and developmental deficits. Whereas a more recent study has confirmed that PHP type Ia patients perform poorly on tests of odor detection, identification, and memory, patients with type Ib PHP, who have no AHO, no generalized hormone resistance, and normal $G_{s\alpha}$ activity, also exhibited olfactory dysfunction relative to matched controls. Furthermore, patients with pseudopseudohypoparathyroidism (PPHP), who have AHO, no generalized hormone resistance, and deficient $G_{s\alpha}$ protein activity, were found to have relatively normal olfactory function. These observations suggest that the olfactory dysfunction associated with PHP is not the result of generalized $G_{s\alpha}$ protein deficiency and imply that other mechanisms are responsible for the olfactory deficits of this disorder. Whether PHP patients are deficient in epithelial G_{olf} has not been determined.

Whereas a second transduction pathway—i.e., that which activates the enzyme phospholipase C to produce the second messenger inositol triphosphate (IP_3)—has been implicated in odor-mediated responses in some vertebrates, its involvement in mammalian olfaction is controversial. Recent studies have employed knockout mice in which genes responsible for both the cyclic-nucleotide-gated ion channel and G_{olf} have been deleted. In this mouse, EOG responses to all odors tested were eliminated, including those previously believed to be mediated by the IP_3 system. To date, IP_3-gated channels have not been demonstrated in mammalian olfactory nerve cells using patch clamp techniques.

III. OLFACTORY RECEPTOR CELL REGENERATION

The sensory cells of the olfactory epithelium have a propensity to replace themselves after damage, unlike the sensory neurons of most other major sensory systems. It was previously thought, on the basis of [^{3}H]thymidine studies, that the entire olfactory neuroepithelium undergoes complete cell turnover approximately every 30 days. However, it is now known that the situation is much more complex. Thus, long-lived receptor cells have been identified despite continuous neurogenesis within the basal sectors of the neuroepithelium, and endogenous and exogenous factors have been found that promote receptor cell death or replenishment from stem cells. There is evidence that some receptor cells of older animals may live longer than those of younger animals.

Subgroups of stem cells can be made to differentiate into mature olfactory receptor cells by stressing them mechanically or biochemically. Furthermore, differentiated neurons, in effect, send back regulatory signals to inform progenitor cells as to the number of

new neurons that need to be produced to maintain cell population equilibrium. Apoptotic cell death has been observed in cells representing all stages of regeneration (e.g., in proliferating neuronal precursors, immature olfactory receptor neurons, and mature olfactory receptor neurons), implying apoptotic regulation of neuron numbers at all levels of the neuronal lineage. The mitral cells of the bulb appear to contain a trophic substance that helps to maintain the survival of olfactory receptor neurons. Currently, there is considerable interest in chemical factors that inhibit (e.g., fibroblast growth factor-2, bone morphogenetic proteins, dopamine) or promote (e.g., transforming growth factor-α, olfactory marker protein) neurogenesis or differentiation or actively produce apoptotic cascades (e.g., tumor necrosis factor-α, Fas ligand).

It is of interest that the olfactory ensheathing cells, which form the bundles of axons that make up the olfactory fila, have unique properties useful in the repair or regeneration of both central and peripheral nerves. Thus, they enhance remyelination and axonal conduction in demyelinated spinal tract nerves, as well as in severed rat sciatic nerves, exhibiting both Schwann-cell-like and astrocyte-like properties.

IV. FUNCTIONAL IMAGING STUDIES OF THE HUMAN OLFACTORY SYSTEM

The development of functional imaging techniques, such as positron emission tomography (PET) and functional magnetic resonance imaging (fMRI), has made it possible to determine, *in vivo*, brain regions influenced by odors, as well as by behaviors associated with smelling (e.g., sniffing). Although the spatial resolution of these techniques limits the degree to which activity in some structures can be visualized (e.g., the olfactory bulbs), odor-induced activation of most of the major olfactory-related cortical structures has been observed, albeit in some cases sporadically. Such activation includes the piriform cortex, the orbitofrontal cortex, and the inferior frontal lobe. Work suggests that the inconsistent fMRI activation of primary olfactory cortex (POC) may reflect sampling factors and temporal components of activation of these regions.

In accord with basic psychophysical observations are fMRI studies demonstrating decreased activation with age, as well as greater activation in women than in men. Also in accord with psychophysical findings are observations that the right hemisphere may be more specialized than the left for central olfactory processing, particularly regions within the orbitofrontal cortex. Odor familiarity judgments have been associated with increased rCBF in the right orbitofrontal area, the subcallosal gyrus, the left inferior and superior frontal gyri, and the anterior cingulate cortices. Edibility judgments made based on odors appear to be associated with selective activation in the primary visual areas, implying that visual imagery is evoked by this task.

An intriguing discovery from several functional imaging studies is that odors reliably and significantly activate regions of the human cerebellum, a structure classically considered to be involved primarily in motor learning. The cerebellar activity was unexpected and serendipitous in light of current knowledge of the olfactory projection pathways and views of the cerebellum, even though functional imaging studies have implicated this structure in a broad range of sensory and cognitive processing tasks. Odors appear to activate, in a concentration-dependent manner, largely posterior lateral areas of the cerebellum, whereas sniffing by itself tends to activate mainly anterior cerebellar regions.

V. PSYCHOPHYSICAL MEASUREMENT OF HUMAN OLFACTORY FUNCTION

During the past few years, psychophysical tests of olfactory function have become commonplace in medical, industrial, and academic centers, largely because of their widespread commercial availability. Traditionally, a determination of the lowest odorant concentration that can evoke a sensation or discernible odor quality was the primary means for assessing the ability to smell. Today, straightforward tests of odor identification and memory are widely used for this purpose.

A. Tests of Odor Identification

The proliferation of easy-to-use commercially available tests of odor identification has significantly increased our understanding of smell function in humans, including the influences of such factors as age, gender, exposure to toxic agents, and various disease states on its function. Such tests, most of which are self-administered, include a 3-odor Pocket Smell Test™, a 12-odor Brief-Smell Identification Test™, and the University of Pennsylvania Smell Identification Test (UPSIT, commercially known as the Smell

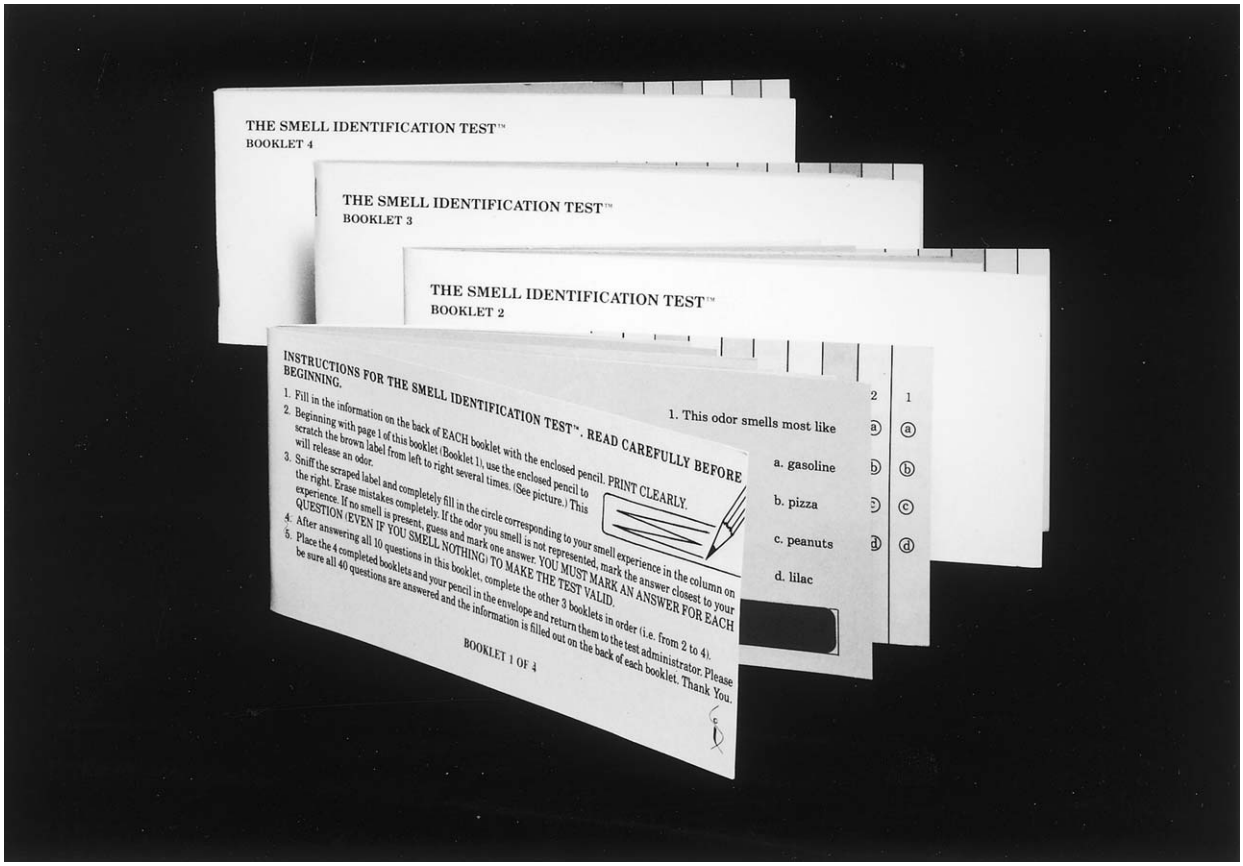


Figure 2 Picture of the University of Pennsylvania Smell Identification Test, the most widely used quantitive test for evaluating smell function. This test has been administered to nearly 200,000 persons over the past 15 years. Photo courtesy of Sensonics, Inc., Haddon Heights, NJ.

Identification Test™). Of these tests, the UPSIT has been employed the most widely, having been administered to ~180,000 people in Europe and North America in the past decade. This microencapsulated odorant test, shown in Fig. 2, employs norms based upon nearly 4,000 persons and is available in English, French, German, and Spanish language versions. In this test, the subject is asked to release, using a pencil tip, a microencapsulated odorant located on a page of a test booklet and to indicate its smell by responding to four alternatives. The subject's smell ability is divided into four categories of dysfunction (anosmia and mild, moderate, or severe microsmia) and one category of function (normosmia). Probable malingering is also defined on the basis of improbable responses (i.e., low test scores suggestive of avoidance of the correct responses). This highly reliable test (test–retest $r = 0.94$) was the impetus for a massive smell function survey sent to nearly 11 million subscribers of *National Geographic Magazine* in 1986.

B. Olfactory Threshold Measures

Before the advent of odor identification tests, olfactory threshold measures were the most common means employed for assessing human olfactory function. Today, such procedures are still used, although not as frequently as odor identification tests.

The lowest concentration of an odorant that can be reliably detected (usually defined as that concentration where detection is midway between chance and perfect detection) is termed the *detection* or *absolute threshold*. At very low concentrations, no odor quality can be discerned, only that something is present that differs from air or the comparison diluent blank or blanks. In modern olfactory detection threshold testing, the subject is asked to report which of two or more stimuli (i.e., an odorant and one or more blanks) smells strongest, rather than to simply report whether an odor is perceived. Such forced-choice procedures are less susceptible to contamination by response biases

(e.g., the conservatism or liberalism in reporting the presence of an odor under uncertain conditions) than non-forced-choice procedures. In addition, they are more reliable and produce lower threshold values. The instructions provided to a subject are critical in measuring a detection threshold, because if the subject is instructed to report which stimulus produces an odor rather than which stimulus is stronger, a spuriously high threshold value may result (odor quality is present only at higher perithreshold concentrations).

The *recognition threshold* is defined as the lowest concentration at which odor quality is reliably discerned. Unfortunately, it is nearly impossible to control criterion biases in recognition threshold measurements. Thus, in a forced-choice situation, guesses are not randomly distributed among alternatives, potentially leading to a spuriously low recognition threshold for the preferred alternative. A classic example of this problem comes from taste psychophysics, where some subjects report ''sour'' much more frequently than the other primary qualities in the absence of a clearly discernible stimulus, resulting in an erroneously low sour taste recognition threshold measure.

A third type of threshold is the *difference threshold*, the smallest amount by which a stimulus must be changed to make it perceptibly stronger or weaker. This is also termed a differential threshold or a just noticeable difference (JND). The size of the increment in odorant concentration (ΔI) required to produce a JND increases as the comparison concentration (I) increases, with the ratio approximating a constant, i.e., $\Delta I/I = C$.

Two types of procedures for determining detection thresholds have received the most use in the past two decades: the ascending method of limits procedure (AMS) and the single staircase procedure (SS). In the AML procedure, odorants are presented sequentially from low to high concentrations and the point of transition between detection and no detection is estimated. In the SS method, the concentration of the stimulus is increased following trials on which a subject fails to detect the stimulus and decreased following trials where correct detection occurs. An average of the up–down transitions (''reversals'') is used to estimate the threshold value. In both the AML and SS procedures, the direction of initial stimulus presentation is made from weak to strong in an effort to reduce potential adaptation effects of prior stimulation.

It has generally been believed that threshold measures exhibit considerable variability. Unfortunately, this conclusion is based mainly upon studies employing unreliable AML techniques for measuring detection thresholds. Procedures employing more trials, such as the SS procedure, produce less variable and more reliable measures than simple AML procedures and negate the notion that thresholds vary markedly from day to day within an individual.

A modern, commercially available threshold test kit that employs a SS procedure is shown in Fig. 3. This kit utilizes squeeze bottles containing various concentrations of an odorant known to stimulate primarily CN I. Norms based upon hundreds of subjects spanning the entire age range allow for the practical application of this test for medical and industrial uses.

Among the major findings derived from modern olfactory tests, primarily the UPSIT, are the following: first, there is a substantial genetic influence on the ability to identify odors; second, women, on average, have a better sense of smell than men, and this superiority is noticable as early as 4 years of age and is culture-independent; third, loss of significant olfactory function occurs after the age of 65 years, with over half of those between 65 and 80 years of age and over three-quarters of those 80 years of age and older having such loss (Fig. 4); fourth, women, on average, retain the ability to smell longer than men; fifth, the decreased smell ability associated with smoking is present in past smokers and recovery to presmoking levels, while possible, can take years depending upon the amount and duration of past smoking; and sixth, olfactory function is compromised in urban residents and in workers in some industries, including the paper and chemical manufacturing industries.

Clinical studies employing odor identification tests during this period have found decreased smell function, relative to matched controls, in dozens of diseases and disorders, including alcoholism, amyotrophic lateral sclerosis (ALS) and related forms of motor neuron disease, the ALS–parkinsonism–dementia complex of Guam (Guam disease or GD), attention deficit hyperactivity disorders, Alzheimer's disease (AD), estrogen-receptor-positive breast cancer, severe stage anorexia nervosa, chronic obstructive pulmonary disease, cystic fibrosis, Down's syndrome, epilepsy and surgical procedures designed to control intractible seizure activity, head trauma, human immunodeficiency virus, Huntington's disease, Kallmann's syndrome, Korsakoff's psychosis, multiple sclerosis, multiple system atrophy, nasopharyngeal carcinoma, nasosinus disease, Parkinson's disease, psychopathy, restless leg syndrome, schizophrenia, schizophrenia-like affective disorders, schizotypy, seasonal affective disorder, and Sjögren's syndrome. Additionally, a

Figure 3 A modern test kit for determining an odor detection threshold. This test kit comprises of bottles containing half-log concentrations of an odorant or blank. The pairs of stimuli (blank vs odorant) at a given concentration are presented in a randomized order, and a single staircase presentation procedure is employed across trials. Photo courtesy of Sensonics, Inc., Haddon Heights, NJ.

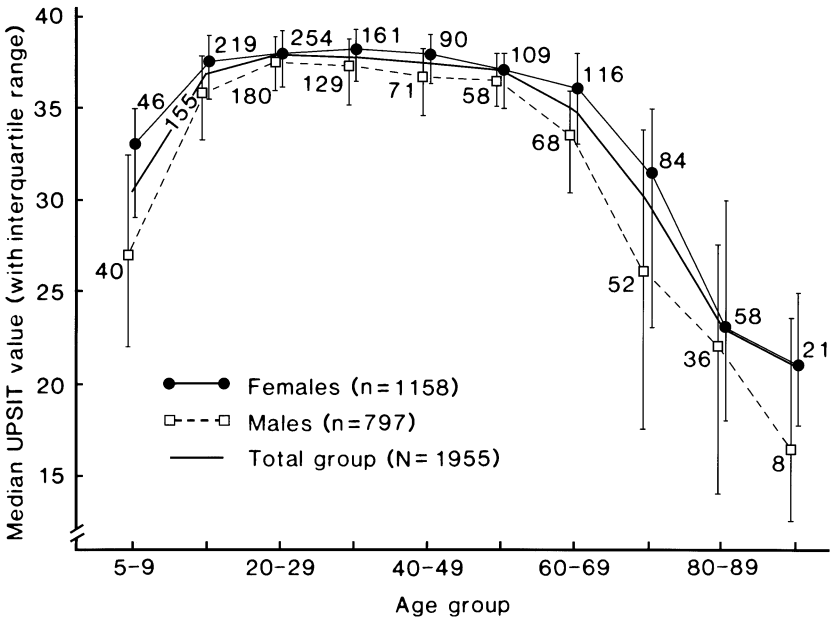


Figure 4 Scores on the University of Pennsylvania Smell Identification Test (UPSIT) as a function of age and gender in a large heterogeneous group of subjects. Numbers by data points indicate sample sizes. From Doty *et al.* (1984). Smell identification ability: Changes with Age. *Science* **226,** 1441–1443. Copyright © American Association for the Advancement of Science.

number of studies have employed these tests in evaluating the effects of surgical or radiological interventions on the ability to smell. Psychophysical testing has found no meaningful smell losses in patients with corticobasal degeneration, depression, panic disorder, progressive supranuclear palsy, MPTP(1-methyl-4-phenyl-1,2,3,6-tetrohydropyridine)-induced parkinsonism, essential tremor, or multiple chemical hypersensitivity (MCS).

The ability to quantify olfactory function using such tests, along with the aforementioned advances in *in vivo* medical imaging, has made it possible to better understand the physiological basis of olfactory loss in some patients. For example, it is now apparent that congenital anosmia is associated with markedly deformed or absent olfactory bulbs and stalks. Furthermore, head-trauma-related smell loss is typically accompanied by contusions of the frontal and temporal poles of the brain, as well as diminuition in the size of the olfactory bulbs and tracts. The latter phenomenon may reflect mitigation of trophic factors from the olfactory receptor neurons, which are often sheared off or otherwise altered in head trauma. The smell loss associated with chronic alcoholism has been found to be correlated with MRI-determined (i) increased cortical and ventricular cerebral spinal fluid volumes and (ii) reduced volumes of the thalamus and of the cortical and subcortical gray matter.

VI. *IN UTERO* LEARNING OF ODORS

It has become increasingly evident, in a wide range of species including *Homo sapiens*, that the olfactory system is functional *in utero* and that intrauterine odor experiences can influence postnatal preferences and behaviors toward odors. Thus, in rats, intrauterine injections of odorants paired with lithium chloride (an agent that induces sickness) result in postnatal avoidance of the injected odorant. The olfactory system is functional by at least 20 gestational days, as rat fetuses transferred from the abdominal cavity of their mothers into saline without interrupting the maternal blood supply exhibit responses to odors. *In utero* learning of chemicals is also believed to occur in humans. Thus, human neonates find the odorous components of human amnionic fluid attractive postnatally. More importantly, the human newborn can distinguish between the odor of the amnionic fluid of his or her mother and that of a strange mother, preferring the former. Because these preferences are equivalent in formula- and breast-fed infants, they cannot be attributed to a generalization of learning of preferences of breast odors. Thus, human neonates selectively respond to the biological fluid they encountered within the prenatal environment.

VII. CONCLUSIONS

Remarkable progress has been made in the past decade in understanding the function of the olfactory system. At the transduction level, the discovery of the gene family that controls the expression of olfactory receptors has been a monumental event. At the ontological level, it is clear that the organism can sense odors long before birth and that experiences with odors in the uterus can alter postnatal odor preferences. At the psychophysical measurement level, the development and proliferation of practical and reliable olfactory tests have spawned hundreds of studies that otherwise would not have been made, demonstrating olfactory dysfunction in a wide range of clinical disorders and leading to the discovery that olfactory loss is a very early clinical sign of several major neurodegenerative diseases. The application of such new technologies as functional imaging is beginning to unravel the mysteries of central olfactory coding and should lead, within the next decade, to a much more complete understanding of the way the brain processes olfactory information.

See Also the Following Articles

AUDITORY PERCEPTION • MULTISENSORY INTEGRATION • TASTE • VISUAL SYSTEM DEVELOPMENT AND NEURAL ACTIVITY

Acknowledgments

This article was supported, in part, by Research Grants PO1 DC 00161, RO1 DC 04278, and RO1 DC 02974 (RLD, Principal Investigator) from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health, Bethesda, MD USA. I thank Lloyd Hastings, David Kareken, Igor Kratskin, Noam Sobel, and Greg Smutzer for their comments on a previous version of the manuscript.

Suggested Reading

Bhatnagar, K. P., and Meisami, E. (1998). Vomeronasal organ in bats and primates: Extremes of structural variability and its phylogenetic implications. *Microsc. Res. Tech.* **43,** 465–475.

Brunet, L. J., Gold, G. H., and Ngai, J. (1996). General anosmia caused by a targeted disruption of the mouse olfactory cyclic nucleotide-gated cation channel. *Neuron* **17,** 681–693.

Buck, L., and Axel, R. (1991). A novel multigene family may encode odorant receptors: A molecular basis for odor recognition. *Cell* **65,** 175–187.

Calof, A. L., Rim, P. C., Askins, K. L., Mumm, J. S., Gordon, M. K., Iannuzzelli, P., and Shou, J. (1998). Factors regulating neurogenesis and programmed cell death in mouse olfactory epithelium. *Ann. N. Y. Acad. Sci.* **30,** 226–229.

Doty, R. L. (1997). Studies of human olfaction from the University of Pennsylvania Smell and Taste Center. *Chem. Senses* **22,** 565–586.

Doty, R. L., Shaman, P., Applebaum, S. L., *et al.* (1984). Smell identification ability: Changes with age. *Science* **226,** 1441–1443.

Farbman, A. I., Bucholtz, J. A., Suzuki, Y., Coines, A., and Speert, D. (1999). A molecular basis of cell death in olfactory epithelium. *J. Comp. Neurol.* **414,** 306–314.

Graves, A. B., Bowen, J. D., Rajaram, L., McCormick, W. C., McCurry, S. M., Schellenberg, G. D., and Larson, E. B. (1999). Impaired olfaction as a marker for cognitive decline. Interaction with apolipoprotein Eϵ4 status. *Neurology* **53,** 1480–1487.

Gold, G. H. (1999). Controversial issues in vertebrate olfactory transduction. *Annu. Rev. Physiol.* **61,** 857–871.

Huard, J. M. T., Youngentob, S. L., Goldstein, B. L., Luskin, M. B., and Schwob, J. E. (1998). Adult olfactory epithelium contains multipotent progenitors that give rise to neurons and nonneural cells. *J. Comp. Neurol.* **400,** 469–486.

Mombaerts, P. (1999). Molecular biology of odorant receptors in vertebrates. *Annu. Rev. Neurosci.* **22,** 487–509.

Smotherman, W. P., and Robinson, S. R. (1987). Psychobiology of fetal experience in the rat. In *Perinatal Development: A Psychobiological Perspective.* (N. A. Krasnegor, E. M. Blass, M. A. Hofer, and W. P. Smotherman, Eds.), pp. 39–60. Academic Press, Orlando, FL.

Sobel, N., Prabhakaran, V., Zhao, Z., Desmond, J. E., Glover, G. H., Sullivan, E. V., and Gabrieli, J. D. E. (2000). Time course of odorant-induced activation in the human primary olfactory cortex. *J. Neurophysiol.* **83,** 537–551.

Yousem, D. M., Maldjian, J. A., Siddiqi, F., Hummel, T., Alsop, D. C., Geckle, R. J., Bilker, W. B., and Doty, R. L. (1999). Gender effects on odor-stimulated functional magnetic resonance imaging. *Brain Res.* **818,** 480–487.

Zhao, H., Ivic, L., Otaki, J. M., Hashimoto, M., Mikoshiba, K., and Firestein, S. (1998). Functional expression of a mammalian odorant receptor. *Science* **279,** 237–242.

Opiates

CLIFFORD M. KNAPP
Boston University School of Medicine and Boston Veterans Affairs Health Care System

GLOSSARY

analgesic A medication used to reduce pain.

brainstem Region consisting of the medulla, pons, and midbrain that connects the rest of the brain to the spinal cord.

medulla oblongata Lowest portion of the brain stem, which extends from the spinal cord to the pons.

nociceptors Peripheral elements of the nervous system (commonly nerve endings) that activate pain pathways when exposed to noxious stimuli.

periaqueductal gray matter Gray matter located in the midbrain that contains nuclei that are involved in modulating the perception of pain.

The term opiates connotes drugs containing opium, derivatives from opium such as morphine, and synthetic compounds that have pharmacological actions that are similar to those of morphine. The term opioids has been taken to mean non-opium derived compounds whose actions are similar to those of morphine, and also to designate endogenous substances such as β-endorphin that act upon the same receptors as do the opiates. These receptors referred to as opioid receptors have been cloned relatively recently, and their structure has become well-characterized. At present, by definition, any compound that interacts with these receptors is classified as being an opioid. Opioid agonists activate opioid receptors, producing effects such as analgesia and sedation that are produced by drug like morphine. Opioid antagonists block agonists from interacting with these receptor sites.

I. HISTORY

The Sumerians and Egyptians may have been familiar with the effects of opium. Theophrastus made one of the first clear references to opium in the third century BC. Opium is derived from the Greek word, opion, for poppy juice. Opium is a complex mixture of substances obtained from the juice of the poppy plant, *Papaver somniferum*. Opium is prepared by drying the milky juice of the poppy. The dried brown material is then powdered.

In 1806, the German chemist Fredrich Saturner isolated the pure alkaloid substance morphine from opium. Morphine was one of the first pure drugs to be isolated from a natural source. Codeine, whose chemical structure is very similar to that of morphine, was found in opium by Robiquet in 1832. The composition of opium consists of approximately 20 alkaloids. The concentration of morphine in opium is about 10%. Opium contains less than 0.5% codeine.

In 1874, the pharmacist C. R. Alder Wright produced heroin (diacetylmorphine) by boiling morphine with acetic anhydride. Heinrich Dreser at Bayer Laboratories in Germany developed a method for the synthesis of large quantities of heroin. This drug was

marketed in 1898 as a cough suppressant and for the treatment of respiratory disorders. At first it was believed that heroin did not have the addictive properties that were associated with morphine.

In the nineteenth century opium was used as an analgesic agent and a sedative and used to treat diarrhea produced by diseases such as cholera and dysentery. Opium was consumed in a variety of preparations. Laudanum is an example of such a preparation. It was prepared by dissolving opium in an alcoholic vehicle such as wine or port. Other ingredients were added to this mixture including cinnamon, cloves, or saffron. Paregoric is another opium-containing preparation that is still used for the treatment of diarrhea. It is a camphorated tincture of opium. Dover's powder consisted of powdered ipecac, a powerful emetic, and powdered opium. It was used as a diaphoretic (sweat-inducing) medication.

Opium was also a common ingredient in a large number of patent medications whose use became increasingly more widespread as society became industrialized. These products included "soothing syrups" that were used to sedate infants, a practice that may have led to the impaired development of many children. With no regulations controlling their sale, opium-containing preparations were consumed on a large scale by adults during the nineteenth century in both the United States and England.

With the introduction of the syringe, morphine could be administered by injection. Individuals with the means to purchase a syringe, i.e., members of the middle and upper classes, began to develop the habit of injecting themselves with morphine. This led to the appearance of an addictive disorder that was called morphinism. During the American Civil War, morphine was used as an analgesic agent and morphine dependence became a problem for many veterans of that conflict.

In 1878, Edward Levinstein wrote a study describing the character of morphine dependence. Toward the end of the nineteenth century and the beginning of the twentieth, the concept of opiate addiction as a specific disorder that needed to be studied and treated by the medical community began to be further developed and elaborated by T. D. Crothers, J. B. Mattison, and others. As the problems associated with addiction to opiates became more apparent, both public and governmental concern about the use of opiates grew. Governments became increasingly involved in attempts to control the problems associated with opiate use for nonmedical purposes. In the United States, the Harrison Narcotics Act was passed in 1914 as one of the first attempts at the federal level to regulate trade in opiates.

II. CHEMICAL CLASSIFICATION OF OPIATES

Morphine is a phenanthrene derivative (see Fig. 1). Many synthetic opiates are structural analogs of morphine. These include heroin, hydromorphone, hydrocodone and oxycodone (see Table I). Codeine, i.e., methylmorphine, although found in nature, is produced commercially by adding a methyl (CH_3) group to morphine. Some analogs of morphine are opioid receptor antagonists (see Table II). That is, they can block and reverse the effects of opiates. Naloxone is an example of an opioid antagonist that is synthesized by replacing the methyl group (CH_3) on the nitrogen (N) found in the structure of morphine with an allyl group ($CH_2CH=CH_2$).

Synthetic opiates also have been developed through the modification of the basic phenanthrene structure of morphine. Pentazocine is an example of such a compound, having a benzomorphan structure. Methadone and meperidine are examples of opiates whose structures are dissimilar from that of morphine (Fig. 2). Meperidine is a phenylpiperidine compound. Many synthetic opiates are piperidines or phenylpiperidines. Methadone is a phenylheptylamine agent, as is propoxyphene. Although methadone and propoxyphene have similar structures, methadone has much more efficacy as an analgesic agent than does propoxyphene, demonstrating that even modest alterations in the structure of an opiate can produce marked differences in the activity of opiate compounds.

III. OPIOID RECEPTORS

Opioid receptors are proteins that extend onto the surface of cell membranes and that act as communication links between the external environment and the cell by interacting with either drugs or neurotransmitters. These receptors are heptahelical structures, i.e., they loop between the inner and outer surfaces of the cell membrane seven times. Opioid receptors are located throughout the brain and spinal cord. They also are found on white blood cells and in the gastrointestinal tract.

The analgesic action of systemically administered opiates results from interaction of these agents with receptors located both in the spinal cord and in the brain. Pain arising from peripheral areas of the body is

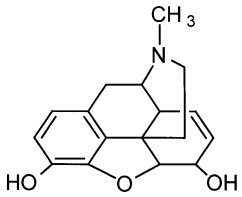


Figure 1 The chemical structure of morphine.

Table II
Opioid Receptor Antagonists

Generic name	Trade name
Nalmefene	Revex
Naloxone	Narcan
Naltrexone	ReVia

produced by the stimulation of nociceptors that arise from dorsal root neurons located in the spinal cord. Pain signals are transmitted via either myelinated axons of the A group or unmyelinated C fibers. Activation of C fibers results in release of the sensory neurotransmitter substance P in the spinal cord. Infusion of morphine and other opiates into the dorsal horn can block the transmission of pain stimuli. This may, in part, be due to morphine-induced inhibition of substance P release. In the brain, morphine and other opiate analgesics block pain by interacting with receptor sites located in the periaqueductal gray and other areas. Opiates, when administered into the brain, may activate inhibitory descending pain pathways that project down into the spinal cord.

Table I
Chemical Classification of Opiates

Chemical class	Generic name (Trade name)
Phenanthrenes	Buprenorphine (Buprenex)
	Codeine
	Hydrocodone (Hycodan, in Vicodin)
	Hydromorphone (Dilaudid)
	Morphine
	Nalbuphine (Nubain)
	Oxycodone (Oxycontin)
Benzmorphans	Dezocine (Dalgan)
	Pentazocine (Talwin)
Morphinans	Butorphanol (Stadol)
	Levorphanol (Levo-Dromoran)
Phenylpiperidines and piperidines	Alfentanil (Alfenta)
	Diphenoxylate (in lomotil)
	Fentanyl (Sublimaze)
	Loperamide (Imodium)
	Meperidine (Demerol)
	Remifentanil (Ultiva)
	Sufentanil (Sufenta)
Phenylheptylamines	Levomethadyl (Orlaam)
	Methadone (Dolophine)
	Propoxyphene (Darvon, in Darvocet)

Opiates have been show to interact with three types of receptors: the μ (mu), δ (delta), and κ (kappa) receptors. Endorphins, enkephalins, and dynorphins, peptide substances that are naturally produced in the body, can also activate these receptors. A fourth type of opioid receptor has been identified, the orphan receptor. The orphan receptor does not appear to play a role in the analgesic actions of opiates. An endogenous substance called nociceptin–orphanin FQ activates this receptor.

The effects of a particular opioid are determined by the nature of its interaction with the opioid receptors, that is, what types of receptors it acts on and to what extent it activates the specific receptor types with which it interacts. Many opiates, including morphine, meperidine, codeine, and oxycodone, exert their analgesic effects primarily by stimulating μ receptors. Opiates differ with respect to the maximal analgesic action (efficacy) that can be produced by the administration

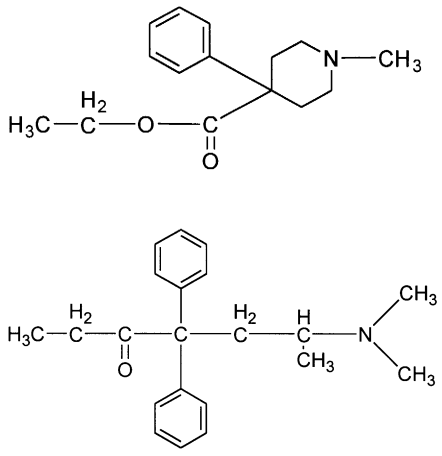


Figure 2 The chemical structures of meperidine (top) and methadone (bottom).

Table III
Maximal Efficacy of Opiate Analgesics

High efficacy (Severe pain)	Moderate efficacy (Moderate pain)	Low efficacy (Mild pain)
Morphine	Hydrocodone	Codeine
Fentanyl	Oxycodone	Propoxyphene
Hydromorphone	Pentazocine	
Levorphanol	Tramadol	
Meperidine		
Methadone		
Butorphanol		
Nalbuphine		
Buprenorphine		

of high doses of them (Table III). Morphine can almost fully activate these receptors and, consequently, in sufficient doses can be used to relieve severe pain. The maximal efficacy of codeine is much lower than is that of morphine. Thus, the administration of even high doses of codeine can not reduce severe pain to tolerable levels.

Opiates that cannot produce maximal activation of opiate receptors are referred to as partial receptor agonists. Buprenorphine is a well-characterized partial, μ-receptor agonist for which there seems to be a limit (ceiling) to the degree to which it can produce adverse effects, such as respiratory depression. This can be contrasted with morphine, which at high enough doses depresses respiration severely enough to cause death.

Antagonists such as naloxone do not activate opioid receptors, but instead block opioid agonists from stimulating receptors. Nalbuphine, butorphanol, dezocine, and pentazocine are examples of mixed agonist–antagonists. These agents are either μ-receptor antagonists or partial agonists but also act as agonists at the κ receptor. Because mixed agonist–antagonist opiates have a pattern of interaction with opioid receptors different from that of μ-selective agonists such as morphine, they produce a distinct range of effects.

IV. THERAPEUTIC USES

Opiates are administered to relieve pain that ranges in intensity from mild to severe. These agents can often reduce pain at doses that do not markedly diminish the perceptions of other kinds of sensations. Pain that results from the activation of pain pathways by noxious stimuli is usually alleviated by the administration of opiates. Neuropathic pain, i.e., pain resulting from nerve damage or irritation, frequently is not readily decreased by opiates unless they are administered in high doses.

Experimental evidence suggests that morphine and other opiates can elevate the threshold level for the amount of noxious stimulation needed for pain to be experienced. There is usually a large emotional component that contributes to the distress and discomfort produced by pain. The experience of pain may involve reactions of panic, anxiety, fear, and depression. The administration of opiates can produce calming and sedative effects that make pain more tolerable. Thus, whereas opiates may not eliminate all sensation of pain patients may find that it is no longer bothersome. After the administration of high doses of opiates, patients may exhibit what appears to be indifference to pain.

Codeine and hydrocodone are used to decrease coughing. These drugs and most of the other opiates will act in the medulla to diminish the effects of stimuli that activate the cough reflex. Reduction of coughing can provide patients with respite from the physical demands of coughing and from its disruptive effects on rest and sleep. Coughing, however, can play a role in the elimination of secretions, so that problems such as airway obstruction may occur if coughing is suppressed in certain situations.

Opium has been used to treat diarrhea for many centuries. Drugs that are currently used for this purpose are listed in Table IV. The administration of opiates decreases propulsive peristaltic waves in the colon and can increase colon tone. These actions reduce the rate of transit of material through the colon, which leads to enhanced resorption of water from fecal matter. The drier fecal matter is retained longer in the colon.

Opiates are used for the induction and maintenance of surgical anesthesia and to reduce the discomfort produced by medical procedures such as endoscopy. The use of these agents for these purposes has increased with the introduction of opiates such as fentanyl and related compounds that enter the brain rapidly and are cleared from the blood rapidly. During induction of anesthesia, opiates are often administered in combination with sedative agents such as midazolam (Versed) and propofol (Diprivan). Coadministration of opiates reduces the amount of anesthetic gas that needs to be administered to maintain anesthesia. When administered during surgery, opiates can reduce

Table IV
Antidiarrheal Opioid Medications

Generic name	Trade name
Diphenoxylate	In Lomotil with atropine
Loperamide	Imodium
Opium	In Paregoric

Table V
Major Adverse Effects Associated with Opiate Administration

Respiratory depression
Nausea and vomiting
Constipation
Pupil constriction
Hallucinations and bizarre thoughts (mixed antagonist–agonists)
Sedation and impaired concentration
Mood alteration
Itching

increases in blood pressure and heart rate that may occur in response to surgical procedures.

As will be discussed in greater detail later, some opiates are used in the treatment of opioid dependence. Methadone and L-α-acetylmethadol (levomethadyl, LAAM) are currently utilized as substitutes for drugs like heroin. Methadone is also used to facilitate detoxification from opiates. Buprenorphine is being investigated as an agent that can substitute for street opiates and that can be used to decrease the severity of symptoms associated with heroin withdrawal that occur during detoxification.

V. ADVERSE EFFECTS

The administration of opiates may produce a variety of toxic effects (Table V). The occurrence of these effects places a limit on the usefulness of opiate agents in the treatment of pain. Intensive efforts have been put into the development of synthetic opiate analgesics that can provide relief from pain while minimizing adverse effects. Morphine, however, still remains one of the best medications available for the treatment of severe pain.

Morphine and other opiates act on respiratory control centers in the brain stem to inhibit respiration. Deaths that result from opiate overdose are often the result of respiratory depression. The administration of higher doses of morphine depresses both respiratory rate and the volume of air exchanged during respiration. Opiates appear to decrease the sensitivity of respiratory centers to carbon dioxide, a compound that, under normal conditions acts as a respiratory stimulant.

Opiate-induced depression of respiration can be markedly enhanced by the concurrent administration of central nervous system depressants, including sedatives, anesthetic agents, or alcohol. Patients with respiratory disorders such as emphysema are particularly sensitive to the respiratory depressant effects of opiates, so that caution needs to be exercised when administering these agents to such patients. Similar caution needs to be exercised when opiates are administered to head injury patients.

Many patients experience nausea and vomiting following the administration of opiates. These effects are more likely to occur in patients when they are ambulatory, with up to 40% of these patients experiencing nausea and 15% vomiting. Morphine and other opiates may produce nausea and vomiting by acting on the chemoreceptor trigger zone located in the medulla. They also may act on elements of the vestibular system, the system that regulates balance.

Opiates act directly on receptors located in the gastrointestinal tract. These actions produce a reduction in the rate of transit of fecal material through the colon. These effects may play a role in the production of opiate-induced constipation. Opiates may blunt the awareness of stimuli that promote defecation. This may be another factor that leads to opiate-related constipation. Only a limited amount of tolerance develops to the constipating effects of opiates. Patients who must receive opiates for prolonged periods of time often experience severe constipation. This problem can frequently be corrected by increasing the amount of bulk fiber that patients consume in their diet and by administering stool softeners and laxatives. Mixed agonist–antagonist agents such as pentazocine and nalbuphine are less likely to produce constipation than are other opiates.

The administration of opiates can produce constriction of the smooth muscle of the gallbladder. Administration of morphine may cause the sphincter of Oddi in the bile duct to constrict and can greatly increase pressure within this duct. Pressure within the gallbladder also may be increased by morphine treatment. These effects of opiates on the gallbladder may cause some patients to experience epigastric discomfort or, in some cases, even the severe pain associated with biliary colic.

Opiate administration can increase the tone of the urinary bladder and the ureters. Treatment with morphine can promote the retention of urine by inhibiting the urinary voiding reflex and by increasing the tone of the bladder's external sphincter. Opiates can have an antidiuretic effect that can lead to reduced urine flow.

Pupil constriction, miosis, is frequently observed after the administration of opiates. This effect results from the activation of parasympathetic nerves. Tolerance to the pupil-constricting effect of opioids does not readily develop. Because of the persistence of opiate-induced pupil constriction, the presence of miosis can be used to establish a diagnosis of opiate-induced toxicity even in individuals who have been using opiates on a chronic basis.

Sedation is often produced by the administration of opiates. Whereas this effect is sometimes of value in helping patients to rest, it can become bothersome particularly in patients who need to receive opiates over prolonged periods of time. Opiate treatment also often results in difficulty in concentrating and other forms of cognitive impairment. Tolerance to the sedative effects of opiates often may develop rapidly.

Opiates can alter mood in a number of ways. Individuals who have abused opiates may experience euphoria and a sense of calm after the administration of heroin, morphine, or related drugs. In contrast, individuals without a history of drug abuse may feel dysphoric when being treated with opiates. For some individuals, opiates may relieve symptoms of depression and decrease feelings of anger and anxiety. It has been suggested that such individuals may seek out opiates to provide relief from symptoms of psychological distress.

Several unpleasant psychological effects, including hallucinations, bizarre thoughts, dysphoria, and anxiety, may occur in patients being treated with the mixed agonist–antagonist pentazocine. Similar effects may be produced by the administration of other mixed agonist–antagonists such as nalbuphine or butarphanol. However, hallucinations and bizarre thoughts occur with greater frequency with pentazocine administration than during treatment with either nalbuphine or butorphanol.

The acute administration of opiates can disrupt hormonal regulation of physiological activity. Endocrine effects of opiates include the inhibition of the release of gonadotropin-releasing hormone (GRH). This hormone facilitates the release of both luteinizing hormone (LH) and follicle-stimulating hormone (FSH), both of which play important roles in reproductive processes. Opiate administration can disrupt the menstrual cycle of women and cause a reduction in testosterone levels in men. Opiates can also inhibit the release of corticotrophin-releasing factor (CRF). This factor promotes the release of adrenocortropic hormone (ACTH), which enhances the release of cortisol, a hormone that helps the body in reacting to stress. Concentrations of the hormone prolactin are increased in the blood by the administration of morphine and heroin. Tolerance develops to the disruptive effects of opiates on hormone release.

Opiates can produce peripheral vasodilation and will decrease peripheral resistance. These effects may cause patients who are being treated with opiates to sometimes experience a sudden drop in blood pressure when they sit or stand up rapidly. Opiate-induced reductions in blood pressure can be a source of risk for patients who have reduced blood volumes produced by internal hemorrhage or other factors.

Administration of the mixed agonist–antagonist agent pentazocine, in contrast to that of μ-receptor agonists like morphine, can result in an elevation in blood pressure and an increase in heart rate. These effects are not produced by the administration of therapeutic doses of the mixed agonist–antagonist nalbuphine.

Morphine administration may lead to the release of histamine which can produce itching and contributes to the morphine-induced vasodilation, allergic reactions such as skin rashes may be produced by the administration of opiates. Such allergic responses to opiates, however, occur only rarely.

The toxic effects of an acute overdose of opiates can include impaired consciousness, which can grade into coma, pupil constriction, and respiratory depression, sometimes occurring in conjunction with fluid accumulation in the lungs. Seizure may occur after high doses of opiates have been administered. Pupils will become dilated if respiratory depression produces marked decreases in blood oxygen levels. Patients who are experiencing symptoms of opiate overdose need to have adequate ventilation reestablished. Administration of the opioid antagonist naloxone will reverse the toxic effects of most opiates. Reversal may be difficult to achieve in the case of certain agents, such as buprenorphine.

VI. OPIATE DISPOSITION AND CLEARANCE

A. Absorption

Many of the opiates are well-absorbed when administered orally. Morphine, when absorbed from the

gastrointestinal tract, is extensively metabolized in the liver. This "first pass" metabolism involves a chemical reaction in which glucuronide is added to morphine to form morphine glucuronide. Oral doses of morphine therefore must be larger than the doses used when morphine is injected to produce analgesia. Chemical modification of the morphine molecule can provide protection from the first pass effect. Morphine analogs such as codeine and oxycodone, for example, are not extensively biotransformed in the liver as they pass from the gastrointestinal tract into the general circulation.

Fentanyl will readily pass through the skin, thus allowing it to be administered transdermally. Fentanyl has been added to patches that can be attached to the skin. These transdermal fentanyl patches can be used to produce analgesia over prolonged periods of time (up to 72 hr).

B. Distribution

In most cases, opiates need to reach sites in the brain, such as the periaqueductal gray, to produce analgesia. The brain is protected by the blood–brain barrier, which acts to hinder many compounds from crossing into the cerebral spinal fluid (CSF) that bathes the brain. Opiates such as codeine, hydromorphone, and heroin readily cross the blood–brain barrier to enter into the CSF. The entry of morphine into the brain is, in contrast, somewhat delayed, and the onset of the central effects of this agent is slower than the onset of action of compounds such as heroin. Another opiate, loperamide, is effectively blocked by the blood–brain barrier. This drug, when administered in the treatment of diarrhoea, does not produce the sedation and euphoria that may result from the administration of morphine and many of the other opiates.

C. Metabolism

The biotransformation of exogenous chemicals often results in a decrease in the activity of the parent compound and frequently leads to the production of compounds that are readily excreted into the urine. Morphine combines with glucuronic acid in the liver to form glucuronides that pass into the urine more readily than does morphine itself. Morphine is primarily converted to morphine 3-glucuronide, but morphine 6-glucuronide is also formed. Interestingly, morphine 6-glucuronide acts similarly to morphine. In patients who are receiving morphine on a chronic basis, the 6-glucuronide metabolite may play a significant role in the production of analgesia. This metabolite is eliminated into the urine, and decreased kidney function can lead to a reduction in the rate at which morphine 6-glucuronide is eliminated from the body.

Normeperidine is produced by the biotransformation of meperidine. This product can accumulate in the body if meperidine is repeatedly administered to patients who have severely compromised kidney function. These patients may experience neurotoxic effects as normeperidine accumulates in the body. These effects can include the development of seizures, tremors, twitching, and mental confusion.

Renal function is often decreased in elderly individuals. Caution must be exercised when agents like morphine or meperidine, drugs that form toxic metabolites that are eliminated via the kidney, are administered to older individuals.

Heroin is rapidly hydrolyzed into 6-monoacetylmorphine, which then undergoes conversion to morphine. Heroin itself may have little activity at opioid receptors. In contrast, 6-monoacetylmorphine readily stimulates these receptors. The euphoria and analgesia produced by heroin administration are due in large part to the actions of its metabolites, morphine and 6-monoacetylmorphine. Similarly, codeine does not readily stimulate opioid receptors, and much of its analgesic actions may be produced by morphine that is formed by the metabolism of codeine.

Codeine, oxycodone, and hydromorphone are metabolized in the liver by the enzyme cytochrome P450. This enzyme exists in many forms, and it is the CYP2D6 form that has been implicated in the biotransformation of these compounds. Some individuals may have mutant forms of CYP2D6 that do not act as efficiently as enzymes as they do in most of the population. Individuals with these forms of CYP2D6 are poor metabolizers of codeine and related drugs. Poor metabolizers of codeine may include between 4 and 10% of populations with European ancestry.

L-α-Acetylmethadol (LAAM) is structurally related to methadone. This drug is used like methadone to reduce the use of heroin and other abused opiates by reducing opiate-related withdrawal symptoms and craving for opiates. The metabolites of LAAM, noracetylmethadol and dinoracetylmethadol, have activity that is similar to that of methadone. These metabolites remain in the plasma for a relatively long period of time. Consequently, LAAM administration

may be effective in preventing opiate withdrawal from developing in heroin addicts for 3–4 days after the administration of a single dose.

VII. TOLERANCE AND PHYSICAL DEPENDENCE

Many of the effects of morphine and the other opiates diminish in magnitude if they are repeatedly administered. This reduction in opiate efficacy is a form of drug tolerance. Opiate tolerance entails both reduced duration and reduced intensity of opiate-induced effects. Profound tolerance can develop to the analgesic and sedative effects of opiates. Marked tolerance can also develop to opiate-induced respiratory depression, nausea, and vomiting. In contrast, tolerance to opiate-induced miosis and constipation tends to be limited in extent.

The development of tolerance to the analgesic effects of opiates can be a problem for patients with conditions, such as cancer, that must be treated with these agents over long spans of time. Very large doses of opiates may have to be administered to produce analgesia in these patients. Patients with malignant disorders, however, may require increased doses of opiates to control pain because of factors other than tolerance. These factors include tumor growth, impingement of tumor onto nerves, and the development of inflammation at the tumor site.

Patients who are treated chronically with opiates and addicts who continually self-administer opiates experience symptoms of withdrawal either when opiate administration is discontinued or when an opioid antagonist such as naloxone is administered. Symptoms of withdrawal may also appear when mixed agonist–antagonist agents such as pentazocine are administered to individuals who have become physically dependent on opiates. Withdrawal symptoms in individuals who have been receiving μ-receptor agonists such as morphine and heroin include a runny nose, tearing, abdominal cramps, muscle pain, gooseflesh, nausea, anxiety, difficulty sleeping, yawning, sweating, restlessness, and dysphoric feelings. This complex of symptoms bears a resemblance to an influenza-like illness.

Addicts often experience intense cravings for opiates during withdrawal that can last for extended periods. In contrast, most patients who have been receiving opiates for therapeutic purposes, particularly those without a prior history of substance abuse, do not experience a compulsive need for these drugs. Some patients display what might be interpreted as addict-like behavior if they experience pain because inadequate doses of opiates are being administered. These patients may become angry and will make frequent requests for pain medications. Such requests are expressions of genuine need for pain relief and should not be confused with the compulsive drug-seeking behavior that is associated with addictive disorders.

The administration of an opioid antagonist such as naloxone can precipitate an intense episode of withdrawal in dependent individuals within a matter of minutes. The time at which withdrawal symptoms occur after the discontinuation of opiate administration is directly related to the rate of clearance of opiate from the body. Withdrawal may occur in 8–12 hr after the termination of rapidly cleared drugs, such as heroin, hydromorphone, or morphine. Symptoms of withdrawal from methadone, which is slowly eliminated from the body, may not appear until 36–48 hr after its discontinuation.

VIII. OPIATE ADDICTION

A. Opiate Dependence and Abuse

Addiction to opiates involves the compulsive use of opiates for nonmedical purposes that may stem from the motivation to experience the pleasant sensations produced by opiates and to prevent the onset of withdrawal symptoms. In an effort to more precisely characterize opiate addiction, the diagnostic categories of opioid abuse and dependence have been developed. Abuse entails recurrent drug use, but the use is not frequent enough to result in marked physical dependence. The opioid-dependent individual compulsively uses opiates and does so frequently enough to become physically dependent on them. Dependent individuals will persist in self-administering opiates even when such behavior results in significant physical, social, psychological, and/or financial harm.

Opiate addiction frequently involves the use of heroin. When administered intravenously, this drug may produce a short-lived, pleasurable sensation known as a "rush" that may occur even in individuals who have developed marked tolerance to the effects of opiates. As an alternative to intravenous administration, the practice of smoking heroin has become more common. In addition to heroin, commercially marketed opiates such as hydromorphone (Dilaudid) and oxycodone-containing products including Percocet,

Percodan, and Oxycontin may be self-administered by individuals with opiate use problems.

The reasons for the development of opioid dependence are, despite extensive investigation, in many respects still poorly understood. The social environment can play a critical role in who will become and remain dependent. This is illustrated by the finding that, whereas heroin was used extensively by American soldiers who fought in the Vietnam Conflict, the incidence of heroin use by these individuals was markedly reduced when they returned to their lives at home in the United States. Patterns of opiate use appear to vary over the course of history. As mentioned earlier, opium and morphine addiction became prevalent during the nineteenth century. Following a wane in opiate addiction, heroin dependence increased in the 1960s and 1970s and began to increase again in the 1990s. Fluctuations in opiate use over time serve as additional support for the view that social factors are involved in regulating the incidence of opiate addiction in a society.

The pleasurable effects that opiate users experience during the self-administration of heroin, morphine, and related agents may play an important role in the development and maintenance of opiate addiction. These pleasurable effects may result from the interaction of opiates with areas of the brain that are responsible for the seeking of rewarding stimuli in the environment. These areas include the ventral tegmental area and the nucleus accumbens of the brain. These structures are part of the mesolimbic system. Dopamine is a primary neurotransmitter within this system that has been linked to the rewarding actions of drugs. Morphine and other μ-receptor agonists act within the ventral tegmental area to enhance the release of dopamine, particularly in the nucleus accumbens. In contrast, compounds that activate κ receptors do not promote dopamine release, which may explain why the administration of κ agonists such as pentazocine produces dysphoria more often than euphoria.

During prolonged abstinence, former users of heroin and other opiates may become depressed and feel heightened distress in stressful situations. They also may feel anxious and experience other forms of psychological discomfort. It has been hypothesized that these feelings of discomfort, which may last for months after opiate withdrawal, may contribute to relapse to drug use. Whereas biological factors may play a role in the persistence of this discomfort, at present the identity of these factors remains to be discovered.

B. Treatment of Opiate Withdrawal and Opiate Dependence

Patients who are not psychologically dependent on opiates can usually be withdrawn from these agents by gradually tapering the daily dose of the drug. Withdrawal can be assisted, if necessary, by the administration of α_2-noradrenergic agents such as clonidine and lofexidine. These drugs can attenuate many symptoms of withdrawal, including nausea, vomiting, sweating, cramps, and diarrhea. Clonidine and lofexidine will not, however, reduce drug craving.

The opiate buprenorphine has been used to facilitate opioid withdrawal. The administration of buprenorphine to opiate-dependent individuals may reduce withdrawal symptoms, including drug craving. Unlike α_2-noradrenergic receptor agents such as clonidine, buprenorphine administration does not produce a marked decrease in blood pressure. Patients will experience mild symptoms of withdrawal after buprenorphine administration has been discontinued.

Many individuals with a history of opioid dependence are unable to remain abstinent following initial periods of withdrawal. Opioid substitution therapy has been devised as an approach to integrate these individuals back into the mainstream of society. Methadone is the drug that has been used most frequently for this purpose. LAAM has also been used in substitution therapy. Because they are slowly cleared from the body, methadone and LAAM allow opiate-dependent individuals to remain free of withdrawal symptoms throughout the day. This allows these individuals to work and engage in other forms of normal daily activities and frees them from the compulsive pursuit of drugs and the disruptive effects of the rapid changes in mood that are associated with heroin use.

IX. MOLECULAR BIOLOGY OF OPIOID SYSTEMS

Genes for the μ-, δ-, and κ-opioid receptors have all been cloned. *MOR-1* is a gene that encodes for the μ receptor. Morphine will not produce analgesia in mice that are missing the *MOR-1* gene. *DOR-1* is a δ-receptor gene and *KOR-1* is a κ-receptor gene. There is evidence that variants (subtypes) of the μ-, δ-, and κ-opioid receptors may exist, but the existence of distinct genes that encode for each of these receptor subtypes has yet to be demonstrated. The possibility also exists, however, that different opioid receptor subtypes could result from processes such as alternative pathways of

RNA splicing that might occur after the messenger RNA is formed by transcription of opioid receptor genes.

Opioid genes contain information for the synthesis of opioid receptors in the cell. These receptors are composed of amino acids that are linked together to form the receptor protein. The μ-, δ-, and κ-opioid receptors share between 50 and 70% of the same amino acid structure.

The activation of opioid receptors by endogenous peptides may serve a variety of important functions, including helping animals to survive the occurrence of traumatic stress and injury. Specific examples of endogenous opioid peptides are β-endorphin, methionine- and leucine-enkephalin, and dynorphins A and B. These peptides are derived from the breakdown of larger precursor molecules. Prepro-opiomelanocortin (POMC) is the precursor of β-endorphin and also several other bioactive substances, including ACTH, melanocyte-stimulating hormone, met-enkephalin, and β-lipotropin. Met- and leu-enkephalin are derived from preproenkephalin, whereas dynorphins A and B are produced from preprodynorphin.

Opioid receptors belong to the G-protein superfamily, meaning that they are coupled to G proteins. These proteins act to link drug-induced changes in receptors to the internal function of the cell. Opioid receptors are coupled to G_1 and G_0 proteins. Opioid receptors interact with these G proteins to inhibit adenylate cyclase, an enzyme that promotes the formation of the intracellular messenger cyclic adenosine 3′,5′- monophosphate (cyclic AMP). G proteins also may mediate the inhibitory effects of opioids on voltage-gated calcium channels and their activating effects on inwardly rectifying potassium channels. Inhibition of calcium channels may lead to a decrease in the release of neurotransmitters such as substance P, which mediates pain signals. Opioids may have a host of other effects on cells, including the activation of phospholipase A_2 (PLA_2) and protein kinase C, enzymes that regulate a number of cellular activities. Opioid-induced activation of PLA_2 in the periaqueductal gray region increases voltage-sensitive potassium conduction. This effect produces inhibition of the activity of cells in this region that can release the inhibitory neurotransmitter γ-aminobutyric acid (GABA). This action, in turn, may produce disinhibition (i.e., activation) of inhibitory descending pathways that block pain signals that ascend through the spinal cord.

Chronic exposure to opiates leads to the development of both tolerance to and dependence on these agents. The exact nature of the changes within neurons that are responsible for the occurrence of opiate tolerance and dependence remains to be fully characterized. Tolerance to the analgesic effects of opiates can be reduced by the administration of agents that block the effects of the excitatory neurotransmitter glutamate at the *N*-methyl-D-aspartate (NMDA) receptor. This suggests that these receptors play some important role in the development of opioid tolerance. NMDA receptors also may be involved in the development of sensitization to the effects of opiates. Opiate-induced sensitization may occur after the repeated administration of drugs like morphine and is characterized by enhanced sensitivity to the motor-activity-enhancing effects of opiates.

The effects of opioid exposure on neurons of the brain structure called the locus ceruleus have been examined in an effort to discover the mechanisms that are involved in the appearance of opiate withdrawal. Compounds that are μ agonists will act to inhibit the activity of neurons in the locus ceruleus. This structure contains many of the brain's norepinephrine neurons, which, when activated, may play a role in the production of anxiety. Opiates may decrease anxiety by inhibiting the firing of cells located in the locus ceruleus. This may occur because opiates decrease the excitability of these cells by opening potassium channels in the cell membrane and by blocking sodium channels. After acute administration, opiates will decrease levels of cyclic AMP. Chronic exposure of locus ceruleus cells to opiates results in an increase in the excitability of cells. This increase in excitability becomes evident following the withdrawal of opioids after periods of prolonged treatment with these agents. Opioid-induced increases in cellular excitability may be produced by increases in cyclic AMP levels, which leads to enhanced activity of sodium channels. Chronic opioid administration may up-regulate cyclic AMP systems by increasing certain forms of adenylate cyclase, the enzyme responsible for cyclic AMP formation. CREB (cyclic AMP response-element-binding protein) is a cyclic-AMP-regulated transcription factor that may play a key role in the opioid-induced increases in adenylate cyclase activity. Opiate withdrawal appears to be less severe in mice that are CREB-deficient.

Following agonist activation of opioid receptors, G-protein-coupled receptor kinases (GRKs) may add a phosphate group to these receptors and facilitate their interaction with β-arrestin 2. When bound to opioid receptors, β-arrestin may inactivate the receptor. It has been hypothesized that β-arrestins may play a role in

the desensitization of opioid receptors that occurs when they are exposed to agonists for prolonged periods of time. It has been shown that the analgesic effects of morphine will persist for much longer than usual in animals that do not have the gene for β-arrestin 2. This finding supports the idea that β-arrestins may play a role in the development of reduced sensitivity to opiates. There is evidence that this form of desensitization is involved in the development of tolerance to the effects of opiates that is associated with the chronic administration of these agents.

X. SUMMARY

Opiates remain the most efficacious agents available for the treatment of severe pain. Unfortunately, they produce a range of adverse effects, including sedation, nausea, and respiratory depression, that make these medications difficult for some patients to tolerate. The development of opiate tolerance and dependence is another concern associated with the use of opiates for therapeutic purposes, particularly when they are needed for the treatment of chronic pain. Problems associated with the abuse of heroin and related drugs have become an increasing concern as infection with the human immunodeficiency virus (HIV) has become widespread in individuals who self-administer drugs by injection or who engage in other high risk behaviors associated with drug cultures.

The use of molecular biological techniques has begun to greatly expand our understanding of how opioid systems function in the brain and other areas of the body. The opioid receptors have been cloned and their molecular structures are now known. Work has begun on how opiates can act through these receptors to influence activity in the cell via intermediary messengers such as cyclic AMP. As knowledge concerning the molecular basis of the interaction between opiates and cell function increases, solutions to the problems currently associated with opiate use and abuse may begin to emerge.

See Also the Following Articles

BEHAVIORAL PHARMACOLOGY • NEUROPHARMACOLOGY • NOCICEPTORS • PAIN • PSYCHOACTIVE DRUGS

Suggested Reading

Berridge, V., and Edwards, G. (1987). *Opium and the People: Opiate Use in Nineteenth-Century England*. Yale University Press, New Haven, CT.

Bohn, L. M., Lefkowitz, R. J., Gainetdinov, R. R., Peppel, K., Caron, M. G., and Lin, F.-T. (1999). Enhanced morphine analgesia in mice lacking β-arrestin 2. *Science* **286,** 2495–2498.

Booth, M. (1996). *Opium A History*. St. Martin's Press, New York.

Hardman, J. G., Limbird, L. E., and Gilman, A.G. (Eds.). (2001). *Goodman & Gilman's The Pharmacological Basis of Therapeutics*. 10th ed. McGraw-Hill, New York.

Lowinson, J. H., Ruiz, P., Millman, R. B., and Langrod, J. G. (Eds.). (1997). *Substance Abuse: A Comprehensive Textbook*. Williams and Wilkins, Baltimore, MD.

Nestler, E. J., and Aghajanian, G. K. (1997). Molecular and cellular basis of addiction. *Science* **278,** 58–63.

Oxytocin

HANS H. ZINGG
McGill University

GLOSSARY

action potential Short-lasting depolarization of the neuronal cell membrane that is propagated along the neuronal processes.

exocytosis The process by which a secretory granule fuses with the cell membrane and releases its content into the extracellular space.

hypothalamo–neurohypophysial system A neuronal system that consists of neurons that are present in two specific nuclei of the hypothalamus (supraoptic and paraventricular nuclei) and that extend their axons to the neurohypophysis.

magnocellular neuron Specific neuron type present in the supraoptic and paraventricular nuclei of the hypothalamus. These neurons are capable of producing and secreting large amounts of secretory peptides (mostly vasopressin and oxytocin) and releasing them into the general circulation. They represent a classical example of neuroendocrine cells. To be distinguished from the smaller parvicellular neurons, present in the paraventricular nucleus which do not project to the neurohypophysis.

neurohypophysis (synonyms: neural lobe of the pituitary, posterior pituitary) Site of release of oxytocin and vasopressin into the general circulation. It consists of the posterior part of the pituitary gland and contains the nerve endings of magnocellular hypothalamic OT- and AVP-producing neurons.

neurosecretion Release of a soluble messenger from a neuron mostly, but not exclusively, from axon terminals.

secretory granule Intracellular particle that contains secretory peptides and is surrounded by a bilayer membrane. Secretory granules are produced at the level of the cell body, involving the so-called Golgi apparatus. To be distinguished from the smaller *secretory vesicles*, which contain small neurotransmitter molecules that are enzymatically synthesized at the level of the nerve terminal.

Oxytocin (OT) is a nonapeptide hormone that is produced in specific hypothalamic neurons and is secreted into the general circulation from the neural lobe of the hypophysis. In addition, OT also acts as a neurotransmitter. As a circulating hormone, OT induces milk ejection from the lactating mammary gland and uterine contractions during parturition. Indeed, OT represents the most potent uterotonic agent known and is commonly used in obstetrical practice to augment labor and prevent postpartum hemorrhage. Moreover, OT antagonists represent promising pharmacotherapeutic tools for the treatment of premature labor. As a neurotransmitter, OT mediates specific behaviors, including sexual, maternal, social, and affiliative behaviors. In addition, an anxiolytic action has also been observed. These effects have led to the speculation that, in humans, defects in the OT system may underlie mood disorders, such as depression and obsessive–compulsive behaviors, as well as disorders in interpersonal relationships, such as autism. Additional proposed functions of OT include a role in natriuresis, pituitary prolactin release, cardiac atrial natriuretic factor release, penile erection, sperm transport, and vasodilation as well as T-cell and bone maturation. This article will review the mechanisms of neural biosynthesis, release, and action of OT and discuss its established and proposed roles in mammalian and human physiology.

I. THE HYPOTHALAMO–NEUROHYPOPHYSIAL SYSTEM

A. Anatomy

Among all of the known neuropeptides, oxytocin (OT) and the related peptide arginine vasopressin (AVP)

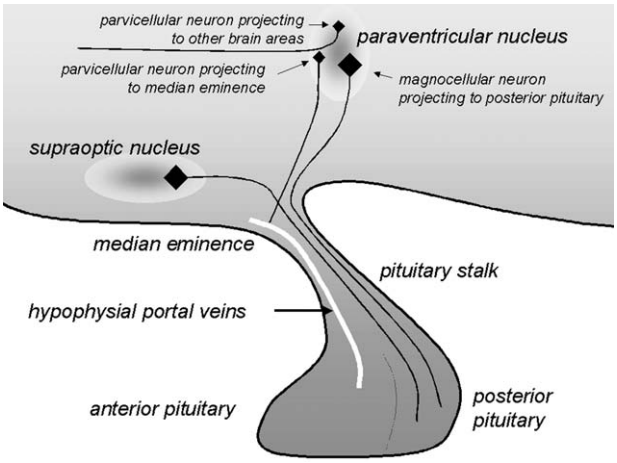


Figure 1 Schematic diagram of the hypothalamo–neurohypophysial system.

share a unique characteristic: on the one hand, both peptides are produced in cells that exhibit all the hallmarks of classical neurons, whereas on the other hand, both neuropeptides are released into, and act via, the general circulation, much like a classical hormone. This feature was recognized long before the precise structures of the two hormones were known and led to the development of the concept of *neurosecretion* during the first half of the last century.

OT is produced in specific hypothalamic neurons present within the supraoptic and paraventricular nuclei of the hypothalamus. The axons of these neurons extend through the internal zone of the median eminence and end in the neural lobe of the pituitary gland, also called the posterior pituitary or neurohypophysis (Fig. 1). The neural lobe represents not only the site of OT release into the general circulation but also the site where microgram quantities of OT are stored in preparation for acute release. The related peptide AVP is produced in a set of different but similar hypothalamic neurons that also extend their axons to the neurohypophysis. Together, these neurons and their axonic extensions constitute the *hypothalamo–neurohypophysial system*. Due to their relatively large size, these specialized neurons have been termed magnocellular neurons. These cells share all of the characteristics of a classical neuron. In addition, they also fulfill the definition of an endocrine cell. Thus, they represent a classical example of a so-called neuroendocrine cell type.

The hypothalamo–neurohypophysial system was initially thought to be unique for its ability to produce and release hormonal substances. We now know that neurons in general are capable of peptide biosynthesis and release and that neuropeptides play a major role as neurotransmitters. The hypothalamo–neurohypophysial system is, however, still unique in the sense that its peptide products act as circulating hormones. As a consequence, the hypothalamo–neurohypophysial system has long served as a classical model system for the study of the mechanisms of neurosecretion and excitation–secretion coupling.

OT, as well as AVP, is also produced in additional, smaller neurons, so-called parvicellular neurons, present within the paraventricular nucleus of the hypothalamus. In contrast to the magnocellular neurons, the axonal processes of these neurons project not to the neural lobe but to different brain areas as well as to the spinal cord, where OT and AVP exert different neurotransmitter functions (see later discussion). Other parvicellular OT and AVP neurons project to the external zone of the median eminence. There, the secretory products are released into the pituitary portal system, a specialized vascular system that allows further transport of the hormones to the anterior pituitary gland, where they exert roles as releasing factors for pituitary hormones (Fig. 1).

B. Electrophysiology

1. Neurosecretion

Excitation of a neuron occurs by the production of action potentials. An action potential is a short-lasting reversal of the electrical potential across the cell membrane. This reversal of the membrane potential originates at the level of the cell body, but then spreads like a shock wave along the different neuronal extensions, including axons and dendrites. At the level of the nerve endings, the arrival of action potentials leads to membrane depolarization, which, in turn, triggers the opening of voltage-sensitive calcium channels. The resulting increase in intracellular calcium concentration stimulates a process termed exocytosis, by which peptide-containing secretory granules fuse with the cell membrane and thereby release their contents into the extracellular space. Through this mechanism that couples excitation to secretion, an increase in the frequency of action potentials arriving at the nerve terminus results in a proportional increase in neuropeptide secretion.

In summary, the axon of a neurosecretory cell fulfills a dual function: Through axonal transport, the secretory granules are delivered to the nerve ending, where they are initially stored. Second, the axonal membrane serves as the medium for the propagation of

the action potentials and thus relays the signal for acute neuropeptide release from the cell body to the nerve ending.

2. Patterns of Electrophysiological Activity

In the lactating rat, continuous suckling of the pups triggers the episodic release of milk from the mammary gland every 5–10 min, a phenomenon known as the milk-ejection reflex. The underlying neuronal mechanism involves a dramatic activation of the entire hypothalamic magnocellular OT neuron population that is induced by the suckling stimulus and that lasts several seconds only. These rhythmic bursts of neuronal activity lead to a rhythmic release of OT from the neurohypophysis into the circulation. The burst of circulating OT induces milk ejection via contraction of the network of myoepithelial cells that surrounds mammary gland milk ducts. In lactating women, milk-ejection activity has also been shown to be rhythmic, and breast feeding is associated with a rise in circulating OT levels.

In all placental mammals studied thus far, parturition represents another condition that is associated with a strong increase in the electrical activity of OT neurons and, as a consequence, an increase in circulating OT levels. Whether this increased activity of OT neurons initiates, or is stimulated by, parturition has remained unclear. On the one hand, electrical stimulation of the neurohypophysis can bring about parturition; on the other hand, parturition-induced distention of the cervix and vagina represents a stimulus for OT release. Detailed measurements of OT secretion in humans revealed that, in parturient women, OT secretion occurs in short pulses. Pulse frequency as well as pulse duration increases during spontaneous labor. These findings suggest that, in humans, pulsatile activation of the OT neuronal system may be involved in triggering and/or augmenting labor. OT secretion also follows a diurnal pattern: in late gestation, OT levels are higher at night than during daytime.

Additional stimuli that affect, seemingly indiscriminately, the activity of both OT and vasopressin neurons include increased serum osmolality and decreased blood pressure, mating, and certain kinds of stress, including fever. In contrast to suckling, which induces short bursts of activation that are punctuated by relative silence, the preceding stimuli lead to a continuous activation of OT neurons.

The patterns of electrophysiological activity displayed by AVP neurons differ very markedly from those of OT neurons. Upon activation by either an increase in blood osmolality or a decrease in blood pressure, AVP neurons display a phasic firing pattern of activity. This pattern is characterized by individual bursts of activity that alternate with periods of silence, each lasting 20–40 sec. In contrast to the highly coordinated activity of OT neurons during lactation, the phasic bursts of AVP neurons are not temporally coordinated between individual neurons, and each neuron follows its own pattern controlled by an intrinsic rhythm generator. These characteristics make it possible to distinguish OT and AVP neurons during electrophysiological investigations on the basis of their distinctive firing patterns.

The mechanisms underlying these differential patterns of activity have been the subject of intense electrophysiological and biochemical investigations. Cell-intrinsic factors include the differential distribution of specific calcium-binding proteins, voltage- and ligand-gated ion channels, internal calcium release mechanisms, and gap junction conductances. In addition, these neurons are synaptically connected via an intrahypothalamic network as well as by afferences from a large number of limbic and sensory structures. Among the neurotransmitter receptors localized to magnocellular neurons, the most numerous appear to be the ones for amino acid transmitters, with glutamate providing excitatory input and GABA providing inhibitory input. The glutamate receptors that mediate excitatory postsynaptic potentials are mainly of the AMPA type. $GABA_A$ as well as $GABA_B$ (but not $GABA_C$) receptors are present on the postsynaptic membrane of supraoptic magnocellular nucleus, where they carry chloride currents inducing inhibitory postsynaptic potentials. In addition, metabotropic GABA receptors have also been shown to be expressed postsynaptically. Finally, OT and AVP release from dendrites and recurring axon collaterals provides the means by which OT- and AVP-containing magnocellular neurons are capable of influencing their own activity. This short-loop feedback control is being referred to as an autocontrol mechanism.

II. BIOSYNTHESIS

A. Peptide Structure

The structure of OT is shown in Table I. The structure is the same in all mammalian species studied. The cysteines in positions 1 and 6 are linked by a disulfide bond. Therefore, the molecule consists of a cyclic part with a C-terminal linear extension. Both cyclization

Table I
Sequences of Oxytocin, Vasopressin, and Vasotocin

	┌──────S-S──────┐
Oxytocin	Cys-Tyr-Ile-Gln-Asn-Cys-Pro-Leu-Gly-NH_2
Vasopressin	Cys-Tyr-Phe-Gln-Asn-Cys-Pro-Arg-Gly-NH_2
Vasotocin	Cys-Tyr-Ile-Gln-Asn-Cys-Pro-Arg-Gly-NH_2

and amidation are important for biological activity. OT differs in positions 3 and 8 from its related molecule AVP. Both peptides have likely evolved from a common precursor molecule. A neurohormone present in several nonmammalian species, termed vasotocin, represents an intermediate between OT and AVP (Table I).

Both OT and AVP are produced as parts of two similar precursor molecules. The OT precursor molecule contains OT at its N-terminus, linked via a pair of basic amino acids to a 10-kDa peptide called neurophysin-I of unclear biological function. Following synthesis on ribosomes of the rough endoplasmic reticulum, the peptide precursors are packaged into secretory granules and transported via fast axonal transport mechanisms to their destination at the nerve terminals, where they are released in response to membrane depolarization. While residing in the secretory granule, the precursor molecules are processed by specific endoproteases (called processing enzymes) that are copackaged into the granules to generate the biologically active nonapeptides.

B. Gene Structure and Regulation

The genes encoding OT and AVP are in close proximity on the same chromosome in all mammalian species studied and have evolved by gene duplication of an ancestral gene. In humans, the OT and AVP genes are in close linkage on chromosome 20p13, separated by only 12 kb of intergenic sequences. The two genes are positioned tail to tail and are thus transcribed from opposite strands. In all species studied, the OT and vasopressin genes consist of three exons that are separated by two relatively small introns. In each gene, the biologically active nonapeptide is encoded by exon 1.

Studies on the human, rat, and bovine OT genes revealed that the OT gene is under positive and negative control by a composite hormone regulatory region that is able to interact with several members of the nuclear receptor superfamily. This element is located about 160 nucleotides upstream of the site of transcriptional initiation and mediates positive transcriptional regulation by interacting with the estrogen, retinoic acid, and thyroid hormone receptors as well as with the nuclear orphan receptor SF1. Negative control is exerted by the nuclear orphan receptors COUP-TF I and II, involving competition with stimulatory factors as well as active silencing. *In vivo*, hypothalamic OT gene expression is stimulated during pregnancy and lactation in response to dehydration and following a combination of estrogen stimulation and progesterone withdrawal.

Outside the central nervous system, the OT gene is expressed in ovarian granulosa cells, fetal membranes and uterine decidua, placenta, testicular Leydig cells, and thymic epithelium. OT released from these sites is likely to have a role as a local paracrine mediator.

III. PHYSIOLOGY

A. Peripheral Effects

OT plays an essential role in mediating milk ejection from the lactating mammary gland. This has been clearly demonstrated in the mouse, where ablation of the OT gene by gene targeting leads to a complete inability to lactate, despite normal mammary gland milk production. In humans, suckling-induced and spontaneous increases in mammary gland milk pressure occur concomitantly with increased circulating OT levels, and morphine administration during lactation attenuates both OT pulses and milk-ejecting activities.

In contrast to the clear-cut, essential role of OT in lactation, there are arguments both for and against the importance of OT in the processes of labor and parturition. Mice in which the gene for OT has been inactivated are able to deliver normally. This indicates that, at least in the mouse, OT is not essential for normal parturition. Considering the multiplicity of interconnected mechanisms that are mediating the important function of parturition, it is not surprising that there is redundancy in this system and that knocking out one of its components does not necessarily lead to the breakdown of the entire process. More importantly, this finding does not rule out that premature overactivation of the OT system may be a determining factor in cases of idiopathic premature labor. Support for this idea stems primarily from a large number of studies that attest to the effectiveness of OT antagonists in inducing uterine quiescence in

normal labor as well as in cases of naturally occurring or experimentally induced premature labor. The effectiveness of OT antagonists, specifically atosiban, to delay premature labor has been demonstrated in many different species, including humans.

B. Central Effects

OT-containing nerve endings as well as functional OT receptors are present at various sites in the mammalian brain, including the ventromedial nucleus of the hypothalamus, the bed nucleus of the stria terminalis, ventral subiculum, lateral septum, caudate nucleus, limbic system, hippocampus, and olfactory nuclei. A potential role of OT in human brain function has mainly been inferred from studies in nonprimate mammals, including rats, sheep, and voles. The combined evidence demonstrates that OT induces specific sexual, affiliative, social, and maternal behaviors. In addition, it reduces anxiety and impairs short-term memory. Several of these behavioral effects depend on previous priming by gonadal steroids, most likely via steroid-induced up-regulation of OT receptors.

1. Female Behavior

Among the sexual behaviors, the lordosis reflex has been intensively studied in the rat. This reflex represents a sign of female sexual responsiveness that can be quantitated. In female rats primed with gonadal steroids, the reflex is strongly enhanced by central OT administration and attenuated by OT antagonists. Concomitantly, central OT administration decreases rejection behavior against male sexual advances. Detailed microinjection studies with OT antagonists injected into discrete brain areas implicate the medial preoptic area and the ventromedial nucleus of the hypothalamus in OT-mediated lordosis, whereas the ventral tegmental area appears to regulate OT-mediated social behaviors.

For studies of OT effects on social attachment and affiliation, two types of voles, the prairie and the montane vole, have provided a valuable model system. Despite their close relation, these two species differ markedly with respect to their social behaviors. The prairie vole manifests the classic features of monogamy. A breeding pair shares the same nest and stays together until "death do them part," and even then a new mate is accepted by the surviving mate only one-fifth of the time. In contrast, montane voles live in isolated burrows, have little social contact, and clearly are not monogamous. Despite an overall similar distribution of neuropeptide receptors, these two species differ very radically in the neural distribution of OT as well as AVP receptors, and both are expressed within very different pathways. Whereas prairie vole OT receptors are found in brain regions associated with reward, such as the nucleus accumbens and the prelimbic cortex, montane vole OT receptors are found in the lateral septum, where they may be responsible for the effect of OT in this species on self-grooming. Proof that OT is involved in mediating these very specialized behavioral patterns came from experiments that demonstrated that OT centrally injected to female prairie voles facilitated the development of partner bonding in the absence of mating and that the mating-induced partner bond formation was blocked by central administration of an OT antagonist before mating. These results indicate that OT released during mating is both necessary and sufficient for the formation of a pair bond in the female prairie vole.

Maternal behavior is another complex behavior that can be induced in virgin, steroid-primed female rats, and the natural occurrence of the behavior following parturition can be blocked by central OT antagonist application. Typical OT-induced maternal behavior includes nest building and recognition and acceptance of the pups, as well as retrieval of the pups into the nest. This function also involves olfactory clues, and a role of OT in olfactory memory has indeed been suggested, consistent with the occurrence of OT receptors in brain areas involved in processing olfactory information.

Because in women OT is released during intercourse and during lactation, it is tempting to speculate that the associated central release of OT may contribute to the strengthening of the emotional bond between sexual partners as well as of the affiliative bond between mother and infant. No direct studies are available in humans, but studies in squirrel monkeys have demonstrated that central OT administration is capable of influencing social interaction also in primates. Preliminary findings in nonlactating women indicate that positive emotions (relaxation, massage) induce a rise in circulating OT levels. This increase appears to be more pronounced in women that display a lesser degree of anxiety in close relationships. In addition, there is also one anecdotal report of increased sexual receptivity in a woman in response to OT administration by nasal spray.

2. Male Behavior

The role of central OT in controlling sexual and social behavior in the male is less well-defined and complex.

In the male rat, sexual activity increases the activity of parvicellular OT neurons in the paraventricular nucleus of the hypothalamus. Some of these OT neurons project to sensorimotor nuclei regulating penile erection, and microinjections of OT into the paraventricular nucleus indeed induce penile erection. Moreover, erections that are physiologically induced due to the presence of a female are inhibited by injection of a specific OT antagonist into the lateral ventricles. Whereas central administration of low acute doses of OT facilitates copulative performance, high acute doses attenuate sexual responses, and chronic intracerebroventricular OT infusion results in markedly enhanced social, but not sexual, contact with the female partner. These observations taken together indicate that OT's effects on social behavior can be separated from its effects on sexual behavior and that these effects are likely mediated by OT receptors located in different brain areas.

Interestingly, in the male vole, it is not OT but rather AVP that influences parental behavior. AVP given centrally increases affiliative behavior in male prairie voles in the absence of mating and induces aggression against other sexually mature males. Administration of a specific AVP antagonist blocks the mating-induced partnership bonding. This effect was only present in the prairie vole but not the montane vole. A most elegant demonstration of the importance of region-specific receptor expression in behavioral patterns has been provided by the group of Insel with the creation of a transgenic mouse in which the prairie vole vasopressin V_{1a} receptor was expressed under the control of its own species-specific promoter. In these transgenic mice, the V_{1a} receptor was expressed at sites within the brain that are typical for the prairie vole. As a result, central injection of AVP increased affiliative behavior in the transgenic, but not the wild-type, mice. These data indicate that the specific pattern of V_{1a} receptor gene expression in the brain is functionally associated with species-typical social behaviors.

3. Disorders of Mood and Behavior

On the basis of the demonstrated effects of OT on affiliative and social behaviors in rodents, it has been speculated that a defect in central OT release or OT responsiveness may underlie the manifestations of autism and/or obsessive–compulsive disorders in humans. The evidence from clinical studies to support this idea, however, is limited. Some clinical data suggest that a subgroup of obsessive–compulsive disorder patients have increased OT concentrations in the cerebrospinal fluid. On the other hand, a strong decrease in OT plasma levels has been observed in autistic children, and these children failed to show the normal developmental increase of plasma OT with age.

An additional central OT effect involves the reduction of anxiety. Central OT administration to estrogen-primed rats or mice induces them to spend significantly longer times in the open arm of a maze. These effects have led to the speculation that, in humans, OT may act as an endogenous antidepressant and anxiolytic hormone and that defects in the OT system may underlie mood disorders that are accompanied by increased anxiety levels, such as depression and obsessive–compulsive behaviors. The presence of OT receptors in the limbic forebrain area in humans represents supportive evidence that OT may be involved in the control of emotional and affective behavior in humans. Because, in laboratory animals, serotonin receptor agonists as well as serotonin re-uptake inhibitors produce a marked increase in plasma OT levels, it has also been proposed that OT may mediate, at least in part, the antidepressant effects of serotonin re-uptake inhibitors in humans.

The marked differences in the OT pathways between different species preclude a simple extrapolation from mice to men. However, the available animal data strongly suggest that OT is an important modulator of several specific behaviors related to reproduction and interindividual relationships. Therefore, their role in human behavioral pathologies, such as autism, depression, and obsessive–compulsive behaviors, deserves to be explored. Because animal studies have shown that discrete changes in the regional expression of OT receptors can have an impact on social behaviors, measurement of levels in CSF or plasma OT concentrations will not be sufficient, and functional or structural variations in OT or OT receptor genes and their specific expression in brain tissues will also have to be investigated.

C. Additional Functions

On the basis of animal studies, roles for OT have been proposed in pituitary, renal, cardiovascular, immune, and male reproductive functions. Although a corresponding role in humans is likely, no evidence is available from clinical studies to support this idea.

OT released from nerve terminals at the median eminence reaches the anterior pituitary via the pituitary portal system where it acts as a releasing factor,

potentiating the release of pituitary corticotropin, prolactin, and gonadotropins. At the level of the kidney, OT induces natriuresis, and this effect involves nitric oxide and cyclic GMP. Additional proposed roles of OT in male reproductive functions include the stimulation of sperm transport and testicular androgen production. The detection of OT receptors in endothelial cells as well as in the atria and ventricles of the heart led to the concept that OT has a role in negatively regulating blood pressure. The presence of functional OT receptors on T-cells implies a role in T-cell function, and OT effects on human and mouse mammary cancer cells suggested an antiproliferative effect of OT.

IV. MECHANISMS OF ACTION

A. Oxytocin Receptors

The various central and peripheral actions of OT that were discussed earlier all appear to be mediated by a single type of OT receptor. The structures of the human, rat, pig, sheep, cow, and mouse OT receptors have been elucidated by molecular cloning. The deduced amino acid sequences indicate that the OT receptor forms part of the family of G-protein-coupled receptors consisting of seven hydrophobic transmembrane α-helices. This receptor is closely related to the three receptors that mediate the actions of AVP, namely, the V_{1a}, V_{1b}, and V_2 receptors. Whereas AVP is a weak activator of the OT receptor, OT does not interact with the V_{1a} or V_2 receptor and interacts only weakly with the V_{1b} receptor. The gene encoding the human OT receptor is located on chromosome 3p25, in close apposition to the caveolin 3 gene.

The distribution of OT receptors in the brain has been carefully mapped in several rodent species. In contrast to the high degree of conservation with respect to the structure of the OT receptor in all mammalian species, important differences in the pattern of OT receptor expression in the CNS have been observed between different species. Therefore, any extrapolation of insights gained from animal studies to humans has to be done with caution. In humans, OT receptors have been described in (but are probably not limited to) the basal forebrain cholinergic nuclei, the nucleus basalis of Meynert, the diagonal band of Broca, and the preoptic area of the hypothalamus. The localization in the diagonal band of Broca is consistent with a role of OT in memory. The preoptic area of the hypothalamus is an area considered important for the mediation of reproductive behaviors; therefore, the presence of OT receptors in this area is consistent with the role of OT in human reproductive behavior.

In the periphery, specific OT-binding sites are present within the uterus at the level of the myometrium as well as the uterine epithelium or decidua. Whereas myometrial receptors mediate uterine contractions directly, activation of the decidual receptor elicits prostaglandin $F_{2\alpha}$ release, which, in turn, not only enhances uterine contractions but also promotes cervical ripening and luteolysis. Additional OT-binding sites have been detected in mammary gland myoepithelial cells, pituitary lactotrophs, cultured astrocytes, certain spinal cord neurons, renal macula densa cells, thymic T-cells, ovarian follicular cells, testicular Leydig cells, cardiac myocytes, vascular endothelial cells, and cultured osteoblasts.

Compared to other members of the OT–vasopressin receptor family, relatively little is known of the specific regions of the OT receptor involved in ligand binding and signaling. It is assumed that the cyclic part of the OT molecule is lodged in the upper one-third of the receptor binding pocket and interacts with transmembrane domains 3, 4, and 6, whereas the linear C-terminal part of the OT molecule remains closer to the surface and interacts with transmembrane domains 2 and 3 as well as with the connecting first extracellular loop. This hypothesis is supported by studies employing site-directed mutagenesis, photoaffinity labeling as well as by experiments involving domain swapping between the V2 vasopressin receptor and the OT receptor.

Activation of the OT receptor leads to phospholipase C-mediated phosphoinositide hydrolysis, followed by activation of protein kinase C and intracellular calcium mobilization. The rise in intracellular calcium is greatly attenuated in the absence of extracellular calcium, indicating that OT receptor activation also stimulates calcium influx, probably via calcium-induced (capacitative) calcium entry. Experiments with antibodies against different G-proteins demonstrated that coupling to phospholipase C occurs mainly via a G-protein of the $G_{q/11}$ family. However, evidence for coupling to other G-proteins, including G_i and G_{ha}, has also been presented. The regions of the OT receptor implicated in G-protein interactions have been delineated by examining the inhibitory effect of coexpression of different intracellular domains of the receptor. The studies indicate that all four intracellular domains are involved in coupling to $G_{q/\alpha 11}$.

Additional effector mechanisms involve coupling to phospholipase A_2 and phosphorylation and activation of MAP kinase. These two pathways are relevant for the stimulation of prostaglandin $F_{2\alpha}$ and prostaglandin E_2 synthesis. The precise mechanisms by which OT receptor activation induces smooth muscle contractions and prostaglandin synthesis remain to be defined. In myometrial cells, calcium-induced activation of the calcium–calmodulin complex leading to activation of myosin light chain kinase is involved in initiating the contraction process. In decidual cells, stimulation of prostaglandin synthesis is the consequence of an OT-induced activation of cyclo-oxygenase II expression. In vascular endothelial cells and renal macula densa cells, OT receptor activation leads to a calcium-dependent activation of nitric oxide synthase, inducing vasodilation and natriuresis, respectively.

B. Oxytocin Receptor Regulation

An outstanding feature of the OT receptor gene is the dramatic change in the expression of this gene in different tissues. This highly regulated expression sets the OT receptor apart from the majority of other G-protein-coupled receptors in general and from the closely related vasopressin receptors in particular. During gestation, uterine OT (but not vasopressin) receptors are up-regulated by one to two orders of magnitude in all mammalian species studied, including humans. This phenomenon induces a strong increase in uterine sensitivity toward OT and is due to a dramatic increase in uterine OT receptor gene expression just prior to parturition. Indeed, the increase in uterine OT receptors is much more striking than, and precedes, the increase in circulating OT. Therefore, it has been proposed that OT receptor up-regulation represents the trigger for parturition. Whereas the number of *uterine* OT-binding sites decreases quickly following parturition, *mammary gland* OT-binding sites reach a maximum shortly after parturition and remain elevated throughout lactation. This differential, tissue-specific regulation of OT receptor expression enables a switch in target organs and allows circulating OT to assume a dual role: uterine contractions during labor and milk ejection during lactation. The same OT receptor gene is expressed in both tissues, and the molecular basis underlying this tissue-specific differential regulation of OT receptor gene expression remains to be elucidated.

Several studies have shown that uterine OT receptors are up-regulated by estrogens and down-regulated by progesterone. Indeed, in most mammals, the situation before term is characterized by high estrogen levels (involving placental conversion of fetally produced dehydroepiandrogen and androstenedione to estradiol) and a sharp decline in progesterone (induced by luteolysis; exceptions include humans and primates, in whom no fall in progesterone occurs). The sudden drop in the progesterone: estrogen ratio is thought to be causally involved in triggering parturition. This idea is further supported by findings in the prostaglandin F receptor "knock-out" mouse. In this mouse model, preterm luteolysis and the ensuing fall in progesterone are prevented. This leads to complete suppression of parturition and suppression of OT receptor gene expression. However, both OT receptor expression and parturition are restored if a fall in progesterone is induced by ovariectomy.

We and others have determined that the estrogen-induced up-regulation is due to a genomic estrogen-mediated increase in OT receptor gene expression. On the other hand, progesterone may act to decrease OT binding via both a genomic effect and a nongenomic effect, leading to the inhibition of OT receptor binding and signaling functions. Moreover, the effect of estrogens on OT receptor expression is region-specific: some, but not all, central OT receptors are up-regulated at the end of pregnancy and in response to estrogen administration. Specifically, OT receptors in regions rich in estrogen receptors, including the ventromedial nucleus of the hypothalamus and the bed nucleus of the stria terminalis, are highly up-regulated in response to estrogens and prior to parturition. Up-regulation of OT receptors in these two key limbic brain regions prior to parturition is coincident with the onset of maternal behavior. On the other hand, OT receptors expressed at other central OT target sites, such as the subiculum or the olfactory nuclei, remain (in the rat) unaffected by gonadal steroids.

In the periphery, estrogens stimulate OT receptor gene expression in the uterus, pituitary, and kidney, but OT receptor expression in the mammary gland remains unaffected by steroid treatment. Although the rat and human OT receptor genes contain potential estrogen-responsive sites in their promoter regions, none of these sites turns out to be estrogen-responsive when analyzed at the original distance from the transcriptional initiation site. Cell-specific estrogen responsiveness may be mediated indirectly or via other functional OT receptor promoter elements, including AP1 sites and a cAMP response element. Interleukins, specifically IL-1β and IL-6, represent additional

potential regulators of OT receptor expression. They are present in high concentrations in amniotic fluid at term labor and are mediators of infection-induced preterm labor. Indeed, the rat and human OT receptor genes contain potential interleukin response elements in their promoter regions; however, proofs of their functionality have not been provided yet. Certainly, the mechanisms regulating OT receptor gene expression in neural and nonneural tissues as well as the precise molecular mechanisms that underlie OT signaling represent exciting areas for future research.

See Also the Following Articles

NEUROTRANSMITTERS • SEXUAL BEHAVIOR

Suggested Reading

Gainer, H., and Wray, S. (1994). Cellular and molecular biology of oxytocin and vasopressin. In *The Physiology of Reproduction*, pp. 1099–1129. Raven Press, New York.

Insel, T. R. (1997). A neurobiological basis of social attachment. *Am. J. Psychiatry* **154,** 726–735.

Insel, T. R., Young, L., and Wang, Z. (1997). Molecular aspects of monogamy. *Ann. N.Y. Acad. Sci.* **807,** 302–316.

Insel, T. R., O'Brien, D. J., and Leckman, J. F. (1999). Oxytocin, vasopressin, and autism: Is there a connection? *Biol. Psychiatry* **45,** 145–157.

Ivell, R., and Russell, J. A. (1995). *Oxytocin: Cellular and Molecular Approaches in Medicine and Research*. Plenum Press, New York.

McCarthy, M. M., and Altemus, M. (1997), Central nervous system actions of oxytocin and modulation of behavior in humans. *Molecular Medicine Today* **3,** 269–275.

McDougle, C. J., Barr, L. C., Goodman, W. K., and Price, L. H. (1999). Possible role of neuropeptides in obsessive compulsive disorder. *Psychoneuroendocrinology* **24,** 1–24.

Pedersen, C. A., Caldwell, J. D., Peterson, G., Walker, C. H., and Mason, G. A. (1992). Oxytocin activation of behavior in the rat. *Ann. N.Y. Acad. Sci.* **652,** 58–69.

Young, L. J., Nilsen, R., Waymire, K. G., MacGregor, G. R., and Insel, T.R. (1999). Increased affiliative response to vasopressin in mice expressing the V1a receptor from a monogamous vole. *Nature* **400,** 766–768.

Zingg, H. H. (1996). Vasopressin and oxytocin receptors. In *Membrane Surface Receptors, Bailliere's Clinical Endocrinology and Metabolism*. (M. C. Sheppard and J. A. Franklyn, Eds.), Vol. 10**,** No. 1, pp. 75–96. Bailliere Tindall, London.

Zingg, H. H., Bourque, C. W., and Bichet, D. G. (1998). *Vasopressin and Oxytocin: Molecular, Cellular and Clinical Advances*. Plenum Press, New York.

Pain and Psychopathology

DAVID FISHBAIN
University of Miami School of Medicine

GLOSSARY

Axis I, II, and III DSM-IV classifications of pain disorders; Axis I is for any acute or chronic mental disorder, Axis II is for personality problems, and Axis III is for non-psychiatric medical disorders, e.g., pneumonia.

soft tissues Tendons, ligaments, and muscles, as opposed to the hard tissues (bones, cartilage).

trigger points A tender point on the soft tissues, pressure on which results in acute pain radiating in a predictable direction.

worker's compensation patients Patients who have been injured while performing their job and as a result of their injury are now receiving worker compensation benefits.

Pain plays a major role in the development of psychopathology in chronic pain. Research also indicates that psychiatric comorbidities often impact negatively on chronic pain and functional status.

I. THE PROBLEM OF NONSPECIFIC LOW BACK AND NECK PAIN AND THEIR IMPACT ON PSYCHOPATHOLOGY IN CHRONIC PAIN

There are three possible relationships between pain and psychopathology. Psychopathology could *cause* the development of pain, it could follow or be caused by the development of pain, or prepain psychopathology could predispose one to the development of pain. There is now good evidence that the development of psychopathology follows–is caused by the development of pain or that preexistent behavior vulnerability predisposes one to the development of psychopathology after the onset of pain. In addition, there is also good evidence that the severity of psychopathology parallels the severity of perceived pain. The fact that pain may be an important etiological factor in the development of psychopathology leads to a major problem in utilizing the current psychiatric diagnostic nomenclature system, the DSM-IV. Here, in a diagnosis termed pain disorder, the psychiatric diagnostician is required to determine whether psychological factors have a major role in the onset, severity, exacerbation, or maintenance of the pain. Yet, as pointed out earlier, pain plays a major role in the development of psychopathology in chronic pain. At the time the psychiatrist sees the patient, he or she is assessing both prepain- and pain-associated psychopathology. As such, at that time it is almost impossible to make a distinction between what was present before the onset of pain and what developed after. There is also a tendency to attribute all of the assessed psychopathology as preexistent to the pain or causing the pain if a cause for the pain cannot be found. Thus,

the issue of "nonspecific" low back pain (LBP) or neck pain is of great importance to pain medicine.

Nonspecific is the opposite of "specific pain," which is defined as pain for which an underlying cause can be defined, such as infection, vertebral fracture, systemic disease, etc. Currently, it is believed that the vast majority of pain problems do not fit into the classical biomedical model that links symptomatology and disease process. As such, a definite somatic cause for pain can only be identified in no more than 10–20% of cases. Thus, the nonspecific pain diagnosis is determined or defined by the process of excluding specific pain. Because so few pain patients can be given a specific diagnosis, many patients are suspected of having psychopathology that causes their pain or makes it worse. If these nonspecific patients do have psychopathology such as depression, the presence of this psychopathology serves as proof that there is not an organic reason for their pain. The nonspecific label can thus color the psychiatric diagnostic process and lead to erroneous psychopathology data.

II. THE PROBLEM OF FIBROMYALGIA–MYOFASCIAL PAIN SYNDROME AND ITS IMPACT ON ALLEGED PSYCHOPATHOLOGY IN CHRONIC PAIN

Fibromyalgia (FMS) and myofascial pain syndrome (MFPS) are a group of soft tissue syndromes characterized by widespread pain and tender points in FMS and regional pain and trigger points in MFPS. The important issue in reference to these syndromes is that they are commonly found in patients with intractable pain. At pain facilities, the numbers range from 30 to 100%. Thus, it is unlikely that a significant number of pain patients presenting to a pain facility will have true nonspecific pain, if an adequate soft tissue examination is performed.

The issue of whether the pain patient has nonspecific pain or actually suffers from a soft tissue syndrome such as FMS or MFPS is important behaviorally in two ways. First, the pain patient needs and awaits an explanation for the pain and feels that the pain needs to be legitimized. Pain patients consider "delegitimization," referring to the repudiation of the patient's pain and suffering, as one of the major perceived difficulties of the pain experience. A pain patient's satisfaction with his or her medical care is dependent on receiving an adequate explanation for his or her care. Pain patients that want more diagnostic tests are less satisfied with their care and are less likely to want the same doctor again. Clearly, the nonspecific label does not provide an explanation nor does it satisfy the patient's expectation of knowing the exact cause of the pain. The patient to whom the nonspecific label applies generally suffers from chronic pain (to be discussed later). The process of learning how to manage this pain can only begin after the patient understands that his or her condition is chronic. FMS and MFPS provide the framework for this understanding, as these syndromes are understood as chronic. Thus, pain patients labeled with FMS or MFPS should, in theory, decrease their health care utilization. This is indeed the case: once the diagnosis is made, hospitalizations and health care utilization drop dramatically for FMS patients. There is, therefore, reason to believe that the labeling of pain patients with an FMS or MFPS diagnosis, if appropriate, can have positive behavioral consequences that can affect pain treatment outcome. The issue of no diagnosis or the nonspecific pain label as the reason for pain is such a common patient complaint that it is considered to be a behavioral comorbidity usually associated with having chronic pain. The various comorbidities usually associated with chronic pain will be discussed in detail later.

The second reason why the presence–absence of a soft tissue diagnosis is important behaviorally is because the absence of MFPS and FMS makes the patient a candidate for a DSM-IV diagnosis of pain disorders on Axis I, which makes it a mental disorder. However, if a patient is diagnosed with MFPS or FMS and has chronic pain, in the DSM-IV this is diagnosed as a pain disorder associated with a general medical condition and is placed on Axis III. This is not a mental disorder. Thus, in the context of chronic pain, the presence of MFPS or FMS can determine whether a patient receives a mental disorder diagnosis and is thus "Labeled."

III. WHAT IS CHRONIC PAIN?

It is to be noted that nonspecific pain is grouped in the ICD-9 (World Health Organization, 1997) classification scheme under the heading "other and unspecified disorders of the back." However, nonspecific pain is not the same and does not necessarily equal chronic pain (CP). The American Medical Association in the *Guides to the Evaluation of Permanent Impairment* (4th ed.) has, however, defined chronic pain and developed operational criteria for its diagnosis. Here chronic pain is defined as "persistent pain… recalcitrant… with

Table I
Diagnostic Characteristics for the Diagnosis of Chronic Pain Syndrome: At Least Four of the Following Are Required for a Presumptive Diagnosis of Chronic Pain Syndrome

1. Persistent pain of greater than 2–4 weeks duration
2. Pain behaviors (verbal and nonverbal)
3 Vague, inconsistent, and inaccurate, indicating nonspecific pain
4. Substance abuse and/or dependence
5. Depression
6. Muscular dysfunction and deconditioning resulting in secondary pain of musculoskeletal origin
7. Withdrawal from work, recreational, and family endeavors
8. Dependence on physicians, spouses, and families

major psychosocial consequences… not a symptom of underlying acute somatic illness but a disease of the whole person… characterized by enhanced pain perception with maladaptive pain-related behaviors… .'' In addition, the AMA *Guides* point out that, if chronic pain is not recognized and properly treated, deterioration of coping mechanisms and limitations of functional capacity will ensue. Other consequences will be despair, alienation from family and society, loss of job, isolation, invalidism, and suicidal thoughts. However, the AMA *Guides* state that chronic pain *is not* a psychiatric disorder. The AMA *Guides*' diagnostic criteria for a presumptive diagnosis of chronic pain syndrome are presented in Table I. Four criteria are required for this diagnosis. It is to be noted that most of these criteria are either psychiatric or other comorbidities (to be discussed later). Thus, *the chronic pain syndrome cannot be defined and characterized without an evaluation for psychiatric or other comorbidities*. In addition, it is to be noted that criterion 3 (Table I) relates to nonspecific pain. Thus, nonspecific pain is encompassed within the chronic pain definition.

IV. PSYCHIATRIC COMORBIDITIES USUALLY ASSOCIATED WITH CHRONIC PAIN

Comorbidity is defined as any distinct clinical entity that has existed or that may occur during the clinical course of the patient who has the index disease under study. It has become clear that the presence of comorbid disease can dramatically affect the treatment of the index disease. Specifically, the presence of comorbid disease can often complicate, interfere with, or make the treatment of the index disease more difficult, making the prognosis worse. In addition, the presence of comorbid disease, because of its impact on the treatment of the index disease, can lead to spurious medical outcome data, especially if the comorbid disease is not classified or analyzed and its effect controlled for. For these reasons, the identification and treatment of psychiatric comorbidities associated with chronic pain have become extremely important to pain medicine.

As discussed earlier, the presence of chronic pain can predispose one to the development of psychiatric psychopathology. As such, one would expect to find significant psychiatric comorbidity within the chronic pain patient (CPP) population. In addition, other reasons why there should be significant psychopathology within CPPs are the following: (1) CPPs consider themselves to suffer from a physical illness for which physicians cannot seem to develop a cure; (2) this physical illness is associated with significant impairment and disability, which has tremendous impact on the CPP's life; (3) in order to control pain, CPPs are often placed on psychoactive substances that have dependence–addiction potential; and (4) because often no apparent tissue damage can be found to explain the etiology of chronic, benign, nonmalignant pain, CPPs are often given the nonspecific pain diagnosis by physicians who then attribute the CPP's pain to underlying psychiatric illness. Because of these and perhaps other less well-defined issues, it is

Table II
Types of Psychiatric Comorbidities Found within CPPs

I	Between the chronic pain state and any psychiatric disorders on Axis I of the DSM-IV
II	Between the chronic pain state and any psychiatric disorders on Axis II of the DSM-IV
III	Within CPPs and between psychiatric disorders on Axis I of the DSM-IV
IV	Within CPPs and between psychiatric disorders on Axes I and II of the DSM-IV
V	Within CPPs and between Axis I psychoactive drug use disorders and other Axis I disorders of the DSM IV
VI	Within CPPs and between various Axis I psychoactive drug use disorders of the DSM-IV

well-established that there is significant psychiatric comorbidity associated with chronic pain. There are six types of psychiatric comorbidities found within CPPs. These are presented in Table II and will be discussed individually next.

A. Comorbidity between the Chronic Pain State and Any Psychiatric Disorder on Axis I of the DSM-IV

This type of comorbidity has been studied extensively in CPPs and is the most exhaustively documented major category of psychiatric comorbidity in CPPs. This type of comorbidity data is presented in Table III. Here data from various research studies for prevalence percentages in CPPs for Axis I disorders are summarized.

Table III indicates that the affective disorders (depression) group is the most commonly found group of comorbid disorders. For example, some authors have reported a prevalence rate for depression approaching 100%. One reviewer concluded that depression is found more commonly in CPPs than in the general population and that the depression seen in CPPs is a consequence of chronic pain and is not

Table III
Percentages of Comorbidly Associated DSM Axis I Disorders with Chronic Pain

Psychiatric disorder	Prevalence within chronic pain patients (%)
Affective disorders (depression):	
Major depression	1.5–54.5
Dysthymia	0–43.3
Adjustment disorder with depressed mood	28.3
All forms of affective disorders	10–100
Psychoactive substance-related disorders:	
Current alcohol abuse–dependence	2–10.6
Current drug dependence (opioids, barbiturates, sedatives, cannabinoids)	5.2–34.0
Current illicit drug abuse (cocaine, cannabinoids, speed)	6.41–12.5
Total current alcohol and other drug dependence	14.9–23.4
Somatoform disorders:	
Somatization disorder	1–16.2
Conversion disorder	2–37.8
Psychogenic pain–pain disorder	0.3–97.0
Hypochondriasis	0.6–1.0
Anxiety disorders:	
Panic disorder	3–11
Panic disorder with agoraphobia	2.0–2.1
Posttraumatic stress disorder	1–9.7
Adjustment disorder with anxious mood	10–62.5
Obsessive–compulsive disorder	1.1–2.0
Phobic disorder	9
Total anxiety disorders	7–62.5
Other diagnoses:	
Psychotic disorders including schizophrenia	0–15.1
Marital problems	7–8.2
Psychological factors affecting medical condition	0–34
Adjustment disorder with work inhibition	5–14
Intermittent explosive disorder	9.9
No diagnosis on Axis I	3–5.2

involved in the cause of chronic pain. There are, however, discrepancies between authors on reported prevalence rates of the various types of affective disorder. This is demonstrated in Table III and is the reason why ranges are reported. These discrepancies relate to three problems: (1) differences in pain center CPP selection criteria; (2) a lack of operationally specified instructions for determining organic factors for the depression as required by the DSM-III-R and DSM-IV; and (3) the effect of pain on mood. The effect of pain on mood may be an extremely important issue and may be the reason for the high prevalence of comorbid affective disorders in CPPs.

The second most common group of Axis I psychiatric comorbidities to be found in CPPs is that of psychoactive substance use related disorders (see Table III). A review of this research area concluded that the prevalence percentages for the diagnoses of drug abuse, dependence, or addiction in CPPs were in the range of 3.2–18.9%. A significant percentage of CPPs, 6.41–12.5%, was reported to abuse illicit drugs (marijuana and cocaine). There was, however, little evidence that addictive behaviors were common.

The third most common group of Axis I psychiatric comorbidities to be found in CPPs is that of somatoform disorders (see Table III). Here, both conversion disorder and psychogenic–pain disorder are reported to have high prevalence rates (see Table III). There are major discrepancies, however, between authors on the prevalence of these two disorders. These discrepancies relate to reliability problems with DSM-III criteria for both of these diagnoses. Both of these diagnoses contain or have contained criteria that require the examiner to make a value judgment about a symptom. As a result of these problems, the DSM criteria for psychogenic pain were changed in the DSM-III-R and further changed in the DSM-IV. In addition, as discussed earlier, the identification of a patient as suffering from conversion disorder or pain disorder is heavily determined by whether the patient has received the nonspecific label as a cause of his or her pain. It is, therefore, likely that the data for the prevalence of these disorders are unreliable.

The fourth most common group of Axis I psychiatric comorbidities to be found in CPPs is that of anxiety disorders (see Table III). The reported high prevalence of anxiety disorders is not surprising because anxiety is the only psychiatric disorder that is commonly associated with any chronic medical condition. It is also reported that this association develops more quickly than any other association between a psychiatric disorder and a chronic medical condition. Although, as demonstrated in Table III, anxiety is common in CPPs, evidence indicates that these reported prevalences are underestimates. The evidence for this statement comes from studies performed with non-CPPs. One study reported that 81% of panic disorder patients had pain as a presenting complaint. In a second study it was reported that, in general practices, low back pain patients are more likely to have a diagnosis of anxiety than non-low back pain patients. Finally, CPPs, when compared with non-CPPs, are more likely to demonstrate avoidance of particular situations [e.g., injections and minor surgery, hospitals, sight of blood, thoughts of injury and illness (blood or injury phobia), being watched or stared at, and speaking or acting to an audience (social phobia)]. It is, therefore, likely that anxiety syndromes are comorbidly associated with chronic pain at a greater frequency than reported.

The fifth most common group of Axis I psychiatric comorbidities to be found in CPPs is that of the other diagnoses group (see Table III). Here, two diagnoses are important: psychotic disorders, including schizophrenia, and psychological factors affecting medical condition. Psychotic disorders, schizophrenia, delusional disorder, and bipolar affective disorder are not overrepresented within CPPs compared with the general population and are probably underrepresented. However, in patients with atypical facial pain, 10.3% have been reported to suffer from either schizophrenic or delusional disorder, and one study reported a 15% prevalence of psychotic disorders in reflex sympathetic dystrophy patients. It is likely that these discrepancies are related to CPP treatment selection criteria and self-selection. It is also possible that some select pain problems, such as reflex sympathetic dystrophy, may not share the same distribution of DSM disorders as other pain problems, such as low back pain. For the diagnosis of psychological factors affecting physical condition, there are major discrepancies between authors. These differences are likely related to reliability problems with the criteria for this diagnosis. These problems are the same as those pointed out for the somatoform disorders.

B. Comorbidity between the Chronic Pain State and Any Psychiatric Disorder on Axis II (Personality Disorders) of the DSM-IV

Comorbidities between chronic pain and Axis II psychiatric disorders are the second best studied group. The results of these studies are summarized in

Table IV
Percentages of Comorbidly Associated DSM Axis II Disorders with Chronic Pain

Psychiatric disorder	Prevalence within chronic pain patients (%)
Paranoid	2.8–33
Histrionic	4–30
Dependent	3–25
Mixed	22.0
Borderline	1–15
Passive–aggressive	12–14.9
Avoidant	14.0
Narcissistic	2.4–10.0
Self-defeating	10.0
Compulsive	6–6.7
Antisocial	5
Schizoid	1–4
Schizotypal	4
Total of samples	31–59.0

Table IV. Here, studies indicated that the prevalence of personality disorders in CPPs may range from 31 to 59%. These rates may be high because of a number of problems related to difficulties in identifying personality disorders when chronic pain is present. This area thus requires further study. In reference to the types of personality disorders comorbidly associated with chronic pain, no consistent trends are evident, and there are discrepancies between authors as to prevalence rates (note the ranges in Table IV). Thus, authors are in conflict as to which personality disorders are most commonly found in CPPs.

C. Comorbidity between the Chronic Pain State and Multiple Psychiatric Disorders on Axis I of the DSM-IV

Psychiatric research has determined that large numbers of psychiatric patients demonstrate more than one Axis I diagnosis, i.e., there is comorbidity between Axis I diagnoses. At issue then is whether chronic pain patients as a group demonstrate this type of comorbidity. Unfortunately, this type of comorbidity has been ignored by chronic pain researchers. There has only been one report on this issue, which reported on the number of Axis I diagnoses by sex within a group of CPPs. For males the distribution was as follows: no diagnoses, 5.7%; one diagnosis, 35.9%; two diagnoses, 33.3%; three diagnoses, 16.0%; and three plus diagnoses, 9.0%. For females, the distribution was as follows: no diagnoses, 4.7%; one diagnosis, 33.9%; two diagnoses, 34.6%, and three or more diagnoses, 7.9%. Therefore, 58.4% of the males and 61.4% of the females *had more than one diagnosis* on Axis I. Therefore, a significant number of CPPs should have Axis I diagnoses comorbidly associated with *other* Axis I diagnoses. The specifics of this comorbidity category have not yet been investigated within CPPs. This category of comorbidity will significantly affect CPP treatment outcome, as has been demonstrated with psychiatric patient groups.

D. Comorbidity between the Chronic Pain State and Psychiatric Disorders on Axis I and Axis II

Comorbidity between Axis I and Axis II (personality disorders) is also frequently found in psychiatric patients. This type of comorbidity is important because it has been demonstrated that it can significantly affect psychiatric treatment outcome. At issue then is whether CPPs as a group demonstrate this type of comorbidity. Unfortunately, this type of comorbidity has also been ignored by pain researchers. There has been only one report that addressed this issue indirectly. In 283 CPPs, 62.3% of the males and 55.1% of the females were assigned an Axis II diagnosis. In the same sample, 94.3% of the males and 95.3% of the females had one or more diagnoses on Axis I. Therefore, it is highly likely that this type of comorbidity is commonly present within CPPs. The exact details of this type of comorbidity have yet to be determined.

E. Comorbidity between the Chronic Pain State and between More Than One Psychoactive Drug Use Disorder on Axis I

Comorbidity between psychoactive drug use disorders on Axis I is also frequently encountered in psychiatric patients. An example of this type of comorbidity is a patient who abuses both alcohol and cocaine. This type of comorbidity is important because it has been shown that it can significantly affect psychiatric pathology in reference to presentation and outcome. At issue then is whether this type of comorbidity is present within CPPs. This type of comorbidity has also

been ignored by pain researchers. However, one study has demonstrated some indirect evidence that points to the presence of this category of comorbidity in CPPs. In this study, two groups of CPPs were compared for the prevalence of Axis I psychoactive drug use disorders: a group positive on urine toxicology giving correct information on current illicit drug use according to urine toxicology and a group positive on urine toxicology giving incorrect information on current illicit drug use according to urine toxicology. The results indicated that the false information group was more likely to have a past history of illicit drug use disorders. This indirect evidence indicates that subgroups of CPPs may suffer from multiple comorbid psychoactive drug use disorders.

F. Comorbidity between the Chronic Pain State and between Psychoactive Drug Use Disorders and Other Axis I Disorders

Comorbidity between psychoactive drug use disorders and other Axis I disorders is frequently found within psychiatric patient groups. This type of comorbidity, e.g., depression and alcohol use, can often complicate treatment and affect treatment outcome. One pain study has addressed this issue. In it, three groups of CPPs were compared for the prevalence of Axis I diagnoses other than psychoactive drug use disorders. The three groups were (1) CPPs with a current alcohol use disorder diagnosis, (2) CPPs with a current drug use disorder diagnosis, and (3) groups 1 and 2 combined. These groups were compared to CPPs without any of the aforementioned alcohol or drug use diagnoses. Some DSM affective and personality disorder diagnoses were found to be more frequently associated with all of the psychoactive drug use disorder groups than with the remaining CPPs. These results indicate that, as for psychiatric populations, this category of comorbidity is to be found in CPPs.

V. BEHAVIORAL COMORBIDITIES COMMONLY ASSOCIATED WITH CHRONIC PAIN

Because of chronic pain and resultant disability, CPPs usually encounter innumerable environmental problems. Some of these are significant enough to cause major stress. These then manifest as a behavioral concern or complaint with resultant anxiety or depression surrounding that issue. Some of these behavioral problems occur frequently enough to be labeled as behavioral comorbidities. These are presented in Table V. The behavioral comorbidities that can be thought of as usually being related to environmental circumstance are labeled nos. 1–7, 12, 13, and 14. Behavioral comorbidities 8–11 can be thought of as being a function of an interaction between the stress that the CPP is experiencing and the CPP's personality structure. Some of these behavioral comorbidities will be discussed next.

The importance of the no diagnosis problem (problem 1, Table V) has already been discussed under nonspecific pain. The vast majority of CPPs, if they are Worker's Compensation (WC) patients, complain of and are adversely stressed by the following stressors. Because of the adversarial nature of the WC system, the vast majority WC CPPs complain of major problems with their disability carriers (Table V). These have been noted by the author to be the following: (1) compensation checks always arriving late; (2) compensation benefits terminated for no reason, according to the CPP; (3) denial of medical care (doctors, medical procedures, or medical supplies) (Table V, 14) that the CPP has requested, wants, or has been recommended by his or her physician; (4) forcing the CPP to go to physicians he or she does not wish to go to; (5) forcing the patient to do a job search when the patient feels he or she is incapable of going back to work; (6) patient unable to reach adjustor; and (7) patient unable to reach an equitable settlement with the carrier. It is to be noted that these problems are the CPP's perceptions, which may be colored by the litigation process. From the nature of these complaints, however, it is clear that for a segment of the WC CPP population, the relationship with his or her carrier is a very stressful one. The author believes that when there is associated litigation (Table V, 3) between the WC patient and the carrier, the stress is even greater. As an indication of the stressful relationship between the carrier and the WC patient, it is not unusual to hear that the patient retained a lawyer because of the perception that he or she was being treated badly by the carrier. Occasionally, these perceptions can interact with paranoid tendencies of the CPP, or this situation can occur to a CPP who has been habitually violent. Under these circumstances, the CPP might begin to harbor fantasies (Table V, 17) about getting even in a violent fashion with the insurance carrier or the "insurance physicians." If there are such fantasies or wishes, the CPP should immediately be referred to a psychiatrist. It is interesting to note here that a multicenter CPP

Table V
Behavioral Comorbidities

(1)	Does not know or understand reason for chronic pain? Has no diagnosis? Conflicting diagnostic opinions?
(2)	Disability carrier problems and stress
(3)	Litigation stress
(4)	Preinjury job stress
(5)	Postinjury involuntary job loss
(6)	Perception of lack of support from spouse
(7)	Financial difficulties
(8)	Guilt over not being able to perform spouse and/or parental functions
(9)	Unrealistic expectations of recovery or function, e.g., ability to weight-lift
(10)	Coping strategies
(11)	Anger–irritability
(12)	Spouse solicitousness
(13)	Perception of employer negligence or other type of negligence for the injury
(14)	Dissatisfaction with medical care
(15)	Fear of pain
(16)	Self-efficacy
(17)	Violence ideation and risk
(18)	Suicidal ideation and risk

study has found that the presence of violent ideation is greater within CPPs than in general medical controls.

Perception of employer negligence or other types of negligence such as reckless driving as a cause of the CPP's pain is a common behavioral comorbidity (Table V, 13). This situation is typified by the patient who has warned his or her superior of a potentially dangerous situation at work or that a job that he or she is being asked to do is too hard or too dangerous, or both. As an example, a WC patient tells his boss that the object that he is being asked to lift is too heavy. He is told to lift this object or he will lose his job. The worker lifts the object and develops LBP, which later becomes chronic. The CPP then develops much anger at his employer and considers some form of litigation. He would then rationalize the litigation with the wish to "get even" with the employer for his or her apparent misconduct in causing the accident.

Often CPPs complain bitterly that, after years of service at a company, he or she was fired soon after the injury (Table V, 5). Thus, many CPPs are subject to involuntary job loss. Because work serves many adaptive functions, involuntary job loss can result in various degrees of loss of self-esteem and resultant depression. In addition, people who are not psychologically healthy may not react to the trauma of job loss in an adaptive fashion. Another issue in involuntary job loss is the loss of retirement benefits. It is not unusual for companies to terminate the CPP after an injury but a short time before the CPP is vested in the plan. Such a situation creates great anger and anxiety in the CPP, raising the issue of monetary gain on the part of the company. In some situations, the CPP's anger is channeled into the litigation process, and the CPP may even admit that he or she is seeking revenge in these actions. CPPs may also complain of anxiety about performing adequately in a new job because of their chronic pain. Finally, many CPPs are often fearful that they will never be able to get another good job because they will forever be labeled a back patient, or someone who has been on WC, or both. This is a correct perception. Often CPPs are denied employment because of their previous injuries.

The perception of preinjury job stress is also often noted to be a common behavioral complaint in CPPs (Table V, 4). This behavioral comorbidity was investigated in a series of four studies for its impact on CPP return-to-work behavior. In the first report it was demonstrated that CPPs not intending to return to work after pain facility treatment were more likely to

complain of job dissatisfaction at their preinjury job. In the second report, an association between nonintent to return to work after pain facility treatment and preinjury job dissatisfaction was similarly found across worker's compensation and non-worker's compensation chronic pain patients. In the third report, it was found that actual return to work was predicted at 1 month after pain facility treatment by intent, perceived job stress, and job like (job dissatisfaction) plus other variables. At 36 months, return to work was predicted by intent and by perceived job stress plus other variables. The final study attempted to predict actual return to work after pain facility treatment in relation to intent to return to work. Intent was predicted by perceived preinjury job stress plus other variables. In addition, those CPPs who intended to return and did not were predicted by whether there was a job to go back to. Also, CPPs not intending to go back to work at the preinjury job but doing so were predicted by having a job to go back to. Overall, this series of studies indicates that there is a strong relationship between preinjury work variables such as job dissatisfaction and intent to return to that job posttreatment. It appears that, in treating the CPP, it is important to understand work-related perceptions such as those of perceived job dissatisfaction and job stress.

Irritability–anger is another common behavioral complaint encountered in CPPs (Table V, 11) but has not been well-studied. Commonly, CPPs verbalize anger against the causal agent of the injury–illness, medical health care providers, attorneys and the legal system, insurance companies, the social security system, and employers but rarely against a significant other, God, self, and the whole world. Anger may represent one of the affective components of pain and may predict perceived pain interference and activity level. Anger expression among males was found to correlate negatively with lifting capacity improvement. Anger may be a function of anxiety and therefore may respond to relaxation coping skills intervention and, if necessary, medication.

The perception of unrealistic expectations for treatment for chronic pain is also often encountered as a behavior comorbidity in CPPs (Table V, 9). This unrealistic expectation applies to surgical treatment and/or pain facility treatment. Often, the CPP wishes to be the "same" as before the injury, i.e., to be able to do all activities including those that may be contraindicated for patients with chronic LBP, e.g., competitive weight-lifting. In the quest to be the same, the CPP will seek more aggressive–risky treatments as the less risky treatments fail. In this quest there is often physician collusion. Rather than provide realistic expectations about a treatment for chronic pain, the physician may provide inflated claims for the possible success of his or her treatment. Thus, some authors have noted that clinicians should be cautions in the prognostic information they give their patients as generally optimistic prognoses may create disappointment and anger. It is important to provide the CPP with information about the limits of the treatment and what can be reasonably expected.

Many CPPs complain or show a significant fear of increased pain (Table V, 15) and thereby fear of any stimuli, such as exercise, that could increase their pain. These fears may be related to measures of catastrophizing and depression and may characterize dysfunctional CPPs. This construct is important because these fears can affect pain facility treatment outcome and should be the focus of treatment efforts if they appear to interfere with treatment.

There has been some dissatisfaction with traditional psychological assessment tools for the behavioral assessment of the CPP, e.g., Minnesota Multiple Personality Inventory. Because of this issue and the development of the Cognitive Behavioral Model of Pain, there has been an attempt to develop new inventories that may be more applicable to the CPP. These new inventories attempt to tap three major concepts: patient pain beliefs (discussed earlier), self-efficacy expectations, and coping strategies. Self-efficacy expectations are beliefs regarding one's capacity to execute the behavior required to produce a certain outcome. Coping strategies are the process of executing a cognitive response to threat. These three concepts are important because there is some evidence that CPPs have some incorrect perceptions regarding their pain and that CPPs' coping behavior is dependent upon judgments regarding their ability to perform that behavior. Finally, training in the use of CPP coping strategies decreased reported pain and increased pain tolerance and threshold. Thus, these three concepts address problem areas where treatment interventions could be initiated.

The behavioral comorbidities of perception of spousal lack of support, guilt over not being able to perform marital and/or parental functions, and alleged spouse over-solicitousness to the CPP's pain behavior are all interrelated. Early in the development of the pain behavioral literature, it was postulated that a solicitous spouse would serve as a reinforcer to continued painful complaints, thus satisfying a secondary gain wish on the part of the CPP. There is some evidence for this assertion. These two approaches,

however, took attention away from the fact that spouses are often much affected by the illness of the CPP. There appears to be significant correlation between the psychiatric distress scores of the CPPs and their spouses, especially when CP is high. Evidence indicates that 28% of CPP spouses report significant depressed mood. This depression is predicted by the CPP's average pain, his or her level of anger and hostility, and the spouse's level of marital satisfaction. The CPP family is prone to two psychological tensions as result of the CPP's pain: uncertainty about the CPP's diagnosis and the authenticity of the pain. This situation arises because for most CPPs the diagnosis is elusive, i.e., nonspecific. Spouse or family perception of nonauthenticity can lead to CPP perception of marital nonsupport and thereby lead to marital–family conflict. This in turn can affect the CPP's pain.

Suicidal ideation and/or attempts are not infrequently encountered with CPPs. Because chronic physical illness is associated with depression, one would expect chronic physical illnesses such as chronic pain to be associated with suicide and/or suicide ideation. Indeed, patients with chronic migraines are at increased risk for suicide attempts. Similarly, the suicide rate for chronic LBP and neck pain populations is twice that of the general population. Therefore, CPPs are at risk for suicidal ideation and suicide attempts–completions. Any evidence of suicide ideation should immediately trigger a request for a psychiatric consultation.

VI. OTHER COMORBIDITIES COMMONLY ASSOCIATED WITH CHRONIC PAIN

There are a number of other comborbidities that are commonly associated with chronic pain. These are listed in Table VI. They are classified as other comorbidities because they do not translate into the DSM psychiatric diagnostic nomenclature and, although some of them appear to be behavior complaints, nevertheless they are more somatic in nature. These other comorbidities dramatically affect the difficulties involved in the care and treatment of the CPP. Usually, treatment plans cannot be developed without taking these problems into account. Each of these other comorbidities will be elaborated on here.

Sleep problems are a universal CPP problem. The vast majority complain of severe sleep disturbance. The sleep problems are usually a result of pain and may be directly associated with pain intensity. Although the CPP can usually fall asleep, he or she is frequently aroused from sleep by pain. Thus, the sleep is fragmented and of poor quality. These sleep disturbances are not a function of depressed mood.

Table VI
Other Comorbidities Commonly Associated with Chronic Pain

(1) Sleep problems
(2) Sexual problems
(3) Fatigue
(4) Somatization
(5) Nonorganic physical findings
(6) Perception of disability
(7) Pain behavior
(8) Memory and concention
(9) Neuropathic pain component
(10) Impaired functional status out of proportion to physical findings
(11) Perception of secondary gain issues
(12) Possibility of malingering

Sexual dysfunction is also a frequent complaint of the CPP. It has been claimed that up to 63% of men complaining of low back pain were found to be impotent. However, according to the author's experience, very few male CPPs are actually truly impotent. The usual complaint is that sexual intercourse is painful, and because of this the CPP is unable to maintain an erection. This serves as a disincentive to further sexual intercourse, which in turn results in decreased interest in sex. The decreased interest in sex can in turn lead to marital difficulties. Complaints of sexual dysfunction should be investigated, and those CPPs whose complaints actually indicate a possibility of impotence should be referred to a urologist.

Fatigue (Table VI, 3) is not an infrequent complaint in CPPs. However, fatigue has not been well-studied. To the author's knowledge there have been only two studies in this area. Both studies indicate that fatigue is commonly found in CPPs and that pain predicts fatigue. Because chronic fatigue syndrome overlaps with fibromyalgia, i.e., most chronic fatigue patients will have chronic pain, it is to be expected that fatigue should be a frequent problem in CPPs.

Somatization may be another other comorbidity that may be a significant problem within CPPs, and CPPs with this problem may be at risk for poor treatment outcome. Somatization has been defined as "a tendency to experience and communicate somatic distress and symptoms which are unaccounted for by

pathological findings, and to attribute them to physical illness, and to seek medical help for them. It is usually assumed that this tendency becomes manifested in response to psychosocial stress brought about by life events and situations personally stressful to the individual'' (Fishbain *et al.*, 1998). Somatization does not represent a specific DSM psychiatric or medical diagnosis and does not necessarily imply that a psychiatric disorder must be present. Because somatization is not a specific diagnosis, it does not have operational criteria by which it can be established. Thus, in CPPs somatization has been studied by questionnaires. These studies indicate that a high percentage of CPPs demonstrate elevated hypochondriasis scores and somatization scores. In addition, when patients with various types of chronic pain are compared to appropriate controls on somatization measures, CPPs frequently have greater somatization scores. This has been demonstrated for CPPs with orofacial pain, migraines, noncardiac chest pain, chronic low back pain, and fibromyalgia. Finally, somatization scores appear to be predictive for treatment outcome in CPPs with temporomandibular disorders and low back pain.

Another common other comorbidity found in CPPs has been termed pain behaviors. CPPs can be characterized as displaying pain behaviors. Pain behaviors are ''any and all outputs of the patient that a reasonable observer would characterize as suggesting pain, such as (but not limited to) posture, facial expression, verbalizing, pain expressions, taking medications, seeking medical assistance and receiving compensation'' (Fishbain *et al.*, 1998). Pain behaviors can often be elicited during physical examination and correlate with physical examination findings, number of operations, and longer pain histories. In addition, pain behaviors correlate with perceived severity of pain and extent of functional impairments. As pain improves, pain behavior diminishes.

Another other comorbidity often demonstrated by CPPs is that of nonorganic physical findings. These are eight physical examination signs found in CPPs with lower back or neck pain. Earlier in this century, these signs were identified as predominantly nonorganic. A large percentage of patients with chronic lower back pain demonstrate these eight nonorganic physical signs. These signs appear to be predictive of treatment outcome, and it has been suggested that their presence or absence should be used as a basis for surgical decisions. Most importantly, these signs correlate with a greater degree of pain behavior and, therefore, a more difficult treatment problem. In addition, the presence of one or more of these signs makes that patient a candidate for the psychiatric diagnosis of conversion disorder, a somatoform disorder. It has been stated that these signs are not a test of credibility or faking but represent a ''psychological yellow flag,'' indicating that psychological factors need to be considered.

CPPs often perceive themselves as disabled. The perception of disability was postulated to be the reason for failure of rehabilitation with some CPPs. However, it has been found that the perceived disability is a result of the pain. The perception of disability as a factor in chronic pain maintenance has not been well-studied.

Concentration and memory problems are another class of other comorbidities often encountered in CPPs. Generally, these patients complain of significant pain and demonstrate severe pain behavior. One study demonstrated that these problems were related to depression, nervousness, family conflicts, dissatisfaction with sexual activities, pain interfering with recreation, irritability, and worse pain when sitting.

At the present time, increasing amounts of evidence indicate that neuropathic pains may be very commonly found in CPPs. Neuropathic pain encompasses a diverse group of syndromes characterized by an injury to the nervous system. The resultant neuropathic pain has several distinctive clinical characteristics: localization in an area of sensory deficit; a burning and/or shooting quality with unusual tingling, crawling, or electrical sensations (dysesthesias); and gentle mechanical stimuli such as bending of hairs may evoke severe pains (allodynia). In some of these CPPs there may be involvement of the sympathetic nervous system called sympathetically maintained pain. Here it has been found that sympathetic activity can sensitize nonceptors following nerve damage. In addition, it has become clear that there may be different subtypes of neuropathic pain, which may have different underlying etiologies. Because myofascial syndrome CPPs often describe their pain as burning, this large group of CPPs could also have an element of neuropathic pain. At the present time this issue has not been researched. There is also some evidence in the literature that CPPs with a neuropathic pain syndrome have significant psychiatric comorbidity, including personality disorders, although not at a greater frequency than other groups of CPPs.

CPPs often demonstrate discrepancies between reported pain level and functional status versus the physician's perception of what the CPP should be able to do functionally according to the physical findings

(Table VI, 10). This perception discrepancy is attributable to a number of complex problems: CPPs perceive their pain as a disability that limits their functional status; the pathology model does not predict back pain, making the reliability and validity of this model questionable; and for a proportion of CPPs nonverbal expression may be discordant with self-reports. The resultant discord in perceptions between CPPs and their physician often leads to the labeling of the CPP as having psychiatric problems, having secondary gain (discussed next), or being a malingerer (discussed later). The key here is for the examining physician to understand that this problem is one of the other comorbidities in CP and that this problem is the *norm* for CPPs. Under these circumstances, it would be expected that fewer of these CPPs would be characterized as having psychiatric problems, having secondary gain, or being a malingerer.

Secondary gain has been defined as "acceptable or legitimate" interpersonal advantages that result when one avoids an activity noxious to him or her or gets support from the environment not otherwise forthcoming. As can be surmised from this definition, secondary gain can encompass many situations in which the patient can receive some gain. Examples of these situations are the following: fulfillment of dependency needs; prevention of desertion by the spouse; desire to get away from a bad job situation; desire to be retired before the injury; increased attention; avoidance of coercive activities (e.g., sexual intercourse); increased self-esteem; increased attachment needs; aggressive gratification (revenge); dissatisfaction with impairment rating before settlement (believes should have a 100% impairment rating); dissatisfaction with settlement terms; desire for retraining or to go back to school (had wanted to increase educational status before the injury); decrease in home–family responsibility; and so on. Secondary gains occur by unconscious mechanisms. It is, therefore, not clear whether desire for financial compensation should be classified along with the other secondary gains. Yet most CPPs who are involved in litigation, may or may not have excessive pain behavior, have significant functional impairment out of proportion to the physical findings, and have nonorganic findings are usually perceived as having secondary gains. This may or may not be the case, which is what the pain physician should keep in mind. Being involved in litigation does not necessarily mean that one has with a secondary gain agenda.

The final other CPP comorbidity is the likelihood of being suspected of malingering. This occurs because of the following other comorbidities described earlier; (1) the organic diagnosis is often elusive or nonexistent, i.e., nonspecific; (2) there is excessive disability versus the medical impairment; (3) nonorganic physical findings are present; (4) CPPs are usually financially dependent on compensation or disability benefits; (5) CPPs are usually involved in litigation; and (6) CPPs may have other secondary gain issues besides number 4 and 5. As such, CPPs are at risk for the accusation that they are malingering their complaints. Malingering is not classified as a mental disorder. The DSM-III-R states that malingering is an "act" and classifies malingering under the V codes that are included for a supplementary classification of factors influencing health status and contact with health services. The DSM-III-R suggests that the possibility of malingering should not be entertained unless there is evidence that physical symptoms are intentionally produced. Rarely is there strong evidence or any evidence for this conclusion. The Institute of Medicine report on pain and disability has also concluded that malingering is rare. The presence of secondary gain, or even monetary gain, is not evidence of malingering. Although malingering is often discussed in reference to CPPs, there are none or few studies of this problem within the chronic pain literature. Pain physicians should not consider this label unless they have definite proof, such as that obtained via covert observation.

VII. IMPORTANCE OF PSYCHIATRIC, BEHAVIORAL, AND OTHER COMORBIDITES TO CHRONIC PAIN TREATMENT

What is the importance of different types of psychiatric comorbidities to chronic pain? Research indicates that psychiatric comorbidities often impact negatively on chronic pain and functional status. As an example, anxiety can change pain threshold and tolerance and increase pain ratings. Comorbidly depressed CPPs demonstrate more automatic hyperactivity (muscle tension) and may be more sensitive to acute pain stimuli. In addition, evidence indicates that depression may magnify medical symptoms. Depression and phobias have also been reported to be associated with role impairment and loss of function. For example, individuals with subsyndrome depressive symptoms of major depression report significantly more impairment in 8–10 functional domains than individuals without subsyndrome depressive symptoms. Comorbidities between Axis I psychiatric

disorders in association with chronic pain can also significantly affect chronic pain treatment and outcome. This is because each of the comorbid Axis I disorders causes more severe symptoms of the *other* Axis I disorders. Because a high percentage of CPPs have more than one comorbid Axis I disorder, they are at risk for this problem. Comorbidities within CPPs between Axis I and Axis II psychiatric disorders can also affect the treatment of CPPs. For example, atypical depression (the type seen in pain facilities) is more frequently associated with personality disorders, making this type of depression more difficult to treat. This type of comorbidity could, therefore, also affect chronic pain treatment. Finally, conceptualization of pain-associated psychiatric comorbidity in terms of the preceding types of psychiatric comorbidities can help pain patient treatment decisions.

Behavioral comorbidities are also important because they can complicate treatment, as most of them are a source of additional stress. As such, these comorbidities can become the focus of psychological intervention. It is most important to understand and remember that these comorbidities should be present in most CPPs.

The other comorbidities characterize the CPP. This is their importance. The understanding that most CPPs will have these other comorbidities can potentially make diagnosis and treatment easier. Their impact on treatment outcome by their presence cannot be overstated.

See Also the Following Articles

ANXIETY • CONVERSION DISORDERS AND SOMATOFORM DISORDERS • DEPRESSION • MIGRAINE • MODELING BRAIN INJURY/TRAUMA • MOOD DISORDERS • NEUROPSYCHOLOGICAL ASSESSMENT • OPIATES • PAIN • STRESS

Suggested Reading

Abenhaim, L., Rossignol, M., and Gobeille, D. (1995). The prognostic consequences in the making of the initial medical diagnosis of work-related injuries. *Spine* **20,** 791–795.

Cedraschi, C., Nordin, M., Nachemson, A. L., and Vischer, T. L. (1998). Health care providers should use a common language in relation to low back pain patients. *Bailliere's Clin. Rheumatol.* **12**(1), 1–13.

Fishbain, D. A., Goldberg, M., Steele-Rosomoff, R., and Rosomoff, H. (1991). Completed suicide in chronic pain. *Clin. J. Pain* **7,** 29–36.

Fishbain, D. A., Cutler, R. B., Steele-Rosomoff, R., and Rosomoff, H. L. (1994). The problem oriented psychiatric examination of the chronic pain patient and its application to the litigation consultation. *Clin. J. Pain* **10,** 38–51.

Fishbain, D. A. (1995). DSM–IV: Implications and issues for the pain clinician. *Am. Pain Soc. Bull.* 6–18.

Fishbain, D. A., Rosomoff, H. L., Cutler, R. B., and Steele-Rosomoff, R. (1995). Secondary gain concept: A review of the scientific evidence. *Clin. J. Pain* **11,** 6–21.

Fishbain, D. A., Cutler, R., Rosomoff, H. L., and Rosomoff, R. S. (1997). Chronic pain-associated depression: Antecedent or consequence of chronic pain. *A review. Clin. J. Pain* **13**(2), 116–137.

Fishbain, D. A., Cutler, R. B., Rosomoff, H. L., and Rosomoff, R. S. (1998). Comorbidity between psychiatric disorders and chronic pain. *Curr. Rev. Pain* **2,** 1–10.

Fishbain, D. A., Cutler, B., Rosomoff, H., and Steele-Rosomoff, R. (1999). Chronic pain disability exaggeration/malingering and submaximal effort research. *Clin. J. Pain* **15,** 244–274.

Main, C. J., and Spanswick, C. C. (1995). Functional overlay, and illness behavior in chronic pain: Distress or malingering? Conceptual difficulties in medico-legal assessment of personal injury claim. *J. Psychosomat. Res.* **39,** 737–753.

Main, C. J., and Waddell, G. (1998). A reappraisal of the interpretation of "non-organic signs". *Spine* **23**(21), 2367–2371.

Rosomoff, H. L., Fishbain, D. A., Goldberg, M., Santana, R., and Rosomoff, R. S. (1989). Physical findings in patients with chronic intractable benign pain of the neck and/or back. *Pain* **37,** 279–287.

Weisberg, J. N., and Keefe, F. J. (1997). Personality disorders in the chronic pain population: Basic concepts, empirical findings, and clinical implications. *Pain Forum* **6**(1), 19.

Pain

A. D. CRAIG
Barrow Neurological Institute, Phoenix

GLOSSARY

ATP Adenosine triphosphate.

bradykinin An agent released from mast cells that activates and sensitizes pain receptors.

homeostasis The general process that maintains optimal conditions for survival.

hyperalgesia State in which increased pain is perceived in response to noxious stimuli.

sensitization Enhanced sensitivity due to peripheral or central changes in sensory processing.

Pain is a distinct, multidimensional sensory experience that is a common clinical complaint but is essential for survival. This article describes the characteristics of acute and chronic pain, the specialized neural elements in the periphery, spinal cord, and brain that subserve pain or that modulate pain, the roles of various neurochemicals, and the changes induced by damage or disease.

I. AN OVERVIEW

Pain is essential for the maintenance and survival of an individual. Pain is a somatic distress signal. It warns of the danger of bodily harm, it alerts to trauma and injury, and it signals unhealthy conditions or inflammation in tissues of the body. The accepted definition of pain is as follows: "an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage" (International Association for the Study of Pain). The alleviation or absence of pain is analgesia. Individuals with congenital insensitivity to noxious stimuli or bodily damage (who lack the requisite small-diameter peripheral nerve fibers) disregard injuries and infections and normally do not survive everyday life.

Pain is integrated within the context of current physiological and environmental conditions, together with past experience and future plans. Aristotle taught that pain is an emotion, but in the late 1800s physiologists began to recognize pain as a specific, discriminative sensation. It can be dissociated from other modalities of somatic sensation (for example, by anterolateral cordotomy), and, though subjective, it can be cognitively evaluated and reproducibly localized, scaled, and timed. Different types of pain can be distinguished qualitatively, for example, burning, stinging, aching, throbbing, or cramping, which reflect subpopulations of receptors and the tissue of origin.

Pain is often thought of as an exteroceptive, discriminative sense, like touch, yet it can originate from most tissues of the body, of which skin is but one vital organ. In addition, like all feelings from the body and in contrast to exteroceptive (touch) and teloreceptive (vision, hearing) modalities, the sensation of pain is innately endowed with a distinct unpleasantness, i.e., a characteristic affect that motivates behavior, which can be distinguished from its discriminative sensory aspect and also from the

long-term psychological experience called "suffering." Pain reflexly generates autonomic responses that signal its primal role in homeostasis (the processes underlying the maintenance of the body). Thus, pain can be regarded as a specific, emergent aspect of interoception, the sense of the physiological condition of the body itself, or, as Sir Charles Sherrington put it in 1900, a feeling from "the material me." The level of unpleasantness that is regarded as painful varies in different individuals (and in different contexts and cultures) and extends across a range of intensities in all individuals; think of the gnawing discomfort of a cold room, the burn of fatigued muscles, the stress of colonic tension, or the sting of a blister and compare these feelings to the excruciating agony of a smashed finger, a severe burn, a deep wound, a colonic spasm, a toothache, a kidney stone, or a migraine headache. The peripheral receptors that cause pain sensation are sensitive to changes in a variety of different tissue conditions (mechanical, thermal, and chemical), and, under normal circumstances, the fibers that report the physiological status of the tissues of the body to the central nervous system engage systemic survival mechanisms at several functional levels, i.e., autonomic, homeostatic, motoric, neuroendocrinologic, motivational, behavioral, and mnemonic. These responses must be integrated with ongoing homeostasis and behavior to maintain the health of the individual's body.

Injury or disease affecting either the peripheral or central nervous system can alter the balance of these integrative systems, and, in some cases, persistent, pathological pain results. There is a great need for the development of better methods for the alleviation and control of both acute, traumatic pain and chronic, long-term pain. Analgesic substances, for instance aspirin or morphine, that interact with the endogenous transmitters and modulators of the pain processing system are helpful for many people with pain, but not all and not always. Acute pain is signaled by specialized pathways that command immediate attention, but chronic pain can involve chemical, physiological, and even anatomical changes in the underlying substrates and these changes can produce a neuropathological state. In some patients with intractable pain, pharmaceutical agents administered orally, intravenously, or intrathecally (in the spinal canal) can produce analgesia. In others, deep brain stimulation or neurosurgical intervention is more effective, and in all cases behavioral management is an important therapeutic adjunct.

The following paragraphs describe the specialized neurons in the peripheral nervous system (primary receptors) and the central nervous system that subserve pain, the interactions with other somatic sensory pathways and with homeostatic processing systems, the neurochemicals associated with these elements, and, finally, the significant changes in these systems that can be induced by injury or disease.

II. PERIPHERAL RECEPTORS

Peripheral receptors localized in nearly all tissues of the body transmit their activity to the spinal cord (and its trigeminal equivalent in the caudal medulla) by way of axons that course in the sensory nerves. These primary afferent fibers have cell bodies in the dorsal root ganglia, and their central terminals enter the spinal cord by way of the dorsal roots. Low-threshold mechanoreceptors and proprioceptors (responsible for touch and movement sensations) generally have large-diameter fibers and large cell bodies, whereas high-threshold receptors (involved in pain sensations) and thermoreceptors (responsible for temperature sensation) generally have small-diameter fibers and smaller cell bodies. These include small myelinated (Aδ or Group III) and unmyelinated fibers (C or Group IV) that conduct slowly (approximately 3–30 and 0.5–3 m/sec, respectively). There is considerable overlap between these categories, however. There are also many C-fibers that innervate distinct low-threshold mechanoreceptors or histamine-sensitive chemoreceptors (that may subserve tickle or itch, respectively). The small dorsal root ganglion cells develop in a later phase than the larger ones and are histologically differentiable as dark neurons.

Receptors that respond selectively to noxious or potentially damaging stimuli are called nociceptors (a term introduced by Sherrington). The receptive elements of nociceptors are free nerve endings without accessory structural specializations, albeit ultrastructural studies suggest that Aδ and C nociceptive endings may be differentiable in fine details. In skin, the Aδ-nociceptors innervate one or a few neighboring spots, whereas most C-fiber nociceptors are sensitive over a 1–4 mm^2 receptive field. Subtypes have been differentiated physiologically in skin and in various somatic tissues (e.g., muscle, joint, gallbladder, colon, genitalia, esophagus, tooth pulp). In general, most Aδ-nociceptors respond only to noxious mechanical stimulation (e.g., pinch) or to both noxious mechanical and noxious heat stimuli, whereas a few respond only to noxious heat. Many C-nociceptors are similarly submodality-selective, but most are polymodal,

responsive to several forms of noxious mechanical, thermal, and chemical stimulation (apparently by separate biophysical transducers). Polymodal C-nociceptors are also responsive to chemical (or metabolic) stimuli, such as acidic interstitial pH, bradykinin, serotonin, ATP, and hypoxia or hypoglycemia. The Aδ-nociceptors in muscle can be activated by pinch though some are activated by contraction, and the C-nociceptors in muscle can be activated by static exercise, pinch, heat, algogenic chemicals, or ischemia.

Nociceptors have been characterized by peripheral nerve recordings not only in animals but also in humans using a technique called percutaneous microneurography, in which a microelectrode is inserted into a peripheral nerve in awake volunteers. Electrical microstimulation during microneurographic recordings supports the specific sensory roles of nociceptive fibers; human volunteers report sharp pain upon stimulation at Aδ nociceptive fiber recording sites, burning pain from stimulation at C-fiber recording sites, and aching or cramping pain upon the stimulation of C-nociceptors innervating muscle, but no pain upon stimulation near low-threshold mechanoreceptive fibers. Blockade of peripheral C-fiber conduction with lidocaine eliminates burning pain sensation. A pressure block of peripheral nerve conduction that blocks Aδ-fiber conduction eliminates sharp pain sensation (and touch and temperature sensations), yet pinch or cold still causes a painful burning sensation. The rapid stinging or pricking sensation associated with Aδ-nociceptors is called first pain, and the slower, burning sensation (or from deep tissues, the dull, aching sensation) associated with activity in C-nociceptors is called second pain.

The mean threshold for mechanical cutaneous pain is approximately 10 gm of force per mm of circumference of a flat probe. The mean heat and cold cutaneous pain thresholds are approximately 45 and 15°C, respectively. The mean activation thresholds and the incremental detection thresholds of nociceptors correspond to human psychophysical sensory measurements; nonetheless, the thresholds for activation of nociceptors in all tissues extend over a broad range encompassing these stimulus intensities. Many are responsive only to overtly damaging stimuli, and some are insensitive (silent) and respond only subsequent to inflammation or actual tissue damage. The discharge activity of C-nociceptors is slow (few fire faster than 20 Hz), and microneurographic stimulation results show that temporal summation of C-fiber activity is required to produce a perception of pain. (Also, C-nociceptors from muscle or joint can have ongoing discharge, consistent with a role in ongoing homeostasis below perceptual threshold.) Above threshold, the intensity of perceived pain is directly related to the number of action potentials that nociceptors fire or the frequency of microneurographic stimulation. Maintained noxious mechanical or heat stimuli, however, generate increasing pain sensations that must result from central summation, because the responses of the peripheral nociceptors to maintained tonic stimuli adapt to low firing levels.

III. PAIN PROCESSING IN THE SPINAL CORD AND BRAIN

The central axons of small-diameter primary afferent fibers enter the spinal cord through the lateral portion of the dorsal roots. Their collateral branches reach 1–3 segments longitudinally in Lissauer's tract. Within the superficial dorsal horn of the spinal cord or the equivalent trigeminal nucleus caudalis in the caudal medulla, cutaneous C nociceptive fibers arborize and terminate in lamina I and outer lamina II, and cutaneous Aδ nociceptive fibers terminate in lamina I as well as deeper in lamina V (the neck of the dorsal horn). Nociceptive afferents from muscle, joint, and viscera terminate primarily in lamina I, but a few also extend into laminae II, V, and X (near the central canal).

Nociceptive cells in laminae I and V are the main second-order pain projection neurons that send ascending axons to the brain. Nociceptive neurons in lamina I have little or no ongoing activity, and they respond specifically to noxious stimuli within a small receptive field (in skin, muscle or viscera). Two distinct functional types of nociceptive cells are present in lamina I, one dominated by Aδ input (fusiform nociceptive-specific cells), whose discharges closely parallel first pain sensation, and the other dominated by C-fiber input (multipolar polymodal nociceptive cells), whose discharges closely parallel second pain sensation. Other lamina I cells respond specifically to innocuous thermal stimuli (cooling or warming) or to histamine application (which causes the sensation of itch). The different physiological types of lamina I cells are also morphologically and pharmacologically distinguishable. The nociceptive neurons in lamina V have large receptive fields and considerable ongoing discharge, and they are multireceptive because they respond to both nonnoxious and noxious mechanical stimuli (often called wide dynamic range or WDR cells). Subclasses of lamina V cells can be distinguished

that are sensitive primarily to low-, medium-, or high-threshold stimulation, but as a population their activity is directly related to the overall intensity or the cumulative integration of afferent activity from all somatic tissues. A few WDR cells can also be found in lamina I. Different neurons in each anatomical location respond to A-fibers only or to both A- and C-fibers (monosynaptically in lamina I, polysynaptically in lamina V). Both lamina I cells and lamina V cells encode the intensity of noxious stimulation with the frequency of their discharge, albeit in a modality-selective and modality-nonselective manner, respectively. Laminae I and V neurons that are activated by noxious stimuli can be observed immunohistochemically by labeling for the immediate early gene *c-fos*; such labeling in lamina I seems to reflect numerically the intensity of graded noxious heat stimulation. The discharge activity of a subset of trigeminal WDR cells in laminae I and V has been closely correlated with the operant detection of intensity changes in noxious heat stimuli by behaviorally trained monkeys. Nociceptive information is conveyed to spinal motoric systems mainly by lamina V cells and to spinal autonomic systems mainly by lamina I cells. Nociceptive cells that respond to visceral stimulation also usually have a cutaneous receptive field. This convergence may be the basis for referred pain, in which activity originating from nociceptors in deep tissues is perceived as arising from (and induces hyperalgesia in) a cutaneous zone represented in the same spinal segments. For instance, in angina pectoris pain is referred to the left shoulder, and pain and hyperalgesia are referred to different portions of the face and head with a toothache in different teeth.

The small neurons in lamina II (the substantia gelatinosa or SG) that receive C-fiber inputs are generally interneurons whose processes extend rostrocaudally at most 2–3 spinal segments. Most of these are inhibitory neurons with varicose dendrites and short or even no axons (Golgi type II neurons) that contain GABA, about half of which also contain glycine. Subclasses of SG cells can be distinguished immunohistochemically by labeling for somatostatin, enkephalin, substance P (SP), nitric oxide synthase, acetylcholine, and many other modulators, as well as particular intracellular messengers such as protein kinase Cγ (PKCγ). At the ultrastructural level, C-fiber terminals in SG are characterized by distinct glomerular structures containing pre- and postsynaptic elements that can be differentially labeled for various peptides, amino acids, or receptor subunits (such as NK1r, mGluR1, or gephyrin). However, the functional, anatomical, and neurochemical organization of the superficial dorsal horn, and particularly the SG, is still a profound mystery that needs to be elucidated in order to understand pain processing at the spinal level.

The ascending projections of laminae I and V spinal nociceptive neurons terminate in several hierarchically organized regions that reflect the various reactions to a painful stimulus. Within the spinal cord, lamina I (and some lamina V) cells project to the sympathetic preganglionic regions in the thoracolumbar cord and to parasympathetic regions in the sacral spinal cord and at the spinomedullary junction. In the lower brain stem, lamina I axons terminate bilaterally in several preautonomic sites, especially the major catecholamine cell groups in the ventrolateral medulla and the dorsolateral pons. These projections provide a basis for modality-selective somatoautonomic reflexes in cardiorespiratory function caused by noxious or thermal stimuli or by adverse tissue metabolic changes. Lamina V axons terminate diffusely in the reticular core of the brain stem and may affect somatomotor integration and behavioral state. There are dense projections from lamina I cells (and a few lamina V cells) to the parabrachial nucleus at the pontomesencephalic junction, which also receives vagal afferent input by way of the solitary nucleus; the parabrachial nucleus is a major viscerosensory (homeostatic) integration site that is interconnected with the periaqueductal gray, hypothalamus, amygdala, and homeostatic cortical regions. The lamina I spinoparabrachial projection provides an indirect spinal pathway to the central amygdala that appears to be involved with fear-associated conditioning by nociceptive activity and probably also with cardiovascular responses and the initiation of opiate-related descending antinociceptive activity. There is a similar spinal projection to the periaqueductal gray, which is the major mesencephalic site for homeostasis, defense reactions, and vocalization. The hypothalamus receives ascending nociceptive activity by way of the parabrachial nucleus, by way of the noradrenergic cells in the caudal ventrolateral medulla, and perhaps also by a direct input from spinohypothalamic neurons. These projections may affect goal-directed aversive and feeding behaviors and neuroendocrine and immune responses to homeostatic or nociceptive activity, as well as thermoregulatory and osmoregulatory control.

The main ascending pathway for pain sensation is the crossed lateral spinothalamic tract (STT), which courses in the middle of the lateral funiculus and consists primarily of lamina I axons from the

contralateral side. This pathway is clinically and behaviorally critical for pain and temperature sensibility, as well as for itch and sexual sensations, based on cordotomy lesions. The main termination of this pathway in primates is a dedicated nociceptive- and thermoreceptive-specific relay nucleus in the posterior thalamus (VMpo), which projects topographically to a cytoarchitectonically distinct field in the dorsal margin of insular cortex and has a collateral projection to area 3a in the fundus of the central sulcus. The homologous pathway in cats and rats exists only in a rudimentary, primordial form; this pathway is proportionately greatly enlarged in humans. Microstimulation of this thalamic region in awake humans produces sensations of graded pain or cold. Lamina I spinothalamic neurons also terminate in a ventral portion of the main somatosensory thalamic relay nucleus (VPI), which in turn projects to the second somatosensory region (SII) in the parietal operculum of the lateral sulcus, and in a posterior portion of the medial dorsal thalamic nucleus (MD), which projects to the anterior cingulate cortex. The axons of spinothalamic lamina V cells ascend in the anterior spinothalamic tract and constitute another major pathway conveying tactile and pain-related activity; these axons terminate in VPI and also in conspicuous patches in the main somatosensory relay nuclei (VPM and VPL) around immunohistochemically distinct (calbindin-positive) cells, where WDR nociceptive neurons can be identified. In contrast to most sensory thalamocortical pathways, these neurons seem to project to the superficial layers of the primary somatosensory cortical region (SI) in the postcentral gyrus. Nociceptive WDR cells have been recorded in SI, and clusters of nociceptive-specific cells have been recorded in area 3a.

There are also other multisynaptic pathways that may provide ancillary pain information to the forebrain. Nociceptive spinothalamic cells in the spinal intermediate zone, especially at upper cervical segments, project bilaterally to the medial thalamus and to motor-related thalamic nuclei. Spinomedullary lamina X neurons that project to the dorsal column nuclei or the nucleus solitarius may be important for visceral pain. A spinoreticulothalamic pathway by way of the dorsal medulla may provide a substrate for widespread activation of supragranular frontal cortex.

The main regions of the human cerebral cortex that are activated by painful stimuli in functional imaging (PET, fMRI) and laser-evoked potential studies are the insula (limbic sensory cortex) and the anterior cingulate (limbic motor cortex). In addition, in many studies activation is also observed in the primary and secondary somatosensory areas. Other regions activated by painful stimuli include dorsolateral prefrontal cortex, striatum, cerebellum, hypothalamus, amygdala, and periaqueductal gray. These areas together form a complex forebrain network involved in pain sensation. Some of these areas may subserve distinguishable roles in pain perception, which are currently being vigorously investigated. Possible associations are as follows: the insula with qualitative sensory differentiation, homeostasis, and memory; the anterior cingulate with affect, motivation, response selection, and attention; and the somatosensory areas with intensity discrimination and somatomotoric integration. The insular cortex seems to be a primary sensory field for pain, temperature, and visceral (interoceptive) sensation, and abnormal activation of this region has been observed in neuropathic pain patients. Activation of the anterior cingulate appears to be particularly important for distressful cutaneous stimulation, because it is selectively associated with the perception of pain and unpleasantness with noxious thermal stimulation and also with itch. Noxious heat inhibits primary somatosensory cortical area 3b as it activates the adjacent area 3a. The distribution of activation of the somatomotor cortex of the central sulcus is related to the occurrence of phantom limb pain in amputees. However, many of these areas of the forebrain are interconnected, and thus the effects of stimulation or lesions in any one of these regions on pain sensation are equivocal and may produce imbalanced integration rather than a discrete change in function. Lesions involving posterolateral thalamus or the parietoinsular cortex can reduce pain and temperature sensation, but they can also result in the central pain syndrome (described later). Large lesions of the postcentral gyrus have no effect on pain, but small lesions (that probably include area 3a in the fundus of the central sulcus) can affect pain. Lesions of the anterior cingulate have had varied effects, including blunting the emotional aspect of pain (also produced by frontal lobotomy) or thermal hyperalgesia. Anatomical variability between human brains has been documented and presents a serious confound to these issues.

IV. MODULATION OF PAIN PROCESSING

Fast endogenous antinociceptive circuits are thought to have evolved to facilitate defense or escape behaviors and thereby enhance survival. The classic example supporting the existence of endogenous central pain control systems is the observation of soldiers with

massive wounds who did not complain of pain. In addition, there are well-documented occurrences of placebo-induced analgesia in patients with organic causes of pain. Central neural controls of the circuits involved in pain transmission make sense because nociceptive processing has significant interactions with integrative homeostatic and immune system function, and because healing following traumatic injury or inflammation requires a prolonged time period.

Descending supraspinal projections modulate nociceptive processing in the dorsal horn. These projections terminate in laminae I–V and originate primarily from homeostatic control regions in the brain stem, that is, the rostroventromedial medulla (including the raphe nuclei) and the enkephalinergic cells and catecholamine cell groups (A5, A6, A7) in the dorsolateral pontine tegmentum. The antinociceptive actions of these systems depend on serotonergic, adrenergic ($\alpha 2$), and enkephalinergic transmission at the spinal level. These descending systems act in part by way of inhibitory interneurons (causing indirect presynaptic actions on primary afferents) and in part by direct postsynaptic inhibition of transmission neurons. However, these projections include pronociceptive (facilitatory) actions on dorsal horn nociceptive processing as well. The descending inhibitory actions can be activated by noxious stimulation over the whole body, which may underlie the phenomenon of counterirritation, and the descending facilitatory actions can be activated by low-threshold mechanoreceptors, which might be involved in touch-evoked pain (mechanical allodynia) in neuropathic conditions. In addition, direct projections to lamina I from the hypothalamus probably underlie the spinal antinociceptive actions of oxytocin, vasopressin, and hypocretin (orexin), although these agents also have actions on thermoregulation and homeostatic integration.

Inhibition of second or tonic pain by low-threshold mechanoreceptive activity (e.g., rubbing, vibration) can be demonstrated with peripheral electrical nerve stimulation. The inhibition of spinal nociceptive processing by low-threshold afferent activity and by descending activity was an important aspect of the gate control theory of pain; this historically important contribution focused attention on spinal integration and endogenous modulation, albeit without regard for the role of nociceptive-specific neurons. Inhibition of pain processing by low-threshold stimulation occurs partially at the spinal segmental level and partially at the cortical level, although the mechanisms are still unclear. Dysfunction of such inhibition is another possible cause of touch-evoked allodynia in neuropathic pain. These inhibitory mechanisms may be engaged by transcutaneous electrical nerve stimulation (TENS) and by dorsal column stimulation, procedures used by some clinicians for pain control.

Innocuous thermosensory (cool or warm) stimulation can inhibit C-fiber-evoked pain. Application of cold or warmth is a well-established therapy that has peripheral effects on inflammation, sensitized nociceptors, and tissue perfusion. In addition, there is a central inhibitory effect of cooling on pain that is demonstrated by the thermal grill illusion, in which a painful, icelike burning sensation is elicited by spatially interlaced innocuous warm and cool stimuli. (This illusion is on display at many science museums.) In this illusion, the reduction of innocuous thermosensory activity (by simultaneously cooling and warming the skin) unmasks the cold activation of the pathway responsible for the burning sensation associated with noxious cold, which is normally inhibited by ongoing innocuous thermosensory activity. Functional imaging results indicate that noxious cold is distinguished by its activation of the anterior cingulate cortex (by way of lamina I spinothalamic input to the medial thalamus), and the thermal grill illusion causes the same pattern of activation in the forebrain (and in lamina I spinothalamic neurons). A conduction block of peripheral nerve fibers that eliminates ongoing activity in Aδ thermoreceptive fibers responsible for cooling sensation produces a similar disinhibition and enables a normally innocuous cool stimulus to produce a burning sensation (via polymodal C-nociceptors and lamina I polymodal nociceptive neurons). The intense burning experienced when lukewarm water is applied to a foot numbed by cold is directly comparable to the thermal grill sensation, and this highlights the thermoregulatory significance of the cold inhibition of burning pain. Descending modulation by the thermosensory region in the insular cortex of lamina I and of integration sites in the brain stem and thalamus may underlie this inhibitory mechanism.

Stimulation-induced analgesia can be produced by the electrical or chemical activation of sites in the brain that are the source of both descending and ascending pathways. Clinical deep brain stimulation of the periaqueductal gray (PAG), the hypothalamus, and other forebrain sites can be behaviorally antinociceptive, can reduce the responses of nociceptive dorsal horn neurons, and can alleviate certain kinds of chronic pain in patients. Stimulation of the periaqueductal gray activates descending pathways that involve serotonergic and peptidergic (substance P) neurons in the rostroventral medial medulla (including the

midline nucleus raphe magnus) and the noradrenergic and enkephalinergic neurons in the dorsolateral pons. These descending systems are also activated by systemic opiates or cannabinoids. Just as pain has direct effects on cardiorespiratory activity, these descending bulbospinal antinociceptive pathways also have direct actions on sympathetic preganglionic neurons and on brain stem autonomic sites. In addition, stimulation in the periaqueductal gray can cause intense negative or positive emotional experiences in awake patients. Nonetheless, stimulation-produced analgesia in human patients is often reported to be selectively effective with long-term use, without any autonomic changes or sensory experiences.

The endogenous pain control systems can, of course, be activated naturally, for example, by environmental stressors (so-called "stress-induced analgesia"), by behavioral danger signals that can be classically conditioned, and perhaps by meditation or biofeedback training. Significantly, there are also endogenous anti-analgesia (pronociceptive) circuits, which can be behaviorally activated by conditioned safety signals that indicate when danger will not occur. These circuits may be distinct from the antinociceptive and pro-nociceptive medullary circuits described earlier. Behavioral pronociceptive actions seem to be powerful enough to reverse the effects of the endogenous antinociceptive systems at the spinal level. Learned safety signals cause the spinal release of peptides (particularly cholecystokinin) that can block analgesia produced by stress, danger, or even morphine. Such antianalgesia mechanisms could have an important role in chronic pain and in the development of morphine tolerance.

V. NEUROCHEMISTRY OF PAIN

The neurochemical attributes of peripheral nociceptors and central pain processing are being studied vigorously. All small-diameter fibers appear to contain the neurotransmitter glutamate, and many contain various peptides and cotransmitters. Several markers have been identified, although none are entirely specific for a particular class of receptor or a particular submodality of pain. Neurochemicals that are not associated with low-threshold receptors are of particular interest. Many, but not all, C-nociceptors contain peptides such as substance P (SP), calcitonin-gene-related peptide (CGRP), neuropepetide Y, galanin, or somatostatin. Some C-nociceptors have the vanilloid receptor (VR1), which responds to protons and heat and is the channel responsible for chemogenic pain caused by the pungent red pepper ingredient capsaicin. Many nociceptive C-fibers also have tetrodotoxin (TTX) resistant sodium channels that are found only in small dorsal root ganglion cells (SNS or PN3 and SNS2 or NaN). In addition, C-fibers that contain SP or CGRP have particular neurotrophin receptors (TrkA and p75) that are sensitive to nerve growth factor (NGF), whereas a subset of C-fibers that can be labeled for the surface lectin IB4 are sensitive to glial-derived growth factor (GDNF). Selective augmentation, blockade, and gene knockout experiments have confirmed that each of these factors affects pain processing.

The peripheral nociceptors respond to a variety of chemicals. Agents that can directly activate nociceptors include glutamate, ATP, potassium, bradykinin, and protons, all of which are released from damaged tissue or mast cells. In addition, these and other agents can sensitize nociceptors, causing spontaneous discharge, increasing their responsiveness to noxious stimulation, and lowering their thresholds to natural stimuli. Sensitization can be produced by repeated noxious stimulation or by chemical agents [such as bradykinin or the pro-inflammatory cytokines, like interleukin 1 (IL-1β) and tumor necrosis factor (TNFα)].

The eicosanoids, derivatives of the cell membrane lipid arachidonic acid, which include the prostaglandins and the leukotrienes, are particularly important sensitizing agents (especially PGE2 and PGI2). The formation of prostaglandins depends on cyclo-oxygenase (COX), which is inhibited by aspirin and other nonsteroidal anti-inflammatory drugs (NSAIDs) such as ibuprofen and naproxen. Most NSAIDs inhibit both COX-1 (constitutive) and COX-2 (inducible), and aspirin does so irreversibly. Whereas COX-1 is essential for the health of the stomach and kidney, the induction of COX-2 has a primary role in inflammation, sensitization, and fever. Thus, newly developed selective COX-2 inhibitors might reduce the deleterious effects of aspirin on stomach and kidney function. Corticosteroids produce a similar anti-inflammatory effect by blocking the induction of COX and the release of arachidonic acid stores, although they also have many other effects on immune function and metabolism.

Antinociceptive chemical agents that might act peripherally have been sought, but few have been found useful so far. Serotonin 5HT1 receptor agonists (such as sumatriptan, which is effective against migraine headaches) block the discharge activity of trigeminal cerebrovascular afferents. They also block

the peripheral release of SP and CGRP from the terminal branches of active nociceptors, which is a significant effect because the release of SP promotes plasma extravasation (edema), and CGRP is a potent vasodilator that contributes to neurogenic inflammatory flare (see Section 6), actions that are thought to contribute to headaches. Adenosine and somatostatin also appear to have direct inhibitory effects on nociceptive terminals. Following inflammation, the opiates have indirect peripheral antinociceptive actions by way of receptors on immunocytes.

The agents that act at the peripheral terminals can also act at the central terminals (presynaptic modulation) of primary afferent nociceptors. The central terminals of nociceptive primary afferent fibers release neurotransmitters, neuromodulators, and trophic agents in the dorsal horn. They all seem to release glutamate and/or aspartate, major excitatory amino acid (EAA) neurotransmitters that are also used by low-threshold primary afferents and many other neurons. Nociceptive (and other) primary afferent fibers may corelease ATP, and some nociceptive C-fibers also corelease peptides, such as SP, CGRP, neuropeptide Y, and galanin. The release of cocontained neurotransmitters may be under differential, independent control. Release at central nociceptive terminals is presynaptically inhibited by GABA, opiates, noradrenalin, adenosine, cannabinoids, and other agents, but enhanced by prostaglandin (PGE2) and cytokines.

The EAAs released by nociceptive fibers act on several types of receptors in the dorsal horn, which include both ionotropic (AMPA, kainate, and NMDA) and metabotropic (mGluR) receptors. Activation of ionotropic glutamate receptors leads to a rapid depolarization in the postsynaptic cell due to the influx of Na^+ through the AMPA receptor or Na^+ and Ca^{2+} through the NMDA receptor. Selective blockade of AMPA receptors (with antagonists such as CNQX) reduces nociceptive activation of dorsal horn neurons and acute pain behaviors in rats. Depolarization can remove the Mg^{2+} blockade of NMDA channels, and NMDA receptor activation produces a more prolonged depolarization and second messenger activation that can lead to long-term changes within dorsal horn neurons. Selective blockade of postsynaptic NMDA receptors (such as with MK-801 or dextromethorphan) can prevent the development of hyperexcitability in dorsal horn neurons (central sensitization) and reduce inflammatory or nerve-injury-induced pain behaviors in animal studies; however, in clinical trials the side effects of these agents are prohibitive. Nonetheless, coadministered NMDA antagonists might permit the use of lower doses of spinal morphine for analgesia and delay the development of opiate tolerance. Antagonists to metabotropic EAA receptors (in particular, mGluR1), which are G-protein-coupled receptors that activate the second messenger protein kinase C (PKC), also seem to reduce or inhibit nocifensive responses; these may also have presynaptic effects.

The SP-containing nociceptive fibers can activate several neurokinin receptors, but especially the neurokinin 1 receptor (NK1r). This G-protein-coupled transmembrane receptor is located almost entirely on lamina I neurons (and a few laminae III–V neurons). Activation of NK1r causes a delayed, long-lasting (tens of seconds) depolarization that can summate, and it results in phosphorylation of the NMDA receptor complex (by way of PKC), rendering it more active. The activation of NK1r in lamina I cells by different noxious stimuli (or its inhibition) can be observed anatomically, because after activation it is translocated from the cell membrane (internalized) into the cytosol of the cells. Abolition of SP fibers, NK1r neurons, or PKCγ in the dorsal horn reduces central sensitization of nociceptive neurons and sensitivity to tonic noxious stimuli in behavioral models.

Inhibitory neurotransmitters are present in intrinsic spinal cord neurons, especially in the substantia gelatinosa (SG). These interneurons are activated by afferent fibers and by descending modulatory fibers. Nociceptive dorsal horn neurons can be inhibited by GABA, glycine, opioids (enkephalin, dynorphin), acetylcholine (epibatadine), adenosine, noradrenaline, dopamine, and serotonin. Both direct synaptic and extrasynaptic actions have been shown. These actions can be modulated locally by steroids, such as cortisol or estrogen. The organization of the intrinsic inhibitory circuits is not well-understood.

Opioid actions have been closely scrutinized, because the classic opiate morphine and its congeners and the opioid peptides with similar actions are still the most potent analgesics known. In the spinal cord, opioids produce presynaptic inhibition of transmitter release from nociceptive primary afferents by the inhibition of calcium entry; this action is antagonized indirectly by cholecystokinin (CCK). Opioids produce postsynaptic inhibition of nociceptive dorsal horn cells via a G-protein-coupled potassium conductance increase. All three types of opiate receptors (μ, δ, and κ) are involved, and the subunit composition of these receptors may be dynamic. Epidural or intrathecal injections of fast-acting μ-selective opiates (such as fentanyl) clinically (in humans) and behaviorally (in

rats) are most effective. Spinal injections usually include bupivicaine, a long-lasting local anesthetic that blocks activity in small-diameter primary afferent fibers, thereby reducing the amount of opiate needed and the danger of respiratory depression due to rostral spread. κ opioids seem to be effective for labor pain in women. δ opioid actions appear to be important for the development of spinal tolerance.

Systemically administered opiates engage endogenous pain control mechanisms in the forebrain, particularly in the brain stem periaqueductal gray, the amygdala, and the anterior cingulate cortex. In the periaqueductal gray, μ-selective agonists disinhibit the endogenous descending antinociceptive pathways involving serotonergic and noradrenergic fibers that are activated by stimulation-induced analgesia. A different opioid antinociceptive pathway normally activated by endogenous β-endorphin (from the hypothalamus) in the periaqueductal gray is not engaged by morphine. Aspirin potentially could synergize with the effects of morphine in the periaqueductal gray, where morphine's mechanism of action involves an eicosanoid pathway. The development of associative (context-dependent) tolerance to morphine can be blocked by microinjection of CCK antagonists in the amygdala. Opiates traditionally have been underprescribed for acute or postoperative pain relief because of the well-publicized fear of dependency; however, usage of opioids for analgesia is now strongly recommended, because controlled studies have shown that fewer than 1% of patients are at risk and because pain itself can affect outcome. Coadministration of synergistic adjuvants that help prevent the development of pharmacological tolerance has further encouraged increased clinical use of opiates, for example, in bedside minipumps for patient-controlled analgesia.

The neurochemistry of supraspinal pain processing regions is not well-known, in large part because many clinically relevant sites in primates have been identified only relatively recently. The terminals of lamina I spinothalamic neurons in primates contain glutamate, and some can be immunohistochemically labeled for calbindin, SP, or enkephalin. Presumably, many of the same neuromodulators found at spinal levels are also involved supraspinally.

VI. PAIN IN INJURY AND DISEASE

Tissue damage or infection elicits a local injury response, which includes the production of eicosanoids (from arachidonic acid), bradykinin (from plasma kalikrein), lactic acid and superoxide free radicals (from leukocytes and mast cells), serotonin, histamine, and various pro-inflammatory cytokines (e.g., IL1-β and TNF-α). All of these inflammatory (and antibacterial) agents have direct effects on nociceptors, resulting in activation or sensitization. Activated nociceptors also contribute a neurogenic component to inflammation by the peripheral release of agents such as glutamate, SP, CGRP, and ATP, which affect sympathetic efferents, mast cells, and small blood vessels, causing vasodilatation and plasma extravasation (increased permeability between capillary endothelial cells). This in turn increases tissue infiltration with blood cells, plasma, cytokines, immune complement agents, and more bradykinin, serotonin, and glutamate. Peripheral inflammation is modulated by ascending nociceptive activation of the hypothalamo–pituitary–adrenal (HPA) axis. Inflammation-induced trophic agents, such as nerve growth factor (NGF) or glial-derived growth factor (GDNF), that are involved in regeneration and healing can have long-term effects on the expression of chemicals and receptors in primary afferent fibers, and they are potentially significant for neuropathic changes.

Sensitization of nociceptors (lowered threshold for activation and increased activity to a noxious stimulus) produces the condition of primary hyperalgesia. An example of primary hyperalgesia is the increased sensitivity caused by a sunburn, where light touch or warmth on the sunburned area hurts (lowered threshold) and a hot shower is more painful than normal. Inflammatory hyperalgesia is an important aspect of pain in many diseases. The pain of arthritis, for example, is due to the activity of sensitized joint nociceptors that otherwise would not be activated by normal movement. The effectiveness of aspirin and other NSAIDs in relieving such pain is based on the role of prostaglandins in producing the sensitization of nociceptors that underlies primary hyperalgesia.

Hyperalgesia results not only from the peripheral sensitization of nociceptors but also from increased excitability of central nociceptive neurons. Sustained activity in nociceptive C-fiber afferents causes chemical changes in dorsal horn neurons so that they become hyperexcitable. Sensitized dorsal horn neurons with increased spontaneous activity, enlarged receptive fields, lowered thresholds, and increased responsiveness to peripheral stimulation are responsible for a zone of secondary hyperalgesia surrounding an injury. Nociceptive-specific cells can become responsive to input from mechanoreceptive A-fibers. Such sensitization may be one basis for the tactile allodynia (pain

from low-threshold stimulation) often observed in neuropathic pain. Central sensitization is dependent on the activation of NK1 and NMDA receptors and the subsequent mobilization of PKC. Activation of immune-competent microglia in the spinal cord may play an important role in central sensitization; spinal pro-inflammatory and anti-inflammatory cytokines can have significant effects. Prevention of the central sensitization initiated by afferent C-fiber nociceptive activity is the justification for preemptive analgesia, in which local and systemic analgesics are administered prophylactically before major surgery.

The NSAIDs also can affect hyperalgesia by a central action. Afferent nociceptive activity causes the release of prostaglandins in the spinal cord, which have presynaptic effects (increasing neurotransmitter release from primary afferent terminals) and probably postsynaptic effects as well. Spinal administration of NSAIDs reduces central sensitization. In addition, they reduce peripheral neurogenic inflammation caused by pathological, centrifugal activity (so-called dorsal root reflexes) carried by nociceptive fibers into peripheral tissues on both sides of the body.

Generalized illness and immune system challenges (infection) can induce hyperalgesia. Circulating pro-inflammatory cytokines have direct effects on primary afferent nociceptors and central neurons. Illness-induced hyperalgesia can be reduced by agents that antagonize pro-inflammatory cytokines (IL-1ra, IL-10). Hyperalgesia may be an important component of immune-activated, neurally organized illness behavior (which includes fever, decreased activity, decreased food and water intake, and increased sleep), consistent with a role for C-fiber afferents in metaboreception and homeostasis. Pain sensation can be strongly influenced by the general health of the body, as well as by hormone levels, nutrition, time of day, and behavioral state. Conversely, pain itself can have strong negative effects on health. Pain, just as stress or surgery, can inhibit immune function, enhance tumor growth, and increase morbidity. Unrelieved pain usually leads to disturbance of sleep, loss of appetite, depression, severe impairment of general health, and even suicide.

Pain due to organic causes (injury or disease) can usually be relieved by treating the source of nociceptive activation, but pain due to injury to the nervous system in some cases becomes chronic and intractable. The nervous system itself is insensate—only the epineurium surrounding peripheral nerves and the meninges surrounding the brain are innervated by C-fiber afferents—and, thus, pain due to neural injury is called neuropathic pain. Peripheral nerve damage due to trauma (e.g., carpal tunnel syndrome, neuroma), vascular malformations (e.g., trigeminal neuralgia), or due to pathogenic processes (e.g., diabetic neuropathy, postherpetic neuralgia, rheumatoid inflammation) can initiate cellular changes in primary afferent fibers that result in ectopic neural activity (originating at the site of nerve injury or in the dorsal root ganglion cells), molecular phenotype changes, or anatomical sprouting in the periphery and in the spinal dorsal horn. For example, nerve lesions can cause mechanoreceptive A-fiber afferents to produce SP or to sprout terminals into laminae I and II, small-diameter afferents to up-regulate peptides (galanin) and trophic (NGF) receptors, and nociceptive fibers to become sensitive to circulating adrenalin and sympathetic efferent activity. The latter effect may be particularly important for a condition known as reflex sympathetic dystrophy (RSD), complex regional pain syndrome or causalgia, which can often be relieved by sympatholysis. Permanent pathological changes may be abnormal sequelae of healing and regeneration, and they can cause a vicious cycle of pain-related activity. In such conditions, cytokines (TNFα), elements of the immune complement cascade (C3a), or antibodies to gangliosides (GD2) that are present on primary afferent fibers can activate C-nociceptors. Central inflammatory (microglial) reactions may be very significant, and the use of intrathecal anti-inflammatory agents (NSAIDS, IL-10) is being studied. Systemic lidocaine and intrathecal gabapentin may relieve neuropathic pain by interfering with such mechanisms peripherally and centrally.

Central spinal hyperexcitability can also result from brachial plexus root evulsion; this form of deafferentation pain can be effectively treated by a dorsal root entry zone (DREZ) lesion, which eliminates hyperactive neurons from the dorsal horn of the cervical spinal cord. Phantom limb pain, which can occur following amputation, particularly if there was significant preoperative pain, may similarly involve changes in both peripheral nerve fibers and central function.

Damage to the central nervous system can also result in intractable pain. In Wallenberg's syndrome (anesthesia dolorosa), an infarct in the caudal medulla produces the loss of evoked pain and temperature sensation in the ipsilateral face (due to damage to the trigeminal dorsal horn) and contralateral body (due to interruption of the spinothalamic tract), and ongoing burning pain can occur in these regions. Similarly, so-called central pain can appear following a spinal lesion

(from trauma, multiple sclerosis, or even an anterolateral cordotomy performed to eliminate cancer pain) or following stroke-induced damage in the forebrain, particularly in the posterolateral thalamus or the parietoinsular cortex. In such cases, ongoing burning pain is reported in deep and cutaneous tissues in a region in which the lesion has eliminated normal pain and (especially) temperature sensation in the contralateral body. Pain from innocuous cooling or touch (allodynia) is typical. Central pain results from the release (disinhibition) of activity in the brain; this could be due to the loss of endogenous analgesia mechanisms like the normal inhibition of pain by cold. Central pain is unresponsive to opiates, but tricyclic antidepressants (amitryptyline) or antiepileptics (carbamazepine) can be effective.

See Also the Following Articles

BRAIN LESIONS • HEADACHES • HOMEOSTATIC MECHANISMS • MILD HEAD INJURY • MODELING BRAIN INJURY/TRAUMA • NOCICEPTORS • OPIATES • PAIN AND PSYCHOPATHOLOGY • PHANTOM LIMB PAIN

Suggested Reading

Belmonte, C., and Cervero, F. (Eds.). (1996). *Neurobiology of Nociceptors*. Oxford University Press, Oxford, UK.

Besson, J. M., Guilbaud, G., and Ollat, H. (Eds.). (1995). *Forebrain Areas Involved in Pain Processing*. John Libbey Eurotext, Paris.

Bonica, J. J. (Ed.). (1990). *The Management of Pain*, 2nd ed. Lea & Fibiger, Philadelphia.

Craig, A. D., Chen, K., Bandy, D., and Reiman, E. M. (2000). Thermosensory activation of insular cortex. *Nature Neurosci.* **3,** 184–190.

Fields, H. L. (1987). *Pain*. McGraw-Hill, New York.

Perl, E. R. (1984). Pain and nociception. In *Handbook of Physiology, Section 1, The Nervous System, Volume III, Sensory Processes* (I. Darian-Smith, Ed.), pp. 915–975. American Physiological Society, Bethesda, MD.

Price, D. D. (1999). *Psychological Mechanisms of Pain and Analgesia*. IASP Press, Seattle, WA.

Sherrington, C. S. (1900). Cutaneous sensations. In *Text-book of Physiology* (E.A. Schäfer, Ed.), pp. 920–1001. Pentland, Edinburgh, UK.

Wall, P. D., and Melzack, R. (Eds.). (1999). *Textbook of Pain*, 4th ed. Churchill-Livingstone, Edinburgh, UK.

Willis, W. D. (Ed.). (1992). *Hyperalgesia and Allodynia*. Raven Press, New York.

Yaksh, T. L., Lynch, C., Zapol, W. M., Maze, M., Biebuyck, J. F., and Saidman, L. J. (Eds.). (1998). *Anesthesia: Biologic Foundations*. Lippincott-Raven, Philadelphia.

Parkinson's Disease

MELVIN D. YAHR and MATTHEW A. BRODSKY
Mount Sinai Medical Center—New York University

GLOSSARY

akinesia An inability to move; the absence of movement.

alpha synuclein A protein found to be misfolded in its expression in the brain in an autosomal-dominant form of familial parkinsonism. It is the product of the gene *PARK1*, which is located on chromosome 4q21-22.

basal ganglia The deep nuclei of the cerebrum, including the caudate nucleus, putamen, globus pallidus subthalamic nucleus, and pedunculopontine nucleus.

bradykinesia Slowness of movement.

carbidopa An inhibitor of the enzyme dopa decarboxylase, which converts levodopa to dopamine and increases its efficacy in the brain. Its action is limited to tissues outside the brain because it does not cross the blood–brain barrier and thus prevents side effects such as nausea.

caudate nucleus One of the two nuclei comprising the corpus striatum; lesions of the caudate rarely cause motor disorders but are more likely to cause behavioral problems.

cortical basal ganglionic degeneration A form of atypical parkinsonism that is highly asymmetric, with primarily features of rigidity, akinesia, and apraxia.

dopamine A neurotransmitter; its depletion in Parkinson's disease is responsible for symptoms of the disorder.

dopamine agonist A compound that binds to dopamine receptors, stimulates them, and mimics the action of dopamine.

dysphagia Difficulty in swallowing.

dystonia Abnormal involuntary muscle contraction and posture, usually action induced.

festination An abnormal rapid forward gait, which leads to falling.

globus pallidus One of the deep nuclei of the brain; it is anatomically divided into a "pars externa" and "pars interna." The pars interna is the final common output from the basal ganglia circuit to the cerebral cortex via the thalamus.

glutathione An important compound in antioxidant defense and in the repair of oxidized proteins. Glutathione levels are decreased in the substantia nigra in Parkinson's disease.

hyperhidrosis Excessive sweating.

hypomimia Decreased facial expression.

hypophonia Decreased volume of speech.

levodopa The molecular precursor to dopamine; it remains the most effective agent for pharmacotherapy of Parkinson's disease.

Lewy body The pathologic cellular marker of Parkinson's disease; it is composed of protein aggregates.

micrographia Small handwriting.

mitochondria The organelle in the cell that provides the majority of energy for cellular function via oxidative phosphorylation; it may be a source of oxidative damage in neurodegenerative diseases such as Parkinson's disease.

monoamine oxidase An enzyme that metabolizes monoamines; it has two subtypes, A and B. Type B metabolizes dopamine in the brain.

multiple system atrophy A progressive neurodegenerative syndrome characterized clinically by autonomic dysfunction, parkinsonism, and cerebellar ataxia.

orthostatic hypotension The abnormal decrease in blood pressure when one goes from a supine to standing position.

oxidative phosphorylation The process by which oxygen is utilized by mitochondria in the cell to form ATP, the basic currency of energy by which cellular processes occur.

parkin The protein found in an autosomal-recessive young-onset form of parkinsonism. It is the product of the gene *PARK2*, which is found on chromosome 6q25.2-27.

progressive supranuclear palsy A neurodegenerative condition with parkinsonian features; progressive gait and balance impairment, downward gaze paralysis, and rigidity are its major features.

putamen One of the two nuclei composing the corpus striatum; its innervation from substantia nigra pars compacta becomes deficient in Parkinson's disease.

retropulsion The tendency to fall backwards.

striatum Composed of the caudate nucleus and putamen; the striatum forms the input zone from the cerebral cortex to the basal ganglia.

substantia nigra Collection of melanin-containing neurons of the midbrain that is divided into two distinct anatomic compartments, the pars compacta and the pars reticulata. The pars compacta contains the dopamine-producing cells that project to the striatum which degenerate and die in Parkinson's disease.

subthalamic nucleus One of the deep nuclei of the cerebrum; it is the only nucleus of the basal ganglia that uses glutamate as its neurotransmitter. It is overactive in the pathological state of Parkinson's disease, and this overactivity has been purported to cause downstream damaging effects in Parkinson's disease excitotoxicity promulgated by glutamate.

thalamus Composed of a series of nuclei in the diencephalon, this anatomic structure forms the major relay for afferent pathways to the cerebral cortex.

ubiquitin A protein that accumulates within pathologic intracellular or extracellular deposits in neurodegenerative diseases, such as the Lewy body in Parkinson's disease. It has multiple functions but it is primarily known for its involvement in a pathway that regulates the bulk of intracellular protein turnover.

Parkinson's disease (paralysis agitans, shaking palsy) is a slowly progressive neurodegenerative disorder of the central nervous system. It is characterized by the presence of bradykinesia (slowness of movements), rigidity, and tremor.

I. INTRODUCTION

First described in 1817 by James Parkinson, a practicing physician in London, in a monograph titled "Shaking Palsy," Parkinson's disease (PD) attracted little attention until approximately 50 years later when the famous French neurologist Jean Charcot more fully defined its phenomenology. Charcot pointed out that there was little if any palsy, that the difficulty in movement related to its initiation and maintenance as well as to stiffness of the musculature, and that the tremor was distinctive in that it occurred at rest rather than with action. He suggested the term "shaking palsy" be discarded and that the disease be given the pseudonym Parkinson's disease honoring its discoverer.

Although many other disorders of the nervous system may show signs and symptoms akin to PD (so-called secondary parkinsonism), PD is a distinctive clinical and pathological entity, accounting for approximately 75% of all cases of parkinsonism.

II. EPIDEMIOLOGY

PD is one of the most common causes of neurological disability, affecting 1% of the population over age 55. The prevalence is 1 in 350, giving an overall lifetime risk of approximately 1 in 40. It is typically a disease of the middle to late years, beginning at a mean age of 50–60 years and progressing slowly over a 10- to 20-year period. The age of onset assumes a broad bell-shaped distribution, with approximately 5% of cases beginning before age 40.

The reported incidence rates of PD range from 4.5 to 21/100,000 population/year in community-based studies. Prevalence rates of PD vary from 18/100,000 persons in a Shanghai, China, population survey to 328/100,000 in the Parsi community in Bombay, India. There are currently at least 500,000 PD cases in the United States. The incidence and prevalence of PD increase with increasing age of the population surveyed. In most studies PD is slightly more common in men than in women.

Identified environmental risk factors include heavy metals such as iron and manganese, drinking well water, farming, living in a rural residence, working in wood pulp mills and steel alloy industries, and exposure to herbicides or pesticides.

III. CLINICAL MANIFESTATIONS OF PD

Parkinsonism, as a clinical syndrome, is readily recognizable by the fixed, expressionless facies and stooped body posture. It is characterized by specific motor deficits: tremor, akinesia (or bradykinesia), muscular rigidity, and postural instability. At least two of these cardinal features should be present to make a firm diagnosis. A wide variety of unrelated disease states can result in parkinsonism; however, the most common form is Parkinson's disease.

PD begins insidiously and is slowly progressive. It is postulated that the underlying pathology (i.e., the loss of pigmented brain stem nuclei) precedes the onset of symptoms by many years. Indeed, at least 80% of dopaminergic neurons must be lost before overt signs and symptoms first appear.

As measured by clinical signs and degree of functional impairment, PD proceeds through five stages.

The first stage shows predominantly one-sided involvement, with little or no functional impairment. Stage II is characterized by bilateral body involvement and minimal functional difficulties. Stage III marks a critical development in that in addition to the symptoms on both sides of the body, postural instability with impaired balance occurs. Some elements of daily living activities may need assistance in this stage. In stage IV, all the previous symptoms are intensified and more reliance on others to perform activities of daily living is needed. Stage V is practically an end stage to the disease, with wheelchair or bed existence and need for nursing care. The rate of progression through these various stages is quite variable and markedly modified by individual tolerance and responsiveness to anti-Parkinson drug therapy.

The signs and symptoms of PD are motor and nonmotor in nature.

A. Motor Features

1. Akinesia or Bradykinesia

The absence, paucity, or slowness of movement is often a disabling sign. Sequential and complex motor acts are particularly difficult. Quick repetitive movements typically show a decrease in both amplitude and frequency. Other manifestations of bradykinesia include drooling due to poor swallowing of saliva and loss of facial expression (hypomimia). Gait disturbance is prominent, with a shuffling quality, decrease in stride length, and reduced arm swing. Attempting to turn becomes difficult, requiring an extra step or two. Freezing, a form of akinesia, is the transient inability to perform active movements. It most often affects the legs and occurs when walking.

2. Muscular Rigidity

On examination, rigidity is noted as increased resistance to passive movement of a limb segment. The amount of resistance remains fairly constant through the entire range of motion, in both flexion and extension, and is not greatly influenced by the speed or force with which the movement is performed. This distinguishes it from spasticity. The rigidity of PD is most often cogwheeling, but such may not be present. Voluntary movement in one limb may induce rigidity in the contralateral limb. The patient may develop truncal rigidity with a flexed or stooped posture.

3. Resting Tremor

In 75% of patients, tremor is the first motor manifestation, usually beginning unilaterally in the distal segment of a limb, in most cases the fingers. The tremor often involves rhythmic, alternating opposition of the forefinger and thumb in the classic stereotypic pill-rolling variety. The characteristic frequency is 3–5 Hz. Over the course of several years, tremor usually spreads proximally in the affected arm before involving the ipsilateral leg and, finally, the contralateral limbs. Although tremor is bilateral in advanced disease, it often maintains some asymmetry throughout the course. In later stages, an accompanying tremor of the face, lips, or chin is not uncommon. The tremor occurs at rest and usually abates when the affected limb performs a motor task.

4. Postural Instability

In general, loss of postural reflexes is the last cardinal sign to appear and often proves to be the most disabling. Patients with postural instability develop festination, manifested by increased rapidity of gait, which leads to falling. Retropulsion, the tendency to fall backwards, is prominent and readily demonstrated in the erect position with a mild pull or push.

5. Other Features

Other motor features of PD include development of smaller handwriting (micrographia), softer voice (hypophonia), start-hesitation and freezing episodes, decreased frequency of eye blinking, and dysphagia. Dystonic phenomena may occur in association with PD, particularly involving the limbs and occurring in early morning hours.

B. Nonmotor Features

1. Autonomic Dysfunction

Autonomic features of PD include orthostatic hypotension, intestinal motility disorders resulting in constipation, bladder and erectile dysfunction, excessive sweating (hyperhidrosis), and pupillary abnormalities.

2. Cognitive Dysfunction

Tests of cognitive function demonstrate mild-to-moderate deficits, including visuospatial impairment,

attentional set-shifting difficulties, and poor executive function. Although not an early finding in PD, dementia eventually occurs in 20–30% of PD patients.

Depression is not an uncommon psychiatric symptom in PD. Prevalences ranging from 45% with minor symptoms to 8% with major symptoms have been reported. It may precede the motor disturbances and be responsive to drugs used for the treatment of PD.

Disturbances in sleep, with altered REM patterns, occur frequently. These vary from frank insomnia to interrupted sleep intervals.

3. Sensory Abnormalities

Paresthetic phenomena as well as impaired olfactory function occur in PD. Indeed, loss of olfaction may be an early sign in Parkinson's patients.

IV. ANATOMY AND PATHOLOGY

A. Anatomy

The motor signs and symptoms of PD result primarily from dysfunction of the basal ganglia. The central mechanism for the physiological dysfunction of the basal ganglia in PD is a progressive decline in the concentration of dopamine. Neuronal loss and depigmentation have been demonstrated in the substantia nigra pars compacta of parkinsonian patients. Profound reductions in the concentration of dopamine in the caudate nucleus and putamen have also been discovered. Pigmented neurons of the substantia nigra project to the striatum and provide it with dopaminergic input. The regional "fallout" of dopaminergic neurons in the substantia nigra pars compacta of patients with PD appears to be specific and selective. Although other neurotransmitters are affected in PD, the nigrostriatal dopaminergic tract is most significantly affected by the neurodegenerative process.

The basal ganglia form a network of parallel loops that integrate cortical regions (motor, limbic, and associative) with the basal ganglia nuclei and thalamus. Cortical motor projections make excitatory glutamatergic synaptic connections with medium spiny neurons of the posterolateral putamen containing GABA. These neurons give rise to two pathways that connect the striatum to the output nuclei of the basal ganglia—the globus pallidus pars interna (GPi) and the substantia nigra pars reticulata (SNr). Neurons in the "direct pathway" project from the putamen to the GPi/SNr, have Dl dopamine receptors, coexpress the peptides substance P and dynorphin, and provide a direct inhibitory effect on GPi/SNr neurons. Striatal neurons in the "indirect pathway" connect the putamen with the GPi/SNr via synaptic connections in the globus pallidus pars externa (GPe) and subthalamic nucleus (STN). They contain D2 receptors and the peptide enkephalin. Projections from putamen to GPe and from GPe to STN are GABAergic and inhibitory. Neurons originating in the STN use glutamate as a neurotransmitter and activate neurons in the GPi/SNr. Stimulation of neurons in the indirect pathway leads to inhibition of the GPe, disinhibition of the STN, and excitation of the GPi/SNr. Thus, the output activity of the basal ganglia is influenced by the opposing effects of inhibitory inputs from the direct pathway and excitatory inputs from the indirect pathway. This, in turn, provides an inhibitory effect on brain stem and thalamocortical neurons involved in motor activities.

In the parkinsonian state, there is increased neuronal activity in the GPi/SNr output nuclei of the basal ganglia, which leads to excessive inhibition of thalamocortical and brain stem motor systems. Reduced activation of dopamine receptors caused by dopamine deficiency results in reduced inhibition of neurons of the indirect pathway and decreased excitation of neurons of the direct pathway. Reduced inhibition from the indirect pathway leads to overinhibition of the GPe, disinhibition of the STN, and increased excitation of GPi/SNr neurons, whereas decreased activation from the direct pathway causes a reduction in its inhibitory influence on the GPi/SNr. The net result is an excessive activation of basal ganglia output neurons accompanied by excessive inhibition of motor systems, leading to parkinsonian motor behavior (Fig. 1).

B. Imaging

1. Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) in patients with typical PD is usually normal, but a high field strength heavily T2-weighted MRI may show a wider area of lucency in the substantia nigra probably indicative of increased accumulation of iron. Prominent hypointensity in the putamen strongly indicates the presence of atypical parkinsonism, such as multiple-system atrophy.

2. Positron Emission Tomography Studies

Using [^{18}F] fluorodopa positron emission tomography (PET) scans to assess the integrity of the striatal

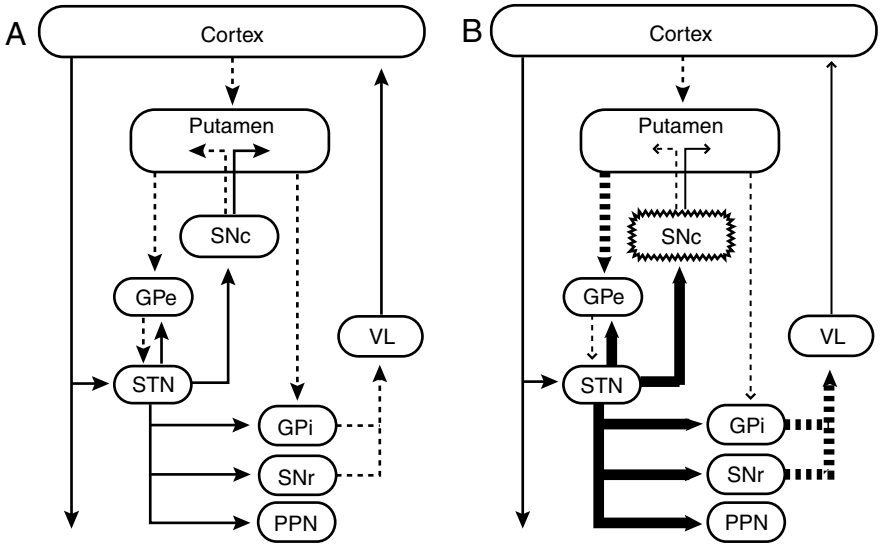


Figure 1 The basal ganglia circuit in normal (A) and Parkinson's disease (B). Solid arrows represent excitatory projection and broken arrows represent inhibitory projections. The thickness of the arrows indicates the degree of activity of each projection.

dopaminergic terminals, characteristic reduction of the [^{18}F]fluorodopa uptake, particularly in the putamen, can be demonstrated in virtually all patients with PD, even in early stages. As estimated on the basis of PET with [^{18}F]fluorodopa, the rate of nigral neuronal loss is faster initially and then tends to approach the normal age-related decline (Fig. 2).

3. Single Photon Emission Computerized Tomography

Single photon emission computerized tomography imaging of the striatal dopamine reuptake sites with ^{123}I-labeled β-CIT [2β-carboxymethoxy-3β-(4{^{123}I}iodophenyl)tropane] and of presynaptic vesicles with [^{11}C] dihydrotetrabenazine may also be helpful in differentiating PD from atypical parkinsonism.

C. Neuropathology

The pathological hallmarks of PD include massive loss of pigmented neurons in the substantia nigra as well as the presence of Lewy bodies. The Lewy body (LB) is a hyaline intraneuronal inclusion, and its presence is essential for pathological confirmation of PD. The LB is not specific to PD because it is found in a number of other conditions including Hallervorden–Spatz disease, ataxia–telangiectasia, and subacute sclerosing panencephalitis. The LB contains protein, free fatty acids, sphingomyelin, and polysaccharides, and it stains for neurofilament (Figs. 3 and 4).

V. ETIOLOGY AND PATHOGENESIS

Despite extensive investigative studies, the cause of PD has not yet been defined. Considerable interest initially centered on an infectious agent (i.e., virus) since a number of instances of parkinsonism followed the pandemic of influenza and encephalitis lethargica in 1917. However, postencephalitic parkinsonism differes from PD in many features and is more in keeping with a secondary type of a disorder. To date, viral DNA probes of tissue from PD brains and serum antibody titers have failed to reveal the presence of any known virus or bacteria.

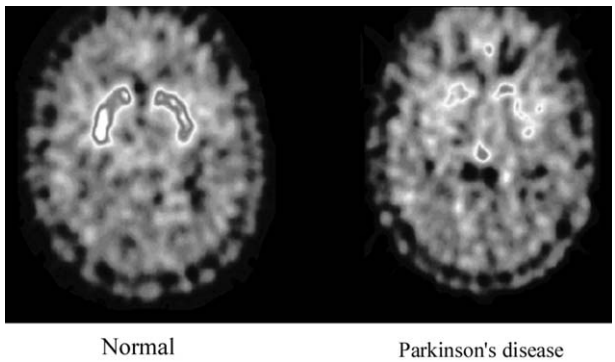


Figure 2 PET study using [^{18}F] fluorodopa.

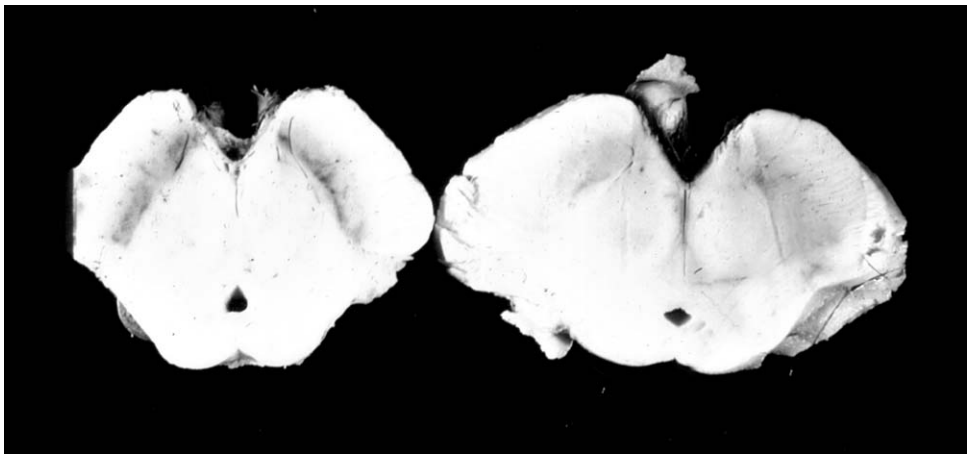


Figure 3 Gross midbrain specimen from a normal (left) and Parkinson's disease (right) brain. Note the decreased pigmentation indicating cellular loss in the substantia nigra pars compacta.

Considerable attention has been directed toward environmental toxins as being causative. The finding that manganese miners develop a form of parkinsonism has led to a search for heavy metal intoxication. This search has been fruitless. The serendipitous finding that the injection of an inadvertently produced meperidine analog, MPTP, led to the development of a relatively acute form of parkinsonism with many characteristics of PD prompted an extensive search for similar agents producing PD, particularly pesticides existing in the environment. None have been found. However, MPTP has become a valuable tool for studying the mechanism of dopaminergic cell death. Systemic injection of MPTP into nonhuman primates also produces parkinsonism with bradykinesia, rigidity, and freezing. Like humans, MPTP-treated monkeys develop severe dopamine depletion in the striatum with marked nigrostriatal neuronal loss.

Evidence supports the view that MPTP causes cell death through inhibition of oxidative phosphorylation and the generation of free radicals. MPTP is a protoxin, being converted to its active form l-methyl-4 phenylpyridinium (MPP^+). MPP^+ is a substrate for the dopamine reuptake system. Thus, MPP^+ is

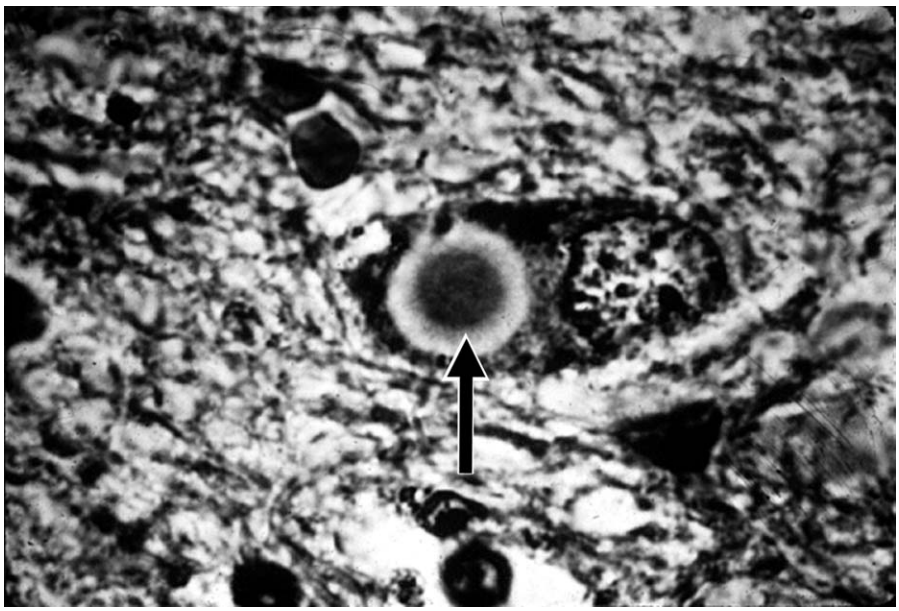


Figure 4 Histological view of a Lewy body (arrow) in the substantia nigra pars compacta stained with H & E.

actively concentrated into dopaminergic neurons. Although taken up via the nerve terminals in the striatum, MPP^+ may be concentrated in the cell body through its affinity with neuromelanin. Once inside the cell MPP^+ is concentrated to millimolar proportions by an energy-dependent mitochondrial ion concentrating system. Once within mitochondria, MPP^+ inhibits NADH CoQl reductase (complex I), the first enzyme of the mitochondrial respiratory chain. There is also evidence that MPP^+ generates free radicals and that the nitric oxide synthase inhibitor, 7-nitroindozole, can prevent MPP^+ toxicity in monkeys. The current assumption is that MPP^+ induces cell death through a combination of inhibition of ATP synthesis and free radical generation.

Since the discovery of MPP^+ toxicity on mitochondria, interest has been focused on mitochondrial functions in PD. Current concepts of nigral cell death emphasize mitochondrial dysfunction and impaired oxidative metabolism. The discovery of complex I deficiency in PD substantia nigra raised the possibility that the mutation of genes (nuclear or mitochondrial) encoding complex I subunits might be involved in determining the enzyme's defective activity. There is a reduction of about 35% in complex I activity of the mitochondrial metabolic chain in the substantia nigra pars compacta. Complex I deficiency is not found elsewhere in the brain in PD.

The substania nigra pars compacta is particularly vulnerable to oxidative stress. Dopamine metabolism by monoamine oxidase forms superoxide and hydrogen peroxide. Excess iron in the substantia nigra may promote the formation of the hydroxyl radical from hydrogen peroxide. Dopamine autooxidation in the presence of iron forms superoxide, reactive quinones, and semiquinones. Neuromelanin in dopminergic cells binds ferric iron and reduces it to the reactive ferrous form.

Glutathione (GSH) in its reduced form is an important compound in antioxidant defense and in the repair of oxidized proteins. It is oxidized to its disulfide, GSSG. High GSH/GSSG ratios are maintained by glutathione reductase, which converts GSSG to GSH. There is evidence for decreased activity of glutathione peroxidase in PD nigra, putamen, and globus pallidus, and evidence that GSH levels are decreased in PD substantia nigra.

Superoxide dismutase (SOD) exists in cytosolic SOD and mitochondrial manganese SOD forms and is important in dismutating superoxide ions. Thus, levels of this enzyme are indicative of superoxide generation. Both forms appear to be increased in PD substantia nigra. High levels of copper/zinc SOD are expressed at the mRNA level in control and PD nigral pigmented neurons. Taken together, these observations suggest that PD nigral neurons in particular are exposed to increased superoxide generation.

Levels of polyunsaturated fatty acids, malondialdehyde, and hydroperoxides are increased in PD substantia nigra. These are the products of free radical damage to lipid membranes and imply oxidative damage in PD. Free radical damage to DNA produces intracellular 8-hydroxydeoxyguanosine. Elevated concentrations of this product are seen in PD in the nuclear DNA and particularly in the mtDNA fractions from patients with PD. Levels of these products are also particularly high in control brains in the substantia nigra and striatum, confirming that, even in controls, this area of the brain is a site of high oxidative stress.

Evidence exists for disturbances in oxidative phosphorylation in PD, particularly a reduction in the activity of complex I of the mitochondrial electron transport chain and increased levels of free iron, which may enhance the formation of toxic free radicals.

A. Genetic Subtypes

The greater than expected incidence of PD within family groups has prompted consideration of a genetic factor as playing a prominent role. Studies of large family kindreds have revealed at least seven mutant genes as being present. These genetic abnormalities may only be applicable to the familial form of PD and not to the largest segment of the Parkinson population. However, they may lead to a better understanding of the pathogenesis that underlies the disease. Currently, it is generally held that a genetic predisposition to PD exists due to multiple gene abnormalities, but an additional factor—either an endogenous or exogenous neurotoxin—is necessary to trigger the disease process.

During the past few years, several genes for inherited forms of PD have been mapped and/or cloned. In a large family with autosomal-dominant inheritance and typical LB pathology, a gene locus has been mapped to the long arm of chromosome 4, and mutations in this and a few other families linked to this locus have been identified in the gene for α-synuclein. α-synuclein is expressed abundantly in brain, including the substantia nigra and LBs. First isolated from plaques in Alzheimer's disease, it is found abundantly in LBs

from both hereditary and sporadic cases, in which its expression overlaps with that of ubiquitin. α-Synuclein normally is a soluble unfolded protein, but it can aggregate to form insoluble amyloid fibrils. This common theme of protein misfolding joins PD with a number of other neurodegenerative conditions in which insoluble protein aggregates eventually lead to neuronal death. The recent elucidation of the proteasomal system in protein degradation, and its dysfunction in substantia nigra in PD, lends further credence to this mechanism of protein misfolding as a crucial cog in the etiology of PD.

A gene causing autosomal-recessive parkinsonism of juvenile onset has been mapped to chromosome 6, and the causative gene has been identified and named parkin. This form of parkinsonism, first identified in a Japanese kindred, clinically resembles PD and is associated with nigral degeneration but without LBs.

A third locus, again in families with dominant inheritance, typical LB pathology and late onset, has been mapped to chromosome 2p13, and two additional genes on chromosome 4p have been linked to other dominantly inherited forms of the disease. A genetic locus on chromosome 2p is associated with more typical PD in a larger number of families. Rare pedigrees show an exclusively maternal inheritance pattern consistent with a defect in mitochondrial DNA. Currently, there is no direct evidence that any of the genes for familial parkinsonian syndromes have a direct role in the etiology of the common sporadic form of PD.

VI. SECONDARY FORMS OF PARKINSONISM

The following conditions have in common many features of PD but differ in a variety of ways:

A. Multiple system atrophy: Several subtypes of this neurodegenerative condition exist—pure autonomic failure (Shy–Drager syndrome), striatonigral degeneration, and olivopontocerebellar atrophy.

B. Progressive supranuclear palsy: Gait disturbance is the presenting feature in more than 60% of cases. Personality change is the next most common presenting feature. Gaze palsy, dysarthria, and dysphagia are also prominent. Contracted facial muscles, retrocollis, predominantly proximal rigidity, predominantly vertical supranuclear gaze abnormality, spastic dysarthria, dysphagia, and behavioral abnormalities complete the list of signs of this progressive neurodegenerative disease.

C. Cortical basal ganglionic degeneration (CBGD): CBGD encompasses features pointing to dysfunction in both the cerebral cortex and basal ganglia. It was first described in the 1960s. Symptoms include those of typical PD and limb dystonia. Tremor may be postural or kinetic, and reflex myoclonus is common. Cortical signs include apraxia, cortical sensory loss, and alien limb phenomenon. Other signs include athetosis, orolingual dyskinesias, frontal lobe reflexes, impaired ocular movements, dysarthria, speech apraxia, dementia, hyperreflexia, and Babinski's signs. Findings are highly asymmetrical. Asymmetric cerebral atrophy is seen on the medial frontal and parietal lobes.

D. Parkinsonism–dementia complex of Guam: A high incidence of parkinsonism has been reported among the Chamorro population of Guam. Signs are similar to those for PD, except that tremor is not prominent. It is invariably accompanied by progressive dementia. The majority of patients also have clinical signs of motor neuron disease, and some have features similar to those of progressive supranuclear palsy.

E. Diffuse Lewy body disease.

F. Drug-induced parkinsonism: Dopamine-blocking agents have been implicated in causing a condition that resembles PD. Such drugs include phenothiazines, butyrophenones, and metoclopramide. This condition represents one of the most common causes of secondary parkinsonism.

G. Toxin-induced parkinsonism (MPTP, manganese, carbon monoxide, and cyanide)

H. Hemiatrophy hemiparkinsonism

I. Postencephalitic parkinsonism: Between 1916 and 1927, a worldwide epidemic of encephalitis lethargica (sleeping sickness) took place, affecting as many as 750,000 people. Approximately one-third of them died acutely, one-third recovered completely, and the remainder were left with chronic neurological deficits. Parkinsonism following encephalitis lethargica can be indistinguishable from the manifestations of PD.

J. Vascular: Multiple lacunar infarctions involving the corpus striatum are seen on MRI. Deficits usually occur in a stepwise progression.

K. Trauma: Parkinsonism is a rare sequela of head trauma. Repetitive trauma seen in boxers may result in progressive parkinsonism as well as dementia (dementia pugilistica). Diffuse neuronal loss and neurofibrillary tangle formation are seen on pathology.

L. Tumor: All the clinical features of parkinsonism may occur as a consequence of brain tumors. They are most commonly supratentorial meningiomas. The

mechanism may be mass compression or invasion of the corpus striatum.

VII. TREATMENT: CURRENT APPROACH, PHARMACOLOGICAL AGENTS, AND SURGERY

Treatment of PD is currently directed to control of symptoms. There are no curative measures, nor any agents that retard the progressive nature of the disorder. As is the case for instances of symptomatic treatment, the decision to utilize therapeutic agents depends on the severity of symptoms and the degree of disability.

A. Medical Treatment

The most effective pharmacological agents for the control of Parkinson's symptoms are those capable of restoring dopaminergic activity in the striatal region of the brain. Of the many drugs having this capability, levodopa administered orally with a peripheral DOPA-decarboxylase inhibitor (e.g., carbidopa) is the most effective. This combination blocks the conversion of levodopa to dopamine outside the nervous system, making it completely available for use in the brain. Given the appropriate dosage more than 80% of patients will experience beneficial effects in the form of symptom control. Such results are usually optimal during the first 3–5 years of treatment and then tend to wane off. Not only does the therapeutic effect become less optimal but also erratic responses occur in addition to the development of adventitious movements of the body and not infrequently behavioral abnormalities. Therefore, it is generally believed that treatment should be withheld until it is a necessity and the dosages used should be kept at a minimum.

The L-DOPA effects may be enhanced by the addition of a number of drugs that inhibit the metabolic pathways of dopamine; for example, *l*-deprenyl, an agent capable of inhibiting monoamine oxidase B, may prolong the effects of a given dose of levodopa, and entacapone or tolcapone, which inhibit the methylation pathway of dopamine degradation, not only prolong its effect but also give a modulated pharmacological response.

Despite complications in the use of levodopa, it has had a profound effect in maintaining the functional capacity of patients with PD for many years beyond that previously experienced. Furthermore, life expectancy has been markedly prolonged for such patients since its introduction.

Other agents utilized to enhance striatal dopaminergic function are agonists of the postsynaptic dopamine receptors. These consist of pergolide, ropinirole, pramipexole, and bromocriptine. They may be used as substitutes for levodopa or in combination with it. These agents may have the advantage of inducing less adventitious movements but are also less potent at restoring dopaminergic function.

A number of ancillary drugs are utilized in a limited fashion and primarily to delay the use of dopaminergic agents. Amantadine is an antagonist at the NMDA-type receptor of the excitatory amino acid system. Amantadine has been shown to have some beneficial effects on tremor in addition to decreasing the severity of L-DOPA-induced dyskinesias. A number of anticholinergic agents, such as trihexyphenidyl and benztropine, have been used for years because they are mildly effective and may be used to delay the use of dopaminergic drugs.

B. Surgery

A number of surgical approaches to the treatment of parkinsonian symptoms have been developed. In general, they have been reserved for patients who have failed pharmacologic agents or whose response to medications has led to a variety of untoward reactions.

The first surgery performed for PD involved excision of parts of the cerebral cortex for parkinsonian tremor and dystonia by Bucy and Case (1939) and Klemme (1940). This surgery produced spastic hemiparesis, however. Corticospinal tract lesioning was performed for relief of dyskinesia (Putnam, 1938), which also produced ipsilateral hemiplegia. Surgical lesioning in deeper structures within the cerebral hemispheres for movement disorders was introduced by Meyers (1940). Lesions in both caudate nucleus and globus pallidus were performed. Initial results found improvement in both tremor and rigidity in 40% of patients with parkinsonism. Dyskinesia improved as well, without accompaniment of weakness or spasticity. In 1953, Cooper accidentally cut the anterior choroidal artery, followed by its ligation, which significantly relieved tremor and rigidity. He subsequently performed anterior choroidal artery ligations in a series of more than 50 patients, with benefit in tremor and rigidity in two-thirds of patients, with a mortality rate of 10%.

Stereotactic neurosurgery, developed in the 1950s, allowed for more accurate targeting of deep brain nuclei. Microelectrode recording techniques, developed

in the 1960s, further improved localization of targets for lesioning. Neurosurgical therapy for PD reached its peak in the 1960s but declined after the introduction of levodopa. The past decade has brought about a resurgence of neurosurgery for PD with improved localization techniques and lower morbidity/mortality rates. Hyperactivity of the subthalamic nucleus in PD leads to increased excitation of the GPi and SNr, which in turn inhibit the motor projections to the thalamus and brain stem. The STN and Gpi have thus become the preferred targets for surgery to treat PD.

There are several procedures currently in use.

1. Pallidotomy

A lesion is made in the ventroposterolateral part of the internal segment of the pallidum (GPi), which is overactive in PD. This lesion also destroys the pallidothalamic pathway (the ansa lenticularis). The major benefit of this procedure is a reduction of contralateral levodopa-induced dyskinesias. Reduction of tremor, rigidity, and bradykinesia is also seen but to a lesser degree.

2. Thalamotomy

Thalamotomy is most efficacious for tremor, with a success rate of approximately 95%. It is much less effective for bradykinesia, which is usually more disabling. The target for treating tremor is the ventral intermediate nucleus of the thalamus. Selective ablation of thalamic nuclei and pallidal regions has had inconsistent beneficial results.

3. Deep Brain Stimulation

Chronic deep brain stimulation (DBS) has been used for treatment of parkinsonian, essential and cerebellar outflow tremors. In addition to parkinsonian tremor, deep brain stimulators have recently been implanted in the GPi and STN to treat other motor signs of PD. They have been reported to improve all the cardinal motor symptoms, including tremor, rigidity, bradykinesia, and gait, as well as drug-induced dyskinesias and motor fluctuations. DBS simulates the effects of a lesion without the need to make a destructive brain lesion. In this procedure, an electrode is implanted in the brain target and connected to a subcutaneous pacemaker. The STN is overly active in patients with PD. Because surgical lesions of the STN run the risk of the patient developing hemiballism, electrical stimulation of the STN has been carried out. The long-term effect of these procedures has not been defined.

4. Transplant

Transplantation of embryonic mesencephalic dopamine cells into the striatum of patients with PD is being investigated as potential therapy. It is not possible to make a definitive conclusion concerning the results. Techniques utilizing stem cell transplantation are also currently under investigation.

See Also the Following Articles

BASAL GANGLIA • CEREBRAL PALSY • DOPAMINE • MOTOR CONTROL • MOTOR SKILL • NEURODEGENERATIVE DISORDERS • THALAMUS AND THALAMIC DAMAGE

Suggested Reading

Ballard, P. A., Tetryd, J. W., and Langston, J. W. (1985). Permanent human parkinsonism due to 1-methyl-4-phenyl-1,2,3,6-tetrahydropyrydine (MPTP): Seven cases. *Neurology* **35,** 949–956.

Cohen, G., and Spina, M. B. (1988). Hydrogen peroxide production in dopamine neurons: Implications for understanding parkinson's disease. In *Progress in Parkinson Research* (F. Hefti and W. J. Weiner, Eds.), pp. 119–126. Plenum, New York.

Costa, E., Cote, L., and Yahr, M. D. (Eds.) (1969). *Biochemistry and Pharmacology of Basal Ganglia Disease*. Raven Press, New York.

Di Rocco, A., Molinari, S. P., Kollmeier, B., and Yahr, M. D. (1996). Parkinson's disease: Progression and mortality in the L-DOPA era. *Adv. Neurol.* **69,** 3–11.

Hoehn, M. M., and Yahr, M. D. (1967). Parkinsonism: Onset, progression and mortality. *Neurology* **17,** 427–442.

Hornykiewicz, A. (2001). Dopamine and Parkinson's disease. *Adv. Neurol.* **86,** 1–12.

Jellinger, K. (2001). Pathology of Parkinson's disease. *Adv. Neurol.* **56,** 55–72.

Jenner, P., and Olanow, C. W. (1996). Oxidative stress and the pathogenesis of Parkinson's disease. *Neurology* **47,** 161–170.

Lang, A. E., and Lozano, A. M. (1989). Parkinson's disease. *N. Engl. J. Med.* **339**(15), 1044–1053; **339**(16), 1130–1143.

Parker, W. D., Byson, S. J., and Parks, W. D., Jr. (1989). Abnormalities in the electron transport chain in idiopathic Parkinson's disease. *Ann. Neurol.* **26,** 719–723.

Schapira, A. H. V., Cooper, J. M., Dexter, D., Clark, J. B., Jenner, P., and Marsden, C. D. (1990). Mitochondrial complex I deficiency in Parkinson's disease. *J. Neurochem.* **54,** 823–827.

Tanner, C. M., and Goldman, S. M. (1996). Epidemiology of Parkinson's disease. *Neuroepidemiology* **14,** 317–335.

Yahr, M. D. (Ed.). (1976). The basal ganglia. *Res. Publ. Assoc. Res. Nervous Mental Dis.* **55**.

Yahr, M. D. (1990). Principles of medical treatment. In *Parkinson's Disease* (G. Stern, Ed.), pp. 495–508. Chapman & Hall, New York.

Yahr, M. D. (1993). Parkinson's disease—The L-DOPA era. *Adv. Neurol.* **60,** 11–17.

Pattern Recognition

WILLIAM R. UTTAL
Arizona State University

GLOSSARY

association A hypothetical second stage of the pattern recognition process in which the represented image is connected to a name or category.

elementalism A theoretical position stressing the precedence of the parts or local features of an object.

holism A theoretical position stressing the precedence of the entire object or the arrangement of its parts.

representation A hypothetical first stage of the pattern recognition process in which the incoming image is transformed or coded for the subsequent association.

stimulus equivalence The ability of the human to recognize objects in the face of severe transformations in shape, orientation, position, or size.

Pattern or form recognition, in the context of this article, is the process by which visually presented objects are identified, categorized, and named. In other words, the challenge of pattern recognition is to answer the question ''What is it that I am seeing?'' This is a different visual question than that asked in detection, ''Is there anything there,'' or in discrimination, ''Are the two things I see the same or different?'' This definition of recognition specifically excludes the even higher cognitive processes by which the categories, inclusive concepts, or names of the objects are themselves created. Rather, the topic is treated purely as a problem in visual perception in this article. Two hypothetical stages of processing can be distinguished: Stage 1, the transformation and representation of the image into a form suitable for stage 2, the comparison or analysis that permits a concept, name, or category to be attached to or associated with the image.

This article is adapted and updated from one of the same title previously published in *The Encyclopedia of Human Behavior* (1994) by Academic Press.

I. INTRODUCTION

The problem of form or pattern recognition is a classic one in visual psychology. From the time of the classic Greek philosophers until today, it has been appreciated that the brain is capable of carrying out amazing feats of object recognition and identification. Despite its antiquity, the psychological problem is still generally considered to be unsolved and a topic of active research in psychology, computer science, and, to an arguable degree, neurophysiology. As we shall, see many controversies and uncertainties currently exist in this field.

It is clear now that one of the easiest ways to conceptualize the problem is to propose that at least two hypothetical functional components—*image representation and memorial comparisons or associations*—must be involved in the recognition process. The object, image, or scene must be represented in some way so that it can be either compared with an existing set of alternative stored categories or processed by some other kind of logic that allows naming or classification to occur. It is important to appreciate that both possible means of association are plausible

explanations—a structural comparison process on the one hand and a logical analysis on the other. If there is any "law" of visual perception, it is that the visual system is capable of using many different strategies to meet the recognition challenges it confronts in daily life.

The true nature and complexity of the visual aspects of the human pattern recognition problem were not clarified until contemporary computer scientists such as D. Marr began to cope with the challenge of producing comparable image processing programs. The requirement that the computer program contain a precise statement of the necessary processing steps forced workers in this field to consider in detail the equivalent processes that are likely to be carried out by the human. In this context, we must appreciate that the study of pattern recognition is a thoroughly modern interdisciplinary issue that transcends psychology, neuroscience, and computer science. Pattern recognition is not, however, a monolithic problem. A number of subsidiary issues have emerged that have attracted significant research attention. In the remainder of this section, some of the most important subtopics will be introduced.

A. The Whole versus Part Controversy

One of the major issues, if not the major issue, in developing a biologically valid theory of pattern recognition concerns the initial image processing strategy used by the viewer to represent the image. The specific question is as follows: Is the image analyzed into its parts, or is it examined as an organized whole prior to and during the recognition process? All computer models and psychological theories must be based on at least an implicit *a priori* judgment that pattern recognition is governed by either the nature of the parts or the arrangement of the parts. In contemporary fact, however, most modern theories are elementalist and assert that the features from which the image is constructed are precedent in recognition. A minority position—holism or Gestaltism—holds that the arrangement of the parts is more important than their nature.

The probable main reason for the predominance of feature theories is that pattern recognition research exists in a context of developments in other highly elementalist sciences. Contemporary theory and experimentation in neurophysiology, on the one hand, and computer technology, on the other, have both had an immense effect on the field. Modern neurophysiology is primarily a science of components and parts—the neurons of which the brain is made and which carry out the mysterious information processing that accounts for pattern recognition. Existing computer programming techniques also tend to force us to examine and manipulate the parts of an image detailed down to the individual feature or even the individual pixel. Conceptually, it "demagnifies" the problem down to one emphasizing elemental parts rather than global patterns. It also must be acknowledged, in the light of this criticism, that another major impetus toward the predominantly part- or feature-based theory orientation in this field is that we do not yet have a satisfactory holistic theory of arrangement, a sufficiently powerful holistic empirical methodology, or even a suitable mathematical tool to provide the bases for a compelling theory of recognition based on global rather than local attributes.

Most perceptual experiments reported in the current scientific literature manipulate features or parts of the stimulus image. Data from such studies, therefore, also conceptually support theories of pattern recognition that are based on the nature of the features or components of which the stimulus image is composed. The intellectual system so created is, therefore, essentially circular. Experimental designs are based on feature-based theoretical premises; the experiments inevitably then produce results that support the initial hypothesis. The initial theory then becomes reified as a valid statement of reality rather than what it was originally intended to be—a working hypothesis or a convenient metaphor.

Nevertheless, there is an increasing, though still relatively small, body of evidence that argues, to the contrary, that global or holistic strategies dominate human pattern recognition instead of feature-based ones. One such body of evidence deals with the manifestation of the final phenomenal outcome itself—the exemplar demonstration or visual phenomenon displaying the outcome of the entire perceptual process. Demonstrations (as opposed to parametric experiments) are all too often ignored by researchers in this field when it comes to generating theory. Nevertheless, this kind of "first-order" phenomenology should not be underestimated. Demonstrations reflect the overall nature of perception in a direct and immediate fashion and should set the stage for research and theory rather than being discarded as irrelevant. The work of V. S. Ramachandran, S. Antsis, and N. Wade, among others, utilizing illusions as probes of recognition mechanisms, is especially notable in running counter to this trend.

One argument against a pure part- or feature-based theory of pattern or object recognition is based on the fact that the specific features of which objects are composed often do not seem to be very important in the recognition process. For example, the letters of the alphabet or chairs of various kinds can be recognized as exemplars of their general class (an A or a chair) regardless of the font or specific shapes of the pieces of which they are made. Caricatures consisting of only a few well-chosen lines can easily be identified as representing a particular person. The essential emerging fact is that no specific feature is necessary for the recognition of an object. The relationship of the features (i.e., the object organization) seems to be the key to pattern recognition, and virtually any feature can serve as the salient cue for recogntion.

Another classic body of evidence that supports holistic or global precedence in pattern recognition can be found in the literature of Gestalt psychology. Many of their rules of the organization of visual perception, in particular the idea that forms have a certain fundamental unity or *Pragnanz*, speak strongly for the priority that the global, whole attributes of an image must have in many aspects of visual perception. Other experiments that deal with stimuli that are virtually free of any continuous features (for example, arrangements of dots or subjective contours) also support a holist point of view.

The ability to recognize an object from a vast number of viewing angles and under a vast number of distortions is referred to as *stimulus equivalence*. Stimulus equivalency reflects the extraordinary capacity of an observer to recognize an object (for example, a plate or a face) as an example of a particular class of objects even though the parts may be substantially distorted by the variations in the viewer's point of view or even partially occluded. It is also considered by many theoreticians to be strong evidence for a holistic interpretive process rather than one driven by local features.

Despite the vigor of the debate between the holist and elementalist theorists, a prudent review of the contemporary scene reveals that the issue is not yet resolved. Repeated demonstrations, experiments, and caveats have supported the precedence of *both* the arrangement of the parts into a meaningful global pattern and the specific nature of the parts. This may signal that we cannot yet definitively specify which of these theoretical points of view best describes this powerful human perceptual skill. As in many human perceptual problems, it may be that there is no single all-inclusive answer to this conundrum. Rather, the method used by the human observer in each case depends upon a number of situational and environmental variables. Another possibility is that the problem is actually irresolvable. Several researchers such as C. Latimer and S. Rakover have considered this possibility.

B. Do Pattern Recognition Theories Explain Human Recognition Activity?

Without delving too deeply into the philosophical foundations of this problem, it should also be noted that cognitive and behavioral psychologies have radically different points of view concerning the accessibility of the underlying human thought processes. Behaviorists look at all psychological theories as being, at best, descriptions of the phenomena. Cognitivists see theories as truly explaining the underlying working mechanisms that are responsible for the observed behavior. Empirical evidence is strongly on the side of the behaviorist position; in general, people cannot give reliable or valid reports of their underlying thought processes any more than they can describe the neural mechanisms that account for those mental events. Many other arguments also assert the neutrality of theories and emphasize their descriptive rather than explanatory nature. Many of the computer-based or mathematical theories discussed in this article must, therefore, be considered to be process descriptions, regardless of how well they fit the empirical data or how useful they are as engineering tools.

There is, it should be noted, a major controversy currently raging concerning the ultimate resolvability of the problem of internal representation that may make all of these models interesting process descriptions rather than true reductive explanations. A number of psychologists have argued that the representation of the image within the visual system is indeterminate through either behavioral or neurophysiological techniques. Behavior is not a satisfactory probe because it is neutral with regard to the internal mechanisms in a fundamental way; neurophysiology is not satisfactory because of the enormous complexity of the coding schemes used by the nervous system to represent spatio temporal patterns beyond the peripheral portions of the sensory pathways. These arguments raise doubts about the utility of, for example, single-cell neurophysiological research for contributing to a satisfactorily complete solution of the human visual pattern recognition problem. Early studies that showed lateral inhibitory interconnections in simple

peripheral structures such as the horseshoe crab eye may have been misleading harbingers for the future. Whereas interneuronal inhibition may have been a good model of the Mach Band, most pattern recognition processes certainly take place at high and complex neural levels and are poorly served by explicitly neural models of this genre. A trend toward computational models led by the thoughtful contributions of computer scientists such as D. Marr, T. Poggio, and W. E. L. Grimson based on transforms, operations, and processes (rather than neural structures) seems more promising. Even the development of neural network or connectionist theories of pattern recognition is now appreciated to be that of process models describing interactions between symbolic nodes rather than between anything close to biologically realistic neurons.

In summary, the problem of going from behavioral observations is what mathematicians denote as an *ill-posed* problem. That is, there is not enough information in the available data to backtrack to the underlying mechanisms. It is only by assuming certain constraints that the problem becomes tractable. That assumption, however, is the equivalent of adding additional degrees of freedom and, thus, detaching any theory from a unique explanation of the underlying psychobiological reality.

II. THEORIES OF PATTERN RECOGNITION

As noted earlier, contemporary theories of pattern recognition are mainly dedicated to answering two kinds of questions. The first concerns the initial transforms that are carried out on the input image. The foundation issue in this case is the nature of the representation of the image information. The second question concerns the process by which the transformed images are linked with a particular prototype name. The foundation issue in this case is the nature of the association between the incoming image, transformed and altered as it may be, and a name or category learned at some previous time.

A. The Representation Problem

Many modern pattern recognition theories that concentrate on the visual process take for granted that, if the image is appropriately represented, the problem is essentially solved, the association of the appropriately represented image with a particular name being a trivial final step. (Of course, for those interested in the higher cognitive and linguistic processes, this second stage of the pattern recognition problem is central.) Much of the emphasis by many contemporary pattern recognition theorists is, therefore, on image transformations and representations prior to the comparison process.

The image transformation process is by no means simple or immediate. It is, itself, a major challenge to explanatory theory. Human vision is wonderfully adaptive. At some stage, it seems that the human pattern recognizer normalizes the stimulus so that, even when an object is rotated, translated, or magnified over wide ranges, it can still be recognized. (This invariance is another way of defining stimulus equivalence.)

Most computer models cum theories, as well as psychological models of perception, usually include some preliminary normalization to a canonical configuration or to an invariant representation. (This is especially true for connectionist or neural net models.) If the normalization is done properly, recognition is not dependent upon the particular situational properties of the stimulus, and the model mimics human recognition invariance in a reasonably complete way. For example, simply transforming a stimulus to a polar, as opposed to a Cartesian, coordinate system is one useful means of establishing invariance to rotation and magnification. It is also possible to transform a stimulus to a completely different representational system such as a spectrum of spatial frequencies, for example, a Fourier or a Walsh transform, to provide another means of precluding any sensitivity to irrelevant spatial translations.

In spite of the ubiquitous nature of this theoretical approach, whether such a standardization or canonical transformation of the stimulus actually occurs in human pattern recognition remains an unresolved question. At a behavioral level, human recognition skills exhibit a profound insensitivity to an object's location or its size. It is conceivable, however, that the organic analysis system is so powerful that this minor miracle can be accomplished without any image normalization of the kind to which computational theorists usually retreat. Such a preliminary modification of the image may merely be a convenience, if not a necessity, for the computer modeler or psychological theoretician because of our incomplete understanding of the later stages of processing. The difficulty of solving problems without some kind of a fixed frame of reference may be much less for the human visual system than for the computer program.

The mathematical problem of defining a canonical coordinate system to achieve good invariance to the various distortions and displacements is not trivial. However effortlessly nervous systems seem to adjust to changes in stimulus position and shape, the general problem posed to the modeler or theoretician whose goal is to describe human pattern recognition is profound, refractory, and clearly not yet solved.

B. The Association Problem

The next stage in the pattern recognition process, following the initial image transformation and representation stage, requires that the modified image be associated with a prestored name or category. This second stage is even more mysterious in terms of the specific underlying mechanisms than the representation stage, even though it has been the object of considerable psychological research. There are, again, two major schools of thought concerning the nature of this second stage. In the first, the process is considered to be a simple correlational comparison with a library of templates, prototypes, or reference images. This type of comparison process depends upon at least a topological kind of isomorphic representation in which the geometrical relationships among the parts of the image are maintained. Depending upon the theoretician's proclivities toward a holist or elementalist approach to the first stage, the second stage may invoke (a) the triggering of a network in which decisions are made concerning the presence or absence of a feature or component until a final "recognition" occurs or (b) the comparison of the whole picture with the library of prototypes until a best fit is achieved. Preservation of the topological relations of the parts of the stimulus image, at least, is assumed in either case, if not the specific Cartesian geometry of the original image. That is, the image objects are assumed to be represented by a nervous system that preserves the spatial characteristics of the original stimulus. The match is made by a maplike superimposition, although on maps that may be very elastic. Size, orientation, and even shape may be distorted, depending upon the association procedure, as long as some semblance of topological order is maintained.

The other approach to name association (the second stage of the pattern recognition process) does not involve preservation of the topological and geometrical relations of the original image and its comparison with the prototypes in an isomorphic manner. Rather, the process may be symbolic and the analysis carried out as a series of logical decisions, analyses, or constructions. That is, an object may be represented in a nonisomorphic code that describes the object as a series of logical steps or in a descriptive language that specifies the rules that would allow the object to be reconstructed. From this point of view, the comparisons are not between maps that are at least topologically constant, but rather between symbol systems. This latter approach, in fact, does not require any kind of isomorphic comparison. It permits an alternative strategy in which the descriptive language or logical construction process acts as a decision tree, directing the recognition process to a category or name for the object. Such a process is intrinsically faster than one that requires an exhaustive search (such as template matching) and can be implemented in a more economical manner than one requiring many multiple processes to be carried out in parallel.

A substantial body of work associated with pattern recognition utilizes the dichotomy of attentive and preattentive processes as an assay of recognition processes. The work of B. Julesz and A. Triesman is especially notable. In each case, they have distinguished between dimensions of a stimulus object that preattentively "pop out" and those that must be attentively scrutinized to become effective in the recognition process.

C. Some Contemporary Pattern Recognition Theories

Theories of pattern recognition come in many forms and stem from many different traditions. The best way to concisely include a broad range of pattern recognition theories in this article is to provide capsule descriptions of the ways in which theories are categorized. The brief typologies presented here are based upon comprehensive taxonomies of pattern recognition methods suggested by J. T. Townsend, D. E. Landon, F. G. Ashby, S. Watanabe, S. Pinker, and the work of other contemporary metatheoreticians.

Townsend and his collaborators, Landon and Ashby, are mathematical psychologists and approach their taxonomies from the point of view of human perceptual theories. All members of the four classes of recognition methods that these authors describe are *descriptive*, in that no allusion to the possible physiological or cognitive mechanisms underlying the pattern perception process is made. They are also mainly based on statistical rather than deterministic kinds of mathematics.

Townsend and his colleagues divide pattern recognition theories into two major subdivisions. The first subdivision includes those pattern recognition theories that are "based on an internal observation." Members of this class of pattern recognition theories contend that each stimulus event is dealt with separately by the perceptual processing system. The probability of a correct recognition (i.e., a response with the correct name of the stimulus or the name of the appropriate category) depends, therefore, upon the evaluation of that stimulus item by a set of internal rules and criteria couched only in the terms of a particular stimulus presentation event. This class of theory attempts to describe the specific processes (e.g., feature detection) that are presumed to exist within the cognitive structure of the observer as an incoming image is processed.

The second major subdivision includes those theories that are designated as "descriptive," with a narrower interpretation of the word than that just used. In this case the role of the individual event is minimized. Instead, the process of recognition is modeled as a kind of guessing or choosing an item from a set of possible responses on the basis of probabilistic rules involving context properties that go far beyond the immediate event. Rather than processing the attributes of a single stimulus, as the internal observation theories did, this second class of models merely uses the stimulus as one of many influences leading to an appropriate guess or choice of the proper response by the observer.

The first major division made by Townsend and his colleagues—the "internal observation" category—is further broken down into two subdivisions, the *general discriminant* models and the *feature confusion* models. General discriminant models are characterized by decision rules and procedures that evaluate the attributes of a particular stimulus and calculate a numerical value or "discriminant" for all possible responses that could conceivably be associated with that stimulus. The largest numerical value associated with any possible response becomes the selection criterion leading to the emission of that correct naming response. Feature confusion models assume that a matrix is actually constructed that tabulates the probabilities of confusing specific responses.

The second major kind of pattern recognition model described by Townsend and his colleagues—feature confusion models—is also divided into two subcategories. The first gathers under a single rubric called sophisticated guessing type models, consisting of a number of different theories including the sophisticated guessing models themselves, *all-or-none models*, *overlap models*, and *confusion-choice* models. The second subcategory, which they designate the choice category, includes but a single exemplar: the *similarity choice model*. As noted, all of these theories are completely descriptive: they make no effort to consider the neural or cognitive processes underlying the statistical formalities. Rather, they involve the statistics of active decision making, choice behavior, or stimulus similarity as their conceptual basis.

Engineers, of course, approach the problem of designing models of pattern recognition from a different point of view. They often invoke a number of different kinds of techniques that do not depend on and are not limited to the observed peculiarities of human perception. Watanabe points out that techniques such as *entropy minimization*, *covariance diagonalization*, and *structural analysis* are often used in that field. Interestingly, though the two traditions have grown up separately, inspection of the respective models indicates that there may be more similarities than their different names and origins suggest. For example, the category of covariance diagonalization is hardly distinguishable from *factor analysis*, terminology with which psychologists are more familiar. In a similar way, some of these methods are comparable to *multidimensional scaling*, which also seeks to reduce large amounts of information to a smaller set of nonredundant measures. It is clear that there has been a vigorous cross-fertilization of pattern recognition models between engineering and psychology; data from psychology often lead to insights for the engineer, and ingenious engineering inventions provide the framework for psychological thinking about the pattern recognition process.

In this context of interdisciplinary interaction, it is interesting to note that the engineering methods are virtually always feature-based. In some cases, such as the structural analysis method, the features are the same as the those defining the geometry of the image—the corners, sides, etc. But in others, such as discrimination and decision-making procedures, a list of geometrical features is not necessary: any collection of attributes, spatially absolute (e.g., square) or relational (e.g., larger), can be used to define a pattern. Patterns may even be dealt with in the abstract. That is, measurements of virtually any kind can be used to set up "vectors" or collections of measurements that can compare with each other or with stored nongeometrical (i.e., symbolic) prototypes to recognize patterns of other kinds.

Other theories of pattern recognition are intuitively more straightforward. For example, the classic

template matching theory assumes that an incoming visual image is compared with a library of prestored dimensionally isomorphic images, each of which has a name already associated with it. Template theories are correlative: the best match of the incoming image with one of the stored library of templates determines which one will be associated with the incoming image. In their most basic form, template models are essentially global or holistic. However, the fact that they require enormous libraries of prestored images has always mitigated against them. Because the phenomenon of stimulus equivalence suggests that it is not even necessary to have previous experience with a particular view of a form (and thus to have an exact prestored image) for recognition to occur, it is not likely that template matching in its simplest version is a likely candidate to be the best model of human pattern recognition.

A theory that models some kind of logical decision-making system is a more plausible candidate to explain human pattern recognition. That is, one that does not require exhaustive searching for the best match in an extensive library would be preferred over one in which the attributes of the incoming image direct the process through a decision-making tree to a final recognition point.

Many neural-network-type theories operate in this mode. One of the classic neural network theories was developed in 1959 by O. Selfridge. His *pandemonium* model assumed that an array of specialized form detectors, sensitive to local geometrical features of the stimulus, become active to the degree that the feature was present. The pattern of activity constitutes a "vector" of activity that directly determines which name will be associated with the stimulus. No correlation with prestored images was necessary—only the response names themselves, which were located, in a conceptual sense, at the end of the decision tree through which the classification logic proceded.

In more recent years (after a hiatus stimulated by an important critique by M. Minsky and S. Papert in 1969), pattern recognition theories based on the operation of neural networks have become increasingly popular. This renaissance in network or connectionist-type theories can be directly attributed to the extraordinary impact of the twin books on parallel distributed processing by D. E. Rumelhart, J. L. McClelland, and their colleagues. Some of these multilayer neural networks or *perceptrons* (a term invented by F. Rosenblatt) type theories include training processes so that a particular output can be associated with a particular input by repetitive reinforcement as the network experiences incoming stimuli. The effect of reinforcement in this type of model is to modify the strength of connections between nodes in the network. (The nodes have been conceptualized as either neurons or higher order symbol processing units at various times in the history of neural networks.) The final outcome of this progressive change in the organization of the network is reminiscent of the pandemonium model: logical routing from multiple, alternate stimulus pattern inputs to appropriate singular response outputs. The distinctive property of modern neural network models, however, is the nature of the adaptive processes by means of which the weights between nodes can be modified.

Most of the theories that have been considered so far use the features or attributes of the stimulus in the same space as originally presented. For example, an angle defined in a Cartesian (x, y) coordinate system is dealt with in terms of that geometry. Currently, there are other very important theories of pattern recognition and vision that base their action on a significantly different foundation: spatial domain transformations. That is, these alternate models transform the stimulus from its original spatial representation to another one that has some other more general, useful, or computationally convenient property. For example, a stimulus originally presented in the familiar (x, y) coordinate space of the visual display may be transformed into a spatial frequency (u, v) space by means of a two-dimensional Fourier transform.

There are two disadvantages to the use of images defined only in the x, y space. First, the original picture may be represented as a collection of component features that can only be laboriously transcribed or exhaustively tabulated. Second, the original picture is a multidimensional pictorial representation that is not quantified in a way that would permit calculations to be made on it. The Fourier-transformed image, however, varies along only three dimensions: spatial frequency, phase, and amplitude. Therefore, this kind of transformation (which usually produces a picture that appears to be more complicated than the original one) also produces a numerically manipulable and computable spectrum or vector (of the frequency components) from the original unquantified pictorial representation. This is the main advantage and *raison d'etre* of the Fourier technique: it takes geometrical forms that are not represented quantitatively and represents them in a form on which computations can be made.

Fourier transforms, or any of the many other similar transforms that are now available, are merely a means

of representing the image. Representation of an image in the (u, v) domain is a process that does not speak at all to the association problem, a matter that must be dealt with separately in a subsequent processing stage. This is not unusual. Theories purporting to be "explanations" of pattern recognition often turn out to be nothing more than methods for representing objects in a computationally useful form. For example, there is a closely related field of computer vision called visualization. In this field, cognate to the pattern recognition field we have been discussing, the task is to draw good images on a display. Theories of image representation aimed at solving this problem describe how complex geometrical forms can be reduced to simple standard shapes. D. Marr and H. K. Nishihara, as well as A. Pentland, have made major contributions to this field by suggesting ways in which images may be represented as sets of basic solid shapes or linear axes.

Other approaches to modeling human pattern recognition and/or developing new engineering techniques have been tried: progress in more advanced forms of neural networks has continued; pruning algorithms and other tree methods have also been advanced; progress has been made in distinguishing between semantic and syntactic methods; and novel new approaches using such mathematical esoterica as fuzzy logic, Bayesian logic, Markov logic, and Dempster–Shafer theory have been explored. Obviously, there is continued development in methodology even if the mechanisms by which the brain recognizes patterns remain obscure.

III. EXPERIMENTAL RESEARCH

The whole–part question continues to be one that attracts a considerable amount of research attention. D. Navon, for example, has worked with stimulus patterns that are composed of two levels of features. The whole stimulus pattern defines one form, whereas the parts of which it is composed consist of another. The question is, "Which level of processing dictates what is recognized?" Navon's experiments, along with those of S. J. Lupker and H. Bourna, all support the idea that the global form is precedent.

Results with stimuli that varied in their degree of symmetry carried out by H. Pashler and B. Jenkins also speak for the holistic theoretical point of view. This type of experiment typically showed that a symmetrical global arrangement of the parts of the stimulus made for much more efficient recognizability. Because the same features were present in each case, once again arrangement is implicated as the key variable in recognition.

The argument for holistic precedence is made even more profoundly made when stimuli are chosen in which the features are not even present. This can be done, for example, if one uses dot patterns, a stimulus form in which there are no continuous geometrical features present, only arrangements of the featureless punctate component. In an extensive series of experiments, my colleagues and I have shown that the arrangement of the dots is critical in pattern recognition. Even more important was the fact that certain specific arrangements were shown to have particular perceptual potency. Specifically, a general rule—*the law of linear periodicity*—which asserts that the best recognized arrangement is a straight line of evenly spaced dots, explained a very large number of different experiments, in both spatial and dynamic environments.

Another class of experiment in which support for the holistic point of view is obtained is generically classified as the *object* or *word superiority* paradigm. In this procedure, a comparison is made between the recognizability of a letter or an oriented line either isolated or in a context—a word or a stick figure. G. M. Reicher, working with letters and words, and N. Weisstein, M. C. Williams, and C. S. Harris, working with lines and outline objects, all found that the lines and letters were better recognized when embedded in a relevant context than when isolated. It seems, therefore, that the arrangement of the stimulus and its context play primary, if not definitive, roles in the recognition process.

During the last decade of the twentieth century, there has been a subtle shift in the research program. Nowadays the words *object perception* (a sidelight only a few years ago) have taken over from the phrase *pattern recognition*, suggesting an increasing emphasis in research on the recognition of three-dimensional objects. One area that has been especially active has been the study of how we recognize faces. Stimuli of this kind are hard to quantify, and much of the salient research has manipulated the integrity of the face, moving, deleting, or rearranging parts and components.

A considerable amount of work is also being done on how attributes such as color, movement, and texture affect recognition. A considerable amount of evidence indicates that coherent motion may make recognizable what is otherwise totally hidden in visual clutter. Color has been shown to have an important influence on objects for which the color is particularly

diagnostic, such as a banana. Other specific parameters and attributes of an object have also been shown to be influential in their recognition. However, much of this research raises new possibilities about the semantic or high-level processing of cues that goes far beyond the prevailing low-level processing theories that were stimulated by important, but possibly irrelevant, discoveries in the neurophysiological laboratory.

For example, one set of experimental findings that the author does not believe helps very much in explaining pattern recognition depends on the recording of single neuron action potentials. Although this is an immensely popular field these days, it seems to some students of the field that such research is aimed at the wrong level of analysis. Pattern recognition, like all other forms of mentation, must arise from the coordinated information processing of many, many neurons. Individual neurons may signal some relevant activity, but are unlikely to be the neural equivalents of pattern recognition or anything else mental.

IV. CONCLUSIONS

This brief article has surveyed some of the theories and empirical data in the field of pattern or object recognition. Obviously, this is a vital and exciting field of contemporary science that has much to gain from the interdisciplinary interaction of biological, psychological, and engineering sciences. Clearly, the generic problem of human pattern recognition (how do we classify an incoming visual image?) has not been solved to the satisfaction of anyone despite continuing progress in the development of engineered systems capable of ever closer approximations to human recognition. We do have a number of interesting theoretical approaches and an abundance of human psychophysical research. We realize what some of the fundamental issues are, but have not yet been able to answer the most fundamental of them.

The field of pattern recognition, as presented in this article, is not, by any means, complete. There are many other important research areas that could have been included if space permitted. We have made no mention of the one-dimensional patterns of speech and music that humans also recognize so effortlessly. Many of the general problems and issues that are identified for vision, however, generalize to the other senses. Even in the context of visual perception, this article could not be all-inclusive.

In conclusion, it does seem as if there is a contemporary contradiction between the majority of the theoretical models that emphasize features and a considerable portion of the findings from psychophysical experiments that attack the pattern recognition problem. Most parametric psychophysical studies of human pattern recognition, and even more strongly most first-order demonstrations, seem to suggest that the human observer recognizes patterns in a manner that is more sensitive to the global and holistic attributes of the incoming image than to the local features, components, or parts. The resolution of this contradiction is one of the major challenges facing this field in the future.

See Also the Following Articles

ARTIFICIAL INTELLIGENCE • INFORMATION PROCESSING • MENTAL WORKLOAD • MOTION PROCESSING • MULTISENSORY INTEGRATION • NERVE CELLS AND MEMORY • NEURAL NETWORKS • OBJECT PERCEPTION • SALIENCE • SPATIAL VISION

Suggested Reading

Chen, C. H., Pau, L. F., and Wang, P. S. P. (Eds.). (1993). *Handbook of Pattern Recognition and Computer Vision*. World Scientific, Singapore.

Grimson, W. E. L. (1990). *Object Recognition by Computer: The Role of Geometrical Constraints*. MIT Press, Cambridge, MA.

Lockhead, G. R., and Pomerantz, J. R. (Eds.). (1991). *The Perception of Structure*. American Psychological Association, Washington, DC.

Schalkhoff, R. (1992). *Pattern Recognition: Statistical, Structural, and Neural Approaches*. Wiley, New York.

Uttal, W. R. (1998). *Towards a New Behaviorism: The Case against Perceptual Reductionism*. Erlbaum, Hillsdale, NJ.

Uttal, W. R. (2002). *A Behaviorist Looks at Form Recognition*. Eplbaum, Mahwah, NJ.

Uttal, W. R., Kakarala, R., Dayanand, S., Shepherd Kalki, J., Lunskis, C. F., and Liu, N. (1999). *Computational Modeling of Vision: The Role of Combination*. Dekker, New York.

Peptides, Hormones, and the Brain and Spinal Cord

LEE J. MARTIN
Johns Hopkins University School of Medicine

GLOSSARY

astrocyte A class of nonneuronal glial cell in the central nervous system that has long, radial processes that ensheathe neurons and synaptic complexes, regulate the extracellular chemical and ionic environment, and secrete brain peptides, growth factors, cytokines, and chemokines.

central nervous system The brain and spinal cord.

chemokine A chemoattractant small peptide or protein cytokine.

cytokine Extracellular signaling peptide or protein that acts as a local mediator in cell–cell communication.

glial cell Nonneuronal cell in the nervous system (e.g., astrocyte, oligodendrocyte, microglial cell, and Schwann cell).

growth factor An extracellular peptide or protein that functions as a cell survival factor that can maintain cell survival or stimulate a cell to grow or proliferate.

hormone A chemical produced by a cell to regulate the functioning of another cell.

microglia the resident small phagocytic cells of the central nervous system that are related to the mononuclear phagocyte lineage and function as immune accessory cells that secrete cytokines and chemokines.

neuromodulator A chemical signal that modulates the response of a neuron to a neurotransmitter.

neuropeptide Peptide secreted by neurons or nonneuronal cells as either a synaptic or nonsynaptic cell–cell signaling molecule.

neurotransmitter Signaling molecule secreted by the presynaptic terminal of a neuron at chemical synapses to relay a signal to a postsynaptic neuron.

oligodendrocyte Glial cells that provide myelin sheaths for axons within the CNS and secrete peptide growth factors.

programmed cell death A form of cell death that is mediated by the activation of intrinsic mechanisms.

Schwann cell Glial cells that provide myelin sheaths for axons within the peripheral nervous system and that secrete peptide growth factors.

tyrosine kinase Enzyme that transfers the terminal phosphate of ATP to a specific tyrosine residue in a target protein.

Brain peptides are amino-acid-comprised molecules that are synthesized and released by neuronal and nonneuronal cells that function as intercellular signaling molecules, serving as neurotransmitters, neuromodulators, neurohormones, cytokines, chemokines, or growth factors.

I. INTRODUCTION

The field of brain peptide research has been revolutionary in neuroscience. Many fundamentally new concepts about brain organization and function, as well as abnormal function and disease, have been revealed by studying brain peptides. Through brain peptides, new interactions have been disclosed between the nervous system and endocrine system, the nervous system and gastrointestinal system, and the nervous system and immune system. New principles about cell–cell signaling and intracellular signal transduction pathways have come to light from studies of brain peptides and related molecules. These studies are particularly important because a variety of these naturally occurring chemicals have been implicated in modulating basic nervous system functions such as sensibility, emotions, and behavior, and they most likely participate in the pathobiology of neurologic and psychiatric diseases in humans. The accumulating information on brain peptides may have ramifications as far-reaching as the further understanding of the pathophysiology of Alzheimer's disease, amyotrophic lateral sclerosis, multiple sclerosis, stroke, acquired immune deficiency syndrome (AIDS), obesity, anorexia, sleeping disorders, and brain–spinal cord trauma.

At the same time, however, advances in the field of brain peptides tend to cloud the conventional classification of intercellular signaling molecules within the body. For example, the tachykinins are a family of peptides including substance P and neurokinins that are present widely throughout the body, including the brain and spinal cord. These peptides have several clearly identified functions outside the central nervous system (CNS) and within the CNS, participating in cardiovascular function and inflammation. In the spinal cord, tachykinins function in the synaptic transmission of nociceptive (pain) information and, thus, are important neurotransmitters. Yet tachykinins (as well as several other well-known brain peptides) also function in neuromodulation, participating in the regulation of neuronal excitability by modulating glutamatergic excitatory synaptic transmission via glutamate receptors, and they also function in neuronal survival akin to neurotrophic factors. Another example that emphasizes the breakdown of the walls of conventional classification of brain peptides is illustrated by the functions of vasoactive intestinal peptide. An early identified function of vasoactive intestinal peptide was the regulation of blood flow in the autonomic nervous system and CNS. Now, it has been shown that vasoactive intestinal peptide has diverse neurotransmitter, neuromodulator, cytokine, and neurotrophic factor actions that regulate neuronal survival and growth as well as glial activation and release of glial-cell-derived soluble factors. Work has revealed exciting new evidence that vasoactive intestinal peptide can prevent neuronal death induced by the AIDS virus envelope protein gp102, and another more recently identified vasoactive intestinal peptide family member (i.e., pituitary adenylate cyclase activating polypeptide) can block apoptosis of cultured granule neurons and necrosis of hippocampal pyramidal neurons after cerebral ischemia. These few examples highlight the fact that the highly diverse actions of brain peptides can make it difficult to discretely classify these naturally occurring brain chemicals. Furthermore, these examples emphasize their pathophysiological and potential therapeutic importance in disorders of the human CNS.

The discovery of each brain peptide involves its isolation, chemical characterization, anatomical localization, gene cloning, receptor identification and cloning, and functional assessment at the micro (electrophysiological and intracellular signaling) and macro-levels (behavior). Substance P was the first neuroactive peptide found in nervous tissue. Originally, there was considerable debate about the suitability of referring to brain peptides as neurotransmitters. Many of these peptides had been identified previously as peripheral tissue hormones with physiological actions at target organs outside the brain. Cholecystokinin, somatostatin, and insulin-like growth factors are only a few examples in this regard (Table I).

The possibility that small peptides acted as neurotransmitters was a new concept in the early emerging field of neuroscience (Fig. 1). Generally, four criteria should be fulfilled for a chemical to be considered a neurotransmitter: (1) a neurotransmitter is synthesized in neurons (Fig. 2); (2) a neurotransmitter is present in the presynaptic terminal and is released after depolarization in quantities sufficient to effect a response postsynaptically (Fig. 1); (3) exogenous application of a candidate neurotransmitter at physiological concentrations mimics the actions of endogenously released chemical; and (4) the action of the candidate neurotransmiter at the synapse is terminated by a specific mechanism. Many different types of brain peptides (old and recently identified) have now been shown to be neurotransmitters (Table I).

More than one neurotransmitter may coexist in neurons. Neuroactive peptides and small-molecule

Table I
Classical Neuroactive Peptides[a]

Peptide	Size[b]	Family	Major functions
Vasopressin	9	Neurohypophyseal hormone	Renal H_2O resorption, vasoconstriction
Oxytocin	9	Neurohypophyseal hormone	Uterine contraction, milk ejection, natriuretic effects
Gonadotropin-releasing hormone	10	Hypothalamic peptide	Stimulates secretion of FSH and LH
Thyrotrophin-releasing hormone	3	Hypothalamic peptide	Stimulates TSH secretion
Corticotropin-releasing hormone	41	Hypothalamic peptide	Stimulates ACTH secretion
β-Endorphin	30	Endogenous opioid	Analgesia, stress response, inhibition of dopamine release
Dynorphin	17	Endogenous opioid	Analgesia, inhibition of dopamine release
Enkephalin	5	Endogenous opioid	Analgesia, primary afferent modulation
Cholecystokinin	8	Gastrointestinal peptide gastrin	Inhibition of feeding
Neuropeptide Y	36	Gastrointestinal peptide	Feeding behavior, modulation of neuronal excitability
Neurotensin	13	Gastrointestinal peptide	Modulation of dopaminergic neurons
Somatostatin	14	Gastrointestinal–hypothalamic peptide	Inhibits growth hormone release, inhibits synaptic transmission, neurotrophin
Substance P	11	Gastrointestinal–tachykinin peptide	Pain transmission, neurotrophin
Vasoactive intestinal peptide	28	Gastrointestinal–secretin peptide	Cerebral blood flow regulation
Insulin-like growth factor	67–70	Insulin	Neuronal survival

[a]Represents only a selection of the more than 50 classical neuroactive peptides that have been identified.
[b]Number of amino acid residues.

neurotransmitters are often found within the same neuron (Fig. 1). This observation was iconoclastic because, until this discovery, it was believed that each neuron used only one neurotransmitter. Now it is known that, at synapses, peptides are stored in dense-core vesicles that contain ATP. These dense-core vesicles (called dense-core because they appear to have a dark core when viewed in the electron microscope) are usually much less numerous than small, clear synaptic vesicles that contain neurotransmitters such as acetylcholine, GABA, and glutamate (Fig. 1). Peptide-containing dense-core vesicles are large (100–150 nm) compared to small, clear synaptic vesicles (40–60 nm). Dense-core vesicles are released at higher frequencies of stimulation of the axon and are not released solely at the presynaptic active zone like small-molecule neurotransmitters, because they do not require the presynaptic membrane specialization for their exocytotic release. The inactivation of the neuroactive peptide at the synapse occurs extracellularly and is slower than the inactivation of the small-molecular-weight neurotransmitter; thus, peptides act over a longer period of time than classical neurotransmitters.

II. BRAIN PEPTIDES OCCUR AS MANY DIFFERENT KINDS WITH MANY DIFFERENT FUNCTIONS

The diversity of brain peptides is extensive (Tables I–IV). These peptides can be grouped into families. One way to classify neuroactive peptides is based on structural similarities, notably, commonalities in amino acid sequences (i.e., primary structure). Similarities in the physiological responses (i.e., biological activities) that are evoked by different brain peptides are another way to group these chemical messengers, although each brain peptide can have multiple physiological actions. More recently, similarities in the nucleotide base sequences in the genes that encode precursors of neuroactive peptides are used identify membership in peptide families.

A. Classical and More Recently Discovered Neuropeptides

The classical brain peptides can be grouped into broad families such as the hypothalamic-releasing hormones,

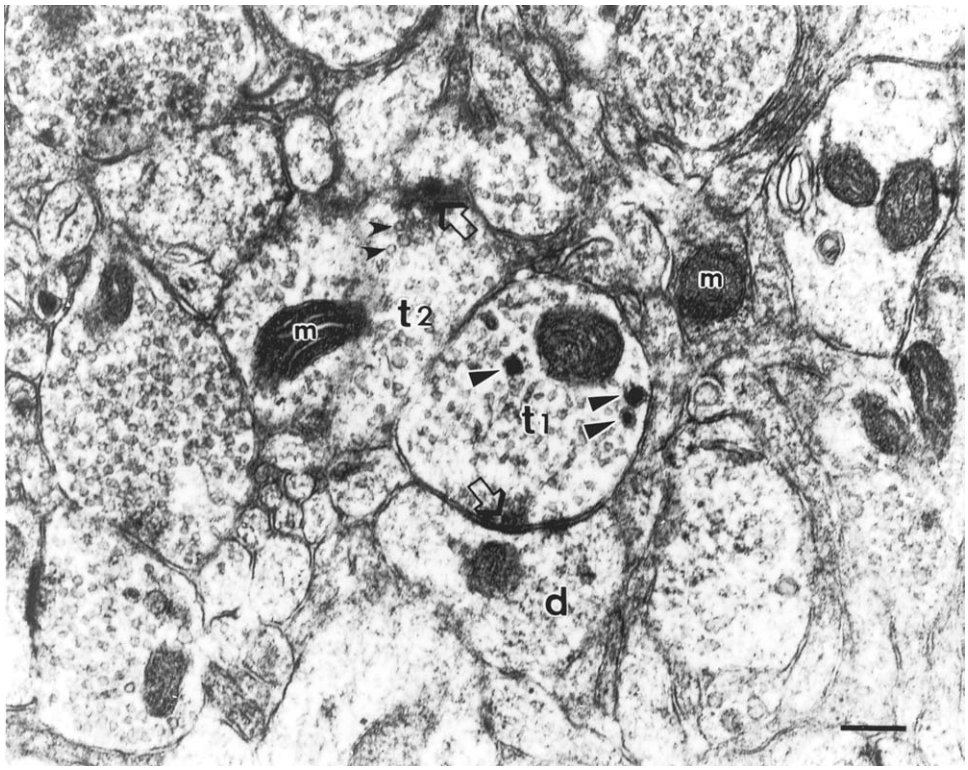


Figure 1 Synaptic structure as revealed by electron microscopic examination of rhesus monkey cerebral cortex. Nerve terminals (t1 and t2) contain numerous small, clear, round synaptic vesicles (t2, small, black arrowheads) but only a few dense-core synaptic vesicles that can contain neuropeptides (t1, black-on-white arrowheads). Mitochondria (m) are present in nerve terminals for the energy-dependent release of synaptic vesicles. Nerve terminals form synaptic contacts with dendrites (d). The active zone of the synapse (t1 and t2, open arrow) is the site of synaptic vesicle exocytosis. Scale bar = 0.4 μm.

the neurohypophyseal hormones, the pituitary peptides, the gastrointestinal peptides, the opioids, and the tachykinins (Table I). The classical brain peptides are located differentially throughout the CNS (Fig. 3). For example, forebrain structures (cerebral cortex, striatum, and amygdala) tend to have greater diversity in neuropeptides than hindbrain structures (cerebellum and spinal cord). Remarkably, regions that have the same embryonic derivation within the diencephalon (thalamus versus hypothalamus) can have vastly different neuropeptide profiles (Fig. 3).

Major advances have been made in understanding brain peptide regulation of feeding and body weight. In the 1950s a recessive obesity mutation was identified that results in profound obesity and adult-onset (type II) diabetes. It was thus postulated that an obese gene product may be a component in a signaling pathway regulating body fat deposition. Mice with mutations in the obese gene are obese and diabetic and are found to have reduced activity, metabolism, and body temperature. It was discovered that leptin, a 16-kDa secreted protein hormone, is the product of the obese gene and is believed to be synthesized only in adipose tissue. Leptin is an integral component in a homeostatic loop that regulates body weight. Leptin acts to control food intake and energy expenditure by both classical and newly discovered neuropeptides in the hypothalamus. One classical neuropeptide that functions in weight control is neuropeptide Y (Table I).

The brain neuropeptides that function in this loop are the orexins (or hypocretins because they are hypothalamic neuropeptides similar to the gut hormone secretin). Orexin-A and orexin-B are peptides of 33 and 28 amino acid residues, respectively, that are derived from the proteolytic processing of a single prepro-orexin precursor protein. These peptides are produced exclusively by a specific group of neurons in the lateral hypothalamus (Fig. 3) called the perifornical nucleus. These orexin-utilizing neurons have widespread projections to the olfactory bulb, cerebral cortex, thalamus, hypothalamus, and brain stem. These brain peptides are endogenous ligands for two

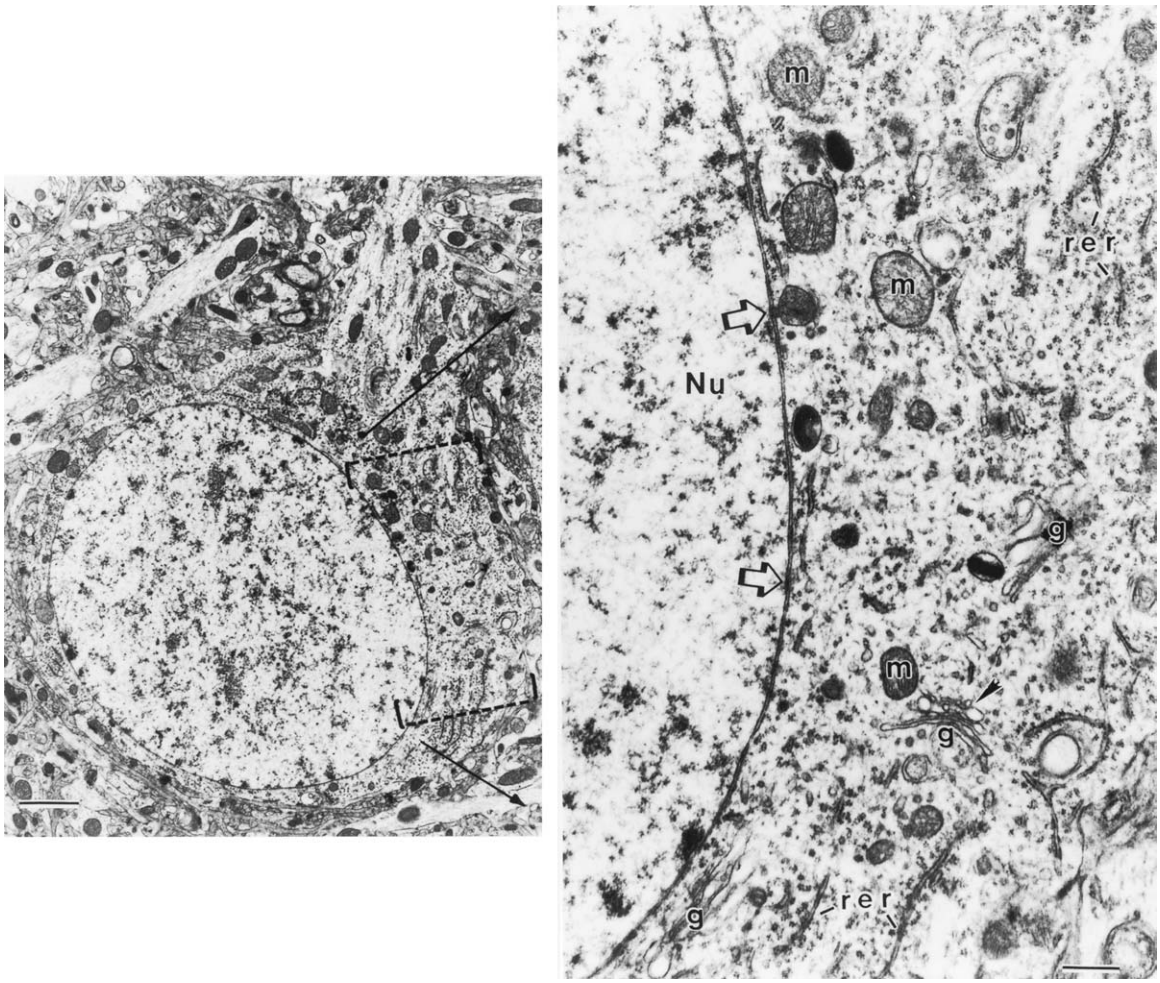


Figure 2 Electron microscopic identification of subcellular organelles in neurons of piglet striatum. Low-magnification view (left panel) of a primary neuron found in the striatum. At higher magnification (right panel), these neurons have abundant polyribosomes distributed throughout the cytoplasmic matrix (small granules), arrays of rough endoplasmic reticulum (rer), stacks of Golgi complex (g) with budding secretory vesicles (black-on-white arrowhead), and mitochondria (m). The nucleus (Nu) has a predominantly pale matrix and is surrounded by a continuous, bilaminar nuclear membrane that contains nuclear pores (arrows). Scale bars = 0.85 μm (left panel) and 0.25 μm (right panel).

G-protein-coupled receptors found in the brain that function at synapses to increase the presynaptic release of GABA and glutamate.

In addition to these peptides functioning in energy homeostasis, it has been discovered that the orexin neuropeptide–receptor system functions in sleep regulation. Two animal models dramatically highlight this point. Mice deficient in the orexin gene display abnormal sleep–wakeful states, and autosomal recessive mutations of the hypocretin (orexin) receptor-2 gene cause a sleep disorder in Doberman Pinschers. Both of these abnormalities are believed to cause sleep disorders similar to narcolepsy in humans. Support for this conclusion is derived from studies showing that orexin is undetectable in many people with narcolepsy.

B. Growth Factors

A growth factor is a secreted peptide or protein that functions as a cell survival factor and can stimulate cell growth, differentiation, or proliferation (Table II). In the nervous system, growth factors are called neurotrophins. They can act locally as autocrine or paracrine regulators of cell function. The first neurotrophin to be isolated, chemically characterized, and physiologically

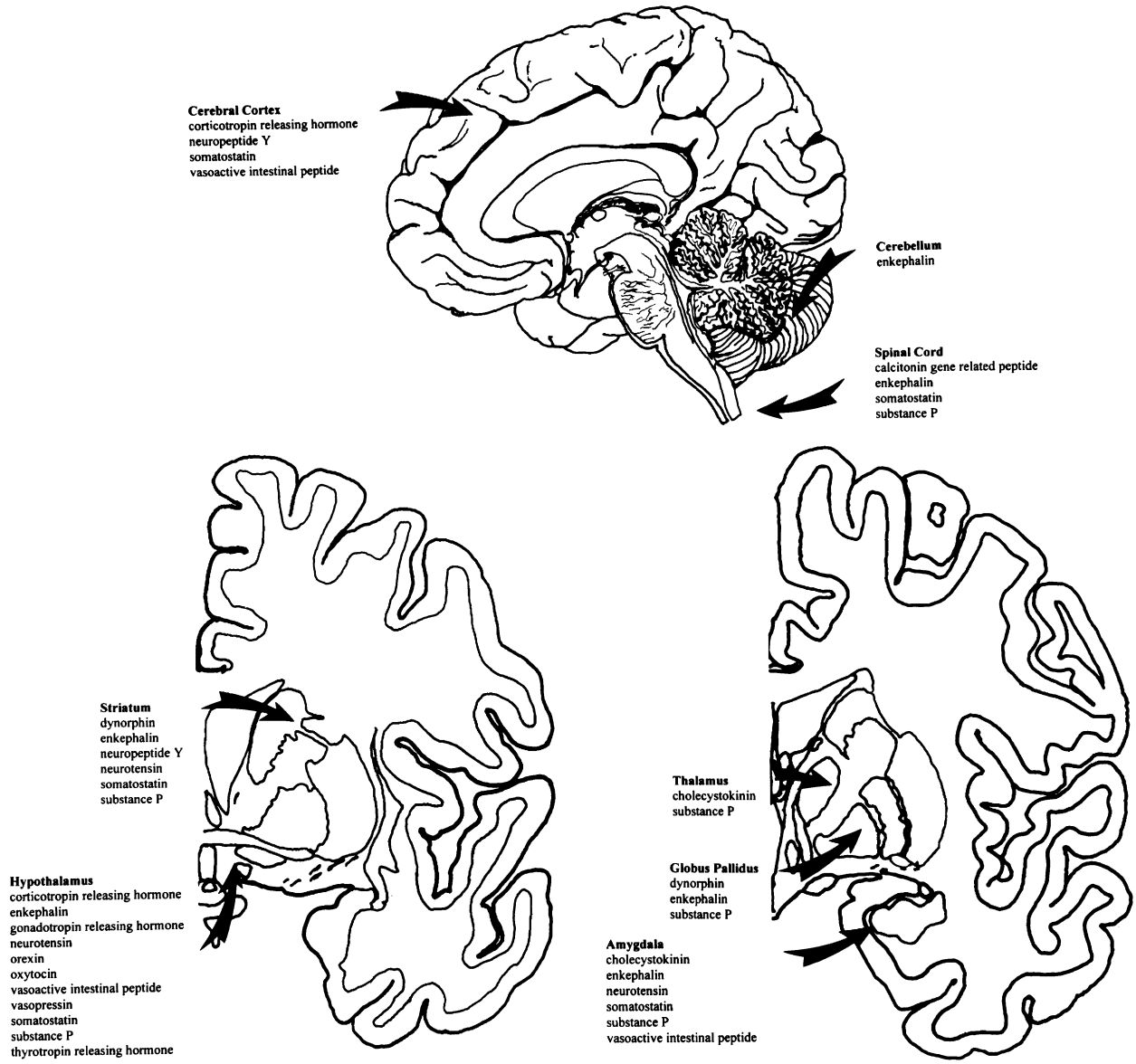


Figure 3 General locations of some classical neuropeptides in the human CNS. The top image is a view of the brain and upper spinal cord from a midsagittal perspective, and the two lower panels are transverse views through the forebrain and anterior or mid-diencephalon. Some of the major neuropeptides found in these regions are indicated.

evaluated was nerve growth factor. Neurotrophins act as target- or afferent-derived growth factors for specific populations of neurons and can have survival-promoting effects on immature and adult neurons. In general, the effects of neurotrophins are mediated by binding to a specific high-affinity tyrosine kinase receptor on the presynaptic terminal (Fig. 1), followed by internalization and then retrograde transport of the neurotrophin–receptor complex to the neuronal cell body (Fig. 2), where it can regulate neuronal differentiation and survival according to the traditional neurotrophin concept. Some neurotrophins (e.g., brain-derived neurotrophic factor) regulate the expression of neuropeptides (e.g., somatostatin and neuropeptide Y) in neurons. More recent discoveries, however, have now altered this concept. Neurotrophins can be transported away from the neuronal cell body down the axon, stored in nerve terminals, and then released

Table II
Survival Peptides and Growth Factors for Neurons

Growth factor	Cellular sources	Responsive cell populations
Nerve growth factor (NGF)	Schwann cells, neurons in hippocampus, neocortex, olfactory bulb	Neurons in peripheral ganglia, basal forebrain magnocellular complex (BFMC), striatum
Ciliary neurotrophic factor (CNTF)	Schwann cells, astroglia	Peripheral ganglion neurons, motor neurons, hippocampal neurons, astroglia, oligodendroglia
Brain-derived neurotrophic factor (BDNF)	Neurons in neocortex, hippocampus, BFMC, striatum, hypothalamus, cerebellum	Neurons in peripheral ganglia, retina, BFMC, substantia nigra, cerebellum
Neurotrophin-3 (NT-3)	Neurons in hippocampus, cerebellum	Neurons in peripheral and enteric ganglia, hippocampus, cerebellum, motor neurons
Neurotrophin-4/5 (NT-4/5)	Neurons in neocortex, hippocampus, BFMC	Peripheral ganglion neurons, BFMC, motor neurons
Glial-cell-derived neurotrophic factor (GDNF)	Neurons in neocortex, hippocampus, striatum, peripheral ganglia, motor neurons, chromaffin cells, muscle	Peripheral ganglion neurons, motor neurons, substantia nigra neurons, cerebellar neurons
Insulin-like growth factor (IGF)	Schwann cells, muscle	Motor neurons, astroglia, oligodendroglia
Leukemia inhibitory factor (LIF)	Astroglia, microglia, Schwann cells	Peripheral ganglion neurons, spinal motor neurons, astroglia, oligodendroglia, microglia
Fibroblast growth factor (FGF)	Peripheral ganglion neurons, motor neurons, muscle	Peripheral ganglion neurons, motor neurons

presynaptically, thus functioning as afferent-derived (instead of target-derived) growth factors or neurotransmitters.

C. Cytokines

Cytokines are a highly diverse group of extracellular signaling molecules that act as local mediators in cell–cell communication (Table III). These molecules are peptide hormones that regulate a wide variety of inflammatory and immune processes. Furthermore, they can function by modulating cellular responses to growth factors by either potentiating or antagonizing signals. Their activity is mediated by binding to specific, high-affinity cell surface receptors on target cells. The production of cytokines is triggered by an activation event, broadly including brain injury or infection. In the nervous system, cytokines are produced by glial cells (astrocytes, microglia, oligodendrocytes, and Schwann cells), and their cellular targets can be other glial cells and neurons. They can initiate, propagate, and suppress inflammatory and immune responses. Some cytokines that are synthesized by astrocytes (e.g., interleukin-1, interleukin-6, and granulocyte macrophage colony stimulating factor) participate in the recruitment of microglia into damaged areas, whereas other cytokines (e.g., interleukin-1α, interleukin-1β, and tumor necrosis factor-α) are produced predominantly by activated microglia.

D. Chemokines

Chemokines are small peptides or proteins that are members of a superfamily of inducible, secreted, proinflammatory cytokines that function as cellular chemoattractant signals (Table IV). These highly diverse inflammatory molecules are classified, according to the topology of cysteine residues, into four groups: C, C-C, C-X-C, and C-X_3-C. The C subfamily lacks the first and third cysteine residues of the conserved motif, whereas the C-C subfamily members have the first two cysteines adjacent. In the C-X-C group, the first two of four cysteine residues are separated by a single amino acid, and in the C-X_3-C group the first two cysteines are separated by three amino acid residues. By activating G-protein-coupled membrane receptors on their target cells, chemokines can signal cells to migrate to or remain at the site of chemokine production. An important example for chemokine activity is the governing of inflammatory cell accumulation during immune-mediated demyelination of axons.

Table III
Cytokines[a]

Cytokine	Cellular sources	Functions
Tumor necrosis factor α (TNF-α)	Astroglia, microglia	Inflammation, cell death through necrosis or apoptosis
Fas ligand (FasL)	T-cells	Promote apoptosis
Interleukins	Astroglia, microglia, endothelial cells	Promote or suppress inflammation, modulate chemokine production, molecule specific (at least 15 interleukins have been identified)
Interferon-γ	T-cells	Antiviral activity, antiproliferation, immunomodulation, microglial activation
Transforming growth factor β (TGF-β)	Neurons, astroglia, microglia, Schwann cells	Regulate extracellular matrix and proliferation of astroglia and Schwann cells, suppress inflammation, modulate chemokine production
Leukemia inhibitory factor (LIF)	Astroglia, microglia, Schwann cells	Promote neuronal repair

[a]Represents a tabulation of only selected cytokines found in brain.

III. NEUROACTIVE BRAIN PEPTIDES ARE MADE FROM PROTEIN PRECURSORS THAT UNDERGO PROTEOLYTIC PROCESSING

Neuroactive peptide hormones are generated from precursor proteins. Several different neuroactive peptides are usually encoded by a single continuous mRNA that is translated into a large, inactive protein precursor. Prior to translation, these mRNA transcripts can undergo alternative RNA splicing within the nucleus to generate different mature mRNAs that encode protein precursors with different amino acid sequences. These secretory proteins are formed in the cell bodies of neurons and nonneuronal cells on polyribosomes attached to the cytosolic surface of the endoplasmic reticulum (Fig. 2). Neuroactive peptides or their precursors, like other secretory proteins, are processed in the endoplasmic reticulum and then shuttled to the Golgi apparatus for further processing (Fig. 2). The processing of larger precursor proteins occurs through specific, regulated proteolytic cleavage by serine proteases, thiol endopeptidases, amino peptidases, and carboxypeptidases. These cleavages are thought to begin in the *trans*-Golgi network, and they continue in the secretory vesicles. Proteins destined for secretory dense-core vesicles (Fig. 1) are packaged into appropriate vesicles in the Golgi apparatus by a sorting signal mechanism involving the selective aggregation of secretory proteins. The peptide-containing secretory vesicles leave the

Table IV
Chemokines in the Brain[a]

Chemokine	Cellular expression	Functions
Macrophage inflammatory protein-1 (MIP1)	T-cells, monocytes, macrophages, platelets, astroglia, microglia	Attraction of T-cells, monocytes, basophils, eosinophils, natural killer cells
Monocyte chemoattractant protein-1 (MCP1)	Monocytes, macrophages, endothelial cells, astroglia	Attraction and activation of monocytes, lymphocytes, basophils
Interferon inducible protein-10 (IP10)	Astroglia	Attraction of T-cells
RANTES[b]	Astroglia, microglia, Schwann cells, endothelial cells	Monocyte chemoattraction, neuronal migration
Fractalkine	T-cells, endothelial cells, neurons, microglia	Suppression of inflammation and microglial activation

[a]Only selected cytokines are shown.
[b]RANTES: regulated upon activation, normal T-cell expressed and secreted.

trans-Golgi apparatus by a process involving clathrin-coated budding (Fig. 2). After the immature secretory vesicles bud from the Golgi apparatus, their contents undergo rapid and extreme condensation resulting from acidification of the vesicle lumen by the activity of a vesicle membrane ATP-driven H^+ pump. In neurons, the mature vesicles are then moved to the presynaptic terminal (Fig. 1) by fast axonal transport. This traveling along the axon occurs by kinesin motor proteins attached to the surface of the secretory vesicle that propel them along microtubules. The release of the peptide-containing vesicle at the axon terminal occurs by exocytosis in response to a depolarization-induced local increase in intracellular Ca^{2+}.

IV. BRAIN PEPTIDES AND HORMONES ACT AT CELL MEMBRANE RECEPTORS

The signal transduction mechanisms through which most neuroactive brain peptides operate involve G-protein-linked receptors or enzyme-linked receptors. Most classical neuroactive peptides (Table I, e.g., substance P–neurokinins, somatostatin, neurotensin, opioid pentapeptides, vasopressin, and angiotensin) and some cytokine peptides (interleukin-8) are ligands for G-protein-linked receptors. These receptors activate a chain of events that alters the concentration of one or more small, intracellular, second messenger signaling molecules that, in turn, amplify the signal and pass it on by altering the functioning of specific proteins. Two of the most widely used second messengers for classical neuroactive peptides are Ca^{2+} and cyclic adenosine monophosphate (AMP). The effects of somatostatin, vasopressin, and adrenocorticotropic hormone are mediated by cyclic AMP.

In contrast, enzyme-linked receptors function directly as enzymes or as receptor-associated enzymes when activated. Neuropeptides, growth factors, cytokines, and chemokines can operate through one of five known classes of enzyme-linked receptors, including the following: (1) receptor guanylyl cyclases that catalyze the production of cyclic guanosine monophosphate (GMP) in the cytosol; (2) receptor tyrosine kinases that phosphorylate specific tyrosine residues on a small set of signaling proteins; (3) tyrosine-kinase-associated receptors that interact with proteins that have tyrosine kinase activity; (4) receptor tyrosine phosphatases that remove phosphate groups (dephosphorylate) from tyrosine residues of signaling proteins; and (5) receptor serine–threonine kinases that phosphorylate serine or threonine residues in specific proteins.

V. BRAIN PEPTIDES PROVIDE A BIOCHEMICAL FINGERPRINT FOR MANY BRAIN REGIONS AND REVEAL NEW FEATURES ABOUT THE PARCELING OF THE BRAIN

Very important and novel information about the nervous system has been obtained by charting the location of brain peptides (Figs. 4–7). Generally, these data are gleaned immunocytochemically by using highly specific antibodies that recognize unique amino acid sequences in brain peptides and proteins. Such experiments can provide critical information on the brain regions that contain peptides–proteins and their detailed cellular and subcellular localizations in experimental animal and post mortem human brain and spinal cord tissues. Over the past two decades, exciting results on brain peptides have been obtained that feature how the brains of humans and animals are organized, how they function, and how they are changed in abnormal conditions. Many prominent examples can be provided about how the charting of brain peptides reveals novel information on brain organization and the abnormalities in brain peptides that occur in experimental animals with profound neurobehavioral disorders. Two forebrain regions, the amygdala–bed nucleus complex and the striatum, provide informative examples.

A. The Amygdala–Bed Nucleus of the Stria Terminalis (BST) Complex

By studying the chemical neuroanatomy of the basal forebrain of humans and monkeys, it was discovered that the bed nucleus of the stria terminalis (BST) and the amygdala form a large, multidivisional complex (Figs. 4 and 5). This complex has been also called the extended amygdala by other investigators. The BST is continuous with the central and medial divisions of the amygdala through the ventral forebrain substantia innominata, and it intermingles with the nucleus basalis of Meynert of the basal forebrain magnocellular complex. The amygdala–BST continuum is a hotbed of neuropeptides, including somatostatin, enkephalins, substance P, neurotensin, cholecystokinin, vasoactive intestinal peptide, and galanin (Figs. 4 and 5). Specific chemically delineated zones and cellular compartments contain these brain peptides. The locations of these brain peptides reveal that the amygdala–BST complex is a very prominent and discretely compartmental structure in the basal

forebrain of humans and monkeys (Figs. 4 and 5). On the basis of neuropeptide distribution and cytology, this complex can be divided into at least 10 subdivisions. Furthermore, both subtle and dramatic differences in brain peptides are found in the amygdala–BST complex in different species, with humans having unique patterns that differ from those of animals.

Prominent gender differences (sexual dimorphisms) occur in the BST of humans and experimental animals. In humans, some divisions of BST are ~ 2.5 larger in the male brain than in the female brain. Major differences in the neuropeptide composition of the BST occur as well. These differences are thought to be caused by gonadal steroid hormones and also by glial-cell-derived peptide growth factors and cytokines such as transforming growth factor-β. Neurons of the BST and centromedial amygdala are among the chief testosterone- and estradiol-concentrating cells in animal brains. The sexual differentiation of the human brain takes place much later than originally believed, suggesting that, in addition to genetic factors, environmental and psychosocial factors may have ample opportunity to exert a profound influence on the sexual differentiation of the brain. This information is important because it could have implications for the determination of sexual orientation, and the BST is known to participate in sexually dimorphic functions, including aggressive behavior, sexual behavior, and feeding behavior. Furthermore, a relatively well-identified role of the amygdala–BST complex is the mediation of stress, fear, and anxiety responses, which

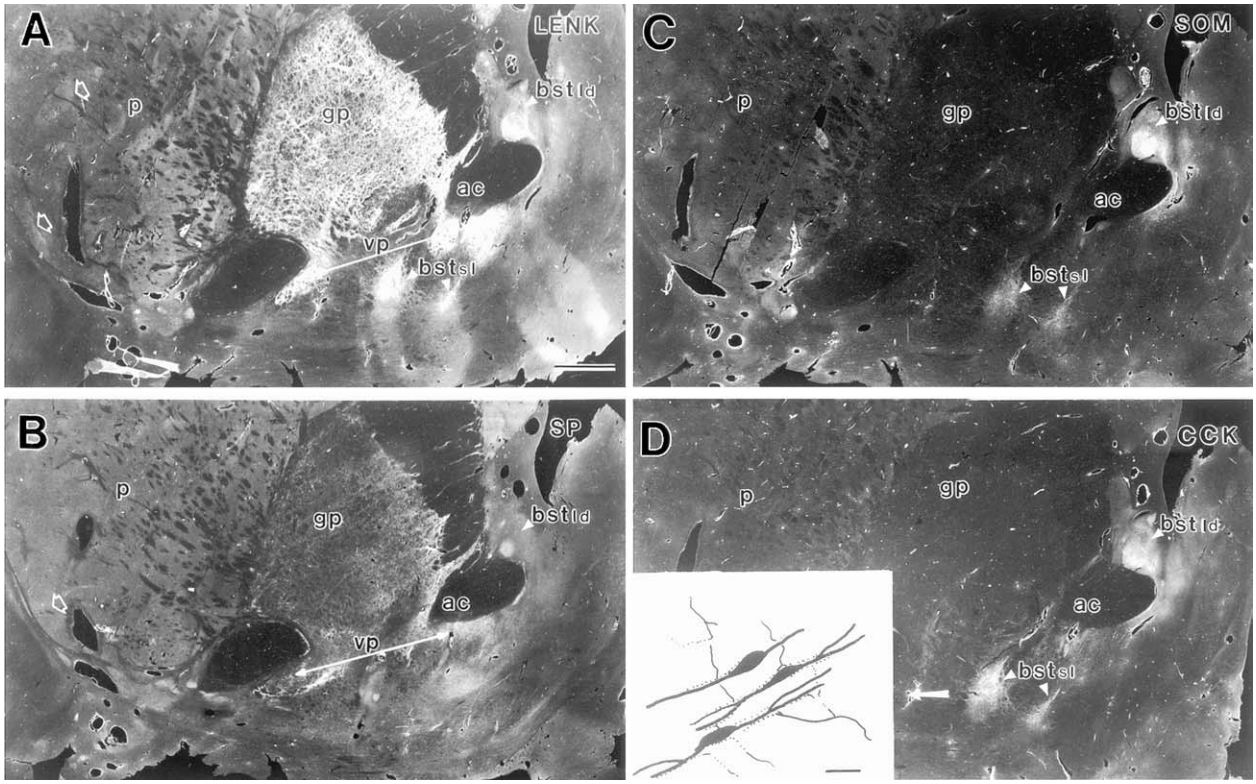


Figure 4 Neuropeptides delineate the compartments within in the basal forebrain and basal ganglia of humans. Antibodies were used to identify leucine-enkephalin (LENK, A), substance P (SP, B), somatostatin (SOM, C), and cholecystokinin (CCK, D) in postmortem human brain sections. Abbreviations: ac, anterior commissure; bstld, bed nucleus of the stria terminalis, lateral dorsal division; bstsl, bed nucleus of the stria terminalis, sublenticular division; gp, globus pallidus; p, putamen; vp, ventral pallidum. White areas represent neuropeptide immunoreactivity. The globus pallidus, its ventral extension (the ventral pallidum), and the bed nucleus are enriched in enkephalin. The striatal mosaic in the putamen (A, white, open, arrowheads) can be identified by enkephalin. The bed nucleus and its sublenticular extension are clearly divisible from the globus pallidus and the ventral pallidum by the presents of somatostatin and cholecystokinin. Somatostatin and cholecystokinin specifically mark the sublenticular division of the BST, so that it can be distinguished from the ventral pallidum. The inset (D) shows the typical distributions of LENK, SOM, and CCK in the BST sublenticular division. These neuropeptides are highly concentrated in many nerve terminals (small dots) that innervate neuronal cell bodies and dendrites that form the cellular columns that bridge the BST and amygdala in the human basal forebrain. Scale bars = 2.5 mm (A, same for B–D) and 40 μm (inset).

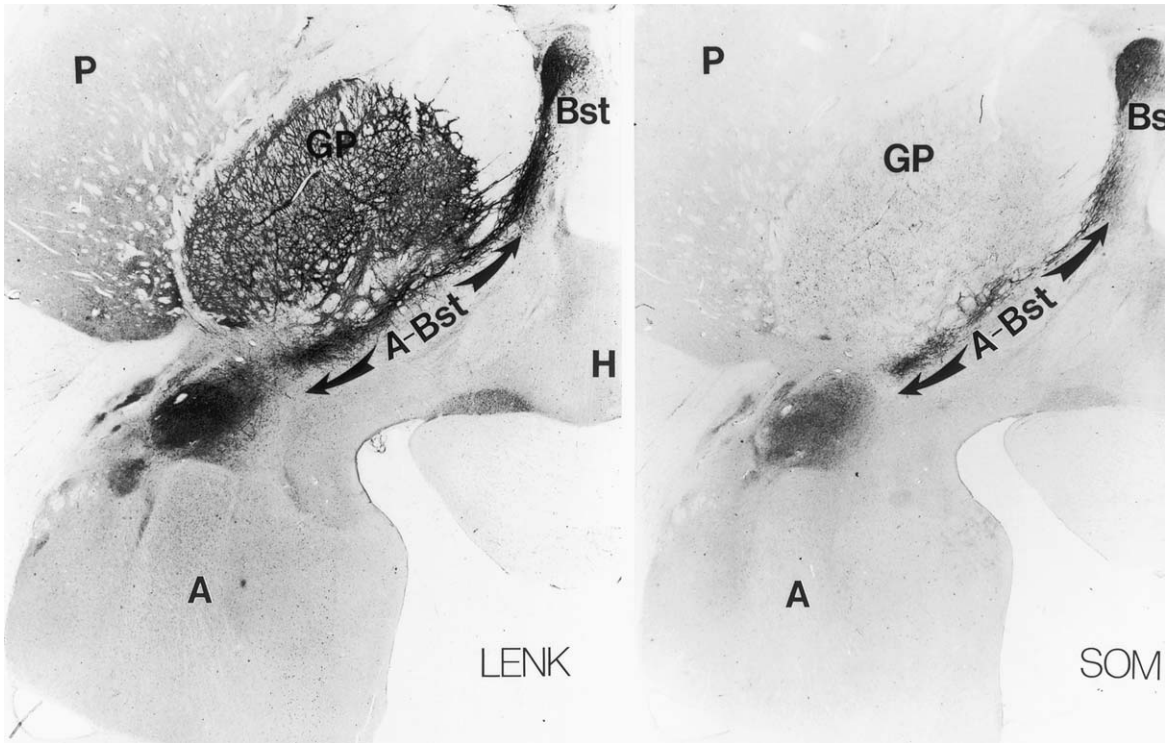


Figure 5 Neuropeptides delineate the amygdala–bed nucleus complex within in the basal forebrain of nonhuman primates. Antibodies were used to identify leucine-enkephalin (LENK, left panel) and somatostatin (SOM, right panel) in brain sections of rhesus monkeys. Abbreviations: A, amygdala; A–Bst, amygdala–bed nucleus continuum; Bst, bed nucleus of the stria terminalis; GP, globus pallidus; H, hypothalamus; P, putamen. Black areas represent neuropeptide immunoreactivity. The amygdala–bed nucleus continuum and the globus pallidus are both enriched in enkephalin. The amygdala–bed nucleus complex is specifically identified by the high concentration of somatostatin. Scale bar = 1.3 mm.

are likely to be regulated by brain peptides such as corticotrophin-releasing hormone.

B. The Striatal Mosaic

The basal ganglia are subcortical brain regions that function in sensory–motor integration, planning and initiation of somatic movements, cognition, and some types of social–emotional and species-typical behaviors. The striatum (caudate nucleus, putamen, and nucleus accumbens), globus pallidus (external and internal divisions and ventral pallidum), substantia nigra (pars compacta and reticulata), and subthalamic nucleus are the major components of the basal ganglia (Figs. 6 and 7). The striatum is divisible into at least two primary compartments designated as striosomes (patches) and matrix (Figs. 6 and 7). This forebrain region, thus, is like a jigsaw puzzle because it has different pieces that fit together, comprising a structure called the striatal mosaic. Neurons within striosomes and matrix differ with respect to their time of embryogenesis, connections with other brain regions, and expression of brain peptides. Striosomes are much more enriched in substance P and enkephalins than the striatal matrix, whereas the matrix is more enriched in somatostatin and neuropeptide Y than the striosomes. These general neuropeptide patterns change when comparing the dorsal striatum (caudate and putamen) with the ventral striatum (nucleus accumbens) (Figs. 6 and 7).

An important relationship exists between the regulation of gene expression for brain peptides within neurons of the striatal mosaic and the expression of excitatory glutamate receptors. Inputs into the striatum participate in the steady-state regulation of peptide expression in medium-spiny, striatal neurons, which are the principal neurons of the striatum. For example, ablation of corticostriatal projections reduces preprotachykinin and preproenkephalin mRNA in neurons of rodent striatum. Synaptic activity and membrane depolarization regulate neural gene

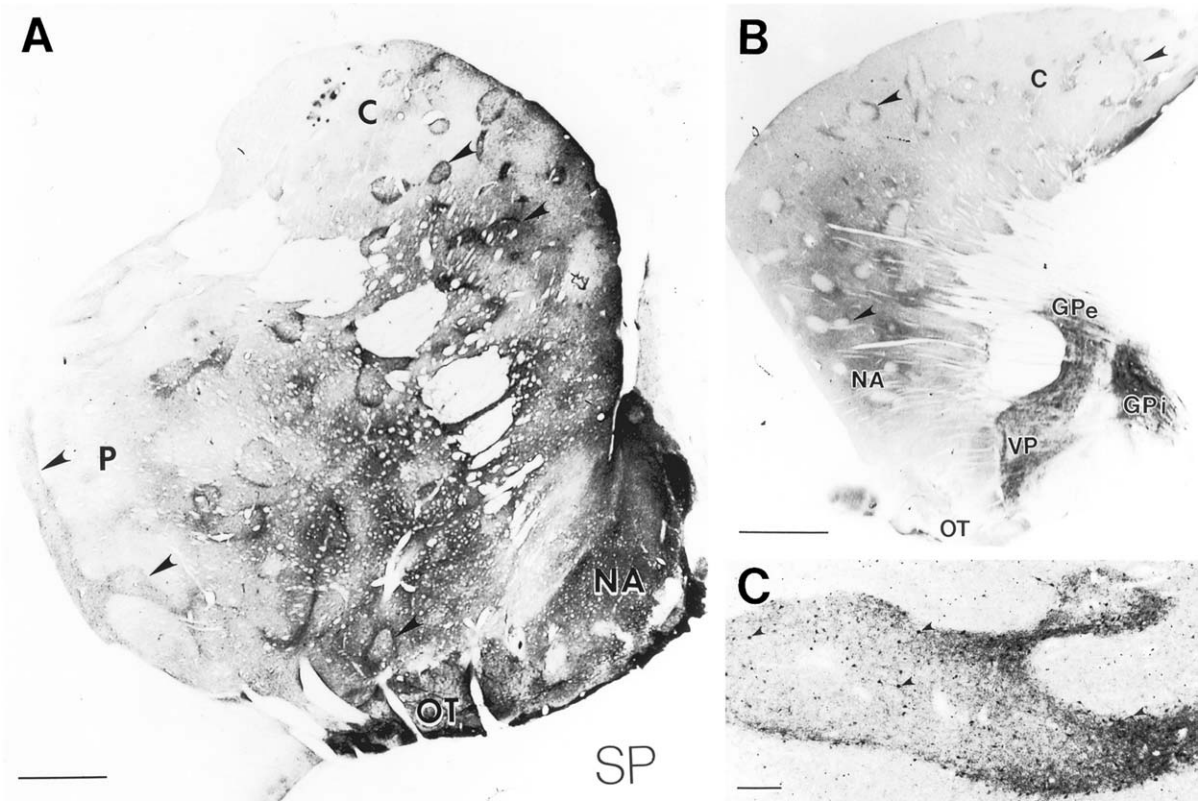


Figure 6 The localization of substance P reveals that the striatum is a mosaic and extends to the ventral surface of the brain. Antibodies were used to identify substance P in brain sections of rhesus monkeys. Shown are photographs of striatum in the transverse (A) and parasagittal (B) planes and a higher magnification view of a striosome (C). Abbreviations: C, caudate nucleus; GPe, globus pallidus external part; GPi, globus pallidus internal part; NA, nucleus accumbens; OT, olfactory tubercle; VP, ventral pallidum. Black–gray areas represent neuropeptide immunoreactivity. These photographs illustrate the discrete topography and heterogeneity of the striosome (A and B, black arrowheads) and matrix compartments throughout the dorsal striatum (caudate nucleus and putamen) and the ventral striatum (nucleus accumbens and olfactory tubercle). The chemical organization of striosomes is different in the dorsal striatum from the ventral striatum. In the caudate and putamen, striosomes are enriched in substance P or they have a ring (dark) and hollow (pale) pattern compared to the surrounding matrix. The more uniform striosomes contain numerous neurons with substance P (C, small black arrowheads) that are embedded in a dense plexus of nerve terminals containing substance P. In the nucleus accumbens, striosomes are poor in substance P compared to the surrounding matrix (B). Scale bars = 1.7 mm (A), 2.2 mm (B) and 130 μm (C).

expression, possibly through a Ca^{2+}-dependent mechanism that may involve calmodulin. Striatal neurons are enriched in calmodulin. Thus, excitatory glutamatergic neurotransmission may directly regulate, in an activity-dependent manner, neuropeptide expression within subsets of striatal neurons. Support for this idea is derived from our studies showing that glutamate receptors are expressed by substance P- and enkephalin-containing neurons within striosomes. This enrichment of glutamate receptors at postsynaptic sites within striosomes of the dorsal striatum may reflect a more dominant glutamatergic regulatory control of neuropeptide expression within striosomal neurons than within matrix neurons. An informative example in this regard is provided by striatal dopamine receptors. Dopamine receptors regulate the relative levels of expression of opiate and tachykinin peptides in subsets of striatal neurons through cyclic AMP, protein kinase A, and cyclic AMP response-element-binding protein. Several genes for striatal neuropeptides contain cyclic AMP response elements, including dynorphin, enkephalin, and somatostatin. Although the activation of different molecular subtypes of glutamate receptors may engage signal transduction pathways for neuropeptide gene expression in striatal neurons, the intracellular signaling pathways have not been identified completely.

For many years, the neuropeptide organization of the forebrain in normal monkeys and humans has been studied, thereby laying a foundation for concepts

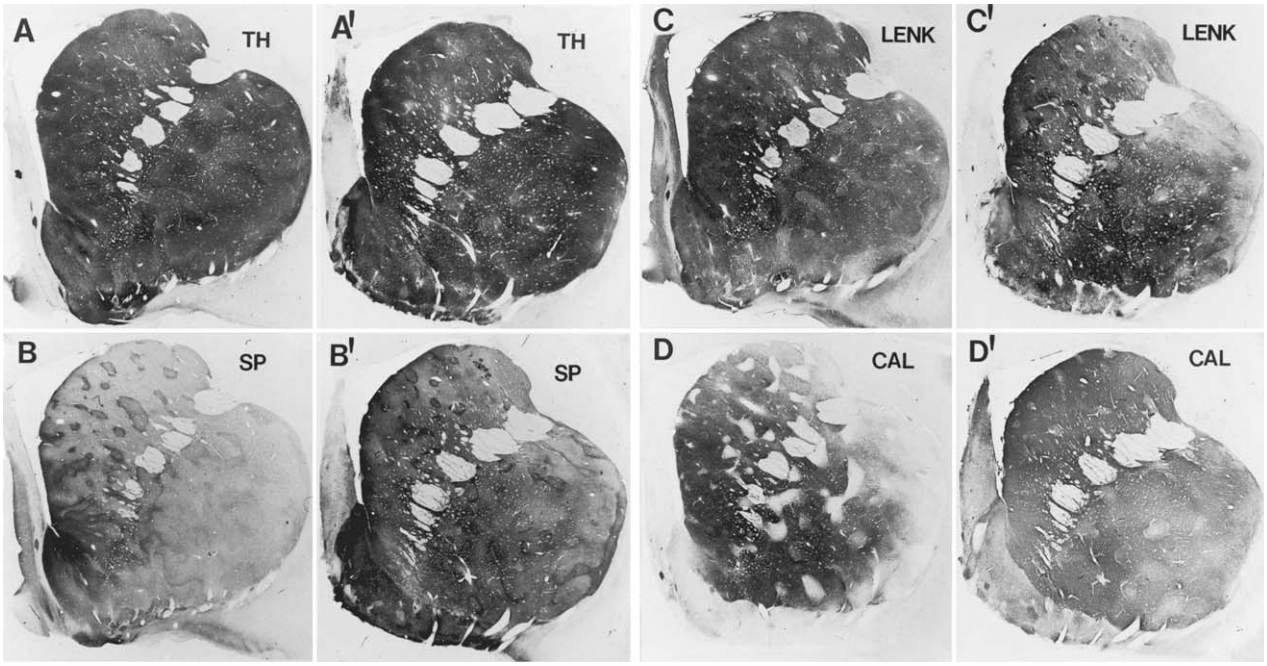


Figure 7 The primate striatum undergoes prominent remodeling of its chemical structure after birth. Antibodies were used to identify tyrosine hydroxylase (TH, a marker of dopamine innervation of the striatum), substance P (SP), leucine-enkephalin (LENK), and the calcium-binding protein calbindin (CAL, a striatal matrix marker) in brain sections of rhesus monkeys. The left image of each pair (A–D) shows the relative amounts and distributions of these striatal peptides–proteins in a monkey 4 months of age. The right images (A′–D′) show the same chemicals in the striatum of an adult monkey. In each pair, the caudate nucleus is on the left and the putamen is on the right. (A, A′) The dopaminergic innervation of the striatum is not fully established at 4 months of age, nor is the adultlike striosome–matrix organization. In the infant, the striosomes are slightly more enriched in dopaminergic innervation than the surrounding matrix. The reverse pattern is present in adults. (B, B′) The substance P pattern at 4 months is immature compared to the adult, but the differences appear to be of degree (intensity) and not in basic pattern (like TH). The caudate nucleus in the infant and adult has numerous striosomes enriched in substance P; however, striosomes in the infant putamen are less developed than the striosomes in the adult putamen. Furthermore, the matrix in the infant caudate and putamen is deficient in substance P compared to adult. Furthermore, the ventral striatum (the nucleus accumbens and olfactory tubercle) has not yet attained adultlike levels of substance P. (C, C′) The distribution and relative amount of enkephalin in the infant striatum are very different from those found in adults. In the 4-month-old monkey, the caudate nucleus has striosomes that are poor in enkephalin and a matrix that is enriched in enkephalin. The reverse pattern is present in the caudate nucleus in adults. (D, D′) The distributions of calbindin in the infant striatum and adult striatum also differ. At both ages the striosomes have less calbindin relative to the matrix; however, in the infant, the striosomes are essentially devoid of calbindin compared to those in the adult, and the infant striosomes are much larger. In the infant putamen (lateral region), the level of calbindin is deficient compared to that in the adult.

about areas of the brain that could be selectively vulnerable to environmental conditions and experiences in early life. We have put forth a hypothesis that neuropeptide transmitter systems in the striatum of infant rhesus monkeys undergo prolonged postnatal maturation, which may render this brain region highly vulnerable to environmental and experiential influences. So far, our studies of developing rhesus monkeys have revealed that peptide neurotransmitters undergo remarkable remodeling during the first year of postnatal life (Fig. 7). Many brain peptides in the striatal mosaic are present in early stages of life, but they need time to develop their proper distributions, amounts, and roles (Fig. 7).

It was hypothesized that an insufficient social environment for infant monkeys might interfere with the development of neurotransmitters in the basal ganglia, because animals raised under such conditions have abnormal movements and social behaviors. The brains of two groups of monkeys with different environmental rearing conditions have been studied: one group of monkeys that experienced social deprivation when they were infants and had many abnormal behaviors, and another group of monkeys that was raised in a highly social environment and behaved like normal monkeys.

It was discovered that the development of the structural organization of peptide neurotransmitter

systems of the basal ganglia is highly vulnerable to early postnatal experience. The normal compartmental arrangements of brain peptides of the major output striatal neurons and the localization of neuropeptides within synaptic targets of the striatum (globus pallidus and substantia nigra) are altered selectively in adult rhesus monkeys that experienced severe sensory–social deprivation during the first year of infancy (Fig. 8). Immature rhesus monkeys that have insufficient infant–mother and infant–peer interactions during early development manifest psychosocial abnormalities and motor impairments, including withdrawal and fearfulness, lack of play, apathy, learning deficits, stereotypic movements, and self-injurious behavior. Furthermore, as adults, these monkeys do not show appropriate parental behavior and are indifferent and abusive to their offspring. Certain brain areas are more vulnerable to the effects of social deprivation than other brain regions. Surprisingly, the neuropeptides in the amygdala–BST complex (Fig. 5) did not appear vulnerable, but the same neuropeptides were profoundly vulnerable in the striatal mosaic (Fig. 8). Because monkeys that experienced social–sensory deprivation have neurochemical alterations in their basal ganglia, and because the behavioral dysfunction induced by social–sensory deprivation of infant primates persists into adulthood and suggests functional abnormalities of the basal ganglia, it has been suggested that the normal postnatal development of peptide neurotransmitter systems in the basal ganglia, particularly the neostriatum, is partly dictated by the environment and experience.

These results were very novel and interesting because they showed, for the first time in a well-controlled experiment, that social deprivation can produce changes in brain chemistry in selectively vulnerable regions in animals that have brains very similar in many respects to the human brain. These changes may underlie the abnormal social, repetitive, and self-harmful behaviors of these monkeys, and they may be related to similar abnormal behaviors of children with developmental problems. Certain peptide neurotransmitters in the striatal mosaic and other basal ganglia areas of monkeys raised with insufficient social contacts do not develop and mature like they do in socially raised monkeys. Some of these altered neurotransmitters have important functions in signaling aversive messages. These findings are important because these animals, as well as children with some developmental disorders, can harm themselves. These experiments are also exciting because similar numbers of neurons are present in both groups of monkeys (i.e., they do not degenerate or die), although some neurons are unable to make the appropriate neurotransmitter. Lithium can increase the expression of enkephalin and substance P in neurons of the striatum of rats. Because the neurons appear to be still alive in these monkeys, they could be stimulated by drugs so that they can make their normal neurotransmitter, thereby possibly modifying the abnormal behaviors of these monkeys.

This seminal study by Martin and colleagues explored important relationships between the early experience of infants, the postnatal maturation of the brain, and the susceptibility of some neurotransmitters in certain brain areas to environmental factors. There are growing concerns about the long-term effects of parental–peer neglect or abuse and exposure to drugs or environmental toxins in early childhood on the development of brain structure and function. As a result of this study, it is now known that the basal ganglia is very vulnerable in infants. The implications may be profound. For example, environmental factors (social–sensory deprivation of varying severity) can interfere with the normal postnatal progression of brain peptides and other neurotransmitters in the maturing basal ganglia. However, the specific events producing these changes are not clear, and considerably more work is needed on how external or environmental experiences can turn on and off vulnerable genes in immature neurons of the brain. The normal maturational events that occur with neuropeptides in the basal ganglia of primates during postnatal development must be understood more completely before this theory of experience-dependent neurodevelopment of the basal ganglia can be clarified and before the possible cellular and molecular mechanisms responsible for the mutability of basal ganglia organization by early postnatal experience can be understood.

Several disorders that affect young people are characterized by abnormal conduct such as repetitive or stereotyped movements, poor skills in communication, self-preoccupation or withdrawal from family and society, and self-inflicted harm. Children with mental retardation and autism, Rett syndrome, and Lesch–Nyhan syndrome show some or all of these abnormal behaviors. In the United States, it was estimated that 160,000 young individuals with developmental disabilities exhibited harmful behaviors involving aggression toward themselves, other people, and property. The self-harmful behaviors of people with these disorders are primary reasons for institutionalization, failure in school environments,

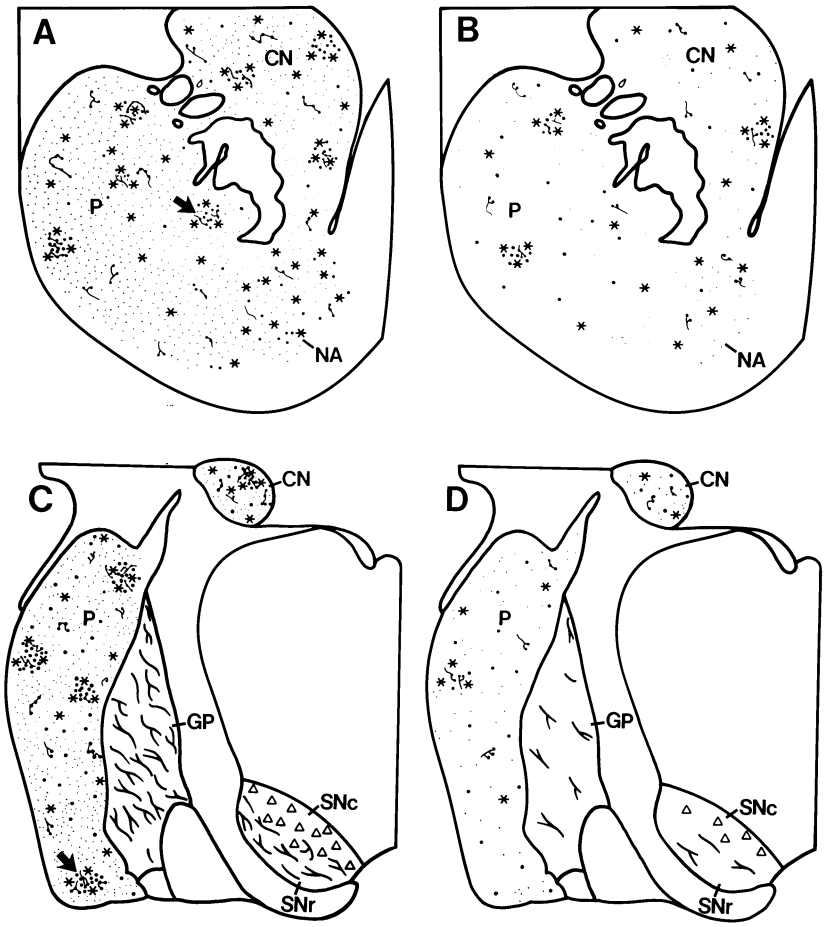


Figure 8 The chemical composition and structural organization of the primate basal ganglia are highly vulnerable to early postnatal experiences and environmental conditions. This diagram depicts some of the organizational features of the primate basal ganglia in monkeys raised in a socially healthy environment and their changes in monkeys raised in an inappropriate social environment. Abbreviations: CN, caudate nucleus; GP, globus pallidus; P, putamen; SNc, substantia nigra compact part; SNr, substantia nigra reticular part. In normal rhesus monkeys (A and C), substance-P-containing neurons (asterisks), processes (short, beaded lines), and axonal terminals (solid circles) are concentrated in some foci forming discrete patches (striosomes, black arrow) in the caudate and putamen. Striosomes are also enriched in enkephalin. The neuronal elements that comprise striosomes are found in lower amounts and diffuse throughout the surrounding matrix (the compartment that makes up the majority of the striatum). The nucleus accumbens has a matrix very enriched in substance P and enkephalin. The striatum receives an extensive synaptic terminal innervation (fine, small dots) from dopamine-utilizing neurons in the substantia nigra compact part (triangles). The neurons in the striatum that contain substance P and enkephalin innervate the globus pallidus and substantia nigra reticular part (smooth, branching lines). In monkeys that were raised in a socially inadequate environment (B and D), this organization of the basal ganglia is abnormal. These monkeys have fewer striosomes, they have less substance P and enkephalin in the matrix, and the dopamine innervation from the substantia nigra is deficient in the striatum. Also, striatal target regions (e.g., the globus pallidus and substantia nigra reticular part) are deficient in substance P and enkephalin due to the loss of innervation from substance-P- and enkephalin-expressing striatal neurons. The ability of neurons in the substantia nigra to make dopamine is impaired, although the neurons remain. [Reproduced with permission from *J. Neurosci.* **11**(11), 3344–3358, 1991, copyright the Society for Neuroscience].

and improper adjustment in communities. In some communities, 6–21 individuals out of 100 people have behavioral and emotional disorders, with a tendency for higher numbers in groups subject to social deprivation. The causes of these behavioral problems are unknown, although animal model experiments suggest a likely connection with brain peptides.

VI. ABNORMALITIES IN BRAIN PEPTIDES OCCUR IN BEHAVIORAL, PSYCHIATRIC, AND NEURODEGENERATIVE DISORDERS OF HUMANS

The participation of brain peptide neurotransmitters, cytokines, and growth factors in many forms of

behavioral and neurological disorders is now very evident (Tables V and VI). The most widely studied disorders in which brain peptides have been implicated are schizophrenia, mood disorders, Huntington's disease, Alzheimer's disease, and multiple sclerosis.

In schizophrenia, cholecystokinin neuronal systems seem to be the most vulnerable by some reports and are prominently depleted in the temporal cortex and hippocampus. Somatostatin, vasoactive intestinal peptide, substance P, neurotensin, and thyrotropin-releasing hormone have been reported to be abnormal in cerebral cortex, amygdala, and basal ganglia. Some of these findings have not been replicated. It is also very possible that ongoing drug treatment of patients diagnosed with schizophrenia will confound observations on brain peptide levels.

In depression, the hypothalamic–pituitary–adrenal axis and the hypothalamic–pituitary–thyroid axis have been examined thoroughly. It is believed that corticotropin-releasing hormone and thyrotropin-releasing hormone are hypersecreted in individuals with major depression. Furthermore, prodynorphin mRNA expression is elevated in the striosomal compartment of the striatal mosaic in individuals that commit suicide. Thus, dysfunction of the endogenous opioid dynorphin system might contribute to depression and the risk of suicide.

In neurodegenerative diseases, the major brain areas affected by the disease process tend to exhibit changes in neuropeptide systems. In Huntington's disease, substance-P- and enkephalin-containing neurons are lost in the striatum. Neurons in neocortex that contain neuropeptides are depleted in Alzheimer's disease, with somatostatin and corticotropin-releasing hormone showing high vulnerability. In neurodegenerative diseases of the human CNS that have prominent inflammatory components, such as multiple sclerosis, Alzheimer's disease, and AIDS, the proinflammatory cytokine tumor necrosis factor-α is up-regulated. Thus, brain and spinal cord inflammation and autoimmune reactions participate in the nervous system degeneration in these diseases. This idea is greatly emphasized by the finding that nonsteroidal anti-inflammatory drugs (such as aspirin, ibuprofen, and indomethacin) strongly protect against Alzheimer's disease.

The observation that anti-inflammatory drugs are helpful in Alzheimer's disease could be related to the discovery that astroglia and microglia are primary generators of amyloid deposits in the aging brain. Amyloid (Aβ) is a protein fragment fibril made up of a 4-kDa peptide consisting of 40–42 amino acid residues that is derived proteolytically from the amyloid precursor protein. Aβ is a major component of senile plaques, which are brain lesions that occur in individuals with Alzheimer's disease, Down's syndrome, and, less frequently, people aging normally. In addition to Aβ, these lesions have a complex composition, consisting of dystrophic neurites (damaged and swollen dendrites or axon terminals) and activated astrocytes and microglia. Our studies of aging monkeys revealed that senile plaques are dynamic brain lesions that evolve from early synaptic defects within the neuropil to mature plaques and extracellular deposits of Aβ. The staging of these lesions is thought to be the degeneration of neuritic structures, followed by the attraction of reactive glia, and the subsequent deposition of extracellular Aβ derived from microglia or

Table V
Behavioral, Psychiatric, and Neurological Disorders Involving Brain Peptides

Abnormality	Possible brain peptides involved
Stress–anxiety	Corticotropin-releasing hormone
Obesity	Neuropeptide Y
Anorexia–cachexia	Orexins, tumor necrosis factor-α, interleukins, leukemia inhibitory factor, ciliary neurotrophic factor
Sleep disorders	Orexins
Epilepsy	Substance P, somatostatin, galanin, neuropeptide Y
Migraine headache	Calcitonin gene-related peptide
Schizophrenia	Cholecystokinin
Mood disorders	Corticotropin-releasing hormone, thyrotropin-releasing hormone, dynorphin

Table VI
Neurodegenerative Disorders Involving Brain Peptides

Disorder	Brain peptide groups involved
AIDS dementia	Classical neuropeptides, cytokines, chemokines
Alzheimer's disease	Classical neuropeptides, growth factors, cytokines, chemokines
Amyotrophic lateral sclerosis	Growth factors, cytokines
Cerebral ischemia (cardiac arrest and stroke)	Growth factors, cytokines
CNS trauma	Growth factors, cytokines, chemokines
Multiple sclerosis	Cytokines, chemokines
Parkinson's disease	Growth factors

astrocytes. These experiments demonstrated that structural and biochemical perturbations within neuronal and nonneuronal cells occur before the deposition of extracellular Aβ fibrils. Furthermore, our results suggest that focal abnormalities in synaptic contacts within the neuropil (synaptic disjunction) may initiate a complex series of inflammatory events resulting in the formation of diffuse senile plaques and deposits of Aβ. In response to synaptic disjunction in the aged brain, astroglia and microglia produce Aβ. This event may be signaled by cytokines.

VII. BRAIN PEPTIDES MODULATE NEURONAL DEGENERATION AND REGENERATION IN EXPERIMENTAL MODELS

Classical neuropeptides as well as brain peptide growth factors and cytokines play important roles in neuronal degeneration and brain repair. It is now very evident that numerous small, diffusible peptide–protein factors modulate the growth, differentiation, survival, proliferation, and migration of neurons and/or glial cells. In the developing nervous system, neuronal death occurs normally and is thought to be necessary for regulating the size of neuronal groups in relation to target size and synaptic inputs. This developmental neuronal death is thought to be partially controlled by growth factors, including the conventional neurotrophic factors (Table II), as well as by some of the classical neuropeptides (Table I, e.g., somatostatin and vasoactive intestinal peptide). After brain and spinal cord injury, many cellular and molecular responses occur simultaneously or in phases, some of which are degenerative and others regenerative events. The vague concept for the release of tissue-soluble factors that may play a role in the degeneration, regeneration, and survival of nerve cells was introduced in the early 1900s by Santiago Ramón y Cajal. Subsequently, Levi-Montalcini, Hamburger, Prestige, and others established this concept. Many of the current ideas about the critical dependence of neurons on survival molecules or trophic factors (Table II) have been developed from experiments of cultured neurons and injured motor neurons *in vivo*.

A. Cell Culture Models

When neurons are removed from the developing nervous system, they require survival factors to grow in a dish. Some of these neuronal survival factors are neuroactive peptides and proteins (Tables I and II). Immature sympathetic ganglion neurons, cerebellar granule neurons, and cortical and hippocampal neurons are used commonly for culture experiments. The classical example for this concept is the survival of sympathetic ganglion neurons being dependent on serum factors such as nerve growth factor. When these cells are deprived of serum or nerve growth factor, they undergo apoptosis by a programmed cell death (PCD) mechanism that is dependent on specific cell death molecules. Bax and caspase-3 are critical for neuronal death by apoptosis, involving Bax translocation to mitochondria and mitochondrial release of cytochrome *c* to participate in the activation of caspase-3, which subsequently activates proteins including DNA fragmentation factor-45 that cleave genomic DNA. Nerve growth factor also controls sympathetic neuron survival by Ras suppression of a p53-mediated cell death pathway. Serum or potassium deprivation of cerebellar granule neurons and cortical neurons also results in apoptosis. Roles for neuropeptides in neuronal survival pathways have come to light. Apoptosis of granule neurons *in vitro* can be delayed by angiotensins II and IV, and angiotensinogen knockout mice have fewer granule neurons in the cerebellum and hippocampus than wildtype-mice, suggesting that the renin–angiotensin system has a role in neuronal survival by preventing apoptosis during the period of developmental cell death. In addition, pituitary adenylate cyclase activating polypeptide can block apoptosis of sympathetic ganglion and granule neurons by activation of the extracellular signal-regulated (ERK) type of MAP kinase via a cyclic-AMP-dependent pathway and proteosome-mediated degradation of caspase-3.

B. Animal Models

In animal models of spinal cord and nerve injury, brain peptide hormones, growth factors, and cytokines can be studied for their roles in degeneration and for their applicability for therapeutics. The extent of neuronal death or of regeneration and survival after spinal cord injury is influenced by the age of the animal at the time of injury and the location of axonal trauma in relation to motor neurons. Motor neurons with axonal damage in the immature CNS die rapidly. In contrast, in some settings, axotomized motor neurons in the adult CNS are more likely to persist in some altered form. This observation

could have major importance with regard to the recovery of neurological function in infants, children, and adults.

Cytokines (Table III) have stimulated much excitement in the pathobiology of spinal cord and peripheral nerve injury. A cytokine called leukemia inhibitory factor might be important for the successful regeneration of injured sensory and motor axons in immature and adult animals by altering neuronal gene expression. Other molecules that may be relevant to neuronal degeneration and regeneration after spinal cord and peripheral nerve injury are inflammatory factors. Tissue inflammation is believed to be a critical component of the degenerative process after nervous system injury. Following CNS damage, a variety of changes occur, including the extravasation of serum proteins through the compromised blood–brain barrier, activation of neuroglial cells, and rapid synthesis and liberation of proinflammatory cytokines. A major proinflammatory cytokine in the CNS is tumor necrosis factor-α, which promotes inflammation and cell death through its p55 receptor. Tumor necrosis factor-α also promotes cell death by inhibiting receptor signals initiated by key survival peptides such as insulin-like growth factor. Thus, brain peptide cytokines can interact antagonistically or synergistically with brain peptide growth factors. Interestingly, the remarkable anti-inflammatory and antipyretic actions of α-melanocyte-stimulating hormone occur by inhibition of glial-derived tumor necrosis factor-α. Other cytokines function as potent anti-inflammatory factors. Interleukin-10 reduces the production of tumor necrosis factor-α by astroglia. Interleukin-10 reduces inflammation and improves functional outcome in humans and in animal models of spinal cord injury. In contrast, interleukin-3 appears to cause spinal cord damage. Overexpression of interleukin-3 in mice leads to motor neuron degeneration and progressive muscular atrophy, suggesting that an autoimmune reaction may participate in the degenerative process in humans with amyotrophic lateral sclerosis.

VIII. THE FUTURE OF CNS PEPTIDE RESEARCH HAS GREAT PROMISE

The use of animal models will be necessary to advance our understanding of brain peptide function and their roles in nervous system development, maturation, degeneration, and repair. An interesting feature of the different classical neuropeptide systems is that many of these brain peptides undergo prominent changes during development (Fig. 7). However, it is not yet clear whether these changes direct maturational events involved in the pattern formation of brain structure or whether changes in the expression are consequences of brain maturation. For example, brain peptides and hormones may be critical for pattern formation during development of the cerebral cortex and striatum, and abnormalities may cause neurodevelopmental defects. With regard to CNS degeneration, paradigms should include animal models of brain and spinal cord injury and aging. Animal models of CNS degeneration are crucial for improving our understanding of the mechanisms and stages of degenerative and repair processes. These models can provide an *in vivo* system to identify how neurons and neuroglia change in paradigms that mirror certain neuropathological and clinical features of neurological conditions that occur in humans (Table VI). With animal models, the process of nervous system degeneration and repair can be studied directly at the structural and molecular levels, the roles for brain peptides and hormones can be identified, and the model subsequently can be used to test new therapies to prevent the neuropathology and improve neurologic recovery in a biologically relevant system. In addition, the application of molecular genetics and transgenic mouse technology will be essential. The development of transgenic mouse systems for peptide growth factors and cytokines should be expanded.

Pharmaceutical drug development centering around neuroactive brain peptides holds immense promise for the future treatment of a wide range of neurobehavioral, neurological, and neurodegenerative disorders (Tables V and VI). For example, the design of drugs that mimic the actions of satiety neuropeptides could be useful for weight management in obesity or blocking the functions of brain satiety peptides for the treatment of anorexia or cachexia in chronically or terminally ill patients. The neuropeptides that function in sleep–wakefulness could be targets for sleep disorders. The anti-inflammatory actions of certain cytokines and the blockade of proinflammatory cytokines could be exploited for the treatment of Alzheimer's disease, multiple sclerosis, head and spinal cord injury, and cerebral ischemia after stroke and cardiac arrest. Furthermore, the well-known neuronal survival actions of neurotrophins should continue to be targets for drug discovery as well as the neuronal survival effects of classical and more recently identified brain peptides. The application of research on survival peptides has enormous potential for the treatment of

Alzheimer's disease, Parkinson's disease, amyotrophic lateral sclerosis, and cerebral ischemia, and CNS trauma (Table VI).

The current inability to effectively and rationally help people with nervous system abnormalities is in large part due to the of lack of understanding of how neurons communicate with other neurons and how neurons communicate with neuroglia or other non-neuronal cells. It is very likely that future discoveries in brain peptides and hormones will add to the formulation of new therapies for humans with neurobehavioral disorders and neurodegenerative conditions.

See Also the Following Articles

ASTROCYTES • CHEMICAL NEUROANATOMY • CIRCADIAN RHYTHMS • ENDORPHINS AND THEIR RECEPTORS • GLIAL CELL TYPES • HOMEOSTATIC MECHANISMS • NEUROPEPTIDES AND ISLET FUNCTION • NEUROTRANSMITTERS • PSYCHONEUROENDOCRINOLOGY • SEXUAL DIFFERENTIATION, HORMONES AND • STRESS: HORMONAL AND NEURAL ASPECTS

Acknowledgments

Dr. Martin is supported by grants from the U.S. Public Health Service, National Institutes of Health (NS34100 and AG16282), and the U.S. Army Medical Research and Material Command (DAMD17-99-1-9553). The photographic assistance of Frank Barksdale is appreciated greatly. This article is dedicated to the tireless efforts of my wife Laura.

Suggested Reading

Heimer, L., de Olmos, J., Alheid, G. F., and Záborszky, L. (1991). "Perestroika" in the basal forebrain: Opening the border between neurology and psychiatry. *Prog. Brain Res.* **87,** 109–165.

Hökfelt, T. (1991). Neuropeptides in prespective: The last ten-years. *Neuron* **7,** 867–879.

Martin, L. J., Spicer, D. M., Lewis, M. H., Gluck, J. P., and Cork, L. C. (1991). Social deprivation of infant rhesus monkeys alters the chemoarchitecture of the brain: I. Subcortical regions. *J. Neurosci.* **11,** 3344–3358.

Martin, L. J., Powers, R. E., Dellovade, T. L., and Price, D. L. (1991). The bed nucleus–amygdala continuum in human and monkey. *J. Comp. Neurol.* **309,** 445–485.

Martin, L. J., Hadfield, M. G., Dellovade, T. L., and Price, D. L. (1991). The striatal mosaic in primates: Patterns of neuropeptide immunoreactivity differentiate the ventral striatum from the dorsal striatum. *Neuroscience* **43,** 397–417.

Martin, L. J., Pardo, C. A., Cork, L. C., and Price, D. L. (1994). Synaptic pathology and glial responses to neuronal injury precede the formation of senile plaques and amyloid deposits in the aging cerebral cortex. *Am. J. Pathol.* **145,** 1358–1381.

Martin, L. J., Portera-Cailliau, C., Ginsberg, S. D., and Al-Abdulla, N. A. (1998). Animal models and degenerative disorders of the human brain. *Lab Animal* **27,** 18–25.

Martin, L. J., Price, A. C., Kaiser A., Shaikh, A. F., and Liu, Z. (2000). Mechanisms for neuronal degeneration in amyotrophic lateral sclerosis and in models of motor neuron death. *Int. J. Mol. Med.* **5,** 3–13.

McGeer, E. G., and McGeer, P. L. (1998). Inflammation in the brain in Alzheimer's disease: Implications for therapy. *NeuroScience News* **1,** 29–35.

Peripheral Nervous System

ANDREI V. KRASSIOUKOV
University of Toronto

GLOSSARY

axon Neuronal process that carries impulses away from the neuronal cell body, usually covered by myelin.

afferent–afferent fiber An axon that carries sensory information to the central nervous system.

dermatome A striplike projection on the skin supplied by sensory neurons of a single spinal segment.

efferent–efferent fiber An axon that carries impulses away from the central nervous system.

effector The target organs, which respond to the stimuli carried by the efferent part of the peripheral nervous system.

ganglion A collection of neuronal cell bodies outside the central nervous system.

Node of Ranvier A gap in the myelin sheath of an axon where the action potential occurs during saltatory conduction.

Schwann cell A myelin-containing cell in the peripheral nervous system.

synapse A specialized structure at the point of contact between a neuron and the next in the circuit, where unidirectional transmission of information occurs.

Wallerian degeneration The changes that occur in peripheral nerves distally from the site of the injury.

The nervous system is made of cells specialized for the transmission of information, called nerve cells or neurons. Neurons are the building blocks of both the central and the peripheral nervous systems. Each neuron has a cell body (soma) and processes (axon and dendrites), which carry information to and away from the cell body. The peripheral nervous system connects the central nervous system to the different tissues and organs of the body. The cranial, spinal, and autonomic nerves and their ganglia form the peripheral nervous system. From this article, the reader will learn about anatomical and physiological subdivisions of the peripheral nervous system. Some practical applications of neuroanatomical facts regarding the peripheral nervous system will also be presented.

I. GENERAL ORGANIZATION OF THE PERIPHERAL NERVOUS SYSTEM

Traditionally, the nervous system is divided into central and peripheral components. The brain (cerebrum, cerebellum, and brain stem) and spinal cord form the central nervous system (CNS). The peripheral nervous system (PNS) connects the CNS to the different tissues and organs of the body. The modern view of the organization of the PNS is based on the work of the English physiologist John Newport Langley (1852–1925). The PNS has two components, a sensory (afferent) and a motor (efferent) component. The sensory component, also known as the afferent component (*ad* meaning to + *ferre* meaning carry), is responsible for conveying information to the CNS from the body itself and from the environment (Fig. 1). We receive sensory information through the sensory endings (receptors) that are scattered throughout the body. These receptors are biological transducers,

associated with the peripheral ends of afferent axons, which transform different stimuli into action potentials in these axons. Sensory nerve fibers are the axons of a group of neurons situated within the dorsal root ganglia (DRG) at the distal end of dorsal roots (Fig. 1). These ganglia are also known as sensory or spinal ganglia and convey information from the body. The sensory information from the face travels through the specialized cranial nerves. The motor or efferent (*ex* meaning from + *ferre*) component carries information from the CNS to the muscles. The target organs, which respond to the stimuli carried by the efferents, are called effectors (Fig. 1). The motorneurons that innervate skeletal muscles are localized within the ventral horns of the spinal cord.

The elements of the efferent component of the PNS can be further categorized into a somatic and an autonomic division. The somatic division controls skeletal muscles and acts mainly under the direction of conscious voluntary control from the brain. The autonomic division innervates smooth muscles, cardiac muscle, and glands. The term autonomic comes from *auto* (meaning self) and *nomos* (meaning law). The autonomic nervous system exerts control over the functions of many organs and brings the fine internal adjustments necessary for the maintenance of the optimal internal environment of the body. This system functions according to its own internal laws, largely unconsciously.

Neurons do not function in isolation; they are organized into specific neuronal circuits that process particular different kinds of information and control particular efferent organs. The simplest type of circuit is called a "reflex" or reflex arc and involves the interaction of sensory and motor neurons. The effector response mediated by such a circuit is called a reflex response or simply a reflex. Some reflexes are extremely important in the diagnosis and localization of neurologic problems. Several neuronal components of the PNS and CNS are involved in a reflex arc. The first component of the reflex arc is a receptor, such as a special cutaneous receptor or a neuromuscular spindle, the stimulation of which initiates an action potential in afferent fibers. The second component is an afferent neuron (sensory neuron with a soma within the DRG), which transmits impulses though the peripheral nerves to the CNS. The third component is an interneuron within the CNS, which relays the information to the efferent neuron. In some reflexes, known as monosynaptic reflexes (such as a stretch reflex, the circuits of which are shown in Fig. 1), this component is missing. The next component of the reflex arc is the efferent (motor) neuron, which delivers information from the CNS to the effector organ. At the point of contact between a neuron and the effector, where unidirectional transmission of information is occurs, there is a specialized structure called a synapse. Finally, the effector, such as a muscle or gland, will respond to the action potential delivered through the efferent fibers. Interruption of the reflex arc at any point will abolish the response.

II. GROSS ANATOMY OF THE PERIPHERAL NERVOUS SYSTEM

The peripheral nervous system consists of the cranial and spinal nerves and their associated ganglia. There are 12 pairs of cranial nerves, which exit the skull through various openings called foramina. There are

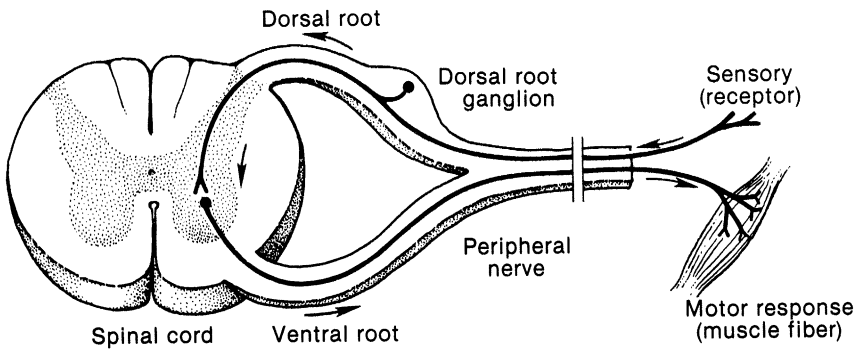


Figure 1 Monosynaptic reflex arc, consisting of a primary sensory neuron activating a motoneuron by a single synapse. Adapted from DeMeyer, W. *Neuroanatomy*, Ch. 2, Copyright © 1988. Reprinted by permission of John Wiley & Sons, Inc.

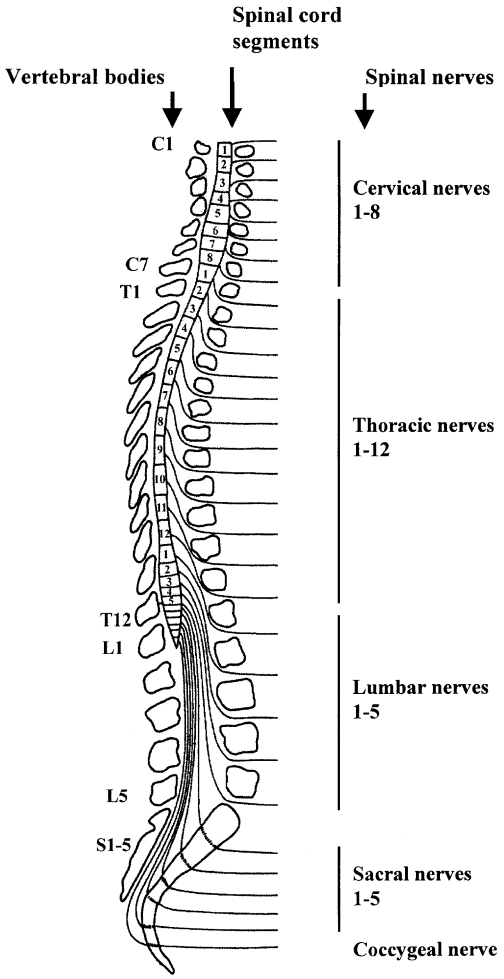


Figure 2 A schematic diagram of the relationship of spinal cord segments and spinal nerves to the vertebral column. The relation between the level of the spinal nerves–spinal segments and the vertebral body is clinically important. Note that the spinal cord extends only to vertebrae L1–L2, but the spinal roots continue down to the appropriate intervertebral foramina. Most spinal levels do not correspond to the same vertebral level. Abbreviations: C, cervical; T, thoracic; L, lumbar; S, sacral; Co, coccygeal.

31 pairs of spinal nerves, each identified by its association with the vertebra where the nerve leaves the vertebral canal (Fig. 2). The spinal cord segment from which a given spinal nerve arises is identified in the same way. The cervical region of the spinal cord gives rise to 8 cervical spinal nerves (C1–C8), the thoracic region to 12 thoracic spinal nerves (T1–T12), the lumbar region to 5 lumbar spinal nerves (L1–L5), the sacral region to 5 sacral spinal nerves (S1–S5), and finally the coccygeal region to 1 coccygeal spinal nerve (Co1). A spinal nerve is made from the confluence of the ventral (efferent, motor) and dorsal (afferent, sensory) roots distal to the DRG (Fig. 1). From the spinal nerve, axons leave through a dorsal and a ventral ramus, which then form peripheral nerves, providing motor and sensory innervation to the whole body.

To free the PNS from the CNS, we cut the spinal nerve roots just at their attachment to the spinal cord (Fig. 1). Each spinal nerve (except for C1 and Co1) is attached to the spinal cord by dorsal (posterior) and ventral (anterior) roots (Fig. 1). Each of these individual roots is broken up into a series of rootlets that are attached to the spinal cord. Spinal nerve C1 often lacks a dorsal root, and the coccygeal nerve may be absent. The dorsal and ventral roots take a lateral and descending course within the subarachnoid space and pierce the arachnoid and dura mater, at which point these meninges become continuous with the epineurium as they approach the intervertebral foramina. The dorsal and anterior roots unite immediately beyond the DRG and form the spinal nerves, which then exit through the intervertebral foramina. Spinal nerves may extend directly to their target tissue, as is seen in intercostal nerves, or they may form a complex structure such as a plexus in which the nerve trunk may interchange fibers with neighboring trunks. The spinal nerves form three major plexuses—the cervical, brachial, and lumbosacral. The cervical plexus (Fig. 3) is formed by the spinal nerves of the upper cervical segments (C1–C4). The phrenic nerve arises from the cervical plexus. This nerve innervates the diaphragm, the most important muscle for breathing. The last four cervical and the first thoracic spinal nerves (C5–T1) form the brachial plexus (Fig. 4). Numerous nerves arise from various parts of the brachial plexus and innervate the skin, muscles, and bone of the shoulder girdle, upper chest wall, arm, and hand. Like the brachial plexus, the lumbosacral plexus (L1–S4) gives off a number of nerves to paravertebral and pelvic girdle muscles and to the lower extremities (Fig. 5).

A. Clinical Note

The peripheral nerves that rise from a plexus convey sensory and motor information. An area of the skin innervated by a single spinal nerve, and, therefore, a single segment of the spinal cord, is called a dermatome (*derma* meaning skin). The distribution of dermatomes on the anterior and posterior surfaces of the body is shown in Fig. 6. The cutaneous areas supplied by adjacent spinal nerves overlap. Therefore, only little

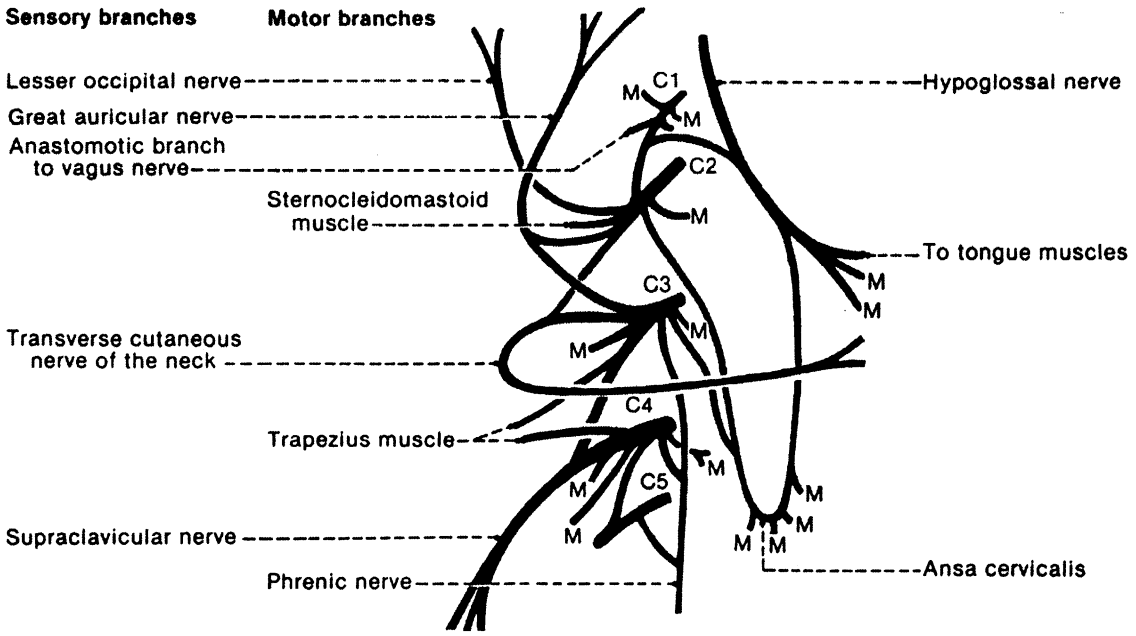


Figure 3 Frontal view of the right cervical plexus. Abbreviations: M, motor to striated muscles; see legend to Fig. 2. From DeMeyer, W. *Neuroanatomy*, Ch. 5, Copyright © 1988. Reprinted by permission of John Wiley & Sons, Inc.

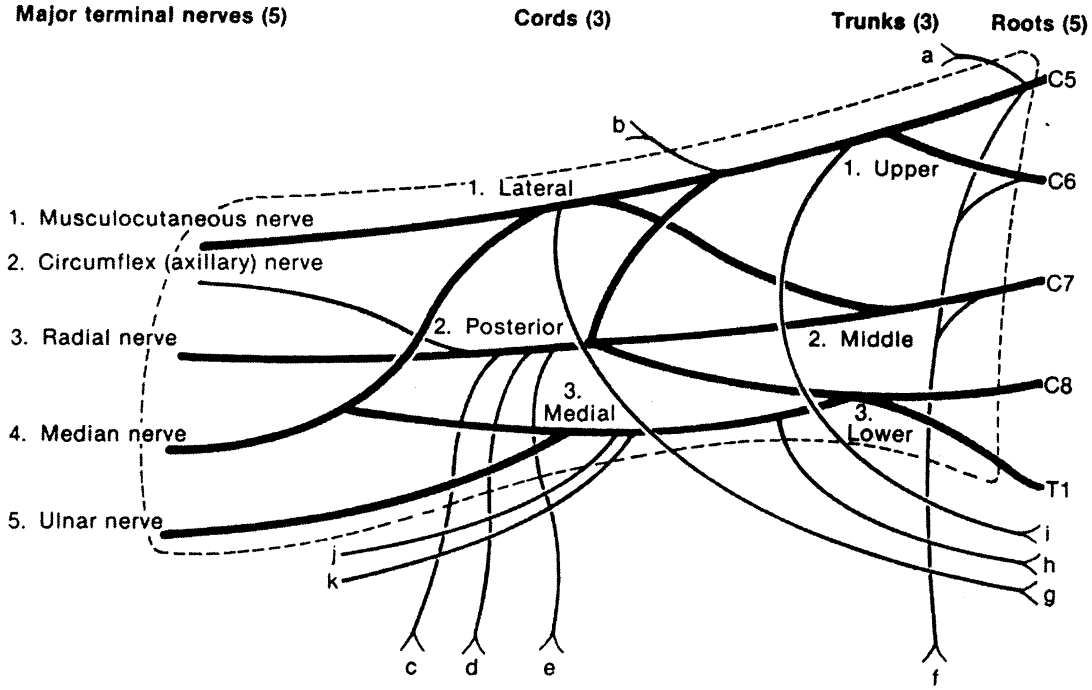


Figure 4 Frontal view of the right brachial plexus. The dashed line encloses the major roots and the largest branches. The trunks are the nerve fibers before they enter the plexus, cords are a redistribution of fibers in the plexus, and terminal nerves are the way out of the plexus. a, dorsal scapular nerve to rhomboid muscle; b, suprascapular nerve to supra- and infraspinatus muscles; c, inferior subscapular nerve to teres major muscle; d, thoracodorsal nerve to latissimus dorsi muscle; e, superior subscapular nerve to subscapular muscle; f, long thoracic nerve to serratus anterior muscle; g, lateral pectoral nerve to pectoralis muscle; h, medial pectoral nerve to pectoralis muscle; i, subclavian nerve to subclavius muscle; j, medial cutaneous nerve to the forearm; k, medial cutaneous nerve to the arm. From DeMyer, W. *Neuroanatomy*, Ch. 5, Copyright © 1988. Reprinted by permission of John Wiley & Sons, Inc.

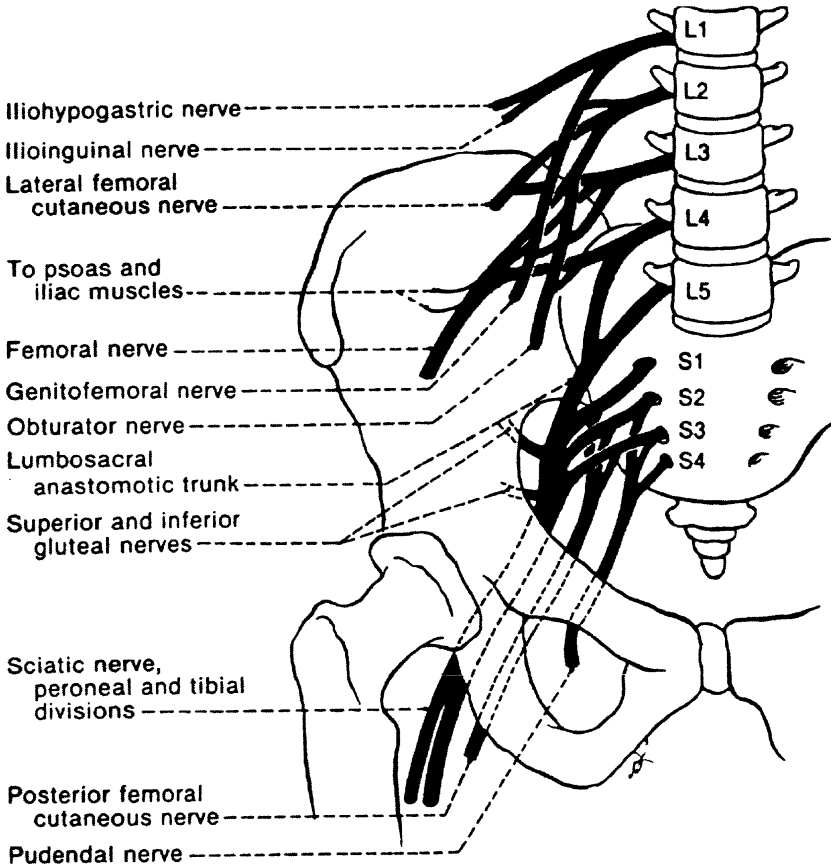


Figure 5 Frontal view of the right lumbosacral plexus, showing its origin in relation to vertebral levels and to the pelvis. From DeMyer, W. *Neuroanatomy*, Ch. 5, Copyright © 1988. Reprinted by permission of John Wiley & Sons, Inc.

sensory loss, if any, is observed following injury of a single dorsal root of the spinal cord. The dermatomes that must be known to solve neurological problems are summarized in Table I.

Skeletal muscles also receive segmental innervation. It is important to recognize that skeletal muscles are innervated by more than one spinal nerve and, therefore, by more than one spinal segment. The segmental innervation of some skeletal muscles is of clinical importance, because it is possible to test them by eliciting simple muscle reflexes (see the section on stretch reflexes) in the patient. Table II summarizes some muscle reflexes and muscle innervation by spinal segments and peripheral nerves. Testing of sensory and motor functions is an important part of the neurological examination and is particularly important in determining the level of injury. For example, the lesion of a segment of the spinal cord, or dorsal root, will result in a sensory loss and/or a motor deficit that is different from that occurring after the lesion of a peripheral nerve.

III. HISTOLOGY OF THE PERIPHERAL NERVOUS SYSTEM

A. Dorsal Root Ganglia

The dorsal root ganglia contain unipolar neuronal cells. A single process leaves rounded or oval-shaped cell bodies and bifurcates into peripheral and central branches or processes. The former terminate in a peripheral sensory ending (sensory receptors), and the latter usually enter the spinal cord through the dorsal root. Neurons within the DRG are closely surrounded by a layer of satellite cells. Each ganglion is covered by a layer of connective tissue that is continuous with the wrapping of the peripheral nerves.

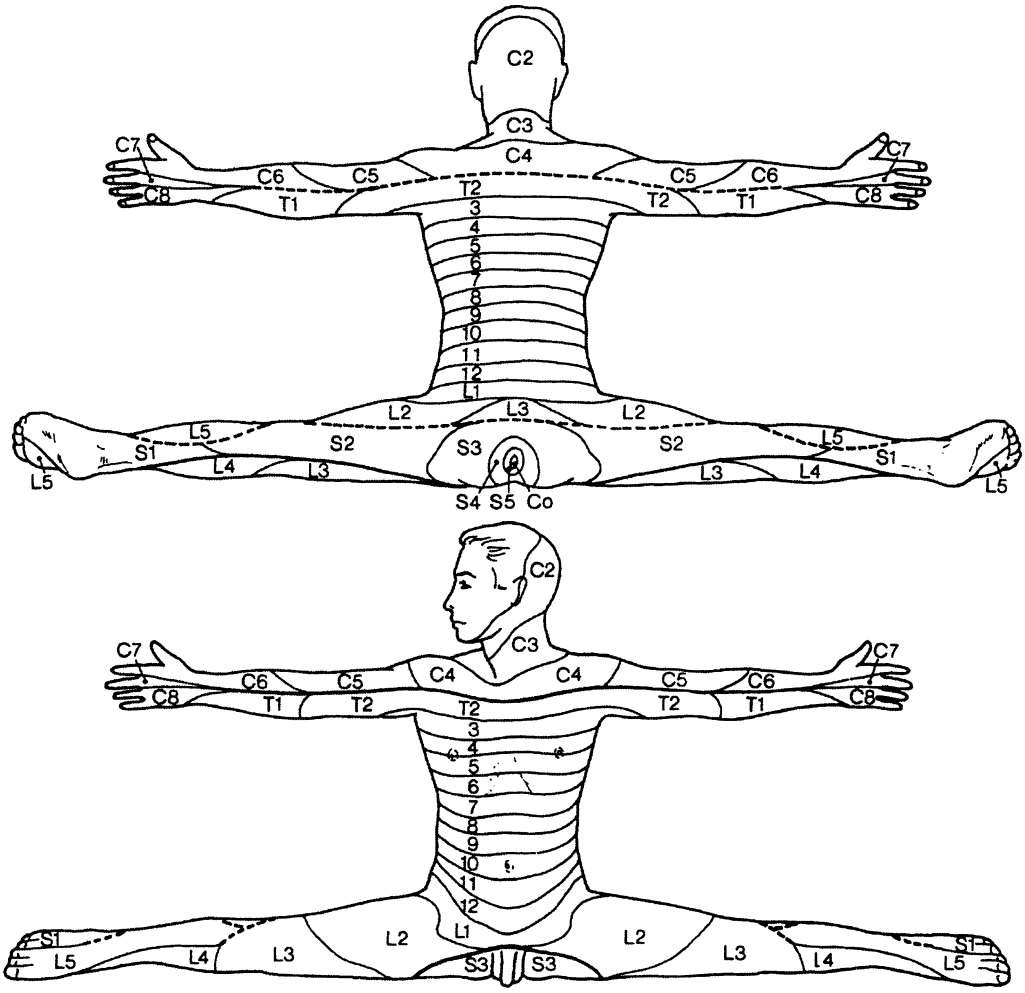


Figure 6 Pattern of dermatomal distribution and innervation by the spinal roots. Abbreviations same as for Fig. 2. From DeMyer, W. *Neuroanatomy*, Ch. 5, Copyright © 1988. Reprinted by permission of John Wiley & Sons, Inc.

B. Peripheral Nerves

Peripheral nerves are bundles of axons of motor and sensory neurons within the PNS enveloped by Schwann cells. Schwann cells form and maintain the myelin sheaths within the PNS. In contrast to an oligodendrocyte (a myelin-producing cell within the CNS), the Schwann cell myelinates only a small segment of an axon (the internode), whereas a single oligodendrocyte can myelinate several axons simultaneously. A majority of axons within the PNS receive circular wrapping by the surface membrane of Schwann cells, forming myelinated axons (Fig. 7). Other axons within the PNS merely indent the surface membrane of Schwann cells, forming unmyelinated axons. The segment of axon myelinated by one Schwann cell constitutes an internode. Between each internode (Schwann cell) is a short, myelin-free segment of axon called a node of Ranvier (Fig. 7). A similar segmentation of myelin sheaths occurs in the CNS. However, the gap between adjacent Schwann cells at the node is much smaller than that between adjacent oligodendrocytes of the CNS. External to Schwann-cell-enveloped axons, a layer of collagen fibers and scattered fibrocytes form a sheath called the endoneurium (Fig. 8). Bundles of endoneurium-ensheathed axons, called fascicles, are enveloped in layers of connective tissue called perineurium. The entire nerve is covered by the epineurium, which consists of fibrocytes and collagen fibers. This loose layer also contains blood and lymphatic vessels. These three sheaths with their interwoven collagen fibers provide

Table I
Essential Dermatomes

Spinal cord segment	Cutaneous innervation
C2	Back of the head
C3	Neck
C6	Thumb
C7	Middle finger
C8	Small finger
T4	Nipple region
T10	Umbilical region
L1	Inguinal ligament
L4	Big toe
S1	Small toe
S5	Perianal region

strength with flexibility and protection to the peripheral nerves.

IV. CLASSIFICATION OF NERVE FIBERS

Myelin has an important effect on the speed of the action potential conduction within the axon. By acting as an electrical insulator, myelin greatly speeds up action potential propagation. The action potential skips electrically from node to node; therefore, the transmission of a nerve impulse along a myelinated fiber is called saltatory conduction (*saltare* meaning to jump). Myelinated axons can conduct at velocities up to 120 m/sec, whereas unmyelinated axon conduction velocities range from about 0.5 to 3 m/sec. Human peripheral nerves contain fibers ranging in size from 1 to 20 μm in diameter (myelin sheath included). A commonly used classification of peripheral nerve fibers is based on fiber diameter, which correlates with the rate of conduction of nerve impulse. Each category and subcategory of nerve fibers arc associated with one particular type of nerve conduction (Table III). In general, the larger the axon, the thicker the myelin sheath and the faster the conduction velocity.

V. SENSORY RECEPTORS AND EFFECTOR ENDINGS

A. Sensory Receptor Endings: General Considerations

We receive information from the outside world or from within the body by special sensory nerve endings or sensory receptors. The sensory endings are scattered widely throughout the body. There are two categories of sensory endings and afferent neurons: somatic and visceral. The somatosensory afferents convey information from sensory receptors of the skin, muscles,

Table II
Summary of Some Muscle Reflexes

Reflex	Spinal segments	Plexus	Peripheral nerve	Expected action
Biceps brachii tendon reflex	C5–C6	Brachial	Musculocutaneous nerve	Flexion of the elbow joint by tapping the biceps tendon
Triceps tendon reflex	C6–C8	Brachial	Radial nerve	Extension of the elbow joint by tapping the triceps tendon
Brachioradialis tendon reflex	C7–C8	Brachial	Radial nerve	Supination of the radioulnar joints by tapping the insertion of the brachioradialis tendon
Abdominal reflexes	T7–T12	N/A	T7–T12	Contraction of underlying abdominal muscles by stroking the skin
Patellar tendon reflex (knee jerk)	L2–L4	Lumbar	Femoral nerve	Extension of knee joint on tapping the patella tendon
Achilles tendon reflex (ankle jerk)	L5–S2	Sacral	Tibial nerve	Plantar flexion of ankle joint on tapping the Achilles tendon
Plantar superficial reflex	S1, S2	Sacral	Tibial nerve	Flexion of toes on firmly stroking the sole of the foot

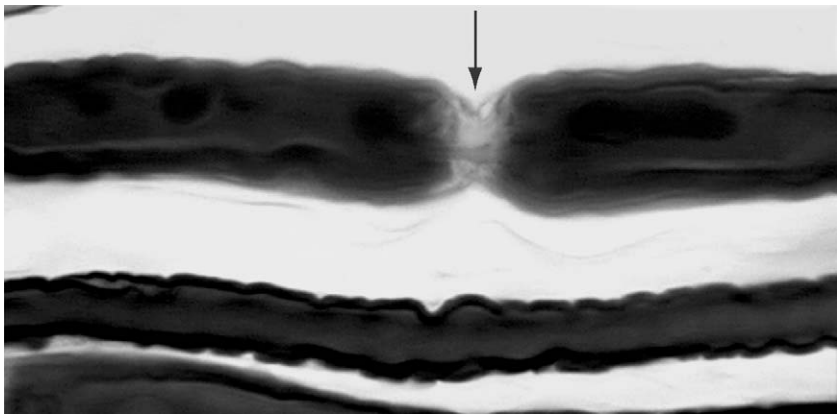


Figure 7 Myelinated nerve fibers (longitudinal section). Arrow points to the node of Ranvier. Stained with osmic acid.

and joints (general somatic afferents). The visceral afferents process information from the sensory receptors of the body cavities and internal organs (general visceral afferents). The general visceral afferents mostly travel within the cranial nerves. On the basis of localization within the body, sensory receptors can also be divided into three groups. Sensory endings that are superficially localized, such as those in the skin, are called *exteroceptors*. These receptors respond to stimuli for pain, temperature, touch, and pressure (Fig. 9 and Table IV). The second group is the *proprioceptors*, localized in muscles, tendons, and joints that monitor the position of our body in space. Finally, the third group of sensory endings is the *interoceptors*, localized in the viscera. On the basis of the sensory stimulus to which receptor endings respond, they can also be classified as mechanoreceptors, photoreceptors, chemoreceptors, thermoreceptors,

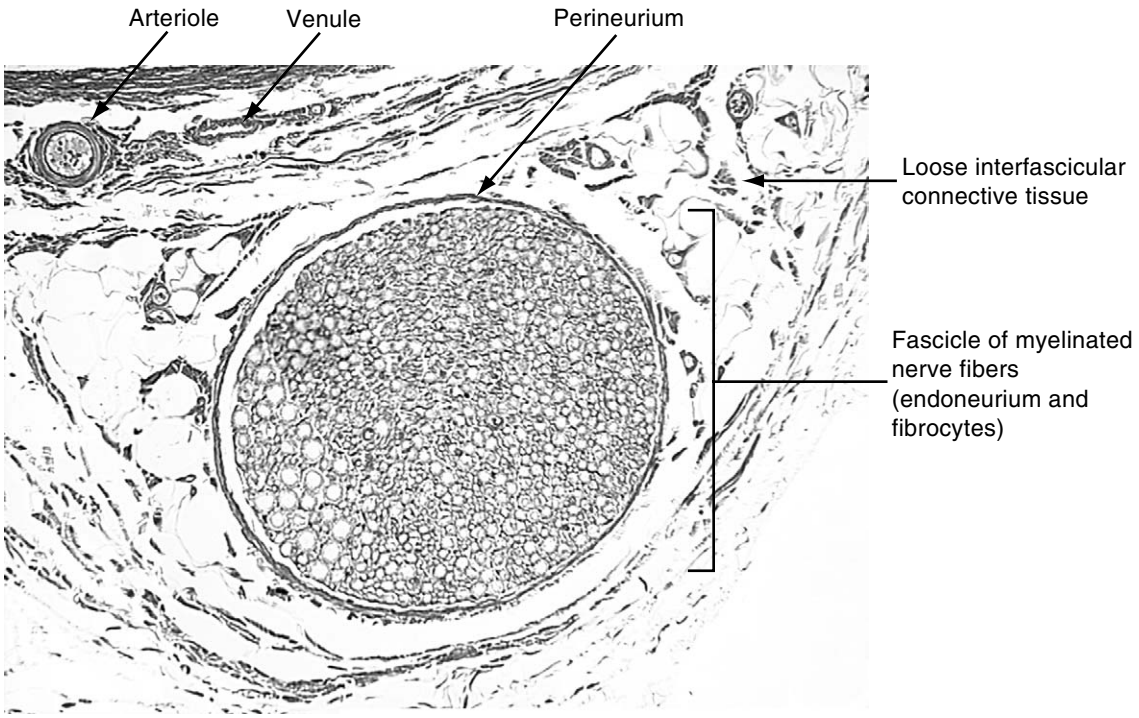


Figure 8 Peripheral nerve (transverse section). Stained with hematoxiline–eosin.

Table III
Classification of Nerve Fibers by Conduction Velocity, Diameter, and Function

Fiber type	Fiber diameter (μm)	Conduction velocity (m/sec)	Function
Somatic and visceral efferents			
A	12–20	70–120	To extrafusal fibers of skeletal muscles from α motor neurons
	2–8	10–50	To intrafusal fibers of skeletal muscles from γ motor neurons
B	<3	3–15	Preganglionic fibers to autonomic ganglia
C	0.2–1.2	0.7–2.3	Postganglionic autonomic fibers to smooth muscles and glands
Cutaneous afferents			
A (α)	12–20	70–120	From joint receptors
A (β)	5–12	30–70	From pacinian corpuscles (vibration) and touch receptors
A (ϵ)	2–5	6–30	From touch, temperature, and pain (sharp, localized) endings
C	<2	0.5–2	From pain and temperature endings
Muscle afferents			
Iα	12–20	70–120	From muscle spindles (anulospiral endings)
Iβ	12–20	70–120	From Golgi tendon organs
II	5–12	30–70	From muscle spindles (flower spray endings)
III	2–6	4–30	From pressure–pain endings
IV	<2	0.5–2	From pain endings

and nociceptors. Finally, on a structural basis, nerve endings can be divided into two classes: nonencapsulated, which may lie freely in the connective tissue, and encapsulated, which are enclosed by a distinctive arrangement of nonneuronal cells.

B. General Sensory Receptors of the Skin

Different types of skin receptors are presented in Fig. 9. For convenience, a detailed comparison of receptors, axon type, and sensory modality is summarized in Table IV.

C. Musculoskeletal Innervation

1. Sensory Endings

Proprioceptors in the joints, muscles, and tendons provide the CNS with the information required for awareness of the position of the body and for coordinated movements (Table IV). These complex actions are possible through the involvement of numerous reflex mechanisms.

a. Neuromuscular Spindles Proprioceptors contained in skeletal muscles are the neuromuscular spindles (muscular spindles). A single spindle is a bundle of specialized muscle (intrafusal) fibers that is a fraction of a millimeter wide, up to 6 mm long, and surrounded by a fusiform capsule of connective tissue (Fig. 10). The spindles lie in the long axis of the muscle and are especially numerous near the tendinous insertion of a muscle. Each spindle has two types of muscle fibers, two types of afferents, and one type of efferent. Within the capsule are numerous (up to 14) intrafusal muscle fibers. The latter are considerably smaller than the extrafusal fibers situated outside the spindle, which comprise the main mass of a muscle. Two types of intrafusal fibers can be identified within

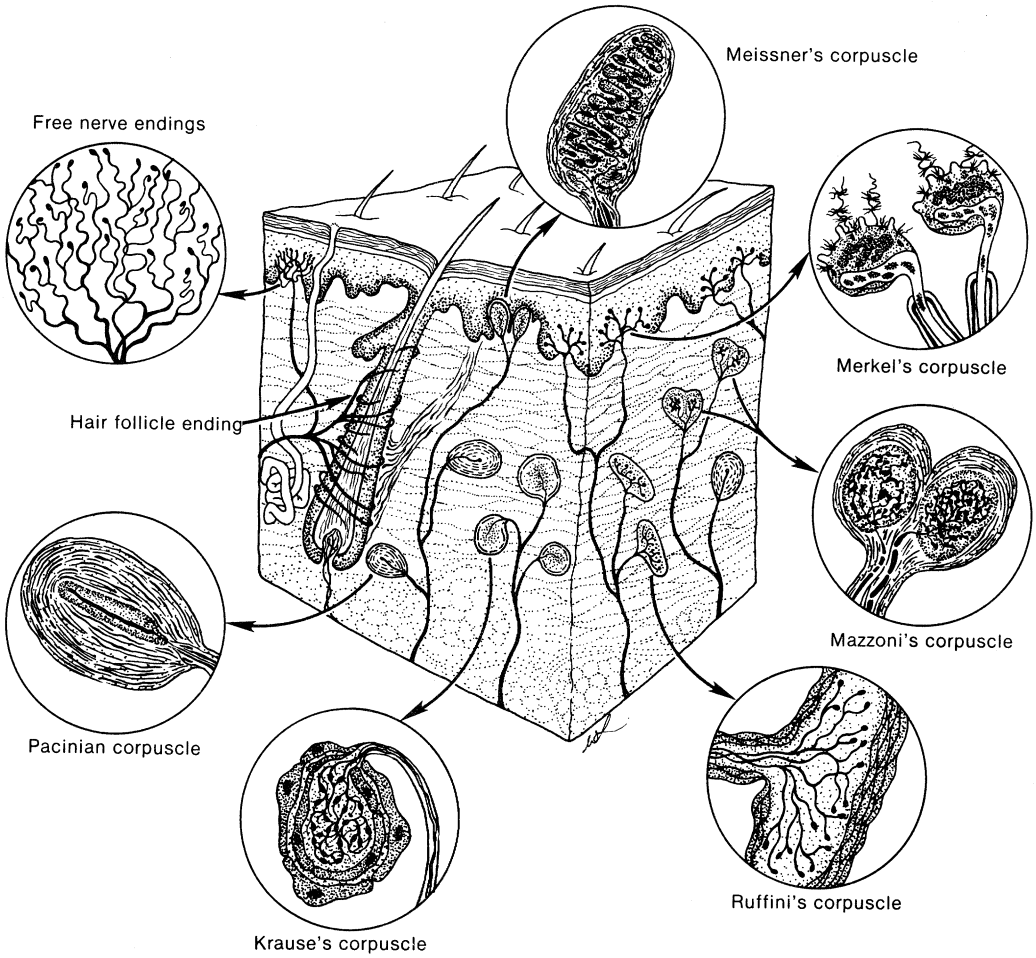


Figure 9 Skin receptors. From DeMyer, W. *Neuroanatomy*, Ch. 7, Copyright © 1988. Reprinted by permission of John Wiley & Sons, Inc.

the spindle: nuclear bag and nuclear chain fibers. The nuclear bag fibers have an expanded equatorial region because of the presence of numerous nuclei. They also extend beyond the capsule at each end before inserting into extrafusal connective tissue or a tendon. The nuclear chain fibers are smaller in diameter than the nuclear bag fibers. In these fibers, the nuclei form a single longitudinal row or chain in the center of each fiber. Finally, the neuromuscular spindles are innervated by smaller γ motor neurons.

b. Neurotendinous Spindles (Golgi Tendon Organs) Tendons are the noncontractile part of the muscles and consist of cords and sheets of collagen fibers attached to bones. Neurotendinous spindles are encapsulated nerve endings woven among the collagen fibers in tendons. The majority of Golgi tendon organs are localized near the junction of tendons and muscles. Information from these receptors is transmitted to the spinal cord by afferents classified as the Ib type.

c. Muscle Spindle–Golgi Tendon Organ Interaction The neuromuscular spindles provide sensory information about the length and the rate of change in the length of the muscle, which is very important in the control of muscle activity. At rest, the muscle spindles constantly send afferent information, and most of the time this information is not consciously perceived. During passive or active stretch of the muscle, spindles are elongated and there is an increase in the firing rate of the afferents. As soon as a muscle relaxes, it results in a decrease in the firing of the afferents to the spinal cord or brain. In order to maintain the sensitivity of the spindle, the γ motor neurons can regulate the stretch or tautness in the central part of the muscle spindle. Increased muscle tension also stimulates the Golgi

Table IV
Receptors, Associated Axons, and Adequate Stimuli

Receptor type	Axon type	Location	Sensory stimulus	Sensory modality
Nonencapsulated receptors				
Free nerve endings	Aδ and C	Skin, cornea, ligaments, joint capsule, bone, etc.	Mechanoreceptor, thermoreceptor, nociceptor	Pain, temperature
Hair follicle endings	Aβ	Hairy skin	Mechanoreceptor	Touch
Merckel's corpuscles	Aβ	Hairless skin	Mechanoreceptor	Touch
Encapsulated receptors				
Meissner's corpuscles	Aβ	Dermal papillae of skin of palm, sole of foot, nipple, and external genitalia	Mechanoreceptor	Touch
Pacinian corpuscles	Aβ	Dermis, ligaments, joint capsule, peritoneum, external genitalia, etc.	Mechanoreceptor	Vibration
Ruffini's corpuscles	Aβ	Dermis of hairy skin	Mechanoreceptor	Stretch
Muscle spindles		Skeletal muscles	Mechanoreceptor	Muscle stretch
Annulospiral sensory endings	Ia			
Flower spray sensory endings	II			
Golgi tendon organs	Ib	Tendons	Mechanoreceptor	Tendon tension

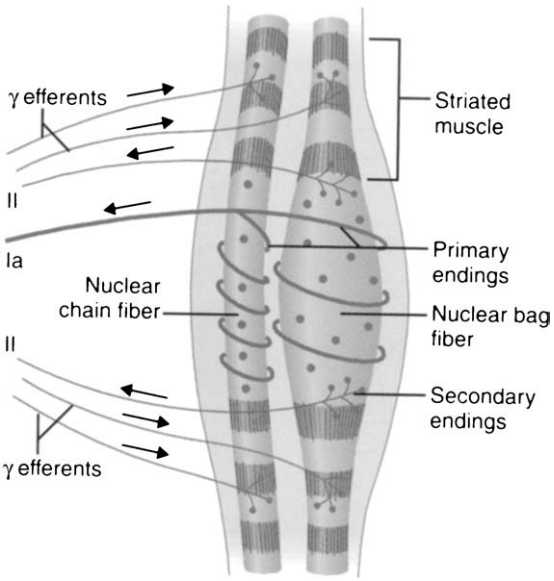


Figure 10 Simplified illustration of a muscle spindle. The intrafusal muscle fibers are nuclear chain and nuclear bag fibers. A stretch of the central region of intrafusal fibers is sensed by primary and secondary endings. The sensory information is conveyed to the central nervous system by groups Ia and II afferents. Efferent control of intrafusal fibers is via γ motor neurons. From Lundy-Ekman, L. *Neuroscience Fundamentals for Rehabilitation*, Ch. 6, p. 90, Copyright © 1998. Reprinted by permission of W.B. Saunders Co.

tendon organs and causes an increase in the firing rate of afferents. These afferent impulses have an inhibitory effect on α motor neurons, causing relaxation of the muscle. The opposing effects of the neuromuscular spindles and the Golgi tendon organs are very important in the integration of spinal reflexes. The myotatic or stretch reflex is one of clinical importance. This reflex is monosynaptic and involves only two neurons (Fig. 1). To initiate the reflex, the muscle is stretched by a sharp tap of the physician's reflex hammer on the tendon of the muscle. This stimulates the stretch receptors of the muscle, and the resultant excitation reaches the spinal cord by way of Ia afferent fibers. These fibers are axons of the sensory neurons of the dorsal root ganglia. The central branches of these neurons have excitatory synapses on spinal α motor neurons. The axons of the motor neurons synapse in the stretched muscle, thereby causing it to contract.

2. Effector Endings

a. Neuromuscular Junctions in Skeletal Muscles A single α motor neuron (located within the ventral horn of the spinal cord; see Fig. 1), with the muscle fibers

that it innervates, is defined as the motor unit. As soon as the large myelinated axon of an α motor neuron enters a skeletal muscle, it branches many times. Each small branch terminates on muscle fibers at the motor end plate or neuromuscular junction (Fig. 11). The plasma membrane of the axon terminal (axolemma) is separated by the synaptic cleft (gap of 20–50 nm) from the plasma membrane of the muscle fiber (sarcolemma). The surface area of sarcolemma (postsynaptic membrane) at a motor end plate is thrown into numerous folds. These serve to increase the contact area of muscle to the naked axon (presynaptic membrane). The neurotransmitter, acetylcholine, is released from synaptic vesicles into the synaptic cleft when a nerve impulse reaches the motor end plate. Once the acetylcholine is released, it will stimulate receptor sites on the postsynaptic muscular membrane, causing the contraction of skeletal muscle fibers. The acetylcholine remains in contact with the postsynaptic membrane for a very short period of time (about 1 msec), and it is rapidly inactivated by the enzyme acetylcholinesterase (AChE). Acetylcholinesterase hydrolyzes acetylcholine into acetic acid and choline. The choline is taken up by the presynaptic terminals for the synthesis of new molecules of acetylcholine.

VI. RESPONSE OF THE PERIPHERAL NERVOUS SYSTEM TO INJURY

Injury to peripheral nerves may cause transection of axons, with resulting degenerative changes in the nerve. Once severed from the cell body, the distal segment of the axon undergoes a process known as Wallerian degeneration. A classic work by Santiago Ramon y Cajal gave a detailed description of degenerative and regenerative changes within the peripheral nervous system after injury. "When one interrupts the continuity of a nerve by the scalpel or destructive agents, the peripheral stump degenerates rapidly, but all its constituent factors do not disappear" (Fig. 12). "Thus those parts, such as the axons and myelin, which trophically depend on the nerve cell, are destroyed and later restored." Clinically, injury to the peripheral nerve manifests in sensory loss and skeletal muscle weakness. Injury to axons of the spinal motor neurons (lower motor neurons) results in flaccid paralysis (paralysis with decreased muscle tone), decreased or absent superficial or deep reflexes, and spontaneous twitches or fasciculation within the muscles. Finally, atrophy in the denervated muscles could occur within a period of time after injury (Fig. 13). These changes occur because the lower motor neurons maintain the

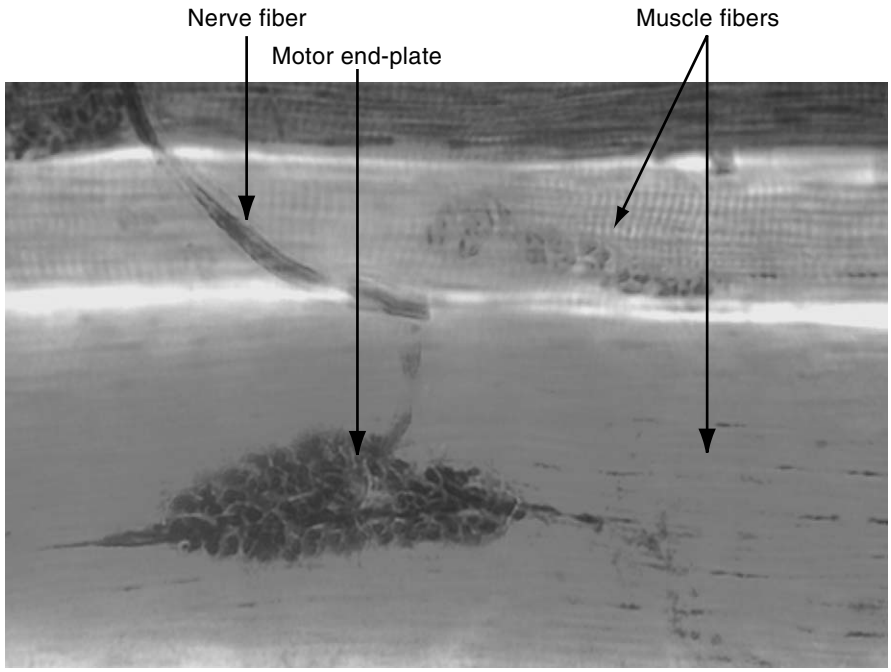


Figure 11 Motor end plate on skeletal muscle. Gold chloride method.

normal tone of the skeletal muscles and also provide trophic support to the muscle fibers.

Degeneration of the peripheral nerves is more likely to be followed by regeneration and recovery of function of the target organ. Many new axonal sprouts of motor or sensory neurons will make a passage through the longitudinal cords of Schwann cells within the distal part of the injured peripheral nerve. When an axon in the PNS is severed, the chain of Schwann cells remains, providing a pathway that guides axon regrowth. A regenerating axon within the peripheral nerve can grow approximately 1 mm per day. In contrast, in the CNS this pathway does not remain after axons are damaged. The tissue often becomes disorganized, and a physical barrier to axonal growth can develop as a result of the proliferation of astrocytes. This process, known as glial scarring, is among the other complex factors inhibiting regeneration within the CNS. Finally, oligodendrocytes and central myelin contain cell surface molecules that inhibit axon outgrowth in culture and *in vitro*. To overcome the nonpermissive environment of the CNS for regeneration, peripheral nerve implants and genetically modified Schwann cells were used in experiments to bridge the gap in the injured spinal cord.

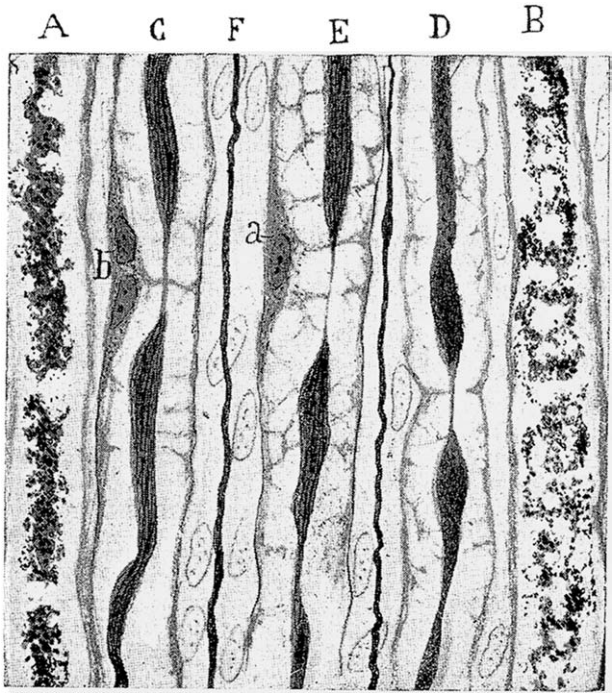


Figure 12 Piece of the peripheral stump of the sciatic nerve of a cat, killed 48 hr after the operation (nerve injury). Region far from the wound. Staining by reduced silver nitrate. (A, B) Axons that were destroyed at an early stage; (C–E) axons in the varicose phase, relatively resistant; (F) barely altered nonmedullated fiber. From *History of Neuroscince*, Corsi, P., *et al.*, Eds., 1991, Oxford University Press, New York. Reprinted by permission of Oxford University Press.

Figure 13 Posterior view of the legs of a 68-year-old male. Severe atrophy is noticeable within the right gastrocnemius muscle. Neurological examination showed the absence of an Achilles reflex on the right side, and the patient also described a decrease in sensation in response to touch and pinprick on the lateral aspect of the right leg. These clinical findings result from the compression and degeneration of the spinal nerves at the L5–S1 level by herniation of the intervertebral disk.

See Also the Following Articles

AXON • BRAIN ANATOMY AND NETWORKS • NERVOUS SYSTEM, ORGANIZATION OF • NEURON • SYNAPSES AND SYNAPTIC TRANSMISSION AND INTEGRATION

Suggested Reading

Bandtlow, C., Zachleder, T., and Schwab, M. E. (1990). Oligodendrocytes arrest neurite growth by contact inhibition. *J. Neurosci.* **10,** 3837–3848.

Cheng, H., Cao, Y., and Olson, L. (1996). Spinal cord repair in adult paraplegic rat: Partial restoration of hind limb function. *Science* **273,** 510–512.

DeFelipe, J., and Jones, E. G. (Eds.). (1991). *History of Neuroscience. Cajal's Degeneration and Regeneration of the Nervous System*. Oxford University Press, New York.

DeMyer, W. (1988). *Neuroanatomy. The National Medical Series for Independent Study*. Wiley & Sons, New York.

Krassioukov, A. V., and Weaver, L. C. (1996). Morphological changes in sympathetic preganglionic neurons after cord injury in rats. *Neuroscience* **70,** 211–226.

Lundy-Ekman, L. (1998). *Neurosceince Fundamentals for Rehabilitation*. W. B. Saunders Company, Philadelphia.

Oudega, M., Gautier, S. E., Chapon, P., Fragoso, M., Bates, M. L., Parel, J. M., and Bunge, M. B. (2001). Axonal regeneration into Schwann cell grafts within resorbable poly(α-hydroxyacid) guidance channels in the adult rat spinal cord. *Biomaterials* **22**(10), 1125–1134.

Phantom Limb Pain

HERTA FLOR
University of Heidelberg

GLOSSARY

cortical reorganization Changes in the functional architecture of the sensory and motor cortex related to injury or stimulation and learning.

nonpainful phantom sensation and nonpainful sensation in the residual limb Sensations in the phantom or the residual portion of the limb that are clearly not painful, such as tingling or itching or a sense of the position and size of the phantom.

preemptive analgesia Analgesia that is applied before the onset of acute pain (e.g., before and during surgery).

referred sensation Sensation in the phantom elicited by stimulation of adjacent but also distal body parts.

residual limb pain Pain in the part of the body adjacent to the amputation site; also referred to as stump pain.

telescoping Retraction of the phantom limb toward the residual limb; sometimes accompanied by shrinking of the limb.

Phantom limb pain is pain in a body part that is no longer present. It is a frequent aftereffect of amputation, occurring in 50–80% of all amputees. Phantom limb pain is commonly classified as neuropathic pain and is assumed to be related to damage of central or peripheral neurons. The French physician Ambroise Paré, who postulated that peripheral factors as well as a central pain memory might be causing the phenomenon, first described it in 1552.

I. EPIDEMIOLOGY, DESCRIPTION, AND COURSE

Although some studies have cited lower rates of prevalence, it is likely that they did not assess the true occurrence of the phenomenon since patients are often reluctant to admit pain in a part of the body that is no longer present. Although phantom limb pain is more common after the amputation of an arm or leg, it may also occur after the surgical removal of other body parts, such as a breast, the rectum, the penis, the testicles, an eye, the tongue, or the teeth. It has also been described subsequent to lesions of peripheral nerves or the central nervous system, such as in brachial plexus avulsion or paraplegia.

Phantom limb pain seems to be more common if the amputee suffered from chronic pain prior to the amputation, may be less frequent if the amputation occurred at a very young age, and is virtually absent in congenital amputees (i.e., those born without a limb). It is usually episodic with brief attacks of pain but may also be continuous and seems to be more intense in the distal portions of the phantom. The pain may have a number of different qualities, such as stabbing, throbbing, burning, cramping, and it may also consist of a painful position or painful movements of the phantom. The long-term course of phantom limb pain is unclear. Whereas some authors reported a slight decline in prevalence over the course of several years, others have not described such changes. Frequently, the pain in the phantom is similar to the pain that existed in the limb prior to amputation. There have been some reports that phantom limb pain is more frequent in male than in female amputees; however, other studies did not confirm these findings. It is also not clear to what extent age of the amputee and

medical status show a relationship to the severity and/or incidence of phantom limb pain.

Phantom limb pain needs to be differentiated from pain in the residual limb, also referred to as stump pain (i.e., pain in the portion of the body adjacent to the amputated or deafferented body part). Patients and physicians alike often have difficulty separating the two types of pain that are only moderately positively correlated. In addition to phantom limb pain, nonpainful sensations may be present in the phantom and/or the residual limb. They may consist of spontaneous movements of the phantom or the residual limb, tingling, cramping, or the sensation of the shape and volume of the limb. Phantom sensations seem to be an invariable consequence of traumatic amputation of a limb, whereas they are very rare in congenital amputees.

Many patients report the phenomenon of telescoping, i.e., the retraction of the phantom toward the residual limb and often the disappearance of the phantom into the limb. It was long assumed that telescoping might be an adaptive phenomenon and be negatively correlated with phantom limb pain and as such an expression of changes in the central nervous system that were beneficial. However, recent evidence suggests that central changes, phantom limb pain, and telescoping are positively correlated, i.e., that telescoping is associated with more phantom limb pain. Patients with phantom limb pain also often report sensations that are referred to the phantom when skin areas adjacent, but also far removed, from the amputated limb are stimulated. They may have a point-to-point correspondence between stimulation sites in other body parts and the phantom, and the modality and quality of the sensations are usually the same. However, in most cases (80–95% of amputees) referred sensations lack this topographic correspondence, which seems to be present in only a small percentage of amputees. Ramachandran suggested that these referred sensations might be a perceptual correlate of cortical changes occurring in the primary somatosensory cortex in amputees. Earlier studies also reported enhanced perceptual acuity of the residual limb as assessed by two-point discrimination and lower perception and pain thresholds compared to those in the contralateral limb in amputees. These changes in sensory perception were also thought to be positively related to the phenomenon of telescoping. Recent studies, however, could not confirm this relationship and have questioned the assumption that stump perceptual acuity might be a correlate of central changes related to the experience of a phantom.

II. PATHOPHYSIOLOGY

Both peripheral and central factors have been discussed as determinants of phantom limb pain. Psychological factors do not seem to contribute to the etiology of the problem but, rather, may affect the course and the severity of the pain. The general view today is that of multiple changes along the neuraxis contributing to the experience of phantom limb pain.

A. Peripheral Factors

Peripheral changes such as nociceptive input from the residual limb have been viewed as an important determinant of phantom limb pain. This is supported by the moderately high correlation (0.40–0.70) between residual limb and phantom limb pain. Ectopic discharge from a stump neuroma has been postulated as one important peripheral mechanism. When peripheral nerves are cut or injured, regenerative sprouting of the injured axon occurs. In this process a neuroma in the residual limb may be formed (i.e., enlarged and disorganized endings of C fibers and demyelinated A fibers that show an increased rate of spontaneous activity). Mechanical and chemical stimulation further increases the rate of discharge, which seems to be mainly related to ectopia (i.e., neuronal discharge that is generated along the axon or in the soma). These ectopic discharges have been related to stimulation of the stump (e.g., by pressure or cold) but can also occur spontaneously as a consequence of nerve injury. In addition, nonfunctional connections between axons (ephapses) may contribute to this spontaneous activity. However, phantom limb pain is often present immediately after amputation, before a neuroma could have formed. Moreover, local anesthesia of the stump does not eliminate phantom limb pain in all amputees.

A further site of ectopic discharge may be the dorsal root ganglion (DRG). This ectopia in the DRG can amplify discharge coming from the residual limb or can lead to cross-excitation and instigate the depolarization of neighboring neurons (Fig. 1). In humans it was found that an anesthetic block of a neuroma eliminates nerve activity related to the stimulation of the stump but not spontaneous activity, which may originate in the DRG. Sympathetic discharge that may also be caused by emotional distress may lead to increased levels of circulating epinephrine that can trigger and exacerbate neuronal activity from neuroma. If a phantom pain eliciting neuroma develops,

Figure 1 This figure shows how hypertonic saline excites ectopic impulses in dorsal root ganglia associated with an injured nerve. (A) Spontaneous activity in a dorsal root ganglion in a rat whose sciatic nerve was subjected to constriction injury. (B) Topical application of 6% saline accelerated the firing and recruited previously silent afferents. All active units were C fibers [reproduced with permission from Devor, M. (1997). Phantom pain as an expression of referred and neuropathic pain. In *Phantom Pain* (R. A. Sherman, Ed.), p. 43. Plenum, New York].

depends on the type of the amputation, exacerbating stimulation, and genetic predisposition to develop neuropathic pain problems.

The role of the sympathetic nervous system in animal models of neuropathic pain such as constriction injury or autotomy has been well documented. However, research on its role in phantom limb pain is scarce. The mechanisms that have been discussed are sympathetically triggered ephaptic transmission, sympathetic activation of nociceptors, or activation of low-threshold mechanoreceptors that trigger sensitized spinal cord neurons. In addition to sympathetic sensory coupling at the level of the periphery (neuroma), sympathetic sensory coupling also occurs at the level of the DRG.

The sympathetic maintenance of some types of phantom limb pain has been supported by evidence that for beta-adrenergic blocking agents, temporary or surgical blockade of the sympathetic activation can reduce phantom limb pain. Injections of epinephrine resulted in an increase of phantom limb pain and paresthesias in some amputees. Katz suggested that paresthesias might be related to postganglionic sympathetic activity acting on primary afferents and that painful sensations might develop when nociceptors are involved and/or central sensitization has occurred. He assumed that the release of acetylcholine and noradrenaline from sudomotor and vasomotor efferents might contribute to the psychophysiological changes observed in amputation stumps and might lead to enhanced discharge from the primary afferent neurons in this region. Stump neuroma may further intensify this process. Katz also pointed out that cognitive and affective processes can trigger phantom limb phenomena and phantom limb pain (Fig. 2) via a hypothalamic–sympathetic route.

Although sympathetically maintained pain does not necessarily covary with regional sympathetic abnormalities, in some patients sympathetic dysregulation in the residual limb is apparent. Reduced near surface blood flow to a limb has been implicated as a predictive physiological correlate of phantom limb pain. A number of studies have demonstrated that (i) the ends of the nerves that used to serve the amputated limb are still sensitive to stimuli; (ii) cooling the nerve ends, especially those of C fiber nociceptors, causes increased firing rates; and (iii) reducing blood flow to the extremity results in the cooling of it. Measurements of skin temperature in amputees by Sherman and collaborators revealed that the residual limbs of amputees with phantom limb pain were cooler at the distal end than paired points on the opposite extremity and that the cooler areas did not warm when attempts were made to increase cutaneous blood flow. Amputees were found to be generally more sensitive to cooling of the affected limb. Most amputees with phantom limb pain showed different patterns of temperature than those without phantom limb pain. Consistent, inverse relationships between intensity of phantom limb pain and temperature in the residual limb relative to that of the intact limb have been demonstrated for burning, throbbing, and tingling descriptions of phantom pain but not for other descriptions. It has also been established that (a) for these descriptors of phantom pain there is a day to day relationship between the relative amount of blood flow in the stump and pain intensity and that (b) there is an immediate change in pain when blood flow changes. This tight relationship has been replicated numerous times and indicates that there is more than a casual relationship between the two. The existence of a vascular-related mechanism for burning phantom pain is also supported by the short-term effectiveness of invasive procedures such as sympathetic blocks and sympathectomies, which increase blood flow to the limb and reduce the intensity of burning phantom and stump pain but not other descriptors. It is indirectly supported by the virtual ineffectiveness of surgical procedures involving severing nerves either in the spinal cord or running between the amputation site and the spinal cord. Beta blockers such as propranolol cause dilation of peripheral blood vessels and have been reported to successfully ameliorate phantom pain at least in the short term. Relationships between muscle tension and burning phantom

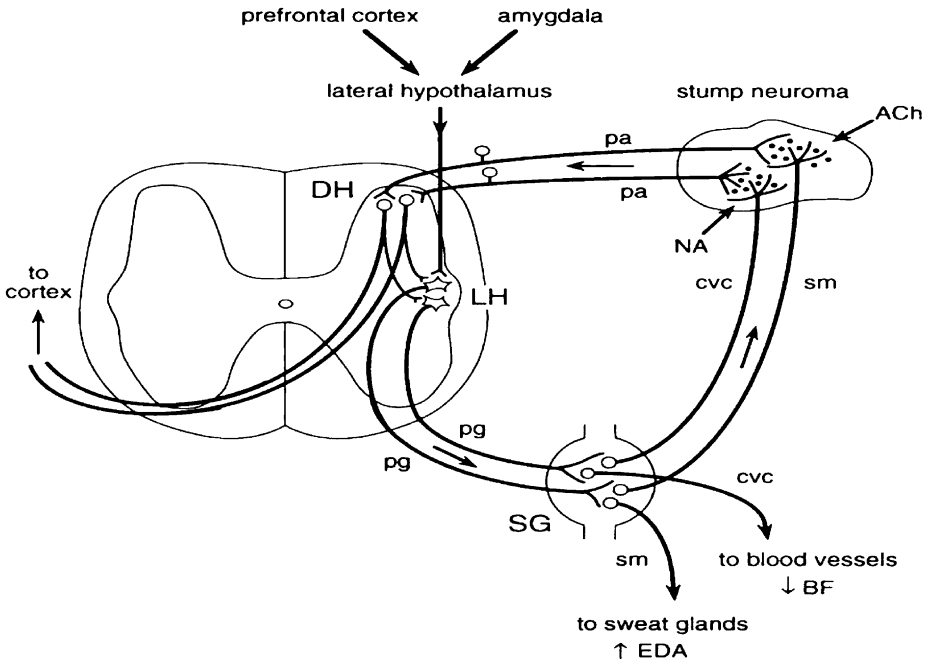


Figure 2 Schematic diagram illustrating a mechanism of sympathetically maintained phantom limb parasthesias. pa; pg, preganglionic sympathetic neurons; DH, dorsal horn; LH, lateral horn; SG, sympathetic ganglion; sm, sudomotor; cvc, cutaneous vasoconstrictor; NA, noradrenaline; Ach, acetylcholine [reproduced with permission from Katz, J. (1997). The role of the sympathetic nervous system in phantom pain. In *Phantom Pain* (R. A. Sherman, Ed.), p. 86. Plenum, New York].

limb pain have been shown to be largely due to the change in near surface blood flow that accompanies increased muscle tension.

Onset and intensity of cramping and squeezing descriptions of phantom pain have been related to muscle tension in the residual limb. A variety of studies have demonstrated that the amount of muscle tension in the residual limb changes before alterations in intensity of cramping phantom pain both from day to day and from moment to moment. Changes in surface electromyographic representations of muscle tension in the residual limb precede changes in cramping and squeezing phantom pain by up to several seconds. This relationship does not seem to hold for any other descriptions of phantom pain. The relationship between cramping phantom pain and muscle tension in the residual limb is supported by the success of treatments resulting in reduction of residual limb muscle tension for cramping phantom pain. Numerous amputees report that cramping phantom pain decreases with any activity that tends to decrease muscle contraction levels in the residual limb and increases with activities that increase overall levels of contraction. Thus, activities such as phantom exercises, which result in changes in muscle tension in the residual limb, can result in temporary changes in intensity of phantom pain.

As noted previously, peripheral factors alone cannot be the primary factor in the occurrence of phantom limb pain. Pain is present even if there is no pathology in the residual limb; it often has an onset immediately after the amputation, and anesthetic blocks do not uniformly eliminate phantom limb pain or they eliminate it for a period of time that clearly exceeds the time the block can be active. This suggests that peripheral factors may be of varying importance in the etiology and modulation of phantom limb pain and that central factors must also play a role.

B. Central Factors

Additional contributions to phantom limbs may come from sensitization of the dorsal horn of the spinal cord. Increased activity of peripheral nociceptors leads to an enduring change in the synaptic structure of the dorsal horn in the spinal cord, a process called *central sensitization*. Central sensitization is characterized by increased excitability of the dorsal horn neurons, the reduction of inhibitory processes, and structural

changes at the central nerve endings of the primary sensory neurons, the interneurons, and the projection neurons. This central sensitization is mediated by the NMDA receptor and its transmitter glutamate. Low threshold afferents may become functionally connected to ascending spinal projection neurons that carry nociceptive information and/or inhibitory interneurons may be destroyed by rapid discharge from injured tissue leading to a hyperexcitable spinal cord. In addition, afferents of the residual limb may invade the regions where the deafferented limb was previously represented. This process may be due to an unmasking of previously silent connections as well as a sprouting of new connections. It has also been proposed that the loss of input related to deafferentation may lead to a general disinhibition of the spinal cord.

Supraspinal changes related to phantom limb pain involve the brain stem, the thalamus, and the cortex. Melzack suggests that there is a neuromatrix (i.e., a network of neurons in several brain areas including the thalamus and somatosensory cortex, the reticular formation and the limbic system, and the posterior parietal cortex) that is the anatomical substrate of the self. The output from this system forms a neurosignature, which is specific for an individual and provides information about the body and its sensations. This neuromatrix is thought to be genetically determined but also modified by experience. Melzack suggests that amputation creates abnormal input into the neuromatrix that stems from either a lack of normal sensory activity or overactivity related to the abnormal firing pattern of damaged nerves. This then leads to an altered neurosignature and the experience of a phantom. He assumes that the reports of phantom phenomena in persons with congenital absence of a limb provide evidence of the genetic determination of the neuromatrix. However, the literature on the presence of phantoms in congenital amputees is controversial, potentially due to problems in the assessment of phantom limb pain. Moreover, the neuromatrix theory is difficult to test because it involves a wide range of brain areas and is not very specific. It also fails to explain why some amputees develop phantom limb pain and why others remain pain free. On the other hand, the brain areas cited as part of the neuromatrix are definitely importantly involved in the experience of phantom limbs and phantom limb pain.

New insights into phantom limb pain have come from studies that showed changes in the functional and structural architecture of primary somatosensory cortex subsequent to amputation and deafferentation in adult animals. It was previously thought that changes in the sensory and motor maps occur only during a limited time window in the development of the organism. In these studies the amputation of digits in an adult owl monkey led to an invasion of adjacent areas into the representation zone of the deafferented fingers. Whereas this type of reorganizational change spanned a distance of several millimeters, recordings from the somatosensory cortex of monkeys that had undergone dorsal rhizotomies 12 years earlier revealed reorganizational changes (invasion of the mouth and chin area into the deafferented arm and hand area) on a scale of several centimeters. Subsequently, Ramachandran *et al.* suggested that the cortical changes observed in these animal studies might be related to phantom phenomena since they noted a point-to-point correspondence between stimulation sites on the face and phantom sensations in upper extremity amputees. Several imaging studies have reported that amputees actually show such reorganizational changes, although they are less related to the perceptual changes observed by Ramachandran *et al.* but rather have a close association with phantom limb pain. It was shown in several studies of human amputees that reorganization in area 3b of primary somatosensory cortex is highly correlated with phantom limb pain (Fig. 3). This result seems to be a very stable finding.

The functional significance of these cortical changes was established by showing that the elimination of peripheral input from the amputation stump by brachial plexus anesthesia completely eliminated cortical reorganization and phantom limb pain in about 50% of the amputees that were studied. In the remaining 50%, both cortical reorganization and phantom limb pain remained unchanged. This result suggests that in some amputees, cortical reorganization and phantom limb pain may be maintained by peripheral input, whereas in others central, potentially intracortical, changes might be more important. It was recently shown that axonal sprouting in the cortex underlies the reorganizational changes observed in amputated monkeys, whereas thalamic reorganization occurs after lesions close to the dorsal horn that is then relayed to the cortex. Similar changes have also been observed in the motor cortex in both animals and human amputees. Computational models of deafferentation and related phenomena have suggested that reorganization of neuronal networks is enhanced when abnormal noise-like input as it might originate from a neuroma is created and fed into the system.

Thalamic stimulation and recordings in human amputees have revealed that reorganizational changes may also occur at the thalamic level and are closely

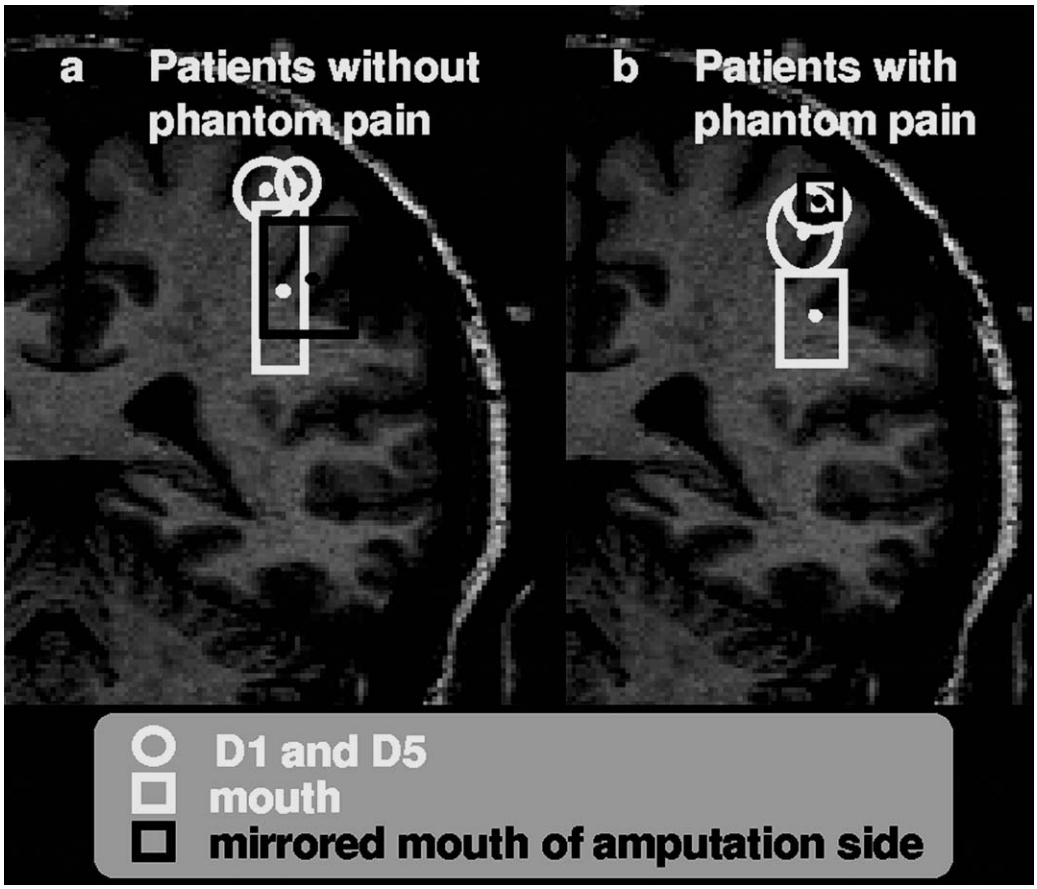


Figure 3 Reorganization in primary somatosensory cortex in unilateral upper limb amputees without phantom limb pain (a) and with phantom limb pain (b). The dots denote the means and the rectangles and circles the standard deviations of the respective localizations. Note that only in amputees with phantom limb pain has a shift of the mouth representation into the hand representation occurred, whereas the amputees without pain do not display a similar shift. Also note the very small standard deviation of the shifted mouth representation in the patients with phantom limb pain suggesting that all persons with pain show reorganization.

related to the perception of phantom limbs and phantom limb pain. Studies in animals have shown that these changes may be relayed from the spinal and brain stem level; however, changes on the subcortical levels may also originate in the cortex, which has strong efferent fiber connections.

A comprehensive model of the development of phantom limbs includes both peripheral and central factors and assumes that pain memories established prior to the amputation may be powerful elicitors of phantom limb pain. This assumption is based on findings that amputees often experience phantom limb pain that is similar in both quality and location to pain experienced before the amputation. Melzack and Katz called these memory traces somatosensory pain memories. In patients with chronic back pain it was shown that increasing chronicity is correlated with an increase of the representation zone of the back in primary somatosensory cortex. These data suggest that long-lasting noxious input may lead to long-term changes at all levels of the neuraxis, including the cortical level. It has long been known that the primary somatosensory cortex is involved in the processing of pain and that it may be important for the sensory-discriminative aspects of the pain experience. There have also been reports that phantom limb pain was abolished after the surgical removal of portions of the primary somatosensory cortex and that stimulation of somatosensory cortex evoked phantom limb pain. If a somatosensory pain memory has been established with an important neural correlate in primary somatosensory cortex, subsequent deafferentation and an invasion of the amputation zone by neighboring input may activate preferentially cortical neurons coding for pain. Since

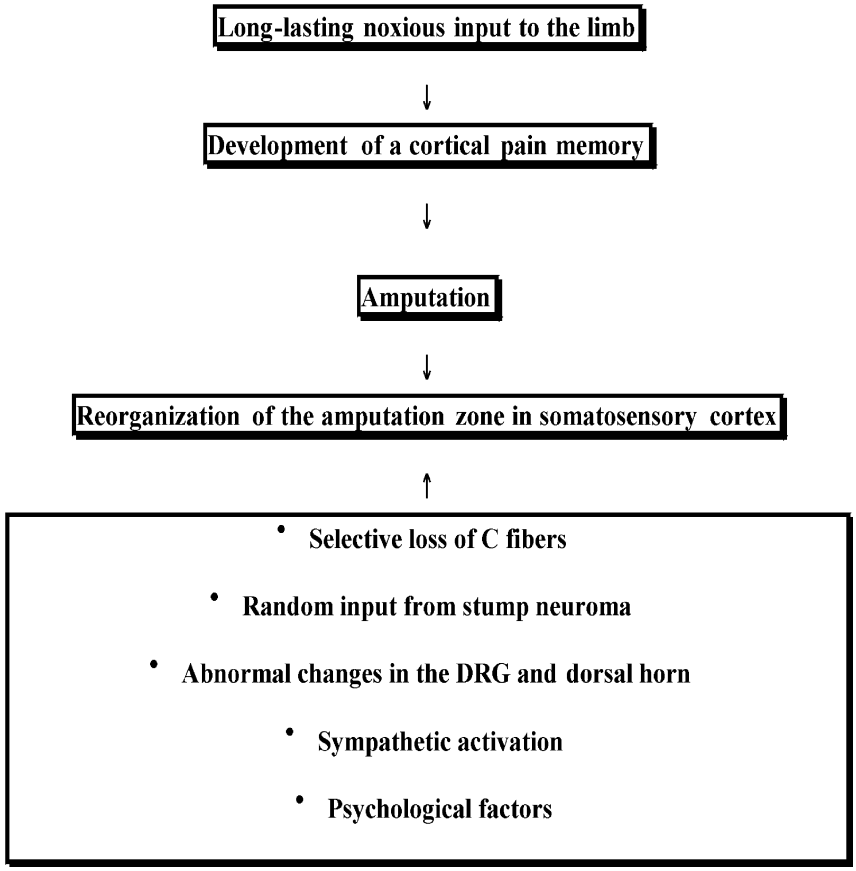


Figure 4 Schematic diagram incorporating the main factors thought to be relevant for the development of phantom limb pain.

the cortical area coding input from the periphery seems to remain assigned to the original zone of input, the activation in the cortical zone representing the amputated limb is referred to this limb and the activation is interpreted as phantom sensation and phantom limb pain. Figure 4 summarizes these change assumed to take place in patients with phantom limb pain.

It is likely that these factors are of varying importance in different patients with phantom limb pain and that subgroups of patients with distinct and differentiable pathologies exist, as suggested by Sherman.

C. Psychopathology

For a long time, it was assumed that phantom limbs and phantom limb pain might be related to unresolved grief over the loss of the limb and these were viewed as a psychosomatic manifestation of a premorbid personality. The idea that phantom limb pain is "just in the head" of the patient is still prevalent and may contribute to the large divergence in reported incidence and prevalence rates. Patients do not readily volunteer information on their phantom phenomena out of fear of being placed in a psychiatric diagnostic category. Empirical studies on psychological characteristics of patients who suffer from phantom limb pain and controls show very clearly that there are no pathological psychological processes in these patients. Phantom limb pain, however, is triggered and exacerbated by psychological factors. Longitudinal diary studies of amputees showed that there is a significant relationship between stress and the onset and exacerbation of episodes of phantom limb pain, probably mediated by sympathetic nervous system activity and increases in muscle tension as outlined previously. Cognitive factors also play a role in the modulation of phantom limb pain: Patients who lack coping strategies when confronted with episodes of pain are more affected by the pain and report more interference than patients who cope well with their problem. Likewise, higher levels of catastrophizing with respect to pain are

associated with higher levels of pain and distress. However, compared to other groups of chronic pain patients, such as patients who suffer from fibromyalgia syndrome or chronic low back pain, phantom limb pain patients do have normal psychological profiles and very little psychopathology. Psychological variables before the amputation have also been found to be predictive of phantom limb pain. Patients who were depressed and experienced little control over their lives tended to complain significantly more often of phantom limb pain 1 year after the amputation.

III. ASSESSMENT

As noted previously, the reliable and valid assessment of phantom phenomena and phantom limb pain is of great importance but has not been evaluated in most empirical studies and even less in clinical practice. Patients do have considerable difficulty in differentiating painful and nonpainful phenomena as well as differentiating to what extent sensations are confined to the residual limb, extend to the phantom, or are related to both. Recent studies have shown that the most reliable and valid way to assess phantom phenomena is by interview, where the interviewer very carefully explains the differences between the various phantom phenomena using visual material. In addition to an interview that should include the quantitative and qualitative assessment of painful and nonpainful phantom and residual limb phenomena, phantom limb pain should be assessed by standardized questionnaires. These instruments also show high stability in retests done weeks or months apart, suggesting that phantom limb pain is less variable than generally believed, at least in chronic amputees. Adjective checklists as included in the McGill Pain Questionnaire are useful since Sherman has shown that different types of phantom limb pain may be related to different underlying mechanisms and require different therapeutic interventions.

Sensory testing should also focus on two-point discrimination in the affected limb and a homologous site, the assessment of perception and pain thresholds, and the occurrence of referred sensation using both painful and nonpainful touch as described, for example, by Ramachandran. Since the use of a prosthesis seems to influence the occurrence of both phantom limb pain and cortical reorganization, details about the use of a prosthesis as well as the amount of use of the intact arm should be assessed. Both questionnaires and interview formats are available for this purpose.

IV. TREATMENT

Several studies, including large surveys of amputees, have shown that most treatments for phantom limb pain are ineffective unless they are related to the mechanisms underlying the production of the pain. Most studies are short-term assessments of small samples of phantom limb pain patients. The maximum benefit reported from a host of treatments, such as local anesthesia, sympathectomy, dorsal root entry zone lesions, cordotomy and rhizotomy, neurostimulation methods, and pharmacological interventions such as anticonvulsants, barbiturates, antidepressiants, neuroleptics, muscle relaxants, and opioids (Table I), seems to be approximately 30%. This does not exceed the placebo effect reported in other studies.

Mechanism-based treatments are very rare but have been shown to be relatively effective in a few small studies. For example, pharmacological and behavioral treatments resulting in vasodilatation of the residual limb are helpful for burning/tingling descriptions of phantom limb pain but not others, even when an individual patient reports several descriptive types. Treatments resulting in decreased muscle tension in the residual limb are helpful for cramping but not other descriptions of phantom pain. There has been no behaviorally oriented treatment identified for stabbing and shooting pains. Behavioral interventions for several descriptive types of phantom limb pain have been in use at least since the late 1970s. Small trials have shown that patients can successfully control burning phantom limb pain to the extent that they learn to control blood flow in the residual limb. Similarly, small trials have shown that amputees with cramping phantom pain can control it to the extent that they learn to control muscle tension in their residual limbs.

Based on findings from neuroelectric and neuromagnetic source imaging, changes in cortical reorganization might influence phantom limb pain. Animal work on stimulation-induced plasticity suggests that extensive behaviorally relevant (but not passive) stimulation of a body part leads to an expansion of its representation zone. Thus, the use of a myoelectric prosthesis might be one method to influence phantom limb pain. It was recently shown that intensive use of a myoelectric prosthesis was positively correlated with both reduced phantom limb pain and reduced cortical reorganization. When cortical reorganization was partialled out, the relationship between prosthesis use and reduced phantom limb pain was no longer significant, suggesting that cortical reorganization

Table I
Commonly Employed Treatments for Phantom Limb Pain

Pharmacological	Surgical	Anesthesiological	Psychological	Other
Conventional analgesics	Stump revision	Nerve blocks	EMG biofeedback	TENS
Opioids	Neurectomy	Epidural blockade	Temperature biofeedback	Acupuncture
Calcitonin	Sympathectomy	Sympathetic block	Cognitive–behavioral pain management	Physical therapy
Beta-blockers	Rhizotomy	Local anesthesia	Sensory discrimination training	Ultrasound
Neuroleptics	Cordotomy		Hypnosis	Manipulation
Anticonvulsives	Tractotomy			Prosthesis training
NMDA antagonists	Dorsal column stimulation			
Antidepressants	Deep brain stimulation			
Barbiturates				
Muscle relaxants				

mediates this relationship. An alternative approach in patients in whom prosthesis use is not viable is the application of behaviorally relevant stimulation. A 2-week training that consisted of discrimination training of electric stimuli to the stump for 2 hr per day led to significant improvements in phantom limb pain and a significant reversal of cortical reorganization. A control group of patients who received standard medical treatment and general psychological counseling in this time period did not show similar changes in cortical reorganization and phantom limb pain. A similar approach used asynchronous stimulation of the mouth and stump region and also achieved alterations in cortical reorganization and phantom limb pain specific for the type of pain. In both cases, the basic idea of the treatment was to provide input into the amputation zone and thus undo the reorganizational changes that occurred subsequent to the amputation. Ramachandran, who used a virtual reality box to train patients to move the phantom and reduce phantom limb pain, described another behaviorally oriented approach. A mirror was placed in a box and the patient inserted both his or her intact arm and the phantom. The patient was then asked to look at the mirror image of the intact arm, which is perceived as an intact hand where the phantom used to be. The patient was then asked to make symmetric movements with both hands, thus suggesting real movement from the lost arm to the brain. This procedure seems to reestablish control over the phantom and to reduce phantom limb pain in some patients.

Although these stimulation-based approaches might be beneficial in many patients who suffer from phantom limb pain, they might be insufficient for patients in whom peripheral factors contribute significantly to the problem. Here, biofeedback might be used as an alternative or adjunct treatment. Sherman reported substantial changes in phantom limb pain related to peripheral biofeedback in those patients in whom peripheral factors were found to be of relevance for the experience of phantom limbs.

Pharmacological interventions might also be useful in the amelioration of phantom limb pain related to central changes. As discussed previously, spinal sensitization as well as cortical and thalamic reorganization might be involved in the experience of phantom limb pain. Animal studies have shown that both spinal sensitization and cortical reorganization can be prevented or reversed by use of NMDA receptor antagonists. Reorganization was also found to be related to reduced GABAergic activity and increased cholinergic activity. Thus, these substances should be beneficial in the treatment of phantom limb pain. However, studies in patients are scarce and the data are controversial. For example, although some studies have reported a positive effect of the NMDA receptor antagonist ketamine on phantom limb pain, others have found no effects of the NMDA receptor antagonist memantine on chronic phantom pain phenomena.

V. PREVENTION

As noted previously, Katz and Melzack emphasized that there are somatosensory pain memories that may

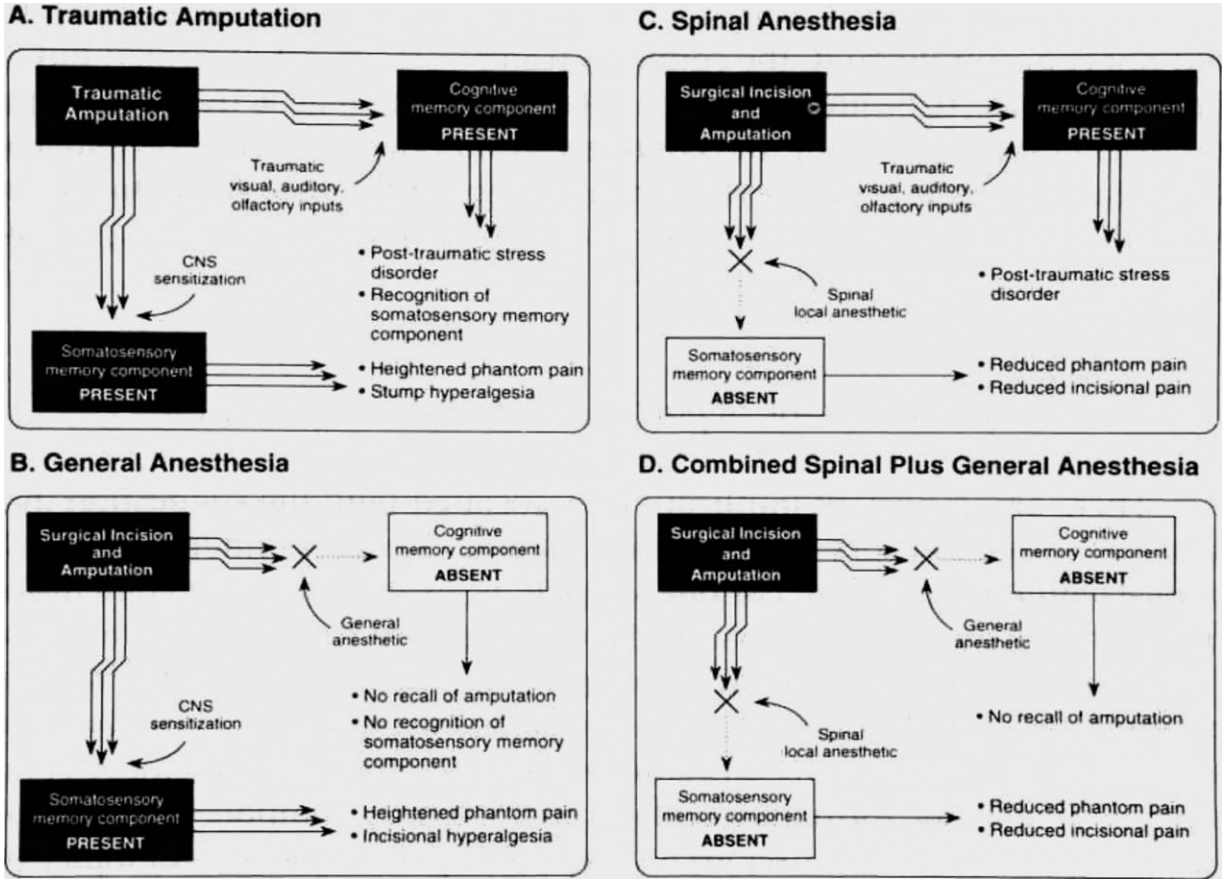


Figure 5 Predicted pain and psychological postoperative status following traumatic amputation as described by Katz [reproduced with permission from Katz, J. (1997). Central nervous system correlates and mechanisms of phantom pain. In *Phantom Pain* (R. A. Sherman, Ed.), p. 98. Plenum, New York].

be revived after an amputation and lead to phantom limb pain. They also noted that implicit and explicit memory components can be differentiated, both of which contribute to the experience of phantom limbs and phantom limb pain. Katz therefore suggested that both memory components (i.e., both general and spinal anesthesia) need to be targeted in preemptive analgesic trials destined to prevent the onset of phantom limb pain (Fig. 5).

Preemptive analgesia refers to the attempt to prevent chronic pain by early intervention before acute pain occurs (e.g., before and during surgery). Based on the data on sensitization of spinal neurons by afferent barrage, it has been suggested that general anesthesia should be complemented by peripheral anesthesia, thus preventing peripheral nociceptive input from reaching the spinal cord and higher centers. However, preemptive analgesia that included both general and spinal anesthesia has not consistently been efficacious in preventing the onset of phantom limb pain. Whereas several studies reported a reduction of the incidence of phantom limb pain when additional epidural anesthesia was used in the pre- and perioperative stage, recent studies failed to find a beneficial effect. A preexisting pain memory that has already led to cortical changes would not necessarily be affected by elimination of afferent barrage. Here, NMDA antagonists as well as GABA agonists—drugs that have been shown to reverse or prevent cortical reorganization—might be beneficial for prevention of phantom limb pain. A recent study used the NMDA receptor antagonist memantine versus placebo in addition to brachial plexus anesthesia in patients undergoing traumatic amputations of individual fingers or a hand. It was found that memantine significantly reduced the incidence of phantom limb pain 1 year after the surgery, whereas placebo failed to show a similar effect.

VI. FUTURE DEVELOPMENTS

Both peripheral and central factors and their interaction need to be examined more closely in animal models of amputation-related pain and in human amputees. To develop more powerful interventions for phantom limb pain, controlled treatment outcome, prospective, and double-blind placebo-controlled outcome research is necessary. Only then will effective evidence-based interventions become available.

See Also the Following Articles

HALLUCINATIONS • PAIN • PAIN AND PSYCHOPATHOLOGY

Suggested Reading

Cronholm, B. (1951). Phantom limbs in amputees. A study of changes in the integration of centripetal impulses with special reference to referred sensations. *Acta Psychiatr. Neuol. Scand.* **72**(Suppl.), 1–310.

Flor, H., Elbert, T., Knecht, S., Pantev, C., Birbaumer, N., Larbig, W., and Taub, E. (1995). Phantom limb pain as a perceptual correlate of cortical reorganization following arm amputation. *Nature* **375,** 482–484.

Flor, H., Denke, C., Schaefer, M., and Grüsser, M. (2001). Sensory discrimination training alters both cortical reorganization and phantom limb pain. *Lancet* **357,** 1763–1764.

Florence, S. L., Taub, H. B., and Kaas, J. H. (1998). Large-scale sprouting of cortical connections after peripheral injury in adult macaque monkeys. *Science* **452,** 388–392.

Hill, A. (1999). Phantom limb pain: A review of the literature on attributes and potential mechanisms. *J. Pain Symptom. Management* **17,** 125–142.

Jensen, T. S., and Nikolajsen, L. (1999). Phantom pain and other phenomena after amputation. In *Textbook of Pain* (P. D. Wall and R. A. Melzack, Eds.), 4th ed., pp. 799–814. Churchill Livingston, Edinburgh, UK.

Katz, J. (1992). Psychophysiological contributions to phantom limbs. *Can. J. Psychiatry.* **37,** 282–298.

Melzack, R. A. (1990). Phantom limbs and the concept of a neuromatrix. *Trends Neurosci.* **13,** 88–92.

Nikolajsen, L., Ilkjaer, S., Christensen, J. H., Kroner, K., and Jensen, T. S. (1997). Randomised trials of epidural bupivacaine and morphine in prevention of stump and phantom pain in lower-limb amputation. *Lancet* **350,** 1353–1357.

Ramachandran, V. S., and Hirstein, W. (1999). The perception of phantom limbs. The D.O. Hebb Lecture. *Brain* **121,** 1603–1630.

Sherman, R. A. (1997). *Phantom Pain*. Plenum, New York.

Phineas Gage

MALCOLM MACMILLAN
Deakin University, Australia

GLOSSARY

antiphlogistic regimen Nineteenth century, pre-germ theory treatment based on the reduction of the inflammatory or phlogistic (Gr. *phlogos*, flame) state, that is, infection. Achieved mainly by bloodletting, purgatives, emetics, reduced diet, and rest.

cerebral fungi Nineteenth century term for extruded and disorganized brain tissue. Produced by infection.

frontal lobes Roughly the upper, forward half of the cerebral hemispheres.

prefrontal cortex Cortex of the anterior portion of the frontal lobes.

sensory–motor physiology The physiology founded in the first third of the nineteenth century that made the link between sensation and movement its basic unit of analysis.

In 1848, as the result of a bizarre accident, Phineas Gage had most of the left frontal lobe of his brain destroyed. Although his surviving the injury by some 11.5 years made him a considerable medical curiosity, it was the changes to his behavior that made him important in the neurosciences. Gage's is actually one of the most important cases in the history of the neurosciences: it revealed for the first time that complex functions might be localized in the brain. Its status is indexed by its still being cited in about two-thirds of all psychology and related neuroscience textbooks and by the fact that studies were still being undertaken some 150 years after the accident to establish which parts of Gage's brain were damaged.

I. GAGE'S BACKGROUND, WORK, AND ACCIDENT

Despite his importance in the neurosciences, very little was written about Phineas Gage at the time of or after his accident. John Martyn Harlow, the physician who treated Gage and followed up his case, wrote two papers about him, one in 1848 and the other in 1868. In between, Henry Jacob Bigelow, Professor of Surgery at Harvard, wrote another in 1850, and John Barnard Swets Jackson recorded additional information among his *Medical Cases*, some obtained from Gage's family. Jackson also recorded a little more, possibly obtained from Harlow, in his catalog of the Warren Anatomical Museum of the Medical School of Harvard University. In 1851, the editor of the *American Phrenological Journal* published a short comment. The written accounts of these four authors are the only contemporaneous ones and constitute the primary record of the case. A mute secondary record is provided by Phineas Gage's skull and the instrument with which it was damaged. It adds particular interest to his case by allowing us to see how his skull was damaged. Once forming part of the Warren Museum collection, both are now held in Harvard University's Francis J. Countway Library of Medicine.

Very little is known about Gage's behavior before his accident, which parts of his brain were destroyed or damaged, or what happened to him afterward. Partly

because so little is definite, a mythology has grown up about him and a large number of the references to his case in the literature are in grievous error. This article draws on the facts about him that were established or clarified only recently, mainly by the author. Except for Phineas' words of greeting to Williams, all of the direct and otherwise unacknowledged quotations are from Harlow's papers of 1848 or 1868.

Because so little is known about what Gage was like before his accident, there is no satisfactory baseline against which to evaluate the changes it produced. Something of his cognitive and organizational abilities may be inferred from the kind of work he was doing, the psychological and physical demands it made on him, and the context in which it was undertaken.

A. Gage's Background

Phineas Gage was born in Lebanon, New Hampshire. Although sometimes described as an Irishman, he was an eighth-generation American of Puritan origin, his remote ancestor being the John Gage of Massachusetts who arrived in the Americas in 1630, probably from Stoneham in Suffolk, England. The fourth generation of the family moved to New Hampshire some time before 1770 and farmed in the towns of Concord, Grafton, Enfield, and Lebanon. Phineas Gage was named after his grandfather and was the first of five children, three of whom were boys, born to Jesse Eaton Gage and Hannah Trussell Swetland. The precise date of his birth cannot be documented but may have been July 9, 1823.

We know nothing about Gage's early years, including his schooling, but during that time literacy rates were high in New Hampshire, and we know he was able to write. We are similarly ignorant about the work he did before his accident. We can be almost certain that he was skilled in such basics of farm work as growing crops and caring for cattle and horses, and probably sheep as well. More, the rocky soils of his family's farms probably meant that he had some knowledge of the use of explosives for obtaining building materials and in excavating cellar holes.

B. Gage's Work

By the time Gage was in his early twenties, a massive development of railroad construction had begun in New England. Phineas may have obtained or extended his knowledge of rock blasting in that industry while employed either on a nearby line, like that of the Northern Railroad Company, or in the equally close isinglass (mica) mines in Grafton. The first definite thing we know about him is that a firm of contractors had employed him as the foreman of a gang working on the route of the Rutland and Burlington Railroad (R&B RR). The gang was blasting a cutting, or cut, through a large rocky outcrop some 0.75 mile south of Cavendish, Vermont.

Although not a skilled occupation requiring formal training, blasting requires high-level cognitive abilities. The rifts or joints along which the rock will fracture have to be assessed, the sites, angles, and depths of the holes to be drilled must be decided, and the right amount of blasting powder has to be packed into place and fired safely. Incorrect judgments in planning or conducting any of these operations not only might result in ineffective or dangerous explosions but also could result in removal of too little rock, its fragmentation into pieces too large for easy removal, or its scattering over too large an area. Work that required considerable labor might then have to be repeated, and excessive damage at one site could increase the amount of work required at another.

In blasting of the kind Gage was undertaking, a hole was laboriously drilled by hand into the rock, a safety fuse (without the yet-to-be invented detonator) was put into it, and the blasting powder was poured over the fuse. Then, with a small crowbar-like tool known as a tamping iron, the powder was packed or tamped into the hole with a series of gentle but firm pats. Even though the charge so tamped contained a fuse, it was usually sufficiently stable not to explode. Sand, clay, or other material would then be placed over the charge and the whole tamped similarly, although more vigorously. The explosive force of a charge so prepared would then be directed into the rock at the end of the drill hole.

Gage's physical and psychological attributes suited him for this work admirably. Physically he was, Harlow tells us, "a perfectly healthy, strong and active young man," with an "iron frame," "vigorous physical organization," and an unusually well-developed muscular system, having had scarcely a day's illness from his childhood. He was of "middle stature" (actually 5 ft, 6 in.), weighed on average 150 lb, and had a "nervo-bilious temperament." He had an iron will, was of temperate habit, "possessed of considerable energy of character," and was "very energetic and persistent in executing all his plans of operation." He had a well-balanced mind, and his contractors said he was the most capable and efficient foreman in their

employ. Harlow described him as being known as "a shrewd, smart, business man." In the New England vernacular of the day, *shrewd* was synonymous with clever, keen-witted in practical affairs, and astute or sagacious in action or speech; *smart* was synonymous with being quick, keen, active, industrious, energetic, clever, and intelligent.

From today's perspective, Harlow's use of the two-word term *business man* may seem to conflict with Gage's position as foreman. But, like many other New England foremen in the nineteenth century, Phineas may well have been an independent subcontractor who tendered competitively for work to be performed by a gang he had recruited. On the other hand, *business man* was just beginning to acquire its modern meanings and then still meant mainly one who organized the work of others. Subcontractor or not, foreman Gage had to be able to plan and maintain an efficient work schedule.

Here the fact that Gage was "a great favorite" with his men is especially worth noting. In whatever position Phineas stood in relation to his men, he would have had to allot tasks to them fairly, record their working times accurately, treat them equally, and pay them properly. Foremen who did not so treat their men were unpopular and sometimes subject to violent and occasionally fatal attacks, and murderous assaults of this type were not unknown in the Cavendish area.

C. The Accident

At 4:30 PM on September 13, 1848, a charge Phineas was tamping accidentally exploded. Precisely how this happened is not known. The details in the primary sources differ, but what is fairly certain is that Gage's attention was distracted between the first and second stages of the tamping and he began the more vigorous tamping before the sand had been added. The tamping iron struck the hole and it caused a spark that set the blasting powder alight. The iron was 3 ft and 7 in. long, 1.25 in. in diameter at the larger end, tapering over a distance of about 12 in. to a diameter of 0.25 in. at the other, and weighed 13.25 lb. This now formidable missile was blown completely through Gage's skull. It entered, pointed end first, under the zygomatic arch or cheekbone, penetrated the base of the skull behind the eye, and emerged near the junction of the sagittal and coronal sutures to land some 20–25 m away.

Gage was knocked over but may not have lost consciousness. He walked, either by himself or with a little assistance, to a nearby ox cart in which, while sitting and supporting himself against the foreboard, he was driven about 0.75 mile to the inn or tavern in Cavendish where he lived. On arriving, Phineas got up, walked to the back of the cart, allowed himself to be helped down, and then walked to the veranda where he sat on a chair. From there he talked to the small crowd that had gathered about what had happened to him until Edward Higginson Williams, the first medical practitioner to reach him, arrived at about 5:00 PM. Phineas then greeted Williams with one of medicine's great understatements, "Doctor, here is business enough for you."

II. THE SEQUELAE OF THE ACCIDENT, 1848–1868

Gage's treatment lasted some 11 weeks, and during it there were minimal signs that his behavior had changed. He seems to have taken at least a further 6 months to recover, and there are no reports of any untoward consequences from that time or from when Henry Jacob Bigelow saw him in Boston 5 months later again. In 1868, Harlow gave a very condensed account of Gage's postaccident years that included about 200 words summarizing the main psychological changes.

A. The Treatment, 1848

Indeed there was business aplenty for Dr. Williams. That was so despite the myths that have Gage recovering with little or no treatment, sometimes with his having the tamping iron in his skull for the rest of his life, or walking several miles to a doctor's office to have it removed. Although Williams did not see the entry wound under the cheekbone until Phineas pointed it out, he could see the inverted funnel shape of the wound on the top of his head even before alighting from his carriage. Gage's major symptoms were profuse hemorrhaging with frequent vomiting of the blood and brain tissue that had drained into his stomach. The "deep burns" on his face, hands, and arms were of less significance.

Harlow arrived at about 6:00 PM and helped Gage walk up the stairs to his room, where he and Williams had the scalp shaved and cleaned the wounds. Harlow removed three small pieces of bone and examined the wound with his fingers, eventually joining the right index finger, which he had entered from the top of the skull, to the left entered through the cheek. So assuring himself that there were no bone fragments still inside

the skull, he and Williams proceeded to dress the wounds. They replaced the two large pieces of bone that had been detached or everted, brought the soft tissue edges together and held them there with adhesive straps, and covered the whole with a wet compress, a night cap, and roller, but they left the wound in the cheek open. They dressed the burns and left Phineas in the care of attendants who had been instructed to keep him in a semirecumbent position with his head elevated.

Harlow visited three times a day to change the dressings and soon had to treat an inflammation (i.e., infection). He based his treatment on the conventional antiphlogistic regimen, rest, and laxatives, prescribing the latter very moderately, but combined them with remedies from the stimulant regimen, probably brandy and milk. Apart from two short episodes of delirium associated with the infection, Gage was quite rational. Soon after the second episode, 11 days after the accident and just as Phineas seemed to be recovering, the fungi that had been sprouting from the wound for 5 days grew much worse. An abscess probably communicating with the left ventricle had formed, and, 14 days after the accident, Harlow was forced to treat Gage more energetically by cutting the protruding fungi away, applying caustic to the rest, and draining the abscess. After remaining semicomatose for about 10 days, Gage then began such steady progress that he was able to walk about the inn and town 56 days after the accident.

Two weeks later, during a short period when Harlow was away from Cavendish, Phineas disobeyed Harlow's (and his attendants') instructions and planned his return to Lebanon. His expedition to the local store, part of this plan, was in cold, wet weather and gave him a chill and such a fever that Harlow again had to resort to energetic treatment. This time purgatives and emetics were administered freely, and he was bled of some 475 ml of blood. Within 3 days Gage had recovered, and was well enough a week later, on November 25 only 73 days after the accident, to be taken home, although in a close carriage.

Harlow completed the first of his two papers on the case 2 days later, on November 27, and he sent it off as a Letter to the Editor of the *Boston Medical and Surgical Journal*, in which it was published on December 13. Much of the paper was based on his daily case notes, from which it is clear that apart from the episodes of delirium associated with the infection, Gage had been substantially rational throughout. Three of Harlow's entries are especially worth noting, however: Phineas had been unable to estimate size or money accurately and would not take $1,000 for a few pebbles; he was very childish; and his recovery was dependent on his being controlled. Harlow concluded by opining that the case would be of interest to "the enlightened physiologist and intellectual philosopher" and foreshadowed a future communication on "the mental manifestations."

B. The Real Gage, 1849–1851

After examining Phineas in Lebanon on January 3, 1849, and again when he came to Cavendish in April that year, Harlow said he was "inclined to say that he has recovered" physically. Phineas told Harlow that he had applied "for his situation as foreman," an intention he also foreshadowed to someone else a few weeks earlier in March. However, in August, J. B. S. Jackson learned from the family that Phineas had not been capable of more than half a day's farm work until May or June. Jackson did not see Gage then as he was away trying to regain his railroad job. And although Phineas seems to have sent his tamping iron to Henry Jacob Bigelow in May 1849, he does not seem to have been well enough to visit Boston and be examined by Bigelow until November. There is also some evidence that he was still physically weak as late as April 1850.

Almost all we know about the marked psychological changes caused by the accident comes from Harlow's 1868 summary. Gage never regained his job as foreman. Twenty years after the accident Harlow explained that his contractors "considered the change in his mind so marked that they could not give him his place again." In about 160 words, Harlow set out the basis for their decision:

> *The equilibrium or balance, so to speak, between his intellectual faculties and his animal propensities, seems to have been destroyed. He is fitful, irreverent, indulging at times in the grossest profanity (which was not previously his custom), manifesting but little deference for his fellows, impatient of restraint or advice when it conflicts with his desires, at times pertinaciously obstinate, yet capricious and vacillating, devising many plans of future operation, which are no sooner arranged than they are abandoned in turn for others appearing more feasible. A child in his intellectual capacity and manifestations, he has the animal passions of a strong man. Previous to his injury, although untrained in the schools, he possessed a well-balanced mind, and was looked upon by those*

who knew him as a shrewd, smart business man, very energetic and persistent in executing all his plans of operation. In this regard his mind was radically changed, so decidedly that his friends and acquaintances said he was "no longer Gage."

Some of the minor ominous changes that Harlow had mentioned in passing in 1848 seemed to have become permanent.

From Harlow we learn that Gage traveled around the larger New England cities, where he exhibited himself and his tamping iron, and spent some time as an exhibit with P. T. Barnum's American Museum in New York. Harlow also tells us that Phineas worked for Jonathan Currier for nearly 1.5 years until August 1852, after which Gage left for Chile. Gage could not, therefore, have begun work for Currier before the beginning of 1850, and that in turn means that his travels and self-exhibition could not have lasted more than 1.5 years, even if they began in mid-1849. We do not know exactly what he did during his period of travel because there seems to be no record of anyone seeing him. On the other hand, we do know that when Phineas worked for Currier it was in the livery stable associated with his Dartmouth Inn in Hanover, New Hampshire. Currier also ran a coach service, and it may be that Gage learned to drive coaches there. Certainly when he left Currier to go to South America it was with a man who intended to found a coach line at Valparaiso. Phineas was engaged to look after the horses, and he did drive coaches there.

C. The Real Gage, 1852–1860

After a period of 6 years in South America, Phineas' health began to fail and he had a long but unspecified illness. He therefore returned to the United States, now to San Francisco, where he lived with his mother, sister, and brother-in-law. By then he had acquired some nieces and nephews whom he used to entertain, his mother told Harlow, with

the most fabulous recitals of his feats and hair-breadth escapes, without any foundation except in his fancy. He conceived a great fondness for pets and souvenirs, especially for children, horses and dogs—only exceeded by his attachment for his tamping iron, which was his constant companion during the remainder of his life

When he regained his health, he began laboring work on a farm in Santa Clara County, south of San Francisco, and worked there until he had his first epileptic convulsion in February 1860.

Phineas then became unsettled and moved around from farm to farm, each time becoming dissatisfied with his employer. On May 18 he returned to his mother's, where, on May 20, he had a severe convulsion. It heralded the end. After a series of increasingly severe seizures, Phineas died at 11 PM on May 21, 1860. No death certificate has been found, but the funeral director's record gives the cause of death as "epilepsy." Harlow calculated that Phineas had survived "twelve years, six months, and eight days" after his accident, a degree of accuracy marred only by the fact that Harlow had been told that the year of death was 1861. We have known only since 1986, when the author published the correction, that 1860 is the correct date. Gage had, therefore, survived 11.5 years.

Harlow does not seem not to have learned of Phineas' death until some time in 1866. He then corresponded with Gage's mother and brother-in-law, eventually persuading them to allow an exhumation so that the skull could be removed for study. No date has been found for the exhumation, but it seems to have taken place in early December 1867, and Phineas' brother-in-law seems to have taken the skull immediately to Harlow, then practicing in Woburn, Massachusetts, where it arrived in 1868. Harlow had photographs made of the skull and tamping iron (Fig. 1a–c), and the woodcuts prepared from them, which he used to illustrate his second paper on Gage presented to the Annual Meeting of the Massachusetts Medical Society on June 3, 1868. It included, for the first time, the full, even dramatic, account of the changes in Phineas' behavior. At the conclusion of his paper, he presented the skull and tamping iron to the Warren Museum.

D. The Mythical Gage, 1849–1860

The myths about Gage tell a very different story. Either because he cannot hold a job down or because he does not want to work, Phineas becomes a vagrant who drifts around the United States, and sometimes South America, eking out a living as a fairground or circus attraction. The fanciful accompaniments to this story include Gage losing his ability to make and carry out plans, his becoming impossible to work with, his turning into an arrogant bully who flies into rages and loses his temper easily, his becoming a psychopath who cannot be trusted to honor his promises, his becoming sexually promiscuous, and his degenerating into a boastful alcoholic who dies of drink.

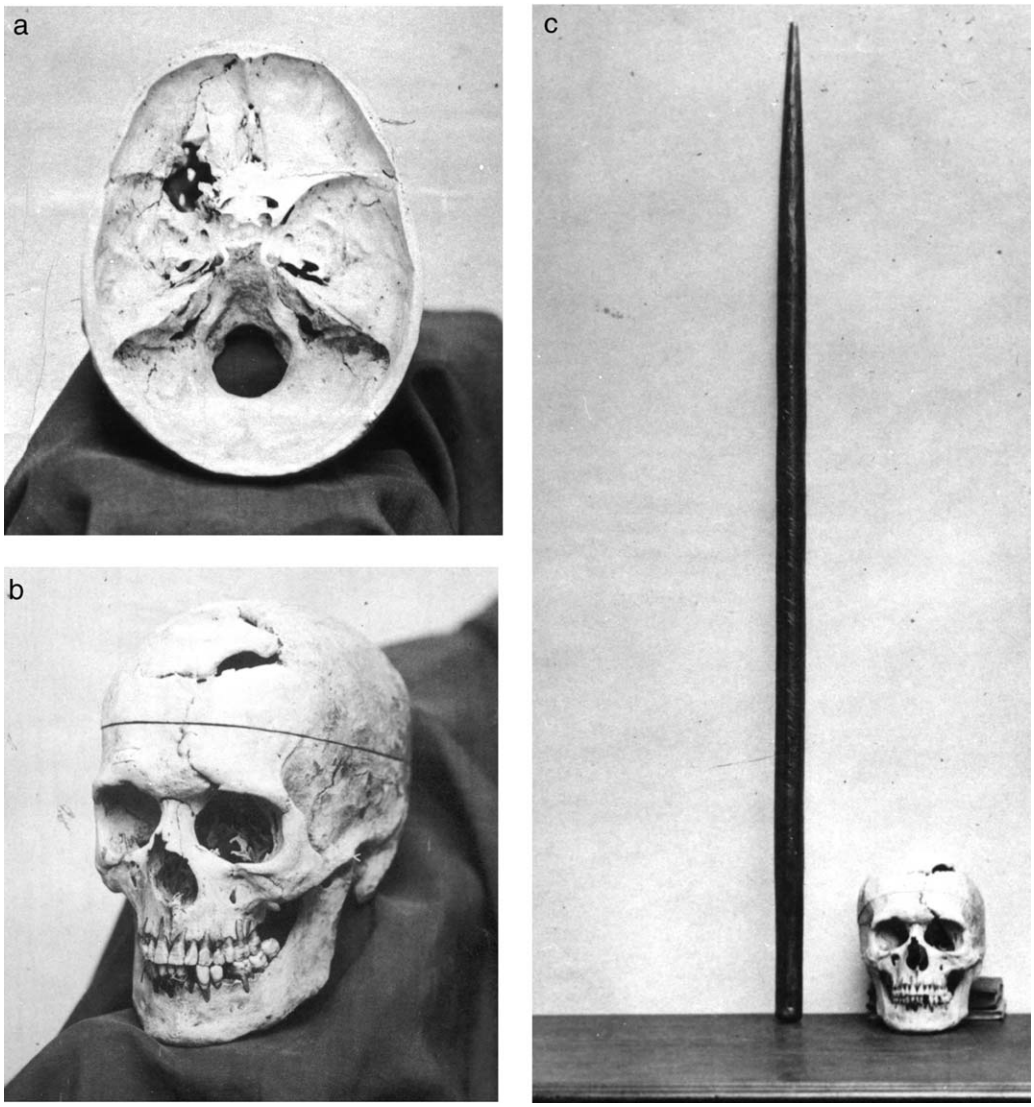


Figure 1 S. Webster Wyman's 1868 photographs of Gage's skull taken for Harlow: (a) base, (b) front left, and (c) in relation to the tamping iron. From the Glennon Archives, Woburn Public Library, Woburn, MA, reproduced by permission of the Trustees of the Library.

Many of the myths are based on a simple lack of knowledge of the primary reports, and many writers appear to have been content to parrot what they had read in versions far removed from the originals. Without the sources, it is not possible to calculate that the maximum time during which Phineas wandered could not have been more than 18 months, but the myth makers have him traveling over whole time of his survival. Similarly, unable to appreciate that Gage probably worked at the one occupation for just two employers during 7 of his 11.5 postaccident survival years, they have him drifting from job to job. And, without the sources, it is the popular memory of P. T. Barnum as the proprietor of a circus that surfaces to interpret Gage's time with him as traveling with Barnum's Circus rather than being in his stationary Museum in the city of New York.

Some parts of the myths do have a basis in fact, but others do not. One wonders, for example, where a Gage who sells his skeleton to two different medical schools, cash in advance, could come from. Where there are facts, they are sometimes not facts about Phineas. Thus, we know nothing about Phineas' sexual behavior, but we do know something about the effects of frontal lobotomy and frontal damage on that behavior in others, and it is that which matches the

Gage of myth. Similarly, none of the primary sources mention Gage's drinking. His reputation as a drunkard seems to have come from a coupling of a parallel some neurosurgeons drew between the euphorias produced by alcohol and frontal lobectomies with some gossip in the Boston medical community that commenced years after Gage's death. Those parts having some connection with Phineas are gross exaggerations of the real Gage. For example, his supposed psychopathic lying seems to be based on the "fabulous recitals" with which he entertained his nephews and nieces.

It must be remembered that, at the time of this writing (February 2002), only four records about Phineas Gage by contemporaneous authors survive. Anything about him not documented in sources as trustworthy as these must be fabrication.

III. THE DAMAGE TO GAGE'S SKULL AND BRAIN

There has always been some debate about the trajectory of the tamping iron through Phineas Gage's skull. In 1848 and the 15 years that followed that debate was fairly academic, but in the 1860s and 1870s it acquired greater significance because definite psychological functions were then assigned to or localized in the brain. From the passage of the iron through the skull, it was hoped to deduce which parts of Gage's brain had been damaged. It was also not until then that it became known in medical circles, although not widely, that his behavior had changed.

A. The Trajectory of the Tamping Iron

The tamping iron caused damage in three places: the area under the zygomatic arch, the base of the skull where it had entered behind the eye, and at the top of the head where it had emerged. Two things came to be important. First, how far forward or behind the junction of the coronal and sagittal sutures had it emerged? The answer would determine the extent to which the damage to Gage's brain was confined to the frontal areas. Second, how far to the left of the midline of the base of the skull had it entered and did it stay left of the midline at the top of the skull when it emerged? This would determine the extent to which the left and/or the right frontal lobes were damaged.

On the top of the skull the bony damage was extensive. There is an area of total bone destruction, roughly triangular in shape about 2.0 in. wide and 4.0 in. in circumference (D in Fig. 2). It is flanked by two semidetached flaps, the one from the frontal bone (B) being a rough quadrilateral about 2.0 in. wide at the widest point and 2.5 in. long and the one behind it about 2.0 in. wide and 1–1.5 in long (C). Williams had noticed their eversion even before alighting from his carriage. Together with the two areas of total destruction (A and D) these two flaps constitute an area of total damage about 2.0 in. by 5.0–6.0 in. Thus, there is not a *point* of emergence; rather, there is a large and irregular *area* of damage within which that point lies (Figs. 3 and 4).

Although the damaged area under the zygomatic arch was circumscribed, it was not at all clear where the iron had penetrated the base of the skull behind the eye. In 1848 Harlow could, naturally, only speculate about the locus and extent of that damage, and Bigelow had the same problem. When he drilled holes in a skull to show that the tamping iron could pass completely through it, he seems to have worked backward from where he estimated it had come out. His judgment of one-third of the exit to lie to the right of the midline on the top of the skull then placed the entry in the base well to the left of the midline. When Harlow examined Gage's skull in 1868, it was clear that the area of basal damage was very irregular and

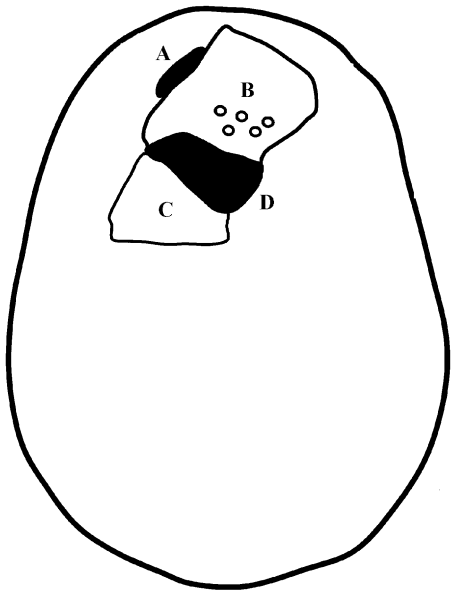


Figure 2 Gage's skull showing approximate dimensions of areas of loss and damage and the Damasio exit points (A) Left frontal (total loss) 2.6-in. circumference; (B) frontal flap (partly reunited) 2.0 × 2.5 in.; (C) rear flap (reunited) 2 × 1.5 × 1.0 in.; (D) top of skull (total loss) about 2.0 in wide and 4.0 in. in circumference. Circles indicate Damasio viable exit points.

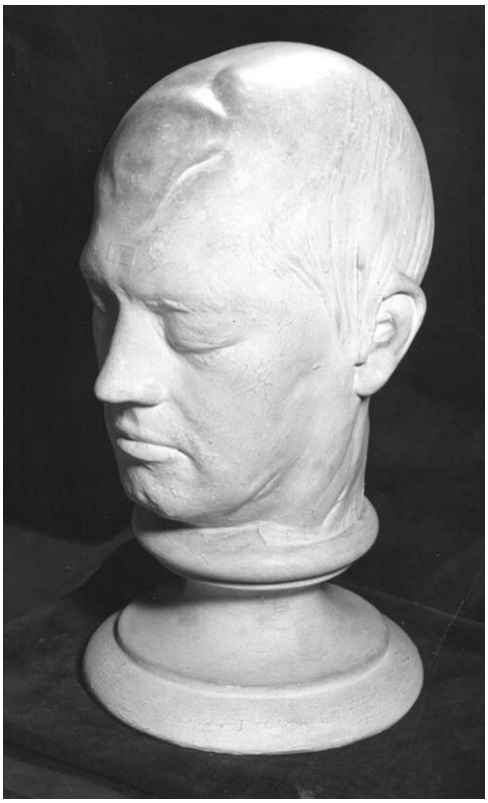


Figure 3 Left frontal view of Phineas Gage's life mask, probably made for Henry Jacob Bigelow in late 1849. From the late Dr. H. M. Constantian of Worcester, MA, and reproduced by courtesy of the Warren Anatomical Museum, Harvard University.

not much bigger than the diameter of the tamping iron at its largest point. But it was bigger, and even after healing its major axes were about 1 in. wide and 2 in. long (Fig. 1a). Harlow concluded that the iron had entered closer than had Bigelow, about 1.25 in. to the left of the midline, but Jackson had placed it even closer, only 1 in. away.

A "true" entry point could be at a number of different places along the axes of the damaged area. This indeterminacy has two consequences: where the entry is placed along the now 2 in. axis affects the location of the exit point on the top of the skull relative to the coronal suture. If the trajectory were vertical, the further forward the entry, the further forward the exit point, and vice versa. Similarly, side to side variation along the now 1 in. axis affects the left–right placement of the exit point. The more the entry is to the left of the midline of the base, the more the exit will be to the right of the midline on the top of the skull, and vice versa.

In 1848, Harlow thought the iron had emerged in the midline and at the junction. Edward Elisha Phelps, Professor of the Theory and Practice of Physik and Pathological Anatomy at Dartmouth College and apart from Harlow and Bigelow the only person known to have examined the live Gage, thought it came out 0.5 in to the front and 1.0 in to the left of the midline. Bigelow, as we have seen, had it emerging in front of the junction but somewhat to the right. In 1868, after examining Gage's skull, Harlow still maintained that it had emerged in the midline but now only "near" to the junction. Jackson, who also used the skull, placed it in front and noted that the area of bone loss crossed the midline, but he was vague about the emergence of the iron in relation to the midline.

B. The Damage to Gage's Brain

None of the preceding investigators explicitly discussed the trajectory in relation to the debate over localization, which, apart from earlier inconclusive arguments over Franz Josef Gall's organology (phrenology), began in earnest in 1861. In that year, Paul Broca, a French surgeon, very tentatively localized a language function in the frontal areas, and by 1864 he had placed it firmly in the third left frontal convolution. That location was to be strengthened with David Ferrier's work, which followed Fritsch's and Hitzig's in demonstrating that limb movements could be elicited by electrical stimulation of precisely delimited areas in the brain on the side opposite the limb. By 1876 Ferrier had shown that areas controlling lip and mouth movements in the monkey's brain were adjacent to the homolog of the human language area and both were near those that controlled the limbs. Where Gage's tamping iron had emerged became important with Eugene Dupuy's claims in 1873 and 1877 that the pattern of damage contradicted both Broca's clinical and Ferrier's experimental data. Dupuy believed that the left-sided trajectory was posterior enough to destroy the language and motor areas, but Gage, of course, had no motor or language impairments.

At that time, Ferrier had no detailed knowledge of Gage's case and wrote to Henry Pickering Bowditch of Harvard in 1877 asking him to examine the skull and help resolve the trajectory of the tamping iron. Bowditch sent Harlow's photographs and his 1868 paper, and Ferrier's analysis of them convinced him that the iron had emerged more frontally without damaging those areas. Ferrier concurred with Harlow that only the left frontal lobe had been damaged, and his opinion prevailed over Dupuy's. The matter essentially well rested there until the 1982 CT scans of

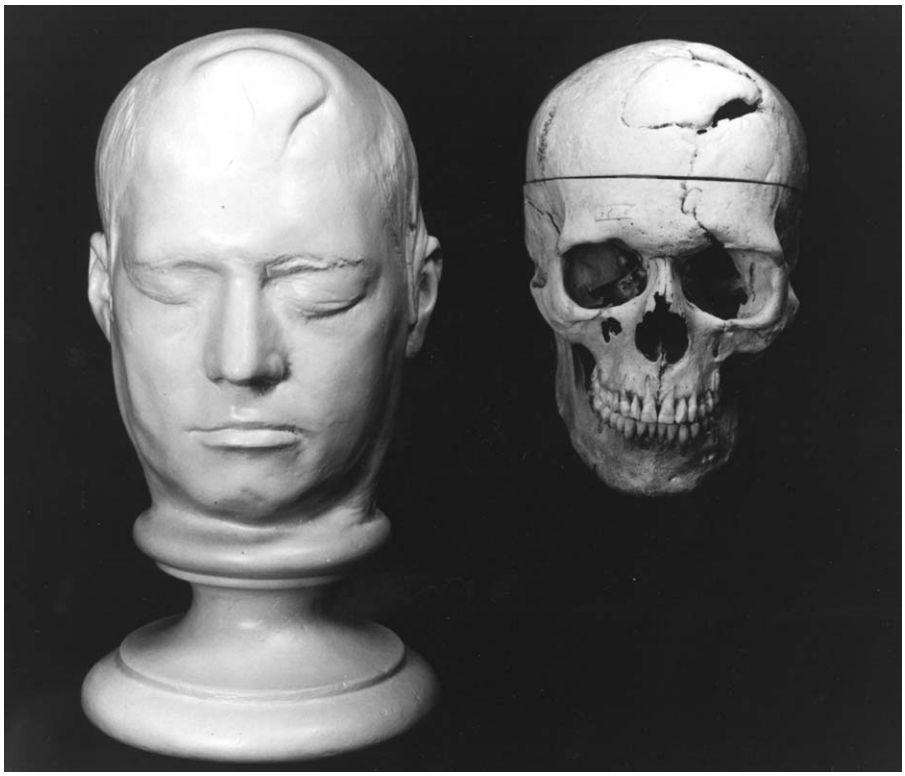


Figure 4 Frontal view Phineas Gage's life mask (probably made for Henry Jacob Bigelow in late 1849) and skull (body probably exhumed in late 1867). Reproduced by courtesy of the Warren Anatomical Museum, Harvard University.

H. Richard Tyler and Kenneth L. Tyler and the 1994 computer modeling of Hanna Damasio and her colleagues.

The Tylers' successive lateral and coronal (frontal) scans were made to determine the bony damage in the entry and exit areas. Imaginary lines joining the sides of the maximum openings at the top and base of the skull as viewed in coronal section defined a funnel shape, narrower at the bottom than at the top. If the funnel does set the limits of bony damage, most of it lies to the left of the midline. The Tylers could not determine precise entry and exit points, but they believed that Harlow was more likely to have been right about the entry and Bigelow about the exit encroaching over the midline. Most of the direct damage the tamping iron did to the brain was to the left anterior frontal lobe, part of the tip of the left temporal lobe, part of the anterior horn of the left lateral ventricle, the head of the caudate nucleus and the putamen, and the superior sagittal sinus. Some direct damage was also done to the right hemisphere, including parts of the right superior frontal and cingulate gyri. With the exception of right hemisphere involvement, their conclusion again is similar to Harlow's, and including the right-sided damage it is very similar to Bigelow's. However, the Tylers stressed that theirs was a minimal estimate of the damage to Gage's brain. Bony fragments being driven through the brain caused more damage, some tissue was lost during the vomiting caused by hemorrhaging, and more damage resulted from later infection.

Hanna Damasio and her group used measurements from Gage's real skull to create a computer model of it and estimate the trajectory of the tamping iron. They began on the top of the skull and confined themselves to what they called the area of bone loss—the area of total loss plus the reunited frontal flap—but they did not include the reunited posterior fragment. Half the diameter of the iron was used to set more precise boundaries to the area within which they tried to estimate its passage: no point could be closer to their margins than it. Within that area they located 15 possible exit points in addition to an undefined *a priori* most likely exit point. From each of these points they projected lines through the center of the area in the base of the skull to the area under and adjacent to the

zygomatic arch where the iron had first entered. There they again used half the diameter of the iron to limit their estimate of its point of entry. From the trajectories so produced, seven seemed anatomically possible, but they rejected two of these because they would have hit the front of the left lateral ventricle, damage they believed Gage would not have survived in the preantibiotic era.

Damasio's group then simulated the passage of the iron along these five trajectories through seven models of real brains closely matching a model of Gage's brain derived from the internal measurements of his skull and given an average position within it. Each simulation produced a similar pattern of damage in both frontal lobes, but it was more pronounced in the left: the anterior half of the orbital and the polar and anterior mesial frontal cortices and the anterior-most part of the anterior cingulate cortex.

Whether these modern methods have settled the pattern of bone and brain damage is doubtful. The Tylers claimed only to have established the maximum and minimum damage to the brain caused by the passage of the iron and pointed to three other factors (bone fragments, hemorrhage, infection) that must have added to that. The Damasio group did not search for trajectories in the rearward areas and rejected those that would have taken the iron through the area of total bone loss, and they settled on trajectories with "points" of emergence under the reunited but otherwise undamaged frontal flap itself (circles, Fig. 2). They also assumed that Gage could not have survived ventricular damage, but the likelihood of damage of that kind had surprised neither Harlow nor Bigelow.

C. The Relation between Gage's Brain and His Behavior

All of the methods used so far to estimate the damage to Gage's brain have been based on two fundamental assumptions. First, all assume that Gage's brain, or its model, lies in an average position relative to the bony landmarks of the skull. Yet it is known that the positions of real brains may vary by as much as 0.5 in from those markers. Second, all assume that there is little variation in the localization of functions in the brain. Yet it is also known that the same functions may be localized in somewhat different places in seemingly similar brains.

Suppose, however, that the brain damage could be estimated accurately. Would that estimate help us understand what functions localized there had been disturbed? Or which of Gage's functions? When the very limited information we have about Gage's postaccident behavior is contrasted with the even smaller amount we know about his preaccident characteristics, it is clear that both answers must be "No". There is another set of problems, however. Harlow noted some deficiency in the ability to judge size and the value of money, and Jackson recorded that the family believed Phineas had a memory defect; however, judging from his apparently long-term employment as a coach driver in a region notorious for its bad roads, high-level cognitive–motor skills may have survived. What can be inferred about the detail of Gage's postaccident cognitive functions is as uncertain as what can be inferred about the damage to his brain.

IV. GAGE AND LOCALIZATION

Four main factors make for complexity in the relation of the case of Phineas Gage to the debate over the psychological functions served by the brain. First, there were profound philosophical and possibly political differences between those supporting and those opposing any doctrine of localization. In the first half of the nineteenth century, admission that the mind was localized in the brain seemed to encourage a godless materialism. Second, there were differences, some legitimate, between medical innovators and the orthodox in evaluating Franz Josef Gall's organology. Were the faculties of the mind fractionated, as Gall argued, or did the brain operate as a whole, as Flourens' experiments seemed to indicate? Third, there was the effect of the 20-year interval between Harlow's first rather sanguine report and his 1868 account of the deleterious psychological effects of the damage. In the beginning the Gage case seemed to weigh against any doctrine of localization, but later there was little doubt of its general importance. Paradoxically, the fourth factor derived from that very significance: other than the limited sensory–motor physiology and psychology of the day, no explanatory framework existed within which the changes in Phineas' behavior could be placed.

A. Gage and Phrenology

Gall had postulated that the various nonmaterial faculties he required for explaining behavior were expressed through material organs localized in the brain under various bony prominences he believed he could detect on the skull. Near the junction of the

coronal and sagittal sutures he located the faculties of Benevolence and Veneration. Benevolence provided the individual's conscience and regulated such feelings as compassion and the moral sense, whereas Veneration, located posterior to it, regulated religious feelings and, at least in the eyes of some of Gall's followers, provided the basis for respecting others, especially those in authority. These feelings were well-developed or even exaggerated in those with large faculties, that is, with large prominences. In those in whom Benevolence and Veneration were underdeveloped, the character was malicious, vindictive, and ungrateful. Together with other anteriorly placed faculties, the two were also components of a system of Moral Sentiments that regulated the posteriorly located Animal Passions that humans shared with animals. These latter included the instincts of Love of Offspring, Propagation, Social Attachment, Self-defense, and Destructiveness.

There is evidence that Harlow accepted much of Gall's doctrine, probably in the form that Caspar Spurzheim had introduced to the United States. In 1848 he used the phrenological concept of nervobilious temperament to encompass the combination of great mental capacity and physical strength basic to Gage's character, and he used terms more common among the phrenologically inclined when he foreshadowed a future communication about those of Gage's "mental manifestations" that would be of interest to "the enlightened physiologist and intellectual philosopher." His drawing attention to the defect in Phineas' ability to value money is also consistent with damage to the Faculty of the Relation of Numbers that Gall located frontally. His 1868 description of Gage's mental operations being deficient in degree rather than in kind and his insistence that only the left lobe was damaged may reflect Gall's basic proposition that all of the faculties were localized equally in both hemispheres. Functions impaired through damage to one lobe were conducted by the other, although less efficiently.

Harlow himself did not explain the changes in Gage by invoking damage to Benevolence and Veneration. In early November 1848, an anonymous letter writer did mention that Veneration had been damaged but said nothing about the effects. In 1851, the editor of the *American Phrenological Journal* published the first description of the changes in Gage and explained them by damage to the two organs. In saying his information came from the best authority in Cavendish, he was very probably attributing them to Harlow himself. The description and phrenological explanation were part of a comment on Bigelow's report on Gage. Bigelow had seen Gage in Boston over a period of some weeks from late 1849 to early 1850, and not only did he say absolutely nothing about Gage's behavior having changed, he implied that there had only been an "inconsiderable disturbance of function." If we consider that Bigelow had also proposed that both frontal lobes had been damaged, his stance in this paper was as decidedly antiphrenological as his beliefs generally are known to have been.

B. Gage and Ferrier

In 1876, Ferrier had been unable to elicit responses by electrically stimulating the frontal areas of the exposed monkey brain, and ablation of the same areas did not seem to produce sensory or motor symptoms. Yet the behavior of the monkeys had changed. Even though they seemed as intelligent as monkeys with intact lobes, they were no longer curiously prying into everything around them but were either apathetic, dull, and dozing or restless, lacking purpose, and responding only to immediate sensations and impressions. The change was best summed up as a loss of the faculty of attentive and intelligent observation. Ferrier placed a hypothetical inhibitory–motor function in the frontal lobes that inhibited immediate motor responses to external stimulation until the most appropriate response could be selected from the pool of potential responses. He was otherwise unable to say what functions the lobes performed. He did not know then that Gage had changed and mentioned him, not because his behavior was like that of the monkeys, but simply because he had also survived similar frontal damage.

Estimation of the trajectory of the tamping iron was central to Ferrier's response to Dupuy's 1877 attack, and that task led him to Harlow's 1868 account of the change. In his *Gulstonian Lectures* of 1878, he now drew a parallel between the changes in the behavior of Gage and that of his monkeys: both had lost the ability to delay responses other than the one most appropriate to a given situation. This inhibitory–motor thesis had a very short life in Ferrier's own theorizing, being withdrawn between 1879 and 1886, but it lived on until some time after 1900 in discussions of the 'mental symptoms' that warranted surgery on the frontal lobes.

V. GAGE AND SURGERY FOR THE BRAIN AND PSYCHE

Brain surgery began almost as soon as the usefulness of localizing signs was recognized and safe, aseptic

operating techniques had been developed. In 1871 and 1876, respectively, Broca himself and William Macewen, the Scottish pioneer of brain surgery, apparently used aphasic symptoms to guide their intracranial operations. Bennett and Godlee, who performed the first operation for the removal of a tumor in the brain in 1885, specifically used localizing signs derived from Ferrier's experimental work and Hughlings Jackson's clinical deductions to plan theirs. Knowledge of Phineas Gage's case was important to the extension of brain surgery to the frontal lobes but not to psychosurgery itself.

A. Gage and Brain Surgery

In 1884, M. Allan Starr, an American neurologist, published the first major series of cases in which an attempt was made to relate symptoms to the site at which the brain was damaged or diseased. Gage figured prominently in Starr's considerations, as did Ferrier's inhibitory thesis (although not connected with his name), and he recommended that frontal lesions be considered when symptoms were present that indicated a lack of self-control: an inability to fix the attention, follow a train of thought, or conduct intellectual processes. In 1893, solely on the basis of such mental symptoms (inability to fix attention, dullness and slowness in thinking, difficulty in expressing ideas, and a general weakness and aversion to work), he and McBurney diagnosed and removed a frontal tumor. It is possible that Macewen had used, even earlier in 1879, what he took to be Gage-like symptoms of frontal impairment in planning a frontal operation. He seems to imply that it was mainly the mental symptoms shown by his patient (her obscured intelligence, slowness in comprehending, and lack of "mental vigour") that led him to explore the frontal lobes. There he found and successfully removed a tumor of the dura mater that had spread over two-thirds of the left frontal lobe.

Although some surgeons did follow Starr's recommendations, it soon became apparent there were major difficulties with them. Definite symptoms were, as Starr had himself noted, present in only about one-third of the cases, physicians were not trained to assess psychological changes, and there was wide variation about which mental symptoms indicated loss of self-control and inability to fix one's attention. There were also practical difficulties. Because they exerted their effects at places remote from their location, tumors localized using the signs were not always at the places indicated, and when they were, they were not always encapsulated enough for easy removal. Soon after the beginning of the twentieth century, Ferrier's thesis as understood by Starr and others ceased to be used diagnostically. In any case, various seemingly more reliable imaging techniques that became available at about that time displaced them.

B. Gage and Psychosurgery

In both popular and professional minds, Gage is firmly associated with the introduction of prefrontal lobotomy. However, apart from his being the main example of survival after frontal damage, there is no evidence that considerations of Gage's case played any part in the development of that form of psychosurgery or its precursors. That is, no one seems to have reasoned that the kind of disinhibited behavior produced accidentally in Gage would bring about benefit if deliberately induced in psychiatric patients.

Psychosurgery began in a small way toward the end of the nineteenth century with operations by Bennett (in collaboration with Gould) and Macewen to relieve psychiatric symptoms in otherwise essentially normal people. They surgically removed tissue and bone fragments that caused symptoms like visual and auditory symptoms. Gottlieb Burckhardt, who operated on psychiatric patients in 1888–1889, was more ambitious. He aimed to disconnect the sensory from the motor areas in the hope that violent behavior seemingly caused by sensory hallucinations then would not occur. In only one of his six cases did he venture into the frontal areas, and it seems to have been because he thought that the patient's ideas located there were the cause of his abnormal behavior. Neither he nor Emory Lanphear (who operated between 1891 and 1895) made any mention of Gage. What Ludwig Puusepp, the other psychosurgical pioneer, actually did and when are not clear, but the changes in Gage do not seem to have formed part of his rationale.

One of the main developments that made psychosurgery possible was the results of the radical resections conducted in the early 1920s, notably by Walter Dandy. He had concluded that, because most brain tumors were not encapsulated, the removal of that part of the brain in which they were located was indicated. The procedures he developed were successful and soon adopted by others, especially for the control of otherwise intractable epilepsy. By the early 1930s, when Egaz Moniz began planning his form of

psychosurgery, it was evident that life was not threatened by the removal of such large parts of the brain, nor did there seem to be any untoward psychological consequences.

The particular sources that Moniz drew on were his own confused ideas about brain function, the changes reported by Brickner in Joe A., the New York stockbroker who had had major bilateral frontal resections for a massive life-threatening tumor, and Moniz's selective and partial consideration of the experiments by Jacobsen and Fulton on the effects of frontal ablation on learning in chimpanzees. Moniz believed that morbid ideas, as he called them, were due to a fixed pattern of neuronal activity and that surgery altered the pattern by disconnecting the neurons. For no apparent reason, he associated psychological activity with the frontal lobes and decided to operate there. Gage did not come into his rationale.

It is true that Walter Freeman included Gage among the topics he mentioned when he spoke to a group of newspaper reporters at a press conference immediately prior to presenting the results of his and Watts' first lobotomies. But his unpublished account of the conference shows that this was part of an impromptu smokescreen to put one of the reporters off the track of a story he had already given another reporter. He said merely that Gage had survived with mental symptoms, which he did not specify, and later made nothing more of it. Freeman did publish a full account of Gage's case in his first book with Watts but made no connection between it and the new operation other than Gage surviving his injury. Rather like the newspaper reporters, none of whom mentioned Gage in their stories, Freeman did not mention the case at all in the second edition of his work.

If Gage is mentioned at all in various symposia devoted to the effects of frontal surgery, particularly those of the (American) Association for Research in Nervous and Mental Disease, it is only in a minor way, usually to reject him as providing a relevant behavioral standard for comparison.

VI. STORIES ABOUT PHINEAS GAGE, 1848–2002

Almost from the beginning, many of the stories about Phineas have included considerable inaccuracies and distortions. As noted earlier, some errors are due to ignorance. Few U.S. libraries hold either the *Publications of the Massachusetts Medical Society* or the pamphlet version of Harlow's 1868 paper, the key follow-up report, and many of the condensed versions in medical journals did not report the changes. Other distortions and inaccuracies are the result of theoretical bias sometimes combined with ignorance.

A. The Essentially Unchanged Phineas

One theoretical position, perhaps the main one, responsible for the stories that Phineas had not been changed by the accident was that of the antilocalizationists, especially the antiphrenologists. The most extreme instance is Bigelow's story. Having read Harlow's first paper, he would at least have known that Phineas had been of temperate habits, and having seen Phineas himself over some weeks, must have at least noticed his profanity. Yet he implied that the change was inconsiderable.

The same theoretical position is behind Dalton's denial of change. Dalton published what appears to be the first discussion in the physiological literature of Gage in relation to localization in his 1859 textbook, placing it between his discussion on the insensibility of the brain to stimulation, which he coupled with an account of Flourens' experiments, and an attack on Gall's organology. In that no account of the changes had been published in the medical literature at that point, this may not have been unreasonable, but what he did in his edition of 1875 was inexcusable. There he drew on Jackson's entry in the Warren Museum *Catalogue* but omitted Jackson's summary of the changes, so that this later version again presented an essentially unchanged Gage.

An even more deliberate antilocalization bias seems to be responsible for a similarly inexcusable treatment of the unchanged Gage in editions of Kirkes's famous textbook. Before 1896 it contained nothing at all about Gage, usually finishing its section on the functions of the frontal lobes with a statement to the effect that they had none. In 1896, Gage was mentioned in the context of frontal ablations having no appreciable effect, and, despite the editors having had their attention drawn to this error, they refused to make any changes. In 1909, when they did give in, they said merely that a recent examination of the case records showed that Phineas had become useless as a supervisor, and they essentially retained that account as late as 1937.

The dominant sensory–motor physiology and psychology that had displaced phrenology by the middle of the nineteenth century provided another theoretical basis for the stories of an unchanged Phineas. Basically, what that theoretical orientation encompassed were losses of sensation and movement and their

psychological consequences. Complex behaviors like those shown by a changed Phineas, or even a normal one, were not. The problem is seen most clearly in Ferrier's short-lived attribution of an inhibitory function to the frontal lobes. He really had no evidence: the facts were "negative" and it was "not actually motor." The very term inhibitory-motor mirrors sensory–motor (which in turn derives from Marshall Hall's earlier excito-motory), and its function of delaying responses could be and was applied to anything that, not being sensory–motor, had to be "mental." Thus, Starr could lump together as "mental symptoms" showing a lack of self-control, Gage, whom he described as having become emotional, easily excitable, and irritable, and another patient, whom he described as stupid and listless. Similarly, Walter Dandy, the pioneer American neurosurgeon, in commenting on what he saw as the psychological changes following radical lobe resection, could only call them "little peculiarities of mental origin."

B. The Changed Gage as a Theoretical Prop

Some of the distortions in the stories about Phineas have come about because their tellers use Phineas to exemplify one or other aspect of their theories of brain function. Luria and Treisman provide the least distorted of these kinds of stories. Luria, who gave the frontal lobes an important role in controlling cognitive processes, describes Gage in almost completely cognitive terms. Gage's deficiency is that he lacks initiative and the ability to critically appraise internally programmed action. Treisman invokes a deficiency of long-term planning in the context of social relations and interactions and turns Phineas into a drunkard.

Bloom, Lazerson, and Hofstadter provide a more distorted story when they use Gage to exemplify their theory that the frontal lobes control the emotions generated in the limbic system. As portrayed in the film made to accompany their book, their Phineas suffers from a lack of balance between his emotions and his intellectual propensities. He is almost the kind of mumbling madman once seen on the stage who sobs with self-pity and hallucinates almost as readily as he flies into fits of aggressive rage. Weisfeld has it that Gage's problems were primarily because he was unable to appreciate the effects of his behavior on others. He links Gage to observations that monkeys with orbitofrontal damage exhibit less appropriate reactions to the ranks of their cagemates and lose their positions in their hierarchies as a consequence, and he presumes that Gage suffered the same kind of damage. However, whereas the operated monkeys become less aggressive, most other storytellers have Phineas becoming more aggressive.

VII. CONCLUSIONS

What can be learned from the case of Phineas Gage? Four things stand out, the first two about Gage himself and the last two about historical figures like him in general (and not only in the neurosciences). We can say about Gage that the uncertainties about which parts of his brain were damaged or destroyed seem destined to remain, and that the chance that we will learn much more about his behavior and history is so slight that the limited comparison we can make between his pre- and postaccident behavior also does not seem likely to change. Of historical figures like him, we can say that his case emphasizes the importance of accurate description and the need to disentangle fact from theoretical preconception almost as much as it points to the need to return to the original sources. On these last two points, it is David Ferrier who provides us with the guide we need. In explaining to Bowditch why he wanted information about Gage he said, "In investigating the reports on diseases and injuries of the brain I am constantly being amazed at the inexactitude and distortion to which they are subjected by men who have some pet theory to support."

What we know about Phineas Gage provides too meager a foundation on which to base hypotheses of the relationship between the frontal lobes and their psychological functions. Is he now fit to be remembered only by such things as the plaque on Cavendish Town Green unveiled on September 13, 1998, to commemorate the 150th anniversary of his accident? I do not believe so. Phineas Gage's lasting importance is because his was the first case to point to a relation between brain and personality functions. All of the reports we have about Gage are reproduced as facsimiles in *An Odd Kind of Fame: Stories of Phineas Gage*.

See Also the Following Articles

BRAIN ANATOMY AND NETWORKS • BRAIN DAMAGE, RECOVERY FROM • EPILEPSY • FRONTAL LOBE • MODELING BRAIN INJURY/TRAUMA • NEUROPSYCHOLOGICAL ASSESSMENT • PREFRONTAL CORTEX

Suggested Reading

Barker, F. G., II. (1993). Treatment of open brain wounds in America, 1810–1880: A survey. *J. Neurosurg.* **78,** 364A.

Barker, F. G., II. (1995). Phineas among the phrenologists: The American crowbar case and nineteenth century theories of cerebral localization. *J. Neurosurg.* **82,** 672–682.

Bigelow, H. J. (1850a). Dr. Harlow's case of recovery from the passage of an iron bar through the head. *Am. J. Med. Sci.* **20,** 13–22.

Bigelow, H. J. (1850b). *Dr. Harlow's Case of Recovery from the Passage of an Iron Bar through the Head.* Collins, Philadelphia, PA.

Damasio, H., Grabowski, T., Frank, R., Galaburda, A. M., and Damasio, A. R. (1994). The return of Phineas Gage: Clues about the brain from the skull of a famous patient. *Science* **264,** 1102–1105.

Harlow, J. M. (1848). Passage of an iron rod through the head. *Boston Med. Surg. J.* **39,** 389–393.

Harlow, J. M. (1849). Letter in "Medical miscellany." *Boston Med. Surg. J.* **39,** 506–507.

Harlow, J. M. (1868). Recovery from the passage of an iron bar through the head. *Pub. Mass. Med. Soc.* **2,** 327–347.

Harlow, J. M. (1869). *Recovery from the Passage of an Iron Bar through the Head.* Clapp, Boston, MA.

Jackson, J. B. S. (1849). *Medical Cases Vol. 4, Case Numbers 1358–1929, pp. 720, 610.* Countway Library Harvard University, Boston MA.

Jackson, J. B. S. (1870). *A Descriptive Catalogue of the Warren Anatomical Museum.* Williams, Boston, MA.

Macmillan, M. (2000). *An Odd Kind of Fame: Stories of Phineas Gage.* MIT Press, Cambridge, MA.

Tyler, K. L., and Tyler, H. R. (1982). A "Yankee Invention:" The celebrated American crowbar case. *Neurology* **32,** A191.

Pick's Disease and Frontotemporal Dementia

ANDREW KERTESZ
St. Joseph's Hospital, University of Western Ontario

DAVID G. MUNOZ
Hospital Universitario "Doce de Octubre," Madrid, Spain

GLOSSARY

corticobasal degeneration (CBD) The extrapyramidal variety of PiD. It is clinically defined as unilateral extrapyramidal symptoms, apraxia, and the alien hand syndrome, but many of these patients develop features of FTD and PPA and they often present with cognitive syndromes. Therefore, the pathological entity of CBD and the clinical syndrome of CBDs are also part of the PC.

frontotemporal dementia (FTD) Clinical PiD or PC. It is also used for the apathy–disinhibition presentation or the pathology without Pick bodies.

FTDP-17 Frontotemporal dementia and parkinsonism linked to chromosome 17. Most of these families have τ mutations and a range of pathological and clinical phenotypes that are similar to sporadic cases of PC.

motor neuron type of FLD This was initially described as a separate entity with ubiquitin-positive, τ-negative inclusions, but there is considerable overlap between these cases and other members of the PC.

Pick bodies Round, argyrophilic, compact inclusions in the dentate gyrus and the neocortex. At one time they were considered the defining pathology of PiD, but there are a variety of inclusions in PC and sometimes there are no inclusions in the clinical syndrome (see Section III).

Pick cells Ballooned neurons. This is a feature of all varieties of the PC and was originally described with PiD but later also as a cardinal feature of CBD.

Pick complex (PC) Clinical PiD encompassing all the syndromes of frontotemporal dementia, primary progressive aphasia, corticobasal degeneration, and frontal lobe dementia with motor neuron disease and all of the pathological varieties, including PiD, CBD, dementia lacking distinctive histology, etc.

Pick's disease (PiD) Clinical frontotemporal dementia or pathology with Pick bodies.

primary progressive aphasia (PPA) This presenting syndrome is also part of the Pick complex or clinical Pick's disease. It has a variety of pathologies just like FTD. By definition, the patient exhibits slowly progressive aphasia before anything else develops.

semantic dementia Semantic aphasia or a fluent type of transcortical sensory aphasia in which the patient has difficulty with comprehension and naming, especially comprehending nouns, with well preserved repetitions, spontaneous speech, and syntax. The loss of meaning extends to visually presented stimuli; therefore, the term semantic dementia is applied. However, because these patients are not truly demented at the beginning and the condition is more related to PPA, semantic aphasia would be a preferable term.

Pick's disease (PiD) has been used to describe either the clinical syndrome of frontotemporal dementia (clinical

Pick's disease), or as a pathological description of frontotemporal atrophy with silver staining, globular inclusion (Pick bodies), swollen neurons (Pick cells), superficial cortical spongiosis, neuronal loss, and gliosis. However, only a small percentage of cases, about 25%, with clinical PiD will have the typical inclusions. This has created a nosological dichotomy, but more recent and advances in neuroimaging, histochemistry and genetics have highlighted the importance of frontotemporal dementia.

I. INTRODUCTION

Arnold Pick described the clinical picture associated with focal atrophy in the frontal and temporal lobes around the turn of the twentieth century. Pick's initial description of a patient with progressive aphasia and behavioral disturbances and cases of frontal lobe dementia included only a gross examination without any microscopic data, but the clinical descriptions and their relationship to focal atrophy form the basis of the syndrome. Gans suggested the eponymic term and considered a predilection for the phylogenetically younger frontal and temporal lobes in the etiology. A reexamination of a series of cases of Pick and others emphasized the histological picture particularly the "pick bodies" described by Alois Alzheimer. However, it soon became apparent that cases of clinical PiD with frontal and temporal lobe symptomatology may not show the typical "Pick bodies" on autopsy. Publications of PiD were often based on post mortem examination, and the clinical features were either variable depending on the locus of disease or described incompletely because of the retrospective nature of these studies. This gave rise to the notion that PiD is difficult to diagnose *in vivo*. The restriction of the pathological diagnosis to the finding of typical Pick bodies created the second misconception that PiD is rare.

Whereas PiD continued to be diagnosed clinically and pathologically, the emphasis on the pathological diagnosis impaired the clinical recognition. Nevertheless, interest in the entity continued, mainly in Europe. Constantinidis *et al.* classified PiD as (A) with Pick bodies, (B) only with swollen neurons, and (C) only gliosis. They felt that "in spite of the dissimilarities between these forms, considering the absence of sufficient knowledge about pathogenesis, it seems prudent at present to maintain the uniqueness of Pick's entity." They thought that the clinical differences between these forms were not related to the nature of histological alterations but rather to the temporal or frontal prominence.

With the development of neuroimaging, frontotemporal atrophy was demonstrated with increasing frequency *in vivo*, first with air studies, then with CT scans, and with MRI scan and SPECT more recently. The clinical diagnosis of PiD is based on recognizing complex frontal and temporal symptomatology: mainly disinhibition dementia and aphasia with relatively preserved memory and visuospatial function (in contrast to Alzheimer's disease, AD) in younger individuals supported by focal atrophy on imaging and normal EEG. However, instead of shifting the diagnosis of PiD back to the clinic, new labels were applied to the syndrome, such as frontal lobe dementia (FLD), primary progressive aphasia (PPA), and frontotemporal degeneration (FTD), as new and separate entities while reserving the diagnosis of PiD for increasingly restricted histological criteria. Further development of histochemistry contributed to the fractionation of the pathological variations neglecting or underemphasizing the similarities. The glossary of the proliferating terminology is summarized in order of historical appearance (Table I).

Two European groups described frontal lobe dementia (FLD) as a distinct entity and contrasted the clinical features with those of AD. They estimated its relative incidence to be 15–20% of degenerative dementias. Both groups recognized that, even though some of the cases had Pick bodies and the majority did not, the clinical syndrome was the same, regardless of the pathological variant. They called the pathology without Pick bodies "frontal lobe dementia type," consisting of neuronal loss and gliosis in the frontal cortex with or without spongiform changes or ballooned neurons. Shortly thereafter, a similar clinicopathological picture was called "dementia lacking distinctive histology (DLDH)". More recently, occasional association with motor neuron disease (MND) has been increasingly recognized. Ubiquitin-positive, τ-negative inclusion bodies in the dentate gyrus and cortex were described as a marker of this syndrome.

The Lund and Manchester groups described dementia of the frontal lobe type changed the terminology to frontotemporal degeneration (FTD), and later to frontotemporal lobar degeneration (FTLD), and summarized the consensus criteria for diagnosis. The term frontotemporal degeneration or frontotemporal dementia does not include the frequent subcortical involvement, parietal pathology, and extrapyramidal symptomatology. Furthermore, it does not distinguish between the behavioral presentation of FLD and

Table I
Pick Complex Related Terminology

1. Circumscribed cerebral atrophy
2. Pick's disease (PiD)
3. Lobar atrophy
4. Progressive subcortical gliosis (PSG)
5. Corticodentatonigral degeneration
6. Generalized Pick's disease
7. Frontal lobe dementia (FLD)
8. Primary progressive aphasia (PPA)
9. Corticobasal degeneration (CBD)
10. Dementia lacking distinctive histology (DLDH)
11. Semantic dementia
12. Frontal lobe dementia with motor neuron disease
13. Frontotemporal dementia (FTD)
14. Nonspecific familial dementia
15. Atypical presenile dementia
16. Spongiform encephalopathy of long duration
17. Frontotemporal lobar degeneration (FTLD)
18. Hereditary dysphasic dementia (HDD)
19. Disinhibition–dementia–parkinsonism–amyotrophy complex (DDPA)
20. Pallidopontonigral degeneration (PPND)
21. Hereditary dysphasic disinhibition dementia (HDDD)
22. Multiple system tauopathy with presenile dementia (MSTD)
23. Frontotemporal dementia and Parkinsonism linked to chromosome-17 (FTDP-17)

aphasic presentation of PPA, which is one of the most valuable contributions of the descriptions of the clinical picture in these conditions. We proposed the term Pick complex (PC) to encompass this family of related disorders. The glossary of these is summarized in order of their historical appearance (Table I).

II. THE CLINICAL PRESENTATIONS OF PICK'S DISEASE

A. The Apathetic–Disinhibition Syndrome of Frontal Lobe Dementia (FLD)

The behavioral and personality changes of the "frontal lobe syndrome" often begin with apathy and disinterest, which may be mistaken for depression. On the other hand, symptoms of disinhibition may suggest a manic psychosis or a personality disorder. Cases of PiD–FTD with behavioral manifestations, therefore, are more likely to be presented to a psychiatrist than to a neurologist.

Socially inappropriate comments, rude, childish, or selfish behavior appear as a significant change of personality. Patients can be strikingly indifferent and unemotional. Stubbornness and insistence on routines resemble obsessive/compulsive behavior. When they are opposed, irritability and aggression appear. Poor judgment in spending, driving, social interaction, and bizarre behavior, such as hoarding or pilfering, can be very disruptive and lead to criminal charges.

Over-eating, cravings of sweets, or other food fads may occur. Table manners deteriorate. In the later stages, patients may touch everything and compulsively put objects in their mouth. Perseverative humming, clapping, and decreased language appear. Incontinence may come relatively early.

Some of the behavioral syndrome of PiD resembled the so-called Kluver–Bucy syndrome, which is produced in monkeys by bilateral ablation of the temporal neocortex and the amygdala and can be seen in humans after encephalitis. It consists of hyperorality, hypersexuality, and compulsive touching or "hypermetamorphosis."

Further distinctions have been made between clinical behavioral syndromes, such as the "apathetic, disinhibited, and stereotypic." The disinhibited type mainly involves the orbitofrontal region of the frontal lobes. In the apathetic type, the dorsal lateral convexity appears to be more affected. The stereotypic type appears to have more extrapyramidal involvement and striatal pathology. These distinctions tend to become blurred as the disease progresses, and not all patients can be easily categorized into these subtypes. Severe disinhibition syndrome may be seen with nondominant temporal lobe atrophy. It is difficult to separate the contributions of the temporal and frontal lobes in these cases.

The highly complex symptomatology requires a pattern recognition. Early cases often remain puzzling for first-time observers. Even dementia authorities claimed at one time or another that AD and PiD are similar clinically and cannot be distinguished reliably by clinical or neuropsychological methods. On the contrary, the pattern of disengagement–disinhibition is characteristic, once it is recognized. Neuropsychological testing often reveals relatively preserved episodic memory and visuospatial function, but a caregiver providing history or response to a questionnaire, such as the *Frontal Behavioral Inventory*, is even more useful. Neuroimaging, especially MRI, completes the differential diagnosis.

B. Primary Progressive Aphasia (PPA)

The dominant temporal lobe variety of PiD presenting with progressive aphasia has been described quite early, and these descriptions are similar to those of primary progressive aphasia (PPA) appearing later. Many cases had histology characterized by gliosis, neuronal loss, and layer II and III spongiosis in the cortex, which was initially considered specific and later described as identical to FLD. Some had classical PiD with Pick bodies and some subcortical involvement with neuronal achromasia similar to CBD.

Several varieties of PPA have been described: (1) the more common nonfluent variety leads to mutism (the majority of published cases); (2) the aphemic variety initially presents with verbal apraxia and stuttering; and (3) semantic aphasia (dementia) in which speech output remains preserved whereas the meaning of objects as tested by naming and comprehension appears to be lost. PPA is defined as a progressive language impairment without dementia for at least 2 years, although it is recognized that other modalities are affected subsequently, particularly behavior changes suggesting frontal deficit. At times extrapyramidal complications or MND supervenes. The variety associated with MND tends to be more rapidly progressive. PPA has been frequently described with progressive apraxia. FTD and CBD, in turn, also have a progressive language disturbance. Eventually most cases of FTD followed long enough develop progressive language loss and mutism. Therefore, not only the pathology but also the symptomatology of PPA overlaps among the entities belonging to the Pick complex.

Those who are more interested in behavioral disorders are less likely to report the language disorder in great detail, often describing the end stage of the progressive language disorder as mutism. Neurologists, on the other hand, may see the primarily aphasic disorder more often, as focal symptoms may be referred for investigation of a slowly growing tumor or even a stroke.

C. Corticobasal Degeneration (CBD)

There have been several case descriptions of PiD in which the patients had prominent extrapyramidal features. This became known as the "Akelaitis" variety of PiD. Sometimes unilateral rigidity and parkinsonism were the first symptoms to attract attention. It was recognized that subcortical changes occur in PiD, even without extrapyramidal symptomatology. Constantinidis *et al.* described extrapyramidal involvement particularly in group B Pick patients and in FLD patients. Changes in the basal ganglia, especially in the striatum and the substantia nigra, in addition to cortical pathology occurred in the majority of 30 cases in one review.

The authors initially describing corticodentatonigral degeneration recognized the similarity of the pathology to PiD, and this was subsequently confirmed by several investigators who also contributed further clinical details. The condition was relabeled corticobasal or corticobasal ganglionic degeneration. Most of the literature concerning this condition acknowledges the clinical and pathological overlap between CBD and PiD. CBD suffers from a dichotomy similar to that of PiD, in that the pathological and clinical descriptions do not fully match. There are some case reports describing patients presented clinically as CBD, as defined by unilateral rigidity, apraxia, and alien hand syndrome, but who have the pathological findings of PiD with Pick bodies. Other cases pathologically typical of CBD have a frontal type of dementia without extrapyramidal features. Typical CBD pathology can be seen with a clinical picture of PPA. We suggested that the clinical syndrome of prominent apraxia, unilateral extrapyramidal syndrome, and alien hand phenomenon should be designated as corticobasal degeneration syndrome (CBDS) and CBD as a pathological description. Progressive supranuclear palsy (PSP) has a great deal of clinical, biochemical, pathological, and genetic overlap with CBD and CBDs and can be considered along with CBD and CBDS as part of the Pick complex.

III. NEUROPATHOLOGY

The key feature of all forms of Pick complex is a gross pattern of focal atrophy that is initially often asymmetrical (Fig. 1). Combined frontotemporal atrophy is most common, followed by predominantly frontal, temporal, and finally parietal atrophy in this order. Involvement of an area does not allow predictions about the degree of involvement of another, in contrast to the sequential involvement of cortical areas in AD. The introduction of the Gallyas silver stain and the ubiquitin and synuclein immunohistological techniques has revealed several unsuspected lesions and allowed the separation of others.

The characteristic lesions affect neurons and glial cells. There is a common underlying theme of neuronal

Figure 1 MRI of a patient with Pick's disease. The arrow shows severe frontotemporal atrophy bilaterally. The posterior half of the brain is normal in appearance. The frontal horns and ventricles are also much enlarged.

loss, gliosis, and superficial linear spongiosis in affected cortical areas in all variations. All can show Pick cells or ballooned neurons, so-called because of the swollen, homogeneously pink appearance of the cytoplasm on routine Hematoxyllin and Eosin (H&E) stains. These cells regularly express Aβ crystallin and phosphorylated neurofilament epitopes, in contrast with the absence or at most weak and scarce immunoreactivity for τ. The consistent location of these cells in the deep layers of the cortex indicates that they likely are not the precursors of the other neuronal inclusions.

The presence of τ immunoreactivity in oligodendrocytes, demonstrated on Gallyas stains (coiled bodies), is also a common feature to all varieties, although their abundance varies. In our series, dementia lacking distinctive histopathology and dementia with ITSNU are the most common variants, followed by CBD, dementia with Pick bodies, and basophilic inclusion body disease contributing to the underlying spectrum of pathology of Pick complex.

A. Pick Body Dementia

Pick bodies are large, round inclusions located in the neuronal cytoplasm (Fig. 2). Barely visible on H&E, they are best demonstrated with traditional silver stains, such as Bodian or Bielchowsky, but do not stain with the Gallyas method. There is considerable heterogeneity in the composition of Pick bodies, not only between cases but even within a single case, as shown in the variable labeling obtained with τ antibodies and antibodies to the synaptic protein chromogranin A. Similarly, electron microscopy has revealed bodies made up of 15-nm straight fibrils, long-period twisted, and even paired helical filaments and neurofilaments. Neocortical Pick bodies are preferentially located in small neurons, as well as in the neurons in the granular cell layer of the dentate gyrus. Their presence in the latter location is pathognomonic of dementia with Pick bodies, whereas traditional silver stains and τ antibodies will identify scattered Pick bodies in the neocortex in CBD as well.

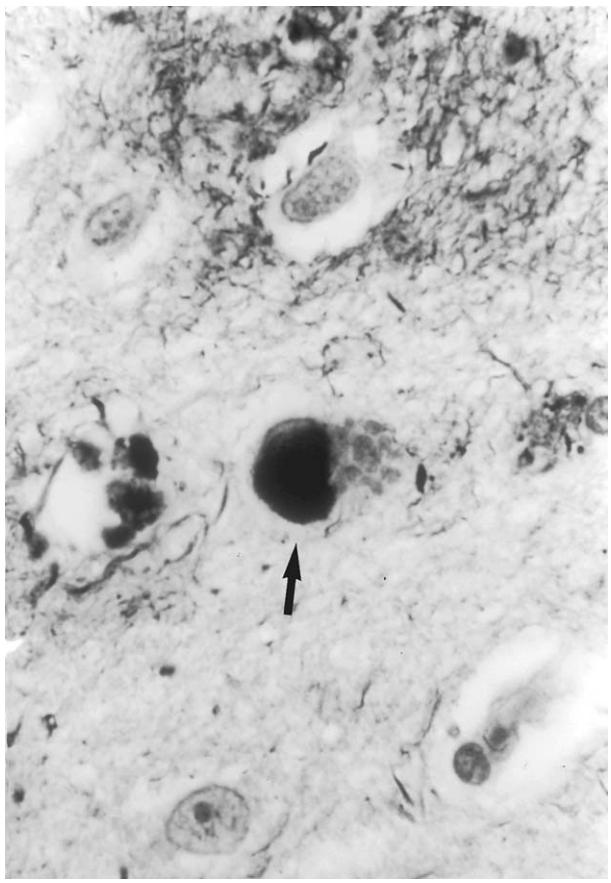


Figure 2 Bielchowsky silver stain shows (arrow) round, compact, neuronal inclusion or Pick body on autopsy.

The role of inflammatory mechanisms in PiD is not well-understood. Complementary proteins and inhibitors are detected in the neuronal cytoplasm, suggesting that complementary activation is interrupted prior to reaching completion and causing neuronal lysis. Additionally, Pick-body-bearing neurons, but not Pick cells, are surrounded by activated microglia cells and T-lymphocytes.

B. Corticobasal Degeneration (CBD)

Since the original pathological description, changes in our understanding of the pathological substrate have been substantial. The neuronal inclusions are best seen in Gallyas stains. In the neocortex, they seem to involve the same neurons as dementia with Pick bodies and adopt a variety of forms, including round lesions like Pick bodies but more characteristically ringlike tangles around the nucleus. CBD inclusions are made up of straight 15-nm fibrils, as in the common form of Pick bodies. Oligodendroglial τ expression and argyrophilic inclusions are substantial, but the pathognomonic glial inclusion occurs in astrocytes, which accumulate aberrantly phosphorylated τ protein in curlicue processes emanating raylike from the cell body, giving rise to the so-called glial plaques that are not associated with amyloid. The neuronal inclusions in the substantia nigra neurons, called corticobasal bodies in the original publications, are now recognized as no more than globose neurofibrillary tangles. They are part of the extensive subcortical involvement that usually accompanies frontal, temporal, or central cortical atrophy in CBD.

Although the absence of immunoreactivity with ubiquitin has been proposed as a method to distinguish the inclusions in CBD from those in dementia with Pick bodies, we have not found this method to be useful because ubiquitin immunoreactivity is often extremely weak in Pick bodies. CBD inclusions lack the chromogranin A immunoreactivity of Pick bodies. Pick bodies' absence of staining in the Gallyas technique may be a more practical criterion of differentiation.

C. Dementia Lacking Distinctive Histopathology (DLDH)

The uninspiring name may be preferable to such confusing terms as frontal lobe dementia or frontotemporal degeneration, which may refer to a clinical presentation, to the other varieties in Pick complex, or to this specific variety. It corresponds to Pick's disease types B and C of Constantinidis. The variation in the distribution of cortical and subcortical atrophy is manifested in clinical syndromes of frontal dementia, aphasia, and extrapyramidal movement disorders.

Histological examination reveals characteristic neuronal loss and gliosis, superficial linear spongiosis, and a variable degree of atrophy and gliosis of the white matter. The presence of ballooned neurons in the deep layers of the cortex and scattered oligodendrocytes expressing τ does not alter the diagnosis. Occasional cases demonstrating diffuse τ immunoreactivity in the neuronal cytoplasm without the formation of Pick bodies or neurofibrillary tangles, as well as the demonstration of abnormal τ proteins, emphasize the uncertainty concerning the precise margins of this entity.

D. Dementia with ITSNU

Inclusion, tau and synuclein negative, ubiquitinated (ITSNU) defines this entity, which otherwise blends with dementia lacking distinctive histopathology. The inclusions bear a resemblance to Pick bodies in size and shape, but so far are revealed only by ubiquitin immunostains. They occur in the dentate gyrus of the hippocampus and the scattered neurons in the atrophic frontal or temporal cortex. The ultrastructural appearance of ITSNU, with fibrils of variable diameter associated with granular material, is reminiscent of Lewy bodies from which they can be distinguished by the absence not only of eosinophilia but also of α-synuclein immunoreactivity. In addition, motor neurons in the brain stem and spinal cord show the skein-like ubiquitinated inclusions characteristic of amyotrophic lateral sclerosis; thus, the term "dementia of motor neuron disease type" has been applied to this condition. However, many of these patients have no record of motor weakness in the course of their disease, and as many as 25% of nondemented patients with amyotrophic lateral sclerosis also show cortical ITSNU. Vertically oriented, ubiquitinated, enlarged neurites can also be seen in the neocortex, and cell loss in the substantia nigra is consistently found. The suggestion of considering dementia with ITSNU as a subset of dementia lacking specific histopathology is not historically unreasonable, because several series of the latter variant were described without the benefit of ubiquitin immunohistochemistry.

E. Basophilic Inclusion Body Disease

This variant was described as the generalized variant of PiD, whereas other authors have reported it as a form of juvenile or adult motor neuron disease. The characteristic inclusion is round in neocortical, hippocampal, neostriatal, subthalamic, and nigral neurons. However, it adopts irregular shapes in the motor neuron. The basophilia is due to the content of RNA, derived from the rough endoplasmic reticulum. The inclusions express ubiquitin immunoreactivity but, unlike Lewy bodies, are not labeled by antibodies to α-synuclein. The ultrastructural appearance is remarkably similar to ITSNU. Although the flamboyant look of the inclusions contrasts with the drab appearance of ITSNU, we have observed two patients of the same family, each showing one form consistently throughout the brain. It is thus likely that basophilic inclusions represent ITSNU with additional RNA.

F. Progressive Subcortical Gliosis

The claim that progressive subcortical gliosis represents a distinct condition is based on the discrepancy between the severe loss of volume and gliosis of the white matter and the relative sparing of the cortex. However, careful reading of modern series reveals no substantial differences with DLDH, in which white matter degeneration can be severe. The variant has been given prominence by a report of a family in which this pathological diagnosis was linked to chromosome 3 and in which abnormal prion proteins were identified by immunohistochemistry and Western blots in the absence of mutations of the prion protein gene.

In conclusion, it seems likely that DLDH, dementia with ITSNU, basophilic inclusion body disease, and progressive subcortical gliosis are manifestations of the same pathological process. The other two pathological substrates of the Pick complex, Pick body dementia and CBD, are distinct, but their lesions show much greater overlap than review articles suggest. For example, most cases of Pick body dementia show neurofibrillary tangles in the substantia nigra indistinguishable from corticobasal bodies, and neocortical Pick bodies are a feature of most cases of CBD. The marked case to case variation in each subgroup of Pick complex (ignored in review articles) and the multiple pathological expressions of chromosome-17-linked dementia further argue for considering the varieties described here as diverse of appearances of shared pathological processes, rather than as unrelated diseases.

IV. BIOCHEMISTRY

Although the etiology and biochemical basis of PiD remain unknown, it has become clear that alterations of the cytoskeletal protein τ play a fundamental role, as they do in several other degenerative disorders. τ proteins are involved in the modulation of microtuble assembly and, thus, in the regulation of axonal transport and neuronal plasticity. Six different τ isoforms are generated in the adult brain by splicing a single gene, located in chromosome 17.

Pathological τ proteins (PTPs), formed by aberrant phosphorylation and other modifications, can be recovered from the brains of patients with AD, PiD, and progressive supranuclear palsy (PSP). Western blots detect distinct profiles of PTP: a triplet at 55, 64, and 69 kDa in AD; a doublet at 55 and 64 kDa in PiD; and a different doublet at 64 and 69 kDa in CBD and PSP. Furthermore, site-specific monoclonal antibodies have shown that, although PTPs in PiD and AD share similar phosphorylated residues, others are distinct such as serine 262, which is phosphorylated in AD but not in PiD. Moreover, PTPs in PiD do not incorporate exon 10, unlike those in AD. The combined use of Western blot of PTPs and antibodies specific for each of the six isoforms of τ has revealed that, whereas all six isoforms are present in AD PTPs, only the three isoforms lacking the domain coded by exon 10 (3-repeat τ) are found in PiD and only the three isoforms incorporating the domain coded by exon 10 (4-repeat τ) are present in CBD. PTPs have somatoaxonal distribution in PiD and somatodendritic distribution in AD.

The demonstration of PTP in DLDH, even in the absence of argyrophilic or τ-immunoreactive inclusions, supports the concept of Pick complex by suggesting that alterations in τ are common to all variants. More recently, cases of DLDH were found to be deficient in normal τ on Western blot.

A. Genetics, Chromosome-17-Linked FTD

It was recognized that PiD at times occurred in families. Wilhelmsen *et al.* have linked an autosomal-dominant family with frontotemporal dementia to chromosome 17. Most patients in this family presented with behavioral disinhibition and subsequently developed a language disturbance, parkinsonism, and amyotrophy; the syndrome was called disinhibition dementia, parkinsonism, amyotrophy complex (DDPAC). Report of a large family with PiD, in

which 25 of 51 examined members were affected with mostly behavioral presentation, was published in Holland. Subsequently, this family was found to have genetic linkage to chromosome 17. Descriptions of what could be classified as familial PiD continue, but there is a tendency to reclassify these because of the lack of Pick bodies. PiD generally has a presenile onset before age 65, in contrast to the majority of AD patients. Familial cases tend to have an even earlier onset in the 40s or 50s.

Other families received various designations, such as pallidopontonigral degeneration (PPND), hereditary dysphasic disinhibition dementia (HDDD2), and multiple system tauopathy with presenile dementia (MSTD). Other families are described by their place of origin such as the Dutch, Australian, Duke Seattle, and Karolinska families. A consensus conference on chromosome-17-linked dementia decided on using the acronym FTDP-17. There is a family with progressive subcortical gliosis (PSG) with probable linkage to chromosome 3.

Several mutations were found in τ in FTD families linked to chromosome 17. A Val337 Met change has been found in exon 12 of the τ gene in the Seattle A family. Other τ mutations have been found in common with PSP in the intron between exons 9 and 10 in association with PSP. Several additional families with P301L mutations on exon 10 have been described with a variety of clinical manifestations, all compatible with Pick's disease. τ polymorphisms, but not mutations, so far have been found in PSP.

The clinical features of chromosome-17-linked dementia are very similar to the sporadic cases of Pick complex and PiD discussed earlier, even though the connection is not always fully recognized or explicitly stated. There is a tendency to report each of these families as being distinct. Disinhibition syndrome and behavioral disturbances are most common. Language difficulties and extrapyramidal symptoms are also frequent. Some patients had signs of motor neuron disease. Prominent psychosis similar to schizophrenia has been reported, which may represent a distinction, but the descriptions are not sufficiently detailed to allow certainty in this regard.

The neuropathology of FTDP-17 is similar to the range of pathological findings described in sporadic Pick complex. τ-positive silver staining neuronal inclusions were numerous in the neocortex, basal ganglia, hypothalamus, and midbrain in some of the families reported. Some of these autopsied cases also had glial cell argyrophilic and τ positive deposits. The most detailed neuropathological studies have been reported for the DDPAC and Seattle family A. All of the pathological reports indicate atrophy of the frontal and temporal lobes of varying degrees and the of parietal lobes to a lesser extent, in addition to atrophy of the basal ganglia such as the caudate, putamen, globus pallidus, amygdala, and hypothalamus. A family with typical Pick bodies has now been reported to have a τ mutation. One of the chromosome-17-linked families had ubiquitin-positive, τ-negative neuronal inclusions, but some τ was found in the glia. There seems to be a whole range of τ deposits in FTDP-17 families from the very severe in the MSTD cases to the very mild or none mentioned earlier. On electromicroscopy, neurofilaments appeared similar to those of AD in the Seattle family, and unique in the MSTD family, suggesting a heterogeneity of τ alterations in the cytoskeleton in FTDP-17.

V. CONCLUSIONS

We suggested the term *Pick complex* to avoid the confusion that continues to surround the terms PiD and FTD. PC designates both the pathological and the clinical overlap between the variations. It avoids the restriction of pathology and clinical symptomatology to the frontotemporal cortex and acknowledges the relationship to PiD. PC is a unifying concept of overlapping clinical syndromes and neuropathological findings underlying the commonalities rather than differences between them.

If one considers all of the cases of FLD, PPA with and without MND, and CBD as part of PC, the entity becomes much less of a rarity. According to some estimates including all pathological variants, the incidence of FTD or clinical PiD may be as high as 20% of degenerative dementias, and PPA reports may represent another 10% even considering the substantial overlap with FTD. This number would be further increased by the inclusion of CBD cases. The incidence matches or surpasses that of pure vascular dementia in our experience. The ratio of PC to AD may turn out to be 1:4 or about 25% of degenerative dementias rather than the estimates of PiD based on autopsy material using restrictive histological criteria. This ratio may be even higher if only patients whose illness starts under age 65 are considered. Apart from a survey of Dutch specialists, which seems to underestimate the incidence probably due to insufficient diagnostic experience, true population-based epidemiological studies are lacking. A selection bias may be playing a role in centers with an interest in the disease.

An increasing number of clinicians and pathologists believe that many, if not all, of the preceding labels describe diseases that clinically and most probably biologically are related to PiD. However, as new immunohistological techniques are applied to varieties of PC, the fractionation of the entity continues. It is difficult to resist the urge to describe a new disease when a new stain or an interesting biochemical marker is discovered. Nevertheless, the clinical, pathological, biochemical, and genetic evidence for the cohesion of the syndrome is compelling, and excessive fractionation may be misleading.

See Also the Following Articles

ALZHEIMER'S DISEASE, NEUROPSYCHOLOGY OF • APHASIA • DEMENTIA • MOTOR NEURON DISEASE

Suggested Reading

Constantinidis, J., Richard, J., and Tissot, R. (1974). Pick's disease. Histological and clinical correlations. *Eur. Neurol.* **11,** 208–217.

Feany, M. B., Mattiace, L. A., and Dickson, D. W. (1996). Neuropathologic overlap of progressive supranuclear palsy, Pick's disease and corticobasal degeneration. *J. Neuropathol. Exp. Neurol.* **55,** 53–67.

Hutton, M., Lendon, C. L., Rizzu, P., Baker, M., Froelich, S., Houlden, H., *et al*. (1998). Association of missense and 5′-splice-site mutations in tau with the inherited dementia FTDP-17. *Nature* **393,** 702–705.

Kertesz, A., and Munoz, D. (1998). *Pick's Disease and Pick Complex*. Wiley-Liss, New York.

Mesulam, M. M. (1987). Primary progressive aphasia—Differentiation from Alzheimer's disease. *Ann. Neurol.* **22,** 533–534.

Pick, A. (1892). Über die Beziehungen der senilen Hirnatrophie zur Aphasie. *Prager Med. Wochenschriff* **17,** 165–167.

Rebeiz, J. J., Kolodny, E. H., and Richardson, E. P., Jr. (1968). Corticodentatonigral degeneration with neuronal achromasia. *Arch. Neurol.* **18,** 20–33.

Snowden, J. S., Neary, D., and Mann, D. M. A. (1996). *Frontotemporal Lobar Degeneration: Frontotemporal dementia. Progresssive Aphasia. Semantic Dementia*. Churchill-Livingston, London.

The Lund and Manchester Groups. (1994). Clinical and neuropathological criteria for frontotemporal dementia. *J. Neurol. Neurosurg. Psychiatry* **57,** 416–418.

Wilhelmsen, K. C., Lynch, T., Pavlou, E., Higgins, M., and Nygaard, T. B. (1994). Localization of disinhibition–dementia–parkinsonism–amyotrophy complex to 17q21-22. *Am. J. Human Genet.* **55,** 1159–1165.

Prefrontal Cortex

ROBERT J. MORECRAFT* and EDWARD H. YETERIAN†
University of South Dakota and Colby College†*

GLOSSARY

association cortex Extremely well-developed and interactive regions of the cerebral cortex involved in the integration of highly processed sensory information and the formation of perception and cognition. The anatomical construct of association cortex is formed in part through local feedforward and feedback projections, as well as widespread projections preferentially interconnecting association areas of the frontal, parietal, temporal, occipital, and limbic cortices.

cognitive functions The mental activities associated with the execution of thinking, learning, and memory requiring computational strategies to acquire knowledge and effectively solve complex problems or situations.

frontal lobe Major subdivision of the cerebral cortex located anterior to the central sulcus, including the medial and ventral surfaces of the cerebrum. The frontal lobe is formed by various components of the primary motor cortex, premotor cortices, frontal eye fields, and prefrontal cortex.

prefrontal cortex The most rostral portion of the frontal lobe, characterized as being involved in cognitive and emotional brain functions.

working memory The ability to plan and sequence an adaptive behavior utilizing on-line or actively stored and processed information from both the external and internal environments.

The prefrontal cortex is a major subdivision of the cerebral cortex, which plays a critical role in mediating complex processes such as attention, planning, decision making, emotion, and personality (Fig. 1). The prefrontal cortex, in part, provides us with our power of intellect, reasoning, and rationality, distinguishing elements of human cognition and behavior. This article will present a general overview of basic structural, functional, and clinical features that characterize the prefrontal cortex based upon an extensive body of information developed from studies of humans and nonhuman primates.

I. ANATOMICAL ORGANIZATION OF THE PREFRONTAL CORTEX

A. Surface Features and Major Subdivisions

The prefrontal cortex is located on the medial, lateral, and orbital surfaces of the most anterior portion of the frontal lobe (Figs. 1 and 2). It has been estimated that prefrontal cortex occupies approximately one-third of the entire human cerebral cortex. Unique cytoarchitectonic, connectional, electrophysiological, behavioral, and clinical features distinguish the prefrontal cortex on each surface. This complex and distributed organization reflects the characterization of this brain area as a protean and heterogeneous entity, dedicated to sustaining rapid computations required to accomplish a wide range of mental activities.

1. Lateral Organization

On the lateral convexity, the frontal lobe lies rostral to the central sulcus, extending to and including the frontal pole. From caudal to rostral, the frontal lobe is subdivided into primary motor cortex, premotor cortex, frontal eye fields, and lateral prefrontal cortex

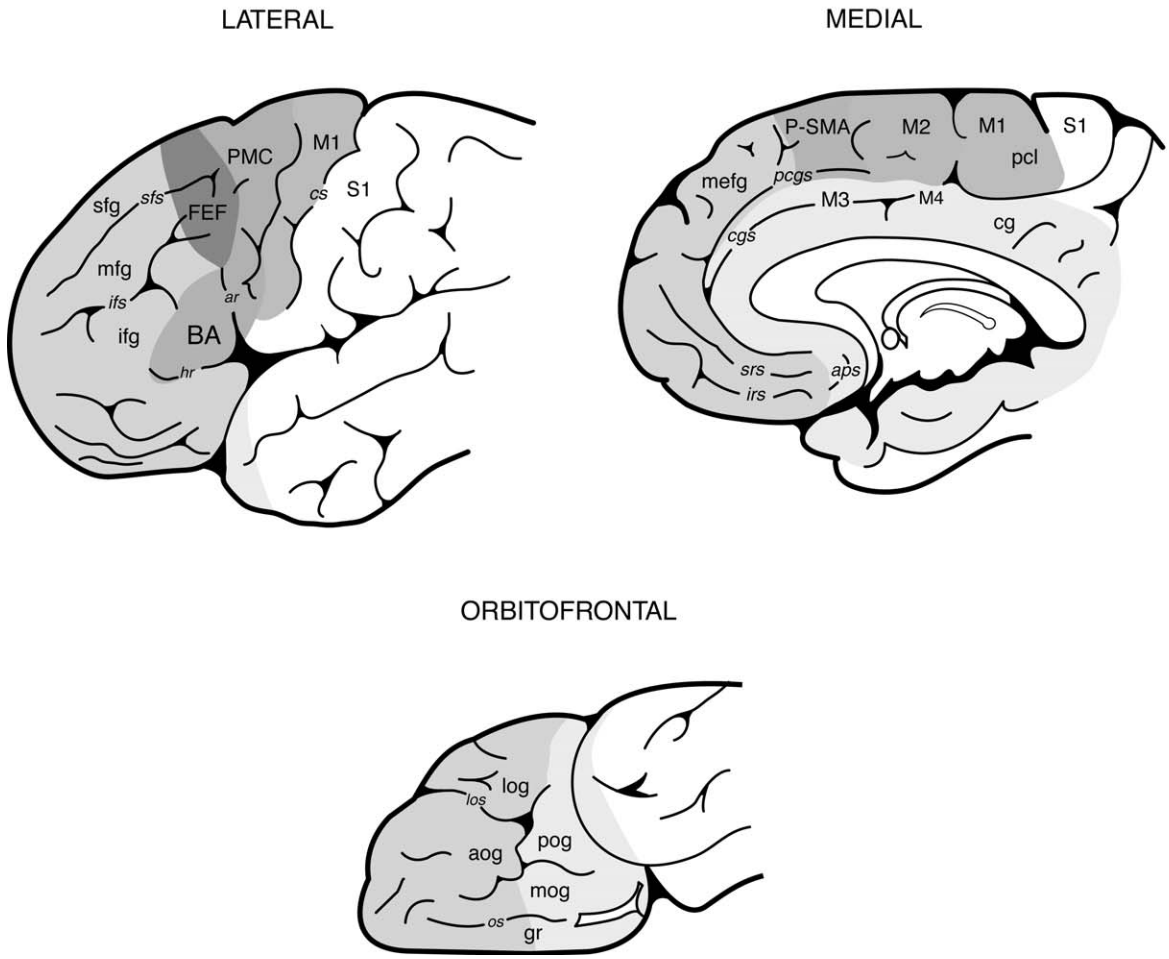


Figure 1 Schematic diagrams of the lateral, medial, and orbital surfaces of the human frontal region illustrating the general anatomical and functional organization. Granular prefrontal cortex is located rostrally and darkly shaded. Limbic cortex is lightly shaded. All other divisions are specifically identified. aog, anterior orbital gyrus; *ar*, ascending ramus; *aps*, anterior parolfactory sulcus; BA, Broca's area; cg, cingulate gyrus; *cgs*, cingulate sulcus; *cs*, central sulcus; FEF, frontal eye fields; gr, gyrus rectus; *hr*, horizontal ramus; ifg, inferior frontal gyrus; *ifs*, inferior frontal sulcus; *irs*, inferior rostral sulcus; log, lateral orbital gyrus; *los*, lateral orbital sulcus; M1, primary motor cortex; M2, supplementary motor cortex; M3, rostral cingulate motor cortex; M4, caudal cingulate motor cortex; mefg, medial frontal gyrus; mfg, middle frontal gyrus; *mog*, medial orbital gyrus; *os*, olfactory sulcus; *pcgs*, paracingulate sulcus; pcl, paracentral lobule; PMC, lateral premotor cortex; pog, posterior orbital gyrus; P-SMA, presupplementary motor area; S1, primary somatosensory cortex; sfg, superior frontal gyrus; *sfs*, superior frontal sulcus; *srs*, superior rostral sulcus.

(Fig. 1). On the lateral convexity, interposed between the dorsolateral prefrontal and the ventral portion of premotor cortex, is Broca's area. This area is involved in the abstract mediation of the verbal expression of language. In humans, there is no primary sulcus demarcating the posterior limit of the prefrontal cortex. However, occasionally there are two distinct sulci within the prefrontal region designated as the superior frontal sulcus and inferior frontal sulcus. They are often discontinuous and extend horizontally to subdivide the rostral part of the frontal cortex, including the prefrontal region into superior frontal, middle frontal, and inferior frontal gyri.

2. Medial Organization

The cortex lining the medial wall of the hemisphere, located directly above the anterior portion of the paracingulate and/or cingulate sulci is divided into the paracentral lobule caudally and the medial frontal gyrus rostrally (Fig. 1). The division between the paracentral and medial frontal gyri is occasionally

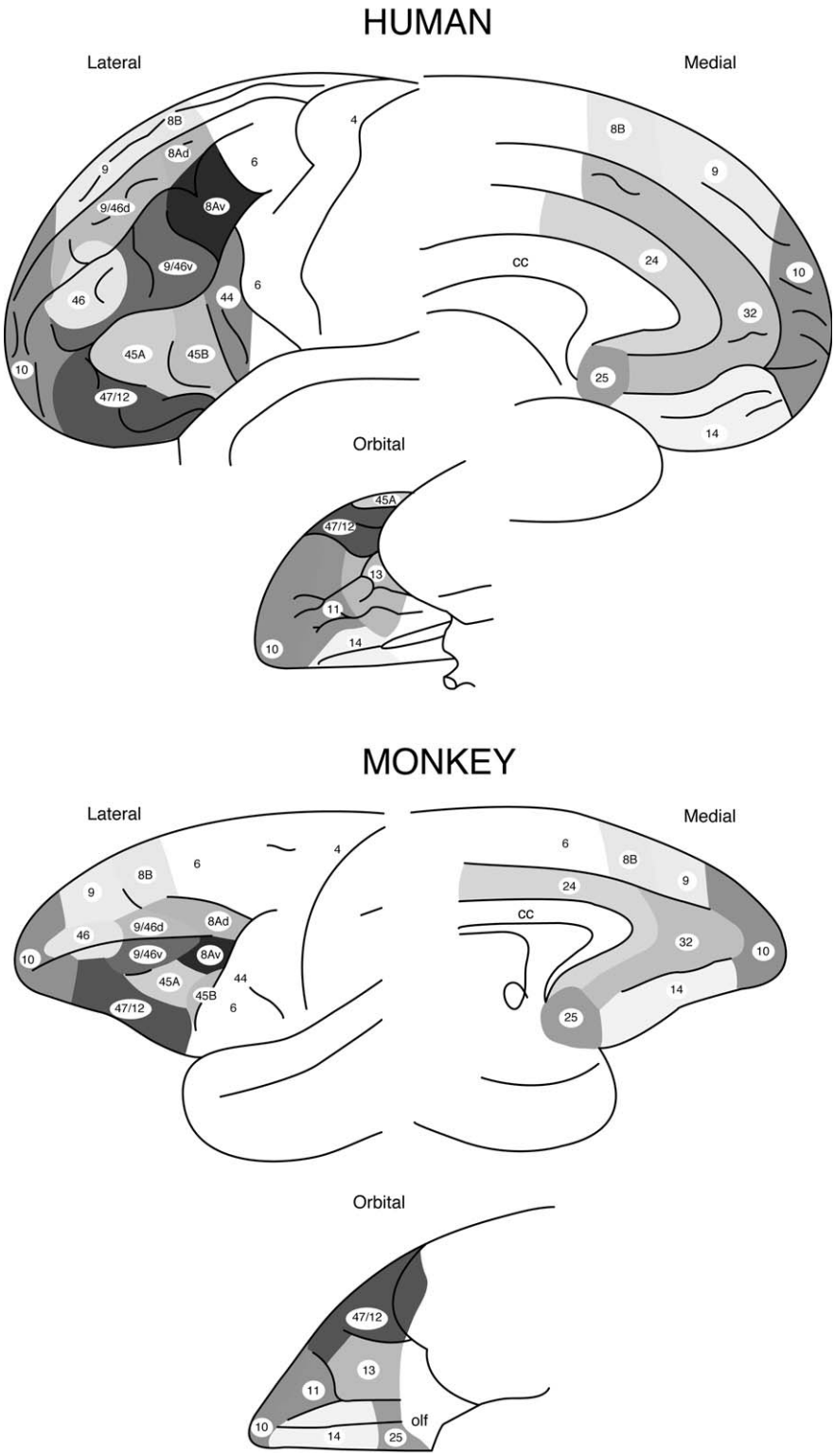


Figure 2 Schematic diagram illustrating the cytoarchitectonic organization of the lateral, medial, and orbital surfaces of the frontal region in the human (top) and monkey (bottom) cerebral cortices. Studies indicate that area 13 is more restricted, or limited in size, than illustrated on the orbital surface of the human brain. [From Petrides and Pandya, 1994, Comparative architectonic analysis of the human and the macaque frontal cortex. *Handbook of Neuropsychology*, Vol. 9. pp. 28, with permission from Elsevier Science.] (See color insert in Volume 1).

formed by a vertical sulcus emerging from the cingulate sulcus above the midpoint of the corpus callosum. The primary somatosensory (S1) and primary motor (M1) cortices occupy caudal and rostral portions of the paracentral lobule, respectively. The supplementary motor cortex (M2), presupplementary motor cortex (P-SMA), and prefrontal cortex comprise the caudal to rostral components of the superior frontal gyrus. Bordering the medial frontal gyrus ventrally is the cingulate gyrus and its affiliated subcallosal extension. In a majority of cases, the paracingulate sulcus roughly defines the border between the frontal lobe and the cingulate cortex of the limbic lobe. Ventral to the paracingulate sulcus is the cingulate sulcus. Rarely, an intralimbic sulcus is located ventral to the cingulate sulcus. Rostral (M3) and caudal (M4) cingulate motor cortices are located in the dorsal portion of the cingulate cortex.

3. Orbitofrontal Organization

The orbitofrontal cortex lines the ventral surface of the frontal lobe in the floor of the anterior cranial fossa (Figs. 1 and 2). The posterior portion of the orbitofrontal cortex is formed by agranular cortex, and the intermediate portion is formed by dysgranular cortex. This general region of cortex is categorized by some authorities as limbic or paralimbic cortex. It is continuous with the agranular and dysgranular cortices of the insula caudally and the temporal pole laterally. It is also continuous with the agranular–dysgranular cortices of the medial wall (e.g., areas 25 and 32). In the caudolateral region of the right dysgranular orbitofrontal cortex is a postulated secondary gustatory or taste region. Caudal to the secondary gustatory is the orbital extension of piriform (primary olfactory) cortex. Granular cortex forms the rostral portion of the orbitofrontal cortex, including the frontal pole. Unlike the lateral and medial surfaces of the frontal lobe, premotor cortex is absent from the orbitofrontal surface. Major sulci in this region include the olfactory sulcus, which is oriented rostrocaudally and demarcates the anatomical boundary between the gyrus rectus medially and orbitofrontal cortex proper laterally (Fig. 1). The orbitofrontal region is subdivided by irregular medial and lateral sulci that occasionally form an H shape. The medial and lateral orbital gyri occupy medial and lateral regions, respectively. Interposed between the medial and lateral orbitofrontal gyri are anterior and posterior orbitofrontal gyri.

B. Cytoarchitecture

Cytoarchitecture is the study of the structural arrangement of neurons within the central nervous system. Neuronal size, shape, packing density, and staining intensity are all features that are used to characterize a specific cytoarchitectural area, region, or trend. Historically, the prefrontal region is characterized cytoarchitecturally as "frontal granular cortex," referring to the fact that a large portion of the prefrontal cortex has a well-developed granular layer IV. Although this view is widely accepted when considering the more rostral parts of the frontal lobe, more complex features reflecting dysgranular and agranular types of cortex are also recognized in the more caudal portions of the transitional prefrontal region. Dysgranular refers to cortex having a poorly developed layer IV, and agranular refers to cortex lacking a layer IV.

Among the many cytoarchitectonic maps of the prefrontal cortices in both humans and monkeys, the general location of medial areas 24, 25, 32, and 10 are very similar in terms of location and nomenclature. However, many differences exist in the various interpretations of the cytoarchitectonic organization of the lateral and orbital areas. On the basis of cytoarchitectural commonalities, Pandya and co-workers have developed a working model of regional cytoarchitectonic homologies in the human and nonhuman primate cortices (Fig. 2). This approach is intended to assist in the interpretation, comparison, and integration of information obtained from the vast number of studies conducted in both species. A condensed version of this view is presented later. Although individual numerically assigned cytoarchitectonic areas are designated, it is important to call attention to the fact that two general, progressive trends in frontal architectonic development can be recognized (Fig. 3). They include the medial and ventral frontal trends. The medial trend originates from the hippocampal system (archicortical lineage) and progresses mediodorsally. In other words, the trend emanates from cortex lining the medial wall along the corpus callosum to cortex forming the dorsolateral frontal surface. The ventral trend originates from the olfactory system (paleocortical lineage) and progresses ventrally from the orbitofrontal region and laterally onto the ventrolateral frontal convexity. Thus, both trends originate from a simple, primitively organized moiety that is relatively undifferentiated in terms of laminae. This is followed by a continuous, stepwise progression of cortical development in which there is a

gradual increase in the number, size, and complexity of cortical layers (from agranular to dysgranular cortex), culminating in the formation of a well-developed six-layered organization (granular cortex) (Fig. 3).

Area 44 is located immediately anterior to the ventral portion of lateral area 6. In humans, area 44 comprises a major portion of Broca's area and is involved in mediating the control of speech. Architecturally, area 44 is dysgranular, referring to the fact that it has a poorly developed layer IV. This cortex is also characterized by the presence of large pyramidal neurons located in the deep portion of layer III.

Area 8 plays a major role in mediating complex eye movements and resides on the dorsolateral surface of the frontal lobe, rostral to the dorsal portion of lateral area 6 and caudal to granular areas 9–46. Area 8 extends medially, where it is located anterior to the presupplementary motor area (pre-SMA) and caudal to area 9. On the lateral convexity, a well-developed layer IV with large pyramidal cells in layer III characterizes the ventral portion of area 8. Progressing dorsally and medially, layer IV becomes less prominent and the number of large pyramids in layer III diminish.

Area 45 corresponds to pars triangularis of the inferior frontal gyrus. Area 45 forms the rostral portion of Broca's expressive speech area, whereas area 44 forms the caudal portion. Structurally, area 45

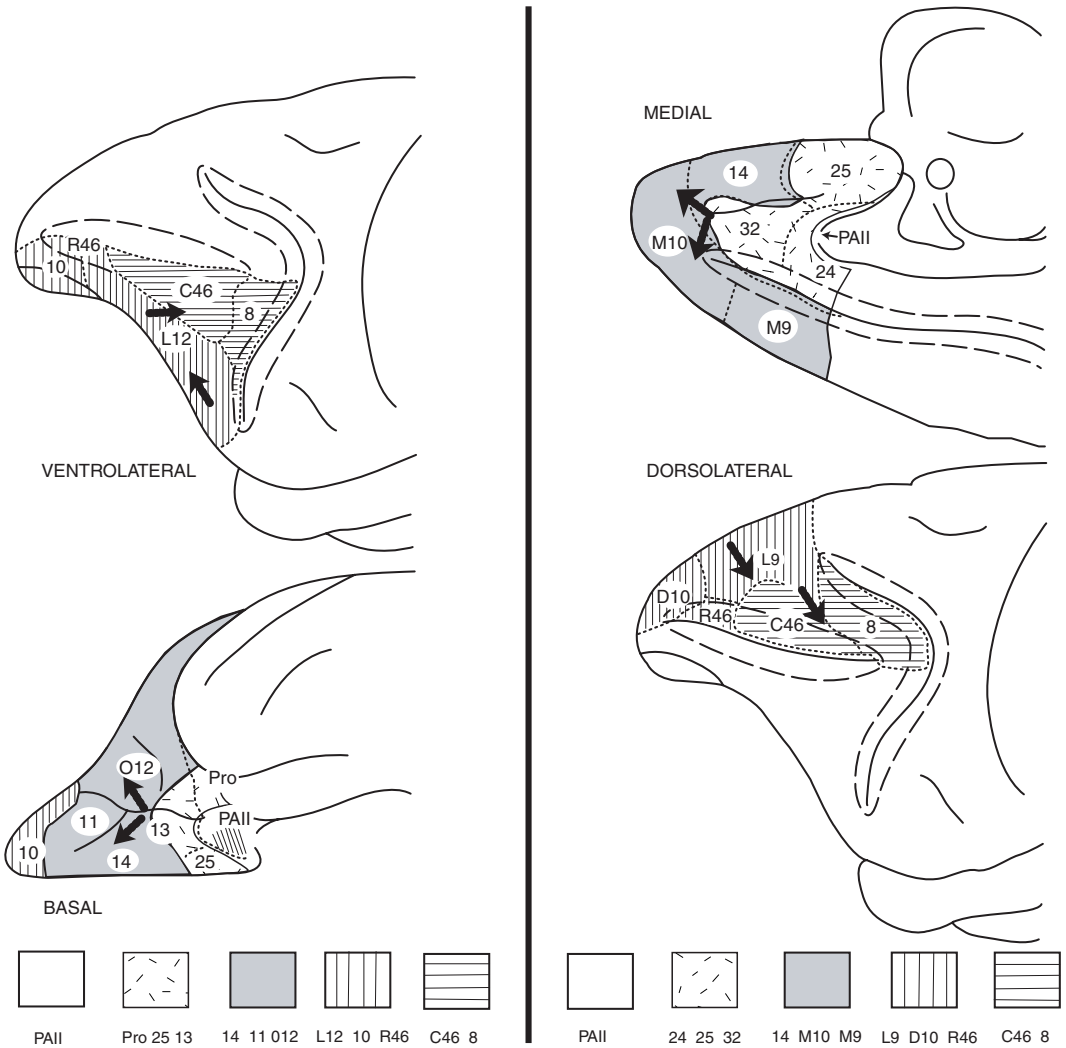


Figure 3 Diagrams depicting the cytoarchitectonic areas in the prefrontal cortex of the rhesus monkey arranged in progressive stages within the orbital and ventral paleocortical trends (left) and medial and dorsolateral archicortical trends (right). [From Barbas and Pandya, 1989, Architecture and intrinsic connections of the prefrontal cortex in the rhesus monkey. *J. Comp. Neurol.* **286,** 372. Reprinted by permission of Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc.]

is characterized by a well-developed layer IV, a hallmark feature of granular frontal cortex. Like adjacent area 44, the lower portion of layer III contains large pyramidal neurons.

Area 9 is located in the dorsal-most portion of the lateral prefrontal cortex and extends medially onto the medial frontal gyrus. **Area 46** is located rostral to area 9, where it occupies the rostral portion of the middle and superior frontal gyri. Morphologically, the lateral portion of area 9 is characterized as dysgranular to moderately granular. Medially, however, granular layers II and IV are less developed. In contrast, area 46 is highly granularized, referring to the fact that layers II and IV are well-developed. Furthermore, area 46 contains medium-sized pyramidal cells in layers III and V.

Area 10 occupies the frontal pole. Layer II, although not prominent, is distinguishable. Layer III is the widest layer in area 10, and layer IV is thin, but prominent. Sublamination is evident in layer V.

Areas 47/12 are located ventrally along the lateral surface and extend onto the lateral portion of the orbital surface. Layer III contains small and medium-sized pyramidal cells and is clearly separated from layer V by layer IV. The lateral and orbital components are further distinguishable. Layer IV is less developed on the orbital surface, and the pyramidal cells are less densely packed in this location.

Area 13 occupies the centrocaudal portion of the orbitofrontal surface. The rostral portion of area 13 is characterized as dysgranular due to its poorly developed layer IV. It contains a poorly differentiated layer II that is difficult to distinguish from adjacent layer III. Similarly, a distinct border between layers V and VI is difficult to recognize. The caudal region is agranular, lacking discernible granular cells in layers II and IV, thus forming a primitive, two-cell-layer arrangement (e.g., bilaminate organization).

Area 14 is located in the gyrus rectus on the orbitofrontal surface. The rostral portion of area 14 is dysgranular. Like adjacent area 13, layers II and IV diminish in size progressing caudally until this cortex becomes agranular. In the agranular sector, inner and outer strata (e.g., bilaminate organization) characterize area 14.

Areas 25 and 32 are located below the genu of the corpus callosum. In both areas, layer I is thick and there is a discernible layer IV. In area 25, layer III is broad, whereas layer V is densely packed. In contrast, layer V in area 32 widens and sublaminates where medium-sized pyramidal neurons are situated superficially.

C. Connections

1. Corticocortical Connections

a. Local Connections Local, or intrinsic, connections of the frontal lobe form important feedforward and feedback circuits, which modulate the neurotransmission of pyramidal and nonpyramidal neuron ensembles. These, in part, form the core elements of prefrontal information processing. These connections are organized such that each architectonic area has a connection with a more differentiated region, on the one hand (e.g., rostral or dorsal), and with a less differentiated region on the other (e.g., caudal or ventral) (Fig. 3). The orbital proisocortical area of the ventral trend projects mainly to orbital areas 11, 13, and 14, as well as to inferior prefrontal area 12. Area 13 in turn projects back to the proisocortical area, as well as to areas 11, 14, and 12. Area 12 projects to orbital areas 11, 13, and 14 and to ventral prefrontal areas 10 and 46. Ventral area 46 sends connections to areas 12 and 10 rostroventrally and to ventral area 8 caudally. Area 8 projects to ventral area 46 rostrally and to dorsal area 8 and lateral premotor area 6 caudally. Similarly, the dorsal trend proisocortical area 32 projects to areas 25 and 14 ventrally, area 24 caudally, and areas 9 and 10 dorsally. Area 9 in turn sends connections to proisocortex (areas 32 and 24) medially and to dorsal areas 10 and 6 laterally. Dorsal area 46 projects to areas 9 and 10, on the one hand, and to dorsal area 8 on the other. In addition, dorsal area 8 projects to areas 46 and 9 rostrally, dorsal area 6 caudally, and area 8 ventrally.

Thus, within both the dorsal and the ventral frontal trends, a systematic pattern of connectivity interrelates more differentiated regions of cortex with less differentiated regions. In addition, it is important to recognize that dorsal and ventral regions are interconnected at certain critical levels. For example, the orbital (ventral trend) and medial (dorsal trend) proisocortical areas are interconnected, as are areas 9 and 12, which reside above (dorsal trend) and below (ventral trend) the principal sulcus, respectively. Finally, there is a specific pattern of connectivity between dorsal and ventral area 8 as well as between dorsal and ventral area 6.

b. Long Association Connections Long association connections represent an integral part of the extended prefrontal system linking the prefrontal cortex with distant and widely dispersed parts of the

association and limbic cortices. With regard to afferent connections of the frontal lobe, the lateral extrastriate and caudal inferotemporal regions project to the ventral part of the arcuate cortex (ventral area 8) and to area 46 below the principal sulcus. In contrast, the rostral inferotemporal region projects to orbital area 11 and area 12 of the inferior prefrontal convexity. The proisocortical area of the superior temporal gyrus (area Pro of the temporal pole) and the caudally adjacent portion of the gyrus (area Ts1) are connected preferentially with the orbital and medial prefrontal cortices. The caudal portion of the superior temporal gyrus (areas Tpt and paAlt) projects to area 8 of the caudal prefrontal cortex. The intermediate portion of the superior temporal gyrus is related to the lateral prefrontal cortices. The superior and medial parietal association regions project preferentially to the dorsal and medial prefrontal cortices, which also receive afferents from visual areas PG-Opt and POa of the inferior parietal lobule. The inferior parietal areas, including the dorsal Sylvian opercular cortex and the gustatory cortex, project mainly to the ventrolateral and orbital prefrontal cortices. The visual association areas of the ventral and dorsomedial occipitotemporal cortices, relating to the peripheral visual field, project mainly to the dorsal prefrontal region. In contrast, the areas of the inferotemporal region involved in central vision project preferentially to the ventral prefrontal region.

The combined entorhinal and perirhinal region, which comprises the rostral part of the parahippocampal gyrus, projects to the orbital and medial (area 25) prefrontal cortices. The caudal part of the parahippocampal gyrus sends efferent connections to the medial prefrontal, caudal orbital, and ventrolateral prefrontal regions. The rostral (agranular–dysgranular) insula projects to the ventrolateral and orbital prefrontal regions (areas 12 and 13). The caudal (granular) insula has efferent connections to the dorsolateral prefrontal cortex (area 8). The cingulate and retrosplenial cortices project to the dorsal prefrontal region.

Within the prefrontal region are specific multimodal areas that receive input from more than a single sensory modality. Area 45 of the ventral prefrontal cortex has afferents from the rostral visual area of the inferotemporal cortex, the ventral somatosensory area, and the auditory cortex of the superior temporal gyrus. In addition, area 45 receives input from multimodal area TPO of the superior temporal sulcus, the insula, and the parahippocampal gyrus. The caudal orbitofrontal cortex receives input from visual association cortices as well as from the olfactory and gustatory areas. Other multimodal frontal regions include dorsal area 8, which receives input from both visual and auditory modalities, and areas 9 and 10, which receive input from multimodal cortices of the superior temporal sulcus.

The efferent connections of the frontal lobe are largely reciprocal to the afferent connections. For example, the dorsal prefrontal cortex projects to cingulate areas 24 and 23 (including the rostral, M3, and caudal, M4, cingulate motor cortices), retrosplenial cortex, as well as the medial parietal and caudal inferior parietal regions. In addition, dorsal prefrontal cortex projects to the presupplementary motor area (P-SMA) and the face representation of the supplementary motor cortex, M2 (Fig. 4). The transitional region between areas 6 V and 12 also projects to the face–head representation of the primary motor cortex, M1. The ventral prefrontal region projects mainly to the inferotemporal cortices, the ventral somatosensory cortex, and the rostral superior temporal region. Finally, the orbitofrontal cortex projects to the entorhinal, perirhinal, and parahippocampal cortices.

2. Corticostriatal Connections

The corticostriatal projection is nonreciprocal and represents the initial stage of the classic striatopallidothalamic pathway, or loop, that is largely directed back to the cortex via thalamocortical projections (see the following section). In general, this pathway is topographically organized throughout the entire loop and is typically directed back to the cortex (e.g., via thalamocortical projections to the dorsal lateral prefrontal region) from which the pathway originated (e.g., lateral prefrontal projections to the caudate nucleus of the striatum). The corticostriatal projection from prefrontal cortex to the caudate nucleus has a distinctive medial–lateral topography that is systematically related to the origin of the projection in the frontal lobe (Fig. 5, top). Medial and orbital prefrontal regions project mainly to the medial and intermediate portions of the head and body of the caudate nucleus. This projection also includes the nucleus accumbens of the ventral striatum. The ventral striatum has been the focus of an increasing number of studies, suggesting that the nucleus accumbens plays a prominent role in mediating reward and motivation. Of clinical relevance is the potential involvement of the nucleus accumbens in drug addiction and mental disorders such as schizophrenia and Tourette's syndrome. The

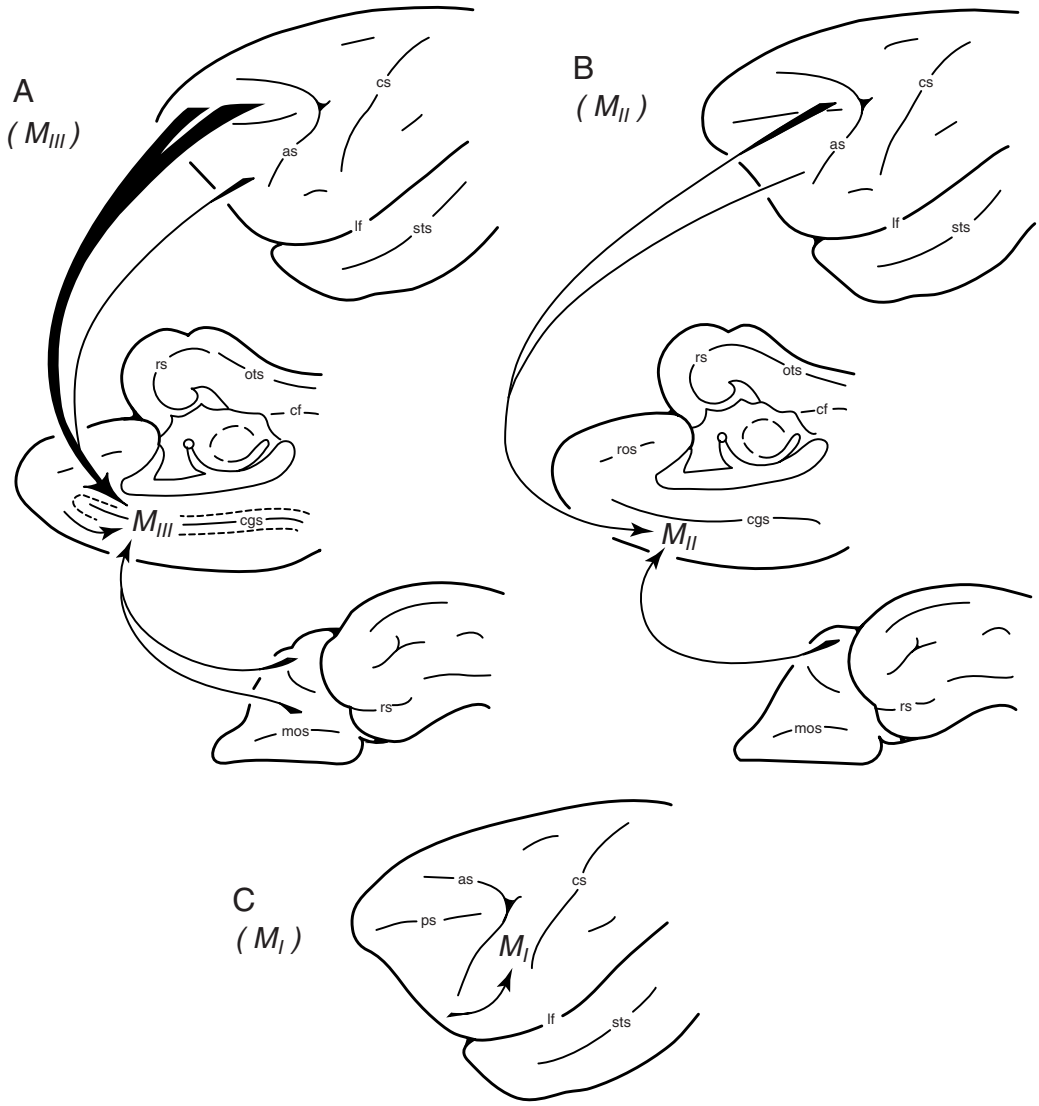


Figure 4 Line drawings summarizing prefrontal inputs to the motor cortices. The rostral (M3) and caudal (M4; not shown) cingulate motor areas receive significant prefrontal inputs primarily from the dorsolateral sector. The supplementary motor cortex (M2) receives weak inputs from the same prefrontal region that targets only the face–head representation. The face–head representation of the primary motor cortex (M1) receives input from the transition prefrontal region that is formed in part by the rostral portion of ventral area 6. as, arcuate sulcus; cf, calcarine fissure; cgs, cingulate sulcus; cs, central sulcus; lf, lateral fissure; M_I, primary motor cortex; M_{II}, supplementary motor cortex; M_{III}, rostral cingulate motor cortex; mos, medial orbital sulcus; ots, occipitotemporal sulcus; ps, principal sulcus; ros, rostral sulcus; rs, rhinal sulcus; sts, superior temporal sulcus. [from Morecraft and Van Hoesen, 1993, Frontal granular cortex input to the cingulate (M3), supplementary (M2), and primary (M1) motor cortices in the rhesus monkey. *J. Comp. Neurol.* **337,** 13. Reprinted with permission of Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc.]

lateral prefrontal cortex is related preferentially to the intermediate portions of the head and body of the caudate nucleus, whereas the caudal prefrontal region (area 8) projects mainly to the lateral and intermediate portions of the nucleus. A dorsal–ventral topography is evident as well (Fig. 5, top). Dorsal prefrontal areas are related predominantly to the dorsal portions of the head and body of the caudate nucleus, whereas ventral prefrontal areas are connected preferentially with the ventral portion. Nevertheless, there is a certain degree of intermingling of projections from the dorsal and ventral prefrontal cortices, specifically within the central sector of the head of the caudate nucleus. The topography of prefrontal projections to the tail of the

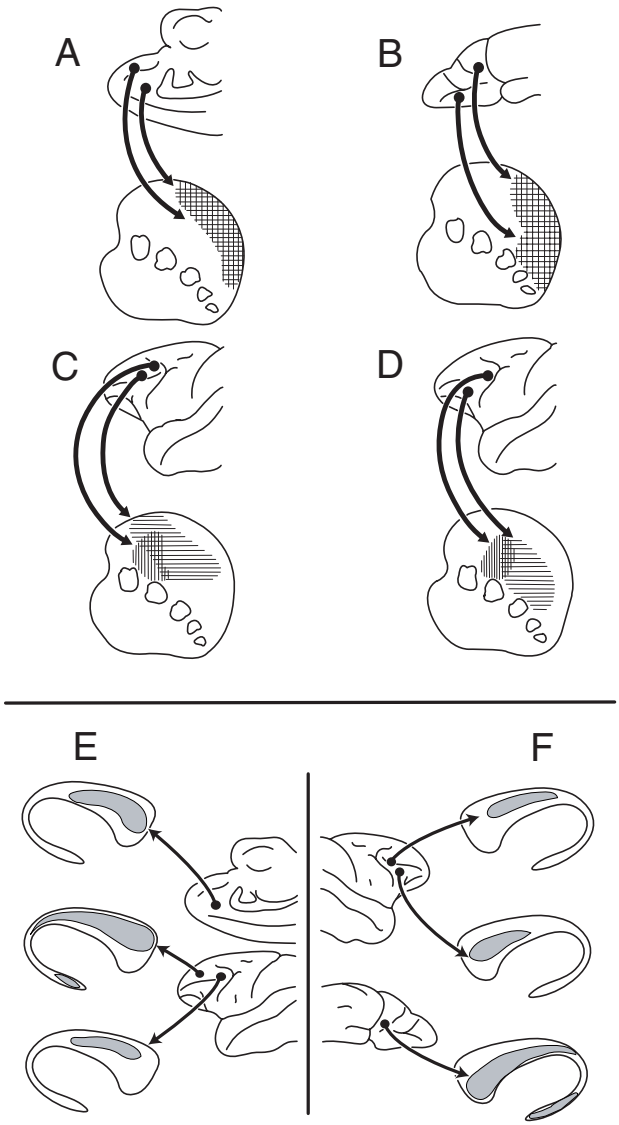


Figure 5 Summary diagrams showing the mediolateral (A–D) and rostrocaudal (E and F) connectional relationships of the dorsal or archicortical (A, C, and E) and ventral or paleocortical (B, D and F) architectonic trends of the prefrontal cortex and the caudate nucleus. [from Yeterian and Pandya, 1991, Prefrontostriatal connections in relation to cortical architectonic organization in rhesus monkeys. *J. Comp. Neurol.* **312,** 60–61. Reproduced by permission of Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc.]

caudate is less distinct. The medial portion of the tail receives projections from the dorsal and ventral proisocortical regions, as well as from the dorsomedial (dorsal area 10) and dorsolateral (dorsal areas 9 and 46) portions of the prefrontal cortices.

In comparison to prefrontocaudate connections, the prefrontal projections to the putamen nucleus are less extensive. In most instances, these appear to be continuous with connections to the caudate nucleus and have a dorsoventral topography. Dorsal trend regions (areas 10, 9, 46, and 8) project mainly to dorsal and central parts of the putamen nucleus. In contrast, ventral trend regions (areas Pro, 13, 12, 14, 10, and 46) project to ventral and central portions of the putamen, with the exception of ventral area 8, which projects to the dorsal part of the putamen. In general, the projections to the putamen from prefrontal regions with highly differentiated laminar organization (e.g., areas 46 and 8) tend to be more dorsal.

The distribution of prefrontal projections to the striatum, when viewed sagittally, appear to be longitudinal, although the projections tend to be patchy

rather than continuous (Fig. 5, bottom). Moreover, not all prefrontal regions have the same rostrocaudal extent of projections. For example, the prefrontocaudate projection from the dorsal arcuate region is less extensive than the projection from dorsal areas 9 and 46. Similarly, the prefrontocaudate projection from the ventral arcuate region is less extensive than those from the orbitofrontal region.

3. Corticothalamic Connections

The prefrontal cortex is interconnected with many thalamic nuclei, but the most prominent prefrontothalamic connection is established with the mediodorsal thalamic nucleus (MD). This connection represents another major outflow pathway of the prefrontal cortex that is largely reciprocated by thalamoprefrontal connections. Studies conducted by many investigators demonstrate that the connections with specific subdivisions of MD follow the mediolateral and rostrocaudal topography of the frontal lobe (Fig. 6). Thus, the medial prefrontal and orbitofrontal proisocortical areas are reciprocally interconnected with the medial portion (magnocellular division) of MD, maintaining their respective dorsal and ventral topographies in this subnucleus. The lateral prefrontal region, above and below the principal sulcus, is linked with the middle part (parvocellular division) of MD, again showing a dorsal to ventral topography. Finally, dorsal area 8 and ventral area 8 are reciprocally interconnected, respectively, with dorsal and ventral portions of the most lateral part (multiformis division) of MD. In summary, prefrontal projections to and from MD correspond to the established patterns of prefrontal cytoarchitectonic differentiation within the dorsal and ventral trends.

4. Corticofugal and Corticopetal Connections

In addition to the corticostriatal and corticothalamic projections discussed earlier, prefrontal interconnections with the amygdala, hypothalamus, midbrain, and pons represent important subcortical linkages of the extended prefrontal neural system. These are likely to integrate higher order brain functions mediated by the prefrontal cortex with more developmentally fundamental brain activities, such as emotion and visceral, or autonomic, functions. Related are important, chemically specific projections to various parts of the prefrontal cortices that arise from the nucleus basalis of Meynert (cholinergic), raphe nuclei (serotonin), hypothalamus (histamine), ventral tegmental area (dopamine), and locus ceruleus (norepinephrine). These projections are thought to play a vital role in cortical arousal, motivational valence, and the overall effectiveness of learning.

a. Amygdala The amygdala plays a central role in processing the emotional significance of complex situations based upon previous and current experiences. Under normal circumstances, cognitive functions such as abstract reasoning, decision making, and planning are greatly affected by previous emotional experiences as well as our current emotional state. The amygdala exerts a direct influence on prefrontal cortex through widespread and reciprocal projections preferentially interconnecting the amygdala with less differentiated cortices of the medial and lateral trends. This cortex includes the agranular and dysgranular portions of the medial and orbitofrontal cortices. Cortex ventrolateral to the principal sulcus is also moderately interconnected with the amygdala. The amygdala projection to the prefrontal cortex arises from the basolateral and accessory basal nuclei and to a lesser extent from the cortical and lateral nuclei. The amygdala projection terminates primarily in superficial cortical layers, namely layers I and II, as well as the superficial portion of layer III. Therefore, the various patterns of connectivity suggest that cognitive operations mediated by the prefrontal cortices are influenced by emotional context through selective amygdaloprefrontal pathways.

b. Hypothalamus The hypothalamus is a major subdivision of the diencephalon that is well-known for mediating core autonomic functions, including feeding, fluid regulation, heart rate, respiratory rate, reproduction, and self-defense. The agranular and dysgranular portions of the prefrontal cortex are reciprocally interconnected with the hypothalamus. For example, medial areas 25 and 32 are reciprocally connected with anterior and ventromedial hypothalamic regions, whereas the posterior orbitofrontal cortex is interconnected primarily with the posterior hypothalamic region. There is also a mediolateral topography suggesting that lateral parts of the hypothalamus are connected with the posterior orbitofrontal cortices, whereas the medial hypothalamus is interconnected with the medial prefrontal cortices. Lateral prefrontal connections with the hypothalamus are weaker than those established with the orbitofrontal and medial prefrontal cortices and primarily target the posterior hypothalamic region. The patterns of descending projections indicate that

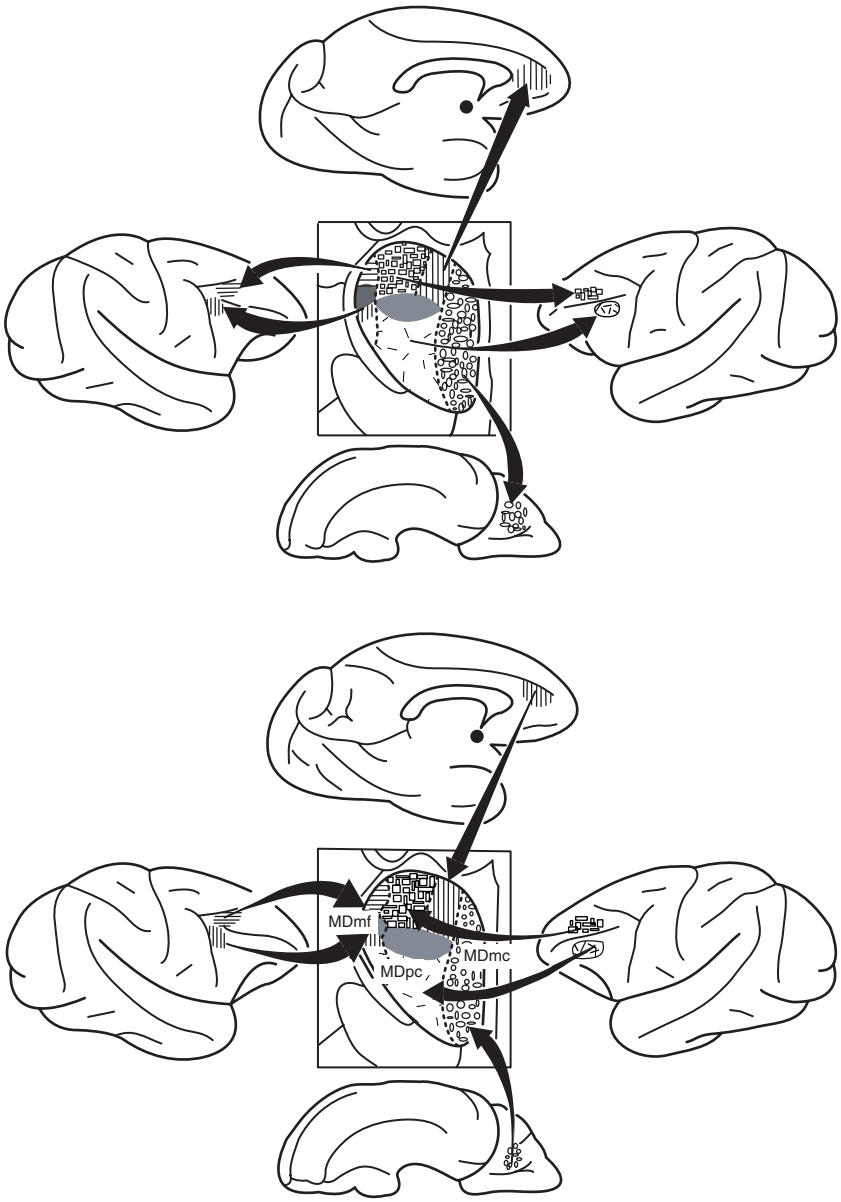


Figure 6 Summary diagram illustrating the topographic organization of thalamocortical (top) and corticothalamic (bottom) connections between the prefrontal cortex and the mediodorsal nucleus. [from Siwek and Pandya, 1991, Prefrontal projections to the mediodorsal nucleus in the rhesus monkey. *J. Comp. Neurol.* **312,** 519. Reprinted by permission of Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc.]

prefrontal activity may directly initiate autonomic responses accompanying stressful and important decision-making processes. Similarly, direct ascending hypothalamic projections to the prefrontal cortex indicate that basic and powerful physiological drives are likely to affect the outcome of higher order behavior. These connections may be involved in recruiting a prefrontal contribution to appropriately address the autonomic needs of the body in general (e.g., maintenance of physiological homeostasis) or to solve a potentially life-threatening situation.

c. Midbrain Prefrontal projections terminate in several major midbrain structures, which include the substantia nigra, periaqueductal gray (PAG), cuneiform nucleus, median raphe nucleus, and reticular formation. Collectively, prefrontomesencephalic projections are thought to play a major role in influencing

vocalization, autonomic mechanisms, and pain-related responses. Medial, lateral, and orbital projections to the PAG are highly organized and terminate in distinct columns. The prefrontal projection to the PAG is heaviest from the medial and posterior orbital regions and terminates, respectively, in the dorsolateral and ventrolateral columns. In addition, projections from dorsomedial areas 9 and 24b end mainly in the lateral column. The prefrontal projection to the PAG from lateral areas 9, 46, and 8 is weak and directed mainly to the dorsolateral column. Finally, the rostral cingulate motor cortex (M3) projects to the lateral and ventrolateral sectors of the PAG.

d. Pons The pontine nuclei receive ipsilateral input from many parts of the cerebral cortex. The pontine nuclei in turn project to the contralateral cerebellar hemisphere. The cerebellum has long been known to mediate aspects of motor behavior, and projections from motor cortex to the pontine nuclei represent a critical contribution to this function in higher primates. However, several new observations suggest a cerebellar role in mediating cognitive processes. Possibly related is the more recent observation of sizeable and highly organized corticopontine projections arising from selected parts of the prefrontal cortex. Massive corticopontine projections originate from dorsolateral prefrontal areas 8d, 9, 46d, and 10 and terminate in the paramedian and medial parts of the peripeduncular pontine nuclei. Medial areas 32, 9, 8B, and possibly 10 give rise to corticopontine projections ending primarily in the medial and paramedian nuclei and the medial portion of the peduncular and peripeduncular nuclei. In contrast, the ventrolateral prefrontal cortices do not project to the pontine nuclei, with the exception of area 45. Similarly, the available evidence indicates that orbitofrontal cortex does not project to the pontine nuclei.

II. FUNCTIONAL ORGANIZATION OF THE PREFRONTAL CORTEX

Prefrontal cortex is often classified as association or multimodal association cortex. This refers to the fact that highly processed information from multiple sensory modalities is integrated or combined within the prefrontal region in a precise, temporally ordered fashion to form the physiological constructs of perception, memory, and complex action. Specific processes mediated by the prefrontal cortex include attention, planning, decision making, emotion, and personality. In general terms, these functions are loosely categorized as higher order, executive, or cognitive functions. Although it is clear that prefrontal cortex as a whole mediates complex behaviors, there is considerable evidence for functionally distinct subregions within the various prefrontal territories.

A. Lateral Prefrontal Cortex

The peri-principalis region in monkeys correlates to the dorsal lateral prefrontal cortex in humans and is generally thought to play a role in tasks that depend upon memory and spatial information, that is, in spatial working or representational memory. Different functional roles for the peri-principalis region and the adjacent dorsolateral prefrontal cortex have been proposed. These include the integration of spatial and visual information, the motivational evaluation of sensory stimuli, the monitoring of the serial order of stimuli during the performance of self-ordered tasks, and the performance of tasks dependent upon tactile and auditory as well as visual working memory. The peri-principalis region has also been shown to have a role in the control of saccadic eye movements and visual attention. The inferior prefrontal convexity, ventral to the peri-principalis region, is known to be involved in processes relating to the behavioral significance of stimuli as well as in reward. Damage to this region results in strong perseverative tendencies, indicating a role in the control of responses to specific stimuli. It has also been suggested that the cortex of the inferior convexity is involved in stimulus selection and attention, immediate visual memory, and the appreciation of faces. In contrast, the cortex of the superior convexity, dorsal to the peri-principalis region, has been implicated in kinesthetic function as well as in the performance of self-ordered and externally ordered tasks.

The caudal-most portion of the lateral prefrontal region includes the frontal eye fields and has a distinctive functional role (Fig. 1). It is well established that this region serves processes important for attention and orientation, particularly in the visual sphere and in the auditory and somatosensory modalities as well. For example, this region mediates smooth-pursuit eye movements, the control of visual scanning and fixation, and auditory orientation as well as aurally guided saccades. In addition, the arcuate region has been implicated in integrative functions such as cross-

modal matching and in intermodal association and conditional associative learning. The ventral periarcuate cortex has also been shown to mediate aspects of vocal control including the laryngeal musculature, which would be consistent with the vocal conduction abnormalities that occur following damage to the homologous region in humans.

B. Medial Prefrontal Cortex

Recent advances in behavioral methodologies, and the ever-expanding application of functional imaging have led to unprecedented advances in our understanding of medial prefrontal region functions. Specifically, the medial frontal region (anterior cingulate area) appears to be involved in bimanual coordination, attention to demanding cognitive tasks, modulation of body arousal, spatial memory, self-initiated movement, conflict resolution, and the retrieval of information from long-term memory (medial prefrontal and medial orbital regions). The anterior cingulate cortex is also involved in the perception of pain (e.g., perceived unpleasantness of painful stimuli) and possibly in mediating the emotional response that is initiated by our perception of pain. The rostral cingulate motor cortex (M3) is a higher order motor area that is involved in selecting an appropriate voluntary motor act on the basis of the consequences and the value of the associated reward. Reward and goal-related activity is thought to correspond to the unique patterns of connections that link the rostral cingulate motor cortex with the prefrontal and limbic cortices, on the one hand, and many classically affiliated motor targets including the isocortical motor areas (e.g., M1 and M2), brain stem (e.g., facial nucleus), and spinal cord on the other (Fig. 7). The presupplementary motor area (pre-SMA) plays a role in generating a motor plan or strategy in response to an instructional signal. In comparison, activity in the adjacent supplementary motor area (M2) is related to the onset or initiation of movement. Finally, it has also been shown that medial frontal damage results in perseverative behavior.

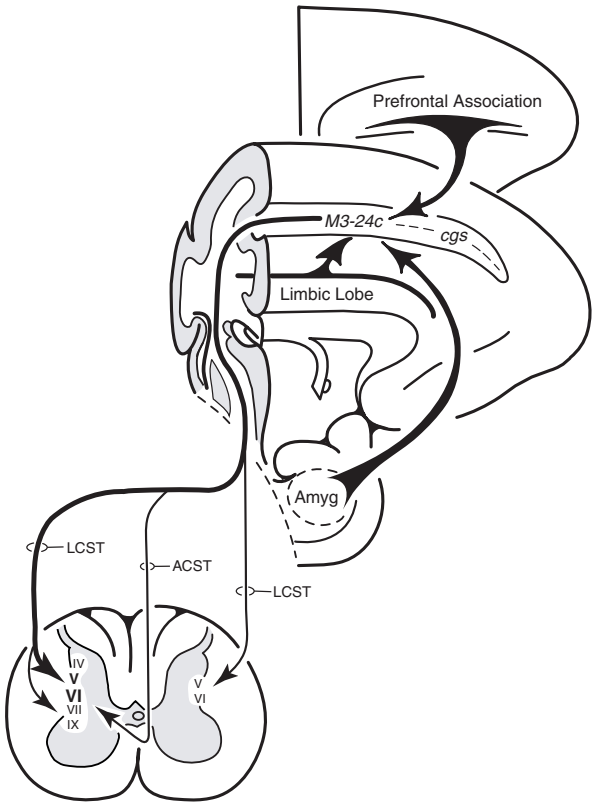


Figure 7 Summary diagram illustrating important higher order inputs from prefrontal and limbic cortices (top, right) to the rostral cingulate motor area (M3–area 24c), which projects directly to the spinal cord (bottom, left) and facial nucleus (not shown). Also illustrated is amygdala input to M3, which may play an important role in coupling emotional relevance and complex motor activity. ACST, anterior corticospinal tract; Amyg, amygdala; cgs, cingulate sulcus; LCST, lateral corticospinal tract. [from Morecraft *et al.*, 1997, Segregated parallel inputs to the brachial spinal cord from the cingulate motor cortex in the monkey. *Neuroreport.* **8**(18), 3933–3938. Copyright Lippincott, Williams and Wilkins.]

C. Orbitofrontal Cortex

The orbital prefrontal region appears to be involved in autonomic, emotional, response inhibition, and stimulus significance functions. This region has been shown to have a significant role in social and emotional behavior and in the formation of new memories. The orbitofrontal cortex is involved in olfactory and gustatory processing and in the appreciation of the emotional significance of stimuli. In addition, this region has been shown to serve the processing of facial information, the relation of specific stimulus features to reinforcement in learning paradigms, and the regulation of behavior in response to reward and punishment. From neuropsychological, neurophysiological, and clinical studies, it can be concluded that these functions are consistent with the distinctive and heavy orbitofrontal connections established with numerous cortical and subcortical limbic system structures.

III. CLINICAL SYNDROMES

The precise origin of the phrase "frontal lobe syndrome" remains obscure, but it has been used as an umbrella term to characterize patients with deficits in executive or intellectually related functions. These deficits cover a wide spectrum of complex behavioral changes, including impairments in foresight, judgment, initiative, creativity, problem-solving, social conduct, and attention. However, experimental data, lesion-based behavioral analyses and functional imaging observations demonstrate that the executive functions managed by the brain depend not only on prefrontal cortex but also on the integrity of other cortical and subcortical structures that are interconnected with the prefrontal cortex (as reviewed in this article). Although the various parts of the extended prefrontal neural system are not equivalent in their contribution to the overall cognitive process, their participation is essential for appropriate and skilled execution of higher order tasks. The clinical relevance of a "distributed neurocognitive network" is that lesions located within the extended cortical and subcortical neural system may produce functional deficits similar to those that occur when damage is isolated to a specific prefrontal region. For example, patients with lesions involving parts of the basal ganglia or mediodorsal thalamic nucleus that are topographically interconnected with dorsolateral prefrontal cortex may present with behavioral disorders essentially the same as those occurring in patients with damage restricted to the dorsolateral prefrontal cortex.

Related is the fact that brain lesions are often difficult to delineate precisely and are rarely confined to the cortical gray matter. Therefore, higher order behavioral deficits associated with prefrontal lesions are often too complex to fit a simple unified description such as frontal lobe syndrome. In the case of prefrontal damage, this has led to the development of more specific terminology based on structural, functional, and clinical features, which recognize *dorsolateral*, *medial*, and *orbital* prefrontal regions as described next. This view not only takes into consideration the unique contributions of the various and highly specific corticocortical and subcortical prefrontal interconnections but also reinforces the functional heterogeneity and specialization of each prefrontal subdivision. This conceptual framework provides a useful reference point for interpreting patient conditions and for advancing our functional understanding of this complex brain region and its inherent circuitry. However, caution should be exercised because lesions are rarely confined to any one prefrontal system, and clinical features of more than one "syndrome" are likely to occur in an individual patient.

A. Lateral Dysexecutive Syndrome

Common sources of damage to the dorsolateral frontal cortex include infarction of the lateral prefrontal branches of the middle cerebral artery, infarction of the internal frontal and frontopolar branches of the anterior cerebral artery, lateral frontal gliomas, convexity meningiomas, and head trauma. Central nervous system infections, demyelinating disorders, and degenerative diseases resulting in neuronal loss and gliosis of the frontal lobe, such as Pick's disease and frontotemporal dementia, alter prefrontal executive functions. In addition, schizophrenia and depression are associated with dorsolateral dysexecutive syndrome characteristics. For example, during verbal episodic memory retrieval tasks, schizophrenic patients show enhanced activity in the dorsolateral prefrontal cortex and reduced activity in the hippocampus when compared to controls. This indicates that normal recruitment of the hippocampus, a central structure in neural systems related to memory, is altered during memory retrieval in schizophrenics.

Patients with dysexecutive syndrome are characterized by deficits in many cognitive domains. This syndrome has been described as a metacognitive disorganization reflecting a reduced state of mental control. Working memory (regulating actions based upon on-line internal and external environmental stimuli) and the temporal ordering of recent events are impaired. The ability to recall or retrieve information is altered despite evidence of intact recognition. Patients present with diminished judgment, planning, insight, and self-care, and there is often a general reduction in verbal and nonverbal fluency.

B. Medial Apathetic Syndrome

The most common cause of medial apathetic syndrome is infarction of the anterior cerebral artery and its branches, which is a rare cardiovascular accident in itself. This syndrome is also associated with midline meningiomas of the falx cerebri, multiple sclerosis, central nervous system infections, and frontotemporal dementia. Occasionally, penetrating wounds whose

trajectory is directed along the midline of the skull, such as that reported in the landmark case of Phineas Gage, damage medial as well as ventromedial sectors of prefrontal cortex.

The hallmark feature of medial apathetic syndrome is a severe reduction in spontaneity and motivation. For example, medial syndrome patients are able to generate internally organized plans for action but lack the impetus to carry them out. This syndrome is also characterized by a reduced interest in the environment. Memory of recent events is relatively intact, as is the ability to recollect principal events. Although few neuropsychological deficits accompany this disorder, the patient has a flattened affect, illuminated by a blunted facial expression. It is thought that the overall alteration in motivation and motor activity is a result of the lesion involving the medial motor cortices. Medial motor areas considered to be of particular importance include the rostral (M3) and caudal (M4) cingulate motor areas and the supplementary (M2) motor area (Figs. 1 and 7). All three motor areas are known to project to the primary motor cortex, lateral premotor cortex, facial nucleus, and spinal cord. In addition, the cingulate motor areas directly receive widespread dorsolateral prefrontal, limbic, and amygdalar inputs, which under normal conditions are likely to provide the appropriate motivational impetus to carry out an internally generated plan of action (Fig. 7).

C. Orbitofrontal Disinhibited Syndrome

The orbitofrontal cortex rests on the floor of the anterior cranial fossa directly above the paired orbital and nasal cavities. It is separated from the orbital cavity by a relatively thin and fragile sheet of bone formed by the orbital plate of the frontal bone and cribriform plate of the ethmoid bone. This spatial association and the potential for fragmentation of the orbital plate render the orbitofrontal cortex highly vulnerable to injury following acceleration–deceleration head trauma, which is common, for example, in automotive collisions. Other common causes of orbitofrontal damage include rupture of the anterior communicating artery (often producing concomitant medial wall damage) and ventral frontal meningiomas. Orbitofrontal cortex is also adversely affected in frontotemporal dementias, viral infections (e.g., herpes encephalitis and Creutzfeldt–Jacob disease), and demyelinating disorders such as multiple sclerosis. Many neuroimaging studies have implicated the orbitofrontal cortex in the pathophysiology of obsessive-compulsive disorder and depression.

Patients with orbitofrontal damage are characterized generally by an acquired disturbance of personal and social behaviors. Remarkably, intellect is not grossly impaired. For example, learning and retention of factual knowledge are relatively preserved. Therefore, working memory is intact. The ability to use logic to solve problems and the ability to articulate verbally are preserved. However, there are marked abnormalities in the realms of reasoning, personal and social decision-making, emotional control, and feeling. Disruption in the ability to control feelings and emotions often results in explosive aggressive outbursts characterized by socially unacceptable, tactless, and vulgar presentation. The overwhelming feature of impulse dyscontrol underscores the classification—disinhibited syndrome. Depending upon the posterior extent of the lesion, patients may have impairments in identifying odors (olfactory damage) and tastes (gustatory involvement).

IV. CONCLUSIONS

The prefrontal cortex is regarded as a critical portion of the cerebral cortex that mediates intellectual or executive functions. Its diverse but specific connectivity patterns, electrophysiological properties, neuroimaging correlates, and associated clinical sequelae all reinforce the complex role of the prefrontal cortex in mediating higher order behaviors. The prefrontal cortex can be viewed as a major component of a large-scale neurocognitive network where complex behaviors are organized at the interactive level of multifocal neural systems. This network contains a variety of anatomical projections for transferring informational content to and from the prefrontal cortex. Also important is complex local circuitry involved in the short-term storage of information, encoding of this information, and synthesis of the associated mental representations to achieve an appropriate goal-directed response. The significance of the prefrontal cortex, which is extensively developed in primates, is related to its involvement in dynamic mental functions that underlie behaviors in response to novel and challenging demands.

See Also the Following Articles

ATTENTION • EMOTION • FRONTAL LOBE • HYPOTHALAMUS • LIMBIC SYSTEM • MIDBRAIN • WORKING MEMORY

Suggested Reading

Barbas, H. (1995). Anatomic basis of cognitive–emotional interactions in the primate prefrontal cortex. *Neurosci. Biobehav. Rev.* **19,** 499–510.

Damasio, A. R. (1994). *Descartes' Error*. Grosset/Putnam, New York.

Fogel, B. S., Schiffer, R. B., and Rao, S. M. (Eds.). (2002). *Neuropsychiatry*. Williams and Wilkins, Baltimore, MD.

Fuster, J. M. (1997). *The Prefrontal Cortex: Anatomy, Physiology and Neuropsychology of the Frontal Lobe*. Lippincott-Raven, New York.

Goldman-Rakic, P. S. (1987). Circuitry of prefrontal cortex and regulation of behavior by representational memory. In *Handbook of Physiology, The Nervous System, Vol. 5*, (V. B. Mountcastle and F. Plum, Eds.), pp. 373–417. American Physiological Society, Bethesda, MD.

Grafman, J., Holyoak, K. J., and Boller, F. (Eds.). (1995). *Structure and Functions of the Human Prefrontal Cortex*. New York Academy of Sciences, New York.

Mesulam, M. M. (1998). From sensation to cognition. *Brain* **121,** 1013–1052.

Miller, B. L., and Cummings, J. L. (Eds.). (1998). *The Human Frontal Lobes: Functions and Disorders*. Guilford Publications, Inc., New York.

Pandya, D. N., and Yeterian, E. H. (1996). Morphological correlations of human and monkey frontal lobe. In *Neurobiology of Decision-Making* (A. R. Damasio, *et al*., Eds.), pp. 13–46. Springer-Verlag, Berlin.

Passingham, R. E. (1995). *The Frontal Lobes and Voluntary Action*. Oxford University Press, New York.

Petrides, M., and Pandya, D. N. (1994). Comparative architectonic analysis of the human and the macaque frontal cortex. In *Handbook of Neuropsychology, Vol. 9* (F. Boller and J. Grafman, Eds.), pp. 17–58. Elsevier Science, New York.

Posner, M. I., DiGirolamo, G. J., and Fernandez-Duque, D. (1997). Brain mechanisms of cognitive skills. *Consciousness Cognition* **6,** 267–290.

Roberts, A. C., Robbins, T. W., and Weiskrantz, L. (Eds.). (1998). *The Prefrontal Cortex: Executive and Cognitive Functions*. Oxford University Press, New York.

Van Hoesen, G. W. (1993). The modern concept of association cortex. *Curr. Opin. Neurobiol.* **3,** 150–154.

Priming

ANTHONY D. WAGNER
Massachusetts Institute of Technology

WILMA KOUTSTAAL
University of Reading

GLOSSARY

declarative memory Forms of long-term memory that may be consciously recollected or explicitly expressed, including memory for facts, ideas, and events, and supported by a medial–temporal and diencephalic neural network. Tests of declarative memory, such as free recall and recognition, are also termed direct or explicit probes of memory.

global amnesia Condition involving marked impairments in declarative memory for recently encountered facts and events, together with intact immediate memory and perceptual and linguistic functions within the normal range.

nondeclarative memory Forms of long-term memory expressed as a change in behavior, rather than by conscious recollection, and including skill learning, conditioning, and priming. Tests of nondeclarative memory, particularly tests of priming, are also termed indirect or implicit probes of memory.

Priming refers to facilitative changes in the ability to identify, generate, or process an item due to a specific prior encounter with the item. This article will consider theoretical accounts of priming, surveying the extensive evidence regarding this form of nondeclarative memory that has been derived from behavioral, neuropsychological, neuroimaging, and electrophysiological approaches.

I. WHAT IS PRIMING?

Historically, the term ''priming'' has been used in several areas of psychology to refer to conditions or stimuli that change an organism or system into a state of increased readiness or preparedness for action or response. Within cognitive psychology and cognitive neuroscience, priming refers to instances in which an earlier encounter with a given stimulus (e.g., a word, face, or other object) alters (''primes'') subsequent responses to that same stimulus or to a related stimulus by increasing the speed of responding, increasing accuracy, or biasing the nature of the response given. For example, participants are more accurate at identifying briefly exposed, masked, or degraded words during a perceptual identification task if they have earlier encountered those words during an unrelated study phase than if the words were not previously exposed. Critically, findings from both brain-damaged individuals and neurologically intact populations demonstrate that this alteration in performance may occur under conditions where the task itself does not involve or require any reference to memory: where the individual has no explicit or overt intention to remember the earlier occurrence and where the individual may remain entirely unaware of the earlier occurrence and its effects on behavior.

A. Initial Insights

Prior to the 1960s, a number of philosophers and psychologists had proposed a distinction between forms of learning involving conscious awareness of the past and forms of learning that occur independent of recollection, using such terms as knowing how vs knowing that (Gilbert Ryle), memory vs habit (William James), explicit vs implicit recognition (William McDougall), and mechanical or sensitive memory vs representative memory (Maine de Biran). However, these distinctions and related clinical findings were based largely on informal observation and introspection and did not clearly map onto questions regarding the possible brain regions that might support these different forms of learning. Two key findings—derived from experimental investigation of patients with global amnesia—radically altered this state of affairs and eventually prompted intensive investigation into multiple forms of long-term "learning without awareness".

The first key finding, reported by Milner and colleagues in the early 1960s, involved the famous amnesic patient H.M. Following bilateral removal of the hippocampi and adjacent medial temporal lobe regions as treatment for intractable epilepsy, H.M. was found to show profoundly impaired memory for recently experienced events: if his attention was momentarily diverted from an event such that it was no longer in working memory, it was, for him, as if the event had never occurred, even though it happened only minutes earlier and had been fully attended and comprehended. Yet, despite this severe impairment in the ability to consciously recollect and report recently experienced events, H.M. showed some forms of preserved long-term memory, including normal learning and retention on a perceptual–motor skill task involving tracing the outline of a figure reflected in a mirror. Even though H.M. had no conscious recollection of having performed this task before, with practice he became increasingly adept at the task, tracing the figure more quickly and more accurately. This practice-related improvement was similar in magnitude to that shown by normal individuals without impaired memory. These observations provided a clear demonstration that different kinds of long-term memory can be distinguished in amnesia.

The second key finding, reported by Warrington and Weiskrantz in 1968 and 1974, showed that amnesics also could (indirectly) manifest evidence of recent learning in a cognitive–perceptual domain. If, instead of asking amnesic patients to explicitly or directly recall or recognize previously encountered words, the task was simply to try to complete a "word stem" with possible words (e.g., TAB__might be completed as TABLE or TABOO), then—just like individuals with intact memory—amnesics more often completed stems with words that had been presented earlier in the experiment. This priming effect, or increased likelihood of completing word stems with studied items, was subsequently clarified by Graf, Squire, and Mandler, who showed that amnesic patients demonstrated normal priming when the task instructions were to complete each stem with the first word that comes to mind but impaired performance when the same cues were given but with the instruction to use the cues to help intentionally remember a previously studied item.

Priming thus refers to instances in which prior experience with a specific stimulus influences subsequent behavior in the absence of intentional remembrance—a form of indirect, nonconscious, nondeclarative, or implicit memory. In everyday life, priming may take many forms. For example, if, in casual conversation, an acquaintance uses a relatively low-frequency word, such as "ebullient,' then, should an appropriate situation later arise, this word might be more likely to simply "spring to mind" and be used than if the earlier exposure had not occurred. Priming may also lead to errors, as when, in written composition, a word may appear particularly apt because it was recently used in an earlier sentence (yielding unintended repetition) or when an idea generated by someone else may be remembered without awareness of its earlier occurrence, leading the "unknowing rememberer" to fail to credit the original source (i.e., cryptomnesia or inadvertent plagiarism).

B. Forms of Priming

Priming is one of several types of learning that do not depend on conscious, explicit recollection of prior episodes, with classical conditioning and skill learning also comprising classes of nondeclarative learning. Further, the term priming is used to refer to several somewhat different phenomena, including repetition priming, semantic priming, and new association priming. Subtypes of priming within these broad classifications may also be distinguished.

1. Repetition Priming

Repetition priming is the chief focus of this article and refers to instances where exactly the same stimulus (or

a very similar stimulus) repeats after an initial occurrence. Tests of repetition priming may be subdivided in multiple ways, but one broad distinction is that between tests that primarily place demands on *perceptual* processes (tasks in which performance depends on analysis of stimulus form) and tests that rely more heavily on *conceptual* processes (tasks in which performance depends on analysis of stimulus meaning).

2. Semantic Priming

Semantic priming refers to a facilitation of responding that occurs as a result of the presentation of a semantically related word, as when presentation of the word ''nurse'' facilitates access to or decisions regarding ''doctor.'' Semantic priming effects are one of the most robust findings in cognitive psychology and have been reported for a variety of tasks, including lexical decision (participants are presented words and nonwords and decide whether each comprises a word), perceptual identification, and speeded word reading (naming). A primary difference between semantic priming and repetition priming concerns the longevity of the effects. Whereas repetition priming may be quite long-lasting, persisting for hours, days, or, in some instances, many months, semantic priming was—until relatively recently—believed to be much more transient, dissipating over the course of several seconds or after more than one intervening item between the prime and target stimulus. However, other data suggest that, particularly under conditions requiring more extensive or ''deeper'' encoding, semantic priming may span several intervening items. Long-term forms of semantic priming also may be more readily observed on tasks that require deeper semantic processing. The literature on semantic priming is not reviewed further here.

3. New Association Priming

New association priming, in contrast to priming of individual stimuli, refers to priming of the relationship or *association* between unrelated stimuli. Priming for new associations is measured by presenting pairs of preexperimentally unrelated stimuli, such as words, in a study phase (e.g., MUSTARD–SATELLITE, TABLE–REASON) and then presenting three types of pairs during a subsequent test phase: intact pairs (words presented in their original context or pairing, as in MUSTARD–SATELLITE), recombined pairs (TABLE–SATELLITE), and new baseline pairs. Measures of new association priming are based on a comparison of performance for intact pairs relative to recombined pairs. If intact pairs show greater facilitation than recombined pairs, this can be attributed to the retention and use of associative information because the individual items in both types of pairs were encountered previously—what differs is their conjunctive or associative pairing.

New association priming can be examined with a variety of verbal and nonverbal materials and tasks (e.g., word-stem completion, lexical decision, naming, color naming of nonwords or novel objects). From both a theoretical and a pragmatic viewpoint, studies of new association priming are important because they may serve to delineate the limits of repetition priming: can priming occur, not only for already familiar or learned individual items but also for new relationships or associations between items? This also has implications for understanding the neural substrates of priming and the extent to which damage to brain substrates believed to contribute to declarative memory may be necessary to support the acquisition of contextual associations between items.

4. Skill Learning vs Priming

Skill learning refers to an improvement in speed and/or accuracy of performance that arises with increased practice on a task, where this improvement not only is observed for specific items that were encountered previously (as in repetition and new association priming) but *generalizes* to new stimuli within a domain of processing. Skill learning may assume several forms, including enhanced performance on sensorimotor tasks (e.g., mirror tracing, rotary pursuit), perceptual tasks (e.g., reading mirror-reversed or inverted text), and cognitive tasks (e.g., artificial grammar learning, probabilistic classification learning). Although several studies comparing skill learning and priming in different neuropsychological populations have demonstrated that these two forms of learning rely on distinct neural mechanisms, it is important to note that these studies have often compared different types of tasks and stimuli for priming vs skill learning. Thus, the observed dissociations may reflect dissociations across domains of processing (e.g., sensorimotor vs lexical–conceptual) rather than between skill learning and priming *within a given domain*. Work by Poldrack and others examining skill learning and item-specific priming within the same task (e.g., mirror reading, digit entry) suggests that —at least under some conditions and for some types of

tasks—skill learning and priming may partially rely on common learning and neural bases.

C. An Important Methodological and Interpretive Caveat

As noted earlier, priming refers to instances in which prior experience with a stimulus influences subsequent behavior in the absence of "conscious, intentional subjective remembering." Initial theoretical considerations and subsequent empirical findings have underscored the need to differentiate between two aspects of subjective remembering: *intention* or effort to remember and *awareness* of remembering. Particularly for individuals with intact memory function, awareness that one has experienced something previously might arise even though one did not intentionally attempt to remember, that is, one may "involuntarily" or spontaneously notice that one's current experience is similar to, or has been influenced by, past experience. Priming involves changes in performance as a result of past experience with a particular stimulus in the absence of an intention to remember. However, should persons become aware of the relation between study and test items, it is possible that they may alter the way that the task is performed, changing a nominally "implicit" test, which makes no reference to the past, into an "explicit" test, in which there are deliberate attempts to retrieve items that correspond to those encountered earlier.

Attention to the possible confounding influences of explicit retrieval is important when designing and interpreting the outcome of priming studies. Researchers have developed a number of informal procedures to reduce the likelihood of contributions from explicit volitional memory on implicit tests and have proposed various criteria to more confidently infer that observed facilitative effects reflect priming rather than explicit contamination (e.g., the effects are observed in amnesics, the effects of an encoding factor differ when retrieval is volitional or intentional relative to when it is implicit, and the effects are observed on speeded tests that make intentional retrieval difficult or less likely). Although no one of these approaches, in isolation, provides unequivocal support for the unintentional nature of memory on a given task or occasion, collectively they can provide convergent evidence of the manner in which nonconscious, unintentionally expressed memory (particularly priming) may differ from intentional, conscious recollection of the past.

II. THEORETICAL ACCOUNTS

Five general classes of accounts of repetition priming have been proposed: activation–elaboration, transfer appropriate processing, memory systems, components of processing, and perceptual or response bias.

A. Activation–Elaboration

Among the earliest accounts of repetition priming were activation theories, which posited that prior experience facilitates subsequent word processing because it temporarily activates preexisting abstract representations in semantic memory. From this perspective, identification of a word was thought to involve activating an abstract lexical–orthographic word code above some threshold, with priming reflecting experience-induced modification in the lexical entry such that repeated words were more readily activated beyond threshold. These early accounts, proposed by Graf, Mandler, Morton, and others, contrasted *activation* of abstract lexical units, which was thought to be especially important for priming, with *elaboration* (e.g., processing the stimulus for meaning or forming associations to it), which was thought to be especially important for the formation, retention, and conscious recollection of events from long-term memory.

Early activation accounts involved three central propositions: (a) the representations involved were abstract (i.e., not modality-specific), (b) activation was temporary, and (c) activation occurred only for preexisting representations. Two empirical findings conclusively refuted the early activation accounts: study-to-test modality changes (e.g., visual presentation at study, auditory presentation at test) reduce or eliminate priming on perceptual tasks, and priming effects can be observed across relatively long delays. However, later activation accounts developed by Morton, J. S. Bowers, and others are not susceptible to these objections because they propose that there are modality-specific representations (that is, different representations for visually and auditorily presented words) and have differentiated between two forms of activation (and repetition priming): short-term priming, lasting only a few seconds and reflecting the activation of abstract orthographic codes or access to abstract lexical entries, and long-term priming, which can persist for minutes, hours, or longer and is viewed as reflecting a structural change in orthographic representations. The revised activation account (see also the reactivation theory proposed by G. H. Bower),

however, has difficulty with evidence demonstrating *within-modality* specificity effects in priming, as in reduced priming for words presented in different type fonts or presented auditorily in different voices relative to items presented in the same font or same voice at study and test. These findings led to the further suggestion that there may be two different perceptual systems involved in identifying information, one abstract and one specific (possibly partially lateralized to the left and right cerebral hemispheres, respectively). Most importantly, the activation account also does not provide a clear explanation for priming effects that have been demonstrated for stimuli other than words, particularly novel forms of information for which no preexisting representation exists, including novel configurations of dots or lines, novel objects, unpronounceable letter strings, novel faces, and novel melodies. This suggests that explanations in terms of activation cannot provide a comprehensive account of all priming effects. Some further mechanism or process that allows for priming involving novel representations is also necessary.

B. Transfer Appropriate Processing

According to this view, the magnitude of priming observed on a given task depends on the ''congruency'' or ''match'' between the forms of cognitive processing that are engaged during an individual's initial encounter with a word, object, or other stimulus (the ''study'' phase) and the forms of cognitive processing elicited during the subsequent (repeated) processing of that stimulus (the ''test'' phase). Greater match or congruency between the forms of processing engaged at study and those demanded at test yields greater priming, whereas mismatch tends to decrease or eliminate priming. (Such transfer appropriate processing—a term proposed by Morris, Bransford, and Franks in the late 1970s—is consistent with the notion of encoding specificity, proposed by Tulving and Thomson in the early 1970s with regard to retrieval processes in episodic memory.)

Much initial work from this framework focused on the broad distinction between ''perceptually driven'' tasks and ''conceptually driven'' tasks. Perceptually driven tasks, such as those requiring the identification of briefly exposed words or fragmented objects, are construed as being heavily reliant on sensory perceptual analyses, whereas conceptually driven tasks, such as producing words from a specified category (category–exemplar generation) or deciding whether words or objects belong to a given semantic category (e.g., making animate–inanimate decisions), are construed as being dependent on conceptual analyses. Roediger and colleagues proposed three primary criteria to operationally define tasks as conceptual vs perceptual, including sensitivity to (a) the ''read–generate'' study manipulation, (b) manipulations of study–test modality, and (c) levels-of processing manipulations at study. Tasks that show (a) a ''generation'' effect, that is, superior performance under conditions where the target items, during study, are not physically presented but must be produced (''generated'') by the participant in response to a conceptual cue (e.g., an antonym), (b) relative immunity to changes in study–test modality, and (c) superior performance for items that were semantically processed relative to items that were nonsemantically processed during study are considered to be conceptual tasks. By contrast, tasks that show (a) a ''reverse-generation'' effect, that is, superior performance for items that are simply read (from physically presented cues) rather than generated, (b) reduced or eliminated priming for items studied in one modality but tested in another modality, and (c) equivalent or near-equivalent levels of priming for items that were semantically vs nonsemantically processed at study are considered to be perceptual tasks.

Early proponents of the data-driven vs conceptually-driven distinction advanced this as an alternative to the implicit–nondeclarative vs explicit–declarative distinction regarding the forms of memory that were preserved vs impaired in amnesia, proposing that the performance of amnesics could be explained as involving the preservation of perceptual memory processes but impaired conceptual processing. However, extant data indicate that this distinction is an inadequate account of memory failure in amnesia (see later discussion). Nevertheless, distinctions relating to transfer appropriate processing continue to inform task analyses, particularly in across-task comparisons (under what conditions does across-task facilitation occur, what aspects of processing might be common across different tasks) and in attempts to specify the neural substrates that support specific task components.

C. Memory Systems

As noted earlier, there have been numerous attempts by psychologists, philosophers, and others to differentiate between different forms of learning and

memory that are subject to conscious awareness and recollection and forms of learning that may occur without such awareness. On the basis of dissociations in the performance of amnesic patients vs normal controls for tasks that require conscious recollection vs motor and cognitive tasks that may show facilitation from past experience without requiring such recollection, Cohen and Squire proposed a distinction between a hippocampally based declarative memory system and a procedural (or nondeclarative) memory system that supports the acquisition of skills and other forms of preserved learning in amnesics. Declarative memory includes memory for events (episodic memory) and factual knowledge (semantic memory) where these forms of information may be consciously recollected, subjected to verbal reflection, or explicitly expressed. Nondeclarative memory includes skill learning, repetition priming, and conditioning.

Other researchers have proposed similar broad divisions, such as that between explicit memory and implicit memory, but with further subdivisions of the latter. For instance, Schacter, Tulving, and others suggest that performance on (implicit) perceptual priming tests is mediated by perceptual representation systems, including a word form system (further differentiated into visual and auditory word form systems representing visual and spoken words, respectively) and a structural representation system for objects. These perceptual representation systems encode and retain presemantic perceptual information (in the form of perceptual records) about words and objects but do not represent semantic or associative information about them (though these systems are connected to semantic systems where such information is represented). By contrast, performance on conceptual priming tests is mediated by a semantic system, separate from the perceptual representation systems, and is sensitive to semantic variables. Within the domain of semantics, further distinctions have been posited such as that between lexical systems and more abstract semantics. Critically, these various memory systems are thought to differ not only in their operating characteristics but also in their underlying neural substrates.

D. Components of Processing

Although researchers began by emphasizing the broad characterization of tasks as perceptually driven vs conceptually driven, evidence from a variety of sources suggested that most tasks involved elements of both types of processing, thereby pointing to the need to move beyond this comparatively coarse characterization to a more precise understanding of which processes are necessary components of which tasks. The components of processing approach holds that memory depends on the operation of various components, depending on the particular processing demands of a task. This approach emphasizes the overlap in critical components of the task at study and test (as in the transfer appropriate processing approach), but also views the components of processing as *structural units*, each performing a particular function with some components shared across tasks and some unique to a given task (as in the memory systems approach). In this approach, the unit of analysis is not a large-scale system, but smaller components that comprise networks of connections with other components, where a single component may belong to a number of different systems.

E. Perceptual or Response Bias

Each of the three preceding theoretical accounts —transfer appropriate processing, memory systems, components of processing—provides a different framework from which to consider priming, but all three agree in their basic characterization of the phenomenon to be explained, specifically: priming reflects facilitation in the processing of repeated relative to nonrepeated items. In contrast, according to a perceptual or response bias account, as proposed and developed by Ratcliff, McKoon, and colleagues, priming does not reflect enhanced sensitivity in processing repeated materials but rather reflects a *bias* to interpret test items as previously studied materials, i.e., a change in the likelihood of guessing the studied item that gives rise both to benefits in performance (when the biased response matches the test item) and to costs (when the biased response leads to the incorrect identification of a test item that, though similar to the studied stimulus, differs from it). This bias is postulated to occur at a perceptual level in tasks such as forced-choice perceptual identification, but it is thought to occur at the response level in other tasks (e.g., a tendency to respond ''possible'' for all studied items in the possible–impossible object decision paradigm developed by Schacter, Cooper, and colleagues). Evidence that bias may *contribute* to priming phenomena has been obtained in several experiments, but there is also evidence to suggest that a ''pure bias'' account —without also allowing for changes in sensitivity—is

incomplete. Theoretical and empirical interpretations of findings in relation to bias continue to be a matter of debate but will not be reviewed further here.

III. BEHAVIORAL EVIDENCE

Studies of priming in neurologically healthy individuals, using behavioral paradigms with the dependent variables of response latency or accuracy, will be considered in relation to four types of priming: perceptual, lexical–phonological, conceptual, and new association priming. Although the distinctions between these types of priming are, in part, relatively crude, and often tasks involve more than one form of priming, they provide an initial classification that can serve as a general guide. More importantly, these distinctions appear to be necessary in mapping to the neural substrates of task performance, as indicated by a convergence of data from neuropsychological, electrophysiological, and neuroimaging methods (reviewed later).

A. Perceptual Priming

As noted earlier, perceptual tasks involve processing of stimulus form (rather than stimulus meaning). Research has primarily focused on perceptual priming for visual stimuli (visually presented words, faces, objects, patterns, nonwords) and, to a lesser extent, auditory stimuli (auditorially presented words or sounds). However, perceptual priming also may occur for tactile stimuli (raised letters, forms) and is reliably found for novel stimuli for which individuals do not have preexisting representations, such as novel dot patterns, novel three-dimensional objects, and novel letter strings that do not obey orthographic or phonological rules.

A given task is thought to involve a substantial perceptual component when priming is reduced or eliminated following changes between study and test modality (modality specificity) or symbolic form (e.g., picture to word changes; format specificity). On the basis of this criterion, a number of tasks have been found to involve a substantial perceptual component, including perceptual word identification, word-stem completion, word-fragment completion (participants are asked to generate a word that conforms to partially provided letter cues, as in –E–S—X, for ''beeswax''), picture-fragment completion (similar to word-fragment completion but involving line drawings of common objects, with some of the contours of the picture omitted), picture naming (participants are shown pictures of common objects and asked to name them), speeded word reading, naming, and lexical decision.

Although priming is reduced by changes in study-to-test modality or stimulus form on these tasks, cross-modal and cross-form priming have also been observed. For example, whereas both positive and negative findings have been reported, significant cross-modal priming has been found when the modality changes both from vision to audition and from audition to vision for perceptual identification, lexical decision, and word-stem completion. Priming from pictures at study to verbal tests (e.g., word-fragment completion) and from words to pictorial tests (e.g., generating words and then later identifying fragmented pictures) has also been found. More recently, several instances of significant cross-modal priming across vision and touch have been documented for words (words were presented as a raised line drawing on a card) and for two-dimensional and three-dimensional objects using such tasks as naming, picture-fragment completion, and object decision.

Cross-modal and cross-form facilitation may be attributable to explicit recollective processes (i.e., ''explicit contamination'') or to nonperceptual abstract aspects of processing that are repeated across the modality change. These latter processes may include lexical access or selection for words and access to structural descriptions for visually and haptically processed common objects (structural representations may be abstracted similarly across these modalities). For haptic-to-visual priming, results showing that cross-modal priming is not affected by elaborative processing, whereas explicit memory is, argue against an explicit contamination account. For words, evidence that cross-modal priming is not enhanced by deep compared with shallow processing—provided that the shallow encoding task requires lexical access—likewise argues against an explicit contamination account and suggests that priming operates at the lexical level. Consistent with this interpretation, neuropsychological data from amnesic patients also indicate that cross-modal priming may not necessarily reflect intentional, conscious retrieval as amnesics show normal cross-modal word-stem completion priming.

The magnitude of perceptual priming may also be evaluated in relation to changes in stimulus form *within* a given modality, as where visually presented words are presented in different type fonts, auditorially presented words are presented in different voices, or alternative exemplars or tokens of an object are presented (e.g., two different chairs). The extent to

which such item-specific changes in stimulus form attenuate priming depends on the task and stimuli. For example, for visually presented common objects, changes in the size, right–left reflection, texture, color, contrast polarity, or illumination of the object have little effect; however, changes in exemplar form (different tokens of an object) consistently produce attenuation of priming (see Fig. 2 for a neural correlate). The insensitivity of object priming to changes in size is particularly noteworthy because changes in this attribute have been found to significantly depress conscious recognition of studied items (though size changes may also affect object priming under conditions where size is highly relevant to the task). For words, changes in font, size, and script often have little effect on such tasks as lexical decision and naming; however, specificity effects may be observed for words presented in unusual or nonstandard fonts or under study conditions that focus attention on the visual appearance of the words. Additionally, the direction of the difference between conditions even in situations involving less salient changes, although not significant, is most often consistent with specificity (see Fig. 1 for relevant neuropsychological data). Specificity effects also tend to emerge more frequently for stimuli that are presented in a degraded or fragmented form.

In contrast to the clear detrimental effects of modality change and stimulus form that are observed on perceptual implicit tests, changes in stimulus form often show relatively little effect on explicit tests of recall and (somewhat less consistently) on explicit tests of recognition, though these tests markedly benefit from semantic processing. However, although many perceptual priming tests show relatively modest or nonsignificant effects of level of processing, particularly compared with the extremely robust effects of this factor on tests of conscious recollection, meta-analyses indicate that it is not possible to conclude that perceptual priming tests are entirely immune to the effects of encoding manipulations: small, often statistically insignificant, but nonetheless relatively consistent benefits for semantic compared with nonsemantic processing on perceptual tasks may, in some instances, reflect explicit retrieval strategies but may also reflect contributions of semantic or conceptual processes to task performance under certain conditions.

B. Lexical–Phonological Priming

Models of single-word reading distinguish between different possible stages of processing, including visual feature analysis, and access to a word-form (lexical) system, semantic system, and phonological output system. Although different models do not necessarily treat these stages as invariably sequential, there is evidence that access to the word-form or lexical system need not be accompanied by access to the semantic system. Consistent with this, repetition priming paradigms have shown asymmetrical transfer on a number of tasks, such that tasks that depend on access to *both* the lexical and the semantic systems (e.g., judging whether something is man-made or is abstract or concrete) may show transfer benefits to a task requiring only access to the lexical system (e.g., lexical classification), but tasks that require *only* access to the lexical system do not show transfer benefits to tasks requiring access to both the lexical system and the semantic system. Transfer across different tasks that require access to the same level, as in two lexical tasks such as lexical classification vs naming, also has been observed, but such across-task facilitation has been found to be less strong than the facilitation found for within-task repetition. Moreover, there is evidence pointing to a specific contribution of phonological similarity to repetition priming: for example, priming transfers across interpretations of a homophone, that is, words that are pronounced identically but that differ in both spelling and meaning (e.g., seeing or hearing "week" primes both "week" and "weak"). Depending on task conditions, multiple aspects of the representation of a word—lexical, orthographic, phonological, and semantic (discussed next)—may contribute to repetition priming.

C. Conceptual Priming

Conceptual priming tasks require access to semantic memory, including abstract knowledge about the structural, functional, and associative information about a stimulus: how the object is constructed or formed, its uses, and the situations or contexts in which it occurs. In a strict sense, conceptual tests might be construed as those in which there is *no* overlap of perceptual information between the target items presented in the initial study exposure and the cues presented in the subsequent test exposure, as in the case of tests of general knowledge (e.g., at study, exposed to the word "cheetah;" at test, queried "What is the fastest animal on earth?") or category–exemplar generation (e.g., at study, exposed to the word "mango;" at test, given the category name "fruit" and asked to generate appropriate exemplars of that

category). However, the term conceptual priming also has been used more broadly to refer to various tasks in which performance during the test phase relies on semantic analysis, including semantic classification or verification tasks (e.g., tasks requiring animate–inanimate or abstract–concrete decisions), and in which, depending on experimental conditions, the item may recur in the same perceptual form at study and test. Support for the notion that it may, nonetheless, be appropriate to consider the latter type of task as "conceptual" is provided by evidence that priming in this instance is largely unaffected by a change in study modality or format but is influenced by levels of processing. Neuropsychological and neuroimaging findings (see later discussion) also support this distinction.

An important question regarding conceptual priming concerns the extent to which it can be dissociated from explicit memory. Because explicit memory performance is often enhanced through deeper, "conceptual" analysis, in many instances manipulations of conceptual processing affect performance on implicit conceptual and explicit tasks in a parallel manner, raising the possibility that there is no need to differentiate between them. However, as noted later, several neuropsychological dissociations between explicit memory and conceptual priming have been observed. Moreover, dissociations between category–exemplar priming and explicit memory performance have been observed in young normal controls relative to older normal controls. Compared to normal young controls, older normal controls show impaired category-cued recall (explicit test) but intact category–exemplar priming (conceptual implicit test), together with normal benefits (increases) in both category-cued recall and category–exemplar priming for semantic relative to nonsemantic encoding.

Further evidence that explicit retrieval (indexed by category-cued recall) and conceptual implicit memory (indexed by category–exemplar generation) may reflect distinct processes even in individuals with intact memory includes the following: (a) findings of a "picture-superiority effect" (i.e., enhanced performance for items studied as pictures relative to items studied as words) during explicit retrieval but not in conceptual priming, (b) results showing that, whereas exemplar frequency and retention interval (varied between immediate testing and 24-hr delay) influenced explicit recall, they did not affect conceptual priming, (c) findings that perceptual interference at study (presenting words very briefly followed by a mask) enhanced explicit retrieval but had no effect on conceptual implicit memory, (d) results showing within-category serial position effects for explicit recall (greater recall for the first and second items that were presented in a category at study than for the later items when categories were studied in blocks) but not for conceptual priming, and (e) pharmacological evidence demonstrating preserved conceptual priming following the administration of the benzodiazepine lorazepam but impaired explicit retrieval.

Earlier, several criteria used to designate a task as conceptual were outlined. The feasibility of this approach, however, has been undermined by the finding that the various experimental manipulations believed to indicate conceptual processing do not invariably converge for a given task. For example, although a number of sources of evidence point to a conceptual processing component in word-stem completion priming (e.g., there is often a strong cross-modal component), this form of priming is not always enhanced by semantic encoding and is not greater after generation than after reading. Other conceptual tasks have also been shown to dissociate in relation to conceptual elaboration, with some tasks showing enhanced priming with conceptual elaboration (e.g., word-cued association with weak associates and category-cued association) but others not showing such enhancement (e.g., word-cued association with strong associates, category verification). This suggests the need to further differentiate types of conceptual processing.

One distinction focuses on tasks that involve semantic *verification* or classification as opposed to the actual generation or *production* of a response. Consistent with this, whereas semantic processing at study enhanced priming on production tasks, it did not facilitate verification responses; likewise, division of attention at study did not affect verification responses. However, there also is evidence of differences within conceptual production tasks: whereas semantic encoding and division of attention may not affect priming on the word association task when the words are highly associated (e.g., TABLE–CHAIR), semantic encoding does yield benefits for words that are weakly related (e.g., TABLE–STOOL).

Rather than distinguishing between verification vs production, an alternative account of these differing patterns focuses on the question of whether the task involves response competition during retrieval: such competition may be minimal or absent in cases where the stimulus is itself presented at test (in verification) or where the target word is strongly associated with the cue word. By contrast, response competition may be considerable in exemplar production tasks where the category cue is consistent with multiple legitimate

responses or for weak associates. Whereas, under conditions where the retrieval cue directly activates relevant knowledge, elaborative processing during study might not be needed or useful, under conditions where the retrieval cue is consistent with multiple responses, elaborative encoding may make a difference. This distinction between competitive and noncompetitive access also coheres with findings in neuropsychology and neuroimaging. For example, compared with matched controls, patients with focal left inferior frontal lesions made more errors on a verb generation task (generate a verb associated with a noun) when the noun had multiple possible responses (high demands on selection or competitive access) relative to when the noun had few possible responses (low demands on selection).

One further posited differentiation between conceptual tasks focuses on the distinction between item-specific encoding (processing related to individual items that tends to enhance the differentiation of that item from others) and relational encoding (processing emphasizing features that are shared among items). It has been suggested that, whereas explicit tasks can benefit from both types of processing, some conceptual implicit tasks rely especially on relational processing. This distinction suggests that minimal effects of divided attention on category–exemplar generation under conditions where the category items are blocked at study, but detrimental effects when category items are presented randomly, may arise because blocking allows the extraction of relational information even under division of attention. A somewhat similar distinction is that between initial interpretative encoding (encoding that occurs during perception) and elaborative encoding.

D. New Association Priming

The question of whether facilitation of performance occurs for "new associations," i.e., novel pairings of stimuli or stimuli that provide a context for one another, is important regarding the generalizability of priming. Measures of new association priming are based on the comparison of performance for repeated items that occur accompanied by the same item as at study (intact pairs) relative to items that are paired with different study items (recombined pairs) with findings of greater facilitation for intact than for recombined pairs taken as evidence for the formation, retention, and use of associative or contextual information.

Early studies of new association priming primarily used a word-stem completion task in which individuals initially study unrelated pairs of words (e.g., MUSTARD–SATELLITE, TABLE–REASON) and then are tested, using either explicit cued-recall instructions or implicit word-stem completion instructions, for intact pairs (MUSTARD–SAT___), recombined pairs (TABLE–SAT___), and entirely new pairs. These initial studies suggested that new association priming only emerged under conditions of elaborative encoding, specifically when the task required relational encoding of the items in the pair (e.g., forming a sentence using the words) rather than encoding that focused on each item individually (e.g., judging each word for pleasantness). These findings suggested that new association priming might require the "unitization" of stimuli in memory, such that items or components formed a single unit. Subsequent work has, however, shown that new association priming can occur even under shallow encoding, at least under conditions where the items in the pair are presented simultaneously rather than sequentially (e.g., for the stem completion task, copying the two words next to one another at study; for lexical decision, deciding whether the two words in a pair have the same number of vowels at study).

Depending on the nature of the task, there is evidence pointing to both perceptual and conceptual contributions to new association priming. For example, evidence of a perceptual contribution has been reported for the relatively perceptual tasks of word-stem completion and lexical decision, including attenuation or elimination of associative priming for unrelated word pairs under cross-modal relative to unimodal study–test conditions, and little effect of levels-processing on lexical decision. Evidence for perceptual or other lower level contributions to new association priming (such as factors relating to the phonological or articulatory production of items) also has been reported for a word reading task, such that interposition of the irrelevant intervening word "and" between the members of each word pair—while keeping the pair and order within the pair intact—virtually eliminated the new association priming effect. However, findings pointing to a conceptual contribution to new association priming have been reported both for the word-stem completion paradigm (performance was influenced by the inclusion of conceptual context words that were not presented during study though they biased interpretation toward one of two possible meanings of a homograph, both of which had been presented at study) and for a more conceptual

task (semantic relatedness judgments: no attenuation of associative priming for unrelated word pairs under cross-modal conditions).

Several early studies, particularly those using the word-stem completion paradigm, raised the possibility that new association priming might only be found in individuals with intact memory who were "test aware" (i.e., aware of the purpose of the task). However, because several subsequent studies have shown new association priming in neurologically intact individuals using tasks involving speeded responses (e.g., masked word identification, reading time for words and nonwords, color naming for compound nonwords and novel line figures), it is unlikely that new association priming entirely or exclusively reflects conscious retrieval processes. Other evidence, including differences in response times as a function of task (longer for explicit recognition than for lexical decision) and as a function of encoding condition and task (little effect of level of processing on associative priming in lexical decision but a strong effect on explicit recognition), also suggests that associative priming does not always reflect explicit contamination. (Even for the associative stem completion paradigm there is evidence of a dissociation between explicit and implicit tests as a function of encoding task.)

That said, evidence for new association repetition priming in neurologically intact individuals is less extensive and more qualified than for item-specific repetition priming, with this form of priming appearing to be quite sensitive to the particular parameters used at study and test (e.g., encoding tasks that encourage some form of relational or at least "integrative" processing; simultaneous presentation of the new associate pair at both study and test; for unfamiliar or novel stimuli, physical integration of the to-be-associated elements). There are also several demonstrations that new association priming—at least in the word-stem completion paradigm—does not occur in densely amnesic patients, although the evidence here, too, is not altogether straightforward, and associative priming *has* been found in amnesics using other tasks such as word identification, color word naming, and reading time (see later discussion).

IV. NEUROPSYCHOLOGICAL EVIDENCE

Arguably, perhaps the most fundamental development in the theoretical understanding of memory's organization was the observation that lesion or insult to specific regions of the brain can result in the impairment of some forms of long-term memory, while leaving other forms preserved. As noted earlier, the initial influential reports of such dissociations between forms of long-term memory began to appear in the 1960s and early 1970s. These studies revealed that, despite considerable impairments on explicit tests of memory, individuals with global amnesia learn new motor skills and sometimes demonstrate intact levels of facilitated word-stem completion due to recent experience with a word. These observations motivated extensive efforts in the 1980s and 1990s to understand the functional and anatomical organization of long-term memory, with significant insights being derived through studying patients with focal lesions and patients with neurodegenerative disorders. Emerging from this literature was evidence indicating that perceptual and conceptual forms of repetition priming are neuroanatomically dissociable from declarative memory, from each other, and from other forms of nondeclarative memory.

A. Perceptual Priming

Beginning with patient H.M., considerable research has revealed that declarative memory is subserved by a medial temporal and diencephalic neural network that includes the hippocampal formation, adjacent parahippocampal, entorhinal and perirhinal cortices, and medial thalamus and mammillothalamic tract. Insult to these structures results in a global amnesia characterized by a specific impairment of declarative memory, as evidenced by deficits on explicit memory tests.

Extensive neuropsychological evidence demonstrates that, in contrast to declarative memory, perceptual priming is not dependent on medial temporal structures. Since Warrington and Weiskrantz's initial reports that amnesics demonstrate intact priming on word-stem completion, numerous studies have extended these observations by revealing intact levels of perceptual priming on various implicit memory measures. For example, amnesics demonstrate normal levels of priming on visual and auditory word identification, picture naming, and word-fragment completion measures. Moreover, on implicit tests that index perceptual representations, amnesics demonstrate an intact sensitivity to the degree of perceptual match between study and test stimuli, showing reduced priming when there is mismatched study–test modality

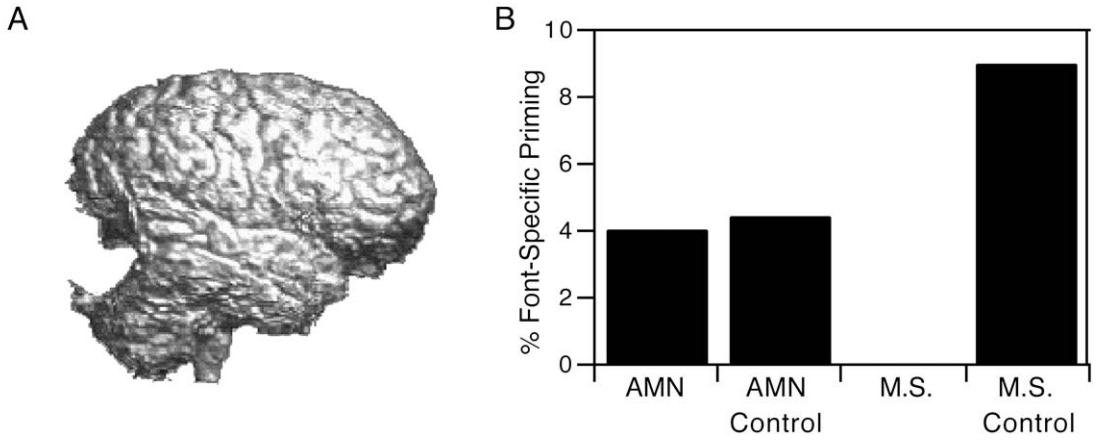


Figure 1 MRI rendering of patient M.S.'s right cerebral hemisphere and measures of font-specific behavioral priming. (A) The MRI reveals the extent of right occipital cortex resected in M.S. (B) Research by Vaidya, Gabrieli, and colleagues has revealed font–specific priming (same font-different font) in amnesic patients (AMN) and their controls (AMN control), as well as in the controls for patient M.S. (M.S. control), but not in M.S. himself.

and when the font in which a word is presented differs across study and test exposures (see Fig. 1).

Perceptual priming in amnesia is not restricted to the priming of stimuli that have preexisting memory representations. For example, levels of priming for novel visual patterns, unfamiliar objects, and novel verbal units (e.g., pseudo-words and orthographically illegal nonwords) are comparable in amnesics and healthy controls. Similarly, amnesics have been observed to demonstrate intact "new association" priming on word identification tasks, where greater priming is observed for previously encountered word pairings. However, priming of novel associations is limited following medial temporal damage. For example, data suggest that healthy individuals demonstrate priming for visual contextual information as revealed through facilitated visual search performance when the context is identical across initial and repeated search trials. Amnesic patients, in contrast, fail to show such priming, which is thought to require the formation of associations across multiple cues.

From a "memory systems" perspective, intact perceptual priming in amnesia has been hypothesized to reflect the operation of a distinct form of long-term memory from that supporting performance on explicit memory tasks. Alternatively, some theorists have suggested that intact priming derives from residual declarative memory in amnesia, rather than from nondeclarative memory. Multiple lines of evidence appear to be inconsistent with the latter interpretation, including the observation by Squire and colleagues that a patient (E.P.) who suffers from such a dense amnesia that he does not demonstrate above chance recognition memory nevertheless shows entirely intact levels of perceptual priming. Thus, perceptual priming can remain intact despite evidence suggesting an absence of residual declarative memory. Such data are consistent with the proposal that perceptual priming derives from perceptual representation systems that process information about the form and structure of stimuli. From this perspective, priming reflects the "tuning" of such a system as a result of prior perceptual processing of a stimulus. Importantly, this tuning or forming of perceptual nondeclarative memory traces does not depend on the medial temporal memory system implicated in declarative memory.

Perceptual priming appears to be subserved by modality-specific sensory cortices, with insult to such regions resulting in a selective impairment of modality-specific nondeclarative memory. For example, lesion of right visual cortex yields spared declarative memory but impaired visual priming. Specifically, in the mid-1990s, Gabrieli and colleagues reported a patient (M.S.) who, due to an extensive lesion of the right occipital lobe, fails to demonstrate visual repetition priming but who nevertheless demonstrates normal visual recognition memory. M.S.'s visual repetition priming deficits include: (a) the failure to demonstrate visual word-identification priming, (b) an absence of modality-specific priming in visual word-stem completion but normal across-modality word-stem priming

(suggesting that priming at an abstract lexical level is intact), and (c) the failure to demonstrate font-specific priming in word-stem completion (see Fig. 1). This dissociation suggests that a nondeclarative perceptual memory system in the right occipital cortex may be involved in memory for visual word form information that is necessary for visual repetition priming but not for visual declarative memory. Moreover, this dissociation between impaired perceptual priming and intact explicit memory complements the reverse dissociation in the densely amnesic patient E.P., providing further evidence against the interpretation that perceptual priming in amnesia reflects residual declarative memory.

B. Conceptual Priming

Whereas some theorists have interpreted intact perceptual priming in amnesia as indicative of multiple long-term memory systems, others have posited an alternative "process" account. This alternative perspective, which derives from the transfer appropriate processing hypothesis, posits that the dissociation between perceptual priming and performance on explicit memory tests reflects a distinction between perceptual and conceptual processes in long-term memory, rather than multiple forms of long-term memory. Accordingly, the explicit memory deficits observed in amnesia are thought to arise from the fact that these memory measures are heavily dependent on conceptually driven processes, which are posited to be impaired following medial–temporal insult, rather than perceptually driven processes, which are posited to be preserved in amnesia.

This position initially appeared viable because most early implicit tests (on which amnesics showed intact performance) were perceptual in nature, whereas the explicit tests (on which amnesics showed impaired performance) were conceptual in nature. However, findings from studies that examined amnesics' performance on all four types of test (i.e., implicit and explicit perceptual tests and implicit and explicit conceptual tests) unambiguously demonstrated that, contrary to this account, amnesic patients show intact conceptual priming (e.g., priming on word–associate generation, category–exemplar generation, and conceptual classification) and impaired performance on perceptually cued tests of conscious recall or recollection (e.g., graphemically cued recall: study the word "treason," at test probe with the visually similar but semantically unrelated word "treasure"). Amnesics also show sensitivity to manipulations of conceptual priming that parallel those found in neurologically intact individuals (e.g., conceptual priming that is uninfluenced by modality change, but sensitive to changes in semantic processing). Thus, extant data indicate that both conceptual and perceptual priming remain intact following medial–temporal and diencephalic insult, revealing dissociations between these memory phenomena and declarative memory.

Conceptual priming appears to be dependent on amodal association cortices. In contrast to amnesia, individuals with Alzheimer's disease (AD) demonstrate impaired explicit memory and impaired conceptual priming together with preserved perceptual priming (two exceptions to impairment of conceptual priming in AD may include category exemplar *verification* and word association for highly associated word pairs). AD is characterized by pathology to frontal, temporal, and parietal neocortical areas, as well as to medial–temporal regions, with a relative sparing of sensory cortices, including occipital regions. This pathology, combined with the accompanying behavioral pattern, suggests that processes mediated by frontal and temporal regions may support conceptual priming phenomena. Thus, across studies of patients with AD, amnesia, and focal lesions to sensory-specific cortices, neuropsychological evidence points to single and double dissociations between conceptual priming, perceptual priming, and declarative memory. Moreover, the preservation of skill learning, which is impaired following the disruption of basal ganglia structures as in Huntington's disease, in amnesia and AD supports the hypothesis that conceptual and perceptual priming reflect the operation of nondeclarative memory systems that are distinct from those supporting skill acquisition.

V. ELECTROPHYSIOLOGICAL AND NEUROIMAGING EVIDENCE

Neuropsychological investigation of the lesioned or injured brain has been the keystone of the cognitive neuroscientific revolution in understanding human memory. Patient studies have provided critical evidence regarding the forms of memory that are impaired and those that remain intact following a specific neural insult. However, whereas lesion studies remain a principal means for determining whether a neural region is necessary for a specific form of memory, the inferences permitted by this approach are limited for a number of reasons, including (a)

difficulties associated with drawing inferences from small samples, (b) interpretative complications arising from possible neural plasticity or reorganization, (c) challenges due to the fact that lesions often do not obey functional boundaries, and (d) the absence of information about the timing of neural processes. In the 1990s, complementary evidence about mind–brain relations, including the neural underpinnings of priming, began to be consistently derived from other approaches, including electrophysiological studies with nonhuman and human primates and neuroimaging studies with humans.

A. Electrophysiological Studies in Nonhuman Primates

One approach to studying memory processes in the primate brain is to use implanted microelectrodes to directly record from single (or, more recently, multiple) neurons while a nonhuman primate is engaged in a memory task. For example, to investigate visual memory, researchers have frequently recorded from neurons in the inferior temporal cortex while a monkey performs some variant of the delayed matching-to-sample (DMS) task. In the DMS task, a sample stimulus is initially presented followed by some number of intervening stimuli and then a repetition of the sample stimulus. Depending on the variant of DMS, all of the intervening stimuli may be novel (i.e., none of them repeat), or some of the intervening stimuli may themselves repeat prior to the repetition of the sample stimulus. In both variants of DMS, the monkey's task is to respond whenever a test stimulus matches the sample stimulus. Thus, only repetitions of the sample stimulus are task-relevant and should elicit a behavioral response from the monkey; repetitions of the intervening stimuli are irrelevant and should be ignored.

With respect to priming, the central outcome from such electrophysiological studies of DMS is the observation that many inferior temporal neurons, and to a lesser extent prefrontal neurons, demonstrate reduced or suppressed neural firing rates during repeated stimulus presentation relative to initial stimulus presentation. These changes in neural firing have been observed very early in the processing stream, with onset occurring within 100 msec or less after presentation of the repeated stimulus. Critically, in the mid-1990s, Miller and Desimone demonstrated that this "repetition suppression" is observed for both task-relevant stimuli (i.e., matches to the sample stimulus) and task-irrelevant stimuli (i.e., matches to an intervening stimulus), indicating that these repetition-induced reductions are not necessarily dependent on goal-related processes. Indeed, repetition suppression is observed even when the primate is not actively engaged in a task (e.g., during passive fixation of visually presented stimuli or when the animal is under anesthesia).

As noted by Wiggs and Martin, accumulating evidence across extant single-unit studies of repetition suppression points to similarities between this neural phenomenon and the behavioral manifestation of priming. First, repetition suppression appears to generalize across retinal locations and is observed following changes in stimulus size, indicating that these reductions, as with behavioral perceptual priming, derive from higher level representations rather than low-level visual representations. Second, repetition suppression is stimulus-specific rather than a general change associated with performance of a task. Third, suppression effects have been observed following long temporal lags between the two presentations of a stimulus, surviving lags of 24 hr (the longest tested) and the presentation of more than 100 intervening stimuli. Fourth, repetition suppression appears to be intact even when operations thought to be dependent on medial–temporal structures are impaired. For example, repetition suppression is intact when monkeys are administered scopolamine, which is thought to hinder declarative memory operations.

One account of repetition suppression is that it reflects a tuning of the neural system such that a smaller population of neurons is activated during the repeated presentation of a stimulus relative to its initial presentation. This trimming of the neuronal representation of a stimulus has been posited to reflect a sharpening within long-term memory due to the initial processing of the stimulus. This sharpening may give rise to more efficient neural computations during subsequent stimulus processing, and this increased neural efficiency may yield the behavioral manifestations of priming (decreased response latencies and increased accuracy). Although additional evidence is required, this putative link between repetition suppression in the nonhuman primate and behavioral and neural priming effects in humans (see later discussion) is theoretically enticing.

B. Functional Neuroimaging

In the human primate, direct recording of single-unit activity is constrained for obvious reasons. However,

investigation of the neural correlates of priming in the human brain has dramatically advanced over the past decade partially due to the application of more recently developed functional neuroimaging methodologies. Two principal neuroimaging methods—positron emission tomography (PET) and functional magnetic resonance imaging (fMRI)—have been adopted to explore the relation between neural and cognitive function, including studies of priming and the brain. As is reviewed elsewhere, PET and fMRI provide indirect measures of neural activity by indexing changes in regional cerebral blood flow (PET) or blood-oxygenation levels (fMRI) that are correlated with changes in neural firing. These techniques have proven to be powerful because they provide a means of determining structure–function relations in the healthy human brain with relatively high spatial resolution. However, unlike electrophysiological and magnetoencephalographic approaches, these methods do not provide millisecond temporal resolution as they depend on a temporally sluggish hemodynamic response.

1. Perceptual and Lexical Priming

Over the past decade, neuroimaging has been adopted to investigate the neural correlates of multiple forms of nondeclarative memory, including phonological, lexical, and conceptual priming. The most frequent observation across PET and fMRI studies of priming is that the specific brain regions that are recruited during initial (unprimed) stimulus processing demonstrate decreased activation during repeated (primed) stimulus processing. For example, in the early 1990s, Squire and colleagues reported the first neuroimaging observation of priming-related reductions in neural activity. In this landmark PET study, neural activations during the performance of two visual word-stem completion conditions were compared. In the primed condition, subjects were presented word stems that could be completed with words that had been encountered visually prior to PET scanning, whereas in the unprimed condition the stems could not be completed with a previously encountered item. As in behavioral studies of word-stem completion priming, here subjects were faster at completing the primed relative to the unprimed stems, and primed stems were completed more often with target items than were unprimed stems. Functional neuroanatomically, the PET data revealed a neural correlate of these behavioral priming effects: there was less activation in extrastriate visual cortex during primed relative to unprimed stem completion.

Subsequent PET and fMRI studies of perceptual priming have extended this initial observation of reduced posterior cortical activation during primed stimulus processing. As with visual stem completion priming, reduced extrastriate activation accompanies primed relative to unprimed visual word-fragment completion. Moreover, decreases in activation have been observed in fusiform and inferior temporal regions during repeated relative to initial processing of visually presented objects. For example, fMRI work in our laboratory has revealed that, when subjects attend to a novel visually presented object, there is greater activation in bilateral fusiform and inferior temporal cortices relative to when they attend to an object that has been encountered recently. However, left and right inferior temporal regions do not demonstrate identical priming patterns. As Fig. 2 illustrates, whereas both left and right fusiform gyri demonstrate priming reductions during repeated relative to initial object processing, the right fusiform gyrus is more sensitive to the recapitulation of specific perceptual details than is the left fusiform gyrus, suggesting a differential role for the two hemispheres in visual object priming.

Posterior cortical activation reductions during repeated stimulus processing have been interpreted as neural correlates of visual repetition priming. As with behavioral priming, these changes in neural firing reflect item-specific mnemonic effects rather than general task learning, because they derive from a specific prior encounter with an item. Moreover, these neural priming effects survive filled retention delays of multiple minutes, indicating that they depend on long-term memory representations. It has been hypothesized that these neural priming effects reflect the facilitated reprocessing of the visual form of words or the structural characteristics of objects as a consequence of nondeclarative memory for prior perceptual processing. These activation reductions resemble the phenomenon of repetition suppression observed in electrophysiological studies with nonhuman primates. As with repetition suppression, neuroimaging demonstrations of priming-related activation reductions are consistent with a representational sharpening mechanism, where the neural representation of an item becomes sparser with experience. This tuning may result in only a subset of the neurons that are initially involved in processing the perceptual properties of the item being engaged during subsequent reprocessing of these perceptual attributes. Thus, one possibility is that the reduction in activation observed in PET and fMRI priming studies reflects a

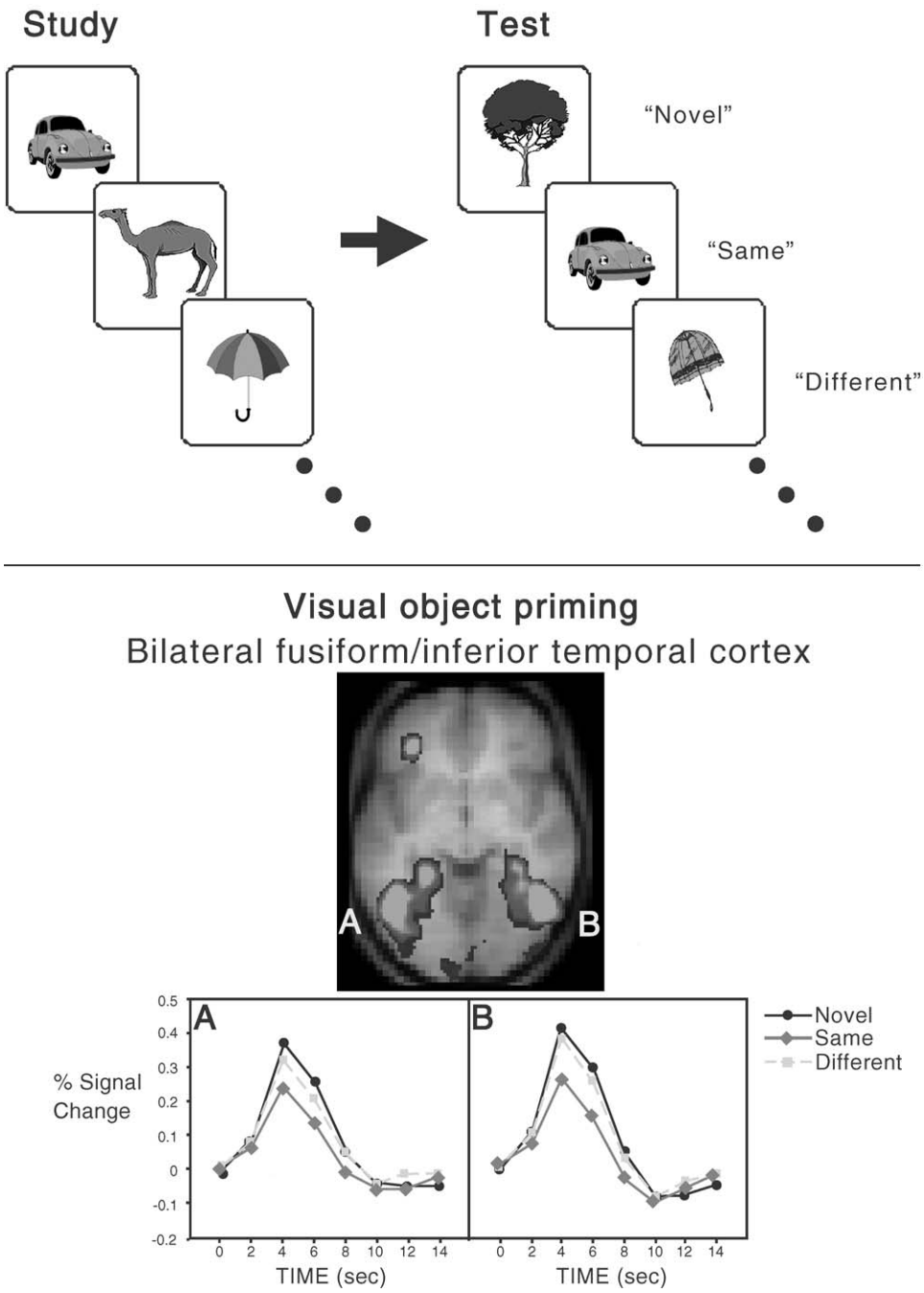


Figure 2 Repetition priming for objects is associated with reduced activation in fusiform and inferior temporal cortices. At top is the task paradigm illustrating the relation between study and test exemplars for the novel, same repetition, and different repetition conditions. At bottom are statistical activation maps showing left (A) and right (B) fusiform–inferior temporal regions that demonstrated reduced activation for both types of repetition trials. The plotted measures of neural activation in these regions revealed that, whereas there was reduced activation in bilateral fusiform regions for both same and different exemplar repetition trials (relative to novel trials), the right fusiform was particularly sensitive to study–test perceptual match (as revealed by very modest priming in this region for different repetition items).

decrease in the number of neurons engaged during repeated stimulus processing. Alternatively, it is possible that the number of neurons engaged does not change, but rather there is a decrease in either the rate or duration of firing of a static population of neurons with repetition.

Neuroimaging has begun to highlight the neural correlates of priming deriving from more abstract, nonperceptual lexical representations. The observation in Fig. 2 that priming in the left fusiform cortex is less sensitive to changes in perceptual detail between the initial and repeated encounter of an object suggests that priming in this region may derive from more abstract, nonvisual representations. Other fMRI evidence converges with this interpretation. For example, activation reductions in a very similar left fusiform–inferior temporal region (as well as in left inferior frontal cortex) have been observed in both visual-to-visual and auditory-to-auditory word-stem completion priming. The colocalization of visual and auditory word priming in a region that also demonstrates priming across perceptually distinct objects that share a common lexical code indicates that priming in this region likely does not reflect (or solely reflect) the recapitulation of visual processing. Rather, left fusiform priming may at least partially derive from facilitated access to abstract lexical representations as a consequence of prior access to these representations. This interpretation is consistent with evidence from human neuropsychological and depth-electrode recording studies, which point to a role for the basal temporal region in lexical processing.

2. Conceptual Priming

PET and fMRI studies have also elucidated the nature of experience-induced changes in neural activation during comparatively more conceptually driven tasks, such as judging whether a presented word refers to an abstract or concrete entity or generating a semantic associate of a stimulus. Considerable evidence indicates that, relative to baseline tasks, semantic processing tasks, which typically require access to and evaluation of both semantic and phonological stimulus attributes, are associated with activation in the posterior and anterior left inferior prefrontal cortices (LIPCs). Posterior LIPC activation has been shown to be more sensitive to the processing of phonological or lexical codes associated with a stimulus, whereas anterior LIPC activation has been shown to modulate as demands on semantic attribute processing vary.

Both posterior LIPC [at or near Brodmann's areas (BA) 44/6] and anterior LIPC (at or near BA 45/47) demonstrate decreased activation during repeated relative to initial semantic processing of a stimulus. In the first neuroimaging report of repetition-induced changes in LIPC activation, Raichle and colleagues observed greater LIPC activation during the initial (unpracticed) generation of semantic associates of nouns relative to that found for highly practiced generation of the associates (participants had previously practiced generating the associates nine times). Subsequent studies extended this result to different semantic processing paradigms and various classes of stimuli and demonstrated that LIPC priming effects can be observed at the individual subject level even following a single presentation. For example, as Fig. 3 illustrates, reduced LIPC activation accompanies semantic classification of a word that was primed by a single prior classification of the item. Whereas these reductions decline with the temporal delay between the initial and repeated presentations of an item, nevertheless they appear to be long-lasting, having been observed even after 24 hr delays (the longest delay tested). Importantly, although only limited evidence is currently available, initial investigations of the status of LIPC priming in amnesic patients indicate that amnesics also demonstrate repetition-related reductions in frontal cortex. These latter data are consistent with the hypothesis that these changes reflect nondeclarative memory.

Available data do not directly implicate a specific underlying mechanism through which prior experience results in decreased LIPC activation. However, a number of theorists have posited that these reductions reflect decreased demands on prefrontally mediated control mechanisms due to nondeclarative memory. Specifically, activation in inferior prefrontal cortex during conceptual processing tasks may reflect directed attention to task-relevant stimulus features (e.g., those semantic features that enable the semantic classification of a word). This allocation of attention to task-relevant features may result in the subsequent priming of these features. During reprocessing of the stimulus, implicit memory representations for the initial processing may serve to make the task- or goal-relevant features more readily available than task-irrelevant features. This greater accessibility of task-relevant features would reduce the computational demands on prefrontally mediated attentional or biasing operations, thus resulting in decreased LIPC activation and faster and more accurate behavioral responses. Thus, neural priming effects may reflect

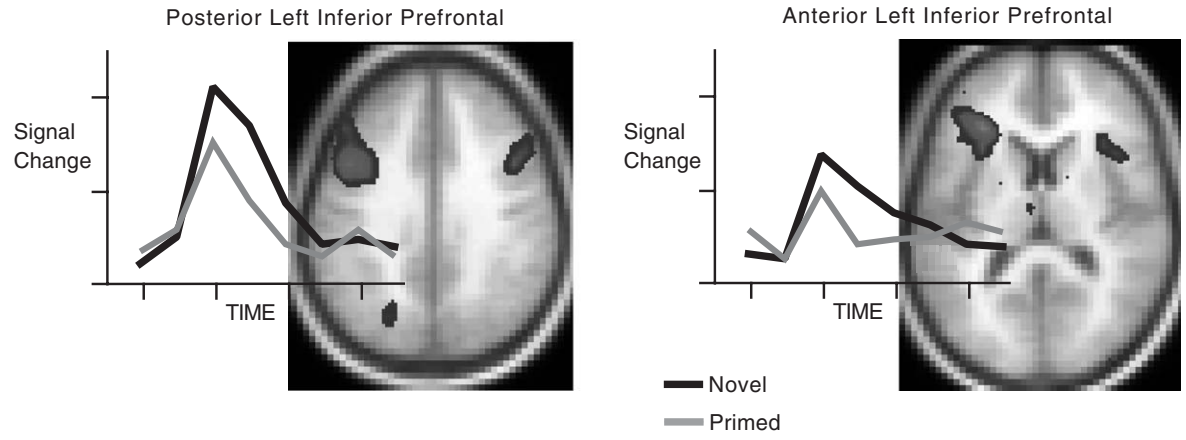


Figure 3 Conceptual repetition priming is associated with reduced activation in left frontal regions, including posterior and anterior left inferior prefrontal cortices. Displayed are the posterior and anterior LIPC regions typically observed to demonstrate greater activation (fMRI signal change) during initial (novel) semantic classification of a stimulus relative to repeated (primed) classification of the same stimulus.

enhanced efficiency in accessing, selecting, and/or evaluating target semantic knowledge necessary to achieve the goal of semantically categorizing a word, with this efficiency deriving from increased availability of the target attributes as a result of earlier processing. From this perspective, priming tends to yield a stereotyped or sparser reprocessing experience because the same semantic features are more likely to be processed during the subsequent encounter with a stimulus as were processed during the initial encounter.

C. Event-Related Potentials

A complementary approach for investigating the neural correlates of cognition in humans is the recording of event-related potentials (ERPs). ERPs, which are recorded using scalp electrodes, reflect changes in neural electrical activity (EEG) that are time-locked to the onset of a specific event. Within the context of priming paradigms, these events typically consist of the presentation of a novel or repeated stimulus. ERPs are complementary to PET and fMRI as they offer high temporal resolution (on the order of milliseconds) but poor spatial resolution. The comparison of ERPs elicited by two different experimental conditions provides measures of if and when (in terms of poststimulus onset time) the underlying neural activity associated with the conditions diverges.

A number of ERP studies have begun to unambiguously identify some of the neural signatures of priming. Common across most extant studies is the recording of ERPs during the performance of implicit or indirect memory tasks and the comparison of ERPs elicited during the processing of repeated items to ERPs elicited during the processing of novel items. Such comparisons provide candidate electrophysiological responses (ERP repetition effects) that might reflect the neural signatures of priming. The inference that a repetition effect reflects priming rather than incidental or intentional explicit remembrance can be drawn when the repetition effect varies with task parameters known to influence priming but not explicit memory (e.g., study–test modality).

Repetition effects often present as a more positive-going shift in ERPs during the processing of repeated relative to novel items (although reduced ERP amplitudes have also been observed during item repetition). For example, in the late 1990s, Paller and colleagues explored the neural correlates of visual word priming by manipulating the perceptual match between the initial and repeated presentation of a word. In this study, words initially were visually presented either in a whole-word format (e.g., APPLE) or in a letter-by-letter format (e.g., exposing each of the letters in a sequential manner: AP__, followed by __P__, __L__, and __E). During the critical poststudy lexical decision test, all words were presented in a whole-word format. The results revealed an ERP repetition effect at occipital electrode sites with onset approximately 400–500 msec poststimulus. Critically, this effect was greatest for studied items that were perceptually identical at study and test (whole-word format items) and was dramatically diminished for items studied in a

letter-by-letter format, thus revealing an effect of the match between study–test perceptual form. Interestingly, other data indicate that ERP repetition effects occur for visually integrated or unitized stimuli (e.g., words, pronounceable nonwords, and pictures) but not for stimuli that lack a higher level unitized representation (e.g., orthographically illegal nonwords and meaningless pictures).

Diminished ERP repetition effects as a result of a mismatch between study–test sensory modality also have been observed at other electrode sites and at earlier temporal latencies. Specifically, Rugg and colleagues reported more modest repetition effects occurring around 200–400 msec poststimulus over frontal and temporal scalp sites when the study–test modality of words was auditory–visual relative to visual–visual. These effects indicate that ERP-indexed neural activity can reflect sensitivities to both repetition and study–test sensory similarity, suggesting that the effects derive from perceptual repetition priming.

Although accurate source localization is difficult with ERP, the observation of ERP repetition effects in occipital and temporal regions broadly converges with visual word-form priming in extrastriate and temporal cortices, as indexed with PET and fMRI, and with impaired visual word-form priming following right occipital resection. Further linkage between these effects comes from a combined neuropsychological–ERP investigation of visual word priming on a lexical decision task, where control subjects and patients with focal lesions in right temporal–occipital cortices were compared. Relative to controls, the repetition-related positive ERP deflections and reaction time priming effects were reduced in the patients, especially when there was a moderate-to-long lag between the initial and repeated occurrences of a word. Collectively, these data further implicate right occipital–temporal cortices in visual word-form priming.

In contrast to perceptual priming, unambiguous ERP indices of conceptual priming remain elusive. The principal challenge for obtaining such evidence is the fact that behavioral task manipulations, such as varying the extent of conceptual elaboration at encoding, often have similar effects on both conceptual priming and explicit remembrance. Given the difficulty of differentially affecting conceptual priming and explicit memory and the absence of relevant ERP data from amnesic patients, researchers have yet to determine whether ERP repetition effects on implicit or indirect tests of conceptual processing reflect priming or explicit contamination. For example, ERP repetition effects have been observed using a semantic generation paradigm previously implemented in PET and fMRI studies of item repetition. As discussed earlier, PET and fMRI studies have revealed that repeated generation of a verb associated with a noun yields reduced left frontal and temporal activation relative to initial generation of the associate. In a subsequent ERP study, it was observed that, relative to ERPs elicited by word reading, ERPs elicited by novel semantic generation were more positive in left frontal regions at approximately 250 msec poststimulus onset and in left temporal–parietal regions at approximately 700 msec poststimulus onset. Broadly consistent with the neuroimaging literature, these ERP repetition effects were reduced during repeated semantic generation. However, repeated semantic generation also was associated with an additional response over right anterior frontal regions previously observed in PET, fMRI, and ERP studies of explicit retrieval, suggesting possible explicit contamination. (To date, no such increase has been reported in PET and fMRI studies of conceptual priming using this paradigm, although similar frontal activations have been reported in cross-modal stem-completion and cross-exemplar object-decision priming paradigms; it is possible that such activation reflects involuntary explicit memory). Although behavioral data indicate that amnesics demonstrate intact behavioral priming in the repeated semantic generation paradigm, in the absence of any ERP data clearly demonstrating that these repetition-related ERP reductions are dissociable from explicit memory effects, it is not possible to conclude that such ERP changes reflect conceptual priming. Thus, unambiguous ERP signatures of conceptual priming await further investigation. Such investigations will ultimately allow an increased understanding of the neural networks and processes that support facilitated performance, not only as a result of prior perceptually based processing of specific stimuli but also from prior conceptually based processing involving the retrieval, manipulation, and representation of the meaning of specific stimuli.

VI. CONCLUSIONS

The past four decades have witnessed a multileveled and ever-deepening convergence of findings that provide insight into the brain processes that subserve the apparently simple manner in which a repeated encounter with a stimulus enhances subsequent processing. Behavioral outcomes documenting increased

accuracy and/or increased speed of response to previously encountered words, objects, faces, and other stimuli may mesh with electrophysiological findings suggesting that repeated presentation of a stimulus may give rise to a form of representational tuning or sharpening, wherein the stimulus is represented in an increasingly sparse or efficient manner. Enhanced neural efficiency is also suggested by findings from functional brain imaging, showing reduced activation in brain regions that are initially involved in stimulus processing and including both modality-specific perceptual processing regions (especially visual word and object processing in, for example, regions of occipitotemporal cortex) and more abstract conceptual processes (e.g., in evaluating stimuli for classification on the basis of meaning, particularly in left prefrontal cortex). However, the precise interrelations between perceptual and conceptual processing still remain somewhat unclear, and refinement and articulation of subcomponents of processing in the conceptual domain are still necessary.

Theoretical analyses have likewise yielded increasing convergence. Although analyses from the perspective of perceptual or response bias underscore the notion that forms of facilitation may not necessarily or exclusively reflect increased "sensitivity," key notions advanced from the transfer appropriate processing perspective are routinely incorporated into task analyses and the interpretation of dissociations across tasks and populations. Theoretical accounts positing the activation of preexisting lexical–semantic representations, separable memory systems, and components of processing all—in different ways—recognize the need for further subdivision and acknowledge the possible interplay and contribution of any one neural or representational component to multiple tasks. The upcoming decades, particularly with further evidence drawn from *conjunctive application* of neuroimaging, neuropsychological, behavioral, and electrophysiological approaches, will continue to increase our understanding of the brain and different forms of long-term memory, including instances where we may benefit from or be influenced by the past—primed for responding—without intentional remembrance.

See Also the Following Articles

COGNITIVE PSYCHOLOGY, OVERVIEW • EVENT-RELATED ELECTROMAGNETIC RESPONSES • INFORMATION PROCESSING • LANGUAGE AND LEXICAL PROCESSING • MEMORY, EXPLICIT AND IMPLICIT • OBJECT PERCEPTION • PROSOPAGNOSIA • WORKING MEMORY

Suggested Reading

Bowers, J. S. (2000). In defense of abstractionist theories of repetition priming and word identification. *Psychonom. Bull. Rev.* **7,** 83–99.

Desimone, R. (1996). Neural mechanisms for visual memory and their role in attention. *Proc. Natl. Acad. Sci. USA* **93,** 13494–13499.

Fleischman, D. A., and Gabrieli, J. D. E. (1998). Repetition priming in normal aging and Alzheimer's disease: A review of finding and theories. *Psychol. Aging* **13,** 88–119.

Gabrieli, J. D. E. (1998). Cognitive neuroscience of human memory. *Ann. Rev. Psychol.* **49,** 97–115.

Moscovitch, M., Goshen-Gottstein, Y., and Vriezen, E. (1994). Memory without conscious recollection: A tutorial review from a neuropsychological perspective. In *Attention and Performance* (C. Umilta and M. Moscovitch, Eds.), Vol. 15**,** pp. 619–660. MIT Press, Cambridge, MA.

Roediger, H. L., III, and McDermott, K. B. (1993). Implicit memory in normal human subjects. In *Handbook of Neuropsychology* (H. Spinnler and F. Boller, Eds.), Vol. 8**,** pp. 63–131. Elsevier, Amsterdam.

Schacter, D. L., and Buckner, R. L. (1998). Priming and brain. *Neuron* **20,** 185–195.

Squire, L. R., and Kandel, E. R. (2000). *Memory: From Mind to Molecules*. Scientific American Library, New York.

Wiggs, C. L., and Martin, A. (1998). Properties and mechanisms of perceptual priming. *Curr. Biol.* **8,** 227–233.

Prion Diseases

HANS A. KRETZSCHMAR
Ludwig-Maximilians-Universität, München

GLOSSARY

BSE Bovine spongiform encephalopathy.

CJD Creutzfeldt–Jakob disease.

FFI Fatal familial insomnia, a hereditary human prion disease.

(n)vCJD New variant of Creutzfeldt–Jakob disease.

prion Infectious agent of prion diseases, derived from *pro*teinaceous *in*fectious particle.

Prnp The prion protein gene of mammals.

***Prnp*$^{0/0}$** Ablation of *Prnp* (in *Prnp* knockout or *Prnp*-null mice).

PRNP *The human prion protein gene.*

PrP^{C} Cellular isoform of the prion protein. PrP^{C} is a normal protein expressed on the surface of many cell types of mammals (and birds). It is a copper-binding protein whose exact function is not known.

PrP^{Sc} Scrapie isoform of the prion protein. This protein shows increased resistance to digestion with proteinase K. It is closely associated with infectivity, and in terms of the prion hypothesis it is part of the infectious agent, the prion.

PrPres Proteinase K (pK) resistant form of PrP derived *in vitro* from PrP^{C}. (PrPres to date has not been shown to be infectious. Unfortunately this term is also used by some groups to denote the pK-resistant PrP found in infectious brain tissue, thus creating confusion with PrP^{Sc}.)

scrapie Prion disease naturally occurring in sheep.

TSE Transmissible spongiform encephalopathy, often used as a synonym for prion disease.

Prion diseases are rapidly progressing, invariably fatal, neurodegenerative diseases associated with dementia and neurological deficits such as ataxia, visual disturbances, or myoclonus. Histologically, nerve cell loss, spongiform change, and various forms of prion protein deposits are found in the brain. They are a heterogeneous group of diseases that can be acquired, hereditary, or idiopathic. All prion diseases are experimentally transmissible with a relatively long incubation time and a comparatively short clinical duration. Human prion diseases were shown to be transmissible after intracerebral inoculation of brain tissue from kuru and CJD patients into chimpanzees in 1966 and 1968. The nature of the infectious agent of prion diseases such as Creutzfeldt–Jakob disease (CJD) in humans, scrapie in sheep, and BSE in cattle has been the subject of numerous studies for many years. The assumption, which seemed natural in the early 1970s, that the agent must be a slow virus or a virino was challenged by the failure to detect viral nucleic acids and by the resistance of the agent to radiation, nucleases, and other reagents that damage nucleic acids. In contrast, infectivity was closely associated with a protein, PrP^{Sc}, and if any nucleic acid is to be associated with the infectious agent it cannot contain more than 50 nucleotides. The term "prion" was proposed by Stanley Prusiner to distinguish the infectious pathogen from viruses and viroids. Albeit not formally proven, the prion hypothesis is supported by many lines of evidence. Prions were originally defined as small, proteinaceous infectious particles that resist inactivation by procedures that modify nucleic acids, and they have been redefined as a proteinaceous particle that lacks nucleic acid (Prusiner, 1998). The change in conformation of a normal protein encoded by the host genome, the cellular isoform of the prion protein (PrP^{C}), into an altered

isoform, the scrapie isoform of the prion protein (PrP^{Sc}), is the core of this hypothesis. The term PrP^{Sc} is used for the isoform of the prion protein that is closely associated with infectivity and that is part and parcel of the prion. Whether PrP^{Sc} is the only necessary constituent of prions is at present unknown. According to the prion hypothesis, the infectious agent, the prion, would require PrP^{C} molecules for its propagation, and organisms devoid of PrP^{C} should not be susceptible to prion diseases. Indeed, this has been shown experimentally by the resistance of PrP gene knockout mice ($Prnp^{0/0}$ mice) to scrapie.

I. THE PRION PROTEIN

A. PrP Structure and PrP Genes

Human PrP^{C} is a glycoprotein of 253 amino acids before cellular processing. There is an 85–90% homology to prion proteins of other mammalian species. PrP^{Sc} is a membrane protein expressed mainly in neurons, but also in astrocytes and a number of other cells. It has an N-terminal signal sequence of 22 amino acids, which is cleaved off the translation product. Twenty-three terminal amino acids are removed when glycosylphosphatidylinositol (GPI) is attached to serine residue 230. Mature PrP^{C} is attached to the cell surface by this GPI anchor and undergoes endocytosis and recycling. It seems, however, that PrP may exist in alternative membrane topologies (^{ctm}PrP and ^{ntm}PrP), whose implications for the function of PrP^{C} and its role in pathogenesis are just beginning to be elucidated. There are two N-glycosylation sites that are glycosylated differently in different human CJD variants. The N-terminal moiety of the protein contains an octapeptide repeat, (PHGGGW-GQ) $\times$ 4, which has been suggested to function in copper binding.

Whereas PrP^{C} is found on the surface of many cell types in all mammals and birds studied so far, PrP^{Sc} is generated from PrP^{C} in a posttranslational process and is closely associated with infectivity. In terms of the prion hypothesis, it is part and parcel of the infectious agent, the prion. NMR structural studies have shown that the C-terminal half of PrP^{C} contains a two-stranded antiparallel β-sheet (S1 and S2) and three α-helices, whereas the N-terminal moiety is thought to have no definite structure in aqueous solution. PrP^{C} and PrP^{Sc} seem to differ mainly in their folded structures. PrP^{C} purified from hamster brain consisted of 42% α-helical and only 3% β-sheet structure, whereas PrP^{Sc} purified from scrapie-infected hamster brain is composed of 30% α-helix and 43% β-sheet. PrP^{Sc} shows increased protease resistance, is insoluble in aqueous solution, and tends to form fibrils that show birefringence after binding of Congo red.

The human PrP gene (*PRNP*) is located on the short arm of chromosome 20. It has a simple genomic structure and consists of two exons and a single intron 13 kb in length. The entire protein-coding region is located in exon 2. In families with inherited prion diseases, a large number of different point mutations and insertion mutations have been described in the open reading frame of *PRNP* (Fig. 1). The insertional mutations are situated in the N-terminal half of the protein in an octapeptide repeat region, whereas the point mutations cluster in the central and C-terminal regions of the protein.

The common polymorphism at amino acid position 129 of the prion protein, where humans carry a methionine (M) or valine (V), clearly influences susceptibility to the sporadic and iatrogenic types of prion diseases and, furthermore, determines in part the phenotype of the sporadic as well as of some inherited prion diseases. Several studies have revealed a marked over-representation of homozygotes (mainly for methionine) at this position in cases of sporadic CJD compared to the normal population. CJD homozygotes at codon 129 also show a higher susceptibility to iatrogenic CJD and a shorter incubation time as well. There is also a strong correlation of codon 129 genotype and clinicopathological phenotype (see later discussion).

A *Prnp*-like gene was identified in the mouse genome. This gene, *Prnd*, is located 16 kb downstream of *Prnp*, and it encodes a protein, named Dpl (Doppel, German for double), that is composed of 179 amino acids. Dpl has a 25% identity with the carboxy-terminal two-thirds of PrP^{C}, but it lacks the N-terminal copper-binding octarepeat and the central region between amino acids 114 and 121 that is needed for prion replication. It is expressed at high levels in testis, at lower levels in other organs, and at very low levels in the brain of wild-type mice. It is overexpressed in certain $Prnp^{0/0}$ mouse lines lacking the entire PrP-coding region and 5′- and 3′-flanking regions. In these animals, overexpression of Dpl leads to ataxia that can be cured by expression of PrP from a cosmid introduced containing *Prnp*. The physiological function of Dpl and its possible role in prion disease are unknown.

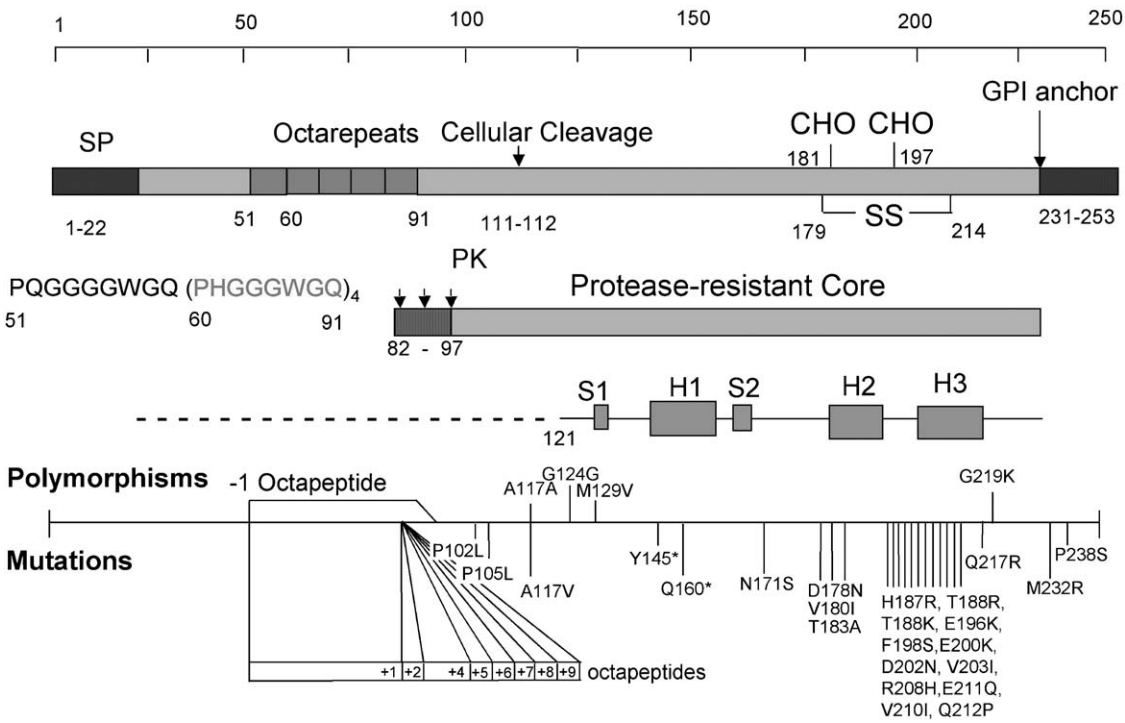


Figure 1 Structural features of the human prion protein. Shown are the cellular isoform of PrP (PrP^C), the protease-resistant core of PrP^{Sc}, and secondary structural features, as well as mutations and polymorphisms of the prion protein gene (*PRNP*). The numbers indicate amino acid residues. Abbreviations: SP, signal peptide; GPI, glycosylphosphatidylinositol anchor; CHO, glycosylation sites; S1, S2, β-sheets; H1, H2, H3, α-helices; PK, proteinase K digestion sites. The dotted line indicates the unstructured N-terminus.

B. The Physiological Function of PrP^C

By using conventional biochemical and molecular biological approaches, it has not been possible to identify a binding partner or a functional role of PrP^C. An attempt to unravel the physiological role of PrP^C in the course of infection and its normal function was undertaken by destroying the murine PrP gene (*Prnp*) by homologous recombination ($Prnp^{0/0}$ mice). $Prnp^{0/0}$ mouse lines proved to be resistant to infection with mouse-adapted scrapie. Therefore, the presence of PrP^C is necessary for agent replication. Also, neuronal loss in prion disease seems to depend on the expression of PrP^C, as shown by two different experimental approaches using transplants of normal mouse brain in $Prnp^{0/0}$ mice or primary neuronal cultures derived from $Prnp^{0/0}$ mice.

$Prnp^{0/0}$ mice appeared to be behaviorally normal. However, several subtle phenotypes were detected. Electrophysiological examination of $Prnp^{0/0}$ mice revealed slight changes in hippocampal long-term potentiation (LTP) and $GABA_A$-receptor-mediated responses. Whereas changes in LTP were confirmed in two $Prnp^{0/0}$ lines, other findings were seen in only one of the $Prnp^{0/0}$ lines created by different research groups. There is, however, morphological, biochemical, and physiological evidence that PrP^C is involved in synaptic transmission. PrP^C is known to be transported axonally to synaptic boutons, where it was localized immunohistochemically and biochemically in normal brains and where PrP^{Sc} was identified in scrapie-infected brain tissue. PrP^C influences the synaptosomal calcium concentration and calcium-dependent potassium channels. Two further phenotypes have been reported. In two of the three $Prnp^{0/0}$ mouse lines there were subtle changes in the circadian rhythm, and in the third line there was loss of Purkinje cells in aged animals.

Because the N-terminal octarepeat, (PHGGGWGQ) $\times$ 4, of recombinant PrP shows cooperative binding of 5–6 Cu(II) ions in the micromolar range compatible with estimates for copper concentrations in the synaptic cleft, comparative studies of wild-type and $Prnp^{0/0}$ mice were undertaken to investigate the

hypothesis of functional synaptic Cu(II) binding to PrP^C. Biochemical preparations of synaptosomes showed a 50% reduction in copper concentration in *Prnp*$^{0/0}$ animals. In brain slice preparations, synaptic transmission is severely disturbed by the addition of excess copper in *Prnp*$^{0/0}$ animals. Synaptic transmission in the presence of H_2O_2, which is known to be decomposed to highly reactive hydroxyl radicals in the presence of iron or copper, was found to correlate with the level of PrP^C expression in the presynaptic neuron. Further evidence of a role for PrP^C in cerebral copper metabolism comes from the finding that the activity of the copper-dependent cytoplasmic enzyme Cu,Zn superoxide dismutase (Cu,Zn-SOD) is significantly reduced in the brains of *Prnp*$^{0/0}$ mice.

C. The Conversion Process of PrP^C to PrP^{Sc}

Conversion of PrP^C to PrP^{Sc} seems to be a late posttranslational process, which in scrapie-infected cells occurs after PrP^C has reached its normal location on the cell surface or even later during endocytosis. Why this is such a rare event and how PrP^{Sc} triggers the conversion of PrP^C are not understood. In the nucleation model, PrP^C and PrP^{Sc} are in equilibrium. PrP^{Sc} is stable only when it adds onto a seed or PrP^{Sc} aggregate, a process that has been compared to crystal formation. The spontaneous formation of an initial PrP^{Sc} aggregate (seed) would be a very rare event. Once the seed has been formed, the ensuing addition of PrP monomers could follow at a fast rate. In contrast, the refolding model holds that PrP^C is unfolded and that the conversion process consists of a refolding of the molecule under the influence of a PrP^{Sc} molecule. In this process, a high activation energy barrier must be overcome and chaperones and an energy source may be required. The two hypotheses, which have also been named the Lansbury and the Prusiner mechanisms, are by no means mutually exclusive.

The conversion process has also been studied in cell-free conditions *in vitro*. Incubation of ^{35}S-labeled hamster PrP^C with a 50-fold excess of PrP^{Sc} from scrapie-infected hamster brain resulted in the conversion of some labeled protein into proteinase K (pK) resistant PrP (PrPres). Because of the large surplus of infectivity associated with PrP^{Sc} necessary to initiate the conversion process, this experimental approach has not been conducive to showing that the abnormal PrP formed *in vitro* (PrPres) is infectious. PrPres, therefore, is not necessarily identical to PrP^{Sc}. A surprising finding was that this cell-free conversion process was species-specific and correlated well with known barriers to cross-species transmission in animals. Thus, PrP^C from mouse was readily convertible to protease-resistant PrPres by murine PrP^{Sc} but poorly convertible by bovine PrP^{Sc} and vice versa.

II. PRION PATHOGENESIS

A. Neuroinvasion

Acquired prion infection such as scrapie is an oral infection; iatrogenic CJD may be peripherally acquired. For invasion of the central nervous system, two possible routes are discussed: (1) transport in blood cells, possibly after amplification of prions in the lymphoreticular system (LRS), and (2) transport in peripheral nerves. Possibly a combination of amplification in the LRS and transport by peripheral nerves is what actually happens in acquired prion disease. The importance of the LRS has been known for a long time.

In experimentally infected mice, infectivity has been shown in the spleen 4 days after intraperitoneal and even after intracerebral inoculation. In these cases, prion replication in the spleen precedes intracerebral replication even after intracerebral inoculation. In nvCJD, PrP^{Sc} accumulates in lymphoid tissues of the tonsils and in the appendix, in one case 8 months before the outbreak of clinical disease. The nature of the cells supporting prion replication in the LRS has not been established beyond doubt. Follicular dendritic cells (FDCs) would be prime candidates. Indeed, PrP^{Sc} accumulation has been observed in these cells. In mouse scrapie, functional B lymphocytes are necessary for neuroinvasion, but PrP^C expression in these cells is not required. B lymphocytes may indirectly influence neuroinvasion by allowing the development of mature spleen FDCs as sites of agent replication. Because lymphocytes do not normally cross the blood–brain barrier, it seems questionable, however, whether immune cells are sufficient to transport the agent from the periphery to the CNS.

Prion replication in the CNS first takes place in areas that relate to the site of peripheral inoculation or oral uptake. This implies that the agent spreads along the peripheral nervous system. The importance for neuroinvasion of PrP^C positivity of peripheral nerves has been demonstrated in experiments where transgenic

mice expressing PrP^C only in neurons developed scrapie after oral or intraperitoneal infection with high doses of the infectious agent. A scenario in which the agent is first transported to FDCs by mobile immune cells, where it amplifies and spreads to the peripheral nervous system, also seems possible and may be of particular importance in cases of low-dose infection.

B. Neuronal Cell Death

Two basic types and mechanisms of cell death have been described in the brain and other organs: necrosis and apoptosis. Necrosis often results from severe and sudden injury and leads to rapid cell lysis and a consecutive inflammatory response. In contrast, apoptosis proceeds in an orderly manner following a cellular suicide program involving active gene expression in response to physiological signals or types of stress. It is usually not accompanied by an inflammatory response. The *in situ* end-labeling technique (ISEL), which is based on the incorporation of labeled nucleotides in fragmented DNA by terminal transferase, is thought to be highly characteristic of apoptosis and has been used to investigate neuronal cell death in a scrapie model in the mouse. ISEL revealed nuclei containing fragmented DNA in the granule cell layer of the cerebellum in mice infected with the 79A strain of scrapie from day 120 onward. The number of labeled cells increased from days 120 to 150 and was highest in terminally ill mice (day 166). Electron microscopy of specimens from the cerebellum of terminally ill mice identified cells that showed homogeneously condensed chromatin, dark cytoplasm, membrane blebbing, and, occasionally, nuclear fragmentation, all of which are morphological changes characteristic of apoptosis. Two mechanisms are discussed as causing neuronal cell death in prion diseases: loss of function of PrP^C and gain of function of PrP^{Sc}. Whereas some of the electrophysiological data in $Prnp^{0/0}$ mice were interpreted as indicating that progressive loss of PrP^C in prion disease might lead to impaired synaptic transmission and neuronal cell death, other experiments showed that PrP^{Sc} has toxic effects on neurons in primary cell culture. Arguing that this toxic effect could reside in a part of the protein that is deposited in brain tissue in human prion disease, Forloni and co-workers identified a peptide corresponding to human PrP amino acid residues 106–126 (PrP106–126), which elicited maximal neurotoxic effects in rat hippocampal cultures. The involvement of different cell types present in mixed cerebellar cultures in the neurotoxic mechanism of PrP106–126 was investigated by using L-leucine methyl ester (LLME), which significantly reduces the number of microglia in mixed cultures. After this treatment, 10-day administration of PrP106–126 was no longer toxic to cerebellar cells. When cerebellar cells from $Prnp^{0/0}$ mice were treated with PrP106–126, the peptide showed no toxic effect. These experiments have also been performed with PrP27–30, the protease-resistant core of PrP^{Sc} isolated from scrapie-infected hamster brains. The results show that cellular expression of PrP^C and the presence of microglia are necessary for the neurotoxicity of PrP27–30 and PrP106–126 *in vitro*. This interpretation was confirmed by experiments using normal brain tissue transplanted into $Prnp^{0/0}$ mice. In these animals, only $Prnp^{+/+}$ cells were prone to the toxic effect of PrP^{Sc}.

To investigate the neurotoxic effect of microglia, the supernatants of microglial cultures were assayed for nitrite production, and a high increase in production after the application of PrP106–126 was found. The increased nitrite most likely is a stable reaction product of sort-lived superoxide radicals and nitric oxide, which can be produced by microglia to induce oxidative stress. This result strongly suggests that PrP106–126 acts directly on microglia and gives strong support to the notion that microglia induce neuronal death by oxidative stress as a result of PrP106–126 stimulation.

As far as the role of microglial activation *in vivo* is concerned, it has been well-documented that there is an intense microglial reaction in mice terminally ill with scrapie, which corresponds to the pattern of neurodegeneration. The question of whether microglial activation is secondary to neuronal cell death or whether microglial activation precedes neuronal cell death and is possibly involved in inducing neuronal cell death has been addressed in a large time course study using three different strains of scrapie in mice. It appeared that microglial activation was present early in the incubation period in all models studied. The pattern of microglial activation closely paralleled the pattern and time course of PrP^{Sc} accumulation. Microglial activation clearly preceded apoptotic neuronal cell death in all strains studied. Furthermore, quantitative analysis of microglial activation and cell death in the cerebellum showed that the time course and extent of neuronal cell death correlated with the time course of microglial activation. Taken together with the data derived from cell-culture studies, this indicates that

microglial activation is also involved in mediating PrP^{Sc} toxicity *in vivo*.

III. THE PRION DISEASES

A. Animal Prion Diseases

Scrapie, a disease that naturally occurs in sheep and goats, has been known for more than 200 years and was the first prion disease to be shown to be infectious in 1936 by two French scientists, Cuillé and Chelle. Affected sheep clinically show abnormal behavior, such as excessive scraping as well as trembling and ataxia and other motor disturbances. There are no known hereditary prion diseases in animals; however, allelic variations in the ovine PrP^{C} sequence exert a strong influence on susceptibility to natural and experimental scrapie. Although the disease has been known for centuries and its infectious nature was recognized more than 60 years ago, the exact mode of natural transmission is not known. Maternal transmission seems to be an established fact, but there are also reports of transmission by keeping sheep on pastures that were previously occupied by scrapie-infected flocks. Scrapie is experimentally transmissible to many mammalian species. Epidemiological studies show that scrapie apparently is not transmissible to humans by the oral route.

BSE is a disease of cattle that was first described in Great Britain in 1986. It is estimated that since then almost 1 million animals have been infected. The mean incubation time is about 5 years; therefore, most animals did not manifest disease because they were slaughtered at age 2 or 3 years. Nonetheless, more than 160,000 affected animals were diagnosed and killed over the years. It is now clear that BSE was spread by feeding cattle with contaminated meat and bone meal (MBM) prepared from the offal of sheep, cattle, and pigs. However, the origin of BSE is still obscure. It is now thought that changes in the hydrocarbon extraction method, including temperature changes, that were made in the rendering of offal in the late 1970s allowed the infectious agent to survive the manufacturing process and pass from sheep to cattle. Because by strain typing and PrP^{Sc} banding pattern in Western blots all scrapie strains tested so far are different from BSE, an alternative hypothesis, i.e., transmission of preexisting natural BSE at low incidence as a consequence of the preceding changes in the rendering process, cannot be dismissed. BSE is experimentally transmissible to many species and seems to have passed the species barrier to humans (see later discussion).

In addition, a number of other prion diseases exist (see Table I), some of which are of unknown origin such as chronic wasting disease in mule deer and elk, others are transmitted by contaminated offal from sheep or cattle such as transmissible mink encephalopathy, feline spongiform encephalopathy, and exotic ungulate encephalopathy.

B. Human Prion Diseases

CJD and Gerstmann–Sträussler–Scheinker syndrome (GSS) were first described as neurodegenerative diseases in the 1920s, whereas kuru was first reported in 1957. CJD and kuru were successfully transmitted to chimpanzees in the 1960s. The exceptional nature of GSS as a hereditary disease that is experimentally transmissible to laboratory animals was discovered in 1981. In 1989, the first mutation of the prion protein gene was identified in a GSS family. Neuropathology has played a lead role in defining the various entities that are now known as human prion diseases and is still of particular importance in routine diagnosis. The classical neuropathological changes consist of (1) spongiform degeneration, (2) PrP^{Sc} deposition, (3) neuronal loss, and (4) astrocytic gliosis (Fig. 2). There is some confusion as to the exact definition of such terms as "spongiform change" and "spongiosis." Spongy changes or spongiois describes small parenchymal vacuoles predominantly in the cerebral or cerebellar gray matter in various diseases. Out of this group, spongiform changes are really specific for prion diseases. They are small, often opaque vacuoles of 2–10 μm in diameter. They are found as single vacuoles or in small groups in the neuropil in practically all human prion diseases (Fig. 2a). Confluent vacuoles are 10–50 μm in diameter, found in grapelike accumulations (Fig. 2b), and characteristic of MM patients with PrP^{Sc} type 2 (see later discussion). Both spongiform changes and confluent vacuoles are only found in prion diseases. In contrast, status spongiosus, which is characterized by almost complete nerve cell loss, extreme astrocytic gliosis, and tissue disintegration with pericellular formation of holes, is found in prion diseases and late stages of various neurodegenerative and metabolic diseases. Spongy degeneration of the first and second cortical layers is a result of vacuolar shrinkage, often in the late stages of neurodegenerative

Table I
Prion Diseases

Type	Disease	Etiology
	Human Diseases	
Idiopathic	Sporadic Creutzfeldt–Jakob disease (sCJD)	Unknown. Possibly spontaneous conversion of PrP^{C} into PrP^{Sc} or somatic *PRNP* mutation.
	Sporadic fatal insomnia (SFI)	Unknown. Possibly spontaneous conversion of PrP^{C} into PrP^{Sc} or somatic *PRNP* mutation. Found only in 129 MM patients with PrP^{Sc} type 2.
Hereditary	Familial CJD	Various *PRNP* mutations.
	Gerstmann–Sträussler–Scheinker syndrome	Various *PRNP* mutations.
	Fatal familial insomnia (FFI)	*PRNP* mutation D178N with M129.
Acquired	Iatrogenic CJD	Accidental transmission through treatment with prion-contaminated preparations of human growth hormone, dura mater grafts, etc.
	(New) variant CJD, (n)vCJD	Infection by bovine prions (BSE-contaminated food or other products).
	Kuru	Infection through ritualistic cannibalism in the Fore population in New Guinea (historical).
Disease	**Host**	**Etiology**
	Animal Diseases	
Scrapie	Sheep (and goats)	Oral–maternal infection in genetically susceptible sheep.
Bovine spongiform encephalopathy (BSE)	Cattle	Infection with contaminated meat and bone meal (MBM).
Feline spongiform encephalopathy (FSE)	Felidae (cats)	Infection with contaminated feed.
Exotic ungulate encephalopathy	Nyala, kudu, onyx	Infection with contaminated feed.
Transmissible mink encephalopathy (TME)	Mink	Infection with contaminated feed.
Chronic wasting disease (CWD)	Mule deer and elk	Unknown.

diseases with strong cortical atrophy such as prion diseases, Alzheimers's disease, and Pick's disease. PrP^{Sc} deposition is found as diffuse delicate accumulations in the gray matter in areas of high synaptic density and in a distribution similar to that of synaptophysin (Fig. 2c). This type of PrP^{Sc} deposition therefore has been called synaptic. Other forms are described as pericellular, perivacuolar, in kuru plaques that are visible in routine H&E stains, and plaquelike, i.e., in small plaques that are only visible after immunohistochemical staining of PrP^{Sc}.

1. Kuru

Kuru was first described as a deadly neurodegenerative disease affecting the Fore people in the eastern highland of New Guinea. This disease was mainly characterized by ataxia and predominantly affected women and children. William Headlow, a veterinary pathologist, noted the neuropathological similarities of scrapie and kuru and suggested that infection experiments be performed in apes. Gajdusek and Gibbs were successful in transmitting the disease to chimpanzees in 1966. Later, ritualistic cannibalism of deceased clan members was identified as the mode of transmission in the Fore population. After cannibalistic practices ceased, the disease disappeared. Kuru incubation times of more than three decades were reported. The pathological similarities of kuru and CJD led Gibbs and Gajdusek to perform further transmission experiments and to show that CJD is transmissible to apes.

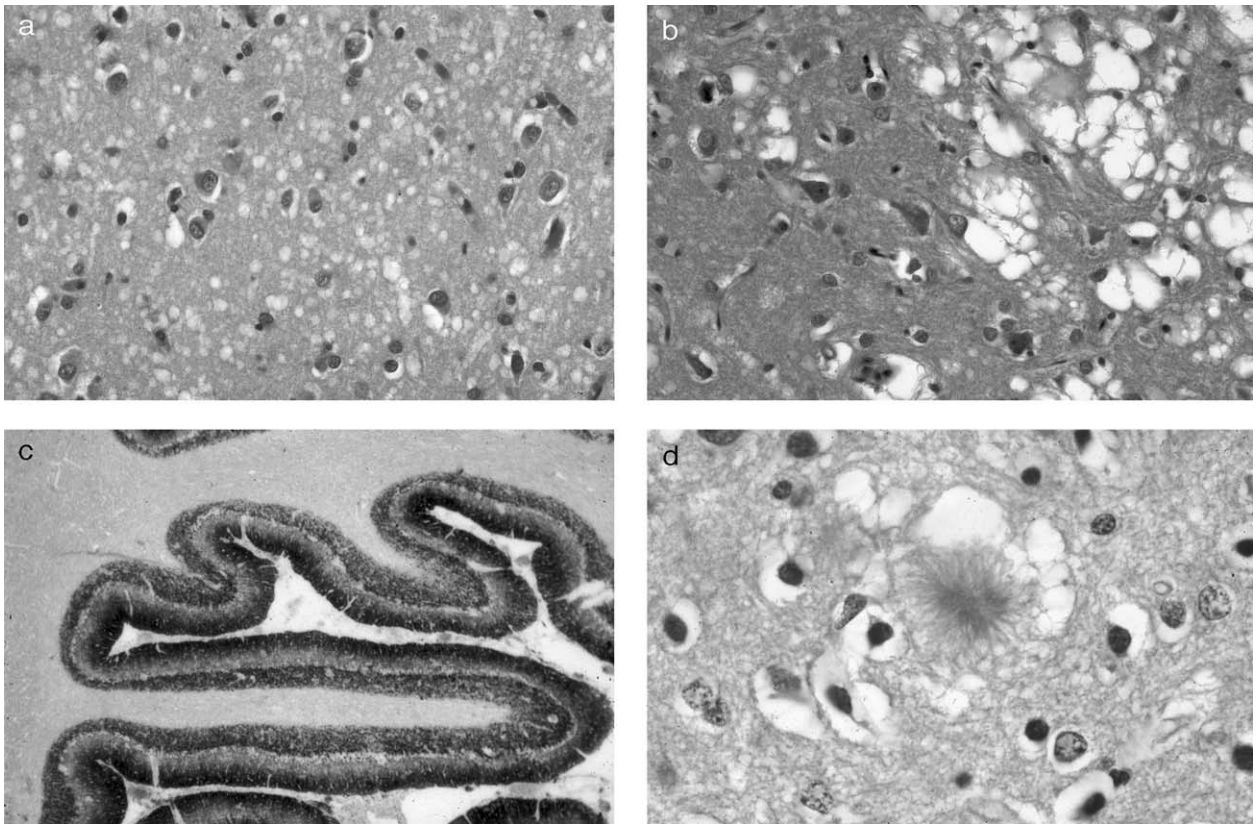


Figure 2 Histological findings in Creutzfeldt–Jakob disease. (a) Typical spongiform changes with small transparent or opaque vacuoles in the neocortex of a sporadic CJD case (MM, PrP^{Sc} type 1). Hematoxylin and eosin stain, × 20 (original magnification). (b) Large, confluent vacuoles are seen in the neocortex of this sporadic CJD case (MM, PrP^{Sc} type 2 cortical). Hematoxylin and eosin, × 40 (original magnification). (c) Paraffin-embedded tissue blot (PET blot) of the cerebellum of a sporadic CJD case (MM PrP^{Sc} type1). In this technique, formalin-fixed and paraffin-embedded tissue is digested with proteinase K and blotted onto a nitrocellulose membrane. PrP^{Sc} detection is then achieved by labeling with PrP-specific antibodies. In this case, synaptic-type deposition in the cerebellum is stained dark brown. (d) A typical florid plaque consisting of delicate strands and surrounded by confluent vacuoles in a case of nvCJD. Hematoxylin and eosin, × 40. (Courtesy of Dr. James Ironside, Edinburgh, Scotland.)

2. Sporadic (Idiopathic) Creutzfeldt–Jakob Disease

Sporadic CJD (sCJD) most often affects patients in their 60s. It usually presents with dementia and various neurological signs and runs a relentless course, leading to death usually within 6 months. Diagnostic criteria established by Masters have been modified (Table II). A definite diagnosis can only be made by neuropathologic or biochemical examination of the brain.

Well-documented phenotypic heterogeneity of sCJD in the absence of a genome of the infectious agent has been a puzzling observation that now seems to be resolved. Parchi and co-workers described two different molecular types of PrP^{Sc} associated with distinct clinical and pathological phenotypes in sCJD. On transmission to laboratory animals, the properties of PrP^{Sc} types 1 and 2 have been shown to propagate faithfully. The two types of PrP^{Sc} are distinguished by their different physicochemical properties, particularly their appearance on Western transfers after digestion with proteinase K. The unglycosylated forms of PrP^{Sc} are seen as proteins of approximately 21 (type 1) and 19 kDa (type 2) relative molecular mass (Fig. 3). Proteinase K has two preferential, if not exclusive, cleavage sites at codons 97 and 82 of PrP^{Sc}, most likely related to two different conformations of PrP^{Sc}. In addition, the common methionine and valine polymorphism at codon 129 of *PRNP* was shown to modify the phenotype of the disease. The full spectrum of sCJD

Table II
Diagnostic Criteria for CJD from the World Health Organization

The clinical criteria define possible and probable cases:

1. Sporadic CJD

   Probable CJD

   Progressive dementia

   Typical EEG (periodic sharp wave complexes) during an illness of any duration and/or

   Positive 14-3-3 assay and clinical duration to death <2 years

   At least two out of four clinical features listed:

   1. Myoclonus
   2. Visual or cerebellar disturbance
   3. Pyramidal or extrapyramidal dysfunction
   4. Akinetic mutism

   Routine investigations should not suggest an alternative diagnosis.

   Possible CJD

   Clinical features identical to those for probable CJD, but no typical EEG and no positive 14-3-3 assay.

2. Accidentally transmitted CJD

   Progressive cerebellar syndrome in a pituitary hormone recipient

   Sporadic CJD with a recognized exposure risk (e.g., dura mater transplant)

3. Familial CJD

   Definite or probable CJD plus definite or probable CJD in a first degree relative

   Neuropsychiatric disorders plus disease-specific *PRNP* mutations

4. NvCJD

   NvCJD cannot be diagnosed with certainty using clinical criteria at present. The diagnosis of nvCJD should be considered as a possibility in a patient with a progressive neuropsychiatric disorder with at least five out of the following six clinical features: (1) early psychiatric symptoms, (2) early persistent paresthesia–dysesthesia, (3) ataxia, (4) chorea–dystonia or myoclonus, (5) dementia, (6) akinetic mutism. The suspicion of nvCJD is strengthened by the following criteria: (7) absence of a history of potential iatrogenic exposure, (8) clinical duration >6 months, (9) age at onset <50 years, (10) absence of a *PRNP* mutation, (11) EEG not showing the typical periodic appearance, (12) routine investigations do not suggest an alternative diagnosis, (13) MRI showing abnormal bilateral signal from the pulvinar on axial T2- and/or proton density-weighted images. A patient with a progressive neuropsychiatric disorder and five out of clinical features 1–6 and all criteria 7–13 should be considered as a suspect case of nvCJD for surveillance purposes. A definite diagnosis is established only by neuropathological examination.

   The definite diagnosis is based on the examination of brain tissue:

   1. Neuropathological examination including immunohistochemistry
   2. Western blot analysis with antibodies against PrP
   3. Isolation of scrapie-associated fibrils or prion rods

variants was defined in a large series of 300 CJD patients and was compared with molecular and pathological phenotypes. Six distinct types of sCJD were described by molecular analysis. Almost 90% of sCJD patients were homozygous at codon 129, the vast majority being MM homozygotes, whereas only half of the normal population in Europe and the United States is homozygous. PrP^{Sc} types 1 and 2 were found to be associated with all *PRNP* genotypes. However, there was a strong association between PrP^{Sc} type 1 and MM homozygosity as well as between PrP^{Sc} type 2 and VV or MV patients.

A detailed analysis of the pathologic lesion pattern (lesion profile) and PrP immunohistochemistry defined six pathological variants of sCJD, which by and large were associated with the six groups of sCJD as defined by molecular analysis (Table III). Among the MM2 patients, one subgroup, termed MM2-C (cortical), had a pathological phenotype that closely resembled that of MM1 patients, with the exception that they lacked cerebellar pathology, whereas another subgroup, termed MM2-T (thalamic), showed a striking resemblance to the fatal familial insomnia (FFI) phenotype. Immunohistochemistry for PrP^{Sc}

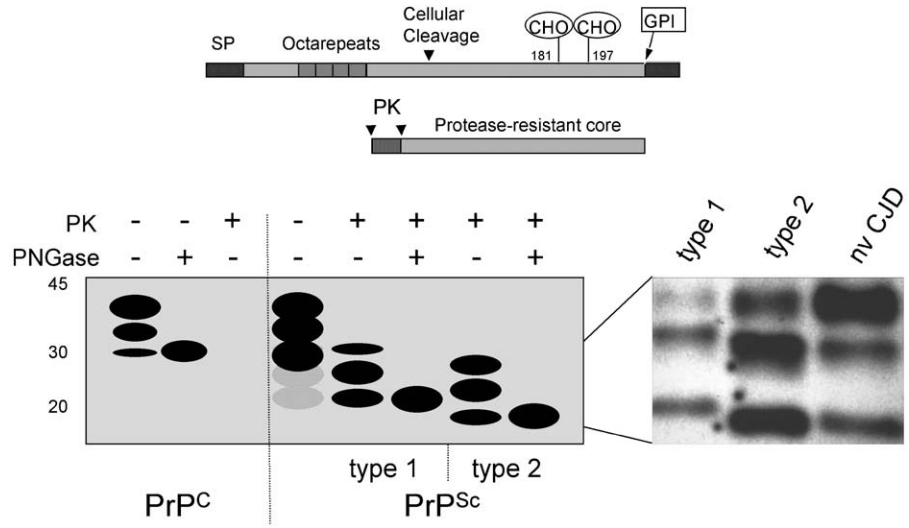


Figure 3 Western blot analysis of human prion proteins. Shown are the primary structure of PrP^{C} and the protease-resistant core of PrP^{Sc}, as well as their appearance on Western blots. The insert in the right lower corner shows an original Western blot with the three PrP^{Sc} bands after pK digestion of two sporadic CJD cases (PrP^{Sc} type 1 with the unglycosylated PrP^{Sc} at 21 kDa, PrP^{Sc} type 2 with the unglycosylated PrP^{Sc} at 19 kDa) and an nvCJD case in which the three PrP bands correspond in size to PrP^{Sc} type 2 but show a different glycosylation pattern.

was positive in all subgroups of sCJD, though it may be extremely weak and hardly recognizable or reliable in VV1 and MM2-T cases.

As for the clinical signs at onset, the absence of cognitive impairment in a large group of VV2 patients and the absence of ataxia in the VV1 and MM2-C subgroups were remarkable. During the course of the disease, dementia was found in all subjects except a small group of patients who rapidly lapsed into a stupor after a short initial phase with exclusively neurological signs. Ataxia was always more common in VV2 and MV2 subjects. Visual signs and myoclonus were prominent in the MM1 and MV1 subgroups.

Electroencephalography (EEG) showed typical periodic sharp wave complexes (PSWC) in MM1 and MV1 subjects, but these were rare in MV2 and VV2 patients and absent in VV1, MM2-C, and MM2-T patients. As a consequence, these patients are classified as possible CJD according to classic clinical diagnostic criteria. The 14-3-3 proteins in the CSF have proven to be good surrogate markers for most variants of sCJD, with the exception of MV2 subjects.

MM1 and MV1 subjects comprise 70% of sCJD cases, exhibiting features that have been recognized as typical of CJD, i.e., these subjects show dementia, myoclonus, and PSWC. On histologic examination, they show delicate spongiform changes in the cerebral cortex, particularly the occipital cortex, neostriatum, thalamus, and cerebellum. Brain stem nuclei and the hippocampus are usually spared. Immunohistochemistry shows synaptic PrP^{Sc} deposition in the cortex and cerebellum. The second most common subgroup is VV2 patients (16%), who are clinically characterized by prominent ataxia and a lack of PSWC upon EEG. Histology is characterized by spongiform degeneration of the deep cortical laminae of the neocortex, limbic cortex, and subcortical nuclei. Immunohistochemically, plaquelike deposits of PrP^{Sc} and prominent perineuronal staining are found. With 9% of all cases, MV2 patients are the third most common group that is clinically similar to the VV2 type. These patients show no typical EEG, and in most cases the proteins 14-3-3 are also negative in the CSF. Histologically there is involvement of cortical and subcortical structures. The most prominent feature is cerebellar kuru plaques, which are homogeneous eosinophilic structures that, in contrast to plaquelike deposits are visible on routine H&E-stained sections. The rare MM2-T phenotype (2%) is indistinguishable from the FFI phenotype clinically and neuropathologically. There is prominent atrophy of the thalamus and the inferior olivary nucleus, minimal focal cortical spongiform change, and minimal and unreliable immunohistochemical staining for PrP^{Sc}. Progressive insomnia, absence of PSWC, and visual hallucinations are found in the MM2-T group, which has also been termed

Table III
Molecular and Phenotypic Features of sCJD Variants

sCJD variant	Other classification	Percent of cases	Duration (months)	Age at onset (years)	Clinical features	Neuropathological features
MM1 or MV1	Myoclonic, Heidenhain variant	70	3.9 (1–18)	65.5 (42–91)	Rapidly progressive dementia, early and prominent myoclonus, typical EEG; visual impairment or unilateral signs at onset in 40% of cases.	"Classic" CJD pathology: Spongiform change with small (2–10 μm) round to oval vacuoles, predominantly located to neuronal processes; often predominant involvement of occipital cortex; "synaptic" type PrP staining; in addition, one third of cases shows confluent vacuoles and perivacuolar PrP staining
VV2	Ataxic variant, Brownell-Oppenheimer variant	16	6.5 (3–8)	61.3 (41–80)	Ataxia at onset, dementia late in the disease, no typical EEG in most cases.	Prominent involvement of subcortical nuclei including brain stem; in neocortex, spongiosis is often confined to deep layers; PrP staining shows plaque-like focal deposits as well as prominent perineuronal staining
MV2	Kuru plaque variant	9	17.1 (5–72)	59.4 (40–81)	Ataxia and progressive dementia, no typical EEG, long duration (>2 years) in some cases.	Cerebellar kuru plaques, other pathological changes similar to VV2 with more consistent plaque-like deposits
MM-2 thalamic	Thalamic variant, sporadic fatal insomnia (SFI)	2	15.6 (8–24)	52.3 (36–71)	Insomnia and psychomotor hyperactivity in most cases, in addition to ataxia and cognitive impairment, no typical EEG.	Prominent atrophy of the thalamus and inferior olive with little pathology in other areas; spongiosis may be absent or focal, minimal PrP^{Sc} deposition
MM-2 cortical	—	2	15.7 (9–36)	64.3 (49–77)	Progressive dementia, no typical EEG.	Large confluent vacuoles with perivacuolar PrP^{Sc} staining in all cortical layers; cerebellum is relatively spared
VV1	—	1	15.3 (14–16)	39.3 (24–49)	Progressive dementia, no typical EEG.	Severe pathology in the cerebral cortex and striatum with sparing of brain stem nuclei and cerebellum; no large confluent vacuoles, and very faint synaptic PrP^{Sc} staining

sporadic fatal insomnia (SFI). The cortical MM-2 phenotype is also rare (2%). It shows progressive dementia and no PSWCs. Histologically there are large confluent vacuoles with coarse perivacuolar PrP^{Sc} deposits. VV1 patients are the smallest subgroup of sCJD (1%). There are very young patients in this group. Clinically they show progressive dementia with no typical EEG. Histologically there is massive spongiform degeneration with numerous small vacuoles in the neocortex and striatum in these cases. Immunohistochemical detection of PrP^{Sc} is minimal and sometimes unreliable.

The detailed analysis of molecular and phenotypic variants of sCJD thus forms the basis for the recognition of unusual subtypes, which in the past may not have been diagnosed as CJD because of their early age at onset, their unusual clinical course, and even their atypical pathology. Thus, there should be

much greater clinical awareness of sCJD cases in young patients with a long clinical course and progressive dementia but no PSWC, as represented by MM2-C and VV1 cases.

3. Hereditary Prion Diseases

About 10–15% of all prion disease cases are hereditary. A large number of different point mutations and insertion mutations of the prion protein gene have been identified in familial prion diseases. Insertion mutations represent additional repeats of the N-terminal Cu^{2+}-binding octapeptide (Fig. 1). Depending on their clinical and neuropathological characteristics, familial prion diseases are designated as familial CJD, GSS, or fatal familial insomnia (FFI).

Familial CJD often is indistinguishable from sporadic CJD and is most often associated with the E200K mutation and less often with D178N-129V, V180I, R208H, V210I, M232R, and insertion mutations. GSS is thought to be distinguishable from CJD by the predominance of ataxia, whereas dementia and myoclonus are more prominent in CJD. Neuropathologically, GSS is quite distinct and is characterized by large, multicentric PrP-containing amyloid plaques; spongiform change is variable. It is found most often in families with the P102L mutation and less often with P105L, A117V, F198S, D202N, Q212P, Q217R, and insertion mutations. Fatal familial insomnia (FFI) often presents with insomnia and dysautonomia and later shows signs of ataxia, dysarthria, myoclonus, and pyramidal tract dysfunction. Neuropathologically, FFI is characterized by neuronal loss and astrocytic gliosis preferentially affecting the ventral anterior and medial dorsal nuclei of the thalamus and the inferior olive. FFI is caused by a D178N mutation of *PRNP* associated with a methionine codon at position 129 of the same allele.

4. Iatrogenic CJD

The first possible person-to-person transmission of CJD was reported in a recipient of a corneal transplant from a donor with CJD in 1974. Other modes of accidental iatrogenic transmission were reported in the following years, including the use contaminated EEG depth electrodes, neurosurgical instruments, cadaveric pituitary-derived gonadotrophin and human growth hormone (hGH), and dura mater grafts. Transmission by contaminated hGH has raised major concern because of the large number of possibly affected persons. Worldwide more than 120 hGH-associated cases have been reported to date, whereas 8000 persons received hGH between 1963 and 1985 in the United States alone. The affected patients commonly present with gait abnormalities and ataxia, whereas dementia is a late manifestation and is mild. The incubation time in these cases has been estimated to be 12 years or longer.

In February 1987, the first CJD case related to a dura mater graft (Lyodura, B. Braun Melsungen, Germany) was reported. The same year, representatives of the company announced that they had changed their procedures for collecting and processing dura mater in order to reduce or eliminate the risk of CJD transmission. Subsequent case reports from Germany, Italy, Spain, New Zealand, the United Kingdom, the United States, and Japan suggested that Lyodura processed before 1987 was indeed associated with an increased risk of developing CJD. Worldwide more than 100 dura-mater-associated CJD cases have been reported to date. The mean incubation time in these cases was 5.8 years, and the patients' mean age was 38.5 years. A total of 57 cases were reported as having received Lyodura, with 43 reported from Japan alone. The Japanese cases go back to a survey study with little information on the details revealed in a publication. Thus, the diagnosis was confirmed in only 10 patients by neuropathological examination. The number of clinically possible and clinically probable cases is not reported. The mean incubation period in the Japanese cases was 89 months, and the mean age was 53 years. In June 1998, 21 unpublished Lyodura-associated cases in various countries were identified. In contrast to the cases transmitted by contaminated hGH (see earlier discussion), there is no difference in the symptomatology of sporadic CJD and dura-mater-associated cases.

5. New Variant CJD and the Transmission of BSE to Humans

A new variant of CJD was described in 10 patients in the United Kingdom in 1996. These patients had a mean age of 29 years and presented with psychiatric disturbances, whereas signs more typical of CJD developed later in the course of disease. Neuropathology at autopsy was exceptional, showing extensive depositions of PrP^{Sc} in various areas of the brain in a fashion that had only been described in hereditary disease before 1996. In addition, there were florid plaques with a central PrP accumulation and surrounding vacuoles (Fig. 2d) that had not been seen in human prion disease before. From 1996 to date, 100 cases of this variant were identified in the United

Kingdom, 3 were found in France, and the search in other countries was not successful. Information on the clinical appearance has been compiled, and a definition for suspect nvCJD cases is now available (see Table II). All nvCJD patients to date have been methionine homozygotes at codon 129 of the prion protein gene; the youngest patient was 15 years old and the oldest was 54 years old at death. It is not possible at present to define the clinical and pathological features to be expected in codon 129 valine homozygotes or in heterozygotes that are likely to be observed in the future. nvCJD has shown significantly new features in the pattern of extracerebral deposition of PrP^{Sc} in the tonsils, lymph nodes, spleen, and appendix. This has raised concern that blood cells may also harbor the infectious agent and that the disease might be spread by blood transfusions. Although this suspicion is unsubstantiated to date, measures have been taken to prevent the spread by blood transfusion in various countries.

It is hypothesized that nvCJD cases have been caused by the consumption of food or other products containing large amounts of the BSE agent (BSE prions). This hypothesis is strengthened by epidemiological and experimental findings. The appearance of nvCJD 10 years after BSE in the country with the highest incidence of BSE is highly consistent with this hypothesis. The banding pattern of PrP^{Sc} in nvCJD has been shown to be different (see Fig. 3) and resembles the pattern in BSE. Although BSE and nvCJD show significant differences pathologically, upon transmission to genetically homogeneous animals (inbred mice) they elicit practically identical patterns, whereas in these strain-typing experiments all tested scrapie strains and sporadic CJD cases were different. These findings have been confirmed in transgenic animals expressing bovine PrP.

If we assume that nvCJD is caused by transmission of the BSE agent to humans, the available epidemiological data are insufficient to predict the number of cases that must be expected to occur in the future, because neither the mean incubation time nor the shape of the epidemiological curve, the exact mode of transmission, or any other parameter that might possibly be important is known at present.

C. Models in Transgenic Animals

Many transgenic mouse models have been developed to study susceptibility, prion pathogenesis, species barrier, prion propagation, and other problems. The nature of the species barrier was examined by introducing the hamster PrP gene into the germ line of mice. These transgenic mice became readily susceptible to the hamster agent and produced the homologous type of agent when infected with the material from either hamster or mouse. The situation was more complicated when human *PRNP* was introduced into mice. Only after a chimeric gene with murine N- and C-termini and the central part of *PRNP* (codons 96–167) was expressed in transgenic animals did these mice become susceptible to human prions. This has led to the postulate of the existence of a species-specific protein X that interacts with PrP. There are conflicting findings from another group who succeeded in breaking the species barrier by introducing unmodified *PRNP* into mice. Transgenic mice expressing the bovine PrP^{C} have shown short incubation times after inoculation with BSE and practically no decrease in incubation time at second transmission, i.e., they seem to have no species barrier and thus should be an ideal model to study the infectivity of various organs of BSE-infected cattle. Exciting data concerning inherited prion diseases were provided by experiments with mice overexpressing a murine *Prnp* with a homolog of the P102L-GSS mutation [Tg(MoPrP-P101L)H]. These mice spontaneously developed a neurodegenerative disease reminiscent of other prion diseases in mice. However, an accumulation of protease-resistant PrP was not found. The infectious nature of the disease generated was shown by serial transmission to hamsters (10% of animals) and transgenic mice expressing low levels of the mutated protein (40% of animals), whereas the disease could not be transmitted to normal mice. Again, protease-resistant PrP was not detectable in the brains of any of the infected animals. Some transgenic lines overexpressing various normal PrP genes develop a lethal spontaneous disease late in life. This disease is characterized by the degeneration of skeletal muscles and peripheral nerves and by spongiform changes in the brain. It was claimed that this disease may be transmissible, yet no protease-resistant PrP was found. To complicate the matter, another transgenic line expressing P101L at normal levels does not spontaneously develop disease.

D. New Diagnostic Approaches

Although transmission studies indicate the presence of low levels of infectivity, it has not been possible to detect PrP^{Sc} aggregates in the CSF using generally

available techniques such as the Western blot or ELISA. On the basis of a setup for confocal dual-color fluorescence correlation spectroscopy (FCS), a technique suitable for single-molecule detection, a novel and highly sensitive detection method, was developed: the scanning for intensely fluorescent targets (SIFT) technique for PrP^{Sc}. In this technique, pathological prion protein aggregates are labeled by specific antibody probes tagged with fluorescent dyes, resulting in intensely fluorescent targets (PrP^{Sc} aggregates) that are measured by dual-color fluorescence intensity distribution analysis in a special scanning setup. In a diagnostic model system, PrP^{Sc} aggregates were detectable down to a concentration of 2pM PrP^{Sc}, corresponding to an aggregate concentration of approximately 2 f*M*. This was more than one order of magnitude more sensitive than Western blot analysis. PrP^{Sc}-specific signal was detected in a number of CSF samples from CJD patients. Thus, for the first time PrP^{Sc} was directly detected in the CSF, providing the basis for a rapid and specific test for CJD and other prion diseases. Improvements in scanning as well as sample handling and preparation may significantly improve sensitivity.

E. Therapy

Interference with the conversion reaction from PrP^{C} to PrP^{Sc} would seem to be the most promising therapeutic approach. This would entail stabilization of PrP^{C} or destabilization of PrP^{Sc} by direct or indirect means. In that vein, experiments in cell-free conversion systems and in scrapie-infected neuroblastoma cell cultures have shown that PrP^{C} conversion can be inhibited by synthetic peptides composed of amino acids 109–141 and 119–136. Because the region from 119 to 136 is conserved in most mammalian species, this peptide may be of practical use for therapy in many prion diseases. In addition, a number of substances such as Congo red, amphothericin B, porphyrins, pentosan phosphate, and phthalocyanins have empirically been found to prolong the incubation time in experimental scrapie systems. Some of these substances have been shown to inhibit PrPres formation *in vitro*. The availability of analogs of macrocyclic compounds such as porphyrins and phthalocyanins in large numbers that vary in structure and that can be modified in a number of ways offers the opportunity for rational drug development and, thus, may become potent therapeutic agents. New avenues of research arise from observations indicating that prions may be transported by cells of the lymphoreticular system. The search for effective approaches to therapy for prion diseases seems most prudent at a time when the total number of nvCJD cases is slowly increasing, whereas we have no means to predict the total number of individuals that may be affected by this new and deadly disease.

See Also the Following Articles

BORNA DISEASE VIRUS • BRAIN LESIONS • CEREBRAL WHITE MATTER DISORDERS • DEMENTIA • LYME ENCEPHALOPATHY • MULTIPLE SCLEROSIS • NEURODEGENERATIVE DISORDERS

Suggested Reading

Bieschke, J., Giese, A, Schulz-Schaeffer, W., Zerr, I., Poser, S., Eigen, M., and Kretzschmar, H. A. (2000). Ultrasensitive detection of pathological prion protein aggregates by dual-color scanning for intensely fluorescent targets (SIFT). *Proc. Natl. Acad. Sci. USA*, **97,** 5468–5473.

Bruce, M. E., Will, R. G., Ironside, J. W., McConnell, I., Drummond, D., Suttie, A., McCardie, L., Chree, A., Hope, J., Birkett, C., Cousens, S., Fraser, H., and Bostock, C. J. (1997). Transmissions to mice indicate that new variant CJD is caused by the BSE agent. *Nature (London)* **389,** 489–501.

Collinge, J. (1997). Human prion diseases and bovine spongiform encephalopathy (BSE). *Human Mol. Genet.* **6,** 1699–1705.

Eigen, M. (1996). Prionics or the kinetic basis of prion diseases. *Biophys. Chem.* **63,** A1–A18.

Herms, J., Tings, T., Gall, S., Madlung, A., Giese, A., Siebert, H., Schurmann, P., Windl, O., Brose, N., and Kretzschmar, H. (1999). Evidence of presynaptic location and function of the prion protein. *J. Neurosci.* **19,** 8866–8875.

Kretzschmar, H. A., Ironside, J. W., DeArmond, S. J., and Tateishi, J. (1996). Diagnostic criteria for sporadic Creutzfeldt–Jakob disease. *Arch. Neurol.* **53,** 913–920.

Parchi, P., Giese, A., Capellari, S., Brown, P., Schulz-Schaeffer, W., Windl, O., Zerr, I., Budka, H., Kopp, N., Piccardo, P., Poser, S., Rojiani, A., Streichenberger, N., Julien, J., Vital, C., Ghetti, B., Gambetti, P., and Kretzschmar, H. A., (1999). Classification of sporadic Creutzfeldt-Jakob disease based on molecular and phenotypic analysis of 300 subjects. *Ann. Neurol.* **46,** 224–233.

Prusiner, S. B. (1982). Novel proteinaceous infectious particles cause scrapie. *Science* **216,** 136–144.

Prusiner, S. B. (1998). Prions. *Proc. Natl. Acad. Sci. USA* **95,** 13363–13383.

Riek, R., Hornemann, S., Wider, G., Glockshuber, R., and Wüthrich, K. (1997). NMR characterization of the full-length recombinant murine prion protein, mPrP(23-231). *FEBS Lett.* **413,** 282–288.

Riesner, D. (1997). Prions and their biophysical background. *Biophys. Chem.* **666,** 259–268.

Weissmann, C. (1999). Molecular genetics of spongiform encephalopathies. *J. Biol. Chem.* **274,** 3–6.

Windl, O., and Kretzschmar, H. A. (2000). Prion Diseases. In *Neurogenetics. Contemporary Neurology Series*. (S.-M. Pulst, Ed.), pp. 191–218. Oxford University Press, New York.

Problem Solving

RICHARD E. MAYER
University of California, Santa Barbara

GLOSSARY

ill-defined problem Problem in which the given state, goal state, and/or operators are not clearly stated.

insight When a problem solver suddenly moves from a state of not knowing how to solve a problem to a state of knowing how to solve a problem.

nonroutine problem Problem for which the problem solver does not know a solution method.

problem When a problem solver wants to change a situation from its given state to a goal state but there are obstacles.

problem solving Cognitive processing directed at transforming a given situation into a goal situation when no obvious solution is available to the problem solver.

routine problem Problem for which the problem solver knows a solution method.

well-defined problem Problem that has a clearly stated given state, a clearly stated goal state, and a clearly stated set of operators.

A *problem* occurs when a person has a goal but does not know how to achieve that goal. This definition has three components: the problem is in the *given state*, the problem solver wants the problem to be in the *goal state*, and there are *obstacles* blocking the path between the given and goal states. Problem solving involves finding a way of moving from the given state to the goal state when there is no obvious path between them. For example, a problem occurs when a person wants to find the area of a parallelogram but does not know the solution formula. Problem solving occurs when the problem solver invents a way to solve the problem, such as cutting the triangle off one side and placing it on the other side to form a rectangle.

I. INTRODUCTION

Problem solving is cognitive processing directed at transforming a given situation into a goal situation when no obvious solution method is available to the problem solver. This definition consists of three parts. First, problem solving is *cognitive*—it occurs in the learner's mind. Because problem solving is an internal event, its occurrence can only be inferred indirectly by observing external events such as the learner's behavior. Observable changes in physiological state that happen while a problem solver is thinking are not the same as problem solving, but rather are physiological correlates or indicators of problem solving. Second, problem solving is a *process*—it involves a series of mental manipulations (or computations) performed on mental representations (or knowledge). A premise of cognitive science—the multidisciplinary study of cognition—is that the working of the mind can be characterized as a series of computations. Third, problem solving is *directed*—it is aimed at accomplishing a goal. Thus, directed thinking—including reasoning—is the same as problem solving, but

Portions of this article are adapted from Mayer, R. E., "Problem Solving," in the *Encyclopedia of Creativity* (1999), pp. 437–447, with permission.

nondirected thinking such as daydreaming is not problem solving. In short, problem solving is directed, cognitive processing. This definition is broad enough to include a wide array of cognitive activities, ranging from solving a puzzle to discovering a cure for a disease or from composing an essay to figuring out how to resolve an interpersonal conflict.

II. TYPES OF PROBLEM SOLVING

A. Routine versus Nonroutine Problems

It is customary in the problem-solving literature to distinguish between routine and nonroutine problems. A *routine problem* is a problem for which the problem solver knows a solution method. For example, for most adults, $567 \times 789 =$ ____, is a routine problem because they know the procedure for long multiplication. Routine problems can also be called exercises because they involve exercising procedures that the problem solver already knows. Routine problems depend on reproductive thinking; in reproductive thinking, problem solvers must reproduce responses that they have used in the past. Technically, routine problems do not fit the definition of a problem because the problem solver knows a solution method, so that there is no obstacle between the given and goal states.

A *nonroutine problem* is a problem for which the problem solver does not know the solution method. For example, for most adults, the following problem proposed by the Gestalt psychologist, Karl Duncker, is a nonroutine problem: "Why is it that all six-place numbers of the type abcabc, for example, 276276, are divisible by 13?" The solution occurs by realizing that all numbers of the form abcabc can be expressed as abc $\times$ 1001 and that 1001 is divisible by 13. Nonroutine problems are called insight problems because the problem solver must invent a new approach to the problem, that is, the problem solver must see the problem in a new way. Nonroutine problems depend on productive thinking in which problem solvers create a novel solution that they have never produced before. The cognitive processes involved in solving nonroutine problems may be different from those involved in solving routine problems, which may be reflected in different patterns of brain activity.

B. Well-Defined versus Ill-Defined Problems

It is also customary to distinguish between well-defined and ill-defined problems. A *well-defined* problem has a clearly stated given state, a clearly stated goal state, and clearly stated set of operations. For example, the game of chess is a well-defined problem because it has a clear given state (the board is set the same way for the start of every game), a clear goal state (the opponent's queen is unable to move), and clear operators (each piece can move in a certain way). In an *ill-defined problem*, the given state, goal state, and/or operators are not clearly stated. For example, the problem of how to have a good life is an ill-defined problem because the goal state (i.e., a good life) is not clearly stated and the operators (i.e., the things you are allowed to do to achieve this goal) are not clearly stated. Well-defined problems can be either routine or nonroutine; ill-defined problems can be either routine or nonroutine.

III. PROCESSES IN PROBLEM SOLVING

A. Problem Representation

There are two major processes in problem solving: problem representation and problem solution. In *problem representation*, a problem solver builds an internal mental representation of the problem based on a statement or presentation of the problem. In short, the problem solver comprehends the problem. Cognitive psychologists have further analyzed problem representation into two subprocesses: translation and integration. *Translation* involves mentally representing each sentence or portion of the problem. For example, we have the problem, "John has two marbles. Pete has three more marbles than John. How many marbles does Pete have?" In listening to this problem, a problem solver may mentally represent the first sentence as, "John's marbles = 2" and the second sentence as, "Pete's marbles = John's marbles + 3." *Integration* involves putting knowledge together into a coherent structure that can be called a situation model; the situation model is the problem solver's mental model of the problem situation. For example, in the marble problem the problem solver may construct a set–subset relation in which one set, "Pete's marbles" consists of two subsets: "John's marbles" and "the difference between Pete's and John's marbles."

B. Problem Solution

In *problem solution*, the problem solver devises and carries out a plan for solving the problem. In short, the

problem solver produces a solution to the problem. The process of problem solution includes the subprocesses of planning, executing, and monitoring. *Planning* involves devising a solution plan, that is, a method for solving the problem. For example, in the marble problem, the problem solver may decide to add 3 to 2. *Executing* involves carrying out the steps in the solution plan, that is, engaging in behaviors based on a plan. In the marble example, this involves determining that 5 is the sum of 3 and 2. *Monitoring* involves awareness and control of one's cognitive processing, including determining the extent to which the problem solution phase is successful and altering one's course if necessary. Monitoring is a metacognitive process because it involves awareness and control of one's cognitive processing. In the marble example, the problem solver may check to see that if Pete has 5 marbles, then the story in the problem makes sense. In addition, after completing the problem-solving process, the problem solver may engage in the subprocess of *evaluating*—reviewing the process of how the problem was solved, reflecting on its merits, and considering how it could be used in the future. For example, the problem solver may note that it is a "compare" problem in which one set is compared to another, so that the same solution procedure will work in other "compare" problems even if the sets do not involve marbles.

C. Example of the Processes in Problem Solving

George Polya analyzed the processes in solving the following geometry problem: "Find the volume F of the frustrum of a right pyramid with square base given the altitude h of the frustrum, the length a of a side of its upper base, and the length b of a side of its lower base." According to Polya, the first step is *understanding the problem*, which corresponds to the translation subprocess in problem representation. In this step, the problem solver asks, "What do I want?" and "What do I have?" In answering these questions, the problem solver mentally represents that the goal is to find the value of F and the givens are the values of a, b, and h. Polya's second step is *devising a plan*, which includes both the integration subprocess of problem representation and the planning subprocess of problem solution. This is the step in problem solving where creative insight occurs. To help the problem solver build a situation model of the problem, the problem solver can ask, "What is a related problem I can solve?" and "Can I restate the goal or givens differently?" In this step, the problem solver might already know a related problem —how to find the volume of a pyramid—and may realize that the goal may be stated differently as the difference between a full pyramid (including the frustrum) and the smaller pyramid on top of the frustrum. After building this coherent representation of the structure of the problem, the problem solver is able to create a plan: compute the volume of the large pyramid and subtract the volume of the small pyramid. Polya's third step, *carrying out the plan*, corresponds to the execution subprocess of problem solution. In this step, the problem solver makes the needed arithmetic calculations to compute the answer. Finally, Polya's last step, *looking back*, involves examining what has been done. This corresponds to the monitoring and evaluating subprocesses. In this step, the problem solver evaluates the logic of the solution.

Although the subprocesses in problem solving can be neatly defined, they rarely occur in a linear order. In most complex problem solving the subprocesses interact with one another; for example, a problem solver may begin by trying to represent the problem, then try to plan, then go back to modify the problem representation, then try to execute the plan, then monitor, then revise the plan, and so on. Most of the difficulty in problem solving concerns building a coherent problem representation (i.e., integrating) and devising and monitoring a solution plan (i.e., planning and monitoring).

IV. APPROACHES TO PROBLEM SOLVING

For more than a century, psychologists have grappled with the thorny question of how the human mind works. What goes on inside someone's mind when they are thinking, that is, when they come up with a solution to a problem? Three different ways of answering this question come from the associationist, Gestalt, and cognitive science approaches.

A. Associationist

In the early 1900s, associationist theory developed as psychology's first large-scale account of how the human mind works. The approach received a large boost from the landmark work of Edward L. Thorndike, including his classic book, *Animal Intelligence*, which was published in 1911. According to this view, knowledge is a network consisting of nodes and associations among them. The strength of the

association between two nodes depends on the experience of the learner. Problem solving involves beginning at one of the nodes and following a chain of associations to other nodes, always taking the association that is the strongest.

For example, in a classic set of studies, Thorndike placed a hungry cat in a puzzle box. If the cat performed a simple response, such as pulling a loop of string, a door would open and the cat could escape to a bowl of food sitting nearby. On the first trial, the cat performed many unsuccessful responses, such as clawing through the bars, pouncing against the walls, and meowing, until it accidentally pulled the string and got out. However, across many trials, the time needed to get out of the puzzle box declined as did the number of unsuccessful responses. On the basis of results like these, Thorndike devised the law of effect: if a response is followed by satisfaction, it is more likely to recur, and if a response is followed by dissatisfaction, it is less likely to recur. Thus, each time the cat engaged in an unsuccessful response, the association between the problem situation and that response was automatically weakened, and each the cat performed the successful response the association between the problem situation and the response was automatically strengthened. Thus, Thorndike characterized the cat's problem-solving process as learning by trial and error and accidental success. According to the associationist view, problem solving is simply a matter of exercising existing associations. One of the major criticisms of this approach to problem solving is that it fails to adequately account for creative problem solving.

B. Gestalt

In the 1930s and 1940s, Gestalt theory provided an important alternative to associationist theory, culminating in psychology's second account of how the human mind works. The approach is reflected in the work of the Gestalt psychologists such as Wolfgang Kohler's *The Mentality of Apes* originally published in 1925, Karl Duncker's *On Problem Solving* published in 1945, and Max Wertheimer's *Productive Thinking* originally published in 1945. According to this view, problem solving occurs when the problem solver mentally reorganizes the problem situation in a new way. Problem solving is not a matter of following preexisting associations but rather requires structural insight—seeing how the parts of the problem fit together to achieve the goal. *Insight* occurs when a problem solver suddenly moves from a state of not knowing how to solve a problem to a state of knowing how to solve a problem.

For example, in a classic set of studies, Kohler observed apes as they attempted to solve problems. In one problem, bananas were hung high in a cage beyond the reach of the ape, and crates were scattered about the floor of the cage. The ape looked around the room and then, in a flash of insight, stacked the crates on top of one another to form a ladder that he then climbed to reach the bananas. According to Kohler, the ape had a flash of insight in which he mentally organized the parts of the problem by seeing how the crates could be used as a ladder. Whereas Thorndike had observed problem solving as following preexisting associations, Kohler observed problem solving by insight. A reconciliation of these apparently conflicting views is that associationist theory explains one kind of problem solving, whereas Gestalt theory explains a different kind of problem solving. A major criticism of the Gestalt approach is that it is too imprecise.

C. Cognitive Science

During the second half of the 20th century, psychology produced a third approach containing elements of the preceding two—the cognitive science approach. Cognitive science is the multidisciplinary study of cognition and is based on the idea that cognition involves mental computations—operations carried out on mental representations. Cognitive science has its roots in the information-processing approach to problem solving, in which humans are assumed to be processors of information. For example, in their classic book, *Human Problem Solving*, Allen Newell and Herbert Simon showed how a computer program could simulate the problem-solving processes of humans on a wide variety of problems. According to this view, an important goal of cognitive science is to understand the ways that knowledge is represented and how it is transformed by mental operations.

V. TOPICS IN PROBLEM SOLVING

A. Problem-Solving Expertise

What do expert problem solvers know that novices to not know? This is the question that motivates research on problem-solving expertise. For example, classic research comparing expert and nonexpert chess

players has shown that they differ on domain-specific skills, such as being able to remember the location of pieces in an actual game, but not on domain-general skills, such as having a better memory in general. Similarly, experts in computer programming outperform beginners in being able to read and reconstruct a computer program from memory but not in being able to remember a random list of programming commands. This line of research is consistent with the idea that problem solving is domain-specific, that is, the knowledge and skills required for problem solving in one domain (such as chess or programming) are not the same as the knowledge and skills required for problem solving in another domain (such as essay writing).

B. Thinking by Analogy

All problem solving involves thinking by analogy. This is the hypothesis that is examined by research on thinking by analogy. For example, when people are asked to solve a problem about electrical flow, such as what happens to the rate of flow of electrons when there are two resistors in series rather than one, they tend to use a familiar analogy such as thinking of the circuit as water flowing in pipes. Research on analogical thinking shows that people often have difficulty in recognizing that the problem they are currently trying to solve is similar to a problem they already know. Thus, a major factor in successful problem solving is the ability to find a problem analog, that is, to be able to think of a related problem that the problem solver already knows how to solve.

C. Teaching Thinking

Another way to study problem solving is to analyze the processes in problem solving and teach them to people. If the instruction improves people's problem-solving skills, then this validates the processes as being important components in problem solving. Research on teaching thinking shows that successful programs generally focus on teaching specific problem-solving processes that fit within a specific domain rather than trying to improve the mind in general for all problem-solving domains.

D. Subject-Based Cognition

How do students think mathematically? How do students think scientifically? How do students think historically? These questions reflect a growing research literature on the psychology of subject matter. Importantly, the cognitive processes involved in each subject appear to be different from one another—mathematical problem solving is different from scientific problem solving, which is different from historical problem solving, and so on. For example, even Polya's analysis of problem-solving processes (summarized in a previous section) seems to be closely tied to the domain of solving geometry problems. Thus, planning a solution to a geometry problem might be quite different from planning how to write an historical essay, even though both involve the problem-solving process of planning.

E. Everyday Thinking

Do people use the same problem-solving processes in their everyday lives as they use in school or laboratory situations? Increasingly, research on problem solving in naturalistic contexts reveals that people's problem-solving strategies are dependent on the situations in which the problems are encountered. When asked to determine the best buy (e.g., 4oz for 45¢ or 10oz for 90¢), most people compute the unit cost in a mathematics class (11.25¢ per ounce versus 9¢ per ounce), but use a ratio strategy in a real supermarket (the larger one costs exactly twice as much but contains more than twice as many ounces). This research contributes to theories of situated cognition in which all cognitive activity depends on its context rather than on applying general cognitive procedures.

VI. CONCLUSION

Some dominant themes in current research on problem solving include the domain specificity of cognition, the focus on component cognitive processes, the situated nature of thinking, and the role of metacognitive control in cognitive processing. Although the field of problem solving has not yet achieved a single perspective or unifying theory, progress continues to be made in understanding how people solve problems.

See Also the Following Articles

ARTIFICIAL INTELLIGENCE • CATEGORIZATION • CREATIVITY • HEURISTICS • INFORMATION PROCESSING • INTELLIGENCE • LOGIC AND

REASONING • NUMBER PROCESSING AND ARITHMETIC • PATTERN RECOGNITION

Suggested Reading

Duncker, K. (1945). On problem solving. *Psychol. Monogr.* **58**(5), 270.

Gilhooly, K. J. (1996). *Thinking: Directed, Undirected, and Creative*. Academic Press, London.

Guenther, R. K. (1998). *Human Cognition*. Prentice-Hall, Upper Saddle River, NJ.

Kohler, W. (1976). *The Mentality of Apes*. Liveright, New York.

Manktelow, K. (1999). *Reasoning and Thinking*. Psychology Press, Hove, UK.

Mayer, R. E. (1992). *Thinking, Problem Solving, Cognition*, 2nd ed. Freeman, New York.

Newell, A., and Simon, H. A. (1972). *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ.

Sternberg, R. J. (Ed.). (1994). *Thinking and Problem Solving*. Academic Press, San Diego, CA.

Thorndike, E. L. (1911). *Animal Intelligence*. Macmillan, New York.

Wertheimer, M. (1959). *Productive Thinking*. Harper & Row, New York.

Prosopagnosia

EDWARD H. F. DE HAAN
Utrecht University

GLOSSARY

affective prosopagnosia An impairment in the appreciation of facial expressions.

agnosia A recognition disorder that cannot be explained by sensory–perceptual, language, or amnestic deficits or general confusion. Agnosias are often modality-specific, i.e., pertaining to one sensory modality (e.g., visual agnosia).

anomia A naming impairment.

apperceptive agnosia A recognition disorder that is caused by a higher order perceptual impairment.

associative agnosia A recognition disorder that is caused by a problem in accessing stored information about objects and faces.

covert face recognition The experimental demonstration of preserved face recognition in prosopagnosia despite a complete absence of acknowledged awareness.

ERP Event-related potential.

fMRI Functional magnetic resonance imaging.

GSR Galvanic skin response.

metamorphopsia The clinical condition in which patients perceive faces in a distorted manner.

object agnosia A recognition disorder for common objects. In pure cases, other stimulus categories such as text or faces may remain unaffected.

person identification deficit The clinical condition in which a patient is able to distinguish between familiar and unfamiliar people but cannot access biographical information.

PET Positron emission tomography.

priming An experimental procedure in which the processing of a particular stimulus is influenced by a previous presentation of that stimulus (repetition priming) or a related stimulus (semantic priming).

speech reading The perception of the movements of the lips and the tongue to aid the understanding of spoken speech.

within-category discrimination The ability to identify specific exemplars of stimulus category with many visually similar items (e.g., motor cars).

The term *prosopagnosia* refers to the inability to recognize familiar people via visual inspection of the face. It is a modality-specific impairment, and other recognition routes, such as via the voice, remain spared. In addition, reading and object recognition may remain unaffected in selective cases. The inability cannot be explained by perceptual, language, or amnestic deficits or by general confusion.

I. BACKGROUND

Anecdotal observations of face recognition deficits date back to ancient Greek literature. For instance, the Greek general Thucydides described intriguing behavior, such as the inability to recognize friends, in soldiers who recovered from the plague. Clinical case studies published in the second half of the nineteenth century sometimes mentioned problems in this domain in patients who have suffered neurological disease. But it was not until the early twentieth century that clinicians suggested that face recognition might be a

separate function that can be selectively impaired after brain injury.

In 1947, the German neurologist Joachim Bodamer published a report on a number of patients who experienced a particular selective difficulty in recognizing faces. Bodamer named the condition prosopagnosia with reference to the Greek *prosopon* (face) and *agnosis* (without knowledge). His patients complained bitterly about the problem. In severe cases, even the faces of family members and close friends were affected and sometimes the patient's own face when seen in a mirror. When Bodamer asked his patient to inspect his own face in the mirror, it was clear that, although he knew he was looking at his own face, there was no sense of familiarity. This problem does not appear to be directly caused by an impairment in the perception of facial features. All of his prosopagnosic patients knew when they were looking at a face, and they could identify and describe separate features such as the eyes and mouth. Thus, although prosopagnosic patients know that they are looking at a human face, it has lost its value as a cue for the identification of a person, and not even a vague feeling of familiarity is evoked when the patient is looking at a known face. Prosopagnosic patients often describe faces as all looking similar, unattractive, and having lost their individuality. Another common comment from these patients is that faces are not perceived as an integrated whole and appear fragmented.

In order to suggest that prosopagnosia constitutes a separate clinical entity, it is important to look in more detail at the claim that these deficits in the recognition of familiar faces dissociate from visual object agnosia in general. There are many case descriptions suggesting that the patient is unable to recognize any familiar faces, whereas his or her ability to recognize common objects is spared. However, it should be noted that the task demands in face and object recognition are not comparable. Instead of recognizing a hammer as a tool, a more comparable situation would be to recognize one's neighbor's hammer among many other exemplars.

The ability to recognize faces is probably the most difficult visual recognition task that we are able to perform, despite the subjective ease with which we do it. In a lifetime, we learn thousands of faces, most of which we are able to recognize relatively unaffected by changes due to age and variable additions such as spectacles and facial hair. These faces, however, only differ very slightly in visual appearance. This is a point well-illustrated by the general experience that faces from another race than one's own are difficult to remember. When prosopagnosic patients are tested on within-class recognition tasks, such as flower or car recognition, that approach the task demands involved in face recognition, they often show impairments. However, there is evidence that within-class discrimination problems do not always co-occur in prosopagnosia.

An extensive investigation of prosopagnosia by the Italian neurologist Ennio de Renzi is illustrative here. He described a prosopagnosic patient who was able to perform recognition tasks that are visually very demanding. For instance, this patient was able to recognize his own car in the car park, his own wallet from an array of similar wallets, and his own handwriting amid that of others. In addition, he was very apt at distinguishing Italian coins from foreign ones. Until very recently, the controversy about a dedicated cortical system for face recognition continued because the other half of the necessary double dissociation was still lacking, despite a number of clinical reports that described a statistical trend toward relatively more severe problems in recognizing objects than faces. Convincing evidence for this position was provided by a number of reports on a patient who performed normally on tests of familiar face recognition while remaining severely impaired in the recognition of common objects.

Other indirect evidence comes from the clinical phenomenon of metamorphopsia, where the patient sees faces—and only faces—in a distorted manner. For example, one patient reported that his visual world appeared to be completely unaffected by his brain damage apart from the fact that all faces looked like "fish heads." Despite this distortion, he was able to recognize most familiar faces without any problems. The fact that this phenomenon is restricted to faces supports the notion of a separate cognitive system that is dedicated to the perception of faces.

II. PROSOPAGNOSIA AND SENSORY STATUS

Visual recognition disorders as a result of brain disease have created theoretical controversies since they were first described in the latter half of the nineteenth century. Regarding impairments in object recognition, Heinrich Lissauer suggested a distinction between apperceptive and associative Seelenblindheit, where the former results from visuosensory deficits, whereas the latter was thought to represent a difficulty in associating meaning with a relatively intact percept. A few years later, Sigmund Freud coined the term

agnosia to describe these recognition disturbances. It is the viability of the concept of associative agnosia, or as Hans-Lukas Teuber put it, "a normal percept stripped of its meaning" that has been questioned. Notably, Eberhard Bay claimed that the so-called higher order recognition deficits are secondary to sensory impairment, general intellectual loss, language problems, or a combination these factors.

More recent investigations have convincingly demonstrated that visual recognition disorders can occur in patients with normal, or even above average intelligence. Also, it has now been clearly established that language difficulties are not instrumental in causing associative agnosias. However, the question of whether subtle sensory impairment or a certain constellation of sensory deficits can produce the clinical symptoms of associative agnosia remains controversial.

The debate is fueled by the fact that even the most "pure" cases of agnosia often show some mild problems on tasks of visual perception. A seminal study by George Ettlinger in 1956 set out to test Bay's hypothesis. He carried out a careful assessment of sensory status in patients with and patients without recognition deficits, and he argued that sensory status alone could not explain the presence or absence of a recognition disorder. The crux of his argument is that, although patients with a recognition deficit may have sensory impairments, other patients who do not experience recognition problems can show equal or worse impairments on the sensory tests. George Ettlinger was subsequently criticized for using a composite sensory score and for using tests that might not have covered all of the relevant sensory abilities for visual recognition. There is now strong evidence for relatively independent functional components in human visual processing.

Neuropsychological research has shown highly selective disorders of color, luminance, shape, location, and movement. In a study with three densely prosopagnosic patients, the evaluation of primary visual proficiency was done with a screening battery comprising all of the relevant visual cues, and the patients showed subtle impairments on several of the screening tasks. These problems were, however, insufficient to explain the pronounced recognition problems, because other patients with unilateral brain lesions who do not have recognition deficits are at least as impaired on these sensory tasks as the agnosics.

The conclusion is, therefore, that the visuosensory impairments cannot explain the agnosic problem. These results support the notion of an associative type of prosopagnosia, which can be distinguished from a perceptive type. This is in line with more recent studies in which it is concluded that there are both perceptual and amnestic forms of prosopagnosia. Thus, prosopagnosia is modality- (visual) and stimulus-specific (faces) impairment that may arise in the context of preserved visuoperceptual processing.

III. PROSOPAGNOSIA AND OTHER FACE PROCESSING DEFICITS

Faces form the source for a multitude of inferences. From a perceived face, we are able to determine gender and age and make more subjective judgments regarding attractiveness, etc. We can assess the emotional state of the person (e.g., happy, sad) by analyzing the facial expression, and by observing the movements of the lips and tongue, we gain additional information regarding the verbal message that the speaker is trying to convey. Finally, the face constitutes the principal cue for visual identification of people we know.

Evidence for a dissociation between the recognition of familiar faces and the processing of facial expressions can be traced back to neuropsychological studies in the 1960s. It was observed that some patients with severe impairments of facial identity recognition showed relatively intact ability to recognize facial expressions. Conversely, a group study with patients suffering from a degenerative illness showed that they were impaired at both recognizing facial expressions and recognizing American Presidents. However, the identity and expression recognition impairments did not correlate with each other. The conclusion is that some of these patients must have had intact familiar face recognition but poor facial expression perception. These patients would provide the double dissociation for the previously mentioned studies.

There is also evidence for distinct processing pathways involved in the recognition of facial expressions and the matching of unfamiliar faces. For example, in a 1985 study by Dawn Bowers and colleagues, patients who had left and right hemisphere damage or no neurological disease were assessed using a series of facial affect tests and tasks for unfamiliar face perception. The patients with right hemisphere lesions were significantly impaired on both types of tasks, but the impairments on the affect tests remained significant even when the face perception deficit was statistically partialed out. Several reports subsequently supported the notion of dissociable impairments of facial expression recognition and unfamiliar face

matching after brain injury. A selective deficit in the analysis of facial expressions is sometimes referred to as affective prosopagnosia. Regarding the neuroanatomical substrate of expression analysis, functional imaging studies suggest that the right posterior hemisphere is obviously instrumental in the perceptual processing of facial information, but the amygdala (bilaterally) plays a crucial role in the experience of emotions, especially for fear and anger.

In 1986, Ruth Campbell and colleagues described two Swiss patients. One patient was able to recognize faces but could no longer read lips, and the other showed the opposite pattern. Lip reading is tested using a recognition task in which the patient is shown photographs of the face of an actor making a speech sound, for instance ee or aa. The other method uses what is known as the McGurk illusion. This involves the videotape recording of a face making a speech sound (e.g., da) and the simultaneous sound recording of a different speech sound (e.g., ta). A normal observer confronted with such a stimulus would "hear" a blend of these two speech sounds (e.g., ga). Subjects who can no longer read lips will report the auditory stimulus and not the blend. Apart from this single paper there is little neuropsychological evidence for this dissociation, although there is some convergent evidence from cognitive psychology.

The literature also supports the notion of distinct pathways for the processing of familiar and unfamiliar faces. Prosopagnosia patients may perform normally on tasks of matching unfamiliar faces. Conversely, patients with unfamiliar face matching impairments often are not clinically prosopagnosic. One of the first clear examples of a double dissociation came from a study by Daniel Malone. Case 1 initially was unable to recognize familiar faces, and he was also impaired on the matching of unfamiliar face photographs. Later testing revealed that the familiar face recognition impairment had improved significantly, whereas his unfamiliar face matching problem remained. Case 2 was unable to recognize both relatives and various familiar personalities by their faces alone 6 weeks after an initial testing. There was a significant improvement, however, in his initially impaired unfamiliar face matching performance, which now fell within the normal range. The idea of independence of familiar face recognition from the perception of unfamiliar faces in terms of gender decision is supported by studies with normal subjects.

In short, prosopagnosia is a selective deficit in the recognition of familiar faces that may arise in the context of preserved recognition of other types of information that can be read from faces, such as expressions and speech gestures.

IV. TOWARD A TAXONOMY OF PROSOPAGNOSIA

It is now becoming clear that there are different forms of prosopagnosia that can be distinguished in terms of the underlying functional deficit. The distinction between apperceptive and associative prosopagnosia has already been introduced. Another important form of face recognition deficit could be termed "person identification deficit." The patient M.E. was a right-handed female who, after a vasculitic disorder, suffered from a selective memory deficit. She performed normal to above normal on tasks for the perception of faces (e.g., Benton facial recognition test) and on tasks that required her to choose the familiar face from an array with unfamiliar faces. However, she was very poor at retrieving personal information about the people whose faces she had recognized. Often she would claim that a face looked familiar but that was all she could say about it. These findings provide a strong indication that her face recognition system is unimpaired to the level of familiarity recognition. The dissociation between face familiarity and access to biographical information was confirmed on formal testing.

A related impairment concerns a specific anomia. In 1989, Brenda Flude and colleagues studied a patient with an impairment at the last stage of face recognition, that is, retrieval of the name. Their patient, E.S.T., showed a preserved ability to (a) match unfamiliar faces on identity, (b) distinguish between familiar and unfamiliar faces and names, and (c) access semantic information from faces and written names of familiar people. However, E.S.T. was severely impaired in naming familiar faces. The finding of this selective functional deficit supports the idea that the mechanism dealing with name retrieval (which was severely impaired for E.S.T.) is separable from mechanisms responsible for the sense of familiarity and access to appropriate semantic information (which were both well-preserved for EST). Flude's investigation of E.S.T., then, demonstrates that name retrieval can be impaired when access to identity-specific semantic information is preserved. A more recent study with Alzheimer's patients showed that there were no patients who could name a face without producing any autobiographical information concerning that person. This again supports the proposal that

name retrieval is a separate stage that takes place after access of personal information.

It is concluded that there are several forms of prosopagnosia that can be distinguished on the basis of the functional locus of the deficit. First, higher order perceptual impairments may lead to face recognition problems. In line with Lissauer's suggestion some 100 years ago, there is an amnestic form. Subsequently, patients may be impaired in accessing biographical information about familiar people, and finally there are patients who are unable to retrieve the names of familiar people. In addition, it is useful to point out that prosopagnosia entails, in fact, two impairments: the inability to recognize familiar faces and the inability to learn new faces or relearn old ones. These problems have subsequently been described in isolation. Transient forms can be interpreted as retrograde prosopagnosia, whereas anterograde prosopagnosia refers to those who are no longer able to learn new faces but can recognize faces learned before the onset of the condition.

V. THE ANATOMY OF FACE PERCEPTION

Face processing disorders have been described in a wide range of patient populations. Patients suffering from diffuse brain damage such as dementia or closed head injury have been shown to suffer from severe deficits in face processing. Progressive visual recognition problems, notably for faces, are the prevailing clinical symptom in patients with a degenerative illness that specifically affects the posterior areas of the cortex. Several studies have shown that autistic patients demonstrate problems in the perception and recognition of faces. A number of patients with visual recognition deficits, including prosopagnosia, have been described after carbon monoxide poisoning. Face processing disorders have also been described in patients with neurological illnesses resulting in more localized damage. For instance, temporal and frontal lobectomy may impair the perception of faces and facial expressions. In addition, the famous prosopagnosic case H.J.A., who has been studied extensively by Glyn Humphreys and Jane Riddoch, had suffered a large posterior stroke.

Prosopagnosia is, in most cases, the result of bilateral posterior damage to the cortex. The left and right occipitotemporal junctions have traditionally been identified as the substrate for face recognition. Most patients with prosopagnosia (and all patients who have come to post mortem investigation) had lesions in those areas. In humans, the fusiform and lingual gyri are most often implicated. These observations are somewhat at odds with the primate single-cell recording studies that localize the cells sensitive to face identity much more anteriorly in the upper bank of the superior temporal sulcus. Obviously, these differences could be explained as differences between species, nonetheless it might be prudent to keep this in mind when comparing the human and the primate data on face perception. In this context it is interesting to note that a study found face-specific cells in the prefrontal area of the macaque monkey.

A number of reports have claimed more recently that a unilateral right hemisphere lesion might be sufficient to cause prosopagnosia, but the status of these demonstrations remains somewhat unclear as long as they depend on neurosurgical reports and neuroimaging instead of post mortem findings. It is known from clinical practice that transient or partial prosopagnosia can occur after a unilateral lesion but that full-blown prosopagnosia appears to require bilateral damage.

The upsurge in functional neuroimaging studies has largely served to confirm the existing evidence from neuropsychology. MEG has supported the notion of bilateral processing as well as the importance of the right inferior temporal lobe for familiar face recognition. Work using PET and fMRI scanning have identified a number of areas involved in the perception of familiar faces in the posterior, occipital–temporal areas of the brain, and again the fusiform gyrus—especially on the right side—appears to play a major role.

In 1999, using fMRI, Nancy Kanwisher and colleagues provided further evidence for a specific face processing system, demonstrating that the face area is not especially activated when the subjects are presented with pictures of animals. Perhaps the most interesting result to come out of neuroimaging until now is the that the area involved in face processing is actually adjacent to the area that is involved in object recognition. This raises the possibility that this area processes visual primitives that happen to be important for face recognition. If these observations are replicated and extended, they might well form the basis for a more tractable definition of what is specific about the face area.

The general consensus is that prosopagnosia is caused by bilateral damage to the ventral fusiform gyrus. There are a number of controversial issues, such as the possibility that a unilateral lesion may be sufficient, and whether there is a specific face area or

whether it is just a part of a general visual object processing area. However, the main challenge for future research must be to use the functional subdivisions of prosopagnosia in order to arrive at a more fine-grained localization.

VI. PROSOPAGNOSIA AND KNOWLEDGE WITHOUT AWARENESS

In 1994, Russell Bauer published his seminal paper that fueled the research on face recognition and awareness. He investigated a patient who had become prosopagnosic after a severe closed head injury. First, he asked his patient to select the correct name from five alternatives to match a photograph of a familiar face. All five alternative names were of celebrities and, therefore, familiar to the patient (who had no problem recognizing names). As expected of a prosopagnosic patient, L.F. performed at chance level. However, skin conductance responses recorded during the experiment occurred significantly more often and with higher amplitude to the correct name than to the other four foils. Daniel Tranel and Antonio Damasio have subsequently replicated these surprising findings. They recorded significantly increased autonomic responses when their prosopagnosic patients looked at slides of familiar faces embedded among those of unknown people.

These observations have been corroborated using other psychophysiological techniques. Differential processing of familiar and unfamiliar faces in prosopagnosia has been demonstrated with ERP measures using an odd-ball paradigm in which the patients looks at a long series of faces. Most of the photographs are of completely unfamiliar people, but once in a while the face of a very famous person is embedded in the series. Each time a famous face appears there is a clear increase in the amplitude of the P300 response.

Finally, the recording of eye movements has suggested spared processing of familiar faces in prosopagnosic patients. It has been demonstrated that the way in which normal people look at faces depends on whether they know the person. Unfamiliar faces are scanned in a rather global manner with about an equal amount of attention given to all aspects of the face. However, with familiar faces the emphasis is clearly on the internal part of the face (eyes–nose–mouth region). Although the prosopagnosic patient is unaware of the fact that he is looking is a familiar face, his scanning behavior is as if he is looking at a familiar face. These demonstrations of differential processing of familiar and unfamiliar faces despite the inability to recognize the familiar faces overtly, thus, appear robust.

These data were first interpreted as indicating two separable recognition systems, one for overt and another for covert recognition. The overt system would result in a conscious experience of recognition, whereas the covert system would feed into the limbic system and serve an alerting function. However, there were also some indications that covert recognition might not be restricted to autonomic or physiological measures. Earlier studies had shown that a partially prosopagnosic patient had shown some degree of preserved knowledge of familiar faces of which he did not appear to be aware. These suggestions have been extensively followed up by our research group in a number of studies with the patient P.H., who is completely unable to recognize familiar faces overtly. The extent of P.H.'s recognition problems is well-illustrated by his performance on a forced-choice task where he had to choose the familiar face from an array of two faces. This test procedure is sensitive to small degrees of residual processing as it is not influenced by possible language or memory dysfunction or subject to the common response bias of agnosic patients who respond "I don't know" in standard line-up confrontation tasks. P.H. performed this task at chance level (correct on 51% of the trials), indicating that he has no access to familiarity information from faces.

Next, in seeking to demonstrate covert knowledge, we used experiments that have been shown to be sensitive to knowledge of face familiarity in healthy individuals but that do not require overt identification of the famous faces used as experimental stimuli. Such experiments used the procedures of matching, interference, associative priming, and paired-associate learning. For instance, in the matching experiment, subjects are required to decide whether two simultaneously presented photographs of faces are taken from the same person or two different people. Normal subjects are faster to match photographs of familiar than unfamiliar people, and P.H. showed exactly the same effect, yet he was unable to identify any of the faces used. Covert processing can also be demonstrated in associative priming experiments. Such experiments show the influence of previously presented stimuli on a subsequent response. Subjects have to decide whether the targets (i.e., written names of familiar and unfamiliar people) are familiar. The responses are influenced by the presentation of a primeing stimulus (e.g., a face) shortly before the target appears. If there is a strong association between the prime stimulus (e.g., a photograph of Prince

Charles) and the target (e.g., the name of Princess Diana), the subject responds faster than when the prime is either the face of an unfamiliar person or that of a familiar person who is not closely associated with the target (i.e., the face of Prince Charles followed by the name of George Bush). The patient, P.H., shows this type of priming effect with faces that he does not overtly recognize. Moreover, we were able to compare the extent of the priming effect with that triggered by name primes (which P.H. recognizes without difficulty). The effects are equivalent: overt recognition of name primes makes no additional contribution to the associative priming effect. This suggests that P.H. not only covertly recognizes faces but that his recognition is normal in the amount of associative priming that it produces.

It should be noted that covert recognition effects are not invariably found in agnosias. Some prosopagnosic patients do not show any covert effects at all, and others have exhibited partial covert recognition effects, such as covert recognition effects for only a subset of the people known to the patient.

The observation of covert face recognition, in its different forms, has extended the taxonomy of prosopagnosia. Apart from the difference between apperceptive and associative prosopagnosia, we can now distinguish between prosopagnosia with and without covert recognition. This difference is possibly related to the issue of whether the stored representations of familiar faces are still spared in a prosopagnosic patient. Several authors have commented on the use of mental imagery tests to evaluate the integrity of the stored representations. A number of patients have been described who are able "picture faces in their mind" that they cannot recognize overtly. Obviously testing of such a subjective ability as mental imagery is not easy, but a number of useful procedures have been developed. The method of odd-one-out, in which the patient has to decide which one of three people looks most unlike the other two (e.g., the patient is given the names: Charlie Chaplin, Adolf Hitler, and Bill Clinton) has been used successfully to demonstrate clinical dissociations between impaired recognition and spared mental imagery of familiar faces.

VII. DEVELOPMENTAL PROSOPAGNOSIA

The ability to recognize familiar faces can, thus, be characterized as modular in terms of the ideas put forward by Jerry Fodor. It is a highly specialized function with a well-circumscribed neurological substrate that appears to be hard-wired at birth. Faculties are modular as a consequence of their importance to the individual and, therefore, are designed to mature and develop in an efficient and relatively protected way. This modular, hard-wired nature of face recognition suggests that the variability in capacity in the general population can be attributed largely to biological factors. A genetic basis for language processing is now generally accepted. Developmental face recognition problems have been described in a handful of case studies. By definition, the problem is present from early childhood. In addition, there should be no known neurological history that could explain the recognition problem. There is one case of developmental prosopagnosia that has been followed extensively. This is an intelligent and verbal woman, who experienced very severe difficulties in recognizing familiar faces. Reading and object recognition were preserved, but she encountered problems on most tasks that depend on the use of facial information, such as the visual analysis of expressions and the short-term memory of unfamiliar faces. Her problem is best described as an inability to form an adequate internal representation of faces. As a result, faces have never gained the significance for her that they have for most of us. Instead, she relies heavily on voices for the recognition of familiar people.

An interesting observation, mentioned in some case studies on developmental face recognition deficit, concerns the anecdotal reference to other family members who are supposedly also poor at recognizing faces. One study concerns a family of two parents, three daughters, and one son. Two of the daughters and the father were very poor at recognizing familiar faces. The other family members did not have any problems. Thus, the face recognition problems occurred not only in two family members of the same generation but also in family members of different generations. These observations present clear evidence for a familial factor in the development of face recognition problems.

VIII. CONCLUSIONS

In clinical terms, it has become clear that, although pure prosopagnosia is rare, face recognition impairments are much more common than has been previously suggested. The reasons that these problems often remain undiagnosed are due to the fact that patients do not spontaneously complain either as a result of lack of insight or because the difficulties are

ascribed to a general memory deficit and/or because the assessment of face recognition deficits is often not included in standardized neuropsychological screening. In most cases, face recognition deficits occur in the context of more widespread perceptual problems. The effects on daily living are often underestimated.

Prosopagnosia severely reduced the professional and social possibilities of the patient and in many cases has led to social isolation. Finally, there have been several attempts at developing rehabilitation programs, but until now without much success. Memory training has been successfully applied in patients who can recognize faces but have problems with retrieving the name. However, even patients with intact covert recognition do not appear to benefit from cognitive training.

In terms of understanding the underlying mechanisms of face recognition, the study of prosopagnosic patients has produced a number of intriguing and often counterintuitive results during the past two decades. If the fractionation of face and object recognition remained controversial until quite recently, we have now seen the general acceptance of a specialized face processing system that itself fractionates further into a system responsible for familiar face recognition, expression analysis, face perception skills, such as gender decision and age estimation, and perhaps lip reading. Within the processing route involved in familiar face recognition, separate sequential processing stages have been identified for the processing of the incoming visual image and access to stored representations of faces, autobiographical information, and names. These processing stages are largely automatic and escape conscious introspection as demonstrated by the phenomenon of covert face recognition where these automatic processes continue to operate despite the fact that the output never reaches conscious awareness. Functional imaging techniques have produced new data suggesting that the apparent selectivity of face processing can actually be traced back to the functional organization of the fusiform gyrus, where there may be a continuous representation of form information that has a highly consistent and orderly topographical arrangement. The face area constitutes the part that entails the form detectors that are particularly important for defining faces. The challenge for future research is to refine the functional architecture of face recognition in order to understand the neuroanatomical basis of the different forms of prosopagnosia that can be distinguished.

See Also the Following Articles

AGNOSIA • ANOMIA • INFORMATION PROCESSING • OBJECT PERCEPTION • PATTERN RECOGNITION • PRIMING • SALIENCE

Suggested Reading

Bruce, V., and Humphreys, G. W. (1995). *Object and Face Recognition*. Lawrence Erlbaum Associates, Hove.

Bruce, V., and Young, A. W. (1998). *In the Eye of the Beholder*. Oxford University Press.

De Haan, E. H. F. (1999). Covert face recognition and anosognosia in prosopagnosia. In *Case Studies in the Neuropsychology of Vision* (G. W. Humphreys, Ed.), pp. 161–180. Lawrence Erlbaum, Hove.

De Haan, E. H. F. (2000). Face perception and recognition. In *A Handbook of Cognitive Neuropsychology* (B. Rapp, Ed.), Psychology Press, London.

Kanwisher, N., and Moscovitch, M. (2000). *The Cognitive Neuroscience of Face Processing. A Special Issue of the Journal of Cognitive Neuropsychology*, Vol. 17. Psychology Press, London.

Psychoactive Drugs

DAVID E. PRESTI
University of California, Berkeley

GLOSSARY

agonist A drug that binds to a neurotransmitter receptor and mimics the action of an endogenous neurotransmitter.

antagonist A drug that binds to a neurotransmitter receptor and has no effect other than to block the action of endogenous neurotransmitters and other receptor agonists.

gamma-aminobutyric acid The major inhibitory neurotransmitter in the human brain.

glutamate An amino acid that functions as the major excitatory neurotransmitter in the human brain.

neurotransmitter A chemical that is released by nerve cells and diffuses extracellularly to bind to receptors on nerve cell membranes, thereby mediating intercellular communication of neural signals.

neurotransmitter receptor Proteins found in nerve cell membranes that specifically bind neurotransmitters, undergo conformational change, and thereby mediate changes in cell membrane electric potential or initiate various intracellular biochemical processes.

psychoactive drug A chemical that in small amounts influences the functioning of the human brain in such a way as to have effects on the psyche or mind.

The human brain is the most complex object in the known universe. Its complexity is apparent in its 100 billion or more neurons interconnected by trillions of synapses. Each synapse is a stage for an intricate molecular drama involving release of chemical neurotransmitters, rapid diffusion of neurotransmitter across the narrow gap separating the presynaptic from the postsynaptic neuron, binding of neurotransmitter to receptor proteins located in the membranes of the postsynaptic and presynaptic neurons, resulting conformational changes of these receptors, and subsequent changes in membrane potential or initiation of various intracellular biochemical processes. The neuronal signaling process is terminated when neurotransmitter is taken back up into the neuron that released it via membrane proteins called reuptake transporters or is otherwise enzymatically degraded.

I. INTRODUCTION

Dozens of different molecules have been identified as neurotransmitters in the human brain. These include glutamate and gamma-aminobutyric acid (GABA), the main excitatory and inhibitory neurotransmitters in

the human brain. Others include serotonin, dopamine, norepinephrine, epinephrine, acetylcholine, histamine, adenosine, glycine, anandamide, and the opioid peptides. Many more neurotransmitter molecules no doubt remain to be discovered. The diversity of chemical neurotransmitters and the numerous different receptor proteins—many different kinds for any one neurotransmitter—are another facet of the vast complexity of the brain.

By psychoactive drug, I mean chemicals that influence the functioning of the human brain in such a way as to have effects on the psyche or mind. It is currently known that psychoactive drugs exert their influence by interacting with the signal transmission process at synapses. There are various ways this can happen. A drug can mimic the effect of a neurotransmitter because sufficient similarity in molecular structure allows it to bind to and activate a receptor in the same way as the natural neurotransmitter. Such a drug is called a receptor agonist. It could also bind to a receptor and block the receptor, preventing its activation by endogenous neurotransmitter. Such a drug is called a receptor antagonist. Some drugs enhance neurotransmitter release or leakage from the presynaptic neuron. Other drugs block release of neurotransmitter from the presynaptic neuron. Still others block reuptake of neurotransmitter from the synapse, interfere with the synthesis or degradation of neurotransmitter, or interact with the chemistry of the brain in ways that have not been elucidated.

Psychoactive drugs include the following chemicals:

- Commonly used nonmedical substances, such as caffeine, alcohol, and nicotine
- Medically used substances, such as morphine and other opioids, amphetamine, and benzodiazepines
- Mind-altering substances that have been excluded from accepted medical use and are declared illegal, such as heroin, marijuana, and psychedelics such as lysergic acid diethylamide (LSD), mescaline, and psilocybin
- Medications used to treat mental-health conditions, such as antidepressants, antipsychotics, and other mood stabilizers

Some of the previously mentioned drugs are frequently used for their pleasure-producing or recreational effects (alcohol, caffeine, etc.), some are used medicinally (antidepressants, antipsychotics, etc.), some are used for both (amphetamine, opioids, benzodiazepines, marijuana, etc.), and some are more prone to fostering problematic relationships such as addiction (nicotine, alcohol, benzodiazepines, etc.).

Historically, psychoactive drugs were all natural products, ingested in the form of plants or fungi or extracts of these organisms. The original meaning of the word "drug" was "dried plant." Beginning in the 1800s, chemists isolated and identified a number of molecules from plants that had psychoactive effects. Morphine, caffeine, nicotine, cocaine, and mescaline were all identified as psychoactive compounds from plants by 19th-century German chemists. In the past century, pharmaceutical chemists synthesized new molecules that have psychoactive effects, some of which have been marketed as drugs for the treatment of various conditions.

In the United States, many psychoactive drugs are classified according to the Controlled Substances Act, a federal law first instituted in 1970 and copied in some form by all of the states. The Controlled Substances Act classifies psychoactive drugs that have been declared to have the potential for abuse (problematic use leading to possible addiction) into five categories or schedules. Schedules II–V are legally available for medical use with appropriate licenses and physicians' prescriptions. Examples are amphetamine, morphine, and the benzodiazepines. Schedule I drugs have been declared categorically illegal and are not available for legal medical use. Examples are marijuana and the psychedelics. Alcohol and nicotine (tobacco), because of the powerful economic interests vested in keeping them easily available, are not regulated by the Controlled Substances Act.

This article only briefly discusses the neurochemical effects of various drugs, but it does contain the essence of what is currently known. As knowledge about the brain increases, we will find that many of the actions of psychoactive drugs are far more complex than is currently understood.

II. SEDATIVE–HYPNOTICS AND RELATED DRUGS

Sedative–hypnotics produce relaxation and often euphoria in low doses, drunkenness and sleep in higher doses, and possible coma and death in overdose. Alcohol (ethyl alcohol or ethanol) is the most widely used sedative–hypnotic. Ethyl alcohol is a natural product produced by yeast via the fermentative metabolism of carbohydrate. Synthetic pharmaceutical sedative–hypnotics include the benzodiazepines [diazepam (Valium), alprazolam (Xanax), lorazopam

(Ativan), temazapam (Restoril), etc.], barbiturates (phenobarbital, secobarbital, amobarbital, etc.), and other substances such as chloral hydrate, meprobamate (Miltown), and methaqualone (Quaalude). Pharmaceutical sedative–hypnotics have been prescribed medically as treatments for anxiety (anxiolytic properties) and as aids to sleep.

General anesthetics are powerful sedative–hypnotics used to induce loss of conscious awareness and insensitivity to pain during surgical procedures. They are often volatile organic compounds inhaled as part of a gas mixture during surgery. Historically, diethyl ether and chloroform were used in this way. Contemporary examples include halogenated compounds such as desflurane, enflurane, halothane, isoflurane, and sevoflurane.

Gamma-hydroxybutyric acid (GHB) is a sedative–hypnotic whose chemical structure is similar to that of the neurotransmitters GABA and glutamate. GHB was recently declared a Schedule I drug in the United States because of its use in recreational settings. Kava kava (*Piper methysticum*, native to South Pacific islands) and valerian (*Valeriana officinalis*) are examples of plants with sedative–hypnotic properties.

Sedative–hypnotics are believed to exert their effect on the brain by interacting with receptors for the neurotransmitter GABA. Their effect at these receptors enhances the action of GABA as an inhibitory neurotransmitter and results in a depression of brain activity. Some (e.g., alcohol) may also have actions at glutamate receptors, reducing the action of glutamate, which is the brain's major excitatory neurotransmitter. In addition, GHB may have direct inhibitory neurotransmitter action in the brain, with its own distinct receptors.

III. CAFFEINE

Caffeine (1,3,7-trimethylxanthine) is by far the most widely used psychoactive drug in the world. It is found in a number of plant sources, the most well-known being coffee (*Coffea arabica*, native to Africa), tea (*Camellia sinensis*, native to China), and cacao (*Theobroma cacao*, native to South and Central America), from which chocolate is made. Other caffeine-containing plants include kola (*Cola acuminata*, native to Africa), guarana (*Paullinia cupana*, native to South America), and yerba mate (*Ilex paraguarienis*, native to South America). Closely related to caffeine are two other botanical methylated xanthines with similar psychoactive effects: theophylline (1,3-dimethylxanthine), found in tea, and theobromine (3,7-dimethylxanthine), from cacao. Caffeine is believed to act by antagonizing (blocking) receptors for the inhibitory neurotransmitter adenosine, thereby producing a stimulating effect on brain activity.

IV. NICOTINE

Nicotine comes from the tobacco plant—*Nicotiana tabacum*, *Nicotiana rustica*, and related species—native to the Americas. It is generally taken into the body by smoking the dried leaves of the tobacco plant. It may also be absorbed through the oral cavity or nasal mucosa if a snuff preparation or chewing tobacco are used. It produces complex effects on the brain that result in relaxation, stimulation, and focused attention. The major known neurochemical effect of nicotine is to bind as an agonist to the nicotinic-type receptor for the neurotransmitter acetylcholine.

V. COCAINE

Cocaine comes from the plant *Erythroxylum coca* and related species, native to South America. Cocaine blocks reuptake transporter proteins for the neurotransmitters dopamine, norepinephrine, and, to a lesser extent, serotonin. In the brain, this results in the activation of circuits that increase alertness and wakefulness, produce euphoria, and decrease appetite. In the peripheral nervous system, cocaine activates the sympathetic branch of the autonomic nervous system, with resultant increased heart rate, blood pressure, pupil size, and bronchial airway size.

VI. AMPHETAMINE AND RELATED MOLECULES

Amphetamine and its close chemical relative methamphetamine are synthetic substances that cause the neurotransmitters dopamine, norepinephrine, and serotonin to leak out of the presynaptic nerve terminal into the synapse. This results in activation of the same brain and peripheral nervous system circuits as with cocaine and substantially similar behavioral effects. There are a number of pharmaceutically manufactured substances that can be considered molecular relatives of amphetamine and that work in similar ways. They have been prescribed as central nervous system (CNS) stimulants for the treatment of narcolepsy, appetite suppressants, and as treatment for attention deficit hyperactivity disorder. Ephedrine (from plants of the

genus *Ephedra*, found worldwide), cathinone (from the plant *Catha edulis*, native to northeast Africa and Yemen), and the synthetic compound phenylpropanolamine are examples of other related molecules with both central and sympathetic nervous system stimulant effects.

VII. OPIOIDS

Opioids (or opiates) are a group of naturally occurring and synthetic chemicals that act on the nervous system in a manner such as opium, the gummy secretion from the unripe seedpod of the opium poppy, *Papaver somniferum*, native to the Middle East and north Africa, now grown throughout the world. Natural opium contains morphine and codeine. In 1803, morphine became the first physiologically active compound to be isolated and identified from a plant. Opioids act on opioid receptors in the brain to reduce the perception of pain (analgesia), suppress cough, constrict pupils, and slow respiration. They may also produce relaxation, euphoria, and a dreamy altered state of consciousness. Medicinally, opioids are invaluable as analgesics and cough suppressants. Acting at opioid receptors in the digestive system, opioids slow intestinal motility and thus are valuable treatments for diarrhea.

A number of opioids either are chemical derivatives of morphine, codeine, or other related molecules in opium or are synthesized from nonopioid precursors. Heroin (diacetylmorphine) is a simple chemical derivative of morphine that allows it to cross the blood–brain barrier into the brain two or three times more efficiently than morphine. Hydromorphone (Dilaudid), oxycodone (Percodan and OxyContin), and hydrocodone (Vicodin) are all derived from opium precursors. Methadone, fentanyl, meperidine (Demerol), and propoxyphene (Darvon) are examples of synthetic opioids.

The endogenous neurotransmitters acting at opioid receptors are a class of small peptide molecules collectively called opioid peptides or endorphins (from "endogenous morphine"). These range in size from the 5-amino acid-long enkephalins to the 31-amino acid-long beta-endorphin.

VIII. CANNABINOIDS

The cannabinoids are a class of chemical substances found in the *Cannabis* plant, the source of marijuana (dried buds and leaves) and hashish (concentrated resinous exudate from the buds and leaves). The primary psychoactive cannabinoid in cannabis is Δ-9-tetrahydrocannibinol, often referred to simply as THC. The cannabis plant is one of the most ancient human cultivars, with medicinal use documented in Chinese medical texts from more than 4000 years ago. Some botanists recognize distinct species of *Cannabis sativa*, *Cannabis indica*, and *Cannabis rudeleris*, whereas others believe that cannabis has been cultivated by humans for so long that no distinct wild-type species remain.

The psychoactive effects of cannabis may include relaxation, sedation, intensification of thoughts and feelings, alterations of perception, and increased appetite. Medicinally, cannabis has been used as an analgesic, an antiinflammatory, a treatment for migraine headache, an antiemetic (to decrease nausea), an anticonvulsant, a muscle relaxant, a treatment for glaucoma, and an appetite stimulant. It was widely used as an accepted pharmaceutical prior to its being declared illegal by the United States government in 1937. Recent years have seen a resurgence of interest in the medicinal properties of cannabis and the discussion and attempted implementation of new laws to allow legal medical research and use to again take place.

For many years, the neurochemical mechanism of action of THC remained obscure. The past decade saw the discovery of specialized receptors responding to THC (cannabinoid receptors). Cannabinoid receptors are widespread in the human brain as well as elsewhere in the body, such as components of the immune and endocrine systems. At least two endogenous compounds have been implicated as neurotransmitters for the cannabinoid system—anandamide and 2-arachidonyl glycerol, both of which are long-chain, polyunsaturated compounds derived from lipids. The function of the cannabinoid neural circuitry remains unknown, although it is likely to be involved in appetite, memory, and the perception of pain.

IX. PSYCHEDELICS OR HALLUCINOGENS

The psychedelics are among the most interesting and ill understood of the psychoactive drugs. They produce a variety of complex effects on the brain and mind, including intensification of thoughts and feelings, alterations of sensory perception, and loosening of psychological defenses. Reported experiences range from euphoria, therapeutic insight, and visions of god, to anxiety and panic. The effects of the psychedelic drugs, more so than with other drugs, may be

significantly influenced by the mental set of the user (expectations, prior experience, mood, etc.) and the physical setting of use (alone, social, therapeutic context with a healer, spiritual ritual, etc.).

In 20th-century medicine, many of these compounds were first called psychotomimetics because it was believed that their effects were like the symptoms of psychosis. This term was eventually discarded in favor of hallucinogen because they may produce alterations of perception that include hallucinations. In 1956, the psychiatrist Humphry Osmond, in correspondence with the author Aldous Huxley, proposed the word "psychedelic" (from the Greek "mind manifesting") to describe the unique effects of these compounds on consciousness. In their plant forms, psychedelics have been used by cultures throughout the world for thousands of years to facilitate spiritual states of consciousness. This has led to the proposal of the word "entheogen" ("generating god within") to emphasize their ritual spiritual use.

Although it has not been discovered as a natural product and its profound psychoactive properties were not known until it was synthesized and tested by Albert Hofmann in 1943, LSD is perhaps the most widely known psychedelic chemical. It is one of the most potent psychoactive compounds known, active in quantities of a few micrograms. Psychedelics identified from plants and fungi include lysergic acid amide from morning glories, psilocybin and psilocin from a variety of mushrooms (many of the genus *Psilocybe*), dimethyltryptamine from dozens of species of plants throughout the world, and mescaline from several species of cacti, one of which is peyote (*Lophophora williamsii*). Numerous chemical relatives of the compounds mentioned previously have been synthesized, tested, and found to have psychedelic effects in humans.

Prior to their being declared illegal in the United States, these compounds were researched and utilized for their psychotherapeutic properties. Research at the time found them to be of value in the treatment of alcoholism and in therapeutic work related to dying in terminally ill patients. There is ample reason to believe that when used with therapeutic intent in carefully controlled settings, they can be of extraordinary value. Such use may also involve some exclusion of individuals having preexisting psychopathology because their powerful effects on the brain and mind may exacerbate preexisting symptoms or tendencies toward psychopathology.

Neurochemically, psychedelics have been found to bind as agonists to various serotonin receptor subtypes, especially type-2 serotonin receptors. Psychedelics also bind to other neurotransmitter receptors, including those for dopamine and norepinephrine. The connection between their neurochemical effects and their profound effects on mental function remains obscure.

X. METHYLENEDIOXYMETHAMPHETAMINE

Methylenedioxymethamphetamine (MDMA), popularly known by the street name "ecstasy," and its close chemical relative methylenedioxyamphetamine (MDA) are often considered together with the psychedelic drugs. Although MDMA does have some of the mind-manifesting characteristics of the psychedelics, its overall properties are distinct enough to warrant a separate categorization. Chemically related to methamphetamine, MDMA produces CNS stimulation, euphoria, and sympathetic nervous system stimulation. The methylenedioxy moiety confers upon the molecule additional psychoactivity that is psychedelic-like, characterized by intensification of thoughts and feelings and enhanced feelings of connection with others. For many individuals, there is a reduction in anxiety and an enhanced ability to verbalize feelings. These qualities led to the use of MDMA as an adjunct to psychotherapy prior to its being declared an illegal Schedule I substance in 1985, an act that followed media attention focusing on its use as a recreational drug in the dance scene. Although illegal, MDMA continues to be used for both its psychotherapeutic and recreational properties by some individuals.

Although MDMA has been found to interact with a number of CNS neurochemical systems, the primary neurochemical effect of MDMA appears to be the release of serotonin and dopamine from nerve terminals into the synapse. Release is not via the usual process of storage vesicle fusion with the cell membrane, as occurs with a nerve impulse, but leakage of neurotransmitter out of the cell via the reuptake transporter.

Animal studies have indicated that MDMA may result in oxidative damage to serotonin nerve terminals. The implications of these findings for human use remain to be determined.

XI. NITROUS OXIDE

Nitrous oxide (N_2O) is a gas (sometimes called laughing gas) widely used for its anesthetic and analgesic properties. It is used by dentists as a sole anesthetic and

is a component of the inhalation anesthesia (together with potent volatile sedative–hypnotic general anesthetics) for major surgeries. In addition, in subanesthetic doses it can produce a powerful psychedelic-like altered state of consciousness. Its various psychoactive properties were documented in 1800 by the British chemist Humphrey Davy, marking the first detailed study of the effects of a defined chemical substance on the human mind. Although it has been found to interact as an antagonist at the NMDA-type of glutamate receptor, the mechanism of its psychoactive effects remains obscure.

XII. DISSOCIATIVE ANESTHETICS

This class of psychoactive drugs includes ketamine and PCP (phenyl cyclohexyl piperidine or phencyclidine). They are synthetic compounds introduced into medicine to produce an anesthetic loss of sensation without depressing respiration and cardiovascular function as do the general anesthetics. Although PCP is no longer used medically, ketamine is used for both human and veterinary surgical procedures. At subanesthetic doses in humans, ketamine produces an altered state of consciousness that has some psychedelic-like characteristics and may be accompanied by a loss of body sensation. The primary neurochemical effect is as an antagonist at the NMDA-type glutamate receptor.

XIII. ANTICHOLINERGICS

These substances act as antagonists at muscarinic-type receptors for the neurotransmitter acetylcholine. Such receptors are found in the brain and in the parasympathetic branch of the autonomic nervous system. Psychoactive effects can range from feelings of disorientation to dream-like alterations of consciousness, powerful hallucinations, and delerium. Effects on the autonomic nervous system produce a cluster of symptoms that include dry mouth, difficulty urinating, constipation, blurry vision, and orthostatic hypotension. Hyoscyamine/atropine and scopolamine are anticholinergics found in a number of species of plants from the family Solanaceae. These include *Atropa belladonna* (deadly nightshade), *Hyoscyamus niger* (henbane), *Mandragora officinarum* (mandrake), *Brugmansia*, *Datura*, and *Brunfelsia*.

Atropine, scopolamine, or tincture of belladonna may be used to treat intestinal motility problems such as diarrhea and irritable bowel syndrome. Several synthetic anticholinergic drugs are used as adjunctive treatments for Parkinson's disease. Several of the phenothiazine antipsychotics and tricyclic antidepressants described later have anticholinergic effects in the brain and autonomic nervous system.

XIV. ANTIPSYCHOTICS

In Indian ayurvedic medicine, from ancient times to the present, extracts of the snakeroot plant, *Rauwolfia serpentina*, have been used to treat symptoms of psychosis. In the 20th century, the chemical reserpine was isolated and identified from *R. serpentina* and found to cause decreases in the activity of monoaminergic neurons using the neurotransmitters dopamine, norepinephrine, and serotonin.

The first pharmaceutical antipsychotics were the phenothiazines such as chlorpromazine (Thorazine), introduced in the 1950s. Some nonphenothiazine antipsychotics introduced to the market in the 1960s and 1970s include haloperidol (Haldol) and thiothixene (Navane).

The primary neurochemical effect of these compounds is to block various types of dopamine receptors. This is believed to lead to a reduction of psychotic symptoms and also to the production of various motor side effects, such as parkinsonian-like symptoms and movement dyskinesias.

Recently, several antipsychotic medications have been introduced that have fewer motor side effects, including clozaril (Clozapine), risperidone (Risperdal), olanzepine (Zyprexa), and ziprasidone (Geodon). In addition to being antagonists at dopamine receptors, these drugs are also antagonists at type-2 serotonin receptors. The serotonin receptor antagonism is believed to compensate for some of the dopamine-blocking effects in the motor circuitry of the brain, thereby reducing problematic movement symptoms.

XV. ANTIDEPRESSANTS

Extracts of the plant Saint John's wort, *Hypericum perforatum*, have been used for centuries in Europe for their antidepressant effects. This plant also facilitates wound healing when preparations are used topically. Its healing properties were mentioned in the ancient medical texts of Hippocrates, Pliny, and Galen.

The first modern pharmaceuticals developed and marketed specifically for their antidepressant effects

were the monoamine oxidase inhibitors (MAOIs), discovered in the 1950s. Examples currently on the market are isocarboxazid (Marplan), phenelzine (Nardil), and tranylcypromine (Parnate). Via inhibition of the enzyme MAO, these compounds may produce enhanced neural activity in circuits utilizing the neurotransmitters serotonin, norepinephrine, and dopamine.

These were followed in the 1960s by the tricyclic antidepressants (TCAs), such as imipramine (Tofranil), amitriptyline (Elavil), desipramine (Norpramine), nortriptyline (Pamelor), doxepin (Sinequan), and clomipramine (Anafranil). TCAs have been found to inhibit monoamine reuptake transporters, primarily for norepinephrine and serotonin.

Newer generation antidepressants include the serotonin selective reuptake inhibitors, represented by fluoxetine (Prozac), paroxetine (Paxil), sertraline (Zoloft), fluvoxamine (Luvox), and citalopram (Celexa), and other compounds such as trazodone (Desyrel), nefazodone (Serzone), bupropion (Wellbutrin and Zyban), venlafaxine (Effexor), and mirtazapine (Remeron). These latter compounds interact in various ways with monoamine neurotransmitter receptors or reuptake transporters.

The prevailing hypothesis regarding antidepressant mechanism is that some sort of change in serotonin and/or norepinephrine synaptic chemistry underlies their clinical action, but exactly what sort of change remains obscure.

XVI. OTHER MOOD STABILIZERS

These compounds are used to treat bipolar or manic–depressive disorder. The most well-known is lithium. The therapeutic effects of lithium in the treatment of manic–depression were discovered in 1949 and lithium has been widely used in psychiatry for more than 40 years. Its mechanism of action remains largely obscure. It is known to influence a variety of different neurotransmitter systems. The most robust neurochemical effect of lithium thus far elucidated is its interference with G-protein-coupled pathways using phosphatidylinositol as an intracellular messenger. Several chemicals, first discovered for their antiseizure properties and used to treat seizure disorders, are also used to stabilize mood in manic–depression disorder. These include carbamazepine (Tegratol), valproic acid (Depakote), and gabapentin (Neurotin). This connection between unstable mood and seizures suggests that there may be so-called kindling processes involved in both, wherein neural pathways become triggered into some sort of chain reaction of overactivity.

XVII. DRUGS AND THE BRAIN

The functioning of the brain depends at least in part on its chemistry and psychoactive drugs work by influencing the chemistry of the brain. This may give rise to short-term behavioral effects, such as relaxation, alertness, euphoria, analgesia, and sleep. Repeated perturbation of the brain's chemistry may bring about changes in the circuitry of the brain that may underlie lasting changes in behavior: improved mood from antidepressants, decreased psychosis from antipsychotics, symptoms of dependence and addiction from a variety of drugs, movement disorders from antipsychotics, and so forth. Changes in the brain's circuitry presumably reflect the strengthening and weakening of synapses through mechanisms such as changes in the number of receptors, changes in the amount of neurotransmitter released, and changes in intracellular biochemical pathways that are activated by membrane receptors. Growth of new dendrites to form new synapses or deterioration of existing synapses may also be involved.

XVIII. WITHDRAWAL, DEPENDENCE, AND ADDICTION

Changes in brain physiology resulting from repeated exposure to psychoactive drugs may result in withdrawal symptoms when use of the drug is decreased or stopped. It is likely that extended use of *any* psychoactive drug will bring about homeostatic changes in brain physiology that will result in symptoms of withdrawal if use of the drug is stopped. In some cases the withdrawal symptoms will be clinically significant, and in some cases they may not be. For example, repeated use of sedative–hypnotics may produce rebound withdrawal symptoms of arousal, anxiety, and insomnia and, in more severe cases, hallucinations, delerium, and seizures. Situations of clinically significant withdrawal are sometimes referred to as marking physical dependence.

Some psychoactive drugs, especially those that produce an experience of reduced anxiety, pleasure, or euphoria, appear to influence neural circuits in the limbic system of the brain. These so-called reward circuits are likely important in the reinforcement of behaviors that are important for survival, such as

eating, sexual activity, and obtaining comfortable shelter. Euphorogenic psychoactive drugs may chemically hijack these circuits and produce anomolously intense stimulation. The resulting homeostatic changes in the circuits might then produce symptoms of withdrawal that include cravings (intense desires) for the drug, dysphoria, anxiety, and a drivenness or compulsiveness directed toward drug use. Clinically significant compulsive drug use is termed addiction.

The reward circuits of the limbic system connect the ventral tegmentum with the nucleus acumbens. The neurotransmitter in these circuits is dopamine. Various drugs known to be euphorogenic and having the potential to produce addiction have been shown to increase the release and activity of dopamine in these circuits. Although different psychoactive drugs (sedatives, stimulants, opioids, etc.) have a variety of different effects on the brain and behavior, it is possible that their actions on the dopaminergic reward circuits may be part of a common neurochemical mechanism underlying addiction.

Whether an individual develops an addictive relationship with a drug is in part dependent on the drug's pharmacology and direct effects on brain neurochemistry. These factors are modulated by genetic proneness for addiction, as well as psychological and sociocultural factors, such as coping strategies for stress and extent of use by others in one's surroundings.

XIX. DRUGS AND THE MIND

I define mind as the collection of mental states, including thoughts, feelings, and perceptions, experienced by the organism. The current working hypothesis in neuroscience is that the mind is associated with the brain, and that neural processes taking place in the brain somehow generate the qualities of mind. I define consciousness as an aspect of mind that involves awareness of these mental states. The neurobiology of mind, of course, is *the* great unsolved problem of neuroscience. We have little clue as to the chemistry and physiology underlying mental states. Drugs that affect consciousness and other qualities of mind can and will be powerful probes to the physiology of mind. To this end, the complex effects of the psychedelics on consciousness may be especially useful probes, but only after our scientific and legal enterprises have risen to embrace the task of such research.

The great scholar William James, whose thinking on the nature of mind was influenced by his self-experiments with nitrous oxide and with cannabis, said a century ago, in words that are every bit as relevant to this day,

> *One conclusion was forced upon my mind at the time, and my impression of its truth has since remained unshaken. It is that our normal waking consciousness, rational consciousness as we call it, is but one special type of consciousness, while all about it, parted from it by the filmiest of screens, there lie potential forms of consciousness entirely different. We may go through life without suspecting their existence; but apply the requisite stimulus, and at a touch they are there in all their completeness, definite types of mentality which probably somewhere have their application and adaptation. No account of the universe in its totality can be final which leaves these other forms of consciousness quite disregarded. How to regard them is the question.…*

See Also the Following Articles

ALCOHOL DAMAGE TO THE BRAIN • CONSCIOUSNESS • DEPRESSION • GABA • MOOD DISORDERS • NEUROPHARMACOLOGY • NEUROTRANSMITTERS • OPIATES

Suggested Reading

Feldman, R. S., Meyer, J. S., and Quenzer, L. F. (1997). *Principles of Neuropsychopharmacology*. Sinauer, New York.

Grob, C. S. (2000). Deconstructing ecstasy: The politics of MDMA research. *Addiction Res.* **8**(6), 549–588.

Grof, S. (2001). *LSD Psychotherapy*. Multidisciplinary Association for Psychedelic Studies, Sarasota, FL. [Original work published 1980]

Hardman, J. G., *et al.* (2001). *Goodman & Gilman's The Pharmacological Basis of Therapeutics*, 10th ed. McGraw-Hill, New York.

Hofmann, A. (1983). *LSD, My Problem Child: Reflections on Sacred Drugs, Mysticism, and Science*. Tarcher, Los Angeles. [Original work published 1979]

Institute of Medicine (1999). *Marijuana and Medicine*. Natl. Acad. Press, Washington, DC.

James, W. (1994). *The Varieties of Religious Experience: Being the Gifford Lectures on Natural Philosophy and Religion Delivered at the University of Edinburgh, Scotland, in 1901–1902*. Modern Library, New York.

Ott, J. (1993). *Pharmacotheon: Entheogenic Drugs, Their Plant Sources and History*. Natural Products Company, Kennewick, WA.

Perrine, D. M. (1996). *The Chemistry of Mind-Altering Drugs: History, Pharmacology, and Cultural Context*. Am. Chem. Soc., Washington, DC.

Psychoneuroendocrinology

SERGE CAMPEAU
University of Colorado

GLOSSARY

adrenal glands Situated just above the kidneys, each gland consists of an inner adrenal medulla, which secretes the catecholamines epinephrine, norepinephrine, and dopamine, and an outer adrenal cortex, which synthesizes and secretes steroid hormones.

circadian rhythm Diurnal (daily) cycles of body functions.

gonads Endocrine structures (testes in males, ovaries in females) with dual functions: the production of germ cells (gametogenesis) and the synthesis and secretion of sex hormones.

homeostasis Complement of various physiological arrangements that serve to restore the normal state, once it has been disturbed.

hormones Chemicals (amino acid derivatives or peptides) released by the endocrine glands into the general circulation.

hypothalamus Area composed of distinct collections of neurons involved in the regulation of basic (homeostatic) vital functions and located in the basomedial region of the brain.

negative feedback Deviations from a given normal set point that triggers opposite compensatory changes, which continues until the set point is again reached.

pituitary gland Located below the brain and suspended on a stalk connecting it to the hypothalamus, it is composed of two major regions: the posterior lobe, which resembles neural tissue, and the anterior lobe, which consists of specialized endocrine secretory cells.

portal blood system Circulatory arrangement that begins in a capillary bed and leads not to vessels that carry blood directly toward the heart, but rather to a second capillary bed.

preprohormones Initial gene products in polypeptide hormone synthesis.

pulsatile release The typical pattern of hormone release, generally lasting only a few minutes, which occurs in large pulses or surges several times a day.

steroids Family of hormone molecules derived from cholesterol, including gonadal steroids, glucocorticoids, and mineralocorticoids, which easily cross lipid membranes.

thyroid gland Gland at the base of the neck that helps maintain the level of metabolism at optimal conditions in all body tissues via the synthesis and secretion of thyroid hormones.

Psychoneuroendocrinology is a multidisciplinary field aimed at elucidating endocrine functions ultimately controlled by the brain and, in turn, how the hormonal products of the various organs influence the functioning of the brain, mood, and cognition. The main endocrine functions of the body include development and growth, reproduction, homeostasis, (temperature, fluids, and minerals), and survival (stress). Although endocrine dysfunctions can be linked to organic disorders, many endocrine disorders have their roots in the brain. Together with accumulating evidence indicating direct hormonal effects upon the brain and the realization that psychological status (the mind or psyche) can dramatically impact endocrine functions, the discipline of psychoneuroendocrinology emphasizes the importance of exploring the interrelationships between mind, brain, organs, and hormones. This article is aimed at describing the major constituents of the neuroendocrine systems and how the brain, moods, and cognition regulate and are regulated by hormones.

I. GENERAL PRINCIPLES OF PSYCHONEUROENDOCRINOLOGY

Although the systems controlling different endocrine endpoints are distinct, several parallels can be drawn in

their overall functioning. The action of hormones on the brain also shows similar mechanisms. The following discussion focuses on the similarities and parallels between the various psychoneuroendocrine systems.

A. The Hypothalamus

The human hypothalamus is very small, accounting for approximately 0.003% of the entire brain mass. However, it controls an array of vital functions without which complex organisms simply could not survive. It is perhaps no accident that the hypothalamus is one of the most deeply located cell-containing regions in the brain, residing in the basomedial part of the head and shielded by the rest of the brain (Fig. 1). It was more than 50 years ago that Geoffrey Harris of Oxford University proposed that *neurosecretory* cells of the hypothalamus release substances in the portal blood system, suggesting for the first time a plausible regulatory mechanism in the release of anterior pituitary hormones.

The hypothalamus is often divided into three major areas: an anterior, a middle, and a posterior region. The middle third of the hypothalamus contains the neuroendocrine components that control most of the posterior and anterior lobes of the pituitary gland. The important cell groups of the middle hypothalamus in

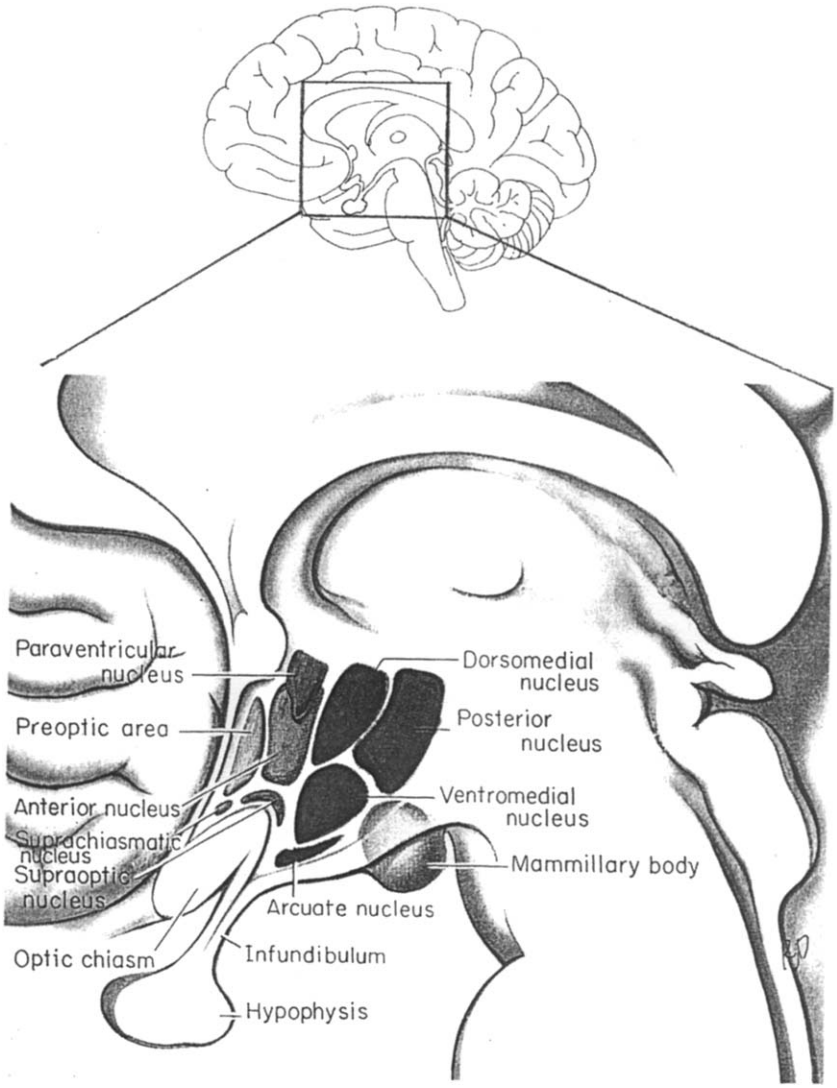


Figure 1 Schematic diagram of a midsagittal section of the human brain showing the location of the hypothalamus and its main subdivisions. Adapted from Carpenter, M. B. (1991). *Core Text of Neuroanatomy*, 4th ed., with permission of Williams & Wilkins.

regard to neuroendocrine control include the paraventricular and arcuate nuclei. The paraventricular nucleus is further broken down into the parvocellular cell group and the magnocellular cell group. In most instances, activity of the parvocellular releasing-factor neurons (also called *hypophysiotropic*) can be linked to the synthesis and secretion of specific pituitary peptides into the general circulation.

B. The Portal Blood System

Unlike the direct projections that exist from the hypothalamus to the *posterior lobe* of the pituitary via axons that join the *internal layer* of the *median eminence* and course through the *infundibular stalk*, the *anterior lobe* of the pituitary does not receive direct axonal projections from the hypothalamus, as shown in Fig. 2. Instead, the connections between the hypothalamus and the anterior lobe are entirely *neurohumoral*. In this process, peptides synthesized in neurons of the middle hypothalamus are carried via their axons to the *external zone* of the median eminence, where they are released into fenestrated capillaries that coalesce to form *portal vessels* that run along the infundibular stalk and terminate in vascular sinuses in the anterior lobe of the pituitary gland.

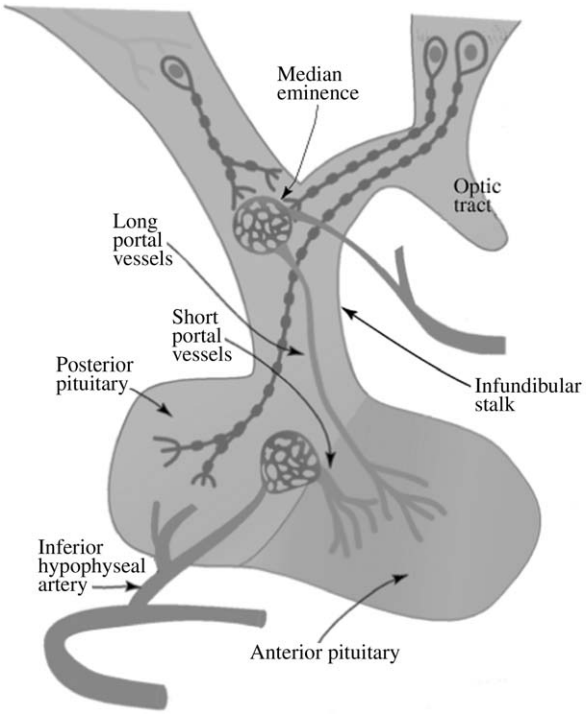


Figure 2 Schematic diagram of the arrangement of the portal blood system connecting the median eminence to the anterior lobe of the pituitary. [From Zigmond *et al.* (eds). (1999). *Fundamental Neuroscience*. Academic Press, San Diego].

C. The Pituitary Gland

Referred to as the "master gland," the pituitary gland, also known as the hypophysis, has been recognized for centuries as a vital organ for survival. The pituitary gland lies beneath the brain at the mesodiencephalic junction and is about the size of a garden pea in humans. Although the pituitary gland, and more specifically the anterior lobe of the gland, is responsible for the release of most hormones and tropic peptides acting on tissues and end organs, the anterior lobe is somewhat enslaved by the various tropic factors produced and released by hypophysiotropic neurons and by feedback effects of the released hormones, as shown in Fig. 3.

D. Hormones

Each neuroendocrine system described here releases chemicals (amino acid derivatives, proteins, or peptides), known as hormones, into the general circulation. Once released from their respective glands, most hormones are bound to specialized plasma proteins that help transport them and increase their half-life by protecting them from breakdown. However, most hormones have their highest biological activity in their free forms. Hormones travel throughout the body to act on body tissues wherever their receptors are located (targets). They have access to most body regions, although steroid hormones penetrate the central nervous system through the blood–brain barrier most readily.

E. Tropic Factors

Most of the hypothalamic peptides released into the external zone of the median eminence act upon the anterior lobe of the pituitary gland to *stimulate* the release of other peptides, so that they are known as *tropic factors* (also termed releasing factors). These should not be confused with *trophic factors*, which are generally involved with *growth*. Because many pituitary peptides released into the general circulation likewise stimulate the release of end-organ hormones, they are also termed tropic factors. As an historical aside, the term factor used to be dedicated for suspected tropic chemicals or peptides that had not yet been isolated. Upon isolation, however, these

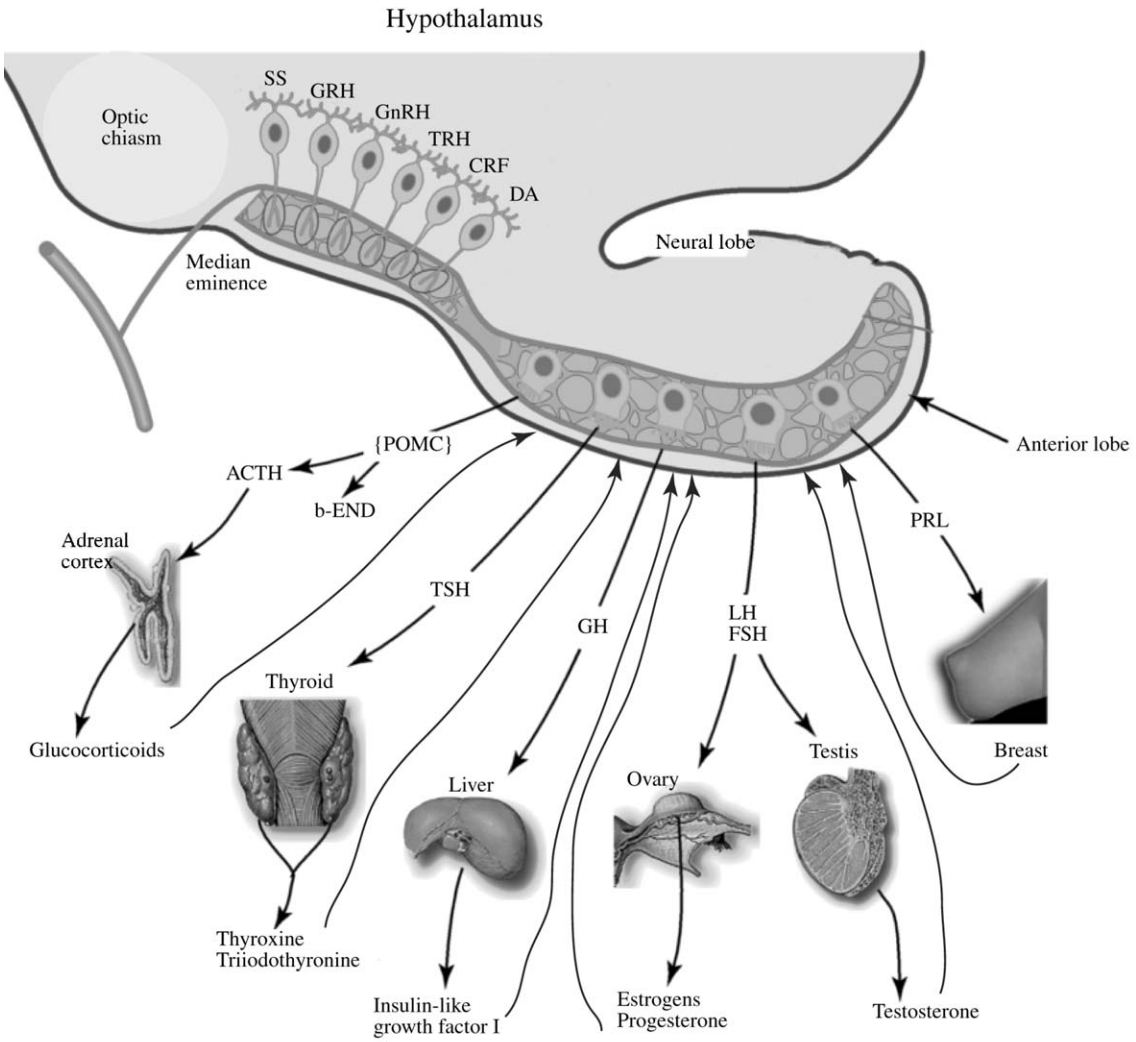


Figure 3 Schematic diagram of the main neuroendocrine systems. See text for abbreviations. [From Zigmond *et al.* (eds). (1999). *Fundamental Neuroscience*. Academic Press, San Diego].

tropic factors were renamed hormones. For instance, corticotropin-releasing factor has also been known as corticotropin-releasing hormone (CRH) since its isolation by Wyllie Vale and collaborators in 1981.

F. Negative Feedback

The circulating levels of several hormones in the body are set to optimal points. Deviations from these points are often detrimental to organisms. Thus, the regulation of circulating hormone levels is tightly controlled by several compensatory mechanisms. An important compensatory mechanism works through *negative feedback inhibition*, whereby hormone release acts at several levels, including the pituitary, hypothalamus, and even brain areas that project to the hypothalamus, to reduce its own further release by inhibiting the synthesis and release of the various hypothalamic and pituitary factors. This is a mechanism observed in all neuroendocrine systems. Negative feedback can take place in different time domains, as will be discussed for the hypothalamo–pituitary–adrenocortical axis.

G. Pulsatile Release and Circadian Rhythm

All of the neuroendocrine systems to be discussed show a typical pattern of activity, with hypothalamic factors and their associated pituitary factors released in only a

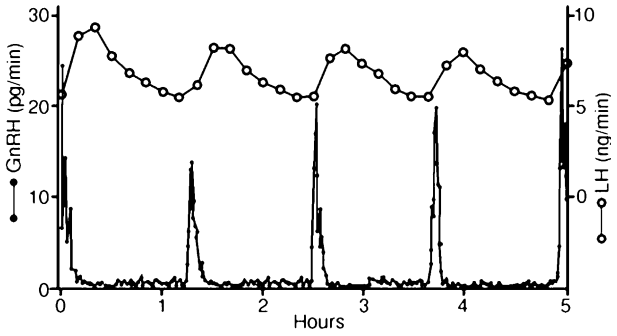


Figure 4 Example of pulsatile release of hypophysiotropic hormone in the hypothalamo–pituitary–gonadal axis. [From Zigmond *et al.* (eds). (1999). *Fundamental Neuroscience*. Academic Press, San Diego].

few minutes, which occurs in large pulses or surges several times a day, as is shown in Fig. 4 for the stimulation of luteinizing hormone by gonadotropin-releasing hormone. In most systems, the exact mechanisms resulting in such a pulsatile release are not known, but several factors have been demonstrated to influence it. Intrinsic characteristics of hypophysiotropic neurons as well as the influence of afferent innervation to these neurons are important factors in pulsatile release. The suprachiasmatic nucleus, for example, is an important biological clock shown to drive pulsatile release in some neuroendocrine systems. Feedback inhibition probably plays a role in the pulsatile release of some systems. In addition, meal time, exercise, and sleep time all appear to provide some influence on pulsatile release. Circadian rhythmicity refers to some daily recurring fluctuations in pulsatile release, which is weaker or stronger during some parts of the day.

II. THE HYPOTHALAMO–PITUITARY–ADRENOCORTICAL AXIS

The functions of the hypothalamo–pituitary–adrenocortical (HPA) axis principally include the restoration of homeostasis following internal or external challenge (stress). It also contributes to the normal regulation of energy metabolism throughout the body. The ultimate control of glucocorticoid levels in the bloodstream is intricately regulated positively and negatively by a series of steps initiated at the level of the hypothalamus. Glucocorticoid levels vary significantly over the course of 24 hr, with the highest levels observed upon waking up in the morning. The functioning of the HPA axis thus needs to be understood in the context of both its tonic activation and its responsive phasic tone.

A. Basic Constituents of the Hypothalamo–Pituitary–Adrenocortical Axis

In humans, cortisol is the major glucocorticoid produced and released by the adrenal glands, with corticosterone as a minor component. As will be discussed following a description of the regulation of the HPA axis, dysregulation of glucocorticoid levels is involved in several physical and psychiatric pathologies. Likewise, endogenous psychiatric disorders can also produce abnormal glucocorticoid secretion.

1. The Adrenal Glands and Cortisol

Glucocorticoids, which are part of a larger family of steroid hormones, are synthesized in the zona fasciculata cells of the cortical (external) layer of the adrenal glands. Cortisol, like the gonadal steroids, is synthesized from a series of enzymatic reactions using cholesterol as the precursor. In contrast to several other secretory cells, adrenocortical cells do not store large amounts of readily releasable glucocorticoids. Secretion is, therefore, tightly linked to *de novo* synthesis. The diffusible glucocorticoids, which easily cross lipid membranes, enter the highly vascularized blood supply of the adrenal glands, where 90–95% of glucocorticoids become bound to *corticosteroid-binding globulin* (CBG), also known as *transcortin*, a carrier protein found in plasma. The half-life of cortisol in blood is approximately 60–90 min. Because glucocorticoids are biologically active only in their free form, normally a small fraction of total glucocorticoids is available for bioactivity. However, CBG levels are significantly reduced under some conditions (stress), thus providing one mechanism regulating biologically active glucocorticoids without altering absolute levels.

Secretion and synthesis of glucocorticoids from adrenocortical cells are triggered by the activation of a specific subtype of melanocortin receptor, the adrenocorticotropin hormone (ACTH) receptor, in response to the anterior pituitary hormone ACTH. These receptors are part of the G-protein-coupled, seven-transmembrane-domain receptor superfamily, and, once activated, they increase the production of the second messenger cAMP via the enzyme adenylate cyclase. The adrenal ACTH receptor is positively regulated, that is, it can increase its number of receptors in response to sustained ACTH levels.

2. The Anterior Pituitary Gland and ACTH

The anterior lobe of the pituitary gland is responsible for the synthesis and release of ACTH. ACTH is a

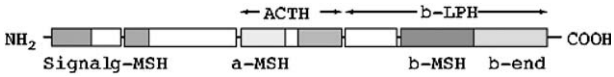


Figure 5 Diagram of one preprohormone, prepro-opiomelanocortin, which is cleaved and processed to give several different products. Abbreviations: ACTH, adrenocorticotropin hormone; end, endorphin; LPH, lipotropin hormone; MSH, melanocyte-stimulating hormone. [From Zigmond *et al.* (eds). (1999). *Fundamental Neuroscience*. Academic Press, San Diego].

39-amino-acid peptide secreted by specialized cells known as corticotropes, which account for approximately 10% of the total cell population of the anterior lobe. ACTH itself is derived from a longer peptide precursor, pro-opiomelanocortin (preproPOMC), as shown in Fig. 5. Regulation of the POMC gene has been studied to a great extent, as it was the first mammalian endocrine precursor molecule to be cloned.

3. The Hypothalamus and Corticotropin-Releasing Factors

The secretion of ACTH from corticotropes is signaled by corticotropin-releasing factors from the medial parvocellular region of the paraventricular hypothalamic nucleus. So far, the most potent endogenous factor involved in ACTH secretion is the 41-amino-acid peptide corticotropin-releasing hormone (CRH), which was discovered and sequenced by Wyllie Vale and colleagues some 20 years ago. Parvocellular neurons send their axons to the external zone of the median eminence where, upon activation, CRH is liberated into the portal blood system. Upon destruction of the PVN, several stressors loose the ability to produce ACTH secretion, suggesting that the CRH-containing neurons of the PVN are obligatory for this function. The secretagogue action of CRH has been shown to synergize with additional factors, such as the peptides arginine vasopressin (AVP), oxytocin, and others, some of which are also coexpressed and released from CRH-containing neurons of the medial parvocellular hypothalamic nucleus. Although the cDNA and genes for CRH and AVP peptides have been cloned for some time, it was only relatively recently that the cDNAs encoding the CRH and AVP receptors were cloned. These receptors are also members of the G-protein-coupled, seven-transmembrane-domain superfamily of receptors and are positively coupled either to the production of cyclic adenosine monophosphate (CRH) or to the phosphatidylinositol (AVP) second messenger intracellular pathways. Multiple receptor subtypes have been cloned for each, cut the CRH_1 and V_{1b} subtypes account for the majority of CRH and AVP receptors localized on anterior pituitary corticotropes. Pituitary CRH and AVP receptor peptide levels are rapidly controlled upon continued activation by CRH and AVP, with evidence of up- and down-regulation in response to different manipulations.

B. Control of the HPA Axis

1. Control of Circulating Glucocorticoids

Neurons of the parvocellular hypothalamic area are responsible for the production of corticotropic-releasing factors. CRH and AVP mRNAs and peptides from parvocellular neurons are significantly regulated in response to a variety of experimental manipulations comprising acute and repeated stress. An additional molecule discovered is the corticotropin-releasing hormone-binding protein (CRH-BP), which is found in blood, the pituitary gland, and various brain areas. This protein is localized in neurons and astrocytes, and it binds and inactivates endogenous CRH and other CRH-like peptides, such as the more recently discovered urocortin, providing an additional mechanism to control the levels of CRH reaching the anterior pituitary.

An important level of control over ACTH secretion is exerted by the glucocorticoids themselves. Because the constant exposure of organisms to high levels can produce deleterious effects, including immune suppression, cardiovascular dysfunction, and even death of certain neuronal populations, glucocorticoids exhibit several mechanisms to restrain their own release. It is generally recognized that glucocorticoids provide feedback inhibition at several levels, including the pituitary, paraventricular hypothalamus, and other brain areas such as the hippocampus (discussed in the following section). These inhibitory influences act in several time domains, including fast, intermediate, and slow (genomic) feedback. The genomic and, to some extent, intermediate feedback inhibition is provided by the action of glucocorticoids upon two well-characterized and specific receptors: the mineralocorticoid (MR) or type I receptor and the glucocorticoid (GR) or type II receptor. They are members of a superfamily of cytoplasmic receptors that act as ligand-regulated transacting factors, as depicted in Fig. 6. Upon activation, these receptors translocate to the nucleus where they may interact with other transacting factors to bind to specific recognition elements on the DNA, effecting changes in the transcriptional rates of several genes, including CRH, AVP, and POMC. However, glucocorticoids are also known to produce feedback

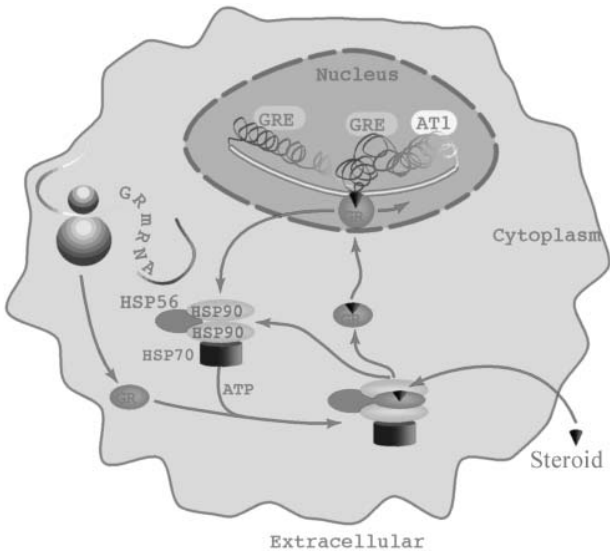


Figure 6 Schematic diagram of glucocorticoid receptor (GR) function. Once translated in the cytoplasm, GR is quickly stabilized by a complex including many heat-shock proteins (HSP). Circulating steroids, including cortisol, easily cross cell membranes to access the cytoplasm from the circulation and bind to GR. Activation of GR by steroid binding induces its separation from the HSP complex and its translocation to the nucleus, where it interacts with DNA (transacting factor) to facilitate or inhibit gene transcription. [From Zigmond *et al*. (eds). (1999). *Fundamental Neuroscience*. Academic Press, San Diego].

inhibition that is too rapid to rely on genomic regulation. This glucocorticoid rate-sensitive regulation has not been well-characterized, but it is likely exerted via either nongenomic effects of MR or GR receptor occupancy or via nonclassical glucocorticoid-sensitive receptors.

Hence, the HPA axis provides a cascade of processes exerting positive and negative control at several levels, ultimately participating in the regulation of energy and metabolic activity throughout the body. Even under conditions of homeostasis, glucocorticoid levels display a diurnal rhythm, with higher levels in the morning than in the evening for humans, anticipating increased bodily demands during the day. This rhythm appears to be driven by one of the brain's main biological clocks, the suprachiasmatic nucleus, as well as by the pattern of food intake. Activation of the HPA axis by events that are deemed stressful by organisms occurs in the context of this diurnal regulation and is mediated by several putative brain pathways with direct connections to hypophysiotropic neurons. The following section briefly describes these neuronal circuits, together with their suggested involvement in the regulation of the HPA axis.

2. Brain Circuits Involved in the Regulation of the HPA Axis

Several reviews have described the neuroanatomical afferents to hypothalamic paraventricular neurons that might be involved in the regulation of the HPA axis, as shown in Fig. 7. The available evidence indicates the existence of several distinct afferents that are suggested to regulate the activity of hypothalamic neurons. Some of these circuits likely interact with other hypothalamic–pituitary endocrine systems.

a. Brain Stem and Medullary Afferents Well-known projections to the hypophysiotropic CRH-containing neurons originate from medullary catecholaminergic neurons of the nucleus of the solitary tract (C2) and the ventrolateral (A1, C1) and dorsomedial (C3) medulla. Most of these nuclei are involved in the processing and transmission of visceral or interoceptive information. These catecholamine-containing neurons appear to play a crucial role in activating most types of hypophysiotropic neurons in response to immune activation (e.g., infection, interleukin-1β, and

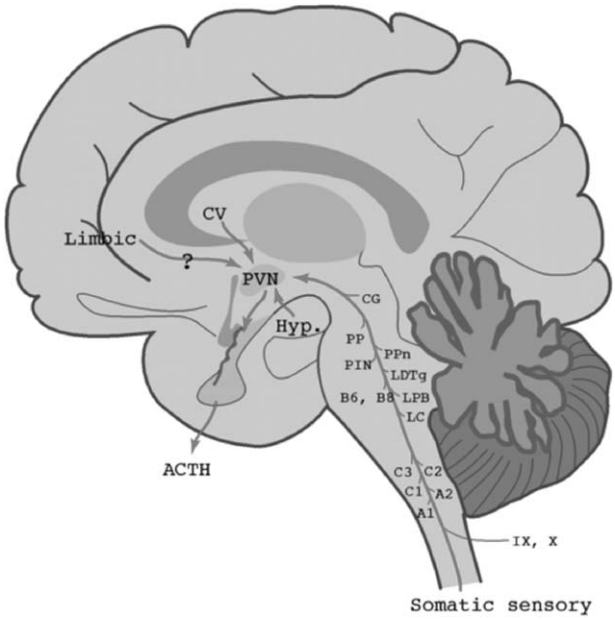


Figure 7 Schematic view of afferent innervation to the paraventricular nucleus (PVN) of the hypothalamus. Abbreviations: A1/2, C1/2/3, LC, catecholamine-containing cell groups of the brain stem; B6/8, serotonin-containing cell groups of the midbrain; CV, circumventricular organs; LPB, lateral parabrachial nucleus; LDTg, PPn, tegmental nuclei; PIN, posterior intralaminar thalamic nucleus; PP, peripeduncular thalamic nucleus; CG, central gray; Hyp, hypothalamic nuclei. [From Zigmond *et al*. (eds). (1999). *Fundamental Neuroscience*. Academic Press, San Diego].

endotoxins) and several types of visceral and somatosensory challenges.

Additional but more limited mesopontine afferents to the hypophysiotropic hypothalamus have been traced from serotonergic cell groups (B7, B8, and B9), some tegmental nuclei (peripeduncular, laterodorsal), the parabrachial nucleus, and some subregions of the central gray. Additional thalamic nuclei, including the peripeduncular nucleus, the posterior intralaminar nucleus, and the intergeniculate leaf, also provide afferents. The proposed function of several of these areas in sensory processing and transmission makes them likely candidates to provide relatively specific sensory information (somatosensory, nociceptive, auditory, and visual) to hypophysiotropic neurons.

b. Circumventricular Afferents Another set of hypophysiotropic afferents, originating from circumventricular organs, includes the subfornical organ, the vascular organ of the lamina terminalis, and the closely related median preoptic nucleus. These nuclei have been implicated in the activation of the HPA axis by blood-borne signals. Such signals include circulating indices of blood volume and ionic concentrations, which also impinge on the magnocellular neurons of the paraventricular and supraoptic nuclei to regulate the release of antidiuretic hormone (vasopressin) and oxytocin involved in water balance and blood pressure.

c. Hypothalamic Afferents An additional important source of afferents to hypophysiotropic neurons originates within the hypothalamus. Indeed, most hypothalamic nuclei contribute afferents to hypophysiotropic neurons with the exception of a few nuclei, namely, the medial and lateral mammillary and supraoptic nuclei. The principal hypothalamic areas providing relatively numerous afferents to the paraventricular nucleus include several preoptic areas, the anterior, lateral, and posterior hypothalamic areas, the dorsomedial and ventromedial nuclei, and the supramammillary nucleus. The suprachiasmatic nucleus, involved in circadian rhythm entrainment, appears to influence the HPA axis via projections to the dorsomedial nucleus. These nuclei are uniquely situated to significantly influence the activity of the neuroendocrine systems. The association of several of these hypothalamic nuclei with basic physiologic functions suggests that they might integrate the sum of all stimuli perceived by an organism at any one moment and then convey this integrated information to hypophysiotropic neurons to keep an optimum level of functioning throughout the neuroendocrine systems, including the HPA axis.

d. Forebrain Afferents Several forebrain areas appear to influence the activity of the HPA axis, including the prefrontal cortex, hippocampus, amygdala, and septum. Surprisingly, none of these areas unambiguously project to hypothalamic paraventricular neurons. Instead, several studies indicate that these areas might regulate the activity of the HPA axis via medial hypothalamic relays or, alternatively, via relays with neurons of the bed nucleus of the stria terminalis, the only additional forebrain structure known to project directly to hypophysiotropic neurons. The preceding forebrain structures are of particular interest because they are all associated with emotionality. Some of these areas might, therefore, be essential in the evaluative process involved in determining or perceiving the stressful character of a situation.

Neuroanatomically, therefore, hypophysiotropic CRH-containing neurons receive a rich set of neural inputs encompassing several categories of stimulus information. In addition, most of these regions, together with the rest of the brain, express one or both glucocorticoid receptors, making the brain an important target for circulating cortisol. The influence of the brain on the activity of the HPA axis, as well as that of cortisol upon brain functioning thus provides a rich set of interactions, the dysregulation of which can have dramatic consequences upon the physical and psychological well-being of individuals.

C. Dysregulation of Circulating Glucocorticoids (Cortisol)

Cushing's syndrome is perhaps the best-known class of diseases involving hypersecretion of glucocorticoids. In the early 1930s, Harvey Cushing described a list of symptoms that was closely associated with pituitary abnormalities. Glucocorticoid hypersecretion can be produced by adrenocortical or pituitary tumors, the latter associated with ACTH hypersecretion. Cases are also known in which increased hypothalamic CRH release produces ACTH hypersecretion. The symptoms of Cushing's disease, characterized by several physical and psychological changes, include thinning of the skin and hair, increased subcutaneous adipose (fat) deposits in the face (moon face), upper back and shoulders (buffalo hump), and abdomen (pendulous abdomen), appearance of reddish purple skin color

from skin stretching, hypertension, and osteoporosis. The psychiatric profile of Cushing's patients includes increased appetite, insomnia, euphoria, anxiety, panic, and unipolar or bipolar disorders. Psychosis (including schizophrenia) is rarely observed in Cushing's patients. Normalization of glucocorticoid levels in these patients greatly improves the physical and mental disorders.

On the other hand, a significant number of individuals seeking help who are diagnosed with major depression present signs of adrenal hypertrophy and increased basal levels (over 24 hr) of circulating cortisol. These endocrine abnormalities can be normalized in a high proportion of patients with a variety of antidepressant treatments (behavioral, pharmacologic, or electroconvulsive therapy), which concurrently improve their psychological status. The dexamethasone suppression test (DST) has been a useful tool to help determine HPA axis dysregulation. Dexamethasone is a glucocorticoid receptor agonist that, when injected intravenously, binds GR receptors at the level of the pituitary corticotropes and perhaps at the level of the hypothalamus and other brain regions and initiates feedback inhibition so that the circulating cortisol levels are suppressed hours after dexamethasone injection. Interestingly, a high proportion of individuals presenting symptoms of endogenous major depression do *not* show the normal suppression of cortisol by dexamethasone; they are said to *escape* the DST effect. Thus, abnormal GR receptor regulation, involved in feedback inhibition, appears to be linked with affective disorders. The most popular mechanisms hypothesized to mediate increased basal cortisol levels and reduced feedback inhibition have implicated an increased drive of central CRH-containing systems and dysregulation of brain serotonergic-containing systems. High levels of circulating cortisol levels have also been linked to neuronal cell death, especially in the hippocampus, a brain region with some of the highest levels of MR and GR receptors that is also thought to play a role in negative feedback of the HPA axis. Cortisol dysregulation thus provides clear example of the interactions between hormones and brain and vice versa.

A reduction in circulating cortisol levels is observed in patients with Addison's disease. This category of diseases includes several pathologies (adrenal autoimmune disease, tuberculosis, and adrenal cancer) that lead to adrenal atrophy and insufficiency. Total adrenal insufficiency is rapidly fatal, but with degenerative diseases, patients develop marked skin pigmentation (melanocyte-stimulating hormone activity of ACTH in blood), hypotension and reduced cardiac size. Even minor stresses can send Addison's patients into fatal shock. Adrenal insufficiency is also correlated with a slower electroencephalogram (EEG of the α rhythm), irritability, apprehension and mild anxiety, inability to concentrate, and increased gustatory and olfactory sensitivity. It is interesting that low circulating cortisol levels produce some of the same affective disorders observed with high circulating cortisol levels, but in milder forms. The mechanisms by which low cortisol levels affect mood and the chemical senses are ill-understood. Learning is also affected by circulating cortisol levels, and evidence of poor memory with either too much or too little cortisol has been documented.

III. THE HYPOTHALAMO–PITUITARY–THYROID AXIS

The hypothalamo–pituitary–thyroid (HPT) axis is the main regulatory system responsible for metabolism in all body tissues throughout life. Metabolism constitutes all biochemical processes that take place at the cellular level and includes functions such as synthesis of protein–lipid–carbohydrate–enzymes, incorporation of nutrients, and breakdown–synthesis of various molecules. Circulating thyroid hormones also stimulate oxygen consumption in most cells of the body. As with the HPA axis, the secretion of thyroid hormones in the bloodstream is intricately regulated at several levels, beginning at the hypothalamus. See Fig. 3 for a summary diagram of the HPT axis. The basic constitutive elements of the HPT axis and thyroid secretion are elaborated first. Central regulation of the HPT axis and dysregulation of thyroid hormones levels are then discussed.

A. Basic Constituents of the HPT Axis

In humans, thyroid hormones, composed mainly of L-thyroxine (denoted as T_4) and triiodothyronine (denoted as T_3), are released by the thyroid gland through a series of steps described next.

1. The Thyroid Gland and the Secretion of Thyroid Hormones

The thyroid gland, located at the base of the neck close to the trachea, synthesizes and stores T_3 and T_4 to large glycoproteins called *thyroglobulins* until they are secreted into the bloodstream. T_3 and T_4 are synthesized from the amino acid tyrosine and involve the incorporation of three (T_3) or four (T_4) iodine atoms,

making it the only molecule in the animal kingdom to contain iodine. Although the only difference between T_3 and T_4 is an additional iodine atom in T_4, this difference provides T_3 with much greater biological activity in all body tissues (3–5 times greater than T_4). In peripheral tissue, T_4 is converted to T_3 by the enzyme 5′-deiodinase type I, whereas T_4 is directly taken up by brain and pituitary cells and transformed into T_3 intracellularly by the type II 5′-deiodinase enzyme. In the bloodstream, thyroid hormones are bound to carrier proteins, including albumin, thyroxine-binding prealbumin (TBPA), and thyroxine-binding globulin (TBG), leaving a very low percentage of free hormones (about 0.2% T_3 and 0.02% T_4). T_4 is carried to the brain across the blood–brain barrier via a specific and saturable carrier molecule, transthyretin (TTR). The half-life of T_4 is approximately 1 week and that of T_3 is slightly shorter.

2. The Anterior Pituitary Gland and TSH

The thyroid-stimulating hormone (TSH), also known as thyrotropin, is responsible for the regulation of thyroid hormone secretion. TSH controls the uptake of iodine by the thyroid gland, regulates the rate of synthesis of thyroid hormones, and controls the storage and release of T_3–T_4 into the bloodstream. As with other neuroendocrine axes, the anterior pituitary is responsible for the production and release of TSH. Thyrotropes, the anterior pituitary cells producing TSH, constitute about 10% of the entire population of the anterior lobe. Through a hypothalamic factor, TSH is released into the bloodstream and eventually reaches the thyroid gland to regulate the synthesis and release of thyroid hormones. TSH release is pulsatile, and in humans the frequency of the pulses varies from 10 to 15 every 24 hr. TSH release also shows a circadian variation, with levels beginning to rise in late afternoon and peaks in the middle of the night (around 4:00 AM). The levels decline until noon and are quiescent until the next rise.

The bloodborne signal from the hypothalamus is mediated by specific membrane receptors on anterior pituitary thyrotropes, the *TRH receptors*. This receptor is a member of the seven-transmembrane, G-protein-coupled receptor superfamily. Binding of TRH to its receptors triggers the activation of the phosphatidylinositol and diacylglycerol second messenger pathways, which eventually leads to increased transcription of the TSHβ gene together with the α gene common to the glycoprotein hormones (TSH, luteinizing hormone, follicle-stimulating hormone, and human chorionic gonadotropin, discussed later in this article) that encode and produce TSH. As with the other glycoprotein hormones, TSH is a heterodimer polypeptide derived from two distinct genes and is heavily glycosylated, which is thought to extend its half-life.

3. The Hypothalamus and Thyrotropin-Releasing Hormone

The hypothalamic factor responsible for the release of TSH is thyrotropin-releasing hormone (TRH). This hormone was the first hypothalamic-releasing hormone isolated and purified from tons of pig and sheep brains in the late 1960s, after many years of hard work, by the research teams of Andrew Schally and Roger Guillemin, for which they would eventually share a Nobel prize. TRH is a small, 3-amino-acid peptide (histidine, proline, and glutamic acid) derived from a larger 242-amino-acid prohormone that contains 6 copies of the TRH sequence. TRH is synthesized in hypothalamic neurons located in the periventricular and parvocellular compartment of the paraventricular nucleus. In the medial parvocellular region, the TRH-containing neurons are observed medial to the CRH-containing neurons. These TRH-containing neurons send their axons to the external zone of the median eminence, where TRH is released into the portal blood circulation bathing the anterior pituitary. Other hypothalamic and extra-hypothalamic regions synthesize TRH, but these neurons do not directly, if at all, control thyroid hormone release.

B. Control of the HPT Axis

1. Control of Circulating Thyroid Hormones

As with the HPA axis and glucocorticoids, the HPT axis exerts rigorous control over thyroid hormone levels. Rising thyroid hormone levels, especially through T_3, act at the level of the pituitary thyrotropes and hypothalamic paraventricular neurons to inhibit further release of TSH and TRH, respectively, in a classic inhibitory feedback mechanism. This action is mediated by thyroid hormone receptors, which are part of a superfamily of cytoplasmic steroid hormone receptors including estrogen, progesterone, androgen, and vitamin D receptors. Several isoforms of the thyroid hormone receptors have been characterized, including Trα1, TRα2, TRβ1, and TRβ2, which act in a manner analogous to the MR and GR receptors as

transacting factors upon genomic response elements. TRα1 and TRβ2 are the most abundant isoforms, found in more than 80% of TRH-producing neurons of the paraventricular nucleus.

2. Brain Circuits Involved in the Regulation of the HPT Axis

Many of the same circuits described in the neural control of the HPA axis also appear to be involved in the control of the HPT axis. The proximity of the medial parvocellular hypophysiotropic neurons containing CRH and TRH can make it difficult, if not impossible, to ascribe distinct sets of afferents to the two endocrine axes. The brain stem catecholaminergic cell groups are known to innervate and play a part in the regulation of the HPT axis. Some of the rostral periventricular TRH-containing neurons also receive a relatively greater input of terminals from the noradrenaline-containing locus ceruleus neurons and from the serotonin-containing neurons of the midbrain raphe nuclei. The thyroid gland is also directly innervated by autonomic fibers (both sympathetic and parasympathetic) that control blood flow and, reportedly, the synthesis of thyroid hormones. These mechanisms can significantly regulate the levels of T_3–T_4 released in the bloodstream. Finally, as is the case with MR and GR receptors, the brain expresses some of the highest levels of the different thyroid hormone receptors. Dysregulation of thyroid hormone secretion can, thus, have dire consequences on physical and psychological status.

C. Dysregulation of Circulating Thyroid Hormones

Consideration of the effects of hypothyroidism or hyperthyroidism can emphasize the importance of normal levels of circulating thyroid hormones. Hypothyroidism is a condition resulting from inadequate thyroid hormone production. If this happens during fetal life, a condition known as *cretinism* develops, in which the afflicted individual remains of short stature, never reaches sexual maturity, and suffers severe mental deficiencies, hearing and speech defects, spastic gait, impaired voluntary motor control, and clonus. This condition can improve if thyroid hormones are provided during the first few months of life.

In adulthood, hypothyroidism is divided into clinical and subclinical types, with clinical hypothyroidism producing several physical and mental conditions. Low thyroid secretion rates produced by reduced thyroid hormone synthesis at the level of the thyroid gland or insufficient dietary iodine can lead to *goiter*, a very apparent enlargement of the thyroid gland produced by sustained TSH stimulation of thyroid growth combined with a lack of adequate thyroid hormone secretion (thus failing to inhibit pituitary TSH secretion). Other symptoms of clinical hypothyroidism include depression, low energy, appetite and sleep changes, poor concentration, memory impairments, apathy, muscle cramps, brittle and thinning hair, husky voice, and elevated levels of cholesterol and triglycerides. The similarity of these symptoms with those of endogenous major depression has prompted clinicians to test thyroid function in order to distinguish between the two conditions. Subclinical hypothyroidism, by definition, does not present the clinical type symptoms, although fatigue, hypothermia, dry skin, memory impairments, and higher scores on ratings of anxiety and depression are reported. The lifetime prevalence of depressive disorders in hypothyroid patients is 56% compared with 20% in euthyroid (high total, but normal free levels of T_4–T_3) individuals. Also, response to tricyclic antidepressant treatment in patients is lower in hypothyroid individuals, but can be improved by cotreatment with thyroid hormones. This may be particularly relevant in women, in whom the incidence of clinical and subclinical hypothyroidism is more than twice that for men, which is also mirrored in the incidence of depressive disorders. A higher incidence of thyroid antibodies is also observed in women, which may be attributed to the higher stimulation of immunologic functions by female gonadal hormones, leading to higher rates of autoimmune diseases. Exactly how hypothyroidism produces affective disorders, particularly major depression and rapid-cycling bipolar disorder, has not been clearly ascertained. Lower thyroid hormone levels have been suggested to reduce β-adrenergic receptor activity and central serotonin activity, effects often associated with endogenous major depression. The higher incidence of both conditions in women might also be linked to the homology and interactions between the estrogen and thyroid receptors, which are known to bind each others' hormones.

The reverse interaction between affective illnesses, particularly major depression, and thyroid hypofunction has also been documented. Patients presenting with major depression often display increased cerebrospinal fluid levels of TRH or a blunting of TSH

secretion in response to TRH challenge, suggesting down-regulation of pituitary TRH receptors in response to chronic high levels of hypothalamic TRH. This also agrees with some reports that depressed patients show blunted circulating plasma levels of TSH at night. There is still scant literature concerning the putative effects of endogenous affective illnesses on the HPT axis, but the preceding preliminary reports clearly indicate the need to better understand this brain–neuroendocrine interaction.

Hyperthyroidism involves conditions in which excess thyroid hormones are secreted. The highest percentage of hyperthyroidism in the United States is caused by Grave's disease, characterized by an autoimmune reaction against the thyroid TSH receptors, which in most cases stimulates the release of excess thyroid hormones. Some of the physical signs of hyperthyroidism include weight loss, palpitations, frequent bowel movements, muscle weakness, thyroid enlargement, and bulging out of the eyes (exophthalmia). Psychiatric symptoms often associated with hyperthyroidism include insomnia, irritability, agitation, general anxiety disorder, major depression, attention deficit disorder, or paranoia. General anxiety is by far the most common psychiatric outcome and generally improves upon the normalization of thyroid hormones. Treatment with antidepressants has also been observed to lower the levels of circulating thyroid hormones.

IV. THE HYPOTHALAMO–PITUITARY GROWTH AXIS

No the hyphen is not missing between pituitary and growth; this is because this endocrine system departs from other neuroendocrine systems by releasing the anterior pituitary growth hormone (GH, also known as somatotropin) directly into the bloodstream to influence body tissues without the intervention of an additional organ (as is also the case with prolactin secretion, discussed later). Growth hormone, as the name implies, is vital for the normal growth and development of humans and most other species, particularly during the perinatal and adolescent periods. However, together with other factors and hormones, including, but not limited to, thyroid hormones, insulin, and dietary nutrients, growth hormone is crucial in the control of carbohydrate metabolism, protein synthesis, body fat, glucose utilization, and other metabolic activity. Without it, abnormal organ, skeletal, and immune functions develop.

A. Basic Constituents of the Hypothalamo–Pituitary Growth Axis

The hypothalamo–pituitary growth axis also departs from other neuroendocrine axes in that two hypothalamic factors mainly regulate growth hormone synthesis and secretion, with one factor demonstrating GH release activity whereas the other inhibits GH secretion.

1. The Anterior Pituitary and Growth Hormone

Growth hormone is released by specialized cells of the anterior pituitary gland called somatotropes. They are located predominantly in the lateral "wings" of the anterior lobe and constitute about 40% of the total anterior pituitary cells (there is a significant gender difference, with women averaging lower percentages than men). In humans, the pituitary gland contains 5–10 mg of somatotropin, a 191-amino-acid peptide derived from a gene located on chromosome 17.

The half-life of GH is relatively short, with an average of approximately 20 min. Upon secretion into the bloodstream, about half of the GH released becomes bound to a protein termed growth-hormone-binding protein (GHBP). GH acts in several different tissues throughout the, body, as discussed earlier, and this action is mediated by specific GH receptors (GH-R) located on the membranes of receptive tissues and organs. Both GH-R and GHBP are derived from the same gene, although alternative splicing of the transcribed mRNAs provides for the two distinct functions of these proteins. Interestingly, to be biologically active, one GH molecule is required to bind sequentially to two GH receptors. This complex then initiates a cascade of events leading to the activation of transcription factors affecting GH-regulated genes. A major effect of GH is to stimulate the production of insulin-like growth factor I (IGF-I), which mediates many of the effects of GH; IGF-I by itself can mimic many effects of GH, although the two molecules are more effective in combination.

GH is secreted by the anterior pituitary upon release of the growth-hormone-releasing hormone (GHRH) signal from the hypothalamus, via the portal blood system, onto GHRH receptors (GHRH-R) of the somatotropes. These receptors are also part of the large family of seven-transmembrane, G-protein-coupled receptors, and they are specifically linked to the stimulatory subtype (G_s) of G-proteins, which activates the cAMP second messenger pathway. Additional evidence also indicates that an as yet

unknown GH-releasing factor from the hypothalamus binds specifically to a different G-protein-coupled receptor, one that is specifically linked to the G_q subtype of G-proteins and appears to synergize, through the inositol triphosphate and diacylglycerol second messenger pathways, with the effects of GHRH-R upon GH secretion. The synthesis of GH in somatotropes is independently controlled by GHRH through a slightly different set of transcription factors associated with cAMP production (cAMP response element binding protein), acting to enhance transcription of the somatotropin gene. Independent regulation of synthesis and release has been shown by the existence of different diseases that are mediated specifically at the level of the pituitary gland upon either GH synthesis or secretion within somatotropes.

The secretion of GH is also negatively regulated by a second hypothalamic factor, somatostatin (SS). Pituitary secretagogues, including somatotropes, contain a variety of SS receptors (at least five different variants with independent genomic origins) that are all linked to the inhibitory G-protein subunit G_i, which inhibits the activation of adenylate cyclase and the production of cAMP. Thus, GH synthesis and secretion are tightly regulated positively and negatively by hypothalamic factors.

2. The Hypothalamus, Growth-Hormone-Releasing Hormone, and Somatostatin

As discussed earlier, the major hypothalamic factor responsible for the synthesis and release of pituitary GH is GHRH. Hypophysiotropic GHRH neurons are located at the base of the brain against the third ventricle in the arcuate nucleus. The human GHRH gene is located on chromosome 20 and gives rise to a preproGHRH precursor of 108 amino acids. Posttranslational processing of proGHRH produces two forms of GHRH: a 44-amino-acid peptide, to which 4 amino acids are found truncated at the C-terminus in some forms. Because the bioactivity of GHRH is restricted to the first 29 amino acids at the N-terminus, this truncation, albeit unique to humans, confers both forms with similar biological activity.

The other hypothalamic factor having a major influence on the control of GH secretion is somatostatin (SS), also known as growth-hormone-release-inhibiting hormone (GHRIH) or somatotrope-release-inhibiting hormone (SRIH). Although SS is found in many cell populations throughout the brain, the important SS-containing neurons in regard to GH release are those in the rostral periventricular region. The human SS gene, located on chromosome 3, encodes a 116-amino-acid preproSS. Posttranslational processing of preproSS in the hypothalamus produces two biologically active forms, SS-28 and SS-14, truncated at the N-terminus. Both forms are released in the external zone of the median eminence and display a short half-life of a few minutes to reach the anterior pituitary somatotropes. A direct inhibitory action of SS upon GHRH-containing arcuate neurons also exists, although the source of this SS input likely is from nuclei other than from the hypophysiotropic SS-containing rostral periventricular region.

B. Control of the Growth Axis

1. Control of Circulating Growth Hormones

As with other neuroendocrine systems, GH and its related product, IGF-I, provide feedback inhibition at both pituitary and hypothalamic levels upon pituitary GH release and upon SS and GHRH synthesis and secretion, respectively. At the level of the pituitary somatotropes, IGF-I is the major inhibitor of GH synthesis and secretion under basal and GHRH-stimulated conditions, with GH displaying little, if any, inhibitory effects. Somatotropes contain IGF-I receptors, which are in a position to respond to both locally produced and circulating IGF-I. Although total IGF-I levels are relatively stable over time, they exist in both free form, which is biologically more active, and in bound form to a plasma protein, IGFBP-3. A number of physiological conditions, such as hyperglycemia, down-regulate IGFBP-3, thus increasing free, biologically active levels of IGF-I, which can then mediate increased inhibition of GH. At the level of the hypothalamus, GHRH- and SS-containing neurons possess both IGF-I and GH receptors, which open these cellular populations to the feedback effects of both molecules. There is limited evidence for IGF-I- and GH-mediated inhibition of GHRH secretion and SS stimulation, which is perhaps linked to the limited crossing of these large peptides across the blood–brain barrier.

Additional neurotransmitter systems are also known to play a role in the secretion of GH, which include, but are probably not limited to, vasoactive intestinal polypeptide (VIP), pituitary adenylate cyclase activating peptide (PACAP), TRH, glucocorticoids, androgens, estrogens, galanin, neuropeptide Y (NPY), interleukins-1 and -2, and the classical neurotransmitters dopamine, noradrenaline, and acetylcholine.

The secretion of GH is also pulsatile, as with most other hormones of the neuroendocrine system. In

humans, there are approximately 12 pulses per 24-hr period, with most of the release (roughly 70%) occurring at night during slow-wave (stages 2–4) sleep. Existing evidence suggests that both GHRH and SS are implicated in the pulsatile release of GH, with GHRH being the dominant regulator in mediating the release of pituitary GH. However, alterations in either GHRH or SS have been shown to block the pulsatile release of GH. In the hypothalamus, GHRH and SS mRNAs also display fluctuating levels, with GHRH and SS mRNA peaks anticipating the GH peaks and troughs by several hours, respectively. There is also a strong sexual disparity in the amplitude of release, with higher levels of estrogen in women contributing to significantly lower GH secretion. Testosterone, which has higher circulating levels in men, has been linked to the higher peak amplitude of GH release in males.

2. Brain Circuits Involved in the Regulation of the HP Growth Axis

Some of the same circuits described previously control the arcuate GHRH-containing neurons as well as the rostral periventricular SS-containing neurons. The brain stem and medullary catecholaminergic cell groups innervate and control the activity of arcuate and periventricular neurons. Other major inputs to the arcuate nucleus originate within the hypothalamus, including the anterior periventricular nucleus, the retrochiasmatic nucleus, the medial parvocellular region, the medial preoptic area, and the ventral premammillary nucleus. The bed nucleus of the stria terminalis also innervates the arcuate nucleus, as is the case for many other hypothalamic subregions. The rostral periventricular region demonstrates a pattern of innervation generally similar to that described for the paraventricular nucleus described earlier, including midbrain, thalamic, hypothalamic, and forebrain afferents. GH and IGF-I receptors are widely distributed throughout the human brain and could mediate some of the effects of circulating growth hormones and their products. However, peripheral GH and IGF-I have limited penetration to the brain due to the blood–brain barrier, and, furthermore, GH and IGF-I are produced in the brain. Thus, direct effects of circulating growth hormones on the brain have been difficult to demonstrate. Similarly, the psychotropic effects of growth hormones have not been well-characterized nor explained. Nevertheless, dysregulation of circulating growth hormones produces marked cognitive and mood alterations.

C. Dysregulation of Circulating Growth Hormones

Physically a very observable effect of growth hormone insufficiency during development is dwarfism (short stature). However, some types of dwarfism result from the insensitivity of tissues to relatively normal circulating GH levels. The psychotropic effects of GH insufficiency in children are not reported to alter intelligence measures, but a higher incidence of attention deficit disorders, anxiety, depressive mood, somatic complaints, impulsivity, distractibility, social adjustment, and attention-seeking are common. The similarity of these psychological effects in idiopathic (not produced by GH dysregulation) short stature vs GH-deficient (GHD) children indicates that environmental and social factors play a role in the development of these traits. Nonetheless, the behavioral, social, and psychological disturbances of GHD children are ameliorated to a greater extent following GH treatment, suggesting a role for circulating GH levels in these disorders. The impact of GH deficiency in adults of normal stature also indicates a higher incidence of affective disorders, lack of energy, and impaired self-control, which are significantly alleviated with GH replacement. Psychosocial dwarfism, a state of short stature sustained by parental abuse, is associated with significantly lower circulating GH levels. Removal of the afflicted children from the abusive environment will often reverse this condition without any other intervention. These examples offer clear instances of the interplay between hormones, psychological status, and the brain.

Just as striking is the effect of excess circulating growth hormones during development, leading to *gigantism* in children and *acromegaly* in adults. These conditions are often associated with pituitary somatotrope adenomas resulting in hypersecretion of growth hormones. Children that grow to be giants have relatively normal features except for their high stature. These children tend to have mood swings, irritability, and lack social skills. They tend to be intellectually normal. The physical and psychological characteristics can be ameliorated with long-acting somatostatin agonists, which reduce circulating GH levels, or with the surgical removal of the adenoma. This growth hormone condition is distinguished from another form of early childhood gigantism known as *cerebral gigantism* or *Sotos syndrome*, which originates from

brain abnormalities and is associated with abnormal EEG, mental retardation, and other physical characteristics, particularly of the facial and cranial bones.

In adults, abnormally high circulating GH levels, also produced by pituitary somatotrope adenomas, present a characteristic pattern of physical changes, including enlarged hands and feet, profuse sweating, protrusion of the lower jaw, overgrowth of facial and cranial bones, hirsutism (increased body hair), osteoarthritic vertebral changes (humpback), headaches, and visual field changes (due to pressure of the pituitary tumor against optic nerves). The psychological symptoms of acromegalic adults include affective disorders (anxiety and major depression), increase in appetite, and loss of drive and libido. These individuals also exhibit normal intelligence and memory functions. Surgical intervention or somatostatin agonists that normalize GH levels ameliorate their physical and psychological symptoms.

V. THE HYPOTHALAMO–PITUITARY–GONADAL AXIS

The hypothalamo–pituitary–gonadal (HPG) axis is responsible for many reproductive functions in humans and other mammals. It plays a pivotal role in the development of mature male and female reproductive organs, secondary sexual traits (breasts, body hair, vocal pitch), and the sexually dimorphic brains allowing appropriate male and female behaviors leading to reproduction, such as courtship and intercourse and probably mate choices and sex preferences. However, many of the stereotypic behaviors attributed to the HPG axis in lower mammals, particularly courtship, mating, and intercourse behaviors, have, for most intents and purposes, been lost in humans. However, sex hormones have been suggested to have some effect on sexual desire, which alternatively has not been investigated to any great extent in animals displaying readily observable stereotypic sexual behaviors. Sexual differentiation is determined very early on during development. The presence of a Y chromosome following fertilization directs the development and maturation of the testes, which begin to produce testosterone early *in utero* and lead to the male phenotype. Without the Y chromosome, the "default" female phenotype emerges. The following section describes the results of this early differentiation and how the various HPG functions are derived.

A. Basic Constituents of the Hypothalamo–Pituitary–Gonad Axis

The HPG consists of a relatively classic neuroendocrine system, in that the various sex hormones are regulated by hypothalamic, pituitary, and end organs (gonads). It is arguably the neuroendocrine system with the most components, with different sex hormones to regulate and track. The different sex hormones, the elements that regulate their secretion, and their effects on various body tissues and reproductive functions are described next.

1. The Gonads and Sex Hormones

The HPG axis is very different between males and females. The sex hormones, also known as the sex steroids or gonadal steroids, consist of *estrogens*, *progesterone*, and the *androgens*, the most well-known of which is *testosterone*. Although all of these hormones are found in both men and women, estrogens and progesterone are observed in much higher levels in women and testosterone in much higher levels in men. The sex steroids are all derived from the same precursor molecule, cholesterol, through a series of aromatic reactions taking place in the gonads (and some in the adrenal cortex and placenta of pregnant women). Most of the sex steroids become bound to plasma proteins [albumin, transcortin, gonadal-binding globulin (GBG), or other proteins] upon release in the bloodstream, leaving approximately 3% of the released pool free. Steroids are relatively small molecules that easily cross lipid membranes. This feature allows sex steroids to act on their respective tissues by interacting with specific receptors (such as the estradiol, progesterone, and androgen receptors) located intracellularly in the cytoplasm of cells, as opposed to binding cell-surface (membrane) receptors. Binding of steroids with their specific cytoplasmic receptors induces a series of reactions that ultimately produces the *translocation* of the steroid–receptor complex (known as a *transacting factor*) from the cytoplasm to the nucleus, where it interacts with specific DNA elements in the promoter region that can increase or decrease the transcription of associated genes.

In females, estrogens and progesterone are synthesized in the ovaries upon stimulation by the anterior pituitary hormones luteinizing hormone (LH) and follicle-stimulating hormone (FSH), together known as the gonadotropins. More specifically, FSH produces the growth of follicles within the ovaries (the *follicular*

phase), which produces increasing amounts of estrogens (estradiol mostly). An LH surge produces ovulation, in which a follicle releases an ovum, the former being transformed in the process to the *corpus luteum*. The *luteal* phase is responsible for continued synthesis and release of estrogens as well as increasing amounts of progesterone. The concerted effects of estrogens and progesterone on the uterine endometrium prepare the uterus for possible implantation by a fertilized egg. Estrogens and progesterone are also responsible for the development and maturation of the female secondary sex characteristics, including breast development and the female pattern of fat deposits. The regulation of synthesis and release of estrogens and progesterone needs to be understood more specifically in the cyclical context (ovarian or "menstrual" cycle in women) in which they occur. This will be discussed after all HPG elements have been introduced.

In males, stimulation by pituitary LH chiefly signals the production of testosterone by the interstitial (Leydig) cells of the testes. In conjunction with FSH, testosterone stimulates spermatogenesis in the seminiferous tubules of the testes. Testosterone also plays a role in the control of penile erection and the synthesis and secretion of prostate and seminal vesicle fluids. Testosterone is also responsible for the development of male secondary sex characteristics, which consist of the general male body form and more massive muscle development, male hair distribution, and enlargement of the larynx, leading to deepening of the voice. The relatively low female sex steroid levels in the male circulation also lead to a different pattern of fat deposition in men. Interestingly, testosterone not only acts through androgen receptors but is also converted to a significant degree to estradiol, particularly in the brain and pituitary, where the converted products act on local estrogen receptors.

2. The Anterior Pituitary Gland and Gonadotropins

As discussed earlier, LH and FSH are gonadotropins that are synthesized and released by the *gonadotropes*, constituting less than 10% of the total cells of the anterior lobe. The half-life of human FSH is about 3 hr, whereas that of LH is approximately 60 min. LH and FSH are glycoprotein hormones derived from a common α gene (located on chromosome 6) but two distinct β genes: LHβ, located on chromosome 19, and FSHβ, located on chromosome 11, respectively. Upon signaling mediated by hypothalamic gonadotropin-releasing hormone (GnRH), also known as luteinizing-hormone-releasing hormone (LHRH), cell-surface GnRH receptors (mapped to chromosome 4 in humans), which are parts of the G-protein-coupled receptor family, activate the cAMP second messenger pathway to stimulate the synthesis and release of LH and FSH. A relatively robust correlation has been observed between pulses of GnRH and LH release in the general circulation. However, the concordance between GnRH pulses and FSH release is variable due to the action of additional factors, such as the peptide inhibin, which is produced in the gonads and appears to repress the synthesis and release of FSH.

3. The Hypothalamus and Gonadotropin-Releasing Hormone

The release of gonadotropins is regulated by hypothalamic GnRH-containing neurons. The GnRH-containing neurons are scattered throughout the extent of the hypothalamus, unlike other releasing hormone cell groups. In humans, they are mostly localized to the medial preoptic area near the optic chiasm and to the arcuate nucleus (ventral) hypothalamus near the median eminence, with some additional cells in the lateral hypothalamus and medial septal complex. Approximately 70% of hypothalamic GnRH-containing neurons send their axons to the external zone of the median eminence, where they release GnRH in the hypophyseal portal blood system and target the gonadotropes. GnRH is a relatively small 10-amino-acid peptide that is also derived from a larger preproGnRH product of 92 amino acids. In humans, a single gene codes for hypothalamic preproGnRH located on chromosome 8. Studies have indicated the existence of a second, distinct GnRH gene on chromosome 20 that is expressed in extrahypothalamic areas and peripheral tissues. The function of this extra-hypothalamic-derived GnRH has not been determined.

4. The Anterior Pituitary Gland and Prolactin

Although not part of the HPG axis, prolactin's best-known function is in lactation and, thus, warrants its inclusion with other reproductive hormones. Prolactin, a 198-amino-acid peptide derived from preproprolactin, is released by anterior pituitary lactotropes (also known as mammotropes). Unlike other anterior pituitary tropic hormones, prolactin is under negative control by an inhibitory factor from the hypothalamus, known as prolactin-inhibiting factor (PIF), likely to be dopamine. Thus, upon transection of the pituitary stalk, removal of PIF actions leads to hyper-prolactinemia. Like other pituitary hormones,

pulsatile release is observed with prolactin, with a circadian rhythm displaying greater release during sleep. Both men and women show circulating prolactin levels, with slightly higher levels observed in women. Prolactin acts through specific, multiple forms of its receptors found throughout the reproductive axis of women and men, in addition to the liver, kidneys, adrenals, pancreas, lymphoid tissues, intestines, hypothalamus, and skin.

During pregnancy, circulating levels of prolactin rise significantly but decline rapidly approximately 1 week following childbirth. Suckling then initiates significant bursts of prolactin secretion, which stimulate the synthesis of maternal milk proteins. Pathologically high circulating prolactin levels are associated with gonadal dysfunctions in both women and men, in addition to headaches and visual field disturbances. Hyper-prolactinemia is reportedly the most common form of hypothalamopituitary disorder in humans, which can normally be treated in several ways to restore normal circulating prolactin levels.

B. Control of the HPG Axis

1. Control of Circulating Gonadal Hormones

Control of circulating sex hormones is determined to a large extent by the sex of the individual, and in women it is determined further by the period within the ovarian cycle. As observed with the other neuroendocrine systems, there is a strong component of feedback inhibition provided by the sex steroids at the pituitary and hypothalamic levels. The pulsatile nature of GnRH release could be mediated by the documented extensive connections between the several GnRH neuron populations or by afferent connections. It is clear, however, that the GnRH population of the arcuate nucleus is necessary to derive the pulsatility of GnRH secretion.

In women, it is customary to define the menstrual cycle as beginning on the first day of the menstrual period, which is easily detected by women experiencing vaginal bloody discharges consisting of the endometrial tissue breaking down and being released (sloughing) because fertilization has not taken place. At this point (beginning of the follicular phase), the hypophysiotropic GnRH-containing neurons generate a pulsatile LH release approximately every hour, which is a time at which estrogens are relatively low. As the ovarian follicles grow, they release increasing amounts of estrogens, which feed back at the level of the hypothalamus and reduce the frequency of LH and FSH release to 1 pulse per 90 min. At the same time, the most developed follicle also synthesizes increasing amounts of LH and FSH receptors, thus providing increased gonadotropin sensitivity in the face of decreasing gonadotropin release from the pituitary. The follicle begins to synthesize and secrete large amounts of estradiol, which, through ill-understood mechanisms, produce a *positive feedback* effect at the level of both the pituitary gonadotropes and the GnRH-containing neurons of the hypothalamus, triggering a gonadotropin surge (large amounts of LH and FSH released in the circulation) at approximately midcycle (14 days). This surge is responsible for ovulation and the transformation of the follicle into the corpus luteum. In the luteal phase of the cycle, increasing amounts of progesterone and moderate levels of estrogens are secreted, which feed back upon the pituitary and hypothalamus to reduce the frequency of LH pulses to 1 every 6–12 hr. The rising levels of progesterone and inhibin lead to an inhibition of FSH release as the ovarian cycle progresses. However, at the end of the cycle, with lack of fertilization of the ovum, the corpus luteum degenerates and significantly reduces its release of estrogens and progesterone. Lower levels of sex steroids release FSH from feedback inhibition and allow the cycle to repeat itself.

In men, LH pulse frequency remains relatively constant throughout adulthood, with a frequency of 1 pulse every 2–3 hr. High testosterone levels, however, can significantly inhibit LH release, but FSH release is not altered by testosterone. This inhibition thus occurs at the level of the anterior pituitary gonadotropes. Inhibin, a 3l-amino-acid peptide produced in the testes and ovaries, inhibits the release of FSH in men, which is thought to act mainly at the pituitary level.

Some of the feedback inhibition mediated by estrogens occurs at the level of the anterior pituitary. Estradiol, through interactions with their receptors, reduces the effectiveness of hypothalamic GnRH upon the gonadotropes, thus reducing the "amplitude" of LH pulses but not their frequency. The feedback inhibition mediated by progesterone, on the other hand, is mediated exclusively at the level of the hypothalamus. Hypothalamic feedback inhibition initiated by gonadal steroids is unlikely to be mediated at the level of the hypophysiotropic GnRH-containing neurons because studies have failed to detect gonadal steroid receptors in these neurons. However, many neural systems contacting hypophysiotropic GnRH neurons express these receptors. Thus, the brain circuits providing innervation to hypophysiotropic GnRH-containing neurons appear to be an important

component regulating circulating gonadal steroids. The large number of developmental (fetal life, childhood quiescence, puberty) and environmental factors (daylight, nutrition, stress, exercise) regulating the HPG axis also point to the importance of the circuitry controlling it.

2. Brain Circuits Involved in the Regulation of the HPG Axis

As discussed earlier, the presence of sex steroids, particularly testosterone, around birth is crucial for the development of the male phenotype and, particularly, the brain regions involved in reproduction. These dimorphic sex differences include larger nuclei in the brains of men, including a region of the medial preoptic, suprachiasmatic, and anterior hypothalamic nuclei. Given the multiple origins of GnRH-containing neurons in the medial septal complex and hypothalamus, the neural regulation of GnRH activity thus could be very complex. There are a few well-characterized inputs that may play an important role in reproduction. For instance, the sexually dimorphic preoptic area receives afferents from the medial part of the bed nucleus of the stria terminalis, the ventral part of the lateral septal nucleus, the medial amygdaloid nucleus, and the amygdalohippocampal area. These areas are rich in sex steroid receptors and, in animals, are known to contribute significantly to copulatory behaviors as well as to sexual interest (perhaps desire). The medial preoptic area also receives projections from the sexually dimorphic medial preoptic nucleus, as well as from some medullary catecholaminergic neurons that contain estrogen receptors, which may mediate, in part, the activational effects of catecholamines on GnRH release. Several other neurotransmitter systems appear to directly stimulate or inhibit GnRH release, including serotonin, GABA, substance P, neuropeptide Y, CRH, and opiate peptides.

C. Dysregulation of Circulating Gonadal Hormones

Observed gonadal hormone dysregulation has several etiological causes. In women, menopause provides a natural cause of long-term reduction in circulating estrogens and progesterone, because the ovaries progressively lose their sensitivity to gonadotropins. Surgical removal of the ovaries and various degenerative ovarian diseases also account for reduced circulating sex steroid levels in women. Low levels of circulating sex steroids lead to larger pulsatile releases of LH and FSH. The LH pulses are associated with "hot flashes," a sensation of warmth spreading from the trunk to the face. However, LH is not directly involved in these hot flashes because hypophysectomy does not abolish them; thus, a hypothalamic signal responsible for both the LH surge and the hot flashes appears to be responsible for this condition. Estrogen replacement usually ameliorates the hot flashes and the psychological fluctuations (irritability, depressed mood) associated with menopause.

It is also the case that women have a much higher incidence of depressive disorders than men. Attempts to correlate reproductive hormone status with depression have generally suggested that women are most susceptible to depression at the times in their ovarian cycles when they experience the largest variations in circulating hormones. Thus, women in the perimenopausal period as well as after childbirth (post partum) show increased risks of depressive illnesses, particularly if they have a prior history of depressive mood. In addition, the late luteal phase is associated with the premenstrual syndrome (PMS), a period characterized by increased cramps, bloating, breast tenderness, and headaches physically and by significant moodiness, anger, anxiety, and even depression psychologically in some 30% of cycling women. In carefully controlled studies, no differences between circulating estrogens, progesterone, or LH could differentiate women with or without PMS.

Nevertheless, effective treatment of PMS involves the elimination of ovarian cycling, suggesting, somehow, the involvement of circulating reproductive steroids in this condition. Disturbances in memory retrieval have also been associated with women experiencing PMS. Estrogens have been shown to interact significantly with subtype-specific serotonin receptors (5-HT_{2A}) in some limbic cortices and the nucleus accumbens, areas putatively involved in mood. This may provide the basis for the observation that serotonin-specific re-uptake inhibitors (SSRIs), one of the latest class of antidepressant drugs, may have beneficial effects in menopausal depressed women, particularly in combination with estrogen replacement.

The testes usually continue to respond to gonadotropins throughout life even if it somewhat declines with age; thus, men do not experience menopause. However, the loss of the testes either surgically (orchidectomy) or by diseases almost invariably leads to a decline in sexual desire and impotence. Physically, the lack of circulating testosterone also reduces some

of the secondary male sex characteristics, including muscle loss, reduced body and facial hair, and deposition of fat around the hips. Hot flashes also are experienced by orchidectomized (castrated) men. Testosterone replacement can reverse all of these conditions, just as estrogen replacement reverses the effects of low circulating estrogen levels in women. Although aggressive behavior has been linked with slightly higher circulating testosterone levels in men, the general belief is that these might be reactive effects, not the mediators of aggression as such. The possibility still exists, however, that some violent–aggressive types of behaviors might be correlated with higher circulating testosterone levels. As is the case with the ovaries, there are virtually no known conditions leading to consistently high circulating testosterone levels.

VI. FINAL COMMENTS

It is obvious from the preceding discussion that there are clear psychologic and psychiatric outcomes associated with endocrine imbalances. It is also remarkable that one of the major psychologic outcomes of hormonal dysregulation, be it due to hypo- or hyperfunction, involves mood disorders. Restoration of hormonal balance ameliorates the associated mood disturbances in many instances. Likewise, endogenous psychiatric conditions encompassing affective disorders such as general anxiety and major depression have a significant impact on several hypothalamo–pituitary endocrine systems. Effective treatment of these conditions also normalizes, in most cases, endocrinologic functions. These observations strongly suggest a close association between the brain substrates underlying affect, on the one hand, and the control of endocrine systems, on the other hand. Exactly how dysregulation in one results in disruption of the other remains to be determined. However, their close interactions further suggest that the same brain systems responsible for affect will also be prominently responsible for the control of psychoneuroendocrine functions.

See Also the Following Articles

CHEMICAL NEUROANATOMY • CIRCADIAN RHYTHMS • COGNITIVE PSYCHOLOGY, OVERVIEW • HOMEOSTATIC MECHANISMS • NEUROPSYCHOLOGICAL ASSESSMENT • NEUROTRANSMITTERS • PAIN AND PSYCHOPATHOLOGY • PEPTIDES, HORMONES, AND THE BRAIN AND SPINAL CORD • PSYCHONEUROIMMUNOLOGY • PSYCHOPHYSIOLOGY • SEXUAL DIFFERENTIATION, HORMONES AND • STRESS: HORMONAL AND NEURAL ASPECTS

Suggested Reading

Akil, H., Campeau, S., Cullinan, W. E., Lechan, R. M., Toni, R., Watson, S. J., and Moore, R. Y. (1999). Neuroendocrine systems I: Overview—Thyroid and adrenal axes. In *Fundamental Neuroscience* (M. J. Zigmond, F. E. Bloom, S. C. Landis, J. L. Roberts, and L. R. Squire, Eds.). Academic Press, San Diego, CA.

Frohman, L., Cameron, J., and Wise, P. (1999). Neuroendocrine system II: Growth, reproduction, and lactation. In *Fundamental Neuroscience* (M. J. Zigmond, F. E. Bloom, S. C. Landis, J. L. Roberts, and L. R. Squire, Eds.). Academic Press, San Diego, CA.

Ganong, W. F. (1999). *Review of Medical Physiology*, 19th ed. McGraw-Hill, New York.

McEwen, B. S. (1994). Endocrine effects on the brain and their relationship to behavior. In *Basic Neurochemistry* (G. J. Siegel, B. W. Agranoff, R. W. Albers, and P. B. Molinoff, Eds.). Raven Press, New York.

Nelson, R. J. (1995). *An Introduction to Behavioral Endocrinology*. Sinauer Associates, Inc., Sunderland, MA.

Nemeroff, C. B. (1992). *Neuroendocrinology*. CRC Press, Inc., Boca Raton, FL.

Nemeroff, C. B. (1999). *The Psychiatric Clinics of North America—Psychoneuroendocrinology*, Vol. 21, No. 2. W. B. Saunders Co., Philadelphia, PA.

Schulkin, J. (1999). *The Neuroendocrine Regulation of Behavior*. Cambridge University Press, Cambridge, UK.

Psychoneuroimmunology

DAVID L. FELTEN
Loma Linda University School of Medicine

MARY E. MAIDA
University of Rochester Medical Center

GLOSSARY

cytokines Signal molecules released from immunocytes, supporting cells, or other cell types (e.g., endothelial cells) that can interact with other cells of the immune system or additional targets and alter their intracellular activity.

immunocytes Cells of the immune system.

innervation The distribution of nerve fibers that end in association with a particular structure, region, organ, or cell.

neurohormone A blood-borne signal of endocrine nature that either is derived from the brain directly or is released from the pituitary gland due to neurally derived factors.

neuromodulation Interaction of a neurally derived signal with a target cell that alters the intracellular capabilities of that cell through either genomic mechanisms or intracellular transduction pathways.

neuropeptides Neural signal molecules derived from larger prohormones that consist of 3 to 100 or more amino acids released from nerve endings, cells of neural origin, or cells of immune or endocrine origin.

neurotransmitter A signal molecule released from nerve endings that can interact with specific receptors on target cells and alter either ion channels or second messengers following transduction.

The field of psychoneuroimmunology has demonstrated that the immune system is not an autonomous, exclusively self-regulating system, neither is the brain devoid of influences emanating from immune signaling. Rather, the CNS and the immune system appear to interact with complex signaling and feedback loops for the protection and homeostasis of the whole organism.

I. THE NERVOUS SYSTEM AND THE IMMUNE SYSTEM INTERACT WITH EACH OTHER

A. Historical, Clinical, and Experimental Perspectives

For centuries, an association has been described between "mind" or "emotions" and the course and consequences of disease. Galen proposed an association between "melancholia" in women and breast cancer almost 2000 years ago. Although suspicions of such mind–body interactions often were voiced, the modern era of the scientific understanding of disease and the discovery of micro-organisms as the

cause of infectious diseases led the medical profession away from a more holistic or integrated approach to medicine. Modern mechanistic medicine moved in a reductionistic fashion toward molecular and cellular explanations for disease, an approach that often did not include psychological and psychosocial factors. Thus, medicine in the modern era focused on the "body" without integrating the "mind" and "spirit."

Immunology grew up as a modern discipline with the underlying premise that the immune system is an autonomous and self-regulating system that responds to its own signals and is not subject to perturbation from neural influences, especially those in the realm of the "mind" or "spirit." However, with the advent of parallel integrative research approaches exploring interactions between the nervous system and the immune system during the past two decades, it has become increasingly evident that these two great communication systems are intimately integrated with each other and have constant, ongoing signaling between them. These interactions have a profound functional influence, including altering the onset, progression, and outcome of many disease processes. This field has come to be called "psychoneuroimmunology" (PNI), a name that acknowledges ongoing interactions among behavioral, neural, endocrine, and immunological processes, whose collective communication networks are integrated for the maintenance of homeostasis within the whole organism. An explosion of information in the past decade has revealed wide-ranging links among these systems using molecular signaling via hormones (endocrine), neurotransmitters (neural), and cytokines (immune). Receptors are present on many cells of all three systems for all of these classes of molecules, indicating the capacity for intersystem communication. Studies from the systemic level to the molecular level have provided evidence that these neural–immune channels of communication not only exist but exert a profound influence on the onset and course of disease and even on the life or death of an organism. Although the public and the popular press have embraced this concept as "common sense" and intuitively obvious, scientific and clinical circles have been far more skeptical and initially were downright hostile to the idea that neural–immune influences could be of functional importance. This article will review the evidence that such communication channels exist, do indeed have important consequences for the understanding of signaling within and among the integrated systems, and have functional significance for the practice of medicine.

1. Early Studies Linked Stressors to Increased Susceptibility and Pathological Effects of Viruses in Experimental Animals

In the late 1950s, Rasmussen and colleagues studied the effects of a behavioral stressor, avoidance conditioning, on the susceptibility of mice to a variety of viral infections. This psychological stressor increased the susceptibility of mice to herpes virus, Coxsackie B virus, and vesicular stomatitis virus. In additional work in the 1960s, Rasmussen and colleagues also measured immunological variables as well as disease outcome in stressed mice, a foreshadowing of the more detailed mechanistic studies that were carried out in the late 1980s and 1990s. These same investigators also showed that a stressor in mice could change the course of malignancy from polyoma virus, another forerunner of more recent studies investigating neural–immune signaling and progression and metastases of specific cancers.

George Solomon, a pioneer in the field of PNI in both human and animal studies, in the 1970s, showed that handling during early life could alter subsequent primary and secondary antibody responses to a bacterial antigen, an important component of the sophisticated acquired immune responses. Solomon also showed that the characteristics of the stressor influence the nature and extent of the subsequent alteration in antibody production. These studies implied a link between the central nervous system (CNS) and the cells of the immune system that are engaged in antibody production. Solomon and colleagues also investigated the effects of stressors on animal models of rheumatoid arthritis, graft verus host response, and virally induced tumors, again showing a link between stressors and immunological outcomes. These findings were not appreciated at that time for their almost revolutionary implications for holistic or integrative medicine.

2. Early Clinical Studies Described an Association between Stress and Human Susceptibility to Immune-Mediated Diseases

Solomon and colleagues studied patients with autoimmune disease (rheumatoid arthritis) and found that psychological well-being could provide a protective influence over expression of this disease. Many other correlational studies in the 1970s and 1980s identified psychosocial factors as important influences on the susceptibility to, clinical course of, or recovery from many types of diseases in which the immune system is

involved, including bacterial infections, viral infections, allergic responses, autoimmune diseases, and cancer. However, because these studies were correlational and were not able to specify the precise alterations in signaling or the biological mechanisms by which such influences were exerted, they were not viewed seriously in scientific circles until further cellular and molecular studies were undertaken in the past few years that pointed toward the mechanistic basis for some of these findings. A further complication of studies involving stressors is that subtle variations in the stressor can influence biological processes in a differential fashion; these effects are superimposed on the pathophysiological alterations produced by the disease itself. Also, the demonstration that a stressor alters the outcome of an immune-mediated disease does not pinpoint or even require a direct effect of a neural signal, provoked by the stressor, on cells of the immune system.

3. Experimental Studies Sought Mechanistic Explanations for Neural–Immune Interactions

Both human and experimental animal studies in the 1970s, 1980s, and even earlier indicated that one consequence of psychological stressors was a change in the outcome or susceptibility to a variety of disease processes that involved the immune system. For decades, steroids (glucocorticoids) were used in clinical medicine to reduce inflammation, suppress immune reactivity in patients with autoimmune disease, and suppress rejection responses in transplantation. Therefore, the most convenient explanation for the influence of stressors on immune responses or disease processes was that a glucocorticoid, induced by the stressor, exerted a general suppression of immune responses. This led to some of the first experimental attempts to substantiate that stress-induced alterations in immune responses or immune-mediated diseases are regulated by the hypothalamo–pituitary–adrenal (HPA) stress axis that links the hypothalamus through the pituitary to the secretion of cortisol and other glucocorticoids from the cortex of the adrenal gland. However, studies over the past two decades that have employed adrenalectomy or glucocorticoid blockade to eliminate glucocorticoid secretion and the influence of the HPA have revealed that glucocorticoids are only one of many classes of signal molecules involved in neural–immune communication.

Signaling from the sympathetic subdivision of the autonomic nervous system [sympathetic postganglionic neurons using norepinephrine (noradrenaline) as the main neurotransmitter] also is involved extensively in neural–immune communication. Other neurohormones (e.g., growth hormone, prolactin) and neuropeptides (e.g., substance P, opioid peptides, vasoactive intestinal peptide) also mediate interactions between the nervous and immune systems. Conditioned immunosuppression, stress-induced alterations in splenic lymphocyte function, and the effects of cytokines or related molecules (IL-1 or the HIV protein gp120) are not eliminated by adrenalectomy or glucocorticoid blockade. In animal models of influenza viral infection in C57Bl/6 mice, undertaken by John Sheridan and colleagues, stress-induced exacerbation of the severity of the viral infection involved both the HPA axis and the sympathetic noradrenergic nervous system; only by blocking the excessive stress-induced activation of both of these systems simultaneously was a normal, robust immune response to the viral infection in young mice restored. These two systems are the two great "stress" axes of the body, and they exert a myriad of effects on virtually all organs and tissues of the body. They influence such processes as inflammation, wound repair, natural and acquired immunity, and a host of vascular and cell adhesion events. The rather straightforward demonstration by Sheridan and colleagues of HPA and sympathetic neural involvement in stress-induced alteration of the severity of viral infection in mice has profound implications for both acute and chronic perturbations of these axes and the consequences in humans for the susceptibility and outcome of many disease processes, particularly those mediated by the immune system. The presence of receptors for many neurotransmitters and neurohormones on many sub sets of cells of the immune system suggests that many influences of the nervous system on immune reactivity may be exerted directly by neurohormone and neurotransmitter signaling.

B. Experimental Evidence for Neural–Immune Interactions

1. Immune Responses Can Be Conditioned and Can Alter the Outcome of Immune-Mediated Disease

Behavioral conditioning is a learning paradigm that requires processing of information in the brain and subsequent outflow to peripheral target tissues involved in the conditioned response. Although some early reports of conditioning of immune responses were made by Metalnikov and Chorine in 1926, the

modern era of conditioning of immune responses began with the carefully controlled and analyzed studies by Robert Ader and Nicholas Cohen. They demonstrated that pairing of saccharine as a novel taste stimulus (a conditioned stimulus) with cyclophosphamide (an unconditioned stimulus), an agent that produces nausea and also induces immunosuppression, resulted in both taste aversion and diminished antibody response to a T-dependent antigen, sheep red blood cells, following exposure to saccharine alone. Saccharine by itself in unconditioned animals had no influence on the antibody response. Furthermore, this conditioned immunosuppression still occurred in adrenalectomized animals, eliminating the possibility that it is merely a conditioned corticosteroid stress response.

The initial response to this work by the immunology community was one of disbelief; even if conditioned immunosuppression had occurred, it was argued, the magnitude of change (approximately 30% diminution) was insignificant and had no clinical significance. However, one factor frequently overlooked is that Ader first discovered conditioned immunosuppression in a follow-up investigation seeking to explain why only conditioned animals died in response to saccharine exposure and not members of the other groups; death from pulmonary infections was indeed a significant biological endpoint. Ader and colleagues then hypothesized that, if conditioned immunosuppression had functional consequences, its application to animals with a genetic autoimmune disease (lupus-like NZB $\times$ NZW F1 mice) should prolong their lives compared with unconditioned controls. Such conditioning did prolong their lives through conditioned immunosuppression, and with drug reinforcement to prevent extinction of the conditioned response, Ader and colleagues were able to keep conditioned autoimmune-prone mice alive for a full life span of 27+ months (compared with less than 8 months in unconditioned mice) by using a total dosage of immunosuppressive drug that was considerably less than the dosage required to keep an unconditioned mouse alive. The implications for these findings is that the brain was somehow aware that immunosuppression had occurred following administration of the drug and, in response to the conditioned stimulus alone, was able to generate specific outflow mediators (other than glucocorticoids) that mimicked the initial immunosuppressive effect of the drug and exerted a life-saving influence on the immune system of these genetically autoimmune animals.

Ader and colleagues have shown that pairing of a novel taste stimulus (chocolate milk) as a conditioned stimulus with antigen itself [keyhole limpet hemocyanin (KLH), a T-cell-dependent antigen] as an unconditioned stimulus results in a robust antibody response, in conditioned animals, to a subthreshold dose of antigen that evokes no antibody response in unconditioned animals. This represents the most definitive evidence for direct conditioning of immune responses by the brain.

Numerous studies have shown that many other manifestations of immunity also can be conditioned using classical behavioral approaches, including antibody response to a T-cell-independent antigen, graft versus host response, generation of cytotoxic T lymphocytes following graft rejection, and natural killer (NK) cell activity. NK cell activity in humans can be conditioned by pairing a novel taste stimulus with low-dose epinephrine administration (releases marginated NK cells and possibly splenic NK cells) followed by reexposure to the novel taste stimulus alone, which results in enhanced NK activity in the blood. Such conditioned responses also occur in animal models. These studies suggest that behavioral conditioning can be used to achieve a specific desired immunological activation or suppression in both humans and animals; the brain generates a response using its outflow mediators, neurohormones, and neurotransmitters to produce the conditioned response.

2. Psychosocial Factors in Humans Are Associated with Altered Immune Reactivity

Adverse psychosocial factors or stressful environments have been reported to increase morbidity and mortality to challenges by some tumors or infectious diseases. Because of the central role of the immune system in response to such challenges, it has been assumed that such stressors exert their effect on cells of the immune system itself. This is reinforced by experimental evidence for the suppression of measures of both natural and acquired immunity by stressors. However, the actual demonstration of the mechanistic basis for such interactions has remained somewhat illusive. Not all effects of stressors on the immune system are suppressive, depending on the timing, nature, route, and magnitude of the challenge and many other parameters. Some stresors can exert remarkably different effects on closely related immunologic processes. A major area of missing evidence is a careful characterization of the neurohormonal and neurotransmitter alterations induced by specific stressors, examined in parametric studies that subtly vary conditions of the stressor and look for changes in the

outflow of mediators from the central nervous system (CNS). This characterization needs to be followed by direct correlation with studies using pharmacologic agonists and antagonists of the major neural–immune mediators, followed by evaluation of specific reactivity of specific subsets of cells of the immune system. We also know very little about the influence of the severity, duration, and exact timing of stressors in response to specific immune challenges or reactivity. There may be a considerable difference between an ongoing period of exposure to stress and a poststress interval, particularly in the availability of the CNS-derived mediators and the expression of the appropriate receptors on potential target cells.

a. Depression Early studies of hospitalized drug-free depressed patients found diminished responses of lymphocytes to mitogen-induced proliferation. This led to the popular notion that depression is associated with immunosuppression. However, ambulatory depressed patients did not show such changes, and the literature on studies of depressed individuals is mixed with regard to the presence of absence of various measures of immune dysfunction. One shortcoming of many observations taken from measures of immune parameters such as mitogen-induced proliferation is that such measures provide little predictive value for susceptibility to infection, course or outcome of infections, and reactivity of acquired immune responses toward cell-mediated immunity versus humoral immunity. The predictive value of blood measures is remarkably poor, given the sophisticated state of cellular and molecular immunology in experimental animals. The characteristics of the state of depression (e.g., dexamethasone response, bipolar versus unipolar), the timing and therapeutic intervention of the depression, the age and gender of the patient, and many other variables appear to have some bearing on whether the depressed individual shows altered measures of immunity. In a careful analysis of the literature, Marvin Stein showed that the likelihood of decreased measures of immunity increases with age and severity of depression, particularly in men. Sheldon Cohen, in a detailed meta-analysis of the literature on depression and immunity, concluded that there is an association of diminished immune responses, including NK cell activity (important for antiviral and some antitumor actions), and lymphocyte proliferation.

b. Bereavement Bereavement has been investigated in humans as a potential severe stressor that should reveal stress-induced immunosuppression if such an association occurs. Such studies have shown that men, but not women, who have lost their spouses (conjugal bereavement) demonstrate diminished measures of immune response (mitogen-induced proliferation, NK cell activity). This literature suggests that the quality of social support and the individual perceptions of control over their situations are buffering events to the stress-induced diminution of immune reactivity. Studies of other forms of bereavement, such as parents who lost a son in Israel in the Six-Day War, showed a lack of altered immune measures.

c. Examination Stress in Medical Students Some of the most carefully studied humans with regard to stress-induced alterations in immune reactivity are medical students undergoing blocks of examinations in the basic sciences. Jan Kiecolt-Glaser, Ronald Glaser, and colleagues demonstrated that medical students showed diminution of mitogen-induced proliferative responses, diminished NK cell activity, altered T-cell subset ratios, and increased antibody to specific viral epitopes of Epstein–Barr virus (suggesting activation of this normally latent virus) during the examination periods, with a return toward baseline in the interim examination-free periods and a return to the population baseline during summer vacation. These studies were controlled for sleep, eating habits, caffeine ingestion, and other possible confounding variables and still showed significant alterations. Those students with the poorest measures of immune reactivity showed a significant increase in clinic visits and self-reported symptoms. In some students, the inability of immune reactivity to return to baseline was associated with loneliness and poor social support. Subsequent studies by others showed that examination stress in medical students shifted T helper (Th) cytokine production away from cell-mediated immunity (Th1 cytokines, interleukin 2 and interferon-γ) and toward humoral immunity (Th2 cytokines, interleukins 4 and 10). This shift could result in a diminished capability to respond to viral infections and to clear some pathogens (e.g., influenza) for which cytotoxic T lymphocyte killing of infected cells is necessary.

d. Stress of Caregiving Kiecolt-Glaser and colleagues also have studied caregivers of patients with Alzheimer's disease and other debilitating illnesses. The caregivers demonstrated diminished lymphocyte responses to mitogens, diminished NK cell activity, and altered T-cell subset ratios, suggesting diminished

immune reactivity. However, it also must be remembered that caregiver exposure to potential infectious organisms may be different from that of their age-matched counterparts because of the restrictions inherent in caregiving. Of considerable interest in these studies is that the suppressed immune measures can remain suppressed for several years, even after the death of the individual to whom they gave care. Caregiving may be one of the most severe stressors in humans, particularly in the elderly who already may have age-related diminution of T-cell immunity from which full recovery may be difficult.

e. Other Stressors Other stressors have been studied in humans, again with measures of immune responsiveness taken from peripheral blood samples. This includes women going through a marital separation or divorce and individuals entering a nursing home. These individuals show the same diminution of measures of immune reactivity mentioned earlier. Kiecolt-Glaser and colleagues have shown that even a brief intervention for nursing home patients, such as a 1-hr visit per week from a college student for a semester, can reverse some of these measures. The most compelling findings from all the human stressor studies are the apparent importance of strong and meaningful social support and the patients' perceptions of being in control of their environment and surrounding events as buffers against stress-induced diminution of immune responsiveness.

f. Positive Factors Can Enhance Measures of Immune Responsiveness Many of the studies in humans have looked at severe stressors such as conjugal bereavement, chronic caregiving, marital difficulties, examination stress in medical students, and other apparently negative factors. A more difficult type of study is that of interventions that use "positive emotions" or approaches intended to enhance immune responsiveness. The lay press is replete with claims of programs and regimens that "enhance" immune responsiveness, with little or no supporting evidence that a specific intervention produces benefits. The best data regarding positive interventions are those showing that moderate exercise can enhance NK cell activity, CD4:CD8 ratios, and lymphocyte mitogen responses, lasting for at least 12 hr after the exercise period. Lee Berk and colleagues have studied laughter and exposure to comedic tapes and programs with many similar findings, including enhanced NK cell activity cell-mediated responses. In laughter studies, baseline cortisol and epinephrine levels are suppressed during and many hours after the laughter episode, whereas norepinephrine remains unchanged (unlike exercise, during which norepinephrine levels rise). Individuals watching videos of nature scenery accompanied by a beautiful musical score also show evidence of diminished production of these "stress mediators." Kiecolt-Glaser's demonstration of nursing home interventions with social support from visiting college students also strengthens the notion that positive benefits can be derived from a behavioral intervention. Additional surgical literature on recovery and infections suggests that the hospital environment (e.g., low ambient levels of noise, a window in the room) may play a role in more rapid recovery and a reduced rate of postoperative infections. These types of study are difficult to carry out and control for intervening variables, but they may be able to shed some light on the range of responsiveness of measures of immune reactivity with both positive and negative environmental and psychosocial factors. It also must be borne in mind that measures of immune reactivity from peripheral blood samples can only suggest, and not absolutely predict, a patient's reaction to challenge. Changes in mitogen responses of lymphocytes, T-cell subset ratios, and NK cell activity may reflect temporary changes in cell trafficking and may not be predictive of strengthening or enhancing of the overall ability of the natural or acquired arms of the immune system to mobilize in response to a bacterial infection, viral challenge, or tumor challenge. Indeed, it may be highly undesirable to enhance the immune reactivity of self-reactive clones on lymphocytes or to enhance interferon-γ production in a patient with an autoimmune disease, such as multiple sclerosis.

3. Psychosocial Factors in Humans Can Influence the Outcome of Immune-Mediated Disease Processes

The question of great clinical importance in this field is whether stress-induced or psychosocially influenced measures of immune responsiveness are predictive of, or reflections of, a disease outcome. This has been considered primarily anecdotally or in outcome studies without concomitant investigation of mechanisms and mediators. Cassileth and colleagues reviewed psychosocial factors, obtained from questionnaires, in patients with terminal cancers that were no longer treatable with surgery, radiation, or chemotherapy and found that psychosocial factors had no predictive

value or protective benefit toward death as an endpoint. This study showed that terminally ill cancer patients indeed die regardless of psychosocial variables, a not terribly surprising finding. If psychosocial factors play an important role in such diseases, they are likely to be most evident at the stages of acquisition of, susceptibility to, or progression of these diseases. The initial editorial response to this study was a blanket closing of the lid on behavioral medicine as an important factor in presumed immune-mediated diseases such as some cancers. However, other studies provided more provocative insights into the role of psychosocial factors and disease.

a. Stress and Colds Sheldon Cohen and colleagues investigated the role of life stresses related to an individual's response to challenge with cold viruses. Increasing stress in the previous 6 months was associated with increased infectivity in subjects administered one of several cold viruses. Other factors such as smoking, alcohol consumption, diet, shared housing with infectious people, personality traits, and assessment of personal control and self-esteem did not influence the increased infectivity. In the initial study, only the rate of infections (infectivity) was increased and not the symptomatology of clinical colds, antiviral antibody titers, or white blood cell counts; subsequent studies have suggested that life stressors may also influence the symptomatology of clinical colds. Thus, chronic stress appears to act at the level of infectivity and may have clinical consequences for these viral infections.

b. Counseling and Peer Support and Metastatic Breast Cancer David Spiegel and colleagues reported one of the most remarkable and provocative findings in the field of psychoneuroimmunology. They randomly assigned women with newly diagnosed, disseminated breast cancer into two groups. Each group received state-of-the-art medical care but was assigned to either receive or not receive weekly sessions of counseling and peer support for 1 year, originally to test the hypothesis that such intervention would enhance the quality of life. Not only did counseling and peer support enhance the quality of life for the women who received them, but they extended average life expectancy from 19 to 37 months. This increased longevity could not be explained by severity of medical parameters or other retrospective aspects of assignment into the groups. Unfortunately, no immune measures were monitored in these patients. Some criticism was leveled at this study because the control group had a shorter average survival time than was seen in other studies of breast cancer. Nonetheless, this controlled study did show a remarkable difference in complementary counseling and peer support added to traditional medical care. A working hypothesis is that this intervention enhanced NK cell activity, which in breast cancer patients can influence metastatic spread; it is interesting to note that NK cell activity is among the most responsive immunological measures reacting to stressors and to influences of the stress hormones and neurotransmitters. Observations by Anderson and colleagues at Ohio State University noted that women with breast cancer showed diminished NK cell activity with higher indices of life stressors and more robust NK cell activity with a lesser extent of life stressors. Follow-up studies on Spiegel's initial report, repeating these interventions and including measurements of immune reactivity, currently are under way.

c. Structured Intervention Studies with Malignant Melanoma Fawzy Fawzy and colleagues undertook a structured intervention study in patients with malignant melanoma. Short-term treatment with psychotherapeutic intervention and social support provided clinical benefit to the progression of this cancer. As in the case of breast cancer, the progression of malignant melanoma also is sensitive to NK cell activity. In addition, in many other cancers, even those that are not particularly immunogenic, it appears that robust cell-mediated immunity is associated with better clinical outcomes, and diminished cell-mediated immunity is associated with a rapid downhill course.

4. Stressors in Experimental Animals Can Alter Measures of Immune Responsiveness and the Outcome of Immune-Mediated Diseases

Many studies of stressors applied to experimental animals have been carried out to explore stress-induced alterations in measures of immune responsiveness. Infant squirrel monkeys separated from their mothers show cortisol-independent depressed lymphocyte responses to mitogens and altered measures of complement proteins, IgG levels, and macrophage functions. In studies by Bruce Rabin and colleagues, rats subjected to unavoidable foot shock showed decreased mitogen responses by blood lymphocytes (cortisol-dependent), decreased mitogen responses in splenic lymphocytes (sympathetically driven), and decreased splenic NK cell activity (dependent on opioid peptide activity). Some studies of avoidable

versus unavoidable foot shock have shown that escapability prevents the suppression of splenic NK cell activity seen with inescapable shock. This is consistent with "learned helplessness" studies that suggest that control over aversive stimuli can buffer stress-induced suppression of immune responsiveness. However, the stress and immune studies are very complex and suggest that many characteristics of the nature, severity, duration, and timing of the stressor, many other host characteristics and variables, and the nature of the immune measure or challenge can influence (1) whether there are alterations in immune reactivity, (2) if so, what directionality those alterations will take, and (3) whether there are disease consequences of such alterations.

Stressors, including psychological stressors, can alter disease progression and outcome, as noted earlier in the introductory section. Among the most careful studies of stress influences on disease outcome are those of John Sheridan and colleagues. They showed that restraint stress can suppress T-cell proliferation and cytokine production in response to influenza viral infection in mice, as well as to herpes infection. Reversal of these stress effects can be achieved only with dual blockade of both glucocorticoids (with RU486) and β-adrenergic receptors (with nadolol). Each of these agents, RU486 and nadolol, when given alone, reversed some of the stress-induced alterations in the mice in response to influenza infections, but it took both agents together to reverse all of the major effects of the stressor. Sheridan and colleagues also showed that injection of herpes virus into the tongue in rats resulted in a very high rate of encephalitis in stressed animals, with little or no such encephalitic response seen in unstressed animals with similar injections. These studies suggest that stressors that activate the HPA axis and sympathetic nervous system can alter both immune responsiveness and disease outcome following a challenge.

A study by Zorzet and colleagues investigated the effects of restraint stress on Lewis lung tumors in rats. These tumors can be treated successfully with cytoxan, eliminating the tumors and restoring health to the rats. Restraint stress during treatment with cytoxan (cyclophosphamide) eliminated the ability of the chemotherapy to destroy the Lewis lung tumor, leading to the death of the rats, as was the case in rats with these tumors who received no chemotherapeutic treatment. This provocative study suggests that a stressor may influence the ability of chemotherapy to destroy some cancers, in addition to possible direct effects on tumor proliferation and dissemination.

5. Hormonal and Neurotransmitter Mediators Link the Nervous System with the Immune System through Known Channels of Biological Signaling

Initially, studies reporting that conditioning or psychological stressors could influence immune reactivity or disease outcome were met with considerable skepticism because there were no well-characterized links between the nervous system and the immune system that could provide a mechanistic explanation for signaling to account for these influences. However, increasing numbers of studies, both *in vitro* and *in vivo*, have indicated that neurohormones and neurotransmitters are abundantly available for interaction with cells of the immune system and that these cells possess receptors that are responsive to these neurally derived signal molecules.

a. Classical Pituitary Hormones Profoundly Influence Immune Responsiveness Classical pituitary hormones can influence immune responsiveness both *in vitro* and *in vivo*, as will be discussed in detail later. Hormones shown to exert such effects include adrenal corticotropic hormone (ACTH), β-endorphin, growth hormone, prolactin, thyroid-stimulating hormone, and others. Hormones from the target glands that respond to classical pituitary hormones also can exert influences on lymphoid cells. These hormones include glucocorticoids (cortisol in humans, corticosterone in rats), thyroid hormone, and gonadal steroids. In addition, posterior pituitary hormones (oxytocin, vaopressin) and melatonin from the pineal gland act on some cells of the immune system.

b. Sympathetic Noradrenergic Nerve Fibers and Neuropeptidergic-Containing Nerve Fibers Directly Innervate Parenchymal Zones of All Lymphoid Organs Sympathetic postganglionic noradrenergic (NA) nerve fibers were described several decades ago as innervating vascular smooth muscle and other smooth muscle (e.g., capsular smooth muscle, splenic trabeculae) in lymphoid organs (thymus, spleen, lymph nodes), but they were thought to mainly control blood flow or capsular contraction. David Felten and colleagues found sympathetic NA nerve fibers extending into the parenchyma of primary lymphoid organs (bone marrow, thymus) and secondary lymphoid organs (spleen and lymph nodes), as well as into mucosal-associated lymphoid tissue of the gastrointestinal tract. The NA nerve terminals formed neuroeffector contacts immediately adjacent to T

lymphocytes, macrophages, granulocytes, reticulocytes, and other cell types. These nerve terminals release norepinephrine (NE) adjacent to these target cells that possess β- and α-adrenoceptors. Communication between the nerve terminals and the cells of the immune system include both the close appositions for direct synaptic-like contacts and more widespread diffusion for paracrine-like interactions. Numerous neuropeptidergic nerve fibers also extend into the parenchyma of primary and secondary lymphoid organs and form similar close appositions. These neuropeptide terminals contain neuropeptide Y (colocalized with NE in most sympathetic postganglionic terminals), substance P and calcitonin gene-related peptide (CGRP) (often, but not always, colocalized with each other), vasoactive intestinal peptide (VIP), somatostatin, enkephalin, and others. It appears that these nerve fibers are separate from each other, with the exception of the colocalized neurotransmitters mentioned earlier. We do not yet know the cells of origin for most of these neuropeptide-containing nerve fibers. Such sources include (1) sensory ganglia, (2) autonomic sympathetic ganglia, and (3) enteric neurons whose processes traverse autonomic ganglia.

c. Some Cells of the Immune System Can Produce Neuromediators Nonneural cells, such as thymic nurse cells, can produce neuropeptides, such as oxytocin and vasopressin, that can be released as mediators for interaction with other cells in the thymus. Ed Blalock, Eric Smith, and colleagues found that some macrophages and lymphocytes can be induced to produce classical pituitary hormones such as the proopiomelanocortin (POMC) products ACTH and β-endorphin, thyroid-stimulating hormone, and others. Some viruses can act on receptors that induce the production of a neurohormone. In addition, some releasing factors, such as corticotropin-releasing factor (CRF), can induce the gene for producing POMC, resulting in the production of ACTH in a manner similar to that occurring in the anterior pituitary gland. In addition, some macrophage-like cells can directly produce neurohormones such as CRF itself. Thus, neuropeptides and neurohormones can be produced by neurons or cells of the immune system.

6. Central Nervous System (CNS) Lesions and Intracerebroventricular Infusions Can Alter Immune Responsiveness

Stereotaxic lesions, precisely placed sites of tissue destruction, in specific portions of the CNS induce both structural and functional changes in the immune system, some of which are transient and some long-lasting. Most of the sites in which lesions induce such changes are in the hypothalamus, limbic forebrain, or brain stem structures associated with the hypothalamus and limbic forebrain. For example, lesions in the anterior hypothalamus resulted in thymic involution and a transient decrease in mitogen-stimulated (conA) proliferation of T-cells. Lesions in the lateral septal area resulted in long-lasting decreases in IgG, IgA, and IgM production, whereas lesions in the hippocampus produced elevated IgM and IgG antibody production. However, care must be taken in interpreting these findings. Lesions show only what the remainder of the nervous system can do in the absence of the lesioned structure, not necessarily what the lesioned structure normally did. These lesion studies do provide sites for further probing for a better understanding of neurochemical mechanisms of regulating neuroendocrine and autonomic outflow from the CNS.

A further approach to studying CNS regions that may play a role in modulating immunological reactivity is intracerebroventricular (icv) infusion of immunologic agents influencing neural outflow to the immune system. For example, infusion of interleukin-1β (IL-1β) icv in femtomolar concentrations resulted in diminished splenic immune responses that were abrogated by the administration of ganglionic blocking agents, indicating that the IL-1β activated central circuits that activated outflow of the sympathetic nervous system. Similarly, icv infusion of the HIV envelope protein gp120 resulted in diminished peripheral immune responses and elevated plasma corticosteroid levels, again indicating a central stimulation of outflow neural mediators that can alter immune responses. The sites and receptors by which these infusions interacted with CNS neurons are not yet fully understood.

7. Immunologically Derived Mediators Can Activate Portions of Both the CNS and Peripheral Nervous System (PNS)

a. Cytokine Induction of Fever, Sleep, and Illness Behavior Cytokines, especially the inflammatory cytokines IL-1β, IL-6, and TNF-α, can act directly or indirectly on CNS neurons to induce fever, slow-wave sleep, lethargy, loss of appetite, and other characteristics of illness behavior. Both icv and peripheral intravenous (iv) injection of these cytokines can induce these behaviors, although the icv route requires lower

concentrations. IL-1β is the most studied and most common cytokine thought to mediate such responses. There are several routes by which such actions may occur: (1) IL-1β may enter directly into the cerebrospinal fluid (CSF) by crossing the blood–brain barrier or by crossing at a circumventricular organ such as the organum vasculosum of the lamina terminalis (OVLT); (2) IL-1β may act on cells at the OVLT to release additional mediators such as prostaglandin E2 (PGE2), which in turn activates CNS neurons; (3) IL-1β may activate small-molecule mediators such as nitric oxide (NO) from the vascular endothelium or cells of the CNS, which in turn activate CNS neurons; (4) IL-1β may act on paraneurons associated with the vagus nerve, thus activating afferent pathways to the CNS that induce illness behavior; or (5) IL-1β may activate other sensory neurons or peripheral mediators that in turn signal the CNS. The interleukins IL-1β and IL-2 may act on presynaptic receptors on sympathetic NA nerve terminals to regulate the release of NE directly, thereby contributing to the regulation of neural–immune signaling at a local level.

b. Cytokine Activation of the Hypothalamo–Pituitary–Adrenal (HPA) Axis and the Sympathetic Nervous System In addition to inducing illness behavior, IL-1β and other inflammatory cytokines also activate the HPA axis by elevating the production of CRF, thereby further stimulating ACTH and glucocorticoid production. Central noradrenergic (NA) input to the paraventricular nucleus, deriving mainly from the A1 and A2 cell groups in the medulla, is necessary for the IL-1β- induced elevation of CRF; ablation of this NA pathway in the brain stem (the ventral NA bundle) abrogates the IL-1β-induced CRF elevation. In addition, central IL-1β activates the sympathetic nervous system, also by a CRF-mediated mechanism through the hypothalamus, which in turn interacts with organs of the immune system. These cytokines that can activate the HPA axis and sympathetic nervous system, therefore, act as stressors, activating the same axes that psychological and physical stressors activate. It should be noted that not all immune responses activate these stress axes. Viral infections, inflammatory mediators such as IL-1β and IL-6, and some stimuli such as lipopolysaccharide (LPS) are particularly potent activators. However, immune responses to some antigens do not activate these axes. It is of some concern that viral challenges, bacterial challenges, and inflammatory mediators can activate the HPA and sympathetic stress axes; these are potentially damaging reactions and, if activated in debilitated elderly patients in a setting where nosocomial organisms are abundant, such as a hospital, could have an adverse impact on morbidity and mortality.

c. Cytokine Regulation of CNS Inflammation, Reaction to Injury, and Regeneration and Plasticity Cytokines, especially IL-1β, IL-6, and TNF-α, have been noted to accumulate at the site of central injuries and inflammatory responses. The presence of IL-1β and other inflammatory cytokines also has been noted at an early phase of plaque formation in the brains of patients with Alzheimer's disease. It is not yet known the extent to which these cytokines can enhance or inhibit repair, regenerative responses, and neuronal plasticity. A complicating factor is the presence of IL-1β from blood, IL-1β from invading macrophages, IL-1β from resident microglial cells, endothelial cells, and possibly astrocytes, and the production of IL-1β by neurons themselves. It is likely that a cascade of interactions exists among inflammatory cytokines, other inflammatory mediators, neurotransmitters, acute phase proteins, and other proteins accumulating at inflammatory sites; these many mediators combine to influence the balance between degeneration and cell death on the one hand and regeneration and plasticity on the other.

II. OVERVIEW OF NEURAL COMMUNICATION CHANNELS TO THE IMMUNE SYSTEM

The central nervous system has several routes of communication available for molecular signaling directed toward cells of the immune system. The most conspicuous routes are connections from the CNS through the anterior and posterior pituitary and direct neural connections from the autonomic and sensory components of the peripheral nervous system. In addition, as noted earlier, some cells of the immune system can synthesize and release neural mediators, whereas others can take up such mediators from the local microenvironment.

A. Hypothalamic–Anterior Pituitary–Target Organ Axes

1. Direct Anterior Pituitary Hormone Actions on Immunocytes

Several of the anterior pituitary hormones have direct actions on cells of the immune system, particularly growth hormone (GH), prolactin, adrenocorticotropic

hormone (ACTH), and β-endorphin. The other anterior pituitary hormones have at least some effects on specific subsets of immunocytes. These hormones are released from the anterior pituitary by releasing factors, or inhibited from release by inhibitory factors, produced by neuronal cell groups in the hypothalamus and other CNS regions. The releasing factors and inhibitory factors are secreted from nerve terminals of these cell groups that end on the hypophyseal–portal vessels, a private vascular communication channel linking the hypothalamus with the anterior pituitary gland that is particularly suited for sending very high concentrations of these releasing and inhibitory factors to interact with pituicytes. The neuronal cell groups producing the releasing and inhibitory factors generally are associated with the hypothalamus, limbic system, and brain stem autonomic centers and can be acted upon by appropriate forebrain and brain stem circuits that participate in behavioral control of the periphery.

2. Target Organ Hormone Actions on Immunocytes

The anterior pituitary hormones are released into the peripheral blood, where they can act upon any cells that bear appropriate receptors for them. Whereas some of the anterior pituitary hormones act directly on immunocytes, they also act on their target organs to further release endocrine hormones into the blood. The most important of these hormones that act on cells of the immune system include the glucocorticoids, dehydroepiandrosterone, the gonadal steroids, and thyroid hormones.

B. Posterior Pituitary Hormones

The posterior pituitary hormones oxytocin and vasopressin are synthesized mainly in cell bodies in the supraoptic (SON) and paraventricular (PVN) nuclei of the hypothalamus and are released from nerve terminals of those cell groups that end in the posterior pituitary. Oxytocin and vasopressin then are released into the general circulation.

C. Autonomic Outflow to Organs of the Immune System

1. Sympathetic Noradrenergic Connections Are Abundant

Extensive sympathetic neural connections form direct links between the brain and cells of the immune system. The preganglionic sympathetic neurons are found in the intermediolateral cell column of the thoracic and lumbar spinal cord and extend their axons out of the CNS to terminate on ganglion cells in the sympathetic chain and collateral (prevertebral) sympathetic ganglia. These ganglion cells generally use norepinephrine (NE) as their principal neurotransmitter and often have colocalized neuropeptides such as neuropeptide Y (NPY). The postganglionic noradrenergic (NA) nerve fibers extend into primary lymphoid organs (bone marrow, thymus), secondary lymphoid organs (spleen, lymph nodes), mucosal-associated lymphoid tissue (MALT) of the gastrointestinal tract and the airway, and lymphoid cells of the skin. In the bone marrow, thymus, spleen, and lymph nodes, the sympathetic nerve fibers enter the organ with the vasculature, extend along the vascular smooth muscle and other trabecular, capsular, or septal channels, and extend directly into the parenchyma of these lymphoid organs. The sympathetic NA nerve terminals form direct appositions or neuroeffector junctions with lymphocytes, macrophages, granulocytes, reticulocytes, and other cell types of the immune system. These contacts are as close as 6 nm (a gap junction is 2 nm and a classical CNS synapse is 20 nm) and form direct links for the sympathetic modulation of lymphoid cells bearing appropriate adrenoceptors responding to the released neurotransmitter. Some of these contacts have been mapped with anatomical tracing techniques; ablation and stimulation studies have shown that sympathetic contacts play an important role in translating CNS activity into immunomodulating signals. Sympathetic postganglionic fibers from the superior cervical ganglion also innervate the pineal gland and can regulate the synthesis and release of another immunomodulating hormone, melatonin.

2. Parasympathetic Cholinergic Connections Are Sparse or Not Present in Many Organs of the Immune System and Regionally Localized in Some Sites

Initially, it was assumed that sympathetic and parasympathetic nerve fibers constituted a balanced yin–yang interaction with the immune system, providing antagonistic functions similar to those described for the heart. However, with the possible exception of sparse parasympathetic fibers to the thymus, parasympathetic cholinergic nerve fibers have not been found extending into the primary and secondary lymphoid organs. There is no evidence for the

synthesis of choline acetyltransferase or high-affinity choline uptake in these organs. Thus, it appears that the main autonomic innervation derives from the sympathetic nervous system. In some of the MALT areas of the gastrointestinal tract and airway, cells of the immune system might be associated with the cholinergic parasympathetic nerve fibers supplying those systems.

3. Many Neuropeptidergic Nerve Fiber Connections to Lymphoid Organs Appear To Derive from Autonomic Ganglia

Abundant neuropeptide-containing nerve fibers also are present in the bone marrow, thymus, spleen, lymph nodes, and MALT. These fibers include substance P (SP), calcitonin gene-related peptide (CGRP), vasoactive intestinal peptide (VIP), somatostatin, neuropeptide Y (NPY), and others. The full "alphabet" of these neuropeptides has not yet been worked out; very careful studies varying the conditions of immunohistochemical staining must be undertaken to reveal the very fine neuropeptide-containing fibers that are present in lymphoid organs. Similar to the NA sympathetic nerve fibers, the neuropeptide-containing nerve terminals end in direct proximity to lymphocytes, macrophages, and other cells of the immune system. At first, it was assumed that these neuropeptide-containing nerve fibers derived from primary sensory ganglia because, for example, substance P is an important neuropeptide in nociceptive C fibers of the "pain" system. Several tracing studies from primary and secondary lymphoid organs have shown that the retrograde tracers show up principally in sympathetic ganglion cells and not in the primary sensory neurons in dorsal root ganglia or cranial nerve sensory ganglia. For example, the superior cervical ganglion and upper sympathetic chain ganglia supply NA fibers to the thymus. The superior mesenteric–coeliac ganglion supplies NA fibers to the spleen. Ganglia of the appropriate sympathetic plexuses supply the lymph nodes. Whereas some of these labeled cell bodies stained positively for tyrosine hydroxylase (TH), the rate-limiting enzyme for NE synthesis in sympathetic postganglionic neurons, others cell bodies stained positively for neuropeptides. It presently appears that a majority of the neuropeptide-containing nerve fibers derive from sympathetic postanglionic neurons. Until more detailed and careful tracing studies are carried out, it remains possible that a few cells of the primary sensory ganglia, and even enteric neurons whose axons pass through peripheral ganglia, might supply neuropeptide-containing fibers to organs of the immune system.

D. Primary Sensory Mechanisms of Neural–Immune Signaling

1. Primary Sensory Innervation of Lymphoid Organs Appears Sparse

Although only a few tracing studies have identified primary sensory ganglion cells of origin for neuropeptide-containing nerve fibers that distribute into the primary and secondary lymphoid organs, it still is possible that sparse numbers of cell bodies can give rise to abundant arborizations that have widespread functional interactions (e.g., a PNS equivalent of the locus ceruleus' widespread influence in the CNS). We do not yet have enough detailed tracing data to make conclusive statements about the possible role of primary afferents as "sensory" signaling devices for potentially passing immunologic information to the brain. However, it is possible for primary sensory signaling to the CNS to play an important role indirectly, as shown by Robert Danzer, Keith Kelley, Linda Watkins, Steve Maier, and others, who have shown that cytokines such as IL-1β in the periphery can stimulate the release of mediators from paraneurons, which in turn activate vagal afferents. These vagal afferents communicate with the nucleus solitarius and other brain stem nuclei that feed information into the brain stem–hypothalamic–limbic network that is active in neural–immune communication.

2. Substance P Nociceptive Fibers and Inflammatory Responses Link Primary Afferents to the Immune System

Substance P nerve fibers link a wide range of tissues with the dorsal horn of the spinal cord or the descending (spinal) trigeminal nucleus via primary sensory ganglia. These systems often contain colocalized neuropeptides such as CGRP. A remarkable property of these substance P fiber systems is their ability to release their neurotransmitter from both the proximal end (into the CNS site of termination) and the distal end (into the peripheral site of innervation). This has been shown definitively in the joints and gastrointestinal tract and suggests that inflammatory lesions that exacerbate nociceptive systems can release substance P into the site of inflammation; this short loop constitutes a positive feedback link and further

reinforces inflammatory effects, such as extravasation, edema, and attraction of additional pro-inflammatory cells. The extent to which these interactions form channels for the communication of information to the CNS is not known, other than nociception and subsequent perceptions of pain.

E. Neurohormones and Neurotransmitters Are Sometimes Synthesized and Sometimes Taken Up by Cells of the Immune System: Paracrine and Autocrine Signaling

A remarkable finding in the mid-1980s from the laboratories of Ed Blalock and Eric Smith was the discovery that lymphocytes and macrophages can synthesize some neuropeptides previously thought to be present only in CNS-associated systems. For example, New Castle disease virus was able to stimulate the production of mRNA for POMC, resulting in ACTH and β-endorphin synthesis in some lymphocytes. Corticotropin-releasing factor (CRF) also can stimulate POMC mRNA in some cells of the immune system. Subsequent studies have shown that a host of anterior pituitary hormones can be synthesized under certain conditions by cells of the immune system. It is most likely that these hormones, if released, would act in a paracrine fashion rather than an endocrine fashion. Some supporting cells, such as epithelial or nurse cells in the thymus, also can produce neuropeptides such as oxytocin and vasopressin. Work suggests that some cells of the immune system can take up norepinephrine or other neurotransmitters from the microenvironment. The extent to which these locally synthesized neuropeptides, or neurotransmitters that are taken up, can be released and can influence immune reactivity directly as paracrine secretions is not yet known.

F. Some Specific CNS Pathways and Sites Can Regulate Neuroendocrine and Autonomic Outflow to the Immune System: Neural Substrate Linking Behavior and Immune Responsiveness

The consensus from a host of "stressor" studies in humans and experimental animals points to the combined interactions of the HPA axis and the sympathetic nervous system as the two most prominent actors in interpreting CNS activity for peripheral immunomodulating actions. Perhaps this is merely a bias of studies that have been most intensely oriented toward these two conspicuous and important axes, although the most definitive studies, such as those from John Sheridan's laboratory, point to a critically important role for the sympathetic and HPA axes acting in concert. This dual working model focuses attention on the hypothalamic mechanisms that regulate the outflow of the HPA and sympathetic axes. Because central CRF is involved in this process, the paraventricular nucleus of the hypothalamus has been the site of greatest interest. It appears that central noradrenergic pathways (as well as serotonergic and cholinergic pathways) play an important role in regulating CRF in neurons of the PVN. Indeed, it appears that the CRF-elevating action of IL-1β depends on the intactness of the central noradrenergic input to the PVN. Other hypothalamic sites also are involved in regulation of the anterior pituitary and the autonomic nervous system, such as anterior–preoptic and posterior hypothalamic areas. Furthermore, lesion studies suggest that the amygdala, hippocampal formation, septal nuclei, and central autonomic brain stem nuclei can influence the "set point" of immunologic reactivity, probably through connections regulating the hypothalamic regions involved in autonomic and neuroendocrine outflow. Therefore, we now have the appropriate target sites to investigate the specific central pathways and their neurotransmitters in the central regulation of outflow influencing the immune system. We do not yet know the details of which of these specific pathways are involved in stress-related alterations of immune response or behavioral conditioning of immune responses.

III. INFLUENCES OF NEUROENDOCRINE HORMONES ON THE IMMUNE SYSTEM

A. The Hypothalamo–Pituitary–Adrenal (HPA) Axis

1. Pharmacologic Doses of Glucocorticoids Have Long Been Used to Reduce Inflammation, Suppress Both Innate and Acquired Immune Responses, and Suppress Autoimmune Diseases

Early studies showed that corticosteroids can exert profound anti-inflammatory and immunosuppressive effects; these effects generally are brought about by supraphysiological doses of glucocorticoids. These steroids can result in lymphoid involution (the thymus

and secondary lymphoid organs), lymphopenia, decreased antibody formation, prolongation of tolerance, suppression of allograft rejection, inhibition of delayed-type hypersensitivity reactions, inhibition of cytotoxic T lymphocyte activity, reduced production of helper T-cells, and decreased NK cell activity. Glucocorticoids also can inhibit numerous macrophage functions such as cytokine production, phagocytosis, intracellular killing, chemotaxic, antigen presentation, and production of inflammatory mediators. Glucocorticoids can alter the trafficking of specific cells, for example, from blood to bone marrow and away from secondary lymphoid organs. Glucocorticoids also are administered to suppress autoimmune diseases such as lupus erythematosis, rheumatoid arthritis, and others. The actions of glucocorticoids are mediated through soluble cytosolic and nucleus receptors, resulting in the inhibition of transcription and activation of several genes.

2. Physiological Concentrations of Glucorticoids Have an Immunoregulatory Role That Includes Both Inhibition and Enhancement of Some Responses

Whereas a majority of the effects of glucocorticoids appear to be inhibitory in nature, studies of physiological actions of glucocorticoids suggest that they may influence the commitment of T helper lymphocytes toward Th2 cytokine production. Glucocorticoids acting on activated T lymphocytes can result in decreased IL-2 production but increased IL-4 production, suggesting that physiological glucorticoids, under some circumstances, can favor humoral immunity and suppress cell-mediated immunity. In addition, other lymphoid cells may influence the actions of glucocorticoids. Activated macrophages can prevent glucocorticoid-mediated inhibition of T helper cell function. Some *in vitro* studies have shown that glucocorticoids in physiological concentrations can enhance B-cell differentiation.

3. The HPA Axis Represents Only One of Many CNS Outflow Systems That Can Alter Immune Responsiveness When Activated by Stressors

When the phenomena of stress-induced and conditioned immunosuppression were first reported, it was assumed that this reflected a centrally activated HPA axis that led to immunosuppression by glucocorticoids. However, conditioned immunosuppression can occur in adrenalectomized animals, and only some components of stress-induced immunosuppression (conA proliferation of peripheral blood T-cells) but not others (splenic conA proliferation of T-cells, splenic NK cell activity) appear to be glucocorticoid-mediated. Therefore, the sympathetic neural connections, neuropeptide neural connections, and other hormonal interactions, in addition to the HPA axis, must be taken into account to fully catalogue the mediators responsible for behaviorally induced changes in the immune system. It is this author's opinion that the task of defining what changes in hormonal and neural activity occur during specific behavioral conditions is essential to understanding neural–immune interactions and needs vastly greater research efforts.

4. The Reactivity of the HPA Axis May Be an Important Determinant in the Susceptibility of an Organism to Autoimmune Reactivity

Although it has long been known that exogenous glucocorticoids, in pharmacologic doses, can suppress some autoimmune diseases, the possible physiological importance of the HPA axis in autoimmune disease has only more recently come to light from the pioneering work of Esther Sternberg and colleagues. She studied a genetic animal model, the LEW/N rat, that is susceptible to a wide range of autoimmune diseases, including adjuvant-induced arthritis, experimental allergic encephalomyelitis, thyroiditis, and other inflammatory diseases. She found diminished CRF responses in female LEW/N rats and proposed that these animals cannot boost glucocorticoid levels appropriately during a massive challenge such as that encountered in the induction of an autoimmune disease, resulting in excessive proliferation of the clones of cells that should not be proliferating, including autoreactive (anti-self) lymphocytes. She demonstrated that, when glucocorticoids were blocked with RU486 in Fischer 344 rats, a genetic counterpart of the LEW/N rats that are *not* susceptible to autoimmune disease, the Fischer 344 rats became fully susceptible to induced autoimmunity. Thus, the HPA axis may play an important role in signal-to-noise ratio, as proposed more than two decades ago by Hugo Besedovsky, and in the loss of counterregulatory suppression of anti-self clones that can lead to autoimmune disease. However, the sympathetic nerve fibers innervating secondary lymphoid organs also play an important regulatory role in neural–immune

influences on susceptibility to autoimmune diseases, as shown by Dianne Lorton, Denise Bellinger, David Felten, and colleagues, discussed later.

B. Growth Hormone (GH) and Prolactin

1. GH and Prolactin Are Essential for the Competent Development of the Immune System

Deficiencies in growth hormone following hypophysectomy result in diminished cellularity of the bone marrow and thymus and diminished T-cell function, antibody responses, and NK cell activity. These deficiencies can be reversed to a large extent by the administration of GH. Exogenous GH also can enhance peritoneal macrophage superanion production in animals. Prolactin exerts mainly an enhancing effect on immune functions. Inhibition of prolactin secretion from the anterior pituitary results in suppressed antibody production, diminished cell-mediated immune responses, and increased susceptibility to infections such as *Listeria monocytogenes*. These inhibited functions can be reversed by the administration of exogenous prolactin or by the inhibition of dopamine (prolactin inhibitory factor) with antagonists. Stressors that result in the release of prolactin from the anterior pituitary provide a counteracting signal that can dampen many of the immunosuppressive effects of glucocorticoids, but the release of prolactin becomes refractory with repeated application of a stressor. High-affinity receptors for GH have been reported on thymocytes, lymphocytes (resting or activated), and monocytes. Prolactin receptors have been reported on B and T lymphocytes and on macrophages.

2. GH and Prolactin Are Important Regulators of Thymic Function and Can Restore Thymic Cellularity and Responses in Old Animals

Keith Kelley and colleagues studied old rats with involuted thymuses and diminished T-cell mitogen-induced proliferation and IL-2 production. They grafted a tumor cell line that produced both GH and prolactin in these animals and found a restoration of these parameters. Ovine GH itself did not restore thymic cellularity or IL-2 production, but it did enhance T-cell mitogen-induced proliferation. GH administered to normal animals can enhance alloantigen-specific cytotoxic T-cell activity.

C. Opioid Peptides

1. Opioid Peptide Neuronal Systems in the PNS Directly Innervate Lymphoid Organs and in the CNS Influence the Outflow of Other Hormones and Neurotransmitters

ACTH itself exerts actions on cells of the immune system, such as the reduction of mouse spleen cell antibody responses and *in vitro* IFN-γ secretion. ACTH also acts in concert with IL-5 to stimulate late stages of B-cell activation and has been reported to enhance NK cell activity *in vivo* in pigs. Both high- and low-affinity ACTH receptors have been reported on T and B lymphocytes, with greater numbers on B-cells. Additional effects of ACTH are exerted through the stimulation of glucocorticoid production, and some immunocytes can directly synthesize ACTH.

α-MSH (ACTH 1–13) is a potent inhibitor of migration in response to IL-1β or other chemotactic agents, prevents spontaneous neutrophil activation, and counteracts many of the effects of IL-1β both in the periphery and in the CNS.

2. Several Opioid Receptor Subsets Are Present on Cells of the Immune System and Mediate Complex and Diverse Responses to Opioid Peptides

The opioid peptides are low-molecular-weight peptides in three families: the endorphins, the enkephalins, and the dynorphins. They are produced in the pituitary gland, the brain, the adrenal gland, peripheral nerves, and even in cells of the immune system. These hormones are cleaved from prohormones synthesized in the cell body; differential splicing can result in a range of peptide end products, depending upon local influences on the cell. The opioid peptides are a heterogeneous group of agents whose metabolism may result in additional biologically active molecules that have further complex functional roles. Similarly, the opioid receptor expression on cells of the immune system is complex. The best evidence suggests that classical μ and κ receptors are found on mouse spleen cells. A functional role for the opioid peptides is contradictory in some of the literature and not straightforward. For example, β-endorphin exerts opioid-mediated effects of diminishing mouse antibody production and inhibiting human lymphocyte chemotactic factor. β-Endorphin exerts non-opioid-mediated effects of enhancing mouse conA T-cell

proliferation, inhibiting human PHA T-cell proliferation, and blocking PGE1-induced suppression of rat lymphocyte proliferation.

Although α-, β-, and γ-endorphin and met-enkaphalin have identical amino termini, they can exert markedly different biological effects. For example, in one study α-endorphin and met-enkaphalin, but not β- or γ-endorphin, decreased antibody production by mouse spleen cells. Enkephalins may play a role in inflammatory responses through the stimulation of neutrophil superoxide anion production and neutrophil migration.

D. Thyroid and Gonadal Hormonal Axes and Other Hormones

Thyroid-stimulating hormone (TSH) in physiological concentrations enhances both T-dependent and T-independent antibody responses of mouse spleen cells. Thyroidectomy of neonatal or adult rats or hypothyroidism in chickens resulted in diminished lymphoid organ weight, decreased circulating lymphocytes, decreased antibody responses, and decreased T- and B-cell mitogen-induced proliferation. These alterations were restored with the administration of T3 or T4 to the thyroidectomized animals. T3 or T4 in euthyroid animals enhanced antibody responses and mitogen-induced proliferation. Administration of T4 to old mice restored NK cell function, but had no effect in young mice.

Gonadal hormones also exert an effect on the immune system. There is clearly sexual dimorphism in the immune system. Females have higher serum immunoglobulin levels of IgG, IgM, and IgA in several species and show more prolonged and robust antibody responses to both T-dependent and T-independent antigen challenges. Females also are subject to a considerably higher incidence of autoimmune diseases, such as systemic lupus erythematosis, than males. Gonadectomy induces hyperplasia of the thymus and other lymphoid organs in both males and females; 5-α-dihydrotestosterone restored thymic weight in males, and prior thymectomy prevented hyperplasia in other lymphoid organs. Manipulation of testosterone or estrogen can alter the progression or onset of autoimmune disease in experimental animal models. Estrogen induces extrathymic T-cell differentiation. Estrogen also enhanced polyclonal Ig, antigen-specific proliferation, and autoantibody production. Testosterone inhibited polyclonal Ig production.

IV. NERVE FIBER CONNECTIONS WITH THE IMMUNE SYSTEM

A. Sympathetic Noradrenergic Innervation of Lymphoid Organs

1. Primary Lymphoid Organs: Bone Marrow and Thymus

a. Noradrenergic Nerve Fibers Distribute into the Parenchyma and along the Vasculature NA postganglionic sympathetic nerve fibers enter the bone marrow with the nutrient arteries, arborize along the vasculature, and distribute among the stem cells between the sinuses. These nerve fibers form close appositions with cells in the parenchyma. In the thymus, NA nerve fibers, derived mainly from neurons in the superior cervical ganglion, enter with the vasculature, distribute along the capsule and in subcapsular and septal regions, and also arborize within the thymic cortex. Some NA nerve fibers also are found in the medulla, but a majority of the sympathetic NA innervation is confined to the cortex. Both bone marrow stem cells and thymocytes possess functional β-adrenoceptors. In these primary lymphoid organs, NA nerve terminals appear to release NE into the microenvironment, where it acts in a paracrine fashion. The significance of close appositions of nerve terminals with parenchymal cells is not yet clear.

b. Norepinephrine Modulates Stem Cell Proliferation, Differentiation, Trafficking, and Adhesion In the bone marrow, NA denervation (chemical sympathectomy) resulted in an increase in the number of peripheral blood leukocytes after syngeneic bone marrow transfer. Prazocin, an α-adrenergic antagonist, mimicked this effect. In a study by David Wu, David Felten, and colleagues using a 3-dimensional bone marrow culture system built on a microsphere scaffolding, the β-agonist isoproterenol was found to enhance hemopoiesis up to 3-fold during 5 days of culture. Isoproterenol stimulated granulopoiesis as effectively as granulocyte colony stimulating factor (G-CSF) and synergized the effects of G-CSF Thus, α- and β-adrenoceptors may exert differential effects on stem cell populations. In the thymus, NE from the sympathetic nerves inhibits thymocyte proliferation and enhances the expression of markers indicating differention. NA denervation in mice results in enhanced thymocyte proliferation, but only in older animals in which the thymus has involuted.

2. Secondary Lymphoid Organs: Lymph Nodes, Spleen and MALT

a. Noradrenergic Nerve Fibers Are Compartmentalized to T-Cell Zones and Macrophage Zones and Do Not Directly Innervate B-Cell Zones Sympathetic NA nerve fibers enter the spleen with the splenic artery, distribute with the capsular–trabecular system, arborize extensively along the central artery of the white pulp, and also branch extensively within the parenchyma of the white pulp. NA nerve fibers that innervate the parenchyma of the white pulp are the first to appear in development; they are present in the parenchyma before smooth muscle cells are present along the vasculature and nerve fibers are present along the vascular compartment. In the parenchyma, the NA fibers distribute among T lymphocytes [both T helper–inducer cells (CD4) and T cytotoxic–suppressor cells (CD8)] in the periarteriolar lymphatic sheath, along the macrophages of the marginal sinus, and within the marginal zone. NA nerve fibers distribute along the outside edge of the B-cell follicles, but do not directly contact B lymphocytes in adults, although B-cells have high numbers of β-adrenoceptors on their surfaces. Electron microscopic studies have shown direct appositions between NA nerve terminals and lymphocytes, macrophages, granulocytes, and reticuloendothelial cells. Extensive evidence indicates that NE is released from these terminals and is available for interaction with cells of the immune system that bear adrenoceptors on their surfaces.

Lymph nodes also are innervated by postganglionic NA nerve fibers. These fibers enter the nodes along the hilar vasculature and along the capsular surface, extend through the medullary cords, and arborize extensively among the T-cell compartments (among both CD4 and CD8 cells) and macrophage compartments, but not the B lymphocyte compartments. However, NE released from the nerve terminals probably has access to B-cells through diffusion. Perhaps the very close apposition (as close as 6 nm) between NA nerve terminals and T lymphocytes or macrophages serves a role to allow cytokines, secreted by these lymphoid cells, to regulate NE release from those terminals. Both IL-1β and IL-2 have been reported to stimulate the release of NE directly from nerve terminals in the spleen.

In gut-associated lymphoid tissue, NA nerve fibers enter with the vasculature, travel along the smooth muscle layers, extend inward through the T lymphocyte zones, and arborize among the lamina propria, adjacent to the immunoglobulin-secreting plasma cells, but they avoid the follicles.

b. Norepinephrine Modulates Innate and Acquired Immune Responses, Including Both Humoral and Cell-Mediated Immunity and Influences Trafficking and Cell Adhesion The role of NE in the regulation of immune responses is complex. β-Adrenoceptors have been identified on many subsets of T and B lymphocytes and macrophages in humans, rats, and mice. α-Adrenoceptors have been reported on activated mouse macrophages and on human and guinea pig T lymphocytes. Infusion of epinephrine or NE into the blood increases both circulating lymphocytes and NK cells transiently and may account for such transient elevations with exercise. There appear to be both α- and β-mediated catecholamine effects on cell migration. NE generally appears to inhibit effector cell functions in cells already activated to this stage (e.g., antibody-secreting plasma cells and cytotoxic T lymphocytes), particularly at peripheral sites where they are recruited (e.g., in the lung in an influenza infection). However, in the early phases of an immune response, NE and other catecholamines can enhance both antibody and cytotoxic T-cell responses. Denervation of NA nerve fibers from the spleen and lymph nodes results in diminished cell-mediated immune responses, has a variable effect on humoral immunity (depending on the predominant T-cell orientation), and exacerbates the severity of induced autoimmunity in autoimmune-prone strains of rodents. Thus, the presence of NE early in an immune response may enhance immune responsiveness, whereas its presence at the effector cell stage may inhibit the ability of the effector cells to act. Also, the collective result of NE influences on the multiple interacting cells in a specific immune response may be complex to interpret. In the inductive phase of an immune response, NE may favor Th1 cytokine production by T helper cells, promoting cell-mediated immunity, consistent with evidence for β-adrenoceptors on Th1 but not Th2 cells. Several studies have suggested that catecholamines may inhibit NK cell activity.

3. Peripheral Sites of Inflammation and Immune Reactivity

a. Noradrenergic Nerve Fibers Are Present at Many Sites Where Inflammatory Responses Occur (e.g., Joints) NA sympathetic nerve fibers are present in many peripheral sites where inflammatory responses

may take place, including joints, the gut, and the skin. These fibers distribute with vascular smooth muscle and other smooth muscle compartments and also innervate secretory glands and metabolic regions such as brown fat and liver.

b. Norepinephrine Can Directly Mediate Inflammatory Responses That Occur at Target Organs of the Inflammation At sites of inflamed joints in rheumatoid arthritis, blockade of NE or ablation of NA nerves resulted in a reduction in inflammation, as did blockade of substance P or destruction of the substance-P-containing nerve fibers; these fibers are nociceptive fibers mediating sensations interpreted by the brain as pain. However, selective destruction of NA innervation of draining lymph nodes prior to immunization in the induction of adjuvant-induced arthritis in LEW/N rats led to a severe exacerbation in the severity, and an earlier onset, of arthritis. This suggests that NA fibers to a joint may play a proinflammatory role, whereas NA innervation to the draining lymph node may dampen the vigor of autoimmune reactivity or processing.

B. Neuropeptide Innervation and Influences on Inflammation and Immune Reactivity

1. A Wide Range of Neuropeptide-Containing Nerve Fibers Innervate Both Primary and Secondary Lymphoid Organs

a. Neuropeptide Y (NPY) Is Often Colocalized with Norepinephrine in Sympathetic Postganglionic Nerve Fibers Although NPY receptors have been reported on some subsets of lymphoid cells, we do not yet have a clear understanding of the role of NPY in immunomodulation. It appears that NPY is extensively colocalized with NE in sympathetic nerve terminals, although some NPY-positive nerve fibers are found that do not colocalize norepinephrine. NPY also is present abundantly in platelets in the spleen. Irwin and colleagues have found that NPY measurement in peripheral blood is a more dependable estimate of sympathetic activity than NE measurement.

b. Substance P (SP) and Calcitonin-G-Related Peptide (CGRP) Are Often Colocalized in Separate Nonnoradrenergic Nerve Fibers, but Can Occur Individually in Nerve Fibers in Lymphoid Organs and Skin Substance P nerve fibers are found in spleen, lymph nodes, thymus, and bone marrow. They distribute with the vasculature and in the parenchyma and appear to be completely separate from the sympathetic NA nerve fibers. CGRP nerve fibers show a distribution similar to that of SP fibers; CGRP is usually, but not exclusively, colocalized with SP. In some sites, such as the skin, CGRP nerve fibers innervate a specific type of antigen-presenting cell, the Langerhans cells, which are markedly inhibited by this neuropeptide. Receptors for SP on lymphocytes have been cloned, and SP receptors have been identified on a human B-cell line, human T-cells, and mouse T- and B-cells. The substance P receptors (and other receptors for tachykinins) differ on nerve cells and lymphocytes; there is greater overlap and cross-talk between substance P and substance K receptors on lymphocytes than on neural tissue.

Functionally, it is clear that SP is a potent regulator of inflammatory responses. SP is proinflammatory, can induce extravasation and edema, can attract additional inflammatory cells to a site, and can stimulate other inflammatory mediators, such as IL-1β, tumor necrosis factor-α, and IL-6, from unstimulated peripheral blood monocytes. SP in a joint can exacerbate rheumatoid arthritis, whereas ablation of SP nerves or blockade of SP can have an anti-inflammatory effect. SP exerts a strong enhancing effect on a wide range of immune responses, including increased polyclonal IgA and IgM production by lymphocytes in Peyer's patch and mesenteric lymph nodes, increased T-cell proliferation and IL-2 production, increased antibody responses *in vivo*, and increased proliferation of unstimulated lymphocytes. SP nerve fibers in draining lymph nodes also contribute to the severity of adjuvant-induced arthritis in LEW/N rats. Many of these enhancing effects can be countered by somatostatin, but somatostatin nerve fibers have not yet been identified or mapped carefully in many sites where SP–CGRP nerve fibers reside. A report noted a marked inhibitory effect of CGRP on antigen-presentation functions of Langerhans cells in the skin.

c. Vasoactive Intestinal Peptide (VIP) Is Present in Nerves in Lymphoid Organs and Particularly Influences T-cell Trafficking and Function VIP-containing nerve fibers have been identified in Peyer's patch, thymus, chicken bursa of Fabricus, and spleen. VIP receptors have been identified on murine T-cells and human peripheral blood lymphocytes. VIP can inhibit T-cell proliferation. In a series of elegant studies by Cliff Ottaway, the expression of VIP receptors on T-cells was shown to enhance their migration to

gut-associated lymphoid tissue and mesenteric lymph nodes, whereas their down-regulation resulted in the failure of T-cells to traffic to these gut-related sites.

d. Many Other Neuropeptide Systems (Opioid Peptides, Cholecystokinin, Neurotensin, Corticotropin-Releasing Factor, Somatostatin) Are Present in Lymphoid Organs, but Their Functions Remain Unknown Neuropeptide-containing fibers that contain opioid peptides such as β-endorphin, cholecystokinin, neurotensin, CRF, or somatostatin have been identified in primary and secondary lymphoid organs with immunohistochemical staining. Although there are some published data suggesting that these neuropeptides may be able to modulate aspects of immune reactivity directly, the actual role of these compounds is not well-understood and the extensive criteria for neurotransmission have not been worked out. Perhaps the best studied of these neuropeptides is CRF, which exerts an enhancing effect on T-cell functions and acquired immune responses. Denise Bellinger, David Felten, and colleagues have identified CRF-containing nerve fibers in primary and secondary lymphoid organs in rodents; the intensity of CRF staining is enhanced by the prior presence of IL-2, suggesting that a cytokine may up-regulate a neuropeptide that, in turn, can be released and act on the target cells of the immune system.

e. The Cells of Origin for Many Neuropeptide Fibers Are Not Known For some SP–CGRP neurons and the sympathetic NE–NPY neurons, the cells of origin have been mapped by tracing studies. However, we do not yet know the site of origin for the cell bodies giving rise to the majority of neuropeptide innervation of lymphoid organs.

C. Presence of Adrenoceptors, Neuropeptide Receptors, and Hormone Receptors on Cells of the Immune System

1. α- and β-Adrenoceptors Are Present on Many Subsets of Lymphoid Cells

As noted previously, T-cells, B-cells, and monocytes–macrophages possess mainly β-adrenoceptors, but also some α-adrenoceptors, as shown by receptor-binding studies. Pharmacological studies with adrenergic agonists and antagonists suggest abundant functional roles for both subsets. It appears that these receptors have differential distribution, with higher numbers of β-adrenoceptors on B-cells compared with T-cells and on CD8 T-cells compared with CD4 T-cells. β-Adrenoceptors are expressed on Th1 but not Th2 helper T-cells and on pre-IgM cells but not pre-IgG cells.

2. Neuropeptide Receptors Are Present on Selected Subsets of Lymphocytes, Macrophages–Monocytes, and Other Lymphoid Cells

The identification of receptors for neuropeptides on various populations of cells of the immune system does not entirely clarify the role of those receptors. We do not yet know whether only a selected subset of such cells contains a high number of receptors, or whether some broader populations contain a lesser number of receptors. More sophisticated cell-sorting approaches will be needed to clarify this issue, an approach only now being used for some α-adrenoceptors.

3. Most Lymphoid Cells Possess Receptors for a Wide Variety of Neurohormones and Target Organ Hormones under Anterior Pituitary Influence

These receptors allow such anterior pituitary hormones as GH, prolactin, and ACTH to act directly on cells of the immune system and to stimulate the release of other hormones from target organs, such as thyroid hormone, gonadal hormones, and glucocorticoids. These hormones derived from target endocrine organs also can act through appropriate cognate receptors also found on or inside of cells of the immune system.

4. The Neurotransmitter and Neurohormone Receptors on Lymphoid Cells Transduce Signaling through Classical Second Messenger Systems and Exert Both Genomic and Nongenomic Effects on Their Target Cells

In detailed studies of receptor–ligand interactions of neurotransmitter and neurohormone actions, a classical second messenger system usually has been identified. These second messenger systems are typical for the receptor (e.g., cAMP response for the β-adrenoceptor in lymphocytes), activate the appropriate protein kinases or other signaling pathways, and provide a classical cascade system for the exertion of genomic and nongenomic effects. In some populations of lymphoid cells, hormone response elements have been identified in the nucleus, providing a further

opportunity for the hormonal regulation of gene expression.

D. Receptors for Neurotransmitters and Hormones Usually Interact in Multiples and Can Exert Synergistic and Nonlinear Interactive Effects with Each Other and with Cytokine Receptors

Lymphoid cells possess receptors for cytokines, neurohormones, and neurotransmitters. These sytems appear to act nonlinearly with each other. For example, β-endorphin and somatostatin can enhance IL-2-stimulated, lymphokine-activated killing (LAK cell activity), but they have no effect alone without IL-2. Occupancy of the β-adrenoceptor during anti-CD3 activation of the T-cell receptor on human T lymphocytes synergizes the production of cAMP intracellularly, but not other intracellular signals generated by activation of the T-cell receptor. Isoproterenol can synergize the effects of G-CSF on bone marrow cell proliferation in a 3-dimensional culture system. It also is possible that colocalized neurotransmitters, such as NE and NPY, may exert synergistic effects with each other on cells of the immune system, as they do on certain vascular beds.

V. IMMUNE SYSTEM COMMUNICATION WITH THE NERVOUS SYSTEM

A. Hypothalamic and Other CNS Alterations Follow Immune Activation

1. Altered Neuronal Electrical Activity

Hugo Besedovsky and others showed that, during the time of peak antibody production in an immune response, increased electrical activity was detectable in several sites of the hypothalamus, including the anterior region, the PVN, and the dorsomedial region. It was assumed that this increased activity reflected either the actions of cytokines on hypothalamic neurons or some indirect signal from the periphery activated by the immune response.

2. Altered Monoamine Release and Turnover

Sonia Carlson and colleagues showed that, during an immune response, NE levels decreased in the PVN, presumably reflecting increased NE turnover. Adrian Dunn subsequently showed that IL-1β, administered either peripherally or centrally, can provoke a wide range of changes in the turnover of NE and serotonin at numerous central sites, including an increase in turnover of NE in the PVN. Work from several laboratories has shown that the release of NE onto CRF neurons of the PVN is necessary for IL-1β to exert its CRF-elevating effect that is involved in activating the HPA and sympathetic axes.

3. Altered HPA Activation

Besedovsky and others have shown that an immune response to an antigenic challenge with sheep red blood cells, administration of lipopolysaccharide (LPS), and the immune response to a range of viruses can result in an elevation of peripheral glucocorticoids in both the early and late phases of immune response to these challenges. The elevation of glucorticoids occurs through the elevation of both CRF and ACTH. However, not all antigen challenges resulted in such elevations of glucocorticoids. It is not yet known what characteristics of an immune challenge can provoke activation of the HPA axis.

4. Altered CNS Cytokine Production

Glial cells (astrocytes and microglia) and some neurons themselves possess the capacity to synthesize IL-1β, and perhaps TNF-α and IL-6. Inflammation in the brain, rejection of an allograft, or even the neurodegenerative process producing placques and tangles in Alzheimer's disease can result in the production and secretion of cytokines such as IL-1β in the brain. Some investigators, such as Carlos Solomon-Plata, have provided evidence that peripheral stimulation of the brain with IL-1β can directly or indirectly lead to the central production of IL-1β, presumably by glial cells.

B. Many Cytokines Can Influence the Nervous System

1. CNS Influences of IL-1β, IL-2, TNF-α, IFN-γ, IFN-β, and Other Cytokines

A wide range of inflammatory cytokines, especially IL-1β, have been shown to exert numerous central effects. IL-1β can induce slow-wave sleep, produce a fever through the alteration of thermoregulatory set points, induce lethargy and reduced locomotor activity, reduce appetite and sex drive, and change electrical

activity and neurotransmitter turnover in the brain. In studies aimed at enhancing lymphokine-activated killing of cancer cells, the administration of IL-2 in chemotherapeutic regimens produced toxic CNS effects, including some of the characteristics listed earlier, as well as depression and altered affective behavior. IFN-γ was administered to patients with multiple sclerosis in an attempt to ameliorate exacerbations, but it was found to actually increase the frequency of demyelinating attacks. Current therapy now uses IFN-β, which inhibits T-cell activity and diminishes exacerbations of multiple sclerosis.

2. PNS Influences of IL-1β and IL-2

In the peripheral nervous system, both IL-1β and IL-2 appear to provoke the release of NE from sympathetic nerve terminals. This may be a mechanism by which the immune system can regulate the availability of a neural signal molecule by cytokine action on receptors that are present on the presynaptic nerve endings. IL-2 also appears to up-regulate the production of CRF in peripheral neurons that innervate secondary lymphoid organs.

C. Lymphoid Cells and Cytokines Have Many Routes of Access and Many Mechanisms of Immune Communication with the Nervous System

1. Immune Trafficking and Cellular Presence in the Brain

William Hickey and colleagues have shown that activated T lymphocytes can traffic through the brain and have at least a 24-hr window of activity during which they may provoke reactivity by other cell types. These T lymphocytes can release cytokines such as IFN-γ, which in some experimental settings can provoke glial cells to become antigen-presenting cells; such a process could remove the partial ''immune privilege'' of the brain in which a therapeutic graft was present.

2. Production of Cytokines by Neurons and Glia

Chris Breder and Cliff Saper reported the presence of IL-1β in neurons in the hypothalamus. A report by Schultzberg and colleagues described IL-1β colocalized with NE in peripheral sympathetic nerve fibers. Numerous studies have identified IL-1β production by microglia *in vivo* in the CNS. Cytokines also can be produced by astrocytes, especially *in vitro*; the extent to which this process occurs *in vivo* is not known.

3. Peripheral Cytokine Access through the Blood–Brain Barrier

Cytokines initially were assumed to be unable to cross the blood–brain barrier because they are large molecules and do not readily equilibrate between the blood and the CSF. However, William Banks and colleagues have reported a crossing of the inflammatory cytokines, IL-1 (α and β), IL-6, and TNF-α. Sundar and Weiss have used intracerebroventricular administration of IL-1β to activate central pathways and have reported effects from the administration of femtomolar concentrations. This concentration certainly is a conceivable concentration of IL-1β to cross into the CNS from the peripheral blood.

4. Cytokine Actions at the Circumventricular Organs (e.g., OVLT)

Cytokines also may act on cells of the circumventricular organs. The most studied of these systems is the OVLT. IL-1β appears to activate PGE2 secretion by cells of the OVLT, thereby providing direct access into the CNS by this secondary signal. The status of cytokines acting at other circumventricular organs is not as clear.

5. Cytokine Actions on the Pituitary and Target Endocrine Organs

Some cytokines, particularly the inflammatory cytokines, may have direct action on some of the supporting folliculostellate cells in the pituitary, thereby providing a stimulus for potential influence over the release of neurohormones at this level. This action is in addition to the actions on releasing and inhibitory factors that activate hormones of the anterior pituitary through the brain.

6. Cytokine Actions on Peripheral Nerves: Modulation of Autonomic Output, Vagal Input, and Somatosensory Input

As noted previously, cytokines may act directly to release NE from sympathetic nerve terminals, may act on paraneuronal cells to stimulate vagal input into the CNS, and may activate some sensory fibers such as the SP nociceptive axons, providing numerous routes for cytokine signaling to the CNS.

D. Regulatory Loops Link the Immune and Nervous Systems

The influences of neurotransmitters and neurohormones on cells of the immune system and, conversely, the influence of cytokines on the CNS do not occur in isolation and are part of feedback circuits that assure appropriate interactions with proper activation, as well as cessation of potent signals. In this manner, the study of neural–immune interactions parallels the work on feedback circuits that characterized the founding of the field of neuroendocrinology a few decades earlier. The following two examples are the major substantiations of this linked circuitry: (1) Cytokines modulate the activity of the HPA axis and sympathetic nervous system outflow through actions impinging on the PVN of the hypothalamus; and (2) glucocorticoids and catecholamines that are produced by the activation of the HPA axis and the sympathetic nervous system, respectively, modulate cytokine production and immunological reactivity.

VI. CLINICAL IMPLICATIONS OF NEURAL–IMMUNE SIGNALING IN BOTH DIRECTIONS

A. Neural Mediators Can Influence Immunologic Reactivity and Response to Viral and Bacterial Infections and Other Pathogens

1. Both Behavioral and Pharmacologic Interventions Can Alter Neural Mediators Involved in Such Interactions

It now is clear that the CNS can influence immune responses, as shown by stressor studies, conditioning studies, human studies correlating behavior and changes in immune responses, and lesion studies in experimental animals. Many neurohormonal and neurotransmitter systems have been identified that can serve as links between the CNS and the immune system, and cytokine feedback to the CNS has been abundantly demonstrated. The two main choices for approaches that could manipulate this interactive circuitry for clinical benefit are behavioral or pharmacological. Use of behavioral interventions can activate the CNS pathways (presumably from cortical and limbic forebrain input) that regulate outflow to the immune system. Pharmacological interventions with agonists or antagonists to neurotransmitters or neurohormones or with other drugs that influence the availability and binding of these mediators could alter one or more of the mediators used by the CNS to interact with cells of the immune system.

2. Neural Mediators Can Influence Susceptibility to, Progression of, and Outcome of Such Infections

Studies by Sheridan and colleagues have shown that both behavioral and pharmacological interventions can change the susceptibility to infection by herpes simplex virus or influenza virus, as well as the subsequent course of the disease. Stressors such as restraint or differential housing paradigms can result in greater susceptibility to infection. Such increased susceptibility can only be reversed by blocking both glucocorticoids and β-adrenoceptors simultaneously; these blockades differentially affect specific immune functions that combine to restore homeostasis.

3. Neural Mediators Can Influence the Latency and Reactivation of Some Viruses

In vitro studies have shown that some latent viruses can be reactivated by the direct application of glucocorticoids. Stressors can activate some herpes viruses, such as herpes zoster. The extent to which glucocorticoids and other mediators of CNS stress responses exert influence on the progression of HIV is not yet known.

B. Neural–Immune Signaling Can Influence Autoimmune Diseases

1. Some Autoimmune Diseases Directly Affect the Nervous System

a. Multiple Sclerosis (CNS) Multiple sclerosis is a remitting and exacerbating demyelinating disease of the CNS. It is characterized by multiple white matter demyelinating lesions over both time and space. A T-cell-mediated attack on oligodendroglia characterize this condition, and β-interferon or other immunosuppressive therapies now characterize clinical approaches to this disease.

b. Guillain–Barre Syndrome (PNS) Guillain–Barre syndrome is an autoimmune attack on the peripheral myelin, frequently following a few weeks after an infectious process such as an upper respiratory infection (possible autoimmunity by molecular mimicry). There is a rapid progression of distal to proximal

weakness, sometimes accompanied by epicritic somatosensory loss (vibratory sensation, position sense, fine discriminative touch) as the disease progresses. This condition can progress all the way to respiratory failure and dependency on respiratory assistance. The reversal of this disease requires remyelination over the course of several months; some residual effects are sometimes found, and slowing of conduction velocity will be present after remyelination.

c. Myasthenia Gravis (Neuromuscular Junction) Myasthenia gravis (MG) is characterized by the production of antibodies against the nicotinic neuromuscular cholinergic receptors, resulting in easy fatiguability of muscles affected by this disease.

2. Neuromediators Can Influence the Onset, Progression, and Severity of Some Autoimmune Diseases (e.g., Lupus, Rheumatoid Arthritis, Autoimmune Demyelination)

Esther Sternberg and colleagues have suggested that glucocorticoids and the HPA axis play a critically important role in modulating the onset, progression, or severity of autoimmunity in susceptible hosts. Barry Arnason, Eva Chelmicka-Schorr, and colleagues have hypothesized that sympathetic ablation or diminished activity can exacerbate the severity of multiple sclerosis in humans or experimental allergic encephalomyelitis in LEW/N rats. This hypothesis has been supported by data from David Felten and colleagues in LEW/N rats with adjuvant-induced arthritis. In genetically susceptible mice with autoimmune disease (e.g., MRL lpr/lpr mice), there is a deficiency in the sympathetic innervation of secondary lymphoid organs, as there is in several other autoimmune-prone mouse strains, suggesting that the loss of sympathetic innervation of secondary lymphoid organs may be a factor contributing to the development of autoimmunity. Further work from Jon Levine, David Felten, and others has suggested that substance P fibers, both in the joints and in the draining lymph nodes, may modulate the severity and progression of autoimmune rheumatoid arthritis.

C. Neural–Immune Interactions Can Influence Responses and Progression of Some Specific Tumors

Work by Shamgar Ben-Eliyhu and colleagues has shown that a stressor can result in more extensive metastatic spread of mammary tumors compared with unstressed controls and that this is mediated through NK cell activity. They have studied NK cell responses in detail and provide the best evidence that NK cell activity exerts potent antitumor effects on some cancers such as breast cancer. It is tempting to extend this hypothesis to the increased survival of breast cancer patients in the study by Spiegel and colleagues, where peer support and psychotherapy as an adjunct to appropriate treatment doubled the survival. It is not yet known whether this was immune-mediated and, if so, whether it was through NK cell ctivity. Studies by S. ThyagaRajan, David Felten, and colleagues have shown that the use of low doses of deprenyl (selegiline), a monoamine oxidase B inhibitor, can inhibit the growth and spread of carcinogenically induced mammary tumors and can suppress the spontaneous appearance of mammary tumors in old age in susceptible strains. It is possible that this agent, acting on both central neural circuits and the peripheral sympathetic nervous system, can exert a protective effect against mammary tumors through neural actions.

D. Immunosenescence May in Part Reflect Neural Dysfunction

1. An Age-Related Severe Loss of Sympathetic Noradrenergic Innervation from Spleen and Lymph Nodes Is Accompanied by Diminished T-Cell Functions, Especially Cell-Mediated Responses

Denise Bellinger and colleagues have reported that secondary lymphoid organs (spleen, lymph nodes) show an age-related loss of NA nerve fibers, accompanied by diminished cell-mediated immunity. This same functional decline is found in young animals following sympathectomy of the secondary lymphoid organs. Restoration of innervation to the spleen in old rats, via treatment with deprenyl as a stimulator of growth factors, resulted in reinnervation of the secondary lymphoid organs by sympathetic noradrenergic nerve fibers and partial restoration of T-cell functions, suggesting that altered neural relationships with cells of the immune system may contribute to immunosenescent changes.

2. Thymic Involution and Loss of T-cell Responsiveness with Age Can Be Restored by GH and Prolactin Administration

The ability of GH and prolactin to restore thymic cellularity and some T-cell functions suggests that age-related neuroendocrine changes may contribute to

immunosenescence. Use of hormonal and/or neurotransmitter pharmacological interventions may be useful for the restoration of T-cell function in aging.

3. An Altered Microenvironment in Both Primary and Secondary Lymphoid Organs May Contribute to Age-Related Deterioration of Some Aspects of Innate and Acquired Immunity

At least some critical aspects of an altered microenvironment in primary and secondary lymphoid organs may be due to changes in neurohormone and neurotransmitter availability and interactions. In addition, lymphoid cells may alter their responses to neural signals (e.g., diminished intracellular generation of cAMP from β-adrenoceptor activation of lymphocytes in aging, despite the increase in actual numbers of β-adrenoceptors on the surface of aging T lymphocytes). The possibility that neuromediators may be involved in such alterations suggests new therapeutic approaches using neuropharmacological and behavioral means of altering immune reactivity.

E. Cytokine Signaling to the CNS Can Provoke a Host of Reactive Responses Related to Illness and Stress

Many major changes in brain function can be evoked by the action of cytokines, including (1) fever, sleep, and illness behavior; (2) activation of stress axes (HPA axis, sympathetics); (3) affective, cognitive, and behavioral changes; and (4) direct immune cell invasion of and interactions in the CNS. The role that cytokines play in altering CNS function in helps part to explain ''illness behavior'' and suggests interactions between health status and CNS functions that may be particularly devastating in the elderly, where hippocampal function may be compromised and CNS perfusion may be diminished.

VII. A FINAL REEVALUATION OF THE AUTONOMY OF INDIVIDUAL SYSTEMS: NEW OPPORTUNITIES FOR A MOLECULAR AND CELLULAR UNDERSTANDING OF MULTISYSTEM INTERACTIONS IN THE WHOLE PATIENT

The field of psychoneuroimmunology has demonstrated that the immune system is not an autonomous, exclusively self-regulating system, neither is the brain devoid of influences emanating from immune signaling. Rather, the CNS and the immune system appear to interact with complex signaling and feedback loops for the protection and homeostasis of the whole organism. Although we how have wonderful opportunities to use molecular and cellular (and, hence, reductionistic) approaches to study neurohormonal and neurotransmitter signaling of cells of the immune system and cytokine signaling of neurons and glia, ultimately we must return to the host, the whole organism. The individual is a complex product of cognition, emotion, and other CNS functions that cannot be studied with full understanding *in vitro* or in isolation. Yet, a holistic approach is extremely challenging in a research setting, particularly where complex variables must be controlled. However, we may have an unprecedented opportunity for a better understanding of behavioral influences on neurohormonal and neurotransmitter responses in the periphery and ultimately may be able to synthesize a more complete understanding of the complexities of biological signaling among the cytokines, neurohormones, and neurotransmitters. We finally are on the threshold of a biologically sound and mechanistically understandable merging of body, mind, and spirit.

See Also the Following Articles

AUTOIMMUNE DISEASES • BEHAVIORAL NEUROIMMUNOLOGY • HIV INFECTION, NEUROCOGNITIVE COMPLICATIONS OF • NEUROTRANSMITTERS • PEPTIDES, HORMONES, AND THE BRAIN AND SPINAL CORD • PSYCHONEUROENDOCRINOLOGY • PSYCHOPHYSIOLOGY

Suggested Reading

Ader, R., Cohen, N., and Felten, D. L. (1995). Psychoneuroimmunology: Interactions between the nervous system and the immune system. *Lancet* **345,** 99–103.

Ader, R., Cohen, N., and Felten, D. L. (2000). *Psychoneuroimmunology*, 3rd ed. Academic Press, San Diego.

Ader, R., Felten, D. L., and Cohen, N. (1990). Interactions between the brain and the immune system. *Annu. Rev. Pharmacol. Toxicol.* **30,** 561–602.

Bellinger, D. L., Madden, K. S., Felten, S. Y., and Felten, D. L. (1994). Neural and endocrine links between the brain and the immune system. In *The Psychoimmunology of Cancer* (C. E. Lewis, C. O'Sullivan, and J. Barraclough, Eds.), pp. 55–106. Oxford University. Press, New York.

Felten, D. L. Neural influence on immune responses: Underlying suppositions and basic principles of neural–immune signaling. *Prog. Brain. Res.* **122,** 379–389.

Felten, D. L., and Maida, M. E. (2000). Neuroimmunomodulation. In *Encyclopedia of Stress* (G. Fink, Ed.), Academic Press, San Diego.

Kiecolt-Glaser, J. K. (1999). Stress, personal relationships, and immune function: Health implications. *Brain Behav. Immunity* **13,** 61–72.

Madden, K. S., and Felten, D. L. (1995). Experimental basis for neural–immune interactions. *Physiol. Rev*. **75,** 77–106.

Madden, K. S., Sanders, V. M., and Felten, D. L. (1995). Catecholamine influences and sympathetic neural modulation of immune responsiveness. *Annu. Rev. Pharmacol. Toxicol*. **35,** 417–448.

McEwen, B. S. (1998). Protective and damaging effects of stress mediators. *New Engl. J. Med*. **338,** 171–179.

Reichlin, S. (1993). Neuroendocrine–immune interactions. *New Engl. J. Med*. **329,** 1246–1253.

Sheridan, J. (1998). Stress-induced modulation of antiviral immunity. *Brain Behav. Immunity* **12,** 1–6.

Watkins, L., Maier, S., and Goehler, L. (1995). Cytokine-to-brain communication: A review and analysis of alternative mechanisms. *Life Sci*. **57,** 1011–1026.

Wilder, R. L. (1995). Neuroendocrine–immune system interactions and autoimmunity. *Annu. Rev. Immunol*. **13,** 307–338.

Psychophysiology

DAVID LYKKEN
University of Minnesota

GLOSSARY

alpha rhythm EEG or brain waves in the frequency range from 8 to 12 Hz, found in most human subjects during a state of relaxed wakefulness.

electrocardiogram (ECG) The electrical signal generated by the heart muscle, which can be recorded between two electrodes on the body surface. The principal components are the P wave, produced by atrial contraction, the QRS complex, produced by contraction of the ventricle, and the T wave, representing repolarization of the muscle.

electroencephalogram (EEG) The electrical signal recorded from the scalp and reflecting the synchronous activity of masses of cortical neurons in the region under the active electrode.

electromyogram (EMG) The electrical signal recorded from the body surface that is produced by the synchronous contraction of a muscle mass.

emergenic trait A characteristic that is determined by independently segregating genes that combine configurally rather than additively. Monozygotic twins are similar with respect to emergenic traits, whereas siblings or parent–child pairs are not.

event-related potential (ERP) An EEG signal representing aspects of the brain's response to discrete stimuli. Normally hidden by on-going EEG activity, ERPs can be visualized by summing 0.5 sec segments of EEG following repeated presentations of the stimulus so that the background activity averages out.

functional magnetic resonance imaging (fMRI) A noninvasive technique for identifying currently activated brain regions by their increased blood flow.

guilty knowledge test (GKT) A method of forensic interrogation that involves recording autonomic or ERP activity elicited by the alternatives in a series of multiple-choice questions presented orally or visually; the purpose is to determine whether the subject responds differently to the correct alternatives, thus revealing that he or she recognizes which alternatives are correct.

monozygotic (MZ) Twins that result from the early splitting of a fertilized ovum and who share an identical genome or DNA blueprint.

plethysmograph A device for recording changes in the volume of a part of the body.

polygraph An instrument for recording several different electrical signals by pens writing in parallel on a moving paper chart.

preception The modulation of the sensory impact of a stimulus, the nature and timing of which can be accurately predicted. Thus, a painful electric shock produces less impact, indicated by an attenuated psychophysiological response, when its time of occurrence can be accurately predicted (''negative'' preception).

skin conductance level (SCL) The average level of electrical conductance of the palmar skin, a quantity that varies monotonically with CNS arousal.

skin conductance response (SCR) A wavelike increase in SCL elicited by stimuli, external or internal. The amplitude of the SCR varies with the subjective impact of the stimulus.

tachycardia Rapid heart rate.

two-flash threshold (TFT) The shortest interval (e.g., 50 msec) between two brief flashes of light that allows the subject to detect the paired flashes as different (e.g., as a flicker) from a single flash.

volar Palm of the hand or sole of the foot.

Psychophysiology is the study of mental or emotional processes as revealed through involuntary physiological reactions that can be monitored noninvasively from an intact subject. The independent variables usually will be psychological. A human subject may be asked a question, given a problem to solve, put under emotional stress, instructed to perform some task, or to attend to a series of simple stimuli, and so on. The dependent variables will be physiological changes that can be recorded peripherally either as electrical signals

(e.g., brain waves, muscle potentials, the electrocardiogram), as pressure, volume, or temperature changes (e.g., blood pressure, skin temperature), or as reflexive movements (e.g., eye blinks, breathing movements). Rarely, the psychophysiologist might use biochemical changes in urine, blood, or sweat as dependent variables.

Research in *physiological psychology* usually reverses this sequence of stimulus and response. There the independent variables are normally physiological (e.g., electrical or chemical stimulation of brain centers), whereas the dependent variables are commonly psychological (e.g., behavior observations or subjective report.) One exception includes the neuroimaging techniques, especially fMRI, which are being used increasingly for the purpose of identifying brain regions involved in specific forms of neurocognitive processing or emotional experience. One anticipates that these techniques will one day be used to identify the cognitive events or emotional experiences that were elicited by some stimulus and, thus, become tools of psychophysiology.

Psychophysiology must be distinguished also from the field of *psychosomatic medicine*, because workers in both areas share an interest in many of the same physiological manifestations of mental and emotional events. For the psychophysiologist, the physiological reaction is a medium that carries a message about events occurring in the mind or brain. The fact that fear can produce peripheral vasoconstriction and a rapid heartbeat is a concern to the student of psychosomatic medicine, who is interested in these physical reactions in their own right. It is the fact that cold hands and tachycardia *betoken* fear that is of interest to the psychophysiologist.

Psychophysiology is not one of the substantive disciplines within psychology, such as psychopathology or cognitive or social psychology, but rather it is a technology that might be (and has been) employed in all of these disciplines. In this respect, psychophysiology is to psychology as, say, microscopy is to physiology. The psychophysiologist's formal training may be in psychology, physiology, medicine, or engineering. Because many of the phenomena studied are electrical in nature and the instrumentation is electronic, a psychophysiologist must understand at least the rudiments of electrical theory. Many psychophysiologists function adequately with minimal training in anatomy and physiology. A typical psychophysiologist will devote part of his or her professional career to improving methods of measurement or inventing new ones and the rest of the time in applying this technology to some substantive problem in psychology, psychiatry, or even such diverse fields as criminology or political science. The principal scientific journals in this field are *Psychophysiology*, the journal of the Society for Psychophysiological Research, and the *International Journal of Psychophysiology*, the journal of the International Organization of Psychophysiology.

I. PSYCHOPHYSIOLOGICAL MEASUREMENT

The immediate object of psychophysiological measurement is to generate an electrical signal that faithfully mimics the manner in which the physiological phenomenon being measured varies over time. Once the phenomenon has been represented as an electrical signal, it can easily be amplified or filtered, visualized as a tracing on a polygraph chart or on a computer monitor, and recorded digitally for later playback or analysis. Some psychophysiological phenomena, such as the electroencephalogram (EEG), the electromyogram (EMG), and the electrocardiogram (ECG), are already electrical signals generated in the body, and their measurement requires only a pair of electrodes appropriately placed to pick up the biological voltage and connected to the input of an amplifier that will boost this voltage until it is strong enough to be recorded in some way. The most versatile method of recording is as a digitized computer file, but many psychophysiology laboratories also make a permanent visual record of the signal using an electrically driven pen that moves up and down with changes in the signal, writing on a paper chart that moves laterally. A *polygraph* consists of several pen motors arranged side by side, each driven by a separate amplifier, so that several separate signals can be recorded on the same time base on one chart. A polygraph with, say, four amplifiers and pen motors is said to have four *channels*.

It is important to note that a single channel can contain numerous variables. For example, the ECG channel provides a complex analog record of the electrical events associated with the cardiac cycle. There are many separate variables that can be quantified from this record: the height of the P, R, and T waves, the time periods between them, the time between successive R waves, which is the reciprocal of heart rate, and so on.

A. Electrodes

Electrodes used in psychophysiology are junctions where the flow of electric current changes from ionic, in

the skin and other tissues, to electronic, in the wires of the external circuitry. Such an interface is subject to electrochemical processes that can produce polarization. A polarized electrode acts like a high-pass filter that discriminates against slow or low-frequency changes. Relatively nonpolarizing electrodes are available, typically made of silver coated with silver chloride. An electrode paste or electrolyte is applied between the skin surface and the electrode; the properties of this conductive paste are also important for successful recording.

B. Transducers

Phenomena such as pressure or temperature changes can be converted into electrical signals by means of an appropriate transducer. A thermister, for example, is a device whose electrical resistance varies reliably with temperature. By passing a weak electric current through a thermister probe and amplifying the voltage developed across it, one can produce a signal that accurately represents changes in temperature. A strain gauge, similarly, changes in resistance as it is flexed and therefore can act as the "sense organ" of a pressure transducer or a device for measuring, for example, breathing movements.

Some psychophysiological phenomena that do not produce signal voltages directly may involve changes in the electrical properties of tissue that can be measured by passing an external sensing current through the tissue. The electrical conductance of the palmer skin is one example. The standard technique for measuring the skin conductance response (SCR) involves applying a constant 0.5 V between two nonpolarizing electrodes attached to the palmar skin surface and measuring the small direct-current flow (less than 10 μA per cm^2 of electrode area) through the tissues. A number of physiological phenomena, including blood flow, heart action, muscle contraction, and respiration, produce changes in the electrical impedance of the associated tissue that can be measured by passing a weak, high-frequency (e.g., 50 kHz) sensing current through the body. These impedance techniques are used infrequently but have considerable promise for specialized applications.

C. Noise

The modern world is literally full of electrical "noise," electromagnetic emanations from television transmitters, electric motors, passing autos, fluorescent lights, and so forth, which the human body picks up as an antenna does. Bioelectric signals originating in the body, similarly, become noise when they are not the signal one wants to measure but appear nonetheless in one's recordings. Where once it was necessary to study weak signals such as the EEG in awkward and expensive shielded rooms, using ranks of storage batteries to provide power untainted by the AC mains, modern amplifiers make this task much simpler. The noise-rejecting capabilities of these amplifiers will be realized, however, only if one (and only one) low-resistance electrode connects the subject to earth or electrical ground.

Noise of biological origin, as when eye movements affect the EEG or when the ECG shows up unwanted in the electrodermal channel, requires special solutions. Sometimes reorientation of the electrodes will suffice. If the noise consists mainly of frequencies outside the bandwidth of the desired signal, a bandpass filter may provide the solution. A third approach is to measure the noise directly in a separate channel and then subtract it from the signal channel by electronic inversion and summation.

D. Safety

Unpleasant or even dangerous electric shocks are always a potential hazard in the vicinity of apparatus connected to the AC power mains. These same shocks can easily be fatal if the recipient happens to be well-grounded. Because subjects in psychophysiological research are almost always well-grounded, it is necessary to take sensible precautions. The voltage with respect to ground of every electrode, wire, or other conductive surface with which the subject might come into contact should be measured, and a variety of other safety measures are available.

E. Computers

Most psychophysiology laboratories now employ small computers for on-line control of experiments and immediate analysis of data and also for more complex subsequent analyses. Laboratory interface systems are available that make it possible to turn things on and off under computer control, to generate stimuli, to time events, and also to provide to the computer data, command signals, and other information from the laboratory. The computer can present pictorial or alphanumeric information to the subject by means of a monitor display and also a variety of

auditory stimuli including spoken words such as "right," "wrong," and "good," which have been digitally recorded and stored in the computer's memory. Psychophysiologists of the past have required a working knowledge of electrical principles, physiology, and statistics and a more than rudimentary understanding of psychology; competent psychophysiologists of the present require a working knowledge of the computer as well.

II. ANALYSIS OF THE DATA

A. Components of the Response

The variance of a sample of scores on some psychophysiological variable, ω, can be partitioned thus:

$$\sigma_\omega^2 = \sigma_\psi^2 + \sigma_\phi^2 + \sigma_\varepsilon^2 \quad (1)$$

where σ_ψ^2 is the variance due to individual differences in the underlying psychological variable of interest, σ_ϕ^2 is the orthogonal component of variance due to physiological differences, and σ_ε^2 represents measurement error. If skin conductance level (SCL), represented by ω, is being measured, for example, ψ might be central nervous system (CNS) arousal or "energy mobilization," ϕ would reflect individual differences in the density and activity of volar sweat glands (which largely determine skin conductance), and ε would increase with variations in cleaning the skin surface, positioning of the electrodes, the area of skin contacting the electrolyte, and so on.

Underlying most psychophysiological measurements is the implicit assumption that ω is a monotonically increasing function—and, it is hoped, a simple linear function—of the underlying variable of interest, ψ, as in

$$\omega = a + b\psi + \varepsilon. \quad (2)$$

By using SCL again as the example, the parameter a would represent this subject's minimum SCL when sudomotor activity is zero, whereas b would be determined by the reactivity of the entire electrodermal system, that is, the increase in conductivity produced by a unit increase in ψ. Very similar assumptions are implicit in most psychological measurements. The problem is that the parameters a and b also vary, often within the same individual from time to time and certainly from one individual to another. This is the variation represented by σ_ϕ^2 in Eq. (1). The job of the psychophysiologist is, first, to ensure that the physiological variable chosen (ω) *is* linearly related to ψ, at least approximately, and then to try to minimize both the measurement error σ_ε^2 and also σ_ϕ^2, the variance due to physiological variability, within subjects or between subjects, which also must be regarded as error variance in this context.

B. The Linearity Assumption

Consider an experiment in which the subject is intensely stressed at the outset and then is allowed to relax and go to sleep, while skin potential level (SPL) is continuously monitored. The SPL will be fairly low under intense stress, then will rise to a maximum while the subject is, say, listening to an interesting story, and then will fall again to a minimum when the subject goes to sleep. These individual curves show us that SPL has an inverted U-shaped relationship to CNS arousal and is therefore a poor index of that variable. Skin conductance level (SCL) and also its reciprocal, skin resistance level (SRL), both normally vary monotonically with CNS arousal, and it appears that the relationship of SCL to arousal is more nearly linear than that of SRL.

Suppose that, in the same experiment, we also measure electrodermal responses: SCRs from one hand and resistance changes or SRRs from the other. Because resistance is the reciprocal of conductance, and the responses are being elicited over widely varying levels of tonic SCL and SRL, the SCRs will be poorly correlated with the corresponding SRRs. Thus, SCR and SRR cannot be equivalent indicants of the same psychological process, for example, of the psychological impact of the stimulus. Which of the two should be used? There is both theoretical and empirical support for the view that conductance is more simply related to central events than is resistance.

These examples illustrate that it is important to investigate the form of the relationship between ω and ψ, that this requires experimental manipulation of ψ (remembering, as we have seen, that the parameters of the function also will vary among subjects), and that the investigation will be illuminated by whatever one knows about the physiological substrate of the variable studied.

C. Minimization of Extraneous Variance

Minimization of variance resulting from error of measurement is largely a matter of competent and

consistent technique; the details will depend on the variable being measured. To minimize variance due to extraneous physiological differences requires a statistical correction for individual differences in the range over which the measured quantity can vary in that subject at that time. The basic idea is to express each individual's score as a fraction of the range over which that person's scores can vary, thus:

$$\text{score}_{RC} = \frac{\text{score}_{raw} - \text{min}}{\text{max}}$$

In the case of SCL, for example, min would be the subject's minimum SCL obtained when relaxed or asleep. The estimate of max might be obtained from that subject's maximum SCL shown under high stress (having the subject blow up a small balloon until it bursts has been found to elicit good estimates of SCL_{max}.) In the case of phasic changes such as the SCR, min or the minimum value is always zero. Phasic response values can therefore be range-corrected merely by dividing by an estimate of that subject's maximum response amplitude.

D. Correction for Unsystematic Variation

Most psychological trait or response variables show intraindividual variability when the measurements are repeated, especially over long intervals (e.g., months or years). Although part of this variation can be attributed to measurement error, some of it reflects variability of the relevant response system or the state of the organism (e.g., mood, fatigue) due to recent events or due to variations in the context of measurement. Thus, for example, whereas MZ twins may show within-pair correlations of, say, 0.50 with respect to certain personality traits, the *between-twin cross-time* correlations over a period of months or years may be almost as high as the *within-twin cross-time* or retest correlation. Such results indicate that the stable component of the trait has a much higher heritability than does a measurement made on a single occasion. These findings suggest that relationships between single-occasion measurements of psychophysiological variables are likely also to underestimate the true or stable correlations of those variables. That being so, the additional cost of retesting the same subjects on different occasions is likely to pay benefits in terms of stronger and more revealing measures of association.

E. Within-Subject versus Between-Subject Correlations

Many psychophysiological studies that are intended to discover within-subject relationships between variables actually employ a between-subject design. Consider, for example, the question of whether a measure of autonomic arousal, such as SCL, is related to measures of cortical arousal, such as reaction time (RT) or the two-flash threshold (TFT). Suppose that, for each of N subjects, one measures the mean SCL during the brief time required to estimate the TFT and then plots these pairs of values over the N subjects. This plot will represent the true within-subject relationship *only* if the within-subject curves all have similar shapes and parameters. This appears to be true for normal subjects because the between-subject correlation of SCL with TFT is strongly negative. For unmedicated schizophrenic subjects, however, the between-subject relationship is negligible. It may be, however, that for schizophrenics TFT decreases linearly with increasing SCL, as it does for normals, but that the individual curves have widely varying slopes. That could yield a negligible between-subject correlation but with very different theoretical implications than if the within-subject plots also showed little correlation.

III. THE CHANNELS OF PSYCHOPHYSIOLOGY

A number of organ systems provide the psychophysiologist with a variety of (clouded) windows through which to observe mental events. This section reviews the most widely studied of these systems.

A. The Cardiovascular System

People have been drawing inferences about each other's mental and emotional processes from cardiovascular changes since the dawn of history because some changes—blushing and blanching of the skin, pounding of the heart, cold hands, and the like—can be detected without instrumental assistance. The important channels are the ECG, arterial pressure, finger pulse pressure, and perhaps digital temperature. The ECG is a sequence of electrical signals generated by the busy heart muscle and radiated throughout the body. The psychophysiological variable most

commonly derived from the ECG channel is heart rate, the reciprocal of the time interval between successive R waves produced by ventricular contractions. The ECG is often fed into a device called a cardiotachometer, which electronically measures this interval, converts it to rate, and outputs a signal proportional to instantaneous heart rate.

In the intact subject, blood pressure can be measured only intermittently by auscultation. A pressure cuff on the upper arm is inflated until the brachial artery is sufficiently compressed to occlude the flow of blood. With a stethoscope over the artery distal to the cuff, the pressure is gradually released until the first Korotkoff sounds are heard. These sounds are caused by the spurting of blood through the arterial occlusion during the peak of the pressure cycle, just after ventricular contraction. The first sounds mark the peak or *systolic* blood pressure. As pressure is relaxed further, the sounds wax and wane until a point is reached at which blood flow continues even at the minimum of the pressure cycle; this is the *diastolic* pressure.

Electronic detectors are now available to replace the ears of the skilled clinician and can measure blood pressure reliably, although their absolute accuracy may be off by several millimeters of mercury. Automatic systems for intermittently inflating and deflating the cuff have been used with some success, but the process is cumbersome and both distracting and uncomfortable to the subject. A possibility currently being explored is that blood pressure may be reliably related to pulse transit time, that is, the time required between the initiation of the pressure pulse when the ventricles contract and the arrival of that pulse at, say, the finger. Peripheral blood flow can be detected either by a pressure transducer (a plethysmograph) on a finger or by means of a photoplethysmograph, a device in which a light is directed into the skin at an angle and the reflections are detected by a photoelectric cell. The reflectance varies as blood surges in and out of the vascular bed of the dermis, yielding a signal that varies with the digital pressure pulse.

One must always remember that the heart has more important things to do than to whisper secrets to the psychophysiologist, and it is under strong homeostatic control. Heart rate and blood pressure tend to obey the law of initial values, which states that the change in either variable produced by a stimulus will be inversely correlated with the prestimulus level of that variable. A pressor stimulus will cause a smaller increase in the rate of an already racing heart than in one beating slowly and calmly.

B. The Electrodermal System

Compared with subdermal tissues, the skin has a relatively high resistance to the passage of electric current. In the latter part of the nineteenth century, it was discovered that the resistance of the thick skin of the palms and soles was extraordinarily reactive to psychological stimulation. It is known that the sweat glands in these volar regions subserve a special function; instead of helping with thermoregulation, they moisten grasping surfaces in preparation for action. Dry palmar skin is both slippery and more subject to abrasion. Neural circuits arising in the activating systems of the midbrain control volar sweating, which increases tonically with CNS arousal and which also shows wavelike, phasic increases in response to any stimulus important enough to produce an orienting response. In part because the sweat gland tubules provide a low-resistance pathway through the epidermis, the electrical resistance of the skin varies with sweat gland activity. Because, in fact, resistance varies inversely with sweating, current practice is to measure skin conductance, the reciprocal of resistance.

Skin conductance level (SCL) is lowest in a drowsy or somnolent subject, rises sharply with awakening, and rises still further during mental effort or emotional storm. Superimposed upon these tidal changes of SCL are the wavelike, phasic skin conductance responses (SCRs) to discrete stimuli. After a latency of perhaps 1.5 sec, conductance rises rather quickly to a peak and then returns more or less rapidly to the prestimulus level. Both the latency period and the size and shape of the SCR are affected by hand temperature; if the hand is cold, SCRs are sluggish and diminished in amplitude. The SCR amplitude seems to vary with the psychological impact of the eliciting stimulus; other things being equal, strong stimuli produce larger SCRs than weak ones, but an unexpected or especially significant weak stimulus will produce a larger SCR than an expected but meaningless strong stimulus.

Probably as a result of inward pumping of sodium ions from secreted sweat, the skin surface tends to be 10–80 mV negative with respect to the underlying tissues. Also a result of sweat gland activity, this endogenous skin potential provides a channel of information that is to some extent parallel with the skin conductance channel, but there are nonlinearities that complicate interpretation. The skin potential response (SPR) may be an increase in negativity, an increase followed by a decrease, or a monophasic wave of decreased negativity, depending on the strength of the stimulus and the prestimulus SPL. For these

reasons, the modern tendency is to employ the skin conductance channel in preference to skin potential.

Because the SCR varies with the psychological impact of the eliciting stimulus, it has been used to assess individual differences in *preception*, the ability to reduce the impact of a predictable noxious stimulus. Because it requires stimulus predictability, such negative preception can be regarded as supplementary to habituation, which requires less predictability but more repetitions of the to-be-attenuated stimulus before it takes effect.

C. Electromyography (EMG)

An electrode on the skin over any muscle mass, referenced against an electrode in some quiescent region such as an earlobe or over the shin, will pick up a high-frequency (100–500 Hz) signal produced by the repeated firing of hundreds or thousands of muscle fibers. This signal can be electronically integrated to yield a simpler curve representing average muscle tension. Except perhaps in deep or Stage 4 sleep, the striate muscles maintain a degree of tonus even at rest, with individual fibers firing asynchronously at a low rate. In a ''tense'' individual, this resting tonus may be quite high, either generally or in specific muscle groups. Surface electromyography provides a means of monitoring such subactive muscle tension. One common application is in relaxation training, in which some easily interpretable indicant of current muscle tension is fed back to guide the subject's efforts to achieve voluntary control.

An important research application of electromyography is the detection and recording of facial expressions having emotional significance. For example, small electrodes positioned above the eyebrow, over the corrugator supercilli muscle, will record frowning, whereas placements on the cheek, over the zygomaticus major muscle, will record smiling.

D. Eye Movements

The eyes, those portals through which the brain receives so much of its information about the external world, are also ''windows of the soul'' through which one obtains glimpses of the workings of that brain. Eye movements and the direction of gaze can be monitored by electrooculography (EOG). The eye is like a little battery with the cornea about 1 mV positive with respect to the back of the retina. If electrodes are positioned adjacent to the outer canthus of each eye, then when both eyes are turned to the right, for instance, the electrode on that side becomes electropositive to the one on the left. Another pair of electrodes above and below one eye will record vertical eye movements and also eye blinks (because the eyeball rotates upward with each blink). The sensitivity of the EOG is illustrated by the fact that, when subjects track a target moving sinusoidally from side to side of a computer screen, the EOG recorded on the polygraph will usually be a nearly perfect sine wave. If the target is then driven by a triangular waveform, the EOG record will reproduce this change.

The EOG has been used to study the saccadic eye movements employed in reading or in searching a visual display. It has also been used in the study of nystagmus and the smooth following movements with which the eyes track a moving target. A defect in smooth tracking performance has been shown to be characteristic of many schizophrenic patients. This deficit appears to result from some central defect in the processing of visual information. Because it appears in schizophrenics in remission as well as in many first-degree relatives of these patients, it is now believed that visual smooth-following dysfunction may prove to be a marker for the genetic predisposition associated with at least one subtype of that illness.

E. The Pupillary Response

The size of the pupil, which can vary from about 2 to 8 mm in diameter, is regulated by the autonomic nervous system so as to tend to hold constant the intensity of light admitted to the retina. However, the pupil is also reactive to psychological stimulation with small (< 1 mm), but regular changes (usually dilations) that follow a stimulus with latencies on the order of 0.2 sec. Of the various techniques that have been used for measuring these pupillary responses, the most sophisticated involves a TV monitoring system using infrared illumination and automatic measurement of the image of the pupil, which yields an output signal proportional to pupil size that can be recorded on the polygraph or in the computer. An accumulation of evidence indicates that the pupillary response measures the proportion of total cognitive processing capacity that has been invested in the analysis of the eliciting stimulus. For this reason, larger responses will be elicited by more complex stimuli or more difficult problems, by more interesting or more important stimuli, and also by those near-threshold stimuli that

are detected (and thus require processing) as compared with those that escape detection. Earlier theories that pupil size varies with the attractiveness of stimuli are subsumed by the processing hypothesis; it is likely that more attractive stimuli tend to be subjected to more intensive processing. Pupillary responses to relatively simple mental tasks are larger for less intelligent subjects, suggesting that the less efficient brain must deploy a greater proportion of its resources to solve a given problem.

F. The Eye Blink

The eye blink is a component of the reflexive startle response. The startle elicited by a sudden-onset auditory stimulus is related to the subject's prestimulus emotional state. For example, a subject viewing a disturbing scene produces a stronger blink response than when viewing a neutral scene, whereas, if the scene is pleasant and engaging, the blink response is attenuated. However, within the first few hundred milliseconds after a picture is presented, there is a general inhibition of the blink reflex to acoustic probes. This inhibition reflects the attenuation of new sensory input in order to ensure adequate processing of the prior stimulus.

This paradigm has been used to demonstrate that, unlike normal subjects, primary psychopaths react to frightening scenes as they react to pleasant, interesting scenes, by showing reduced startle responses in both conditions. Similarly, if the acoustic probes are presented while subjects are anticipating a painful electric shock rather than while viewing slides, psychopaths show less startle potentiation than do nonpsychopaths.

G. Electroencephalography (EEG)

Whereas the EEG might appear to provide the most direct "window" of all through which to observe mental events, the electroencephalographer has been likened to a spy, prowling outside the concrete walls of a great factory complex, trying to infer from the din of noises reaching the outside what is going on within. The electrical activity of the brain is far more complex than that in the most elaborate manufactured computer; only a minute part of this information is available at the brain's surface and still less at the scalp. Because an electrode on the scalp integrates electrical activity over a considerable area of cortex, a reasonably comprehensive record of the total EEG can be obtained from about 20 electrodes distributed systematically over the head. A set of standard placements has been defined, called the International Ten–Twenty System. The complete montage will be used by clinicians looking for EEG evidence of tumors or epileptiform activity, whereas the researcher more commonly uses only one or a few EEG channels.

The most common use of the spontaneous EEG is in sleep research where, with additional channels recording lateral eye movements and muscle tension, it is possible to identify, with considerable reliability, the stages of sleep. Computer programs are available for doing a fast Fourier transform of a segment of EEG from, say, a relaxed waking subject, thus producing a spectrum or graph of the frequency content of the EEG, most of it contained in the band from 0 to 20 Hz. An individual's EEG spectrum is relatively distinctive and also stable, as long as it is obtained under similar conditions each time. The spectra of "identical" or monozygotic twins tend to be as similar as that of one person measured on two occasions. Fraternal or dizygotic twins, on the other hand, show negligible similarity with respect to spectral parameters. Thus, it appears that the features of the resting spectrum, the relative amounts of alpha, beta, or theta activity, the midfrequency of the alpha rhythm, and so on, though strongly determined by genetic factors, are emergenic traits that are not commonly passed from parent to offspring. Whereas it seems reasonable to expect that these parameters ought to have interesting psychological correlates, as yet none have been reliably demonstrated.

H. Event-Related Cortical Potentials (ERPs)

Virtually any stimulus sensed by the subject will produce an effect upon the EEG; indeed, much of the apparently spontaneous EEG may be simply the composite effect of the flux of stimulation, external and internal, that continuously bombards the sensorium. To detect the effect of all but the most intense stimuli, against the background of EEG activity, requires repeated presentations of the stimulus so that the immediate poststimulus segments of the EEG record can be averaged together. If 100, 0.5-sec EEG segments are randomly selected and then averaged, the mean will tend toward a straight line. But the 0.5-sec segments that follow 100 presentations of, say, an auditory click will each contain the ERP elicited by that click, a relatively complex train of waves that is

time-locked to the stimulus. The averaging process minimizes the random background activity and reveals the features that are consistent in each sample.

The earlier components of the ERP seem to represent earlier stages of cerebral processing. Evidence suggests the possibility of a relationship between the speed (latency) of these components and some basic dimension of intelligence. Later components, especially a positive wave about 300 msec poststimulus, seem to reflect the completion of a process of stimulus identification or classification. The actual latency of this wave varies with reaction time, and its amplitude varies with the information content of the stimulus; unexpected, important, or possibly memorable stimuli produce larger P300 components. The amplitude of the P300 wave also decreases as part of the aging process.

Other types of ERP that have received research attention include *contingent negative variation* (CNN) and *mismatch negativity* (MMN). In studies of reaction time, where a warning signal precedes the imperative stimulus (the "Go!" signal) by one or a few seconds, the EEG signal measured at the top of the head moves in a negative direction, and the magnitude of this CNN is related to the degree of attentional focus and, thus, to the speed or accuracy of the response. MMN is a negative ERP seen most strongly in frontal areas, which is elicited by a deviant stimulus presented infrequently in a train of repeated standard stimuli.

The study of the psychological correlates of components of the ERP and the use of these data in formulating and testing models of the way in which the brain processes information constitute one of the most active and promising areas of current psychophysiological research.

IV. APPLIED PSYCHOPHYSIOLOGY

A. Biofeedback

Psychophysiological techniques are sometimes used for purposes of relaxation training. Skin conductance or heart rate can be measured and converted to a visual or auditory signal that is fed back to the subject. The subject uses this signal during relaxation training in order to monitor success. Because the alpha rhythm of the EEG occurs during relaxed wakefulness but gives way to higher frequencies when the subject is aroused or anxious, a feedback signal that is proportional to the average amplitude of this activity is sometimes used for the same purpose. Lower frequency (1–5 Hz) EEG waves, as well as sporadic spiky waves called *sleep spindles*, indicate incipient sleep onset. An EEG monitoring system that feeds back an alarm signal in response to such activity has been proposed for use as an accident prevention tool by long-distance truck drivers or persons subject to attacks of sleep apnea.

B. Polygraphic Detection of Deception

The most extensive and socially important use (or misuse) of psychophysiology is the polygraph or "lie detector" test. Since the early twentieth century, primarily in the United States, police and various federal agencies increasingly have employed polygraphs to monitor blood pressure changes, breathing movements, and palmar skin resistance during the questioning of criminal suspects or job applicants. The federal employee Polygraph Protection Act of 1988 prohibits requesting or requiring job applicants or employees to submit to polygraph testing in most areas of the private sector. Ironically, however, this practice continues to be used by federal police and security agencies and by the military. On evolutionary grounds, it would be unlikely that our species would have acquired by natural selection some involuntary response that we show when, and only when, we attempt to deceive. No such specific lie response has ever been demonstrated.

Most people, however, will tend to show some perturbation in their breathing, blood pressure, and/or in the sweating of their palms after falsely answering an accusatory question. The problem is that most people will also evince similar autonomic disturbance even when answering truthfully. Beginning in the 1950s, polygraph examiners began interspersing their accusatory or *relevant* questions with what were called "probable lie control" questions. These questions refer in a vague and general way to probable misdeeds in the examinee's past, e.g., "Have you ever told a lie to someone in authority?" The examiner induces the subject to deny each such question by implying that admitting such past misdeeds might lead to his or her being diagnosed as deceptive in respect to the relevant questions. In fact, however, it is assumed that a subject who can answer the relevant questions truthfully will be less disturbed by them than by the probable lie questions. A subject who shows stronger polygraph disturbance after the relevant than after the probable lie questions is classified as deceptive.

The test format just described is known as the Control Question Test or CQT, but it will be apparent

that the probable lie questions used in such tests are not controls in the scientific sense of that term. If Jones is accused of crime X and has reason to believe that he may be prosecuted, even convicted of crime X, then—whether Jones is innocent or guilty—it is not reasonable to suppose that a question of the form, "Prior to last year, did you ever tell a lie to someone in authority" can be expected to have stimulus value equal to the relevant questions of the form, "Did you commit crime X?" Therefore, it is to be expected that many innocent persons will display stronger polygraphic reactions to the relevant than to the "control" questions and become false-positive errors. This, in fact, is what the few validity studies of the CQT indicate, studies done of actual criminal suspects in real life circumstances.

Unfortunately, however, no wholly satisfactory study of lie test validity has ever been accomplished. Existing studies have used polygraph-induced confessions as their criterion of ground truth. One pragmatic virtue of the polygraph test is that guilty suspects, when told that the "impartial, scientific instrument has revealed" that they were lying, often confess their guilt. Validity studies have used other polygraphers to blindly score such confession-verified charts in order to estimate the sensitivity or true-positive validity of the CQT. However, even if we assume all such polygraph-induced confessions to be valid (although it is known that some innocent suspects, feeling that all hope is lost, will make false confessions), the charts thus rescored are not representative of all charts produced by guilty suspects. Some guilty suspects "fail" the test, are interrogated, but do not confess. Moreover, some guilty suspects will "pass" the test, therefore will not be interrogated, and will not confess.

Some polygraph-induced confessions will clear other suspects in the same crime. However, it almost always happens that the only innocent suspects thus cleared will be suspects tested before the guilty suspect, who then confesses. This is because, if the first-tested innocent suspect had failed his test, then it is most unlikely that the second and guilty suspect would ever have been tested, the examiner having assumed that the first test solved the crime. This means that charts of innocent suspects, cleared by another's confession, are almost always charts that examiners will classify as truthful. Therefore, as is true for the confession-verified deceptive charts, blind rescoring of confession-verified truthful charts will necessarily overestimate the validity of the CQT.

The only acceptable scientific assessment of CQT validity would require that all possible criminal suspects be given polygraph tests (e.g., by the FBI) and then the charts left unscored until the case had been definitively cleared on the basis of extra-polygraphic evidence. Then the relevant charts could be blindly scored and the results compared against ground truth thus determined independently of the polygraph results. No such proper study has ever been done of any method of polygraphic interrogation.

The CQT requires that the examiner be able to ask "control" questions, which the suspects will answer deceptively, and also that the examiner can persuade the suspect that he or she is in real jeopardy if it is concluded that he or she is answering these "control" questions deceptively. Because both of these assumptions are implausible, a different question format is being increasingly substituted for the CQT. This is the Directed-Lie Test or DLT. In this method, the "control" questions are similar to those used in the CQT, but they are presented differently. The examiner says, for example, "You must have told a lie sometime in your life, haven't you?" When the suspect admits to this, the examiner continues: "Well, I am going to ask you that question during this test and I want you to answer it 'No.' Then you and I will both know that your answer is not true, and I will be able to see from your polygraph chart what your response looks like when you lie." The DLT is scored in the same way as the CQT.

Thus, the DLT assumes that an innocent person will be more disturbed by a question about some minor misdeed that, on instruction, he or she answers deceptively, than by an accusatory question about some real crime of which he or she is suspected, which is answered truthfully. No good evidence of the validity of the DLT has ever been published in a refereed scientific journal, but the assumptions of the technique seem so implausible that one wonders why the FBI, the NSA, the Secret Service, and other government agencies continue to rely upon it.

Another problem with any form of lie detector test, in addition to the high rate of false-positive errors that result, is the fact that any guilty suspect who understands the simple principles on which these tests are based can easily employ countermeasures designed to beat the test. If one can recognize the "control" questions, then one need only to self-stimulate after answering such questions—by tightening one's toes or one's anal sphincter or by biting one's tongue—so as to augment one's polygraph reactions selectively to these questions. The examiner cannot detect such countermeasures and, if one is successful in augmenting the responses to the "control" questions so that

they average larger than one's responses to the relevant questions, then one will pass the test.

Polygraph tests such as the DLT are now used extensively by U.S. government agencies in order to screen job applicants and to periodically screen employees in sensitive positions, such as the scientists working in the weapons laboratories at Lawrence Livermore, Sandia, and Los Alamos. Due to the assumptions of these tests, as outlined earlier, one would expect that such screening would produce a very high proportion of false-positive errors. Because of this, it is a fact that many innocent and patriotic young applicants for positions with the FBI or other agencies have been turned away, their career aspirations smashed. However, when these methods are used to periodically screen employees in the nation's weapons laboratories, or military personnel involved in the nuclear missile system, it would obviously be unacceptable to "fail" 30–40% of them. Therefore, government examiners have wisely (but covertly) modified their scoring protocols in order to avoid this result. Instead of failing everyone who is more disturbed by the relevant than by the "control" questions, these periodic screening tests are scored by requiring a *much* stronger response to the relevant than to the "control" questions. *How much* stronger has never been made public. However, in as much as there is no acceptable scientific evidence that the probability of guilt increases monotonically with the difference between the amplitudes of the responses to the two types of question, there is no reason to suppose that this higher cutting score actually increases lie test validity.

C. Detecting Guilty Knowledge

The Guilty Knowledge Test or GKT is a method of polygraphic interrogation that attempts to determine whether the suspect possesses knowledge (e.g., of a crime scene or of military secrets) that he or she should not possess if innocent (e.g., of the crime or of espionage). Facts that an innocent person should not recognize are presented, verbally or pictorially, together with several alternative "facts" chosen to appear equally plausible to an innocent person in the form of a multiple-choice question. For example, "Where in the house did the murderer leave the body of the victim? Was it on the porch? In the kitchen? In a bedroom? In the basement? In the living room? In a hallway?" The GKT assumes that a suspect who possesses guilty knowledge will recognize the correct alternative and therefore will respond more strongly to it than to the others. If each GKT question contains five scoreable alternatives (the first alternative is always incorrect and left unscored), then the probability that an innocent suspect will chance to "hit" on a single question equals 0.2. However, the probability that an innocent suspect will chance to hit on each of 10 questions is 0.2^{10} or 1 in 10 million. Well-constructed GKTs have been shown to have high validity in experiments involving mock crimes and are currently used by police agencies in Israel and Japan, but not in the United States.

See Also the Following Articles

BIOFEEDBACK • CONVERSION DISORDERS AND SOMATOFORM DISORDERS • ELECTRICAL POTENTIALS • ELECTROENCEPHALOGRAPHY (EEG) • EVENT-RELATED ELECTROMAGNETIC RESPONSES • PSYCHONEUROENDOCRINOLOGY • PSYCHONEUROIMMUNOLOGY

Suggested Reading

Anderson, N. B., and Scott, P. A. (1999). Making the case for psychophysiology during the era of molecular biology. *Psychophysiology* **36,** 1–13.

Andreassi, J. L. (1995). *Psychophysiology: Human Behavior and Physiological Response*. 3rd ed. L. Erlbaum Associates, Hillsdale, NJ.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychol. Bull.* **91,** 276–292.

Cacioppo, J., and Tassinary, L. (1990). *Principles of Psychophysiology: Physical, Social and Inferential Elements*. Cambridge University Press, New York.

Christian, J. D., Morzorati, S., Nortin, J. A., Jr., William, C. J., O'Commor, S., and Li, T. K. (1996). Genetic analysis of the resting electroencephalographic power spectrum in human twins. *Psychophysiology* **33,** 584–591.

Fowles, D., Christie, M. J., Edelberg, R., Grings, W., Lykken, D. T., and Venables, P. (1981). Publication recommendations for electrodermal measurements. *Psychophysiology* **18,** 232–239.

Iacono, W. G. (1998). Identifying psychophysiological risk for psychopathology: Examples from substance abuse and schizophrenia research. *Psychophysiology* **35,** 621–637.

Lykken, D. (1998). *A Tremor in the Blood: Uses and Abuses of the Lie Detector*. Plenum Press, New York.

Lykken, D. T., Bouchard, T. J., Jr., McGue, M., and Tellegen, A. (1992). Emergenesis: Genetic traits that may not fun in families. *Am. Psychol.* **47,** 1565–1577.

Rugg, M. D., and Coles, M. G. H. (Eds.). (1995). *Electrophysiology of Mind: Event-Related Brain Potentials and Cognition*. Oxford University Press, Oxford, UK.

Reading Disorders, Developmental

VIRGINIA A. MANN
University of California, Irvine

GLOSSARY

decoding The act of converting a written word into its spoken counterpart.

dyslexia A specific learning disability of neurological origin that hinders the ability to learn to read fluently and accurately. It affects approximately 4% of the population and does not imply low intelligence or poor educational potential. It is also independent of race and social background.

morphophonological orthography An orthography such as English or French that preserves both the phonemes and morphemes of spoken words.

orthography A symbol system used to transcribe or represent a spoken language. Logographies (i.e., Chinese and Japanese Kanji) use symbols to represent morphemes, syllabaries (i.e., Hebrew and Japanese Kana) use symbols to represent syllables, and alphabets (i.e., English and Spanish) use symbols to represent phonemes.

phoneme The minimal segment of sound that distinguishes two words or syllables (e.g., "cat" and "cap" differ in their third phoneme). Phonemes are the units of language represented by the letters of an alphabet.

phoneme awareness Sensitivity to the phoneme-sized units in spoken words, measured as the ability to analyze (e.g., do "cap" and "tap" start with the same sound?) or manipulate (e.g., say "cat" without the first sound) words in terms of the phonemes they contain.

syntax The grammar of language; the system of rules by which words are combined and moved to form phrases and sentences.

working memory A temporary memory system in which material is held for a short time and analyzed. Different working memory systems exist for different types of information and they use different types of codes. The working memory for language uses a phonological code that represents words in terms of the syllables and phonemes they contain.

I. INTRODUCTION: THE LINK BETWEEN READING PROBLEMS AND LANGUAGE PROBLEMS

What makes a poor reader a poor reader? Why do some children fail to learn to read when others succeed and even excel? Although learning to read is a task that most children accomplish relatively early in their education, it poses a specific difficulty for approximately 4% children who may be labeled as "dyslexic" or "reading disabled." The syndrome of dyslexia is now widely recognized as a specific learning disability of neurological origin that does not imply low intelligence or poor educational potential and which is independent of race and social background.

Research during the past few decades has increasingly identified the neurological and behavioral factors that are responsible for difficulties in learning to read. Psychologists, educators, and neurologists have all, in one way or another, tried to identify the basis of early reading difficulty, and their efforts have been guided implicitly, if not explicitly, by some basic assumptions about skilled reading and its demands on perceptual and cognitive abilities. Those researchers who have

succeeded were able to do so because they had a framework in which to interpret their results as well as a meticulous methodology that allowed them to separate spurious observations from replicable generalities. In recent years, the most successful approaches to the problem of early reading disability have been guided by the assumption that reading is first and foremost a language skill. Thus, a primary goal of this article is to review both the theoretical underpinnings and the empirical research that show how reading is tied to the integrity of language skills.

II. PRIOR ACCOUNTS OF POOR READING

A time-honored means of discovering the problems that underlie reading disorders is to compare children who differ in reading skill. This approach has a long history with many false turns, and has only recently converged on the link between reading and language problems. This section reviews a sample of the most influential accounts as a preface to current research.

Learning to read is a specific example of a complex learning task that correlates about 0.6 with IQ. However, a low IQ cannot be the sole basis of reading problems since there are children who have low reading ability but average intelligence, just as there are children of low intelligence who are nonetheless able to read. Children who possess a seemingly adequate IQ (typically 90 or higher) but nonetheless encounter reading problems are said to have a specific reading difficulty. Their actual reading ability lags between 1 and 2 years behind that which is predicted on the basis of their age, IQ, and social standing. For these children, something other than general intelligence must be the primary ''cause'' of many instances of poor reading.

In attempting to discover the cause of early reading problems, many early theories were biased by a popular assumption that influenced psychologists and educators alike. It was commonly believed that reading is first and foremost a complex visual skill that demands differentiation and recognition of visual stimuli. Owing to this view, models of skilled reading have often been biased toward clarification of how readers see and recognize the various letter and word shapes, causing many studies of the cause of poor reading to conclude that early reading difficulty is due to some problem in the visual domain. Recently, however, visual theories of reading disability have become less popular because it seems that, at best, only a few of the children who are poor readers actually suffer from perceptual malfunctions that somehow prevent recognition, differentiation, or memory of visual forms. In short, many visual perceptual skills cannot very reliably distinguish children who differ in reading ability, so visual problems may not underlie most reading problems. The only exception involves recent evidence that poor readers tend to have impaired contrast sensitivity to transient (i.e., flickering) low-luminance visual stimuli. Although the existence of the impairment is increasingly clear, its causative role in reading disability is not. The flickering conditions of low-illumination under which dyslexic individuals ''see'' less well are not characteristic of printed text read under normal conditions. It can thus be argued that this particular visual impairment may be a concomitant factor in reading problems but not necessarily a causal one.

Additionally, two other pieces of evidence show just how unfair it is to attribute the majority of early reading difficulties to visual deficits. First, 5- and 6-year-old children identified as having deficient visual perception and/or visuomotor coordination skills show no more instances of reading difficulty at age 8 and 9 than do matched controls who possess no such deficits. Second, although it is true that most young children tend to confuse spatially reversible letters such as ''b,'' ''d,'' ''p,'' and ''g'' until they are 7 or 8 years old, letter and sequence reversals actually account for only a small proportion of the reading errors that are made by children in this age range. Even children who have been formally diagnosed as dyslexic make relatively few letter and sequence reversal errors, as shown by I. Liberman and D. Shankweiler.

Theories that place primary emphasis on ''cross-modal'' integration have also been popular at one time or another. One particular misconception posited that reading involved translating visual information into auditory information, and that dysfunctions of this cross-modal match served as the source of reading problems. However, when investigators carefully examined the behavior of skilled readers, they realized that the translation of stimuli did not proceed directly from the visual to auditory domain. Instead, visual information was first translated into an abstract linguistic code. Further studies made it clear that poor readers tended to make inefficient use of an abstract linguistic code in a variety of circumstances, including cross-modal integration. Consequently, researchers noted that visual-auditory integration problems were often coupled with auditory–auditory problems and, in some cases, visual-visual ones as well. Thus, the poor readers' problems with visual–auditory integration

have come to be viewed as one of the many consequences of a more general linguistic coding problem that impairs short-term or "working" memory and negatively affects integration both within and between different sensory modalities.

Other theories have suffered from similar attempts to explain observations about poor readers in terms that are somehow too general. For example, certain theories were preoccupied with the fact that reading involves remembering an ordered sequence of letters in a word, words in a sentence, etc. Hence, it was suggested that poor sequential order memory or poor short-term memory (STM), in general, might be a cause of poor reading. Other observations about the specific pattern of deficits that are characteristic of poor readers prompt a need for refinements in the methodology/logical foundation on which STM theories of dyslexia are based. For example, good and poor readers do not differ on all tasks that require temporary memory for items or their order. Likewise, good and poor beginning readers are equivalent, for example, in their ability to remember faces or visual stimuli that cannot readily be assigned verbal labels. Only when the to-be-remembered stimuli can be linguistically coded do children who are poor readers consistently fail to perform as well as good readers.

Various other general or visual accounts of reading disability have been offered in the literature. These tend to be inadequate because they fail to explain why poor readers often do as well as good readers on nonlinguistic tasks, but lag behind good readers in performance on many linguistic tasks. For the sake of brevity, such general accounts will not be discussed here. Instead, I turn to the more positive task of reviewing evidence that links language and reading problems. I begin with a justification of why a language-based approach to reading is the most appropriate approach to take.

III. WHY SPOKEN LANGUAGE IS CRITICAL TO READERS

Two insights underscore the role of language problems in poor reading. One stems from the fact that writing systems function as written language by transcribing the units of spoken language. Another concerns the active role of language skills in skilled reading. We often think of reading as a visual skill, but it is so much more. For readers to successfully perceive, recognize, remember, and interpret various letters and the words, sentences, etc. that they form, readers must first map their written language onto their spoken language. Spoken language comes first in both the history of the individual and the history of our species; furthermore, spoken language is far more universal and far more natural than written languages. The latter constitute a human invention. Thus, it should come as no surprise that written language makes use of some of the very same language skills that allow us to be speakers and hearers of our language.

A. How Writing Systems Represent Language

Writing systems (or orthographies) are symbol systems by which we are able to write spoken language. They "write" language by representing or "transcribing" the units of spoken language. Since most of the readers of this article will be interested in children who are having problems with an alphabetic writing system (English being a case in point), I focus on alphabets and on the English alphabet in particular. All writing systems represent units of a spoken language, but there are differences between writing systems and even between alphabets that turn on which type of linguistic units are being represented. Ideographies, such as American Indian petroglyphs or the universal set of road signs that a driver encounters in a daily commute, represent language at the level of ideas, and logographies such as the Chinese writing system and Japanese Kanji represent units called morphemes. Syllabaries such as Hebrew and Japanese Kana represent syllables. Alphabets represent units called phonemes. Each of the different types of writing systems makes different demands on the beginning reader because of the differences in the type of the unit they transcribe.

It is important that the would-be reader of a writing system have some appreciation of the existence of whatever units that writing system happens to transcribe. Otherwise, it will be difficult to understand how written words relate to their spoken language counterparts. Since alphabets represent phonemes, someone who wishes to learn how an alphabet functions should be sensitive to the fact that spoken language can be broken down into phonemes, the minimal units of sound that are referred to as consonants and vowels. Section VI presents evidence that this sensitivity, referred to as phoneme awareness, is a very important trait of successful beginning readers. It is also a major problem for many young children and for poor readers, in particular.

More precisely, the English alphabet is a less than perfect one. It does not provide the consistent

one-to-one mapping of letter to phonemes that one finds in Spanish or German, for example. Rather, it provides a ''deeper,'' more abstract level of representation that goes beyond phonemes and sounds to the morphemes and meanings of words. As such, it has been referred to as a morphophonological transcription because it combines the transcription of morphemes and phonemes. Morphophonological transcription corresponds not so much to the consonants and vowels that speakers and hearers think they pronounce and perceive as much as it does to the way theoretical linguistics assumes that words are abstractly represented in the ideal speaker–hearer's mental dictionary or lexicon. According to this view of the mental lexicon, words are represented in terms of the basic units of meaning that we refer to as morphemes as well as in terms of phonemes. When they produce or perceive language, speakers and hearers convert the morphophonological representations of the words in their lexicon to less abstract, phonetic representations by using an ordered series of phonological rules that alter, insert, or delete phonemes.

If a writing system correctly represents words in a deep manner that approximates the mental lexicon, it will sometimes fail to represent the phonetic representations with which we are most familiar. Witness the spellings of words in pairs such as ''*atom*'' vs ''*atomic*,'' ''*heal*'' vs ''*health*,'' and ''*relate*'' vs ''*relation*.'' In each of these pairs of words, the similar spelling of the base and derived form captures the relatedness of their meanings. However, the similar spelling comes at the cost of having a letter or letter sequence represent different phonemes in different words. Preservation of common roots and of common word ancestries is one of the reasons why the English vowel system is so complicated and one of the reasons why English uses nearly 120 spelling patterns for 40 phonemes. It is also the source of the notorious evil triplets, ''*to*''–''*two*''–''*too*'' and ''*their*''–''*there*''–''*they're*.'' The different spellings of these homonyms reflect their different meanings, the common pronunciation is what makes their usage so difficult. The important point to be remembered is that the English alphabet represents phonemes and morphemes and there is a certain trade-off between the two types of representation.

Why should English use a writing system that transcribes both phonemes and morphemes? One general benefit of systems that transcribe phonemes is that they are economical. The transcription of phonemes greatly reduces the number of individual characters that a child must learn to recognize and reproduce. Consider the ease of learning the 26 letters used in the English alphabet as opposed to learning the 2000–3000 characters needed to read a newspaper written in the Chinese logography. Aside from being economical, phoneme transcription is highly productive. It allows for a highly rule-governed relationship between written and spoken words. Knowledge of the rules that relate written words and spoken words rests on the rules that relate letter sequences to their pronunciation. This allows the reader to read not only highly familiar words but also less familiar ones such as ''skiff'' and even nonsense words such as ''ifts'' or ''polypluckable'' that are lawfully constructed. A skilled reader of the Chinese logography must have memorized thousands of distinct characters, and even then may encounter difficulty in reading a new word. On the other hand, skilled readers of English need know only a limited set of phoneme to grapheme correspondences and the phonological rules of their spoken language in order to decode most words (and any phonologically plausible nonwords such as ''bliggle'').

There is also a benefit to the English morphophonological system as opposed to the more transparent alphabetic systems used in Spanish and German. By transcribing a ''deep,'' relatively abstract level of phonological structure where units of meaning are represented, the English writing system helps convey cues to meaning as well as to sound. If one recognizes ''heal'' in ''health,'' ''atom'' in ''atomic,'' or ''relate'' in ''relation,'' this can facilitate the recovery of those words' meanings. Another advantage to the transcription of morphemes is that it can avoid the need to create different spelling patterns for people who speak with different accents. Were English to use a more shallow alphabet, speakers from Boston would spell ''cot'' and ''cart'' the same way, speakers from the South would spell ''pen'' and ''pin'' the same, and speakers from Brooklyn would spell ''oil'' and ''earl'' the same—imagine all the inconveniences this could cause. A final advantage of the English alphabet is that its morphophonological spelling can disambiguate homonyms. Indeed, the key to disentangling ''there,'' ''their,'' and ''they're'' depends on their respective meanings and the context in which they occur in both spoken and written language.

This is not to imply that there is anything inherently undesirable about reading a syllabary or a pure logography. Ultimately, the utility of a given orthography rests on the nature of the spoken language it transcribes. For example, a logography is appropriate for Chinese because it allows people to read the same

text even though they cannot understand each other's speech. Likewise, for Japanese, the Kana syllabaries are quite well suited to the approximately 100 syllables in the Japanese language. English, however, has a less profound dialectical variation than Chinese, and the English language employs more than 1000 syllables. Hence, an alphabet is appropriate, and it would be less efficient and even a disservice to present the English writing system otherwise. However, this will require that a child who is learning to read possess two things: language processing skills and phonological sophistication.

B. Phoneme Awareness: A Requirement of Alphabetic Systems

Consider that almost all normal children "know" spoken English when they start to read. If reading is a language skill, why is it that not every speaker of English automatically becomes a successful beginning reader? The answer to this question is that knowing spoken English is a necessary but not sufficient requirement for skilled reading. Would-be readers must go one step further than merely being a speaker/hearer of their language: They must be able to consciously analyze and manipulate the units that their writing system represents. For readers of an alphabet, this means that they must be aware of phonemes. Phoneme awareness, as first discussed by Mattingly and later developed by many others, is not something that we use in the normal activities of speaking and hearing. We use it in certain secondary language activities such as appreciating verse (i.e., alliteration), making jokes (i.e., "Where do you leave your dog? In a barking lot"), and talking in secret languages (i.e., pig Latin). Such activities require that we consciously compare and manipulate the consonants and vowels that comprise spoken words.

One problem with the term phoneme awareness is that it is often used interchangeably with several other terms, *phonemic awareness*, *phonological awareness*, *metalinguistic awareness*, and *linguistic awareness*. By using the term phoneme awareness (or phonemic awareness) we confine the issue to sensitivity about phonemes. Phonological awareness could also include sensitivity to syllables, morphemes, and the phonological rules that operate on them; linguistic awareness and metalinguistic awareness would further include sensitivity to syntax (i.e., grammar), semantics (i.e., meaning), and their rules. To date, awareness of phonemes has been most often studied and deficient phoneme awareness is a major factor in reading disorders.

C. Language Skills That Skilled Readers Use

Considerations about the way in which the English alphabet transcribes language offer one form of deductive evidence about the importance of spoken language skills to reading. A second source of evidence comes from studies of skilled readers. These studies show a clear involvement of certain spoken language processes in the skilled reading of words, sentences, and paragraphs. Furthermore, these studies have shown that reading is quite parasitic on spoken language processes and have inspired subsequent investigations that contrast good and poor beginning readers.

The question of whether or not written words must be recoded into some type of silent speech has been a topic in much of the research on the psychology of skilled reading. It has especially preoccupied those who study the processes that facilitate word recognition, (a.k.a. lexical access or word perception). In some circumstances, successful word recognition may not require silent speech. Some words may be directly perceived as visual units instead of being decoded into a string of phonemes. However, there is also clear evidence implicating at least some speech code involvement in word perception, making many psychologists favor models in which both phonetic and visual access occur in parallel. Some believe that access via the speech code or phonetic route may be most heavily used in the case of less frequent words and unfamiliar ones, and that the visual or whole word route is most important for very familiar words and words with irregular spelling patterns. Still others regard the phonetic route as playing an early, dominant role in all lexical access.

Regardless of how a word is recognized, its subsequent processing will involve the use of speech or phonological coding. It may not be necessary to recode print into a speech code in the process of gaining access to the mental lexicon. Lexical access could occur in terms of morphemes or syllables. At some level, however, phonological recoding clearly must occur; otherwise, reading phrases, sentences, and paragraphs would not be possible. This is because, from the point of word perception onward, the involvement of speech processes in reading is most clear. First, there is considerable evidence that temporary or working memory for written material involves recoding the

material into some kind of silent speech or phonological representation. This type of representation is used whether the task requires temporary memory for isolated letters, printed nonsense syllables, or printed words. In all these cases, both the nature of the errors that subjects make in recalling such material and the experimental manipulations that help or hinder their memory performance have indicated that a phonological representation is being used. That is, subjects are remembering the items in terms of the consonants, vowels, and syllables that form the name of each item rather than the visual shape of the letters, the shape of the words, etc. Furthermore, subjects appear to rely on phonological representation when they are required to comprehend sentences written in either alphabetic or logographic orthographies. This is one reason why we may observe such significantly high correlations between reading and listening comprehension across a variety of languages and orthographies, including English, Japanese, and Chinese.

Thus, it is apparent that regardless of the way in which the reader recognizes each word, the processes involved in reading sentences and paragraphs place certain obvious demands on temporary memory, and temporary memory for language appears to make use of phonological representation in short-term memory. In Section V, it will be shown that problems with the use of phonological representation in working memory are often found among poor beginning readers.

IV. TWO TYPES OF LANGUAGE SKILLS THAT ARE ESSENTIAL TO BEGINNING READERS

What skills does a child need in order to learn to read well? Obviously, would-be readers need to possess the visual skills that allow them to differentiate and remember various letter shapes, but this is not a very prevalent source of difficulty. The deficits that tend to trouble disabled readers are classified into two areas of language skill: language processing and phoneme awareness. Readers also need language processing skills in order to be able to perceive and recognize their teacher's spoken words and sentences as well as to combine written words into phrases, sentences, and paragraphs, as discussed in Section III. They need to possess phoneme awareness if they are to make any real sense of the way in which the alphabet works.

In principle, beginning readers should possess language processing skills at four different levels. First, they need the speech perception skills that make it possible to distinguish the words of their vocabulary so that they can hear the difference between "cat" and "hat," for example. They also need vocabulary skills to understand what words mean and morphological skills to understand how words are built from roots, suffixes, and prefixes (although they need not necessarily possess fully mature morphophonological representations in their lexicons since the experience of reading, in and of itself, serves to stimulate and further development). Beginning readers should also have an adequate working memory for language since this is not only critical to skilled readers but also supports retention of sufficient words to understand the sentences and instructions that their teacher is saying. Finally, they should be able to recover the syntactic and semantic structure of phrases and sentences (although their mastery of these aspects of language, like their mastery of phonology, may be facilitated by the experience of reading).

Language processing skills, however, are only one aspect of the language skills needed by would-be readers of English. As noted in Section III, successful readers need to be sensitive to phonemes; otherwise, the alphabet will make no sense as a transcription of spoken English. The achievement of phoneme awareness proves an initial obstacle for most young children and an enduring one for those who become disabled readers.

Why should this be the case? One reason is that phonemes are quite abstract units of language, more so than either words or syllables. As noted by Alvin Liberman and colleagues at Haskins Laboratories, humans reflexively and unconsciously perceive phonemes when listening to the speech stream because they possess a neurophysiology uniquely and elegantly adapted to that purpose. However, phonemes cannot be mechanically separated like beads on a string. Most of them cannot be produced in isolation as can syllables and words. This was a groundbreaking discovery of the speech research team at Haskins Laboratories, headed by A. Liberman and Cooper. There are some very interesting indications that infants may distinguish phonemes and preschool-aged children most certainly employ phonological representation when holding linguistic material in short-term memory. However, these are automatic, tacit aspects of language processing ability. The child who tacitly knows his or her language well enough to hear that "*cat*" and "*hat*" are different words may perceive and remember phonemes while being blissfully unaware of the fact that it is the initial phoneme that differs in "*cat*" and "*hat*." Young children can be as unaware of

phonemes as you and I are unaware of the rods and cones that allow us to see.

The problem with using written language is that the tacit perception and remembering of phonemes must become an explicit ability to analyze and manipulate. Successful beginning readers must know not only the difference between words such as "*cat*" and "*hat*," but also how to hold these words in memory. They must also possess the awareness of phonemes that allows them to appreciate the fact that "*cat*" and "*hat*" differ in one phoneme, namely the first, and share a final phoneme that is the initial one in "*top*." Otherwise, the alphabet will remain a mystery to them, and its virtues will be unrealized.

V. DEFICIENT LANGUAGE PROCESSING AS A CAUSATIVE FACTOR

Section II mentioned several studies that demonstrate how good and poor readers differ in their performance on certain linguistic tasks but not in their performance on comparably demanding nonlinguistic ones. That evidence receives further support from a consideration of the frequency of reading difficulties in children with various sorts of handicaps. Children deficient in visual–perceptual and/or visual–motor skills do not encounter reading difficulty any more frequently than matched controls, but speech- and language-retarded children encounter reading problems at least six times more often than do controls. Deaf individuals may have the greatest difficulties of all. However, we can ask whether there is a more fine-grained analysis of the language problems found among poor readers. Are some areas of language skill more problematic than others? Considered broadly, the language disabilities that tend to be found among poor beginning readers can be classified into two categories of language processing and phoneme awareness. Next, I examine the evidence within each area.

Without spoken English, there would be nothing for the English orthography to transcribe; the well-known difficulties of prelingually deaf readers attest to the importance of spoken language skills for successful reading. However, deaf children are not the only ones for whom deficient language abilities are a cause of reading problems. As discussed later, many of the hearing children who are poor readers also suffer from spoken language problems. Although their problems are considerably subtler than those of the deaf, they are no less critical.

Since the mid-1970s, there has been considerable activity in the psychology of early reading problems, and many studies have uncovered some link between difficulties in learning to read and difficulties with some aspect of spoken language processing. This connection is clearly established beyond question, not only in English but also in Swedish, Hebrew, and Chinese. In the case of English, there have also been considerable attempts to more precisely specify the nature of the language problems that typify poor beginning readers. These attempts can be organized in terms of the four levels of language processing that were identified in Section IV as being important to beginning reading: speech perception, vocabulary skills, linguistic short-term memory, and syntax and semantics.

A. Speech Perception

The possibility that some aspect of speech perception might be a special problem for poor readers has come from a variety of sources. At least a subgroup of poor readers may have phonological categories that are more prone to disruption during perception in noise or gating conditions that present parts of words. For example, Brady and colleagues report that poor readers are less able to identify spoken words under a "noisy" listening condition. They also present evidence that this lack of ability is not simply a hearing problem: The poor readers had audiometry scores in the normal ranges and were just as capable as good readers when the stimuli were environmental sounds instead of speech sounds. Their problem was limited to the perception of speech.

Other studies have investigated the categorical perception of synthetic speech stimuli by good and poor beginning readers. These studies report that categorical perception is evident in both groups of subjects, but a subgroup of poor readers either fail to meet the level of intercategory discrimination predicted on the basis of their identification responses or fail to give as consistent a pattern of identification responses. These findings have also been interpreted as reflecting deficient speech perception processes on the part of poor readers. Some debate surrounds the possibility that the speech perception difficulties of poor readers may be explained by a more general problem with the perception of rapidly changing auditory events. Whatever the basis of the impairment, the point is that when listening conditions are noisy or otherwise degraded, at least some poor readers do not perceive speech as well as do good readers.

B. Vocabulary Skills

There are several indications that reading ability is related to vocabulary skills, depending on how reading ability is measured and on what type of vocabulary skill is at issue. Recognition vocabulary tests require the child to point to a picture that illustrates a word. These sometimes relate to early reading ability, especially if measured in terms of comprehension. Naming or productive vocabulary tests require the child to produce the word that a picture illustrates and these are more clearly linked to both decoding and comprehension. Tests of continuous naming (sometimes called rapid automatized naming) require children to name a series of repeating objects, letters, or colors. These tests have demonstrated that children who are poor readers take longer to name the series than do good readers.

A causal link between naming problems and reading problems is indicated by the discovery that performance on naming tests can predict future reading ability. Problems with rapid letter recognition and with naming also play a more prolonged role in the reading of severely impaired readers. Since letter naming predicts future reading ability more consistently than current reading ability predicts future letter naming ability, it is likely that something above and beyond a lack of educational or home literacy experience is preventing reading impaired children from naming the letter names as fast as other children can. That "something" may entail problems with productive vocabulary skills.

Another pertinent piece of evidence about the vocabulary problems of poor readers is that children who perform poorly on a decoding test are particularly prone to difficulties in producing low-frequency and polysyllabic names. Such children may possess less phonologically complete lexical representations than good readers. However, because speech rate may not reliably predict naming problems, researchers such as Wolf and Bowers have come to regard difficulties with naming speed as reflections of some deficiencies in the precision and timing of rapid processes.

C. Working Memory

The observation that poor readers perform worse than good readers on a variety of short-term memory tests has given rise to one of the more fruitful lines of research in the field. It has often been noted that poor readers tend to perform worse on the digit span test and are deficient in their ability to recall ordered strings of letters, nonsense syllables, or words, whether the stimuli are presented by ear or by eye. Additionally, poor readers may not recall the words of spoken sentences as accurately as do good readers. Evidence that these differences are not merely a consequence of differences in reading ability comes from longitudinal studies that have shown that problems with working memory can precede the attainment of reading ability and may actually serve to presage future reading problems.

To explain this pattern of results, researchers employed the aforementioned investigations into short-term memory (namely the research that indicated that linguistic materials such as letters and words are held in short-term memory through use of phonological representation, see Section III). Observations about the errors that poor readers make and the conditions under which they succeed or fail have led to the postulation that poor readers—and children who are likely to become poor readers—are for some reason less able to use phoneme and syllable structure as a means of holding material in short-term memory. They do not avoid phonological representation altogether. They merely use it less proficiently, a problem that may persist even into adulthood.

D. Syntax and Semantics

Do poor readers have a problem with the syntax (the grammar) and the semantics (the meaning) of language in addition to their problems with speech perception, vocabulary, and the use of phonological structure in short-term memory? The observation that poor readers cannot repeat sentences as well as good readers has prompted some obvious questions regarding these two higher level language skills and their involvement in reading problems. However, researchers have found no conclusive evidence supporting the notion that syntactic and semantic problems arise as a direct consequence of reading disability. Although there is an accumulating body of evidence that good and poor readers differ in the ability to both repeat and comprehend spoken sentences, these problems seem to reflect difficulties with using phonological coding in working memory rather than a form of agrammatism. As discussed by Shankweiler and colleagues, it appears that poor readers are no less sensitive to syntactic structure than good readers, but they fail to understand because they cannot hold an adequate representation of the sentence in short-term memory.

Another higher level language task that proves problematic for poor readers is the cloze task, in which subjects must supply a missing word. My colleagues and I have found that when we ask children to choose among derivational forms that complete a sentence, as in "He was blinded by the…: bright, brighten, brightly, brightness," poor readers perform worse than normal readers. Moreover, this is true whether they are reading the sentences or hearing them aloud, and whether the word that fills in the blank is a real word or an appropriately derived nonsense word such as "froodness." Here, the poor readers' difficulty in distinguishing among noun-, verb-, adverb-, and adjective-forming suffixes may be attributed to a problem with derivational morphology rather than a syntactic problem. Given the discussion in Section III regarding the transcription of morphemes by English orthography, the fact that reading problems are linked to problems with morphology should come as no surprise. What we find particularly striking in our own research is the fact that morphological skills in both written and spoken language relate to both the comprehension and decoding of written language.

Although it is clear that poor readers do have sentence comprehension problems, there is little reason to think that their difficulties reflect a problem with the syntax of language. Problems with working memory morphology seem a more likely source of the difficulty. Moreover, such syntactic differences that have been evident among good and poor readers are relatively subtle, with poor readers performing only as well as children younger than the good readers. These syntactic problems could be either the cause of reading difficulty or a consequence of different amounts of reading experience. Regarding the question of semantic impairments among poor readers, there is no reason to presume any real deviance exists. If anything, poor readers place greater reliance on semantic context and semantic representation than do good readers, perhaps to compensate for their other language deficits.

VI. DEFICIENT PHONEME AWARENESS AS A CAUSATIVE FACTOR

Inadequacies in phonetic perception, naming ability, and working memory for language are only part of the story behind reading failure. As noted in Section III, successful readers of the alphabet must go beyond these tacit language processing abilities to achieve an explicit awareness of phonemes. Here, I discuss studies concerned with the pertinence of phonological sophistication to success in learning to read an alphabetic orthography.

A. Evidence from the Analysis of Reading Errors

The errors that a person makes can be informative about the difficulties that produce those errors, and oral reading errors are a particularly important source of evidence about the cause of reading problems. I. Liberman and colleagues' consideration of these errors has shown that a lack of phoneme awareness is responsible for making beginning reading difficult for all young children, including dyslexic ones. Such errors do not tend to involve visual confusions and letter or sequence reversals to any appreciable degree. Instead, they reflect a problem with recovering and integrating the phonological information that letter sequences convey. Hence, children often tend to correctly pronounce the first letter in a word but experience increasing difficulty with subsequent letters and exhibit a particular problem with vowels as opposed to consonants. These decoding problems stem from a problem with understanding how the spoken word is transcribed by a string of letters representing the sequence of phonemes that form the word. In other words, they stem from deficient phoneme awareness.

B. Evidence from Tasks That Measure Awareness Directly

Most of the studies linking reading ability to phoneme awareness have concerned tasks that measure awareness directly. These tasks require children to play language games that manipulate the phonemes within a word in one way or another: counting them, deleting them, choosing words that contain the same phoneme, etc. The use of these tasks has revealed that phoneme awareness develops later than phonetic perception and the use of phonological representation, and it remains a chronic problem for those individuals who are poor readers. Research has shown that the awareness of phonemes and syllables develops considerably between the ages of 4 and 7. However, a child's phoneme awareness undergoes a slower and less natural development than the awareness of syllables. Both types of awareness markedly improve at just the age when children are learning to read, but phoneme awareness is more critically dependent on instruction in an alphabetic orthography.

Numerous experiments involving widely diverse subjects, school systems, and measurement devices have shown a strong positive correlation between a lack of awareness about phonemes and current problems in learning to read. There is also evidence that lack of awareness about syllables and such subsyllabic units as phonemes, onsets, and rhymes is associated with reading disability. Finally, studies of kindergarten children provide evidence that problems with phoneme segmentation, onset-rhyme segmentation, and syllable segmentation can all presage future reading difficulty. For example, we found that 85% of a population of kindergarten children who went on to become good readers in the first grade correctly counted the number of syllables in spoken words, whereas only 17% of the future poor readers could do so. Stanovich and colleagues found that a kindergarten battery of tests that assessed phoneme awareness accounted for 66% of the variance in children's first-grade reading ability.

C. Morpheme Awareness: Another Problem for Poor Readers

Although deficient phoneme awareness is the most reliable attribute of disabled reading in the early elementary grades, deficient morpheme awareness begins to play an increasingly important role as children reach the later grades. Several researchers have shown that disabled readers in later grades have problems with the morphological aspects of both written and spoken language. The difficulties of the disabled readers are seen in spelling errors, in performance on cloze tasks, and in vocabulary exercises such as defining a word or giving its derivational forms. The difficulties are linked to poor phoneme awareness and poor vocabulary, but we have found that, after the third grade, they play a role of their own when phoneme awareness and vocabulary are controlled. The link between morphological abilities and reading ability should come as no surprise given the morphophonological nature of the English alphabet (see Section III).

VII. EXPLANATIONS OF THE LANGUAGE PROBLEM: NEUROLOGICAL AND EXPERIENTIAL

Both theoretical and practical matters are at stake when we ask why poor readers are lacking in certain language skills. Much of the available literature on the causes of language processing problems is centered on what can be referred to as neurological causes. These involve factors that are somehow intrinsic to the child, owing to his or her brain structure. Problems with phoneme awareness have also been explained in these terms, but it has been more common to attribute problems in this area to an insufficient exposure to instruction in the use of an alphabetic writing system.

A. Neurological Accounts

Why should a neurological account be considered as a cause of reading disorder? Perhaps the strongest reason is that reading disability appears to have a genetic basis. It consistently runs in certain families and has been linked to certain autosomal-recessive chromosomal abnormalities. In these families, the nonaffected members of the family tend to show a higher incidence of language and speech problems. It has a higher concordance rate in identical twins.

Five neuropsychological theories are representative of the types of constitutional explanations currently favored. They are offered as a representative but by no means exhaustive summary of the literature in this area and need not be taken as mutually exclusive. In fact, current trends in research lean toward an emphasis on structural and functional abnormalities in the left cerebral hemisphere. This corresponds well with the language disabilities described previously.

Orton offered one of the first neuropsychological accounts in his theory of strephosymbolia. In this theory, mirror reversals (which Orton erroneously thought to be the predominant symptom of reading disability) were attributed to insufficiently developed cerebral dominance. This insufficiency further manifests itself, according to Orton, in such abnormalities of lateral preference as mixed dominance. Wealthy in its own heuristic value, Orton's theory inspired considerable research as well as criticism. On the one hand, it has been falsified by findings that reading difficulty is not associated with any particular pattern of handedness, eyedness, or footedness. It has also motivated a number of studies of cerebral lateralization for language processing among good and poor readers, though with mixed results. Some such studies have provided evidence that poor readers show a reversal of the normal anatomical asymmetries between the left and right hemispheres, in conjunction with a lower verbal IQ. Others have reported that poor readers may show a lack of cerebral dominance for language processing. To date, research has yielded only a weak association between abnormal

lateralization and poor reading since not all individuals who display abnormal cerebral lateralization are poor readers. It must also be recognized that several other studies have failed to find that good and poor readers differ in the extent or direction of the lateralization for language processing.

The data are not particularly supportive of Orton's theory about incomplete cerebral dominance as an explanation of reading difficulty. However, Orton may still have been correct in the spirit, if not the letter, of his explanation. If we accept the left hemisphere to be the mediator of language processing (in the majority of individuals), and if we accept that language processes are deficient among poor readers, then certainly we may suppose that some anatomical or neurochemical abnormality of the left hemisphere is involved in early reading difficulty. This is the position taken by three contemporary researchers: Geschwind and Galaburda in their theory of cerebral lateralization; Pugh and colleagues in their emphasis on the integrity of posterior left hemisphere circuits; and Temple and colleagues, who focus on myelination and functional integrity within the left hemisphere's temporoparietal regions. Geschwind and Galaburda attempted to explain developmental dyslexia as a consequence of slowed development of the left hemisphere, owing early exposure to testosterone. They point to the greater instance of reading problems in males and evidence from a variety of sources, including autopsies of the brains of several adult dyslexics and population studies that indicate a certain profile of disabilities (language problems), abilities (spatial skills), and other traits (left handedness and allergies) that distinguish the population of dyslexics from the general population. However, some evidence, challenges this theory. There are claims that the ratio of male to female dyslexics is actually quite close to 1:1 and claims that handedness, immune disorders, and dyslexia are not always correlated.

An emphasis on the integrity of the left hemisphere is shared by recent functional magnetic regonance imaging and DTI (diffusion tensor imaging) studies of dyslexic individuals. Based on findings of reduced activation and reduced functional connectivity during reading tasks, Pugh and colleagues in the Yale–Haskins reading group suggest that fluent word reading is related to the integrity of the dorsal (temporal parietal) circuit and the ventral (occipital–temporal) circuit of the left hemisphere. As a compensatory consequence, disabled readers demonstrate heightened reliance on left inferior frontal and right hemisphere posterior regions. The initial cause of the disabled readers' problems is said to lie in a disruption of the integrity of the temporoparietal system, although its genetic or teratogenic basis remains unspecified. As a consequence of this disruption, phonological and lexical semantic features of words are not well organized and integrated. This slows the acquisition of rule-based analyses of the relations between orthography, phonology, morphology, and meaning and hinders basic decoding and word analysis. As a consequence, fluent word identification by the left occipitotemporal region suffers and other areas (left frontal and right posterior circuits) take over inappropriately.

Temple and colleagues have a related but slightly different perspective. One observation concerned the use of diffusion tensor imaging to study the integrity of white matter in dyslexic individuals. This relatively new technique, which measures the diffusion of water molecules, suggests that dyslexic adults show evidence of decreased myelination within anterior-posterior circuits of the left hemisphere, and that the extent of decrease correlates with measures of decoding ability. Thus, there seems to be impairment in the strength of communication between cortical areas. These researchers have also documented that dyslexic children show a disruption in the left temporoparietal response to letter rhyming and matching tasks, consistent with the observations of Pugh and colleagues. They have also shown that intensive training may help reduce this disruption.

A fifth neuropsychological account that has received attention in the literature owes to Tallal, who regards the phonological difficulties of poor readers as a symptom of an underlying auditory temporal processing deficit. It has been suggested that an inability to recognize the very short-duration sounds of speech is at the root of some poor readers' phonological difficulties. Originally an account of the perceptual deficits of dysphasic children, Tallal's account states that the left hemisphere is specialized for the rapid auditory temporal processing that is essential to speech processing and that the phonological deficits seen among some dysphasic children and some dyslexic children are due to left hemisphere-based deficits in auditory temporal processing. Her theory draws on several forms of results, predominantly on studies of dysphasic children but also on some research involving aphasic adults and reading impaired children. She and her colleagues have shown that some spoken language impaired children show deficits in performance on (1) discrimination and temporal order judgments of long versus short stimuli separated by long versus short

intervals and (2) in performance in identifying synthetic speech in which the relevant cues involve rapid changes in acoustic spectra. They have developed training programs that give experience with enhanced speech and have shown that these improve the language skills of some language learning impaired children.

The rapid temporal processing approach is controversial. On the one hand, it seems consistent with evidence that dyslexics have a defect in the magnocellular system (see Section II) responsible for temporal contrast sensitivity and temporal order resolution in visual stimuli as well as auditory ones. On the other-hand, it finds a strong and compelling critic in the research of Studdert-Kennedy and Mody, who point out that the errors that Tallal ascribes to poor temporal judgment reflect problems with rapidly identifying similar stimuli (i.e., naming problems) rather than problems with temporal order per se. They also note that certain findings about the processing of nonspeech stimuli fail to support the contention that the left hemisphere is specialized for rapid temporal processing as opposed to language processing. Perhaps most compelling is their result involving a group of impaired readers in the second grade who were selected on the basis of having made many errors in Tallal's tonal order judgment task. The children who showed problems with tonal order judgment were impaired, but only when the stimuli were highly similar pairs of syllables. Moreover, although the children had problems with discriminating similar speech sounds (i.e. "*ba*" and "*da*") their problems did not extend to nonspeech analogs of the rapidly changing parts of the difficult syllables, nor did they extend to another speech task that involved brief transitional cues. Thus, Mody and Studdert-Kennedy conclude that poor readers' problems with speech perception and their diminished ability to access speech codes have nothing to do with auditory temporal processing, in general, so much as with phonological skills. The problems appear to be domain specific and phonological rather than general and auditory.

B. Experiential Accounts

Both genetic and environmental influences can factor into reading difficulties. On the one hand, there are clearly heritable aspects of reading disability. Disruptions of neurological integrity seem to prevail in the left hemispheres of reading disabled individuals. However, poor reading and low verbal intelligence also tend to be associated with low socioeconomic status (SES) and large family size. By the time they are in kindergarten, children in lower SES families tend to have weaker spoken language skills and weaker phonological awareness skills, both of which place them at risk for reading problems in the future. Experience is no doubt important to the child's development of the language processing skills involved in speech perception, vocabulary, etc., but the role of the experiential environment in poor readers' problems with speech perception is only beginning to receive attention and remains a topic for future research. However, there is a wealth of evidence about the role of the environment—specifically the educational environment—in the development of phoneme awareness.

Phoneme awareness is a cognitive skill of sorts and, as such, must surely demand the attainment of a certain degree of intellectual maturity. However, phoneme awareness undergoes a spurt when children begin formal instruction in learning to read. For example, American children undergo the spurt earlier than German ones due to the fact that they receive literacy instruction in kindergarten, whereas literacy instruction in Germany is delayed until the first grade. Likewise, illiterate adults are unable to manipulate the phonetic structure of spoken words. It seems that awareness of phonemes is enhanced by (and perhaps dependent on) reading instruction methods that direct the child's attention to the phonetic structure of words. However, inadequate experience alone cannot globally account for some children's failure to achieve phoneme awareness. This is aptly shown by Bryant and colleagues' finding that among a group of 6-year-old skilled readers and 10-year-old disabled readers who were matched for reading ability, the disabled readers performed significantly worse on a phoneme awareness task, even though they would be expected to have had even more reading instruction than the younger children. Here, it could be argued that some constitutional factor limited the disabled readers' ability to profit from instruction and thus limited their attainment of phonological sophistication. This problem with achieving phoneme awareness is a primary trait of individuals who are familial dyslexics. It may relate to a more basic problem with using phonological representations as discussed by Elbro and colleagues, for example.

VIII. SUMMARY AND CONCLUDING REMARKS

The International Dyslexia Association defines dyslexia as a neurologically based, often familial disorder

that interferes with the acquisition of language. It varies in degrees of severity but manifests itself in difficulties with receptive and expressive language, including phonological processing, reading, writing, spelling, handwriting, and sometimes arithmetic. Dyslexia is not the result of a lack of motivation, sensory impairment, inadequate instructional, or environmental opportunities, but it may occur together with these conditions. Although dyslexia persists through life, individuals with dyslexia frequently respond successfully to timely and appropriate intervention.

This article attempted to give substance to this definition by proceeding from a consideration of the importance of certain language skills to reading, to a survey of evidence that links problems with these language skills to early reading disability, and to a consideration of some plausible origins of these problems. By way of a conclusion, some generalizations are now offered.

A. Conceptualization of the Problem

The literature on the relation between language processing skills and reading problems indicates that poor readers—and children who are likely to become poor readers—tend to have problems with phoneme awareness and also with three aspects of language processing skill: speech perception under difficult listening conditions; vocabulary, especially when it is measured in terms of naming ability; and using a phonological representation in linguistic short-term memory. They also possess less well-integrated neural structures in the left hemisphere. There is a logical interrelation between the behavioral difficulties of poor readers because they all involve phonological processes that concern the sound pattern of language. Hence, we may speculate that the cause of many instances of reading disability involves a certain dysfunction within the phonological system—a phonological core deficit.

This article focused primarily on the language problems of reading disabled children in general, and there has been no attempt to differentiate between dyslexic children and so-called garden variety poor readers. Recent reviews have seriously discredited the use of a discrepancy between IQ and reading ability as a definitive characteristic that sets dyslexic children and their difficulties apart from other poor readers and their difficulties. From the perspective of current research, the two groups of children seem to form a continuous distribution. Both have problems with the language skills reviewed in Sections V and VI, and the phonological core deficit seems just as characteristic of dyslexic children as of garden variety poor readers. To date, the only language measure that distinguishes the two groups of children is receptive vocabulary, which may account for the lack of consensus about the role of receptive vocabulary problems in reading, as discussed in Section V. All other differences between the groups seem to involve real-world knowledge and strategic abilities: Generally, dyslexic children possess superior skills in these nonlinguistic areas, hence the discrepancy between their IQ and their reading ability, whereas the garden variety poor readers may show a developmental delay in these skills as well as in their phonological ones.

Future research can help us to better approximate a more accurate description of the phonological core deficit and its role in the reading problems of different groups of children. For example, it may help us contrast children within each group, informing us as to whether there are subtypes in terms of different problems or different clusters of problems. In this regard it is interesting to note Pennington and colleagues' observation that the language problems of adults from dyslexic families tend to be restricted to phoneme awareness, whereas those of individuals from nondyslexic families demonstrate problems with linguistic short-term memory as well as problems with phoneme awareness. In the future, researchers will surely attempt to determine whether this distinction applies to the population of young children who have reading problems as well as it applies to that of adult disabled readers.

Future researchers may also draw a closer link between the neurological problems currently being identified and the language deficiencies of poor readers. The strongest consensus of current neuropsychological research is that the reading problems of disabled readers are somehow linked to the integrity of structures and systems in the left cerebral hemisphere. Current hypotheses range from considerations of the integrity of systems, especially the left posterior systems, to considerations of the connections among systems, especially anterior and posterior systems. Linking these to behavioral deficits and effective remediation are important goals as well.

B. Practical Applications and Implications for Future Research

One of the practical benefits of the research described in this article is its potential for suggesting ways of

predicting and remediating early reading difficulty. One obvious benefit concerns screening devices for identifying children at risk for early reading problems. Such phonological processing skills as the ability to rapidly access the names of objects and the ability to make effective use of phonological representation in short-term memory have already proven effective kindergarten predictors of first-grade reading success. It is the task of future research to consider other tests of speech perception and tests of sentence comprehension that place special demands on short-term memory for language. It is also important for future studies to address the very practical matter of how to administer such tests to groups of children. To date, the studies that have successfully predicted future reading problems have tediously involved two or more sessions of individual testing—hardly practical for large-scale use in public school systems.

The research surveyed in this article may also be of interest to those who are concerned with the remediation of reading problems. As we begin to identify the linguistic problems associated with specific reading difficulty, and to refine their causes, we should also be able to develop more effective procedures for their remediation.

Certainly the brightest prospects for remediation are offered by research that has shown that various types of training can facilitate phoneme awareness. The best favor we can do for all children is to promote their phoneme awareness so that we may let them in on the secrets of the alphabetic principle as early as possible. Some very interesting and very practical advice on how to facilitate phoneme awareness is currently available from researchers such as Liberman, Blachman, Torgesen, and Bradley and colleagues. There exist a variety of word games, nursery rhymes, and other prereading activities that can help nurture a child's awareness of the way in which words break down into phonemes. Such activities will undoubtedly pave the way for phonics-oriented methods of instruction obviously favored by current research and so obviously in keeping with this article's focus on the importance of phoneme awareness in early reading.

See Also the Following Articles

ADOLESCENT BRAIN MATURATION • AGRAPHIA • BILINGUALISM • BRAIN DEVELOPMENT • DYSLEXIA • LANGUAGE ACQUISITION • LANGUAGE, NEURAL BASIS OF • LATERALITY • NEUROPSYCHOLOGICAL ASSESSMENT • SEMANTIC MEMORY • SPEECH • WORKING MEMORY

Acknowledgments

This work was partially supported by a grant from the Orange County Children and Families Commission. Joshua Ramirez is gratefully acknowledged for his comments on an earlier draft of the manuscript.

Suggested Reading

Adams, M. J. (1990). *Beginning to Read: Thinking and Learning about Print*. MIT Press, Cambridge, MA.

Andrews, J. J. W., Saklofske, D. H., and Janzen, H. L. (2001). *Handbook of Psychoeducational Assessment: Ability, Achievement, and Behavior in Children*. Academic Press, San Diego.

Brady, S., and Shankweiler, D. (1991). *Phonological Processes in Literacy*. Erlbaum, Hillsdale, NJ.

Geschwind, N., and Galaburda, A. M. (1987). *Cerebral Lateralization*. Bradford, Cambridge, MA.

Klein, R. M., and McMullen, P. A. (1999). *Converging Methods for Understanding Reading and Dyslexia*. MIT Press, Cambridge, MA.

Liberman, A. M. (1999). The reading researcher and the reading teacher need the right theory of speech. *Sci. Stud. Reading* **3,** 95–111.

Liberman, I. Y., Shankweiler, D., and Liberman, A. M. (1989). The alphabetic principle and learning to read. In *Phonology and Reading Disability: Solving the Reading Puzzle* (D. Shankweiler and I. Y. Liberman, Eds.), pp. 1–33. Univ of Michigan Press, Ann Arbor.

Mann, V. A. (1998). Language problems: A key to early reading problems. In *Learning About Learning Disabilities* (B. Wong, Ed.), 2nd ed. pp. 163–202. Academic Press, San Diago.

Perfecta, C. A., and Sandal, R. (2000). Reading optimally builds on spoken language: Implications for deaf readers. *J. Deaf Stud. Deaf Ed.* **5,** 32–50.

Pugh, K. R., Mencl, W. E., Shaywitz, B. A., Shaywitz, S. E., *et al.* (2000). The angular gyrus in developmental dyslexia: Task-specific differences in functional connectivity within posterior cortex. *Psychol. Sci.* **11,** 51–56.

Stanovich, K. E., and Beck, I. (2000). *Progress in Understanding Reading*. Guilford, New York.

Studdert-Kennedy, M., and Mody, M. (1995). Auditory–temporal perception deficits in the reading-impaired: A critical review of the evidence. *Psychonomic Bull. Rev.* **2,** 508–514.

Temple, E., Poldrack, R. A., Salidis, J., Deutsch, G. K., *et al.* (2001). Disrupted neural responses to phonological and orthographic processing in dyslexic children: An fMRI study. *Neuro. Report.* **12,** 299–307.

Torgesen, J. K., and Davis, C. (1996). Individual difference variables that predict response to training in phonological awareness. *J. Exp. Child Psychol.* **63,** 1–21.

Willcutt, E. G., Pennington, B. F., Boada, R., Ogline, J. S., *et al.* (2001). A comparison of the cognitive deficits in reading disability and attention-deficit/hyperactivity disorder. *J. Abnormal Psychol.* **11,** 157–172.

Receptive Field

RAJESH P. N. RAO
University of Washington

GLOSSARY

classical receptive field The region of sensory space that activates a given neuron.

color-opponent receptive field A type of center-surround receptive field characterizing a cell that increases its response when light of a particular wavelength (e.g., red) is turned on in the center, and decreases its response when light of a different wavelength (e.g., green) is turned on in the surrounding region.

Hebbian learning An algorithm for modifying the connection between two neurons in proportion to the correlations between the activity of the input neuron and the output neuron.

nonclassical receptive field The region of sensory space surrounding the classical receptive field of a neuron that allows stimuli placed within it to modulate the responses of the neuron to stimuli placed within the classical receptive field.

"on"-center, "off"-surround receptive field A type of receptive field characterizing a cell that responds best to a circular stimulus being turned on in the center of the cell's activating region and an annular stimulus being turned off in the surrounding region.

oriented receptive field A receptive field whose structure is oriented in at least two of its stimulus dimensions, for example, x and y spatial coordinates, space and time, or frequency and time. Cells with oriented receptive fields enable detection of features such as oriented edges, motion in a particular direction, or auditory features exhibiting increasing or decreasing frequencies.

spectrotemporal receptive field Receptive field of an auditory neuron that depicts the response of the neuron as a function of sound frequency and time.

The receptive field of a neuron is traditionally defined as the region of sensory space whose stimulation influences the response of the neuron. Recent definitions also include the specific properties of the stimulus that generate a strong response in addition to the region of sensory space causing the response. By mapping the receptive field of a neuron, one obtains crucial insights into the role that the neuron plays in transforming sensory information into perceptually useful quantities. This article reviews methods for mapping receptive fields, the types of receptive fields found in visual, auditory, somatosensory, olfactory, and gustatory neurons, modulation of these classical receptive field responses by stimulation of surrounding regions and by attention, plasticity of receptive fields, and computational models for receptive field development.

I. MAPPING RECEPTIVE FIELDS

A. Traditional Mapping Techniques

The notion of receptive fields was first explored systematically in the visual system by Hartline, Kuffler, Barlow, and Hubel and Wiesel in the 1940s, 1950s, and 1960s. To determine the receptive field of a sensory neuron, one probes the responsiveness of the neuron by presenting the animal with specific stimuli. For example, to determine the receptive field of a retinal ganglion cell in an anesthetized cat, the eyes of the cat are focused on a screen and spots of light are projected for brief durations at different spatial locations on the

screen. The responses of the retinal cell are recorded using a microelectrode. Spots flashed within a restricted location on the screen may increase or decrease the firing rate of the cell. The receptive field can thus be mapped by determining all areas on the screen that influence the firing rate of the cell either by increasing it or by decreasing it. Figure 1 illustrates this mapping procedure for retinal ganglion cells.

B. Mapping by Reverse Correlation

Recent methods for receptive field mapping are based on using white noise stimuli and reverse correlation analysis (borrowed from linear systems theory) to determine receptive field properties of a neuron in both space and time. For example, to map receptive fields in primary visual cortex, the stimuli used consist of sequences of small bright or dark bars whose orientation matches the preferred orientation of the recorded neuron. The bars are flashed one after another at random locations on a uniformly gray screen. The contrast of each bar (bright or dark) is chosen randomly and a given randomized sequence contains bright and dark bars presented once at each discrete screen location. The spike train of the neuron is recorded for each randomized sequence of bar stimuli. For each spike, the stimuli that occurred a brief period of time (typically a few hundred milliseconds) before that spike are summed, adding 1 or -1 at location (X, Y) and time T if the stimulus at time T before the spike contained a bright or dark bar at location (X, Y). This process is repeated for 20–40 random stimulus sequences and the resulting summated stimulus is divided by the total number of spikes to obtain the average spatiotemporal stimulus profile that causes the given neuron to spike. This spatiotemporal average is known as the space–time receptive field of the neuron. The value at (X, Y, T) in the space–time receptive field map represents the contribution made to the generation of a spike by a stimulus presented T milliseconds ago at spatial position (X, Y). Reverse correlation analysis thus allows one to determine the size, shape, and dynamics of a neuron's receptive field in a quantitative fashion, assuming that the relationship between the input stimulus and the response of the neuron is linear. Figure 2 depicts the process of reverse correlation for a neuron in cat visual cortex.

II. VISUAL RECEPTIVE FIELDS

A. Retina

Visual information reaching the eye is first processed by the circuitry of the retina. The output of the retina is conveyed to the lateral geniculate nucleus by the axons

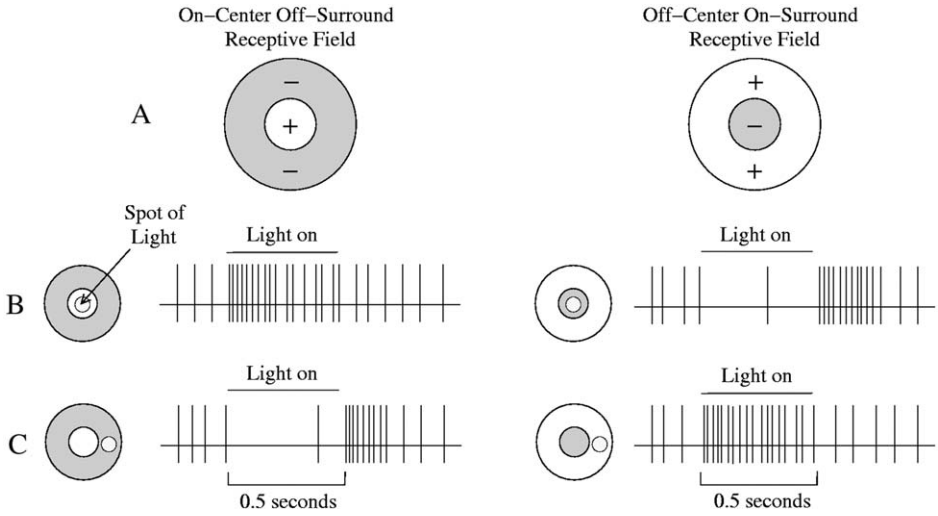


Figure 1 Mapping a receptive field. (A) An "on"-center "off"-surround receptive field (left) and an "off"-center "on"-surround receptive field (right) in the retina. These receptive fields are mapped by determining all areas on the screen that influence the firing rate of the cell either by increasing it or by decreasing it. Spots flashed within a restricted location on the retina increase or decrease the firing rate depending on the location of the spot. (B) "On"-center refers to an increase in the firing rate when a spot of light is turned on in the center (left). "Off"-center refers to a decrease in the rate when the spot is turned on in the center (right). Note that in the latter case, the firing rate also increases when the spot is turned off. (C) "Off"- and "on"-surround refer respectively to a decrease or increase in the rate when a spot of light is turned on in the region surrounding the center of the receptive field (after Nicholls *et al.*, 1992).

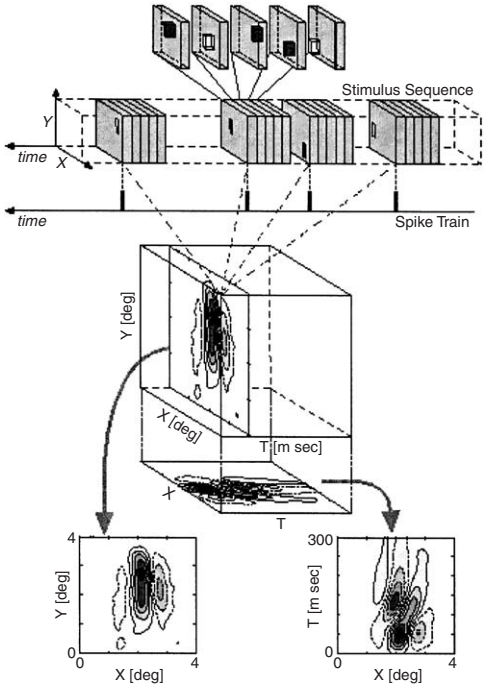


Figure 2 Using reverse correlation to map receptive fields. The top of the figure represents a segment of a randomized stimulus sequence that can be thought of as a time series of two-dimensional spatial patterns, as shown by a collection of square "slabs" stacked from right to left. These slabs are shown separately at the top in an exploded view for five consecutive stimuli. Each slab represents a single stimulus presentation containing a bar stimulus that is either bright or dark and flashed on an otherwise uniformly gray screen somewhere within the square region. Some of these stimuli fall on excitatory regions of the receptive field and spikes are elicited. A spike train is depicted below the stimulus sequence. For each spike generated, we look back in time (rightward in the figure, shaded cube) for stimuli that are likely to have elicited the spike, starting at the instant of spike occurrence. The stimuli in the shaded cubes are summed for all spikes generated in a stimulus sequence, counting each bright bar stimulus as 1, each dark bar as -1, and the background as 0. This produces the average stimulus profile (after division by the number of cubes added) that elicits a spike. The average profile is depicted within a cube shown in the center. An element within this cube at (X, Y, T) represents the contribution that a stimulus makes to the generation of a spike, i.e., a stimulus presented T milliseconds ago at spatial position (X, Y) contributes to spike generation by the amount indicated by the value at (X, Y, T) in the averaged cube. A cross section at a given T is a spatial map of the receptive field at that instant, as shown at the bottom left. A cross section along the Y dimension gives the space–time $(X–T)$ receptive field, as depicted by the inset at the bottom right (reproduced with permission from Ohzawa *et al.*, 1996, Encoding of binocular disparity by simple cells in the cat's visual cortex. *J. Neurophysiol.* **75,** 1779–1805).

of ganglion cells, which form the optic nerve. The receptive fields of ganglion cells provide insight into the type of processing and transformations occurring in the retina. The typical receptive field of a ganglion cell is a circular area on the retina, consisting of a central region and an annular surrounding region. Some cells respond best when a bright spot of light is shone in the center of the receptive field surrounded by darkness: These cells are known as "on"-center cells. Shining light in the annular surround region of these cells decreases their response. Other cells decrease their activity or stop firing when light is shone in the central area but fire vigorously when the light is turned off: These cells are known as "off"-center cells. "Off"-center cells typically have "on"-surrounds (i.e., they increase their response when light is turned on in the annular surrounding region of their receptive field) (Fig. 1). Both types of cells do not respond at all or respond weakly to uniform illumination covering both the center and the surround, as expected from their antagonistic receptive field structure. In addition, the size of the receptive fields increases as one progresses from the center of the retina (fovea or area centralis) to the periphery. In the primate, the two major classes of ganglion cells are P cells and M cells. P cells are sensitive to color and have "color-opponent" receptive fields: They receive antagonistic center-surround input from the long-, middle- and short-wavelength-sensitive (L, M, and S) cones in the retina. For example, red–green P cells (or equivalently, red-on green-off cells) are excited by red (or L-wavelength) light shining in the center of the receptive field and inhibited by green (or M-wavelength) light in the surround. Blue–yellow cells are excited by blue light in the center and inhibited by yellow light (which is the sum of inputs from L- and M-wavelength cones) in the surround. In contrast to P cells, M cells have a larger receptive field and are insensitive to color but highly sensitive to low contrasts. They receive mixed L, M, and S cone inputs to both the center and surround of their receptive fields. Figure 3 illustrates the center-surround color-opponent receptive fields of P and M retinal ganglion cells.

B. Lateral Geniculate Nucleus

The receptive fields of neurons in the lateral geniculate nucleus (LGN) in the thalamus, which receives the outputs of retinal ganglion cells via the optic nerve, are in many ways similar to retinal receptive fields. They have "on"-center, "off"-surround-type and "off"-center, "on"-surround-type concentric receptive fields, but compared to ganglion cells, the inhibitory interaction between the center and the surround is more finely

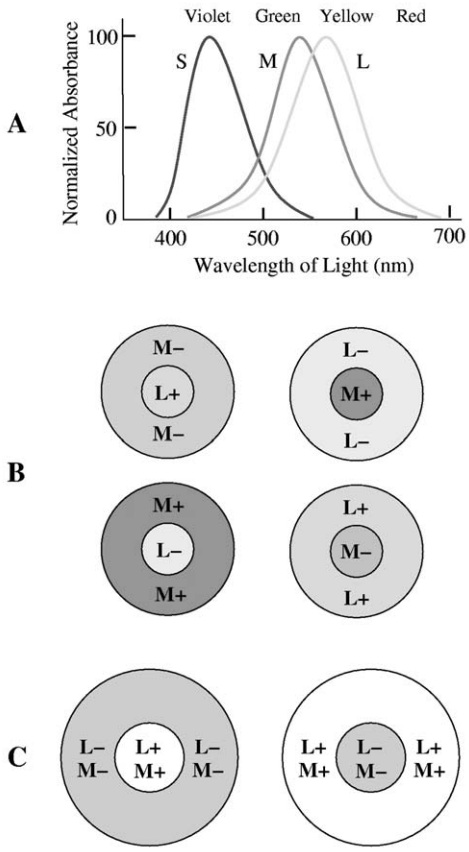


Figure 3 Color-opponent receptive fields. (A) Absorption spectra of long, middle, and short wavelength-sensitive cones (L, M, and S) in the human retina. (B) Center-surround color-opponent receptive fields of "red–green" parvocellular ganglion cells in the retina (projecting to the parvocellular layers of the LGN). Turning on long-wavelength (red) light in the receptive field center excites L "on"-center cells (L+) but inhibits M "off"-center cells (M−). Four possible combinations of L or M and "on"- or "off"-center-surround receptive fields are shown. (C) Center-surround receptive fields of magnocellular retinal ganglion cells. The receptive fields are not color sensitive, receive both L and M inputs, and are either "on"-center (L+M+) or "off"-center (L−M−) (after Zigmond *et al.*, 1999).

matched, resulting in much weaker responses to uniform illumination. In the primate, cells in the parvocellular layers (found dorsally in the LGN) receive inputs from retinal P cells and exhibit color-opponent receptive fields, whereas cells in the ventral magnocellular layers receive inputs from M cells and are not selective for color. In addition, unlike retinal ganglion cells, the responses of LGN neurons can be influenced by feedback from the visual cortex. For example, some LGN neurons are sensitive to the length of a moving bar, responding more weakly for longer bars than shorter bars, a property known as end inhibition or endstopping. Inactivating cortical feedback may disinhibit these neurons, causing them to lose their sensitivity to bar length. The influence of cortical feedback on LGN receptive fields is still unclear due to incomplete knowledge about the precise nature of cortical feedback and its interactions with inhibitory interneurons in the LGN.

C. Primary Visual Cortex

Axons from the LGN terminate in the primary visual cortex (V1 in primates and area 17 in the cat). Although cells in layer 4C of the cortex, where most of the geniculate fibers terminate, possess concentric center-surround receptive fields, the vast majority of neurons in other cortical layers exhibit receptive field properties that are much richer than those of their retinal and geniculate counterparts. Most of these neurons respond best not to spots of light but to elongated bars or edges that are flashed or moved in a particular direction. In addition, receptive field sizes in V1 can reach up to 8° or more, compared to a maximum size of 2 or 3° in the LGN, suggesting that these neurons spatially integrate the outputs of their antecedent LGN neurons, yielding receptive fields that are combinations of LGN receptive fields. Unlike LGN cells, many cells in V1 have binocular receptive fields (i.e., they are driven by stimuli presented to both eyes).

1. Simple Cells

Two major classes of cortical neurons were identified by Hubel and Wiesel in their seminal studies of the visual cortex. One class of cells, known as simple cells, possess receptive fields consisting of elongated "on" and "off" subregions oriented at a particular angle and localized to a small region of visual space. For example, a simple cell with a receptive field consisting of a vertically oriented elongated central "on" region flanked by two elongated "off" regions (Fig. 4A, left) responds best to a vertical bar of light that entirely fills the central "on" region. Figure 4B illustrates the responses of this neuron to bars at different orientations. Figure 4A also shows other simple cell receptive fields selective for dark bars or edges in images. Binocularly driven simple cells fuse information from both eyes and act as "disparity detectors," with selectivity for near, far, or zero disparity depending on whether they respond most to stimuli nearer than, farther than, or on the plane of fixation. They typically have receptive fields with the same preferred

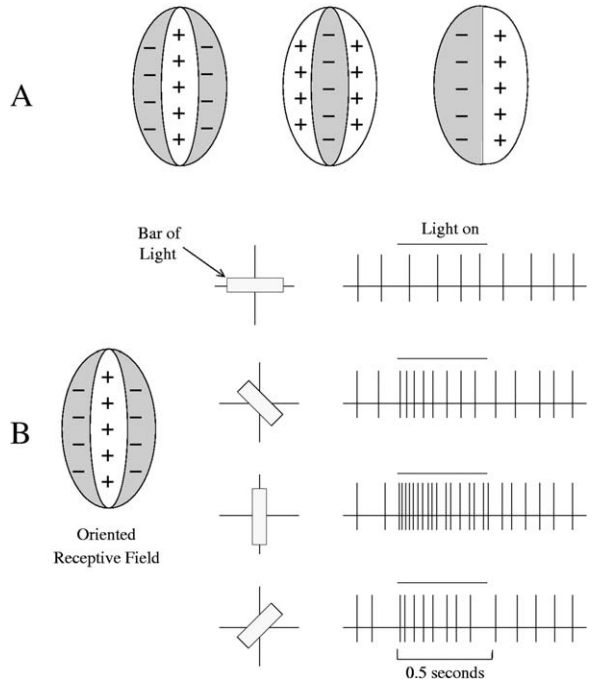


Figure 4 Oriented receptive fields. (A) Examples of oriented receptive fields of three simple cells selective for (from left to right) a bright vertical bar in the center, a dark vertical bar in the center, and a bright edge turned on in the right half of the receptive field. Although these receptive fields are all at a single orientation (vertical), a wide range of orientations are found in the cortex for each type of receptive field (as illustrated in Fig. 15). (B) (Left) A receptive field with an elongated "on" area (+) surrounded by two antagonistic "off" areas (−). A cell with such a receptive field responds best to a vertically oriented bar of light turned on in the center of its receptive field, other orientations producing little or no response (see examples on the right).

orientation in both eyes. Receptive field locations may be identical or differ in the two eyes. In the latter case, disparity may be detected due to a phase shift in the receptive field structure. Presenting optimally oriented stimuli simultaneously within the receptive fields of both eyes typically yields a much larger response than stimulating each eye alone, indicating a nonlinear facilitatory interaction between the two eyes (Fig. 5). Simple cells are generally found in layers 4 and 6 of primary visual cortex.

2. Complex Cells

A second major class of neurons, found primarily in layers 2, 3, and 5 of primary visual cortex, differ from simple cells in that they respond to both bright and dark bars (or edges) and are insensitive to their position within the receptive field. These cells are known as complex cells. The receptive fields of complex cells cannot be separated into distinct "on" and "off" subregions as in the case of simple cells but can be conceptualized as consisting of many subregions integrating both "on"- and "off"-type responses; this prompted Hubel and Wiesel to propose that complex cell receptive fields were constructed by combining several antecedent simple cell receptive fields (Fig. 6). Most complex cells have binocular receptive fields that are nearly identical for dark and bright stimuli. Due to their invariance to stimulus

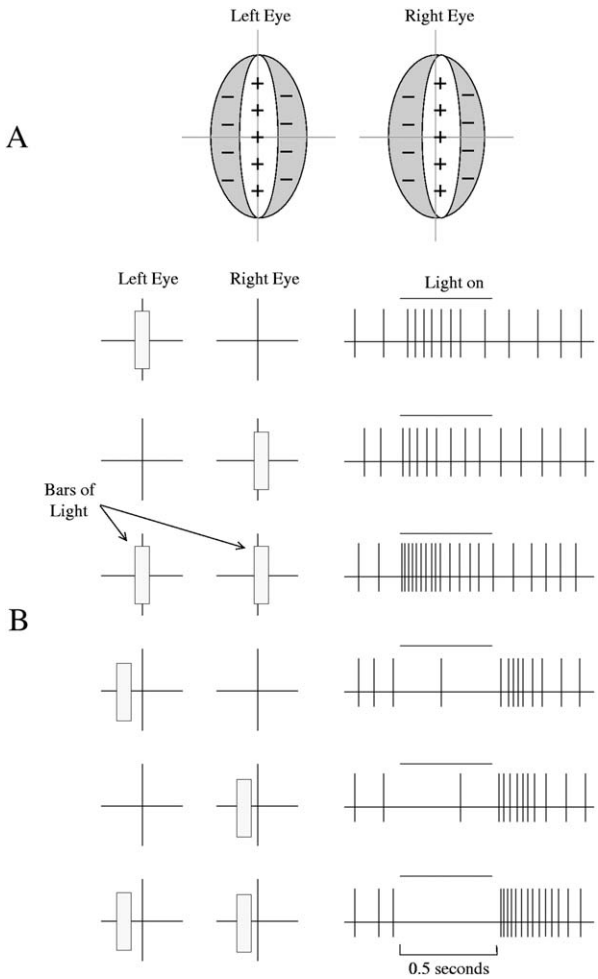


Figure 5 Binocular receptive fields. (A) Receptive fields of a cortical cell for the left and right eye. The receptive field for the right eye is shifted slightly to the right relative to the receptive field for the left eye. (B) Simultaneous presentation of bright bars in the "on" region of the receptive fields produces a much larger response from the cell (see third row) than presentation of the stimulus to either eye alone (first two rows). A similar nonlinear interaction is observed in the "off" responses of the cell when a bright bar is flashed in the left "off" region of the receptive fields. Such binocular receptive fields contribute to depth perception.

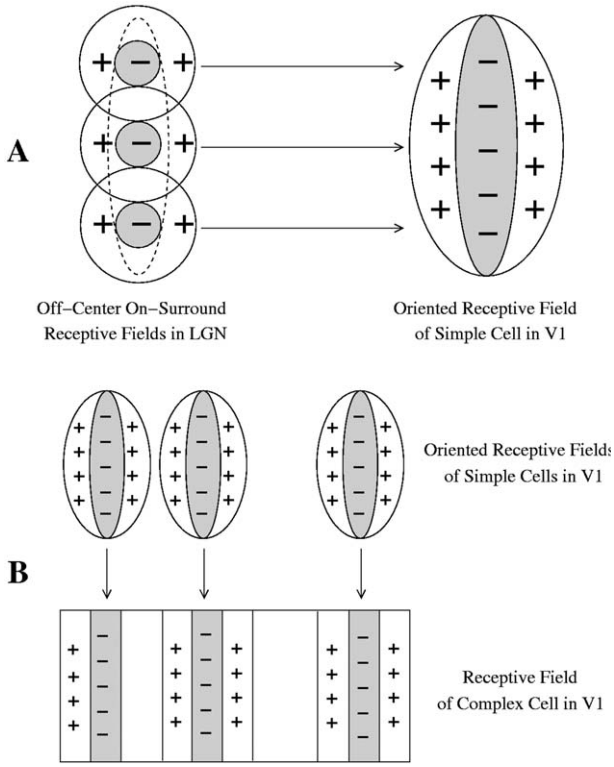


Figure 6 Model of simple and complex cell receptive fields. The figure depicts the hierarchical model proposed by Hubel and Wiesel for the construction of simple and complex cell receptive fields from center-surround LGN receptive fields. (A) A simple cell receptive field is synthesized from converging inputs from LGN cells whose center-surround receptive fields are displaced along the preferred orientation of the simple cell. (B) The receptive field of a complex cell is constructed from converging inputs from many simple cells that respond best to vertically oriented bars at different spatial positions. This allows the complex cell to respond to vertical edges anywhere within its receptive field.

position, they are thought to play an important role in computing stimulus depth.

3. Endstopped Cells

Some simple and complex cells respond optimally to stimuli of a particular size (e.g., a bar of a particular length); their response is reduced or even eliminated when the stimulus size is further increased beyond the optimal length. Such cells are known as endstopped cells because the optimal stimulus is an oriented bar or edge that "stops" at a particular place beyond the receptive field of the cell. In other words, there exists an "off" region just beyond the top and/or bottom ends of the receptive fields of these neurons that tends to inhibit the firing of the neuron. Figure 7 illustrates the response properties of an endstopped neuron in primary visual cortex. Note that endstopping or end inhibition is also found in the LGN (see Section II.B) as well as higher cortical areas and can be regarded as a type of nonclassical receptive field effect (see Section IV.A).

4. Double-Opponent Color Coding Cells

An important class of color-sensitive cells found in V1 but not in the retina or the LGN is that of double-opponent cells. These cells have concentric center-surround receptive fields with red–green or blue–yellow opponency, but unlike color-opponent cells in the LGN, these cells exhibit antagonistic responses in the center and surround for both colors. For example, in the case of a red–green "on"-center "off"-surround double-opponent receptive field, a red spot in the center excites the cell, and a red spot in the surround reduces its response; a green spot in the surround excites the cell but green in the center reduces the cell's response. Double-opponent receptive fields may play a crucial role in mediating color constancy, the phenomenon that allows us to perceive the color of objects despite large variations in ambient lighting conditions. Figure 8 illustrates the receptive field and responses of a red–green double-opponent cell.

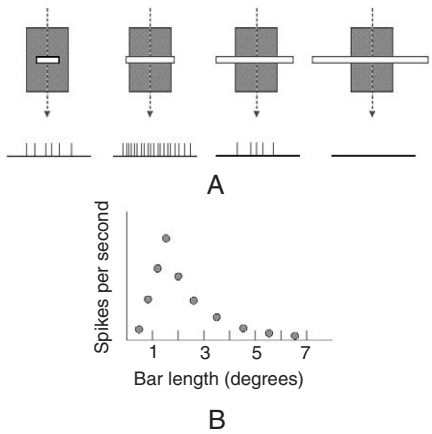


Figure 7 Endstopping. (A) The responses of an endstopped neuron to vertically moving bright bars of increasing length. The classical receptive field of the neuron is shown as a shaded rectangle. As the length of the bar is increased, the response of the neuron first increases and then decreases to zero. The peak response occurs when the size of the bar matches the width of the classical receptive field (second column). (B) The average response of the neuron (in spikes/second) plotted as a function of bar length (in degrees of visual angle).

D. Space–Time Receptive Fields of LGN and V1 Neurons

We have thus far focused on the spatial structure of visual receptive fields, as determined by flashing spots of lights or bars at different retinal locations. Recent techniques such as reverse correlation (see Section I.B) allow receptive field structure to be mapped in both space and time. Reverse correlation studies have revealed that the spatial structure of visual receptive fields may change over time. For example, an "on"-center "off"-surround-type receptive field in the LGN may gradually transform into an "off"-center "on"-surround-type receptive field over time. The evolution of such a receptive field is depicted in Fig. 9 as a space–time plot. Similarly, neurons in V1 may switch their "on" and "off" subregions over time. Some of these V1 receptive fields can be modeled by multiplying an oriented spatial receptive field by a temporal sinusoidal function with an initial positive lobe and a later negative lobe (Fig. 10). Such receptive fields are called separable. Many space–time receptive fields in V1, however, are inseparable (i.e., they cannot be expressed as a product of a spatial function and a temporal function). Such receptive fields generally have an orientation not only in space but also in time, as shown in Fig. 11. This orientation in time is due to a gradual shift in the phase of the oriented "on"–"off" subregions of the spatial receptive field over time: the spatial receptive field "moves" in a particular direction for a brief duration of time. The faster the shift, the greater the slope of orientation and the greater the velocity preferred by the cell. The direction of shift matches the preferred direction of motion of the cell. The space–time receptive field of a simple cell thus accurately predicts the orientation, direction of motion, and the preferred velocity of stimuli that best activate the cell. Space–time receptive fields of complex cells obtained through reverse correlation are limited in their usefulness due to the nonlinear summation inherent in these cells. Second-order methods involving the interaction of multiple stimuli presented at different positions and times are currently being explored to understand their receptive field properties.

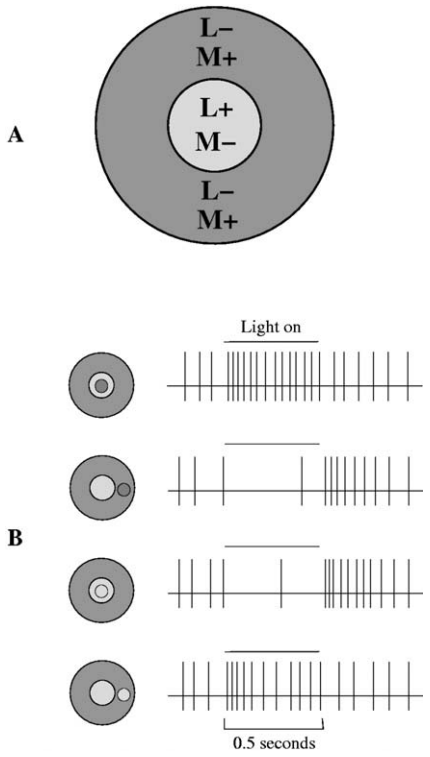


Figure 8 Double-opponent receptive fields. First discovered in the goldfish retina and later in the primate visual cortex, double-opponent receptive fields consist of concentric center-surround fields with L–M (red–green) or S–(L + M) (blue–yellow) antagonism. However, unlike color-opponent receptive fields in the retina (Fig. 3), each color produces antagonistic effects both in the center and in the surround. (A) An L–M (red–green) double-opponent receptive field. (B) Shining red light in the center of the receptive field produces an "on" response (first panel), whereas shining red light in the surround produces an "off" response (second panel). Green light in the center causes an "off" response (third panel), whereas in the surround it causes an "on" response (fourth panel). Double-opponent cells are believed to play a major role in color constancy, the phenomenon by which we perceive objects to be of the same color regardless of background illumination.

E. Higher Visual Cortical Areas

Based on anatomical considerations, visual cortical areas can be classified into a rough hierarchy with a coarse dichotomy between a dorsal visual processing pathway, involving areas V1, V2, MT, MST, and areas in the parietal cortex, and a ventral processing stream, involving areas V1–V4 and areas in the inferotemporal (IT) cortex. The receptive field properties of V1 neurons were discussed previously. V2 neurons share many of the receptive field properties of V1 neurons, including orientation selectivity and endstopping, but integrate information over a larger portion of the visual field. In addition, many V2 neurons respond to complex visual stimuli such as wedges, concentric circles (Fig. 12), or illusory contours of appropriate orientation, the latter being an example of contextual influences from stimuli lying beyond the classical

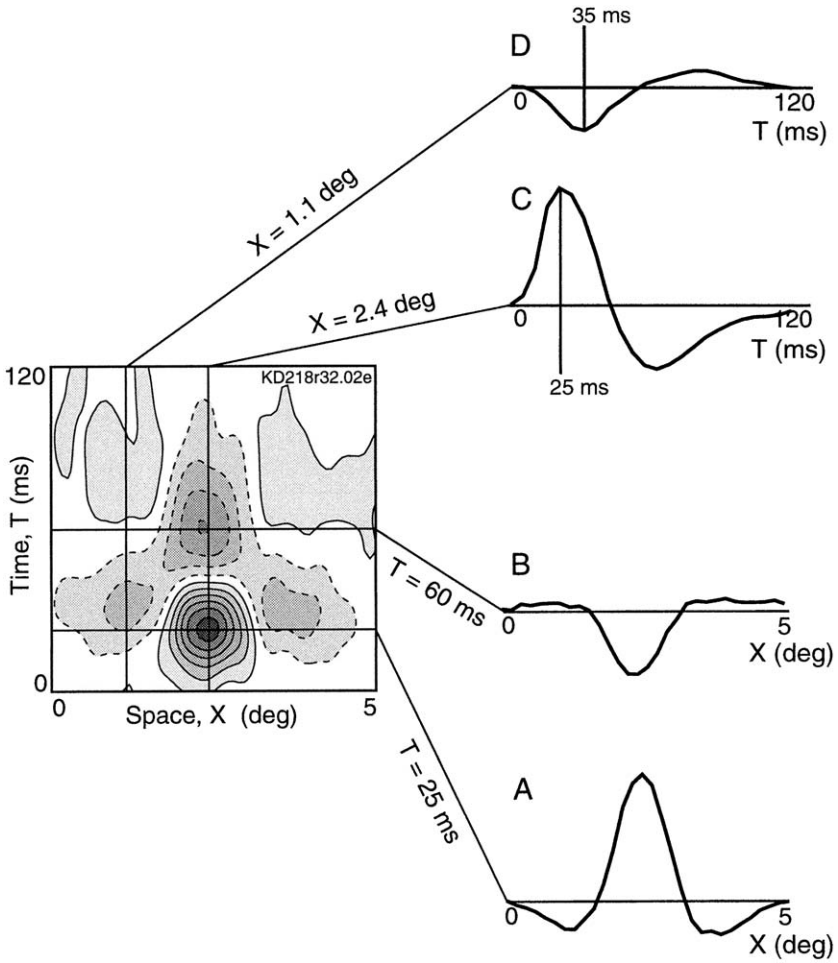


Figure 9 LGN space–time receptive field. The X–T contour plot on the left depicts the space–time receptive field of an LGN neuron: areas enclosed by solid contours correspond to bright-excitatory regions (bright bars at the these locations excite the neuron), whereas areas enclosed by dashed contours correspond to dark-excitatory regions (dark bars at these locations excite the neuron). It can be seen that the receptive field of this neuron transforms over time from an on-center, off-surround-type spatial receptive field (profile shown in A for $T = 25$ msec) to an off-center, on-surround-type spatial receptive field (shown in B for $T = 60$ msec). (C and D) Evolution of two different spatial values in the receptive field ($X = 1.1$ and 2.4°) over time (reproduced with permission from Cai *et al.*, 1997, Spatiotemporal receptive field organization in the lateral geniculate nucleus of cats and kittens. *J. Neurophysiol.* **78,** 1045–1061).

receptive field of the neuron (see Section IV.A). Comparably little is known about the receptive field properties of V3 neurons, except that they are selective for stimulus orientation and direction of motion, with a receptive field size even larger than those of V2 neurons. On the other hand, much is known about the properties of MT neurons. Most of these neurons are highly selective for stimulus velocity, with preferred speeds ranging from 2 to 256°/sec compared to a range of 1 to 32°/sec in V1. MT neurons are relatively insensitive to stimulus form (as defined by luminance, texture, color, or relative motion), but many are selective for binocular disparity, suggesting a role for these neurons in detecting motion in depth. An interesting receptive field property found in MT but not V1 is selectivity for pattern motion (as opposed to component motion): MT neurons respond to the direction of motion of the plaid pattern resulting from superimposing a pair of sinusoidal gratings moving in two different directions, whereas V1 neurons do not. Neurons in area MST are selective for more complex patterns of motion, such as rotations, dilations, and translations, within a receptive field that is larger than those of MT neurons. Most are disparity selective but largely insensitive to form and color.

Neurons in V4 are known to be selective for color as well as shape of stimuli. Many V4 neurons remain selective for length and width of bar stimuli over a

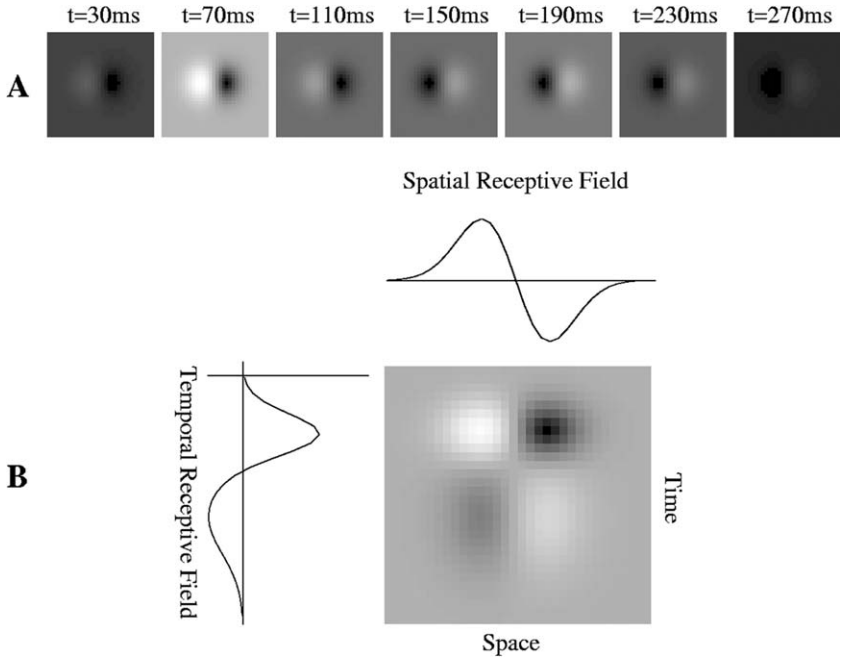


Figure 10 Separable space–time receptive fields. (A) An oriented spatial receptive field that reverses the polarity of its on and off regions (bright and dark regions) over time. The evolution of the receptive field is shown as a series of snapshots in increments of 40 msec. (B) The space–time evolution of the receptive field in A can be captured in a space–time plot obtained by integrating along the axis of orientation (vertical in this case) for each moment in time. As shown in the figure, the space–time plot for this receptive field can be characterized as the product of a spatial receptive field profile (top) and a temporal receptive field profile (left). Receptive fields with such space–time plots are said to be "separable." Some simple cells in primary visual cortex possess such separable space–time receptive fields.

much larger receptive field than do V2 or V3 neurons. Some V4 neurons prefer more complex shapes, such as conjunctions of edges or bars, star-like patterns, and concentric regions of contrast (Fig. 12). Neurons in areas TEO and TE of IT cortex have large receptive fields (20° in width or more) that often span both contralateral and ipsilateral visual fields. They show a selectivity for a range of complicated 2-dimensional and 3-dimensional shapes, from stars and concentric circles to hands and faces (Fig. 12). In addition, responses of IT neurons are often invariant to the position, size, and orientation of their preferred stimulus shapes. Finally, some neurons in IT cortex respond not to the stimulus but to its memory. Some of these neurons reduce their response gradually as the stimulus is repeatedly presented, whereas others continue to fire after the stimulus has been removed, essentially holding a memory trace of the stimulus.

F. Summary of Visual Receptive Field Properties

As one progresses from lower visual processing centers, such as the retina and LGN, to higher cortical areas, one witnesses a gradual increase in receptive field size (Fig. 13). Simultaneously, there is an increase in both response invariance and selectivity for higher order stimulus features: while complex cells in V1 respond to oriented edges or bars and remain invariant over a spatial extent of a degree or two, cells in IT maintain their response selectivity for complex stimuli such as faces across several tens of degrees of visual space. In addition, as one progresses along the visual hierarchy, one begins to encounter cells that store information regarding the recency and familiarity of a stimulus rather than responding to the stimulus per se.

III. RECEPTIVE FIELDS IN OTHER SENSORY MODALITIES

The study of receptive fields in sensory modalities other than vision is still in its infancy. Here, I briefly review some of the known properties of somatosensory, auditory, gustatory, and olfactory receptive fields, with an emphasis on the similarities in receptive field structure across modalities.

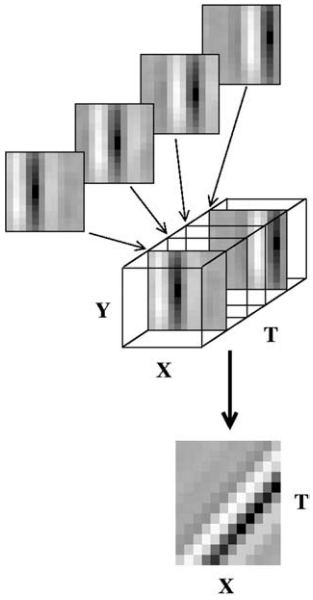


Figure 11 Inseparable space–time receptive fields. The X–Y–T cube depicts the space–time receptive field of a direction-selective cell that responds best to a dark bar moving rightwards. Four spatial cross sections of this cube are shown at the top for four different time values (time values increase from front to back). The bright regions in the receptive field represent spatial locations where a bright bar elicits a response from the cell, whereas the dark regions denote locations where a dark bar causes a response. The degree of brightness indicates the magnitude of responses. An X–T plot is obtained by integrating the three-dimensional cube along the Y axis, which is normalized to be parallel to the preferred orientation of the cell. The gradual shift in the spatial receptive field (in this case, to the right) as a function of time is responsible for the rightward slope in the space–time receptive field (X–T plot) and for the direction selectivity of the neuron. Note that unlike the receptive field in Fig. 10, the receptive field of this simple cell is space–time inseparable: its X–T plot cannot be approximated as the product of a spatial and a temporal receptive field profile (after DeAngelis *et al.*, 1993).

A. Somatosensory Receptive Fields

Somatosensory receptive fields are defined in terms of areas of skin within which a mechanical stimulus such as a touch causes a neural response. The receptive fields of afferent fibers carrying information from mechanoreceptors in the skin are primarily excitatory and only a few millimeters in diameter, with size and location determined by the receptor. Such receptive fields are analogous to the receptive fields of photoreceptors in the retina.

In the primate primary somatosensory cortex (S1), there are antagonastic center-surround receptive fields: the somatosensory neuron produces a robust response when a particular area of skin is touched but the neuron is inhibited when any part of the surrounding area is touched (Fig. 14). This is reminiscent of the center-surround organization of receptive fields in the retina and LGN. Some cells in somatosensory cortex are direction-selective (i.e., they respond only to movement of touch in a particular direction on the skin). Such cells have an inseparable oriented space–time receptive field, analogous to the receptive fields of direction-selective cells in visual cortex.

B. Auditory Receptive Fields

The receptive fields of neurons in the early auditory pathway have traditionally been studied using tone bursts at particular sound frequencies. These studies reveal that early auditory neurons, such as the type I neurons that transmit information from the inner hair cells in cochlea, show sharp tuning for particular characteristic frequencies: The response diminishes

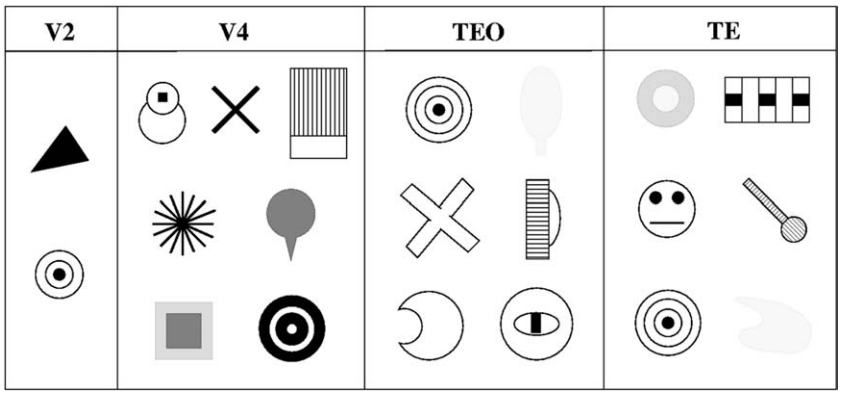


Figure 12 Receptive fields in higher visual areas. The figure shows some of the complex object features that elicit the best responses from different neurons in the higher visual cortical areas V2, V4, TEO, and TE (after Zigmond *et al.*, 1999).

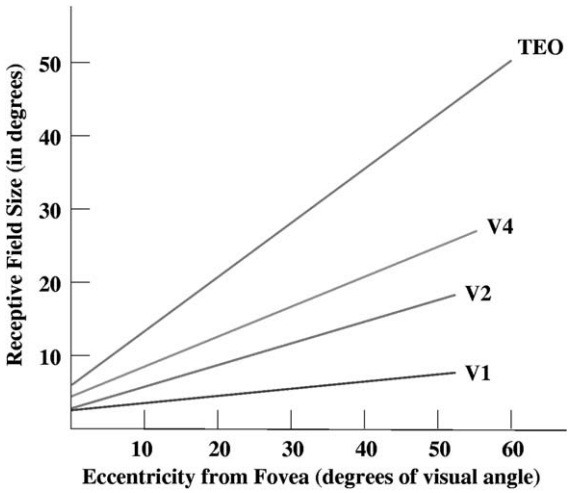


Figure 13 Receptive field sizes in different visual cortical areas. The graph plots receptive field size, defined as the square root of receptive field area, as a function of receptive field eccentricity for four different visual areas. The lines are based on linear regression from individual data points, replotted from Zigmond *et al.*, 1999. Two trends are apparent: (i) There is a linear increase in receptive field size with eccentricity for each area and (ii) receptive field sizes in higher visual areas (TEO, V4, and V2) are consistently larger than receptive field sizes in corresponding lower areas (V4, V2, and V1, respectively), reflecting the anatomical convergence of information from many neurons in a lower area onto a single higher area neuron.

rapidly for frequencies that are not close to the characteristic frequency. Other neurons, such as those in the cochlear nucleus, exhibit center-surround-type receptive fields when their responses are mapped as a function of sound pressure level and frequency. As one proceeds to higher auditory centers, the range of frequencies that characterize the receptive field becomes narrower. Many neurons in the auditory cortex do not respond to pure tones but do respond to changes in frequency in a particular direction (up or down) and at a particular rate. The spectrotemporal receptive fields of these neurons are oriented when plotted as a function of frequency and time; they are thus analogous to the oriented space–time receptive fields of direction-selective neurons in the visual and somatosensory cortex.

C. Olfactory and Gustatory Receptive Fields

Researchers have only recently begun to explore the structure of olfactory and gustatory receptive fields at different levels of processing. Much of the past research focused on elucidating the mechanisms by which taste receptors on the tongue and epithelia in the mouth and odor receptors in the epithelia of the nasal cavity convert chemical substances into neural signals. Most neurons in the early stages of processing have been found to be broadly tuned to a range of stimuli (e.g., salty, bitter, or sweet taste qualities) and are also modulated by intensity. Some also respond to tactile and thermal stimuli. However, it has been suggested that some olfactory receptive fields are analogous to retinal center-surround receptive fields: Mitral cells in the olfactory bulb exhibit excitatory responses to certain chemical compounds in a homologous series of compounds and inhibitory responses to neighboring compounds that flank the excitatory compounds in the series. The presence of antagonistic center-surround receptive fields in the olfactory system suggests that higher level receptive fields, such as those of neurons in the olfactory cortex, may possess an oriented structure

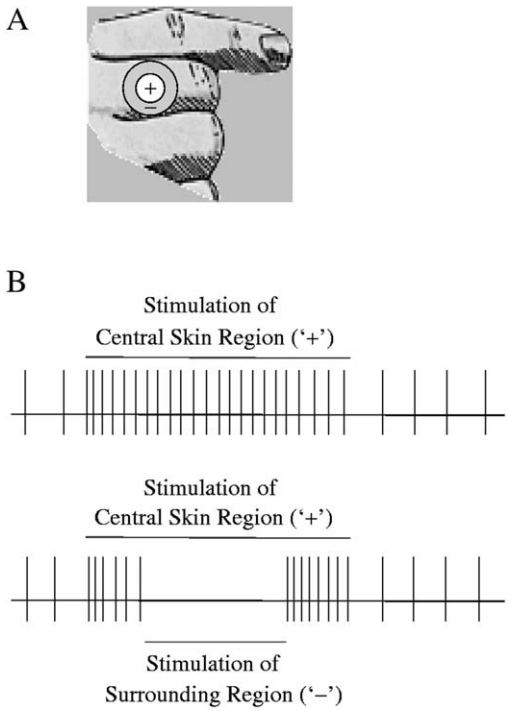


Figure 14 Somatosensory center-surround receptive field. (A) An example of a center-surround antagonistic receptive field of a neuron in the somatosensory cortex. The receptive field is located on the right hand and consists of an excitatory central region (+) surrounded by an annular inhibitory region (−). (B) The spike trains produced by a cell with such a receptive field for two cases: (i) stimulation of the skin in the central receptive field region (top) and (ii) stimulation of the skin in the central region together with stimulation in the surrounding region for a brief period of time (bottom). These spike trains illustrate the center-surround antagonism characterizing the cell's receptive field.

analogous to visual and auditory cortical receptive fields. However, the orientation would be in the space of chemical concentration and time, implying a sensitivity toward increasing or decreasing amounts of particular chemical compounds at a particular rate.

IV. NONCLASSICAL RECEPTIVE FIELDS AND ATTENTIONAL MODULATION

A. Nonclassical Receptive Field Effects and Contextual Modulation

The responses of many neurons can be modulated by presenting stimuli in the region surrounding the receptive field of the neuron. For example, a visual cortical neuron that responds to a moving bar or an oriented grating within its receptive field may no longer respond when the bar extends into the surrounding region or when the surrounding region contains a grating oriented in the same direction as the central grating. The region of space surrounding the "classical" receptive field that modulates the responses of the neuron to the central stimulus is known as the neuron's nonclassical or extraclassical receptive field. The modulatory effect is sometimes referred to as contextual modulation.

B. Attentional Modulation

The receptive fields of certain cortical neurons are also modulated by attention. For example, some neurons in areas V2 and V4 of the monkey visual cortex that respond vigorously to a bar stimulus in their classical receptive field reduce their response when a second stimulus is placed inside the receptive field. If, however, the monkey is trained to attend to the bar stimulus, the introduction of a second stimulus does not reduce the response. The neuron responds as if its receptive field has shrunk to include just the bar stimulus, ignoring the distracting second stimulus. Similar attentional effects have also been observed in imaging studies involving human visual cortex.

V. RECEPTIVE FIELD PLASTICITY

In addition to short-term modulation of receptive fields due to context and attention, the receptive fields of many neurons are labile and may be subject to long-lasting modifications in their structure. Some neurons exhibit a "critical period" during which exposure to stimuli from the external world causes changes in their receptive fields. For example, in kittens that are raised in a visual environment containing only vertical stripes, there is a predominance of visual cortical neurons with vertically oriented receptive fields. Similarly, although the receptive fields of many neurons in kitten visual cortex are directionally selective before eye opening, these receptive field lose their directionality when the animal is placed in a strobe environment, in which flashes occur a frequencies less than 10 Hz.

Receptive field plasticity has also been observed in adult animals. For example, in monkeys, if the skin from two adjacent fingers is surgically joined so that the two fingers now behave as one functional unit, the receptive fields of neurons in the somatosensory cortex that initially coded for just one of the fingers now extend across both fingers. If the fingers are then separated and used individually again, the receptive fields reorganize and start coding for the individual fingers. Similarly, in monkeys that have lost a limb, the receptive fields that were initially selective for stimuli on the lost limb may reorganize and start coding for other parts of the body. This reorganization of receptive fields after changes in the pattern of inputs is thought to occur through "Hebbian" adaptation of horizontal connections between cortical neurons. Hebbian mechanisms play an important role in models of receptive field development.

VI. COMPUTATIONAL MODELS OF RECEPTIVE FIELDS

Computational models of receptive fields can be divided into two categories: models that seek to explain receptive field properties based on known anatomical connectivity and physiological data, and models of the development of receptive fields based on, for instance, the statistics of natural images or sounds. An example of a model of the first type is Hubel and Wiesel's hierarchical model of the receptive fields of simple and complex cells in primary visual cortex (Fig. 6). In this model, the oriented receptive field of a simple cell is assumed to be constructed by aligning the center-surround receptive fields of a set of antecedent LGN cells in the proper orientation. In other words, the receptive fields of simple cells are assumed to be determined by pooling the outputs of LGN cells. The

receptive fields of complex cells are in turn determined by pooling the outputs of a set of simple cells with receptive fields of similar orientation. Such a "feedforward" model has received some support in both anatomical and physiological studies. However, other studies support a competing model based on local feedback connections between neurons, which amplify weak biases in the feedforward inputs. Clarifying the role of feedforward and feedback connections in determining cortical receptive field properties remains an active area of current research. A second class of models ask not *how* receptive fields are constructed but *why* they exhibit the properties that they do. These models seeks to explain the receptive field properties of sensory neurons in terms of the statistics of natural sensory signals. For example, the antagonistic center-surround receptive fields in the retina can be explained in terms of predictive coding theory: since natural images are not random but contain statistical regularities such as correlations over space and time, the value of an image pixel can usually be predicted quite accurately using a weighted average of the surrounding pixels. Thus, the output of a retinal ganglion cell, as determined by its center-surround receptive field, encodes the difference between the actual pixel values in the center and the predictions of these pixel values based on surrounding pixel values. Such a strategy minimizes redundancy and enhances dynamic range. It can also be interpreted in terms of the statistical operations of whitening and smoothing of input signals based on the frequency spectra of natural images. A similar explanation applies to color-opponent receptive fields and the center-surround receptive fields in the LGN.

Cortical receptive fields appear to be the result of a more sophisticated predictive coding strategy, which takes into account properties such as orientation, disparity, and direction of motion of objects. Models of cortical receptive field development have typically invoked Hebbian learning mechanisms that are based on correlations between inputs and outputs of neurons. Computer simulations of these models show how oriented receptive fields may develop from initially random receptive fields. Recent simulations using collections of natural images have shown that oriented receptive fields very similar to those measured in cortical neurons develop as a consequence of learning rules that maximize the statistical independence between outputs of cortical neurons (Fig. 15). Divisive normalization of neural activity based on the responses of neighboring neurons has also been suggested as an additional mechanism for achieving independence. Other models have shown that nonclassical receptive field effects may be a consequence of predictive coding in abstract parameter spaces such as orientation rather than the raw image space as in the case of the retina. The implementation of such models using lateral and/or corticocortical feedback connections remains a subject of current research.

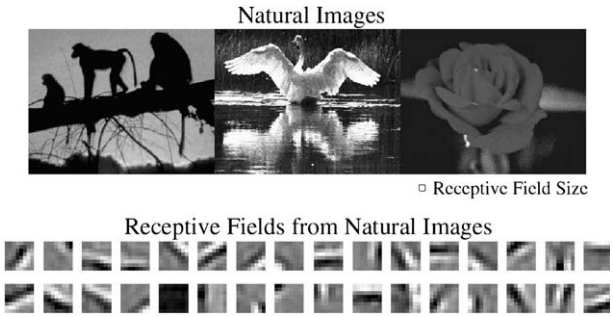


Figure 15 Oriented receptive fields from natural images. (Top) Three of the natural images used in training an artificial neural network whose goal is to reconstruct patches of these natural images using a linear combination of a given set of receptive fields. An additional constraint that the responses of model neurons be as statistically independent as possible helps to minimize the redundancy in the outputs of the neurons. Starting from a random set of receptive fields, the network learns a set of receptive fields that simultaneously optimize the goals of information preservation (reconstruction) and statistical independence for a large set of randomly extracted natural image patches. The relative size of the patches and receptive fields (8×8 pixels) is shown below the natural images as a square box labeled "Receptive Field Size." (Bottom) The receptive fields learned by the network after exposure to several thousand natural image patches. Bright pixels denote "on" regions of the receptive field, whereas dark pixels represent "off" regions. These receptive fields, which are both localized and oriented, resemble those of simple cells in primary visual cortex, as first noted by Olshausen and Field (1996). Results from such simulations support the hypothesis that receptive fields in sensory systems are often a consequence of efficient coding strategies applied to natural signals.

VII. CONCLUSIONS

The receptive field of a neuron provides a succinct characterization of the properties of the sensory world that the neuron is most selective for. Neurophysiologists have characterized a wide range of receptive fields in visual, auditory, somatosensory, and other sensory modalities. Many of these receptive fields are plastic and can be modulated by contextual information and attention. Receptive fields across different sensory

modalities share many similarities in their structure, such as center-surround opponent organization (Fig. 16) and orientation along their defining axes (such as space and time). These similarities in organizational principles are thought to reflect a set of common goals related to the efficient coding of natural sensory signals. Computational models of efficient coding based on the ideas of redundancy reduction, statistical independence, and predictive coding have provided new functional interpretations of classical and non-classical receptive field properties.

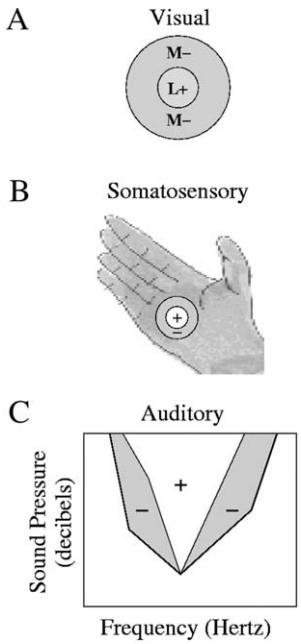


Figure 16 Similarities in receptive fields across sensory modalities. (A) Color-opponent center-surround receptive field in the retina. (B) Center-surround receptive field of a somatosensory neuron with an "on" response for skin stimulation in a circular region on the palm and an "off" response for stimulation in the annular surrounding region. (C) Center-surround receptive field of an auditory neuron in the cochlear nucleus. The center-surround antagonistic mechanism operates within the 2D space defined by sound pressure level, or loudness, and the frequency of the sound (a tone). The neuron is excited (+) by tones of a characteristic frequency (near the center of the frequency axis) but is inhibited (−) by surrounding frequencies.

See Also the Following Articles

AUDITORY PERCEPTION • COLOR VISION • FUNCTIONAL MAGNETIC RESONANCE IMAGING (fMRI) • MOTION PROCESSING • OLFACTION • RETINA • TASTE • VISION: BRAIN MECHANISMS • VISUAL CORTEX

Suggested Reading

Allman, J., Miezin, F., and McGuinness, E. (1985). Stimulus specific responses from beyond the classical receptive field: Neurophysiological mechanisms for local–global comparisons in visual neurons. *Annu. Rev. Neurosci.* **8,** 407–430.

Bell, A. J., and Sejnowski, T. J. (1997). The "independent components" of natural scenes are edge filters. *Vision Res.* **37**(23), 3327–3338.

Cai, D., Deangelis, G. C., and Freeman, R. D. (1997). Spatiotemporal receptive field organization in the lateral geniculate nucleus of cats and kittens. *J. Neurophysiol.* **78,** 1045–1061.

DeAngelis, G. C., Ohzawa, I., and Freeman, R. D. (1993). Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. I. General characteristics and postnatal development. *J. Neurophysiol.* **69,** 1091–1117.

DeAngelis, G. C., Ohzawa, I., and Freeman, R. D. (1995). Receptive-field dynamics in the central visual pathways. *Trends Neurosci.* **18**(10), 451–458.

Hubel, D. H. (1988). *Eye, Brain, and Vision*. Sci. Am., New York.

Kandel, E. R., Schwartz, J. H., and Jessell, T. M. (1991). *Principles of Neural Science*, 3rd ed. Appleton & Lange, Norwalk, CT.

Miller, K. D., Erwin, E., and Kayser, A. (1999). Is the development of orientation selectivity instructed by activity? *J. Neurobiol.* **41**(1), 44–57.

Nicholls, J. G., Martin, A. R., and Wallace, B. G. (1992). *From Neuron to Brain*, 3rd ed. Sinauer, Sunderland, MA.

Ohzawa, I., DeAngelis, G. C., and Freeman, R. D. (1996). Encoding of binocular disparity by simple cells in the cat's visual cortex. *J. Neurophsiol.* **75,** 1779–1805.

Olshausen, B. A., and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381,** 607–609.

Rao, R. P., and Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neurosci.* **2**(1), 79–87.

Reynolds, J. H., Chelazzi, L., and Desimone, R. (1999). Competitive mechanisms subserve attention in macaque areas V2 and V4. *J. Neurosci.* **19**(5), 1736–1753.

Simoncelli, E. P., and Schwartz, O. (1999). Modeling surround suppression in V1 neurons with a statistically-derived normalization model. In *Advances in Neural Information Processing Systems 11* (M. S. Kearns, S. A. Solla, and D. A. Cohn, Eds.), pp. 153–159. MIT Press, Cambridge, MA.

Zigmond, M. J., Bloom, F. E., Landis, S. C., Roberts, J. L., and Squire, L. R. (1999). *Fundamental Neuroscience*. Academic Press, San Diego.

Recovered Memories

HEIDI SIVERS
Stanford University

JONATHAN SCHOOLER
University of Pittsburgh

JENNIFER J. FREYD
University of Oregon

GLOSSARY

betrayal trauma theory A theory that predicts that the degree to which a negative event represents a betrayal by a trusted, needed other will influence the way in which that event is processed and remembered.

dissociation A psychological state involving alterations in one's sense of reality and one's identity, hypothesized to influence the storage and accessibility of event memories.

encoding specificity The tendency for a prior experience to come to mind when one encounter cues in the environment that correspond to the conditions in which the experience was originally encoded.

false memory A memory for an event that did not occur.

forgot it all along effect The claim that various biases sometimes cause individuals to overestimate the degree to which they had previously forgotten an experience.

memory persistence The degree to which a memory has remained available over time.

metaconsciousness One's explicit appraisal of his or her own's phenomenological experience. Variations in metaconsciousness of abuse experiences have been hypothesized to contribute recovered memories.

prospective trauma studies A research methodology in which participants are identified on the basis of their known history of trauma and are then contacted in order to determine their subsequent memory for the trauma.

psychogenic amnesia The report of profoundly forgetting experiences hypothesized to be due to psychological rather than physiological factors.

recovered memory The recollection of a memory that is perceived to have been unavailable for some period of time.

repression A defense mechanism hypothesized to keep unwanted information out of an individual's awareness.

Recently, there has been disagreement in the public, legal, and academic communities regarding if, how, and why people forget and later recover memories of traumatic life events. This article addresses the specific situation of remembering experiences of childhood sexual abuse, one of the most controversial types of trauma.

I. INTRODUCTION

There are many experiences in life that we expect to forget: a name, a birthday, perhaps even a trip we once

took. Typically, however, we expect to remember our most significant life experiences: one's wedding day, the birth of a child, or the death of a parent. Although very significant life events are generally remembered (albeit often far from perfectly) there is at least one type of situation in which people sometimes report forgetting seemingly unforgettable experiences. Specifically, for centuries, people have reported forgetting, and later recovering, memories for traumatic events. For almost as long it seems people have been discussing if, how, and why this phenomenon actually occurs. By the end of the 1900s, this discussion appeared in the popular and academic press as a heated and polarized debate, focused primarily on memory for childhood sexual abuse. On one extreme, it was argued that not only can significant events be forgotten and later recovered but also such recovered memories are intact and accurate. On the other extreme, it was argued that a developing community of hysteria, based on the assumption that almost any adult difficulties were due to forgotten childhood trauma, could convince otherwise normal adults that they were abused as children. All involved in the discussion emphasized the emotional significance of such claims and the importance in furthering our understanding of how trauma can affect memory. Fortunately, much of the fervor seems to have quieted, and what remains are a number of interesting and important scientific questions on the nature of human memory.

This article attempts to give the reader an overall picture of the issues surrounding recovered memories in which individuals report having remembered long-forgotten experiences of trauma. Although this article uses the standard term "recovered memory," we note that some have expressed concerns that this term makes undue assumptions. In particular, it implies that a memory had been entirely "lost," and that the memory that was ultimately recovered is the same as that which was lost. As will be noted, there are many cases in which individuals appear to sincerely believe they have discovered long-lost memories of trauma, even though the accuracy of the memories and/or the degree to which the memories had been entirely lost are unclear. Jonathan Schooler suggested that the term "discovered memory" may be more appropriate because it maintains agnosticity regarding both whether the memory was truly forgotten and the degree to which the memory corresponds to an experience that really occurred. At the same time, however, it respects the integrity of the individual's experience of having made a profound memory discovery. Clearly, the field would be well served by a careful consideration of the most appropriate term for this contentious construct. Here, we use the standard term recovered memory in deference to the fact that it is currently the most commonly used term in this context.

We begin by briefly summarizing the proposed social situation in which the current debate arose. Then, we present a sample of published research that has attempted to document the forgetting and recovery of memories for childhood sexual abuse. This is followed by a discussion of possible mechanisms for recovered memories, including ones proposed to be specific to traumatic events and others that are more standard mechanisms for remembering and forgetting. Finally, we discuss issues of accuracy in recovered memories.

II. SOCIOHISTORICAL CONTEXT

The idea that memories of painful events can be selectively forgotten and later recovered has been around at least since the late 19th century, when it was discussed in association with hysteria by Charcot, Janet, and Freud. This concept of repression experienced renaissance periods after each World War, when numerous soldiers reported experiencing an inability to remember and later recovery of parts or all of traumatic combat experiences, then considered a symptom of combat neuroses. Although the topic of war neuroses was periodically addressed and forgotten, it does not appear that the question of whether or not war experiences could be forgotten or recovered was particularly controversial. Furthermore, since it was well-known that the men had been in horrible combat situations, there was little need to doubt their reports of exactly what occurred. When the diagnosis of posttraumatic stress disorder (PTSD) first entered the third *Diagnostic and Statistical Manual of the American Psychological Association*, it included psychogenic amnesia as a symptom of the disorder.

In the 1980s and beyond, individuals began making claims regarding the forgetting and discovery of memories for a topic that was controversial—sexual abuse. The feminist movement of the 1960 and 1970s increased awareness of the prevalence of physical and sexual abuse in the lives of women and children. Furthermore, it was noted that the response to such trauma was similar to the response of the combat veterans to war; the women were displaying the same symptoms of PTSD as the men. Both of these claims in and of themselves were debated. First, it has been

argued that the prevalence of sexual abuse—variously defined—is actually very low. Second, it has been argued that many childhood sexual abuse experiences may not actually be traumatic. Needless to say, determining the precise incidence of sexual abuse and the concomitant trauma that is experienced at the time are complicated issues. Nevertheless, there is a large body of evidence indicating sexual abuse does occur with alarming regularity, and that many experiencers of such abuse find it extremely upsetting.

The major issue in the current controversy surrounds memory. In the 1980s and 1990s, the notion that traumatic events could be forgotten and later accurately recovered received a great deal of public support and attention. Along with this movement came several ideas that would cause the greatest amount of controversy. First, some groups asserted that an individual might have been abused even if he or she did not remember the abuse. Possessing symptoms from a broad list of possible consequences of trauma was proposed to be enough to assume that abuse did occur. Second, therapeutic techniques arose with the prime goal of recovering "repressed" memories. Free association, dream interpretation, hypnosis, and sodium amyatal, among other techniques, were used to uncover memories for abuse. Perhaps most critically, a number of individuals were encouraged to bring criminal or civil law suits against their perpetrators based on their recovered memories. Furthermore, numerous states passed special statutes of limitations for cases in which the plaintiffs purport to have only recently remembered the crime. These ideas were soon heavily contested. The inability of researchers to find unique and consistent symptoms of a history of trauma argued against making assumptions that an individual was abused based on his or her current psychological profile. In addition, concern was raised in research, practitioner, and legal communities that the strong influence of therapist beliefs, especially in combination with highly suggestive and aggressive techniques such as hypnosis, could lead to the creation of false beliefs and memories about childhood trauma. The acceptance of recovered memories as legal testimony further increased the emphasis on the accuracy of such memories.

Interestingly, the shift in public interest was reflected in popular media. In an analysis of the content of four popular magazines—*Time*, *Newsweek*, *US News and World Report*, and *People*—sociologist Katherine Becker documented this change. According to Becker, in 1991, 80% of articles on childhood sexual abuse cases weighted toward the survivors, with their memories and therapeutic histories virtually unquestioned. By 1994, 80% of the articles focused on false accusations of abuse and supposedly false memories. A major influence in this shift of media attention was credited to an organization named the False Memory Syndrome Foundation, a support group for parents who claimed to be falsely accused of abusing their children. The foundation, whose board members include many eminent scientists and scholars, argued that there was little evidence that extensive traumatic events could be forgotten and later recovered. Moreover, individuals expressed concern that the social and therapeutic climate was causing an epidemic of false recovered memories.

For a number of years, the controversy was often treated as a black-and-white issue of whether recovered memories should be considered real or false. However, in recent years a more nuanced perspective has emerged in which most researchers and practitioners seem willing to accept that both false memories and authentic recovered memories do in fact occur. The distinct questions of the degree to which a traumatic memory is available over time and the degree to which that memory is accurate are beginning to be disentangled in human memory research. Within the topic of memory availability, the distinction between claims that events can be forgotten and claims that the memory for such events can be recovered has also been noted. Disagreements still exist, particularly with respect to estimation of the relative frequency of authentic and false recovered memories and the precise therapy practices that are or are not appropriate. Nevertheless, a spirit of moderation has emerged in which individuals from alternative perspectives have begun to work collectively to assess the current evidence and identify promising areas for future research.

III. EMPIRICAL RESEARCH ON RECOVERED MEMORIES

Numerous methodologies have been employed to study forgotten and recovered memories. Here, we present a brief sampling of the different studies. We begin with interview studies on patient populations and survey studies, discussing the advantages and disadvantages of these methods. We then discuss several prospective and case studies that have attempted to address some of the difficulties found in the earlier research.

A. Studies on Patient Populations

In one of the earliest empirical studies on forgetting of sexual abuse, Judith Herman and Emily Schatzow interviewed women in group psychotherapy for incest. In this sample, 38% reported that they had never experienced amnesia for the abuse, 36% reported experiencing moderate memory deficits, and 26% reported severe memory deficits. Relatedly, Elizabeth Loftus and colleagues interviewed a number of substance abuse patients, most of whom had experienced sexual abuse. Although most of those women reported always remembering the abuse, 12% reported consistently remembering only part of the abuse and 19% said they forgot abuse for period of time only later to have the memory return. However, results from this study must be taken with caution due to the effect of drug abuse on memory. In an additional study, Steven Gold and colleagues surveyed individuals entering therapy for sexual abuse. Approximately one-third reported that they "always contained a fairly complete memory of all or most episodes of abuse" and an additional 16% "remembered at least one episode of abuse in its entirety, but not all of them." About 14% reported having only a partial memory or flashes of recollection for some aspects of the event only. Ten percent "had a vague sense or suspicion but no definite memory" and almost another one-third "completely blocked out any recollection of the abuse."

B. Survey Studies

In an often-cited study, John Briere and Jon Conte surveyed hundreds of individuals in outpatient therapy, all of whom reported a history of child sexual abuse (CSA). One critical question on this survey asked "During the period of time between when the first forced sexual experience happened and your 18th birthday, was there ever a time when you could not remember the forced sexual experience?" Almost 60% of the respondents answered yes to this question. Similarly, Shirley Feldman-Summers and Kenneth Pope published the results of a randomized national survey of psychologists. Of the respondents who reported experiencing CSA, approximately 40% reported forgetting some or all of the abuse for a period of time. Non-sexual abuse was also reported to be forgotten. Interestingly, about half of these respondents said their abuse memories were triggered by therapy. Half reported having corroboration.

C. Methodological Issues in Patient and Survey Studies

These studies demonstrate that some people do indeed believe that they experienced abuse as children and subsequently had partial or no memory for the abuse. Furthermore, although there were a few cases of people still reporting difficulty in remembering abuse experiences, most believed they later recovered the previously inaccessible memories. Once again emphasizing the complexity of memory, awareness ranged from vague suspicions to "flashes of recollection" and full knowledge. The patient studies have the advantage of being more in-depth than standard survey studies, although surveys allow for the collection of vast amounts of data.

Although compelling, there are problems with these studies on therapist or patient populations. First, one might be concerned with surveying beliefs about memory in any individual who has been exposed to therapeutic notions of forgetting and recovery, as may happen in any therapy session. Only the study by Gold and colleagues attempted to account for this by conducting the memory interview at the intake interview. However, even they did not indicate which of their participants had received prior therapy. None of the studies can rule out the possibility that therapist or patients' beliefs about memory and trauma have tainted patients' recollection of their past memory. In fact, all retrospective studies suffer from reliance on the biased and difficult task of estimating past states of knowledge, especially in response to one or two nondetailed questions. Imagine trying to estimate if there was ever a period of time in which you had forgotten your sixth birthday party. There was probably a long period of time during which you did not think of the party, but it is very difficult to answer the question, "Could you have remembered if you tried?"

In addition, critics of these studies have raised the issue of corroboration for the abuse. Sexual abuse in particular is one of the most difficult crimes to find objective evidence for, and it is obvious that perpetrators have a great deal of motivation to lie and not confess their guilt. Additionally, one could argue that attempts to seek out corroboration in the therapeutic setting are bad for an atmosphere of trust in the patient. However, in scientific research one wants corroboration. Although Herman and Schatzow reported that three-fourths of the women were able to find independent corroboration, the authors relied on the patient reports of that evidence. Similar to the requirement of corroboration for the abuse, it is

difficult to give evidence that the memory was actually inaccessible had the individual tried to remember the event. Indeed, the study by Gold and colleagues reported their clinical observation that many individual's beliefs about their memories can change over the course of therapy (though this was in the context of an observation that most came to believe they had less complete memories than before). As discussed later, it is possible for individuals to believe they forgot something during a period of time in which they demonstrated awareness of that event.

D. Prospective Studies

Studies employing prospective methodologies have attempted to address the issue of corroboration by identifying individuals with documented abuse experiences and following them to later test their memory for the event. For example, Linda Williams published a study of 129 women whose childhood abuse experiences were documented due to required treatment at a hospital emergency room. In an interview conducted on average 17 years after hospitalization, the women were asked in detail about their abuse histories. Thirty-eight percent of the participants did not recall or did not report the abuse. Importantly, these women were no less likely to disclose personal information than the women who did report the specific abuse experience, suggesting this finding was not just due to an unwillingness to report the event. Williams also ruled out the possibility that all of the nonreporters were simply too young at the time the abuse occurred to remember it, and that the nonreporters likely represented original false reports. In addition to the group that seemed to have forgotten the abuse experience, approximately 10% of the participants reported recovered memory experiences. Based on an analysis of the hospital records, the women who reported recovery experiences had no more inaccuracies in their accounts than the women who reported continuous availability of the memory.

This type of study has the benefit of corroboration for the abuse. Furthermore, the interviews conducted in the Williams study and others like it allow for more in-depth questioning about the person and his or her experiences. Some have criticized the Williams study because it did not specifically ask about the abuse experience in question, thus why participants did not report the abuse experience cannot be definitively argued. In this study, the possibility cannot be ruled out that some individuals either felt disinclined to report the abuse or conflated the target incident with some other incident that they did recall. However, as described previously, the author made great efforts to rule out this possibility. Another limitation to prospective memory studies is that the evidence that they provide for forgetting only applies to those memories that *were not recalled* at the time of the interview. Since the recovered memories identified in these studies *were recalled* during the interview, these studies do not speak to the degree to which recovered memories are ever fully forgotten. Indeed, as discussed later, case study analyses have provided evidence that individuals may sometimes overestimate the degree that they had previously forgotten memories that they describe as recovered. A final limitation of all of the approaches outlined previously is that they aggregate data across cases, possibly obscuring potentially critical details of individual cases that may provide important clues about the nature of the phenomenon. Case study analyses help to fill this critical gap.

E. Case Studies

Another approach for investigating recovered memories is to engage in detailed retrospective analysis of the various elements of individual cases. For example, Jonathan Schooler and collaborators interviewed a number of individuals who reported having discovered previously forgotten memories of abuse. In one case, JR reported that following viewing a movie involving sexual abuse, he suddenly remembered incidents of being molested by a priest on camping trips when he was an adolescent. In another case, DJ reported that following a dinner with a neighbor whom she had not seen in many years, she suddenly recalled that this individual molested her numerous times between the ages of 5 and 7. In each of seven cases collected thus far, the investigators sought independent corroboration of the abuse, usually by contacting other individuals who the victim indicated had prior knowledge of either the abuse or the abusive tendencies of the alleged perpetrator. For example, in JR's case another individual indicated that the priest had attempted to molest him also, and in DJ's case, her mother reported that the alleged perpetrator had confessed. In each of these cases, independent evidence of the abuse was found. Moreover, in each case the individual reported a strong belief that there was a period of time in which he or she did not remember the abuse experience. The individuals also provided numerous details regarding their "memory discovery experiences," which were typically

described as involving shock, surprise, and a sudden unfolding of events. On the basis of the details extracted from these cases, Schooler proposed a number of possible mechanisms that may help to explain what he terms "discovered memories" of trauma.

In addition to Schooler's case studies, at least two other case studies have recently been published. Interestingly, both of these cases are prospective in nature. David Corwin and Erna Olafson published an incredibly detailed case history (including interview transcripts) of a young girl referred to Corwin for a court-appointed evaluation of allegations of physical and sexual abuse against the mother. At the age of 6, this girl made consistent allegations of abuse against her mother. At the age of 17, she requested to see the videotape of herself because she claimed she could not remember what had actually happened to her. During a subsequent interview between Corwin and the girl, she seemed to have a recovery experience, in which the allegations of abuse suddenly, and with great emotion, came back to her. Interestingly, although some of her memory corresponded to what was on record, she seemed to "remember" other information for which there was no evidence and no previous claims.

Sunita Duggal and Alan Stroufe also published a prospective study on a young girl in Minnesota who reported being abused by her father at age 4. Although the father always denied the abuse and the police were not able to obtain enough evidence to press charges, the girl was believed by her therapist, mother, and caseworker. The girl displayed memory for the abuse during all yearly research interviews conducted through the third grade. However, during annual interviews at ages 16 and 17, in structured interviews she denied experiencing sexual abuse in response to general "negative event" questions and questions specifically about sexual abuse. At age 18, she reported recovering her memory for the abuse.

In each of these cases there are compelling reasons to believe that the individual's memory recovery experiences corresponded to actual abuse events because at least some independent corroboration was obtained. However, the toughest critics may point out that corroboration does not absolutely guarantee that the events took place. Even in the prospective cases, there is no definitive evidence of the abuse. Furthermore, there is no way to prove forgetting occurred. Interestingly, in two of Schooler's case studies, participants judgments about their forgetting were contradicted by close others. In both cases, ex-husbands reported that they had been told about the event; however, the disclosure was remarkably devoid of affect. Although it is possible that the husbands are the ones who are wrong, the individuals could have made errors in the degree to which they forgot the event or the period of time during which the forgetting occurred. Collectively, the reports and the corroborative evidence suggest the strong likelihood that recovered memories really can occur and can correspond to actual events.

F. Summary of Research Documenting Discovered Memories

Despite the methodological difficulties involved in documenting recovered memory experiences, there seems to be at least reasonable agreement that individuals can have sincere recovery experiences corresponding to actual abuse that was perceived to have been entirely forgotten. Although the precise degree of forgetting in these cases may be unclear, the fact that individuals regularly perceive themselves to have forgotten the abuse indicates that at a minimum the memory was less accessible (i.e., less likely to come to mind) prior to the recovery, relative to after the recovery. In other words, even the most conservative assessment of the evidence indicates that traumatic memories can fluctuate in their persistence, with periods of time in which the memories are relatively less accessible. It is therefore appropriate to consider the various mechanisms that might in principle influence the accessibility of traumatic memories. In addition, given that estimations of forgetting are often retrospective, it is also important to consider the factors that might influence the judgment of past accessibility.

In reviewing variables that might contribute to the actual and perceived reduction in the accessibility of memories, it is important to emphasize that much research is still needed in these areas before definitive conclusions regarding the relative contribution of various mechanisms will be possible. Nevertheless, it is useful to outline the various mechanisms that current evidence suggests might be involved.

IV. THEORIES AND MECHANISMS OF RECOVERED MEMORIES

A. Mechanisms Specific to Emotional Events

1. Psychodynamic Theories of Repression

Freudian notions of repression argue that painful or threatening material can be selectively, though

effortfully, kept out of conscious awareness. According to Freud, the memory or impulse still exists as a disconnected idea that can affect the person despite lack of conscious awareness. Importantly, the material can later return to conscious awareness. In this view, repression is a protective device. Forgetting occurs to ease pain. Whether Freud considered repression to be an intentional or unintentional occurrence is still a matter of debate since Freud was inconsistent in his writings. Similarly, how this selective awareness occurs is also unknown. Regardless, the notion of repression remains well-known by the layperson and is popular within many mental health communities.

In 1990, David Holmes published an article titled "The Evidence for Repression: An Examination of Sixty Years of Research." Although he noted the observation that Freud was often ambiguous and inconsistent about the exact definition of repression, Holmes decided to investigate experimental evidence for the "conventional use" of the concept. According to Holmes, this conventional definition contains three necessary elements: (i) the selective forgetting of painful information, (ii) it must be involuntary, and (iii) the information can be recovered under the right conditions. His review addressed several different experimental approaches to finding a mechanism for this definition of repression. These areas included the differential recall of pleasant and unpleasant memories, differential recall of completed and incompleted tasks, changes in recall associated with the introduction and elimination of stress, individual differences in repressive tendencies, and perceptual defense. Based on the results of his literature review and some resulting investigations, Holmes concluded that there is no laboratory evidence of a mechanism for Freudian repression.

Holmes' article has been frequently cited as evidence that recovered memories cannot occur. However, it is important to point out that his work investigates mechanisms for repression, not the phenomenon of memory inaccessibility. Furthermore, since the publication of Holmes' studies there have been a number of laboratory studies that offer evidence of repression-like mechanisms. Several studies have investigated the memory performance of individuals who are classified as "repressors"—people who report low anxiety and simultaneously report using various defensive strategies (such as trying not to think about things that bother them). In a study by Penelope Davis it was found that repressors recall fewer negative childhood memories than do nonrepressors. In a related study by Lynn Myers and colleagues, it was found that when instructed, repressors were better than nonrepressors at forgetting negative words that they had recently read. Of course, findings such as these do not necessarily demonstrate that individuals can massively repress severely emotional experiences, but no laboratory experiment could ethically be expected to demonstrate such repression.

Currently, the following conclusions regarding evidence for repression are probably warranted. First, the question of whether or not individuals can forget extensive incidents of trauma is independent of whether the specific mechanism of repression contributes to such forgetting. Second, although laboratory evidence has been difficult to obtain, there have been a number of studies that can be interpreted as supporting the notion that defensive strategies may lead to the forgetting of some negative material. Finally, regardless of one's opinions on the laboratory evidence for repression, its applicability to the forgetting reported in recovered memories is limited because traumatic experiences are ultimately much more emotional, and the alleged forgetting is reportedly much more extensive, than anything that can be expected to be produced in the laboratory.

2. Dissociation during Traumas

Similar to the notion of repression, theories on traumatic dissociation propose that some individuals may psychologically separate themselves from overwhelming negative experiences. Pierre Janet first identified the phenomenon in the late 1800s and proposed that the intense emotion aroused during trauma could interfere with the assimilation and integration of perceptions, thoughts, and experiences. Current understanding of the phenomenon argues that there are three components to an acute dissociative response: derealization (alteration in one's perceptions), depersonalization (alteration in one's sense of self and connection to one's own body), and memory disturbances. Further research has identified a persistent, dissociative personality trait that seems to lie on a continuum throughout all members of the population. In addition to having a joint component of derealization and depersonalization and a memory disturbance component, this concept includes a measure of absorption, the ability to become "lost" in one's thoughts or activities.

The concept of dissociation is central to current research and psychiatric theorizing about trauma. Acute dissociative responses have been identified in survivors of overwhelming traumas such as combat,

sexual abuse, accidents, natural disasters, and fires. Furthermore, it has been argued that immediate dissociative responses to trauma predict poorer psychological recovery. Additionally, trait dissociative tendencies seem to be higher in individuals with a traumatic history. How acute dissociative responses and trait dissociation are related is unclear. Researchers have distinguished between normal and pathological dissociation in the trait variable, arguing that the absorption factor does not seem to be related to pathological dissociation. Thus, it may be that a preexisting dissociative tendency can lead to an acute dissociative response to trauma, which in turn increases one's dissociative tendencies. Of primary interest to the current discussion is how an acute dissociative response or dissociative tendencies may affect memory.

One possibility is that an acute dissociated state leads to poor encoding of a traumatic event. In this situation, trauma memories could be fragmentary or missing but could not be completely recovered. The second possibility is that the dissociated state is functionally distinct from the normal state of mind, leading to state-dependent effects in which dissociated material cannot be retrieved until one once again enters the dissociated state of mind. Experimental research has demonstrated that memory can improve if one is in the same mood or physical location at the time he or she tries to recall something as he or she was at the time he or she learned the information. Thus, material learned while dissociated may be difficult to remember when one is in a normal state of mind but more easily retrieved when one returns to the dissociated state. In a study using highly hypnotizable undergraduates, Heidi Sivers and Gordon Bower gave participants a series of items to learn while in either a hypnotically induced ''dissociated'' or ''normal'' state. They then asked the participants to later recall the items, again while in a dissociated or normal state. Although they had hypothesized that state-dependent memory would be found, instead they found what appeared to be very poor recall of the learn-dissociated items regardless of retrieval state. This suggests that dissociation was indeed negatively affecting the original encoding of the material learned.

Jennifer Freyd and colleagues relatedly found that dissociative trait can affect attention and memory. Using a measure called the Dissociative Experiences Scale (DES), they identified members of the general population who were either high or low on dissociative tendencies. They discovered that highly dissociative individuals are worse at filtering out irrelevant material in tasks that require attending to a select portion of incoming information. However, these same people are better at tasks that require attending to more than one thing at a time. In a different task, high and low DES individuals were asked to learn portions of a list containing neutral and trauma-relevant words. If asked to simply recall everything, there were no differences between the two groups. However, when their attention was divided at learning by an additional task, high DES participants recalled fewer trauma-related and more neutral words than did low DES participants. In both learning conditions, high DES participants recognized fewer trauma-related and more neutral words than did low DES participants. These findings suggest that dissociation may be adaptive in keeping threatening information from awareness in certain circumstances. In particular, attentional context may be a central factor in understanding when dissociative tendencies are most likely to help people keep threatening information from awareness. Thus, the lack of integration of experiences, memories, and thoughts creates an environment that requires constant divided attention and encourages cognitive strategies for functioning efficiently in such environments.

Our understanding of state and trait dissociation is thus still growing. For now, it remains a concept based primarily on clinical observation and self-report, lacking a known cause or mechanism. This had led many research psychologists to view the concept and its attendant hypotheses skeptically. Although much more research is needed before we fully understand what dissociation is, at a minimum measures that purportedly measure dissociation have proven to be a useful tool for predicting both the response to trauma and performance on a number of standard cognitive tasks. This is currently an area being heavily researched, and it is hoped that our understanding of the concept and its relation to trauma and memory will continue to grow.

3. Brain Theories on ''Processing''

There now exists a reasonable amount of evidence to support the idea that emotional information is processed and remembered through partially different pathways than nonemotional memories. Highly arousing situations may increase the involvement of an area of the brain called the amygdala. Interestingly, Joseph LeDoux has discovered evidence that there are two primary information pathways involved with the amygdala. One pathway is fast, has only generic

information, and does not involve interaction with the areas of the brain involved in the higher processing systems of thinking, reasoning, and consciousness (the cortex). A second, slower and more refined pathway sends input to the amygdala through the sensory cortex. LeDoux suggests that the fear reaction system involves parallel transmission to the amygdala from these two pathways. The subcortical pathway provides a crude image of the external world, whereas a more detailed, accurate, and perhaps conscious representation comes from the cortex. Interestingly, a third input system comes from the sensory-independent hippocampus. The hippocampus is the area of the brain argued to be primarily responsible for normal long-term memory. This hippocampal pathway seems to be involved in the integration of individual stimuli (sights, sounds, and sensations) during learning in order to create a broader context. The hippocampus also seems to play a role in integration at retrieval.

There is evidence that extreme and continuous stress damages the hippocampus. For example, monkeys who died due to extremely stressful living conditions showed damage to subregions of the hippocampus. Relatedly, many different research labs have found decreases in hippocampal volume in Vietnam War veterans with PTSD compared to noncombat control subjects. Based on this research, it is possible that the fear-conditioning pathways may correspond in some way to claims made regarding traumatic memory. The first two pathways are sensory specific, with one occurring completely without cortical involvement, suggesting it may be possible to process some fear-related information with little or no conscious awareness. The second pathway involves some areas of higher processing. Thus, it is possible that the sensory-dependent memories (memories that do not arise unless one is exposed to a specific, sensory cue) are related to representations held within this second system. Because this system is responsible for stimuli discrimination, it may have important ramifications for cue sensitivity. Processing in these cortical areas would allow the individual to be consciously aware of sensory information during the experience. However, extreme terror and fear would interfere primarily with the third pathway projecting to the hippocampus, perhaps the area responsible for integrating those representations into a complex whole. This would result in the isolated sensory recollection posited by some models of traumatic memory. Although these suggestions are highly speculative, there is currently a great deal of research in investigating the neural correlates of traumatic forgetting and memory recovery.

4. Betrayal Trauma Theory

Betrayal trauma is a theory proposed by Jennifer Freyd that addresses both the how and why issues of forgetting of traumatic experiences. In this theory, she argues that amnesia for childhood abuse exists, not for the reduction of suffering but because not knowing about abuse by a caregiver is often necessary for survival. From a logical analysis of developmental and cognitive research, she argues that a cognitive information blockage under certain conditions, such as sexual abuse by a parent, can be expected. This is the "why" portion of the argument that childhood abuse is forgotten.

Betrayal trauma theory proposes a two-dimensional model of trauma. One dimension addresses terror, the emotional state required in the definition of traumatic response. This dimension corresponds to threats to life—things that actually can cause one bodily harm and often do. Another dimension is that of betrayal and threats to social relationships. In this case, the event involves a treacherous act by someone depended on for survival. Some traumas are high on both these dimensions; for example, sadistic abuse by a caregiver, the Holocaust, some combat experiences, and many childhood sexual abuse situations. These events are both terrorizing and involve a betrayal of a relationship. Although the fear dimension is important for some of the anxiety responses found in PTSD, amnesia is especially likely to occur for the events that are high in betrayal.

Betrayal trauma theory leads to specific predictions about the factors that will make amnesia most probable. One notable factor is the individual who is perpetrating the abuse. According to this theory, childhood sexual abuse is more likely to be forgotten if it is perpetrated by a parent or other trusted caregiver. If a child processes the betrayal in the normal way, he or she will be motivated to stop interacting with the betrayer. Essentially, the child needs to ignore the betrayal in order to preserve the attachment. Thus, for a child who is dependent on a caregiver, the trauma of abuse, by the very nature of it, demands that information about the abuse be blocked from mental mechanisms that control attachment and attachment behavior. How is a child to manage this on a long-term and sometimes nearly daily basis? How is the child to succeed at maintaining this necessary relationship when a natural response is to withdraw from the source of the pain? Betrayal trauma theory proposes that the child blocks the pain of the abuse and betrayal by isolating knowledge of the abuse/betrayal

from awareness and memory. There are various avenues for achieving this isolation, one being conscious memories without affect and another being the isolation of knowledge of the event from awareness.

Freyd relies on numerous concepts from cognitive psychology to support the "how" argument of betrayal blindness. She points to mental mechanisms for processing information in parallel, selective attention, sharability of information, and the time course of complex information processing to support the fact that knowledge can be isolated by interrupting the extended processing of complex events. Furthermore, she mentions research on inhibition and recovery of well-formed memories. In summary, there are multiple ways for the abused child to disrupt knowledge integration and awareness of the abuse while facilitating the important and crucial relationship. Furthermore, there are multiple ways for the adult survivor of childhood abuse to recover these memories, and these will depend in part on how the memories were isolated in the first place. At the same time, this cognitive plausibility does not negate the potential for false memories to occur. Indeed, the cognitive mechanisms that support knowledge isolation and recovery may be in part the same mechanisms that may support memory errors.

To support the notion that amnesia will be more likely the more dependent the victim is on the perpetrator, Freyd reanalyzed three sets of extant data, including those described in the Feldman-Summers and Pope and Williams papers articles discussed previously. This investigation indicated that amnesia rates are higher for parental or incestuous abuse than for nonparental or nonincestuous abuse. Furthermore, she and her students collected survey data questioning individuals' memory for a wide array of specific situations of physical, emotional, and sexual abuse in childhood. The preliminary results support the prediction that the greater the victim's dependence on the perpetrator, the less persistent are memories of abuse. Together, these data sets suggest that social dependence may play an important role in memory for traumatic events.

Although the betrayal trauma theory has considerable potential, current evidence in support of it is largely preliminary and exclusively correlational in nature. Although a relationship has tentatively been observed between reported memory persistence and the relationship of the victim to the alleged pertpetrator, it does not necessarily follow that the cause of this relationship is betrayal trauma processes. In principle, a variety of other potential factors could account for these correlations, including age at the time of the event, differences in the interpretations of abuse associated with caretaker vs stranger abuse, differences in the likelihood of talking about the two types of abuse, and/or differences in the likelihood that the memories of the two types of abuse may be fabricated. Freyd and colleagues are currently measuring some of these potentially confounding variables and will be able to evaluate statistically the contribution of these covarying factors in predicting memory impairment. Preliminary analyses indicate that one factor, age at the time of the event, does not account for memory persistence over time. Some issues will require specialized populations. For instance, to evaluate the possibility that there is a difference in the likelihood that memories of types of abuse are fabricated, it will be necessary to use a prospective methodology with documented abuse samples. In correlational research there is always the possibility of unmeasured confounds; because we cannot ethically vary many of the factors of interest related to real abuse, the best we can currently do is to systematically evaluate the contribution of covarying factors that we identify as possibly accounting for differences in rates of reported forgetting.

None of the theories discussed here—repression, dissociation, or betrayal trauma—are exclusive of many standard memory mechanisms that can affect whether any type of information is encoded, stored, or recalled over time. In fact, betrayal trauma theory incorporates many of them. In the next section, we review a number of more standard memory mechanisms that have been proposed to play a role in traumatic amnesia.

B. Well-Established, or Non-Trauma-Specific, Memory Mechanisms

Although many accounts of recovered memory have focused on processes that may be unique to trauma, other approaches have emphasized the various general memory/forgetting mechanisms that could be involved. There are several well-documented general memory mechanisms that seem readily applicable to the current discussion. Here, we review the mechanisms we believe are most relevant.

1. Simple Forgetting

Many observers have noted that we routinely forget all sorts of experiences in life. In one survey, Don Read

found that a significant proportion of people reported recovering memories for all sorts of significant but nontraumatic life experiences. Thus, in many cases the forgetting of childhood abuse may simply reflect the passage of time and the fact that we simply cannot constantly remember the plethora of experiences from our past.

2. Directed/Intentional Forgetting

In many cases individuals may try to forget their unpleasant abuse experiences. In fact, there is a large body of laboratory research indicating that people can intentionally forget information when they try. In numerous word-list learning paradigms it has been repeatedly demonstrated that individuals can intentionally forget something. In a typical experiment, participants are given an initial list of words and then told to forget that list and focus on remembering a new list of words. They are then surprised with a memory test for all of the items, including items presented on the first list. The results demonstrate that memory for items on the first list is worse in individuals who were instructed to forget the list than in individuals who were told to remember both lists. Thus, individuals can intentionally forget information.

Many have interpreted these results as due to selective rehearsal. It could be that the first list is simply rehearsed less in the "forget" than the "remember" condition. Relating this to forgotten memories of childhood sexual abuse, a lifetime of avoiding the thought of the abuse or keeping the abuse secret would lead to poor memory due to a lack of rehearsal. Additional directed forgetting studies have attempted to equate intentional rehearsal or use incidental memory and have still found poorer memory for the first list. Furthermore, if nonrecalled forget and remember items are later re-presented for learning, evidence for a "release from inhibition" is found in that original forget items are recalled better than original remember items. This line of cognitive research has been used as support for the notion that information can be intentionally inhibited from conscious awareness and later recovered.

3. Interference Theories of Memory

Countless examples of word-list learning experiments have demonstrated that if two related pieces of information are learned, practice of one piece of information can interfere with the ability to remember the other piece of information. This can take the form of prospective interference, in which past information interferes with the ability to retrieve new information. One example of this phenomenon is having difficulty remembering a friend's married name because her maiden name keeps popping to mind. Alternatively, retrospective interference can occur, in which the learning of new information interferes with the ability to recall old information; the new married name gets in the way of recalling the friend's maiden name. In either case, recall of one set of information "interferes" with the ability to recall the other. When considering traumatic experiences, imagine the child who's favorite uncle sexually abused him on one occasion and takes him to a ball game on another. The child may rehearse the uncle–ball game association repeatedly while never rehearsing the uncle–abuse experience due to pressure not to disclose, threats or denial from the uncle, or numerous other reasons. According to standard interference theories of memory, the strengthening of the uncle–ball game association would actually decrease the ability to recall the uncle–abuse situation.

4. Change in Understanding/Reinterpretation

It is frequently proposed that an individual who experiences CSA may not fully understand the event at the time it originally occurs. Knowing when an unfamiliar type of touch is acceptable instead of abusive, for example, may require understanding of social norms and the intentions of the individual doing the touching—a difficult task for an adult, let alone a young child. It has been demonstrated that individuals have very poor memory for information for which they do not have a "schema" or knowledge system. For example, in one research study, participants who read an ambiguous passage had very poor memory for the content of the passage unless a title presented ahead of time indicated what the story was about. In this laboratory study, it was found that if the title was presented after the story, memory was even worse than if no title was provided. Thus, if a person has no way to label, understand, or describe an experience, memory for that experience may suffer.

Even if an event is learned and understood in one manner, a later reinterpretation of the event as "abusive" may cause one to feel like the event is being recalled for the first time. Furthermore, changes in the interpretation of an event may activate previously inaccessible information. Again turning to psychological research studies, it was demonstrated that individuals who were told to read a description of a house from the perspective of a robber remembered

different information than did individuals who read the same description from the perspective of a home buyer. Interestingly, when told afterwards to take on the other role, participants recovered previously unremembered information in line with the new perspective. Therefore, an individual who interprets an event in a different way could actually retrieve additional information related to the new interpretation.

5. Encoding Specificity/State Dependency

As mentioned previously, an additional relevant mechanism is found in encoding specificity or state dependency theories. Encoding specificity theory states that the probability of retrieving a memory is maximized when retrieval conditions correspond to the encoding conditions. Similarly, state dependency argues that memory improves if one is in the same state of mind at the time of recall as when he or she originally learned the information. For example, psychologists Baddeley and Godden had scuba divers learn two lists of words, one while under water and one while on dry land. On a subsequent memory test, it was discovered that individuals recalling information while on dry land remembered more words from the list learned in the same physical location. Likewise, individuals recalling under-water recalled more under-water-learned words. Similar results have been found in studies comparing happiness to sadness and drunkenness to sobriety, among other mind states. Strikingly, when considering the case examples given in support of recovered memories, in all the cases collected by Schooler and the Corwin case there was notable correspondence between the original abuse situation and the situation in which the memory was ultimately recalled. Furthermore, it is possible that the proposed knowledge isolation, or the fragmentary nature of traumatic memory, could make such memories even more dependent on highly specific cues, thus making the events irretrievable except in very limited circumstances.

C. Speculative, Non-Trauma-Specific Memory Mechanisms

In addition to the previously mentioned well-established, non-trauma-specific memory mechanisms, there are more speculative memory mechanisms that may play an important role in recovered memories. Although these mechanisms remain to be definitively established, they are largely consistent with extant evidence and are worthy candidates as possible accounts for at least some recovered memories.

1. The ''Forgot It All Along'' Effect

In several of his cases, Schooler found that the individuals underestimated their prior knowledge about the event as evidenced by the fact that others reported they had talked about the abuse during the time that they thought they were being amnesic. Schooler likens this to a similar bias in individual's estimates of past knowledge, the ''knew it all along'' effect. The premise of the knew it all along effect is that a person who is told something new comes to believe that he or she knew it all along. This happens because the current knowledge state is used to infer the earlier knowledge state. Although there has been little research to date on cases of underestimations of prior knowledge, it seems reasonable to suppose that if one can use one's current knowledge state to overestimate prior knowledge, one may also use it to underestimate prior knowledge. In the context of an emotional onrush associated with thinking about memories of abuse, individuals may assume that they had no previous knowledge about their abuse. They may reason, ''If I'm this shocked and surprised now, then I must have previously completely forgotten about the experience.'' In short, individuals may misattribute the emotional onrush associated with thinking about the event to the emotional onrush of discovering the memory. Future research is needed to determine whether and, if so, to what degree this intriguing mechanism, which is consistent with apparent mischaracterizations of forgetting in several of Schooler's cases, applies more generally to recovered memory cases.

2. Precipitous Forgetting of Nocturnal Experiences

Recently, Schooler proposed that there may be something ''special'' about nocturnal experiences that could lead them to be forgotten almost immediately after occurring. Although characterizations of forgetting as precipitous are certainly not ubiquitous, it has been reported. One possible explanation is that various physiological processes that contribute to dream forgetting may also contribute to the forgetting of nocturnal abuse. Individuals often, indeed usually, forget dreams, even traumatic and disturbing ones. Schooler argues there are a number of striking

parallels between dream forgetting and allegations of forgetting of sexual abuse. First, sexual abuse, like dreams, often occurs at night while the individual is in bed. Second, like dreams, sexual abuse experiences are often bizarre, occur in isolation, and may be difficult to reconcile with preexisting schemata and other events. In fact, descriptions of dissociated experiences often include the statement "it was as if I was dreaming." These parallels between dreams and nocturnal abuse may both contribute to the forgetting of such abuse and to the dismissal of such recollections as being merely "bad dreams," especially in children, who have lesser ability to distinguish between reality and fantasy. In support of this notion, laboratory studies have demonstrated increased forgetting for materials presented immediately prior to sleep onset and immediately after awakening. However, there is no direct empirical evidence that such memories can later be recovered.

3. Metaconsciousness

Recently, Schooler proposed a theory of metaconsciousness that assumes that experiential awareness (i.e., the contents of phenomenological experience) can be distinct from metaconsciousness (i.e., one's explicit understanding of his or her phenomenological experience). In this context, recovered memories involve changes in individuals' metaconsciousness of the abuse. In some cases, they may involve the gaining of a different metaconsciousness of the meaning of an experience, which may become confused with the discovery of the memory. The result of such confusion would be the sometimes erroneous belief that the memory is just now being accessed for the first time (similar to the notion of reinterpretation discussed earlier). In other cases, the memory discovery may involve regaining a prior metaconsciousness that was avoided for some time. In still other cases it may involve the gaining of a previously nonexistent metaconsciousness of the experience. A variety of factors ranging from the very straightforward (e.g., age, lack of discussion, and stress) to the more esoteric (e.g., dissociation and nocturnal cognitive processing) may prevent incidents of abuse from being initially encoded with metaconsciousness. Such nonreflected memories, particularly when they are aschematic and disjunctive with other experiences, may continue to elude metaconsciousness until a specific contextual retrieval cue is encountered. Once recalled in the light of metaconsciousness, individuals may understand what happened to them, and this discovery may fundamentally change their view of their personal histories. Again, research is needed to establish the role that changes in metaconsciousness of abuse may have in contributing to recovered memories.

D. Conclusions Regarding Theories and Mechanisms of Recovered Memories

Currently, understanding of the mechanisms behind recovered memory experience has resulted in many advances, but there is much to uncover. It seems likely that no one cognitive process will be able to explain all forgetting and recovery of awareness for traumatic events. Indeed, a range of devices, from standard memory mechanisms to processes unique to trauma, should come into play, influenced by the nature of the trauma, the situation in which it occurs, and the immediate and subsequent reaction of the survivors and their social network. Thus, scientific research and discussion on the topic of recovered memories must reflect this complexity and not attempt to reduce the answer of whether traumatic experiences can be forgotten or if they can be accurately recovered to a simple "yes" or "no."

Memory in general is a reconstructive process. We use our current knowledge and understanding to recreate our knowledge of the past. Thus, all memory has the potential for inaccuracy. For centuries, the reliability of our knowledge of the past has been discussed. Of particular relevance to the current discussion is how emotion influences memory. We know that emotion influences the persistence or availability of events over time. Within the range of everyday events, we tend to remember emotionally arousing events more than the mundane. For extraordinary events, extreme emotion is claimed to make an experience impossible to forget (as seen in traumatic flashbacks of PTSD) or difficult to remember (as discussed in this article). What is currently unclear is the degree to which emotion influences the probability of inaccuracies in memory.

V. ISSUES OF ACCURACY AND INACCURACY IN RECOVERED MEMORIES

Although there are good reasons to believe that many recovered memories of abuse correspond to actual events, there are also compelling reasons to be concerned that some recovered memories may be false. A variety of lines of research raise the specter of false recovered memories.

A. Research Supporting False Memories

First, individuals can remember, sometimes in excruciating detail, memories of events that are extraordinarily unlikely to have occurred. For example, Michael Persinger found that individuals can recover memories of alien abductions in a manner that at least superficially resembles that associated with some recovered memories of sexual abuse.

Second, under certain experimental conditions, subjects can be induced to recall "memories" of disturbing events that never happened. For example, Ira Hyman and colleagues planted, in a sizeable minority of participants, a variety of mildly upsetting and somewhat bizarre memories, such as spilling punch on a bride's parents at a wedding.

Third, a variety of psychotherapeutic techniques such as visualization repeated retrieval attempts, dream interpretation, and hypnosis can increase individual's beliefs that unlikely events actually occurred. For example, Mary Anne Garry and colleagues found that visualization techniques increased many peoples' beliefs that they might have experienced events (such as putting a hand through a window) that they previously reported were very unlikely to have happened.

Fourth, these techniques correspond, with disturbing closeness, to those argued to be used by a sizeable minority of clinicians in their aggressive efforts to recover memories of abuse. For example, in a national survey of licensed practitioners, Mellisa Polusny and Victoria Follete found that more than 25% of therapists reported using guided imagery, dream interpretation, bibliotherapy regarding sexual abuse, referral to sexual abuse survivors' group, and free association of childhood memories as memory retrieval techniques with clients who had no specific memory of childhood sexual abuse.

Finally, many individuals with recovered memories conclude that their memories are false. For example, a review by Joseph de Rivera reported that more than 300 people have retracted charges of sexual abuse based on memories recovered in psychotherapy.

B. Discussion of False Memory Research

The previously mentioned findings raise the real concern that individuals may, as a consequence of aggressive memory therapy techniques and other social pressures, develop recovered memories for events that never happened. However, just as some cautions are appropriate in interpreting the evidence in support of authentic recovered memories, so too some caveats are in order in interpreting the evidence for false memories.

A central potential limitation of the evidence for false memories is that for ethical reasons it is simply not feasible to attempt to induce false memories that are as emotionally disturbing as sexual abuse. Thus, some have raised the question of whether research on the more benign false memories that have been produced in the lab (such as being lost in a mall or spilling punch on someone in a wedding) would generalize to falsely recalling being abused by one's parents. Although the issue of generalization is important, it should be pointed out that real-world recollections of highly unlikely events, such as past life, prenatal, and UFO-associated traumas, suggest that false memories of even highly disturbing experiences are possible.

Furthermore, it has been argued that false memory research investigating the creation of memory errors, including those that have been characterized as involving the creation of false memories, has not been focused on the phenomenology of recovered memories per se. In other words, it has not fully addressed the experience of the individual at the time of recollection. Thus, is it unclear if the false memory is experienced as recovered from a previously inaccessible state or as a continuous memory. On the basis of such observations, it has been suggested that false memory research may only be applied to the question of memory accuracy when individuals believe they have always remembered something. However, in some laboratory studies, such as that of Ira Hyman mentioned previously, participants typically denied the suggested experiences when first asked about them. Only after engaging in extensive visualization did some subjects come to remember the suggested events. This process of nonrecall followed by considerable effort and eventual recall is in fact akin to recovered memories. Therefore, although this is an area of dispute, it seems clear that false memories of both continuous and recovered memories can occur and are important topics for further investigation.

Another concern regarding false memory studies involves the degree to which the ideas that individuals generate are best described as false memories. Many studies that have been characterized as involving the creation of false memories have not actually caused individuals to specifically recall events that never occurred but rather have caused them to believe that such events might have occurred. For example, as

noted previously, when individuals are encouraged to imagine various unlikely events, they subsequently estimate that such events are more likely to have occurred. However, in research on the topic, individuals did not actually report remembering these events. This and other studies of its type might be better characterized as involving false beliefs about memory rather than false memories per se. Although false beliefs about memory are clearly relevant to this discussion, future research is needed to determined whether a false belief can transform into a full-fledged false recollection.

Critics have also noted that in much of the false memory research there is no way to be certain that the allegedly false events did not actually take place. For example, in one very influential false memory study by Elizabeth Loftus and Jacqueline Pickerel, college student participants came to remember being lost in a mall, even though a parent had indicated that such an experience had not occurred. Given how common it is to be lost in a mall (probably many readers have had this experience), these findings could be interpreted as cases in which the parents forgot the critical incident. Indeed, little research has specifically examined whether parents' memories are necessarily any more accurate than their adult children's. Research that increases the estimated probability that highly unlikely events occurred beyond what can be reasonably expected (e.g., 25% of students remembering spilling punch on a bride at a wedding) provides more conclusive evidence that memories can be planted for specific events that never occurred. Nevertheless, in these cases it seems quite plausible that individuals' false memories might incorporate details from related events that did occur (spilling a drink at some other social occasion). In more naturalistic cases it is similarly possible that individuals who report UFO traumas with sexual elements experienced and forgotten more mundane sexual abuse. Research is needed to investigate the relationship between actual experiences and the probability of accepting a false event as having occurred.

Relatedly, some have suggested that in order for individuals to develop false memories they must have some sort of "schema" or "knowledge structure" through which to create the false information. For example, Kathy Pezdek and colleagues were unable to plant false memories for experiences that individuals had little knowledge about (e.g., receiving a rectal enema) or that they perceived as being highly implausible (e.g., Jewish students receiving communion). Importantly, however, one of the central components of many therapy practices, self-help groups, talk shows, and books is that they provide individuals with a clearer picture of abuse scenarios and persuade them that such scenarios might in fact apply to them. The resulting increased knowledge and plausibility of the abuse scenarios may be just what it takes for the abuse suggestion to take hold.

Ultimately, our understanding of the recovered memory phenomenon will require us to develop a broader appreciation of the various ways in which individuals can acquire false beliefs about their personal memories. We need to continue to investigate the various conditions that can lead individuals to accept false memories as being true. At the same time, we also need to understand the likely (though less often considered) factors that may cause individuals to reject true memories as being false, as for example might occur following the application of pressure from an authority figure insisting that an event never occurred (such as a perpetrator denying abuse). In addition, we need to attend to the various mechanisms that may lead individuals to generate false beliefs about the degree to which a memory had or had not persisted over the years.

VI. CONCLUSIONS

In the new millennium, we can be hopeful that the old polemics regarding whether recovered memories are false or authentic will increasingly be replaced by a more nuanced understanding of the issue. It will be understood that recovered memories may vary in their degree of accuracy, ranging from largely accurate to entirely false, with many gradations of gray in between. Moreover, the issue of accuracy in recovered memories will be carefully separated from consideration of the degree to which they are forgotten. Careful analyses of the variables that may lead individuals to perceive themselves to have recovered long-lost memories of trauma seem likely to identify a plethora of important mechanisms, some trauma specific and others general to all memories, and some known for years and others only recently conjectured. Such advances will help to reveal the important insights into the human mind that can be gained by exploring this unequivocally remarkable phenomenon. It is hoped that these insights will help and perhaps even provide some consolation to the untold numbers whose personal tragedies serve as the inspiration for this research.

See Also the Following Articles

CONSCIOUSNESS • DREAMING • MEMORY, OVERVIEW • SEXUAL BEHAVIOR • UNCONSCIOUS, THE

Suggested Reading

Bower, G. H., and Sivers, H. (1998). The cognitive impact of traumatic events. *Dev. Psychopathol.* **10**(4), 625–653.

Freyd, J. (1996). *Betrayal Trauma: The Logic of Forgetting Childhood Abuse*. Harvard Univ. Press, Cambridge, MA.

Lindsay, D. S., and Read, J. D. (1994). Psychotherapy and memories of child sexual abuse: A cognitive perspective. *Appl. Cognitive Psychol.* **8,** 281–338.

Loftus, E. F. (1997). Memories for a past that never was. *Curr. Directions Psychol. Sci.* **6,** 60–65.

Nadel, L., and Jacobs, W. J. (1998). Traumatic memory is special. *Curr. Directions Psychol. Sci.* **7**(5), 154–157.

Pezdek, K., and Banks, W. P. (Eds.) (1996). *The Recovered Memory/False Memory Debate*. Academic Press, San Diego.

Putnam, F. (1997). *Dissociation in Children and Adolescents: A Developmental Perspective*. Guilford, New York.

Readm, J. D., and Lindsay, D. S. (Eds.) (1997). *Recollections of Trauma: Scientific Research and Clinical Practices*. Plenum, New York.

Schacter, D. L. (1996). *Searching for Memory*. Basic Books, New York.

Schooler, J. W. (2001). Discovering memories in the light of meta-consciousness. *J. Aggression Maltreatment Trauma* **4**(2), 105–136.

Shobe, K. K., and Kihlstrom, J. F. (1997). Is traumatic memory special? *Curr. Directions Psychol. Sci.* **6,** 70–74.

Singer, J. L. (Ed.) (1990). *Repression and Dissociation*. Univ. of Chicago Press, Chicago.

Reinforcement, Reward, and Punishment

DANIEL T. CERUTTI
Duke University

GLOSSARY

contingency A conditional "if–then" relation between events such as responses and reinforcers.

discrimination Differential responding to stimuli that results when the stimuli signal the presence of a reinforcement contingency.

operant Behavior that is modified by its consequences, as in reinforcement and punishment.

punishment A reduction in the frequency of a behavior that results from its consequences. In positive punishment, behavior produces aversive stimuli such as illness; in negative punishment, behavior causes the loss of appetitive stimuli such as food. Punishment refers to both a procedure and a process—one in which the consequences are punishers.

reinforcement As in punishment, reinforcement refers to both a procedure and a behavioral process. The procedure involves providing reinforcers for responses; the process is the increase in the frequency of the behavior that results from the procedure. The consequences arranged in a reinforcement contingency are called reinforcers or rewards. In positive reinforcement, behavior produces appetitive stimuli such as food; in negative reinforcement, behavior eliminates or prevents aversive stimuli such as predators.

schedule Programs defining the necessary conditions for responses to produce consequences. Schedules arrange consequences as a function of time, number of responses, or both; they can be signaled by discriminative stimuli, and they can deliver appetitive or aversive stimuli.

Behavior interacts with the environment in a few important ways. Frequently, the environment supplies a stimulus that elicits a response; alternatively, the environment directs behavior by providing consequences for a response. The concepts of reinforcement and punishment deal with the effects of the latter; responses can be made more likely when they produce certain kinds of consequences and less likely when they produce other kinds. Sensitivity to the consequences of one's own behavior is among the most significant evolutionary innovations—a sort of ontogenetic natural selection to local aspects of ecology—extending the process of adaptation throughout the life of the individual.

Reinforcement and punishment entail both an environmental operation, arranging a consequence for a response, and a behavioral process, the effect on the frequency of the response. Like the reflex, the concepts are descriptive rather than explanatory—to observe an instance of reinforcement or punishment is not to explain it. However, the tautology implied in the definitions is easily refuted by findings showing that changes in the operation are always followed by related changes in the process.

Despite their apparent simplicity, processes of reinforcement and punishment are affected in complex ways by numerous variables and are capable of producing exquisitely coordinated behavior. The effect of a particular reinforcer depends on prior reinforcers, other sources of reinforcement, and many

other variables. Nor does reinforcement influence only the simplest varieties of behavior. Basic research on reinforcement is concerned with problems ranging from timing behavior in animals to concept learning in humans.

The challenge for the psychology of learning has been to understand the dynamic mechanisms of reward, the functions relating changes in the availability of reinforcers to changes in behavior. Knowledge of these functions will be essential to discover the physiological basis underlying the reinforcement process. Meanwhile, the pragmatic concerns of clinical psychologists and educators have benefited greatly from applications of basic findings helping them to develop effective behavior management programs for education, industry, and individuals with behavior problems.

I. HISTORY

The earliest concepts of reinforcement originated within experimental neurobiology and philosophy, but psychological investigation is largely responsible for unraveling its complexity.

In the middle of the 18th century, spinal reflex arcs had been isolated and used to support one-half of the Cartesian distinction between reflexive behavior and a spiritual mind. However, the latter half, Rene Descarte's assumption of native mental knowledge, was more difficult to endorse. Empiricist philosophers such as Locke proposed that complex knowledge could be explained by the accumulated experience of associated ideas. In the mid-19th century, Alexander Bain conceived the first modern idea of reinforcement by extending empiricism to action. He proposed that in addition to reflexes, organisms spontaneously emitted random, nonpurposive movements. When one of these actions produced pleasurable consequences, it would be repeated because of the association of a "muscle sense" with pleasure. This early conception of habit, behavior acquired by experience, gained acceptance in the late 19th century amidst the rising tide of Darwinism. Importantly, it provided a viable alternative to the creationist spirit implicit in Descarte's assumption that rational knowledge preceded complex action.

Approximately 40 years later, Edward Thorndike provided the first experimental demonstration that behavior could be affected by its consequences. One of his cat "puzzle boxes" is shown in Fig. 1. The top of Fig. 1 shows that the latency to escape from the box was long at first but gradually shortened. He called the behavior "trial-and-error learning" and formulated the law of effect, stating that pleasurable consequences "stamped in" responses and annoying consequences "stamped out" responses. Shortly thereafter, Pavlov's research on conditioned reflexes showed how reflexes could be conditioned to previously neutral stimuli. Thorndike and Pavlov had provided strategic concepts and tools for the new independent field of experimental psychology.

Pavlov's work initially overshadowed that of Thorndike because of arguments that Thorndike's finding could be explained by conditioned reflexes. However, in the 1930s, B. F. Skinner promoted the currently held view that Thorndike and Pavlov had discovered different processes. Skinner went on to design a highly successful experimental chamber to study reinforcement with rats and a cumulative recorder to record behavior in time (Fig. 2). His simplest design involved fitting a box with a lever, a food dispenser, and a light to signal the availability of food. He used his box to conduct numerous critical experiments on reinforcement, including several influential studies of discrimination learning and reinforcement schedules. Unlike puzzle boxes and mazes that studied learning in a trial-by-trial format, where the experimenter determined when a subject was placed in the apparatus, Skinner's automated apparatus permitted analysis of "free-operant behavior," where the animal could respond without an experimenter's intervention. Free-operant procedures are ideally suited to reveal the real-time dynamic interaction between behavior and environment, such as how quickly an organism adapts to changes in the availability of food. Skinner disdained statistical research designs comparing average performances of groups of animals and promoted single-subject designs in which very few subjects, exposed repeatedly to different levels of an independent variable, serve as their own controls. The latter approach seeks to reveal sources of variations in an individual's behavior rather than to deal with them simply as average performances subject to "error variance."

Researchers and theorists have since broadened the relevance of reinforcement to economics, language, problem solving, and other varieties of complex behavior. These advances have occurred simultaneously with the recognition that the effects of reinforcement are constrained by biological endowment: Ultimately, the organism constrains both the responses that are available to be modified by consequences and how they can be modified. Now the

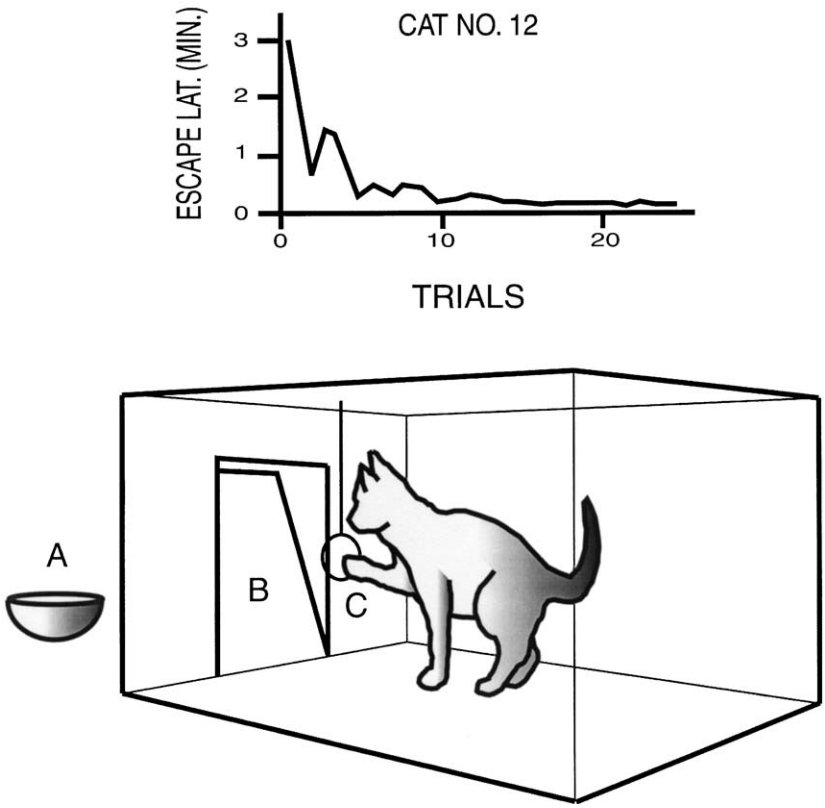


Figure 1 Thorndike's 1898 puzzle box. (A) Food placed outside of the box could be reached through a door (B) by pulling a loop (C) (Top) Time to escape as a function of trials.

challenge is to understand this interaction between the organism and its environment.

II. OPERANTS, REFLEXES, AND CONTINGENCIES

The reinforcement process can be illustrated by contrast with the reflex. In a reflex, a stimulus elicits a response and the response is said to be elicited. As shown in Fig. 3, reinforcement reverses the role between stimuli and responses an emitted behavioral response produces a reinforcing stimulus and the response is said to be operant. Operant and reflex responses are frequently linked by the same eliciting stimuli. For example, foods that elicit ingestion reflexes can vary in their form and distribution; these variations in foods can reinforce novel forms of operant handling and search trajectories. The effects of reinforcers and punishers do not require conscious deliberation, as illustrated by the nightly unconscious restraint of our movements that keep us from falling out of our beds.

Operants and reflexes are defined empirically by computing the conditional probabilities between stimuli and responses. In a reflex, response probability must be higher given a stimulus than given no stimulus. In an operant, response probability must be higher when responses produce reinforcers than when responses have no consequences or when reinforcers are delivered independently of responses.

A. Contiguity and Contingency

Reinforcers are most effective if they are contiguous with responses and less effective if they are delayed (Fig. 4). In animals, delays of just a few seconds are very detrimental. Even in the case of immediate reinforcement, however, reinforcers may not affect responding. There must also be a positive contingency, where the probability of a reinforcer is greater following a response. For example, rats will stop pressing a lever if the probability of reinforcement is the same with and without reinforcement, even if response-produced reinforcers are contiguous with

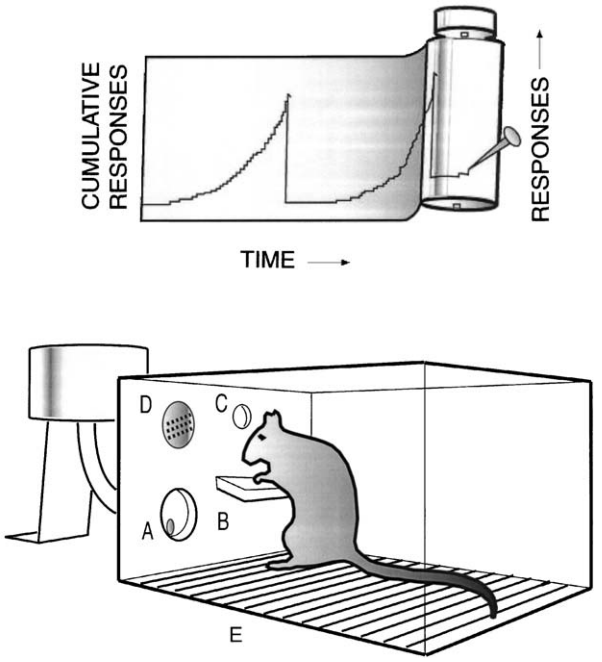


Figure 2 Skinner's 1938 electromechanical rat chamber. (A) Food is produced by pressing on a lever. (B) Stimuli signaling the availability of food or other events can be presented via a light (C) or a speaker (D). (E) The metal grid floor can be used to present electrical shock. (Top) A cumulative record: A clock motor feeds paper past a pen that increases with responses.

responses. Reinforcers that are delivered independently of responses diminish the effect of contiguous reinforcers, reducing their effect. Negative contingencies, sometimes called differential reinforcement of other behavior, are an effective way to reduce problem behavior in applied settings.

Responding can be maintained by slightly delayed reinforcers but only if there is a positive contingency.

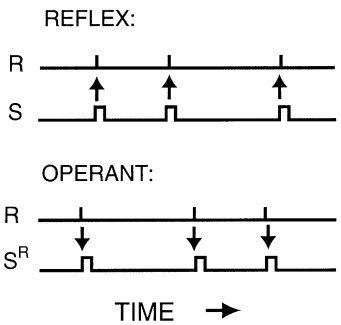


Figure 3 Causal relations between stimuli and responses in reflex and operant behavior. Arrows indicate dependencies between behavior and environment.

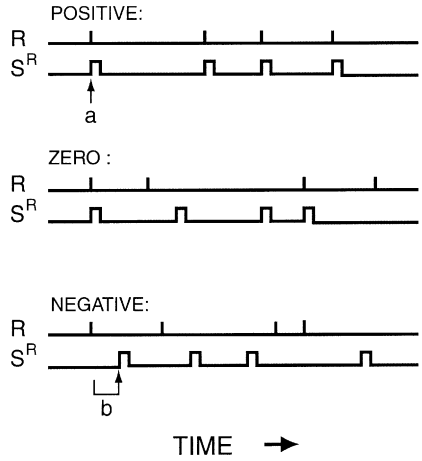


Figure 4 Contiguity and contingency in operant behavior. Contiguity is shown at a, when a response (R) is immediately followed by a reinforcer (S^R); a delay between a response and reinforcer is shown at b. Contingency is defined by (i) the probability of a reinforcer given a response and (ii) the probability of a reinforcer given no response. In a positive contingency, $1 > 2$; in a zero contingency, $1 = 2$; and in a negative contingency, $1 < 2$.

The detrimental effect of delay is dramatically reduced if the delay between responses and reinforcers is signaled by a stimulus that bridges the gap between a response and its delayed consequence. In the absence of signals, delayed reinforcers maintain more responding with variable delays than with fixed delays.

B. Extinction

Operant responses can be reduced in frequency by extinction (i) by terminating the reinforcement contingency or (ii) by delivering reinforcers independently of responses. Extinction procedures do not eliminate operants; rather, they produce additional learning that the operant contingency is broken. This can be seen in the rapid recovery of an operant when a reinforcement contingency is reinstated after extinction.

C. Conditioned Reinforcement

Reinforcers can be produced by procedures that generate conditioned reflexes, which are learned reflexes that result from a contingency between neutral stimuli and elicitors. In a conditioned reflex, a neutral stimulus becomes a conditioned stimulus and the response it elicits is a conditioned response. For example, if a light is paired with food, the light will elicit salivation as a conditioned stimulus but it can

also serve as a conditioned reinforcer. The effects of conditioned reinforcers change when the association between the conditioned stimulus and the reinforcer is changed.

D. Relativity of Reinforcement

Early theories of reinforcement argued that reinforcers served to reduce drives. Although some behavior could be explained by these motivation accounts, much behavior is not driven by need reduction. Many activities, such as play, seem to be maintained independently of their immediate consequences; moreover, a child might be induced to eat dinner in order to be able to play. Observations such as this led David Premack to suggest that if might be easier to understand the reinforcement process if reinforcers were identified by activities rather than stimuli. He proposed that relative time spent in different activities could be used to predict their value as reinforcers. For example, if a person spent more time in activity A than in activity B, the rate of B could be increased by reinforcing B with opportunities to do A. Figure 5 presents data from a human in a long-term observational study. During baseline conditions, this subject read more often than worked on art. When reading was restricted, and the opportunity to read was made contingent on art, the amount of time devoted to art increased. In control conditions in which reading was restricted without a contingency, the rate of art increased, but not to the levels seen with the contingency. The control demonstrated that the increase in artwork activity could not be explained only in terms of the restriction procedure; the increase required a contingency.

The fact that some infrequent activities such as sex can still be very reinforcing led to a modification of Premack's theory proposing that organisms work to maintain a balance between activities. This theory predicts that any activity can serve as a reinforcer if it is restricted: The individual will work to restore the balance. In the previous example, restricting the time available for B would cause B to be reinforced by A. This prediction has been experimentally supported.

III. VARIATION AND SELECTION

Reflexes are innate, stereotyped mechanisms selected by invariant phylogenic contingencies on survival. They serve prey capture, feeding, defense, escape, reproductive behavior, and care of offspring. In contrast, emitted responses contribute adaptive variations in behavior that assist in dealing with unpredictable features of the environment: locating food sources, nesting sites, mates, and other valuable resources.

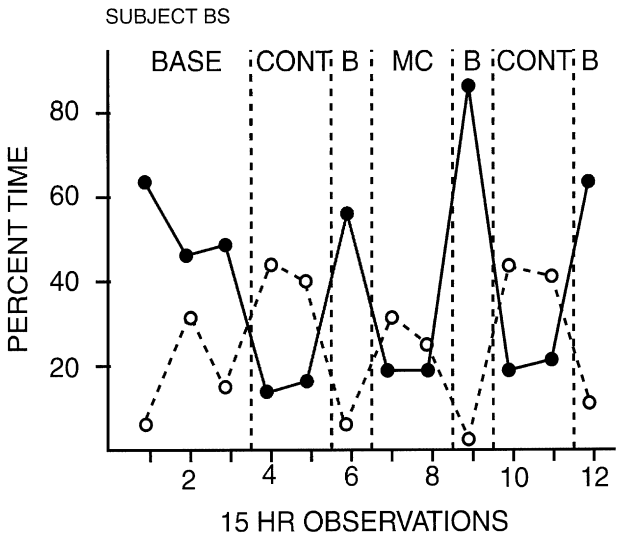


Figure 5 Premack contingencies. Open circles indicate time devoted to an instrumental response (artwork); solid circles indicate time devoted to a contingent response (reading). Vertical dashed lines indicate condition changes: baseline observations when activities were not restricted (BASE and B), contingency conditions when reading was restricted and access to reading depended on doing artwork (CONT), and matched control conditions in which reading was restricted (MC) (data from Bernstein and Ebbesen, 1978, Reinforcement and substitution in humans: A multiple response analysis. *J. Exp. Anal. Behav.* **30,** 243–253).

The selection of behavioral variations by reinforcement is illustrated in Fig. 6. In this puzzle box experiment, movements of a pole opened the box door. Camera records of the cats' behavior showed that a single response, emitted early in test trials, came to dominate. The response form differed between cats, suggesting that reinforcers were selecting random variations. Later experiments revealed that the escape responses these cats emitted, such as nuzzling the pole and rubbing the pole with their backside, were social responses elicited by the experimenters. The finding illustrates two important points about reinforcement. First the behavior a reinforcement contingency captures is constrained by the range of response variation provided by the organism in a particular situation. Second operant responses can have features that are reinforced but that are not essential to the contingency; for example, a rat might learn to operate a lever by pressing it with two paws even though one might suffice.

Reinforcement contingencies serve to narrow the range of variability in emitted behavior by selecting its most successful forms, making them more frequent and thereby displacing less effective forms. Juvenile hawks and wolves, for example, show adjustments in hunting strategies as a result of their effectiveness. The differential reinforcement of some responses to the exclusion of others is called shaping. In a dramatic example, invisibly small thumb twitches in college students were trained with a shaping procedure by connecting subjects to recording instruments and telling them that by relaxing they could earn money. When a particular thumb-twitch amplitude was emitted points were added on a counter in front of the subject, indicating money. The procedure quickly increased frequency of the target response. In debriefing it was found that students were never aware of how they produced money, experimentally demonstrating learning without awareness in humans.

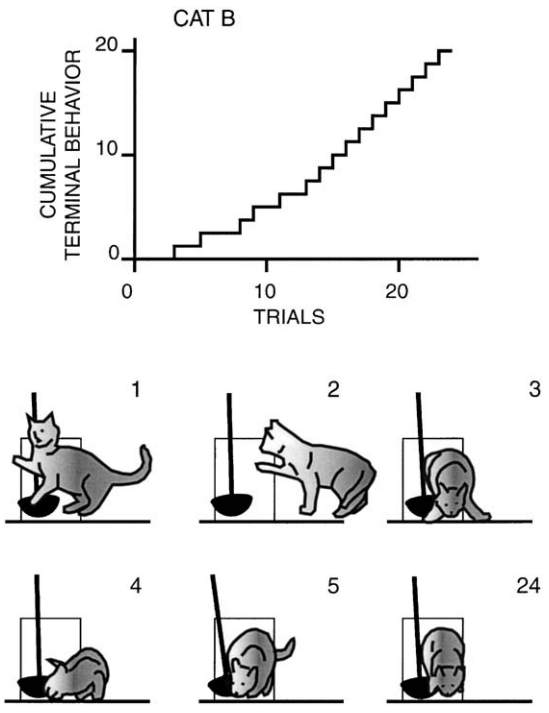


Figure 6 The stop-action principle of reinforcement. The cat's task involved escaping from a puzzle box by moving a pole. The cat's escape responses for trials 1–5 and 24 are shown below an inset showing the cumulative occurrences of the terminal response (pushing the pole with its hindquarters) (adapted from Guthrie and Horton, 1946, *Cats in a puzzle box*. Rinehart & Company. Copyright Edwin R. Guthrie and George P. Horton).

Shaping is closely analogous to natural selection, where different responses are analogous to phenotypic variations. Shaping illustrates the importance of variation in emitted behavior. Without variation, changes in reinforcement contingencies could not select new forms of behavior. This points out the trade-off in the variability reducing effects of reinforcement since less variation means diminished adaptability to future changes in contingencies. However, reinforcement never reduces variations in behavior entirely. For example, variability returns to emitted behavior in extinction, a potentially adaptive reaction to dwindling resources. Also, several studies have shown that variability in behavior can serve as the basis for reinforcement, as in the reinforcement of creative behavior (e.g., each reinforcer is contingent on the emission of a novel pattern of behavior).

A. Superstition in Animals

Skinner arranged response-independent food for a pigeon at 13-sec intervals and found that it developed stereotypical patterns of behavior. He called the behavior "superstition," arguing it was adventitiously reinforced by the accidental contiguous pairing of a response with a reinforcer and suggested that some human superstition could be explained in the same way. An important implication of the experiment was that contiguity was the necessary and sufficient condition for reinforcement. However, experiments have since shown that the behavior Skinner observed is not a simple reinforcement effect but instead complexly determined by the food and the experimental chamber. The behavior of a typical pigeon in such a

procedure is shown in Fig. 7: Some types of behavior reliably occur early in the interval (interim behavior) and other types occur later (terminal behavior). Also unlike the reinforced behavior shown in Fig. 6, the patterns are nearly identical in different pigeons. Evidence suggests that some responses such as pecking the wall are conditioned reflexes; others seem to be food-elicited foraging responses. The conclusion from this and other studies is that nonhuman animals are very sensitive to the difference between response-produced and response-independent reinforcers.

IV. ELICITING EFFECTS OF REINFORCERS AND PUNISHERS

The eliciting effects of reinforcers can reduce variability in behavior by restricting the range of responses available for reinforcement. The most dramatic examples arising from the animal-training literature have been misnamed "misbehavior." An example is illustrated in Fig. 8. The operant contingency involved reinforcing putting coins in a bank. Shortly after learning to do this, the raccoon began to misbehave, rubbing the coins together for long periods of time. However, the rubbing is actually a raccoon species-typical food washing behavior elicited by food. As a result of earning food by depositing coins, the coins became a conditioned stimulus for food and elicited washing.

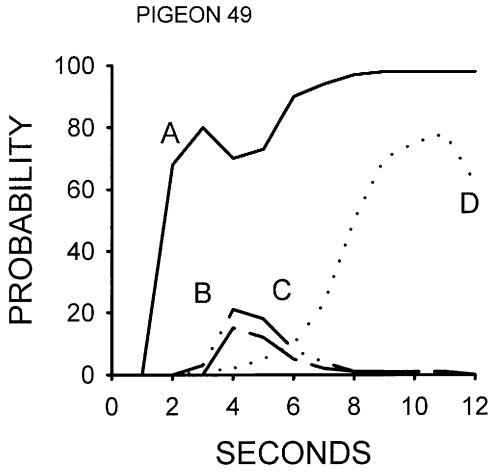


Figure 7 Probabilities of pigeon 49's behavior in successive seconds under a fixed-time 12-sec schedule. The letters indicate behaviors: A, facing the feeder wall; B, pecking the floor; C, quarter turns, and D, pecking at the feeder wall (data from Staddon and Simmelhag, 1971, The "superstition" experiment: A reexamination of its implications for the principles of adaptive behavior. *Psych. Rev.* **78,** 3–43).

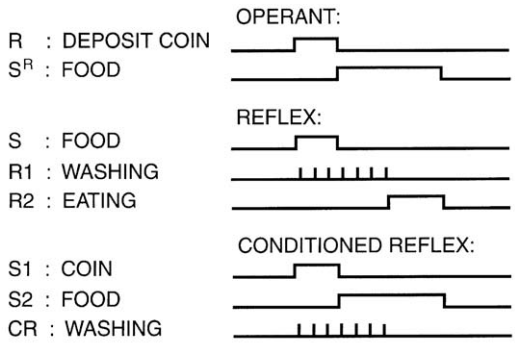


Figure 8 "Misbehavior" in the raccoon. The operant contingency reinforced depositing money in the bank. However, the additional contingency between coins and food resulted in the coins becoming conditioned elicitors of washing, interfering with the operant performance.

Many instances of punishment reduce the frequency of behavior in a similar manner: An emitted response produces a painful stimulus that elicits defensive behavior such as escape and avoidance. The result is that "punishment" reduces the target behavior but not just by being a consequence of behavior. As a rule, reinforcers are most effective when the responses they elicit are identical to or compatible with a reinforced response and punishers are most effective when the responses they elicit are incompatible with the punished response.

Conditioned aversive stimuli, stimuli that are paired with pain-eliciting stimuli, can suppress operant behavior. They can do so by being made contingent on operant responses or by their mere presence because they elicit defensive behavior that is incompatible with the operant. Figure 9 illustrates the interaction between conditioned-aversive stimuli and food-reinforced responding. As the tone + shock trials proceed, less operant responding is seen during the

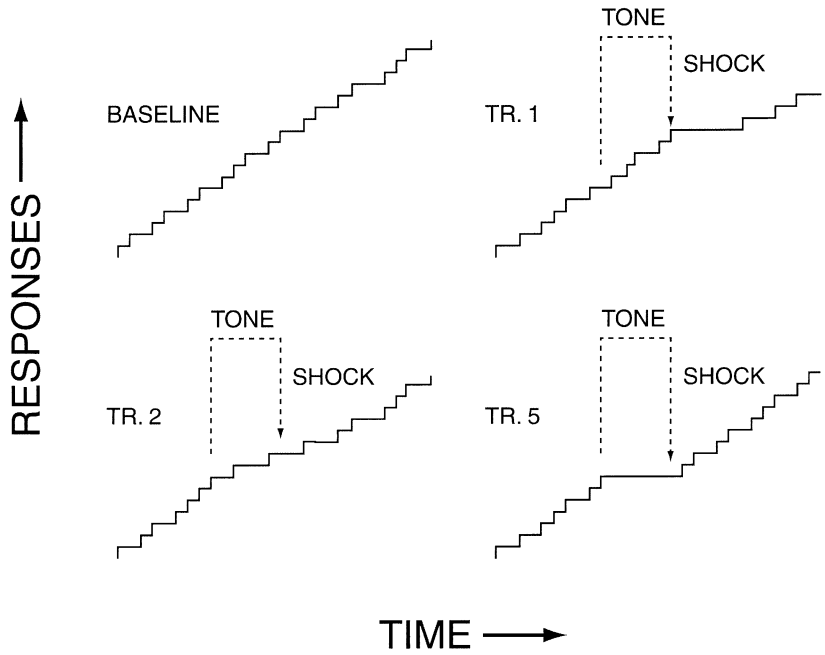


Figure 9 The development of a conditioned-aversive response as indicated by disruption of ongoing operant behavior.

tone. The amount of suppression generated by a conditioned-aversive stimulus depends greatly on the level of deprivation maintaining the operant behavior —a very hungry rat will show less suppression of food-reinforced lever pressing than a less hungry rat.

V. PUNISHMENT

Punishment contingencies are in opposition to reinforcement; they reduce the frequency of responses. Many of the variables that determine the effects of reinforcement similarly affect punishment. These variables are the following: (i) Immediate punishers are more effective than delayed punishers, (ii) strong punishers are more effective than weak punishers, and (iii) punishers are most effective that are presented only for target responses and not at other times.

The concept of punishment includes two operations: positive punishment, in which behavior produces aversive events such as hitting, and negative punishment, in which behavior causes the loss of a currently available appetitive stimulus such as food. However, these need not imply the same behavioral process even though they are both called punishment. For example, the former operation involves the presentation of eliciting stimuli with responses that habituate over repeated presentations, making punishers less effective over time.

Positive punishment has two effects that account for the reduction in behavior: (i) elicitation of species-typical escape or defensive behavior and (ii) the effect of the punishment contingency. These can be distinguished by comparing the effects of response-independent with response-contingent punishment. For example, a rat's lever pressing for food will be reduced by simply presenting occasional response-independent shocks, but it will be reduced more if shocks are contingent on lever presses. This finding shows that not all the effects of punishment are due to the elicitation of escape or defensive behavior; some are due to the punishment contingency.

Punishment can be seen as the opposite of reinforcement, but if differs from reinforcement in another important way. Punishment operations are always superimposed on a baseline of existing responding, usually maintained by reinforcement. In the laboratory, for example, punishment is usually studied on responses maintained by food. The result is that the effect of a given punishment is determined largely by the target response's level of motivation.

In extinction from punishment, responses recover to previous levels. Although the effects of punishment can be long-lasting, there are only a few experimental

examples in which a punishment eliminates responding permanently.

VI. SCHEDULES OF REINFORCEMENT

The same reinforcer can have very different effects on behavior depending on the reinforcement schedule that arranges it. Very few naturally occurring reinforcement contingencies arrange reinforcement for every response. Instead, the rule is partial reinforcement, in which responses are reinforced on the basis of time, number of responses, or some combination of time and responses. Figure 10 shows typical patterns of behavior on four simple schedules—the where, when, and how much of operant behavior is determined by schedules. These patterns and their dynamics are the subject of much quantitative modeling in search of a simple set of principles that will generate the gamut of schedule-typical performances. An adequate model must encompass complex behavior such as choice, self-control, and the temporal patterning of behavior.

Skinner pioneered the study of schedules, documented many of their effects, and developed a highly successful graphical form to display changes in behavior over time, the cumulative record. The discovery of the orderly and distinctive patterns of behavior on different schedules of reinforcement led Skinner to argue that rate of responding is the primary datum of the science of behavior. However, attention to the molar property of rate has had the unfortunate effect of obscuring investigation of the moment-to-moment molecular dynamics of behavior.

A. Interval versus Ratio Contingencies

In an interval schedule, a reinforcer is produced by the first response after the passage of a specified interval of time. In a ratio schedule, a reinforcer is produced after the emission of a specified number of responses. Time or response requirements can either be fixed or variable around some mean value (of time or responses). Figure 11 shows typical data from a pigeon responding on variable interval (VI) and variable ratio (VR) schedules. When rate of reinforcement in interval and ratio schedules is equated, organisms respond at higher rates on ratio schedules (although the difference is

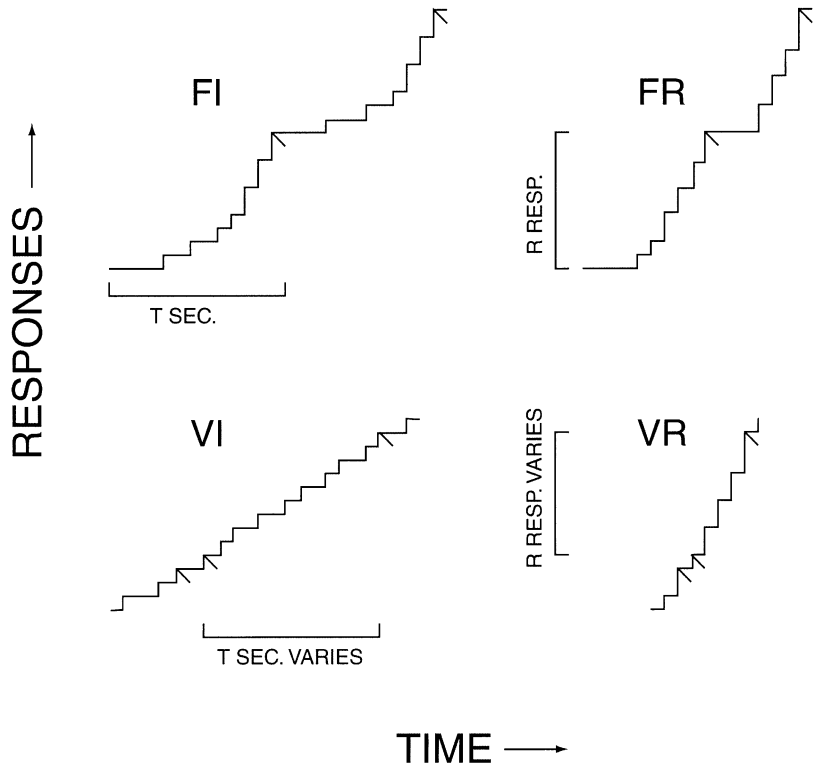


Figure 10 Schedule-typical response patterns on four simple schedules: fixed interval (FI), fixed ratio (FR), variable interval (VI), and variable ratio (VR). Ratio schedules produce higher rates of responding than interval schedules at a given rate of reinforcement, and fixed schedules show more pausing after reinforcement than variable schedules.

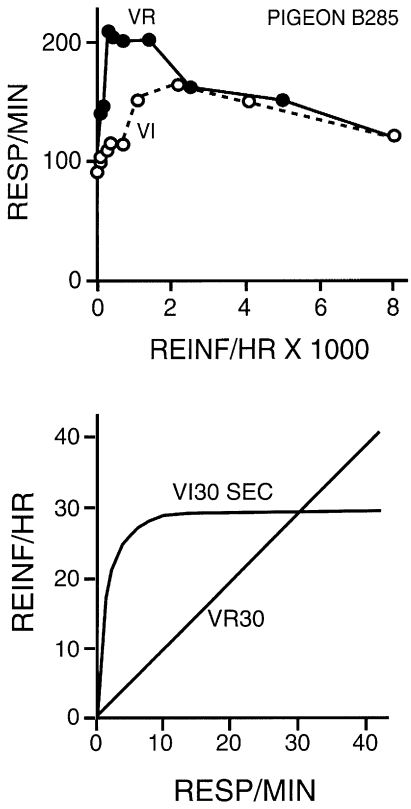


Figure 11 Interval and ratio schedule feedback contingencies. (Top) Response rates of a pigeon on variable interval (VI) and variable ratio (VR) schedules as a function of reinforcement rate.

small at very high rates of reinforcement). The reason is illustrated by the feedback functions in the bottom of Fig. 11: Rate of reinforcement in ratio schedules increases directly as a function of response rate but not in interval schedules. Fixed interval (FI) and fixed ratio (FR) schedules show similar rate functions.

Ratio and interval schedules produce distinctive declines of responding in extinction. In interval schedules, responding declines gradually; in ratio schedules, responding alternates between periods of high rates and no responding.

B. Waiting: The Timing of Operant Behavior

Organisms show characteristic pausing before responding, or waiting, after earning a reinforcer. Figure 12 shows pause distributions for simple interval and ratio schedules. Pausing is greatest on FI and FR schedules and minimal on variable interval VI and VR schedules. In VI and VR schedules, wait times are heavily influenced by the shortest interreinforcement intervals or ratios in a schedule: Organisms wait somewhat less than the shortest interfood interval.

Figure 13 shows characteristics of waiting on FI schedules in a typical pigeon, patterns that are seen in a broad variety of species, including monkeys, fish, and humans. The top plot in Fig. 13 shows that in an FI, wait time is reliably a fixed proportion of the FI value. Proportional timing is sometimes complemented by scalar variability in wait times, where variance in waiting is a constant proportion of interval value. The bottom plot in Fig. 13 shows that variability in this pigeon's wait time increases with FI duration. The proportional and scalar properties of waiting are so ubiquitous that they have led to several timing models.

Waiting is obligatory: Organisms will pause after earning a reinforcer even when doing so reduces the immediacy and overall rate of reinforcement. This is illustrated in experiments with response-initiated delay (RID) schedules. The contingency is shown in the bottom of Fig. 14: After a reinforcer, a response initiates a delay of T seconds that ends in reinforcement. Instead of responding immediately after reinforcement (an optimal strategy), pigeons treat the RID like an FI with a duration equal to the wait plus T. Moreover, waiting on the current trial is a function of the preceding trial, a linear-waiting rule. A linear-waiting rule has been demonstrated by adding an adjusting feature to the RID such that the value of T depends on the duration of the preceding interfood interval, i.e.,

$$T(N+1) = \alpha(t(N) + T(N)) \qquad (1)$$

where t is the time until the first peck (i.e., pause or wait time), α is a proportionality constant, and N is the time between food deliveries. If the animal behaves in such a way that pause in interval $N+1$ is simply proportional to the preceding interreinforcement time, this process is unstable: t should decrease to 0 if α is less than 1, but increase without limit if α is greater than 1. The top of Fig. 14 shows data that illustrate this instability.

C. Differential Reinforcement of Rate

Reinforcement can be arranged to follow a pause followed by a response, an interresponse time (IRT). If the IRT must be less than a maximum value, response rate is increased; if the IRT must be greater than a minimum value, response rate is decreased. Differential reinforcement of IRTs has been proposed to

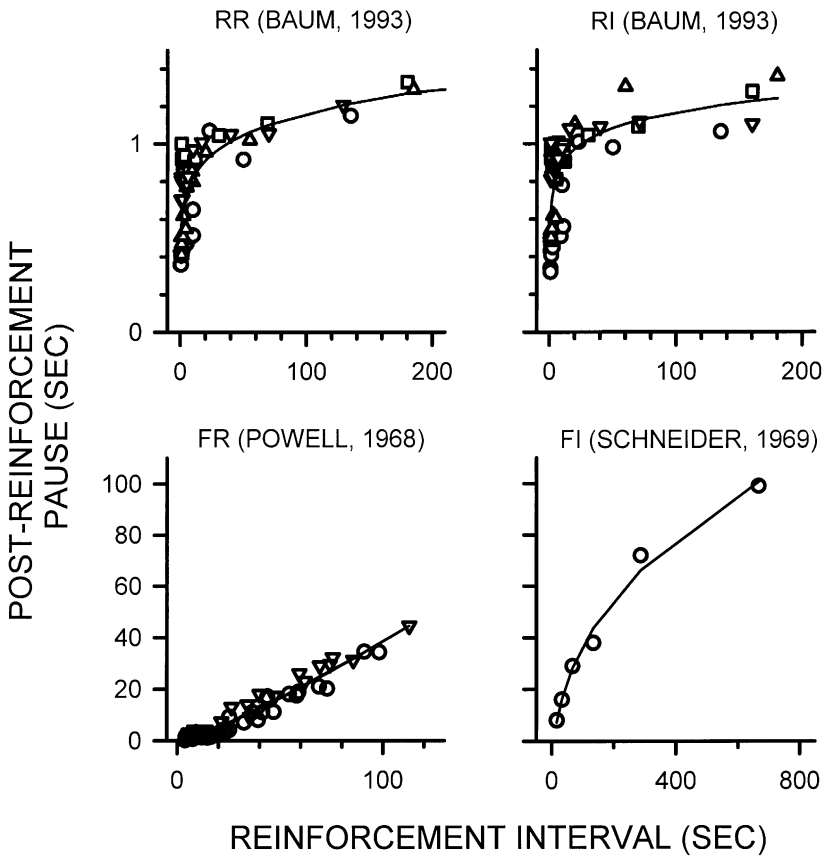


Figure 12 Pause distributions on the four basic schedules. FI data points are means; other plots show individual organisms. [Data from: Baum, W. M. (1993). Performance on ratio and interval schedules of reinforcement: Data and theory. *J. Exper. Anal. Behav.* **59,** 245–264; Powell, R. W. (1968). The effect of small sequential changes in fixed-ratio size upon the post-reinforcement pause. *J. Exp. Anal. Behav.* **11,** 589–593; Schneider, B. (1969). A two-state analysis of fixed-interval responding in the pigeon. *J. Exp. Anal. Behav.* **12,** 677–687].

explain rate differences in interval and ratio schedule behavior since the probability of reinforcement in interval schedules is greater for responses following long IRTs, and ratio schedules reinforce short IRTs. However, behavior produced by IRT schedules is unlike that seen under interval and ratio schedules. For example, it is difficult to differentially reinforce wait times longer than about 20 sec, even though waiting on FI schedules can extend to many minutes. This is because the proportional aspect of timing, pausing for a proportion of a reinforcement interval, preempts the wait time required for reinforcement.

D. Behavior Chains, Sequences, and Multiple Schedules

Much complex behavior, such as building a house, is made up of elementary tasks that must be executed in a fixed order to obtain a reinforcer. This is the essence of a behavior chain: One contingency is made a consequence of another until a terminal reinforcer is procured, and each contingency is signaled by a unique stimulus. The bottom diagram in Fig. 15 shows a three-link chain—cumulative records from tandem and chain schedules made up of three FI 60-sec schedules. Original accounts of how behavior in early elements of a chain was maintained appealed to conditioned reinforcement. However, the data shown at the top of Fig. 14 clearly show that response rates in the early links of the tandem schedule (a chain in which the links are not signaled by distinctive stimuli) are higher than in a chain schedule. Although conditioned reinforcement effects are well documented, they do not seem to apply to behavior chains. A completely satisfactory explanation has not yet been obtained, but there is a similarity between chain and RID schedules that can explain a large part of the effect: Any tendency of the

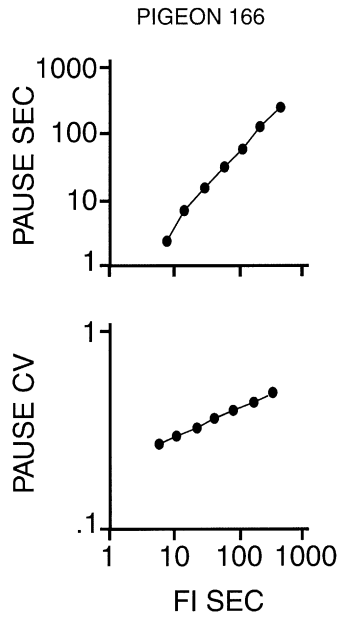


Figure 13 Properties of pausing on FI schedules. (Top) Pausing. (Bottom The coefficient of variation (standard deviation of wait times divided by their mean: CV). Data are on log–log coordinates

animal to wait in the first link of a chain in proportion to the delay to the terminal reinforcer would produce behavior like that seen in Fig. 15.

E. Multiple Schedules

The rate of responding on a schedule is affected by the existence of alternating schedules. This is seen in multiple schedules in which a schedule identified by stimulus A alternates with another schedule identified by stimulus B. For example, if a multiple VI–VI schedule is changed to a multiple VI–extinction, the rate in the unchanged VI will increase. This effect is called behavioral contrast. Several variables appear to contribute to contrast; one important reason is simply that the rate of one operant is suppressed by the availability of alternative sources of reinforcement for other operants.

F. Choice

Reinforcements can occur simultaneously for different responses. In concurrent schedules, choices between alternatives are a function of the types of competing schedules. Studies of choice have found that the behavior is complexly determined and not just a simple function of competing rates or magnitudes of reinforcement.

Humans and other organisms faced with two VI schedules will choose in proportion to the relative rates of reinforcement they provide. Figure 16 presents data from an experiment with pigeons in which the

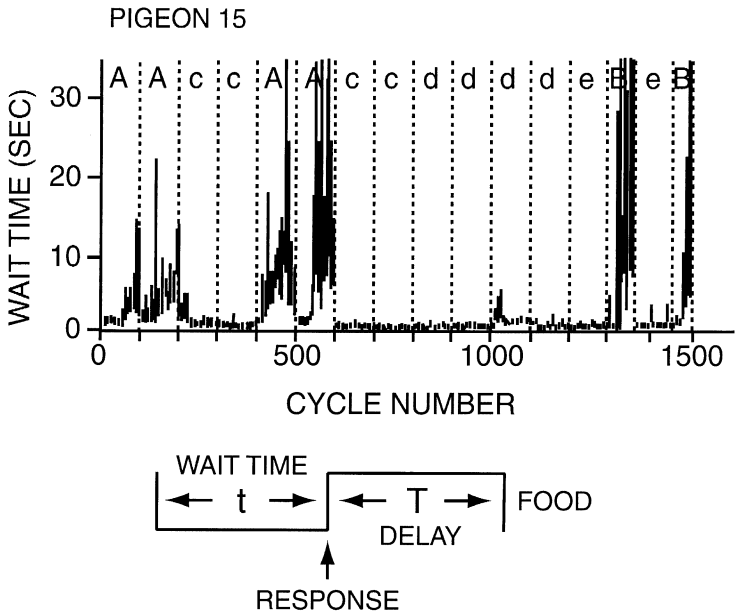


Figure 14 Dynamics of waiting on RID schedules. Vertical dashed lines show sessions during different values of the gain parameter α: A = 7, B = 4, c = 0.5, and d = 0.25 (data from Wynn and Staddon, 1988, Typical delay determines waiting time on periodic schedules: Static and dynamic tests. *J. Exp. Anal. Behav.* **50,** 197–210).

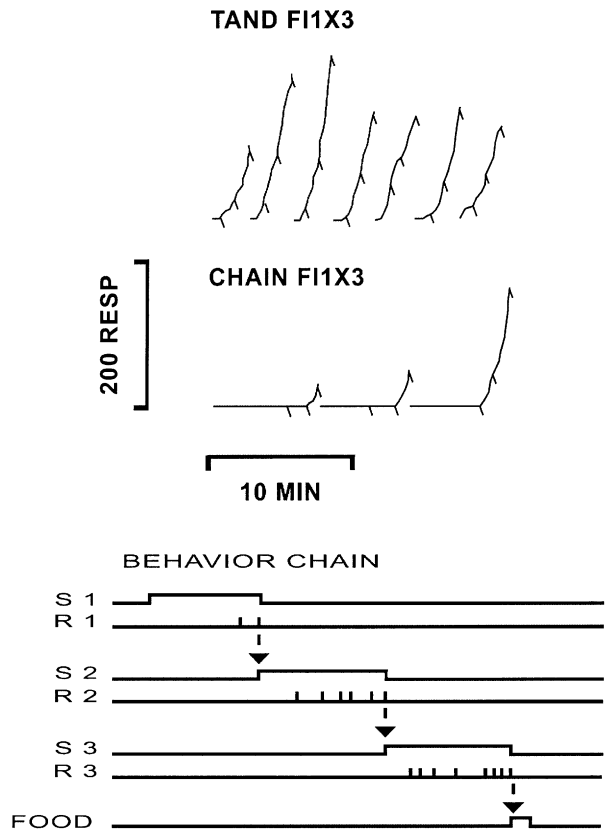


Figure 15 Chain and tandem schedules. In both schedules, reinforcement is obtained after completing a succession of three FI 60-sec schedules. In the chain, the stimuli identifying each FI (S1–S3) are distinctive; in the tandem, they are identical. (Top) Cumulative records from one pigeon (downward pips of the response pen indicate the last response in each FI; the response pen was reset after the third link of each schedule) (data from Kelleher and Fry, 1962, Stimulus functions in chained fixed-interval schedules. *J. Exp. Anal. Behav.* **5,** 167–173).

proportion of reinforcement available on left and right VI schedules was varied over sessions. Data points along the diagonal of the figure indicate that the pigeons' choices tracked reinforcement; for example, they selected the alternative producing 20% of the total reinforcers about 20% of the time. Richard Herrnstein formalized this observation in the matching principle, the first quantitative formulation of choice:

$$\frac{R_L}{(R_L + R_R)} = \frac{r_L}{(r_L + r_R)} \qquad (2)$$

where R_L and R_R are response rates on left and right alternatives and r_L and r_R are reinforcement rates on left and right alternatives. The matching principle has been demonstrated in a variety of species, tasks, and circumstances, including in conversation between a speaker and two listeners who reinforced verbalizations at different rates.

Matching assumes that behavior is controlled by the molar properties of rate of reinforcement. Molecular models, attempting to account for choice in terms of the moment-to-moment tracking of reinforcement probabilities for different alternatives, have not been supported by experimental tests. However, the effects of reinforcers can be seen at the molecular level. In simple FI schedules, for example, shortening a single interval will produce a shorter waiting time in the next interval (i.e., linear waiting). Models that predict such dynamic properties of behavior have only just appeared and have not yet been applied to choice behavior.

G. Choice and Conditioned Reinforcement

Despite questions about whether stimuli in behavior chains are conditioned reinforcers, many experiments have studied choice in concurrent chain schedules. The procedure is illustrated in Fig. 17. In a typical concurrent chain schedule with pigeons, two white response disks are presented. Pecking on each disk is reinforced according to a VI schedule, the same for each key. This is the first link of a two-link chain. When the schedule on a key "times out," the next peck causes a change in both keys: For example the pecked key changes to red (the second link on the left), and the

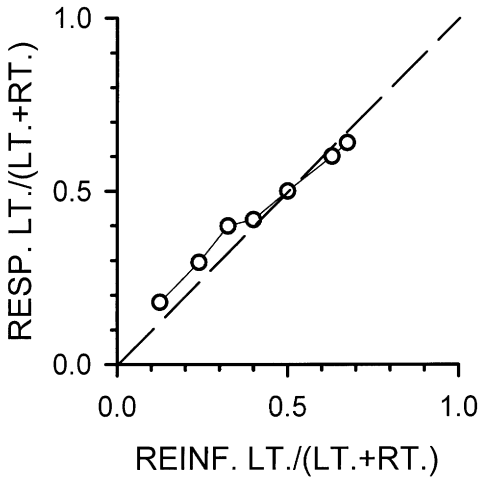


Figure 16 Choice on concurrent VI–VI schedules (data from Catania, 1962, Concurrent performances: Reinforcement interaction and response independence. *J. Exp. Anal. Behav.* **6,** 253–263).

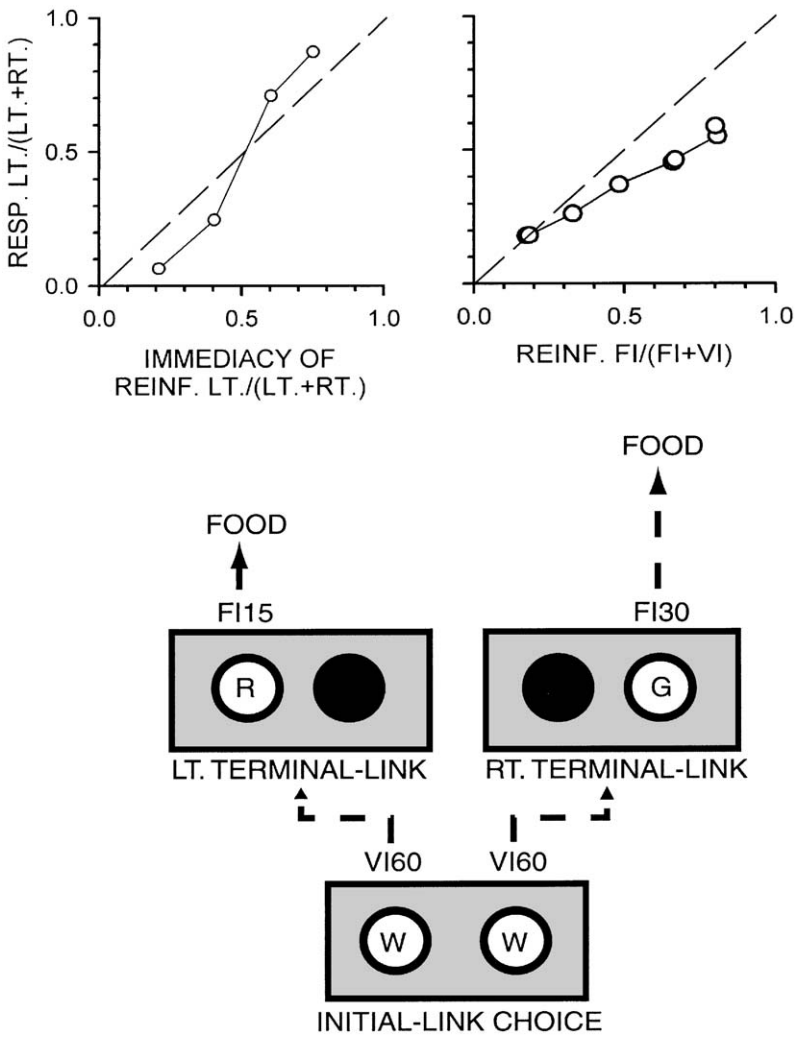


Figure 17 Concurrent chains schedule. (Top) Data from pigeons with FI schedules in second links on the left (data from Killeen, 1970, Preference for fixed-interval schedules of reinforcement. *J. Exp. Anal. Behav.* **14,** 127–131) and with VI and FI second links on the right (data from Trevett *et al.*, 1972, Performance in concurrent interval schedules. *J. Exp. Anal. Behav.* **17,** 369–374).

other is blacked out. In the presence of the red key, food reinforcement is delivered according to another schedule, (e.g., FI, VI, VR, or FR). A similar sequence operates for the other second-link stimulus (green, on the right).

The usual dependent measure in these experiments is steady-state relative response rate, $x_L/(x_L+x_R)$, in the first link. Typical independent measures in the second links are absolute and relative reinforcement rates (R_{1L} and R_{1R}) and the associated average delays to reinforcement ($1/R_{1L}$ and $1/R_{1R}$). In the second links, in addition to absolute reinforcement rates, studies have varied the number of links with or without distinctive stimuli, size, and delay (VI or FI value) of food reinforcement. The following are reliable results from these experiments:

1. First-link preference tends to indifference as the absolute value of the first-link VI is increased: The longer the delay to the second link, The more indifferent the animal is to which link is chosen.
2. A single FI second link is preferred to an equal-rate two-link FI chain.
3. A short FI second link is preferred over a long FI second link more than predicted by the matching principle.
4. A VI second link is preferred to an equal-reinforcement-rate FI.

One of the important generalizations that can be made from these findings is that organisms value immediate reinforcers more than delayed reinforcers more than predicted by rate of reinforcement, even in the case in which the immediate reinforcers are infrequent (3 and 4). Several quantitative theories have been proposed to account for these and other results from concurrent chain schedules.

Delay-reduction theory (DRI) proposes that the value of a conditioned reinforcer, $V(S_i)$, is directly related to the difference to delay to reinforcement signaled by the current stimulus and the succeeding stimulus (i.e., to the first derivative of expected reinforcement rate). In its simplest form, DRT is $V(S_1) = k(t_1 - t_2)$, where t_1 is average time to reinforcement in the first stimulus, t_2 is average time to reinforcement in the succeeding stimulus, and k is a constant. Preference (relative response rate) is determined by relative V values:

$$\frac{x_{L1}}{x_{L1} + x_{R1}} = \frac{\Delta_L}{\Delta_L + \Delta_R} \tag{3}$$

where Δ_L is $d_{L1} - d_{L2}$, the delay reduction on the left, and similarly for Δ_R. A later version of DRT, modified to incorporate reinforcement rate, describes a number of experimental findings in the concurrent chains literature but it has difficulty if the initial links are not VI schedules and difficulty with terminal links made up of complex schedules, such as two-link chains or multiple schedules. DRT can account for properties 1 and 2 but has difficulty with 3 and 4.

Incentive models of conditioned reinforcers share the assumption of a nonlinear relation between stimulus value and reinforcement rate. For a simple two-link concurrent chain, incentive theory (IT) reduces to the relation

$$x_{L1} = \frac{1}{d_{L1} + d_{L2}} \left(e^{-qt_{L2}} + \frac{1}{d_{L2}} \right), \tag{4}$$

and similarly for the other choice. This equation, with decay (q) and a bias variable as free parameters, could fit data from concurrent chain experiments with two-link second links (which pose problems for DRT) as well as a number of other studies.

Melioration theory differs from DRT and IT in that it makes predictions directly about the value of a stimulus (i.e., without concern about the overall rate of reinforcement). Melioration theory assumes that conditioned stimulus value is a function of the value and rate of the primary reinforcer identified with the stimulus, and that there is a hyperbolic growth function relating rate to value. This model can make similar predictions to DRT but fails to handle complex second-link concurrent chains (e.g., a behavior chain).

None of these theories account for all the data on concurrent chains. Recently, researchers have turned to economic models of choice. These models treat responses per reinforcer as cost and incorporate factors such as elasticity and demand in their predictions. Economic models work well with simple procedures involving cost and demand, but organisms deviate substantially from predictions of performance on complex schedules. Moreover, economic models have demonstrated that animals do not make optimal choices (or, as suggested by RID schedules, show optimal behavior).

The most successful quantitative accounts of reinforcement, such as variants of the matching law, involve molar averages of behavior and consequences. This has led to speculation that reinforcement mechanisms operate on molar principles. Despite the success of these models, molecular effects of reinforcement are abundantly obvious. Recent models of reinforcement suggest that there are two processes that determine the effect of a reinforcer, one "slow" and one "fast." The slow process accumulates the effect of reinforcers over long periods of time, reducing fast changes in behavior that occur when contingencies change.

VII. DISCRIMINATION AND EXTINCTION

In early experiments on the law of effect, Thorndike recognized that the effect of reinforcers was to strengthen behavior in a particular situation. In formal terms, this is the three-term contingency relation of discriminative stimulus (S^d), response, and consequence. The discrimination procedure arranges for responses to have consequences in the presence of an S^d and not at other times; the S^d is said to occasion responses. Discrimination learning underlies much behavior said to involve cognition, including important aspects of concept learning, verbal behavior and problem solving.

A. Generalization, Concepts, and the Quantal Nature of Discrimination

Once responses are reinforced in the presence of a stimulus, generalized responding will be occasioned by

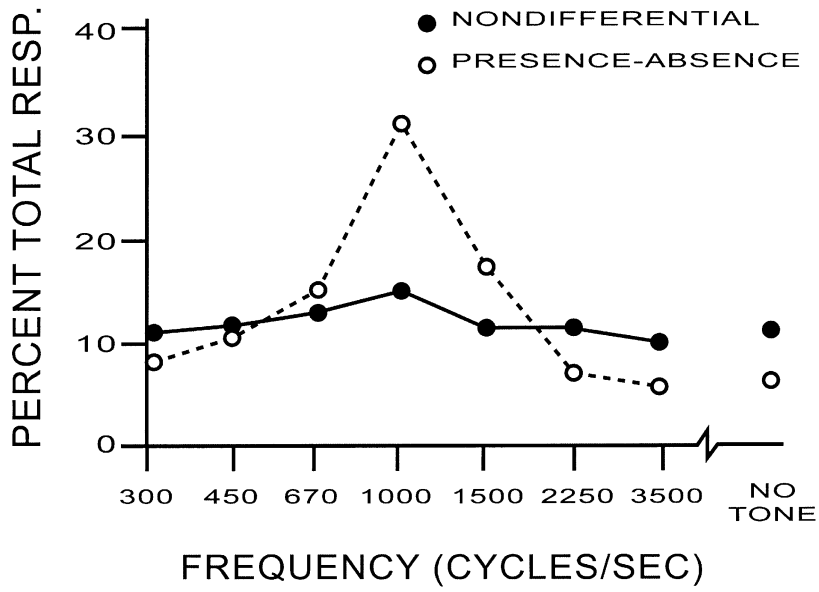


Figure 18 Generalization gradients from pigeons after nondifferential training with a 1000-Hz tone and presence–absence training with and without the tone (data from Jenkins and Harrison, 1960, Effect of discrimination training on stimulus generalization. *J. Exper. Psych.* **59,** 321–334).

similar stimuli without additional reinforcement. With tone and color S^d, generalization shows a peak around the S^d that is called a generalization gradient: The probability of responding decreases as a function of the difference with the training stimulus. Figure 18 shows that discriminative stimulus control is tighter (the gradient is peaked) following training in which reinforcement in the presence of the S^d was alternated with periods of extinction without the S^d (presence–absence training) in comparison to training without extinction periods (nondifferential training).

Figure 19 shows the effect of training with an S^d and extinction stimulus that lie on the same stimulus dimension of light (wavelength). In this case, the postdiscrimination gradient is steeper and the peak is displaced away from the training frequency. A satisfactory explanation for the shift has not been determined but it is possible that the extinction stimulus produces an inhibitory gradient that subtracts from the excitatory generalization gradient.

Discriminative stimuli can be very complex. Natural concepts, such as trees, water, and flowers, can form the basis of discriminations. In a concept, perceptually similar stimuli occasion a response such as saying "tree, "water," or "flower." Concepts are trained by reinforcing a response in the presence of concept exemplars and extinguishing the response in the presence of other stimuli. Demonstrating a concept requires testing with novel stimuli and finding generalization to untrained examples of the concept but not to other stimuli. Research investigating the basis of concept learning in animals has found that some concepts such as cars require training with several examples, but others, such as an oak leaf, can be learned after a single exemplar. The findings suggest that some concepts involve learning about correlation of stimulus features and other concepts contain a single perceptually distinct feature.

Generalization is an adaptive process since discriminative stimuli in the natural environment are likely to vary from instance to instance. However, if discriminative stimuli vary, reinforcement ensures that responses do not. This has been demonstrated in tasks in which both responses and stimuli are continuous. In one study, pigeons were trained to peck at one end of a 25-cm strip when the strip was brightly illuminated and to peck at the other end when it was dimly illuminated. In tests with intermediate levels of illumination, the pigeons pecked only at the end points where responses had previously been reinforced. Moreover, the end they pecked depended on the similarity of the test stimulus to the training stimulus. This property of behavior is seen in naming. For example, a child who has learned the concepts of "dog" and "horse" will not say "dorse" in the presence of a llama (assuming that a llama has intermediate dimensions); instead, the child

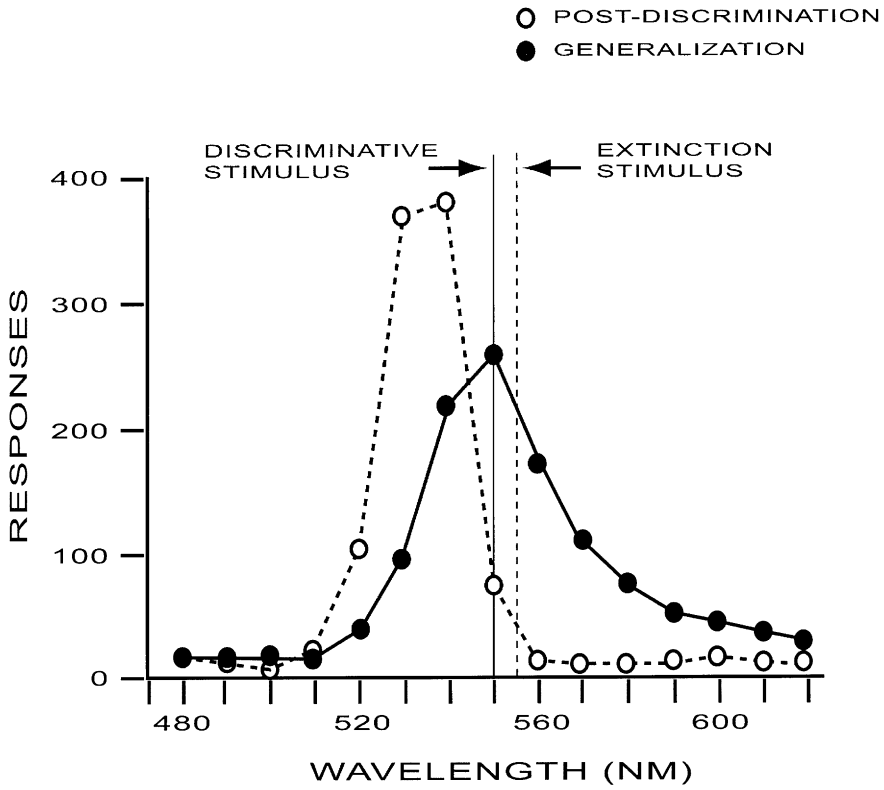


Figure 19 Postdiscrimination and discrimination gradients in pigeons for wavelength of light (data from Hanson, 1959, Effects of discrimination training on stimulus generalization. *J. Exper. Psych.* **59,** 246–253).

will emit one or the other of the previously learned responses. The invariance of discriminated responses has been called the quantal nature of discrimination to emphasize the fact that response forms remain discrete.

B. Learning Set and Learning by Exclusion

Discrimination learning can be generalized. The point is illustrated by experiments with monkeys on learning set. In the typical procedure, a monkey is presented with two objects; under one is a piece of food. When it learns the discrimination, two new objects are presented, and so on. Learning proceeds more rapidly after each problem until it learns new discriminations in a single trial. Learning set has also been demonstrated in other species and with other tasks.

Learning set is evident in a word-learning discrimination called learning by exclusion. In typical experiments employing nonverbal humans with mental retardation, subjects are first taught names for several foods. Then a new food is placed among foods with known names. Without any further training or prompts, subjects are asked to point to the new food that is given a never before heard name. Experienced subjects correctly select the novel food on the first test, and in subsequent tests with that food and new foods they demonstrate retention of the new name. These experiments show that after learning several names, individuals learn the contingency that exists between new names and new objects.

C. Conditional Discrimination

Reinforcement can be made contingent on conditional relationships between stimuli. Figure 20 shows a pigeon in an identity matching task in which stimuli are matched on the basis of similarity: In the example, selecting the "square" comparison is conditional on the square sample. Conditional discrimination is the basis of symbolic behavior. For example, in the presence of a red triangle, we say "red" when prompted with "color?" and "triangle" when prompted with "shape?"

It is in this domain of conditional discriminations that clear differences between humans and other

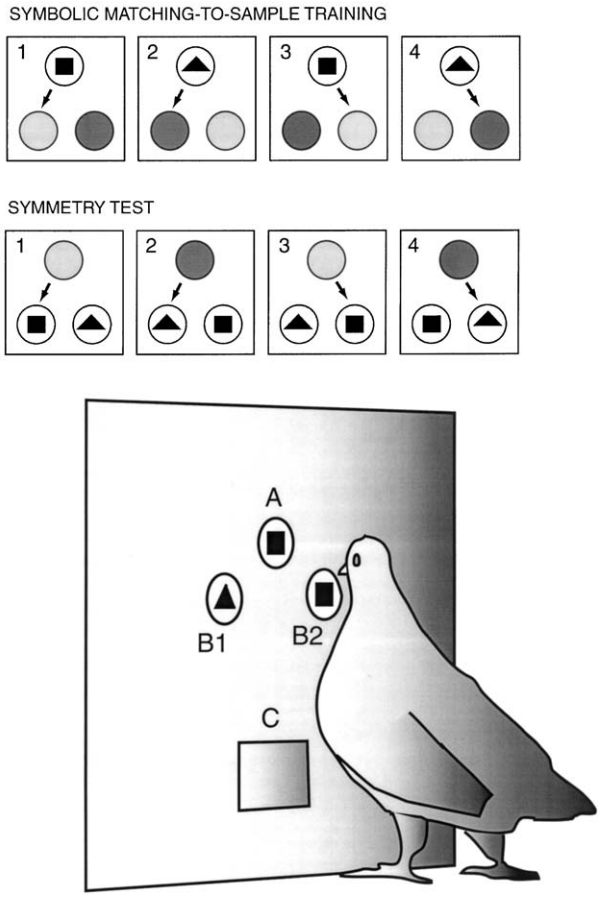


Figure 20 Matching-to-sample procedure. The pigeon is shown matching the comparison stimulus, B2, to the sample stimulus, A, on the basis of identity. (Top) Symbolic matching training trials between sample shapes and comparison hues and a symmetry test in which the hues are samples and the shapes are comparisons. Arrows indicate correct selections.

species arise. Animals and humans can learn conditional discriminations, such as "same" and "different," as well as symbolic matching between arbitrarily related stimuli, such as words and their referents. However, only humans can demonstrate some kinds of emergent conditional discriminations. For example, once humans learn the symbolic stimulus relation, "if stimulus A then B", they can respond correctly to a symmetry test, "if B then A", without further training (the simplest training and test relations necessary to evaluate symmetry are shown at the top of Fig. 20). The variables that explain symmetry are not yet understood; the ability may be a uniquely human accomplishment or emerge from experience with language. Interestingly, both humans and other species show transitivity: training the relations, "if A then B" and "if B then C"; finally, test "if A then C."

D. Recombination of Discriminations

Novel behavior results when discriminative stimuli controlling different responses or response dimensions are combined in novel ways. Recombinant generalization underlies some instances of problem solving, instruction following, and productivity in language. Figure 21 illustrates an example taken from the literature on language training for children with mental retardation. In this study, the child was first taught to push the 12 objects when given the instructions "push the glass," "push the scissors," and so on for the remaining nouns. Then, the next verb, "drop," was taught in the same way. As more verbs were taught, the child soon generalized new verbs to the objects after training with only one noun. Recombination processes have been demonstrated in a variety of species and in other contexts, including productivity and problem solving (e.g., cases in which a solution requires testing different combinations of elements).

E. Fading, Attention, and Blocking

Difficult discriminations, such as subtle differences between hues, can be established by initiating training with very different stimuli and progressively making them more similar. This procedure, called fading, is analogous to shaping. Fading can establish a difficult discrimination in fewer trials than is possible by prolonged training with the more subtle target stimuli.

When discriminative stimuli contain multiple dimensions, such as shape and color, organisms may show attention to one dimension and ignore the others. Selective attention is common in nonhuman animals and is more common in humans with mental retardation. For example, a child with autism may fail to recognize a teacher who just got a haircut, indicating that the child identified the teacher by her hair. Some attention effects in children are due to language exposure. For example, 3-year-old children who learn the name of a novel object show attention to its shape, not other attributes such as color or texture. This shape bias (or shape generalization) results because of the prevalence of object names in early language training and not because of a perceptual bias.

Attention to a stimulus can also be blocked by an established S^d. For example, after training a discrimination to a green circle, further training with the green circle plus a white cross will result in no learning to the white cross. Blocking is easily shown in nonhuman animals but difficult to demonstrate in humans.

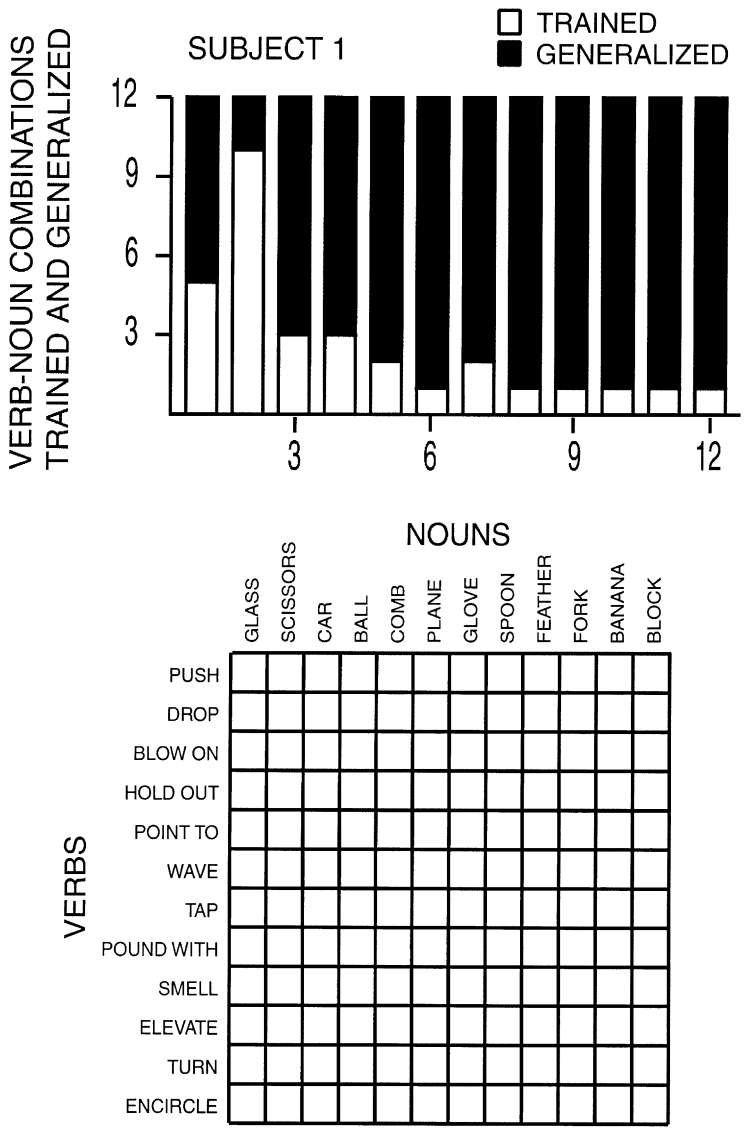


Figure 21 A factorial discrimination contingency. The noun–verb matrix shows the combinations that were arranged for training and testing the receptive skills of two children with mental retardation. A verb was first tested with 12 nouns, and if the child failed to follow the verb–noun instruction he received training. (Top) Learning for one of the children. As the child learned more verbs, his performance reached the point where learning a new verb with only one noun was sufficient to produce generalized compliance with the remaining nouns (data from Striefel *et al.*, 1976, Establishing generalized instruction-following skills in retarded children. *J. Exper. Child. Psych.* **22,** 247–260).

F. Preparedness and Latent Learning

Some three-term contingencies are more readily established than others for reasons that have to do with the species' ecological niche. For example, rats are biologically prepared to learn the spatial locations of food sources and to avoid previously depleted food locations. Similarly, experiments with food-storing birds have found that they selectively attend to the spatial locations of food sites and not their nonspatial aspects such as colors. Preparedness in spatial learning is also evident in latent learning, in which rats have been found to learn mazes in the absence of food reinforcement. Latent learning and related experiments show that rats are prepared to learn the spatial relations between cues and paths. Humans show analogous behavior by remembering the locations of stimulus objects without reinforcement, sometimes called automatic processing.

G. Extinction, Spontaneous Recovery, and Contingency Discrimination

Operant responding declines once a reinforcement contingency is discontinued. However, several findings show that operants are not erased by extinction. For example, after being removed from a situation in which an operant has undergone extinction, a return to the situation will produce a burst of responding called spontaneous recovery. Also, even if a response is completely extinguished, a single reinforcer will reestablish even a complex performance rapidly.

Spontaneous recovery and other findings suggest that extinction is not a distinct learning process but, rather, extinction procedures actually result in discrimination learning. Several lines of evidence support this interpretation of extinction. For example, extinction is more rapid if it is arranged in different conditions than those existing during reinforcement. If instead conditions during extinction are the same as those during reinforcement, responding decreases more rapidly with successive extinction conditions, showing a discrimination of the presence versus the absence of the contingency.

VIII. NEGATIVE REINFORCEMENT: AVOIDANCE AND ESCAPE

Responses can be reinforced by the production of appetitive stimuli or by the removal of aversive stimuli. Negative reinforcement operations include avoidance, in which responses prevent the occurrence of aversive stimuli, and escape in which responses terminate an existing aversive stimulus. Responses reinforced by escape are subject to effects of contiguity as in positive reinforcement; for example, escape is maintained best by immediate removal of aversive stimuli and less well maintained with delayed removal. Avoidance, however, is maintained in the absence of contiguity (responses prevent reinforcement) and has therefore attracted far more attention.

A. Theories of Avoidance

Avoidance can be maintained when responses remove an immediately present conditioned aversive (warning) stimulus. This is called two-factor avoidance because it invokes both operant and respondent conditioning; this reduces avoidance to a variant of escape—the conditioned stimulus elicits fear from which the organism escapes (and thus invokes contiguous reinforcement). However, experiments show that avoidance can be maintained in the absence of conditioned aversive stimuli and experienced organisms do not show fear to conditioned aversive stimuli in avoidance procedures. Figure 22 diagrams a free-operant avoidance procedure in which responses occurring before a fixed time serve to postpone a shock. Without a warning stimulus, rats in this procedure responded with equal probability throughout the interval; with the addition of a warning stimulus toward the end of the stimulus, the rats were more likely to wait for the stimulus and then respond. This and other similar findings run counter to predictions of two-factor theories since subjects waited for the warning stimulus to respond rather than preventing it from occurring.

As in the case of positive reinforcement, avoidance is sensitive to contingency effects. For example, responding is maintained when the probability of an aversive stimulus is lowered following a response but not if the probability of the stimulus is the same regardless of whether responses occur. Avoidance and punishment are closely linked because organisms will learn to avoid situations in which responses are punished.

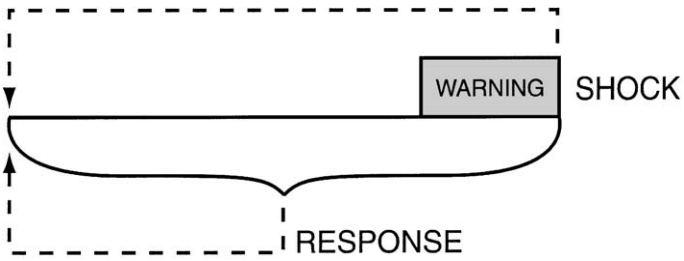


Figure 22 Stimulus function in negative reinforcement.

B. Learned Helplessness

Prior exposure to unpredictable noncontingent aversive stimuli can retard or prevent learning of avoidance responding. Learned helplessness theory holds that organisms can learn about the absence of a contingency, and this retards learning about the presence of a contingency. Some theories of functional depression in humans hold that people who experience many aversive, uncontrolled life events will show learned helplessness.

IX. LANGUAGE AND CULTURE

The social environment is rich in reinforcement contingencies. It provides models to imitate, opportunities to cooperate, and the possibility to exchange reinforcers. Instructional control by verbal discriminative stimuli vastly extends the power of individuals. For the speaker, instructions instantly generate target behavior in the listener; for the listener, instructions generate complex adaptive behavior that might have taken a culture centuries to perfect.

A. Verbal Behavior

Reinforcement theories of language have been proposed from the outset. Skinner outlined the most complete program based entirely on his concept of the discriminated operant. His account distinguished between the behavior of speakers and listeners, parsing their interactions in terms of discriminative stimuli, responses, and social reinforcers. Skinner identified contingencies operating in processes analogous to naming, requesting, and the grammatical processes such as assertion. The account was initially severely criticized by linguists but it continues to be influential and to provoke research. At the heart of the controversy was the question first raised in the opposing views of Descartes and Locke: Is language a unique accomplishment of humans or is it the result of social contingencies of reinforcement?

Early experiments with children were concerned with simple problems such as showing that reinforcement could increase the rate of target vocalizations. Later research turned to the problem of training language in individuals with developmental disabilities. These studies were motivated by both practical and scientific concerns. For example, studies with typical children showed that correspondence between things said and things done, as in ''telling the truth,'' had to be established by differential reinforcement. Another important finding from language programs for children with mental retardation was the importance of noun–verb recombinant generalization training to establish generalized instruction-following repertoires (Fig. 21).

However, these experiments did not answer the important question about reinforcement and language. Some researchers turned to nonhuman primates for a solution. Early efforts to teach chimpanzees (Pan troglodytes) vocal speech failed because their vocal process only allows them to emit a few vowels. Researchers then turned to training sign language in juvenile chimps, but with limited success. Chimps easily learned to sign object names but tended to imitate their trainers and showed little evidence of grammar. In contrast, chimps and bonobos (Pan paniscus) raised from birth in language-rich environments have shown comprehension of simple spoken English grammar. The interpretation of these experiments continues to be controversial, but the apes have demonstrated rudimentary language competencies such as symbolic behavior, grammar, comprehension, and verbal exchanges between themselves.

Reinforcement is also apparent in the language learning of typical children. Early developmental studies with children appeared to show that language developed without parental feedback. However, recent reanalysis of these data show that parents' verbal interactions included large amounts of modeling and differential approval and correction. These findings do not imply that reinforcement processes are sufficient to explain human language but, rather, that social reinforcement plays a role in language development.

B. Generalized Imitation

Almost as important as language is imitation of a model. In fact, the two frequently occur together, such as in ''Watch how I do it.'' Infants show some elicited imitation of facial expressions at birth, but research shows that a generalized repertoire of imitation of skeletal motions, such as that in dance, must be established by reinforcing many instances of imitation.

C. Human Schedule Behavior and Superstition

Infants show schedule performances very much like those seen in other species. However, laboratory studies with human adults show a variety of response

patterns, many of which suggest insensitivity to schedule contingencies. Some of these patterns can be traced to discriminative control by instructions. For example, human subjects earning money by pressing a button according to a FI schedule will respond at very high rates if they are told that the schedule is a FR; they will wait out the entire interval before making a response if they are told it is a FI schedule. Neither pattern is seen in nonhuman organisms: Their behavior is under discriminative control of the instructions and not under direct control of the FI contingencies. Many experiments have shown this effect—the insulation from contingency control by instructional control: Even simply telling humans to press a button to earn money will make them insensitive to reinforcement schedules. Ironically, the very strength of language, its instructional function, interferes with control by the nonsocial environment.

Enduring superstitions have never been documented in other species, but they are extremely common in humans. Laboratory studies of humans suggest that instructions are not sufficient to produce insensitivity; the instructions must be combined with ambiguous tasks. For example, in one study, subjects were told that they could press a button to prevent a computer's speaker from beeping. In reality, the computer presented beeps independently of subjects' button pressing; presses had no consequences. However, subjects who were presented with randomly timed beeps diligently followed the instructions to press and even erroneously reported having prevented the beeps, whereas subjects presented with fixed timed beeps quickly stopped following the instruction to press and correctly identified the inaccuracy of the instruction. Control experiments showed that the superstition did not depend on fortuitous contingencies but on the interaction between instruction and environmental ambiguity. The finding is consistent with the human tendency to show superstitious behavior in the face of unpredictable illnesses, births, weather, sports, and the like. The mechanism is unclear; however, the findings that the superstitions were maintained in the absence of a reinforcement contingency do show that the environment exerts strong discriminative control over instruction following in humans.

X. THE PHYSIOLOGY OF REINFORCEMENT

All attempts to match physiology to behavior have been based on the assumption that the description of behavior will guide the discovery of neurological mechanisms. In the case of reinforcement, the principle of selective neural strengthening has dominated neurological theories since the time of Bain and Thorndike.

Converging evidence suggests that the brain reinforcement system, the selective strengthening process, is mediated by the neuromodulator dopamine. Reinforcers such as food and addictive drugs cause the release of dopamine, and animals will work to obtain electrical brain stimulation of the medial forebrain bundle, the main ascending path for dopamine. These pathways project to various brain regions, including the motor association areas that contain both sensory inputs and connections to motor systems. The effect of dopamine may be to selectively sensitize those glutaminergic processes between sensory and motor neurons that are active at the time of reinforcement. Once the synaptic efficacy of the postsynaptic neuron has been increased in this manner, the neuron will be more likely to respond to glutamate whether or not dopamine is present. Support for this theory is provided by the fact that drugs that block glutamate receptors also interfere with long-term memory in many learning tasks.

The amygdala is implicated in conditioned reflexes involving aversive stimuli, but its role in operant performances involving aversive control has not been determined. Much more is known about the role of the hippocampus in learning. Damage to the hippocampus in rats interferes with memory of the most recently visited locations in a maze but not with memory of the areas in the maze that contain food day after day. This is approximately analogous to the effect in humans with hippocampus lesions, who cannot remember the flow of daily events but can remember that chairs are to sit in, beds are to sleep in, and so on. It is also consistent with behavioral models of reinforcement schedule effects showing that learning must be characterized by both fast (short-term) and slow (long-term) processes.

XI. APPLICATIONS

Knowledge of reinforcement and punishment has guided the evolution of a vast behavior modification industry. The concept of conditioned reinforcement gave rise to token economies used in schools and correctional institutions, shaping suggested the development of programmed instruction now implemented with computers, the matching principle has been applied to the reduction of maladaptive behavior in individuals with mental retardation without the need for punishment, various principles have been applied to

problems of treatment adherence in drug dependency programs, and so on. Among the most innovative and successful applications to date has been the introduction of the Premack principle to language instruction programs for children with autism. These children typically do not develop language without intense training efforts. Previously, training them involved sitting them at a desk, presenting them with objects to name, and reinforcing their efforts with food. The latest and most successful programs for these children employ a method called incidental teaching. The children are placed in a room containing many objects, such as toys, tools, and sandboxes. When a child shows interest in an object, the teacher uses access to the object as a reinforcer for talking. The method guarantees that the teacher is always working with the momentarily most potent reinforcer. The results are rapid and show far greater generalization to new situations than previously reported. Our understanding of the reinforcement mechanisms makes such simple yet sophisticated approaches to complex problems possible.

See Also the Following Articles

AGGRESSION • BEHAVIORAL NEUROGENETICS • CLASSICAL CONDITIONING • COGNITIVE PSYCHOLOGY, OVERVIEW • PROBLEM SOLVING

Suggested Reading

Azrin, N. H., and Holz, W. C. (1966). Punishment. In *Operant Behavior: Areas of Research and Application* (W. K. Honig, Ed.). Prentice Hall, Englewood Cliffs, NJ.

Baum, W. M. (1993). Performance on ratio and interval schedules of reinforcement: Data and theory. *J. Exp. Anal. Behav.* **59,** 245–264.

Cerutti, D. T. (1989). Discrimination theory of rule-governed behavior. *J. Exp. Anal. Behav.* **51,** 259–276.

Donahoe, J. W., and Palmer, D. C. (1994). *Learning and Complex Behavior*. Allyn & Bacon, Boston.

Dragoi, V., and Staddon, J. E. R. (1999). The dynamics of operant conditioning. *Psychol. Rev.* **106,** 20–61.

Fantino, E. (1981). Contiguity, response strength, and the delay-reduction hypothesis. *Adv. Anal. Behav.* **2,** 169–201.

Hammond, L. J. (1980). The effect of contingency upon the appetitive conditioning of free operant behavior. *J. Exp. Anal. Behav.* **34,** 297–304.

Herrnstein, R. J. (1970). On the law of effect. *J. Exp. Anal. Behav.* **13,** 243–266.

Hineline, P. N. (1981). The several roles of stimuli in negative reinforcement. In *Predictability, Correlation, and Contiguity* (P. Harzem and M. D. Zeiler, Eds.), pp. 203–246. Wiley, New York.

McGee, G. G., Krantz, P. J., Mason, D., and McClannahan, L. E. (1983). A modified incidental-teaching procedure for autistic youth: Acquisition and generalization. *J. Appl. Behav. Anal.* **18,** 17–31.

Moerk, E. L. (1983). A behavioral analysis of controversial topics in first language acquisition: Reinforcement, correction, modeling, input frequencies, and the three-term contingency. *J. Psycholinguistic Res.* **12,** 129–155.

Premack, D. (1962). Reversibility of the reinforcement relation. *Science* **136,** 255–257.

Sidman, M. (1994). *Equivalence Relations and Behavior: A Research Story*. Author's Cooperative, Boston.

Staddon, J. E. R. (1983). *Adaptive Behavior and Learning*. Cambridge Univ. Press, Cambridge, UK.

Williams, B. A. (1989). Reinforcement, choice, and response strength. In *Stevens' Handbook of Experimental Psychology: Vol. 2. Learning and Cognition* (R. C. Atkinson, R. J. Herrnstein, G. Lindzey, and R. D. Luce, Eds.), pp. 167–244. Wiley, New York.

Respiration

HOMAYOUN KAZEMI and DOUGLAS C. JOHNSON
Harvard Medical School and Massachusetts General Hospital

GLOSSARY

aortic and carotid bodies Chemoreceptors at the arch of aorta and bifurcation of carotid artery and sensors of PO_2, PCO_2, and pH of arterial blood.

minute ventilation Volume of air inhaled or exhaled per minute.

nucleus tractus solitarius A nucleus in the midbrain in which afferent impulses from aortic and carotid bodies terminate.

respiratory drive Central neural output to drive muscles of respiration.

ventral medullary surface Area near the ventral surface of the medulla that is critical in central chemical drive of ventilation.

Respiration, as considered here, is an automatic function in which the level of ventilation, minute ventilation (volume of air per minute inhaled or exhaled) is set to meet the metabolic needs of the whole body in order to provide adequate quantities of oxygen for metabolism (O_2 consumption) and remove the primary by-product of metabolism, carbon dioxide (CO_2 output). Respiratory control mechanisms that regulate this essential function match respiration to metabolic needs through a complex and diverse system of receptors and feedback loops under control of the brain. In the brain impulses arising from different sites and receptors are integrated and a final coordinated command is issued through the respiratory centers in the medulla to the motoneurons of nerves of respiratory muscles, primarily the phrenic nerves to the diaphragm, nerves of intercostal muscles, and nerves of the muscles of the upper airway. However, there is reflex modulation of this central command by stimuli arising from the respiratory apparatus, which also affect the central output. In this article, we review the current knowledge about the stimuli to respiration, the receptors and mechanisms and sites within the central nervous system (CNS) that control respiration, acid–base and ionic composition of fluids bathing the brain that affect respiration, and neurotransmitters involved centrally in respiratory drive.

I. RESPIRATORY CENTERS AND RESPIRATORY CONTROL

Since one of the primary functions of the lungs is gas exchange and maintenance of acid–base homeostasis, the respiratory control system is exquisitely responsive to changes in O_2 tension (PO_2), CO_2 tension (PCO_2), and hydrogen ion concentration ($[H^+]$) in the blood. The respiratory receptors must initiate rapid responses to transient changes in internal and external environments of the subject and adapt to sustained changes as occurs during exercise, sleep, residing at high altitude, or illnesses that affect not only the lungs and the respiratory system but also the whole body.

Receptors for this elaborate control system vary in structure and function and include mechanical and irritant receptors in the lung and airways, receptors in respiratory muscles, and chemoreceptors in the carotid body and aortic body (peripheral chemoreceptors) and, most important, in the brain stem (central chemoreceptors). Receptors send signals to the brain

respiratory centers based on changes in their chemical or physical environment or their metabolism. The centers in the brain integrate these impulses and produce appropriate neuronal responses that then set the level of ventilation. There is great variation in the types of stimuli and the response they elicit. Since ventilation is essential to life, considerable redundancy exists in the system. In addition to these receptors and their feedback loops, ventilation is also under volitional control. Namely, one can voluntarily increase the level of ventilation (hyperventilation) or reduce ventilation to an extreme (breathold). This implies cerebral cortical control of ventilation, but clearly it has finite limits. The known factors that control ventilation broadly speaking can be divided into two categories: chemical factors and neural factors. Chemical factors are stimuli that arise from changes in PO_2, PCO_2, and pH in the arterial blood and are most critical in respiratory control. Neural factors are neural stimuli that arise from the lungs, pleura, muscles of respiration, pulmonary vessels, joints, and other structures. Figure 1 depicts the known mechanisms of respiratory control, various receptors, and coordination of the respiratory command from the medulla to the muscles of respiration. For this system to work properly the respiratory centers within the

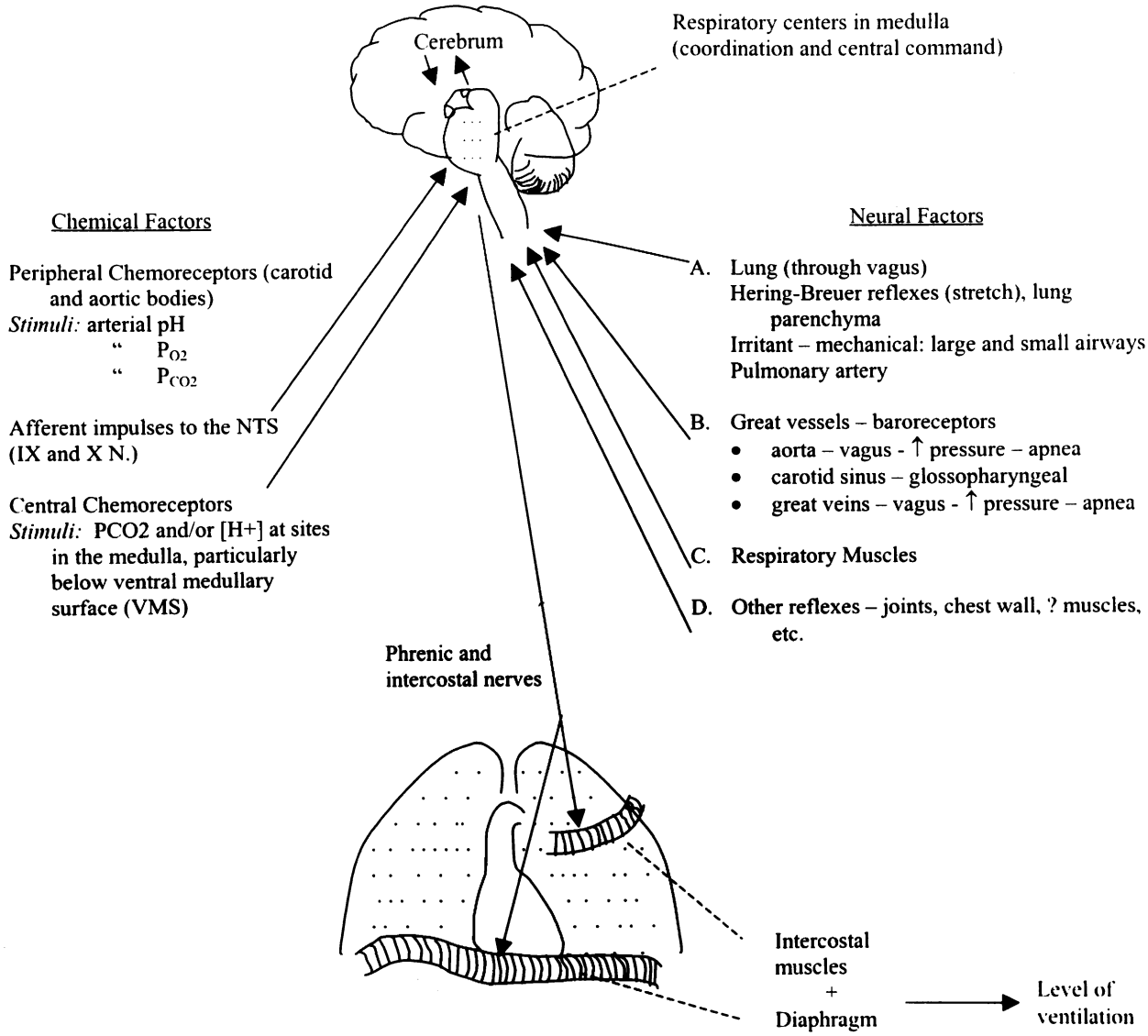


Figure 1 Schematic presentation of both chemical and neural mechanisms in control of ventilation.

brain need to receive the stimuli, integrate them, and send appropriate impulses to inspiratory and expiratory neurons; proper transmission of the impulses from the brain to the respiratory muscles must occur; and there must be a properly functioning lung and thoracic cage capable of responding to the action of respiratory muscles.

A. Chemical Factors

Chemical control of ventilation is the primary control mechanism. Classically, chemical factors influencing ventilation are O_2, CO_2, and pH. Chemoreceptors sensing changes in these variables reside in the carotid body and, to a lesser extent, the aortic body. Carotid and aortic bodies are called peripheral chemoreceptors. Changes in $[H^+]$ and CO_2 in the brain extracellular fluid are sensed by a group of chemoreceptors that are present at several sites in the medulla but predominantly very close to the ventral medullary surface (VMS). These sites are the central chemoreceptors.

1. Peripheral Chemoreceptors

The carotid and aortic bodies are the peripheral chemoreceptors. Carotid body chemoreceptors are located at bifurcation of common carotid arteries and the aortic body chemoreceptors between the arch of the aorta and pulmonary artery. Both sets of chemoreceptors are distinct and separate from baroreceptors that also reside in these bodies. Afferent impulses from aortic chemoreceptors reach the medulla via the vagus nerve, and those from the carotid chemoreceptors reach the medulla via the nerve of Hering to the IX cranial nerve, the glossopharyngeal. In respiratory control carotid chemoreceptors are much more significant that aortic chemoreceptors, particularly in the adult human. Carotid body stimulation increases ventilation several folds. Decreases in arterial PO_2 or increases in arterial PCO_2 increase carotid body firing, as does an increase in $[H^+]$ (i.e., a decrease in pH). Because of its high rate of blood flow in relation to its mass and metabolism, carotid chemoreceptors closely monitor and respond to changes in the arterial blood gases and pH. Five different cell types have been identified in the carotid body, but type I cells, or glomus cells, are thought to possess chemoreceptor function. The exact mechanism of signal transduction in the chemoreceptor cells is not established, and evidence indicates that dopamine and possibly acetylcholine are involved. The afferent impulses from the carotid body reach the caudal part of the nucleus tractus solitarius, where the excitatory amino acid neurotransmitter glutamate is released, increasing ventilation.

Whether hypoxia, hypercapnia, or acidosis stimulate the peripheral chemoreceptors through identical mechanisms is not clear. Evidence suggests that signal transduction by hypoxia may be different from that by hypercapnia. The peripheral chemoreceptors are the primary oxygen sensors in the body and ventilation increases significantly in man when arterial PO_2 decreases to values below 60 Torr. A decrease in arterial pH is associated with an increase in ventilation because of stimulation of peripheral chemoreceptors. The increase in ventilation is relatively modest as arterial pH decreases from 7.40 to 7.25, but once pH decreases below 7.20 there is a brisk and significant increase in ventilation.

Peripheral chemoreceptors also respond to changes in arterial PCO_2, where an increase in PCO_2 stimulates the chemoreceptors and ventilation increases. However, peripheral chemoreceptors have a modest role in CO_2 homeostasis. Peripheral chemodenervation in man causes a modest increase in resting arterial PCO_2 of about 5 Torr. The CO_2 effect is predominantly central by stimulating the central chemoreceptors that reside in the medulla.

2. Central Chemoreceptors

It is well established that central chemical control of ventilation is at the apex of hierarchy of respiratory drive, and that the acid–base status of extracellular fluid (ECF) of the brain bathing the "chemosensitive areas" near the VMS is key in central ventilatory drive. Studies in unanesthetized goats have shown that perfusion of the brain ECF in the medullary region with acid solutions increases ventilation, and that with alkaline solutions depresses ventilation. In unanesthetized animals, other investigators have shown that ventilation can be described as a single function of $[H^+]$ of brain interstitial fluid (ISF) over a large change in $[H^+]$. Changes in PCO_2 instantly affect brain pH and thus ventilation changes to correct the pH. When there is central acidosis, ventilation is stimulated, PCO_2 decreases, and central pH returns toward normal. The reverse occurs when there is central alkalosis. In many physiological situations, acid–base status of the arterial blood mirrors that in the brain ECF, particularly in the steady state.

The exact location of the medullary chemosensitive areas has been studied extensively. By direct application of minute quantities of acid to the VMS, three chemosensitive areas have been identified: the caudal, intermediate, and rostral areas. There are no distinct anatomical landmarks at these sites, but receptors responsive to changes in $[H^+]$ reside approximately 200 μm below the medullary surface in close proximity to brain ISF. Microinjection techniques using acid solutions, CO_2, and acetazolamide in animals have demonstrated several other sites in addition to the areas below the VMS that are involved in respiratory control and chemosensing. The exact mechanism of neuronal stimulation by CO_2 and/or H^+ is unknown. Evidence from several sources has shown that the neurotransmitter acetylcholine is essential in chemoreception centrally in the response to changes in CO_2 and/or H^+. In particular, increases in $[H^+]$ lessen the activity of the enzyme acetylcholinesterase, which degrades acetylcholine; thus, in the presence of central acidosis acetylcholine degradation is slowed, allowing the excitatory neurotransmitter to increase its duration and intensity of action.

In man a number of clinical disorders whose hallmark is diminished ventilatory response to CO_2 (reduced CO_2 sensitivity) have been associated with neuropathological findings in the respiratory-related nuclei in the midbrain, indicating diminished cholinergic activity. These studies have been performed almost entirely in infants and children since lack of CO_2 response increases the risks of hypoventilation, leading to death. In sudden infant death syndrome, there is hypoplasia of arcuate nucleus and diminished expression of acetylcholine receptors, and identical findings have been reported in the brain of children dying from congenital central hypoventilation syndrome (CCHS). Family studies of CCHS have revealed an associated increase in CCHS, suggesting a genetic link, probably related to abnormal development of the cholinergic system. Many patients with CCHS have aganglionic colon (Hirschprung's disease) and a mutation in the *ret* protooncogene.

ret gene encodes for a receptor tyrosine kinase that transduces signals for cell growth and is essential in neural crest development. Since cholinergic pathways are essential for CO_2/H^+ sensitivity, homozygotic *ret* mutant mice show essentially no CO_2 sensitivity and die within 48 hr of birth, indicating the importance of the *ret* gene in autonomic nervous system development, particularly the cholinergic system and the CO_2/H^+ axis in the central chemical control of respiration. Further molecular biology and genetic data in man should help decipher the molecular basis of respiratory drive in the brain.

3. Neural Factors

Neural factors are important in respiratory control but secondary to chemical factors. They become prominent in certain disease states or in special circumstances. Many of the neural stimuli are transmitted to the brain through the vagus nerve.

a. Hering–Breuer (Stretch) Reflexes These reflexes arise from the lung parenchyma. There are two primary reflexes: the inhibitoinspiratory reflex and the excitoexpiratory reflex. Afferent stimuli travel through the vagus. When lung is inflated, the inhibitoinspiratory reflex is stimulated and inspiration is terminated. When lung volume is reduced, excitoexpiratory reflex is stimulated, inspiration is initiated, and respiratory frequency is increased. The excitoexpiratory reflex is stimulated not only when lung volume is reduced but also by pulmonary congestion and inflammatory processes in the lung parenchyma—all leading to an increase in respiratory frequency.

b. Irritant and Mechanical Reflexes Receptors are located near the mucosal surface in the tracheobronchial tree and their stimulation by inflammation or irritant gases such as cigarette smoke leads to cough or forceful expiration as well as the sensation of dyspnea.

c. Pulmonary Vessels and Great Vessels Congestion and distention of pulmonary vessels lead to increased ventilation, mostly an increase in respiratory frequency. An increase in transmural pressure of the aorta and carotid sinus can cause a reduction in ventilation or apnea. These reflexes become important in respiratory control primarily in disease states.

d. Respiratory Muscles Muscles of respiration possess a stretch reflex similar to that in all striated muscles. This reflex through the gamma fibers interacts with alpha fibers coming from the anterior motoneurons in the spinal cord. Thus, there is interaction between local muscle reflexes through the gamma efferent system and impulses arriving from higher centers through the alpha system to set the level of muscle contraction and thereby ventilation.

e. Other Reflexes Impulses arising from systemic muscle groups, joints, and possibly other visceral organs can influence the level of ventilation, but their

roles and mechanisms of action are not well understood, as is the case with hyperventilation associated with exercise.

II. CENTRAL FLUIDS AND RESPIRATION

The brain and spinal cord are bathed in the cerebrospinal fluid (CSF). Ionic composition and acid–base balance of CSF play an important role in the central control of respiration, and "acidity" of this fluid has been thought to be a major stimulus to ventilation since Haldane and Priestly's suggestion in 1905 that CO_2 drove ventilation by making brain "acid."

A. CSF Formation

CSF is a protein-free fluid secreted in the CNS by the choroid plexus and brain tissue. The choroid plexus is responsible for two-thirds of the fluids secreted. A number of factors are responsible for the differences in electrolyte composition and acid–base values in cerebral fluids compared to those in plasma, including the brain–blood barrier, the nature of brain blood vessels, and the mechanisms of CSF formation. The choroid plexus is important in CSF composition since it is a major source of CSF production. The choroid plexus is composed of microvilli and a core containing capillaries and fenestrated endothelial cells with loose connective tissue in between. The cells of the choroid plexus are also endowed with a number of ionic pumps and transport systems. Furthermore, cells of the choroid plexus also contain abundant quantities of carbonic anhydrase, which catalyzes the CO_2 hydration reaction. Carbonic anhydrase catalyzes the reaction $CO_2 + H_2OHCO_3^- + H^+$. In the CNS, glia cells and the choroid plexus contain carbonic anhydrase in high concentrations. The carbonic anhydrase inhibitor acetazolamide, when given intraventricularly or intravenously, reduces CSF formation.

B. CSF Ionic Composition and Acid–Base Balance

There are differences in the ionic composition, pH, and PCO_2 between CSF and plasma. These differences are maintained by active regulation of cerebral fluid composition. CSF is secreted as an ultrafiltrate of the plasma and has Na^+ and Cl^- concentrations similar to those of plasma. The cisternal CSF, however, has a $[Cl^-]$ that is about 10–15 mEq higher than that of the nascent CSF and a $[HCO_3^-]$ that is 10–15 mEq lower than that of the nascent CSF. The mean CSF PCO_2 is about 4–11 mmHg higher than arterial PCO_2, and this gradient is relatively well maintained in respiratory acidosis and alkalosis in the steady state.

Regulation of CNS $[H^+]$ results from changes brought about by the independent variables of PCO_2 and the strong ion difference (SID). SID, which is the difference between cations and anions in the cerebral fluids, determines the $[HCO_3^-]$ since these fluids contain little protein. Of the various strongly dissociated ions, chloride is the most significant in the regulation of CSF acid–base status. Na^+ concentration in the CSF is kept constant in a variety of conditions, probably because of its role as an "osmoregulator." The other cations are also well regulated and kept constant despite major alterations in their plasma concentration. Chloride, whose concentration varies considerably in CSF in acid–base distress, becomes the variable ion in determining SID in cerebral fluids. In metabolic and respiratory acid–base disorders there is a reciprocal relationship between CSF $[Cl^-]$ and $[HCO_3^-]$. Ionic regulation of cerebral fluids is a dynamic process that includes movement of ions to and from blood as well as to and from brain.

In systemic acid–base disorders there are also changes in CSF $[H^+]$ and $[HCO_3^-]$; however, for any given change in plasma $[H^+]$ and $[HCO_3^-]$ there is less of a change in CSF $[H^+]$ and $[HCO_3^-]$. The change in CSF $[HCO_3^-]$ is about 30–40% of the change in plasma $[HCO_3^-]$ in metabolic acidosis and alkalosis and 65–95% in respiratory acidosis and alkalosis in the steady state. With respiratory acidosis, the increase in arterial PCO_2 is rapidly reflected in brain and CSF PCO_2 because CO_2 is freely diffusible across membranes. CSF pH becomes acidic during hypercapnia, and there is then compensation with a variable increase in CSF $[HCO_3^-]$ and decrease in CSF $[Cl^-]$. In high-altitude acclimatization, CSF pH becomes alkaline during the first few hours; by 24 hr, CSF pH becomes less alkaline but changes little thereafter.

The acid–base status of the brain ECF and CSF plays a significant role in determining the ventilatory output for a given PCO_2. At constant temperature, ventilation depends on brain interstitial fluid pH. However, as temperature varies, ventilation no longer tracks pH but the level of α-imidazole or the fractional dissociation of the imidazole moiety of histidine. This results in a decrease of 0.010 pH U/°C and thus helps maintain protein charge states and enzymatic functions constant as temperature and pH change.

In addition to PCO_2, the pH of brain ECF also depends on CSF electrolyte composition, which varies within hours of arterial acid–base changes. Respiratory or metabolic acidosis or alkalosis lead to changes in brain CSF composition that help keep brain ECF pH within a narrow range. Chronic respiratory acidosis leads to a higher CSF [HCO_3^-] and lower CSF [Cl^-], which lowers the central drive of ventilation. With a decrease in PCO_2 in respiratory alkalosis, such as occurs with the respiratory acclimatization to high altitude, CSF pH becomes alkaline and during the first 24 hr it returns toward normal. Regulation of brain ECF electrolyte composition is a complex process and involves a number of cells that are metabolically active, including the glia cells, the cells of the choroid fluxes, and cells at the blood–brain barrier.

III. NEUROTRANSMITTERS AND RESPIRATION

Neuronal firing in the midbrain is essential in the central drive of ventilation. The importance of O_2, CO_2, and H^+ in respiratory control was previously discussed. However, a number of neurotransmitters and neuromodulators centrally regulate neuronal firing and thus the respiratory drive. Changes in O_2 and CO_2 affect brain metabolism; as a consequence, the concentrations of certain amino acid neurotransmitters are changed, which affects the level of ventilation. Broadly speaking, amino acid neurotransmitters and acetylcholine are fast-acting neurotransmitters; their effect on neuronal firing is instantaneous. A diverse group of other substances, such as adenosine, substance P, dopamine, progesterone, and thyroid hormone, also affect ventilation, but their time constants are much slower minutes to hours and they are thus termed neuromodulators.

Among amino acid neurotransmitters involved in respiration are the excitatory amino acids, glutamate and aspartate, and inhibitory amino acids, GABA, glycine, and taurine. The effect of neurotransmitters centrally on ventilatory drive parallels their effects on neuronal function. Excitatory neurotransmitters stimulate ventilation and inhibitory neurotransmitters depress ventilation. Of particular significance is the fact that receptors for many of these neurotransmitters and neuromodulators have been identified in the brain in respiratory-related nuclei, especially in the midbrain. Studies in animals have shown that both GABA and glutamate are operative in setting the resting level of ventilation. Inhibiting GABA increases the resting ventilation, whereas inhibiting glutamate markedly reduces resting ventilation.

Both oxygen and CO_2 affect the concentration of amino acid neurotransmitters in the brain. During hypercapnia, brain free amino acids and acetylcholine concentrations change. Acetylcholine, GABA, and glutamine levels increase, whereas glutamate and aspartate levels decrease. These changes affect neuronal firing and the resulting ventilation is the product of this interaction. The role of acetylcholine in the CO_2/H^+ drive of ventilation was discussed previously. In hypoxia, amino acid neurotransmitters are of particular significance in setting the level of ventilation. Their role in acute hypoxia ventilatory response has been studied extensively, whereas their role in ventilatory adaptation to chronic hypoxia is less clear.

Hypoxia stimulates the peripheral chemoreceptors, and in adult humans primarily the carotid body, whose afferent impulses terminate in the caudal part of the nucleus tractus solitarius and cause the release of the excitatory amino acid glutamate. Brain hypoxia at the same site causes the release of the inhibitory amino acid neurotransmitter GABA. Release of GABA in the brain during hypoxia is due to conversion of glutamate to GABA by the enzymatic action of glutamic acid decarboxylase (GAD): glutamic acid $\xrightarrow[\text{GAD}]{}$ GABA + CO_2. GAD is one of the primary anaerobic enzymes in the brain and it continues to function during hypoxia, allowing for a significant increase in the brain's GABA content. Thus, during acute hypoxia the initial hyperventilatory response in the first 2–5 min is due to the excitatory effects of glutamate; subsequently, with the increase in GABA there is a decrease in ventilation. The level of ventilation at any point in time during acute hypoxia is determined by the interaction between the opposing effects of these amino acids. In addition, there are also increases in aspartate, which stimulates ventilation, and an initial decrease and subsequent increase in taurine, which is a respiratory depressant and thus has a variable effect based on whether it is increasing or decreasing.

Glutamate and GABA centrally also affect cardiac output, systemic blood pressure, oxygen consumption, and CO_2 output in a manner similar to their effects on ventilation. Therefore, these two amino acids in the brain are important in coordinating ventilation, cardiac function, and metabolic demands and maintaining overall homeostasis in the body.

IV. DISORDERS OF CENTRAL RESPIRATORY CONTROL

A. Abnormal Patterns of Breathing

There are cells in the brain stem, mainly in the medulla, that act as central pattern generators, helping set the breathing frequency, inspiratory time, and expiratory time. Abnormal patterns of breathing can occur with certain brain stem lesions. Ataxic or grossly irregular breathing can occur with medullary lesions. Apneustic breathing, characterized by prolonged expiratory pauses, can occur with pontine lesions.

B. Depressed CO_2 Ventilatory Response

The central ventilatory drive can be assessed by studying the change in ventilation with increasing levels of PCO_2. In normal individuals, the slope of the CO_2 ventilatory response is approximately 2 or 3 liters/mm increase in PCO_2 (i.e., ventilation increases by 2 or 3 liters/min for every millimeter increase in PCO_2. However, about 15% of adults have a diminished response to CO_2, with the slope being approximately 1 liter/mm PCO_2. These individuals are likely to develop CO_2 retention when additional problems occur, such as obesity, obstructive lung disease, or status asthmaticus. Many normal family members of patients with CO_2 retention also have a depressed ventilatory response to CO_2, again indicating a possible genetic basis for the depressed CO_2 ventilatory response. One beneficial effect of depressed CO_2 ventilatory response is seen in endurance athletes, many of whom have a CO_2 ventilatory slope of approximately 1 liter/mm PCO_2. The reasons for this are not clear, but it is probably due to a combination of genetic factors and training.

Depressed ventilatory response to CO_2 can also occur with elevation of serum and brain ECF bicarbonate, as in metabolic alkalosis or chronic CO_2 retention, because an increase in PCO_2 causes a reduced increase in hydrogen ion concentration with the increased bicarbonate concentration.

C. Ondine's Curse

The term Ondine's curse refers to patients with alveolar hypoventilation due to impaired autonomic control of ventilation. Classically, they "forget to breathe" when they fall asleep but maintain relatively normal blood gases while awake. Ondine's curse is usually due to CCHS, but it can be caused by surgery-induced incisions into the second cervical segment to relieve intractable pain.

D. Congenital Central Hypoventilation Syndrome

CCHS is associated with a nearly absent respiratory response to hypoxia and hypercapnia, mildly elevated arterial PCO_2 during wakefulness and markedly elevated PCO_2 during non-rapid eye movement (REM) sleep, and no respiratory discomfort during CO_2 inhalation. CCHS often occurs in association with Hirshsprung's disease, which is characterized by abnormalities of the cholinergic system in the gastrointestinal tract. This supports the view that the cholinergic system is crucial for central chemical drive, as discussed previously.

E. Cheyne–Stokes Respiration

Cheyne–Stokes respiration is a form of periodic breathing in which apnea is followed by gradually increasing breaths and then gradually decreasing breaths to the next apneic period. Patients with neurologic cause for Cheyne–Stokes respiration generally show increased respiratory sensitivity to CO_2, likely related to depressed cortical inhibition. Cheyne–Stokes respiration is also present in some patients with congestive heart failure.

F. Myxedema

Hypoventilation can occur in patients with hypothyroidism (myxedema). This appears to be due in part to both a depressed drive to breathe centrally and respiratory muscle weakness.

G. Sleep and Respiration

Sleep has profound effects on respiration and is mentioned briefly here. Breathing disorders during sleep are relatively common and include obstructive sleep apnea, hypopneas, hypoventilation, Cheyne–Stokes respiration, and central apneas. The central drive to breathe (CO_2 response) decreases during sleep, more so during REM than non-REM sleep. Muscle tone also decreases during sleep, promoting a tendency

for the upper airway to collapse and develop obstructive hypopneas or apneas.

See Also the Following Articles

CHEMICAL NEUROANATOMY • HOMEOSTATIC MECHANISMS • NEUROTRANSMITTERS • PSYCHOPHYSIOLOGY • SLEEP DISORDERS

Suggested Reading

Burton, M. D., and Kazemi, H. (2000). Neurotransmitters in central respiratory control. *Respir. Physiol.* **122,** 111–121.

Coates, E. L., Li, A., and Nattie, E. E. (1993). Widespread sites of brain stem ventilatory chemoreceptors. *J. Appl. Physiol.* **75,** 5–14.

Coleridge, H. M., and Coleridge, J. C. G. (1986). Reflexes evoked from tracheobronchial tree and lungs. In *Handbook of Physiology. Section 3, The Respiratory System. Vol. II, part I: Control of Breathing* (N. S. Cherniack, and J. G. Widdicombe, Eds.), pp. 395–429. American Physiological Society, Bethesda, MD.

Dempsey, J. A., and Forester, H. V. (1982). Mediation of ventilatory adaptation. *Physiol. Rev.* **62,** 262–346.

Housley, G. D., and Sinclair, J. D. (1988). Localization by kainic acid lesions of neurons transmitting the carotid chemoreceptor stimulus for respiration in rat. *J. Physiol. (London)* **406,** 99–114.

Kazemi, H., and Hitzig, B. (1994). Control of ventilation: Central chemical drive. In *Maxwell and Kleeman's Clinical Disorders of Fluid and Electrolyte Metabolism* (R. Narins, Ed.), 5th ed., pp. 175–186. McGraw-Hill, New York.

Kazemi, H., and Hoop, B. (1991). Glutamic acid and γ-aminobutyric acid neurotransmitters in central control of breathing. *J. Appl. Physiol.* **70,** 1–7.

Kazemi, H., and Johnson, D. C. (1986). Regulation of cerebrospinal fluid acid–base balance. *Physiol. Rev.* **66,** 953–1037.

Leusen, I. R. (1954). Chemosensitivity of the respiratory center. Influence of changes in the H^+ and total buffer concentrations in the cerebral ventricles on respiration. *Am. J. Physiol.* **176,** 45–51.

Mitchell, R. A., Leoschcke, H. H., Massion, W. H., and Severinghaus, J. W. (1963). Respiratory responses mediated through superficial chemosensitive areas on the medulla. *J. Appl. Physiol.* **18,** 523–533.

Pappenheimer, J. R., Fencl, V., Heisey, S. R., and Held, D. (1965). Role of cerebral fluids in control of respiration as studied in unanesthetized goats. *Am. J. Physiol.* **208,** 436–450.

Schlaefke, M. E. (1981). Central chemosensitivity: A respiratory drive. *Rev. Physiol. Biochem. Pharmacol.* **90,** 171–244.

Retina

JOHN E. DOWLING
Harvard University

GLOSSARY

amacrine cells Neurons whose processes are confined to the inner plexiform layer.

bipolar cells Output neurons of the outer plexiform layer that carry information to the inner plexiform layer.

fovea Specialized retinal region of highest visual resolution, containing only cones.

ganglion cells Third-order neurons in retina whose axons form the optic nerve and carry visual information from the eye to the rest of the brain.

horizontal cells Neurons whose cell bodies lie along the distal margin of the outer nuclear layer and that extend processes mainly in the outer plexiform layer.

interplexiform cells Neurons whose cell bodies reside in the proximal part of the inner nuclear layer and that extend processes in both plexiform layers.

plexiform layers Regions consisting principally of neuronal processes, where synaptic interactions take place.

receptive field Area of retina that when illuminated influences the activity of a cell.

visual pigments Light-sensitive molecules in photoreceptors consisting of 11-*cis*-retinal (vitamin A aldehyde) and protein (opsin).

The retina is a thin (~0.25-mm) layer of neural tissue that lines the back of the eye. It is a true part of the brain displaced into the eye during development. In addition to the light-sensitive photoreceptor cells, the retina contains five basic classes of neurons and one principal type of glial cell, the Müller cell. The neurons are organized into three cellular (nuclear) layers, which are separated by two synaptic (plexiform) layers. Virtually all the junctions (synapses) between the retinal neurons are made in the two synaptic layers, and all visual information passes across at least two synapses, one in the outer plexiform layer and another in the inner plexiform layer, before it leaves the eye.

I. INTRODUCTION

Processing of visual information occurs in both plexiform layers of the retina. The outer plexiform layer separates visual information into "on" and "off" channels and carries out a *spatial*-type analysis on the visual input. The output neurons of this layer, the "on" and "off-" bipolar cells, demonstrate a center-surround antagonistic receptor field organization. The inner plexiform layer is concerned more with the *temporal* aspects of light stimuli. Many cells receiving input in this layer respond with transient responses and respond better to moving stimuli than to static spots of light. The output neurons of this layer, the ganglion cells, tend to reflect the processing of information in either the outer plexiform layer (i.e., the cells respond in a sustained fashion to appropriately positioned stimuli) or the inner plexiform layer (i.e., the cells respond better to moving stimuli than to static ones).

Figure 1 shows a light micrograph of a piece of human retina. The photoreceptors are located furthest from the front of the eye, at the top of the micrograph.

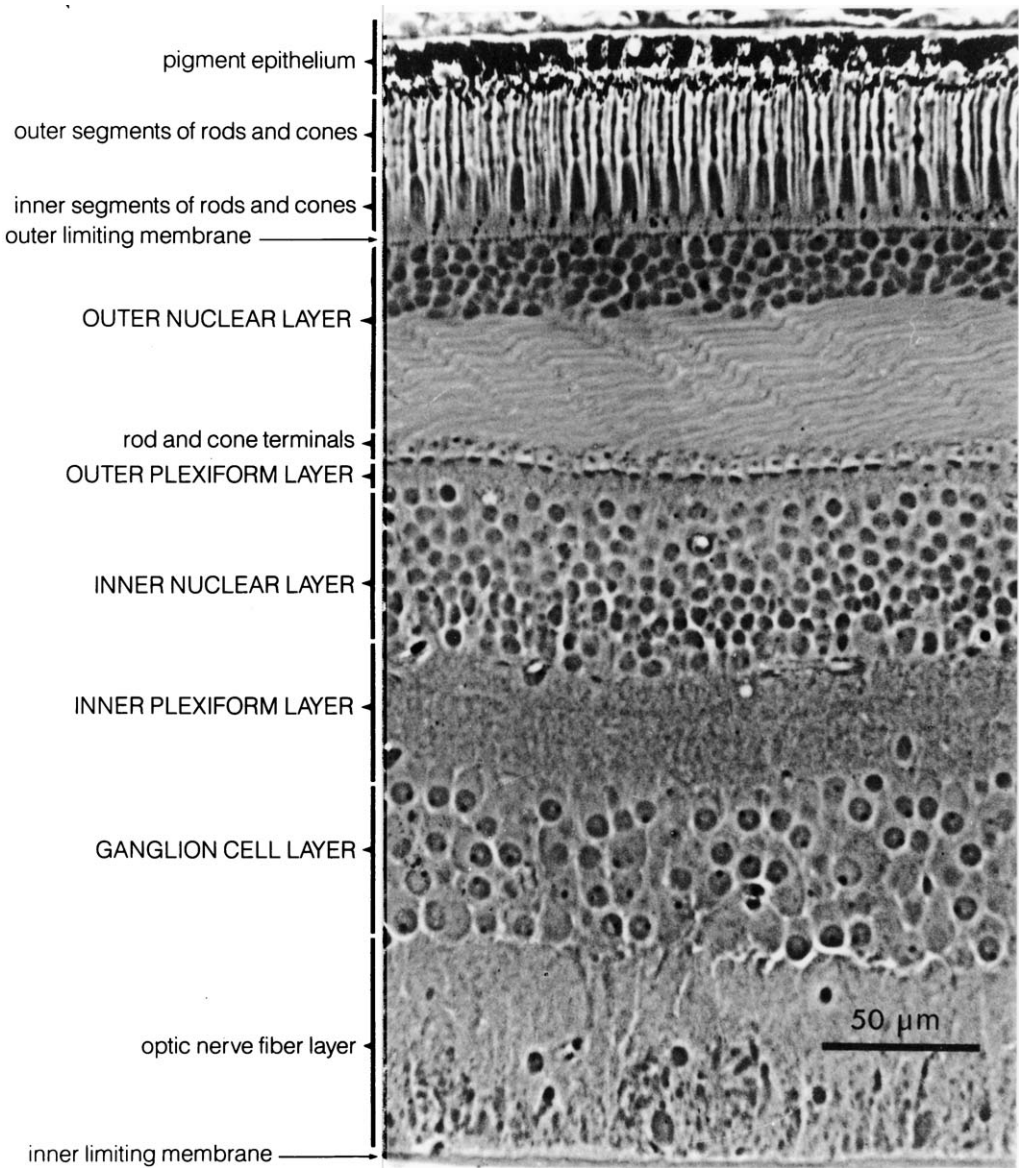


Figure 1 Vertical section through the human retina. Micrograph shows an area approximately 1.25 mm from the center of the fovea. In the foveal region of the retina, the inner layers of the retina are pushed aside so that light can impinge more directly on receptors. Thus, around the fovea and for some distance away (as shown here), receptor terminals are displaced laterally from the rest of the photoreceptor cells.

Light entering the eye passes through the transparent retina and is captured by the pigment-containing outer segments of the photoreceptors. (Overlying the photoreceptors is the pigment epithelium, which serves to absorb stray light and to prevent backscatter of light into the retina.) The cell bodies of the photoreceptors are located in the outer nuclear layer, whereas the cell bodies of three of the basic classes of retinal neurons—horizontal, bipolar, and amacrine cells—are in the inner nuclear layer. The cell bodies of the ganglion cells make up the most proximal cellular layer. The outer and inner plexiform layers are interspersed, respectively, between the outer and inner nuclear layers and the inner nuclear and ganglion cell layers.

In many primates, including humans, a small region of the retina is specialized for high-acuity vision. It is called the fovea and is centrally located (i.e., on the

visual axis of the eye). The layers of the retina below the photoreceptor inner and outer segments are displaced from the fovea so that light can impinge directly on the photoreceptors. Only cones are present in this area, and the foveal cones are the thinnest and longest photoreceptors in the retina. The rod-free area is about 0.3 mm in diameter and contains approximately 35,000 cones. No blood vessels are found in the fovea, and there are few, if any, blue-light absorbing cones in the center of the fovea. These specializations serve to improve the visual resolution of the fovea.

II. PHOTORECEPTORS

The human retina contains two types of photoreceptors, rods and cones, differentiated on the basis of their outer segment shape (see Fig. 1). This criterion is not always reliable; for example, the outer segments of the cones found in the fovea show no significant taper (i.e., they are rod shaped). Rods mediate dim-light vision, whereas cones function in bright light and are responsible for color vision.

A. Visual Pigments

Light sensitivity of the photoreceptors results from the presence of visual pigment molecules contained within their outer segments. One rod pigment, called rhodopsin, and three cone pigments are in the human retina. Rhodopsin absorbs light maximally in the blue-green region of the spectrum ($\sim$500 nm), whereas the human cone visual pigments absorb maximally in the blue ($\sim$430 nm), green ($\sim$530 nm), and red-yellow ($\sim$560 nm) regions of the spectrum. The cone pigments are segregated into separate classes of cones; thus, blue-, green-, and red-yellow-sensitive cones are in the human retina. In the human eye, there are more red and green cones than blue cones and, as noted previously, blue cones are extremely rare in the high-acuity foveal region of the retina.

The genes encoding for rhodopsin and the cone pigments have been identified and isolated in humans and in a number of other species. The red- and green-sensitive pigment genes in humans are on the X chromosome (i.e., they are sex-linked), whereas the genes for the blue-sensitive pigment and rhodopsin are on autosomes. The red- and green-sensitive visual segments are highly homologous ($\sim$95%), whereas there is approximately 40% homology between the red and green pigments and the blue pigment and approximately the same homology between all the cone pigments and rhodopsin. Studies of the color vision pigment genes in red- and green-blind individuals have shown that color blindness is caused by a loss or alteration of one or another of the genes. Red-blind individuals (protanopes) have an altered gene for the red-sensitive pigment, whereas green-blind individuals (deuteranopes) have an altered green-sensitive pigment gene or are lacking the gene altogether. Blue-blind individuals (tritanopes) likewise are missing or have an altered blue visual pigment. When the gene for the red or green pigment is altered, usually no red- or green-sensitive cone cells develop in the retina. Sometimes, red- or green-sensitive cones form, but the pigment within the cell, and therefore color vision, is abnormal (a condition termed anomalous trichromacy).

The visual pigment molecules are concentrated to a high degree in the outer segments of the photoreceptors. The outer segments contain numerous transverse membranous discs (Fig. 2), and virtually all the visual pigment molecules are contained within the disc membranes. A typical outer segment may have as many as 2000 transverse discs and may contain 10^9 visual pigment molecules.

1. Visual Pigment Chemistry

All visual pigments have a similar chemistry. They consist of two components: retinal (vitamin A aldehyde), termed a chromophore, which is bound to a protein called opsin. Different visual pigments have different opsins, and this accounts for the variations in their color sensitivity. The light sensitivity of the visual pigments is due to the retinal chromophore. When a visual pigment molecule absorbs a quantum of light, several molecular transformations occur, first in the chromophore and then in the protein (opsin) part of the molecule. These transformations lead to the excitation of the photoreceptor cell and also to the separation of the retinal chromophore from opsin. This latter process is called bleaching because it results in the loss of color of the visual pigment molecules and their ability to absorb visible light.

Retinal can exist in different shapes (i.e., several *cis–trans* isomers of the molecule are possible). All visual pigments require one particular isomer, the 11-*cis*, for their synthesis, and this form of the chromophore combines spontaneously with opsin to form visual pigment. When a visual pigment molecule absorbs a

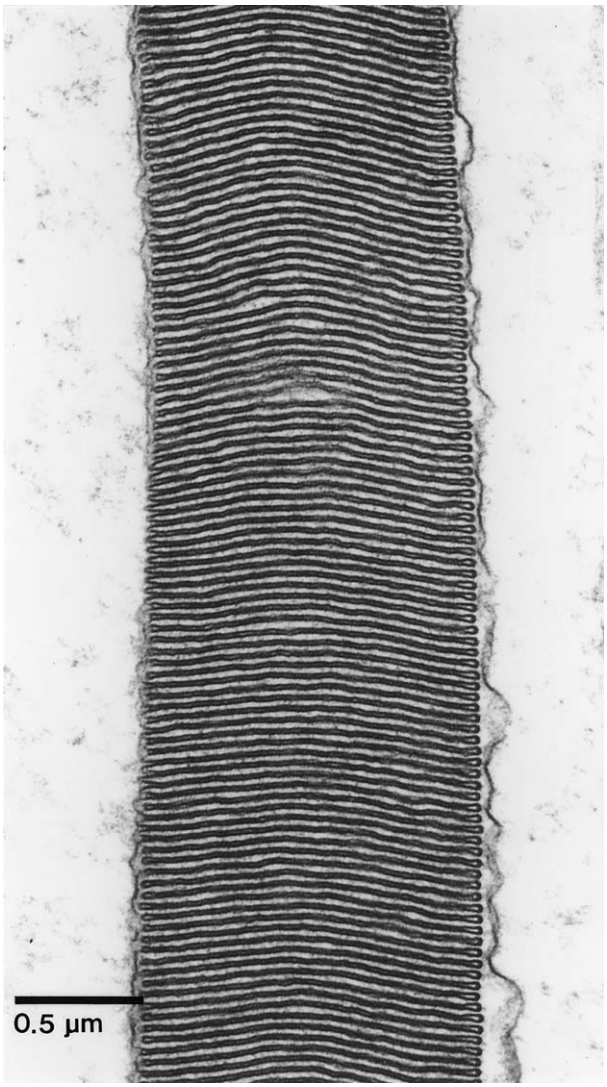


Figure 2 Electron micrograph of a portion of a cone outer segment. Contained within the structure are numerous transverse membranous discs. The visual pigment molecules and associated proteins that lead to excitation of the photoreceptors are contained within or in association with the disc membranes.

quantum of light, the first transformation is the isomerization of the chromophore from the 11-*cis* to the all-*trans* form. Indeed, the only action of light in the visual process is to change the shape of the chromophore. The chromophore shape change initiates a series of conformational changes in the opsin, and this leads to excitation of the photoreceptor cell and release of the chromophore from opsin.

A series of intermediates have been identified between the absorption of light by visual pigment molecules and the release of the chromophore from opsin. One of the intermediates, metarhodopsin II, is responsible for excitation of the photoreceptor cells (i.e., it is the photoactive intermediate). Figure 3 shows a simplified scheme of the visual cycle for the rod visual pigment rhodopsin.

The retinal chromophore of rhodopsin derives from vitamin A, and it represents a slightly oxidized (aldehyde) form of the vitamin. During the operation of the visual cycle, some retinal is lost and must be replaced from body stores of the vitamin. In vitamin A deficiency, this replenishment fails. A full complement of visual pigment can no longer be synthesized in the photoreceptors, and light sensitivity of the rods and cones is decreased. The loss of light sensitivity is more obvious in the dark; hence, the condition is known as night blindness. Refeeding of vitamin A to a vitamin A-deficient animal or patient usually restores visual sensitivity.

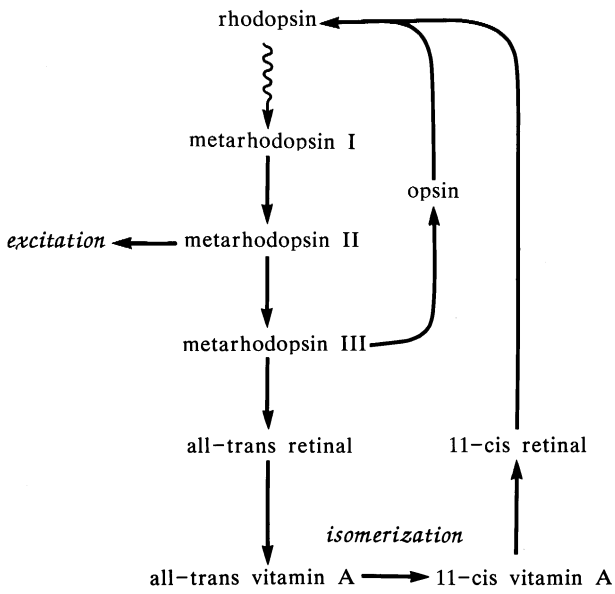


Figure 3 Scheme of the sequence of events that occur following the absorption of a quantum of light by the rod visual pigment, rhodopsin. Light initiates the conversion of rhodopsin to retinal and opsin through a series of metarhodopsin intermediates. Metarhodopsin II is the active intermediate leading to excitation of the photoreceptor cell. Eventually, the chromophore of rhodopsin, retinal, separates from the protein opsin and is reduced to vitamin A (retinol). For the resynthesis of rhodopsin, the shape of vitamin A must be changed (isomerized) from the all-*trans* to the 11-*cis* form, and this isomerization takes place in the pigment epithelium overlying the receptors (see Fig. 1). Vitamin A is replenished in the eye from the blood.

B. Photoreceptor Responses

Excitation of visual pigment molecules leads ultimately to a change of potential across the membrane surrounding the photoreceptor cell. Membrane potential controls the release of neurotransmitter from synapses, and in this way light influences the exchange of information between photoreceptors and second-order neurons in the retina.

All vertebrate photoreceptors hyperpolarize in response to light (i.e., the membrane potential becomes more negative), for the following reason: In darkness, the membrane of the outer segment is permeable (leaky) to Na^+. Because Na^+ levels are higher outside the cell than inside, positive Na^+ ions enter the cell in darkness, causing the cell to be partially depolarized (i.e., the membrane potential is more positive than is typically the case for neurons at rest). Light decreases the conductance of the outer segment membrane to Na^+, thereby decreasing the flow of positive ions into the cell and causing the cell to become more negative (i.e., to hyperpolarize).

The conductance of the outer segment membrane to Na^+ is controlled by a second messenger molecule, cyclic GMP, which maintains channels in the membrane in an open state. In the light, levels of cyclic GMP decrease because an enzyme (phosphodiesterase), which breaks down cyclic GMP, is activated, and this causes the channels in the outer segment membrane to close. Phosphodiesterase, in turn, is activated by a protein called transducin (a G protein), and transducin is activated by metarhodopsin II, the photoactive visual pigment intermediate. The cascade between the visual pigment molecule and the Na^+ channel in the outer segment membrane is shown in Fig. 4.

The cascade between the light activation of visual pigment molecules and the closing of Na^+ channels in the outer segment membrane is important for amplification of the signal. That is, one photoactive visual pigment molecule can interact with many transducin molecules (as many as 500), and one phosphodiesterase molecule can break down about 2000 cyclic GMP molecules per second. The cascade of reactions between photon absorption and cyclic GMP inactivation can result in an amplification of about 10^6.

The channels in the outer segment membrane controlled by cyclic GMP also allow some Ca^{2+} to enter the photoreceptor cell, and Ca^{2+} plays an important regulatory role in the phototransduction process. For example, Ca^{2+} entering the cell inhibits an enzyme, guanylate cyclase, which synthesizes cyclic GMP. Thus, in the dark, cyclic GMP synthesis is relatively low. In the light, when the channels in the outer segment membrane are closed, Ca^{2+} entry into the cell decreases and intracellular Ca^{2+} levels decline. This results in an increased synthesis of cyclic GMP, which serves to counter the effect of light lowering cyclic GMP levels. Thus, in continuous light, the photoreceptor response recovers partially to dark levels, enabling the photoreceptor to continue to respond even in bright light. This process is termed adaptation and photoreceptor light and dark adaptation play an important role in the ability of the visual system to respond over a wide range of ambient illumination. Ca^{2+} appears to mediate many aspects of photoreceptor light and dark adaptation.

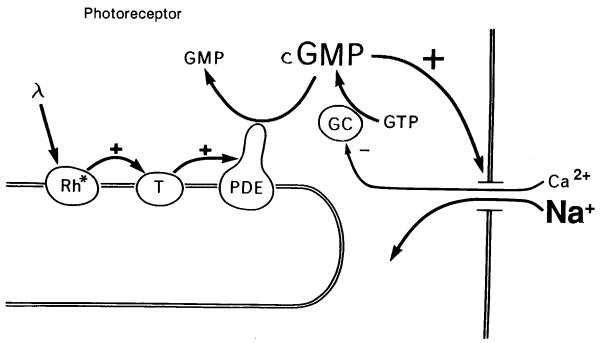


Figure 4 Summary diagram of interactions occurring in the rod outer segment during phototransduction. Light-activated rhodopsin (Rh*) activates a G protein transducin (T), which in turn activates the enzyme phosphodiesterase (PDE). These interactions occur in the disc membrane. Activation of PDE leads to breakdown of cyclic GMP (cGMP) to an inactive product (GMP). Cyclic GMP maintains channels in the outer segment membrane in an open configuration, thereby allowing both Na^+ and Ca^{2+} to enter the cell in the dark. With a decrease in cGMP levels in the light, channels in the outer segment membrane close. The resulting decrease in Na^+ levels causes the cell to hyperpolarize. Decrease of Ca^{2+} levels enhances guanylate cyclase (GC) activity, an action that counters the effects of light and increases GMP levels in the outer segment.

III. CELLULAR AND SYNAPTIC ORGANIZATION

A. Cellular Organization

Most of what is known about the classes of retinal cells in all vertebrates has come from light microscopic studies of retinas processed by the silver-staining method of Golgi. This technique enables investigators

to see the extent and distribution of the processes of the cells within the retina, to classify the cells, and to construct schemes of the cellular organization of the retina such as that shown in Fig. 5 for the primate retina. Although there are just five major classes of neurons in the retina, there are many morphological subtypes in each major cell class, and even today the total number of morphological subtypes is not known. It is clear that some cell subtypes are much more common than others, and the focus here is on the major subtypes of cells in the mammalian and primate retinas.

In the outer plexiform layer, two cell classes—horizontal and bipolar cells—receive input from the photoreceptors. The bipolar cells are the output neurons for the outer plexiform layer; all information passes from outer to inner plexiform layers via these neurons. Horizontal cells, however, extend processes widely in the outer plexiform layer, but their processes are confined to this layer. Their role is to mediate lateral interactions within this first synaptic zone.

Bipolar cell terminals provide the input to the inner plexiform layer, and two cell classes—amacrine and ganglion cells—are ultimately activated. Amacrine cells, like horizontal cells in the outer plexiform layer, spread processes widely in the inner plexiform layer, but their processes are usually confined to this layer. Ganglion cells are the output neurons for the retina; their axons run along the margin of the retina, collect at the optic disc to form the optic nerve, and carry all visual information to higher visual centers.

1. Outer Plexiform Layer Neurons: Horizontal and Bipolar Cells

Most mammalian retinas contain two basic subtypes of horizontal cells; a cell with a short axon that runs approximately 400 μm before ending in a prominent terminal expansion and an axonless cell. The axonless cell has not been seen in the primate retina; rather, the second horizontal cell type in primates has a short axon, although its synaptic connections are like those of the axonless cell—they are all cone.

Horizontal cells of the cat have been studied in detail, and Fig. 6 shows examples observed by looking down on a flat mount of the whole retina. Electron

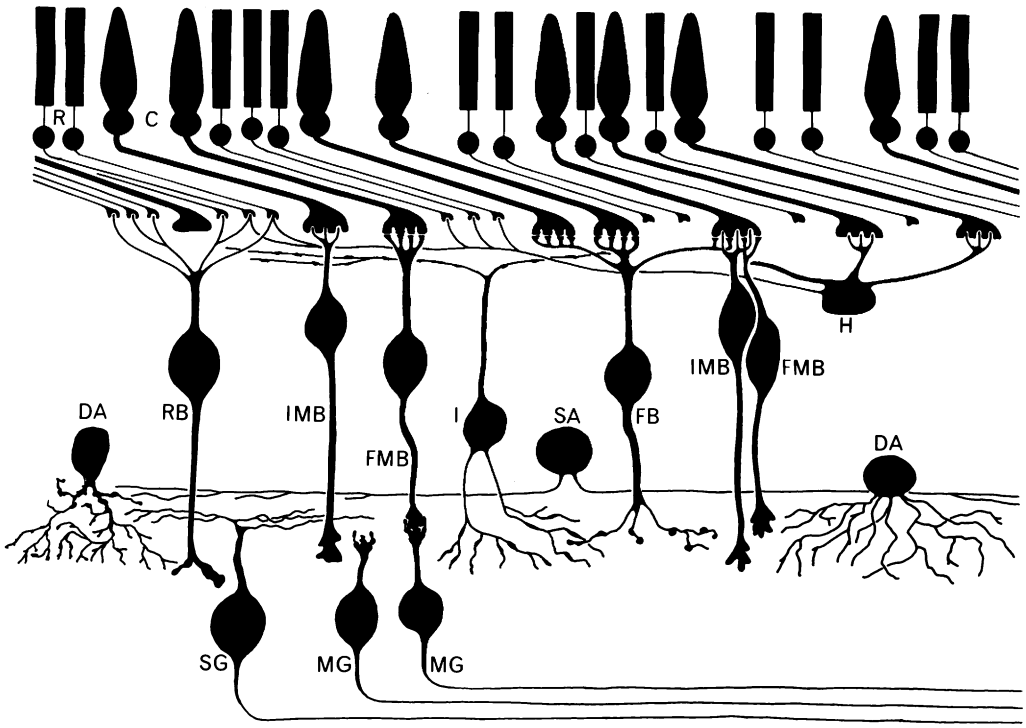


Figure 5 Major cell types found in primate retina as viewed in vertical sections of Golgi-stained retinas. See text for description of cells. R, rods; C, cones; RB, rod bipolar cells; IMB; invaginating midget bipolar cell; FMB, flat midget bipolar cells; FB, flat bipolar cell; H, horizontal cell; I, interplexiform cell; DA, diffuse amacrine cell; SA, stratified amacrine cell; MG, midget ganglion cell; SG, stratified ganglion cell.

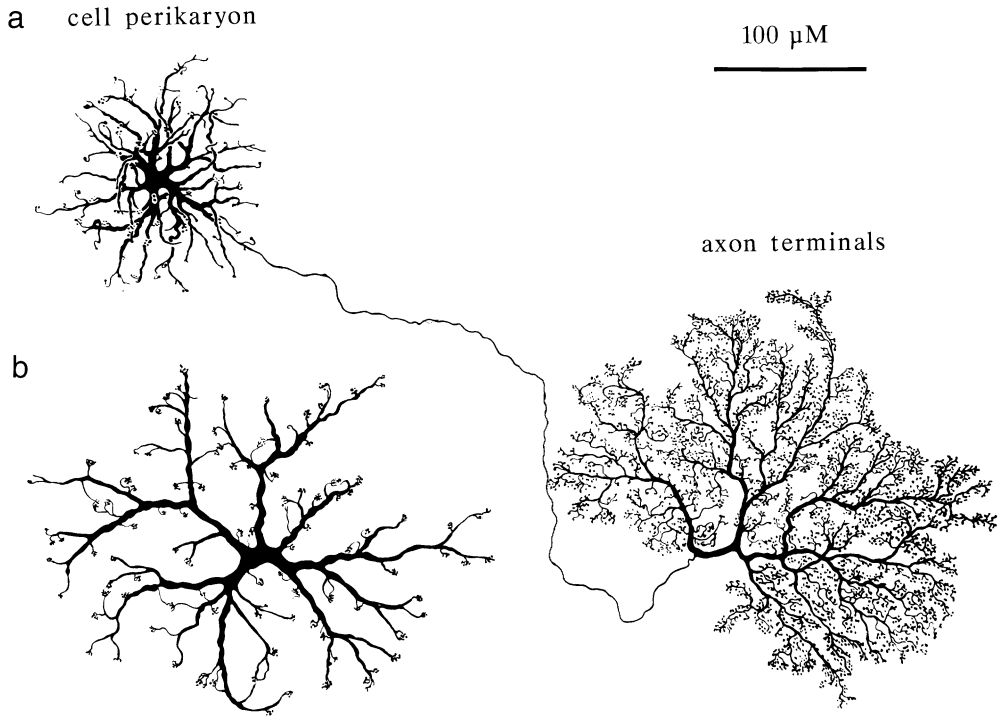


Figure 6 Drawings of Golgi-stained horizontal cells from the cat, viewed by looking down on a flat mount of retina. There are two types of horizontal cells in cat retina: an axonless cell (b) and a cell with a short axon (a).

microscopy has shown that the processes of the axonless cell (Fig. 6b) and the proximal dendritic processes of the short axon cell (Fig. 6a) connect exclusively with cone synaptic terminals, whereas the axon terminal processes of the short axon cell end exclusively in the rod terminals. Why the input to the short axon cells is organized this way is not clear; an appealing suggestion is that this allows for segregation of rod and cone responses in different regions of the cells. In fish, for example, there are separate rod and cone horizontal cells.

In primates, up to 11 subtypes of bipolar cells have been distinguished: One type is exclusively connected to rods and the rest are connected to cones. The rod-related bipolar cells extend their dendrites into the rod synaptic terminals, and their axon terminals end deep in the inner plexiform layer. The rod bipolars contact between 15 and 50 rod terminals depending on eccentricity. More peripherally located rod bipolar cells contact more rods.

Two of the cone-related bipolar cell types contact only a single cone terminal. These cells, called midget bipolar cells, make different kinds of synaptic contacts with the cone terminals: invaginating and flat contacts (see Section III.B). Their axon terminals end at different levels within the inner plexiform layer, and they appear to be related to the generation of either "on" or "off" responses to light in the retina. Every cone terminal in the primate retina probably makes connections with both kinds of midget bipolar cells (see Fig. 5). In addition to the cone midget bipolar cells, there are also bipolar cells that contact several cones, probably as many as six or seven. They are called diffuse flat bipolar or diffuse invaginating bipolars based on the type of connection they make with the photoreceptors (see Section III.B).

2. Inner Plexiform Layer Neurons: Amacrine, Interplexiform, and Ganglion Cells

Most amacrine cells have no axonal processes; all their processes look similar. Some amacrine cells do show two sets of processes emerging from the cell body. The thinner of the two has a wider extent and has been termed an axon. Many such processes (up to 40) can emerge from a single cell body and it is not known

whether they have axon-like properties (i.e., conduct action potentials). Amacrine cells are diverse in terms of the extent and distribution of their processes, and it is possible to describe a large number of amacrine cell subtypes in most species (up to 30–40). Investigators typically classify amacrine cells into two major types: diffuse and stratified amacrine cells. Diffuse amacrine cells extend their processes throughout the thickness of the inner plexiform layer, whereas the stratified cells extend their processes on one or a few levels in the layer. This simple classification scheme can be expanded to include narrow- and wide-field diffuse or stratified amacrine cells, depending on how far their processes extend, and mono-, bi-, or multistratified cells, depending on whether their processes are confined to one, two, or several levels in the inner plexiform layers, respectively.

A cell termed the interplexiform cell deserves special mention. Some investigators classify these cells as a separate class of retinal neuron, but others believe they are a subtype of amacrine cell. Their perikarya (cell bodies) sit among the amacrine cells, but unlike typical amacrine cells they send processes to both plexiform layers. Input to these cells is in the inner plexiform layer, whereas most of their output is in the outer plexiform layer. Hence, they appear to be mainly a centrifugal type of neuron, carrying information from inner to outer plexiform layers.

Ganglion cells, like amacrine cells, are also diverse in their morphology. In most vertebrate retinas as many as 20 morphological subtypes can be distinguished. They are mainly stratified cells. In primates, particularly in the central region of the retina, ganglion cells are observed with limited dendritic fields. These cells, termed midget ganglion cells, receive input from one midget bipolar cell. The dendrites of these midget ganglion cells ramify in either the upper or the lower parts of the inner plexiform layer so that they receive input from one or the other of the two midget bipolar cells described earlier. This implies that one cone in the central (foveal) part of primate retina can send two separate messages to the rest of the brain via two midget ganglion cells. One cell is believed to signal increases in illumination of the cone (i.e., it is an ''on'' cell); the other signal decreases in cone illumination (i.e., it is an ''off'' cell).

The other major ganglion cell subtype in the primate retina is called the parasol ganglion cell. It receives most of its bipolar cell input from diffuse bipolar cells, extends its processes in either the ''on'' or ''off'' laminae of the inner plexiform layer, and is physiologically either an ''on'' or ''off'' cell.

3. Retinal Processing of Visual Information: Ganglion Cell Responses and Receptive Field Organization

As noted earlier, ganglion cells signal brightness and darkness information to the rest of the brain, but they also communicate much more. Indeed, two basic kinds of processing appear to be occurring in the retina: one carried out mainly in the outer plexiform layer and the other in the inner plexiform layer. Ganglion cells convey information to the rest of the brain that reflects these two stages of processing.

Ganglion cells typically respond to illumination of a restricted but relatively large region of the retina. This region is called the receptive field of the cell and is typically about 1 mm in diameter. The most common ganglion cell in the mammalian retinas shows evidence of spatial processing of visual information in its responses. Ganglion cells of this type are often called contrast-sensitive cells, and they are subdivided into two mirror-image classes: ''on''-center, ''off''-surround cells and ''off''-center, ''on''-surround cells. Each has a receptive field that is organized into two concentric zones that are antagonistic to each other (Fig. 7a).

''On''-center cells respond to an increase in the illumination of the receptive field center with a sustained burst of nerve impulses, whereas illumination of the surround inhibits the firing of nerve impulses by the cell for as long as the light is on. ''Off''-center cells respond in the opposite way: Illumination of the receptive field center inhibits the cell, whereas surround illumination provides a sustained excitation of the cell. In both cases the center and surround zones are antagonistic; if center and surround areas are simultaneously illuminated, the cell responds only with a weak response that usually reflects the central response. These ganglion cells mainly reflect the processing that occurs in the outer plexiform layer of the retina. In cat these ganglion cells are termed X cells; in primate, they are called P cells.

Other types of ganglion cells respond with more transient responses to retinal illumination, regardless of the position of the illuminating spot in the receptive field, and they reflect more responses of inner plexiform layer neurons. These cells typically respond with more vigorous responses to moving stimuli than to static spots of light. The receptive fields of some of these cells are organized into antagonistic center and surround regions, like the sustained contrast-sensitivity cells described previously. The transient ganglion cells with a center-surround organization are termed Y

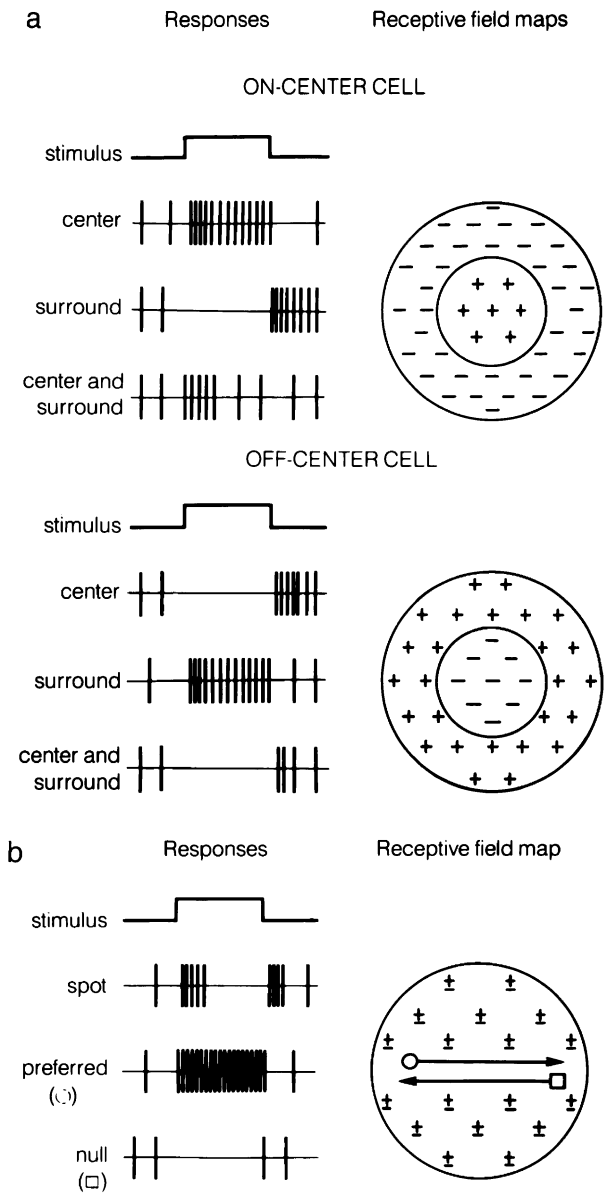


Figure 7 (a) Idealized responses and receptive field maps for on-center (top) and off-center (bottom) contrast-sensitive ganglion cells. Drawings on the left represent hypothetical responses to a spot of light presented in the center of the receptive field, in the surround of the receptor field, or in both center and surround regions of the receptive field. +, an increase in firing rate of the cell (i.e., excitation); –, a decrease in firing rate (i.e., inhibition). (b) Idealized responses and a receptive field map for a direction-sensitive ganglion cell. Such cells respond with a burst of impulses at both onset and termination of a spot of light presented anywhere in the cell's receptive field. This response is indicated by ± symbols all over the map. Movement of a spot of light through the receptive field in the preferred direction (○) elicits firing from the cell that lasts for as long as the spot is within the field. Movement of a spot of light in the opposite (null) direction (□) causes inhibition of the cell's maintained activity for as long as the spot is within the receptive field.

cells in cat and M cells in primate. Other ganglion cells give short on–off bursts of impulses to spots of light positioned anywhere in the receptive field. Some of the latter cells (particularly in nonprimate species) show direction-selective properties (Fig. 7b). Movement of a light spot in one direction vigorously excites the cell, whereas movement in the opposite direction inhibits the cell.

As noted earlier, midget ganglion cells receive input from a single midget bipolar cell, which in turn receives input from a single cone. The center of the receptive field corresponds to this direct pathway and is therefore cone specific. The antagonistic surround response, however, reflects mainly horizontal cell activity, and the horizontal cells receive input from many cones. Therefore, the center of the receptive field, is very small and has a spectral sensitivity different from that of the surround. Such ganglion cells are termed color-opponent cells and are thought to be involved in the early stages in color vision processing.

B. Synaptic Organization

Two types of chemical synaptic contacts are observed in both plexiform layers. One of these is similar in morphology to known chemical synapses seen throughout the brain, and it is termed a conventional synapse. It is characterized by an aggregation of synaptic vesicles in the presynaptic terminal clustered close to the membrane. In the retina, conventional synapses are made by horizontal, amacrine, and interplexiform cells. The other type of synapse is characterized by an electron-dense ribbon or bar in the presynaptic process and is called a ribbon synapse. The ribbon is thought to act like a conveyer belt, directing and allowing a large number of synaptic vesicles to discharge their contents into the synaptic cleft at the base of the ribbon (Fig. 8). Photoreceptor and bipolar cells make ribbon synapses in the retina. Electrical (gap) junctions are also observed in both plexiform layers of the retina, and they are believed to mediate direct electrical interactions between certain retinal neurons.

Figure 8a shows a drawing of a bipolar cell ribbon synapse and an amacrine cell conventional synapse in the inner plexiform layer. Typically, there are two postsynaptic processes at the ribbon synapses of bipolar cells, whereas at conventional synapses only one postsynaptic process is found. In this drawing, the amacrine cell synapse is made back onto the bipolar

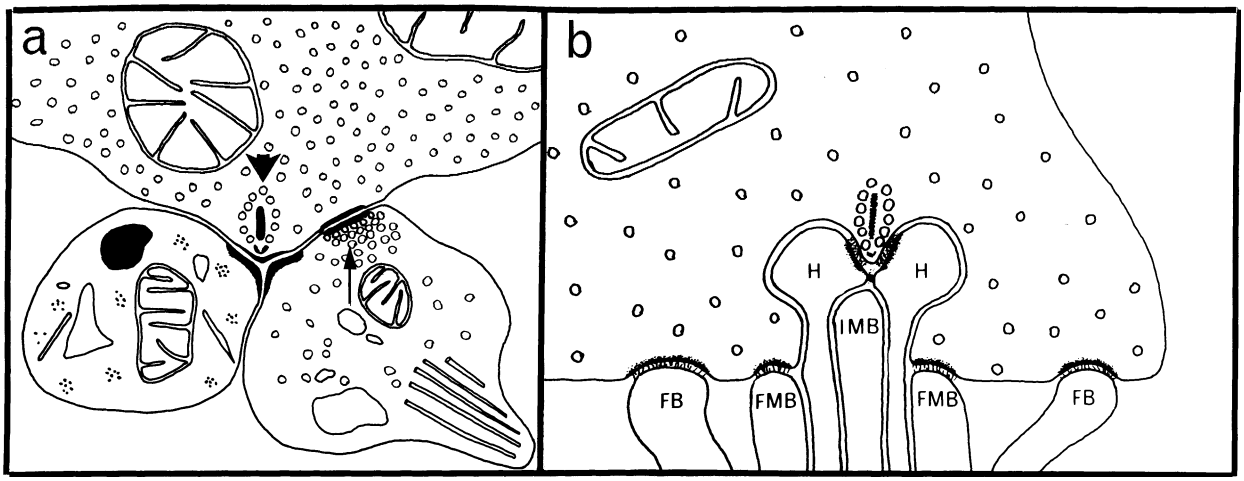


Figure 8 (a) Schematic drawing of a bipolar cell ribbon synapse (arrow) back onto the bipolar cell terminal. One postsynaptic process at the ribbon synapse is an amacrine cell process (right) and the other is a ganglion cell dendrite (left), a typical arrangement at cone bipolar cell terminals in primate. (b) Schematic drawing of synapses made by cone terminals in primates. See text for details. H, horizontal cell process; FB, flat bipolar cell dendrite; FMB, flat midget bipolar cell dendrite; IMB, invaginating bipolar cell dendrite.

terminal that is making a synapse on it. Thus, a reciprocal or feedback synaptic arrangement is suggested, which is commonly seen between bipolar cell terminals and amacrine cell processes. Amacrine cell synapses are also made on the processes and cell bodies of ganglion, interplexiform, and other amacrine cells.

Photoreceptor cell synapses are particularly complex, and a summary drawing of a part of a cone synaptic terminal is shown in Fig. 8b. The synaptic ribbons are found above invaginations of the basal surface of the terminal. Processes from horizontal cells and invaginating midget and invaginating diffuse bipolar cells penetrate into the terminal invaginations. Horizontal cell processes lie lateral to the synaptic ribbons, whereas the dendrites of the invaginating bipolar cells are centrally positioned. In addition to the ribbon synapse, the cone photoreceptor terminals make a second, unusual synaptic contact with the flat bipolar cells, which is called a flat or basal junction. No ribbon or vesicle cluster is associated with this junction, but some specializations of the membranes on both sides of the synapse are seen. The flat midget bipolar processes are typically found immediately adjacent to the invaginating midget bipolar processes, whereas the processes of the other flat bipolar cells are positioned away from the invaginations. Rod photoreceptor terminals do not make basal junctions; all rod bipolar dendrites penetrate into invaginations of the rod terminal.

Figure 9 is a simplified summary diagram of the synaptic organization of the primate retina. Each cone in the primate retina makes connections with two midget bipolar cells—one that makes invaginating-type junctions and a second that makes flat junctions (Fig. 9, left). In the central region of the primate retina, these two bipolar cell types synapse on separate midget ganglion cells. In the periphery of the primate retina, midget bipolar cells synapse on ganglion cells that receive input from a few to many midget bipolar terminals.

All cone terminals in primates also make synapses with diffuse cone bipolar cells and the proximal (dendritic) processes of horizontal cells. The axonal processes of the horizontal cells usually extend to the rod terminals. Synapses made by the horizontal cells have been observed on bipolar cells in many species but very rarely back onto the photoreceptor terminals. (However, there is good physiological evidence for feedback from horizontal cells into photoreceptors in many species.) Another interesting feature of the cone photoreceptor terminals is that they make junctions with each other and with adjacent rod terminals. Evidence in many species, including primates, suggests that these junctions are small electrical synapses.

Rod bipolar terminals in the mammalian retina do not contact ganglion cells directly (Fig. 9, right). Rather, amacrine cell processes are always postsynaptic at the ribbon synapses of the rod bipolar terminals. One of these makes a feedback synapse onto the terminal, whereas the other belongs to a special subtype of amacrine cell called the All cell, which makes both gap junctional (electrical) and

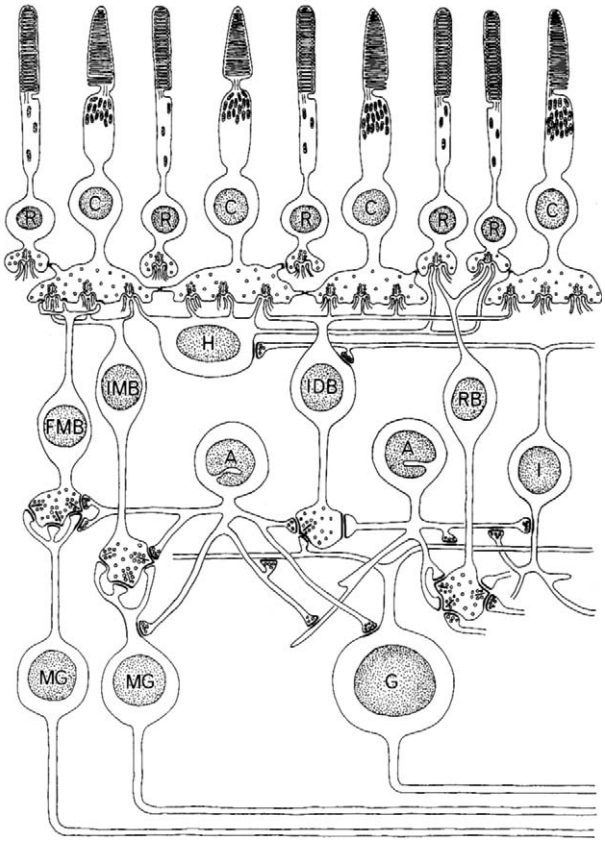


Figure 9 Summary diagram of the synaptic organization of the primate retina. See text for details. R, rods; C, cones; FMB, flat midget bipolar cells; IMB, invaginating midget bipolar cell; H, horizontal cell; IDB, invaginating diffuse bipolar cell; RB, rod bipolar cell; I, interplexiform cell; A, amacrine cell; G, parasol ganglion cell; MG, midget ganglion cells.

conventional (chemical) synapses with cone bipolar terminals. These cone bipolar terminals then contact the ganglion cells. Thus, rod information in the mammalian retina usually passes through an amacrine cell before it is transmitted to the ganglion cells via the cone bipolar cells. Why this is so is not clear: It has been suggested that the amacrine cell serves to amplify and quicken the rod signal.

Finally, interplexiform cells receive their input from amacrine cells, and they make some synapses in the inner plexiform layer on amacrine and ganglion cell processes. Most of their synapses, however, are made in the outer plexiform layer on bipolar and horizontal cells.

It should be noted that Fig. 9 is highly simplified. It is unlikely that any one amacrine cell makes the variety of contacts shown for either of the amacrine cells drawn in the figure.

IV. NEURONAL RESPONSES

In many nonmammalian species, intracellular recordings can be made routinely from most of the retinal cells. In mammalian retinas, intracellular recordings are more difficult to make, and relatively few have been reported from the primate retina. The following discussion is based mainly on recordings from nonmammalian species, but the recordings made so far from mammalian retinal neurons are similar.

The distal retinal neurons—receptors, horizontal cells, and bipolar cells—respond to light with sustained, graded membrane potential changes (Fig. 10). Unlike most neurons found elsewhere in the brain, they do not generally generate action potentials. This may be the case because these neurons have relatively short processes, and they do not need to transmit information over long distances; in other words, passive spread of potential along the cell membrane is sufficient to transmit information from one end of the cell to the other. The distal retinal cells may also function with graded potentials because such potentials are capable of discriminating a wider range of signals than can all-or-none events (i.e., action potentials).

Another unusual feature of the distal retinal neurons is that most respond to light with hyperpolarizing potentials; during illumination, the cell's membrane potential becomes more negative. Elsewhere in the nervous system, hyperpolarizing potentials are usually associated with inhibition because they prevent neurons from generating action potentials. In the distal retina, the neurons do not generate action potentials; thus, hyperpolarizing potentials can reflect excitation, and the photoreceptors are a good example.

A. Photoreceptors

Both rod and cone photoreceptors only hyperpolarize in response to illumination, but rods are more sensitive to light than are cones by about 25 times in the primate and other species. The reasons photoreceptors hyperpolarize in response to light and why in darkness the membrane potential of the cell is partially depolarized were discussed earlier. A typical resting value for the membrane potential of photoreceptors in the dark is

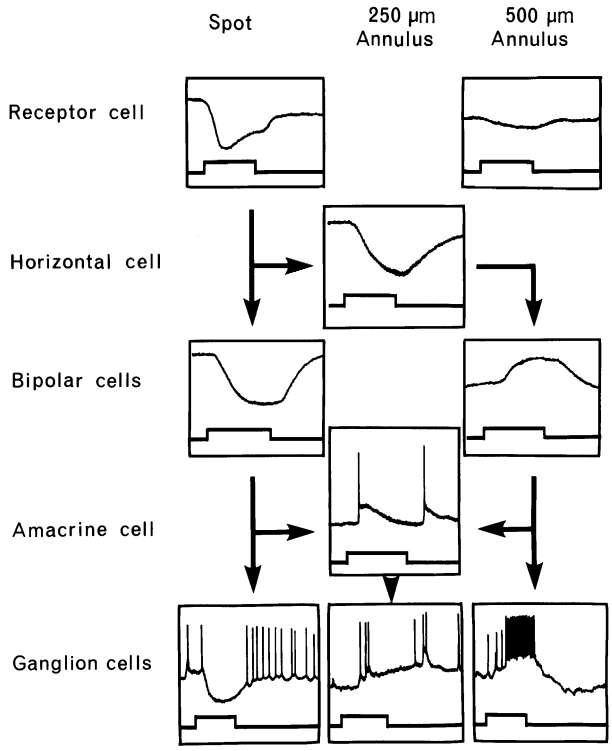


Figure 10 Intracellular responses from receptor, horizontal, bipolar, amacrine, and ganglion cells of mudpuppy retina. Distal retinal neurons (receptor, horizontal, and bipolar cells) respond to illumination with sustained graded potentials; proximal retinal neurons show both sustained and transient potentials and action potentials. Receptor, bipolar, and ganglion cells respond differently to center (spot) and surround (annular) illumination. Horizontal and amacrine cells usually respond similarly to spot and annular illumination; here, responses to a small annulus (250 μm) are shown that stimulate both the center and surround of the receptive field. The bipolar cell illustrated is a center-hyperpolarizing cell, the amacrine cell shown is a transient amacrine cell, and the ganglion cell is an off-center cell. Arrows indicate in a general way how the responses are synaptically generated (but see Fig. 11).

−30 mV, and in response to bright illumination the membrane potential is −60 mV, a typical resting potential for most neurons at rest. Thus, in the distal retina, photoreceptors and other neurons behave as though darkness is the stimulus and light turns them off (i.e., it hyperpolarizes them). It is not known why the system works is this way.

As noted earlier, there are electrical synapses between photoreceptors, but their role is not well understood. They may relate to the functioning of the photoreceptors or as a pathway for information flow. For example, they may provide an alternative pathway for rod signals to reach the cone bipolar cells. On the other hand, it has been proposed that electrical coupling between photoreceptors can reduce membrane noise and thus improve signal detection and also that it may increase the amplification of signals transmitted from the photoreceptors at their synapses. Whatever the functions of the electrical coupling between photoreceptors, it does increase the receptive field size of the cell. That is, the receptors respond to illumination over a wider area of the retina than that of a single receptor. Nevertheless, photoreceptors typically have the smallest receptive fields of any of the retinal neurons (Fig. 10).

B. Horizontal Cells

Horizontal cells, like photoreceptors, have relatively low resting membrane potentials in the dark (−30 mV), and often they only hyperpolarize in response to light (see Fig. 10). Some horizontal cells respond with small depolarizing responses to certain wavelengths of light and with hyperpolarizing responses to other wavelengths. These cells appear to be involved in color processing and are termed chromaticity cells. Horizontal cells that only hyperpolarize to illumination are called luminosity cells.

The receptive fields of horizontal cells are characteristically large, often several millimeters in diameter. Thus, the receptive field size usually exceeds the dendritic spread of the cells. Horizontal cells typically make extensive electrical junctions with each other, and the large receptive fields of these cells can be explained on the basis of the extensive electrical coupling between horizontal cells.

C. Bipolar Cells

Bipolar cells, like receptors and horizontal cells, respond to light with sustained graded potentials (see Fig. 10). In the mammalian retina, all rod bipolar cells depolarize to light. However, these are two physiological types of cone bipolar cells: those that depolarize in response to central spot illumination and those that hyperpolarize to such stimuli. Furthermore, the cone bipolar cell receptive field is organized into antagonistic zones such that illumination of the surround antagonizes the response to spot illumination. Thus, cone bipolar cells show a center-surround receptive field organization. "On" (center-depolarizing) and "off" (center hyperpolarizing) bipolar cells have been

found in all species so far examined, both mammalian and nonmammalian.

D. Amacrine Cells

In most retinas two basic types of amacrine cell responses are observed: transient and sustained. Transient amacrine cells usually give "on" and "off" depolarizing responses to illumination presented anywhere in their receptive field (see Fig. 10), but there are also transient amacrine cells that respond only at the "on" or "off" of illumination. Transient amacrine cells typically generate action potentials, and these potentials are observed on the transient "on" and "off" depolarizations. Usually only one to three action potentials are observed on the transient depolarization; thus, it has been proposed that the action potentials may serve as a local amplifying mechanism for the amacrine cell potentials rather than as the signal transmitted along the cell, as is the case for many neurons.

Sustained amacrine cell responses often resemble horizontal cell responses. They may be either hyperpolarizing or depolarizing in polarity, and the amplitudes of the two response types are comparable. Usually, sustained amacrine cells, like transient amacrine cells, give similar responses to spot illumination anywhere in their receptive field. There are also amacrine cells whose responses are a mix of transient and sustained potentials.

E. Interplexiform Cells

Only a few recordings from interplexiform cells have been reported, and they have all been from the retinas of nonmammalian species. In some cases the potentials are sustained with transient components. In other words, they are like amacrine cell responses.

F. Ganglion Cells

The receptive field properties of ganglion cells were discussed earlier, and it was pointed out that two basic types of ganglion cell responses are recorded in most retinas: sustained and transient responses. Intracellular recordings reveal the underlying potentials that give rise to these two response types. For example, Fig. 10 illustrates a sustained "off"-center, "on"-surround ganglion cell. With central spot illumination (Fig. 10, left), a sustained hyperpolarizing potential was evoked in the cell, and the cell was inhibited from discharging action potentials for the duration of the stimulus. With annular illumination (Fig. 10, right), a sustained depolarizing potential was produced, which elicited a steady discharge of action potentials from the cell for the duration of the stimulus.

Figure 10 also illustrates an "on"–"off" transient ganglion cell response (center response). With small annular illumination, the cell responded with transient depolarizing potentials at the onset and offset of the light. Each depolarization evoked a short burst of action potentials, but the durations of both the depolarizing potentials and action potential discharge were always shorter than that of the light stimulus. Such cells give similar responses to illumination anywhere in the receptive field.

G. Functional Organization of the Retina

Figure 11 shows in a simplified way how the potentials shown in Fig. 10 and the receptive fields of the retinal neurons may be produced by the synaptic interaction occurring within the retina. The figure correlates the basic connections of the retinal cells with intracellular responses recorded from an amphibian (mudpuppy) with a spot of light, small annulus, or large annulus.

As noted earlier, receptors have small receptive fields. They respond well to spots of light centered over the receptor but poorly to surrounding annuli (i.e., to light that does not directly strike the cell). The anatomy of the retina indicates that bipolar and horizontal cells are both activated by the receptors and Fig. 11 shows that both cell types respond with sustained graded potentials that resemble in waveform the photoreceptor response. Figure 11 also shows that the horizontal cells interact with bipolar cells and that this interaction is opposite in terms of receptor sign from the receptor–bipolar interaction. In Fig. 11 the receptor causes the bipolar cell to hyperpolarize, whereas the horizontal cell causes the bipolar cell to depolarize. Horizontal cells have a much larger lateral extent than do bipolar cells; thus, a center-surround receptive field organization is observed in the bipolar cell response. The center response of the bipolar cell (Fig. 11, left) is mediated by direct receptor–bipolar cell interactions, whereas the surround response (Fig. 11, right) reflects input to the bipolar cell via the horizontal cells.

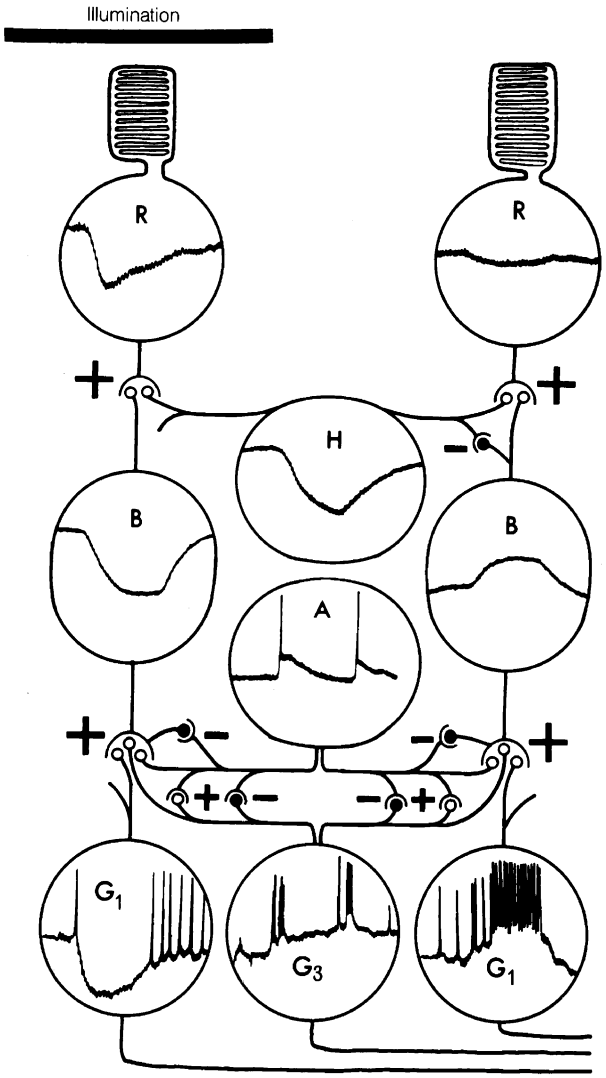


Figure 11 Summary diagram correlating the synaptic organization of the vertebrate retina with some of the intracellularly recorded responses from the mudpuppy retina. This figure shows how the receptive field organization of the hyperpolarizing bipolar cells, "off"-center ganglion cells, and "on"-"off" ganglion cells is established. The responses occurring in the various neurons upon illumination (bar) of the left receptor are indicated. The hyperpolarizing bipolar cells and "off"-center ganglion cells (G_1) respond to direct central illumination (left) by hyperpolarizing and to indirect (surround) illumination (right) by depolarizing. Note that the switch from hyperpolarizing to depolarizing potentials along the surround illumination pathway occurs at the horizontal–bipolar cell junction. The "on"-"off" ganglion cell (G_3) receives strong inhibitory input from amacrine cells; the figure indicates that these cells receive their excitatory input from both amacrine and bipolar cells. Inhibitory feedback synapses from amacrine cells onto the bipolar terminals are also indicated. R, receptors; H, horizontal cell; B, bipolar cells; A, amacrine cells; G, ganglion cells; + with open circles, excitatory synapses; − with filled circles inhibitory synapses.

How horizontal cells mediate the surround response in bipolar cells is not well understood. In most species, feedback from horizontal cells to cone photoreceptor cells has been observed, and it has been proposed that the surround antagonism is generated in this way. Feedforward inhibition of horizontal cells onto bipolar cells has also been seen in some species, and no feedback inhibition has been observed on rod photoreceptor cells. Thus, the surround response of the bipolar cells probably results mainly from feedback interactions, but feedforward interactions may contribute as well. Furthermore, the mechanism underlying the feedback or feedforward interaction is unclear. It may represent classic chemical synaptic inhibition or complex electrical interactions between the cells in the outer plexiform layer.

As noted earlier, two types of synapses made by the cone terminals have been observed: invaginating and flat junctions. It appears that in many cases (but not always) the flat synapses result in the center-hyperpolarizing responses in bipolar cells, whereas the invaginating junctions result in the center-depolarizing bipolar cell responses.

The bipolar cell terminals carry the visual signal from the outer to inner plexiform layers. Depolarizing or "on" bipolar cell terminals are found in the inner half of the plexiform layer, whereas hyperpolarizing bipolar cell terminals are in the outer half of the layer. Thus, there is a division of the inner plexiform layer into "on" and "off" laminae or strata. "On" responses of amacrine and ganglion cells are generated in the lower part of the layer, whereas "off"-responses are generated in the upper part. Thus, "on"-center ganglion cells have their processes in the lower or "on"-strata of the inner plexiform layer, "off"-center cells in the upper or "off" layer, and "on"–"off" ganglion cells spread processes in both halves of the layer. In addition to the "on" and "off" laminae, there is evidence for as many as 10–12 sublaminae in the inner plexiform layer. The cells contributing processes to these sublaminae have distinct properties with regard to such things as rod or cone dominance, response waveform, and excitatory and inhibitory input. Each sublamina thus appears to process visual information somewhat differently.

The responses of the two basic types of ganglion cells found in the mudpuppy (and other retinas) appear to relate to the responses of the input neurons to the ganglion cells. The sustained "on"- or "off"-center ganglion cells (G_1 cells in Fig. 11) appear to receive much of their synaptic input directly from the bipolar cells; their responses resemble those of the "on"- or

"off"-bipolar cells, and their receptive fields thus primarily reflect the processing that occurs in the outer plexiform layer. The "on"–"off" transient ganglion cells, on the other hand, resemble in their properties the transient amacrine cells, and they appear to receive more of their input from amacrine cells (G_3 cell in Fig. 11). These cells reflect more the processing that occurs in the inner plexiform layer. As noted earlier, the transient ganglion cells respond better to moving than to static stimuli, and some of these cells show complex receptive field properties such as direction selectivity. Substantial anatomical evidence shows that complex ganglion cell receptive field properties, such as motion and direction sensitivity, are mediated in the inner plexiform layer, predominately as a result of interactions between amacrine and ganglion cells. That is, in those species that have many motion- and direction-selective cells, there are more amacrine cell synapses per unit area of the inner plexiform layer than there are in species that have relatively few motion- and direction-selective ganglion cells.

In addition to the two basic types of ganglion cells described here, many ganglion cells have a mix of transient amacrine and bipolar cell characteristics. The Y-type ganglion cell of the cat and M-type primate ganglion cell are examples. They show a center-surround receptive organization like the bipolar cells, but their responses to both center and surround stimulation are quite transient and they are sensitive to moving stimuli. These cells appear to receive more of a mix of bipolar and amacrine cell synaptic input than the more sustained "on"-center or "off"-center ganglion cells of the cat and primates, the X and P cells.

In primates the P cells correspond to the midget ganglion cells, whereas the M cells are parasol cells. The designation P and M cells derives from the layers of the lateral geniculate nucleus (LGN) to which these cells project; P cells go to the parvocellular layers and M cells to the magnocellular layers of the LGN.

V. PHARMACOLOGY

The retina, like other regions of the brain, contains a large number of neuroactive substances. Currently, the following substances are believed to be released from retinal neurons during retinal activity:

- Amino Acids
  - γ-Aminobutyric acid (GABA)
  - L-Glutamate
  - Glycine
- Amines
  - Acetylcholine
  - Dopamine
  - Serotonin
- Peptides
  - Cholecystokinin
  - Enkephalin
  - Glucagon
  - Neurotensin
  - Neuropeptide Y
  - Somatostatin
  - Substance P
  - Vasoactive intestinal peptide
- Unconventional modulators
  - Nitric oxide
  - Retinoic acid
  - Zinc

Also, like in other brain regions, these substances may be classified into two general categories, neurotransmitters and neuromodulators, based on their mode of action. Relatively few substances appear to serve as neurotransmitters in the retina. L-Glutamate and acetylcholine mediate fast excitatory pathways in the retina, whereas GABA and glycine mediate fast inhibitory pathways. The bulk of the neuroactive substances released from retinal neurons appear to be neuromodulatory in nature, although little is known about the action or role of most of these substances, especially the peptides.

A. Amino Acids

Both photoreceptors and bipolar cells employ L-glutamate as their transmitter. L-Glutamate depolarizes both horizontal cells and "off"-center (hyperpolarizing) bipolar cells, whereas it hyperpolarizes the "on"-center (depolarizing) bipolar cells. Neurotransmitter is released when neurons are depolarized, and because photoreceptors are maintained in a depolarized state in the dark they release L-glutamate in the dark. When illuminated, photoreceptors hyperpolarize, and transmitter release is decreased. Thus, the light responses of the horizontal and bipolar cells reflect the withdrawal of transmitter from the cell. Horizontal cells are depolarized in the dark because of a dark release of L-glutamate from the photoreceptors; their light response is a hyperpolarization, reflecting the decrease in transmitter release from the photoreceptors in light. The same explanation holds for the off-center bipolar cells. These cells are depolarized in the

dark; they hyperpolarize in light as transmitter release from the photoreceptor decreases. The ''on''-center cell, however, is hyperpolarized in the dark, and it depolarizes in the light in response to the decreased release of L-glutamate from the photoreceptor.

Although the same substance, L-glutamate, mediates the photoreceptor input to horizontal cells and both types of bipolar cells, the receptor proteins with which the glutamate interacts differ in the three types of cells. ''Off''-bipolar and horizontal cells possess ionotropic glutamate receptor channels, whereas many ''on'' bipolar cells have metabotropic glutamate receptor proteins. Thus, it is possible to block the responses of one or another of the three cells with pharmacological agents while leaving the responses of the other cells intact. One substance, 2-amino-4-phosphonobutyric acid (APB), blocks specifically the metabotropic receptors of the ''on''-center bipolar cells, and this results in the loss of ''on'' activity throughout the visual system in many species. Monkeys treated with APB are unable to distinguish increases of illumination, although they can discriminate decreases of retinal illumination.

The amino acids GABA and glycine serve as the main inhibitory neurotransmitter agents in both the inner and outer plexiform layers. Many horizontal cells contain GABA, and approximately 80% of the amacrine cells in many retinas contain either GABA or glycine, equally divided. Both GABA and glycine powerfully inhibit ganglion cells by opening channels in the cell membrane that cause hyperpolarization of the cell and inhibition of action potential generation. GABA and glycine act on many retinal neurons, mediating specific inhibitory effects. For example, by blocking the effects of GABA in a retina with pharmacological agents, direction-sensitive ganglion cells lose their directional selectivity. Such cells then respond to spots of light moving in any direction across the retina. Therefore, it is believed, that direction sensitivity is mediated by GABAergic amacrine cells in the inner plexiform layer.

B. Amines and Peptides

Most amines and neuropeptides appear to function as neuromodulatory agents in the retina, although little is known about the function of most of these agents, particularly the neuropeptides. The exception is acetylcholine, which functions in the retina as an excitatory neurotransmitter in the inner plexiform layer. Acetylcholine is found in a specific type of amacrine cell, the starburst cell, and half of these cells spread processes in the ''on'' region of the inner plexiform layer and respond with transient depolarizing responses at the onset of illumination. The other half extend processes in the ''off'' region of the inner plexiform layer and respond with transient depolarizing responses at the offset of illumination. The cell bodies of the on acetylcholine-containing amacrine cells are found among the ganglion cells, whereas the cell bodies of the ''off'' acetylcholine-containing amacrines are in the inner nuclear layer. The acetylcholine-releasing amacrine cells are believed to provide excitatory input to the transient ''on''–''off'' ganglion cells, especially those that are direction selective. They also provide input to the parasol (M-type) ganglion cells in the primate and may play an important role in the development of the retina. Interestingly, the starburst amacrine cells also contain GABA, and both acetylcholine and GABA are released by this neuron. The significance of a neuron releasing simultaneously an excitatory and inhibitory neurotransmitter is not understood.

The other amines, dopamine and serotonin, and the neuropeptides have also been localized principally to amacrine cells. For the most part, these cells are relatively scarce, accounting for no more than a few percent of the total number of amacrine cells. These cells usually spread their processes widely in the inner plexiform layer; thus, they are capable of exerting wide modulatory effects. It is also the case that coexistence of neuroactive substances occurs in many amacrine cells. It has been reported that two peptides can coexist in the same amacrine cell, that a peptide and a monoamine coexist, or that a peptide or monoamine and an inhibitory amino acid (usually GABA) are in the same cell. Some evidence has been provided that three or even more agents may be colocalized in the same neuron. Currently, the significance of the colocalization of two or more neuroactive agents in a single retinal neuron is not understood, although an obvious hypothesis is that it permits multiple effects to be exerted on postsynaptic neurons.

1. Interplexiform Cells and Dopamine

Much of what we know about the action of neuromodulators in the retina has come from the study of dopamine in cold-blooded vertebrate retinas, especially fish and amphibian retinas. In the teleost, dopamine is present in interplexiform cells and so these studies have also shed light on the role of these cells in retinal

function. A brief summary of the findings is presented as a model for the action of neuromodulators in the retina.

In teleosts, the synaptic output of the interplexiform cells is mainly on the cone-related horizontal cells. Two effects of dopamine on these cells have been observed; a loss of light responsiveness (i.e., light responses are reduced in amplitude after dopamine application to the retina) and a decrease in electrical coupling between horizontal cells. Dopamine does not exert these effects by acting directly on horizontal cell membrane channels; rather, it interacts with a membrane receptor protein linked to the enzyme adenylate cyclase that catalyzes the formation of cyclic AMP (Fig. 12). A specific kinase (PKA) activated by cyclic AMP in horizontal cells phosphorylates both the glutamate channels (i.e., the channels activated by the photoreceptors transmitter) and the gap junctional channels. The phosphorylation of these channels serves to alter their properties. In the case of the gap junctional channels, phosphorylation mainly decreases the time that the channels remain open, thereby decreasing the flow of current that passes across the junction. Phosphorylation of the glutamate channels modifies the frequency of opening of the channels, which again modulates ion flow across the membrane.

The action of dopamine on teleost horizontal cells is therefore not to initiate activity but to modulate chemical synapses made onto the cell and the electrical synapses made between horizontal cells. Similar effects of dopamine on chemical and electrical synapses have been seen in a variety of animals, including turtles, toads, tiger salamanders, and rabbits. Dopamine modulates gap junctions between horizontal cells in turtle, between rod–cone photoreceptors in the toad, and between amacrine cells in rabbit. In tiger salamander, dopamine modulates glutamatergic chemical synapses onto bipolar cells, whereas in the toad, *Xenopus*, dopamine alters the rod–cone balance of input to the horizontal cells presumably through a similar mechanism. Dopamine modulation of GABAergic synapses in the retina has also been reported. Dopamine also has a variety of other actions in the retina, including the control of retinomotor movements of the photoreceptors in fish and the regulation of Ca^{2+} currents in turtle ganglion cells.

One major effect of dopamine in the teleost retina is to decrease the effectiveness of the horizontal cells in mediating lateral inhibitory effects in the outer plexiform layer. Decreasing light responsiveness of the cell and shrinking its receptive field size are effective ways of lessening its influence. As noted earlier, horizontal cells form the antagonistic surround response of bipolar cells, and thus a decrease in bipolar cell surround response is observed after dopamine application to the retina.

What might be the significance of the modulation of lateral inhibition and surround antagonism by

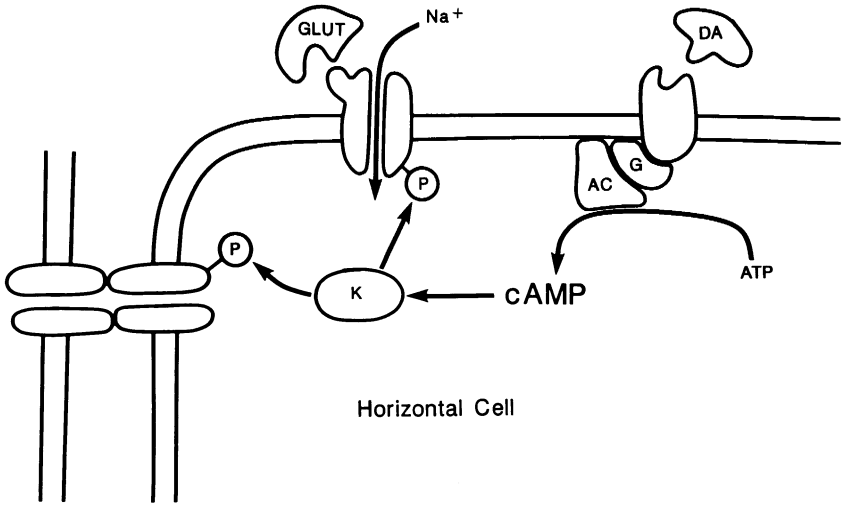


Figure 12 Summary scheme showing how dopamine (DA), acting via cyclic AMP, may influence responsiveness of horizontal cells to L-glutamate (the photoreceptor transmitter) and electrical coupling between horizontal cells. DA interacts with receptors that are linked to the enzyme adenylate cyclase (AC) via a G protein. Activation of AC results in conversion of ATP to cyclic AMP. Cyclic AMP interacts with a kinase (K) that phosphorylates (P) glutamate (Glut) channels or the gap junction channels.

dopamine and the interplexiform cells in the retina? It has long been known that after prolonged periods of time in the dark the antagonistic surround responses of ganglion cells are reduced in strength or even eliminated. An obvious speculation is that interplexiform cells and dopamine play such a role and regulate the strength of lateral inhibition and center-surround antagonism in the retina is a function of adaptive state. In fish, there is evidence supporting this view. After periods of prolonged darkness, horizontal cell receptive field size and light responsiveness are substantially decreased.

C. Unconventional Modulators

Three unusual substances have been shown to modulate retinal neurons, and they appear to play roles as neuromodulators, especially in the outer retina. Nitric oxide (NO) is a gas that has the interesting property of being able to diffuse from cell to cell. NO activates the enzyme guanylate cyclase, which produces the second messenger cyclic GMP. Cyclic GMP, in turn, activates the kinase PKG. NO, through PKG, has similar effects as dopamine on horizontal cell coupling, but it does so via a different mechanism. Whereas dopamine reduces coupling mainly by decreasing gap junction channel open time, NO reduces coupling by reducing channel opening frequency, suggesting that PKC interacts with a different site on the gap junctional channel than does PKA. NO, like dopamine, also modulates glutamate receptors on horizontal cells, but its effects are complicated. It decreases the affinity of the receptors for glutamate, but it increases the maximal current that can pass through the glutamate receptors on horizontal cells.

The enzymes that produce NO are found in many cells throughout the retina, including several types of amacrine cells; thus, it is believed that NO exerts a variety of effects in the retina in addition to those shown on horizontal cells.

Retinoic acid (vitamin A acid) is usually considered to be a morphogen, playing an important role in the development of many tissues including the retina. For example, retinoic acid is important for ventral retinal development and it is also involved in photoreceptor maturation. Retinoic acid is known to regulate a number of genes, which it does by interacting with specific retinoic acid receptors known as RARs or RXRs. On horizontal cells, retinoic acid, like dopamine and NO, uncouples horizontal cells. It appears to act directly on the gap junctional channels and not via second messengers or kinases. Retinoic acid levels in the retina are modified by light, suggesting that retinoic acid is a third substance involved in the regulation of horizontal cell gap junctions. Whether retinoic acid plays other neuromodulatory roles in the retina is unknown.

Finally, zinc, a trace metal, has been shown to be colocalized with glutamate in the photoreceptor synaptic terminals, and it is thought to be released with glutamate from the photoreceptor terminals. Zinc has been found to modulate both glutamate and GABA receptors in the retina. In both cases, zinc decreases the membrane current flows elicited by glutamate or GABA. Again, it is possible that zinc plays a role at other glutamatergic and GABAergic synapses in the retina.

VI. CONCLUSIONS

The retina is a relatively well-understood part of the central nervous system, but its complexity is overwhelming. Although we have long recognized that it consists of five major classes of neurons, it is now apparent that each class of neuron consists of many cell subtypes, perhaps as many as 30–40 in the case of amacrine cells. Furthermore, although we can classify the responses of most retinal neurons as ''on'' cells, ''off'' cells, or ''on''–''off'' cells, and as giving principally sustained or transient responses, it is now recognized that a variety of subtle image representations are sent to the brain via the optic nerve and that this variety reflects complex synaptic excitatory and inhibitory synaptic interactions occurring among various neuronal subtypes at different levels of the retina, particularly in different laminae in the inner plexiform layer. Finally, although relatively few substances appear to serve as the principal excitatory and inhibitory neurotransmitters in the retina, a much larger number of neuromodulatory substances are known to exist in the retina, including several unconventional substances that have been shown to have significant effects on the retinal cells. Thus, the responses of individual retinal neurons reflect not only the direct excitatory and inhibitory synaptic interactions between neurons but also an extensive modulation of these synaptic interactions by multiple substances acting in different ways on the neurons and their channels.

In this article, I presented a broad introduction to the retina and how it functions. I hinted at the complexity of interactions occurring in the retina,

but I hardly did them justice. An example is the multiplicity of channels and receptor proteins that any one neurotransmitter or neuromodulator can activate, such as the three types of GABA receptors—$GABA_A$, $GABA_B$, and $GABA_C$—thought to be key in shaping the responses of inner retinal neurons.

See Also the Following Articles

COLOR VISION • EYE MOVEMENTS • INFORMATION PROCESSING • RECEPTIVE FIELD • VISION: BRAIN MECHANISMS • VISUAL DISORDERS • VISUAL SYSTEM DEVELOPMENT AND NEURAL ACTIVITY

Suggested Reading

Baylor, D. A. (1987). Photoreceptor signals and vision. *Invest. Ophthalmol. Vis. Sci.* **28,** 34–49.

Boycott, B. B., and Dowling, J. E. (1969). Organization of the primate retina: Light microscopy. *Philos. Trans. R. Soc. London Ser. B* **255,** 109–184.

Djamgoz, M. B. A., Archer, S. N., and Vallerga, S. (Eds.). (1995). *Neurobiology and Clinical Aspects of the Outer Retina*, Chapman & Hall, London.

Dowling, J. E. (1987). *The Retina: An Approachable Part of the Brain*. Harvard Univ. Press, Cambridge, MA.

Ehinger, B., and Dowling, J. E. (1987). Retinal neurocircuitry and transmission. In *Handbook of Chemical Neuroanatomy, Vol. 5: Integrated System of the CNS Part I* (A. Bjorklund, T. Hokfelt, and L. W. Swanson, Eds.), pp. 389–446. Elsevier, Amsterdam.

Gouras, P. (Ed.) (1991). *The Perception of Colour: Vision and Visual Dysfunction*. Macmillan, London.

Kolb, H. (1994). The architecture of functional neural circuits in the vertebrate retina. *Invest. Ophthalmol. Vis. Sci.* **35,** 2385–2404.

Kolb, H., Ripps, H., and Wu, S. (Eds.) (2001). *Concepts and Challenges in Retinal Biology*. Elsevier, Amsterdam.

Polyak, S. L. (1941). *The Retina*. Univ. of Chicago Press, Chicago.

Rodieck, R. W. (1998). *The First Steps in Seeing*. Sinauer, Sunderland, MA.

Stryer, L. (1986). Cyclic GMP cascade of vision. *Annu. Rev. Neurosci.* **9,** 87–119.

Wässle, H., and Boycott, B. B. (1991). Functional architecture of the mammalian retina. *Physiol. Rev.* **71,** 447–480.

Werblin, F. S., and Dowling, J. E. (1969). Organization of the retina of the mudpuppy, *Necturus maculosus*. II. Intracellular recording, *J. Neurophysiol.* **32,** 339–355.

Salience

SHIH-CHENG YEN
Montana State University at Bozeman

LEIF H. FINKEL
University of Pennsylvania

GLOSSARY

contour integration The process of grouping of edge fragments into an extended contour.

distractor Items that are presented together with the target in a search task that share some but not all of the features of the target.

outliers Observations that are significantly different from the majority of the observations.

pop-out The degree to which a target can be detected in brief presentations.

preattentive Operations that are rapidly accomplished without apparent conscious effort.

visual search Psychophysical task of finding a target embedded in an array of distractors.

Salience refers to the perceptual prominence of an object relative to its background. The word "salience" derives from roots connoting an assault or sally, in this case upon the senses, and it implies a quality of leaping or springing forth from the stimulus. The salience of a visual target is determined by the degree to which the features of the target differ from those of the surroundings. However, it is not completely understood which "features" of the stimulus are relevant for the visual system. Measurement of salience requires a quantitative scale for assessing differences in features, sometimes across several dimensions. A red circle in a field of green squares is nominally salient, but is it more or less salient than a fast-moving dot among slow-moving dots? From an ecological perspective, animals must be rapid and reliable at detecting certain types of targets—an object directed at one's head is salient regardless of its texture. Essential targets—food, predators—must induce rapidly discriminable patterns of neural activation. An understanding of salience thus provides a glimpse into how the brain prioritizes the world. The salience of a target can be evaluated by preattentive processes, or operations that are carried out rapidly without apparent conscious effort, and is often measured by "pop-out," the degree to which the target can be detected in brief presentations. However, salience is also closely related to attentional processes; for example, search targets can be found more easily if subjects are provided with instructions relevant to the task. The mechanisms by which salience is generated in the nervous system thus involve the integration of both bottom-up and top-down influences.

I. SALIENCE, POP-OUT, AND VISUAL SEARCH

Although salience is a property of stimuli in all sensory modalities, previous studies have focused primarily on

vision. Salience is often measured by using a two-interval, forced-choice paradigm. Two stimuli are briefly presented in succession, separated by a short interval; one stimulus contains the target, and the other contains a generically similar stimulus, but lacking a target. Over multiple trials with different stimuli, the results of these trials generate a psychometric function in which the accuracy of detecting the target is plotted versus changes in stimulus parameters (e.g., contrast, size, spacing). The detection accuracy reflects the salience of the target. Brief presentations (<100 msec) restrict visual processing to preattentive vision, and targets detectable in such conditions are said to "pop out."

Alternatively, the salience of a target can be measured by recording the time required to detect the target in a visual search task. In general, the time required to detect a target increases with the number of distractor elements. However, the slope of this function—reaction time (RT) versus number of distractors—depends upon the salience of the target relative to that of the distractors. Highly salient targets pop out and can be found effectively instantly, regardless of the number of distractors. When the target has less salience, detection requires a serial, attention-based scan of putative target locations.

Interestingly, in many cases, there is an asymmetry in salience when the roles of target and distractors are exchanged. For example (as shown in Fig. 1), a long line is salient when placed amid shorter lines, but a short line is less salient among longer lines. A tilted line is salient among vertical lines, but a vertical line pops out to a lesser degree from a field of tilted lines. A circle with a gap pops out from a field of closed circles, but less so the converse. On the basis of dozens of such examples, Anne Treisman concluded that salience resembles a signal detection problem and must depend upon a signal-to-noise ratio. The degree to which the distractor elements activate the relevant feature detectors defines the noise level. If the relevant neural feature detectors are strongly activated by the distractor elements, then a target that less strongly activates the neural detectors will be below the noise level. The less strongly distractors activate the neural detectors in comparison to the target, the more salient the target will appear.

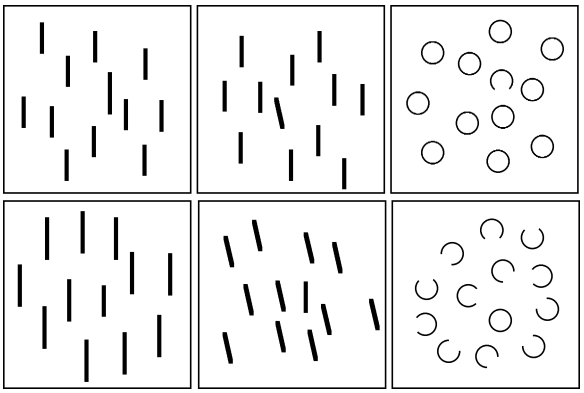


Figure 1 Asymmetries in salience. Left column: A long line is salient when placed amid shorter lines (top), but a short line is less salient among longer lines (bottom). Middle column: A tilted line is salient among vertical lines (top), but a vertical line is less salient among tilted lines (bottom). Right column: A circle with a gap is salient among closed circles (top), but a closed circle is less salient among circles with a gap (bottom). Adapted from Treisman and Gormican (1988).

Treisman used this approach to attempt to identify what features (e.g., closure, line intersections) are actually employed by the early visual system and to substantiate her Feature Integration theory of visual function. According to this theory, visual features are extracted in multiple parallel maps, but the *locations* of these features are independently coded in separate maps. Attention is posited to be necessary to bind features with their spatial location. Thus, visual search can operate in a scanning mode, where attention is serially directed to each element, or search can involve attending to larger areas containing several elements and measuring the total amount of feature activation within each area. If the wider area contains the target and several distractors, then the amount of feature activation differs from that measured in areas of the image containing only distractors. Treisman's point is that signal-to-noise constraints make it easier to detect a large target signal on top of small distractor signals than to detect a small target signal mixed in with a number of large distractor signals. Search asymmetries would then follow from Weber's law: the minimal detectable increase in signal energy over the noise level is proportional to the noise level.

Search asymmetries provide clues to the feature detectors used by the visual system because the signal-to-noise ratio will depend on which particular feature detectors are used. However, we have to be careful about making the distinction between features of a stimulus and the postulated feature detectors in the visual system. Because the visual system has multiple feature maps, target activation may be above the noise level in some maps, but below the noise level in other maps.

As an example, it has been found psychophysically that a field of L's surrounded by a field of $+$'s is nearly twice as salient as a field of $+$'s embedded in a field of

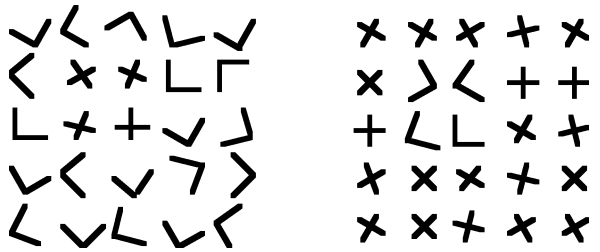


Figure 2 Asymmetry in texture pop-out. A field of L's within a larger field of +'s (shown on right) is more salient than the converse (shown on left).

L's (see Fig. 2). These texture elements are similar: both contain horizontal and vertical segments; the L contains two line endings and a single corner junction, whereas the + contains four line ends, four corner junctions, and a line intersection. According to Bela Julesz, these elements differ in the number and kind of "textons" and should be discriminable—as indeed they are. However, Treisman argued that the pop-out is due to global aspects of the elements, namely, despite the fact that they are composed of identical lines, the L's appear bigger than the +'s. Treisman carried out trials using portions of the elements (e.g., just one quadrant of the +) and found that none of these component elements pop out. In fact, she concluded that no features based on line arrangement (angles, intersections, etc.) are detected preattentively.

Interestingly, Jitendra Malik and Pietro Perona measured the response to these two texture elements by using an array of spatial filters at different spatial frequencies and orientations, resembling those in the visual pathway. The filters included circularly symmetric on- and off-center filters (corresponding to the receptive fields of retinal ganglion cells or cells in the lateral geniculate nucleus) and oriented difference-of-offset Gaussian (DOOG) filters (corresponding to the receptive fields of striate cortical simple cells). They found that most filters showed differential responses to the two elements; however, the maximum difference in response arose from the circularly symmetric filters, which responded more vigorously to the +'s. Similar results were reported by Edward Adelson and James Bergen. Note that these reported responses are the spatially averaged responses of a bank of non-interconnected filters centered at different positions with respect to the stimulus. If one considers the response of individual oriented DOOG or Gabor-type filters within a cortical network, then the L might evoke a greater peak response than the + because there is less cross-orientation inhibition. Thus, multiple neural dimensions appear to conflict regarding which stimulus is more salient. The question is whether we can identify the rules by which the visual system integrates this information.

II. SALIENT TARGETS AS STATISTICAL OUTLIERS

In our initial definition of salience, we observed that an object has salience only in comparison to a population of other objects. A 6′7″ tall basketball player stands out in his college class but is less salient on an NBA team. Weber's law implies that the degree of salience depends upon the distribution of feature properties. In some sense then, salience determination is similar to testing for outliers in a statistical distribution. The distribution of feature values over the distractor population provides a basis for determining whether the feature values of the target are statistical outliers. Ruth Rosenholtz has elegantly formulated this idea in terms of a statistical parametric test for outliers. Given a one-dimensional normal (Gaussian) distribution, a target is an outlier if its feature value lies several standard deviations away from the mean. To determine an outlier in multiple feature dimensions, Rosenholtz developed the use of the Mahalanobis distance between the target and the mean of the distractor distribution.[1]

Rosenholtz's approach is powerful because the Mahalanobis distance can provide a *quantitative* measure of salience and target discriminability. With a few additional assumptions, it can be extended to allow comparisons across feature dimensions. For example, how does the salience of a red target among yellow distractors compare to the salience of a 45° line among vertical lines? The use of a parametric test is not meant to imply that the underlying feature distributions must be Gaussian or are known *a priori* by the visual system. Rather, Rosenholtz suggests that a parametric test is simply an efficient means of rapidly determining whether a point is an outlier. This efficiency might come into play in brief stimulus

[1] Δ is defined as $\Delta^2 = (\mathbf{v}-\mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{v}-\mu)$, where $\mathbf{v}$ is a vector containing the feature values for the target in each dimension, and μ is the mean and Σ the covariance of these feature values across the population. In one dimension, the covariance matrix Σ reduces to the standard deviation, and so Δ measures the distance in terms of standard deviations between the target's feature values and the mean of these values across the population. T signifies the transpose.

presentations (<100 msec) in which the preattentive visual system may only have time to determine general properties of the object population, such as its mean and covariance.

Rosenholtz has used the Mahalanobis metric to explain asymmetries in search difficulty reported in experiments involving target and distractor motion. For example, the Mahalanobis metric correctly predicts that it is rather easy to detect a single moving target in a field of stationary distractors, but that it is more difficult to detect a stationary target in a field of moving distractors. Similarly, search for a fast target among slow distractors is more efficient than search for a slow target among fast distractors. Treisman's explanation based on Weber's law can qualitatively account for these findings, but the Mahalanobis metric quantifies the difference in difficulty. For example, in the search for fast versus slow elements, the stimuli consist of objects oscillating back and forth horizontally, with the target and distractors differing only in speed. A small amount of noise is added to the velocity profile of each object. As shown in Fig. 3 the distribution of velocities is thus centered on two values for each class: v_t and $-v_t$ for the target versus v_d and $-v_d$ for the distractors. When the target is fast and the distractors are slow, $v_t > v_d$, the target velocity lays outside the 1σ covariance ellipse (one standard deviation from the mean) of the velocities of the distractors. However, when the target is slow compared to the distractors, its velocity is within 1σ of the mean of the distractor distribution. The Mahalanobis measure clearly shows that only in the former case is the target an outlier compared to the distractor distribution.

Rosenholtz demonstrates that this approach can account for all of the well-known cases of asymmetries in motion search. She argues that the method holds as well for salience in the color dimension. In one dimension the Mahalanobis metric is similar to the Weber law approach. But when targets and distractors differ in multiple dimensions, intuition is less clear. For example, Rosenholtz considers the case in which targets and distractors can move with varying speeds in different directions. Psychophysics shows that it is easy to find a moving target among distractors with varying speeds if the distractors all move in the same direction as the target. But it is difficult if the distractors have both varying speeds and varying directions. On the other hand, if the distractors move in different directions, it is equally difficult to spot the target if the distractors all have the same speed or each has different speeds. The Mahalanobis model predicts the observed result in all cases.

III. FEATURAL AND CONFIGURAL SALIENCE

Treisman's Feature Integration theory was originally developed to explain a set of findings using stimuli in which target and distractors differ only in the *conjunction* of features. For example, the target might be a horizontal blue line within a field of horizontal green

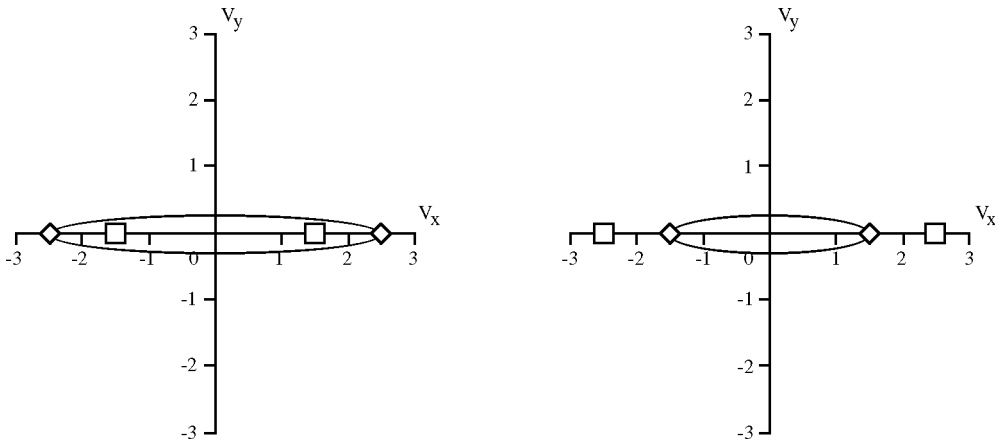


Figure 3 Search asymmetry and the Mahalanobis measure. The two graphs plot the x and y component vectors of the target (plotted with squares) and the distractors (plotted with diamonds). The ellipse delimits one standard deviation from the mean of the distractor velocities. The graph on the left illustrates search for a slow target among fast distractors, and the Mahalanobis measure correctly predicts a difficult search task. The graph on the right illustrates the search for a fast target among slow distractors, and the Mahalanobis measure again correctly predicts an easier search task. Adapted from Rosenholtz (1999).

lines and vertical blue lines. Such tasks are difficult; they generate reaction time plots whose slopes increase sharply with increasing number of distractors. On the basis of such tasks, Jeremy Wolfe has proposed models for attention-based search that incorporate more sophisticated grouping and segmenting mechanisms than the simple pooling used by Treisman. In the preceding example, all horizontal lines could be grouped and segmented from lines of other orientations. Similarly, all blue lines could be grouped and a selected search could be directed within these maps, making target detection more efficient.

Theories of attention have focused on binding features across dimensions (orientation, color, line endings, etc); however, binding also occurs across spatial locations. As Gaetano Kanizsa and others have shown, preattentive vision carries out an array of grouping and segmentation processes that organize the scene. In Fig. 4, we perceive alternating horizontal rows of open and closed circles. Despite the myriad other possible ways these elements could be grouped, we cannot consciously desist from organizing the elements into groups on the basis of similarity and proximity. This illustrates the central problem of Gestalt psychology: what object characteristics determine figure versus ground. The Gestaltists identified several characteristics that cause elements to be grouped together: similarity, proximity, good continuation, closure, symmetry, and common motion. Elements that share these characteristics tend to be grouped together, segmented from other elements, and tend to be perceived as more salient because the difference between figure and ground *is* salience. Camouflage attempts to manipulate these Gestalt characteristics so that the parts of the object are bound to the background rather than with each other thus decreasing object salience and pop-out.

Thus, in some stimulus domains, objects are salient based on the *relations* between stimulus elements

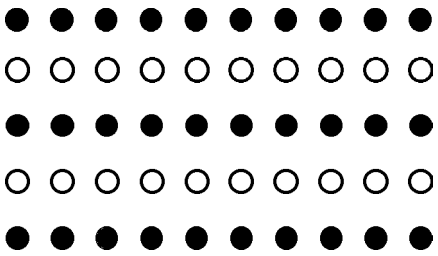


Figure 4 Perceptual grouping. The predominant percept is to group the array of dots into alternating rows of black and white dots. According to the Gestalt psychologists, this is an example of grouping using the similarity cue.

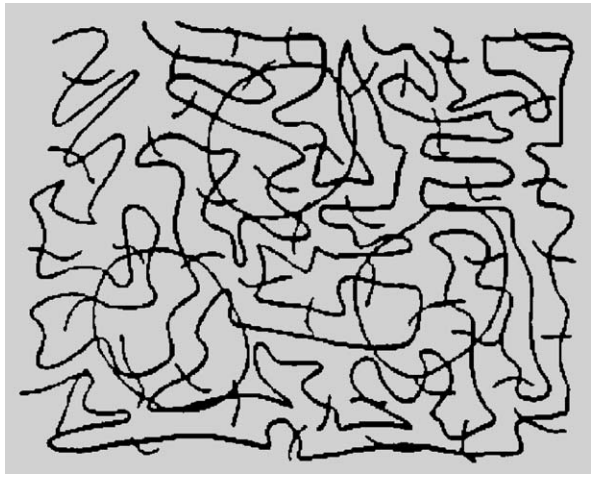


Figure 5 Salient contours. The three circular contours pop out despite the fact that there is nothing locally salient about any part of the contour as compared to the contours that make up the background. From Ullman and Sha'ashua (1988).

rather than on the features themselves. The prime example of this configural salience is contour detection and the issue of which contours in an image are most salient. Contour detection also offers the most intensively examined paradigm for understanding the link between neural and psychophysical mechanisms.

IV. CONTOUR SALIENCE

The study of contour salience was established by the work of Shimon Ullman and his colleagues in the late 1980s. Ullman was the first to articulate the global or configural nature of contour salience. Fig. 5 from Ullman's original paper, shows an image consisting of various contours that are roughly equal in contrast, spatial frequency, and orientation. Why do the three circular contours pop out? The Gestaltists would argue that the circles are most salient because they have greater continuity (longer extents with less abrupt changes in curvature) and because they are closed. Ullman attempted to quantify this Gestalt explanation by developing a simple but rigorous model, consistent with known cortical mechanisms but also computationally efficient and analyzable, that could be tested on a variety of images to evaluate its performance relative to human psychophysics.

Ullman proposed that the salience of a contour is due to three main characteristics: overall length, degree of continuity, number and size of any gaps. He proposed a weighting function to combine these contour features into an overall salience measure.

The salience measure increases monotonically with length of contour, decreases monotonically with total squared curvature of contour, and decreases with amount of fragmentation. Salience is measured iteratively at each pixel in an image, and a salience map is constructed—an image where the intensity of each pixel corresponds to the salience score of the most salient curve emanating from that pixel. Once computation of the salience map stabilizes, the most salient contours are then picked out by tracing the contours starting from the most salient points.

As shown in Fig. 6, the contours judged most salient by the algorithm correspond, at least qualitatively, with intuitive perceptual judgment. In this case, the most salient contours in an image of an automobile are produced by the wheels, the roof, the hood, etc. The fact that one can identify the object shows that, like in a caricature, these contours contain key information about object identity. The network is capable of filling in gaps and completing curves in the presence of noise. By using this measure of salience, the network is also capable of carrying out perceptual grouping. Under attentional scrutiny, we might wish such an algorithm to provide additional contour information, for example, to allow us to recognize the make and model of the car. Ullman's algorithm ranks all contours in the image, and thus attention could determine the threshold for including additional contours in further processing.

The computational power of the Ullman salience network arises from its instantiation as a relaxation algorithm: it recasts the search for the globally most salient contour into an optimization problem over local connections. The resulting algorithm is extremely efficient. Whereas the number of contours grows exponentially with image size, the time for the algorithm to find the most salient contour only grows quadratically with image size.

In a subsequent study, Tao Alter and Ronen Basri identified two critical elements of the Ullman algorithm that underlie its efficiency: (1) extensibility—the property that the most salient curve of length N contains, as subcurves, the most salient curves of lengths $N-i$ (where $N>i>0$) pixels, and (2) geometric convergence—the choice of exponential fall-off in the effects of distant portions of the curve, which limits the range of interactions particularly on closed curves.

However, Alter and Basri found that these two properties that underlie the efficiency of the algorithm also give rise to certain shortcomings. For example, the salience measure is not completely scale-invariant. For short contours, the algorithm ranks straight lines as more salient than circles, whereas for long contours circles are ranked more salient. This is because, according to the algorithm, the salience of a line grows linearly with length, whereas the salience of a circle grows quadratically with the length of its perimeter. Thus, there is a trade-off for contours above a certain size. A second problem uncovered by Alter and Basri is that the algorithm occasionally ranks a particular contour segment as the most salient, even if it does not lie on what the algorithm ranks as the most salient contour. For example, if the end of a straight line touches a circle such that the line is tangent to the circle, the line will be ranked as more salient than the circle, even if the line is quite short and on its own would be ranked far less salient. This is because there is no mechanism for segmenting contours into separate objects; thus, osculating contours (those that intersect and whose tangents agree at that point) can appropriate salience from the other curve.

Alter and Basri also point out difficulties that arise with gaps in the contour and with discretization errors. For example, the algorithm ranks a circle with one large gap as more salient than a similar circle with several smaller gaps (where the total gap size is the

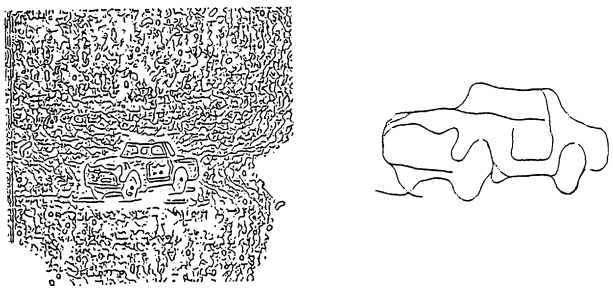


Figure 6 Result of the saliency algorithm. The figure on the left shows the input image to Ullman's saliency network. The figure on the right shows the most salient contours traced out from the points on the saliency map that have the highest values. From Ullman and Sha'ashua (1988).

same in both cases). Psychophysical studies indicate that humans perceive the opposite ranking to hold. Finally, in some cases simple aliasing in a contour, due to sampling on a x–y grid can cause an aliased straight line to be ranked less salient than an actual corner of a polyhedron.

These technical shortfalls in the algorithm arise from the simplifications adopted for computational efficiency. They do not detract from the central point that a network based on local interactions can account for global contextual effects. However, by illuminating the detailed and idiosyncratic nature of human perceptual salience, these studies suggest that the cortex employs additional mechanisms.

Can the statistical outlier idea be used to improve things? In the spirit of Ullman's approach, we might characterize a contour in terms of several dimensions: length, gap size, total curvature, etc. The target contour can be compared to all possible distractor contours in the image or, more efficiently, to the distribution predicted from knowledge of how the distractors are generated (e.g., random orientations). One would then expect that the Mahalanobis distance between a target contour and the overall distribution would correlate with the perceptual salience of the contour.

There are several impediments to developing such a measure for contour salience. Gestalt psychology provides hints as to the dimensions along which contour salience should be evaluated, but the exact measures are not known. Ullman's choices (length, curvature, etc.) are reasonable and correlate to an impressive degree with perceptual judgments. However, as Alter and Basri have shown, the correspondence is not perfect. In addition, because the dimensions are diverse in nature, e.g., length versus deviation from constant curvature, they must be normalized in order to generate a covariance matrix Σ. Finally, and most critically, the contours in such images must be actively generated by the subject through a process of perceptual grouping. Unlike the moving dots of Rosenholtz's motion stimuli, here both the target and the distractors must first be generated before their salience can be evaluated. Yet, at the same time, the salience of the contour appears to affect the grouping process.

The path out of this paradox is not totally clear, but significant insight is available by examining what is known of the psychophysics and underlying neurophysiology relating to contour detection. The quest of all of these analyses, theoretical and experimental, after all is to understand the *neural mechanisms* that mediate perceptual salience. The goal is to appreciate how the cortical machinery could carry out a process akin to distinguishing statistical distributions.

A. Psychophysical Studies of Contour Salience

Because the salience of a target depends upon context, psychophysical studies of salience have employed stimuli in which the structure of the background elements can be carefully controlled. In 1993, David Field, Anthony Hayes, and Robert Hess introduced a stimulus in which a contour is formed by a number of discontinuous, oriented elements (Gabor functions) embedded in a field of randomly oriented elements (see Fig. 8). A two-alternative, forced-choice procedure was used in which one stimulus contained a contour embedded in the random background, whereas the other stimulus contained only randomly oriented elements. Detection threshold was set at 75% accuracy for choosing which image contained the contour (presentation order was randomized). Field and colleagues found that contour salience, as measured by the detection threshold, depends on the relative position and orientation of the contour elements (all elements had the same, relatively high, contrast). The effect was scale-invariant, i.e., when the separation between all of the elements was reduced, the dependency on relative position and orientation remained unchanged.

Because each Gabor function optimally stimulates a local set of orientation columns in cortex, Field and colleagues argued that their results reflect the wiring of an underlying cortical "association field." Each cell must receive long-distance cortical connections from other cells tuned to a specific range of orientations depending upon the relative positions of the two cells. The shape of the association field closely resembles cortical connection patterns proposed in earlier computational studies of Stephen Grossberg and colleagues (see Fig. 7). In their models of boundary completion, as occurs in illusory contours and real images, Grossberg proposed a similar, bipolar-shaped connection structure.

The Gestalt law of good continuation is vague regarding what makes a particular contour better or worse. An elegant solution was put forward in 1989 by Steven Zucker and colleagues—cocircularity. Given a contour element of orientation θ_1 at position (x_1, y_1), then the optimum orientation of a contour element at position (x_2, y_2) is determined by the unique circle

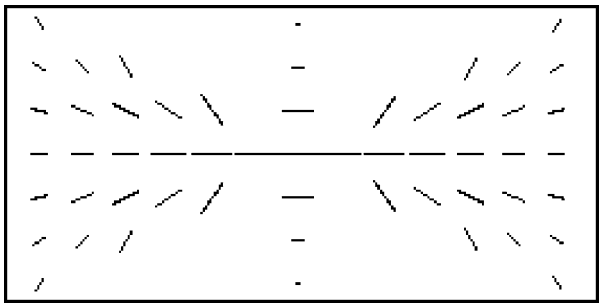


Figure 7 Presumed cortical connection pattern. Illustration of the cocircular connection pattern put forward by Zucker and colleagues. The length of the line segment indicates the strength of the connection, and the orientation of the line segment indicates the preferred orientation at each spatial location. The connection pattern is consistent with the association field suggested by Field and colleagues, as well as the bipolar connection structure proposed by Grossberg and colleagues.

passing through both points whose tangent at (x_1, y_1) has orientation θ_1 (see Fig. 7). Note that cocircularity allows the curvature to change along the course of the contour and that a straight line is cocircular (with curvature = 0). The association field conforms, at least qualitatively, to cocircularity. Moreover, cocircularity provides a metric for evaluating salience by associating a defined penalty as a function of the deviation from cocircularity.

Contour salience has also been shown to depend on the ratio of the distance between contour elements versus the distance between distractor elements. Thus, elements can be spaced over a wide range of absolute distances, and the perceptual salience remains invariant as long as the normalized distance ($d_{\text{contour}}/d_{\text{distractor}}$) remains fixed. This finding suggests the presence of a cortical normalization mechanism, akin to that controlling contrast invariance.

By using stimuli similar to those of Field and colleagues, Ilona Kovács and Bela Julesz studied another cue affecting contour salience: the Gestalt property of closure. They first established the maximum interelement spacing at which open contours and closed contours were just detectable (75% accuracy in two-interval trials). They found that the threshold spacing for closed contours was approximately twice that for open contours. Next, they plotted psychometric functions for detection accuracy versus number of elements on the contour. Starting with four elements spaced at the threshold for open contours, Δ_{Open}, and four elements spaced at the threshold for closed contours, Δ_{Closed}, additional elements were appended to both types of contours at their respective threshold separations. The surprising finding was that, whereas the psychometric curve for the open contours monotonically increased with increasing number of contour elements, the psychometric curve for the closed contours remained at chance (50% accuracy in a two-alternative choice) until the final (twelfth) element was added, at which point accuracy nonlinearly jumped to threshold (75% accuracy; see Fig. 8). Note that the "closed" contours were not closed until composed of all twelve elements, so that before the last element was added the contours were open and were effectively spaced at twice the threshold distance for open contours. Thus, it might not be too surprising that these contours were not detectable. What was unexpected, however, was the sudden jump in salience that arose with closure. We will consider possible mechanisms for this phenomenon later.

One caveat to these findings comes from more recent work by Jochen Braun. Braun has argued that the methods used to generate the Field-type stimuli introduce a small, but significant difference in the distribution of distances between elements on the contour versus the distribution of distances between distractor elements. This implies that the local context around a contour element differs from that around a distractor element, and thus their activation may differ simply on the basis of proximity cues. Braun has developed a statistical method for generating contour displays that controls contour–background element densities. By using these stimuli, Braun finds only a small effect of closure (6–14%) on contour salience. In addition, Braun suggests that the stimuli of Kovács and Julesz confound closed versus open contours with differences in eccentricity (open contours extended further) and a difference in element number. He reports a large nonlinear increase in salience as the number of contour elements increases in the range of 10–12. Thus, the change in the number of elements may have contributed to the change in salience, in addition to the difference between closed and open contours.

Braun also determined the stimulus offset asynchrony (SOA) (time between stimulus presentation and a subsequent masking stimulus) to detect stimuli with elements spaced at various distances. For very salient stimuli, the SOA was approximately 100 msec, suggesting that the visual system only needs 100 msec to detect the contour. However, as the contours became less salient and approached perceptual threshold, the SOA increased nonlinearly, with a value of 250 msec near threshold. Braun suggests that this timing

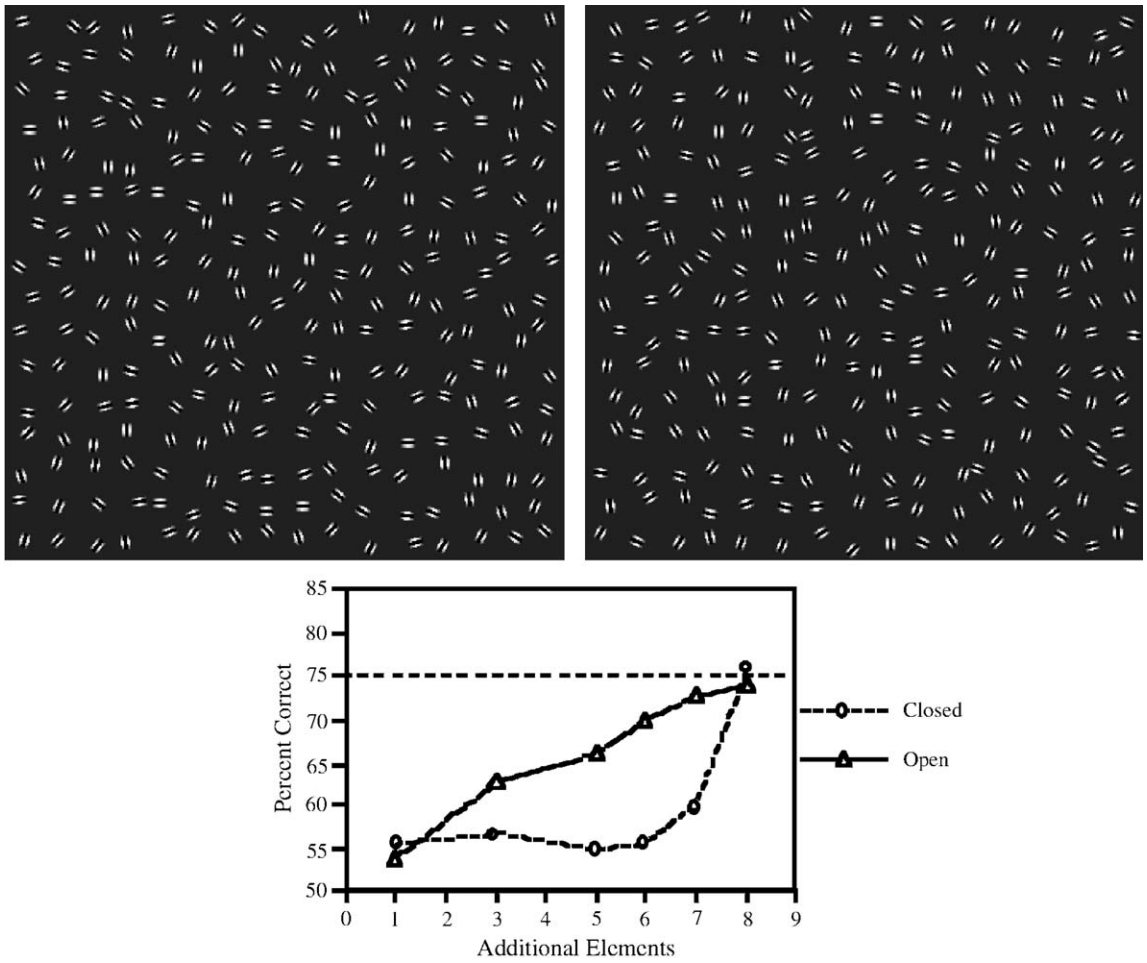


Figure 8 Psychophysical effect of closure. The figures on the top show examples of a stimulus with an open contour (left) and one with a closed contour (right), embedded in an array of randomly oriented elements. The bottom graph, adapted from Kovács and Julesz (1993), shows the performance of subjects as a function of the number of elements on the contour. The points plotted with triangles show that the salience of the contour with elements spaced at Δ_{Open} increased monotonically with each additional element. The points plotted with circles show that the salience of the contour with elements spaced at Δ_{Closed} did not increase with each additional element until the last element was added and the contour was closed.

information places a constraint on computational models of salience. He makes the argument that, if information must be transmitted sequentially between neurons responding to the elements, then in order for the elements at the ends of an eight-element open contour to transmit information to each other, 250 msec represents approximately 30 msec per iteration. Perhaps not coincidentally, cells spiking with this interspike interval would produce strong energy in the γ frequency range (roughly 30–80 Hz). We will return to the possible significance of this frequency range later.

B. Anatomical and Physiological Basis of Salience

These observations bear on the most important question: What is the neural representation of salience? Most attempts to answer this question have targeted the first visual cortical area, also known as primary visual cortex, where segmentation and grouping are also hypothesized to take place. Two aspects of this question have been addressed: How do neurons representing different parts of visual space interact with each other? What is the form of this interaction?

These two aspects are analogous to the physical wiring used to connect telecommunication devices and the transmission code (e.g., Morse code, digital encoding) used to relay the information over those wires.

The neurons in the primary visual cortex are organized topographically, with each neuron representing a part of visual space. Integration of a set of discontinuous oriented elements into a contour requires communication between spatially segregated sets of cortical cells. Charles Gilbert has proposed that this communication may be mediated, in part, by horizontal connections between pyramidal cells in superficial layers of primary visual cortex. These long-distance connections can travel several millimeters in cortex, corresponding to receptive field locations up to 6° apart (1° of visual space is approximately equivalent to the area spanned by the thumbnail when held at arm's length). Work by David Fitzpatrick and colleagues has shown that these connections are spatially organized, connecting cells whose receptive fields and orientation preferences are collinear. These connections are remarkably reminiscent of the Gestalt law of good continuation, the cocircular connectivity rule, as well as the association field, and these cortical connections may well underlie many of the feature relationships that are perceptually salient.[2]

The function of these connections has also been studied by the Gilbert group. By recording from awake monkeys, Gilbert and colleagues examined the response of cells in the superficial layers to sets of oriented bars placed in various configurations inside and outside the receptive field and compared the neuronal firing rates to perceptual responses. In agreement with their own psychophysical studies, as well as those of Field and colleagues, they found that cell responses diminish as the alignment of the oriented bars deviates from collinearity or cocircularity and also with increases in separation between the bars. The interaction between cells can be facilitatory or suppressive, depending upon the cells' relative positions and orientations.

Gilbert, together with Mitesh Kapadia and Gerald Westheimer, has also mapped the spatial structure of the facilitatory/suppressive regions using an ingenious technique based on the tilt illusion. The tilt illusion refers to an illusory perceptual tilt of a vertically oriented bar placed between two flanking bars oriented at a different orientation. The vertical target appears to tilt either toward (attractive) or away from (repulsive) the flanking bars. Attractive or repulsive tilts are observed as the flankers are placed at different spatial locations and orientations with respect to the target. Gilbert and colleagues found a beautiful correspondence between the direction of the tilt illusion and the presence of facilitation and suppression, effectively demonstrating the basis of the tilt illusion. Figure 9 shows the spatial structure of these facilitatory and inhibitory sites as derived from physiological recordings in the primary cortex of alert monkeys. Again, the agreement with the cocircular connection pattern is remarkable.

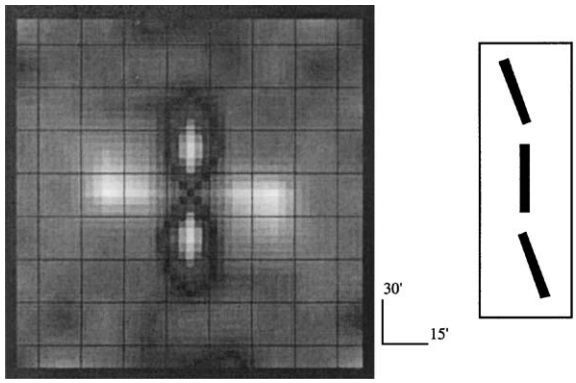


Figure 9 Map of contextual effects surrounding a neuron in primary visual cortex. The two vertical blobs in the middle indicate positions around the neuron where the introduction of a bar produced facilitation, whereas the regions in white surrounding the two vertical blobs indicate positions around the neuron where the introduction of a bar produced suppression. The scale bar indicates that the vertical grid points are separated by 30′ (0.5′), whereas the horizontal grid points are separated by 15′. The inset shows an example of a tilt illusion stimulus used in the studies. From Kapadia *et al.* (2000). Reprinted with permission from The American Physiological Society.

The studies described earlier suggest that primary visual cortical neurons show elevated firing rates when they are stimulated with aligned stimuli, but it is not clear whether this elevated firing rate is used to represent salience. The firing rate of a neuron is well-correlated with a number of different stimulus parameters, including the contrast, orientation, and spatial frequency of the stimulus. How would a neuron receiving input from these primary visual cortical neurons distinguish between elevated firing rates that

[2]It should be noted that extensive horizontal connections are also found in infragranular layers and in many extrastriate visual areas, as well as in higher cortical areas including prefrontal cortex. In fact, loss of horizontal connections in layer III is a common finding in the prefrontal cortex of schizophrenic brains, suggesting perhaps a more general role for such connectivity in cortical integration.

represent higher salience and elevated firing rates that represent, for instance, contrast? It has been shown that the performance of subjects in a contour detection task is not affected when the Gabor elements that make up the contour as well as those that are distractors have different contrasts. In this case, there would be neurons stimulated by high-contrast distractors that have higher firing rates than some of the neurons stimulated by elements on the contour. The salience representation would be confused in this situation.

One alternative is that, because contrast, orientation, and spatial frequency are represented using population codes, it might be possible to represent salience using a subpopulation of cells without perturbing the other representations. Another possibility that has been suggested is that the initial phase of the response of a neuron (less than 80 msec after the stimulus is presented to the retina) could represent local stimulus parameters like contrast, orientation, and spatial frequency, whereas the delayed response phase (80–150 msec) could represent global stimulus properties like salience. In fact, a number of research groups have shown that elevations in the firing rate during the delayed response phase appear to be correlated to global properties such as figure–ground segmentation and contour tracing. These effects have mostly been attributed to feedback from higher cortical areas, but it is certainly plausible that the horizontal connections might provide a similar effect. However, the effect is relatively small, and it is not clear that the cells could encode different levels of salience within such a small range in firing rates.

Separate evidence suggests that salience may be encoded in a different dimension of the neural response. Since the late 1970s, a number of groups have postulated that the timing of a neuron's response might be used to represent global properties. The idea was first postulated to be a solution to the puzzle that, although brain areas are highly specialized with different cortical areas responding to color, motion, shape, and location, our percept is perfectly unified. In addition, we have little difficulty paying attention to several objects with different properties. That meant that neurons in different areas of the brain not only have to be able to signify that they belong together as an object but they also have to be able to signify that they are distinct from the neurons representing other objects. This puzzle has often been referred to as the binding problem. The proposed solution postulated that, if groups of neurons representing the same object fired together, or in synchrony, and were distinct, or desynchronized, from other groups of neurons representing other objects, that would allow the brain to keep track of multiple objects with multiple stimulus properties. There has since been a large body of experimental evidence in many different brain areas and in many different animal species that supports this hypothesis, although it remains hotly debated in neuroscience. Interestingly, most of the experimental evidence has been correlated with increased energy in the γ frequency range.

A similar idea has been proposed to account for the representation of salience. For instance, we have proposed that the salience of a contour is represented by the degree of synchrony among the responding neurons. Our modeling studies have shown an interesting parallel between changes in synchrony and the changes in salience observed in a number of psychophysical studies. Of special interest is the ability of the models to account for the difference in the changes in salience between open and closed contours described earlier.

We modeled the neurons that were stimulated by Gabor elements as oscillators that were coupled in the cocircular connection pattern shown in Fig. 7 via the horizontal connections found in the superficial layers of primary visual cortex. (Cells exhibiting oscillatory or periodic firing have been observed in a number of physiological experiments and are thought to be closely linked to synchrony.) In our model, salience was measured using the amount of synchronized activity. It has been shown, both in simulation and analytically, that, whereas a chain of strongly coupled oscillators will reliably synchronize, a chain of weakly coupled oscillators will not approach synchrony. However, if the ends of the same chain are coupled together, the oscillators will reliably synchronize. Weak and strong coupling corresponds to connections only between nearest neighbors or connections between next nearest neighbors as well.

Our simulation results led us to propose the following account of the closure effect (also shown in Fig. 10): when the contour elements were separated by Δ_{Open}, the horizontal connections were strong enough to cause the neurons stimulated by the contour elements to synchronize. As a result, each additional element adds to the synchronized activity and, thus, the salience of the contour. When the contour elements were separated at Δ_{Closed}, the reduced strength of the horizontal connections was no longer effective in synchronizing the neurons stimulated by the contour, and as such there was no increase in synchronized

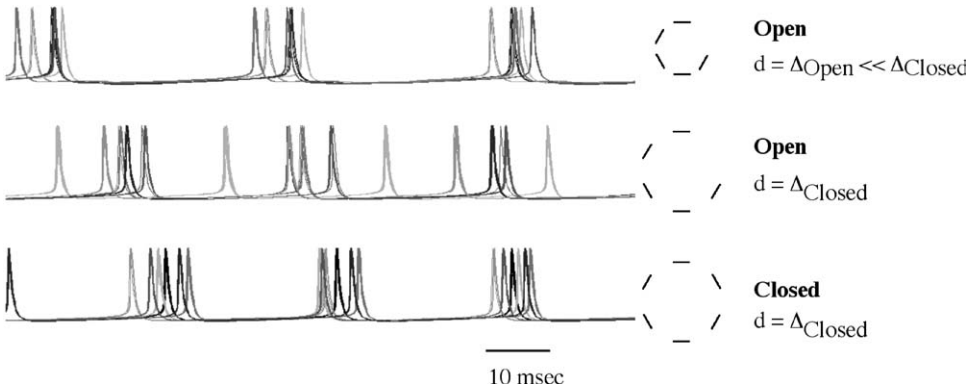


Figure 10 Simulation results illustrating the closure experiment. The different gray level traces represent the membrane potentials of groups of pyramidal cells in an orientation column stimulated by the different Gabor elements that make up the contour. Neighboring groups are connected via horizontal connections, and these connections decrease in strength as the separation, *d*, between the elements increases. The set of traces plotted at the top illustrates that, at the separation Δ_{Open}, the cells are synchronized, as they fire within 10 msec of each other. As a result, each additional element adds to the synchronized activity of the population and, thus, to the salience of the contour. The traces plotted in the middle show that, at the separation Δ_{Closed}, the cells are not synchronized, and as such each additional element does not add to the amount of synchronized activity. However, when the last element is added, which changes the contour from an open contour to a closed contour, the traces at the bottom show that the cells are again synchronized. This then leads to an increase in the synchronized activity and, thus, the salience of the contour.

activity with each additional element. However, when the last element was added, the topology of the connections between the neurons changed from an open chain to a closed chain, and this allowed the neurons to synchronize and, thus, led to an increase in the salience of the contour.

The use of synchrony to represent salient contours provides a way for contours in close proximity to be represented distinctly from one another. It also avoids the problem of confusing representations of salience with other feature dimensions. However, although we have been able to reproduce this effect using a wide variety of oscillator models, the time needed for synchrony to emerge in the models is substantially higher than the 100- to 250-msec time window estimated from the psychophysical studies discussed earlier. This suggests that, if the brain uses synchronized oscillations to represent salience, the oscillations may be of a very different nature from the oscillators that we have used in our models.

Although there is indirect evidence for the different hypotheses, the question of how the brain represents salience is still very much an open question. There are a number of plausible hypotheses, but the definitive experiments have yet to be done. The results of these experiments not only will provide insights into the representation of salience in the brain but also will shed light on an even more fundamental issue, the form of the codes used by the brain to process and transmit information.

V. THE ONTOGENY OF SALIENCE

Evolutionary biologists distinguish between proximate and ultimate causation of phenotypic properties. The properties of salience thus must have both proximal (physiological) and ultimate (evolutionary) components. Cortical interactions provide the proximal mechanisms underlying salience, but what determines the ultimate cause?

One idea proposed in the literature is that Gestalt properties such as continuity and cocircularity abound in natural images, and particularly in the types of objects of adaptive interest. Charles Gilbert and colleagues analyzed the statistics of a set of natural images (pictures of farmland, woods, etc.) and determined the probabilities of contour elements lying at various relative orientations as a function of their relative position in the images. They found that cocircularity was the most probable relationship, and there was a high degree of cocircular contours in the images analyzed. In a related study, William Geisler and colleagues identified all contours in a set of natural images and found that the statistically most likely relations between contour elements were described well by an association-field-like structure. Thus, collinear and cocircular arrangements are most common.

These and similar studies suggest that the visual system has evolved to detect and amplify the presence of continuous contours in the environment, perhaps

because of their association with the boundaries of objects of interest. Salience emerges from the resonance of naturally occurring continuity with the cortical mechanisms evolved based on these stimuli. In addition to the sun and the moon, biological organisms tend to be distinguished by continuous contours.

An additional consideration, however, arises from the active nature of vision. Babies have an intrinsic fascination with high-contrast salient edges—shadows, corners, etc. Most parents have had the experience of watching an infant focus back and forth along a shadow edge on the wall or stare, somewhat disturbingly, at the edge of one's face. Perhaps certain innate preferences are built into the nervous system for high-contrast edges. High contrast induces higher signal-to-noise ratios in the firing rates across a cortical region, and the system may try to maximize this type of stimulation through eye and head movements. What may be relevant then is not the statistics of natural images *on their own*, but rather the statistics of foveated regions of these images. Hebbian mechanisms underlying connection formation might also be sensitive to neuromodulatory inputs that reflect the innate degree of arousal by images, so that a high-contrast edge or a mother's face may induce more plasticity. In addition, moving stimuli evoke more activity than static images and the statistics of spatiotemporal imagery show higher degrees of continuity, so that these properties might also have an influence on the development of cortical connections.

Eye movements might induce much greater coactivation of collinearly or cocircularly arrayed cells than static stimuli, and the series of fixations might activate cells in a temporal pattern as required for synaptic plasticity. Thus, an innate preference for one type of salience, high-contrast edges or smooth motion, might drive cortical development, and the resulting network connectivity might then respond to other types of salient stimuli. In amblyopia, where the retinal images are discordant, cortical connectivity develops in a disorganized manner, and subjects display higher thresholds for detecting salient contours. The ability to influence the spatial and temporal pattern of stimulation, through active vision, may continue through life and may underlie our aesthetic appreciation for certain images over others.

VI. CONCLUSIONS

An object appears salient based on its features relative to those in the surroundings. As the Gestalt psychologists discovered, certain features and configurations of features are naturally more salient, and more recent work indicates that these correspond to intrinsic selectivities and connectivities in visual cortex, respectively. It is possible to quantify the salience of a target using psychophysical paradigms, and salience should correspond to the degree to which the target is a statistical outlier from the distractor distribution.

Despite these conclusions, the neural representation of salience remains an open question. A link between salience and synchronization would be theoretically satisfying, but the definitive experiments have yet to be done. The rapidity of perception, however, does put a possibly unattainable time constraint on any synchronization process.

We have concentrated on perceptual salience in the context of striate visual cortex; however, it is clear that many brain regions play a role in controlling salience. One of the primary effects of a lesion in extrastriate visual area V4, for example, is an inability to notice less salient objects in the presence of more salient objects. The hippocampus also has been proposed to restructure the representation of neocortical information to reflect its salience to the animal. Sustained activity in prefrontal cortex in delay-period working memory tasks has been shown to be robust against distractor inputs, and thus attractor-type behavior in prefrontal circuits maintains goal-directed behavior in the face of less salient competing drives. It appears that many, if not all, higher brain areas face similar problems of integrating bottom-up, top-down, and horizontal inputs and of regulating which among many competing parallel processes should receive priority. Many of these brain regions make use of similar anatomical architectures, including specific long-distance connections. Salience may be the perceptual manifestation of this process of integration and gating.

See Also the Following Articles

ATTENTION • COLOR PROCESSING AND COLOR PROCESSING DISORDERS • INFORMATION PROCESSING • MOTION PROCESSING • MULTISENSORY INTEGRATION • OBJECT PERCEPTION • PATTERN RECOGNITION • SPATIAL COGNITION

Suggested Reading

Alter, T. D., and Basri, R. (1997). Extracting salient curves from images: On analysis of the saliency network. *Int. J. Computer Vision* **27**(1), 51–69.

Braun, J. (1999). On the detection of salient contours. *Spat. Vis.* **12**(2), 211–225.

Field, D. J., Hayes, A., and Hess, R. F. (1993). Contour integration by the human visual system: Evidence for a local "association field." *Vision Res.* **33**(2), 173–193.

Gilbert, C. D., Das, A., Ito, M., Kapadia, M. K., and Westheimer, G. (1996). Cortical dynamics and visual perception. *Cold Spring Harbor Symp. Quant. Biol.* **61,** 105–113.

Kapadia, M. K., Ito, M., Gilbert, C. D., and Westheimer, G. (1995). Improvement in visual sensitivity by changes in local context: Parallel studies in human observers and in V1 of alert monkeys. *Neuron* **15**(4), 843–856.

Kapadia, M. K., Westheimer, G., and Gilbert, C. D. (2000). Spatial distribution of contextual interactions in primary visual cortex and in visual perception. *J. Neurophysiol.* **84**(4), 2048–2062.

Kovács, I., and Julesz, B. (1993). A closed curve is much more than an incomplete one: Effect of closure in figure–ground segmentation. *Proc. Natl. Acad. Sci. USA* **90**(16), 7495–7497.

Rosenholtz, R. (1999). A simple saliency model predicts a number of motion pop out phenomena. *Vision Res.* **39**(19), 3157–3163.

Treisman, A., and Gormican, S. (1988). Feature analysis in early vision: Evidence from search asymmetries. *Psychol. Rev.* **95,** 15–48.

Ullman, S., and Sha'ashua, A. (1988). Structural saliency: The detection of globally salient structures using a locally connected network. *AI Memo 1061*. Massachusetts Institute of Technology, Cambridge, MA.

Yen, S.-C., and Finkel, L. H. (1998). Extraction of perceptually salient contours by striate cortical networks. *Vision Res.* **38**(5), 719–741.

Yen, S.-C., Menschik, E. D., and Finkel, L. H. (1998). Cortical synchronization and perceptual salience. *Com-putational Neuroscience: Trends in Research*. J. Bower, Plenum, New York.

Schizophrenia

LISA T. EYLER ZORRILLA and DILIP V. JESTE
University of California, San Diego and VA San Diego Healthcare System

GLOSSARY

atypical antipsychotics A newer class of antipsychotic medications available since 1988 that act at both 5-HT_2 and D_2 receptors and carry a lower risk of motor side effects, e.g., clozapine.

conventional, or typical, antipsychotics A class of antipsychotic medications available since 1952 with a high affinity for D_2 dopamine receptors that carry a high risk for either motor side effects (with high-potency agents such as haloperidol) or sedation and anticholinergic side effects (with low-potency agents such as chlorpromazine).

dementia praecox Kraepelin's term for a psychotic disorder, eventually known as schizophrenia, characterized by progressive functional decline beginning in adolescence or early adulthood.

diathesis–stress model An etiologic theory in which environmental events interact with an underlying vulnerability (usually genetic) to produce a disorder.

executive functions Cognitive functions involving higher order processing such as abstraction, planning, problem solving, and initiation.

hypofrontality A state characterized by less blood flow or metabolism in the frontal lobes relative to posterior brain regions or a state of lower activity in the frontal cortex relative to healthy individuals.

negative symptoms Symptoms, including poverty of emotions, speech, and intention relative to healthy individuals, seen in patients with schizophrenia.

neurodegenerative disorder A disorder characterized by progressive deterioration of brain systems from a healthy baseline state, often marked by the presence of gliosis.

neurodevelopmental disorder A disorder caused by abnormal development of specific neural structures.

paranoid A common subtype of schizophrenia characterized by delusions, especially of persecution, a preponderance of auditory hallucinations, and few negative or disorganized symptoms.

positive symptoms Symptoms present in patients with schizophrenia and absent from healthy individuals, including delusions, hallucinations, and disorganized thought and behavior.

tardive dyskinesia A disorder characterized by abnormal, involuntary hyperkinetic movements that can develop in patients on long-term neuroleptic treatment.

Schizophrenia is a psychopathological disorder characterized by symptoms of delusions, hallucinations, disorganized thought and behavior, and poverty of emotion, speech, and intention. Features associated with the disorder include abnormalities of neurotransmitter systems, brain structure, neurophysiology, and cognition. Although the etiology of the disorder is not yet understood, genetic and perinatal predisposition appears to play a critical role. Treatment for the disorder involves pharmacological and psychosocial interventions.

I. INTRODUCTION

A. History

Schizophrenia is likely an ancient disorder. Despite many obstacles to identifying the disorder in historical records (e.g., lack of detailed descriptions and clear terminology, a tendency to ignore psychiatric diseases or consider them in the domain of religion), early accounts seem to contain descriptions of schizophrenia-like disorders. It was not until the nineteenth century, however, that schizophrenia as we know it

today was first defined. French physician Benedict Morel described *demence precose* in 1857, and Emil Kraepelin later formalized the diagnosis of *dementia praecox*, which was characterized by a progressive downward course with symptoms such as hallucinations, delusions, thought disorder, catatonia, blunted affect, and lack of insight. The term "schizophrenia" was coined in 1911 by Bleuler, who brought psychological theory to bear on the disorder and also broadened the concept to include thought-disordered individuals with progressive decline in functioning without evidence of delusions and hallucinations. Until the 1960s, there was a great deal of confusion regarding the diagnosis of schizophrenia due to conflicts between these two conceptualizations and evidence that some individuals with characteristic symptoms did not decline progressively. Eventually, awareness of substantial differences in rates of diagnosis of schizophrenia between the United States and other countries stimulated efforts to operationalize the criteria for the disorder. This led to development of many diagnostic schemes, which, though more structured, still differed considerably from one another. The two most widely used classification systems today are the Diagnostic and Statistical Manual of Mental Disorders, 4th ed. (DSM-IV) and the International Classification of Diseases, 10th ed. (ICD-10).

B. Symptoms

Diagnosis of schizophrenia is based on a constellation of signs and symptoms; there is no laboratory test that can definitively indicate the presence of the disorder. The DSM-IV criteria for schizophrenia include presence of certain characteristic symptoms for longer than 6 months and evidence of impairment in social or occupational functioning. In the ICD-10, symptoms must persist for only 1 month and functional consequences are not specified. Symptoms of schizophrenia can be broadly classified into "positive" (i.e., present in patients and absent from healthy individuals) and "negative" (i.e., reflecting a deficiency of normal psychological functions in patients). Positive symptoms include hallucinations, delusions, disorganized or catatonic behavior, and some forms of thought disorder. Hallucinations are most frequently auditory and may be voices of people talking to one another or commenting on the patient's behavior. Visual, tactile, olfactory, and gustatory hallucinations can be present but are less common. Delusions include paranoid ideation (i.e., the belief that others intend to harm the patient), delusions of reference (i.e., the belief that others are talking about the patient or that the television is communicating directly with the patient), somatic or grandiose delusions, and delusions of mind control, thought insertion or withdrawal, or thought broadcasting. Disorganized behavior includes silliness and unpredictability and may be reflected in disheveled or inappropriate dress and poor hygiene. Catatonia, including stupor, rigidity, negativism, posturing, and excitement, may also be observed but is much less common now than historically. Thought disorder is a disruption of the form of thought that is primarily evident in speech characterized by loosening of associations, disorganization, wandering or off-topic discourse, and use of idiosyncratic terms (i.e., neologisms). Although traditionally grouped with the positive symptoms, factor analytic studies of schizophrenic symptoms have often found that disorganized thought and behavior load on a separate factor from hallucinations and delusions. Negative symptoms form yet a third cluster, including flat affect, poverty of speech (i.e., alogia), and lack of will (i.e., avolition), and they are significant contributors to deficits in everyday functioning. Finally, although not part of most classification systems, cognitive deficits are now generally accepted to be characteristic symptoms that also seem to have a large impact on a patient's daily life.

Symptoms similar to those seen in schizophrenia are present in many other disorders. In schizoaffective disorder, symptoms of an affective disorder are present at the same time as active-phase psychotic symptoms and the mood component is substantial in duration, but hallucinations and delusions also must be present for at least 2 weeks in the absence of mood symptoms. Psychotic symptoms also can be found in primary mood disorders, brief psychotic disorder, delusional disorder, and substance abuse and dependence. There is considerable debate about the degree of overlap in the etiology of these potentially related disorders.

C. Subtypes

In response to the heterogeneity of symptoms, treatment response, and functional outcome in schizophrenia, a number of subclassifications of the disorder have been proposed. Subtypes based on symptom profiles include predominantly positive vs predominantly negative, as well as subtypes specified by the DSM-IV and ICD-10, some of which date back to Kraepelin and Bleuler's original formulations. These include

simple, hebephrenic, catatonic, paranoid, disorganized, undifferentiated, and residual. The symptom profiles associated with each of these subtypes are listed in Table I. Accumulating evidence suggests that symptom subtypes generally are not stable throughout the course of illness, thus reducing their usefulness for determining prognosis or etiology. In response, there has been a move away from symptom categorization toward a dimensional approach in more recent years. In addition, subtyping schemes that rely on distinctions regarding age of onset (e.g., early onset vs typical onset vs late onset), course [e.g., Kraepelinian (progressive) vs non-Kraepelinian (stable); reactive vs endogenous], or presence vs absence of associated features (e.g., family history positive vs negative; with or without cerebral ventricular enlargment) have become more prominent.

D. Onset, Course, and Prevalence

The typical age at onset of schizophrenia is in late adolescence or the early 20s for males, with a somewhat later age at onset for females. Although onset may be acute, often development of the disorder is preceded by a prodromal phase of illness that includes gradual development of symptoms such as social withdrawal, loss of interest in activities, deterioration in grooming, and unusual behavior. Personality disturbances including paranoid or schizotypal (e.g., unusual perceptions, thinking, and behavior) traits are common premorbidly.

Rarely, florid symptoms of schizophrenia manifest for the first time in childhood or late adulthood. Childhood-onset schizophrenia (i.e., onset before age 12) is qualitatively similar to adult-onset schizophrenia in symptoms, developmental precursors, and associated features. Prognosis, however, is generally worse, and evidence suggests that childhood-onset patients show prolonged deterioration in intellectual function and progressive ventricular enlargement. Schizophrenia can also manifest for the first time after age 45. Although precise estimates are lacking, approximately 15% of all the patients with schizophrenia may have onset after age 45. The prevalence of late-onset schizophrenia is difficult to determine due to poor sampling of older adults in epidemiologic studies and past criteria prohibiting the diagnosis of schizophrenia with onset in late adulthood. Research suggests that, whereas late-onset schizophrenia shares core features with earlier-onset schizophrenia, differences exist that raise the possibility of differing disease-modifying factors. Table II presents a comparison of features of childhood-, late-, and typical-onset schizophrenia.

There is considerable variation in the course of schizophrenia. After an initial acute stage during which the risk for suicide is quite high, a small proportion of patients suffer a declining course similar to that described by Kraepelin, a majority evidence a stable or fluctuating course, and a smaller percentage appear to remit to a clinically significant degree (Fig. 1). Several factors appear to influence prognosis and course in schizophrenia. Features associated with

Table I
Symptom Profiles Associated with Subtypes of Schizophrenia

Subtype	Symptom profile	Classification system
Catatonic	One or more of the following: stupor or mutism, excitement, posturing, negativism, rigidity, waxy flexibility, automatic imitation of sounds or gestures (i.e., echolalia, echopraxia).	DSM-IV[a] and ICD-10[b]
Disorganized	All of the following are prominent: disorganized speech, disorganized behavior, flat or inappropriate affect. Criteria for catatonic type are not met.	DSM-IV
Hebephrenic	Disturbances of affect, volition, and thought disorder are prominent. Hallucinations and delusions are not prominent. Onset between ages 15 and 25 years.	ICD-10
Paranoid	Preoccupation with one or more delusions or frequent auditory hallucinations. Disorganization, catatonia, and flat affect are not prominent.	DSM-IV and ICD-10
Residual	Absence of prominent hallucinations, delusions, disorganization, or catatonia, in the context of continuing negative symptoms or attenuated positive symptoms.	DSM-IV and ICD-10
Simple	Slowly progressive development of negative symptoms without a history of positive symptoms.	ICD-10
Undifferentiated	Symptom profile does not fit any of the other subtypes.	DSM-IV and ICD-10

[a]Diagnostic and Statistical Manual of Mental Disorders, 4th ed.
[b]International Classification of Diseases, 10th ed.

Table II
Childhood- and Late-Onset Schizophrenia: Similarities with and Differences from Typical-Onset Schizophrenia

Similarities	Differences
Childhood Onset	
Symptom profile	More chronic course and poorer outcome
Deficits in attentional processing	Continued declines in intellectual functioning 2–3 years after onset
Decreased total brain volume, increased volume of lateral ventricles, decreased thalamic volume	Larger effect sizes for decreases in total brain and thalamus volume and progressive longitudinal changes
Qualitative response to antipsychotics, including posttreatment homovanillic acid levels	
Premorbid developmental delays and neurologic signs	
Eye-tracking abnormalities	
Metabolic hypofrontality in adolescents	
Late-Onset	
Degree of psychopathology	Preponderance of women
Severity of positive symptoms	Less severe negative symptoms
Family history	Mostly paranoid subtype
Qualitative response to antipsychotics	Lower daily antipsychotic doses
Early childhood maladjustment	Better early adulthood psychosocial functioning
Overall pattern of cognitive deficits	Milder impairments in learning, retrieval, and abstraction–cognitive flexibility
Disrupted quality of well-being	Less disturbance of semantic network
Degree of nonspecific structural brain abnormalities (ventricular enlargement, white matter hyperintensities)	Larger thalamus (?)
Minor physical anomalies	Unique electrophysiological findings (e.g., normal P300 amplitude, longer latency of N400 congruity effect)
Constitutional (uncorrected) sensory impairment	
Increased mortality	

poorer outcome include hebephrenic or simple subtype, early and insidious onset, delayed neuroleptic treatment, more negative symptoms, greater cognitive deficits, and poor premorbid adjustment.

The prevalence of schizophrenia is approximately 1%. Because of the morbidity associated with the disorder, however, it is the most expensive psychiatric disorder. The total economic burden of schizophrenia is estimated to be $40–60 billion, including both direct (approximately $30 billion) and indirect (approximately $20 billion) costs.

II. SCHIZOPHRENIA AND THE BRAIN

Despite a tradition of using psychological theories to explain the psychopathology of schizophrenia that began with Bleuler and Freud, most contemporary research into the etiology of the disorder has focused on the role of brain abnormalities. Deficits at all levels of analysis, from the synaptic cleft to cognitive performance, have been associated with schizophrenia. Unfortunately, none of these abnormalities has yet proven sensitive or specific enough to implicate a clear etiopathology of schizophrenia.

A. Neurotransmitter Abnormalities

1. Dopamine

One of the most researched biochemical theories of schizophrenia is the "dopamine hypothesis," which posits that the disorder results from dopamine (DA) overactivity in the brain. This theory was developed in response to evidence that the administration of DA agonists induced psychotic symptoms and research showing that neuroleptic drugs with efficacy in the treatment of schizophrenia were those that effectively

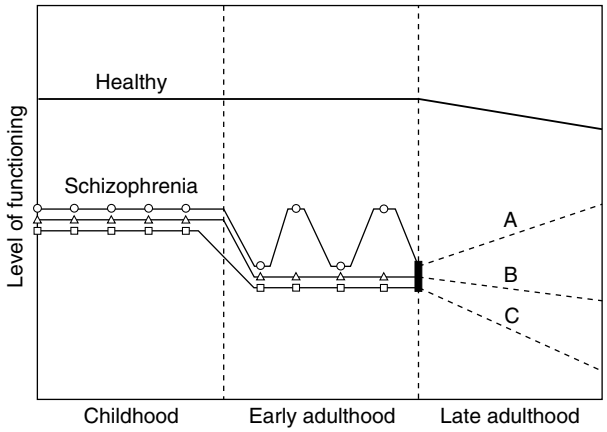


Figure 1 Possible functional courses of schizophrenia in childhood, early adulthood, and late adulthood, compared to healthy individuals. Healthy individuals (black line) show good functioning in childhood and early adulthood with mild declines in late adulthood. In childhood, patients with schizophrenia may have mild functional deficits (○, △, □) compared to healthy individuals, and some patients (□) may begin to show prodromal signs of schizophrenia. Onset of the disorder typically occurs in adolescence or early adulthood and may be acute (△, ○) or more gradual (□). Course can be stable (□, △) or fluctuating (○). In late adulthood, most patients remain relatively stable with only mild age-related declines (B), whereas a smaller proportion may show clinically significant remission (A) or continued degeneration as proposed by Kraepelin (C).

blocked DA receptors in the brain. Further work, however, failed to support an increase in DA turnover as indexed by levels of the metabolites homovanillic acid (HVA) and 3,4-dihydroxyphenylacetic acid (DOPAC), and studies of DA levels in post mortem brains were equivocal. Instead, the evidence suggests that DA overactivity may be the result of increased receptor sensitivity, as post mortem studies have consistently demonstrated increased numbers of and binding to D_2 receptors in the brains of patients. Unfortunately, interpretation of post mortem studies is complicated by the long history of neuroleptic exposure in most patients. Whereas neuroimaging using radioactive DA receptor ligands holds great promise for overcoming this difficulty, the results so far have been conflicting. Some studies of neuroleptic-naive patients have shown the up-regulation of receptors, whereas others do not find differences. The myriad receptor subtypes for DA further complicates the investigation of the role of this neurotransmitter. In addition, studies have suggested that positive symptoms may be related to hyperdopaminergia, whereas negative symptoms are related to hypodopaminergia in different areas of the brain. Clearly, more research is needed to specify the details of the dopamine hypothesis and to explore the usefulness of this theory for understanding the neurochemistry of schizophrenia.

2. Serotonin

Partly in response to the shortcomings of the dopamine hypothesis, studies have revived an interest in the role of serotonin (5-HT) in schizophrenia that began with the observation of the hallucinogenic properties of structurally similar lysergic acid diethylamide (LSD). Alterations in 5-HT receptors have been found in limbic and prefrontal cortices, and atypical antipsychotics with an affinity for 5-HT_2 receptors have been reported to improve negative symptoms of schizophrenia. Work has emphasized the importance of interactions between 5-HT and DA, because studies of 5-HT metabolite levels have been equivocal and selective 5-HT_2 antagonists do not ameliorate positive symptoms.

3. Glutamate

The excitatory amino acid glutamate, which is the primary transmitter of pyramidal neurons, may also play a role in the neurochemistry of schizophrenia. Although direct examination of glutamate levels has yielded conflicting results, studies of the *N*-methyl-D-aspartate (NMDA) receptor have proved more fruitful. Generally, it has been shown that NMDA receptor levels and binding are reduced in medial temporal lobe structures and increased in orbital frontal cortex. These findings are consistent with etiologic theories of early developmental insults to the medial temporal lobe and failure of proper neuronal pruning in the frontal lobe. Also, early lesions of glutamate receptors in rats produce behavioral changes only after puberty, which models the usual onset of schizophrenia. Finally, glutamate and DA interactions in the striatum, with subsequent influence on the activity of the thalamus, have been proposed to underlie the defective sensory filtering seen in patients with the disorder.

4. Others

Investigations of the role of other neurochemicals, such as norepinephrine (NE), monoamine oxidase (MAO), γ-aminobutyric acid (GABA), and acetylcholine (ACh), have provided mixed results. Early studies that demonstrated a reduction in dopamine-β-hydroxylase, a marker for NE neurons, have not been replicated. Similarly, some studies have found

alterations in the distribution of NE receptor subtypes in limbic regions, whereas others failed to show a difference between patients and controls. Platelet levels of MAO, which inactivates DA, NE, and 5-HT in the brain, have been proposed to mark genetic vulnerability to schizophrenia because reduced activity was found in discordant twins. Other studies have not supported an MAO abnormality, however, and research has suggested that MAO levels are influenced by neuroleptic exposure. In addition, cigarette smoking, which is widespread among patients with schizophrenia, has been found to reduce MAO levels, which may also account for the apparent heritability of this putative marker. Finally, MAO activity in brain tissue has been shown not to be deficient. The interactions between GABA and DA in the striatum and limbic system led to an interest in this inhibitory amino acid. Studies of glutamate decarboxylase, a marker of GABA activity, have not found abnormalities in patients with schizophrenia, however. Results of receptor-binding studies have been equivocal as well, and GABA agonists do not appear to have any therapeutic efficacy. Finally, a possible role for ACh deficiency in schizophrenia has been examined. ACh agonists do not seem to have efficacy in the treatment of positive and negative symptoms, and investigations of brain receptor levels have yielded conflicting results. An association between cigarette smoking and schizophrenia, however, suggests a possible role for nicotinic ACh receptors. Ongoing work is centered on the importance of ACh in the cognitive deficits of schizophrenia.

B. Structural Abnormalities

1. Macroscopic

Studies of the volume of brain structures using post mortem samples as well as *in vivo* neuroimaging techniques have found differences between patients and controls in both generalized and specific features. The most consistent finding has been that patients with schizophrenia show lateral and third ventricle enlargement (Fig. 2), although the magnitude of these changes is small and there is considerable overlap between the patient and control distributions. In addition, schizophrenia has been associated with overall loss of cortical gray matter, decreased volume of the temporal lobe, especially in medial temporal structures, and more modest decrements in frontal lobe volume. In addition to these alterations, growing evidence suggests that normal brain asymmetries are disrupted in schizophrenia, with somewhat greater pathology on the left side. Volumetric abnormalities do not appear to be restricted to one subgroup of patients, and they are present in first-episode, unmedicated patients. In addition, individuals at high risk for developing schizophrenia (e.g., children of patients with schizophrenia) show similar structural brain abnormalities. The course of macroscopic abnormalities in patients is controversial: for some, values remain stable over time, whereas others show a deteriorating course.

2. Microscopic

A great deal of neuropathological research in schizophrenia has addressed the question of whether there is evidence of gliosis in the brains of patients. The weight of the evidence suggests that there is no significant gliosis, thus lending support to a neurodevelopmental, rather than neurodegenerative, etiopathology of schizophrenia. Studies also have examined the incidence of Alzheimer's-like pathology and found that, even among elderly schizophrenia patients with severe dementia, no such pathology is evident. Further support for a neurodevelopmental view of schizophrenia comes from cytoarchitectural studies. Decreased neuronal size in the hippocampus and frontal lobes has been found, although neuronal density studies have been equivocal. Synaptic density, as reflected by synaptic protein levels, is reduced in the hippocampus, frontal and cingulate cortex, and thalamus of patients with schizophrenia. Despite continuing controversy and some failures to replicate, the literature also supports some degree of abnormality of the structure of the entorhinal cortex, disarray of hippocampal pyramidal neurons, and deeper distribution of subplate neurons in the white matter of the frontal and temporal cortices. Finally, magnetic resonance spectroscopy (MRS) studies have demonstrated abnormalities in cell membrane phospholipids in the frontal lobes and decreases in *N*-acetylaspartate, a putative marker of neurons, in the temporal lobe.

C. Neurophysiological Abnormalities

1. Metabolism and Blood Flow

Positron emission tomography (PET) and single photon emission computed tomography (SPECT) have been used extensively to study differences between patients with schizophrenia and healthy

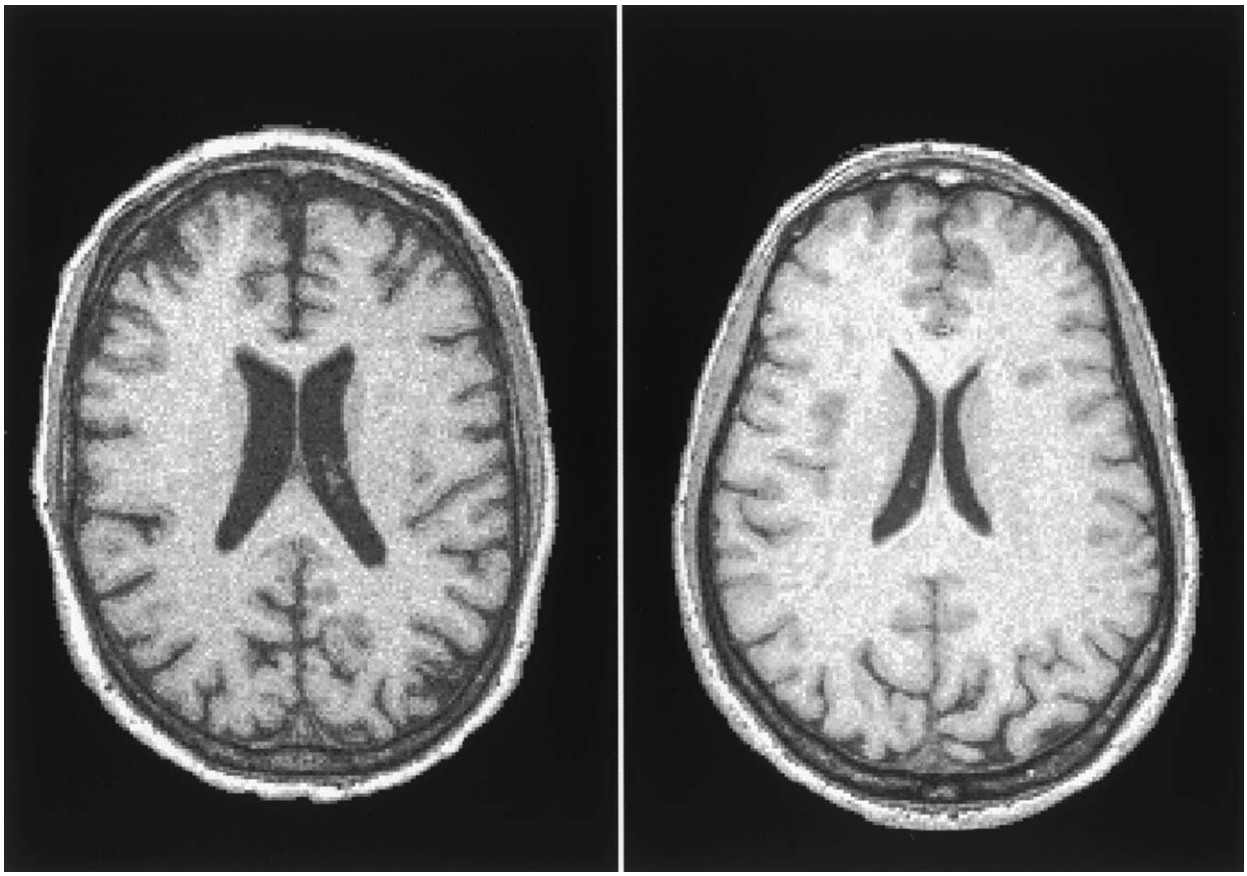


Figure 2 Magnetic resonance images of the brain at the level of the lateral ventricles in a patient with schizophrenia (left) and a healthy comparison subject. Both subjects were 57 years old at the time of the scan. Notice the greater size of the ventricles as well as more prominent sulci in the patient's image.

individuals in both brain metabolism and regional cerebral blood flow (rCBF). Resting metabolism, as measured by the uptake of radioactively labeled fluorodeoxyglucose (FDG), has been found to be lower in the frontal lobes than in the occipital lobes in many studies. This contrasts with healthy individuals, who generally demonstrate higher frontal metabolism at rest. The term "hypofrontality" has been used to describe both the frontal–occipital imbalance seen in patients and the relative deficiency compared to healthy individuals in frontal activity. Hypofrontality has also been observed in many, but not all, studies of rCBF during rest, and the degree of abnormality appears to be related to the severity of negative symptoms. Metabolic reductions in the striatum of unmedicated patients have been demonstrated repeatedly, and left greater than right temporal lobe metabolism in the context of overall lower temporal metabolism is a common finding. More recent studies using PET, SPECT, or the newer functional magnetic resonance imaging (fMRI) have tended to examine rCBF of patients during cognitive tasks. Although complicated by performance differences between patients and controls, possible effects of structural abnormalities, and the impact of neuroleptic medication, these functional neuroimaging studies have consistently demonstrated reduced blood flow to frontal regions during cognitive challenge. Abnormalities of blood flow and metabolism appear to be stable, but the few published twin studies suggest that they are state rather than trait deficits. Because hypofrontality has been observed consistently in affective disorders as well, the specificity of blood flow and metabolism deficits for understanding the etiopathology of schizophrenia remains to be seen.

2. Electrophysiology

Another approach to understanding the role of the brain in schizophrenia has been to study evoked

response potentials (ERPs) measured on the scalp using an electroencephalograph (EEG). By far, most research has focused on the positive-going electrical impulse that occurs in healthy individuals approximately 300 msec following an unusual target stimulus. The P300 (or P3) component is reduced in amplitude in patients with schizophrenia, even those who have never been medicated, and the reduction is often greatest over the left temporal lobe. Whereas P300 abnormalities are found in other psychopathological conditions, the lateralized findings appear to be more specific to schizophrenia. Evidence from depth recordings, lesion studies, fMRI investigations, and new methods of determining ERP localization has pointed to the superior temporal gyrus and inferior parietal lobe as primary sources of the P300 component. Other ERP research has focused on the P50 response. Patients with schizophrenia fail to show a suppression of the P50 response that occurs in healthy individuals when two auditory stimuli follow each other in rapid succession. This effect appears to be trait-related, as it is also reliably observed in the patients' nonpsychotic biological relatives. Several negative-going ERPs also have been investigated. Patients show abnormalities in the N100 evoked response to auditory stimuli, in the mismatch negativity seen approximately 200 msec following an unusual or novel stimulus presented in the background, and in the N400 component elicited in response to semantic incongruity. Further research is needed to determine the sensitivity and specificity of these less-studied ERPs.

D. Behavioral and Cognitive Abnormalities

1. Startle Response

The startle reflex is a cluster of defensive responses to a sudden, intense stimulus. In healthy individuals, the magnitude of startle diminishes with each stimulus presentation. Patients with schizophrenia, however, show reductions in the normal habituation to startle. In addition, they do not exhibit prepulse inhibition (PPI). PPI refers to the normal suppression of the startle response by the introduction of a weaker prestimulus and is thought to reflect behavioral gating processes that are hard-wired in the mammalian brain. Individuals with schizotypal personality traits show similar deficits. Deficits in PPI are not specific to schizophrenia, however, as they have been found in other disorders with putative abnormalities in the cortico–striato–pallido–pontine circuitry thought to be involved in the startle response. Because the startle response is seen in all mammals, this behavioral abnormality has proved useful for the development of some animal models of schizophrenia.

2. Oculomotor

Many studies have demonstrated that patients with schizophrenia show abnormalities in smooth-pursuit (SP) voluntary eye movements required to follow a moving target. Although the source of some continuing debate, these deficits appear to result from low-pursuit gain (the ratio of eye velocity to target velocity) with a concomitant increase in the number of corrective saccades. SP deficits do not appear to be affected by neuroleptic medication and are present in nonpsychotic biological relatives of patients. Impairments of SP eye movements are also found in other psychopathological disorders, though further detailed analysis of the nature of the deficits may reveal a more specific marker of schizophrenia. In contrast, abnormalities in the performance of an antisaccade task appear to be relatively specific to patients with schizophrenia and their relatives. This inability to inhibit a reflexive saccade to a location while making a voluntary saccade to an opposite location is thought to reflect functional deficits of the frontal lobe.

3. Cognition

Patients with schizophrenia suffer from cognitive deficits that are stable across changes in clinical status. Thus, studies of neuropsychological abnormalities are thought to provide information about the brain systems involved in schizophrenia. Several areas of particular impairment exist above and beyond the oft-observed deficits in global intellectual functioning. Patients consistently perform poorly on tasks of sustained and selective attention. Deficits on the same tasks are found in individuals at risk for developing schizophrenia (i.e., patients' biological relatives). Learning and memory are impaired too, and deficits in these abilities are also apparent in patients' first-degree relatives. Acquisition of new information and free recall of verbal and visual information are both impaired. In contrast, patients are not likely to forget information rapidly once it is learned, and procedural learning appears to be relatively intact. Working memory, or the on-line manipulation or maintenance of information in a short-term store, has been shown to be abnormal. These results point to a disruption of frontal lobe mechanisms known from animal studies to be involved in working memory tasks. Further

evidence for the role of frontal dysfunction in schizophrenia comes from studies of executive functions. Patients perform more poorly than healthy individuals on tasks that assess planning, strategy, set shifting, and abstraction, such as the Wisconsin Card Sorting Test and the Tower of London test. Neuroimaging studies have suggested that failure on these tasks is related to underactivation of the frontal lobes. The weight of the evidence from longitudinal studies and family studies supports the idea that cognitive deficits in schizophrenia are stable, trait-related features of the disorder. Research also suggests that these abnormalities are related to functional deficits such as quality of life, independence, and work status. Unfortunately, most current treatments do not significantly impact cognitive deficits in patients with schizophrenia.

III. THEORIES OF ETIOLOGY

A. Genetics

Schizophrenia has a strong familial component. Siblings and parents of patients have an 8–10% chance of developing the disorder and offspring have a 12–15% risk, compared to a 1% risk in the general population. Evidence from twin and adoption studies suggests that a portion of the familiality of the disorder is genetic. Specifically, monozygotic (MZ) twins have a 40–60% concordance rate, whereas dizygotic (DZ) twins have a 12–15% concordance. In addition, the offspring of schizophrenic mothers who are adopted away have a higher risk of schizophrenia than those from nonschizophrenic mothers.

Despite strong support for the role of genes in the disorder, the search for a particular gene related to schizophrenia has been disappointingly unfruitful. Linkage analyses have revealed several candidate markers that have not been confirmed in subsequent studies. The most studied have been the long arm of chromosome 22, chromosome 6p, and genes for various dopamine receptors.There are several methodological issues that complicate the search for a ''schizophrenia gene.'' First, there is great heterogeneity of symptom presentation among those diagnosed with schizophrenia. It is quite possible that different symptom profiles are related to abnormalities in different sets of genes. Thus, by combining all subtypes into one category, potentially relevant genes may be missed. Relatedly, the boundaries of schizophrenia are not well-defined. Results of linkage analyses are likely to be strongly influenced by how broadly the schizophrenia spectrum is defined. Thirdly, schizophrenia is not likely to be caused by just one gene. Multiple genes, interacting with one another, are the appropriate targets for genetic research in this disorder. Finally, although genes clearly play some role in the etiology of schizophrenia, results from twin and adoption studies also point to the importance of environmental factors. For instance, the concordance rate for MZ twins is only half of that expected in a purely genetic disorder, and adoption studies have found that characteristics of the adoptive parents contributed substantially to the risk of developing schizophrenia. Thus, even if genes related to schizophrenia are discovered, they are unlikely to explain fully the etiology of the disorder.

B. Environment

Early theories regarding environmental risk factors for schizophrenia focused solely on the role of parenting styles (especially, ''schizophrenogenic'' mothering). These theories were eventually supplanted, however, by hypotheses centering on biological precursors. Most current research has focused on influences in the perinatal period of development. There are three main reasons for this emphasis. First, as reviewed earlier, the nature of the brain abnormalities in schizophrenia suggests neurodevelopmental aberrations instead of neurodegenerative processes. Second, patients with schizophrenia have an increased prevalence of minor physical anomalies and neurological soft signs, pointing to a gestational origin. Third, birth-cohort and high-risk studies that followed subjects into the risk period for development of symptoms have consistently demonstrated that patients with schizophrenia show deficits long before the onset of the disorder. By using techniques such as the retrospective review of home movies and records as well as prospective studies, it has been shown that individuals who later develop schizophrenia demonstrate at an early age motor and psychomotor abnormalities, cerebral ventricular enlargement, neurological soft signs, cognitive deficits including lower IQ and attentional dysfunction, and social and emotional disturbances at school.

Exploration of perinatal environmental influences on the development of schizophrenia has centered on two main candidates: obstetric complications and viruses. History of obstetric complications (OCs) is more common in patients with schizophrenia than in healthy individuals. Specifically, prematurity, low birth weight, and complications involving trauma or

oxygen deprivation are frequently found. Because the medial temporal lobe is known to be particularly sensitive to hypoxia, the prevalence of OCs among patients provides a potential origin of the neurodevelopmental deficits seen in this region. A viral origin of schizophrenia has also been proposed. Support for this hypothesis comes from epidemiological studies showing an increased rate of schizophrenia among the offspring of women who were in the second trimester of pregnancy during an influenza epidemic. In addition, it has been argued that the increased risk of schizophrenia found among those born in the winter, in urban areas, and during times of famine supports a viral etiology. Finally, adult patients with schizophrenia show a wide variety of immunologic abnormalities, although the specificity of these deficits is not determined. Despite these pieces of circumstantial evidence, no virus has been identified that explains most cases of schizophrenia.

C. Interactive

Due to the complexity of schizophrenia, it is unlikely that genes or environmental factors alone cause the disorder. Rather, some interaction of these factors probably leads to the heterogeneity of symptom profiles, course, and outcome. The diathesis–stress theory posits that environmental factors combine with an underlying genetic vulnerability to produce schizophrenia. In support of this theory, research suggests that the offspring of mothers with schizophrenia who also suffer delivery complications are the most likely to develop schizophrenia later in life. In addition, early separation from caregivers increases the risk of schizophrenia only among those individuals with a family history of the disorder. Finally, studies have suggested that individuals at genetic risk for the disorder who receive poor parenting are more likely to develop schizophrenia. Thus, environmental risks at many stages of development appear to interact with genetic factors to promote the genesis of schizophrenia.

D. Developmental

A complete understanding of the etiopathology of schizophrenia must include an explanation of the age at onset of characteristic symptoms. As reviewed earlier, most known risk factors have their effects perinatally. In addition, subtle abnormalities of functioning are evident in childhood and adolescence in those individuals who are eventually diagnosed with schizophrenia. The pathognomonic signs and symptoms of schizophrenia, however, do not usually arise until late adolescence or early adulthood and typically erupt rather forcefully in the form of a "first break." In addition, in some individuals onset is earlier (childhood onset), and in others, especially women, onset is later in life (late onset). One class of theories that attempts to explain these facts hypothesizes that perinatal events (either genetic, environmental, or an interaction) create a brain vulnerability that only becomes apparent after normal developmental changes occur near the age at onset. In search of support for these hypotheses, the role of developmental hormones in the onset of schizophrenia has been investigated. In childhood-onset schizophrenia, the onset of puberty is related to the onset of the disorder, though only in girls. In addition, the preponderance of women among late-onset patients with schizophrenia has led to the suggestion that female reproductive hormones, such as estrogen, may be protective factors against onset of the disorder. It is posited that the decrease in estrogen levels during menopause promotes the onset of schizophrenia. Related theories suggest that the nature of the congenital abnormality is a *disruption* of normal developmental changes, such as a delayed expression of abnormal genes or failure of normal synaptic pruning that should occur around the second decade of life. Finally, other theories emphasize an accumulation of adverse events that eventually reach a threshold, leading to onset of the disorder. At this time, no definitive evidence has been found that favors any of these theories over others.

IV. TREATMENT

Until the etiology of schizophrenia is well-understood, curative treatment may not be possible. Current palliative treatments of schizophrenia include both pharmacological and psychosocial approaches and are most effective for positive symptoms. More research on interventions aimed at the improvement of negative symptoms, cognitive deficits, and functional status is needed.

A. Pharmacological

Since the 1950s, when chlorpromazine was first shown to be effective for the treatment of patients with schizophrenia, antipsychotic medications have been used to treat the psychotic symptoms of the disorder

Table III
Selected Antipsychotic Agents Used in the Treatment of Schizophrenia

Drug	Possible mechanism of action	Clinical benefits	Important side effects
Conventional:			
Chlorpromazine	D_2[a] blockade (low potency)	Positive symptoms, calming effect	Autonomic, sedation, EPS,[b] TD[c]
Haloperidol	D_2 blockade (high potency)	Positive symptoms, calming effect	EPS, TD, NMS[d]
Atypical:			
Clozapine	Selective blockade of mesolimbic neurons via D_2[e] occupancy and/or combination of lower D_2 blockade with 5-HT_{2A}[f] antagonism	Positive symptoms, negative symptoms, improved cognition and functioning	Agranulocytosis, sedation, anticholinergic, orthostatic hypotension
Risperidone	Antagonism of D_2, 5-HT_{2A}, α_1[g], and α_2[h]	Positive symptoms, negative symptoms, improved cognition (especially working memory) and functioning	Somnolence, orthostatic hypotension, EPS, hyperprolactinemia
Olanzapine	Greater 5-HT_{2A} than D_2 antagonism, affinity for D_3[i], D_4, 5-HT_3[j], 5-HT_6[k], H_1[l], α_1, and M_{1-5}[m]	Positive symptoms, negative symptoms, improved cognition (especially verbal fluency) and functioning	Sedation, orthostatic hypotension, weight gain
Quetiapine	High 5-HT_{2A} relative to D_2 and D_1[n] antagonism	β positive symptoms, β negative symptoms	Sedation, orthostatic hypotension, increased risk of cataracts (?)

[a]Dopamine subtype 2 receptor.
[b]Extrapyramidal symptoms.
[c]Tardive dyskinesia.
[d]Neuroleptic malignant syndrome.
[e]Dopamine subtype 4 receptor.
[f]Serotonin subtype 2A receptor.
[g]Noradrenaline subtype α_1 receptor.
[h]Noradrenaline subtype α_2 receptor.
[i]Dopamine subtype 3 receptor.
[j]Serotonin subtype 3 receptor.
[k]Serotonin subtype 6 receptor.
[l]Histamine subtype 1 receptor.
[m]Acetylcholine muscarinic subtype 1–5 receptors.
[n]Dopamine subtype 1 receptor.

and to prevent relapse. Conventional or typical neuroleptics, including the low-potency agent chlorpromazine and high-potency drug haloperidol, repeatedly have demonstrated efficacy compared to placebo, especially in the treatment of positive symptoms such as hallucinations, delusions, thought disorder, and bizarre behavior. In addition, these compounds have been shown to prevent relapse: a majority of untreated patients relapse, whereas only about 20% of treated patients do so over a 1-year period. Efficacy appears to be related to the affinity of these compounds for the D_2 dopamine receptor (see Table III), although the onset of antipsychotic effect is generally slower (5–10 days) than the time course of receptor blockade (hours). Unfortunately, this same affinity results in a high risk of motor side effects, such as extrapyramidal symptoms (EPSs) and potentially persistent tardive dyskinesia (TD). EPSs include akathisia (i.e., restlessness), dystonia, tremor, and rigidity and are generally treated with anticholinergic agents, such as benztropine. TD is a movement disorder characterized by abnormal, choreiform movements of the mouth, face, limbs, or trunk, which can develop after chronic neuroleptic exposure. The benefits and risks involved in long-term treatment with typical neuroleptics complicate decisions regarding the optimal duration of treatment and whether discontinuation is recommended. In addition to these issues, conventional neuroleptics do not seem to reduce the chronicity of the disorder, do not ameliorate negative symptoms, and have little impact on cognitive deficits.

In contrast, the newer atypical antipsychotics, such as clozapine, risperidone, olanzapine, and quetiapine, reduce the intensity of positive symptoms, carry a

lower risk for EPSs and TD, and seem to demonstrate better efficacy in treatment of negative symptoms. The atypicality of these medications appears to stem from their affinity for non-D_2 receptors, such as the D_1 and D_4 dopamine receptors and 5-HT_2 serotonin receptors (see Table III). Claims that newer atypical medications also improve cognitive deficits and everyday functioning require further investigation. In addition to antipsychotic medications, other psychotropic agents are sometimes employed in the treatment of schizophrenia. Mood stabilizers such as lithium and antiseizure medications, antidepressants, and anxiolytic agents have all been used. None of these compounds has proved superior to neuroleptics in the management of psychosis, but their use as adjunctive agents, especially in patients with schizoaffective disorders, has been supported.

B. Psychosocial

Whereas antipsychotic medication is the first line of treatment for individuals with schizophrenia, several observations point to the important role that psychosocial and rehabilitative interventions can play. First, as noted earlier, medication treatment does not improve all types of symptoms and often does not result in improvements in functional status, such as returning to work or decreasing level of care. In addition, medication compliance among patients with schizophrenia is often poor. Because medications can only be effective if they are actually taken by the patient, improving adherence is one of the targets for psychosocial intervention. Finally, studies suggest that relapse among medicated patients is more likely following stressful life events. Thus, interventions aimed at coping with such situations may reduce relapse rates.

Some psychosocial interventions have targeted symptom reduction. Behavioral and cognitive–behavioral techniques aimed at reducing hallucinations, delusional beliefs, and bizarre behavior generally show efficacy, but many of these interventions have failed to generalize outside the treatment setting. Other interventions have focused on improving problem solving, social skills, independent living skills (including medication management), coping skills, and cognitive deficits. In general, studies have shown that these treatments decrease symptoms and relapse rates and improve the everyday functioning of patients with schizophrenia. Further research is necessary, however, to identify the necessary features of these treatments and to determine whether all, or only specific, patients benefit.

See Also the Following Articles

BEHAVIORAL NEUROGENETICS • CATECHOLAMINES • CREATIVITY • DEMENTIA • DOPAMINE • HALLUCINATIONS • MANIC–DEPRESSIVE ILLNESS • MOOD DISORDERS • NEURODEGENERATIVE DISORDERS • NEUROPHARMACOLOGY • NEUROPSYCHOLOGICAL ASSESSMENT

Suggested Reading

Andreasen, N. C. (1995). Symptoms, signs, and diagnosis of schizophrenia. *Lancet* **346,** 477–481.

Braff, D. L., Swerdlow, N. R., and Geyer, M. A. (1995). Gating and habituation deficits in the schizophrenia disorders. *Clin. Neurosci.* **3,** 131–139.

Buchsbaum, M. S., and Hazlett, E. A. (1998). Positron emission tomography studies of abnormal glucose metabolism in schizophrenia. *Schiz. Bull.* **24,** 343–364.

Cannon, T. D. (1996). Abnormalities of brain structure and function in schizophrenia: Implications for etiology and pathophysiology. *Ann. Med.* **28,** 533–539.

Friedman, D., and Squires-Wheeler, E. (1994). Event-related potentials (ERPs) as indicators of risk for schizophrenia. *Schiz. Bull.* **20,** 63–74.

Harris, M. J., and Jeste, D. V. (1988). Late-onset schizophrenia: An overview. *Schiz. Bull.* **14,** 39–55.

Harrison, P. J. (1999). The neuropathology of schizophrenia. A critical review of the data and their interpretation. *Brain* **122,** 593–624.

Heaton, R. K., Paulsen, J., McAdams, L. A., Kuck, J., Zisook, S., Braff, D. L., Harris, M. J., and Jeste, D. V. (1994). Neuropsychological deficits in schizophrenia: Relationship to age, chronicity, and dementia. *Arch. Gen. Psychiat.* **51,** 469–476.

Hutton, S., and Kennard, C. (1998). Oculomotor abnormalities in schizophrenia: A critical review. *Neurology* **50,** 604–609.

Jacobsen, L. K., and Rapoport, J. L. (1998). Research update: Childhood-onset schizophrenia. Implications of clinical and neurobiological research. *J. Child Psychol. Psychiat.* **39,** 101–113.

Portin, P., and Alanen, Y. O. (1997). A critical review of genetic studies of schizophrenia. I. Epidemiological and brain studies. *Acta Psych. Scand.* **95,** 1–5.

Portin, P., and Alanen, Y. O. (1997). A critical review of genetic studies of schizophrenia. II. Molecular genetic studies. *Acta Psych. Scand.* **95,** 73–80.

Tsuang, M. T., and Faraone, S. V. (1995). The case for heterogeneity in the etiology of schizophrenia. *Schiz. Res.* **17,** 161–175.

Weinberger, D. R., and Berman, K. F. (1996). Prefrontal function in schizophrenia: Confounds and controversies. *Philos. Trans. R. Soc. London B* **351,** 1495–1503.

Yolken, R. H., and Torrey, E. F. (1995). Viruses, schizophrenia, and bipolar disorder. *Clin. Microbiol. Rev.* **8,** 131–145.

Semantic Memory

MURRAY GROSSMAN and PHYLLIS L. KOENIG
University of Pennsylvania

GLOSSARY

aphasia Acquired disturbance of language that can include difficulty with semantic memory. Aphasia can be seen in the setting of a stroke or a neurodegenerative condition such as Alzheimer's disease. Transcortical sensory aphasia, for example, is a language disturbance that includes fluent speech that is often empty of content, associated with poor comprehension of language.

category-specific An impairment in meaning for a particular domain of knowledge, such as animals or tools, with minimal impairment of other semantic domains.

functional magnetic resonance imaging (fMRI) A technique used to study brain activation during cognitive challenges, such as understanding word meaning.

semantic memory The long-term representation of meaningful information and the processes we use to make this information available for a multitude of cognitive tasks.

Semantic memory is the long-term representation of meaningful information and the processes we use to make this information available for a multitude of cognitive tasks. Unlike episodic memory, we cannot necessarily recall the specific experiences associated with the acquisition of information represented in semantic memory. Semantic memory allows us to recognize a fork as a tined implement used for transferring food (without having to recall how we learned what a fork is) and to know that an apple is a sweet, red fruit with a pitted core that grows on trees in moderate climates (apart from our memory of the last time we ate one). The goal of current research efforts at several laboratories has been to establish the cognitive and neural bases for the representation of semantic knowledge and to define the set of cognitive processes and neural structures that support semantic memory in communication and thought. This article will summarize several advances in the area of semantic memory. We focus on results that have contributed to our understanding of the neural substrate for semantic knowledge and associated processes. It is important to recognize that these studies have been motivated by one of several theories of semantic memory, that theories of semantic memory play a crucial role in helping us interpret the results of studies investigating neural mechanisms associated with semantic memory, and that the results of such neural-based investigations in turn help constrain the nature of the cognitive models we develop for understanding semantic memory.

I. SEMANTIC MEMORY: A BRIEF OVERVIEW

In brief, semantic memory involves at least two key elements. First, there is the representation of semantic knowledge. This includes facts about the perceptual features (e.g., shape, color) and functional features (e.g., intended use of an implement, inherent activity of a beast of burden) associated with objects. We consider other facts beyond object recognition as well, such as the origins and biological properties of natural kinds such as ANIMALS (we use capitals to denote concepts) and FOODS and the range of perceptual variability displayed by manufactured artifacts such as TOOL and WEAPON, while still retaining the

essence of the object's meaning. Some of these features are relatively necessary components of a concept (e.g., apples grow on trees), whereas others are characteristically associated with a concept even if they are not necessary (e.g., many apples are red). Whereas we consider knowledge in semantic memory generally to be modality-neutral, allowing it to be represented visually, auditorially, or in any other fashion, there are certainly constraints on the manner in which some types of information can be represented. For example, semantic knowledge extends to nonobject concepts that are best represented propositionally, such as JUSTICE, or that depend on analog representations such as a visual image (e.g., RED). Semantic memory also includes actions, manners of thought, and emotions that are quite plastic in their manifestations and often entail relational information.

The mere existence of semantic knowledge is not sufficient to guarantee its effective use. The second key element of semantic memory involves the processes required to implement the contribution of semantic knowledge in our thoughts and actions. For example, we must be able to organize this vast array of knowledge for it to be used in a rapid and coherent fashion during thought and communication. Some of the properties of the concepts represented in semantic memory may cluster themselves in a categorical manner that groups like features and objects, but such "autoassociation" does not explain how the massive volume of our meaningful experiences coheres into concepts. Specific processes used for categorizing objects may help organize the immense amount of information about our meaningful experiences. One such process is thought to be "rule-based" and involves an analysis of a test object for the necessary and sufficient features of a concept; a second categorization process is based on "similarity" and involves a comparison of a test object with a prototype or with remembered instances of a concept. Moreover, we must be able to access and retrieve semantic knowledge, and this conceptual information must then be represented in a material-specific symbol system, such as writing or speech, for the purpose of communication. We also put semantic knowledge to many uses beyond an encyclopedic collection of facts for concept identification. For example, we make inferences about our world that are not readily apparent from the superficial appearance and function of an object, and we often acquire new knowledge on the basis of its relationship to established knowledge.

We are faced with the problem of mapping a semantic memory system such as this onto the brain, a 3-lb gelatinous mass composed of billions of neurons and a greater number of support cells. Broadly speaking, there are at least two approaches to the neural bases for semantic memory. First, there is a distributed account, in which the information in semantic memory is represented in a diffuse fashion throughout the superficial cortical gray matter of the brain. Several biochemical and microanatomic changes have been described during learning in simple organisms like aplysia that result in greater connective strength among neurons. Complex knowledge in semantic memory may be represented in the massively interconnected nature of neural elements bearing these microscopic changes. During learning, for example, we can imagine a specific network of connections between nodes representing particular features of a concept being facilitated by these microscopic changes. This allows the neural network to settle into a solution that represents the specific knowledge of a concept. From this perspective, a category—a collection of similar concepts such as FRUIT—may be a family of similar network solutions. The fact that a stable solution to a concept has been achieved is equivalent to the process of retrieving a concept. This hypothesis about the neural basis for semantic memory has been difficult to test directly, but researchers have attempted to simulate this distributed approach with computers using neural nets: computer simulations of cognitive functions that involve large arrays of interconnected nodes. These simulations are acknowledged by their developers to represent only a pale metaphor for the truly massive complexity of the central nervous system, but they represent an important start. Additional support for this approach comes from neuroimaging studies that fail to find distinct activation patterns for specific categories of knowledge.

A second approach to the neural basis for semantic memory hypothesizes the localized representation of semantic knowledge and semantic processes in specific parts of the brain. For example, the specific features of a concept in semantic memory may constrain the anatomic distribution of this concept in the brain because its representation must be processed by a particular modality (e.g., the visual representation of color concepts). Imaging tools such as positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) are available for directly studying the neural substrate of cognitive processes involved in semantic memory *in vivo*. These techniques allow us to image the gross spatial and topographic distribution of the brain working to solve a cognitive challenge, but these tools give us little insight into the

microscopic workings of human neural tissue when it is considering the meaning of an object. We have learned about the temporal characteristics of thought from cognitive event-related potential (ERP) studies. This technique uses powerful electrical detectors at the surface of the skull—and even directly on the cortical surface during the course of neurosurgery—to examine the spatial and temporal distribution of the electrical potentials generated by neurons during cognitive activity. The great achievement of high-resolution temporal information about semantic memory unfortunately comes at the cost of poorer spatial resolution. Studies of subhuman species have monitored neuronal activity directly with implanted electrodes, but comparable electrical studies in humans (as a prelude to the surgical management of intractable epilepsy) generally involve relatively primitive extracellular monitoring under highly restricted circumstances. New techniques such as magnetoencephalography and evoked-potential studies performed in the bore of a magnet during fMRI offer the potential of excellent temporal resolution with improved spatial resolution.

In other words, we are only at the very beginnings of our exploration of the neural basis for semantic memory, quite analogous to the European voyagers of the fifteenth century. These explorers hypothesized a vast world beyond their shores, and they forged ahead in the face of frequent false steps, poor tools, and misleading theories. Despite this grim picture, we are beginning to understand the gross lay of the land, as outlined in this article.

II. SEMANTIC KNOWLEDGE IN BRAIN-DAMAGED POPULATIONS WITH SEMANTIC MEMORY IMPAIRMENT

A. Investigations of Clinical Patient Groups

Much of the early work devoted to investigating semantic memory from a neural perspective focused on groups of patients with aphasia or a language disturbance following a stroke or other discrete cerebral injury. One subgroup of fluent aphasics—transcortical sensory aphasics—is said to be particularly impaired on tasks requiring semantic memory, such as word definition and picture naming. Their spontaneous speech typically is empty, with difficulty retrieving content words (e.g., the names of objects and actions), and is circumlocutory, with vague descriptions of the intended target word. These aphasics also substitute verbal or semantic paraphasias for many of the content words that they produce [e.g., for the intended target *egg* (we use italics to indicate a stimulus), they might produce "chicken" or "milk"], yet their utterances often appear to be grammatically intact. Evidence that this is not a deficit restricted to speech comes from the observation of similar difficulties in writing. Moreover, errors are found on comprehension measures that do not require expression, including word–picture matching, synonym judgments, and other tasks that require knowledge of word meaning. On the basis of these observations of oral and written difficulty with word meaning in comprehension and expression, researchers have concluded that patients with a transcortical sensory aphasia have a central deficit in semantic memory.

Whereas it is possible that these patients are impaired at some level associated with the concept underlying a word, an alternative account is that fluent aphasics have a language disturbance that is material-specific, that is, an impairment restricted to the language domain. This would be consistent with models of semantic memory that partition acquired knowledge into different, material-specific segments, including concepts that are said to be most amenable to representation in a propositional format (e.g., language-based semantic memory) and concepts with analog features that can be represented best as an image (e.g., visual-based semantic memory). Support for this material-specific model of semantic memory comes from observations of patients with a visual agnosia who appear to have a material-specific deficit for the meaning of visual material. Visual agnosics thus have profound difficulty understanding line-drawings of objects. These patients nevertheless have relatively intact elementary visual perceptual functioning that allows them to identify the location of a small dot on a page and even to specify its color. Visual agnosics often are able to recognize, copy, and match simple regular geometric shapes such as a square or a circle. A subgroup of visual agnosics—visual associative agnosics—are even capable of copying a complex line-drawing quite accurately, although this is accomplished in a slavish, line-by-line fashion that includes the unintended imperfections of the model. The material-specific nature of the deficit is emphasized by the fact that visual agnosics are relatively intact in their comprehension of objects presented linguistically and in other nonvisual modalities such as by hearing or feeling. Observations such as these suggest that all concepts represented in semantic memory are not compromised following focal cerebral injury, although

there may be selective difficulty understanding concepts that are particularly dependent on a specific type of material.

Transcortical sensory aphasics and visual agnosics may have material-specific difficulty understanding some representations of concepts, but other groups of patients are said to have difficulty with meaning that appears to be material-neutral in nature. Many of these patients have a dementia, that is, a progressive neurodegenerative disease that increasingly interferes with cognitive functioning over time. For example, Progressive Fluent Aphasia or Semantic Dementia is a neurodegenerative condition that results in fluent spontaneous speech with empty content but well-formed grammatical features that resembles transcortical sensory aphasia in many ways. Additional examination of these patients demonstrates profound difficulty on tasks such as picture–word matching, word sorting, picture sorting, and definition generation. They are limited on verbal and nonverbal measures assessing lexical and picture associations (e.g., Pyramid and Palm Trees Test, in which the subject decides which of two available choices, such as a *palm tree* and a *pine tree*, goes best with a target stimulus, such as a pyramid), and their coherence judgments of pictures composed of elements from two different objects (e.g., the head of a cow on the body of a monkey) also is compromised. Whereas these patients are relatively impaired in their knowledge of information at a basic object level (e.g., semantic knowledge associated with COW characteristics), their appreciation of superordinate category knowledge appears to be relatively preserved (e.g., knowledge that a cow is a kind of ANIMAL).

Similarly, some patients with Alzheimer's disease are thought to have semantic memory difficulty that transcends material and modality. As with transcortical sensory aphasia and Semantic Dementia, Alzheimer's disease results in fluent, grammatically well-formed speech that is empty and contains few content words. These patients also have difficulty on tasks that require semantic memory, such as matching pairs of words and pairs of pictures, judgments of anomalous pictures, category membership judgments for words and pictures, definition generation, and knowledge of object attributes.

Group studies such as these are immensely valuable because they demonstrate a consistent pattern of difficulty across a large number of similar patients, a process of generalization that adds confidence to our observations. However, there are several problems associated with investigations of semantic memory in a clinical group. One practical limitation is that it is difficult to test all members of a large group of subjects on an extensive series of measures. More importantly, there can be considerable variability within the population of patients bearing a similar diagnosis. Within Alzheimer's disease, assessments of individual patient profiles reveal an apparent semantic memory impairment for words as well as pictures in only about 50% of these patients. Different patterns of results thus can emerge in a group assessment, depending on the specific subgroup of patients constituting the cohort under study. For example, some investigators find greater semantic memory difficulty in Alzheimer's disease (AD) as a function of their impairment with the specific perceptual characteristics associated with a particular subset of words and objects, whereas other researchers AD associate the semantic memory impairment with a limited appreciation of the functional features associated with concepts. Our understanding of semantic memory through these group studies thus can be complemented by careful case studies of informative patients with discrete deficits, as will be discussed later.

B. The Neural Basis for Semantic Memory Impairment in Clinical Patient Groups

Observations of groups of similar patients with difficulty understanding and expressing meaningful material provide important confirmation that semantic memory can be compromised following central nervous system disease. But what is the neural basis for semantic memory? Although we can speculate about the architecture of a model of semantic memory—whether we are more accurate to use an information processing model or a parallel distributed processing model, for example—the accuracy of our models can be significantly enhanced by direct investigations of the neural basis for semantic memory. Moreover, information about the neural basis for semantic memory can help constrain the nature of a cognitive model of semantic memory. This two-way interaction is a keystone of cognitive neuroscience.

The process of studying brain–behavior relationships in semantic memory has been significantly advanced by the advent of modern neuroimaging techniques. Rather than waiting for a patient to die, as was the case until relatively recently in the history of neuroscience, we can immediately observe the locus of cerebral insult with any of several modern

neuroimaging techniques. For example, injury to the posterior superior temporal and inferior parietal regions of the left hemisphere appears to play a crucial role in the difficulties of patients with a transcortical sensory aphasia. Imaging studies of patients with visual agnosia emphasize bilateral dysfunction of visual association cortex in the context of the relative preservation of primary visual cortex and posterior superior temporal regions of the left hemisphere. Studies such as these suggest that there may be distinct neural substrates associated with representing concepts propositionally and visually.

Group assessments also have been performed that directly correlate a semantic memory deficit with a focus of neurological dysfunction. In a study of Alzheimer's disease, for example, patients were asked to judge the membership of words and pictures in a category: When given a picture of an apple or the word "apple," for example, patients decided whether the test item was a FRUIT. Judgment accuracy correlates with regional cerebral activity in the superior temporal and inferior parietal regions of the left hemisphere, and Alzheimer's disease patients with significant difficulty making these judgments have significantly less activity in this area than patients who are successful at making these judgments. This corresponds to the distribution of increased cerebral activity of healthy subjects during some functional neuroimaging studies of category membership judgments. Imaging studies also associate Semantic Dementia with disease in portions of the left temporal lobe. For example, analyses of the distribution of structural changes on MRI demonstrate significant atrophy in the anterior portions of the left temporal lobe in these patients, whereas functional neuroimaging with PET during the performance of an association task (Pyramids and Palm Trees Test) shows reduced activity in the posterior inferior portion of the left temporal lobe. This corresponds to the locus of cortical recruitment in healthy subjects while performing this association task. The investigators conclude that a disconnection within the left temporal lobe contributes importantly to the breakdown of semantic memory in patients with Semantic Dementia. Findings such as these appear to underline the contribution of various regions within the left temporal lobe to semantic memory. However, these studies leave open the question of exactly what role these brain areas play in semantic memory. Is it the facts represented in semantic memory that are degraded by disease, or is it the processes making use of these facts that are compromised? We address this issue later.

It is convenient to assess information processing models of semantic memory in stroke patients because the focal cerebral insult caused by a stroke may interfere with a particular component of semantic memory such as a specific process or a specific kind of semantic knowledge. A distributed approach to modeling semantic memory has been proposed as well, but this kind of theory is difficult to test in patients suffering from focal neurologic injury. Investigators thus study dementing patients from the perspective that their illness causes diffuse damage to the distributed network underlying semantic memory. According to this approach, the semantic memory deficit in Alzheimer's disease is due to nonfocal damage that loosens axonal projection patterns and degrades the synaptic densities of interneuronal connections in a massively complex neural network. As the disease progresses, the neurodegenerative process results in greater semantic memory difficulty due to continuous degradation of the neural network supporting semantic memory. However, evidence supports the claim that many patients with a dementia have a focal neurodegenerative disease that disproportionately compromises the functioning of specific cortical regions.

Yet another source of information is available for assessing the neural basis for semantic memory—a highly focal technique whose results surprisingly support the distributed approach to semantic memory. Specifically, many patients requiring surgical treatment for pharmacoresistant epilepsy have direct electrocortical stimulation obtained during the intraoperative evaluation prior to the removal of neural tissue. This allows the neurosurgeon to map out "eloquent" cortical regions and avoid damage to the patient's language system. These stimulation studies suggest that components of language systems are widely distributed throughout the frontal and temporoparietal cortices of the left hemisphere. For example, electrocortical stimulation in a wide area of the left hemisphere—including frontal, temporal, and parietal regions—disrupts naming. To be sure, the patients examined during these studies have sustained significant brain injury, leading to the need for a surgical cure for their intractable epilepsy. This can have a significant impact on the representation of language in the brain. These findings nevertheless contrast with the localization studies described earlier and emphasize the possibility that the neural basis for word meaning that is associated with naming is distributed in a relatively undifferentiated fashion throughout a large area of the left hemisphere.

C. Category-Specific Semantic Memory Impairment: Selective Deficit for Natural Kinds or Manufactured Artifacts

Recent findings in groups of patients with apparent semantic memory difficulty discussed earlier raise important questions about the nature of semantic memory and its neural representation. One model of semantic memory proposes that specific types of visual and linguistic semantic knowledge can be selectively compromised. The limits of this compartmentalized approach to semantic memory are tested with detailed case studies of patients who have highly selective semantic memory impairments. In particular, semantic memory deficits are detailed that are putatively category-specific because they are associated with relative difficulty in only one domain of knowledge. These observations prompt widespread excitement and controversy because of their potentially important contribution to a theory of semantic memory and the neural representation of the semantic memory system.

Whereas observations of selective difficulty understanding specific categories of information have been noted for many decades, Warrington and her colleagues published several case studies that provide the path-breaking impetus for our renewed interest in this focal approach to the neural representation of knowledge in semantic memory. These patients have relative difficulty on tasks requiring knowledge of natural kinds such as ANIMALS, VEGETABLES, and PLANTS, although performance on comparable tasks involving manufactured artifacts such as TOOLS, VEHICLES, and WEAPONS is relatively preserved. For example, these patients provide impoverished definitions of living things (e.g., *daffodil*: ''plant;'' *duck*: ''an animal'') compared to their definitions of manufactured artifacts (e.g., *compass*: ''tool for telling the direction you are going;'' *umbrella*: ''object used to protect you from water that comes''). The patients of Warrington and Shallice, as well as additional patients described since this seminal report, also have greater difficulty on measures of word–picture matching, picture naming, naming to definition, attribute judgment, and category naming fluency (e.g., ''name as many different ANIMALS as you can in 1 minute'') for stimuli involving natural kinds, compared to manufactured artifacts.

The basis for this intriguing deficit is the subject of intense investigation, and several questions have since arisen regarding the reliability of these findings. For example, one objection is that the stimuli used to test notions about natural kinds and manufactured artifacts are not carefully matched for frequency, familiarity, visual complexity, and other ''nuisance'' variables that can potentially affect performance. For example, a case of herpes encephalitis has been reported in which the patient demonstrates a clear category-specific deficit for living things compared to manufactured artifacts. When the lists representing each category are matched for frequency, the category-specific advantage for living things remains. However, when the stimulus lists are jointly matched for frequency, familiarity, and visual complexity, the category-specific advantage disappears. To be sure other cases are reported in which the category-specific effect remains despite matching lists of stimuli jointly for frequency, familiarity, and visual complexity. Attention to details such as this nevertheless helps underline the variety of subtle individual differences that exist across the cases said to have a category-specific semantic memory impairment.

Another more substantive issue related to case-by-case variability is the precise nature of the categories that are compromised. We often think of terms such as ''living things'' and ''natural kinds'' to be coreferential, for example, but there appear to be subtle and potentially important differences in the scope of the deficit across individual patients with a category-specific impairment. A relatively large number of patients thus have a category-specific deficit for natural kinds that includes both categories of ANIMALS and FOODS. Other patients may be impaired for the category of FOODS but not ANIMALS, or for ANIMALS but not FOODS. Similar controversies arise regarding the status of artifact categories such as MUSICAL INSTRUMENTS and BODY PARTS. This kind of variability makes it difficult to identify a theoretically coherent account that is explanatory across the spectrum of individuals with a category-specific deficit for natural kinds.

Despite these case-by-case differences, some theorists relate category-specific impairments to the model of semantic memory that differentiates between visual semantics and propositional semantics mentioned earlier. From this perspective, conceptual knowledge is organized in the brain according to the sensory (visual, auditory, kinesthetic, etc.) and motor modalities that contribute features to the concept. This model holds that the same neural elements responsible for perceiving an object's attributes are also partially responsible for their long-term neural representation. The various properties of an object are said to be distributed across these sensory–motor

representational domains, and the conceptual representation of an object is an "autoassociated" collection of its attributes. Warrington and Shallice thus argue that the concepts underlying natural kinds hold in common their dependence on visual properties, for example, whereas the concepts associated with manufactured artifacts are more dependent on functional properties. From this perspective, the visual sensory attributes of natural kinds are crucial for identifying and distinguishing between items such as strawberries and raspberries or between zebras and horses, although these pairs of items are not readily distinguished in terms of their functional attributes. On the other hand, a screwdriver and an ice pick look very similar visually, just as a minibus and a van closely resemble each other, but these pairs of items are said to be readily distinguished on the basis of their functional characteristics.

This theory of semantic memory receives some support from computational models. For example, one simulation represents semantic knowledge in separate but interconnected networks of visual and functional attributes, and the ratio of visual to functional properties in living things is set to be much greater than in nonliving things. The ratios are derived from subjects who inspect dictionaries for the relative numbers of visual and functional attributes associated with living things and nonliving things. When Farah and McClelland simulate a brain lesion by damaging the visual components of their model, the resulting pattern of performance resembles that seen for a category-specific impairment for living things. Conversely, when they lesion the functional components of their model, a category-specific impairment for nonliving things emerges. Caramazza and Shelton as well as other investigators have objected that the instructions for culling information from dictionaries does not reasonably ascertain the frequency of the nonsensory properties of objects. When Caramazza and Shelton replicate the Farah and McClelland experiment by rephrasing the instructions for identifying functional attributes to include all nonsensory properties, they find a ratio of sensory to nonsensory properties that approaches 1:1 and is essentially identical for living things and nonliving things.

This sensory–motor model for semantic memory is put to the test in patients with a category-specific impairment. For example, knowledge of natural kinds and manufactured artifacts in patients with a category-specific impairment can be probed for visual attributes and functional attributes. Evidence to support the sensory–functional model would come from the finding that patients with a category-specific impairment for natural kinds have relatively impoverished knowledge of the visual attributes of these objects compared to their functional knowledge of the same objects. Indeed, a patient with a category-specific impairment for living things is described with disproportionate difficulty identifying the visual feature that distinguishes between two objects. However, these investigators use animal stimuli that are less familiar than their artifacts. Another reported case of category-specific impairment for natural kinds has much more difficulty with visual than with functional properties of natural kinds, and these are more difficult than both visual and functional properties of artifacts. However, this pattern of difficulty is evident only when the patient is probed verbally: The patient is considerably less impaired when asked to evaluate the visual properties of visually presented objects. For example, the patient has difficulty deciding verbally whether an elephant is orange-colored or gray-colored, but he or she is virtually error-free when asked to judge the coherence of pictures of an orange-colored elephant and a gray-colored elephant. Yet other investigators object that patients with a category-specific impairment for natural kinds have equal levels of difficulty with visual and functional probes of natural kinds and are equally superior in their responses to probes of visual and functional properties of manufactured artifacts.

Another concern about the category-specific impairment for natural kinds is that natural kinds are somehow more difficult or less familiar than manufactured artifacts. However, Warrington and others describe patients with the inverse pattern of impairment, namely, greater difficulty with manufactured artifacts compared to natural kinds. As with category-specific difficulty for natural kinds, patients with selective difficulty for manufactured artifacts are less than consistent from case to case: One case has difficulty predominantly with small manipulable objects, whereas another patient has difficulty with artifacts of any sort (e.g., TOOLS, VEHICLES) as well as BODY PARTS.

Difficulty with manufactured artifacts is attributed to an impairment in appreciating the functional properties of objects that are thought to be so crucial to the semantic feature representation of these objects. However, studies focusing on the functional properties associated with manufactured artifacts are difficult to interpret, possibly because functional attributes have a much more complicated mental organization than perceptual attributes. For example, there appear to be

functional attributes that are common only to natural kinds (e.g., BREATHES, EATS) or manufactured artifacts (e.g., STAPLING, CARVING), there are additional functional features that help distinguish among natural kinds (e.g., EXTRACTS OXYGEN FROM WATER or FROM AIR) or among manufactured artifacts (e.g., CUTS BY SAWING or CHOPPING), and there are distinct action schemata for expressing a function.

Some evidence to support dissociations among these elements of function is seen in patients with impaired semantic memory. For example, patients with difficulty appreciating fine-grained functional features that distinguish among natural kinds nevertheless are able to appreciate functional features common to all natural kinds. In another study, a double dissociation is found between knowledge of functional attributes in a Semantic Dementia patient and errors of object selection and usage on a naturalistic action test in a patient with impaired executive function.

Given the many concerns about the sensory–motor account for category-specific impairments, several alternative models have been forwarded to account for semantic memory deficits in patients with a category-specific deficit. One approach suggests that category-specific deficits cannot be reduced to sensory and functional limitations, but that categories of knowledge themselves are directly compromised. Patient E.W. has a category-specific impairment for natural kinds that narrowly affects the category of ANIMALS. Her difficulty spans naming, visual and auditory recognition, and visual and functional statements about animals. On the basis of these observations as well as those of other investigators, a domain-specific basis for the organization of conceptual knowledge in the brain is proposed. In this view, the categories of ANIMALS, FRUITS AND VEGETABLES, and ARTIFACTS can be damaged independently of each other due to an evolutionary adaptation of dissociable neural circuits dedicated to responding selectively to instances of these domains.

In contrast to this sensory–motor approach based largely on an information processing model, an alternative theory focuses on the distributed nature of the characteristics contributing to semantic knowledge. According to this model, there is a high intercorrelation among features of natural kinds, allowing collateral support from other features and similar objects even if there is some degradation of natural kind knowledge. For example, many fruits are round, and this is highly correlated with other features of fruits such as color, size, and taste. Even if the round shape feature of an orange is destroyed, the semantic memory system can "infer" on the basis of the high intercorrelations among fruit features that oranges are round. By comparison, knowledge of each manufactured artifact is said to be relatively more dependent on a unique feature that does not have much collateral support, and destruction of the neural representation of this feature will consequently lead to reduced knowledge of this object. Following the destruction of the blade feature of a saw, for example, the semantic memory system cannot easily infer from other manufactured artifacts that a saw must have a blade for cutting because this feature is relatively unique to saws.

This distributed model can be assessed in patients with Alzheimer's disease. Mildly demented patients are found to have greater impairment for manufactured artifacts than for natural kinds. The investigators argue that random destruction of neural elements representing specific manufactured artifact features results in a decline in knowledge of these objects, but an equal burden of disease with randomly destroyed natural kind features has less of an effect on natural kind knowledge. The authors attribute this finding to the higher intercorrelation among natural kind features that allows collateral support from other natural kinds. By comparison, more severely demented patients have more difficulty with natural kinds than manufactured artifacts. The investigators attribute this to the progressive level of damage sustained by the brain that is sufficient to interfere with the feature system of natural kinds. This hypothesis is implemented in a connectionist model that emphasizes the distributed nature of correlated features in natural kind and manufactured artifact categories.

D. Neural Basis for Category-Specific Semantic Memory

What can we learn about category-specific impairments from direct studies of the brains of these patients? Can such studies discriminate between the localized, information processing approach and the distributed, nonlocalized approach that are proposed to account for these intriguing deficits? By focusing on differences in the statistical distribution of features associated with natural kinds compared to manufactured artifacts, some investigators demonstrate distinct patterns of relative impairment in cross-sectional studies and in a limited longitudinal study of Alzheimer's disease. This disease condition is examined because the investigators claim that these patients have

diffuse, nonfocal disease. In essence, these researchers assert that the disturbed connectivity pattern among categories of objects—rather than the anatomic locus of disease—defines the nature of category-specific representation of semantic information in the brain and explains the deficit in Alzheimer's disease.

A major difficulty for this approach is that Alzheimer's disease appears to have a specific anatomic distribution of disease. Thus, other assessments of the category-specific deficit in Alzheimer's disease directly challenge the distributed account of higher intercorrelation of features among natural kinds compared to manufactured artifacts. These investigators find the same pattern of cognitive impairment described by Gonnerman *et al.*: greater difficulty with artifacts in mildly impaired patients, but greater difficulty with natural kinds in moderately impaired patients. However, they attribute the distinct impairment patterns to the specific anatomic distribution of neuronal dysfunction seen at different stages of Alzheimer's disease.

An alternative hypothesis proposes that category-specific difficulty is attributable to a specific, structural locus of disease. Patients with category-specific difficulty for natural kinds often sustain damage to inferior temporal lobe structures. A common cause of disease in this location is herpes simplex encephalitis, although other patients with a category-specific impairment for natural kinds suffer from a neurodegenerative condition or head trauma involving this brain area. By comparison, most patients with a category-specific deficit for manufactured artifacts have larger lesions such as a stroke affecting lateral aspects of the left hemisphere in the distribution of the middle cerebral artery.

Whereas this approach to localizing a category-specific impairment is highly appealing because of its direct, one-to-one mapping of a category onto a neural substrate, several shortcomings are associated with this approach. The most important of these is the difficulty in establishing the precise locus and extent of disease, particularly when disease quantification is ascertained only with a structural imaging study. For example, diseases causing structural lesions such as a stroke often are indiscriminant in the extent of damage to gray matter, and there may be insult extending into functionally unrelated cortical regions that are interfering in a "nuisance" fashion with some aspect of task performance. Further, damage to the underlying white matter may interrupt projections that have little to do with the cortical area of concern, resulting in a remote functional defect that confounds the localizing value of the lesion.

More recently, functional neuroimaging studies of healthy adults have significantly aided our analysis of the neural basis for category-specific semantic knowledge. Functional neuroimaging studies have the distinct advantage that regional recruitment of the brain in response to a cognitive challenge can be studied in a healthy, neurologically intact individual. Perhaps the most elegant study involves the examination of patients with lexical retrieval difficulty who have focal insult involving the left temporal lobe, compared with the assessment of healthy adults exposed to natural kinds and manufactured artifacts while regional cerebral blood flow is monitored with PET. These investigators define the damaged neural substrate common to a series of patients with selective difficulty naming manufactured artifacts—the ventral midtemporal region of the left hemisphere—and the brain region compromised in patients with a category-specific deficit naming natural kinds—the posterolateral midtemporal region of the left hemisphere. In their PET study challenging healthy subjects with natural kinds and manufactured artifacts, these investigators observe a distinct distribution of activation associated with each category that corresponds to each of the damaged areas responsible for a category-specific deficit. The involvement of visual association cortices in these category-specific phenomena is interpreted to support the sensory–functional account for the representation of knowledge in the brain.

Additional support for the sensory–motor model comes from a PET study that monitors regional cerebral activity while subjects judge whether pairs of line-drawings are different exemplars of the same kind of object or are two different objects. These investigators find that subjects recruit bilateral inferior temporal regions when animals are presented, but they activate a left hemisphere network including left dorsolateral prefrontal cortex when artifacts are presented. Another PET study directly investigates the pattern of regional cerebral recruitment associated with sensory and functional characteristics. These investigators find recruitment of left ventral temporal regions during naming of a color associated with a word or a picture representing an object, but they observe activation of left frontal and temporal–parietal areas during naming of an action associated with a word or a picture of an object. In sum, observations of brain-damaged patients with a category-specific impairment and functional neuroimaging studies of healthy adults challenged with specific categories of knowledge provide some support for the hypothesis that the neural representation of semantic

information is organized in a fashion that is sensitive to the cortical region processing the features critical to the mental representation of a specific category. It remains to be determined whether the neural representation of information within these regions is distributed or topographically localized.

E. Other Forms of Category-Specific Difficulty: Action Verbs, Proper Nouns, and Abstract Nouns

Many investigators argue that the category-specific difficulties evident in studies of natural kinds and manufactured artifacts provide important insights into the neural representation of knowledge. However, these hopes must be tempered by the fact that this theory of category-specific knowledge is based on concrete objects such as natural kinds and manufactured artifacts, which represent only one kind of knowledge, and categories such as these have relatively unique sensory and functional properties that may not generalize to other domains of knowledge. This note of caution is emphasized by the fact that we use terms with concrete reference relatively rarely in our day-to-day conversations: our speech consists of words like "word" and "example," for example, much more often than "apple" and "screwdriver."

A handful of studies has begun to test the hypothesis that category-specific difficulties can be found in other domains of knowledge. Consider actions, a domain of knowledge that is associated with a different major word class—verbs—and has different kinds of perceptual and functional features. Verbs are not static, visuoperceptual entities, for example, but instead are dynamic, relational concepts with highly malleable visual perceptual manifestations that have an important motor-related component. Some evidence has emerged that patients with disease in motor-related cortices, such as in frontotemporal forms of degeneration, are relatively impaired with action verbs compared to concrete nouns. Observations such as this raise the possibility that domains of knowledge labeled by verbs also have a distinct, category-specific neural representation and potentially provide important evidence that the sensory–motor model of category-specific knowledge is not simply a special case relevant to specific kinds of objects or parts of speech.

However, there are several confounds that must be addressed before this line of thinking receives unequivocal support. For example, verbs have a rich family of associated grammatical features (e.g., whether the verb takes a direct or indirect object), and it may be these grammatical attributes that determine in part the neural representation of verbs. In this context, a considerable body of work demonstrates that the left prefrontal regions of the brain play a special role in grammatical aspects of comprehension. Another issue has to do with the stimuli that the patients are asked to name. For the most part, static line-drawings are used to illustrate the verbs, and it may be that the relative difficulty inferring an action from such a representation plays a role in the verb-specific impairment of these patients. Indeed, there appear to be brain regions in the posterolateral temporal–occipital area that are specialized for the perception of motion that may play an important role in action verb knowledge. In comparison to generation of a color term associated with pictures and words denoting object stimuli, generation of an associated action term results in the recruitment of posterolateral temporal–occipital cortices. Another issue has to do with the sheer amount of information represented in a verb compared to a noun. The coordination of processes underlying verb use may depend in part on working memory resources that are supported by prefrontal cortices. In separate subgroups of patients with a frontotemporal form of dementia, for example, a dual-task study of word–picture matching with verbs and nouns finds that patients with a nonfluent form of progressive aphasia have greater difficulty with verbs than nouns regardless of the condition during which verb comprehension is assessed. However, frontotemporal dementia patients with executive difficulty but no aphasia have difficulty with verbs depending on the simultaneous performance of a resource-demanding secondary task.

Some evidence has accumulated to indicate that frontal brain regions contribute specifically to action verb comprehension. In a study assessing regional cerebral activity during subjects' judgments of words from various verb categories, motion verbs recruit left inferior frontal as well as right striatal regions in comparison to judgments of cognition verbs (e.g., mental state verbs like "discuss"), whereas the reverse contrast shows recruitment of left posterolateral temporal regions. Both verb categories also recruited right ventral temporal, bilateral temporal–occipital, and striatal regions. A direct contrast of motion verbs and cognition verbs demonstrated specific activation of inferior frontal regions for motion verbs.

Another category with its own neural representation may be proper nouns. These words have a concrete referent—the person, geographic location, or object that each proper noun names—and thus resemble common nouns such as natural kinds. However,

proper nouns differ from common nouns such as natural kinds because of the arbitrary relationship that exists among the set of features associated with the referent. Several cases are reported with selective proper noun impairment. These patients are not able to produce the names of famous individuals in response to a picture or to verbal description, even though they are able to use the same phonologic shape accurately to name a picture or a description of a common object. The reverse pattern is reported as well, that is, greater difficulty with common nouns than proper nouns. This suggests that proper nouns are not simply more difficult to name and understand than common nouns, but that proper nouns may represent a distinct category of knowledge.

The neural representation of proper nouns is quite controversial. Damasio and his colleagues describe seven patients who have selective difficulty in naming with proper nouns. These subjects have disease that affects the left temporal pole. These investigators also find that healthy individuals recruit anterior portions of the left temporal lobe when naming with proper nouns. However, a summary of the lesion localization of patients with proper name difficulty reveals widespread insult in many areas of the left hemisphere. Whereas the neural representation for proper nouns remains unclear, all of these investigators interpret their results in the context of an information processing model that localizes proper nouns to some specific brain region.

A stronger challenge to the sensory–motor model comes from studies of the neural basis for abstract words. These words have few sensory–motor features and thus are difficult to link to a specific brain region concerned with processing visual perceptual or motor–kinesthetic information. Nevertheless, the concepts underlying abstract words are readily understood due to their propositional content. Perhaps the most intriguing evidence concerning the psychological reality of abstract concepts comes from the demonstration of the ''reversal of the concreteness effect'' in patients with Semantic Dementia. Whereas we typically understand concrete nouns more rapidly and accurately than abstract nouns, these patients are more rapid and accurate at understanding and naming abstract nouns.

The neural basis for abstract nouns has been investigated with functional neuroimaging studies as well. One study examines the distribution of neural activation associated with the passive presentation of abstract nouns or concrete nouns, using as a baseline an ''anticipation'' condition that involves the presentation of instructions without the actual presentation of stimuli during imaging. This study reports activation in left middle temporal, left inferior frontal, and left occipital cortices in association with abstract nouns. However, a similar pattern of activation in left middle temporal, left inferior frontal, and left middle and inferior occipital regions is seen during passive presentation of concrete nouns. Another fMRI study reports left posterolateral temporal and left prefrontal activation during abstract noun stimuli as well as IMPLEMENTS. On the basis of observations such as this, the sensory–motor approach to category-specific knowledge does not appear to account fully for the neural representation of semantic memory because similar brain regions are recruited for concrete and abstract nouns, despite minimal sensory and motor attributes associated with the latter. One possibility is that these brain regions are activated to make decisions about any type of word, regardless of its content, rather than because the neural representation of knowledge represented in specific categories is represented in this portion of neocortex. In sum, some evidence supports the claim that category-specific knowledge is topographically distributed in the brain, depending on the nature of the category. However, important questions remain to be resolved before this model can be fully endorsed.

III. PROCESSING COMPONENTS OF SEMANTIC MEMORY

The studies reviewed previously provide considerable evidence to support the claim that some form of category-based organization of long-term semantic knowledge is honored by the brain, although the precise basis for the category-specific form of neural organization remains to be established. Is this determined by the sensory–motor features that must be recruited in association with a concept? Is the category-based organization due to some genetically determined predisposition that grants specific categories a special status? Or is there some other variable remaining to be uncovered that will encompass abstract concepts as well as the categories of knowledge with sensory–motor features that have been so well-studied? Regardless of the ultimate outcome of this debate, concerns about the representation of categories of knowledge in the brain rarely address a crucial issue in our understanding of semantic memory, namely, how these concepts are *processed*. For example, it may be that the distinct status of natural

kinds is determined only in part by the neural representation of knowledge of the visual properties of these objects, but natural kinds additionally may be related to the visual–perceptual processing of a concept that is recognized by its visual attributes or to some other aspect of processing these concepts.

Some investigators suggest that patients with Alzheimer's disease do not have an impairment in the neural representation of semantic knowledge. Instead, Alzheimer's patients may have difficulty accessing long-term semantic knowledge that is otherwise relatively intact. This assertion is based on the results of priming studies. In this work, a semantically related priming word facilitates performance on a lexical decision task or a timed reading task when the priming word immediately precedes the target word, but this facilitation does not occur when the preceding word is unrelated to the target word. For example, the speed at which we initiate our oral reading of the word "doctor" is more rapid when preceded by the semantically related priming word "nurse" than by the semantically unrelated foil "purse." This phenomenon is equally robust in some studies of patients with Alzheimer's disease, suggesting that their semantic knowledge is fundamentally preserved. This line of investigation leads to the conclusion that the semantic deficit in Alzheimer's disease is due to an impairment in the effortful processes needed to access and retrieve information from semantic memory.

Whereas this phenomenon is reported in several studies of patients with Alzheimer's disease, support for this approach is far from universal. For example, some investigators demonstrate hyperpriming in Alzheimer's disease, that is, word recognition in a priming paradigm that is even more rapid than in control subjects. Moreover, the specific items with which Alzheimer's disease patients have the greatest hyperpriming are the items that the same Alzheimer's patients have the greatest difficulty understanding in nonpriming studies. It is hypothesized that there may be "pruning" of semantic relationships in an associative network, thus leading to very fast priming in the setting of sparse semantic associative knowledge that may otherwise "distract" patients from the associative link.

An alternative account has been forwarded. This considers the nature of the semantic relationship being tested in priming studies. The high associative strength between prime–target pairs of words often used in priming studies is not necessarily semantic in nature, but may be due to a high frequency of lexical co-occurrence that is not necessarily semantic. For example, "cottage" and "cheese" are words that co-occur frequently but have no true semantic relationship that binds them together. Compare these pairs with words that have a low frequency of co-occurrence but a high degree of semantic relatedness, such as members of the same superordinate semantic category like "rooster" and "canary." A priming study examines pairs of words that co-occur at a high frequency but have a low index of semantic associativity and word pairs that are semantically related by virtue of their membership in a common superordinate category but rarely co-occur. The findings demonstrate that Alzheimer's disease patients prime for word pairs that co-occur at a high frequency, but not for semantically related word pairs.

Functional neuroimaging studies can examine the neural basis for retrieving semantic knowledge. This work demonstrates that inferior frontal as well as temporal regions of the left hemisphere are recruited during tasks assessing the retrieval of semantic knowledge. Thompson-Schill and her co-workers attempt to establish the role of this inferior frontal recruitment by varying the selection demands across three semantic tasks. They observe changes in inferior frontal recruitment as a function of task demands, associating inferior frontal recruitment with the retrieval process involved in semantic memory.

There is another, more fundamental way in which a processing impairment can interfere with the comprehension of word meaning. If information about concepts is distributed in the brain, then there must be some way in which this information is synthesized into a coherent whole that can be useful to us in our communication with others and in our internal mental musings. Allport and others allude to this issue when they discuss the "autoassociation" between elements of an object that are distributed across many modality-specific segments of conceptual knowledge. However, the nature of this autoassociative process remains to be defined, and no independent evidence supports the claim that autoassociation in fact is the mechanism by which elements of a concept are synthesized into a coherent whole representing an object. Studies of categorization in fact propose at least two distinct ways in which we organize semantic information into coherent categories. One process for organizing distributed information about semantic memory is said to be *rule-based* in nature. According to this approach, there are specific attributes that are generally necessary and sufficient for identifying an instance of a category, and an object must have these attributes in order to be a member of the target category. This is an analytic,

effort-demanding process that requires us to hold in mind a list of attributes that are selected to be necessary features of a concept and to verify the presence of each and every one of these attributes in a test object while nonnecessary features are inhibited. A second process for organizing semantic knowledge is said to be *similarity-based* in nature. According to this approach, a global comparison based largely on perceptual features is made between a test object and an ideal instance (or remembered instances) of the category. If the similarity between the test object and the ideal instance is close enough, then the test object is said to be a member of the category.

One study attempts to test the integrity of categorization processes such as these by asking patients to judge the category membership of familiar objects taken from well-defined or nominal categories such as FEMALE and MALE, or from categories with poorly defined or fuzzy boundaries such as VEGETABLE and FRUIT. A subgroup of Alzheimer's disease patients has a significant impairment in deciding the category membership of test items from fuzzy categories. Among these patients with a deficit pattern said to be typical of a semantic memory impairment, there is no deficit in category decisions about items taken from well-defined categories. Similarity-based categorization and rule-based categorization thus are equally effective at supporting decisions about well-defined categories, but similarity-based categorization can be misleading and unreliable for decisions about fuzzy categories. Therefore, we appear to be relatively dependent on rule-based categorization for poorly defined categories. These findings thus are thought to be most consistent with the hypothesis that AD patients have a rule-based categorization impairment that selectively limits their judgments of objects taken from categories with fuzzy boundaries.

A neuroimaging study assessed the neural representation of similarity-based and rule-based processes. These investigators taught subjects about categories of artificial animals using one of two strategies: memorizing instances of each category (i.e., similarity-based categorization) or using a set of rules for determining category membership (i.e., rule-based categorization). Subjects subsequently were scanned while they attempted to generalize their knowledge to a new set of these artificial animals. The results revealed a different distribution of neural activation depending on the process used initially to learn category membership. The subjects asked to memorize the instances of each category recruited visual association cortices, consistent with the hypothesis that these subjects are making perceptually based similarity judgments between a test object and a remembered instance. The subjects using a rule-based approach also activated some frontal regions, consistent with the contribution of executive processes to rule-based categorization decisions. These findings suggest that the categorization processes implicated in organizing knowledge for the purpose of understanding concepts are honored by the brain and appear to be supported by dissociable neural systems.

IV. SUMMARY

Where have we come by the end of this brief review? It should be apparent that more questions remain to be answered than have been addressed at this early stage of investigation. The barest of outlines nevertheless is beginning to emerge from the haze to indicate that the semantic memory system has a store of knowledge that sorts meaningful experiences in a systematic fashion and that several kinds of processes are available to make use of this information. This body of work begins to demonstrate the way in which cognitive investigations of semantic memory can constrain our interpretation of neurologically based investigations, and direct assessments of the neural basis for semantic memory can help develop a model of the way in which the brain supports our representation and use of concepts. We can be hopeful, moreover, that new, *in vivo* techniques are continuing to emerge that will help us assess the neural basis of semantic memory during life with increasing detail.

We believe that the data discussed earlier are consistent with a model of semantic memory in which information is represented in a distributed network. Each concept contains information describing at least perceptual and functional features that are stored in modality-specific, sensory–motor association cortices, as well as propositional knowledge possibly in language-related brain regions. This massive network of information is organized with the aid of at least two processes: rule-based categorization implemented in part in frontal heteromodal association cortex that uses working memory to process the necessary features defining a concept, and similarity-based categorization implemented in part in posterior heteromodal association cortex that evaluates the graded resemblance between a stimulus and an ideal (or remembered) representative of the concept. Concepts are retrieved so that they can be represented by a phonological, orthographic, or visual structural description that is

processed in the appropriate modality-specific association regions.

See Also the Following Articles

ALZHEIMER'S DISEASE, NEUROPSYCHOLOGY OF • ANOMIA • APHASIA • MEMORY DISORDERS, ORGANIC • MEMORY, EXPLICIT AND IMPLICIT • MEMORY NEUROBIOLOGY • NERVE CELLS AND MEMORY • SHORT-TERM MEMORY • WORKING MEMORY

Acknowledgments

This work was supported in part by funding from the U.S. Public Health Service (AG15116, NS35867, and AG17586) and the American Health Assistance Foundation.

Suggested Reading

Breedin, S. D., Saffran, E. M., and Coslett, H. B. (1995). Reversal of a concreteness effect in a patient with semantic dementia. *Cogn. Neuropsychol.* **11,** 617–660.

Caramazza, A., and Shelton, J. R. (1998). Domain-specific knowledge systems in the brain: The animate–inanimate distinction. *J. Cogn. Neurosci.* **10,** 1–34.

Damasio, H., Grabowski, T. J., Tranel, D., Hichwa, R. D., and Damasio, A. R. (1996). A neural basis for lexical retrieval. *Nature* **380,** 499–505.

Farah, M. J., and McClelland, J. L. (1991). A computational model of semantic memory impairment: Modality specificity and emergent category specificity. *J. Exp. Psychol.: Gen.* **120,** 339–357.

Garrard, P., Patterson, K., Watson, P. C., and Hodges, J. R. (1998). Category specific semantic loss in dementia of Alzheimer's type: Functional–anatomic correlations from cross-sectional analyses. *Brain* **121,** 633–646.

Gonnerman, L. M., Andersen, E. S., Devlin, J. T., Kempler, D., and Seidenberg, M. S. (1997). Double dissociation of semantic categories in Alzheimer's disease. *Brain Lang.* **57,** 254–279.

Grossman, M., Koenig, P., DeVita, C., Glosser, G., Alsop, D., Detre, J., and Gee, J. C. (2002). The neural representation of verb meaning: An fMRI study. *Human Brain Mapping.* **15,** 224–234.

Grossman, M., Payer, F., Onishi, K., D'Esposito, M., Morrison, D., Sadek, A., and Alavi, A. (1998). Language comprehension and regional cerebral defects in frontotemporal degeneration and Alzheimer's disease. *Neurology* **50,** 157–163.

Hart, J., and Gordon, B. (1990). Delineation of single word semantic comprehension deficits in aphasia, with anatomical correlation. *Ann. Neurol.* **27,** 226–231.

Martin, A., Haxby, J. V., Lalonde, F. M., Wiggs, C. L., and Ungerleider, L. G. (1995). Discrete cortical regions associated with knowledge of color and knowledge of action. *Science* **270,** 102–105.

Mummery, C. J., Patterson, K., Wise, R. J. S., Vandenbergh, R., Price, C. J., and Hodges, J. R. (1999). Disrupted temporal lobe connections in semantic dementia. *Brain* **122,** 61–73.

Ojemann, G. A. (1991). Cortical organization of language. *J. Neurosci.* **11,** 2281–2287.

Semenza, C. (1997). Proper-name-specific aphasias. In *Anomia: Neuroanatomical and Cognitive Correlates* (H. Goodglass and A. Wingfield, Eds.), pp. 115–136. Academic Press, San Diego.

Smith, E. E., Patalano, A., and Jonides, J. (1998). Alternative strategies of categorization. *Cognition* **65,** 167–196.

Tyler, L. K., and Moss, H. E. (2001). Towards a distributed account of conceptual knowledge. *Trends Cogn. Sci.* **5,** 244–252.

Sensory Deprivation

JOSEF P. RAUSCHECKER
Georgetown University Medical Center

GLOSSARY

amblyopia Reduced vision not correctable by glasses (thus of central origin), which can be caused by various forms of visual deprivation at an early age.

cerebral cortex Outer sheet of the mammalian brain consisting of several dozen functionally specialized areas, each consisting of vertical columns with various functional properties as well as six horizontal layers with connections to other cortical areas, to the thalamus, and to other subcortical nuclei.

compensatory plasticity Improved performance of a sensory system in response to sensory deprivation in another; one system compensating for the loss of function in the other.

deafferentation Severing of afferents to a particular region of the central nervous system.

enucleation Surgical removal of an eye.

Hebbian synapses Synaptic connections that are strengthened as a function of coincident activation in pre- and postsynaptic neurons.

monocular–binocular deprivation Visual deprivation in one or both eyes.

representational plasticity Shrinkage or expansion of brain maps as a result of sensory deprivation or training.

sensitive period Time period of postnatal development during which a brain region or perceptual function is particularly sensitive to the effects of sensory deprivation (analogous to "critical period:" the time period of postnatal development during which exposure to adequate stimuli is particularly critical for normal development).

sensorineural hearing loss Hearing loss due to the destruction of hair cells in the inner ear, often associated with aging.

synaptic competition Competition between synapses for space at their common target neurons, which is controlled by activity in the respective pathways, with active synapses winning over silent ones.

tinnitus Ringing or buzzing in one or both ears, presumably of central origin, but usually initiated by peripheral damage.

Sensory deprivation studies consider the role that sensory experience plays in the development, reorganization, and plasticity of the brain and behavior by withholding or restricting that experience. This article will review animal studies in higher vertebrates in which sensory deprivation is brought about experimentally as well as some studies in congenitally blind or deaf humans. Most of this work assesses the effects of deprivation with either neurophysiological–anatomical or neuroimaging techniques, both within the same sensory system and across systems. The overarching role that neuronal activity plays in the modification of individual synapses and distributed representations is discussed.

I. INTRODUCTION

The question of whether nature or nurture determines our sensory and perceptual abilities is as old as human culture and goes back to Plato and Aristotle. An Egyptian pharaoh purportedly separated identical twins at birth in order to study the nature–nurture question, and Kaspar Hauser became a household name in the eighteenth century when he was found after having been deprived of language from birth. Whereas the role of nature is usually subsumed under the terms "innate" or "genetic," nurture has many different connotations. The most common one implies a role of sensory experience or the active use of an

individual's sensory systems. Withholding of that experience by sensory deprivation thus directly tests the influence of the nurture component on sensory development. The answer to the question of how important nurture is for brain development is likely to influence our thinking about the importance of education and environmental enrichment in the upbringing of our children. These issues therefore are often politically loaded, and results of research on these topics are hotly debated to this day.

Before modern neurobiological techniques became available, an assessment of the effects of sensory experience or deprivation was restricted to the behavioral level. To measure the effects of sensory experience on the brain, the first anatomical studies on visually deprived animals were performed in the nineteenth century. When Hubel and Wiesel and their contemporaries took up the theme of sensory deprivation in the last third of the twentieth century, electrophysiological single-unit recording, supplemented with more refined neuroanatomy, became the method of choice. Most recently, biochemical and molecular biological techniques have been applied to these questions as well.

At the single-neuron level, the effects of sensory deprivation can be observed by changes in sensory input pathways and their central representations. For example, temporary loss of pattern vision early in life can lead to permanently reduced vision because of neural changes in the brain's representations of vision. In normal animals, neurons are highly selective for certain properties of sensory stimuli, depending on the brain region in which these neurons are found. Neurons in the visual cortex, for instance, are "tuned" to the orientation of visual bars or gratings; neurons in the primary auditory cortex are tuned to tones of particular frequencies. Rearing of kittens in a striped environment or with strong cylindrical lenses that blur vision in all orientations but one leads to a distortion in the number of visual cortical neurons tuned to particular orientations. Thus, deprivation within a parameter domain, such as orientation selectivity, leads to changes of tuning within that domain. An altered performance of the neural system leads to corresponding changes in behavioral performance.

One of the major effects of sensory deprivation on brain circuitry is a loss of neural connections. Neural circuitry in newborn animals is overdefined and is pruned during development in a use-dependent manner. Sensory experience plays an important role in this selection process by feeding highly specific neural activity into the system. By contrast, sensory deprivation reduces this activity and therefore renders the brain circuitry less selective. Sensory experience is also sometimes capable of inducing the growth of new connections ("sprouting"), and sensory deprivation prevents this from happening. Whether one looks upon the role of sensory experience as maintaining neural selectivity that is otherwise lost or as actively instructing specificity depends on one's perspective. The amount of selectivity found at birth ("innately") varies across species and brain regions. Therefore, one may look at the effects of sensory deprivation as supporting both nativist and empiricist views of brain development.

Whereas sensory deprivation often has a detrimental effect on circuitry within the deprived modality, cross-modal effects have also been reported in which other, nondeprived modalities can actually improve their performance, so as if to "compensate" or substitute for the loss of the other modality. The existence of such "compensatory plasticity" or "sensory substitution" has been much debated over the years but now seems well-established. Experimental work presently focuses on the mechanisms governing such compensatory changes that can obviously be highly beneficial to the individual. It appears that the same synaptic mechanisms may be invoked cross-modally that govern adaptive changes within a single modality. The activity-dependent modification of synaptic connections originally proposed by D. O. Hebb in 1949 for learning-related changes in brain connections seems also to be the underlying mechanism for cross-modal effects.

II. VISUAL DEPRIVATION

A. Animal Studies

One of the ingenious ideas of Wiesel and Hubel in their studies of visual deprivation was the use of a reversible technique that permitted later testing of the deprivation effects on the visual system through the very end organ through which the animals had been deprived. By sewing the eyelids of young kittens or monkeys together before the time of natural eye opening, input to the visual system was prevented, but the optics of the eyes and their sensory receptor surface remained unharmed. With most deprivation studies in other sensory modalities (e.g., auditory or somatosensory, see later discussion), irreversible deafferentation of the sensory pathways is the only practical way to investigate the effects of sensory deprivation.

1. Effects on Retina and Lateral Geniculate

Whereas early studies of visual deprivation denied any effect on the physiology of the retina or lateral geniculate nucleus (LGN), later scrutiny found some subtle changes. Anatomically, changes in cell size are found in the layers of the LGN after monocular deprivation, with cells in the nondeprived layers being larger than those in the deprived layers. However, these changes were shown to be retrograde changes resulting from binocular competition at the cortical level. This was demonstrated first by the fact that cells in the monocular segments of the deprived layers, which do not have to compete with cells from the other eye, were of normal size. In a second series of experiments, an artificial monocular segment was created in the LGN by making a defined lesion in one retina with a laser (thus removing binocular competition specifically for this retinal location). In these cases, a shrinkage in cell size in the deprived laminae of the LGN was observed only in those sections that had an intact binocular representation and not in those areas that corresponded to a part of the visual field not represented in the other eye. Thus, it is the binocular competition of these projections at the cortical level that leads to retrograde changes in LGN cell size, not the direct effect of visual deprivation.

2. Effects on Visual Cortex

a. Monocular and Binocular Deprivation Extensive effects of visual deprivation during a critical period of visual development have been demonstrated on the primary visual cortex (V1, A17, striate cortex) of kittens or infant monkeys. Experimentally, visual deprivation can be achieved by occlusion with opaque contact lenses or by lid suture, which completely abolishes patterned visual input and drastically reduces the luminance on the deprived retina.

Although profound effects on neural response selectivity were reported after binocular deprivation (see later discussion), very robust consequences for the binocularity of cortical neurons are found by withholding normal visual input to only one eye (monocular deprivation). After just a few hours of monocular deprivation at the height of the critical period (in kittens around 1 month of age), many neurons in the primary visual cortex showed reduced excitability when tested through the previously deprived and then reopened eye. After progressively longer deprivation periods, the effects become more and more pronounced until, with about 1 week of monocular deprivation, virtually all cortical neurons (which normally receive binocular input) are driven only through the experienced eye. This means that, as a result of monocular deprivation, the input from the deprived eye to the cortical neurons becomes disconnected. These effects are reversible when, during a restricted sensitive period, the deprived eye is reopened and, at the same time, the experienced eye is closed. Simple reopening of the originally deprived eye is usually ineffective, which is interesting from a mechanistic point of view (“competition,” see earlier discussion), but also has practical consequences for the treatment of amblyopia in human infants (see later discussion).

The anatomical correlates of monocular deprivation are equally striking. Cells in the striate cortex of rhesus monkeys are organized into a band- or stripelike pattern of ocular dominance columns that are normally of equal width for either eye. Monkeys with long-term lid suture in one eye are found to have ocular dominance columns that strongly favor the nondeprived eye by expanding its representation at the expense of the deprived eye and show only small islands of input from the latter.

Shutting of both eyelids (binocular deprivation) during visual development does not create an ocular input asymmetry to the cortex, and the effects of binocular deprivation therefore cannot be measured by determining a shift in ocular dominance. However, its effects on other physiological response properties of visual cortical neurons are similarly dramatic: general responsiveness to visual stimuli drops (i.e., excitability thresholds rise) and the tuning of neurons for pattern-selective properties, such as orientation and direction selectivity, suffers drastically. Effects similar to binocular lid suture can be found with rearing kittens in total darkness.

Monocular and binocular deprivation meanwhile have become the most robust paradigms for the testing of changes in neural connectivity of the brain. Monocular deprivation is a standard bioassay for the synaptic mechanisms of cortical plasticity. Pharmacological and biochemical procedures for the development of new drugs are using the synaptic changes during monocular deprivation as their main testing ground, and the primary species for testing these effects are now rodents.

b. Pattern-Selective Deprivation The manipulations described so far have been rather crude. Closing of an eye lid or rearing in darkness deprives an animal of total pattern vision. Vision comprises many aspects such

as color, motion, contrast, spatial frequency, and depth. Experimental designs to withhold some specific aspect of normal visual experience while leaving the others more or less intact have originated from an understanding about the parameter domains within which cortical neurons are "tuned." These include orientation and motion-direction selectivity. The assumption behind such "selective-rearing" (or "selective-deprivation") studies was that neural networks encoding each selective property require experience in that particular domain to either establish or confirm their selectivity.

i. Orientation-Selective Deprivation: The earliest attempts to influence the development of orientation selectivity were made in the early 1970s and continued for more than a decade. Various techniques were used with the aim of rearing kittens with visual input restricted to only a narrow range of orientations. One group reared their kittens in striped drums, similar to what had been used earlier in behavioral studies of sensorimotor development. Another group used goggles that had contours of one orientation carved onto their semitranslucent occluders. A third group used cylindrical lenses, which blur vision in all orientations except for one narrow sector, and this sector remains constant relative to retinal coordinates regardless of head motion. The main outcome of all of these studies was that more cortical neurons were found with preferences for orientations that the animals had experienced. Thus, it seems that orientation tuning in visual cortical neurons requires sensory experience to maintain or fully develop that selectivity.

ii. Direction-Selective Deprivation: Selective deprivation of direction-selective mechanisms in the visual cortex has been performed in two ways: (1) Rearing of kittens in an environment without visual motion by using stroboscopic light. This procedure dramatically increases the number of neurons without direction selectivity. (2) Striped drums that move in only one direction have been used successfully to bias the distribution of direction-selective neurons in the visual cortex of kittens.

iii. Deprivation of Depth Vision: Various methods have been used to interfere with the normal means of binocular interaction in the visual cortex. Cutting of an eye muscle leads to strabismus or squint, i.e., a divergence or convergence of the visual axes. If strabismus is induced in young kittens or monkeys, the result is a reduction in binocularity of cortical neurons. This is caused by a loss of the normal binocular facilitation that occurs when the eye axes are aligned. Similar effects are found when kittens are reared with alternating occlusion of the two eyes.

Rearing of kittens in artificial environments that are moving toward or away from the animals causes visual association areas involved in the processing of optic flow to develop abnormally. Few other attempts have been made to deprive higher visual areas of their normal visual input, but the expectation is that adequate stimulation is required for normal development or maintenance of visual functions and that cortical areas with different specializations may possess different critical periods.

3. Cross-Modal Effects of Visual Deprivation

Behavioral observations of visually deprived animals demonstrate that they are highly adept at a variety of behaviors, including the localization of objects. Uninformed observers could not infer the blindness of these animals. This suggests that other sensory systems compensate for the loss of vision. In neural terms, the increased usage of these other systems (relative to sighted animals) may lead to an enlarged and refined representation of these senses in the brain of visually deprived (blind) animals.

a. Auditory Compensation for Visual Deprivation

i. Behavioral Evidence: Auditory spatial testing after visual deprivation has been performed in two nonhuman mammalian species: cats and ferrets. Ferrets are particularly useful because their brains are highly immature at birth. Both sets of data demonstrate increased auditory spatial acuity (i.e., decreased sound localization error) in early blind animals, binocularly lid-sutured from birth. The most pronounced effects were found in lateral and rear positions of azimuth, where the differences to sighted controls were highly significant.

ii. Neural Changes: Neurophysiologically, the most extensive studies on cross-modal effects of visual deprivation have been performed in cats and monkeys. Changes have been observed in the superior colliculus (SC) and in the parietal and occipital cortices. Overall, an increased number of auditory neurons compared to normal controls was found in visually deprived animals. In the SC, a higher incidence of auditory neurons was found in intermediate and deep layers, but what appeared to be new auditory neurons were also occasionally found in the superficial layers, where normally only visual cells are present. In the cortex of visually deprived monkeys, cross-modal changes were found primarily in association regions (area 7 in parietal cortex and areas 18 and 19 in occipital cortex).

In cats binocularly deprived from birth, visual responses of area AEV in the fundus of the anterior ectosylvian sulcus (AES) (the probable homolog of parietal area 7) virtually disappeared. Neurons in this region, however, did not become unresponsive, but were replaced by neurons with brisk responses to auditory and tactile stimuli. Apparently, auditory and somatosensory areas within the AES had expanded at the expense of formerly visual territory.

The response properties of the expanded auditory ectosylvian area (AEA) and those of neighboring auditory fields in the AES region were homogeneous. Auditory spatial tuning (the tuning for the location of a sound source in free field) was significantly sharper in the whole AES region when compared to sighted controls. Visually deprived cats had close to 90% spatially tuned cells (with a spatial tuning ratio of better than 2 : 1 between best and worst location). In addition, neurons with spatial tuning ratios of 10 : 1 or better were more abundant in blind cats. The increased number of auditory cortical neurons, together with their sharpened spatial filtering characteristics, is likely to improve the sampling density of auditory space and provide the neural basis for the improved spatial abilities of early blind cats and ferrets.

b. Tactile Compensation for Visual Deprivation - Cross-modal changes also occur in the somatosensory system of visually deprived cats and rodents. Visually deprived cats grow longer and thicker facial vibrissae. The same is true for mice and rats in which both eyes were removed (binocularly enucleated) at birth. Binocularly enucleated mice also show an expansion of the whisker barrels, the neural representation of the vibrissae in the primary somatosensory cortex. Barrels corresponding to whiskers in lateral positions show the most significant expansion, and whiskers in these same positions show the greatest hypertrophy, thus increasing the lateral range of the vibrissae as a tactile "organ." Increased usage and stimulation of the whiskers seem to be the common cause of both processes, but different signals may be responsible on the two levels.

Whereas the representation of nonvisual modalities in the cortex expands, the total visual cortex of rhesus monkeys after binocular enucleation is reduced in size and contains smaller cell bodies. Taken together, these findings suggest that cross-modal compensatory plasticity is guided by the same principles of activity-dependent synaptic competition as unimodal forms of neural plasticity, where active representations expand at the expense of inactive ones.

B. Human Studies

1. Visual Deprivation and Amblyopia

Visual deprivation in humans due to congenital cataract, if not corrected early in life, has the same devastating consequences as deprivation due to lid suture in cats and monkeys: adequate stimulation is withheld from cortical neurons and will inevitably lead to the same negative effects on neural selectivity. At the behavioral level, the neural deprivation effects can be recognized by reduced visual acuity that is uncorrectable through optical means (even if the cataract is removed or bypassed with laser optics). This deficit is characterized by the term "deprivation amblyopia" (amblyopia = weak vision).

Congenital cataracts that can cause deprivation amblyopia, previously not uncommon due to eye infections and oxygen toxicity for premature babies in incubators, have become quite rare in the Western world because of increased monitoring and even more rarely remain uncorrected due to improved surgery techniques. There are other forms of visual deprivation, however, that still pose a threat to normal vision and whose neural bases are less well-understood. Anisometropic amblyopia, which is caused by unequal refractive power in the two lenses, is related to deprivation amblyopia, because the unfocused image generated by one eye reduces the responses of cortical neurons in ways similar to those of diffuse light.

The most common form of amblyopia, strabismic amblyopia, is caused by a deviation of one eye axis from its regular position and a subsequent mismatch of the two retinal images at their corresponding cortical locations. The missing binocular facilitation also leads to reduced firing of cortical neurons and a resulting disconnection of visual inputs from their target neurons. Furthermore, a desynchronization of the inputs from the two eyes results in the loss of stereopsis and reduced visual acuity in the amblyopic eye.

2. Compensatory Plasticity in Blind Humans

As in visually deprived animals, blind humans have been reported to compensate for their loss of vision by increased use of their remaining senses and by correspondingly improved capacities in these senses. In the following, evidence in favor of these observations will be reviewed.

a. Auditory Compensation for Early Blindness *i. Behavioral Evidence*: Early studies of compensatory

plasticity in blind humans yielded controversial results. Some studies found evidence for an improvement in nonvisual functions after early blindness, whereas others found just the opposite: a deterioration in hearing compared to sighted controls. Several factors contributed to the confusion in the field: inhomogeneous patient populations with diverse etiology and unknown neurological status [partly due to the unavailability of objective tests such as magnetic resonance imaging (MRI)], different ages of onset and duration of blindness, and small numbers of patients combined with possible experimenter bias.

Studies with large patient populations have been undertaken that tested subjects with similar histories under stringent conditions. None of these studies found a disadvantage for the blind in their sound localization abilities, and most showed them to be superior. Most interestingly, another study found patients with partial vision to be the worst of all three groups (fully sighted, completely blind, and partially sighted) at localizing sounds. The same study provided valuable hints as to the neural basis for the improvement in spatial tuning in the blind: the biggest improvement was found when monaural spectral cues had to be used for sound localization. Other studies found the greatest improvements in sound localization in blind humans for lateral and rear positions of azimuth, precisely as had been demonstrated in visually deprived cats.

ii. Auditory Neuroimaging Studies in Blind Humans: Modern imaging techniques permit the mapping of neural activity during auditory stimulation in blind and sighted subjects directly in the human brain. Studies using event-related potentials (ERP) have demonstrated that the extent of cortical activation by changes in the frequency, intensity, and location of a sound is expanded in blind people, and the center of gravity of cortical activation is shifted posteriorly toward occipital areas. Comparison of patients blind from birth with those who became blind later in life reveals that a posteriorly directed expansion is also found in late blind patients, although it is smaller in extent than the shift found in the early blind. This expansion in late-blind adults implies at least partial cross-modal plasticity in the adult human, which is consistent with the behavioral and neurophysiological findings in visually deprived cats described earlier.

Investigations with positron emission tomography (PET) comparing congenitally blind and sighted subjects showed massive activation of the occipital cortex in the blind during a sound localization task in virtual auditory space. The sounds were presented via headphones, and the spatial cues were simulated using standardized head-related transfer functions, which present monaural spectral as well as binaural cues. Other PET studies have shown that the occipital cortex in blind subjects has metabolic rates that are as high as during visual stimulation in sighted controls. This is compatible with the notion that the occipital cortex in the blind is not idle but is recruited for functions other than vision.

Localization of sounds in virtual auditory space by sighted subjects leads to the activation of specific foci in the inferior parietal lobule (IPL) and in the frontal cortices, with a bias toward the right hemisphere as measured by PET. The same foci light up in blind subjects but are vastly expanded toward parieto-occipital (area 7) and occipital locations (areas 18 and 19), which normally are not active during sound localization, and the right-hemisphere bias is even more pronounced. The areas of expansion are probably homologous, in part, to higher visual areas shown to be newly activated by auditory or somatosensory stimuli in visually deprived cats and monkeys (see previous discussion).

In neuroimaging studies, regional activation alone, expressed as increased activity relative to a control state, does not necessarily indicate functional involvement. Interregional correlation analysis between activated brain regions provides additional information regarding the function of the right occipital cortex during auditory localization in the blind. By using the right IPL as a reference region, correlation analysis reveals a functional network of connections involving inferior and posterior parietal and occipital areas of the right hemisphere. Comparison of blind and sighted subjects revealed significantly greater interregional correlations in the blind between the right IPL and the right parieto-occipital (areas 7–19), right peristriate (BA 18), and right superior temporal cortices (area 22). This suggests that auditory signals from the temporal areas reach the occipital (formerly visual) cortex via parietal and parieto-occipital areas. Whether this involves the formation of new or the strengthening of existing connections remains to be elucidated. The animal work from binocularly deprived cats would suggest that the original cross-modal expansion happens in a competitive fashion in parietal cortex and that the back-projections from parietal to occipital cortex already carry an enhanced auditory signal.

b. Human Studies of Tactile Compensation Interest in the question of whether blind humans develop

enhanced tactile capacities has been great at least since the days of Louis Braille, mainly because of the practical implications of such potential improvements. Early behavioral studies in the tactile domain have been as controversial as those described earlier for the auditory domain. It is very difficult to control for practice effects, because very few sighted individuals would devote as much time and effort to working with Braille script. Thus, even those studies that demonstrated an improvement in spatial acuity in blind Braille readers often attributed the improvement to training effects, i.e., greater opportunity to practice Braille reading.

PET studies have begun to shine some light on whether blind people have improved tactile capacities at least in neurobiological terms. The activation of occipital areas by tactile Braille reading has been demonstrated to be higher in blind Braille readers than in normally sighted people trained to read Braille. Transcranial magnetic stimulation (TMS) of the occipital cortex disrupts the ability to recognize Braille characters. Even a case of "Braille alexia" has been reported, in which a blind subject became unable to decipher Braille after a stroke in the occipital cortex. The tactile consequences of focal damage to the visual cortex demonstrate unequivocally that the expanded region of cortex actually participates in the processing of tactile information in blind people. It appears that somatosensory regions normally participating in this task have expanded into formerly visual territory. How these newly invading somatosensory and auditory inputs coexist in occipital cortex and share this territory remains to be elucidated.

A more philosophical problem concerns the question of how blind individuals "see" their world. Does the activation of occipital cortex by auditory and somatosensory stimuli cause a "visualization" of the world? Similarly, does this activation evoke an "image" in the blind that can somehow be compared to vision because it is mediated by a brain region that was designed to process visual information? Or does the altered input change the function of these areas so fundamentally that they, in essence, turn into an extension of auditory or somatosensory cortex and provide sensations equivalent to those generated by these areas? In other words, is the perceptual and cognitive function determined by the recipient structure or by its inputs? The visual cortex in blind individuals certainly participates in auditory and somatosensory perception, but the studies reviewed here do not provide a final answer about the qualities of the experience.

C. Synthesis of Human and Animal Data: Mechanisms and Loci

Whereas many similarities between the animal and human responses to deprivation exist, as described earlier, one major difference stands out: an expansion of the auditory-responsive cortex in visually deprived cats and monkeys was reported only for higher association areas. By contrast, ERP and neuroimaging studies of blind humans additionally demonstrate a vast expansion of auditory activation into the occipital cortex, corresponding to primary and secondary visual areas. The explanation for this seeming discrepancy may be quite simple, however. Auditory responses were never tested in the primary occipital areas of blind cats and monkeys, because it seemed unlikely that auditory input could expand so far into normally visual territory. Cross-modal expansion was thought to be limited to neighboring areas that already had multimodal overlap and where competition could occur between overlapping input from different modalities. In light of the PET data, a possible reexamination of this view appears to be warranted. The invasion of new areas with minimal functional input from the auditory modality would at first imply a different mechanism from cross-modal competition. On the other hand, multimodal overlap may exist even in occipital cortex, especially during early postnatal stages when occipital areas still receive input from auditory cortex, as has been reported in newborn kittens. Such transitory input may be stabilized in blind animals and humans. Another, perhaps very likely, possibility is that nonvisual activation of the occipital cortex simply reflects the changes in the modality of the feedback projection from multimodal association cortex, in particular parietal cortex, where the initial modifications have taken place on the basis of cross-modal competition.

III. AUDITORY DEPRIVATION

A. General Remarks

Compared to the vast number of experimental studies performed in the visual system, sensory deprivation of the auditory system has received much less attention. This is surprising because a lack of normal auditory experience in early childhood is known to have profound and permanent effects on hearing. Most importantly, normal speech can only be acquired with auditory feedback. Likewise, sound localization

depends to a large extent on experience and can be modified by it. The extraordinary precision with which this system operates can only be accomplished with tuning mechanisms that recalibrate the system continually, especially during the growth phase of the head and outer ears.

At subcortical levels, there have been relatively few experimental studies of auditory deprivation, learning, and plasticity, but even less work has been performed on deprivation effects in the auditory cortex. To be sure, auditory deprivation cannot be accomplished as easily nor as fully as visual deprivation. Plugging an external ear canal can, at best, achieve an attenuation of sound by 20–30 dB. Rearing in controlled or restricted auditory environments is difficult, although some studies have shown that altered activity patterns during development can indeed reduce the frequency tuning of central auditory neurons. Most other methods for creating auditory deprivation, such as partial deafening with antibiotics or loud sound exposure, exert their effects by destroying the sensory receptor cells in the inner ear. However, the receptors need to be intact if the influence of deprivation on the central auditory system is to be tested. This is different from deprivation of the visual system by lid suture, which leaves the sensory end organ intact.

B. Effects of Auditory Deprivation on Speech and Language

As stated earlier, the literature does not provide such an abundance of examples of experimentally induced auditory deprivation studies in animals as it does for visual deprivation. However, there are plenty of examples of ''natural deprivation'' in humans in which modified or incomplete experience with the natural auditory environment leads to permanent deficits in hearing and, consequently, in speech. Because of their practical importance, we will deal with these studies first.

1. Specific Language Impairments and the Influence of Early Phonetic Environment

Congenital deafness or severe hearing impairment in early childhood leads to irreversible deficits in the perception and production of speech. Even transient elevations in hearing thresholds, such as during otitis media, if frequently recurring, may lead to permanent deficits. Although genetic influences may play a role as well, specific language impairments have been attributed to deprivation in the domain of rapid transients in the acoustic speech signal, which are important carriers of information. Language-impaired children often have severe perceptual deficits for brief tones when they are masked by noise. The fact that auditory feedback is essential for the development of normal vocal production is also suggested by studies of the song system in birds. Deafening at an early age leads to permanent deficits.

The phonetic environment in which a child is reared during early language acquisition undoubtedly has a profound influence on what he or she can or cannot perceive later in life. As such, this selective experience can also be seen as selective deprivation. Indeed, the inability of many speakers of Asian languages, in particular Japanese, to recognize distinct sounds for ''r'' or ''l'' is probably due to the absence of the distinguishing features in their early auditory environment. Once acquired through early exposure, such a predisposition can only be reversed through intensive training. Infants in different countries (Sweden, Russia, and the United States) initially do not show a preference for phonemes unique to their own language. By about 6 months of age, however, they suddenly develop this bias. Language-specific preferences for prosodic cues, which are necessary for the segmentation of the speech stream into perceptual units, also develop between 6 and 9 months of age.

Specific experience during a critical period of early postnatal development also seems to be necessary for the development of better-than-average musical skills. Early musical training in children is closely related to the development of absolute pitch and a concomitant expansion of auditory cortex. The conclusion is not far-fetched, therefore, that lack of such training (or partial deprivation) leads to an impoverished or less than fully developed auditory cortex. Experience-dependent plasticity for the perception of harmonic sounds is greatest before the age of 8 or 9, which coincides with the critical period for development of absolute pitch. This is in striking parallel with findings on phonological development, which also demonstrate that foreign accents in a second language invariably develop if the latter is acquired after the age of 8.

C. Cross-Modal Effects in Deaf Subjects

Processes analogous to auditory and tactile compensation for sensory deprivation in the blind can be demonstrated in congenitally deaf subjects. It can be shown with event-related potentials (ERP), as well as with neuroimaging techniques, that the brains of deaf

subjects are reorganized profoundly. Visual motion areas in the right parietal cortex thought to be involved in the initial decoding of visual motion in American Sign Language (ASL) are expanded. At the same time, "auditory" areas in the superior temporal (ST) cortex also become activated by sign language, but not by the presentation of English words as in hearing subjects. However, areas responsible for the processing of language-specific contents in the anterior ST and inferior frontal (Broca's) regions develop normally, even though they must be fed through a different input system. Hearing children of deaf parents, whose first language is ASL but who also develop normal English skills, do not show the same deprivation effects. Neurobiologically speaking, the mechanisms responsible for cross-modal reorganization during visual or auditory deprivation must be very similar. As has been argued earlier, the neural mechanisms are even likely to be the same for reorganization within and across modalities.

D. Effects of Auditory Deprivation on Sound Localization

Localization of sounds by the brain requires either the evaluation of binaural cues (intensity or time differences between the two ears that depend on the angle from which the sound comes) or monaural spectral cues (that are created by the filter characteristics of the head and pinnae). Some of these cues are already extracted at the brain stem level and are carried on into higher areas of the auditory cortex, where they create neurons with pronounced spatial tuning and possibly specialized representations for the perception of auditory space. As the complete deprivation of sound is close to impossible, most studies have sought to perform a selective deprivation of certain parameters to determine their influence on the development of binaural or spectral cues.

Ligation of one ear canal in rats leads to latency increases in neurons in the inferior colliculus (IC). After an ear had been deprived of sound from 10 days after birth, response latencies of high-frequency (but not low-frequency) units in the contralateral IC were 2–3 times higher than normal. This result is intriguing because high-frequency units in the IC are thought to participate in the coding of interaural intensity differences (IIDs), one of the major binaural cues for the coding of auditory space. Ears sound-deprived later in life were also associated with changes in latencies at the IC, but these were much less than in those deprived from birth, which hints at the existence of a critical period. Latencies of gross potentials at the auditory nerve were not affected by early deprivation, indicating a central origin for the latency changes.

Some of the most interesting experiments on the plasticity of sound localization have been performed in barn owls, which show exceptional precision in this domain. Plugging of one ear causes a change in the IID, which is one of the main binaural cues to sound source azimuth. Ear plugging leads to systematic behavioral sound localization errors, which can be predicted from the experimental change in the IID. The map of auditory space in the superior colliculus, which is considered the neural basis for sound localization behavior in owls, is systematically distorted in the direction of the changed IID. More specifically, implantation of an acoustic filtering device that alters both the timing and the level of sound reaching one eardrum in a frequency-dependent fashion induces frequency-specific adjustments in the auditory spatial tuning of collicular neurons. These results demonstrate that the response properties of higher-order auditory neurons in the optic tectum of the barn owl can be adjusted during development to reflect the influence of frequency-specific features of the binaural localization cues experienced by the individual. Most of the effects are reversible during a critical period, similar to that for the effects of monocular deprivation in the visual cortex.

The importance of auditory experience for the development of sound localization in humans is underscored by the observation that familiarity with a sound's spectrum makes it easier to localize the sound in elevation, and filtering of a sound in specific ways can bias its perceived elevation. This is due to the fact that elevations are encoded by notches in the frequency spectrum created by the head and pinnae. These notches seem to become associated with certain elevations by learning-induced mechanisms. The comparison with a stored sound template works best if the sound is familiar. If the spectrum is not known, the brain seems to work under the assumption of a flat frequency distribution, which leads to larger errors.

E. Central Effects of Peripheral Lesions

Small lesions in the cochlea of the inner ear (in a restricted region of the frequency axis that is represented along the basilar membrane) lead to a missing frequency representation in the contralateral auditory cortex. Input from neighboring frequencies expands

into the vacated cortical region that would normally process the frequencies from the lesioned part of the cochlea. A similar frequency expansion is not observed in the brain stem, which indicates that the reorganization occurs more centrally, perhaps as central as the auditory cortex itself.

The same mechanisms may lead to tinnitus, a sensation of ringing or buzzing in one or both ears, after sensorineural hearing loss due to aging or loud noise exposure. A particular frequency representation, deafferented by a peripheral lesion, is taken over by input from neighboring frequencies, which leads to an overabundance of activity from these frequency ranges. Human brain mapping studies using magnetoencephalography (MEG) and PET provide evidence that such cortical reorganization and the resulting frequency shift do indeed occur in tinnitus patients. Support for a central origin of tinnitus also comes from the fact that tinnitus persists in patients with acoustic neuroma even after transection of the auditory nerve. Furthermore, studies using 2-deoxyglucose autoradiography in gerbils treated with salicylate (which is known to generate tinnitus) demonstrate reduced activity in the inferior colliculus, but increased activation in portions of the auditory cortex.

Very similar reorganization of central representations after small peripheral lesions is found in the visual system. Laser lesions of the retina in animal studies lead to filling in of the resulting scotoma; neglect of the blind spot of the eye is the perceptual counterpart in human perception. Both phenomena can be considered to be analogous to tinnitus.

Reorganization of the auditory cortex after changes in peripheral activity is also observed, in a more positive fashion, after chronic auditory stimulation of a particular frequency region in owl monkeys. This leads to an expansion of the corresponding frequency representation in the auditory cortex and an enhanced discriminability of the corresponding frequencies. These findings underscore the idea that cortical representations are dynamic, continually shaped by sensory experience and deprivation, and that these dynamic map changes form the basis for experience-dependent behavioral effects, in both a positive and a negative sense.

IV. SOMATOSENSORY DEPRIVATION

The conclusions drawn in the preceding section on map changes in auditory plasticity apply in analogous ways to the effects of somatosensory experience and deprivation on the organization and reorganization of the somatosensory cortex. Partial denervation in monkeys by transection of peripheral nerves or at the level of the dorsal roots in the spinal cord leads to shrinkage in the corresponding cortical field representations and expansion of neighboring points, whereas increased peripheral stimulation causes the expansion of the corresponding representations. Just as in the auditory system, most work on somatosensory deprivation has been done on the effects of peripheral deafferentation. It is just as difficult to deprive the somatosensory system of its natural stimulation as it is to selectively deprive the auditory system, in contrast to the visual system in which this can be done so effectively.

One of the few classic examples in which somatosensory deprivation can be accomplished without deafferentation or neural injury is the clipping of whiskers in rodents. The effects of this procedure on the "barrel cortex," the representation of the whiskers in the somatosensory cortex, are noticeable but not as profound as after eradicating one of the whiskers altogether. The latter leads to a complete disappearance of the corresponding barrel, whereas clipping leads only to an attenuation or partial shrinkage of the barrel.

There are, however, several studies on the role of specific sensory experience that demonstrate the importance of peripheral stimulation for normal somatosensory development. Repetitive, long-term stimulation of a specific finger in monkeys leads to an expansion of that finger's representation in the somatosensory cortex. In syndactyly, a condition in which several fingers are pathologically joined at birth, receptive fields (RFs), as determined with MEG in humans, stretch across all of these fingers. When the fingers are surgically separated, the cortical MEG activation also begins to divide into separate zones. Experiments in monkeys in which several fingers were surgically joined together by suture have the opposite effect: RFs in the somatosensory cortex, which are normally well-separated into distinct finger representations, become larger and stretch across fingers. Once the fingers are separated, the RFs separate again as well. Coselection of coincident inputs according to Hebb's rule again is the principle after which the cortex becomes reorganized.

Whereas the results of these studies show that reorganization takes place largely at the cortical level (as in the visual and auditory systems), other experiments suggest that somatosensory subcortical structures, including the thalamus and spinal cord, can also be reorganized in response to deafferentation or

deactivation. Whether this occurs under cortical control is currently an active area of research. It will be interesting to determine whether a real difference exists between the somatosensory and the visual–auditory systems in that respect.

Cross-modal compensatory plasticity as described earlier for congenitally blind or deaf individuals, who show an improved representation of their other senses, is more difficult to discern with somatosensory deprivation because it is never complete. However, a rare form of severe pan-sensory neuropathy leads to an almost complete absence of proprioceptive input to the cerebral cortex. Imaging studies have shown that this somatosensory deprivation is partially compensated by increased visual activation in the parietal cortex.

V. CONCLUSIONS

Sensory deprivation studies in all three major modalities (visual, auditory, and somatosensory) demonstrate the importance of sensory experience for normal brain development. The activity-dependent reorganization of the brain (and in particular the cerebral cortex) follows rules postulated by Hebb for associative learning. In their most general form, these rules state that synaptic connections are strengthened in which the pre- and postsynaptic neurons are active together. Many modifications of these rules have been explored, in both theoretical models as well as experimental studies: Coactivtation is sometimes defined as a "firing together" of the neurons (as envisioned by Hebb) or merely as a joint rise in membrane potential. In addition to a strengthening of synapses by temporally coincident activation (long-term potentiation), a weakening of synapses by anti-correlation is now also generally accepted ("anti-Hebb," long-term depression). Both rules together also incorporate the important concept of synaptic competition, mentioned variously throughout this review, in which connections are strengthened (and maps expand) at the expense of others. Cross-modal compensation by the nondeprived senses is also a natural consequence of deprivation in one sensory modality, which is governed by the same rules for synaptic modification.

See Also the Following Articles

AGNOSIA • AUDITORY AGNOSIA • BRAIN DEVELOPMENT • CEREBRAL CORTEX • HEARING • NEUROPLASTICITY, DEVELOPMENTAL • TACTILE PERCEPTION • VISUAL DISORDERS • VISUAL SYSTEM DEVELOPMENT AND NEURAL ACTIVITY

Suggested Reading

Cruikshank, S. J., and Weinberger, N. M. (1996). Evidence for the Hebbian hypothesis in experience-dependent physiological plasticity of neocortex: A critical review. *Brain Res. Rev.* **22,** 191–228.

Doupe, A. J., and Kuhl, P. K. (1999). Birdsong and human speech: Common themes and mechanisms. *Annu. Rev. Neurosci.* **22,** 567–631.

Gilbert, C. D. (1998). Adult cortical dynamics. *Physiol. Rev.* **78,** 467–485.

Hamilton, R. H., and Pascual-Leone, A. (1998). Cortical plasticity associated with Braille reading. *Trends Cogn. Sci.* **2,** 168–174.

Hebb, D. O. (1949). *The Organization of Behavior*. Wiley, New York.

Jones, E. G., and Pons, T. P. (1998). Thalamic and brain stem contributions to large-scale plasticity of primate somatosensory cortex. *Science* **282,** 1121–1125.

Kaas, J. H. (1991). Plasticity of sensory and motor maps in adult mammals. *Annu. Rev. Neurosci.* **14,** 137–167.

Kujala, T., Alho, K., and Näätänen, R. (2000). Cross-modal reorganization of human cortical functions. *Trends Neurosci.* **23,** 115–120.

Merzenich, M. M., Recanzone, G., Jenkins, W. M., Allard, T. T., and Nudo, R. J. (1988). Cortical representational plasticity. In *Neurobiology of Neocortex* (P. Rakic and W. Singer, Eds.), pp. 41–67. Wiley, New York.

Rauschecker, J. P. (1991). Mechanisms of visual plasticity: Hebb synapses, NMDA receptors and beyond. *Physiol. Rev.* **71,** 587–615.

Rauschecker, J. P. (1995). Compensatory plasticity and sensory substitution in the cerebral cortex. *Trends Neurosci.* **18,** 36–43.

Rauschecker, J. P. (1999). Auditory cortical plasticity: A comparison with other sensory systems. *Trends Neurosci.* **22,** 74–80.

Sherman, S. M., and Spear, P. D. (1982). Organization of visual pathways in normal and visually deprived cats. *Physiol. Rev.* **62,** 738–855.

Wiesel, T. N. (1982). Postnatal development of the visual cortex and the influence of environment. *Nature* **299,** 583–591.

Sex Differences in the Human Brain

LAURA S. ALLEN and ROGER A. GORSKI
University of California, Los Angeles

GLOSSARY

critical period In terms of sexual differentiation, this is the period during perinatal life that levels of gonadal hormones impose permanent modifications on the structure and functional potential of the brain. It varies from species to species and between structures and functions.

interstitial nuclei of the anterior hypothalamus-3 (INAH-3) A cell group in the human brain that is about 2–3 times larger in males than in females or homosexual males. It is a possible homolog to the SDN-POA in the rat brain.

sexual differentiation The process by which structures and functions of the body and brain become different, largely as a result of circulating gonadal hormones.

sexually dimorphic A structure or function that is, on the average, different between males and females.

sexually dimorphic nucleus of the preoptic area (SDN-POA) A nucleus in the preoptic area of the rat that is about 5 times larger in males than in females; its size is determined by levels of gonadal hormones during perinatal life. This nucleus has served as the most studied model for the process of sexual differentiation.

Every school child knows that there is a difference between boys and girls and between men and women. What remains novel to many adults—to those in the general public as well as to many *biological* scientists—is that there are measurable sex differences in the human brain as well, in terms of both structure and function. Some of these differences have been well-established, whereas many more recent findings have not been consistently replicated or even replicated at all. Some of these studies, after publication in the popular press, have been used to form or support political and religious viewpoints. More importantly, however, they contribute significantly to our fundamental biological understanding of human development, educational issues, disease, variations in human sexuality, and the process of aging. Studies in laboratory animals have led to concepts about the biology of sex differences in the human brain. Clearly, it is not sociocultural factors in the rat cage that determine sex differences in maze performance and sexual behavior. However, in humans, the separation of biological factors, sociocultural influences, and the interaction of nature and nurture are more difficult to dissect, because we cannot experimentally manipulate the human being and because one's environment without question modifies certain sex differences in human behavior. Research findings of sex differences in the human brain demand replication. Even in studies of laboratory animals, subtle, even unrecognized differences in technical approach, subject selection, inherent biological variations, experimental design, and interpretation of the results may produce data that cannot be replicated or are contradictory to other findings. With research involving human beings this problem is increased by orders of magnitude. The format of this article necessitates statements without detailed citations. The generalizations in this review reflect the research

results that appear most predominantly and/or that are consistent with the prevailing concepts. The field of sex differences in the human brain should be further researched in part because many classic and very important studies are awaiting replication for the first time.

I. SEXUAL DIFFERENTIATION OF THE MAMMALIAN BODY

In the late 1940s, scientists began to understand the mechanism by which male and female mammals become different, or "sexually differentiate." In a landmark experiment, caesarian-like operations were performed early in the pregnancy of rabbits to expose their fetuses. The abdomens of the fetuses were then opened to remove their gonads, which had not yet differentiated into either ovaries or testes. The incisions in the fetuses and mothers were carefully stitched back together, and the pregnancies continued. After birth, all of the gonadectomized infants, whether genetic males or females, developed as normal-appearing females. Without the testes, even the internal sex organs of both the males and females were feminine. It was concluded that male development depends on male sex hormones called androgens, which are primarily secreted by the testes. Without androgens, regardless of genetic sex, the body develops in a feminine direction. Subsequently, pregnant female rabbits were injected with a drug called cyproterone acetate, which prevents the influence of the androgens on sexual development. Although the testes continue to manufacture androgens, this drug blocks testosterone's entrance into cells, thereby preventing its masculinizing influence. As expected, all of the infants born—even the genetic males—looked like females.

Scientists turned to the cattle industry to ask the opposite question: If a lack of testosterone results in the feminine development of genetic males, does the presence of testosterone result in the masculine development of genetic females? Usually cows have one calf per pregnancy, but on occasion twins or triplets are born. If the calves are both brothers or both sisters, they can grow up to be normal bulls and cows. But if the siblings are of opposite sex, the female is frequently masculinized by the androgens secreted by her brother *in utero*. These cows are known as *freemartins*. Although they are sexually normal on the outside, on the inside their ovaries, fallopian tubes, and uterus are usually stunted, and they may have internal male sex organs such as an epididymis and seminiferous tubules. Thus, these females exposed to testosterone were masculinized. Do these principles of sexual differentiation apply to humans as well?

II. SEXUAL DIFFERENTIATION OF THE HUMAN BODY

The sexual differentiation of the human body remarkably resembles that of the rabbit, cattle, and other mammals in that it is the level of androgens during a critical period of early development that determines the genital phenotype. Until the 8th week of fetal life, the female and male fetus look identical. Each fetus has all of the anatomical structures to become either male or female. On the inside of their bodies they have an identical gonad, which will become either an ovary or a testis. There are two pairs of ducts: *Müllerian ducts*, which form the female's internal reproductive organs (the fallopian tubes, uterus, cervix, and upper vagina), and the *Wolffian ducts*, which can form the male's internal reproductive organs (the epididymis, vas deferens, ejaculatory duct, and seminal vesicles). Normally, only one pair of these ducts actually develops; the other disappears. On the outside of their bodies the genitals appear identical. At about 8 weeks, however, sexual differentiation of the body begins. If the fetus is genetically male, then a gene on the Y chromosome called the sex-determining region of the Y chromosome (SRY) causes the undifferentiated gonads to become the testes. The testes begin to produce two hormones required for masculine development: Müllerian-duct-inhibiting hormone (MIH), a polypeptide that prevents the development of the Müllerian ducts, and the androgens, primarily testosterone, which promote the development of the Wolffian ducts. Testosterone is converted by the enzyme 5α-reductase to dihydrotestosterone (DHT) to masculinize the external genitalia, including the penis and scrotum. A baby boy is born. However, if there is no Y chromosome and no SRY, the gonad automatically becomes an ovary, and the fallopian tubes, uterus, vagina, and labia form. The baby is born a girl. Without testosterone, regardless of genetic sex, the fetus develops in the feminine direction. Metaphorically, the body is inherently female.

A. Variations in Sexual Differentiation of the Body

Although we cannot experimentally manipulate the sexual differentiation of human beings, we can turn to

individuals who, for medical reasons, have had altered hormone levels during development, such as those with genetic variations that result in either high or low levels of androgen action. Three commonly studied conditions include the androgen insensitivity syndrome, 5α-reductase deficiency, and congenital adrenal hyperplasia.

1. Androgen Insensitivity Syndrome (AIS)

This is a genetic variation whereby receptors for testosterone do not function. Because SRY is present, the gonads become testes. The testes secrete MIH, preventing the development of the Müllerian ducts. At puberty, the testes produce increasing amounts of testosterone, some of which is aromatized to estrogen, resulting in feminine breast development. In this case, the genetic male develops into a normal-appearing psychosexual female.

2. 5α-Reductase Deficiency

This genetic variation involves a deficiency in the enzyme 5α-reductase that converts testosterone into DHT, the hormone largely responsible for masculinizing the external genitalia. With a deficiency in DHT, the external genitalia appear feminine and the urethra may open into a vagina-like urogenital sinus. Because MIH is produced, the Müllerian ducts involute and the Wolffian ducts develop normally. At birth these children are often identified as girls; however, at puberty, rising levels of testosterone masculinize the body, including the external genitalia.

3. Congenital Adrenal Hyperplasia (CAH)

In contrast to AIS and 5α-reductase deficiency, which impair the masculinity of males, CAH results in exposure to high levels of androgens, thus masculinizing females. This rare genetic disease involves one or more defects in the enzyme system of the adrenal glands necessary for the synthesis of cortisol. Cortisol normally provides negative feedback to the hypothalamo–hypophyseal axis. However, with reduced cortisol and, thus, reduced negative feedback, there is excess secretion of adrenocorticotropic hormone, stimulating the adrenal cortex to increase steroidogenesis and resulting in the excessive formation of adrenal androgens both before and after birth.

Like other female fetuses, those with CAH exhibit normal development of the Müllerian ducts with regression of the Wolffian system. However, their clitoris may be so masculinized that it appears to be a penis, and they are sometimes mistakenly identified at birth as boys. Their external appearance is usually corrected surgically, and both boys and girls with this syndrome are given life-long medication to control their condition.

III. SEXUAL DIFFERENTIATION OF THE BRAIN OF NONHUMAN MAMMALS

Because the reproductive organs that are required for bisexual reproduction are sexually dimorphic, it is logical that the brain, *which controls the behavior and hormonal regulation of bisexual reproduction*, also exhibits sex differences. In fact, a similar process of sexual differentiation that occurs in the body and reproductive organs in animals including humans has been well-established to occur in the *brains* of many animals, including amphibians, birds, and mammals. With respect to mammals, high levels of sex hormones —whether secreted by the testes or administered by a scientist—result in masculine brain development; low levels of androgens result in sexually dimorphic brain functions and structures that remain feminine.

A. Gonadal Hormones Influence the Brains of Nonhuman Mammals

Rats are the most studied species in terms of sexual differentiation of the brain. The principles that have been observed in rats often apply to other species and have stimulated studies in other animals, including human beings. The most studied functions controlled by the brain that differ between males and females are reproductive physiology and sexual behavior.

In terms of reproductive physiology, the female rat has a 4-day ovarian cycle. At first, the ovaries secrete increasing amounts of estrogen that result in an abrupt surge of leutinizing hormone-releasing hormone from the brain, causing the release of leutinizing hormone from the pituitary, which in turn triggers ovulation. This process is termed *positive feedback*. Positive feedback is not present in the normal adult male rat, even when given injections of estrogen; testosterone early in life has permanently prevented positive feedback in response to estrogen. Similarly, males and females exhibit differences in partner preference and sexual behavior. In rats, basic sexual behavior involves mounting and penile intromission by the male and an arching of the back, termed lordosis, in the female.

1. Historical Perspective

In the late 1930s and early 1940s, a series of experiments was performed in newborn rats: in males the testes were removed and replaced with transplanted ovaries; in females the ovaries were removed and replaced with testes; and in some pups, additional gonads were added so that they had both ovaries and testes. When these rats became adults, the males with ovaries and no testes displayed the female ability to produce corpora lutea, whereas those with both ovaries and testes failed to form any corpora lutea. Among the females with intact ovaries and transplanted testes, many failed to either show estrous cycles or form corpora lutea. From these results it was deduced that, regardless of genetic sex, if testosterone was present during a critical period, then in adulthood the *pituitary* would produce secretions in the noncyclic male mode, and if testicular androgens were *absent* during early life, then in adulthood the pituitary would be cyclic and feminine in nature.

However, the hypothesis that androgen levels had sexually differentiated the pituitary was challenged by the observation that when the pituitary of a male rat was transplanted under the hypothalamus of a female, her reproductive functions remained feminine. Likewise, when the pituitary of a female was transplanted under the hypothalamus of a male, his reproductive functions remained masculine. These experiments suggested that it was not the pituitary but the *brain* that was subject to sexual differentiation by levels of testosterone during early life.

By the late 1950s it was demonstrated that sexual differentiation of the brain not only applied to physiology but to behavior as well. Pregnant guinea pigs injected with large quantities of testosterone propionate (TP) during pregnancy gave birth to females with external genitalia that were macroscopically indistinguishable from those of their brothers. In addition, these females demonstrated a reduced capacity to display lordosis and an increase in mounting behavior. Thus, it was concluded that gonadal hormones influence the brain during both development and adulthood. Such effects have been distinguished into two categories: organizational and activational. Androgens act to *organize* the brain during perinatal life to impose permanent modifications in sexually dimorphic structures and functions. The mechanisms of organization include the growth of nerve processes, prevention of neuronal death, and promotion of differentiation. In adulthood, gonadal hormones exert *activational* effects, which involve more transient, reversible, and modifiable influences that may include temporary changes in structure and neurotransmitter systems and the regulation of ovulation and sexual behavior. The ability of a male or female rat to exhibit sex-stereotyped sexual behavior is first determined during development by the organizational influence of gonadal hormones, but later activated in adulthood by circulating levels of sex steroids.

A close relationship has long been observed between reproduction and the external environment. The female reproductive cycle in a variety of mammals studied thus far is influenced not just by early hormone exposure but also by a wide range of environmental factors, such as by light, diet, temperature, and emotional stress. Moreover, electrical stimulation of the hypothalamus can induce ovulation, and lesions of the hypothalamus can block ovulation. Thus, in adulthood, the environment can influence sexually dimorphic functions in the brain.

B. Structural Sex Differences in the Brain

In the last three decades, it has been established that there are not just functional sex differences but also anatomical sex differences in the brain. These structural differences in all species studied thus far are influenced not by genetic sex directly but by sex hormones primarily during development and occasionally during adulthood. The discovery of sexually dimorphic structures in regions controlling sexually dimorphic functions has served as a springboard for the concept that such structures may underlie differences in physiology and behavior.

1. Historical Perspective

In the early 1970s the electron microscope was used to observe a structural sex difference in a region of the brain called the preoptic area (POA) located in the anterior hypothalamus that is involved in sexual and reproductive functions. This difference involves synaptic input from various regions of the brain other than the amygdala to the POA: in the normal female, the number of nonamygdaloid synapses in dendritic spines in the POA is greater than in the male. This sex difference in brain structure, like that of brain function, follows the principles of sexual differentiation: castration of the male within 12 hr after birth causes a female pattern in the number of spine synapses and permits the cyclic pattern of gonadotropin release and the ability to show a progesterone-

facilitated increase in receptivity. Likewise, females treated on day 4 of life with TP have a masculine number of spine synapses and tonic gonadotropin release. Thus, for the first time, it was shown that there is a *correlation* between sexually dimorphic functions and structures of the brain, not necessarily because these structures determine these functions but because the same factor, the level of androgens during perinatal life, determines whether the brain develops in a masculine or feminine direction for both functional and structural characteristics.

In rapid succession came reports of more neuroanatomical sex differences: in the dendritic branching patterns of neurons in the preoptic area of the rat, hamster, and macaque monkey; in the synaptic organization of the arcuate nucleus and the medial amygdala of the rat; and in the volumes of nuclei in the bed nucleus of the stria terminalis of rats and guinea pigs and in the preoptic area of gerbils, guinea pigs, ferrets, quail, and rhesus monkeys. Sexual dimorphism also exists in regions of the brain apparently not directly related to reproduction: in rats there are sexually dimorphic patterns of cortical and hippocampal asymmetries; and possibly in the number of axons, extent of myelination, and size and shape of the corpus callosum. At the molecular level, there are sexual dimorphisms in terms of neurotransmitter systems and receptors for gonadal hormones, in regions of the brain that both do and do not influence reproductive functions.

Perhaps the most studied sexual dimorphisms are in terms of the volumes of several nuclei that are present in regions of the brain related to reproductive functions. Several sexually dimorphic cell clusters control vocal behavior in canaries and zebra finches. In these species, males sing and females do not, and the male possesses a network of vocal nuclei that is as much as 6 times larger than that of females. In the rat, there is a distinct nucleus in the preoptic area, called the sexually dimorphic nucleus of the preoptic area (SDN-POA), that is about 5 times larger in males. In the spinal cord of rats, there is a sex difference in the number of motor neurons in the spinal nucleus of the bulbocavernosus (SNB), which innervates penile muscles.

2. The Medial Preoptic Area (MPOA) and Sexually Dimorphic Nucleus of the Preoptic Area (SDN-POA)

The most studied sexual dimorphism in the mammalian brain is the SDN-POA in the rat, located in the medial preoptic area (MPOA) of the anterior hypothalamus. The SDN-POA is sexually dimorphic in terms of both volume and the number of neurons. The MPOA influences sexually dimorphic functions in both rodents and primates, including gonadotropin regulation and maternal and sexual behavior. Although we still do not know the function of the SDN-POA, there is a *correlation* in terms of the volume of this structure and male sexual behavior. This correlation exists because gonadal hormone levels in the brain during a critical period determine whether brain function and structure develop in a masculine or feminine direction.

Like other sexually dimorphic areas in the body and brain, the cells of the SDN-POA contain receptors for steroid hormones. In the presence of testosterone during a critical period of early life, the SDN-POA develops into a larger structure typical of males; relatively low levels of androgens result in a smaller structure typical of females. It is believed that androgens are required during a critical period to prevent cell death of the young neurons within the SDN-POA; likewise, steroid hormones influence the survival of neurons during development in other sexually dimorphic structures in the CNS of various species, probably including the human being.

Scientists, by altering the androgen levels during perinatal life, have determined the critical period of sexual differentiation of the SDN-POA to be between day 18 of gestation and day 5 after birth. Castration of the newborn male rat produces a reduction in the volume of the SDN-POA in adulthood that can be completely prevented by the administration of androgen 1 day later. Similarly, a single injection of TP into a newborn female significantly increases the volume of the SDN-POA in adulthood. Moreover, the volume of the SDN-POA can be *completely* sex-reversed in males by prenatal antiandrogen injections and castration on day 1 and in females by pre- and postnatal testosterone injections.

The mechanisms by which androgens act to masculinize structures and functions vary. For example, the external genitalia are sexually differentiated by the conversion of testosterone to DHT; in the SDN-POA and other structures, testosterone is aromatized to estrogen within the neurons. Therefore, injection(s) of estradiol benzoate or diethylstilbesterol can masculinize the SDN-POA in the female. The volume of the SDN-POA in the adult rat does not appear to depend strongly on gonadal hormones; however, adult levels of gonadal hormones do influence the volumes of other sexually dimorphic structures, such as certain spinal cord nuclei in the rat, an apparent SDN-POA

equivalent in the gerbil, and nuclei involved in singing in canaries and finches.

The critical period of sexual differentiation of the brain is after that of the body. Therefore, by manipulating sex hormones in laboratory animals during the period of sexual differentiation of the brain, one can have an individual with a body concordant with the genetic sex but a brain, in terms of both structure and function, characteristic of the opposite sex. Similarly, the critical periods of sexual differentiation of the brain vary between structures and functions and from species to species. It is essentially unknown when during perinatal life the sexual differentiation of the human brain occurs.

IV. SEXUAL DIMORPHISM IN THE HUMAN BRAIN

A. Sex Differences in Overall Brain Size

For over a century, it has been known that the brains of men weigh about 15% more than those of women. Historically, this anatomical dimorphism had been used to promote the notion of intellectual dimorphism, with a larger brain positively correlating with greater intelligence; however, the functional significance of sex differences in brain weight is unknown. In fact, the functional significance of every anatomical sex difference in the human brain identified thus far is unknown. The generally accepted explanation for sexual dimorphism in brain size is that it is dependent upon body size. However, brain size, although correlated with body size, is determined by other factors as well. For example, sex differences in brain size actually precede sex differences in overall height. Brain size is similar between boys and girls until 2 or 3 years of age, at which time it grows at a faster rate in boys until 5 or 6 years of age, when most of the full brain weight is achieved. The sex difference in brain size then remains relatively constant despite the fact that a sex difference in height increases throughout adolescence. Increases in gross brain weight, which may occur earlier in boys, do not necessarily correlate with brain *maturation*; in fact, *functional* differences between boys and girls during childhood and adolescence suggest that the brains of girls may actually mature earlier than those of boys.

Reports suggest that, overall, there are more neurons in the brains of men than in those of women. However, in certain regions, there are reports that women may contain more brain cells and/or neural activity than men: (1) a greater density of neurons in the planum temporale was found in women; (2) in cortical regions involved in verbal behavior, magnetic resonance imaging demonstrated that women had a greater percentage of gray matter in the dorsolateral prefrontal cortex and in the superior temporal gyrus; and (3) in certain regions women exhibit greater cerebral blood flow rate and regional cerebral glucose metabolism. However, more brain cells or neural activity does not necessitate higher brain functioning. For example, the developing human contains more neurons than the adult; however, the infant is not more neurologically capable than the adult until he or she matures through dramatic cell loss and the development of considerable connectivity.

B. What Causes Sexual Dimorphism in the Human Brain?

The presence or absence of gonadal hormones during a critical period of early life determines the sexual differentiation of the genitalia in nonhuman and human animals and of the brains of nonhuman animals. However, the role of steroid hormones in the sexual differentiation of the human brain is less understood for two reasons: (1) until relatively recently, there were few reports of sex differences in the human brain, and (2) we cannot experimentally manipulate the brains of humans during perinatal life to determine hormonal influence on structure and function.

Therefore, scientists searched the human brain for anatomical sex differences so that later the identified sexually dimorphic structures could be examined in individuals who may have been exposed to altered hormone levels during early life, including those with AIS, 5α-reductase deficiency, and CAH, and in individuals who exhibit altered sexually dimorphic behavior such as homosexuals and transsexuals, who may also have been exposed to atypical gonadal hormone levels during development.

C. Sex Differences in Regions Related to Reproduction

Because much has been learned about the SDN-POA in the rat, researchers attempted to identify a human homolog. From autopsy samples, hypothalamic areas were age-matched and prepared much like those in the rat studies. The search for the "SDN-POA" in the preoptic–anterior hypothalamic area in the human

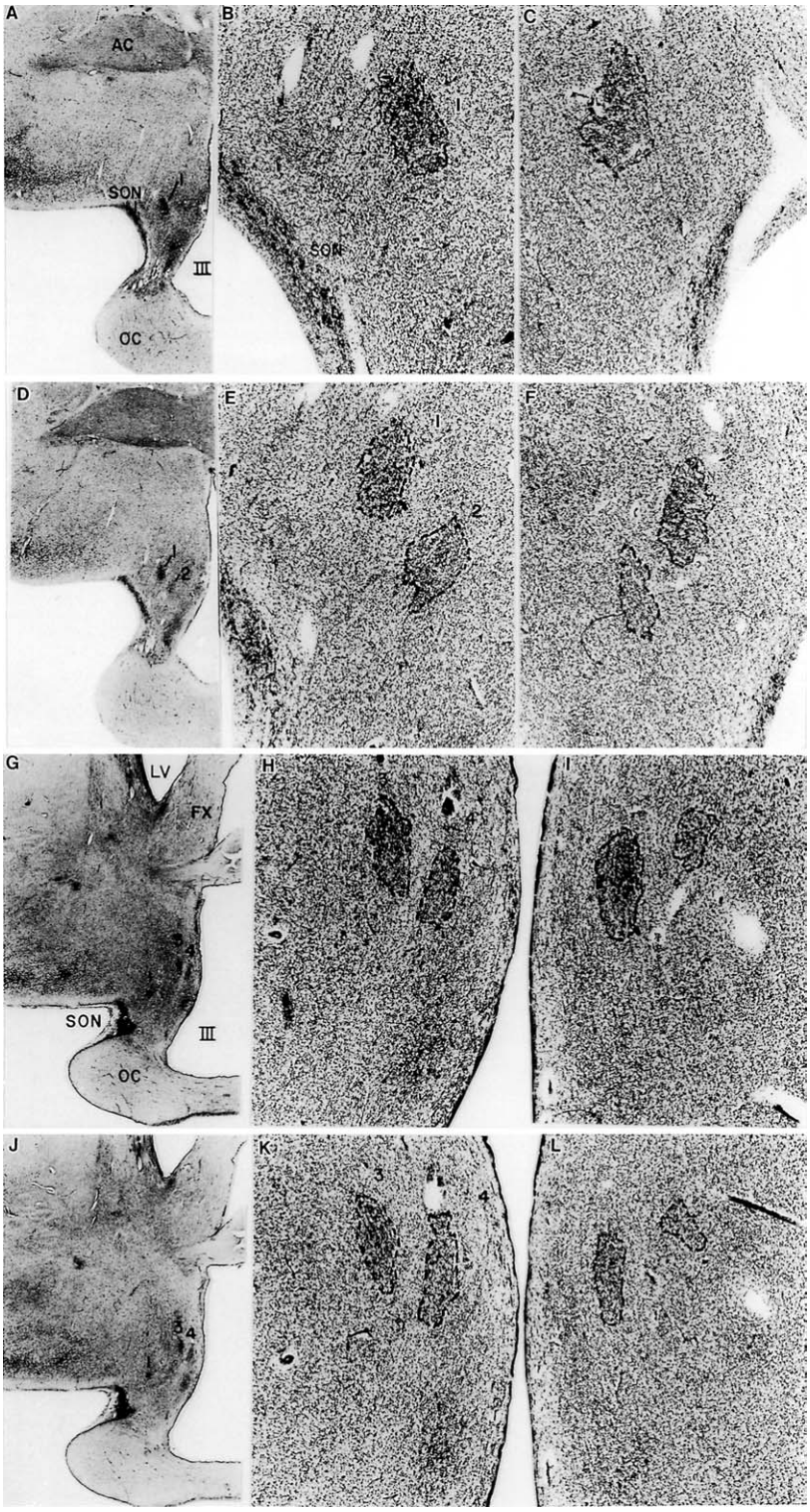


Figure 1 (A–F), Photomicrographs of thionin-stained coronal sections of the anterior hypothalamus in the human brain. A, D, G, and J are photomicrographs from a 58-year-old male, from anterior to posterior. B, E, H, and K are higher power photomicrographs of A, D, G, and J, respectively. Likewise, C, F, I, and L are female counterparts to B, E, H, and K, respectively, from a 63-year-old woman. Abbreviations: AC, anterior commissure; 1, 2, 3, and 4, INAH 1–4; III, third ventricle; FX, fornix; LV, lateral ventricle; SON, supraoptic nucleus; OC, optic chiasm. Reprinted and modified by permission from Allen, L. S., Hines, M., Shryne, J. E., and Gorski, R. A. (1989). Two sexually dimorphic cell groups in the human brain. *J. Neurosci.* **9**(2), 497.

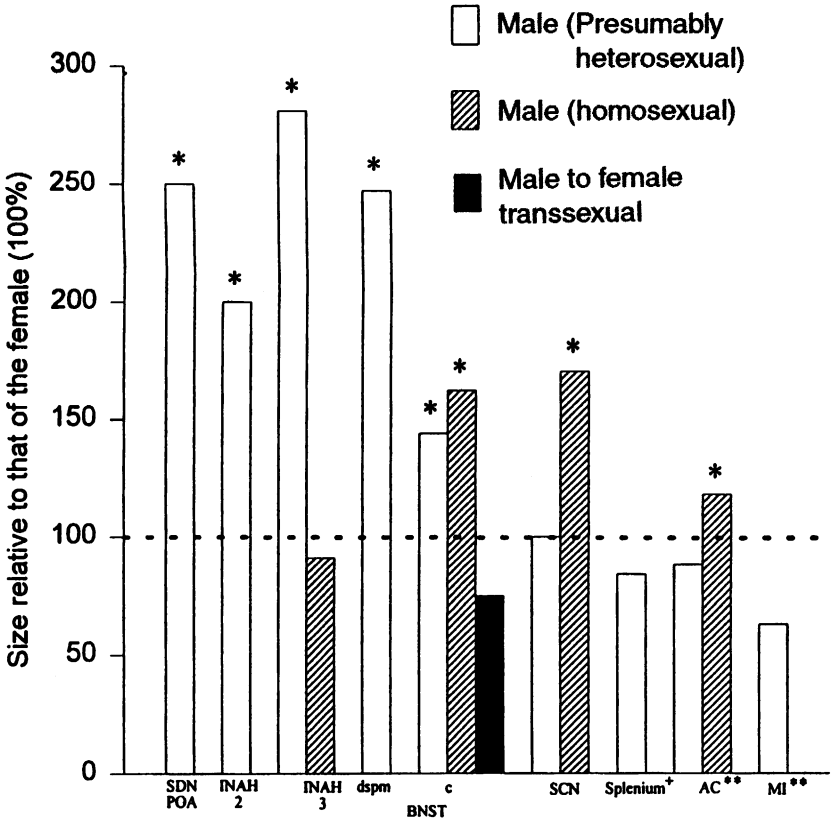


Figure 2 Size relative to that of the female in terms of volume (SDN-POA, INAH-2, INAH-3, BNST-dspm, BNSTc, SCN) or midsagittal area (splenium, AC, and MI) of structures that have been reported to differ between heterosexual males and females, homosexual males, and male-to-female transsexuals. Abbreviations: AC, anterior commissure; BNST, bed nucleus of the stria terminalis (dspm, darkly staining posteromedial component; c, central subdivision); INAH-2 and INAH-3, interstitial nuclei of the anterior hypothalamus 2 and 3; SCN, suprachiasmatic nucleus; SDN-POA, sexually dimorphic nucleus of the preoptic area (in the human also termed INAH-1 and the intermediate nucleus). *Significant difference from that of the female. **Midsagittal area. +Bulbosity. Reprinted from Gorski, R. A. (1998). Sexual differentiation of the brain. In *Principles of Medical Biology 12: Reproductive Endocrinology and Biology* (E. E. Bittar and N. Bittar, Eds.), p. 1. JAI Press, a subsidiary of Elsevier Science.

resulted in the quantification of four nuclei that were termed the interstitial nuclei of the anterior hypothalamus 1–4 or INAH 1–4. INAH-2 and INAH-3 were 2–3 times larger in males (Figs. 1 and 2). Interestingly, in a very small sample it appeared that the sex difference in INAH-2 might be related to reproductive age.

Subsequently, the human bed nucleus of the stria terminalis was examined for sexual dimorphism because this area was found to be sexually dimorphic in several rodent species. Like the SDN-POA of the rat, this area has been implicated in sexually dimorphic functions, including gonadotropin regulation and sexual and maternal behavior, it concentrates gonadal hormones, and it is anatomically connected to several other sexually dimorphic regions including the MPOA. A region termed the darkly staining posteromedial region of the bed nucleus of the stria terminalis (BNST-dspm) is about 2.5 times larger in males than in females (Fig. 2).

1. Is INAH-1 the SDN-POA?

In one laboratory, a sex difference has been reported in several studies in the volume and cell number of the intermediate nucleus of the hypothalamus, subsequently called the SDN-POA and INAH-1. Because this nucleus was found to be 2–3 times larger in males than in females and is located in an area similar to the SDN-POA of rats, this laboratory named it the SDN-POA of the human. Several other laboratories have found neither INAH-1 nor INAH-2 to be sexually dimorphic. The discrepancy in findings might be partly due to the fact that INAH-1 undergoes dramatic

changes in volume and cell number throughout life, beginning at birth, and the rate at which this process occurs during adulthood is greater in males. Discordant findings in terms of the sexual dimorphism of INAH-1 remind us how vulnerable human brain research is to variations among individuals, ranging from nuances in the environment to something so fundamental as age. Indeed, examination of particular age groups or the failure to age-match males and females could contribute to a variation in results.

The suprachiasmatic nucleus (SCN) is reportedly more *elongated* in women and more *spherical* in men. Unlike the INAH, the SCN has a known function—to participate in the regulation of circadian rhythmicity. Indeed, certain circadian cycles are regulated differently in males and females. The SCN, like the INAH 1–4, has been examined for differences between heterosexuals and homosexual men, and an intriguing result was obtained (see later discussion).

2. Sexually Dimorphic Neural Circuits

In the rat, there appears to be a sexually dimorphic and interconnected network that begins with incoming chemical information from the environment to the vomeronasal organ that projects to the medial amygdala, which in turn projects to the BNST and the MPOA. Interestingly, each structure in this neural circuit appears to concentrate gonadal hormones and to be sexually dimorphic. Indeed, such a neural circuit is possibly present in human beings and may include the BNST-dspm, INAH-3, and other as yet unreported regions in the vomeronasal organ and medial amygdala.

3. Other Nuclei Located in Regions Related to Reproduction

Perhaps the only sexual dimorphism in the human nervous system with a known function is Onuf's nucleus, called the SNB in rats, which is located in the sacral spinal cord of dogs, cats, and primates and which innervates striated penile muscles, including the bulbocavernosus and ischiocavernosus muscles. In the homologous nucleus in rats, androgens have been shown to save the neurons of this nucleus from cell death during perinatal life. In humans, there was a significant decline in motoneuron number during a period of gestation that coincides with the initiation of androgen production by the human fetal testes, suggesting, perhaps, that levels of androgens in humans as well may determine this sex difference by altering cell survival during early life.

D. Sexual Dimorphism in Regions of the Brain Not Related to Reproduction

In contrast to sexually dimorphic structures in regions related to reproductive function that exhibit little overlap between males and females, those in areas with no currently known sexual function appear to exhibit considerable overlap between the sexes and, thus, require larger sample sizes to demonstrate differences. There are sex differences in non-reproductive regions, including the cerebral cortex and structures that connect the two hemispheres.

1. Structures Connecting the Two Hemispheres

Communication between the two hemispheres of the brain is limited to several structures, three of which have been reported to be sexually dimorphic, including the massa intermedia (MI), corpus callosum (CC), and anterior commissure (AC). In contrast to the sexually dimorphic nuclei that are larger in males, each of these structures has been reported to be larger, shaped differently, and/or more often present at the midsagittal plane of the brain in females.

2. The Corpus Callosum (CC)

The major communication between the two hemispheres of the brain is orchestrated by the CC, which is composed of axons that cross the midline. In the early 1980s researchers reported that the CC at the midsagittal plane exhibits sexual dimorphism. In a small sample of 14 brains, they found that the *shape* of the posterior part of the CC, the splenium, was sexually dimorphic: it was more bulbous-shaped in women and more tubular-shaped in men, and the maximum splenial width was greater in women. However, there was no significant sexual dimorphism in terms of *size*: the area of the splenium, defined as the posterior one-fifth of the CC, was *nonsignificantly* larger in females (Figs. 3A and 3B).

Despite many subsequent studies, the findings of sexual dimorphism of the CC have remained controversial for several reasons: (1) Measurements have been performed in different ways in various laboratories, in part because published reports frequently do not describe methodology in detail. Moreover, approaches to the measurement of the CC rely primarily on measures based on arbitrary subdivisions, which may mask the particular regions that exhibit the greatest dimorphism by including areas that exhibit no difference. For example, some researchers have

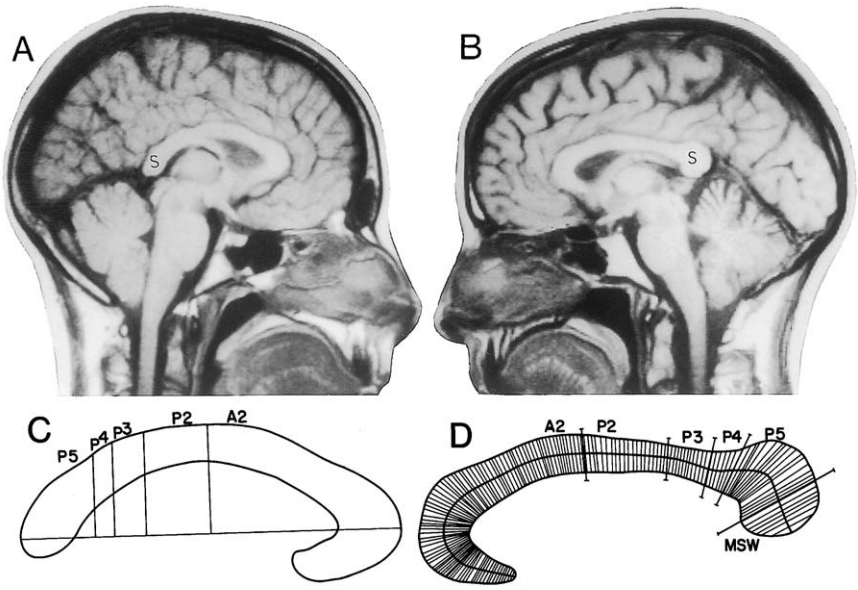


Figure 3 Magnetic resonance images of midsagittal sections of the CC through a male (A) and a female (B) brain demonstrating, in this pair, a difference in the shape of the splenium (S), and the "straight-line" (C) and "curved-line" (D) methods of dividing the CC into sections. Other abbreviations: P5, posterior fifth; P4, posterior fourth; P3, posterior third; P2, posterior half; A2, anterior half; MSW, maximum splenial width. Reprinted and modified by permission from Allen, L. S., Richey, M. F., Chai, Y. M., and Gorski, R. A. (1991). Sex differences in the corpus callosum of the living human being. *J. Neurosci.* **11**(4), 933.

partitioned the CC on the basis of its most anterior and posterior points utilizing a "straight-line method" (Fig. 3C), whereas others partitioned it on the basis of its curvature, utilizing a "curved-line method" (Fig. 3D). Although neither method is correct or incorrect, both methods arbitrarily partition the CC into slightly different components. (2) Despite age-related changes during childhood and adulthood, few investigators have explicitly age-matched subjects. (3) The shape and size of CC vary considerably among individuals, requiring large sample sizes to demonstrate significant sex differences. (4) Many people overinterpreted the original report of a *nonsignificantly* larger splenium in females to mean that the splenium *is* significantly larger in women. At its extreme, the original report was interpreted to suggest that the overall area of the CC is larger in women, and thus there is greater connectivity between the two hemispheres in females. (5) It can be difficult to obtain a precisely midsagittal section of the CC, resulting in less accurate results. Although some studies do find the splenium to be larger and/or shaped differently in women, most studies agree that one cannot correctly classify every brain as either male or female on the basis of the size or shape of the CC. In the literature, controversy remains as to whether the absolute overall midsagittal area of the CC is significantly larger in males, essentially similar in males and females, or significantly larger in females particularly when adjustments are made for sex differences in brain size.

The functional significance of differences in the CC remains unknown. It has been suggested that a larger CC subregion might reflect more axons connecting the two sides of the brain, corresponding to more cerebral symmetry. However, a sex difference in the midsagittal area of the CC does not mean that there is a sex difference in the number of axons. Moreover, more axons do not necessarily correspond to greater cerebral symmetry. Variations in (areas of) the CC could correspond to differences in myelination, differences in the number of fibers, and/or a varying arrangement of axons coursing through the CC. For example, in the splenium of rats, there are more axons in females but more myelinated axons in males, resulting in an area similar in males and females. In rhesus monkeys there is an approximately 2-fold variation among animals between the density of axons and the midsagittal area of the CC; therefore, there is no correlation between the area of the CC and the number of axons. In contrast, several human studies did not find a sex difference in the number of fibers per area when examining relatively larger fibers. However, smaller,

unmyelinated axons, whereas functionally quite significant, are difficult to measure in post mortem human tissue. Thus, the number of larger, more easily quantifiable axons does not necessarily indicate the quantity of connections between the hemispheres.

3. The Massa Intermedia (MI) and the Anterior Commissure (AC)

For decades, it has been observed that the MI is more often present in women than in men. In addition, the MI is larger in women whether all males are considered or only those that possess an MI (Fig. 2). It is possible that the sex difference in the MI is due to a progressive involution of the MI that occurs with age more rapidly in males than in females. If true, then the younger the subjects that are examined, the less of a sex difference that may be found. Although the functional significance of these differences is unknown, pneumoencephalograms demonstrated that men without an MI exhibited higher performance on the nonverbal portion of the *Wechsler–Bellevue Intelligence Scale* than men with this structure, although no association was seen in women.

In addition, the midsagittal area of the AC is reportedly larger in women, although not all studies agree (Fig. 2). It is unclear whether sex differences in the MI and AC are due to the number of axons, the thickness of myelination, or, in the case of the MI, the number of neurons, glia, and connective tissue in the thalamic nuclei that are present at the midsagittal plane. The AC of the primate brain is a tract of axons that primarily connects the left and right neocortices. In a small sample of human subjects, variations in the area of the AC reflect differences in the number of axons but not their density. In contrast, in a study of rhesus monkeys, the midsagittal surface area of the AC did not correlate strongly with the number of axons.

E. Hemispheric Specialization

The two hemispheres of the human brain exhibit differences in anatomy and function by birth that continue throughout life. An anatomical or functional difference between the two hemispheres is termed *hemispheric specialization* or *cerebral lateralization*. The greater involvement that a given hemisphere has for a given structure or function relative to the opposite hemisphere, the more lateralized or specialized it is said to be.

In terms of hemispheric specialization, anatomical differences appear to underlie functional differences. In most people, the overall size of the right hemisphere is larger than that of the left, although particular regions, such as those involved in language, may be larger on the left side. In terms of function, electrophysiological studies show that most infants exhibit left-ear superiority for musical stimuli and right-ear superiority for verbal stimuli. As one matures, the left hemisphere of a majority of right-handed people is involved in language, the execution of learned manual activities such as handwriting, and analytical, symbolic, logical, and abstract thinking. In contrast, the right hemisphere has a greater role in emotion, perceptual, literal, concrete, wholistic, and synthetic thinking, and the processes of visual imagery—the kind in which a single picture or mental image is worth a thousand words.

There are also *sex differences* in both the anatomy and function of hemispheric specialization. The left planum temporale is generally larger than its right hemisphere counterpart. However, as a group, there is greater symmetry of the planum temporale in women: more females have significantly less leftward asymmetry, and more females have a reversed pattern of asymmetry with a larger right planum temporale. The planum temporale is one of the regions of the brain involved in verbal abilities: in women, the tendency for greater symmetry of the planum temporale may underlie greater symmetry of language function.

Likewise, the primary transverse gyrus (Heschl's gyrus), which contains the primary auditory cortex, exhibits sexually dimorphic patterns. Normally, Heschl's gyrus arches onto the lateral surface of the superior temporal gyrus more frequently on the left than on the right hemisphere; however, it does so more often in men than in women. In addition, there are sex differences in the pattern of the fissurization of the Sylvian fissure.

The results of studies suggest that males have greater cerebral lateralization for both verbal and nonverbal functions than do females. Data supporting this contention come from three sources: (1) studies that involved IQ tests administered to people following stroke or tumor confined to one hemisphere have demonstrated sexually dimorphic patterns in cognitive deficits; (2) psychological tests that assess the functioning of both hemispheres suggest dimorphisms in hemispheric processing; and (3) more recently, *in vivo* imaging techniques such as positron emission tomography (PET) and functional MRI (fMRI) demonstrate hemispheric dimorphism during certain task performances.

Following unilateral brain damage by stroke, trauma, or surgical removal, there are sex differences in the pattern of impairment on the performance intelligence quotient (PIQ) and verbal intelligence quotient (VIQ) components of the *Wechsler Adult Intelligence Scale* (*WAIS*). In most adults, damage to the left hemisphere impairs VIQ more, whereas damage to the right side decreases PIQ to a greater extent. The IQ deficit in either the VIQ or PIQ, depending upon which hemisphere is damaged, is much greater in males than in females. Males exhibit a greater reduction in either the VIQ or PIQ because each hemisphere is more specialized for that particular function. In females, because both hemispheres have a similar functional role, damage to either hemisphere impairs VIQ and PIQ to a similar extent; moreover, the opposite, undamaged hemisphere, which possesses both verbal and visuospatial capacities, can compensate for the loss to a greater extent than in males.

Some psychological studies such as tachistoscopic, dichotic, and monaural listening experiments, and EEG analysis and magnetoencephalographic (MEG) auditory evoked fields combined with neuropsychological test performance, suggest greater functional hemispheric specialization in males. Similarly, some fMRI and PET studies during language tasks also show sex differences depending on the particular language task tested.

In addition to the tendency for women to exhibit more functional symmetry between hemispheres, there are also sex differences within a single hemisphere, determined by patterns of hemispheric cerebral metabolism using PET and fMRI. For example, particular language and praxic functions may be represented more diffusely within the left hemisphere in males. Likewise, males exhibit auditory evoked fields more anteriorly within the superior temporal lobes, particularly in the right hemisphere. By using fMRI, brain activation in males is found to be lateralized to the left inferior frontal gyrus, whereas in females the pattern of activation engages a more diffuse neural system that involves both the left and right inferior frontal gyri. However, there are discrepancies between different studies, which may be due to the particular test given.

F. Handedness

A complex relationship exists between sex, handedness, and brain organization. Most right-handed people have left-sided language lateralization and right hemisphere specialization for visuospatial processing, whereas as many as half of left-handed people have right-sided language dominance and left-hemisphere visuospatial processing. Less than 1% of right-handed people have the reversed pattern, whereas about half of left-handed people exhibit such a reversal.

These patterns exhibit sex differences. In terms of function, right-hemispheric language localization is more frequent in left-handed females than in left-handed males. Additionally, there are anatomical relationships between sex and handedness: (1) The structure of the Sylvian fissure correlates with handedness in men but not in women. In contrast to the generally held view that right-handed males have greater cerebral asymmetry, consistently right-handed males have longer horizonatal segments in both hemispheres than males that use the left hand for certain tasks. (2) Right-handed men and left-handed women, in comparison to left-handed women and right-handed men, have stronger rightward asymmetry of the planum parietale. (3) The planum temporale exhibits less leftward asymmetry in females and left-handed individuals than in right-handed males.

G. The Corpus Callosum, Cerebral Lateralization, and Handedness

Relationships exist between the areas of particular regions of the CC, handedness, and cerebral lateralization. In the 1980s Sandra Witelson identified the most consistent and convincing sex difference in the CC to date: certain regions vary in midsagittal area with *handedness* in a sexually dimorphic manner. In these studies, subjects were given a questionnaire to determine not just which hand they used to write with but which hand they used to perform a variety of other tasks, including drawing, brushing their teeth, throwing a ball and using scissors. Subjects were classified accordingly as consistent right-handers (CRH) (those who use their right hand to perform each item on the questionnaire), left-handers (LH), and nonconsistent right-handers or mixed-handers (NCRH) (those who use their left hand to do at least one of the tasks). In males, the CC is larger in NCRH and LH, who also have greater cerebral symmetry than CRH in a posterior region of the CC called the *isthmus* that connects the posterior parietal and temporal cortical regions. Similarly, regions of the anterior CC were larger in LH males than in CRH males. In contrast, in females, regions of the anterior CC were larger in CRH than in LH. This is reminiscent of rats, where

left-biased animals exhibit the greatest sexual dimorphism in the size of the CC.

If there are sex and handedness dimorphisms in the CC, then there may be corresponding differences in regions of the cerebral cortex with which these dimorphic areas of the CC interconnect. Some researchers have suggested that a larger region of the CC correlates with greater symmetry. Thus, there would be an inverse relationship between interhemispheric connectivity (as defined by the area of a region of the CC) and cerebral asymmetries. Such negative correlations have been reported in the following regions of the CC and the cerebral cortex: the size and total number of fibers of the isthmus and the Sylvian fissure (planum temporale) asymmetry in males but not in females; the population of relatively large fibers in the isthmus and pre-Sylvian asymmetry in males but not females; the number of fibers in the splenium and the value of Sylvian fissure asymmetry in females; and the size of the midbody of the CC and Sylvian fissure asymmetry in males but not females. The hypothesis that the greater connectivity between the two hemispheres correlates with a more symmetric brain is explained somewhat by the assumption that the axons of the CC connect structurally homologous regions of the cerebral cortex such that asymmetric brains with fewer homologous regions would, therefore, have fewer connections coursing through the CC. However, this hypothesis has not been demonstrated anatomically.

Irrespective of sex, there are correlations between the size of areas of the CC and cerebral lateralization: the CC was found to be larger in subjects with right-hemisphere speech dominance; the posterior CC, including part of the splenium, correlated with language symmetry; and the CC correlated with measures of asymmetry in dichotic listening performance in some, but not all, studies.

H. Sex Differences in Cognition

Sex differences in cognitive measures have been documented for several decades. In general, males perform better on tasks of visuospatial skills, such as the alignment of a rod to the vertical, left–right discrimination, mental rotation, point localization, and disembedding figures, and mathematical reasoning. Females, on average, score higher on tests of language fluency, grammar, speed of articulation, phonetics, and fine motor skills. The mean differences in performance between the sexes typically are not large, and individual male and female scores show extensive overlap; however, the sex differences are reliable for certain tests and are statistically significant when large enough samples are considered.

Perhaps the most striking cognitive sex difference is seen in male and female adolescents with outstanding *Scholastic Aptitude Test* (*SAT*) scores involving mathematical reasoning ability. The more outstanding the student, the greater the male prevalence. Among the top students examined, the ratio of boys to girls climbed to 13:1. Researchers have been unable to identify any *environmental* factor that accounts for this difference.

An understanding of sex differences in cognition is probably of little practical importance in terms of the individual person; moreover, it must be remembered that stereotypes applied to any given individual can be damaging. However, the knowledge regarding sex differences may have major theoretical significance in terms of understanding brain development and the physiology of cognition. Whereas the dissection of nature and nurture for these differences is difficult, there is considerable evidence that cognitive sexual dimorphism is not dependent entirely upon sociocultural factors and is at least partly innate.

(1) Certain sex differences are exhibited early in life. For example, in newborn infants, girls spend more time than boys making eye contact with a silent adult caregiver, and only in females did the eye contact increase when the person was speaking. (2) Sex differences are observed in complex learned behaviors in laboratory animals, such as maze learning in rats, that cannot be explained by cultural influences. (3) In a "natural experiment" provided by "Women in the Kibbutz" in Israel, the women chose to go against the ideology of their communities in search of traditional female roles. (4) Performance in motor and spatial–perceptual skills fluctuates across the menstrual cycle, implying an influence of sex hormones during adulthood. (5) Some sex differences interact with factors associated with neurological variations. For example, although men perform better on certain spatial–perceptual tests, spatial ability is associated with hand preference, which in turn correlates with cerebral asymmetry, and this relation differs between the sexes. Likewise, right-handed men performed better on spatial tasks than left-handed men, with the reversed pattern for women with *above* average reasoning ability. (6) Various neuropsychiatric disorders such as schizophrenia, autism, and developmental dyslexia exhibit marked sex differences in incidence, symptoms, and possibly etiology.

I. Connections: Hormones, Anatomy, Hemispheric Specialization, Handedness, and Cognition

There are correlations between sex differences in gonadal hormones, anatomy, hemispheric specialization, handedness, and cognition. Hypothetically, gonadal hormones during perinatal life result in sex differences in anatomy, both in terms of the connectivity between the hemispheres and in the structure of regions within the cerebral cortex. In turn, these anatomical differences underlie functional differences in cerebral lateralization that may lead to differences in handedness and cognition.

In nonhuman animals, perinatal hormone levels appear to influence cerebral lateralization. Hormonal manipulation influences structural aspects of lateralization in rats and asymmetrically represented motor skills in songbirds. The mechanism by which early hormone levels affect cerebral lateralization is unknown. However, studies during the development of both rats and monkeys have identified asymmetries in testosterone receptors in regions of the cerebral cortex. It is unknown how or why such asymmetries for testosterone receptors arise. If asymmetries in testosterone receptors are also present in the developing human cerebral cortex, this may explain why males in general have greater cerebral lateralization.

The number of neurons in any region of the brain depends upon the number initially produced during proliferation and that survive early cell death. Testosterone, known to influence both cell survival and cell death in laboratory animals, might similarly affect the number of neurons in the developing human cerebral cortex. Thus, a greater number of testosterone receptors on one side of the brain may enable testosterone to influence cell survival or death on that particular side, resulting in asymmetry of neuronal number and, thus, both anatomical and functional asymmetry, ultimately underlying the greater cerebral lateralization in males. These findings in nonhuman animals are consistent with the hypotheses in human males that testosterone may delay the growth of the left hemisphere, resulting in dimorphic patterns in language functions, or that it may increase the development of the right hemisphere, resulting in enhanced visuospatial abilities.

There are suggestions that some sex differences in cerebral lateralization underlie differences in cognition. For example, in children with left-hemispheric language dominance, males show an asymmetry on spatial tasks but not on verbal tasks, whereas females show the reverse pattern. In children with right-hemispheric language, males display an asymmetry on the verbal task but not on the spatial task, and girls show the reversed pattern. Girls with left-hemisphere language exhibit better scores for verbal function than for spatial ability, and boys exhibit the opposite profile. The well-known sex-related differences in cognitive function, with males being relatively superior in visuospatial skills and females in verbal fluency, may pertain only to people having language localization in the left hemisphere. If the left hemisphere of females and the right hemisphere of males are functionally enhanced regardless of their specialties, then the reverse cognitive pattern may prevail in individuals with right-hemisphere language. Furthermore, sex-related differences in cognition might be increased in populations having weak hemispheric specialization.

V. SEX DIFFERENCES IN AGING

Despite many studies evaluating sex differences in the brain during development, relatively few have addressed such changes during aging. In humans, the process of neural aging has been evaluated using post mortem tissue, MRI, and PET. Advancing age has been associated with decreased brain tissue and increased cerebrospinal fluid. In general, this process occurs faster, earlier, and/or to a greater extent in men than in women. Like other studies of sex differences in the human brain, results have not been consistent partly due to variations in the designs of studies in terms of the number and age range of subjects and methods and regions of measurements.

Several sexually dimorphic structures exhibit sex differences in size changes with advancing age. The CC and INAH-1 decrease in size with age to a greater extent in men. The sex differences in neural atrophy in INAH-1 in part may account for why there is controversy as to whether it is sexually dimorphic. In the SCN, the number of neurons containing vasoactive intestinal polypeptide (VIP) exhibits a complete inversion during life: between 10 and 30 years the SCN of males contained twice as many VIP neurons as that of females, with a subsequent decrease at a more rapid rate in males such that between 40 and 60 years of age there are more VIP neurons in women. With respect to the MI, it is possible that sexual dimorphism in atrophy may account for why this structure is both present more often and larger in women than in men.

Other neural areas that have been reported to exhibit sexually dimorphic brain atrophy, which in general is greater in males, include an increase in cerebrospinal fluid volume particularly in sulcal and ventricular areas, a decrease in overall brain volume, and volumetric reductions in the frontal and temporal lobes, temporoparietal area, parieto-occipital region, and regions of the anterior CC. Like sexual dimorphism in functional hemispheric asymmetry, elderly males exhibit more atrophy in the *left* hemisphere, whereas women exhibit more symmetric hemispheric atrophy. At the microscopic level, there is significantly greater neurofibrillary degeneration of the medial basal hypothalamus in men. Likewise, cerebral blood flow in left frontal regions decreases with age to a greater extent in men.

These data provide evidence that sex differences in human brain structure are not fixed but continue to change throughout adulthood. Neuroanatomical sex differences in aging suggest that there may be cognitive sex differences in aging as well. There are sex differences in declines in memory functions, speech capacity, and visuospatial abilities. Women may have less age-related decline in verbal memory than men, whereas men may have better preservation of visuospatial abilities.

A. Hormone Changes during Aging

It is unknown why more age-related atrophy occurs in the brains of men. It has been suggested that women show less effects of neural aging due to the protective influence of female sex hormones prior to menopause. Considerable data support the role of sex hormones and corticosteroids in both development and aging. Gonadal hormones have been shown to stimulate neural growth and synapse formation, and their removal has resulted in retarded development and atrophy. The topographical distribution, binding capacity, and associated enzyme levels of sex hormones and their receptors have been shown to vary as a function of age in a variety of nonhuman species. Thus, neural atrophy with aging may also be due to varying hormone levels and changes in their receptors.

However, the fact that women exhibit an abrupt change in circulating sex hormones during menopause, whereas men exhibit a more gradual decrease—they actually have proportionately more of their gonadal hormone levels remaining for a longer time in older age—raises doubts about this notion. Perhaps the process of neural atrophy is slow enough such that it is not apparent for many years after the menopausal reduction in estrogen. If estrogen and/or progesterone are protecting the female brain from atrophy relative to the male brain, then several possibilities exist: (1) women on long-term postmenopausal estrogen therapy might have even less brain atrophy than women who are not on such therapy; (2) estrogen and/or progesterone are protective against aging, regardless of sex; (3) estrogen–progesterone protect against aging only in the female brain; (4) androgens do not protect against aging in men or do so to a lesser extent than estrogen and/or progesterone; and (5) androgens may actually promote certain facets of aging.

B. Hormone Replacement Therapy (HRT)

Age-related symptoms have been treated with HRT since the nineteenth century. By 1900 the life expectancy was only 40 years, so that the average woman throughout most of history has not lived beyond the reproductive years. Changes in female sex hormones during the menstrual cycle, the post partum period, and menopause may result in difficulties in attention, concentration, and memory, as well as disturbances such as irritability and depression. Many, but not all, studies suggest that administration of estrogen can improve cognitive abilities and emotional function.

The results of some studies suggest that changes in estrogen levels may affect psychological characteristics by action upon various neurotransmitter systems. Hormones such as estrogen influence neurochemistry by genomic and nongenomic mechanisms, modulating receptors, enzymes, synthesis, re-uptake, and degradation. Neurotransmitter systems shown to be responsive to estrogen in laboratory animals include the opiate, GABAergic, adrenergic, cholinergic, dopaminergic, and serotonergic systems, which also influence mental status in humans. Although estrogen-sensitive neurons represent a small fraction of the total neuronal population within the brain, their action can be relatively profound due in part to their interaction directly and indirectly with other neurons.

C. Alzheimer's Disease

Although anatomical studies in general suggest greater neural atrophy in men, Alzheimer's disease (AD) appears to be more frequent in women, even after accounting for sex differences in longevity. Estrogen in women may protect against AD. Likewise, the fall in

estrogen that occurs during menopause may increase the risk for this disease. Several studies indicate that HRT improves social, emotional, and cognitive functions in women with AD. Likewise, HRT has been associated with a decreased risk and delayed onset of AD. By utilizing functional MRI, postmenopausal women on HRT exhibited brain activation during memory tasks in different regions than when they were not on HRT.

The use of estrogen for AD in elderly women may aid in its prevention, because estrogen deficiency may contribute to the development of neuronal dysfunction and cell death that occurs in AD. In the rat, estrogen clearly influences hippocampal neurons: ovariectomy in the female decreases the density of dendritic spines in hippocampal pyramidal neurons, an effect prevented by estradiol treatment.

In AD, there are pathological changes in the hippocampus and basal forebrain, particularly in terms of overall neuronal loss, cholinergic cell loss in the basal forebrain, and neurofibrillary tangle formation. Typically, the basal forebrain is the major source of cholinergic innervation to the cerebral cortex, limbic system, hippocampus, and hypothalamus. In laboratory animals, estradiol increases cholinergic markers in the hippocampus and basal forebrain, where estrogen receptors colocalize on the magnocellular cholinergic neurons that express receptors for brain-derived neurotrophic factor (BDNF). Estrogen regulates the expression of BDNF that promotes the survival of basal forebrain cholinergic neurons and cerebral cortical neurons and plays a protective role after brain injury. In AD, mRNA for BDNF is markedly reduced in the cerebral cortex and the hippocampus. Estrogen's ability to enhance the transcription of molecules such as BDNF may partly underlie the molecular mechanisms by which HRT benefits individuals with AD.

VI. SEXUAL DIMORPHISM OF NEUROLOGICAL DISORDERS

A number of other neurological disorders exhibit sex differences, and an understanding of these dimorphisms gives scientists and physicians clues to their origins and treatment. During childhood, boys are slightly more prone than girls to exhibit disorders such as mental retardation, cerebral palsy, and febrile convulsions. Males, beginning in childhood and continuing into adulthood, are several times more likely than females to be diagnosed with attention deficit hyperactivity disorder. They are 3–5 times more likely to exhibit developmental language disorders such as infantile autism, delayed speech acquisition, developmental dyslexia, and stuttering. At higher levels of intelligence, the proportion of male dyslexics increases: among children with IQs under 90, boys slightly outnumber girls at a ratio of 1.2:1; at IQs between 90 and 99 the ratio climbs to 5:1; and at IQs 100 and above the ratio reaches 10:1. It has been suggested that the greater hemispheric specialization in boys could add to their risk of dyslexia because of more exclusive reliance on the left hemisphere, thus diminishing the effectiveness of other regions critical to verbal abilities. Likewise, the higher verbal abilities in girls may be related to their ability to utilize both cerebral hemispheres.

Women tend to have a greater incidence of depression, eating disorders, and obsessive–convulsive disorder (OCD). Depression affects about 3 females for every 1 male, and eating disorders such as bulimia and anorexia nervosa affect about 10 women for every 1 man. Tourette's syndrome, which sometimes accompanies OCD, is more often present in males and is probably influenced by both autosomal-dominant transmission and androgens. In fact, it is often aggravated by anabolic steroids.

Depression, eating disorders, and OCD are frequently treated by medications that influence the serotonergic system. In humans there are sex differences in the serotonergic system in terms of receptors and synthesis rate, and in laboratory animals estrogen levels influence this system. Likewise, synthetic gonadal hormones have been used to treat women with depression and other mood disorders, some of whom are unresponsive to state-of-the-art pharmaceuticals. Depression and other mood disorders have been treated with oral contraceptives in cycling women, and HRT has been effective in menopausal and postmenopausal women, perhaps by stimulating neurotransmitter systems such as the serotonergic system, which can influence one's mental and emotional status. Women tend to have a greater incidence of multiple sclerosis and myasthenia gravis.

Sex differences in schizophrenia have been identified in terms of the premorbid history, symptoms, brain morphology and functioning, neurochemistry, genetic transmission, course, treatment response, and epidemiology. In fact, being a man or woman probably places one at differential risks for particular forms of schizophrenia. Schizophrenic men have an earlier age of onset, the preponderance of olfactory disturbances, poorer premorbid history, more negative symptoms, poorer response to neuroleptics, lower family

morbidity risk, and a 3-fold decrease in fertility relative to schizophrenic women. The earlier brain *maturation* in terms of function observed in girls relative to boys may protect females from this disease during youth, presuming that an immature brain, typical of the young male, is more susceptible to complications during childbirth, trauma, and infections. This is consistent with the fact that birth complications are more frequently linked to schizophrenia in males. In contrast to a prevalence of schizophrenia among males at an early age of onset, there is a prevalence of women among late-onset schizophrenics: among people over 75, the female to male ratio is 20:3. Again, it has been suggested that declining estrogen levels and the resulting changes in neurotransmitter systems may influence the risk and course of schizophrenia in women at this age.

More women than men on neuroleptic medications to treat schizophrenia and other psychoses exhibit movement disorders such as tardive dyskinesia and parkinsonism. These problems are often aggravated when estrogen levels drop during the menstrual cycle and menopause, suggesting a "protective" role for estrogen; likewise, they may be improved by estrogen treatment. These movement disorders involve the dopaminergic system, upon which estrogen may act. In laboratory animals, estrogen receptors are present on cells that contain tyrosine hydroxylase; therefore, by modulating tyrosine hydroxylase, the dopaminergic system may be modulated to prevent movement disorders.

VII. SEX DIFFERENCES IN CHILDHOOD

Speech and language develop earlier in girls than in boys, and girls remain ahead of boys on tests of articulation through much of grade school. Sex differences favoring females that begin in early childhood and persist into adulthood are found on tests of clerical speed and accuracy, spelling, grammatical ability, and verbal fluency. However, consistent sex differences in vocabulary and verbal memory and reasoning ability have not been established. Males perform substantially better than females on several tests of visuospatial abilities, including certain mazes, right–left discrimination tasks, judgment of horizontality, and mechanical reasoning. It is unclear whether these differences are as substantial in childhood as in adolescence and adulthood. It is believed that sex differences in childhood cognitive skills are partly related to sex differences in the rate of brain maturation.

Sex differences in childhood play behavior and interests may enhance the development of sex differences in both verbal and visuospatial skills. Males typically engage in more rough-and-tumble play and play fighting, whereas girls more often play with dolls and "dress-up" articles. There are striking sex differences in certain toy preferences that even toy companies have attempted and yet failed to alter.

Childhood play behavior clearly differs between boys and girls, and this appears to be influenced by hormone levels perinatally. Play behavior in animals, including monkeys and rats, exhibits sex differences that are dependent upon androgens during perinatal life. For example, male rats treated with antiandrogen during neonatal life engage in less rough-and-tumble play. Likewise, female monkeys exposed *in utero* to androgens exhibit more male-like play behavior as juveniles.

Groups of individuals who may have been exposed to altered hormone levels during perinatal life also exhibit sex-reversed patterns of play behavior. Homosexual men and women report a higher incidence of gender-atypical behavior during childhood than do heterosexual people. Girls with CAH who have been exposed to higher androgen levels during development exhibit more masculine patterns of childhood play behavior, including a greater degree of outdoor play activity and expenditure of physical energy.

VIII. HOMOSEXUALITY AND TRANSSEXUALITY

Homosexuals have sex-atypical sexual orientation: they are erotically attracted to people of the same sex. Transsexuals have sex-reversed gender identity: they identify themselves as members opposite of their genital and legal sex, although they may be hetero-, bi-, or homosexual in sexual orientation. The nature–nurture debate about the origin of homosexuality and transsexuality has been ongoing for over a century. Although past attention has focused on sociocultural factors, more recent studies have identified neuroanatomical, genetic, hormonal, and nonsexual behavioral differences between homosexuals, transsexuals, and heterosexuals. However, no social or biological characteristic is unique to all groups, apart from their self-proclaimed sexual orientation and gender identity.

Sex differences in the human brain are particularly relevant to the etiology of homosexuality and transsexuality because it is believed by many scientists and physicians that certain regions of the brain may differ structurally in these individuals. A prominent hypothesis

holds that the exposure to sex hormones, which sexually differentiate brain structure and function in a variety of species during a critical period in perinatal life, may have been altered during development in homosexuals and transsexuals, influencing both sexually dimorphic brain structure and the functions of sexual orientation and identity. This would result in a *correlation* between sexually dimorphic structures and functions such as sexual orientation and/or sexual identity, not because any single brain structure determines human sexuality but because the same factor—sex hormones during a critical period of early life—sexually differentiates the brain on a more *global* level.

A. Anatomical Differences

In the first published report of an anatomical difference between the brains of homosexual males and heterosexuals, two nuclei were examined: INAH-1, which had been named the SDN-POA, and the suprachiasmatic nucleus (SCN) as a control. Unexpectedly, researchers did not find that INAH-1 varied with sexual orientation. However, to their surprise, the SCN was *larger* in homosexual men than in heterosexual men and women, despite the fact that there was no sex difference between men and women in terms of volume (Fig. 2).

The following year, in a highly publicized report, LeVay reported that INAH-3 was smaller in women and homosexual men than in heterosexual men, whereas neither INAH-1 nor INAH-2 varied with sex or sexual orientation (Fig. 2). This was a classic discovery, but not for the reason that many believed. There was an underlying assumption that INAH-3, being in a region of the brain that influences sexual behavior, may also determine sexual orientation. Whereas this may be true, we still know nothing about INAH-3 except that it varies in volume with sex and sexual orientation. What is important is that INAH-3 correlates not just with sex but also with a highly sexually dimorphic behavior, namely, sexual orientation. This correlation is reminiscent in rats of the relationship between the volume of the SDN-POA and sexual behavior, both determined by sex hormones. Of course, rat sexual behavior and human sexual orientation are not the same, but both are sexually dimorphic behaviors and both correlate with sexually dimorphic brain structures. The question remains: Are both determined by sex hormone levels during perinatal life?

The following year, another structure that was sexually dimorphic, the AC, was reported to be larger at the midline in heterosexual women and homosexual men than in heterosexual men (Fig. 2). This was important for several reasons. First, there was a second structure that varied not just with sex but more specifically with sexual orientation toward a given sex. Second, the AC is not known to be related to sexual behavior, suggesting that sexual orientation involves regions in the brain not directly related to sex or reproduction. This underscores the probability that differences in sexual orientation, like differences in genetic sex, involve not just sexuality but the overall way an individual processes information. Third, it suggests that no single structure underlies sexual orientation; rather, the homosexual brain may be organized differently on a global level. No doubt, other differences in the structure of the brains of homosexuals will be discovered.

In a small sample of male-to-female transsexuals, the volume of the central division of the BNST innervated by vasoactive intestinal polypeptide was greater in heterosexual and homosexual men than in male-to-female transsexuals and heterosexual women. In this sample, the difference appeared to be present regardless of the sexual orientation of the transsexuals (Fig. 2).

There are several caveats to these four studies: (1) in the sexual orientation studies, most of the homosexual subjects died of AIDS, whereas while most of the heterosexuals did not. AIDS is associated with severe neuropathology such as atrophy. However, if mere AIDS-related complications explain why INAH-3 is *smaller* in homosexual men, it is unlikely to account for the fact that the AC is *larger* in homosexual men. (2) The homosexuals and transsexuals probably had significantly different hormone levels relative to the heterosexuals: individuals with AIDS generally have depressed testosterone levels, whereas transsexuals are treated with the gonadal hormones of the opposite sex. Circulating levels of sex hormones influence several sexually dimorphic structures in various species. (3) Sexual orientation was determined not directly from the individual but from medical records. Some subjects classified as heterosexual may have been bi- or homosexual. This problem would not create false-positive results but could obscure a difference that may even be more dramatic. (4) Homosexual women were not evaluated because sexual orientation is not stated in the autopsy records of females, because their sexual orientation is not associated with illness. These problems reinforce the importance of examining healthy living human subjects by using modern imaging technology.

In contrast to laboratory animals, it is believed that hormones play a proportionally lesser role in sex differences in the human brain. However, in view of the findings that gonadal hormones sexually differentiate nearly every sex difference in neuroanatomy and function studied thus far in laboratory animals, the correlation between sexually dimorphic brain structure and sexuality supports the notion that gonadal hormones may also sexually differentiate sexual orientation and identity during early life in human beings. An important question remains: If altered levels of sex steroids during early life can result in altered sexuality, what causes these alterations in steroid hormone levels in the first place?

B. Prenatal Stress

It has been speculated that prenatal stress slightly increases an individual's chance of becoming homosexual. This hypothesis has stemmed from studies in rats. Exposure of female rats during the last trimester of pregnancy to stressors such as heat, restraint, and/or bright light results in modifications of brain structures and function. This has been termed the prenatal stress syndrome. Researchers believe that prenatal stress increases the secretion of adrenal androgens by the mother that, in males, leads to a shift in the peak of testosterone necessary for sexual differentiation of the brain from day 18 or 19 to day 17. This change in testosterone levels influences prenatally stressed males in several ways: as juveniles, they exhibit a more feminine pattern of play behavior, with a decrease in rough-and-tumble play; as adults they show less male copulatory behavior and increased rates of lordosis; their brain structure develops in a feminine direction (the SDN-POA is significantly smaller than in the control males); and regions of the cerebral cortex that normally exhibit asymmetry in males are more symmetric as in females. Prenatal stress in females disrupts adult estrous cycles. Like prenatally stressed male rats, some male homosexuals exhibit less rough-and-tumble play during childhood than heterosexual males, and a female pattern of brain structure is observed in terms of INAH-3 and the AC.

Of course, pregnant women are not physically stressed in the same way as experimental rats, but many undergo psychological stress that can also alter adrenal function. In humans, it has been suggested that prenatal stress increases the incidence of left-handedness. Interestingly, the results of some studies suggest that there is a higher incidence of left-handedness in male and female homosexuals and transsexuals. Of course, the mothers of most homosexuals and transsexuals were not stressed while pregnant, nor are most of their homosexuals and transsexual children left-handed. Likewise, most women stressed during pregnancy do not have homosexual or left-handed children. However, the concept that hormonal alteration by a factor such as stress can lead to other clues regarding the origins of homo- and transsexuality. For example, genetic factors, which might alter prenatal adrenal activity, the timing of the testosterone surge, or the synthesis of steroid hormone receptors, might also contribute to variations in sexuality.

C. Genes of Sexual Orientation

Several studies suggest that homosexuality may be genetically influenced: homosexuals have more homosexual siblings than do members of the general population, and genetic markers have been associated with male sexual orientation. However, a family pattern of homosexuality could be caused by a similarity in genes and/or shared home environment. Researchers turned to families with twins and adopted siblings to perform studies examining the genetic component to this pattern. They reasoned that, if genes are involved in homosexuality, then the more closely related two siblings are genetically, the more likely they should be concordant for sexual orientation. Indeed, for both men and women this holds true: identical twins exhibit the most concordance, followed by fraternal twins, and finally adopted siblings. However, in only about half of the identical twins are both homosexuals; in the other half they are discordant for sexual orientation, suggesting that genes alone do not determine sexual orientation.

DNA studies in homosexual and heterosexual individuals provide the greatest chance of identifying genetic determinants of sexual orientation. A linkage between DNA markers on the X chromosome in a region called Xq28 and male homosexuality has been reported; however, not all researchers seeking such a linkage have been successful.

IX. CONCLUSION

Sex differences in the human brain have broad implications for understanding neural development and maturation, reproductive and sexual functions, cerebral lateralization, cognition and information processing, disorders and their treatment, and the

process of aging. Research in laboratory animals has profoundly influenced our understanding of the biological component of sex differences in the human brain.

In nonhuman animals, both the body and the brain are sexually differentiated by gonadal hormones during a critical period of early life. Likewise, the body of the human is sexually dimorphic as a result of androgens perinatally. Sex differences in both the structure and function of the human brain have been identified, and, in all likelihood, they too are determined not just by culture that may accentuate or diminish certain sex differences but also by levels of sex hormones during early life. In humans, variations in levels of sex hormones during development are likely to have organizational effects on the brain, possibly resulting in variations in cerebral lateralization, cognition, sexual identity, and orientation. In adulthood, sex hormones may have activational effects on the brain, influencing reproduction and sexuality, emotions, cognition, aging, and a variety of diseases.

There is a complex relationship between sex, development, gonadal hormones, brain structure, cerebral lateralization, cognition, and handedness. Although certain correlational patterns exist, the considerable overlap between males and females suggests that factors other than genetic sex and hormones participate in the determination of these characteristics and their interrelationships.

See Also the Following Articles

AGING BRAIN • CORPUS CALLOSUM • DEPRESSION • EVOLUTION OF THE BRAIN • LATERALITY • LEFT-HANDEDNESS • OXYTOCIN • SEXUAL BEHAVIOR • SEXUAL DIFFERENTIATION, HORMONES AND • SEXUAL DYSFUNCTION • SEXUAL FUNCTION • TOURETTE SYNDROME AND OBSESSIVE COMPULSIVE DISORDER

Suggested Reading

Allen, L. S., and Gorski, R. A. (1991). Sexual dimorphism of the anterior commissure and massa intermedia of the human brain. *J. Comp. Neurol.* **312,** 97.

Coffey, C. E., Lucke, J. F., Saxton, J. A., Dphil, G. R., Unitas, L .J., Bilig, B., and Bryan, R. N. (1998). Sex differences in brain aging. *Arch. Neurol.* **55,** 169.

Cowell, P. E., Turetsky, B. I., Gur, R. C., Grossman, R. I., Shtasel, D. L., and Gur, R. E. (1994). Sex differences in aging of the human frontal and temporal lobes. *J. Neurosci.* **14,** 4748.

De Lacoste-Utamsing, C., and Holloway, R. L. (1982). Sexual dimorphism in the human corpus callosum. *Science* **216,** 1431.

De Vries, G. J., De Bruin, J. P. C., Uylings, H .B. M., and Corner, M. A. (Eds.). (1984). Sex differences in the brain. *Progress in Brain Research* **61**. Elsevier, Amsterdam.

Gorski, R. A. (1996). Gonadal hormones and the organization of brain structure and function. In *Nobel Symposium. Lifespan Development of Individuals: Behavioral, Neurobiological, and Psychosocial Perspectives* (D. Magnusson, Ed.), p. 315. Cambridge University Press, Cambridge.

Gorski, R. A. (2000). Sexual differentiation of the nervous system. In *Principles of Neural Science* (E. R. Kandel, J. Schwartz and T. Jessel, Eds.), p. 1130. McGraw-Hill, New York.

Hier, D. B. (1979). Sex differences in hemispheric specialization: Hypothesis for the excess of dyslexia in boys. *Bull. Orton Soc.* **29,** 74.

LeVay, S. (1993). *The Sexual Brain*. MIT Press, Cambridge, MA.

Matsumoto, A. (Ed.) (1999). *Sexual Differentiation of the Brain*. CRC Press, Boca Raton, FL.

Witelson, S. F. (1989). Hand and sex differences in the isthmus and genu of the human corpus callosum. *Brain* **112,** 799.

Sexual Behavior

JOHN D. BALDWIN and JANICE I. BALDWIN
University of California, Santa Barbara

GLOSSARY

erotophiles People who like and have positive attitudes about sex.

erotophobes People who have a fear or phobia about sexuality and sexual information.

operant learning Behavior that is followed by rewards tends to become more frequent in the future, and behavior followed by punishment tends to become less frequent.

Pavlovian conditioning Stimuli that precede and predict many reflexes take on the power to elicit reflexive responses.

psychogenic stimulation Excitation that has its origins in mental activity.

scripts Patterns of talk and action that people learn from other people and from books, movies, and other media. There are sexual scripts for dealing with every stage of relationships, from meeting a potential partner to engaging in intercourse.

The human brain is complexly involved in the production and enjoyment of sexual behavior. During the last 50 years, scientific research has shown how sexual behavior is influenced by multifaceted interactions of nature and nurture, brain and peripheral reflexes, thoughts and emotions, and learning and planning. We present an empirically based analysis that integrates many of the better understood elements of this complex equation along with their interactions. This discussion of sexual behavior includes emotions and subjective experiences, even though these topics are sometimes neglected in biological models: cognitive events are open to scientific study and can be included in comprehensive theories. Because sexuality is an important topic that deserves to be widely understood, we avoid overly technical vocabulary when possible.

I. INTRODUCTION

From a bioevolutionary perspective, sexual behavior functions primarily to assure reproduction, and throughout most of human evolution, sexual activity was closely related to pregnancy and childbearing. Through evolutionary processes, numerous physiological structures have emerged that usually make sexual behavior pleasurable and easy to do, though sexual pain and problems are not uncommon. The thoughts, emotions, and physiological responses involved in sexual behavior are mediated by the brain and spinal cord. The *limbic system*, which encircles the upper end of the spinal cord below the cortical hemispheres in the brain, plays a crucial role in regulating emotions and sexual behaviors. Stimulation of parts of the limbic system can produce sexual arousal. The *cerebral cortex*, which processes memory, fantasy, language, and thinking, has multiple connections with the limbic system and plays multiple roles in sexual behavior. These neural structures are, in turn, influenced by countless biological, psychological, and social factors.

Although neural and hormonal mechanisms help to explain many aspects of sexuality, they are also influenced by social and cultural factors. Countless mixes of social and cultural inputs—mediated by the brain at conscious and unconscious levels—make human sexuality far more complex than the more basic sexual activities seen in other species.

Before the development of reliable, modern contraceptive techniques, sexual behavior was closely linked

with reproduction in most societies. Whereas people in many parts of the world attempted to unlink sex from pregnancy by using potions, magic incantations, and pessaries (inventions worn in the vagina), most had little success. Given the close connection between sex and procreation, most cultures surrounded sexuality with rituals and morals that demanded and/or sanctified family commitments, even though many people failed to follow those rules in their entirety. In the last four decades, the development of increasingly effective contraception and abortion techniques has allowed increasing numbers of people to avoid "procreational sex" (except when they want babies) and explore "recreational sex" (the fun side of sex). This adds to the already considerable variability within and among cultures in their sexual values and practices.

In order to appreciate the vast complexity of human sexual behavior, let us begin with the most basic elements of sexual behavior and then repeatedly add layers of extra information. One of the human brain's main functions is to process learning, and we will see how Pavlovian, operant, and social learning all add to and complicate sexual behavior and experience.

II. THE SEXUAL REFLEXES

There are reflex mechanisms located in the lower half of the spinal cord that mediate vaginal lubrication, penile erection, and other basic genital responses that can lead to arousal, orgasm, and resolution (after orgasm). The sexual reflexes function from birth: soon after birth, boys may have penile erections and girls may have vaginal lubrication and clitoral tumescence. Even before birth, ultrasound studies suggest that erections occur in male fetuses. The reflexes are simple stimulus–response actions at first, and years of sexual learning are needed to add the multiple layers of complexity that are typical of adult sexuality.

A. Before Conditioning

The basic sex reflex occurs when touch to the genitals triggers penile erection and vaginal lubrication. With prolonged tactile stimulation of appropriate intensities, the reflex mechanisms eventually produce all of the genital changes that lead up to orgasm and resolution (which is the end of the sex reflex).

Starting from birth, the sexual reflexes in the lower spinal cord send signals to and receive signals from the brain. The brain can (1) sense inputs from the genitals and sexual reflexes and (2) cause sexual responses in the genitals. The connections between the brain and the spinal reflex centers allow sexual responses to influence and be influenced by cortical and subcortical processes, including learning, memory, emotions, thoughts, and fantasies. Nerve fibers from the autonomic nervous system connect the internal and external sexual organs to the central nervous system.

Reflexes are neural mechanisms that allow stimulus–response (S–R) activities to occur without any prior learning. Touch-sensitive nerves in the genitals send afferent signals to the reflex centers located in the lower spinal cord. Tactile stimulation of the genitals serves as the *unconditioned stimulus* (US) that activates the reflexes; the word "unconditioned" indicates that no prior conditioning is needed for these stimuli to activate the reflex. All of the sexual responses are called *unconditional responses* (URs) to indicate that no prior learning is needed for them to occur.

From birth, the sexual reflexes can be activated by gentle touch to the genitals, without inputs from the brain. By adolescence, males tend to prefer stronger tactile stimulation of the genitals than do females, but both males and females have thresholds beyond which additional stimulation elicits pain. The most sensitive parts of the female body are the glans and shaft of the clitoris (even though the shaft is hidden beneath the skin), the minor lips (surrounding the outside of the vaginal opening), and the frenulum of the clitoris (located where the minor lips connect with the glans of the clitoris). The most sensitive parts of the male body are the glans and shaft of the penis, along with the frenulum (the small strip of skin that extends along the lower side of the penis, where the glans and shaft join). Touch of these structures, with the appropriate pressure and patterns of movement, activates the spinal reflex centers, and continued stimulation produces a long chain of several distinct responses—called excitement, plateau, orgasm, and resolution—that often take from 2 to 20 min for a person to experience all four of these phases. Hence, the sex reflexes are quite different from such quick S–R reflexes as the patellar knee jerk or jerking away from electric shock. The following paragraphs present a simple summary of our current knowledge about the ways in which the sexual reflexes function.

Appropriate tactile stimulation to the genitals first activates parts of the sacral reflex mechanisms (in the lower vertebrae of the spinal cord). Early sacral activities relax the muscles in the arterial walls, allowing the arteries to dilate and carry blood into the genitals faster than usual. The excess blood in the

arteries compresses the veins, slowing the flow of the blood from the genitals. As a result, blood accumulates in the pelvic area, creating *vasocongestion* of the genitals. This vascular engorgement of the genitals produces the earliest phase of the sexual response, called the excitement phase. Vasocongestion causes the male's penis to become larger and erect. It also causes the female's clitoris to enlarge by a small amount and fluids to exude through the vaginal wall, lubricating the inside of the vagina (making it more receptive to penetration).

After the sacral reflexes are activated in the excitement phase, continued genital stimulation transfers reflex control to the thoracic and lumbar sections of the spinal cord (located above the sacral area). This initiates the plateau phase of the sexual response, when the thoracic and lumbar reflexes cause even stronger vasocongestion, resulting in additional vaginal lubrication for females and stronger erections for males.

Continued sexual stimulation eventually causes the thoracic and lumbar reflexes to initiate a several-second transition period that leads to orgasm. In males, the thoracic and lumbar reflexes cause a 2- or 3-sec-long response called "ejaculatory inevitability" during the first several seconds of the orgasm phase. At this time, men can feel pleasurable genital sensations that come from rhythmic muscular contractions in the walls of the male internal reproductive organs—the vas deferens, seminal vesicles, and prostate gland—as these organs push their contents into the tubes from which they will be ejaculated. Then reflex control is transferred to the sacral reflexes, which initiate ejaculation and orgasm. Most men learn they cannot do anything to stop ejaculation from occurring once the 2- to 3-sec period of ejaculatory inevitability starts, hence its name. After this brief preejaculatory period, reflexive control is transferred to the sacral reflexes, which now cause rhythmic muscle contractions around the base of the penis (inside the body), starting at intervals of 0.8 sec and then gradually slowing to longer intervals. These muscle contractions force the ejaculate out of the male's body at the time of orgasm, and they cause orgasm to be experienced as rhythmic pulses or surges of pleasure.

In females, the transition from plateau to orgasm is similar to that seen in males, but more subtle. During the plateau phase of the sexual response, the thoracic and lumbar reflexes produce increasing vasocongestion, which augments vaginal lubrication. Then, for about 2 or 3 sec before the sacral reflexes cause contractions in the pelvic muscles at 0.8-sec intervals, women report the onset of pleasurable experiences (at the time when males feel the pleasurable sensation of ejaculatory inevitability). After this 2- to 3-sec period, females begin to experience the rhythmic surges and pulses of orgasmic pleasure that occur as their genital muscles begin to contract at intervals of 0.8-sec, gradually slowing to longer intervals. Thus, both females and males have 2–3 sec of pleasure before reflexive control is transferred to the sacral reflexes (which cause the muscle contractions that begin at 0.8-sec intervals). Even though females do not ejaculate semen, the contractions of the pelvic muscles are experienced as rhythmic pulses of pleasure. Many females and males report that orgasm lasts for about 15 sec, though this timing shows considerable variability.

After orgasm, the final phase of the sexual reflex is resolution, in which all of the genital responses seen in the prior three phases subside. There is a reversal of vasocongestion, and blood flows out of the genitals faster than it flows in. Most muscle contractions slow and weaken, though small movements may continue for a while. Gradually, the genitals return to the relaxed condition seen before the onset of sexual stimulation. After orgasm, most males experience a refractory period during which no additional stimulation can activate the sexual reflexes. In adult men, the refractory period tends to last 30–90 min, but the period becomes longer with age. At this time, the fluids that will be ejaculated at the time of the next orgasm flow into the internal male reproductive organs. Females do not have an equivalent process and can have multiple orgasms without several-minute pauses between them.

Although the sexual reflexes usually function rather automatically, as described earlier, they can be partially or completely inhibited by pain, fear, depression, fatigue, exhaustion, illness, biomedical problems, and certain drugs, even when effective tactile stimulation is present. Before sexual therapy, a medical exam is usually conducted to determine whether a sexual problem is the result of medical or drug problems. If these causes are ruled out, psychological treatment is in order.

B. Pavlovian Conditioning

In humans, the sexual reflexes are among the reflexes that are influenced by Pavlovian conditioning, which is also known as classical and respondent conditioning. This type of conditioning allows stimuli other than the tactile genital stimuli that are the USs (unconditioned

stimuli) for the sex reflex to gain the power to elicit sexual responses. Pavlovian conditioning is adaptive from a biological perspective, because it allows an individual to learn to respond to many stimuli that are associated with crucial USs, all of which are important for survival and reproduction. Thus, the brain has evolved to be quite sensitive to correlations—or contingent relations—between reflexes and the stimuli that regularly precede them.

Pavlovian conditioning of sexual responses occurs in the brain when some neutral stimulus repeatedly precedes and predicts the onset of the US of appropriate touch to the genitals. The brain can associate stimuli from all sense modalities with the USs that trigger the reflexes in the lower spine. Neutral stimuli that predict the onset of tactile genital stimulation (the US) gradually become *conditioned stimuli* (CSs) capable of eliciting sexual reflexes and pleasurable sexual feelings. For example, the sound of an electric vibrator is a neutral stimulus that does not elicit sexual responses in children, but adults who have orgasms with vibrators learn that the sound of these electric devices is closely associated with sexual pleasures. After such sexual conditioning, some people become sexually aroused when they hear the vibrator's sounds (the CSs) before having any tactile stimulation to the genitals (the US). The cues that most reliably precede and predict the onset of USs are the stimuli that become conditioned stimuli via Pavlovian conditioning. Once a CS has been created, it can elicit *conditioned responses* (CRs), which are similar to the URs of the original reflex from which they take their power, though CRs are usually weaker than URs and slower to appear.

Although some people think that seeing naked bodies might be a US for the sexual reflexes, research does not support this supposition. Young children who see naked bodies do not respond with sexual reflexes, but they do have sexual responses to genital touches (the US for the sexual reflex). When people learn that the sight of naked bodies can precede and predict the onset of the US of tactile genital stimulation, they learn to respond to the sight of naked bodies as a CS that elicits sexual responses. In common parlance, CSs are called "erotic stimuli" or "sexual turn-ons," and they can elicit conditioned sexual responses even when the US of touch to the genitals is not present. Conditioned sexual responses (CRs) are not identical with the URs of the unconditioned reflex, because the CSs activate different brain circuits than the USs. If the sight of naked bodies is a CS for a person, it stimulates visual centers, whereas the US of genital stimulation directly activates the spinal reflexes, along with nonvisual brain areas. In addition, CSs elicit responses that are weaker and slower to appear than those elicited by the USs, which directly stimulate the spinal reflex mechanisms.

The effects of CSs can vary from weak to moderately strong. Usually, the more frequently that any given CS has preceded and accurately predicted the onset of the US of genital stimulation, the more power that CS develops for eliciting conditioned sexual responses. People who have had numerous orgasms with vibrators are most likely to respond to the hum of a vibrator as a CS for sexual arousal. However, the effects of strong sexual CSs can be opposed and suppressed by pain, fear, fatigue, illness, or other CSs that elicit aversive emotions.

Through Pavlovian conditioning, almost any stimulus that frequently precedes and accurately predicts the onset of sexual arousal can become a sexual CS. The specific stimuli that people respond to as sexually exciting differ from person to person, depending on each individual's unique history of sexual experiences. For example, foul language can affect people quite differently. Some people learn to find "talking dirty" to be sexually arousing, if such talk has been a common part of their more exciting sexual experiences. Although many people are offended by vulgar words and wonder how anyone could find obscenities to be sexually exciting, the principles of Pavlovian conditioning clarify the phenomenon: almost *any* stimulus can become a sexual CS—an erotic stimulus or sexual turn-on—for those individuals who have experienced it in conjunction with intense sexual excitement.

Because the cerebral cortex is connected with the sexual reflex mechanisms in the lower spinal cord, sexual thoughts and fantasies that include sexual turn-ons (CSs) can elicit vaginal lubrication, penile erection, and other sexual responses without any tactile stimulation of the genitals (the US). When the brain's sexual thoughts are the only causes of sexual reflexes, the sexual reactions are called *psychogenic* responses, to indicate that they are purely psychological in origin. People with little sexual learning often find that psychogenic stimulation only produces mild sexual responses, such as slight vaginal lubrication or partial penile erection. But individuals who have had many hours of intense sexual experiences can have strong psychogenic responses to the CSs present in their sexual lives.

Many people find that fantasies help their sex lives in two ways. First, during tactile sexual activities, such as masturbation or partner interactions, fantasies and

psychogenic sexual stimulation can add to the total excitement that enhances and intensifies the sexual response. Second, fantasies also help distract people from thinking about some of the negative things that can suppress their sexual reflexes, thereby reducing sexual inhibitions. Fantasies can distract people from unwanted fears of not being attractive enough to their partners, doubts about the adequacy of their sexual performance, or guilt about engaging in sexual activities. Unfortunately, some people feel guilty about having sexual fantasies while having sex with a partner: in their lives, fantasies elicit negative thoughts and emotions, which interfere with the two benefits of fantasy.

Sex is not always associated with pleasure. When sexual activities precede and predict the onset of aversive experiences, such as loss of erections or failure to reach orgasm, sexual activities can be conditioned into CSs that elicit fears that suppress the sexual reflexes. Any of a variety of sex-related stimuli can become associated with fear, worry, or anxiety. For example, fears of becoming pregnant can inhibit a person's sexual response. In this case, the onset of sexual activities can trigger thoughts about pregnancy, and these are CSs that elicit fears and worry that can interfere with the sexual response, decreasing the chances of reaching orgasm. People can combat this fear by learning about highly effective contraceptive methods, which offer genuine protection from pregnancy and reduce their users' fears of fertility that inhibit their sexual responses. Other thoughts that can become CSs associated with physical intimacy and inhibit sexual responses are fears of not pleasing one's partner, not being attractive enough, and not being the best sexual partner one's partner ever had. Both males and females can have these fears and never know that one's partner has some of these feelings, because many people are reluctant to communicate about such topics.

''Performance anxiety'' is a common sexual fear that arises when people worry that their sexual performance may not be as ideal as they would like. She may fear that she may not reach orgasm; he may fear that he will ejaculate too early to please his partner. To combat performance anxiety, people first need to realize that their sexual responses are based on reflexes that usually appear rather automatically: by increasing or decreasing the number of USs and CSs that enter the sense receptors of the body and brain, people can intensify or delay the sex reflexes. Second, knowledge of this fact and adjustment of their sexual activities accordingly will convince them that their performance anxiety is irrational, and that they can forget it.

III. OPERANT LEARNING

In humans and many other species, the sexual reflexes and Pavlovian conditioning are not adequate—in the absence of other forms of learning—to assure intercourse, successful reproduction, and the continuation of the species. Humans need considerable *operant learning* to coordinate successful sexual activities with a partner. Namely, they have to learn numerous skills for ''operating'' their way through a complex series of activities and interactions that are needed to work up to and perform sexually activities, at which time the sex reflex produces penile erection, vaginal lubrication, and orgasm.

A. First-Hand Experience with Reinforcers and Punishers

Operant conditioning, also known as instrumental conditioning, occurs when any behavior is modified by the consequences that follow it, be they pleasant or painful, or more technically, be they reinforcers or punishers. People learn to do behaviors that are instrumental in attaining pleasure and avoiding pain (attaining reinforcement and avoiding punishment). If two people try ten different sexual positions and find that three lead to pleasurable outcomes, but two other positions led to painful consequences, the couple is especially likely to repeat the three positions that brought reinforcement and avoid the two followed by pain. During operant learning, behavior that is followed by reinforcers is strengthened and made more likely to occur at future times, and behavior that leads to punishers is suppressed and becomes less likely to occur in the future.

From an evolutionary perspective, it makes sense that humans would be biologically prepared to learn countless behaviors for attaining sexual pleasures. The human cerebral cortex, which plays a central role in learning, is quite large in proportion to the human body size, and learning how to have sex in almost any kind of ecological situation guaranteed the reproduction of the species throughout most of human history. It could be argued that humans have, in fact, been too successful in reproducing to the point where our large population is destroying many ecosystems and causing other species to become extinct. Only relatively recently, with the development of modem birth control technologies have humans been able to learn how to enjoy sexual reinforcers without risking pregnancy.

During sexual activities, all of the external and internal sense organs activate the sensory centers in the brain and arouse sexual perceptions and motor activations. In addition, some sensory inputs activate the pleasure or pain centers in the brain, causing certain types of sexual stimulation to feel pleasurable, whereas other types feel aversive. The human brain is very sensitive to (though not perfectly perceptive of) contingent relationships between behavior and its consequences, especially pleasure and pain. Special neural circuits associate rewards and punishers with the behaviors that precede and predict them, strengthening behaviors associated with reinforcers and suppressing behaviors associated with punishment. This makes the sensations of pleasure and pain the prime movers that produce operant learning. Through the hippocampus, basolateral amygdala, and other areas, reinforcement and punishment also facilitate the development of long-term memories that influence later thoughts and actions.

Most operant behavior occurs as chains of activities, and sex is no exception. Two adults may spend 10, 20, or 60 min performing a long chain of varied activities before they reach orgasm or decide to end sexual activities. It usually takes years of experience to learn long behavioral chains that smoothly link many different operant activities. The following paragraphs provide examples of some of the countless routes by which children and adults learn longer and more complex chains of sexual behavior as they gain increasing operant learning experience with numerous behaviors and their consequences.

1. Infancy

Starting in the early months of life, some infants learn simple chains of sexual behavior that lead to sexual pleasures. Out of simple curiosity, infants use their hands to explore touching many parts of their bodies and things in their environments. After touching and feeling anything nearby, infants often—by accident, with no plan—touch their genitals. Touch to the genitals is the US that elicits the sex reflex and activates the pleasure centers in the brain. Next, the brain mechanisms that sense contingent relationships between behavior and rewards—and underlie operant conditioning—reinforce and strengthen the motor mechanisms that produced touching and feeling the genitals. If infants are not punished for self-stimulation, they gradually learn to repeat the operant behavior of touching and feeling the genitals.

Infant self-stimulation consists of simple chains of behaviors of repeated touches and manipulation of the genitals, and infants can learn these chains easily. Adults call the behavior masturbation, but infants learn the behavior without the adult vocabulary or any verbal awareness of society's views or values about masturbation. As the months pass, some infants gain sufficient experience to learn long enough chains of touching, feeling, pressing, and rubbing that they reach orgasm by their second or third birthday (though boys, at the time of orgasm, do not ejaculate any fluid until after the onset of puberty).

There is considerable variation among infants in early sexual self-stimulation, depending largely on their parents' responses to the activity. Some parents punish and suppress early sexual behavior, whereas others are more accepting. Children who are punished for genital exploration may learn not to touch their genitals. In fact, they may become so well-inhibited that they do not learn enough skills for tactile self-stimulation to become sexually aroused, much less reach orgasm. Once these children reach adolescence, they will not remember how they learned to feel inhibited—and perhaps even fearfull—about masturbating, because few of us can recall experiences from the first couple years of our lives.

2. Childhood

As children play with each other and learn increasing social skills, some discover activities that can lead to sexual interactions. During early social exploration and play, without any intention of discovering coitus, some children undress and feel each other's bodies, simply out of curiosity, without any sexual motivations. Children are inquisitive and eager to learn about their world, including what other children's bodies look like. Because children's play is not always monitored closely by adults, sexual exploration can lead in many possible directions. Often it leads to nothing more than one child's simply looking at and briefly touching the other child's body and then losing interest, but other outcomes are possible. A 9-year-old brother and his younger sister may innocently engage in mutual sexual exploration and discover that sexual play is pleasurable. Any sex play that happens to involve genital touching is rewarded by the positive reinforcers of genital stimulation, and this strengthens any behavioral patterns that lead up genital stimulation. Due to reinforcement and operant learning, children may learn to play these "touching games"

with increasing frequency, unless they are taught to avoid these kinds of "naked games."

After weeks or months of repeated sex play, children may discover intercourse if adults have not inhibited sex play. If children learn behavioral chains that are effective in eliciting the sexual reflexes of penile erection and vaginal lubrication, they have the opportunity for discovering penile–vaginal sex. The brain has no innate sexual taboo or inhibitory mechanisms that prevent children from exploring sexual behavior, and childhood sexual interactions are not rare. Even children who are raised by parents who have strict sexual standards may accidentally discover sex play before it occurs to their parents to teach their values to the children. Only later may the children discover that these activities are called "incest," "sinful," or "dirty," and these social labels can make them feel shameful afterward.

What happens in homes where parents systematically punish early childhood sexual exploration and give their children little information about the sexual activities and experience that many children have? Such children can grow up with considerable guilt and little knowledge about their sexual organs and self-touching. Small doses of sexual inhibition may not have long-lasting effects on the person's sexuality, but high levels of guilt and sexual ignorance can make it difficult for a person to develop a happy, uninhibited sexual life in subsequent years. Children who have punitive socializations about sex often develop erotophobia, which is a fear or phobia of sexuality and sexual information, leading to negative attitudes about sex. In contrast, a more positive sexual socialization tends to lead to erotophilia, with positive attitudes about sexuality and sexual information.

In most societies, including ours, there is a double standard that imposes more sexual inhibitions, guilt, and erotophobia on girls than on boys. Parents are more likely to disapprove of masturbation and sex play for girls than for boys. They tend to withhold sexual information about sexual anatomy, feelings, and responses from girls more than from boys. Girls are often told about the joys of motherhood but given little information about the clitoris, genital stimulation, or orgasm. As a result, girls and women often start life with more erotophobia than males of their same age have, and this can limit both their Pavlovian and operant learning about sex.

3. All through Life

Operant learning can occur at any age, whenever behavior leads to reinforcers or punishers. During sexual activities, a woman may have her clitoris and frenulum touched by her partner, leading to her moaning with pleasure, and her response reinforces her partner's use of that style of stimulation. His jerking back and saying "ouch" may punish his partner for pushing too strongly against his testicles. All of the presexual behaviors, such as kissing and undressing, that occur as early links in the behavioral chains that lead up to sexual acts are also influenced by reinforcement and punishment.

B. Observational Learning

Operant learning is often hastened by *observational learning*, whenever observers learn how to do some new behavior by watching a person, who is called a *model*, perform the behavior. Observational learning has two phases: acquisition and performance. People can acquire information about a behavior but never perform it. Viewers of a modern movie may see the actors engage in sadomasochistic sexual activities, and some observers may perform the behavior within 24 hr, whereas others do it within the next month and yet others never try the behavior. People are most likely to perform a modeled behavior if (1) they would not feel guilty about doing the behavior, (2) it is clear that the behavior brings considerable pleasure and little pain to the model, (3) they feel they are similar to the model, (4) the crucial details of the modeled behavior are clearly visible, and (5) the behavior is easy to imitate. Next, the performance of an imitative act may be followed by reinforcement or punishment, and these consequences affect the future chances that the behavior will be repeated.

If attempts to replicate a sexual position modeled in a movie bring pleasure, the reinforcement speeds operant learning. If the performance of a new behavior causes one's partner to feel pain and complain, the behavior is less likely to be repeated due to punishment.

Although sexual intercourse is usually not done publicly in our society, people can learn countless things about presexual and sexual behavior from models. In private, they can learn from observing and interacting with sexual partners, who provide *real-life models* (dressed or undressed) for new sexual and presexual behaviors. People also learn sex-related information from *symbolic models* whose behavior is described in conversations, magazines, and books, or shown in movies, TV, videos, web sites, or any other symbolic medium.

1. Childhood

In our society, children learn activities that can influence sex play when they observe doctors, nurses, TV, and more. Because doctors are high-status people, some children observe and learn from them and then later imitate by playing doctor, which may involve undressing a playmate. Other children innocently imitate physical activities they see modeled on TV, without guessing what pleasures they might discover if their imitative acts lead to genital stimulation and the reinforcers that strengthen operant behavior.

Anthropologists studying several preindustrial societies have observed childhood sex play that leads all the way to copulation. In cultures where a whole family sleeps together in a hut or 1-room house with no walls between the adults and their offspring, children may see their parents engaging in coitus and learn the positions that are involved. The next day, the children may imitate their parents' coital activities during play with other children, and discover some of the pleasures of genital stimulation. In some societies, childhood sexual interactions are punished, but there are societies in which adults respond to child sex play as "natural" or "cute," and children are allowed to learn about sexuality early in life. Of course, parents in societies that do not approve of child sexual activity could create enormous problems for their child and family by condoning childhood sexual activities that the neighbors and community would condemn.

A great deal of variability exists among cultures in the treatment of childhood sexuality, and sexual exploration and play can continue throughout childhood, even in societies where parents do not approve of it. Although Freud and other Victorians knew that infants explored sexually, many thought that children between age 5 and the onset of puberty experienced a "latency period" in their sexual development—when sexual motivation was thought to be dormant—before puberty triggered the rise in adult sexual hormones and sexual interest. Modern research shows that many children in our society and other societies do not cease being sexual after early childhood, though childhood sexuality can be suppressed to a significant degree by high ratios of punishment to reinforcement.

2. Adolescence

During puberty, biology and learning interact in ways that increase the chances that teens will think, wonder, and fantasize about sex, leading increasing numbers of teens to experiment with sex with each passing year of age. Let us deal with biology first and then learning. During puberty, rising levels of adult sex hormones increase the sensitivity of the genitals and the sexual reflexes in both male and female bodies. Boys experience spontaneous penile erections during dream sleep, wet dreams, and various times during the waking day. Having a spontaneous erection can lead to masturbation. Adolescent boys report more spontaneous sexual arousal than do girls. This is in part because the penis is large enough that boys can easily notice their erections, whereas the female sexual responses of clitoral enlargement and vaginal lubrication are more subtle and difficult to notice.

Because the sexual hormones make the genitals more sensitive, stimulation to them is more pleasurable and reinforcing than it was in childhood. As a result, adolescents who explore sexual activities receive stronger rewards than they did in childhood, which leads them to think and fantasize about sex more often. These sexual thoughts and feelings lead many teens to seek sexual information from real and symbolic models or look for first-hand experience via sexual activities. As a consequence, many teens (especially boys) experience a rapid surge in learning more complex and lengthy chains of sex-related activities: how to talk with potential sexual partners, how to suggest physical intimacy, and how to explore all sorts of sex-related activities.

During adolescence, the differences between people who have learned to be erotophiles or erotophobes become especially noticeable. It is also clear that there is a continuum of responses to sex between erotophilia and erotophobia—between those people who like sex (and feel little guilt about it) and those people who feel guilt, anxiety, shame or phobia about sex. Generally, the more phobic people have so many aversive associations related to sex that they feel inhibited about seeking sexual information or education. However, their avoidance of sexual knowledge does not guarantee that they will know how to resist seductive sexual approaches by or from insistent partners. Many erotophobes have been talked out of their virginity by verbally persuasive individuals. Then we discover that erotophobes are not immune to the painful consequences of uninformed sex, such as pregnancy and sexually transmitted diseases (STDs), and their avoidance of sexual education actually puts them at a greater risk of problems than their better educated peers. All through history, many teens who have been told to remain virgins until marriage have disobeyed, experimented with sex, become pregnant and been forced into marriage, abandoned by their families, or worse.

3. All through Life

People who are exploring physical intimacies with a new partner may see sexual behavior they have never done before and acquire information about it via observational learning. Affluent nations provide their citizens with countless magazines, books, videos, and movies from which teens and adults can acquire information about sex by observational learning.

As people learn longer chains of sexual behavior, they discover the importance of many of the early presexual links of behavioral chains that come before genital stimulation and the sex reflexes of erection, vaginal lubrication, and orgasm. Through observational learning, most people discover countless ways to deal with the long chains of presexual behaviors that begin with meeting, starting a relationship, becoming intimate, suggesting sexual activities, discussing methods of protection from pregnancy and STDs, and then engaging in sexual behavior. Some links of this chain of activities are modeled in public and in the media, but others are seldom discussed openly, which deprives people of valuable information that could be useful in solving some of the problems that sexual people eventually confront. Countless people have problems in learning all of the long chains of presexual and sexual behavior needed to have optimally gratifying intimate and loving relationships, yet all of the information about each link in these long chains is not broadly available.

C. Words, Rules, and Scripts

Language facilitates the learning of operant sexual behavior, not only by presenting symbolically described models for observational learning but also by providing words, rules, and scripts that influence sexual behavior. *Words* are more important than is often recognized, as studies on sexuality clearly show. Most parents follow a double standard and give boys more useful words for the male genitals than girls receive for describing their genitals. Few young girls are told they have a clitoris, but almost all boys learn a name for their penis. Because the clitoris is smaller than the penis and most girls are not given a word to draw attention to it, few girls are empowered to ask questions such as, ''What is the clitoris for?'' As a result, girls are less likely than boys to talk or think about their genitals. In contrast, boys are verbally aware of the presence of and names for their penis, which helps them verbally describe and think about penile responses, ask for information about these things, talk with their parents and peers, and learn about their sexuality. The double standard in teaching boys more words about sex augments sexual curiosity and desire for sex information in boys and suppresses it in girls. Later, as females gain increasing access to books and symbolic media, which are great equalizers, females can begin to catch up with males in gaining access to sexual words and knowledge.

Rules are statements that tell us to perform specific behaviors or chains of behavior. Women's magazines often suggest rules for chains of presexual acts, such as, ''Give a long gaze to the person you like and when your eyes meet, give a warm and friendly smile.'' Both women's and men's magazines give rules for doing the behavior needed to help females reach orgasm, ''Using gentle circular motions, touch the frenulum, glans, and shaft of her clitoris, and gradually increase the pressure of stimulation as she becomes more aroused.''

In order to follow rules, people must understand the language and the words that others use to present rules. The words, ''Appuie moins fort, s'il te plaît,'' are meaningless to people who have not learned that these words mean ''less pressure, please,'' in French. Books, magazines, web sites, and knowledgeable friends often provide rules for directing sexual behavior, though not all sources of rules are equally accurate. People who wish to benefit from rules need to learn which sources of information are most useful.

Scripts involve long chains of rules that tell us how to talk and act with others. In one sense, we are all like actors on a stage, and each society provides countless scripts that help us learn how to act our sex and gender roles, and all other aspects of life. Society gives us countless sexual scripts, which are chains of phrases and sentences along with hints about postures, gestures, and tone of voice, that can influence every phase of a person's actions. For example, there are countless ''opening lines'' that a person can use to meet another person, ''Hi, I've noticed you in a couple of classes and would like to introduce myself.'' There are also scripts for attracting a potential partner, starting a relationship, developing intimacy, and having sexual relations. Children begin to learn male and female gender scripts early in life, long before they can see the link between female and male gender roles and sexual activities. With the onset of puberty, teens become increasingly motivated to learn the sexual scripts of their society, because mastery of smooth lines and scripts can open opportunities for rewarding presexual and sexual interactions. Peers, magazines, TV, videos, and movies provide countless models for these scripts and lines. Both females and males often repeat and creatively

alter lines they have found to be useful and rewarding in the past.

Scientific studies show that women tend to be more discerning among scripts and lines than men. Men like almost all types of friendly looks and opening lines that women use to initiate conversations. In contrast, women tend to like direct and innocuous lines from men and are turned off by cute and flippant lines.

The countless scripts for dealing with sex vary from society to society and from subculture to subculture. This can create confusion when people move from one social sphere to another or try to become intimate with someone from a different subculture. On the other hand, many people are quite creative in picking snippets from numerous scripts and chaining them together in novel and interesting ways. Thus, sexual behavior is far more varied than if each society provided only one sexual script and no one innovated or created new chains of acts and words.

Many teens are exposed to countless models, rules, and scripts that promise rewarding outcomes for hanging out, kissing, necking, hooking up, playing around, exploring, pushing limits, and seeing how far they can go with sex play. Some teens adhere to the scripts that their parents advocate, but many explore rebellious scripts if their subculture glamorizes these and provides more reinforcers than their parents can. Thus, adolescent society is a complex mixture of scripts, from tame to wild. From this tumultuous mélange, teens piece together ever longer chains of presexual and sexual behavior, though many also learn to set limits about how far they are willing to go at each point in a developing relationship.

Rewarding experiences help people learn to like behavioral links in chains that lead to good experiences in developing intimacy—from meeting and relating well to engaging in gratifying sexual acts. In contrast, punishment and painful experiences tend to inhibit future repetitions of behaviors that lead to pain. Many a teen has been punished for using "dumb come-on lines" or clumsy bedroom techniques. Especially painful are actions that expose one to being taken advantage of, hurt, or raped. Punishment motivates people to avoid future pain by setting limits beyond which they will not go at any phase of a newly developing relationship, "I don't want to have sex until I know you better." All during life, rewards and punishers lead people to learn more polished behavioral chains over time, though many people never learn as many presexual and sexual skills as they could if more accurate information were readily available.

People's preferences for scripts tend to change as they get older and move from being adolescents to young adults, then develop long-term relationships, wed, and create successful or fractious marriages. Each new generation experiences different historical settings —from Victorian properness (in the late 1800s) to e-mail and cyber-sex (today)—and these historical inputs interact with each person's unique learning experiences as the years take them from teens to twenties and beyond.

D. Cognition

Where in the body does learning leave its traces as people learn from Pavlovian and operant conditioning, assisted with information from models, words, rules, and scripts? The brain is the organ that processes and stores the inputs gained through all types of learning, including sexual learning. During learning experiences, perceptual inputs from all sense modalities trigger thoughts and feelings in both primitive and advanced brain areas. Sensory and motor experiences that are associated with reinforcement or punishment are especially likely to be entered into the brain's long-term memory systems.

Our thoughts are filled with two basic components: (1) perceptual images from present sensory inputs, along with the memory images that they arouse, and (2) the "inner words" that we "hear" in our minds as we use language to label perceptual and memory inputs in the privacy of our brains. Both nonverbal and verbal inputs can elicit emotional responses that heighten the "feelings" of inputs that are associated with reinforcers or punishers, whereas other inputs are neutral, with little or no emotional association.

Sexy images and words can come from real-life personal experiences or from indirect social learning via movies, TV, videos, and photos. When people see TV shows, videos, or movies in which certain lines, scripts, postures, or clothing lead to sexually exciting outcomes, they may find themselves repeating and rehearing those scripted events in their brains and then using parts or combinations of those scripts when interacting with others. The more that scripted words, actions, and images are associated with rewards, the more people come to like and use them.

Each person's vocabulary and verbal knowledge are gifts from society. Some people receive generous verbal gifts that allow them to cogitate about sex—and all the rest of life—in great detail. Other people receive limited or highly censored verbal labels and knowledge

about sex. People who have many detailed words for describing most aspects of sexuality can have more complex "inner verbalizations" for analyzing their sexuality in the privacy of their own brains than can people with limited vocabularies. Highly verbal people can cogitate for hours about the details of a sexual experience that a less verbal person might summarize in a few words, "It was good." Unfortunately, if the verbal information a person is given about sex is wrong, the many hours of self-analysis can lead to faulty conclusions, such as, "As long as I douche afterward, I won't be able to get pregnant."

Every human brain resides in a unique body that, in turn, is located in a singular place in time and space. Hence, each individual is influenced by his or her special position in the world. Two people living in different parts of the same city are exposed to different subsets of the multiple subcultures existing in that city; hence, they learn different information about sexual conduct from their different symbolic and real-life models. Some people obtain vast amounts of information about sex and pregnancy, relationships, birth control, and STDs, whereas others obtain little.

Each person has a singular perspective on life, and each brain learns unique thoughts about sex and gender; thus, 100 people can perceive the same event and interpret it 100 different ways. After two people see the same erotic movie, one describes it as interesting and sexually exciting, whereas the other calls it repulsive and pornographic. One person wants to imitate the behavior modeled in the movie, and the other sets limits and refuses to do the behavior. As they gain experience, people who live in pluralistic societies with many different views and values about sex often become more tolerant of sexual diversity as they hear many different points of view about life.

The brain is not a passive organ or a simple memory bank that is filled by personal and social experience. The brain is an active, dynamic, and creative organ that can modify and reconstruct the scripts and scenes it perceives in the world around it. Most people learn how to combine and recombine various inputs in creative new ways, play with ideas, free-associate, and then talk and act in creative ways. Problem solving and fantasy often enhance the brain's creative processes. Problems often cause people to search for creative solutions. When people confront new difficulties, such as increasing boredom with sex after 3 years of marriage, the problem may stimulate creative efforts to devise novel and interesting sexual experiences. Fantasies are a second stimulant of creativity. As people imagine sexual activities they have never done before, they may consciously attempt to create special sexual experiences that are uniquely their own. Through the creative recombination of different rules, scripts, fantasies, and novel ideas, people can construct sexual plans and actions that differ from anything they have ever done in the past or seen others do. Although most creative acts are not radically different from things that other people in their culture do, they do contribute to the cultural innovations that add spice to life.

E. Society

Human creativity causes societies to change and evolve as the years pass, as can be seen by contrasting Victorian sexual constraints of the late 1800s with modern sexual practices. As millions of creative people use their brains to invent novel sexual practices and share them with others, some of these innovations change the erotic practices of the larger society. Many new ideas are lost after being tried a few times, but others—usually the most rewarding ones—spread and become widely popular. Magazines and other information media may help popularize certain sexual positions or fantasies. Public debate may lead creative people to challenge old taboos or invent safer, more rewarding sexual practices.

People also change democratic societies by voting, for example, using the ballot to express their views on abortion, nude beaches, pornography, and other sex-related topics. Democratic societies need to educate each new generation to use their brains to make democracy work well. Concerned citizens need to listen to the many views voiced in their society, evaluate all options suggested by others, seek creative new directions, and take active roles in improving society, including those social conditions related to sexuality and gender.

Societies are not static, and we are in the midst of some of the most rapid social changes ever seen in history regarding the treatment of both sex and gender. The creative contributions of millions of individuals provide the raw material for these social changes. Those social practices that are most rewarding tend to gain favor, whereas practices that produce pain fall out of favor. In a large democratic society, people can have very different values about pornography, abortion, prostitution, and countless other sexual topics, so that social change concerning sexual practices is not always rapid. (The process of social evolution has parallels with biological evolution: The raw materials for biological evolution are mutations, whereas social

evolution draws on the raw materials of human creativity. Biological evolution is steered by the natural selection of the fitter individuals, whereas social evolution is shaped by the reinforcers and punishers that follow social behaviors.)

Among the countless debates about sex in our society, let us examine two that occupy many academics: the first is constructionism versus essentialism, and the second is the validity of cultural relativism. Some people argue that all human sexuality and gender roles are mental constructions, inventions of the brain that are free from any biological constraints of anatomical maleness or femaleness. Pure constructionists usually take a subjective or phenomenological point of view and state that ''everything is in the mind;'' hence, we cannot actually know the ''real world'' outside our skins because all our perceptions of the entire world are constructions located completely inside our brains. People who study human fantasies can tell you that the brain can invent almost any type of constructions about sex and gender. Constructionists also point out that our inner images of the world are biased by the words and images we inherit from our society, and, worse, there is no way to know whether these social inputs are true.

People who are critical of the scripts and images provided by their society can use the constructionist perspective for political ends. Unfair social practices (such as the subjugation of females by males) can be exposed as mere social constructions, not concrete and immutable facts. Constructionists often seek to empower people to create more liberated and utopian sex and gender roles by using their critical analyses to deconstruct and destroy old, ''status quo'' scripts and images.

Because everything we believe is basically ''in our minds,'' extreme constructionists assume that biological ''facts about sex'' are trivial—or even misleading—information for shaping our brain's constructions of sex and gender. Constructionists describe purely biological views about sex and gender as ''essentialism''—a naive belief that sex and gender are shaped solely by biological essentials, such as sex chromosomes, sex hormones, and sex differences in the brain. At its extreme, pure essentialism makes all behavior appear to be biologically determined, leading to the view that ''biology is destiny.'' For example, sociobiologists sometimes claim that ''selfish genes'' control our sex and gender behavior. This offends social constructionists for two reasons: it is a form of biological determinism, and it provides support to political causes that support very traditional and old-fashioned views of sex and gender roles. Both of these are anathemas to social constructionists.

Can our human brains determine which view is correct, constructionism or essentialism? The answer depends on the methodology we use to approach the question. If one uses introspection, one may conclude that one's brain can create the world to be anything it wants. For example, one may note that pessimists look at their world and see doom and gloom, whereas optimists look at the same physical world and see hope and happiness. Perhaps sex and gender are constructions that each brain can freely create. The methods of transcendental philosophies also suggest that brains can transcend their locations in female or male bodies and create subjective realities that are not constrained by those physical bodies.

The scientific method offers a different perspective, the one taken in this article. Scientists should use their brains to examine all types of hypotheses about sex and gender. This reveals a continuum of hypotheses, with constructionism and pure essentialism at the two extreme ends. Between the two extremes are numerous other hypotheses that interweave biological and social factors in various combinations, some being more biological than social and others having other leanings. After testing all of these hypotheses against the empirical data available from several disciplines, many scientists agree that most data on human behavior do not support theories at either extreme of the continuum of hypotheses. Most data support theories that show that both biological and social factors are important, and each influences the other. The view presented in this article is that sex and gender are shaped by biology (including sexual reflexes, neurophysiology, and hormones) and by years of learning experiences (including Pavlovian conditioning, operant learning, and social scripts). From this scientific perspective, both biological and social factors interact with the brain's creative abilities to invent new behavior patterns and social practices. Biosocial theories that blend an empirically defensible mixture of both biological and social variables may be more defensible than theories at either extreme—either total constructionism or pure essentialism.

Cultural relativism is a second major topic of academic debate. Anthropologists have conducted field studies on many societies and documented considerable variation in sexual and gender practices among cultures. For example, some societies have painful rites of passage that boys or girls have to experience before they can be treated as adults. Other societies have no such practices. Some societies allow

young people to select the person they wish to marry, whereas other societies have arranged marriages in which the parents select the spouse for their child to wed. In Japan today, many marriages are still arranged, though Western-style "love matches" are becoming more common.

Extreme cultural relativists argue that there can be countless different cultural variations on sex and gender behavior, and none is more valid than others. All through history, human brains have created a large number of different societies on our planet—with different sex and gender practices—so who can say which one is "right" or "better than others"? Hence, extreme relativists argue that no culture should sit in judgment of the sexual or gender practices of other cultures. When one nation takes the role of judging other cultures and finds faults, those other cultures often return with critical assessments of the judgmental nation, asking what right it has to condemn them. "Judge not, least ye be judged." It is so easy for people in one culture to be biased by their own values and social practices when assessing other cultures.

As powerful as this argument appears to be, some cultures have practices that many other societies consider to be quite harsh, if not brutal. Some societies virtually enslave women and restrain their movement as if they were in jails. Some societies mutilate children's genitals so badly during their rites of passage to adulthood that some children become infected and a few die, whereas the survivors suffer pain and loss of sexual pleasure when they engage in sexual intercourse after they recuperate from the ritual. According to extreme cultural relativism, there is no way to criticize these practices.

Debates about cultural relativism have resonated in the halls of academe for many decades. Again, one's conclusions about this topic depend on the methodology one uses to analyze it. People who take phenomenological perspectives—emphasizing introspection, constructionism, or transcendentalism—are reluctant to claim that any one individual or society can condemn the sexual and gender practices of any culture. Some scholars see this topic as a politically charged issue, because they fear that dominant and wealthy nations in the developed world may force their values and practices on smaller and more vulnerable cultures. Westernization of the whole world could obliterate many ancient cultural practices and reduce the cultural diversity of the planet.

The scientific method provides an alternate perspective. Anthropologists, physicians, epidemiologists, and other scientists can collect data on different cultural practices and evaluate the data according to various criteria. Does genital mutilation increase the risk of infection, sexual dysfunction, or death? Are the effects trivially small or quite noticeable? Next, these data can be made available to anyone who is interested in these issues.

There is no question that debates about cultural relativism have been useful in training people to think twice about imposing their personal values on people of different cultures. It has taught us to question individuals who criticize or disparage other societies merely because they differ from their own society. Fortunately, the scientific method allows us to go further than this. It provides numerous techniques for evaluating hypotheses about all the social practices of all cultures. Societies with painful rites of passage that mutilate the genitals do have higher infection and death rates, and some people are scarred for life, with reduced ability to enjoy sexual relations. If people want to minimize the risk of infection, death, and sexual dysfunction, they have data they can use to argue against those practices. Many cultures have practices that produce more pain than necessary, practices that increase the risks of AIDS infections, unwanted pregnancies, or sexual abuse. The scientific method can help identify those cultural practices that are best at enhancing sexual pleasure while reducing the risk of pain, STDs, and other aversive outcomes. Science can also evaluate gender roles and identify which cultural practices allow both women and men to have lives that are pleasurable and fulfilling, while minimizing suffering and pain. There are several scientific disciplines that collect data on these topics, and the Internet is making such studies increasingly available to people all around the world. We shall see what they do with the information.

See Also the Following Articles

AGGRESSION • BEHAVIORAL NEUROGENETICS • CLASSICAL CONDITIONING • EMOTION • REINFORCEMENT, REWARD, AND PUNISHMENT • SEX DIFFERENCES IN THE HUMAN BRAIN • SEXUAL DIFFERENTIATION, HORMONES AND • SEXUAL DYSFUNCTION • SEXUAL FUNCTION

Suggested Reading

Baldwin, J. D., and Baldwin, J. I. (1998). Gender differences in sexual interest. *Arch. Sex. Behav.* **26,** 181–210.

Crooks, R., and Baur, K. (2002). *Our Sexuality*, 8th ed. Wadsworth, Pacific Grove, CA.

Domjan, M., and Holloway, K. S. (1998). Sexual learning. In *Comparative Psyhology: A Handbook* (G. Greenberg and M. M. Haraway, Eds.), pp. 602–613. Garland Publishing, New York.

Lalumiére, M. L., and Quinsey, V. L. (1998). Pavlovian conditioning of sexual interests in human males. *Arch. Sex. Behav.* **27,** 241–252.

Rescorla, R. A. (1991). Associative relations in instrumental learning: The Eighteenth Bartlett Memorial Lecture. *Quart. J. Exp. Psychol.* **43B,** 1–23.

Rescorla, R. A. (1999). Summation and overexpectation with qualitatively different outcomes. *Anim. Learn. Behav.* **27,** 50–62.

Rosen, R. C., and Beck, J. G. (1988). *Patterns of Sexual Arousal.* Guilford Press, New York.

Stoléru, S., Grégoire, M.-C., Gérard, D., Decety, J., Lafarge, E., Cinotti, L., *et al.* (1999). Neuroanatomical correlates of visually evoked sexual arousal in human males. *Arch. Sex. Behav.* **28,** 1–21.

Walker, P. L., and Cook, D. C. (1998). Brief communication, Gender and sex: Vive la difference. *Am. J. Phys. Anthropol.* **106,** 255–259.

Sexual Differentiation, Hormones and

SUSAN L. ZUP and NANCY G. FORGER
University of Massachusetts

GLOSSARY

androgen insensitivity syndrome (AI) A condition caused by a genetic defect in the androgen receptor, creating individuals who produce androgens but are unable to respond to them.

congenital adrenal hyperplasia (CAH) A condition in which a genetic defect in one of several adrenal enzymes causes insufficient production of glucocorticoid hormones and excessive production of androgens by the adrenal gland, beginning *in utero*.

gonads The primary sex organs; ovaries (female) and testes (male).

hypothalamus A small structure at the base of the brain that sits just above the pituitary gland; it consists of many different clusters of cells that control functions such as sexual behavior, the secretion of hormones, and the regulation of drinking and feeding.

magnetic resonance imaging (MRI) A noninvasive procedure using radio waves and a strong magnetic field for two- and three-dimensional imaging of living tissue, including the brain.

motoneuron A specialized nerve cell in the spinal cord or brain stem that directly connects to and controls muscle.

nucleus As used in this article, an identifiable cluster of neural cell bodies in the central nervous system.

organizational hypothesis The concept that testosterone permanently "organizes" the brain during fetal or neonatal development to create a male-typical pattern of brain structures.

sexual dimorphism A sex difference in form or structure.

Sexual differentiation is the process by which the two sexes become different. For all mammals, including humans, sexual differentiation of the body is driven primarily by hormones produced by the gonads. This article will review the process of sexual differentiation and will present evidence that hormones cause sex differences in the human brain.

I. SEXUAL DIFFERENTIATION: AN OVERVIEW

Almost everyone agrees that, on average, the behavior of men and women differs. By contrast, there is considerable controversy about *how* behavioral sex differences come about. The evidence is quite strong for nonhuman animals that sex differences in behavior are attributable to specific structural or functional sex differences in the brain. Logically, this must be true for humans as well: because all behaviors are controlled by the nervous system, stable, predictable sex differences in human behavior must be due to sex differences in the structure or function of some neural circuit(s). More problematic are determining the extent to which sex differences in the human nervous system are innate and sorting out the relative contributions of genes, hormones, and socialization to the development of male and female brains.

Among laboratory animals, a vast literature indicates that the brain becomes sexually differentiated, in large part, as a result of exposure to gonadal steroid hormones. The same hormones that determine the appearance of the genitalia and the structure of the urogenital tracts during embryonic development concomitantly influence the developing brain. For this reason, an overview of the process of sexual differentiation of the body will be presented first, followed by a discussion of neural sex differences and their origins in animals and humans.

A. Sexual Differentiation of the Body

The fundamental mechanism of sexual differentiation of somatic tissues (''the body'') is essentially identical for all eutherian mammals, including humans. The genetic sex of an individual is determined at the moment of fertilization. The unfertilized ovum, or egg, contains a single sex chromosome, X, inherited from the mother. If a sperm carrying a second X chromosome fertilizes the ovum, then the embryo is a genetic female (XX); a genetic male (XY) will result if a sperm carrying a Y chromosome fertilizes the egg. However, as far as anyone can tell, early development proceeds identically for XX and XY embryos. In humans, the sexes cannot be distinguished (except by a chromosome test) until the eighth week of gestation. Prior to that time, both sexes possess a pair of so-called undifferentiated or ''indifferent'' gonads located in the abdominal cavity. The undifferentiated gonads are capable of becoming either ovaries or testes. If a normal Y chromosome is present (i.e., in the vast majority of genetic males), testes will develop; the indifferent gonads will become ovaries if the Y chromosome is absent (as in genetic females).

The gene on the Y chromosome responsible for testis development is known as SRY (*S*ex-determining *R*egion of the *Y* chromosome) and was identified by studying so-called ''XX-males.'' Occasionally, men (or mice) are found that appear to be genetically female (XX) but that nonetheless are phenotypically male (i.e., possess testes, a scrotum, and a penis). In almost every case, careful scrutiny reveals that a small piece of the Y chromosome is attached to one of the X chromosomes of such XX males. This piece of the Y presumably was inherited due to an abnormal chromosome crossover event during meiosis (a step in gamete development). By progressively pinpointing the piece of the Y chromosome common to all XX males, scientists eventually isolated the SRY gene. Of course, normal development of the testes and ovaries depends on the coordinated expression of many genes. SRY may be thought of simply as a switch that turns on a genetic cascade of other essential, secondary genes.

Although chromosomes and genes are responsible for the crucial first step of sexual differentiation (i.e., differentiation of the gonads), subsequent steps are driven *not* primarily by sex differences in genes but by differential exposure to *hormones* in males and females. Soon after the embryonic testes differentiate, they begin producing two hormones: *anti-Müllerian hormone* (also known as Müllerian inhibitory substance or Müllerian regression factor) and *testosterone*. Together, these hormones determine the subsequent differentiation of the internal reproductive tract and the external genitalia. As will be discussed later, the androgenic hormone, testosterone, also has wide-ranging effects on the developing nervous system.

Prior to sexual differentiation, all human embryos possess two sets of ducts, which are the precursors of both the male and female urogenital systems (Fig. 1). The Wolffian duct has the capacity to develop into the male urogenital tract, and the Müllerian duct is the anlagen of female structures. Anti-Müllerian hormone, secreted by the embryonic testes, causes the Müllerian ducts to regress in males. Testosterone, the other major testicular secretion, causes the Wolffian ducts to develop into the epididymis, vas deferens, and seminal vesicles. The absence of anti-Müllerian hormone in females allows the Müllerian ducts to persist and give rise to the fallopian tubes, uterus, and upper vagina (Fig. 1). The Wolffian ducts degenerate in females as a result of the simple absence of testosterone. Note that, in this scenario, female differentiation occurs ''by default'' in the absence of specific hormones, and it is the presence of hormones (anti-Müllerian hormone and testosterone) that drives the differentiation of males.

In addition to its effects on the Wolffian ducts, testosterone is also responsible for differentiation of the external genitalia. The primordia of the external genitalia are indistinguishable in early male and female embryos, consisting of a genital tubercle, genital folds, and genital swellings surrounding a single urogenital opening (Fig. 2). In the absence of hormonal stimulation, these structures form the clitoris and labia of females, whereas testosterone directs these same primordia to form the penis and scrotum in males. In this case, testosterone must first be converted at the genital skin to the more potent androgen, dihydrotestosterone, for complete virilization of the external genitalia. In humans, these events are completed by the end of the first trimester of pregnancy (approximately 12 weeks gestation).

B. Mechanisms of Hormone Action

How do the testicular hormones exert their profound effects on the developing body? In general, in order for a hormone to elicit a response in a given cell, that cell must have corresponding hormone ''receptors.'' This is true regardless of whether the cell in question is

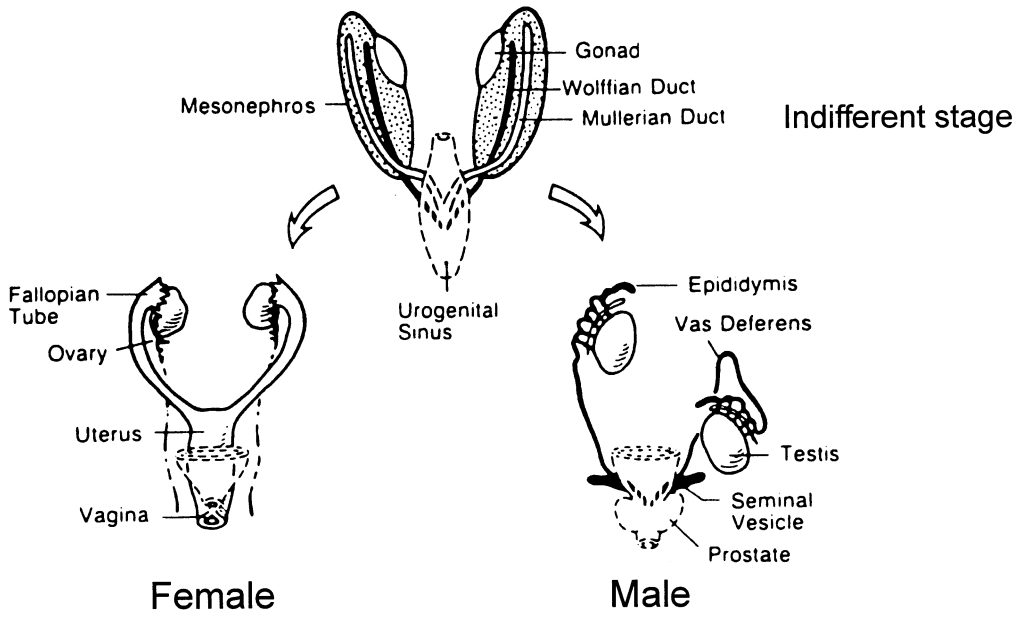


Figure 1 Differentiation of the internal genitalia in humans. Adapted, with permission from George, F. W., and Wilson, J. D. (1994). Sex Determination and Differentiation. In *The Physiology of Reproduction*, 2nd ed., pp. 3–28, (E. Knobil and J. D. Neill, Eds.). Raven Press, New York.

located in the genital skin, the embryonic urogenital ducts, or the brain. Hormone receptors are specialized proteins that bind to and interact with hormones to ultimately cause changes in target cells. Such receptors may be embedded in the plasma membrane on the surface of cells or may be found intracellularly, within the cytoplasm or cell nucleus. Anti-Müllerian hormone is itself a protein that exerts its effects by binding

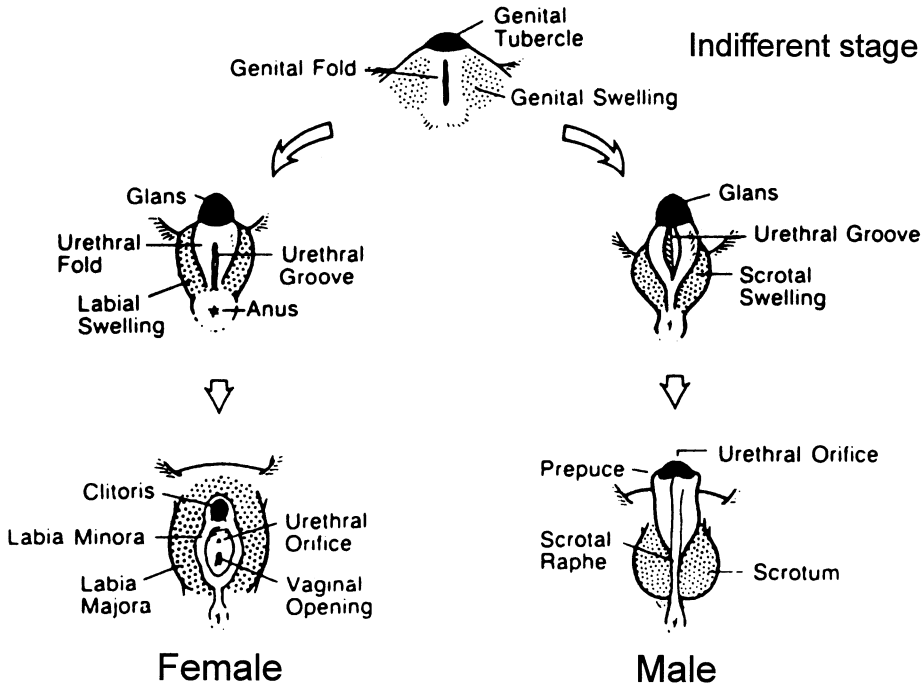


Figure 2 Differentiation of the external genitalia in humans. Adapted, with permission, from George, F. W., and Wilson, J. D. (1994). Sex Determination and Differentiation. In *The Physiology of Reproduction*, 2nd ed., pp. 3–28, (E. Knobil and J. D. Neill, Eds.). Raven Press, New York.

to cell surface membrane receptors. By contrast, most of the actions of the sex steroids are thought to be mediated by intracellular receptors.

Steroid hormones, such as testosterone, are lipophilic and can passively diffuse through the lipid-rich plasma membrane of cells. Once inside a target cell, the steroid may directly bind its receptor, or it may first be converted to another hormone metabolite. For example, after it enters a target cell, testosterone can be converted into dihydrotestosterone by the enzyme 5 α-reductase, or it can be metabolized to estradiol by the aromatase enzyme. The fate of testosterone will depend upon which enzymes are present in a given target cell. The "active" steroid metabolite then binds the corresponding receptor protein (an androgen receptor for testosterone or dihydrotestosterone, an estrogen receptor for estradiol) found in the cytoplasm or cell nucleus. These intracellular steroid hormone receptors are ligand-dependent transcription factors, which means that upon binding hormone they interact with specific DNA sequences to directly alter gene transcription (Fig. 3).[1]

If receptors to a given hormone are defective or absent, then cells cannot respond to the hormone. For example, males that, due to a genetic defect, lack receptors to anti-Müllerian hormone develop as pseudo-hermaphrodites. These individuals possess testes and fully differentiated Wolffian duct derivatives, but also a uterus and fallopian tubes. Similarly, XY individuals that lack functional androgen receptors are said to be "androgen-insensitive." The testes develop in these genetic males (due to the presence of the SRY gene on the Y chromosome) and produce plenty of testosterone. However, in the absence of androgen receptors, testosterone elicits no response: the Wolffian ducts fail to develop and the external genitalia are completely feminized in androgen-insensitive males. Such individuals are usually raised as girls and may not be recognized as androgen-insensitive genetic males until puberty, when they fail to menstruate. The foregoing examples underscore the principle that it is hormones, not sex chromosomes, that primarily drive somatic sexual differentiation. An individual's phenotype follows closely to his or her "hormonal sex," regardless of genetic sex.

[1]Although the "classical" mechanism of steroid hormone action involves binding to intracellular receptors and direct regulation of gene transcription, some effects of steroids occur too rapidly to involve changes in gene expression. These nongenomic actions of steroid hormones apparently involve hormone receptors located on the membranes of cells. At present, the extent to which nongenomic actions of steroid hormones contribute to sexual differentiation is not known.

C. The Brain as a Site of Steroid Hormone Action

Although early endocrinologists initially focused on the mechanisms of hormone action in peripheral tissues, it became clear in the 1960s and 1970s that *neurons* also express receptors for gonadal steroid hormones. Androgen and estrogen receptors are widely distributed throughout the brains of mammals. Whereas the highest concentrations of steroid receptors are found within the hypothalamus and in limbic regions providing input to the hypothalamus, brain regions associated with higher cognitive functioning (e.g., the cerebral cortex and hippocampus) also bind androgens and estrogens. This basic pattern of steroid hormone receptor distribution in the brain is present in both sexes and is similar across a wide range of species, from rats, to guinea pigs, to monkeys, to humans. Perhaps not surprisingly, many of the sex differences described so far in vertebrate brains are found in regions, such as the hypothalamus, that are rich in androgen and/or estrogen receptors. Importantly, neural receptors for gonadal steroids also are expressed in early development. Thus, the brain must be considered a target site for direct steroid hormone action during perinatal life and in adulthood.

II. SEXUAL DIFFERENTIATION OF THE NERVOUS SYSTEM: PRINCIPLES DERIVED FROM ANIMAL WORK

Although the role of hormones in sexual differentiation of the body was understood by the 1940s, the brain was initially considered immune to hormone action. In the late 1950s and 1960s, however, work on sexual behavior in rodents first led to the prediction that sexual differentiation of the brain, like other parts of the body, was controlled by gonadal steroid hormones. For example, an adult female guinea pig will normally exhibit female-typical behaviors during mating, including a stereotypical receptive posture called lordosis. However, a female guinea pig given testosterone early in development will exhibit lower levels of lordosis in adulthood and will exhibit higher levels of male-typical sexual behaviors (e.g., mounting other animals). These observations led Phoenix, Goy, Gerrall, and Young to propose the "organizational

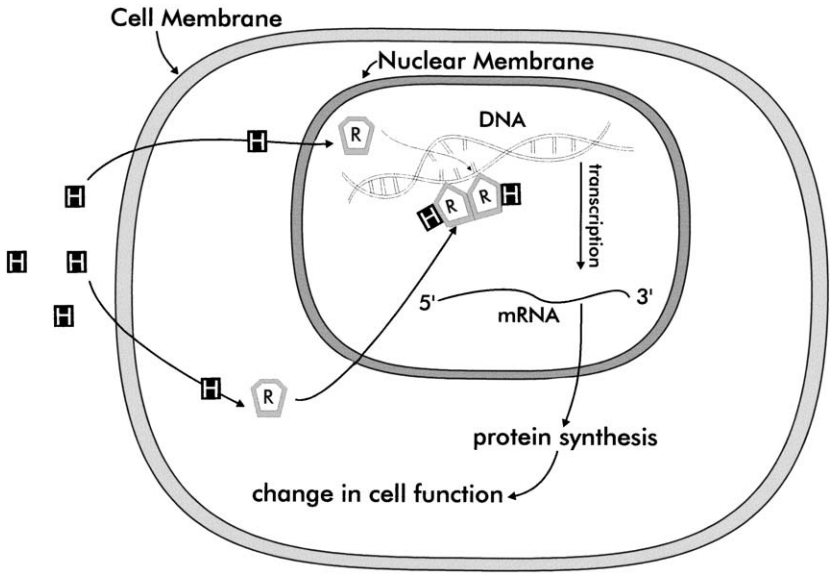


Figure 3 Simplified model of the classical mechanism of steroid hormone action. The steroid hormone (H) diffuses across the cell membrane to bind a hormone receptor protein (R) located in the cytoplasm or nucleus. Hormone–receptor complexes then bind to specific sequences on DNA to regulate gene transcription.

hypothesis:'' testosterone, present in the developing male, permanently ''organizes'' the brain during fetal or neonatal development to create a male-typical pattern of brain structures. In adulthood, gonadal steroids then ''activate'' the previously organized neural substrates, resulting in male-typical behaviors. Females are not normally exposed to high levels of testosterone and, thus, develop female-typical behaviors. The theory of hormone-directed neural differentiation was strengthened when, as previously discussed, receptors to gonadal steroid hormones were shown to exist in the brain. However, at the time the organizational hypothesis was first proposed and for a number of years afterward, no neural sex differences had actually been discovered.

One of the first examples of a sex difference in the brain came from a 1971 study by Raisman and Field, who used electron microscopy to characterize more than 25,000 synapses in the preoptic area of the rat brain. They found a sex difference in neural connectivity, such that male rats had more of a particular type of synapse than did female rats. If all sex differences had been this elusive, the field of neural sex differences might have died in its infancy. In 1976, however, a sex difference in songbird brains was discovered that was so large, it could be seen in brain sections with the naked eye. The higher vocal center (HVC) controls singing in songbirds such as zebra finches. Nottebohm and colleagues demonstrated that the HVC of male zebra finches is 5–6 times larger in volume than that of females. Because male zebra finches sing and females do not, this was an exciting finding lending credence to the idea that sex differences in the brain cause sex differences in behavior. Since then, an astonishing (and ever growing) number of sexual dimorphisms in the nervous system have been discovered in many species. Here we present just two of the best characterized sex differences in the rodent nervous system, for which possible human homologs have been identified.

A. The Spinal Nucleus of the Bulbocavernosus

The spinal nucleus of the bulbocavernosus (SNB) is a sexually dimorphic group of motoneurons in the lumbar spinal cord of rodents (Fig. 4). These motoneurons innervate muscles (including the bulbocavernosus muscle) surrounding the base of the penis, which help control erection and ejaculation. Although the muscles and motoneurons are present in both sexes during fetal development, the bulbocavernosus muscle and SNB motoneurons degenerate in females around the time of birth. Adult female rats completely lack the bulbocavernosus muscle and have only one-fourth as many SNB motoneurons as do adult males. Sexual differentiation of the SNB conforms nicely to the organizational hypothesis. Female rats given testosterone around the time of birth are permanently ''masculinized,'' in that they retain the bulbocavernosus musculature and many more motoneurons in the

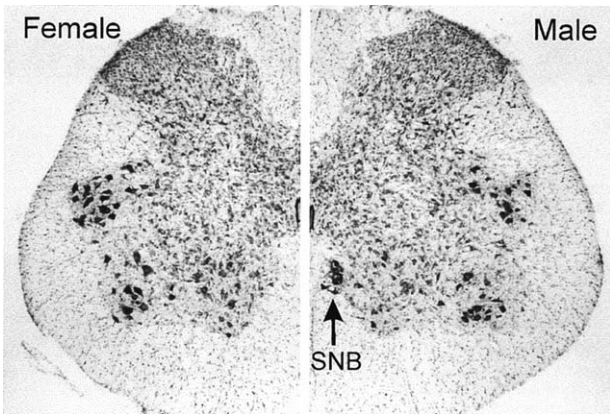


Figure 4 Photomicrographs depicting the spinal nucleus of the bulbocavernosus (SNB) in adult female (left) and male (right) rats. Androgens produced by the perinatal testes rescue SNB motoneurons in males, whereas most SNB cells degenerate in females.

SNB than do control females. Likewise, blockage of the actions of testosterone in a perinatal male rat causes the bulbocavernosus muscle to atrophy and most SNB motoneurons to die. For example, male rats with a genetic defect of the androgen receptor have no more SNB motoneurons than do normal females; this finding indicates that androgens direct the sexual differentiation of the SNB. Although first discovered in rats, homologs to the SNB have since been found in many different species, including mice, gerbils, hyenas, dogs, monkeys, and humans.

B. The Sexually Dimorphic Nucleus of the Preoptic Area

The sexually dimorphic nucleus of the preoptic area (SDN-POA) is the best characterized sex difference in the mammalian brain, and homologs have been found in species ranging from rats and gerbils to hyenas, macaques, and humans. SDN-POA volume and neuronal number are 5–7 times larger in adult male rats than in females. Similar to the SNB, SDN-POA volume is permanently increased in female rats treated with testosterone during a critical neonatal period, and castration of male rats on the day of birth permanently decreases the size of the SDN-POA. In this case, however, testosterone itself is not the active hormonal metabolite; rather, it is the conversion of testosterone to estradiol (a process known as aromatization) within individual neurons that masculinizes SDN-POA volume and neuron number. Although the aromatization of testosterone to an estrogen is crucial for a number of neural sex differences in rodents, the role of estrogens in the sexual differentiation of primate brains is less clear.

The SDN-POA is located in a region of the brain (the preoptic area of the hypothalamus) that controls male sex behavior. Large lesions of the preoptic area decrease male-typical sex behavior in snakes, lizards, birds, rodents, ungulates, and primates. Lesions confined to sexually dimorphic regions of the preoptic area disrupt scent marking and sexual behavior in male gerbils and alter sexual partner preference in male ferrets. However, the function of the rat SDN-POA has not been clearly demonstrated. Lesions limited to the SDN-POA of rats result in only subtle decrements in male sex behavior. The lack of a clear behavioral correlate for even so striking a sex difference as that in the SDN-POA illustrates the difficulty in demonstrating a cause and effect relationship between brain morphology and behavioral differences.

C. How Do Hormones Cause Neural Sex Differences?

Many sex differences in the nervous system (including the SNB and SDN-POA) involve a difference in neuron number. Sex differences in neuronal number could result from sex differences in neurogenesis (the "birth" of neurons), neuronal migration, neuronal differentiation, or neuronal cell death. At present, the evidence is strong that differential rates of cell death between the sexes contribute to neural sex differences. During nervous system development, many more neurons are born than are ultimately needed. Through a decision-making process not yet fully understood, the "extra" neurons die off while other neurons live. Gonadal sex steroids can influence this process in hormone-sensitive brain regions so that more neurons die in one sex than in the other, resulting in regional sex differences in neuron number.

In the SNB and SDN-POA, testosterone, or its hormonal metabolites, *decreases* cell death, thereby creating a sex difference in neuron number "favoring" males. This is not always the case, however. For example, in the anteroventral periventricular nucleus (AVPV), a cell group in the hypothalamus crucial for the control of ovulation, female rats and mice have more neurons than do males. This sex difference in neuronal number apparently is also due to hormonal control of cell death, although here testosterone *increases* cell death. Therefore, the same hormone seems to decrease cell death in some neural regions (e.g., SNB and SDN-POA) while increasing it in others

(e.g., AVPV). Testosterone may also cause sex differences by influencing neurogenesis, neuronal migration, and/or neuronal differentiation, although this has not yet been demonstrated conclusively.

D. Sex Differences Due to "Activational" Effects of Hormones

The organizing effects of hormones engender many structural sex differences in the developing nervous system. However, sexually dimorphic *behavior* often requires the presence of hormones in adulthood to "activate" previously organized brain circuits. For example, a normal adult female rat requires stimulation by estrogens and progestins to exhibit sexual receptivity; if she lacks these hormones in adulthood (e.g., if the ovaries are removed), then sexual behavior is abolished. However, even large amounts of estrogens and progestins may not cause an adult female rat to show normal levels of sexual receptivity, if she had previously been treated with testosterone during perinatal development.

This distinction between organizational and activational effects of hormones is a useful heuristic device, but it may be an oversimplification. Sex steroid hormones secreted in adulthood not only activate existing neural circuits but can also change neural morphology. For example, the volume of the medial amygdala is normally larger in adult male rats than in females. Breedlove and colleagues have shown that this sex difference in the medial amygdala can be eliminated by equalizing the circulating testosterone levels in male and female rats *in adulthood* (either by castrating males or by giving females exogenous testosterone). In addition, the same gonadal hormones that influence cell death during development may also influence that process during adulthood. Adult birds and mammals continue to make new neurons in specific parts of the brain throughout life, and evidence shows that gonadal hormones can modulate the life span of newly born cells. Thus, structural sex differences in the nervous system may be due to differential exposure to sex steroid hormones in adulthood as well as in development.

III. SEX DIFFERENCES IN THE HUMAN BRAIN

The prediction that sex differences *must* exist in the human brain comes from at least three different types of observations: men and women (or boys and girls) differ predictably in (1) behavioral traits (e.g., aggression, juvenile play); (2) average cognitive abilities; and (3) the incidence of neurological and psychiatric disorders. Regardless of whether these differences are due to genes, hormones, socialization, or the confluence of all three, sex differences in behavior, neuropathology, and cognitive performance must be reflected by sex differences present someplace in the nervous system. As described next, anatomical and functional differences have, in fact, been identified in the human brain, although thus far it has proven difficult to definitively link the neuroanatomical findings with specific behavioral sex differences.

A. Anatomical Sex Differences in the Human Brain

In comparison to the rich animal literature on neural sex differences, reports of sexual dimorphisms in the human brain are relatively sparse. This may suggest that sex differences in the human nervous system are not as numerous or as pronounced as they are in laboratory animals. However, animal studies routinely compare the brains of large groups of animals precisely matched for age, genetic background, and rearing conditions. Similar control over human subjects obviously is not feasible, and experimenters working with humans often must content themselves with relatively small, heterogeneous samples. The difficulty of conducting well-designed studies of the human brain almost certainly contributes to the relative lack of information about sex differences in our own species. Despite the limitations, a number of anatomical sex differences have been reported in the human nervous system (Table I). These findings have predominantly come from the examination of post mortem material, although investigators are increasingly making use of magnetic resonance imaging (MRI) to study structural sex differences in the living human brain. A caveat that must be kept in mind when examining Table I is that several of the listed findings are not firmly established. For a number of the sex differences listed, the original report has not been followed up, so that the observation is neither refuted nor confirmed. In other cases, multiple studies of a single brain area have yielded conflicting conclusions (see later discussion).

1. Overall Brain Size and Neuron Number

Perhaps the most obvious, and best established, neural sex difference in the human brain is that of overall size.

Table I
Anatomical Sex Differences Reported in the Human Nervous System

Neural structure	Nature of the difference	Reference[a]
Whole brain	Larger in men	Many reports spanning over 100 years.
Neocortex	More neurons in men	Pakkenberg & Gundersen (1997). *J. Comp. Neurol.* **384,** 312.
Language-related areas		
Planum temporale	Size of left and right more symmetrical in women; neuronal density greater in women	Wada *et al.* (1975). *Arch. Neurol.* **32,** 239; Witelson *et al.* (1995). *J. Neurosci.* **15,** 3418.
Dorsolateral prefrontal cortex and superior temporal gyrus	Proportionally larger in women	Schlaepfer *et al.* (1995). *Psychiartry Res.: Neuroimaging* **61,** 129.
Structures connecting the left and right brain		
Corpus callosum	Posterior portion is larger and/or more bulbous in women	De Lacoste-Utamsing & Holloway (1982). *Science* **216,** 1413.
Anterior commissure	Larger, and more fibers, in women	Allen & Gorski (1991). *J. Comp. Neurol.* **312,** 97; Highley *et al.* (1999). *Biol. Psych.* **45,** 1120.
Massa intermedia of thalamus	More often present, and larger, in women	Morel (1948). *Acta Anat.* **4,** 203; Allen & Gorski (1991). *J. Comp. Neurol.* **312,** 97.
Regions of the hypothalamus		
SDN-POA[b]	Larger in men	Swaab & Fliers (1985). *Science* **228,** 1112.
INAH-3[c]	Larger in men	Allen *et al.* (1989). *J. Neurosci.* **9,** 497.
BNST[d]	Subarea is larger in men	Allen & Gorski (1990). *J. Comp. Neurol.* **302**, 697.
SCN[e]	Sex difference in shape	Swaab *et al.* (1985). *Brain Res.* **342,** 37.
Spinal cord		
Onuf's nucleus	More motoneurons in men	Forger & Breedlove (1986). *Proc. Natl. Acad. Sci. USA* **83,** 7527.

[a]Where multiple references exist, the earliest reports are cited.
[b]SDN-POA: sexually dimorphic nucleus of the peoptic area.
[c]INAH-3: third interstitial nucleus of the anterior hypothalamus.
[d]BNST: bed nucleus of the stria terminalis.
[e]SCN: suprachiasmatic nucleus.

In studies spanning many decades, from laboratories all over the world, brain size is found to be about 10% greater in men than in women. Although this sex difference is minimal when brain size is mathematically "corrected" for body size, the rationale for such a correction is not immediately obvious. Body weight and body height each account for only a tiny proportion of the variance in brain size *within* each sex. Moreover, by dismissing the sex difference in brain size on the basis of a correction for body size, we risk overlooking some interesting issues: Is the relationship between body and brain size *causal*, and, if so, how do large bodies "cause" large brains? If the relationship is not causal, is there some common factor (e.g., gene or hormone) independently modulating both brain and body size?

Only very recently have scientists addressed the question of whether the larger brains of males actually contain *more cells*. For at least some neural regions, the answer appears to be "yes." Two different research groups have concluded that men have about 15% more neurons in the neocortex than do women, although there is disagreement as to whether there is a difference in neuronal *density* between the sexes. The consequences of a larger brain and more neocortical neurons are not immediately obvious, however, because general cognitive abilities such as overall intelligence and memory do not appear to differ significantly between

men and women. In addition, there is some evidence that the male advantage in cortical neuron number may be offset by an increase in neuronal processes (and, by extension, synapses) in females. The overall greater number of neurons in the neocortex of men also may belie some interesting regional sex differences in the opposite direction. For example, the volumes of several language-related cortical areas are proportionally larger in women than in men (Table I). Specifically, the planum temporale is about equal in absolute volume in men and women, despite the overall larger brains of men. This, combined with the report by Witelson and colleagues that neuronal density in the planum temporale is greater in women than in men, leads to the prediction of a greater total number of neurons in the planum temporale of females. Thus, the seemingly simple observations of sex differences in overall brain size and neocortical neuron number illustrate the difficulties encountered in attempting to ascribe a functional role to gross differences in neuroanatomy.

2. Structures Connecting the Left and Right Brain

A number of structures connecting the left and right sides of the brain have been reported to be larger in women than in men (Table I). These observations have generated considerable interest, because several neural functions, including speech and visuospatial function, appear to be less lateralized in female than in male brains. The corpus callosum is the major fiber tract connecting the left and right cerebral hemispheres, and a report of almost 20 years ago indicated that the posterior portion of the corpus callosum (the ''splenium'') is larger and more bulbous in women than in men. However, not all subsequent studies have replicated the finding, and despite a large number of reports examining the corpus callosum in men and women, no clear consensus regarding sex differences has emerged. A literature search indicates that the controversy continues to the present day, with about equal numbers of papers published in the last 5 years confirming and refuting the existence of a larger or more bulbous splenium in the corpus callosum of women. Because many studies on both sides of the argument appear to be well-designed and well-executed, the reason for the conflicting outcomes is not clear, although the great variability in size and shape of the corpus callosum *within each sex* no doubt contributes to the confusion. One must conclude that, if a sex difference in the splenium exists, it must be quite subtle.

The massa intermedia of the thalamus and the anterior commissure are also reported to be larger in women than in men. The massa intermedia is composed of neurons, neuropil, and loosely organized axons that connect the left and right thalami. This structure is completely absent from about 30% of all humans, and reports dating back to the 1940s indicate that women are significantly more likely than men to retain a massa intermedia. More recently, Allen and Gorski found that, when present, the massa intermedia is about 50% larger in women than in men. The anterior commissure is a fiber tract that primarily connects the left and right temporal lobes. This structure is reported by two independent groups of investigators to be significantly larger in cross-sectional area and to contain a greater number of fibers in women than in men.

3. The Hypothalamus

As discussed earlier, the hypothalamus is the site of many neural sex differences in laboratory animals, and several differences have also been reported in the hypothalami of men and women. Among these, the finding that appears most solid, in that it has been replicated by independent groups of scientists, is a sex difference in a small cell group known as INAH-3 (third interstitial nucleus of the anterior hypothalamus). INAH-3 contains about twice as many neurons in men as in women and is located within a region of the brain implicated in the control of male sexual behavior in virtually all vertebrates. On the basis of cytoarchitectonic criteria, Byne has argued that INAH-3 may be the primate counterpart of the SDN-POA of rats. However, nothing is known about the development of INAH-3, its function, or the role of gonadal hormones in sexual differentiation of this area. INAH-1 (also known as ''the SDN-POA of humans'') is another cell group in the anterior hypothalamus that has been proposed as the human homolog of the rat SDN-POA. Although it shares several neurochemical markers with the rat SDN-POA, not all studies find a sex difference in the overall volume of human INAH-1.

4. The Spinal Cord

There is a clear human homolog to the spinal nucleus of the bulbocavernosus (SNB) of rats, described in Section II. Motoneurons innervating the bulbocaver-

nosus muscle of humans are located in a cell group in the sacral spinal cord known as Onuf's nucleus, and men have more neurons in Onuf's nucleus than do women. The sex difference in human Onuf's nucleus (approximately 30% more neurons in men) is much less pronounced than the 3- to 4-fold sex difference in motoneuron number of the rat SNB. Should this relatively subtle sex difference in human Onuf's nucleus be taken as an indication that sex differences in the nervous systems of humans are less pronounced than they are in nonhuman animals? Probably not. Unlike female rats, women *do* have a bulbocavernosus muscle, although it is smaller and modified in form compared to that in males. The modest sex difference in motoneuron number of human Onuf's nucleus probably reflects the retention of the perineal musculature in women. In support of this conclusion, a bulbocavernosus muscle is also retained in female dogs, hyenas, and monkeys, and in each of these species males have about 20–50% more motoneurons in Onuf's nucleus than do females.

B. Sex Differences in Connectivity and Neurochemistry

In addition to the sex differences in gross anatomy listed in Table I, many sex differences are likely to exist in more subtle features of the brain, such as neural connectivity or neurochemistry. From animal work, it is clear that the size of projections from one neural region to another can vary markedly by sex and that many of these sex differences can be eliminated by manipulating gonadal hormones early in development. Sex differences in neurotransmitter levels within particular neural circuits also contribute importantly to behavioral differences in male and female laboratory animals and are modified by both organizational and activational effects of steroid hormones. Such sex differences would not be visible from routine inspection of post mortem brain tissue, which formed the basis of virtually all studies on the human brain until quite recently. Only in the last several years have investigators begun to tackle the question of ''functional'' sex differences in human neural tissue by adapting neurochemical staining techniques for post mortem human brains or by taking advantage of the development of techniques allowing for the functional imaging of the living human brain.

For example, positron emission technology (PET) can be used to examine the accumulation and distribution within the brain of radioactively labeled tracers injected into human subjects. PET studies report that the rate of synthesis of the neurotransmitter, serotonin, is significantly higher in men than in women. Men apparently also have an increased number of type-2 serotonin-binding sites in several brain regions. Because perturbations in the serotonin system are implicated in several psychiatric disorders, including unipolar depression, findings such as these may help to explain sex differences in the incidence of psychopathology (see Section III.D).

Functional MRI (fMRI) is another technique that has contributed significantly to the study of human sex differences. Based on the detection of regional changes in blood oxygenation induced by neural activity, fMRI is a noninvasive method for identifying brain regions involved in specific tasks. By using fMRI, Shaywitz and colleagues found a marked sex difference in the pattern of brain activation in men and women performing a phonological task (specifically, deciding whether two strings of nonsense words rhyme). Men predominantly used a language area in the left cerebral hemisphere, whereas more diffuse activation of both the left and right sides of the brain was observed in women performing the same task. A similar sex difference in lateralization of neuronal activity was seen in a PET study of men and women performing a grammatical task. Because the overall performance (speed and accuracy) of men and women on both tasks was equivalent, these studies raise an interesting point: neural sex differences may exist even when the behavior or performance of men and women is similar. MRI and fMRI hold tremendous promise for the study of human neural sex differences, because (1) they permit noninvasive, *in vivo* examination of carefully screened, well-matched subjects; (2) both anatomical structure and functional activation can be compared in the same subjects; and (3) it is possible to repeatedly scan the same individual. This last feature allows for longitudinal studies of the same groups of subjects, as well as before-and-after experimental designs.

C. Sex Differences in Behavior and Performance

Most of the evidence for sex differences in the human brain is indirect and comes from observations of sex differences in human behavior. Countless claims of behavioral and/or cognitive differences between men and women can be found scattered throughout the literature and in the popular press. However, the number of human behavioral sex differences that have been repeatedly confirmed in carefully designed studies

Table II
Effect Sizes for Sex Differences in Selected Human Behaviors

Behavior	Approximate effect size[a]
Aggression	Moderate
Childhood play (level of activity, degree of rough-and-tumble play, selection of toys and playmates)	Moderate to large
Tasks on which males typically outperform females:	
Visuospatial	
Three-dimensional visual rotation	Large
Two-dimensional visual rotation	Small
Spatial perception	Small to moderate
Spatial visualization	Negligible
Quantitative	
Overall ability	Small to moderate
Problem solving	Small to moderate
Tasks on which females typically outperform males:	
Verbal	
Overall ability	Negligible to small
Verbal–association fluency	Moderate
Speech production	Small
Perceptual Speed	Moderate

[a]Magnitude of effect size was defined according to J. Cohen (*Statistical Power Analysis for the Behavioral Sciences*, Academic Press, 1997). Effect sizes of approximately 0.8 standard deviation or greater were considered large, effect sizes of approximately 0.5 were considered moderate, effect sizes of approximately 0.2 were considered small, and effect sizes of less than 0.19 were considered negligible. Table adapted from Collaer and Hines (1995) with permission from the American Psychological Association. See Collaer, M. L., and Hines, M. (1995). Human behavioral sex differences: A role for gonadal hormones during early development? *Psychol. Bull.* **118,** 55–107, for supporting references.

is actually relatively small. Table II lists several of the most reliably observed sex differences in human behavior or performance. On average, males are more aggressive than females and exhibit more active, ''rough-and-tumble'' play in childhood. The sex difference in play is particularly interesting, in that it is found across cultures and can be observed in very young children. In the cognitive realm, males excel at visuospatial tasks, particularly those requiring three-dimensional mental rotation, whereas females typically excel at certain verbal skills, especially verbal fluency and perceptual speed. It is important to note that all of the sex differences described here are differences in the *average* behavior or performance of males and females. In all cases there is considerable overlap between the sexes, and the range of behavior within one sex usually is larger than the mean difference between the sexes. With large enough sample sizes, even tiny, arguably meaningless differences may be statistically significant. For this reason it is useful to categorize sex differences by ''effect size,'' as in Table II. Effect size is a measure of the size of the difference between two groups relative to the variability within groups. By one commonly accepted standard, when the mean difference between groups exceeds 0.8 times the standard deviation of scores *within* each group, the effect is said to be ''large.'' Note that, by this standard, the only sex differences listed in Table II that can be considered ''large'' are differences in three-dimensional mental rotation and childhood play. Arguably the greatest sex difference in human behavior is not listed in Table II. Most, but not all, men are sexually attracted to women, whereas most, but not all, women are sexually attracted to men. Although this gender difference is so large as to be often taken for granted, sexual partner preference must also have a neural basis.

Have any of the sex differences in behavior or performance been linked to observed sex differences in the brain? At this point, the answer is ''no;'' we cannot definitively point to any neural sex difference as the basis for any specific sex difference in human behavior. One difficulty in this endeavor is in moving from correlation to causation in human studies. For

example, LeVay has reported that the size of INAH-3 is larger in heterosexual men than in homosexual men, but it is not clear whether there is any causal relationship between INAH-3 size and sexual orientation or even how such a causal relationship could be tested. A second challenge stems from our lack of sophistication regarding the neural underpinnings of human behaviors. A case in point is the sex difference in juvenile play. Although some neural areas have been implicated in play behavior in animals (e.g., the amygdala in rats), in general we have only a weak grasp of the neurocircuitry underlying complex behaviors such as "play" in humans. Scientists cannot begin to explain sex differences in a given behavior until the neural basis for the behavior is understood.

Human language is one behavior that has received intense scrutiny and is relatively well-understood. In most people, neural regions involved in speech production and recognition are located predominantly in several interconnected cortical structures of the left frontal and temporal lobes. Several language-related areas are reported to be absolutely or proportionally larger in the brains of women than those in men (Table I). These anatomical sex differences favoring women may underlie the superior performance of women on several verbal measures. Note, however, that the relationship is purely correlational at this point. It will be very difficult to prove that larger language-related cortical areas "cause" enhanced verbal performance, but one step in that direction would be to determine whether there is a positive relationship between the size of specific language-related neural areas and verbal performance *within* each sex.

Table III
Neurological and Psychiatric Disorders That Exhibit a Sex Difference in Incidence[a]

Disorder	Ratio male:female
Primarily diagnosed in infancy or childhood	
Attention deficit–hyperactivity	80:20
Autism	80:20
Dyslexia	70:30
Gilles de la Tourette	70:30
Mental retardation	60:40
Stuttering	75:25
Primarily diagnosed in adulthood	
Alzheimer's disease	30:70[b]
Amyotrophic lateral sclerosis	62:38
Anorexia nervosa	7:93
Anxiety disorder	33:67
Bulimia	10:90
Depression	33:67
Gender identity disorder	75:25
Multiple sclerosis	33:67
Sleep apnea	90:10

[a]Data compiled from: *Diagnostic and Statistical Manual of Mental Disorders*, 4th ed., 1994, American Psychiatric Association, Washington, DC, and reviews of the neurological literature. Where original sources report a range of gender ratios, the approximate midpoint of the range is given.

[b]Estimates of gender-specific incidence for Alzheimer's disease vary widely. After accounting for sex differences in life span, risk is between 1.5 and 3.0 times higher in women than in men.

D. Sex Differences in Neurological Diseases and Psychiatric Disorders

The marked sex differences that exist in the occurrence of many neurological and psychiatric diseases provide some of the strongest evidence for sex differences in the human brain. It has been argued that *most* major neurological disorders exhibit a sex difference in prevalence, disease course, or symptomatology, and some of the differences are quite striking (Table III). For example, about 80% of those suffering from autism are male; an even higher percentage of all cases of sleep apnea occur in men or boys. Conversely, females make up the large majority of cases of anorexia nervosa, and essentially all anxiety disorders are more prevalent in women than in men (i.e., simple phobia, social phobia, agoraphobia, panic disorder, and posttraumatic stress syndrome). For some disorders, such as schizophrenia, overall incidence does not vary by gender, but the disease course does.

One would think that if we understood why one sex represents the majority of those suffering from a given neurological condition or why one sex presents a different disease course than the other, we would be much closer to understanding the etiology and possible treatment of the problem. Nonetheless, surprisingly little is known about the neuroanatomical or neurochemical bases for sex differences in neurological and psychiatric diseases. Schizophrenia and dyslexia are two disorders that have received some attention in this regard. Although the overall lifetime risk for schizophrenia is about equal in men and women, the average age of onset is 3–5 years earlier in men. Women suffering from schizophrenia also exhibit less severe

symptoms and respond better to neuroleptic medication than do men. These sex differences may be due to differences in the neuropathology characteristic of male and female schizophrenics. A number of structural brain abnormalities are associated with schizophrenia (e.g., increased volume of the cerebral ventricles, reduced volume of the hippocampal area, and alterations in the corpus callosum), and each of these is seen more frequently in male than in female schizophrenics.

Dyslexia is a disorder characterized by a specific retardation in reading and spelling skills despite otherwise normal intelligence and is more prevalent in boys than in girls. In fact, as evidenced in Table III (top) and commented on by many investigators, males are at greater risk than females for almost all childhood neurodevelopmental disorders. Autism, attention deficit disorder, stuttering, dyslexia, Tourette syndrome, mental retardation, cerebral palsy, and epilepsy are all significantly more prevalent in boys than in girls. The sex difference in dyslexia is particularly interesting in light of the growing body of research suggesting that areas of the brain responsible for language are organized differently in males and females. As discussed earlier, functional imaging studies indicate that brain activation in males and females differs during phonological processing. Gender differences also are found in the morphology of language-related brain areas, and there are sexually dimorphic consequences of brain injury to the retention of language skills. In light of all this, it might be surprising if there were *not* differences in the incidence, symptoms, or causes of dyslexia in boys and girls. Post mortem studies conducted by Galaburda and colleagues indicate that dyslexia is associated with clusters of incorrectly positioned neurons, known as ectopias. Although the number of subjects examined so far has been small, ectopias appear to be much more prevalent in the brains of male than those of female dyslexics. Because malpositioned cells are thought to be indicative of disrupted neuronal migration during fetal life, this work suggests that a prenatal event may be responsible for the increased susceptibility to dyslexia in males.

IV. ARE HUMAN NEURAL SEX DIFFERENCES DUE TO HORMONES?

Virtually all of the neural sex differences described in nonhuman animals are ascribed to the actions of gonadal steroid hormones. Do sex differences in the human brain similarly depend on sex steroids? Although it would be parsimonious to assume so, the evidence thus far remains largely indirect. For example, it currently is not known whether any of the neuroanatomical sex differences listed in Table I are due to differential exposure to gonadal steroids in males and females. Many of the entries of Table I consist of tiny cell groups too small to be detected in the living brain by even the most sophisticated neuroimaging technology available today. In order to test whether there is a hormonal basis for these sex differences, scientists would have to rely on post mortem brains of individuals exposed to sex-atypical hormone levels. Such individuals are rare, and it would take years to compile an adequate sample. On the other hand, some structural sex differences (e.g., the sex difference in overall brain size) can readily be imaged in the living human brain, and it should be feasible to obtain scans from clinical populations of interest. For example, one could address whether the sex difference in brain size is androgen-dependent by comparing the brain scans of androgen-insensitive individuals with those of normal males and females. The next few years may yield investigations of this kind.

If sex steroids do cause sex differences in the human brain, they are likely to do so during one of three developmental windows when sexually dimorphic "peaks" of testosterone are observed in males: from about weeks 10 to 20 of gestation, during the first few postnatal months, and after puberty (Fig. 5). Although the ovaries are also capable of steroidogenesis beginning in midfetal life, androgen production by the ovaries is low and all mammalian fetuses are presumably exposed to estrogens and progestins of *maternal* origin. The role of *fetal* ovarian secretions in female brain development is not clear. Postpubertally, estrogen (and progestin) levels fluctuate markedly across the menstrual cycle and are higher in females than in males (Fig. 6). Ovarian androgen secretion also varies with the menstrual cycle, but at all points in the cycle, average circulating androgen levels are lower in adult females than in adult males.

By analogy with the animal literature, the perinatal exposure to testosterone in human males might be expected to be "organizational" in nature and to cause relatively permanent changes in brain structure, whereas postpubertal differences in hormone exposure are predicted to be short-lived and to "activate" existing neural circuits. The remainder of this section presents evidence for organizational and activational effects of gonadal steroids on human behavior,

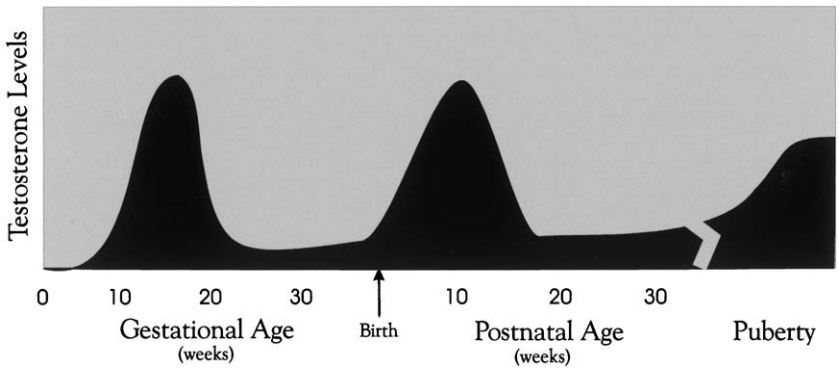


Figure 5 Human males are exposed to high levels of testicular testosterone during three different developmental windows: weeks 10–20 of gestation, the first few postnatal months, and throughout adulthood.

cognition, and psychopathology, although the reader is reminded that the distinction between organizational and activational effects of hormones can be blurry.

A. Organizational Effects of Hormones

Organizational effects of gonadal steroids are suspected when a neural sex difference is found early in life. For example, because human brain size is already larger in males than in females at birth, this sex difference cannot reasonably be attributed to differences in adult hormone levels or differences in socialization between the two sexes. Prenatal hormones, genes, or some combination of the two must be responsible for the sex difference in brain size. Similarly, several of the neuropsychiatric disorders discussed earlier (e.g., cerebral palsy, mental retardation, autism) can be diagnosed in infancy and are almost certainly due to anomalies of brain development occurring in the pre- or perinatal period. Because each of these disorders is more prevalent in boys, a sex difference in the susceptibility to developmental perturbation is indicated. Several investigators have proposed that perinatal testosterone exposure predisposes males to neurodevelopmental disorders by altering the pace of brain development. Specifically, because cerebral development is slower in human males than in females, males may experience an extended window of susceptibility to insult. Interestingly, the pace of functional cortical development is also sexually dimorphic in nonhuman primates, and this sex difference can be eliminated by treating female monkeys with testosterone during pre- and perinatal life.

Perinatal hormonal events may contribute to the sex difference in human juvenile play behavior, discussed earlier. Sex differences in play can be observed very early in life and occur during a time when gonadal hormone secretions are low and not sexually dimorphic. As mentioned previously, boys tend to exhibit more active play than do girls, and this difference is seen

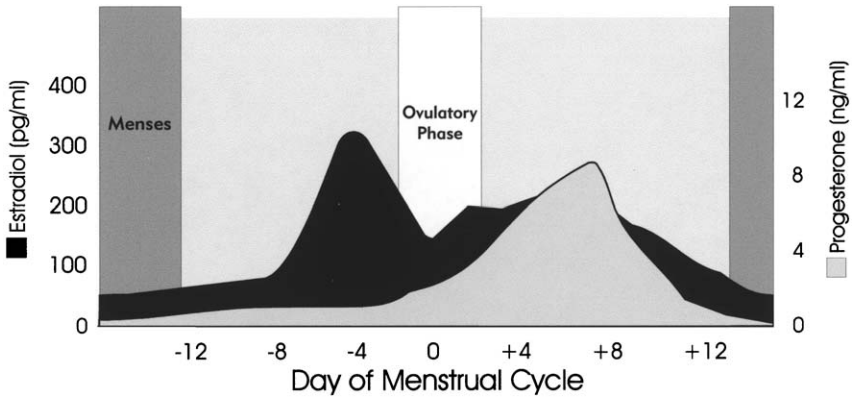


Figure 6 Schematic depiction of fluctuations in estrogen and progestin levels across the menstrual cycle in humans.

across cultures. Many nonhuman mammals, including rhesus monkeys, also exhibit a sex difference in juvenile play, with males displaying more rough-and-tumble play than females. Goy and colleagues have shown that this sex difference in monkey play behavior is due, at least in part, to hormone exposure *in utero*. When female rhesus macaques are exposed to testosterone during late fetal life, their play behavior is masculinized and they exhibit more rough-and-tumble play than do unexposed females. Similarly, play behavior in female rats is masculinized by early exposure to testosterone. Because of obvious ethical constraints, similar experimental manipulations of pre- and perinatal hormone exposure are not possible in humans. Most of the information relevant to the possible organizational effects of gonadal steroid hormones on human brain development, instead, has come from patients who experienced sex-atypical hormone exposure *in utero*, either by virtue of naturally occurring disorders ("experiments of nature") or as a result of treatment for a medical condition.

1. Excess Androgen Exposure of Females *in Utero*

One "experiment of nature" is the condition known as congenital adrenal hyperplasia (CAH). CAH is precipitated by a genetic defect in one of several adrenal enzymes. This defect causes insufficient production of glucocorticoid hormones and concomitant overproduction of androgens, beginning *in utero*. The increased level of androgens masculinizes the external genitalia of affected females to varying degrees, ranging from a mildly enlarged clitoris to fused labia and a pseudo-penis. Normally, this condition is detected at birth, and the individual has corrective surgery on the genitalia, as well as cortisol replacement therapy to correct the underlying problem. Thus, female patients with CAH are exposed to high levels of androgens only during development and are usually raised and socialized as girls.

Research on CAH girls and women offers some support for organizational effects of androgens on human behavioral development. Young girls with CAH tend to exhibit more masculine behavior and/or less feminine behavior on a number of measures. For example, some studies have shown that CAH girls exhibit higher levels of aggression, more rough-and-tumble play, better spatial ability, and greater preference for male-typical toys than their unaffected female relatives. CAH girls also tend to show less interest in infants, marriage, and motherhood, as well as a higher incidence of homosexuality and/or bisexuality than do control girls. However, several investigators have raised concerns that simply having knowledge of the condition may affect the behavioral expectations of both CAH girls and their parents. For example, parents who know their daughter was born with a pseudo-penis might expect more "tomboyish" behavior. This confound could influence retrospective reports of a child's behavior, which is a problem because most studies of CAH girls have relied on child and parent interviews rather than on direct behavioral observation. It should be noted, however, that the few studies directly examining behavior do support the finding of more male-typical play in CAH girls than in their unaffected female relatives. For example, Berenbaum and colleagues videotaped play sessions of CAH and control girls. They found that both play activity preference and playmate preference were masculinized in CAH girls in early childhood, although only play activity preference remained masculinized throughout childhood.

Another strategy for evaluating the effects of gonadal hormone exposure on female brain development comes from studies of "intrauterine position." In litter-bearing mammals, female fetuses gestating next to males are exposed to more testosterone than are female fetuses with female uterine neighbors. The testosterone exposure from male uterine neighbors somewhat masculinizes later sex behavior in female rats, mice, and gerbils. A similar mild masculinization may occur in human females with a male twin. Several studies have reported increased "tomboyism" in females with male twins. Although such observations could easily have a social explanation, intriguing evidence for *in utero* masculinization of female cotwins comes from the study of otoacoustic emissions. The human inner ear produces weak clicks, called otoacoustic emissions, that can be measured by inserting a tiny microphone into the ear. The rate of these clicks is modulated by neural input. Although the function of otoacoustic emissions is not clear, these emissions exhibit a sexual dimorphism measurable as early as 1 month of age: females produce more and stronger clicks than do males. McFadden has shown that human females with a male twin have fewer and weaker otoacoustic emissions (i.e., they are more masculinized) than females with either a female twin or no twin. Because it is difficult to see how socialization could affect otoacoustic emissions, this observation suggests that intrauterine testosterone alters development of the auditory system in humans.

2. Excess Estrogen Exposure of Females *in Utero*

As stated previously, many of the neural sex differences seen in rodents are due not to testosterone per se but to the neural conversion of testosterone to estrogenic metabolites. Children who have been exposed to excess estrogens *in utero* may allow us to determine whether prenatal estrogen similarly masculinizes the human brain. Diethylstilbestrol (DES), a nonsteroidal estrogen, was given to many pregnant women in the 1940s through the 1960s to prevent miscarriage or other pregnancy complications. In female rats exposed to DES during fetal life, juvenile play and some aspects of adult sex behavior are masculinized. *In utero* DES exposure also results in a permanent change in the gonadotropin secretion pattern in rhesus monkeys. The results from studies of the human children born to women given DES during pregnancy, however, are more ambiguous. DES exposure does increase the risk of genital tract problems (e.g., menstrual irregularities, a high infertility rate, and risk for certain types of cancers), but DES-exposed women do not differ reliably from their unexposed sisters in visuospatial ability, verbal ability, or incidence of marriage and motherhood. These findings suggest that *in utero* estrogen exposure does not masculinize a human fetus. However, there are caveats to interpreting these negative findings. Only a few published studies are available, and most of these had relatively small sample sizes. Moreover, the timing and amount of DES exposure can differ dramatically among subjects of a single study. Because different aspects of sexual differentiation have different critical periods, the various levels and durations of hormone exposure could obscure any behavioral effects of DES exposure.

3. Blockade of Androgens in Genetic Males

According to the organizational hypothesis, females can be masculinized by perinatal exposure to testicular hormones, but the hypothesis also predicts that males will be feminized or demasculinized when they are *not* exposed to testosterone. The study of individuals with androgen insensitivity (AI) is relevant to this latter prediction. As described previously, AI results from a mutation of the androgen receptor gene. Patients with complete AI are incapable of responding to testosterone or other androgens and are phenotypically female. The cognitive abilities and other behaviors of genetic males with AI also appear to be feminized. In fact, a few studies have found that these AI individuals exhibit higher verbal comprehension and lower visuospatial skills than their unaffected sisters. In other words, in the absence of any androgen stimulation, cognitive abilities may be "superfeminized" in AI males. Again, however, very few well-controlled studies have been performed. A more serious constraint is that any effect of AI on cognition or behavior could not *a priori* be ascribed to an organizational effect of hormones, because genetic males with complete AI are almost always reared as females, and sex of rearing is therefore a confounding variable.

B. Activational Effects of Hormones

Sex differences exhibited in adulthood may be caused, or modulated, by activational effects of gonadal steroid hormones. In this section, we consider evidence for the the effects of circulating hormones on human brain function and/or structure by considering activational effects of androgens and estrogens on aggression, cognition, and neuropsychiatric disorders in adults.

1. Aggressive Behavior

Males are more aggressive than females in many mammalian species, including humans. An activational effect of testosterone is implicated in this sex difference because intermale aggression increases at puberty, and castration of adult males reduces aggression across a wide range of species. In rodents, testosterone also has organizational effects on aggression, apparently by influencing the sensitivity to androgen stimulation in adulthood. However, the role of androgens in aggression among primates is somewhat equivocal. Manipulations of adult testosterone levels do not consistently alter aggression in nonhuman primates, leading investigators to conclude that either androgens are not involved or that androgens exert their primary influence on primate aggression early in development.

Studies of human aggression generally do not take into account organizational effects of steroid hormone exposure. At least five studies of male prisoners have found evidence of a positive association between adult testosterone levels and aggression, and women convicted of violent crimes also may have higher salivary testosterone levels than those convicted of nonviolent crimes. In the general population, there seems to be a small, positive correlation between circulating

testosterone levels and aggression (usually as measured on a paper-and-pencil test). The overall impression is that current circulating androgen levels account for some of the differences in aggression between sexes but only weakly predict variations in aggression *within* each sex. One problem with most studies on testosterone and human aggression, however, is that they are correlational in nature. Because violent or competitive behavior can itself increase androgen secretion, the direction of causality between aggression and testosterone levels is unclear.

2. Cognition

Several investigators have taken advantage of the fact that sex steroid hormone levels change markedly across the menstrual cycle to examine possible activational effects of ovarian steroids on cognitive abilities. In one report, Hampson found that women's articulatory and fine motor skills were facilitated at the midluteal phase of the menstrual cycle when estrogens and progestins are high. By contrast, best performance on a perceptual–spatial test was observed during menses, when estrogens and progestins are low. In other words, high levels of ovarian hormones facilitated skills normally favoring females but were detrimental to skills favoring males. If so, then activational effects of ovarian hormones may contribute to some of the sex differences in cognitive abilities described earlier. However, not all subsequent reports have corroborated Hampson's finding. Some of the discrepancies in this literature may be due to the fact that investigators have measured a single cognitive trait (e.g., visuospatial ability) using various instruments. Different tests probably tap into different aspects of a complex cognitive function such as visuospatial ability, not all of which may be responsive to hormonal variations.

At least two studies have examined the activating effects of steroid hormones on cognitive abilities in transsexuals. Male-to-female transsexuals often are treated with antiandrogens and/or estrogens, and female-to-male transsexuals frequently receive androgens as part of the sex-reassignment process. By examining cognitive abilities in the same individuals before and after cross-sex hormone treatment, one can test for an activational role of sex steroid hormones in cognition. Van Goozen and colleagues report that 3 months after androgen treatment, visuospatial ability was increased and verbal fluency decreased in female-to-male transsexuals. In other words, the androgen treatment was associated with a shift to a "male" cognitive style. Smaller changes in the opposite direction were reported in male-to-female transsexuals 3 months after antiandrogen plus estrogen treatments. In a second study, Miles and colleagues compared cognitive performance in transsexuals currently receiving estrogens (treatment duration of at least 4 months) with that of transsexuals awaiting hormone treatment. Increased performance on a memory task (Paired Associate Learning) that is normally performed better by women than men was observed in the estrogen-treated group. No effect of estrogen treatment was seen on another memory test (Digit Span), for which there is not normally a sex difference in the general population. Estrogen did not affect visuospatial ability, as measured by a Mental Rotations test in this study.

In adult rats, surgical removal of the ovaries results in a decrease in dendritic spines (a site of synapse formation) on hippocampal neurons, and this decrease can be prevented by treating the ovariectomized rats with estradiol. Because the hippocampus is critically involved in memory formation, these observations indicate a possible neuroanatomical basis for estrogen modulation of memory function. Changes in synapse density in the hippocampus are also seen over the course of the rats' 4- to 5-day-day estrous cycle, with a higher density of synapses seen on days of high estradiol levels. This latter result indicates that morphological effects of estradiol on synapse formation can be remarkably rapid. Such effects of estrogens on neuronal morphology and/or function may be direct or may be mediated by protein neurotrophic factors such as nerve growth factor and brain-derived neurotrophic factor, both of which are up-regulated by estrogens in regions of the rat brain involved in cognition. It is not known whether estrogens similarly affect the morphology of the human brain. The anatomical resolution of current neuroimaging techniques is not nearly sensitive enough to detect microscopic features such as dendritic spines in the living human brain.

3. Neurological and Psychiatric Disorders

In keeping with the salutary effects of estrogens on cognitive function in healthy women, described earlier, small clinical trials indicate that estrogens may improve the memory of women suffering from Alzheimer's disease (a neurological disorder marked by cognitive decline and profound memory loss). Postmenopausal estrogen therapy also has been reported to reduce the risk of Alzheimer's disease, leading several

researchers to postulate that the higher age-specific prevalence of Alzheimer's disease in women than in men may be related to the relative lack of estrogens in postmenopausal women. Men presumably retain access to estrogens throughout life, either via direct secretion from the testes or via locally converted testosterone. Indeed, at least one study has found a positive correlation between circulating estradiol levels and performance on a visual memory test in healthy young men.

Similarly, the well-established sex differences in schizophrenia (i.e., later onset and less severe symptoms in females) may be accounted for by a "protective" effect of estrogens in premenopausal women. In support of this idea, several studies report variations in schizophrenic symptoms across the menstrual cycle, with the lowest level of psychopathology observed at the time of peak estrogen levels. In addition, hospitalizations of women experiencing an acute schizophrenic episode are highest during the perimenstrual period, when estrogen levels are low. Finally, whereas most cases of schizophrenia are first diagnosed in late adolescence and early adulthood in both males and females, only females exhibit a second peak of new diagnoses between 45 and 49 years of age (i.e., around the time of menopause). A link between unipolar depression and the menstrual cycle has also been noted, with increased vulnerability to the onset of a depressive episode or the worsening of ongoing depression during the premenstrual phase of the cycle.

These various findings obviously do not "prove" that gonadal sex steroids, secreted in adulthood, cause any of the many sex differences in the incidence of neurological or psychiatric disorders (Table III). However, it seems clear that sex steroid hormones modulate the neuro- and psychopathology of adults and may act in concert with other factors (including, possibly, organizational effects of hormones) to influence the susceptibility to disease.

V. GENES, HORMONES, AND EXPERIENCE: INTERACTIVE EFFECTS ON BRAIN DIFFERENTIATION

A. Genes

The role of hormones in sexual differentiation has been emphasized in the foregoing discussion, and, indeed, hormones appear to be primarily responsible for sexual dimorphism of neural and somatic tissues in mammals. However, there may also be direct genetic contributions to sexual differences. For example, the work of Renfree and colleagues indicates that development of the scrotum, pouch, and mammary glands of tamar wallabies is under direct genetic control. Normally, a pouch and mammary glands are present only in female wallabies, whereas only genetic males have a scrotum. However, the scrotal anlagen develops in fetal males, and a rudimentary pouch is identified in females *before* differentiation of the gonads, indicating that gonadal hormones cannot drive the differentiation of these structures in wallabies. In this case, genes on the X chromosome apparently control sexual differentiation, with pouch and mammary glands developing whenever two or more X chromosomes are present.

Are any *neural* sex differences under direct genetic control? A few sex differences have been identified in the brains of birds and rodents that so far defy a simple hormonal explanation. A genetic basis for these sex differences therefore is suspected, but not yet proven. In principle, any gene that is present in one sex but absent from the other could contribute to neural sex differences. Because only genetic male mammals have a Y chromosome, any Y-chromosome gene that influences nervous system development or maintenance could be responsible for neural sex differences. A few Y-chromosome-specific genes (including SRY) have been shown to be expressed in the brain, although, thus far, the link to sex differences has not been made. Genes on the X chromosome may also contribute to sex differences in neural development, because females have two copies of each X-chromosome gene, whereas males have only one.

Rett syndrome provides a dramatic illustration of how genes can contribute to sex differences in a neurological disorder. Children with Rett syndrome are apparently normal at birth, but undergo developmental regression beginning at about 18 months of age, allegedly due to an arrest of normal neuronal maturation. Essentially 100% of all Rett syndrome patients are female. How can such an absolute sex difference in a neurodevelopmental disorder be explained? Rett syndrome has been linked to a gene on the long arm of the X chromosome. Genetic males have only one copy of all X-chromosome genes, and males who inherit the gene responsible for Rett syndrome apparently die *in utero* or soon after birth. Because females have two copies of each X-chromosome gene, those who inherit a single defective copy of the gene causing Rett syndrome may compensate for the defect to some extent. Viewed in this light, females are the only long-term survivors of Rett syndrome, and

males are more vulnerable to this syndrome, as they are to most neurodevelopmental disorders. Although few psychiatric or neurological disorders are exclusive to only one sex, X- and/or Y-linked genes may contribute in important ways to the likelihood of exhibiting a given disorder, to the severity of symptoms, or to the disease course. Gonadal steroid hormones likely interact with sex-linked genetic predispositions to shape sex differences in clinical profiles of many disorders.

B. Experience

In both popular and scientific thought, a distinction is often made between biological and social causes of sex differences. Although it may not be articulated, implicit in this line of thinking is the notion that hormones and gene expression affect the structure of the brain, whereas experience influences the more intangible realm of behavior. This dichotomy, however, is a false one. By definition, people exhibiting different behaviors must have differences in the function or structure of the brain, regardless of whether that difference is due to genes, hormones, experience, or the interaction of all three. The animal literature provides us with several concrete examples of how hormones and experiences may interact to alter neuroanatomy and behavior. For example, when a rat is exposed to stress during pregnancy (by restraining her under bright lights for a few hours a day), the prenatal hormone exposure and future behavior of her male offspring are altered. Specifically, the prenatal testosterone surge is blunted in the male offspring of stressed mothers, and these males are more likely to display female-typical behaviors in adulthood than are the male offspring of nonstressed mothers. Sexually dimorphic neural structures, such as the SDN-POA and SNB, are also affected in the male offspring of stressed dams.

Hormones may also influence how an individual interacts with the environment. For example, testosterone may predispose males to more active, rough-and-tumble play during childhood, and this predisposition may then be magnified or diminished by societal pressures. Regardless of how it originates, the differing play behavior of girls and boys will lead to very different childhood experiences, which can be expected to shape the brain and behavior. In this scenario then, early hormone exposure results in sex differences in adulthood behavior, but only via a circuitous route that is strongly modified by experience. Another example comes from the work of Juraska and colleagues. Male rats, when raised alone in standard cages, have more hippocampal dendrites than do females or males castrated at birth. However, this sex difference is actually reversed when the animals are reared in large cages with many cagemates and toys, indicating that the environment must interact with hormones and/or genes to produce neuroanatomical sex differences. This type of interaction is likely to be even more prevalent in a species such as humans, who are social, long-lived, and exposed to complex and varied environments over the course of a lifetime.

See Also the Following Articles

HYPOTHALAMUS • PEPTIDES, HORMONES, AND THE BRAIN AND SPINAL CORD • PSYCHONEUROENDOCRINOLOGY • SEX DIFFERENCES IN THE HUMAN BRAIN • SEXUAL BEHAVIOR • SEXUAL DYSFUNCTION • SEXUAL FUNCTION • STRESS: HORMONAL AND NEURAL ASPECTS

Suggested Reading

Archer, J. (1991). The influence of testosterone on human aggression. *Brit. J. Psychol.* **82,** 1–28.

Breedlove, S. M. (1994). Sexual differentiation of the human nervous system. *Ann. Rev. Psychol.* **45,** 389–418.

Collaer, M. L., and Hines, M. (1995). Human behavioral sex differences: A role for gonadal hormones during early development? *Psych. Bull.* **118,** 55–107.

Forger, N. G. (2001). The development of sex differences in the nervous system. In *Developmental Psychobiology, Volume 13 of The Handbook of Behavioral Neurobiology* (E. Blass, Ed.), pp. 153–208. Plenum, New York.

George, F. W., and Wilson, J. D. (1994). Sex determination and differentiation. In *The Physiology of Reproduction*, 2nd ed. (E. Knobil and J. D. Neill, Eds.), pp. 3–28. Raven Press, Ltd., New York.

Lambe, E. (1999). Dyslexia, gender, and brain imaging. *Neuropsychologia* **37,** 521–536.

Murray, R. M. (1991). The neurodevelopmental basis of sex differences in schizophrenia. *Psychol. Med.* **21,** 565–575.

Peters, M. (1991). Sex differences in human brain size and the general meaning of differences in brain size. *Can J. Psychol.* **45,** 507–522.

Sexual Dysfunction

LISA REGEV, CLAUDIA AVINA, and WILLIAM O'DONOHUE
University of Nevada, Reno

GLOSSARY

dopaminergic Relating to, participating in, or activated by dopamine or related substances.

incidence The occurrence of new disease per unit of person–time.

prevalence The percentage of a population that is affected with a particular disease or disorder at a given time.

psychogenic Originating in the mind or in mental or emotional conflict.

vasodilation Widening of the lumen of blood vessels.

Sexual functioning is a complex series of biopsychological events, which in turn is affected by other series of complex biopsychological events. Much research is needed to understand the phenomena involved. This article reviews what is known biologically and psychologically regarding male and female sexual dysfunction.

I. INTRODUCTION

From adolescence onward, sex is a very important activity for most individuals. Sex can lead to the experience of great pleasure and the giving of great pleasure. It can result in a special bond. Sex can lead to sons and daughters—another one of life's greatest joys. If can also have direct connections to our personal identities and to our self-esteem.

However, there are a number of difficulties associated with sex. Finding the right number of partners and finding partners who are satisfying and whom we satisfy comprise one set of problems. Finding certain proscribed activities arousing, such as sex with children or coerced sex, comprise another set of problems commonly called the paraphilias. However, in this article, we focus on the sexual dysfunctions, problems with desire, arousal, orgasm, or pain, as identified by the *Diagnostic and Statistical Manual of Mental Disorder* (DSM).

DSM is the generally accepted nosology when classifying mental health problems. Problems of desire, arousal, orgasm, and genital pain are currently the disorders classified as sexual disorders. When diagnosing these problems, DSM identifies three criteria for each disorder. The first criterion for each disorder is the distinguishing criterion and will be discussed later for each disorder. The remaining two criteria are similar for each of the sexual disorders. One criterion is that the problem must cause significant distress or interpersonal difficulty. For example, if an individual is not bothered by his or her inability to reach orgasm, then it is not diagnosed as a problem. The last criterion requires the clinician to rule out other possible sources for the problem. This includes another Axis I disorder (e.g., Posttraumatic Stress Disorder, Obsessive–Compulsive Disorder, and a mood disorder), the direct physiological effects of a substance (e.g., alcohol, drugs, medications), or a general medical condition.

If an individual meets criteria for a sexual disorder, it is essential to specify the onset, context, and etiology of the disorder. Regarding onset, an important distinction is between patients who have always experienced

the problem (lifelong) and those who did not experience the problem in the past but more recently have begun to experience the problem (acquired). Regarding context, it is important to specify whether the person experiences the problem in any context (generalized) or in limited situations (situational). Regarding etiological factors, it is important to specify whether the problem is due to psychological factors or a combination of psychological and organic factors. If the etiology of the disorder is found to be due exclusively to the physiological effects of a general medical condition, the individual is diagnosed with Sexual Dysfunction due to a General Medical Condition. DSM is in its fourth revision and is continually changing. The current classification will likely change in the not-too-distant future as more research in this area continues to inform clinical assessment. Whereas psychological and medical problems continually interact with one another in the development and maintenance of a sexual problem, this article will discuss these issues separately to help the clinician identify appropriate assessment and treatment modalities.

II. DESIRE DISORDERS

In 1977, sexual desire was explicated as significantly influencing sexual functioning by both Lief and Kaplan. Definitions of desire vary in terms of whether a biological component is involved and the importance of psychological, biological, and environmental components. Difficulties in defining desire arise from its subjective nature. In 1998, Leiblum and Rosen viewed sexual desire as a "subjective feeling state that may be triggered by both internal and external cues, and that may or may not result in overt sexual behavior." This definition involves neuroendocrine functioning and exposure to individual and environmental sexual cues. Since the introduction of desire as an integral component of the sexual response cycle, problems relating to sexual desire have been found to be the most common sexual disorders. DSM-IV identifies two sexual desire disorders: 302.71 Hypoactive Sexual Desire Disorder (HSD) and 302.79 Sexual Aversion Disorder (SAD).

HSD is characterized by "persistently or recurrently deficient (or absent) sexual fantasies and desire for sexual activity." Additionally, the clinician is advised to take into account factors that affect sexual functioning (e.g., age and the context of the person's life). As indicated earlier, when diagnosing sexual problems, the second and third criteria are similar for all of the sexual dysfunctions.

Epidemiological data suggest that 1–15% of males and 1–35% in females suffer from HSD. Assessment of prevalence is complicated because it is likely that only those HSD patients with an unsatisfied partner who motivates the patient to seek treatment will be evaluated. Due to the nature of this disorder, HSD patients may not have a sexual partner, and some may credit low desire to the lack of sexual opportunities and, thus, may not seek treatment.

According to DSM-IV, Sexual Aversion Disorder is characterized by "persistent or recurrent extreme aversion to, and avoidance of, all (or almost all) genital sexual contact with a sexual partner." This disorder was introduced in DSM-III-R to distinguish between those who did and did not experience severe disgust and avoidance regarding sexual activity. The disorder was added to the diagnostic manual on the basis of clinical experience that victims of rape and child sexual abuse experienced this sexual problem. Sexual aversion disorder describes an individual's unwillingness to participate in a sexual activity. This disorder is distinguished from sexual phobias in that it describes an intense dislike or disgust but not an excessive fear.

An adequate research base regarding the incidence and prevalence of sexual aversion does not exist. These data are unknown because sexual aversion is a poorly recognized syndrome, sexual panic has received little professional attention, and epidemiological studies on sexual disorder have not included sexual aversion. Another problem is that sexual aversion problems may be comorbid with other psychological problems. Aversion may occur with another more prominent sexual disorder such that it is undiagnosed or with other disorder such as panic so that the sexual problem is missed if clinicians fail to ask about sexual stimuli. At this point we can only make inferences from research studying general sexual disorders and likely comorbid problems.

A. Etiology and Treatment of Desire Disorders

There are multiple etiological factors that may be involved in disorders of sexual desire. These include both medical and psychological factors. The following will discuss possible etiological factors associated with each of these as well as possible interventions.

1. Medical Aspects of Desire Disorders

The pituitary–gonadal system is the major endocrine system involved in sexual behavior. Sexual desire

results from diverse hormones, neurochemicals, and neuropeptides. The hypothalamus significantly impacts sexual desire as it mediates central nervous system effects on the anterior pituitary–gonadal system. The hypothalamus is responsible for producing and releasing hormones and neurotransmitters, including gonadotropin-releasing hormones (GnRH), luteinizing-releasing hormones (LHRH), thyrotropin-releasing hormones (TRH), oxytocin, and vasopressin.

a. Neurophysiology in Males Hormones allow one group of cells to influence the activity of other groups of cells. Gonadal hormones influence sexual desire because sex hormones affect cells in the brain by way of the hypothalamus, which has receptors for gonadal steroids. The hypothalamus affects the anterior pituitary–gonadal system by secreting and transporting GnRH. GnRH then stimulates basophilic cells of the anterior pituitary to produce LHRH and follicle-stimulating hormone (FSH). LHRH is secreted from the pituitary and travels to the testes, where it stimulates the production of testosterone (also a hormone).

b. Neurophysiology in Females The endocrinology of females is analogous to that of males but is influenced by changes in the menstrual cycle. There is an increase in FSH during the first half of the menstrual cycle. A new ovarian follicle also begins to grow that secretes increasing amounts of estradiol. A surge in LH provokes ovulation. The ovarian follicle produces estrogen and progesterone after ovulation. Although one would infer that desire fluctuates with changes in estrogen and progesterone, there is minimal support for this conclusion. Additionally, prolactin, which is capable of initiating and sustaining lactation in females, is considered to be potentially important in the sexual desire of females. However, the relationship between prolactin and desire is disputable.

c. Clinical Implications for Hypoactive Sexual Desire Disorder There are a large number of etiological factors involved in sexual dysfunction. However, there is an inadequate amount of research studying the relationship between medical conditions and desire. Studies of sexual desire are limited because desire can only be assessed through self-report, and these measures have unknown or poor psychometrics.

Women with Turner's syndrome, in whom one X chromosome is missing and as a result they do not develop gonads, have confused and diminished desire due to their delayed growth and social maturity. Whereas many studies evaluate the process of menopause and its effects on women, researchers do not agree that postmenopausal women have diminished sexual functioning. In 1953, Alfred Kinsey found that up to 48% of postmenopausal women reported that their sexual response and frequency of intercourse had decreased, whereas studies in 1985 demonstrated that 50% of the sample did not report a decrease in sexual interest after menopause and less than 20% of the women in the sample indicated a significant decrease. Findings are unclear as to whether losses in sexual desire in postmenopausal women are estrogen-dependent. In a 1999 study, sexual problems in women tended to decrease with increasing age. These findings did not investigate the onset of menopause or estrogen levels in their sample.

Sexual interest and activity decrease after exogenous androgen removal in hypogonadal men. Also, ejaculation and orgasm stop after this procedure. The relationship between testosterone and desire has been supported by findings that testosterone injections increase sexual interest but do not affect erectile functioning. Dialysis patients with sexual dysfunction have a significantly higher level of prolactin, a hormone secreted by the anterior pituitary gland into the bloodstream. Endocrine diseases have been suggested to affect libido and menses during later stages of these diseases.

Many medical treatments decrease sexual desire. Antihypertensive medications such as methyldopa, clonidine, reserpine, propanolol, and spironolactate decrease sexual desire in patients taking these drugs. Some antipsychotic medications, including thioridazine and fluphenazine, decrease sexual drive and interest. This may occur by blocking dopamine effects and suppressing the hypothalamic–pituitary–gonadal axis.

It is still largely unknown whether drugs can possibly enhance sexual desire in humans. Generally, studies have been conducted with animals, and it is difficult to determine what components of an animal's sexual behavior are equivalent to human sexual desire. Specific drugs such as L-Dopa and amphetamines may potentially increase sexual desire due to their influence on dopaminergic neurotransmission. L-Dopa is used to treat Parkinson's disease, and there is some evidence that it increases sexual interest. Difficulties in interpreting these results include distinguishing whether the increase is due to the medication, amelioration of the disease, a decrease in inhibitions, or a hypomanic syndrome produced by the drugs. There are only a few preliminary reports that suggest that amphetamines may enhance sexual desire. Amphetamines release

dopamine and are dopaminergic agonists. It is difficult to draw conclusions about their effect on sexuality because a large percentage of amphetamine users have personality disorders and/or other psychological or medical problems.

d. Clinical Implications for Sexual Aversion Disorder Biological treatment for this disorder consists of medications that decrease anxiety and suppress the panic response associated with sexual aversion. Medications such as tricyclic antidepressants, MAO inhibitors, and benzodiazepines are used because they reduce the firing rate of the locus ceruleus (LC), which is potentially hyperactive in anxious individuals. It is recommended that medications be used only when the panic response is so strong that it interferes with psychological treatment.

2. Psychological Aspects of the Desire Disorders

There are a number of possible etiological factors for the desire disorders that are psychological in nature. For HSD, these factors may include poor marital adjustment, low levels of emotional closeness and romantic feelings, religious orthodoxy, affective disturbances, obsessive–compulsive personality, problems concerning gender or sexual identity, fear of loss of control, fear of pregnancy, and fear of contracting a sexually transmitted disease. Of these, anxiety has been the most frequently cited factor affecting desire. Etiological factors of a psychological nature contributing to SAD are not clear, and sexual assault may result in an aversion to sex.

There is no standard assessment procedure for assessing sexual desire disorders. Current assessment practices include the use of self-report questionnaires, interviews of the individual and/or the couple, and medical evaluations. Some of the more well-known assessment protocols for sexual desire are discussed next. These assessments may be used when evaluating any of the sexual dysfunctions.

Studies in 1983 describe a semistructured clinical interview comprising seven components: (1) assessment of the chief complaint, (2) a sexual status exam, (3) assessment of medical status, (4) assessment of psychiatric status, (5) family and psychosexual history, (6) evaluation of the current relationship, and (7) summation and recommendations. The interview involves an assessment of relevant symptoms, an assessment of onset and progress of the sexual problem, an analysis of behavior and emotional conflict, a full medical evaluation, and a meeting with the clients to share the clinician's findings and treatment recommendations. The interview is based on 12 years of experience as a sex therapist but has not been empirically supported.

Others recommended the use of interviews with the individual experiencing the sexual problem as well as with the partner. The interview inquires about the following: (1) Does the individual experience complete or intermittent loss of desire? (2) Is the loss solely partner-related or relevant to all sexual stimuli? (3) What are the other symptoms? (4) Are there other relationship changes associated with the loss of sexual desire? The interview also asks about family background, sexual knowledge, satisfaction with current relationships, medical, drug, and psychiatric histories, and sexual development and experiences. The goals of the assessment interview are to identify the problem, develop hypotheses about etiology, and commence the therapeutic alliance.

Later came recommendations for an interview that examines biological, cognitive, and motivational factors of sexual desire, with suggestions that the clinician assess the partner's reaction to low desire, past trauma, abuse, and physical disorders, the psychiatric history of the dysfunctional partner, past trauma, and gender or sexual orientation conflict. Others suggest conducting interviews with the individual and the couple that assess sexual and nonsexual stressors in the relationship, complete sexual histories, onset and history of the problem, and maintaining factors.

Additionally, a series of self-report measures may be used when assessing problems with sexual desire. These include the Sex History Form and the Locke–Wallace Marriage Inventory. Commonly used assessment instruments for sexual functioning include the Derogatis Interview for Sexual Functioning (DSFI), the Golombok Rust Inventory of Sexual Satisfaction (GRISS), and the Sexual Interaction Inventory (SII). These measures are self-report questionnaires that generally provide an overall measure of the quality of sexual functioning. They are not appropriate for diagnosing specific sexual dysfunction disorders.

Currently there is no standard assessment procedure for sexual aversion disorder. This disorder may be difficult to address because clients are not aware of their aversion or are unwilling to disclose this information. There are two self-report questionnaires that may be helpful in assessing sexual aversion: the Sexual Anxiety Inventory (SAI) and the Sexual Aversion Scale (SAS).

Whereas the SAI does not specifically assess the aversion or avoidance associated with this disorder, this measure inquires about anxiety and worry about a

number of sexual situations. The SAS assesses avoidance of sexual situations and the desire for help with a sexual problem. It also asks about appropriate fears to situations such as contracting AIDS or becoming pregnant. The SAS may be used as a preliminary screening device for identifying the specific fear underlying the aversion.

These questionnaires have not been tested with a clinical population of individuals suffering from sexual aversion. Therefore, the SAI and SAS are not appropriate for making a diagnosis of sexual aversion. Assessment should include a thorough sexual history as well as aspects commonly associated with the disorder, such as child sexual abuse, panic attacks, familial history of panic disorder, self-blame, and worry about sexual performance.

Orgasm consistency training has been found to be effective when treating women with HSD. It is presumed that by teaching women to reach orgasm more frequently, women would be more likely to enjoy sexual interactions and therefore desire to engage in these interactions more frequently. Additionally, marital therapy may be effective in treating low desire if couples problems, including emotional intimacy problems, are present within the couple.

For individuals with SAD, systematic desensitization has been shown to be an effective form of treatment. This treatment consists of desensitizing the individual to the feared stimulus. This may include fear or disgust relating to genitalia, the sexual act itself, as well as other more preliminary sexual activities. The individual is gradually desensitized by first focusing on less fear-provoking stimuli and eventually focusing on the most fear-inducing stimuli. These stimuli are repeatedly presented to the individual while the individual does not experience adverse consequences. Subsequently, the fear or anxiety is reduced. By reducing the fear, the person is able and more willing to engage in sexual behavior.

III. SEXUAL AROUSAL DISORDERS

Arousal consists of a set of physiological responses and subjective physiological responses. Problems with male arousal include difficulties attaining or sustaining an erection and is commonly termed "impotence," whereas arousal problems in females involves the lubrication–swelling response.

DSM-IV identifies two sexual arousal disorders: 302.72 Female Sexual Arousal Disorder (FSAD) and 302.72 Male Erectile Disorder (MED). According to the DSM-IV, Female Sexual Arousal Disorder is characterized by "persistent or recurrent inability to attain, or to maintain until completion of the sexual activity, an adequate lubrication–swelling response of sexual excitement."

If a woman develops this sexual dysfunction, she is probably more likely to also lose interest in sex or have orgasm difficulties. It is not clear whether this occurs because the latter problems are more prevalent, or whether clinicians and scholars place more attention on problems of desire and orgasm difficulties than they do on problems of female arousal. However, this may be partly due to the complexities in defining female arousal. Specifically, problems in sexual arousal may be present even if the female engages in sexual intercourse, and one cannot assume that if she is engaging in coitus she is sufficiently aroused to reach orgasm. Epidemiological data are lacking because studies have failed to inquire about arousal when measuring sexual functioning. Also, the existing estimates vary greatly due to differences in the criteria being used.

According to the DSM-IV, Male Erectile Disorder is characterized by "persistent or recurrent inability to attain, or to maintain until completion of the sexual activity, an adequate erection." The American Psychiatric Association has described this problem as a sustained inability to maintain an erection that allows penetration and ejaculation to take place and is caused by psychological factors. Approximately 10% of a sample of young adults in Europe and the United States suffer from MED. Erectile difficulty is estimated to be prevalent in between 3% and 9% of males. Also, this problem is increasingly found in specific populations, such as the elderly and those who suffer from diabetes, cardiovascular disease, and chronic renal failure.

A. Etiology and Treatment of Sexual Arousal Disorders

Problems of arousal may be the result of physiological or psychological problems. When presented with an individual suffering from this problem, it is important to determine which of these etiologies is the culprit. The following will discuss possible medical and psychological aspects of the disorder as well as possible treatments to pursue.

1. Medical Aspects of Sexual Arousal Disorders

a. Neurophysiology of the Male The penis becomes erect through the supply of blood from the

internal pudenal artery to that genital area. The penis has three cylinders. The two paired bodies are known as the corpora cavernose (CCP) and are located in the dorsal and lateral parts of the penis. When the penis is flaccid, the CCP are empty and collapsed. The third body is on the ventral part of the penis and is the unpaired corpus spongiosum (CSP). The inner spaces of the CSP are partly filled during the flaccid stage.

Erection occurs when neurotransmitters, including acetylcholine and vasoactive intestinal peptide, are released by higher central nervous system activation or viscerosomatic reflex activation. These neurotransmitters provoke smooth muscle relaxation, resulting in decreased resistance to arterial blood flow. The sinusoidal spaces are enlarged and compress emissary veins. Blood is then trapped in the CCP. When the penis is erect, the CCP increases in diameter and the CSP increases in the bulb region of the penis. During this time, the CCP increases more in diameter than in length, whereas the CSP increases more in length than in diameter. Erection requires an intact functioning of the arterial–venous network.

b. Neurophysiology of the Female Erectile tissues in females include the vulva, vagina, surrounding vasculature, and the clitoris. During the excitement phase, the clitoris, hidden by the anterior ends of the labia minora, undergoes a vasocongestive response and the labia minora increases in diameter. The dorsal artery of the clitoris is a terminal branch of the internal pudendal artery. The dorsal artery provides the clitoris with its blood supply. Nerves originating from T12–L1 supply the labia majora. The pudental nerve originating from S2–S3 terminates in branches to the glans, corona, and prepuce and is responsible for the innervation of the clitoris. Pelvic vasocongestion results in lubrication, a clear viscuous fluid from the circumvaginal vasculature, and any malfunction in this area results in disorder. Reflex vasodilation, labia and circumvaginal tissue swelling, heightened labial coloring, and lubrication accompany the arousal phase in women.

c. Clinical Implications for Male Erectile Disorder Impotence can be caused by any disease that impedes the blood flow through the hypogastric–cavernous arterial tree. Diabetes commonly results in MED. Men are unable to attain and sustain erection due to the impediment of arterial blood flow caused by small-vessel and large-vessel disease induced by diabetes. Neural malfunctioning, including peripheral denervation, pudendal sensory diabetic neuropathy, and nerve damage in CCP tissue, accounts for the majority of erectile disorder in this population. Other disorders such as scleroderma (a connective tissue disease involving thickening of the skin and fibrotic changes), chronic renal failure, chronic lung disease, bronchitis, asthma, emphysema, and neurogenic problems, including multiple sclerosis (MS) and spinal cord lesions and injuries, are also associated with a high incidence of MED.

Medication effects resulting in MED are comparable to findings for sexual desire. Antihypertensive agents, cardiovascular drugs, and psychiatric medications impair erectile ability. Antihypertensives and cardiovascular drugs either affect autonomic innervation by interfering with the neurovascular response to stimulation or can decrease blood pressure in diseased cavernous arteries. Psychiatric medications including antidepressants and antipsychotics may antagonize neurotransmitters by anticholinergic effects.

Internal iliac endarterectomy, transluminal balloon, and aortobifemoral graft anastomoses are medical procedures used to improve blood flow in the internal pudendal and penile arteries. Microsurgical techniques anastomosing the inferior epigastic artery to the deep dorsal artery or the cavernous artery are successful for MED caused by lesions in the internal pudendal artery or its branches. Intracavernous injection therapy consists of either instructing the patient to inject the lateral aspect of the CCP or having the physican inject papaverine every 2–3 weeks for at least three injections.

Several types of prostheses are available for men suffering from MED. Suction devices apply negative pressure to the penile shaft so that an erection can be induced by blood that fills the CCP. A tourniquet is applied to the base of the penis so that the erection can be maintained for coitus. Implants consisting of inflatable cylinders, an indwelling reservoir, and a pump in the scrotum increase girth during erections and are fairly natural looking when the penis is flaccid. Also available are prostheses consisting of a pair of silicone rods that occupy the length of the CCP but that do not detumesce.

A complete sexual history should be collected that inquires about onset, sexual practices, and quality of sexual interactions. Risk factors such as trauma, diabetes, hypertension, and smoking should also be assessed. A thorough medical examination should be employed. This may include an assessment of genital perineal sensations, tone, and reflexes. Physiological instruments for MED include assessment of nocturnal penile tumescence (NPT), which monitors erections occurring during the rapid eye movement phase of

sleep, the penile–brachial index (PBI) in which a pneumatic cuff is applied to the base of the penis and penile systolic pressure is measured, and penile pneumoplethysmography, which also utilizes a pneumatic cuff at the base of the penis and measures waveforms produced by arterial pulsation.

d. Clinical Implications for Female Sexual Arousal Disorder Estrogen deficiency results in atrophy of the vaginal endothelium (responsible for lubrication) and the underlying vasculature. Impairment in estrogen production is most commonly caused by menopause. An insufficient amount of estrogen will not support the lubrication response and decreases vaginal blood flow in older menopausal women. Diabetes also causes arousal difficulties in women. Arousal problems may be induced by clitoral nerve degeneration, blood vessel damage, and hyperargentophilia of neural fibers in the clitoris found in diabetic women. Other medical problems that can affect female excitement are gynecologic cancer, chronic liver disease, and chronic renal disease.

Vaginal dryness occurs in women with MS where parts of the spinal cord are demyelinized. These women may also experience absent clitoral sensitivity. Lubrication does not occur in spinal cord injury, disease, and/or lesions affecting the spinal cord at T10–T12. Also, when lesions around T10–T12 exist, there are no clitoral sensations during arousal.

To date, the literature does not support the relationship of certain classes of medications (discussed earlier) and the absence of lubrication. However, it is plausible that to the degree that the neurophysiology of arousal is comparable in men and women, drugs affecting MED can have similar deleterious effects and result in FSAD. Ethinylestradiol hormone replacement is generally effective in increasing sexual excitement and vaginal lubrication specifically in menopausal women. Vaginal cream or oral administration can reverse the damaging effects of estrogen deficiency on the vagina but increases the risk of endometrial cancer. A transdermal estrogen patch may also be helpful in treating difficulties with vaginal dryness. Thus far, there are no medications that increase female arousal.

FSAD is assessed by use of a vaginal photometer, which measures vaginal pulse amplitude and vaginal pulse volume through a probe and photocell, cardiovascular autonomic nerve function tests monitoring heart rate and blood pressure during specific exercises; and a heated oxygen electrode placed on the vaginal wall, which uses an index of blood flow and content obtained through the energy required to maintain the heat setting. The lubricant fluid produced by the vasocongestive response can be evaluated for its quantity, electrolyte content, and pH. Also, assessment can include the study of changes in estrogen and serum gonadotropin levels.

Pubococcygeal (PC) muscle excercises, also known as Kegel excersises, are voluntary vaginal contractions originally developed to treat stress incontinence. Consistent use of these exercises has been shown to increase intravaginal pressure as well as to increase subjective reports of sexual arousal. These reports are preliminary but suggest that the prescription of Kegel exercises for women suffering from a sexual dysfunction may be appropriate.

2. Psychological Aspects of Arousal Disorders

Arousal is the result of physiological as well as cognitive and affective changes that indicate that the person is ready to engage in sexual activity. The cognitive component consists of focusing on erotic stimuli, including fantasies and sexual cues. The affective component includes a subjective sense of sexual excitement. Psychological factors contributing to Male Erectile Disorder may include negative affect, specifically, performance anxiety and depressed affect, as well as cognitive interference. Cognitive interference may consist of underestimating the degree of erectile response. Additionally, chronic or acute stress as well as relationship problems may play a central role in erectile problems.

Psychological factors contributing to Female Sexual Arousal Disorder may include anxiety, fear, relationship problems, lack of adequate stimulation, as well as a history of sexual abuse. For both males and females it is important to understand which of these factors is the primary contributing factor so as to know what to target in treatment. This may be accomplished by obtaining a detailed description of a recent problematic interaction, including thoughts and feelings during the interaction. This may provide the clinician with invaluable information concerning problem maintenance. How does each partner feel about the other partner? Was there adequate stimulation? Was either partner experiencing thoughts that interfered with arousal? Was either partner anxious or fearful?

To date, there are no controlled psychotherapy outcome studies to attest to the efficacy or effectiveness of treatment for females with arousal problems. For men with these problems, correction of cognitive distortions associated with sexual functioning,

systematic desensitization, and psychoeducation were found to be effective treatments when compared to a control condition.

IV. ORGASMIC DISORDERS

Whereas many sex therapists and researchers do not consider orgasm to be the epitome of a sexual experience, many people do. Problems regarding orgasm may influence the individual's self-esteem, mood, sexual satisfaction, quality of relationship with sexual partners, and desire to engage in sexual activity. These problems may include the absence of orgasm, orgasm only after prolonged stimulation, or reaching orgasm quickly with little stimulation. Epidemiological data suggest that orgasmic disorders affect approximately 24% of the population.

DSM-IV identifies three orgasmic disorders: 302.73 Female Orgasmic Disorder, 302.74 Male Orgasmic Disorder, and 302.75 Premature Ejaculation. According to DSM-IV, Female Orgasmic Disorder is characterized by "persistent or recurrent delay in, or absence of, orgasm following a normal sexual excitement phase." The clinician is instructed that women vary in the type or intensity of stimulation that leads to orgasm. Additionally, the clinician must use judgment to determine that the woman's capacity for orgasm is less than would be reasonable given her age, sexual experience, and the adequacy of the stimulation she receives.

According to DSM-IV, Male Orgasmic Disorder is characterized by a "persistent or recurrent delay in, or absence of, orgasm following a normal sexual excitement phase during sexual activity that the clinician, taking into account the person's age, judges to be adequate in focus, intensity and duration."

According to DSM-IV, Premature Ejaculation is characterized by "persistent or recurrent ejaculation with minimal sexual stimulation before, on, or shortly after penetration and before the person wishes it." Additionally, the clinician is instructed to take into account factors that affect the duration of the excitement phase, including age, novelty of the sexual partner or situation, and recent frequency of sexual activity.

A. Etiology and Treatment of Orgasmic Disorders

A person's ability to reach orgasm depends on the interplay of two primary factors: physiological and psychological. Physiologically, the orgasmic reflex consists of sensory components, motor components, and the sensation of orgasm. Psychologically, the ability to reach orgasm depends on the individual's cognitions pertaining to sex in general and achieving orgasm during the sexual encounter.

1. Medical Aspects of Orgasm Disorders

Little information is currently available regarding organic factors that contribute to orgasmic disorders. With few exceptions, laboratory procedures to confirm an organic basis for orgasm problems are unavailable. To date, protocols for the medical evaluation of orgasmic disorders have not been developed. Most of the information available concerns male orgasm difficulties. Much of the information that is currently thought to be true of female orgasm has been extrapolated from studies of male orgasm.

a. Neurophysiology of Male Orgasm Repeated tactile stimulation of the penis can evoke reflex vasodilatation of the penile vasculature, resulting in penile erection, seminal expulsion from contraction of the smooth muscle of the internal reproductive organs, glandular secretion, and rhythmic contraction of the striated musculature. This sequence is subject to inhibitory and/or excitatory influences from the brain. The ejaculatory reflex consists of emission and bladder neck closure, ejaculation, and the experience of orgasm.

Sensory impulses eliciting the ejaculatory reflex enter the sacral spinal cord via the pudendal nerve. The ejaculate is then emitted into the pelvic urethra. This is caused by contraction of the vas deferens, seminal vesicles, and smooth muscle of the prostate. Inflowing seminal fluid stimulates the urethral bulb, which elicits reflex closure of the bladder neck. This reflex closure prevents retrograde ejaculation. The motor fibers mediating emission are contained in the hypogastric nerve. These motor fibers consist of preganglionic fibers that arise from the thoracolumbar spinal cord, segment T12–L3. These preganglionic fibers synapse with short adrenergic nerves that lie close to the innervated organ.

Expulsion of the ejaculate through the external urinary meatus is called *ejaculation*. The ejaculate is propelled along the length of the urethra via rhythmic contractions of the urethral bulb and perineal musculature. Sensory efferents from the urethra travel in the pudendal nerve to the sacral cord, where they synapse with somatic efferents that travel in the pudendal nerve. This causes contraction of the ischio- and bulbocavernosus muscles.

Orgasm is a subjective sensory experience associated with the rhythmic contractions of ejaculation and

emission. Presumably, the sensation of orgasm is produced via tactile sensory impulses that travel to the thalamus and then to the limbic lobe.

b. Neurophysiology of Female Orgasm Female orgasm appears to be a genital reflex. The sensory component consists of tactile stimulation of the external genitalia, which travels via the pudendal nerve to spinal centers. Hormonal influences on brain centers that connect with spinal centers influence the threshold level required for orgasmic reflex. The motor component consists of efferent fibers from the T12–L1 area of the spinal cord, which travel via the hypogastric nerve to genital organs, including fallopian tubes and uterus, and which cause rhythmic contractions of these structures. Sensory fibers from these structures travel to the sacral cord via the pudendal nerve, where they synapse with motor fibers that cause contractions of the peritoneal muscles. The sensation of orgasm results when sensory stimuli from the contractions of the sexual organs travel to the spinal cord, thalamus, limbic lobes, and sensory cortex, which is presumed to produce the experience of orgasm.

c. Clinical Implications There are multiple points at which organic factors can cause orgasm disorders. When presented with an anorgasmic individual, it is important to conduct a biomedical evaluation to rule out organic factors. For example, it is important to assess for any injury to the cerebellum, including a stroke, tumor, or trauma. These injuries may result in decreased ability to reach orgasm. In addition, spinal cord injury, such as due to a tumor, trauma, multiple sclerosis, and tabes dorsalis, might result in orgasm disorders. It is easier to predict the sexual effects of a complete cord lesion than an incomplete lesion, given that in an incomplete lesion it is difficult to determine which tracts are spared. Spinal cord lesions are generally more destructive to ejaculatory function than erectile function in males. For females with spinal cord injury, relatively little information is available regarding its impact on sexual functioning. To date, it has been found that if the lesion is below T1, orgasm is not possible.

Additionally, damage to the peripheral nerves may negatively impact the ability to reach orgasm, such as due to pelvic surgery, trauma, and diabetes mellitus. For males, surgical procedures including retroperitoneal lymphadenectomy, protocolectomy, abdominoperitoneal resection, aortoiliac surgery, prostatectomy, and surgical procedures for bladder cancer can disrupt ejaculatory function. For females, less information is available regarding the impact of surgical procedures that interrupt the sympathetic innervation of the genital organs on a woman's ability to orgasm. Some surgical procedures that may impact females' ability to orgasm include proctocolectomy, abdominoperitoneal resection, pelvic exenteration, radical vulvectomy, and cystectomy. Many of these impact other aspects of sexual functioning as well.

Additionally, pharmacologic agents may impact an individual's ability to orgasm. For males, these primarily include antihypertensive and psychiatric drugs: lorazepam, reserpine, quanethidine, quanoxan, quanoclor, bethanidine, de brisoquine, phenoxybenzamine, phentolamine, thioridazine, chlorpromazine, clorprothexine, mezoridazine, fluphenazine, thiothixine, perphenazine, trifluoperazine, haloperidol, phenelzine, pargyline, mebanazine, imipramine, amitryptiline, chlomipramine, amoxapine, desipramine, chlordiazepoxide, and alprazolam.

Less is known regarding the effects of drugs on female anorgasmia. Most of the literature available concerns the effects of psychiatric drugs, including hetereocyclic and monoamine oxidase inhibitors. Anorgasmia has been associated with phenelzine, isocarboxazid, tranylcypromine, imipramine, chlomipramine, amoxapine, nortriptyline, thioridazine, trifluoperazine, fluphenazine, alprazolam, and diazepam.

Endocrinopathy, including hypogonadism and hyperprolactinemia, should be evaluated in males with orgasmic dysfunction. The removal of androgens results in the decline of sexual interest and activity. In addition, ejaculation and the capacity to orgasm cease. With the use of exogenous androgens in hypogonadal men, sexual functioning is restored. There appears to be a minimum level of androgen necessary to sustain sexual functioning. Additional androgen beyond this level does not impact male sexual functioning. Studies have demonstrated the association between hyperprolactinemia and severe difficulty ejaculating. When this condition was effectively treated, sexual functioning was restored. For men complaining of ejaculatory failure, routine serum prolactin and testosterone levels should be obtained. For females, the role of androgens and hyperprolactinemia in anorgasmia is currently unclear.

2. Psychological Aspects of Orgasm Disorders

The inability to reach orgasm or reaching orgasm prematurely may be related to the individual's thoughts and/or feelings during the sexual encounter. Negative feelings toward sex, oneself, or the sexual

partner, fear of losing control, fear of pregnancy, performance anxiety, thoughts of a previous trauma (such as a sexual trauma), and discomfort with increased vulnerability may negatively impact sexual functioning, including orgasm problems. It is important to assess the thoughts and feelings the client experiences during these encounters so as to determine what is interfering with the progression of the sexual experience.

It is possible that the individual has unrealistic expectations or goals related to sex, such as the expectation that multiple orgasms occur in each sexual interaction. If this is found to be the case, it is important to educate the individual regarding reasonable expectations. If a previous trauma appears to be influencing the individual's sexual experiences, it is important to treat the person's trauma first, the success of which may indirectly influence the orgasm problem. In addition, the individual's relationship with his or her partner may influence their sexual interactions. If this is found to be true, it is appropriate to address relationship problems prior to addressing the sexual relationship.

If the person's expectations are reasonable, there is no trauma influencing the sexual experience, and the relationship is not problematic, directed masturbation has been demonstrated to be effective in treating women with orgasmic disorder. This treatment can be effective if the woman and her partner consider masturbation to be an acceptable form of treatment. The book *Becoming Orgasmic* delineates a program for women to learn to masturbate and become orgasmic.

For men with orgasmic disorder or premature ejaculation, performance anxiety may play a central role in problem maintenance. If during the assessment phase he describes thoughts indicating anxiety related to reaching orgasm, it is important to attend to these thoughts in treatment. These thoughts may inhibit him from reaching orgasm or they may result in his reaching orgasm prematurely. Treatment may focus on educating him that orgasm is not the goal and endpoint of sex. Instead, the focus is shifted to sensual pleasure as related to arousing other parts of the body, which Masters and Johnson termed sensate focus.

Additionally, premature ejaculation may be treated by teaching him to control ejaculatory latency by means of the "squeeze" technique or the "pause" technique. The squeeze technique involves instructing him to masturbate to a point that he feels would result in ejaculation if he continued and squeeze the head of his penis for approximately 10 sec. This is continued a number of times over several occasions until he gains control over his ejaculatory latency. The same is true of the pause technique, whereby he pauses prior to ejaculation and then resumes masturbation.

V. SEXUAL PAIN DISORDERS

Sexual pain may interrupt sexual activity, reduce future desire for sexual activity, prevent sexual satisfaction, result in relationship problems due to problematic sexual experiences, and influence the individual's mood and self-image. Problems concerning sexual pain include male or female genital pain that occurs before, during, or after sex as well as the involuntary spasm of the pelvic muscles for females that interferes with penetration. Epidemiological data suggest that 7% of the nation suffer from these problems. DSM-IV identifies two sexual pain disorders: 302.76 Dyspareunia and 306.51 Vaginismus. According to DSM-IV, Dyspareunia is characterized by "recurrent or persistent genital pain associated with sexual intercourse in either a male or a female."

According to DSM-IV, Vaginismus is characterized by "recurrent or persistent involuntary spasm of the musculature of the outer third of the vagina that interferes with sexual intercourse." If an individual meets criteria for either of these disorders, it is important to determine the onset (lifelong or acquired), context (generalized or situational), and etiological factors (due to psychological factors or due to combined factors) of the disorder. Vaginismus and Dyspareunia are often linked. Whereas both of these disorders pertain to genital pain, the distinction between them is that Dyspareunia is pain associated with intercourse and Vaginismus is pain as a result of an inability to penetrate. If Dyspareunia is caused by Vaginismus, the primary diagnosis is Vaginismus. As will be discussed later, a gynecological or urological evaluation must be conducted in order to adequately diagnose the problem.

A. Etiology and Treatment of Pain Disorders

As is true of other sexual disorders, the sexual pain disorders may result from physiological and/or psychological problems. It is important to determine which of these factors is contributing to the problem.

1. Medical Aspects of Pain Disorders

Individuals who experience pain during intercourse are more likely to present to gynecologists and urologists for symptom relief. The sexual pain

disorders are the most common sexual problems that physicians encounter, possibly because individuals view these more as medical problems.

a. Neurophysiology of Pelvic Pain The spinal segments responsible for the somatic nerve supply to the groin, vulva, and perineum are predominantly L1–L2, whose fibers form the ilioinguinal and iliohypogastric nerves, and S2–S4, whose fibers form the pudendal nerve. These nerves provide sensory innervation to the skin of the groin, mons pubis, and anterior vulva as well as the clitoris, perineum, and posterior labia. Additionally, the pelvic organs send sensory impulses via afferent fibers, both sympathetic and parasympathetic.

Pain resulting from Vaginismus is due to an involuntary spasm response of the pelvic muscles located in the outer third of the vaginal barrel. The muscle groups affected include the perineal muscles and the levator ani muscles. In several cases, the adductors of the thighs, the rectus abdominis, and the gluteus muscles may be involved. This spasm response prevents penetration and leads to pain when intercourse is attempted.

b. Clinical Implications For males, genital pain may be associated with a urinary tract infection. The pain may occur during urination as well as ejaculation. Men who are not circumcised may also describe pain during intercourse. Additionally, Peyronie's disease is sometimes associated with pain during erection. As such, it is essential that men with genital pain be referred to a urologist for a thorough examination.

Scarring following either episiotomy or vaginal repair operations may lead to painful intercourse. This includes hymeneal remnants, a pelvic tumor, prolapsed ovaries, endometriosis, childbirth pathologies, and pelvic inflammatory disease. Other possible organic influences include clitoral inflammation and adhesions, lesions of the vulva, vaginal infection, cystitis, constipation, proctitis, and allergic reactions to contraceptive and douching materials. Additionally, vulvar vestibulitis, which consists of tiny erythematous sores in the vulvar vestibule, may lead to intense superficial pain. Women may also experience pain as a result of insufficient vaginal lubrication upon penetration attempts. The risk of genital pain for this reason is increased in women using oral contraceptives, pregnant women, or premenopausal women.

c. Definition of the Pain It is important to understand the nature of the pain when assessing these sexual problems. For example, it is important to understand the quality of the pain, including whether the pain is sharp or dull, diffuse or localized, burning or itchy, superficial or deep. Superficial pain tends to be sharp, burning, and well-localizable. Deep somatic or visceral pain is usually dull, diffuse, and poorly localizable. The severity of the problem is also important. How intense is the pain? To what extent does the pain interfere with sexual activity? The location of the pain should be assessed as well. A physical examination may help determine whether the location conforms to a dermatome, to the distribution of a specific peripheral nerve, or to the position of a specific anatomic structure. The duration and periodicity of the pain should be discussed in detail. Is the pain always present? Are there times when the pain is gone? Responses to these questions may provide treatment utility. For example, anticonvulsant agents may be more effective for pain syndromes that are paroxysmal, whereas antidepressants may be more effective for pain syndromes characterized by constant burning. Severe, but short-lived pain evoked by mechanical manipulations such as vaginal intercourse with little or no pain between episodes will respond poorly to nonspecific treatments.

2. Psychological Aspects of Sexual Pain Disorders

Psychosocial problems may result in a complaint of sexual pain or may exacerbate existing pain. Possible risk factors include anxiety, fear of pregnancy, fear of contracting a sexually transmitted disease, depression, decreased self-esteem, poor body image, and religious orthodoxy. If this is the case, it is important to address the issue. For example, anxiety may be alleviated by using sensate focus. Poor body image may be targeted through psychoeducation. Negative attitudes about sex may be overcome by reviewing and processing these attitudes. Additionally, relationship problems should be considered when assessing these problems. Anger at one's partner, inadequate communication, and distrust should be considered. If this is found to be true, it is appropriate to initiate couples therapy if the couple agrees. Sexual abuse or trauma may also result in a pain disorder. If this is the case, it is important to treat the trauma prior to treating the sexual problem, which may alleviate the sexual problem indirectly.

Psychogenic Vaginismus has been explained by classical conditioning. When first attempting intercourse, a woman may experience significant pain. This

pain, the unconditioned stimulus, may lead to a natural self-protective tightening of the vaginal muscles, the unconditioned response. With repeated attempts, stimuli associated with vaginal penetration or even the thought of intercourse can become conditioned stimuli and lead to the conditioned response of reflexive muscle spasms. In the process, the woman may avoid intercourse in order to avoid the pain associated with sex.

A commonly used treatment for women experiencing these problems is desensitization using a graduated set of dilators or the gradual insertion of the woman's fingers. This treatment begins with sensate focus, whereby the client is instructed to focus on pleasurable bodily sensations. This exercise is expected to reduce the client's anxiety and underscore the notion that sexual pleasure can be achieved in many ways other than penetration. Once the client is comfortable with her body, the gradual insertion of her fingers or of graduated dilators is initiated. The exercise begins with insertion of the smallest finger or dilator, usually with the use of a lubricant. The client is instructed to control the depth of penetration and length of time of penetration. Once the client is comfortable with inserting the dilator or finger for 5 min, the next larger size should be inserted, again while the client maintains control over insertion. As she becomes comfortable with insertion, her partner can be included in the exercise. She should maintain complete control throughout, including control over the depth and length of time of penetration. It is expected that this sense of control will help alleviate the problem.

See Also the Following Articles

AROUSAL • DOPAMINE • HYPOTHALAMUS • SEX DIFFERENCES IN THE HUMAN BRAIN • SEXUAL BEHAVIOR • SEXUAL DIFFERENTIATION, HORMONES AND • SEXUAL FUNCTION

Suggested Reading

Bachmann, G. A., Leiblum, S. R., Sandler, B., Ainsley, W., Narcessian, R., Sheldon, R., and Nakajima Hymans, H. (1985). Correlates of sexual desire in post-menopausal women. *Maturitas* **7,** 211–216.

Becker, J. V., Skinner, L. J., Abel, G. G., and Cichon, J. (1986). Level of postassault sexual functioning in rape and incest victims. *Arch. Sex. Behav.* **15**(1), 37–49.

Becker, J. V., Skinner, L. J., Abel, G. G., and Treacy, E. C. (1982). Incidence and types of sexual dysfunctions in rape and incest victims. *J. Sex Marital Ther.* **8**(1), 65–74.

Hawton, K. (1983). Behavioural treatment of sexual dysfunction. *Br. J. Psychiatry* **140,** 94–101.

Hawton, K. (1985). *Sex Therapy: A Practical Guide*. Oxford University Press, Oxford, UK.

Heiman, J., and LoPiccolo, J. (1988). *Becoming Orgasmic: A Sexual and Personal Growth Program for Women*, rev. ed. Prentice Hall, New York.

Hurlbert, D., White, L., Powell, R., and Apt, C. (1993). Orgasm consistency training in the treatment of women reporting hypoactive sexual desire: An outcome comparison of women-only groups and couples-only groups. *J. Behav. Ther. Exp. Psychiatry* **24,** 3–13.

Katz, R. C., Gipson, M. T., Kearl, A., and Kriskovich, M. (1989). Assessing sexual aversion in college students: The Sexual Aversion Scale. *J. Sex Marital Ther.* **15,** 135–140.

Laumann, E. O., Paik, A., and Rosen, R. C. (1999). Sexual dysfunction in the United States: Prevalence and predictors. *J. Am. Med. Assoc.* **281,** 537–544.

Leiblum, S. R., and Rosen, R. C. Introduction: Changing perspectives on sexual desire. In *Sexual Desire Disorders* (S. R. Leiblum and R. C. Rosen, Eds.), Guilford Press, New York.

Letourneau, E., and O'Donohue, W. (1993). Sexual desire disorders. In *Handbook of Sexual Dysfunctions* (W. O'Donohue and J. H. Geer, Eds.), Allyn & Bacon, Needham Heights, MA.

O'Donohue, W., Dopke, C., and Swingen, D. (1997). Psychotherapy for female sexual dysfunction: A review. *Clin. Psychol. Rev.* **17,** 537–566.

Rosen, R. C., and Leiblum, S. R. (1988). A sexual scripting approach to problems of desire. In *Sexual Desire Disorders* (S. R. Leiblum and R. C. Rosen, Eds.), Guilford Press, New York.

Segraves, R. T., and Schoeberg, H. W. (Eds.). (1985). *Diagnosis and Treatment of Erectile Disturbances*. Plenum, New York.

Screiner-Engel, P., and Schiavi, R. C. (1986). Life psychopathology in individuals with low sexual desire. *J. Nervous Mental Dis.* **174,** 646–651.

Sexual Function

MARK S. GEORGE and JEFFREY P. LORBERBAUM
Medical University of South Carolina and Ralph H. Johnson Veterans Hospital

GLOSSARY

ejaculation When the penis expulses semen.

genital tumescence (penile erection or clitoral enlargement) When the penis or clitoris fills with blood and becomes rigid or elevated.

limbic system Anatomically, the limbic system refers to a large cerebral convolution that lies medially and envelops the brain stem. Although there is no clear consensus, the following regions are generally considered part of the limbic system. Cortical structures include the cingulate gyrus, subcallosal gyrus, hippocampus, and olfactory cortex. Subcortical regions include the amygdala, septum pellucidum, epithalamus (habenula), anterior thalamic nuclei, hypothalamus, and parts of the basal ganglia. In addition, several closely linked cortical structures that appear important in emotional behavior are also considered part of this circuit and are often referred to as paralimbic. These regions include the anterior temporal polar cortex, medial–posterior orbitofrontal cortex, and insular cortex. Evolutionarily, limbic structures appear to be common to all mammals. Behaviorally, limbic structures are thought to be important in emotional behavior, although this point is strongly debated.

sexual arousal Pertaining to the emotional and cognitive state in which one is interested in having sexual relations with another. Arousal is difficult to measure or quantify in animals. Courtship displays and genital tumescence are likely the most objective markers of arousal, although these behaviors can also occur in the absence of sexual arousal–interest.

sexual function A complex series of orchestrated behaviors consisting of courtship, arousal, intercourse, ejaculation, and recovery.

social bonding Affiliative behaviors that promote proximity and contact between individuals and can be divided into several types: (1) parental behavior, such as nursing and retrieving offspring; (2) infant attachment behavior, such as clinging and signals of protest or despair (i.e., crying) on separation; and (3) adult–adult affiliative behaviors, such as playing, grooming, a propensity to be near another much of the time, and monogamous pair bonding.

Normal human sexual function involves a complex series of orchestrated behaviors consisting of arousal, intercourse, ejaculation, and recovery. The regions and circuits involved in these behaviors are becoming known from analogous studies in animals, from epilepsy and brain lesion case studies in humans, and, most recently, from human functional neuroimaging studies. Importantly, however, sexual function commonly involves the selection and courting of a mate, which can result in forming a strong emotional bond with the other person, sometimes producing the birth of children. Thus, the sexual act in humans is inextricably linked with brain systems involved in other appetitive behaviors and in the formation of bonding and social attachment (or "love"). Thus, sexual function sets the stage for procreation and the birth of children. The maternal compulsion to answer the needs of infants might be regarded as representing the germ of responsibility that in human beings generalizes to become what we call "conscience."

Therefore, for most people, sex is linked with love and love is often intermingled with sex. In this article, we review the functional neuroanatomy involved in the basic behaviors of human sexual function and then discuss brain regions and neurochemicals involved in more complex aspects of sexual function, such as courtship and pair bonding. Human sexual function thus involves the entire nervous system—from

peripheral nerves in end sexual organs, through the spinal cord and brain stem for some aspects of copulation, to limbic regions involved in pleasure, orgasm, bonding, and love.

The human brain mediates a multitude of complex behaviors, and none is more important from an evolutionary standpoint than the behaviors associated with sexual function and reproduction. Some researchers in the field of sociobiology have argued that our bodies and lives are nothing but the mechanisms by which our genes replicate and extend themselves forward in time. Viewed from this somewhat radical and distorted perspective, our brains therefore are nothing more than the complex organs that orchestrate normal sexual function. However, this view from a purely reproduction standpoint does little justice to much of the complex higher cortical functioning of our brains and how this more complex functioning relates to sexual function.

I. BASIC PRINCIPLES OF BRAIN ORGANIZATION

Different components of sexual function involve distinct brain regions. Before we address the specific aspects of sexual function and the relevant neuroanatomy, it is important to outline a few basic principles of brain organization that will be referred to throughout the article.

P. D. MacLean introduced the concept of *the triune brain*, an attractive heuristic to organize thinking about the neural mechanisms involved in sexual function (both sex and love). The triune brain consists of three concentric brain systems, resting on top of each other yet intermeshed, with each successive layer representing an evolutionary advance (see Fig. 1). Each advance simultaneously allows for new behaviors, while recapitulating older ones. Thus, the human brain can be thought of as representing a composite of three brains in one (i.e., the triune brain). The deepest layer, found in reptiles, is the reptilian brain or basal ganglia (olfactory tubercle, nucleus accumbens, caudate, putamen, and globus pallidus). This part of the brain allows for simple actions (alertness, eating, basic acts of procreating) in preprogrammed rigid routines, which lack flexibility and emotion. The next layer (paleomammalian) arose with early mammals and corresponds to the great limbic lobe of Broca and connected brain stem structures. This extension is only rudimentary in reptiles and birds. The final part of the human triune brain is the neomammalian brain, comprising of the neocortex and connected brain stem structures. These structures are thought to have an important part in organizing behaviors that are unique to mammals, such as nursing, care of the young, bonding, and play.

The term "limbic" requires more elaboration. In 1878, Pierre Paul Broca first coined the term limbic,

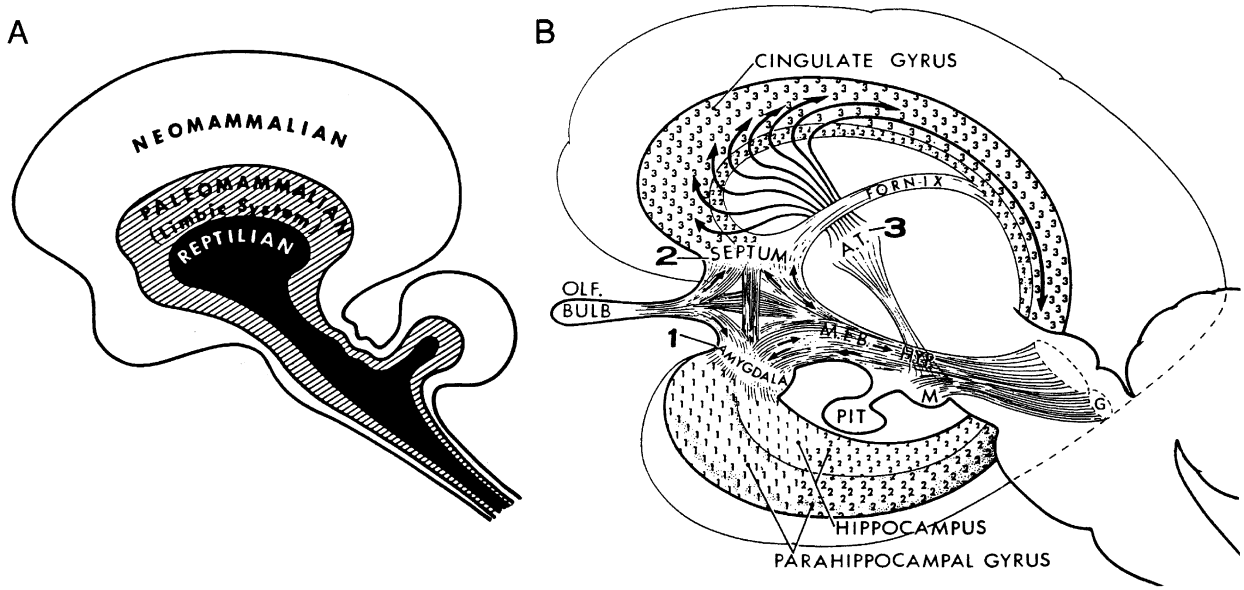


Figure 1 (A) The triune brain concept, from MacLean (1968). Alternative neural pathways to violence. In *Alternatives to violence*. (L. Ng, ed.), pp. 24–34, Time-Life Books, New York. (B) The three main subdivisions of the limbic system, from MacLean, P.D. (1990). *The triune brain in evolution: Role in paleocerebral functions*. Plenum Press, New York.

using it to refer to the structures forming a frame around the brain stem that he posited were important in regulating emotion (see Fig. 1). Early in the twentieth century, J. W. Papez detailed the circuit that he thought was important in regulating emotion. Thereafter, MacLean further elaborated the notion of a limbic brain involved in recognizing and regulating emotion and other complex aspects of behavior within the larger concept of the triune brain.

An important part of this heuristic scheme involves *the notion of the recapitulation of function within the brain*. Just as the motor homunculus is represented at the level of the spinal cord, cerebellum, thalamus, basal ganglia, and cortex, the brain structures involved in sexual function are also represented in all three components of the triune brain. Thus, the triune brain concept would be consistent with the notion that direct damage to some reptilian brain structures, such as the caudate, is associated with defects in basic sexual function. More importantly, damage to the early mammalian brain (limbic system) would more likely derange more complex aspects of sexual function, such as bonding. Mammals, in contrast to lizards, engage in three family-related behaviors: nursing and parental care, vocal communication with offspring, and play. Thus, phylogenetically, the newly emergent limbic system would be more important in producing and regulating these characteristic mammalian behaviors that involve recognizing and displaying affect in order to build social relationships. Finally, direct damage to the neocortex, which is greatly expanded in humans, might also disrupt sexual function, largely through the interruption of frontal cortical regulation of the limbic and reptilian brains.

The limbic system can be further subdivided into three divisions: the (1) amygdalar, (2) septal, and (3) thalamocingulate divisions. *Amygdalar division*: This region is involved in self-preservation behaviors such as those required in the search for food, including fighting and self-defense. Stimulation of this area in humans may produce fear and anxiety. *Septal division*: This division likely subserves behaviors related to primal sexual function and procreation. Septal stimulation in humans can produce pleasurable sensations and in animals can elicit social grooming as well as genital tumescence. *Thalamocingulate division*: This region represents the phylogenetically newest subdivision of the limbic system and is present in mammals, but not in reptiles even in rudimentary form. Several typical mammalian social behaviors are associated with this area, including extensive mother–infant bonding, pair bonding, infant crying, and play. Lesions of this region in nonhuman mammals often produce social apathy and mothers will neglect their offspring.

II. REGIONAL BRAIN BASIS OF SEXUAL INTERCOURSE

An attractive way to approach the neuroanatomy of sexual function is to divide the sexual act into its discrete phases and discuss the relevant neuroanatomy by each phase. This approach has some value, although many of the stimulation studies in animals have shown effects on several aspects of sexual function, arguing against their anatomic discreteness. As a way of organizing and discussing each aspect of sexual function, we will interleave information from animal stimulation studies with human neuroimaging and lesion data when available.

A. Arousal

The first part of the human sexual act is perhaps the most difficult to measure or quantify in animals. Stimulation of several regions, including the hippocampus, is associated with courtship displays or genital tumescence, which are likely objective markers of arousal. However, in humans, erections and genital tumescence can occur in the absence of pleasure or arousal as in states of priapism.

In humans, arousal states can now be studied directly using *functional neuroimaging*. Researchers have used functional neuroimaging to investigate which brain regions are activated during sexual arousal. In 1999, Stoleru and colleagues used oxygen positron emission tomography (PET) to image brain activity in eight healthy adult men while they viewed sexually explicit film clips (heterosexual coitus), emotionally neutral clips, and humorous control clips. During sexual arousal compared with the neutral resting state, the men had significantly more blood flow in the inferior temporal cortex (a visual association area), the right insula and inferior frontal cortex, and the left anterior cingulate. These researchers also serially measured plasma testosterone and found that activity in several of these regions increased in proportion to increased testosterone.

The insula and inferior frontal cortex are limbic regions that relate processed sensory information with motivational states. Several authors have found

a right-sided bias for processing and mediating emotional stimuli. As outlined earlier, the cingulate gyrus is an important region in which four forms of behavior are represented that distinguish the evolutionary transition from reptiles to mammals (milk secretion, infant crying, extensive maternal behavior, and play). The cingulate gyrus similarly has been implicated in studies of directed attention, particularly when there is competition for attention or divided attention.

By using a methodology similar to that of the Stoleru study, Rauch and colleagues studied brain activity in eight healthy men during script-driven imagery to produce sexual arousal (response to an autobiographical script of thinking about a sexually arousing experience) or competitive arousal (response to an autobiographical script of thinking about success in a competitive arena such as winning a sports event). In both of the aroused states compared with rest, blood flow increased in the anterior cingulate and the anterior temporal cortex, as well as in the ventral globus pallidus. The brain stem was activated uniquely in the sexually aroused state. Unfortunately, the PET camera used in this study was not able to resolve nuclei within the brain stem.

Thus, the anterior cingulate gyrus was activated in both of these PET studies, which attempted to measure brain correlates of arousal using two different stimulus forms (silent visual film clips or script-driven imagery of past experiences). In the second study, the globus pallidus also was activated, which is interesting in that this region of the R complex has been implicated in courtship displays demonstrating interest in a potential mate during experiments on nonhuman mammals (see Section IV.A on courtship). Further studies are needed before a consensus can emerge about the functional neuroanatomy of human sexual arousal. It also remains unclear which method (visual or, script-driven) is best for eliciting and measuring arousal during functional imaging studies.

B. Erection, Vasodilatation, and Intercourse

Brain pathways involved in genital tumescence have been studied mostly in males where it is evidenced by penile erection, although genital tumescence (clitoral enlargement) in females appears to involve the same pathways. Except for the posterior medial prefrontal cortex (an area that may represent a transition zone between limbic and neocortex), electrical stimulation of the neocortex generally has not been effective in eliciting genital tumescence. Within the limbic system, penile erection appears to be primarily represented in

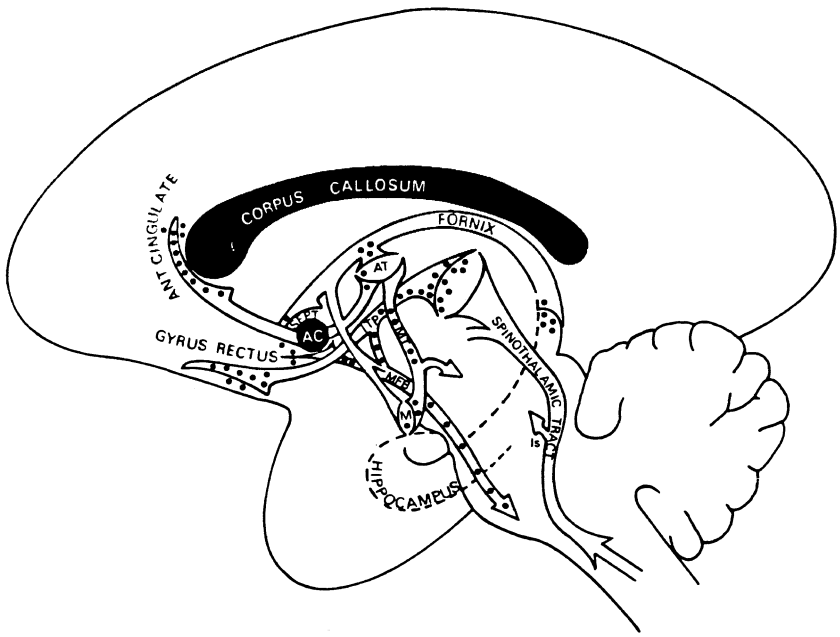


Figure 2 A sagittal diagram of brain regions summarizing regions in the squirrel monkey where electrical stimulation is highly effective in eliciting erections [after MacLean (1990). *The triune brain in evolution: Role in paleocerebral functions*. Plenum Press, New York].

the septal and thalamocingulate subdivisions, which are involved mainly in procreation and social bonding behaviors. This is an interesting finding given the intertwining of primary sexual function and social bonding in mammals. The pathways for penile erection are summarized in Fig. 2.

Regarding pathways from the septal subdivision, penile erection occurs with stimulation of *the medial septal region or its contiguous medial preoptic region of the hypothalamus*, as well as along this region's efferent white matter tract called the *median forebrain bundle* to the ventral tegmental area in the midbrain. The other major efferent white matter tract pathway from the medial septal preoptic region is the periventricular fiber system. This does not appear to be a major pathway for penile erection. Another area that consistently produces penile erection with electrical stimulation is the *dorsomedial hypothalamus*, which is bounded rostroventrally by the fornix and dorsocaudally by the mammilothalamic tract. The dorsomedial hypothalamus may also connect to the median forebrain bundle. Finally, electrical stimulation of the *proximoseptal hippocampus* also appears to be effective in eliciting penile erection. Regarding the septal subdivision, it is also of interest that electrical stimulation of the septum produces pleasurable, erotic sensations in humans. In male cats, septal and proximal hippocampal electrical or chemical stimulation produces pleasurable reactions, such as loud purring, turning over on its back, rubbing against many objects, and an invitation to be stroked by an observing experimenter while the cat gently bites and licks the experimenter's hand. Further, erections frequently occur in response to septal stimulation as well as self-grooming—behaviors that are found in feline courtship.

Regarding pathways from the *thalamocingulate subdivision*, penile erection is elicited by stimulation of the *anterior thalamic nuclei* and the *dorsomedial nucleus of the thalamus*, as well as by *stimulation of the anterior cingulate–posterior medial prefrontal cortex*, regions to which these nuclei project (i.e., the anterior cingulate and posterior gyrus rectus, respectively). Further, electrical stimulation along the entire length of the mammilothalamic tract from the anterior thalamic nuclei to the mamillary bodies elicits penile erection in squirrel monkeys. It is unclear how this pathway for penile erection enters the midbrain or brain stem, but it does not appear to depend on the median forebrain bundle (MFB). Also, stimulation of the dorsomedial nucleus of the thalamus and along its white matter projection, the inferior thalamic peduncle, elicits penile erection at least to the level at which this tract connects to the MFB.

At the level of the midbrain, the major effector pathway for penile erection follows the MFB to the ventral tegmental area (VTA). From this level, the pathways for penile erection quickly turn laterally, touching the dorsal aspect of the substantia nigra before descending to the pons (see Fig. 3). In the pons, the pathway descends through the region of the lateral tegmental process. In this region, there are positive loci for penile erection in the brachium pontis, providing possible evidence of cerebellar mechanisms in penile erection. MacLean and colleagues did not explore the erection pathways below the level of the upper medulla where the erection pathways enter the medulla just lateral to the pyramids and just medial to cranial nerve VI.

C. Ejaculation and Orgasm

Stimulation in monkeys along the spinothalamic tract and its projections to the thalamus produces genital scratching and seminal discharge. Although these behaviors are sometimes linked, detailed studies have shown that MD intralaminar thalamus stimulation can produce seminal discharge without preceding genital scratching or even penile erection. Stimulation near this spot might also elicit other visceral-related effects, such as salivation, vomiting, urination, or defecation.

In 1994, Tiihonen and colleagues used perfusion SPECT to try to image the brain regions activated during ejaculation. They studied adult men and imaged brain activity around the moment of ejaculation. With perfusion SPECT, one can inject the radiotracer during a state with a great deal of movement (like an ejaculation or a seizure) in someone outside of the scanner and then actually develop the brain image later. The SPECT radiotracer injected through a vein travels to the brain within several seconds and then undergoes metabolic changes and rests in those neurons that were most active at the time of injection. The technetium tracer continues to emit photons for several hours thereafter, allowing later positioning in the camera. This research group, with low spatial resolution, found increased activation during orgasm in the right prefrontal cortex. The crude spatial and temporal resolution of this technique precludes real comparison with the direct stimulation studies in animals noted earlier.

D. Postintercourse, Recovery, or Refractory Phase

Very little is known about the central brain mechanisms involved in postsexual recovery. The new popular medication for the treatment of male impotence, sildenafil citrate or Viagra, works by inhibiting this recovery phase. Sildenafil is a selective inhibitor of cyclic guanosine monophosphate (cGMP) specific phosphodiesterase type 5. The physiologic mechanism of erection of the penis involves the release of nitric oxide (NO) in the corpus cavernosum during sexual stimulation. NO then activates the enzyme guanylate cyclase, which results in increased levels of cyclic guanosine monophosphate (cGMP), producing smooth muscle relaxation in the corpus cavernosum and allowing the inflow of blood. Sildenafil has no direct relaxant effect on isolated human corpus cavernosum, but it enhances the effect of nitric oxide (NO) by inhibiting phosphodiesterase type 5 (PDE5), which is responsible for the degradation of cGMP in the corpus cavernosum. When sexual stimulation causes local release of NO, inhibition of PDE5 by sildenafil causes increased levels of cGMP in the corpus cavernosum, resulting in smooth muscle relaxation and the inflow of blood to the corpus cavernosum. Sildenafil at recommended doses has no effect in the absence of sexual stimulation.

III. MORE COMPLEX ASPECTS OF SEXUAL FUNCTION

A. Courtship Behavior

In MacLean's triune brain model, the reptilian complex (striatal complex) is important for ritualistic display behavior, such as challenge and courtship displays. In common green anolian lizards, partial lesions of the paleostriatum (considered by many to be the counterpart of the striatum and globus pallidus) greatly impairs the challenge display, a display similar to their courtship display in several respects. In contrast, lesions of the overlying forebrain (including an area believed to be a counterpart of the amygdala) do not inhibit courtship displays. Further, MacLean and colleagues conducted extensive experiments on squirrel monkeys to understand which regions may be responsible for challenge and courtship displays. To accomplish this, these researchers studied the *mirror display*, which is a highly predictable variation of a naturally occurring display used by male gothic-type squirrel monkeys in a show of aggression, in courtship, and in greeting. Vocalization, thigh spreading, and forward thrusts of the erect phallus are the major and most regularly occurring manifestations of the 5 features of the display. MacLean found that, in a series of lesion experiments involving more than 100 monkeys, the medial part of the globus pallidus in the R complex is a main site of coordination for evocation of the mirror display. Lesion of the medial part of the globus pallidus (especially its rostral two-thirds) resulted in either a sustained loss, decreased occurrence, or fragmentation of the mirror display. On the basis of globus pallidus lesions, it appears that the critical outputs for the display arise from the rostral two-thirds of the medial segment of the globus pallidus. Regarding outputs from the globus pallidus, destruction of the transhypothalamic pathways results in a lasting interference with the display. This is not seen with destruction of the other major pathways associated with the hypothalamus. Further, globus pallidus projections to the thalamus and midbrain tegmentum were more important than those leading to the ventral thalamus for enduring performance of the display. A wide variety of lesions, including lesions of limbic and neocortical structures, had no effect on the expression of this coordinated display.

Whereas the preceding lesion data indicate structures that appear to be important for the coordinated performance of the mirror display, specific pathways are important in individual aspects of the display: penile erection, vocalization, and thigh spread. When these other pathways are interrupted, the display becomes fragmented in that a specific element of the display such as erection is not well-performed, although the attempt to make the display as a whole still exists. In this regard, lesions destroying the inferior thalamic peduncle as well as those destroying the MFB result in a decreased magnitude of penile erection (entirely consistent with the stimulation studies discussed earlier). Decreased penile erection also occurred with lesions of the medial preoptic nucleus and adjacent anterior hypothalamus as well as lesions of the dorsomedial hypothalamus. Regarding vocalization, lesions in the tegmentum (at or near the origin of the medial longitudinal fasciculus) appeared to decrease the display vocalization.

It is unclear how homologous the squirrel monkey display is to courtship displays in humans, which may be different and depend on language and many other aspects of social context. However, in 1970 while filming a number of societies around the world,

Eibl-Eibesfeldt, a German ethologist, found what he thought was a universal flirtation display in humans. In this display, one's gaze alternates from the face of the other person to the side and down. This may be a classic example of an approach–avoidant display.

B. Affiliative Behavior

Affiliative behaviors promote proximity and contact between individuals and can be divided into several types: (1) parental behavior, such as nursing and retrieving offspring; (2) infant attachment behavior, such as clinging and signals of protest or despair (i.e., crying) on separation; and (3) adult–adult bonding behaviors, such as playing, grooming, a propensity to be near another much of the time, and monogamous pair bonding.

In MacLean's triune brain model, the brain region viewed to be central in affiliative behaviors is the limbic system's thalamocingulate division, which includes the cingulate gyrus and anterior thalamic nuclei. These are the phylogenetically newest part of the limbic system and are present in mammals but not in lizardlike reptiles. Extensive maternal behavior, infant crying, and play are three very basic affiliative behaviors that are found in mammals but not in lizardlike reptiles. The thalamocingulate gyrus may be an important brain region for all three of these behaviors.

1. Thalamocingulate's Role in Infant Crying

When separated, all mammalian infants that have been studied emit a species-specific isolation call that is an affiliative behavior because it causes mammals to search and care for the infant. Early mammals are believed to have been nocturnal creatures living on the forest floor. Infant separation from a parent therefore would be catastrophic and audiovocal communication would be important for reestablishing contact in the dark. Lesions of the anterior cingulate and the immediately adjacent medial prefrontal cortex (which in turn cause retrograde degeneration of anterior thalamic nuclei) impair crying upon separation in infant and adult squirrel monkeys, but more anterior prefrontal lesions do not impair crying. Structures strongly connected to the thalamocingulate circuit near the thalamic–midbrain junction (i.e., the thalamic tegmentum, thalamic periventricular gray matter, tegmentum, and the midbrain's central gray matter) are important as well not only in the actual production of the isolation call but also in forming its distinctive vocal characteristics.

2. Thalamocingulate's Role in Maternal Behavior

Animal studies and preliminary results in humans suggest that the brain's cingulate and anterior thalamic nuclei (a region we collectively refer to as the thalamocingulate circuit) may be involved in mammalian maternal behavior. First, cingulate lesions impair maternal behavior in rodents. Impaired behaviors include (1) nest-building, (2) adopting the correct nursing position, (3) retrieving pups when separated, and (4) sustaining pups through the time of weaning. Interestingly, female hamsters deprived of a neocortex soon after birth have seemingly normal maternal behavior, whereas those also deprived of their cingulate gyrus and underlying dorsal part of the hippocampus have impaired maternal behavior. This suggests that, in MacLean's triune brain model, the thalamocingulate circuit is necessary for maternal behavior. This experiment also suggests that both an intact reptilian brain and a paleomammalian brain are sufficient for normal maternal behavior in the rodent. Second, the anterior cingulate cortex is abundant in both opiate and oxytocin receptors, which are thought to be involved in social bonding. In fact, electrical stimulation of the anterior cingulate in female rabbits and cats can cause oxytocin release, milk ejection, and uterine contractions. Third, two functional MRI studies in humans by our group suggest that the thalamocingulate circuit activates when human mothers listen to recorded, standardized infant cries. This is a potentially important measure of maternal behavior because mammalian mothers usually respond to infant crying with care-taking behavior. In fact, in humans, the more responsive a human mother is to infant crying, the more likely she is to be responsive to other infant cues, and the more likely her child will become securely attached. Thus, examination of a mother's brain activity when she hears crying infants is potentially one way to measure the brain basis of maternal behavior.

Specifically, in 1998, Lorberbaum and colleagues measured regional brain activity in healthy mothers using BOLD–fMRI while they listened to recorded infant cries and white noise control sounds matched for intensity and temporal pattern with the cries. On the basis of the animal studies mentioned earlier, they hypothesized that mothers would have more thalamocingulate circuit activity in response to the cries in

comparison to the control noises. A first study consisted of 7 premenopausal, physically healthy mothers aged 21–46 having a youngest child less than 4 years old (range = 3 wks to 3.5 years). Four subjects had fMRI data suitable for analysis. When fMRI data were averaged over these 4 subjects, the anterior cingulate and right medial prefrontal cortex showed statistically increased activity with the cries compared with the control sounds.

We have replicated and extended this finding in a larger cohort and have found that the degree of self-reported bonding corresponds to increased blood flow while hearing cries. Further imaging studies are warranted to understand the role of the thalamocingulate circuit in both maternal behavior specifically and social bonding more generally. Given the importance of bonding for all later behaviors, some would argue that improved understanding of the regional brain basis of social bonding would likely aid in research into normal emotions such as love, separation anxiety, and grief.

There is also a large literature on the potential facilitary role that the septal subdivision of the limbic system and its projections (medial preoptic region of the hypothalamus, ventral bed of the stria terminalis, and ventral tegmentum) may play in mammalian maternal behavior. Lesions of these structures appear to diminish maternal behavior. The medial preoptic region of the hypothalamus and ventral bed of the stria terminalis are likely important. These regions have a great proliferation of oxytocin receptors just before delivery in the female rat, an event that is believed to be important in maternal behavior. The amygdala subdivision of the limbic system has also been implicated in maternal behavior. Some evidence suggests that it may promote, whereas other data suggest that it may inhibit, maternal behavior. In wild, free-ranging Old World monkeys, lesions of the amygdala or strongly interconnected regions (the anterior temporal region and posterior orbitofrontal cortex) may decrease maternal and affiliative behavior. On the other hand, in the rat, medial amygdala lesions (as well as lesions of its projections to the bed nucleus of the stria terminalis and ventromedial nucleus of the hypothalamus) appear to increase maternal behavior, suggesting an inhibitory role for the amygdala in maternal behavior. Given the amydala's role in aversive and defensive behavior, such lesions perhaps may facilitate maternal behavior because nulliparous female rats generally have an aversive reaction to pups. This is thought to be mediated by the smell of the pups because, if the adult female is made anosmic, she will display maternal behavior.

3. Thalamocingulate's Role in Play Behavior

Removal of the cingulate (as well as the underlying dorsal part of the hippocampus and neocortex) impairs play-fighting behavior in hamsters. Conversely, removal of the neocortex alone does not impair this species-typical play behavior in both male and female hamsters.

So far in this article we have concentrated on studies of the functional anatomy of sexual function as discovered through stimulation or lesion studies in nonhuman animals or in relatively crude measures of brain activity (blood flow or glucose use) in humans. Obviously, within this neuroanatomy is an important neurochemistry, and ultimate understanding of sexual function will involve an integration of neuroanatomy with regional neuropharmacology. The pharmacology of sexual function is a relatively new area but one that is capturing popular attention with new medications such as Viagra (sildenafil), which can improve sexual function in humans. Next we outline some of the more recent studies of the regional pharmacology of sexual function, realizing that an exhaustive discussion is beyond the scope of this article.

4. The Neurochemistry of Affiliation

a. Pair Bonding Only 3% of mammals are monogamous, although 15% of primate species are monogamous. Much of the work on the neurochemistry of monogamy has involved the role of oxytocin and vasopressin, two nine amino acid length peptides that are synthesized in the hypothalamus and that differ by only two amino acids. Although these nonapeptides are part of a family of nonapeptides that can be traced phylogenetically to invertebrates, oxytocin and vasopressin are found only in mammals. Oxytocin appears to be important for monogamy in females, whereas vasopressin appears to be important for monogamy in males.

Much of the work on the neurochemistry of monogamy has been performed on voles, a mouse-sized rodent. Two types of voles, called prairie voles and montane voles, physically look very similar and have similar nonsocial behavioral repertoires. They differ in that prairie voles are monogamous, whereas the montane vole is promiscuous. For the monogamous female prairie vole, puberty does not occur at a particular age, but rather when she is exposed to a chemical signal in an unrelated male's urine. This causes her to become sexually receptive within 24 hr of

exposure. When she then mates with a male, they form an enduring monogamous pair bond and become in a sense "addicted" to one another. They share the same nest and territory, remain near each other much of the time, prefer their mate to an intruder, and become attentive parents. Further, the male attacks adult intruders if they come near. The promiscuous montane vole, on the other hand, is a loner. Adult montane vole males and females do not share a nest and come together only briefly to mate. Montane vole females are parental only briefly, and males do not demonstrate parental behavior.

Oxytocin and vasopressin appear to be the important factors in forming the monogamous pair bonds of female and male prairie voles, respectively. For a female prairie vole, if an oxytocin antagonist is centrally administered just prior to mating, mating proceeds normally but she is no more interested in being near her mate than a stranger, just like the montane vole. If oxytocin is given in the absence of mating to a mature, virgin female prairie vole, she prefers a currently present male. For a male prairie vole, if a vasopressin antagonist is centrally administered prior to mating, mating proceeds normally but he is no more interested in being near his mate than a stranger and does not attack adult intruders who come near his mate, just like the montane vole. If vasopressin is given in the absence of mating to a mature, virgin male prairie vole, he prefers a currently present female and will attack intruders. Of importance, centrally administered oxytocin to female montane voles and centrally administered vasopressin to male montane voles does not make them less promiscuous or more social.

Montane and prairie voles have brain oxytocin and vasopressin receptor patterns that are widely different, which likely accounts for their differing monogamy and promiscuity. These vole species differ in *oxytocin receptors* in the prelimbic region, lateral septum, anterior olfactory nucleus, accessory olfactory bulb, nucleus accumbens, and thalamus. Species differences in vasopressin exist in the main olfactory bulb, lateral septum, and thalamus. Other types of monogamous voles have brain receptor patterns similar to those of the prairie voles, and other types of polygamous voles have brain receptor patterns similar to those of the montane voles. In a fascinating experiment, the gene for the vasopressin type 1a receptor in the prairie vole was transplanted into male mice that are usually not very social, and they became much more social than usual, although they did not become monogamous. Interestingly, the brain vasopressin 1a receptor pattern in these transgenic mice was transformed into a pattern similar to that of the normally social and monogamous prairie voles.

The specific implications of these studies for human sexual function remains unclear, but these studies begin to explain the neurochemical bases of bonding and love and perhaps of psychiatric disorders of decreased social bonding such as autism or schizoid behavior. Both vasopressin and oxytocin are released into the blood during human sexual behavior. In males, vasopressin peaks during arousal, whereas oxytocin peaks with ejaculation. In humans, oxytocin receptors are concentrated in the basal forebrain cholinergic nuclei, the nucleus basalis of Meynert, the diagonal band of Broca, and the preoptic areas of the hypothalamus. Vasopressin receptors are concentrated in the lateral septum and the amygdala. This is a pattern of receptor location different from that of either the montane or prairie voles. Therefore, it cannot be assumed that oxytocin and vasopressin have the same exact functions and effects in humans as in voles.

b. Parental Bonding In the laboratory rat, oxytocin appears to be important for maternal bonding with offspring. Prior to the end of pregnancy, female rats show little interest in rat pups. Starting about a day before delivery and persisting through lactation until the time of weaning, they change from avoiding their pups to showing marked interest, including nest-building, retrieval behavior, and guarding of infant rats in experiments. These behavioral changes occur simultaneously with a proliferation of oxytocin receptors in two key regions—the bed nucleus of the stria terminalis and the ventromedial nucleus of the hypothalamus. If a central oxytocin antagonist is given prior to a mother showing interest in her pups, then she does not bond well with her pups. If the oxytocin antagonist is given after she bonds with the pups, maternal behavior is not affected, suggesting that oxytocin is important in the onset but not the maintenance of maternal behavior. Further, when oxytocin is given to estrogen-primed, nulliparous female rats, maternal behavior is facilitated.

Whereas paternal care is minimal in the laboratory rat, vasopressin appears to be important for paternal behavior in the prairie vole. Central injection of vasopressin into the lateral septum increases, whereas central injection of a vasopressin inhibitor into the lateral septum decreases, the time a male prairie vole spends with a pup.

c. Infant Bonding Behaviors During the first 2 weeks of life in both rodents and nonhuman primates, there is a proliferation of oxytocin and vasopressin receptors in limbic regions (especially the cingulate, the region postulated to be involved in separation cries) relative to the adult. These receptors largely disappear by the time of weaning. Interestingly, administration of oxytocin or vasopressin reduces separation responses in rat pups. However, in oxytocin receptor knockout mice, there is a decrease in the separation call rate, suggesting that, perhaps in the absence of oxytocin, rat pups do not experience separation as stressful. Whereas it is intuitively appealing that oxytocin secreted during a mother's lactation may decrease infant crying, this has not been well-studied. Further, oxytocin appears to facilitate, whereas an oxytocin antagonist inhibits, the emergence of conditioned associations and preferences for things linked to one's mother, such as maternal odor.

Regarding mammalian infant bonding behavior, opiate-blocking agents (i.e., naloxone, naltrexone) increase, whereas opiates (such as morphine) given in low, nonsedating doses, decrease pup isolation calls. Regarding maternal behavior, opiate-receptor-blocking agents appear to increase maternal motivation to be near separated crying pups but decrease the maternal competence in retrieving these pups.

Rat pup isolation calls are decreased by serotonergic lesions or by the administration of selective serotonin re-uptake inhibitors and 5-HT1a receptor agonists. Isolation calls appear to be increased by 5-HT2 antagonists and 5-HT1b agonists. In nonhuman primates, grooming of other individuals appears to be important for forming and maintaining both alliances and dominance hierarchy ranking. Studies in various social types of male monkeys have found that the higher the level of cerebrospinal fluid 5-hydroxyindoleic acid (5-HIAA), which is a serotonin metabolite that is a rough measure of brain serotonin activity, the more time a monkey spends grooming others and the higher the dominance status. Further, in vervet monkeys, drugs that increase serotonergic transmission increase dominance status, whereas drugs that decrease serotonin transmission decrease dominance status. In this study, male vervet monkeys that were on fluoxetine or tryptophan (which increased serotonin transmission) appeared to become dominant by increased grooming and affiliation with the females in the group. Of note, in humans, serotonin re-uptake inhibitors are effective in treating social phobia, a condition in which individuals often avoid social encounters.

IV. COMPLEX HUMAN SEXUAL BEHAVIOR

In humans, basic sexual function is carried out within the concept of self-identity and then integration of tastes and preferences about arousing stimuli. Other articles within this encyclopedia describe aspects of this complex behavior in a more detailed fashion. It is important to note, however, that one's own sense of sexual identity is likely determined at an early age, probably within the first year or two of life. The complex brain regions involved in this sense of self-gender are beyond the scope of this article.

Similarly, little is known about the functional neuroanatomy of one's choice of sexual partners or sexually arousing objects. For example, there is considerable controversy about whether there are brain differences between homosexual and heterosexual men, and whether critical exposure to gonadal hormones during brain development may influence later aspects of sexual partner preference. There are numerous examples in nonhumans in which exposure to changes in gonadal steroids for even a brief time during development has lifelong impacts on sexual behavior.

In humans as well there are now a host of classified and recognized disorders of sexual behavior, including compulsive and paraphilic sexual behaviors. The functional and structural neuroanatomy of these areas is not well-understood, although it is of note that pharmacological treatments with antiandrogens, dopamine agonists, or serotonin-modulating medications may alter some of these behavior patterns.

V. SUMMARY AND CONCLUSIONS

The studies described earlier convincingly demonstrate that the limbic cortex receives visual connections as well as input through the vagus nerve about the viscera. Thus, the limbic system serves as an assimilating region in which visual and visceral information is compared and bound. This integration of higher cortical input with visceral and primitive reflexes and urges appears to be essential for what is visually remembered—memory and dreaming. Through the articulation of the anterior thalamus, the limbic system extends its relationship to the phylogenetically new prefrontal cortex, which is involved in insight and in foresight and planning for ourselves and others. Thus, in the complex organization of old and new structures, the human brain has a neural ladder for ascending from the most primitive sexual feelings to the highest altruistic elements.

Thus, human sexual function involves a complex, orchestrated series of behaviors that involve many areas of the brain. Animal stimulation work has identified many of these key regions for the basic aspects of sexual function. In humans, as in most mammals, sexual function also involves selecting and choosing a mate and bonding (love). The functional neuroanatomy of these more complex aspects of sexual function is becoming better understood. Knowledge of the triune brain with the overlapping and interconnected nature of limbic and more basic brain functions reveals why, for most people, sex and love are inextricably linked.

See Also the Following Articles

AROUSAL • BEHAVIORAL NEUROGENETICS • LIMBIC SYSTEM • SEX DIFFERENCES IN THE HUMAN BRAIN • SEXUAL BEHAVIOR • SEXUAL DIFFERENTIATION, HORMONES AND • SEXUAL DYSFUNCTION

Acknowledgment

The authors acknowledge Dr. Paul D. MacLean for helpful comments on this article. Dr. MacLean is located at the Clinical Brain Disorders Branch of the Division of Intramural Research at the National Institute of Mental Health, NIH, Bethesda, MD.

Suggested Reading

Bowers, D., Blonder, L. X., Feinberg, T., and Heilman, K. M. (1991). Differential impact of right and left hemisphere lesions on facial emotion and object imagery. *Brain* **114,** 2593–2609.

Cutler, W. B., Friedman, E., and McCoy, N. L. (1998). Pheromonal influences on sociosexual behavior in men. *Arch. Sex. Behav.* **27,** 1–13.

Davidson, R. J. (1994). Asymmetric brain function, affective style, and psychopathology: The role of early experience and plasticity. *Dev. Psychopathol.* **6,** 741–758.

George, M. S., Ring, H. A., Costa, D. C., Ell, P. J., Kouris, K., and Jarritt, P. (1991). *Neuroactivation and Neuroimaging with SPET*. Springer-Verlag, London.

George, M. S., Ketter, T. A., Parekh, P. I., Rosinsky, N., Ring, H. A., Casey, B. J., Trimble, M. R., Horwitz, B., Herscovitch, P., Post, R. M. (1994). Post RM:Regional Brain Activity When Selecting a Response Despite Interference: An H215O PET study of the Stroop and an Emotional Stroop. *Human Brain Mapping* **1,** 194–209.

George, M. S., Parekh, P. I., Rosinsky, N., Ketter, T. A., Kimbrell, T. A., Heilman, K., Herscovitch, P., and Post, R. M. (1996). Understanding Emotional Prosody Activates Right Hemisphere Regions. *Arch. Neurol.* **53,** 665–670.

Levin, R. J. (1992). The Mechanisms of Human Female Sexual Arousal. *Annu. Rev. Sex. Res.* **3,** 1–48.

Lorberbaum, J. P., Newman, J. D., Dubno, J. R., Horwitz, A. R., Nahas, Z., Teneback, C., Johnson, M. R., Lydiard, R. B., Ballenger, J. C., and George, M. S. (1998). Feasibility of Using fMRI To Study Mothers Resonding to Infant Cries. *Neuroimage* **7,** S907 (abstract).

Maclean, P. D., Newman, J. D. (1988). Role of midline frontolimbic cortex in production of the isolation call of squirrel monkeys. *Brain Res.* **45,** 111–123.

Meston, C. M., and Frohlich, P. F. (2000). The Neurobiology of Sexual Function. *Arch. Gen. Psychiatry* **57,** 1012–1030.

Pardo, J. V., Pardo, P. J., Janer, K. W., Raichle, M. E. (1990). The anterior cingulate cortex mediates processing selection in the Stroop attentional conflict paradigm. *Proc. Natl. Acad. Sci. USA* **87,** 256–259.

Pfaus, J. G. (1999). Neurobiology of sexual behavior. *Curr. Opin. Neurobiol.* **9,** 751–758.

Rauch, S. L., Shin, L. M., Dougherty, D. D., Alpert, N. M., Orr, S. P., Lasko, M., Macklin, M. L., Fischman, A. J., and Pitman, R. K. (2000). Neural Activation during Sexual and Competitive Arousal in Healthy Men. *Psychiatry Res. Neuroimaging*, {com}(in press){/com}.

Stoleru, S., Gregoire, M.-C., Gerard, D., Decety, J., Lafarge, E., Cinotti, L., Lavenne, F., Le Bars, D., Vernet-Maury, E., *et al.* (1999). Neuroanatomical Correlates of Visually Evoked Sexual Arousal in Human Males. *Arch. Sex. Behav.* **28,** 1–21.

Tiihonen, J., Kuikka, J., Kupila, J., Partanen, K., Vainio, P., Airaksinen, J., Eronen, M., Hallikainen, T., Paanila, J., Kinnunen, I., and Huttunen, J. (1994). Increase in cerebral blood flow of right prefrontal cortex in man during orgasm. *Neurosci. Lett.* **170,** 241–243.

Short-Term Memory

GIUSEPPE VALLAR

Università degli Studi di Milano-Bicocca and IRCCS Fondazione S. Lucia, Roma

GLOSSARY

articulatory suppression A paradigm whereby subjects are required to utter an irrelevant speech sound (e.g., *''the, the, the''*) while engaged in another task, such as the short-term retention of a list of words. This concurrent articulatory activity disrupts the process of verbal rehearsal and has little, if any, general interference effects.

buffer/store A limited capacity system concerned with temporary retention; with reference to short-term memory, the term is frequently used as a synonym of ''store.''

double dissociation A pattern of impairment whereby (i) a patient or a group of patients exhibits a defective performance in task A (with respect to normal control subjects) but not in task B, in which the performance level is within the normal range, and (ii) another patient or group shows an opposite disorder (impairment in task B and normal performance in task A). In its ''strong'' or ''classical'' form the double dissociation concerns tasks of comparable difficulty. The double dissociation allows the inference that two functions (e.g., short- and long-term memory) are independent. If this pattern of impairment is associated with different neural correlates (e.g., defective performance in task A but not in B in patients with parietal damage and impaired performance in task B but not in A in patients with frontal damage), the double dissociation is anatomofunctional. The same logic may be applied to neuroimaging activation studies.

recency effect A phenomenon whereby the final events in a series are recalled better than the preceding ones; in immediate free recall the recency effect reflects retention in a short-term memory system.

rehearsal A process that supports temporary retention in a short-term store, reviving a memory trace; rehearsal may involve translation between two different representations or codes.

similarity effect A phenomenon whereby recall of lists of items similar for some specific coding feature (e.g., phonological, visual, and semantic similarity) differs from that of lists including items dissimilar with respect to that particular feature. In short-term memory paradigms similarity effects typically disrupt performance and provide indications as to the representational format of the memory trace.

span A classical measure of short-term memory involving the immediate serial recall of a sequence of events, such as a list of digits, words, visual patterns, and locations of objects.

store A component of short-term memory that secures the temporary retention of a limited amount of material.

trace A term that denotes the event stored in memory and is frequently used in a physiological context.

unattended speech An interfering auditory stimulation consisting of phonological material presented to the subject, who is engaged in another task such as immediate memory span.

working memory This term is widely used in the domain of cognitive psychology, neuropsychology, and cognitive neuroscience. In a more restricted sense, working memory refers to limited capacity systems involved in temporary retention (short-term stores) but with emphasis also on the cognitive operations performed on the stored material. The concept of working memory, however, has expanded to include the efficient monitoring and coordination of mental activity during the simultaneous execution of two or more tasks, the planning of complex actions, and, more general, the organization of behavior. In the latter broader sense, the relationships of the term working memory with its original roots in the specific domain of memory are less close.

The term short-term memory refers to a number of memory systems with limited capacity, concerned with the temporary (in the range of seconds) retention of a variety of materials. This article presents current knowledge concerning the functional and anatomical organization of short-term memory in humans. Behavioral and functional neuroimaging activation studies in normal human subjects, engaged in tasks assessing

short-term retention, as well as neuropsychological evidence in patients with brain damage are considered. The article is mainly concerned with the more extensively investigated aspects of short-term memory: verbal and visuospatial. Much of the data come from studies in the adult, but some pertinent evidence from different populations (children and elderly subjects) is also mentioned.

I. INTRODUCTION

The general idea that the faculty of memory is not unitary and monolithic, comprising instead multiple systems, dates back at least to the "Essay Concerning Human Understanding" by the British philosopher John Locke (1700), who suggested a broad distinction between two types of memory, one concerned with temporary retention and the other a storehouse of materials that have been laid out of sight. This distinction between two memory systems, one shorter and one longer term with respect to the duration of the trace, was revived by the North American psychologist William James (1895), who suggested the existence of a limited-capacity "primary memory," embracing the present and the immediate past and supporting consciousness. Psychological research in the 19th century and in the first half of the 20th century was mainly concerned with the diverse factors affecting learning and retention, in the context of a basically unitary view of human memory. It was only in the 1950s that short-term memory became the object of systematic scientific investigation in normal subjects. Figure 1 shows a model of short- and long-term memory popular in the 1970s.

II. FUNCTIONAL ARCHITECTURE

A. Behavioral Studies in Normal Subjects

The main empirical evidence suggesting the existence of a discrete limited-capacity system, concerned with short-term retention, was provided in the 1950s and 1960s by the extensive investigation in normal human subjects of three behavioral phenomena, which subsequently provided reliable paradigms for exploring the behavior of pathological populations. The majority of empirical studies focused on verbal material, such as words or letters, and therefore on the hypothetical construct of "verbal" or "phonological" short-term memory. Evidence was also provided, however, for the existence of a separate visual short-term store. Examples of stimuli used to investigate short-term retention are shown in Fig. 2.

First, short-term forgetting is assessed by the recall or recognition of individual items after short time intervals filled by interfering activity. The accuracy of

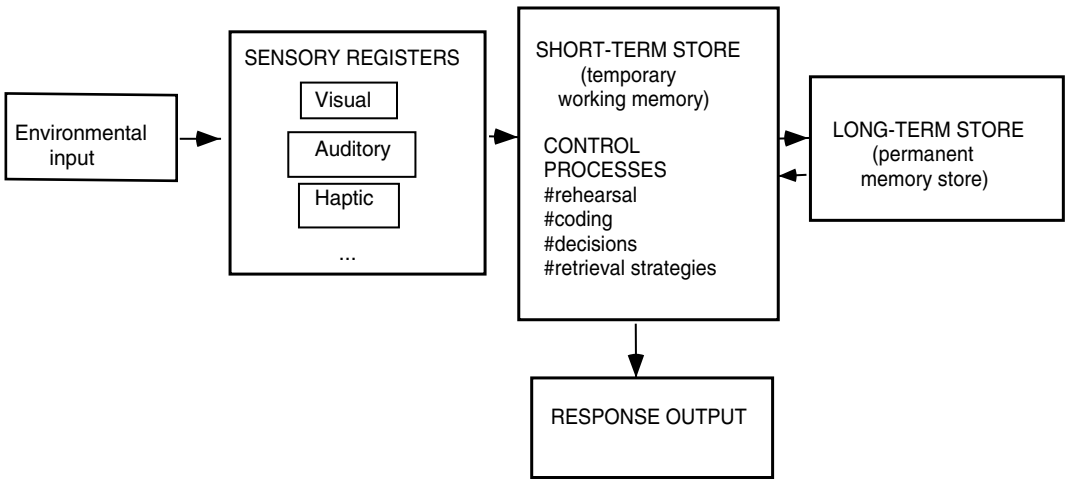


Figure 1 Short-term and long-term memory. This flowchart illustrates a functional architecture of human memory proposed in the late 1960s by Richard Atkinson and Richard Shiffrin that summarizes the main features of a number of models put forward at that time. The model draws a distinction between short- and long-term memory. The two systems are organized serially, with temporary storage in short-term memory being a necessary condition for retention in long-term memory. The short-term store is a unitary system that is not specific for sensory modality, receiving input from different sensory registers, and includes a number of control processes such as rehearsal. This store plays a central role in cognitive activity, being equated with consciousness [based on Atkinson, R. C., and Shiffrin, R. M. (1971). The control of short-term memory. *Sci. Am.* **225,** 82–90].

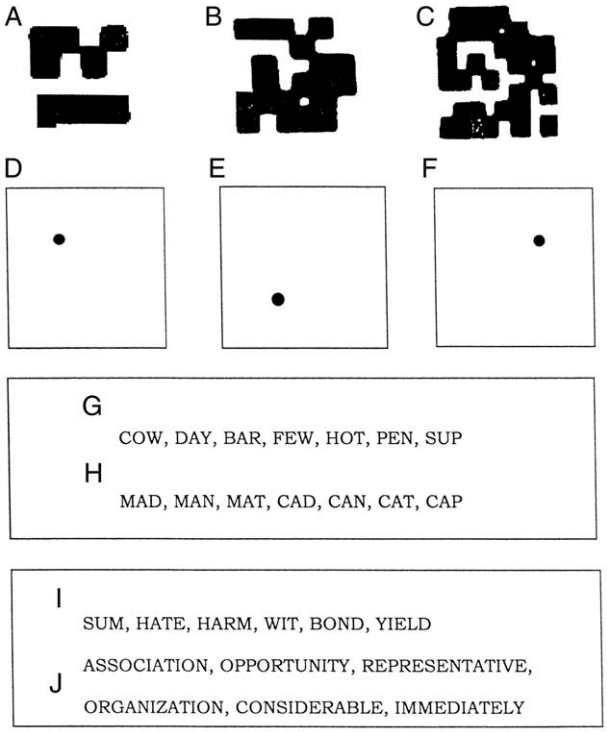


Figure 2 Stimuli used in short-term memory tasks. (A)–(C): patterns used to assess visual short-term memory [based on Phillips, W. A. (1974). On the distinction between sensory storage and short-term memory. *Perception and Psychophysics* **16,** 283–290]. (D)–(F): stimuli used to assess visuo-spatial memory (location of a dot in a square) [Dale, H. C. (1973). Short-term memory for visual information. *British Journal of Psychology* **64,** 1–8]. (G)–(J): stimuli used to assess verbal short-term memory; phonologically dissimilar (G) and similar (H), short (I) and long (J) sets of words. [("G" and "H" sets) Baddeley, A. D. (1966). Short-term memory for word sequences as a function of acoustic, semantic and formal similarity. *Quarterly Journal of Experimental Psychology* **18,** 362–365. ("I" and "J" sets) Baddeley, A. D., Thomson, N., and Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior* **14,** 575–589].

recall of a short list of stimuli (e.g., trigrams of consonants or of words) decreases dramatically in a few (<10) seconds if the subjects' repetition of the memory material (rehearsal) is prevented by a distracting activity such as counting backwards by threes (Fig. 3). A broadly similar forgetting occurs for visual and spatial stimuli (e.g., the location of a dot and random patterns) and auditory nonverbal stimuli (e.g., a tone). This rapid short-term forgetting may be contrasted with the long-lasting persistence in memory of other types of records, such as autobiographical events, which may last for years to the entire life.

Second, in immediate recall, normal subjects presented with a sequence of events such as a list of words or visual patterns, which they are required to recall in any order they wish (free recall), produce first and best the final, more recent stimuli. A comparatively minor advantage of the first one or two items is also present, giving to the recall curve a nonsymmetrical U-shaped form. A similar curve occurs when the subject produces the stimuli of the list in the presentation order (serial recall). The recall superiority of the final items, a phenomenon termed the recency effect (Fig. 4), vanishes after a few seconds of distracting activity. In a typical free recall experiment with verbal material, such as a list of 10–15 words, the recency effect involves the final 5–7 stimuli. The recency effect for verbal stimuli is minimally influenced by factors such as age, rate of presentation, and word frequency, which in turn affect recall performance in the pre-recency positions. This pattern is taken as evidence of a dissociation between short- and long-term memory processes, with the former being associated with the recency effect and the latter with recall in the preceding positions.

Finally, in immediate serial recall (memory span), the ability of normal subjects to recall a sequence of events in their presentation order (e.g., a list of digits, letters, words, or a sequence of locations of objects) is limited (George Miller's "magical" number 7 ± 2 items). Regarding the paradigm of short-term forgetting, if a succession of different lists is presented, subjects become unable to remember the previous ones in a few minutes.

These three sets of empirical observations illustrate the main characteristics of short-term memory: a retention system with limited capacity, holding recent (in the time range of seconds) events, in which the memory trace decays in a few seconds unless this is prevented through rehearsal.

The material stored in short-term memory has a specific representational format. In the verbal domain this involves phonological codes, which may be distinguished from lexical–semantic representations stored in long-term memory. In immediate verbal span, the subjects' performance is primarily affected by factors such as phonological similarity and word length (Fig. 5), with the effects of lexical–semantic factors being comparatively minor. The phonological similarity effect is illustrated by the superior immediate serial recall of sequences of dissimilar words compared to lists comprising similar items (Fig. 2, "G" vs "H" sets). The word length effect is illustrated by the superior immediate serial recall of sequences of short words compared to lists comprising longer items (Fig. 2, "I" vs "J" sets). Also, lexical factors contribute to

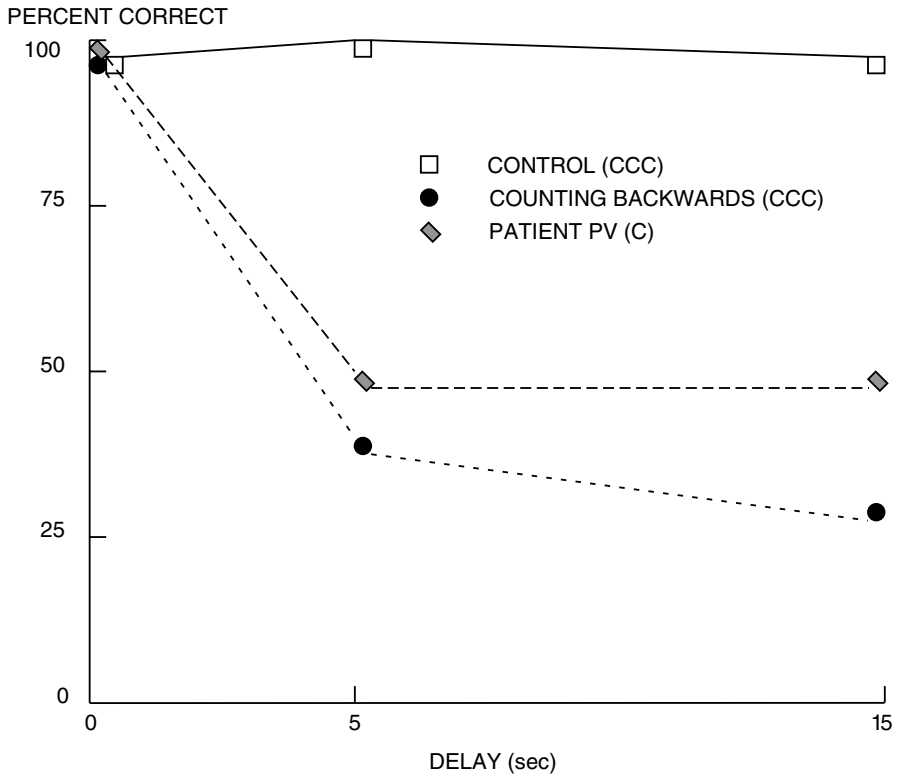


Figure 3 Short-term retention of individual letters. Normal subjects: control, recall of consonant trigrams (CCC) with no interpolated distracting activity; counting backwards, recall of CCC with interpolated distracting activity. Patient PV, recall of single letters (C) with no interpolated distracting activity. In normal subjects the counting task has dramatic detrimental effects on short-term retention of CCC sequences. In the patient who has an auditory span of 2.5 digits, a C stimulus is forgotten in a few seconds, even under conditions of no interference. The errorless performance at the 0-sec delay rules out the possibility that the patient's deficit reflects a perceptual impairment [data from Basso, A., Spinnler, H., Vallar, G., and Zanobio, M. E. (1982). Left hemisphere damage and selective impairment of auditory–verbal short-term memory. *Neuropsychologia* **20,** 263–274; Vallar, G., and Baddeley, A. D. (1982). Short-term forgetting and the articulatory loop. *Q. J. Exp. Psychol.* **34A,** 53–60].

immediate retention, with subjects showing a higher level of performance in span tasks involving word stimuli compared to nonwords or words in a foreign language unknown to the subject, namely, meaningless pronounceable sequences of letters that do not possess a representation in the lexicon.

The three paradigms of short-term forgetting, free recall, and immediate memory span provided the initial impetus for studies in normal subjects aimed at dissociating short- vs long-term memory processes. They also provided the basic tools for assessing this putative dissociation in brain-damaged patients. Each of the phenomena mentioned earlier subsequently proved to be considerably more complex than initially thought. In short-term forgetting the counting backwards task has more complex disruptive effects than the mere interference with the process of rehearsal. In addition, although forgetting in the first 5 sec is likely to reflect the decay of a short-term memory trace, subsequent forgetting involves other mechanisms such as interference. The recency effect is a general memory phenomenon, which may also be observed under conditions involving only long-term memory processes, such as recall of remote autobiographical events. Furthermore, factors such as the relative temporal interval among the events in the sequence influence the occurrence of the recency effect in long-term retention. Immediate memory span is also influenced by lexical factors, which reflect long-term memory processes. Initially, attempts were made to explain the three phenomena in terms of the then dominant view of human memory as a unitary system. Such a monolithic account proved untenable, however, mainly on the basis of neuropsychological evidence from patients with brain damage.

The functional architecture of the system concerned with the short-term retention of verbal material (phonological short-term memory) has been

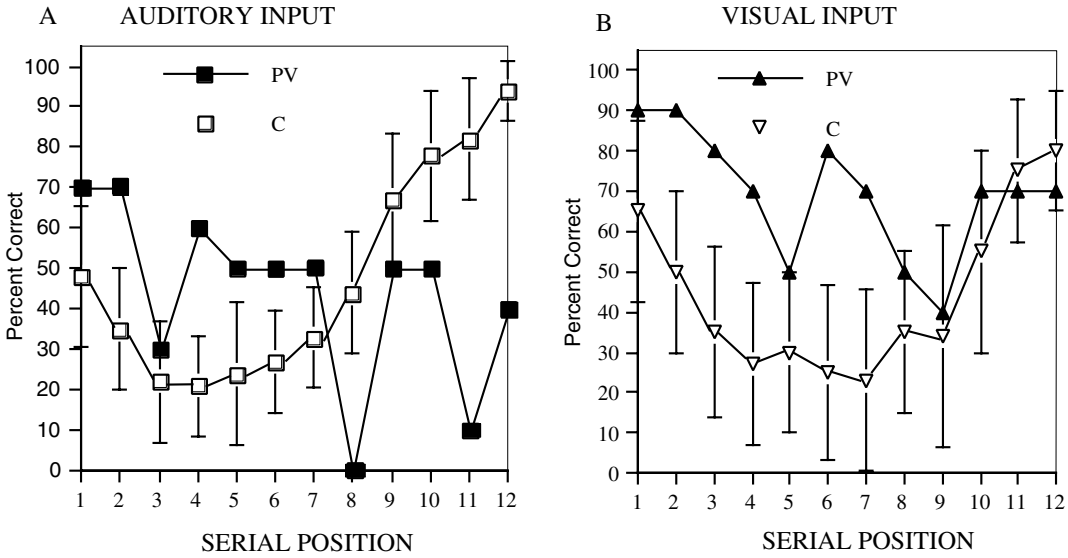


Figure 4 Immediate free recall of sequences of words. Average free recall performance with lists of 12 Italian words by patient PV (see Fig. 2) and normal subjects (C) by serial position of each stimulus in the presentation order and modality of presentation [auditory (A) and visual (B)]. In this task subjects are presented with lists of words, which they subsequently recall in any order they wish. Normal subjects show a higher level of performance in the final positions of the list (recency effect), with recall accuracy steadily increasing in the last six serial positions (in this particular sequence from about the sixth to the final 12th stimulus). The recency effect is present with both auditory and visual presentation of the stimuli, even though an auditory advantage (modality effect) is typically found. Normal subjects also show a minor advantage of the initial one–two positions in the sequence. In immediate free recall, the recency effect reflects the output of short-term memory systems, whereas recall from the preceding positions is mainly based on long-term memory systems. With auditory presentation of the stimuli, patient PV shows no recency, but this is within the normal range with visual input. Patient PV also shows a well-preserved recall performance in the initial and middle positions of the recall sequence [based on Vallar, G., and Papagno, C. (1986). Phonological short-term store and the nature of the recency effect. Evidence from neuropsychology. *Brain Cognition* **5,** 428–442].

investigated in further detail by studying the effects of phonological similarity and word length and the concurrent task of articulatory suppression, teasing apart storage vs rehearsal components. Suppression has three relevant effects on immediate retention in span tasks: (i) It brings about an overall reduction of memory performance; (ii) it abolishes the effect of phonological similarity when the stimuli are presented visually, but not with auditory input; and (iii) it eliminates the effect of word length both with auditory and with visual presentation of the stimuli.

This pattern suggests a fractionation of phonological short-term memory into two components, both contributing to immediate retention in span tasks: (i) a passive input system, the phonological short-term store, to which auditory input has a direct access—this phonological system is nonarticulatory in nature, as witnessed by the persistence of the effect of phonological similarity during suppression; and (ii) a more controlled process of articulatory rehearsal, which has the twofold function of refreshing the phonological trace, preventing its decay, and of conveying visual–verbal information to the phonological short-term store, after recoding, or phoneme-to-grapheme conversion. Suppression, interfering with the operation of rehearsal, reduces span performance and prevents access of visual–verbal material to the phonological store, as witnessed by the disappearance of the effect of phonological similarity. The word length effect, being abolished by suppression, likely reflects articulatory factors, such as the spoken duration and complexity of the stimuli, and may be taken as an index of the activity of the rehearsal process. A model of phonological short-term memory is shown in Fig. 6.

In line with the phonological store/rehearsal fractionation illustrated in Fig. 6, unattended auditory speech disrupts immediate memory span for verbal material, presented in both the visual and the auditory modality, preempting the storage capacity of the phonological short-term input store. Suppression, however, abolishes the disruptive effect of unattended speech on immediate retention of verbal material presented in the visual modality. This effect occurs because suppression, interfering with the operation of rehearsal, prevents visual–verbal material from entering the phonological short-term input store, which is

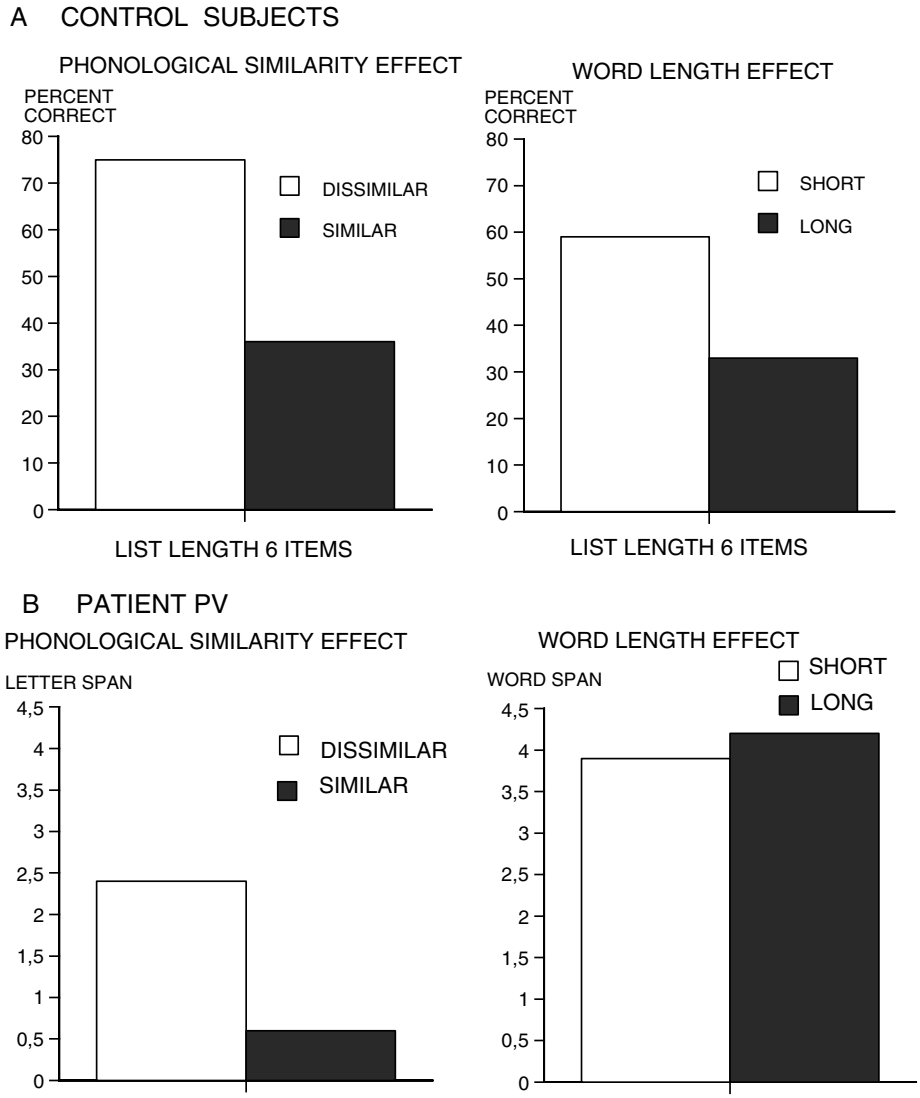


Figure 5 Coding in phonological short-term memory. (A) In immediate serial recall the memory performance of normal subjects is better when the stimuli are phonologically dissimilar rather than similar (phonological similarity effect) and shorter rather than longer (item length effect). Normal subjects exhibit these effects with both auditory and visual presentation of the stimuli. For each effect, the recall performance of normal subjects with six-item auditory sequences of similar and dissimilar letters and long (five-syllable) and short (two-syllable) words is shown. These effects are taken as evidence of phonological coding in verbal short-term memory. (B) Patient PV shows a dramatic reduction of auditory span performance (less than 2.5 letters) and a preserved effect of phonological similarity, which indicates that auditory–verbal material is, as in normal subjects, coded in a phonological format, even though the capacity of the phonological short-term store is dramatically reduced. The absence of any effect of word length with auditory input indicates that the process of articulatory rehearsal is not used by the patient [data from Vallar, G., and Baddeley, A. D. (1984). Fractionation of working memory: Neuropsychological evidence for a phonological short-term store. *J. Verbal Learning Verbal Behav*. **23,** 151–161].

corrupted by unattended speech. Finally, some perceptual phonological judgments are held to specifically involve the articulatory components of rehearsal because the performance of normal subjects is slightly but significantly impaired by suppression. The term phonological judgment refers to the conscious processing (phonological awareness) of a number of phonological aspects of strings of letters, either words or nonwords. In most studies the material is presented visually in order to prevent the direct activation of acoustic–phonological representations. These judgments include deciding whether or not two letter

strings "rhyme" (i.e., their final segments sound the same) or whether their initial sounds are identical or not.

In the nonverbal domain a similar distinction is drawn between visual and spatial short- and long-term memory systems, with the relevant representational format being in terms of the shape or the spatial location of the stimulus. Visuospatial short-term memory is likely to comprise storage and rehearsal (pictorial and spatial) components. In the case of spatial locations, rehearsal may be conceived in terms of planned movements (e.g., ocular, manual reaching, and locomotion) toward a target coded in a spatial reference frame (e.g., egocentric).

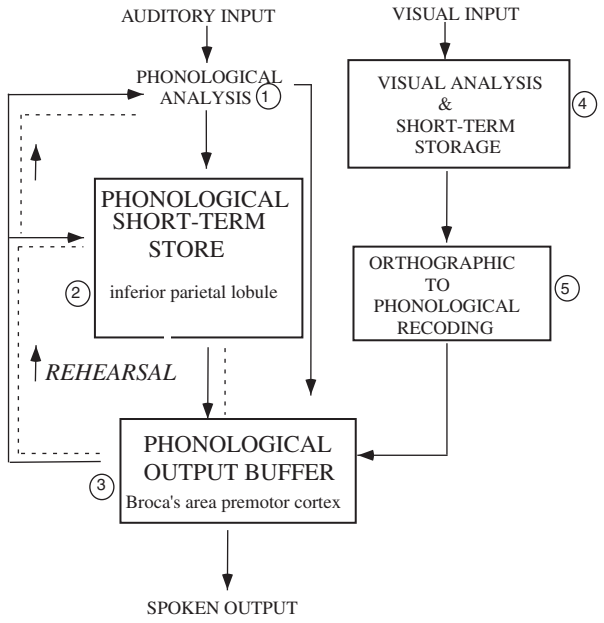


Figure 6 A functional model of phonological short-term memory. After early acoustic and phonological analysis (1), auditory–verbal material enters the main retention component of the system, the phonological short-term store (2), where material is coded in a phonological format. The phonological short-term store is an input system to which auditory material has a direct and automatic access. The process of rehearsal is conceived as involving a recirculation of the memory trace between the phonological short-term store and a phonological output system, the phonological output buffer or phonological assembly system (3), primarily concerned with the articulatory programming of speech output, with a recurring translation between input (acoustic) and output (articulatory) phonological representations. The phonological output buffer provides access of visually presented verbal material to the phonological short-term store after phonological recoding or grapheme-to-phoneme conversion (5). The model also illustrates the multiple-component nature of short-term memory, showing a visual short-term store (4), where material is likely to be encoded in terms of shape [based on Vallar, G., Di Betta, A. M., and Silveri, M. C. (1997). The phonological short-term store-rehearsal system: Patterns of impairment and neural correlates. *Neuropsychologia* **35,** 795–812].

B. Neuropsychological Studies in Brain-Damaged Patients

Studies of patients with memory disorders associated with brain damage provide unequivocal evidence that supports the independence of short- and long-term memory systems, conjuring up a double dissociation of deficits. The hallmark of global amnesia, characterized as a selective impairment of the declarative or explicit component of long-term memory, is the inability to learn verbal and visuospatial material as assessed by recall or recognition tasks (anterograde amnesia). The deficit of these patients also concerns memory for autobiographical events and facts that occurred prior to the onset of the disease (retrograde amnesia). This dramatic memory deficit, which disrupts patients' everyday lives, contrasts sharply with their unimpaired performance in the three paradigms mentioned earlier. Short-term forgetting and immediate memory span are within the normal range for both verbal and nonverbal material. In immediate free recall, amnesic patients show a preserved recency effect associated with a defective performance in the pre-recency positions of the list. These results suggest a functional characterization of amnesia in terms of defective long-term memory, with a preserved function of short-term memory. Second, the pattern of impairment of amnesia is compatible with a serial organization of the two systems, with temporary retention in short-term memory being a necessary condition for long-term storage (see Fig. 1).

Since the late 1960s selective impairments of different components of short-term memory have been reported. The more extensively investigated areas concern auditory–verbal (phonological) and visuospatial short-term memory.

1. Impairments of Phonological Short-Term Memory

Neuropsychological deficits of phonological short-term memory comprise two different types of impairment: a reduced capacity of the store component of the system and a dysfunction of the process of rehearsal.

a. Disproportionately Reduced Capacity of the Phonological Short-Term Store Although it had long been known that patients with dysphasia may show a

disproportionate impairment of repetition of verbal material, such as a sentence or a list of words or digits, it was only in the late 1960s that this deficit was interpreted as a specific disorder of verbal short-term memory rather than as a linguistic impairment (conduction aphasia). The hallmark of this memory disorder is a selective and disproportionate deficit of immediate repetition of all kinds of verbal material (digits, letters, words, and sentences). Patients with this short-term memory deficit, engaged in the three tasks discussed earlier, exhibit the following pattern of performance: (i) Short-term forgetting is abnormally rapid (Fig. 3), (ii) the recency effect in immediate free recall of auditory–verbal lists of words is abolished or severely reduced (Fig. 4), and (iii) auditory–verbal span is disproportionately reduced to an average of less than three items (digits, letters, or words; see Figs. 5b and 7). The retention deficit cannot be explained in terms of defective sensory or perceptual analysis. The patients' immediate reproduction of individual items is entirely spared, as is their ability to perform tasks requiring phonological discriminations, under conditions of minimal memory load. Speech production is also entirely preserved in a number of patients, ruling out interpretations in terms of defective response (articulatory) processes. In addition, the memory deficit does not improve when a response modality that does not require speech production (e.g., recognition by pointing) is used. These features distinguish this "mnestic" disorder of repetition from a type of conduction aphasia in which both the repetition deficit and spontaneous speech are characterized by phonological errors (phonemic paraphasias). Taken together, these findings suggest the impairment of a short-term storage system.

In the majority of these patients the disorder is modality specific, with the level of performance being better when the material is presented visually (Fig. 7). This input-related dissociation has two main implications: (i) Discrete phonological and visual short-term memory components exist, and (ii) in the input–output processing chain, the phonological short-term store should be conceived as an input perceptual system rather than as an output production buffer store. This fractionation also argues against a monolithic view of the system as a single amodal store, which is not specific for the different sensory modalities. Additional support for this input locus of the system comes from the observation that in some of these patients the production of articulated speech is entirely preserved.

Patients with defective phonological short-term memory may show unimpaired long-term verbal

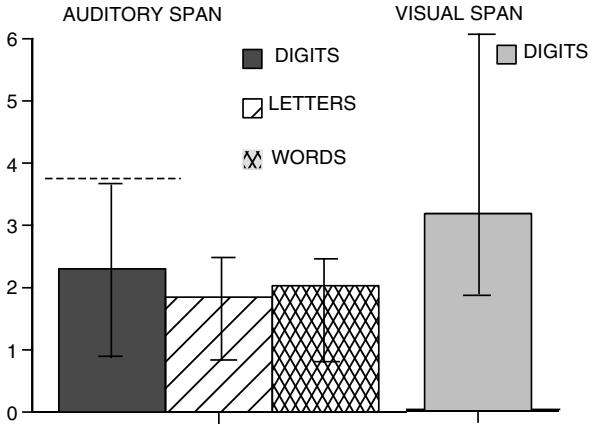


Figure 7 Auditory and visual span in patients with defective phonological short-term memory. Auditory repetition span for all types of verbal material (digits, letters, and words) is defective, whereas the level of performance is higher with visual presentation. In contrast, a minor advantage of the auditory modality is typically found in normal subjects (modality effect). The histograms indicate average spans, the vertical bars indicate the ranges of the patients' span performances, the horizontal dashed line indicates the cut-off for auditory digit span [meta-analytic data from Vallar and Papagno (1995)].

[e.g., paired-associate learning (Fig. 8) and learning of a short story] and visuospatial (e.g., visual maze learning) memory. This observation supports the notion of complete independence of short- and long-term memory systems. However, it is incompatible with a serial organization in which defective short-term memory entails a long-term memory impairment (Fig. 1), suggesting instead a parallel architecture: After early perceptual analysis, information may enter short- or long-term memory, either of which may be selectively disrupted by brain damage. The learning abilities of these patients are dramatically impaired when the phonological material to be learned does not possess preexisting lexical–semantic representations in long-term memory. This is the case for pronounceable meaningless sequences of letters—nonwords or words in a language unknown to the subject (e.g., "svieti" and "tevome") (Fig. 9). This deficit of phonological learning indicates that within a specific representational domain, temporary storage in short-term memory is necessary for stable long-term learning and retention, as predicted by the serial model illustrated in Fig. 1.

b. Impairment of the Rehearsal Process Investigations concerning the precise nature of verbal rehearsal and of its impairments have been performed more recently, with major advances in the past 20 years. The

process of rehearsal has traditionally been described as an activity that, through rote repetition, refreshes the short-term memory trace, preventing its decay. Rehearsal may be conceived in terms of the recoding of the memory trace from input (auditory–verbal) to output-related (articulatory) representations and vice versa (Fig. 6). Specifically, rehearsal of verbal material has long been regarded as an "articulatory" activity ("inner" or "internal" speech), which involves output-related verbal processes, including the motor programming of speech production in a phonological assembly system (Fig. 6), the actual articulation of the material to be rehearsed, making use of the peripheral musculature ("subvocal" rehearsal), or both. Anarthric subjects who, due to congenital disorders or brain damage acquired as adults, are unable to utter any articulated speech sound may show an entirely preserved immediate memory, including verbal rehearsal, as revealed by normal span, similarity, and length effects. This suggests that the process is "central" and does not necessarily require the activity of the peripheral musculature. Brain-damaged patients with a selective impairment of rehearsal show a defective immediate verbal span, as do patients with damage to the phonological short-term store, because both components of phonological memory contribute to immediate retention. Damage confined to the rehearsal process (but not to the phonological store) disrupts the ability to perform phonological judgments, such as rhyme and initial sound. This result, together with the previously mentioned interference effects of articulatory suppression found in normal subjects, suggests that phonological processes such as segmentation and deletion involve output phonological codes. Conversely, a defective rehearsal leaves largely unimpaired the recency effect in immediate free recall of auditory lists of words. This recency effect is grossly reduced or absent in patients with damage to the phonological short-term store (Fig. 4). This dissociation elucidates the precise mechanisms underlying the recency effect in immediate free recall, as reflecting the output of the phonological input short-term store.

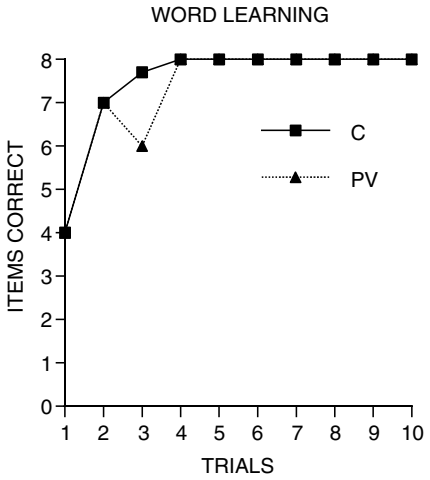


Figure 8 Long-term verbal learning in a patient with defective phonological memory. Patient PV shows unimpaired paired-associate learning of Italian words (C, control subjects). In a paired-associate learning paradigm, subjects are presented with a list of pairs of stimuli (e.g., unrelated words such as "dog"–"pint" and "hammer"–"truth"). Immediately after presentation, or after a delay, subjects are presented with the first member of each pair in an order different from presentation and are required to provide the second member (e.g., "hammer" … → "truth" and "dog" … → "pint") [based on Baddeley, A. D., Papagno, C., and Vallar, G. (1988). When long-term learning depends on short-term storage. *J. Memory Language* **27,** 586–595].

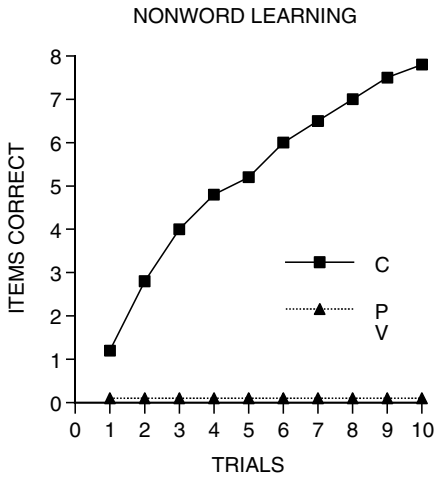


Figure 9 Long-term phonological learning in a patient with defective phonological memory. Patient PV is entirely unable to learn pronounceable nonwords (Russian words transliterated into Italian, e.g., "orange"–"apielsin" and "bear"–"miedvied") in a paired-associate word–nonword paradigm. C, control subjects [based on Baddeley, A. D., Papagno, C., and Vallar, G. (1988). When long-term learning depends on short-term storage. *J. Memory Language* **27,** 586–595].

2. Impairments of Spatial and Visual Short-Term Memory

Two main patterns of short-term memory impairment have been described in brain-injured patients: impairments for spatial locations and for visual patterns. The relationships of these disorders with visuospatial learning have also been investigated. A number of caveats should be considered, however. These deficits

have been explored less extensively than the phonological disorder, and much fewer patients have been adequately described. In addition, the distinction between visual and spatial component processes in memory may be intrinsically blurred because visual objects have spatial properties (e.g., their location in egocentric space and the spatial relationships between their component parts). Finally, no definite distinction between storage and rehearsal components has been drawn.

a. Impairments of Short-Term Memory for Spatial Locations These patients show a defective immediate retention of short sequences of spatial locations, as assessed by a block tapping task (a spatial analog of digit span; Fig. 10a), with verbal memory being unaffected, in both its short- and long-term components (Fig. 10b). Regarding the deficits of phonological short-term memory, the impaired retention of spatial locations cannot be accounted for in terms of visuoperceptual or response-related disorders (e.g., unilateral spatial neglect and misreaching). These patients may show no evidence of topographical disorientation or impairment in long-term learning of visuospatial information, such as the path of a visual maze. These findings suggest a complete independence of short- and long-term memory spatial systems concerned with spatial locations.

b. Impairments of Short-Term Memory for Visual Patterns The short-term visual recognition of unfamiliar faces, objects, voices, and colors may be impaired following brain damage. These deficits may be associated with defective short-term memory for spatial locations. The disorder of patients with defective visual imagery (e.g., as assessed by tasks requiring color or size comparisons) may also be interpreted in terms of a defective visual short-term memory store. A deficit of visual short-term memory may disrupt long-term learning of unfamiliar nonverbal material, as assessed by recognition memory for unfamiliar faces and objects. This extends to the visuospatial domain the conclusion that long-term acquisition requires short-term storage.

III. NEUROLOGICAL ARCHITECTURE

Investigation of the neural correlates of short-term memory systems takes advantage of two main sources of evidence: (i) the traditional anatomoclinical correlations between the site and the size of the cerebral lesion and the behavioral impairment and (ii) the recent functional neuroimaging methods involving the measurement of regional cerebral activity in normal subjects engaged in behavioral tasks probing the function of the component of interest.

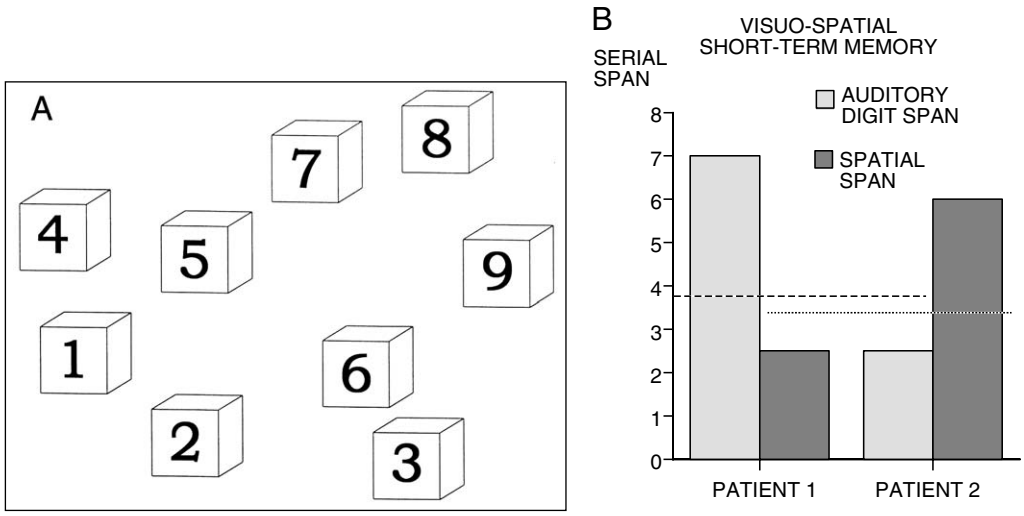


Figure 10 Impairments of short-term memory for spatial positions. (A) The board of the block tapping test. The patient's task is to reproduce the sequence of spatial positions pointed to by the examiner. Like verbal span, sequences of increasing length are presented. In the actual board, the numbers are on the examiner's side. (B) Patient 1 shows a defective visuospatial span associated with a normal auditory digit span; patient 2 shows the opposite pattern of performance. The dashed line indicates the cutoff for digit span, and the dotted line indicates the cutoff for visuospatial span. This double dissociation illustrates the independence of short-term memory for spatial locations from phonological short-term memory [data from De Renzi, E., and Nichelli, P. (1975). Verbal and non-verbal short-term memory impairment following hemispheric damage. *Cortex* **11,** 341–354].

A. Phonological Short-Term Memory

1. Anatomoclinical Correlation Studies in Brain-Damaged Patients

Correlations performed in a large number of patients with defective auditory–verbal span suggest a definite association between damage to a number of areas in the left cerebral hemisphere, located around the Sylvian fissure, and the dysfunction of phonological short-term memory. The parietal association cortex in the left hemisphere [supramarginal gyrus of the inferior parietal lobule (Brodmann's area, BA, 40) at the temporoparietal junction] represents the main neural correlate of the short-term store component of phonological memory. The frontal premotor regions in the left hemisphere and other structures such as the insula constitute the major neural correlates of the rehearsal component, even though the available anatomoclinical data are much more limited. Damage to the right hemisphere does not affect immediate verbal memory, suggesting a left-sided lateralization of the system.

2. Measurement of Regional Cerebral Activity in Normal Subjects

Studies in normal subjects using functional neuroimaging methods support the previously mentioned pathological evidence suggesting a left hemisphere-based network. The short-term store component of phonological memory is associated with activation in the left supramarginal gyrus (BA 40) at the temporoparietal junction, and the rehearsal component is associated with activation in the left frontal premotor regions (BA 44, BA 6, and supplementary motor area) and in the left insula. In these studies, in line with behavioral evidence from normal subjects and brain-damaged patients, an immediate verbal span task activates both the inferior–parietal region at the temporoparietal junction (phonological short-term store), and the premotor cortex (rehearsal process) in the left hemisphere. Conversely, a rhyme judgment task selectively activates the left premotor regions but not the supramarginal gyrus.

As usually found in neuroimaging studies in normal subjects, the network activated by an immediate verbal retention task tapping phonological short-term memory is more extensive and bilateral compared to the anatomoclinical observations in brain-injured patients. The activation observed in normal subjects extends to Wernicke's area (BA 22) in the left hemisphere and to a number of regions in the right hemisphere homologous to the ones active in the left hemisphere (inferior frontal gyrus, BA 44 and BA 6; supramarginal gyrus, BA 40; superior temporal gyrus, BA 20; insula). Similarly, a rhyming task in normal subjects also activates Wernicke's area in the left hemisphere and the insula in the right hemisphere. The difference between the evidence from lesion and activation studies may be related to the redundancy of any neural system subserving a given function. Although the active network is bilateral, its major component is lateralized to the left hemisphere, where the regions necessary to the execution of the task of interest (in this case, immediate memory span) are localized. Accordingly, only damage to the necessary regions brings about behavioral consequences in brain-damaged patients. This approach implies that the complete understanding of the neural correlates of cognitive functions requires evidence from both lesion studies in brain-injured patients and activation experiments in normal subjects.

The activation and lesion-based data support, from an anatomofunctional perspective, the behavioral distinction between a storage component and a rehearsal process in phonological short-term memory. Furthermore, they qualify rehearsal as a process that makes use of components (Fig. 6) also concerned with the planning (i.e., programming in the left premotor cortex) of articulated speech. Finally, from this perspective, phonological memory may be regarded as a component part of the language system.

The neural correlates of the phonological short-term store–rehearsal system are shown in Fig. 11. The system comprises posterior–inferior parietal and premotor components, concerned with the storage and rehearsal aspects of short-term retention. When subjects are engaged in tasks that require more "executive" processes in addition to storage (e.g., free recall and matching a target one, two, or three back in a sequence), the pattern of activation becomes more anterior, extending to the dorsolateral prefrontal cortex, and bilateral.

B. Visual and Spatial Short-Term Memory

1. Anatomoclinical Correlation Studies in Brain-Damaged Patients

Short-term memory for spatial locations, as assessed by tasks requiring the reproduction of sequences of positions in near extrapersonal space, is disrupted by damage to the posterior (parietooccipital) regions of both the left and the right hemisphere, but the role of

right hemisphere damage appears to be more relevant. Damage to the right hemisphere may also affect short-term memory for unfamiliar faces. In contrast, phonological short-term memory is not impaired by lesions in the right hemisphere.

Some deficits of visual short-term memory are associated with damage to the posterior (parietooccipital) regions of the left hemisphere. One such impairment concerns the immediate retention of sequences of visual stimuli, such as straight or curved lines. This deficit may occur in the presence of a normal auditory–verbal span. A second type of impairment concerns the recognition of more than one visual stimulus at a time (defective simultaneous form perception): Patients who are able to identify single meaningful forms visually (letters, numbers, and geometric figures) are disproportionately impaired if two stimuli are presented, misidentifying one of them.

In summary, lesion studies in brain-damaged patients provide support for the distinction between short-term memory for spatial locations and for visual patterns, the former being more closely associated with right hemisphere damage and the latter with left hemisphere damage. The lateralization of visuospatial short-term memory systems appears less pronounced than that of phonological memory.

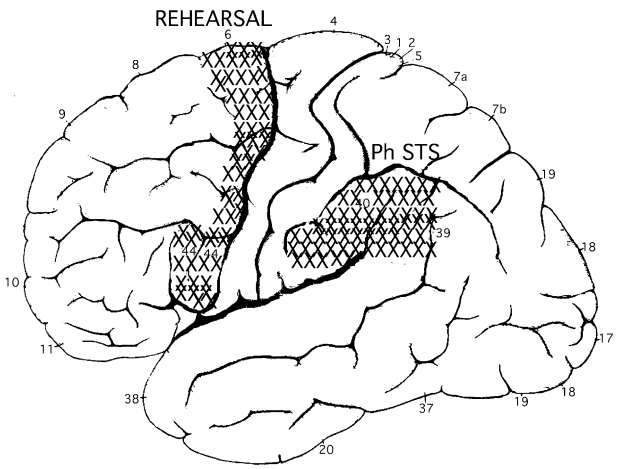


Figure 11 The anatomical basis of phonological short-term memory. This system comprises two components: the phonological short-term store (ph STS), whose main neural correlate is the left inferior parietal lobule (supramarginal gyrus, BA 40) at the temporoparietal junction, and the process of "rehearsal," whose neural basis includes Broca's area (BA 44), the premotor area (BA 6), and the supplementary motor area in the left hemisphere. These two neural short-term memory systems are likely to be connected through the arcuate fasciculus and white matter fiber tracts in the insular region.

2. Measurement of Regional Cerebral Activity in Normal Subjects

Neuroimaging activation studies in humans support and further qualify the distinction between primarily spatial (location) and visual (recognition) short-term memory components. The anatomical differences between them concern both the relevant cerebral areas and their prevailing lateralization. When normal subjects are engaged in tasks involving short-term memory for spatial location, the activated areas include the occipital extrastriate (BA 19), posterior–inferior parietal (BA 40), dorsolateral premotor (BA 6), and prefrontal cortices. Short-term recognition memory for visual patterns (e.g., faces and designs) is associated with activation in a more ventral network, including the occipital and temporal (BA 37) regions, the posterior–inferior parietal region (BA 40), and the prefrontal cortex. Within the prefrontal cortex (BAs 45, 46, 47, and 9), more dorsolateral areas have been associated with short-term memory for spatial location and more ventral areas with short-term memory for patterns. Activation has frequently been found to be bilateral, but there is evidence suggesting more sustained activity in the right hemisphere when the task assesses recognition memory for spatial location and in the left hemisphere when the task probes recognition memory for visual patterns.

These nonoverlapping patterns of activation indicate an association between the dorsal visual stream and short-term memory for spatial locations and between the ventral visual stream and recognition memory for visual patterns. These findings also suggest the existence of some hemispheric asymmetry, with right-sided areas being more concerned with the spatial (location) aspects of short-term retention and left-sided areas with the visual (pattern recognition) aspects. The relative contribution of different cerebral regions of the two sides of the brain may also be related to the format of the cerebral representation involved. For instance, a left lateralization for visual object working memory may reflect a more symbolic or verbal encoding and a right-sided lateralization a more image-based encoding.

Finally, the different cerebral areas participating in the network may provide distinct contributions to the short-term retention process. The occipital and temporal regions may participate in perceptual analysis, the posterior parietal cortex in computing the coordinates of the stimulus, and the premotor and prefrontal cortices in the retention process (storage and rehearsal) proper.

IV. SHORT-TERM MEMORY AND COGNITION

In the preceding sections, behavioral and anatomical evidence for the existence of a number of discrete systems concerned with short-term retention was considered. In this section, short-term memory systems are discussed from the perspective of their relationships with other components of the cognitive system and their role in cognitive activity. Is there a use for a system providing the temporary retention of a limited amount of stimuli, besides infrequent situations, such as the following: a friend tells you an unfamiliar eight-digit number that you have to dial on a telephone on the other side of a major street and you have no paper and pencil to write down the number.

Serial models of memory (Fig. 1) imply that short-term retention is a necessary stage for the stable acquisition of new information in long-term memory. Specifically, there is definite evidence that phonological short-term memory plays an important role in learning new vocabulary. In addition, this system participates in the processes of speech comprehension and production. The evidence concerning visual and spatial short-term memory is less definite.

A. Long-Term Learning

The observation that patients with defective auditory–verbal span are also impaired in learning unfamiliar

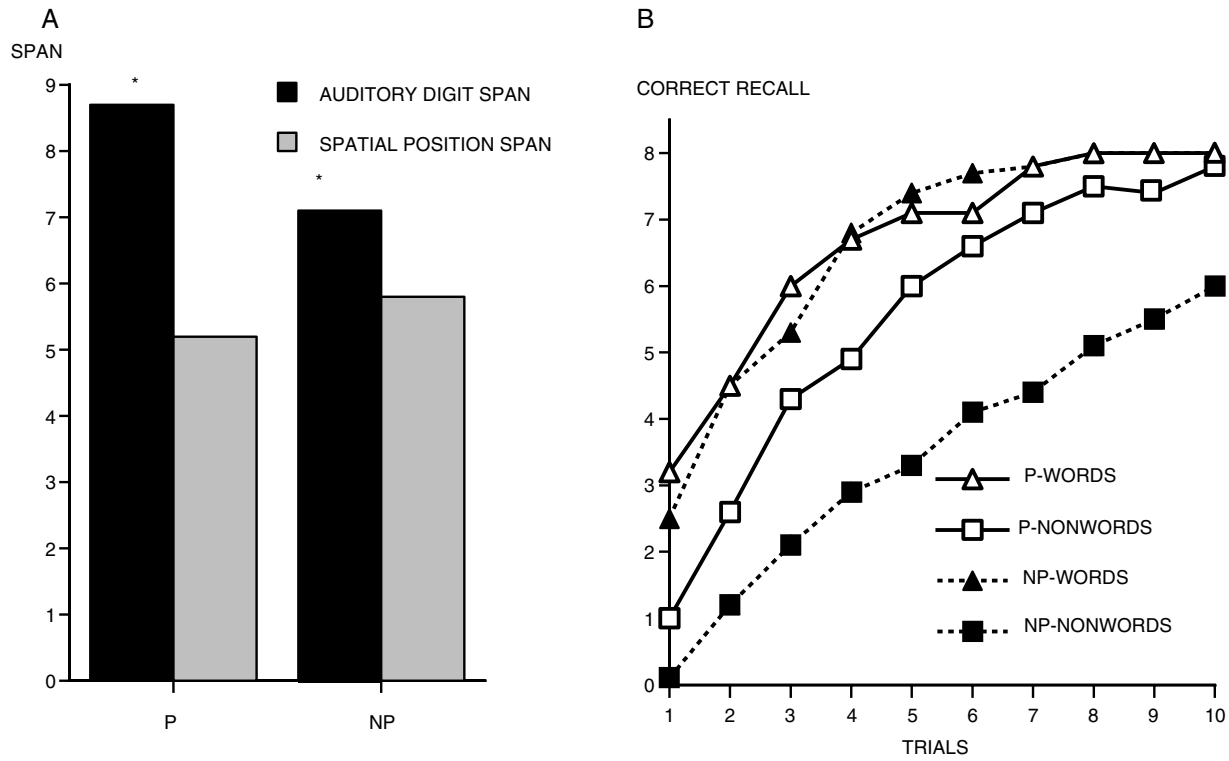


Figure 12 Phonological short-term memory and learning of nonwords by polyglot and nonpolyglot subjects. (A) Polyglot (P) subjects have a higher average immediate auditory–verbal digit span (*), whereas their average spatial span is comparable to that of nonpolyglot (NP) subjects. (B) In a word–nonword paired-associate learning task, P subjects show better learning of pronounceable strings of letters not corresponding to any lexical item known to them. The P subjects' superior learning of new phonological information does not reflect higher learning skills in general, as shown by their word learning performance, which is comparable to that of NP subjects. Similarly, the comparable performance of the two groups in spatial span indicates that the short-term memory advantage of P subjects is confined to the phonological system [based on Papagno, C., and Vallar, G. (1995). Verbal short-term memory and vocabulary learning in polyglots. *Q. J. Exp. Psychol.* **48A,** 98–107].

pronounceable letter sequences raises the possibility that phonological memory may contribute to a relevant aspect of language development—the acquisition of vocabulary (Fig. 9). Similarly, subjects with a developmental deficit of phonological memory are impaired in vocabulary acquisition and in nonword learning. An opposite pattern occurs in subjects with a congenital cognitive impairment that selectively spares phonological short-term memory: Acquisition of vocabulary, foreign languages, and nonword learning are also preserved. Evidence from different subject populations supports this view. First, correlational studies in children have shown that the capacity of phonological memory is a main predictor of the subsequent acquisition of vocabulary, both in the native and in a second language. Second, in normal adult subjects, the variables that disrupt immediate memory span (phonological similarity, item length, and articulatory suppression) also impair the acquisition of nonwords. Third, polyglot subjects have a higher auditory–verbal span compared to nonpolyglots and a better ability to learn novel words (Fig. 12). This advantage of polyglots in nonword learning is not related to differences in general intelligence and long-term verbal or spatial memory, and it likely reflects a greater capacity and superior function of phonological short-term memory. This system may be considered as a learning device for the acquisition of novel phonological representations and the building up of the phonological lexicon. Some observations in brain-damaged patients suggest a similar role for visuospatial short-term memory in the acquisition of new visual information, such as unfamiliar faces and objects. The relevant evidence is much more limited, however.

B. Language Processing

1. Sentence Comprehension

The general idea that temporary retention in phonological memory may contribute to aspects of speech comprehension is far from novel, dating back at least to the 1960s. Phonological memory may hold incoming auditory–verbal strings while syntactic and lexical–semantic analyses are performed. Patients with defective phonological memory exhibit preserved comprehension of individual words and many sentential materials and also a normal ability to decide whether or not sentences are grammatically correct. This may reflect the operation of on-line lexical–semantic processes, heuristics, and pragmatics with a complete independence of syntactic and lexical–semantic processes from phonological memory. However, patients are impaired with complex sentences, where complexity refers to a number of non-mutually exclusive factors, such as (i) a high rate of presentation of the material, which prevents the immediate building up of an unambiguous cognitive representation; (ii) word order conveying information crucial for meaning (e.g., in sentences in which a semantic anomaly is introduced by a change in the linear arrangement of words: ''The world divides the equator into two hemispheres, the southern and the northern''); and (iii) extralinguistic presuppositions biasing the interpretation of the spoken message. Under conditions of this sort an adequate interpretation may require backtracking to the verbatim (phonological) representation of the sentence temporarily held in phonological memory. This may provide a ''backup'' or ''mnemonic window'' resource for performing supplementary cognitive operations necessary for comprehension.

2. Speech Production

In spontaneous speech some errors involve the sequential misplacement of consonant or vowels (phonemic Spoonerism, e.g., ''You have *m*issed all my *h*istory lectures'' → ''You have *h*issed all my *m*istery lectures''). This type of error provides evidence that parts of speech several words in length are planned in advance and stored in a phonological format before actual articulation. The short-term retention system (phonological output buffer or phonological assembly system) in which such errors are generated may be conceived as a working memory space, in which phonological segments are temporarily stored prior to the application of various output processes, such as planning and editing of the articulatory procedures needed to produce speech. In addition, this output retention system, participates in the process of rehearsing material held in the phonological short-term input store (Fig. 5). Seen in this perspective, rehearsal is an activity that makes use of components primarily concerned with a basic linguistic function—speech production. Finally, the phonological output buffer provides a working memory space where operations such as segmentation and deletion may be performed. These may support phonological skills involving a comparison between segments of pairs of stimuli (e.g., rhyme judgments).

C. Spatial Orientation and Navigation

Spatial orientation and navigation in three-dimensional space are complex human skills that require, among other things, keeping track of the current position of the body with reference to previous locations in the environment. A spatial short-term memory system is likely to be involved in the retention of the successive body displacements occuring during navigation. Such a system is likely to integrate multiple sensory inputs (visual, proprioceptive–somatosensory, vestibular, and auditory) and to interact with relevant spatial information stored in long-term memory. These spatial memory systems likely make use of multiple reference frames along the egocentric–allocentric dimension.

V. CONCLUSIONS

Behavioral observations and neuroanatomical evidence from normal subjects and brain-injured patients suggest that short-term memory should be conceived as a multiple-component system with specific functional properties and discrete neural correlates. Current evidence suggests the existence of phonological, spatial, and visual short-term memory systems, each including storage and rehearsal components. These systems secure retention of a limited amount of material in the time range of seconds and contribute to relevant aspects of cognition, such as long-term learning. Specifically, the phonological short-term memory system is closely related to a number of aspects of language: Vocabulary acquisition, speech production, and sentence comprehension.

See Also the Following Articles

AGING BRAIN • ALCOHOL DAMAGE TO THE BRAIN • MEMORY DISORDERS, ORGANIC • MEMORY, EXPLICIT AND IMPLICIT • MEMORY NEUROBIOLOGY • MEMORY, NEUROIMAGING • MEMORY, OVERVIEW • MENTAL WORKLOAD • NERVE CELLS AND MEMORY • SEMANTIC MEMORY • WORKING MEMORY

Acknowledgments

This work was supported in part by grants from the MURST and the Ministero della Sanità.

Suggested Reading

Baddeley, A. D. (1986). *Working Memory*. Clarendon, Oxford.

Baddeley, A. D., Gathercole, S., and Papagno, C. (1998). The phonological loop as a language learning device. *Psychol. Rev.* **105,** 158–173.

Burgess, N., and Hitch, G. J. (1999). Memory for serial order: A network model of the phonological loop and its timing. *Psychol. Rev.* **106,** 551–581.

Burgess, N., Jeffery, K. J., and O'Keefe, J. (Eds.) (1999). *The Hippocampal and Parietal Foundations of Spatial Cognition*. Oxford Univ. Press, Oxford.

Courtney, S. M., Petit, L., Maisog, J. M., Ungerleider, L. G., and Haxby, J. V. (1998). An area specialized for spatial working memory in human frontal cortex. *Science* **279,** 1347–1351.

Gathercole, S. E. (Ed.) (1996). *Models of Short-Term Memory*. Psychology Press, Hove, UK.

Glanzer, M. (1972). Storage mechanisms in recall. In *The Psychology of Learning and Motivation. Advances in Research and Theory* (G. H. Bower, Ed.), Vol. 5**,** pp. 129–193. Academic Press, New York.

Hitch, G. J., and Logie, R. H. (Eds.) (1996). *Working Memory*. Psychology Press, Hove, UK.

McCarthy, R. A., and Warrington, E. K. (1990). *Cognitive Neuropsychology. A Clinical Introduction*. Academic Press, San Diego.

Paulesu, E., Frith, U., Snowling, M., Gallagher, A., Morton, J., Frackowiak, R. S. J., and Frith, C. D. (1996). Is developmental dyslexia a disconnection syndrome? Evidence from PET scanning. *Brain* **119,** 143–157.

Smith, E. E., and Jonides, J. (1999). Storage and executive processes in the frontal lobes. *Science* **283,** 1657–1661.

Vallar, G., and Papagno, C. (1995). Neuropsychological impairments of short-term memory. In *Handbook of Memory Disorders* (A. D. Baddeley, B. A. Wilson, and F. Watts, Eds.), pp. 135–165. Wiley, Chichester, UK.

Vallar, G., and Shallice, T. (Eds.) (1990). *Neuropsychological Impairments of Short-Term Memory*. Cambridge Univ. Press, Cambridge, UK.

Vallar, G., Di Betta, A. M., and Silveri, M. C. (1997). The phonological short-term store-rehearsal system: Patterns of impairment and neural correlates. *Neuropsychologia* **35,** 795–812.

Sleep Disorders

CLETE A. KUSHIDA
Stanford University School of Medicine

GLOSSARY

actigraphy Method of estimating sleep–wake patterns by activity/movement-sensitive devices, usually placed on the wrist.

apnea Cessation of airflow lasting at least 10 sec associated with a decrease in blood oxygen saturation.

continuous positive airway pressure A device consisting of a mask, usually placed over the nose, connected via a hose to a pump, which draws in room air and gently pushes the air into the patient's airway to prevent it from collapsing during sleep.

hypopnea Decrease, but not complete cessation, of airflow lasting at least 10 sec and typically associated with a decrease in blood oxygen saturation.

non-REM sleep Comprises the majority of a night's sleep and is divided into four stages based on electroencephalograph (EEG) frequency, amplitude, and specific EEG waveforms.

polysomnogram or sleep study A test consisting of assessment of breathing, heart rate, snoring intesity, and activity of the brain, eyes, and chin and leg muscle activity.

REM sleep A type of sleep in which the majority of dreams occur, the EEG patterns are similar to those of wakefulness, and most of the muscles under voluntary control are paralyzed.

sleep latency Time to onset of sleep after the lights are turned off in the bedroom.

slow-wave sleep Deep sleep, also called non-REM stages 3 and 4 sleep, during which time the brain electrical activity is characterized by high-amplitude, low-frequency waveforms.

Sleep disorders are highly prevalent but are largely unrecognized, misdiagnosed, and untreated. The *International Classification of Sleep Disorders, Revised: Diagnostic and Coding Manual* by the American Sleep Disorders Association classifies sleep disorders into four broad categories: (i) The dyssomnias are disorders of either difficulty initiating or maintaining sleep or excessive daytime sleepiness; (ii) the parasomnias are disorders that encroach into the sleep process; (iii) sleep disorders associated with mental, neurologic, or other medical disorders; and (iv) proposed sleep disorders, in which insufficient data are available to confirm the existence of these disorders. Based on this classification system, there are currently 77 recognized sleep disorders and 11 proposed sleep disorders. This article describes these sleep disorders and identifies key elements of the history, physical examination, and diagnostic evaluation for the patient presenting with a sleep complaint.

I. EVALUATION OF A PATIENT PRESENTING WITH A SLEEP COMPLAINT

A. Sleep History

A thorough sleep history is important to assess the patient's chief complaint and to derive a differential diagnosis. Typical questions useful to a clinician evaluating a patient with a sleep problem are listed in Table I. The disorders described later are recognized sleep disorders from the *International Classification of Sleep Disorders*. Although the majority of patients with sleep problems typically present with dyssomnia complaints, it is not unusual for a given patient to have more than one type of sleep disorder.

An interview with the bed partner at the same time as the patient is important; the partner can frequently provide important symptoms about which the patient is not aware. For example, the bed partner can report

Table I
Typical Sleep History Questions[a]

- Bedtime and awakening times; is this schedule consistent during the week? (Consider circadian rhythm disorders)
- Works a nighttime or rotating shift? (Consider circadian rhythm disorders)
- Uses medications to fall asleep or stay awake? (Consider insomnias and hypnotic- or stimulant-dependent sleep disorders)
- Time to fall asleep after lights turned out? (Consider insomnias vs disorders with excessive daytime sleepiness)
- Number of awakenings during the night, cause (if known), and time to fall back asleep? (consider dyssomnias)
- Feels refreshed in the morning? (Consider disorders with excessive daytime sleepiness)
- Number of naps per week; timing and duration of naps? (Consider disorders with excessive daytime sleepiness)
- Amount, timing, and frequency of caffeinated beverages? (Consider disorders with excessive daytime sleepiness)
- Falls asleep inappropriately during the day or while driving? (Consider disorders with excessive daytime sleepiness)
- Snoring, witnessed breathing pauses, or awaking with gasping, gagging, choking, snorting, or dyspnea? (Consider obstructive sleep apnea syndrome)
- Frequent difficulty falling asleep or staying asleep? (Consider insomnias vs disorders with excessive daytime sleepiness)
- Unusual sensations in the legs that are relieved by movement? (Consider restless legs syndrome)
- Leg twitching or kicking during the night? (Consider periodic limb movement disorder)
- Unusual movements or walking during sleep? (Consider parasomnias)
- Any limb movements associated with dreams? (Consider REM sleep behavior disorder)
- Unable to move during the transition from sleep to wakefulness? (Consider narcolepsy)
- Auditory, visual, tactile, or kinetic hallucinations at sleep onset? (Consider narcolepsy)
- Episodes of sudden muscle weakness precipitated by strong emotion? (Consider narcolepsy)

[a]This is not intended to be a comprehensive screening questionnaire; however, some of these questions, in addition to a thorough clinical evaluation, may be useful in screening a patient for a specific sleep disorder.

that the patient has gasping or choking episodes during sleep, which would be characteristic of a sleep-related breathing disorder such as obstructive sleep apnea syndrome.

Children with sleep complaints deserve special mention. A higher prevalence of parasomnias, such as sleep terrors and sleepwalking, is found in children compared to adults, and children often do not manifest daytime sleepiness in the same manner as adults. For example, instead of appearing drowsy or sleepy, children may be hyperactive and irritable, even to the degree of being misdiagnosed as having attention-deficit hyperactivity disorder.

At the time of the initial evaluation, sleep specialists frequently use questionnaires, such as the Sleep Disorders Questionnaire by Douglass *et al*., to aid in the diagnosis of these disorders. The Stanford Sleepiness Scale (SSS) by Hoddes *et al.* and the Epworth Sleepiness Scale (ESS) by Johns have been used to assess the subjective level of daytime sleepiness, although there has been criticism that these scales do not correlate well with objective measures of daytime sleepiness. Daily sleep logs are helpful in documenting subjective sleep–wake patterns. These logs are frequently useful for the clinician in evaluating patients suspected of having irregular sleep–wake schedules or circadian rhythm disorders. The logs are typically 2-week sleep diaries in which the patients record their bedtime and awakening times, estimated sleep-onset time, time and duration of nocturnal awakenings and daytime naps, and time of medication use. Lastly, quality of life questionnaires have been used to assess the impact of treatment of sleep disorders.

1. Dyssomnias

a. Insomnias Patients with insomnia have difficulty initiating and maintaining sleep, characterized by difficulty falling asleep and frequent awakenings; they also typically suffer from chronic fatigue and tiredness, irritability, and concentration and memory impairment. Patients who have inadvertently conditioned themselves to have difficulty falling or staying asleep in their normal bedroom environment indicate a diagnosis of psychophysiological insomnia. A lifelong history of insomnia, typically starting in childhood, characterizes idiopathic insomnia. Insomnia in relation to a significant stress event, such as divorce or death of a parent, typifies adjustment sleep disorder. Patients with poor sleep habits may suffer from inadequate sleep hygiene, and sleep state misperception, in which the patient's sleep complaints do not accurately reflect objective measures of sleep, is commonly found in insomniacs. Insomnia due to elevations above 4000 m and sleep disturbances due to

environmental factors are hallmarks of altitude insomnia and environmental sleep disorder, respectively. Lastly, insomnia found in children may be characteristic of limit-setting sleep disorder (unenforced bedtimes), sleep-onset association disorder (the absence of certain conditions, e.g., watching television), food allergy insomnia, or nocturnal eating (drinking) syndrome.

The completion of daily sleep diaries may be useful in detecting the sleep–wake patterns of insomniacs; it is particularly beneficial in demonstrating poor sleep habits indicative of inadequate sleep hygiene as well as circadian rhythm disorders coexisting with insomnia. Actigraphy, the use of movement-sensitive devices to assess sleep–wake cycles, may be beneficial in screening for sleep state misperception and circadian rhythm disorders in insomniacs. However, an unknown number of insomniacs may remain in bed without moving or restlessness; this behavior may incorrectly register as sleep by actigraphy since it may be below the movement threshold of the actigraphic device for recognition of wakefulness. The coexistence of psychiatric conditions, such as depression or anxiety disorders, as well as circadian rhythm disorders should be screened for in all patients presenting with insomnia since the proper identification and treatment of these conditions usually hastens the patient's recovery from insomnia.

The mainstay of insomnia treatment is implementation of behavioral techniques. These include the adoption of a strict sleep–wake schedule of regular bedtimes and awakening times coupled with the avoidance of naps. Patients should avoid remaining in bed for extended periods of time while awake since this may result in the patients being conditioned to being awake in their bedroom environment. Decreased caffeine, nicotine, and alcohol use, especially prior to bedtime, and limiting heavy exercise, meals, and fluid intake near bedtime frequently aid in reducing the time to fall asleep as well as the frequency and duration of nocturnal awakening. Avoidance of reading and watching television in bed, unless these activities result in drowsiness, may also help initiate sleep. Other specific behavioral approaches, such as stimulus control therapy, sleep restriction, and relaxation therapy, can be beneficial. Stimulus control therapy involves reassociating the bedroom environment with rapid sleep onset. Sleep restriction is the programmed reduction in the duration of time spent in bed, with the goal of increasing sleep efficiency (time spent asleep/time in bed). Relaxation therapy includes meditation, self-hypnosis, progressive muscle relaxation, and electromyographic biofeedback. Lastly, a mild hypnotic medication with a short elimination half-life (e.g., zolpidem) may be used as a supplement to behavioral therapy in intractable cases of insomnia.

b. Disorders with Excessive Daytime Sleepiness Patients with excessive daytime sleepiness have decreased cognitive functioning during the day as well as frequent periods of falling asleep at inappropriate times, such as during meetings or while driving. Although patients with the following disorders may present with insomnia, the more typical presentation is that of excessive daytime sleepiness.

Daytime sleepiness combined with a history of witnessed nocturnal breathing pauses or awaking with gasping, gagging, choking, snorting, or dyspnea characterize the obstructive sleep apnea syndrome. Based on epidemiology studies by Terry Young and colleagues in 1993, it is estimated that up to one-fourth of men and one-tenth of women between the ages of 30 and 60 years have polysomnographic evidence of this condition. Besides placing the patient at greater risk for motor vehicle and work-related accidents, this disorder is associated with cardiovascular conditions such as hypertension, stroke, myocardial infarction, dysrhythmias, and cardiac failure as well as neurocognitive problems such as decreased short-term memory. The diagnosis of obstructive sleep apnea syndrome is confirmed by a sleep study (Fig. 1), which reveals the repetitive abnormal breathing events (apneas and hypopneas) characteristic of this disorder. Treatment consists of weight loss, cervical positioning devices, oral appliances (devices that move the tongue or mandible anteriorly), nasal continuous positive airway pressure (CPAP), or upper airway surgery. The upper airway surgery may involve a uvulopalatopharyngoplasty (i.e., removing redundant tissue from the soft palate and tonsillar pillars), mandibular osteotomy with genioglossus advancement (moving the tongue base anteriorly), and a hyoid myotomy (stabilizing the tongue muscles). Newer technologies such as radiofrequency surgery may offer patients a near painless, rapid, outpatient procedure for reducing snoring and obstructive sleep apnea. Seasonal allergic rhinitis and significant nasal obstruction (e.g., nasal polyps and adenoidal hypertrophy) may worsen the patient's obstructive sleep apnea syndrome; the patient may need antihistamines, nasal steroids, or corrective surgery to alleviate these conditions. The upper airway resistance syndrome is a mild variant of obstructive sleep apnea syndrome; this disorder is characterized by excessive daytime sleepiness resulting from frequent

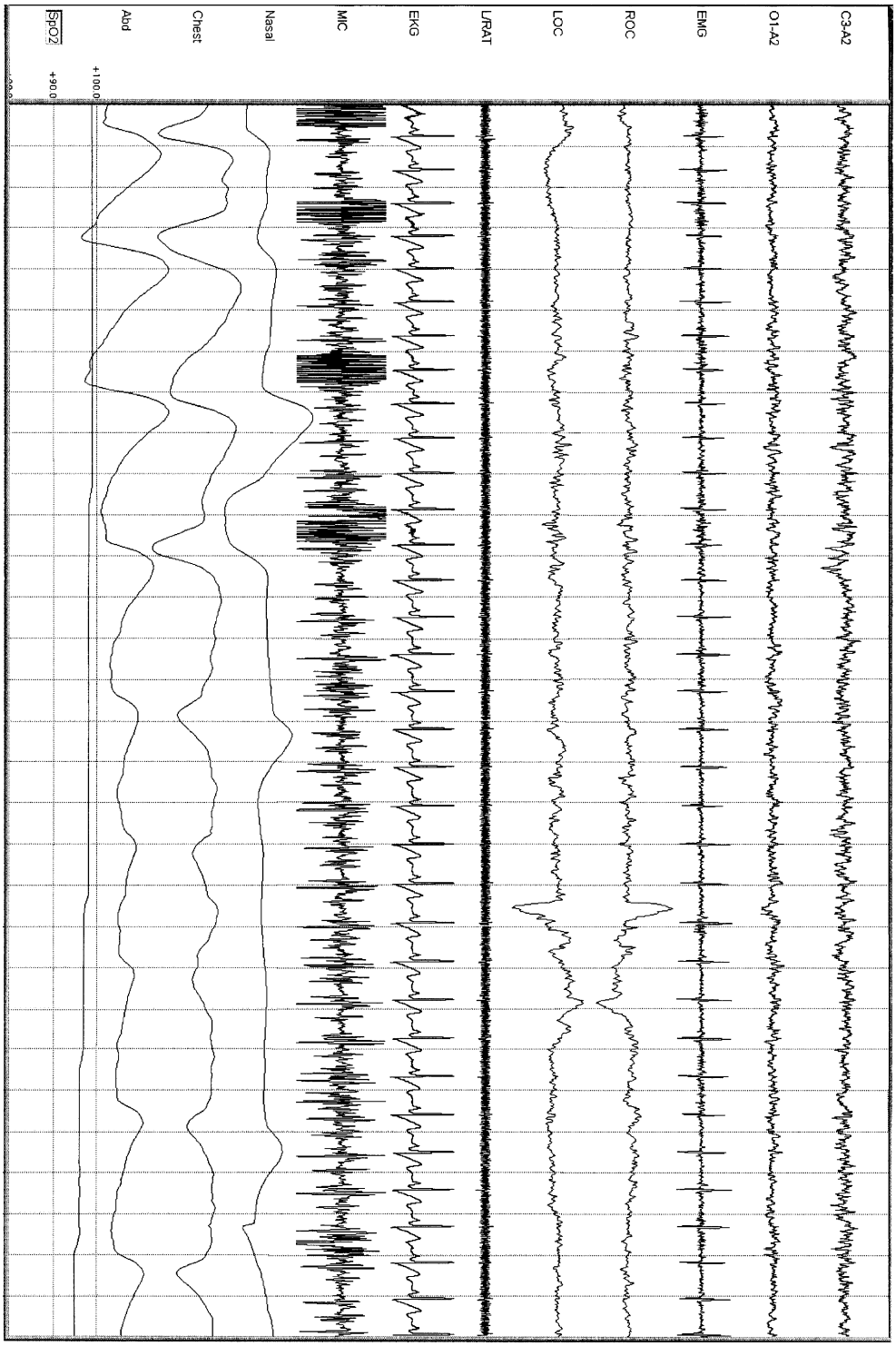


Figure 1 Polysomnographic recording for a patient with obstructive sleep apnea syndrome. This represents a 30-sec period of a polysomnogram. The patient is in REM sleep, as evidenced by the EEG (C3-A2 and O1-A2 electrode derivations) with a low-amplitude, mixed frequency pattern; the EMG (EMG representing chin muscle activity, and L/RAT representing left and right anterior tibialis leg muscle activity) with relatively suppressed muscle activity; and the EOG (ROC and LOC representing right and left eye muscle activity, respectively) with rapid eye movements. Snoring is evidenced in the MIC (microphone) channel. An obstructive apnea is apparent in the right half of the figure, with a cessation of nasal airflow despite chest and abdominal wall movement. As a consequence of this apnea, the SpO_2 (oxygen saturation) decreases from approximately 99 to 95% by the right edge of the figure.

arousals that, in turn, are a consequence of increased work of breathing (detected by esophageal pressure monitoring). Patients with the rarer central sleep apnea syndrome or central alveolar hypoventilation syndrome may present with complaints or a history similar to those of obstructive sleep apnea syndrome. CPAP or bilevel pressure (increased pressure during inspiration and decreased pressure during expiration) may similarly help patients with these conditions.

Narcolepsy is a sleep disorder estimated to affect approximately 200,000 Americans, and there appears to be a genetic predisposition for the development of this condition. This disorder commonly occurs in the second decade of life with an equal distribution in men and women. Narcolepsy is suspected in individuals with the following symptoms: (i) attacks of daytime sleepiness, which typically appear suddenly and irresistibly throughout the day; (ii) cataplexy, or the sudden loss of postural muscle tone typically precipitated by strong emotion (subtle cases of cataplexy may involve only slight weakness of the neck muscles); (iii) hypnagogic (sleep-onset) hallucinations, which may be fleeting sounds (e.g., voices), visions, or full-blown auditory, visual, tactile, or kinetic hallucinations; and (iv) sleep paralysis (which may occur in individuals without narcolepsy), that occurs at the transition from sleep to wakefulness. This paralysis may frequently accompany the hypnagogic hallucinations. The diagnosis of narcolepsy is established by symptoms and confirmed by the multiple sleep latency test. This disorder is typically treated by scheduled naps, stimulants (e.g., methylphenidate, modafinil, and dexedrine) to ward off daytime sleepiness, and anticataplectic medications (typically tricyclic antidepressants).

Daytime sleepiness combined with hypersomnia (prolonged sleep periods) characterize idiopathic, recurrent, and posttraumatic hypersomnia. These disorders may be differentiated from narcolepsy and other disorders of excessive daytime sleepiness by symptoms, polysomnography, and multiple sleep latency tests. Treatment typically consists of scheduled naps and stimulant medication in order to control the irresistible sleepiness throughout the day. Lastly, daytime sleepiness with evidence of voluntary sleep deprivation suggests insufficient sleep syndrome; lifestyle modification to reduce work or family commitments may be necessary to increase the patient's sleep requirements.

c. Restless Legs Syndrome and Periodic Limb Movement Disorder Restless legs syndrome is characterized by dysesthesias or unpleasant sensations in the legs that result in an irresistible urge to move the legs. The dysesthesias may be described as a "creeping," "crawling," "itching," or pain and are relieved by movement. Although these symptoms can occur anytime throughout the day, they may worsen at night, thus interfering with the ability to initiate sleep and ultimately resulting in insomnia symptoms or daytime sleepiness. It commonly occurs in middle age, affects equally both men and women, and pregnant women and individuals with uremia and rheumatoid arthritis appear susceptible to developing this disorder. The diagnosis is typically made by history, although a sleep study will frequently detect the limb movements resulting from the dysesthesias. First-line treatment consists of dopaminergic agents (e.g., carbidopa and levodopa); benzodiazepines (e.g., clonazepam), opiates, and anticonvulsants (e.g., tegretol and gabapentin) have variable effectiveness in treating this disorder.

Periodic limb movement disorder is characterized by daytime sleepiness resulting from repetitive limb movements during sleep. The leg movements are typically stereotyped, with extension of the big toe and partial flexion of the ankle, knee, and hip. Although peak onset of this disorder is in middle age, the incidence increases with age. Patients suspected of having this disorder are evaluated by polysomnography, and specific polysomnographic criteria for the frequency and duration of individual limb movements have to be met for confirmation of the diagnosis. The clinical decision to treat this disorder is usually made by the severity of the symptoms and frequency of the periodic limb movements associated with arousals. Treatment consists of the same medications used to treat restless legs syndrome; as in restless legs syndrome, dopaminergic agents remain first-line therapy for periodic limb movement disorder.

d. Circadian Rhythm Disorders Circadian rhythm disorders should be considered in patients who have an unusual sleep–wake pattern, and these patients may present with either insomnia or excessive daytime sleepiness. Time zone change (jet-lag) syndrome and shift work sleep disorder are characterized by the effect of external factors on the patient's sleep. These disorders are usually easy to diagnose by patient history; strict application of sleep hygiene recommendations and, if needed, a short-acting hypnotic medication (e.g., zolpidem) may be beneficial in helping these patients. Delayed or advanced sleep-phase syndromes correspond to disorders of the timing

of the sleep period, in which there is either a chronic delay or advance, respectively, in the major sleep period with respect to the desired bedtime and awakening time. For example, an individual with delayed sleep-phase syndrome may desire to sleep at 10 PM and awaken at 6 AM but finds that he or she usually becomes sleepy at 2 AM and spontaneously awakens at 10 AM on a chronic basis. These disorders are frequently recognized by careful history taking; sleep logs and actigraphy may be useful in establishing the diagnosis. Treatment consists of behavioral conditioning techniques such as stimulus control therapy and relaxation therapy, and a mild hypnotic medication may be necessary to initiate sleep at the desired bedtime. Irregular sleep–wake pattern and non-24-hr sleep–wake disorder are characterized by a disorganized and a progressively delayed sleep–wake pattern, respectively. These disorders also rely on careful history taking, sleep logs, and actigraphy to establish the diagnosis. The prognosis for these latter disorders is poor, often relying on major lifestyle accommodations (e.g., adapting the work schedule to fit the wake–sleep schedule) in addition to the behavioral and pharmacologic treatment, described previously used to treat other circadian rhythm disorders.

e. Disorders Associated with Exposure to Toxins, Substances, or Drugs Exposure to substances or toxins (e.g., organic chemicals and heavy metals) can produce insomnia or excessive daytime sleepiness. The toxin-induced sleep disorder and hypnotic-, stimulant-, and alcohol-dependent sleep disorders are characterized by sleep complaints arising from the use or withdrawal of these compounds. In particular, hypnotic and stimulant disorders frequently coexist with other sleep disorders. It is not uncommon to find patients with insomnia who have abused various hypnotics for many years and to find narcoleptics or patients with obstructive sleep apnea syndrome who have abused stimulants in order to maintain daytime alertness. Patients with alcohol-dependent sleep disorder often find that their time to sleep onset is reduced; however, the frequency and duration of nocturnal awakenings increase.

2. Parasomnias

Unusual behaviors, sensations, or movements are characteristic of parasomnias. Since the patient is often unaware of these abnormal episodes, the patient's bed partner is often more helpful in describing these events.

a. Sleep Terrors and Sleepwalking Sleep terrors and sleepwalking (somnambulism) occur during slow-wave sleep typically during the first third of the night and are characterized by screaming and ambulating, respectively, in addition to amnesia for the episodes. The individual with sleep terrors will frequently awaken screaming and confused, unresponsive to the environment and intensely fearful, with dilated pupils and rapid respiratory and heart rate. These disorders are thought to be related, and it is not uncommon for individuals to have both of these conditions. Typically, the incidence of these disorders is between 1 and 15% of the general population, and it is estimated that 15–30% of healthy children have had at least one episode of sleepwalking. Both disorders are thought to be familial, with a peak age of onset of approximately 4–8 years, and above the age of 15 years the incidence of sleepwalking and sleep terrors is about 1%. The etiology of sleep terrors and sleepwalking is unknown, but certain factors, such as antipsychotic or antidepressant medications, fever, sleep deprivation, and other sleep disorders that fragment sleep, increase the probability of these episodes in susceptible individuals. In the case of sleepwalking, sleep-related violence, such as assault or homicide, may occur; the legal system classifies true sleepwalking as a "noninsane automatism," in which the behavior is not under volitional control. The diagnosis is established by clinical history and polysomnography, although the chance of witnessing sleep terrors or sleepwalking episodes in the sleep laboratory is slim. However, polysomnography is important to exclude nocturnal seizures, which mimic sleepwalking, and certain polysomnographic features, such as repetitive arousals out of slow-wave sleep, frequent sleep stage shifts, and hypersynchronous delta activity, are often found in sleepwalkers. Nocturnal safety precautions (e.g., removal of nearby weapons or dangerous objects, alerting house guests, and assigning the sleepwalker to a ground floor bedroom) and treatment by medications (benzodiazepines and tricylic antidepressants), and occasionally hypnosis and psychotherapy, may control these conditions.

b. Other Parasomnias REM sleep behavior disorder is a condition that may be confused with sleepwalking. However, individuals with this disorder are typically men older than the age of 60 years and they frequently "act out" their dreams during REM sleep (as opposed to sleepwalking, which occurs during slow-wave sleep, typically without dream mentation). The etiology of the majority of cases is unknown,

although vascular, degenerative, and neoplastic conditions affecting the brain may be associated with this condition. The diagnosis is established by history and polysomnography, and nocturnal safety precautions combined with medications (clonazepam and desipramine) are usually effective in decreasing the frequency of this condition.

Rhythmic movement disorder is a disorder of infancy or early childhood and is characterized by repetitive movements, typically of the head and neck, that occur prior to sleep onset and in the transition to sleep. Typical movements are repeated banging of the head on the sleep surface, rolling of the head, or rocking of the torso or entire body; polysomnography can be useful in distinguishing this condition from epilepsy. Nocturnal safety precautions combined with behavior modification and benzodiazepine or tricyclic medications may be useful in controlling this condition.

The following parasomnias are specific phenomena witnessed by the patient, bed partner, or parent and, as with other parasomnias, can often be confirmed with a sleep study. Sleep-related abnormal swallowing syndrome is a rare condition characterized by awaking with coughing or choking from sleep as a result of aspirated saliva, which in turn may be due to deficient swallowing. Nocturnal paroxysmal dystonia manifests as stereotyped, repetitive dystonic or dyskinetic episodes lasting up to 1 hr per episode, with a possible familial pattern but without identifiable pathology. Anticonvulsants such as carbamazepine may be helpful in patients with this condition. Sudden unexplained nocturnal death syndrome is the occurrence of sudden death during sleep, particularly in healthy young Southeast Asian men. Abnormal respiratory symptoms such as gasping or choking may precede the cardiorespiratory arrest, although postmortem examinations have failed to reveal significant pathology in the victims. REM sleep-related sinus arrest is characterized by repeated periods of asystole during REM sleep in otherwise healthy, typically young adults. No significant cardiac pathology has been identified in these cases, and polysomnography is essential in documenting this phenomenon. Impaired sleep-related penile erections (organic/physiologic impotence or the failure to sustain erections during sleep) and sleep-related painful erections are most frequently found in patients older than age 40, and in the case of impaired sleep-related penile erections, polysomnography is critical in evaluating this condition.

A few parasomnias are observed during infancy. Infant sleep apnea is characterized by central or obstructive apneas occurring during sleep in infants; apnea of prematurity afflicts infants younger than 37 weeks of gestation, apnea of infancy afflicts infants older than 37 weeks, apparent life-threatening event associated with an apnea is typically accompanied by cyanosis and limpness, and obstructive sleep apnea syndrome involves repetitive upper airway obstruction. Polysomnography is essential to the diagnosis of these conditions, and continuous positive airway pressure or upper airway surgery may be necessary to treat these infants. Congenital central hypoventilation syndrome, or failure of the control mechanisms of breathing, is a rare condition affecting otherwise normal-appearing infants without identifiable lung disease and may require nocturnal or 24-hr ventilatory support. Benign neonatal sleep myoclonus is a rare, self-limited, and benign condition without identifiable pathology, usually involving the arms and legs and consisting of frequent jerking movements during sleep in neonates. Sudden infant death syndrome (SIDS) is characterized by unexplained death in infants, usually while asleep. It is estimated to occur in less than 1 of every 1000 live births, and factors such as sleeping prone, preterm birth, multiple births, affected siblings, maternal drug abuse, tobacco smoke exposure, lower socioeconomic status, and prior history of apneic events increase the risk for occurrence of SIDS.

Sleep paralysis (at sleep onset or upon awakening), confusional arousals (brief periods of disorientation following arousals from sleep), sleep talking, sleep starts (sudden contractions of the body at sleep onset), nocturnal leg cramps, nightmares, bruxism (teeth grinding), enuresis (bedwetting), and primary snoring are typically more nuisances rather than clinically significant problems, unless the episodes are frequent, severe, and/or disruptive to the patient or bed partner's sleep.

B. Past Medical History

Significant medical, neurologic, and psychiatric problems must be thoroughly evaluated, since these conditions typically adversely affect sleep. Medical problems such as nocturnal asthma, chronic obstructive pulmonary disease, cardiac ischemia, and peptic ulcer disease disturb sleep; neurologic conditions such as Parkinson's disease and other dementing illnesses, neuropathies, degenerative arthritis of the spine and joints, seizure disorders, and migraine headaches similarly disrupt sleep. Although behavioral and pharmacologic treatments, may be beneficial in

reducing the symptoms of these conditions, close monitoring and treatment of the patient's underlying medical problem by the patient's internist often helps the patient's sleep. Pierre Robin, Treacher Collins, hypothyroidism, Cushing's syndrome, acromegaly, and Marfans syndrome are disorders that have been associated with obstructive sleep apnea syndrome. In the case of Pierre Robin and Treacher Collins, corrective surgery by maxillofacial and oral surgeons may be necessary to correct the craniofacial abnormalities predisposing the patient to obstructive sleep apnea syndrome. A history of surgery to remove lower teeth (indicating a small mandible), tonsillectomy and adenoidectomy, nasal septoplasty, jaw reconstruction, cleft palate repair, or bariatric surgery (for weight loss) are important factors to note for the patient suspected of having obstructive sleep apnea syndrome. Closed-head injury may result in hypersomnia or, rarely, narcolepsy. A history of uremia or low iron stores are significant in a patient with suspected restless legs syndrome or periodic limb movement disorder; awareness and treatment of these predisposing factors may improve the latter two sleep disorders.

Depression and bipolar disorder are frequently associated with patients who present with excessive daytime sleepiness or insomnia; a patient with seasonal affective disorder may frequently experience an increased need for sleep and a significant delay in morning awakening times. Patients with a diagnosis of posttraumatic stress disorder, or who have experienced significant trauma, abuse, or stress in the past, may suffer from lifelong insomnia. Anxiety, panic, and psychotic disorders can severely impair sleep and result in insomnia. A team approach by the sleep specialist and psychologist or psychiatrist to effectively manage the patient's sleep disorder and coexisting psychiatric problem is frequently beneficial for the patient.

Frequently, patients with insomnia have tried a variety of nonprescription and prescription hypnotics; as discussed previously, a diagnosis of hypnotic-dependent sleep disorder is common among this patient population. It is also not uncommon for these patients to abuse stimulants, ranging from over-the-counter appetite suppressants to prescribed stimulants such as pemoline, methylphenidate, and amphetamine, in order to counter their daytime fatigue. Patients with disorders with excessive daytime sleepiness (e.g., obstructive sleep apnea syndrome and periodic limb movement disorder) may also abuse these stimulants. Medications such as those used to treat asthma and chronic obstructive pulmonary disease may independently disrupt sleep, particularly the xanthine derivatives such as theophylline. It is important to evaluate all medications and substances the patient is currently using; for example, cough medication may contain alcohol, which can fragment sleep, and herbal preparations may contain sleep-promoting substances such as melatonin.

C. Family History

There are several sleep disorders in which familial or genetic influences have been proposed. Genetic markers for narcolepsy have been identified. HLA DQB1-0602 is the most specific HLA marker associated with narcolepsy across all ethnic groups; however, this haplotype is neither sufficient nor necessary for the diagnosis of narcolepsy. There is also evidence that restless legs syndrome, periodic limb movement disorder, enuresis, sleep terrors, sleepwalking, snoring, and obstructive sleep apnea syndrome may be subject to familial influences. Thus, patients with one or both parents or siblings afflicted with these conditions may be at a greater risk for developing these disorders compared to individuals without a family history of these conditions. Physician awareness of the family history of the patient with respect to sleep disorders may help to reduce the differential diagnosis of the patient and may be helpful in special cases (e.g., if a patient being evaluated for restless legs syndrome has a son with insomnia, it would be prudent to evaluate the son for possible restless legs syndrome as well).

D. Social History

The impact of the patient's sleep problems on his or her life must be thoroughly assessed. In the case of children, school performance is frequently affected by disorders of excessive daytime sleepiness; school-authorized naps and stimulant administration during school hours may be necessary. The adult patient with excessive daytime sleepiness or a shift work sleep disorder, for example, may be on probation at work or disability. It may be necessary to become a patient advocate in the these situations, particularly when the patient's school or work performance is severely affected by the sleep disorder. Additionally, a strain on the patient's marriage may be the result of the patient's loud, disrupted snoring or decreased libido as a consequence of obstructive sleep apnea syndrome. However, if the patient is at risk for drowsiness while driving or operating heavy machinery, it is prudent to expedite the diagnosis and treatment of the underlying

sleep disorders. In cases in which the patient already has a history of falling asleep at the wheel or is at great risk for harming himself or herself as well as others while driving, the physician should notify the state public health department, which in turn will notify the state department of motor vehicles for consideration of suspension of the patient's driver's license until the patient's daytime sleepiness is effectively treated.

Heavy caffeine use can result in difficulty initiating and maintaining sleep. A patient may also use caffeine to mask excessive daytime sleepiness from other causes, such as obstructive sleep apnea syndrome. Heavy tobacco use may result in chronic obstructive pulmonary disease, which in turn may result in nocturnal awakenings due to coughing or dyspnea. Nicotine, as a stimulant, may also result in episodes of insomnia. Alcohol use frequently results in shortened, disrupted, and nonrefreshing sleep. In a patient with sleep-disordered breathing, alcohol can worsen the abnormal respiratory events, resulting in a vicious cycle in which the patient has progressively disrupted sleep and daytime sleepiness. Use of illicit stimulants, hallucinogens, and hypnotics can greatly influence sleep; chronic usage of one or more of these classes of drugs frequently results in chronic insomnia.

E. Review of Systems

A complete system review will frequently provide evidence that implicates medical problems as an underlying source of the patient's sleep disorder or may provide additional symptoms that can aid the clinician in arriving at a sleep diagnosis. Significant nasal congestion and obstruction during allergy season and recent weight gain or increased neck circumference can place the patient at risk for obstructive sleep apnea syndrome. Complaints of obstructive apneas in addition to lethargy, constipation, cold intolerance, muscle cramps and stiffness, menorrhagia, intellectual and motor activity decline, poor appetite, weight gain, dry skin, and hair loss would indicate a diagnosis of hypothyroidism. In this situation, confirmatory laboratory tests are important since the patient's sleep apnea may well resolve with thyroid supplements. Respiratory (dyspnea and wheezing) and cardiac (palpitations, angina, and nocturnal paroxysmal dyspnea) symptoms typically result in nocturnal awakenings and may suggest an undiagnosed case of obstructive sleep apnea syndrome. Gastroesophageal reflux disease, in addition to being an associated symptom of obstructive sleep apnea syndrome, may independently produce nocturnal awakenings from choking or gagging episodes at night. Pain due to conditions such as sciatica, lumbar strain, or disk disease may disrupt sleep. Sensitivity of teeth to cold or frequent temporomandibular joint pain upon awakening may indicate bruxism, which may fragment the patient's sleep. Pregnancy, uremia, and iron deficiency can be risk factors for restless legs syndrome and/or periodic limb movement disorder. Nocturia, frequently the result of benign prostatic hypertrophy or detrusor instability in the elderly, can produce nocturnal awakenings. However, nocturia can also be a common complaint of patients who awaken from sleep disorders such as obstructive sleep apnea syndrome or periodic limb movement disorder; these patients often mistakenly believe they awaken because of the urge to urinate rather than due to sleep-disordered breathing or leg movements.

F. Physical Examination

A complete and thorough physical examination is important in establishing a differential diagnosis or confirming a specific diagnosis. Measurement of the height and weight is important for determining the body mass index [BMI; weight (kg)/height (m^2)] described by Khosla; patients with a BMI higher than 25 are considered obese, although frame size and body fat percentages should be taken into consideration. This is important in the assessment of patients suspected of having obstructive sleep apnea syndrome, although it is estimated that 30–67% of patients with obstructive sleep apnea syndrome are obese. Pulse and respiration rate as well as blood pressure are important for assessing cardiopulmonary status and the impact of conditions such as sleep-related breathing disorders on the cardiac and respiratory systems.

Examination of the head is critical for evaluation of sleep-related breathing disorders. Craniofacial abnormalities, specifically a narrow, posteriorly displaced mandible and a high, arched soft palate, are characteristic of obstructive sleep apnea syndrome. In fact, a morphometric model recently developed by Kushida and colleagues enables the prediction of obstructive sleep apnea syndrome based on measurements of these typical craniofacial features. Hypertrophied adenotonsillar tissue, particularly in children, produces increased airway resistance, which in turn may result in sleep-related breathing disorders. Redundant, erythematous, and edematous peritonsillar

tissue, a long, thick uvula, and/or an obscured posterior wall of the pharynx by an elongated soft palate or an enlarged tongue are typically found in patients with snoring, obstructive sleep apnea syndrome, or upper airway resistance syndrome (a milder variant of obstructive sleep apnea syndrome). A deviated nasal septum and erythematous and edematous nasal mucosa may predispose the patient to developing sleep-disordered breathing; these conditions may also worsen an existing case of obstructive sleep apnea syndrome.

As described by Young and colleagues, a neck circumference more than 17 in. (43 cm) is associated with a higher likelihood of obstructive sleep apnea syndrome in habitual snorers, but it is by no means necessary nor sufficient for this diagnosis since the exact relationship of neck circumference to body weight is unknown and not all patients with obstructive sleep apnea syndrome are obese. Examination of the thyroid may enable the assessment of the goitrous variety of hypothyroidism, which may result in obstructive sleep apneas. Assessment of jugular venous pressure and pulses may provide additional information on the cardiovascular status of patients with suspected obstructive sleep apnea.

Examination of the chest and lungs by ascultation is important to detect wheezing and airflow limitation, which may be diagnostic of sleep-related asthma or chronic obstructive pulmonary disease, both of which result in nocturnal awakenings. A cardiac impulse that is increased in amplitude, prolonged in duration, and located at the third, fourth, or fifth interspace, and also subxiphoid, can be found in patients with right ventricular enlargement secondary to pulmonary hypertension, a late cardiac sequela of obstructive sleep apnea syndrome. A right-sided fourth heart sound and a pulmonic ejection sound best heard in the second and third left interspaces are also characteristic of right ventricular enlargement.

Abdominal examination may reveal hepatomegaly, suggestive of alcohol-dependent or toxin-induced sleep disorders. Focal or diffuse muscle tenderness may be indicative of fibromyalgia, which may produce chronic fatigue and nonrefreshing sleep. Degenerative joint disease, resulting in tenderness and inflammation of the joints as well as decreased range of motion, may result in disrupted sleep. A digital examination of the rectum may detect prostatic enlargement for a male patient who reports nocturia. In the case of a male patient who presents with a complaint of impotence, baseline measurement of the penile base and tip is important for comparison if the patient undergoes a nocturnal penile tumescence test, an important tool for distinguishing physiologic versus psychologic causes of impotence.

A mental status examination can assess short-term memory and attention, which are impaired in sleep-deprived subjects, such as those with sleep-disordered breathing or periodic limb movement disorder. An anxious, manic, or depressed mood state can assist in the diagnosis of an underlying psychiatric disorder affecting sleep. Speech and sensory changes may underlie chronic substance abuse. Alterations in coordination, muscle strength, and reflexes, in addition to mental status changes, may indicate a degenerative or dementing illness. Back pain and tenderness may be an important cause of insomnia; these symptoms combined with leg pain, paresthesias, and weakness may be indicative of a peripheral neuropathy, which may also awaken the patient at night. Focal neurologic deficits combined with a history of unusual movements at night may indicate a lesion or stroke resulting in nocturnal seizures.

G. Diagnostic Tests

Diagnostic tests are frequently useful in confirming the patient's preliminary diagnosis. Polysomnographic testing can be used to confirm the presence of a sleep disorder, particularly those that fragment sleep, such as obstructive sleep apnea syndrome, sleep terrors or sleepwalking, nocturnal epilepsy, or periodic limb movement disorder. This test can also be used to determine the severity of the condition by the frequency of occurrence of the abnormal events per recording night or per hour of recorded sleep. The subject is connected to electrodes that measure activity of the brain, eyes, heart, and chin and leg muscles in addition to snoring intensity, oronasal airflow, chest and abdominal impedance, and oxygen saturation throughout the night. Esophageal manometry is useful in measuring transmitted intrathoracic pressure or the work associated with breathing; these pressure measurements can aid in the diagnosis of mild sleep-related breathing disorders, such as upper airway resistance syndrome.

Assessment of the level of daytime sleepiness can be objectively measured by tests such as the multiple sleep latency test, in which a subject's tendency to fall asleep is assessed with polysomnography during five daytime naps lasting 20 min each, with naps separated by 2 hr. This test can also be used to confirm a diagnosis of narcolepsy, in which the mean sleep latency (average of

the time from lights out to sleep onset for the five naps) is in the pathologic range of sleepiness (less than 5 min) and REM sleep is present in two or more of the naps. A urine toxicology screen is obtained during the day of testing to detect any medications or substances that may influence the test results. A maintenance of wakefulness test uses the same equipment and schedule with the exception that the subject is sitting and instructed to resist sleep. Pupillography and continuous polygraphic monitoring are other methods that are less commonly used to assess daytime sleepiness, and vigilance tests such as the psychomotor vigilance task are used by some investigators.

Lastly, other tests used by sleep specialists to aid in the diagnosis of specific sleep disorders include pulmonary function tests, for patients with nocturnal asthma or chronic obstructive pulmonary disease; thyroid function tests for patients with obstructive sleep apnea and symptoms of hypothyroidism; renal function tests and ferritin levels, useful in identifying causes of restless legs syndrome; HLA typing, helpful in a diagnosis of narcolepsy, using markers such as HLA-DQB1-0602; actigraphy, or detection of activity typically by a wristwatch-like device, which is used in evaluating insomnia, misperception of sleep states, or circadian rhythm disorders; video monitoring and sleep-deprived electroencephalography to aid in the diagnosis of nocturnal seizures; and nocturnal penile tumescence test to distinguish between physiologic and psychologic impotence.

II. CONCLUSIONS

Sleep medicine is a relatively new subspecialty but has experienced steady growth in numbers of patients, physicians, and sleep centers. The number of recognized sleep disorders and the high prevalence of these disorders in the general population attest to the fact that the appropriate diagnosis and management of a patient presenting with a sleep complaint are far from trivial. Nevertheless, although a patient presenting with a sleep complaint is frequently a diagnostic challenge, successful evaluation and management of the sleep disorder, the majority of which are treatable, typically results in a dramatic improvement in the patient's quality of life.

See Also the Following Articles

CIRCADIAN RHYTHMS • DREAMING • HOMEOSTATIC MECHANISMS • RESPIRATION • STRESS

Suggested Reading

American Sleep Disorders Association (1997). *International Classification of Sleep Disorders, Revised: Diagnostic and Coding Manual*. American Sleep Disorders Association, Rochester, MN.

Douglass, A., Bornstein, R., and Nino-Murcia, G. (1986). *Sleep Disorders Questionnaire*. Univ. of Michigan Press, Ann Arbor.

Guilleminault, C., Mignot, E., and Grumet, C. (1989). Family study of narcolepsy. *Lancet*. **11,** 1376.

Hoddes, E., Dement, W. C., and Zarcone, V. (1972). The development and use of the Stanford Sleepiness Scale (SSS). *Psychophysiology* **9,** 150.

Johns, M. W. (1991). A new method for measuring daytime sleepiness: The Epworth Sleepiness Scale. *Sleep* **14,** 540–545.

Khosla, T., and Lowe, C. R. (1967). Indices of obesity derived from body weight and height. *Br. J. Prevention Soc. Med.* **21,** 121–128.

Kushida, C. A., Efron, B. E., and Guilleminault, C. (1997). A predictive morphometric model for the obstructive sleep apnea syndrome. *Ann. Internal. Med*. **127,** 581–587.

Kushida, C. A., Guilleminault, C., Dement, W. D., Simon, R. D., Jr., and Ball, E. (1998). Diagnostic and treatment strategies for the obstructive sleep apnea syndrome. *JCOM* **5,** 49–65.

Young, T., Palta, M., Dempsey, J., Skatrud, J., Weber, S., and Badr, S. (1993). The occurrence of sleep-disordered breathing among middle-aged adults. *N. Engl. J. Med*. **328,** 1230–1235.

Young, T., Palta, M., and Badr, S. (1992). Sleep-disordered breathing (letter to the editor). *N. Engl. J. Med*. **329,** 1429.

Spatial Cognition

BARBARA LANDAU
Johns Hopkins University

GLOSSARY

dead reckoning A process by which small changes in position are summed as an animal moves through space, resulting in continual updating of the animal's representation of its current position.

neglect A neurological syndrome in which the patient ignores or fails to acknowledge the presence of stimuli in a region of space, usually contralateral to the site of the person's brain lesion.

reference system A set of orthogonal axes with a common origin, used to specify the location of some target. Examples include location in retina-centered, body-centered, and environment-centered reference systems.

Spatial cognition is the capacity to discover, mentally transform, and use spatial information about the world to achieve a variety of goals, including navigating through the world, identifying and acting on objects, talking about objects and events, and using explicit symbolic representations such as maps and diagrams to communicate about space. Although some functions of spatial cognition such as language and mapping use are exclusively human, other aspects of spatial cognition can be found in all mobile species. Growing evidence suggests that the multiple functions of spatial cognition in humans may be represented in specialized subsystems of the mind and brain. Such specialization raises important questions about how different subsystems interact with each other in order to support our capacity to act on and communicate about the spatial world. Furthermore, these issues raise important questions about how spatial cognition develops in young children (e.g., whether specialization is present from the earliest points of development and how the specialized subsystems come to interact with each other). This article discusses spatial cognition in the context of specialization, interaction, and development.

I. CENTRAL ISSUES AND CONCEPTS

Perhaps the most fundamental capacity of all mobile species is their negotiation of the spatial world. For nonhumans, the range of activities that engage some aspect of physical space includes locating and gathering food, searching for mates, reproducing and caring for young, and virtually all other activities fundamental to survival. Each of these requires that the animal obtain and store spatial information about the world that will be used at some later point to make predictions that will allow the animal to generate some new activity in space. The form of this spatial information in the mind and brain is referred to as a *spatial representation*. Spatial representations support our capacity to perceive, act upon, and think about space. Human spatial activities include the ones already discussed as well as many other functions that are species specific. Chief among these are the use of language and external physical representations of space (such as maps) to communicate about space.

A. Preliminary Arguments for Specialization and Interaction

The three-dimensional physical world in which we live serves as the stage on which the diverse set of spatial

activities is carried out. Because of this, each aspect of spatial cognition must map onto one and the same physical reality. However, the functional requirements of spatial activity differ considerably across domains, raising the question of whether spatial cognition draws on one monolithic set of spatial representations or on multiple separate representations that are specific to different domains. This article reviews growing evidence suggesting that specialized spatial representations underlie different kinds of spatial activities. For example, in order to identify an object, we must represent the parts of the object and their spatial relationships, and this spatial structure is invariant as observers and objects move through space. Thus, a representation that preserves the enduring shape of the object would be useful for object recognition. Such a representation of the object's enduring shape may also be useful for object naming; however, naming also requires that we group together under the same name objects that vary in their precise shape. Thus, the visual system might compute the shape of a particular dog in order for recognition to proceed, but the object name "dog" will be used as a cover term for many objects whose shape commonality is quite abstract because it must cover, for example, both Dalmatians and Pekinese dogs. In contrast with either object recognition or naming, the spatial system that supports our capacity to reach and grasp the object that we see might have quite different functional requirements: Depending on our current posture, and whether the object is moving, the spatial system guiding action will have to account not only for the object's enduring shape but also for its current disposition relative to the moving effectors of our body (hands, arms, torso, etc.). The different functional requirements of recognizing objects vs acting on them suggest the existence of separate representational systems dedicated to each (see Section II.B). Similarly, the different functional requirements of object recognition vs naming suggests only a partial mapping between these two systems. Navigating through space may have different constraints, and accordingly, some have suggested that it is guided by a highly specialized representational system (see Section II.A).

Thus, functionally different requirements may give rise to differences in the form of spatial information that is preserved by spatial systems across different domains, such as perception, action, and language. However, these differences must also be reconcilable: In order for us to recognize and name the objects we can reach or to talk about the spatial layout of an environment through which we have moved, there must be some connection or mapping among the different kinds of spatial representation. Such mapping among systems is necessary in order for the spatial systems to communicate with each other as information is passed between them.

Both specialization and interaction are fundamental aspects of human spatial cognition that emerge early and easily, with little formal tutoring. For example, infants learn to reach and grasp the objects they see, and toddlers learn to name those objects. Toddlers learn to navigate their environments, and early productive language is dominated by talk of where things are in space. By the age of about 3 years, children can use external representations of space (such as simple physical maps) to locate objects in a corresponding space. This is long before "map use" is formally taught in school. This evidence suggests that the multiple systems of spatial cognition emerge as a consequence of native biases to represent spatial information in particular ways. These native biases provide the foundation for early development of spatial cognitive systems. At the same time, they leave room for further developments that depend on particular kinds of experience that may modulate the foundational structures. Variations in early experience that can modulate foundational structures include congenital deafness or blindness and brain damage due to certain genetic syndromes or environmental insult (see Section IV).

If the precise nature of spatial representations differs considerably across domains, then what holds them together? Two important organizing ideas pervade the literature on spatial cognition and permit us to think across domains to the kinds of properties that are shared by (and differ across) the different domains. One concerns those geometric properties that are preserved in our spatial representations. The second concerns reference systems, which serve as further organizers of geometric properties.

B. The Geometries of Spatial Representation

We have already observed that spatial cognitive systems must resonate to spatial aspects of the physical world in which we live. However, the physical world can be formally described in terms of many different kinds of spatial–geometric properties. The critical question is which of these many different geometric properties best captures our mental representations of space.

Several theorists have drawn on formal theories of geometry as the framework within which this question

can be addressed. Two key examples of this approach can be found in work by Jean Piaget, whose goal was to describe the development of spatial cognition in the child, and that by C. R. Gallistel, whose goal has been to describe the representational system guiding navigation in mobile species. Both have drawn on Felix Klein's Erlangen program, which considers geometries in terms of the sets of properties left invariant under transformation. Different geometries are characterized by different kinds of transformations and thus leave different sets of properties unchanged. If an organism represents the geometric properties left invariant in a particular geometric system, then it can be said to have a representation that can be formally characterized by that geometry.

An intuitive sense of what this means can be illustrated using the example of a simple shape, for example, say a 2-in square (Fig. 1). By definition, a square is a figure composed of four equal sides; each pair of sides intersects at a 90° angle. The properties of this geometric object—its four equal sides (each 2-in.), and its four 90° angles—are left invariant under certain geometric transformations but not others. For example, if we translate, rotate, or reflect the square, the lengths of its sides and their angles of intersection will remain the same. Thus, the geometric properties of distance and angle remain invariant under the transformations of translation, rotation, and reflection: The 2-in. square that has undergone these transformations can be said to be the "same" square with respect to each transformation. These are part of the Euclidean metric geometry, and so our representation of the 2-in. square can be defined within that geometry.

However, the properties left invariant under Euclidean transformations are not necessarily preserved under the transformations of other geometries. For example, topological geometries preserve only quite general geometric properties, such as the closedness or openness of a segment (contrast a closed loop with a curved but nonclosed line), but not properties such as distances or angles. Topological geometry is often called "rubber sheet" geometry because the properties that it preserves are those that would be retained if one drew a figure on a rubber sheet and then applied nonuniform stretches to the sheet. If we did this with our 2-in. square, the result would be a range of figures that are enormously distorted, as shown in Fig. 1. Among the only geometric properties that would be preserved under such a transformation would be the closedness of the segment (i.e., the edges of the original square). The only way that this property could be destroyed would be if the rubber sheet were cut. For example, a single cut could produce an open line formed from the (closed) edges of the original square.

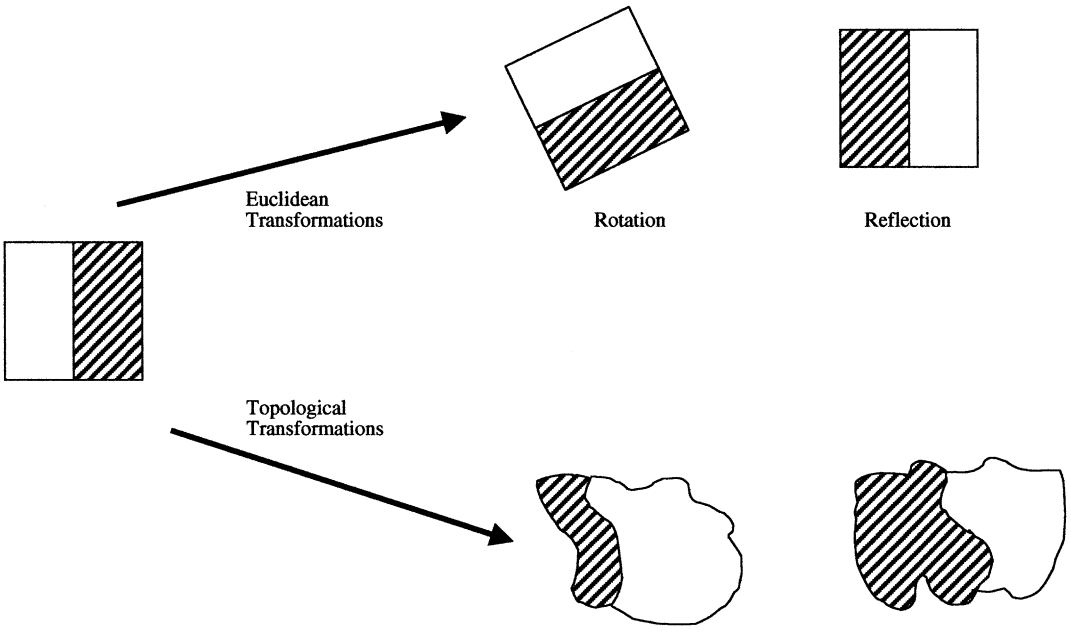


Figure 1 Two different kinds of geometrical transformation. Under Euclidean transformations such as rotation or reflection, the lengths of the figure's sides and its angles of intersection are preserved. However, under topological transformations, the figure's size and shape may be destroyed, preserving only highly general geometric properties such as whether the figure is closed or open.

Table I
Examples of Geometries and the Properties That Are Preserved under Each[a]

Geometry	Properties
Metric	Linear distance
	Angular distance
	Congruence of line segments
	Congruence of angles
Projective	Straightness (of "curves") (i.e., the property of being a line)
	Collinearity
	Type of conic section (ellipse, hyperbola, or parabola)
	Betweeness
Topological	Concurrence of curves at a point
	Openness/closedness of curve

[a]The geometries are hierarchically related to each other, with each successive geometry preserving all of the properties of the first plus new ones. For example, topological geometry preserves only highly general properties such as whether a figure is closed or open, projective geometry preserves these properties plus straight lines and collinearity, and metric geometry preserves these properties plus angles and distances.

Thus, although we would not normally consider the original square to be equivalent to a radically distorted closed figure, these two figures are, in fact, topologically equivalent.

Euclidean geometry and topological geometry illustrate two extremes in the set of geometries that Klein described, with the metric geometry preserving the most and most specific properties and the topological preserving the fewest and most general properties. Within the entire set, the geometries are hierarchically related to each other. This fact allows us to conduct strong tests of the nature of the geometric representation guiding a particular aspect of spatial cognition (Table I). As we move from topological to metric geometries, each successive geometry preserves all the properties of the preceding geometry plus some new ones. Thus, for example, whereas topological geometry preserves openness/closedness of segments, projective geometry includes this property plus straight lines and collinearity. Farther up the hierarchy, metric geometry includes all properties preserved under the others plus the new properties of linear distance (or length) and angular separation of lines.

The importance of Klein's Erlangen program for understanding the nature of spatial cognition derives from the fact that it provides a formal, testable theory that can be used to understand the nature of spatial representations. Sections II.A and II.C discuss evidence within this framework. Briefly, the evidence for both nonhuman species, such as rats, and human adults and young children shows that the spatial representations underlying navigation preserve metric properties of space. Moreover, evidence on young children shows that metric properties are preserved in early use of simple maps. Finally, evidence on spatial language in children and adults suggests that natural languages may preferentially encode nonmetric aspects of space in their basic vocabulary, using specialized vocabulary to fill in metric details. Questions about the links between different spatial cognitive systems can profitably be organized around the issue of how different geometric properties can be converted or translated into others.

C. Reference Systems

The notion of a *reference system* is the second crucial aspect of spatial representation that is pervasive in discussions of spatial cognition. The representation of location in terms of reference systems is common to many different spatial domains; hence, it is likely to be crucial to our understanding of how spatial representations come to be coordinated across domains. In this article, reference systems are defined as a set of orthogonal axes with an origin. Note that this kind of representational system requires preservation of the geometric properties "straight line" and "intersection" as well as the properties of "angle," which define the intersection of the axes. Within such a reference system, a target may be located by a precise set of coordinates, representing its distance and direction (or angle) from the origin. It is also possible for a target's location to be defined in terms of more general regions within each quadrant or within each half (vertically or horizontally divided). In much research, the distinction between reference systems and coordinate systems is not made carefully, and it is important to note that much of the evidence supporting use of a particular reference system does not necessarily also provide support for use of a coordinate system. It is an open question whether all spatial cognitive functions that engage reference systems also represent location in terms of metrically specified coordinates for distance and angle.

Given that spatial reference systems are crucial organizers of spatial information, how are their origins

determined? Objects have definite locations in space, but as we move our eyes, head, hands, or body through space, the location of the object will shift relative to our perceptual and motor effectors. Therefore, it is critical to specify the reference system in terms of its origin or the location that serves as the point of intersection of the axes. Axes and origins can be centered on a wide range of different spatial regions, and each reference system must therefore be defined in terms of its referent region. For example, if I observe an apple on a table in front of me, I can define the location of that apple in terms of a variety of reference systems: Its position may be defined in terms of the retina (retinocentric, with the origin at the center of the retina), the head (head-centered, with the origin at the center of the head), the body (egocentric or body-centered), the arm, or other parts of the body. I can also define the apple's location in terms of the table, other objects on the table or other objects in the world. The apple has some definite location on the table, in the room, in the building, and on the earth. In each case, a different reference system is used. This means that an object's location can only be defined relatively—with respect to a particular reference system (Fig. 2).

In principle, there exist infinitely many different reference systems—as many as there are objects that can serve as an origin. However, the ones most commonly considered in discussions of spatial cognition include retinocentric, head-centered, body-centered (or egocentric), object-centered, and environment-centered. The latter is used to refer to both local environments (such as rooms, buildings, or campuses) and larger environments, such as the earth, for which we can specify coordinates in a north/south/east/west reference system.

The logical requirement for reference systems in defining locations does not answer the question of how such systems are represented by the brain or, indeed, even whether the notion of a reference system as a set of Cartesian axes with coordinates is psychologically meaningful across all tasks and all domains. Moreover, one might wonder whether the entire range of different kinds of reference systems—centered on the retina, the eye, the head, the body, etc.—is actually deployed in our spatial representations. Evidence from a wide range of areas suggests that most of the reference frames commonly discussed are indeed represented in the mind and brain at some level.

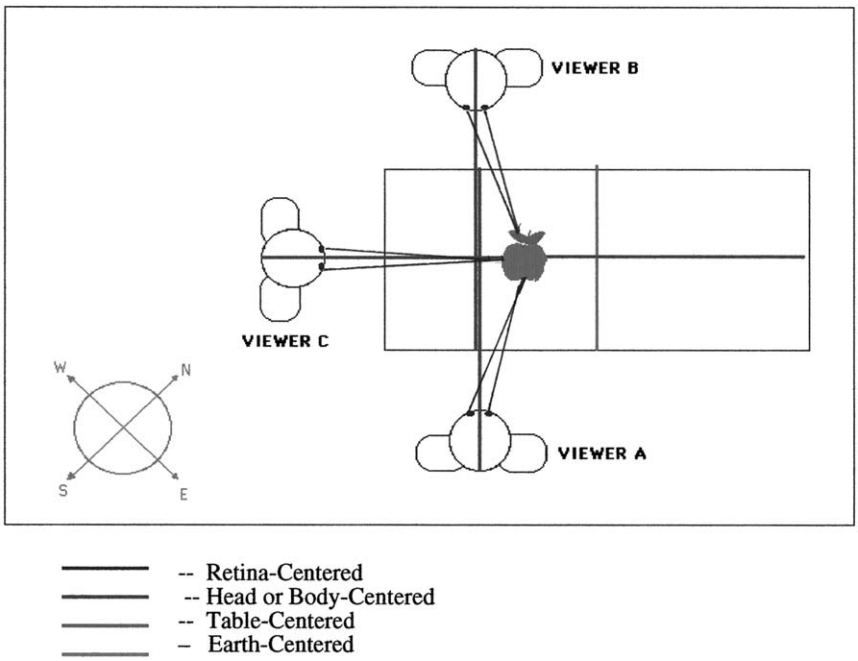


Figure 2 Where is the apple? The apple is in front of viewer A in a retina-centered frame of reference, to the viewer's right in a head- or body-centered frame of reference, and on the left side of the table from his viewpoint. For viewer B, the same apple is in front of him or her using a retina-centered frame of reference, to his or her left in a head- or body-centered frame of reference, and on the right side of the table from his or her viewpoint. The locations are defined differently for viewer C. The apple's location can also be defined in terms of reference systems centered on the room, the building, the earth, etc.

1. Multiple Spatial Reference Frames in Perception and Imagination

One source of evidence is neuroscientific studies showing that the brain represents space in terms of multiple reference systems. For example, studies of alert monkeys performing spatial tasks show that certain neurons in one area of parietal cortex respond selectively to stimuli in receptive fields defined by a head-centered system, whereas neurons in an adjacent area respond selectively to stimuli in receptive fields defined by an eye-centered system. Moreover, certain of the latter neurons are responsible for updating the location of the stimulus over eye movements. They are activated at the start of saccades that bring the receptive field onto the location of the stimulus or when a saccade brings the receptive field to a location that was previously occupied by the stimulus. The latter shows that the neurons essentially update the location of the stimulus from initial to final eye position, whether the stimulus is present or a recent memory.

Studies in human neuropsychology also reveal that the brain represents location in terms of multiple reference frames. Pertinent evidence comes from the study of neglect, a syndrome in which the patient ignores or fails to acknowledge the presence of stimuli in the region of space contralateral to the lesion. Neglect patients often show lesions in the parietal lobe, especially in the right hemisphere. Thus, the patients often neglect the "left" region of space. In one classic task used to diagnose neglect, patients are shown a sheet of paper with a large set of randomly oriented lines. When they are asked to cross out all the lines, patients often cross out only the lines on one side of the paper (e.g., the right side), leaving the lines on the left side intact. Thus, the patients are said to neglect the left side. However what does "left" mean? In the context of the current discussion, it should be clear that left must be defined relative to a reference frame in order to be meaningful: Is the region left relative to a retina-centered frame of reference, a head-centered frame, a body-centered frame, an object-centered frame, and so forth?

Investigators have found that neglect can affect different kinds of reference frames. For example, Edoardo Bisiach and colleagues described a neglect patient who was asked to imagine himself in a salient location in his hometown (e.g., on the steps of a church in the center of a piazza) and then was asked to describe the rest of the piazza. The patient described only the buildings and landmarks that were on one side of the piazza, omitting those that were on the side of his body contralateral to his lesion. Then he was asked to imagine himself at a location, at the other end of the piazza, facing the opposite direction, toward his original location, and he was again asked to describe the piazza. From this location, all buildings originally to his right (hence included in the description) would now be to the left of his body midline in the layout and vice versa. The patient now tended to describe the buildings and landmarks that he had neglected in the first condition; that is, he again described only those buildings and landmarks imagined on his "right" side, omitting those now imagined on his left side. These striking observations show that neglect affects not only the current perception of space but also remembered space, and that it can affect locations defined in a reference system that is centered on the body, head, or retina but not the environment as a whole.

Other studies have shown that neglect can affect object-centered representations. For example, Marlene Behrmann and Steven Tipper asked neglect patients to detect a flashing light presented at one of

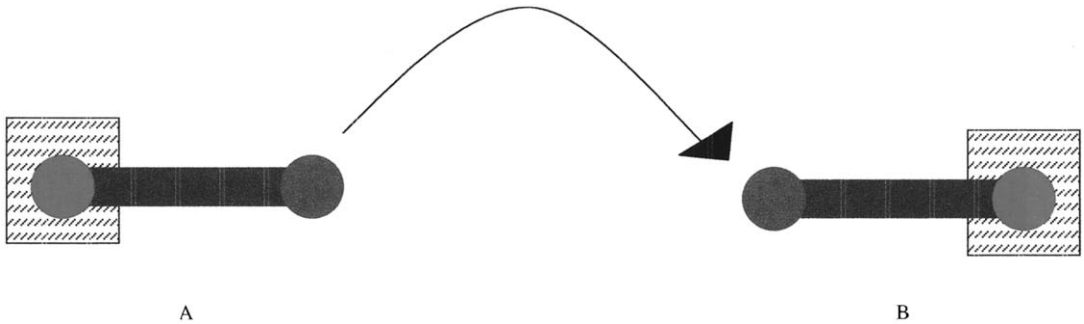


Figure 3 Behrmann and Tipper asked neglect patients to detect a flashing light presented at one end of an object both before and after rotating the object (A and B, respectively). Patients showed neglect for the same side of the object, indicating that they represented the location of the flashing light in terms of an object-centered frame of reference. (Reprinted by permission from Behrmann and Tipper, 1994, *Nature* **370,** 57–59, copyright 1994, Macmillan Magazines Ltd.)

two ends of a single object (a "dumbbell"; Fig. 3) that was presented in the center of their visual field. Patients showed the typical neglect for the side that was currently contralateral to their lesion. Behrmann and Tipper then rotated the dumbell around the center-point, with the result that the side of the object that had been neglected now moved to the other side of the viewer's visual field. Patients continued to neglect the exact *side* of the object that they had originally neglected, even though it was no longer in the visual field contralateral to the lesion. This result suggests that the patients were neglecting one side of the object as defined in an object-centered reference system.

Although the foregoing studies do not provide evidence for spatial coding in terms of specific coordinates within each reference system, recent studies by Michael McCloskey and Brenda Rapp have revealed the existence of coordinate representations. These investigators studied a young woman who had a developmental deficit in the visual perception of location. Although she showed no frank neurological insult, the woman systematically erred in indicating the location of visual stimuli presented on a computer screen in front of her. Her errors systematically preserved the distance of the stimulus from the origin of the reference system but not its direction. In particular, when asked to indicate the location of a stimulus, she typically erred by indicating locations that were the coordinate equivalents of the target under reflection around either the horizontal or vertical axis. Thus, for a stimulus that was located 9 in. *above* the origin, she would indicate a location 9 in. *below* the origin; if the stimulus was location 7 in. to the *left* of the origin, she would indicate a location 7 in. to the *right* of the origin. This evidence shows that the target's location was represented in terms of coordinates within the relevant reference system, preserving the distance along each axis but not preserving direction. Thus, it appears that the two parameters of the coordinate system—distance and direction—are separable in at least some visual representations of space.

2. Multiple Spatial Reference Frames in Navigation and Search

Classical studies of spatial learning in rats have documented the capacity to represent space in terms of either body-centered (egocentric) or environment-centered (sometimes called "allocentric") frames of reference. Early in the 20th century, evidence for each was gathered in the context of a debate between behaviorists (such as C. L. Hull) and cognitivists (such as E. C. Tolman), who sought evidence in support of more general learning theories. In key studies, rats were trained to run a simple maze and were rewarded at the end of the maze with food. The mazes were often simple, in the shape of a cross or a T. Initially, rats were trained to run the maze from a starting point (e.g., point A in Fig. 4) and then "turn right" at the intersection of the arms to reach the food at the end of the next arm (Fig. 4, point B). Rats learned this quite rapidly, indicating that something had been stored in memory that could guide their behavior. The key question was what had been learned. Behaviorists argued that the rats had learned a set of specific chained motor responses, which could then be run off in the context of the right set of cues (e.g., hunger and being placed in the maze) and that would result in success reaching the goal. In contrast, cognitive theorists argued that the rats had learned a "cognitive map"—a layout of the maze that allowed them to generate actions appropriate to reaching the goal, which was defined as a "place" in space. The critical tests were ones in which the rats' initial position was changed. For example, if the rats had been trained in a cross-maze, starting at point A, they could now be tested with a starting point at point C. If the rats were simply running off a sequence of motor responses, then

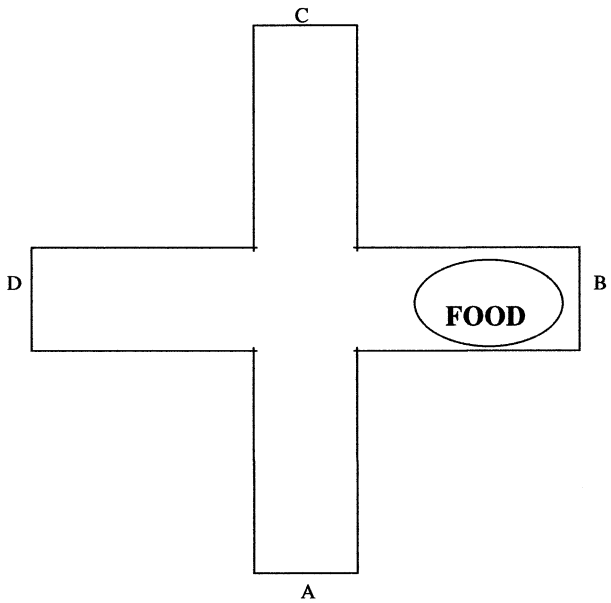


Figure 4 When rats are trained to run a maze, starting at point A and turning to find the food at B, they encode the location of the food in either a body-centered or an environment-centered frame of reference, depending on the circumstances of training and testing.

they would continue to turn right as they had done during training, now reaching D (which had no food). If, in contrast, rats were running toward a place, then they would have to turn left from their new starting point, reaching B (the location of the food).

The outcome varied depending on the particular circumstances of testing. In some cases, rats moved to the "same" location defined in terms of an egocentric, body-centered code (whereby they would continue turning right in our example); in other cases, they moved to the same location defined in terms of an allocentric, object-centered code (whereby they would now turn left from the new starting point, always reaching the correct place). In retrospect, the differences in results were due to a number of factors. One of the most critical, however, was the extent to which the animals could use information about the environmental layout beyond the maze to guide their behavior. For example, if the rats were trained and tested in a maze that was completely enclosed within a homogeneous environment (e.g., a heavily curtained space within which the maze sat) and the maze was rotated between training and testing, the animals tended to use the body-centered reference system. In contrast, if trained and tested in an open environment, such as a typically cluttered laboratory, the animals tended to run to the place where the food was located, thereby using the room layout as the reference system. Although much ink was spilled on using these different outcomes to defend different theories of learning, it appears that the best interpretation of the evidence suggests that, unless deprived of information about external layout, rats will use environmental frames of reference.

Abundant behavioral evidence exists to show that animals represent their environments in terms of both global and local layout, and that these representations guide their capacity to generate novel routes to known places in space. Moreover, work initiated by John O'Keefe and Lynn Nadel has shown that selective cells in the rat's hippocampus and surrounding regions respond when the rat is in specific places. Thus, the brain represents *places* in space. Layouts representing places are bound together as "cognitive maps," which share a number of similarities with real physical maps and are likely to be constructed by specialized mechanisms in the brain (see Section II.A). Because these cognitive maps must represent stable properties of the world through which the animal moves, they are not likely to be anchored to the body (which moves) and instead are anchored to frames of reference pertinent to the environmental layout.

Similar issues regarding use of reference frames were considered in some of the earliest studies of spatial learning among human infants. Jean Piaget proposed that the infant's earliest actions in space were confined to repetitions of sensorimotor activities that resulted in desired goals. Goal-directed spatial behavior, such as searching for objects, was thought to be guided by memories of the motor programs that had been successful in the past. A natural extension of this view was that the infant's representations of places in the world were intimately linked to the actions of its body in space—a different way of saying that place was defined in terms of the infant's body-centered reference system. According to Piaget's theory, it was only later in development that the child would come to represent object locations in a form other than a body-centered system.

Although the parallel was never sharply recognized, the formal experimental tests of this proposed developmental sequence were strikingly similar to those that had been used decades before in the animal literature. For example, in several studies Linda Acredolo and colleagues trained 6-, 12-, and 18-month-old infants to look, on cue, to either their right or left in order to gain a reinforcement, which was an animated face and voice. The environment was homogeneous, except that the left and right sides were marked with "windows" from which the animation would come. Infants rapidly learned to turn on cue to the target side. After this learning period, infants were moved along a circular path from their starting point, ending up at a point directly opposite their starting point, where what had been on their right was now on their left and vice versa. The question was whether infants would now turn their head in the same direction as before (e.g., looking to their right if trained to look to their right before) or toward the same place in space (now to their left, if originally to their right, and vice versa). Functionally, this is the same question as researchers asked regarding rats in the maze experiments. Results showed that 6-month-old infants tended to turn their heads in the same *direction* as they had before, whereas 18-month-olds turned their heads toward the same *place* as before. This pattern was interpreted as confirmation of the idea that infants initially act in space guided by egocentric frames of reference, which in this case could not take into account self-movement through space, only later coming to represent space in terms of other kinds of reference frames.

Later experiments by many investigators, however, revealed that such a developmental progression is unlikely. Rather, like the rat, the infant's and toddler's

responses to spatial tasks such as the one just described typically depend on the environmental layout available to them as well as the particular method used to test their spatial knowledge. The method used in the training study may have encouraged the development of body-centered representations because the infants underwent numerous training trials in which this kind of response resulted in an affectively positive event. In other studies, when infants are not trained extensively at an initial location, their performance at new locations improves. Moreover, in the studies carried out by Acredolo, infants were moved along a curved path, which is a combined rotation and translation, possibly obscuring their capacity to update their position over movement. Evidence shows that 6-month-old infants can update their own position in space over simple rotations (without translations) and simple translations (without rotations) but much less well over combined rotation–translations. It thus appears likely that infants have the capacity to represent places in space in terms of multiple reference systems; the choice of reference system is probably a complex function of task requirements and available information. However, the basic capacity to use different reference systems is unlikely to change substantially over development.

3. Multiple Spatial Reference Frames in Language

Multiple reference frames are represented in the mind and brain; however, different sets of reference frames are deployed for different spatial functions. Languages have the resources to encode location relative to any of the reference systems described; for example, one can describe a location as "to the left, in a *retinocentric* frame of reference." However, such words are technical, specialized terms used in restrictive settings. The basic terms of a language—those that are morphologically simple, widely used, and learned early in life—do *not* include *separate* terms that specially mark each reference system. Thus, for example, we can speak of "retina-centered," "head-centered," or "body-centered" reference systems, but in each case we refer to locations within these reference systems using the same basic set of terms, i.e., "*above*," "*below*," "*right*," and "*left*." That is, this basic set of terms is used to describe location over a number of reference systems, without further special marking. This might suggest that languages do not mark reference systems at all. However, this is false. There are two clear exceptions in English and other languages. First, locations defined in an object-centered system draw on a special set of terms—the *top*, *bottom*, *front*, *back*, and *side* of objects. These terms are used to describe locations within objects, not between objects. For example, if we wish to refer to the "*top*" of a sugar bowl, this region is the same regardless of the bowl's location or orientation in space. Similarly, the "*bottom*" is usually at the opposite end of the main axis of the object from the *top*. Terms marking these object-centered regions are seen across a variety of languages and are sometimes derived from the names of body parts (e.g., in Tzeltal and related languages). In these cases, as with the English terms *top*, *bottom*, etc., the application of the terms appears to be guided by perceptual representations of the object's shape, particularly, its main and subsidiary axes. This is an example of an interaction between two spatial systems, in this case, object perception (which also encodes axes) and language (see Section II.C).

A second set of terms specially marks location relative to the earth's coordinate system. In English, these terms are *north*, *south*, *east*, and *west*; many other languages specially mark these locations as well. Some languages also have terms describing location relative to local geographic features, e.g., *uphill*, *downhill*, or *seaward* (though the latter tend to be morphologically complex, i.e., compounds). Thus, languages appear to engage representations of a variety of reference systems, as do other spatial cognitive functions. However, locational terms in natural languages do not specially mark certain of the reference systems, even though these play a major role in action (e.g., retinocentric and body-centered).

II. SPECIALIZATION

The notion of a specialized cognitive system assumes that the structures and computations characterizing the system are unique, compared to other cognitive systems. For example, the structures underlying various aspects of spatial cognition depend on representation of geometric properties of the world, compared to the structures underlying aspects of language, which depend on representing units such as noun, verb, and subject of sentence. Similarly, the computations that allow us to manipulate geometric information might include the possible mental transformations that we can apply to our spatial representations (e.g., translation and rotation), whereas these transformations are simply irrelevant to our knowledge of language (except insofar as the two systems interact). If a system is

specialized, then it may qualify as a "module" in Jerry Fodor's sense, especially if other conditions are met. These conditions include developmentally early emergence of the system under conditions of informal (i.e., nontutored) learning, universality, and "impenetrability." The latter is the system's essential "blindness" to information that is not of the proper form for processing by the system, even though the information might be available in the environment and thus could in principle be used by the organism. One example of this is the rat's cognitive blindness to certain properties of spatial layouts that are not computed by the system dedicated to navigation (see Section II.A).

The idea of specialization in spatial cognition contrasts sharply with a different view that has recently gained currency. That is, all cognitive systems share both formal and computational characteristics, and they emerge in development through the operation of domain-general acquisition processes. Abundant evidence suggests that this latter view is wrong. Rather, many (if not all) foundational cognitive systems, including space, time, number, and systems of communication, appear to possess a rich internal structure that serves as a specialized foundation for further refinement and learning.

A. Navigation

All mobile species navigate through their environment to find food, care for young, etc. Abundant evidence shows that animals of diverse species have the capacity to construct and use cognitive maps of their environment, which are then used to guide their search in the world and return them safely home. The impressiveness of such spatial activity in animals can be conveyed with a few examples. One is the case of the desert ant, which travels many meters from home to search for food. The paths that the ant travels can be enormously long and indirect. Once the animal reaches the food, it must return to the home nest. Tracking the path of the ant on its return route, compared to its venture out from the nest, shows a striking fact: Although the initial path was long, indirect, and geometrically quite complex, the path home is direct, along a straight-line path. Experimental evidence shows that this route is not guided by "beacons" (i.e., perceptually available information emanating from the nest and serving as a continually present guide). For example, if the animal is picked up at the food source and displaced to a remote location, it will immediately move directly along a path whose angle and distance would have been appropriate for reaching its nest had it never been displaced (Fig. 5). Furthermore, when the ant has moved along the path for the appropriate distance and does not find the nest, it begins to search the area by tracing continuously enlarging circular paths whose end point falls each time in the vicinity of its initial guess.

The capacity of the ant to explore its environment and then return directly home rests on the brain's capacity to carry out "dead reckoning" (or "path integration"), a process by which small changes in position are summed as the animal moves through space, resulting in a continual updating of the animal's current position in space. This capacity is shared by many species, including humans. Recently, Gallistel proposed that spatial navigation in many mobile species rests on the capacities of dead reckoning together with "piloting," the capacity to use landmarks to correct the errors that can accumulate during the process of dead reckoning. Moreover, Gallistel

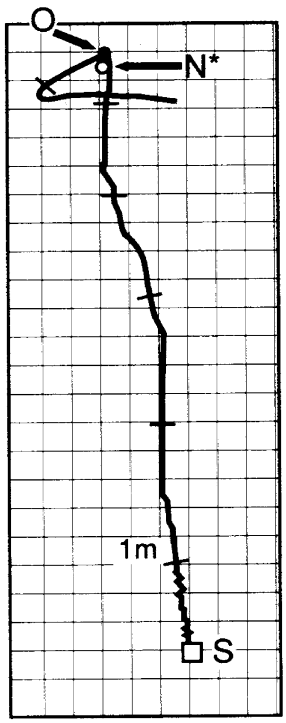


Figure 5 Homing behavior in desert ants. Ants traveled from their home nest to a feeding station. After they emerged from the feeding station, they were displaced to point S, far from their true nest. From point S, the ants traveled to their "home" nest, moving along a path that would have had the correct distance and direction had they never been displaced. The path shown is that of a single ant [N is the location of the fictive nest (i.e., where they ended up after moving along the path)]. (Reprinted with permission from Gallistel (1990)).

proposed that animals use dead reckoning to construct a cognitive map, or spatial representation in the brain, of the environments through which they move. These cognitive maps are thought to preserve the metric properties of angle and distance as well as the property of "sense"—the left–right relationships among parts of the layout. Once the map is constructed, the animal moves through space guided by dead reckoning in conjunction with this mental representation of space. Periodically, the animal checks to determine that its estimated position (based on dead reckoning) is consistent with its actual position in the environment. It does this by consulting the visible environment (i.e., salient aspects of the physical layout) and determining whether what it sees is consistent with where it thinks it currently is based on its mental map and its record of its own movements.

The capacity of the brain to represent the animal's location in space has been shown in extensive neurophysiological research. For example, John O'Keefe and Lynn Nadel found that neurons in the rat's hippocampus respond selectively when the animal is in a particular spatial location, documenting the existence of place cells and the capacity to use dead reckoning as well as information from landmarks to establish its current location. Recent work by Bruce McNaughton and colleagues has further defined the neural circuits responsible for dead reckoning.

Recent studies have investigated the possibility that the representations constructed during human navigation are also based on dead reckoning. Evidence from Jack Loomis and colleagues has shown that adults can keep track of their own location in space using mechanisms of dead reckoning. This mechanism is also available to children by the age of approximately 2 years and possibly earlier. Barbara Landau and colleagues showed that the blind and sighted toddler can represent spatial layouts in terms of angle and distance relationships and that these representations are most likely constructed by mechanisms of path integration. Figure 6 (left) shows an experimental layout in which a young congenitally blind child was walked along three paths between objects in a novel layout and then asked to travel between these objects along paths she had never traveled before. The resulting paths (Fig. 6, right) indicate that she constructed a cognitive map from her experience traveling the first set of paths, and that she was able to generate novel paths between the objects. The only information available to her to construct the map was her own guided movement during the initial walks; therefore, it is likely that mechanisms of path integration were

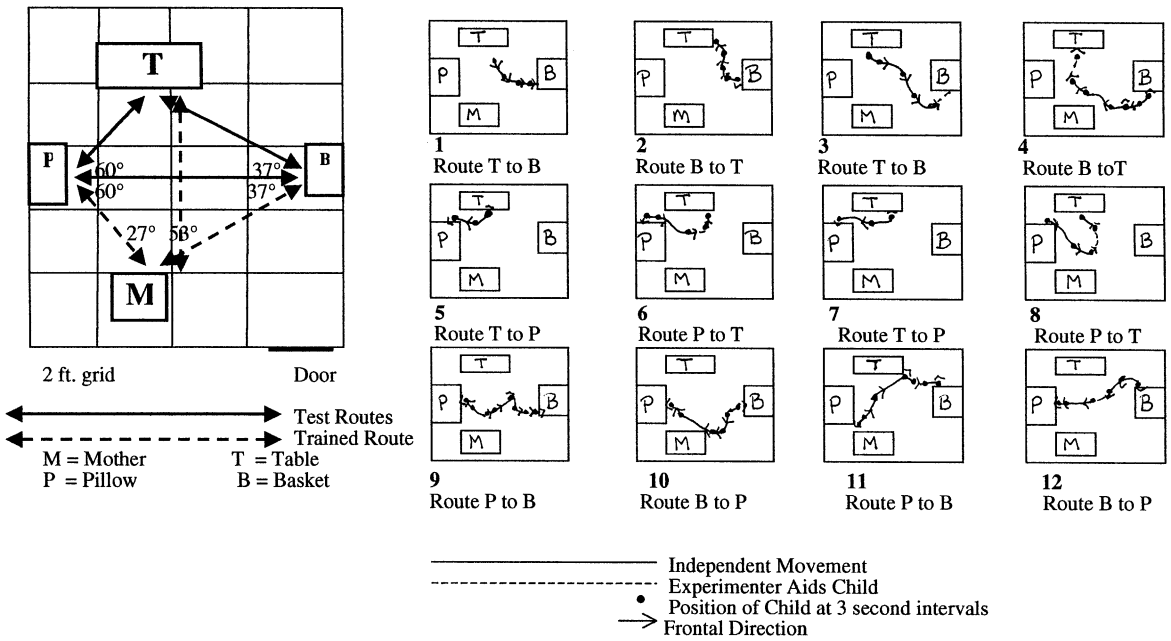


Figure 6 Layout of room used for testing the blind child's construction of a cognitive map (left). The child was trained by guiding her along paths indicated with dotted lines. She was then tested on her ability to independently move between landmarks along the test routes indicated by solid lines. Paths of movement shown in the individual panels (right) indicate that she used information from her own movement to construct a representation of the overall layout of the space and could use this representation to guide further movement among landmarks.

crucial to the development of her cognitive map of the layout. The accuracy of the novel paths that she generated immediately thereafter suggests that the cognitive map preserved both angles and distances between the pairs of landmarks.

Gallistel's theory of navigation also proposes that the representations that constitute the cognitive map qualify as a module in Fodor's sense. In particular, they are impenetrable to information that is not of the proper form for processing by the mechanisms dedicated to map construction. The hypothesized module in the animal's brain constructs a geometric map that preserves angle, distance, and sense relations using mechanisms of dead reckoning.

Behavioral evidence for the geometric module hypothesis has been found by Gallistel and colleagues in a series of elegant experiments. In one series of experiments, Cheng and Gallistel allowed rats to search for food in a homogeneous rectangular space, in which one container was placed in each corner. Each corner was distinguished by a variety of local cues, including surface textural and odor differences. Rats learned to search for the food and then were disoriented and returned to an identical but separate search space (which the rats assumed was the original, as shown by search patterns). The rats searched in the correct corner of the space on about half of the trials, searching the majority of the remaining trials in the corner that was diagonally opposite to the correct one (Fig. 7). This pattern held even when one of the walls had a distinctive covering, and it could in principle provide information critical to correct searching. Cheng and Gallistel concluded from this pattern of errors that the rat had constructed a geometric representation of the space in which sense was preserved along with distances and angles. Moreover, they concluded that the representation was modular, in that it was blind to information not properly admitted by the module, though clearly available in the environment.

In particular, if the rat's cognitive map preserved the different lengths and angles of intersection of the walls, and their sense relationships relative to each other, they would have represented the correct location of the food as something like "the corner in which the long wall is the left of the short wall." This relationship holds for two corners in the experimental spaces: the correct one and its rotational equivalent. Thus, the rat's geometric representation would lead it to search in the corner that yields a match between the location in its cognitive map (long wall left, short wall right) and the location in the current array. This will lead it to equally divide guesses about the location of the food between the two rotationally equivalent corners. A modular interpretation of the results emphasizes the fact that the additional, nongeometric cues about the target location, such as surface texture and odor, were available to the rat and used in certain contexts but were *not* used by it during search (compare Figs. 7A and 7B).

Other studies by Linda Hermer and colleagues have provided tentative evidence that the system guiding search after disorientation in humans is modular as well as metric and sense–preserving. In a modified replication of Cheng and Gallistel's experiments, Hermer tested toddlers' capacity to find a target object after disorientation in a featureless rectangular room. The toddlers, like rats, made clear errors confusing the two rotationally equivalent corners of the room (Fig. 8). In contexts in which surface information was provided to distinguish the two corners, the toddlers ignored this information (although they, like the rats, were capable of detecting it in other circumstances). The striking similarity between the pattern of responses of toddlers and rats suggests that some foundational aspects of navigation systems may be shared across species. The pattern shown by humans, however, changes by approximately 3 years of age, with increasing use of surface characteristics under conditions of disorientation. It is currently an open question whether such a developmental pattern is robust and, if so, what kinds of learning mechanisms might support the change.

B. Perceiving and Acting on Objects

In a different domain of inquiry, there is also considerable evidence in favor of specialization of spatial functions in the mind and brain. A key starting point is the work of Leslie Ungerleider and Mortimer Mishkin, who proposed in 1982 that the visual system in the macaque monkey brain is divided into two "streams," which they dubbed the "what" and "where" systems. Ungerleider and Mishkin proposed that visual information is processed beyond the primary visual cortex along two distinct anatomical pathways, each of which processes information differently. The "what" system processes information along the ventral stream of the brain that projects to the inferotemporal cortex, whereas the "where" system processes information along the dorsal stream of the brain, projecting to the posterior parietal cortex. Early evidence in support of these two streams came from

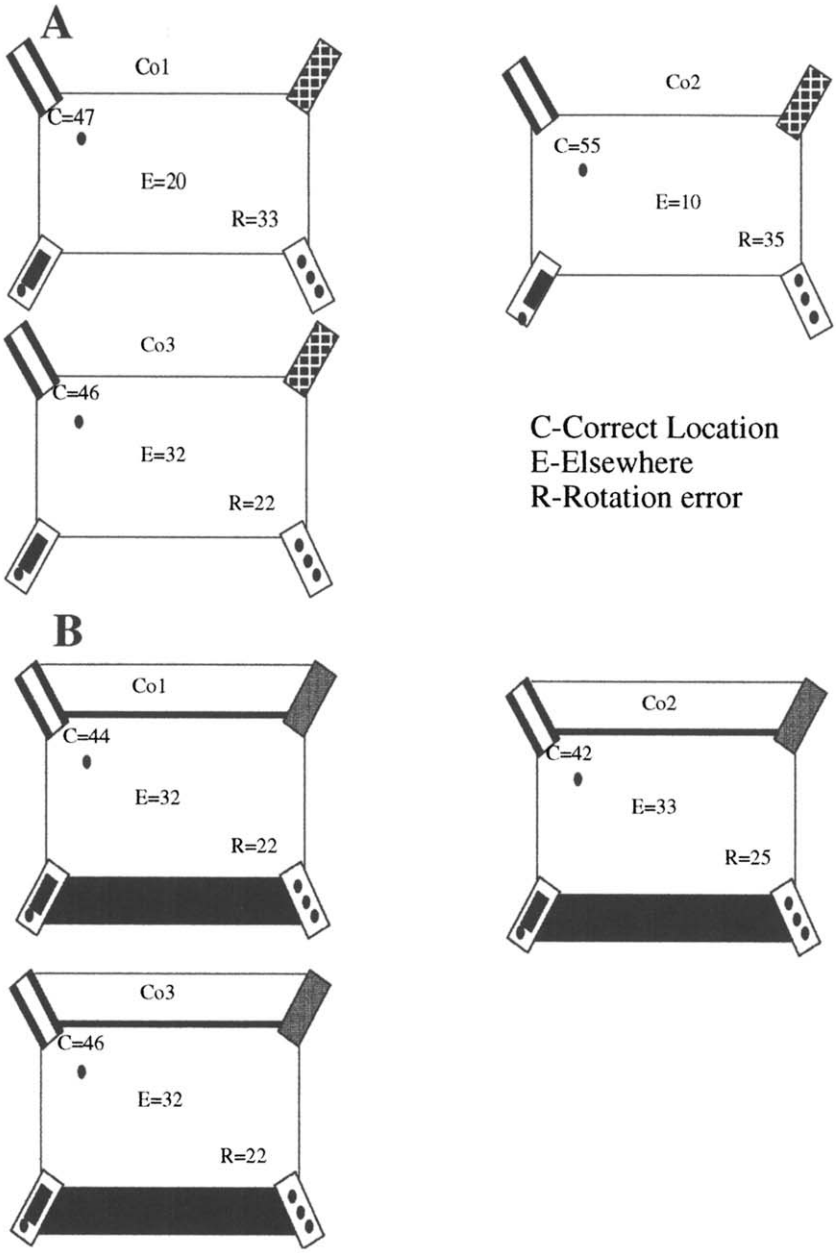


Figure 7 Results of Cheng and Gallistel's experiments on navigation in rats. (A) The search results for three rats that were tested in completely homogeneous rectangular spaces. (B) The search results for rats tested in a rectangular space that included one distinctive wall. In both cases, the rats tended to search at the correct location or its rotational equivalent. (Reprinted with permission from Gallistel (1990)).

studies of the behavioral effects of lesions in monkey cortex: Lesions of the inferotemporal cortex led to impairment in the visual discrimination of objects (on the basis of object properties such as form) but no impairment in a task requiring identification of objects on the basis of their spatial location. In contrast, lesions of the posterior parietal cortex led to impairment on object location tasks but no impairment in object discrimination based on (nonspatial) object properties. Related evidence from patients sustaining damage to corresponding locations of the brain suggests that a similar distinction between the two processing streams may exist for humans.

Recently, a different hypothesis regarding division of labor in the visual cortex has been proposed by Melvyn Goodale and A. David Milner. They suggest

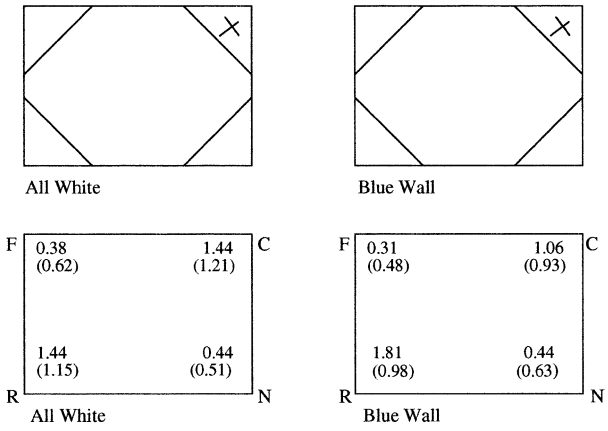


Figure 8 Results of studies by Hermer and Spelke showing that toddlers who are disoriented in space use geometric properties of the layout to search for hidden objects. Toddlers observed an object being hidden (X), were disoriented, and then allowed to search for the object. They searched primarily in the correct location or its rotational equivalent (bottom), whether or not the room had distinctive surface markings (compare right and left panels). [from L. Hermer and E. S. Spelke (1994). A geometric process for spatial reorientation in young children, *Nature* **370,** 57–59. Reprinted with permission from Macmillan Magazines Ltd.].

that the functions of the ventral and dorsal streams might better be characterized as the guidance of visual perception vs visually guided action, respectively. Specifically, the ventral system is thought to support perceptual judgments underlying object recognition and explicit judgments of location, and its contents are accessible to consciousness (required for explicit reporting). The dorsal system is thought to guide visually directed action, including reaching, grasping, and other action tasks, and the contents of this system are not accessible to consciousness. The specialization characteristic of each system means that the visual information arriving at primary visual cortex will later be represented differently; that is, different aspects of the same stimulus will be preserved.

One of the key distinctions here can be described in terms of different reference systems. As noted earlier, the action system requires that the shape, size, and orientation of an object be coded in terms of its location relative to the effectors as they move through space. As we reach and grasp, the location of our arm, hand, and fingers changes relative to the target. The action system must therefore be capable of constantly updating the representation of the target's location as an actor moves through time and space. Milner and Goodale suggest that this characteristic of the "how" system naturally entails that it will have little or no memory—there is no point in retaining information about an object's orientation and location relative to the momentary disposition of one's hand, and so this information is not retained over time. In contrast to this momentary nature of the action system, Milner and Goodale propose that the "what" system is dedicated to representing the enduring qualities of objects, including their permanent size and shape. These qualities are retained over time as a more permanent property of the object's representation.

Evidence supporting this functional distinction between the two visual streams has come both from adults sustaining brain damage and normal adults. For example, Milner and Goodale reported a case study in which they observed a patient, D.F., who had sustained brain damage due to carbon monoxide poisoning. D.F. was severely impaired in perceiving objects; for example, she could not judge the size, shape, or orientation of simple objects. However, she was quite capable of acting on objects in a way that suggests that she does represent these characteristics. In a frequently cited experiment, D.F. was asked to judge whether the orientation of a card was appropriately positioned to put through a mail slot. She was incapable of doing this accurately. However, when asked to "post" the card through the slot as if mailing it, she was quite accurate. These findings, along with complementary ones from studies of patients sustaining damage to other areas of the brain, suggest that the two streams can be dissociated under different conditions of pathology.

Earlier results from normal adults also provide support for such a dissociation between the two visual systems. For example, Bruce Bridgeman and colleagues showed that people can experience visual illusions that lead to inaccurate reports of a target location, but at the same time they can accurately point to the same location. Such findings show that the two visual systems operate in at least partial independence under normal (i.e., nonpathological) conditions.

Recently, the two-visual system distinction has come to the attention of cognitive developmentalists, in the context of the infant's representation of objects and space. Decades ago, Piaget observed a striking pattern of behavior among infants who were enticed to search for an object that disappeared from their view. Before the age of 6 months, infants show very little reaction at all, simply staring at the spot where the object disappeared or, at best, searching in inconsistent locations for it. By 9 months, infants actively search in these contexts. However, they still often search at the

wrong location, and this is often the location where they have successfully retrieved the object on past occasions rather than the location where they saw the object disappear. Piaget theorized that infants do not, at this developmental stage, understand that objects are independent entities that occupy stable locations in space and can move or be caused to move to other stable locations. Piaget proposed that this concept—part of the notion of a "permanent object"—develops slowly over the first 2 years.

Many researchers have noted, however, that the tasks Piaget used to test object permanence rely on two capacities: The infant must first observe the object being hidden, and then he or she must perform an action based on the representation of the object's location before its disappearance. In the context of the two-visual systems approach, a developmental independence between the products of perception and the system used for action could lead to a pattern in which the infant could establish and maintain a representation of a permanent object but not act appropriately on the basis of that representation. Research by Renee Baillargeon and colleagues, among others, has shown that infants can indeed represent hidden objects by the age of 4 months (and perhaps earlier). The tests depend on infants' looking patterns during and after events in which an object is hidden. For example, in one experiment, the infant observes an object that is subsequently occluded by a barrier (thus, the object goes out of view). If the occluder then continues to move, passing through the space previously occupied by the object, the infant will show surprise, indexed by lengthy looking at the event. In contrast, if the occluder stops moving at the point where it would have contacted the hidden object, the infant does not show surprise. Thus, by 4 months—considerably earlier than active and accurate search—the infant apparently knows that the object still exists when hidden. There is apparently a contrast between the knowledge shown by the infant when the perceptual system is engaged and that shown when the action system is engaged. This is consistent with the idea that the infant can represent an object's existence, but that this representation is not engaged by the action system until considerably later.

This and other findings have led a number of researchers to propose that there is a developmental dissociation between the systems guiding perception and those used to guide action in infancy. The development of action is considerably complex and, importantly, is driven by functional constraints that are different from those engaged by the perceptual systems. The fact of different developmental timetables is consistent with the idea of functional specialization, possibly corresponding to the two different visual streams.

C. Symbol Systems

The construction and use of symbolic representations of space is one of the hallmarks of human cognition, permitting the transmission of spatial information over time and space. One clear example of this capacity is the formation and use of physical maps to convey spatial information to oneself and others. Another example is the use of spatial language to talk about the objects, motions, and spatial relationships around us. The symbolic representation of space by language seems to have special priority for humans, providing a natural format for the expression of concepts and relationships that are not inherently spatial. For example, linguists such as Ray Jackendoff have shown that our spatial vocabulary is easily extended to temporal concepts: Words such as "ahead" and "behind" are predominantly spatial in meaning but can be used to express temporal meanings such as "ahead of/behind schedule." Spatial representations also provide natural means for solving problems, even those that are not inherently spatial. Thus, people seem to find it natural to solve nonspatial problems by converting the informational content of the problem into a spatial format—a fact that has significant applied importance.

1. Maps

A map is a physical layout of symbols that stands for a set of objects and places. Critically, the spatial relationships among these symbols stand for the spatial relationships in some corresponding layout in the physical space around us. The particular formal means for expressing these objects, places, and spatial relationships varies widely over history and cultures. Thus, it is clear that some aspects of map use must be learned with each new generation. However, it is also clear that certain fundamental properties are general to the very notion of a map. These properties are presumably universal among human cultures. Recent evidence suggests that appreciation of certain fundamental properties of maps emerges early in development, without explicit tutoring.

The functional requirements of maps suggest that, like other systems of spatial knowledge, only certain kinds of information will be preserved. For example,

because maps are highly condensed forms of information, they will not necessarily preserve the detailed microstructure of paths or the detailed geometry of landmarks and places. Nevertheless, in order to be used for locating objects, maps regularly preserve the angular and distance relationships among places in space—information that is also represented by our cognitive mapping system. Angle and distance are preserved in metric geometries; hence, the capacity to use a map entails the capacity to represent metric information. Because maps usually preserve metric information using a function (i.e., a scale), map use also entails the capacity to recognize the conversion from represented distances and angles (on the map) to their correspondents in real space.

Although cognitive maps and physical maps share many characteristics, being able to construct a cognitive map of a territory through navigation does not guarantee the capacity to use a real map. Many species construct cognitive maps to guide their navigation, but only humans are capable of using external physical representations to guide their locomotion and search. Careful experiments by David and Ann Premack showed that chimps cannot learn to use even a simple map to guide spatial search.

To understand the nature of this peculiarly human representational system, we can ask what kind of mental representation would be required to understand the most basic properties of physical maps. First, the map contains a set of symbols in which it is understood that each "stands for" a real object, place, or other part of the space it is meant to represent. Second, the observer must be able to represent the spatial relationships among symbols and use these to locate objects in the represented space. Angular relationships among objects are usually preserved, thereby allowing the user to compute the angular direction required in order to get from one point to another. Distance must also be represented. However, because maps are compressed representations of physical space, actual distances among objects and places are almost always altered and must be reinstated using a metric conversion. Thus, the user must be capable of understanding the "stand for" relationship between maps and the physical layouts that correspond to them; preserving angular information contained on the map and using it to predict corresponding angular relationships in the physical layout; and computing distances on the map and converting them to another metric distance, determined by the proportional relationship between map and physical space.

Knowledge of these fundamental aspects of maps emerges relatively early in life and does not appear to rely on formal tutoring. By the age of approximately 2½ years, children appear to understand the "stand for" relationship between symbols for objects and the real corresponding objects. Judy DeLoache showed children miniature three-dimensional models of real rooms, including scale furniture arranged in exactly the same layout as the represented room. She went to great pains to show the children each object and how they corresponded to each represented object. She then showed the children a small toy being hidden in the model room and asked them to find the corresponding toy, which was hidden in the same place in the represented room (Fig. 9). DeLoache found that 3-year-olds were quite accurate at finding the object in the represented room, but that 2-year-olds were quite poor, searching at chance in the represented room. After their attempted (but unsuccessful) search, children were asked to find the toy in the location where it had been hidden in the model. Both 2-year-olds and 3-year-olds performed at ceiling in this task, showing that the 2-year-olds' problem in retrieving the toy from

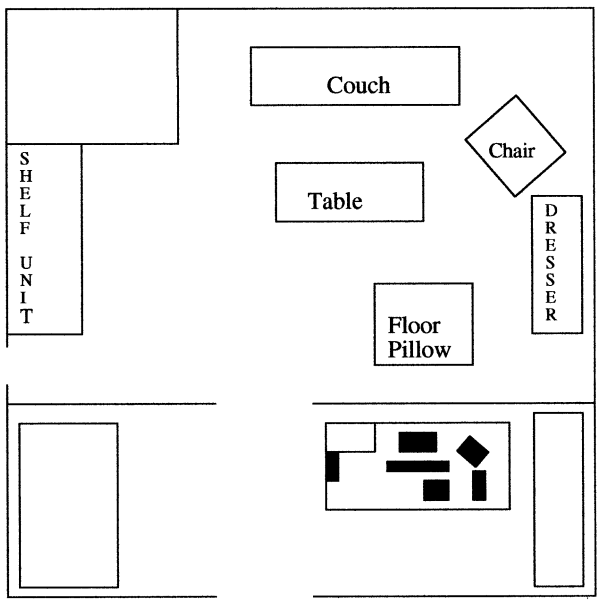


Figure 9 DeLoache showed toddlers a room and its scale model (bottom right). She hid an object in the model and then asked toddlers to find the corresponding hidden object in the real room (or vice versa). Two-and-a-half-year-olds had great difficulty using the model as a representation of the room, but 3-year-olds found the task trivial [from DeLoache, J. (1989). Young children's understanding of the correspondence between a scale model and a larger space. *Cognitive Dev.* **5,** 121–139. Reprinted with permission of Elsevier Science].

the represented room was not due to their failure to remember the hiding location in the model.

DeLoache argued that the rapid change between 2 and 2½ years was due in part to "representational insight"—the ability to recognize the relationship between a symbol and what it stands for. She also argued that the capacity to use this insight might be facilitated or dampened by changing the extent to which the children regarded the model as a symbol rather than an (nonsymbolic) object per se. In several clever experiments, she was able to confirm this hypothesis. For example, she found that, although 2½ year-olds failed to find the toy in the represented room using an explicit three-dimensional model of the room, they were successful when the model was a photograph of the room. The photograph, argued DeLoache, is a more felicitous symbol of the room because it is less like a real object. Similarly, DeLoache found that children were more successful in using the three-dimensional model when they observed it behind glass, and were not able to interact with it, than when they were allowed to touch and explore it. This too shows that young children are capable of using symbols to guide search, but only when they can best be interpreted as symbols. Thus, the ability to recognize the "stand for" relationship between map and physical layout appears early in life, well before children are formally taught about these correspondences.

Other findings show that young children are capable of using metric information, including angle and distance, to locate objects using rudimentary maps. In one study, Barbara Landau showed a 4-year-old congenitally blind child a small layout including several wooden blocks representing key landmarks and the child in a room. The child had never been exposed to tactual representations of objects or layouts. Nevertheless, the child understood the "stand for" relationship between the abstract objects on the map and the corresponding objects in the represented layout and could use the "map" to find the landmarks in the represented room. The symbols on the map were arranged to preserve angular information among objects, and the child successfully used this information to guide her search for the represented objects in space.

Young children can also use distance information conveyed in simple maps. Nora Newcombe and Janellen Huttenlocher showed 3- and 4-year-olds a piece of paper on which they had drawn a rectangle with a small dot positioned inside. The children were told that the drawing represented the location of a toy that had been hidden inside of a rectangular box of the same proportional lengths as indicated on the drawn rectangle. Children were quite successful at retrieving the toy, showing metric errors that were quite small relative to the entire space. Thus, by the age of 3 or 4 years, children are capable of representing the angular and distance relationships among objects in explicit representations of space.

2. Language

Like the construction and use of maps, our capacity to represent objects, motions, and their spatial relationships through language is species specific and emerges early in development without formal tutoring. Because spatial language must encode our spatial experience, one might expect to find exact reflections of nonlinguistic spatial representations in language. However, like other specialized systems of spatial cognition, language selectively encodes only certain properties of space.

The fundamental linguistic tools for talking about space include expressions for objects, motions, and spatial relationships. In English, objects are typically represented by noun phrases, motions are typically represented by verbs, and spatial relationships are represented by spatial prepositions (including, "in," "on," "under," and "between"). In English, there are approximately 85 spatial prepositions, which encode basic spatial relationships. By assembling these basic spatial terms in various combinations, we can construct an infinite set of descriptions that encode detailed spatial relationships (e.g., "in the middle of the front edge of the table"). Furthermore, we can add metric information to spatial descriptions by using the number system (e.g., "3 inches left of the middle of the front edge of the table").

Linguists such as Leonard Talmy and Ray Jackendoff have proposed that the basic linguistic expression used to encode a spatial relationship contains three fundamental elements: One represents the "figure" object—that is, the object that is to be located by the expression; a second represents the "ground" or "reference" object—that is, the object in terms of which the figure is located; and a third represents the geometric relationship between figure and ground. In English, these relationships are usually encoded by spatial prepositions, which are highly selective in the kinds of relationships they encode.

For example, common spatial prepositions such as "in" and "on" encode relationships that, informally, can be dubbed "containment" and "support." In fact,

these labels understate the complexity and subtlety of the relationships that are encompassed by the terms. For example, the term "in" can describe the relationship between two concrete objects, where one object sits in the cavity of another (e.g., "the apple in the bowl"), one object sits in the virtual volume created by the other (e.g., "the owl in the tree"), or one object sits completely outside of the volume of the second (e.g., "the apple in the bowl" when it is on top of other fruit and extends above the bowl's lip). In English, the term "on" can describe a relationship of gravitational support (e.g., "the jar on the table"), nongravitational support (e.g., "the gum on the table"), or transitive support (e.g., "the book on the table," when the book sits on the top of a pile of other books). The very same physical array can be aptly described by more than one spatial preposition, depending on what kind of relationship the user has in mind. For example, one can say "the fruit on [not "in"] the *plate*" or "the fruit in [not "on"] the *bowl*." The use of one preposition rather than the other depends on the conceptual nature of the reference object: We conceptualize bowls as containers, hence we must use the preposition "in"; we conceptualize plates as surfaces, hence we use "on."

Because of these constraints, different spatial prepositions "select" different kinds of objects. The examples just discussed show that "in" selects reference objects that can be conceptualized as "containers" (albeit in some very abstract sense), whereas "on" selects reference objects that can be conceptualized as "surfaces." Leonard Talmy described this as the "schematic" nature of spatial language: Spatial terms select only certain geometric properties as relevant and discard others. Talmy suggested that spatial prepositions in English (and corresponding terms in other languages) encode only nonmetric properties of objects, for example, objects as surfaces, containers, and lines. Talmy also suggested that the range of relationships encoded by languages is limited to nonmetric properties, such as support or contact, containment, parallelism, and intersection. Of course, metric information can be encoded by language, but this is done by a different subset of the vocabulary, particularly metric terms (feet, inches, and degrees) and the number system. The spatial prepositions thus might be regarded as a special subset of terms that are dedicated to expressing only selective aspects of location. Notably, the prepositions are part of the "closed class" system of English. That is, there is a restricted number of terms, and new terms are not freely admitted to the class. Thus, for example, English has terms expressing support and containment but no terms specifying exact metric relationships such as "2 inches from a reference object's border." Such a term is unlikely to be invented by speakers, and if it were it would be difficult to learn. This contrasts with the vocabulary of object words, which is an "open class" and thus readily admits new members (e.g., fax and e-mail).

Recently, Barbara Landau and Ray Jackendoff argued that the division of labor in language between the coding of objects and the coding of places parallels and possibly stems from the division of labor in the brain between the nonlinguistic representation of objects and places. Landau and Jackendoff observed that although spatial terms typically respect only nonmetric properties of objects, more detailed geometric information is respected by the naming of objects as figure and ground. Thus, in the sentence "The pear is in the bowl," the representations of "pear" and "bowl" might be linked to shape-based representations of the objects, possibly carried by the ventral stream in the brain. The spatial relationships encoded by the spatial prepositions, however, encode the very same objects without reference to detailed geometric structure—objects as figure and ground are represented as "blobs" or as containers, surfaces, lines, etc. The division of labor between the encoding of objects vs places (i.e., the nouns vs the prepositions in English) might therefore correspond to the distinction made in the brain between representations of objects (in the "what" system) and representations of location (in the "where" system) as conceptualized by Ungerleider and Mishkin.

To summarize, spatial language reveals a high degree of selectivity in its representation of spatial information about objects and locations. Only certain geometric properties of objects and only certain properties of locations and spatial relationships are respected by the language's stock of basic spatial terms. This information is clearly different from that required for navigation, for reaching or perceiving, or for map use. The distinctive nature of linguistic representations of space raises the intruiging question of how it maps onto other systems in order to allow us to talk about what we see and to use language to direct activity.

III. INTERACTION

Having considered several systems of spatial cognition that appear to engage different kinds of spatial information, one can ask how these systems interact. We do not consciously experience space as a set of separate systems, each preserving different kinds of

geometric information. Rather, our experience is of a seamless spatial world in which we navigate, reach and grasp objects, construct and use maps, and talk about objects and their spatial relationships relative to ourselves and to each other. How, then, are these interactions carried out in the mind and brain?

A distinction can be made between two different kinds of interaction, which are not always clearly specified in the literature. One kind of interaction supports the translation of information from one system into another. Because spatial systems differ, the translation is unlikely to be a one-to-one correspondence. Rather, there must be some function mapping the spatial information in one system to that of another. An example is the mapping between spatial language and other spatial cognitive systems, in which information from perception and action must be converted into a format suitable for linguistic description, in part by stripping away the metric properties of space. Such conversion is necessary in order to support our capacity to talk about our spatial experience.

A second kind of interaction, which is be considered here, is the capacity of one system to penetrate and thereby either guide or, more strongly, modulate and change the representations in another. One arena in which such interactions have been examined is navigation, considering the effects of imaginal and linguistic representations on our capacity to update our own location in space. A second, hotly debated arena concerns the relationship between cross-linguistic differences in spatial language and spatial cognition. Here, researchers have tried to determine whether language-specific encoding of space can affect the nature of nonlinguistic spatial representations or, more broadly, whether spatial language can mold spatial thought.

A. Imagery and Language as Modulators of Navigation

Although humans and other species readily update their own location as they move through space, the deployment of this capacity depends heavily on actual movement through space. Studies by Roberta Klatzky and colleagues have shown that there are substantially more errors in updating in humans when it is guided by language and imagery, by simulated motion, or by the observation of another person moving along the target path. In Klatzky's studies, people were asked to solve the following problem: Given two segments of a path, with a turn between segments, they had to turn to face the origin, as they would if they had walked along the third route back to home (Fig. 10). In one group, people were blindfolded and walked along the first two segments and then were asked to turn to face the origin (i.e., their starting point). People in a second group were given a verbal description, with which they were to construct an image of walking along the two segments. For example, people were told to imagine they would "Go forward 3 m, turn clockwise 90°, then go forward 3 m." Then they were blindfolded and told "Now face the origin." People in a third group observed another person walking along the first two segments and then were blindfolded and instructed to turn as the person would turn in order to face the origin.

Results showed that people who had physically walked along the segments were accurate, turning by the correct number of degrees to face the origin from their testing point. This indicates that they had updated an internal representation of their heading and location and could use this to generate the final

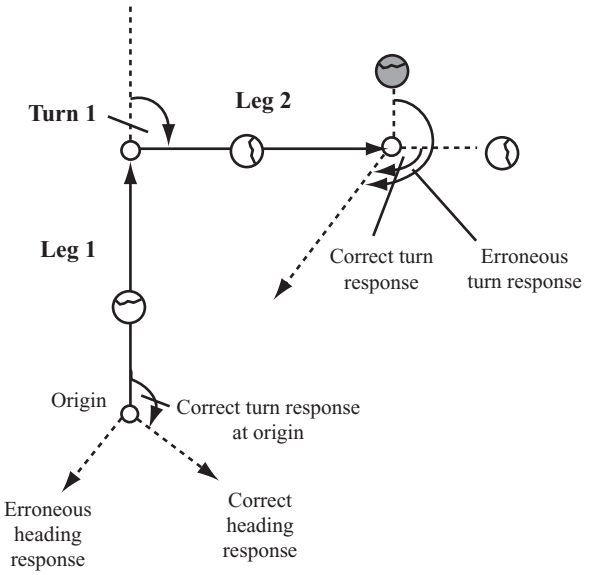


Figure 10 Schematic of the task used by Klatzky and colleagues. The participant is shown a path comprised of Leg 1, Turn 1, Leg 2 and is then asked to turn and face the origin. If people update their own location over movement, they should turn to face the origin accurately. Klatzky and colleagues found that people are accurate when they have moved through the array themselves. However, when instructions about the paths are given via language or observation of another person, they fail to update [from Klatzky, R., Loomis, J., Beall, A., Chance, S., and Golledge, R. (1998). Spatial updating of self-position and orientation during real, imagined, and virtual locomotion. *Psychol. Sci.* **9**(4), 293–298. Reproduced with permission of the American Psychological Society].

change in their own orientation so as to face the origin. Clearly, their representations preserved angle and distance information. However, people in the verbal/imagery condition and the observation condition failed to accurately update. Instead, people in both groups systematically overturned by the magnitude of the turn between segments one and two. Thus, as Klatzky and colleagues argue, in the imaginal and observational conditions, subjects did not encode the degree to which they had ''turned'' between the first and second segment. This is a striking finding, in view of the fact that the verbal instructions clearly specified exactly how much should be turned and that the observer witnessed someone carrying out the 90° turn, something that could be represented quite clearly by language. The findings suggest that neither linguistic information nor perceptual observation can modulate the representation of one's location in space. Specifically, these types of information cannot update one's heading. In contrast, the capacity to update is a foundational characteristic of the system underlying human navigation through real space.

Updating through one's own imagined movements is not impossible, however, and compared to other mental manipulations of spatial scenes it may even be relatively accessible. Early studies by Janellen Huttenlocher and Clark Presson showed that when children are instructed to imagine that they have moved to another location, they are more successful at predicting what the layout will look like than when they are instructed to imagine that the layout has moved. This finding has recently been discussed in the context of scene perception and offered as evidence supporting the view that one's representation of scenes relies on updating the viewer's representation of his or her own location in space. However, the Klatzky studies reveal that actual movement through space engages updating mechanisms that yield much more accurate representations of one's location in space. A reasonable conclusion from these studies is that the systems of spatial perception and spatial language cannot, with any precision, penetrate and modulate the system of updating used during real navigation.

B. Cross-Linguistic Differences in Spatial Language as Modulators of Spatial Thought

Spatial language is generally assumed to rest on nonlinguistic representations of space, which are presumably universal. However, recent cross-linguistic studies have revealed substantial differences in the encoding of space, some of which are quite surprising. For example, as shown in Fig. 11, English uses the spatial term ''on'' to encode the relationship of a bowl on a table and a handle on a door but distinguishes these relationships from that of an apple ''in'' a bowl. To English speakers, this partitioning is perfectly natural, which makes it surprising that the three relationships are categorized differently in other languages. For example, Dutch categorizes each of the three relationships differently, using three distinct terms, and Spanish collapses them together, using a single term to express all three relationships. Finnish partitions the handle-on-door and apple-in-bowl together and separates these from the cup-on-table. These differences, along with many others that have been recently documented, have naturally led to a resurrection of the classic question raised by Benjamin Whorf: Does the language we learn affect the way we think? In the case of spatial cognition, the question is whether early learned differences in the linguistic coding of space can modify or change aspects of nonlinguistic spatial cognition. If the answer is yes, then this would suggest that the relationship between

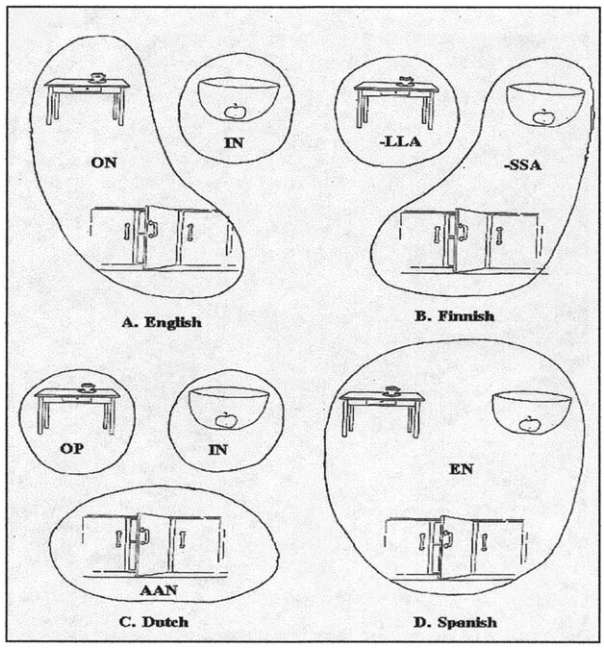


Figure 11 There is substantial cross-linguistic variation in the encoding of spatial relationships. The physical relationships shown in the figure are categorized differently by the languages shown [Reproduced with permission from Bowerman, M. (1996). Learning how to structure space for language: A cross-linguistic perspective. In *Language and Space*. (P. Bloom, M. Peterson, L. Nadel, and M. Garrett, Eds.). MIT Press, Cambridge, MA].

spatial language and nonlinguistic spatial cognition is one of substantial interaction and, more important, that this interaction can result in changes to nonlinguistic spatial cognition. If the answer is no, it would suggest the absence of such interactivity, at least in the strongest sense.

The empirical evidence to date does not support the strong form of interactivity. For example, in one widely cited study reported by Stephen Levinson and colleagues, people were asked to solve a variety of spatial problems that required them to code the location of objects relative to one or more reference systems. As discussed in Section I, languages regularly distinguish between earth-centered reference system (using terms north, south, east, and west in English) and environment-centered or body-centered reference systems (using terms such as right and left in English). Speakers of English typically reserve the former for cases in which they wish to describe large geographic layouts (e.g., ''north of New York'' but not ''north of the cup''). Levinson observed that speakers of Tzeltal follow quite a different pattern, regularly using the terms of the earth-centered reference system to encode the locations of most objects, including small moveable ones. Thus, a speaker of Tzeltal might describe the location of an apple on the kitchen table as ''to the north'' rather than ''to my left'' or ''to the right of the sink.'' Levinson speculated that lifelong usage of this reference system in linguistically coding location might modify people's spatial representations such that they would solve *nonlinguistic* spatial problems using this reference system as well. The results from a variety of spatial problems were that speakers of Tzeltal showed a bias—though far from an absolute bias—to code spatial relationships in terms of the earth-centered rather than environment-centered reference system. This contrasted with the performance of Dutch speakers, who tended to encode location in an environment-centered reference system.

Such findings might suggest a bias toward the use of one kind of reference system, but they do not prove that the underlying nonlinguistic representations have been modified. Moreover, biases to use one reference system rather than another are common in many spatial tasks; for example, evidence reviewed in Section II showed that navigation in animals and humans can engage different reference systems, depending on the context of the task. Recent research by Peggy Li and Lila Gleitman has shown that biases such as those shown by Tzeltal speakers can also be induced in native English speakers if the context is suitably altered to enhance the likelihood of encoding location in terms of a geographic reference system. This suggests that the effects of linguistic experience in Levinson's task are shallow.

A number of other studies have shown that certain aspects of nonlinguistic spatial representation are immune to the effects of linguistic experience. In one study, Edward Munnich and colleagues gave native English and Korean speakers a nonlinguistic spatial task that tested their memory for the location of a ball, either on or above a table. Adults of both speaking communities were much more accurate in remembering the locations that were in contact with the table (''on'' it) than those that were not in contact with it (''above'' it). This pattern of performance contrasted sharply with the people's *linguistic* categorization of the same locations: Whereas English speakers uniformly distinguished the two categories of relationships by using two different spatial terms (e.g., ''on'' vs ''above''), the Korean speakers did not. Consistent with their native language, Korean speakers used just a single term across both contact and no-contact locations, only occasionally marking the distinction with different terms. Thus, a clear, lifelong difference in the linguistic encoding of contact vs no contact apparently had no impact on the two groups' nonlinguistic memory for location, which showed a strong contact/no-contact distinction regardless of whether it was coded by their language.

Results such as these indicate that at least some aspects of nonlinguistic spatial representation are immune to the effects of cross-linguistic differences. It remains to be determined whether some aspects of nonlinguistic spatial representation can indeed be restructured by linguistic experience.

IV. DEVELOPMENT

The early emergence of specialization in spatial cognition is consistent with the idea that there are innate constraints on the kinds of spatial information that the brain can represent, and that these vary for different aspects of spatial cognition. At the same time, the existence of constraints does not preclude learning, development, or change. The developing brain is a dynamic structure, and there is much evidence for its plasticity during the early years of life. However, this plasticity is not unlimited. Rather, studies showing effects of early experience often suggest change that is constrained by the lines of normal cognitive architecture. Studies of variation in early experience have led to insights about the early localization of spatial

cognitive functions in the brain, the existence and nature of specialization across spatial cognitive domains, and the causes of reorganization.

A. Localization

Studies of adults sustaining lesions to the brain suggest that the two hemispheres preferentially represent different aspects of spatial information. People with right posterior lesions tend to show disorders in which they have difficulty assembling parts into a coherent global whole. People with left posterior lesions tend to show a different pattern, with difficulties representing individual elements of global patterns. The question then arises whether these patterns of hemispheric difference are shown early in development, which would suggest that the two hemispheres have native preference for processing different aspects of spatial information. Joan Stiles and colleagues studied young children who sustained early injury to the right vs left hemisphere. They found that the patterns seen in adulthood are also present in early childhood. For example, children with right hemisphere damage typically perform more poorly than those with left hemisphere damage on tasks requiring them to assemble blocks to copy an existing spatial design. In contrast, children with left hemisphere damage tend to have more difficulties breaking up a global pattern into its individual components. These different patterns are not as pronounced as in adults, but according to Stiles they are qualitatively similar. Moreover, children recover to some extent, ultimately producing constructions that are qualitatively similar to those of normally developing children. The mechanisms by which such recovery takes place are not well understood, but compensation by recruiting new kinds of strategies is likely to be partly responsible for recovery of function. The early existence of some degree of hemispheric specialization for spatial cognitive functions suggests that the right hemisphere may be innately privileged in its capacity to represent the spatial relationships among elements.

B. Specialization

Early specialization is also suggested by recent studies of individuals with Williams syndrome (WS), a rare genetic disorder caused by a hemizygous submicroscopic deletion of chromosome 7q11.23, which includes the gene for elastin (ELN) as well as a number of other genes, including the gene for protein LIMK1. Children and adults with WS show a characteristic cognitive pattern that includes overall mild to moderate mental retardation (mean IQ = 55–60) as well as a unique profile of profoundly impaired spatial cognition, together with relatively spared language. These relative strengths and deficits are shown from quite early in development, suggesting a link between the deleted genes and the development of the brain's functional capacities. The gene producing the LIMK1 protein is thought to be implicated in the spatial disorder because it is strongly expressed pre- and postnatally in the brain, whereas ELN is not.

The spatial impairment shown by individuals with WS has been demonstrated most often in tasks requiring them to copy a model, either by drawing it in an adjacent space or by assembling individual blocks to create an overall design. Typically, individuals with WS have great difficulty in creating the appropriate spatial relationships between individual elements, as shown by the sample drawings in Fig. 12. Although individuals with WS do show some developmental progression, they do not show the same degree of recovery as children with right hemisphere lesions; even adults with WS show striking and profound spatial impairment in this task.

Currently, the causes and nature of the spatial impairment in WS are not well understood. Hallmark tasks such as drawing and block construction most likely draw on numerous capacities that are difficult to isolate for tasks of such complexity. However, there is growing evidence that the spatial impairment does not affect all domains of spatial organization equally. For example, individuals with WS do not appear to be significantly impaired in the capacity to recognize faces, even under conditions of unusual orientation and lighting. They are capable of recognizing familiar objects under nonoptimal viewing conditions, such as when the images are blurred, or present objects in unusual orientations. They are capable of perceiving biological motion displays, which require the construction of coherent global figures from information conveyed through the motions of individual points of light. The uneven profile within spatial cognition thus suggests that the genetic deficit may target certain spatial cognitive capacities while leaving others unaffected.

C. Causes of Reorganization

There is strong evidence for plasticity in the developing brain. Recovery of function among children with right hemisphere damage is one example of such plasticity.

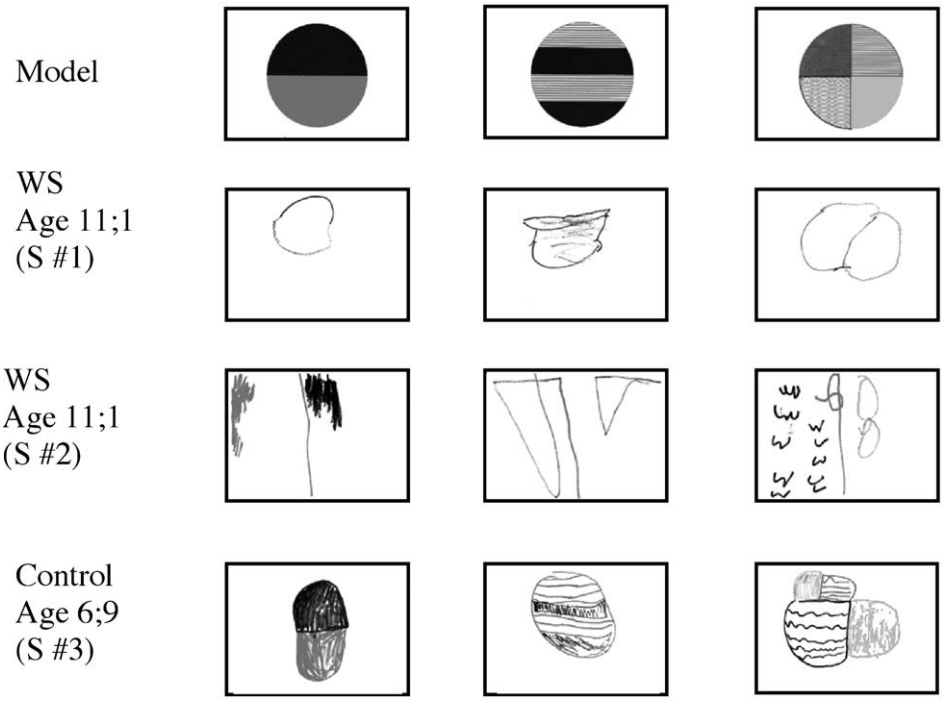


Figure 12 Children with Williams syndrome have great difficulty copying visible models. The top row shows three models that were copied by two children with Williams syndrome (rows 2 and 3) and one normally developing child who was their mental age match (row 4).

Other evidence suggests that changes in early experience can have a major impact on the organization of the cortex. A number of studies have shown that depriving the organism of input from a given set of sensory receptor surfaces leads to reorganization in which the cortical areas normally representing the input are "taken over" by other kinds of input. For example, in adults, amputation of a limb can lead to reorganization of the cortex whereby the area of the cortex that previously represented the limb becomes responsive to stimuli from inputs that are normally represented in adjacent areas of the cortex. Congenitally blind individuals show activation of areas normally considered visual cortex during tasks that engage the somatosensory systems. This suggests that the area normally "designated" as visual cortex may be commandeered to represent other systems when a person has no vision. Thus, the cortical representation of a particular input may become expanded or contracted through major alterations in the organism's experience.

A clear example of such reorganization in the spatial domain has been described by Helen Neville and colleagues, who studied spatial attention in congenitally deaf individuals compared to hearing individuals. They found that responses to stimuli presented *centrally* (i.e., to the fovea) in the visual field were similar in both groups, both in terms of behavioral accuracy and event-related potential (ERP) responses. In contrast, responses to stimuli presented *peripherally*, either in the right or left visual field, showed significant differences across groups. Deaf individuals showed increased ERP responses to peripheral stimuli, relative to hearing people, and the distribution over scalp regions was different. Whereas hearing individuals predominantly showed responses over the parietal region contralateral to the stimulus, deaf individuals also showed bilateral responses over the occipital region. Complementary studies of hearing individuals who were born to deaf parents and whose native language was a sign language showed that congenital deafness, and not the experience of acquiring a spatial–manual language, was responsible for the pattern of brain activation among the deaf individuals. These results suggest that congenital auditory deprivation can lead to cortical reorganization in which there is enhanced responsiveness to peripheral stimuli.

Additional evidence, from both animal models and studies of the human visual system, suggests that the representation of peripheral visual space is carried along the dorsal pathway toward the posterior parietal cortex, which is important for the representation of spatial location. In contrast, the representation of central visual space is carried along the ventral pathway, which is important for the representation of form.

Based on evidence from comparative studies of deaf and hearing individuals, as well as evidence from animal models and neuroanatomy, Neville and colleagues suggested that the development of the dorsal stream may be more subject to effects of early experience than that of the ventral stream. The modifiability of peripheral visual attention consequent on deafness (relative to central visual attention) is consistent with this notion. Additional studies have shown that congenitally deaf individuals exhibit selectively enhanced processing of stimuli that are normally processed by the dorsal stream.

V. CONCLUSIONS

Spatial cognition is a broad area of inquiry that can be informed by studies of the mind and brain, including their organization, function, and development. This article provided a selective summary of evidence supporting a high degree of specialization within the domain of spatial cognition, limited and highly specialized interactions among these subsystems, and a strong native basis through which developmental process can act to modify spatial organization. We are only beginning to understand the detailed nature of the spatial systems the nature of the mechanisms that support the interface among these systems. However, the progress that has been made shows that rich and detailed organization in the mind and brain supports our capacity to negotiate the spatial world.

See Also the Following Articles

CONSCIOUSNESS • INFORMATION PROCESSING • LANGUAGE AND LEXICAL PROCESSING • MEMORY, NEUROIMAGING • MOVEMENT REGULATION • MULTISENSORY INTEGRATION • OBJECT PERCEPTION • PATTERN RECOGNITION • SALIENCE

Acknowledgments

Preparation of this article was supported in part by Grants 12-FY99-670 from the March of Dimes Birth Defects Foundation, 1 R55 NS37923 from the National Institutes of Neurological Disorders and Stroke, and SBR-9808585 from the National Science Foundation. I thank Michael McCloskey and Ed Munnich for helpful comments on the text and Nicole Kurz for help in preparing the figures.

Suggested Reading

Behrmann, M. (2000). Spatial reference frames and hemispatial neglect. In *The New Cognitive Neurosciences*, (M. Gazzaniga, Ed.), 2nd ed. MIT Press, Cambridge, MA.

Bertenthal, B. I., and Clifton, R. K. (1998). Perception and action. In *Cognition, Perception and Language, Vol. 2: Handbook of Child Psychology* (D. Kuhn and R. S. Siegler, Eds.), 5th ed. Wiley, New York.

Bloom, P., Peterson, M. A., Nadel, L., and Garrett, M. F. (1998). *Language and Space*. MIT Press, Cambridge, MA.

DeLoache, J., Miller, K. F., and Pierroutsakos, S. L. (1998). Reasoning and problem-solving. In *Cognition, Perception and Language, Vol. 2: Handbook of Child Psychology* (D. Kuhn and R. S. Siegler, Eds.), 5th ed. Wiley, New York.

Gallistel, C. R. (1990). *The Organization of Learning*. MIT Press, Cambridge, MA.

Goodale, M. A. (2000). Perception and action in the human visual system. In *The New Cognitive Neurosciences* (M. Gazzaniga, Ed.), 2nd ed. MIT Press, Cambridge, MA.

Johnson, M. H. (1997). *Developmental Cognitive Neuroscience*. Blackwell, Cambridge, MA.

Landau, B., and Jackendoff, R. (1993). "What" and "where" in spatial language and spatial cognition. *Behav and Brain Sci* **16,** 217–266.

McCloskey, M. (2000). Spatial representation in mind and brain. In *What Deficits Reveal about the Human Mind/Brain: A Handbook of Cognitive Neuropsychology*. (B. Rapp, Ed.), Psychology Press, Philadelphia.

Neville, H. J., and Bavelier, D. (2000). Specificity and plasticity in neurocognitive development in humans. In *The New Cognitive Neurosciences* (M. Gazzaniga, Ed.), 2nd ed. MIT Press, Cambridge, MA.

Newcombe, N., and Huttenlocher, J. (2000). *Making Space*. MIT Press, Cambridge, MA.

O'Keefe, J., and Nadel, L. (1978). *The Hippocampus as a Cognitive Map*. Clarendon, Oxford.

Spelke, E., and Newport, E. L. (1998). Nativism, empiricism, and the development of knowledge. In *Theoretical Models of Human Development. Vol. 1: Handbook of Child Psychology* (R. M. Lerner, Ed.), 5th ed. Wiley, New York.

Stiles, J. (1998). Early brain injury: Rolling with the punches. *Curr Directions Psychol Sci* **7**(1), 21–25.

Spatial Vision

RUSSELL L. DE VALOIS and KAREN K. DE VALOIS
University of California, Berkeley

GLOSSARY

adaptation Change in sensitivity to a pattern resulting from prolonged exposure to that pattern.

columnar organization The cells extending in a column from the surface of the cortex to the white matter, all of which share common stimulus selectivities.

modular organization The subdivision of the striate cortex into about 1500 modules, each containing approximately 200,000 cells analyzing information from a specific region of the visual field.

receptive field The receptive field of a cell consists of all those receptors that provide input to the cell or, equivalently, the delimited directions in visual space from which stimuli activate the cell.

receptive field organization The spatial and temporal arrangement of the excitatory and inhibitory portions of a cell's receptive field.

retinotopic organization An arrangement of the cells at some processing level in which the general topographic arrangement with relation to the photoreceptors is maintained, even though some portions might be magnified with respect to others.

spatial frequency The frequency of sinusoidal variations in light across space, specified in cycles per degree visual angle.

visual angle The angle at the pupil subtended by an object in the visual field.

The information obtained by the photoreceptors in the eye provides the primary basis for our information about objects in the world around us, about their characteristics and their spatial locations. Obtaining useful information about the world from the receptor outputs, however, involves a number of very complex computational tasks. The visual information from objects, most of which do not emit light but only reflect a certain proportion of light that falls on them, needs to be disentangled from that from the illuminant. There is a massive amount of information in the visual image and the computations need to be completed in a very brief amount of time. The image of the environment on the receptors is greatly underspecified since there are many possible stimuli that could have produced any particular image. Approximately one-third of the brain is devoted to vision, and many of the processes by which it solves these and other problems are not well understood. We discuss how the initial stages of spatial vision appear to operate.

I. PROBLEMS TO BE SOLVED

One of the most difficult problems in understanding how the visual system works is in realizing that there is a problem to be solved. We open our eyes and instantly see in front of us a scene made up of, for example, trees and grass and people coming out of a building, objects of different colors and shapes at different distances, some moving and some stationary. We perceive all this so effortlessly that it feels as if it must be a simple process. Nothing could be further from the truth. It is in fact an incredibly complex process, by far the most difficult problem the brain must solve. Tens of billions

of cells, about 35% of the whole neocortex, are devoted to analyzing and interpreting the output of the visual receptors. Although we perceive objects of various shapes and characteristics, those percepts are constructs of the brain. The only signal coming from each photoreceptor (100 million rods and 3–5 million cones) is in effect a single number corresponding to how many photons that receptor has absorbed at a particular instant. The brain must determine what in the outside world could have reflected light into the eye to produce that particular enormous set of numbers. Furthermore, the brain has only about 200 msec or less to carry out this computation before an eye movement leads to a new set of millions of numbers.

The problem facing the visual system is made even more daunting by the fact that the numbers put out by the receptors are by themselves largely meaningless. Unlike some other senses, the visual system does not obtain information directly from objects in the world but only visual echoes from light reflected from objects. Objects make sounds and emit odors and can be identified by the auditory and olfactory systems from these emissions, but with few exceptions objects in the natural world do not emit light; they merely reflect a certain proportion of whatever light falls on them. Furthermore, almost all the variations in photon absorption over a day are due to the variations in the intensity of the illuminant and have nothing to do with the color, brightness, or any other identifying characteristic of visual objects. The intensity of illumination from the sun rises and falls by huge amounts over the course of the day and as clouds pass by. The difference in the amount of light reflected from one object as opposed to another is very much smaller. The whitest object in a scene reflects perhaps 80% of the incident light and the blackest object perhaps 10%. The variations due to the object of interest are thus at most only about eightfold, whereas the illuminant may vary by a factor of 10 million. Strictly speaking, this is an unsolvable problem because each receptor puts out only one number (corresponding to the number of photons it has absorbed) and this is a product of both the illuminant and the reflectance of the object at that point in the image. Since we have no independent knowledge of the illuminant, this is in effect one equation with two unknowns. The only way the visual system can gain information about visual objects is to compare the outputs of different receptors and make certain assumptions about how objects differ in their visual properties from the illuminant. Much of the early processing of spatial information reflects the way the visual system deals with this problem.

The analysis of spatial information, from the capture of light by the photoreceptors to the cortical levels associated with perception and the organization of a motor response, is a sequential process involving about a dozen synaptic levels—a dozen successive levels of analysis. At the early stages, there is massively parallel processing: Each small region of visual space is analyzed largely independently by a few of the millions of neurons at that level. At successive early stages, the information is combined over increasingly larger regions but still in a massively parallel way. Such parallel processing enables the nervous system, with its relatively slow processing abilities, to nonetheless carry out the immense computations required in the brief time available between fixations.

The first five or six processing stages [in the retina, the lateral geniculate nucleus (LGN) of the thalamus, and the striate cortex] isolate and encode in a useful form that component of the retinal image that is due to reflective objects. We understand fairly well the anatomical structures, the physiological processes, and the contribution to vision of these early processing levels. Beyond the striate cortex, however, are 25 or more later visual areas. Here, the visual system must be dealing with the problems of segregating the visual information from different parts of the scene into separate objects and identifying the nature of each object. We currently have only very fragmentary evidence about and a very incomplete understanding of how this is accomplished.

II. OPTICS, SAMPLING, AND VISUAL ACUITY

The cornea and, to a lesser but variable extent, the lens form a two-dimensional image of the visual world at the level of the receptors at the back of the eye. The receptors sample the image as the first step in spatial vision. The optical quality of the human eye is not very good—well below that of the cheapest camera lenses. Two of the most prominent imperfections are spherical and chromatic aberrations. The former results from the cornea's imperfect shape, and the latter occurs because the cornea and lens differentially refract light of different wavelengths. Both of these aberrations become worse as the pupil of the eye increases in diameter, as it does in order to capture more light under dim light conditions. An imperfect image also results from the imaging of a three-dimensional world on a two-dimensional surface since objects at different distances are not all simultaneously in focus. Finally, the eye is mechanically unstable. It floats in fat within

the orbit, held fixed only by the tension in six relatively massive eye muscles. These muscles can rotate the eye in different directions very rapidly, but they cannot hold it very still because of the random firing of the individual muscle fibers. The eye thus oscillates rapidly in a random fashion and drifts about to some extent as well. The resulting blur is made worse by observer movement, both translating the eye through space (which blurs the image) and jarring it.

Visual stimuli can be equivalently described in the space domain (i.e., by how much light is present at each location in space) or the spatial frequency domain (by the frequency, amplitude, and phase of sinusoidal variations in light intensity across space). Although space–domain descriptions are more familiar, there are sometimes heuristic advantages to examining visual problems in the spatial frequency domain instead. The various optical imperfections of the eye mentioned previously all blur the image on the retina, and blur selectively attenuates high spatial frequencies. Our ability to detect sinusoidal gratings of different spatial frequencies is described by the spatial contrast sensitivity function (CSF) (Fig. 1). In photopic (daylight) vision under optimal conditions, the CSF function peaks at about 4 cycles per degree, and a person with excellent acuity can resolve gratings as high as about 60 c/deg visual angle at the fixation point. "Normal" acuity (20/20 in American notation) is equivalent to a high spatial frequency cutoff of 30 c/deg. Under many viewing conditions, however, such as when the illumination level decreases (see Fig. 1), with increasing retinal eccentricity, or when we are moving and thus blurring the retinal image, the resolution of a normal observer does not exceed 10–20 c/deg. In these circumstances, tasks such as reading that require high acuity become impossible, but we still have excellent ability to segment the visual world into different objects and to recognize these objects and their spatial relationships.

The CSF shows both high and low spatial frequency attenuation. The decrease in sensitivity of the CSF at high spatial frequencies is in part a consequence of the various types of optical blur discussed previously, but it is also related to sampling by receptors and later neural components. The reduction in sensitivity at very low spatial frequencies has another, purely neural, explanation (discussed later).

Studies of the spatial characteristics of natural scenes have shown that most of the important visual information lies not in fine detail but at much coarser scales. The amplitude spectrum of most natural scenes decreases with increasing spatial frequency by approximately 1/f. Thus, most of the information is at

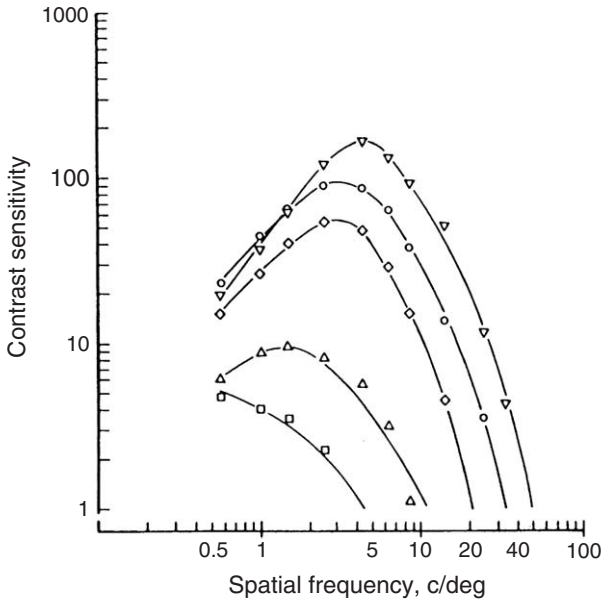


Figure 1 The mean spatial contrast sensitivity of five observers for gratings of different spatial frequencies at each of five different luminance levels, in one log unit steps from photopic (top curve) to scotopic (bottom curve) levels (reproduced with permission from De Valois *et al.*, *Vision Res.* **14,** 75–81, 1974).

mid to low spatial frequencies, and gaining information about the low spatial frequency content of the image is doubtless much more important in most circumstances than detecting the very highest spatial frequencies. This is fortunate because the optic nerve bottleneck would not allow the huge number of fibers required to transmit all the high spatial frequency information in any case.

Just more than two samples per cycle of a sine wave of a particular frequency are necessary and sufficient to resolve it and all lower frequencies. Thus, if the optics of the eye pass up to 60 c/deg, sampling by 120 receptors/deg should capture all the information in the retinal image. This is in fact approximately the spacing of cones in the very center of the human fovea, the foveola, where the cones are very thin and tightly packed together. Furthermore, there are advantages in having the neural sampling comparable to, neither better nor worse, the highest frequencies passed by the optics. Thus, our eyes' optics are approximately optimally matched to the density of cones in the fovea.

Outside the central 20–30′ of the foveola, the cones become increasingly larger and widely spaced. Near the edge of the fovea, rods begin to appear interspersed among the cones, and by 2° from the center rods

surround each cone, further separating them. With increasing eccentricity, there is also increasing convergence of cones onto later neural elements, with a further loss of detailed vision. The highest spatial frequencies the system can resolve photopically (when only cones are functional) must and does therefore decrease as retinal eccentricity increases, but low spatial frequency information is captured even out to the far periphery. This system of gaining high spatial frequency information only within a very limited region of the image only works because we can move our eyes and inspect in greater detail any desired part of the visual world. Animals with less mobile eyes have quite a different organization.

Rods are as small as foveal cones and are densely packed between cones so they could potentially sample the image with high resolution. However, our acuity under scotopic conditions (dim light levels when only rods are functional) is very poor, with contrast sensitivity cutting off at less than 10 c/deg. In this case (and also to a lesser extent for cones in the far periphery), it is clear that the effective sampling unit is not the receptors but the later neural elements. Anatomical observations support this, showing considerable convergence of rods onto rod bipolars and rod bipolars onto ganglion cells. Two adjacent scotopic "samples" of the image, then, would each cover an area containing dozens to hundreds of rods, not just one, and would be separated by the diameter of dozens or hundreds of rods.

III. TRANSDUCTION AND ADAPTATION

Two of the most critical retinal functions are to transduce light energy into neural activity and to transform the information into a usable form. The transduction of light energy into the flow of ions across nerve membranes, the language of the nervous system, takes place in the outer segments of the receptors. The capture of a single photon of light by a uniquely sensitive photopigment, rhodopsin (or the similar cone opsins), converts rhodopsin into an enzyme and initiates a cascade of chemical reactions within the cell. The cell membrane is maintained in a partially depolarized state at any illumination level. Variations in the chemical cascade produced by momentary increases or decreases in the amount of light absorbed produce momentary changes in polarization of the membrane. A decrement in light absorbed depolarizes the membrane; an increment in light absorbed hyperpolarizes the membrane. The resulting current flow is conducted to the synaptic region of the receptor, which then secretes synaptic transmitter in proportion to the amount of depolarization.

The cascade of chemical reactions within the outer segment provides the amplification necessary to bridge the enormous energy gap between a single photon and the flow of ions across a membrane. It also has other effects directly relevant to spatial (and temporal) vision. It is a slow process spread out over several milliseconds. This introduces a significant delay into the visual response, but it also provides for useful temporal integration to increase sensitivity. The resulting smearing in time filters out high temporal frequencies that, in the case of vision (as opposed to audition), are not informative because they are mostly noise due to quantum fluctuations. Few objects in the natural world produce informative variations in light at very high temporal frequencies. Temporal information in vision is particularly useful for detecting motion, but this can be accomplished without sensitivity to high temporal frequencies.

The chemical cascade within the receptor outer segment also contributes to the solution of the problem of separating the signal due to reflective objects from that due to the illuminant. Although we cannot know the illuminant, there are certain differences between visual objects and illuminants that the visual system exploits to deal with this problem (although not without occasional large errors). The visual system correctly makes the assumption that light from the illuminant tends to be uniform over large spatial and temporal extents, whereas reflective objects of visual interest usually vary across space and also across time as we move our eyes. Therefore, the visual system in effect treats any part of the image that is largely invariant in space and time as if it were due to the illuminant and ignores it, whereas parts of the image that change more rapidly in space and time are considered to come from reflective objects and are further analyzed. Any process that strongly attenuates very low spatial and temporal frequencies will selectively attenuate signals from the illuminant, thereby allowing the visual system to detect and process the small variations due to objects in the presence of larger variations in light intensity due to the illuminant.

The attenuation of very low temporal frequencies starts within the photoreceptors. A feedback circuit in the receptor outer segment reduces responses to steady illumination while passing responses to transient increments and decrements. Functionally similar interactions occur at later neural levels as well so that most cortical cells respond poorly to any unchanging

stimulus. The attenuation of very low spatial frequencies is also a multistage process, and it starts at the first synaptic level.

To deal with the vast, but largely irrelevant, changes in light level over a day, not only low temporal and spatial frequency attenuation but also variations in sensitivity with light level are required. Light and dark adaptation refer to the processes by which the visual system adjusts its sensitivity as the illumination level increases and decreases. A white object might reflect 80% of incident light and a slightly grayer one 70%. In bright daylight, there might be 10 million incident photons/sec/unit area so these two surfaces would reflect 8 million and 7 million photons/sec/unit area, respectively. However, on a dark night, the illumination might be only 1000 photons/sec/unit area and the two surfaces would then reflect only 800 and 700 photons/sec/unit area, respectively. To distinguish between the two surfaces under these dim conditions, the visual system would need to be 10,000 times more sensitive. Furthermore, if the visual system did not change its sensitivity as illumination levels changed, objects would not maintain constant appearance.

Many of the adaptational sensitivity changes required to maintain object constancy take place within the receptors. The decrease in sensitivity going from dark to light is fast, but the reverse change can be very slow (as one notices on entering a dark theater from a brightly lit street). The breakdown of photopigment molecules upon light capture is fast but the regeneration of the photopigment is very slow, and the presence of unregenerated photopigment molecules sends a signal that slows the increase in sensitivity when one suddenly enters a dark environment.

IV. RETINAL ORGANIZATION

The optic nerve is a serious bottleneck in the visual path, constraining early processing. This is largely a consequence of the inversion of the retina in the vertebrate eye. The photoreceptors lie at the very back of the eye, distal to all the retinal circuitry. The ganglion cell axons, which carry the eye's output to the brain, are inside the eye. They exit through a hole in the receptor layer of the retina (at the optic disk, forming the blind spot). There are approximately 100 million receptors in each eye, and perhaps 2000 million cells in the cortex that process their output, but each optic nerve contains only 1 million axons. Getting even this relatively small number of fibers out of the eye produces a blind spot large enough to encompass more than 50 nonoverlapping images of the moon. Any larger number of ganglion cell axons would be self-defeating. This organization puts a considerable constraint on the visual system, limiting the extent and nature of processing that takes place in the retina.

The output of the receptors is processed at two synaptic levels in the retina and at one level in the LGN of the thalamus before arriving at the cortex. There is a direct anatomical path from receptors to bipolars to ganglion cells, with lateral connections made by horizontal and amacrine cells at their respective levels bringing in information about neighboring regions. At each stage, the output of a small group of receptors (to which the bipolars, and through them the ganglion cells, are directly connected) is compared with the output of a larger group of receptors (to which the bipolars and ganglion cells are connected through the horizontal and amacrine cells). The lateral inputs at each level are opposite in sign to the direct inputs, so each ganglion cell reports the difference between the extent of activation of two different populations of receptors. All the receptors that feed into a particular cell constitute its receptive field (RF). The RF of a ganglion cell thus consists of a center region responding to light decrements and a surround responding to light increments or vice versa. The spatial extent of an RF can vary from a tiny region in the foveal representation to a much larger region in the periphery.

A neural organization that is seen only to a limited extent in the retina but very frequently at cortical levels is a second type of parallel processing. In addition to parallel processing of information from different retinal regions, the visual system separately processes different types of information within each small region. The output of each receptor feeds up to the brain along several paths, in each of which its output is compared with different combinations of other receptors. There are three main paths or channels in the retina, with various subtypes of cells within each. The magnocellular or (Mc) pathway leads from receptors to diffuse bipolar cells to parasol ganglion cells. Their axons, in turn, project to the Mc layers of the LGN. The parvocellular (Pc) path leads from receptors to midget bipolars to midget ganglion cells, the axons of which project to the Pc layers of the LGN. All the long (L)- and middle (M)-wavelength-sensitive cones feed into both of these paths, but different analyses of their outputs are carried out in the two cases. In addition, the relatively few short-wavelength-sensitive (S) cones feed down a separate koniocellular (Kc) path. This system is involved almost exclusively with color vision and will not be further discussed here.

A diffuse bipolar cell in the central retina makes direct contact with about six cones, but it receives indirect input from many other cones by way of the lateral contacts made by the horizontal cells. The direct bipolar connections and the horizontal inputs come from both L and M cones but not from S cones. The important point is that the direct and indirect inputs produce opposite effects; that is, if the direct input from receptors depolarizes the bipolar, then the indirect input from receptors by way of the horizontal cells hyperpolarizes the bipolar. This variety of bipolar thus sums the output of a small group of L and M cones in a region and subtracts that from the sum of the outputs of a larger group of L and M cones.

There are two subtypes of diffuse bipolars that feed separately up the Mc pathway. One subtype receives direct excitatory input from receptors. Its center mechanism is depolarized by decrements in light absorption, just as are receptors. The other bipolar subtype inverts the receptor signal, so it is depolarized by increments in light. Each L or M cone feeds into one or more of each of these bipolar types. Recall that the crucial information for vision is not the absolute light level, which is due mainly to the illuminant, but the increments and decrements from this mean level that characterize reflective objects. For a single cell to encode both increments and decrements equally well would require a very high (and thus metabolically expensive) average maintained firing rate. By having some cells that respond with excitation to increments and others that are excited by decrements, the visual system doubles its capability to transmit the information, eliminates the need for a high maintained firing rate, and ensures that the important signals about increments and decrements are equally represented in the message to the brain.

In addition to the synapses each L and M cone makes with diffuse bipolars, each L and M cone also feeds into two midget bipolar cells, one sign retaining and one sign inverting, in the Pc pathway. In the whole central 70° or more of the retina, each midget bipolar cell makes direct contact with just a single cone. Of course, other cones are contacted indirectly by way of the horizontal cells. Thus, there are four types of midget bipolars and four types of midget ganglion cells into which the midget bipolars feed. The midget bipolars also make contact with several horizontal cells, which bring in opposing information from neighboring L and M cones. These cells thus report the difference between the activation in one particular L or M cone and the activation of a small group of neighboring L and M cones.

Cells in the Mc and Pc pathways thus differ with respect to their chromatic properties since color information lies in the difference between the activation of receptors with different spectral sensitivity. The cone inputs to RF center and surround are the same in Mc cells (M + L in each case), whereas the RF center of each Pc cell (in the central retina) is derived from one cone type and its surround from a different cone type. The Mc cells can thus convey information about increments and decrements from the mean luminance level, luminance contrast, but they cannot convey information about chromatic differences. Cells in the Pc pathway, on the other hand, convey information about both color and luminance. They multiplex these parameters. Because of their radially symmetric RFs, cells at this level have minimal selectivity for the shape of objects, nor do they have selectivity for motion or, of course, for depth. All of this changes at the cortex.

V. RETINOTOPIC ORGANIZATION AND CODING

The optics of the eye forms an image of the world on the receptors. Except for the fact that the image represents a two-dimensional projection of a three-dimensional world, spatial relationships in the visual world are faithfully represented in the receptor activation pattern. Adjacent receptors receive light from adjacent visual directions, and two receptors spaced, for example, 100 μm apart are stimulated by points about equally far apart in visual angle in the image, regardless of whether the two are in the fovea or in the far periphery. Both of these properties break down in the retinal processing and projection to the cortex—the adjacency relationship to a relatively minor extent and the equal separation one to a major extent.

There is an approximate retinotopic organization through the retina and geniculate in terms of adjacency relationships, with neighboring retinal regions projecting to neighboring cortical areas. However, the topographical map becomes grossly distorted because of the far greater number of cells processing information from the fovea than from the peripheral retina. To a first approximation, the number of ganglion cells decreases approximately proportional to the log of the distance from the fovea. Thus, an annulus on the retina extending from 1° to 2° contains about the same number of ganglion cells as an annulus from 10° to 20°, although the latter covers a region 100 times as large in visual space.

The striate cortex, as opposed to the retina, contains many cells that are selective for the presence of certain

precise local stimulus characteristics. They show excitation only if the local stimulus contains their particular set of required features. Retinal information, however, is compactly coded for maximum efficiency of transmission through the bottleneck of the optic nerve. Each ganglion cell responds to many different spatial patterns, independent of the local orientation and largely independent of spatial frequency, whether the object is moving or stationary, etc. The fact that a particular ganglion cell responds thus indicates relatively little about the nature of the stimulus. The information is carried instead in the relative firing rates and locations within the array of each of a number of cells. In the striate cortex, on the other hand, the information is more sparsely coded. Each cell responds if and only if the information in its local region is within the particular range of spatial frequencies, orientations, etc. to which the neuron is tuned. The majority of the cells within a cortical region will not fire at all to most stimuli (hence the term "sparse coding"), but when a particular cell fires it conveys a great deal of information about the stimulus in that region: It must be a pattern of *this* orientation and *this* spatial frequency, and so forth.

VI. MODULAR RETINOTOPIC AND COLUMNAR ORGANIZATION

There is a purely spatial representation of visual information at the level of the receptors, in which adjacent receptors respond to light from adjacent points in visual space. The first visual cortical area (V1) is broken up into approximately 1500 cortical modules, each containing about 200,000 neurons that deal with a particular separate subregion of the visual field; thus, the cortical locus of activity also specifies to some extent the location of the stimulus in visual space. But within each cortical module, there is a functional organization, as a consequence of which adjacent neurons may be receiving input from the same region in visual space but differ from each other in their functional properties. Each striate cortex cell receives input from a particular combination of geniculate fibers, with neighboring cells receiving different combinations of outputs from the same group of fibers. Thus, if one cell as opposed to an adjacent one fires, this does not necessarily signal that the stimulus is in one location rather than another but rather that the stimulus in a particular location has a certain orientation or spatial frequency rather than another. This modular arrangement in the cortex can be seen as a compromise between a purely spatial, retinotopic representation and a functional representation. Each individual V1 neuron thus carries information both about the spatial location of the stimulus and about the nature of the stimulus at that location. The cells within each of the V1 modules perform the same sorts of stimulus analyses, although over different scales depending on eccentricity.

VII. SELECTIVITIES

The most important spatial processing within the striate cortex is the production of cells selective for only certain values along each of various spatial stimulus parameters. Correlated with this stimulus selectivity is the production of multiple channels of a group of neurons each responsive to a different portion of a given stimulus dimension. Cells in the striate cortex receive input from larger spatial regions than do cells earlier in the path, but they still perform a local analysis of the visual information, being only minimally responsive to whatever is present in any but their relatively small part of the visual field. However, they do not respond to every pattern within their local region. Rather, the effective stimulus for any given cell must have certain specific properties.

Most neurons in V1 are selective for stimulus orientation. They respond to a bar or grating within their RF only if it is within a certain range of orientations, with the optimum orientation varying from cell to cell. There is a considerable range of orientation bandwidths among cells in the primate striate cortex, with the median bandwidth at half-maximum-amplitude response being about 45°. Therefore, although LGN cells have little or no orientation tuning, responding equally well to patterns of any orientation, a typical striate cortex cell will respond only to patterns of about one-fourth of the total range of possible orientations. Within each cortical module are cells tuned to each of many different orientations.

Most V1 cells are also quite narrowly tuned to sinusoidal gratings of different spatial frequencies. LGN cells have a small degree of spatial frequency selectivity for luminance-varying patterns, but a typical neuron in V1 is much more selective, with a median bandwidth of about 1.4 octaves, which (like their orientation selectivity) encompasses about one-fourth of the total range to which we are sensitive. Cells tuned to each of many different spatial frequencies are found within each cortical module. The overall CSF is thus the envelope of the contrast sensitivity functions

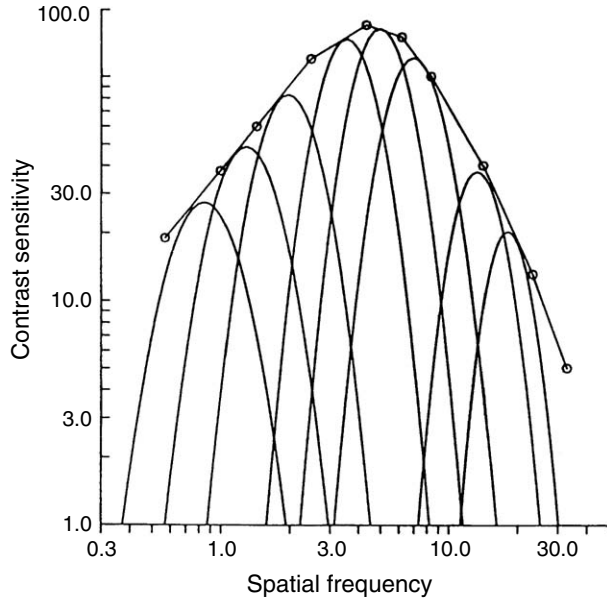


Figure 2 A diagram of how the overall contrast sensitivity function reflects the sensitivities of multiple separate spatial frequency channels whose selectivities correspond to that of various cells within a single module of a foveally related region of striate cortex.

of these multiple units tuned to various different spatial frequency ranges (Fig. 2). Cells in the foveal representation are tuned to various points covering the total spatial frequency range to which we are sensitive, but those cells tuned to the highest spatial frequencies are progressively lost with increasing retinal eccentricity. In V1 are also found, for the first time, neurons selective for the spatial frequency of chromatic- as well as luminance-varying patterns.

Evidence for the kinds of selectivity found in V1 neurons comes from psychophysical as well as physiological studies. After inspecting a high-contrast grating of a particular spatial frequency and orientation for approximately 1 min, an observer becomes less sensitive to that pattern for a short period of time. This reflects a type of gain control process by which the visual system deemphasizes unchanging aspects of the environment to facilitate the capture of information about change. However, adaptation to a grating of one spatial frequency produces no loss in sensitivity to patterns of a quite different spatial frequency, thus indicating the presence of cells with fairly narrow spatial frequency tuning and of multiple spatial frequency channels. This loss in sensitivity to a grating of one spatial frequency may also result in a temporary change in apparent spatial frequency of patterns seen shortly afterwards. A grating of a slightly higher spatial frequency than the adapting pattern will appear to be higher still, and one lower in frequency will appear to be lower still. Similar selective sensitivity losses and aftereffects are seen following adaptation to a pattern of a particular orientation.

Masking experiments also reveal spatial frequency and orientation selectivity. Detection of a pattern of a particular spatial frequency and/or orientation is interfered with (masked) by a high-contrast grating of similar spatial frequency and orientation but not by gratings of very different spatial frequency or orientation. The relatively narrow bandwidths of the masking and adaptation losses in spatial frequency and orientation sensitivity are similar to the bandwidths of spatial frequency- and orientation-selective cells in the striate cortex.

Cells with both particular orientation and particular spatial frequency tuning encode the two-dimensional spatial frequency spectrum of the pattern within the small region of the visual world with which they deal. At the receptors, spatial information is encoded purely in the space domain. Through the retina and particularly at the striate cortex, spatial information is transformed from the space domain into the local spatial frequency domain (combined space and spatial frequency) by filtering through these multiple cells within each cortical module that are selective for a certain narrow range of spatial frequencies and orientations (Fig. 3). Given the typical orientation and spatial frequency bandwidths of cells, the two-dimensional spatial frequency spectrum of a pattern could be specified by the relative activity of cells tuned to each of about six different spatial frequencies at each of six orientations, thus about 36 cells in each region. A sine-wave grating pattern of a particular spatial frequency and orientation might significantly activate only 1 of these 36 cells; more complex patterns would activate a small number of cells. Therefore, which few cells in a region fire would uniquely define the spectrum of the local stimulus in that region.

Spatial frequency and orientation selectivity not only play important roles in encoding useful spatial aspects of a stimulus but also contribute crucially to two other types of selectivity seen at the striate cortex level: selectivities for motion and for stereoscopic depth.

About 20% of V1 cells will respond to a pattern of a particular two-dimensional spatial frequency only if that pattern is moving in the cell's preferred direction. Such neurons thus extract a local motion signal. A moving object is by definition at one location at time 1 and at a different location at time 2. The motion can be

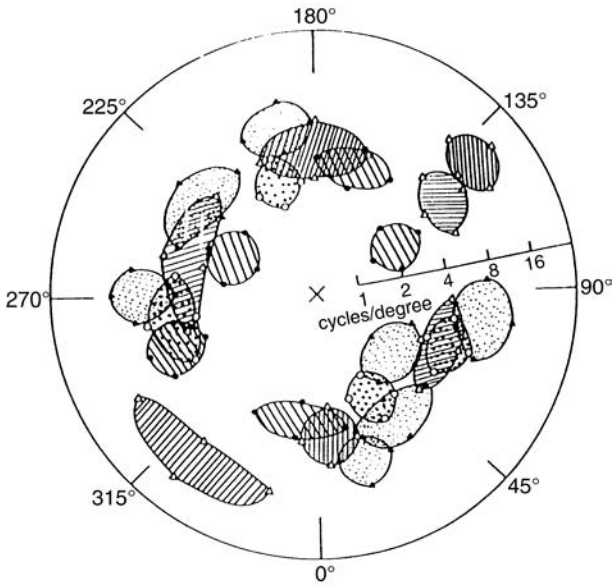


Figure 3 The two-dimensional spatial frequency response fields of 20 neurons recorded from within a single region of macaque monkey striate cortex. It can be seen that each of these cells responds to only a limited range of spatial frequencies and orientations (reproduced with permission from De Valois *et al*., *Vision Res*. **22,** 545–559, 1982).

detected by comparing the activity at these two locations at these two times. However, deciding what to compare with what at the two time intervals is a nontrivial task known as the correspondence problem. It can be demonstrated by considering how one might detect the motion of certain dots in a random dot pattern, in which each dot at time 1 could potentially be related to any one of several different identical dots at time 2, with each combination signaling local motion in a different direction. With prefiltering of the pattern into multiple local two-dimensional spatial frequency channels, the problem is vastly simplified. Neighboring regions would differ in their two-dimensional spatial frequency spectra rather than having numerous identical dots. Because only two local patterns of the same orientation and spatial frequency at the two different times would be combined, there would thus be many fewer false matches. Within each cortical module, dealing with stimuli from a certain small region of the visual field, there are thus multiple local motion detectors. Each responds only to patterns of a particular spatial frequency moving in a particular direction.

The light from a point that is binocularly fixated stimulates corresponding locations in the two eyes (e.g., a point 2° to the right of fixation in the visual field would stimulate a point 2° to the temporal side of the fovea in left eye and a point 2° to the nasal side in the right eye). Objects in the visual field that are closer or further than the fixation plane, however, stimulate noncorresponding, disparate points in the two eyes, with the direction and amount of disparity signaling whether the point is nearer or further than the fixation plane and by what distance. The visual system uses this stereoscopic information as one of many cues to depth. The local binocular disparity is computed in the striate cortex by cells that combine input from the two eyes. Some cells receive input from corresponding regions in the two eyes and respond maximally to objects in the fixation plane. Other cells receive binocular input from various noncorresponding regions in the two eyes and respond best to objects in front of or behind the fixation plane. Again, as in the case of motion, only cells with similar two-dimensional spatial frequency tuning are combined, thus simplifying the correspondence problem. As with motion, there are separate systems processing stereoscopic information over each of a number of different spatial frequencies.

The spatial frequency and orientation filtering of inputs to the motion and stereoscopic systems not only contributes to solving the correspondence problem —that of knowing what to combine from one eye or at one moment with what from the other eye or next moment—but also allows for the perception of transparency. Often, there are different objects at different depths within a given region, such as a lion seen through grass in a field, or two objects moving in different directions in a region. The presence of multiple motion and stereoscopic systems, each selective for a different spatial frequency range, allows the visual system to deal with these situations.

Those with trichromatic color vision (about 99% of the human population) have in effect at least three of each of the previously mentioned sets of maps; we can discriminate the spatial frequency, the orientation, the motion direction, and the depth of patterns along each of three color axes (e.g., red-green, yellow-blue, and black-white). In fact, there are cells in striate cortex that have a variety of color specificities, and these color-specific cells also have spatial frequency and orientation selectivity, suggesting that local color-spatial maps (perhaps more than three) are produced at this level.

There are two quite distinctive cell types in the striate cortex, termed simple and complex cells, and these reflect two successive levels of processing. Like cells earlier in the pathway, simple cortical cells are color specific (using color in the broad sense to include black and white), and thus in response to grating patterns they are selective for spatial phase. The RF of a simple

cell might consist of an elongated region responsive to light decrements (black), with flanking regions responsive to white. This cell will be most sensitive to a black-white grating of a particular spatial frequency and orientation, in such a location that the black bars fall on the RF center and the white bars on the flanking sidebands. If black and white are reversed (if the grating is shifted 180°), the cell will be inhibited and a 90° shift will produce no response.

Complex cells within a given region of the striate cortex have much the same orientation and spatial frequency tuning as simple cells in the same cortical column, but they lack phase selectivity. In the previous example, a complex cell would respond to the black-white grating regardless of its spatial phase—to either a white or a black bar (or to red or green) in any position within its RF. Complex cells are first found in the striate cortex, and they are the predominant cell type in later cortical regions.

Simple cells have little if any maintained discharge in the absence of stimulation. Since they cannot fire less than zero even to a highly inhibitory stimulus, they give a half-wave rectified output to a grating pattern in different locations. Complex cells appear to sum the half-wave-rectified (and approximately squared since most simple cells have an expansive response nonlinearity) outputs of simple cells in different spatial phases in a local region. They thus maintain the two-dimensional spatial frequency selectivity of the simple cells but lose the phase selectivity.

Complex cells undoubtedly play many important roles in spatial vision. One probable role is in texture perception. Although some objects are spatially uniform over their extent, like a sheet of white paper, most objects and many backgrounds are textured to various extents. By characterizing the spatial spectrum of such textured regions, the visual system can differentiate an object from its background or distinguish one background from another. However, it is not necessary to identify the location of all the millions of edges in a patch of sand to distinguish it from say a patch of grass. Which cell, out of an array of complex cells tuned to different two-dimensional spatial frequency ranges, is most active would provide a very economical abstraction of the needed information over a small region.

VIII. LOCAL VERSUS GLOBAL PROCESSING

The RFs of bipolar cells extend over only a tiny area comprising just a small number of receptors; the RFs of ganglion cells are only slightly larger. At these retinal levels, each element in the visual system processes information locally, over just a few minutes of arc in the fovea. One of the primary changes that occur at the striate cortex is the development, in the foveal representation as elsewhere, of RFs of various sizes. These form the substrate for multiple spatial frequency channels. The largest of these striate cortex RFs can subtend a degree or more, even in the foveal representation. An area 1° in diameter in the fovea encompasses inputs from approximately 20,000 cones, but this is still a very small region compared with the sizes of many visual objects that the system must analyze. For instance, a person's face at ordinary conversational distance may subtend 10° or more. The processing beyond V1 must have a more global character, and one of the few clear facts we know about the properties of cells in various later cortical regions is that the RFs become increasingly larger at successive processing stages. Many neurons in the parietal and temporal lobes have RFs that encompass half the visual field or more. From the stand-point of the earlier discussion of retinotopic versus functional organizations, with successive synapses through the visual path there is a progressive decrease in a spatial, retinotopic representation and an increase in a functional representation.

More global processing is required in later cortical areas not only to encompass large objects but also to resolve ambiguities in local signals. One example is that of separating luminance variations due to reflected objects from those due to the illuminant. As discussed earlier, the primary assumption the visual system uses in accomplishing this task is that light from the illuminant is uniform across space. The attenuation of responses to spatially uniform patterns begins in the retina, but the larger RFs of neurons later in the pathway allow the system to discount the illuminant over larger areas than was possible in the retina.

A second type of stimulus ambiguity that can only be resolved at higher cortical levels is related to motion. Within the image of an object moving in a particular direction, local contours may be moving in a variety of different directions. A more global analysis is required to detect the overall direction of pattern movement.

IX. SEGMENTATION OF IMAGE INTO OBJECTS AND BACKGROUND

When we look at a natural visual scene, we perceive it to be made up of various objects: in a rural scene,

perhaps a group of trees, a cow or two in a grassy field, a barn, and a stream. It is so easy and natural for us to see the world in this way that it is difficult to realize that segmentation of the image into separate objects against a background is not automatically given by the visual input to the eye but rather is a product of visual processing. It represents a theoretical conclusion about the world reached after an elaborate analysis of the input by the visual system. The vast array of numbers put out by the receptors are transformed by the processing up through the striate cortex into the local two-dimensional spatial frequency domain. Although this domain has certain advantages for representing the spatial information, it is clear that the cells at this level "know" almost nothing about objects. They respond to the pattern in their local regions in much the same way whether that pattern is part of one object or another or just part of the background. A particular striate cortex cell might respond virtually identically to a vertical trunk of a tree, a vertical stripe on a zebra, or a vertical shadow of a branch as long as each of these patterns had the same two-dimensional spatial frequency content within the local region covered by the cell's RF. To parse a visual scene into objects requires a higher level, a more global analysis than takes place in the striate cortex. There is increasing evidence that the RFs of V1 cells are more complex than initially thought, with the responses to patterns within the classical RF being influenced to some extent by stimuli in neighboring regions and by feedback from later cortical areas. However, the influences of the "nonclassical" RFs of striate cortex cells are at best the first stage in object segregation. There must be further processing.

The problem of image segmentation, the grouping of local image components into visual objects, was first recognized and extensively studied in perceptual experiments by the Gestalt psychologists at the beginning of the 20th century. They explicitly raised the question of the principles by which the visual system puts together some but not other local components to form the percept of an object and to recognize its shape. Among the principles they identified were those of proximity (nearby components are more likely to be grouped together and seen as parts of one object than components further apart), similarity (components similar in characteristics such as color or orientation tend to be perceptually grouped together as parts of the same object), common fate (grouping can be based on common motion direction), and good continuation (a smooth contour is more likely than one with an abrupt change in curvature to be seen as forming the contour of a single object). The relative depths of different parts of a scene, based on stereopsis or other kinds of depth information, also provide powerful cues for breaking a scene into different objects and separating objects from the background. Although there have been many modern investigations of grouping principles, we understand relatively little about the rules for how the different grouping principles combine in image segmentation and about what happens when the different grouping cues conflict.

The early stages of neural processing through V1 provide analyses necessary for object segregation, with cells selective for spatial frequency, orientation, motion, color, and stereopsis. We can distinguish thousands of different colors and brightnesses, and this allows us to distinguish the contours of a chair of almost any color, for instance, on almost any color of background. However, it would not be feasible to have separate object segregation and form recognition systems for chairs of each of thousands of different colors against each of thousands of different backgrounds. Furthermore, upholstered chairs may differ from their backgrounds in thousands of different textures as well as in color. Required for scene segmentation and object recognition is a system that uses the information from the selectivities of V1 cells but then generalizes across various dimensions and combines like information from neighboring regions.

Complex cells in V1 generalize across dimensions to a limited extent in that they respond to a pattern of a particular spatial frequency and orientation regardless of color or spatial phase. This type of processing is further extended in V2, in which some cells have an orientation and direction tuning to contours regardless of whether the contour is formed by luminance or texture. The circuitry by which such cells are constructed has not been established, but a plausible mechanism is that these V2 cells sum the outputs of those cells in extended regions of V1 that have the same orientation and/or direction tuning but different spatial frequency and color tuning.

A common problem faced by the visual system in image segmentation and object identification is that an object is often partially occluded by other objects. A deer in a forest, for instance, might be partially obscured by a nearby branch. Such occlusion not only changes the shape of the object's image, thus making object recognition more difficult; it may also break the image of a single object into separate pieces, thus making it difficult to see all the separate parts as

belonging to a single object. Some cells in the second cortical visual area (V2) appear to deal with the problem of occlusion over limited regions. They respond to two colinear borders even if there is a gap in the middle, as would be produced by a nearby occluding object.

X. TWO PROCESSING STREAMS

Approximately 35% of the human cortex is involved in visual processing (depending on how one wants to categorize the late stages of the process), including all of the occipital cortex and large parts of the parietal and temporal cortices. Approximately 25 visual cortical areas have been identified anatomically, with V1 and V2 being the largest. Some of these areas are labeled on the basis of presumed sequential order (V1–V5), and others are labeled on the basis of anatomical location with respect to various other anatomical landmarks such as gyri (in the macaque monkey). If one diagrams the interconnectivity among these regions, the resulting map resembles something that might have been produced by a drunken spider. Everything is connected to everything else. However, a consideration of just the largest pathways reveals evidence for sequential processing and for two main processing streams from the striate cortex going anterior (hence "prestriate") in the brain. One of the streams is from cells in V1 layer IVB to MT, MST, and from there to various regions in the parietal lobe. The other path is from cells in V1 layers II and III to V2, V4, and from there to various regions in the temporal lobe.

The temporal stream appears to be principally involved in processing information about form and color, and the parietal stream processes information about motion and spatial location, but this division of labor should not be taken too literally. Not only are there numerous anatomical connections between cells in these two paths but also it is clear from functional considerations that the two systems cannot be independent. For instance, motion is one of the strongest cues for the segregation and identification of form, so its analysis cannot be completely independent of the form system. Nonetheless, there is considerable evidence for some degree of specialization of function among the various prestriate visual regions and for this rough division between "what" (temporal) and "where" (parietal) systems.

Lesions in the temporal lobe of humans sometimes lead to loss of the ability to recognize certain types of objects or specific characteristics of objects. For instance, lesions in a particular temporal lobe region may lead to a loss of the ability to recognize the color of objects (although basic color discrimination ability may be unaffected). Strokes affecting certain parts of the temporal lobe produce a loss of the ability to recognize faces. For both monkeys and humans, hands and especially faces are of great importance, both for recognition of different individuals and for social interactions and communication. To distinguish one person from another on the basis of the slight differences in facial features is surely among the most complex of visual tasks. It is perhaps not too surprising that there are many cells in the temporal lobe that seem to be involved specifically in processing the images of faces and hands. The visual system seems to treat these particular stimuli in a special way. Unfortunately, there are no good theories or data on the transformations involved in going from cells in V2 and V4 (which appear to be only marginally different from striate cortex cells except for their larger RFs) to the neurons in the temporal lobe that have such complex tuning properties.

At early and middle stages in visual processing (through the striate cortex and perhaps V2), the visual system processes information from the whole visual field simultaneously and in parallel. It is of course such massive parallel processing that allows the very slow neural elements to carry out the huge computations required in real time. It is apparent from psychophysical experiments, supported by recordings from cells in V4 and later, that at later stages the system continues to process only selected portions of the scene—those parts to which one is attending. There is in fact some indication from patients with Balint's syndrome that object recognition may operate on only one object at a time.

Lesions in the parietal lobe produce quite different types of visual losses from those in the temporal lobe. Lesions in a particular occipitoparietal region that presumably correspond to area MT in macaque monkeys can lead to a loss of the ability to see motion, with no loss in the ability to recognize objects. Lesions in other parietal regions (presumably later in the processing stream) can leave an individual with the ability to recognize whatever object is attended to but an inability to determine where that object is in the visual field. Such an individual may be totally unaware of the presence of other, unattended objects, including ones that had been attended to and recognized just a moment earlier. This suggests that the temporal lobe systems identify objects one at a time and feed this information to the parietal lobe regions, which then fit

them into their appropriate relative locations in the visual field, like a collage of separate photographs, to construct our perceptual world.

One reason for having multiple visual areas is the need for analysis over large visual extents in order to accomplish object segregation and image recognition. Cells in V1 carry all the information about all aspects of visual perception since the whole path goes through this area. A typical visual object may extend over several degrees, and the neurons in V1 responding to different portions of it may be separated by tens of millions of other cells. For all cells to interact with each other over such a vast array would require an enormous set of interconnections. However, if the cells from all parts of V1 processing a certain type of information (e.g., motion) were brought together in a separate region, they would comprise a sufficiently small subpopulation to permit interactions over greater distances in visual space without requiring an impossible number of interconnections.

The visual system may also have evolved separate parallel processing areas to bring together cells processing common features so as to organize them in optimal ways for their particular functions. In the striate cortex, cells are organized in columns selective for orientation and spatial frequency. Within a motion area, however, a columnar organization by motion direction, which is seen in area MT, is presumably more useful. In a color area, a columnar organization by hue might be optimal.

See Also the Following Articles

COLOR VISION • OBJECT PERCEPTION • RECEPTIVE FIELD • SPATIAL COGNITION • VISION: BRAIN MECHANISMS • VISUAL DISORDERS • VISUAL SYSTEM DEVELOPMENT AND NEURAL ACTIVITY

Suggested Reading

De Valois, R .L., and De Valois, K. K. (1988). *Spatial Vision*. Oxford Univ. Press, New York.

Graham, N. (1989). *Visual Pattern Analyzers*. Oxford Univ. Press, New York.

Movshon, J. A., and Landy, M. S. (1991). *Computational Models of Visual Processing*. MIT Press, Cambridge, MA.

Spillmann, L., and Werner, J. S. (1990). *Visual Perception: The Neurophysiological Foundations*. Academic Press, New York.

Wandell, B. A. (1995). *Foundations of Vision*. Sinauer, Sunderland, MA.

Speech

KEITH R. KLUENDER
University of Wisconsin

GLOSSARY

categorical perception Stimuli that vary systematically along one or more dimensions are perceived relatively discretely. Identification of stimuli shifts abruptly at some point along one or more dimensions, and discrimination is much better for two stimuli that are identified differently than for two physically equidistant stimuli that are labeled the same.

formant A resonance of the vocal tract that is reflected in the speech signal as a frequency band of relatively higher energy. This band of energy is described in terms of both the center frequency and the bandwidth.

fundamental frequency The lowest frequency of a periodic signal. For speech sounds, the fundamental frequency ($f0$) reflects the frequency at which vocal folds vibrate. The harmonic spectrum consists of the fundamental frequency (first harmonic) with higher harmonics at multiples of the fundamental.

lack of invariance There is no simple one-to-one correspondence between attributes of speech sounds (or articulations) and linguistic units such as phonemes. Thus, phonemes are variably specified. The lack of necessary and sufficient attributes specifying phonemes is classically described as the problem of lack of invariance.

The most important sounds for humans are speech. Speech sounds are defined within the design and capabilities of vocal tracts, and languages have come to use different collections of sounds that are relatively distinct from one another in multiple ways. Perception of speech entails using multiple characteristics of the signal, and experience with reliable patterns across sounds within a language profoundly affects how speech sounds are heard. Understanding speech perception requires knowing both how complex signals such as speech are encoded in the auditory system, and how experience with speech shapes the use of speech sounds within a language.

I. RELATIONSHIP BETWEEN PRODUCTION AND PERCEPTION

In some ways, one could address the production and perception of speech much like one would consider other motor and perceptual behaviors. A talker executes some motor sequence engaging various structures, including the diaphragm, vocal folds, and speech articulators such as the tongue and lips. One could then study perception of the resultant sounds much as one would investigate perception of other complex sounds. In large part, this has been a fruitful approach to understanding speech perception. However, there is one aspect of speech communication that is not typical of most other motor and perceptual events. The relationship between production and perception of speech is an especially intimate one. All perceptual systems must maintain relatively close agreement between organisms and their environments, but the relationship between things that are perceived and the act of perception is not usually as close as it is for speech. A tree does not strike a pose that deliberately conveys strength and longevity, but talkers do speak so that they can be understood. The structure of speech is presented for the listener, and

listeners' perception of speech hews closely to dependable characteristics of speech production. One cannot know much about either speech production or speech perception without knowing a fair amount about the other. Most of this article focuses on aspects of speech perception, but first, some understanding of speech production is necessary.

II. SPEECH PRODUCTION

There are three basic components to production of speech: respiratory (lungs), phonatory (vocal chords), and articulatory (vocal tract). First is the respiratory system through which air is pushed out of the lungs, through the trachea, and up to the larynx. At the larynx, air must pass through the two vocal folds that are made up of muscle tissue that can be adjusted to vary how freely air passes through the opening between them. In some cases, such as voiceless sounds like [p] and [s], vocal folds are apart; they do not restrict airflow and do not vibrate. For voiced speech sounds such as [b] and [z], vocal folds are closer together, and the pressure of airflow causes them to vibrate with a periodicity referred to as the fundamental frequency ($f0$) much like a string on a musical instrument vibrates at a particular frequency. Owing to variations in the pressure of airflow from the lungs and muscular control of vocal fold tension, talkers can vary the fundamental frequency of voiced sounds. If one were to consider the nature of speech at this point, it could be depicted as a spectrum of energy spread across frequency and concentrated at the fundamental frequency and at multiples of the fundamental with decreasing energy at each successive multiple as seen in Fig. 1a.

The area above the larynx, the oral tract and the nasal tract combined, is referred to as the vocal tract. The nearly continuously varying configuration of the vocal tract is responsible for shaping the spectrum differently for different speech sounds. There are many ways the jaw, lips, tongue body, tongue tip, velum (soft palate), and other vocal tract structures can be manipulated to shape the sounds that emanate from the mouth and nose. Widening and narrowing of points along the vocal tract selectively attenuate some frequencies while making others more prominent. Figure 1b illustrates the filtering effects of the vocal tract for the vowel [a] as in "father." Figure 1c portrays the net result from passing the glottal source (Fig. 1a) through the vocal tract (Fig. 1b).

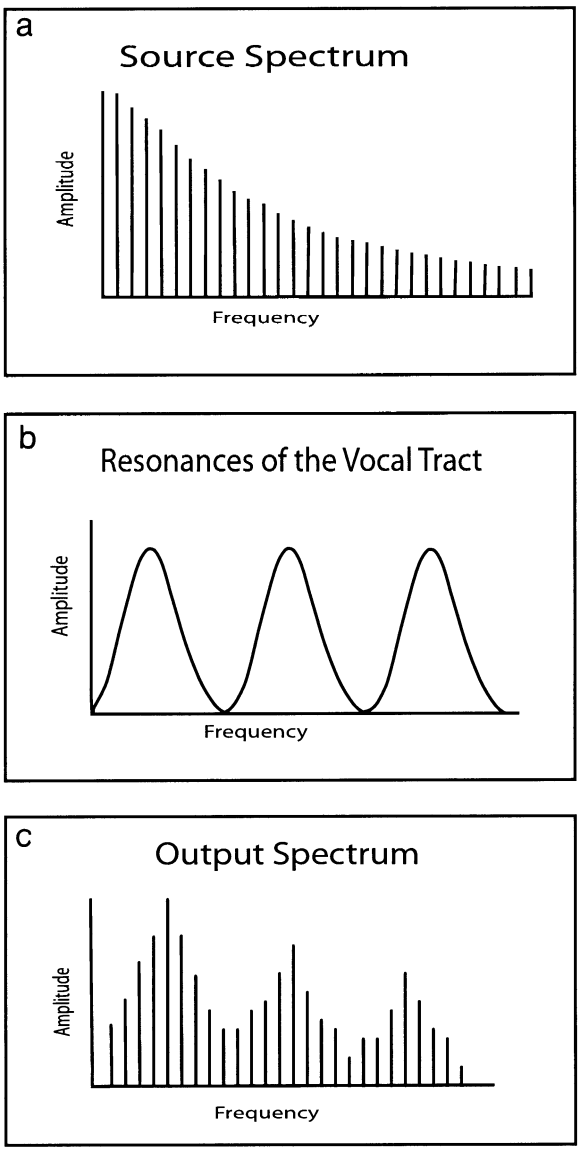


Figure 1 (a) Speech depicted as a spectrum of energy spread across frequency and concentrated at the fundamental frequency and at multiples of the fundamental with decreasing energy at each succesive multiple. (b) Filtering effects of the vocal tract for the vowel [a] as in "father." (c) Net result from passing the glottal source (a) through the vocal tract (b).

The peaks in the resultant spectrum are referred to as formants, and formants are described by number, lowest to highest (F1, F2, F3,...). Only the first three formants are depicted in Fig. 1, and for the most part, speech sounds can be identified on the basis of energy in the region of these lowest three formants. However, additional formants exist with lower amplitudes at higher frequencies (F4, F5, F6, etc.) and are

relatively more acoustically prominent in the speech of children.

Airflow can be channeled, constricted, or obstructed to produce different vowel and consonant sounds. Vowels are made with a relatively open unoccluded vocal tract. In terms of articulation, vowels vary mostly in how high or low and how forward or far back the tongue body is in the oral tract. In addition, some vowels are produced with rounded lips (e.g., [u] as in "boot") or with modestly different fundamental frequencies among other variations. Figure 2 displays the vowel sounds of English with respect to how high or low the tongue is placed, which corresponds to the frequency of the first formant (F1), and how far back or forward (front) the tongue is placed, which corresponds to the frequency of the second formant (F2). Not shown is the fact that high vowels such as [i] and [u] typically are made with higher *f*0 (more vocal fold tension).

In part because early linguists had greater access to their own vocal tracts than to sophisticated audio analysis equipment, it is a general convention that speech sounds are described in terms of the articulations necessary to produce them. In addition to variation in tongue height, frontness/backness, and lip rounding as descriptions for vowels, consonants also are described by articulatory characteristics. For example, consonants are described in terms of the manner in which constrictions are introduced along the vocal tract. Stop consonants or plosives such as [b], [p], [d], [t], [g], and [k] include complete constriction such that no air may pass through. Nasal consonants such as [n], [m], and [ŋ] (as in "sing") are like [b], [d], and [g], respectively, with complete constriction at some point in the oral tract, but air is allowed to escape through the nasal tract because the velum is lowered. Fricative consonants are caused by nearly complete obstruction of the vocal tract, with a noisy sound being produced by turbulence of airflow passing through a very small opening. Some examples of English fricatives are [s], [z], [š] (as in "ash"), and [ž] (as in "azure"). Affricate consonants are produced by a combination of complete occlusion (like a stop) followed by nearly complete occlusion (like a fricative). Examples of

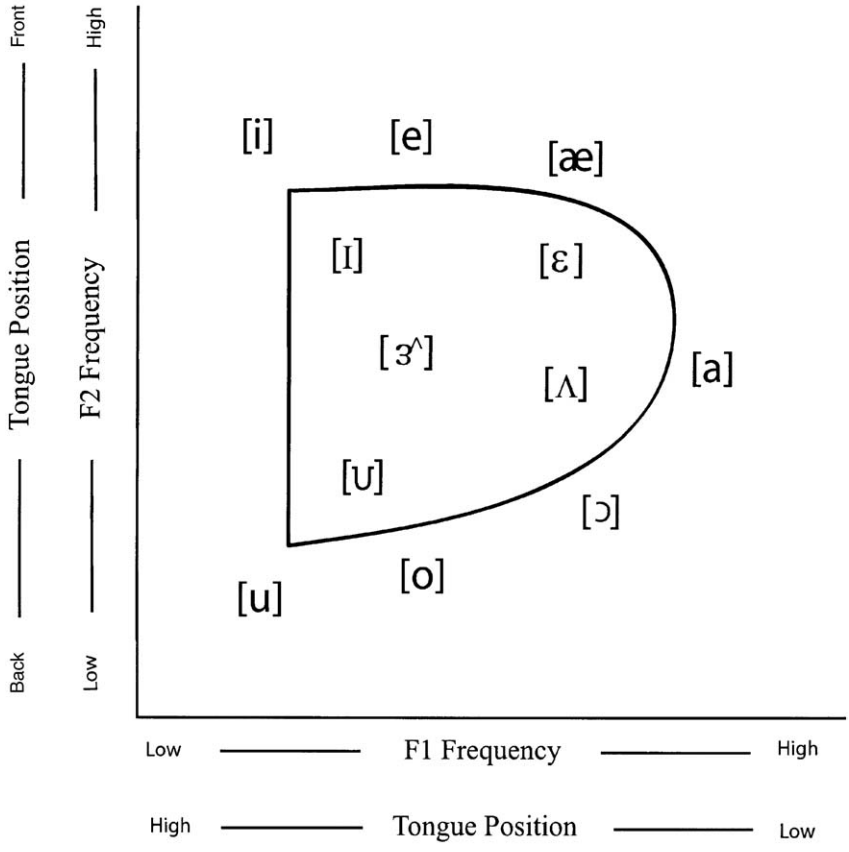


Figure 2 Vowel sounds of English defined by tongue position and acoustics.

affricates in English are [tš] (as in "chug") and [dž] (as in "jug"). The least constricted consonants in English are laterals and semivowels, such as [l], [w], [r], and [j] (as in "yet").

Consonants are also described in terms of whether the vocal folds are close and vibrating (voiced) or further apart and not vibrating (voiceless). Thus, sounds such as [b], [d], [g], [z], [ž], [dž], [l], [w], [r], and [j] are voiced, And [p], [t], [k], [s], [š], and [tš] are voiceless. Finally, consonants are described on the basis of place of articulation. Constrictions can be placed along a number of places in the oral tract. In English, the three major places of articulation are bilabial (lips: [p], [b], and [m]), alveolar (alveolar ridge behind teeth: [t], [d], and [n]), and velar (soft palate or velum; [k], [g], and [N]). The major anatomical features of vocal tracts are shown in Fig. 3.

III. LINGUISTIC SOUND SYSTEMS

This description does not exhaust all the distinctions among the approximately 40 sounds used by English, and it is a vast underdescription of variation among languages more generally. Humans are capable of producing an impressive range of sounds with potential to be included within the speech sounds of a language. Owing primarily to unique characteristics of supralaryngeal anatomy, the adult human possesses sound-producing abilities unrivaled among other organisms. This capacity is revealed in a grand assortment of more than 850 different speech sounds used contrastively by the more than 5000 distinct languages used throughout the world—more than 550 consonants and 300 vowels (including diphthongs such as [e^{y}] and [o^{w}]). Such capacity dwarfs that of other animals, being more than an order of magnitude larger than the largest inventory of nonhuman primate calls. Although the comparison between nonhuman primate vocalizations and human speech sounds is not straightforward, it can be suggested that, owing to supralaryngeal adaptations, the number of functionally different mouth sounds for humans is more than an order of magnitude larger than for any other primate. The number and variety of speech sounds producible by human vocal tracts has been taken as evidence of human specialization for speech production.

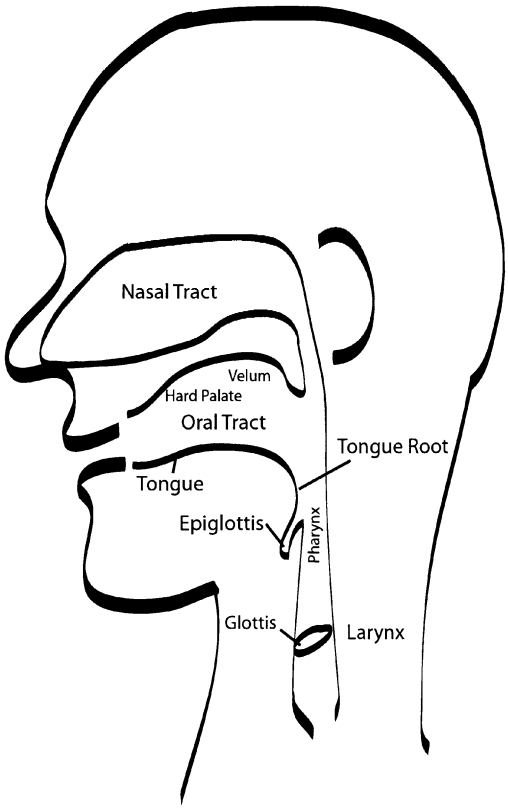


Figure 3 Selected anatomical structures of the vocal tract.

In contrast to this diversity in potential speech sounds, systematic inspection reveals that the collections of consonants and vowels used by individual languages are anything but random. The vast majority of speech sounds are relatively rare, whereas a handful are extremely common. For example, all known languages have stop consonants. The overwhelming majority of languages have three places of articulation for stop consonants, typically the three described previously for English. More than 80% of languages include a distinction in voicing (e.g., [p] and [b] vs [s] vs [z]), usually at these three main places of articulation. Regarding consonants, diversity does not imply entropy.

The structure of vowel systems is as much or more orderly than that for consonants. There is a fair amount of variety in the particular number of vowels used by languages, with some languages using as few as 3 vowels, whereas others use as many as 24. English uses about 15 depending on dialect. However, the most common number of vowels used by languages is only five, and other numbers of vowels appear to be relatively favored. Especially for the five- to nine-vowel systems that predominate across languages, the particular sets of five, seven, or nine vowels are typically used. Figure 4 displays some of the more common three, five, and seven vowel systems. Although there are relatively fewer languages that use more than seven

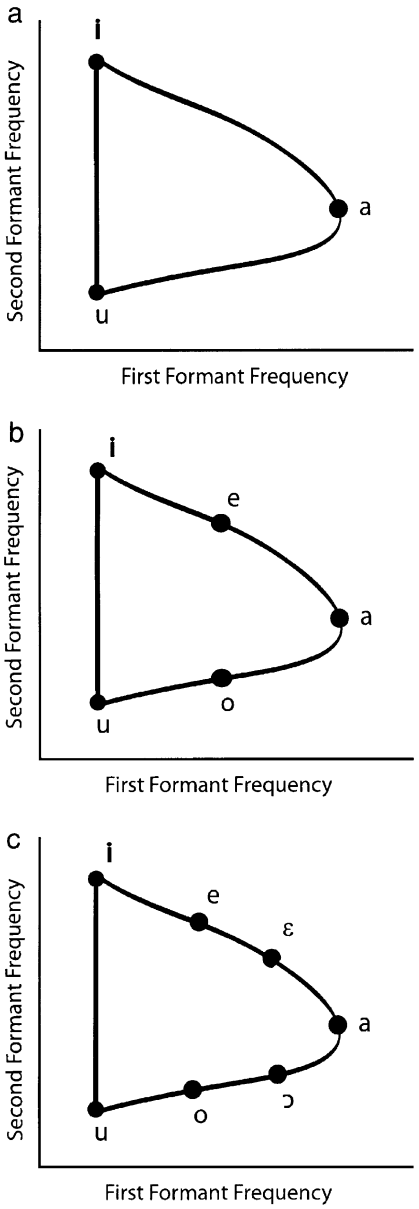


Figure 4 Common three- (a), five- (b), and seven-vowel systems (c).

vowels, there remains a good deal of commonality among systems with the same number of vowels.

What are the forces acting on languages that encourage the selection of some sounds and groups of sounds over others? It may be that some sounds are easier to produce than others. Although the number of possible speech sounds is prodigious, one guiding factor explaining regularities is how easy some sounds are to produce either in isolation or in sequence with others. The role of articulatory ease is perhaps best evidenced by the fact that languages tend to use articulatorily simpler consonants before incorporating more complex consonants. A least effort principle contributes to the selection of sounds used commonly by languages, although the precise degree to which producing some sounds may require more energy or greater coordination has proven very difficult to quantify.

Consistent with the close tethering of speech production and speech perception, there are many ways in which languages have come to use sets of speech sounds that enhance perceptual effectiveness. It is quite clear that talkers are willing to expend effort for communicative robustness. For example, the tense high vowels [i] and [u] (as in "beet" and "boot") require more effort to produce relative to their lax counterparts [ɪ] and [ʊ] (as in "bit" and "book"). The vowels [i] and [u], however, are acoustically more distinct, not only from each other but also from other possible vowel sounds. Also, across languages, these tense vowels [i] and [u] occur five times more frequently than lax vowels [ɪ] and [ʊ].

Aside from music and visual art, there may be no other instances of perceptual events being created with such deference to the perceiver's interests as is human speech. Common examples of talkers molding their utterances to the needs of the listener include instances in which conditions for communication are not optimal. Talkers speak more clearly to young or nonnative listeners for whom distinctions are not obvious. When environments are noisy or reverberant, talkers strive to produce contrasts that are maximally distinctive. Generally, speech sound repertoires of languages have developed over generations of individuals toward greater communicative effectiveness. Humans talking are not like leaves rustling or turbines whining. An important part of understanding speech perception rests on learning the ways generations of talkers have come to capitalize on auditory predisposition of listeners.

Sound systems of language communities have adapted to be fairly robust signaling devices by exploiting general characteristics of auditory systems whenever possible. An obvious way that a language community achieves such robustness is by developing an inventory of phonemes so as to optimize distinctiveness acoustically and auditorily. Inspection of regularities in vowel sounds used by languages, such as those shown in Fig. 4, provides some of the most illuminating examples of auditory processes operating as a driving force. Different languages use different sets of vowel sounds, and languages use subsets of vowels that are most easily discriminated from one

another. In particular, those vowels favored for languages with five vowels are vowels that are as acoustically distant as possible from one another. As a general rule, the set of vowels selected by a language, whether it uses 3 or 10 vowels, is composed of sounds that tend toward maximal discriminability.

There is another way that differences between speech sounds are perceptually more dependable. The speech signal is very rich owing to speech production giving rise to multiple acoustic consequences from the same basic speech articulation. Acoustic products of the larynx and vocal tract are quite complex. In many cases, multiple acoustic attributes are effortless consequences of passive interactions between articulators and/or airflow. In other cases, languages use systematic variation between relatively independent articulatory maneuvers to enhance auditory distinctiveness. One well-known example of multiple attributes and redundancy involves the distinction between voiced [b] and voiceless [p] when they occur between two vowels such as in [aba] and [apa]. There are at least 16 identifiable acoustic differences between these two utterances. The following are among the most significant for [aba]: The duration of silence (corresponding to closure of the lips) between the two syllables is shorter, the initial vowel is longer, there is more low-frequency energy (corresponding to vocal fold vibration) between the vowels; and the frequency of F1 both ends lower and begins lower around the medial silence. There are many other instances such as this, with the most consistent theme being that there are no distinctions that are signaled by only a single acoustic attribute.

In general, there are a large number of ways sound systems of languages have come to be structured in a manner that exploits characteristics of auditory systems. It bears note that ease of production and perception do not explain all the variance across inventories of speech sounds because many cultural or sociolinguistic factors play a significant role. However, beginning with the most basic descriptions of speech, the synergy between talkers and listeners is a robust characteristic of speech communication.

IV. CONTRIBUTIONS OF THE PERIPHERAL AUDITORY SENSORY SYSTEM

Languages have developed to be robust signaling systems, and distinctions between speech sounds do not rely on the ability to make fine-grained discriminations bordering on thresholds of auditory systems. In this way, perception of speech bears little likeness to classic psychophysical studies of absolute thresholds and just noticeable differences. Determining the differences between vowels such as [æ] (as in ''bat'') and [ɛ] (as in ''bet''), on the basis of gross spectral differences, shares little in common with psychoacoustic studies that demonstrate humans' abilities to detect changes as small as 1 Hz from a 1000-Hz sinusoid tone. It is widely appreciated that the performances of listeners detecting tones in quiet or detecting differences in pitch between two tones are poor predictors of the ability of the same listeners to understand speech. Listeners who suffer significant hearing loss can nonetheless manage to understand speech until the level of impairment becomes severe or there is substantial background noise competing with the speech signal. Perhaps predictably, the ability of normal-hearing listeners to understand severely degraded speech is impressive. If the entire signal is filtered away above 1.5 kHz, listeners can still understand it nearly perfectly. If, instead, one filters away all energy below 1.5 kHz, listeners can also understand it nearly perfectly. Also, across a fairly wide range of frequencies, one can present only a relatively narrow frequency band of the speech signal and still achieve high intelligibility. Profoundly deaf people who have received cochlear prosthetics (electrodes implanted into the cochlea) and who consequently receive extremely degraded signal via as few as one to four electrode contacts, can often understand speech, sometimes sufficiently well to talk on the telephone.

Because languages use distinctions that maximize acoustic differences and exploit auditory capacities, several observations follow. First, human infants are quite proficient at discriminating differences between speech sounds from a very early age. Three decades of studies document the impressive abilities of human infants, some less than 1 week old, to discriminate a wide variety of consonants and vowels from across many languages. A plethora of positive findings indicate that infants have the discriminative capacity necessary for most or all of the speech distinctions they will need to use in their language. Languages tend to use distinctions that are relatively easy to detect, and infant auditory abilities appear to be quite well developed. By 3 months of age, the human auditory system is nearly adult-like in absolute sensitivity and frequency-resolving power within the frequency range of most speech sounds.

A second observation, perhaps more telling than the first, is that discrimination of speech contrasts by nonhuman animals appears to be generally quite good.

There have been a number of demonstrations that animals can distinguish human speech sounds with facility. Differences between vowel sounds present no apparent difficulty for nonhuman animals. Several species have been shown to distinguish voiced from voiceless stop consonants or to distinguish consonants on the basis of place of articulation (even beyond the basic three places described previously for English). Although such findings are expected on the basis of languages tending to use distinctions that are relatively easy on ears, currently the number of speech distinctions tested with animals is small relative to the very large set of possible contrasts. Also, the number of animal species used is limited.

Because nonhuman animals appear to provide an adequate model for simpler aspects of speech perception, such as distinguishing one sound from another, many neurophysiological studies have been conducted in the interest of describing neural representation of speech sounds. Most of this work has focused on neural responses in the auditory (VIIIth) nerve, but there have been a number of studies concerning successive stages of the auditory system, such as ventral cochlear nucleus, inferior colliculus, medial geniculate, and auditory cortex. Owing to the preponderance of work having been done with auditory nerve, there is greater understanding of the neural representation of speech at this level.

For the most part, neural representations of speech at peripheral levels [e.g., auditory nerve (AN)] are fairly well predicted by neural responses to other sorts of simpler sounds. Individual AN fibers have their origin at inner hair cells in the basilar membrane of the cochlea. Just as the basilar membrane embodies a tonotopic mechanical sensitivity, AN fibers are most sensitive to particular frequencies [characteristic frequencies (CFs)] as a function of the region of the basilar membrane from which they originate. AN fibers have greater neural firing rates at their CF, and the rate of neural firing diminishes for frequencies further from CF. Attenuation firing to frequencies not at CF is much more pronounced for frequencies higher than CF relative to those lower than CF. In addition, the extent to which energy lower than CF encourages firing is highly dependent on the intensity of the signal. Although individual fibers appear to be fine-tuned to a limited frequency range at or near threshold, this is not the case for higher intensity sounds. This broadening of tuning at higher stimulus levels is most pronounced toward lower frequencies. For example, at 70-dB sound pressure level (SPL), which is representative of conversational speech, an AN fiber with CF of 8 kHz may respond reliably to energy below 1 kHz, possibly including energy in the region of F1 or even *f*0. Neural firing rates across frequencies for an AN fiber with CF of 2 kHz are shown in Fig. 5, which shows that, at intensity levels common for speech perception, individual fibers are not narrowly tuned to detect fine-grained spectral differences. Because the intact organism is not limited to responses of an individual fiber, greater spectral resolution could be gained by a pattern of responses across a population of AN fibers.

Although frequency selectivity becomes broadened at higher levels, rate as a measure of neural response also has the appearance of being compromised at intensity levels common for speech. The effective range of hearing is much greater than the dynamic range of firing for individual AN fibers. Although human listeners can hear loudness differences over a range from 1 to 120 dB (about 10^{12}:1), the range of neural firing rates is only about 0–1000 spikes per second (10^3:1). For most neurons, increases in rate of neural firing occur over only the first 30–40 dB above threshold before a steady maximum is reached. Because not all AN fibers have the same threshold and are sensitive to changes in intensity over different intensity ranges, a broader range of intensities can be encoded across a broader population of neurons.

Changes in firing rate are not the only source of information in the firing patterns of AN fibers. The firing of neurons is not temporally random. Instead, at a given signal level, the probability of firing at any given point in time is related to the phase of some frequency or frequencies in the signal to which the neuron is sensitive. The term ''neural synchrony'' is typically used to refer to synchronization between the firing pattern of a neuron and the frequency or frequencies of the signal. In response to lower frequency signals (<4 kHz) fibers tend to fire predominantly during a restricted portion of the stimulus cycle. Synchrony encoding of lower frequency energy can be observed in AN fibers with CFs at much higher frequencies, presumably because at moderate to high SPLs typical of conversational speech filtering characteristics of AN fibers are broadened toward lower frequencies. Consequently, synchrony (versus rate) encoding of spectral information, particularly for *f*0 and F1, could be of substantial value for speech. However, there is a complication. Although synchrony encoding in the auditory nerve appears robust for lower and mid frequencies, this temporal encoding seems to become more diminished in ascending auditory structures. By the level of the cortex, thalamocortical temporal encoding appears to have an

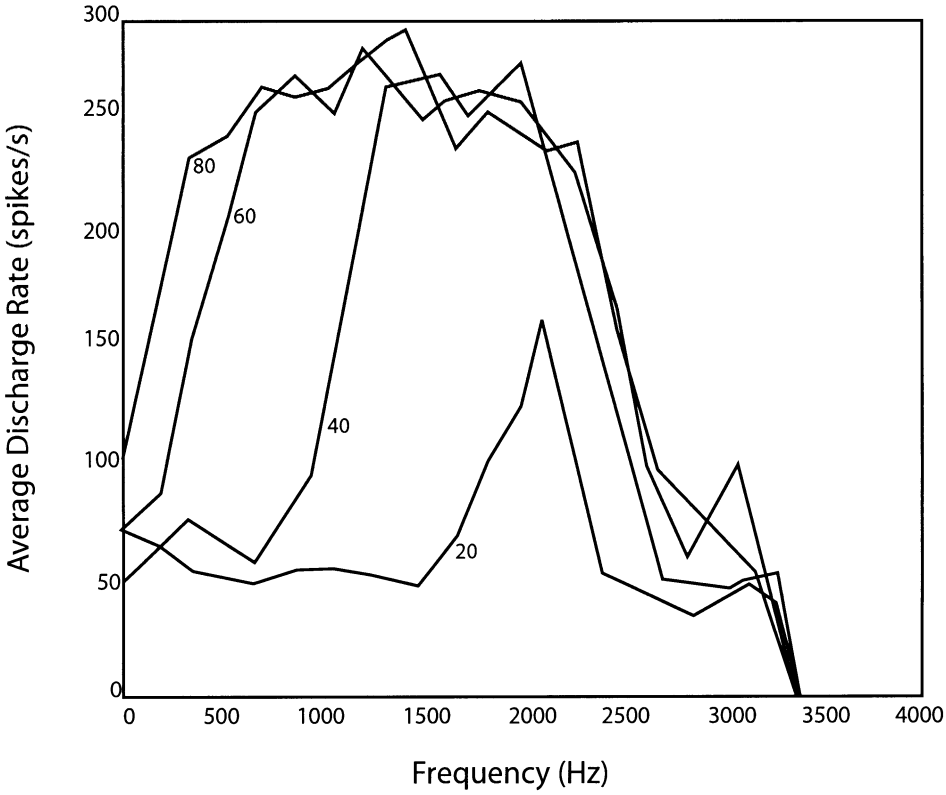


Figure 5 Neural firing rates across frequencies for auditory nerve fiber with characteristic frequency of 2000 Hz.

upper limit of 400–600 Hz. Either the auditory system does not exploit temporal encoding despite apparent virtues at intensity levels characteristic of speech or synchrony encoding is translated to some other form of neural code at higher levels of processing.

Taken together, there are a number of important issues to consider with respect to encoding of speech and other complex sounds in the auditory periphery. There is sufficient information about the ability of humans and animals to make use of acoustic information necessary for speech perception that it is clear that apparent limitations on the separate contributions of individual neurons are not realized as real obstacles to understanding of speech. In part, this is because ample information is conveyed across the assemblage of 15,000 AN fibers. In part, it is because high-fidelity representations are unnecessary. Finally, redundancy in the signal makes it possible for one or more of the details of the stimulus to be compromised with little or no decrement in performance. Subsequent work will better illuminate the contributions of populations of neurons at low and at higher levels. In the remainder of the article, discussion of cortical organization for speech perception is presented following the depiction of more sophisticated aspects of speech perception, including the use of multiple stimulus attributes.

V. CONTRIBUTIONS OF EXPERIENCE

Producing and perceiving differences between speech sounds is only a part of understanding how speech is perceived. One reason is because the mapping between production and perception of speech is complex. As noted previously speech production results in a host of identifiable acoustic attributes. Although one can view this as a positive state of affairs owing to redundant acoustic attributes, the matter is complicated by the fact that no single acoustic attribute dependably signals a given consonant or vowel. It is generally accepted that consonants and vowels do not have unique identifying properties, and this is referred to as the classic problem (and double negative) of "lack of invariance." The problem of variance or variability is that, for most or all phonemes, there are no individually necessary and sufficient cues that uniquely identify

speech sounds. Although there have been many attempts to identify invariant cues, these have not been broadly successful. In general, multiple acoustic attributes of the speech signal must be combined in some way to identify speech sounds. Even the most definitive description of individual acoustic or auditory attributes falls short of explaining how multiple attributes are combined to perceive the consonants and vowels of one's language.

Consequently, the ability to discriminate speech sounds on the basis of some acoustic attribute is not synonymous with the functional use of speech as part of language. Although a relatively complete understanding of the auditory representation of speech sounds is an essential step in understanding how consonant and vowels are recognized, mapping acoustic information onto either neural representations or some more abstract perceptual space is only part of understanding speech perception.

Among the many variations in the speech signal, some are relevant to the linguistic message and many are not. Part of speech perception includes ignoring linguistically irrelevant variation while using relevant changes. For example, different talkers produce different renditions of speech sounds. Males have lower pitched voices not only because their vocal folds are larger following puberty but also because males are taller than women in general and their vocal tracts are longer resulting in lower formant frequencies. Children, of course, produce speech that is acoustically distinct from that of adults, and the acoustics of their speech change through age. Listeners must recover sequences of consonants and vowels despite substantial changes to the speech signal owing to differences in talker characteristics. Other changes that do not change the linguistic content of speech include acoustic consequences of distance, room reverberation, or competing sounds in the environment including the voices of other talkers.

Finally, because many speech sounds are used across the languages of the world, another very substantial problem arises. Perceiving distinctions in one's language requires becoming attuned to one's native language. Acoustic differences that matter critically for one language may be irrelevant or even distracting with respect to distinctions from another language. In order to perceive speech in a way that is appropriate to one's language, speech perception must be tuned in a way that most of the many possible differences between speech sounds are relatively ignored. In contrast, distinctions between speech sounds that are used by one's native language must be preserved or even enhanced. This ability to use some acoustic distinctions between sounds but not others is part of what defines being a native listener.

Many readers have had the experience of listening to speech produced by native speakers of Spanish, one of the many languages that uses only the five vowels [i] (an in "beet"), [u] (as in "boot"), [a] (as in "hot"), [e] (as in "bake"), and [o] (as in (boat"). Consequently, a native Spanish speaker may produce English [ɪ] (as in "bit") as [i] or English [ɛ] (as in "bet") as [e]. These differences in speaking are not due simply to being incapable of producing English vowels. In part, these differences are due to difficulty clearly perceiving distinctions between English vowels following experience with only Spanish. Many can remember such difficulties as high school students learning a new language, with both perceiving and producing differences between sounds in the new language being difficult. Perhaps the most widely appreciated effects of experience are those found for native speakers/listeners of Japanese whereby both perception and production of the distinction between English [r] and [l] can be dramatically diminished.

Now, consider the prospects for an infant. How is it that, whether born in New Delhi, New England, or New Guinea, an infant comes to hear speech contrasts in a fashion appropriate to Hindi, English, or Tok Pisin, respectively? In addition to becoming attuned to useful contrasts within the native language, infants must come to understand the consonants and vowels of their language independent of whether father, mother, sister, brother, or stranger utters them.

Many of the most difficult problems in understanding speech perception concern how listeners come to be able to understand speech on the basis of multiple acoustic attributes, some of which are helpful in a given language, some of which are not, and none of which are wholly reliable.

A. Lack of Invariance

Even for utterances by the same speaker, acoustic qualities of a consonant or vowel could be different, thus giving rise to the concept of lack of invariance. The major reason for this is that production of a consonant or vowel varies depending on the consonants or vowels that precede and follow. This context sensitivity is a consequence of coarticulation, the temporal spatial overlap in production of successive sounds. One classic example is the stop consonant [d] as depicted schematically in Fig. 6. Depending on the following vowel,

changes in formants across time (referred to as formant transitions) can differ dramatically. This is only one of many examples of lack of invariance, and this context sensitivity resulting from coarticulation has presented one of the most substantial difficulties in developing computer recognition of speech.

For some time, the fact that human listeners effortlessly understand speech despite context sensitivity suggested to many investigators that there might be something unique about human speech perception relative to other forms of perception. In particular, it was suggested that because acoustics were not sufficiently invariant to signal reliably vowels or consonants such as [d] across different vowel contexts, listeners must instead make access to something that was assumed to be more nearly invariant—the articulation of speech sounds. During the past 40 years, a number of theories of speech perception have been proposed that, in one way or another, have suggested that perception must be accomplished through some sort of reference to articulation. The most significant and widely influential is the motor theory of speech perception advocated by Alvin Liberman. In several ways, such an idea is very attractive. As noted previously, humans are the only organisms with such an exquisite vocal repertoire, quite likely owing to vocal tracts that are specially adapted for production of speech. Even if perceptual theories in other domains do not rely on some link between producing some object or event and perceiving it, it seems reasonable that speech may provide a special case for which humans have evolved specialized perceptual processes complementary to speech production. There have even been attempts to develop computer speech recognition based on working back from the speech signal to articulations instead of the more traditional approaches based on some type of acoustic pattern matching; however, these articulation-mediated attempts have not been as successful as more traditional approaches.

Although theories of speech perception that make reference to articulation have been attractive, there are problems with such approaches. There is a lawful relationship between acoustics and the structures that produced the sound. For a given state of mechanical affairs in any sound-producing structure including vocal tracts, there is a single predictable outcome. The physics of sound are constant, so the acoustic products of the same sound-producing event are always the same. Thus, variance in the acoustic speech signal must be the direct result of variance in production. If one were to rely on production for the defining features for speech perception, then the problem is at least as complex for articulation as it is for acoustics unless one adopts a level of description for production that is considerably more abstract than actual physical articulations.

In Liberman's last revision of the motor theory during the mid-1980s, this is evident from his claim that speech perception is the result of human specializations that serve to detect intended (not actual) gestures. Given the variance in physical realization, articulatory gestures have been difficult to measure precisely, and intentions for gestures are even more difficult to define or measure. Appeals to production do not provide as much simplification to models of speech perception as was initially hoped. Another line of evidence against uniquely human specializations for perception of speech can be found in demonstrations that nonhuman animals can learn to sort consonants (e.g., [d] versus [b] and [g]) across the same sort of acoustic variation depicted in Fig. 6. Although descriptions of speech contrasts in either articulatory or acoustic terms appear quite complex, this apparent complexity does not seem to provide insurmountable obstacles for perception by human or nonhuman biological systems.

Issues concerning context sensitivity and lack of invariance remain among the most central to

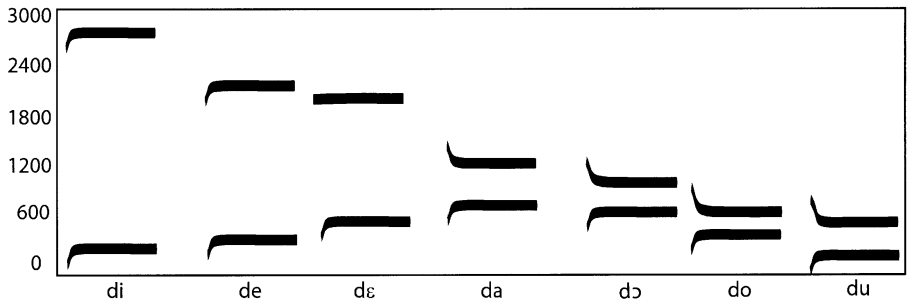


Figure 6 Stop consonant [d] depicted schematically.

understanding speech perception. It is well-known that in the absence of necessary and sufficient stimulus attributes, multiple stimulus attributes contribute to perception of speech sounds. Many experiments have revealed a phenomenon referred to as a "trading relation" between stimulus attributes. In these experiments, attributes of the signal that naturally covary with one another are altered independently. When one attribute is held constant, often at some intermediate value, the value of the second attribute is sufficient to tilt perception one way or the other. Also when this second attribute is held constant, variation in the first attribute can dictate perception. Across such experiments, it was shown how listeners use multiple stimulus attributes for speech perception.

Studies of trading relations in speech perception bear a striking likeness to issues in other domains of perception. One of the central classic problems for theories of perception, extending to the British Empiricists of the 17th and 18th centuries, is perceptual constancy (e.g., color, shape, and size). Common to instances of visual perceptual constancies is that objects or scenes are recognized as maintaining identity despite transformations in the optical array owing to changes in distance, lighting, perspective, etc. Each of the visual constancies represents an instance for which the sensory attributes can vary, but perception is relatively stable across such variation. No single aspect of the stimulus can be counted on as a necessary and sufficient condition for perception. For more than three centuries, a pervasive approach to understanding perceptual constancy has been to assume that experience is required to develop perceptual constancy, and that multiple aspects of the signal are used together via associationist processes. It is now known that one of the things that neural systems do best is to use multiple sources of modestly reliable information. Recent simulations of neural connectivity and activity in so-called neural network or connectionist models are designed to mimic the use of multiple attributes and associations between attributes as a function of experience with natural covariation. Among some of the most interesting findings for computer simulations of vision are demonstrations that performance can be very robust across degradations of the signal such as adding noise, spectral filtering, or deleting portions of the signal—a close analogy to examples of the robustness of speech perception described previously.

There have been many efforts to simulate speech perception within connectionist simulations; although there has been a fair degree of success, these simulations vary in the extent to which they include authentic auditory representations of speech or operate on stimuli that are representative of infant experience. Studies of development using human infant subjects address some of the same issues, including several reports that infant perception of speech and words can be driven by statistical regularities in the speech signal. In the meantime, there have been a small number of studies in which animals, instead of infants or computers, serve as surrogates to investigate how experience with natural covariation between acoustic attributes of speech serves to maintain perceptual constancy for speech sounds.

B. Becoming a Native Listener

Given the enormous variation in the particular subsets of the hundreds of sounds that are used by individual languages, the next major question concerning experience is how infant listeners come to understand speech in a fashion appropriate to their language community. In the multidimensional space of stimulus attributes, experience plays a critical role in parceling speech sounds into functional classes with linguistic significance—the consonants and vowels of one's language. The simple fact that different languages partition the domain of possible speech sounds differently implies that functional use of speech sounds depends on linguistic experience.

There have been ample demonstrations that during the first year of life, before infants produce much speech, infants become attuned to the differences between vowels and consonants that are functionally appropriate to their language environment based on their experience with speech sounds. By 6 months of age, infants are more likely to respond to acoustic differences that distinguish two vowels in their native language than to respond to equivalent differences that do not distinguish vowels in their language. This is true even though the differences to which infants do not respond are linguistically significant in other languages. Infants have been shown to respond to differences between speech sounds that would signal a linguistically significant change, such as from one vowel to another, while apparently ignoring greater acoustic variation relating to changes in talker.

At least for the relatively small number of contrasts that have been tested thus far, it appears that perception of analogous changes in consonants is developmentally more extended than for vowels. There have been several demonstrations that, at approximately 6 months of age, infants respond fairly

equivalently to differences between stop consonants independent of whether or not the differences are functionally significant within the native language. However, by the time infants are about 1 year old, they exhibit a tendency to respond more reliably to acoustic differences that distinguish consonants from within their native language. This pattern of performance rests on 1-year-old infants not responding to the same acoustic differences to which 6-month-old infants respond. What may be most surprising about the data from infant listeners is that the well-known adult difficulty in perceiving differences between nonnative speech appears to develop very early, before infant speech (babbling) includes most of the differences they will use in their language.

In addition to many studies concerning the ability of adult Japanese listeners to perceive or learn to perceive the difference between English [r] and [l], there have been many studies of perceiving and producing contrasts when learning a second language. Although the basic effects concerning initial inability to detect and produce differences among sounds in a second language are relatively consistent across other consonants and vowels, difficulty with [r] and [l] appears to be much more recalcitrant than contrasts more generally. A number of models have been proposed to explain processes underlying second language acquisition, some more explicitly developmental than others. Currently, such models can be broadly summarized in a fashion that acknowledges the importance of prior experience. They all attempt to account for patterns of performance showing that perception of speech sounds in a second language can be predicted on the basis of similarity or dissimilarity with sounds from the native language. When two sounds from a second language are both most similar to only a single sound in the native language, there will be difficulty perceiving that difference, with both being heard as renditions of the single sound from the native language. For example, when a native-Spanish listener confronts English vowels, for which English [i] and [ɪ] are both most similar to only [i] in Spanish, the Spanish listener will have difficulty with the distinction between [i] and [ɪ].

When the distinction between two sounds from a second language is most similar to a distinction between two sounds in the native language, listeners will tend to perceive differences between the two new sounds in a pattern much like that for the two sounds in the native language. Finally, in cases in which two new sounds in the second language bear little to no likeness to sounds used in the native language, listeners typically have no difficulty perceiving these differences in the absence of interference from sounds in the first language. In all cases, the effects of early experience with speech sounds in the native language predict the ease with which new speech contrasts can be adopted for the second language. There are a number of additional observations that have been made depending on duration of experience with the first or the second language, and of course, things become more complicate when more than two languages are considered.

C. Categorical Perception

Among experimental phenomena related to speech perception, perhaps none is more widely known than categorical perception. This classic pattern of perceptual data is studied in the laboratory using a series of stimuli that vary systematically in one or more acoustic attributes that distinguish two speech sounds, typically consonants. Listeners are asked to label each stimulus in the series as one or the other speech sound, and they are asked to discriminate between pairs of speech sounds drawn from series. Discrimination is usually measured using tasks that require subjects to listen to three sounds and report either which of three stimuli is different (AXB) or which of the first two stimuli sounds most like the third (ABX). Across these identification and discrimination procedures, three features define categorical perception: a sharp labeling (identification) function, discontinuous discrimination performance (near perfect across identification boundary and near chance to either side), and the ability to predict discrimination performance on the basis of labeling data. The classic patterns of identification and discrimination are displayed in Fig. 7.

For a while, such patterns of data stood in contrast to the way investigators thought about perception of continuously varying stimuli in other domains. For example, if one were to ask subjects to judge the pitches of tones that varied in frequency, it may be possible to have listeners divide the series in half, labeling some low and some high. However, listeners would not appear to lose their ability to detect the difference in pitch of two stimuli that both were labeled low or high. Many speech scientists initially took categorical perception to be indicative of perceptual processes that were unique to human perception of speech.

Throughout the years, however, a number of additional findings conspired to indicate that categorical perception was not unique to human perception

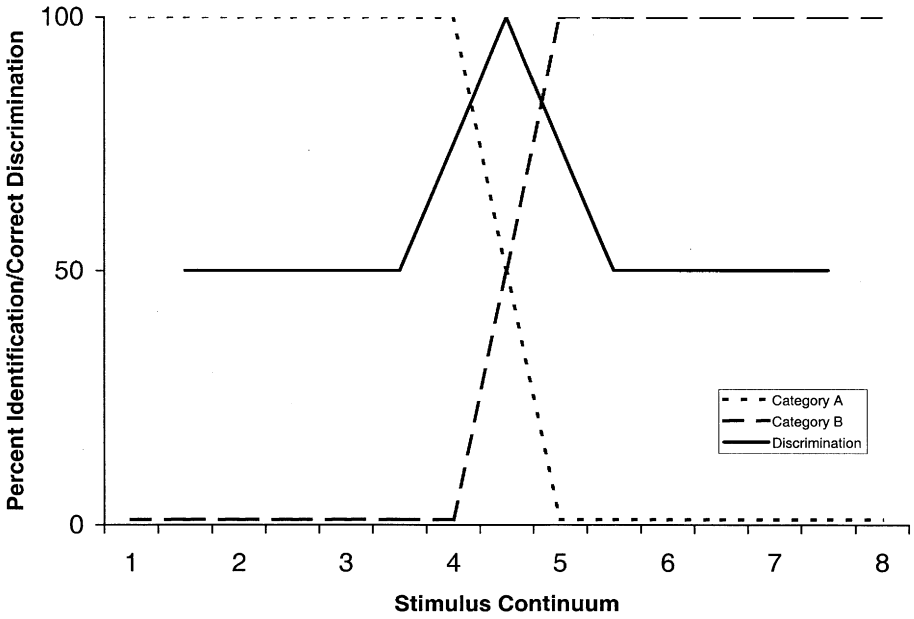


Figure 7 Categorical perception: idealized patterns of identification and discrimination.

of speech. First, it was demonstrated that nonhuman animals, responding to the same speech stimuli used in some of the experiments with human listeners, demonstrated the classic criteria defining categorical perception. In addition, it was demonstrated that with training, human listeners could accurately discriminate between stimuli that were labeled the same before training, contrary to the criterion that discrimination of speech sounds should be near chance if these sounds are labeled the same. Similar to learning a new language, which requires distinguishing between two new sounds that are perceived as examples of the same consonant or vowel in the native language, experience or training with the new distinction improves discrimination. As might be expected given the wealth of experience with natural distributions of speech sounds, the addition of memory variables also serves to predict performance on categorical perception tasks. When associative memory is simulated using neural network models, it appears that signature response patterns for categorical perception may be an emergent property of learning following exposure to distributions of speech sounds. Consistent with such a conclusion is the fact that categorical perception has been observed with a number of visual and nonspeech auditory stimuli with which observers and listeners have much experience.

Despite the fact that categorical perception does not distinguish speech perception from perception in other domains, it does illuminate principles of perception more generally. Psychophysicists often interrogate the operating potential of sensory systems by measuring thresholds at the limits of detection or discrimination, and such studies provide important data demonstrating the absolute potential of sensory systems, usually under ideal conditions for listening or observing. In such experiments, participants typically have little or no experience with the stimuli prior to coming to the laboratory. Most functional purposes of perception, however, do not require making minimal distinctions between unnatural stimuli. Instead, they require providing sufficient information to function adaptively, and this must be accomplished under conditions that are often not optimal. This usually requires using multiple stimulus attributes in many cases, this requires not using some differences between physical stimuli that do not have functional consequences. In addition, contrary to psychophysical studies that typically use unfamiliar stimuli such as pure tones for hearing or sine-wave gratings for vision, perception is very much driven by experience with a structured world. Speech perception is emblematic of both these points. Sound systems of languages are structured to avoid minimal differences, and many of the most important issues concerning speech perception highlight the powerful effects of experience for adapting perception to fulfill its functional role.

VI. CONTRIBUTIONS OF THE CENTRAL AUDITORY SYSTEM

In Section IV, some observations about peripheral neural encoding of speech and other complex sounds were described. Here, processing of speech at higher levels of the central nervous system is considered. Two of the most classic early observations of cortical functions related to speech were made by surgeon Pierre Paul Broca and neurologist Carl Wernicke in the 19th century. Broca conducted an autopsy on a patient who apparently could understand everything that was said to him but could not speak more than a few syllables. Autopsy revealed a large lesion to the third convolution of the frontal lobe on the left side of the brain. This area in left frontal cortex, shown in Fig. 8 anterior to motor cortex, became known as Broca's area. Damage to this area results in expressive aphasia —difficulty in producing speech despite maintaining the ability to understand spoken language. Approximately a decade after Broca's observations, Wernicke reported that damage to the posterior superior left temporal lobe (Fig. 8) resulted in a complementary disorder. Patients with damaged Wernicke's area suffered from receptive aphasia, being able to produce fluent but meaningless speech, but left relatively unable to comprehend language.

Owing to the spatial proximity of Wernicke's area to posterior auditory cortex, Wernicke suggested this is where speech sounds are converted into meaning. Because Broca's area is close to areas of motor cortex related to face and neck, Wernicke suggested that disconnecting Broca's area from Wernicke's area should result in intact language comprehension and fluent but meaningless speech. There is a fiber tract, the arcuate fasciculus (Fig. 8), that does make such a connection. In fact, Wernicke's predicted outcome, known as conduction aphasia, is obtained when the connection between Broca's and Wernicke's areas is severed.

There are several obstacles to making strong claims concerning speech perception on the basis of observations such as those made by Broca and Wernicke. The first is a quite general concern regarding the accuracy with which one can draw inferences between substantial brain areas containing a lesion and a specific process. The most common lesions from stroke follow patterns of vascularization, not function. Thus, a lesion can disable only part of some putative process or an entire region plus other regions that may or may not be related to the process of interest. Further weakening

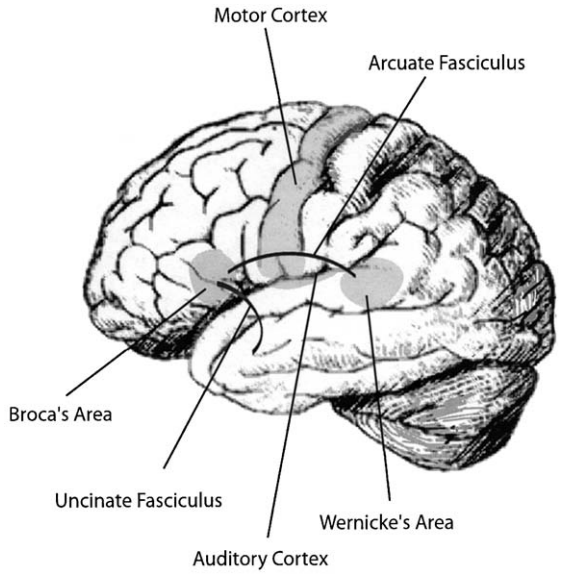


Figure 8 Broca's and Wernicke's areas: damage resulting in either receptive or expressive aphasias, respectively.

the ability to make strong claims from lesion data is the difficulty dissociating between localization of some process per se and damage to some pathway between two areas of primary interest. In later years, it became increasingly apparent that precisely defining either Broca's or Wernicke's areas is very difficult owing to the previously mentioned considerations and to considerable differences between individuals' cortical organization, and these terms are now often abandoned in lieu of more precise anatomical description.

A more specific difficulty for assessing cortical neural processes and locations for speech perception is revealed by Wernicke's hypothesis regarding conversion of speech sounds into language with meaning. Although speech without meaning (nonsense words or pseudowords) can be perceived, perceiving speech that has no meaning does not imply that neural processes related to higher linguistic processes and meaning are not engaged. Listeners can perceive the sounds of pseudowords such as "tig" and "gup." However, use of pseudowords may not eliminate involvement of higher language processes because such processes may be drawn to the task.

Another approach to distinguishing processing of speech and nonspeech sounds might be to use nonspeech sounds that maintain the most important structure and complexity of speech sounds without being speech—something more complex than noise or simple tones but not speech. This is difficult to do in

practice. First, owing to the great diversity of sounds used by languages (or even for a single language such as English), the range of acoustic attributes for speech is so large that constructing acceptably complex nonspeech stimuli without any such attributes is very difficult. Second, listeners can understand extremely degraded speech. How does one construct nonspeech stimuli that escape being heard by some listeners as speech based on even the most limited resemblance?

These last two experimental concerns extend beyond the neuropsychological study of patients who have suffered brain insult to modern methods for noninvasive study of cortical processing by unimpaired listeners. Modern methods of electrophysiology, magnetoencephalography, positron emission tomography (PET), and functional magnetic resonance imaging (fMRI) can be compromised by trade-offs between temporal and spatial resolution. Although measures of neural activity via electrophysiology have good temporal resolution, it is very difficult to be confident about the location of activity one is measuring. Magnetoencephalography is a related measure that offers better spatial resolution, but application has been fairly limited to date. Imaging methods that measure changes in cerebral blood flow or oxygen use (PET and fMRI) have much greater spatial resolution at the expense of temporal resolution. Consequently, the duration of a trial often exceeds 1 min, during which the acoustic stimulus of choice is presented dozens of times. This presents a clear limitation for understanding the relatively rapid processing of acoustic signals of any kind, and when one considers that the duration of a syllable can be less than 200 msec shortcomings of such a time base are apparent. Consequent to these difficulties, most conclusions regarding cortical processing of speech are tentative. However, some broad findings can be summarized.

As expected, presentations of sounds of any kind tend to result in activation of primary auditory cortex. In addition to auditory cortex, ventral, anterior, and posterior, there is a belt of cortex that is activated by more complex sounds. Ventrally, a second belt, sometimes referred to as a parabelt, can be described. For these belt areas, often referred to as ''secondary'' or ''associational'' areas, there is decreasing activation in response to very simple sounds such as sine waves and white noise, particularly when stimuli have limited temporal structure. This appears analogous to the visual system for which there is a hierarchy such that as processing proceeds from simple to more complex stimuli, areas that respond selectively to more complex stimuli tend to respond less to simpler stimuli. In addition to responding to more complex acoustic structure, there is evidence of cross-modal encoding (e.g., acoustic and optic), particularly in parabelt areas. In contrast to multiple observations of language usually being lateralized in the left hemisphere, activation in response to speech in core, belt, and parabelt areas of auditory cortex is usually found to be relatively balanced bilaterally in the absence of higher level linguistic effects.

This bilaterality, in contrast to laterality for language, does not imply that there is nothing exceptional about processing of speech in the human cortex. There are other ways that processing of speech can be distinguished from perception of other sounds. Cortical organization is critically linked to the amount and nature of experience, and there are no acoustic signals with which humans have more experience than with speech. An example from vision is perception of faces, which appears to have some special status in the middle fusiform gyrus of the cortex. This area is disproportionately activated when subjects view faces versus viewing other objects or scenes. Later work has demonstrated that similar patterns of activation can be found in response to other types of stimuli with which subjects have a great deal of experience. One may well expect that as more becomes known about cortical processing of complex sounds, cortical areas that appear dedicated to speech sounds will be revealed.

There is a second way in which cortical processes underlying speech processing could turn out to be atypical. If much of speech perception must be explained in terms of experience and, by extension, associationist processes, two types of cross-modal organization might be predicted. First, owing to a wealth of experience simultaneously hearing speech and viewing talkers' faces, one may expect substantial interaction between auditory and visual speech. The so-called McGurk effect, achieved by placing auditory and visual speech information in conflict, provides a powerful behavioral demonstration of this interaction, and there is brain imaging evidence for cortical areas related to these perceptual effects.

Second, there is reason to expect that some of the classic ideas concerning interactions with production processes may be true, albeit in a form different from what was considered in the latter part of the past century. Since the time of Wernicke's hypothesis concerning the conduction from Wernicke's area to Broca's area via the arcuate fasciculus, many who study neuropsychology have focused most attention on that conduit. There is another connection between

temporal and frontal areas of cortex, however. The uncinate fasciculus extends from anterior temporal cortex to inferior frontal gyrus within Broca's area. Thus, anatomical connections are in place if one were to hypothesize associations between areas involved in perception of speech with areas involved with production. Given that speech provides an unusual case for which, at least when one is talking, there are simultaneous activities of producing and perceiving one's own speech, the potential for such an association existing and being significant for speech perception is substantial but undocumented.

See Also the Following Articles

APHASIA • AUDITORY PERCEPTION • BILINGUALISM • BROCA'S AREA • INTELLIGENCE • LANGUAGE ACQUISITION • LANGUAGE AND LEXICAL PROCESSING • LANGUAGE DISORDERS • LANGUAGE, NEURAL BASIS OF • NUMBER PROCESSING AND ARITHMETIC • WERNICKE'S AREA

Suggested Reading

Atal, B., Miller, J., and Kent, R. (1991). *Papers in Speech Communication: Speech Processing*. Acoustical Society of America, Woodbury, NY.

Kent, R., Atal, B., and Miller, J. (1991). *Papers in Speech Communication: Speech Production*. Acoustical Society of America, Woodbury, NY.

Liberman, A. M. (1996). *Speech: A Special Code*. MIT Press, Cambridge, MA.

Miller, J., Kent, R., and Atal, B. (1991). *Papers in Speech Communication: Speech Perception*. Acoustical Society of America, Woodbury, NY.

Stevens, K. N. (1998). *Acoustic Phonetics*. MIT Press, Cambridge, MA.

Stress: Hormonal and Neural Aspects

BRUCE McEWEN
Rockefeller University

SONIA LUPIEN
McGill University

GLOSSARY

adrenalin A catecholamine; released by the adrenal medulla.

allostasis Achieving stability through change; active process that leads to adaptation of many body systems in the face of challenge.

allostatic load The cost to the body of responding to challenges and/or mismanaging the adaptive response to challenge.

amygdala A brain region involved in fear and emotional expression.

cortisol A glucocorticoid; a product of the adrenal cortex.

hippocampus A region of the brain involved in declarative, spatial, and contextual memory

hypothalamic–pituitary–adrenal axis The neuroendocrine system that produces cortisol in response to stress and the internal clock that determines diurnal rhythm.

prefrontal cortex Region of the cerebral cortex linked to integration of information between emotion and cognition.

socioeconomic status Social status of an individual based on income and educational attainment; often determined by type of job.

Stress is an almost universal topic in the modern world. There are many uses of the word "stress," including its usefulness in promoting conversation and empathy among friends and colleagues and even among strangers. In this article, stress is defined as a threat, real or implied, to the psychological or physiological integrity of an individual. As a result of the perception of threat, an individual makes behavioral and physiological responses that are intended to protect and defend the body and psyche from damage. Many factors contribute to individual differences in interpretation of and response to threat. Early life events, experiences throughout life, genetic factors, and the actions of hormones on the brain combine to produce major differences among individuals in how they react to potentially stressful situations; for example, the physical condition of the body determines how "stressful" shoveling snow may be. Because psychological and physiological systems are linked through the brain, the behavioral and physiological consequences of stress reflect two sides of the same coin.

Stress involves a stressor and a stress response. A stressor may be physical, such as trauma or injury, or physical exertion, particularly when the body is being forced to operate beyond its capacity. Physical stressors include environmental factors such as noise, overcrowding, and excessive heat or cold. Stressors also include primarily psychological experiences such as traumatic life events as well as daily hassles in the family and workplace, but many of these stressors have real physiological consequences, such as increased blood pressure and altered functions of the metabolic system, immune system and brain.

A broader concept that includes stress is that of "allostasis" and "allostatic load," which take into consideration aspects of lifestyle, genetic factors, and developmental influences that determine the degree to which life experiences exact a toll on the body over the

life course. Allostasis and allostatic load help to resolve the ambiguity inherent in the multiple uses of the word stress and they emphasize that there is an almost inevitable cost to the body of adapting to challenges of many kinds.

Stress is a term popularized by Hans Selye in 1936 that has taken on great significance in modern life. Many people refer to being "stressed out" by the pressures of commuting, working, studying, caring for children, and the generally increasingly faster pace of life. However, stress as a subjective experience does not always predict increased activity of the hormonal systems that are linked to stress, namely, the sympathetic nervous system and the hypothalamic–pituitary–adrenal (HPA) axis. In addition to the reporting of subjective stress, challenges to the individual that may be stressful lead to physiological and behavioral responses that can be measured.

The physiological stress response can be measured by monitoring levels of cortisol and ACTH in blood or cortisol in saliva or urine, whereas activity of the sympathetic nervous system is measurable as levels of catecholamines in blood, urine, or saliva or by monitoring heart rate and blood pressure. The stimuli that activate these systems involve novelty and unpredictability, a lack of sense of control, and threats to the sense of self. In addition, for the sympathetic component, arousal and physical activity are potent stimuli.

When stress hormones are released into the bloodstream, they reach tissues and organs throughout the body, on which they produce a variety of effects. These effects are mediated by receptors residing on the cell surface for the catecholamines and inside the cell for cortisol. Both types of receptors alter expression of genes in the cell nuclei of target cells, and the catecholamines also signal changes in activities of enzymes and ion channels and receptors in the cell membrane or cytoplasm.

The physiological stress response appears to be one aspect of a broader range of biological and behavioral processes that can protect, on the one hand, and cause damage and facilitate disease, on the other hand. For example, the "stress mediators," such as cortisol and catecholamines, also vary in their basal secretion according to a diurnal rhythm that is coordinated by the light–dark cycle and sleep–waking patterns. For example, chronic elevation of diurnal levels in cortisol is associated with pathophysiological states such as accelerated bone mineral loss and hyperglycemia. Therefore, over a long period of time, the measurement of hormone levels and the processes that they control, such as cholesterol, blood pressure, and waist–hip ratio, constitutes the primary means of connecting experience with resilience or the risk for disease. It also represents a final common path for assessing the impact of both the perturbations in diurnal rhythm and the stress response.

The process of adaptation has been termed allostasis, and the cost the body pays for this adaptation is referred to as allostatic load. Many systems of the body are subject to allostatic load, and the brain is a particularly important target because of its coordinating role with respect to the physiological stress response and also the adaptive and maladaptive behaviors that are associated with the response to stressful situations. This article discusses the hormonal and neural aspects of the response to stressful events in terms of allostasis and allostatic load.

I. PROTECTIVE AND DAMAGING EFFECTS OF STRESS MEDIATORS

A. Allostasis and Allostatic Load

Understanding the nature of the final common biological path for resilience or disease requires a discussion of a number of contributing factors, such as the protective and damaging role of the physiological mediators of stress and the diurnal rhythm, as well as the role of behavior, the importance of genetic factors, the role of gender, and the biological features of resilience.

An attempt at formulating the relationship among environmental challenges, physiological responses, and outcomes such as resilience or disease uses two terms: allostasis and allostatic load. Allostasis, literally meaning "maintaining stability (or homeostasis) through change," describes the process of adaptation to challenge. It was introduced by Sterling and Eyer in 1988 to describe how the cardiovascular system adjusts to resting and active states of the body. This notion was generalized to other physiological mediators, such as the secretion of cortisol and catecholamines, and the concept of allostatic load was proposed to refer to the wear and tear that the body experiences due to the repeated use of allostatic responses as well as the inefficient turning on or shutting off of these responses. One example of allostatic load is the persistent activation of blood pressure in dominant male cynomologus monkeys vying for position in an unstable dominance hierarchy, which has been shown to accelerate atherosclerotic plaque formation. This can

be prevented by β-adrenergic blocking drugs, which suggests that repeated activation of β receptors by blood pressure surges accelerates the atherosclerotic process.

The concepts of allostasis and allostatic load apply to all systems of the body and focus attention on the protective (adaptive) and damaging roles of primary stress mediators, such as catecholamines and cortisol. There are four important aspects of this formulation that are summarized in Fig. 1. First, the brain is the integrative center for coordinating the behavioral and neuroendocrine responses (hormonal and autonomic) to challenges, some of which qualify as "stressful" and others of which are related to the diurnal rhythm and its ability to coordinate waking and sleeping functions with the environment. Second, there are considerable individual differences in coping with challenges that are based on interacting genetic, developmental, and experiential influences. Third, inherent within the neuroendocrine and behavioral responses to challenge is the capacity to adapt (allostasis); indeed, the neuroendocrine responses are set up to be protective in the short term. For the neuroendocrine system, turning on and turning off responses efficiently is vital (Fig. 2); inefficiency in allostasis leads to cumulative effects over long time intervals. Fourth, allostasis has a price (allostatic load, referring to cumulative negative effects) that is related to how inefficient the response is or how many challenges an individual experiences (i.e., many stressful events). Thus, allostatic load is more than chronic stress and can also reflect a genetically or developmentally determined failure to efficiently handle the normal challenges of daily life related to the sleep–wake cycle and other daily experiences.

For behavioral responses to challenge, there are also protective and damaging aspects. Individuals can act to increase or decrease further risk for harm or disease: for example, antisocial responses such as hostility and aggression vs cooperation and conciliation; risk taking behaviors such as smoking, drinking, and physical risk taking vs self-protection; and a high-calorie and/or fat-rich diet and poor health practices vs a moderate calorie, lower fat diet and regular exercise. The linkage of allostasis and allostatic load probably applies to behavioral responses as well as to physiological responses to challenge insofar as the behavioral response, such as smoking or alcohol consumption, may have at least perceived adaptive effects in the short term but damaging effects in the long term. Behavior also plays a role in modifying the physiological response (e.g., eating rich food synergizes with elevated physiological mediators to increase the metabolic syndrome, whereas even a brief period of exercise can reduce insulin resistance of muscle and enhance glucose uptake).

Although allostasis and allostatic load are general concepts, it is essential to describe and understand the mechanism of the protective and damaging effects that take place in each system of the body. The mediators [adrenal steroids, mainly catecholamines, but also

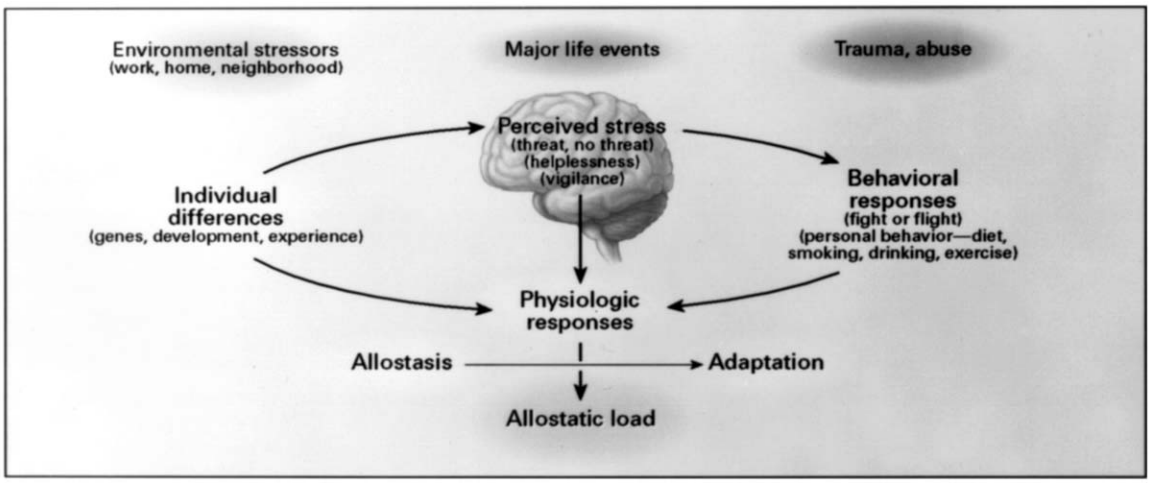


Figure 1 The stress response and development of allostatic load. Perception of stress is influenced by one's experiences, genetics, and behavior. When the brain perceives an experience as stressful, physiologic and behavioral responses are initiated leading to allostasis and adaptation. Over time, allostatic load can accumulate, and the overexposure to neural, endocrine, and immune stress mediators can have adverse effects on various organ systems, leading to disease (reproduced with permission from McEwen (1998). Copyright © 1998 Massachusetts Medical Society. All rights reserved).

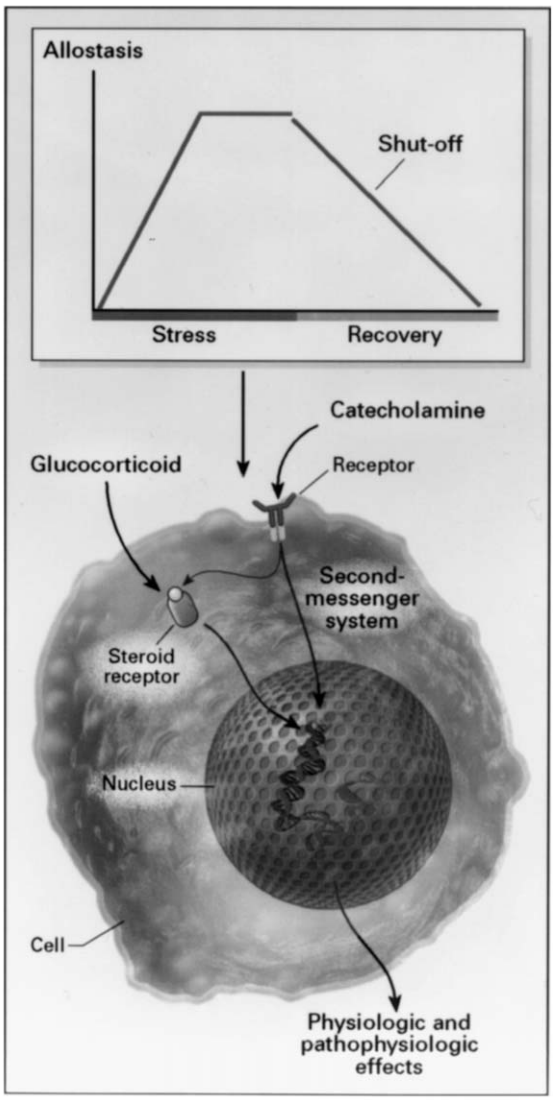


Figure 2 Allostasis in the autonomic nervous system and the hypothalamic–pituitary–adrenal axis. Allostatic systems respond to stress (top) by initiating the adaptive response, sustaining it until the stress ceases, and then shutting it off (recovery). Allostatic responses are initiated (bottom) by an increase in circulating catecholamines from the autonomic nervous system and glucocorticoids from the adrenal cortex. This sets into motion adaptive processes that alter the structure and function of a variety of cells and tissues. These processes are initiated via intracellular receptors for steroid hormones, plasma membrane receptors, and second messenger systems for catecholamines. Cross talk between catecholamines and glucocorticoid receptor signaling systems can occur (see text) (reproduced with permission from McEwen (1998). Copyright © 1998 Massachusetts Medical Society. All rights reserved).

other hormones such as dehydroepiandrosterone (DHEA), prolactin, and growth hormones and the cytokines related to the immune system] affect many systems, but their effects are organ or system specific in many cases. Once the mediators are secreted, they produce effects on systems and organs of the body by acting on receptors. The effects can be classified as primary effects, secondary outcomes that are risk factors for disease, and tertiary outcomes that are the diseases.

In Fig. 2, the actions of two mediators, the glucocorticoids and the catecholamines, are shown. The mediators act via receptors that trigger changes throughout the target cell in processes, including changes in gene expression that have long-lasting consequences for cell function. Thus, every time a hormone is secreted, it is important to consider the short-term and long-term consequences of hormone action on cell function.

The following are adaptive (allostasis) and damaging (allostatic load) aspects of four systems, and the roles of some of the mediators are presented. Later, we discuss these systems in-depth:

Cardiovascular. *Adaptive*: The autonomic nervous system, involving catecholamines, promotes adaptation by adjusting heart rate and blood pressure to sleeping, waking, and physical exertion. *Damaging*: Repeated surges of blood pressure or failure to shut off blood pressure surges efficiently accelerate atherosclerosis and synergize with metabolic hormones to produce type II diabetes.

Metabolic. *Adaptive*: Adrenal steroids promote food intake and promote replenishment of energy reserves. *Damaging*: Elevated evening cortisol and hyperreactivity of the HPA axis promote insulin resistance and accelerate progression toward type II diabetes, including abdominal obesity, atherosclerosis, and hypertension.

Brain. *Adaptive*: Adrenal steroids, together with catecholamines, promote retention of memories of emotionally charged events, both positive and negative. *Damaging*: Adrenal steroids, acting with excitatory amino acid neurotransmitters, promote cognitive dysfunction by a variety of mechanisms that involve atrophy and, in extreme cases, death of brain cells, particularly in the hippocampal region of the brain.

Immune. *Adpative*: Adrenal steroids, together with catecholamines, promote ''trafficking'' (i.e., movement) of immune cells to organs and tissues in which they are needed to fight an infection or other challenge, thereby enhancing immune responses in those sites. Adrenal steroids also modulate the expression of the hormones of the immune systems, the cytokines and chemokines. *Damaging*: Adrenal steroids and

catecholamines both have immunosuppressive effects when these mediators are secreted chronically or not shut off properly. On the other hand, the absence of sufficient levels of these mediators allows other immune mediators to overreact and increases the risk of autoimmune and inflammatory disorders. An inadequate response of the HPA axis and autonomic nervous system is thus another type of allostatic load, in which the dysregulation of other mediators, normally contained by cortisol and catecholamines, is a primary factor in a disorder.

The long-term cost to the body (allostatic load) may be categorized into at least four subtypes, as summarized in Fig. 3. The first type is simply too much stress in the form of repeated, novel events that cause repeated elevations of stress mediators. For example, the amount and frequency of economic hardship predict decline of physical and mental functioning as well as

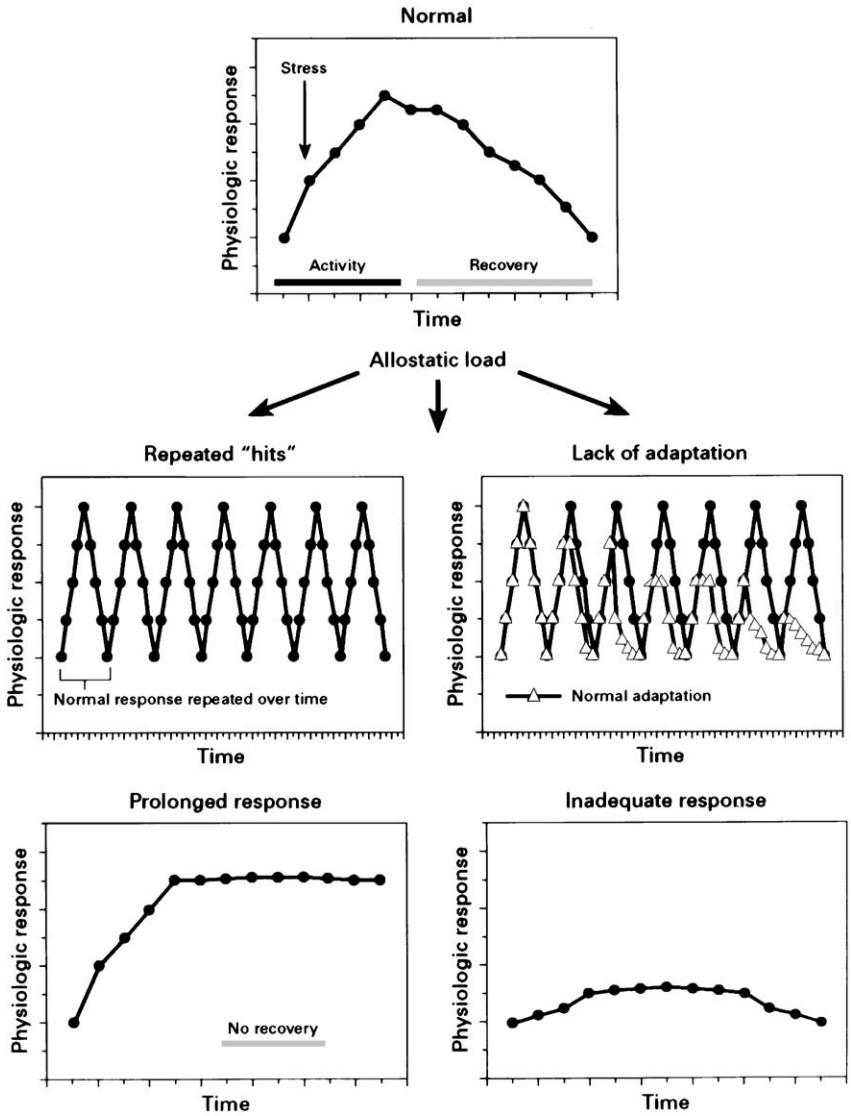


Figure 3 Four types of allostatic load. (Top) The normal allostatic response, in which a response is initiated by a stressor, sustained for an appropriate interval, and then turned off. (Bottom) Four conditions that lead to allostatic load: (i) repeated "hits" from multiple novel stressors, (ii) lack of adaptation, (iii) prolonged response due to delayed shut down, and (iv) inadequate response that leads to compensatory hyperactivity of other mediators (e.g., inadequate secretion of glucocorticoid, resulting in increased levels of cytokines that are normally counterregulated by glucocorticoids) (drawn by Dr. Firdaus Dhabhar, Rockefeller University; reproduced with permission from McEwen (1998).

mortality. A second type of allostatic load involves a failure to habituate or adapt to multiple occurrences of the same stressor. This leads to the overexposure to stress mediators because the body fails to dampen or eliminate the hormonal stress response to a repeated event. For example, when confronted with repeated public speaking challenges, most individuals habituated their cortisol response but a few individuals continued to show cortisol response. A third type of allostatic load involves the failure to shut off either the hormonal stress response or the normal trough of the diurnal cortisol pattern. Examples of this include blood pressure elevations in work-related stress, sleep deprivation leading to elevated evening cortisol and hyperglycemia within 5 days, and depressive illness leading to chronically elevated cortisol and loss of bone mineral mass.

The fourth type of allostatic load involves an inadequate hormonal stress response that allows other systems, such as the inflammatory cytokines, to become overactive. The Lewis rat is an example of an animal strain in which increased susceptibility to inflammatory and autoimmune disturbances is related to inadequate levels of cortisol.

Allostatic load can be measured. In the initial validation study, data were used from the MacArthur Successful Aging Study that pertain to levels of physiologic activity across a range of important regulatory systems, including the HPA and sympathetic nervous systems as well as the cardiovascular system and metabolic processes. The measure of allostatic load reflects only one of the two aspects of physiologic activity postulated to contribute to allostatic load—namely, higher, chronic, steady-state levels of activity related to the diurnal variation and any residual activity reflecting chronic stress or failure to shut off responses to acute stressors. The parameters that have been measured for determining allostatic load scores include urinary cortisol, adrenalin, noradrenalin; serum DHEA, waist–hip ratio, diastolic and systolic blood pressure, glycosylated hemoglobin, serum HDL, and total cholesterol.

One of the problems with the original conceptualization of allostatic load and its measurement is that the components are not organized and categorized with regard to what each measure represents in the cascade of events that lead from allostasis to allostatic load. Nor is there any suggested organization in choosing those original measures that would facilitate systematically relating measures to specific disease outcomes or systematically adding new measures. Allostasis and allostatic load are concepts that are mechanistically based and only as good as the information about mechanisms that lead to disease.

A new way of classifying the measures must provide a way to relate what is measured to a pathophysiological process and allow for the incorporation of new measures as more is known about underlying mechanisms leading to disease. This new formulation is based on the notion of primary mediators leading to primary effects and then to secondary outcomes, which finally lead to tertiary outcomes that represent actual diseases. In the original allostatic load measures, urinary cortisol, adrenalin, noradrenalin, and serum DHEA are primary mediators, whereas the other six (waist–hip ratio, diastolic and systolic blood pressure, glycosylated hemoglobin, serum HDL, and total cholesterol) represent secondary outcomes.

B. Modulators of Protection and Damage

There are a number of modulators of the processes of protection and damage, including genetic factors and developmental processes. These are briefly outlined before we discuss the impact of gender and socioenvironmental influences on the ability of the body to show allostasis and resist allostatic load in terms of resilience.

1. Role of Genetics

Genes are a major factor in risk for many disorders, but the environment and experiences of the individual play a major role in regulating the genetic traits throughout the life span. Many of the genetic traits are complex and reflect the operation of many genes. There has been enormous progress in the identification of genes and the mapping of the human genome. However, identifying a gene is only the first step, and understanding its regulation is critical. Fortunately, there is rapid progress in understanding how genes are regulated by hormones and other environmental signals such as neurotransmitters.

2. Developmental Influences

The susceptibility of an individual to stress-related disorders is likely to reflect developmental influences as well as genetic risk factors. Early life experiences play a major role in determining the long-term cost to the body, or allostatic load, over a lifetime. Animal models provide a useful way of understanding some environmental and developmental influences on individual differences that are relevant to human stress reactivity.

3. Gender Differences

An important aspect of development deals with sex differences in brain function and behavior. Sex differences are developmentally regulated and reflect not only the presence or absence of gonadal hormones during sensitive periods in pre- and postnatal life but also the interactive effects of a lifetime of experiences related to being male or female. Hormones, particularly testosterone, act during early pre- and postnatal development to shape the structure and function of brain regions, including both the hypothalamus and extrahypothalamic brain regions such as the hippocampus. Testosterone secretion by the testes during gestation and early neonatal life masculinizes and defeminizes the brain, and these processes affect the connectivity between nerve cells, the number of nerve cells, and the functional capacity to respond to hormones and other chemical messengers. However, hormones and experience can change brain structure and function in adult life; thus, there is a considerable range of plasticity throughout the life span.

Behavioral sex differences develop as a result of the interaction of a sexually differentiated neural substrate with experiences, and these sex differences pertain to reproduction-related events, such as establishment and defense of territory and mating and sexual behavior, as well as to nonreproductive behaviors such as strategies used to solve spatial learning tasks. Sex differences are also recognized in the prevalence of psychiatric disorders, with females showing higher rates of anxiety and depression and males having higher incidences of substance abuse and antisocial behavior. Moreover, in animal models and possibly also in humans, there are sex differences in the vulnerability to brain damage or brain remodeling. Although the role of hormones is a hallmark of sexual differentation, the role of experience and social factors cannot be underestimated, particularly in infrahuman primates and humans. As noted previously, the impact of experience to change brain structure and function is one of the emerging areas of neurobiological research. However, in humans as in lower animals, there are certain basic patterns or tendencies that characterize each gender.

In animals, and possibly also to some extent in humans, the role of basic sex differences in attitudes toward ''nurturing'' may provide a fundamental difference in response to a variety of life events and therefore in the types of long-term cost, or allostatic load, that are manifested. Within the concept of ''inclusive fitness,'' the male is more prone to show a ''fight or flight'' response to stressful challenges, whereas female responses to such challenges are more likely to follow a pattern referred to as ''tend and befriend.'' ''Tend'' refers to nurturant activities designed to protect self and offspring, promote safety, and reduce distress, whereas ''befriend'' means the formation and maintenance of largely female networks that aid in protecting and nurturing offspring and self. Possible underlying mechanisms for these differences include oxytocin, female and male reproductive hormones, and endogenous opioid mechanisms as well as neurotransmitters such as serotonin, low levels of which promote hostility, aggression, suicide, and depression.

Regarding allostatic load, the tend and befriend behavior promotes social support, which reduces the hormone levels associated with allostatic load. In contrast, the fight or flight response pattern can either put the individual at greater risk for harm and futher allostatic load (fight or confrontation) or remove the individual from danger, which may be beneficial for health (flight or withdrawal). On the other hand, defeat, which can result from confrontation and fighting, can have devastating effects on health and survival among males even in the absence of physical injury. Subordination and conflict in social groups of animals reveal more subtle sex differences. For example, unstable social hierarchies among male cynomologous monkeys increase the rates of atherosclerosis in coronary arteries, although subordinate status of males in a stable hierarchy does not increase atherosclerosis; however, subordinate female cynomologous monkeys are known to have increased risk for atherosclerosis and depression in both stable and unstable social hierarchies.

4. Socioenvironmental Conditions

We previously discussed what happens to the individual and how the individual's behavioral and physiological responses and risk for disease are determined by genetic, developmental, and experiential factors. Individuals make up groups that differ, for example, in education and income [referred to as socioeconomic status (SES)]. SES appears to be an important factor in disease, and new evidence points to gradients of health across SES in the British Civil Service as well as in other Western countries. Hardship also compromises health, including cognitive and physical function and mental health. The collapse of communism resulted in worse health in Eastern Europe, particularly in Russia, between 1989 and 1993, and behavioral problems (alcoholism, suicide, homicide, and cardiovascular disease) accounted for much of the increased mortality.

5. Resilience

From a physiological standpoint, we know very little about positive influences on health (e.g., resiliency in recovery from illness, injury, or surgery and the overcoming of adversity). Undoubtedly, resilience is more than just the absence of allostatic load but, rather, the product of cellular processes that protect and build cells and tissues and provide some reserve capacity and resistance to the damaging effects of stress mediators. In particular, the role of anabolic hormones, such as growth hormone and insulin, and the neurotrophic factors in the brain needs to be explored in relation to such phenomena as voluntary exercise and recovery from illness, injury, or surgery. One example of this is the ability of voluntary exercise in rats (running on an activity wheel) to increase the expression of the messenger RNA for a neurotrophin, which protects nerve cells from dying and promotes aspects of cellular plasticity and synaptic transmission. It is not known what advantages this increased level of neurotrophins may confer on the brains of the exercising animals, for example, whether they are more resilient in the face of severe stress or whether their brains will age more slowly. One possible consequence of this neurotrophin regulation is that voluntary exercise has been reported to increase production of new nerve cells in the dentate gyrus of the hippocampal formation, a brain region that is important in spatial and declarative memory.

More research is needed to relate the production of neurotrophins and other mediators related to growth and repair of tissues to human life histories and to identify the role of social support mechanisms and individual attitudes in promoting beneficial physiological states associated with the ability to repair damaged tissues and to protect against pathogens and toxic agents such as free radicals. Though poorly defined, this area of inquiry is very important as a complement to the more traditional approach of studying the damaging aspects of stress mediators.

II. ALLOSTATIC LOAD IN THE BRAIN

A. Protective and Damaging Effects of Stress Mediators in Brain

Stress mediators have both positive and negative effects on the brain, just as they do on other systems of the body. The stress mediators enhance formation of the so-called ''flashbulb memories'' of events associated with strong emotions, including fear but also positive emotions. These involve the amygdala, and the pathway for encoding these memories involves the interaction between neurotransmitters in the amygdala and in related brain areas such as the hippocampus along with circulating stress hormones of the adrenal cortex and adrenal medulla. Indeed, encoding of these memories is strengthened by glucocorticoids acting in the amygdala and hippocampus, among other brain regions, and epinephrine acting in the sensory vagus outside of the blood–brain barrier, with information transmitted into the brain via the nucleus of the solitary tract. These findings may have relevance to posttraumatic stress disorder and also to symptoms of depression, in which an overactive amygdala appears to be involved.

At the same time as the brain encodes information and controls the behavioral responses, it is also changed structurally and chemically by those experiences. Studies of learning and memory have revealed levels of plasticity involving structural changes in brain cells and changes in gene expression. On the one hand, this can be seen by the remodeling of neuron structure brought about by training. On the other hand, transcription factors involved in regulating expression of groups of genes in brain cells appear to be essential for the formation of long-term memories in species ranging from fruit flies to mice.

Although short-term response of the brain to novel and potentially threatening situations may be adaptive and result in new learning and acquired behavioral strategies for coping, as may be the case for certain types of fear-related memories, repeated stress can cause cognitive impairment via at least four different mechanisms:

1. Impairing neuronal excitability: Adrenal steroids biphasically modulate long-term potentiation (LTP), with low levels enhancing it and high levels impairing LTP in regions of the hippocampus that use NMDA receptors; other measures of excitability are also affected by adrenal steroids.
2. Causing atrophy of nerve cells in the Ammon's horn region of the hippocampus: Adrenal steroids facilitate a remodeling of apical dendrites of pyramidal neurons in the CA3 region of the hippocampus that is caused by excitatory amino acids; such remodeling is reversible as long as stress is terminated after a number of weeks.
3. Inhibiting neurogenesis in the dentate gyrus region of the hippocampus: The adult hippocampus continues to produce nerve cells in adult life, and this

process is inhibited by certain stressors and by activation of NMDA receptors as well as by elevated glucocorticoids.

4. Causing permanent loss of nerve cells in hippocampus: Prolonged psychosocial stress causes damage and apparent neuron loss in the hippocampus.

These processes may occur somewhat independently of each other and contribute in various degrees to different pathophysiological situations involving traumatic stress, depression, or aging.

B. Effects of Stress and Stress Hormones on Cognitive Function

Having reviewed the potential mechanisms by which stress and stress hormones can biphasically modulate learning and memory processes, we now consider the information available on stress and glucocorticoid effects on cognitive function in human subjects.

The cognitive effects of elevated concentrations of glucocorticoids in human populations have been studied in disorders affecting corticosteroid levels and by using exogenous administration of the synthetic compound to healthy subjects. Mental disturbances mimicking mild dementia (such as decrements in simple and complex attentional tasks, verbal and visual memory, encoding, storage, and retrieval) have been described in depressed patients with hypercortisolism and in those with steroid psychosis following corticosteroid treatment. Similar cognitive deficits are also reported in patients suffering from Cushing's disease. During human aging, a significant proportion of elderly individuals present an endogenous increase of glucocorticoid levels, and this increase has been related to impaired memory performance. Moreover, many investigators have reported inverse relationships between mean 24-hr cortisol levels and severity of cognitive decline in Alzheimer patients.

Studies in both animals and humans have shown that the glucocorticoid-induced memory impairment is related to an atrophy of the hippocampus. Hippocampal atrophy associated with chronic exposure to high levels of glucocorticoids is reported in Cushing patients, elderly individuals, depressed patients, and individuals suffering from posttraumatic stress disorders. This is a significant finding and implicates the hippocampus since the declarative memory impairments that are induced by chronic exposure to high levels of glucocorticoids are those attributed to the hippocampus in memory function.

It is known that the hippocampus plays a significant role in declarative memory function, whereas it has little function in nondeclarative memory function. Declarative memory refers to the conscious and voluntary recollection of information that was previously learned, whereas nondeclarative memory function refers to the facilitation in performance observed after exposure to a given information, without necessary consciousness of recall of this information. Many studies have shown that hippocampal damage in animals and humans leads to declarative memory impairments, whereas nondeclarative memory is unimpaired. This is the pattern of memory dysfunction reported to occur in all cases of chronic exposure to high levels of glucocorticoids.

However, studies of endogenous disorders generally fail to discriminate the cognitive deficits related to HPA hyperactivity from those due to the underlying illness. Thus, most of the cognitive deficits associated with corticosteroids are derived from those observed during acute exogenous administration of synthetic glucocorticoids to healthy subjects. In general, studies measuring the acute impact of glucocorticoids on cognitive function report that this steroid impairs selective attention (i.e., the ability to discriminate relevant from irrelevant information), which thus impairs encoding of incoming information. This finding is in accordance with electrophysiological results showing that acute administration of cortisol to human subjects reduces the average evoked potential response to relevant but not to irrelevant stimuli. These findings are also consistent with studies showing that glucocorticoids can impair neuronal electrophysiology and hippocampal long-term potentiation.

Recent studies have reported that an acute increase of glucocorticoids also impairs working memory function. Working memory is the cognitive mechanism that allows us to keep a limited amount of information active for a limited period of time. Working memory impairments have been found in several experiments using a variety of delay task procedures. In these tasks, a temporal gap is introduced between a stimulus and a response, which creates the need to maintain the stimulus in a temporary memory storage. Interestingly, data obtained in monkeys show that cells in the lateral prefrontal cortex become particularly active during delayed response tasks, suggesting that these cells are actively involved in holding on to the information during the delay. This result is in accordance with studies reporting a high density of corticosteroid receptors in the cerebral cortex of both rat and human. Receptor binding studies in rats have

shown the presence of adrenal steroids in the cortex, particularly in the medial prefrontal regions. Further studies in rats and humans have shown that the prefrontal cortex is a significant target for the negative-feedback actions of circulating glucocorticoids, which suggests that this area could play a significant role in the acute effects of corticosteroids on cognitive function.

Thus, the hippocampus is not likely to be the only brain area affected in this way since atrophy of the amygdala and prefrontal cortex has also been reported in depressive illness. Reversibility and/or preventability of such atrophy is a major topic for future research, as is the implication of such treatment for cognitive function. A recent study showed that treatment of Cushing's patients induces a 10% reversibility in the hippocampal atrophy that was induced by chronic exposure to high levels of glucocorticoids.

III. DEVELOPMENTAL INFLUENCES ON ALLOSTATIC LOAD

The susceptibility of an individual to stress-related disorders is likely to reflect developmental influences as well as genetic risk factors. A major advance in behavioral neuroscience during the past decade has been the establishment of a life course perspective for understanding environmental influences on stress reactivity during development that persists for a lifetime. Prenatal stress increases emotionality and stress hormone reactivity for the life of the individual; conversely, postnatal "handling" of neonates leads to reduced emotionality and reduced stress hormone reactivity for a lifetime. In animal models, brain aging is increased by elevated stress reactivity and reduced by lowered stress reactivity. In addition, studies in infrahuman primates have shown that early maternal neglect increases anxiety and emotional lability in the offspring.

These animal studies are also directly relevant to and should encourage further studies of the adverse consequences for mental and physical health of abuse and neglect during childhood. Although data on human subjects are limited, there is a report that abuse in childhood is associated with substantial increases in substance abuse, depression, and suicide and also increased incidence of heart disease, cancer, chronic lung disease, skeletal fractures, and liver disease. There are also indications that individuals with posttraumatic stress disorder who were abused as children have a significantly smaller hippocampal volume. In addition to its relevance to structural plasticity within the brain, this finding raises questions about when in an individual's life these changes actually occur—whether they occur early in life or are initiated by the traumatic experience later in life.

Besides exposure to traumatic events during childhood, another pathway by which individual differences in physical health and/or cognitive function could be slowly established is through differential exposure to stress. In children, it has been hypothesized that stress experienced in the early years may have long-term effects on future development of psychosomatic diseases. However, immediate effects of stress on children's health have also been reported. In classic study, Cohen *et al*. compared children from schools that were beneath the flight path of the Los Angeles International Airport with children from schools in quieter areas. These investigators showed that children from the noisy schools presented higher blood pressure than children from the quieter schools. Moreover, they reported that children from the noisy areas had difficulty learning how to discriminate between irrelevant noise and the relevant task, which is in agreement with results of studies showing that stress affects selective attention. In a subsequent study, Cohen *et al*. evaluated the effects of living in apartments near congested highways in New York City. They showed that despite the relatively higher noise levels on lower floors of the buildings, residents of lower floors did not differ from neighbors on upper floors in their subjective ratings of noise level or degree of annoyance. However, children on the lower floors adapted to the noise by "tuning out" sounds, which resulted in a lower responsiveness to differences in spoken words. This further affected their reading ability, as demonstrated by long-term reading impairments. These results suggest that stress at an early age may affect health (e.g., blood pressure) as well as cognitive processing.

As noted previously, SES is another pathway by which significant individual differences in stress reactivity might evolve. It has been shown that individuals from lower SES report greater exposure to stressful life events and a greater impact of these events on their lives than do individuals from higher SES. Given the known impact of glucocorticoids on immune function, this suggests that individuals from lower SES may have greater vulnerability to stress and, subsequently, to disease. The association between SES and stress may also stem from environmental and social/psychological factors. With regard to environmental factors, it is known that as one moves down the

SES ladder, residential choices become more limited and many of the environments in which individuals lower on the SES hierarchy live are associated with increased mortality rate and crime. It has been suggested that these differences in environments vary objectively in chronic exposures to stressor events. With regard to sociological/psychological factors, it has been shown that higher SES decreases the likelihood of exposure to negative events such as social aggression and risk behaviors. Members of lower SES are exposed to a higher rate of change and/or instability in their lives and this instability has been found to produce a higher level of individual distress in lower SES individuals.

The relationship between SES and health begins at the earliest stages of life. In children, an SES–health linkage has been found with the following health problems: lead poisoning, vision problems, otitis media and hearing loss, cytomegalic inclusion disease, and iron deficiency anemia. In addition, mental retardation, learning disorders, and emotional and behavioral problems occur at greater frequency among children of lower SES. In order to assess whether the development of the relationship between SES and mental health is sustained, at least in part, by exposure to high levels of glucocorticoids, Lupien and collaborators measured morning salivary cortisol levels as well as cognitive function (memory, attention, language, and emotional processing) in 396 children (6–16 years of age) from low versus high SES in the Montreal area. Depressive symptomatology and exposure to stressful events were also assessed in each child's mother through a semistructured phone interview. The results revealed that low SES children from 6 to 10 years of age presented significantly higher salivary cortisol levels when compared to children from high SES. This difference disappeared at the time of school transition (i.e., at 12 years of age), and no SES differences were observed at 14 and 16 years of age. Neuropsychological data showed that although children from low and high SES do not differ with regard to memory, attentional, and linguistic functions, children from low SES (at all ages) presented a significantly higher level of negative emotional processing. Scores on the depression and stress questionnaires were significantly higher for low SES mothers compared to high SES mothers. More important, the authors reported a significant positive correlation between the mother's depressive symptomatology and her child's cortisol level. This showed that the higher the mother's score on the depressive scale, the higher the cortisol level of her child. Altogether, these results showed that low SES in young children is related to increased cortisol secretion and negative emotional processing. Moreover, they showed that depression in the mother, in conjunction with low SES, is a significant factor predicting high cortisol levels in children.

Further study of these topics is important because data indicate potentially beneficial effects for antisocial behavior and success in school of early life interventions by nurses visiting the home and providing support and education to the parents.

IV. CONCLUSIONS

The concepts of allostasis and allostatic load are inclusive of what we mean by stress, but they are much broader because they include aspects of lifestyle as well as genetic influences and developmental effects, including early life experiences and gender differences. In this way, allostasis and allostatic load provide a general conceptual framework in which to evaluate the overall impact of the physical and social environment on individuals and groups of individuals. It should be emphasized that although most of the work done so far has focused on the role of HPA activity and autonomic nervous system reactivity in these nature–nurture interactions, the allostasis/allostatic load model can be generalized to other physiological systems that respond to environmental stimuli. In other words, allostasis and allostatic load attempt to embody a general biological principle that the systems that help protect the body and promote adaptation in the short term can also participate in pathophysiological processes when they are overused or inefficiently managed.

The most important feature of allostatic load is that it operates gradually over long periods of time in the life of an individual. In fact, as summarized in this article, influences of genetic factors and early experiences, when coupled with the subsequent life experiences of each individual, exert a life-long effect on the physiology of an individual and alter the risk for developing a variety of pathophysiological conditions and diseases later in life as well as the rate of certain aspects of the aging process.

This multidisciplinary view of stress as a form of allostatic load should eventually allow us to identify very early in the life of an individual the potential risk factors for the negative effects of stress on physical and mental health. For example, personality traits such as anger, negative affect, and emotional inhibition

increase HPA and autonomic nervous system reactivity on an acute and/or chronic basis. These elevations, in turn, contribute to allostatic load when they represent a life-long pattern of response to challenges. We have seen that the personality traits cited previously may be the result of early life experiences as well as reflections of genetic factors. Fortunately, none of these traits, or the genetic and environmental factors that produce them, reflect an irreversible change in either behavior or allostatic load, and there are many intervention strategies that are likely to be very helpful early in development as well as later in life.

See Also the Following Articles

BIOFEEDBACK • CHEMICAL NEUROANATOMY • ENDORPHINS AND THEIR RECEPTORS • HOMEOSTATIC MECHANISMS • NEUROTRANSMITTERS • NOREPINEPHRINE • PEPTIDES, HORMONES, AND THE BRAIN AND SPINAL CORD • PSYCHONEUROENDOCRINOLOGY

Suggested Reading

Adler, N. E., Boyce, T., Chesney, M. A., Cohen, S., Folkman, S., Kahn, R. L., and Syme, L. S. (1994). Socioeconomic status and health: The challenge of the gradient. *Am. Psychol.* **49,** 15–24.

Anderson, N. B., and Armstead, C. A. (1995). Toward understanding the association of socioeconomic status and health: A new challenge for the biopsychosocial approach. *Psychosom. Med.* **57,** 213–225.

Becker, J., Breedlove, S. M., and Crews, D. (1992). *Behavioral Endocrinology*. MIT Press, Cambridge, MA.

Bobak, M., and Marmot, M. (1996). East–West mortality divide and its potential explanations: Proposed research agenda. *Br. Med. J.* **312,** 421–425.

Cahill, L., and McGaugh, J. L. (1998). Mechanisms of emotional arousal and lasting declarative memory. *TINS* **21,** 294–299.

DeKloet, E. R., Vreugdenhil, E., Oitzl, M. S., and Joels, M. (1998). Brain corticosteroid receptor balance in health and disease. *Endocrine Rev.* **19,** 269–301.

Fink, G. (2000). *The Encyclopedia of Stress*. Academic Press, New York.

Lupien, S. J., and McEwen, B. S. (1997). The acute effects of corticosteroids on cognition: Integration of animal and human model studies. *Brain Res. Rev.* **24,** 1–27.

Lupien, S. J., Brière, S., McEwen, B. S., Nair, N. P. V., and Meaney, M. J. (1999). Increased cortisol levels during human aging: Implication for the study of depression and dementia in later life. *Rev. Neurosci.* **10,** 117–140.

McEwen, B. (2000). Allostasis and allostatic load: Implications for neuropsychopharmacology. *Neuropsychopharmacology* **22,** 108–124.

McEwen, B. S. (1998). Protective and damaging effects of stress mediators. *N. Engl. J. Med.* **338,** 171–179.

McEwen, B. S. (1999). Stress and hippocampal plasticity. *Annu. Rev. Neurosci.* **22,** 105–122.

McEwen, B. S., and Seeman, T. (1999). Protective and damaging effects of mediators of stress: Elaborating and testing the concepts of allostasis and allostatic load. *Ann. N.Y. Acad. Sci.* **896,** 30–47.

Ryff, C. D., and Singer, B. (1998). The contours of positive human health. *Psychol. Inquiry* **9,** 1–28.

Taylor, S. E., Klein, L. C., Lewis, B. P., Gruenewald, T. L., Gurung, R. A. R., and Updegraff, J. A. (2000). Biobehavioral response to stress in females: Tend-and-befriend, not fight-or-flight. *Psychol. Rev.* **107,** 411–429.

Stress

D. JEFFREY NEWPORT and CHARLES B. NEMEROFF
Emory University School of Medicine

GLOSSARY

corticotropin-releasing factor Both a neurotransmitter and a neurohormone; the primary secretagogue of the hypothalamic–pituitary–adrenal axis.

diathesis A vulnerability or predisposition to illness; dependent on both genetic and acquired factors.

homeostasis Maintenance of constant conditions within the internal environment of the body.

limbic system Complex of central nervous system structures responsible for the regulation of emotions and behavior; consists of the amygdala, hippocampus, certain regions of the thalamus, and basilar cortex.

locus coeruleus The site of origin of nearly all norepinephrine-secreting neurons in the central nervous system.

norepinephrine The primary neurotransmitter utilized in activation of the sympathetic nervous system.

stress Any challenge to homeostasis that requires an adaptational response.

Stress as a concept has in recent years so preoccupied our collective imagination that it has altered the very way we speak and think. Consequently, we customarily talk of being stressed out, being under stress, taking a stress test, managing stress, relieving stress, reducing stress, suffering a stress fracture, enduring corporate stress, practicing stress medicine, or suffering from stress incontinence or posttraumatic stress disorder. However, moving beyond common discourse to a scientific discussion of stress is not a simple straightforward transition. Considerable debate remains as to the definition of stress, the nature of normal versus pathological responses to stress, and the elucidation of a neurobiology of stress that is distinct from the neurobiology of stress-related or stress-responsive psychiatric and medical disorders. In this article, we endeavor to provide a relatively comprehensive yet concise discourse on this topic. First, we offer a definition of stress that is operationally suitable for scientific discussion. Second, we describe the phenomenology of stress and the stress response in normal individuals. This includes both behavioral and biological alterations during the response to a stressor. Third, we discuss the role that stress plays in the pathogenesis and ongoing pathophysiology of stress-responsive psychiatric disorders. This incorporates a discussion of so-called diathesis:stress models of the disease process, the role that stressful life experiences play in precipitating specific psychiatric syndromes, and the functional alterations in stress-respondent biological systems observed in particular psychiatric illnesses. Finally, we briefly discuss the impact of psychopharmacological intervention on stress-induced neurobiological changes with a view to the development of novel antidepressant and anxiolytic agents.

I. STRESS: TOWARD AN OPERATIONAL DEFINITION

Most people have a reasonable understanding of what is meant by stress. When using the term in everyday

conversation, we are seldom misunderstood. Nevertheless, attaining universal agreement on a precise yet comprehensive definition is more problematic. This should not be surprising. Stress is an inherently nonspecific term that is applied in an overarching manner to a heterogeneous array of life experiences. We intuitively associate stress with negative life experiences, such as life-threatening illnesses, loss of employment, or impending divorce. However, intrinsically positive experiences such as moving into a new home, beginning a new job, or having a baby can be stressful as well.

How can we arrive at a scientifically meaningful definition of stress when such diverse life experiences are encompassed by the term? When communicating in a scientific forum, it is imperative that we speak with accuracy and precision. Unfortunately, ill-defined terms such as stress do not readily lend themselves to the empirical rectitude demanded by science. It is precisely for this reason that some investigators have advocated abandoning the concept altogether on the grounds that we too often resort to the nonspecificity inherent in abstract terms such as stress. However, the tenaciously enduring presence of stress as a scientific concept speaks to its ultimate relevance. If there were not some underlying commonality to the myriad experiences that we deem stressful, then there would be no impetus for applying the single word stress to such a broad range of life events. Similarly, there would be no reason to formulate the term "mind" had we not recognized the connections between the concepts of memory, cognition, emotion, and volition.

Just what is it that is common to the experiences that we deem as stressful? Most formulations incorporate the concepts of change and adaptability. However, before we can properly understand stress, we must first describe the concept of homeostasis. Homeostasis refers to the maintenance of constant conditions within the internal environment (milieu) of the body. It is necessary for survival because cells are capable of living, growing, reproducing, and performing their specialized functions only so long as proper concentrations of nutrients (i.e., energy sources) and other precursors to constituent molecules are readily available.

The external environment, however, is not subject to such constancy. In fact, change is often the only certainty in our surroundings. Perturbations in the external environment frequently impact the body's internal milieu. In order to maintain the constancy of the body's internal environment, we must be equipped to make adaptive changes in response to these external forces. Otherwise, homeostasis is compromised, and survival is in jeopardy. These adaptive responses include conscious efforts to manipulate the external environment so that the challenge to homeostasis is reduced. Adaptation also includes unconscious internal mechanisms that alter the function of one or more body systems in an effort to maintain constancy for the body as a whole. A simple example is our body's requirement to maintain a constant temperature of 37°C. When we walk outside on a cold, wintry day, we are faced with an environmental challenge to our ability to maintain that constant body temperature. We respond with conscious manipulation of the external milieu when we drink a cup of hot coffee, wear an overcoat, or build a fire. Our bodies also respond unconsciously by activating the sympathetic nervous system and perhaps by increasing thyroid hormone secretion. The increase in sympathetic activation in turn produces shivering that generates body heat aimed at maintaining the optimal body temperature.

In this context, we can construct an operational definition of stress that is suitable for scientific inquiry. Stress is defined as any challenge to homeostasis that requires an adaptational response. Often, stress is a consequence of a change in the external environment that perturbs the internal milieu. However, stress can be a direct result of dysfunction in one or more organ systems within the body's internal environment.

The stressfulness that we assign to an event is not solely contingent on the innate characteristics of the event. Coping mechanisms exist such that two individuals exposed to the same stressor will experience different degrees of stress. The stress experienced by a particular individual in a given circumstance depends both on the demands placed by the event and on the individual's sense of control in that situation. For example, a child's first attempt at riding a bicycle is highly stressful, but as the child learns to balance the bicycle, a sense of control is gained. Eventually, the child develops mastery over bike riding and virtually no stress is encountered. Consequently, it is meaningless to ask, "How stressful is riding a bike?" However, it may be reasonable to ask, "How stressful is learning to ride a bike?"

II. PHENOMENOLOGY OF STRESS AND THE STRESS RESPONSE

A. Adaptive and Maladaptive Stress Responses

In discussing the phenomenology of stress, the focus obviously cannot be exclusively on the stressor.

Equally important is the stress response, the biobehavioral sequelae of stress. All stress responses are intended to preserve homeostasis. In reality, however, stress responses may be adaptive or maladaptive. An adaptive response promotes the maintenance of homeostasis. A maladaptive response fails to achieve homeostasis and illness may result. When this occurs, the maladaptive response to a stressor becomes a stress and a pathophysiological cascade may ensue. For example, a stressful life event may precipitate an episode of major depression. Depression, in turn, is associated with changes in certain neurobiological systems that often have persistent adverse consequences, thereby increasing the vulnerability to future episodes of depression. The depressive episode is then partially a result of a maladaptive response to an initial stressor, but it may evolve into a stressor in its own right that persists long after the triggering life event has resolved.

The adverse consequences of stress with a particular emphasis on its psychiatric sequelae are highlighted in the ensuing discussion. This is certainly consistent with the typical negative connotations of undue stress in modern society. However, stress often has beneficial effects. For example, tempered steel is exposed to the stress of intense heat followed by sudden cooling to purge impurities. This process improves the durability, hardness, and strength of steel. Without undergoing the stress of the tempering process, steel is not a reliable building product.

Stress can offer similar benefits to organisms. Darwinian theory is predicated on the notion that stress plays a vital role in ensuring the fitness of a species by eliminating unfit organisms from the reproductive pool. Overcoming stress confers strength. As Nietzshe stated, ''That which doesn't kill you makes you stronger.'' We argue that stress is likewise the impetus behind much human ingenuity. The proverb ''necessity is the mother of invention'' illustrates this beneficial function of stress. The demands placed on human adaptability by stressful life events instigate human resourcefulness and eventuate in technological and philosophical advances. In fact, the unprecedented success of the human species is certainly, at least in part, a consequence of our superior capacity for adaptation in the context of environmental change.

It is a misnomer to think of a single invariable response to stress. Recognizing that stress is a heterogeneous phenomenon, a monolithic stress response would hardly be adaptive. Instead, an adaptive stress response is by definition tailored to respond appropriately to the particular intensity and quality of the stressor. For example, the stress of infection is met with a swift response by the immune system. In contrast, immune system changes are present but less prominent in the response to psychosocial stress. Nevertheless, there are certain behavioral characteristics common to most stress responses. In particular, physical manifestations suggestive of increased sympathetic tone are often seen in both humans and animals under stress. This has been the basis for a plethora of psychophysiological research that has demonstrated physical responses to stress, such as tachycardia, hypertension, diaphoresis, dizziness, elevations in skin temperature and skin conductance, and increased electromyographic activity.

Here, three points of paramount importance should be considered. First, there are profound differences in the central nervous system (CNS) response to physical (e.g., hypovolemia) and psychological stressors. Second, although adaptation to homologous stressors has been demonstrated, this does not generally hold true for heterologous stressors. Third, the developmental/life cycle period in which the stressor is applied is critical to the nature and course of the stress response.

B. Stress Patterns and the Characteristics of the Stress Response

In life, the events that constitute stress are often largely beyond our control. In the laboratory setting, however, the characteristics of the stressor can be delicately manipulated. This includes both qualitative and quantitative aspects of the stressor. Qualitatively, experimental stressors are composed of those that represent an addition to the environment and those that consist of a subtraction from the environment. Additions to the environment include unpleasant stimuli, such as air puffs to the face, electric shock, temperature extremes, disturbing loud noises, or intense lights. Stressful subtractions from the environment entail some form of deprivation, including food deprivation, maternal deprivation, and isolation (i.e., deprivation of stimuli).

Equally important is the quantitative presentation of the experimental stressor. This refers to the temporal schedule of stressor presentation and includes single short-term exposure, single long-term exposure, and repeated exposure. The temporal characteristics of stressor presentation often exert a marked impact on the stress response. For example, a single short-term stress exposure initially elicits an

orientation response. During orientation, the experimental subject ceases its prestress activity and directs its attention toward the source of the stress stimulus.

Single long-term stress exposure typically consists of placement into a novel environment. The initial response is commonly fear and inactivity (i.e., freezing) that is followed by exploration. This exploration is less a product of curiosity than an effort to reduce anxiety by eliminating the uncertainty precipitated by placement into an unknown environment.

Repeated stress exposure can elicit two primary patterns of stress response: sensitization and habituation. During the initial repetitions of a stressful stimulus, the characteristic orientation response may be exaggerated. Although stress sensitization can be persistent [as is perhaps the case in some patients with posttraumatic stress disorder (PTSD)], continued repetition of the stressor usually results in a habituated response. Through habituation, the ability of a stressor to elicit the characteristic orientation response is lost. Repeated presentation of the stressor thus results in an adaptive response that enables the organism to gain a sense of control over the stress stimulus. Thus, the requirement for an orientation response is diminished.

An important variant is repeated uncontrollable stress exposure. The best example is the uncontrollable shock paradigm used in rodents. This experimental animal model is believed to be homologous to depression in humans and is popularly known as the learned helplessness model. In this application, the initial exaggerations in the orientation response are also lost after repeated administrations of shock. The test animals instead become less active after repeated shocks, but the conception of this as a manifestation of learned helplessness is far from certain. The animal's inactivity may instead be a consequence of extreme fear and anxiety and thus represent an extension of the freezing response typically seen in the initial aftermath of a single long-term stress exposure. The distinction is that learned helplessness would essentially represent surrendering to the uncontrollable stress, whereas prolonged freezing is a consequence of anxious indecision.

III. STRESS AND PSYCHIATRIC ILLNESS: DIATHESIS/STRESS MODEL

Stress is generally acknowledged to play a paramount role in the pathogenesis of many psychiatric disorders. However, not everyone exposed to a given stressor generally considered likely to precipitate a psychiatric syndrome becomes ill. For example, pancreatic cancer has long been believed to convey a substantial risk for developing a major depressive episode. In fact, depression commonly predates the onset of physical symptoms of the cancer. Nevertheless, one-half of patients with pancreatic cancer do not become depressed. Similarly, PTSD remains a significant burden to many Vietnam combat veterans nearly 30 years after the conclusion of that war. However, most Vietnam combat veterans (85%) do not suffer from PTSD. Why do some individuals succumb to a particular stressor and become ill while others do not?

Stressful life events can obviously serve as acute precipitants to psychiatric and medical illness. However, some individuals tolerate stress of great magnitude and long duration without becoming ill. Others exhibit a constitutional vulnerability to the effects of stress (i.e., a lower threshold of tolerance for stress that predisposes them to stress-induced illness). This inherent vulnerability to the adverse effects of stress is known as a diathesis and provides the basis for the *diathesis/stress disease model*.

The diathesis/stress model has in recent years been applied to a broad range of psychiatric and medical disorders, including major depression, schizophrenia, chronic fatigue syndrome, PTSD and other anxiety disorders, sexual disorders, and pain disorders such as fibromyalgia and arthritis. This model theoretically has practical application to other psychiatric syndromes, including somatoform disorders, eating disorders, attention-deficit hyperactivity disorder, and impulse control disorders. In addition, neurological disorders including epilepsy and migraine headaches, rheumatological diseases such as systemic lupus erythematosus, and other illnesses such as irritable bowel syndrome and diabetes mellitus may be appropriate to consider from the framework of a diathesis/stress model.

What is the origin of the diathesis? The relative contributions of genetic inheritance and environmental exposure to the susceptibility to illness have been deliberated in often contentious nature versus nurture debates. Such arguments are often couched in overly absolute terms, but the diathesis/stress model permits a more balanced consideration. Diathesis/stress models recognize that both inherited and acquired factors may contribute to the vulnerability to stress.

A full accounting of the genetic contribution to the predisposition to stress-related illness is beyond the scope of this article. Nevertheless, human twin studies have clearly revealed a significant genetic contribution to the vulnerability to many psychiatric syndromes,

including depression, bipolar disorder, and schizophrenia. To date, it has not been possible to identify with certainty the chromosomal localization much less the precise gene(s) that forms the basis of the genetic contribution to the diathesis for psychiatric disorders. From years of largely unsuccessful research by psychiatric geneticists, it is clear that the genetic contribution to the vulnerability for psychiatric illnesses is unlikely to arise from simple single gene Mendelian transmission. It is more likely that the heritable vulnerability to psychiatric illness arises either from complex polygenic patterns of inheritance or from even more complex epigenetic modification of genotypic risk. For example, in recent years, the existence of resistance genes that interact with susceptibility genes has been postulated.

Epidemiological research also indicates that the environment makes a substantial contribution to the vulnerability to psychiatric illnesses. Environmental contributions to the diathesis may emerge from any of a variety of biopsychosocial stressors. Consequently, stress plays a dual role in the pathogenesis of psychiatric illness. When stress is coincident with the onset of an illness, it serves as an acute precipitant to the disorder. In contrast, when stress predates the onset of a disorder, it may well shape the predisposition to future illness. Although the major environmental contributions to the diathesis occur during the formative childhood years, the diathesis remains mutable throughout adult life. Stresses during adulthood continue to modify the predisposition to illness. In fact, disease is a stressor that can increase the risk for future episodes of illness. Theoretically, the stress-induced predisposition to illness should be both psychometrically and biologically measurable. It is this impression that underlies a burgeoning line of research investigating the persistent neurobiological sequelae of early life adverse experiences.

IV. NEUROBIOLOGY OF THE STRESS RESPONSE

A. Norepinephrine and the Sympathetic Nervous System

Among classic neurotransmitters, norepinephrine (NE) is by far the most thoroughly studied and arguably the most crucial in the biological response to stress. Long-standing psychophysiological research has demonstrated physical responses to stress suggestive of increased sympathetic nervous system activity. Like dopamine (DA) and epinephrine (EPI), NE is a catecholamine that is formed from the enzymatic conversion of the precursor amino acid tyrosine. The rate-limiting enzyme in the production of the catecholamines is tyrosine hydroxylase, which converts tyrosine to L-DOPA. L-DOPA is subsequently converted in the neuronal cytosol to DA by the enzyme DOPA decarboxylase. DA is then packaged into storage vesicles in both DA-secreting and NE-secreting neurons. In the NE neurons, another enzyme, dopamine β-hydroxylase, resides in the storage vesicles and transforms the stored DA into NE.

There are two stress-respondent components to the NE neuronal system. One resides in the CNS, and the other resides in the peripheral sympathetic nervous system (SNS). Within the CNS, the cell bodies of NE-containing neurons are primarily located within the locus coeruleus, a small area in the brain stem at the posterior junction between the pons and mesencephalon. The axons of these neurons extend from the locus coeruleus to distribute widely throughout the cerebral cortex, subcortical limbic structures, cerebellum, and brain stem. The vast majority of the CNS NE neurons project unilaterally from the locus coeruleus. The phenomenological significance of this unilateral distribution remains obscure; however, it may in part explain the clinical discrepancy in psychiatric outcome between patients with right versus left hemispheric strokes.

NE neurons are also a key component in the SNS. The SNS is composed of a chain of ganglia outside the CNS. Each ganglion is an aggregation of neuronal cell bodies. The SNS is therefore composed of preganglionic and postganglionic neurons. The preganglionic sympathetic (and parasympathetic) neurons utilize the neurotransmitter acetylcholine to transmit signals from the CNS to the ganglion. Each postganglionic neuron transmits signals from a ganglion to receptors on nonneuronal cells. All postganglionic sympathetic neurons secrete NE. These neurons are distributed throughout much of the body and are responsible for mediating the acute physical manifestations of the stress response.

Each ganglion has a well-defined anatomical distribution. Certain of the postganglionic SNS neurons projecting from the celiac ganglion extend to the adrenal medulla, a region of the adrenal gland that is of embryonic neuroectodermic origin. When stimulated by the preganglionic splanchnic nerve, the adrenal medullae, a homolog of the postganglionic neuron, secretes vast amounts of EPI (80%) and NE (20%)

into the peripheral circulation. The SNS and the adrenal medulla together have also been termed the sympathoadrenal system. The biobehavioral effects of adrenally secreted catecholamines are fundamentally identical to those of SNS activation; however, the effects produced by the circulating EPI and NE last up to 10 times longer.

Circulating NE causes vasoconstriction. In addition, NE and EPI increase the contractility and rate of the heart, dilate the pupils, and diminish the activity of the digestive system. EPI, however, does have a more significant impact on cardiac output than NE because it increases cardiac contractility to a greater degree than blood vessel constriction. The marked increase in cardiac output induced by EPI provides a critical increase in blood supply to the muscles during times of stress. Finally, EPI has profound effects on metabolism, triggering the release of glucose into the bloodstream from glycogen stores in liver and muscle.

B. Corticotropin-Releasing Factor, the Limbic System, and the Hypothalamic–Pituitary–Adrenal Axis

The hypothalamic–pituitary–adrenal (HPA) axis ultimately regulates the secretion of glucocorticoids, adrenocortical steroids, that act on target tissues throughout the body in order to preserve homeostasis during stress. Properly understood, the HPA axis is not simply an endocrine system but a neuroendocrine system. The hypothalamus, long considered the pinnacle of the HPA axis, is in reality a relay point for information exchange between limbic components of the CNS and the HPA axis. Because the hypothalamus has numerous bidirectional communication pathways with the limbic system that modulate HPA axis activity, the HPA axis in its broader sense might be better labeled the limbic–hypothalamic–pituitary–adrenal axis.

Literally meaning border, the term limbic was originally used in reference to the anatomical structures bordering the interface of the cerebrum with phylogenetically more primitive CNS regions. It has since been discovered that one of the primary functions of the structures in the limbic region is the modulation of emotions and motivational drives. Thus, the typical conception of the limbic system has evolved to refer more globally to the complex of CNS structures responsible for the regulation of emotions and behavior. The key components of the limbic system include the amygdala and hippocampus, certain regions of the thalamus, and a circle of basilar cerebral cortex including the cingulate gyrus, the orbitofrontal cortex, the prefrontal cortex, the parahippocampal gyrus, and the uncus. The limbic structures each have a unique function in the regulation of emotion and behavior. The amygdala in lower animals (that are dependent on a well-developed sense of smell for survival) facilitates the sensory association of olfactory stimuli. In humans, the role of the amygdala in olfaction is negligible. Instead, the primary function of the human amygdala is to modulate behavioral responses to emotionally charged stimuli. Coordinating the interplay between thoughts and the environmental context at a preconscious level, the amygdala shapes the intensity and direction of human behavioral responses to esure that they are contextually appropriate. The hippocampus also processes sensory input. However, the hippocampus's primary function is to consolidate emotion-laden and other experiences into long-term memory. It is important to recognize that the hippocampus is a remarkably plastic structure. Recent evidence from both rodents and nonhuman primates indicates that granule cell neurogenesis in the hippocampus persists well into adulthood and likely throughout life. Equally important is the fact that glucocorticoids, in conjunction with the excitatory neurotransmitter glutamate, interfere with hippocampal neurogenesis. This may well explain, in part, the oft-observed decremental effect of stress on memory and the reported association between chronic stress and hippocampal atrophy.

The ring of limbic cortex that encircles these subcortical limbic structures is perhaps the least understood component of the system. As a whole, the limbic cortex serves as an association area for behavioral regulation acting to coordinate the exchange of information between higher regions within the neocortex and the subcortical limbic structures. The prefrontal cortex plays a key role in the working memory component of explicit memory and may perform a counterregulatory function in the stress response through inhibitory effects on the amygdala. The anterior cingulate gyrus is responsible for the maintenance of social mores, fear-related behavior, and selective attentional processing. Finally, the orbitofrontal cortex is involved in conditioned fear processing and its extinction.

Although anatomically segregated from the remainder of the limbic structures, the hypothalamus, from a physiological standpoint, is a central component of the limbic system. Neuronal transmission from the hypothalamus travels in three directions: downstream

within the CNS to autonomic control regions within the brain stem such as the locus coeruleus, upstream within the CNS to the cortical and subcortical limbic structures, and outside the CNS via the infundibulum to the pituitary gland.

The HPA axis is the primary biological stress response system in mammalian species. A broad range of experimental stressors reliably induce the cascade of HPA axis activation. Stressors shown to induce HPA activation in human and animal studies include psychosocial stress, thermal extremes, infection, surgery or anesthesia, physical or emotional trauma, forced restraint, and the administration of NE and serotonin (5-HT) agonists. Although other biological systems often participate in the stress response, the HPA axis is unique in its reliable induction by such a wide assortment of stressors. It may be assumed that HPA axis activation represents the final common pathway of the mammalian stress response.

A cascade of biological events presumably beginning in the hypothalamus and proceeding through the anterior pituitary gland and ultimately the adrenal cortex comprises HPA axis activation. Within the context of the HPA axis, the primary CNS chemical mediator in the hypothalamus is corticotropin-releasing factor (CRF). CRF is a 41-amino acid neuropeptide that was structurally characterized in 1981. It is well established as the primary secretagogue stimulating HPA axis activity. CRF functions as both a CNS neurotransmitter and a hormone. In its latter role, CRF is secreted from hypothalamic neurons with nerve terminals in the median eminence, an enlarged area of the infundibulum just inferior to the hypothalamus. From the median eminence, CRF enters an extensive capillary bed that flows into a venous network known as the hypothalamic hypophysial portal system. This portal system travels directly to the anterior pituitary gland, where another capillary bed enables the CRF to gain access to a class of secretory cells in the anterior pituitary known as corticotrophs. Comprising about 20% of the anterior pituitary gland's overall volume, the corticotrophs increase the production and release of adrenocorticotropin hormone (ACTH) when stimulated by CRF. Although other hormones, including vasopressin, oxytocin, and somatostatin, act either to enhance or diminish ACTH release, CRF is by far the primary regulator of ACTH secretion.

ACTH is a 39-amino acid peptide that is derived from a larger precursor molecule known as proopiomelanocortin (POMC). The enzymatic cleavage of POMC produces three bioactive peptides: ACTH, β-endorphin, and melanocortin. Consequently, stress exposure typically induces parallel increases in serum concentrations of ACTH and β-endorphin. ACTH travels via the circulatory system to the adrenal cortex, where it stimulates the production and release of glucocorticoids such as cortisol.

The glucocorticoids are a group of structurally similar adrenocortical hormones that are synthesized from cholesterol. In humans, 95% of the circulating glucocorticoid is cortisol, whereas corticosterone comprises most of the remaining 5%. More than 90% of the circulating glucocorticoids are bound to serum proteins such as cortisol-binding globulin (CBG), leaving less than 10% in the unbound bioavailable state. In addition to increasing the adrenal secretion of cortisol, stress-induced HPA axis activation also decreases the circulating concentration of CBG. Therefore, the circulating level of the biologically active free form of the hormone is increased by two distinct mechanisms.

The name glucocorticoid is derived from the demonstration that these hormones increase glucose production. Glucose is a ready energy source that is used by nearly all cells of the body. When responding to stress, it is obviously adaptive to increase the availability of this fuel source in a rapid fashion. However, increasing the circulating levels of glucose comes at a price. Glucose production is increased by cortisol through the catabolic mobilization of protein stores from muscle and by the concomitant activation of hepatic enzymes that convert glycogen into glucose. Cortisol also increases the availability of a secondary energy source by inducing the mobilization of fatty acids from adipose tissues. Thus, HPA axis activation sacrifices muscle mass and adipose tissue in order to provide energy sources necessary to meet the challenges of stress.

In the short term, cortisol-induced catabolic effects are biologically critical to the mounting of an adaptive stress response. However, persistent HPA axis hyperactivity can lead to a depletion of protein, carbohydrate, and fat stores. This would eventually produce adverse (and hence maladaptive) consequences. The HPA axis is therefore served by negative feedback mechanisms that decrease its activity once sufficient glucocorticoid release has been attained. This negative feedback effect is mediated by glucocorticoid receptors at all levels of the HPA axis, including the hippocampus, the hypothalamus, and the pituitary and adrenal glands. Once the cascade of HPA axis activation has triggered sufficient cortisol release from the adrenal cortex, this hormone is carried via the circulatory

system to target tissues throughout the body. Some of the cortisol finds its way to these "feedback" receptors. Cortisol feedback at the hypothalamus reduces CRF release, at the pituitary it inhibits ACTH release, and at the adrenal gland it inhibits further cortisol release. Cortisol feedback at the hippocampus inhibits CRF secretion from the hypothalamus (and decreases granule cell neurogenesis). In addition, there are other feedback and feedforward mechanisms that render the HPA axis an exquisitely tuned and balanced stress response system.

V. NEUROBIOLOGY OF STRESS-RELATED PSYCHIATRIC ILLNESS: FOCUS ON DEPRESSION AND PTSD

As we examine the neurobiology of psychiatric disorders, the clinical significance of stress is evident. The very biological systems that are most intimately involved in the mammalian stress response (i.e., the HPA axis and NE neuronal systems) are those that are clearly awry in psychiatric patients. By far, the psychiatric disorder most extensively studied is depression. Consequently, our discussion emphasizes the neurobiological alterations in depressed patients. A second stress-related psychiatric disorder, PTSD, is also associated with neurobiological changes in these same systems, although the pattern of alteration differs from that seen in depression. Therefore, the neurobiology of PTSD is compared to that of depression in an effort to provide a comprehensive view of the range of dysfunction in biological stress response systems that occur in the context of psychiatric illness.

A. Neurobiology of Depression

The bulk of the extant data regarding the pathophysiology of depression address the role of monoamines, including NE, 5-HT, and, to a lesser extent, DA. Accordingly, modulation of the synaptic activity of these neurotransmitters is the principal pharmacodynamic mechanism thought to underly the action of all currently available antidepressants. Dating back to the mid-1960s, the earliest modern neurobiological hypothesis of depression proposed that the illness (or at least some forms of it) is a consequence of a deficiency in catecholaminergic neurotransmission, particularly NE. An abundance of research regarding the role of NE in the pathophysiology of depression has since accumulated. Although controversial, the majority of data still suggest a relative NE deficiency, although discordant results are available.

Recently, neurobiological inquiry has increasingly focused on changes in neuroendocrine function in depressed patients. The most intensely scrutinized of the neuroendocrine systems has been the HPA axis. The earliest reports of HPA axis dysfunction in depression emerged more than 40 years ago when it was recognized that many depressed patients have higher baseline serum cortisol concentrations. Shortly thereafter, the dexamethasone suppression test (DST), the first functional assessment of HPA axis physiology, was developed as a diagnostic tool in the evaluation of Cushing's Syndrome. A synthetic analog of cortisol, dexamethasone, is used in the DST to evaluate the activity of the HPA axis. Dexamethasone activates HPA negative feedback and thus suppresses the circulating concentrations of ACTH and cortisol. However, cortisol release is not suppressed by dexamethasone in many depressed patients. This suggests that the HPA feedback mechanism is dysfunctional.

A myriad of studies have subsequently confirmed these early reports of HPA axis hyperactivity in depressed patients. Several plausible theories of the ultimate cause of hypercortisolemia have been proffered. It may be that adrenal ACTH receptors are upregulated in depressed patients and therefore hypersensitive to ACTH stimulation. Administration of pharmacological doses of exogenous ACTH does in fact elicit an exaggerated cortisol response in some depressed patients. However, administration of lower physiological doses of ACTH to patients with depression after pretreatment with dexamethasone does not potentiate increased cortisol release. This suggests that adrenal sensitivity to ACTH stimulation is not appreciably elevated in depressed patients, although adrenocortical capacity may be increased.

Indeed, anatomical studies of depressed patients, utilizing either imaging techniques or postmortem dissection of the adrenal glands in suicide victims, demonstrate adrenal gland enlargement. Pituitary gland enlargement in depressed patients has also been reported. Future studies combining volumetric pituitary and/or adrenal imaging with physiological dose CRF and/or ACTH stimulation tests may help clarify the relative contribution of glandular hypertrophy to HPA axis hyperactivity.

However, impressive accumulation of data indicates that the predominant cause of HPA axis hyperactivity in depressed patients is likely hypersecretion of CRF. Several lines of evidence indicate that depressed patients hypersecrete CRF. First, administration of

CRF directly into the CNS of laboratory animals produces behavioral alterations homologous to those observed in depressed humans. These changes include decreased appetite, disrupted sleep, and diminished libido. Second, similar behavioral changes are witnessed in transgenic (i.e., genetically engineered) mice that overexpress CRF. Third, baseline CRF concentrations in the CSF of depressed patients have repeatedly been reported to be increased. Fourth, depressed patients commonly demonstrate a blunted ACTH response in the CRF stimulation test. In this test, serum samples of ACTH and cortisol are collected at baseline prior to the administration of exogenous CRF and at regular intervals post-CRF administration. The change in ACTH and cortisol concentrations induced by CRF provide an indication of pituitary and adrenal responsivity to hormonal stimulation. The blunted ACTH response in depressed patients is believed to be a consequence of pituitary CRF receptor downregulation that occurs as a result of chronic CRF hypersecretion. However, it should be recognized that pretreatment with metyrapone (a potent inhibitor of cortisol production that in effect temporarily eliminates HPA axis negative feedback) before CRF administration increases the ACTH response in depressed patients. This implies that mechanisms other than CRF receptor downregulation contribute, at least in part, to the blunted ACTH response to CRF stimulation. Nevertheless, potential genomic effects of metyrapone on glucocorticoid receptor density make these data difficult to interpret. Finally, the most persuasive evidence of CRF hypersecretion comes from the postmortem histological analysis of depressed patients. Patients who died with untreated depression exhibited a threefold increase in the number of hypothalamic CRF-secreting neurons and an increase in CRF mRNA expression, indicating that neurons that do not normally secrete CRF are recruited to do so during episodes of depression.

These data raise the crucial question of whether CRF hypersecretion only occurs during a depressive episode or exists during periods of remission. At issue is whether CRF hypersecretion is a trait marker for the risk of becoming depressed or a state marker for active illness. This question is important on two counts. First, it is critical to our understanding of the pathophysiology of depression. If CRF hypersecretion is a trait-dependent variable, then alterations in CRF activity likely contribute to the vulnerability for depression but other biological mechanisms induce an episode of illness. If, however, CRF hypersecretion is state dependent, then it is likely necessary in the development of the disorder. If CRF hypersecretion is state dependent, then the next step in studying the neurobiology of depression is to investigate the mechanisms that induce persistent CRF hypersecretion. If CRF hypersecretion is trait dependent, then the next step is to determine why someone who chronically hypersecretes CRF suddenly becomes depressed. This question is also important from a clinical standpoint. If CRF hypersecretion is trait dependent, then it may afford an opportunity to devise a screening test that reliably identifies those at risk for becoming depressed. If it is state dependent, then we may ultimately be able to devise tests to guide medication selection and monitor treatment response.

A definitive answer to the state vs trait question could be obtained by measuring CRF activity in healthy volunteers who had never been depressed and then repeating the measure if and when any of those subjects ever become depressed. Unfortunately, such research would be costly, unwieldy, and is unlikely to ever be undertaken. Short of this, assessing CRF activity during an episode of depression and repeating that measure once the depression has resolved can provide some insight. These studies do exist. In fact, both elevations of CRF concentrations in the CSF and ACTH blunting in the CRF stimulation test have been shown to normalize when depression resolves after treatment with an antidepressant or a course of electroconvulsive therapy. Of course, this indicates that CRF hypersecretion is likely state dependent in patients with depression. Some have argued, however, that indirect effects of these somatic therapies on HPA axis activity in depressed patients may confound these results. In other words, did the CRF hypersecretion resolve because the depression remitted or because the patient was treated with an antidepressant (a fine distinction but a fair and important one nonetheless)? Although HPA axis activity has been demonstrated to remain normal in successfully treated patients even after the course of antidepressant medication has been discontinued, a more satisfying answer to this question will no doubt be provided when future studies investigate HPA axis activity before and after a successful nonpharmacological treatment (e.g., psychotherapy). However, using a very sensitive measure of HPA axis activity, the combined dexamethasone–CRF stimulation test, Holsboer and colleagues not only confirmed the HPA axis abnormalities in depressed patients but also, most important in relationship to the state-trait issue, identified abnormalities in this test in asymptomatic first-degree relatives of a cohort of depressed patients.

We have long assumed that stress has a detrimental impact on immune system function; however, research results have been inconsistent. Investigations of the association between stress and immune system disturbance include measures of both cellular and humoral immunity. Cellular immunity studies measuring cellular enumeration and immune cell proliferative responses in depressed subjects suggest a pattern of cellular immunosuppression. The most consistent finding has been decreased natural killer cell activity and decreased T cell responses to mitogen stimulation. However, components of the humoral immune system that are activated as part of an acute phase response are enhanced in depressed patients. In particular, depressed patients exhibit increases in serum concentrations of numerous proinflammatory cytokines, including IL-1, IL-2, and IL-6.

How can immunosuppressive measures such as decreased natural killer cell activity be reconciled with this apparently contradictory evidence of immunoactivation? One hypothesis is that stress induces bidirectional, homeostatic interactions between the HPA axis and the immune system. In depression, increased CRF secretion has been associated with humoral immunoactivation as evidenced by increased proinflammatory cytokine release. These cytokines in turn appear to augment HPA axis function by promoting additional CRF release and by inducing glucocorticoid resistance that impairs HPA axis negative feedback. Conversely, HPA hyperactivity, presumably via a direct effect of glucocorticoids, also elicits immunosuppression. In summary, depression produces a mixed picture of immunological activation, immunological suppression, and HPA axis hyperactivity.

B. Neurobiology of PTSD

NE is certainly the most thoroughly studied and undoubtedly a most critical neurotransmitter involved in the biological response to traumatic stress. An extensive array of evidence indicates that psychophysiological measures suggestive of SNS activation are accompanied by NE hypersecretion in patients with PTSD. For example, many studies have reported increased urinary concentrations of both NE and EPI in numerous groups of patients with PTSD, including combat veterans, civilian assault victims, and abused children.

However, the pattern of NE hyperactivity in PTSD is not characterized by elevations in baseline plasma NE concentrations. Instead, markedly abnormal increases in NE and MHPG plasma concentrations occur in response to provocative stimuli such as exercise or exposure to trauma-related cues. Interestingly, both oral and intravenous administration of the α_2-adrenergic receptor antagonist yohimbine (which increases NE activity by blocking presynaptic negative feedback) exacerbates reexperiencing symptoms and triggers panic attacks in many patients with PTSD. Noradrenergic dysfunction remains a major focus of PTSD research, but forthcoming studies will increasingly investigate interactions between NE activity and other biological systems.

Not surprisingly, the HPA axis has also been relatively well studied in PTSD patients. HPA axis dysfunction in these patients appears to distinguish them not only from healthy volunteers but also from patients with major depression. Like patients with depression, PTSD patients hypersecrete CRF as evidenced by elevated levels of CRF in cerebrospinal fluid (CSF) and a blunted ACTH response to exogenous CRF. However, despite CRF hypersecretion, most, but not all, studies of PTSD have paradoxically demonstrated not elevated but rather reduced serum cortisol concentrations. The pathophysiology underlying the relative hypocortisolemia in patients with PTSD remains obscure; however, some have postulated this to be a consequence of exaggerated HPA axis negative feedback. Two lines of evidence support this contention. First, DST studies of HPA axis feedback indicate that patients with PTSD may be supersensitive to cortisol suppression, in marked contrast to cortisol nonsuppression often seen in depressed patients. Second, enhanced negative feedback may be a consequence of glucocorticoid receptor upregulation as evidenced by their increased density on peripheral lymphocytes. These receptor changes in humans are paralleled by findings from preclinical laboratory animal research utilizing a stress sensitization model. In this model, rodents exposed to an acute stress paradigm exhibited the enhanced HPA axis feedback observed in PTSD. Furthermore, these acutely stressed animals demonstrated increased hippocampal glucocorticoid receptor mRNA expression.

The limited immunological research in PTSD suggests immune system alterations in this disorder are as distinct from those seen in depressed patients as the HPA axis alterations. PTSD research indicates that circulating proinflammatory cytokine concentrations (particularly IL-6) are increased in patients with PTSD, as also reported in depression. However, cellular immune measures in PTSD and depression appear to differ. Although most cellular immunity

studies in depression indicate immunosuppression, the limited available data suggest that cellular immunity may in fact be enhanced in PTSD. Increases in total leukocyte count, total T lymphocyte count, CD4 and CD8 T cell counts, and the proportion of activated to nonactivated lymphocytes have all been reported in PTSD patients.

Although immunological findings in PTSD patients are at best preliminary, they suggest potential differences in the immune system responses between depression and PTSD that are of heuristic value. CRF hypersecretion and humoral immunoactivation are common to both disorders and may represent a shared pathophysiology. In depression, CRF hypersecretion precipitates hypercortisolemia that may in turn produce cellular immunosuppression. However, most PTSD studies have demonstrated low to normal cortisol levels despite increased CRF secretion. Consequently, the relative hypocortisolemia of PTSD may preserve or even exaggerate cellular immune activity.

In comparing the biological findings of depression and PTSD, we can learn much regarding stress-associated pathophysiology. Several key biological changes are common to the two disorders (e.g., CRF hypersecretion and humoral immunoactivation). However, other alterations are unique to each particular syndrome. Their common cooccurrence represents another difficult complexity in this field. Just as the phenomenological response to stress may vary, the neurobiological consequences of stress exposure may take distinct paths as well.

VI. TOWARD A NEUROBIOLOGY OF DIATHESIS

There is convincing evidence that stressful early life experiences confer a heightened vulnerability to psychiatric illness during adulthood. The contribution of stress during the formative years to a psychiatric diathesis appears to be particularly relevant for the depressive disorders, certain anxiety disorders such as PTSD, and perhaps substance use disorders. Consequently, an expanding line of research has searched for evidence of persistent neurobiological alterations induced by early adverse experiences that may represent the biological substrate of illness predisposition. Such neurobiological changes would theoretically represent a trait marker for a subset of patients vulnerable to depression and other forms of stress-related psychiatric illness. To date, the bulk of this data derive from animal research in which young animals are exposed to a stressor and then assessed at a later date for evidence of enduring behavioral and neurobiological changes precipitated by the earlier stress. Recent work has extended this research to human populations that have suffered traumatic experiences during childhood.

A. Laboratory Animal Studies of Early Life Stress

Among the earliest animal studies of early life stress were those that implemented handling of neonatal rodents by laboratory personnel as the initial stressor. In these protocols, infant rats or mice are removed each day from their cages and handled briefly by laboratory personnel. The handled pups are subsequently returned to the maternal cage. If it is our contention that early life stress increases the vulnerability to subsequent stress, we would expect heightened biological measures of stress responsivity in animals subjected to the stress of handling as infants. However, neonatal handling paradoxically decreases anxiety-like behaviors and lowers the magnitude of the biological response to later stress. Handled pups are resistant to stress-dependent decrements in learning ability and demonstrate increased exploratory behavior in novel environments. Although neonatal handling induces transient increases in noradrenergic neuronal activity, it reduces HPA axis activity. Indeed, HPA axis sensitivity to stress is lower in rodents that were exposed to the stress of neonatal handling than in those exposed to no neonatal stress. This appears to be a consequence of decreased CRF secretion and enhanced HPA axis negative feedback.

Certainly, these results are counterintuitive and seem to undermine the notion of a diathesis/stress model. However, on closer inspection, the findings from the neonatal handling studies do not necessarily refute this model. Although posthandling stimulation of maternal care may explain the apparently beneficial effect of neonatal handling, another potential explanation is that rodent pups do not experience handling as a particularly unpleasant experience. Neonatal handling introduces an environmental change that in turn induces homeostatic responses (as evidenced by acute changes in noradrenergic activity), but it is not an empirically noxious laboratory stimulus. It may be that the mild stress associated with a novel but nonnoxious environmental perturbation such as handling may improve the animal's adaptability to subsequent environmental changes. In that case, the mild

stress of handling may be beneficial to the development of the young animal's adaptive repertoire.

Other attempts to study the effects of early life stress utilized protocols patterned after those previously implemented to induce stress in adult animals. In these models, young animals (typically rodents) are exposed to a noxious stimulus, such as temperature extremes, pin-prick, surgical procedures, or surgical anesthesia. Interestingly, such noxious stimuli evoke a subnormal HPA axis response during the first 2 weeks of life in the rat. During this so-called stress hyporesponsive period, noxious stimuli do not induce CRF hypersecretion and only minimally increase ACTH and corticosterone (the principal glucocorticoid in the rat). This indicates that the HPA axis is a dynamic system that (in the rat) continues to undergo maturational adaptations during postnatal development. Although the young rat's failure to mount a full stress response during early development may increase its immediate vulnerability, this likely provides an adaptive evolutionary trade-off that protects the growing pup from the potentially devastating catabolic effects of heightened glucocorticoid secretion.

The two early life stress paradigms that have produced the most robust findings are maternal deprivation studies and variable foraging studies. In maternal deprivation studies, immature animals are separated from their mothers for a defined interval (or repeated intervals). These studies have been conducted using both rodent and nonhuman primate models. Not only are the separated pups deprived of maternal care during the separation period but also the separation frequently produces aberrant changes in maternal behavior even after the pups are returned. Maternal deprivation produces abrupt increases in HPA axis activity as indicated by elevations in the serum concentrations of corticosterone (the rodent equivalent to cortisol) and ACTH, in conjunction with CRF hypersecretion.

More relevant to the diathesis/stress model, the HPA axis changes induced by maternal deprivation persist into adulthood. When exposed to a stressor during adulthood, rats that were maternally deprived as pups exhibit an increased ACTH and corticosterone response compared to nondeprived controls. Furthermore, maternally deprived adult rats demonstrate increases in hypothalamic expression of CRF mRNA in the hypothalamus, increased CRF concentrations in the median eminence, and increased CSF CRF concentrations.

Maternal deprivation research has also been conducted in nonhuman primates, although these are perhaps better termed social deprivation studies because the young animal is separated from the social group as a whole. Later refinements in which the infant's absent mother is replaced by one of a series of incrementally more realistic maternal surrogates have provided an opportunity for assessment of varying degrees of early life social deprivation in these animals. These older studies did not take place in the modern era of neurobiology; as such, few validated neuroendocrine measures are available from such experiments.

The most significant recent advance in early life stress research in non-human primates is the variable foraging studies. This protocol has the advantage of minimizing direct human contact with the young animal. Instead, the researchers modulate an environmental factor that does not directly stress the infant but tests the mother's ability to provide adequate maternal care. The purpose is not to starve the animals but to increase the work of parenting while allowing sufficient nutrition for both mother and infant. Three experimental conditions exist in the variable foraging paradigm: low foraging demand (LFD), high foraging demand (HFD), and variable foraging demand (VFD). In the LFD condition, food is always readily available with minimal effort. In the HFD condition, the primate mother must always expend some effort (e.g., digging through wood chip bedding) to obtain food. Finally, in the VFD condition, the amount of maternal effort necessary to obtain food varies randomly from the LFD to the HFD condition.

Adult primates raised under VFD conditions demonstrate higher baseline CSF CRF concentrations than either the HFD and LFD groups. However, the VFD offspring exhibit lower CSF cortisol concentrations than the comparator groups, a finding that is strikingly similar to the HPA axis changes reported in PTSD. CSF concentrations of somatostatin, another peptide with HPA axis activity, are also elevated in adult primates raised under VFD conditions.

B. Human Studies of Early Life Stress

Clinical research on the impact of early life stress on HPA axis function is scant. There is increasing evidence, however, that children exposed to traumatic events (e.g., childhood sexual or physical abuse) exhibit alterations in HPA axis activity. It appears that these HPA axis changes persist into adulthood. For example, nondepressed adult women who were abused as children exhibit enhanced ACTH responses to the administration of exogenous CRF and in

response to performance of a standardized psychosocial stress test. However, these same women demonstrate normal cortisol responses to both the hormonal and the psychosocial challenges. If these women are currently depressed, then the ACTH and cortisol responses to the CRF stimulation are blunted but the hormonal responses to the psychosocial stressor are dramatically enhanced.

These findings suggest that traumatic early life experiences may precipitate a persistent sensitization of the anterior pituitary with a counterregulatory adrenocortical adaptation. However, the adrenal adaptation is apparently overridden when survivors of child abuse become depressed. Perhaps the most intriguing finding arises when we compare the pituitary responses of depressed women who were abused during childhood to the two types of stressors (i.e., hormonal and psychosocial). Administration of exogenous CRF produces a blunted ACTH response in depressed survivors of child abuse, but exposure to a psychosocial stress precipitates marked increases in ACTH concentration in this same group. How can we understand this seeming contradiction? If CRF is not driving the ACTH response to the psychosocial stressor, as the results of the CRF stimulation test suggest, then what biological substrate is inducing the release of ACTH in the depressed/abused group? There is no clear answer to this question, although potentiation of CRF release by vasopressin is one possibility.

These represent the first human studies to follow up on animal research in moving beyond the neuroendocrinology of psychiatric disorders to address instead the neurobiology of diathesis. The importance of this line of research should not be underestimated. First, diathesis research has the potential to answer questions not only about the pathophysiology of psychiatric disorders but also about the pathogenesis of these disorders. Second, as future studies more clearly characterize the neurobiology of diathesis, the possibility of finding not only biological state markers for illness but also trait markers for the risk of illness is within our reach. Such markers would be useful in screening at-risk populations such as abused children, assault victims, combat veterans, and public servants who contend with emergency situations. Finally, improved understanding of the neurobiology of diathesis may ultimately guide prophylactic treatment decisions for those exposed to early life stress before the neurobiological changes that convey the risk for adulthood psychopathology become irreversible.

VII. THE PSYCHOPHARMACOLOGY OF STRESS

A variety of psychopharmacological agents are used to treat patients with stress-related disorders and those vulnerable to the adverse consequences of stress. Foremost among these medications are the anxiolytics and antidepressants. The most widely used anxiety-reducing medications have traditionally been the benzodiazepines, which act primarily by enhancing GABA neurotransmission. GABA is an inhibitory neurotransmitter that reduces neuronal activity. The principal therapeutic benefit of benzodiazepines lies in their calming effect that alleviates anxiety. The effects of a benzodiazepine can be experienced within minutes. For this reason, benzodiazepines can be used acutely on an ''as-needed'' basis to reduce the immediate consequences of stress. In addition, these medications can be used long-term on a prophylactic basis for patients especially prone to stress-induced symptomatology. Unfortunately, benzodiazepines are not without problems. They may cause drowsiness and motor and cognitive impairment, and they may be abused by individuals with substance use disorders. Furthermore, physical dependence can develop with prolonged usage. In addition, benzodiazepines are not effective antidepressants.

Antidepressants are the most widely used agents in the treatment of stress-related psychiatric illness. These medications act primarily by increasing the neuronal transmission of NE and 5-HT, or some combination of the two. Although antidepressants are primarily used to treat depression, they have a wide array of uses. Antidepressants have demonstrated efficacy for certain anxiety disorders (e.g., PTSD, obsessive–compulsive disorder, generalized anxiety disorder, social anxiety disorder, and panic disorder), eating disorders, impulse control disorders, and chronic pain disorders. Remarkably, preclinical animal research indicates that administration of selective serotonin reuptake inhibitors confers some degree of protection from the adverse impact of stress-induced hypercortisolemia on hippocampal neurogenesis. In addition, antidepressants reverse the alterations in HPA axis activity that occur during episodes of depression. This is likely an indirect effect on the HPA axis that is mediated via 5-HT/CRF interactions.

Advances in the treatment of stress-related illness will utilize agents that directly modulate CRF activity. In fact, numerous CRF receptor antagonists are currently under development and testing. These agents appear to possess both potent antidepressant and anxiolytic properties. In addition to alleviating psy-

chiatric illness, CRF antagonists may provide prophylactic benefit for those with a predisposition to stress-related psychopathology.

Prophylactic intervention is of paramount importance for those exposed to the traumatic stress of combat, violent assault, child abuse, natural disaster, etc. If we can respond with appropriate pharmacological interventions in the acute aftermath of a trauma, might we be able to circumvent the pathogenesis of trauma-induced psychiatric illness? Recognizing that CRF and NE systems play a pivotal role in the pathophysiology of stress-related psychiatric illness, these systems are likely targets for treatment. Indeed, limited research on the viability of clonidine and propranolol (agents that interfere with NE) and serotonergic antidepressants (which have been shown to reduce serum levels of NE and MHPG) for such prophylactic benefit is under way. This will doubtlessly be a fruitful area for research with CRF receptor antagonists when they become available.

See Also the Following Articles

ANXIETY • EVOLUTION OF THE BRAIN • HEADACHES • HOMEOSTATIC MECHANISMS • INHIBITION • LIMBIC SYSTEM • NEUROPSYCHOLOGICAL ASSESSMENT • NOREPINEPHRINE • PAIN AND PSYCHOPATHOLOGY • STRESS: HORMONAL AND NEURAL ASPECTS • SUICIDE

Suggested Reading

Bremner, J., Southwick, S., and Charney, D. (1999). The neurobiology of posttraumatic stress disorder: an integration of animal and human research. In *Posttraumatic Stress Disorder: A Comprehensive Text* (P. Saigh and J. Bremner, Eds.), pp. 103–143. Allyn & Bacon, Boston.

Francis, D., and Meaney, M. (1999). Maternal care and the development of stress responses. *Curr. Opin. Neurobiol.* **9,** 128–134.

Francis, D., Caldji, C., Champagne, F., Plotsky, P., and Meaney, M. (1999). The role of corticotropin-releasing factor–norepinephrine systems in mediating the effects of early experience on the development of behavioral and endocrine responses to stress. *Biol. Psychiatr.* **46,** 1153–1166.

Gould, E., and Tanapat, P. (1999). Stress and hippocampal neurogenesis. *Biol. Psychiatr.* **46,** 1472–1479.

Graham, Y., Heim, C., Goodman, S., Miller, A., and Nemeroff, C. (1999). The effects of neonatal stress on brain development: Implications for psychopathology. *Dev. Psychopathol.* **11,** 545–565.

Koob, G. (1999). Corticotropin-releasing factor, norepinephrine, and stress. *Biol. Psychiatr.* **46,** 1167–1180.

Levine, S., and Ursin, H. (1999). What is stress? In *Stress: Neurobiology and Neuroendocrinology* (M. Brown, G. Koob, and C. Rivier, Eds.), pp. 3–21. Dekker, New York.

Nemeroff, C. (1998). Psychopharmacology of affective disorders in the 21st century. *Biol. Psychiatr.* **44,** 517–525.

Newport, D. J., and Nemeroff, C. B. (2000). Recent advances in the neurobiology of posttraumatic stress disorder. *Curr. Opin. Neurobiol.* **10**(2), 211–218.

Newport, D. J., and Nemeroff, C. B. (2001). Hypothalamic–pituitary–adrenal (HPA) axis: Normal physiology and disturbances in depression. In *Physical Consequences of Depression* (J. H. Thakore, Ed.), pp. 1–22. Wrightson Biomedical, Petersfield, UK.

Owens, M., and Nemeroff, C. (1999). Corticotropin-releasing factor antagonists: Therapeutic potential in the treatment of affective disorders. *CNS Drugs* **12,** 85–92.

Paris, J. (1999). *Nature and Nurture in Psychiatry: A Predisposition–Stress Model of Mental Disorders*. American Psychiatric Press, Washington.

Plotsky, P., Owens, M., and Nemeroff, C. (1998). Psychoneuroendocrinology of depression: Hypothalamic–pituitary–adrenal axis. *Psych. Clin. North Am.* **21,** 293–307.

Ressler, K., and Nemeroff, C. (1999). Role of norepinephrine in the pathophysiology and treatment of mood disorders. *Biol. Psychiatr.* **46,** 1219–1233.

Yehuda, R. (1998). Psychoneuroendocrinology of post-traumatic stress disorder. *Psych. Clin. North Amer.* **21,** 359–379.

Stroke

ALLYSON R. ZAZULIA
Washington University, St. Louis

GLOSSARY

autoregulation The physiological process by which blood vessels change caliber to maintain constant cerebral blood flow over a wide range of cerebral perfusion pressures.

infarction Permanent tissue damage and death of all cellular elements (neurons, glia, and vessels) due to prolonged or severe ischemia.

ischemia Impairment of tissue function due to a reduction in blood supply relative to metabolic demand.

neuroprotection Strategies that ameliorate the biochemical and metabolic derangements induced by cerebral ischemia, thus limiting delayed neuronal injury.

penumbra The area surrounding the dense core of irreversibly damaged cells that has preserved ionic homeostasis and reduced neuronal electrical activity but that is capable of recovery.

transient ischemic attack Abrupt focal loss of neurologic function caused by reduction in blood flow that persists less than 24 hr and clears without residual disability.

vasospasm Blood vessel constriction in response to irritative stimuli.

Stroke is the third leading cause of death in the United States and is responsible for more hospitalizations than any other neurological disorder in adults. The term "stroke" encompasses three vascular disorders: (i) cerebral ischemia, which is due to interference with normal brain circulation; (ii) intracerebral hemorrhage, which results from blood vessel rupture within the brain substance; and (iii) subarachnoid hemorrhage, which results from bleeding of surface arteries located in the space between the pia and arachnoid membranes. This article focuses on the pathophysiology, clinical features, and management of each of the stroke categories. Since an understanding of these concepts requires knowledge of normal cerebral circulation and metabolism, cerebrovascular anatomy and physiology is reviewed first.

I. CEREBROVASCULAR ANATOMY

The brain receives its blood supply from two pairs of arteries: the right and left internal carotid arteries, which supply the anterior two-thirds of the corresponding cerebral hemisphere, and the right and left vertebral arteries, which converge to form the basilar artery and supply the brain stem and posterior portion of the hemispheres. The two internal carotid arteries and the basilar artery unite via anastomotic channels at the base of the brain to form the circle of Willis.

The internal carotid artery originates in the neck from the bifurcation of the common carotid arteries into internal and external branches. Each internal carotid artery enters the skull through the ipsilateral foramen lacerum, traverses the petrous bone and cavernous sinus, and gives off a major branch to the eye, the ophthalmic artery, before dividing lateral to the optic chiasm into its two terminal branches, the anterior cerebral artery and the middle cerebral artery (Fig. 1). The anterior cerebral artery passes rostromedially and approaches the corresponding artery of the opposite side with which it connects via the anterior communicating artery. It then gives off branches that supply the inferior surface of the frontal lobe, the

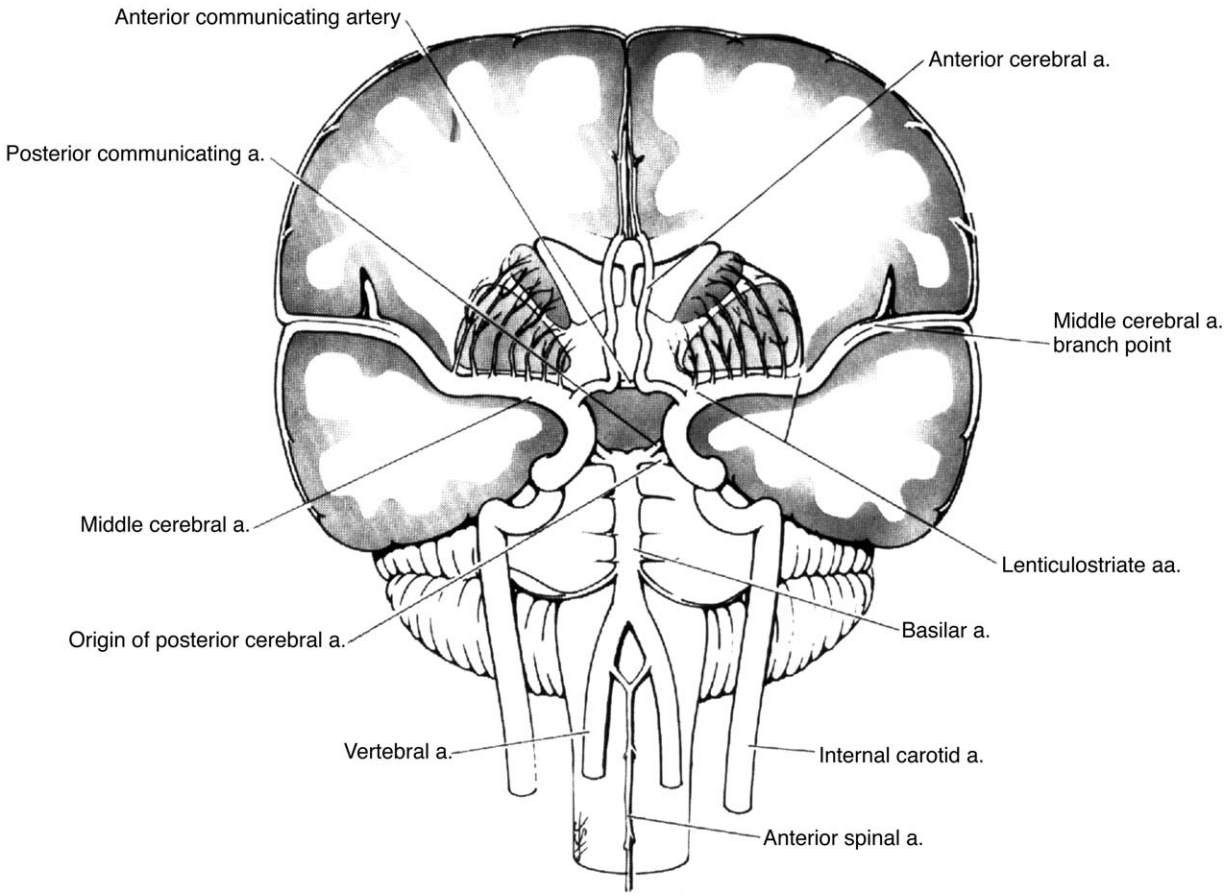


Figure 1 Coronal section through the front of the brain with the anterior portions of the frontal and temporal lobes removed. The major arteries of the anterior (internal carotid, middle, and anterior cerebral arteries) and posterior circulation as well as their interconnecting vessels (anterior and posterior communicating arteries) at the circle of Willis can be seen [from Powers, W. J. (1989). Stroke. In *Neurobiology of Disease* (A. L. Pearlman and R. C. Collins, Eds.), Copyright © 1989 by Oxford University Press, Inc. Used by permission of Oxford University Press, Inc.].

interhemispheric fissure, and the medial surface of the parietal lobe, thus supplying primary motor and sensory cortex for the contralateral lower extremity (Fig. 2). The middle cerebral artery travels laterally, immediately giving off a series of small lenticulostriate arteries, which supply the deep hemispheric gray and white matter that include the basal ganglia and internal capsule. It then continues laterally into the Sylvian fissure, where it divides in a fan-like fashion to carry blood to the lateral surface of the frontal, parietal, and temporal lobes (Fig. 3), thus supplying primary motor and sensory cortex (except for the lower extremity), the speech areas in the dominant hemisphere, and motor, sensory, and visual fiber tracts.

The vertebral arteries arise from the subclavian arteries in the neck, traverse the foramina of the cervical vertebrae, and enter the skull via the foramen magnum. They then course along the anteroleteral surface of the medulla, giving off branches that supply the medulla and inferior cerebellum before joining together at the pontomedullary junction to form the basilar artery (Fig. 1). The basilar artery, which lies on the ventral surface of the brain stem, supplies the pons, cerebellum, and midbrain and then bifurcates into the posterior cerebral arteries, which supply the posterolateral midbrain, thalami, and medial temporal and occipital lobes (Fig. 2).

Three collateral pathways can supply blood to ischemic brain in the event of compromise of one of the major vessels. The most important of these, the circle of Willis, allows for communication between all major vascular trees, thus serving as a potential bypass in the event of vascular compromise in the neck. Significant anatomic variation exists in the circle of

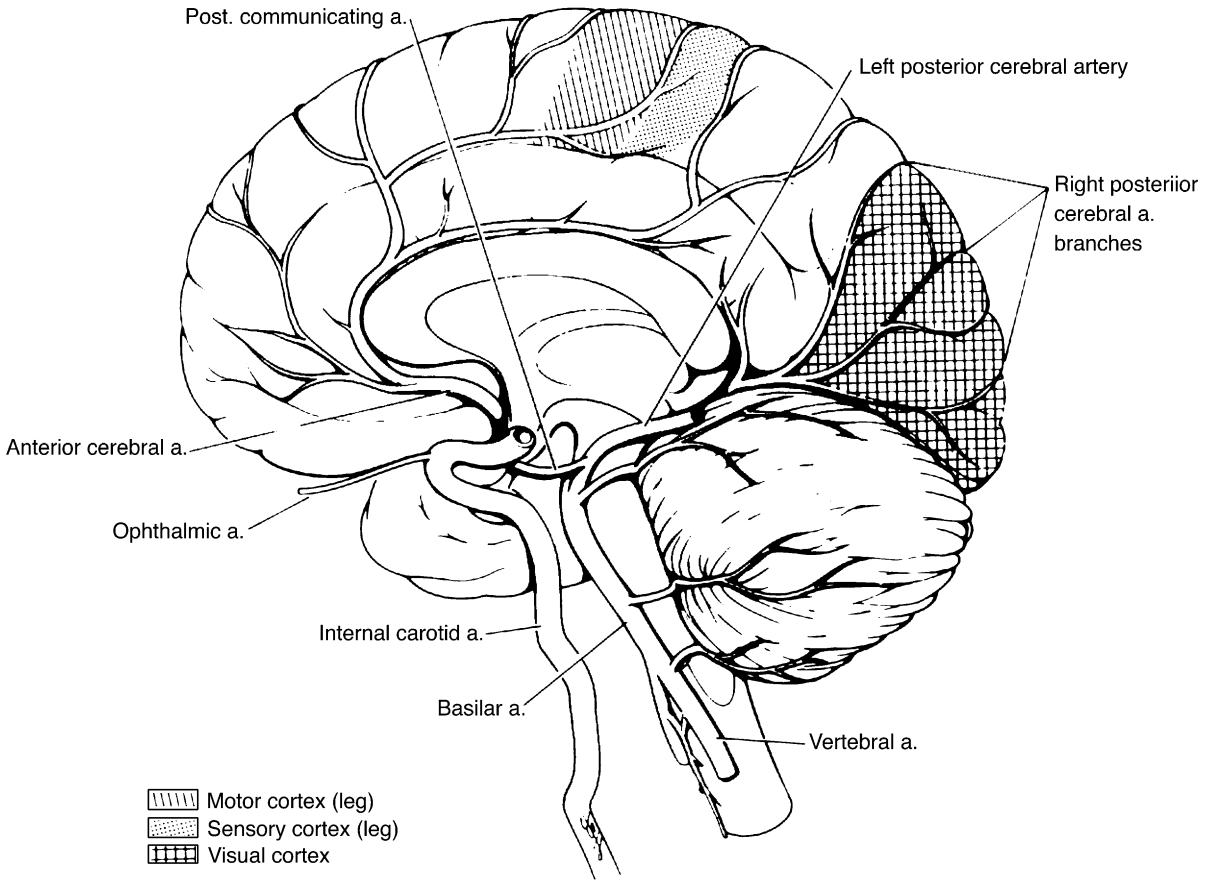


Figure 2 Saggital section through the corpus collosum revealing the medial aspect of the hemisphere. The anterior cerebral artery can be seen coming off the internal carotid artery and running in the interhemispheric fissure along frontal and parietal cortex. The posterior cerebral artery can be seen coming off the top of the basilar artery and traveling posteriorally to the occipital cortex. Also note the circumferential branches of the basilar artery supplying the lateral aspect of the brain stem and cerebellum [from Powers, W. J. (1989). Stroke. In *Neurobiology of Disease* (A. L. Pearlman and R. C. Collins, Eds.), Copyright © 1989 by Oxford University Press, Inc. Used by permission of Oxford University Press, Inc.].

Willis, however, and a communicating artery may be absent or inadequate in up to one-third of people. Another collateral pathway involves communication between the extracranial and intracranial circulations. Anastomoses between branches of the external carotid artery supplying the orbit and branches of the ophthalmic artery link the external carotid artery with the internal carotid artery. Anastomoses between branches of the external carotid artery supplying the posterior scalp muscles and branches of the vertebral artery link the carotid and vertebral systems. Finally, end-to-end arterial anastomoses of the superficial cortical branches of the anterior, middle, and posterior cerebral arteries form the border zone or "watershed" areas between vascular territories (Fig. 3).

The venous drainage of the brain consists of superficial veins draining the surface of the hemispheres into the superior sagittal sinus and deep veins draining the deep structures into the straight sinus. These sinuses then converge at the torcular, where blood flows into the right and left lateral sinuses and finally through the jugular foramina to form the internal jugular veins in the neck (Fig. 4). Collateral pathways exist both within and between the superficial and deep systems.

II. CEREBROVASCULAR PHYSIOLOGY

A. Cerebral Blood Flow and Metabolism

Under normal resting conditions, the adult human brain receives about 20% of the cardiac output, resulting in an average cerebral blood flow (CBF) of $50\,\mathrm{ml}\,100\,\mathrm{g}^{-1}\,\mathrm{min}^{-1}$. Flow to gray matter is

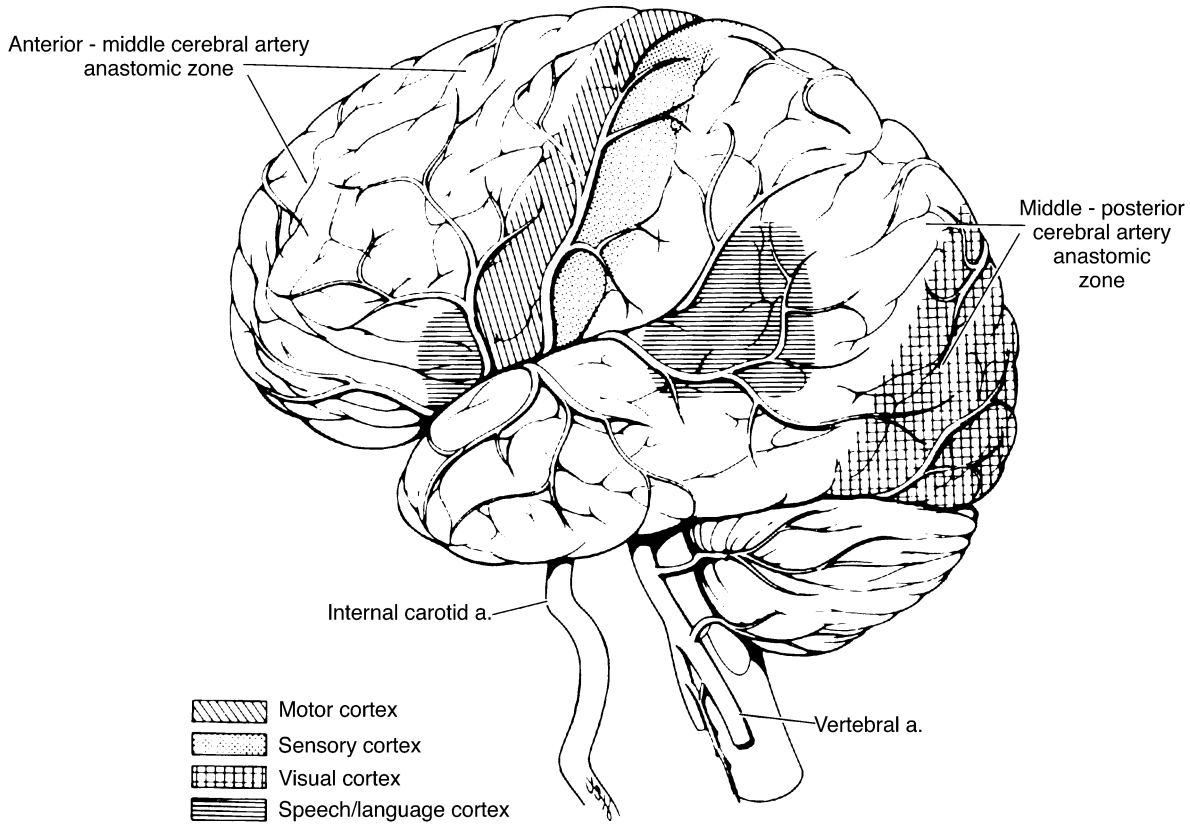


Figure 3 Lateral view of the brain showing the middle cerebral artery emerging from the Sylvian fissure and supplying branches to the lateral aspects of the frontal, parietal, and temporal lobes. Also shown are the watershed or border zone areas between the anterior and middle cerebral arteries and between the middle and posterior cerebral arteries [from Powers, W. J. (1989). Stroke. In *Neurobiology of Disease* (A. L. Pearlman and R. C. Collins, Eds.), Copyright © 1989 by Oxford University Press, Inc. Used by permission of Oxford University Press, Inc.].

approximately four times that of white matter, reflecting the higher metabolic rate and oxygen requirement of neurons. The brain relies almost entirely on oxidative metabolism of glucose for its energy needs in the resting state. Since it stores very little glucose, the brain is critically dependent on a continuous supply of blood-borne glucose and oxygen to maintain its functional integrity. The resting brain consumes approximately $3.5\,\text{ml}\,100\,\text{g}^{-1}\,\text{min}^{-1}$ of oxygen, which represents about one-forth of total body consumption, and approximately $5\,\mu\text{g}\,100\,\text{g}^{-1}\,\text{min}^{-1}$ of glucose, 90% of which is metabolized to carbon dioxide and water.

CBF is regulated by the relationship between cerebral perfusion pressure (CPP) and cerebrovascular resistance (CVR) by the equation

$$\text{CBF} = \frac{\text{CPP}}{\text{CVR}}$$

CPP represents the difference between arterial pressure forcing blood into the cerebral circulation and venous backpressure, which is mainly dependent on intracranial pressure (ICP). In normal circumstances, venous pressure is negligible, and CPP is equal to arterial pressure. CVR is determined primarily by vessel caliber and is defined by the equation

$$\text{CVR} = \frac{L\eta}{\pi r^4}$$

where L is vessel length, η is blood viscosity, and r is vessel radius. It bears emphasis that CVR is related to the fourth power of vessel radius, such that even small changes in luminal caliber can have significant effects on CVR; thus, in normal circumstances, vessel caliber is the most important determinant of blood flow. Factors that regulate vessel caliber include arterial carbon dioxide tension (P_aCO_2), arterial oxygen content (C_aO_2), perivascular ionic concentrations,

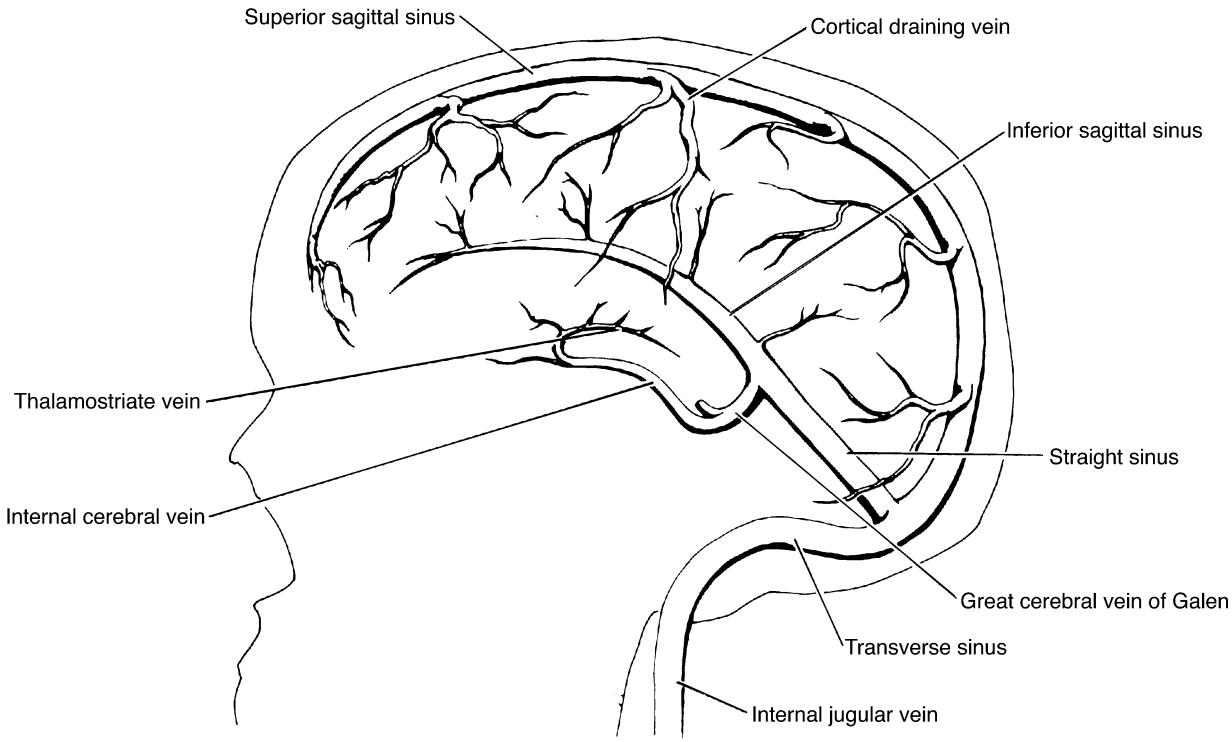


Figure 4 Sagittal section through the brain demonstrating the superficial and deep venous systems. The hemispheres drain superiorly into the superior sagittal sinus. The deep structures drain into the straight sinus. The superior sagittal and straight sinuses converge to form the two lateral sinuses, which ultimately drain into the internal jugular veins [From Powers, W. J. (1989). Stroke. In *Neurobiology of Disease* (A. L. Pearlman and R. C. Collins, Eds.), Copyright © 1989 by Oxford University Press, Inc. Used by permission of Oxford University Press, Inc.].

and innervation by sympathetic, parasympathetic, and trigeminal nerves. In addition, a host of diverse mediators that alter vessel tone and function, including vasoactive peptides, amines, lipids, phospholipids, and vasoactive gases, have been identified. Some of these (e.g., nitric oxide) are suspected to play a role in ischemia.

Carbon dioxide exerts a profound influence on cerebral vasculature, with hypercapnia (increased P_aCO_2) producing pronounced vasodilation through preferential effects on small cerebral arterioles, and hypocapnia (reduced P_aCO_2) producing vasoconstriction of all cerebral vessels. Inhalation of 5% CO_2 increases CBF by approximately 50%, and reducing P_aCO_2 from 45 to 25 mmHg lowers CBF by approximately 35%. Prolonged hypo- or hypercapnia alters CBF responsiveness to acute changes in P_aCO_2, such that CBF eventually returns toward normal despite continuous exposure. The effect of CO_2 on CBF appears to be primarily mediated by the hydrogen ion concentration of extracellular fluid, but prostaglandins, nitric oxide, and neural pathways may also play a role.

C_aO_2, which is dependent on oxygen tension (P_aO_2) and hematocrit, is another important determinant of CVR, with hypoxia (reduced P_aO_2) and anemia (reduced hematocrit) inducing vasodilation and increased blood flow. Cerebral vasculature is not as sensitive to changes in P_aO_2 as it is to changes in P_aCO_2 because of the shape of the oxygen–hemoglobin dissociation curve. Nevertheless, when P_aO_2 decreases below a threshold of approximately 50 mmHg, CBF begins to increase exponentially. The mechanisms by which hypoxia alters CVR include a direct effect on vessel relaxation, stimulation of oxygen-sensitive neurons in the ventral medulla, and possibly through accumulation of local metabolites. Prolonged hypoxia induces compensatory mechanisms, such that CBF eventually returns to normal.

Cerebral blood vessels receive innervation from sympathetic, parasympathetic, and trigeminal (serotonergic) nerve fibers, but the physiological role

of these systems is unclear. They appear to have minimal effect on CBF in normal circumstances but may serve a protective role at extremes of blood pressure (sympathetic system) or in response to ischemia or subarachnoid hemorrhage-induced vasospasm (parasympathetic and serotonergic systems).

B. Autoregulation

Autoregulation of blood flow is the well-developed mechanism whereby cerebral resistance arteries dilate during reductions in CPP and constrict during increases in CPP. As a result, blood flow to the brain remains constant over a wide range of pressures. The normal range for autoregulation is approximately 60–150 mmHg, but is higher in chronically hypertensive subjects.

Global reductions in CPP may be caused by decreased systemic arterial pressure or increased ICP. Regional reductions may occur in arterial occlusive disease or venous thrombosis. As CPP decreases, vasodilation maintains CBF until the capacity for autoregulation has been exceeded. At that point, the brain begins to extract increasing amounts of oxygen and glucose to maintain normal metabolism and brain function. It is not until oxygen extraction fraction (OEF) has been maximized that further decreases in CBF result in reduced metabolism and clinical symptoms of ischemia (Fig. 5). This occurs at a CBF of approximately 15–20 ml $100\,g^{-1}\,min^{-1}$.

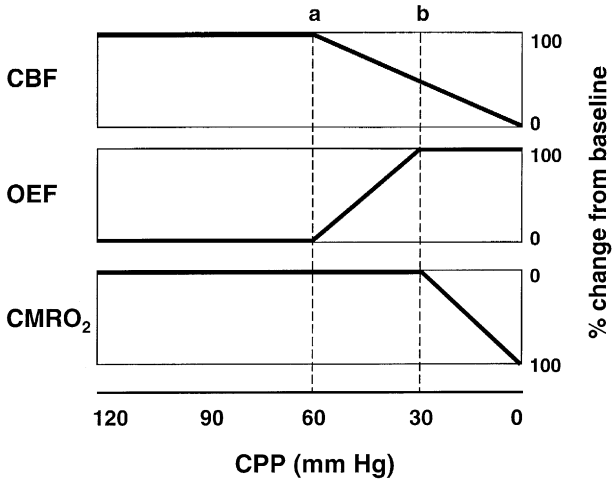


Figure 5 Compensatory responses to reduced cerebral perfusion pressure (CPP). As CPP decreases, cerebral blood flow (CBF) is initially maintained by dilation of resistance arteries (not shown). When vasodilation can no longer compensate, autoregulation fails and CBF begins to decrease (a). A progressive increase in the oxygen extraction fraction (OEF) now maintains cerebral oxygen metabolism ($CMRO_2$). Once OEF is maximal (b), further declines in CBF disrupt cellular metabolism and function (reproduced with permission from Powers, W. J., Cerebral hemodynamics in ischemic cerebrovascular disease. *Ann. Neurol.* **29,** 231–240, 1991).

III. CEREBRAL ISCHEMIA/INFARCTION

A. Pathophysiology

Based on the previous discussion, ischemia can be defined as an impairment of tissue function due to a reduction in blood supply relative to metabolic demand. Experimental models have provided a detailed description of the changes in CBF and metabolism that occur during focal ischemia. Occlusion of a vessel produces a severe reduction in blood flow to a central core of tissue and a less severe reduction in blood flow to a zone surrounding this core, where perfusion is maintained by collateral circulation. As long as CBF remains above approximately 20 ml $100\,g^{-1}\,min^{-1}$, no changes in brain metabolism or function occur. Below this level, brain electrical activity fails, and neurological symptoms appear. As the supply of oxygen becomes insufficient to maintain normal cellular biochemistry, stores of high-energy phosphates are quickly depleted. Anaerobic metabolism of the small amount of residual glucose in intracellular stores results in lactic acidosis. With more severe reductions in CBF to approximately 10–12 ml $100\,g^{-1}\,min^{-1}$, cell membrane integrity is lost and depolarization causes leakage of intracellular K^+ into the extracellular space and influx of extracellular Ca^{2+} into neurons. Spreading depression, a depolarization wave running over the brain cortex at a rate of 3 or 4 mm/min, is elicited by increases in extracellular K^+ and leads to increased energy consumption, as the Na^+–K^+ pump is activated in an attempt to restore intracellular K^+. Since the resultant increased substrate demand cannot be compensated for by increases in CBF, further ischemia occurs.

Surrounding the core of dense ischemia is the ischemic penumbra, in which blood flow is reduced to the point of interrupting neuronal electrical activity but ionic homeostasis is preserved. If blood flow is restored, neurons in the penumbra can be salvaged. If

reperfusion does not occur, the neurons usually die. Therefore, it is the fate of the neurons in the ischemic penumbra that often determines the final size of an infarct.

The ability of neurons to survive a period of transient cerebral ischemia depends on both severity and duration and varies in different brain regions, with cerebellar Purkinje cells, hippocampal CA1 cells, and cells in cortical layers 3, 5, and 6 being particularly vulnerable. Once a region's "ischemic threshold" is reached, cells become irreversibly damaged. The finding that in some cells death is immediate whereas in others it is delayed until after restoration of perfusion suggests that energy failure and loss of ionic homeostasis do not inevitably lead to cell death. Delayed cell death occurs as a result of a complex series of secondary events that are triggered by the ionic derangements, including release of excitotoxic neurotransmitters, calcium influx, activation of proteases, free radical production, programmed cell death (apoptosis), acidosis, and inflammatory mechanisms.

B. Causes/Risk Factors

The specific causes of cerebral ischemia can be divided into two general categories: those that cause global ischemia (a decrease in blood flow to the entire brain) and those that cause focal ischemia (a decrease in blood flow to a specific area within the brain).

Global ischemia occurs when decreased cardiac output (e.g., secondary to cardiac arrest or hypovolemic shock) or decreased peripheral vascular resistance (e.g., associated with vasovagal syncope or septic shock) lead to a reduction in systemic blood pressure. The most common sign is loss of consciousness. Only rarely does hypotension produce a focal deficit, even in the presence of a stenotic artery. If ischemia is moderate, damage is often restricted to the watershed or border zone areas between the major vascular territories. If severe or prolonged, it can result in infarction of the entire brain.

Focal ischemia occurs as a result of a local vascular disturbance. Causes of focal ischemia can be divided into three general categories: primary disease of the cerebral blood vessels, embolism from extracerebral sites, and hematological disorders (Table I). Regardless of the underlying etiology, the final common pathway is vascular compromise resulting in inadequate blood supply to a region of the brain. The primary vascular diseases and the hematological disorders produce stroke through thrombosis of cerebral blood vessels. Thrombosis is initiated by alterations in endothelial integrity and disruptions in the normal laminar flow of blood, with activation of the coagulation cascade and subsequent production of

Table I
Causes of Focal Cerebral Ischemia

Vascular disease
Atherosclerosis
Cerebral autosomal-dominant arteriopathy with subcortical infarcts and leukoencephalopathy
Connective tissue disease
Drugs (cocaine, sympathomimetics)
Fibromuscular dysplasia
Infective arteritis
Lipohyalinosis/fibrinoid necrosis
Metabolic diseases
Migraine
Moya-moya disease
Radiation vasculopathy
Sickle cell anemia vasculopathy
Vascular dissection
Vascular trauma
Vasculitis
Vasospasm
Venous thrombosis
Systemic embolism
Air embolism
Aortic arch atheroma
Cardiac thrombus (atrial fibrillation, atrial myxoma, dilated cardiomyopathy, myocardial infarction, ventricular aneurysm)
Fat embolism
Infective endocarditis
Paradoxical embolism
Tumor embolism
Valvular heart disease
Hematological disorders
Hypercoagulable states (antiphospholipid antibody syndrome, antithrombin III deficiency, disseminated intravascular coagulation, factor V Leiden mutation, heparin cofactor II deficiency, paraneoplastic, paroxysmal nocturnal hemoglobinuria, protein C deficiency, protein S deficiency, thrombotic thrombocytopenic purpura)
Hyperviscosity syndromes (dysproteinemia, myeloproliferative syndromes, polycythemia vera, thrombocytosis)
Other
Nephrotic syndrome
Oral contraceptives
Pregnancy/puerperium

thrombin. Thrombin, the key regulator of thrombosis, activates platelet aggregation and cleaves fibrinogen to fibrin, which then cross-links to form the basis for the clot matrix. Thrombi can either occlude the vessel at the site of endothelial injury or embolize distally through a combination of shear stress and the action of intrinsic antithrombotic mechanisms. With time, thromboemboli to the cerebral vessels spontaneously lyse, but this often occurs too late to prevent permanent infarction.

The most common cause of focal ischemia is atherosclerosis. Atherosclerosis is a multifactorial process by which plaque (''atheroma'') containing cholesterol, lipid-laden macrophages, inflammatory cells, abnormally proliferated smooth muscle cells, and dense connective tissue forms within arterial walls. Endothelial injury is one of the key events in the pathogenesis of atherosclerosis since it leads to fatty streak formation and contributes to smooth muscle cell proliferation. Stroke can occur if the atheroma produces sufficient arterial stenosis and/or promotes sufficient thrombus formation to reduce blood flow distally. More commonly, thrombus formed at the site of atheromatous plaque serves as a source for emboli that are carried distally where they occlude smaller intracranial vessels (artery-to-artery embolism). The factors that determine when atherosclerotic plaque will produce symptomatic ischemia are poorly understood but may include rupture of fragile plaque vessels resulting in hemorrhage, calcification of the fibrous surface leading to plaque rupture, and plaque ulceration. These plaque complications, which are related to local hemodynamic forces and increasing degree of luminal narrowing, expose the highly thrombogenic contents of the necrotic core and are therefore likely the basis for atherosclerosis-induced ischemia. Atherosclerosis has a predilection for bifurcation points and curves in vessels. Within the cerebrovascular system, it occurs most commonly at the origin of the internal carotid arteries in the neck, at the carotid siphon, in the distal vertebral arteries, and in the basilar artery. Its prevalence increases with age, and it affects men at a younger age than women. In addition, there are racial differences, with extracranial and large artery disease being more common in whites than in blacks and Asians. Hypertensive patients often have more advanced atherosclerotic disease for age as well as disease extension into more distal vessels. Primary disease of the small penetrating arteries and arterioles is most commonly related to the hypertension-associated conditions termed arteriolosclerosis, in which microatheroma form in the same manner as the atheroma affecting larger arteries, and lipohyalinosis, in which a fibrohyaline substance accumulates in the vessel wall. These conditions are the most frequent causes of the small infarcts (''lacunes'') that predominantly occur in the basal ganglia and white matter of the internal capsule and pons.

Among the systemic causes of cerebral embolism, the most clinically important are those arising in the heart since cardiac embolism is implicated in 15–25% of ischemic strokes. Embolic material is usually composed of fibrin and/or platelet thrombi, but it may also include debris from calcified or degenerated heart valves, fragments of heart valve vegetations, tumor cells, fat, air, or other foreign substances.

Primary hematological disorders can predispose to ischemia by increasing thrombotic potential (hypercoagulable states) or decreasing blood flow (hyperviscosity syndromes), but they underlie only a small minority of ischemic strokes. Hypercoagulable states are more convincingly related to venous thrombosis than arterial, and patients harboring these defects will often have a history of peripheral venous thrombosis, pulmonary embolism, or recurrent fetal loss.

Prospective epidemiological studies have identified a number of risk factors for stroke. Nonmodifiable risk factors include increasing age, black race, male gender, and family history of stroke. More important from the standpoint of stroke control, several modifiable risk factors have been established. Hypertension is the most powerful of these since it occurs in 25–40% of the U.S. population and increases stroke risk approximately three-fold. Even borderline hypertension (systolic 140–159 mmHg and diastolic 90–94 mmHg) is associated with a 50% increase in stroke risk. Pooled data from 14 treatment trials of hypertension indicate that a blood pressure reduction of 5.8 mmHg corresponds to a 42% reduction in stroke incidence. Diabetes confers a 1.8–3 times increased stroke risk, and increasing degrees of glucose intolerance correspond to increasing risk of stroke. Cardiac disease is a frequent comorbid factor in stroke by virtue of shared risk factors, but it also predisposes to stroke by serving as a source for embolism. Atrial fibrillation is associated with as much as a five-fold increase in stroke incidence depending on associated factors. In both primary and secondary prevention trials, anticoagulation treatment of nonvalvular atrial fibrillation reduces stroke risk by approximately 70%. In contrast to the known impact of blood lipids on coronary heart disease, the relationship between lipids and stroke is unclear, with inconsistent results among

epidemiological studies. Serum lipid levels have been directly related to atherosclerosis of extracranial and intracranial vessels, however, with low-density lipoprotein and total cholesterol increasing vessel wall thickness and high-density lipoprotein cholesterol protecting against the development of atheroma. In addition, clinical studies of the cholesterol-lowering HMG-CoA reductase inhibitors (statins) in patients with coronary artery disease have demonstrated a significant reduction in ischemic stroke. The mechanism by which statins exert this neuroprotective effect is unclear. Cigarette smoking is a significant independent risk factor for stroke. Risk increases with the number of cigarettes smoked per day, and inferences from observational studies suggest that cessation of smoking results in a decrease in stroke risk within 5 years to nearly that of persons who have never smoked. Since cigarette smoking is correlated with degree of carotid stenosis, the likely mechanism by which smoking increases stroke risk is through acceleration of atherosclerosis. The relationship between alcohol use and stroke appears to be U shaped: Total abstinence and excessive consumption are associated with the greatest risk, and moderate use is associated with the lowest risk. Elevated plasma homocysteine levels, which are inversely related to intake of folic acid and vitamins B_{12} and B_6, have been demonstrated to increase stroke risk in some studies but not in others. Other factors cited as contributors to stroke risk include elevated serum fibrinogen, low level of physical activity, and oral contraceptive use (especially in older women who smoke or suffer from migraine).

C. Clinical Features

The clinical hallmark of cerebral ischemia is the sudden onset of a focal neurological deficit. The onset is usually abrupt and the deficit maximal within minutes. When progression does occur, it too tends to be abrupt and stepwise, rather than a slow, steady decline. Other causes of sudden focal neurological deficit, including seizure, migraine, and trauma, should always be considered. The deficits caused by a stroke reflect the region of the brain (or retina) involved, thus allowing for localization of the lesion by neurological examination.

Lesions in the carotid circulation produce motor and/or sensory deficits on the opposite side of the body. Language dysfunction may occur with ischemia of the dominant (almost always left) hemisphere and impaired spatial perception with ischemia of the nondominant hemisphere. If deep white matter is involved, dysfunction generally affects face, arm, and leg and may be accompanied by visual field defects. Superficial lesions in the territory of the middle cerebral artery generally spare the leg, whereas those in the territory of the anterior cerebral artery primarily affect the leg. Retinal ischemia due to a lesion of the ophthalmic artery causes transient monocular blindness (amaurosis fugax).

Ischemia within the vertebrobasilar system characteristically produces cranial nerve or cerebellar dysfunction in addition to motor and/or sensory deficits affecting either the opposite or both sides of the body. Isolated unilateral motor deficits can occur with small lesions of the brain stem, creating diagnostic confusion with lesions of the carotid system, but more often neighboring signs suggest the correct localization. Involvement of the posterior cerebral artery causes visual symptoms due to ischemia of the occipital lobe. Although the posterior cerebral artery also supplies the inferior temporal lobe regions involved in memory, whether or not ischemia of this artery causes memory disturbance (e.g., transient global amnesia) is controversial. Ischemia in the territory of the vertebral and basilar branches to the cerebellum often produces ataxia. Thrombosis of the basilar artery may produce bilateral infarction of the brain stem or thalamus, resulting in coma or a "locked-in" state characterized by complete loss of speech and quadriplegia with retained consciousness.

The temporal course of ischemia is variable. Ischemic deficits may resolve completely, remain stable, or worsen. If complete resolution occurs within 24 hr, the event is called a transient ischemic attack (TIA). In actuality, most TIAs last less than 20 min. Persistence of deficits beyond 24 hr is labeled a completed stroke or infarction. The terms "progressing stroke" and "stroke in evolution" have been applied when clinical deficits are observed to worsen (increased severity of deficit or development of new deficits), whether this is a gradual decline or a stuttering progression over minutes to hours. The lack of uniformity to the definition of progressing stroke precludes meaningful comparisons between reports and makes interpretation of incidence and prognosis unsatisfactory. In addition, numerous mechanisms may underlie or mimic progressing stroke. Thrombus enlargement or distal progression along a vessel may extend the area of damage either by converting a partial occlusion to a complete occlusion or by obliterating connections to collateral circulation. Recurrent embolization may occur. With

clot lysis or upon establishment of reperfusion by collateral vessels, hemorrhage into the infarct may occur. Cerebral edema, which is maximal 3–5 days after stroke onset, may cause clinical deterioration and death due to herniation. Systemic factors such as fever, infection, electrolyte abnormalities, acid–base disturbances, and altered cardiopulmonary function may interfere with cerebral metabolism, leading to worsening of neurological deficits. Finally, stroke progression may occur as the ongoing cascade of biochemical events induced by ischemia leads to delayed cell death.

1. Diagnostic Studies

Diagnostic studies, including neuroimaging, analysis of blood and cerebrospinal fluid (CSF), and cardiac testing, are used to distinguish stroke from nonstroke entities, to exclude the presence of cerebral hemorrhage, to assist in localization of lesions when exam findings are ambiguous, and to help establish etiology and appropriate treatment. The neuroimaging manifestations of cerebral ischemia vary significantly with time. Within the first 24 hr after symptom onset, computed tomography (CT) can detect only subtle signs of infarction, such as blurring of the usually distinct border between gray and white matter or loss of sulci due to early edema. After 24–48 hr, most infarcts are visible as wedge-shaped areas of decreased attenuation that involve both gray and white matter in a typical vascular distribution. Increasing mass effect and edema become evident over the next few days. By the end of the first week, contrast enhancement can often be seen in a gyral pattern and may persist for months. Months to years later, there is encephalomalacic change and volume loss in the area of the infarct. Magnetic resonance imaging (MRI) of the brain is more sensitive to ischemic changes. Most infarcts can be recognized within 24 hr as bright signals within a vascular territory on conventional MRI sequences sensitive to water content (T2 and FLAIR). During the next week, intense parenchymal contrast enhancement develops. Edema and mass effect become prominent, then decrease during the second week. Encephalomalacia is seen in the chronic stage. Increasing MRI technology has allowed for even earlier detection of ischemia. Diffusion-weighted MRI can detect restricted water movement in brain tissue within hours of ischemia, but this finding is not specific. Initially thought to indicate permanent neuronal damage, reduced diffusion has been reported to occur in TIA, migraine, and seizure without the expected evolution of abnormalities typical of infarction on conventional MRI sequences. Attempts to predict final infarct volume using diffusion imaging in conjunction with perfusion imaging, an MRI technique that provides qualitative information on CBF, remain experimental. Although rarely performed in acute ischemia, cerebral angiography demonstrates vessel occlusion appropriate to neurological symptoms in more than 50% of patients studied within a few hours of stroke. Approximately half of the occluded vessels spontaneously recanalize within 1 week. Noninvasive techniques are available to assess for disease of the extracranial and intracranial vessels (carotid and vertebral Doppler sonography, transcranial Doppler, magnetic resonance angiography, and CT angiography), but conventional angiography remains the gold standard.

2. Prognosis

The possible outcomes in patients with recurrent TIA include spontaneous cessation of attacks, continued attacks, or cerebral infarction. The risk of stroke within the first year after TIA is about 10%, declining to 5% per year thereafter. Twenty to 50% of patients with an untreated TIA who have a subsequent stroke will do so during the first month. The presence of ipsilateral carotid stenosis $>70\%$ increases the risk of subsequent stroke to 15% per year. Retinal TIA (transient monocular blindness) carries a lower stroke risk than hemispheric TIA.

Mortality from stroke has declined dramatically during the past 30 years, variably attributed to reduced stroke incidence, declining stroke severity, improved sensitivity of diagnosing smaller infarcts with milder deficits, and improved stroke survival. Approximately 5–15% of patients die within the first 30 days of ischemic stroke, most commonly as a result of cardiopulmonary complications or brain death. Neurological deficits due to stroke tend to improve over time, with most recovery of function occurring in the first few months. The mechanisms underlying this recovery are poorly understood. No large-scale regeneration of infarcted tissue occurs, but undamaged brain may assume lost function or facilitate function of any remaining neurons. However, about one-third of stroke survivors will not return to independent function.

The longer term prognosis after completed stroke depends on the epidemiological factors that originally predisposed to the stroke. The 5-year mortality rate is approximately 50%, with one-third of deaths due to initial or recurrent stroke, one-third due to cardiac

disease, and one-third due to other causes. The risk of recurrent stroke is about 5–10% per year, but there is no universal agreement on the factors that predict this recurrence.

D. Treatment

Treatment of ischemic stroke aims to limit the extent of infarction and reduce disability by (i) augmenting fibrinolysis (thrombolytic therapy), (ii) salvaging neurons in the ischemic penumbra (neuroprotective agents), and (iii) preventing/treating stroke-related complications. The role of antiplatelet and anticoagulant agents in preventing thrombus progression in acute stroke is limited. Another major goal in the treatment of patients with ischemic stroke as well as those with TIA is to reduce the risk of subsequent stroke. This is accomplished via (i) antiplatelet or anticoagulant agents to decrease thrombotic potential, (ii) surgical treatment of extracranial carotid stenosis, and (iii) risk factor modification.

1. Acute Stroke Therapy

a. Thrombolytic Agents Since the extent of stroke-induced brain injury is related to the duration and severity of ischemia, agents that produce clot lysis and restore cerebral perfusion may limit the degree of injury and thus improve outcome. Intravenous thrombolytic therapy has been reported to achieve some degree of angiographic clot lysis in 30–60% of patients. In the mid-1990s, three trials of intravenous streptokinase in acute ischemic stroke were stopped because of excessive intracranial hemorrhage. A concurrent trial of tissue plasminogen activator (tPA) yielded positive results, leading to the 1996 approval by the U.S. Food and Drug Administration of tPA in acute ischemic stroke. When administered within 3 hr of symptom onset, tPA was associated with a 30% greater likelihood of excellent recovery at 3 months despite an increase in early symptomatic intracranial hemorrhage. Three randomized trials in which the treatment window was extended failed to demonstrate benefit of tPA when given after 3 hr.

Thrombolytic agents can be infused directly into occlusive thrombi via microcatheters, with the potential advantages of higher recanalization rates and improved safety (because of lower drug dose required) but the disadvantages of limited availability of interventional neuroradiology teams and delays in drug administration. One recent randomized open-label trial using pro-urokinase within 6 hr of symptom onset in middle cerebral artery occlusion showed marginally significant benefit, but because of methodological problems and small sample size, a larger trial is warranted before the drug can be recommended for clinical use.

b. Neuroprotective Agents The goal of neuroprotective therapy is to prevent cellular death in the ischemic penumbra, thus limiting infarct size and improving outcome. The most important factor in successful neuroprotection appears to be time because ischemia becomes irreversible within minutes to hours of onset. Thus, many agents that have shown marked efficacy when administered before or immediately after stroke induction in experimental models have failed to demonstrate any benefit in subsequent clinical trials. Potential neuroprotective strategies have targeted various steps in the ischemic cascade, but no neuroprotective drug has successfully completed phase III clinical efficacy trials.

c. Antiplatelet and Anticoagulant Therapy Although aspirin has a clear role in the treatment of acute myocardial infarction, it is of limited benefit in acute stroke. A recent randomized controlled trial of aspirin given within 48 hr of stroke onset revealed a modest reduction in early mortality (9 per 1000) among the aspirin-treated patients. Similarly, heparin (a parenterally administered polysaccharide that inhibits thrombin activation) prevents progression of thrombosis in experimental models and in patients with acute myocardial infarction, but six randomized trials of immediate anticoagulation with heparin or its related low-molecular-weight compounds after stroke revealed no consistent benefit in any stroke subgroup. Although heparin may reduce the risk of recurrent stroke, this benefit is offset by an increase in hemorrhagic stroke and extracranial hemorrhage.

d. Prevention/Treatment of Stroke-Related Complications Stroke-related complications and comorbid disease are a major cause of morbidity and mortality among stroke patients. Hypertension commonly occurs after stroke. Even in patients without a previous history of hypertension, blood pressure is often elevated acutely and typically returns to baseline spontaneously over the first week. The acute treatment of hypertension is controversial because of concerns that impaired autoregulation in the periinfarct area will result in further reduction of CBF if blood pressure is lowered. Most patients have coexisting cardiac disease, and concomitant cardiac ischemia or

arrhythmia occur frequently. Deep venous thrombosis occurs in up to 75% of hemiplegic patients in the first week after stroke and may lead to pulmonary embolism and death. The risk of venous thrombosis is greatly reduced by the use of low-dose subcutaneous heparin, elastic stockings and pneumatic compression devices on the lower extremities, and early mobility. Pulmonary function may be compromised by aspiration due to impaired cough reflex or hospital-acquired pneumonia. Dysphagia may impair adequate nutrition and hydration, so intravenous fluids, nasogastric tube feeding, or diet modifications may be necessary. Urinary retention and incontinence may require bladder catheterization. If the patient is unable to roll over in bed, pressure sores are a risk. Poststroke depression is common and usually responds to antidepressant medications.

2. Secondary Stroke Prevention

a. Antiplatelet Agents Platelets contribute to thrombosis as catalysts for thrombin generation and as constituents of thrombi. Antiplatelet agents inhibit platelet function by preventing platelet aggregation and adhesion to activated endothelium. Aspirin, an irreversible inhibitor of the platelet enzyme cyclooxygenase, is the best studied of the antiplatelet agents and remains the most commonly prescribed medication for stroke prophylaxis. Aspirin has been shown to reduce stroke risk among patients with recent TIA or stroke by 15–58%. Dipyridamole, a cAMP phosphodiesterase inhibitor, has been shown to reduce stroke risk by approximately the same degree as low doses of aspirin. The combination of dipyridamole and low-dose aspirin provides an additive effect, reducing stroke risk by 37%. Ticlopidine, an irreversible inhibitor of adenosine diphosphate-induced platelet aggregation, has been demonstrated to be slightly superior to aspirin in secondary stroke prevention, reducing the yearly stroke rate from 4.3% with aspirin to 3.3% with ticlopidine. This benefit is offset by a higher incidence of side effects. Clopidogrel, an agent chemically related to ticlopidine but with a lower incidence of side effects, is not superior to aspirin, but it remains a treatment option for aspirin-intolerant individuals.

b. Anticoagulants The clinical efficacy of anticoagulants has been well established for a number of systemic disorders, including prevention of venous thromboembolism and prevention of myocardial infarction. Despite having long been advocated for stroke prevention, the effectiveness of anticoagulants in this regard has been demonstrated only in select cardiac diseases. Long-term treatment with warfarin, an oral anticoagulant that antagonizes vitamin K, has been proven to reduce stroke risk in patients with nonvalvular atrial fibrillation (70% risk reduction) and in those with mechanical prosthetic cardiac valves. Because of a high risk of embolism, several other cardiogenic sources for stroke are often treated with warfarin, including mitral stenosis, recent myocardial infarction, left ventricular thrombus, and dilated cardiomyopathy. The value of oral anticoagulation in atherosclerotic stroke is currently being evaluated in several randomized trials, but one recent multi-center randomized trial in patients with non-cardioembolic stroke failed to show any benefit of warfarin over aspirin in any subgroup of patients.

c. Carotid Endarterectomy Clinical trials have established that carotid endarterectomy is the treatment of choice for patients with retinal or hemispheric TIA or stroke accompanied by severe stenosis of the appropriate carotid artery, as long as the perioperative complication rate is low. Such patients with symptomatic carotid stenosis of 70–99% have a 17% absolute risk reduction of ipsilateral stroke when treated with carotid endarterectomy and medical treatment compared to medical treatment only. This benefit is reduced or lost if surgical morbidity and mortality increase above 6%. In patients with symptomatic moderate carotid stenosis (50–69%), the benefit of endarterectomy is less clear-cut, with an absolute risk reduction of ipsilateral stroke of only 6.5%. Therefore, decisions about surgical treatment in this category must consider the patient's risk factor profile and the surgeon's complication rate. Patients with stenosis of less than 50% do not obtain any benefit from surgery.

The value of endarterectomy in asymptomatic carotid stenosis has also been evaluated. Among patients with $>60\%$ carotid stenosis who have not had an ipsilateral stroke or TIA, endarterectomy reduces the risk of subsequent stroke from about 2% per year to 1% per year. The risk reduction appears to be much less for women, however. In order for surgery to be beneficial, patients must have a reasonable life expectancy, and operative morbidity and mortality must be $<3\%$.

d. Angioplasty and Stenting Endovascular procedures, including percutaneous transluminal angioplasty and arterial stenting, are sometimes considered for surgically inaccessible stenotic lesions, restenosis after carotid endarterectomy, inflammatory vascular

disease, and for patients who are hemodynamically unstable or poor surgical candidates. Although case series have reported good results, the efficacy of these treatments for stroke prevention remains to be determined by randomized, controlled trials.

E. Cerebral Venous Thrombosis

Obstruction of cerebral venous drainage represents a special category of cerebral infarction. In cerebral venous thrombosis (CVT), occlusion of a cerebral vein or sinus produces an increase in venous pressure as well as an increase in tissue pressure within its drainage territory. Unless alternative venous pathways are available, tissue perfusion is reduced and ischemia occurs. Causes of CVT may be divided into local and systemic processes, with local causes consisting primarily of trauma and intracranial spread of infection or tumor, and systemic causes including the wide gamut of factors known to predispose to deep venous thrombosis. The most common of these are puerperal state, infection, and dehydration (in developing countries) and hereditary coagulation disorders such as factor V Leiden mutation and prothrombin gene mutation. The clinical presentation of CVT is varied but classically consists of evidence for increased ICP (e.g., headache, papilledema, and altered consciousness), focal neurological deficits, and seizures. Neuroimaging hallmarks of CVT include the empty delta sign on contrast-enhanced CT (in which a triangular pattern of enhancing dilated venous channels surrounds a nonenhancing clot in the superior sagittal sinus), multifocal hemorrhagic and nonhemorrhagic infarcts that do not conform to arterial vascular distributions on CT or MRI, and lack of venous filling on angiography. Treatment is directed at the underlying disorder (e.g., antibiotics for septic thrombosis), symptomatic control (e.g., anticonvulsants for seizures), and antithrombotic therapy (e.g., anticoagulants and thrombolytics). Because the disease is uncommon and has wide-ranging causes and variable outcome, the benefit of these treatments remains uncertain. The prognosis of CVT is highly variable but is influenced by such factors as age, rate of evolution and location of thrombosis, and nature of the underlying disease.

IV. INTRACEREBRAL HEMORRHAGE

Spontaneous intracerebral hemorrhage (ICH) accounts for approximately 10% of cerebrovascular disease and is associated with greater mortality and more severe neurological deficits than any other stroke subtype.

A. Pathophysiology

The pathophysiological mechanisms of brain injury due to ICH are complex. The primary injury is one of direct tissue destruction through mechanical compression. Rupture of a cerebral blood vessel introduces a sudden force of blood into the brain parenchyma where it acts as a mass to destroy tissue locally. In more than one-third of patients, continued bleeding or rebleeding result in hematoma enlargement and further mechanical injury within the first few hours after onset. The mass of blood produces tissue shifts within the intracranial cavity since the rigid skull prevents distension. If the volume of blood is large enough, brain herniation and death ensue.

In addition to the mechanical disruption produced by the hematoma, further damage is believed to occur after the bleeding stops. The mechanisms underlying this secondary injury are unknown, but ischemia and edema have been implicated. The importance of ischemia in the pathophysiology of ICH remains unsettled. Some experimental models of ICH have demonstrated a perihematomal zone of reduced CBF that may be below that necessary for neuronal viability. On the other hand, since ischemia requires insufficient blood flow to sustain metabolism, another possible explanation is that the reduced CBF simply reflects reduced metabolic demand of the damaged tissue surrounding the hematoma. Recent positron emission tomography (PET) studies in humans performed 10–22 hr after symptom onset have shown that CBF and metabolism are both reduced around the clot, with no evidence of ongoing ischemia. Cerebral edema has been demonstrated to occur within hours of induced ICH in some animal models, variably thought to result from the toxic effects of blood-derived enzymes, from increased osmotic pressure exerted by clot-derived serum proteins, or from ischemia. In humans, signal changes on radiographic studies after ICH indicate increased water content in the area surrounding the clot, but the clinical and pathophysiological significance of this is not known.

Hemostasis after hemorrhage is rapidly achieved at the site of vascular injury by the formation of a platelet-fibrin plug. Concurrent with the activation of local hemostatic processes, however, a localized fibrinolytic response is also activated to limit the extent of

coagulation, repair vascular injury, and reestablish normal blood flow. After several days, red blood cells within the clot begin to lyse, cellular infiltrates appear, and the process of reabsorption begins. Months later, a residual collapsed cavity is all that remains.

B. Causes/Risk Factors

The most common form of intracerebral hemorrhage, accounting for more than half of all cases, is hypertensive hemorrhage. Hypertensive ICH occurs predominantly in the deep portions of the cerebral hemispheres (putamen, thalamus, and deep white matter) and in the pons and cerebellum. What these sites have in common is that they are all supplied by small penetrating arteries—end vessels that have no collaterals and that are therefore vulnerable to the effects of increased blood pressure. Chronic hypertension contributes to the development of lipohyalinosis, fibrinoid necrosis, and microaneurysms (Charcot–Bouchard aneurysms). Although Charcot–Bouchard aneurysms have been demonstrated in the weakened vessel walls of patients with ICH, their pathogenetic role in vascular rupture is uncertain.

The occurrence of ICH in an atypical location, in multiple locations, or in association with subarachnoid hemorrhage raises suspicion of a nonhypertensive etiology, such as a cerebral vascular anomaly, blood dyscrasia, or trauma (Table II). Other risk factors for ICH include increasing age, black race, alcohol abuse, and low serum cholesterol. The relationship between smoking and ICH has not been convincingly established. Similarly, the impact of diabetes on risk of ICH is disputed.

Table II
Causes of Intracerebral Hemorrhage

Chronic hypertension
Coagulopathy
Drugs (anticoagulant therapy, cocaine, sympathomimetics, thrombolytic therapy)
Hemorrhagic transformation of cerebral infarcts
Thrombocytopenia
Trauma
Tumor
Vascular anomalies (aneurysms, arteriovenous malformations)
Vasculitis
Vasculopathy (amyloid angiopathy, moya-moya disease)

C. Clinical Features

The clinical presentation of ICH is often indistinguishable from that of ischemic stroke. As with ischemic stroke, the clinical manifestations of ICH depend on the location and size of the lesion, but they also reflect the presence of increased ICP. Thus, ICH more commonly produces alterations in level of consciousness as well as headache and vomiting. Symptoms are maximal at onset or develop over minutes to hours. Regression of deficits in the acute phase generally does not occur. Seizures may occur at onset, especially with lobar hemorrhages and underlying vascular or neoplastic lesions, but delayed seizures are rare.

Arterial blood pressure is elevated on admission in the majority of ICH patients, even in those with no history of hypertension, and generally declines to premorbid levels after 7–10 days without pharmacological intervention. Although this acute increase in blood pressure is often implicated as the cause of the hemorrhage, it may simply be a reflection of the brain's attempt to maintain CPP in response to the sudden increase in ICP.

Neurological deterioration after hospital admission has been reported to occur in 33–61% of patients with ICH. The cause for clinical worsening is not always evident, but increased hematoma size is found in more than one-fourth of cases. The role of edema in clinical deterioration is much more elusive.

1. Diagnostic Studies

Noncontrast CT scanning is the most widely used imaging modality in the diagnosis of ICH. Since acute blood has a high density, CT is sensitive in detecting even the smallest of hematomas. The typical CT appearance of an acute hematoma consists of a well-defined area of increased density surrounded by a rim of decreased density. The etiology of the hypodense rim is controversial, but it is at least partially explained by the process of clot retraction, in which the clot separates into a central mass of red blood cells and a surrounding area of serum. Also often apparent on the initial scan is mass effect of the hematoma on neighboring structures. Over time, the borders of both the high and low attenuation regions become increasingly indistinct, such that the hematoma is isodense with adjacent brain parenchyma by 2–6 weeks. A decrease in mass effect lags behind the signal changes, indicating that the apparent resolution of the hematoma is merely a manifestation of morphological changes within the extravasated blood. Peripheral contrast

enhancement can often be seen at this time and is thought to be due to angiogenesis and breakdown of the blood–brain barrier. By 2–6 months, true resolution of the hematoma and mass effect occurs, leaving an area of hypodensity, a slit-like scar, or no evidence of the previous hemorrhage. MRI of ICH is complex, reflecting the temporal evolution of hemoglobin degradation products. Because each hemoglobin oxidation state produces a predictable pattern of signal intensity, MRI is better able to determine approximate age of a hematoma than is CT. Its sensitivity to detect acute hematomas is controversial, however. Although MRI can be useful to visualize vascular abnormalities underlying a hematoma, angiography is indicated if a vascular lesion such as an aneurysm or arteriovenous malformation (AVM) is suspected.

2. Prognosis

Mortality following ICH is high (25–50%), with more than half of the deaths occurring in the first 48 hr. Transtentorial herniation accounts for a greater proportion of deaths in ICH than it does in ischemic stroke. Although patients who have small hemorrhages and mild deficits may recover completely, the majority of ICH survivors have significant residual disability. Many clinical and radiographic prognostic indicators have been identified, but all are not consistently recognized across studies. Clinical factors reported to predict poor prognosis include increasing age, reduced level of consciousness on admission, elevated blood pressure on admission, and in-hospital neurological deterioration. Radiographic features include large initial hematoma size, intraventricular spread of the hemorrhage, midline shift, hydrocephalus, and hematoma growth.

D. Treatment

Since the primary injury has already occurred by the time the patient obtains medical attention, treatment strategies have attempted to prevent or limit secondary injury. Unfortunately, no medical or surgical therapy has been shown to improve outcome after ICH. Trials of clot evacuation, aimed at achieving hemostasis and alleviating the potential for ischemia, have been disappointing. Ventriculostomy placement in patients with concomitant hydrocephalus or intraventricular blood has not been evaluated in a randomized trial, but retrospective data suggest that it does not improve outcome. Similarly, a benefit of osmotic therapy to reduce increased ICP has not been shown. Trials of corticosteroids to reduce edema have demonstrated no benefit but an increased complication rate. Therefore, the treatment of ICH is currently limited to supportive care. Trials of ultraearly clot evacuation and attempts to reduce the incidence of hemorrhage growth are ongoing. As with ischemic stroke, however, prevention through risk factor management remains the most effective means of reducing mortality and morbidity due to ICH.

The management of blood pressure in acute ICH is highly controversial, reflecting the competing theories that elevated blood pressure predisposes to hematoma growth and that lowering blood pressure exacerbates ischemic damage by impairing CBF. Data regarding the relationship between increased blood pressure and hematoma growth and mortality have been inconsistent. In addition, support for the theory that acutely lowering blood pressure reduces the likelihood of hematoma growth has not been demonstrated. Recent PET studies in patients with acute ICH and very elevated blood pressure (mean arterial pressure 120–150 mmHg) showed that a 15% reduction in blood pressure did not alter CBF in the periclot area.

V. SUBARACHNOID HEMORRHAGE

A. Pathophysiology

In subarachnoid hemorrhage (SAH), the primary site of bleeding is within the subarachnoid space; however, depending on etiology, it may also involve hemorrhage into the brain parenchyma, ventricular system, or subdural space. Rupture of an intracranial saccular aneurysm is by far the most common cause of spontaneous SAH. Saccular or berry aneurysms are small, rounded protrusions of the arterial wall occurring predominantly at arterial bifurcations at the base of the brain. Aneurysmal pathogenesis remains controversial, with the importance of developmental versus acquired factors in dispute. Proponents of the congenital theory suggest that aneurysms arise at sites of faulty fusion between muscular segments within the arterial wall. Supporters of the acquired-degenerative theory, on the other hand, focus on the role of vascular damage caused by hemodynamic stress. Citing the age-dependent occurrence of arterial defects and the predilection for aneurysms at arterial bifurcations (sites of maximal hemodynamic stress), those favoring the importance of acquired features postulate that degenerative changes within the internal elastic lamina

and media result in a local weakness that allows for aneurysm formation. A third possibility is that aneurysms develop at sites harboring congenital defects with superimposed degenerative changes.

What leads to aneurysmal growth and rupture is also debated. Hemodynamic stress and other factors intrinsic to the involved vessels may play a role. The time course over which aneurysms grow and subsequently rupture is unknown. One hypothesis is that most aneurysms that grow do so over a relatively short period of time (hours to weeks). If the limits of elasticity of the vessel wall are exceeded, the aneurysm ruptures; if not, the wall is stabilized by connective tissue deposition.

B. Causes/Risk Factors

Rupture of a saccular cerebral aneurysm is the cause of hemorrhage in approximately 80% of patients with nontraumatic SAH. Thus, genetic conditions that predispose to aneurysm formation, such as polycystic kidney disease, connective tissue disorders, and coarctation of the aorta, predispose to SAH. Other types of aneurysms that less commonly underlie SAH include atherosclerotic, mycotic, and traumatic aneurysms. Among the etiologies of nonaneurysmal SAH, trauma is the most common. AVMs, cocaine and stimulant abuse, neoplasia, and vasculitis account for the bulk of the remainder.

The risk of SAH increases with age, peaking at 55–60 years. There is a slight male predominance in younger age groups and a slight female predominance among older patients. Potentially reversible risk factors for SAH include cigarette smoking, oral contraceptive use, alcohol abuse, and hypertension. Prospective cohort studies have reported a relative risk of SAH as high as 5.7 for female and 4.7 for male smokers. No increased risk has been shown in former smokers, however. Oral contraceptive use, in addition to being an independent risk factor for SAH, dramatically increases the risk among smokers. A dose–response relationship exists between alcohol consumption and incidence of SAH. Finally, hypertension appears to increase the risk of SAH and may be the mechanism by which conditions such as polycystic kidney disease and stimulant drug use increase the risk for SAH.

C. Clinical Features

The most common initial symptom of SAH, occurring in more than 90% of patients, is sudden severe headache. Less severe warning (“sentinel”) headaches may precede the presenting event in as many as half and are thought to represent minor hemorrhages. In 45% of patients, transient or persistent loss of consciousness accompanies the headache. The mechanism responsible for this acute loss of consciousness is the sudden surge in ICP at the moment of hemorrhage that approaches systemic arterial pressure, resulting in inadequate cerebral perfusion. Vomiting can be prominent in awake patients. Seizure activity may be reported, but it is unclear whether this represents true epileptic seizures or reflex posturing related to the sudden increase in ICP. Focal deficits at onset are infrequent, but when present they may indicate the location of the lesion. After a few days, fever and stiff neck can develop, reflecting the sterile meningeal inflammation induced by the presence of blood in the subarachnoid space.

A worsening of neurological status following stabilization or improvement of symptoms often indicates one of the three major complications of SAH: rebleeding, hydrocephalus, or vasospasm. Rebleeding occurs in up to one-third of patients and is often fatal. The risk of rebleeding is greatest during the first 24 hr, declining rapidly after 1 or 2 weeks. It is heralded by a sudden worsening of headache, vomiting, development of a new neurological deficit, or arrhythmia. Hydrocephalus occurs if subarachnoid blood impairs CSF reabsorption at the arachnoid granulations or if ventricular blood obstructs its flow. It may develop acutely or gradually at any time after hemorrhage and most commonly produces a change in level of consciousness. Vasospasm, defined as segmental or diffuse narrowing of the arteries at the base of the brain, is a leading cause of morbidity and mortality following SAH. It can be detected angiographically in up to 70% of patients, of whom 20–30% become symptomatic. The pathogenesis of vasospasm is complex and not fully understood, but sustained exposure of vessels to

Table III
Fisher Grade of SAH on CT

Grade	Description
1	No blood detected
2	Diffuse or vertical layers < 1 mm thick
3	Localized subarachnoid clot and/or vertical layers $\geqslant 1$ mm thick
4	Intraparenchymal or intraventricular clot with diffuse or no SAH

extraluminal blood constituents and catecholamines is thought to play a role. It involves structural changes in the vessel walls and in adrenergic nerve fibers. The onset of vasospasm is delayed, commonly developing in the first week after initial hemorrhage, and it may persist for up to 3 weeks. The strongest predictor of subsequent vasospasm is the amount and distribution of subarachnoid blood on CT (Table III), with the greatest risk occurring in those having subarachnoid clots larger than 5×3 mm in the basal cisterns or layers of blood $\geqslant 1$ mm thick in the cerebral fissures (Fisher grade 3). Deficits reflect the territory of the involved arteries and may appear abruptly or gradually. Infarction may occur.

In addition to the neurological complications of SAH, several medical complications are common. Hypertension after SAH is associated with an increased risk of vasospasm and higher mortality. Multiple factors may underlie the increase in blood pressure, including increased catecholamine production induced by hypothalamic dysfunction, physiological response to increased ICP in an effort to maintain CPP, agitation, and pain. Blood pressure often returns to normal after hospital admission and/or after analgesics are administered. Disturbances in sodium and water balance occur in approximately one-third of patients, and hyponatremia and volume depletion after SAH are correlated with an increased risk of symptomatic vasospasm and poor outcome. Although hyponatremia was previously thought to be due to inappropriate secretion of antidiuretic hormone and was therefore treated with fluid restriction, later evidence suggested that both sodium and water are lost. The mechanisms underlying the volume contraction and inability to conserve sodium (cerebral salt wasting) are unknown. Finally, cardiac complications, thought to be due to catecholamine excess, are common in the first 48 hr after SAH. Arrhythmias are typically benign but can be severe or fatal.

Electrocardiographic abnormalities and elevated cardiac enzymes, resembling the changes in myocardial ischemia can be seen in more than half of patients. Rare patients may develop stunned myocardium, with transient cardiac pump failure and pulmonary edema.

1. Diagnostic Studies

CT is the imaging modality of choice in screening for SAH, having a sensitivity of $>90\%$. Blood appears as high attenuation within the subarachnoid cisterns. When blood is spread diffusely throughout the CSF spaces, its site of origin may be difficult to detect. Certain bleeding patterns are associated with specific aneurysm locations, however, such as that of blood in the anterior interhemispheric fissure commonly occurring with anterior communicating artery aneurysms. If CT scanning fails to demonstrate SAH and clinical suspicion is high, lumbar puncture is indicated to examine the CSF for blood. Following SAH, CSF contains red blood cells indicative of the hemorrhage, as well as a few white blood cells reflecting the secondary inflammatory response. A yellow pigment (xanthochromia), resulting from red cell breakdown, can be detected in the centrifuged fluid 2–6 hr after hemorrhage, allowing for differentiation from a traumatic spinal puncture. Xanthochromia persists for 1–4 weeks after SAH. Once SAH has been diagnosed, cerebral angiography is performed to identify the responsible vascular lesion, search for other lesions (multiple aneurysms are found in 20–30% of patients with aneurysmal SAH), and assist in operative management. Angiography is negative in 10–15% of patients with nontraumatic SAH. If the blood on CT is localized to the perimesencephalic cisterns and angiography is negative, prognosis is excellent and repeat angiography is almost always negative. Angiography is also used after the acute period in the detection of vasospasm. Another means of evaluating for vasospasm is transcranial Doppler.

2. Prognosis

Untreated aneurysmal SAH carries a poor prognosis, with an estimated mortality rate of approximately 50%. Ten percent of patients die before obtaining medical attention, and 25% die within the first 24 hr. Among survivors of the initial hemorrhage, 30–50% rebleed, an event that increases mortality even further. The risk diminishes after the first month, but even long-term survivors suffer a 3% annual risk of rehemorrhage. Overall, less than one-third of patients achieve good neurological recovery. Predictors of poor prognosis include loss of consciousness or poor neurological condition (i.e., high Hunt and Hess grade; Table IV) on admission, older age, hypertension, preexisting medical illness, $\geqslant 1$ mm thickness of subarachnoid blood on CT (Fisher grade 3), aneurysm location in the basilar artery, and symptomatic vasospasm.

D. Treatment

Management of SAH involves symptomatic treatment and prevention of secondary complications.

Table IV
Hunt and Hess Clinical Classification of SAH

Grade	Description
I	Asymptomatic or mild headache and neck stiffness
II	Moderate to severe headache and neck stiffness ± cranial nerve palsy
III	Mild focal deficit, lethargy, or confusion
IV	Stupor, moderate to severe hemiparesis, early decerebration
V	Deep coma, decerebrate rigidity

Analgesics, antiemetics, and sedatives relieve pain, nausea, and agitation, factors that can contribute to hypertension and rebleeding. Although elevated blood pressure increases the risk of rebleeding, lowering it may reduce cerebral perfusion in patients who have elevated ICP and may worsen neurological symptoms in patients who have vasospasm. Because seizures can increase the risk of rebleeding, anticonvulsants are indicated if seizures occur. The value of prophylactic anticonvulsants in patients who have not had a seizure is unknown. Multiple clinical trials have demonstrated that antifibrinolytic agents reduce the risk of rebleeding, but this benefit is offset by an increased incidence of vasospasm and hydrocephalus. Whether a shorter course of antifibrinolytic therapy (before the risk period for vasospasm begins) might eliminate the negative effect is unknown. In aneurysmal SAH, the most effective way to prevent rebleeding is to obliterate the aneurysm by surgical clipping of its neck. The optimal timing of surgery is controversial, but in patients with good clinical grade (Hunt–Hess grades I–III), favorable neurological outcome is seen more commonly when surgery is performed early (within 72 hr). The role of endovascular repair, using either detachable balloons that trap the aneurysm (but also occlude the parent artery) or electrolytically detachable coils that thrombose the aneurysm, awaits data on long-term efficacy.

Management of vasospasm is largely preventative, involving avoidance of volume contraction and hyponatremia (through hydration with isotonic saline) and administration of the calcium channel blocker nimodipine. Although its mechanism of action is unclear, nimodipine has been demonstrated in multiple controlled trials to significantly reduce the risk of symptomatic vasospasm and poor outcome after SAH. If symptomatic vasospasm develops, treatment strategies include hemodynamic augmentation and directly opening constricted vessels. In hemodynamic augmentation, blood volume and cardiac output are increased via aggressive hydration and blood pressure is increased via administration of vasoactive agents in order to enhance CBF and prevent cerebral infarction. Potential complications of the therapy include heart failure and myocardial ischemia. The second strategy involves endovascular treatment of constricted vessels with either balloon angioplasty or intraarterial infusion of vasodilating agents such as papaverine. Potential complications of endovascular treatments include arterial perforation, cerebral infarction, and aneurysm rebleeding. Each of these strategies has shown efficacy in case series (although the response to papaverine is often transient, with vasospasm returning after 24–48 hr), but controlled trials are lacking.

See Also the Following Articles

BRAIN DAMAGE, RECOVERY FROM • BRAIN LESIONS • CEREBRAL CIRCULATION • CEREBRAL EDEMA • CEREBROVASCULAR DISEASE • COGNITIVE REHABILITATION • MODELING BRAIN INJURY/TRAUMA

Acknowledgment

The author gratefully acknowledges the support and advice of William J. Powers, M. D., in preparing the manuscript.

Suggested Reading

Barnett, H. J., Mohr, J. P., Stein, B. M., and Yatsu, F. M. (Eds.) (1998). *Stroke: Pathophysiology, Diagnosis, and Management*, 3rd ed. Churchill Livingstone, New York.

Ginsberg, M. D., and Bogousslavsky, J. (Eds.) (1998). *Cerebrovascular Disease: Pathophysiology, Diagnosis, and Management*. Blackwell Science, Malden, MA.

Welch, K. M. A., Caplan, L. R., Reis, D. J., Siesjo, B. K., and Weir, B. (Eds.) (1997). *Primer on Cerebrovascular Diseases*. Academic Press, San Diego.

Suicide

BETH S. BRODSKY and J. JOHN MANN

New York State Psychiatric Institute and Columbia University College of Physicians and Surgeons

GLOSSARY

borderline personality disorder A disorder of the personality that is characterized by dysregulation in emotional and interpersonal functioning and self-destructive and impulsive behavior.

cerebrospinal fluid 5-hydroxyindole acetic acid The major metabolite of serotonin.

dialectical behavior therapy A type of psychotherapy in which cognitive behavioral techniques are aimed at the reduction of suicidal behavior.

major depression A mental disorder characterized by prolonged sadness, lack of pleasure, appetite and sleep disturbance, impaired concentration, extreme feelings of low self-worth and/or guilt, and sometimes suicidal thoughts or behaviors.

psychosis An emotional state characterized by unusual perceptual experiences, such as visual or auditory hallucinations, by delusional thoughts or beliefs, and/or by disorganized thinking or behavior.

serotonin A brain chemical (neurotransmitter) that is believed to have a role in depression and impulsivity.

stress–diathesis model A conceptual framework in which suicidal behavior is seen as a result of the interaction between a trait predisposition with an environmental event.

suicide attempt A self-injurious act committed with an intent to die.

Theories addressing the phenomenon of suicide have been put forth from numerous academic disciplines. However, in the past few decades, empirical research efforts toward improved understanding, prediction, assessment, treatment, and prevention of suicide have come mainly from within the fields of psychiatry and neurobiology. This article summarizes the research findings that both identify risk factors and challenge common myths regarding causal factors for suicidal behavior. A comprehensive hypothetical explanatory and predictive model that incorporates many of these correlates of suicide risk is described. The application of these findings toward improved prevention, treatment, and further research is outlined.

I. INTRODUCTION

A. Suicide Rates

Suicide accounts for more than 30,000 deaths in the United States annually. Men are two or three times more likely to complete suicide than women. Rates of completed suicide in males are higher in the age range of 65–80 years, although during the past 30 years the rates of adolescent suicide have reached levels seen in adults. It is estimated that there are about 10 suicide attempts for every completed suicide. Suicide attempts are more common in women.

B. Definition of Suicidal Behavior

In order to increase specificity in prediction of suicidal behavior, it is necessary to consider the spectrum of suicidal behavior that ranges in severity from completed suicide to suicide attempts and, finally, suicidal ideation. This article encompasses completed suicide and suicide attempts defined as a self-damaging act carried out with at least some intent to end one's life. We do not address various other forms of

self-destructive behavior, such as self-mutilation, failure to cooperate with medical treatment in severely ill patients, cigarette smoking and substance abuse.

Nonlethal suicide attempts are more complex to assess than completed suicide since they include a medical damage dimension. This can range from individuals who survive by chance despite inflicting extensive injuries requiring intensive care treatment to individuals who sustain little or no injury. The assessment of suicidal intent is another dimension that might be ambiguous in light of the individual's survival. Although medical lethality and intent to die are often closely related, this is not necessarily the case in a nonlethal suicide attempt, in which it is possible to have very little medical damage, despite high suicidal intent, or vice versa.

Thus, in the case of nonlethal suicide attempts, broad and not completely distinct categories of suicidal behavior can be generated along the dimensions of medical lethality, level of suicidal intent, and planning. The first category is that of failed suicide, in which a highly lethal method is used that inflicts considerable medical damage and is accompanied by careful preparation and planning. At the other end of the attempt spectrum is a suicide attempt that involves a low lethality method, inflicts little medical damage, and the level of objective intent or planning is relatively low.

In researching nonlethal suicide attempts, the most common predictive factor for suicidal behavior is whether or not an individual has made a suicide attempt (attempter vs nonattempter). Other potential outcome measures include the number of lifetime suicide attempts, the age at first suicide attempt, medical lethality of the most lethal attempt, suicide intent associated with the most lethal attempt, and level of suicidal ideation at different time points.

II. STRESS–DIATHESIS MODEL OF SUICIDAL BEHAVIOR

From the perspective of a stress–diathesis model, suicidal behavior results from the interaction of a behavioral/biologic/genetic predisposition to act on self-destructive urges, together with a stressor or trigger such as an acute psychiatric illness and perhaps an environmental or psychosocial factor (Fig. 1).

A. Predisposing Trait

The criteria for a predisposing trait are stability over time and a consistent correlation with suicidal behavior. Thus, by identifying certain predisposing personality, biological, or genetic traits, the model also further specifies individuals at risk for suicidal behavior within the context of a mood state, environmental stressor, or other state trigger. It also provides a theoretical framework for further research that will increase prediction and recognition of suicidal risk. Viewing risk factors in interaction with each other also allows a vantage point from which to reconsider prevailing misconceptions regarding suicidal behavior. For example, although it makes intuitive sense that a severe life stressor might trigger a suicide attempt, this interpretation does not explain why most individuals with the same objective life circumstances do not make a suicide attempt. Viewing the stressor in interaction with a predisposing trait within the individual can further clarify the role of the stressor without mistakenly concluding that the stressor is the sole explanatory factor or not relevant.

In addition, the recognition of a trait disposition to act on suicidal urges may help identify individuals at risk for suicide across diagnostic categories. Most studies of potential risk factors have examined a single diagnostic group, preventing the generalization of findings to a wider range of clinical patients. Moreover, most studies have examined a narrow range of potential risk factors, preventing the determination of their relative importance and the development of a comprehensive explanatory or predictive model.

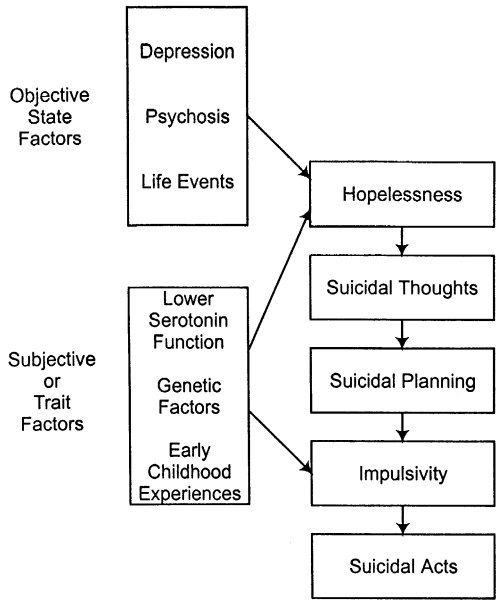


Figure 1 A model of suicidal behavior.

B. State Factors

A stress–diathesis model of suicidal behavior would define stressors as the recurrence of depressed mood or psychotic symptoms within the course of a psychiatric illness or a negative stressful life event, such as the death of a loved one or losing one's job. In the model, such a stressor might lead to the subjective experience of depressed mood, feelings of hopelessness, and suicidal thoughts. The suicidal thinking might lead to more active suicidal planning. At this point, these states would then interact with a trait disposition that might both increase the subjective experience of the depression, hopelessness, and suicidal ideation and interact with the lower threshold to act on the urges or thoughts (Fig. 1).

III. CLINICAL AND BIOLOGICAL CORRELATES OF SUICIDAL BEHAVIOR

A. Clinical Correlates

A previous suicide attempt is the best predictor of a future completed suicide or suicide attempt. However, only 20–30% of individuals who commit suicide have made a previous suicide attempt; therefore, to predict suicide attempts or suicide in the other 80% whose first suicide attempt is successful, it is necessary to identify other predictors of risk.

Much research has been devoted to identifying clinical correlates of suicidal behavior. In terms of demographic characteristics such as age and educational level, studies report conflicting findings. Generally, age tends to be higher among suicide completers but lower in those who attempt suicide than in comparable nonattempter patient populations. Different studies report both higher and lower educational levels in suicide attempters versus nonsuicidal patients. Income level has not consistently emerged as a correlate, although socioeconomic factors such as unemployment and lack of job security have been associated with higher suicide rates. The relationship between marital status and suicide is confounded by an interaction effect with gender. There are higher rates of suicide among whites than blacks. Women attempt suicide more often than men, but men employ more lethal methods and are more likely to complete suicide.

1. Psychiatric Disorders

Approximately 90% of suicide cases have a psychiatric diagnosis at the time of death. Mood disorders generally account for 60% of cases, with major depression being the most common. Lifetime mortality due to suicide is approximately 15% in individuals suffering from major depressive disorder, 20% in those with manic–depressive illness, 18% in alcoholics, and 10% in schizophrenia. In addition, certain personality disorders are known to be associated with high risk for suicidal behavior. In particular, 5–10% of individuals diagnosed with borderline personality, a disorder characterized by impulsive and self-destructive behaviors as well as emotional and interpersonal dysregulation, die by suicide.

Thus, suicide is not merely a complication of a personal crisis or a psychosocial setback; it almost always occurs within the context of a psychiatric illness. Suicide may be understood as an attempt to seek relief from the emotional pain and hopelessness regarding recovery that is often present in individuals with severe, recurrent major depression.

However, most patients with a psychiatric illness do not commit suicide. Moreover, objective severity of depression does not distinguish those individuals who suffer from recurrent major depression who make a suicide attempt from those who do not. This suggests that suicide is not just a logical response to extreme stress or depressed mood state, but that some patients have a vulnerability or predisposition to suicidal behavior. Finding ways of evaluating this predisposition or diathesis can be important for the detection of high-risk patients, and at the same time the diathesis is another potential target for a therapeutic intervention.

2. Personality Traits

Certain personality traits appear to contribute to a predisposition to suicidal behavior. Individuals with major depression, psychotic disorders, and personality disorders that have a history of suicidal behavior also have a greater lifetime history of aggression. Criminals who have a history of suicidal behavior have a history of more severe aggression than criminals who do not have a history of suicidal behavior. Thus, there may be a fundamental predisposition to aggressive behavior, whether it is self-directed, in the form of suicidal behavior or externally directed in the form of aggression against property of other persons.

Impulsivity, a personality trait or cognitive style characterized by a tendency to act more quickly on urges or to stimuli, has been found to be associated with suicidal and self-destructive behaviors within various adult psychiatric populations. Individuals who self-injure or make suicide attempts score higher on

clinical measures of impulsivity. In individuals with borderline personality disorder, the presence or absence of impulsive behaviors in at least two areas of life, such as binge spending, binge eating, reckless driving, gambling, and unsafe sexual practices, is the clinical feature most predictive of suicidal behavior.

Individuals with a past history of a suicide attempt exhibit greater lifetime aggression and impulsivity than nonattempters with the same psychiatric illness. Clinical studies that assess lifetime externally directed aggression and impulsivity have identified a trait aggression/impulsivity factor that distinguishes between suicide attempters and nonattempters regardless of psychiatric diagnosis, level of objective depression, and negative life events. Thus, the association of suicidal behavior and traits of aggression and impulsivity suggests there is a common underlying predisposition to suicidal and aggressive behavior that may relate to impulsivity.

3. Hopelessness versus Reasons for Living

Studies have shown that levels of hopelessness, subjective depression, and suicidal ideation are higher in individuals with previous suicide attempts, and that suicide attempters report fewer reasons for living. This is so despite a lack of significant difference in the objective amount of stressful or negative life events between suicide attempters and nonattempters. In addition, hopelessness is greater in suicide attempters compared to nonattempters during an acute depression, after successful treatment, and between episodes of depression. Hopelessness can predict the risk for suicide over a 10-year follow-up period. These observations suggest that the degree of hopelessness may be influenced by a trait as well as a state vulnerability.

4. Family History and the Familial Transmission of Suicide

Clinical studies show that a family history of suicide is associated with a greater risk of both suicide attempts and completed suicide. Genetic and familial factors have been investigated in association with psychiatric diagnoses in general and suicide risk in particular. Twin studies show that monozygotic twins have a greater concordance for suicidal behavior than dizygotic twins. Adoption studies report the transmission of a genetic risk factor for suicide, independently of psychiatric illness. This is illustrated by the Copenhagen Adoption study, which found a significantly higher percentage of suicide in the biological families of the suicided adoptees compared to matched adoptees who did not commit suicide. Recently, molecular genetic studies have reported that a polymorphism in the tryptophan hydroxylase gene is associated with suicidal behavior.

There is also evidence that suicidal behavior transmitted in families independently of psychiatric diagnoses is related to transmission of impulsive aggression. It is possible that genes might be responsible for lower serotonergic functioning, which is associated with trait aggression and impulsivity. Genetic transmission of a propensity for externally directed aggression and suicidal behavior, independent of transmission of major depression, is consistent with evidence that the diathesis for suicidal behavior is transmitted in families independent of psychiatric diagnosis.

Trait aggression and impulsivity may also be a result of early life family experiences, including a history of physical or sexual abuse. Indeed, adults who have made at least one suicide attempt are more likely to report a history of childhood abuse, independent of their psychiatric diagnosis.

5. Neuropsychological Correlates

Recent neuropsychological studies have begun to identify certain features of cognitive processing that distinguish individuals who have made a suicide attempt from those who have not. High lethality suicide attempters do worse than nonattempters on tasks that measure cognitive flexibility, which may affect decision making. On timed tasks that measure impulsive cognitive style, suicide attempters make more commission errors, suggesting that they may be more likely to act inappropriately under conditions of uncertainty. These neuropsychological measures of impulsivity and impaired decision making may ultimately prove to be more sensitive than a clinical history risk assessment.

B. Biological Correlates

Investigation of biological correlates of suicidal behavior has generated evidence for a trait predisposition that distinguishes between individuals at high risk versus low risk for suicidal behavior. Neurobiologic studies have also identified biological correlates of the clinical risk factors for suicide mentioned previously, such as depressed mood, trait impulsivity, and aggression. These findings have come from postmortem studies on the brain tissue from suicide victims as well as from studies of neurotransmitter function in live individuals with a history of serious suicide attempts.

1. Neurotransmitters

A large number of studies of postmortem brain tissue from suicide victims have been carried out examining indices of the serotonergic, noradrenergic, and dopaminergic neurotransmitter systems.

a. Serotonin Serotonin is a neurotransmitter that is implicated in depression and impulsivity. Studies of the serotonergic system have generally found evidence of reduced serotonergic activity in suicide victims, as indicated by decreases in presynaptic serotonin nerve terminal binding sites. Autoradiographic studies of postmortem brain tissue suggest that these abnormalities in suicide victims are more pronounced in the ventral prefrontal cortex than in the dorsolateral prefrontal cortex.

Similarly, postsynaptic serotonin receptors, such as the 5-HT1A and 5-HT2A receptors, appear to be increased in number in the prefrontal cortex of suicide victims. One possible explanation for the increases in these subtypes of postsynaptic serotonin receptors is compensatory upregulation in response to reduced serotonin neuron activity. Most of these altered serotonin receptor binding indices are detected in the ventral prefrontal cortex, which indicates that this brain region is of particular importance in relation to the risk for suicide. The reduction in the presynaptic serotonin nerve terminal transporter binding and the increase (possibly compensatory) in postsynaptic receptor binding are consistent with the notion of reduced serotonin input into this brain region.

The ventral prefrontal cortex is involved in the executive function of inhibition, and injuries to that area of the brain have been shown to result in cognitive and behavioral disinhibition. Reduced serotonergic input into this part of the brain may also result in impaired inhibition and a greater propensity to act on powerful feelings such as suicidal or aggressive feelings. In support of this hypothesis, most postmortem studies of brain stem levels of serotonin or its major metabolite, 5-hydroxyindole acetic acid (5-HIAA), have found modest reductions in suicide victims, suggesting that serotonergic activity may be reduced in these individuals.

Serotonin function has also been measured in survivors of nonlethal suicide attempts using several techniques. Samples of cerebrospinal fluid (CSF) are assayed to determine levels of serotonin metabolite 5-HIAA. Most studies report that about two-thirds of major depressives who have attempted suicide have lower CSF 5-HIAA levels compared to nonattempters with the same psychiatric diagnoses. In addition, low CSF 5-HIAA predicts future suicide attempts and suicide completion. Similarly, lower CSF 5-HIAA has been found in schizophrenics with a history of suicidal behavior compared to schizophrenics without such a history and also in individuals with personality disorders who make suicide attempts compared to patients with the same diagnosis who do not. Therefore, low CSF 5-HIAA, consistent with the notion of a biochemical trait, appears to be a relatively stable marker that is associated with suicide attempt behavior, independent of psychiatric diagnosis.

The medication fenfluramine, known to stimulate the release of serotonin, has been widely used in challenge studies in which prolactin levels in the blood (a proxy measure for serotonin response) are measured and compared in the same individual after administration of the active fenfluramine medication or a placebo. Individuals with major depression or personality disorders who have attempted suicide have a blunted prolactin response elicited in response to fenfluramine. The more lethal the suicide attempt, the more likely there is to be low CSF 5-HIAA or a blunted prolactin response to fenfluramine. Brain imaging techniques such as positron emission tomography (PET) scans provide an even more sensitive measure of serotonin functioning by evaluating regional changes in serotonin receptors or regional responses to fenfluramine. Such studies are currently in progress.

In addition, serotonin mediates a number of platelet functions, some of which are altered in suicide attempters. A relationship between lethality and higher platelet 5-HT2A receptor number has been reported. Thus, three different indices of serotonergic function appear to correlate with suicidal behavior. The serotonergic correlations with suicide appear to be equally strong regardless of the associated psychiatric disorder. Therefore, the serotonergic abnormality in the ventral prefrontal cortex of suicide victims may be related to the predisposition to suicidal behavior rather than to the psychiatric illness that may have triggered it.

b. The Noradrenergic System A number of alterations in noradrenergic indices are also reported in suicide victims, although the results are less consistent than those for the serotonergic system. The noradrenergic system is implicated in the neurobiological response to stress. In postmortem brain tissue of suicide victims, there is evidence of a reduced number of noradrenergic neurons in the locus coeruleus. In addition, levels of norepinephrine appear to be lower

in the brain stem of suicide victims and α_2- adrenergic autoreceptor numbers are increased. We have proposed that norepinephrine may be depleted from the smaller numbers of noradrenergic neurons present in the brain stems of suicide victims. Studies of chronic stress in rodents report depletion of norepinephrine. There is evidence of hyperactive stress response systems in depression. Thus, these biochemical findings could potentially be a result of the stress preceding a suicidal event in a serious psychiatric illness.

c. Dopamine Fewer studies have been carried out on the dopaminergic system. There is postmortem evidence of decreased dopaminergic binding in the prefrontal cortex of adolescent suicide victims. In live suicide attempters with major depression, there have been reports of reduced CSF homovanillic acid (the major metabolite of dopamine) as well as reports linking reduced dopaminergic function to major depression. Studies are under way to evaluate the dopaminergic system in suicide attempters.

C. Suicide, Aggression, Impulsivity and Neurobiology

Recent reports indicate that impulsivity and externally directed aggression are associated with low serotonin activity. Low CSF 5-HIAA has been shown to be present in impulsive murderers and arsonists compared to nonimpulsive murderers and other criminals as well as control subjects. A blunted prolactin response to fenfluramine (an indication of low serotonin activity) has been reported in association with aggressive behavior and impulsivity in individuals with personality disorders and major depression and medically healthy volunteers. Thus, reduced serotonergic function appears to be associated with impulsive aggression. This relationship is analogous to that between serotonergic function and suicidal behavior and leads to the more general hypothesis that serotonergic function supports a restraint mechanism, and a deficiency of serotonergic function results in greater impulsivity and aggression that also include self-directed aggression in the form of suicidal behavior.

D. Environmental Factors

Genetic and familial factors, as well as early childhood familial environmental factors such as childhood abuse, may be responsible for the transmission of suicidal behavior via trait aggression and impulsivity. Early childhood environmental factors might either contribute to the lower serotonergic functioning or personality traits of impulsivity/aggression or be correlated with them, with the lower serotonergic functioning as the underlying correlate of the trait factor. Demonstrating the interaction of genetics and environment, peer-reared monkeys when compared to maternally raised monkeys have their serotonergic activity reset to a lower level that persists into adulthood. The rank order of their serotonergic activity is not altered within the peer-raised monkeys. Thus, the effects of rearing are superimposed on the genetic effects. Given that a history of child abuse is associated with a greater risk for suicidal behavior in adulthood, it may be useful to extrapolate from the findings of these monkey studies that examine environmental deprivation. It is possible that child abuse resets serotonergic function at lower levels and this neurochemical effect persists into adulthood, contributing to the increased risk for suicidal behavior.

Other environmental suicide risk factors that might correlate with lower serotonergic function include alcoholism, substance abuse, smoking, and head injury. These risk factors might also interact with each other. Alcoholism and substance abuse are associated with head injury and higher scores on clinical measures of impulsivity. In fact, approximately 50% of head injuries are sustained while using alcohol, thereby resulting in disinhibition and a greater probability of suicidal behavior.

Potentially related to the increased rate of substance abuse and alcoholism in the suicide attempters is the observation that cigarette smoking is more common in individuals who have made a suicide attempt. Smoking is associated with higher rates of completed suicide as well, independent of any association of psychiatric illness with smoking.

In nonhuman primates, low levels of cholesterol or cholesterol-lowering treatments are associated with an increase in suicidal behavior and a decrease in serotonergic function. In addition, a low-cholesterol diet in nonhuman primates is associated with an increase in aggressive behaviors and a decrease in social contact. Thus, cholesterol levels appear to have a measurable, although small, effect on behavior involving aggression and suicidality. The link between the serotonergic system and cholesterol levels remains to be demonstrated in humans.

Given the evidence linking low serotonergic activity separately to suicidal behavior, aggression, and alcoholism, it is conceivable that low serotonergic activity may, to some degree, underlie all three problems and

that low serotonergic activity may mediate genetic effects on suicide, aggression, and alcoholism. Reports of increased aggression, impulsivity, cocaine and alcohol consumption in mutant mice lacking the 5-HT1b receptor indicate that a genetically-mediated serotonin abnormality can result in both increased impulsive aggression as well as possibly alcoholism and substance abuse. Although suicidal behavior is never observed in a rodent model, other elements of disinhibitory psychopathology are observed in this model.

In summary, much of clinical suicide research has focused on the identification of isolated demographic, diagnostic, and biological correlates of suicidal behavior. In addition, clinical assessment of suicide risk has focused on objective severity of symptoms as well as life stressors. However, recent findings identify an aggressivity/impulsivity trait that along with a biochemical trait of lowered serotonergic functioning are predictive of suicidal behavior independently of psychiatric diagnostic groups, depressed mood state, and level of external stressors.

IV. CLINICAL ASSESSMENT, INTERVENTION, AND PREVENTION

A. Risk Assessment

Recognition of risk is vital to suicide prevention. Conceptualizing risk as an interaction of trait predisposition and state mood and/or environmental trigger can lead to better detection of high-risk patients. Accordingly, a clinical history should focus on the parameters that reflect a diathesis or predisposition for suicidal behavior. These include the following: (i) impulsive behaviors such as binge spending, binge eating, substance abuse, reckless driving, and unsafe sexual practices; (ii) other outwardly directed aggression across the life span on a continuum from less severe incidents, such as temper outbursts, conflicts with authority figures such as teachers and supervisors, and arguments that involve verbal abuse, to more severe physical fights, legal problems, threatening behavior, assault, and use of weapons; (iii) family history of a suicide attempt and/or suicide and any report of physical and/or sexual abuse in childhood are associated with suicidal behavior in adulthood; and (iv) levels and patterns of hopelessness and suicidal ideation that are disproportionate to the objective severity of depression or life events in individuals with a history of suicidal acts. In terms of state, the sense of hopelessness, subjective depression, perception of fewer reasons for living, and especially severity of suicidal ideation correlate with suicide attempts.

In the near future, perhaps biological measures of serotonin and other neurotransmitter activity, as well as neuropsychological measures, will supplement clinical history to increase sensitivity of assessment and risk prediction. These measures currently involve the measurement of CSF 5-HIAA; however, newer techniques involving functional brain imaging of serotonergic activity and candidate gene markers hold promise. Given the rapid advances in PET imaging of serotonergic function *in vivo*, and the identification of polymorphisms in serotonin-related candidate genes, we now have the tools for developing neurobiological tests to detect patients at high risk for suicide.

B. Psychopharmacology

Unfortunately, despite the great progress made during the past 35 years in the development of effective medications for the treatment of major psychiatric illnesses, there has been little reduction in suicide rates. This is not due to a failure to contact health care professionals, because most studies report that the majority of patients (between 50 and 80%) have seen a health care professional within 30–90 days of suicide. However, among individuals who have committed suicide in the context of major depression, most studies find that only 10–14% of these individuals received adequate doses of antidepressant treatment prior to suicide. Thus, it appears that there is the potential for more effective recognition of psychiatric conditions such as major depression and the treatment of these conditions with adequate doses of antidepressants in terms of reducing suicide rates. A Swedish study in which primary care physicians were educated on the diagnosis and treatment of depression resulted in an increase in prescription rates of antidepressants and a decrease in suicide rates.

Choosing the best medication partially depends on the psychiatric disorder associated with suicidal ideation. Some studies suggest that selective serotonin reuptake inhibitors (SSRIs) produce a more rapid amelioration of suicidal symptoms in patients with major depression. This observation is consistent with a model of suicidal behavior that posits reduced serotonergic activity as underlying the vulnerability to suicidal behavior in the presence of major depression or any other psychiatric disorder that is accompanied by suicidal ideation. Lithium, which enhances

serotonergic activity, appears to reduce suicide risk in bipolar patients, providing further support for a role of serotonin in moderating suicide risk.

Treating patients with the newer SSRI antidepressant medications instead of older tricyclic antidepressants, which are more lethal on overdose, has the potential for reducing suicide rates. Non-antidepressant sedative medications, such as barbiturates or benzodiazepines, should be avoided in the treatment of patients with mood disorders. Failure to improve the mood disorder, combined with potential disinhibition by these sedatives and their lethality on overdose, makes for a risky treatment strategy.

C. Psychotherapy

Treatment should target impulsivity and aggression as well as depressed mood and psychotic symptoms. A promising psychotherapeutic approach to the treatment of self-injurious behavior in individuals with borderline personality disorder is dialectical behavior therapy (DBT), a form of cognitive behavioral therapy designed to specifically target self-injurious behaviors and the personality traits that contribute to self-injury, such as impulsivity. This therapy is based on the biosocial theory, in which borderline personality disorder is described as a disorder of emotional, cognitive, and behavioral dysregulation. According to this theory, dysregulation in these areas of function arises from a genetic predisposition to an emotionally vulnerable temperament in interaction with an emotionally invalidating environment. Such a conceptualization of interaction between biological and environmental factors is similar to a diathesis–stress model of suicidal behavior. DBT therapy involves the teaching of skills to regulate emotions and behavior as well as to reframe thinking that leads to hopelessness and suicidal ideation. Interpersonal skills target the types of life events that tend to trigger suicidal behavior in these individuals. Empirical studies testing the efficacy of DBT indicate that the therapy decreases self-injury without a concomitant decrease in depressive symptoms. These findings also support the idea of self-injury being related to behavioral traits that are independent of a diagnosis of depression.

Other interventions for suicide prevention include psychoeducation for primary care physicians to recognize the signs of depression and for the general public to be more familiar with the warning signs of suicide. An increase in the availability of parenting classes would help identify and address the stressors that often lead to early life experiences of physical and sexual abuse.

V. DIRECTIONS FOR FUTURE RESEARCH

The stress–diathesis model of suicidal behavior provides a comprehensive framework in which the interaction of various risk factors for suicidal behavior can be further explored in order to increase specificity in the prediction of suicidal risk. Prospective studies are under way to increase understanding of the interaction between biological and environmental factors. Newer technologies are being employed to measure biological and neuropsychological traits that interact with mood states and environmental triggers in the development of more sensitive tools for suicide risk assessment. Treatment outcome research with high-risk subjects is being conducted to compare the efficacy of serotonin reuptake inhibitor medications, DBT, and other psychotherapy approaches in decreasing suicidal behavior.

See Also the Following Articles

AGGRESSION • BEHAVIORAL PHARMACOLOGY • COGNITIVE PSYCHOLOGY, OVERVIEW • DEPRESSION • MANIC–DEPRESSIVE ILLNESS • MOOD DISORDERS • NEUROPSYCHOLOGICAL ASSESSMENT • STRESS • VIOLENCE AND THE BRAIN

Acknowledgments

This work was supported by Public Health Service Grants MH46745 and MH48514.

Suggested Reading

Arango, V., Underwood, M. D., Gubbi, A. V., and Mann, J. J. (1995). Localized alterations in pre- and postsynaptic serotonin binding sites in the ventrolateral prefrontal cortex of suicide victims. *Brain Res.* **688,** 121–133.

Asberg, M., Nordstrom, P., and Traskman-Bendz, L. (1986). Cerebrospinal fluid studies in suicide. An overview. *Ann. N.Y. Acad. Sci.* **487,** 243–255.

Beck, A. T., Steer, R. A., Kovacs, M., and Garrison, B. (1985). Hopelessness and eventual suicide: A 10 year prospective study of patients hospitalized with suicidal ideation. *Am. J. Psych.* **142**(5), 559–563.

Higley, J. D., *et al.* (1981). Paternal and maternal genetic and environmental contributions to cerebrospinal fluid monamine metabolites in rhesus monkeys (*Macaca mulatta*). *Arch. Gen. Psych.* **38,** 15–22.

Kaplan, J. R., Shively, C. A., Fontenot, M. B., Morgan, T. M., Howell, S. M., Manuck, S. B., Muldoon, M. F., and Mann, J. J. (1994). Demonstration of an association among dietary cholesterol, central serotonergic activity, and social behavior in monkeys. *Psychosom. Med.* **56**(6), 479–484.

Kapur, S., Mieczkowski, T., and Mann, J. J. (1992). Antidepressant medications and the relative risk of suicide attempt and suicide. *J. Am. Med. Assoc.* **268**(24), 3441–3445.

Linehan, M. (1993). *Cognitive Behavior Therapy of Borderline Personality Disorder*. Guilford, New York.

Malone, K. M., Oquendo, M. A., Haas, G. L., Ellis, S., and Mann, J. J. (2000). Protective factors against suicidal acts in major depression: Reasons for living. *Am. J. Psychiatry* **157**(7), 1084–1088.

Mann, J. J. (1998). The neurobiology of suicide. *Nature Med.* **4**(1), 25–30.

Mann, J. J., Waternaux, C., Haas, G. L., and Malone, K. M. M. (1999). Towards a clinical model of suicidal behavior in psychiatric patients. *Am. J. Psych.* **156**(2), 181–189.

Roy, A., Segal, N. L., Centerwall, B. S., and Robinette, C.D. (1991). Suicide in twins. *Arch. Gen. Psych.* **48,** 29–32.

Stoff, D. M., and Mann, J. J. (1997). *Neurobiology of Suicide. From the Bench to the Clinic*. N. Y. Acad. Sci., New York.

Superior Colliculus

BARRY E. STEIN, MARK T. WALLACE, and TERRENCE R. STANFORD
Wake Forest University School of Medicine

GLOSSARY

anuran A tailless amphibian such as a frog or toad.

contralateral Located on the opposite side of the body.

encephalization The evolutionary and developmental processes by which the cerebral cortex takes over the functions of lower brain structures.

geniculostriate Relating to the visual pathway connecting the lateral geniculate nucleus of the thalamus to the primary visual cortex.

habituation The diminution or loss of a response to a frequently repeated stimulus.

homolog A structure or organ similar in form and origin to that in another organism.

ipsilateral Located on the same side of the body.

lamina Layer.

motor error The distance of a target stimulus from the current position of gaze.

movement field The range of amplitude and direction (saccade vector) to which the motor-related discharge of a superior colliculus saccade-related neuron is associated (the motor counterpart of a neuron's sensory receptive field).

receptive field The area of the sensory field (or the receptor epithelium) that when stimulated influences the activity of a single neuron.

retinal error The difference between the location of a visual stimulus on the retina and the fovea.

saccade A very rapid, conjugate movement of the eyes (such as those used in reading).

somatotopy The topographic representation of the body in the brain.

spatiotopic A topographic representation of space (e.g., a map of auditory space).

stereopsis The ability to use binocular cues in order to determine depth.

tonotopy The topographic representation of frequency in the brain.

topography The delineation or description of a map of the features of a sensory representation in the brain (e.g., the arrangement of the visual receptive fields into a map of visual space).

vector The amplitude and direction of a movement.

ventricle One of a system of interconnected cavities in the brain that is filled with cerebrospinal fluid.

visuotopy The topographical representation of the retina, and hence visual space, in the brain.

The superior colliculus plays its primary role in generating and controlling orientation behaviors. It is best known for its role in initiating eye and head movements (collectively referred to as the control of gaze) made for the purpose of bringing an object of interested onto the fovea, the region of the retina that offers the greatest spatial acuity.

I. INTRODUCTION

The name superior colliculus (SC) is derived from the shape and location of this structure on the surface of the midbrain: ''colliculus'' because it reminded early

neuroanatomists of a small hill (the root "collis" is the Latin term for hill), and "superior" because it is found immediately above a second rounded structure, the inferior colliculus. Although most often associated with the visual modality, the SC receives information from multiple senses (visual, auditory, and somatosensory) and, like most structures in the central nervous system, it is paired: There is one SC on each side of the brain and each receives information primarily about contralateral space. Simply stated, sensory inputs originating from the left reach the right SC, which in turn generates leftward orienting movements. Similarly, the left SC receives input from right space and accordingly produces orientation shifts to the right.

Although a great deal has been learned about how the SC processes the sensory information and initiates the motor responses that comprise its behavioral role, overall views of its impact on overt behavior, its intrinsic organization, the functional properties of its constituent neurons, its interactions with other brain structures, and its evolutionary transition from an earlier premammalian form have changed significantly as a result of the intense scrutiny that has been directed toward this structure during the past few decades. It seems unlikely that current concepts of SC organization and function will change substantially in the future (and this will almost certainly be reflected in some of the categorical statements made in this article), but a note of caution is appropriate here because this also may have been the view of earlier researchers.

II. EVOLUTIONARY CONSIDERATIONS

In terms of the evolutionary history of the brain, the SC is a very old structure. Its nonmammalian homolog is the optic tectum, a structure that antedates the development of primitive mammals and the appearance of the neocortex. The optic tectum, like the SC, derived its name in part from its position in the brain (Fig. 1). The word "optic" pays heed to its massive retinal input, and the word "tectum," Latin for "roof," refers to the fact that it covers a large ventricle in the midbrain. Given the close homology between the optic tectum and the SC, it is not exactly clear why it is necessary to refer to them by different names. Presumably, part of the explanation lies in the fact that the SC does not overlie a large ventricle but rather a narrowed aqueduct between two ventricles and, thus, is not as rooflike in appearance as is the optic tectum. To add to the confusion, there are a number of general descriptive sources that continue to refer to the SC as the tectum. Sometimes there is no accounting for the oddities in anatomical nomenclature.

The optic tectum is considered the principal visuomotor structure in the central nervous system of nonmammalian vertebrates, and an excellent example of its sensory and motor roles is apparent in the prey-catching behavior of frogs and toads. This behavior can be readily evoked in alert animals by stimulating the structure through permanently implanted microelectrodes. The electrical stimulus mimics the effects of a natural sensory stimulus and elicits a complete prey-catching sequence. The behavior has a number of very specific components: The frog or toad turns toward the presumptive prey in contralateral space, flicks out its tongue to capture the prey, grasps this illusory prize with its jaws and brings it into its mouth, gulps, and then finally cleans its snout. Electrophysiological studies of the responses of individual frog or toad tectal neurons have shown that they respond best to the same kinds of stimuli that elicit natural prey-catching responses: dark objects that approximate the size of prey animals and move at appropriate speeds. These studies also indicate that the overall organization of tectal neurons, as well as their individual response properties, is well suited to the task of coding stimulus location and, via connections with motor structures, they are involved in initiating the motor programs through which the prey is captured. The stereotypical nature of tectal-mediated behaviors in the adult frog and toad, and their comparative insensitivity to alteration through learning, have made the anuran optic tectum a convenient and productive neural model for studies of visuomotor function. Behavioral and electrophysiological studies in other nonmammalian vertebrates, such as salamander, iguana, snake, pigeon, and barn owl, have collectively enhanced our knowledge of the stimuli that gain access to this circuitry and to the behaviors that it mediates.

Many of the characteristics of the optic tectum were altered during the evolutionary transition to mammalian forms nearly 200 million years ago. It was during this time that a host of radical changes were taking place in much of the brain that is believed to have characterized the premammalian reptile. The most prominent of these changes was the emergence of a neocortex (hereafter referred to simply as cortex). New capacities began to appear that were dependent on this new cortex and that were greatly elaborated with the appearance and increasing encephalization of new and more complex species. Brain circuits were redesigned, thereby transforming the sharing and/or localization

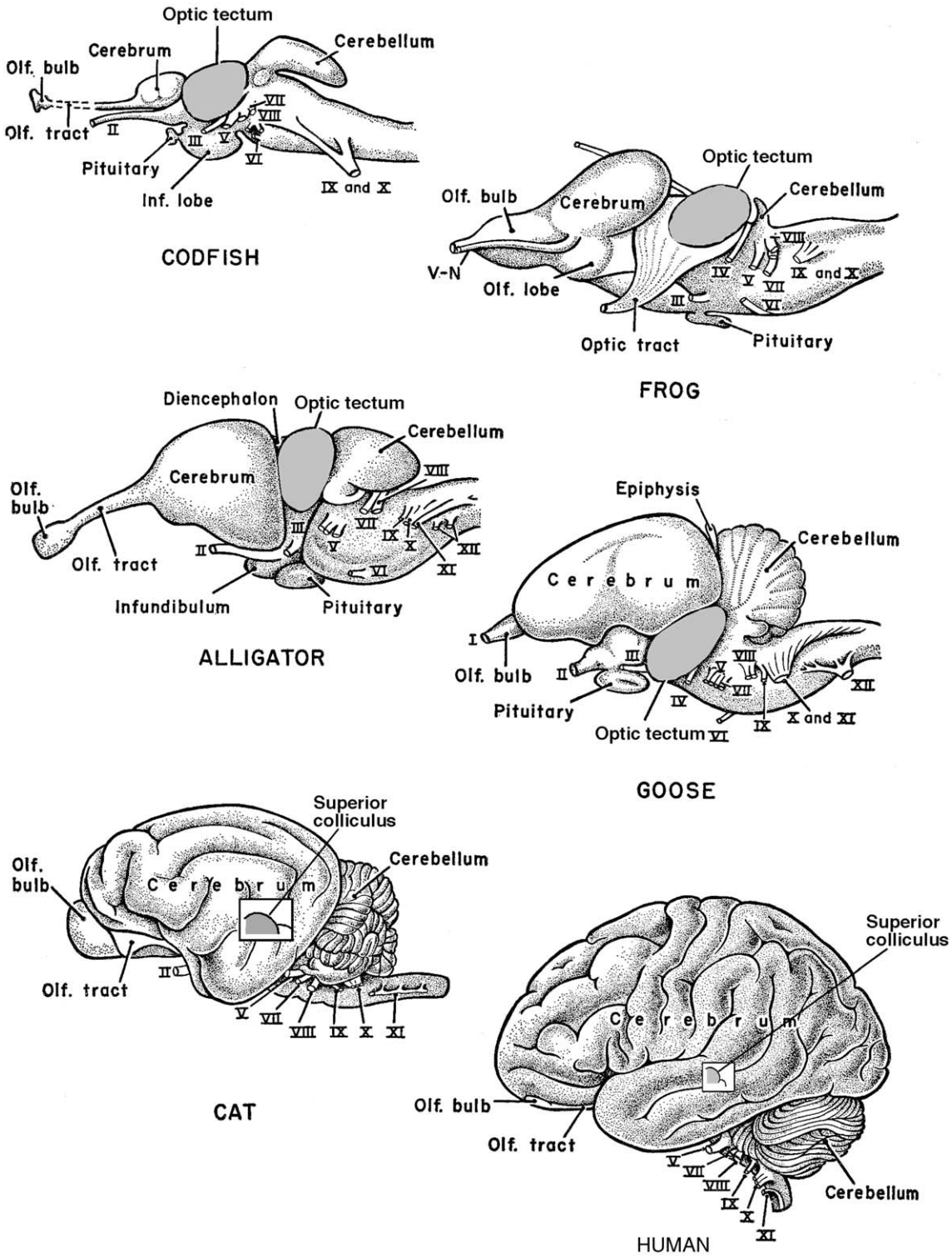


Figure 1 The optic tectum and superior colliculus in a variety of species. Note that in cat and man, an illustrative "window" is cut into the cortex in order to view the underlying superior colliculus (adapted with permission from Truex and Carpenter, 1964).

of older functions. The midbrain did not escape these evolutionary alterations. Not only did the anatomical appearance, relative size, lamination pattern, and internal organization of the optic tectum change but also its preeminent role in vision declined and some of its functional roles were altered and/or shifted to geniculostriate and extrastriate cortical systems. Although the stereotyped and complete behavioral sequences that were coded in the optic tectum (and released by an electrical stimulus) were lost during the evolution to the SC, its role in coordinating sensory inputs with motor commands was retained and adapted to changes that were taking place in peripheral sensory and motor organs. Indeed, at least some of the visual response properties of optic tectum neurons that served its functions so well are also recognizable in the properties of SC neurons that are described in this article.

III. CHANGING CONCEPTS OF THE ROLE OF THE SUPERIOR COLLICULUS

For some time it was believed that the role of the SC was restricted to reflex orientation of the eyes in response to visual stimuli, and some modern medical dictionaries still refer to it simply as the center of visual reflexes. The heavy direct retinal input to the SC and its homology with the optic tectum guaranteed that its visual role would be emphasized (and that its nonvisual inputs would receive considerably less early attention), and the recognition of its role in eye movements was a consequence of neurological studies in anesthetized animals that began more than 100 years ago. In these early studies investigators employed direct and localized electrical (and later chemical) stimulation to a host of brain structures. The idea was to determine the functional significance of these areas on the basis of the overt responses that could be evoked by directly activating them, a motivation shared by modern researchers interested in frog and toad tectum. Stimulation of the SC in this way produced rapid, contralateral movements of both eyes and began a long tradition of studies designed to determine how the activity of SC neurons is triggered by natural stimuli and how its neurons code the metrics of an eye movement or shift in gaze. The more sensory aspects of visual function, such as the ability to provide the subjective sensory experiences that are peculiar to this sensory modality or the ability to discriminate the form, pattern, or identity of a visual stimulus, were thought to be a consequence of activation of the geniculostriate system: The projection from the retina to the lateral geniculate nucleus in the thalamus and from there to primary visual cortex (also known as "striate" cortex).

Views of the role of the SC began to change in the mid-to-late 1960s when it was found that animals that had the SC removed exhibited a striking sensory neglect. An animal with the SC damaged on one side behaved as if it were unaware of, or totally disinterested in, sensory stimuli in its contralateral sensory field, even though its geniculostriate system was intact (Fig. 2). Although this deficit was not restricted to vision (auditory and somatosensory defects were noted as well), the visual impairment was the most prominent and received the greatest attention. In large part this was because the visual deficit did not fit the established dogma of how visual function was supposed to be segregated in the brain, and this provoked considerable surprise. Animals with SC lesions had no obvious motor impairments and could easily move about a room without banging into obstacles. However, they did not react to stimuli in the contralateral visual field, even those that were clearly rewarding or threatening. Although, given enough time after the lesion, these deficits were gradually ameliorated, presumably as a consequence of compensation through changes in other brain areas, contralateral visual stimuli never regained their prelesion effectiveness. Similar observations were made in rodents and primates. Experiments in which the same animals were tested on different visual tasks after experiencing damage limited to either

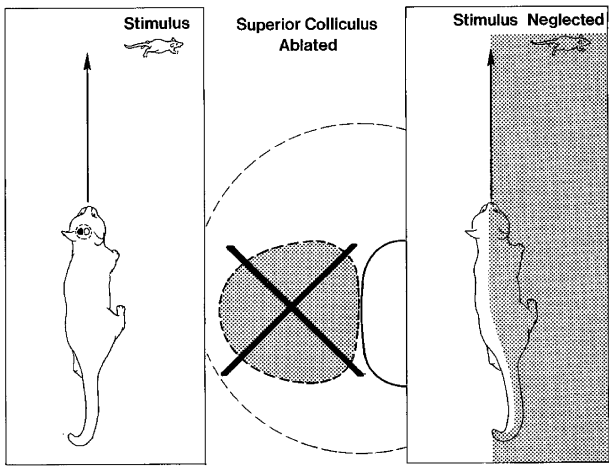


Figure 2 Ablation of a single SC results in contralateral sensory neglect. In this example, ablation of the left SC results in a neglect of stimuli in the right half of sensory space (reproduced with permission from Stein and Meredith, 1993).

the visual cortex on one side of the brain or to one SC yielded data consistent with the presence of two largely independent visual systems. The first of these, based in cortex, was involved primarily in evaluating the features of a stimulus. The second, based in the SC, was involved primarily in the initiation of an orientation to that stimulus. For simplicity, these were referred to as the "what" and "where" divisions of the visual system, highlighting the differential contributions of these regions to visual function (a dichotomy that has been revisited in terms of how visual information processing is now believed to be segregated in the primate neocortex).

The results of recent studies of the SC have tended to minimize this conceptual dichotomy and have stressed the anatomical and functional interdependence, rather than independence, of these structures. Without denying their differing roles, the SC and visual cortex are now believed to work cooperatively in order to ensure that a sensory stimulus is attended and scanned and its features are processed. The particular reliance of the SC on influences from the visual cortex has become evident from studies documenting the changes in SC-mediated behaviors in animals in which visual cortex has been damaged.

IV. INTERACTIONS BETWEEN VISUAL CORTEX AND THE SUPERIOR COLLICULUS: THE SPRAGUE EFFECT

One of the most striking examples of the delicate balance that normally is maintained among structures at different levels of the visual system was provided by James Sprague in 1966. Sprague used the cat's normal tendency to turn and examine a novel stimulus in order to map its visual field. He showed that when facing forward, the cat's visual field extended about 90–105° into peripheral space on either side of fixation. He then used this technique to evaluate the effects of large lesions of visual cortex on visually guided behavior. He found that large lesions that eliminated all visual cortex produced a profound contralateral visual neglect, similar to that produced by SC lesions. However, when he removed the SC on the opposite side, the defect was reversed (Fig. 3). Now the animal was able to react to stimuli in the previously "blind" hemifield but not to stimuli in the previously normal hemifield.

This finding became known as the Sprague effect and represents a dramatic demonstration of the

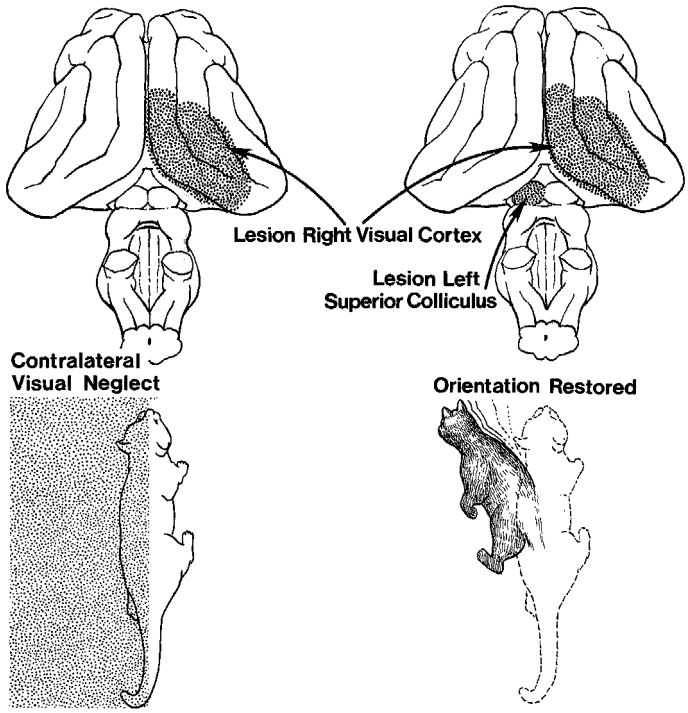


Figure 3 The Sprague effect. A large lesion of the right cerebral cortex, encompassing striate and extrastriate visual areas, results in a neglect of visual stimuli in the left hemifield. However, when this cortical lesion is paired with a lesion of the opposite (i.e., left) SC, orientation behavior is restored in the previously neglected hemifield (reproduced with permission from Stein and Meredith, 1993).

amelioration of a lesion-induced defect by another lesion (though a new defect is simultaneously induced). The explanation lies in the balance of excitatory influences the SC receives from the visual cortex on the same side of the brain and the inhibitory influences that are routed to it via the SC on the opposite side. Removing the cortical source of excitation allows SC neurons to become dominated by inhibition from the opposite side of the brain. These inhibitory inputs are derived from the basal ganglia, specifically from its output structure, the substantia nigra. The fibers carrying this inhibition pass through the SC on the opposite side, enter the intercollicular commissure, and finally synapse on their target neurons. Removing their inhibitory influences by removing the opposite SC or, as was shown later, by cutting the intercollicular commissure, restored a balance to the system, thereby freeing the SC to regain some of its function. Undoubtedly, its neurons did not recover all of the characteristics they had before visual cortex was removed, but enough of its neurons were functioning sufficiently well to support gross orientation behaviors.

At the heart of the Sprague effect is the importance of symmetry between the two halves of the brain. This interhemispheric balance appears to be crucial in some circumstances and may help explain the seemingly odd observation that removal of one SC produces a more pronounced visually guided defect than does removal of both. Apparently, it is easier for the brain to deal effectively with their combined loss than to compensate for the imbalances associated with one functional and one nonfunctional SC.

V. ANATOMICAL AND FUNCTIONAL ORGANIZATION OF THE SUPERIOR COLLICULUS

A. Laminar Organization

The SC is a layered structure consisting of seven alternating cellular and fibrous laminae. Its layering reflects the organization of its inputs, outputs, and interconnections, but it has generally been divided into two functional divisions: superficial (layers I–III) and deep (layers IV–VII). Sensory neurons in the superficial layers are responsive only to visual stimuli, whereas deeper layer neurons may be responsive to visual, auditory, or somatosensory stimuli (i.e., modality-specific neurons) or to inputs from two or three of these sensory modalities (i.e., ''multisensory'' neurons). Furthermore, the outputs of superficial layer neurons are directed primarily to sensory nuclei in the thalamus, whereas those of the deep layers project heavily to brain stem and spinal cord areas involved in motor responses. It is the deeper laminae that are most directly linked to the immediate overt behavioral responses mediated by the SC.

The possible behavioral differences served by the superficial and deep layer neurons were underscored in a study by Casagrande and colleagues in the 1970s in which selective lesions to the superficial layers in young tree shrews did not induce any obvious disruption of overt behaviors. Such overt behavioral defects were induced only by lesions that included the deeper layers. Extrapolations from the results of these experiments to adult animals were complicated by the well-known ability of the brain to compensate for early localized damage. Replications have not been forthcoming primarily because it is very difficult to produce lesions that are restricted to the superficial layers of the SC in any animal. There are unpublished observations suggesting that a superficial layer lesion in an adult cat that produced only minor damage in deeper layers did induce orientation defects. Although this must be replicated in other animals, the observation is consistent with findings suggesting that the superficial layer neurons can influence neurons in deeper layers via interlaminar connections. It is also possible that ascending projections from superficial layer neurons could exert their influence indirectly through a series of connections in thalamus and then to those cortical territories that target the deep SC. Though speculative, the heterogeneity of output pathways could allow superficial layer neurons to participate in the sensory discriminative functions carried out within thalamocortical systems as well as contribute to the orientation functions of the deep SC.

B. Visual Receptive Field Properties

Initial attempts to understand the properties of SC neurons were begun before superficial–deep layer distinctions had substantial impact on concepts of SC function. The objective of these early electrophysiological studies was to determine how the visual properties of SC neurons differed from those in other areas of the brain. Because visual neurons were most heavily represented in the superficial layers, these were the neurons most often examined. The general assumption guiding these studies was much the same as that guiding similar electrophysiological studies in other brain areas, specifically that those stimuli that

have the best access to the circuitry of the SC also have the best access to SC-mediated behaviors. In a quirk of science, initially the least amount of attention was devoted to the deeper layer neurons that were most relevant to the investigators' motivation, a situation that has since changed. Fortunately, however, superficial and deep layer neurons share many of the same visual receptive field properties, a surprising observation considering their different sources of visual information. Superficial layer neurons get heavy projections from the retina and from striate cortex, neither of which heavily target deeper layer neurons. In contrast, deep layer neurons receive their principal visual inputs from extrastriate cortical areas. Nevertheless, both superficial and deep layer neurons transform the signals they receive from their converging inputs into a set of response properties that distinguish them from their afferent sources, sometimes in subtle ways and sometimes in obvious ways.

By recording the activity of individual neurons, early investigators were able to identify these neuronal properties. The first step was to determine where a stimulus had to be placed in visual space in order to activate a given neuron (the neuron's visual receptive field). It was then possible to determine which visual stimuli, when presented within this region of visual space, would most reliably and most vigorously activate that neuron. The strategy was to use the same sorts of approaches that had proved so successful in the pioneering work of Hubel and Wiesel in the 1960s in which they studied the organization of visual cortex, a strategy that was also being used in studies of other geniculostriate structures. This entailed the presentation of a variety of simple visual stimuli whose characteristics were easily quantified and could be systematically varied to examine their effects on a neuron's responsiveness. Stationary flashed spots, edges, and bars of various sizes, intensities, orientations, and positions were presented within each neuron's receptive field, and similar stimuli were moved across the receptive field in different directions and at various speeds. The composite results describing the selective response properties for any given neuron were referred to as that neuron's receptive field properties. By studying many hundreds or thousands of neurons, it would be possible to determine the kinds of stimuli that were most effective in activating this population and then contrast them with the properties of neurons in other visual structures.

The receptive field properties of SC neurons proved to be unlike those of neurons in the retina, lateral geniculate nucleus, and the first stage of visual processing in striate cortex (i.e., the so-called simple cells). Retinal neurons and their target neurons in the lateral geniculate nucleus are readily activated by stationary flashed lights. Their receptive fields are characterized by a central core that is activated either by light onset or offset ("on" or "off") and a surrounding region with the opposite response ("off" or "on"). Larger stimuli within the "on" or "off" regions produce greater responses, but these different receptive field areas are mutually antagonistic so that when they are stimulated simultaneously they inhibit one another. A change in this receptive field structure takes place at the first stage of information processing in primary visual cortex, where the outputs of the lateral geniculate nucleus synapse on their simple cell targets. Although simple cells also have "on" and "off" regions that antagonize one another, these regions are now juxtaposed, forming an elongated receptive field for which angles and edges have now become the most effective stimuli. Although these properties are believed to be critical for the brain to ultimately construct a mechanism for detecting the angles, edges, and contours necessary to discriminate among shapes and patterns, these are not primary functions of the SC. SC neurons respond poorly to stationary flashed lights, and those that do respond to these stimuli do not show separate "on" or "off" receptive field subregions. Rather, they prefer stimuli that move across their receptive fields regardless of their shape or orientation. In addition, such neurons are tuned for the speed and direction of stimulus movement. Furthermore, unlike some of their geniculostriate counterparts, they always respond in transient fashion even to maintained stimuli. Thus, they seem adapted to signaling rapidly changing events.

Many of the receptive field properties of SC neurons are dependent on influences derived from cortex. Removal of cortex disrupts the ability of many of these neurons to respond selectively to specific directions of movement and decreases the proportion of binocular SC neurons. However, it does not change the transient nature of SC responses or their ability to signal the presence and location of a stimulus.

C. Auditory and Somatosensory Response Properties

The deep layers of the SC, unlike the superficial layers, receive considerable projections from auditory and somatosensory structures. The nonvisual signals

carried by these input systems are capable of reliably activating many SC neurons, and they can also initiate gaze shifts and other orientation movements via the circuitry of deeper SC. Many of these neurons respond to inputs from more than one modality, and they, like multisensory neurons elsewhere in the brain, have the remarkable ability to synthesize cross-modal inputs and thereby enhance (or degrade) their responses and the behaviors that depend on these responses. Multisensory neurons in the SC have served as a general model for understanding the principles governing this fundamental capability of the brain. However, because the mechanisms by which this multisensory integration takes place and the behavioral consequences of this integration are considered in detail elsewhere in this encyclopedia, this aspect of SC function is not discussed here.

Auditory-responsive neurons are well represented in the SC, and their organization is quite distinct from that found along the auditory thalamocortical or primary auditory projection pathway (the auditory equivalent of the geniculostriate system). Whereas the neurons in the thalamocortical system are specialized to represent the frequency content of stimuli, SC neurons do a poor job of discriminating stimulus frequency. Most auditory-responsive SC neurons respond best to complex sounds containing a broad range of frequencies, such as the sounds made by the voice, an object moving, the rustling of paper, a click, or a hiss. They are not particularly good at coding the intensity of the stimulus; they will respond with only a brief train of discharges even if the stimulus is maintained, and they respond well to a moving stimulus. Somatosensory-responsive neurons are also best activated by moving stimuli—in this case, stimuli moving across the hairs and/or skin. They, too, respond with a transient response even if the stimulus is maintained, and they can code the velocity of the stimulus.

Compared to visual receptive field properties, the nonvisual receptive field properties of SC neurons are less dependent on cortex. For example, removal of the somatosensory cortical areas that project to the SC has comparatively minor influences on these neurons, and these primarily involve changes in receptive field size. Similarly, removal of the most relevant areas of auditory cortex raises the threshold of auditory-responsive neurons in the SC but does not appear to alter their fundamental stimulus preferences.

The particular sensitivity of SC neurons to moving stimuli and their ability to code the parameters of movement, such as speed and direction (most evident in visual neurons), intuitively seem well suited for their role in detecting and orienting to sudden changes in the environment—changes that may signal important events for the animal and, therefore, should result in their examination. The presumptive importance and/or novelty of the stimulus are particularly significant in this context. Innocuous environmental stimuli that have little apparent significance to the organism but are presented repeatedly quickly lose their ability to activate SC neurons, a process known as habituation. Visual, auditory, somatosensory, and multisensory neurons exhibit response habituation. However, in experiments in which a neutral stimulus such as a flashing light was coupled repeatedly with a reward so that the animal associated one with the other, the stimulus not only became more effective than it was before but also failed to induce habituation even after many repetitions. Apparently, learning and experience can play substantial roles in crafting the vigor of SC neuronal responses and their resistance to response habituation. There is reason to suspect that this flexibility depends, in part, on influences from the cortex. Cortical neurons are particularly well suited to altering their properties as a consequence of experience and have been shown to be capable of imposing acquired characteristics on neurons in their target structures.

D. Sensory Topographies

1. Visual

For the SC to play its role in guiding orientating behavior, its neurons must be capable not only of detecting important events but also of specifying their locations so that appropriately directed movements can be generated. The need to localize stimuli is not unique to the SC, nor is its solution to the problem: the construction of a map, or topography, in which space is represented in a systematic manner within a structure. The visual map reflects the fact that the outputs from neighboring retinal neurons (ganglion cells), whose receptive fields are in neighboring regions of visual space, project to neighboring neurons in their target structures. The visual map in the SC (and in other central structures) has been recreated from experiments in which the receptive fields of neurons are plotted along vertical electrode penetrations that are made systematically at different locations. Generally, a grid-like pattern of electrode penetrations is made in order to sample the rostral-to-caudal (front-to-back) and medial-to-lateral (in the right SC this

would be midline-to-right) axes of the structure. Then the clusters of receptive fields that are recorded in each electrode penetration, and that define the regions of visual space represented there, are compared. This reveals that the two axes of visual space are nearly perpendicular to one another: The central-to-peripheral axis of visual space is represented across the rostral-to-caudal axis of the SC, and the superior–inferior axis of visual space is laid out along the medial-to-lateral axis of the SC (Fig. 4). Consequently, the receptive fields of a neuron at the rostral pole of the SC and in the middle of the medial–lateral axis will have a receptive field representing the foveal or parafoveal region. Neurons medial and lateral to this neuron will have receptive fields higher or lower in visual space, respectively. This visual topography, or visuotopy, is seen in the SC despite the variety of visual input sources, a result of the fact that all its inputs are distributed according to the same topographic scheme.

Although this visuotopy represents a fundamental vertebrate plan, there are considerable variations in the plan that reflect species differences in visual behavior. For example, the maps do not represent a simple linear transformation of the retina but, rather, geometric distortions of the retinal sheet: Some regions of the retina capture greater areas and, hence, more neurons in the target structure than do others. Cats, monkeys, and humans make heavy use of the central retina in fixating and evaluating a visual stimulus, and they require the greatest resolution in this region. In each of these animals there is an expansion of that region of the map devoted to central vision so that the neurons representing the foveal and parafoveal regions have smaller receptive fields and proportionally more of them are involved in representing a stimulus in comparison to regions of the map devoted to more peripheral regions of space (Fig. 4). Another obvious species variation is whether one or both eyes have direct input to the structure and, thus, whether individual neurons have direct access to binocular cues. In nonmammalian vertebrates, each optic tectum receives inputs from nearly the entire retina of the eye on the opposite side and therefore represents almost exclusively contralateral visual space; little binocularity is achieved. In rodents, the eyes are set further forward on the head and the retinal inputs from the opposite eye are nearly, but not completely, crossed. A region of the rostral SC is now devoted to representing the area of visual space common to both eyes (the binocular region) and has neurons activated by signals originating in both eyes (i.e., binocular neurons). In cat and monkey the eyes are set in the front of head and there is a large region of binocular overlap. This is represented centrally by having the outputs from the half of the retina near the nose (nasal) crossing the brain to terminate in the opposite hemisphere. The temporal half of the retina keeps its outputs uncrossed and directed to structures on the same side of the brain. This provides the right SC with a ''view'' of the left visual field, and it also endows many of its neurons with direct binocular input and the possibility of stereopsis. The reverse situation is present in the left SC. The binocularity of SC neurons is enhanced by projections from binocular neurons in cortex.

2. Auditory

As noted previously, the thalamocortical auditory projection pathway does an excellent job of coding for frequency. Neurons with similar frequency selectivities are organized into ''isofrequency'' bands in auditory cortex, thereby producing a ''tonotopic'' map. However, no clear map of auditory space is present there. This is not surprising given the fact that the cochlea codes sounds in terms of their frequency, not their spatial location. Consequently, when the outputs of the cochlea project topographically to their target structures, the result is a tonotopy. In the SC, however, it is the coding of space rather than frequency that is critical, and SC neurons are organized to represent the locations of auditory stimuli topographically. This spatiotopic map is unique because it, unlike the visual (or somatosensory) map, requires a computation for its creation. The construction of an auditory map relies, at least in part, on a comparison of the sound as it arrives at the two ears. Differences in both relative timing and intensity are cues to the location of a sound. For example, sounds coming from the right arrive sooner and are louder in the right ear. These inter-aural comparisons are made by binaural neurons that reflect in their discharge rate the relative timing and/or intensity of sound at the two ears. Although not the primary site of binaural interaction, spatially restricted receptive fields in the SC are presumed to reflect the results of this neural computation. With the eyes directed forward in the head, one finds auditory receptive fields that are organized in a manner that parallels the visuotopy in the SC: Auditory neurons with receptive fields in central space are rostral in the SC and those with progressively more peripheral receptive fields are found at progressively more caudal sites; neurons with receptive fields in upper space are medial, and those with receptive fields in lower space

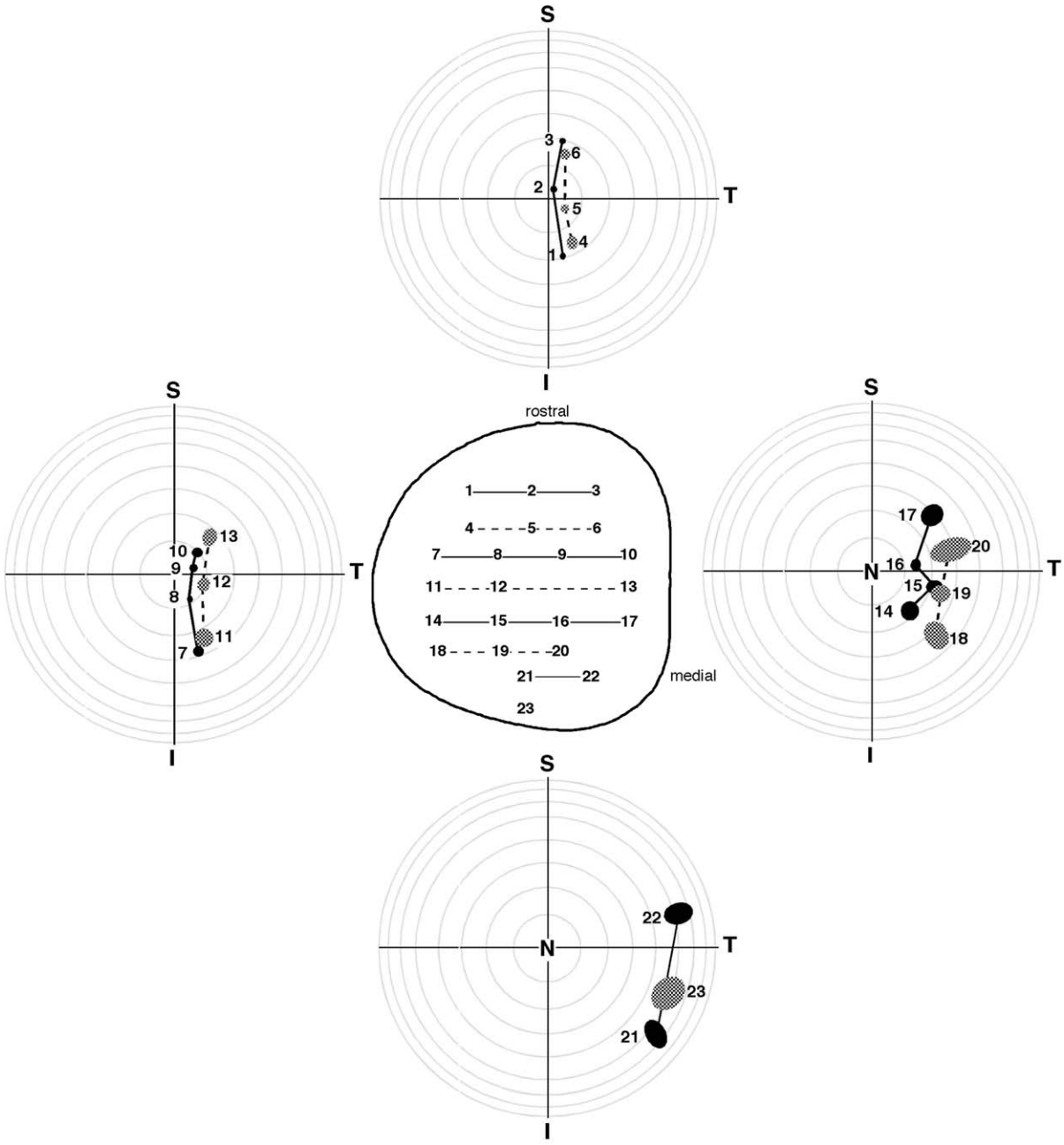


Figure 4 The SC is characterized by a topographic visual organization. In the center is shown a representation of a dorsal view of the left SC. The locations of a series of systematic electrode penetrations are plotted here. The corresponding receptive fields mapped in each of these electrode penetrations are shown on the four surrounding representations of visual space. In this standardized representation, the horizontal and vertical meridians are shown by the thick lines, and each of the concentric circles represents 10°. The size and location of receptive fields are shown by shading. Note the systematic shift of receptive fields from nasal (e.g., receptive fields 1–3) to temporal (e.g., receptive fields 21–23) as progressively more caudal regions of the SC are sampled. Similarly, note the superior-to-inferior shift as the sampling sites move laterally. N, nasal; T, temporal; S, superior; I, inferior (data from Wallace *et al.*, 1997).

are lateral (Fig. 5). As discussed in more detail later, the relationships among auditory, somatosensory, visual, and motor maps have important implications for the SC's role in the generation of orienting behaviors.

3. Somatosensory

The somatosensory representation follows the same general format, and its receptive fields are arranged in

a "somatotopic" map (Fig. 5). Again, given alignment of the eyes, ears, head, and body, one finds that this somatotopy registers with the visual and auditory topographies. Thus, neurons with receptive fields on the front of the body are located rostral in the SC and neurons with receptive fields progressively further back on the body (toward the tail) are located progressively more caudal in the SC. Similarly, upper portions of the body are represented medially and lower portions of the body are represented laterally in the structure.

Innocuous cutaneous stimuli have been found to be effective in activating many SC neurons in every species examined. However, based on early clinical observations, anatomical connections, and stimulation studies, early investigators speculated that the SC was also involved in processing information relevant to pain. One would therefore expect that some of its neurons would respond preferentially or solely to noxious stimuli. Confirmation of this speculation came in the late 1970s from experiments with hamsters. As one might expect, the neurons responsive to noxious stimuli had receptive fields, and later experiments showed that these receptive fields were organized into a map that paralleled the maps in each of the other sensory representations. Although these observations were confirmed in other rodents, they have not been reported in carnivores or primates.

It should be apparent from the previous descriptions that by aligning the individual sensory representations, a multisensory map is created. In principle, this simplifies considerably the process of localizing an event and activating the motor circuitry of the SC in order to initiate a response to that event. In fact, as discussed later, the SC sensory topographies are not only in register with one another but are also in register with the SC motor topography.

E. Motor-Related Function of the SC

Although electrophysiological studies of the sensory properties of SC neurons have a long history, the advent of similar studies to examine the motor-related activities of SC neurons has been much more recent. This delay was not due to a lack of interest in the motor properties of SC neurons but to the lack of a satisfactory means to study them. The technological advances that finally made it possible to examine the activity of SC neurons in association with a motor response in an awake behaving animal were a boon to motor physiology and led to a host of studies in awake behaving cats and monkeys. Although the term "motor" will be used here to describe the effector properties of SC neurons, some researchers prefer to describe these properties as "premotor" because SC neurons do not provide direct innervation to the muscles. Rather, they influence the final motor acts in which they are involved via direct or indirect connections to the motor structures of the brain stem and spinal cord. With this caveat in mind, it is fair to say that these studies have revealed that many SC neurons are both responsive to sensory stimuli and activated in a time-locked fashion before a motor response,

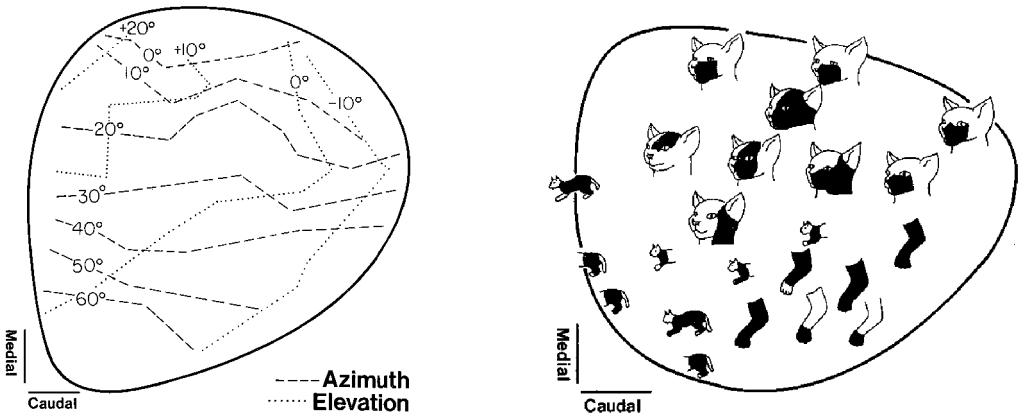


Figure 5 Auditory (left) and somatosensory (right) maps in the SC. In both maps, sensory space is depicted on a dorsal view of the SC. In the case of the auditory map, for each neuron studied the best area of its receptive field is related to its location in the SC. The map is derived from the raw receptive field data so that the dashed and dotted lines represent the progression in azimuth and elevation of receptive field best areas. Thus, auditory neurons with best areas centered on 60° of azimuth are found in the caudal SC, and those with best areas in inferior space or below (e.g., −10°) are found in the lateral SC. In the case of the somatosensory map, representative receptive fields, illustrated by shading on the figurines, are plotted directly onto the SC (adapted with permission from Stein and Meredith, 1993).

providing the impetus for the term "sensorimotor" to describe them.

Coupling sensory and motor properties in many SC neurons seems particularly appropriate in the case of visuomotor control, for which the SC is best known and in which it is most often involved, since visual cues are used continually to shift gaze in order to bring the most sensitive area of the retina (e.g., the fovea) to bear on an object of interest. However, historically, researchers have chosen to focus on either the sensory information processing capabilities of SC neurons or the manner in which they initiate and direct an overt motor response. This created two largely independent bodies of knowledge. The information gathered on the sensory organization and sensory properties of SC neurons (discussed previously) was most often obtained from anesthetized animals and emphasized the relationships between neural responses and the physical attributes of the stimulus. By comparison, studies that have dealt with SC motor function have more often been conducted in awake behaving animals, have employed more rudimentary sensory stimuli, and have focused more on the linkage between neural activity and the metrics of movements directed toward those stimuli. The vast majority of these studies have concentrated on saccadic eye movements (rapid movements of both eyes) or combined eye/head gaze shifts. Recently, however, there has been an emphasis on aspects of SC activity that resist the general classifications of sensory or motor, being attributed more often to mechanisms of attention, memory, target selection, movement selection, or motor intention—terms suggestive of more cognitive processes. These studies attempt to bridge the gap between the mainly sensory and motor studies of the SC and provide insights into the neural computations that are ongoing during the interval between the early sensory-evoked activity and later motor-related activity that precedes any goal-directed orientation. Thus, although this section deals primarily with the motor organization and motor-related properties of SC neurons, no discussion of the SC can be considered complete without some treatment of the more "cognitive" aspects of SC activity that have been the focus of many recent studies. Such a discussion will follow the consideration of the motor properties of the SC.

1. Motor Topographies

Unlike studies of sensory topographies in the SC, early efforts to reveal a systematic motor representation relied more on electrical stimulation than on single neuron recording. That the application of stimulating current to the SC can produce contralateral movements of both eyes has been known since the late 19th century. Although these and stimulation studies conducted in the 1940s demonstrated a topographic motor organization, the issue was revisited in the 1970s using modern methods of microstimulation and eye movement measurement. In his classic 1972 study, David A. Robinson systematically varied the position of a stimulating electrode within the SC of an alert monkey and demonstrated an orderly "motor map" wherein the amplitude of contralaterally directed saccades was represented along the rostrocaudal axis (small to large movement) and the upward–downward component of the movement was represented along the mediolateral axis of the SC (Fig. 6). Although first demonstrated in the monkey, analogous eye movement maps have been shown in the SC of other species (e.g., cat and rat). Furthermore, observations from a number of species have suggested the possibility of multiple motor maps in the SC. Indeed, a well-organized ear movement map has been demonstrated in the cat, and it has also been reported that stimulation of the SC can evoke movements of the limb (monkey), head (cat, rat, monkey, and owl), whiskers (cat and rat), and trunk (rat).

It is given that the motor topographies in the SC are tailored in a species-specific fashion (e.g., one would not find a detailed whisker representation in monkeys). With this in mind, it is interesting to note that despite early expectations to the contrary, the motor topographies controlling shifts of gaze in the two most intensely studied animals, cat and monkey, are quite similar despite the considerable phyletic differences of these species. Nonhuman primates, like humans, rely heavily on vision for maintaining an awareness of the environment. Accordingly, the range of ocular motility in monkeys rivals that of humans. In contrast, cats possess a more limited oculomotor range and consequently recruit head movements for all but the smallest changes in gaze. Unlike the cat SC, until very recently, the monkey SC was considered to be involved primarily in the generation of saccadic eye movements and not to be involved in producing combined eye/head gaze shifts. This notion of a significant species difference survived for some time without challenge because studies of nonhuman primates had traditionally been carried out in animals whose heads were restrained, thus precluding the possibility of evoking head movements with SC stimulation. However, in recent microstimulation and recording studies of monkeys, the head has been free to move, and these

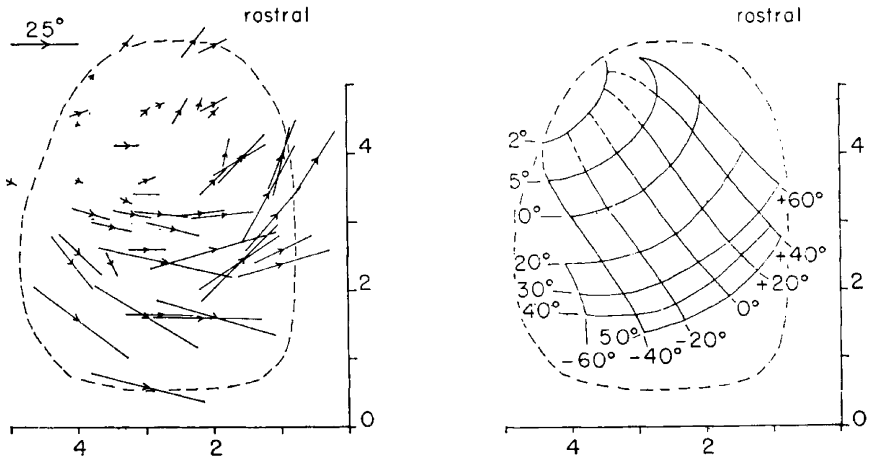


Figure 6 The motor map in the SC. (Left) Eye movements evoked by electrical stimulation at different SC sites are depicted by vectors (length represents amplitude and arrow orientiation represents direction) plotted onto a dorsal view of the SC. (Right) The data have been transformed to generate a map of eye movement amplitude and direction. Amplitude is represented from rostral (top) to caudal (bottom) and direction from medial (right) to lateral (left) within the SC. From rostral to caudal within the SC, the representation is from small movements to large movements. From medial to lateral, the representation is from movements with large upward (direction = $+60°$) components to movements with large downward components (direction = $-60°$). Thus, for example, microstimulation at the intersection of the $10°$ isoamplitude line and the $+20°$ isodirection line would evoke an eye movement of $10°$ in amplitude and inclined $20°$ upward (adapted with permission from Robinson, 1972).

studies have shown that the assumed differences between cat and monkey were, in large part, more apparent than real. Thus, although the relative contributions of eye and head movement to producing a gaze shift vary across species as well as circumstances (e.g., the position of the eye in the orbit at the start of the gaze shift), there seems little doubt that the motor map in the monkey SC, as in the cat, codes for changes in gaze and not simply changes in eye position.

Regardless of the specific kinds of movements one considers, it is thought that there are fundamental commonalities in the physiological characteristics of the neurons responsible for initiating an action. However, because the overwhelming amount of information available about the motor organization of the SC has been gathered in studies of eye and eye/head movements, this will be the focus of the following discussion.

Early studies established that an SC neuron that discharges in association with saccadic eye movements has a movement field. A movement field refers to the set of all saccade vectors (amplitudes and directions) that the neuron's activity contributes to producing. Consequently, SC movement-related neurons were found to discharge maximally for movements of a particular amplitude and direction (i.e., its preferred vector) and progressively less vigorously as a function of increasing deviation from the preferred vector (Fig. 7). Not surprisingly, experiments that have combined microstimulation and single-neuron recording have established that the vector of an electrically evoked movement agrees well with the preferred vector (and thus the movement field centers) of the neurons in the stimulated region.

2. Population Coding in the SC: Vector Averaging

Motor-related neurons are thought to provide a command signal to move the eyes and/or head. This signal plays a part both in initiating and in specifying the vector of a movement. Accordingly, saccade-related neurons are characterized by a high-frequency burst of action potentials that begins approximately 20 msec before the movement. Historically, the SC stands out as a model for understanding how information is represented within a topographic map. As discussed later, not only is it clear "what" the SC motor map encodes but also inferences can be made about "how" this information is "read out" by downstream circuits. In a series of elegant experiments, pharmacological blockade of small areas of the SC motor map provided significant insights into these issues and unexpectedly changed established views of how a motor map operates. These experiments demonstrated that the vector of SC-mediated saccadic eye movements is determined by the weighted *average* of activity across the motor map rather than by the activity of any

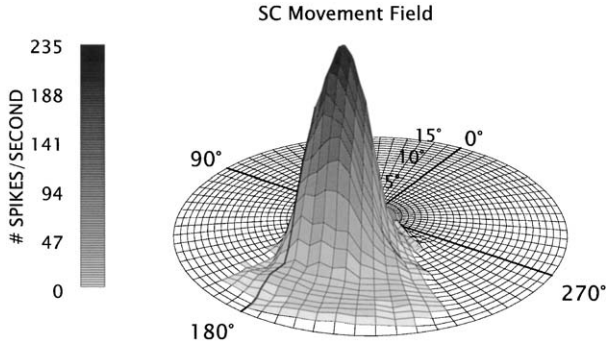


Figure 7 Motor-related neurons in the SC have movement fields that are defined by the relationship between the strength of motor-related activity and the vector (amplitude and direction) of movement. The 3D polar plot depicts a movement field for one saccade-related neuron in the right SC. Saccadic eye movement directions from 0 to 360° are represented on the graph. As a matter of convention, the 0 and 180° spokes represent saccadic eye movements that are straight right and left, respectively. Likewise, the 90 and 270° spokes represent movements that are straight up and down, respectively (note, however, that the plot is rotated and tilted for display purposes). Saccade amplitude is represented as the distance from the center of the plot (0 to 15°). Thus, each grid intersection (intersection of the amplitude and direction axes) corresponds to a vector, and the level of activity associated with making a saccade of that vector is given by the height of the plot at that point. For this particular neuron, most of the activity is associated with movements near 180° but biased toward the lower quadrant (i.e., toward 270°). Thus, this neuron discharged most vigorously for slightly down and left movements, with the peak rate corresponding to an amplitude of approximately 5–7°. Activity declines gradually for saccades that deviate from this vector (from Stanford and Sparks, unpublished).

restricted group of neurons within the topographic map. Although this conclusion seemed counterintuitive at first, it was the most reasonable conclusion that could be drawn from experiments in which deactivating a circumscribed region within the map did not render a trained monkey incapable of producing saccades that matched the preferred vector of neurons at the center of the now inactive population. In fact, these saccades were quite accurate, although they were slower to begin and, because of their lower overall velocity, took longer to reach their final position. That the appropriate vector can be computed in the absence of what would normally be the most active neuronal population indicated that the influences of surrounding neurons (i.e., those maximally responsive to different vectors) are averaged to produce the actual movement. On the other hand, the loss of the contribution of the deactivated population in the computation of other, *nonpreferred* vectors leads to a systematic pattern of movement errors, and this is also consistent with an averaging operation (Fig. 8).

It is typically stated that the *site* of activity within the SC motor map codes for the movement vector (e.g., of a gaze shift). Although true in a sense, it is also an incomplete explanation of how the movement vector of a gaze shift is programmed by the SC. The deactivation experiments referred to previously illustrate that there really is no singular correspondence between the *site* of activity in the SC and the actual movement vector produced because a vector-averaging scheme dictates that equivalent vectors can result from any number of activity patterns across the SC motor map. Thus, for example, gaze shifts of comparable metrics can be produced by activity tightly constrained around a particular locus in the SC, which we can refer to as A, or by a broad symmetrical swath of activity with point A at its center. More precisely, the desired change in gaze is represented by the spatial distribution of activity across a substantial fraction of the SC motor map. In this scheme, the contribution of any given neuron (whose preferred movement vector is determined by its position in the map) to the overall average (i.e., the ''weight'' of its contribution) is determined by the vigor of its discharge.

Interestingly, in the past decade there has been growing interest in understanding how the activity of the SC contributes to the time course of a gaze shift. Several proposals have been put forth in this regard. Although specific schemes differ substantially in detail, many have hypothesized that SC activity determines the moment-by-moment progress of a gaze shift by placing the SC within a feedback loop. The feedback loop compares the desired final gaze position with the current gaze position so that the appropriate signal can be generated to complete the movement. In other words, it is thought that at any given moment, activity in the SC reflects the difference between the current and desired gaze displacement, a ''motor error'' that must be progressively reduced in order to achieve the final desired gaze position. Some studies have suggested that somehow the nervous system represents motor error by the progressive changes in the peak *position* of activity on the SC motor map—a message that is predominantly spatial in nature. Others have suggested that the key index is the moment-to-moment changes in the firing rate of SC neurons during the gaze shift—a message that is predominantly temporal in nature. In neither case, however, has a causal relationship between SC activity and the progress of a gaze shift been established. As might be expected, this remains a topic of debate.

However, there seems to be little debate that the speed of a gaze shift depends on the overall vigor of SC

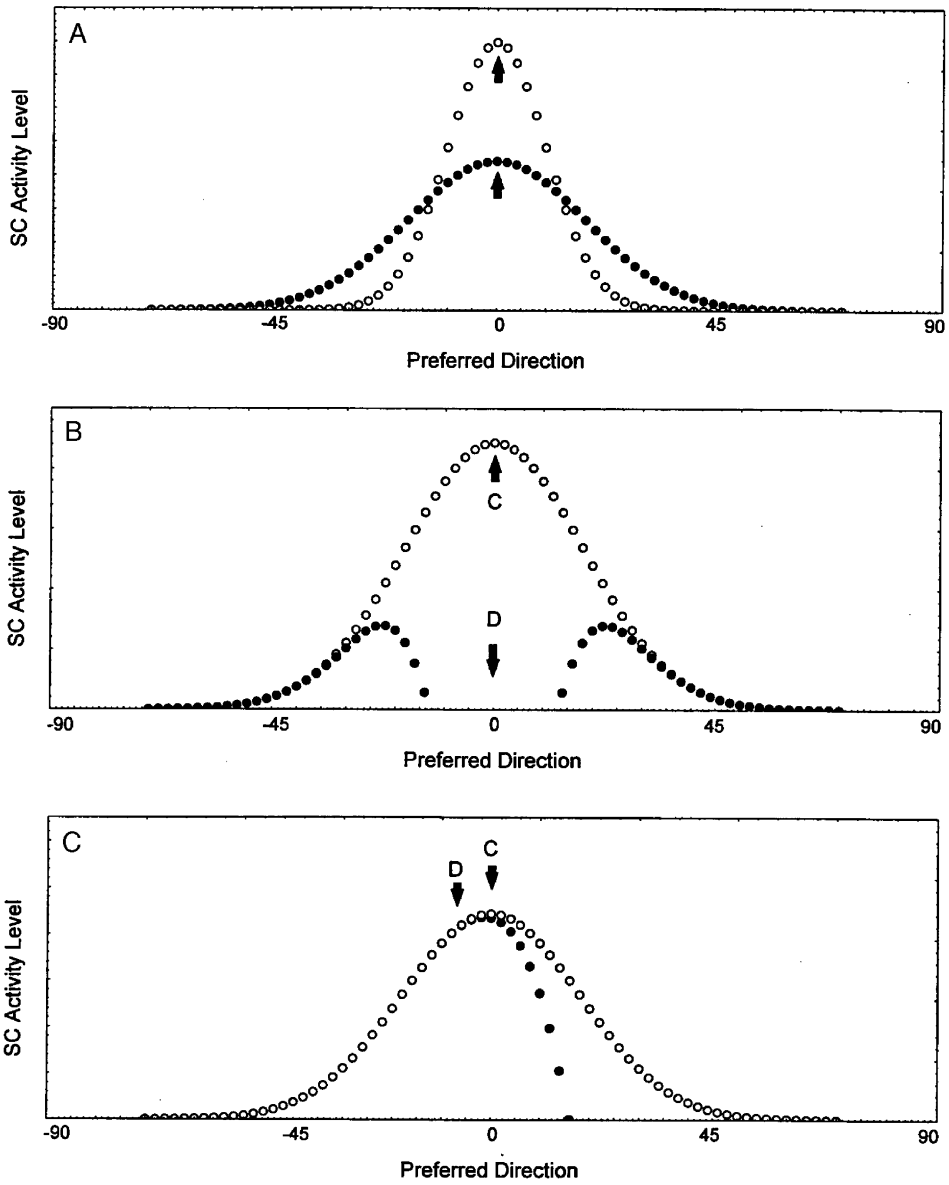


Figure 8 The vector of a saccadic eye movement or gaze shift is computed by averaging the contributions of active neurons within the SC motor map. To illustrate this concept, we show how hypothetical profiles of SC activity contribute to the computation of saccadic direction. For example, in A two different hypothetical activity profiles are shown. Consider each curve as representing the spatial distribution of activity across a lateral to medial slice through the SC (i.e., parallel to the direction axis). With this in mind, adjacent dots represent the activity of adjacent neurons (each with a slightly different preferred direction) within the motor map. Thus, in A, both activity distributions are centered about neurons having preferred directions of 0° (purely horizontal movements) with the activity profile declining symmetrically in both directions across neurons having progressively more upward (positive values) and downward (negative values) preferred directions. The key point of A is to illustrate how very different activity profiles (open and solid symbols) could result in a saccade of similar metrics. A vector-averaging algorithm for computing the direction of movement predicts an eye movement of 0° in each case. (Computed eye movement direction corresponds to the mean of each distribution, as indicated by an arrow in each case.) (B) A hypothetical deactivation experiment. Deactivation eliminates activity at the center of the population of SC neurons that would be active for a purely horizontal saccade (●). Nevertheless, computing the direction of the eye movement by averaging the contributions of the remaining active population of surrounding neurons predicts a saccade of the same direction (arrow D) as in the control condition (O; arrow C), (C) A hypothetical deactivation experiment that preferentially eliminates the activity of neurons with upward preferred directions (●). Loss of the "upward" neurons to the population average would lead to a "downward" bias on an attempt to make purely horizontal saccades. D, deactivation; C, control.

motor activity. The deactivation studies described previously are consistent with a positive relationship between the overall level of SC activity and the speed of the gaze shift that is produced. For example, local inactivation of regions of the SC, which reduces the overall average of activity in the SC, results in slower saccades. Similar results are obtained in electrical stimulation studies when the level or frequency of the stimulation is reduced. It is important to emphasize that changes in speed can occur without differences in movement vector. Consequently, a rule of thumb regarding an eye movement is that the spatial distribution of SC activity determines its vector, whereas the level of SC activity determines its speed.

3. SC Sensory Information Is Represented in Motor Coordinates

As discussed previously, some SC neurons discharge in response to sensory stimuli and some discharge in association with movements made for the purpose of orienting to sensory stimuli. However, many of the same neurons that respond to the onset of a sensory stimulus are later active in association with a movement directed toward that stimulus. Thus, the SC is at the interface of sensory and motor processing.

At first glance, the process of translating an incoming sensory signal into an outgoing motor command seems straightforward. Because the SC possesses both sensory and motor topographies, it is reasonable to assume that a simple point-to-point correspondence between the sensory and motor maps could support an accurate sensory-to-motor transform. This is certainly easy to understand if one considers gaze shifts to visual stimuli. Thus, for example, a visual stimulus 10° to the right (relative to the fovea) would require a gaze shift of 10° to the right to bring the image of the target onto the fovea. Because the fovea is used as a point of reference in each case, there is agreement between the position of a stimulus in retinal coordinates (retinal error) and the distance and direction of a gaze shift required to look at it (motor error); each is 10°. This reasoning formed the basis of the foveation hypothesis, a proposed mechanism for how the SC transforms a visual signal into a motor command for a saccadic eye movement. According to this scheme, visually driven activity among neurons in the topographically organized superficial layers leads to excitation in deeper layers where motor-related neurons are found. Although attractive in its simplicity, the foveation hypothesis cannot account for some recently demonstrated aspects of saccade programming. In particular, behavioral studies have shown that saccades can be accurate even in the absence of nominal agreement between retinal error and motor error. For example, saccades can be directed accurately to the *remembered* location of a previously flashed visual stimulus, even if an intervening saccade (e.g., the eyes are moved 5° upward) alters the required vector so that it no longer matches the original retinal error. To do so accurately requires that the intervening movement be taken into account and precludes a direct relationship between retinal error and initial motor error. Although dissociation between retinal error and motor error may not be the norm for visually guided saccades, these experiments demonstrate that saccades are not programmed on the basis of retinal coordinates. Rather, at the time (or before) the SC motor-related neurons become involved, the information about stimulus location has been transformed from sensory coordinates (e.g., retinocentric) to motor coordinates (oculocentric).

The issue of whether information is represented in a sensory versus a motor frame of reference in the SC is more readily appreciated when considering a *nonvisual* modality. For example, consider the problem of making a saccadic eye movement to an auditory or somatosensory target. Unlike the axes of visual space (centered on the retina), which are yoked to the position of the eye, the axes of acoustic space (head centered) and those of tactile space (body centered) are independent of eye position. Thus, there is no unique relationship between the position of a stimulus in either auditory or somatosensory coordinates and the vector of a movement that would be required to look at it. For example, the amplitude and direction of a gaze shift required to look at a stimulus on the forearm (or forepaw) depend on many factors, including the relative positions of the limb, the head, and the eyes. Similarly, the gaze shift required to look toward an auditory stimulus changes as a function of the position of the eyes in the orbit. Knowing the location of an auditory stimulus with respect to the head or a tactile stimulus with respect to the body surface (e.g., this knowledge is in terms of sensory coordinates) is not sufficient information to program a gaze shift to look at the stimulus; one needs to know where the stimulus is with respect to the current position of gaze. In other words, the referent must be eye centered or oculocentric. In fact, studies have shown that both auditory and somatosensory information in the SC have been translated from their native reference frames to a common motor coordinate frame. Rather than represent the position of a stimulus in auditory or somatosensory space, auditory- and

somatosensory-evoked activity represents stimulus location with respect to the current position of gaze.

F. A "Cognitive" Superior Colliculus?

In preceding sections, we focused on the sensory-related and motor-related aspects of SC function. In particular, we discussed how the SC represents information about sensory stimuli and how activity within the SC contributes to the programming of movements to orient to these stimuli. It may strike the reader that the majority of our discussion has been devoted to how the SC represents and guides simple actions to very simple stimuli. Whereas simple experimental paradigms afford the stimulus and behavioral control essential to addressing many fundamental issues, all would agree that the task of looking toward a bright spot of light in an otherwise dark environment does not approach the contextual complexity of orienting to stimuli in a natural setting. For example, given a complex scene, what factors dictate which of many potential targets becomes the goal of an orienting movement? Clearly, the choice of action is dictated by both the physical attributes of stimuli and the current set of circumstances (e.g., behavioral objectives). Recently, there has been a great deal of interest in investigating the neural basis of the decision processes that lead to a choice of a particular action among competing alternatives. Although the majority of these studies have focused on cortical structures, several have sought to determine if and how activity in the SC might reflect such higher order processes.

Although we briefly consider some of the recent developments concerning the influence of cognitive factors on SC activity, the fact that experimental context can modulate the responsiveness of SC neurons has been known for decades. Relatively early in the course of recording from awake behaving monkeys, it was noted that some SC neurons responded more vigorously to a visual stimulus if the stimulus was the anticipated goal of a saccadic eye movement (as opposed to being inconsequential to correct/rewarded performance on a behavioral trial). Although a clear-cut interpretation of this so-called response enhancement is difficult owing to the fact that many cognitive factors could be at play (e.g., anticipation of reward, allocation of spatial attention to a predictable location, and advance preparation of a predictable response), the important point is that the activity of these neurons was the joint product of the physical stimulus and behavioral context. In general, recent studies attempting to examine issues related to cognitive processes have used behavioral tasks designed specifically to manipulate behavioral context independently of sensory input or motor output.

As discussed in preceding sections, SC neurons may have activity that is temporally linked to the onset of a sensory stimulus and/or to the onset of a motor event (e.g., saccadic eye movement). Thus, temporal correlation is the main criterion for classifying activity as either sensory contingent or motor related. The focus of many cognitive studies has been on activity that cannot be so classified using a simple temporal criterion. In contrast to the transient, often high-frequency discharges closely associated with the occurrence of sensory or motor events, many SC neurons demonstrate a sustained, lower frequency discharge during the period that intervenes between early sensory and later motor-related bursts. Modulation of this low-frequency activity by contextual factors has been shown in a number of studies. Briefly, activity during the period has been linked to the maintenance of a stimulus in spatial memory, to the allocation of spatial attention, to the discrimination between relevant and irrelevant stimuli, to the likelihood of making a particular movement, and to the current state of response readiness. From these and other studies it is clear that processing within the SC is neither purely sensory nor purely motor, but the product of measurable events (e.g., stimulus onset and movement onset) and less tangible (and less readily quantified) internal factors.

Interest in the influence of cognitive factors on activity in the SC is a comparatively recent development and there remain many unsettled issues. As noted previously, low-frequency SC activity has been related to numerous covert processes. As noted by others, at least two caveats must be considered along with these findings. First, many of the factors mentioned previously are not mutually exclusive and are bound to covary in most experimental contexts (e.g., spatial attention and response readiness). Second, unlike the case for motor output, it is much more difficult with covert processes to establish a causal linkage between SC activity and overt behavior. With this in mind, it may be difficult to establish whether the SC activity is a significant contributor to the decision processes that lead to a specific action or simply reflects the outcome of computations carried out in higher centers.

VI. CONCLUDING REMARKS

The foregoing discussion has touched on many of the issues germane to the organization and functional role

of the SC. The history of research in this area has had obvious phases in which issues most closely tied to its sensory or motor roles appeared to dominate, and it is not difficult to predict that in the immediate future much of the SC scholarship will mirror the recent trend toward issues falling under the general heading cognitive neuroscience. The continued interest in this structure is prompted, in part, by its critical involvement in the immediate reactions to sensory stimuli that are so critical to survival and by the elegant simplicity of its sensory and motor roles. The SC is designed to process sensory cues so that motor commands can be generated to orient the organism to the very cues that began the process. Determining the degree to which the SC participates in the decisions that make these motor actions both timely and appropriate for a given circumstance is likely to remain a topic of interest for the foreseeable future.

See Also the Following Articles

EVOLUTION OF THE BRAIN • EYE MOVEMENTS • INHIBITION • MIDBRAIN • MULTISENSORY INTEGRATION • RECEPTIVE FIELD • RETINA • VISUAL AND AUDITORY INTEGRATION

Suggested Reading

Casagrande, V. A., Harting, J. K., Hall, W. C., and Diamond, I. T. (1972). Superior colliculus of the tree shrew: A structural and functional subdivision into superficial and deep layers. *Science* **177,** 444–447.

Gazzaniga, M. S. (Ed.) (2000). *The New Cognitive Neurosciences*. MIT Press, Cambridge, MA.

Hubel, D. H., and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Physiol. (London)* **195,** 215–243.

Kuffler, S. W., Nicholls, J. G., and Martin, A. R. (1984). *From Neurons to Brain*. Sinauer, Sunderland, MA.

Lee, C., Rohrer, W. H., and Sparks, D. L. (1988). Population coding of saccadic eye movements by neurons in the superior colliculus. *Nature* **332,** 357–360.

Robinson, D. A. (1972). Eye movements evoked by collicular stimulation in the alert monkey. *Vision Res.* **12,** 1796–l808.

Schall, J. D., and Bichot, N. P. (1998). Neural correlates of visual and motor decision processes. *Curr. Opin. Neurobiol.* **8,** 211–217.

Sparks, D. L. (1986). Translation of sensory signals into commands for control of saccadic eye movements: Role of primate superior colliculus. *Physiol. Rev.* **66,** 118–171.

Sparks, D. L. (1999). Conceptual issues related to the role of the superior colliculus in the control of gaze. *Curr. Opin. Neurobiol.* **9,** 698–707.

Stein, B. E., and Meredith, M. A. (1993). *The Merging of the Senses*. MIT Press, Cambridge, MA.

Truex, R. C., and Carpenter, B. A. (1964). *Strong and Elwyn's Human Neuroanatomy*, 5th ed. Williams & Wilkins, Baltimore.

Vanegas, H. (Ed.) (1984). *Comparative Neurology of the Optic Tectum*. Plenum, New York.

Wallace, M. T., and Stein, B. E. (1997). Development of multisensory neurons and multisensory integration in cat superior colliculus. *J. Neurosci.* **17,** 2429–2444.

Wallace, M. T., McHaffie, J. G., and Stein, B. E. (1997). Visual response properties and visuotopic representation in the newborn monkey superior colliculus. *J. Neurophysiol.* **78,** 2732–2741.

Wilson, R. A., and Keil, F. C. (Eds.) (1999). *The MIT Encyclopedia of Cognitive Science*. MIT Press, Cambridge, MA.

Synapses and Synaptic Transmission and Integration

MICHEL BAUDRY
University of Southern California

GLOSSARY

active zones Specialized membrane structures in presynaptic terminals in which vesicles are docked and clustered and neurotransmitter is released.

long-term potentiation/long-term depression Long-lasting increase/decrease in synaptic transmission as a result of specific patterns of electrical activity.

neurotransmitter A molecule synthesized and released by an action potential and activating a postsynaptic receptor, thus producing a change in membrane potential.

receptor A transmembrane protein that is activated by a neurotransmitter and produces a modification of some postsynaptic function.

signal transduction Biochemical cascade that translates receptor activation by a neurotransmitter into a cellular signal.

synaptic dynamics Time-dependent modification of synaptic transmission.

The term synapse was coined by Sherrington to designate the points of contact between the terminals of a neuron and another cell or neuron. Synaptic transmission represents the process by which communication between neurons occurs—that is, when an action potential is propagated along the axon of one neuron and produces a postsynaptic response in the target cell or neuron. This process is a specialized form of intercellular communication and exhibits many unique features. In particular, it depends on a relatively large number of chemical messengers, or neurotransmitters, and their specific receptors and associated signal transduction systems. Synaptic transmission spans several time domains ranging from a few milliseconds to hours or days. Synaptic integration refers to the dynamics of synaptic transmission because the efficacy of transmission is dependent on the frequency of the signals in the presynaptic elements as well as the past history of activity. These characteristics, frequency dependency and activity dependency, form the basis for the adaptive properties of synaptic transmission and are responsible for the critical role of synapses in information processing and storage.

I. INTRODUCTION

Although transmission between nerve cells occasionally takes place at electric synapses, the most common form of neuronal communication involves chemical synapses. Synapses are structures specialized for interneuronal communication that have to satisfy a number of difficult requirements: (i) They have to be optimally designed for fast as well as slow synaptic transmission; (ii) they have to be modifiable for both

short-term and long-term periods of time as a result of activity; and (iii) once modified, they have to remain stable for very long periods of time. Because it is widely believed that learning and memory involve some forms of synaptic modifications, the extremely long duration of memories implies that such modifications could last one's lifetime. Recent data as well as computer models have emphasized the importance of the morphological characteristics of synaptic contacts in determining their functional properties. Synapses contain complex biochemical machineries on both the presynaptic and postsynaptic sides that perform the synthesis and release of specialized molecules (neurotransmitters and neuromodulators) as well as generate complex responses to these molecules that span many time domains (milliseconds, seconds, minutes, and hours). Furthermore, synaptic transmission at a number of synapses exhibits many adaptive features that are generally considered to form the basis of information processing and storage processes. In this article, I briefly describe the morphological features of synapses. The characteristics of neurotransmitters and neuromodulators are then presented, and the mechanisms of transmitter release are described. This is followed by a presentation of the properties and functions of neurotransmitter receptors as well as those of active channels located at synapses. The article concludes by discussing the mechanisms involved in synaptic dynamics and synaptic plasticity and the functions of these processes in synaptic integration.

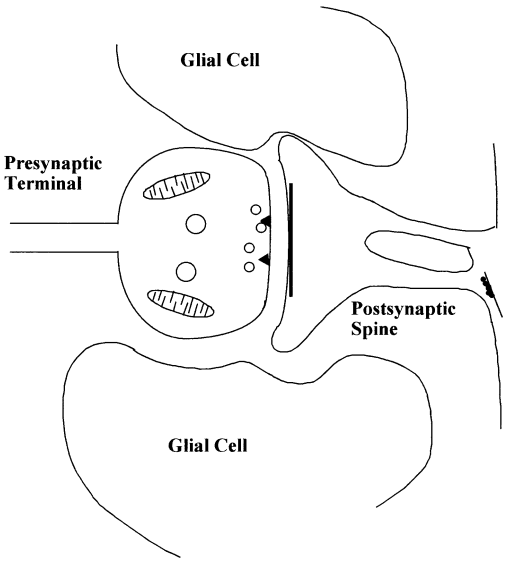


Figure 1 Morphological features of synapses. Schematic representation of an excitatory spine synapse. Note the presence of mitochondria in the presynaptic terminal, of small and large vesicles, and of active zones in the presynaptic membranes. A synaptic vesicle is depicted docked in an active zone. On the postsynaptic site, note the thickening of the postsynaptic membrane (corresponding to the postsynaptic density), the presence of a spine apparatus, generally thought of as a variant of the endoplasmic reticulum, and the presence of polyribosomes at the base of the spine. It is often assumed that signals originating from the postsynaptic density in response to neurotransmitter release might control the local translation of a small number of mRNAs present in dendrites. The whole synaptic complex is embedded in astrocytic processes.

II. MORPHOLOGICAL FEATURES

Synapses are highly specialized structures that are adapted to interneuronal communication. It was not until the advent of electron microscopy that the debate between Cajal and Golgi was definitely settled by showing the existence of numerous discontinuities between the axons and the dendrites of target neurons. Sherrington coined the term synapses to describe these points of contacts between one neuron and another cell or neuron. Synapses are characterized by a presynaptic element, generally loaded with small and large electron-dense vesicles (synaptic vesicles) that contain high concentrations of neurotransmitter and neuromodulator molecules, a postsynaptic element that exhibits a large variety of shapes, and a narrow synaptic cleft—a small 50-nm-wide space separating the pre- and postsynaptic elements. Synapses have been broadly categorized as asymmetric or excitatory and symmetric or inhibitory (Fig. 1). The presynaptic element often contains mitochondria, specialized membrane-associated structures believed to be the sites for neurotransmitter release (active zones), as well as various types of vesicles (small or large dense core vesicles). The postsynaptic element is characterized by a membrane thickening, the postsynaptic density, and occasionally by polyribosomes (especially at the basis of dendritic spines), which are believed to be a site for local protein synthesis. The postsynaptic density is assumed to contain numerous receptors for the presynaptically released neurotransmitters as well as associated transduction pathways and a whole set of anchoring proteins and the biochemical machinery required to regulate receptor properties.

Synapses exist in a variety of shapes and locations and include spine synapses, dendritic shaft synapses, cell body synapses, and axoaxonic synapses. Spine synapses are generally believed to be glutamatergic and excitatory (i.e., they use the excitatory amino acid

glutamate as a neurotransmitter), and recent results indicate that they also possess excitable channels, providing for a much more complex mode of functioning than previously thought. Axoaxonic synapses are less frequent and are presumed to provide for presynaptic regulation of synaptic transmission. A morphological feature that is becoming much more recognized is the close association of glial processes with synaptic contacts. Several functions have been attributed to such an association, including neurotransmitter metabolism, ion buffering, growth and trophic factor production, and morphological stabilization. Furthermore, it has been shown that glial cells exhibit several receptors for neurotransmitters, thereby providing for a much more active role for glial cells in the regulation of synaptic transmission. It is quite likely that the regulation of the shape of synaptic contacts as well as their stabilization involve complex interactions between cell surface adhesion molecules present in neuronal membranes and the extracellular matrix. Moreover, various types of enzymes, including proteases, lipases, and possibly protein kinases, have been shown to act extracellularly to regulate the interactions between neurons as well as between neurons and glial cells.

III. NEUROTRANSMITTERS AND NEUROMODULATORS

Synaptic transmission relies on a number of specialized molecules referred to as neurotransmitters and neuromodulators, which function in an anterograde direction—that is, from the presynaptic to the postsynaptic sites. In addition, neurotransmitters can also exert a presynaptic effect by stimulating receptors located in the presynaptic terminals (presynaptic receptors for the presynaptically released neurotransmitters are called autoreceptors). These molecules have to satisfy a number of criteria and have been divided into two major classes—classical neurotransmitters (mostly nonpeptidergic molecules) and peptide neurotransmitters. In addition, a small group of molecules have been identified that appear to function in a retrograde direction and influence the presynaptic site in response to postsynaptic activation. In all cases, neurotransmitters share a number of features with chemical messengers found in other forms of cellular communication. Not surprisingly, a number of neurotransmitters also participate in endocrine or paracrine communication.

A. Criteria

For a molecule to be considered a neurotransmitter, the following criteria should be satisfied: (i) The molecule should be present and/or synthesized in neurons that use it as a neurotransmitter; (ii) the molecule should be released as a result of an action potential (or as a result of depolarization), (iii) the molecule should produce appropriate pre- and postsynaptic responses, i.e., it should activate specific receptors, the activation of which by application of the exogenous molecules should mimic the effects of the endogenous neurotransmitter; and (iv) the molecule should be rapidly eliminated from the synaptic cleft. These criteria have been verified only rarely at synapses of the central nervous system (CNS), and a set of less stringent criteria has been developed to accommodate the technical difficulties of studying single synapses in the CNS.

B. Classical Neurotransmitters

The main feature of classical neurotransmitters is their local synthesis in the presynaptic terminals from precursor molecules provided by cellular metabolism or specialized transporters located in the plasma membrane. There are nine such molecules, amino acids or derivatives of amino acids, and acetylcholine, the neurotransmitter of the neuromuscular junctions in mammals. Local synthesis provides for a rapid and activity-dependent regulation of neurotransmitter availability in synaptic terminals. Specialized transport systems located in synaptic vesicle membranes are responsible for generating high neurotransmitter concentrations in synaptic vesicles. Once released in the synaptic cleft, neurotransmitters are rapidly inactivated by diffusion, reuptake in neurons or glial cells by potent transport systems, or by enzymatic inactivation. Several families of transporters for various neurotransmitters have been identified. Some are localized in neurons, whereas others exhibit mostly a glial localization. Transporters are generally responsible for the rapid elimination of neurotransmitters from the synaptic cleft and alterations in the function of transporters can result in dramatic modifications of synaptic transmission, as in the case of amyotrophic lateral sclerosis. In this disease, alteration in glutamate transport is responsible for the progressive degeneration of motoneurons.

C. Peptide Neurotransmitters

In contrast to classical neurotransmitters, peptide neurotransmitters are synthesized as large precursor proteins in the neuronal cell bodies; these precursors are transported down the axons by axonal transport and cleaved by appropriate peptidases in nerve terminals to generate the active neuropeptides. There are about 70 known neuroactive peptides and it is frequently the case that one or several neuropeptides are colocalized with a classical neurotransmitter. It has generally been proposed that neuropeptides are released under certain conditions along with the classical neurotransmitter and, as a result of activation of their postsynaptic receptors, modify the response to the classical neurotransmitter with which they are released. However, relatively little is known regarding the conditions under which neuropeptides are released and how their release affects synaptic transmission mediated by classical neurotransmitters. The peptide neurotransmitters belong to several families of peptides that sometimes share the same precursor protein. Several key differences exist between peptide neurotransmitters and classical neurotransmitters. First, as mentioned previously, peptide synthesis is nonlocal and the availability of neuropeptides has a much more limited local regulation than that of classical neurotransmitters. Second, neuropeptides are present in a much lower concentration than classical neurotransmitters, are released at a much lower concentration than classical neurotransmitters, and bind to postsynaptic receptors with high affinity compared to the low affinity of the classical neurotransmitters for their receptors. Finally, neuropeptides are stored and released from large, dense core vesicles as opposed to small secretory vesicles for classical neurotransmitters. They are degraded by a number of peptidases located extracellularly, and peptidase inhibitors have been found to alleviate a number of problems associated with disregulation of essential body functions. For all these reasons, neuropeptides are often considered to play a neuromodulatory role rather than a neurotransmitter function.

D. Retrograde Messengers

Recently, a small number of molecules have been identified and proposed to play the role of retrograde messenger at synapses; that is, these molecules are synthesized and released from the postsynaptic site, diffuse in the synaptic cleft, and act presynaptically to modify the properties of neurotransmitter release. In particular, arachidonic acid and certain derivatives of arachidonic acid could be produced as a result of postsynaptic phospholipase activation. Similarly, two diffusible gases nitric oxide (NO) and carbon monoxide, are produced postsynaptically and can affect neurotransmitter release. Whether these molecules qualify as neurotransmitters is a debated issue, and to many they belong to a broad category of chemical signaling molecules and have a limited role in synaptic transmission. In particular, these molecules could have a broader spatiotemporal domain of action and could influence many cellular types (e.g., glial cells and blood vessels).

IV. TRANSMITTER RELEASE

The mechanisms underlying neurotransmitter release represent a complex cascade of protein–protein interactions triggered by the influx of calcium resulting from the presynaptic activation of voltage-dependent calcium channels. A large number of proteins translate the influx of calcium into the fusion of synaptic vesicles to the so-called active zones of presynaptic membranes, which are presynaptic structures composed of voltage-dependent calcium channels and proteins required to dock synaptic vesicles against the presynaptic membranes. The majority of synaptic vesicle proteins and the proteins with which they are interact in the presynaptic terminal have been identified. One of the most remarkable features of transmitter release is its quantal nature, which originates from the storage of neurotransmitters in synaptic vesicles.

A. Quantal Release

Soon after electrophysiologists began investigating synaptic transmission, they realized that synaptic responses at individual synapses were not continuous but were distributed in a quantal fashion, representing multiple of a single size event, the quantal unit, which appears to be the amplitude of spontaneous potentials. In fact, when the distribution of evoked synaptic responses was analyzed statistically, the remarkable finding of Katz and colleagues was that this distribution followed a binomial law, with parameters n and p, where n is the number of quanta released and p the release probability. With the elucidation of the ultrastructure of presynaptic terminals, and the finding that synaptic vesicles were clustered near the active zones, it has been proposed that n could represent the number

of active zones. It is now clear that synapses in the CNS have different properties than those found at the neuromuscular junction. In particular, the majority of synapses in the CNS might have only one active zone, and the amplitude of postsynaptic potentials might reflect the number of postsynaptic receptors rather than the amount of transmitter released.

B. Release Machinery

A large number of vesicle-associated proteins have been identified and shown to participate in synaptic vesicle movement in presynaptic terminals, docking on the active zones, and fusing with presynaptic membranes. Several models of neurotransmitter release have been proposed to incorporate all these proteins and to account for all the salient features of transmitter release. One of the most recent models suggests that synaptobrevin and synaptophysin are bound to each other before synaptic vesicles fuse with the plasma membrane. Following an influx of calcium, synaptotagmin-1 (the calcium sensor of synaptic vesicles) interacts with two proteins from the plasma membrane, syntaxin and SNAP-25. Synaptobrevin dissociates from synaptophysin and interacts with the synaptotagmin-1–syntaxin–SNAP-25 complex. The interaction with SNAP dislodges synaptotagmin, and a new intermediate complex is formed. Subsequent steps leading to the fusion of the synaptic vesicle with the plasma membrane have not been completely identified. In addition, a family of vesicle-associated proteins, the synapsins, have long been identified and been shown to play a critical role in regulating the distribution of synaptic vesicles between two pools, the readily releasable vesicles and the reserve pool. Greengard and collaborators have shown that phosphorylation of synapsins is responsible for regulating the interactions between vesicles and actin filaments and thus for regulating the movement and the localization of synaptic vesicles in presynaptic terminals.

V. RECEPTORS AND SECOND MESSENGERS

All neurotransmitters exert their effects by activating membrane-associated proteins called receptors. When receptors are ligand-gated ionic channels, they are ionotropic receptors, whereas metabotropic receptors generally refer to receptors that are coupled to G proteins and produce a modification of the concentration of a second, intracellular messenger or the modulation of an ionic channel. Each neurotransmitter stimulates an array of receptors, thereby eliciting a number of postsynaptic responses (Fig. 2). Moreover, responses to a given neurotransmitter can be modified as a result of the simultaneous activation of receptors for another neurotransmitter on the same target cell or of the past activity of the target cell. Finally, transmitter receptors are not exclusively located on neuronal cells; they are also present on glial cells, thus establishing a direct communication between neurons and glia.

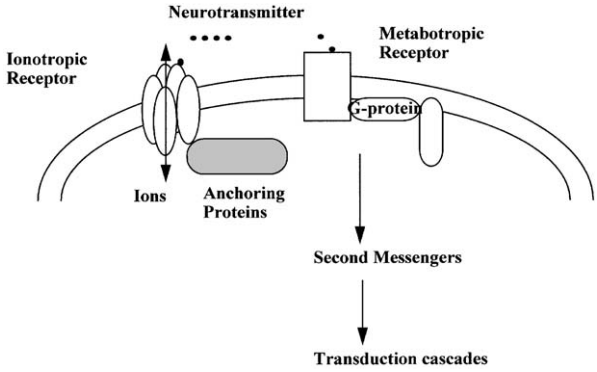


Figure 2 Schematic representation of neurotransmitter receptors. Both ionotropic and metabotropic receptors are schematically represented. Upon binding of the neurotransmitter, ionotropic receptors are activated and produce an increased permeability to specific ions, thereby resulting in a brief modification of the membrane potential. Activation of metabotropic receptors results in the activation of G proteins and their associated target ion channels or enzymes and the modification of the concentration of a variety of second messengers. This leads to the activation of transduction cascades that can modify the state of the neurons for minutes, hours, or weeks.

A. Ionotropic Receptors

The best known ionotropic receptor is the nicotinic receptor, which has served as the prototypical model for ionotropic receptors (this has occasionally led to some significant setbacks because some receptors were found to have properties distinct from those of the nicotinic receptors). It is a heteromeric pentameric protein, and each subunit is a transmembrane protein with four transmembrane domains. There are two binding sites for acetylcholine and binding of acetylcholine to the extracellular domain of the receptor results in a change of configuration of the channel domain and in a brief opening of a transmembrane ionic pore. This produces a brief modification of the membrane potential, the so-called postsynaptic potential. The same scenario seems to generalize to all

neurotransmitters, with hyperpolarization being produced by inhibitory neurotransmitters (generally due to the opening of chloride channels) or depolarization by excitatory neurotransmitters (generally due to the opening of sodium channels). This is generally followed by a desensitization phase during which the channel remains closed despite the binding of the ligand to its recognition domain. Numbers and kinetics for opening and closing of the receptor channels are critical parameters to determine the main features of synaptic transmission. A number of issues remain highly debated, particulary the saturability of the postsynaptic receptors by synaptically released neurotransmitter. In other words, are all the available receptors occupied following a release event? A number of laboratories have attempted to estimate the number of postsynaptic receptors for various neurotransmitters at different synapses. At glutamatergic synapses, recent studies estimate this number to be approximately 100 receptors. This obviously contrasts with the tens of thousands of nicotinic acetylcholine receptors found at the mammalian neuromuscular junction and emphasizes the point that synapses can have radically different properties in different structures. The answer to the question of postsynaptic saturability has obvious implications for mechanisms of synaptic plasticity. Furthermore, the mechanisms involved in the postsynaptic localization, anchoring, internalization, and turnover of the receptors remain the object of intense investigation. Recent studies have identified families of proteins that interact with the intracellular C-terminal domains of various receptors and channels and appear to anchor the receptors or channels in synaptic membranes as well as to link the receptors to various intracellular signaling processes. These proteins belong to a family of PSD-95 proteins, which are multifunctional proteins because they exhibit a guanylate kinase activity. In addition, members of the typical exocytosis machinery such as *N*-ethylmaleimide-sensitive factor and associated proteins have also been found to participate in receptor targeting and insertion mechanism. As discussed later, it is widely assumed that rapid regulation of receptor properties and numbers is an essential feature of synaptic modifications as a result of various patterns of activity.

B. Metabotropic Receptors

In addition to ionotropic receptors, neurotransmitters activate transmembrane receptors coupled with G proteins that generate a variety of cellular responses. These metabotropic receptors belong to a superfamily of proteins with seven transmembrane domains, with an extracellular ligand binding domain and an intracellular C-terminal domain that interacts with G proteins. Depending on the type of G proteins activated by these receptors, metabotropic receptor activation results in a variety of cellular responses, particularly in the modification of the concentration of several second messengers (e.g., calcium, cyclic AMP, cyclic GMP, inositol-triphosphate, and other phospholipid derivatives). Like ionotropic receptors, metabotropic receptors exhibit desensitization, generally mediated by phosphorylation reactions. Whereas ionotropic receptor-mediated responses generally last less than 1 sec, metabotropic receptor-mediated responses last for considerably longer periods of time, from seconds to minutes. Furthermore, some of these responses involve activation of transcriptional processes, resulting in the modification of specific proteins, and therefore produce functional modifications of neurons that last for hours to days (depending on the half-life of the newly translated proteins). As in the case of ionotropic receptors, several proteins involved in the targeting and anchoring of metabotropic receptors have been identified.

C. Second Messengers and Their Functions

Most second messengers activate enzymatic processes and are thus part of biochemical cascades that serve to amplify the responses to the first messenger. A number of protein kinases are the targets of second messengers, although protein phosphatases can also be activated by second messengers. As a result, a large variety of proteins can be functionally modified following the activation of metabotropic receptors. These include cytoskeletal proteins, neurotransmitter receptors, and ionic channels as well as transcription factors that regulate the expression of specific genes. Thus, the properties of ionotropic receptors can be regulated by phosphorylation reactions triggered by activation of a number of metabotropic receptors. In particular, channel kinetics, desensitization parameters, and even localization of a particular receptor can be modified as a result of the previous activation of metabotropic receptors for the same or a different neurotransmitter in its vicinity. As discussed later, some of these biochemical cascades have properties of biochemical switches and have been postulated to play important roles in synaptic plasticity processes. In any event,

second messenger-mediated cascades are highly regulated and provide for rich interactions between receptors for several neurotransmitters.

VI. SYNAPTIC INTEGRATION

It is well established that synaptic transmission at a given synapse is highly variable because it depends on a variety of parameters, including the state of the postsynaptic neuron (depolarized, hyperpolarized, firing, etc.) and the history of the synapse that influences both pre- and postsynaptic properties. These two properties, time- and state-dependent fluctuations and activity-dependent modifications of synaptic transmission, are considered the basis for most of the features of information processing and storage. It is generally agreed that synaptic dynamics regulate the efficacy of synaptic transmission on a short-term timescale, whereas synaptic plasticity operates on a longer timescale to modify synaptic strength.

A. Synaptic Dynamics

Synaptic transmission at CNS synapses is highly variable. This variability is the result of multiple processes that can alter any step involved in this highly complex and regulated mechanism (Fig. 3). The effectiveness of synaptic transmission depends on the frequency of stimulation and on the past activity of both pre- and postsynaptic elements. In general, repetitive high-frequency stimulation results in a rapid growth of postsynaptic potentials, a process called synaptic facilitation. A slower phase of increased transmission with a time course of several seconds follows facilitation and is called augmentation. A phase of decreasing transmission can also be superimposed to these processes and is referred to as synaptic depression. Other processes resulting from presynaptic feedback inhibition contribute to further regulate synaptic strength during a train of presynaptic action potentials. Several of these processes are due to the dynamics of $[Ca^{2+}]$ in the presynaptic elements and to the links between $[Ca^{2+}]$ and neurotransmitter release. Thus, synaptic strength is constantly modulated by the temporal pattern of presynaptic activity and the past history of the network in which the synapse is embedded. It is widely believed that this property of synaptic transmission results in the extraction of significant signals from noisy and variable temporal patterns. It has recently been recognized that dendrites and dendritic spines also

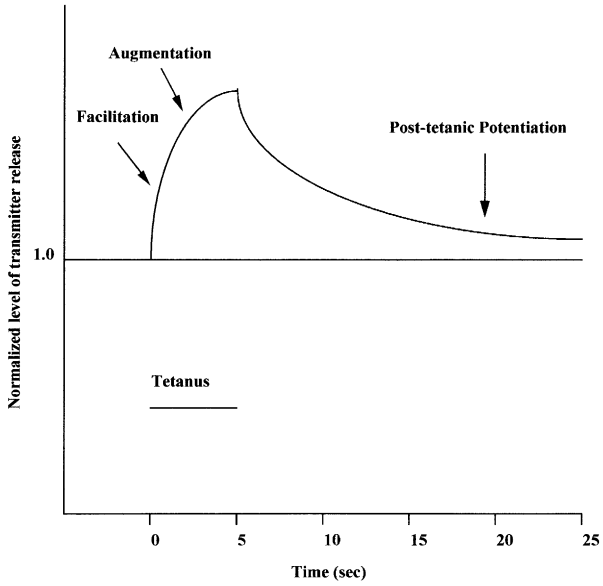


Figure 3 Multiple forms of synaptic modifications. Schematic representation of various forms of short-term synaptic plasticity due to changes in probability of release in response to a train of action potentials. Facilitation is a rapid increase in transmitter release occurring after a few pulses. Augmentation follows facilitation and has a longer decay period than facilitation. Finally, posttetanic potentiation is generally characterized by an increase in transmitter release lasting 20–40 sec.

contain active channels (in addition to neurotransmitter receptors), such as voltage-dependent sodium, calcium, and potassium channels. This means that firing of an action potential by the neuron can produce a dendritic action potential that is retrogradely propagated and thereby modifies the membrane potential of dendrites and their future responses to incoming signals. This property introduces additional complexities to synaptic transmission and integration. Numerous computer simulations of models of synapses incorporating rules of synaptic modifications based on experimental data have been developed and used to explore the computational properties of neuronal networks.

B. Synaptic Plasticity

In addition to the short-term variability of synaptic strength just described, several forms of long-term modifications of synaptic transmission, referred to as synaptic plasticity, have been described for a variety of synapses. Long-term potentiation (LTP) and

long-term depression (LTD) have received the most attention because they represent potential cellular models of information storage. LTP was first described in 1973 by Bliss and Lomo in rabbit hippocampus and has since been reported to occur in a wide range of structures and species. Its most common form is present at glutamatergic synapses, where stimulation of the NMDA subtypes of glutamate receptors triggers an influx of calcium and the activation of a complex intracellular cascade leading to increased responsiveness of the AMPA subtypes of glutamate receptors. Although much has been learned about the mechanisms involved in LTP during the past 10–15 years, several issues remain unanswered, particularly regarding the roles of gene expression and protein synthesis in LTP maintenance as well as the nature of the process involved in LTP stabilization. Two types of mechanisms by which synaptic activity could produce localized modifications of synaptic properties restricted to activated synapses have been proposed. In the first one, there is a small population of mRNAs constitutively present in dendrites; in response to an appropriate signal, these mRNAs are locally translated, and the newly synthesized proteins are incorporated into the activated synapses. Note that this process is rapid and does not require gene expression because it depends only on constitutively present mRNAs. In the second one, some locally generated signals are transmitted to the cell nucleus, triggering gene transcription and the synthesis of mRNA. This mRNA is targeted to the dendrites and is locally translated at the appropriate site. Note that this requires the existence of a signal or signals that remain present at activated synapses for a long period of time because the transcription of genes and their targeting to the dendrites necessitate a significant period of time. This mechanism is related to the synaptic tagging described by Frey and Morris. In their model, these authors propose that LTP induction is accompanied by the activation of signals they call "synaptic tags," which interact with plasticity-related newly synthesized proteins and stabilize the formation of LTP.

Many models of LTP propose that changes in synaptic efficacy are best accounted for by changes in synaptic structure. Although the issue is still highly debated, a large literature supports the notion that LTP is associated with modifications of synaptic structure and especially that an increased number of perforated synapses might represent an intermediate stage in the establishment of long-lasting potentiation. Again, various models have been proposed to incorporate changes in synaptic structure with changes in synaptic function, ranging from an increase in the number of functional synaptic contacts to an increase in axodendritic synapses. A popular version of these models is the notion of silent synapses, i.e., synapses with functional NMDA receptors but lacking functional AMPA receptors. Such synapses could become functional as a result of the coincidence of presynaptic activity with postsynaptic depolarization (conditions necessary and sufficient to activate the NMDA receptors) and the appearance of functional AMPA receptors due to their insertion in postsynaptic membranes from an intracellular pool or their redistribution from extracellular locations. Studies of simulation of transmitter release and receptor activation based on a synapse model that incorporates the geometry of synaptic contacts as well as the distribution of receptor proteins in postsynaptic densities have been performed to examine the role of various parameters in LTP. The results were quite astonishing because they indicated that the geometry of synaptic contacts has a much greater impact on synaptic function than previously thought. This is true not only on the presynaptic terminals, where geometry determines the probability of transmitter release, but also on the postsynaptic sites, where receptor location in concert with the location of neurotransmitter release determine the probability that a transmitter molecule activates a postsynaptic receptor. Thus, a silent synapse could simply represent a synapse for which the receptors are located too far from the site of transmitter release to be activated. Relocation of transmitter receptors closer to the release site could transform a silent synapse into an active synapse.

In contrast to LTP, LTD has a much shorter history and has only recently been studied extensively and incorporated into models of learning and memory. LTD at glutamatergic synapses between the parallel fibers and the Purkinje neurons in the cerebellum was proposed to account for the adaptation of the vestibuloocular reflex by Ito and colleagues. It was later proposed to account for learning of classically conditioned motor responses. Cerebellar LTD occurs when parallel fiber activation is closely associated with climbing fiber stimulation, and it appears to be due to a long-lasting decrease in AMPA receptor responsiveness at the parallel fiber to Purkinje cell synapse. Several second messengers have been proposed to participate in LTD induction, including cGMP and NO. LTD has also been found at hippocampal glutamatergic synapses and could represent the reversal of LTP and, in particular, the transformation of active synapses into silent synapses. Whether

cerebellar LTD and hippocampal/cortical LTD represent the same or different processes remains to be determined.

VII. CONCLUSIONS

Synaptic transmission is a highly complex and regulated process. Not surprisingly, numerous neurological diseases are thought to result from alterations in one or several steps involved in synaptic transmission. In particular, alterations in dopaminergic transmission are widely believed to contribute to schizophrenia, whereas abnormalities in biogenic amine transmission might contribute to affective disorders. Furthermore, learning and memory are generally considered to be due to activity-dependent modifications of synaptic efficacy along the networks involved in the processing of information. Although the roles of LTP/LTD in learning and memory remain highly debated, it is likely that multiple forms of synaptic plasticity are responsible for the rich variety of memory types exhibited by mammalian brains. Neuronal networks incorporating synaptic dynamics and synaptic plasticity rules derived from experimental neurobiology have been shown to exhibit rich computational capabilities, and silicon instantiation of such networks is being developed for performing human-like computational operations.

See Also the Following Articles

AXON • CHEMICAL NEUROANATOMY • NERVOUS SYSTEM, ORGANIZATION OF • NEURON • NEUROTRANSMITTERS • PERIPHERAL NERVOUS SYSTEM • VENTRICULAR SYSTEM

Suggested Reading

Baudry, M., and Davis, J. L. (Eds.) (1991, 1994, 1999). *Long-Term Potentiation*, Vols. 1–3, MIT Press, Cambridge, MA.

Baudry, M., Thompson, R. F., and Davis, J. L. (Eds.) (1993). *Synaptic Plasticity*. MIT Press, Cambridge, MA.

Kandel, E. R., Schwartz, J. H., and Jessell, T. M. (Eds.) (1991). *Principles of Neural Science*, 3rd ed. Appleton & Lange, East Norwalk, CT.

Siegel, G. J., Agranoff, B. W., Albers, R. W., Fisher, S. K., and Uhler, M. D. (Eds.) (1998). *Basic Neurochemistry*, 6th ed. Lippincott–Raven Press, New York.

Zigmond, M. J., Bloom, F. E., Landis, S. C., Roberts, J. L., and Squire, L. R. (Eds.) (1999). *Fundamental Neuroscience*. Academic Press, San Diego.

Synaptogenesis

FREDERIQUE VAROQUEAUX and NILS BROSE
Max Planck Institute for Experimental Medicine, Göttingen, Germany

GLOSSARY

active zone Presynaptic membrane specialization that the transmitter release process is restricted to; contains a scaffold of proteins involved in synaptic vesicle exocytosis.

cell adhesion Binding between cells, mostly mediated by cell surface molecules that interact homotypically or heterotypically with other cell surface proteins or with components of the extracellular matrix.

postsynaptic density Electron-dense postsynaptic membrane specialization containing a scaffold of receptors, ion channels, and signal transduction proteins.

synapse Contact between the sending nerve cell and target cell, typically formed by a presynaptic axon terminal that is specialized for transmitter secretion and a postsynaptic cell body or dendrite that is specialized for signal reception and transduction.

Synaptogenesis, the formation of functional chemical synapses between neurons, represents the climax of brain development. It results in the establishment of an ordered neuronal network that is later fine-tuned or remodeled to allow dynamic and flexible information processing in the brain. This article describes the characteristics of synaptogenetic processes with an emphasis on their possible molecular basis.

I. INTRODUCTION

Information processing in the brain is based on a complex and specific network of synaptic connections that in humans consists of approximately 10^{11} neurons connected by 10^{15} synapses. Synapses are asymmetric, strongly adhering, morphologically and functionally specialized contact zones where transfer of information takes place between a sending nerve cell and a receiving neuron, muscle fiber, or other target. The presynaptic axon terminal is specialized for the complex membrane trafficking mechanisms that underlie regulated secretion of transmitter. Many of the proteins it contains are highly specific for presynaptic structures. The axon terminal faces a postsynaptic structure that is uniquely specialized for signal reception and transduction. The postsynapse is typically built as a scaffold of proteins involved in clustering and anchoring neurotransmitter receptors, ion channels, and intracellular signal transduction molecules, many of which are specifically targeted to this structure and not present in other parts of the cell. The extraordinary specificity of connections in the adult brain is achieved by an elaborate developmental program that is regulated by a combination of genetic and epigenetic processes. Embryonic stem cells generate a population of neural precursor cells that proliferate and differentiate into glial cells and immature neurons. Immature neurons migrate to their final destination, where they form processes, axons, and dendrites, through which they are able to make synaptic contacts with target cells. Neurons can grow very long and ramified processes allowing them to connect with target structures over considerable distances. Once growing axons have reached their target areas, the process of brain development culminates in synaptogenesis, that is, in the formation and maturation of specific functional synapses between defined sets of neurons, resulting in an amazingly complex neuronal network. Even after

the network is formed, it retains its ability to undergo dynamic changes. Existing synaptic connections can be strengthened, weakened, or eliminated, and new synapses can be generated as a consequence of endogenous or environmental stimuli and experience, allowing the brain to adjust and optimize its responses and adapt to novel tasks throughout life. Similar to the situation in the brain, synapses in the periphery (e.g., at the neuromuscular junction) are generated with extreme accuracy, mature after formation, and can be modified or eliminated in response to various stimuli.

Centrally and peripherally, synaptogenesis and activity-dependent modulation of synaptic strength or connectivity are the final steps in the formation of a functional neuronal network. On an operational level, two temporally distinct synaptogenetic processes can be distinguished. First, a nascent axonal growth cone specifically recognizes its appropriate target neuron and make an initial contact. Second, pre- and postsynaptic proteins have to be recruited to the initial contact site in order to assemble a functional synapse. Accumulating evidence from convergent biochemical, morphological, electrophysiological, and genetic studies in model organisms (*Drosophila*, *Caenorhabditis elegans*, and mouse) indicates that both processes are mediated by a complex cascade of protein–protein interactions, involving cell adhesion molecules, intracellular adaptor proteins, and components of the secretory machinery or signal transduction cascades. Although the basic principles of synaptogenesis are similar for most synapses at the phenomenological level, ultrastructure and protein composition of synaptic contacts as well as molecular mechanisms of synaptogenesis differ strikingly between different types of synapses and seem to be only partially conserved during evolution. This suggests that multiple and, to a certain degree, redundant molecular mechanisms can contribute to the process of synapse formation.

This article discusses synapse structure and its synaptogenetic requirements, basic characteristics of synaptogenetic processes, and molecular mechanisms of synaptogenesis with respect to the four most prominent types of synapses: the excitatory glutamatergic synapse (type I), the inhibitory glycinergic synapse (type II), the inhibitory GABAergic synapse (type II), and the cholinergic neuromuscular junction.

II. STRUCTURE OF THE SYNAPSE

Neurotransmission at central nervous system and neuromuscular synapses occurs by the same basic mechanisms. In response to an action potential, the highly specialized presynaptic terminal releases neurotransmitter into the synaptic cleft by exocytotic fusion of synaptic vesicles. Transmitter then diffuses to the postsynapse, where an elaborate signal transduction apparatus mediates reception of the neurotransmitter and propagation of the signal. Synapses are asymmetric cell–cell contacts that consist of three core compartments: the presynapse, the synaptic cleft, and the postsynapse (Fig. 1).

The presynaptic compartment is morphologically very similar in all types of synapses (Fig. 1). Ultrastructurally, it is characterized by an accumulation of transmitter containing small (50 nm), clear synaptic vesicles. These are clustered near a thickened stretch of presynaptic plasma membrane, the active zone, to which synaptic exocytosis is restricted. Additional material protrudes from the active zone into the presynaptic lumen and partially fills the space between the vesicle cluster, suggesting that a dense network of scaffold molecules, together with components of the cytoskeleton and membrane skeleton, controls and optimizes vesicle trafficking at the active zone. In addition to small, clear vesicles, most presynapses contain large, dense-core vesicles (100 nm) as well as clathrin-coated vesicles, the latter reflecting the permanent membrane recycling activity in the presynaptic terminal. The high energy requirement of presynapses is reflected by the presence of numerous mitochondria.

In contrast to the similarities at the presynaptic level, different types of synapses differ dramatically in the structure and composition of the synaptic cleft and postsynaptic compartments. At central nervous system synapses (Figs. 1A and 1B), the synaptic cleft is formed by a regular parallel apposition of pre- and postsynaptic membranes. It is characterized by a widening of the intercellular space (20 and 30 nm in symmetric and asymmetric synapses, respectively, vs 15 nm outside of synaptic areas) that contains dense material but is devoid of a basal lamina and its molecular constituents. At the neuromuscular junction, a structured basal lamina ensheaths each muscle fiber and extends through the synaptic cleft and into the junctional folds, separating pre- and postsynaptic membranes (Fig. 1C). The synaptic part of the basal lamina has a specific molecular composition. At the postsynaptic level, structures are also diverse. Excitatory/glutamatergic type I synapses, the most abundant in the brain, are mostly located on dendritic spines. These represent a separate cellular compartment with autonomous biochemical activity that contains endoplasmic reticulum, polysomes, and mitochondria.

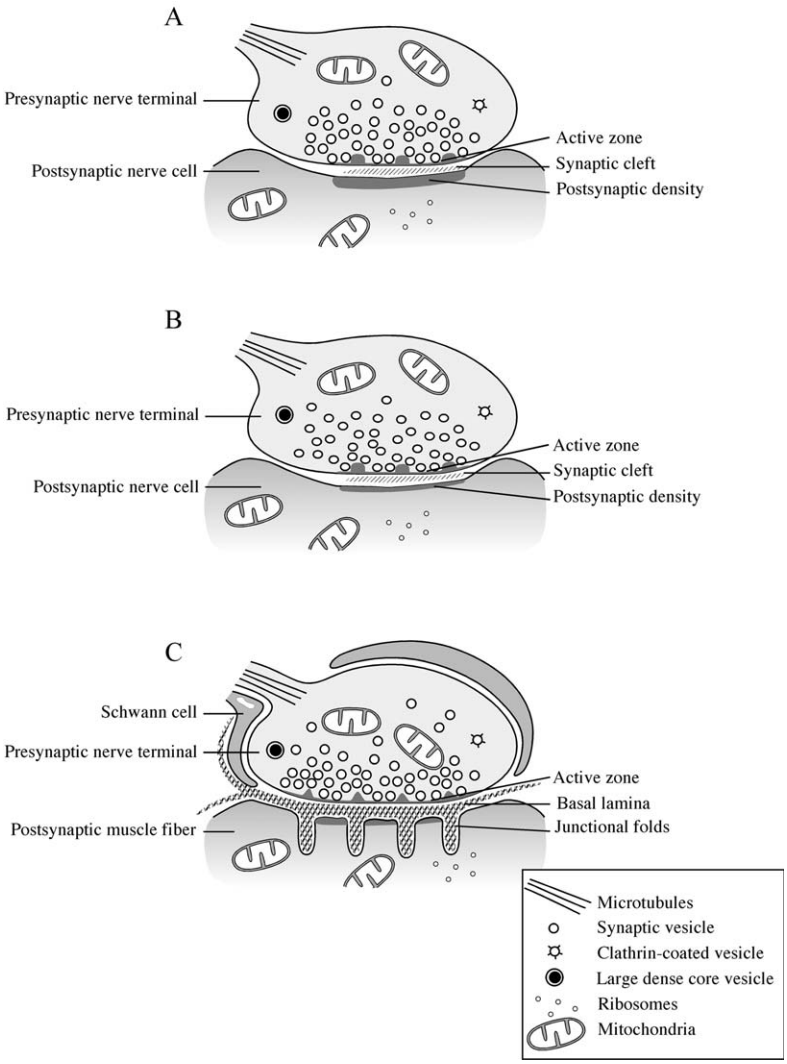


Figure 1 Morphology of chemical synapses. (A) Central asymmetric synapse (or type I); (B) central symmetric synapse (or type II); (C) neuromuscular synapse. Synapses in the central nervous system are established between two neuronal elements (A and B) and at the neuromuscular junction between a motor neuron and a muscle fiber (C). Asymmetric (or type I) synapses (A) use glutamate as major neurotransmitter. Symmetric (or type II) synapses (B) utilize GABA or glycine, whereas acetylcholine is the main chemical messenger delivered at the neuromuscular junction (C). All three types of synapses have developed a highly specialized and polarized presynaptic apparatus, where neurotransmitter-containing synaptic vesicles dock at the active zone of the presynapse and release their transmitter content into the synaptic cleft in response to an arriving action potential. At the neuromuscular junction only, this cleft is occupied by a dense organized network of material, the basal lamina. Facing the presynaptic release zone, the postsynaptic membrane is locally thickened and built differentially in the three synapse types. At central synapses, it is parallel to the release site and characterized by an intracellular accumulation of dense material, the so-called the postsynaptic density, characterized by an electron-dense appearance in electron microscopy. The postsynaptic density is prominent in asymmetric synapses (A) but barely detectable in symmetric synapses (B). At the neuromuscular junction, the postsynaptic membrane exhibits deep folds and intracellular accumulation of dense material in its uppermost area (C). Free or endoplasmic reticulum-linked ribosomes are observed in the proximity of the postsynapse. Under aldehyde fixation conditions with high osmolarity buffers, synaptic vesicles appear more ovoid or almost flattened in symmetric synapses, whereas they remain round shaped at asymmetric sites.

Type I postsynapses (Fig. 1A) are characterized by a prominent disc-shaped thickening underneath the membrane, the postsynaptic density, and by dense structures emanating from both sides of the junction, the puncta adhaerentia. Inhibitory/GABAergic or glycinergic synapses are mostly formed between axons and cell somata. They have a much less prominent postsynaptic thickening (Fig. 1B) and have therefore

been referred to as symmetric synapses (type II) in contrast to the prominent asymmetric appearance of the synaptic membranes in type I excitatory synapses. At the neuromuscular synapse (Fig. 1C), the postsynaptic membrane forms deep infoldings opposite to the active zone and exhibits thickenings at its uppermost portion.

The ultrastructural diversity of synapses reflects their fundamentally different molecular compositions, which will be discussed later. Their complex structure is based on an elaborate arrangement of a large number of proteins. During synaptogenesis, these have to be assembled in a coordinated and efficient manner—a task that is achieved by an interplay of cell adhesion molecules, scaffold proteins, and functionally relevant pre- and postsynaptic proteins.

III. RECOGNITION AND CONTACT FORMATION

A. The Problem of Specificity and Adhesion

Any given brain region contains multiple classes of neurons that can be distinguished on the basis of their localization, morphology, molecular composition, functional characteristics, and connectivity. In the case of the cerebral cortex, neuronal diversity may amount to approximately 500 different types of neurons. To produce specific connectivity despite this diversity, a cascade of pathfinding and recognition processes comes into play that allows a given neuron to find its partners.

Long before a nascent axonal growth cone establishes contact with its partner cell and synaptogenesis can commence, several classes of hierarchically organized cell surface and secreted proteins mediate and modulate axon extension and guidance, ushering the axon to its appropriate target area. In some cases, guidepost cells serve as temporary scaffold and transient synaptic targets in these processes. Such pathfinding mechanisms significantly reduce the specificity problem of synaptogenesis by confining possible partners of an ingrowing process to a defined compartment of the brain. However, the establishment of specific synaptic connections remains a daunting task because several levels of specificity have to be met by newly formed synapses. First, mechanisms of synaptogenesis have to mediate specificity at the cellular level, allowing an axon to distinguish between different cell types or even between individual cells of the same type. Second, mechanisms of synaptogenesis have to be specific at the subcellular level, identifying axon, dendrite, and soma of a target cell as functionally distinct input areas. Finally, mechanisms of synaptogenesis have to be specific for the different types of transmitter that will be used at the newly forming synapse.

Similar to axon pathfinding, synaptic recognition and synaptogenesis processes are thought to be mediated by cell adhesion molecules. In fact, different types of cell adhesion proteins must be present to mediate different levels of synaptic specificity, determining target region and target cell identity, subsections of the neuronal membrane surface, pre- and postsynaptic compartments, transmitter specificity of a synapse, or synapse strength, size, and stability. In addition to synapse specificity, cell adhesion molecules also fulfill the more mundane task of synaptic adhesion, keeping pre- and postsynaptic compartments in close apposition and register.

In contrast to synaptogenesis in the central nervous system, the specificity requirements during the formation of a neuromuscular junction are less complex. Once guided to the appropriate target area, the ingrowing axon of a motor neuron contacts its target structure near its site of entry into the muscle. This initial unspecialized contact is only capable of rudimentary neurotransmission, reflecting the absence of pre- and postsynaptic specializations, and matures into a functional synapse as the presynaptic compartment assembles and triggers postsynaptic clustering of acetylcholine receptors by secreting a soluble clustering factor. The fact that neuromuscular junctions contain a basal lamina while synapses in the central nervous system do not suggests that membrane–matrix interactions may predominate adhesion mechanisms at the neuromuscular junction, whereas typical cell–cell adhesion molecules may be more crucial in the brain. Such differences in adhesive mechanisms also imply that signals involved in target recognition and synapse maintenance may be different at neuromuscular junctions compared to synapses in the central nervous system.

B. Cell Adhesion Molecules in Synaptic Contact Formation and Specificity

Cell adhesion molecules form a diverse group of cell surface proteins that are specialized in mediating adhesion at the cell–cell and cell–substrate interface and play important roles in a variety of physiological processes, ranging from organogenesis to adhesion in cells of the immune system. In addition, many cell

adhesion molecules also serve signal transducing purposes. Although several hundred putative cell adhesion proteins have been described, only very few may be present in synapses. In fact, only five cell adhesion complexes have been found in electron microscopic studies to be unequivocally synaptic: cadherin/catenin, protocadherins/cadherin-litre neuronal receptors/(CNRs), neurexin/neuroligin, neural cell adhesion molecule (NCAM), and integrins (Fig. 2).

1. Cadherins

Cadherins were originally discovered as vertebrate cell surface molecules that mediate Ca^{2+}-dependent cell adhesion. They are expressed by almost all cell types in which, they are enriched in cell–cell junctions of different types (e.g., zonula adherens and desmosomes). Cadherins are essential for cell adhesion during development as well as in mature tissues. In addition to "classic" cadherins, the cadherin protein family contains desmosomal cadherins, protocadherins, and cadherin-related molecules. Classic cadherins (>15 isoforms) are composed of a large N-terminal extracellular domain containing five repeated subdomains (EC1–EC5), a single transmembrane domain, and a highly conserved C-terminal cytoplasmic region (Fig. 2). With few exceptions, cadherins interact homotypically, preferentially binding to proteins of the same subtype. This homotypic interaction and its specificity are mediated by multiple portions of the N-terminal domain. Intracellularly, classic cadherins interact with catenins, a group of proteins that attach the cytoplasmic domain of cadherins to the actin cytoskeleton and function as intracellular signaling molecules.

In the central nervous system, cadherins are involved in axon outgrowth, axon fasciculation, target recognition, and maintenance of neuronal connections in the adult brain. The differential region-specific and cell-type-specific expression patterns of certain cadherins was interpreted as a possible molecular basis for the specificity of synaptic connectivity in the brain, where the binding specificity of cadherins expressed on invading growth cones could determine synaptic partner choice. However, the number of classic cadherins expressed in the brain may not be high enough to account for the observed specificity in the synaptic network. Nevertheless, a synaptic role of cadherins is evident from the observation that neural (N) and epithelial (E)-cadherins are found on both sides of many synapses in the brain. Furthermore, their

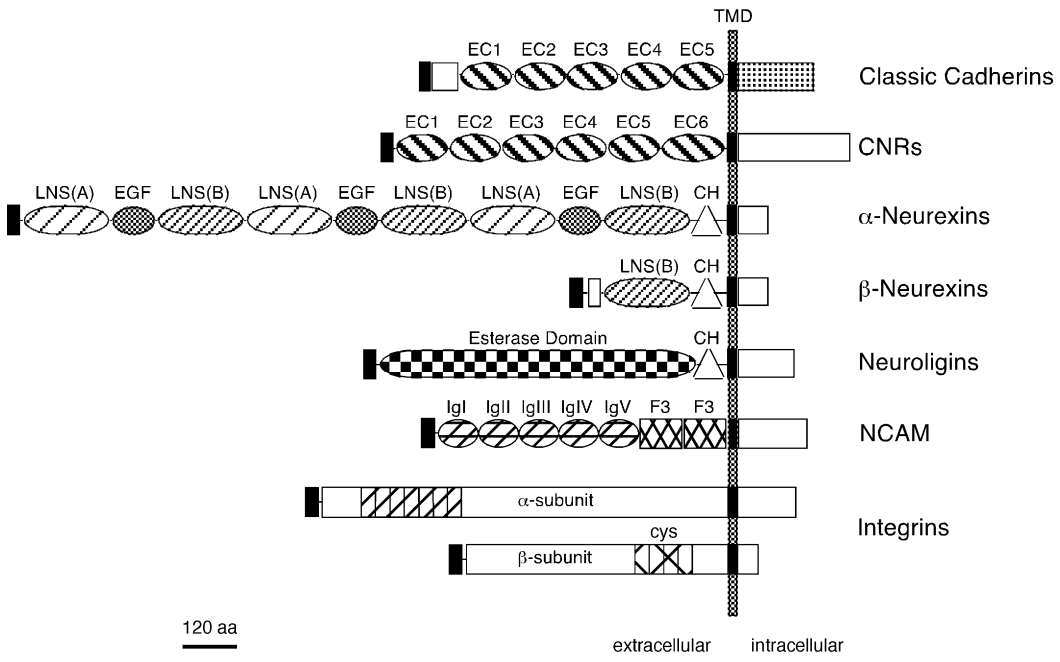


Figure 2 Domain structures of synaptic cell adhesion molecules. CH, O-linked carbohydrate attachment sequence; cys, cystein-rich repeats; EC, cadherin-like domain; EGF, EGF-like domain; F3, fibronectin type III domain; IgI–IgV, immunoglobulin domains I–V; LNS, laminin A/neurexin/sex hormone binding globulin repeats; TMD, transmembrane domain; *black* boxes, signal peptide sequences; *left*, N termini.

intracellular adaptor proteins, αN- and β-catenins, are present in asymmetric (type I) and symmetric (type II) synapses. In type I synapses catenin-containing regions are localized in a parasynaptic manner, flanking presynaptic release sites and postsynaptic densities, whereas in type II synapses catenin localization overlaps with release zones. The differential association of N- and E-cadherins with distinct synapses suggests that the homotypic binding preferences of different cadherin isoforms contributes to the specificity of synapse formation.

In addition to a role in synaptogenesis, functional cadherin-mediated synaptic cell adhesion appears to influence mature synapses as well in that it may be involved in the activity-dependent strengthening of synapses during memory formation by enhancing synaptic adhesion and efficacy.

2. Cadherin-like Neuronal Receptors and Protocadherins

The discovery of a novel family of cadherin-related proteins, the cadherin-like neuronal receptors (CNRs), has provided additional evidence for a synaptic function of the cadherin family of proteins. CNRs constitute a family of more than 50 different gene products. They are composed of a large N-terminal extracellular region containing six EC domains (EC1–EC6) with high homology to cadherins, a single transmembrane domain, and an intracellular C terminus that is highly conserved within the CNR family but has no homology with known cadherins and therefore does not bind catenins (Fig. 2). Some CNRs were shown to be present at the synapse, forming transsynaptic connections.

CNRs belong to the family of protocadherins whose common features are six or seven extracellular cadherin ectodomains and a divergent cytoplasmic region. Like CNRs, the initially identified protocadherins are highly expressed in brain, where they are localized mainly at cell–cell contact sites. One isoform, OL-protocadherin, was shown to be enriched in the synapse-rich glomeruli of the olfactory bulb, suggesting a role in synaptic adhesion. Recently, a large family of more than 50 human neural cadherin-like cell adhesion genes, including CNRs, was described. These genes are organized in three tandem linked clusters, in which the N termini of different cadherin-like proteins, containing all six cadherin ectodomains, are encoded by a large exon. In contrast, the C termini of all cadherin-like proteins are identical and each terminus is encoded by three exons located downstream from the tandem array of N-terminal exons. Although it is not known whether these cadherin-like genes recombine in an immunoglobulin-like manner, the genomic organization of protocadherin genes allows for a considerable combinatorial variability of gene products that may form the basis for the cell-specific expression of certain isoforms.

CNRs in the brain have multiple functional roles. Extrapolating from published data on classic cadherins, CNRs are likely to interact homotypically. Assuming that up to 100 different CNR and protocadherin genes are expressed in brain in a partially cell-type-specific manner, the CNR proteins may play a role in synaptic recognition, allowing an arriving growth cone to distinguish its appropriate partner cell from a large selection of possible candidates. Additionally, CNRs are likely to participate in Fyn-mediated intracellular signaling, suggesting an interesting functional parallel between CNRs and cadherins. Data on deletion mutant mice demonstrate that Fyn is critical for LTP induction and spatial learning; given that CNRs bind to Fyn, these novel cadherin-like proteins may also be instrumental in the induction of LTP, resembling functional characteristics of cadherins. Finally, members of the CNR family interact with reelin, a protein that is an essential cue during neuronal migration.

3. β-Neurexins and Neuroligins

Although homotypic cell adhesion molecules such as cadherins or CNRs may play a critical role in synaptic recognition and adhesion, the asymmetry of synapses is likely to be mediated in part by heterotypic transsynaptic signaling. In fact, heterotypic cell adhesion has to be postulated to explain certain aspects of synaptogenesis, such as the specific and differential recruitment of pre- and postsynaptic protein components to their respective subcellular compartments. The β-neurexin/neuroligin junction represents the only known heterotypic cell adhesion system present in synapses.

Neurexins constitute a family of brain-specific cell surface molecules. The mammalian genome contains at least three homologous neurexin genes (neurexins I–III). Each of these is controlled by two different promoters, resulting in the generation of two primary transcripts per gene (α and β isoforms). α-Neurexins are composed of a signal peptide, a large extracellular N terminus, a single transmembrane domain, and a highly conserved intracellular C-terminal tail. In addition to an O-linked sugar attachment sequence

close to the transmembrane domain, the extracellular region of α-neurexins is assembled from three overall LNS(A)–EGF–LNS(B) repeats, each of which carries a central epidermal growth factor (EGF)-like sequence flanked by two subdomains [LNS(A) and LNS(B)] with homology to repeats found in the G domain of laminin A, in sex hormone-binding globuline, agrin, protein S, and several other proteins (Fig. 2). β-Neurexins are truncated forms of the respective α variants, containing a unique signal peptide and short N-terminal sequence stretch that is spliced into the α-neurexin sequence after the third EGF domain (Fig. 2). Extensive alternative splicing of the six primary neurexin transcripts at five splice sites (that can carry 2–12 alternatively spliced inserts) results in a combinatorial polymorphism of neurexin gene products that is unparalleled in cell adhesion molecules. Because many of these neurexin variants exhibit differential expression patterns in the brain, it was suggested that a given neuronal subpopulation may express a unique set of neurexins, which in turn could contribute to the specificity of neuronal cell–cell contacts in the brain.

Neuroligins were originally discovered as isoform-specific and splice site-specific ligands of β-neurexins (Δ4) that constitute a family of three brain-specific type I membrane proteins (neuroligin 1–3). Their large extracellular, β-neurexin-binding N-terminal domain is homologous to serine esterases such as acetylcholine esterase but lacks catalytic activity. It is followed by a transmembrane segment and a shorter intracellular tail (Fig. 2). Several prominent homologs of neuroligins are involved in cell adhesion in *Drosophila*. In addition, the intracellular tail of neuroligin binds to PSD95, a PDZ domain-containing membrane-associated guanylate kinase that is thought to be involved in the assembly and organization of signal transduction complexes in postsynaptic densities by acting as a multifunctional scaffold protein. This suggests that neuroligins may not only serve as adhesion molecules between cells but also are able to influence the assembly of protein complexes involved in signal transduction.

The characteristics of β-neurexins and neuroligins gave rise to a model according to which the two proteins form intercellular junctions in the brain. In the rat central nervous system, neuroligin 1 resides in postsynaptic densities, its extracellular tail reaching into the synaptic cleft, placing the β-neurexin/neuroligin junction at the synapse. Like glutamate receptors, neuroligin 1 is specifically localized to excitatory synapses, suggesting that it may not simply serve to form transsynaptic contacts between pre- and postsynaptic compartments. Rather, it may be involved in the determination of synapse specificity, distinguishing excitatory from inhibitory contact sites and recruiting protein components that are specific for excitatory synapses. The highly specific localization of neuroligin 1 is paralleled by its interaction partner PSD95 and several other postsynaptic proteins. Given that all three neuroligins are coexpressed in almost all neurons of the rat brain, a role of the β-neurexin/neuroligin junction in the determination of cellular synaptic specificity is very unlikely. Rather, neuroligins may function in a late step of synaptogenesis and/or the modulation of mature synapses. Indeed, neuroligins may not be essential for axonal pathfinding or initial synaptogenesis *in vivo*, although they are sufficient to trigger the generation of synapse-like structures between neurons and "postsynaptic" fibroblasts that overexpress neuroligins.

4. Neural Cell Adhesion Molecule

NCAM is a member of the immunoglobulin superfamily. It is a glycosylated protein expressed in most cells of the central and peripheral nervous system, muscle, heart, and gonads. NCAM is concentrated at contact sites between cells and mediates cell adhesion and recognition by intercellular homophilic interactions. Three major isoforms of NCAM are expressed: NCAM-A, -B, and -C. The A and B isoforms are single-pass transmembrane proteins with different intracellular domains, whereas NCAM-C is attached to the extracellular membrane surface by a glycosylphosphatidylinositol anchor. The extracellular part of all NCAM isoforms is composed of five N-terminal immunoglobulin domains (IgI–IgV), followed by two fibronectin type III domains (Fig. 2). Alternative splicing and posttranslational modification generate additional structural and functional diversity. Particularly, glycosylation with polysialic acid is an important regulatory process in NCAM function.

NCAM is present at synaptic junctions in several areas of the brain. *In vitro* studies demonstrate that NCAM stimulates axonal growth. In contrast, NCAM is not essential for most axon pathfinding and synaptogenesis processes in the intact brain. Nevertheless, interference with NCAM function by various methods has interesting consequences for several aspects of brain function, most notably deficiencies in neuronal migration, axon pathfinding or fasciculation, and physiological (LTP) as well as behavioral correlates of learning and memory. Evidently, NCAM function is not central to synaptogenesis but does

influence dynamic changes in synaptic strength and adhesion.

With respect to a possible functional role in synaptogenesis, fasciclin II, a close *Drosophila* homolog of NCAM, exhibits unique characteristics. Like NCAM, fasciclin II is involved in axon fasciculation and sprouting. However, it has an additional role in synaptogenesis and synaptic plasticity. Fasciclin II likely forms transsynaptic contacts and associates with PDZ domain-containing proteins and receptors through their intracellular C terminus. During evolution, the fasciclin II system may have been replaced by other molecular processes.

5. Integrins

Integrins are cell surface proteins responsible for cell–cell and cell–matrix interactions. In addition to the families of Ig-like cell adhesion molecules, cadherins, and selectins, they represent one of the largest families of cell adhesion proteins. Integrins are heterodimeric transmembrane proteins consisting of an α and a β subunit. Each subunit has a large N-terminal extracellular domain, a single transmembrane region, and a cytoplasmic C terminus. In vertebrates, 15 α and 8 β subunits are known, which together can form more than 20 heterodimeric integrin complexes. The α subunits are characterized by seven homologous repeats in their N-terminal extracellular region, some of which have functional EF-hand regions that bind divalent cations. The β subunits carry a cation binding site in their N-terminal extracellular region that is followed by four cysteine-rich repeats (Fig. 2). Intracellularly, integrins bind to and recruit linker proteins and components of signal transduction modules that link the proteins to the cytoskeleton and allow integrin-mediated signaling to various subcellular compartments.

Depending on their subunit composition, integrins are essential for many physiological processes most commonly by mediating cell–matrix interactions. Binding of extracellular matrix components induces integrin-mediated signaling to regulate cytoplasmic kinases, growth factor receptors, or ion channels and to control cytoskeleton dynamics or cell cycle progression.

Several integrins are found to be specifically enriched in synapses, mostly at neuromuscular junctions. However, the role of integrins in synapses is only beginning to emerge. Recent studies suggest that they are involved in synaptogenesis and synaptic specificity at *Drosophila* and vertebrate neuromuscular junctions as well as in the formation, maintenance, or plasticity of central nervous system synapses. Signaling mechanisms that mediate these effects are poorly understood, particularly in synapses of the central nervous system. They may involve the regulation of the pre- and postsynaptic cytoskeleton as well as receptors and ion channels.

6. Hierarchical Organization of Synaptic Cell Adhesion Molecules

The synaptic expression of a large number of cell adhesion molecules from different protein families suggests that these proteins function, at least in part, in consecutive steps of synaptogenesis. The unusually large families of CNRs and protocadherins provide sufficient combinatorial diversity to serve as identifiers of neurons. They are good candidates for mediators of initial contact formation and synaptic specificity. On the other hand, cell adhesion systems such as the β-neurexin/neuroligin junction are linked to intracellular protein–protein interaction cascades that are thought to be involved in the specific recruitment of synaptic proteins and synapse maturation. Finally, cell adhesion molecules such as NCAM and integrins are likely to function in synapse maturation and plasticity by regulating synaptic adhesion and intracellular signaling cascades.

IV. ASSEMBLY AND MOLECULAR ARCHITECTURE OF THE SYNAPSE

Essentially supporting the operational view of two phases in synaptogenesis, initial contact formation and assembly, morphological and functional studies demonstrate that a nascent growth cone forms an immature synaptic structure on initial contact with its target cell, which then matures into a fully functional synapse (Fig. 3).

As mentioned previously, assembly is most likely initiated and regulated by synaptic cell adhesion molecules. Given the asymmetric character of synapses with morphologically, functionally, and molecularly distinct pre- and postsynaptic compartments, heterophilic transsynaptic interaction of adhesion molecules is likely to be an important contributor to synapse assembly.

Analysis of the molecular architecture of synapses has revealed an extremely complex cascade of protein–protein interactions that are thought to be the molecular basis of the synapse assembly process. It is

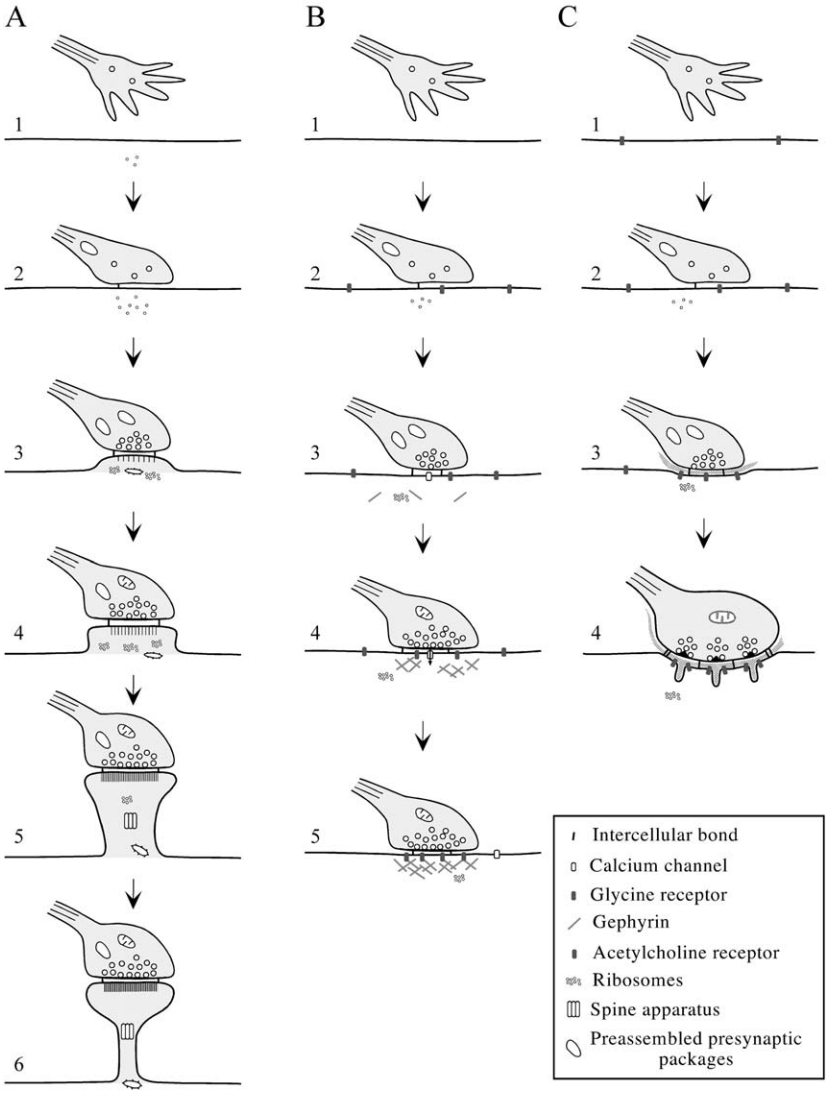


Figure 3 Formation of synapses. (A) Central asymmetric synapse (or type I) on a dendritic spine; (B) central symmetric synapse (or type II) on a cell soma; (C) neuromuscular synapse on a myotube. (A) Glutamatergic synapse. The growth cone approaches the membrane (1) and forms a specific but initially unspecialized contact (2), likely driven by a homotypic adhesion system (CNRs or cadherins). Presynaptically, putative preassembled active zones and scarce synaptic vesicles are present. Soon thereafter, Bassoon is recruited to the presynaptic site. A heterotypic adhesion system (β-neurexin/neuroligin) may coordinate and drive pre- and postsynaptic differentiation (3). Numerous polyribosomes may reflect high local protein synthesis. Progressively (4 and 5), functional presynaptic active zones are formed. Synaptic vesicles accumulate. The spine is taking shape, and a spine apparatus is organized. First-order scaffold proteins (e.g., SAP90/PSD95) accumulate at postsynaptic sites, followed by recruitment of specific receptors, ion channels, and second-order interaction partners. Glutamate receptors accumulate at postsynaptic sites (6). (B) Glycinergic synapse. The growth cone approaches the membrane (1) and forms a specific but initially unspecialized contact, possibly triggered by a homotypic adhesion system (CNRs and cadherins). At that point, glycine receptors are freely diffusing in the postsynaptic membrane (2). Calcium influx induced by presynaptic release triggers gephyrin accumulation (3). Gephyrin aggregates (4) and clusters glycine receptors at the subsynaptic membrane (5). (C) Cholinergic neuromuscular synapse. The growth cone approaches an already formed muscle cell (1) and forms an unspecialized contact capable of rudimentary neurotransmission. Acetylcholine receptors are diffusely distributed over the muscle cell membrane (2). The presynaptic terminal differentiates and accumulates synaptic vesicles. In parallel, the basal lamina forms and the muscle cell matures. Electric activity induces further synthesis of acetylcholine receptors near the synaptic input and suppresses synthesis in nonsynaptic areas (3). The junction matures, bonds between pre- and postssynapse induce the formation of folds, and different postsynaptic components are differentially distributed. In particular, acetylcholine receptors cluster at the edges of the folds (4).

likely that assembly of active zones and postsynaptic signal transduction systems is not just occurring at the level of the synapse. Rather, various protein complexes or even preassembled active zones may be put together at the level of the cell body and transported *in toto* to the developing synaptic contact, where they can be integrated into the newly forming synapse in response to the activity of an organizer protein or other stimulus.

Presynaptically, synapses of different types are very similar with respect to their protein composition, and only very few presynaptic proteins are known to be specifically expressed in only one type of synapse but not in others. This indicates that the assembly of presynaptic compartments follows the same general mechanisms, irrespective of the cell type under investigation. In contrast, the postsynaptic compartments of different types of synapses are composed of strikingly different sets of proteins, suggesting that alternative pathways of postsynaptic assembly are used.

A. Presynapse

In most cases studied, presynaptic differentiation precedes postsynaptic differentiation during synaptogenesis. Even before synapse formation, axonal growth cones exhibit vesicular fusion and irregular quantal release of transmitter from vesicles. This phenomenon likely contributes to growth cone extension during axon elongation. In addition, the resulting electrical activity may aid in early phases of synaptogenesis. However, mice with a completely dysfunctional synaptic secretion apparatus initially form a normal neuronal network that later degenerates. This suggests that sporadic neurotransmitter release from nascent presynaptic terminals/growth cones may not be required for synaptogenesis but instead be important for synapse maturation and maintenance. Upon contact between growth cone and target structure, a specific molecular adhesion reaction is thought to trigger synapse formation. Functional presynaptic boutons can form within 30 min of the establishment of an axodendritic contact, and 1 or 2 h may suffice to form a glutamatergic synapse.

Immature presynaptic compartments contain a small number of synaptic vesicles, cytoskeletal components, dense-core vesicles, tubulovesicular structures, and pleiomorphic vesicles. Some of the larger vesicular structures contain active zone and synaptic vesicle proteins. They are thought to represent preassembled presynaptic active zones that can be rapidly integrated into the presynaptic plasma membrane following contact formation between axon and target structure or an appropriate signal. Bassoon or Piccolo, two multidomain proteins of the presynaptic active zone, may function as organizers of active zone transport vesicles. As the synapse matures, synaptic vesicles accumulate, probably recruited from the cell soma or neighboring terminals (Fig. 3).

The physical contact between axon and target membrane appears to stabilize the presynapse and coordinate the development of pre- and postsynaptic specializations. Although the underlying coordinating signal is not known in central nervous system synapses, it is likely to involve cell adhesion molecules. Good candidates are cadherins and protocadherins as well as the β-neurexin/neuroligin system. The latter in particular is capable of coupling cell adhesion to intracellular protein interaction cascades that could ultimately lead to the assembly of presynaptic release sites or the integration of preassembled active zone complexes at the site of transsynaptic contact (Fig. 3A). β-Neurexins bind to the PDZ domain of calmodulin-dependent protein kinase, which is thought to serve as a first-order organizer molecule by recruiting/binding to components of the secretory apparatus via Mint1, Munc18, and Velis. Indeed, experiments using cocultures of neurons with neuroligin-overexpressing fibroblasts indicate that, at least *in vitro*, postsynaptically expressed neuroligin in fibroblasts may be sufficient to trigger presynaptic accumulation of synaptic vesicles in a contacting axon.

B. Postsynapse

1. Asymmetric/Type 1

Asymmetric synapses in the central nervous system are almost exclusively formed between axons and dendritic spines. Formation and maturation of synapses is concomitant with a striking morphological change. Most type 1 synapses are thought to be induced by contact between an axon and a dendritic filopodium. After retraction of the filopodium, a synapse onto the shaft of the dendrite is created. Spines are then formed from stubby protrusions emerging beneath the synapse (Fig. 3A). During growth of the protrusion and spine, polyribosomes accumulate in the postsynaptic compartment, indicative of local protein synthesis and its regulation by activity. The polyribosome number is reduced again in mature synapses. The mature spine

sometimes contains a spine apparatus that is a specialized form of smooth endoplasmic reticulum and also a prominent postsynaptic membrane specialization, the postsynaptic density, that is barely detectable in early shaft synapses and develops as the spine matures.

The postsynaptic specialization of asymmetric type 1 synapses contains a large number of proteins (Fig. 4). These include primary signal transduction molecules such as receptors and ion channels in the plasma membrane, secondary signal transduction proteins such as kinases and other components of intracellular signal transduction pathways, as well as organizer and scaffold proteins. This elaborate protein network is thought to be arranged in different organizational layers. CNRs and/or other protocadherins are believed to mediate the initial specific contact between axon and target membrane. Cell surface molecules of the cadherin family may then mediate synaptic adhesion and serve as cytoskeletal anchors. For example, N-cadherin and its intracellular catenin interactors are localized in a parasynaptic manner, flanking the presynaptic release sites and postsynaptic densities in type 1 synapses, compatible with a role in synaptic adhesion but not with a function in the formation and organization of the postsynaptic specialization. The β-neurexin/neuroligin system is thought to initiate the assembly of the postsynaptic signal transduction apparatus. Neuroligin 1, which resides in the postsynaptic density of type 1 synapses and is thought to form a transsynaptic connection with β-neurexin at the initial synaptic contact, could recruit a set of first-order organizing/scaffold proteins (PSD-95, S-SCAM, GRIP, and others). Their common feature is the presence of multiple PDZ domains that essentially serve as insertion sites for various postsynaptic proteins. PSD-95 and S-SCAM, for example, bind to neuroligins with one of their PDZ domains and to NMDA receptor subunits with another, thereby allowing the synaptic cell adhesion molecule neuroligin to recruit and aggregate NMDA receptors at the developing synaptic contact. Other domains of PSD-95 could recruit kainate receptors in a similar manner. Moreover, AMPA receptors may be recruited through their interaction with GRIP, although it is not clear whether GRIP can simultaneously bind to neuroligins. In addition to receptors, first-order organizing proteins of the PSD-95/S–SCAM/GRIP type bind to

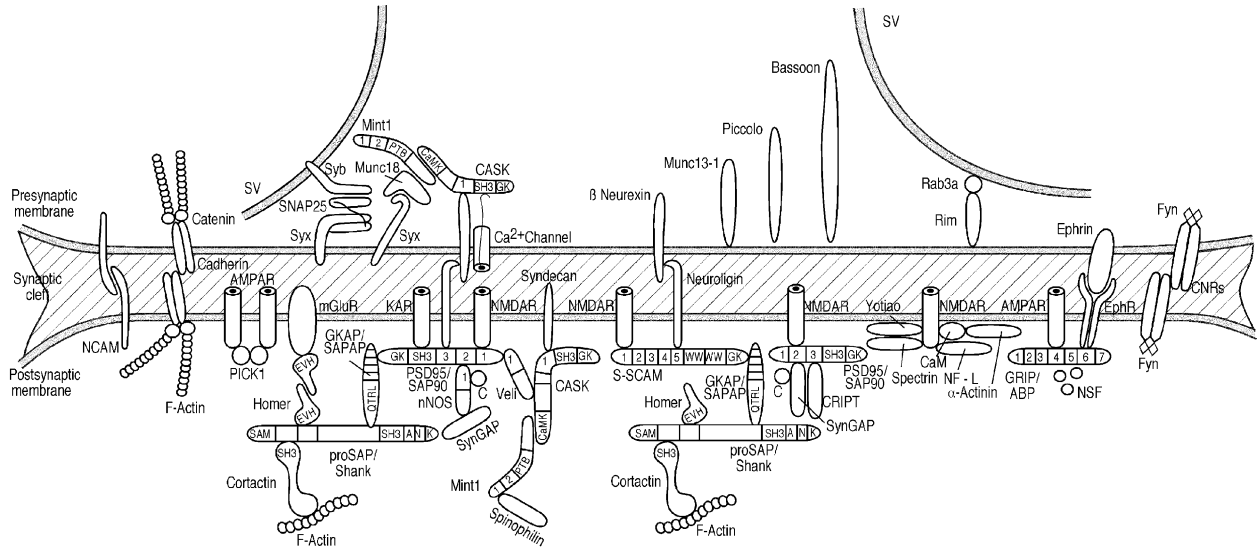


Figure 4 Molecular architecture of the glutamatergic synapse. Schematic representation of the major synaptic proteins with an emphasis on postsynaptic components. F-actin, filamentous actin; AMPAR, α-amino-3-hydroxy-5-methyl-4-isoxasole propionic acid receptor; CaM, calmodulin; C, capon; CASK, calmodulin-dependent protein kinase; CNRs, cadherin-like neuronal receptors; CRIPT, cysteine-rich interactor of PDZ3; EphR, ephrin receptor; GK, guanylate kinase; GKAP/SAPAP, guanylate kinase-associated protein/synapse-associated protein-associated protein; GRIP/ABP, glutamate receptor-interacting/AMPA receptor-binding protein; KAR, kainate receptor; mGluR, metabotropic glutamate receptor; NCAM, neural cell adhesion molecule; NF-L, neurofilament L; NMDAR; *N*-methyl-D-aspartate receptor; nNOS, neuronal nitric oxide synthase; NSF, *N*-ethylmaleimide-sensitive factor; PICK1, PKCα-interacting protein 1; proSAP/Shank, proline-rich synapse-associated protein; PSD95/SAP90, postsynaptic density protein of 95 kDa/synapse-associated protein; SH3, src homology domain 3; SNAP 25, soluble NSF attachment protein; S-SCAM, synaptic scaffolding molecule; SV, synaptic vesicle; Syb, synaptobrevin; SynGAP, synaptic GTPase-activating protein; Syx, syntaxin.

several cytoplasmic proteins, including scaffold proteins and components of signal transduction cascades. In turn, some of these interact with second-order organizing proteins such as ProSAP/SHANK, which provide a link to metabotropic glutamate receptors and the cytoskeleton.

In principle, the protein interaction cascades depicted in Fig. 4 provide a framework for recruitment processes that ultimately result in the assembly of the postsynaptic specialization. An important open question concerns the specificity of PDZ domain/target protein interaction *in vivo*. In many *in vitro* studies, these interactions are rather promiscuous, and as yet uncharacterized differences in affinities between PDZ domains and their respective targets as well as subcellular compartmentalization processes are likely contributors to an ordered assembly of the complex postsynaptic protein network. As with presynaptic active zones, it is possible that parts of this protein network are already assembled at the level of the cell body and transported to the developing synapse and inserted into the postsynaptic membrane *in toto*. Interestingly, the generation of asymmetric synapses appears to be dependent on and regulated by synaptic activity. The molecular mechanisms of such regulation are not known, but they may involve signaling through NMDA receptors.

2. Symmetric/Type 2

Symmetric synapses are mostly formed between axons and cell somata. In comparison to asymmetric synapses, they develop a much less pronounced postsynaptic thickening (Figs. 1B and 3B). These differences in morphology are paralleled by striking differences in molecular composition. Symmetric synapses exhibit a unique set of proteins in their postsynaptic compartment whose assembly into a functional postsynaptic specialization follows similarly unique molecular mechanisms.

Adhesion processes at developing symmetric synapses are likely to involve cadherins and related proteins. Catenins, for example, are found at symmetric synaptic contacts, suggesting that classic cadherins are acting as synaptic adhesion molecules in these synapses. Interestingly, the only characterized neuroligin isoform, neuroligin 1, appears to be specific for asymmetric synapses.

With respect to molecular mechanisms of assembly, glycinergic synapses are the best characterized inhibitory synaptic contacts. However, GABAergic synapses may be formed according to similar assembly mechanisms. Gephyrin, an intracellular scaffolding protein and enzyme cofactor, is essential for the formation of postsynaptic glycine receptor and GABA receptor clusters (Fig. 5). Gephyrin is thought to be recruited to synapses and assembled into a submembranous scaffold in response to glycine receptor activation and Ca^{2+} influx following presynaptic glycine release. Due to the altered Cl^- equilibrium potential, embryonic neurons respond to glycine with depolarization, resulting in Ca^{2+} influx through voltage-gated channels. An additional factor contributing to gephyrin recruitment may be activation of phosphatidylinositol 3-kinase through an unknown mechanism. This would lead to increases in membrane 3′-phosphorylated phosphoinositide (PIP_3) content and recruitment of collybistin and profilin, a regulator

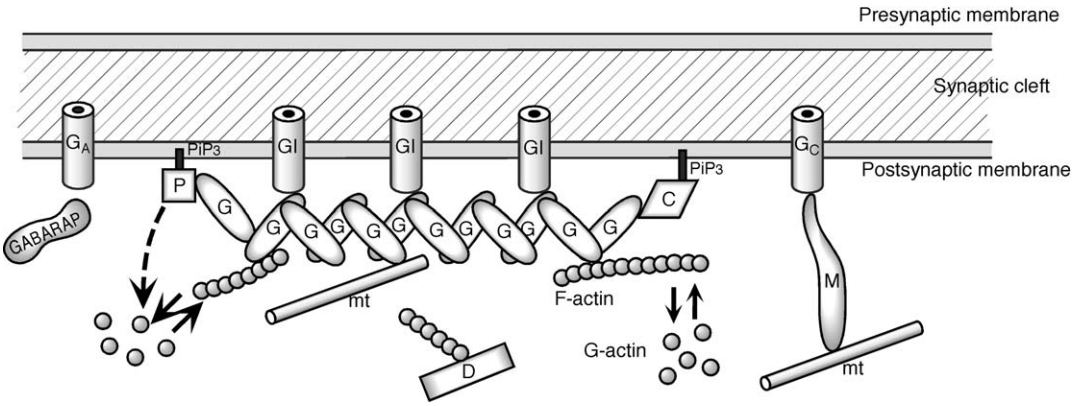


Figure 5 Molecular architecture of the GABA/glycinergic postsynapse. C, collybistin; D, dystrophin; F-actin, filamentous actin; G, gephyrin; GA, $GABA_A$ receptor; GABARAP, $GABA_A$ receptor-associated protein; G-actin, globular actin; GC, GABAC receptor; Gl, glycine receptor; M, MAP-1; mt, microtubules; P, profilin; PIP_3, 3′-phosphorylated phosphoinositides.

of the actin cytoskeleton, to the activated membrane domains. In turn, these induce cytoskeletal changes and apposition of gephyrin to the PIP_3-rich plasma membrane by binding to its PI-anchored interaction partners. Accumulation of gephyrin at activated postsynaptic membranes would then lead to recruitment and clustering of glycine receptors. A similar gephyrin-dependent mechanism may be active in $GABA_A$ receptor clustering, although direct interaction between $GABA_A$ receptors and gephyrin has not been demonstrated. In addition, a novel $GABA_A$ receptor binding protein, GABARAP, appears to be important in $GABA_A$ receptor transport to the synapse.

3. Neuromuscular Junction

The postsynaptic compartment of neuromuscular junctions undergoes significant structural changes as the synapse develops and matures. The most striking features are the presence of a basal lamina in the synaptic cleft and the formation of folds in the postsynaptic plasma membrane (Fig. 3C). Folds are created as the muscle fiber grows but is constrained by transsynaptic bonds between pre- and postsynaptic membrane via the basal lamina.

The specificity of neuromuscular connections is thought to be mediated by a combination of guidance, recognition, and synapse elimination/conversion mechanisms. Recognition and adhesion may involve ephrins and their receptors as well as integrins and NCAM. The assembly of a postsynaptic specialization at neuromuscular junctions, particularly the accumulation of acetylcholine receptors at subsynaptic sites, is influenced by three distinct signaling processes (Fig. 6).

Receptor clustering at the developing synaptic contact is induced by agrin, a proteoglycan that is synthesized by motor neurons, released from motor nerve terminals, and inserted into the basal lamina of the synaptic cleft. Agrin binds to a muscle-specific kinase (MuSK) in the subsynaptic membrane, most likely via an unidentified coreceptor (MASC). Activated MuSK leads indirectly (e.g., via a src family kinase) to the recruitment of a critical downstream effector, rapsyn, which contains domains responsible for membrane binding, oligomerization, and acetylcholine receptor binding. In turn, rapsyn is essential for receptor clustering. It associates with receptors in a 1:1 stoichiometry and accumulates them in the subsynaptic membrane. In addition, MuSK appears to be an important structural component of the primary synaptic scaffold.

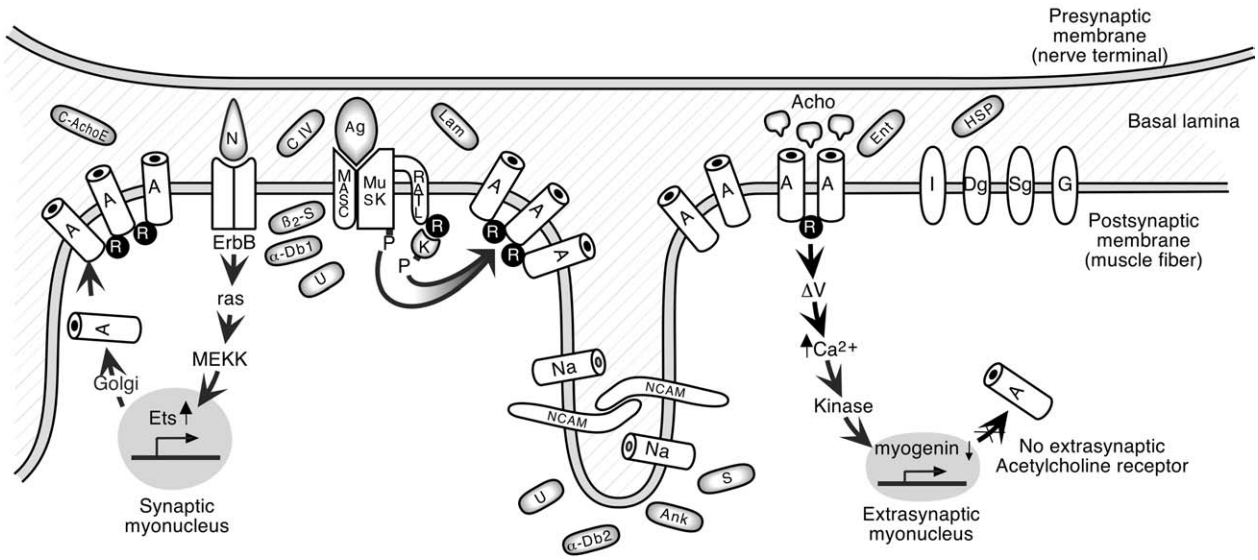


Figure 6 Molecular architecture of the neuromuscular postsynapse. A, acetylcholine receptor; Ach, acetylcholine; Ag, agrin; Ank, ankyrin; C-AchE, collagen-tailed acetylcholine esterase; C IV, collagen IV; α-Db1/2, α-dystrobrevin 1 and 2; D, dystrophin; Dg, dystroglycan; Ent, entactin; G, Gal *N*-acetyl transferase; HSP, heparan sulfate proteoglycans; I, integrin dimer; K, MuSK-activated src-like kinase; Lam, laminin; MASC, myotube-specific accessory component; MEKK, MAP/ERK kinase kinase; MuSK, muscle-specific kinase; N, neuregulin; Na, sodium channel; NCAM, neural cell adhesion molecule; R, rapsyn; RATL, rapsyn-associated transmembrane linker; β_2-S, β_2-syntrophin; Sg, sarcoglycan; U, utrophin.

Apart from agrin-mediated receptor clustering, the induction of synapse-specific transcription of acetylcholine receptors contributes to the development of postsynaptic specializations at neuromuscular junctions. In this context, the most important factor is neuregulin, which, like agrin, is released from motor nerve terminals and incorporated into the synaptic basal lamina. Neuregulin binds to receptor tyrosine kinases of the erbB family, some of which are enriched in the postsynaptic membrane. These receptor tyrosine kinases are the triggering components of an intracellular signaling cascade involving ras, raf, erk, PI_3 kinase, and transcription factors of the ets family. The latter are able to induce transcription of acetylcholine receptor subunit genes by binding to specific elements in the respective genomic sequence. The synapse specificity of this process results from the localization of ligands, receptors, and signaling molecules as well as additional transcriptional control mechanisms. Indeed, transcription of nicotinic receptor genes is restricted to subsynaptic nuclei.

In contrast to the synapse-specific induction of acetylcholine receptor gene expression, extrasynaptic receptor expression is repressed by a third regulatory mechanism at neuromuscular junctions. Here, the signaling molecule is acetylcholine. Binding of acetylcholine to its receptors leads to a depolarization of the muscle fiber and thereby triggers an action potential that is propagated. As a consequence of the action potential, Ca^{2+} levels increase in the cytosol mainly due to efflux from the sarcoplasmic reticulum. Apart from inducing muscle contraction, these Ca^{2+} ions are now able to repress acetylcholine receptor transcription. The key target is protein kinase C, which in turn inactivates/downregulates myogenic factors such as myoD, myf5, MRF4, and myogenin, resulting in reduced receptor expression.

V. CONCLUSION

The establishment of a functional neuronal network in the nervous system follows a complex developmental program that culminates in synaptogenesis, the formation of specific functional contacts between nerve cells. The specificity of synapse formation is thought to be caused by synaptic cell surface molecules that are responsible for appropriate target recognition and synaptic adhesion. Assembly of functional synapses at initial contact sites is most likely achieved by an interplay of several factors, including transsynaptic adhesion systems capable of recruiting pre- and postsynaptic proteins as well as input-dependent and activity-dependent clustering and recruiting mechanisms. Depending on the type of synapse, these processes follow strikingly different molecular principles.

During the development of the nervous system, synaptogenesis is paralleled by massive synapse elimination. Synapse elimination is, at least in part, regulated by exogenous stimuli and synaptic activity patterns. It results in the fine-tuning of the neuronal network. Furthermore, learning and memory storage processes in the central nervous system lead to the formation of new synapses and the modification or elimination of existing synapses. At the molecular level, these processes are likely to involve, among others, the same proteins that are necessary for synaptogenesis during development, including cell adhesion proteins such as cadherins, protocadherins, integrins, and NCAM, receptors and their trafficking or clustering, as well as scaffold and organizer proteins or cytoskeletal components. Thus, further molecular analysis of synaptogenesis will not only provide insights into the basic principles of nervous system development but also contribute to our molecular understanding of plastic processes in the brain.

See Also the Following Articles

BRAIN DEVELOPMENT • EVOLUTION OF THE BRAIN • INFORMATION PROCESSING • NEURAL NETWORKS • NEUROPLASTICITY, DEVELOPMENTAL • SYNAPSES AND SYNAPTIC TRANSMISSION AND INTEGRATION

Suggested Reading

Brose, N. (1999). Synaptic cell adhesion molecules and synaptogenesis in the mammalian central nervous system. *Naturwissenschaften* **86,** 516–524.

Garner, C. C., Kindler, S., and Gundelfinger, E. D. (2000a). Molecular determinants of presynaptic active zones. *Curr. Opin. Neurobiol.* **10,** 321–327.

Garner, C. C., Nash, J., and Huganir, R. L. (2000b). PDZ domains in synapse assembly and signaling. *Trends Cell Biol.* **10,** 274–280.

Jacobson, M. (1993). *Developmental Neurobiology*. Plenum, New York.

Kim, J. H., and Huganir, R. L. (1999). Organization and regulation of proteins at synapses. *Curr. Opin. Cell Biol.* **11,** 248–254.

Kneussel, M., and Betz, H. (2000). Clustering of inhibitory neurotransmitter receptors at developing postsynaptic sites: The membrane activation model. *Trends Neurosci.* **23,** 429–435.

Lee, S. H., and Sheng, M. (2000). Development of neuron–neuron synapses. *Curr. Opin. Neurobiol.* **10,** 125–131.

Sanes, J. R., and Lichtman, J. W. (1999). Development of the vertebrate neuromuscular junction. *Annu. Rev. Neurosci.* **22,** 389–442.

Tactile Perception

CATHERINE L. REED
University of Denver

GLOSSARY

exploratory procedures Stereotypical hand movements used to extract distinctive object dimensions or qualities (e.g., pressure is used to identify hardness).

haptics Active touch; typically used for object recognition by touch.

homunculus Map of the body in somatosensory cortex; "little man."

secondary somatosensory cortex Cortical area located on the upper bank of the Sylvian fissure, ventrolaterally to primary somatosensory area, that processes information about touch, pain, and temperature.

somatosensory cortex Cortical strip located on the postcentral sulcus that receives input from the skin and is responsive to somatosensory stimulation.

somatosensory system Sensory system that includes the cutaneous senses, proprioception (the sense of limb position), and kinesthesis (sense of limb movement).

somatotopic map A map in somatosensory areas of the brain organized so neurons responsive to particular body parts are located next to neurons representing adjacent body parts.

Our sense of touch connects us physically with the external world. Tactile perception, or somesthesis, refers to our ability to apprehend information through touch. Not only do objects in the world touch us but also we explore our environment actively with our hands, fingers, and bodies. We use the motor capabilities of our hands to extract important characteristics necessary for identifying and using objects. Thus, tactile perception results from the stimulation of receptors in the skin and from proprioception. Our understanding of the neural bases of tactile perception is based on the study of perceptual and neurophysiological responses in animals and humans.

I. PSYCHOPHYSICS, ANATOMY, AND PHYSIOLOGY

A. Peripheral Somatosensory Receptors and Afferent Nerve Fibers

Tactile perception begins with the mechanical stimulation of the skin. Tactile information travels from cutaneous receptors to the afferent somatosensory pathways, the spinal cord and thalamus, and the brain. It is first collected and grouped by peripheral receptors in the skin, joints, and muscles. The hand is the primary organ for acquiring tactile information, although receptors are located throughout the body in both glabrous (nonhairy) and nonglabrous (hairy) skin. This article focuses on the receptors located in the glabrous skin of the hand that process nonnoxious information.

1. Psychophysics

To study the relationship between tactile perception and its neural mechanisms, psychophysical and physiological experiments have used several stimulating techniques designed to distinguish the response properties of the mechanoreceptors and their afferent nerve fibers. These techniques include the cooling of the skin and the presentation of masking vibrations to decrease

the sensitivity of one type of tactile channel to study another one. Electrophysiological studies using microneurography record the activity of the nerve fibers innervating the hand. Vibration has been applied perpendicularly and tangentially to the skin by pins and probes. Periodic and aperiodic gratings have been moved across the skin. Other stimuli include air puffs, embossed letters, steel wool, sandpaper, and cloth.

2. Mechanoreceptors and Afferent Nerve Fibers

Cutaneous mechanoreceptors transduce mechanical energy into electrochemical energy to create the neural signal. In glabrous skin, four major mechanoreceptors have been identified to respond to specific types of tactile information: Merkel disks, Meissner corpuscles, Pacinian corpuscles, and Ruffini cylinders. Each receptor corresponds to a specific type of afferent nerve fiber. The receptors and their associated nerve fibers respond selectively to certain types of mechanical stimulation, in that each has a lower threshold to a specific stimulus relative to the others.

Mechanoreceptors can be classified by the type of fiber, optimal stimulus, adaptation properties, and size of receptive field (Table I). Slowly adapting (SA) fibers signal the presence of stimulation and fire steadily while the stimulus is applied. Merkel discs and SA1 fibers respond best to 0.3- to 3-Hz frequencies that are perceived as pressure. They are slow to adapt and have small receptive field sizes. SA1 fiber activity is associated with the feeling of texture or fine detail. Ruffini cylinders and SA2 fibers respond best to 15- to 400-Hz frequencies that are perceived as buzzing. They are slow to adapt and have large receptive fields. SA2 fibers respond to the stretching of the skin or movements of the joints.

Rapidly adapting (RA) fibers signal stimulus onset by firing strongly and briefly when a stimulus is applied and when it is removed. Meissner corpuscles and RA1 fibers respond best to 3- to 40-Hz vibrations or taps on the skin that are perceived as flutter. They adapt rapidly and have small receptive fields. RA1 fiber activity is associated with rapid changes of pressure that occur when the hand feels a textured surface. Last, Pacinian corpuscles and RA2 fibers respond best to 10- to 500-Hz frequencies that are perceived as vibration. They adapt rapidly and have large receptive fields. RA2 fibers are specialized to respond to changes in stimulation.

3. Relationship between Nerve Fiber Response and Tactile Perception

How do the receptors and their nerve fibers give rise to our perceptions of touch? Although tactile perception is correlated with the properties of individual receptors and afferent fibers, it results primarily from activity across multiple nerve fibers. When we interact with common objects, such as a cold wet cup, different combinations of neural fibers tend to be activated. The brain combines these temporal and spatial patterns of impulses from a large number of different types of receptors to create complex tactile percepts.

4. Proprioception

In addition to cutaneous information, proprioceptive information is essential to tactile perception. Proprioception is signaled by specific receptors in muscles and joints to provide information about joint movement, position, and related forces. Without proprioception, the hand could not perform any exploratory procedures to determine the specific shape of an object or the size of a large object.

5. Temperature

Our perception of temperature also contributes to tactile object recognition. The perception of temperature is produced by the heating or cooling of the skin. Thermoreceptors respond to specific temperatures and changes in temperature but do not respond to

Table I
Properties of Mechanoreceptors and Afferent Nerve Fibers

Mechanoreceptor	Afferent nerve fiber	Optimal stimulus	Adaptation rate	Receptive field size
Merkel receptor	SA-I	Pressure	Slow	Small
Meissner corpuscle	RA-I	Taps on skin	Rapid	Small
Ruffini cylinder	SA-II	Skin stretch, joint movement	Slow	Large
Pacinian corpuscle	PC	Rapid vibration	Rapid	Large

mechanical stimulation. There are separate thermoreceptors for warm and cold.

B. Neural Pathways

Tactile signals from the body travel through the peripheral nerves to the dorsal root ganglia in the spinal cord. Once they enter the spinal cord, the nerve fibers go up the spinal cord in one of two pathways: the medial lemniscal pathway and the spinothalamic pathway. On the way to the thalamus, nerve fibers in both pathways cross over to the other side of the body so that signals originating from the left side of the body are represented in the right hemisphere of the brain and signals from the right side of the body are represented in the left hemisphere.

1. Medial Lemniscal Pathway

The medial lemniscal pathway consists of large fibers that convey information about aspects of touch and proprioception used to identify objects and surfaces (e.g., form, position, and temporal change). The neural signals travel via the medial lemniscal pathway to synapse in the ventroposterolateral thalamic nuclei. From the thalamus, signals travel to the post-central gyrus [somatosensory cortex (SI)], secondary somatosensory cortex (SII), and other higher order areas.

2. Spinothalamic Pathway

The spinothalamic pathway consists of smaller fibers that convey information primarily about pain and temperature. The neural signals travel via the spinothalamic pathway to the reticular formation, the intrinsic thalamic nuclei, and the SII.

3. Thalamus

All somatosensory signals are processed by specific thalamic nuclei. The ventroposteromedial nucleus receives cutaneous inputs from the face and the ventroposterolateral nucleus receives inputs from the rest of the body. These nuclei have precise somatotopy to maintain the stimulus' location on the body. The ventroposteroinferior (VPI) and the medial posterior nucleus (Pom) may also have functional importance for the processing of tactile information in that they have cutaneous receptive fields and project to cortical somatosensory areas.

C. Somatosensory Cortices

Much of our knowledge regarding the anatomy and function of cortical areas comes from research on animals and brain-injured humans. However, recently noninvasive neuroimaging techniques have permitted the investigation of the relationship between stimulus characteristics (e.g., frequency and amplitude) and physiological responses in the brains of non-brain-damaged humans. These techniques include magnetoencephalography/electroencephography, positron emission tomography, and functional magnetic resonance imaging. Brain activity is indexed by changes in cerebral blood flow, blood oxygenation levels, and magnetic field generation.

When the skin is touched, tactile information is sent to somatosensory areas of the brain (Fig. 1). If neurons in a particular cortical area have predominant or

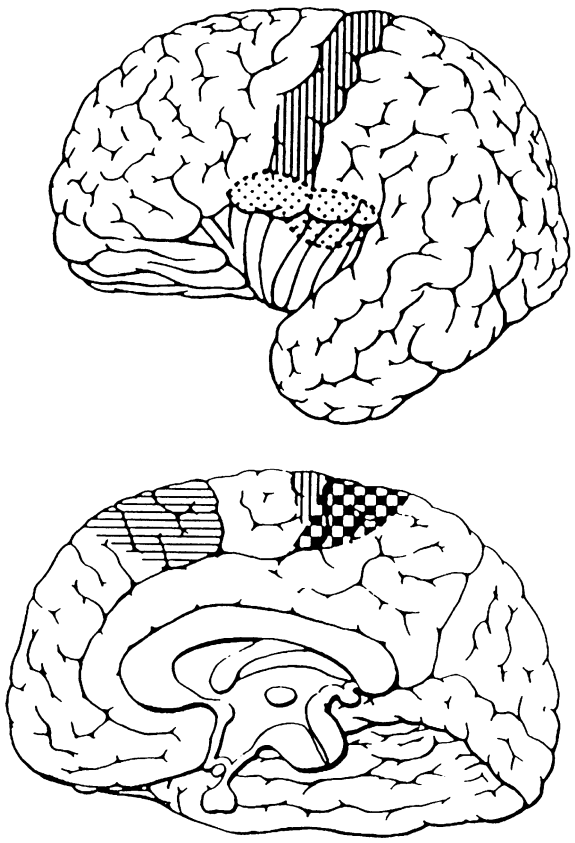


Figure 1 Schematic diagram of ventrolateral (top) and dorsomedial (bottom) somatosensory areas in the human brain. Vertical lines indicate SI, and small dots indicate SII, parietal operculum, and posterior insula. Checkerboards indicate SSA, and horizontal lines indicate SMA (from Caselli, R. J., 1997, by permission of Mayo Foundation).

exclusive responses to somatosensory stimuli, the area is considered to be involved in tactile perception. In monkey and human brains, the major cortical areas considered to be somatosensory cortices are SI, SII, parietal areas (areas 5 and 7b), the retroinsular cortex (Ri), and the insula.

1. Primary Somatosensory Cortex

Primary somatosensory cortex, or SI, is a region of postcentral cortex activated by tactile somatic stimuli. It includes Brodmann areas 1–3 (3a and 3b). These areas of cortex are activated almost exclusively from the contralateral side of the body. The neurons in primary somatosensory cortex are arranged in columns so that neurons with receptive fields processing information from like areas of the body are located together.

a. Somatotopic Organization Each of the four subdivisions of SI has a complete representation of the body, or a homunculus (Fig. 2). The homunculus has a medial-to-lateral organization, with the representations of the inferior limbs at the vertex, the upper limbs and mouth at the convexity, and the tongue at the bottom of gyrus. In the homunculus, adjacent body parts are represented primarily next to each other, but the cortical organization does not exactly match the body surface. For example, the representation of the hand lies between the back of head and the face.

b. Cortical Magnification There is a relative magnification of the amount of cortical area devoted to body parts with the highest tactile acuity and sensory specialization (Fig. 2). For example, the lips, fingers, and hands are the primary organs for extracting tactile information and they have proportionately larger representations. Cortical magnification means that there are many cortical cells for each afferent nerve fiber. This increase in the number of cells processing the same sensory input suggests that incoming

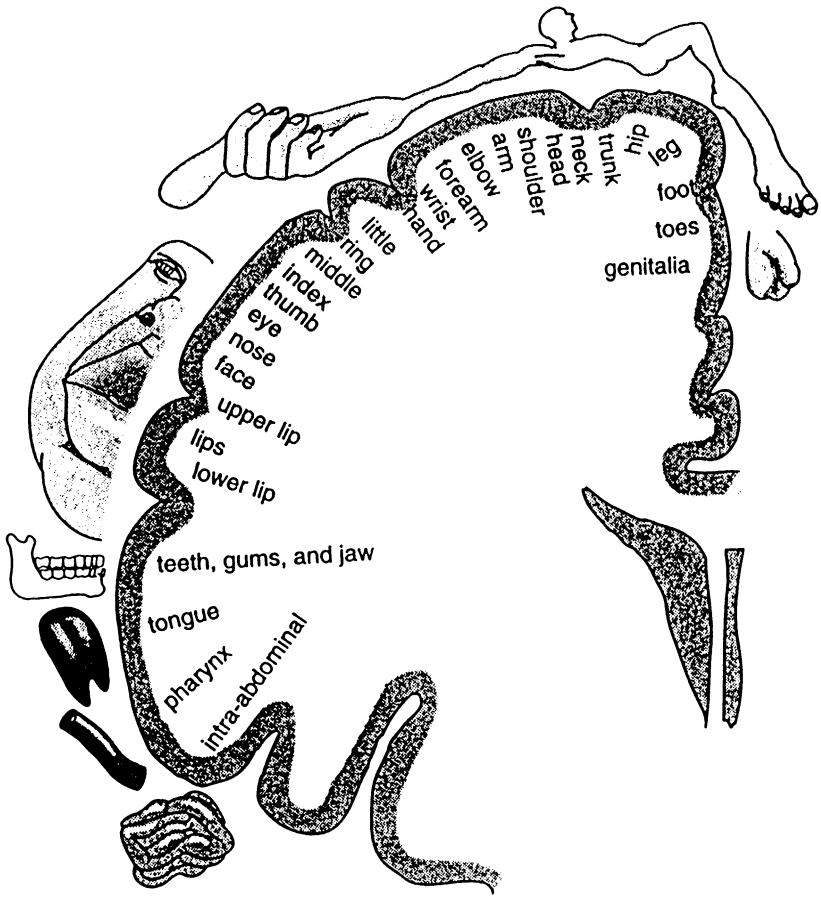


Figure 2 The homunculus on somatosensory cortex. Body parts with the highest tactile sensitivity occupy larger areas of cortex. [from Penfield and Rasmussen, (1950). *The Cerebral Cortex of Man: A Clinical Study of Localization and Function.* Macmillan, New York].

information may be elaborated and recoded. Higher level processes, such as object identification, may require information in a different format from which it was encoded.

c. Multiple Body Representations The different body representations are not just duplications of each other; areas 1 and 3b have two mirror-reversed body maps of cutaneous receptors. Areas 2 and 3a respond primarily to deep stimulation of the joints and muscles, although area 2 does receive cutaneous inputs. One reason for multiple body representations is that different areas within the somatosensory cortex have different functions.

d. Function To determine the function of SI, monkeys with damage to postcentral cortex have been studied. Lesions in area 3b affect all tactile discrimination, with the exception of crude touch. Lesions to area 1 impair texture discrimination. Lesions to area 2 impair the discrimination of shape and size as well as finger coordination. Supporting these functions, single-cell recording research has located orientation- and direction-selective neurons in area 2. In humans, damage to SI is rarely restricted to one area. Clinical data from extensive postcentral lesions suggest that damage to SI produces lasting impairments of pressure sensitivity, two-point discrimination, point localization, and the discrimination of object shape, size, and texture. In contrast, cortical lesions rarely affect the detection of touch, pain, and temperature.

2. Secondary Somatosensory Cortex

SI is bounded ventrolaterally by secondary somatosensory cortex (or the parietal operculum). In monkeys it is located in the parietal bank of the Sylvian sulcus behind the insula. In humans, SII includes the parietal operculum, and it appears to extend more posteriorally to the inferior parietal cortex. SII receives its peripheral inputs from SI and from the VPI thalamus. There are reciprocal connections between SI and SII. It has many direct connections with all somatosensory areas with the exception of area 5. Unlike SI, in which inputs are almost entirely contralateral, some proportion of SII neurons receive input from both sides of the body. Thus, SII can be activated bilaterally with unilateral stimulation.

a. Somatopic Organization SII also has somatotopic organization. Although the entire body is systematically represented in SII, the digits of hand and the face representations occupy most of the total area. According to cortical stimulation studies, human SII cortex has the face area in the most superficial part, the foot area in the most medial part, and the hand area in between.

b. Function The role of SII in tactile perception is not certain. Evidence suggests that SII is part of a higher-order association center for tactile object recognition and learning. In monkeys, damage to SII produces severe impairments of shape discrimination, impairments of learning, and altered size and roughness discrimination. In humans, lesions of SII impair texture and shape discrimination that require active haptic exploration. Another role of SII is to integrate tactile and motor actions. In humans, neural activity in SII can be modulated by concurrent somatosensory stimulation and finger movement. This modulation may reflect an improved analysis of tactile signals during movements.

3. Retroinsular Cortex and Somatosensory Insula

Ri and the posterior insula also have robust somatosensory properties. They are contiguous with SI and are located within the lateral Sylvian sulcus. Ri is adjacent to SII. In addition to SII and posterior SI, many neurons in Ri have bilateral receptive fields. Neurons in these areas respond primarily to contralateral tactile stimuli. Somatotopic representation is only rough. In primates, the majority of neurons in the insula (Ig) also have bilateral receptive fields. The insula may be involved in tactile learning and its role may be similar to the role of area TE in inferior temporal cortex for visual learning. In contrast to TE, which is purely visual, the insula is multimodal. In humans, lesions involving area 40 (putative area SII) and somatosensory insula can cause disorders of tactile object recognition.

4. Somatosensory Areas of the Posterior Parietal Cortex

In humans, SI is bounded dorsomedially by another somatosensory association cortex called supplementary sensory area (SSA) encompassing Brodmann's area 5 (inferior parietal cortex) and area 7b (supramarginal gyrus). The stimulation of supplementary motor area (SMA) during brain surgery sometimes elicits tactile sensations, suggesting that SMA may be a dorsomedial somatosensory association cortex. Posterior parietal cortex is believed to integrate sensory and motor processing and to combine tactile and proprioceptive information with other sensory

modalities. In monkeys, damage to area 5 affects proprioceptive inputs and impairs the nonvisual guidance of arm movements. In humans, damage to the right superior parietal lobe can produce similar impairments. Area 7b may have a role in higher order integration within the somatosensory system. The majority of neurons in area 7b have bilateral receptive fields. Furthermore, the area has multiple intra- and interhemispheric connections. They are activated by nonpainful tactile stimuli and, to a lesser extent, by visual and painful stimuli.

5. Summary

The multiple somatosensory cortical areas form a complex network. Somatosensory receiving areas are functionally involved with cortical areas subserving motor and spatial functions. The existence of such a complex system in the human brain is important for intentional, spatially guided motor movements that allow us to interact with tactile stimuli. The organization of the somatosensory system emphasizes the fact that perception is an active process.

II. PRIMARY AND COMPLEX TACTILE SENSATIONS

What parts of the neural system are involved in processing tactile information and our perceptions of touch? Primary tactile sensations refer to percepts arising from stimuli that vary along a single dimension. Often, primary somatosensory sensations can be elicited through passive stimulation. Complex tactile sensations arise from stimuli that vary along multiple dimensions. Typically, complex tactile sensations are enhanced by active exploration of the stimuli.

A. Primary Somatosensory Functions

1. Tactile Acuity and Two-Point Discrimination

The sensitivity of tactile perception depends on the relationship between mechanoreceptors and somatosensory cortex. Tactile acuity or sensitivity is typically defined as the ability to distinguish between two points of stimulation. Different parts of the body have greater sensitivity to tactile stimulation than others. For example, the fingertips and the lips are more sensitive than the back. Tactile acuity can be correlated with the properties of the afferent nerve fibers and of neurons in SI. At the afferent nerve fiber level, tactile acuity corresponds to parts of the body that have the greatest density of nerve fibers with both small and large receptive fields. However, better two-point discrimination is found for nerve fibers with small receptive fields (SA1 and RA1 channels): Two points of stimulation are more likely to fall in two separate receptive fields and be distinguished from each other. Thus, there are more small receptive field nerve fibers innervating the sensitive fingertips than the less sensitive palm. A similar relationship exists at the cortical level. Larger regions of the SI are devoted to sensitive parts of the body. The cortical area of the hand is relatively greater than that of the back, presumably to provide additional neural processing necessary for the perception of fine detail.

2. Light Touch

Light touch refers to the sensation of near-threshold tactile stimulation. Nerve fibers carrying light touch information are part of both the medial lemnescal system and the spinothalamic system. The exact contribution of different cortical areas to these sensory functions is not established, but 3- to 5-Hz pressure stimulation on various parts of the body produces contralateral activation in SI and bilateral activation in SII.

3. Vibration

Vibration activates most of the mechanoreceptors. Furthermore, it is easily manipulated experimentally and produces robust changes in cortical activity. Typically, neuroimaging experiments compare simple vibratory stimulation with no stimulation (i.e., a resting state). Stimulating either the hand or the foot with vibration produces bilateral activation in SII, the parietal operculum, and the insula. However, activation in SI is greater in the contralateral hemisphere. Vibration also activates contralateral motor cortex, supplementary motor area, and anterior parietal areas.

4. Summary

Primary somatosensory functions are largely processed by SI and, to some extent, by SII. Neuropsychological data support this conclusion. In patients who underwent surgery to relieve epileptic symptoms, the most severe disorders of tactile sensitivity, two-point discrimination, point localization, and position sense were produced by lesions in the contralateral, postcentral gyrus. The most severe defects occurred in patients whose lesions encroached the hand area of SI.

Furthermore, unilateral cortical lesions of SI produced clear bilateral sensory defects.

B. Complex Somatosensory Functions

Primary tactile perception varies along a single attribute and requires primarily passive touch. In contrast, most tactile experiences involve three-dimensional objects that vary on multiple attributes. The perception of complex objects is enhanced by active touch, or haptics, to help apprehend characteristic features that distinguish one object from another. When we explore objects, we use stereotypical hand movements called exploratory procedures. People use exploratory procedures to obtain particular types of information from objects. Lateral motion or rubbing is used to extract texture, pressure is used to extract hardness, and contour following is used for detailed shape. Enclosure, or a grasp, is often used to extract global shape and size. In addition, exploratory procedures are often executed simultaneously to extract multiple object features. Thus, the achievement of tactile object recognition results from the integration of information from cutaneous, motor, and cognitive systems.

1. Texture

The hand, with its mechanoreceptors and motoric abilities, is specialized to process information about an object's material substance. This information includes texture (roughness), hardness, and temperature. Sometimes, it is referred to as microgeometric information because it does not change the global shape of an object. Compared to vision, touch is superior for the discrimination of surface texture. When multiple object properties within an object are equated for perceptual discriminability, people prefer to categorize objects by touch on the basis of texture rather than shape. Furthermore, blind and sighted people are equally good at discriminating texture.

The perception of surface texture includes attributes such as roughness, hardnesssoftness, elasticity, and viscosity. The most widely examined texture dimension is roughness. To investigate roughness, studies use surfaces with regularly spaced ridges (gratings) and surfaces with raised dots to systematically and independently manipulate the relevant physical features. The spacing between the ridges (groove width) and the amount of force applied to the stimulus correlate highly with perceived roughness. The speed at which stimuli are moved across the skin, actively or passively, has little effect if the forces are equal. These results suggest that mechanoreceptors provide the signals for roughness perception independently of kinesthetic input.

However, roughness perception is not a simple function of any particular mechanoreceptor and its afferent nerve's response rate. Instead, roughness perception is related to spatial variation in responses between nerve fibers. Specifically, it is the between-fiber spatial variation in firing responses of the SA fibers that corresponds with perceived roughness. For example, in the fingertip, between-fiber spatial variation is the difference in the discharge rate over a distance of 1 or 2 mm.

At a cortical level, roughness perception can be associated with groove width and, to some extent, with force and velocity. In monkeys, strong responses in areas 1 and 3b of SI are produced from the active stroking of the fingertips over gratings with constant ridge and varying groove. In these areas, neurons respond differentially for rough stimuli, smooth stimuli, and combinations of the roughness, the force applied, and the velocity of stroking. In SII, neurons respond to the beginning, middle, and end of the strokes. Passive and active touching of the gratings activates different proportions of neurons. In humans, neuroimaging studies of passive roughness perception demonstrate corresponding SI and SII activation.

2. Complex Pattern and Shape

Shape and complex patterns can be considered part of an object's macrogeometry. In contrast to texture, the hand is less well suited for the extraction of these global properties. The tactile perception of pattern and shape involves the integration of input from multiple receptors in the skin and sometimes the integration of contour information over space and time. It appears that somatosensory cortex has specialized regions for processing different types of tactile information.

The identification of complex forms, such as Braille and alphabet letters, requires the recruitment of information from a number of afferent nerve fibers. Pattern perception appears to rely on the pattern of SA1 nerve fiber responses over a region of skin. The pattern of SA1 responses, often illustrated using spatial event plots, is similar to the stimulus pattern, indicating that they are important for the tactile perception of detail. Confusion in the perception of two patterns correlates with the similarity between their spatial event plots. For example, C's are confused with O's.

Pattern and shape information is processed by different areas of cortex from texture and hardness information. In monkeys, the discrimination of hardness, texture, and shape can be differentially impaired depending on which part of SI is ablated. Lesions of area 2 impair shape and contour discriminations but spare texture and hardness discriminations. The opposite pattern is found for lesions of area 1. Excisions of area 3b produce impairments in all aspects of tactile discrimination learning.

Neuroimaging studies in humans provide converging evidence that different areas of the somatosensory cortices are involved in the tactile perception of shape and size from texture and hardness. Contralateral activation of SI is produced by the discrimination of texture, shape, and hardness compared to a grasp. Although both shape and texture activate contralateral inferior temporal regions, texture differentially activates the parietal operculum. Shape and length discrimination activates the contralateral supramarginal gyrus, contralateral premotor cortex, and bilateral angular gyri. Thus, specific parts of the somatosensory cortices beyond SI are responsible for the perception of particular tactile properties. Global or macrogeometric features (shape and size) require additional integration of somatosensory information over space and time as well as more extensive somatosensory processing.

3. Object Recognition

The recognition of everyday, common objects through touch is fast and accurate, often requiring less than 2 sec. Haptic object recognition plays a frequent role in our lives. We use it every time we reach into our pockets for keys or coins. Real objects vary along multiple object dimensions and typically have characteristic features to help identify them. To determine the cortical areas involved in tactile object recognition, I turn to studies of patients with selective brain lesions and neuroimaging studies of normal individuals.

a. Astereognosia Tactile object recognition can be defective if primary sensory perception is impaired. Astereognosis is defined as the general inability to recognize objects by touch in the absence of vision. Although damage to higher level somatosensory regions can produce tactile object recognition deficits, astereognosis refers most commonly to tactile object recognition deficits resulting from severe primary somatosensory imperception. It reflects damage to any level of the somatosensory system from the peripheral nerves, spinal cord, brain stem, and thalamus to SI. Patients with astereognosis typically have difficulty perceiving light touch, vibratory sensation, proprioception, superficial pain, temperature, two-point discrimination, weight discrimination, texture, substance, double simultaneous stimulation, and shape. The impairment is usually restricted to one hand. SI is considered to be the cortical substrate of astereognosis.

Furthermore, when trying to identify objects by touch, patients with astereognosia do not tactually explore the object. This lack of appropriate exploration suggests that primary tactile functions are not adequately perceived or integrated with motor information during tactile object processing. Although sensorimotor integration and proprioception are important for tactile object recognition, it should be noted that tactile exploration is not necessary. Paralysis by itself does not produce defective tactile object recognition. In addition, objects can often be identified by touch if they are passively moved over a person's hand.

b. Tactile Agnosia Impaired tactile object recognition can arise from deficits in higher order tactile processing. In contrast to astereognosia, tactile agnosia is the inability to recognize objects by touch despite adequate primary somatosensory functions, intellectual ability, attentional capacity, and linguistic skill. Although patients with tactile agnosia cannot recognize objects by touch, they can often draw an accurate picture of an unrecognized tactually perceived object or match it to another object. The parietotemporal cortices, possibly involving SII, are considered to be cortical substrates for tactile agnosia. One patient with unilateral tactile agnosia had a left inferior parietal lesion involving the SII complex. She was able to discriminate and categorize objects on the basis of stimulus properties, but she was unable to recognize objects with the contralateral hand. It appeared that she was impaired in her ability to integrate information about the various object attributes to create a representation of the object as a whole.

Evidence from neuroimaging studies with non-brain-damaged individuals has also implicated SII as being important for tactile object recognition. When activation from exploration is subtracted, SII is activated bilaterally, with stronger activation on the contralateral side. Superior parietal and medial temporal cortices are also involved. Together, these cortical regions suggest that information about tactile features, spatial properties, and object identity is being integrated during tactile object recognition.

Tactile agnosia may also be produced from bilateral lesions in the subcortical part of the angular gyrus. The subcortical lesions presumably disconnect somatosensory association cortex from semantic memory stores located in the inferior temporal lobe.

III. CORTICAL ORGANIZATION OF TACTILE PERCEPTION

A. Plasticity and Reorganization of Somatosensory Cortex

Tactile perception can be influenced by changes occurring in somatosensory cortex as a result of trauma or experience. Several factors can influence the relative proportion of somatosensory cortex devoted to a particular body part or parts: the lack of sensory input due to the loss of a limb, the amount of stimulation or practice on a particular region of skin, and the active nature of the tactile task.

1. Changes in Sensory Input and Phantom Limbs

The somatosensory cortical maps can change when peripheral input changes or when signals from a specific part of the body no longer reach cortex. Signals can be prevented from reaching cortex due to the amputation of a body part or damage to the peripheral nerves conducting the signal from the skin. In monkeys, cutting nerves from the fingers eliminates responses in the corresponding SI region, but over time, stimulation of adjacent fingers may activate cortical cells in the lost digit's cortical area. The area of cortex devoted to that body part reduces over time.

The presence of phantom limbs provides good evidence for the plasticity of the human somatosensory cortex and how experience can change the functional organization of the brain. In humans, amputation often produces a phenomenon called the phantom limb; amputees experience sensations as if the limb existed. The plasticity of the synaptic connectivity of neurons in SI is one explanation for the cause of phantom limbs. For example, phantom sensation is related to changes in SI. Some studies support a remapping of somatosensory cortex in which the portion of SI devoted to the face (an active area) encroaches or is remapped into the adjacent area devoted to the absent limb (an inactive area). Some amputees can relieve an itch in their phantom limb by scratching the body region that is located in the homunculus next to the cortical area of the amputated limb. Neuroimaging experiments of patients with phantom upper limbs confirm an enlargement of the primary sensory field of the face.

2. Effects of Training

Somatotopic maps can be altered by increasing the stimulation to a particular body part location. Stimulating a specific region of skin can produce an expansion of the cortical area receiving its input. Furthermore, practice in the use of a particular fingertip can increase the cortical area representing that fingertip.

Evidence that training can effectively modify cortical representations comes from neuroimaging studies comparing string players and nonmusicians. Musicians had stronger cortical responses than nonmusicians when the fingers of their left hands were stimulated. Furthermore, the effect was strongest in musicians who began their careers early in life. It can also be demonstrated that the somatotopic cortical representations of the finger areas were altered in blind Braille readers who used three fingers of both hands to read. Instead of the typical homucular pattern observed in the neural responses of finger stimulation in sighted subjects, Braille readers had distorted organization in one or both hands. This disorganization had perceptual correlates. The same Braille readers also had difficulty identifying which finger was touched.

3. Activity of Body Part

The active nature of perception appears to influence cortical representation. Before surgery, patients with webbed hands had decreased cortical representations for the fingers and a reduced hand area in SI. After surgery, when the fingers functioned independently, the cortical hand area expanded and a greater distance was found between the representations of the thumb and little finger. Thus, somatosensory cortical maps are not static.

B. Dual Streams of Processing for Object Recognition and Localization in Somatosensory Cortices

When we recognize objects, not only do we have to identify the objects but also we must localize them in space. In visual cortical areas, two separate processing streams have been identified. The dorsal stream, leading from visual cortex to the parietal lobes,

conveys information about an object's spatial location ("where") and how to interact with it. The ventral stream, leading from visual cortex to the temporal lobes, conveys information about object identity ("what"). A similar dual stream of processing is proposed for the somatosensory system. Physiologic studies and behavioral studies suggest a functional distinction between ventrolateral ("what") and dorsomedial ("where") somatosensory association cortex (Fig. 1). Ventrolateral somatosensory association cortex includes SII, the parietal operculum, and the posterior insula. Dorsomedial somatosensory association cortex involves SSA and SMA.

1. The "What" Pathway

The ventrolateral stream processes somatosensory information regarding object recognition, tactual learning, and memory. Lesions of this pathway result in tactile agnosia, independent of other tactile or spatial abnormalities. Specific evidence for a distinction between somatosensory "what" and "where" systems in humans comes from a patient with unilateral tactile agnosia. Her selective impairment of tactile object recognition arose from a unilateral lesion involving the left inferior parietal lobe, thought to be part of the ventral pathway. She was unable to recognize objects with her right hand despite normal exploration procedures. In contrast, she was completely normal in her ability to perform visual and tactile spatial tasks. Thus, she was selectively impaired in "what" but not "where."

2. The "Where" Pathway

The dorsomedial stream is concerned with sensorimotor integration and tactile spatiotemporal functions. In particular, an apraxia–astereognosis syndrome has been identified in patients with damage to dorsomedial somatosensory association areas. Patients with extensive damage to SMA and SSA have moderate to severe impairment of primary and complex tactile functions. They have severe limb apraxia with extremely disordered tactile search strategies. In addition to the faulty spatial and temporal control of movement, they have an analogous spatiotemporal defect of tactile perception: They have difficulty localizing a stimulus within a limb or determining whether they have been touched once or twice. Other patients with damage to the superior parietal lobe have profound impairments in shape discrimination and spatial orientation without tactile sensory impairments. From these patients, it is suggested that there may be a nontactual, supramodal factor in tactile impairments. In summary, patient data suggest that the somatosensory "where" system can be differentially impaired.

C. Tactile Learning and Memory

Once tactile information is perceived, what parts of the brain are involved with its storage and retrieval? The hands feel object surfaces to sample and discriminate particular object properties. This information is sent to the brain so that it can be retrieved later. Most of our knowledge about tactile memory comes from experimental lesion studies of learning in the monkey, but recent functional neuroimaging studies have provided supporting evidence in humans. Tactile learning and recognition appear to be mediated by a circuit going from SI to SII, the insular granular cortex, the amygdala, hippocampus, and perirhinal cortex. This pathway is organized in a similar fashion as the visual learning pathway that accesses the limbic structures of the temporal lobe through inferior temporal cortex.

The anatomical organization of tactile learning in monkeys and in humans appears to be similar. Tactile learning and recognition activate areas involved in motor function (premotor areas, SMA, and the cerebellum) and in tactile perception (SI, SII, SSA, and superior parietal lobule). In addition, the tasks activate limbic and paralimbic structures (hippocampus, amygdala, insular cortex, orbitofrontal cortex, and cingulate gyrus), prefrontal areas, and the striatum. The major difference between tactile learning and recognition tasks is the relatively higher activation of the striatum and cerebellum during learning.

D. Selective Attention in Touch

Selective attention has been shown to influence tactile perception. It not only affects the ability to discriminate among stimuli but also appears to modulate the properties of somatosensory cortical neurons by either enhancing or suppressing their activity. In monkeys, the firing of neurons in SI and SII is dependent on whether the monkey is paying attention to a stimulus. Attention can increase the neuron's response to a stimulus. Relatively larger responses are found in somatosensory neurons when the animal is attending to the tactile stimulus than when it is not. Similarly,

neuroimaging studies in humans demonstrate that attention to vibratory stimulation can enhance signal processing in SI.

Attention can also change the signal-to-noise ratio in neuronal response. Attention may be a mechanism for reducing distractions through the suppression of signals from noncued channels. This produces a selective increase of signal magnitude in an attentionally cued channel. In monkeys, attending to one of several simultaneous vibrotactile stimuli suppressed firing rates in populations of neurons in SII and area 7b prior to the presentation of the target stimulus and enhanced activity during and after the target. A similar finding can be demonstrated in humans anticipating a stimulus. Behaviorally, individuals respond more quickly to a site when a vibratory stimulus is expected. Neurophysiologically, anticipation of either a focal touch or painful shock produces decreases in activation for regions of SI located outside of the representation for that skin area of the expected stimulus. Although the mechanisms underlying increases and decreases of activation of somatosensory cortical areas are not fully understood, it is clear that attention modulates the activity of the neurons.

E. Hemispheric Specialization

As noted previously, tactile perception is largely lateralized. Do both hemispheres equally process tactile information? To demonstrate hemispheric function and interhemispheric communication, two types of patients have been studied: patients with unilateral hemispheric lesions and "split-brain" patients who have undergone surgery dividing the cerebral commissures. Most interhemispheric transfer occurs via the corpus callosum (a large collection of nerve fibers that carry information from one side to the other).

In intact humans, it remains unclear whether hemispheric specialization exists for primary somatosensory functions (e.g., pressure sensitivity, vibration sensitivity, two-point discrimination, or point localization). However, unilateral SI lesions frequently result in severe and long-lasting defects in the contralateral hand. There is growing evidence from brain-damaged patients that the two brain hemispheres are specialized for certain higher tactile functions requiring the spatial exploration of objects or fine temporal analysis. The left hand–right hemisphere combination is better able to perform tasks with a spatial component, such as reading Braille or tactually determining the orientation of rods. The right hand–left hemisphere combination can perform these tasks if familiar stimuli are used, only a few objects are presented, or when linguistic processing is possible.

Given the lateralization of the tactile system, what information perceived by one hand is available to other if the two hemispheres cannot share information? The lack of interhemispheric transfer of tactile information (touch, pressure, and proprioception) can be demonstrated in the performance of split-brain patients. In a tactile task in which a set of objects are first felt using one hand and then retrieved from a larger set of objects, split-brain patients can accurately retrieve the objects with the same hand but not the other hand. Furthermore, these patients can name and describe objects explored tactually by the right hand but not by the left hand. In a tactile task in which the patient is touched on the fingertip of one hand and has to touch that fingertip with his or her thumb, the split-brain patient can perform the task with the same hand's thumb but not with the other hand's thumb. Last, in a tactile task in which the split-brain patient's hand is pressed into a posture, the patient can mimic the posture with the same hand but not with the other hand. In summary, tactile perception and proprioception are largely lateralized.

IV. CONCLUSIONS

The understanding of tactile perception requires an evaluation of tactile object recognition as well as an evaluation of basic tactile responses. Most tactile object recognition tasks contain sensory, spatial, proprioceptive, constructive, and motor components. As a result, tactile perception involves functional interconnections between SI, SII, premotor regions, and more posterior portions of the parietal cortex. There is still much to be learned about the neural bases of tactile perception. In particular, we need to know more about the roles of somatosensory cortices for the processing of complex tactile stimuli, such as real objects. Recent evidence suggests that the functional organization of the neural pathways is similar to that of vision.

See Also the Following Articles

BODY PERCEPTION DISORDERS • HAND MOVEMENTS • MOTOR SKILL • MULTISENSORY INTEGRATION • OBJECT PERCEPTION • SENSORY DEPRIVATION • TASTE

Suggested Reading

Burton, H., and Sinclair, R. (1996). Somatosensory cortex and tactile perceptions. In *Pain and Touch* (L. Kruger, Ed.), pp. 105–179. Academic Press, New York.

Caselli, R. J. (1997). Tactile agnosia and disorders of tactile perception. In *Behavioral Neurology and Neuropsychology* (T. E. Feinberg and M. J. Farah, Eds.), pp. 277–288. McGraw-Hill, New York.

Craig, J. C., and Rollman, G. B. (1999). Somesthesis. *Annu. Rev. Psychol.* **50,** 305–331.

Greenspan, J. D., and Bolanowski, S. J. (1996). The psychophysics of tactile perception and its peripheral physiological basis. In *Pain and Touch* (L. Kruger, Ed.), pp. 25–104. Academic Press, New York.

Heller, M. A., and Schiff, W. (Eds.) (1991). *The Psychology of Touch*. Erlbaum, Hillsdale, NJ.

Johnson, K. O., and Hsiao, S. S. (1992). Neural mechanisms of tactual form and texture perception. *Annu. Rev. Neurosci.* **15,** 227–250.

Klatzky, R. L., and Lederman, S. J. (1993). Toward a computational model of constraint-driven exploration and haptic object identification. *Perception* **22,** 597–621.

Loomis, J. M., and Lederman, S. J. (1986). Tactual perception. In *Handbook of Perception and Human Performance: Cognitive Processes and Performance* (K. R. Boff, L. Kaufman, and J. P. Thomas, Eds.), pp. 1–41. Wiley, New York.

Ramachandran, V.S., Rogers-Ramachandran, D. C., and Stewart, M. (1992). Perceptual correlates of massive cortical reorganization. *Science* **258,** 1159–1160.

Roland, P. E. (1993). *Brain Activation*. Wiley, New York.

Taste

MICHAEL A. BARRY* and MARION E. FRANK
**Seton Hall University and University of Connecticut Health Center*

GLOSSARY

branchiomeric nerves Nerves that innervate tissues derived from branchial (gill) arches.

best response Refers to the stimulus that elicits the highest spiking rate from a neuron and is used in defining classes of taste neurons, such as sucrose-best.

chemotopy Spatial separation of taste receptors within the oral cavity or taste neurons within an area of the central nervous system, according to chemosensory sensitivities.

common chemical sensation Elicited by activation of somatosensory pain pathways with a chemical stimulus.

dysgeusia A distorted, frequently unpleasant, taste perception, whereas aguesia is a loss of taste sensation.

flavor The combination of taste, smell, and other sensory aspects of food.

gastrointestinal malaise A feeling of illness that is attributed to a disorder of the gastrointestinal tract. It is the unconditioned stimulus that is associated with a conditional stimulus, the taste or smell of a food item, in learning a flavor aversion.

hedonic value Either positive or negative. A pleasurable experience has positive hedonic value; a repulsive experience has negative hedonic value.

iatrogenic Conditions caused by medical treatment, such as the mucositis associated with chemotherapy.

laterality of projections Refers to the side of the brain, left or right hemisphere, that is involved in the neural pathways. For example, touch is represented primarily on the opposite, contralateral hemisphere relative to the side of the body surface stimulated. The taste pathway is primarily ipsilateral, on the same side as the side of the tongue stimulated.

modality A sensation associated with one sensory system, such as vision, hearing, or taste. Multimodal responses are responses to stimuli that activate different sensory systems.

Old World primates Primates that evolved from the same progenitors as humans, not those living in the New World (North and South America).

prototypal taste stimuli Stimuli that represent taste qualities in relatively pure form, such as sucrose for sweet or quinine for bitter.

The sense of taste evaluates nutritional or harmful chemicals taken into the mouth and contributes to the pleasure of eating. Taste senses sources of nutrients, such as sugars, salts, and amino acids, as well as artificial sweeteners, which have been developed to satisfy the ''sweet tooth'' without causing health problems that come with overindulgence. Taste also senses poisonous and injurious substances. The brain associates an agreeable hedonic value to nutrients and a repulsive hedonic value to poisons. Nutrient sources are sought after and reflexively ingested. Poisons are reflexively expelled and their sources avoided. Internal physiological states and past experience influence the pleasure associated with taste sensations. The taste system, which is also known as the gustatory system, includes the taste receptor cells located in taste buds in the oral and pharyngeal cavities, the cranial nerves that innervate these receptors, and central nuclei and cortical areas that process the information transmitted by these nerves. The term ''taste'' here refers only to sensations mediated by this system. It should not be confused with flavor, the total experience of a food stimulus, which involves the senses of temperature, texture, common chemical, smell, and taste.

I. STIMULUS QUALITIES

Humans experience a limited set of taste qualities that each vary in perceptual intensity. The four elementary qualities are salty, sweet, sour, and bitter. *Umami* (savory), the meaty taste of monosodium glutamate (MSG), is considered to be a fifth taste quality. Complex tastes of various chemical stimuli can be thought of as combinations of the basic qualities. For example, potassium chloride is salty and bitter. Prototypical gustatory stimuli sodium chloride (salty), sucrose (sweet), citric acid (sour), quinine–HCl (bitter), and MSG (savory) each primarily evoke a single taste quality. However, particularly at just detectable or very high concentrations, the quality evoked by even these stimuli is perceived as a combination of qualities. At high concentrations, electrolyte stimuli activate the common chemical sense resulting in further perceptual complexity. A variety of chemicals can elicit the same taste quality. Chemicals having a number of hydrophilic molecular structures stimulate the sweet taste quality and include sugars, certain amino acids, and artificial sweeteners. Salty taste is elicited by sodium, lithium, potassium, and ammonium salts, although the latter two chemicals are also bitter. Acids have a sour taste. Chemicals with diverse chemical structures, such as salts, amino acids, and a variety of lipophilic compounds including alkaloids, thioureas, and bile acids, elicit the bitter taste.

II. PERIPHERAL TASTE SYSTEM

A. Taste Buds

Taste buds contain receptor cells and supporting cells that are both derived from epithelial cells. Taste receptor cells (TRCs) have an average life span of about 10 days and are replaced from a population of stem cells located within the taste bud. TRCs are depolarized by chemical stimuli that either enter the cells via channels (e.g., Na^+) or are bound by receptors located in the apical TRC membrane (e.g. sucrose, quinine, and glutamate). Peripheral axonal processes of neurons located in cranial nerve ganglia innervate the taste buds. The central axonal processes of the ganglionic neurons project to the brain. These neurons are mostly afferent; that is, they carry information to the brain. There is no purely efferent innervation of the taste buds. Efferent neurons modify peripheral receptors based on central commands. However, somatosensory afferent fibers and possibly gustatory fibers can release chemicals peripherally in response to stimulation and thus modify the response of TRCs. An example of a chemical released by somatosensory fibers is the peptide substance P. Taste papillae also receive efferent innervation from autonomic fibers, which could also influence gustatory responses. Finally, processing of gustatory information within taste buds may also be mediated by interactions between TRCs.

B. Taste Bud Fields and Innervation

The taste buds are located in a number of fields in the oral cavity. Taste buds in the several fields differ in distribution, morphology, sensitivity, and innervation. On the anterior two-thirds of the tongue, taste buds are located in fungiform papilla that are most dense at the anterior tip and are innervated by the chorda tympani branch of cranial nerve VII (cN-VII), the facial nerve. The greater superficial petrosal branch of cN-VII innervates palatal taste buds, found on either side of the border between hard and soft palates. The lingual and palatal branches of the trigeminal nerve (cN-V) provide somatosensory innervation to these regions and mediate the oral common chemical sense. Much of the epidermal trigeminal innervation of these areas is concentrated in or around the taste buds.

The lingual branch of the glossopharyngeal nerve (cN-IX) provides the gustatory innervation of taste buds that line the trenches of the foliate and circumvallate papilla located on the posterior one-third of the tongue. The superior laryngeal branch of the vagus nerve (cN-X) innervates taste buds located on the larynx. Nearby taste buds scattered on the surface of the epiglottis, pharynx, and uvula are probably also innervated by the cN-X, but the tonsilar branch of cN-IX may also contribute to the innervation of these as well as some palatal taste buds. Neuronal fibers for pain, temperature, and touch on the posterior third of the tongue and the pharynx and larynx are contained within cN-IX and cN-X.

C. Functional Differences among Taste Bud Fields

Prototypical taste stimuli can all be detected in any of the taste fields on the tongue or palate. Within the fungiform taste field there is little evidence for chemotopy, a mapping of chemical sensitivities based on spatial position. However, for NaCl and sucrose, perceptual intensity ratings are higher and detection

thresholds are lower for the anterior tongue and palate compared to other taste fields. For citric acid and particularly quinine, intensity ratings are higher and thresholds are lower for the posterior tongue. Stimulus sensitivities of the laryngeal and pharyngeal taste receptors are not well-known in humans or primates. In other mammals, they appear to play a role in airway protection and other reflexes. Many of the receptors are sensitive to water.

D. Afferent Nerve Physiology

Electrophysiological recordings have been made from single nerve fibers in the chorda tympani and glossopharyngeal nerves in a variety of species, including some Old World primates. The gustatory afferent neurons have low spontaneous rates and respond to taste stimulation with an increase in spiking (action potential) rate. The neurons are slowly adapting; the initial phasic response is followed by a tonic response that slowly decays over the course of stimulation. The response latency is longer for stimuli that bind to apical receptors, such as sucrose, than for ionic stimuli such as NaCl. Stimulus intensity is encoded by rate of neuronal response because spike rates are directly correlated with stimulus concentration. The firing pattern is not regular, and in response to sweet chemicals it often consists of irregular bursts of spikes. It is not known if this temporal information is utilized by the nervous system. Rather, the primary mechanism of taste quality coding is thought to be present in the differential response of neurons. Some classes of afferent taste neurons are specific in that they respond only to chemical stimuli that elicit one taste quality (Fig. 1). S fibers, a class of neurons in the chorda tympani nerve of the chimpanzee, respond to all compounds known to be sweet to humans, and they show little response to other stimuli. N fibers, which show a best response to NaCl, are not as specific as a group. However, there are subsets of N fibers specific for sodium or lithium salts, others responsive to NaCl, KCl, and other salts, and still others that respond to NaCl and Na-glutamate. A class of chorda tympani fiber with a best response to acids is found in macaques but not in chimpanzees. Quinine-best Q fibers are the least specific class of afferent taste neuron in the chimpanzee chorda tympani. In contrast, the glossopharyngeal nerve contains quinine-best neurons quite specific for stimuli that taste bitter to humans. In

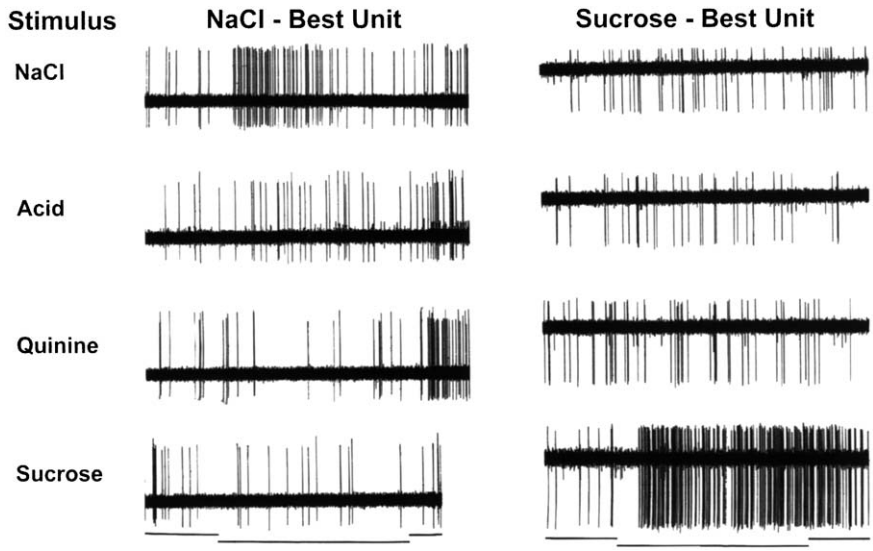


Figure 1 Response of two nerve fibers (units) in the chimpanzee chorda tympani nerve to the application of taste stimuli to the anterior tongue. Each unitary excursion from the baseline represents a single action potential. The stimuli were 70 m*M* NaCl, 5 m*M* quinine–HCl, and 0.3 *M*, sucrose for both units. For the (sour) stimulis, 50 m*M* aspartic acid was used for the unit on the left, and 40 mM citric acid was used for the unit on the right. The stimuli were applied to the anterior tongue for 8 sec as indicated by the displaced line at the base of the figure, followed by an artificial saliva rinse. The delay in the response is partially due to the time it took the stimulus to reach the taste buds on the tongue. Note that the NaCl-best unit appears to be inhibited by quinine. This is because, unlike the other stimuli, quinine was dissolved in water rather than the artificial saliva. Thus, the NaCl-best fiber was excited by the 10-m*M* sodium content of the saliva rinse following quinine (courtesy of Viktoria Danilova and Göran Hellekant).

addition, there is a group of glossopharyngeal fibers that responds best to MSG and glutamate and another group that responds best to stimuli that are sweet to humans. Although there are quantitative regional differences in taste sensitivities, there is significant overlap. Loss of any one receptive field does not seriously impair discrimination among prototypal taste stimuli in humans.

III. CENTRAL TASTE SYSTEM

The location of central processing of taste information in the brain reflects the distribution of taste bud fields in the peripheral taste system and their relationships to other sensory systems (Fig. 2). Taste buds are located in the oral and pharyngeal cavities. The oral cavity marks a transition between the external sensory world, which is largely conscious, and the internal sensory world of the pharyngeal cavity and more internal parts of the body, which is reflexive and largely unconscious. Thus, gustatory information is processed in regions of the brain that are transitional between areas that process exteroreceptive versus interoreceptive inputs. Peripherally, the taste buds are closely associated with receptors for pain, temperature, and touch. The processing of taste and these somatosensory modalities in adjacent or overlapping central nuclei occurs throughout the central nervous system (CNS). On the other hand, olfactory information is not integrated with taste and somatosensory information until they converge within association areas of the cerebral cortex, the amygdala, or hypothalamus. It is also in these areas that taste information is integrated with information on internal states, such as the state manifested as satiety.

A. Functional Anatomy of Taste Pathways

1. Nucleus of the Solitary Tract

The central axonal processes of primary afferent gustatory neurons enter the brain via three cranial nerves: cN-VII, cN-IX, and cN-X. All of these are mixed branchiomeric nerves, which contain gustatory and nongustatory sensory and motor components. The cell bodies of the sensory nerve fibers are located in cranial nerve ganglia as shown in Fig. 2. Gustatory afferent fibers terminate in the superior part of the ipsilateral nucleus of the solitary tract (NST) in the medulla. Nongustatory primary afferent fibers in the facial, glossopharyngeal, and vagus nerves terminate more caudally in the NST as well as in other nuclei, including the caudal spinal trigeminal nucleus and even the dorsal horn of the spinal cord. These nuclei are involved in somatosensory processing from areas including the oral cavity and may influence gustatory processing. The spinal trigeminal nucleus projects to the NST as well as to higher order gustatory nuclei including the thalamic gustatory nucleus.

Primary afferent gustatory fibers from the greater superficial petrosal and chorda tympani nerves enter the pons between the vestibulocochlear (cN-VIII) and motor cN-VII nerve roots via a separate rootlet of cN-VII, the intermediate nerve of Wisberg. The fibers course posteriorly and medially and descend over a short distance through the pons, trapezoid body, spinal trigeminal tract, and interpolar spinal trigeminal nucleus to reach the solitary tract. Fibers descend in the solitary tract or ascend for a few millimeters to terminate in the NST superior to the tract. Glossopharyngeal primary afferent fibers enter the medulla in the cN-IX cranial nerve root and course posteriorly and medially through olivocerebellar fibers, the descending trigeminal tract, and interpolar nucleus to reach the solitary tract, where they may ascend or descend. More inferiorly, vagal afferent fibers course posteriorly and medially through the inferior cerebellar peduncle and trigeminal tract and nucleus to reach the solitary tract. Within the NST, gustatory fibers terminate largely in the interstitial nucleus, which wraps around the solitary tract and forms most of the NST superior to the tract. The interstitial nucleus extends from the level of the accessory motor trigeminal nucleus to just inferior to the obex. It is characterized by dense acetylcholinesterase staining, which is associated with the extracellular surface of primary afferent terminal fibers. Acetylcholinesterase is an enzyme that hydrolyzes the neurotransmitter acetylcholine, but its function in the NST is unknown. The sequence of projections of gustatory afferent axons into the solitary tract is from superior to inferior: facial, glossopharyngeal, and vagal. However, there is substantial overlap and all three nerves have projections into the superior pole of NST. The oral and interpolar sections of the descending trigeminal nucleus lie immediately anterolateral to the solitary tract (Fig. 2, inset). Trigeminal axons penetrate the NST at midsuperior levels where glossopharyngeal axons terminate most heavily. The trigeminal terminal fields overlap with more lateral glossopharyngeal and vagal axonal terminal areas that contain neurons responsive to somatosensory stimuli. The clearly separate medial areas contain taste-responsive

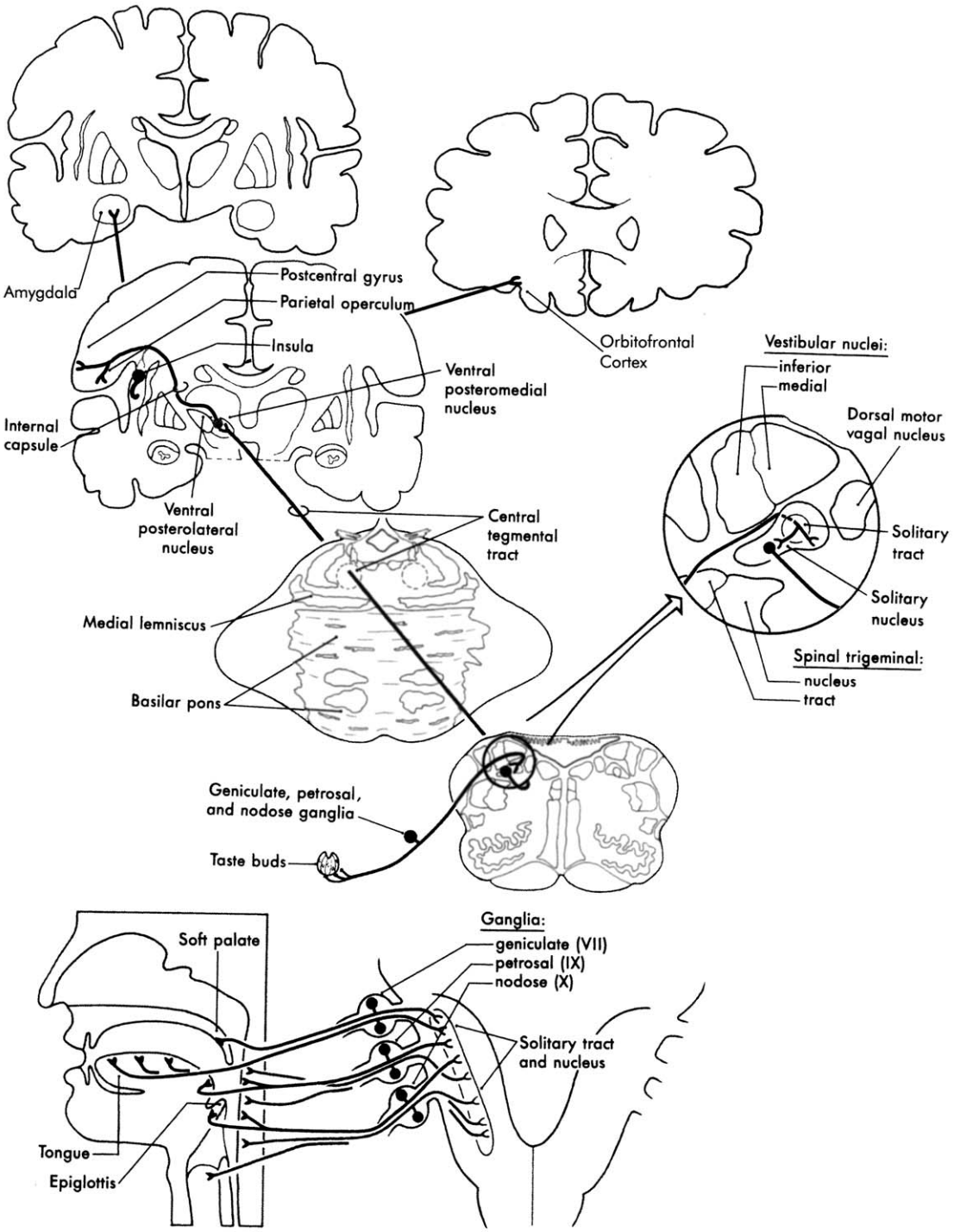


Figure 2 Summary of peripheral and central taste pathways. Gustatory, afferent fibers terminate in the superior part of the nucleus of the solitary tract, whereas general visceral afferent fibers terminate in the inferior portion of the nucleus. Gustatory neurons in the nucleus of the solitary tract project to the parvicellular part of the ventral posteromedial nucleus in the thalamus. Thalamic gustatory neurons project to the insular cortex and parietal operculum. Gustatory neurons in the insular cortex project to the amygdala and orbitofrontal cortex [modified with permission from Sweazy, R. D. (1997), Olfaction and taste. In *Fundamental Neuroscience* (D. E. Haines, Ed.), pp. 321–333. Academic Press, San Diego].

neurons. In primates, responses to stimulation of the tongue with taste stimuli are found primarily in the superior part of the afferent projection zone within the interstitial nucleus.

Primary afferent axonal endings form synapses on dendrites of NST neurons in the gustatory zone and release the excitatory neurotransmitter glutamate. Many small intrinsic NST neurons use the inhibitory neurotransmitter γ-aminobutyric acid (GABA) to mediate inhibitory interactions among primary afferent and other inputs. Other neurotransmitters that influence NST gustatory neurons include substance P and possibly enkephalins. Inputs to the gustatory zone include projections from trigeminal nuclei, other parts of NST, the reticular formation, pons, amygdala, and insular cerebral cortex. NST neurons project to sites such as oromotor centers in the reticular formation, implying involvement in oral reflexes. The brain stem, including NST and pontine parabrachial nucleus, without involvement of higher neural centers, can generate behavioral motor responses, such as sucking and rejection, and exocrine responses, such as salivation, in response to taste stimulation.

Ascending projections from the gustatory zone of NST course in the ipsilateral central tegmental bundle and terminate in the ipsilateral small-celled [parvicellular (pc)] portion of the ventral posteromedial nucleus (VPMpc) of the thalamus. In rodents and lagomorphs, NST thalamopedal projections terminate in the parabrachial nucleus of dorsal pons. This nucleus surrounds the brachium conjunctivum, also known as the superior cerebellar peduncle. In Old World primates, and probably humans, there is apparently no projection from the NST gustatory zone for the anterior lingual taste buds to the pons. There may be pontine projections from NST gustatory zones for posterior tongue or pharynx. The pontine parabrachial nucleus projects to the thalamus and to brain regions that process visceral information, such as the hypothalamus and amygdala. Destruction of the thalamic gustatory relay has little effect on taste preferences for taste stimuli by primates. This observation might be explained by indirect pathways that may bypass the VPMpc, such as those that utilize the pontine parabrachial nucleus.

2. The Parvicellular Ventral Posteriomedial Thalamic Nucleus

The gustatory pathway to the thalamus is largely ipsilateral; VPMpc thalamic neurons respond only to stimulation of the tongue on the same side in most species. However, the situation in humans is not as clear. In the few available clinical cases, ipsilateral gustatory losses usually follow pontine or midbrain lesions, but there are some contradictory examples and some experimental studies suggest that projections from the NST to the thalamus are bilateral. VPMpc lies immediately medial to the VPM and is contiguous with the facial–oral somatosensory representation in VPM. Neurons in VPMpc are smaller than in the VPM. VPMpc neurons respond to taste stimuli, but unlike the taste part of the NST, neurons that respond to somatosensory stimuli or somatosensory and taste stimuli are also present. General viscerosensory information may also be processed in VPMpc. Thus, unlike somatosensory, visual, and auditory thalamic nuclei, VPMpc, the thalamic nucleus that processes taste information, also processes information from other modalities.

3. Primary Gustatory Cortex

By definition, a primary sensory cortex is the cortical area that receives most direct projections from the specialized sensory relay nucleus in the thalamus. Thus, the primary gustatory cortex (PGC) receives direct projections from the thalamic gustatory relay, VPMpc. In Old World monkeys, the most extensive VPMpc projections are to the dorsal (superior) part of the anterior insular cortex and adjacent frontal–parietal lobes covering the insular cortex, the opercula. The projection zone is centered on granular insular cortex but extends ventrally (inferiorly) into the dysgranular insular and dorsally and laterally into dysgranular opercular cortex. Granular cells are small neurons that are most prominent in cortical layer IV in granular cortex, but in dysgranular cortex the granule cell layer is thinner and contains fewer cells. The VPMpc projection area is characterized by increased cytochrome oxidase staining, but it is not as dense as it is in regions receiving primary thalamocortical inputs for other sensory modalities. Cytochrome oxidase is a mitochondrial enzyme. Increased levels of this enzyme are indicative of increased capacity for oxidative metabolism. A second projection area from the VPMpc is more lateral, at the convexity of the parietal operculum in somatosensory cortical area 3b, just anterior to the area where the tongue is represented.

Cortical neurons that respond to gustatory stimulation are found in both of these areas and in adjacent opercular and insular cortical regions not receiving direct projections from VPMpc. These areas, such as the inferior agranular insula, are secondary or higher

order gustatory areas. Less than 10% of neurons in PGC respond to taste stimuli; many others respond to other intraoral stimuli, but most are unresponsive to these stimuli. Thus, the PGC and adjacent higher order areas sensitive to gustatory stimuli cover a large region of cortex, but gustatory neurons contribute to but do not dominate the processing taking place in these regions. As in VPMpc, the thalamic relay, taste-responsive neurons in the primary gustatory projection zone in the cerebral cortex are intermingled among nontaste neurons, unlike in other sensory modalities.

Much of what is known about the primary gustatory cortex is based on studies of nonhuman Old World primates. Although the insular and adjacent opercular cortex is similarly organized in humans, the exact location of PGC is unknown. PGC and adjacent taste-responsive areas occupy large regions of cortex and precise tracing of VPMpc projections has not been possible in humans. Clinical cases with lesions have placed the PGC in the insular cortex, anterior temporal lobe, or parietal operculum (i.e., the inferior part of the post-central gyrus). The parietal operculum is adjacent to the somatosensory representation of the tongue, which is shifted posteriorly in humans compared to Old World monkeys.

Modern imaging techniques localize human cortical gustatory areas in the general region of the insular cortex and adjacent parietal or frontal operculi (Fig. 3). Specifics differ depending on recording technique and stimulus paradigm.

Magnetoencephalography identifies the posterior insula and adjacent parietal operculum as the first cortical site of activation in humans, which is consistent with a posterior shift of the homunculus compared to monkeys. Functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) also detect insular activation anteriorly in positions similar to the PGC of Old World monkeys.

Neurons in the VPMpc project to the ipsilateral PGC, but the laterality of gustatory processing in the PGC is not established. As is the case for the brain stem, the results of most studies of gustatory loss following insular lesions in humans suggest that taste information is initially processed ipsilaterally to the stimulus. However, the initial wave of the electroencephalogram evoked by a unilateral taste stimulus occurs in both hemispheres of the cortex simultaneously, suggesting projections to PGC are bilateral. Many PET and fMRI studies localize activation (or a stronger activation) to the right superior and anterior insular cortex following whole tongue stimulation.

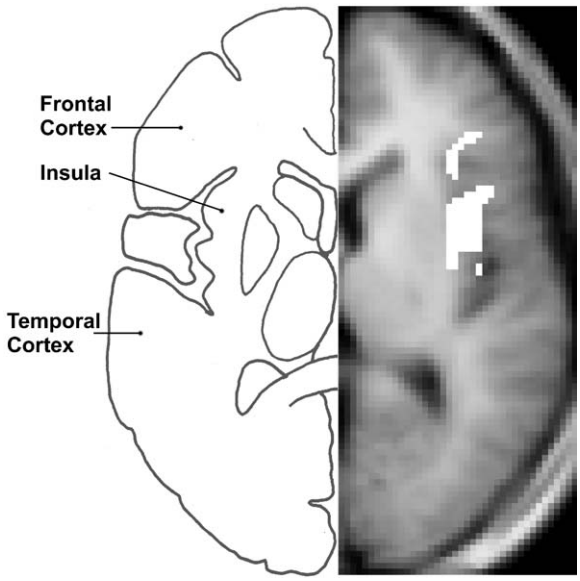


Figure 3 Response of the human cerebral cortex to electric-taste stimuli applied to the anterior tongue. The right hemisphere of a horizontal section is shown. Areas with a significantly ($p<0.01$) elevated hemodynamic response are shown superimposed on an anatomical image. The activation and anatomical images represent the average of 11 subjects. The activated regions at this level are in the insular cortex and adjacent frontal cortex. Electric taste is a phenomenon in which small currents applied across the tongue mucosa activate taste receptor cells and produce a metallic/sour taste sensation.

Thus, cortical taste processing occurred predominantly in the nondominant hemisphere of the right-handed subjects studied, but more work is needed to clarify lateralization of taste function.

The many taste-responsive neurons in the insular and opercular cortices that do not receive direct projections from the VPMpc by definition comprise gustatory association areas, which also lie outside this periinsular region, mostly in the limbic system. Neurons in the PCG project to the amygdala, entorhinal cortex, perirhinal cortex, superior temporal sulcus, secondary somatosensory area, and frontal cortex. The connections are reciprocal because neurons in these regions project back to the PGC. Gustatory association areas in the amygdala, a precentral frontal opercular area, and orbitofrontal cortex have been electrophysiologically verified. PET imaging has revealed taste-evoked activation in the orbitofrontal cortex bilaterally and right anterior medial temporal cortex, including the amygdala. Epileptic patients who are treated with right temporal lobectomy have difficulty with taste quality recognition but have normal taste thresholds.

In turn, orbitofrontal and amygdalar neurons project to other nuclei, including the striatum and lateral hypothalamus, in which responses to gustatory stimuli have been recorded. These secondary and tertiary association areas are multimodal limbic centers in which information about internal state, taste, and smell may be integrated. Activation of other limbic centers, such as the hippocampus or anterior cingulate cortex, by gustatory stimuli depends on the hedonic value of the stimulus.

B. Gustatory CNS Physiology

Neurons in the rostral (superior in humans) part of the NST in mammals respond to gustatory stimulation of the ipsilateral tongue and palate. Excitation and inhibition are seen, but due to low spontaneous rates inhibition is clearer in response to stimulus mixtures. NST neurons are less specific for stimuli of one taste quality than peripheral neurons but show quality tuning across stimulus classes. Additional taste information may be provided by differences in the locations of neurons processing gustatory stimuli of different perceptual qualities along the long superior–inferior NTS axis, an organizational plan known as chemotopy. There is a rough chemotopy imposed on the NTS by the rostral to caudal (superior to inferior in humans) projection pattern of the intermediate, glossopharyngeal, and vagal nerves, which have characteristic sensitivities to taste stimuli, as noted previously. In addition, NTS neurons that are most sensitive to sour acids are located more caudally than neurons tuned to stimuli with other taste qualities. This chemotopy is not seen in the gustatory thalamus and cortex, in which there are few neurons that respond best to acids. Unlike taste-responsive NTS neurons, many thalamic and cortical neurons respond to both gustatory and nongustatory stimuli. Neural response patterns across different taste stimuli at brain stem, thalamic, or cortical levels take on one of four or five general response patterns that show correspondence to the four or five perceptual taste qualities. Neuronal responses to mixtures of taste stimuli are more complex. Insular cortical neurons respond less to a mixture of several prototypal stimuli, including the best stimulus, than to the best stimulus alone. This result suggests that primary cortical neurons do not identify a specific mixture, such as lemonade, but are tuned to detect a single taste quality, such as sweet; their responses are inhibited if the mixture contains stimuli with other qualities.

Responses of neurons in gustatory association areas are modifiable by internal states. Taste information is used for making decisions on acceptance/rejection of food, the hedonic value of which depends on general internal state (i.e., hunger) and even specific nutritional deficits. In monkeys, satiety modifies responses of neurons in gustatory association areas, such as orbitofrontal cortex, amygdala, and lateral hypothalamus. For example, neurons in the amygdala respond to taste stimuli only if the monkey is hungry. In the association areas, taste information is combined with the olfactory, visual, and somatosensory information about foods. Thus, many of the neurons respond in effect to the flavor of foods, and it is the response to flavor that is influenced by satiety. In the orbitofrontal cortex, the responses of single neurons can show specific, satiety-related decreases in firing rates that occur only for the stimulus for which the monkey is satiated and not for other stimuli. In Fig. 4, the response of a single orbitofrontal neuron to the taste of glucose decreases as the monkey is fed to satiation with a glucose solution. However, the firing rate in response to black currant juice is unchanged. In other experiments in which monkeys were fed to satiation with black currant juice, neurons were found that showed a decreased response to the juice but not to glucose.

In addition to being multimodal, neurons in the orbitofrontal cortical gustatory association area often respond more strongly to complex taste mixtures (e.g., black currant juice) than to any single stimulus (e.g., the sweet sugar) contained therein. This is in contrast to PGC neurons that respond more to a best single taste stimulus than to a taste mixture containing that stimulus. The tuning of multimodal association neurons to a mixture may simply reflect an additive response to the taste, olfactory, and textural stimuli that activate the neurons. Most natural taste stimuli have smells and textures as well as tastes that contribute to flavor perception.

Neurons in the lateral hypothalamus are also excited by specific gustatory stimuli, which suggests that they may be involved in taste discrimination. However, there are also neurons in this nucleus that respond nonspecifically to taste stimuli and may influence other aspects of taste processing besides discrimination. The nonspecific neurons have high spontaneous spiking rates, which are reduced in response to taste stimuli. They are (i) spatially separated from the taste-excited neurons; (ii) sensitive to the internal milieu, as demonstrated by an increase in spiking rate when glucose is locally applied; and (iii) involved in sensorimotor integration.

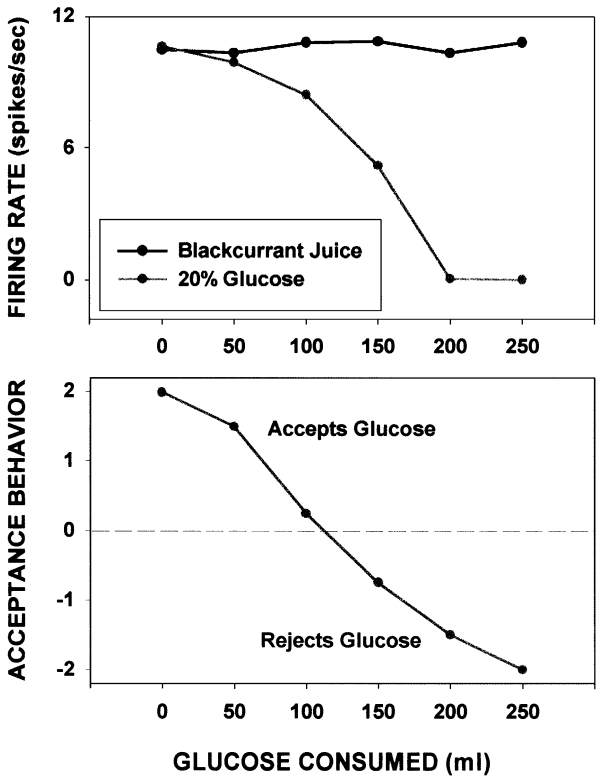


Figure 4 (Top) The effect of feeding to satiety with a 20% glucose solution on the response of a neuron in the orbitofrontal cortex to the taste of glucose and black current juice. (Bottom) The acceptance or rejection of the glucose solution on a scale from 2 to −2, respectively; a value of 0 indicates neither acceptance or rejection. The monkey was fed 50 ml of the glucose solution at each stage, as indicated along the abscissa, until he was satiated as shown by whether or not he rejected the solution [modified with permission from Rolls, E. T. (1997). Taste and olfactory processing in the brain and its relation to the control of eating. *Crit. Rev. Neurobiol.* **11,** 263–287].

IV. TASTE PSYCHOPHYSICS

A. Spatial Localization

Much of what we know about human tastes derives from psychophysical studies, which scientifically demonstrated fundamental taste qualities as noted previously. Precise localization of a taste stimulus was thought to require specific, associated somatosensory cues in humans. In contrast, goldfish have an exquisite ability to separate food pellets that have taste from inorganic particles, such as sand or pebbles, within the mouth. Although this ability may not be as developed in people, we can easily tell which side of the tongue a taste stimulus is on without specific touch cues. Thus, location-specific taste information is available to humans.

B. Integration of Information from Several Peripheral Fields

As noted previously, taste buds located on the anterior and posterior tongue have different sensitivities to prototypal taste stimuli in humans. Thus, whole-mouth perception of taste stimuli is the result of CNS integration of information from the various taste fields. For example, specificity of neurons mediating responses to salts and *umami* in the glossopharyngeal and chorda tympani nerves differs and thus provides different information about taste quality to the CNS from anterior and posterior lingual sites. However, because stimulation of either site with sweet stimuli elicits a similar response profile, in this case, stimulation of a second field provides added intensity and not new quality information. People rarely report deficits in whole-mouth taste following unilateral loss of a single peripheral taste field, which may be related to the poor localization of taste noted previously or central compensatory mechanisms. For example, if CNS neurons that are excited by posterior tongue stimulation are inhibited by anterior tongue stimulation, eliminating the inhibitory inputs would result in heightened responses to posterior tongue stimulation. For the bitter taste of quinine, there is evidence for such compensation. When the chorda tympani nerve was anesthetized, the perceived intensity of quinine applied to the posterior tongue increased.

C. Variability in Human Taste Abilities

In any human population, there is a unimodal normal distribution of taste detection thresholds or ratings of perceived suprathreshold taste intensity for many taste stimuli. This variation may be correlated with anatomical variation, such as the number of fungiform taste buds on the anterior tongue. For the bitter-tasting thioureas, such as propylthiouracil (PROP) or phenylthiocarbamide (PTC), there is a bimodal (or trimodal) distribution of sensitivities. Subjects are divided into tasters and nontasters of 1 m*M* PROP or PTC. Individual differences in sensitivity are hereditary and the ability to taste PROP is mediated by an autosomal-dominant allele.

V. PHYSIOLOGICAL AND DEVELOPMENTAL EFFECTS

Internal states and drives, such as hunger, modify preference for foods with specified taste qualities. For

example, after ingesting untasted NaCl tablets, humans on a low-sodium diet add less salt to unsalted tomato juice, demonstrating decreased salt preference. They also show an increase in NaCl recognition threshold, the concentration at which its taste is correctly identified. Thus, based on internal NaCl levels, physiological mechanisms, which likely involve the CNS, modify salt intake and salt recognition. As noted previously, lateral hypothalamic neurons respond to internal levels of glucose and oral taste stimuli. These neurons may form part of the pathway mediating effects of internal states on taste function. In contrast to physiologically adaptive mechanisms, increased habitual tasting and consuming of salted foods results in decreased salt intensity ratings and increased salt preference and consumption. Similar effects of habitually satisfying the sweet tooth lead to increased sugar consumption. Such eating habits establish a tolerance that exacerbates common health problems, such as hypertension and tooth decay.

Human newborns are very responsive to sweet stimuli. Their appetitive hedonic responses are readily observed in their sucking rates, preferences, facial expressions, and heart rates. Newborns' rejection of MSG, acid, and quinine is just as clear, demonstrating negative hedonic value on many of the same measures. However, sensitivity to the taste of some bitter compounds and NaCl may develop during the first 3 or 4 months after birth. Newborns hardly detect the taste of NaCl, but if detected it has negative hedonic value. However, sensitivity increases rapidly; by the time they are 4 months of age, NaCl is readily detected and preferred over water. The earlier development of sensitivity for sweet compared to salty stimuli may reflect differential development of taste bud receptors. Changes in preference during early postnatal development may reflect developing physiological controls involving the CNS.

VI. TASTE DISORDERS AND AGING

Dysgeusia is a distorted, frequently unpleasant taste perception. A "metallic" perception, perhaps the most common taste quality reported for dysgeusia, is a side effect of many medications, including antibiotics, antipsychotics, antiarthritics, and antihypertensives. Dysgeusia accompanies illnesses such as gastroesophageal reflux disease, oral yeast infections (candidosis), and burning mouth syndrome (BMS), a pain disorder primarily affecting postmenopausal women. Dysgeusia is also an iatrogenic result of treatment of head and neck cancer with radiation or chemotherapy, which also causes oral mucositis. Distorted unpleasant tastes may also result from treatment of oral infections with antiseptic mouth rinses that contain chlorhexidine. Taste-quality distortions may have peripheral or central origins. BMS dysgeusia, which is often eliminated by topical anesthesia of the tongue, may be due to a malfunction of the taste buds and peripheral nerves. Radiation-induced dysgeusia may encompass learned flavor aversions involving the CNS; food may be associated with the gastrointestinal malaise accompanying radiation therapy.

Taste losses without distortions also accompany illnesses or medical procedures and may indicate injuries to the taste nerves. The symptoms of BMS, for example, include loss of taste, which can be verified by clinical taste testing. High (1 *M*) concentrations of sweet sucrose and salty NaCl are rated weaker by BMS patients than by normal people. Also, extraction of the third molar teeth may result in weakened taste perceptions. The amount of taste loss is correlated with severity of tooth impaction and likelihood of nerve damage from the oral surgery. Losses are associated with damage to the chorda tympani nerve for lower third-molar extractions and with damage to the greater superficial petrosal nerve following upper third-molar extractions. Increases in taste detection thresholds and decreases in taste intensity ratings are also associated with aging. However, age-related losses in taste function are generally less severe than those for olfaction, audition, or vision.

Medications that many elderly people use result in taste losses, complicating the question of the effect of aging on taste. For example, antiarthritic, antihypertensive, and antineoplastic drugs have been implicated in taste loss. Nonetheless, aging has been shown to affect detection thresholds for taste stimuli, as well as ratings of perceptual intensity of taste stimuli that are well above threshold, in healthy adults. Average threshold increases are significant; for example, the NaCl detection threshold is about 2 m*M* NaCl at 25 years but about 4 m*M* NaCl at 65 years. It may be argued that the minimal concentration detected is not that relevant to "real-world" taste and thus the ratings for above-threshold stimuli are of greater importance. Again, there are significant decreases with age, but they are smallest for sweet sucrose and largest for bitter quinine. The intensity of 0.56 *M* sucrose is rated 10% weaker by those older than 70 years compared to those younger than 40 years. The intensity of 0.1 m*M* quinine–HCl is judged 40% weaker by those older than 70 years. The physiological mechanisms by which

aging affects taste perception are unknown. In rodents, aged individuals have about 10% fewer taste buds in their fungiform papillae on the anterior tongue than the young, but this trend has not been verified in humans. It is possible that taste buds remain but the function of this regenerating epithelial population is compromised with age.

See Also the Following Articles

AUDITORY PERCEPTION • CHEMICAL NEUROANATOMY • NEUROBEHAVIORAL TOXICOLOGY • OLFACTION • RECEPTIVE FIELD • TACTILE PERCEPTION • VISION: BRAIN MECHANISMS

Suggested Reading

Bernstein, R. L. (1999). Food aversion learning: A risk factor for nutritional problems in the elderly? *Physiol. Behav.* **66,** 199–201.

Besile, F. (1999). Glutamate and the UMAMI taste: Sensory, metabolic, nutritional and behavioral considerations. A review of the literature published in the last 10 years. *Neurosci. Biobehav. Rev.* **23**(3), 423–438.

Bradley, R. M., King, M. S., Wang, L., and Shu, X. (1996). Neurotransmitter and neuromodulator activity in the gustatory zone of the nucleus tractus solitarius. *Chem. Senses* **21,** 377–385.

Doty, R. L. (Ed.) (1995). *Handbook of Olfaction and Gustation*. Dekker, New York.

Getchell, T. V., Doty, R. L., Bartoshuk, L. M., and Snow, J. B. (Eds.) (199l). *Smell and Taste in Health and Disease*. Raven Press, New York.

Norgren, R. (1990). Gustatory system. In *The Human Nervous System* (G. Paxinos, Ed.), pp. 845–861. Academic Press, San Diego.

Rolls, E. T. (1997). Taste and olfactory processing in the brain and its relation to the control of eating. *Crit. Rev. Neurobiol*. **11,** 263–287.

Scott, T. R., and Plata-Salaman, C. R. (1999). Taste in the monkey cortex. *Physiol. Behav*. **67,** 489–511.

Small, D. M., Zald, D. H., Jones-Gotman, M., Zatorre, R. J., Pardo, J. V., Frey, S., and Petrides, M. (1999). Human cortical gustatory areas: A review of functional neuroimaging data. *NeuroReport* **10,** 7–14.

Temporal Lobes

HENRY A. BUCHTEL
Ann Arbor VA Health Care Center and University of Michigan

GLOSSARY

Klüver–Bucy syndrome Behavioral changes after bilateral damage to mesial temporal lobe tissues.

material-specific memory The lateralization of memory functions such that the learning of different kinds of materials depends on different degrees of left or right temporal lobe mechanisms.

temporal lobe syndrome A concept that patients with temporal lobe pathology have a particular way of behaving, usually involving a personality change with religiosity or sexual dysregulation.

The temporal lobes are important brain structures comprising about 20% of the total volume of the cerebrum. This part of the brain is important for complex visual and linguistic analyses and for the formation of new memories.

I. INTRODUCTION

The temporal lobes are readily recognizable brain structures with a thumb-like appearance when viewed from the side (Fig. 1). Their name simply reflects their location beneath the temporal bone on the side of the head. The temporal bone, in turn, receives its name from the fact that this is the place in which graying of hair starts, indicating aging with the passage of time (*L. tempus*). In some ways, the temporal lobes are more a convenient fiction than anatomical entities. They share borders with the occipital and parietal lobes, but the precise boundaries are not clearly defined by landmarks. A better definition of the anatomical limits of the temporal lobe would come from thalamic and intracortical projections and a functional analysis of the various subunits within the lobe. Because excision of the anterior temporal lobe is often used to help control medically intractable seizure disorders, much of our knowledge of the effects of damage to (or stimulation of) this area comes from studies of persons with epilepsy.

II. CYTOARCHITECTURE, PATHWAYS, AND CONNECTIONS

The posterior limit of the inferior and inferior-lateral temporal lobes is defined by the anterior border of area 19 of the occipital lobe. The Sylvian fissure (sometimes called the lateral cerebral fissure) defines the dorsal limit, whereas the posterior border at this dorsal level is defined by the beginning of parietal lobe Brodmann areas 39 and 40. Areas 22 and 37 (laterally) and 27 and 36 (mesially) belong within the temporal lobe, whereas areas 39 and 40 (inferiorally and laterally) and areas 23, 26, 29, and 30 (mesially) belong within the parietal lobe. There are three lateral gyri, delineated by the *superior, middle, and inferior sulci*, that parallel the Sylvian fissure. In the left hemisphere, Wernicke's area, which includes temporal lobe area 22 (as well as parietal areas 39 and 40), lies at the posterior end of the superior temporal gyrus where it meets the parietal lobe. Just in front of this region, bilaterally, is Heschel's gyrus (primary auditory cortex).

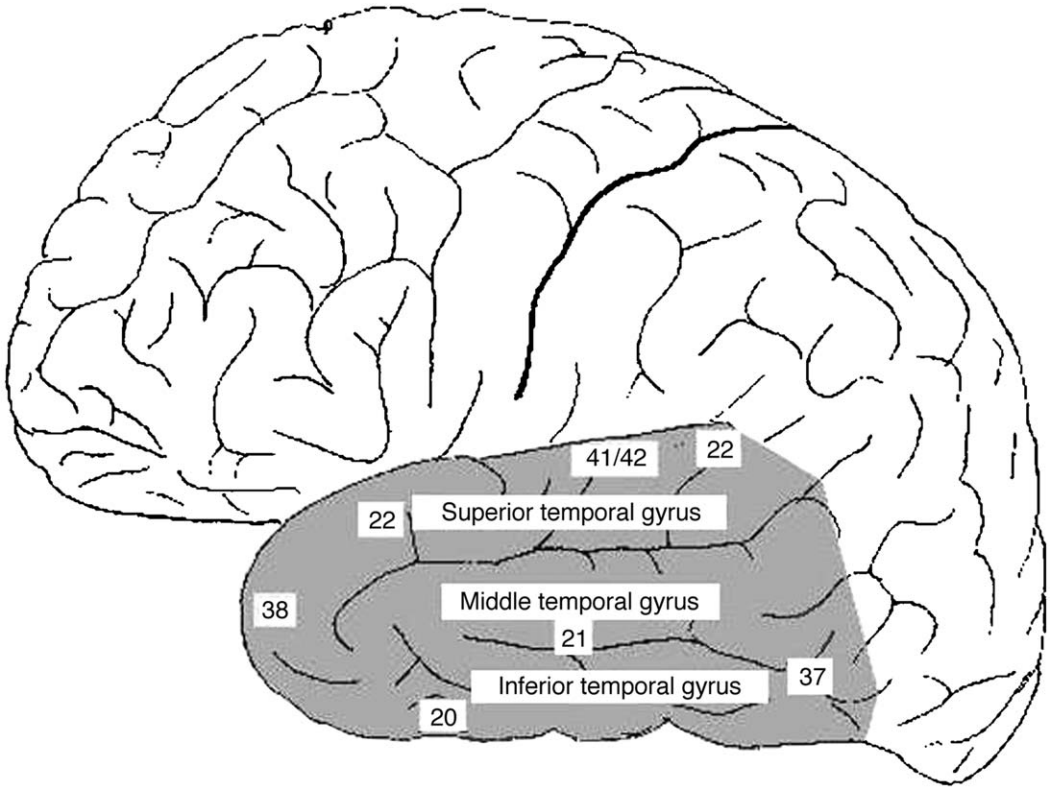


Figure 1 Side view of the brain showing temporal lobe (shaded) and Brodman areas.

III. HEMISPHERIC SPECIALIZATION

In humans, the functions of the temporal lobe can best be understood through the organizing principle of lateralization of function. It has been clearly established that the temporal lobes make important contributions to memory and perception, but the types of materials that are optimally handled by right and left temporal lobe mechanisms depend on the side being considered. In order to simplify the description of these functions, references to the right and left hemispheres in the following discussion are based on a typical representation, as seen in the majority of both right- and left-handers. In other words, for the purposes of this article, the left cerebral hemisphere will be assumed to be the one responsible for most speech functions and the right cerebral hemisphere to be the one responsible for most visual–spatial functions. There are clear exceptions among both right- and especially left-handed individuals, and even the designations visual–spatial functions and speech functions are not simple and unidimensionally lateralizable. This needs to be stated because for the temporal lobes, as well as for significant portions of the frontal and parietal lobes, many of the functions subserved need to be discussed in the context of the functional organization of the hemisphere within which the lobe is located. Evidence for this assertion comes from observations of patients with brain damage, from the effects of direct electrical stimulation during brain surgery, and from functional imaging studies [e.g., positron emission tomography (PET) and functional Magnetic resonance imaging (fMRI)]. In terms of gross anatomical structure, the striking differences between the functions of the right and left temporal lobes are not reflected in equally striking morphological differences. The Sylvian fissure rises more sharply on the left than on the right, but this pattern is by no means invariable and its probable underlying cause (the size of the planum temporalis) is only weakly correlated with the side of speech mechanisms (about 60% of right-handers have a larger planum temporalis on the left, but more than 97% of normal right-handers appear to have left hemisphere speech mechanisms). Interestingly, these morphological differences between the hemispheres appear to have predated the acquisition of speech abilities. Higher apes show the same pattern of morphological asymmetry and possess lateralized

differences in the motor system (preferring the right hand for manipulation tasks and the left for spatial localization), but it is clear that speech is not developed in these animals and "true" language abilities in nonhuman primates remain controversial.

IV. CLINICAL AND FUNCTIONAL ASPECTS

Functionally and neuroanatomically, the posterior parts of the inferior temporal lobe (especially the fusiform gyrus, area 37) can be considered to be extensions of the occipital lobe visual system. Microelectrode recordings from this region in animals in visual tasks and functional brain imaging findings with humans show that cells in these regions are actively involved in the analysis of visual images: In animals, cells have been found that are particularly sensitive to the presence of faces and other body parts.

Because of its position in the cranium, the temporal lobe is vulnerable to damage from closed-head injuries and increased intracranial pressure (sometimes induced during the process of vaginal delivery); its proximity (vascular and linear distance) to the face and to openings into the cranium from the nose leads to a major involvement in diseases such as herpes simplex virus. Most of what we know about the functional significance of temporal lobe mechanisms comes from the effects of lesions on behavior. Although much of this knowledge comes from animal studies, the higher cognitive skills subserved by the temporal lobes mean that it has been more rewarding to work with humans who, in addition to having psychometric deficits, can report their subjective impressions of the changes brought about by the damage. Cognitive changes occurring after damage to the temporal lobes are mainly seen in the domains of memory and perception. Again, the concept of hemispheric specialization is relevant because of the material-specific nature of some of the deficits.

A. Memory

Mesial structures of the temporal lobe (*amygdala, hippocampus*, and *rhinal cortex*) all have demonstrated roles in some aspects of the establishment of new memories. The relative roles of these portions of mesial temporal lobe are being worked out; however, interpretation is difficult because it is unusual for one region to be damaged without a disturbance of the functions of the others, and animal studies, in which precise lesions are possible, do not have exact analogs to the kinds of memory tasks given to humans.

B. Religiosity

There is a clinical impression that some patients with right hemispheric temporal lobe lesions undergo an increase in religiousness, sometimes to the extent that the term "hyperreligiosity" is applicable. The theoretical basis for such a change has resisted definition, but if the phenomenon does exist, its origin may lie in an attempt by the patient to explain the presence of unusual and "unworldly" feelings that arise from the effects of his or her lesions or seizure activity, especially if arising from the right hemisphere.

C. Emotion

The hemispheric organization of emotion has received renewed attention since the strengthening of the biological revolution in psychiatry, and the limbic system has traditionally been accorded a role in the experience and expression of emotion. The amygdala, in particular, has been seen as contributing to normal and abnormal emotional responses and experiences (leading, in extreme and misguided cases, to psychosurgical excision of these structures in individuals who are seen to have unacceptable emotionally driven behavior). Bilateral amygdaloid destruction causes a severe disturbance of normal affective behavior (Klüver–Bucy syndrome); damage in humans is usually unilateral and often incomplete, but even unilateral amygdaloid damage has led to changes in emotional experience. Recently, the amygdala has been implicated in the manifestations of schizophrenia and bipolar disorder.

D. Visual Perception

The contribution of the temporal lobe to visual perception has been clarified by the division of visual projections from the occipital lobe in two directions or "streams": a dorsal route into parietal lobes that is involved in spatial localization and a ventral route into the temporal lobe that is involved in the appreciation of the qualities of objects (and their identification) in the visual scene (Fig. 2). Early and controversial evidence from single-cell recordings that neurons in the inferior temporal lobe of monkeys respond to complex visual stimuli has thus been confirmed and extended.

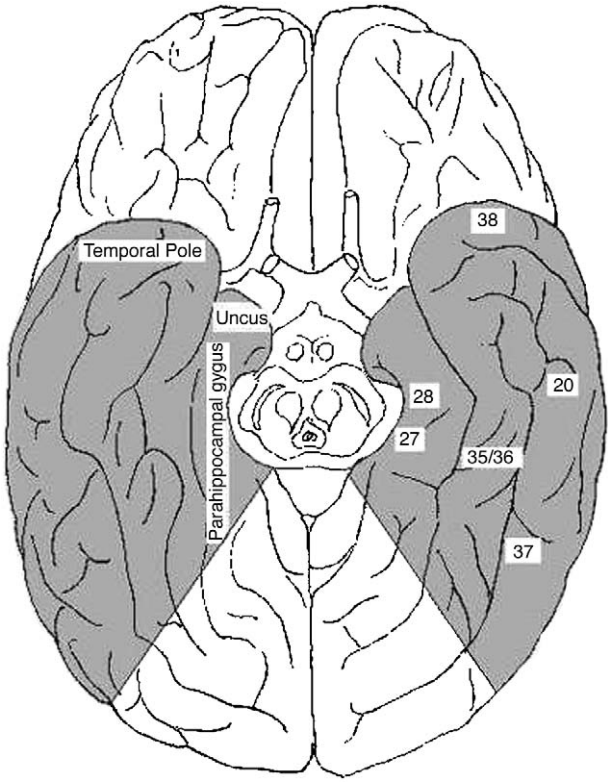


Figure 2 Base of the brain showing temporal lobe (shaded) and Brodman areas.

E. Face and Object Recognition

Damage to posterior inferior temporal (IT) lobe, especially when bilateral and when portions of inferior areas 18 and 19 are involved, can lead to a condition called *prosopagnosia*. Patients with this disorder can see a face but do not recognize the person from the face alone. This affects the recognition of friends, relatives, and famous persons alike. Interestingly, as soon as the person speaks, his or her identity may be immediately recognized. Prosopagnosics can usually recognize and correctly interpret facial expressions, and they may also be able to recognize a person from other visual but nonfacial cues, such as gait or posture. It is clear that the disorder caused by IT lesions has to do with conscious recognition: Recordings of galvanic skin conductance (which reflect emotional reactions) show that emotional responses occur with presentations of photographs of relatives and friends but not of photographs of strangers, despite the prosopagnosic's claim that all faces are equally unknown. Persons with prosopagnosia are usually also unable to learn new faces, at least when tested using conscious recognition. The disorder may be accompanied by difficulty recognizing (naming) famous buildings, and such individuals may also have difficulty with texture discriminations and color perception. In addition, attempts to name objects may result in the use of a general category rather than the object's unique name (e.g. "bird" rather than the more specific "robin" or "peacock"), with some kinds of objects appearing to be affected more than others (e.g., the naming of tools may be relatively good compared to the naming of natural things such as vegetables or animals). Prosopagnosia has also been seen with unilateral IT lesions, but the nature of the disorder appears to be different, depending on the side of the damage. When the damage is on the left, the person may identify the face incorrectly but still give an answer in the correct category (e.g., naming a different figure in politics or entertainment). This is called "deep" prosopagnosia, in analogy with "deep" dyslexia in which the person reads a word semantically related to a word on the page. When the damage is on the right, the person may be slow and erratic at recognizing faces but not profoundly prosopagnosic. The conjunction of prosopagnosia and other recognition deficits has been taken as evidence that faces and other complex visual stimuli share neurological substrates. However, it has been found that the regions responsible for these different categories of perceptual recognition are close neuroanatomically and thus it is difficult to damage one while leaving the other intact.

F. Language

The *superior temporal sulcus* lies above the *superior temporal gyrus* and electrical stimulation of this gyrus (especially in the posterior half) in the left hemisphere of typical right-handers often causes speech arrest or other alterations of language abilities. PET findings show bilateral activation of Heschel's gyrus and unilateral activation of Wernicke's area (left hemisphere only) when typical right-handed subjects listen to a verbal material such as a narrative.

The *middle temporal gyrus* lies between the middle and superior temporal sulci. Speech-related areas are occasionally found here, although it is possible that such unusual patterns of speech representation might be found predominantly in patients with preexisting brain damage (e.g., the patients described previously who were being evaluated for seizure surgery).

Damage to the anterior parts of the inferior and lateral portions of the temporal lobe of the speech

Table I
Effects of Temporal Lobe Damage

Region	Damaged hemisphere		
	Left	Right	Both
Lateral (inferior and/or inferior lateral)	Mild deficits in naming object categories; "deep" prosopagnosia	Mild deficits in naming object categories; brief prosopagnosia	Severe deficits in naming object categories; severe persistent prosopagnosia
Anterior	Deficits in finding proper names; anomia or restricted naming	Difficulty naming facial expressions	Anomia; retrograde memory deficits
Medial	Poor memory for new verbal information	Poor memory for new nonverbal information	Severe memory deficit for all new information

hemisphere causes problems with confrontational naming (e.g., naming pictures of common objects), but repetition and grammar are usually unaffected. An inferior speech area has been seen in some patients (speech arrest following electrical stimulation) and surgeons have usually assumed that these areas can be removed without significant loss of language ability. The existence and persistence of naming deficits after temporal lobectomy despite the sparing of the posterior speech zone raise the possibility that some patients need this inferior language area. Some patients have isolated naming deficits (e.g., proper names) with unilateral lesions of the anterior and lateral tip of the temporal lobe. Damage to the anterior parts of the inferior and lateral portions of the temporal lobe of the right hemisphere has been associated with memory for autobiographical episodes, although other retrograde memory functions may remain intact (Table I).

V. FUNCTIONAL IMAGING AND FUTURE DIRECTIONS

Most of the studies detailing the perceptual functions of the temporal lobes have been carried out in primates using microelectrode recordings and lesions, but new studies using functional activation with PET or fMRI have confirmed the hypothesis that visual analysis extends into the inferior portions of the temporal lobes. The activation appears to be bilateral in most cases, but there is a predominance of left-sided activation if the object's name is being sought. In animal studies, posterior lesions of IT lobe tissue disturb visual discrimination, whereas more anterior lesions disturb visual recognition. In this context, it is worth mentioning that one of the disadvantages of functional brain scans is that patterns of activation reveal what is activated, not what must be activated for successful completion of the task. For example, if a subject finds a task difficult, portions of the brain associated with frustration and discontent may be activated even though they are not a part of the neural substrate for the task. The design of control tasks is even more crucial than usual in such tasks.

See Also the Following Articles

ANOMIA • AUDITORY PERCEPTION • BRAIN ANATOMY AND NETWORKS • EPILEPSY • HEARING • OBJECT PERCEPTION • OCCIPITAL LOBE • PROSOPAGNOSIA • TIME PASSAGE, NEURAL SUBSTRATES

Thalamus and Thalamic Damage

CHIHIRO OHYE
Hidaka Hospital

GLOSSARY

cytoarchitecture Morphology based on a cellular component.

gamma knife A medical device used to treat deep-seated brain lesions by gamma-ray from 201 cobalt sources focused on a pinpoint.

lamellar distribution Distribution of cell dendrites directed in an anteroposterior direction to form a parallel thin layer.

microrecording Recording of neuronal activity by a microelectrode.

myeloarchitecture Morphology based on a myelin component.

positron emission tomography scan Computerized imaging of the living brain by detecting energy emitted by positrons from injected or inhalated short-lived radioactive substances.

stereotactic surgery Neurosurgical technique used to locate a given deep cerebral point using three-dimensional coordinates.

The thalamus is a mass of neurons that receives external and internal information and sends it to the cerebral cortex or basal ganglia after selection or modulation. Thus, the inner environment of human life is kept in harmony. If the thalamus is damaged, sensory, motor, or mental disorders may result depending on the site of thalamic damage.

I. INTRODUCTION

The thalamus is a mass of various types of neurons located in approximately the center of the brain underneath the cerebral cortex. It is covered completely by the cortical hemisphere, and therefore it is not visible from the surface of the brain. Its shape is ellipsoid, like a rugby ball, separated by the third ventricle in the middle except for small connecting fibers of massa intermedia. Thus, in relation to the third ventricle, it is symmetric, with a half ellipsoid on each side (Fig. 1). In a normal adult, its longer axis is about 30 mm, the maximum width of one side is about 20 mm, and the height about 20 mm. It contains approximately 10^7 neurons, and anatomists classified them into about 50–80 nuclei. However, this kind of classification is more or less artificial, depending on the diifferent methods and criteria used. Therefore, the same area could be divided or joined to make a new percellation. In fact, recent progress due to new staining techniques has shed new light on this field, delineating new boundaries previously not obvious by other methods.

Because the fundamental function of the thalamus is to relay and modulate peripheral information to the cerebral cortex and to the basal ganglia, keeping the somatosensory, mental, and emotional activity of a living individual in harmony, the ideal parcellation probably depends on the input–output relation from periphery via thalamus to higher centers. However, in the human thalamus, it is difficult to apply many modern methods, and most of the thalamic nuclei are still poorly understood. Thus, this article describes neurological findings correlated with restricted thalamic damage, stereotactic surgery with microrecording (even though it is biased to the ventral tier nuclei), and some relevant data in nonhuman primates.

The term thalamus, which was used by Galen approximately 1700 years ago, originated from the

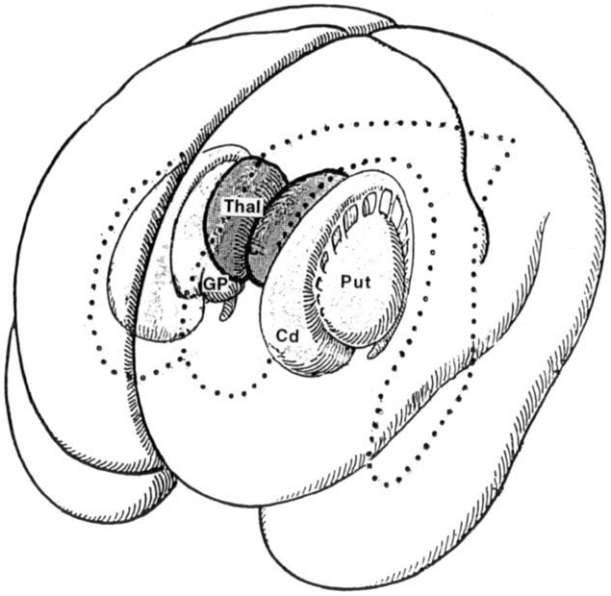


Figure 1 Transparent overview (anterolateral view) of the thalamus (Thal) (hatched area) in the brain. Cd, caudate nucleus; GP, globus pallidus; Put, putamen; dotted line, lateral ventricle. (modified with permission from Duus, 1989). Copyright George Thieme Verlag.

Latin or Greek word that means "inner chamber." The corresponding French word, *couch optique*, may be related to the fact that the optic tract can be traced by gross anatomy to the thalamus; in fact, the optic tract is the only visible input pathway from the periphery to the thalamus that can be recognized from the base. The term optic thalamus has also been used as a synonym for thalamus in old English literature. It is interesting to note that in Japanese terminology, thalamus is called *shisho*, which means "optic bed."

II. ANATOMY

A. Development

The thalamus is derived from the rostral part of neural tube together with the cerebral hemisphere. In the course of the early embryonic developmental stage, the first bending of the rostral neural tube toward the ventral side (flexura cephalica) separates the prosencephalon and mesencephalon. In the second bending, the flexura cervicalis further separates the caudal part of the neural tube into the rhomencephalon and spinal cord. From the rostral neural tube (prosencephalon), the telencephalon (future cerebral hemisphere) and diencephalon (future thalamus) develop. At an early stage of 6 mm crown–rump length (CRL), when the optic vesicle becomes visible from the outside, a swelling or thickening of the wall of the neural tube at the level of the diencephalon occurs (origin of the thalamus), together with a swelling of the neural tube canal (origin of the third ventricle) (Fig. 2A). In this stage, the diencephalon consists of four main parts: the epithalamus, dorsal thalamus, ventral thalamus, and hypothalamus (from dorsal to ventral). Each part is separated by the sulcus diencephalica dorsalis, medius, and ventralis, respectively. At the stage of 18 mm CRL (about 6 weeks gestational age), rapid growth of the dorsal thalamus with migration of the neuroblast of the dorsal thalamus surpasses that of the ventral thalamus, thus changing the relative portion and size (Figs. 2B–2D). The dorsal thalamus becomes larger, and the ventral thalamus is displaced and compressed to the lateral and ventral part. Further cellular differentiation at 25 mm CRL (about 7 weeks) induces nuclear primordis in each dienephalic part. For example, the pineal body, habenula, and posterior commissure originate from the epithalamus; the lateral part of the lateral geniculate body originates from the dorsal thalamus; the nucleus reticularis thalami, zona incerta, nucleus reuniens ventralis, and so on originate from the ventral thlamus; and the subthalamic nucleus and probably globus pallidus originate from the hypothalamus. As anticipated, the dorsal thalamus finally becomes the so-called thalamus, the main part.

Most thalamic cells migrate from the wall of the third ventricle at an early stage untill about 10 weeks gestation. Then, the path of migration is more active from the ganglionic eminence, which is a huge mass of highly proliferating cells located on the floor of the lateral ventricle–ditelencephalic border. This supplies cells to the caudate, putamen, and amygdala, and from 16 to 37 weeks of gestation the telencephalic part of the ganglion eminence (corpus gangliothalamicum) becomes a source of cells in pulvinar, which develops late in the gestation stage. It is interesting to note that the cells in the largest nucleus of the thalamus, the pulvinar, are telencephalic in origin.

Until approximately the 50-mm CRL stage, cellular differentiation is relatively poor in the dorsal thalamus. One can distinguish only lateral and medial geniculate body, parafascicular nucleus, and pulvinar. At the 100-mm CRL stage (about 15 weeks), various nuclei and fiber tracts develop. Also, the internal medullary lamina separates the dorsal thalamus into three the dorsal, ventral, and lateral parts. Thus, from the dorsal part originates centrum medianum and nucleus medialis, from the ventral part embryonic

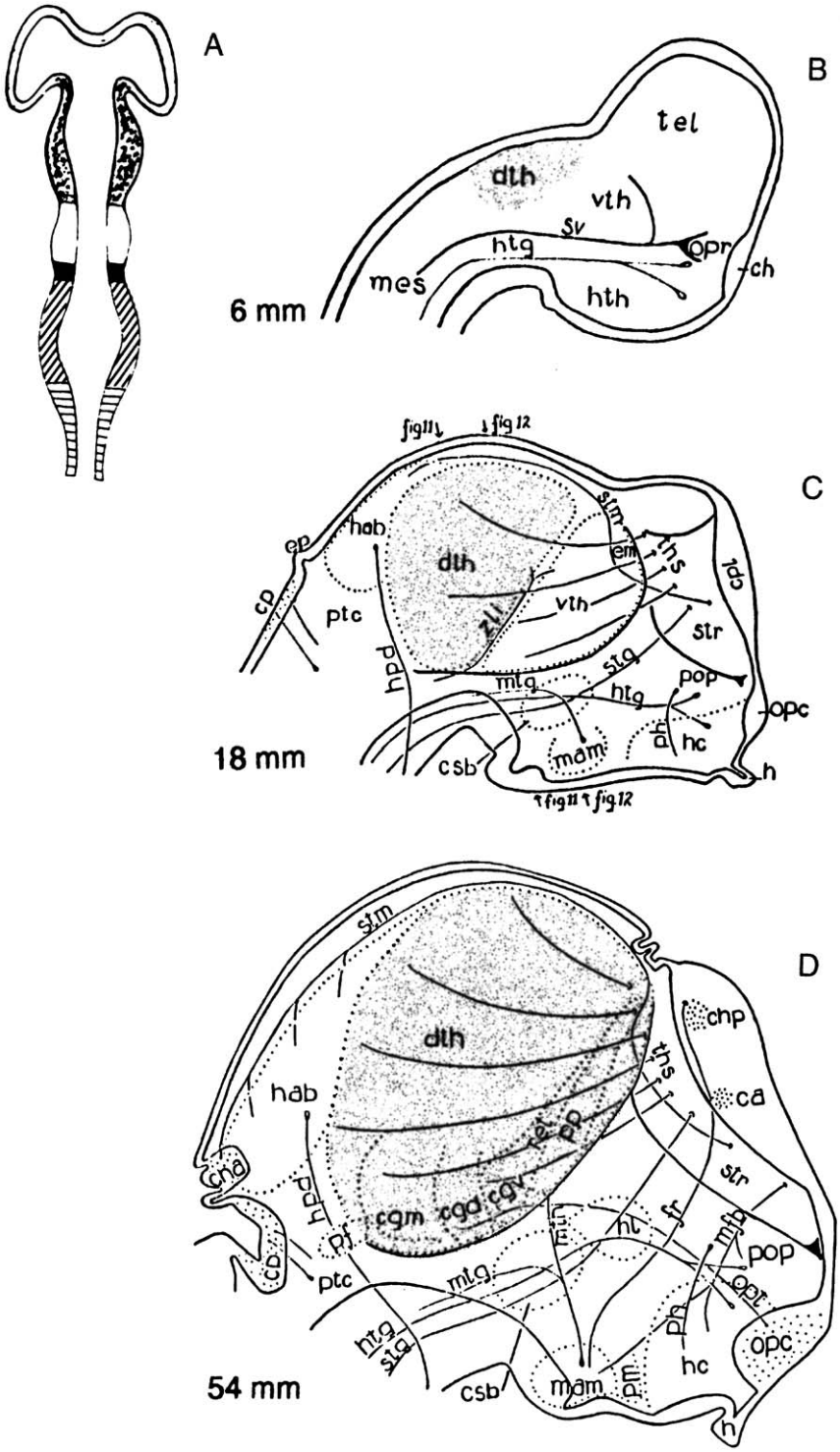


Figure 2 Schematic illustration of embryonic development of the human thalamus. (A) Axial section of the neural tube showing the origin of the main element of the central nervous system. From top to bottom, telencephalon (bilateral swellings), diencephalon (dotted area), mesencephalon (white zone), isthmus of rhombencephalon (black area), metencephalon (hatched area), and myelencephalon (horizontal bars). (B–D) Sagittal views of diencephalon at 6 mm (B), 18 mm (C), and 54 mm crown–rump length (D). Abbreviations are given only for relevant parts: dth, dorsal thalamus; vth, ventral thalamus; hth, hypothalamus; ep, epithalamus; htg, tractus hypothalamicus tegmentalis; ch, chiasmatic plate; opc, optic chiasma. Lines inside the thalamus are several fiber bundles (modified from Gilbert, 1934).

nucleus ventralis and arcuate nucleus, and from the lateral part lateral geniculate body (pars lateralis), medial geniculate body, and nucleus lateralis including pulvinar. Therefore, at about 4 months genstatinal age, most of the main thalamic nuclei and fibers are very similar to the complete form. Further cytoarchitectural differentiation of the diencephalon is essentially completed at birth. On the other hand, the development of the telencephalon is remarkable, and in a fully developed state each thalamus locates almost at the center of each hemisphere, sandwiching the third ventricle but completely covered by the cerebral cortex.

B. Gross Anatomy

The coronal section (a vertical plane through the two ears) has been used since the early era of anatomy of the human thalamus. However, horizontal or sagittal sections can clearly demonstrate nuclei that are not easily distinguished in coronal section. Furthermore, for practical purposes, a stereotactic section using an intercommissural line (an imaginary line connecting the anterior and posterior commissures through the third ventricle) is now widely used for intervention into the thalamus, which lies just adjacent to the third ventricle. Therefore, a three-dimensional (coronal, sagittal, and horizontal) stereotactic atlas is commonly used. Today, in the era of computerized three-dimensional imaging, the idea of the stereotactic method is receiving increasingly more important meaning.

Looking at a horizontal section of the brain shown in Fig. 3, the large ellipsoid mass of gray matter in the center is the thalamus. It is surrounded by the basal ganglia, with the intermediate of the internal capsule: anteriorly by the canudate nucleus and laterally by the lenticular nucleus. Within the thalamus, one can easily

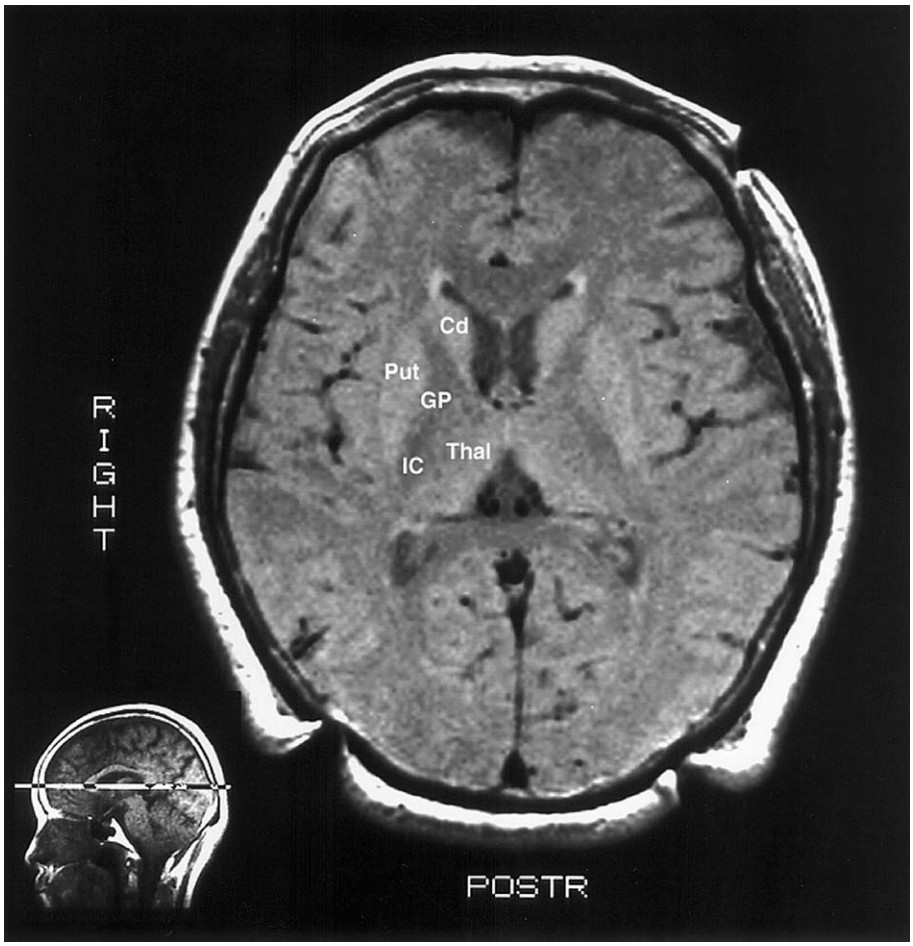


Figure 3 MR axial image showing the thalamus (Thal) and its surroundings. Cd, caudate nucleus; GP, globus pallidus; IC, internal capsule; Put, putamen. (Bottom left) A scout view for cutting level.

identify the fiber-rich longitudinal structure named the internal medullary lamina, which separates the thalamus into three parts—anterior, lateral, and medial groups. Therefore, the internal medullary lamina is a good landmark for orientation in the thalamus, and the thalamic nucleus is generally named after the relative geographic location: anterior nucleus, posterior nucleus, medial nucleus, midline nucleus, intralaminar nucleus, dorsal nucleus, and ventral nucleus.

In reality, however, the nomenclature of the human thalamus is complicated. Traditionally, there are at least two different streams, German and Anglo-American; the former depended on myeloarchitecture and the latter on cytoarchitecture. Moreover, it is often the case that each anatomist uses his or her own naming, and there is no uniform system for the nomenclature of animal thalamic nuclei. Therefore, this confusion has resulted in further misunderstanding and controversy. One should be careful in reading the literature or in referring to an atlas of the human thalamus. In this article, one of the most prevalent styles in the field of human stereotaxy since the publication of *Stereotactic Atlas of the Human Thalamus*, the German style of Hassler's naming, is used, especially for the ventrolateral part. An example of the stereotactic atlas for three different sections (horizontal, coronal, and sagittal) is shown in Fig. 4.

The whole thalamus is surrounded by a thin wall of reticular nucleus, which is more conspicuous anteriorly. Also, the internal medullary lamina is easily distinguished. The most anterior is the lateropolar nucleus. The lateral group consists of (from anterior to posterior) the ventral anterior, ventral oral, ventral intermediate, and ventral caudal nucleus. Each nucleus is usually further divided into several subparts. The main medial group consists of the dorsomedial nucleus, center median nucleus, and parafascicular nucleus. The internal medullary lamina has its own nuclear group, centralis lateralis together with the previously mentioned center median–parafascicular

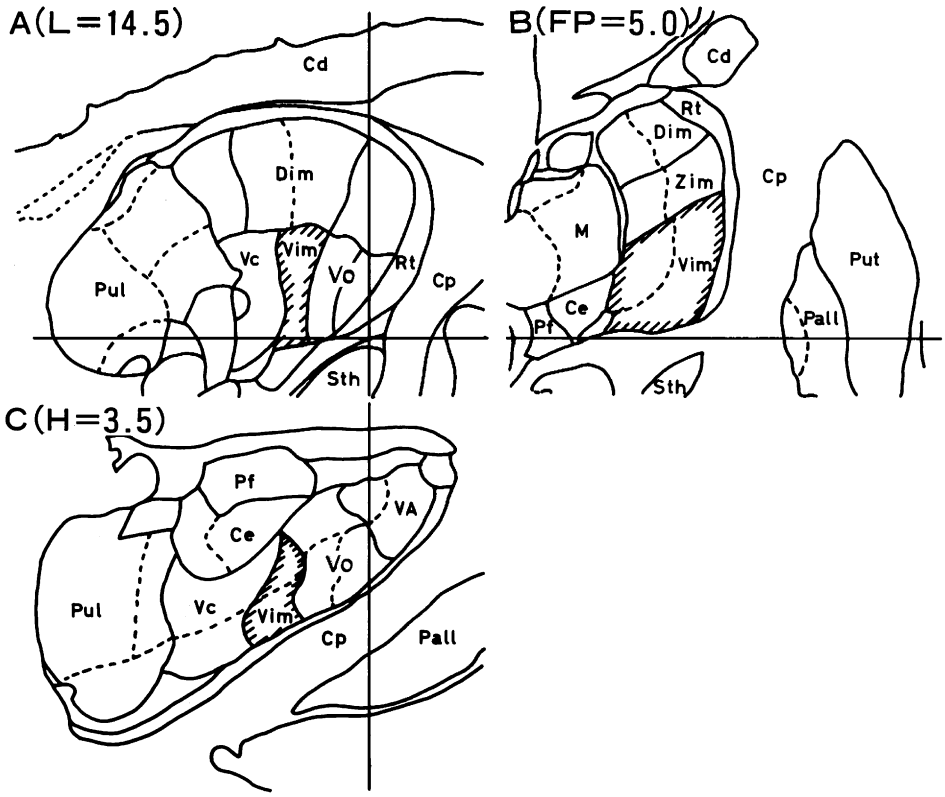


Figure 4 Examples of the stereotactic atlas in common coordinates, with main nuclei. A, sagittal section; B, coronal section; C, axial section. Horizontal line is drawn at the level of the intercommissural (IC) line. Vertical line is drawn at the midpoint of the IC line. Cd, caudate nucleus; Ce, centromedian nucleus; Cp, internal capsule; Dim, dorsal intermedius nucleus; M, dorsomedial nucleus; Pall, globus pallidus; Pf, parafascicularis; Pul, pulvinar; Put, putamen; R, reticular nucleus; Sth, subthalamic nucleus; Vc, ventral caudal nucleus; Vim, ventral intermediate nucleus Vo, ventral oral nucleus; Zim, zentral intermediate nucleus (naming after Hassler).

complex. The most posterior part is occupied by the pulvinar, the largest nucleus in the human thalamus.

C. Thalamic Cells

Each thalamic nucleus consists of characteristic cellular components, neurons. The morphology of neurons is variable; for example, angular, round, spherical, or oval in shape and large, medium, and small in size. Usually, the similar types of neurons gather together in one nucleus, providing the basis for drawing a line at the frontier to classify subnuclei. It is generally the case that the larger neurons are relay neuron, transmitting peripheral information to the higher centers, and the smaller neurons are intrinsic interneurons connecting intranuclear neurons.

Precise cytometric studies have determined the characteristics of the various nuclei. For instance, in the ventral oral nucleus, medium-sized and small neurons are densely packed, with the cell density being about 110 cells/mm^2/50 μm thickness. The subdivisions of the nucleus, ventral oral anterior and posterior, are almost the same in cytoarchitecture but the ventral oral internus is composed of larger neurons with a less dense distribution. The ventral intermediate nucleus is characterized by large cells and by an area in which cells are sparse (Fig. 5). The cell size peaks at 600–700 μm^2, the largest neuron group in the human thalamus, and at 400 μm^2. The cell density is about 60 cells/mm^2/50 μm thickness, contrasting to that of the rostral nucleus of ventral oral group. The ventral caudal nucleus, located further caudally, consists of a mixture of large and small cells. The histogram of cell size shows a different distribution according to subnuclei (anterior, posterior, externus, and internus), varying from 100 to 550 μm^2. The mean cell density is approximately 120 cells/mm^2/50 μm thickness. These histological findings correlate well with electrophysiological findings discussed later.

D. Chemical Anatomy

Recent advances in neuroscience have opened a new horizon to visualize particular chemical substances, proteins, or peptides related to synapse, receptor, etc. on the cell membrane or in the cell content. For instance, dopamine, acetylcholine, peptides, calcium-binding proteins, and so on are visualized by immunochemical and histochemical staining methods or by radioimmunoassay. In the literature, however, most animal experiments have been performed in the rat, and only a small amount of direct study on human thalamus has been performed. Figure 6 shows an example of a histological section of the human thalamus stained by acetylcholine estelase legant. By this method, a new parcellation of the ventrolateral nuclei was proposed.

In Table I, the distribution of chemical substances in the human thalamus collected from 15 different sources is shown. From these results obtained using techniques, there emerge several interesting hypotheses about the function of the particular thalamic nucleus, which was hitherto unable to be seen.

III. PHYSIOLOGY

Physiological understanding of the human thalamus is limited. Only restricted areas such as ventrolateral tier nuclei (ventral anterior, ventral oral, ventral intermediate, and ventral caudal) and nuclei of the medial group (central lateral, centromedian, and parafascicular) are explored in the course of stereotactic thalamotomy for treating various kinds of involuntary movement and intractable pain. Electrodes are guided into the specific parts, and the signals from neurons responding to stimulation of peripheral body areas are recorded. The electrical stimulation of the specific thalamic point elicits sensation or movement at the periphery.

Because the subjects are inevitably in a more or less abnormal state, one should exercise caution in interpreting the results. For convenience, here description will be made from caudal to rostral.

A. Ventral Caudal Nucleus or Nucleus Ventralis Caudalis

This nucleus is one of the most thoroughly explored, probably because it gives the same sensory response as that obtained from animal experiments. Light touch on a small contralateral body part gives rise to a sharp response of spike discharges (Fig. 7), which is the same as the well-known thalamic sensory response in animal experiment.

The peripheral receptive field is very small, usually several square millimeters in the distal extremity, especially at the finger tip and face and lip. In the proximal parts, the receptive field is relatively large, sometimes several square centimeters. Once a response to natural stimuli is found, another response in an

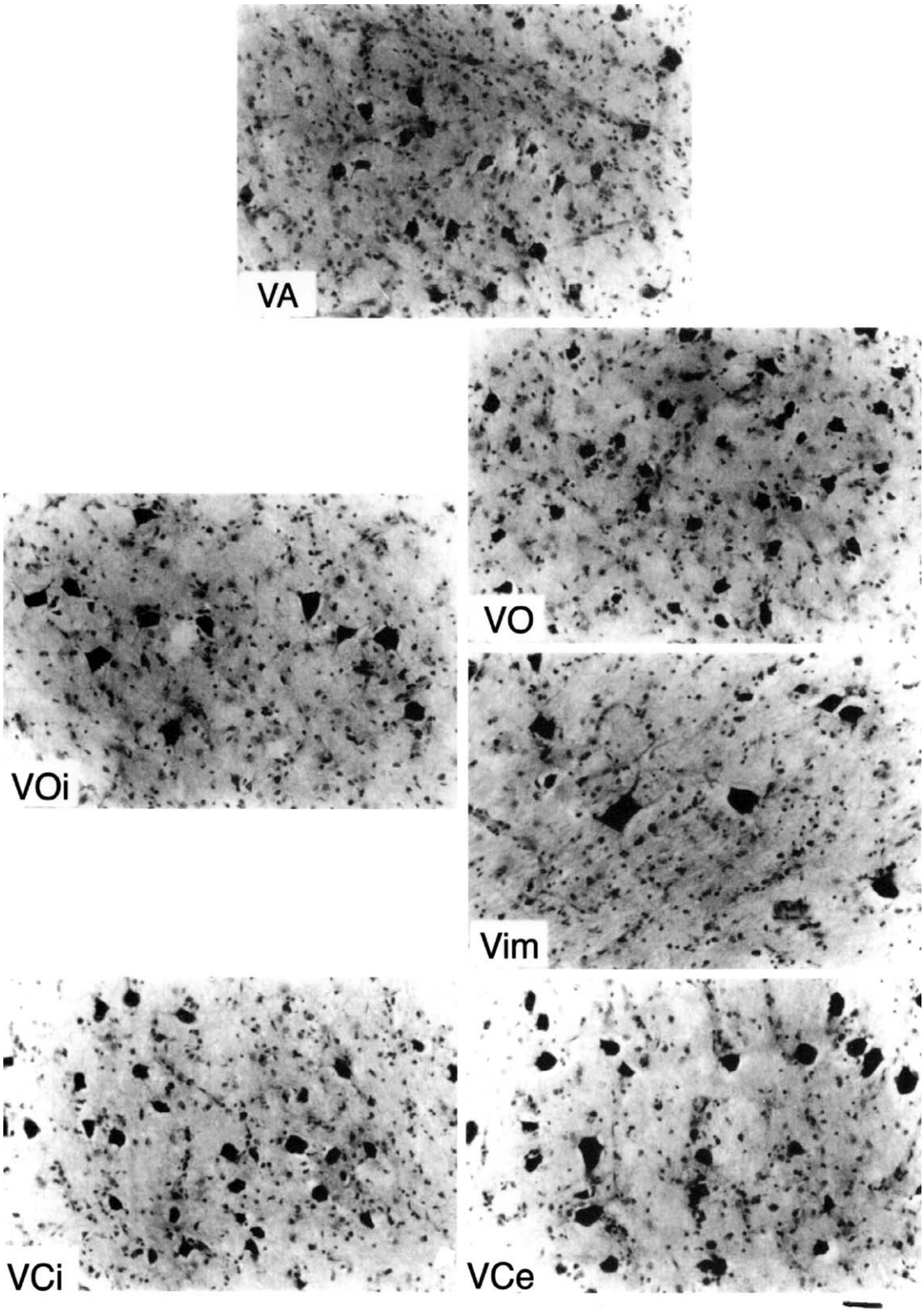


Figure 5 Microphotographs of neurons of the thalamic ventral tier nuclei shown in appropriate positions relative to each other. VA, ventral anterior nucleus; VO, ventral oral nucleus; VOi, ventral oral internus nucleus; Vim, ventral intermediate nucleus; VOe, ventral oral externus nucleus; VCi, ventral caudal internus nucleus. Scale bar = 50 μm.

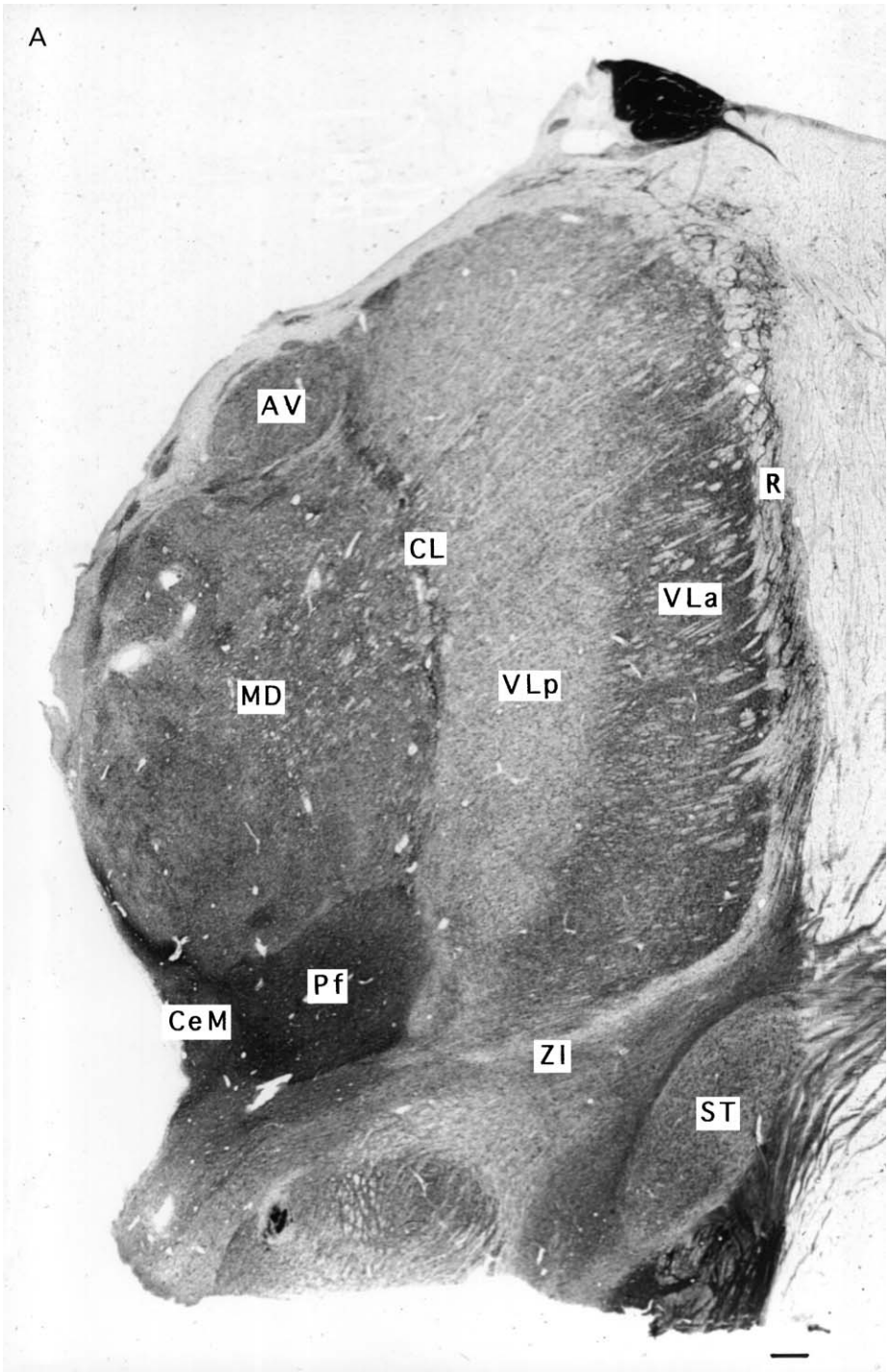


Figure 6 Comparison of two different stainings of almost the same coronal section of the human thalamus. (Left) Histochemical staining by acetylcholine estelase. (Right) Myelin staining AV, anterior ventral nucleus; CL, central lateral nucleus; CeM, center median nucleus; MD, medial dorsal nucleus; Pf, parafascicular nucleus; R, reticular nucleus; ST, subthalamic nucleus; VLa, ventral lateral anterior nucleus; VLp, ventral lateral posterior nucleus; ZI, zona incerta (naming after Jones) scale bar = 1 mm (courtesy of Dr. T. Hirai).

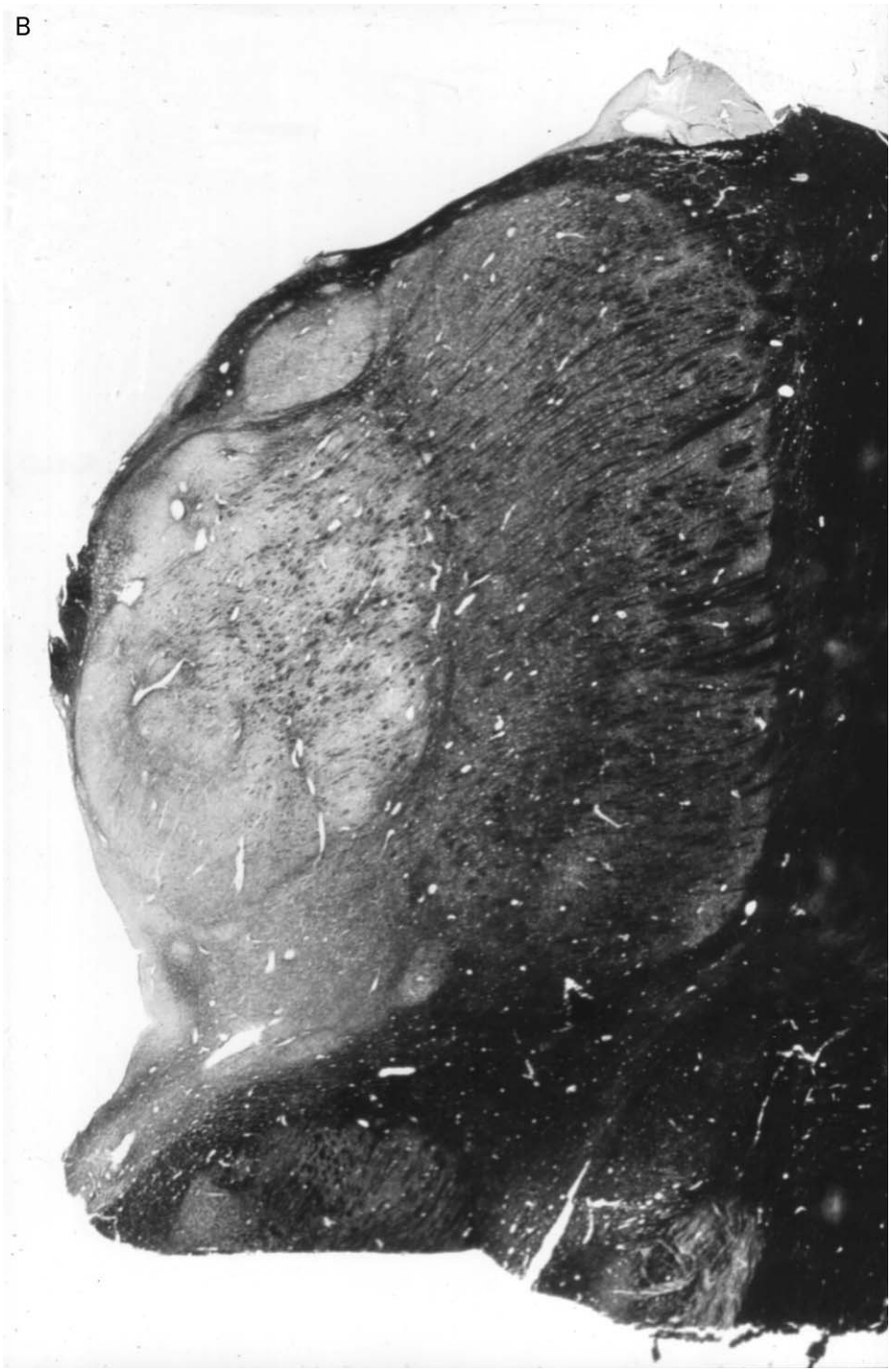


Figure 6 *(continued)*

adjacent area of a given receptive field follows as the electrode descends. For example, if the first response is to light touch on the chin area, the next response is from the buccal area, followed by responses from the nasal labia, along several millimeters of descent. In this sense, there is a topographic representation within the ventral caudal (VC) nucleus. Generally, the leg area is represented in the dorsolateral part of the nucleus, the

Table I
Distribution of Chemical Substances in the Human Thalamus

Classification	Substance	Location	Remarks
Acetylcholine	CHAT	All thalamic nuclei	Affiliation with limbic structure and influence of ascending reticular system
		CHAT-positive axons and varicosities	
		Medial habenula	
		CHAT-positive perikaria	
		Some intralaminar nuclei	
		Dorsomedial N	
		Patch dense, matrix weak	
	[^{3}H] Nicotine	Anteroventral N and dorsomedial N	Heterogeneous distribution
	[^{3}H] Ach		
	[^{3}H] Nicotine	Lateral dorsal N, medial and lateral geniculate bodies	Different from animal
		Anterior nuclei	
Monoamines	Norepinephrine	Medial group, intralaminar nuclei-high pulvinar-low level	
			Dopamine (-)
	Serotonin	Similar as above but pulvinar- considerable increase	
Amino acids	GABA	Local circuit neurons in thalamic visual domain	Species difference
		Dorsal thalamus	
		Reticular N	
	Glutamate	Caudal part of medial group	
		Rostral part of pulvinar	
	Asparate	Middle part of the thalamus	
Gut–brain peptides	Cholecystokinin	Dorsal thalamus, intralaminar nuclei	
	Neurotensin	Periventricular nuclei	Widespread in CNS
	Neuropeptides Y	Periventricular zone	Fiber reaction
	Tachykinin	Immunoreactive fiber	Related to pain
		Posterior group (posterior nucleus, limitans nucleus, suprageniculate nucleus)	
		Intralaminar nuclei	
		Submedius nucleus	
Hypophysiotropic peptide	Somatostatin	Reticular nucleus, dorsal nuclei	
	DO	Anterior nuclei	Wide distribution
	Somatostatin		Septum
			Corpus striatum
Calcium binding proteins	Calbindin D-28	Central lateral nucleus, paracentral nucleus, central medial nucleus, parafascicular nucleus	All intralaminar nuclei
	Calretinin	DO Central lateral nucleus, paracentral nucleus, central medial nucleus parafascicular nucleus	
	Palvalbumin	Centromedian nucleus	
	Calretinin (gene)	Reticular nucleus, dorsomedial nucleus, intralaminar nucleus (rostral)	Heterogeneous
	Calretinin (immunohistochemical)	Very intense: submedius nucleus	Related to emotional and motivational state
		Strong: midline group	
		Moderate: reticular nucleus and anterior medial and lateral group	
		Weak: medial and lateral geniculate bodies	

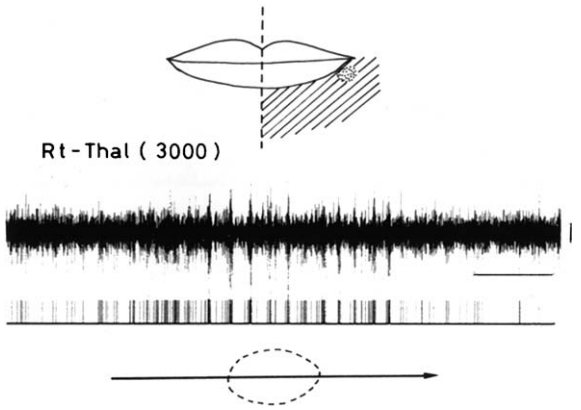


Figure 7 Cutaneous response of a thalamic VC neuron. (Top) Peripheral receptive field just under the left lower lip is illustrated. The small dotted area is the most sensitive. (Bottom) Thalamic responses are shown. Repetitive light touch is applied passing through the most sensitive area (dotted circle) from left to right (arrow).

hand area in the medioventral part, and the face area in the most medial part. No neuron has been shown to respond to different receptive fields or to different modalities (such as joint movement, heat, or pain). Inhibitory responses including so-called lateral inhibition that could be revealed by suppression of spontaneous spike discharge seem to be very rare.

When a stable response to natural tactile stimuli is recorded, electric stimulation on the same skin surface elicites the same spike response in the thalamus, with a fixed latency of approximately 10 msec (conduction velocity about 40–50 m/sec) in a given case. In analogy with animal experiments, this response is mediated via the lemniscal pathway and further projects to sensory cortex of Brodmann's area 3–1–2 although there is no direct evidence in humans.

In certain cases, at the thalamic point at which a cutaneous response is recorded, electric stimulation is given to test the sensations thereby elicited. Subjective perceptions are paresthesia, numbness, tingling, or electric sensation in the receptive field. Such unique observations may be realized only in human cases. However, it is not known how and where touch sensation arises. Also, it is noteworthy that destruction of a part of the VC nucleus results in paresthesias in the contralateral corresponding receptive field. Therefore, VC nucleus is an area that cannot be invaded for therapeutic purposes.

Anteroposterior lamellar distribution of dendrites of a sensory relay cell demonstrated by single cell staining in experimental animals has not been anatomically determined in human thalamus. However, a recent physiological study in the human thalamus demonstrated a possibility of similar lamellar organization in this nucleus.

B. Ventral Intermediate Nucleus or Nucleus Ventralis Intermedius

Rostral to the VC nucleus, between the ventral oral (VO) and VC nuclei, there is a wedge-shaped zone named the ventral intermediate (Vim) nucleus that contains clusters of large cells, scattered throughout. This fact is related to electrophysiologically very high spontaneous background activity. Also, recent histochemical staining revealed a cholinesterase-poor characteristic, in contrast to the cholinesterase-rich nature of VO and VC nuclei neurons.

In this zone with particularly high spontaneous activity, rhythmic grouped discharge time locked with contralateral tremor was found in Parkinson's disease in the early 1960s. It was a remarkable discovery made by French investigators and subsequently confirmed by many who used microrecording during the course of stereotactic surgery.

The early assumption was that the rhythmic discharge might be of a tremorogenic nature (origin of tremor); however, it is mainly the results of afferent impulse originating from trembling muscle. The most pertinent evidence is that the same unit discharge involved in rhythmic discharge also responds to passive stretch of the corresponding muscle and compression on that muscle can produce the same spike discharge (Fig. 8).

The peripheral receptive field is strictly contralateral, and there is no convergence from a different receptive field or different modality. In this sense, it is termed a "kinesthetic" response. The distribution of kinesthetic cells is mostly confined within the anatomically defined Vim nucleus or slightly extends caudally and dorsally. They are always found among a very high background activity zone (suggesting a zone of large cells) and the recorded points on X-ray film correspond well to the Vim nucleus zone in the stereotactic standard atlas. However, it should be noted that only the most lateral part of the nucleus is involved.

The most posteroventral area of the kinesthetic zone may extend slightly into the anteriormost part of the VC nucleus, namely the ventrocaudalis externus anterior, but there is no mixture of cells responding to kinesthetic and cutaneous stimuli, the former always located anterior to the latter. This fact is very important for practical, operative purposes because a

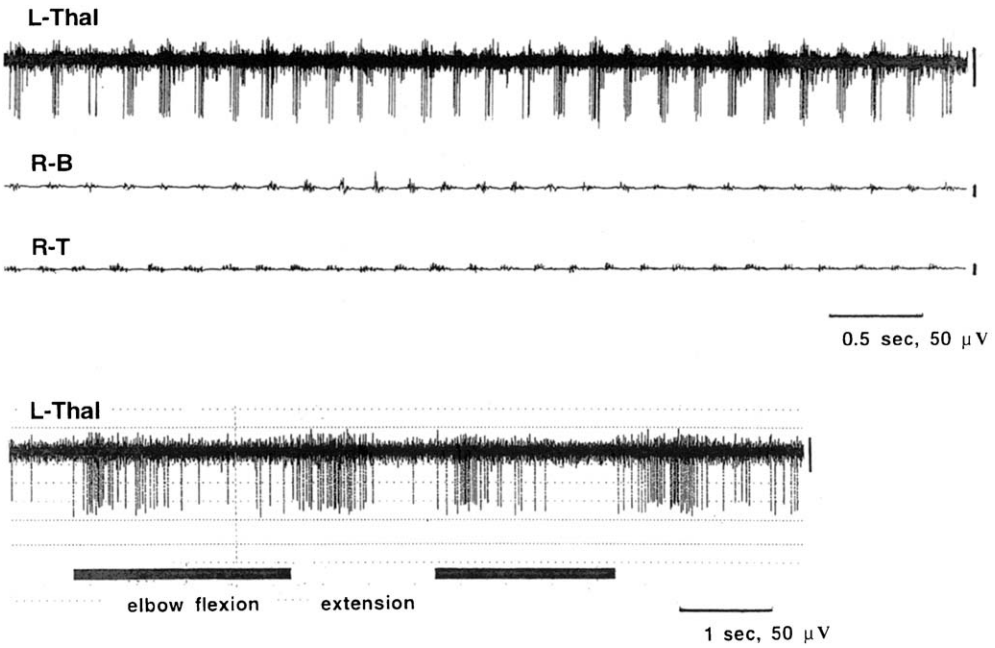


Figure 8 Rhythmic grouped discharge time locked with the contralateral peripheral tremor (top) and kinesthetic response (bottom) in the thalamic Vim, recorded in a case with Parkinson's disease. In the top set of recordings, the first trace is from the left thalamic Vim nucleus, and the second and third EMG traces are from the right upper arm. B, biceps muscle; T, triceps muscle. In the bottom set of recordings, the first trace is from the left thalamic Vim (the same neuonal activity as above), the thick lines depict the moment of elbow flexion, and spaces between thick lines depict the moment of elbow extension. The Vim neuron responded at both flexion and extension.

final therapeutic lesion (minimal effective volume is about 40–60 mm^3) including this kinesthetic region results in complete arrest of tremor, irrespective of its cause, without affecting cutaneous sense. It seems plausible that a point-to-point correspondence exists between peripheral tremor and thalamic kinesthetic cells.

Like the thalamic cutaneous cell in the VC nucleus, the kinesthetic cell in the Vim nucleus responds to electric stimulation given to the contralateral peripheral nerve innervating its receptive field. The shortest latency is about 10–12 msec from upper limb stimulation, and there is slightly faster conduction velocity than that of cutaneous origin (60 m/sec). The spinothalamic tract is postulated to be a possible ascending pathway mediating kinesthetic sense, but another possibility via the cerebellum cannot be denied. From Vim, it is thought to project to the cotical 3a area (Brodmann). However, there is no direct evidence in humans; hence, this should be clarified.

Acute electric stimulation of this kinesthetic point usually suppresses tremor. This fact is used widely for the treatment of parkinsonian tremor and other types of tremor. For this purpose, stimulation is applied through the chronically implanted electrode. The tremor stops only during repetitive (approximately 100 Hz) electric stimulation while the voluntary movement is maintained normally. Because this method is reversible in one sense, it is accepted for therapeutic purposes.

Careful observation on electric stimulation revealed that it often elecits a sensation of numbness or electric sensation at the peripheral receptive field. Also, stimulation of the medial part induces a turning sensation or elevator feeling, suggesting vestibular activation. It is claimed that the vestibular system projects to the medial part of Vim, the so-called Vim internus. In this regard, it can be argued that the main, medial part of the Vim nucleus (kinesthetic zone is only the lateral part) may be the receptive area of the cerebellar afferent, being the equivalent area of the so-called VL nucleus relay station of dentatothalamocortical motor area.

C. Ventral Oral Nucleus or Nucleus Ventralis Oralis

Early attempts (in the 1960s) to make a stereotactic therapeutic lesion in the VO area [ventralis oralis

anterior (Voa) and ventralis oralis posterior (Vop) of Hassler] were based on anatomical knowledge that pallidal fibers terminate in this thalamic area of Voa and cerebellar projection in Vop. The coagulative lesion was in fact large enough to cover the whole VO and may include the Vim nucleus. As a result of using microrecording techniques during the course of stereotactic surgery, the functional organization of this particular part is now fairly well understood.

Myeloarchitecturally, the VO nucleus is divided into two parts, anterior (Voa) and posterior (Vop), the former being more darkly stained. The Voa nucleus is believed to receive fibers originating from the internal segment of the globus pallidus and to project to the premotor cortex of Brodmann's area (BA) 6; the Vop is considered to receive cerebellar projection fibers and to send axons to motor cortex of BA 4, although direct evidence is lacking. Currently, however, there is ample evidence against the classical idea. Namely, as mentioned previously, both Voa and Vop consist of small-sized angular neurons; cell density is almost homogeneous. Moreover, VO neurons as a whole receive projections from the striatum, from the internal segment of globus pallidus. In an awake state, this nucleus exhibits low spontaneous background electrical activity superimposed by small-amplitude spike discharges. There is no sensory response to natural stimuli given to the contralateral body part, but activity related to voluntary movement occurs, as mentioned later. These features are recognized as characteristics of the VO nucleus, in marked contrast to its caudal nuclei, Vim and VC.

In the VO nucleus, so-called voluntary movement units exist, as shown in Fig. 9. The unitary spike activity increases firing only to an active movement of a part of the contralateral extremity and not to passive movement or to light touch on the corresponding skin surface. Response initiation time analysis revealed that the thalamic unit started firing before execution of the contralateral muscle contraction. The voluntary movement unit may transfer the signals to motor cortex. However, there remains a delicate problem of initiation of voluntary action.

In this region, other complex responses have been reported, including bilaterally activated cells, multi-unitary activity associated with arousal responses, and activation of cells of a complex polyfunctional type.

Electrical stimulation of the VO nucleus also produces changes in tremor rhythm in the contralateral extremity, if tremor exists. However, the response is variable; sometimes there is an increase of tremor and sometimes a suppression of tremor, suggesting that the VO nucleus is not directly related to tremor mechanism. Other motor effects, such as arrest of ongoing speech, errors in naming objects, and a kind of perseveration (repetition), were also noticed.

Clinically, a small stereotactic coagulation restricted to the VO nucleus results in alleviation of rigidity (muscle tone) but not of tremor in Parkinson's disease. Thus, the same procedure is effective for the treatment

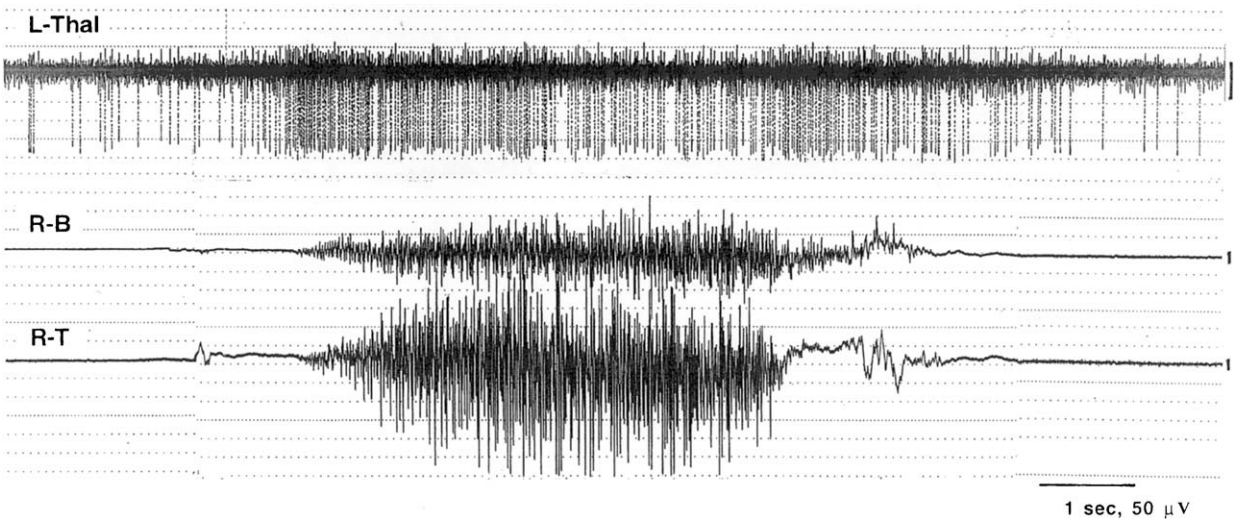


Figure 9 A left thalamic VO neuron related to voluntary action. The neuronal discharge increased (first trace) slightly before the start of elbow movement (attempt to raise arm) against examiner's resistance, indicated by EMG discharge of the contralateral upper arm. B, biceps muscle; T, triceps muscle.

of partial dystonia, several kinds of dyskinesia, especially DOPA-induced dyskinesia, and choreic movement. It should be noted that these disorders are related, more or less, to abnormal muscle tone.

Finally, the Voa and Vop nuclei have been regarded (according to Hassler's concept) as different components in the motor thalamus; curiously, however, both are considered to have almost the equivalent structure as the well-known motor thalamic nucleus in experimental animals, the ventrolateral (VL) nucleus. However, it would be reasonable to revise this concept: Voa and Vop nuclei are homogeneous with one of the thalamic motor nuclei, receiving striatal input. We assume that the majority of the Vim nucleus (except the lateral part) may correspond to the VL nucleus.

To summarize the current understanding of the input–output relations in the main ventral tier nuclei of the human thalamus, a simplified schema is shown in Fig. 10.

D. Other Nuclei

Apart from the previously mentioned ventral nuclei that have been investigated extensively because they are the target of choice for stereotactic surgery, there are several other thalamic nuclei that have attracted attention.

1. Ventral Anterior Nucleus

The ventral anterior (VA) nucleus located just rostral to VO, divided at least in two parts: VA magnocellular (VAmc) and VA parvocellular (VApc). VAmc receives projection from the substance nigra reticulata and probably sends fiber to the prefrontal cortex of areas 8–10, whereas VApc receives the pallidal projection and sends fiber to the cerebral cortex of area 6. It should be noted that there is some discrepancy in the naming of the pallidal projection nucleus. Experiments in monkeys showed that the pallidothalamic projection fiber terminates in two part of VA: VApc and VA densicellular (VAdc). According to this classification, VAdc may correspond to VLo of Olszewski or Voa of Hassler, a member of the VO nucleus.

Few physiological studies on these nuclei have been conducted. A Canadian group found rhythmic discharge time locked to the contralateral tremor in the VA zone (subdivision not defined) and a coagulation in this area was also effective for parkinsonian tremor. If

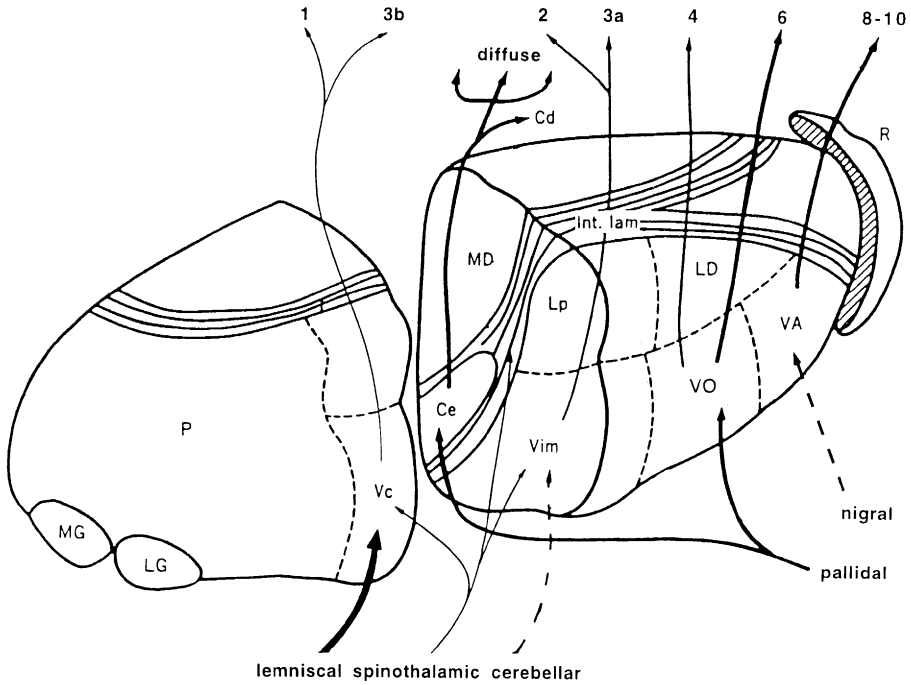


Figure 10 Simplified schema of input–output relations of the thalamic ventral tier nuclei. Numbers indicate Brodmann's cortical numbers. Cd, caudate nucleus; Ce, centromedian nucleus; Int. lam, internal medullary lamina; LD, lateral dorsal nucleus; LG, lateral geniculate body; Lp, lateral posterior nucleus; MG, medial geniculate body; P, putamen; R, reticular nucleus; VA, ventral anterior nucleus; Vim, ventral intermediate nucleus; VO, ventral oral nucleus.

pallidal surgery is effective for tremor as recently claimed, for coagulation of the VA nucleus, a part of which receives pallidal projection, it is not surprising to find tremor rhythms in VA. Unfortunately, there are no follow-up studies.

2. Reticular Nucleus

The reticular nucleus wraps around the thalamus. As described previously, it originates from the ventral thalamus embryologically, being differnt from other neurons of the so-called dorsal thalamus. It has a wide network with the other thalamic nuclei as well as cerebral cortex and brain stem reticular formation. This nucleus is considered to be related to arousal, attention, cognitive function, etc. Also, as discussed later, it plays a role in maintining cortical activity in a disease state of epilepsy.

A Russian group studied the human thalamus using microrecordings during stereotactic thalamotomy for dyskinesia and found verbal command neurons in this nucleus and adjacent area. They classified three types of neurons: A-type neurons exhibited irregular sporadic spikes that were usually activated by imperative verbal command, B-type neurons showed spontaneous rhythmic burst that was inhibited during the command presentation, and C-type neurons showed aperiodic long-lasting burst discharge without responding to verbal command. They discussed a possible role for a basic regulatory mechanism allowing the performance of speech-mediated voluntary movements.

3. Noxious Neurons

It has long been debated whether there is a particular group of neurons responding solely to noxious or thermal stimuli in human thalamus as well as in animals. Such nociceptive neurons have been discovered. Neurons of the nucleus submedius located under the VC nucleus are activated by noxious stimuli.

Recently, exploring the human VC nucleus during the pain surgery, a Canadian group discovered a small restricted zone, microstimulation of which evoked pure cold sensation and neurons there responded to innocuous cooling. They claimed that it is the first direct electrophysiological evidence for a human thalamic relay site for the cold sense pathway. This nucleus may correspond to the ventral medial posterior (VMpo) nucleus situated just ventral to the centromedian parafascicular (CM–Pf) complex and caudal to CM and ventral posteromedial nucleus (VPM). The VMpo was described in 1996 by American investigators as a parvalbumin-rich zone in rat and monkey. The VMpo neurons responded solely to noxious and thermal stimuli. The clinical significance of this nucleus is not clear.

4. Pulvinar

Pulvinar is the largest nucleus in the thalamus, located most caudally. On the philogenetical scale, it is a newly developed nucleus, simultaneously developing with an increase in association cortex in the primate. Therefore, interspecies difference is marked. In primate and human, it is roughly divided into four parts: oral (or superior), medial, lateral, and inferior pulvinar. Several different functions have been suggested, but recently it has been considered to be related more to the visual system, playing a role as the secondary visual system or extrageniculate visual system. Especially in the inferior pulvinar of experimental animals, there is a direct connection from the contralateral retina (not via lateral geniculate body). Also, there are close reciprocal connections between visual cortex (BA 17), superior colliculus, prefrontal cortex, parietal association cortex (angular gyrus and supramarginal gyrus), and temporal lobe, being involved in the higher cognitive functions such as symbolic experession. Pulvinar is correlated with attention behavior; therefore, if it is damaged, visual neglect often occur.

Using stereotactic surgery, a group tried to make a lesion here in cases with intractable pain or a kind of involuntary movement. While microrecording during the course of surgery, several pulvinar neurons fired synchronously with the rhythmic electroencephalogram of the pulvinar. Also, some pulvinar neurons increased firing during voluntary action or during combined stimuli, such as voluntary action plus visual or auditory stimuli.

5. Dorsomedial Nucleus

This nucleus is a large mass located medial to the internal medullary lamina. It is divided into different subdivisions, but probably its medial part, the magnocellular part of the nucleus, is relatively well-known because it receives direct input from the rhenal cortex and amygdala, both related to cognitive memory function. It also has connections from the prefrontal cortex, temporal lobe, and basal forbrain area. Hence, it may play an important role in higher nervous function. Clinically, it is well-known that damage to this area induces personality changes and memory disturbance known as antegrade amnesia.

In this connection, it should be remembered that the first human stereotactic surgery was performed to destroy a part of this nucleus in a patient with mental illness by Spiegel *et al*. in the mid-20th century. Their original idea was to eliminate the uncertainty of frontal lobotomy frequently used at that time and to avoid undue damage of overlying healthy brain structures by open surgery.

6. Intralaminar Nuclei

This nucleus, embedded in the internal medullary lamina, consists of centralis lateralis, paracentralis, central medial nuclei (anterior group), and centromedian and parafasciculus nuclei (posterior group). The latter is often called the centromedian–parafascicular complex. The anterior group receives diverse projections from the spinothalamic tract, deep cerebellar nucleus, brain stem reticular formation, etc. The posterior group has a reciprocal connection with the basal ganglia. The efferent connection with the cerebral cortex is very wide and was once thought to be a diffuse projection. However, it is now known that each intralaminar nucleus has its own topographic projection area and projection pattern is somewhat heterogeneous.

The intralaminar nuclei has been a main focus of anatomophysiological investigations on the organization of the thalamocortical system. It was classified as a representative of the ''nonspecific system'' rather than the ''specific system,'' such as the thalamic relay station for the visual, auditory, or somatosensory system with definite modality-specific peripheral input. Due to many multidisciplinary studies performed during the past 50 years, the meaning of nonspecific system has changed because the morphological understanding of the thalamocortical connection is far advanced and the physiological mechanism of recruiting response, for example, has been precisely analyzed into synaptic events. Functionally, on the basis of this recent knowledge, the mechanism of sleep–wakefulness, the dual behavioral basis of animal life including humans is gradually becoming clearer.

IV. THALAMIC DAMAGE

There are several different kinds of thalamic damage that induce particular clinical symptoms. Vascular damage, thalamic tumor, and degeneration are the most common causes. The clinical features are closely related to the site of affected thalamic nucleus. Therefore, precise clinical study, especially clinicopathological analysis, has contributed greatly to understanding the function of the particular thalamic nucleus thus involved. It can be understood that medial thalamic damage leads to disturbance of consciousness, cognitive function, personality, etc. more or less related to higher nervous activity, whereas the lateral thalamic damage results in the disturbance of somatosensory function.

A. Thalamic Syndrome

At the end of the 19th century, a particular syndrome with sensory disturbance, movement disorder, and often severe pain in one side of the body without any damage to the affected area attracted the attention of German neurologists and thalamic vascular damage was suspected as the responsible pathology. Later, French neurologists Dejerine and Roussy described its exact pathological lesion: the thalamic ventralis posterior lateral nucleus (VC of Hassler).

The core symptoms of thalamic syndrome are (i) mild hemiparesis, ususally without contracture, and early recovery; (ii) continuous hypesthesia of superficial sense, sometimes replaced by hyperesthesia of superficial sense, always accompanied by disturbance of deep sense; and (iii) slight hemiataxia and astereognosia. These symptoms are often accompanied by continuous severe pain developing later on the paretic side (it is intolerable but there is no effective remedy) and choreoathetotic movement developing on the side of motor paresis.

The previously mentioned clinical description has been accepted and the vascular damage (either hemorrhage or infarction) in the VPL (VC) was established, as the responsible pathology, together with its responsible artery, the so-called thalamogeniculate artery.

Regarding the developmental course of thalamic syndrome, it was found that the major vascular impact was ususally not related to the later development of thalamic pain but rather minor damage induced pain after several months to 1 year after stroke (Fig. 11). This may be due to the fact that a large lesion destroys the main sensory nucleus of VC and surrounding area concomitantly, whereas a small lesion affects only a part of VC and causes irritaion in the adjacent area. In fact, microrecording during the course of stereotactic thalamotomy showed many abnormal irregular burst discharges in and around the VC nucleus and Vim nucleus as well (Fig. 12). Within the supposed VC

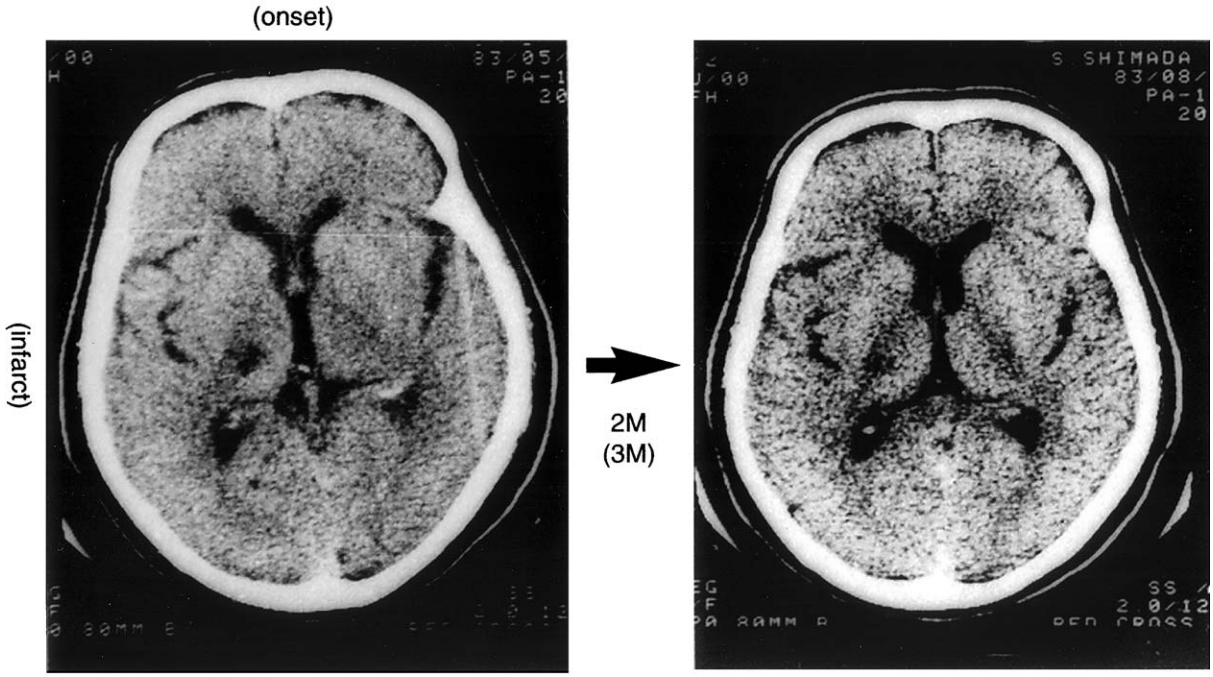


Figure 11 Vascular damage that caused typical thalamic pain. (Left) CT image of the infarction (low-density area) in the right posterior thalamus just after the stroke (onset). (Right) Follow-up CT image in the same patient 2 months (2 M) after stroke. Thalamic pain gradually developed after 3 months (3 M) in this case. The remaining responsible lesion (infarct: low-density area) was much reduced.

nucleus, the background activity was low and the sensory response to cutaneous stimuli was disorganized, with the usual topographic representation being lost and often the face area occupying a wider distribution than that of the usual VC nucleus. Moreover, it was often the case that convergent response to peripheral stimuli was encountered. Therefore, in cases with thalamic pain syndrome, the thalamic VC nucleus is disorganized after vascular damage and some plastic change may take place during the recovery course, leading to the reorganization. Microstimulation of the VC nucleus in the thalamic pain patient induces pain sensation of the patient's own pain. This is noteworthy because usually stimulation of the thalamic VC nucleus results in the sense of paresthesia but not pain.

Clinically, we tried to make a coagulation including the abnormal VC area and a part of Vim where the neuronal activity was also different from that in the usual Parkinson's disease. As mentioned previously, Vim neurons are assumed to receive muscle sense; this operation is more effective for the thalamic pain patient with deep pain than the patient with predominant superficial pain.

The mechanism of the thalamic pain has not been clarified. Dejerine and Roussy discussed the possible abnormal excitation at the level of terminal ascending fibers and terminal neurons of the pain system. There are other hypotheses, for this mechanism, including the released corticothalamic inhibitory system, a disorder of the filter function of the thalamus, and the problem of isolation in the damaged pain conducting fibers. It is interesting to note that a recent study put forward the possible upregulation or downregulation of transmitter receptors (probably noradrenaergic). We assume that the abnormal irregular burst activity of the viable VC neurons caused by the adjacent affected part of the VC nucleus is certainly responsible for production of the chronic pain, but there is no reasonable explanation. Because the development of pain takes time after stroke (often several months), the strong influence from these abnormal zones over the central nervous system, directly or indirectly connected, may cause unusual neural networks to be established in due course.

B. Other Vascular Disorders

Apart from the well-known thalamic syndrome described previously, other types of thalamic hemorrhage and infarction may occur. Using computerized three-dimensional imaging, different kinds of vascular

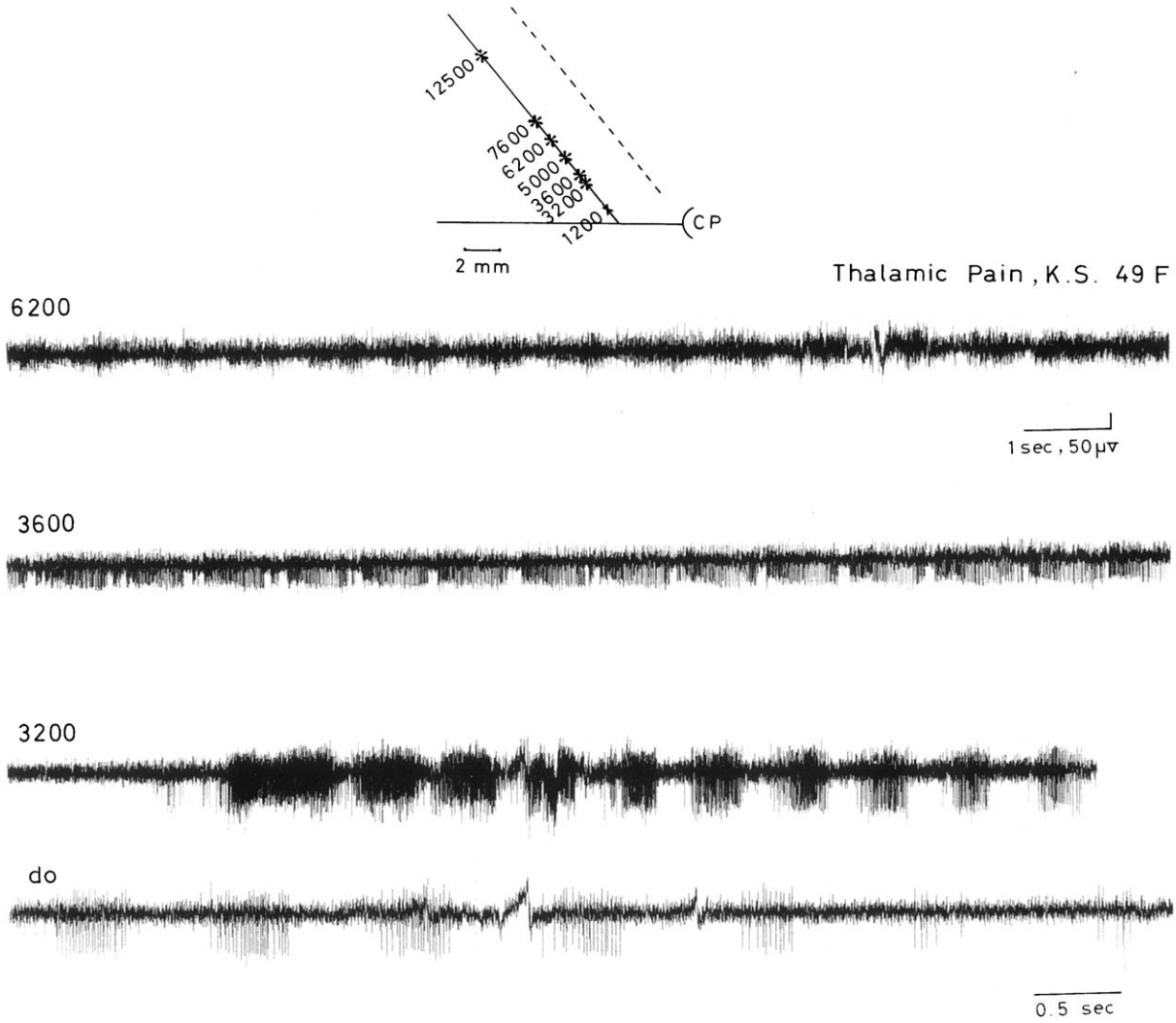


Figure 12 Abnormal irregular burst discharges recorded in a case with thalamic pain. (Top) A lateral view of recording track (an oblique line) referring to the intercommissural (IC) line (a horizontal line) to show the recording site (asterisks) along the anterior track. There was no recording from the posterior track (dotted line) in this case. CP, posterior commissure; Numbers indicate distance in micro meters from zero at the level of the IC line.

damage are easily visualized and these lesions are correlated with clinical symptoms. Thus, it is reasonable to understand thalamic vascular damages with respect to the specialized blood supply to the thalamus. However, there are problems of nomenclature of blood vessels and the variety of blood vessels (origin and course of small thalamic vessels). Therefore, one should be careful to refer to the literature. There are at least five main vessels. Here, the most common naming and synonyms are listed with location of blood supply.

- Anterior choroidal artery, originating from carotid artery

  — Posterolateral part of thalamus
- Tuberothalamic artery (anterior thalamoperforating artery, polar artery), originating from posterior communicating artery (this artery is often lacking)

  — Reticular nucleus, part of VL, dorsomedial nucleus, mammillothalamic tract
- Paramedian artery (posterior thalamosubthalamic paramedian artery, posterior thalamoperforating artery), originating from P1 part of posterior cerebral artery (PC)

  — Internal medullary lamina, dorsomedial nucleus, VPL and VPM (internal side), upper mesencephalon

- Thalamogeniculate artery (principal inferolateral artery), originating from P2 part of PC
  — VPL and VPM, VL, pulvinar
- Posterior choroidal arteries (medial and lateral branch), originating from P2 part of PC
  — Lateral and medial geniculate body, pulvinar, subthalamic region, part of substantia nigra

In Fig. 13, these arteries are illustrated schematically with their respective main locations.

Recent clinicopathological studies of cases with thalamic hemorrhage have demonstrated its clinical features in relation to the responsible artery. Results are summarized in Table II. In cases with infarction, the clinical features are essentially the same as those in cases with hemorrhage but usually less severe because of more restricted damage compared to that of hemorrhagic lesion.

It should be mentioned that contrary to the classical thalamic syndrome, analgesia and thermoanesthesia of thalamic origin have been observed clinically. These clinical symptoms have been known since the early 20th century as poststroke symptoms but their pathophysiological analysis was somewhat behind that for the thalamic pain. As mentioned previously, the recently reported group of specific nociceptive and thermoceptic neurons of VMpo may be responsible for these symptoms, although no direct evidence has been provided.

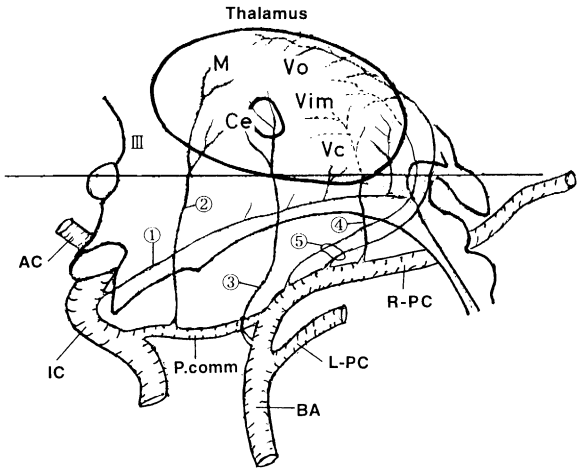


Figure 13 Schematic illustration of the blood supply (1–5) in the thalamus. Right thalamus and the third ventricle are viewed from the medial side. In the thalamus, only representative nuclei are marked. M, dorsomedial nucleus; Ce, centromedian nucleus; Vo, ventral oral nucleus; Vim, ventral intermediate nucleus; Vc, ventral caudal nucleus; 1, anterior choroidal artery; 2, tuberothalamic artery; 3, paramedian thalamic artery; 4, thalamogeniculate artery; 5, posterior choroidal arteries. AC, anterior cerebral artery; IC, internal carotid artery; P. comm, posterior communicating artery; L-PC, left posterior cerebral artery; R-PC, right posterior cerebral artery. Dotted parts of the small arteries indicate lateral, hidden side of the thalamus. Horizontal line is the intercommissural line.

C. Movement Disorder

Contrary to the previously mentioned sensory disturbances caused by damage in and around the VC nucleus, the pathophysiology of movement disorders of thalamic origin is not always clearly understood. Although mild motor paresis, ataxia, choreoathetotic movement, and dystonia are claimed to occur after thalamic damage (by vascular, tumoral, or degenerative lesion), it is doubtful that a restricted lesion in the thalamus produces such motor disturbance. Because these symptoms are otherwise caused by capsular, cerebellar, and basal ganglia and brain stem lesion, they could be explained as secondary symptoms due to the disruption of a specific neural loop connected with one of the thalamic ventral tier nuclei (VA, VO, Vim, and VC).

Moreover, it should be remembered that the stereotactic intervention in certain thalamic nuclei (mainly VO and/or Vim) is useful for the treatment of movement disorders such as tremor and rigidiy of Parkinson's disease and other kinds of tremors, dyskinesia, choreic movement, dystonia, etc.

D. Thalamic Tumor

The incidence of thalamic tumor is low, about 1% of all intracranial primary tumors. In fact, thalamic tumor is usually classified as an item of tumor of rostral brain stem and basal ganglia. There is no specific primary brain tumor that develops within the thalamus, but glial tumors (e.g., astrocytoma, oligodendroglioma, and glioblastoma often in children or adolescents) and malignant lymphoma may grow in the thalamus.

Often, the thalamic tumor grows to a large size, occupying mass without noticeable signs and symptoms. With the progress of tumor size, headache and nausea caused by obstructive hydrocephalus due to compressed foramen Monro are the most common symptoms (Fig. 14). Some specific symptoms indicate the possible location of the tumor. For example, movement disorder is caused by damage or compression of the VO area and invasion into the internal

Table II
Types of Thalamic Hemorrhage and Responsible Arteries

Type	Responsible artery	Main symptoms	Prognosis
Anterior	Tuberothalamic artery	Acute behavioral abnormality	Benign prognosis
Posteromedial	Paramedian artery	Marked hydrocephalus due to bleeding into the third ventricle	Worst prognosis
Posterolateral	Thalamogeniculate artery	Large hematoma	
		Marked sensory motor signs	
		Dejerine–Roussy syndrome	Common
		Right-side hemineglect; left, language problem	
Dorsal	Posterior choroidal artery	Moderate hematoma	Excellent recovery
		Sensorimotor signs	
Global	Combined	Massive hemorrhage	
		Severe sensorimotor signs	Death
		Conscious disturbance	

capsule; sensory disturbance is caused by VC invasion; and mental confusion, personality change, and memory disturbance are related to the medial nuclei and internal medullary lamina. Diagnosis is easily made by computed tomography (CT) scan or magnetic resonance imaging, and malignancy is suspected by positron emission tomography (PET) scan when higher uptake of glucose metabolism or methionin is found. An example is shown in Fig. 15.

As the secondary brain tumor that emanates from other organs besides the brain, the metastatic tumor is increasingly found and the thalamus is also involved. Tumors of the third ventricle (plexus papilloma, ependymoma, and astrocytoma) or basal ganaglia can compress or invade the thalamus.

Because the thalamus is deep in the brain, direct attack for removal of the tumor is a challenging strategy. A stereotactic approach with the aid of three-dimensional imaging and microscope is one a possible procedure. As adjuvant therapies, stereotactic radiosurgery by Gamma Knife or LINAC irradiation often with some kind of chemotherapy are performed. Complete remission is almost impossible and the prognosis is generally poor.

E. Degeneration

Degenerative disease specific for the thalamus is rare. In the literature, only approximately 10 cases have been reported since the first description by Stern. In these cases, the most characteristic clinical symptoms were progressive dementia and conscious disturbance often accompanied by the cerebellar symptom involuntary movement. In most cases, in the thalamus, anterior group and dorsomedial nucleus were symmetrically affected, showing severe cell loss and gliosis. Once, this disease was considered to be a thalamic type of Creuzfeld–Jacob disease (Stern–Garcin syndrome), but now it is recognized as primary thalamic degeneration (without spongiosis). Even in these cases, mild degeneration was noticed in other nucleus, especially in the inferior olivary nucleus. Similar primary degenerative changes in certain thalamic nuclei were reported in Japan. In cases with rapidly progessing mental deterioration (amnesia and dementia) and movement disorder, autopsy revealed symmetric degeneration in the thalamus, especially in the dorsomedial nucleus, lateral posterior and dorsal posterior nuclei, and pulvinar. It was emphasized that these nuclei were a relatively late developing group in the ontogenetic course.

Some of the systemic degenerative process involves thalamic neurons as well. For example, it was noticed that smaller neurons of the VL nucleus (not defined clearly) decreased in Huntington's chorea, in which degeneration of the smaller neurons of the caudate nucleus is essential. Also, a similar change was observed in the VA nucleus. In a particular type of amyotrophic lateral sclerosis, eosinophilic inclusion body was found in the thalamus. Also, in pallido–nigro–Luysian atrophy, Lewy body-like inclusion was found in the thalamic parafascicular nucleus and the centromedian nucleus, in addition to other parts of the central nervous system. Other types of systemic degenerative disease, such as Pick's disease, olivopontocerebellar atrophy, and Friedleich's ataxia, are also known to exhibit thalamic degeneration

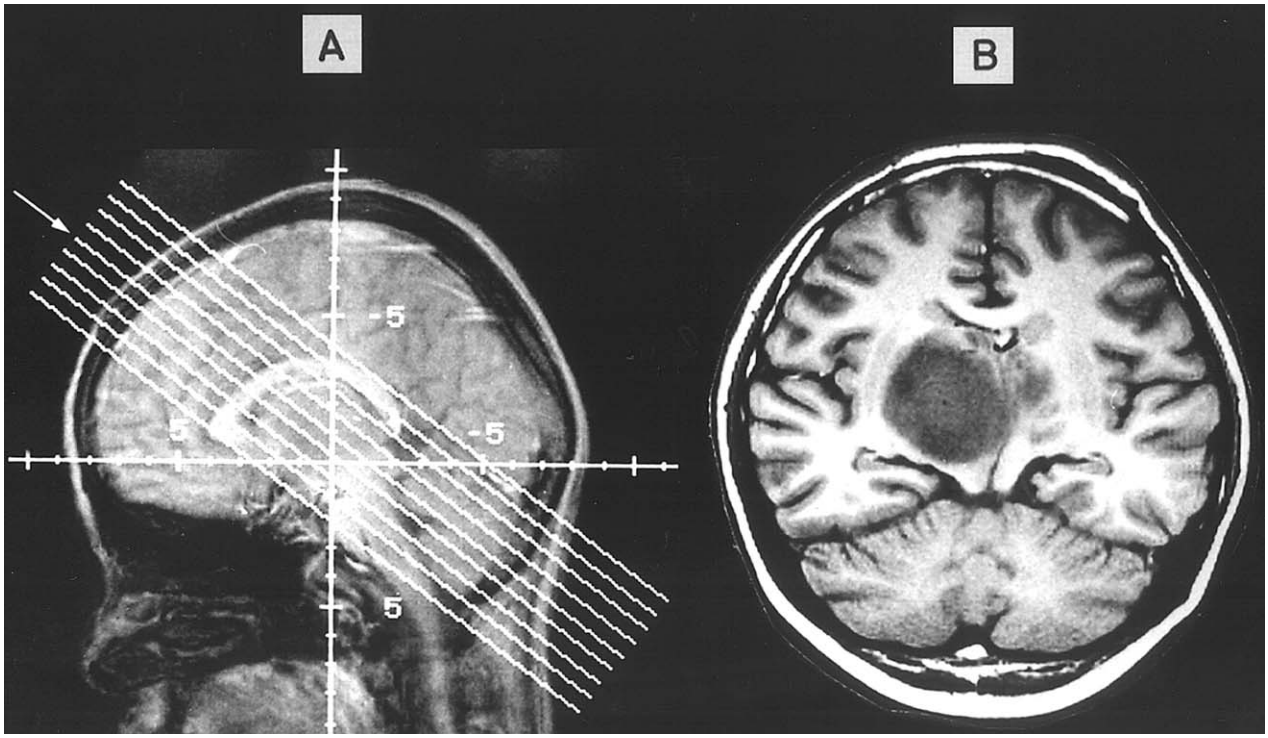


Figure 14 MRI of the right thalamic tumor (malignant glioma) in a young girl. (A) Scout view showing the direction of the brain slice. Arrow indicates the semiaxil slice corresponding to the image in B. (B) A large thalamic tumor in the right thalamus, compressing the lateral ventricles.

concomitantly. In the literature, however, description of the degeneration in the thalamus is not frequent.

F. Others

In certain conditions of thalamic damage by vascular, tumoral, or degenerative effects as mentioned previously, several types of abnormal state are recognized.

1. Thalamic Amnesia

When the anterior part of the thalamus is destroyed, so-called thalamic amnesia (antegrade type) may occur. Precise analysis (lesion that may cause amnesia vs lesion that may not cause amnesia) of such cases using CT scan revealed two important elements. One is the mammillothalamic tract connecting the mammillary body and the thalamic anterior nuclei, forming a loop from hippocampal formation to mammillary body and anterior thalamic nucleus. The latter tract is also known as Vic d'Azur. The other is the amygdalothalamic projection (ventroamygdalofugal pathway), connecting amygdala and thalamic dorsomedial nucleus. In the anteromedial thalamic area, the latter pathway runs adjacent to the mammillothalamic tract. Therefore, it is interesting to note that only when these two tracts are interfered does thalamic amnesia occur; if one of them is spared, no amnesia is noticed.

Therefore, the essential element for thalamic amnesia is not a lesion of the thalamic nucleus but fiber tract connecting extrathalamic components related to the memory system. Hence, if the thalamic damage is repaired, amnesia may disappear.

2. Thalamic Aphasia

As in the case of cortical aphasia, it is the posterior part of the left thalamus that may cause thalamic aphasia. Language disturbance of primary thalamic origin has been discussed for a long time; if it really exists, what nucleus is responsible? Although recent imaging and autopsy studies list a part of the dorsomedial nucleus, the centromedian–parafascicular complex, a part of the VL nucleus (VO), the pulvinar, etc. as the responsible thalamic lesion, the exact lesion has not been clarified. For example, the role of the dorsomedial nucleus was strongly suspected, but there were cases without aphasia in which the dorsomedial nucleus was damaged.

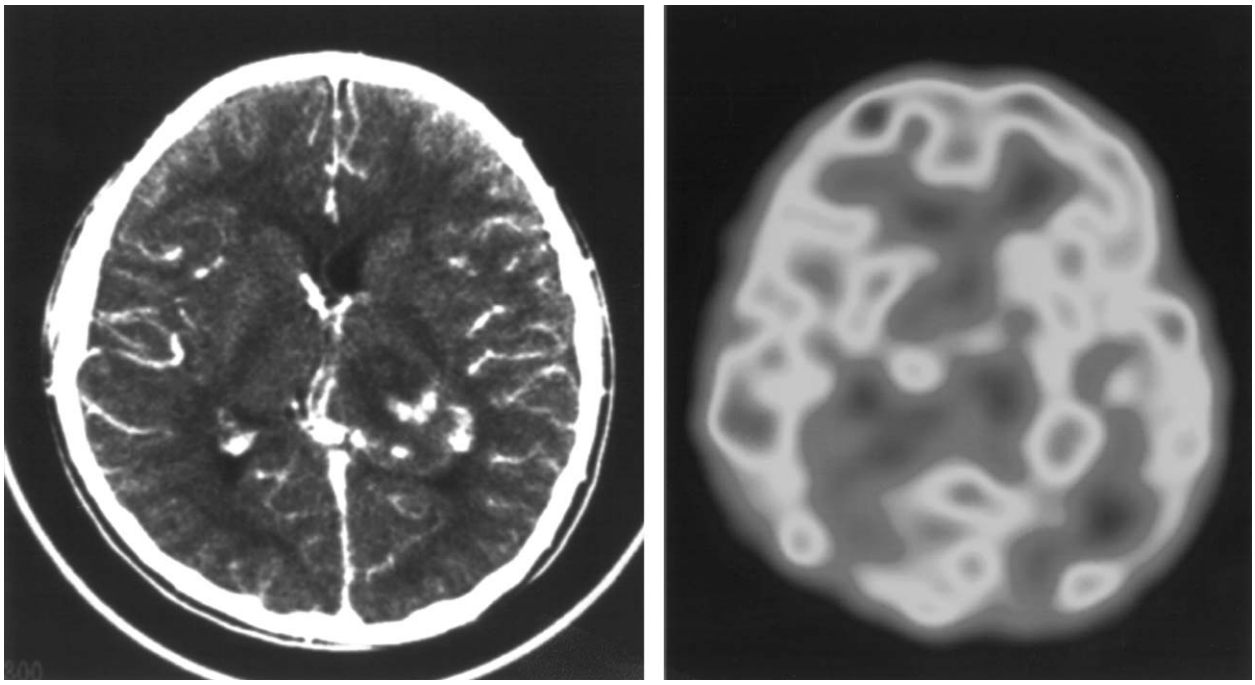


Figure 15 A left thalamic tumor demonstrated by CT image (left) and PET scan (right). The thalamic tumor is seen in the posterior part of the left thalamus as a low-density shadow with an irregular high-density area at its caudal end, the latter showed high uptake of FDG in the right PET image. Diagnosis was anaplastic astrocytoma in this 11-year-old boy. (courtesy of Dr. T. Shibasaki).

Another possible explanation for thalamic aphasia is the decreased thalamic influence on the related cortical area. In fact, cortical blood flow measurement by PET or single photon emission computerized-tomography has often revealed a marked decrease of blood flow in the left cortical area in cases of thalamic damage with aphasia. In such cases, thalamic damage causes a decrease in excitatory input to the cortical activity, which in turn reduces its blood supply. Further analytical study is expected.

The thalamic aphasia is similar to a type of so-called transcortical aphasia. In this category, repetition is preserved, although other language functions, such as spontaneous speech, comprehension, reading, and writing, are more or less disturbed.

It is relevant to mention studies of stereotactic thalamotomy on language function. The arrest of speech by electrical stimulation of the VO area was reported several times. It occurs only during repetitive electrical stimulation, for example, when a patient is counting numbers or telling a story. Whether it is due to an effect on the specific language mechanism or just tremor arrest on movement in general is not clear.

Also, it is well-known that bilateral thalamotomies (not in one procedure) induce speech disturbance, often referred to as dysarthria and dysphasia (mild aphasia). Sometimes, only one-sided thalamotomy results in the same kind of speech disturbance (left side is dominated). However, with the aid of microrecording to identify the exact target point in the Vim nucleus for tremor surgery, for example, postoperative speech disturbance is much reduced even after bilateral operation. Probably, a relatively large lesion including the thalamic pharyngial representative zone is responsible for inducing such speech problems.

3. Epilepsy

Among various epileptic seizures, petit mal or absence is considered to be closely related to the thalamus, in the sense that projection neurons of the thalamus play an important role in cortical seizure activity. According to basic studies, the thalamocortical triangular circuit that consisted of cortical neurons of the superficial layer, thalamic reticular neurons, thalamic diffuse projection neurons of the intralaminar nucleus, is the essential element. The synaptic events in this thalamocortical reciprocally connected loop maintain the seizure activity.

Clinically, absence seizure is characterized, in a typical case in children, by abrupt (no aura) arrest of movement or behavior, with head drop posture or upward gaze in vain with eyes open. Within a few second, consciousness returns completely to the normal level without any postictal deficit. The patient has no memory of the seizure. During epileptic seizure, bilaterally synchronized and generalized (but frontal dominant) "three-cycle spike and wave" in EEG is well-known and has diagnostic value. Also, three-cycle spike and wave synchronized with cortical EEG has been recorded from the thalamic depth. Thus, it was once called a centroencephalic seizure, which includes the generalized type of epileptic seizure.

V. CONCLUSION

Anatomy, physiology, and disease of the human thalamus were described briefly. Although the modern techniques of neuroscience are limited to use in humans, meticulous clinical observation and stereotactic surgery experience with microrecording provide useful information about the structure and function of the human thalamus. Many problems remain to be solved, but with the development of new techniques the future of thalamic study is promising.

See Also the Following Articles

ALERTNESS • BASAL GANGLIA • CEREBRAL CORTEX • EPILEPSY • HYPOTHALAMUS

Suggested Reading

Duus, P. (1989). *Topical Diagnosis in Neurology*. Thieme, Stuttgart.

Gilbert, M. S. (1934). The early development of the human diencephalon. *J. Comp. Neurol.* **62,** 81.

Hassler, R. (1977). Architechtonic organization of the thalamic nuclei. In *Atlas for Stereotaxy of the Human Brain* (G. Schaltenbrandt and W. Wahren, Eds.). Thieme, Stuttgart.

Jones, E. G. (1985). *The Thalamus*. Plenum, New York.

Ohye, C. (1977). Thalamus. In *Encyclopedia of Human Biology*, 2nd ed., pp. 381–391. Academic Press, San Diego.

Ohye, C. (1990). Thalamus. In *The Human Nervous System* (G. Paxinos, Ed.), pp. 439–468. Academic Press, San Diego.

Ohye, C. (1998). Thalamotomy for Parkinson's disease and other types of tremor. Part 1: Historical background and technique. In *Textbook of Stereotactic and Functional Neurosurgery* (P. L. Gildenberg and R. R. Tasker, Eds.), pp. 1167–1178, McGraw-Hill, New York.

Parent, A. (1996). *Carpenter's Human Neuroanatomy*. Williams & Willkins, Baltimore.

Sidman, R. L., and Rakic, P. (1973). Neural migration, with special reference to developing human brain. A review. *Brain Res.* **62,** 1.

Steriade, M., and Jones, E. G., and McCormick, D. A. (Eds.) (1997). *Thalamus. Vol. II: Experimental and Clinical Aspects*. Elsevier, Oxford.

Time Passage, Neural Substrates

DEBORAH L. HARRINGTON
Albuquerque Veterans Affairs Medical Center and University of New Mexico

STEPHEN M. RAO
Medical College of Wisconsin

GLOSSARY

basal ganglia A group of nuclei situated in the internal portion of each cerebral hemisphere. The definition of basal ganglia varies but typically includes the caudate nucleus, putamen, globus pallidus, subthalamic nucleus, and substantia nigra. Different terms are applied to various subdivisions of the basal ganglia. The putamen and globus pallidus comprise the lenticular nucleus. The caudate nucleus and putamen comprise the striatum.

Brodmann areas Neuroanatomically distinct areas of the cerebral cortex derived by Korbinian Brodmann, a German neurologist, based on the organization of cells or the cytoarchitecture of the human brain. Functional distinctions of brain regions correlate remarkably well with this neuroanatomic differentiation.

cerebellum The cerebellum is attached to the posterior surface of the brain stem by fiber bundles called cerebellar peduncles. The midline of the cerebellum forms the vermis and the larger portions on each side are the lateral hemispheres. The cerebellum receives sensory input from most of the cerebral cortex, which then projects back from the deep cerebellar nuclei by way of the brain stem and thalamus. The deep cerebellar nuclei consist of the dentate nuclei in the lateral hemispheres, the emboliform and globus nuclei, which are medial to the dentate, and the fastigial nuclei, which are the most medial of the nuclei.

dopaminergic neurons Most dopaminergic neurons are located in the midbrain, including the dorsal portions of the substantia nigra. Fiber pathways from the substantia nigra project to the caudate and putamen. In Parkinson's disease there is neuronal loss in the substantia nigra, which reduces activity in receptors that terminate in the caudate and the putamen.

frontal lobes The frontal lobes are located anterior to the central sulcus and consist of four major areas: the primary motor cortex, the premotor area, Broca's area, and the prefrontal cortex. General functions subserved by the frontal lobe included movement initiation, the production of language, executive processes involved in working memory, and aspects of personality.

functional imaging methods Scientific methods for investigating brain–behavior relationships. Two main functional imaging techniques are positron emission tomography (PET) and functional magnetic resonance imaging (fMRI). PET techniques involve intravenously injecting various short-lived radioactive agents and then imaging subjects in a specialized detector system to examine biochemical or physiological processes involved in cerebral blood flow and metabolism. Like PET, fMRI techniques are noninvasive and produce images of brain activation by detecting the indirect effects of neural activity on local blood volume, flow, and oxygenation. The most common fMRI technique for brain mapping involves blood oxygen dependent-level contrast that detects endogenous changes in local concentrations of paramagnetic deoxyhemoglobin associated with local increases/decrease in neural activity. Relative to PET, fMRI has superior temporal and spatial resolution.

lesion method A scientific method for investigating brain–behavior relationships by studying patients with focal brain lesions or neurodegenerative disorders that primarily affect a specific brain region.

long-term memory A repository where information can be stored for long periods of time and retrieved or recalled in accord with current goals.

parietal lobes The parietal lobes are located posterior to the central sulcus and are divided into the primary somatosensory, the

superior parietal, and the inferior parietal lobes. The inferior parietal lobe is further subdivided into the supramarginal gyrus and the angular gyrus. These structures are important for many functions, including processing sensory information, language comprehension, attention, and spatial perception.

prospective timing A term used to describe estimation of current time.

temporal lobes The temporal lobes are located inferior to the lateral sulcus and are divided into the superior, middle, and inferior temporal gyri. Various portions of the temporal lobe are important for memory, higher order processing of visual information, language comprehension, and audition.

timekeeper mechanisms Explicit control of prospective timing is thought to be implemented by a hypothetical timekeeper mechanism. A clock metaphor is used to describe the timekeeper, which represents subjective time through the accumulation of periodic pulses, possibly generated by oscillators.

working memory A short-term repository where information is temporarily stored before it is transferred to long-term memory. Working memory is composed of two main processes. Rehearsal processes repeat information to maintain it in working memory. Executive processes manipulate information in accord with current goals (e.g., comparing two intervals of time).

Psychologists' long fascination with the study of time is evident from the immense body of research that has grown over the years. In our daily life, we are so accustomed to experiencing time that frequently there is no conscious awareness of it. For example, the perception of time in the range of tenths of milliseconds is necessary for detecting formant changes in speech, but we have no awareness of these intervals. Longer intervals are more easily accessible to consciousness, but their processing differs depending on whether time perception is retrospective or prospective. Our memory for time in the past is thought to be estimated using information in the environment to reconstruct elapsed time. Therefore, for example, when reflecting on how much time has gone by since last seeing a close friend, one might use notable events such as a holiday or a place of encounter to reconstruct the passage of time. This contrasts with our subjective perception of the present in which events in the hundredths of milliseconds up to about 3 sec appear to require a central timekeeping or pacemaker mechanism. A clock metaphor is commonly used to describe the timekeeper mechanism, which represents subjective time through the accumulation of periodic pulses within a particular physical time. This article focuses on prospective timing of these relatively short intervals because they are thought to structure actions and our perception of objects and events, both of which can be abnormal in patients with central nervous system damage. In view of the extensive literature on temporal processing, this article is further confined to the study of time perception, except to briefly note that although support exists for analogous timekeeper mechanisms in movement, timing can also emerge from properties of movements, such as their force or trajectory.

I. INTRODUCTION

In this article, we explore the mechanisms by which the brain times events and stores them in memory for later use. This is of great interest to neuroscientists not only because there are a variety of neurological disorders in which the timing of behaviors appears abnormal (e.g., Parkinson's disease, Huntington's disease, and cerebellar damage) but also because insight into the underlying neural substrate(s) has important implications for understanding everyday perceptual and motor activities (e.g., playing a musical instrument and driving an automobile) that depend on timing precision. A prevailing model of temporal processing is first reviewed, along with a description of paradigms commonly used for investigating these phenomena. We then discuss the insights into the neural underpinnings of time perception provided by studies of patients with neurological disorders. Although significant contributions have been gained from this line of investigation, the emergence of functional neuroimaging technologies has offered a converging methodology, which may help clarify some of the issues in this area. This literature is discussed, including a recent study in which we charted the evolution of brain activation patterns during time perception using event-related functional magnetic resonance imaging (fMRI). Although important research has also been conducted using electrophysiological and pharmacological approaches, this relevant work is acknowledged only in brief due to necessary limitations in the scope of this article. Our current understanding of the neural systems that modulate time perception from converging lines of research will then be integrated into a framework that we hope will be useful for guiding future investigations.

II. COGNITIVE MODELS OF TEMPORAL PROCESSING

Many theories have been put forth to explain prospective timing, but most embody similar processes

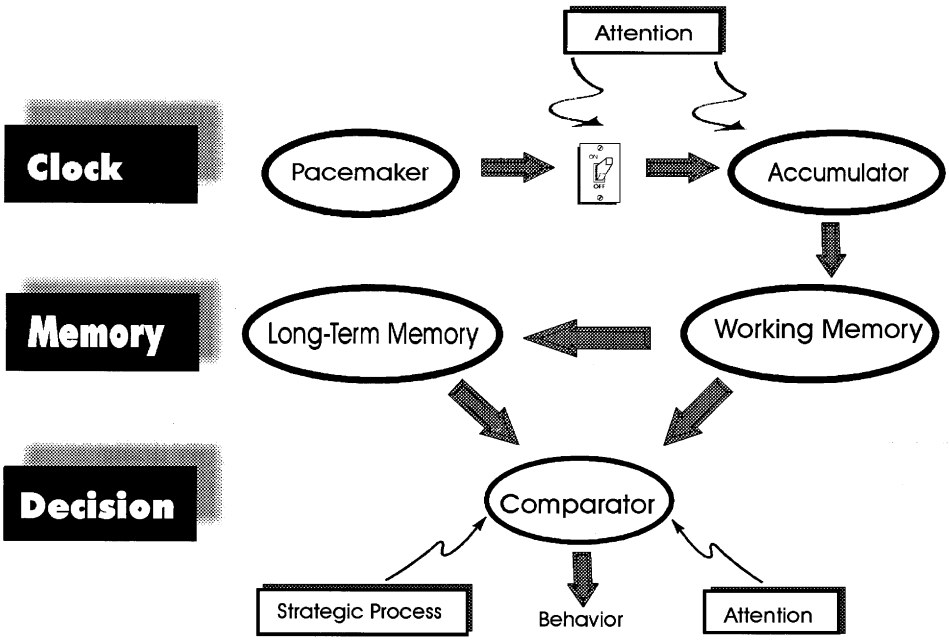


Figure 1 Diagram of the key components of temporal processing. The clock component consists of a pacemaker, which represents time through the accumulation of internal pulses. Pulses are turned on and off by a switch and then passed into an accumulator to be counted. Accumulated pulses are passed on or encoded into working memory, where interval representations are temporarily maintained and manipulated in accord with current goals. Over time, more enduring interval representations are stored in long-term memory. Decision process compares pulse counts from the accumulator with ones in memory to determine when or how to respond.

and there is a remarkable resemblance among them in their predictions. Figure 1 illustrates the key component processes of an influential theory of John Gibbon and colleagues, scalar timing theory, which was derived from an impressive series of animal studies. In this theory, central control of timing is implemented by a pacemaker mechanism that represents time through the accumulation of internal pulses, possibly generated by oscillators. Subjective perception of time, however, is intimately related to the interplay between the pacemaker and the level of attention given to the passage of time. Attention affects the switch that controls the flexible starting and stopping of pulses from the pacemaker, enabling anticipation of predictable events. The cost of this flexibility, however, is less accurate and more variable timing than is found, for example, with circadian timing mechanisms. Attention also appears to play a role in the accumulation of pulses because when attention is diverted there is a systematic shortening of subjective duration, suggesting that pulses from the pacemaker may be lost. Once a representation of time is formulated, it is passed on to working memory, which is a short-term repository in which interval representations are maintained and manipulated in accord with current goals. Over time, more enduring interval representations are stored in long-term memory. Subjective perception of time can therefore be altered in working or long-term memory. In some renditions of the model, the representation of time in memory is also determined by the rate by which the pulse count in the accumulator is transferred or encoded into memory. Finally, a decision process compares pulse counts from the accumulator, which represents current time, with ones in memory to determine when or how to respond. Strategic or attentional factors can influence decision processes by biasing response thresholds that determine when the current time is sufficiently close to the remembered time. The coupling of these different component processes gives rise to our perception of time.

The three component processes depicted in Fig. 1 have been studied using a variety of procedures, which vary in their ability to test the finer points of scalar timing theory. Figure 2a illustrates one widely used paradigm for investigating time perception—the comparison procedure. In this procedure, a standard interval to be timed is presented on each trial followed by a comparison interval, which the subject then judges as longer or shorter than the standard. Typically, subjects are not given feedback about their

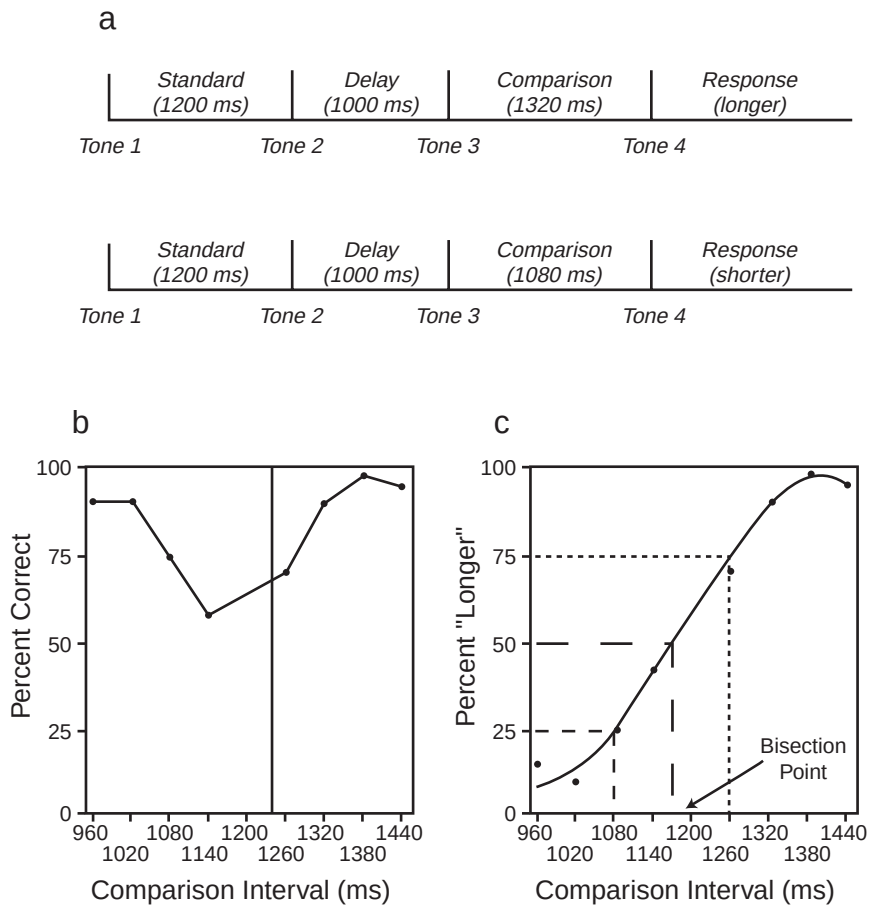


Figure 2 Diagram of a time perception paradigm and hypothetical results. (a) The time perception paradigm using the comparison procedure. In this procedure, a standard interval to be timed is presented on each trial followed by a comparison interval, which the subject then judges as longer or shorter than the standard. (b and c) Hypothetical data generated from this method. (b) The percentage correct responses as a function of the duration of the comparison interval. (c) The percentage of "longer" responses as a function of the duration of the comparison interval. The bisection point (dashed line) is derived through interpolation of the point at which approximately 50% of the trials are classified as longer. The difference limen is the difference between the intervals classified as longer in 75 and 25% of the trials (dotted lines) divided by 2.

response. Some procedures utilize a fixed set of comparison intervals wherein half are longer and half are shorter than the standard interval.

Figure 2b shows hypothetical data generated from this method. Here, a standard interval of 1200 msec is represented on the graph by a vertical line intersecting the *x* axis. The *y* axis designates the mean percentage correct responses, which are plotted for each of eight comparison intervals on the *x* axis. This graph illustrates a classic discrimination function wherein accuracy is worse the closer the comparison interval is to the standard. In this example, accuracy tends to be better for longer than shorter comparison intervals, which implies that the duration of the standard is underestimated. This is more evident in Fig. 2c, in which the same data are plotted as a function of the percentage of "longer" responses for each comparison interval. Here, time perception accuracy is designated by the bisection point, which is most directly derived through interpolation of the point at which approximately 50% of the trials are classified as longer. In this example, the bisection point is approximately 1170 msec, indicating underestimation of the 1200-msec standard interval. Asymmetrical discrimination functions are common because perception of the standard interval can be biased by various circumstances. For example, we tend to perceive an interval as being shorter than its actual physical duration when it is maintained for a longer period of time in working memory. Hence, as time passes our memory for a

standard interval will be more difficult to distinguish from shorter than longer comparison intervals. Another measure of time perception performance is the difference limen, which is the difference between the intervals classified as longer on 75 and 25% of the trials divided by 2. In the example in Fig. 2c, this would be calculated by subtracting 1080 msec from 1260 msec and then dividing this quantity by 2, which yields a difference limen of 90 msec. Larger values reflect greater variability in time perception performance. When the difference limen is divided by the bisection point, this produces a coefficient of variation, which represents processing sensitivity or efficiency. Finally, the precision or consistency by which intervals are judged over trials is reflected in the residual variance, which is the portion of variance unexplained by changes in the comparison interval.

A similar paradigm for studying time perception is the time bisection method. Here, subjects are first trained to discriminate two standard intervals, one short and one long. Once this is mastered, both standard intervals are randomly presented along with other intervals that fall in between the two standards. The subject's task is to classify intervals in terms of whether they are more similar to the short than the long standard or vice versa. There are many variations on this approach, including those in which subjects judge whether a comparison interval is longer, shorter, or the same as the standard. Unlike the comparison procedure, subjects must remember the standard intervals over trials such that subjective perception of time can be more influenced by working and long-term memory. Although the time bisection method generates data similar to those shown in Fig. 2c, in humans the bisection point is typically closer to the arithmetic than the geometrical mean of the two standard intervals. The bisection point is also sensitive to the spacing of intervals, which suggests that the difficulty of the discrimination task can influence accuracy.

Other variations on these procedures employ step-down approaches to estimate time perception acuity. This is commonly done in comparison paradigms by systematically reducing the difference between the standard and comparison intervals by small increments (e.g., 7 msec) to assess the least amount of time an individual can reliably perceive between two intervals. This method generates lower and upper thresholds for shorter and longer responses, respectively. The main measure of interest is the difference threshold, which is computed by subtracting the lower from the upper threshold and dividing the difference by 2. Difference thresholds computed from step-down methods are highly accurate measures of perceptual acuity for time, but the data do not easily lend themselves to computations of variability that are equivalent to those described previously (Fig. 2c). Rather, the variability of time representations is implicit in the difference threshold because the threshold estimate is set based on a predetermined number of standard deviation units (e.g., 1 standard deviation) from the interval at which an individual is equally likely to respond shorter or longer.

These and other methods reveal two fundamental properties that are thought to regulate the timing of behaviors by the internal clock. One property is mean accuracy, which is assumed to reflect a representation of time close to the actual physical time. The second is called the scalar property, which specifies that the standard deviation of time representations grows linearly with the mean of the interval being timed. This is a form of Weber's law and indicates that the precision of the timekeeper mechanism is constant irrespective of the duration being timed.

Scalar timing theory predicts that clock, memory, and decision processes should be represented by specific patterns of accuracy and variability. Perhaps some of the most compelling evidence for a distinction between clock and memory components comes from pharmacological investigations in animals conducted by Andres Maricq, Russell Church, Warren Meck, and colleagues. In these studies, animals were first pretrained to discriminate a standard interval(s) and then administered either a dopaminergic (DA) agonist (e.g., methamphetamine) or antagonist (e.g., haloperidol) to observe the effect on timing performance. Timing variability (coefficient of variation) remained scalar, but accuracy (e.g., bisection point) was abruptly altered. DA antagonists produced an *immediate* overestimation of time, which was attributed to the clock slowing down so that the pulse count from the pacemaker accumulated later in physical time relative to pretraining. In contrast, DA agonists produced an underestimation of time presumably due to the clock speeding up so that the pulse count accumulated earlier in physical time. Interestingly, when animals were given feedback about their timing performance, these distortions gradually disappeared over trials because the clock can be recalibrated to adjust a slower or a faster pulse count to the values stored in memory. This pattern of results contrasts with those produced by manipulating cholinergic activity, which is thought to affect memory processes. Cholinergic challenges alter timing accuracy and variability, and the effects are gradual because it takes time to establish new pulse

counts in memory. Cholinergic agonists (e.g., physostigmine) produce an underestimation of time and reduced variability presumably because they speed up the encoding of the pulse count into memory (Fig. 1). Cholinergic antagonists (e.g., atropine) have the opposite effect, causing an overestimation of intervals and increasing variability because encoding of the pulse count into memory is ostensibly slowed down. These effects are long-lasting because memory cannot be recalibrated quickly.

These findings are intriguing and suggest that scalar timing theory and other similar models may be conceptually useful for understanding temporal processing. Still, in most research it has proven difficult to isolate the independent effects of the clock from other component processes in Fig. 1 because they interact to bring about timed behaviors. Experimental manipulations of attention or working memory, pharmaceutical challenges, or the use of neurological patients can alter variability in one or more component processes. The reasons for changes in variability, however, are often difficult to determine because there are subtle differences in the behavioral patterns predicted for clock, memory, and decision processes. Still, this remains a prevailing model and has been a springboard for investigations into the neural underpinnings of temporal processing. The remainder of this article discusses investigations in this area that have used focal lesion and functional imaging methods in humans, which to date are the predominant neuroscience approaches. Although few of these studies are able to directly specify the source of behavioral changes resulting from manipulations of brain states (i.e., clock, memory, or decision processes), a picture of the distributed neural systems that give rise to precisely time behaviors is beginning to emerge. This work, combined with a preliminary understanding of the functional significance of different neural systems, extends current models and points to new approaches for dissecting the component processes underlying timed performance and their neural substrates.

III. LESION STUDIES OF TIME PERCEPTION

In humans, one approach to investigating the role of a neural site in a behavior is to study patients with focal brain damage as a consequence of a stroke, tumor resection, or neurodegenerative disorder. This approach can validate areas essential for a behavior by showing, for example, that damage to a particular site disrupts the clock but not memory or decision processes. This is a definite strength of the lesion method compared to others because there is a direct link between brain dysfunction and specific behavioral consequences of theoretical significance. In practice, however, caution must be taken in interpreting these studies because it is possible that a distant site interconnected with the damaged area is the neural system primarily responsible for sustaining a particular behavioral process. Additionally, inferences about brain–behavior relationships can be complicated by recovery of function and the fact that brain damage often affects multiple sites. For these reasons, it is important to cross-validate findings with converging methods.

Despite these cautionary remarks, the results from studies of time perception using the lesion method have been remarkably consistent with those derived from other methods. The basal ganglia and the cerebellum have been most studied because they are logical candidates for timekeeper operations. Damage to these brain regions commonly disrupts behaviors that depend on precise timing, such as rhythmic movements in Parkinson's disease (PD) and regulation of agonist–antagonist muscle activity (e.g., dysmetria) after cerebellar damage. Although these movement abnormalities could be attributed to a disruption of more generalized motor execution functions classically attributed to these systems, the basal ganglia and cerebellum mediate the perception of time as well. The role of interconnected cortical sites in temporal processing has been less well studied, but an emerging literature indicates that our perception of time is also regulated by various regions of the cerebral cortex.

A. The Role of the Basal Ganglia

The previously mentioned pharmacological studies in animals provided some of the first evidence for the regulation of timekeeper operations through DA neurotransmission in the striatum. Subsequently, Warren Meck showed that the physiological effect of DA agents on pacemaker operations was due to their binding affinity for dopamine D2-type receptors rather than type D1, noradrenergic, or serotoninergic receptors. These studies implicated the basal ganglia in timekeeper operations because DA receptors are abundant in the caudate, putamen, and substantia nigra (SN).

PD has been the dominant model of dopaminergic dysfunction in humans. In PD, dopaminergic neuronal loss in the SN reduces activity in receptors that

terminate in the caudate and the putamen, which increases inhibitory output from the striatum. The hallmark symptoms of PD include tremor, postural rigidity, and slowness in movement (bradykinesia). Additionally, the disorder is commonly associated with bradyphrenia or a slowing in cognition. One proposal is that cognitive slowing is due to the reduction in rhythmic neuronal firing of basal ganglia–thalamocortical circuits, causing a disruption in the timing of behaviors. This has been studied by comparing performance on timing tasks when PD patients are withdrawn from dopaminergic replacement therapy (''off'' medication) to their performance when therapy is reinstated (''on'' medication). Although the study of unmedicated patients offers a more sensitive appraisal of basal ganglia dysfunction, withdrawal of medication for sufficient periods of time is often difficult or impractical. For this reason, studying medicated patients has advantages because normal dopaminergic functioning is not fully restored by replacement therapy.

Several studies have shown that PD disrupts the timing of movements, but this could be due to the motoric functions traditionally ascribed to the basal ganglia. For this reason, time discrimination procedures have advantages because the motor requirements are minimal. Relatively few studies, however, have used these procedures. In 1989, Richard Ivry and Steven Keele were the first to report that time perception was normal in PD patients who were on medication. This was surprising given previous work in animals, and it raised the possibility that a sensitive assessment of basal ganglia functioning may not always be obtained in patients on medication. Julio Artieda, Maria Pastor, and colleagues reexamined this issue in 1992 using a temporal discrimination threshold (TDT) paradigm, which measured the least amount of time subjects could discriminate between two successive auditory, visual, or somaesthetic stimuli in order to perceive them as separate. They reported prolonged TDTs in PD patients tested off medication, irrespective of the stimulus modality. Moreover, reinstatement of dopaminergic replacement therapy substantially reduced TDTs, although whether performance actually returned to normal levels was not evaluated. These findings were not easily explained by possible primary sensory deficits because indices of such deficits, including sensory nerve action potentials from the wrist and cortical evoked potentials recorded over the parietal lobe, were normal in PD patients. Although the results implied that dopamine neurotransmission in the striatum regulated time perception competency, arguably this task may not be a pure measure of explicit timing because subjects made decisions about whether they perceived one or two stimuli rather than about the length of time between presentation of two stimuli. Additionally, impaired time perception in this study could be due to a disruption in other processes that interact with the timekeeper mechanism.

We attempted to address these and other concerns in a study of medicated PD patients using the comparison procedure. Our study employed a step-down approach, in which the difference threshold was the measure of perceptional acuity for time. Unlike previous studies using the comparison procedure, we employed two different standard intervals (300 and 600 msec) to assess whether time perception acuity remained scalar in patients. When variability is nonscalar, especially due to reduced processing efficiency for longer time intervals, this is sometimes attributed to the role of memory or decision processes in timing performance rather than to changes in clock speed, which are thought to alter timing accuracy in proportion to the interval being timed (i.e., scalar). However, because patterns of accuracy and variability are not understood well with respect to specific mechanisms, we also controlled for the effect of memory and decision processes on discrimination performance using a pitch perception task. In this task, the trial events were the same as in the time discrimination task, except that subjects judged the pitch of the comparison tone pair relative to the standard tone pair. We found that PD patients were impaired on the time (i.e., larger difference thresholds) but not the pitch perception task. Timing performance was also scalar, suggesting that timing sensitivity remained constant across the two standard intervals despite the elevated difference thresholds in PD patients. Moreover, these same patients were impaired on a converging measure of clock processes from a motor timing task. Together, these results suggest that timekeeper operations, rather than other cognitive processes required for the performance of temporal processing tasks, are disrupted in PD patients.

Although the previously mentioned studies support the basal ganglia's role in timekeeper operations, surprisingly few studies have directly examined whether basal ganglia dysfunction compromises processes that interact with timekeeper mechanisms to bring about representations of subjective time. To investigate this possibility, research will be required that analyzes patterns of performance as a function of manipulating memory, decision, and other processes

that may be regulated by the basal ganglia. For example, the time perception paradigms utilized in most human neuroscience studies minimize long-term and working memory demands so that their effect on performance may not be as evident as in other paradigms, such as the time bisection procedure. This avenue of research will be essential for more critical tests of the basal ganglia hypothesis of timing. It is also worth considering that diminished timing competency could also arise from the effect of basal ganglia dysfunction on interconnected cortical sites, some of which are strongly associated with working memory and attention functions. Specifically, the basal ganglia project via separate pathways of the thalamus to at least five prefrontal cortical areas: the supplementary motor area, premotor cortex, dorsolateral prefrontal cortex (DLPFC), anterior cingulate cortex, and the frontal eye fields. Fiber pathways also connect the basal ganglia with the parietal and the temporal cortex. Consequently, performance could be altered by abnormalities in the functioning of any of these systems due to reduced basal ganglia output.

B. The Role of the Cerebellum

Human lesion studies also support the role of the cerebellum in temporal processing. Patients with cerebellum damage due to a variety of etiologies, including cerebellar atrophy, soft signs of cerebellar damage (e.g., clumsy children), tumor resection, and stroke, have been studied. These cerebellar disorders can produce a variety of motor symptoms that impair gait (ataxia), accurate reaching to targets (dysmetria), postural stability, tremor, slowed or slurred speech (dysarthria), and eye movement abnormalities (nystagmus and ocular dysmetria). Interestingly, damage to focal structural regions of the cerebellum, such as the vermis or lateral hemispheres, does not correlate strongly with the manifestation of specific symptoms. Additionally, patients with cerebellar damage sometimes show cognitive deficits on clinical testing resembling those classically associated with cortical damage. This is consistent with mounting evidence from cognitive neuroscience investigations, which implicate the cerebellum in cognitive operations.

Many studies have documented disruptions in the timing of movements after cerebellar damage; however, as in PD, this could be attributed to the motor functions of this system. Nonetheless, several studies have reported that cerebellar damage also impairs performance on time perception tasks. In 1989, Richard Ivry and Steven Keele reported that time perception performance was impaired (i.e., larger difference threshold) in patients with cerebellar damage due to stroke, tumor, or degenerative disease. These patients also showed impairments in timing movements. In another study, these investigators reported that patients with focal damage to the lateral cerebellum and its primary output, the dentate nucleus, showed deficits in timing movements that appeared independent of potential motor implementation deficits. This contrasted with damage to the medial aspect of the cerebellum (vermis) that produced motor implementation problems, but the timing of movements was normal. Although these findings implicated the lateral cerebellum in timekeeper mechanisms of motor timing, a similar lesion analysis has not been conducted for time perception competency. Additionally, whether motor timing competency was scalar could not be assessed since only a single standard interval was used in these studies.

In 1996, Paolo Nichelli and colleagues studied time perception in patients with cerebellar degenerative disorders using the time bisection method, wherein subjects classified intervals as short or long after exposure to two standard intervals. Time perception was studied in four different conditions, each of which used different standard intervals (100 and 325 msec, 100 and 600 msec, 100 and 900 msec, and 8 and 32 sec) that varied in interval range. To control for possible deficits in discrimination processes unrelated to explicitly timing stimulus events, subjects also performed a spatial bisection task in which they classified horizontal lines as short or long relative to two previously practiced standard line lengths (6 and 54 mm). These investigators reported that spatial bisection performance was normal in patients with cerebellar degeneration, despite deficits in time perception in some but not all of the experiments. Specifically, processing efficiency was diminished (i.e., larger coefficient of variation) in cerebellar patients relative to the normal control subjects for the long-interval discrimination condition (8 and 32 sec). It was speculated that this finding was due to deficits in sustaining attention or the use of different strategies to maintain temporal information over long periods of time. Processing efficiency was also diminished in one of the short-interval discrimination conditions (100 and 600 msec), but in another condition (100–900 msec) only accuracy (bisection point) differed between the groups, such that patients underestimated intervals. The investigators concluded that the results for short-interval conditions supported the

cerebellum's role in clock processes. Although these findings were intriguing, the type of deficits (bisection point and coefficient of variation) in patients depended on the range of intervals used in each experiment in ways that were not necessarily predicted by scalar timing theory.

This issue was revisited in 1998 by Jennifer Mangels, Richard Ivry, and colleagues in a study of nine patients with focal lesions of the cerebellum. A step-down comparison procedure was adopted to estimate perceptional acuity for time using two different standard intervals (400 msec and 4 sec). Cerebellar patients showed impaired time perception (i.e., larger difference thresholds). Their deficits also appeared greater for the longer than the shorter standard interval, which might imply deficient working memory or attention. Alternatively, strategic processing deficits could also be an important factor if patients fail to employ normal strategies for sustaining information in working memory, especially over longer intervals. This latter possibility was investigated by examining the role of counting strategies in time perception because counting helps to divide longer intervals into a series of shorter ones, which improves timing accuracy. Subjects performed a time perception task with and without strategic support (i.e., a 4-sec interval was divided into 10 subintervals using auditory clicks) and with and without instructions to use the subintervals to count out time. Time perception acuity was impaired in the cerebellar patients irrespective of the condition and could not be attributed to counting deficits. Although the authors concluded that the cerebellum regulates clock processes, this is arguable because patients used explicit counting strategies, which involve timing, as efficiently as the control subjects to improve their performance. Moreover, patients with cerebellar damage also showed significant deficits on a pitch perception task, which raised the possibility that their time perception impairments might be explained by deficits in other processes that interact with pacemaking operations. A recent study by Laurence Casini and Richard Ivry indicates that working memory deficits do not explain timing impairments after cerebellum damage. In this study, seven patients with focal cerebellar damage and one with cerebellar atrophy showed impaired time perception (i.e., larger difference thresholds). However, performance deteriorated to the same extent in the healthy control and the cerebellar groups when performing under dual-task conditions, which places a greater demand on working memory. Nonetheless, pitch perception was still impaired in patients, raising the possibility that deficits in other undefined processes might explain impaired time perception after cerebellar damage.

The previous studies indicate that cerebellar damage usually disrupts time perception performance; however, the specificity of this system for explicit timing operations is questioned by findings of impairments in other types of discrimination that are not time dependent. More work is necessary to tease out the fundamental mechanisms of time perception deficiencies in these patients. Abnormalities could be due in part to a disruption in the functioning of cortical systems, given that reciprocal fiber pathways interconnect the cerebellum to most regions of the frontal, temporal, and parietal cortex. This is especially likely inasmuch as most research has studied cerebellar degenerative disorders, which are rarely focal and typically involve the cerebral cortex. For this reason, studies of patients with focal cerebellar lesions more directly assess the function(s) of the cerebellum. To critically evaluate the cerebellar timing hypothesis, it will also be important to further validate whether the lateral hemispheres and the vermis play functionally different roles in building representations of subjective time. Most functional imaging investigations of time perception challenge this proposal.

C. The Role of the Cerebral Cortex

Relatively limited attention has been given to the role of specific cortical systems in time perception, despite the multiple interconnections of the cerebral cortex with the basal ganglia and the cerebellum. Richard Ivry and Steven Keele first investigated the role of the frontal cortex in time perception in 1989. They employed a step-down comparison procedure to estimate perceptual acuity for time using a 400-msec standard interval. Eight patients with lesions in undefined areas of the frontal cortex were studied and these investigators reported normal time perception performance. Recent studies, however, indicate that the frontal cortex is involved in time perception processes. Paolo Nichelli and colleagues studied time perception in 11 patients with frontal lesions in undefined and heterogeneous regions of the frontal lobe using the time bisection task. Time perception was studied in short-(100–900 msec) and long-interval (8–32 sec) discrimination conditions. In both conditions, patients with frontal lobe damage were as accurate as controls in estimating intervals but processing efficiency (i.e., larger coefficient of variation) was diminished. Additionally, patients demonstrated less

precision or consistency for the long- but not the short-interval discrimination condition relative to the control subjects. Although this pattern of results could be attributed to impairments in many processes, the investigators speculated they were due to the role of the frontal lobes in memory or sustained attention since time perception deficits were more striking for long-interval discriminations.

The discrepant results from the previous two studies could be due to a variety of factors, including differences in patients (e.g., lesion location or size) or the use of time perception tasks that differed, especially in their memory requirements. However, recent studies using procedures similar to those of Richard Ivry and Steven Keele provide convincing support for the frontal lobe's role in working memory functions, including the short-term maintenance of intervals and executive processes involved in comparing intervals, both of which modify subjective representations of time. Jennifer Mangels and Richard Ivry used a step-down comparison procedure for estimating perceptual acuity for time in seven patients with frontal lobe damage. They reported impaired time perception acuity (i.e., larger difference threshold) when the standard and comparison intervals were separated by 4 sec but not 400 msec. Likewise, pitch perception acuity was impaired in these patients only when the standard and comparison pitches were separated by 4 sec. Thus, frontal lobe damage produced deficits related to maintaining information in working memory over longer periods of time, irrespective of whether the auditory events were explicitly timed. In another study, Laurence Casini and Richard Ivry reported that executive functions of working memory are also impaired in patients with frontal lobe damage. In their study, subjects made time or pitch discriminations in a comparison procedure that had a single- and a dual-task condition. Perceptual acuity for time and pitch deteriorated in the five patients with frontal lobe damage, whereas only time perception acuity declined under dual-task conditions in the normal control group. Hence, frontal lobe damage disrupts the division of attentional resources in working memory, irrespective of whether events are explicitly timed and even when the demands on working memory resources are minimal.

Although most of the previously mentioned studies implicate the frontal lobe in various functions of working memory during temporal processing, the specific frontal sites supporting these functions were not delineated. Importantly, the role of other cortical areas was not examined. In particular, the parietal cortex might be expected to support timing operations because of its interconnectivity with the basal ganglia and the cerebellum. The inferior parietal cortex has abundant, bilateral projections to the putamen and caudate nucleus in monkeys and there are also cerebellar–thalamic projections to the inferior and the superior parietal cortex. Moreover, damage to the right parietal cortex often produces neglect, a disorder in attention, which might interrupt communication between timekeeping and attention processes.

To address these issues, we investigated the role of the cerebral cortex in time perception in a study of a large sample of patients with unilateral cortical lesions in the left ($n = 19$ patients) or the right ($n = 20$ patients) hemisphere. Patients and control subjects were studied using time and pitch perception tasks (comparison procedure), and the interval separating the standard tone pairs was either 300 or 600 msec. We also investigated whether time perception competency correlated with an independent measure of attention to determine whether timing deficits were related to diminished attention. This task required subjects to switch their attention between one of two key presses when a preceding cue did not match the impending response. Right or left hemisphere damage compromised attention to a similar degree. Pitch perception was also impaired after left or right hemisphere damage, but the incidence of deficits was considerably higher in patients with left frontal lesions (67%) than in the other patient groups (20% or less), suggesting a left hemispheric bias for processing frequency information. When deficits in pitch perception were controlled by including only subjects with normal pitch perception, only the right hemisphere-damaged group showed time perception deficits (larger difference thresholds), which were greater for the 600- than the 300-msec interval condition (nonscalar). Poorer time perception acuity also correlated with more severe deficits in attention in the right but not the left hemisphere group, despite similar attention deficits in both groups. Interestingly, attentional capacity did not correlate with pitch perception acuity in any patient group, implying that its relationship to discrimination performance was more specific to explicitly timing events.

We then examined whether damage to specific regions of the cerebral cortex was associated with time perception deficits. Patients' lesions were reconstructed from MRI scans onto computerized templates, which portrayed the neuroanatomy for different horizontal (axial) sections of the brain. Next, the templates from each patient were overlaid to

examine the extent to which lesions overlapped in patients with and without time perception deficits. Figure 3 displays the lesion overlap in patients with damage primarily anterior or posterior to the central sulcus, who exhibited impaired or normal time perception. The left and right halves of the brain show the lesion overlap for patients with left and right hemisphere damage, respectively. The section displayed in Fig. 3 showed the greatest lesion overlap in patients with impaired time perception, and the lower right side of the figure gives the location of this section on a lateral view of the brain. All patients with right anterior lesions in the vicinity of the middle and the superior frontal gyri [Brodmann areas (BA) 8 and 46] exhibited time perception deficits. In contrast, time perception deficits were not found when this same region was damaged in the left hemisphere. All right hemisphere patients with posterior damage and time perception deficits had lesions involving the supramarginal (BA 40) and the angular gyrus (BA 39). Time perception was normal in three patients with right posterior lesions; one had occipital damage (not shown in Fig. 3) and although the other two had lesions involving BA 40, performance was borderline normal in one of these patients.

Together, these results implicate a right prefrontal–inferior parietal network in time perception, which is associated with the right hemisphere's bias for controlling attention. Moreover, our results suggest that the right inferior parietal cortex is not limited to spatial attention, as commonly reported, but also participates in attentional functions involved in formulating representations of time. Attention is intimately associated with working memory functions, which presumably are required to compare intervals in time perception tasks. Others have implicated the middle and superior prefrontal cortex in working memory, which may explain the behavioral basis for impaired time perception in patients with damage to this system. It is worth mentioning that we failed to uncover evidence of a common left hemisphere network for time perception when carefully inspecting lesion sites in patients with left hemisphere damage who exhibited both impaired time and pitch perception. Nonetheless,

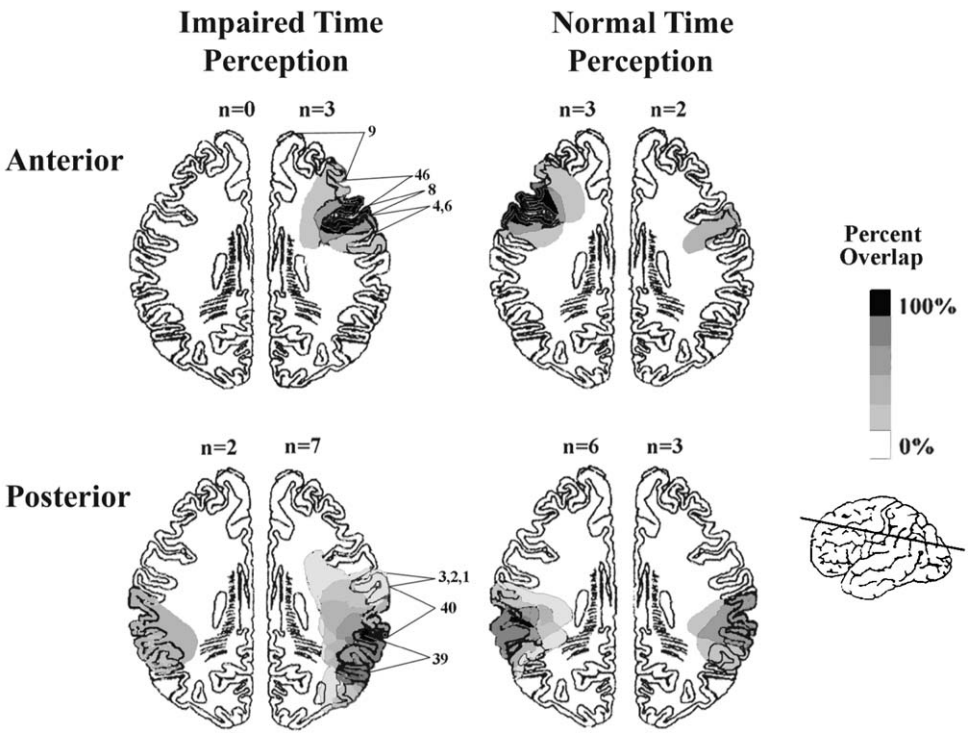


Figure 3 Lesion overlap in stroke patients with damage anterior or posterior to the central sulcus. The axial sections show the overlap for patients with left (left side of brain) and right (right side of brain) hemisphere damage with impaired and normal frequency perception. (Bottom right) The location of this section on a lateral view of the brain. The bar designates the percentage of patients with damage to a particular area.

it remains possible that the left hemisphere supports representations of time by way of other processes that were not identified in our study.

The previous work is consistent with clinical neuropsychological views of the functions of the frontal and the parietal lobes. This avenue of research, however, will be strengthened by further studies that directly manipulate variables presumed to alter clock, memory, and decision processes, with the goal of better delineating the functional significance of neural systems that regulate temporal processing. It also appears that there may be a right hemispheric bias for attending to temporal information, at least for intervals in the range of hundredths of milliseconds. Electrophysiological recordings in humans have also shown a right hemisphere bias for temporal processing, especially in the parietal cortex. In the next section, we discuss converging results from a recent functional imaging study that are in agreement with this proposal.

IV. FUNCTIONAL IMAGING INVESTIGATIONS INTO TIME PERCEPTION

Recently, the neural substrates of time perception have been studied using functional imaging tools in healthy individuals. Positron emission tomography (PET) and fMRI are widely used methods for examining hemodynamic changes in blood oxygenation, flow, and volume associated with regional increases or decreases in neural activity in the brain. PET scanning involves injection of a short-lived isotope produced in a cyclotron. PET scanners, which have rings of crystal detectors arranged around a central bore, create images by identifying areas of increased radioactivity in the brain. The most commonly used isotope for conducting brain mapping experiments, ^{15}O-water, has a relatively short half-life (2.1 min) and provides a measure of both global and local changes in cerebral blood flow. ^{15}O-water is injected into a subject while he or she performs a task for approximately 40–60 sec. In a typical brain mapping experiment, a subject may receive 4–10 injections per session. Subtraction techniques are used to compare images obtained in response to contrasting tasks in order to identify brain regions critically involved in each task.

fMRI comprises a wide range of techniques for imaging brain function. The most commonly used technique for brain mapping experiments, referred to as blood oxygen level-dependent (BOLD) imaging, involves a high-speed, T2-weighted echo-planar pulse sequence capable of generating an image in 40–100 msec. The endogenous BOLD technique involves collection of a series of images to detect decreases in local concentration of paramagnetic deoxyhemoglobin, reflected in an increase in magnetic resonance signal intensity. These signal changes take approximately 5–7 sec to reach a peak; with cessation of the activation task, the signal returns to baseline in 7 or 8 sec. In addition to not having to inject agents, BOLD imaging offers advantages in temporal and spatial resolution relative to PET imaging. Whereas PET experiments typically examine the cumulative activity associated with a block of trials administered over a 40- to 60-sec interval, BOLD experiments can examine changes associated with a single trial as brief as 1 sec and, as shown later, can image the evolution of brain activity in trials that span several seconds. With spatial resolutions that can approach a cubic millimeter, BOLD imaging, unlike PET, can pinpoint activity resulting from relatively small subcortical structures, such as distinguishing between putamen versus globus pallidus activation in the basal ganglia.

The first PET experiment to examine time perception was conducted by Jueptner and colleagues in 1995. Three conditions were imaged: a relatively easy temporal comparison procedure (standard tone pair interval = 300 msec; test tone pair = 200 or 400 msec), a control condition with two isochronous tone pairs, and a rest condition. Greater activation in time perception task was observed in the superior parts of the lateral cerebellum bilaterally and in the vermis when compared to the control condition. The authors argued that their data supported a cerebellar timing process, despite the fact that greater activation in the basal ganglia (right caudate and bilateral putamen/globus pallidus), bilateral thalamus, and cingulate cortex (BA 24) was also observed in the timing condition. It should also be noted that this study was limited by the lack of an active comparison control task (e.g., pitch perception). The low-level sensorimotor control task cannot determine whether the activated structures were specific to time perception or were associated with more generalized encoding, memory, and decision processes involved in any discrimination task. Finally, the limited coverage of the PET camera was focused on examining activity in the cerebellum; as a result, potential activation of superior cortical structures (dorsolateral prefrontal and parietal regions) was not measured.

Two PET studies by Maquet, Lejuene, and colleagues were subsequently reported in 1996 and 1997. In

the 1996 study, three conditions were contrasted: (i) a visual time-bisection procedure with a 700-msec standard and 490- to 910-msec comparisons, (ii) a visual intensity bisection procedure, and (iii) a sensorimotor control task. The contrast comparing the first and third conditions resulted in greater activation in the time discrimination task in the right DLPFC (BA 44, 45, and 47), right inferior parietal (BA 40), anterior cingulate areas, and the cerebellum (vermis and left lateral cerebellar hemisphere). However, no significant differences were observed between the first (time bisection) and second (visual intensity) conditions, suggesting that time and visual discriminations were supported by similar neural systems. In the 1997 study, subjects were required to judge the duration (2.7-sec interval) separating successive illumination of a light stimulus. The sensorimotor control task consisted of responding to a light stimulus that appeared on average every 2.7 sec (range, 2–3.4 sec). Four regions demonstrated greater activation in the duration than control task: right insula/inferior prefrontal (BA 47), right inferior parietal (BA 40), anterior cingulate (BA 32), and the cerebellar vermis. The authors also performed a meta-analysis that examined conjointly activated regions associated with both PET studies. In addition to the regions noted previously, they also observed activation in the right DLPFC (BA 46) and the left putamen.

Taken together, these three PET studies provide evidence for cerebellar and basal ganglia involvement in time perception. In addition, primarily right-sided cortical activations were observed in the middle and inferior prefrontal and inferior parietal regions. Unfortunately, the temporal resolution of PET scanning is limited to blocked-trial designs that cannot disentangle timekeeper and attention operations from working memory and response implementation processes. In a recently completed study, we exploited the finer temporal resolution of event-related fMRI to study the time course of brain activation patterns associated with different stages of a time perception task. Specifically, our study was designed to isolate patterns of brain activation that correlated with encoding time intervals from those associated with comparing two time intervals and implementing a response. Timing theory suggests that activation should develop earlier in systems integrally involved in encoding or formulating a representation of time (i.e., timekeeper and attention operations) and later in systems concerned with manipulating information in working memory (i.e., comparing intervals) and implementing a response.

fMRI scans were obtained for three different tasks. In the time discrimination condition, two tones (50 msec) separated by 1200 msec (standard tone pair) were presented, followed by a 1 sec delay and then a comparison tone pair. Subjects indicated whether the comparison tone pair was longer or shorter than the standard. To better separate neural systems specific to timing operations, subjects also performed a pitch discrimination condition in which the auditory events were similar except that subjects indicated whether the fourth tone was higher or lower in pitch than the first three tones. Imaging runs in each discrimination condition were contrasted with a sensorimotor control condition in which subjects responded after the presentation of two isochronous tone pairs of identical pitch. Figure 4 (top) illustrates the temporal relationship among the trial events, the acquisition of images, and the expected hemodynamic response function. The first scan was acquired at the onset of the first tone (T1). The fourth tone (T4) was presented an average of 3.4 sec post-trial onset. The typical key press response occurred 4.5 sec post-trial onset. Idealized early and late MR signal changes pegged to T1 and T4 are modeled using a gamma variate distribution to allow for the delay in the rise and fall times of the hemodynamic response. The time and pitch conditions were compared to the control condition at each of four post-trial onset scanning intervals or epochs (2.5, 5.0, 7.5, and 10.0 sec). Allowing 5 sec for the hemodynamic response to peak, we hypothesized that the 2.5- and 5.0-sec post-trial onset intervals should reveal brain activation patterns specific to encoding time intervals. In contrast, the 10.0-sec scanning interval should also include activations associated with contrasting the standard and comparison intervals and implementing the response. Overlap between these processes should be particularly evident during the 7.5-sec scan.

Figure 4 (bottom) shows that time discriminations produced unique and relatively early subcortical activations within the right putamen (2.5 sec), the head of the caudate nucleus bilaterally (5.0 and 7.5 sec), and the right centromedian and ventroanterior thalamic nuclei (5.0 and 7.5 sec). In contrast, cerebellar activation, which was located in the posterior vermis (tuber) of lobule VII, occurred relatively late (7.5 and 10.0 sec) in the time discrimination condition. During the first two epochs (2.5 and 5.0 sec), cortical foci specific to the time condition included the right intraparietal sulcus (IPS; BA 40), right inferior frontal gyrus (BA 44/45), bilateral dorsal and left ventral premotor (BA 6; not shown in Fig. 4

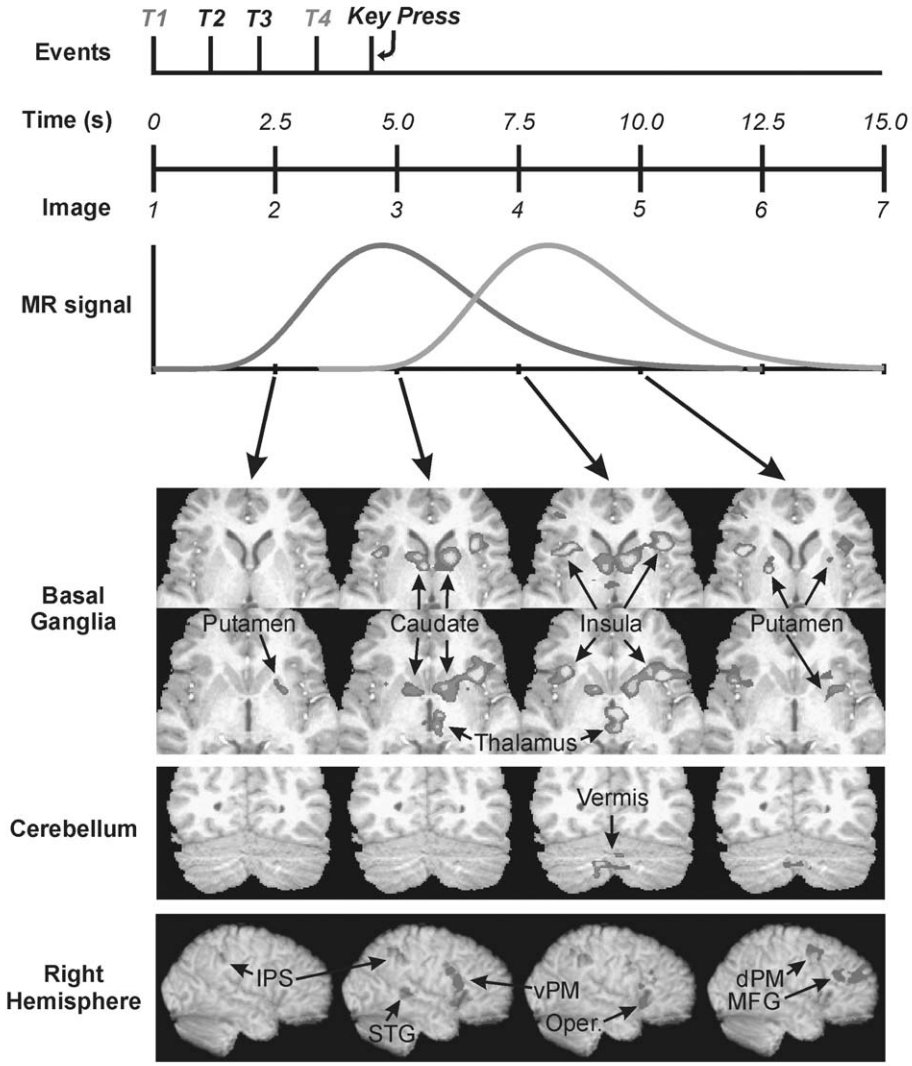


Figure 4 Illustration of the temporal relationship among the trial events, the acquisition of images, the expected hemodynamic response function, and the patterns of brain activation. (Top) The first scan was acquired at the onset of the first tone (T1). The fourth tone (T4) was presented an average of 3.4 sec post-trial onset. The typical key press response occurred 4.5 sec post-trial onset. (Middle) Idealized early and late MR signal changes pegged to T1 (left function) and T4 (right function) are modeled using a gamma variate distribution to allow for the delay in the rise and fall times of the hemodynamic response. These functions indicate that the 2.5- and 5.0-sec scans are influenced by the encoding of the standard interval, whereas the 10.0-sec scan reflects decision and response preparation processes. Overlap between these processes is evident during the 7.5-sec scan. (Bottom) The patterns of activation that were unique to the time perception task. Significant activation foci ($p < 0.001$) indicate greater activation for the time and pitch discrimination conditions relative to the control condition at 2.5, 5.0, 7.5, and 10.0 sec post-trial onset. dPM, dorsal premotor cortex; IPS, intraparietal sulcus; MFG, middle frontal gyrus; Oper, frontal operculum; STG, superior temporal gyrus; vPM, ventral premotor cortex.

for the 2.5-sec epoch), and bilateral lateral temporal (BA 21 and 22) regions. No such foci were observed in the pitch condition. Unique activations of the right DLPFC and right dorsal premotor areas also occurred in response to the time condition during the 7.5- and 10.0-sec epochs. Common activation foci (time and pitch conditions) were observed within the left hemisphere and included the inferior frontal gyrus (BA 44/45), IPS (BA 40), superior parietal lobule/precuneus (BA 7), and DLPFC (BA 9, 10, and 46). In addition, common areas were observed in the anterior supplementary motor area (pre-SMA; area 6), anterior cingulate (BA 32), and anterior insula/frontal operculum (BA 47).

From this fMRI investigation, we concluded that early sustained activation in the bilateral basal ganglia and right inferior parietal cortex during the encoding of time intervals suggests that these systems may regulate the hypothetical timekeeper and attentional operations put forth by theories of timing. Activation was also observed in regions commonly associated with various working memory functions, including the right premotor cortex and right DLPFC. However, the time course of activation in these two systems was remarkably different. This finding suggests that the two regions serve different purposes in temporarily storing time information. PET and fMRI studies have implicated the premotor cortex in the temporary maintenance of information, whereas the DLPFC subserves comparative processes that involve manipulation of information in working memory. Contrary to Ivry's theory, the lateral cerebellum was not associated with temporal processing. We did observe relatively late activity specific to the temporal processing condition in the vermis, suggesting that this structure is not critical to timekeeper or attentional factors associated with time perception. This study, as well as the three previously mentional PET studies, provides additional support for a right hemispheric bias for attending to temporal information.

Clearly, more research is needed to directly specify the proposed brain–behavior relationships we suggest account for our fMRI findings. It is also important to note that the interpretation of functional imaging results is often indirect, frequently relying on findings from other methods (e.g., lesion, pharmacological, and electrophysiological), in which brain function is more directly associated with behavioral measures that represent different hypothetical processes. Functional imaging approaches also currently depend heavily on the subtractive logic, wherein the subtraction of two tasks is understood to reflect brain activity unique to a particular mental operation. Unfortunately, this assumption is often not tenable or difficult to test, such that interpretations of causal linkages between brain structure and behavior may not be valid. Hence, findings from functional imaging approaches need to be weighed and verified by studies using other converging methods.

V. CONCLUSIONS

An understanding of the neural systems that give rise to our perception of time is beginning to emerge from converging lines of research. There is mounting evidence from lesion, pharmacological, and functional imaging studies for the basal ganglia's role in regulating hypothetical timekeeper operations. It also appears that this system supports prospective timing of intervals as short as hundredths of milliseconds and as long as many seconds, although more work is clearly needed to verify this conclusion. Although timing functions have been attributed to the cerebellum in some very influential work, the findings are weakened by reports that the cerebellum also participates in discrimination processes and working memory functions that do not depend on explicit prospective judgments of time. Likewise, our fMRI study suggests that the cerebellum is activated at a point in time closely correlated with selecting a response. Taken together, these findings are compatible with the proposal that the cerebellum is involved in monitoring and adjusting cognitive and sensory input from the entire cerebral cortex rather than performing a specific mental operation per se.

The right inferior parietal cortex is also intimately involved in regulating our perception of the passage of time. The strong association of this system with attention functions in lesion studies, combined with coupled activation in the basal ganglia, raises the prospect that this region may be crucial for controlling the flexible starting and stopping of pulses from the hypothetical timekeeper mechanism. There is also consensus that the frontal cortex is important for temporal processing. The specific role it plays needs clarification because different regions likely sustain different processes. Some insight into this issue was provided by our finding that the time course of activation differed in the premotor cortex and the right DLPFC during time discriminations. Based on other work, we speculated that this may be due to the role of the premotor cortex in maintaining an active representation of the time interval in working memory and the role of the DLPFC in manipulating or comparing interval representations in working memory. Other functional imaging work has reported activation of the inferior–prefrontal superior–temporal network during temporal processing. However, we and others have also shown that this network is activated during pitch discriminations, silent rehearsal of letter strings, and auditory imagery. Hence, this system appears to support a more general function, perhaps in the retrieval and rehearsal of auditory information. This is compatible with our lesion work showing that damage to this network does not disrupt time perception competency, indicating that it is not

essential for formulating or maintaining representations of subjective time.

It is particularly striking that lesion, electrophysiological, and fMRI studies report a right hemispheric bias for processing current time. This is not to say that the left hemisphere does not participate in prospective timing. We found activation in left frontal, temporal, and parietal regions during time perception in our fMRI study, but similar regions were also activated during pitch discriminations. Likewise, both time and pitch perception performance were abnormal after left hemisphere damage. Thus, unidentified processes common to both forms of discriminations appear to be biased for left hemispheric processing.

The previously mentioned research illustrates the many potential ways in which neural systems might come to regulate different component processes implicated in prospective timing. Nevertheless, more research is needed to directly test some of the ideas put forth in this article. Differentiating the reasons for impaired timing in focal lesion or functional imaging studies will require more refined paradigms that are capable of isolating clock mechanisms from those involved in working memory or decision processes. This endeavor should facilitate an understanding of how timing accuracy and variability are reflective of manipulations in some but not all component processes of time perception. This approach combined with investigations of the time course of brain activation should prove enormously beneficial in furthering an understanding of how the brain processes time.

See Also the Following Articles

ATTENTION • BASAL GANGLIA • BRAIN LESIONS • CIRCADIAN RHYTHMS • MEMORY, OVERVIEW • MUSIC AND THE BRAIN • MOTION PROCESSING • TEMPORAL LOBES • WORKING MEMORY

Acknowledgments

This article was supported in part by grants from the Veterans Affairs Medical Center, the MIND Institute (6), the National Institutes of Health (P01 MH51358, R01 MH57836, and M01 RR00058), and the W. M. Keck Foundation. The authors thank Gabrielle Mallory for her technical assistance in preparing the manuscript.

Suggested Reading

Block, R. A., and Zakay, D. (1997). Prospective and retrospective duration judgments: A meta-analytic review. *Psychonomic Bull. Rev.* **4,** 184–197.

Gibbon, J., Church, R. M., and Meck, W. H. (1984). Scalar timing in memory. *Ann. N. Y. Acad. Sci.* **423,** 52–77.

Gibbon, J., Malapani, C., Dale, C. L., and Gallistel, C. R. (1997). Toward a neurobiology of temporal cognition: Advances and challenges. *Curr. Opin. Neurobiol.* **7,** 170–184.

Harrington, D. L., and Haaland, K. Y. (1999). Neural underpinnings of temporal processing: A review of focal lesion, pharmacological, and functional imaging research. *Rev. Neurosci.* **10,** 91–116.

Harrington, D. L., Haaland, K. Y., and Knight, R. T. (1998). Cortical networks underlying mechanisms of time perception. *J. Neurosci.* **18,** 1085–1095.

Ivry, R. B. (1996). The representation of temporal information in perception and motor control. *Curr. Opin. Neurobiol.* **6,** 851–857.

Ivry, R. B., and Keele, S. W. (1989). Timing functions of the cerebellum. *J. Cogn. Neurosci.* **1,** 136–152.

Meck, W. H. (1984). Attentional bias between modalities: Effect on the internal clock, memory, and decision stages used in animal time discrimination. *Ann. N. Y. Acad. Sci.* **423,** 528–541.

Meck, W. H. (1996). Neuropharmacology of timing and time perception. *Cogni. Brain Res.* **3,** 227–242.

Rao, S. M., Mayer A. R., and Harrington, D. L. (2001). The evolution of brain activation during temporal processing. *Nature Neurosci.*, **4,** 317–323.

Wearden, A. J. (1999). "Beyond the fields we know…": Exploring and developing scalar timing theory. *Behav. Processes* **45,** 3–21.

Wittmann, M. (1999). Time perception and temporal processing levels of the brain. *Chronobiol. Int.* **16,** 17–32.

Tourette Syndrome and Obsessive–Compulsive Disorder

VALSAMMA EAPEN
United Arab Emirates University and Royal Free and University College Medical School, London

MARY M. ROBERTSON
University College, London

GLOSSARY

dopamine An amine neurotransmitter; the dopaminergic system is the pathway.

genotype Genetic makeup of an individual.

noradrenalin A catecholamine neurotransmitter; the noradrenergic system is the pathway.

perfusion The amount of blood that is used to supply a certain area of the brain per unit of time in order to supply oxygen, glucose, etc. Hypoperfusion refers to a decrease in perfusion, whereas hyperperfusion refers to an increase in perfusion.

phenotype Physical expression or observable characteristics of the individual. The striatum and pallidum are parts of the basal ganglia. The thalamus, limbic system, anterior cingulate, medial temporal, prefrontal, and midbrain are all specific areas of the brain.

serotonin (5-hydroxytryptamine) An amine neurotransmitter; the serotonergic system is the pathway.

Gilles de la Tourette syndrome (TS) is a movement disorder characterized by multiple motor tics (involuntary muscle twitches) and one or more vocal tics (involuntary noises) that have been present for more than a year, with age of onset before 18 years. The *International Classification of Diseases*, 10th version, and the *Diagnostic and Statistical Manual of Mental Disorders*, fourth edition (*DSM-IV*), define obsessive–compulsive disorder (OCD) as a condition characterized by recurrent obsessions or compulsions that cause marked distress, are time-consuming (take more than 1 hr a day), or significantly interfere with the person's normal routine functioning, social activities, or relationships. Obsessions are recurrent, intrusive, and unwelcome ideas, thoughts, images, or impulses, which are recognized by the individual as absurd. Compulsions are repetitive purposeful behaviors performed in response to an obsession. The diagnosis of OCD as per *DSM-IV* also incorporates an additional criteria that "at some point during the course of the disorder, the person has recognized that the obsessions or compulsions are excessive or unreasonable; however, this does not apply to children." The occurrence of OCD in the context of TS has received a great deal of attention and research in the recent past. Although there is increasing evidence from epidemiological, phenomenological, neurobiological, and genetic research about the association between TS and obsessive–compulsive behaviors (OCB), the exact nature of such an association is still debated.

I. HISTORY

The occurrence of obsessive–compulsive symptoms in the context of TS was first recognized by Gilles de la Tourette. While describing the case of Marquise de Dampierre, a French noblewoman, he reported that in addition to the tics and vocalizations, obsessive thoughts tormented her. Subsequently, Charcot identified involuntary "impulsive" ideas, such as arithmomania (counting obsessions), doubting, checking, and touching in the context of TS. Subsequent writings by other authors referred to the presence of obsessions as an accompaniment of the tic disorder, representing psychical tics. These early clinicians documented the psychopathology of people of "convulsive tic disorder" with special reference to OCB, such as checking, arithmomania, folie du doute, delire de toucher (forced touching), and folie du pourqoui (the irresistible habit of seeking explanations for the most commonplace insignificant facts by asking perpetual questions). It has been speculated with some conviction that Dr. Samuel Johnson, the 18th century literary figure, had TS, and in a book on the life of Dr. Johnson it was noted that "he often had convulsive starts and odd gesticulations." Furthermore, Johnson felt compelled to measure his footsteps, perform complex gestures when he crossed the threshold of a door, and involuntarily touch specific objects, thus suggesting that he also had OCB.

In the early 20th century, experts in the field began to note that "the frequency with which obsessions, or at least a proclivity for them, and tics are associated, cannot be a simple coincidence" and acknowledged a relationship between tics and OCD: "No feature is more prominent in tics than its irresistibility.... The element of compulsion links the condition intimately to the vast group of obsessions and fixed ideas." Robertson and colleagues documented in detail the historical aspects of the link between TS and OCD.

II. EPIDEMIOLOGY

TS is no longer considered to be rare. The accepted prevalence rate for some time has been 0.5/1000, but recent studies have suggested that it affects between 1 and 8 of every 1000 boys and between 0.1 and 4 of every 1000 girls. Even rates of up to 3% have been reported, with higher rates reported in children with special educational needs (approximately 25%) and autism (approximately 6%). Similarly, OCD was considered an uncommon psychiatric disorder until recently, but it is now recognized to be a more common psychiatric illness affecting approximately 1–3% of population. To date, nine population surveys have been conducted using the Diagnostic Interview Schedule and the lifetime risk of OCD has been estimated to be 1.9–3.2%, with a 6-month prevalence rate of 0.7–2.1%; furthermore, the British National Survey of Psychiatric Morbidity gave a 1-month prevalence of 1% in males and 1.5% in females. Other community-based studies have noted that the prevalence estimates are similar for males and females, whites and blacks, and across different socioeconomic status. However, the central issue in epidemiological research in OCD is that of case definition. Even if there were a reasonable consensus regarding case definition, difficulties would remain because of comorbid conditions, such as generalized anxiety disorder, depressive disorder, and TS. This is further complicated by the lack of consensus as to whether a separate diagnosis of OCD is justified in these situations. Although different solutions have been proposed, none has gained sufficient acceptance to dispel ambiguity from epidemiological studies.

Robertson and associates, reviewing the prevalence of OCB in the context of TS, mention that high rates of OCB have been reported by several investigators, with the rates ranging from 11 to 90%. It has also been shown that these symptoms vary based on the severity of TS; the frequency of echolalia has been reported to be 9.3% in grade 1 (mild) TS and 48.3% in grade 3 (severe) TS. It has been suggested that OC symptoms are more pronounced and severe in older TS patients, and in a study of 30 patients with TS older than 21 years of age, 27 patients (90%) were found to have OCD. Other investigators, however, have failed to find any relationship between age and OC symptomatology. Thus, the variations in the reported frequencies of OC symptoms in TS may well be due to a bias in sample selection.

Earlier empirical studies in the 1960s based on clinic samples reported that 12–35% of patients with TS also had OCB. However, these estimates have been considered to be underestimates since the absence of a report of symptoms could not be equated with the absence of the symptoms. Higher rates of 55–80% were reported in the 1980s, with one study documenting that 37% of 90 TS patients not only reported OC behavior but also obtained higher scores on standardized rating scales than normals. In the early 1990s, Robertson and colleagues interviewed 85 members of a multiply affected TS family and reported that distinction between cases and noncases could be made on the

basis of OC features and the trait score of the Leyton Obsessional Inventory. Similarly, in an interview of an Arab family, 4 TS cases were found, all of whom had OCB. In a study in a New Zealand cohort of 40 TS cases, 20 of whom were evaluated in detail, half of the cases were found to have OCB. Pauls and associates also reported that OCB occurs in approximately 40% of TS cases.

In this context, only few studies have included control groups. Fifty-two percent of TS patients in one study were found to have OCB compared to 12.2% of the controls; another study found 45% of TS patients had OCB compared to 8.5% of controls. In yet another controlled study, TS patients were found to be disproportionately obsessional, which was not accounted for by depression. Similarly, it has been documented that TS patients have significantly more obsessionality than controls. Given that the population prevalence of OCD is between 1 and 3%, the prevalence of OCD in TS patients seems to be much greater than is expected by chance.

While examining epidemiological evidence, it is equally important to consider the relationship from the other perspective, i.e., the occurrence of tics and TS among OCD patients. In this regard, open studies in the 1980s reported a rate of 5% for TS among OCD patients, whereas in a controlled study of 16 TS, 16 OCD, and 16 normal subjects, 37% of the OCD subjects were found to have a tic disorder compared to only 1 subject from the normal controls. In the 1990s, an epidemiologic sample of 861 Israeli adolescents found 40 individuals with OC spectrum disorders. After reviewing the literature on OCD, Rapoport and colleagues suggested that approximately 20–25% of OCD patients have tics.

III. PHENOMENOLOGY

It has been observed that TS and OCD share some clinical features, such as waxing and waning of symptoms, early age of onset, lifelong course, egodystonic behavior, and worsening of symptoms with stress and anxiety. However, despite some overlap in clinical features, available evidence suggests that the two conditions are not phenomenologically identical and that OCD may well be a heterogeneous entity, with one subtype related to TS.

Several investigators have tried to identify OC symptom subtypes in an attempt to find homogeneous subgroups of OCD patients. In this regard, four factors have been described using the Maudsley Obsessive Compulsive Inventory (MOCI): checking, cleaning, slowness, and doubting in descending order of variance. However, use of the MOCI has an inherent problem of item selection bias in that although checking (9 items) and cleaning (10 items) symptoms are well represented, others, such as aggressive obsessions (2 items), symmetry, ordering, and hoarding compulsions (no items), are underrepresented. In another cluster analysis of 410 OCD patients assessed using a checklist with similar symptom selection bias as the MOCI, five symptom subgroups were reported; checking, washing, the past, death, and sex.

The Yale–Brown Obsessive Compulsive Scale (YBOCS) symptom checklist overcomes the symptom selection bias of the MOCI and provides a comprehensive list of more than 50 obsessions and compulsions, comprising 15 general symptom categories. Using this checklist, it has been found that in a series of 250 OCD patients, 60% had multiple obsessions, whereas 48% had multiple compulsions. Another study using the same instrument assessed 79 children and adolescents with OCD and found that 47% had both washing and checking compulsions at some time and that none maintained the same symptom constellation at follow-up 2–7 years later. This calls into question the validity of categorizing patients into mutually exclusive subgroups, such as cleaners or checkers, based on symptoms at a given point.

It has also been reported that nearly 80% of the patients could be correctly classified as having TS or primary OCD diagnosis based on two items of the Hamburg obsessive–compulsive inventory, namely "fearful obsessive thoughts" found in 90% of OCD patients and echophenomenon present in 56% of TS patients. Differences between OCD and TS patients were evident on the MOCI subscales checking and slowness/repetition as well as on the total scores. Although TS patients scored higher than controls, they reported fewer symptoms when compared to OCD patients.

Although several studies have attempted to define the precise phenomenology of OC symptoms in TS, few have attempted to compare them to the symptomatology that occurs in OCD patients without a tic disorder. It has been documented that certain kinds of compulsions, such as touching and symmetry behaviors, occurred more often in TS patients than in the OCD group and that TS patients had more counting compulsions and impulses to hurt themselves, whereas OCD patients had compulsions to do things in a specified order and to arrange things; other symptoms,

however, showed an overlap between the two disorders. These clinicians also noted that the symptoms changed with age, with the younger patients exhibiting compulsive behaviors related to impulse control, whereas the older patients had more checking and arranging compulsions. Furthermore, it was suggested that the frequency of OC symptoms increases with the duration of TS. The role of gender and comorbid OCD on the phenomenology of TS has also been studied, and it was found that probands with OCD were more likely than those without OCD to onset with complex tics, and that females onset with compulsive tics more often than do males. In a study of TS children and adolescents with TS, it was observed that the presence of comorbid obsessions contributed to the prediction of learning problems, perfectionism, and antisocial behavior, whereas compulsions contributed to the prediction of hyperactivity, psychosomatic symptoms, and muscular tension.

Robertson and associates noted that coprolalia and echolalia were significantly related to OCB and that self-injurious behaviors in TS were associated with obsessionality. However, nonobscene complex socially inappropriate behaviors occurring in TS were not found to be associated with obsessions or compulsions. In another study by the same group comparing 10 OCD patients to 15 TS patients with comorbid OCD, it was found that violent, sexual, and symmetrical obsessions, as well as forced touching, counting, and self-damaging compulsions, were more common in comorbid OCD/TS subjects. On the other hand, obsessions concerning dirt or germs and cleaning compulsions were more commonly encountered in OCD subjects. Furthermore, the TS group reported that the compulsions occurred spontaneously or *de novo*, whereas in the OCD group this was preceded by guilt or worry. Other studies comparing TS patients with and without OCB found that, in contrast to OCB-free TS patients, TS + OCB patients had a higher incidence of volatile temper, compulsive tics, perinatal disorders, and brain wave changes, as well as a higher prevalence of developmental disorders and a higher severity of TS. They suggested that TS + OCB may be more strongly associated with organic cerebral disorders.

Eapen and colleagues compared the distribution of OC symptoms in 16 patients with OCD and 16 patients with TS and associated OCB. Among the obsessional symptoms, sexual and violent themes were more common in the TS group, whereas concern about contamination and fear of something going wrong or something bad happening were more common among the OCD group. With regard to compulsions, symmetry/evening up behaviors, saying or doing things "just right," and forced touching were more prevalent in the TS group, whereas washing and cleaning were more common in the OCD group. Sex of the proband did not account for any of these differences. Other investigators reported that patients with a history of tic disorder had significantly more touching, repeating, self-damaging, counting, and ordering compulsions. Similarly, in an epidemiological study of OCD, it was reported that OCD subjects with tics had significantly more touching, ordering, repeating, counting, violent, sexual, and aggressive behaviors. In yet another study, TS patients were found to have obsessions concerning harming self or others, intrusive nonsense words, music, or sounds, thoughts about something bad (fire, death, and illness) happening, and compulsions such as checking, excessive washing and tooth brushing, cleaning rituals, counting, and hoarding or collecting things. It was also found that TS patients with comorbid OCD were significantly more likely to report obsessions involving nonviolent images, need for symmetry and excessive concern with appearance, and compulsions such as touching, blinking or staring, and counting. The authors suggested that the differences in symptomatology between TS and primary OCD may be linked to putative differences in pathophysiology.

Thus, it has been suggested that there are at least two subtypes of OCD: one associated with abnormal risk assessment, high levels of anxiety, and pathologic doubt involving worry that something terrible might happen that may be relieved by checking and washing compulsions, and the other type associated with incompleteness and sense of imperfection that is relieved when a compulsion is performed just right. The latter was often associated with tics, TS, onychophagia, and trichotillomania. It has also been proposed that all these conditions belong to a family of OCD spectrum disorders, with the TS subtype having predominantly impulsive features.

However, a problem inherent in this approach of comparing OCD patients with or without tic disorder is the disagreement regarding whether certain symptoms should be classified as compulsions or complex motor tics. Furthermore, Leckman and colleagues reported considerable overlap between TS and OCD symptoms; 93% of their 135 subjects with a tic disorder reported experiencing premonitory urges before performing tics. Thus, it seems that the common distinction based on the involuntary nature of tics compared to OC symptoms is not entirely valid. It has been

reported that intentional repetitive behaviors in OCD differ from those in TS in that the former is preceded by cognitive phenomena and autonomic anxiety and the latter by sensory phenomena. In another study by the same investigators, it was found that, like the TS group, the OCD + TS group reported more sensory phenomena and fewer cognitions than did the OCD group.

An alternate approach to distinguish TS + OCB subjects from pure OCD individuals is to search for a symptom profile rather than individual symptoms. By identifying such phenomenological similarities and differences, it may be possible to gain a better understanding of the underlying etiology and pathogenesis of both TS and OCD. Thus, in a factor analysis of symptom subtypes, three factors (symmetry/hoarding, contamination/cleaning, and pure obsessions) were found, of which only the first factor was associated with a lifetime history of tics or TS. In a cluster analysis of OC symptoms, Eapen and associates found that the TS cluster was characterized by fear of harming self/others, obsessions with violent/aggressive themes, symmetry/evening up, saying/doing things just right, forced touching, and arranging, whereas the OCD cluster was characterized by obsessions about dirt, germs, and contamination as well as the need to tell/ask/know, fear of something bad happening, the need to be neat and clean, and excessive washing and cleaning. In addition to such differences in the symptom profile, both research and clinical experience suggest that the OCB occurring in the context of TS is ego-syntonic, causing less distress to the patients, and is often associated with lesser levels of anxiety. Primary OCD, on the other hand, seems to be ego-dystonic, causing considerable distress to the patients, is preceded by obsessive cognitions, is associated with higher levels of autonomic anxiety, and has fewer or no prior sensory phenomena.

IV. NEUROBIOLOGY

The role of serotonin in the etiology of OCD and that of dopamine in the etiology of TS have been well documented. The strongest body of evidence favoring the serotoninergic and dopaminergic basis of OCD and TS, respectively, comes from studies indicating the effectiveness of selective serotonin reuptake inhibitors (SSRIs) in OCD and dopamine antagonists in TS. Reviewing the available neurobiological and pharmacological data in OCD and TS, Eapen and Robertson mention that the basal ganglia and its frontal cortical circuits are the prime anatomic loci that are implicated in TS and OCD. Furthermore, there is growing evidence of the existence of anatomic and functional interactions between 5-hydroxytryptamine (5-HT) and dopaminergic systems, especially in the basal ganglia areas. 5-HT neurons are believed to maintain a tonic inhibitory influence on dopaminergic function in some regions of the brain, especially the midbrain and brain stem projections to the forebrain. Thus, the neuroanatomic hypothesis that the basal ganglia and its orbitofrontal connections may form the neuronal circuit that subserves TS and OCB, and the available evidence about the interaction between 5-HT and dopamine, is compatible with the role of the previously mentioned structures and neurotransmitters in the pathophysiology of TS spectrum OCB.

Additional evidence for the previously mentioned hypothesis comes from pharmaclogical data on the beneficial effects of neuroleptic augmentation in the treatment of OCD with SSRIs, especially when there are comorbid tics/TS. It has also been suggested that although there may be abnormalities in the serotonergic system in primary OCD patients, both serotonergic and dopaminergic systems may be involved in those with TS + OCD. Furthermore, Leckman and associates suggested that noradrenergic mechanisms are also involved in the pathobiology of TS, and that alterations in the balance of noradrenergic, dopaminergic, and serotonergic systems may be involved in the pathobiology of OCD.

Although no definitive abnormalities have been found, there is a tentative consensus regarding the brain areas involved in TS and OCD, and particularly implicated are the striatal and frontal areas. There is growing evidence of the role of multiple parallel neuronal circuits involving cortex, basal ganglia, and thalamus (fronto-striato-pallido-thalamo-frontal circuits) in the pathophysiology of TS and OCB, and that a failure of inhibition in these circuits may be the cause of inappropriate release of fixed action patterns and OCB. The central role of the basal ganglia in the storage and regulation of complex motor schemas and the existence of contiguous projection areas within the striatum from both the motor and sensorimotor cortical areas suggest that these projections are involved in TS. Leckman and colleagues reviewed the evidence for the previous hypothesis from animal immunohistological studies, postmortem studies of TS patients, as well as neuroanatomical, neurochemical, and neuroimaging studies. Neuroimmunological studies have also suggested that Sydenhams chorea, TS, and OCB might occur following streptococcal

infection through a mechanism of molecular mimicry whereby antibodies are produced that mistakenly recognize cells within the basal ganglia and cause an inflammatory response. Neuropsychiatric manifestations occuring as a result have been termed pediatric autoimmune neuropsychiatric disorders associated with streptococcal infection (PANDAS) by Swedo and associates. However, there is continuing debate about the exact role of such a mechanism in TS and OCD.

Complex processes, such as generation, maintenance, switching, and blending of motor, mental, or emotional sets, take place at the cortical level, whereas the planning and execution of these behaviors are coordinated in the basal ganglia. Given the diversity and complexity of activity within the basal ganglia, the consequence of disruption will depend largely on the site of the lesion and the associated interplay of neurochemical factors. Thus, the anatomical distribution of tics as well as the theme and content of obsessions and compulsions in OCB may be determined by these aspects of the basal ganglia circuitry and the site and extent of its involvement. It is known that an abnormality in the function of dopamine in the basal ganglia structures could affect wide areas of corticostriatal loops modulating and integrating both limbic and frontal motor circuits, thus influencing the clinical presentation and symptom profile in TS. Furthermore, Leckman and colleagues suggested that other modifying factors, such as age, gender, and other environmental factors including perinatal events and exposure, may play a role in determining the final symptom expression in TS.

It has been proposed that the previously mentioned basal ganglia cortical circuits may be hyperactive in OCD as evidenced by the hyperperfusion and increased metabolic changes in the orbitofrontal cortices and basal ganglia in OCD patients. Furthermore, it has also been suggested that the frontal hyperperfusion seen in OCD may be linked to increased arousal and anxiety in these cases. Although limited, the available literature on perfusion patterns in TS suggests hypoperfusion rather than increased blood flow. For example, increased relative metabolic activity and regional cerebral blood flow in certain frontal cortical areas have been documented in OCD when compared to TS. Preliminary studies in TS suggest involvement of caudate, anterior cingulate, medial temporal, and dorsolateral prefrontal areas in the form of hypoperfusion. In a family study using single photon emission computed tomography imaging, Moriarty and colleagues noted that the ''affected'' family members of TS probands showed hypoperfusion irrespective of whether they had tics, TS, or OCD. None of the affected members, including those who had only OC symptoms, showed hyperperfusion. Furthermore, the affected family members were noted to have lower anxiety scores. This, coupled with the finding by other investigators of a correlation between improvement in anxiety and a reduction in the orbitofrontal metabolism in OCD subjects, suggests that the difference in perfusion findings between OCD subjects and TS + OCD patients may be related to the difference in anxiety levels, which in turn may be determined by the specific OC symptom profile in these two groups of patients. Thus, it seems that even at a functional level, there are differences between primary OCD subjects and those with TS spectrum OCB.

V. GENETICS

In addition to the clinical and neurobiological relationship between TS and OCB, studies by Pauls and associates have suggested that OCB may be genetically related to TS. Family studies have shown that TS and OCB are genetically related, with the frequency of OCD in the absence of tics among first-degree relatives (FDRs) being significantly elevated in families of both TS + OCD and TS−OCD probands. The rate of OCD among FDRs was significantly increased compared to estimates of the general population and a control sample of adoptive relatives. The rates of TS, tics, and OCD were the same among relatives of TS probands with OCD (TS + OCD) when compared with families of probands without OCD (TS−OCD). Other family studies by Comings and colleagues also found that many relatives of TS patients have OC symptoms even in the absence of motor and vocal tics.

Segregation analyses studies of TS families by the previously mentioned groups of researchers found that TS and OCB may be an alternative expression of the putative TS gene(s). Furthermore, in a TS twin study, both twins were reported to be obsessional, and in a study of triplets, one was reported to be obsessional.

Genetic mechanisms have been found to be important in OCD as well, with a heritability rate of 44% for OC traits and 47% for OC symptoms, thus suggesting that genetic and specific environmental factors were both important for the manifestation of OCD. In a study on identical twin pairs who were separated prior to the onset of symptoms, neither being aware of the other's problems, it was found that the OC symptoms started at similar ages and followed a similar course in

both pairs. It is interesting that the fathers of both sets of twins had obsessional traits. Furthermore, one of these twins had childhood tics and two of the four sets of identical twins with OCD had TS. However, in all the twin data reported, the concordance for MZ twins was less than 1.0 and heritability estimates were consistently less than 1.0. In addition, an analysis of OC traits in twins has shown a strong interaction between genetic and environmental factors. Thus, although genetic factors are important in the expression of OC symptomatology, these behaviors are also influenced by environmental factors.

Based on evidence from two twin studies, it has been suggested that although genetic factors are important for anxiety disorders in general, this contribution is obscured by the grouping of anxiety symptoms into specific disorders. In this regard, in a study of FDRs of 32 adult OCD probands and 33 psychiatrically normal controls, it was found that the morbid risk for anxiety disorders was increased among the relatives of OCD subjects when compared to the relatives of controls, but the risk for OCD was not. Risk for a more broadly defined OCD was increased among the parents of OCD probands but not among the parents of controls (16 vs 3%). These findings suggest that an anxiety disorder diathesis is transmitted in families with OCD but that its expression within these families is variable.

When compared to population prevalence estimates, several family studies have reported significantly higher rates of OCD in parents and siblings of OCD probands, with rates among parents being 5–10 times higher. Two family studies found remarkably similar risk rates in FDRs of young OCD probands. One study of 145 FDRs of 46 children and adolescents with OCD reported an age-corrected morbid risk of 35% in FDRs when subclinical OCD was included. Another study examined families of 21 children and adolescents with OCD and found that 35.7% of parents received a diagnosis of clinical or subclinical OCD. In yet another study of 92 adult OCD probands, it was found that the rate of OCD among FDRs was only 3.4%. However, when probands were separated into two groups based on whether the age of onset was before or after 14 years, it was noted that the morbid risk for OCD among relatives of the early onset probands was 8.8% compared to 3.4% among the relatives of probands with a later age of onset.

However, the results from these investigations do not suggest that all subjects with OCD have a disorder that is genetically determined. It has been suggested that there are subgroups within OCD: those with a family history of OCD and those without such a history. Furthermore, there may be a familial subgroup that is genetically related to TS. Available evidence suggests that the OC symptoms observed in members of families with TS comprise a milder form with less distress and anxiety, and that the range and character of symptoms are somewhat different and specific when compared to those observed in clinical patients with OCD in the absence of tics or TS. Further evidence for heterogeneity in OCD comes from descriptive studies of childhood OCD and family studies. Although some investigators did not find an increased incidence of tic disorders among adult OCD patients, others, such as Riddle and colleagues, suggested that childhood-onset OCD may be different from adult-onset OCD and that the childhood-onset subtype may be more closely related to TS.

In a recent segregation analysis study of OCD families, a dominant model of transmission was suggested. When the phenotype was widened to include TS and chronic motor tics, an unrestricted model of transmission became the best fit, and it was proposed that the OCD phenotype probably presents a higher level of heterogeneity than the TS phenotype. Furthermore, the clinical and phenotypic heterogeneity may be linked to genetic heterogeneity, with some individuals having inherited the TS + OCD genotype and others the classic OCD genotype. In a family study of OCD probands and TS + OCD probands, Eapen and colleagues found that all the OCD probands who shared a similar symptom profile to that of TS probands had at least one first-degree relative with OCD, whereas none of the OCD probands with classic OCD symptoms had a positive family history. These authors concluded that the latter could be regarded as sporadic or nonfamilial cases. These observations may be consistent with genetic heterogeneity within both OCD and TS. In this regard, it is interesting to note the findings of a recent segregation analysis study by Pauls and colleagues, who used factor-analytic symptom dimensions to subset the family sample based on probands' symptom factor scores. They found that families with a high symptom score of symmetry and ordering had a higher risk of OCD in relatives and suggested that this may constitute a genetically significant subtype of OCD. This is in keeping with the findings of another study that reported that OCB is less prominent in sporadic than in familial TS, perhaps reflecting a more restricted pathophysiology in this subgroup. These investigators also found that although bilineal transmission of tics is relatively infrequent in consecutive TS pedigrees, cotransmission of OCB from an otherwise unaffected parent is

common and significantly influences the development of OCB and self-injurious behaviors, but not tics, in the offspring.

Thus, according to the collective knowledge, it seems that although there is little doubt about a familial relationship between OCD and TS, not all cases of OCD are either familial or etiologically related to TS. In this regard, three subtypes of OCD have been described by Pauls and associates: a familial type related to tic disorders, a familial type unrelated to tics, and a nonfamilial type. It is hoped that further refinements at various levels of diagnostic hierarchies will ultimately lead to an improved definition of what constitutes the genetic spectrum of TS. Furthermore, molecular genetics offers a promising method for exploring the underlying etiologic heterogeneity of TS and related conditions such as OCD.

See Also the Following Articles

CATECHOLAMINES • DOPAMINE • MOVEMENT REGULATION • NEURODEGENERATIVE DISORDERS • PSYCHOPHYSIOLOGY • SEX DIFFERENCES IN THE HUMAN BRAIN

Suggested Reading

Alsobrook, J. P., and Pauls, D. L. (1997). The genetics of Tourette syndrome. *Neurol. Clin.* **15**(2), 381–393.

Comings, D. E., and Comings, B. G. (1987). Hereditary agoraphobia and obsessive–compulsive behaviour in relatives of patients with Gilles de la Tourette's syndrome. *Br. J. Psychiatr.* **151,** 195–199.

Eapen, V., and Robertson, M. M. (2000). Gilles de la Tourette syndrome and co-morbid obsessive compulsive disorder. Therapeutic interventions. *CNS Drugs* **13**(3), 173–183.

Eapen, V., Pauls, D. L., and Robertson, M. M. (1993). Evidence for autosomal dominant transmission in Gilles de la Tourette syndrome—United Kingdom cohort. *Br. J. Psychiatr.* **162,** 593–596.

Eapen, V., Robertson, M. M., Alsobrook, J. P., *et al.* (1997). Obsessive compulsive symptoms in Gilles de la Tourette syndrome and obsessive compulsive disorder: Differences by diagnosis and family history. *Am. J. Med. Genet.* **74,** 432–438.

Leckman, J. F., Peterson, B. S., Anderson, G. M., *et al.* (1997). Pathogenesis of Tourette's syndrome. *J. Child Psychol. Psychatr.* **38**(1), 119–142.

Moriarty, J., Eapen, V., Costa, D. C., *et al.* (1997). HMPAO SPECT does not distinguish obsessive compulsive and tic syndromes in families multiply affected with Gilles de la Tourette syndrome. *Psychol. Med.* **27,** 737–740.

Pauls, D. L., Alsobrook, J. P., II Goodman, W., *et al.* (1995). A family study of obsessive compulsive disorder. *Am. J. Psychiatr.* **152,** 76–84.

Rapoport, J. L., Leonard, H. L., Swedo, S. E., *et al.* (1993). Obsessive compulsive disorder in children and adolescents: Issues in management. *J. Clin. Psychiatr.* **54,** 27–29.

Riddle, M. A., Scahill, L., King, R., *et al.* (1990). Obsessive compulsive disorder in children and adolescents: Phenomenology and family history. *J. Am. Acad. Child. Adolescent Psychiatr.* **29,** 766–777.

Robertson, M. M. (1989). The Gilles de la Tourette syndrome—The current status. *Br. J. Psychiatr.* **154,** 147–169.

Robertson, M. M. (1994). Annotation: Gilles de la Tourette syndrome—An update. *J. Child Psychol. Psychiatr.* **35,** 597–611.

Robertson, M. M. (1995). Relationship between Gilles de la Tourette's syndrome and OCD. *J. Serotonin Res.* **1**(Suppl.), 49–62.

Robertson, M. M. (2000). Tourette syndrome, associated conditions and the complexities of treatment. *Brain* **123,** 425–462.

Swedo, S. E., Leonard, H. L., Mittleman, B. B., *et al.* (1997). Identification of children with pediatric autoimmune neuropsychiatric disorders associated with streptococcal infections by a marker associated with rheumatic fever. *Am. J. Psychiatr.* **154**(1), 110–112.

Transplantation

TIMOTHY J. COLLIER
Rush Presbyterian–St. Luke's Medical Center

JOHN R. SLADEK, Jr.
University of Colorado Health Sciences Center

GLOSSARY

allograft Transplant consisting of cells or tissues from a donor of the same species.

apoptosis Cell death occurring due to a specific series of genetically orchestrated intracellular reactions.

autograft Transplant consisting of cells or tissues from the recipient's own body.

neurotrophic factor Molecules in the environment of central nervous system cells that support their survival and growth.

plasticity The capacity of nerve cells to alter their structure and function.

rejection The response of the immune system that coordinates the destruction of transplanted cells or tissues.

stereotaxic surgery Procedure that allows accurate targeting of deep brain structures within a system of three-dimensional coordinates.

xenograft Transplant consisting of cells or tissues from a donor of a different species.

We generally think of transplantation as a surgical approach for the treatment of diseases in which an entire organ is transferred from one body to another. As applied to the brain, however, transplantation refers to the surgical replacement of specific cells or clusters of cells of the brain, instead of the whole organ. Thus, this approach replaces brain cells that have died or stopped functioning due to injury or disease. This article reviews the theoretical background of neural transplantation, its history, current and future applications to the treatment of brain damage as developed from studies of animals and human patients, and the societal issues associated with the procedure.

I. WHY TRANSPLANT BRAIN CELLS?

Most of the organs in our body are made up of cells that are capable of constantly repairing and renewing themselves, maintaining the vital functions of the organs. For example, when your skin is damaged the cells can multiply and migrate to form new skin; thus, healing occurs over a short period of time. The cells lining the digestive system are constantly being renewed as old cells succumb to the harsh environment of digestive enzymes. However, there are certain types of cells in the body that exhibit little or no ability to multiply in order to renew the population. These include muscle cells of skeletal muscle and heart and the cells of the brain that process and transmit information—the neurons. Therefore, when neurons die there exists an extremely limited capacity for neuronal replacement. When neurons die in large

numbers due to injury or disease, permanent problems arise in the normal function of the brain. In Alzheimer's disease, for example, problems arise in the ability to learn and remember information. In Parkinson's disease, movement and balance are disturbed. These problems, and others associated with other types of brain injury or disease, can be related directly to death of specific populations of neurons that normally function to produce the mental or physical activities that become disturbed.

Given that the brain has a limited capacity to repair itself, how can brain disorders be treated? For some problems, there is no treatment currently available. For problems that can be linked to depletion of the chemicals used by neurons to transmit information, the "neurotransmitters" of the brain, the traditional treatment has been to replace lost activity by administering drugs that produce a similar effect. However, treating brain disorders with drugs can be problematic. Some potentially therapeutic substances may never gain access to the brain. Instead, they either are rapidly broken down by enzymes in other parts of the body or are unable to pass the physical barrier that exists between the blood supply and the brain, the so-called blood–brain barrier. In addition, it currently is not possible to restrict the action of a drug to the particular problem in the brain that needs treatment. Neurotransmitters used by neurons to perform their functions often exist in other systems of the body, where they perform different functions. Thus, although a drug might be administered to treat a problem in the brain, it circulates through the rest of the body and may act at other organs to produce side effects, for example, changing blood pressure or affecting digestion. Even within the brain a given neurotransmitter often exists in many different parts of the brain, each subserving a different function. Consequently, a drug that gains access to the brain will affect all neural systems that use the chemical, again producing side-effects that may be undesireable, such as confusion, anxiety, or sedation. Finally, the effects provided by drugs usually are not long-lasting. Even if a drug can be identified that treats the problem in the brain effectively and with acceptable side effects, the effects of the drug vary with its amount present in the body over time. Thus, high levels and strong effects occur immediately after ingestion, whereas low levels and loss of effect are seen several hours later as the drug disappears from the body.

Transplantation is a treatment strategy that aims to replace the function of neurons that die by replacing the neurons. With a combination of surgical skill and an understanding of brain anatomy and physiology, new cells can be placed into the specific location they are needed, become part of the brain, replace neurotransmitters, and function for the lifetime of the individual (Table I).

II. HISTORY OF NEURAL TRANSPLANTATION

For many years, neural transplantation was used in biological research as an experimental tool for understanding the development of the nervous system. Beginning in the 1920s, a number of now classic studies employed transplantation, removal and reimplantation, and transposition of nervous tissue to study the specificity of the developing interconnections of the nervous system and the capacity of this tissue for regeneration. Nearly all this work was done in amphibians, fish, and birds. Indeed, the robust growth capacity of nervous tissue in cold-blooded vertebrates, coupled with the limited success of early neural transplantation experiments in mammals, probably

Table I
Comparison of Neural Transplantation to Drug Therapy

Transplantation	Drug therapy
Targeted delivery: Replacement of neurotransmitter/other molecules at the specific site in brain required	Diffuse delivery: Effects produced throughout the body; side effects potentially problematic
Biological replacement: Multiple actions of natural cell products possible	Synthetic replacement: Highly specific drug action
Regulated delivery: Cellular release of molecules potentially controlled by surrounding cells	Delivery based on pharmacokinetics: Variable concentrations over time due to ingestion/absorption/circulation
Lifetime replacement: Cells integrate with host, becoming part of the tissue	Temporary replacement: Drug response varies over time

contributed to the scientific bias of the time that the regenerative properties of central nervous system (CNS) tissue were fundamentally different among species and deficient in mammals. In fact, it can be argued that the most important concept that has emerged from neural transplantation studies is an appreciation of the considerable capacity for regeneration, growth, and functional integration that can be exhibited in the complex CNS of mammals.

The first neural transplantation experiment reported in the scientific literature appeared in 1890. The study by W. G. Thompson at New York University examined the survival of pieces of cerebral cortex from adult cats transplanted into cortical cavaties of adult dogs. By 6 weeks after grafting essentially no transplanted neurons were observed to survive. In 1917, Elizabeth Dunn at the University of Chicago published the first study to convincingly demonstrate survival of transplanted CNS cells. Earlier, in 1909, S. Walter Ranson of Northwestern University successfully grafted neurons from the peripheral nervous system into the CNS of rats, documenting a form of morphological plasticity associated with younger tissue. In Dunn's experiments, cortical tissue from 9- and 10-day-old rats was grafted into cortical cavaties of littermates. At survival intervals of as long as 66 days, Dunn reported that grafts retained some of their normal cellular architecture. In addition to providing the first indication that CNS tissue could survive transplantation, the study provided more evidence that the developmental stage of the tissue was related to graft survival, and that transplantation of tissue in contact with a rich vascular bed may improve transplant outcome. In the 1940s, W. E. LeGros Clark at Oxford University began to report studies of embryonic CNS tissue transplants in rabbits that indicated that tissue from donors of even earlier developmental stage had greater capacity for survival and integration with the host nervous system. In 1945, Harry Greene and Hildegarde Arnold at Yale University provided the first report of successful transplantation of human embryonic CNS tissue to the anterior eye chamber of guinea pigs. Comparison of grafts of embryonic tissue from humans, rabbits, and guinea pigs revealed comparable excellent survival from all donors for periods of up to 2 years. With the demonstration that transplanted human tissue behaved similarly to transplants derived from other mammalian species, the slowly emerging concepts being developed in experimental animals became directly applicable to the human nervous system.

Although the roots of neural transplantation can be traced back to 1890, the approach largely remained a scientific curiosity until the late 1960s. At this time, a variety of new techniques were developed that enhanced the ability of scientists to study transplanted tissue. Interest in transplantation experienced a resurgence, with the number of research articles dealing with the topic that were published in 1983 alone equaling the number published in the entire century preceding 1970. In rapid succession, experimental findings documented the conditions that favored transplant survival and demonstrated that transplanted neurons exhibit a variety of important characteristics of mature neurons including manufacture and release of specific neurotransmitters, normal patterns of electrical activity, expression of cell surface receptors, the capacity to establish appropriate synaptic connections with the host brain, and influence the behavior of the host organism in a manner consistent with the biology of the transplanted cells.

Although transplantation experiments in animals have yielded several findings important to our current understanding of the capacity of neurons to alter their structure and function, properties termed "plasticity," transplantation as applied to the human brain exists in the realm of evolving medical therapies. The remainder of this article focuses on the development of neural transplantation as a treatment strategy for CNS injury and disease in humans.

III. WHICH CELLS CAN BE TRANSPLANTED TO THE BRAIN?

In animal experiments, many different types of cells have been transplanted into the brain or spinal cord with varying degrees of success. For treatment of disorders of the human CNS, the majority of transplantation trials to date have used nervous tissue or paraneural cells such as the chromaffin cells of the adrenal gland. However, there is reason to believe that many types of cells could be transplanted successfully to the brain. This contention is based on what is known about the interactions between the immune system and the CNS.

One problem associated with transplantation of tissues and organs is the potential for rejection. This is because the immune system functions to distinguish foreign organisms, cells, and tissues from those associated with one's own body. Cells of the immune system can identify "self" and "nonself" and coordinate the destruction of foreign substances. Transplants are classified as autografts when the cells or tissues are taken from other parts of the transplant recipient's

own body, homografts or allografts when taken from an individual of the same species, and heterografts or xenografts when taken from a member of a different species. Autografts tend to survive transplantation well. These transplants consist of cells that are genetically identical to the host and recognized by the immune system as self. In contrast, allografts and xenografts are composed of cells that are detected as foreign by the immune system, triggering rejection.

As neural transplantation experiments progressed, it became clear that transplant rejection was not as serious a problem as originally anticipated when tissues were transplanted to the CNS. As early as 1921, evidence was beginning to emerge that suggested that xenografts survived for relatively long periods of time in the brain while the same tissue implanted in muscle was actively rejected in a short period of time. Findings such as these led to the concept of the brain as an immunologically privileged site. Recently, this concept has been revised to incorporate research findings that indicate the brain is a site of reduced surveilance by the immune system, but that immunological responses can be provoked in brain, albeit often delayed and less vigorous than those that occur in other parts of the body. Part of this diminished immune surveilance may be attributable to the blood–brain barrier. The same barrier that shelters the brain from organisms and substances circulating in the bloodstream that are potentially toxic or damaging reduces the ability of immune cells in the circulation to access the brain. Although this barrier is physically opened by transplant surgery, it is reestablished relatively rapidly over a period of several days. This allows physicians to treat patients receiving neural transplants with drugs that suppress the immune response for a relatively short time after surgery, until the blood–brain barrier can be reestablished, whereas recipients of transplants in other parts of the body often take drugs for immune suppression for the rest of their lives.

Although the brain as a site for transplantation presents fewer problems than other parts of the body with respect to rejection, in theory making it possible to successfully transplant a variety of types of cells, most success has been achieved with implants of neural tissue taken from embryonic donors. From a commonsense perspective, the best cell to replace the function of a neuron is another neuron. It became clear in early neural grafting experiments that transplantation of adult CNS tissue seldom yielded surviving neurons, whereas implants of neonatal or embryonic tissue exhibited much better results. This is probably due at least in part to the anatomy of neurons. Neurons consist of a cell body that extends a number of fine cytoplasmic processes that serve as sites for transmitting, receiving, and integrating neurochemical/electrical signals from other cells in the CNS. In particular, the neuron's axon, the process subserving signal transmission, can extend for long distances in the CNS, intermingling with large numbers of processes of other neurons and glial support cells. This anatomical complexity of mature CNS tissue presents a problem for the harvest of neural tissue for transplantation. During dissection of donor tissue, mature neurons inevitably will be damaged by severing their long processes from the cell body, inducing molecular events leading to cell death. Neural tissue in the early stages of development does not present this problem. In fact, it has been found that neurons collected at a time during development just prior to extension of processes provide the best chance of optimal transplant survival. Thus, the appropriate time interval for collection of neurons for transplantation varies with the specific regional development of the neurons of interest. Knowledge of the timing of neural development and the landmarks and architecture of the immature brain becomes crucial to the collection of appropriate donor cells for transplantation. The features of embryonic neural cells considered optimal for transplantation are listed in Table II.

IV. HOW CELLS ARE TRANSPLANTED TO THE BRAIN

In comparison with many types of neurosurgery, transplantation surgery is a relatively straightforward procedure. In general, collection of embryonic neural tissue for transplantation is the most technically challenging aspect of the approach. In surgery using this type of donor tissue, the appropriate region of the developing CNS, from donors of an appropriate developmental stage, is collected via microdissection. Working with the aid of a microscope, the surgeon removes the region of brain containing the appropriate population of neurons for transplantation by utilizing the external and internal landmarks of the brain as reference points. Once the donor tissue is collected, it is usually prepared in one of two ways for transplantation into the host brain: either as small pieces of tissue (i.e., so-called solid tissue grafts) or as a physically and enzymatically dispersed solution of cells in the form of a "cell suspension." Tissue pieces or suspensions are held in an oxygen-rich fluid medium that can be

Table II
Optimal Characteristics of Neural Cells for Transplantation

Characteristic	Description
Postmigratory	Cells have reached their specific anatomical location in the developing brain that allows reproducible dissection and collection
Postmitotic	Neurons have reached a stage of maturity in which they no longer divide; provides a measure of safety against tumor formation by transplants
Differentiation	Cells have reached a stage of maturity in which they express the chemical machinery specific for production of neurotransmitters or other molecules needed for therapeutic replacement following transplantation
Prior to neurite extension	Neurons have not yet extended long processes making them prone to injury during collection, but they are poised for rapid growth of processes, which facilitates integration with the host brain following transplantation

supplemented with other ingredients that promote the health of the cells. Implantation of tissue or cells occurs by drawing the material into sterile syringe needles, or cannulae, and slowly infusing the material into the appropriate brain site of the host. The target for transplants can be determined with medical imaging techniques, such as magnetic resonance imaging, and the infusion cannulae can be accurately positioned using stereotaxic surgery—a method of accurately locating deep brain structures within a three-dimensional system of coordinates (Fig. 1). After surgery, drugs are administered to prevent infection and suppress the immune response in the case of allografts or xenografts.

The major problem in transplant surgery is not the delivery of donor cells to the appropriate location but survival of the transplanted cells. Transplanted cells are under assault at every stage of the procedure. The issue of cell survival has been studied in greatest detail for embryonic neurons making the neurotransmitter dopamine transplanted as an experimental therapy for Parkinson's disease. Factors influencing survival of these developing neurons are summarized in Fig. 2 but may be generalized to other donor cells. It is estimated that 80–95% of immature dopamine neurons harvested for transplantation die within 4 days of implantation into the host brain. As indicated in Fig. 2, factors influencing cell survival probably occur during three main stages of the transplantation procedure.

Stage 1 includes the time during which donor tissue is collected and prepared for implantation. It has been found that a significant proportion of cell death following transplantation is triggered during the preparation of tissue prior to transplantation. Among the potential triggers of cell death occurring during this interval are (i) physical trauma to the tissue as a consequence of dissection and handling; (ii) anoikis, a type of genetically programmed cell death triggered by detaching immature neural cells from the matrix of extracellular molecules in which they are embedded; (iii) ischemia—disruption of the supply of blood-borne oxygen and nutrients; (iv) oxidative stress—exposure to oxygen radicals that damage cell membranes and are generated by cells undergoing metabolic stress; and (v) withdrawal of trophic factors, molecules that exist in the environment of the developing brain that are known to promote survival and growth of immature neural cells. Stage 2 includes the 4 days immediately following transplantation. In this stage many of the same triggers of cell death exist but are in part generated by the environment of the host brain, not just removal of immature tissue from the developing brain. For example, withdrawal of trophic factors for the immature neural cells is compounded by the markedly lower concentrations of these molecules in the environment of the mature or aging host brain. Ischemia is likely to exist for several days for neural transplants. Unlike organ transplantation in the periphery, which often includes meticulous reconnection of the vascular supply of the transplanted organ, tissue transplants in brain are entirely dependent on growth of new blood vessels connecting the transplant with the host. This process takes place over 4–7 days, leaving the transplanted tissue dependent on diffusion of oxygen and nutrients from nearby host brain blood vessels for an extended time. Stage 3 includes the time after the first few days following transplantation. Most death of transplanted cells has already occurred, but delayed death of cells can be triggered by continued deprivation of trophic molecules and a delayed rejection response.

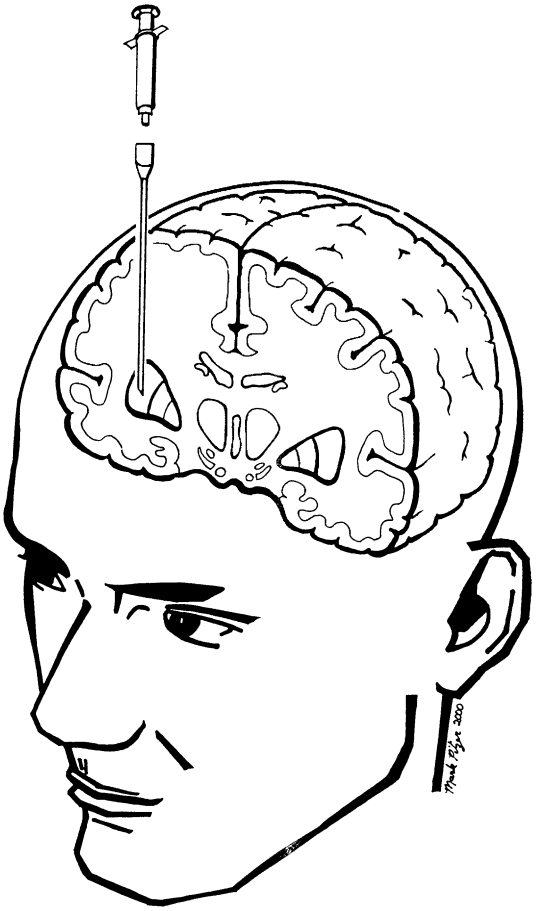


Figure 1 Diagram illustrating the principles of stereotaxic implantation of cells into the brain. Medical imaging techniques, such as MRI, provide visual cross sections through the brain, providing a map from which measurements are made that allow localization of deep brain structures for placement of transplants. The cross section shown approximates the detail provided by MRI and shows the putamen being targeted for transplantation. The putamen is a structure in the striatum and is a target of transplants for treatment of Parkinson's disease. A hollow cannula is lowered into position using stereotaxic surgery, a method for accurately locating deep brain structures within a three-dimensional system of coordinates. Cells for transplantation are slowly infused via the cannula, the cannula is withdrawn, and the surgery is completed.

Taken together, these elements present a formidable obstacle to graft survival. Thus, development of strategies designed to prevent death of cells transplanted to brain represents one of the keys to more widespread therapeutic application of this technique. Increased understanding of the triggers and mechanisms of cell death provides clues to potential methods for augmentation of survival of transplanted cells. For example, the appreciation that many cells are dying via the type of genetically programmed cell death observed during development of the nervous system, apoptosis, provides a framework for designing interventions. The molecular events participating in apoptosis have been defined, first by studies in invertebrates and then in mammals. Accordingly, strategies aimed at

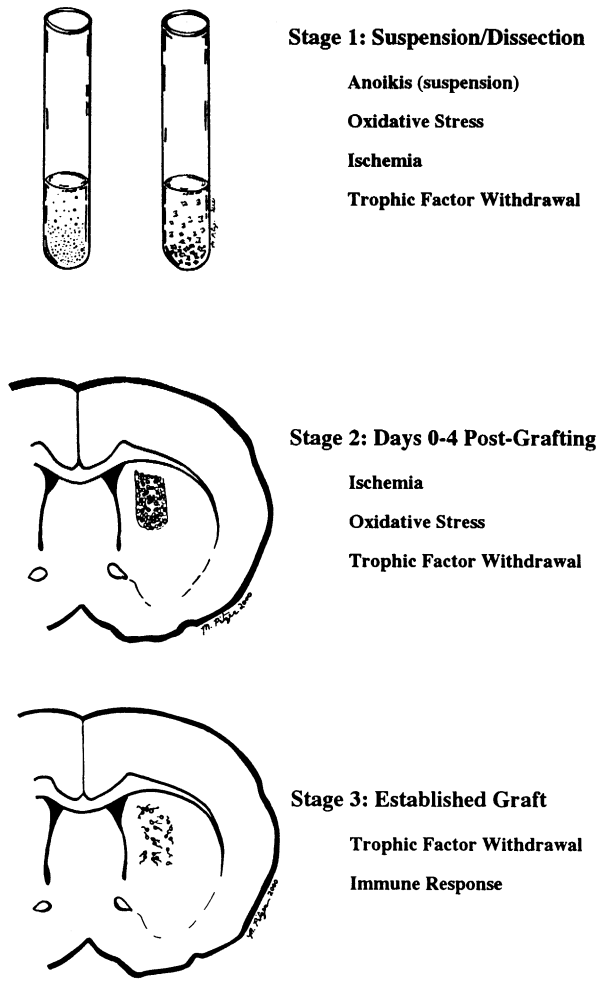


Figure 2 Factors associated with the transplantation procedure that may negatively affect survival of immature neural cells. The transplantation procedure has been divided into three stages. Factors that challenge cell survival for each stage are discussed in the text. Stage 1 covers the process of dissecting, collecting, and preparing the cells for transplantation. The diagram depicts a dissociated cell solution or small pieces of neural tissue held in test tubes prior to transplantation. Stage 2 covers the first few days after transplantation of cells. The diagram depicts a cross section through the rat brain at the level of the striatum. A newly implanted group of cells is shown as a cylindrical deposit on the right side of the brain. Stage 3 covers long-term survival of transplanted cells. Again, a cross section through the rat striatum is shown, with transplanted cells, now more integrated with the host brain, on the right side of the diagram.

interfering with this molecular cascade, such as pharmacological blockade of the action of the ultimate executioner molecule, caspase, have been demonstrated to enhance grafted cell survival in animal experiments. Similarly, approaches that increase the blood supply to transplanted cells, buffer environmental free radicals with antioxidants, and recapitulate the trophic environment of the developing nervous system by local replacement of survival- and growth-promoting molecules all show promise in ongoing studies.

V. TRANSPLANTATION AS A TREATMENT FOR PARKINSON'S DISEASE

In large part, the history of neural transplantation in humans is the history of transplantation as a treatment for Parkinson's disease (PD). PD is the human neurodegenerative disease most amenable to a neural transplantation approach for several reasons: (i) Symptoms of the disease are mainly due to loss of a single population of neurons, the dopamine-producing cells of the substantia nigra; (ii) death of these neurons depletes dopamine in specific locations of the brain, of which the regions of the striatum appear to be most important for the disruptions of movement that debilitate patients; and (iii) dopamine release in these brain regions is widespread and relatively constant, requiring no reconstruction of highly specific point-to-point connections that characterize some other systems of the brain. Thus, cell transplants that replace dopamine in the striatum, in a relatively nonspecific manner, have a high likelihood of being therapeutic in PD.

Additional support for considering transplantation therapy for PD is provided by the history of treating the disease with drug therapy. In the early 1960s, scientists and physicians began developing a drug therapy for PD consisting of the precursor molecule of dopamine, levodopa, which was known to cross the blood–brain barrier. Levodopa was administered in combination with carbidopa, which prevented breakdown of levodopa in the body, allowing it to access the brain. This combination of drugs revolutionized treatment of PD and remains an effective treatment of symptoms for several years early in the progression of the disease. However, with continued degeneration of dopamine neurons and long-term use of the drugs, therapeutic benefit decreases and undesireable side effects increase. Patients with advanced PD are left with unmanageable symptoms for most of their daily lives. The transplantation approach is an attractive alternative, potentially providing a cellular source of dopamine, implanted directly into the striatum to deliver dopamine locally to the structure important for the movement disorders and lasting for the lifetime of the individual.

Clinical application of neural transplantation to PD was preceded by more than 15 years of studies in rodents and nonhuman primates. These studies demonstrated that immature dopamine neurons could survive transplantation into the model environment of the anterior eye chamber of rats, with subsequent verification of survival of grafted dopamine neurons and the dopamine-producing paraneural cells, chromaffin cells, of the adrenal gland transplanted into the brain ventricles and striatum of rats. These studies documented that grafted cells produced and released dopamine and could ameliorate behavioral symptoms of dopamine depletion in rodent models of PD. In the case of transplants of immature dopamine neurons, it was also documented that cells extended processes into the host brain, made appropriate contacts with other neurons in the host, and maintained electrical properties typical of dopamine neurons. With the notable successes of rodent studies, investigations were carried over to nonhuman primates exposed to the neurotoxin 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP), which produced specific degeneration of substantia nigra dopamine neurons highly reminiscent of PD. Systematic study of transplants in this setting allowed development of parameters applicable to clinical trials by approximating the complexity of neural circuitry of the motor system in a species more closely related to humans.

The first neural transplants performed as a therapy for PD were undertaken in the early 1980s. Although transplants of embryonic dopamine neurons showed promise in animal experiments, the first tissue transplants in humans consisted of chromaffin cells of the adrenal gland implanted into the striatum. These cells normally produce small amounts of dopamine and larger amounts of the chemicals norepinephrine and epinephrine, which are synthesized from dopamine. However, when removed from the regulatory influences of other cells in the gland, chromaffin cells increase production of dopamine. In addition to making the appropriate neurochemical, chromaffin cells can be collected from the patient's own gland as an autograft, minimizing the chances of rejection. The use of chromaffin cells for transplantation was also not encumbered with the ethical issues associated with the use of tissue from embryonic donors. In 1985, the first

report of chromaffin cell transplants for the treatment of PD was published by investigators from Sweden, followed soon after by reports of trials conducted in Mexico. During the next few years, hundreds of patients received these transplants worldwide. Careful evaluation of 126 patients operated on in the United States revealed only modest improvements in symptoms, often lasting only 1 or 2 years after surgery, accompanied by significant medical or surgical complications in more than 40% of the cases. Of the cases that came to autopsy, most reported no surviving transplanted chromaffin cells expressing the biochemical signs of dopamine production. Studies of chromaffin cell transplants in animals demonstrated that despite poor survival of these cells, there was evidence of stimulation of surviving host dopamine neurons, perhaps via release of neurotrophic molecules by transplanted cells. This transplant-associated stimulation may account for the modest improvements in symptoms observed in some patients. In 1990, chromaffin cell transplants for the treatment of PD were discontinued in the United States.

With support for the therapeutic efficacy of chromaffin cell transplants waning, initial reports of human embryonic dopamine neuron transplants in patients with PD appeared in the literature in 1988. All of these initial clinical studies were performed outside of the United States because links between the issue of abortion and the use of embryonic tissue for transplantation sparked a societal debate in the United States that lasted for nearly 4 years. By 1996, published articles accounted for approximately 200 patients receiving embryonic tissue grafts worldwide. It is likely that many more patients have received transplant therapy and have not been accounted for in the scientific literature. Reports of effects on PD symptoms have been highly variable. In general, modest improvements have been observed, primarily expressed as a reduction in so-called "off time," the time during which medication provides no benefit for PD patients, and a reduction in dyskinesias, which are involuntary movements that are a side effect of medications for PD. In general, reports suggested that transplants failed to replace sufficient dopamine innervation to allow significant reduction of levodopa intake by patients to manage their symptoms. However, it is noteworthy that considerable variation existed around this average outcome, with some patients showing dramatic improvements.

Transplantation of embryonic dopamine neurons in PD patients continued at a cautious pace as animal experiments sought to inform physicians about strategies to optimize the transplantation approach. In 1992, an article was published by Scandinavian scientists indicating much greater therapeutic effects than previously had been reported following dopamine neuron transplants in two patients suffering from PD-like symptoms produced by exposure to the neurotoxin MPTP. In 1995, a group of scientists in the United States published a similar report of improved outcome following transplantation of immature dopamine neurons in PD patients, followed in 1996 by autopsy findings from two patients providing the first demonstration of good survival of large numbers of transplanted dopamine neurons over an 18-month period following surgery. In 1999, a report was published by the Scandinavian group on positron emission tomography imaging of a PD patient that suggested survival of transplanted embryonic dopamine neurons and continuous replacement of dopamine activity over a period of 10 years after surgery. A common feature of these studies was a revision of the transplant procedure to implant tissue from several donors into each patient. Animal experiments had begun to describe the problems associated with survival of transplanted immature neurons and the need for an initial strategy for compensating for loss of cells by increasing the amount of tissue transplanted. Although the relatively poor survival of transplanted neurons remained problematic, it is clear that when more cells were implanted a sufficient number survived to produce a better therapeutic effect.

The exchange of information between experimental findings and issues raised in animal studies and clinical trials has been one of the defining characteristics of neural transplantation as an experimental therapy. To date, every important variable associated with the transplantation procedure has been accurately predicted by animal experiments. The sequential examination of the properties of immature neurons grown and maintained in tissue culture, followed by animal experiments with transplantation in rodents and ultimately tests of safety and feasibility with transplants in the best model of the human brain, the nonhuman primate, has directly led to therapeutic application of neural transplantation and continues to provide strategies to improve the current approach. The microscopic photographs of dopamine neuron grafts presented in Fig. 3 illustrate the similarity of outcome that has been achieved in animal studies and human clinical trials. Some of the important issues for transplants of embryonic neurons that have been studied in animals and confirmed in patients are listed in Table III.

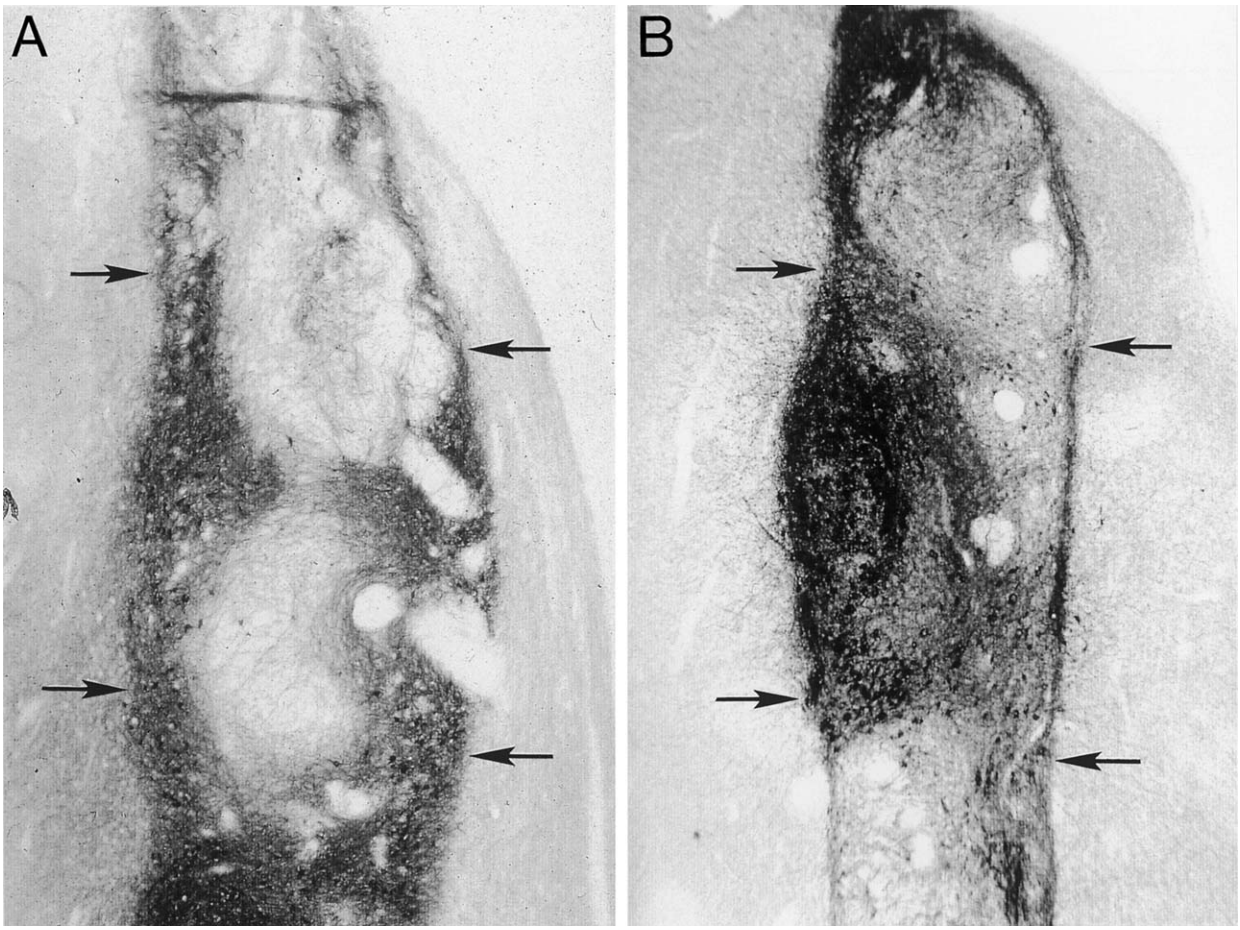


Figure 3 Comparison of transplants in human and monkey. Photographs of transplanted dopamine neurons visualized postmortem in the striatum of a human patient with Parkinson's disease at 18 months after transplantation (A) and in an MPTP-treated monkey 7 months after transplantation (B). Transplants (arrows) appear as similar cylindrical deposits of cells. Dark staining is for the enzyme tyrosine hydroxylase, which is necessary for the synthesis of dopamine and a characteristic marker for dopamine neurons. Note the nearly identical appearance of the transplants, illustrating one aspect of the predictive value that animal experiments have had for the development of transplant procedures for clinical application.

VI. THE SOCIOLOGY OF TRANSPLANTATION

The use of transplantation as a treatment for human neurological disorders carries with it a variety of ethical and societal concerns. At an abstract level, manipulations involving the brain as the organ of being human, the source of intellect and personality, may give rise to ethical and religious concerns. However, individuals afflicted with neurological disorders are often faced with desperate circumstances in which any hope of treatment overwhelms such philosophical issues. The more difficult societal issue for neural transplantation revolves around the current understanding that the best therapeutic results are provided by implants of embryonic tissue. In this context, the medical therapy is confronted with issues surrounding abortion.

This debate between science and ethics spilled over into politics with a request by scientists of the United States National Institutes of Health for permission to implant human embryonic dopamine neurons into patients with PD. In a climate of political conservatism, a moratorium of indefinite duration was placed on the use of federal research funds for transplantation of human embryonic tissue obtained from elective abortions into human patients in March 1988. The moratorium did not preclude the study of human embryonic tissue not used for transplantation in humans, and it did not target embyronic tissue obtained from spontaneous abortions. During the

Table III
Issues in Neuron Transplantation

Issue	Description
Donor age	Relatively narrow range (a few days) based on stage of development when specific characteristics of neuron type are expressed, but large-scale extension of neurites has not occurred
Amount of tissue	Currently, relatively poor survival of transplanted embryonic neurons requires implantation of tissue from multiple donors.
Target/distribution of tissue	In general, transplanted embryonic neurons extend neurites for short distances (2–3 mm) in the adult host brain. Thus, cells are transplanted within the brain region targeted for replacement therapy, and multiple transplants spaced close together provide the most complete replacement.
Form of transplant (pieces, suspensions)	Both have advantages. Both have been effective in human clinical trials.
Tissue storage (culture, hibernation, cryopreservation)	Storage of tissue prior to transplantation is desireable to allow testing for infection/contamination, best tissue type match, and transportation/distribution of tissue. Cryopreservation allows storage for very long periods of time (years), but cells are lost with freezing and thawing. Hibernation for short periods (2–3 days) does not reduce cell survival. Storage in tissue culture for short periods (4–8 days) with rotary motion and exposure to antioxidants and/or neurotrophic factors does not reduce cell survival
Immune suppression	Rejection is reduced in the CNS. Drugs for suppression of immune response have been administered for up to 6 months after transplantation, then discontinued, with no detected adverse effects
Pharmacotherapy	Patients receiving transplants often take medications to manage symptoms. For PD, animal experiments indicate mixed findings regarding whether levodopa therapy reduces survival or function of transplanted neurons. No conclusions have been drawn from human trials
Disease progression	Will transplanted cells be attacked by the neurodegenerative disease process existing in the patient? Studies using medical imaging techniques in PD patients indicate stable survival of transplanted neurons 10 years after implantation
Age of transplant recipient	Advanced age in the transplant recipient reduces therapeutic efficacy of transplanted neurons

moratorium, a government-appointed panel on human fetal tissue transplantation research met to assess both the science and the morality of using tissue from deceased aborted embryos. The panel, consisting of members representing a range of views on abortion, religion, and research, concluded by a vote of 17 to 4 that embryonic tissue transplantation was morally acceptable and that the issue of abortion and the use of tissues after death of the embryo were separable. In consensus agreement with similar debates that were carried out in Sweden, the United Kingdom, The Netherlands, Canada and elsewhere, it was believed that research of potential benefit to so many people should proceed as long as important safeguards were in place. Among these were prohibition of the sale of tissue and procedures for isolating the decision to have an abortion from consent to use the tissue for research.

Despite the recommendations of the panel, no immediate action was taken to change the moratorium. During this time, first reports of neural transplantation for PD were published by investigators in Sweden, Mexico, the United Kingdom, and China. Progress in embryonic tissue transplants for PD and promising work on transplantation of embryonic pancreatic islet cells for treatment of diabetes were slowed in the United States and moved forward only to the extent that private funds could be garnered to support the research. In January 1992, 2 days after a change of administration in Washington, DC, the government lifted the moratorium on human embryonic tissue transplantation research. The research is now governed by strict guidelines that borrow from those already accepted as the standard for acquisition of other organs for transplantation from human cadavers. Although work in the area of embryonic tissue transplants proceeds, the association of this approach with abortion remains a contentious issue that will remain unacceptable to some individuals. Accordingly, one challenge for the field of neural transplantation is the development of approaches that take advantage of the important characteristics of immature cells but move away from a continuing need

for tissue derived from aborted embryos. Several approaches are being explored, including (i) the use of gene transfer techniques to alter adult cells, such as skin fibroblasts, to express therapeutic molecules and allow them to be maintained in cell culture for an indefinite period of time, providing a continuous source of cells for transplantation; (ii) the use of neural stem cells or progenitor cells—cell types that are more immature than embryonic neurons and glia currently used for transplantation—that maintain the capacity to proliferate in cell culture as a source of cells for transplants and potentially can respond to specific chemical cues to differentiate into the cell type of therapeutic choice; and (iii) the use of viral vectors, a strategy to "infect" the nervous system with therapeutic molecules carried in the genetic material of experimentally manipulated viruses.

VII. OTHER USES OF TRANSPLANTATION

Experimental evidence suggests that neural transplantation will be most effective in neurodegenerative diseases or injury that involves the focal loss of neurons. It appears that in the environment of the adult or aging host brain, molecular cues may not be present in sufficient quantity to provide for complete long-distance reconstruction of neural circuits or the establishment of highly specific, complex, point-to-point connections by implanted neurons. Transplantation of neural stem cells may represent an exception to this, as discussed later. Use of nonneuronal cells for transplantation requires an even less demanding set of conditions in which mere secretion of a chemical in the local environment of the transplant could be expected to be therapeutic. Given these constraints, some types of disease or injury remain reasonable candidates for transplantation therapy. We have described the characteristics of PD that make this disease particularly appropriate for transplant treatment. Another neurodegenerative disease that has been studied in this context is Huntington's disease. In this fatal disease, neurons localized within the striatum die in large numbers, yielding significant atrophy of this brain region accompanied by abnormal body movements and psychological disturbances. The replacement of lost tissue by transplantation of embryonic striatal cells into the striatum appears to be another logical application of the transplantation approach. Replacement of striatal neurons via transplantation has been studied with some success in animal models, and initial trials in patients have been reported. Although establishing the feasibility and safety of the transplant surgery, these trials are too recent to provide results relevant to therapeutic outcome.

Transplantation to correct neurological deficits relevant to Alzheimer's disease has also been studied in animal models. Problems associated with learning and remembering new information that can be linked to loss of neurons that secrete the neurotransmitters norepinephrine and acetylcholine have been individually corrected with transplants of embryonic neurons in both young and elderly rodents. Replacement of the function of acetylcholine neurons with transplants has also been successful in monkeys. Although these therapeutic effects have sometimes been dramatic, Alzheimer's disease affects multiple brain regions and populations of cells, making it difficult to treat with the transplant approach. While the relatively diffuse actions of norepinephrine and acetylcholine have been successfully replaced in animal models, the more complex restoration of point-to-point connections that characterize the cells of the cortex and hippocampus, which are also affected by the disease, has been less successful in transplant experiments.

One recent option for the neural transplantation approach is the implantation of neural stem cells. As mentioned previously, these cells are more immature than the primary embryonic neurons and glia that have been employed in most neural transplantation therapy to date. Stem cells are self-maintaining, self-renewing, and multipotential. Accordingly, they provide a continuously available source of cells for transplantation, limiting the continuous need for embryonic tissue, and they may respond to specific molecular cues to mature into multiple types of neurons and/or glial cells that may be of therapeutic importance. Of particular interest is the finding that stem cells can migrate away from the site of implantation, be attracted to injured brain regions, and mature into the types of neural cells required to repopulate the damaged area. For example, following induction of experimental stroke in mice, transplanted stem cells were found to repopulate the damaged area of cerebral cortex and mature into neurons and oligodendrocytes, the primary cells lost following interruption of the blood supply. Similar transplantation of stem cells into the spinal cord of mice with peripheral nerve damage that leads to degeneration of lower motor neurons yields maturation of stem cells into cells with characteristics of motor neurons. These early studies are consistent with the view that local environmental signals generated during injury and degeneration of CNS regions may attract and cause site-specific maturation of implanted

stem cells in a manner consistent with reconstitution of the injured neural region.

Cell transplants have also been used as a vehicle for local administration of neurotrophic factors in the CNS. For this approach, cells are modified in tissue culture, using gene transfer techniques, to produce and release molecules that support survival and growth of specific neurons that are endangered by disease or injury. The cells that have been used most frequently are skin fibroblasts or baby hamster kidney cells. Following insertion of genes coding for growth factor production, the cells are either implanted directly into the brain site of interest or encapsulated in a semipermeable membrane prior to grafting. This approach has been successful in rescuing vulnerable neurons in animal models of Alzheimer's disease using nerve growth factor (NGF)-secreting cells and in animal models of Huntington's disease using cells engineered to produce either NGF or ciliary neuronotrophic factor.

VIII. CONCLUSIONS

During the past 25 years, cell transplant therapies for the treatment of neurological injury and disease have moved from science fiction to science fact. The safety, feasibility, and efficacy of transplant therapies have been firmly established by detailed, systematic studies in animal models of injury and disease and confirmed in clinical trials, primarily in the treatment of PD. Serious limitations and concerns still exist that have prevented more widespread clinical application of neural transplantation. These include limited survival and growth of immature neural cells transplanted into the adult and aged nervous system and the sometimes highly polarized views concerning the morality of using embryonic cells for transplants. As the intricate cellular mechanisms of life, death, and growth are better understood, it is likely that graft survival will not be an issue. With the emerging inquiry into neural cell lineage, including the study of stem cells, progenitor cells, and the cues necessary to drive differentiation into specific neural cell types, the possibility of a safe, dependable, and virtually inexhaustable supply of cells for transplantation may become an acceptable resource for medicine and society. If these problems can be successfully addressed, the original promise of the neural transplantation approach can be realized: replacement of neurotransmitters or other cell products lost or compromised by disease or injury at exactly the site required by replacing the cells themselves, which integrate with the host CNS, providing therapeutic benefit for the lifetime of the individual.

See Also the Following Articles

ALZHEIMER'S DISEASE, NEUROPSYCHOLOGY OF • BRAIN DEVELOPMENT • NEUROPLASTICITY, DEVELOPMENTAL • PARKINSON'S DISEASE • PHINEAS GAGE • PSYCHONEUROIMMUNOLOGY

Suggested Reading

Freed, W. J. (2000). *Neural Transplantation. An Introduction*. MIT Press, Cambridge, MA.

Olanow, C. W., Kordower, J. H., and Freeman, T. B. (1996). Fetal nigral transplantation as a therapy for Parkinson's disease. *Trends Neurosci.* **19**(3), 102–109.

Sladek, J. R., Jr., and Gash, D. M. (Eds.) (1984). *Neural Transplants. Development and Function*. Plenum, New York.

Tuszynski, M. H., and Kordower, J. H. (Eds.) (1999). *CNS Regeneration. Basic Science and Clinical Advances*. Academic Press, San Diego.

Unconscious, The

JOHN F. KIHLSTROM
University of California, Berkeley

GLOSSARY

automatic processes Perceptual–cognitive operations that are initiated involuntarily, executed outside phenomenal awareness, and consume no attentional resources.

dichotic listening A technique in which different auditory messages are presented over separate earphones; the subject is instructed to repeat (shadow) one message but ignore the other.

dissociation A statistical outcome in which one variable, either a subject characteristic (such as the presence of brain damage) or an experimental manipulation (such as the direction of attention), has different effects on two dependent measures (such as free recall or priming). In psychiatry, the "dissociative disorders" include psychogenic amnesia, psychogenic fugue, and multiple personality disorder: All three disorders involve disruptions in consciousness involving memory and identity.

explicit memory Conscious recollection, as manifested in a person's ability to recall or recognize some past event. Explicit memory is the model for the definition of explicit perception, learning, thought, emotion, and motivation.

functional magnetic resonance imaging A brain-imaging technique using magnets to measure blood oxygenation due to brain activity.

Implicit memory Any effect on task performance that is attributable to a past event, independent of conscious recollection of that event. Implicit memory is the model for the definition of implicit perception, learning, thought, emotion, and motivation.

limen From the German, "threshold." In classical psychophysics, the threshold referred to the minimum amount of energy required for an observer to detect the presence of a stimulus (the absolute threshold) or a change in a stimulus (the relative threshold). Processing of stimuli, which are undetected, is referred to as subliminal. In modern signal-detection theory, the concept of the threshold has been replaced by that of zero detectability ($d' = 0$).

preattentive processing Perceptual–cognitive operations occurring before attention has been paid to a stimulus.

priming The facilitation of perceptual–cognitive processing of a target stimulus by prior presentation of a priming stimulus. "Negative" priming refers to the inhibition of processing occurring under the same conditions.

Consciousness has two aspects: (i) monitoring ourselves and the world around us so that percepts, memories, thoughts, and other mental states are represented in phenomenal awareness and (ii) controlling ourselves and the environment so that we are able to voluntarily initiate and terminate behavioral and cognitive activities. In psychology, the unconscious refers to the idea that cognitive, emotional, and motivational states and processes can influence ongoing experience, thought, and action outside of phenomenal awareness and voluntary control.

I. CONSCIOUSNESS IN PSYCHOLOGY

In the last twenty-five years of the 19th century, as the new science of psychology began to emerge from its roots in philosophy and physiology, consciousness was at the center of the enterprise. Beginning with Wilhelm Wundt and E. B. Titchener, the whole structuralist school of psychology attempted to analyze conscious mental states in terms of their constituent sensations, images, and feelings. Its preferred method of introspection assumed that people had accurate

introspective awareness of their own mental states. In 1927, E. G. Boring summarized the achievements of structuralism with a monograph titled *The Physical Dimensions of Consciousness*.

Even William James, opposed as he was to the doctrines of structuralism, embraced a version of introspection as his preferred research method (he had a collection of brass instruments, but he hated using them). James began his *Principles of Psychology* with the assertion that "psychology is the science of mental life." By this he meant *conscious* mental life, as he made abundantly clear in the *Briefer Course*, in which he adopted G.T. Ladd's definition of psychology as "the description and explanation of states of consciousness as such."

II. THE INITIAL DISCOVERY OF THE UNCONSCIOUS

At the same time, both James and the structuralists understood that there was more to mental life than was accessible to introspection. The notion that unconscious processes are important elements of mental life is commonly ascribed to Sigmund Freud, the founder of psychoanalysis, but in fact it was an old idea, originating before Freud was born.

A. The Unconscious in Prepsychological Philosophy and Physiology

At the beginning of the 18th century, long before psychology split from philosophy and physiology, the German philosopher Leibniz wrote that our conscious thoughts are influenced by sensory stimuli of which we are not aware. At the close of that century, Immanuel Kant's *Anthropology from a Pragmatic Point of View*, the philosopher's last work and arguably the first comprehensive textbook of psychology, devoted a major section to a discussion of "the ideas which we have without being conscious of them."

In the 19th century, Johann Friedrich Herbart, drawing on the views of Gottfried Wilhelm von Leibniz, defined the "*limen*," or sensory threshold, as a mental battleground where various perceptions, themselves mostly unconscious, competed for representation in consciousness. In the *Treatise on Physiological Optics*, Hermann von Helmholtz argued that our conscious perceptions are determined by unconscious inferences.

The pre-Freudian analysis of unconscious mental life reached its apex with Eduard von Hartmann and his *Philosophy of the Unconscious* (1868), an extremely popular work whose three volumes, comprising more than a 1000 pages, went through a total of 12 editions. For Hartmann, the universe was ruled by the unconscious, a highly intelligent dynamic force composed of three layers: the absolute unconscious, accounting for the mechanics of the physical universe; the physiological unconscious, underlying the origin, evolution, development, and mechanisms of life; and the relative unconscious, which Hartmann considered to be the origin of conscious mental life. Hartmann's relative unconscious is what we nowadays call the psychological unconscious—a term referring to those mental states and processes that influence our experience, thought, and action outside phenomenal awareness and independent of voluntary control.

We owe to Hartmann the Romantic notion, still with us in some quarters today, that the unconscious possesses capacities and powers that are superior to those available to consciousness. As Hartmann stated, "the Unconscious can really outdo all the performances of conscious reason." In the end, however, Hartmann's ideas proved to be too speculative for the first generation of scientific psychologists. Hermann von Ebbinghaus, in discussing Hartmann's book, concluded that "what is true is not new, and what is new is not true." William James also offered a warning that would reverberate throughout the 20th century exploration of the unconscious: "The distinction … between the unconscious and the conscious being of the mental state … is the sovereign means for believing what one likes in psychology, and of turning what might become a science into a tumbling-ground for whimsies."

B. The Psychoanalytic Unconscious

Nevertheless, all of this activity, from Leibniz and Kant to Helmholtz and Hartmann, laid the foundation for what Henri Ellenberger, the great historian of psychiatry, called the discovery of the unconscious. This discovery was consolidated with what Ellenberger called a new dynamic psychiatry—the psychiatry of Sigmund Freud and his sometime compatriots, C. G. Jung and Alfred Adler.

Based on their clinical observations, for example, Freud and collaborator Josef Breuer concluded that the symptoms of hysteria were produced by unconscious memories of traumatic events. They stated, "Hysterics suffer mainly from reminiscences," with the proviso that these memories are unconscious,

emerging only after skilled intervention by a therapist. In *The Interpretation of Dreams* (1900), Freud proposed a topographical division of the mind into conscious, preconscious, and unconscious compartments. Later, he proposed that our conscious experiences, thoughts, and actions are determined by the interaction between unconscious sexual and aggressive impulses, on the one hand, and defense mechanisms such as repression arrayed against them in order to ward off anxiety caused by the conflict between these primitive motives and sociocultural demands and strictures, on the other hand.

C. The Impact of Behaviorism and the Revival of Consciousness

Unfortunately, just when the concept of the psychological unconscious was gaining support, the behaviorist revolution occurred. Interest in consciousness disappeared virtually overnight, and interest in the psychological unconscious went with it. For John B. Watson and comrades in arms, the only way to make psychology truly scientific was to abandon the mental. It was bad enough to try to explain behavior in terms of mental states that could not be publicly observed; it was even worse to try to explain behavior in terms of mental states that could not be privately observed. Of course, psychological interest in consciousness did not die completely with the triumph of behaviorism. So, too, some psychologists maintained an interest in the psychological unconscious. Still, a full-scale revival of academic interest in consciousness had to wait until behaviorism was overthrown by the cognitive revolution.

Ironically, the early cognitive psychologists rarely used the term consciousness. This is a reflection of two tendencies described by the philosopher Owen Flanagan: On the one hand, there is a positivistic reserve reflecting our persisting reluctance to use mentalistic language, and on the other hand, there is a piecemeal approach that assumes that major problems such as consciousness can be solved by working up from the bottom. However, consciousness was there anyway, in the guise of such topics as selective attention, primary or short-term memory, and mental imagery.

III. THE REDISCOVERY OF THE UNCONSCIOUS

A revival of interest in unconscious mental life followed shortly thereafter. The seeds for this revival had been planted at the very beginnings of the cognitive revolution, when the linguist Noam Chomsky argued that human language was mediated by "deep" grammatical structures that are inaccessible to conscious introspection and can be known only by inference. Along the same lines, the philosopher Jerry Fodor argued that many mental functions, such as visual perception, were mediated by dedicated structures that were impenetrable by conscious awareness and voluntary control. Cognitive approaches to perception, as exemplified by the work of Irvin Rock on visual illusions, entailed a version of Helmholtz's notion of unconscious inference. Finally, the classic multistore model of memory invoked a concept of preattentive, or preconscious, information processing.

We are now at a point, however, where interest in the psychological unconscious runs wide and deep within psychology. This happy state of affairs is the end-product of at least four independent strands of investigation, which together converge on our modern conception of the psychological unconscious: automaticity, cognitive neuropsychology, subliminal perception, and hypnosis.

A. Automatic and Controlled Processes

One research tradition contributing to the modern interest in the psychological unconscious is the distinction commonly drawn between "automatic" and "strategic" cognitive processes. Skilled reading provides one example of automaticity: We recognize certain patterns of marks on the printed page as letters and certain patterns of letters as words, and we decode the meanings of words in light of the words near them, but we rarely have any conscious awareness of the rules by which we do so. It just happens, as an automatic consequence of having learned to read.

The power of these processes is illustrated by the color–word effect discovered by J. R. Stroop. In the basic "Stroop" experiment, subjects are presented with a list of color names printed in different colors and are asked to name the color in which each word is printed. This task is easy if the ink color matches the color name (e.g., the word yellow printed in yellow ink), but if the word and its color do not match (e.g., yellow printed in green ink) it is very difficult. Despite the subjects' conscious intention to name ink colors and to ignore the words, they cannot help reading the color names, and this interferes with naming of the colors. It just happens automatically.

According to traditional formulations, automatic processes are inevitably engaged by the appearance of

specific environmental stimuli, regardless of the person's conscious intentions. Once invoked, they proceed inevitably to their conclusion, and (in theory at least) their execution consumes no attentional resources. Because they consume no attentional resources, automatic processes leave no traces in conscious memory. Some automatic processes are innate, or nearly so, whereas others are automatized only after extensive practice of a task. Recently, some of these properties have been called into question by revisionist, memory-based views of automaticity. It is no longer clear that ostensibly automatic processes are really executed involuntarily and really consume no attentional resources. However, even these revisionist views concur that some mental processes, represented in procedural memory, are unconscious in the strict sense of the term: They are inaccessible to phenomenal awareness in any circumstances and can be known only by inference from task performance.

B. The Rise of Neuropsychology

Both traditional and revisionist approaches to automaticity assume, at least tacitly, that the mental contents on which these processes operate are accessible to conscious awareness. However, it is now clear that our experiences, thoughts, and actions can be influenced by mental contents—percepts, memories, thoughts, feelings, and desires, of which we are unaware. Compelling evidence for this proposition began to accumulate about 30 years ago, as cognitive psychology turned into cognitive neuropsychology and researchers began to obtain evidence of the psychological unconscious in the behavior of brain-damaged patients.

Pride of place in this history goes to studies of the amnesic syndrome resulting from bilateral damage to the hippocampus and related structures in the medial temporal lobe or, alternatively, to the diencephalon and mammillary bodies. On clinical observation, such patients show a dense anterograde amnesia: After only a few moments of distraction, they cannot consciously remember events that occurred recently. However, as the study of these patients shifted from clinical description to controlled laboratory investigation, it became apparent that the events apparently covered by the amnesia nonetheless influenced the patients' ongoing experience, thought, and action.

For example, Elizabeth Warrington and Lawrence Weiskrantz showed that amnesic patients who could not remember having studied a list of words were nonetheless biased to complete ambiguous word stems or fragments with items from the previously studied list. Past experience influenced their subsequent task performance, even though they had no conscious recollection of the experience. Based on "priming" effects such as these, Daniel Schacter and others drew a distinction between two expressions of memory, explicit and implicit. Explicit memory refers to one's conscious recollection of the past, as manifested on tasks such as recall and recognition. In contrast, implicit memory refers to any change in experience, thought, or action that is attributable to a past event, regardless of whether that event is consciously remembered.

Priming effects in which prior exposure to a word such as "assassin" makes it easier to complete a fragment such as "a__a__i_" than one such as "t_p__r_" are good examples of implicit memory because the priming effect obviously depends on memory, but the task does not, logically, require conscious recollection of any past event. The subject must only generate an acceptable word that fits in the spaces provided. The sparing of implicit memory in amnesia shows that some representation of a prior event has been encoded and stored in memory and influences ongoing experience, thought, and action, even though that event cannot be consciously remembered. Implicit memories are unconscious memories.

Neuropsychological research has also revealed unconscious influences in the perceptual domain. Perhaps the most dramatic example is the phenomenon of blindsight documented in some patients with damage to the striate cortex of the occipital lobe. Such patients experience a scotoma—a portion of the visual field where they have no visual experience. When a stimulus is presented to their scotoma, they see nothing at all. However, when encouraged to make guesses about the properties of the stimulus, their conjectures about the presence, location, form, movement, velocity, orientation, and size prove to be more accurate than would be expected by chance.

Something similar occurs in at least some cases of visual neglect arising from lesions in the temporoparietal region of one hemisphere (usually the right) that do not affect primary sensory or motor cortices. These patients appear to neglect the corresponding portion of the contralateral sensory field (usually the left). Thus, a patient asked to bisect a set of horizontal lines may ignore the ones on the left side of the page; and for the remainder, the pencil strokes tend to be located about one-fourth of the way in from the right. It is as if the left half of the page, and the left half of each line, is

not seen at all. However, in some cases, it can be shown that these patients respond to information available only in the neglected field. For example, pictures presented in the neglected portion of the visual field prime lexical decisions concerning semantically related words presented in the intact portion. Preserved visual functioning in blindsight and in neglect is unconscious perception.

C. The Subliminal

Perception without awareness can also be observed in neurologically intact subjects in the form of "subliminal" perception. Before World War II, the question of subliminal perception was raised mostly in regard to psychoanalysis; after the war, the notion was revived as part of Jerome Bruner's "New Look" in perception, only to be shot down by the withering critiques of Israel Goldiamond and C. W. Eriksen. However, in the early 1980s Anthony Marcel presented solid evidence of subliminal semantic priming effects on lexical decision: Presentation of a word such as "doctor" primed lexical decisions of a semantically related word such as "nurse," even though an intervening mask prevented subjects from consciously perceiving the prime.

Other investigators soon confirmed and extended Marcel's findings, but his studies raised a firestorm anyway, with a number of critics essentially repeating the criticisms that Goldiamond and Eriksen had made of Bruner and colleagues. A major reason for this response may have been the association of subliminal perception with psychoanalytic theory. However, another reason was the simple fact that the cognitive theories of the time tended to describe cognition in terms of a series of ever more complicated processes and had no room for the possibility that the meanings of words could be analyzed unless conscious attention was paid to them. The criticisms had the character of the apocryphal entomologist who found a bug he could not classify so he stepped on it.

In any event, things are vastly different now. Signal detection theory has replaced the notion of an absolute limen, or threshold, with the statistical concept of zero sensitivity (where $d' = 0$). A wealth of evidence now supports the validity of subliminal perception, defined broadly as the influence of stimuli that are too degraded by their conditions of presentation to be accessible to conscious perception. The debate now is not so much over whether subliminal perception (so defined) occurs as over the extent of subliminal processing.

D. The Role of Hypnosis

The fourth line of research contributing to the rediscovery of the unconscious was hypnosis, a social interaction in which the subject acts on suggestions for experiences involving alterations in perception, memory, and the voluntary control of action. As William James recognized more than a century ago, many of these phenomena involve a division in consciousness such that memories, percepts, and the like influence experience, thought, and action outside of phenomenal awareness.

Consider posthypnotic amnesia, the phenomenon that gave hypnosis its very name. After receiving appropriate suggestions, many highly hypnotizable subjects come out of hypnosis unable to remember the events and experiences that transpired while they were hypnotized. For example, subjects who studied a list of animal names while hypnotized will be unable to remember them afterward. However, these unremembered items will also give rise to priming effects: If asked to generate names of animals, subjects will be more likely to give the names they studied while under hypnosis. Moreover, after the amnesia suggestion has been canceled by a prearranged reversibility cue, the subject will regain perfect conscious memory for the studied list. The reversibility of amnesia indicates that, in contrast to the organic amnesias associated with hippocampal damage, posthypnotic amnesia reflects a deficit in retrieval rather than encoding. However, the preserved priming effects show that the retrieval deficit affects explicit, but spares implicit, expressions of memory.

Hypnotic suggestion can also affect perceptual functions. When given hypnotic suggestions for blindness, many hypnotizable subjects have the subjectively compelling experience that they no longer can see. However, Richard Bryant and Kevin McConkey presented subjects with cards on which were printed homophones (i.e., words that have the same sound but two different spellings, such as "pain" and "pane") together with disambuguating words (e.g., "body" or "window"). The hypnotically blind subjects did not see the cards, but on a subsequent test they spelled the homophones in accordance with the disambiguating associates with which the words had been paired. This is another kind of priming effect, and it clearly indicates that the words in question were perceptually processed outside of awareness.

A third example of unconscious processing in hypnosis is provided by posthypnotic suggestion—the phenomenon that, our mythology tells us, gave Freud

his first good insight into the psychological unconscious. In some sense, posthypnotic suggestion is a special case of implicit memory because subjects act on a suggestion given during hypnosis, despite the fact that they cannot remember the suggestion. Posthypnotic behavior is typically experienced as an involuntary, quasi-automatic response to the cue, but it has had no opportunity to be automatized by extensive practice, and detailed experimental analysis shows that it does not fit the conventional criteria of automaticity. Despite consuming considerable attentional resources, posthypnotic suggestions are executed, outside of awareness, thus extending the boundaries of the psychological unconscious beyond the automatic.

IV. IMPLICIT COGNITION

The rediscovery of the unconscious may have gotten a late start, but it is now well along.

A. The Automaticity Juggernaut

The concept of automaticity has proved to be particularly popular across a wide variety of fields within psychology. For example, a set of articles in the July 1999 issue of *American Psychologist*, published under the title of "Behavior—It's Involuntary," illustrated the increasingly powerful role played by the concept of automaticity in personality, social, and clinical psychology. The general argument is that some attitudes, impressions, and other social judgments, as well as aggression, compliance, prejudice, and other social behaviors, are typically mediated by automatic processes, which operate outside phenomenal awareness and voluntary control. To some extent, what might be called an automaticity juggernaut seems to represent a reaction to a cognitive view of social interaction, which some consider inappropriately emphasizes conscious, rational, and cognitive processes at the expense of the unconscious, irrational, emotive, and conative. In addition, the popularity of automaticity seems to represent a reversion to earlier, precognitive, situationist views within social psychology.

After all, the concept of automaticity is at least tacitly modeled on innate stimulus–response connections, such as reflexes, taxes, and instincts, as well as those acquired through classical and instrumental conditioning. The automaticity juggernaut is not exactly a reversion to Skinnerian behaviorism because it entails internal mental representations and processes intervening between stimulus and response, but it is close: If the cognitive processes underlying social cognition and social behavior are indeed largely automatic, then not too much thought has gone into them.

B. Implicit Memory

Like automaticity, the concept of implicit memory has received a huge amount of attention in the field. A whole industry has developed around implicit memory, involving amnesic and demented neurological patients; dissociative disorders such as psychogenic amnesia, fugue, and multiple personality; children and the healthy aged; depressed patients receiving electroconvulsive therapy; surgical patients receiving general anesthesia or conscious sedation; hypnotized subjects (but apparently not sleepers); and even college students who have all their wits about them.

However, there are some important issues that have to be addressed by further research. For example, almost all the evidence on implicit memory has been collected within a single narrow paradigm, repetition priming, leading to theories of implicit memory that emphasize relatively low-level perceptual processes. However, semantic priming also occurs, not just in posthypnotic amnesia but also in organic amnesia, suggesting that these perception-based theories are not adequate for the phenomenon. Similarly, although explicit and implicit memory are dissociable, they also interact, requiring revisions in theories that hold that these two expressions of memory are mediated by separate memory systems in the brain.

Despite these and other persisting questions, the general acceptance of the distinction between explicit, conscious and implicit, unconscious expressions of memory opens the door to extensions of the explicit–implicit distinction to other domains of mental life.

C. Implicit Perception

By analogy to implicit memory, implicit perception may be defined as the influence of a current event, or an event in the very recent past (what William James called the specious present), on experience, thought, or action, in the absence of conscious perception of that event. Implicit perception subsumes so-called "subliminal" perception, involving the processing of stimuli that are degraded beyond conscious perception by low intensities, brief durations, or masking stimuli. However, it goes beyond the subliminal to include neurological syndromes such as blindsight and neglect,

where the stimuli are in no sense subliminal: They are perfectly visible to everyone but the brain-damaged subject. This is also the case for the conversion syndromes of "hysterical" deafness, blindness, and anesthesia. Similarly, in hypnotic blindness, deafness, anesthesia, and analgesia, the subjects would be clearly aware of the stimuli in question were it not for the hypnotist's suggestion. On the fringes of consciousness are cases of so-called preattentive processing, in which the stimulus in question is nominally supraliminal but escapes focal awareness by virtue of parafoveal presentation or presentation over the unattended channel in the dichotic listening or shadowing paradigm. Thus, the term implicit perception captures a broader range of phenomena than is covered by the term subliminal perception because it covers the processing, outside of conscious awareness, of stimulus events that are clearly perceptible in terms of intensity, duration, and other characteristics. It also has the advantage of skirting the difficult psychophysical concept of the limen.

What all these phenomena have in common is a dissociation between explicit and implicit perception, analogous to the dissociation between explicit and implicit memory: The subject's experience, thought, and action are affected by some event in the current stimulus environment, in the absence of conscious perception. The distinction between implicit perception and implicit memory is not always easy to make because both phenomena are revealed by postexposure priming effects (i.e., by performance on a nominal test of implicit memory). Arguably, however, the term implicit memory should be reserved for cases in which the stimulus event was consciously perceived at the time of encoding; where there is no conscious awareness at the time of encoding, we can consider priming effects as evidence of implicit perception. Thus, on the assumption that adequately anesthetized patients are really unaware, during their procedures, of what our medical colleagues disarmingly call "surgical stimuli," evidence collected in the recovery room of priming effects attributable to events presented during surgery constitutes evidence of implicit perception, not just implicit memory. In contrast, priming effects observed in conscious sedation, in which subjects are fully aware of the study trials, are pure implicit memory effects.

Adopting the implicit–explicit distinction may help resolve a persisting controversy over the scope of unconscious perception. For example, Anthony Greenwald and colleagues asserted that subliminal perception is analytically limited: Some semantic processing of a subliminal stimulus is possible, but not too much. On the other hand, advocates of subliminal symbiotic stimulation—the "mommy and I are one" experiments—assume more processing than Greenwald's arguments would predict; so do those who employ subliminal techniques in advertising and psychotherapy. Most arguments for the analytic power of subliminal processing appear to be based on Romantic or psychoanalytic notions of the unconscious. On the other hand, there are many ways to render percepts implicit, and how this is done may make a major difference in regard to what one can do with them. Philip Merikle and colleagues distinguished between the objective threshold, where all responses to the stimulus decrease to chance levels, and the subjective threshold, where the subject simply does not experience the stimulus consciously. Greenwald's experiments take subjects as close to the objective threshold as possible, and with such degraded presentations it is not surprising that perception, and memory encoding, is analytically limited. When subjects get closer to the subjective threshold, more extensive analyses, still outside awareness, might be possible. In hypnosis, in which the stimulus is in no sense degraded, the possibilities for analysis might be unlimited.

D. Implicit Learning

Continuing the elaboration of the explicit–implicit distinction to other domains, implicit learning can be defined as the acquisition of new knowledge and patterns of behavior through experience, in the absence of awareness of the knowledge or behavior so acquired. As it happens, the term implicit learning antedates implicit memory, having been coined by Arthur Reber in 1967. Reber demonstrated that subjects who studied a set of letter strings generated by a Markov process artificial grammar could distinguish between new grammatical and ungrammatical strings without being able to articulate the grammar in question.

In some respects, the learning of artificial grammars appears similar to the acquisition of syntax in natural language. After all, we are perfectly fluent speakers and interpreters of our native language long before we learn the rules of grammar in elementary school. However, implicit learning has also been observed in a wide variety of other paradigms, including classical and instrumental conditioning, control of complex systems, and the learning of categories and sequential relationships. In each of these cases, the claim is that people's behavior is shaped by prior experience (the

classical definition of learning), even though they are unaware of what they have learned.

As with implicit perception, however, the border between implicit learning and implicit memory is vague. Of course, this is as it should be: Memory provides the cognitive basis for learning in the first place, and whatever is learned has to be stored in memory. However, the problem goes deeper than that. When normal subjects learn an artificial grammar, they certainly remember being asked to study the sample strings, and they may even remember the strings themselves, even if they are unaware of what they have evidently learned about the structure of the grammar. In contrast, when brain-damaged amnesic patients acquire new patterns of behavior from experience, as exemplified by mirror tracing and pursuit-rotor learning in patient H.M., they are amnesic for the whole learning experience. In amnesia, the occurrence of implicit learning also provides evidence of preserved implicit memory, but as in the case of implicit perception the term implicit memory is best reserved for effects that occur in the absence of conscious memory for the original experience. By the same token, implicit learning refers to abilities and patterns of response that are acquired through learning experiences, in the absence of conscious awareness of what has been learned.

A small industry has developed around implicit learning, culminating in the 1998 publication of an entire handbook devoted to the topic. However, the claim of implicit learning remains controversial even after more than 30 years of work. There is continuing debate over whether implicit learning is really unconscious in any meaningful sense of the term. It may be too much to expect subjects to be able to articulate an entire Markov process artificial grammar, but subjects might be consciously aware of enough of the rule to permit them to discriminate at above chance levels between grammatical and ungrammatical strings. Perhaps this debate will be resolved if we pay more attention to the format in which the newly acquired knowledge is represented. Perhaps, as implied by the original artificial grammar studies, the subject acquires a whole system of ''if–then'' productions and this procedural knowledge, like all procedural knowledge, is inaccessible to conscious introspection.

On the other hand, perhaps the knowledge acquired during implicit learning is not procedural at all, but declarative in nature. For example, subjects might abstract from the learning trials a prototype of a grammatical string; alternatively, they may simply memorize the instances on the study list. In either case, they may make relatively accurate grammaticality judgments by consciously comparing test items to the summary prototype or to the specific exemplars they have memorized. In any event, amnesic patients can learn from their experience without remembering the learning experience, and in this sense, at least, implicit learning provides evidence of unconscious influence.

E. Implicit Thought

If the concept of implicit learning is more controversial than those of implicit memory or implicit perception, the concept of implicit thought is even more so. However, the literature contains some favorable evidence. For example, Kenneth Bowers and associates found that subjects can choose which of two problems is soluble without knowing the answer to the soluble one. In a variant on Sarnoff Mednick's Remote Associates Test (RAT), called ''Dyads of Triads,'' subjects were presented with two RAT-like items. The task in the traditional RAT is to generate a fourth word that is associatively linked to the other three. In Bowers's modification, one triad coheres on a common (if remote) associate, and the other does not (at least barring psychotically loose associations). Bowers *et al.* found that subjects could choose the coherent triad at better than chance levels, even when they could not say what the solution was.

Subsequent research has found that soluble RAT items primed their solution words for lexical decision, even when the subject failed to produce the solution. This priming effect, and the effect on choice observed by Bowers, seem to reflect the preconscious activation of a mental representation of the solution to the problem. However, this mental representation is not a percept, and it is not a memory. The mental representation in question is an idea or an image or, more broadly, a thought.

Accordingly, implicit thought may be defined as the influence of some cognitive representation, itself neither a percept nor an episodic memory, on experience, thought, or action in the absence of conscious awareness of that representation.

Implicit thought may well underlie some of the most interesting facets of creative thought. In this view, intuition reflects a priming-based ''feeling of knowing'' similar to what we commonly see in studies of memory, incubation reflects the gradual accumulation of strength of this primed idea, and insight reflects the emergence of the preconscious idea into the full daylight of consciousness.

V. IMPLICIT MOTIVATION AND EMOTION

Along with automaticity, implicit memory, implicit perception, implicit learning, and implicit thought comprise the cognitive unconscious. However, cognition is not all there is to mental life, and so we are led to ask whether there is an affective unconscious and a conative unconscious as well. Of course, emotional and motivational states may arise automatically and in this sense result from unconscious processes. However, can motives and emotions be unconscious in the same way that implicit memories can?

In fact, the late David McClelland and associates articulated a concept of implicit motives—interestingly, without overt reference to the concept of implicit memory. Explicit motivation might be defined as the conscious representation of a conative state or the desire to engage in some particular activity, as represented by a craving for food, yearning for love, and the like. In contrast, implicit motivation refers to any change in experience, thought, or action that is attributable to one's motivational state in the absence of conscious awareness of that state. We are admittedly on the verge of Freudian territory here, but the motives in question are not seething sexual and aggressive impulses arising from the id; they are motives for achievement, power, affiliation, and intimacy.

Moreover, McClelland and colleagues keep us on a firm empirical base. For them, explicit motives are self-attributed: The person is aware of the motive, can reflect on it, and can report its presence in interviews or on personality questionnaires. Implicit motives, in contrast, are inferred from the person's performance on such tasks as the Thematic Apperception Test. On the basis of an extensive program of empirical research, McClelland and associates concluded that explicit and implicit motives influence different classes of behavior and respond to different types of social influence. More work needs to be done, but the evidence indicates that explicit and implicit motives are indeed dissociable, in much the same way that explicit and implicit memories are dissociable.

Regarding the affective domain, it is possible that dissociations occur between explicit and implicit expressions of emotion, just as they occur between explicit and implicit expressions of memory. Again paralleling the vocabulary of the cognitive unconscious, we define explicit emotion as the person's conscious awareness of an emotion, feeling, or mood state; implicit emotion refers to changes in experience, thought, or action that are attributable to one's emotional state in the absence of conscious awareness of that state.

Note that conscious emotional responses can serve as expressions of implicit memory and perception. Most of the relevant studies make use of the mere exposure effect documented by Robert Zajonc, in which exposure to an object increases one's preference for that object on a subsequent choice task. According to one theory, the mere exposure effect is a variant on the priming effect familiar from studies of implicit memory and implicit perception. On the memory side, it seems that brain-damaged, amnesic patients show the mere exposure effect, even though they cannot recognize the objects to which they had previously been exposed. With respect to perception, normal subjects show mere exposure effects even when the exposures in question were subliminal and thus not consciously perceived. In both cases, the subjects preferred previously exposed objects to new ones, even though they had no conscious perception, much less conscious recollection, of the visual stimuli.

With respect to the proposition that people can be unaware of emotional states, which nonetheless influence their ongoing experience, thought, and action, the empirical evidence is regrettably sparse. However, there remain good theoretical reasons for thinking that it might be true. For example, Peter Lang's multiple-systems theory of emotion postulates that every emotional response consists of three components: verbal–cognitive, corresponding to a subjective feeling state such as fear; overt motor, corresponding to a behavioral response such as escape or avoidance; and covert physiological, corresponding to a change in some autonomic index such as skin conductance or heart rate. These three components or systems usually covary, but in some circumstances they can move in different directions—a state that Stanley Rachman and colleagues labeled desynchrony.

Of special interest in the current context is a particular form of desynchrony in which explicit emotion, as represented by the conscious, subjective feeling state, is absent, but behavioral and somatic signs of emotion persist. Currently, most evidence for this desynchrony comes in the form of clinical anecdote rather than controlled experiment, but it is comforting that just such a desynchrony is predicted by a neuropsychological model of fear recently presented by Joseph LeDoux. LeDoux proposes that environmental stimuli are first processed by sensory centers in the thalamus that then pass information about emotional events to the amygdala, which in turn generates appropriate behavioral, autonomic, and

endocrine responses. Information about these responses is also passed to cortical centers supporting working memory, where it is integrated with information provided by thalamic centers about the fear stimulus, thus generating the full-blown subjective experience of being afraid of something. However, if a disconnection between thalamus and cortex prevents the fear-eliciting stimulus from being represented in working memory, the person will experience fear without being aware of the fear stimulus. In this case, emotion will serve as an implicit expression of perception or memory, as described earlier. Alternatively, if there is a disconnection between the amygdala and the cerebral cortex, the person will behave in a fearful manner without feeling fear or anxiety. In this case, there will be a dissociation between explicit and implicit emotion.

Neuroscientific theory aside, a potentially interesting approach to implicit emotion has been offered by Anthony Greenwald and Mahzarin Banaji in their application of the explicit–implicit distinction to the social–psychological concept of attitude. Attitudes are affective dispositions to like or dislike certain things; like motives, they are usually measured explicitly by self-report scales. However, Greenwald and Banaji suggested that people may possess positive and negative implicit attitudes about themselves and other people, which can affect ongoing social behavior outside of conscious awareness. Like McClelland's implicit motives, implicit attitudes are assessed in terms of task performance rather than self-report. When implicit attitudes diverge greatly from their explicitly expressed counterparts, this indicates a dissociation between explicit and implicit emotion. Unfortunately, the experimental literature on implicit attitudes rarely offers a direct contrast between explicit attitudes, so it is unknown whether such dissociations actually occur, in what kinds of people, and in what circumstances. However, this experimental approach to implicit emotion is very promising.

VI. THE UNCONSCIOUS AND THE BRAIN

In the 19th century, Hartmann distinguished between the relative (psychological) unconscious and the physiological and physical unconscious, and we can make a similar distinction between unconscious mental life and unconscious physiological and biochemical processes. Most of our bodily functions proceed unconsciously, without any direct introspective awareness of them or direct voluntary control over them, and this is as true of the nervous system as it is of the other systems in the body. At the molecular and cellular levels of analysis, we have no awareness of the depolarization of cell membranes, the transmission of action potentials down the axon, or the processes of synaptic transmission. We are not aware of the transmission of neural impulses up and down the spinal cord or of the patterns of neural activity that constitute the cortical representation of experience, thought, and action afferent.

We are not aware of the saccades by which we refresh our retinal images or of the opponent processes that allow us to see various combinations of red, yellow, green, and blue. Spinal reflexes are spared in neurological patients who have been rendered paraplegic or quadriplegic by virtue of a complete break in the spinal cord. However, the patients have no direct awareness of these responses, no ability to voluntarily inhibit them, and no ability to initiate similar actions, at will, in the absence of effective stimuli. In the absence of biofeedback technology, we have no awareness of our blood pressure or whether our brains are generating alpha, beta, delta, or theta activity. These physiological and electrochemical processes in the body are as unconscious as photosynthesis and cell division, the tides, and plate tectonics. However, these things are not what "the unconscious" is all about. To refer to them as unconscious, as Hartmann did, is to make what the philosopher Gilbert Ryle called a category mistake—the error of ascribing to one domain a feature attributable only to another. Consciousness belongs to the domain of the mental, and so does the psychological unconscious. It makes no sense to apply the distinction to anything other than mental life.

The revival of interest in consciousness among psychologists and other cognitive scientists has led to a revival of interest in its physiological substrates. A number of neural correlates of consciousness (NCCs) have been proposed, such as activity in the reticular formation or the left hemisphere. However, the reticular formation appears to be responsible for maintaining cortical arousal, which is not the same as consciousness, and careful studies of split-brain patients make it clear that the right hemisphere has a consciousness of its own, even if lacks the ability to communicate its experiences verbally. Recent proposals have similar problems. For example, Francis Crick and Christof Koch proposed that the NCC consists of synchronized firings (e.g., at 40 Hz) of different various cortical neurons representing features of an object; however, it may also be that such firings serve merely to

bind various features of an object together to form a unified mental representation, and that other neural processes determine whether that representation will be consciously accessible. Giulio Tononi and Gerald Edelman proposed that consciousness results from the firings of neurons distributed widely over the thalamocortical system; however, this may be true of any mental representation, conscious or not.

The 19th-century phrenologists did not assign the "faculty" of consciousness to any portion of the brain, perhaps because they viewed consciousness as an intrinsic property of mental life. However, documented dissociations between explicit and implicit cognition, emotion, and motivation offer the possibility of finding portions of the brain that are involved in conscious but not unconscious, or unconscious but not conscious, mental life. For example, Schacter hypothesized the existence of a conscious awareness system (CAS), a brain module that supports conscious awareness in various domains, such as perception and memory. Connection between a module supporting memory and the CAS would render memories accessible to conscious awareness; a disconnection (to use a term coined by Norman Geschwind) between these modules would impair conscious recollection but spare implicit expressions of memory.

Schacter's proposal is congruent with a general principle of the modularity of the mind and brain, but some considerations suggest that there might be multiple CASs, each supporting conscious awareness in different mental domains. Thus, although it appears that the hippocampus plays a special role in explicit memory, it plays essentially no role in conscious vision. The striate cortex (area V1) mediates conscious visual experience but plays no special role in conscious recollection. As noted earlier, LeDoux proposed that conscious emotional experiences (or at least the experience of fear) are mediated by connections between the amygdala and cortical structures supporting working memory, whereas unconscious behavioral and physiological expressions of emotion are mediated by connections between the amygdala and subcortical structures. Perhaps newly emerging brain imaging techniques applied to patients or subjects who display a dissociation between explicit and implicit memory, or the like, will be able to reveal cortical structures that are differentially involved in conscious and unconscious mental life.

On the other hand, it may be that brain imaging approaches will not reveal the neurophysiological differences between conscious and unconscious mental life. For example, it has been proposed that all conscious mental states refer to the self as the agent or patient of some action or the stimulus or experiencer of some state; unconscious mental states lack this kind of self-reference. If so, then conscious awareness is not a matter of the activity of one or more brain modules but rather of a connection of some sort between the mental representation of the action or state in question and a mental representation of the self. It is now believed that such knowledge representations are generated by ensembles of neurons distributed widely across the cortex rather than by specific cortical loci (e.g., "grandmother" cells). If this notion is correct, then brain imaging techniques such as functional magnetic resonance imaging, which are designed to identify specific neural loci involved in one kind of mental activity or another but cannot discriminate between neural representations of specific mental contents, cannot reveal the NCC.

VII. WHITHER THE UNCONSCIOUS?

The initial discovery of the unconscious, which was consolidated at the turn of the 20th century, has been revived, and the process of rediscovery is well along at the beginning of the 21st century. There is incontrovertible evidence for automatic mental processes and for implicit memories. Implicit perception is perhaps less convincingly established, and implicit learning remains controversial as well. However, the evidence favoring both concepts cannot be dismissed out of hand. Research on implicit thought is admittedly immature, but the evidence in hand is quite provocative. Based on the evidence for the cognitive unconscious, implicit motivation and implicit emotion cannot be dismissed out of hand, but we still require convincing evidence that they can be dissociated from their explicit counterparts. However, it is clear that the paradigms developed in the study of implicit memory provide a vehicle for exploring all aspects of the psychological unconscious. In response to Immanuel Kant, we can say that priming and other methodologies do in fact enable us to infer that that we have ideas, even though we are not conscious of them. Also, in response to William James, we can say that these same methodologies, rigorously applied, will prevent us from believing whatever we like about the unconscious mind.

In that respect, it must be emphasized that the scope of the psychological unconscious, broad as it is, does not appear to be so broad as to encompass the unconscious of psychoanalytic theory. There is no

evidence favoring Freud's view that the unconscious is the repository of primitive, infantile, irrational, sexual, and aggressive impulses, repressed in a defensive maneuver to avoid conflict and anxiety. Nor is there any evidence to support the more extreme clinical lore concerning unconscious representations of trauma or the excesses of the recovered memory movement in psychotherapy. In this case, as James warned, the unconscious does indeed seem to be a tumbling ground for whimsies.

See Also the Following Articles

COGNITIVE PSYCHOLOGY, OVERVIEW • CONSCIOUSNESS • CREATIVITY • DREAMING • INHIBITION • MEMORY, EXPLICIT AND IMPLICIT • NEUROPSYCHOLOGICAL ASSESSMENT • PRIMING • PSYCHOPHYSIOLOGY • RECOVERED MEMORIES

Acknowledgment

Preparation of this article was supported by Grant MH-35856 from the National Institute of Mental Health.

Suggested Reading

Bargh, J. A. (1997). The automaticity of everyday life. *Adv. Social Cognition* **10,** 1–61.

Bornstein, R. F., and Masling, J. M. (Eds.) (1998). *Empirical Perspectives on the Psychoanalytic Unconscious.* American Psychological Association, Washington, DC.

Bornstein, R. F., and Pittman, T. S. (1992). *Perception without Awareness: Cognitive, Clinical, and Social Perspectives.* Guilford, New York.

Conway, M. A. (Ed.) (1997). *Recovered Memories and False Memories.* Oxford Univ. Press, Oxford.

Kelly, W. L. (2001). *The Psychology of the Unconscious: Mesmer, Janet, Freud, Jung, and Current Issues.* Prometheus, Buffalo, NY.

Kihlstrom, J. F. (1997). Consciousness and me-ness. In *Scientific Approaches to Consciousness* (J. D. Cohen and J. W. Schooler, Eds.), pp. 451–468. Erlbaum, Mahwah, NJ.

Kihlstrom, J. F. (1999). The psychological unconscious. *Handbook of personality: Theory and research* (2nd ed.). pp. (424–442). The Guilford Press, New York.

Levy, D. (1996). *Freud Among the Philosophers: The Psychoanalytic Unconscious and Its Philosophical Critics.*Yale Univ. Press, New Haven, CT.

Reber, A. S. (1993). *Implicit Learning and Tacit Knowledge: An Essay on the Cognitive Unconscious.* Oxford Univ. Press, New York.

Reder, L. S. (Ed.) (1996). *Implicit Memory and Metacognition.* Erlbaum, Hillsdale, NJ.

Shevrin, H., Bond, J. A., Brakel, L. A. W., Hertel, R., and Williams, W. J. (1999). *Conscious and Unconscious Processes: Psychodynamic, Cognitive, and Neurophysiological Convergences.* Guilford, New York.

Stein, D. J. (Ed.) (1997). *Cognitive Science and the Unconscious.* American Psychiatric Press, Washington, DC.

Underwood, G. (Ed.) (1996). *Implicit Cognition.* Oxford Univ. Press, Oxford.

Weiskrantz, L. (1986). *Blindsight: A Case Study and Its Implications.* Oxford Univ. Press, Oxford.

Weiskrantz, L. (1997). *Consciousness Lost and Found: A Neuropsychological Exploration.* Oxford Univ. Press, New York.

Unilateral Neglect

ELI WERTMAN
Hadassah Medical School and Hebrew University, Jerusalem

KENNETH M. HEILMAN
University of Florida College of Medicine

GLOSSARY

agnosia Loss of comprehension at the level of the central nervous system of any of the senses; the sensory sphere is intact but the patient is unable to assimilate the meaning of the sense.

dyslexia A disorder, occurring more frequently in males, characterized by the inability to read, spell, and write words despite the ability to see and recognize letters; a typical manifestation is the reversal of letters within a word.

somatosensory of or relating to bodily sensations not necessarily associated the eyes, ears, tongue, and other primary sense organs.

Neglect is a deficit in spatial distribution of attention or exploration that is manifested in the failure to report, consciously respond, and/or orient to novel or meaningful stimuli, when this failure cannot be attributed to either sensory or motor defects. Neglect may be spatial or personal: One may be neglecting stimuli in space or on the person and one may fail to act in a portion of space, in a spatial direction, or on a portion of one's body. The neglected stimuli are usually situated in the side of space that is opposite a brain lesion (contralesional) but might be on the same side (ipsilesional). This disorder was first described in the last 25 years of the 19th century. Most of the early reports were clinicoanatomical studies of unilateral horizontal right/left, and many of the behavioral manifestations of neglect were defined throughout this period. During the past 25 years, neglect has been the subject of a multitude of studies. It is difficult to integrate all these observations into a unified model. In this article, we first discuss the elementary behavioral components of neglect and related disorders. We will then introduce some recent observations, and finally we attempt to integrate these observations into a unified model.

I. THE CLINICAL SYNDROMES

The major behavioral manifestations of neglect include the syndromes discussed in the following sections.

A. Sensory Neglect

Sensory neglect, or inattention, refers to a deficit in awareness of contralesional stimuli that is not related to afferent defects. Inattention might be found in one or more sensory modalities—visual, tactile, and

auditory. Sensory neglect must be differentiating from sensory disorders. This is easier for auditory inattention because the auditory pathways ascend to the cortex bilaterally, and unilateral brain lesions cannot impair hearing in one ear. Thus, the inability to hear stimuli on one side is usually due to inattention. Inattention is more difficult to discriminate from sensory deficits in the visual and tactile modalities. Visual hemiinattention may be dissociated from hemianopia by having subjects move their eyes to the ipsilesional side of space. Similar methods can be used to distinguish tactile inattention from a true sensory defect.

B. Motor Neglect

Motor neglect (action intentional) is often seen in patients who are aware of contralesional stimuli but fail to respond normally to these stimuli (exo-evoked akinesia). Even in the absence of weakness, patients may have aspontaneity of the contralesional arm. Motor neglect might present not only as a failure to initiate movement (akineia) but also as a delay, slowing of initiation (hypokinesia), or inability to sustain (impersistence) of a movement. The body parts that are involved in the motor neglect might differ between patients and include the eyes, the head, a limb, or the whole body. The akinesia might be directional (e.g., there is a reluctance to move in a direction that is contralateral to the lesion) or spatial (e.g., the contralesional arm is more akinetic in the contralesional hemispace independent of the direction of movement).

C. Personal Neglect

Patients with this syndrome might not interact with the contralesional parts of their body. For example, they do not dress or groom one side (e.g., shaving or combing only the ipsilesional side of the face and hair). A patient might only put the right half of his or her body on a chair. Patients with personal neglect might have asomatognosia, which is the failure of a patient with personal space neglect to accept ownership of his or her contralesional limbs. Patients might think that their limbs are owned by someone else, even when confronted with objective evidence that they are mistaken (e.g., following the course of the limb from its distal end toward the point where it is attached to the patient). In severe cases, the limb might be personified (e.g., related to as a person or a thing) or be the object of hatred (misoplegia). More frequently, in milder forms of personal neglect, patients will be aware that their extremities belong to them, but they still refer to them as though they were objects. Contralesional supernumerary limbs have also been described with this syndrome.

D. Spatial Neglect

Spacial neglect is a failure to respond to stimuli in hemispace. Contralesional hemispace is most often involved. Spatial neglect may occur in all three directions of space; horizontal, radial, and vertical.

E. Imagery Neglect

When asked to image an object, scene, or part of the body, some patients fail to image one half. Although patients with imagery neglect often have hemispatial neglect or asomatagnosia, imagery neglect and spatial or personal neglect can be dissociated. Imagery neglect, like spatial neglect, can be viewer, environmentally, or object centered. For example, patients might have difficulty spelling from memory the neglected side of a word, both forwards and backwards.

F. Memory Neglect

In memory neglect, patients who can recognize contralateral stimuli may have difficulty recalling them.

G. Extinction

In this disorder, patients detect and respond to the contralesional stimuli when they are presented alone, but the patients are unaware of these stimuli when presented simultaneously with ipsilesional stimuli. Extinction might be sensory (e.g., failure to detect a contralesional stimulus during bilateral simultaneous stimulation in the visual, tactile, auditory, and even olfactory modalities). Extinction might be found either in patients who recover from inattention or in patients who never had inattention. Patients with motor extinction can detect bilateral simultaneous stimuli but cannot move both sides of the body simultaneously.

II. EVALUATION PROCEDURES

As a result of the number and heterogeneity of the behavioral deficits associated with the neglect syndromes, there are an enormous the number of possible tests. In the following sections, we describe the most important evaluation procedures.

A. Observation of Spontaneous Behavior

Spontaneous directional orienting of the eyes, head, or body should be observed to determine if there is a directional bias. There may be a decreased amount of movement of the limbs on one side. Daily activities are usually very informative. For example, patients might be able to communicate with visitors and family only from one side. They might eat and drink only from one side of the tray or plate. Patients might neglect dressing and grooming (e.g., shaving and combing) in this same side.

B. Tests for Sensory Inattentional Tasks

Inattention is tested by asking the patient where he or she was stimulated. This should be done in the visual (either in bedside confrontation tests by using a cotton-tipped applicator or a moving finger or in the laboratory by perimetry or a tangent screen), auditory (either by rubbing or snapping the fingers or by audiometric techniques), and somesthetic (either by touching with a finger or a cotton applicator or by Von Frey hairs). Stimuli of other somatosensory modalities (pin, temperature, etc.) can also be used.

C. Motor Intentional Tests

Hypo- or akinesia in one or more body parts (eyes, head, body, or limbs) should be examined by observing spontaneous behavior and behaviour in response to commands. Failure to move or a delay in movement in the face of relatively preserved motor strength suggest motor neglect. Crossed-response tasks might be used to differentiate motor-intentional neglect from elemental sensory disorders and sensory inattention.

D. Extinction Tasks

If the patient responds normally to unilateral stimuli, he or she is intermittently presented with bilateral stimuli and asked either to verbally report the side of the stimuli or to move the stimulated body part (e.g., hand or arm). The bilateral stimuli should be randomly mixed with unilateral stimuli to the right and left sides. Correct response to the unilateral stimuli with contralesional unawareness in the bilateral stimulation suggests contralesional sensory extinction. A failure to move or a delay in moving the contralesional limb with preserved verbal reports of bilateral stimulation is a sign of motor extinction.

E. Spatial Neglect Tasks

The following are among the many tests that can be used to assess spatial neglect:

1. Cancellation tasks: Lines circles, letters, or stars are drawn in random positions on a sheet of paper and presented to patients who are asked to cancel or cross out all the targets. Spatial neglect patients may fail to cancel targets on the contralesional side.
2. Line bisection task: Patients are asked to bisect or find the middle of lines 6–10 in. long. Patients with spatial neglect make their mark to the ipsilesional side of the real midline. When copying a figure, patients with neglect might omit the contralesional parts of the figure (e.g., omitting the contralesional petals while copying a daisy).
3. When drawing spontaneously, a patient with spatial neglect might omit the contralesional parts of the drawing (e.g., in drawing one half of a daisy). When given a frame to fill with typically related data, such as numbers on an empty clock face or cities and states on an empty map, the patient might omit the contralesional data or draw all the data on the ipsilesional side. Norms are available for a few of these.

F. Representational Neglect

Representational neglect can be detected by asking the patient to recall details of mental images (e.g., places, geographical areas, or familiar routes) from a specific point of view and comparing the recalled details on each side. The same might then be done from oppositely located points of view.

III. EPIDEMIOLOGY OF NEGLECT SYNDROMES

The following data in this section concern patients with acute stroke.

Contralesional neglect: The frequency of contralesional visual spatial neglect has been investigated

extensively. Recent studies have shown: the frequency of visuospatial neglect to be significantly higher after right (70–82%) compared to left (40–65%) hemispheric lesions. These rates are higher than those indicated by previous studies, probably because of better definition of the syndrome.

Extinction: The frequency of visual extinction was found to be 23% with right and 2% with left hemispheric lesions. Tactile extinction was found in 65 and 35% of patients with right and left hemispheric stroke, respectively. Another study found the rate of visual extinction after right hemispheric stroke to be 9% and that of tactile extinction to be 14%. An additional 6% had combined visual/tactile extinction. The difference in the rates between the two studies seems to be related to the differences in the criteria of extinction, but in both studies the frequency of tactile extinction is higher than that of visual extinction.

Ipsilesional neglect: The rate of ipsilesional neglect was found in one study to be approximately 16% of right hemispheric lesioned patients.

Clustering of the neglect phenomena: It was found that between 49% and 73% of neglect patients present three or four coexisting neglect features. In addition, many double dissociations were found between these features. For example, hemiinattention was dissociated from all other components of neglect.

IV. HETEROGENEITY OF BEHAVIORAL DEFICITS ASSOCIATED WITH THE NEGLECT SYNDROME

Previously, we discussed many of the behavioral deficits associated with the neglect syndrome. Several studies have revealed that although some patients may have all or most of these deficits, others have only one or some deficits but not others, suggesting that many of these behavioral deficits are dissociable. In the following sections, we present these clinical dissociations.

A. Hemineglect Syndromes

1. Representational/Non-Representational Neglect

Most of the patients that manifest representational neglect also have some other hemispatial perceptual deficits. Thus, they do not fully show that representational and perceptual neglect are dissociated. However, recent studies report cases in which patients had representational but not perceptual neglect or the opposite.

2. Attentional/Intentional Neglect

The failure to cancel contralesional stimuli and the error toward the ipsilesional space in bisecting a line may be related to either a failure to allocate attention to contralesional hemispace, with a resulting failure to fully perceptually process the sided stimulus (a sensory-attentional deficit), or a failure to initiate or execute movements toward or in the neglected hemispace (a motor-intentional deficit).

Several methods have been used to test this dissociation in patients with neglect. Tests for motor interference reduce the need for blindfolded manual exploration and also reduce measuring time to initiate movement (moving a handle along a track with the right hand) as well as other indirect time measurements. Investigators have tried to create conditions in which attentional and intentional conditions are dissociated. Such paradigms include using a pulley device, a mirror, or closed-circuit TV. For example, a recent study evaluated 10 patients with right hemispheric injury and left hemispatial neglect in both line bisection and cancellation tasks while they viewed stimuli on closed-circuit TV. The patients could only view the stimuli and their hands indirectly through a closed circuit TV monitor. Direct viewing was precluded by a drape. Manipulation of the video camera created two different conditions. In the direct condition the left and right sides of the paper in actual work space were on the left and right sides of the monitor scan. Hand movements on the table (work space) were congruent with those displayed on the monitor (perceptual space). Inverting the camera 180° created the indirect condition, in which the left side of the paper projected to the right side of the monitor and vice versa. Consequently, leftward movements in the work space appeared as rightward movements in perceptual space and vice versa. In the direct condition of the line bisection task, all subjects bisected, as anticipated, to the right of the midline. The critical condition for discriminating between attentional and intentional factors is a comparison between indirect conditions. If the patient has primarily an attentional deficit or bias, then in the direct condition this would have a rightward bias but in the indirect condition the bias would reverse and be leftward. Rightward deviations on the actual line during both conditions suggest a failure to move leftward (intentional deficit). Whereas patients with temporoparietal posterior lesions had primarily attentional deficits, patients with frontal lesions had primarily intentional deficits. Patients with primarily

attentional or intentional deficits often had secondary intentional or attentional deficits. The attentional–intentional dissociations were also found in the auditory modality.

3. Personal vs Extrapersonal (Spatial) Neglect

Personal space refers to the body and extrapersonal space refers to the space outside the body. Extrapersonal space is divided into near (peripersonal) and far. Personal and spatial neglect were found to double dissociate. Whereas spatial neglect in the absence of personal neglect is common, the opposite is less common.

4. Domains of Spatial Neglect

Neglect might occur in different spatial domains. Several dissociable subtypes have been reported. First, hemineglect is commonly reported in the horizontal direction (right/left axis); however, neglect has also been observed in the radial (near or far) and vertical (high or low altitudinal) directions of space. Horizontal neglect may also be related to the position of the stimulus in the peripersonal versus extrapersonal space. For example, some patients will fail to correctly bisect horizontal lines in near space but perform correctly in far space and vice versa. In diagonal neglect, patients frequently fail to cancel targets primarily in the left part of the page nearest the body. This seems to be related to a combination of horizontal and near radial neglect. Second, neglect is reported in different reference frames. In the egocentric reference frame, neglect is relative to the body of the viewer (i.e., left or right of the viewer). Within the body-centered coordinates, there might be a dissociation between retinotopic; head- and trunk centered-neglect. An allocentric or environmental reference frame locates an object in relation to its spatial environment. For example, the stimulus might be located in the left of a scene, independent of the viewer reference frame. Egocentric and allocentric neglect can be dissociated by changing the position of the patient relative to the environment. An object-centered reference frame codes intrinsic spatial aspects of the object relative to the object's principal axis. The object-centered reference frame remains stable when the stimulus is moved in relation to the body or environment. By manipulating the relative position of the patient's body, visual scene, and object neglect could be shown in each of these reference frames. Third, some types of neglect are stimulus specific. For example, neglect was shown to the left side of the face or to either the right or left side of a word. Complex objects or scenes have global and local features. In Navon's figures, a larger (global) form is composed of smaller (local) forms (e.g., a large "H" and small "o"s). There are patients with left hemispatial neglect who might fail to recognize either the global or the local elements on the neglected side.

a. Auditory Neglect Although auditory spatial neglect is not uncommon, there is a paucity of reports in the literature on the subject. Also, most of the reports are of auditory extinction (with its unresolved connection with neglect).

b. Tactile Neglect The dissociable described subtypes are those that are related to orthogonal axes of neglect. Here also, near and far radial and high and low altitudinal neglect are seen.

c. Olfactory Neglect This was reported in right parietal lesioned patients. Olfactory detection was comparable to that in controls. Both lateralization and identification of unilaterally presented odorants were decreased on the left contralesional nostril.

d. Motor Neglect The following dissociations have been reported for motor neglect:

1. Orthogonal axes of neglect: These dissociations might be attentional, representational, or intentional. In order to show a specific dissociation of motor neglect along the orthogonal axes, specific paradigms should be used. Using a fixed-window paradigm, a near radial intentional neglect was identified.
2. Involved part of the body: This might be eyes, head, limbs, or the whole body.
3. Type of action: For example, writing (producing errors on the contralesional side of the word), reaching with the ipsilesional limb, and manual exploration.
4. Direction of movement: Directional or spatial deficit.
5. Part of movement: There might be difficulty in initiating (akinesia and hypokinesia) or in maintaining (impersistence) a movement.

e. Contra/Ipsilesional Neglect Whereas most patients with viewer-centered or environmental-centered neglect fail to be aware of or act on contralesional stimuli, some patients with neglect fail to recognize or act on ipsilesional stimuli.

B. Extinction Syndromes

Neglect and extinction are clinically related phenomena. They frequently coexist in the same patient. Also, extinction appears when there is an improvement in the signs of neglect. There is even a tendency to view extinction as "mild neglect." Although these syndromes are related, there are significant differences between them. The main similarity is the failure to report, respond, or orient to a contralesional stimulus, with the difference being the bilateral stimulation condition in the extinction syndrome. Many of the dissociations found in spatial neglect have counterparts in the extinction syndrome, but dissociations as representational/nonrepresentational and personal/extrapersonal have not been reported in extinction, maybe because extinction is of stimuli in space and not of space. The following dissociations have been reported:

1. Motor/sensory: Showing extinction in either sensory or motor tasks.
2. Within/between modalities: Usually, extinction is reported to be between two stimuli of the same modality. Recently, cross-modal extinction was reported between visual and tactile stimuli.
3. Orthogonal axes Vertical extinction has been reported. That the distinction between near and far peripersonal spaces is also relevant for extinction is suggested by the observation that modulation of visuotactile cross-modal extinction is found when both stimuli are in the near peripersonal space but not when only one of them is in the near peripersonal space and the other is in the far peripersonal space.
4. Reference frames: The previously mentioned ego/allocentric distinction is not described in the extinction literature. However, it seems that there is a place for such a dichotomy. For example, double simultaneous tactile stimuli were given to the ipsilesional thumb and little fingers of the ipsilesional hand, in the ipsilesional hemispace, when the palm was up or down. In the two cases the medial contralesional stimulus was extinguished. This means that some allocentric reference frame is affecting the extinction. Also, posterior truncal tactile extinction was found in parkinsonian patients.
5. Hemispatial location of stimuli: Extinction might occur not only when stimuli are presented in both hemispaces but also when they are in the ipsilesional hemispace. This was shown for visual and for tactile domains.
6. Type of stimuli: For example, patients with visual extinction were found not to extinguish stimuli presenting motion.

V. ISSUES WITH SIGNIFICANCE FOR THE MECHANISMS OF NEGLECT SYNDROMES

A. Variability of Neglect Due to Task-Specificity

In addition to the previously mentioned variability of the different syndromes of neglect, its appearance and degree are affected by the demands of the task by which it is evaluated. For example, the line bisection and cancellation tests, both used to diagnose neglect, may show a double dissociation. In a study that investigated 120 patients with neglect after right hemispheric infarcts, 14 of 83 patient who were impaired in a letter cancellation task had no neglect on line bisection, whereas 29 of 37 patients who cancelled normally showed signs of neglect in a bisection task. Interestingly, the side of the neglect relative to the side of the lesion was also found to dissociate with these tasks. Some patients were reported to show contralesional hemineglect on cancellation and ipsilesional hemineglect on line bisection tasks. These task-specific dissociations are explained by specific task demands (e.g., in terms of needed computations of target location or for intention, moving attention, etc.) and locations.

Several factors were found to affect the degree of neglect even within the same task, including location of the target (lesser neglect in ipsilesional bisection), length of the line (more neglected on longer lines), higher discrimination demands (e.g., in cancellation), levels of perceptual parsing (e.g., in copying complex drawings), hand used for accomplishing the task, its starting point, and the type of response (e.g., pointing vs grasping the midline).

Drawing tasks were also found to be affected by general factors, such as verbal IQ and simplification of instruction. Fluctuations in visual neglect were related to fatigue, time of day, and other nonspecific factors. This variability suggests that neglect is not a unitary syndrome, and that its components in a specific patient reflect the specific neuroanatomical injury.

B. Amelioration of Neglect

Neglect can be temporarily ameliorated by a number of procedures. These treatments can be generally

classified according to their approach and possible mechanism. The first group includes vestibular (by irrigation of the left external ear with cold water or the right ear with warm water, inducing nystagmus movements of the eyes, with the slow phase to the left), optokinetic (by horizontally moving bars to the left, also inducing nystagmus with the same features), posterior neck proprioceptive (by vibratory stimulation to the left side), and transcutaneous somatosensory electrical (to the left side of the neck) stimulation. Such stimulations were found to improve visuospatial, motor, and personal neglect, anosagnosia, and representational neglect. The time range of the improvement is between the duration of the stimulation and about 20 min. The exact mechanisms of the effects of these manipulations are not understood.

Other procedures include altering perceptual components (e.g., perceptual cueing) and moving stimuli from the neglected to the normal side by using prisms, reorienting attention to the neglected side by motor–intentional manipulations (e.g., looking to the neglected side, limb activation, both contra- and bilateral, and provoking an intention to act; changing the relative positions of the body-centered and environmental-centered reference frames; and increasing arousal), and transcranial magnetic stimulation of the unaffected hemisphere. These procedures altered performance in a variety of neglect tasks and their effects cannot be explained by a single mechanism.

C. Neuroanatomical Correlates

Neglect is often associated with right than with left hemisphere lesions. The mechanisms for this asymmetry might be related to right hemispheric dominance for arousal, space representation, and attentional and intentional computations.

Analysis of anatomical correlations reveals that left hemispatial neglect can be associated with both cortical and subcortical lesions. The cortical lesions are mainly posterior (e.g., inferior parietal) or both anterior and posterior (e.g., frontal and inferior parietal). However, neglect might be associated with isolated anterior (frontal) lesions. The subcortical structures that neglect when injured include the basal ganglia, the thalamus, and subcortical white matter. Lower diencephalic and upper mesencephalic lesions were also reported to cause neglect both in human and in animals. The size of the lesion does not always correlate with the degree of the neglect. Specific behavioral deficit may be associated with specific brain lesions. Near radial and lower vertical neglect were found after bilateral posterior parietal, whereas far radial and high vertical neglect were found after bilateral inferior temporal lesions. Neglect on the cancellation task is more likely to be associated with frontal lesions and that on line bisection with parietal lesions. Intentional (motor) neglect and ipsilesional neglect are also more frequently associated with frontal than parietal lesions. Sensory-attentional neglect is more likely to be associated with parietal lesions. Whereas spatial hemineglect is associated with inferior parietal lesions, visual extinction seems to be more associated with superior parietal lesions.

VI. UNAWARENESS OF NEGLECTED/EXTINGUISHED STIMULI

When extrapersonal or personal hemispatial neglect are severe, unawareness seems general, as if the contralateral sides of the world or of the body do not exist. In less severe cases, the picture is usually mixed: The unawareness is more variable. The patient might be aware of certain stimuli but not others. The picture is even more surprising in extinction since the general behavior of the patient (if having only extinction) does not include any signs of unawareness except during double simultaneous stimulation. Although most patients insist on seeing only a single stimulus, some observations suggest that in some circumstances the ''extinguished'' stimulus is perceived. The unawareness of neglect/extinction syndrome should be differentiated from other disorders associated with unawareness, including blindsight and simultanagnosia. Blindsight refers to the ability to perform some visual (e.g., detecting movements) and visuomotor tasks (e.g., pointing or grasping a target) in the areas of visual scotoma that resulted from damage to the primary visual cortex. The patient does not have a visual experience of the stimulus on which he or she acted. In simultanagnosia, the patient has difficulty simultaneously perceiving more than a single object.

Another type of deficit of awareness in some patients with neglect might be described as ''hyperawareness,'' in the form of completion phenomena. The patient copies or draws from memory only the ipsilesional half of a drawing (e.g., a butterfly with only the left wing) but believes that what he or she drew is the complete picture (e.g., a butterfly with two wings). Directing his or her attention to the neglected side or to the missing parts might not help. This

phenomenon might provide an indication of the veridicality of the patient's visual experience despite the deficit of awareness.

A. Implicit Processing of Contralesional Stimuli in Neglect

To be aware of contralesional stimuli does not mean that these contralesional stimuli are not processed. Several studies report that even in the presence of unawareness there may be a significant contralesional processing.

The processing of contralesional stimuli in hemineglect was suspected in patients with left neglect dyslexia who make errors that affect the left side of the word but maintain the length of the misread word (e.g., "pillow" for "yellow"), showing an implicit detection of the full word. Using a forced-choice paradigm, patients were able to correctly select between two pictures that were identical on the normal side but different on the neglected side (e.g., two identical houses with red flames coming out of the left side of one house). Patients were aware of the right but not the left side of stimuli (e.g., only the kangaroo in a picture that was contained a deer on the left and a kangaroo on the right side) even after correctly tracing on command, the silhouette of the whole picture. Patients who neglected pictures of objects on the left side of a display claimed to have already seen a significant number of them when these pictures were later presented intermixed with foils in a forced-choice paradigm. The presence of a significant context in the neglected contralesional side of a visual scene increased the range of exploratory eye movements toward the neglected hemifield, despite unawareness of that side. Electrophysiological and psychophysiological measurements indicated that contralesional visual and somatosensory stimuli were perceived, although the patients were unaware of these stimuli.

These observations raise the following question: What is the level of processing that the contralesional stimulus of which the patient is unaware is achieving? The study of patients with visual object agnosia and patients with naming disorders has helped to identify two levels of processing that are essential for recognizing and naming visual stimuli. The first and more elementary is the perceptual level, in which the sensory components of the presented stimulus are determined (e.g., color and form). The output of the perceptual level is transferred to the semantic level, which attaches the meaning to the stimulus based on the knowledge represented in the brain. An appropriate specific response to the stimulus (e.g., naming and motor act) as well as full awareness to it are dependent on the completion of the processing in these two levels. These principles are also valid for the auditory and the somatosensory stimuli.

Several studies have been performed to define the level of processing that is achieved by the contralateral stimulus of which the patient is unaware. Perceptual presemantic processing was reported in some studies. For example, the mere presence and the number of contralesional stimuli were found to affect ipsilesional reaction time and the degree of neglect. The segregation of figures from their background was found to occur on the neglected side. These studies indicate that the stimulus of which the patient is unaware is perceived to a level that it can modulate the severity of the neglect. Semantic processing was also reported in the neglected hemifield in semantic priming (e.g., where the contralesional stimulus should be processed to a semantic level in order to modify the response to an ipsilesional stimulus) and other tasks.

Reports of extinction indicate that it mainly occurs on the perceptual level. They include modulation of extinction by contralesional identical and similar stimuli and also contralateral stimuli that affect grouping processes. Perceptual processing of extinguished stimuli was shown by other paradigms. The anatomical level of processing of the extinguished stimulus was investigated by event-related functional magnetic resonance imaging (fMRI) in a patient with right inferior parietal lesion and left–sided extinction. Even when the left object was extinguished in the double simultaneous presentation, fMRI showed activation of the striate and early extrastriate cortex of the extinguishing hemisphere. In another study, it was shown that the contralesional processing of the extinguished stimulus occurs to the level where it affects motor responses to the ipsilesional stimulus, namely, an advanced perceptual level. Contralesional processing was also reported to extinguished somatosensory stimulus.

In summary, both in neglect and in extinction, there is a significant processing of the neglected stimulus. Any hypothesized mechanism (e.g., attentional or representational) for these syndromes should relate to the phenomenon of unawareness in the face of an advanced level of processing of the neglected stimulus.

VII. MECHANISMS OF NEGLECT AND EXTINCTION

Many mechanisms have been hypothesized to explain the neglect/extinction syndromes, including perceptual, specific representational, attentional, and executive mechanisms. Although neglect and extinction impair awareness, to some extent they might have different mechanisms because they are dissociable both clinically and anatomically.

The clinical heterogeneity of the syndromes and the abundance of related observations have resulted in a multitude of possible mechanisms. Although many of the posited mechanisms are supported by research and can explain several clinical observations, their relation to the core deficit of the neglect/extinction syndromes, which is the unawareness of contralesional stimuli, is not always explicit. Therefore, the suggested mechanisms should be understood as parts of a modular system that creates awareness in general, explain the spatial localization of the deficit, and offer a basis for the heterogeneity and specific dissociative features of the relevant syndromes.

Awareness, as a phenomenon of subjective consciousness, is associated with several cognitive domains, including arousal, the nature of the perceived stimuli and its processing, binding, space representation, selective attention, and working memory. Unawareness of unilateral stimuli might result from a deficit in any of the previously mentioned domains or a combination of them.

A. Explanations of Extrapersonal Hemineglect

The main explanations of the hemispatial unawareness in neglect include attentional, intentional, and representational mechanisms. These mechanisms are not mutually exclusive, and they might even be related or interact with each other.

1. Mechanisms Associated with Asymmetric Deployment of Attention

The attentional deployment bias mechanisms are based on the assumption that the unawareness is caused by a difficulty in allocating the appropriate level of attentional resources to process the presented stimuli. There are two main theories to explain the nature of these deficits. The attention-arousal theory claims that each hemisphere is responsible for orienting responses to the contralateral hemispace. In addition, the right hemisphere has dominance in mediating bilateral cerebral activation and arousal. In left-sided neglect, the lesioned right hemisphere is less aroused than the nonlesioned left hemisphere. As a result, there will be hypoorienting to the contralesional hemispace. Evidence supporting this theory includes the effect of the increased arousal procedure to ameliorate neglect, the effect of contralesional novel stimuli, hypometabolism of the contralesional left hemisphere in neglect patients, and the failure of right neglect patients to show pupilary dilation when they look at the nonneglected stimuli in the ipsilesional left hemispace. The ipsilesional attentional bias theory claims that each hemisphere activates contralateral orienting through corticotectal connections. Each hemisphere inhibits the opposite one through callossal connections. When one hemisphere is injured, the other (the left, in the case of left hemispatial neglect) becomes hyperactive and attention is biased contralaterally (to the right hemispace in this case). This mechanism is indicated by the "magnetic attraction" of gaze to the right, the exacerbation of neglect by general activation of the left hemisphere (e.g., by a language task), amelioration of neglect by decreasing saliency of right-sided stimuli (e.g., when canceling by erasing and not by drawing over), detecting rightmost targets faster than controls, the lack of effect of dimensions of distribution of attention (e.g., entire display vs rightmost position), and the presence of a steep gradient from left to right (as opposed to sharp discontinuity) in a visual search task.

2. Mechanisms Associated with Asymmetry of the Spatial Distribution of Intention

The process of acting on stimuli requires setting the goal of the action and preparing, initiating, executing, and monitoring the action. Patients with intentional hemispatial neglect fail to act on contralesional stimuli of which they are aware. Some methods used to differentiate between intentional and attentional neglect were discussed previously. Spatially directed intention is an integrative process that incorporates multimodal extroceptive stimuli, internal drives, and goal selection and develops coordinated output programs. The patient with the intentional deficit does not have the normal level of mental effort to support these action-related processes. The intentional deficit might affect the different components of the action-producing processes. It might be presented either in simple (e.g., pointing) or in complex actions, in highly

voluntary or automatic actions, or when the intentional demands are phasic or tonic.

Neglect intentional deficits are associated with lesions in dorsolateral and medial (SMA and anterior cingulate gyrus) frontal lobe, inferior parietal lobe, thalamic (ventrolateral, anterior lateral, and medial nuclei), and basal ganglia (striatum and substantia nigra) regions of the ipsilesional hemisphere. Based on the transmodal gateway approach, every region should have a specific role in the intentional system (these roles are not known). Although the net result of the deficit is a decrease in the level of contralesional deployment of intention, it is not known whether this bias is related to the hypointentional level of the stimulus in contralesional space or the hyperintentional level of the stimulus in ipsilesional space.

3. Mechanisms Associated with Disturbance of Shifting Spatial Attention

These mechanisms are based on the three-stage model of attention. According to this model, in order to shift attention from one spatial location to another, one first disengages attention from a currently attended stimulus, moves attention to the new stimulus, and engages that stimulus. These stages can be dissociated using a valid/invalid cueing paradigm, and disengagement deficit was found in patients with right parietal lesions.

In addition to the spatially asymmetric attentional operations described previously, some nondirectional attentional deficits have also been reported in neglect, including a smaller floodlight for distributed attention, a smaller attentional focus, and a nonspatial attention deficit (reflected by a strong relationship between the degree of neglect and digit span discrepancy). A global effect was found in a patient with left hemispatial neglect. The patient was requested to cancel targets of a standard cancellation task in the left and right side alternatively. The patient was able to overcome her right-sided spatial bias, but she did not cancel more targets even though she was successful in overcoming the bias. These results suggest a limited capacity system that might be either attentional (e.g., rapid habituation) or intentional.

4. Mechanisms Associated with Space Representation

The syndrome of representational neglect is based on the assumption that the environment is internally represented as a spatial map in the brain. This mental map enables the brain to calculate spatial coordinates for both perceiving (attending) and acting (intending) on our environment by relating stimuli to spatial referent. A critically localized brain lesion might damage specific parts of these space representations. As a result, selected locations for mental processing will be impaired (representational scotoma). Neglect of left hemispace after right posterior parietal lesion might be caused by a representational deficit.

During the past decade, neuropsychological research both in animals and inhumans has advanced our understanding of the role of the parietal cortex in constructing multimodal distributed representations of space. Such representations should be able to supply the subject with a stable spatial representation for daily functioning. Paradoxically, the parts of space to which we are exposed are continuously changing during normal perceptual experience. For example every movement of our eyes while making saccades toward stimuli in the environment activates different retinotopic neurons. However, the space around us remains stable. The posterior parietal cortex has the processing mechanisms to solve this paradox. Studies in monkeys indicate the following:

1. Multiple modalities are converged in the parietal cortex, including visual, somatosensory (e.g., neck, arm, and eye positions), auditory, vestibular, and efference copies of motor commands. A high proportion of its neurons response to bimodal stimulation [e.g., visual/tactile in the ventral intraparietal area and visual auditory in the lateral intraparietal area].
2. Specific areas are associated with spatial neuronal computations of specific activities (e.g., the LIP area, related to space exploration by eye movements, and the medial intraparietal area, related to space associated with reaching).
3. Posterior parietal neurons encode the locations of stimuli in specific reference frames (e.g., retinotopic, eye centered, head centered, body centered, and world centered). An area might encode in a single reference frame, and a stimulus might be coded in more than a single reference frame. There is debate whether a single neuron can simultaneously encode spatial location of an object in multiple reference frames. This combinatorial multiple reference frame encoding might permit high-efficiency processing of spatial information relevant to selection of stimuli and actions on them.
4. Parietal areas execute transformations of spatial information between specific reference frames (e.g., from retinotopic to head centered) and between modalities and reference frames (e.g., auditory and visual signals to eye-centered coordinates).

5. The parietal cortex contains mechanisms for shifting attention, stimulus selection, and movement planning.

Relevant spatial reference frames and locations are selected according to their salience and relevance to specific sensori motor transformation.

B. Explanations of Extinction

The mechanisms of extinction have been investigated much less intensively than those of neglect. The main reason for this seems to be that the understanding that extinction is a syndrome on its own and not a ''mild neglect'' is relatively new. As a result, only recently have independent mechanisms been sought. Thus, for example, a dichotomy such as attentional/intentional extinction does not exist as it does in spatial neglect. Therefore, we are able to give only a general view of the subject.

The main mechanisms suggested to account for extinction are either sensory or attentional. The sensory theories claim that pure contralateral deafferentation might cause extinction. Although this can be the mechanism for some patients, it seems unlikely to explain the syndrome in most patients, mainly because of recent reports suggesting that the extinguished stimuli is processed to an advanced level. Also, sensory deficit cannot explain extinction of the left stimulus of a pair that is presented in the basically intact ipsilesional hemifield. An attentional explanation seems to be more adequate.

The unique feature of extinction is that it appears only upon bilateral double simultaneous stimulation. This raises the hypothesis that there is a competition between the stimuli—that in the patient, because of an attentional deficit, there is a failure of the contralesional stimuli to successfully compete for limited processes resources. In fact, an effect of competition can even be found in normals. In a position emission tomography study of normals, it was shown that simultaneous bilateral visual stimulation reduced the activation of bilateral striate and extrastriate cortices compared to the level at which each side was activated in the single stimuli condition. Interestingly, an extinction-like phenomenon appeared on one of the sides that was not attentively prioritized (e.g., the subject was instructed to begin the report either from the left or the right side of the array of stimuli). Thus, with the bilateral stimulation, there is a hemispheric rivalry. The extinction appears because there is an attentional bias to the ipsilesional hemifield, with a resulting suppression of the lesioned hemisphere. The two main hypotheses used to explain this attentional bias are the same as those used to explain spatial neglect. The limited capacity theory claims that the lesioned hemisphere cannot fully attend to the contralateral stimuli. The capacity of the remaining attention is generally limited and thus results in extinction of the less attended stimulus. In support of this theory is the finding that patients report bilateral stimulation in conditions in which only the ipsilesional arm was touched but not vice versa, and also the fact that the reaction times to visual stimuli in extinction patients were higher than those of normals even in the intact hemifield. The other explanation is that an ipsilesional bias is caused by the release of the left hemisphere contralateral-orienting system from the right hemisphere inhibitory influence. In support of this theory is the gradient of increased efficiency of stimulus detection from left to right and the improvement of the extinction by nonverbal report in right-handed patients.

In contrast to neglect, much less is known about the possible dissociation between attentional and intentional features of extinction. A recent study of extinction in parkinsonian patients who had an anterior flexed posture and were simultaneously stimulated on the anterior and posterior surfaces of the upper part of the trunk revealed that they extinguished the posterior stimuli. This might suggest an intentional effect on attentional bias.

A disengagement deficit was described in patients with visual extinction. However, this cannot explain contralesional extinction when the contralesional stimulus was presented before the ipsilesional one. An effect of specific spatial reference frames on extinction is suggested by some studies for both visual and tactile extinction in a way similar to neglect. It seems that in extinction there is also a deficit in computations of reference frames that are needed for various tasks, but the representations are not abolished.

C. The Relation between Mechanisms of Neglect and Unawareness

Each of the mechanisms mentioned previously might interfere with awareness of perceptual experience in the neglected contralesional hemispace. For example, although the exact role of selective attention in visual awareness is not fully understood, some of its effects seem to be correlated with awareness (to visual

stimulus in this case). Among them are binding components of the visual scene, increasing activity in the different visual areas, and linking information from the ventral and dorsal streams. Selective attention was found to have a powerful effect on recognition potentials. Intentional processes are part of motor planning. It was suggested that visual awareness is associated with motor programming. Thus, an intentional deficit might be related to unawareness. The feedforward theory of motor-intentional deficit in anosagnosia to hemiplegia is also related to the intention–awareness relationship.

A deficit in the representation of a portion of space might cause a deficit in perceptual awareness by more than a single mechanism. For example, with a degraded representation one may be unable to compute the spatial coordinates that are needed to spatially direct attention. Another possibility is a deficit in visuospatial working memory. Such a deficit will prevent the incorporation of the spatial background that is needed for the short-term storage and processing of newly perceived and already represented information. Interestingly, working memory activity is needed for perceptual awareness.

Recently, it was shown that the use of nonverbal response in testing for extinction reduced visual extinction. Conscious awareness is deepened by language and by activation of language-bound representations. It might be that there was a disconnection of language networks in the left hemisphere from the visual presentation of the extinguished stimulus in the right hemisphere. This of course will not explain extinction of contralesional stimulus of double stimuli that are presented in the ipsilesional hemifield.

Clearly, the spectrum of the hypothesized mechanisms of neglect and extinction reflects the better understanding of these syndromes. There is no doubt that further study of the relation between phenomena such as attention, intention, and awareness will enable an integration of many of these mechanisms into a more comprehensive model of spatial awareness and its deficits.

VIII. CLINICALLY IMPORTANT ISSUES

A. Course

In stroke patients, neglect usually improves in weeks to months. The rate of recovery is fastest in the first 10 days. Left hemispheric/right hemispatial neglect improves at a faster rate and to a higher level than does right hemispheric/left hemispatial neglect. Intentional deficits, extinction, and anosodiaphoria might be present for long periods of time. The natural history reflects the spontaneous recovery of the brain, but the severity of neglect can also be influenced by training.

B. Rehabilitation

Rehabilitation of neglect means improving the patient's awareness and orienting to the neglected hemispace beyond that expected by its natural recovery. The improvement should be prolonged and should generalize to real–life situations. Some studies provide evidence for such an effect. Until recently, three main approaches were used to get patients to direct their attention to the neglected side; cueing, the use of salient stimuli, and contralesional limb activation. Altering a spatial attention bias was achieved by training visuospatial scanning, sometimes with addition of optokinetic stimulation, by turning the head to look to the neglected side, by using Fresnel prisms to shift images from the neglected side toward the normal side, and by feedback procedures. Increasing nonspatial attention was achieved, for example, by training vigilance by a self–alerting procedure.

Rehabilitation procedures are associated with functional improvement of lesioned brain areas. A recent PET study in neglect patients after right hemispheric lesions showed that in the acute phase, hypometabolism was found in both the ipsi- and contrahemispheres. The degree of this transhemispheric diaschisis was correlated with the size of the involvement of the right cortical and subcortical regions and with the degree of improvement in the chronic phase. A recent single photon emission computerized tomography (SPECT) study investigated cerebral blood flow changes in seven patients with left hemineglect from right hemispheric lesions who had rehabilitation for 2 months, beginning 4 to 14 months post-CVA. The training included visual scanning, reading and copying, copying line drawings on a dot matrix, and descriptions of a scene. SPECT was done before and after rehabilitation. Increased perfusion was found in the posterior parts of the lesioned right hemisphere and in the anterior areas of the nonlesioned left hemisphere. This increment in perfusion was correlated with the behavioral improvement shown by the patients in a visual exploration task. These results suggest that the training of space exploration helps patients to develop strategies to increase attention and exploration of the left hemispace. Although in normals attending and exploring to left hemispace involves the

right prefrontal areas, these areas were either destroyed or disconnected in the neglect patients. Consequently, the contralesional left hemisphere prefrontal areas might have taken over this function. In addition to intrahemispheric and interhemispheric effects, behavioral improvements might also be achieved by subcortical mechanisms. For example, a study in the cat showed that neglect associated with cortical lesions was reduced by destroying the ipsillateral superior colliculus or the intercollicular commissure. Recent studies have improved our insight into the plasticity of brain changes that are related to the rehabilitation of neglect. For example, an ipsilesional gaze-related enhancement in blood flow in the lesioned hemisphere was recently reported. This enhancement was broadly distributed and reflected an increase over normal values, suggesting the involvement of generalized diffuse synaptic activity in functional inprovement. The observation that the effect of prisms in reducing neglect is maintained for at least 2 hr after removal of the goggles also suggests that treatment can induce changes in synaptic connectivity.

The understanding of neglect as a heterogeneous deficit in specific computations of components of spatial behavior enables better understanding and development of potential rehabilitation strategies. A recent study aimed to analyze whether the improvement of line bisection errors in neglect is due to improvement of the causative sensory-attentional mechanism (defined by the Landmark task) or because of a learned behavioral strategy to counteract these mechanisms. In the two studied patients, it was shown that despite marked recovery on line bisection, they still showed a significant sensory-attentional deficit. Thus, they adapted some other mechanisms to overcome the deficit. The fMRI finding that in normals there is an activation of the cerebellum as well as other cortical areas in line bisection judgments (in addition to the inferior parietal area) supports the possibility of other task-related areas taking over the inferior parietal function. Another study reported a patient who had right frontotemporal CVA, with resulting ipsilesional neglect. The deficit was primarily motor intentional with a motor-action bias to the left and a secondarily sensory-attentional bias for stimuli to the right (defined by the video apparatus described previously). Rightward motor-intentional cue (touching the end of the bisected line) improved the primary left-sided intentional bias, and the leftward sensory-attentional cue improved the secondary right-sided attentional bias. Using such methods to define the exact deficient mechanism according to the model of spatial behavior might help in designing rehabilitational-specific strategies and increase the efficiency of the treatment.

C. Pharmacology

Although only very little is known about the neurochemistry of spatial attention, several observations suggest that deficits in some neurochemical systems, especially the dopaminergic, might be related to the syndrome of spatial neglect. The effect of dopamine might be either specific (to improve the hemispatial intentional deficit by enhancing the ability to intend to act) or more general (to increase the modulatory effect of dopamine on the brain distributed network that moderates spatial distribution of attention). A recent study reported an improvement in neglect behavior in a patient who was treated with the dopamine agonist bromocriptine. However, the same agent was found to increase the ipsilesional attentional bias. Thus, the effect of dopamine to improve the behavior of neglect might be more specific to the attentional and not to the intentional component of the deficit.

In addition to dopamine, an effect of norepinephrine has been suggested to improve spatial neglect, but there is no direct clinical evidence. Clomethiazole, a modulator of the GABA(A)ergic system, has been reported to specifically reduce the severity of neglect behavior in monkeys that had an experimental unilateral stroke. It seems that a pharmacological treatment for hemispatial neglect should be actively searched for.

See Also the Following Articles

AGNOSIA • ATTENTION • BRAIN LESIONS • DYSLEXIA • LATERALITY • SPATIAL COGNITION • SPATIAL VISION • VISUAL DISORDERS

Acknowledgment

The authors thank Ms. Yael Arnon for help in preparing the manuscript.

Suggested Reading

Adair, J. C., Na, D. L., Schwartz, R. L., and Heilman, K. M. (1998). Analysis of primary and secondary influences on spatial neglect. *Brain Cognition* **37,** 351–367.

Andersen, R. A., Snyder, L. H., Bradley, D. C., and Xing, J. (1997). Multimodal representation of space in the posterior parietal cortex and its use in planning movements. *Ann. Rev. Neurosci.* **20,** 303–330.

Anderson, B., Mennemeier, M., and Chatterjee, A. (2000). Variability not ability: Another basis for performance decrements in neglect. *Neuropsychologia* **38,** 785–796.

Barrett, A. M., Beversdorf, D. Q., Crucian, G. P., and Heilman, K. M. (1998). Neglect after right hemisphere stroke: A smaller floodlight for disturbed attention. *Neurobiology* **51,** 972–978.

Benson, F., and Geschwind, N. (1969). The alexia. In *Handbook of Neurology* (P. J. Vinken and G. W. Bruyn, Eds.), Vol. 4. North-Holland**,** Amsterdam.

Bisiach, E., and Vallar, G. (1988). Hemineglect in Humans. In *Handbook of Neuropsychology* (F. Boller and J. Grafman, Eds.), Vol. 1**,** 195–222. Elsevier, Amsterdam.

Husain, M., Mattingley, J. B., Rorden, C., Kenneth, C., and Driver, J. (2000). Distinguishing sensory and motor biases in paritela and frontal neglect. *Brain* **123,** 1643–1659.

Na, D. L., Adair, J. C., Williamson, D. J., Schwartz, R. L., Haws, B., and Heilman, K. M. (1998). Dissociation of sensory-attentional from motor-intentional neglect. *J. Neurol. Neurosurg. Psychiatr.* **64,** 331–338.

Stone, S. P., Halligan, P. W., Marshall, J. C., and Greenwood, R. J. (1998). Unilateral neglect: A common but heterogeneous syndrome. *Neurology* **50,** 1902–1905.

Wertman, E., Reches, A., Heilman, K. M., and Linetsky, E. (2000). Posterior truncal tactile extinction in anteriorly flexed Parkinsonian patients. *Neurology* **54,** 206.

Ventricular System

DAVID E. SCOTT
Eastern Virginia Medical School

GLOSSARY

cuboidal ciliated ependymal cells Involved in the dynamics of local formation and flow of cerebrospinal fluid.

ependymal cells Subtypes that line the cerebral ventricular system.

tanycyte Specialized cells that constitute the lining of circumventricular organs.

The cerebral ventricular system is regarded as the core of the mammalian central nervous system and plays multiple physiological roles. It is important in the physical support of the brain through buoyancy effects; it is a trophic mediator for the global distribution of a broad array of neuromodulators, neuropeptide hormones, and neurotransmitters. The cerebral ventricular system also harbors the choroid plexus, which is responsible for the production of cerebrospinal fluid. The mammalian cerebral ventricular system possesses specialized regions called circumventricular organs that allow the brain not only to perceive the peripheral endocrine and physiological milieu but also to respond by changing it with precision.

I. HISTORICAL INTRODUCTION

It is certain that we are not the first to inquire or analyze the mammalian cerebral ventricular system, which has been regarded for centuries as the core of the central nervous system. Alexandrian anatomists associated the cerebral ventricles with various mental functions. In 1487, Leonardo DaVinci turned his attention to the mammalian cerebral ventricular system. Employing the brain of an ox and his skills as a sculptor, he injected the cerebral ventricular system with hot wax using a syringe after making two vent holes in the anterior horns of the lateral ventricle. When the wax cooled, he carefully dissected away the overlying gray and white matter to reveal a relatively accurate waxen cast of the mammalian cerebral ventricular system. He regarded the cerebral ventricular system as a repository for fantasy, cognition, and memory. In 1503, a Carthusian scientist named Gregor Reich asserted that the freely communicating cavities of the cerebral ventricular system possessed different functions. His thesis inculcated the idea that the first ventricle (the anterior horn) sequestered imagination. The second cavity (the rest of the lateral cerebral ventricle) was the seat of higher associations and the midline third cerebral ventricle was the repository for memory. In his *Treatise on Man*, (1624), Rene Decartes suggested that the ''spirits'' from the body entered the brain via the pineal gland and were distributed throughout the cerebrospinal fluid (CSF) to various regions of the brain in a global fashion. Despite the fact the these early scientists–philosophers did not posses the high-tech imaging equipment and biochemical technology of today, it is surprising how close they came to modern thinking about the

Table I
Overview of Development

Level	Pr. Div.	Subdivision	Derivative	Ventricular system
Supratentorial	Prosencephalon	Telencephalon	Cerebral cortex	Lateral cerebral ventricle
			Limbic structures	
			Basal nuclei (ganglia)	
		Diencephalon	Epithalamus	Third cerebral ventricle
			Thalamus	
			Hypothalamus	
			Optic nerves	
Post fossa	Mesencephalon	Mesencephalon	Cerebral peduncles	Cerebral aqueduct
			Corpora quadrigeminal	IV cerebral ventricle
Post fossa	Rhombencephalon	Metencephalon	Pons cerebellum	IV cerebral ventricle
		Myelencephalon	Medulla	
Spinal	Myelon	Neural tube	Spinal cord	Central canal of spinal cord
	Neural tube			
Peripheral		Neural crest	Peripheral ganglia	

structure and function of the cerebral ventricular system.

This article focuses upon current knowledge that deals with the development (Table I), regional organization, specializations, and structural and functional changes that occur in the normal mammalian cerebral ventricular system or following pathophysiological events as observed with scanning or transmission electron microscopy and immunocytochemistry.

II. EMBRYOGENESIS OF THE CEREBRAL VENTRICULAR SYSTEM

From a structural standpoint, the cerebral ventricular system can be considered the functional core of the brain and begins to make its appearance during the 18th day of life. The flattened neural plate, no more than 70 μm in diameter, appears at approximately 18 days postcoitus (PC). At 20 days PC, the neural groove and folds appear rostrally. Exuberant growth of the neural folds forms a complete neural tube with closure of the caudal neuropore by 25 days in the normal developing brain. The neural tube exhibits three enlargements along its linear axis. The most rostral of these is the prosencephalon, which is destined to differentiate into the telencephalon and diencephalon. The telencephalon expands rapidly and by 12 weeks there is further differentiation into the forerunners of the cerebral cortex, basal ganglia, and the more complex limbic structures. At this time, the corpus callosum begins its inexorable caudal migration, forcing the hippocampus ahead of it into each newly forming temporal horn of the lateral ventricle. The diencephalon, whose walls constitute much of the third cerebral ventricle, begin its differentiation into the epithalamus, thalamus, and hypothalamus during the 12th week of gestation. The mesencephalon is slow to differentiate and does not develop into its definitive derivatives, the corpora quadrigemina and the crus cerebri, until the 20th week of gestation. The cerebral aqueduct (iter) is surrounded by the aforementioned structures and is continuous with the third cerebral ventricle rostrally and the fourth cerebral ventricle caudally, whose dorsal and ventral surfaces are formed by the metencephalon and myelencephalon. The metencephalon differentiates into the cerebellum and the pons by the 16th week of gestation and the myelincephalon differentiates into the medulla oblongata by 18 weeks PC. Despite the fact that the outer surface of the mammalian brain remains lissencephalic for 28 weeks PC, the cerebral ventricular system is fully formed and lined by various types of ependymal cells that exhibit an unusually broad array of functional and structural differences.

Although this article focuses on the three-dimensional topography of the cerebral ventricular system, it is important to discuss the pleuripotential neuroepithelial cells that constitute the simple neural tube and particularly the neuroblasts that give rise to neurons

and the spongioblasts (gliablasts) that evolve into glia and ependymal cells. Many types of ependymal cells exist, and although their lineage is poorly understood, certain data point to the fact that they may be derived from wandering neuroblasts. In the adult definitive brain, they represent a population of postmitotic cells that lack the capability of renewing themselves and reepithelializing their surfaces of the ventricular wall, which may become denuded by various pathological processes and remain so permanently.

III. NEUROANATOMICAL ORGANIZATION OF THE ADULT CEREBRAL VENTRICULAR SYSTEM

A. The Lateral Cerebral Ventricular System

The lateral cerebral ventricular system is the most expansive part of the mammalian cerebral ventricular system. It consists of an anterior projection, with the anterior horns bounded by the septum pillucidum medially, the genu of the corpus callosum rostrally, and the head of the caudate nucleus laterally. It is continuous with the body of the lateral ventricle, whose medial and lateral walls are composed of the body of the caudate nucleus and the lateral aspect of the thalamus. A thick veil of overlying choroid plexus, which is derived from the germinal matrix of the choroidal fissure during the seventh week *in utero*, commonly obscures both of these structures. The body of the lateral cerebral ventricle is continuous, with the posterior horn and the beginning of the temporal horn, an area that is collectively referred to as the collateral trigone. The posterior horn is quite variable in mammals and particularly in the human brain. The most distinct feature of the temporal horn is the presence of the massive hippocampus, with the alveolus of the fornix arising from it as a flattened filet of fibers. Both the hippocampus and the fornix are commonly obscured by a thick felt work of chord plexus, which is highly vascularized and irrigated in large part by the anterior choroidal artery coupled with some branches of the posterior cerebral artery. The lining cells of the lateral cerebral ventricle are chiefly cuboidal ciliated ependymal cells (Figs. 1 and 2). The cilia that constitutes the surfaces of these ependymal cells beat in a metachronal fashion and are regarded as a mechanism that alters the local flow and dynamics of CSF, which is produced in large part by the choroid plexus. It should be noted, however, that cilial density alters from region to region and if a region is denuded of ciliated ependymal cells, no reepithelization will occur. Hence, ciliated ependymal cells of the lateral cerebral ventricle are regarded as postmitotic.

B. The Third Cerebral Ventricle

The mammalian third cerebral ventricle is bounded dorsally by the thalamus. The lower walls and floor are formed by the endocrine hypothalamus. The actual floor of the hypothalamus is formed by the dorsum of the median eminence (Fig. 3), which is a major component of the neurohypophyseal system. The third cerebral ventricle functionally interconnects with the lateral cerebral ventricular system via the two interventricular foramina. The leading edge of both foramina is formed by the columns of the postcommissural fornix that are heavily myelinated fibers that arise from neurons in the hippocampus. The roof of the third cerebral ventricle is formed by the velum interpositum, which separates the body of the fornix from the cavity of the third cerebral ventricle below. The two internal cerebral veins pass through the velum interpositum to fuse as the great vein of Galen over the pineal gland caudally. The velum interpositum gives rise to choroid plexus, which is continuous with that found in both lateral ventricles and courses through the interventricular foramen. The interventricular foramen is a critical area because of its small size and can be easily obstructed by pathological changes in ependymal cells that constitute the choroid plexus, such as ependymal choroid plexus papillomas or ependymomas. Due to the narrowness of this foramen, obstruction can trigger the development of a noncommunicating hydrocephalus and the buildup of CSF in either one or both lateral cerebral ventricles.

Ependymal cells that line the dorsal aspect of the hypothalamic wall are heavily ciliated and posses gap junctions; they are strikingly similar to those found in the lateral cerebral ventricular system. However, as one descends toward the floor of the third cerebral ventricle (dorsum of the median eminence), there is a gradual transition with increasingly fewer ciliated ependymal cells (Figs. 3–5). As one reaches the lateral recess of the third cerebral ventricle, there is a replacement of cuboidal ciliated cells with a specialized type of ependymal cells called tanycytes. These specialized ependymal cells have relatively few cilia, if any, but instead exhibit a rich nap of microvilli that form distinct hexagonal arrays (Fig. 6). On the floor of the third cerebral ventricle, which is the dorsum of the median eminence, beneath these arrays of microvilli there exist tight junctions referred to as zonulae

Figure 1 Scanning electron microgram (SEM) of ependymal surface of the head of the caudate nucleus forming the lateral wall of the anterior horn of the lateral ventricle. Both cilia (CL) and micro villi (MV) are notable in this region of the cerebral ventricular system. × 8000.

occludentes (Fig. 7), that prevent even the smallest proteins or macromolecules from migrating interstitially from the blood vascular system into the cerebral ventricular system from the median eminence below. The barrier properties of tanycytes are discussed later.

C. The Cerebral Aqueduct

The mammalian cerebral aqueduct is clearly one of the most critical regions of the cerebral ventricular system. Through this narrow channel, only 2 mm in diameter, more than 500 ml of CSF must pass in a 24-hr period in the human brain. Pathophysioloical changes in the ciliated ependymal of this region due to fetal infections or teratogens can cause atresia and/or obstruction with the formation of a noncommunicating hydrocephalus. Lining cells of the cerebral aqueduct (Fig. 8) are uniformly ciliated and demonstrate regular metachronal synchrony in their movement.

D. The Fourth Cerebral Ventricle

The roof of the fourth cerebral ventricle is formed by the cerebellum (Figs. 9 and 10). The rostral floor is formed by the pons, whereas the caudal floor is formed by the medulla oblongata. From the opening of the cerebral aqueduct to the caudal portion of the obex of the fourth ventricle, the ependymal cells that line the surfaces of the fourth cerebral ventricle are uniformly ciliated, even in the central depression, which is linear

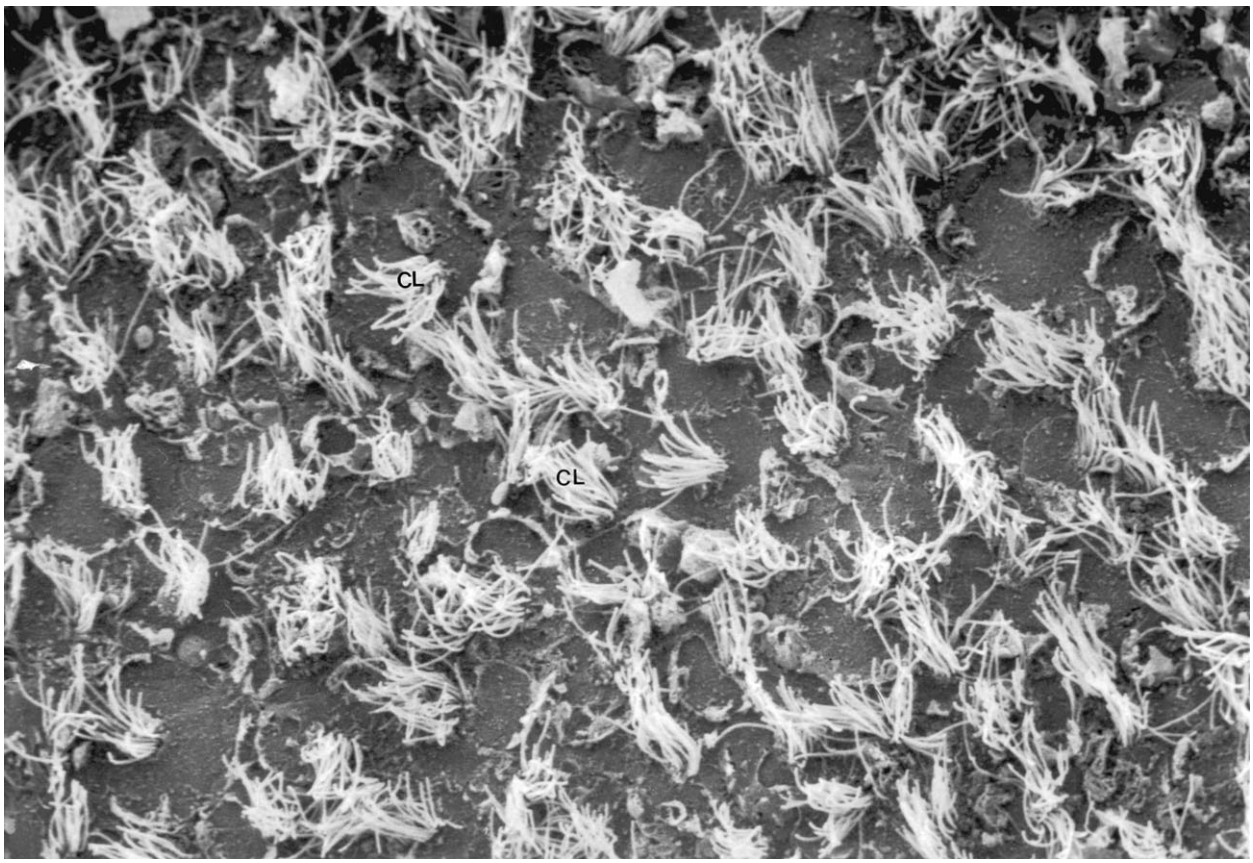


Figure 2 SEM of cerebral ventricular wall of the human hippocampus 28 days postcoitus. Ependymal cells demonstrate patchy cilia (CL) interspersed by areas devoid of any membranous modification. × 6000.

and is referred to as the medial eminence of the fourth ventricle. Frequently, isolated populations of cells can be observed on the surface of these ciliated ependymal cells. Careful analysis has revealed them to be histocytes (Fig. 9).

E. The Choroid Plexus

The choroid plexus arises from the choroid fissure at approximately 7 weeks *in utero*. It is normally found in all the cavities of the cerebral ventricular system, with the exception of the anterior horn and posterior horn of the lateral ventricle, the cerebral aqueduct, and the rostral component of the fourth ventricle. The choroid plexus is formed by a bilayer of ependyma and pia, which is referred to as a tela. The tela is invaded at 7 weeks *in utero* by primitive mesodermal elements, which are the forerunners of noncontinuous fenestrated capillaries. The pia eventually becomes atrophic, leaving a residuum of collagen and fibroblasts sandwiched between an inner layer of porous fenestrated capillaries and an outer ventricular layer of cuboidal ependymal cells covered by dense microvilli and sparse cilia in contact with the CSF of the ventricular system (Figs. 9 and 10). An important neuroanatomical specialization occurs between the apical surfaces of choroid plexus ependymal cells. The apposed apical membranes of adjoining cells posses zonula occludentes (tight junctions), which prevent the passage of even the smallest protein or metabolite from the blood vascular system, through the fenestrated capillaries, into the CSF. Hence, unlike the rest of the brain, the choroid plexus has a blood–CSF barrier rather than a blood–brain barrier and is more similar to the specialized regions of the brain that constitute the unique circumventricular organs where the barrier has been shifted from the vascular to the ventricular side. The fetal choroid plexus is significantly larger during development, than in the adult definitive brain. This

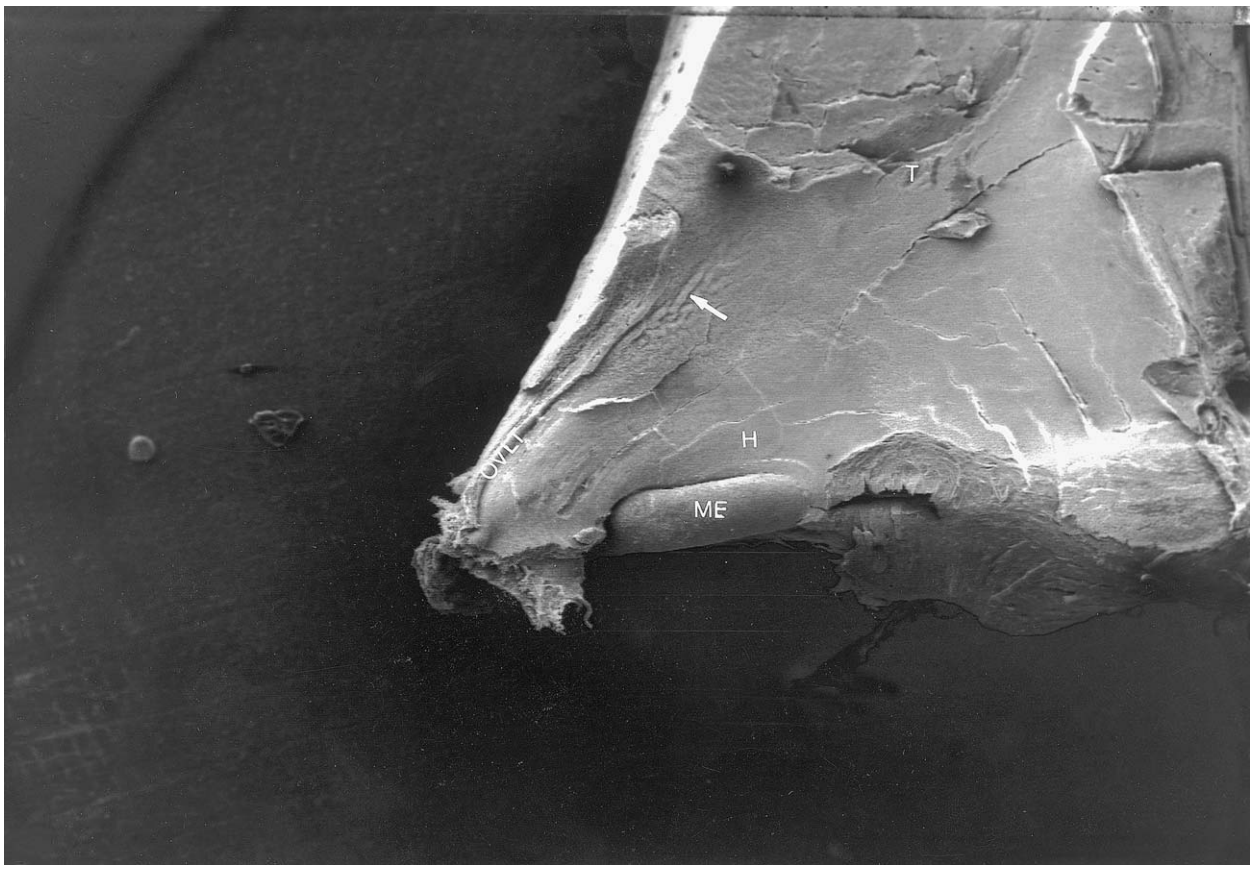


Figure 3 Low magnification SEM of the entire wall of the thalamus (T) and endocrine hypothalamus (H). There is a distinct transition in the neuroanatomical organization of the thalamic ependymal cells versus hypothalamic ependymal cells that line this expansive region of the diencephalon. Rostrally, the thalamic region is lined by heavily ciliated cuboidal ependymal cells that are thrown up into ridges (arrow) and possess gap junctions. As one descends toward the floor of the third cerebral ventricle, there is a transition toward less ciliated ependymal cells with more microvilli (compare with Fig. 4). Upon reaching the median emience (ME), which forms the floor of the third cerebral ventricle, ciliated ependymal cells are essentially replaced by tanycytes that possess apical tight junctions, have relatively few cilia, and whose foot processes stretch through the entire parenchyma of the basomedial hypothalamus to terminate in large numbers on fenestrated capillaries of the hypophyseal portal system. OVLT, organum vasculosum of the lamina terminals. $\times 280$.

relative enlargement and rich vascularity suggest the concept of early prenatal function for this critical organ system within the developing brain and connote a trophic and nutritional contribution during fetal growth and development. The structural organization of human fetal ependymal cells, as observed with scanning electron microscopy, makes it clear that they are aptly suited for the bidirectional transport of a variety of ions and other solutes that comprise normal CSF, which plays a multipotential role in the developing brain. Because the parenchyma of the central nervous system has the same specific gravity as CSF, CSF may act as a buoyancy control mechanism and provide the brain with a certain degree of protection against sudden forces of trauma. In addition, CSF is thought to play a nutritive role for developing glia, ependyma, and neurons in the fetal brain. Since the mammalian cerebral nervous system has no lymphatic drainage system, it is speculated that CSF production, movement, and reabsorption may serve as a mechanism for the removal of cerebral metabolites back into the blood vascular system at the level of the arachnoid granulations that penetrate into the superior sagittal sinus. Since the extracellular space of the developing and adult brain freely communicates with the CSF, this may allow for the constant control of ionic composition of both compartments, with the exception of circumventricular organs. The immense array of biologically active factors, such as neurotransmitters, neuromodulators, and neuropeptide hormones, that may arise from the periventricular neuropil, as well as the established presence of intraventricular

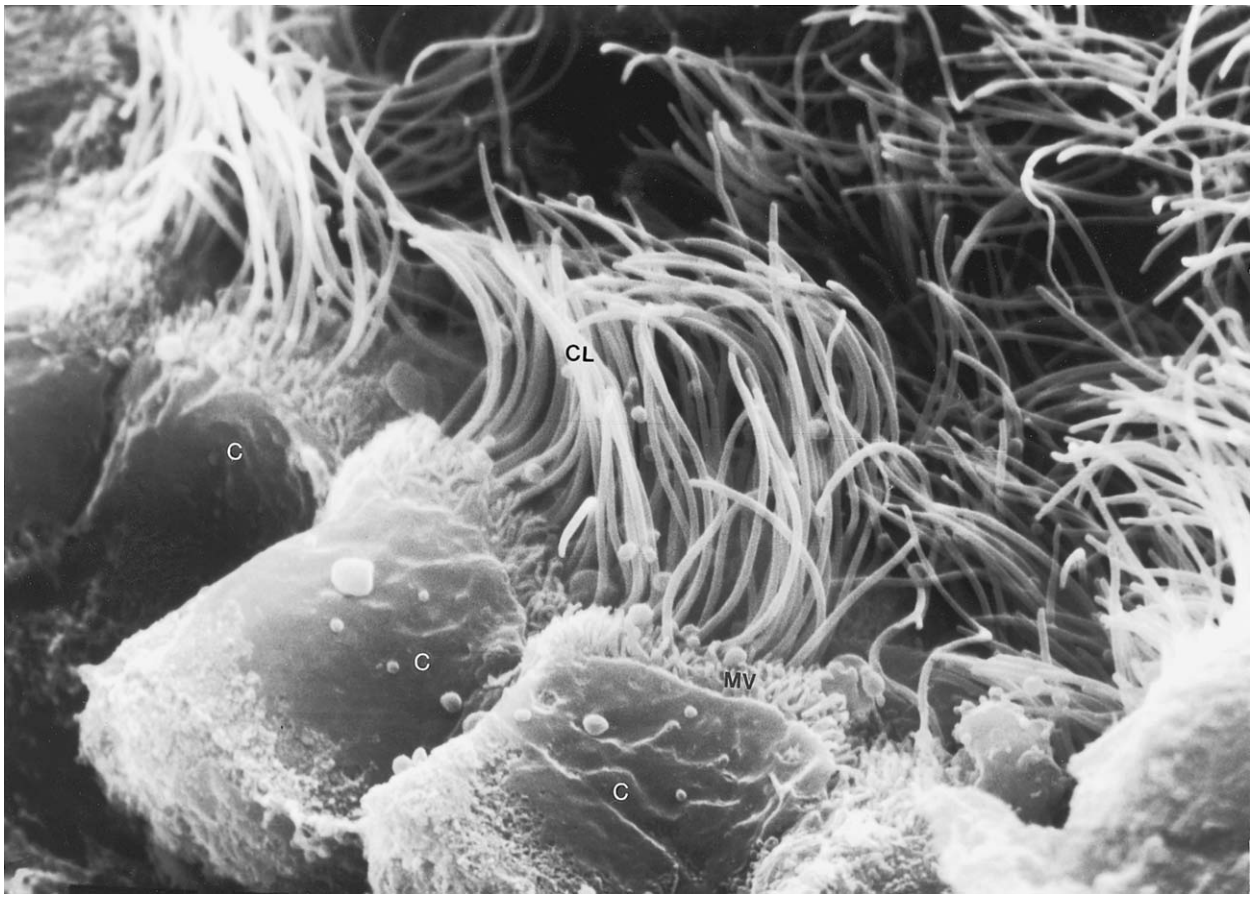


Figure 4 Midrange SEM of dorsal thalamic wall. Cuboidal ependymal cells (C), which lack tight junctions, exhibit a rich nap of cilia (CL) and microvilli (MV). × 6800.

axons and neurons in the cerebral ventricles, has triggered consistent speculation that the CSF may function as a trophic mediator for the distribution of such around the mammalian brain. Due to its strategic location and the lack of intrinsic enzyme systems in the CSF, this makes such an hypothesis very attractive. Finally, because of its predictable composition as well as its chemical and cellular consistency, any alteration in the CSF becomes an excellent diagnostic tool to demonstrate a wide spectrum of pathophysiological changes that can occur within the mammalian central nervous system.

IV. CIRCUMVENTRICULAR ORGANS (THE WINDOWS OF THE BRAIN)

The mammalian central nervous system would be a relatively useless organ if it were not capable of constantly assessing the peripheral physiological and endocrine milieu of the body that it exercises such precise control over. However, despite the fact that most of the brain and spinal cord possess a blood–brain barrier that consists of continuous capillaries with tight junctions protecting brain parenchyma from a variety of blood–bourne compounds that may be toxic to it, there are a small number of highly specialized areas found chiefly around the third cerebral ventricle and in the caudal region of the fourth cerebral ventricle that are unique with respect to their neuronal, glial, and ependymal organization. These specialized regions of the brain are referred to as circumventricular organs and, like the choroid plexus, their barrier properties have been shifted from the vascular to the ventricular side of the brain, which allows them the capacity to assess the peripheral physiological changes by their parenchyma, which is in direct contact with molecular components carried in the bloodstream. They are collectively referred to as circumventricular organs or the windows of the brain

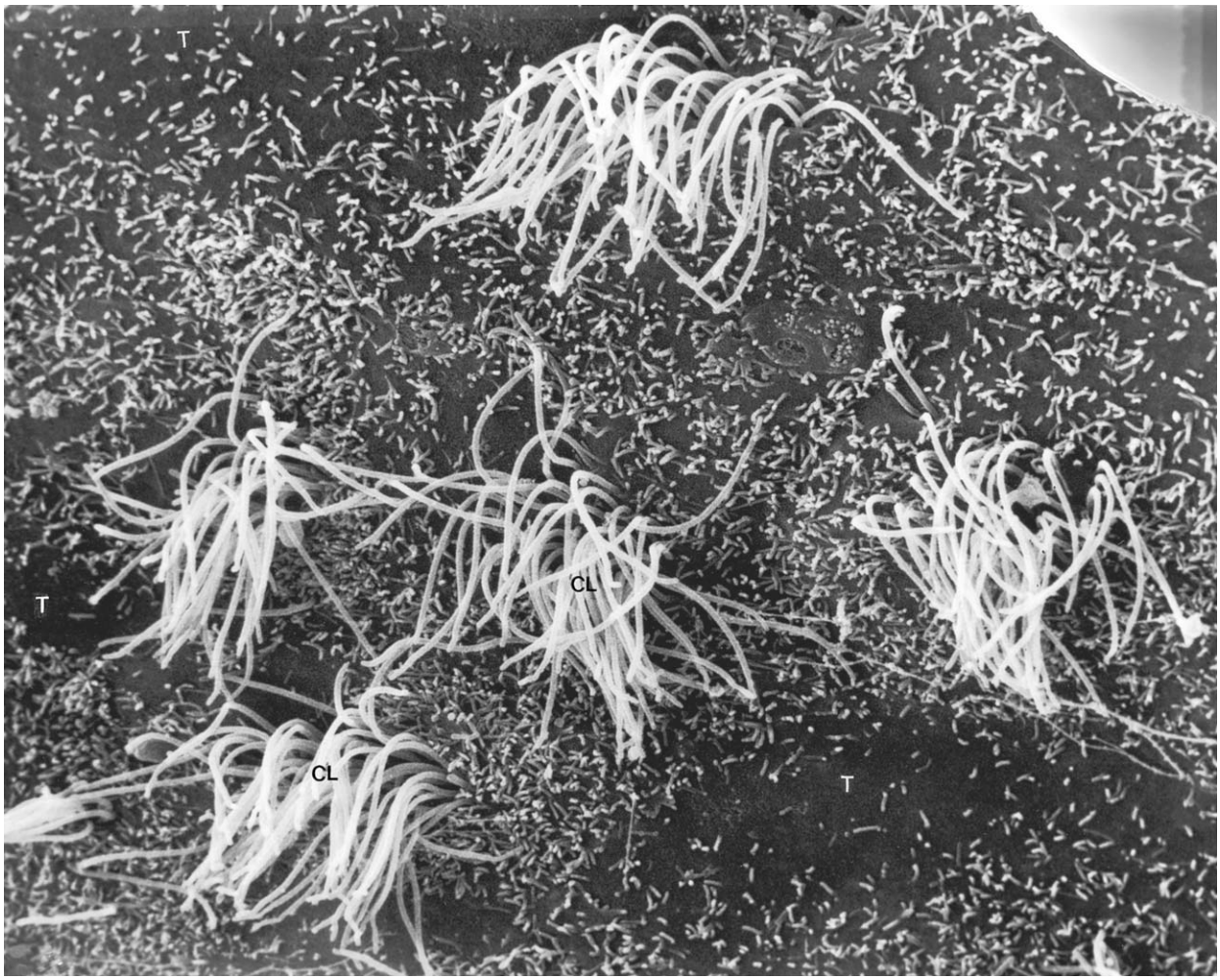


Figure 5 Transitional zone of lower thalamic–upper hypothalamic wall as observed with SEM. Ependymal cells with small amounts of cilia (CL) are being replaced by tanycytes (T), which exhibit microvilli on their apical (ventricular) surfaces. × 7500.

and are composed of the organum vasculosum of the lamina termialus, the median eminence (ME), the neural stem (NS) and neural lobe (NL), the subfornical organ (SFO), the pineal gland (PG), and the area postrema (AP) of the fourth cerebral ventricle. The organization of these specialized areas is unique from that of the rest of the brain. The barrier properties of these areas are fundamentally different and allow for the diffusium of bioactive macromolecules, peptides, and other blood-borne signals to enter their parenchyma in order to integrate homeostatic mechanisms, such as water balance, hypovolemia/hypervolemia, and changes in the peripheral endocrine milieu. With the exception of the PG, NS, and NL, all these circumventricular organs posses a highly specialized type of ependymal cell, the so-called tanycyte (Figs. 6 and 7). Tanycytes posses zonulae occludentes (tight junctions) between their apical zones, which face the ventricular lumen, and much like the ependymal cells of the choroid plexus, these tight junctions act as a blood–CSF barrier to prevent the transport of metabolites from the bloodstream into the cerebral ventricular system. Despite the fact that the lineage of tanycytes is not well understood, like pituicytes of the NS and NL, they are all regarded as highly specialized ependymal cells. Compelling evidence suggests that these unique ependymal cells are neurally innervated and exhibit synaptic specializations. Moreover, these specialized ependymal cells may produce specific neurotrophic factors that serve to attract both hypothalamic and extrahypothalamic neurites. Recent data also strongly suggest that transected supraoptic and

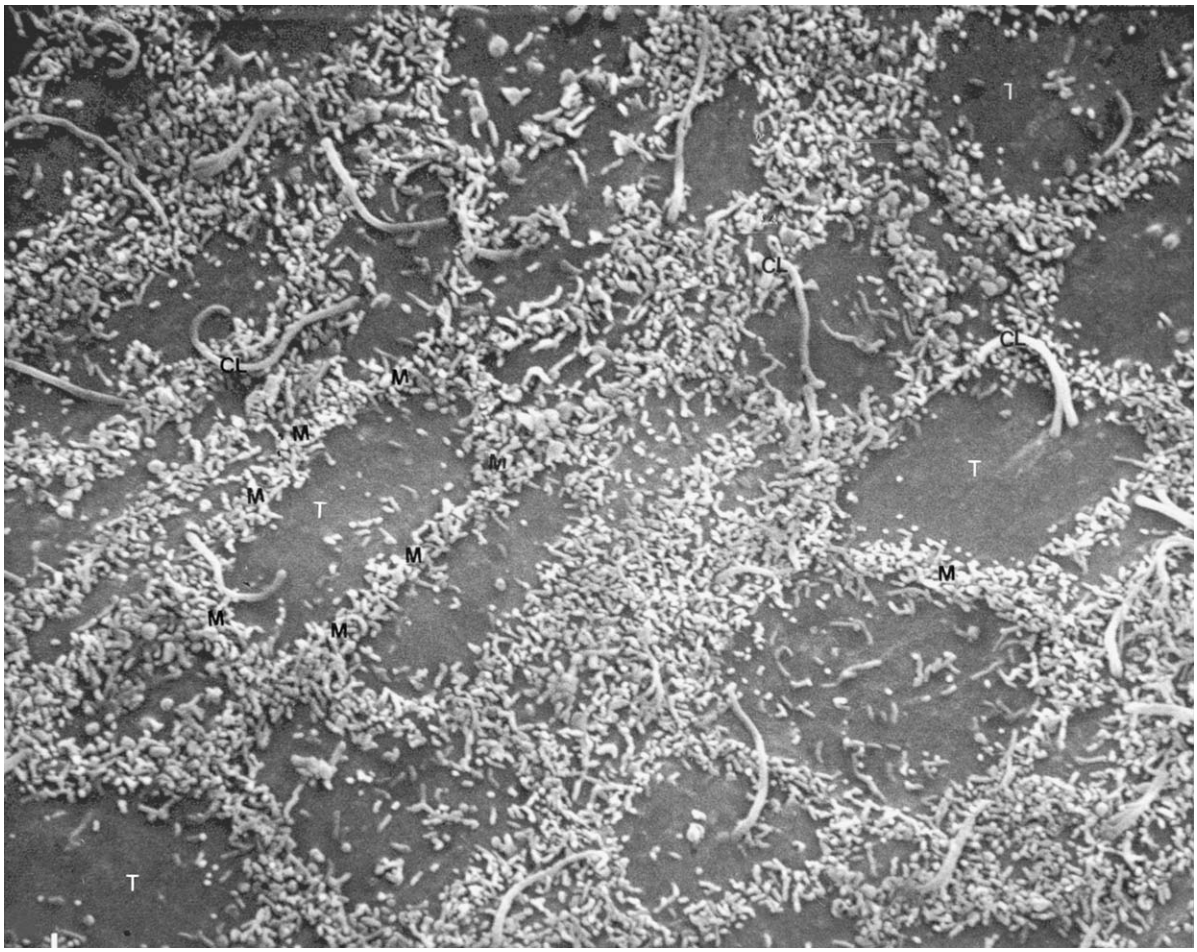


Figure 6 SEM of floor of the third cerebral ventricle (dorsum of the median eminence). Notable are the apical (ventricular) surfaces of tanycytes (T) that are rimmed by ridges of microvilli (M). This gives a hexagonal mosaic appearance to the surface of this circumventricular organ. A few isolated cilia (CL) are notable. This region of the brain is regarded as a target area for the absorption of biologically active molecules from the CSF into the parenchyma of the median eminence. $\times$ 6000.

paraventricular neurits regenerate along tanycyte processes employing them as a template for growth and regeneration. Tanycytes of the ME consistently demonstrate the accumulation of dense core vesicles and microvesicles in their terminal foot processes that abut upon the fenestrated capillaries of the hypophyseal portal plexus. This observation has led to speculation that this specialized ependymal cell type may release various macromolecules into the portal plexus and may directly influence adenohypophyseal metabolism. In control animals sparse numbers of neuritic processes can be observed on the surface of the median eminence (Fig. 11). Although relatively rare in normal animals, these neurites are considered to harbor catecholemines.

It appears that circumventricular organs have a great deal of functional connectivity and integration and not only contribute to the composition of the CSF but also to the composition of the blood. Also, along with other areas of the brain, they posses a high degree of plasticity and germinal potential.

A. The Median Eminence–Neurohypophyseal System

The ME, NS, and NL play an important role in the release of both arginine vasopressin (AVP) and oxytocin (OXY) into the blood vascular system. Both these occtopeptide hormones are noncovalently bound to a carrier molecule neurophysin. However, in the case of the ME, hypothalamic releasing factors from the surrounding endocrine hypothalamus enter a unique capillary bed called the primary hypophyseal

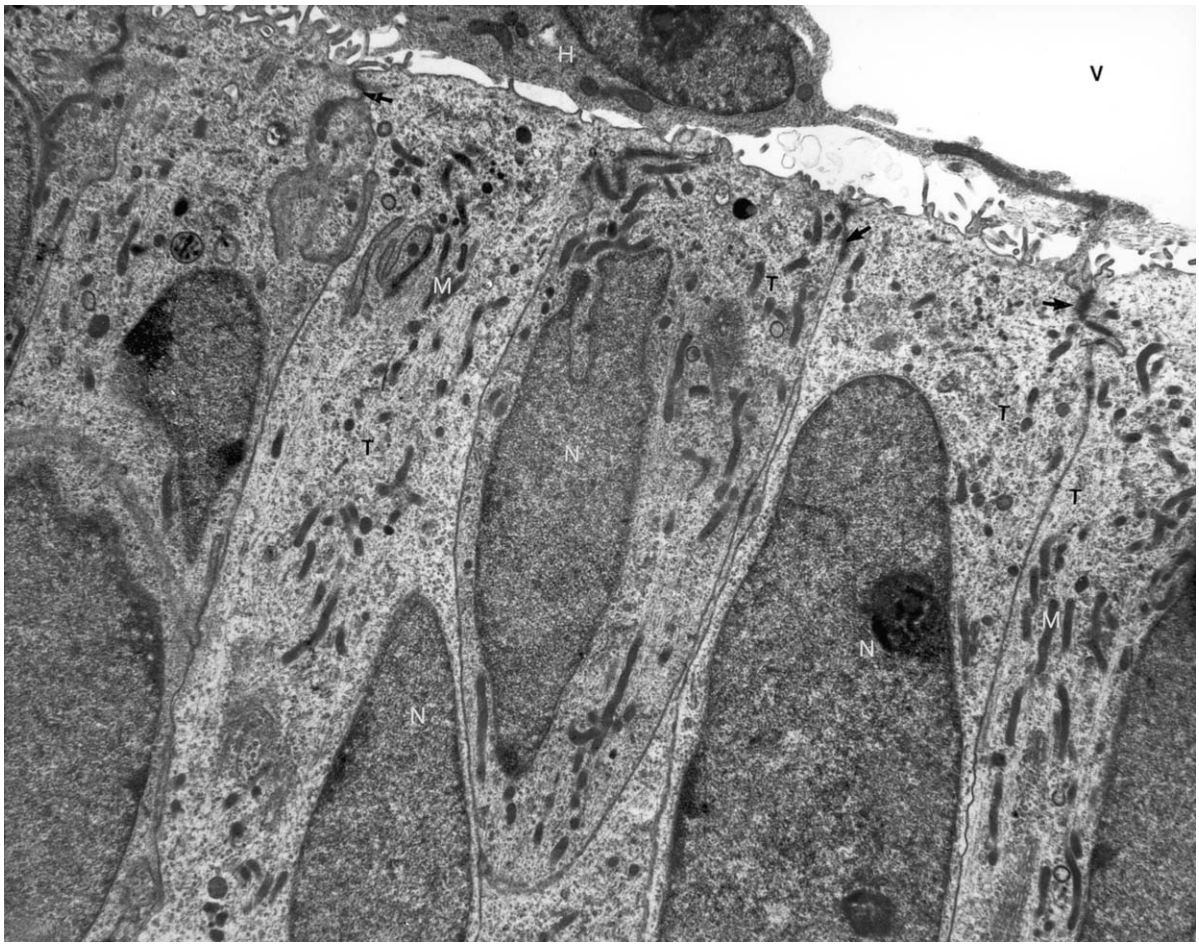


Figure 7 Transmission electron micrograph (TEM) of tanycytes (T) that constitute the floor of the third cerebral ventricle (dorsum of the median eminence). These specialized ependymal cells posses tight junctions between the apical cell collars that face the lumen of the ventricle (V) and, like other circumventricular organs, are regarded as a blood–CSF barrier to the movement of proteins and other macromolecules. H, histiocyte; N, nuclei; M, mitochondria. × 12,500.

portal plexus, which is found at the base of the ME. These delicate capillaries ultimately become the long portal veins that recondense to form the portal sinusoids, which are the sole and exclusive blood supply of the anterior pituitary gland. The anterior pituitary gland is a source of trophic hormones that serve to maintain the integrity of the peripheral endocrine milieu that feeds back on both the pituitary and the median eminence and other portions of the brain and endocrine hypothalamus. Modified glial cells, the so-called pituicytes, exercise dynamic control over NS and NL axons that sequester AVP and OXY during states of hypovolemia or during lactation and suckling. In their passage from the supraoptic (SON) and paraventricular (PVN) neurons of the endocrine hypothalamus, axons that pass through the median eminence (which are destined to terminate in the NS and NL) become myelinated as they enter it. When SON and PVN axons ultimately leave the median eminence to enter the NS and NL, where they form neurovascular zones for the systemic release of neuropeptide hormones, they again become unmyelinated. Although not fully understood, the myelinization of neurohypophyseal axon fibers in the median eminence may serve to confer a degree of protection in a region that otherwise lacks a bona fide blood–brain barrier and instead possesses a blood–CSF barrier. All of the circumventricular organs posses a blood–CSF barrier and all exhibit axon terminals that abut upon fenestrated capillaries for the quantal release of various bioactive hormones and neuropeptides into the blood–vascular system. This concept of a neurovascular

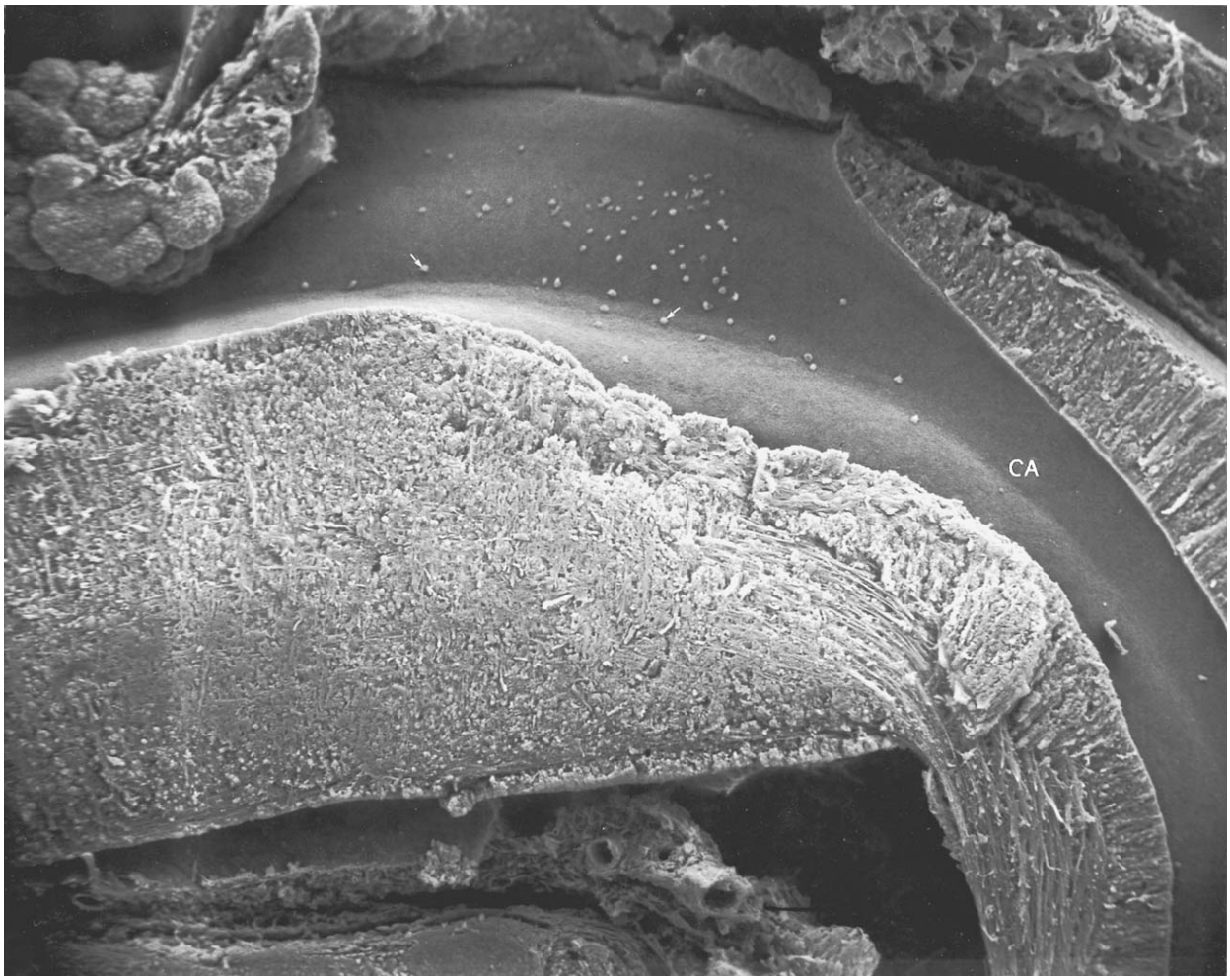


Figure 8 Low-magnification SEM of cerebral aqueduct (CA). The ependymal cells are uniformly ciliated. Numerous macrophages (arrows) stud the rostral end of the fourth cerebral ventricle. × 600.

(neurohemal) function is a pivotal concept in neuroscience and has been described as the mechanism of neurosecretion. This fundamental concept enunciated more than 69 years ago by Ernst Scharrer suffered a long apprenticeship of skepticism and doubt before it was finally accepted as a major tenent of neuroendocrinology in the early 1950s.

B. The Subfornical Organ

Unlike other circumventricular organs (CVOs), the subfornical organ possesses a variety of different neuronal types, which are integrated with other regions of the hypothalamus to include the SON and PVN. There are essentially four different types of neurons that reside within the subfornical organ. Two types are in contact with the cerebral ventricular lumen and the CSF and the other two types are deep within the parenchyma of the SFO. Consistent with the neuronal subtypes are axonal subtypes that are found in the SFO. A number of glial subtypes are also found in the SFO, including tanycytes, whose apical surfaces face the ventricular lumen and whose end feet terminate on perivascular spaces that surround fenestrated capillaries. Like other CVOs, the SFO possesses a blood–CSF barrier. Numerous investigations have firmly established that neurons of the SFO in the third cerebral ventricle function as central dipsogenic receptors. These investigations have shown that the intraparencymal injection of a number of substances, including carbacol, acetylcholine, and physiostigmine, triggers drinking behavior. Conversely, ablation of the SFO inhibits chemically induced drinking behavior.

Figure 9 Caudal end of the fourth ventricle (V) just ventral to the area postrema (AP). The roof of this part of the ventricle is formed by the developing choroid plexus (CP). The floor exhibits numerous isolated histiocytes. Compare with Fig. 5. × 550.

Such conclusions are simply reinforced by the finding of rich reciprocal innervation that exists between the SFO and the SON and PVN of the endocrine hypothalamus.

C. The Pineal Gland

The PG, common to all mammalian species, is a relatively simple circumventricular organ with respect to its histological organization. It is routinely found at the dorsal–caudal region of the third cerebral ventricle just above the posterior commissure and medial to both habenular nuclei. The PG possesses essentially one predominant cell type, the so-called pinealocyte. The basic function of the normal intact, PG is antigonadotrophic and it actively synthesizes and releases the indole, melatonin. The neural regulation of the normal intact PG has been well investigated. Primary control of the PG is exerted by the prevailing light–dark environmental photoperiod acting through the suprachiasmatic nuclei (SCN). Environmental information perceived by the eyes is transferred to the PG over a complex series of neuronal pathways, which include the SCN and the PVN of the hypothalamus, pregangionic sympathetic neurons in the upper thoracic spinal cord, and postgangionic sympathetic neurons in the superior sympathetic ganglion. Direct innervation of the PG is from the superior cervical sympathetic ganglion. Production of the pineal hormone melatonin is cyclic, with high levels of synthesis occuring at night (dark environment) and low levels occuring during the day (light environment). This cyclic synthesis is controlled by the neural pathways mentioned previously coupled with direct retinohypothalamic pathways. The synthesis of melatonin at

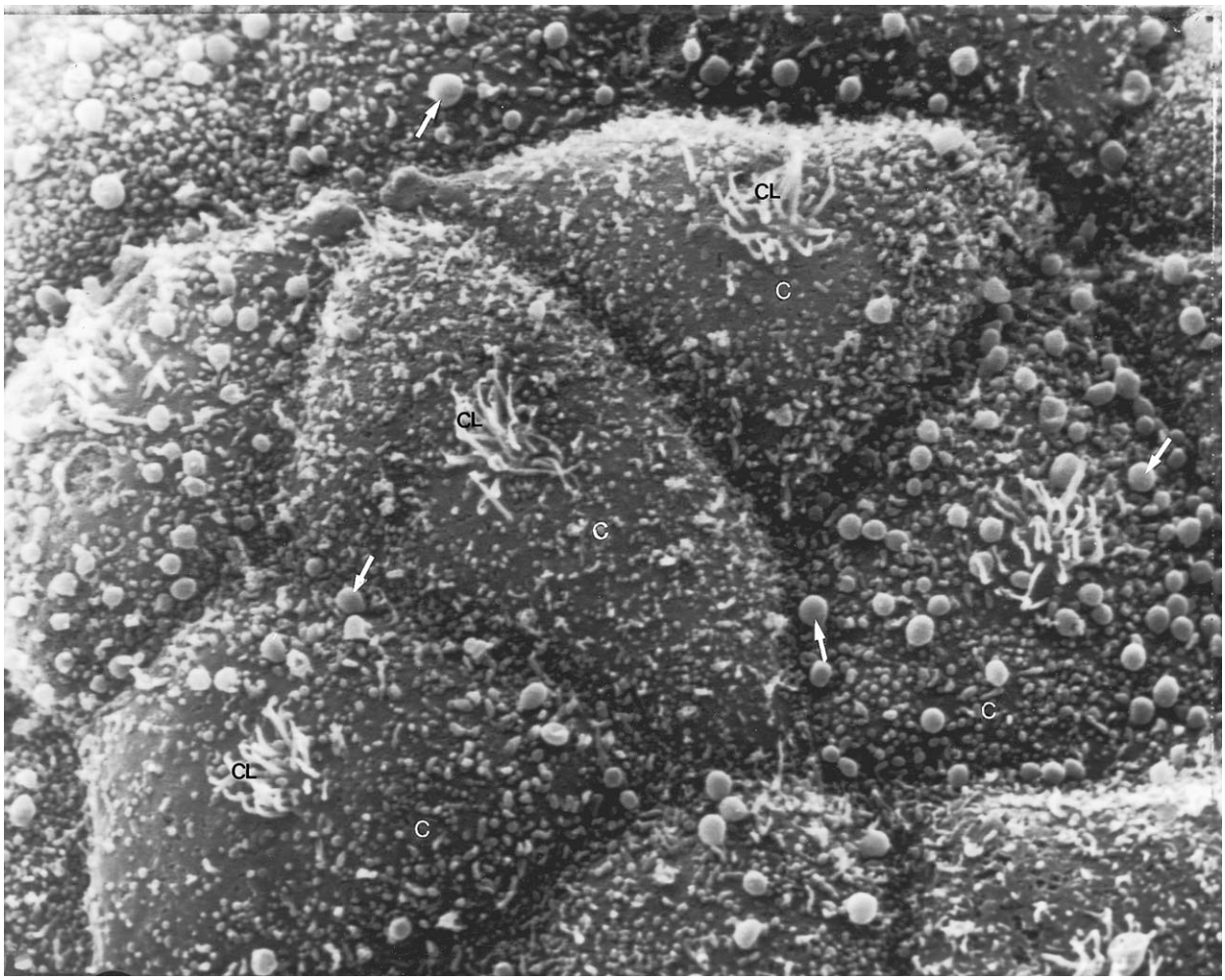


Figure 10 High-magnification SEM of the choroid plexus in the fourth cerebral ventricle. Note a distinct separation of the apical surfaces of the choroid ependymal cells (C) that in the living state are bathed by the cerebrospinal fluid and to which they contribute numerous bioactive molecules ions and peptides. These choroidal ependymal cells exhibit numerous apical secretory blebs (arrows) and populations of isolated cilia (CL). Also evident on the surface of these secretory cells are numerous microvilli. × 7000.

night is a consequence of norepinephrine (NE) released from postganglionic sympathetic nerve endings that terminate in the vicinity of the pinealocyte processes. The PG possesses a pineal recess, which is continuous with the lumen of the third cerebral ventricle. This recess is lined by ependymal cells and occasional small axonal varicosities can be seen on their surfaces. These varicosities have been interpreted to be monoaminergic.

D. The Area Postrema

The AP is found in the caudal end of the fourth ventricle just rostral to the obex. The cerebral ventricular lining of this CVO is composed of tanycytes that have relatively short cell bodies and terminate on underlying fenestrated capillaries. The apical surfaces of these cells lack cilia and posses a rich nap of microvili; they are also quite uniform in their organization (Fig. 9). The AP is rich in both serotimergic and noradrenergic fibers, some of which penetrate between lining tanycytes and sequester both clear and dense core vesicles that measure between 20 and 40 nm in diameter. Large numbers of neuronal elements can be noted in the parenchyma of the AP. Many of these are in close physical contact with the abluminal basal lamina of perivascular spaces that surround fenestrated capillaries. This close neuroanatomical juxtaposition of neurons with fenestrated capillaries suggests a

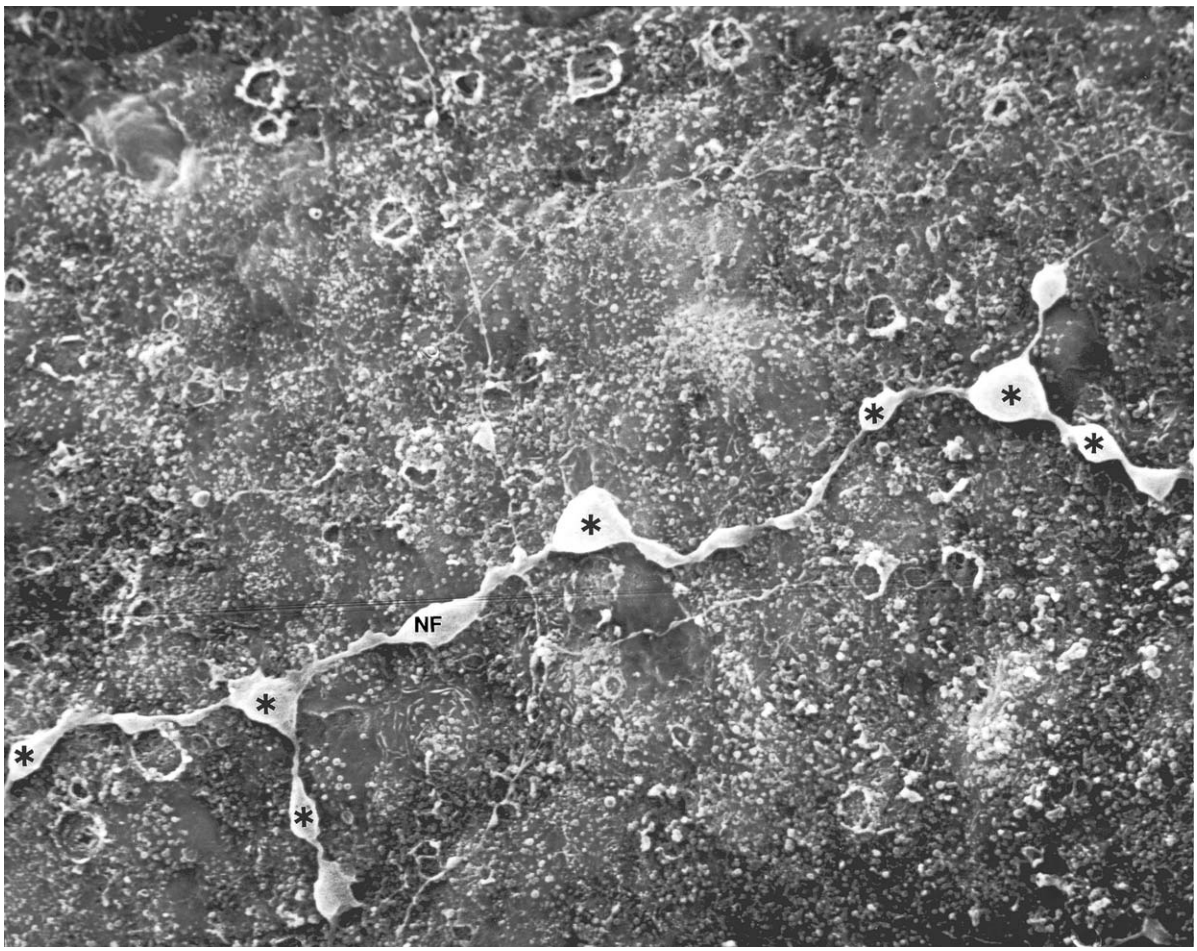


Figure 11 High-magnification SEM of a nerve fiber (NF) on the floor of the third ventricle. This fiber demonstrates distinct enlargements or varicosities (asterisks) along its linear axis. Such fibers are occasionally observed on the ventricular surface of circumventricular organs in the third and fourth ventricles and are regarded to be serotinergic in nature. × 15,000.

neurosecretory function for these neurons of the AP. The AP is regarded as a center of the brain that is a chemoreceptor trigger zone involved in body fluid homeostasis and may play an important role in the physiology of emesis, water balance, and blood pressure due to its critical location and the presence of a blood–CSF barrier.

V. STRUCTURAL–FUNCTIONAL CHANGES IN THE CEREBRAL VENTRICULAR SYSTEM

A. Influence of Neural Transplantation

During the late 20th century, there were many reports dealing with the neural transplantation of various fetal and adult tissue types into the central nervous system. Many of these studies focused on the transplantation of normal tissue into the brains of both humans and other mammals that suffered from neurodegenerative disorders such as Parkinson's disease, or, in the case of animals, autosomal homozygous diabetes insipidus (the Brattleboro rat). A large number of human patients underwent steriotaxic surgery for the placement of their own adrenal medulla into the region of the corpora striata of the telencephalon. The rational for this approach was based on the concept that the cells of the adrenal medulla, derived originally from the neural crest, would, in the environment of the central nervous system, undergo phenotypic differentiation and revert back into neural elements. By doing so, these autografts could secrete sufficient amounts of dopamine to overcome the deficit of Parkinson's disease due to the loss of the nigrostriatal

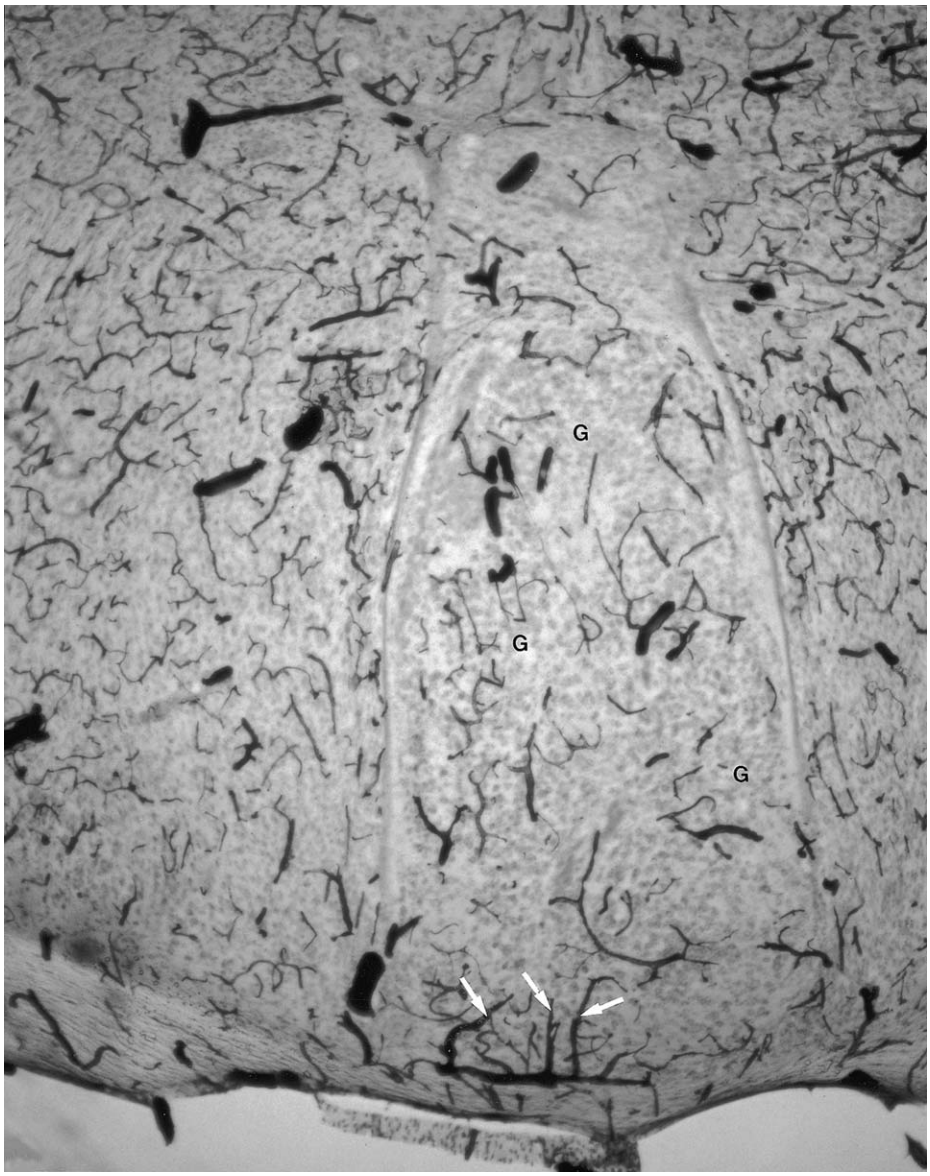


Figure 12 Fifty-micro meter-thick vibratome section of a 20-day-old fetal neural graft (G) in the hypothalamus of a Brattleboro rat with autosomal homozygous diabetes insipidus. The host–graft preparation was stained with methylene blue and reacted for endogenous peroxidase in blood vessels with diamino benzadine. Note especially how delicate branches of the primary portal plexus (arrows) make entry into the base of the graft parenchyma. $\times 400$.

dopaminergic input in parkinsonian patients. Most, if not all, of these procedures failed to produce any positive long-term effects. However, in certain species of monkeys, whose dopaminergic nigrostrigal systems were destroyed chemically, the stereotaxic placement of fetal neurotransplants into the ventricular region of the caudate–putamen proved successful and long lasting.

In the early 1980s, a number of experiments were conducted in which adult rats with homozygous autosomal diabetes insipidus were utilized. These rats received stereotaxic placement of fragments of normal fetal supraoptic hypothalamus. The SON of the Brattleboro rat is incapable of producing viable AVP because it's carrier molecule, neurophysin, contains a mutant cystine end terminal residue that fails to project AVP against enzymatic degradation. Hence, these rats suffer with profound polyuria and polydypsia, produce large amounts of

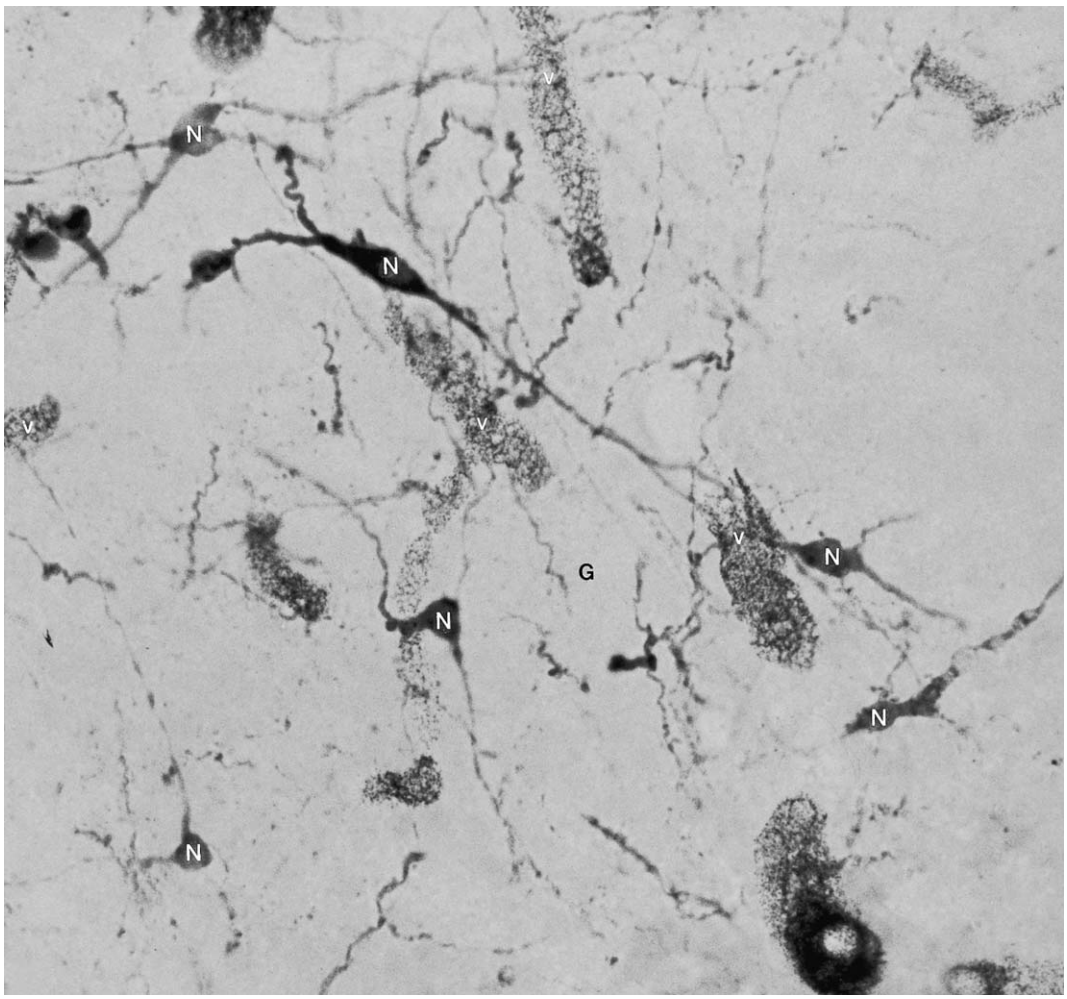


Figure 13 Immunohistochemical preparation of a fetal neurograft 20 days post-transplantation into the third cerebral ventricle of a Brattleboro rat with diabetes insipidus. Neurons (N) and their neurites that are reactive to antisera against AVP appear dark and are observed anatomically juxtaposed to blood vessels (V) that have been infused with the polymerized silastic microfil. × 700.

low-molecular-weight urine, and commonly can consume their own body weight of water in a 24-hr period. Without AVP, they are incapable of absorbing water through the epithelium of the renal distal convoluted tubules and loops of Henle. When adult Brattleboro rats received viable normal fetal hypothalamic implants via steriotaxic surgery into the third cerebral ventricle, several notable events took place. Approximately 40% of the adult Brattleboro rats began to demonstrate normal water balance properties, with a decrease in polyuria and polydipsia. By 20 days post-transplantation, many of these grafted animals had returned to normal. Upon histologic examination, a large proportion of recovering rats demonstrated the invasion of capillaries from the surrounding host parenchyma, which became anastomotic with vessels of the transplanted fetal hypothalamic fragment. These vessels commonly arose from the underlying hypophyseal portal plexus in the median eminence below that formed the floor of the third cerebral ventricle (Fig. 12). Furthermore, surviving neurons that stained positively with antisera against AVP in the fetal transplant were observed to grow their axonal processes toward capillaries of the host portal system (Fig. 13). At the ultrastructural level, these appeared to establish bona-fide neurovascular zones (Fig. 14). Hence, new neurosecretory zones were established for the transport of viable AVP from normal fetal transplants into the vasculature of adult hosts with diabetes insipitus.

Transplantation of the pineal glands also met with similar success. After adult rats were pinealectamized, all cyclic melatonin secretion was inhibited. However,

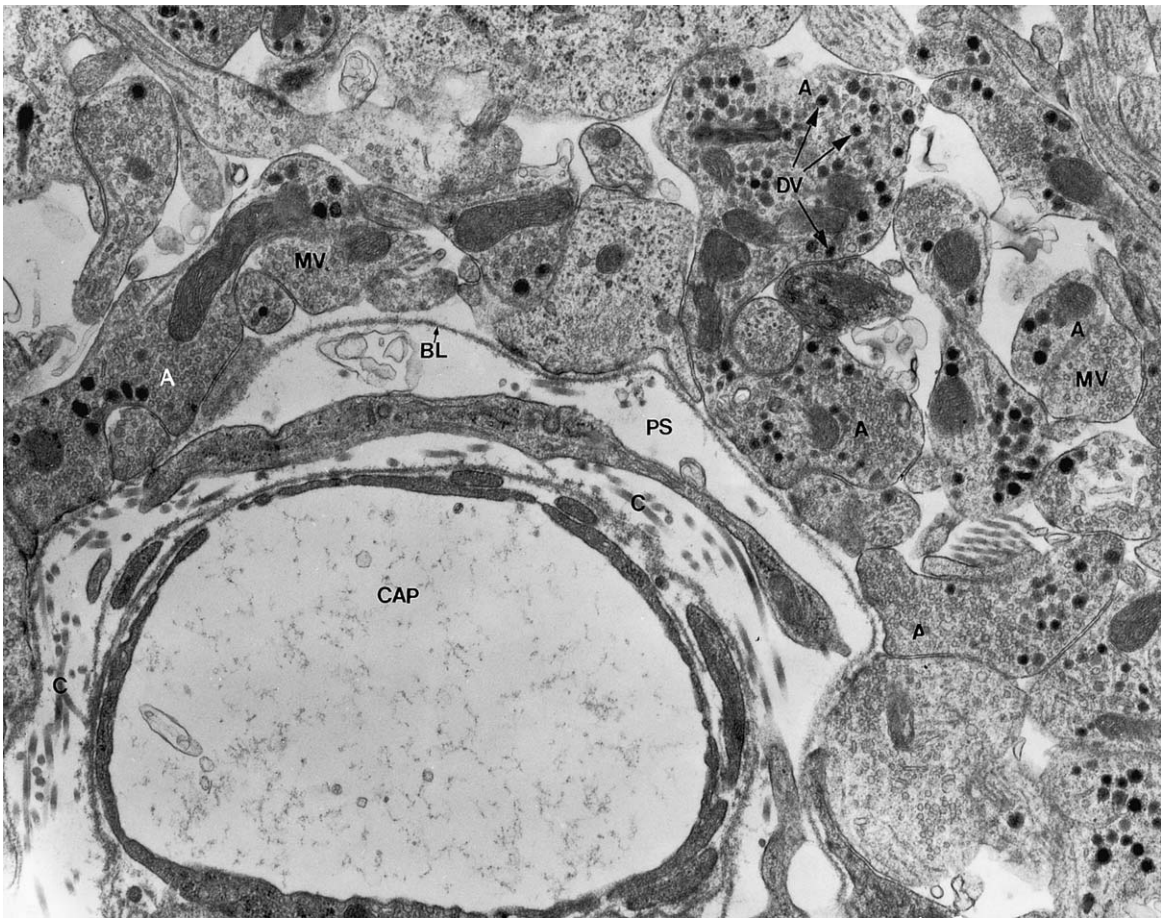


Figure 14 TEM of neurovascular zone at the base of a neural graft 20 days following neural transplantation into the third cerebral ventricle of a Brattleboro rat. Notable are numerous axon terminals (A) with dense core (DV) and microvesicles (MV) that terminate near or on the abluminal basal lamina (BL) of a large perivascular space (PS). This space is filled with collagen fibrils (C) and fenestrated capillaries (CAP) and constitutes a bonafide neurovascular zone for neurosecretion and the delivery of AVP to the host circulation of the adult Brattleboro rat. $\times$ 24,000.

when pineal glands from neonates were transplanted into pinealectomized adults, these grafts became well vascularized by fenestrated capillaries surrounded by large perivascular spaces. Two secretory processes were observed to occur in these grafted pineal transplants. The first type was neurosecretory-like, involving the formation of dense core vesicles in axon terminals. The second was ependymal in origin, involving vacuoles formed by specialized ependymal cells. Neonatal pinealocytes survived transplantation and exhibited the ultrastructural correlates of active neurosecretary processes. The synthesis and release of melatonin was reestablished in pinealectomized adults following transplantation of neonatal pineal glands into the third cerebral ventricle.

During the past 25 years, in all the transplantation paradigms involving the cerebral ventricular system little attention was ever paid to the incidence of mechanical trauma and perturbation that occurred to the cerebral ventricular wall when various fragments of normal brain tissue were stereotaxically placed into the third or fourth lateral cerebral ventricles of diseased brains. Little was mentioned about structural changes or alterations of ependymal surfaces of the cerebral ventricle in which the transplants were stereotaxically placed. In the early 1980s, a number of studies demonstrated that by 20 days following the steriotaxic transplantation of fetal hypothalamic fragments, large numbers of axon fibers and neuron-like cells were observed to emerge on the surface of the third cerebral ventricle following transplantation (Figs. 15–18). Cerebrospinal fluid contacting neurons had been previously described in a variety of vertebrate species. However, the numbers reported were far less

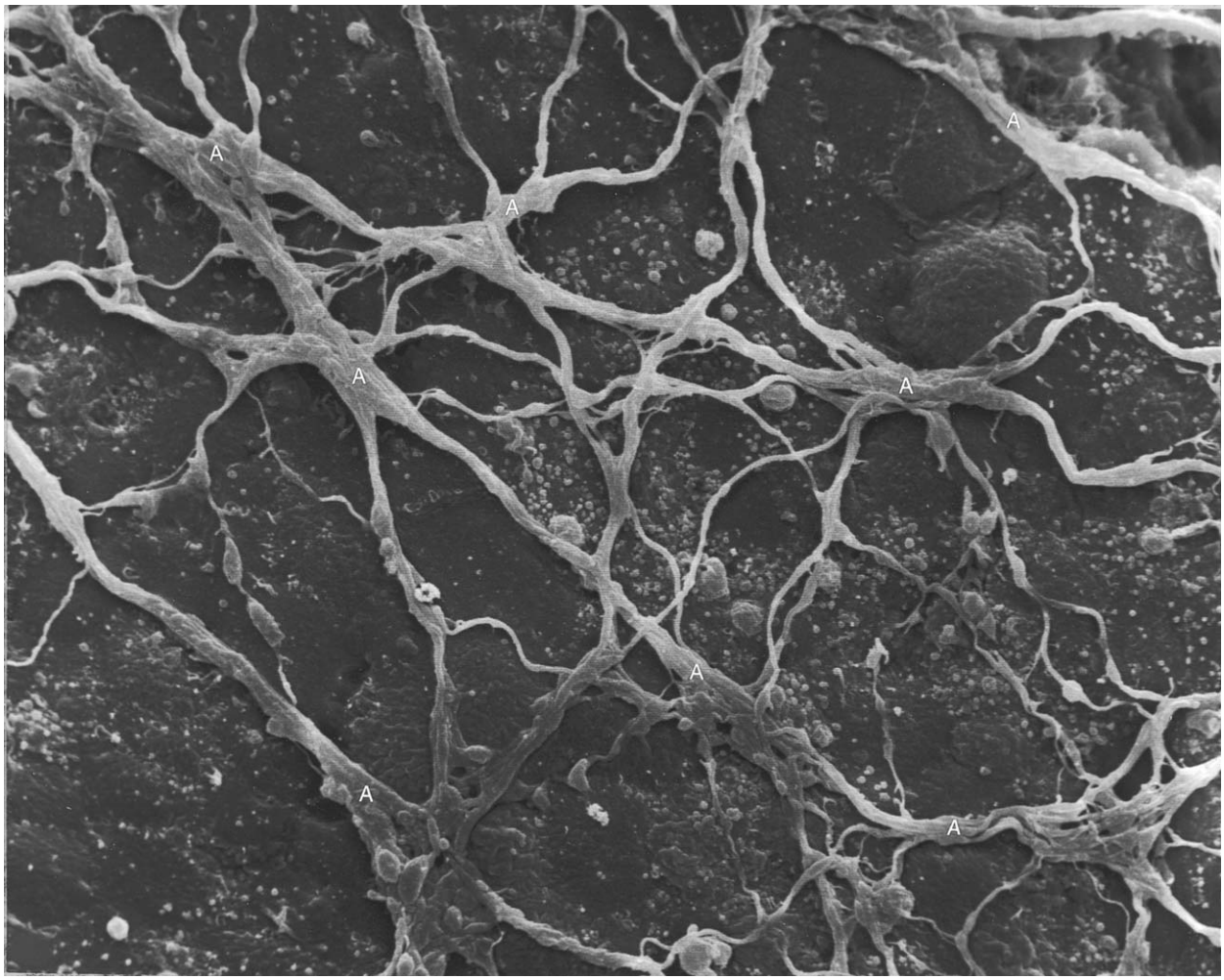


Figure 15 SEM of the third ventricular floor (median eminence) of a rat 20 days posthypophysectomy (axotomy). Numerous axon fibers (A) carpet the floor of this circumventricular organ in response to the upregulation of NOS due to trauma to the neural stalk from axotomy. Unlike the axon fiber in Fig. 11, these axon fibers are immunoreactive for antisera against AVP and are regarded as supraoptic or paraventricular in origin. × 10,500.

than those observed following transplantation or operative manipulation. Investigation of neonatal rats also described limited numbers of CSF contacting neurons that emerged by Day 3 postpartum, only to disappear by Day 7. This phenomenon was regarded as a part of the critical period of sexual differentiation in the rat, and their disappearance by Day 7 postpartum was due to a process of dying back or apoptosis. Large sample sizes were available through the use of scanning electron microscopy coupled with transmission electron microscopy on the same tissue sample, which made it easy to identify the phenotype of this supraependymal neuron. It was hypothesized that upon stereotaxic transplantation, there might have occurred mechanical perturbation and injury to the cerebral ventricular walls and floor. Such a mechanical insult or stimulation was thought to have led to the release of tissue injury factors, neuronotrophic factors, or other powerful cytokines that stimulated the migration and emergence of axonal and neuronal elements to the ventricular surface. However, other factors had to be considered. For example, it was demonstrated that in acute stress models, neuron-like cells emerged upon the surface of the third cerebral ventricles, which are ordinarily devoid of surface cells, with the exception of scattered macrophages. These neurons appeared in large numbers and were never observed prior to the stress stimulus. It was also

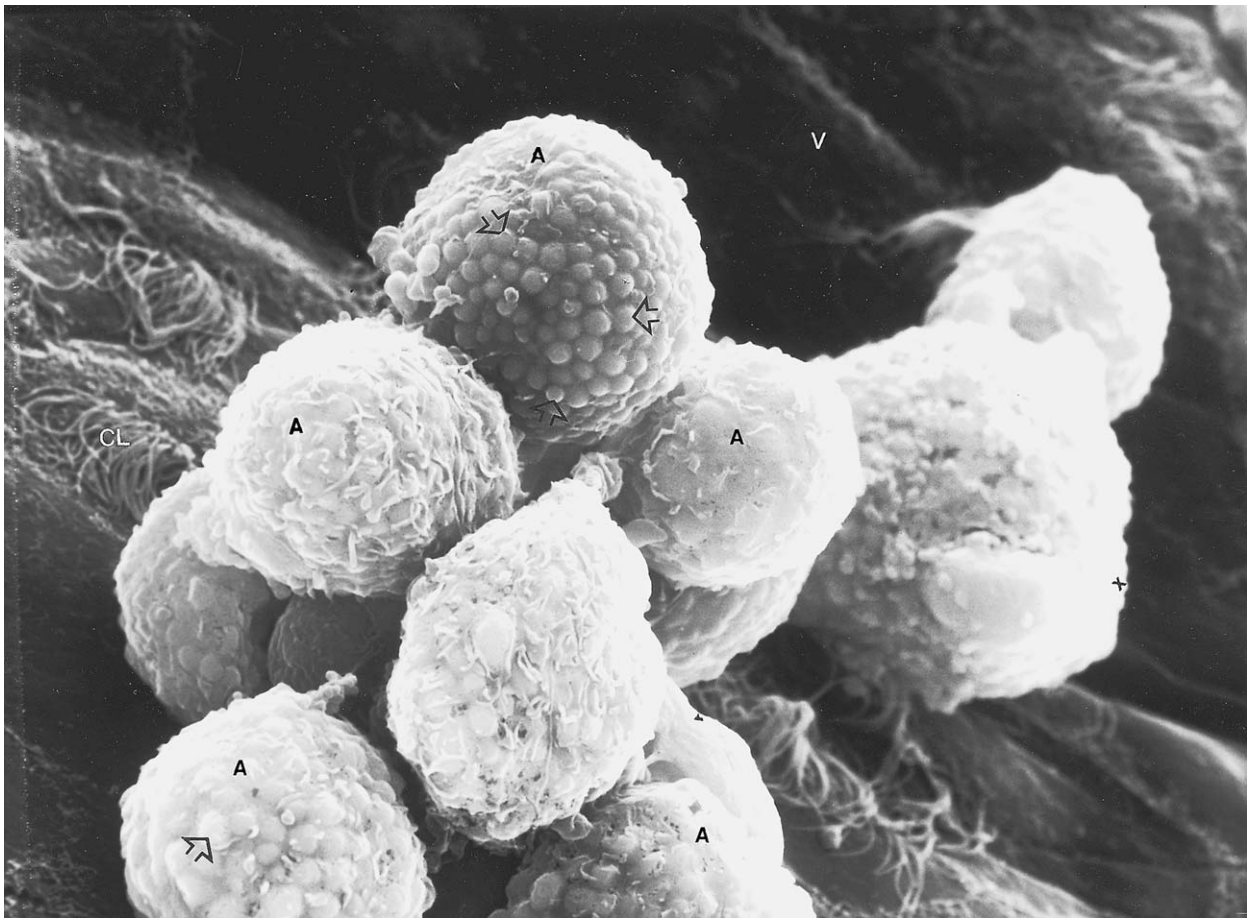


Figure 16 SEM of the cerebral ventricular floor of a rat that received the intraventricular grafting of a normal fetal hypothalamic fragment containing functional supraoptic neurons. Individual axon fibers (A) appear to contain uniform vesicular components (arrows) beneath the plasmalemma, which has shrunk from technical processing. These vesicles measure from 130 to 180 nm in diameter and fall into the size range for neurohypophyseal peptide hormones. CL, cilia; V, cerebral ventricular lumen. × 13,500.

demonstrated that following ovariectomy or the infusion of dopamine into the third cerebral ventricle, these two procedures stimulated the emergence of supraependymal neurons in large numbers coupled with numerous axonal arrays. By the mid 1990s, new techniques were being employed to assess the role of trauma and its effect on the walls of the cerebral ventricular system.

B. Trauma-Induced Changes in Ventricular Ultrastructure

In By mid-1990s, a number of investigations reconfirmed the hypothesis that the neurohypophyseal system (ME, NS, and NL) was capable of regeneration following high stalk section due to parapharyngeal hypophysectomy. Investigations employing *in situ* hybridization demonstrated that there was a significant upregulation of nitric oxide synthase (NOS) in SON and PVN following high stalk section (axotomy). NOS is the enzyme that catalyzes arginine to the ubiquitous gas nitric oxide (NO). Ordinarily, regeneration of the neurohypophyseal system took approximately 4 weeks in otherwise normal rats. The regeneration process involved the proliferation of pituiticytes, modified glial cells that acted as a structural template for the regrowth of SON and PVN axons, much as Schwan cells do for regenerating axons in the peripheral nervous system. However, the entire process of neurohypophyseal regeneration was completely inhibited by the infusion of the NO antagonist nitroarginine. This became a pivotal finding in light of the fact that during upregulation of NO,

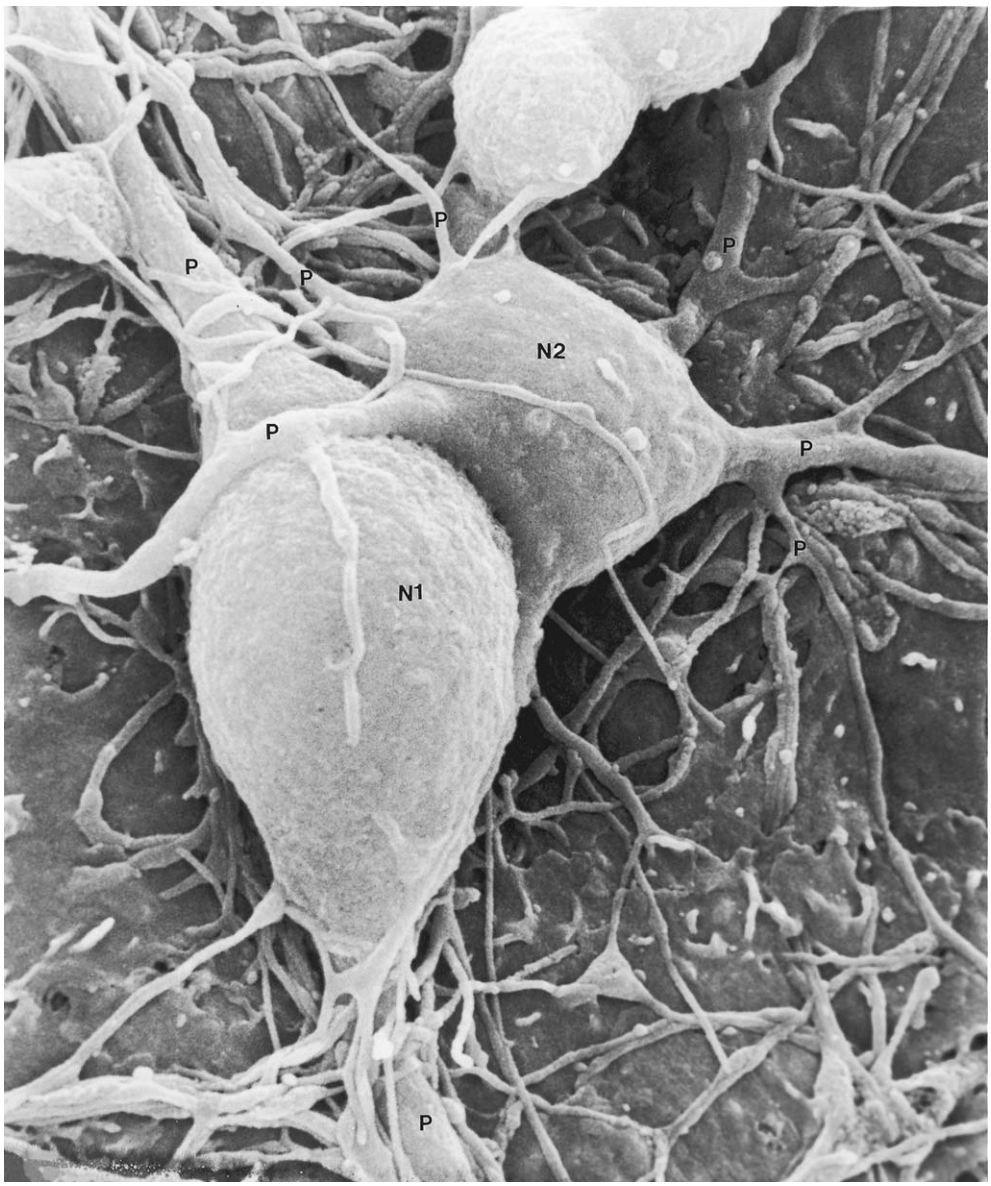


Figure 17 High-magnification SEM of two supraependymal neurons (N) that have just migrated to the cerebral ventricular surface of the median eminence. Both are relatively featureless, but both are distinctly different from one another. Cell type 1 is a bipolar neuron with two main processes (P) arising from its cell body. Cell type 2 is a multipolar neuron with numerous neuritic processes. Neither cell has undergone synaptogenesis. × 21,000. [From Scott, D. E. *Front. Horm. Res.* **9,** 15–35. By permission of S. Karger AG, Basel].

following the trauma of hypophysectomy, there was a remarkable degree of migration and appearance of supraependymal neurites (Fig. 15) and neurons on the ventricular surface of the adjacent median eminence (Fig. 19). Hundreds of axons and nerve cell bodies emerged within 5 days after hypophysectomy (Figs. 20 and 21). This appeared to be driven by the upregulation of NO. Furthermore, as in the case of neurohypophyseal regeneration, when the antagonist of NO (nitroarginine) was employed in the same animal model, these migratory neuroblasts and axonal arrays were inhibited from emerging onto the floor of the third cerebral ventricle (dorsum of the median eminence), which retained its relatively barren appearance and mimicked the picture observed in control rats. The current hypothesis being tested by a number of laboratories throughout the world is that these neuroblasts may arise from pluripotential neuronal stem

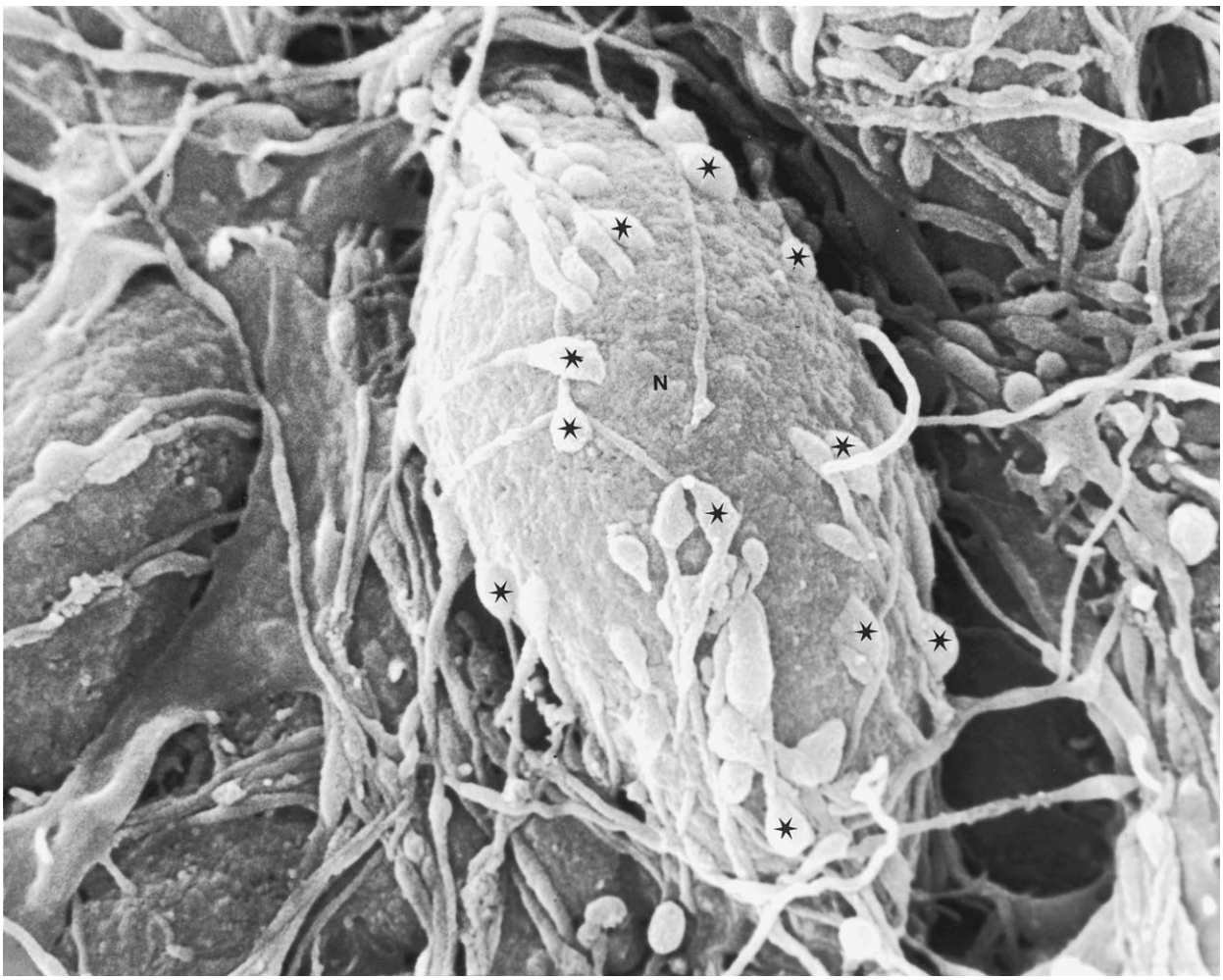


Figure 18 High magnification SEM of a supraependymal neuron (N) that has migrated and has been on the surface of the median eminence for 20 days posttransplantation. This neuron has not migrated any further. Instead, it has undergone synaptogenesis characterized by numerous axosomatic synapses (stars) on its cell body. If this population of migratory neuroblasts fail to migrate any further after the injury of axotomy, then invariably they will remain on the surface of the ventricle and undergo compensatory synaptogenesis. × 20,000. [From Scott, D. E. *Front. Horm. Res.* **9,** 15–35. By permission of S. Karger AG, Basel].

cells, which are found in germinal matrix regions throughout the adult mammalian brain.

C. Neuronal Stem Cell Migration in the Cerebral Ventricular System

The lineage and formation of neural stem cells and their relationship with the cerebral ventricular system can be traced to scientific observations in the early 1980s. However, at that time, NO, the ubiquitous central nervous system molecule, was not recognized as a mechanism that controlled migration of neuroblasts (or daughter cells) from these neural stem cells. Clearly, NO was inculcated in a diverse number of other physiological events. In the case of spinal cord injury, the presence of NO was viewed as a harbinger of apoptosis of alpha motor neurons and widespread neuronal death. Studies performed in the early 1990s revealed that it served as a powerful endothelial-relaxing factor. However, its role and upregulation in post-traumatic alterations in the third cerebral ventricle of the endocrine hypothalamus became well established as a beneficial mechanism surrounding regeneration of the neurohypophyseal system and recovery of function.

In the early 1960s, a body of evidence began to accumulate that neurogenesis persisted through

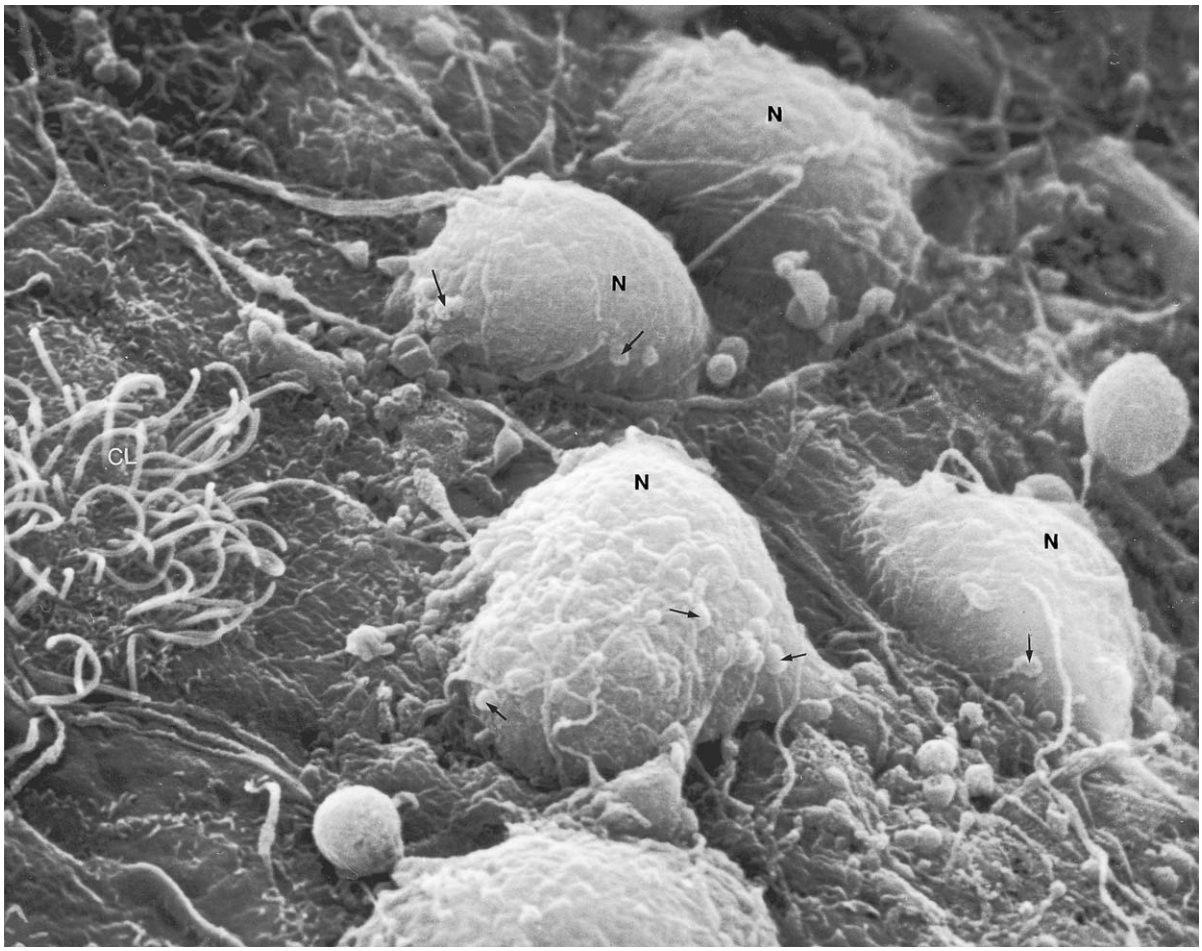


Figure 19 SEM of four supraependymal neurons (N) on the surface of the ventricle 8 days following hypophysectomy. Early signs of synaptogenesis have become apparent, with small axosomatic bouton-termineaux (arrows) beginning to appear on their cell bodies. × 6500.

adulthood in the central nervous system. Recent evidence has demonstrated that neural stem cells in the subventricular zone of the lateral ventricle of both rats and mice give rise to daughter cells (neuroblasts). In the rat model, these stem cells are stimulated to emerge by the intraventricular infusion of brain-derived nerve growth factor. In the murine model, these stem cells give rise to neuroblasts that reportedly stream rostrally beneath the ventricular ependymal cells of the lateral ventricle and ultimately replace cells that have been lost to natural turnover processes in the olfactory region. The emergence of migratory nerve cell bodies to the surface of the ventricle has recently been described following the placement of lesions in the underlying median eminence (the floor of the third cerebral ventricle). These elegant studies served to confirm early observations following neural transplantation as well as those following the high stalk neural sections done in the early to mid-1990s. Migrating neuroblasts that emerge from putative stem cells following lesions placed in the median eminence express neural cell adhesion molecules such as PSA-NCAM and B-50/GAP43, which are regarded as mechanisms that serve to expedite their migratory dynamics through the neuropil of the endocrine hypothalamus to the surface of the third cerebral ventricle above. Recent evidence strongly supports the presence of stem cells in the subventricular zone of the lateral cerebral ventricle in the adult primate brain. These newer data simply confirm earlier investigations in the murine model that daughter cells derived from stem cells migrate to the olfactory system and replace nonfunctional cells. Finally, it has recently been shown that when human neural stem cells are introduced into

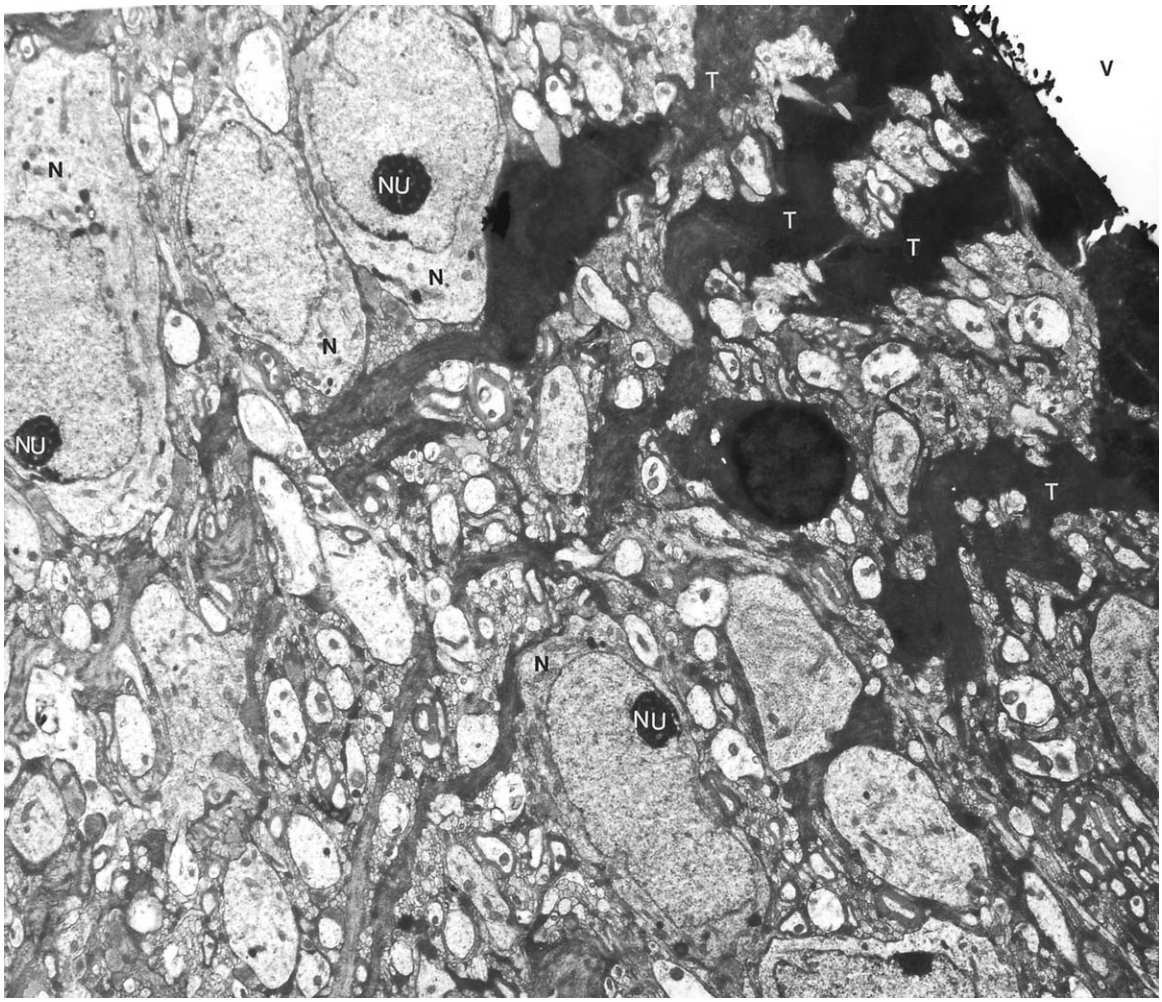


Figure 20 TEM of previously scanned median eminence 5 days following hypophysectomy (axotomy). Notable are numerous neuroblasts (N), with distinct nucleoli (NU), relatively clear nucleoplasm, and rudimentary organelles and inclusions. This sort of ultrastructural picture supports the concept of neuroblast migration from germinal zones in this circumventricular organ toward the ventricular surface above (V). T, tanycyte cytoplasm turned dark due to technical elution from previous scanning electron microscopy. × 6000.

the cerebral ventricular system of the neonatal Shiverer mouse, these cells distribute globally and use the cerebral ventricular system as a low-resistance pathway for migration to many different regions of the mouse brain. It is of interest that these human stem cells, regarded at first to be "neuronal," differentiated into oligodendroglial cells in these transgenically dysmyelinated mice. Sixty percent of the mice that received intraventricular infusion of human neuronal stem cells remained normal and exhibited none of the neurological deficits associated with the Shiverer syndrome. If indeed these were neuronal stem cells to begin with and not an earlier, more primordial undifferentiated type of human stem cell, then the concept of molecular recognition due to the lack of oligodendroglia and its product basic myelin protein must be considered.

VI. THE CEREBRAL VENTRICULAR SYSTEM AS A NEUROENDOCRINE TRANSDUCER

Since the early 1960s, the mammalian cerebral ventricular system has been regarded as a neuroendocrine transducer and its fluid contents, the CSF, a trophic mediator and delivery system for a vast array of hormones to target sites around the brain. Lines of evidence supporting such a supposition include (i) mechanisms of entry of bioactive molecules into the CSF, (ii) the transmission of such throughout the CSF,

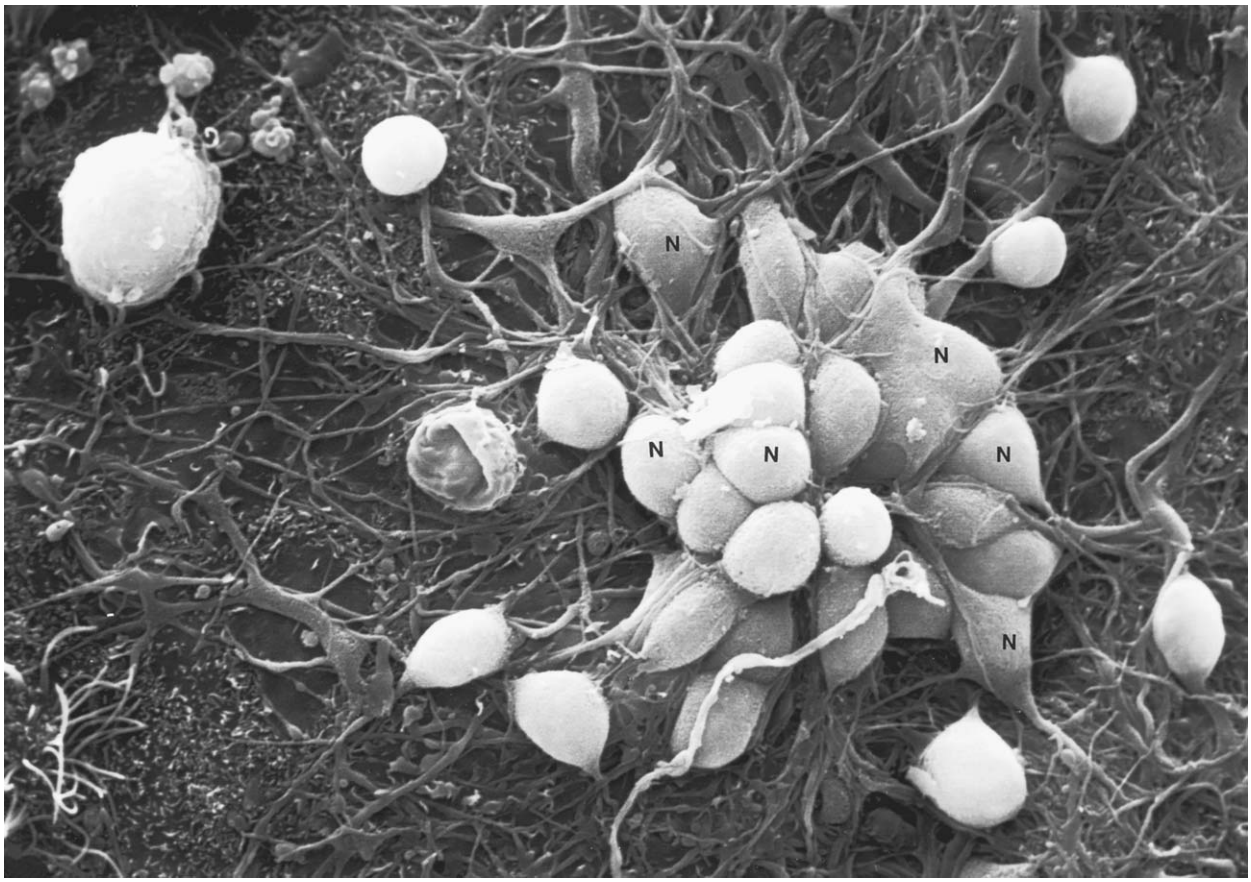


Figure 21 SEM of the surface of the median eminence 7 days posthypophysectomy. Notable is a cluster of what are regarded as juvenile neuroblasts (N) newly emergent into the third ventricular lumen. These juvenile cells are smooth but possess distinct neuritic processes. × 4000. [From Scott, D. E. (1999). Post-traumatic and emergence of a novel cell line upon the ependymal surface of the third cerebral ventricle in the adult mammalian brain. *Anat. Record* **256,** 233–241. Copyright 1999 Wiley-Liss Inc., a Subsidiary of John Wiley & Sons, Inc.]

and (iii) mechanisms of exit or delivery to specific target areas responsive to biologically active molecules. There is ample evidence that there is a wide range of biologically active factors that can be found in the CSF of normal mammals using a variety of techniques, including radioimmunoassay and other molecular techniques. The appearance of these bioactive substances may in large part be due to direct secretion by supraependymal nerve terminals from brain stem neurons of the brain stem raphe nuclei, which are known to secrete serotonin into the CSF. Others may be derived from the choroid plexus, which actively secretes AVP, angiotension, T3, and T4 as well as IGF2. However, perhaps the most important contributors to the CSF are all the CVOs that secrete a variety of neuropeptide hormones, catecholeamines, and indoleamines into the CSF. CVOs as well as the choroid plexus are also the target organs for their various bioactive molecules, which they secrete; hence, they may function in a paracrine fashion to alter the internal biochemical milieu of the brain and the CSF that bathes it. Hence, the cerebral ventricular system is not only the neuroanatomical core of the brain but also it becomes the physiological basis for the control of the internal neuroendocrine milieu of the mammalian central nervous system.

See Also the Following Articles

CEREBRAL CIRCULATION • NERVOUS SYSTEM, ORGANIZATION OF • NEUROANATOMY • PERIPHERAL NERVOUS SYSTEM • SYNAPSES AND SYNAPTIC TRANSMISSION AND INTEGRATION

Acknowledgment

Supported in part by Grant J-531 from the Jeffress Foundation.

Suggested Reading

Dellmann, H. D., and Stahl, S. (1984). Fine structural cytology of the rat subfornical organ during ontogenesis brain. *Res. Bull.* **13,** 135–145.

Gross, P. M. P. (1987). *Circumventricular Organs and Body Fluids* Vol. 3. CRC Press, Boca Raton, FL.

Johnson, C. E. (2000). Ventricules and cerebrospinal fluid. In *Medical Neurosciences* (P. M. Conn, Ed.), pp. 1–43. Lippincott, Philadelphia.

Kmgge, K. M., Scott, D. E., Kobayashi, H., and Ishii, S. (Eds.) (1974). *Brain–Endocrine Interaction II. The Ventricular System in Neuroendocrine Mechanisms*. Karger, Basel.

Nilsson, C., Lindvall-Axelsson, M., and Owman, C. (1992). Neuroendocrine regulatory mechanisms in the choroid plexus cerebrospinal fluid system. *Brain Res. Rev.* **17,** 109–138.

Scott, D. E., Kozlowski, G. P., and Sheridan, M. N. (1974). Scanning electron microscopy in the ultrastructural analysis of the mammalian cerebral ventricular system. *Int. Rev. Cytol.* **37,** 349–388.

Spector, R. (1989). Micronutrient homeostasis in the mammalian brain and spinal cord. *Neurochemistry* **3,** 1667–1674.

Williams, J. L., Thebert, M. M., Shalk, K. A., and Heistad, D. D. (1991). Stimulation of the area postrema decreases blood flow to the choroid plexus. *Am. J. Physiol.* **260,** H902–H908.

Vigilance

MARTIN SARTER and JOHN P. BRUNO
Ohio State University

GLOSSARY

acetylcholine A neurotransmitter in the peripheral and central nervous system. Acetylcholine is released in all cortical areas and layers by neurons originating in the basal forebrain. This neuronal projection represents a crucial component of the circuitry mediating sustained attention or vigilance.

anterior attention system A term proposed by Posner and Petersen and now used to conceptualize the results from mostly blood flow studies indicating that right medial–frontal and prefrontal areas are consistently activated in subjects performing vigilance tasks. The anterior attention system is also thought to act downstream via the posterior (parietal) attention system to facilitate sensory processing, particularly visual information processing.

arousal Desynchronization of forebrain activity or, functionally, a general increase in the state of mental and physical activity has usually been associated with the term arousal. Although arousal has never been well defined as a psychological construct, it is used here, and thereby dissociated from attentional processes, to describe the general activational effects of novel, stressful, or emotionally strong stimuli on forebrain information processing. Recent imaging studies in humans support the role of ascending noradrenergic projections in mediating arousal and suggest that the noradrenergic activation of the thalamus represents a critical link mediating the interactions between attentional processes and states of arousal.

basal forebrain A region in the forebrain that refers to cholinergic neurons originating in the nucleus basalis of Meynert and the substantia innominata and that project to the cortex.

noradrenaline A neurotransmitter in the peripheral and central nervous system. Noradrenergic neurons originating in the locus coeruleus and terminating in widespread telencephalic and diencephalic regions are conceptualized to mediate arousal.

positron emission tomography A functional imaging method that generates images of the distribution of a radionucleotide (such as $H_2\ ^{15}O$ or [^{18}F]-2-fluoro-2-deoxy-D-glucose) to detect changes in blood oxygenation or regional cerebral metabolic rates. Positron emission tomography is also used to monitor the distribution of more selective radiotracers for receptors and other markers of neurotransmission.

posterior attention system A component of the attentional network as conceptualized by Posner and Petersen consisting mainly of the posterior parietal lobe and thalamic and collicular circuits and mediating the spatial orientation of attention.

vigilance A component of attention, also termed sustained attention, that describes the subject's readiness to detect rarely and unpredictably occurring signals over prolonged periods of time.

The psychological construct vigilance, or sustained attention, describes a fundamental component of attention characterized by the subject's readiness to detect rarely and unpredictably occurring signals. Human imaging studies revealed frontal and parietal cortical activation associated with vigilance performance. Converging animal research has focused on cortical afferent systems, particularly on the cholinergic inputs originating in the basal forebrain, as a crucial component of the neuronal network mediating vigilance performance. Collectively, the findings from human and animal studies provide the basis for the description of the main components of the neuronal circuits mediating sustained attention.

I. INTRODUCTION

Beginning with Mackworth's experiments in the 1950s, the assessment of vigilance has typically utilized

situations in which an observer is required to keep watch for inconspicuous signals over prolonged periods of time. The state of readiness to respond to rarely and unpredictably occurring signals is characterized by an overall ability to detect signals (termed vigilance level) and, importantly, a decrement in performance over time (termed vigilance decrement). The psychological construct of "vigilance" (or "sustained attention") has been greatly developed in recent decades and provides the basis for recent studies on brain mechanisms mediating vigilance performance in humans and laboratory animals.

Vigilance represents a fundamental attentional function that determines the efficacy of the "higher" aspects of attention (selective attention and divided attention) and of the overall cognitive capacity in general. Although impairments in the ability to detect and select relevant stimuli or associations are intuitively understood to tremendously affect modern living skills (e.g., driving a car), cognitive abilities (e.g., acquiring novel operating contingencies of automatic teller machines or detecting social cues important to communicate effectively), and possibly even consciousness, psychological research on vigilance has largely focused on parametric, construct-specific issues and only rarely addressed the essential significance of vigilance for learning and memory. In fact, the evidence in support of the fundamental importance of sustained attention for general cognitive abilities has largely been derived from studies in neuropsychiatric populations. Thus, determining the brain networks mediating vigilance or sustained attention represents a crucial step in the development of cognitive neuroscience-inspired theories of neuropsychiatric disorders.

II. VIGILANCE: COMPONENTS OF A PSYCHOLOGICAL CONSTRUCT

A. Vigilance versus Arousal

Because the terms vigilance and arousal have been used interchangeably, particularly in clinical contexts and in interpreting electroencephalographic data, the specific attentional meaning of vigilance needs to be stressed and dissociated from the more global classification of brain states that includes arousal. Obviously, the ability to perform in monitoring tasks requires an activated forebrain and thus depends on arousal. Likewise, the "arousing" consequences of novel, emotional, or stress-like stimuli interact with vigilance performance. However, the operational definition of vigilance and the measures generated to describe vigilance performance are specific and unrelated to the concept of arousal. Although changes in vigilance typically are deduced from electroencephalograph (EEG) data, an interpretation of data in terms of vigilance or sustained attention is necessarily based on performance data, and the validity of hypotheses about the neuronal mediation of vigilance depends on the validity of the behavioral measures of vigilance employed to study the mediating brain mechanisms.

B. Variables Affecting Vigilance

1. Variables Taxing Vigilance Performance

Similar to other attentional capacities, the capacity to sustain attention has been considered to represent a limited resource. Several variables have been demonstrated and conceptualized as taxing vigilance performance. First, the successive (as opposed to the simultaneous) presentation of signal and nonsignal features taxes vigilance performance. Second, high (as opposed to low) event rate (i.e., the frequency of signal events) combined with unpredictability of the time of the presentation of the event (event asynchrony) and of the event type (e.g., signal vs nonsignal) enhance the demands on vigilance performance. High event rates also represent a critical variable in the manifestation of a vigilance decrement. Third, spatial uncertainty (as opposed to certainty) about the locus of event presentation also promotes the manifestation of vigilance decrements. Fourth, the use of dynamic (as opposed to static) stimuli, such as signals with variable luminance or presentation time, also fosters the manifestation of a vigilance decrement, partly because presentation of dynamic stimuli is associated with decreased discriminability. Fifth, demands on working memory (as, for example, occur in tasks with successive event presentation) tax vigilance performance. Finally, using signals with conditioned or symbolic significance, thus requiring additional processing to report detection (as opposed to pure detection of signals), is thought to foster the exhaustion of the vigilance capacity. Such signals are considered to increase the demands for the controlled processing of signals and thus to increase the allocation of resources consumed by the vigilance task.

2. Practice

In experiments using human subjects, practice of vigilance tasks typically is kept at a minimum in order to limit the degree to which the task performance is mediated by highly automated attentional processing. Prevention of the disappearance of the vigilance decrement is a related reason for allowing only minimal practice in many studies. However, practice also serves to produce stable measures of performance and, in interaction with the variables listed previously, does not necessarily attenuate the vigilance decrement. Different levels of practice may account for a substantial proportion of conflicting data in the literature, and although the issue is widely recognized, the effects of practice on vigilance performance remain insufficiently investigated. Extensive practice and the resulting high level of automatism may not decrease the demands on processing resources to the extent assumed in earlier theories. Recent conceptualizations of sustained attention as a process in which an attentional supervisory system maintains target schemata that correspond with the actual detection requirements predict that the functions of this system consume processing resources even in well-practiced vigilance tasks.

3. Reaction Time versus Number of Detected Signals and False Alarms

In highly practiced vigilance tasks in which subjects exhibit high levels of detections of signals and low levels of false alarms, reaction time may become the critical measure of performance and performance change. Increased reaction times usually correlate with decreased detection rates, supporting the hypothesis that the former measure also indicates vigilance decrements. In humans, reaction times may increase more than several hundred milliseconds during monitoring tasks that last more than 1 hr. In animals, reaction times have been analyzed and interpreted with great caution, especially following manipulation of neuronal functions, because they are potentially confounded by a multitude of sensory and motor variables and competing behavioral activities.

4. Changes in Sensitivity versus Shift in Criterion

Data generated by traditional vigilance tasks have been routinely analyzed using signal detection theory, to the extent that vigilance and signal detection theory have been considered interchangeable terms and ignoring the fact that the latter in essence represents a statistical method. However, an analysis of vigilance performance data using signal detection theory may provide the quantitative basis for an interpretation of the changes in vigilance performance. Specifically, the number of signals detected is a combined result of signal detectability (or sensitivity) and the subject's criterion or "willingness" to report detection of a signal. The former depends largely on the psychophysical characteristics of the signals relative to nonsignals, whereas the latter is a more complex function of the subject's general strategy, task instructions (e.g., performing in accordance with a conservative versus a risky criterion), and task parameters such as cost/benefit considerations for reporting signals and false alarms, respectively. For example, a decline in the number of hits (i.e., the correct detection of signals) over a prolonged period of time may be due to an increasingly conservative criterion. Experimentally induced increases in the probability for a signal typically result in a decrease in the criterion that is reflected by increases in hits and false alarms. Conversely, changes in the performance in high event rate, successive discrimination paradigms may be due to a decline in sensitivity. Signal detection theory-derived analyses potentially assist in dissociating the contribution of shifts in sensitivity and criterion to changes in vigilance performance.

III. MACROANATOMICAL CORRELATES OF VIGILANCE PERFORMANCE

A. General Heuristic Issues

Macroanatomical correlates of vigilance performance have been determined mainly by two lines of research. First, findings from neuropsychological studies have identified areas of brain damage or degeneration that are correlated with impairments in vigilance performance. Second, recently, functional imaging studies have located areas in the intact brain in which changes in metabolic activity correlate with vigilance performance. The interpretation of these findings in terms of neuronal systems mediating vigilance performance is limited by several issues. Complexities in the interpretation of brain damage in terms of determining the normal functions of the brain region of interest, although a persistent topic in the literature, deserve continued and careful scrutiny. Typically, the

functional consequences of the absence of a brain region routinely have been interpreted as indicating this area's functions in the intact brain, ignoring the more appropriate view that residual functioning reflects the processing maintained by the residual brain. This interpretational concern arises in part from, and is further convoluted by, the often implicit conception of a particular brain region as a functionally distinct unit rather than representing a functionally critical "knot" or intersection of multiple neuronal networks. For example, the prominent (and almost compulsory) involvement of prefrontal cortical areas in diverse cognitive functions may be due to their central position within cortical associational, limbic, and paralimbic neuronal networks rather than, or at least in addition to, the cognitive operations mediated via intraprefrontal neuronal circuits.

Although the assumption that attentional (or other cognitive) functions can be localized is considered to be valid, attempts to map complex functions to brain structures assume, often implicitly, that the psychological function of interest (e.g., vigilance) represents a relevant functional unit that maps on, or is isomorphic with, a traditionally defined neuroanatomical region or a neuronal system of interest. Usually, however, such an assumption is without basis (for illustration, consider the historical ideas versus the actual conceptualization of the analyses of visual information by the primary visual cortex). Neuropsychological evidence or data from imaging studies that suggest that vigilance maps onto macroanatomical correlates (such as the prefrontal or parietal cortex) will have to converge with experimental evidence to allow "strong inference" in concluding that vigilance is mediated via these structures. The evidence discussed later focuses on such converging avenues of research and presents the view that such converging evidence forms the basis for a theory that attributes vigilance performance to defined neuronal projection systems and their afferent circuits rather than to the activity within macroanatomically defined regions of the cortex.

B. Vigilance Following Brain Damage and Degeneration

Damage to the frontal cortex (anterior to the central sulcus and dorsal to the Sylvian fissure), in particular, as well as lesions in the (inferior) parietal lobe result in decreases in the number of hits, increases in reaction time, and in the manifestation or augmentation of a decrement in performance over time. The detrimental effects of distractors on sustained attention performance appear to be particularly prominent in patients with right frontal lobe damage, whereas the performance of patients with parietal lesions is generally more sensitive to the effects of high event rates. Experiments in split-brain patients support the hypothesis that vigilance processes are primarily mediated via right hemispheric mechanisms. The specificity of interpretations of the effects of right frontoparietal damage in terms of vigilance impairments has been supported by studies that excluded the possibility that increased fatigue, differential effects of practice, or differences in motivational processes significantly confound impairments in vigilance performance in these patients.

Aged humans and, more dramatically, patients with age-related dementias exhibit lower detection rates in continuous performance tasks and other standard neuropsychological tests that assess an insufficiently defined blend of arousal and vigilance functions. Because the majority of the available studies conducted in patients with Alzheimer's disease did not selectively assess sustained attention, the reliability of demonstrating a vigilance decrement in these patients has remained unsettled. Most likely, the demonstration of such effects depends on the degree to which tasks demand the sustained and effortful monitoring and processing of stimuli, thereby presumably taxing frontoparietal functions. Likewise, the extent to which impairments in sustained attention can be demonstrated, similar to impairments in selective and divided attention, in early stages in Alzheimer's disease is unclear. In general, the determination of such early attentional impairments may provide a neuropsychological measure that assists in the determination of initial pathological processes that set off the accelerating and escalating cognitive decline in these patients.

Assumptions concerning the neurobiological bases for the attentional impairments in Alzheimer's disease, including the decline in sustained attention, have generally followed the frontoparietal conceptualizations of attentional functions. It must be noted, however, that impairments in sustained attention in neurodegenerative disorders have been reflexively attributed primarily to frontal cortical degeneration, based on a circular logic that considers attentional performance to depend on the integrity of frontal cortical supervisory attentional systems and thus attributes impairments in attentional abilities in patients with neurodegenerative disorders to frontal dysfunction.

C. Macroanatomical Correlates Determined by Functional Neuroimaging Studies

The degree of agreement of evidence from lesion studies and functional imaging studies is remarkable. Anterior cingulate and dorsolateral prefrontal as well as parietal cortical regions, primarily but not exclusively in the right hemisphere, have been consistently found to be activated in subjects performing sustained attention tasks, irrespective of the modality of stimuli used in these tasks (Fig. 1). Furthermore, decline in activity in fronto-parietal-temporal regions in the course of task performance has been suggested to mediate vigilance decrements. The findings from functional neuroimaging studies have suggested that the neuronal circuits mediating the necessary arousal for proper sustained attention performance can be dissociated from those correlated with sustained attention, and that the former may be mediated via activation of the thalamus, presumably including thalamocortical projections.

In addition to the modality-independent activation of right frontoparietal regions during sustained attention, demands on monitoring and discriminating stimuli of a particular modality secondarily activate sensory cortical regions as well as sensory associational areas. These findings correspond with increases in responsiveness and selectivity of single cell firing activity observed in sensory associational areas, predominately in extrastriate regions of nonhuman primates, under conditions of increased demands on attention.

Collectively, the findings from functional imaging studies have supported a general model that describes sustained attention as a "top-down" process that begins with the subject's general readiness to detect and discriminate information, mediated largely via right frontoparietal regions, and that enhances the spatial and perceptual processes contributing to vigilance performance via activation of posterior parietal areas and facilitation of the processing of sensory inputs in primary and secondary sensory and sensory-associational regions. The available data generally support Posner and Petersen's conceptualization of a right hemispheric anterior attention system mediating modality-independent aspects of vigilance and recruiting the more posterior and temporal cortical areas to mediate relevant spatial processes and modality-specific components of vigilance performance.

IV. NEURONAL CIRCUITS MEDIATING VIGILANCE PERFORMANCE

A. "Bottom-Up" Research Approaches

The determination of neuronal circuits mediating vigilance or sustained attention is typically based on

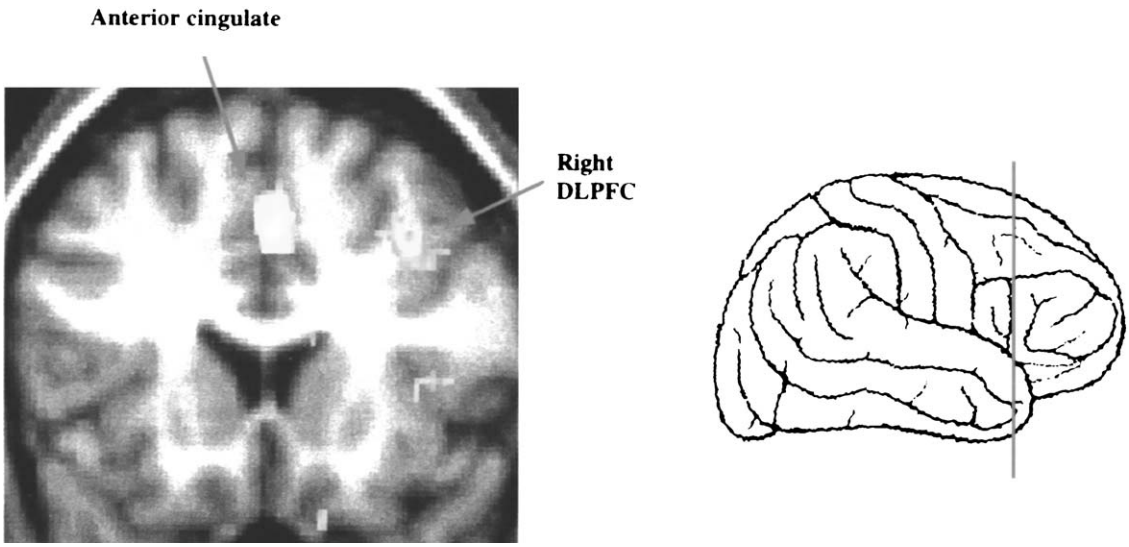


Figure 1 Regional cerebral blood flow in subjects performing a rapid visual information processing task. Right dorsolateral prefrontal cortex (DLPFC) and anterior cingulate were activated in subjects required to respond to target stimuli but not to distractors, compared to nonselective responding to all stimuli. The data from this study generally support the concept of a right frontoparietal network mediating sustained attention performance (reprinted from *Prog. Neurobiol.* **55,** J. T. Coull, Neural correlates of attention and arousal: Insights from electrophysiological, functional neuroimaging and psychopharmacology, 343–361. Copyright 1998, with permission from Elsevier Science).

(bottom-up) research approaches that manipulate the functioning of defined neuronal circuits and assess the consequences for attentional performance or record from neurons constituting such circuits in subjects that perform tasks designed to tax attentional abilities. As already mentioned, evidence from such approaches ideally converges with the evidence from human (top-down) neuropsychological and imaging studies to allow robust formulations about the neuronal mechanisms mediating sustained attentional abilities. In actual terms, experimental evidence from bottom-up research approaches is expected to explain the prominence attributed to frontal cortical and parietal regions in neuropsychological studies on sustained attention. Such reductional explanations are expected to specify the nature of the processing mediated via frontoparietal cortical areas in sustained attention, and they are intended to determine the specific neuronal circuits responsible for the activation of these areas observed in imaging studies or for the impairments in sustained attention as a result of damage or degenerative processes in these areas. Although the focus of the current discussion is the neuronal mechanisms mediating vigilance in the human brain, evidence from animal experimental approaches represents a necessary component of such an analysis.

B. Animal Behavioral Tests of Sustained Attention

Animal experimental approaches to the study of the neuronal mechanisms of attention have only recently become viable as tasks for the measurement of different aspects of attention have become available. These tasks were designed and demonstrated to be valid in accordance with the criteria constituting the psychological construct. Two tasks for use in rats have been extensively used for the study of sustained attention. First, Robbins, Everitt, and coworkers employed the "five-choice serial reaction time task" that requires animals to monitor the location of a briefly presented light in one of five spatially arranged target areas. This task combines aspects of the continuous performance tasks used in human studies and primarily tests sustained spatial attention. Second, an operant task requiring rats to detect signals and to discriminate between signals and nonsignals was designed to generate a complete set of response categories (hits, misses, false alarms, and correct rejections). Importantly, a false alarm in this task represents a discrete "claim" for a signal in a nonsignal trial, therefore overcoming the limitations of the response rate-confounded calculation of false alarm rates in more traditional signal detection tasks used in animal research. Among the parametric manipulations shown to vary task performance as predicted by the construct "sustained attention," the effects of distractors and variations in event rate or the probability for signals supported the validity of the measures of performance generated by this task in terms of indicating sustained attention. Similar to tasks used in human research, significant vigilance decrements in intact animals were not always reliably observed but usually emerged in interaction with detrimental neuronal manipulations.

C. Ascending Noradrenergic Projections: The Arousal Link?

Although noradrenergic projections originating from the locus coeruleus (LC) have been traditionally considered to mediate increases in arousal and aspects of attentional processing, their specific roles in sustained attention and arousal have remained less well defined. Psychopharmacological studies in humans and animals traditionally did not dissociate between the effects of systemic manipulations of noradrenergic neurotransmission, particularly the effects of the administration of the α_2 receptor agonist clonidine, on arousal and attention. However, findings from recent studies support the hypothesis that rather than mediating specific attentional functions, ascending noradrenergic projections regulate the more general activation or arousal of the forebrain that is required for proper sustained attention performance. Anatomically, this noradrenergically mediated foundation for attentional performance may be attributed primarily to thalamic mechanisms. Furthermore, Coull and coworkers demonstrated that the administration of clonidine in humans performing a rapid visual information processing task resulted in the reduction of thalamic activity at baseline but not during performance. The data from these studies collectively support the hypotheses that the attentional effects of clonidine depend more on the state of arousal rather than revealing a primary noradrenergic component mediating sustained attention and that thalamic circuits mediate the interactions between arousal and attention.

Aston-Jones and coworkers developed a comprehensive model of the noradrenergic contributions to attentional processes. Based largely on the sympathoexcitatory afferents of the LC, noradrenergic

projections are proposed to serve as the cognitive limb of a rapid response system. Thus, the attentional processing triggered by emotional, novel, or stress-related stimuli is "initiated" via ascending noradrenergic projections, but it is not mediated by this system. For example, the privileged attentional processing of fear- and anxiety-related stimuli has been theorized to depend on the noradrenergic activation of basal forebrain corticopetal projections. Evidence in support of regulation of the LC excitability by the prefrontal cortex indicates a top-down control of such gating mechanisms mediated via ascending noradrenergic projections. This hypothesis suggests that the attentional functions triggered, but not mediated, by noradrenergic ascending activation refer more to selection of stimuli and associations and to the reallocation of processing resources for the processing of selected stimuli, whereas sustained attention performance may less critically depend on variations in activity in ascending noradrenergic projections. This hypothesis is also supported by results from several studies indicating that lesions of the dorsal noradrenergic bundle, which deplete cortical noradrenaline contents, do not affect the performance of animals in well-practiced vigilance tasks unless the state of arousal is manipulated by the presentation of salient, stressful stimuli.

D. Basal Forebrain Corticopetal Cholinergic Projections: Sufficient Neuronal Mechanisms Mediating Sustained Attention

For several decades, human and animal psychopharmacological experiments on the effects of nicotine and muscarinic receptor antagonists (such as scopolamine and atropine) have strongly implicated cholinergic systems in sustained attention. Beginning with the

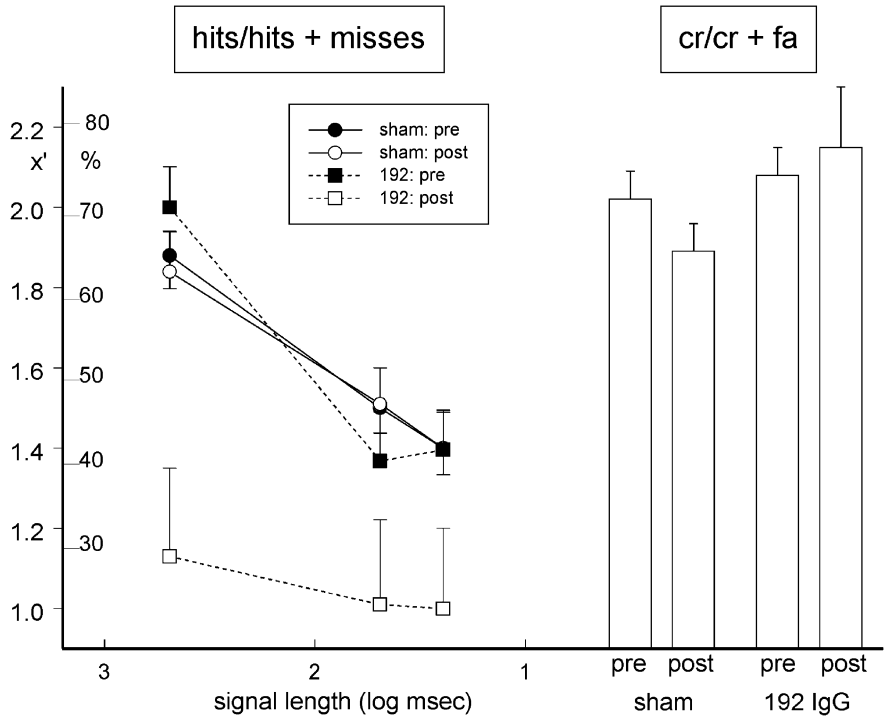


Figure 2 Effects of ^{192}IgG–saporin-induced lesions of basal forebrain cholinergic projections on the performance of rats in an operant sustained attention task. The figure depicts the effects of the lesion and the sham lesion on the relative number of hits and correct rejections (cr). The ordinate depicts the transformed percentage values (x') that were used for statistical analyses (mean $\pm$ SEM). To improve readability, the ordinate also shows percentage values assigned to appropriate transformed values. The right part of the abscissa depicts log signal length (signals were presented for 500, 50, or 25 msec). The relative number of correct rejections (i.e., the correct response in nonsignal trials) could not vary with signal length. fa, false alarms; pre, preoperative baseline; post, postoperative baseline. Preoperative baselines did not differ between the groups, and the lesion did not affect the animals' ability to correctly reject nonsignals (right). However, the lesion decreased the relative number of hits across all signals. The effects of the lesions correlated highly with the loss of cholinergic inputs to frontal and frontodorsal cortical areas (from McGaughy *et al.*, 1996. *Behav. Neurosci.* **110,** 253. Copyright 1996 by the American Psychological Association, Reprinted with permission).

demonstration of the effects of excitotoxic lesions of the basal forebrain on animals' performance in the five-choice reaction time task, the crucial dependency of attentional abilities on the integrity of this system has been extensively explored. Collectively, the available evidence demonstrates the following: First, selective lesions of the basal forebrain corticopetal cholinergic projections, produced by infusions of the cholinoimmunotoxin ^{192}IgG–saporin into the region of the nucleus basalis of Meynert and the substantia innominata in the basal forebrain, are sufficient to produce profound impairments in sustained attention. Second, the lesion-induced impairment in performance is restricted to signal trials, whereas correct rejections remain unaffected (Fig. 2). Moreover, such lesions decrease the vigilance levels as well as augment the vigilance decrement. Third, loss of cortical cholinergic inputs alone, as opposed to loss of all basal forebrain cholinergic efferents, suffices to produce such impairments. Fourth, the impairments in sustained attention observed following cortical cholinergic deafferentation are persistent and do not recover, not even following extended periods of daily practice of performance. Finally, the extent of the impairments in sustained attention correlates highly with the degree of loss of cortical cholinergic inputs, particularly in frontodorsal cortical areas.

Numerous experiments have provided strong and consistent support for the hypothesis that loss of cortical cholinergic inputs, or neuropharmacological manipulations of the excitability of basal forebrain corticopetal cholinergic projections, may represent a sufficient and possibly even necessary neuronal manipulation to affect sustained attention performance. The focus on cortical cholinergic inputs in the mediation of sustained attention is not based solely on the effects of lesions or of manipulation of the excitability of corticopetal projections. Recent studies demonstrated that acetylcholine (ACh) release in the cortex in animals performing a sustained attention task is correlated with demands on sustained attention. The hypothesis that distractor-induced increases in the demands for sustained attention are mediated via cortical cholinergic inputs has also been supported by electrophysiological studies that observed distractor-induced increases in neuronal activity in the medial prefrontal cortex that were cholinergically mediated. It should also be noted that the fact that higher forms of attention (selective and divided) also depend on the integrity of cortical cholinergic inputs is not in conflict with the hypothesis that this neuronal system primarily mediates sustained attention, because the selection of relevant and the rejection of irrelevant stimuli as well as the allocation of available processing resources to these tasks depend fundamentally on the ability to sustain attention.

The basal forebrain corticopetal cholinergic projections terminate in practically all areas and layers of the cortex. The question of whether the cholinergic inputs into a particular cortical region(s) are more crucial than inputs to other cortical regions in mediating sustained attention has remained unsettled. It is important to keep in mind that although mechanical and excitotoxic damage to medial prefrontal areas impair sustained attention performance, such effects were suggested to be due to more fundamental executive problems rather than specifically reflecting sustained attentional impairments, and they cannot be attributed primarily to the loss of cholinergic inputs. Data from studies on the consequences of loss of cortical cholinergic inputs to defined, restricted cortical regions (produced by intracortical infusions of ^{192}IgG–saporin) suggest that more widespread loss in frontal and frontodorsal areas is required to produce reliable and persistent impairments in sustained attention.

Assuming that sustained attention performance is mediated, minimally, by cholinergic inputs to widespread areas of the cortex, the issue of how the specific functions of ACh in these areas mediate sustained attention performance requires further comment. Electrophysiological recordings from neurons in sensory cortical areas following the local application of ACh or the stimulation of the basal forebrain have uniformly suggested that ACh amplifies the processing of sensory inputs. For example, such manipulations result in an increase in the responsivity of sensory cortical neurons to stimuli, an enhancement of their response selectivity to sensory stimuli, and generally in a facilitation of the detection and discrimination of sensory inputs. In the visual cortex, evidence in support of a cholinergically mediated enhancement of intralaminar transfer of information between cortical columns suggests that the effects of ACh go beyond primary sensory processes and include facilitation of higher perceptual processes. Thus, the enhanced processing of information in sensory cortical areas may not depend solely on the top-down regulation of these areas by the anterior attention system but may reflect the direct effects of ACh in sensory regions that occur in parallel and in interaction with the ACh-mediated activation of the anterior attention system (Fig. 3). It must be stressed that loss of cholinergic inputs, for example, to primary and secondary visual cortical areas alone does not result in an impairment in

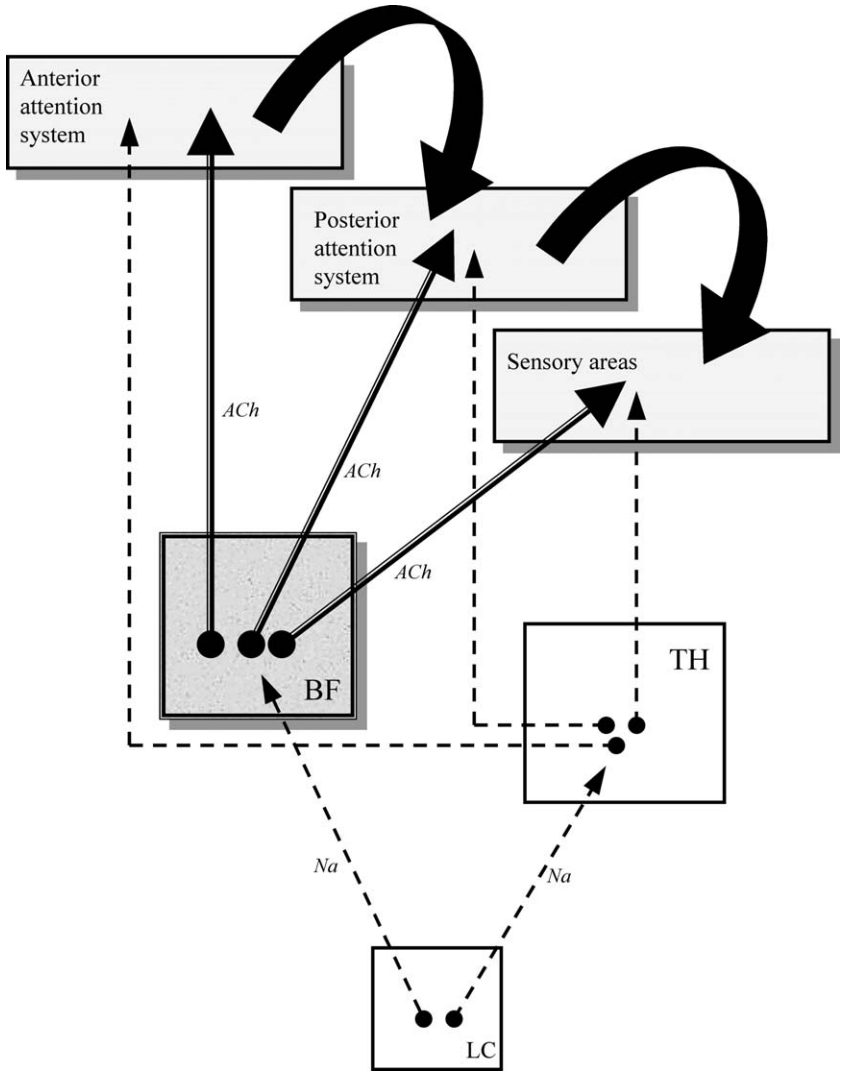


Figure 3 Schematic illustration of the major components of a neuronal network mediating sustained attention performance. The figure combines anatomical and functional relationships and represents a conceptual summary of the evidence from human neuropsychological and imaging studies and animal experimental approaches. Neuroimaging studies demonstrated consistent activation of right medial frontal and dorsolateral prefrontal cortical regions as well as parietal cortical regions in subjects performing sustained attention tasks. Activation of the anterior attention system has been suggested to modulate the functions of the posterior attention system, thereby supporting selective and spatial attentional processes, and sensory information processing in primary sensory through sensory-associational regions, thereby enhancing the sensory components of vigilance performance (curved arrows). Animal experiments determined the crucial role of cortical cholinergic inputs in sustained attention (ACh). Activation of basal forebrain (BF) corticopetal cholinergic projections is necessary for sustained attention performance, and these cortical inputs may mediate, or at least critically contribute to, the activation of frontoparietal regions. Cholinergic inputs also mediate the facilitation of sensory information processing in sensory cortical regions and may do so in interactions with the downstream modulation of sensory information processing by the anterior attention system. The necessary arousal for proper sustained attention performance is mediated largely via noradrenergic (Na) projections originating in the locus coeruleus (LC) and terminating in the thalamus (TH) and the basal forebrain. Aberrations in the crucial cholinergic activation of the anterior attention system are thought to mediate the attentional dysfunctions in major neuropsychiatric disorders (see text).

visual sustained attention performance, supporting the view that the significance of the direct sensory effects of ACh in mediating sustained attention performance depends on the converging activation of the anterior attention system and its downstream modulation of the ACh-mediated enhancement in sensory information processing. These necessary interactions between ACh-activated anterior cortical attention systems and sensory cortex (Fig. 3) may also explain the observation that loss of cholinergic inputs into single

cortical areas does not suffice to impaire sustained attention performance.

Compared to the effects of ACh in sensory cortex, information about the effects of ACh on the firing properties of individual neurons in associational areas is scarce. Generally, however, ACh appears to produce a combination of intrinsic cortical inhibition and postsynaptic increases in excitability that collectively enhance the ability of cortical neurons to process subcortical or associational inputs and thus the processing of conditioned or behaviorally significant stimuli. These cellular effects in associational areas, combined with the enhanced sensory processing, may form the basis for a cholinergic mediation of the subject's readiness to monitor the source(s) of signals, to detect such signals even if they occur rarely and unpredictably, and to initiate the higher cognitive processes that result in a decision for, and the execution of, a response that indicates detection and processing of a signal.

The prevailing focus on the cholinergic system in sustained attention should not be confused with the suggestion that sustained attention is "localized" within this neuronal system. Rather, the anatomical and electrophysiological characteristics of this neuronal system make it an ideal mechanism for gating information processing in the entire cortical mantle. Thus, in terms of localizing sustained attention, sustained attention is more adequately described as being mediated via the activation of the cortex by cholinergic inputs, specifically anterior–parietal circuits, and the interactions between the modulation of sensory processing by the downstream projections of anterior attention system, including the recruitment of interhemispheric circuits, the direct cholinergic stimulation of sensory areas and thalamic input to these regions (Fig. 3). Thalamic inputs import sensory information and, activated via noradrenergic afferents, a general state of arousal that is a prerequisite for proper attentional performance. Thus, sustained attention is a function based on distributed, parallel circuits, with the cortical cholinergic inputs being most selectively involved in the mediation of sustained attention (whereas, for example, prefrontal regions are involved in a multitude of cognitive operations) and representing a necessary limb of these circuits.

E. Convergence of Bottom-Up and Top-Down Evidence: Predictions and Gaps

The hypothesis that the activation of the frontoparietal areas observed in humans performing sustained attention tasks reflects cholinergic stimulation of these regions is attractive but remains unsubstantiated. The hypothesis would predict more widespread areas of cortical activation than reported in most of the available studies, although the relative selectivity of the activity changes reported in these studies may in part by due to the methods used to isolate regions of interest. Furthermore, the predominant role of the right hemisphere for sustained attention observed in human studies has not been addressed in the experimental literature on the functions of cortical cholinergic inputs. However, right cortical lesions produce greater sensory neglect in rats than do left lesions, and lateralized cortical muscarinic receptor densities have been reported, raising the possibility that the prominent right hemispheric involvement in sustained attention is associated with lateralized cholinergic function. Alternatively, evidence in support of a cortical lateralization of sustained attention may be associated with the type of tasks used because spatial, nonverbal sustained attention tasks by default may foster the demonstration of right hemispheric dominance in the mediation of sustained attention. As radiotracers for the visualization of presynaptic cholinergic activity and muscarinic and nicotinic receptor stimulation become available for human PET studies, future experiments will be capable of testing the hypothesis that sustained attention performance is associated specifically with increases in cholinergic transmission in the cortex.

V. ABERRATIONS IN VIGILANCE AND THE MANIFESTATION OF DEMENTIA AND SCHIZOPHRENIA

A. Sustained Hypo- and Hyperattention and Escalating Cognitive Consequences

Impairments in sustained attention detrimentally impact other aspects of attention and higher cognitive functions. For example, impairments in the ability to monitor traffic or the functions of common interfaces (such as automatic teller machines) not only acutely diminish everyday skills and performance but also, if persistent, result in the weakening of operational routines and their underlying mnemonic processing that rapidly escalates into the reduction of behavioral competence in complex situations. Because it is likely that impairments in the ability to sustain attention generalize to associational capabilities (such as the

ability to sustain chains of associations for proper processing), impairments in sustained attention proficiently and directly advance cognitive dysfunction, particularly if such impairments converge and interact with other limitations in executive functions (as is the case, for example, in aging and dementia).

In addition to impairments in the detection of signals, impairments in sustained attention may result from diverse behavioral and cognitive aberrations and manifest in a variety of changes in performance. Specifically, overattentional processes that can be explained in terms of a pathological shifts toward a "riskier" criterion, a biased monitoring of a particular source of information (e.g., anxiety-related stimuli), or an excessively comprehensive monitoring of the environment, including the subjects' own proprioceptive inputs and associational processes, may yield tremendous impairments in sustained attention. Such sustained hyperattentional impairments effectively contribute and interact with abnormalities in executive functions in human neuropsychiatric disorders, such as the exhaustion of processing resources for the unconstrained monitoring of all potential sources of information. Impairments in sustained attentional functions, broadly classified into sustained hyper- or hypoattentional impairments, represent the core symptoms of several neuropsychiatric disorders and, as discussed later, may be attributed in part to aberrations in the integrity and/or the afferent regulation of corticopetal cholinergic projections.

B. Senile Dementia

The significance of impairments in sustained attention in senile dementia, specifically Alzheimer's disease, and the sufficiency of a loss of cortical cholinergic inputs to disrupt sustained attention performance were previously addressed. The so-called cholinergic hypothesis of senile dementia—that is, the hypothesis that loss of basal forebrain cholinergic projections represents the main neuronal event mediating the emergence of dementia—continues to be debated. The available evidence strongly supports the assumption that degeneration in this system is primarily responsible for the attentional impairments in these patients that at least contribute to, and possibly initiate, the development of dementia.

Morphological as well as neurochemical measures of the loss of basal forebrain cholinergic neurons, and specifically the loss of cholinergic inputs to cortical areas, have consistently revealed significant correlations with the intellectual status of patients with Alzheimer's disease. Such correlations have not been obtained concerning the activity or density of other cortical input systems, such as noradrenergic or serotonergic inputs. Recent *in vivo* PET mapping studies of cerebral acetylcholinesterase activity confirmed the early widespread loss of cholinergic inputs in the entire neocortex in patients with Alzheimer's disease. The experimental evidence discussed previously indicates that profound impairments in attentional functions are the primary result of the cortical cholinergic deafferentation in these patients. Although treatments of patients with Alzheimer's disease with drugs that inhibit the metabolizing enzyme acetylcholinesterase have yielded heterogeneous results, including the enhancement of attentional functions, it is important to understand that traditional cholinomimetics have only a very limited capacity to "replace" the functions of absent presynaptic cholinergic neurons. Because the phasic nature of the cholinergic signal encodes crucial cognitive information, and because traditional agonists at postsynaptic cholinergic receptors and (indirectly) cholinesterase inhibitors result in excessive stimulation of these receptors that is largely unrelated to the activity of presynaptic (residual) cholinergic neurons, it is unlikely that such drugs effectively replace a partly degenerated presynaptic cholinergic system. Therefore, the limited therapeutic significance of such drugs may not serve to reject the hypothesis that the attentional impairments in senile dementia are primarily due to loss of cortical cholinergic inputs.

C. Schizophrenia

Impairments in the attentional functions, including sustained attention, have been hypothesized since Bleuler's early descriptions to constitute the core cognitive symptoms of this disorder. Unfortunately, clinical research has focused on the demonstration of impairments in standard neuropsychological tests such as the continuous performance test and only rarely attempted to identify the nature of the attentional dysfunctions in schizophrenia. The frequently used description of the breakdown in filtering functions of schizophrenics and the resulting indiscriminate monitoring of multiple sources of information (i.e., the sustained hyperattention), combined with the resulting exhaustion of processing resources, yields a diminished ability to reject irrelevant sources from being monitored in these patients (e.g., patient 13 in McGhie and Chapman stated, "If I am talking to

someone they only need to cross their legs or scratch their heads and I am distracted and forget what I was saying''). The assumption that sustained hyperattention represents a component of the fundamental cognitive impairments in schizophrenia predicts that, if sources of information are experimentally restricted, schizophrenics should show higher detection rates and reduced vigilance decrements when compared to normal people. Accumulating evidence supports the hypothesis that the sustained hyperattention in schizophrenia is mediated via a disinhibition and thus an increased excitability of cortical cholinergic inputs.

Schizophrenia is generally believed to result from developmental dysmorphogenesis. Results from recent PET studies strongly support the long-standing hypothesis that increases in dopaminergic transmission in ventral striatal regions represent a hallmark of the brains of schizophrenics. Although the cognitive dysfunction of schizophrenia has frequently been attributed to the aberrations in mesolimbic dopaminergic systems, particularly in the nucleus accumbens (NAC), neuropharmacological studies support the hypothesis that NAC dopaminergic receptor stimulation disinhibits basal forebrain corticopetal cholinergic projections via projections to the basal forebrain. The resulting overactive cortical cholinergic inputs system represents a prime candidate for a neuronal mechanism mediating the greater activation of cortical areas in response to sensory stimuli observed in functional imaging studies and, cognitively, for the hyperattentional dysfunctions in schizophrenia. Evidence from animal studies supports this assumption because sustained hyperattention, indicated by increases in the number of false alarms, was demonstrated to result from manipulations that disinhibit or even sensitize the excitability of corticopetal cholinergic projections. It should be noted that studies that assess traditional neurochemical measures indicating static levels of cholinergic transmission or that test the acute effects of traditional cholinergic drugs on schizophrenic symptoms may have very limited power in testing this hypothesis. A direct test of the hypothesis of hyperattentional dysfunctions in schizophrenic patients awaits PET studies using markers that provide sensitive indications of the phasic nature of cholinergic transmission.

VI. CONCLUSIONS

Sustained attention represents a fundamental component of the cognitive capacities of humans. Aberrations in the ability to properly monitor significant sources of information rapidly develop into major cognitive impairments. Human neuropsychological and functional imaging studies have indicated frontoparietal areas, particularly in the right hemisphere, as being prominently involved in the mediation of sustained attention or vigilance. Animal experimental evidence strongly supports the basal forebrain corticopetal cholinergic projection as a major component of the neuronal circuits mediating sustained attention performance. Future research will test the prediction that sustained attention performance-associated increases in cortical acetylcholine release mediate the activation of the anterior attention system and, in parallel, sensory cortical areas, and that the interactions between the direct effects of acetylcholine in sensory regions and the top-down modulation of sensory processing by the anterior attention system are required for proper sustained attention performance. Furthermore, hypotheses suggesting that aberrations in the integrity or afferent regulation of corticopetal cholinergic systems mediate the alterations in sustained attention as they contribute to schizophrenia, dementia, as well as other disorders, require investigations in humans, particularly using PET combined with radiotracers for markers for cholinergic transmission and muscarinic and nicotinic receptor function, as well as the generation of valid animal models on the long-term attentional effects of changes in cortical cholinergic inputs. Such studies will also clarify the significance of the hypothesis that the right frontoparietal circuits dominate over left cortical areas in the mediation of sustained attention performance.

Given the close theoretical relationships between different aspects of attention (sustained, selective, and divided), it should not be surprising that the available data from neuropsychological, functional imaging, and animal experimental studies suggest extensive overlaps in the circuits mediating different aspects of attention. For research purposes, it is important to maintain clear definitions of the aspects of attention studied. However, with respect to real-life performance, and to the underlying brain mechanisms, those differentiations may turn out to be of more limited significance. Monitoring a particular source of information requires the selection of such a source and the rejection of competing sources and also the allocation of processing resources to this task. Obviously, such tasks cannot be performed without mnemonic processing, taxing additional executive functions. Likewise, impairments in attention are difficult to conceive as

remaining restricted to a particular aspect of attention, and suggestions for such specific impairments in patients have overly relied on dissociations in performance on standardized tests. Thus, the current reductional attempts to determine the neuronal circuits mediating specific aspects of attentional functions represent a first step toward a more comprehensive understanding of the multiple circuits that allow us to effectively attend to our environment and our cognitive operations.

See Also the Following Articles

ALERTNESS • AROUSAL • ATTENTION • CONSCIOUSNESS • MENTAL WORKLOAD • NOREPINEPHRINE • STRESS

Acknowledgments

The authors' research was supported by NIH Grants NS32938, MH57436, NS37026, and AG10173.

Suggested Reading

Berntson, G. G., Sarter, M., and Cacioppo, J. T. (1998). Anxiety and cardiovascular reactivity: The basal forebrain cholinergic link. *Behav. Brain Res.* **94,** 225–248.

Coull, J. T. (1998). Neural correlates of attention and arousal: Insights from electrophysiological, functional neuroimaging and psychopharmacology. *Prog. Neurobiol.* **55,** 343–361.

Coull, J. T., Frith, C. D., Dolan, R. J., Frackowiack, R. S. J., and Grasby, P. M. (1997). The neural correlates of the noradrenergic modulation of human attention, arousal and learning. *Eur. J. Neurosci.* **9,** 589–598.

Coull, J. T., Frith, C. D., Frackowiak, R. S. J., and Grasby, P. M. (1998). A fronto-parietal network for rapid visual information processing: A PET study of sustained attention and working memory. *Neuropsychology* **34,** 1085–1095.

Kuhl, D. E., Koeppe, R. A., Minoshima, S., Snyder, S. E., Ficaro, E. P., Foster, N. L., Frey, K. A., and Kilbourn, M. R. (1999). In vivo mapping of cerebral acetylcholinesterase activity in aging and Alzheimer's disease. *Neurology* **52,** 691–699.

Lawrence, A. D., and Sahakian, B. J. (1998). The cognitive psychopharmacology of Alzheimer's disease: Focus on cholinergic systems. *Neurochem. Res.* **23,** 787–794.

Paus, T., Zatorre, R. J., Hofle, N., Caramanos, Z., Gotman, J., Petrides, M., and Evans, A. C. (1997). Time-related changes in neural systems underlying attention and arousal during the performance of an auditory vigilance task. *J. Cog. Neurosci.* **9,** 392–408.

Portas, C. M., Rees, G., Howseman, A. M., Josephs, O., Turner, R., and Frith, C. D. (1998). A specific role for the thalamus in mediating the interaction of attention and arousal in humans. *J. Neurosci.* **18,** 8979–8989.

Robbins, T. W. (1998). Arousal and attention: Psychopharmacological and neuropsychological studies in experimental animals. In *The Attentive Brain* (R. Parasuraman, Ed.), pp. 189–220. MIT Press, Cambridge, MA.

Sarter, M., and Bruno, J. P. (1999). Abnormal regulation of corticopetal cholinergic neurons and impaired information processing in neuropsychiatric disorders. *Trends Neurosci.* **22,** 67–74.

Sarter, M., Berntson, G. G., and Cacioppo, J. T. (1996). Brain imaging and cognitive neuroscience. Toward strong inference in attributing function to structure. *Am. Psychol.* **51,** 13–21.

Sarter, M., Bruno, J. P., and Himmelheber, A. M. (1997). Cortical acetylcholine and attention: Neuropharmacological and cognitive principles directing treatment strategies for cognitive disorders. In *Pharmacological Treatment of Alzheimer's Disease: Molecular and Neurobiological Foundations* (J. E. Brioni and M. W. Decker, Eds.), pp. 105–128. Wiley–Liss, New York.

Sarter, M., Givens, B., and Bruno, J. P. (2001). The cognitive neuroscience of sustained attention: Where top-down meets bottom-up. *Brain Res. Rev.* **35,** 146–160.

Turchi, J., and Sarter, M. (1997). Cortical acetylcholine and processing capacity: Effects of cortical cholinergic deafferentation on crossmodal divided attention in rats. *Cogn. Brain Res.* **6,** 147–158.

Xiang, Z., Huguenard, J. R., and Price, D. A. (1998). Cholinergic switching within neocortical inhibitory networks. *Science* **281,** 985–988.

Violence and the Brain

KAREN E. ANDERSON
University of Maryland

JONATHAN M. SILVER
Lenox Hill Hospital

GLOSSARY

catastrophic reaction Unusual behavior, usually seen after left hemispheric strokes and often seen in those individuals with aphasia. Patients may display anxiety, aggression, and extreme agitation.

explosive No gradual buildup; instead, sudden and unpredictable outburst.

ictal Behavioral changes occurring during an epileptic seizure.

interictal Behavioral episode occurring between seizure episodes.

nonreflective Does not involve planning; is not premeditated.

postictal Behavioral change occurring between seizure periods.

Explosive and violent behavior can occur with focal brain lesions as well as with diffuse damage to the central nervous system. In this article, we review neuroanatomical structures implicated in violence and the neurotransmitters believed to modulate aggressive behavior. We then discuss common neurological and medical conditions related to aggressive behavior.

I. NEUROANATOMY OF AGGRESSION

Many brain areas are involved in the production and modulation of aggression (Fig. 1). Damage due to a particular condition may be diffuse, as seen in traumatic brain injury or infection, or focal, as may occur in a small stroke, tumor, or herpes encephalitis.

A. Limbic System

The limbic system is composed of a complex set of structures, including the temporal lobes, frontal lobes, thalamus, hypothalamus, and amygdala, with extensive and far-reaching connections in the brain. The temporal lobes, especially in the amygdaloid complexes, are involved in regulation of response to basic emotions, such as fear and anger, and to drives, such as hunger, pain, and sexual behavior. The limbic system has an important role in the incorporation of sensory experience into memory so that previously unpleasant experiences can be avoided and pleasurable ones sought out.

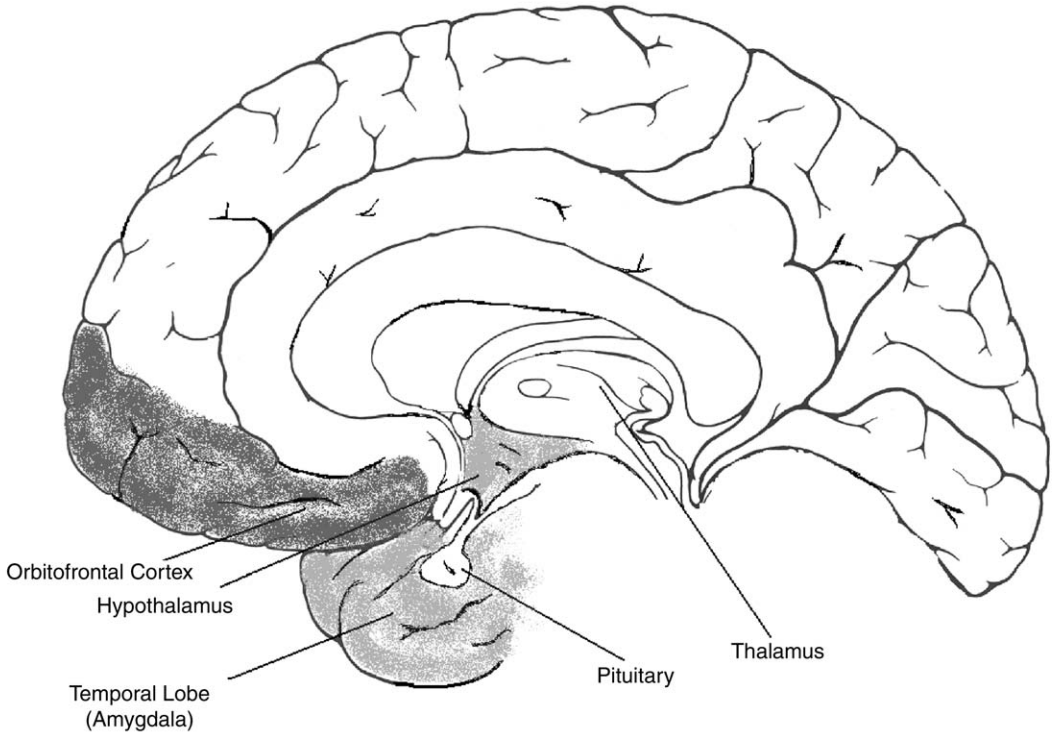


Figure 1 Regulation of aggression by brain regions: Orbitofrontal cortex modulates decreases in aggression. Limbic regions and hypothalamus modulate increases in aggression.

Early studies with animals indicated that the amygdala provides a mediating response on aggressive behavior. Klüver and Bucy first demonstrated that bilateral lesions of the amygdala produced passive behavior in untamed animals. Their work delineated the Klüver–Bucy syndrome of hyperorality, hypersexuality, the absence of fear response, increased touching (hypermetamorphosis), and visual agnosia in monkeys that had undergone bilateral temporal lobectomy. The animals could not identify which objects were associated with which drives, and they would, for example, attempt to eat nonfood items, including feces. Later work revealed that amygdalectomy in submissive monkeys could result in aggressive behavior, which concurs with the modulatory role of the amygdala on other brain functions rather than acting as a simple "aggression center."

Humans with bilateral temporal lobe damage can display behaviors similar to those seen in monkeys with Klüver–Bucy syndrome. They show placidity, apathy, bulimia, and aphasia but may also become quite aggressive at times. Temporal lobe damage in humans may occur as part of surgery or in traumatic brain injury.

Some medical conditions also cause disruption of temporal lobe function. The herpes simplex virus (HSV-1) can produce a severe form of encephalitis that, for unclear reasons, localizes in the temporal lobes. These patients often have severe neurological sequelae, including profound behavioral disturbances. They may display a Klüver–Bucy syndrome in which they are hyperoral, hypersexual, and hypermetamorphic.

Although Klüver–Bucy syndrome generally produces passive behavior, some patients with temporal lobe damage from the encephalitis are aggressive. Since aggression does not occur in Klüver–Bucy monkeys, the aggression in some patients with herpes simplex encephalitis may be due to only partial involvement of limbic areas in these cases, whereas patients with a classical Klüver–Bucy syndrome have complete involvement of both temporal lobes.

Recent functional studies using positron emission tomography (PET) or single photon emission computerized tomography (SPECT) indicate that particular aggressive individuals may have focal abnormalities in the temporal lobes, often on the left side. Some data suggest that the amygdala may be particularly prone to

kindling, in which repeated stimulation of a neural pathway leads to a lower threshold for activation of those areas. Animal models of temporal lobe epilepsy have shown evidence of kindling phenomena. Some authors have suggested that a kindling effect is responsible for the emotional outbursts seen interictally (between seizures) in a few patients with temporal lobe epilepsy.

Damage or dysfunction in the limbic area, and not temporal lobe epilepsy per se, is the significant factor in predisposing to violent behavior. Therefore, temporal lobe epilepsy may be related to aggression as a result of pathological abnormalities or dysfunction of the temporal lobe.

Numerous studies with experimental animals have demonstrated that damage to particular areas of the hypothalamus can result in rage. Decorticate cats display "sham rage" when all areas rostral to the hypothalamus are lesioned. The posterior lateral hypothalamus is a site that, when stimulated, causes the decorticate animals to hiss, extend their claws, and display other signs of aggression. Ventromedial hypothalamic stimulation may inhibit aggressive behavior in some instances, and its removal or injury have been associated with an increase in aggression, even in previously tame animals.

Several similar case reports of patients with hypothalamic damage and aggression are also cited in the neurological and psychiatric literature. In the early 20th century, tumors of the hypothalamus were found to be associated with unpredictable rage outbursts in a few patients who eventually came to postmortem. Some authors proposed that these aggressive behaviors are actually in response to heightened perception of internal stimuli, such as hunger, rather than completely unprovoked actions. With the advent of neuroimaging as a common diagnostic tool, it is now rare for patients with dramatic personality change not to undergo brain imaging early on so that tumors are discovered and treated.

B. Prefrontal Cortex

The prefrontal cortex regulates both limbic and hypothalamic activity. Thus, damage to this area may cause aggression due to lack of limbic inhibition. This area evaluates social cues and mediates an individual's judgment concerning behavior. Lesions of the prefrontal cortex may result in an exaggerated response to seemingly minor irritations. These patients may exhibit little remorse over their aggressive acts, in contrast with temporal lobe epilepsy patients, who often are quite apologetic when they have acted violently. Many people with violent or antisocial behavior have had damage to the frontal areas of the brain.

Damage to specific areas of the prefrontal cortex produces varying behavioral changes. Dorsal prefrontal lesions generally cause apathy and inability to plan. Orbital frontal lesions result in impulsive behavior. Patients with these lesions respond to environmental stimuli without evaluating the possible consequences. The most famous case of this syndrome is Phineas Gage, a railroad worker whose left orbitofrontal lobe was severely injured when a tamping iron was driven through his skull in an accidental explosion. Whereas he had previously been a sober, diligent employee, after the accident he was disruptive, rude, and garrulous. A recent case series of murderers demonstrated poor behavioral performance on classical "frontal" psychological tests and a high incidence of frontal lobe abnormalities, both structurally, [on magnetic resonance imaging (MRI) or computed tomography (CT)], or functionally (on PET or SPECT). Preliminary neuroimaging data indicate that some murderers, especially those with histories of poor impulse control, have decreased prefrontal glucose metabolism.

C. Brain Stem

Animal studies indicate that the brain stem may mediate some parts of aggressive display but does not seem to modulate complex aggressive actions. Threatening facial expressions have been seen in subhuman primates when brain stem nuclei are stimulated. Humans rarely display more than fragments of aggression resulting from brain stem lesions. Patients with rapid eye movement (REM) sleep behavior disorders are an exception to the rule that the brain stem does not mediate aggressive action.

D. Diffuse Injury/Dysfunction

Although aggressive behavior can be localized to injury or activation of some particular brain areas, confusion, delirium, and psychosis produced by general medical conditions such as anoxia or infection can also potentiate aggression. Both acute cerebral hypoxia, a deficiency of oxygen in tissue, and hypoxemia, deficiency of oxygen in the blood (usually resulting from cardiac or pulmonary compromise), may also

cause delirium. As with any delirious state, these patients may become combative. Numerous other medical conditions, such as hypertensive encephalopathy, hypoglycemia, or intracranial bleeding, can also cause delirium, occasionally with concomitant aggression.

II. NEUROTRANSMITTER MODULATION OF AGGRESSION

A. Serotonin

Serotonin exerts an inhibitory influence over aggression in animal studies. Low serotonin levels are associated with higher rates of shock-induced fighting in rats and aggression in rhesus monkeys. Studies in humans reveal an association between decreased serotonin and both aggression and impulsivity. Low serotonin has a strong association with self-destructive acts such as suicide attempts, particularly of the violent type. The serotonin system projects to the frontal cortex and thus may be particularly susceptible to damage from traumatic brain injury. However, studies after traumatic brain injury fail to show consistent decreases or increases in serotonin.

Levels of 5-HIAA, the primary serotonin metabolite, considered to be a reflection of presynaptic brain serotonergic activity, correlate with central serotonin activity. Studies of cerebrospinal fluid (CSF) 5-HIAA levels in prisoners accused of murder or attempted murder found that murderers whose acts were more impulsive had lower CSF 5-HIAA levels. Prisoners who had a history of suicide attempts or repeated violent acts had lower CSF 5-HIAA levels than those without these factors. Low serotonin levels have also been reported in persons with personality disorders who have a long-standing history of violent behavior. Thus, low serotonin may lower the threshold for impulsivity, which can lead to violence.

B. Norepinephrine

Animal studies suggest that norepinephrine enhances aggressive behavior, including sham rage and shock-induced fighting. The major norepinephrine tracts in the brain originate in the locus ceruleus and lateral tegmental system and continue to the forebrain. They are thus among the neuroanatomical regions most vulnerable to traumatic brain injury, as discussed later. Indeed, some studies have demonstrated markedly elevated norepinephrine levels after traumatic brain injury. Some, but not all, humans who exhibit aggressive or impulsive behavior have been shown to have increased levels of norepinephrine metabolites.

Monoamine oxidase (MAO) and catechol-*O*-methyltransferase (COMT) are two of the major enzymes involved in breakdown of norepinephrine and other catecholamines. Low levels of these enzymes may reflect high norepinephrine activity since catabolism is reduced. Animal studies implicate low MAO levels in aggressive behavior since MAO knockout mice, which lack the MAO-A gene, show increased aggression. Low MAO activity has been found in platelet studies of violent criminals. Data on COMT levels are less conclusive. Male knockout mice that lack COMT are more aggressive, but this increase is not seen in female COMT knockouts. Human studies are inconclusive with regard to the role of COMT and violence.

C. Dopamine

Dopamine pathways course through mesolimbic and mesocortical regions. Increases in dopamine have been shown to lead to aggressive behavior in animal studies. Dopamine increases in the prefrontal cortex have been linked to stress in humans (e.g., in studies of posttraumatic stress disorder). Chronic stress may increase dopamine in other brain regions. Agitation and, at times, aggression may be seen in schizophrenia, which is treated with antidopaminergic medications.

III. MEDICAL AND NEUROLOGICAL ILLNESS ASSOCIATED WITH AGGRESSION

Numerous conditions may cause or exacerbate violent behavior. Characteristic behavioral features occur in many patients with brain pathology who are violent. Typically, violence seen in these patients is reactive (i.e., triggered by modest or trivial stimuli). It is nonreflective, in that it does not involve premeditation or planning, and nonpurposeful, in the sense that the aggression serves no obvious long-term aims or goals. The violence is periodic, with brief outbursts of rage and aggression interspersed between long periods of relatively calm behavior. The aggression is ego-dystonic, such that patients are often upset or embarrassed by it. Finally, it is generally explosive, occurring suddenly with no apparent buildup.

Some relatively common conditions appear to have a relationship to aggressive behavior but have not been

studied empirically. Other illnesses may have a strong association with violence but occur too infrequently to permit large case studies. Studies of aggression often vary greatly in the reported frequency of this behavior in various populations. There are several factors that may influence these findings. Definitions may differ as to requirements of frequency or severity for inclusion in criteria of aggression. Some studied separate self-injurious behavior from aggression. Also, cutoff rates vary for inclusion of behavior such as damage to property or injury to others in studies of aggression. We can report which populations are more likely to show aggressive behavior and can predict some situations in which violence is more likely to be provoked in these patients. However, risk of violent behavior for an individual patients can be difficult to anticipate.

A. Dementia

Behavioral disturbances are common in individuals with dementia. Up to 70–90% of patients with Alzheimer's disease have some form of emotional or behavioral problem. Families are able to tolerate their relative being forgetful; it is more difficult to manage the sudden episodes of anger. Aggressive behavior is one of the main factors that lead to placement of a demented person in a nursing home.

In a study of agitation and cognitive impairment in nursing home residents, Cohen-Mansfield studied 408 residents of a large suburban nursing home. Impairment of activities of daily living (ADL) and cognitive impairment were associated with increased aggression. Aggressive behavior was associated with the individual being a married man, having poor social interactions, being more likely to fall, and exhibiting more aggressive behavior before admission to the nursing home. Most of the aggressive behavior was directed at caregivers and often occurred consistently around morning assistance with ADL.

Several predictors of aggressive and violent behavior have been found in Alzheimer's disease. Greater functional impairment, but not severity of cognitive decline, is a predictor of aggressive behavior in Alzheimer's patients. Delusions and misidentifications (e.g., inability to recognize one's self in a mirror or believing that someone else is living in the house) are both associated with an aggressive episodes. In a recent multicenter study of 235 outpatients with Alzheimer's disease, Devanand and others found physical aggression to be an uncommon symptom of Alzheimer's disease, persisting in only 2.8% of the sample for over 2 years. However, if physical aggression was reported during one visit, there was more than a 50% probability that it would be reported by a caretaker at the next visit. The authors concluded that physical aggression may become a more persistent problem at later stages of Alzheimer's disease since the patients became more demented as the study progressed. Changes in multiple neurotransmitter systems, including serotonin and norepinephrine, may play a role in the development of aggressive behavior in patients with Alzheimer's.

Behavioral disturbances in a subcortical dementia (Huntington's disease) and a cortical dementia (Alzheimer's disease) have been compared. Huntington's patients are significantly more aggressive than Alzheimer's patients. It is possible that damage to deep subcortical structures, such as the hypothalamus and limbic system, may result in aggression and behavioral dyscontrol.

B. Traumatic Brain Injury

Aggression and agitation commonly occur after traumatic brain injury (TBI). These behaviors, which occur during the acute stages of recovery from brain injury, can endanger the safety of the patients and their caregivers. In the acute phase after brain injury, patients often experience a period of agitation and confusion that may last from days to months. Agitation usually appears in the first 2 weeks of hospitalization and resolves within 2 weeks. Restlessness may appear after 2 months and may persist for 4–6 weeks. Subsequently, patients may develop low frustration tolerance and explosive behavior that can be set off by minimal provocation or occur without warning. These episodes range in severity from irritability to outbursts that result in damage to property or assaults on others. In severe cases, affected individuals cannot remain in the community or with their families and often are referred to long-term psychiatric or neurobehavioral facilities.

Studies of mild TBI have evaluated patients for much briefer periods of time; 1-year estimates from these studies range from 5 to 70%. Other work has examined the relationship between the number of TBIs associated with loss of consciousness (LOC) and various symptoms and has demonstrated that irritability increases with subsequent injuries. Of those men who did not have head injuries with LOC, 21% reported irritability, whereas 31% of men with one

injury with LOC and 33% of men with two or more injuries with LOC admitted to having this symptom.

C. Congenital Brain Disorders and Developmental Disorders

People with intellectual deficiencies who engage in aggression toward others or self-injurious behavior are more likely to require intensive supervision and management. Aside from the severity of intellectual impairment, aggression is the most important reason why patients are institutionalized. The mounting costs of caring for such individuals makes improved understanding of such behavior a crucial concern.

1. Inpatient Populations

Several recent studies reviewed rates of aggressive behavior among institutionalized people with intellectual disabilities. Linaker compared 57 people in an institution who had shown violence in the past year with 57 people in the same institution who had not shown violence. The groups were matched for age, sex, and level of retardation. The aggressive residents had significantly more resources allocated to them and more space available per client on the wards. Communication skills and the ability to care for themselves did not differ between the two groups. Neuroleptic drugs were given in a higher percentage in the assaultive group, and a higher level of psychopatholgy was found in this group. The number of patients with specific organic brain disorders, such as Down's syndrome or infantile autism, did not differ significantly, except that Down's syndrome tended to be more common in the control group.

Sigafoos and others studied a population of 2412 people with intellectual disability in Queensland, Australia. Of the sample, 48% were severely to profoundly mentally retarded, whereas 24% were moderately impaired and 9.6% were mildly impaired. Most individuals (59%) were male, and 16% of this population lived in institutions. Aggressive behavior was rated by a questionnaire given to those who cared for the patients. For the entire sample, hitting others with a closed or open hand was the most common aggressive behavior (68%), followed by pushing others (64%). Three or more forms of aggression were displayed by 80% of those who were violent (e.g., hitting others, kicking, verbal abuse, and pinching others). Persons who were identified as aggressive had more profound levels of retardation and lower verbal abilities. Self-injurious behavior was less common than aggression toward others.

2. Outpatients

Those people with severe intellectual impairment are more severely and frequently aggressive. In a study by Sigafoos and coworkers, aggressive behavior occurred in 17% of those in group homes and 3% of those in community-based facilities. The remainder of those studied were either in family care or living independently or semi-independently. Intake evaluations, historical data from agency records, and medical records were reviewed as sources of data on aggressive behavior. After the cohort was sorted by the presence or absence of primary referral involving explosive or assaultive behavior toward property or other individuals, 131 individuals were classified as aggressive at the time of the study. This study found that aggressive and nonaggressive patients had similar neurological histories and medical status. Central nervous system disorders, including seizures, were seen with a similar prevalence in both groups. The authors concluded that current aggressive behavior was best predicted by past aggressive behavior when found in males with lower cognitive functioning who might have been previously institutionalized. Consistent with these findings, studies of prison inmates with intellectual disabilities suggest that they have a higher incidence of offenses such as fighting and destruction of property when compared with inmates without documented intellectual deficiencies.

D. Epilepsy

Epilepsy has long been believed to be a cause of, or at least a contributor to, acts of violence. There has been substantial use of the "epilepsy defense"—persons who claim they are not guilty of certain crimes, usually violent in nature, because they were seizing or in a period of confusion following a seizure (i.e., "postictal confusion"). An examination of 75 cases in the United States, reviewed by state or federal courts and in which the epilepsy defense was used by an offender charged with commission of a violent crime ranging from rape to murder, showed that the defense was successful in only 1 case.

Estimates of interictal aggressive acts vary from 5 to 50% of individuals with epilepsy. There are several possible explanations for this variability. Many

studies of general populations of epileptics are performed at large teaching hospitals, where the patients who are referred likely have more severe disorders than those seen in a general population of people with epilepsy. Aggressive behavior can result from structural lesions that cause seizures. Medications used to treat epilepsy can alter mood and behavior. Lastly, many reports of violent behavior in persons with epilepsy are anecdotal and do not rely on standardized rating scales.

In analyzing the occurrence of aggressive behavior in individuals with epilepsy, it is important to note when the behavior occurs. Aggressive acts can be ictal (during a seizure), postictal (immediately after a seizure), or interictal (in the period between seizures).

1. Ictal Aggression

Ictal aggression occurs most often during attempts to assist or restrain a patient during a seizure. Resistive violence has also been observed in animal seizure models. Ictal aggression is rarely directed. Delgado-Escueta and colleagues reported that on video electroencephalograph(EEG) recordings of ictal behavior in 5400 patients, only 19 showed aggression. Only 1 of these 19 demonstrated actual aggression directed toward another person. The others yelled loudly, spat, or destroyed property.

2. Postictal Aggression

Postictal aggression involves violent acts that occur when a patient is still confused following a seizure. Postictal aggression is commonly seen following a general tonic–clonic seizure, and it occurs less often following a complex partial seizure. Attempts at restraint are the most typical cause of aggression during this time. Postictal psychosis is another likely cause of aggression that occurs during this period, especially if the patient experiences frightening hallucinations or paranoia during that time.

3. Interictal Aggression

The occurrence of aggressive behavior between seizures is more controversial than ictal or postictal aggression because there is no direct relationship between the aggression and the seizure event. Most epileptologists agree that the majority of patients with epilepsy are psychologically normal between seizures. It is unclear whether a small subset of persons with epilepsy behave differently between seizures as a result of brain alterations caused by the ictal events. Interictal violence is associated more with underlying psychopathology, such as schizophrenia or mental retardation, than with seizure activity.

4. Temporal Lobe Pathology

Although studies that examine general populations of individuals with epilepsy do not find a significant correlation with aggressive behavior, there is a large body of literature that links the interictal behavior of individuals with temporal lobe epilepsy to aggression. This may be due to the association of the "limbic lobe" with emotion and aggression. Patients with temporal lobe epilepsy may exhibit emotional lability, impairment of impulse control, and suspiciousness.

It was found that male sex, early seizure onset, and a history of long-standing behavioral problems were associated with aggression. Aggression was not associated with a history of psychosis.

Stevens, in reviewing several large, older studies of aggression in temporal lobe epilepsy, concluded that people with temporal lobe epilepsy are not, as a group, more violent than those with other forms of epilepsy. Damage or dysfunction in the limbic area, and not temporal lobe epilepsy per se, were the significant factors in predisposing to violent behavior. Therefore, these studies suggest that temporal lobe epilepsy may be related to aggression as a result of pathological abnormalities or dysfunction of the temporal lobe.

E. Encephalitis

The influence of viral and other forms of encephalitis on behavior was first noticed during the pandemic of encephalitis lethargica during World War I. Also known as Von Economo's encephalitis, the illness was noted to produce a plethora of psychiatric symptoms, including behavioral changes such as aggression in previously normal persons. Other forms of encephalitis, most notably herpes encephalitis, are now also known to result in aggressive behavior.

1. Encephalitis Lethargica

Encephalitis lethargica was first described in detail by Constantine von Economo in 1917. The agent that causes encephalitis lethargica has not been isolated. Besides producing physical symptoms of an acute central nervous system infection, encephalitis lethargica can, at times, progress to coma or death. Survivors were sometimes afflicted with parkinsonism or with

bizarre behavioral disturbances. Some survivors, mostly adolescents, suffered pseudopsychopathic states. In a review of early work on the illness, children who had been normal became uninhibited, damaging property and attacking strangers in the street. Few neurological abnormalities were seen in these cases. Sporadic cases of encephalitis lethargica are still seen throughout the world.

2. Herpes Simplex Encephalitis

The herpes simplex virus (HSV-1) can produce a severe form of encephalitis. It is probably the most common cause of nonepidemic encephalitis in temperate zones. Mortality rates are as high as 70%. However, with new antiviral treatments, many patients are living longer. These people often have severe neurological sequelae, including profound behavioral disturbances. For unclear reasons, the herpes virus tends to specifically destroy the temporal lobes. This can produce a Klüver–Bucy syndrome in which patients are hyperoral, hypersexual, and may have an abnormal desire to explore objects (hypermetamorphosis).

3. Other Forms of Encephalitis

Since herpes simplex encephalitis is the only encephalitis known to localize to a particular brain area, it is the only encephalitis in which neuropsychiatric symptoms may be predictable. Other types of encephalitis produce more diffuse central nervous system damage.

a. Autoimmune Deficiency Encephalitis Autoimmune deficiency syndrome (AIDS) causes numerous behavioral deficits, including dementia. Another common behavioral disturbance associated with agitated behavior in AIDS patients is AIDS-related mania. Patients can display symptoms seen in typical mania, including irritability and psychosis. AIDS encephalitis has been reported to cause depression and psychotic symptoms in some patients.

b. Limbic Encephalitis Paraneoplastic syndromes, a remote effect of malignancy generally believed to be autoimmune in nature, are the most common cause of limbic encephalitis. Due to its involvement with the limbic system, aggression can be seen as a result of the condition.

F. Movement Disorders

Movement disorders often present with behavioral changes (Table I). Increased irritability and angry outbursts are reported in many movement disorders. Psychotic symptoms, which exacerbate underlying aggression, are also seen.

1. Gilles de la Tourette Syndrome

Gilles de la Tourette syndrome is a condition in which patients display both verbal and motor tics. These tics change over time in localization and presentation. Patients with Tourette's may display coprolalia (obscene speech) and copropraxia (obscene gestures). Numerous authors have cited behavioral problems as part of the clinical picture in some cases of Tourette's syndrome. Indeed, Tourette mentioned behavioral problems in his original description of the syndrome. Although obsessions, compulsions, and disorders related to attention deficit hyperactivity disorder are generally the most commonly described psychiatric symptoms in Tourette's, aggressive outbursts have also been described. Aggressive behavior was significantly associated with symptoms of being forced to touch and with copropraxia. There was no significant association between aggression and age of onset, personal or family history of psychiatric illness, EEG or neurological abnormalities, medication, distribution of tics, hyperactivity, or difficulty in concentration or attention as a child.

2. Wilson's Disease

Wilson's disease, or hepatolenticular degeneration, is an autosomal-recessive disorder involving dysregula-

Table I
Movement Disorders Associated with Aggressive Behaviors

Disorder	Comments
Tourette's syndrome	Aggression often associated with attention deficit or thwarting of urge to touch/manipulate
Wilson's disease	Psychiatric symptoms may be presenting complaint, including changes in personality and temperament
Huntington's disease	Irritability in more than 40% of patients, often associated with rigid behavioral patterns
Parkinson's disease	Aggression rare, generally associated with confusion or psychosis due to anti-Parkinson's medications

tion of copper metabolism by the liver. Psychiatric symptoms in Wilson's disease were recognized as early as 1912 by Samuel Alexander Kinnier Wilson in his description of the disorder. Neurological, renal, and hepatic abnormalities are the usual findings in the disease. Kaiyser–Fleischer rings can often be seen on fundoscopic examination. All symptoms are caused by abnormal deposition of copper in organ systems. Some studies have reported that more than half of all Wilson's disease patients have psychiatric symptoms as their presenting complaint. Personality changes are generally the most common psychiatric complaint and may include aggressive behavior. Family members may describe a patient as having more "temper" than usual or yelling and throwing things when angry. Psychosis, which could contribute to aggression, may also occur. There may be an association between incongruous behavior, aggression, and personality change and certain neurological symptoms, such as dysarthria, dysphagia, drooling, and rigidity.

3. Huntington's Disease

Huntington's disease is an autosomally dominant disorder that can present with choriform movements and/or psychiatric symptoms. Aggression and irritability may be seen in these patients and are frequently reported to be among the most common psychiatric features of the illness. Violence may occur in these patients due to degeneration of frontal–striatal pathways and subsequent behavioral disinhibition. Psychosis is also occasionally seen in Huntington's patients.

4. Parkinson's Disease

Although aggression may not be a common manifestation of Parkinson's disease, it can develop as a result of treatment with dopaminergic medications, which may precipitate psychosis.

G. Brain Tumors

Aggression secondary to intracerebral malignancy has long been a part of the differential diagnosis considered by psychiatrists. However, it is rare for violent behavior to manifest itself as the first symptom of a brain tumor. We believe that the more frequent use of brain imaging techniques has resulted in earlier detection of tumors, resulting in less frequent behavioral presentation. To date, there have been few studies of central nervous system malignancies as a cause of aggressive behavior. Generally, behavioral symptoms attributed to brain tumors can be difficult to correlate with location since tumor growth, especially when rapid, can cause widespread edema, bleeding, and large mass effects. Often, patients present with lethargy as the predominant behavioral problem. There are a few case series, especially in the older literature, in which attempts have been made to correlate tumor location with manifestation of behavioral disturbance.

Patients with tumors in the temporal lobe may have irritable, angry, or actual assaultive episodes during their clinical courses. Tumors causing aggression vary in histological type but are all relatively slow growing, disrupting mainly limbic system structures, and not causing rapid shifts in intracranial pressure. Limbic system tumors may cause seizures, and these patients' psychiatric symptoms may be related to the behavioral disturbances seen in temporal lobe epilepsy. Numerous studies with experimental animals have demonstrated that damage to the ventromedial hypothalamus can result in hyperphagia and, occasionally, rage.

H. Stroke

Mood disorders following stroke have been reported in numerous case studies and reviews. Robinson's extensive studies on the relationship between lesion location and depression or mania suggest that patients with left hemisphere lesions are more likely to become depressed, whereas those with right-sided lesions are more likely to be unusually cheerful or even manic.

Catastrophic reaction to stroke is a relatively rare mood disorder seen most commonly after left hemispheric strokes, although it is also described in other types of brain damage. It is often seen in those individuals with aphasia. Patients may display anxiety, aggression, and extreme agitation. Some authors attribute this to frustration at not being able to communicate due to the aphasia.

I. Hypoglycemia

A series of studies conducted in Finland examined biologic correlates of aggression in a group of violent prisoners. One consistent finding was that this group is

prone to hypoglycemia, and they have an increase in irritability during these episodes.

The main criticism of studies of violent populations is the confounding effect of alcohol. Since many of these patients have a history of extensive alcohol abuse, especially while perpetrating violent acts, it is unclear what contribution alcohol makes to the overall picture of violence. Some authors have questioned the usefulness of the oral glucose tolerance test as a measure of blood glucose levels in these circumstances. The differences in diet of patients compared with that of controls can be quite disparate for institutionalized persons. Currently, these results are not useful predictors of violent behavior in the general population.

J. Sleep Disorders

There have been a few reports of aggressive behavior during parasomnias, including violent attacks. Patients often seem to be fleeing from imaginary pursuers and may assault family members who happen to be in the room with them. Formal sleep studies can help determine the cause of this sleep-related behavior, which includes non-REM parasomnias, sleep-related seizures, REM behavior disorder, or psychogenic dissociative phenomena.

K. Endocrine Disorders

In addition to systemic effects, several endocrine disorders are known to produce behavioral symptoms and often have psychiatric symptomatology as the presenting complaint.

1. Thyroid Disorders

Hyperthyroid states can produce agitation and psychosis. Thyrotoxicosis is caused by increased production of thyroid hormone by the thyroid gland, inflammation of the thyroid, or sources of thyroid hormone from outside the gland (i.e., exogenous thyroid hormone consumption or tumor production of thyroid hormone). Neuropsychiatric symptoms of thyrotoxicosis include anxiety, insomnia, and intense dysphoria. Memory and concentration can be impaired. Psychosis is sometimes present; when it is, violence generally occurs since patients may react dramatically to perceived threats or command hallucinations. When psychosis is seen, it may be a sign that the patient is progressing toward the more serious state of thyroid storm.

Thyrotoxicosis in children can present in numerous ways, including an attention deficit hyperactivity disorder-like picture, affective symptoms, and psychosis. Severe aggression is rarely reported in children with thyrotoxicosis.

Hypothyroidism is usually primary; pituitary or hypothalamic disease causing hypothyroidism is rare. In psychiatric patients, lithium can cause hypothyroidism. Patients with hypothyroid conditions are usually characterized as depressed, apathetic, and lethargic. Elderly patients may look demented. However, there are case reports of agitation leading to violent behavior in persons with hypothryroidism.

Approximately half of the patients with hypothyroidism will have some affective disturbance, and a small percentage of these cases will show signs of psychosis. Long-standing hypothyroidism can lead to myxedema. Patients with myxedema psychosis (or "myxedema madness", as severe psychiatric disturbance in hypothyroid states was once termed) can become quite agitated.

2. Parathyroid Disorders

It has long been recognized that parathyroid disorders, and the resulting calcium abnormalities, can cause psychiatric symptoms. Since calcium is involved in neurotransmission, it is not surprising that hypercalcemia can have many psychiatric manifestations. Primary hyperparathyroidism can be caused by a tumor of the parathyroid or by parathyroid hyperplasia. The condition usually causes depression, fatigue, and confusion. However, psychosis has been reported in rare cases and can lead to violent behavior. Patients with hypercalcemia are also more likely than those with other endocrine disorders to display psychiatric symptoms with concomitant fully intact consciousness. There is some indication that patients with higher calcium levels are more likely to have psychotic symptoms.

Hypoparathyroidism can cause hypocalcemia. Hypoparathyroidism is usually associated with autoimmune disorders or other destruction of the parathyroid. Hypocalcemia has not been reported to cause violence. Patients with this disorder are often anxious and demented.

3. Cortisol Disorders

Patients with Cushing's syndrome due to adrenal tumors or adrenocorticotropin hormone (ACTH) secreting tumors have high levels of cortisol, which

can cause metabolic encephalopathy and various psychiatric symptoms. Psychotic symptoms can be precipitated by hypercortisolism. Older literature cites an example of homicidal psychosis after exogenous ACTH use. Patients with Addison's disease, who have abnormally low cortisol levels, are generally depressed.

IV. MEDICATIONS AND DRUGS

Drug effects and side effects can result in disinhibition or irritability. In any individual with brain dysfunction, there is increased susceptibility to adverse behavioral effects of medication. By far the most common drug associated with aggression is alcohol, during both intoxication and withdrawal. Stimulating drugs such as cocaine and amphetamines, as well as stimulating antidepressants, may produce agitation. Antipsychotic medications often increase agitation through anticholinergic side effects, and agitation and irritability usually accompany severe akathisia. Many other drugs may produce confusional states, especially anticholinergic medications that cause agitated delirium. Other drugs that may produce aggressive behavior include steroids, analgesics (opiates and other narcotics), and anxiolytics (barbiturates and benzodiazepines).

V. SEX HORMONES

A. Testosterone

High testosterone levels are often blamed for aggressive behavior in men, although the actual relationship between testosterone levels and violence is far from clear. Several controlled studies of testosterone levels in serum, saliva, and spinal fluid have attempted to correlate these values with various measures of violence.

Several groups have examined testosterone levels in male prisoners who committed violent acts, finding some correlation between elevated testosterone and aggression. Some studies have found higher serum testosterone levels in continuously violent groups of prisoners. The socially dominant prisoners have also been shown to have higher testosterone levels than the nondominant, nonaggressive prisoners.

In contrast, other studies found no correlation between plasma testosterone and incidences of fighting and verbal aggression in prisoners. Those prisoners with a more prominent history of violent crime in adolescence did have a higher level of testosterone than those without a significant early history of violence. The authors propose that high levels of testosterone may have a permissive effect on aggression early in life in some individuals who are more prone to commit crime in the first place. However, it is unclear whether high testosterone levels in adulthood correlate with increased testosterone early in life. Testosterone may exert a modulatory effect in decreasing the threshold for violent crimes. However, most violent crime is not directly correlated with testosterone levels. Future studies using saliva or CSF testosterone may be more conclusive.

B. Premenstrual Dysphoric Disorder

Popular wisdom has linked the menstrual cycle to mood swings and unusual behavior since the early 1900s. However, there are few controlled studies that address the question of whether violence in women is linked to a certain phase or phases of the menstrual cycle.

The term premenstrual syndrome, which has been replaced with the term premenstrual dysphoric disorder, is used to describe a number of physical and/or mood changes observed in a small percentage of women in association with their menstrual cycles. The psychiatric symptoms include tiredness, depression, and, in some women, irritability and anxiety.

Several studies have examined the relationship between stage of menstrual cycle and commission of a crime. These studies, which rely on retrospective self-report by women, may be confounded by popular opinion, which may cause women to falsely recall that they were paramenstrual at the time a crime was committed as an explanation or excuse for the behavior.

Thus, the relationship between the menstrual cycle and aggression is unclear. Women who commit crimes may be more likely to do so during the premenstrual or menstrual phase of their cycle. However, these data may be influenced by suggestibility because women may retrospectively attribute their actions to a particular phase in their menstrual cycle.

C. Prenatal Exposure to Exogenous Hormones

Prenatal exposure to androgens occurs mainly when synthetic androgens are used to prevent abortion. It has been suggested that the exposure to androgens would masculinize females and could lead to higher

levels of aggression. Fortunately, these cases are quite rare. Exposed children were compared with their same-sex siblings, who had not been exposed, using questionnaires about aggression. Both exposed boys and girls scored higher on the questions that predicted potential for aggression. However, no actual observation of whether the children displayed more aggression than their siblings was reported. Prenatal diethylstilbestrol exposure was examined No increase in verbal or physical aggression was found in exposed females when compared with controls. In summary, human studies on prenatal exposure to sex hormones do not provide conclusive evidence that exposure to androgens leads to an actual increase in violent behavior.

VI. NUTRITIONAL FACTORS

Vitamin deficiencies generally present with apathy as the foremost behavioral symptom, but some nutritional deficiencies are reported to cause agitation and violence, usually as a result of psychosis (Table II). Pyridoxine (B_6) deficiency due to homocystinuria may cause psychosis. Folic acid deficiency in homocystinuria and hypomethioninemia may cause a schizophrenia-like syndrome.

There are a few case reports of dietary vitamin deficiencies causing agitation. Niacin deficiency can cause depression with psychotic features, including hallucinations and thought disorder. Patients with more severe psychotic symptoms may become agitated. As the disorder progresses, patients will become more demented and, ultimately, stuporous.

B_{12} deficiency leading to pernicious anemia causes psychotic symptoms in approximately 16% of patients. It is also reported that 30–80% of patients with pernicious anemia may have an organic brain syndrome including paranoia, violence, and depression. These symptoms all respond to correction of vitamin deficiencies. Psychiatric symptoms can be seen without neurological deficits or anemia.

VII. ELECTROLYTE IMBALANCE AND HYPOXIA

Electrolyte imbalance can precipitate delirium and confusional states. Patients may become acutely agitated and aggressive as a result of various electrolyte abnormalities. Sodium and calcium are probably the most common abnormalities to cause agitation since either state can result in disturbance of cell membrane equilibrium in the central nervous system.

Both acute cerebral hypoxia, a deficiency of oxygen in tissue, and hypoxemia, a deficiency of oxygen in the blood (usually resulting from cardiac or pulmonary compromise), may also cause delirium. As with any delirious state, these patients may become combative. Patients with chronic cerebral anoxia may compensate for the condition at baseline but can become quite confused and aggressive if the anoxia is worsened by infection or other systemic illness. Numerous other medical conditions, such as hypertensive encephalopathy, hypoglycemia, or intracranial bleeding, can also cause delirium, occasionally with concomitant aggression.

VIII. TOXINS

Toxins will generally cause stupor and lethargy with acute intoxication. However, numerous toxins have

Table II
Nutritional Deficiencies Associated with Aggressive Behavior

Deficiency	Comments
Pyridoxine (B_6)	Due to homocysteinuria; causes psychosis that may lead to violence; also seen in patients treated for tuberculosis with isoniazid who are not given pyridoxine supplementation
Folic acid	Due to homocystinuria and hypomethioninemia; may cause schizophrenia-like syndrome with agitation
Cobalamin (B_{12})	Irritability, emotional lability, and suspiciousness may lead to violence; neuropsychiatric symptoms occasionally present without peripheral neurological signs
Niacin	Deficiency can cause pellegra, seen commonly in alcoholics; depression, psychosis (hallucinations and thought disorder), more severe psychotic symptoms may cause agitation

Table III
Toxin Exposures That May Cause Violent Behavior

Toxin	Typical source
Mercury	Felt manufacture (past), hat makers
Organophosphates	Pesticides
Lead	Adults exposed occupationally (e.g., construction workers), children by ingesting paint
Manganese	Usual exposure is in miners, processors of manganese; parkinsonsonian symptoms with agitation
Arsenic	Chronic exposure may cause encephalopathy, confusion, agitation

been reported to cause irritability and even aggression in extreme cases, especially when exposure is gradual (Table III). As with general medical illnesses discussed previously, encephalopathy and psychosis are the main reasons for aggression. Chronic mercury exposure produces the most well-known neuropsychiatric side effects, leading to the expression "mad as a hatter" since it was formerly used in the manufacture of felt hats. Afflicted persons have mood lability and are shy, with decreased sleep and memory problems. Actual violence has not been reported. Organophosphates act by binding to acetylcholinesterase and inactivating the enzyme. The immediate effect is accumulation of large amounts of acetylcholine in the synapse. Organophosphate toxicity, usually following pesticide use, commonly presents with neuropsychiatric symptoms. Aggression has been described in rare cases, including one incidence of a man and his pet cat who both became quite violent after organophosphate exposure in their home. The aggression resolved after the pesticides were discontinued.

Lead has been reported to cause mood disorders and irritability, among other neuropsychiatric symptoms, but aggression per se has not been reported. Adults are usually exposed to lead through occupational contact, whereas children commonly ingest lead paint. At highly toxic levels, lead intoxication causes encephalopathy. Encephalopathy is seen most commonly in children since they absorb lead more readily. Encephalopathy can progress to psychosis. Chronic arsenic poisoning can also cause encephalopathy and, in some case, psychotic symptoms. "Manganese madness" occurs most often in miners and processors of manganese. Besides parkinsonian symptoms, patients may have behavioral disturbances, including psychosis.

IX. TREATMENT

In establishing a treatment plan for patients with agitation or aggression, the overarching principle is that diagnosis comes before treatment. It is essential to determine the mental status of the patient before the agitated or aggressive event, the nature of the precipitant, the environment in which the behavior occurs, the ways in which the event is mitigated, and the primary and secondary gains related to the problem behaviors. We suggest utilizing the *Consensus Guidelines for the Treatment of Agitation in the Elderly with Dementia* as a framework for the assessment and management of agitation and aggression. After assessment of possible etiologies of these behaviors, treatment is focused on the occurrence of comorbid neuropsychiatric conditions (depression, psychosis, insomnia, anxiety, and delirium) (Fig. 2). Evaluation and documentation with objective measures such as the Overt Aggression Scale and the Overt Agitation and Severity Scale are often helpful. Other factors that influence the choice of medications include whether the treatment is in the acute (hours to days) or chronic (weeks to months) phase and the severity of the behavior (mild to severe). Table IV presents an overview of medications used to treat aggressive behaviors.

Although no medication has been approved by the Food and Drug Administration specifically for the treatment of aggression, medications are widely used (and commonly misused) in the management of patients with aggression. We divide the use of pharmacological interventions for aggression to two major categories: (i) the use of the sedating effects of medications, as required in acute or severe situations so that the patient does not harm him- or herself or others, and (ii) the use of

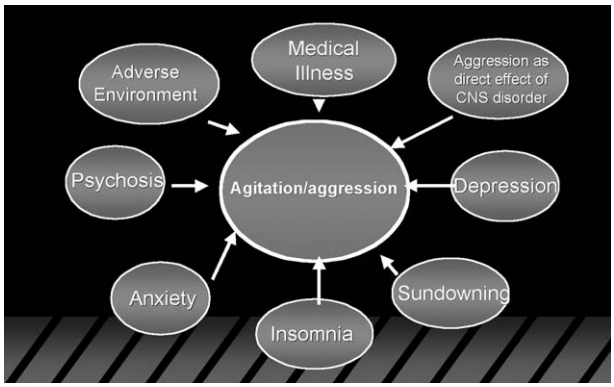


Figure 2 Neuropsychiatric factors associated with agitation and aggression.

nonsedating medications for the treatment of chronic aggression.

A. Acute Aggression and Agitation

Medications that are sedating may be indicated for the treatment of agitation and for treating severe episodes of aggressive behavior. Because these drugs are not specific in their ability to inhibit aggressive behaviors, however, there may be detrimental effects on arousal and cognition. They may also cause seriously disabling side effects including tardive dyskinesia. Unless aggressive behavior is clearly related to psychosis, we prefer to limit the use of both antipsychotic agents and benzodiazepines for ''sedating'' aggression to a maximum time period of 4 weeks. Beyond this time, the clinician must consider whether or not the aggression is chronic and alter the treatment plan accordingly to use medications recommended in the following sections for chronic treatment.

1. Antipsychotic Drugs

Antipsychotics are the most commonly used medications in the treatment of aggression. Although these agents are appropriate and effective when aggression is a derivative of active psychosis, the use of neuroleptic agents to treat chronic aggression is often ineffective and entails significant risks for the patient. Often, patients develop tolerance to the sedative effects of the neuroleptics and therefore require escalating doses. As a result, especially with the older neuroleptics, extrapyramidal and anticholinergic-related side effects occur. Paradoxically (and frequently), because of the development of akathisia, the patient may become more agitated and restless, especially when a high-potency antipsychotic such as haloperidol is administered.

Evidence from studies of injury to motor neurons in animals indicates that haloperidol decreases recovery. It is possible that the effect on decreasing dopamine and inhibiting neuronal function, which may be the mechanism of action to treat aggression, may have other detrimental effects on recovery. Whether this finding is generalizable to humans remains unclear. However, these studies raise important potential risk/benefit issues that must be considered in the management of aggression.

In patients with acute aggression, we recommend starting a high-potency neuroleptic such as risperidone

Table IV
Medications to Treat Aggressive Behavior

Class (Examples of Agents)	Indications	Side effects/Considerations
Antipsychotics (atypicals include: risperidone, olanzapine, quetiapine)	Psychosis, need for sedation, hallucinations	Multiple side effects (decrease in extrapyramidal side effects with atypicals)
Benzodiazepines (lorazepam)	Acute management	Extremely sedating; may be cause of violence by paradoxical disinhibition and rage
Anticonvulsants (carbamazepine, valproate)	Seizure disorder, bipolar disorder	Bone marrow suppression (CBZ), hepatotoxicity (VPA) May be very sedating
Buspirone	Chronic anxiety/depression	Long onset to therapeutic action
Antidepressants (fluoxetine, paroxetine, citalopram, sertraline)	Depression, anxiety, or labile mood	Selective serotonin reuptake inhibitors may be most useful
Beta blockers	Chronic or recurrent aggression	4–6 weeks to onset of efficacy; high doses needed may cause blood pressure effects in some patients

at low dosages of 0.5 mg orally with repeated administration every hour until control of aggression is achieved. If intramuscular medication is required, haloperidol should be utilized. If, after several administrations of risperidone the aggressive behavior does not improve, the hourly dose may be increased until the patient is so sedated that he or she no longer exhibits agitation or violence. Once the patient is not aggressive for 48 hr, the daily dosage should be decreased gradually (i.e., by 25%/day) to ascertain whether aggressive behavior reemerges. In this case, consideration should then be made as to whether it is best to increase the dose of risperidone and/or to initiate treatment with a more specific antiaggressive drug.

2. Sedatives and Hypnotics

There is an inconsistent literature on the effects of the benzodiazepines in the treatment of aggression. The sedative properties of benzodiazepines are especially helpful in the management of acute agitation and aggression. Most likely, this is due to the effect of benzodiazepines on increasing the inhibitory neurotransmitter GABA. Paradoxically, several studies report increased hostility, aggression, and the induction of rage in patients treated with benzodiazepines, possibly due to disinhibitory effects. However, this phenomenon is rare. Preexisting memory dysfunction can be exacerbated by the use of benzodiazepines.

For treatment of acute aggression, 1 or 2 mg of lorazepam may be administered every hour by either oral or intramuscular route until sedation is achieved. Intravenous lorazepam is also effective, but caution must be taken with intravenous administration, and it should be injected in doses less than 1 cc (1 mg) per minute to avoid laryngospasm. Intravenous lorazepam does not have faster onset of action than intramuscular administration. As with neuroleptics, gradual tapering of lorazepam may be attempted when the patient has been in control for 48 hr. If aggressive behavior reoccurs, medications for the treatment of chronic aggression may be initiated. Other sedating medications, such as paraldehyde, chloral hydrate, or diphenhydramine, may be preferable to sedative antipsychotic agents.

B. Chronic Aggression

If a patient continues to exhibit periods of agitation or aggression beyond several weeks, the use of specific antiaggressive medications should be initiated for prevention.

1. Antipsychotic Medications

If, after thorough clinical evaluation, it is determined that the aggressive episodes result from psychosis, such as paranoid delusions or command hallucinations, then antipsychotic medications will be the treatment of choice. Risperidone has been used to treat agitation in elderly patients with dementia with good results. Olanzapine appears to be more sedating, and quitiepine may have fewer extrapyramidal side effects (EPS) than does risperidone. Clozapine may have greater antiaggressive effects than other antipsychotic medications. However, the decrease in seizure threshold must be carefully assessed, along with the need for monitoring blood counts.

2. Antianxiety Medications

Serotonin appears to be a key neurotransmitter in the modulation of aggressive behavior, as discussed previously. In preliminary reports, buspirone, a 5-HT 1A agonist, has been reported to be effective in the management of aggression and agitation for patients with head injury, dementia, and developmental disabilities and autism. Rare cases of paradoxical increase in aggression have been reported with buspirone use. We usually initiate buspirone at 7.5 mg bid for 1 week and then increase the dosage to 15 mg bid. Dosages of 45–60 mg/day may be required before there is improvement in aggressive behavior, although we have noted dramatic improvement within 1 week.

Clonazepam may be effective in the long-term management of aggression, although controlled, double-blind studies have not been conducted. We use clonazepam when pronounced aggression and anxiety occur together or when aggression occurs in association with neurologically induced tics and similarly disinhibited motor behaviors. Doses should be initiated at 0.5 mg twice daily bid and may be increased to as high as 2–4 mg bid, as tolerated. Sedation and ataxia are frequent side effects.

3. Anticonvulsive Medications

Carbamazepine has been proved to be effective for the treatment of bipolar disorders and has also been advocated for the control of aggression in both epileptic and nonepileptic populations. Several open studies have indicated that carbamazepine may

be effective in decreasing aggressive behavior associated with TBI, dementia, developmental disabilities, schizophrenia, and patients with a variety of other organic brain disorders.

In our experience and that of others, valproic acid may also be helpful to some patients with organically induced aggression. For patients with aggression and epilepsy whose seizures are being treated with anticonvulsant drugs such as phenytoin and phenobarbital, switching to carbamazepine or to valproic acid may treat both conditions. Gabapentin has been used effectively for the treatment of agitation in patients with dementia.

4. Antimanic Medications

Although lithium is known to be effective in controlling aggression related to mania, many studies suggest that it may also have a role in the treatment of aggression in selected, nonbipolar patient populations. Included are patients with TBI as well as patients with mental retardation who exhibit self-injurious or aggressive behavior, children and adolescents with behavioral disorders, and patients with other organic brain syndromes. Reports also suggest that lithium is useful in reduction of aggression in prisoners. Because of lithium's potential to cause neurotoxicity and dehydration, and its relative lack of efficacy in many patients, we limit the use of lithium in those patients whose aggression is related to manic effects or recurrent irritability related to cyclic mood disorders.

5. Antidepressants

The antidepressants that have been reported to control aggressive behavior are those that act preferentially (amitriptyline) or specifically (trazodone and the SSRIs) on serotonin. Trazodone has also been reported to be effective in the treatment of aggression. Sertraline and fluoxetine have been used in numerous cases of aggression due to neurological and psychiatric disorders. We have used selective SSRIs with considerable success in aggressive patients with brain lesions. The dosages used are similar to those for the treatment of mood lability and depression.

We emphasize that for many patients it may be necessary to administer these medications at standard antidepressant dosages to obtain full therapeutic effects, although response may occur for others within days of initiating treatment at relatively low doses. In select patients with obsessions and compulsions due to organic disease, aggression is seen when the patients are prevented from acting on these symptoms. In these cases, higher doses of SSRIs, such as those used in the treatment of obsessive–compulsive disorder, may be more effective (e.g., 60–80 mg of fluoxetine).

6. Antihypertensive Medications: Beta Blockers

Since the first report of the use of β-adrenergic receptor blockers in the treatment of acute aggression in 1977, more than 25 articles have appeared in the neurological and psychiatric literature reporting experience in using beta blockers for more than over 200 patients with aggression. Most of these patients had been unsuccessfully treated with antipsychotics, minor tranquilizers, lithium, and/or anticonvulsants before treatment with beta blockers. The beta blockers that have been investigated in controlled prospective studies include propranolol (a lipid-soluble, nonselective receptor antagonist), nadolol (a water-soluble, nonselective receptor antagonist), and pindolol (a lipid-soluble, nonselective beta receptor antagonist with partial sympathomimetic activity). The effectiveness of propranolol in reducing agitation has been demonstrated during initial hospitalization after TBI. When a patient requires the use of a once-a-day medication because of compliance difficulties, long-acting propranolol (i.e., Inderal LA) or nadolol (Corgard) can be used. When patients develop bradycardia that prevents prescribing therapeutic dosages of propranolol, pindolol (Visken) can be substituted using one-tenth the dosage of propranolol. Pindolol's intrinsic sympathomimetic activity stimulates the beta receptor and restricts the development of bradycardia.

The major side effects of beta blockers are a lowering of blood pressure and pulse rate. Because peripheral beta receptors are fully blocked in doses of 300–400 mg/day, further decreases in these vital signs usually do not occur, even when doses are increased to much higher levels. Despite reports of depression with the use of beta blockers, controlled trials and our experience indicate that it is a rare occurrence. Because the use of propranolol is associated with significant increases in plasma levels of thioridazine, which has an absolute dosage ceiling of 800 mg/day, the combination of these two medications should be avoided.

X. SUMMARY OF PHARMACOTHERAPY AGGRESSION

Table V summarizes our recommendations for the use of various classes of medication in the treatment of

Table V
Pharmacotherapy of Agitation

Acute agitation/severe aggression	
High-potency antipsychotic drugs (haloperidol, risperidone)	
Benzodiazepines (lorazepam)	
Chronic agitation	
Drug	**Primary indication**
Atypical antipsychotics (risperidone, olanzapine, quetiapine, clozapine)	Psychosis
VPA, CBZ,? gabapentin	Seizure disorder, severe aggression
Serotonergic antidepressants (SSRI, trazodone)	Depression, mood lability
Buspirone	Anxiety
Beta blockers	Aggression without concomitant neuropsychiatric sequelae

acute and chronic aggressive disorders. If partial response without undue side effects is seen with adequate dosing of one agent, augmentation with a second agent should be considered. Control of aggression with the lowest dose of the medication (or medications) with the fewest potential unwanted side effects should be the goal of pharmacotherapy.

XI. CONCLUSIONS

Aggressive and violent behavior may result from many disorders of the central nervous system. These problems are common and constitute a major source of morbidity and mortality. Disruptive behavior is often the largest barrier to reintegration into the community for those with severe neurological illnesses. Many medications are effective in the treatment of aggression, although accompanying side effects must be weighed carefully for each particular patient (Table IV).

Lesions that involve the limbic lobes may be more frequently associated with aggression, but violent behavior is also seen with diffuse cortical disease. Patients with central nervous system disease may be violent in an unpredictable manner, but aggression often occurs when others attempt to restrain or assist a patient. This resistive aggression is reported frequently in studies of demented patients, those with intellectual deficiency, and those with epilepsy who are postictal. Aggression is also seen as a result of severe medical illness, especially in individuals with chronic illness. Encephalopathy as a result of a medical condition can lead to agitation and aggression. Systemic medical illness can also cause psychosis, which may lead to violent behavior in response to paranoid ideation.

Although aggression may appear to be unpredictable in these populations, patterns are evident in many forms of disruptive behavior. If a sudden change in behavior occurs, neurological or medical illness should be ruled out as a cause of the aggression. This may require a fairly extensive workup and consideration of uncommon conditions. Description and evaluation of aggression in the aggressive behavior is the first step toward treatment and prevention.

In treating aggression, the clinician, when possible, should diagnose and treat underlying disorders and use antiaggressive agents specific for those disorders. When there is partial response after a therapeutic trial with a specific medication, adjunctive treatment with a medication with a different mechanism of action should be instituted. In all cases, the clinician should attempt to treat aggression with the lowest effective dose of medication and to avoid using treatments with severe potential side effects.

See Also the Following Articles

AGGRESSION • ANGER • BEHAVIORAL NEUROGENETICS • EVOLUTION OF THE BRAIN • NEUROTRANSMITTERS • SEXUAL DIFFERENTIATION, HORMONES AND • SUICIDE

Suggested Reading

Cohen-Mansfield, J., Marx, M. S., and Rosenthal, A. S. (1990). Dementia and agitation in nursing home residents: How are they related? *Psychol. Aging* **5**(1), 3–8.

Delgado-Escueta, A. V., Mattson, R. H., King, L., Goldensohn, E. S., Spiegel, H., Madsen, J., Crandall, P., Dreifuss, F., and Porter, R. J. (1981). The nature of aggression during epileptic seizures [special report]. *N. Engl. J. Med.* **305**(12), 711–716.

Devanand, D. P., Jacobs, D. M., Tang, M. X., Del Castillo-Castaneda, C., Sano, M., Marder, K., Bell, K., Bylsma, F. W., Brandt, J., Albert, M., and Stern, Y. (1997). The course of psychopathologic features in mild to moderate Alzheimer disease. *Arch. Gen. Psychiatr.* **54**(3), 257–263.

Klüver, H., and Bucy, P. C. (1939). Preliminary analysis of functions of the temporal lobe in monkeys. *Arch. Neurol. Psychiatr.* **42,** 979–1000.

Linaker, O. M. (1994). Assaultiveness among institutionalised adults with mental retardation. *Br. J. Psychiatr.* **164**(1), 62–68.

Sigafoos, J., Elkins, J., Kerr, M., and Attwood, T. (1994). A survey of aggressive behaviour among a population of persons with intellectual disability in Queensland. *J. Intellect. Disabil. Res.* **38**(Pt. 4), 369–381.

Silver, J. M., and Yudofsky, S. C. (1994). Aggressive disorders. In *Neuropsychiatry of Traumatic Brain Injury* (J. M. Silver, S. C. Yudofsky, and R. E. Hales, Eds.). American Psychiatric Press, Washington, DC.

Stevens, J. R. (1988). Psychiatric aspects of epilepsy. *J. Clin. Psychiatr.* **49**(Suppl.), 49–57.

Volavka, J. (1995). *Neurobiology of Violence*. American Psychiatric Association Press, Washington, DC.

Yudofsky, S. C., Sliver, J. M., Jackson, W., *et al*. (1986). The Overt Aggression Scale for the objective rating of verbal and physical aggression. *Am. J. Psychiatr.* **143,** 35–39.

Yudofsky, S. C., Kopecky, H. J., Kunik, M., Silver, J. M., and Endicott, J. (1997). The Overt Agitation Severity Scale for the objective rating of agitation. *J. Neuropsychiatr. Clin. Neurosci.* **9**(4), 541–548.

Vision: Brain Mechanisms

ROBERT SHAPLEY and NAVA RUBIN
New York University

GLOSSARY

magnetic resonance imaging Imaging of the internal structure of body tissue by the observation of magnetic resonances induced in them.

occipital cortex Cerebral cortex belonging to the occiput or back part of the head.

retinotopic map A map that preserves the spatial relations of the sensory receptors of the retina.

segmentation The process of division into segments.

stereopsis The ability to perceive depth and relief by stereoscopic vision.

temporal cortex The lowest lobe of the brain lying below the Sylvian fissure.

Vision appears to be a simple matter, which does not require any mental effort: We open our eyes and see. However, the subjective ease of seeing masks a tremendous amount of brain processing required to allow us to perceive the world through the visual sense. The most direct evidence for this is the large portions of cortex that are devoted to visual processing. The entire occipital lobe, as well as parts of the temporal and parietal lobe, is concerned with interpreting the image impinging on the retina. The operations done by the numerous visual cortical areas are understood only partially. Part of the difficulty stems from the lack of good theoretical understanding of visual processing. For example, research on artificial vision has not been able to produce "seeing machines" that would perform visual tasks that humans perform effortlessly, such as navigating through a crowded room or recognizing a familiar face. Better theoretical understanding would have allowed to map the various brain processes onto known computational stages of visual processing. Instead, the study of the brain mechanisms of vision and research on the computational aspects of vision have progressed in parallel, each field drawing inspiration from findings in the other. In this article, we survey some of the major findings on the function of the various areas of the visual cortex.

I. INTRODUCTION

The visual cortex takes signals about the retinal images that are relayed to it through the lateral geniculate nuclei (LGN) and seeks to recognize objects and navigate in the world based on vision. Visual signals are processed in V1 (also known as the striate cortex) and V2, the primary and secondary visual cortical areas. From V1 and V2, signals are sent to multiple visual areas anterior and lateral to these early visual areas in the brain. Some of these extrastriate visual areas seem to be very specialized for certain aspects of the visual image, such as motion or color. How these different areas contribute to visual perception has become much clearer in recent years.

Knowledge about the visual system's functional anatomy has expanded tremendously in the past decade through the use of brain imaging techniques.

Optical imaging of the upper layers of the cerebral cortex in nonhuman primates, using activity-dependent changes in reflectance, can be used to study structure on the scale of several square millimeters. Magnetic resonance imaging (MRI) both in nonhuman primates and in human subjects allows scientists to study the structure of the visual system noninvasively. Functional MRI (fMRI) enables researchers to study changes in the brain's blood oxygenation that are caused by variations in neuronal activity, thus enabling noninvasive functional mapping of human brain activity. Most of the results we cite about human brain activity correlated with perception are derived from fMRI experiments. fMRI has lower spatial resolution than optical imaging, but it allows the experimenter to study brain volumes from 100 to 1000 ml and reflects activity not only from the cortex's superficial layers but also through the depth of the cortex.

It appears that the brain must find meaningful patterns in order to compute visual attributes correctly as well as to be able to place these attributes in and around perceived objects in a scene organization that is consistent with the visual image. This view is based on experimental observations by, among others, Hans Wallach that visual attributes of perceived objects, such as color, shape, and motion, are linked, and they all need to be computed by the brain. Thus, we are compelled by the evidence from human perception, from fMRI, and from physiological experiments on animals to conceive of perception as an active process in which visual image information is combined with memory to compute "why things look as they do." This is most easily understood when visual images are ambiguous—for example, when selected visual images produce figure/ground reversals or binocular rivalry. Then it becomes clear that there are multiple states of brain activity corresponding to multiple "interpretations" that are consistent with the visual image. What one sees in this situation changes with time and the state of the brain and must involve time variations in the computations that the visual system is doing based on the visual image.

The first neural computation we consider is the mapping of the visual world onto the cortex in V1 and adjacent visual cortex.

II. V1–V*n* AND RETINOTOPY

In order to understand appearance, we must account for the location of perceived objects in the world. Perceived location depends on the mapping from location in the world to location on the cortex. A schematic diagram of human visual cortex is depicted in Fig. 1. This diagram emphasizes the visual field

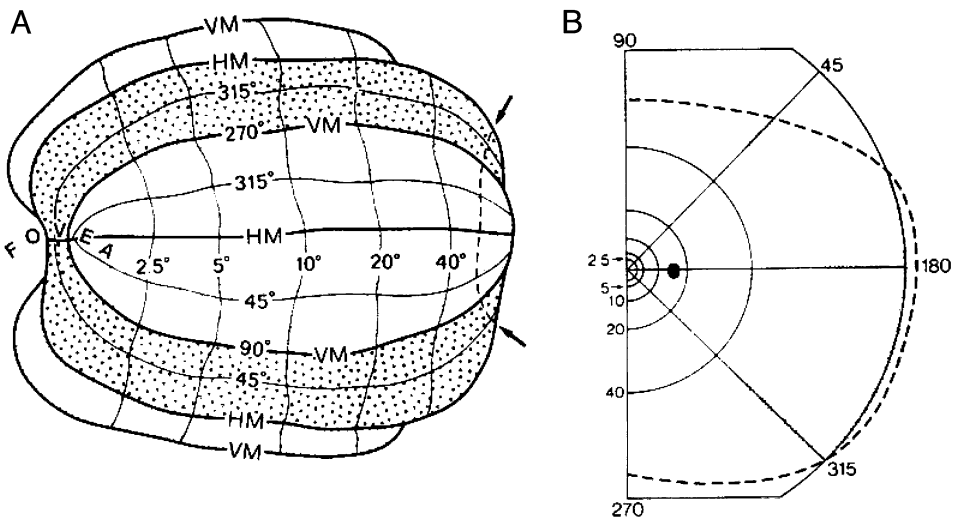


Figure 1 (A) Retinotopic mapping in human V1–V3. Mapping of the visual field on an artificially flattened human visual cortex. V1 is the central clear zone, V2 is marked by the small dot stippling, and V3 is the outermost zone. (B) Visual field coordinates corresponding to those drawn in A. Note that V2 and V3 are split along the representation of the horizontal meridian into separate upper field and lower field halves (reproduced with permission from Horton and Hoyt, 1991). Quadrantic visual field defects. A Hallmark of lesions in extrastriate (V2/V3) cortex. Brain **114,** 1703–1718.

mapping onto different brain regions, V1–V3. Human primary visual cortex V1 is located deep in the calcarine fissure, and it extends slightly onto the cortex around the posterior pole. Most of the visual cortex located on the posterior surface of the human cerebral cortex is in fact extrastriate cortex; that is, V2, V3, V4, and up to V*n*. V1 is continuous above and below the calcarine fissure and forms a continuous boundary with V2, which surrounds it completely. V2 is composed of two disjoint cortical regions, upper and lower V2, as is the case for V3 and V4. From lesion studies we know that the upper half of V1 receives LGN input that is looking only at the lower visual field, whereas lower V1 is looking at the upper visual field. This is true also of V2, V3, etc. It is well-known that the early, retinotopic areas V1–V3 represent the visual field contralateral to the cerebral hemisphere in which they are located. Thus, a lesion in the upper region of V2 will cause a visual field defect only in the lower contralateral quadrant of the visual field.

The map of the visual field in V1 was demonstrated elegantly in the monkey visual cortex by Roger Tootell and colleagues. They used the 2-deoxyglucose activity-labeling technique as illustrated in Fig. 2. The stimulus used was a "ring and ray" pattern presented for approximately 1 hr while the radioactively labeled 2-deoxyglucose circulated through the bloodstream of the monkey. Active cells picked up the label and appear dark in the autoradiograph in Fig. 2. The center of the ring and ray pattern was placed where the foveae of the eyes were looking, and the pattern in the contralateral hemifield (Fig. 2) activated the neurons in the hemisphere that is imaged in the figure. The cortex was flattened physically to enable the flat two-dimensional (2D) image of the cortical map to be made. The mapping of the ring and ray pattern is smooth and relatively precise onto V1 cortex. The circular rings were spaced at distances that grew geometrically larger, and their map onto V1 is approximately onto lines that are approximately equidistant from each other. This means the mapping is approximately logarithmic, as has often been noted. The stimulus rays also map into approximately straight lines that are separated from each other by a fixed distance. The vertical meridian is mapped into the V1/V2 border, whereas the horizontal meridian divides V1 in half, with the upper field mapped to lower V1 and vice versa. Figure 2 is a succinct summary of the main facts about retinotopic mapping in primate V1 cortex.

Retinotopic mapping in the human V1 cortex, and in extrastriate visual areas as well, has been demonstrated with fMRI. The technique developed by Engel and colleagues is shown in Fig. 3. The basic idea is to use a periodic stimulus, such as an expanding ring or a rotating wedge. Then neurons that are functionally connected to the retinal region excited by the stimulus will be activated only during one phase of the stimulus cycle. Location will be encoded in the phase of the fMRI response with respect to the stimulus cycle. This technique is therefore called phase mapping. The phase of fMRI response with respect to the expanding ring's cycle encodes radial distance from the point of expansion (usually the fovea). The phase of fMRI response with respect to the rotating wedge encodes the azimuth of the response location in the frontoparallel plane. From Fig. 2, one might expect that the active regions of V1 excited by a given azimuth should lie along straight lines in a flattened map of the cortex since this is what was observed in the 2-deoxyglucose maps. Also, one should expect that regions of a given retinal eccentricity in flattened-cortex phase maps should lie along straight lines that are approximately perpendicular to the cortical projection of the vertical meridian, as demonstrated by the monkey 2-deoxyglucose data.

Data on the retinotopic map in human visual cortex are given in Fig. 4. Here, the representation of the visual cortex as a flat sheet was not achieved by physical flattening as in the monkey experiments since these were *in vivo*, noninvasive measurements and the cortex was not available for physical flattening. Rather, the cortex was virtually flattened by computer image processing of the 3D image of the cortex obtained from MRI measurements, using an algorithm that preserved local distances. Separate eccentricity and azimuth maps were measured to create the mapping in the figure. As shown, only azimuth is coded by color. The V1 map is consistent with the monkey V1 map in terms of both eccentricity and azimuth. The map indicates that human V1 also represents only the contralateral hemifield, consistent with the effects of lesions on human visual cortex.

The results summarized in Fig. 4 also reveal important facts about the mapping of visual space in extrastriate cortex for example, V2 cortex adjacent to V1. Also, there is a retinotopic map in V2. The V1 and V2 maps are separated by the representation of the vertical meridian in each cortical area, so the vertical meridian projection serves as a marker for the V1/V2 boundary. This also means that V2, like V1, is only receiving visual input from the contralateral visual hemifield. Furthermore, the boundary between V2 and V3, a distinct visual cortex region more lateral than V2,

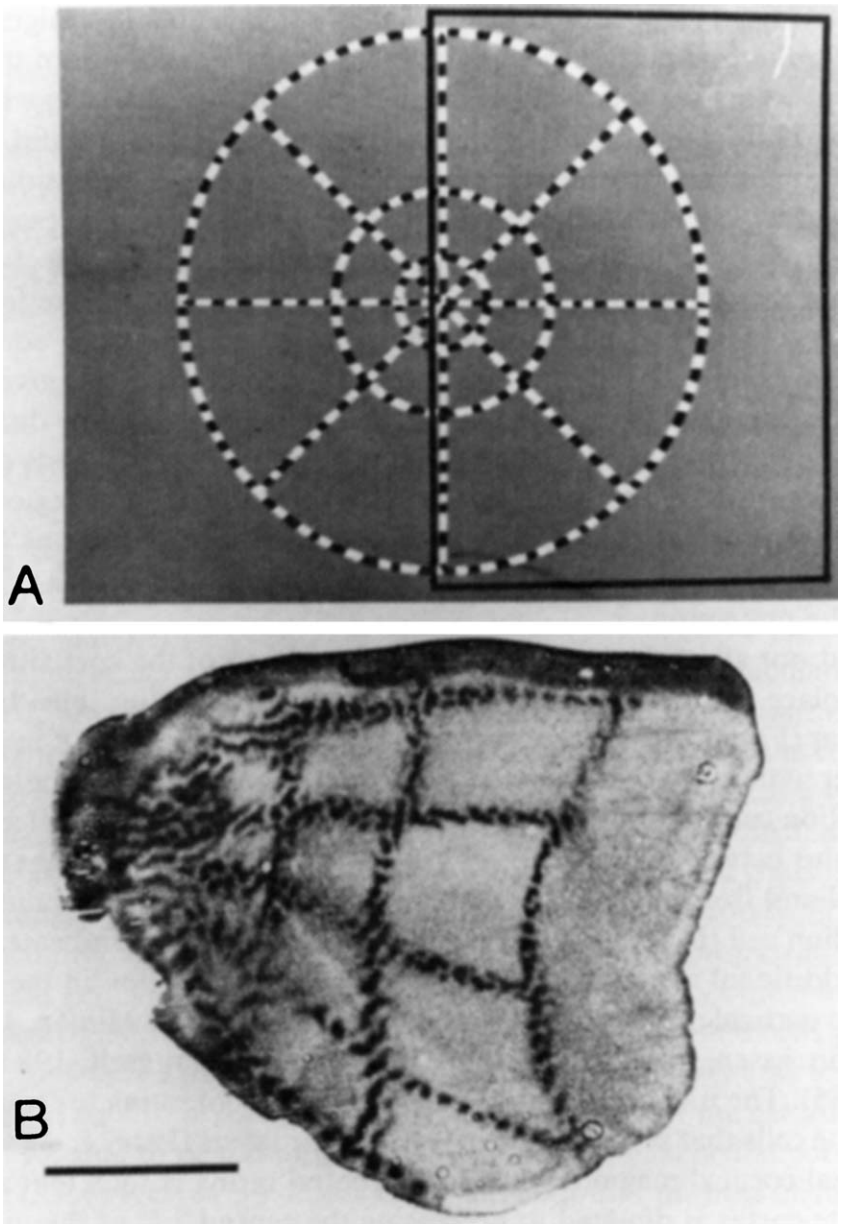


Figure 2 Retinotopic mapping of macaque monkey V1 cortex measured with radioactively labeled 2-deoxyglucose. (Top) The ring and ray pattern was presented to an anesthetized monkey for a long duration. The pattern that was exposed to the left visual cortex is shown. (Bottom) Autoradiograph of V1 showing the regions of V1 that took up the labeled 2-deoxyglucose (dark regions). This shows the mapping of the rings and rays into straight lines on the cortical surface (reproduced with permission from Tootell *et al.*, 1982). Deoxyglucose analysis of retinotopic organization in primate striate cortex. *Science* **218,** 902–904.

is the projection of the horizontal meridian. It is important that the projection to V2 (and also V3) of visual image regions just above and just below the horizontal meridian projects to cortical subregions that are widely separated. This means that local circuits that may compute spatial linking between local stimuli encounter a wide chasm in extrastriate cortex at the horizontal meridian; therefore, it is likely that visual signals from upper and lower visual fields are processed separately in extrastriate cortex. This does not apply to V1, in which neurons are mapped smoothly across the horizontal meridian. This can be

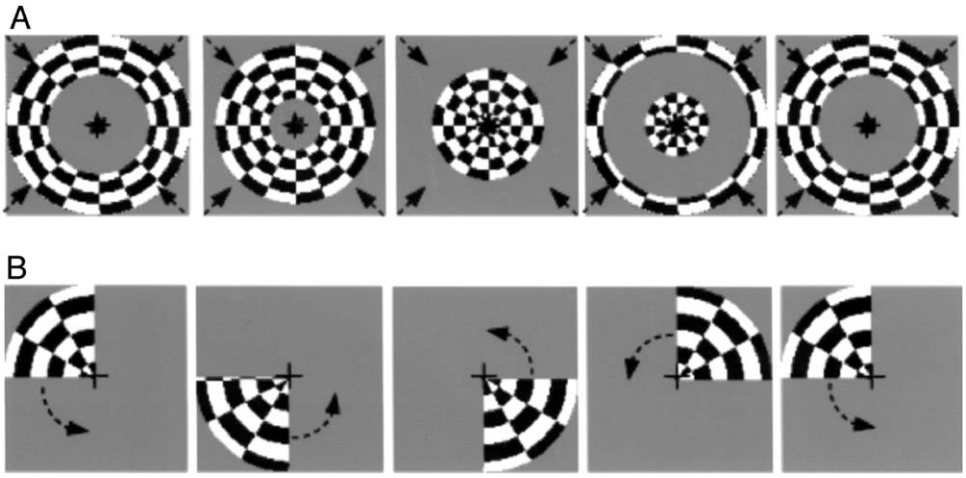


Figure 3 Diagram of the stimuli used in the phase mapping technique for fMRI. Radial checkerboards are the basic stimuli. In A, used for eccentricity mapping, rings of the underlying checkerboards are propagating inwards to a focus of contraction. In B, a rotating wedge of checkerboard is shown, and the wedge's azimuth varies with time [reproduced with permission from Wandell (1999). Computational neuroimaging of human visual cortex. *Annu. Rev. Neurosci.* **22,** 145–173.]

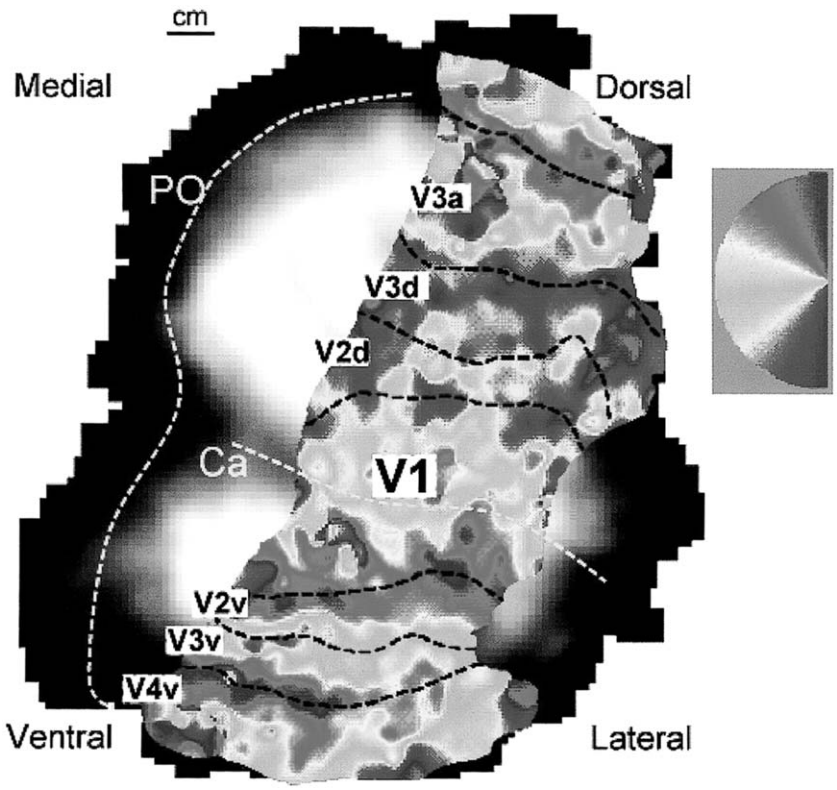


Figure 4 Retinotopic mapping in the human cortex measured with fMRI. Phase mapping as in Fig. 3 was used. In this example, only azimuth is mapped, as indicated in the inset, which gives the key to angle. The map is onto a mathematically flattened cortex and shows the transitions between retinotopic areas along the meridians [reproduced with permission from Wandell (1999). Computational neuroimaging of human visual cortex. *Annu. Rev. Neurosci.* **22,** 145–173.] (See color insert in Volume 1).

seen best in Fig. 1. The continuous mapping in V1 and the discontinuous mapping in V2–V4 are likely to have functional consequences for vision.

In the human visual cortex, mapping experiments such as those in Figs. 3 and 4 indicate that the right and left hemifields are kept distinct through several extrastriate regions, including V1–V4. Signals from the right and left visual fields are not combined until one reaches the lateral occipital cortex (LOC), an area discussed later. Therefore, there are two "seams" in the mapping of the visual world onto early visual cortical processing areas from V1–V4 and possibly beyond. One seam is the vertical meridian that separates the right and left visual fields. The second seam is the horizontal meridian that separates the upper from the lower visual fields. Our visual perception appears to be seamless, subjectively. The unified worldview we have must be the result of a combination of visual signals across these seams in the mapping in a manner that remains to be discovered. The interhemispheric mapping of the visual world in LOC and anterior visual areas is likely to be significant in producing the seamless appearance of the visual world.

III. SEGMENTATION

There is a great transformation that takes place in visual perception between the analog representation of the visual image in V1 and the symbolic representation of surfaces and objects as they appear to us. In V1, as in the retina, there seems to be an analog map of the brightness and colors in the image. However, when we perceive a meaningful natural image, we see a countable number of surfaces and discrete objects that are segregated from the background and from other objects. There are probably many stages of this transformation, but one stage of this process is known to be of great importance: visual segmentation. Segmentation is a process that involves the grouping of object parts together in a separated figure that is distinct from what is around or near it. Often, this process must group together parts of an object in an image that are separated from each other by an occluding object in front of the object to be segmented. Segmentation is necessary for correct organization of a scene because it allows the organizing process to know which are the surfaces that must be ordered in depth. Also, the action of segmentation could contribute to the scene organization computation because it needs to resolve occlusions that are a strong clue to depth order and therefore, scene organization.

As with many brain computations, we can understand segmentation better by observing its action when it deals with an exceptionally difficult task. Usually, segmentation is so smooth and efficient that we are unaware it is happening. However, for certain special visual images, the segmentation process becomes evident. This is the reason for the fascination with these special images, the so-called illusory contours (ICs). An example of such a visual image is shown in Fig. 5, which presents an image that is referred to as a Kanizsa triangle, named after the Italian Gestalt psychologist Gaetano Kanizsa. In Fig. 5, the perception of a bright white triangle is very strong, but if one scrutinizes the boundaries of the triangle it becomes evident that there is no difference in the amount of light coming to the eye from the regions inside and outside the perceived triangle. However, we see the inside as a bright surface segmented from its background by sharp contours along the boundary of the triangle. In this sense, the boundary between inside and outside the triangle is an illusory contour. This image is a classic example in favor of the basic concept of the Gestalt psychologists—that the brain is searching for meaningful patterns. In this case, the brain manufactures a perceptual triangle from fragmentary information because a meaningful pattern, an occluding triangle, is consistent with the available image information even

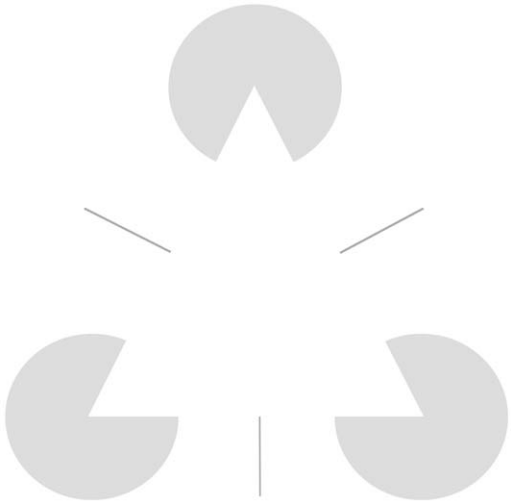


Figure 5 Kanizsa triangle. The occluding triangle that appears in front of the three circles and the three line segments is of the same physical brightness as the surroundings. However, it appears brighter and appears to be a solid surface in front because of perceptual processes.

though other perceptions are possible. It is reasonable to believe that the segmentation computations that the visual system performs on these exceptional Kanizsa images are the same as for more typical images.

From electrophysiological single-cell recordings in awake monkeys, von der Heydt, Peterhans, and colleagues found that Kanizsa-type images and other illusory contour images could excite spike activity in neurons in early visual cortex. An example of one of their experiments is shown in Fig. 6. They recorded from a single neuron in area V2 of the macaque visual cortex. The neuron's activity is depicted as a raster plot of spikes vs time during repeated presentations of the stimuli. The neuron responds with excitation to a luminance contour that crosses its receptive field. When an illusory contour (as perceived by us) crosses its receptive field, the cell produces a slightly delayed excitatory response resembling the response to a real contour. As a control that the response is not merely a weak response to the remote features of the IC stimulus, the investigators made a small image manipulation (closing the inducing boundary) and this eliminated the neuron's response. Von der Heydt and Peterhans also performed several quantitative studies on these IC-responsive V2 neurons, particularly the measurement of the orientation tuning for illusory and real contours on the same population of V2 neurons; they found that real and illusory contours produced similar orientation tuning in IC-responsive neurons in V2. Thus, these neurons seem to be a candidate neural substrate for illusory contour perception. There have also been reports of IC responses in neurons in V1. This is a controversial issue since von der Heydt, Peterhans, and colleagues maintained that they observed very few V1 neurons that produced IC responses. The discrepancy may occur in part perhaps because of the use of different stimuli and in part because of different views of what constitutes an illusory contour. For present purposes, it is enough to conclude that IC responses can be observed in

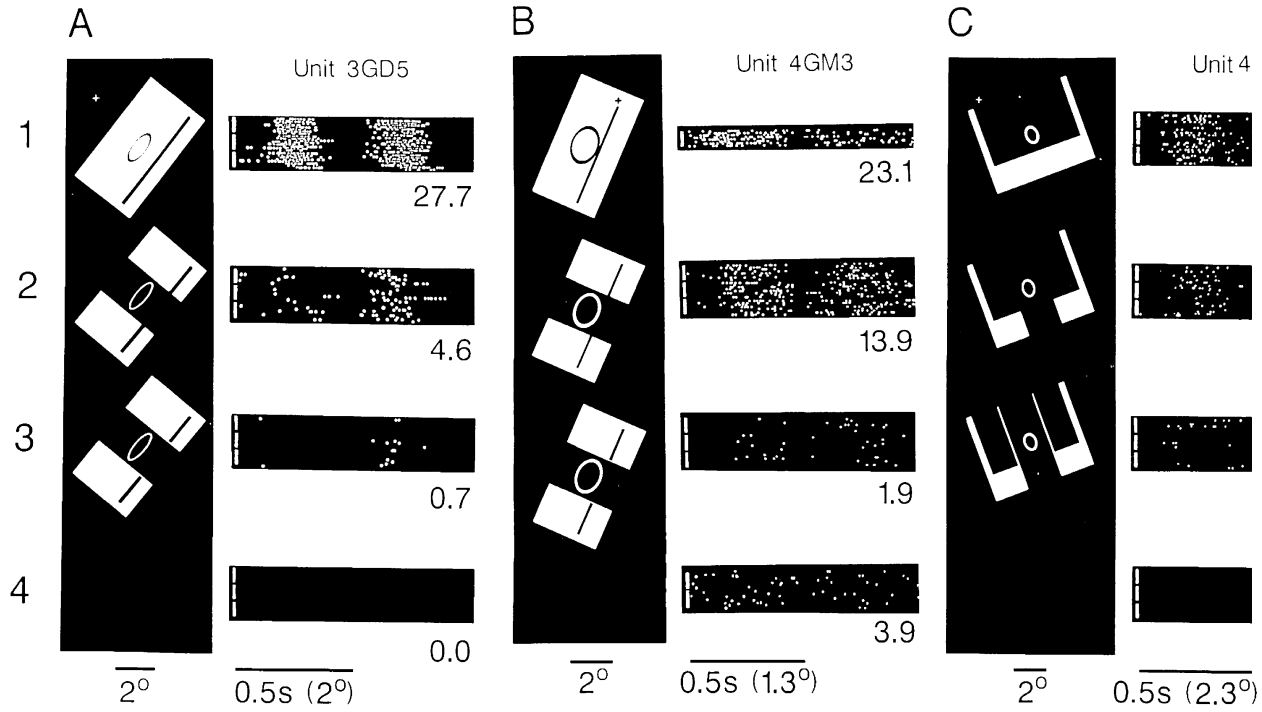


Figure 6 Responses of three macaque V2 neurons to ICs. Stimuli are shown in the insets next to raster plots of neuronal responses, in which each white dot is a nerve impulse. The raster shows multiple repeats of the same stimulus swept over the receptive field repeatedly. The crosses indicate the fixation point, and ellipses mark the classical receptive field of each neuron. In each case, row 1 is a conventional luminance stimulus, row 2 is the IC stimulus, and row 3 is a control stimulus, where a small image manipulation destroys the IC percept. Row 4 is a blank control to show the spontaneous firing rate [reproduced with permission from Peterhans and von der Heydt (1989). Mechanisms of contour perception in monkey visual cortex. II. Contours bridging gaps. *J. Neurosci.* **9,** 1749–1763.]

retinotopic areas in the monkey's brain—areas that are traditionally thought of as stimulus driven.

The connection of the monkey V2 neurons with IC perception in humans needs to be established more firmly. We do not know the nature or quality of IC perception in these animals because there are insufficient animal data from rigorous experiments to test for IC perception. Once we know whether monkeys can exhibit behavior that proves they see illusory contours in Kanizsa-type images, further experiments would be necessary to determine whether or not the V2 neurons have the same sort of parameter dependence on size, contrast, and retinal location as the behavior. How do the V2 neurons respond to an illusory contour that crosses the horizontal meridian? Humans respond as well to such illusory contours as they do to ICs that do not cross the horizontal or meridian. The reader can observe this for himself or herself by fixating the middle of the Kanizsa triangle in Fig. 5 and observing the robust lateral ICs that traverse the horizontal meridian in his or her visual field. However, in V2, there is a marked separation between neurons that represent the visual field just above and just below the meridian. Therefore, one might expect some deleterious effect on IC responses for meridian-crossing ICs in V2 neurons. If that decrease in IC sensitivity in V2 neurons were observed, it might cast doubt on the role of V2 alone in IC perception. Moreover, as discussed later, the human data point to other brain areas as the major processing sites for ICs. The monkey results could be interpreted to indicate that similar IC-related activity is occurring in human V2, but fMRI and other techniques used on humans are too insensitive to measure it. A second possibility is that the V2 activity seen in monkeys is related to IC perception but is not the central mechanism involved in the percept. Another possibility is that human and monkey perception and neural mechanisms are fundamentally different at this midlevel stage of visual processing.

Human responses to ICs have been measured with fMRI techniques. Most fMRI studies have involved the measurement of the activation of Kanizsa squares or diamonds compared with the same pacman-shaped inducers rotated outwards or all in the same direction. Pacman is a term that refers to the shape of an agent in a video game from the early 1980s. The shape of "Pacman" was exactly the same as the cut-off circles used by Kanizsa in the IC figures he originated. Early studies by Hirsch and colleagues and by Ffytche and Zeki found that there was activation of the occipital cortex lateral to V1 by Kanizsa-type figures, but they could not pinpoint the cortical location because the IC experiments were not combined with retinotopic mapping. Therefore, these studies established that signals related to segmentation were present in occipital cortex, but further work was needed to determine more precise localization.

The extensive research of Janine Mendola and colleagues established that IC-related signals were observed in retinotopic area V3 and also in LOC, the lateral occipital area previously discovered by Malach and colleagues. Figure 7 is an fMRI image indicating the large region of cortical activation evoked by the Kanizsa diamonds used as stimuli in Mendola's study. The early retinotopic areas V1 and V2 did not produce statistically significant activation. Mendola and colleagues also used different inducers for ICs, such as aligned line endings, and found a similar pattern of brain activation in V3 and LOC. These results are important in implicating extrastriate cortex in the process of visual segmentation in humans.

Although ICs are often chosen for studying visual segmentation, other visual phenomena can lead us to an understanding of this very important process. The assignment of an image region as figure or ground, is one such phenomenon. As Edgar Rubin, the famous perceptual psychologist, pointed out in 1921, such assignment is automatic and inescapable. However, ambiguous figures exist in which figure and ground assignments flip back and forth, and perception changes when that happens. The familiar face/vase figure from E. Rubin is the most reproduced example, but there are other examples from Rubin that illustrate the consequences of figure/ground assignments even better. One of these is the Maltese cross figure shown in Fig. 8. This example is described in Koffka's 1935 book but not depicted in it. The diamond-shaped arms of the cross can appear grouped in fours, with a vertical and horizontal pair grouped together as figures in front (resembling a propeller in shape) and then two diagonal pairs grouped together as figures in front (and the vertical–horizontal pairs are then in back). The contrasts in the figure are arranged such that the vertical–horizontal propeller shape looks darker in front than it does when it is perceived "in back," appearing as a light gray diamond behind the white tilted propeller. This is because of the enhanced effect of brightness contrast across borders that define a figure and on the regions to which such borders are attached, as Rubin noted. Similar effects can be seen in color. This is only one of many illustrations of the deep consequences of figure/ground assignment. For instance, another consequence of the importance of

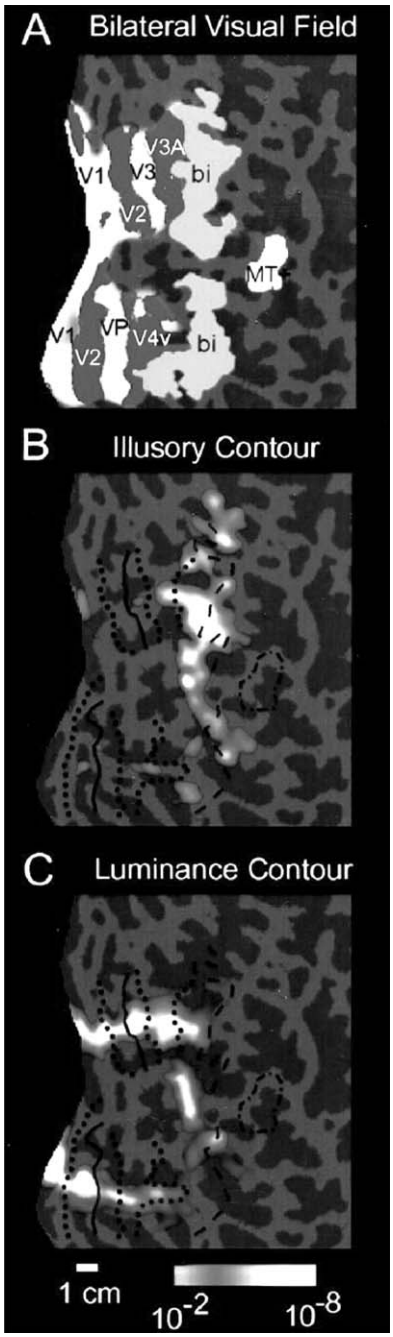


Figure 7 Mapping of IC responses in human cortex with fMRI. (A) Map in human visual cortex of retinotopic visual areas and also the interhemispheric region as obtained by use of phase mapping as in Figs. 3 and 4. (B) The region of differentially high activity when Kanizsa illusory contours are compared with activation produced by the unorganized pacman inducers. The main activation is in V3A and in the interhemispheric region. (C) Activation produced by a square defined by luminance contours [reproduced with permission from Mendola *et al.* (1999). The representation of illusory and real contours in human cortical visual areas revealed by functional magnetic resonance imaging. *J. Neurosci.* **19,** 8560–8572.]

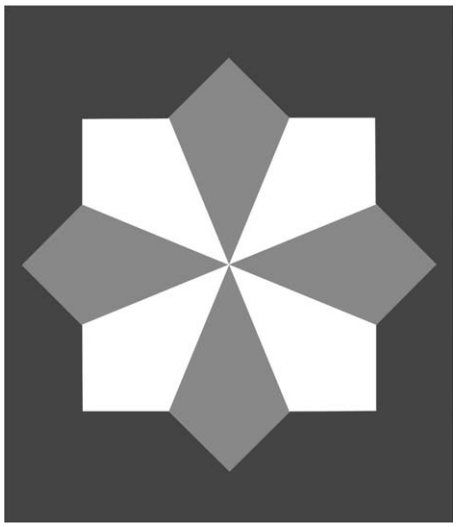


Figure 8 Maltese cross figure from E. Rubin. There are two sets of vanes, white and gray. They group together to produce "propeller" figures that alternate front and back. When the propeller of a given color is seen in back, it tends to complete into a gray diamond or a white square, respectively. Some observers report that the perceived value of the gray changes, becoming darker when the gray regions form a propeller in front and lighter when they form a diamond in back (drawn from a description in Koffka, 1935).

figure/ground is that people remember the shapes of figures, not grounds. Therefore, understanding the neural basis for this phenomenology is an important clue to the function of the visual system.

Figure/ground assignment is a special case of a more general problem in vision—the assignment of border ownership. Assignment of a region as figure or ground is all one has to do if there is only one figure surrounded by the background. However, if there are many figures, and if one is in front of another so that it partly occludes the shape of the second figure in the visual image, then the visual system must decide on the basis of image information which surface is in front along the boundary between the two figures in the image. Briefly, the brain has to decide which figural region owns the border between them (i.e., the front surface). Assignment of border ownership is a problem that must be solved for almost every visual image.

There have been only a few investigations of neural mechanisms for border ownership and figure/ground assignments. Zhou and colleagues studied single cells in V1 and V2 cortex of macaque monkeys. By keeping local edge contrast the same but varying the global stimulus so that different regions own the boundary between them perceptually, Zhou *et al.* tested

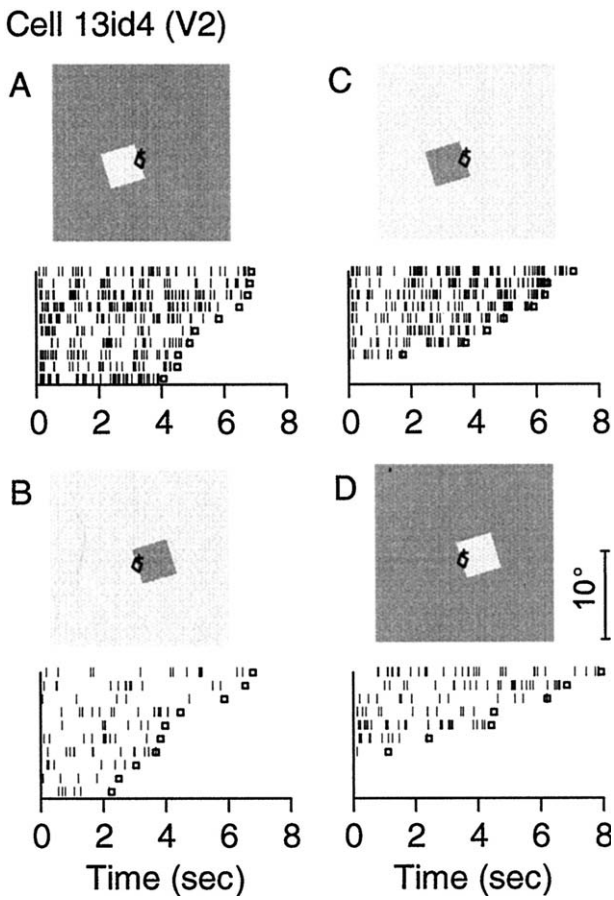


Figure 9 This is a representative neuron from V2 cortex in a macaque monkey that reveals sensitivity to border ownership. The same polarity edge border is in the pairs A and B and C and D, but the cell responds most strongly when the border "belongs" to a figural region down and to the left of the cell's receptive field (black line). Each vertical stroke in the raster plots stands for a nerve impulse [reproduced with permission from Zhou *et al.* (2000). Coding of border ownership in monkey visual cortex. *J. Neurosci.* **20,** 6594–6611.]

sensitivity to border ownership in single cortical neurons. The experimental design and results in an archetypal border ownership cell are shown in Fig. 9. A substantial number of cells such as this are encountered in monkey V2 cortex. Baylis and Driver reported that many neurons in monkey inferotemporal (IT) cortex respond differentially to figure or ground; thus, these also must reflect signals about border ownership. Since the IT cortex is supposed to be involved in object recognition, it is reasonable that neurons in this area should be affected by border ownership that is necessary for accurate object recognition in the real world.

Studies by Kleinschmidt and colleagues of figure/ground reversals in human cortex using fMRI revealed activation over a number of areas in occipital, temporal, parietal, and even frontal cortex. The involvement of temporal, parietal, and frontal cortex seems to imply that activation of top-down influences from "high-level" cortical areas could be necessary for figure/ground reversal of border ownership. However, as in the case of ICs, it is also possible that there may be signals associated with figure/ground assignment in "early" retinotopic areas, such as V1 or V2, that are undetectable with fMRI.

An important part of segmentation in human visual perception is the phenomenon of amodal completion—that is, completion and grouping together of the parts of a partially occluded object that are visible. This completion process is crucial for normal object perception in the real world. Evidence that amodal completion affects the firing rates of V1 neurons in macaque V1 was obtained by Y. Sugita by manipulating apparent occlusion using stereopsis (Fig. 10). Only a small fraction of V1 neurons were affected by amodal completion, but it is still a significant result.

IV. MT AND MOTION

Two of the strongest stimulus cues for segmentation are motion and color. This makes sense ecologically since it is unlikely that separate things will move together for any length of time, and similarity in color is associated with similarity in surface properties. The visual cortex appears to agree with this reasoning because it devotes a large amount of specialized cortical processing for motion and for color. We first consider motion processing in the middle temporal area MT (or V5). This brain area was initially defined by Zeki in his early work in the macaque extrastriate visual cortex. Zeki noted the large number of directionally selective cells in what he called V5, the middle temporal area later called MT by others. In human cortex, a motion-sensitive homolog to macaque MT was suspected from neuropsychological work on brain-damaged patients. Zihl and colleagues described a patient who had a lesion in dorsolateral occipital cortex and who had lost the ability to see motion. The motion-blind patient reinforced the concept of Zeki and others that there was a discrete brain region, assumed to include MT, that was obligatory for the perception of motion.

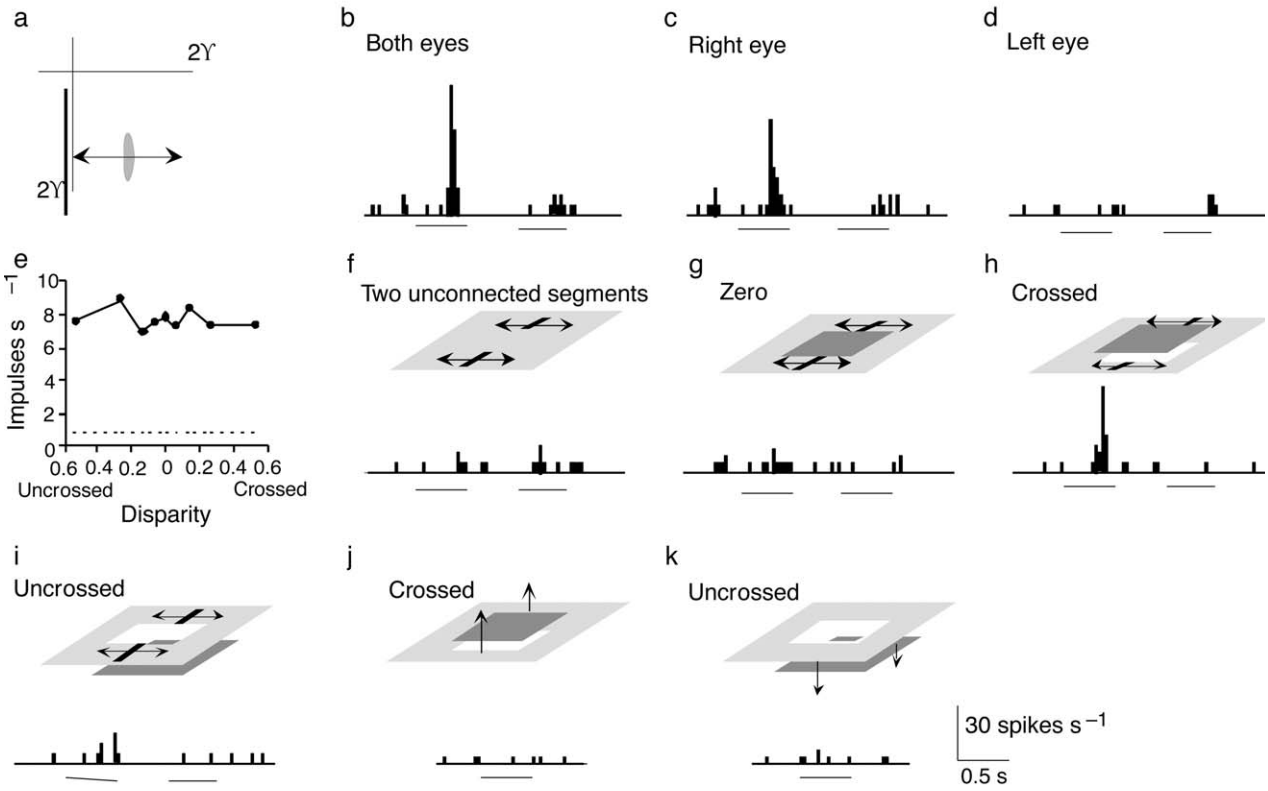


Figure 10 Macaque V1 neurons that respond to amodal completion. The neuron responds to a long contour in its receptive field as in a. It responds to both eyes (b) and to the right eye alone (c) but not to the left eye alone (d). The interesting manipulation is in f–h. In f, the neuron does not respond to two unconnected segments. In g, it does not respond to the same two segments when they are perceived as being in front of a gray region. However, there is a response when the retinal disparity is such that the gray region is in front and occluding the two line segments (h) [reproduced with permission from Sugita (1999). Grouping of image fragments in primary visual cortex. *Nature* **401,** 269–272.]

Functional imaging in normal human subjects, with visual stimulation by moving dots compared to stationary flashing dots, indicated that there was a discrete region in dorsolateral occipital cortex (lateral, dorsal, and anterior to LOC) that had most differential activity. Figure 11 shows the location bilaterally of what is now assigned to be human MT (V5). This is shown in a drawing of a slice of brain, not a flat map. Many other groups have confirmed this localization of MT in human cortex. The location of MT in the flat map representation of fMRI images of occipital cortex is given in Fig. 7 (top), in which MT's 1ocation is mapped and the locations of the retinotopically mapped areas are also indicated.

Classic results on MT in monkey cortex tended to confirm the notion that MT neurons were devoted to extracting velocity (direction and speed) of contours or patterns over the large receptive fields of the MT neurons. A significant number of MT neurons are directionally selective when stimulated with drifting bars or grating patterns. Using plaid patterns similar to those that Hans Wallach introduced to study motion coherency and transparency in human perception, Tony Movshon and colleagues found that some MT neurons were tuned to the coherent motion direction of a plaid pattern, whereas other MT cells were only selective for the motion directions of the component gratings that summed to make the plaid. These two kinds of neurons were given the labels "pattern" and "component" neurons, respectively. The assignment of two separate names might be supposed to carry the implication that there are distinct classes of neurons in MT. However, this was not claimed. In fact, existing data tend to suggest that there are not two distinct types of neurons in MT but rather a continuum of pattern vs component

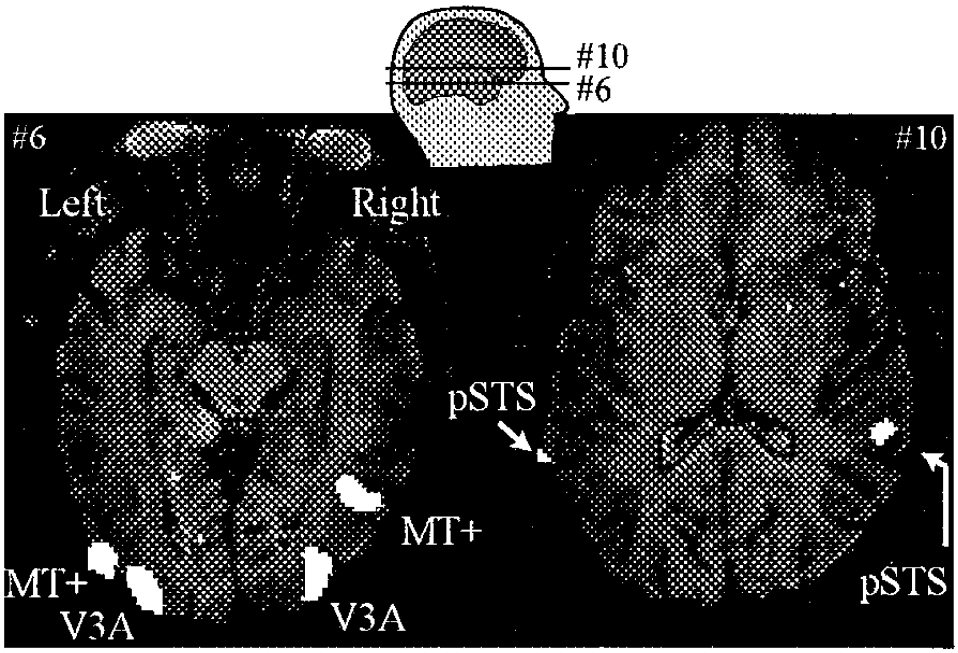


Figure 11 Mapping of human MT with fMRI. Moving vs stationary dot patterns elicit two active regions in the posterior cortex. These are labeled V3A and MT in slice 6 from comparisons with retinotopic maps. A parietal cortex region is also activated by moving vs stationary stimuli and this is shown in slice 10 [reproduced with permission from Ahlfors *et al.* (1999). Spatiotemporal activity of a cortical network for processing visual motion revealed by MEG and fMRI. *J. Neurophysiol.* **82,** 2545–2555.]

selectivity. There is definitely a large group of "mixed" cells that are partially component and partially pattern selective.

Although many macaque MT cells respond to grating patterns and contours, most MT cells respond most vigorously to small dots or populations of dots moved through their receptive fields. Also, it was shown that MT neurons would respond to coherent motion of a proportion of dots embedded within random dot kinematograms, and that the coherency thresholds for individual MT neurons were similar to behavioral thresholds for the monkey. This work in Bill Newsome's laboratory suggested that MT was involved in the perception of coherent motion in a definite direction. The work on cortical microstimulation in macaque MT cortex by Salzman and colleagues in Newsome's laboratory showed that differential activation of subpopulations of macaque MT neurons could bias perceptual judgments of coherency. Indeed, when coherency was low, MT microstimulation could cause the monkey to behave as if it perceived the random dot kinematogram with no coherent dot motions to have significantly suprathreshold coherency. These results were consistent with the interpretation that MT activity determined the perceptual judgment of coherency and direction.

Braddick and colleagues studied human fMRI response to coherency of motion in random dot kinematograms, in analogy to the experiments done by Newsome's group on macaque MT. Their experimental design was broader in that they were comparing the pattern of activation in different cortical areas in response to different kinds of stimuli that were designed to elicit responses to motion coherency and to form coherency. However, for the current discussion, we focus on the motion coherency results. Their stimuli for motion coherency were in essence the same as those used by Salzman *et al.* The comparison stimulus in their fMRI experiments was totally incoherent random motion of a field of dots. They found a widespread pattern of activation caused by motion coherency including MT (V5), the retinotopic area V3, as well as isolated regions of temporal and parietal cortex.

The role of human MT in motion perception has been examined in another kind of experiment: the correlation of brain activation with the motion aftereffect by Roger Tootell and colleagues. Prolonged

viewing of motion in one direction leads to a familiar perceptual aftereffect when the stimulus motion ceases. Stationary objects appear to move in the opposite direction. Tootell *et al.*'s fMRI study of the motion aftereffect used expanding concentric rings as a motion stimulus and stationary concentric rings as a control. They found that at the termination of the motion stimulus, MT activation continued for another 20–30 sec. This prolonged afteractivation paralleled almost exactly the subjective perception of the motion aftereffect (apparent contraction of the concentric rings). Because there was no physical motion during the aftereffect period, but there was subjective experience of motion counter to that seen during the stimulation period, this is a very strong result indicating that MT activation is correlated with the subjective experience of perceived motion. Other studies have confirmed this result, finding motion aftereffect activation in MT.

A number of recent studies indicate that MT cortex is not simply responding to motion direction or speed but, rather, may be playing a role in visual segmentation. For example, at the Salk Institute, Stoner and Albright, working with Ramachandran, initiated a series of studies combining experiments on human perception and on single cells in macaque MT that indicate that MT neurons are affected by form cues as well as by motion. They used plaid patterns and studied human perception of coherency as they changed the brightness at the intersections between the lines of the plaid. When the intersection brightness was not compatible with optical transparency of the perceived overlaying line, the probability of perceiving coherency was high and the probability of perceiving transparency was low. When the intersection brightness was compatible with transparency, the probability of perceiving motion transparency became greater. This perceptual effect was also confirmed in macaque MT neurons, as shown in Fig. 12. The direction selectivity of neurons in MT was changed by the same manipulation of intersection brightness and hence of consistency with physical transparency. Thus, in macaque monkey MT there is evidence that individual neuron activity is determined by segmentation (or combination) of motion signals contingent on cues for transparency.

Another line of evidence that macaque MT neurons are responding to moving surfaces and not simply a

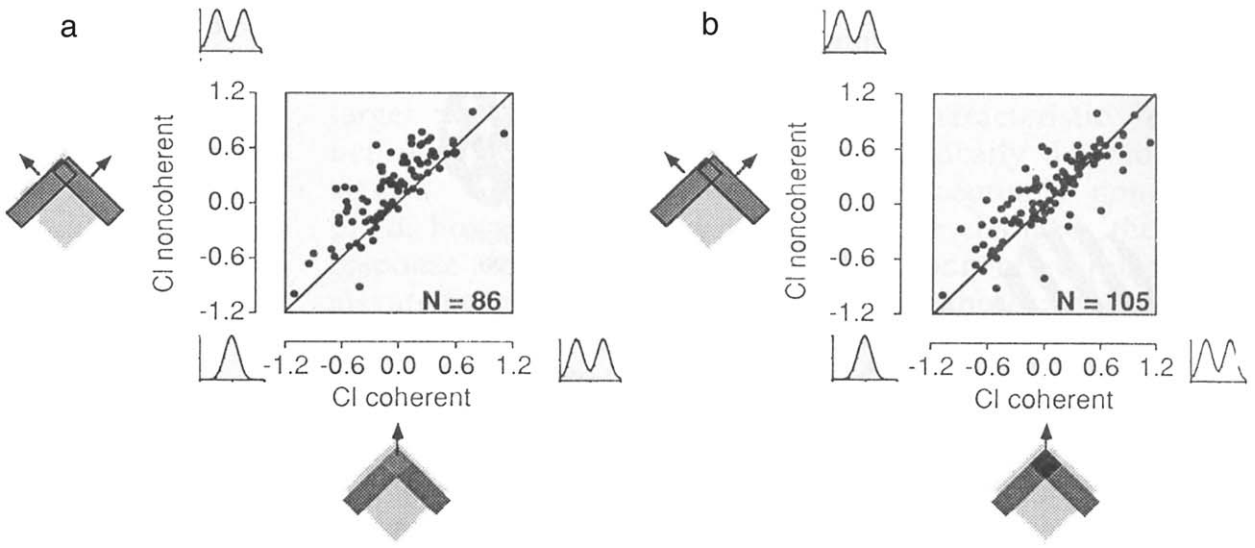


Figure 12 The effect of static form cues on motion coherency in MT neurons in macaque visual cortex. Plaid patterns were moved through MT receptive fields. Each point in each graph represents a single neuron and the scatter plots represent the entire population of MT neurons studied. The graphs plot the calculated component index (CI), which estimates the amount of a neuron's directional tuning curve that is explained by transparent (or "sliding") motion as opposed to coherent plaid motion. The horizontal axis in each graph is the CI measured and calculated for stimuli that should be perceived as coherent because the static brightness of the intersection is not consistent physically with transparency. The vertical axis is for a case in which the intersection brightness is consistent with transparency. The tendency of the points to lie above the line indicates that when physical transparency is possible, motion transparency is perceived more often [reproduced with permission from Stoner and Albright (1992). Neural correlates of perceptual motion coherence. *Nature* **358,** 412–414.]

velocity vector derives from experiments on the phenomenon of the kinetic depth effect (KDE). This is a classical perceptual phenomenon introduced and explored perceptually by Hans Wallach and one of his students in 1953. A visual image formed by the projection of a moving 3D object on a flat screen elicits a strong perception of 3D structure. The perceptual ability to see KDE is the most important means for judging 3D shape for objects that are too distant for a subject to use binocular stereopsis for depth and shape judgments (i.e., for distance >5 m). Such depth from motion is combined with stereo for depth judgments at any distance. David Bradley and colleagues found that many macaque MT neurons were tuned for depth. This finding was confirmed and amplified by the later work of DeAngelis and co-workers, who showed that there was an orderly mapping of depth in MT cortex. For example, some MT neurons would be responsive only when the moving target was in front of the fixation plane. Bradley then showed that such depth-sensitive MT neurons were not only responsive to depth based on stereopsis but also signaled depth from KDE (Fig. 13). The KDE stimuli used by Bradley *et al*. were random dot kinematograms like those used in the 1970s for studying KDE perception by Shimon Ullman and others. The moving dot flow fields are ambiguous stimuli that could be perceived in a number of ways. Bradley *et al*. demonstrated that the neurons responded to KDE when the monkey perceived the KDE as creating a surface at the appropriate depth for the neuron's depth tuning. This is an important result that indicates the perceptual sophistication of MT neurons. However, in this case the homology with human MT may be less close than in other interspecies comparisons. In an experiment on KDE in humans, Paradis and colleagues used fMRI. It revealed that many areas of the brain responded differentially to KDE, but MT was not among those KDE regions. Rather, V3 and regions in parietal and temporal cortex were activated more by KDE than by random dot motion. These authors speculated that in humans, the higher form-related functions of MT were taken over by parietal cortex.

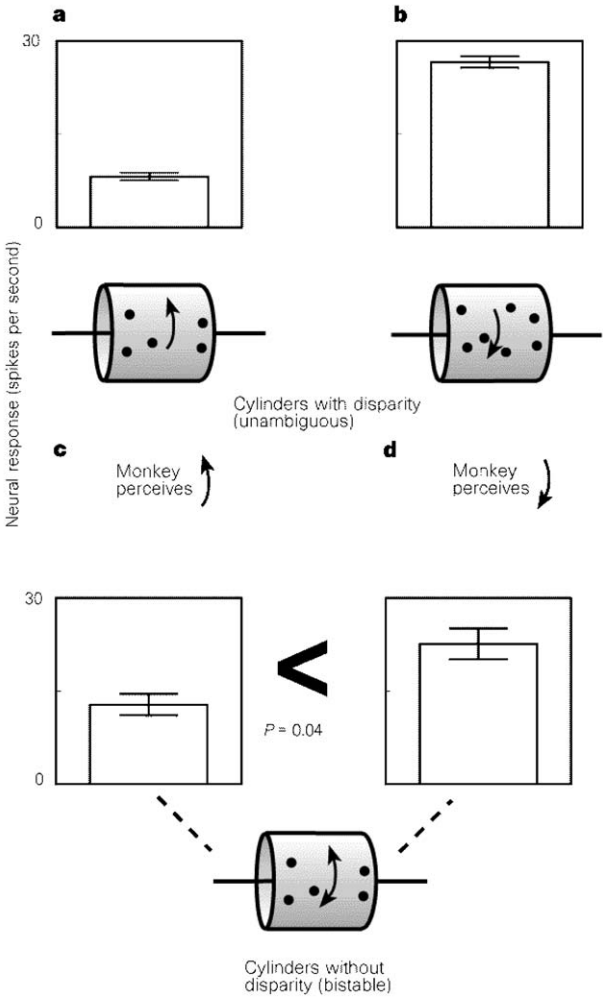


Figure 13 KDE in macaque MT neurons. Monkeys viewed two-dimensional projections of three-dimensional, revolving cylinders. They reported the direction of rotation they perceived by choosing a target moving in the same direction as the perceived front surface. The depth of the cylinders could be determined by stereopsis using disparity or by the motion field (KDE). Data are from an MT neuron. (a, b) Responses to 100%-disparity cylinders (using only trials with correct responses). (c, d) Responses of the same cell to zero-disparity cylinders, when the monkeys reported them as going front-up (c) or front-down (d) [reproduced with permission from Bradley *et al*. (1998). Encoding of three-dimensional structure-from-motion by primate area MT neurons. *Nature*, **392,** 714–717.]

Another important result more directly indicates the linkage between form and motion in human MT. Rainer Goebel and colleagues, studying fMRI activation in humans by different apparent motion stimuli, found that apparent motion of organized Kanizsa squares elicited much more fMRI activation than did apparent motion of shuffled pacmen. This result also corresponds with perception: There is a much stronger percept of apparent motion with moving Kanizsa squares than with the same pacmen pointing outwards. These results suggest that the nature of human MT's response to moving visual surfaces needs to be further investigated. For instance, experiments such as those by Bradley combining motion flow fields with

variations in depth organization should be attempted on human MT and compared with Bradley's results in macaques.

V. COLOR IN VENTRAL OCCIPITAL BRAIN AREAS

The idea that there is a separate localized area in human extrastriate cortex that is specialized for mediating color perception is derived from the phenomenon of cerebral achromatopsia (cortical color blindness). Achromatopsia is usually caused by stroke lesions in ventral occipital–temporal cortex. It is a variable clinical syndrome, with the common feature that patients cannot pass tests indicating they can perceive colors normally. The critical test is the Farnsworth–Munsell 100 hue test, which involves color arrangement. Normal humans perform this test very accurately and achromatopsics can be at chance in this test of arranging hues in an orderly sequence. Failure on the 100 hue test correlates very well with subjective reports that the patients cannot tell what color they are looking at. Because this neurological syndrome is usually accompanied by lesions in a particular region of cortex, there is strong suspicion that this extrastriate cortex subregion is the color-specialist area, a view stated forcefully by Semir Zeki. However, we review and modify this concept of a color module in the following discussion.

Human imaging studies comparing activation by color vs achromatic patterns are quite consistent in identifying a ventral occipital–temporal region as the color area. Originally, this region was dubbed V4 by Lueck and colleagues, who used positron emission tomography (PET) to find the regions of the brain that were preferentially activated by color. This was later confirmed with fMRI by Zeki and colleagues. The naming of the color region as V4 cortex was done in analogy with Zeki's previous studies on macaque cortex, in which he proposed that macaque V4 was specialized for color processing. This proposal of color specialization in macaque V4 has been challenged many times and is still controversial, especially due to the research of Heywood and Cowey, who found that lesions in V4 did not have a major effect on color discriminations in monkeys. Hadjikhani, Tootell, and colleagues mapped the color area in the same manner as Zeki and colleagues, but they also measured retinotopic mapping in the same subjects to find out if the color area Zeki and colleagues discovered was indeed V4 in the retinotopic framework (Fig. 14). They found that it was quite anterior to V4 as measured with retinotopic mapping. There is no disagreement on the location of the color-preferring area, just on its proper name. Hadjikhani and colleagues added an elegancy to their study by showing that the color region they

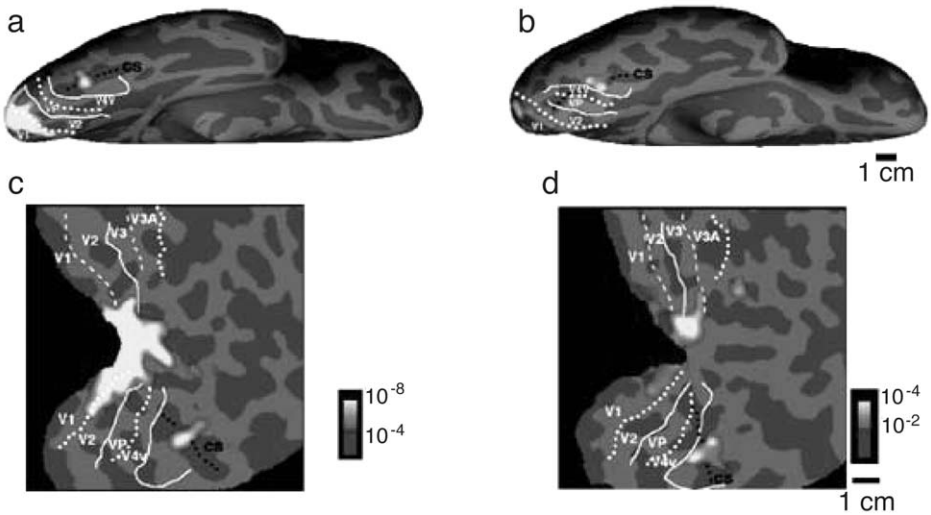


Figure 14 Mapping of human V8 cortex in ventral visual cortex with fMRI. The visual stimuli used for mapping were radial gratings that were either black–white or equiluminant color (red–green). The activation maps are for the difference between these two stimuli. The important point is the placement of the ventral activation outside the region of retinotopically mapped V4 [reproduced with permission from Hadjikhani *et al.* (1998). Retinotopy and color sensitivity in visual cortical area V8. *Nat. Neurosci.* **1,** 235–241.]

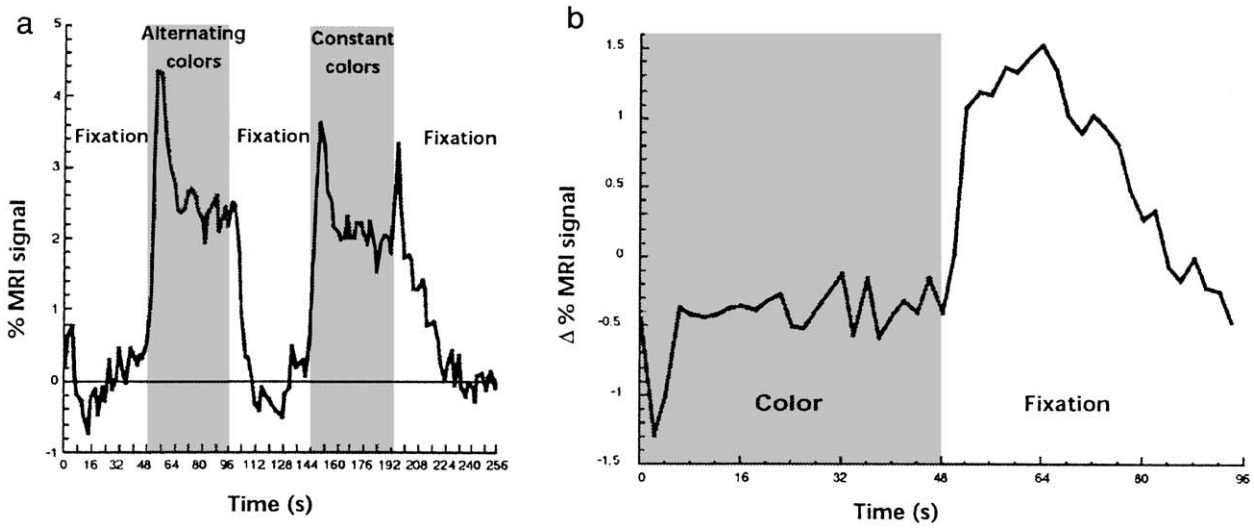


Figure 15 Color afterimages in fMRI of V8 cortex. (a) Activations caused by perceiving stimuli to create color afterimages. The fMRI activation for the constant colors condition outlasts the stimulus by 10–20 sec compared to the response to the alternating colors condition because of the color afterimage. (b) The alternating colors response has been subtracted from the constant colors response to illustrate the time course of the afterimage activation during the fixation period. Only regions of cortex activated by the color mapping stimulus exhibit these long afterresponses to color afterimages [reproduced with permission from Hadjikhani *et al.* (1998). Retinotopy and color sensitivity in visual cortical area V8. *Nat. Neurosci.* **1,** 235–241.]

found, which they named V8, responded persistently after a color stimulus ceased (Fig. 15). This poststimulus activity correlated with the perception of color afterimages. In this way, it resembled the motion aftereffect activity earlier observed in MT by Tootell and colleagues.

There are other methods for testing for color activation besides the differencing method used by Zeki and by Hadjikhani. The differencing method assumes that color perception will be associated with activation that is linked to color but that disappears when achromatic stimulation is used. Implicitly, this method includes the assumption that the neural mechanism for color perception is composed of neurons that only respond to color and do not respond to achromatic stimuli. Engel *et al.* showed that many other areas of visual cortex, including V1, will respond to purely chromatic modulation but will also respond to chromatic modulation. This finding in human fMRI is related to recent single-cell studies in macaque V1 that indicate that there are many neurons in V1 that respond to both color and brightness. Such neurons could play an important role in color perception but would be ignored by the difference image methods. Therefore, the amount of cortex involved in color perception may have been significantly underestimated by these difference methods, a point made by Brian Wandell. Nevertheless, it is important to note that the region of cortex involved in achromatopsia lesions and the region of cortex mapped with PET and fMRI with the differencing methodology are similar.

VI. OBJECT RECOGNITION IN LOC AND VENTRAL OCCIPITAL–TEMPORAL CORTEX

One of the chief functions of visual perception is object recognition. Although it is still too soon to tell the whole story of how the visual system recognizes and categorizes objects, recent research indicates remarkable specialization of specific regions in visual cortex for recognizing objects and even classes of objects. This topic can be subdivided into a discussion of LOC and ventral occipital–temporal cortex.

Lateral and slightly anterior to retinotopic areas V3A and V4V, there are regions of human occipital cortex that are activated preferentially by pictures of objects compared with pictures of random textures. This has been shown repeatedly in fMRI experiments

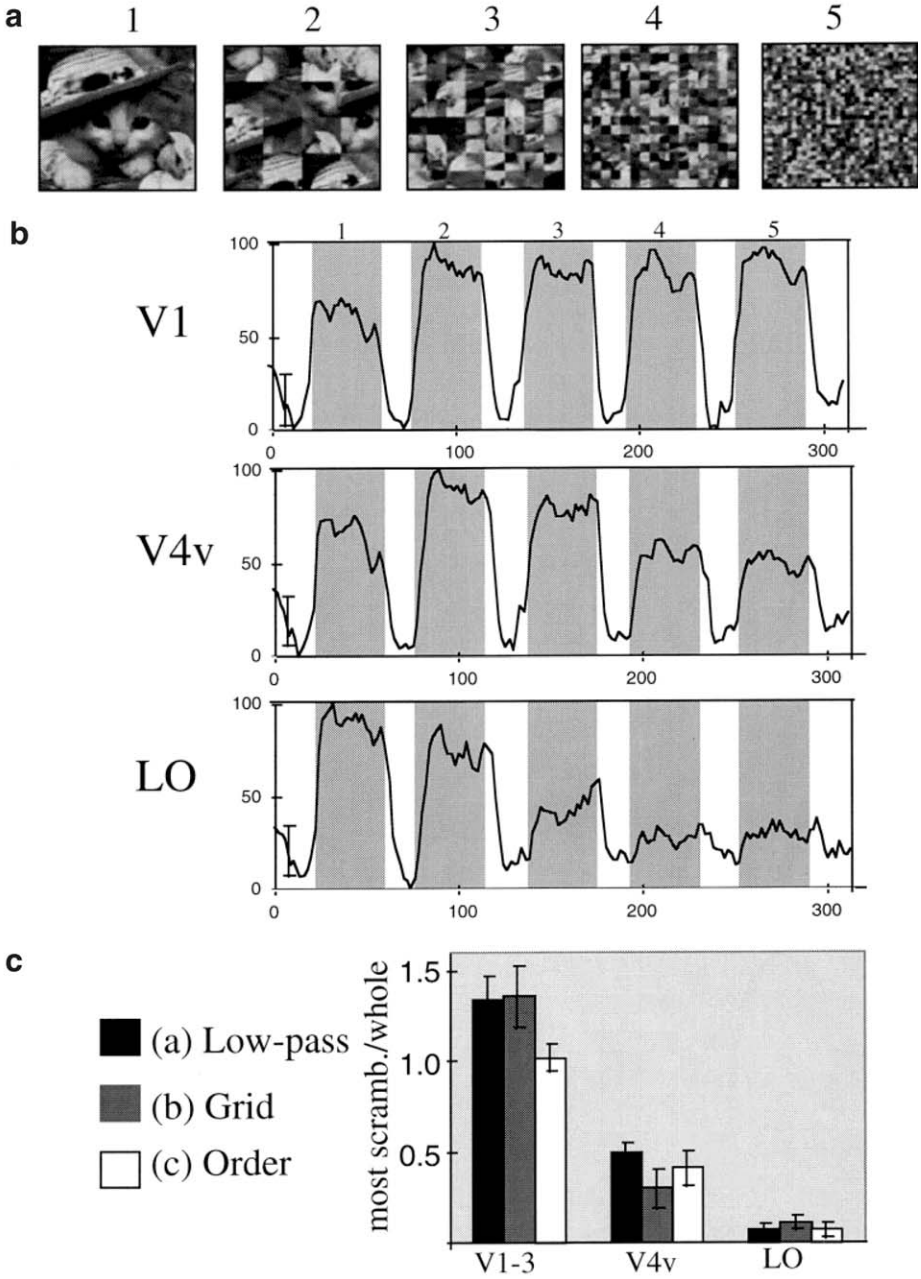


Figure 16 Responses in human fMRI to scrambled pictures of objects in LO and V1 and V4 cortex. Pictures such as those shown in row a were used as stimuli. (b) Time series of responses in LO, V1, and V4v. LO shows the largest effects of scrambling of the object pictures, indicating it is responding to large subregions of each picture and that it cannot tolerate disarrangement of the parts [reproduced with permission from Grill-Spector *et al.* (1998). A sequence of object-processing stages revealed by fMRI in the human occipital lobe. *Hum. Brain Mapping* **6,** 316–328.]

on human subjects, first by Rafi Malach working with Tootell and then by Malach and colleagues. The ventral region most strongly activated is called LOC. LOC activation has many distinguishing characteristics. In experiments with cut and scrambled pictures of objects, LOC activation declined only slightly when the pictures were cut into four pieces and scrambled, but it declined precipitously when the pictures were cut

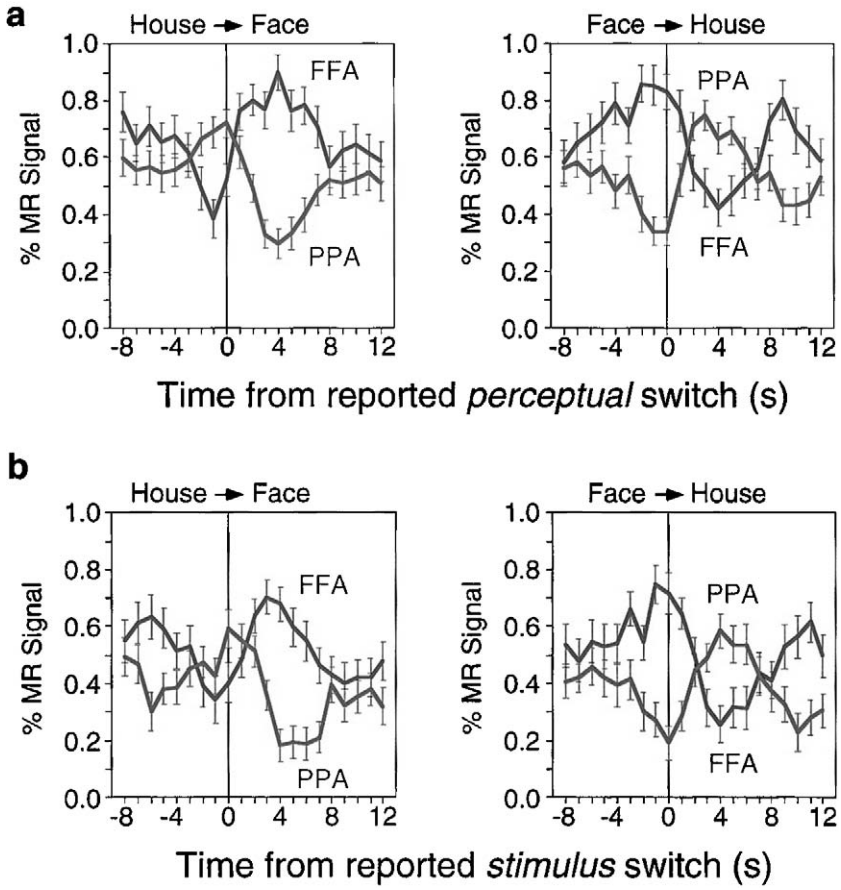


Figure 17 Binocular rivalry between pictures of faces and of houses. The fusiform face area (FFA) and parahippocampal place area (PPA) were monitored with fMRI in human subjects during binocular rivalry between pictures of faces and of houses. There was a clear modulation of the amount of activation during and after the switch from house to face and from face to house, as seen in a. (b) A control by stimulating both eyes with pictures of faces and of houses and measuring the magnitude of fMRI amplitude modulation caused by switching the stimuli and percepts [reproduced with permission from Tong *et al.* (1998). Binocular rivalry and visual awareness in human extrastriate cortex. *Neuron* **21,** 753–759.]

and scrambled into smaller pieces (Fig. 16). LOC activation was relatively invariant to change in location and to changes in the size of pictures of the same object. Therefore, activation of LOC seems to resemble neuronal activation in IT cortex of monkeys. LOC seems to have a bias toward visual information from the center of the visual field. From fMRI studies on retinotopic mapping, Tootell and colleagues concluded that the region in which LOC is located receives interhemispheric inputs, meaning that it represents visual signals coming from both right and left visual fields. LOC is also an area activated by the IC stimuli used by Mendola (Fig. 7). These various experimental results suggest that LOC could be an important area for intermediate-level vision, where segmentation and grouping are made explicit in the activity of neurons and neuronal populations.

Further specialization for the recognition of classes of objects has been found in occipital–temporal areas of human cerebral cortex. These areas are located in the ventral cortex and more anteriorly. A much studied region is the anterior fusiform gyrus, where pictures of faces activate the region preferentially, as first shown by Nancy Kanwisher and colleagues. It was shown that pictures of faces compared to pictures of houses activate this fusiform face area (FFA) differentially. Also, pictures of faces compared to inverted face pictures also produce differential activation in a

comparison paradigm. Nearby but distinct from the FFA is the parahippocampal place area (PPA), a region that prefers pictures of houses to those of faces. This more posterior region is located in the collateral sulcus, also in the ventral occipitotemporal cortex. Malach and colleagues showed that the FFA was biased more toward central vision, whereas the PPA in the collateral sulcus was biased more to the periphery of the visual field. Also, these occipitotemporal regions overlap with the interhemispheric region mapped by Tootell and thus represent both right and left visual hemifields.

The specialization of the face and house regions of the occipitotemporal cortex was shown in an elegant experiment by Frank Tong and colleagues. They used the phenomenon of binocular rivalry combined with fMRI measurements to explore the modulation of activity with change of perceptual state. Subjects viewed face and house pictures that were present monoptically—for example, a picture of a house to the right eye and a picture of a face to the left eye. In these circumstances, perception is bistable: First one sees a face, and then after some time the picture of the face fades and one sees a house, and the two percepts repeatedly alternate in perception. Tong and colleagues asked subjects to press different keys when they saw a face or house, and they averaged the fMRI signal with respect to the key presses. They observed the results in Fig. 17. FFA activation increased when subjects perceived faces in the rivalrous situation, and PPA activation increased when they saw houses. The respective activations were comparable to what was measured when only faces or only houses were shown. This experiment reveals in a very direct manner that activation of these specific cortical areas is highly correlated with the perception of specific categories of objects.

VII. CONCLUSION

Perception requires location and mapping of space, segmentation, sensitivity to motion and color, and recognition of familiar or unfamiliar shapes. We discussed all these aspects of visual perception and the areas of the cerebral cortex that are activated when these are seen. It can be concluded from the experiments we reviewed that the neural basis of visual perception is based on specialized modules in the visual cortex. However, in perception there can be an overriding influence of part of a visual scene on the perception of other parts of the scene, as in the effects of visual cues for segmentation. Therefore, it cannot be definitively concluded that specialized modules are all there is to neural mechanisms of visual perception. A major role may be played by the interaction between different visual areas and between the visual networks in occipital cortex and memory- or decision-related cortical networks elsewhere in the brain. As the Gestalt psychologists used to say about visual images, so also we can say about the visual cortex: The whole is greater than the sum of its parts.

See Also the Following Articles

AREA V2 • COLOR VISION • EYE MOVEMENTS • MOTION PROCESSING • MULTISENSORY INTEGRATION • RETINA • SENSORY DEPRIVATION • SPATIAL VISION • VISUAL CORTEX • VISUAL DISORDERS • VISUAL SYSTEM DEVELOPMENT AND NEURAL ACTIVITY

Suggested Reading

Ahlfors, S. P., Simpson, G. V., Dale, A. M., Belliveau, J. W., Liu, A. K., Korvenoja, A., Virtanen, J., Houtilainen, M., Tootell, R. B., Aronen, H. J., and Ilmoniemi, R. J. (1999). Spatiotemporal activity of a cortical network for processing visual motion revealed by MEG and fMRI. *J. Neurophysiol.* **82,** 2545–2555.

Bradley, D. C., Chang, G. C., and Andersen, R. A. (1998). Encoding of three-dimensional structure-from-motion by primate area MT neurons. *Nature* **392,** 714–717.

Engel, S. A., Rumelhart, D. E., Wandell, B. A., Lee, A. T., Glover, G. H., Chichilnisky, E. J., and Shadlen, M. N. (1994). fMRI of human visual cortex. *Nature* **369,** 525.

Goebel, R., Khorram-Sefat, D., Muckli, L., Hacker, H., and Singer W. (1998). The constructive nature of vision: Direct evidence from functional magnetic resonance imaging studies of apparent motion and motion imagery. *Eur. J. Neurosci.* **10,** 1563–1573.

Grill-Spector, K., Kushnir, T., Hendler, T., Edelman, S., Itzchak, Y., and Malach, R. (1998). A sequence of object-processing stages revealed by fMRI in the human occipital lobe. *Hum. Brain Mapping* **6,** 316–328.

Hadjikhani, N., Liu, A. K., Dale, A. M., Cavanagh, P., and Tootell, R. B. (1998). Retinotopy and color sensitivity in visual cortical area V8. *Nat. Neurosci.* **1,** 235–241.

Malach, R., Reppas, J. B., Benson, R. R., Kwong, K. K., Jiang, H., Kennedy, W. A., Ledden, P. J., Brady, T. J., Rosen, B. R., and Tootell, R. B. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proc. Natl. Acad. Sci. USA* **92,** 8135–8139.

Mendola, J. D., Dale, A. M., Fischl, B., Liu, A. K., and Tootell, R. B. (1999). The representation of illusory and real contours in human cortical visual areas revealed by functional magnetic resonance imaging. *J. Neurosci.* **19,** 8560–8572.

Paradis, A. L., Cornilleau-Peres, V., Droulez, J., Van De Moortele, P. F., Lobel, E., Berthoz, A., Le Bihan, D., and Poline, J. B.

(2000). Visual perception of motion and 3-D structure from motion: An fMRI study. *Cereb. Cortex* **10,** 772–783.

Stoner, G. R., and Albright, T. D. (1992). Neural correlates of perceptual motion coherence. *Nature* **358,** 412–414.

Sugita, Y. (1999). Grouping of image fragments in primary visual cortex. *Nature* **401,** 269–272.

Tong, F., Nakayama, K., Vaughan, J. T., and Kanwisher, N. (1998). Binocular rivalry and visual awareness in human extrastriate cortex. *Neuron* **21,** 753–759.

Tootell, R. B., Reppas, J. B., Dale, A. M., Look, R. B., Sereno, M. I., Malach, R., Brady, T. J., and Rosen, B. R. (1995). Visual motion aftereffect in human cortical MT revealed by fMRI. *Nature* **375,** 139–141.

Wandell, B. A. (1999). Computational neuroimaging of human visual cortex. *Annu. Rev. Neurosci.* **22,** 145–173.

Zhou, H., Friedman, H. S., and von der Heydt, R. (2000). Coding of border ownership in monkey visual cortex. *J. Neurosci.* **20,** 6594–6611.

Visual and Auditory Integration

JENNIFER M. GROH and URI WERNER-REISS
Dartmouth College

GLOSSARY

binding The mutual association of all the component features of an individual object or event. Within the visual domain, the features that must be associated include shape, color, and location. When objects are detected by more than one sense, binding refers to the association of all the different sensory cues generated by the same object or event.

frame of reference The particular set of axes used to describe a location in space.

linear function A function $f(x) = mx$, where m is a constant. Linear functions have the property of superimposibility: $f(a + b) = f(a) + f(b)$.

microstimulation A technique for activating a small population of neurons (*in vivo*) by passing electrical pulses into the neural tissue through an electrode.

movement field The range of eye movements in space that are associated with a modulation in the activity of a neuron. A movement field is essentially a receptive field for motor-related activity.

receptive field The range of locations from which a stimulus must originate in order to modulate the activity of a neuron.

saccade A rapid eye movement that brings the fovea to bear on a stimulus of interest. Saccades can be guided by visual, auditory, tactile, or remembered stimuli.

spatial correspondence A systematic relationship falling short of registry between the unimodal receptive field locations in individual multimodal neurons. Correspondence can also occur between the receptive fields of neighboring unimodal neurons that respond to different modalities.

spatial registry Perfect alignment between the locations of the unimodal receptive fields in a multimodal neuron; also, alignment between the receptive fields of neighboring neurons responsive to different modalities.

Our experience of the world as a unitary place requires that our brains meld together the information gleaned from our different sense organs. We can then act on this information, regardless of the source. The multisensory neurons in the brain may contribute to this process. In this article we review the properties of multisensory neurons in areas of the brain that participate in one kind of multisensory behavior: saccadic eye movements.

I. OVERVIEW

Knowledge of the world comes to us through our different senses. Our eyes monitor light patterns, our ears detect air pressure changes, our skin reports the pressure of objects impinging on it, and our noses and mouths measure the chemical content of the sensory scene. Despite the variety of routes that sensory information takes to reach our brains, we tend to experience the world as one place. Furthermore, we are capable of using different types of sensory information in equivalent ways, regardless of which sensory system was initially responsible for detecting the stimulus. For example, object shape can be detected either visually or through our sense of touch. Saccadic eye movements can be directed to the locations of visual, auditory, or tactile stimuli. When discrepancies occur between different sensory cues, one sense can dominate

another. When visual and auditory stimuli are mismatched in space, vision sometimes "captures" the location of the sound, a phenomenon known as the ventriloquism effect.

How is the mental experience of a combined multisensory event constructed from a collection of neurons sensitive to the individual components? Called the binding problem, this class of issues has received considerable attention within a single sensory domain. Vision scientists wonder how our percept of objects is synthesized from the activity of separate populations of neurons encoding individual visual features such as color, contours, and motion. For multisensory binding, the information to be bound together comes from many different sensory systems. Many neurons in the brain respond to more than one sensory modality; such neurons serve as a potential neural substrate for this binding. Multisensory neurons are actually common, having been identified in the parietal, temporal, and frontal lobes as well as in various brain stem areas. In this article, we discuss primarily visual and auditory integration. We take as our model system the areas of the brain that are responsible for guiding saccadic eye movements to visual and auditory stimuli. We review the response properties of neurons in these areas as well as those in selected regions of the brain that provide input to these areas. All the experiments we describe were conducted in either cats or monkeys unless otherwise noted. We seek to answer the following questions:

1. How do the responses of multimodal cells respond to each of the modalities alone? Are the receptive fields in the same frame of reference? Are the visual and auditory receptive fields of individual neurons in spatial register with one another?
2. How do they respond to combinations of stimuli? Is the response to a combined stimulus the linear combination of the responses to the components, for example?

In many instances, the complete answers to these questions are not known. Nevertheless, they will serve as a useful framework for exploring the current state of multisensory research.

II. THE REFERENCE FRAME PROBLEM

How does the brain determine that two pieces of sensory information arise from a common source? Visual and auditory stimuli are more likely to be the result of the same event if they arise from the same location, so determining whether two stimuli of different modalities are coming from the same place is an important task for the brain. However, this task is complicated by the fact that the visual and auditory systems employ different reference frames for their initial encoding of stimulus location. Visual stimuli are initially represented in retinal coordinates. Light is reflected from objects in the world, enters the eye, and activates a particular location on the retina. The site of retinal activation informs the brain concerning the location of the visual stimulus with respect to the direction of gaze. When the eye moves, the site of activation on the retina shifts.

The locations of auditory stimuli are computed based on interaural timing and level differences as well as spectral cues, yielding a head- and ear-centered frame of reference. If the source of a sound is located to the right, the sound will arrive in the right ear first, and it will be louder in the right ear. These interaural timing and level differences can be used to compute the location of the sound with respect to the orientation of the head and ears. Obviously, the position of the eyes would be irrelevant.

In species whose eyes have significant mobility (humans, monkeys, cats, but not barn owls, for example), these different reference frames pose a problem. Moving the eyes without moving the head will affect the location of retinal activation without influencing the auditory cues generated by a given object. On the other hand, moving the head while keeping the eyes pointed in a stable direction in space will alter the interaural acoustic cues without affecting the site of retinal activation. The solution to this problem appears to be a transformation of auditory signals into an eye-centered frame of reference so that comparison with eye-centered visual information can take place more easily.

III. BRAIN AREAS INVOLVED IN SACCADES TO VISUAL AND AUDITORY STIMULI

Saccades are rapid eye movements that direct the fovea to the positions of stimuli of interest. Saccades may be elicited by visual, auditory, or tactile stimuli. Our discussion focuses on some of the motor areas that have been implicated in the control of saccades to these visual and nonvisual stimuli as well as one of the areas that provides auditory input to these areas. The superior colliculus (SC), frontal eye fields (FEF), and

lateral intraparietal areas (LIP) are the major motor or motor-related areas that we discuss. One way that auditory signals reach this pathway is through an auditory cortical area in the anterior ectosylvian sulcus (AES, cat; the location is unknown in primate). Other sensory areas such as the inferior colliculus undoubtedly also contribute, but little is known about the role they play in multisensory processing so they will not be discussed.

A. Superior Colliculus

1. Oculomotor Function

The SC is a laminated structure located on the roof of the midbrain. The superficial layers receive a direct projection from the retina, but the intermediate and deep layers are primarily motor in function. Microstimulation in these layers elicits saccadic eye movements with a very short latency. The direction and amplitude of the evoked movement depend on the location of stimulation in the SC. In primates, the left SC encodes conjugate eye movements toward the right and vice versa. Movement size increases as the stimulating electrode is moved from rostral to caudal positions. Movements with an upward component are represented medially, whereas downward movements are represented laterally. Recording studies reveal that a prominent feature of the discharge characteristics of SC neurons is the vigorous burst of activity that they generate just prior to saccades. This activity is selective for the direction and amplitude of the saccade. The most vigorous activity occurs when the saccade will be directed to the center of the cell's so-called movement field. Movement fields are topographically organized and correspond to the direction and amplitude of the saccade evoked by stimulation at that site. Lesions of the SC, when combined with lesions of the FEF, abolish saccades completely. Taken together, these features place the SC convincingly within the saccade pathway.

2. Sensory Properties

The SC is not just a motor structure, however. Neurons in the intermediate and deep layers of the SC discharge in response to sensory events in three different sensory modalities: vision, hearing, and touch. These responses are sensory in the sense that they occur time locked to the onset of the sensory stimulus and they do not appear to actually trigger an eye movement. In other words, the neural responses occur even if the animal is trained to delay its saccade to the location of the stimulus until some kind of "go" cue is received. Although the responses can be dissociated from saccades in this fashion, the responses are considerably more robust if the animal subsequently makes a saccade to the stimulus than if the animal is asked to ignore the stimulus. Thus, the true character of the discharge is intermediate between purely sensory and purely motor. Indeed, individual neurons can show both sensory-related and motor-related properties. A schematic of an idealized SC neuron's response is shown in Fig. 1. When an animal is performing a delayed-saccade task, neurons in the SC may show a brief burst when the sensory stimulus is first turned on, followed by some level of sustained activity, and culminating in a second burst beginning just prior to the saccade. Actual individual neurons do not necessarily show activity during all three of these epochs.

Many different but overlapping schemes have been used to classify different cell types found in the intermediate and deep SC. Neurons that exhibit a burst of activity at approximately the time of the saccade are usually referred to as saccade-related burst neurons. Neurons that have both visual and motor-related activity are frequently referred to as visuomotor cells. Neurons that have both sustained and motor-related activity have also been called prelude-bursters. Finally, neurons showing any type of activity prior to

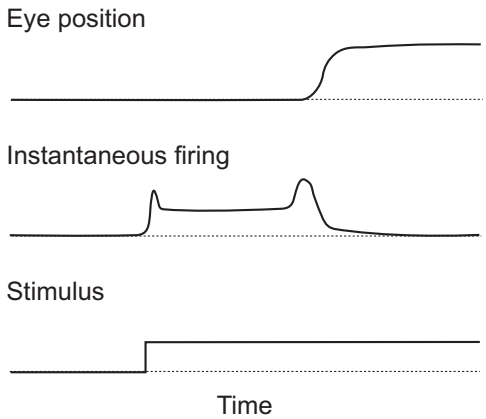


Figure 1 Response profile of an idealized SC or FEF neuron. The top trace illustrates the movement of the eyes, and the bottom trace shows the onset of a stimulus. (Middle) The discharge of the neuron is correlated with both the sensory stimulus and the movement. A burst of activity occurs at target onset and is sustained until another burst occurs at movement onset.

the saccade-related burst have been referred to as "buildup" neurons in many studies. However, this term is not particularly apt because the activity has not been shown to build up—the discharge is typically better characterized as being sustained at a steady level until the saccade-related burst occurs. All of these terms describe overlapping populations of SC neurons.

Early studies of the multisensory properties of these neurons were conducted primarily in anesthetized, paralyzed animals such as cats, rodents, and barn owls. Under these conditions, any motor-related properties of the cells cannot be determined. Instead, cells were classified based on their responses to visual, auditory, or tactile stimuli. Unimodal, bimodal, and trimodal cells have all been identified. Circumstantial evidence suggests that the multimodal cells studied in anesthetized animals might be part of the population of cells that show motor-related activity in awake animals: Multimodal neurons can usually be activated antidromically by microstimulation in the efferent output pathway from the SC. This suggests that they are a source of output from the SC, a role that is likely to involve saccade-related activity.

3. Receptive Field Registry and/or Correspondence

When sensory responses occur, the existence and location of the receptive field(s) become an important issue. Before delving further into this topic, it is important to define our terminology. At issue is the question of the relative locations of visual, auditory, and tactile receptive fields of individual SC neurons. We define the term "spatial registry" to mean that an individual SC neuron responds to stimuli of these different modalities at the same location in space. For example, does the cell respond to the sight, sound, and touch of a mosquito on the back of the hand? If so, then the visual, auditory, and tactile receptive fields may be said to be in register. In contrast, we define "spatial correspondence" to mean that the unimodal receptive fields of multimodal neurons bear a systematic relationship to one another but are not necessarily in truly the same location in space. For example, if most neurons with tactile receptive fields on the hindquarters (outside the field of view) had visual receptive field in the lower eccentric region of the visual scene, the tactile and visual receptive fields would clearly be correlated in a nonrandom way, but because they are not actually in the same position in space they would not be in true register.

In anesthetized cats, visual and auditory receptive fields are usually in spatial register. Alex Meredith and Barry Stein compared the visual and auditory receptive field locations in multimodal neurons and reported that the area of spatial overlap averaged 86% of the smaller receptive field Somatosensory receptive fields show a spatial correspondence with visual or auditory receptive fields: Tactile receptive fields on the face tend to be correlated with central visual receptive fields, whereas tactile receptive fields on the flank or hindquarters tend to cooccur with more peripheral visual receptive fields (Fig. 2). These representations are topographically organized, creating maps of the

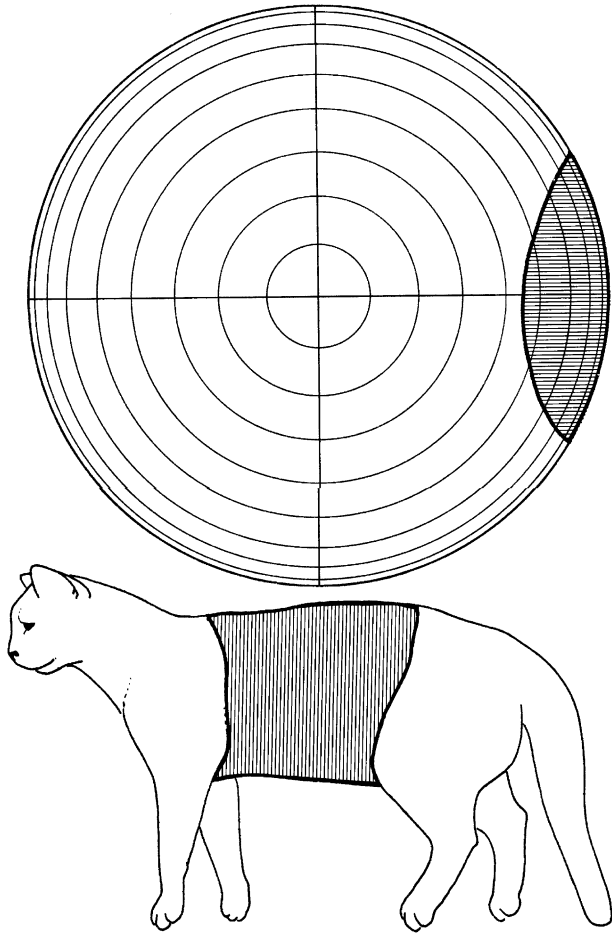


Figure 2 Spatial correspondence between the visual and tactile receptive fields in cat SC. The shaded regions indicate the receptive field locations. The visual (top) and tactile (bottom) receptive fields are closely related to one another but not in register because the body surface containing the tactile receptive field is actually out of the animal's field of view [courtesy of M. A. Meredith and B. E. Stein].

visual, auditory, and tactual scenes. In the rostral part, the auditory and visual receptive fields are centrally located, and the somatosensory receptive fields are on the face or forelimb. Neurons located more caudally have receptive fields located more peripherally in the contralateral space, with tactile receptive fields on the back, hindlegs, or tail. The sensory maps also generally correspond with the motor map for eye movements.

4. Frames of Reference

A priori, one might have expected that registry between receptive fields would be quite poor, due to the fact that the different sensory systems employ different frames of reference. Thus, a visual receptive field anchored to the retina and an auditory receptive field anchored to the position of the head and ears would only match up if the eyes happened to be in the one and only position that would result in alignment. The convergence of visual, auditory, and tactile information onto the SC therefore prompted investigations into the frame of reference of these signals.

Martha Jay and David Sparks were the first to test the influence of eye position on collicular auditory receptive fields. They trained monkeys to make saccades to the locations of auditory stimuli from different initial eye positions. The animals' heads were held fixed so that eye movements were dissociated from head movements. Jay and Sparks found that many of the auditory receptive fields shifted when the eyes moved (Fig. 3), a result that has been confirmed in subsequent studies involving awake cats. Although in principle this should permit perfect registry between visual and auditory receptive fields to be maintained when the eyes move, the alignment of visual and auditory receptive fields was not tested in this experiment. Paradoxically, their findings may actually cast doubt on the possibility of perfect alignment: The receptive fields did not shift as much as they should have to maintain a perfect eye-centered representation of the location of sounds. Instead, the receptive fields shifted an average of only about half the amount of the eye movement, which should disrupt the registry with visual receptive fields. Thus, the alignment of visual and auditory receptive fields in the SC of the awake primate is deserving of further investigation to resolve this apparent conundrum.

An investigation into the frame of reference of tactile signals in the SC was conducted by Jennifer Groh and David Sparks. Monkeys were trained to make saccades from a variety of different initial eye positions to the locations of vibrotactile stimuli delivered to the palms of the hands. The locations of the somatosensory stimuli on the body surface were held constant, as was limb position. Under these conditions, the positions of the eyes influenced the responses of many SC neurons to tactile stimuli. The effect of eye position was suggestive of an eye-centered frame of reference: Cells tended to respond best to the tactile stimulus when the initial position of the eyes brought the location of the tactile stimulus into the visual receptive field or movement field of the neuron under study.

Although the frame of reference of visual and saccade-related activity in the SC is indeed eye

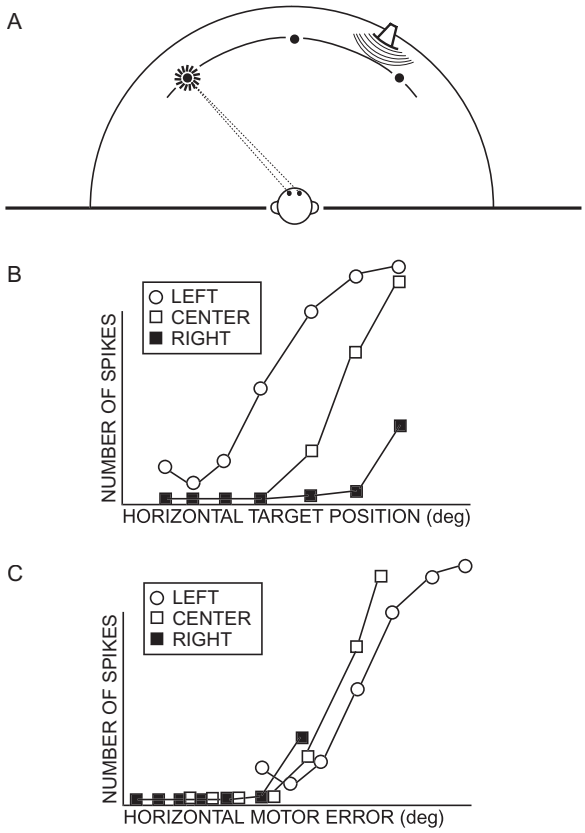


Figure 3 Effect of eye position on auditory responses of a neuron in the SC. (A) Experimental design. Monkeys fixated one of three LEDs while sounds were presented from a range of locations using a speaker mounted on a rotating hoop. Neural responses were better correlated with the location of the sound with respect to the eyes, or motor error (C), than with the location of the sound in space (B) [adapted with permission from Jay, M. F., and Sparks, D. L. (1984). Auditory receptive fields in primate superior colliculus shift with changes in eye position. *Nature* **309,** 345–347].

centered, it is more properly described as a motor error or updating eye-centered frame of reference than a strictly retinal frame of reference. This was demonstrated in 1980 by Larry Mays and David Sparks. They recorded from neurons in the SC that were visually responsive in an ordinary visual saccade task. They then asked monkeys to make saccades to a sequence of two briefly flashed visual stimuli. Both stimuli were illuminated and then extinguished before the first saccade began (Fig. 4). Thus, the direction and amplitude of the saccade to the second target were dissociated from the original retinal locus of that target. Monkeys can perform this task quite well.

If the saccade vector between the first and second targets corresponded to the location of the receptive field, then many cells discharged between the first and second saccades, even though the retinal locus that would normally be associated with that saccade vector had never been activated. Thus, although these cells respond to visual stimuli, in fact they can also represent the remembered locations of visual stimuli in a frame of reference anchored to the eyes—a frame of reference that is updated when the eyes moved. These characteristics led them to be called quasi-visual cells.

5. Multisensory Integration

If a cell responds to both visual and auditory stimuli separately, what happens when the two stimuli are presented together? One might expect individual neurons to treat stimulus energy as equivalent, regardless of the source of the energy. Thus, the response to a joint stimulus might be approximately equal to the sum of the responses to the individual components. In

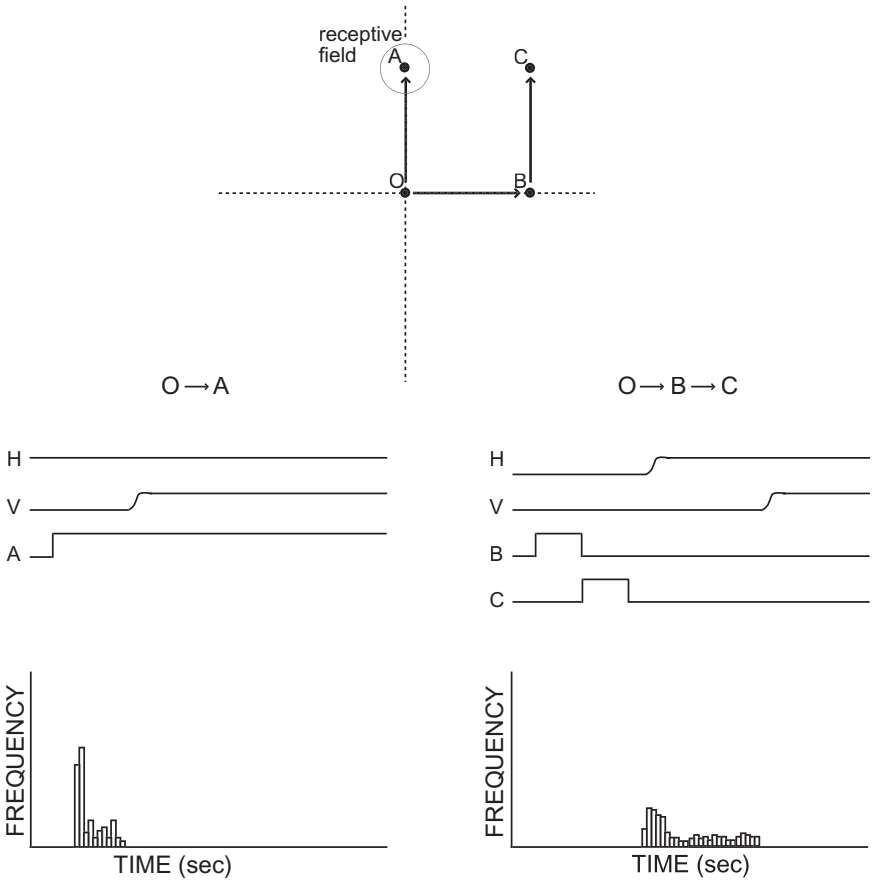


Figure 4 Idealized response of a quasi-visual cell in the SC. (Top) The locations of receptive fields and various visual stimuli. (Bottom left) The response of the cell when the visual stimulus is presented in the receptive field and the monkey makes a saccade directly to it. (Bottom right) The response of the cell when no stimulus is presented in the retinal location corresponding to the receptive field. Rather, a sequence of two stimuli is presented. After the animal makes a saccade to the first, the remembered location of the second lies in the new position of the receptive field and the cell becomes active [adapted with permission from Mays, L. E., and Sparks, D. L. (1980). Dissociation of visual and saccade-related responses in superior colliculus neurons. *J. Neurophysiol.* **43,** 207–232].

other words, SC neurons might behave linearly. Violations of linearity should occur for weak stimuli near the threshold of the neuron, where the combined response should exceed the sum of the two responses, and again for strong stimuli, when responses to combined stimuli might saturate at a level below the sum of the unimodal responses.

An alternative means of characterizing multisensory interactions has been employed by Meredith, Stein, and colleagues. These researchers emphasized the potential benefit for stimulus detection that may be provided by combined sensory stimuli. They therefore compared the joint response to the larger of the two unimodal responses using the following formula:

$$\% \text{ Interaction} = \frac{(\text{joint response} - \text{maximal unimodal response}) \cdot 100}{\text{maximal unimodal response}}$$

Response enhancement is defined as a joint response exceeding the maximal unimodal response, and depression is defined as a joint response that is less than the maximal unimodal response. If the joint response equals the maximal unimodal response, this is defined as "no interaction." Comparing this classification scheme with linearity, any interaction that exceeds 100% must be supralinear: If the joint response is more than twice as large as the largest unimodal response, it must also be larger than the sum of the two unimodal responses. Similarly, any interaction that is less than 0% must be sublinear: If the joint response is less than the maximal unimodal response, it must also be less than the sum of the two unimodal responses (assuming the weaker response is not inhibitory). However, between these two extremes a comparison between the percentage interaction and linearity cannot be made without knowledge of the individual component responses.

When Meredith, Stein, *et al.* characterized the multimodal responses in terms of percentage interaction, they found evidence of all three types of effect: enhancement, depression, and no interaction. The type of interactions depended on three factors: the effectiveness of the component stimuli at driving the cell, the spatial locations of the component stimuli, and the timing between the component stimuli. Response enhancement tended to be largest when the component stimuli evoked only modest responses on their own. Enhancement tended to occur when both stimuli were in the center of their respective receptive fields, whereas depression was more likely to occur when one of the two stimuli was in the suppressive area of its receptive field. The degree of enhancement or depression was largest when the stimuli were presented within the same time frame (with less than 100 msec separating the two). Specifically, stimulus-onset asynchronies small enough that the responses to the individual components would overlap in time tended to produce maximal enhancement. Enhancement decreased with greater stimulus-onset asynchrony, and when asynchrony was increased even further depression often occurred.

Both methods of characterizing multisensory interactions imply certain assumptions. As mentioned previously, comparing the joint responses to the sum of the individual responses suggests a null hypotheses that stimulus energy is stimulus energy, regardless of which sensory system detected the event. Two stimuli have more energy than one, so the response to two should be close to the sums of the individual responses (i.e., linear). In contrast, the percentage interaction (enhancement/depression) analysis emphasizes the relationship between the combined response and the stronger unimodal response. When the two are equal, this is defined as no interaction. However, failure of the joint response to exceed the maximal unimodal response might actually indicate that the stronger unimodal stimulus suppresses any response to the weaker unimodal stimulus. This would hardly be an inconsequential interaction between the sensory signals. Large enhancements (those that exceed linearity) clearly indicate synergistic facilitation between sensory inputs, but anything less than a large enhancement may indicate equally interesting competition between the different types of inputs to individual neurons. Further work is needed using both types of analysis to clarify these issues.

B. Frontal Eye Field

The FEF is a cortical region that is involved in controlling saccadic eye movements. The FEF is similar to the SC in many ways. In awake monkeys, many FEF neurons discharge in conjunction with saccades. Cells have movement fields and discharge most vigorously if the saccade will be directed to the center of the movement field. The idealized FEF neuron has an activity profile with sensory and motor components similar to the idealized SC neuron shown in Fig. 1. Microstimulation in FEF evokes saccades. As mentioned earlier, combined lesions of FEF and SC completely abolish saccades.

Like the SC, the FEF has been implicated in the control of saccades to sounds as well as visual stimuli.

Gary Russo and Charles Bruce investigated the motor-related activity of neurons in the FEF to determine if they were active for saccades to auditory stimuli as well as visual stimuli. Monkeys were trained to make saccades to the locations of both lights and sounds. They found that all the cells in their sample responded prior to saccades to both kinds of targets. Furthermore, the movement fields for both were anchored to the initial position of the eyes. These results, were similar to those of Jay and Sparks for primate SC.

Neurons in FEF also respond to sensory events. Many respond to visual stimuli and a few respond to sounds. Neurons that respond to sounds can be driven by a variety of different types of sound stimuli, and receptive fields tend to lie on the contralateral side. About half of auditory neurons also respond to visual stimuli. Acoustically responsive cells tend to be found primarily at locations in FEF that represent large amplitude movements. Beyond this, little is known. The frame of reference of these sensory responses has not been examined, nor have the interactions between visual and auditory stimuli been characterized. Do bimodal cells have receptive fields that are in register with one another? Do they stay in register when the eyes move? Do the cells show response enhancement/depression to combined stimuli? How do the responses to combined modality stimuli compare with the linear sum of the unimodal responses? Further experiments are needed to answer these questions.

C. Lateral Intraparietal Cortex

Posterior parietal cortex is a region of the brain that has been implicated in spatial attention and eye movements. In humans, lesions of the parietal lobe cause patients to ignore sensory stimuli located in the contralateral half of space, a phenomenon known as neglect. What is the frame of reference of this neglect? One intriguing experiment that addressed this question was conducted by Edoardo Bisiach and Claudio Luzzatti in 1978. They asked patients with parietal lobe lesions to imagine themselves facing the cathedral at the Piazza del Duomo in Milan and report all the landmarks they could recall. From this imaginary vantage point, the patients named primarily landmarks on the side of the piazza ipsilateral to the side of their brain lesion. Then, they were asked to imagine themselves standing on the cathedral steps, facing in the opposite direction. From this opposite perspective, the deficit was reversed. They could now recall landmarks on the opposite side of the piazza, but they failed to name landmarks on the first side—landmarks that they had been able to recall from the first imaginary vantage point.

In monkeys, the lateral intraparietal region of posterior parietal cortex, or area LIP, is believed to play a role in guiding saccades to the locations of visual, auditory, and remembered stimuli. Various groups have debated whether the function of LIP is better characterized as attentional, overseeing the sensory processing of stimuli at a particular location, or "intentional," preparing the motor circuitry to make a saccadic eye movement to a stimulus at that location. Clearly, these two concepts are closely related, perhaps so much so that no experiment could adequately distinguish between them. We therefore remain agnostic on this point and will merely review the relevant literature concerning the representations of visual, auditory, and saccade-related information in LIP.

Neurons in LIP resemble those in both SC and FEF in having both sensory and motor-related activity. For many neurons in LIP, the visual responses are not strictly dependent on the location of the stimulus on the retina. The position of the eye is also important. Richard Andersen, Vernon Mountcastle, and colleagues found that visual stimuli at a given retinal location evoked different responses depending on where the animal's eyes were pointing (Fig. 5). The effects of eye position were dubbed "gain fields" because the level of the visual response was affected by orbital position. Although the term gain normally implies a multiplicative interaction, whether this eye position effect is truly multiplicative has not been studied. Nevertheless, the location of the receptive field stays anchored to the retina when the eyes move, but the responsiveness of the cell varies. Thus, the frame of reference of visual activity is not strictly retinal but reflects the combination of retinal and eye position signals.

Larry Snyder, Peter Brotchie, Richard Andersen, and colleagues further investigated these gain fields to determine whether the position of the eyes with respect to the head, body, or world was the important factor. They varied the positions of the eyes with respect to the head, the position of the head with respect to the body, and the position of the body with respect to the world while keeping the visual stimulus at a fixed position on the retina. These manipulations revealed that the magnitude of the response to a given retinal stimulus depended on the position of the eyes with respect to the body. In LIP, it mattered little whether gaze shifts with respect to the body were accomplished largely by

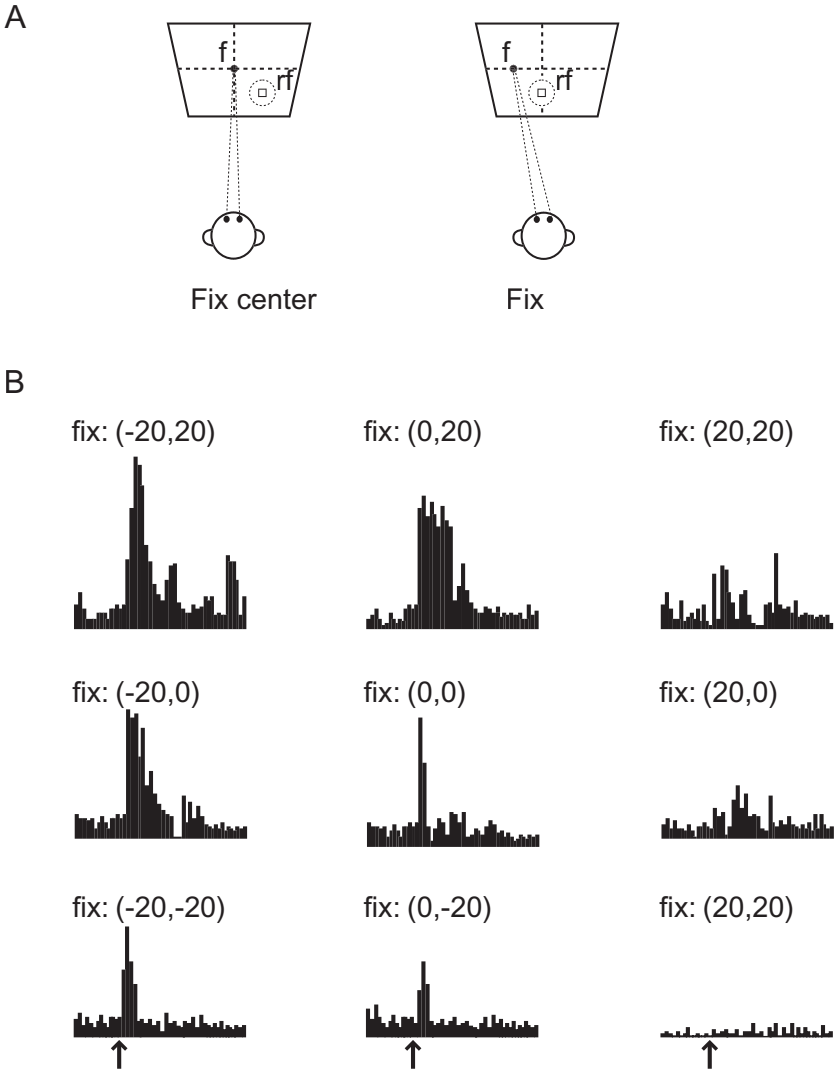


Figure 5 Effects of eye position on visual responses in posterior parietal cortex. (A) Schematic of experimental design. The responses of neurons to visual stimuli at a particular retinal location were examined for a range of orbital positions. (B) Responses varied with eye position, although retinal location was held constant [adapted with permission from Andersen, R. A., Essick, G. K., and Siegel, R. M. (1985). Encoding of spatial location by posterior parietal neurons. *Science* **230,** 456–458].

eye-in-head rotations or head-on-body rotations. (In nearby area 7a, however, more cells were affected by the position of the eyes with respect to the world than by the position of the eyes with respect to the body.)

What about auditory signals in LIP? Pietro Mazzoni, Richard Andersen, and colleagues examined how LIP represents sound location. They presented sounds from two speakers, one located contralaterally and the other in the ipsilateral field. They found that approximately half of neurons in their population responded to sounds at one or both of these locations. Most responded more vigorously from one speaker than from the other. Most neurons also responded to visual stimuli and tended to prefer visual stimuli located near the preferred speaker. However, receptive fields were not mapped in this study, so it is not known how extensive the visual and auditory receptive fields are, nor how closely they correspond with one another.

The frame of reference of auditory signals was examined by Brigitte Stricanne, Richard Andersen, and Pietro Mazzoni. These researchers trained monkeys to make saccades to the locations of five speakers separated 12 degrees apart along the horizontal meridian. Three fixation positions 12 degrees apart

were used. An analysis of variance confirmed that the location of the sound was an important determinant of neural activity. The location of the sound with respect to the head, the location of the sound with respect to the eyes, or both of these factors were predictive of how cells would respond. However, if receptive fields are large relative to the separation between fixation positions, it would be difficult to determine from this analysis whether the activity is better correlated with head vs eye-centered frames of reference because the head- and eye-centered locations are correlated with one another. Nevertheless, this study does suggest that eye position influences the responses of at least a subset of auditory neurons in LIP.

To summarize, LIP contains neurons that respond to auditory stimuli, but many important questions remain unanswered. What is the relationship between visual and auditory receptive fields? Do they overlap? Are they in register? What is the frame of reference? Furthermore, nothing is known about how the cells respond to combined visual and auditory stimuli. Do the cells sum their inputs linearly? Do they show the kind of response enhancement/depression that is found in the SC? These issues await further inquiry.

D. Anterior Ectosylvian Sulcus

In cat, a major source of descending input to the SC derives from the AES. This descending input is of particular interest because this cortical region (whose primate homolog is unknown) contains subdivisions for different sensory modalities. Visual responses are present in area AEV, somatosensory activity occurs in area SIV (the so-called fourth somatosensory cortex, pronounced "S4"), and auditory signals are found in field AES. At the periphery of these unimodal subdivisions are multimodality neurons that do not project directly to the SC. Overall, approximately one-third of the cells are auditory, one-fourth somatosensory, and one-fifth visual. The remaining cells are bimodal or, rarely, trimodal, with visual–auditory cells being most common in this group.

Lee Wilkinson and colleagues investigated the functional role of AES in integrating multisensory information by deactivating it and testing the performance of cats on a task involving multisensory cues. The cats were trained to orient toward a visual target. When the visual target was dim enough to be near threshold for detection, orientation performance could be improved by pairing the visual stimulus with a sound at the same location. Pharmacological deactivation of the AES prevented this facilitation by the paired auditory cue. Orientation to a visual stimulus alone was not affected by the deactivation. Inactivation of primary auditory cortex had no such effect. These results suggest that area AES plays an important role in facilitating orienting movements based on auditory information.

Inactivation of AES has also been shown to affect the responses of SC neurons in an interesting way. When AES is cooled, the responses of SC neurons to unimodal stimuli are unaffected. However, the responses to combined stimuli cease to show the same type of interaction (enhancement or depression) that they showed prior to the inactivation. This result is particularly surprising given that the projection from AES to SC derives from the unimodal cells, not the multimodal cells in that region.

Although they do not project to the SC, the multimodal cells in AES do share many of the properties of multimodal cells in SC. Receptive field locations for the different sensory modalities show a correspondence in space. Visual–somatosensory cells with tactile receptive fields on the forelimbs tend to have visual receptive fields in the lower quadrant of space and so forth. This correspondence is surprising because the unimodal subdivisions of the AES region do not seem to have a topographic organization. As in the SC, multimodal cells show response enhancement when the relevant stimuli are presented simultaneously in their respective receptive fields. How their responses to combined stimuli compare with the linear sum of the responses to the individual components has not been tested.

Our knowledge of area AES is currently quite preliminary. A number of questions remain to be answered. How do these cells respond in the awake, behaving animal? What is the frame of reference of these signals? Are the receptive fields of multimodal cells in true spatial register, and how is the registry influenced by movement of the eyes, ears, head, and body? What is the primate homolog? Experiments in awake monkeys are needed to address these issues.

IV. ALGORITHMS FOR COORDINATE TRANSFORMATIONS

Despite the fact that many of these areas show signs that auditory signals have been translated into an eye-centered reference frame, little is known about how the brain does this. Two models for how this may be accomplished have been proposed by Groh and Sparks.

A. Computation

Transforming signals from one frame of reference into another is mathematically straightforward. If the location of an auditory target with respect to the head ($\boldsymbol{A}_h$) is known, and the position of the eyes with respect to the head ($\boldsymbol{E}_h$) is also known, then subtracting the eye position signal $\boldsymbol{E}_h$ from the head-centered target location signal $\boldsymbol{A}_h$ will yield a signal of the location of the sound with respect to the eyes ($\boldsymbol{A}_e$) (Fig. 6A). The first model, the vector subtraction model (Fig. 6B), implements an algorithm resembling this mathematical calculation using neurons as the computational elements. The second model, the dendrite model (Fig. 6C), uses a different algorithm: a neural analog of a multidimensional lookup table, which is implemented through local circuitry at the dendrites of the units in the eye-centered auditory map.

B. Input and Output Signals

As far as possible, the component signals are provided by units that resemble known populations of neurons. Eye position signals are carried by neurons from a variety of brain areas, such as the brain stem gaze centers, nucleus prepositus hypoglossi, the flocculus of the cerebellum, and posterior parietal cortex. Each of these areas contains neurons whose firing rate is typically linearly related to eye position. This information is most likely derived from a copy of the motor command to move the eyes, although proprioceptive feedback from the extraocular muscle spindles is also possible. No evidence of a place code of eye position, with neurons having response fields for eye position, has been reported. Thus, the eye position units in the model signal the position of the eyes in the orbits in a linear or rate-coded fashion.

For lack of any evidence to the contrary, the representation of the inputs to the model is assumed to be an auditory map of space in a head-centered frame of reference. Units in this map have receptive fields defined with respect to the head. However, convincing evidence for an auditory map of space in a head-centered frame of reference has not been reported—the frame of reference has either not been studied (classical auditory areas such as IC and auditory cortex) or the results have indicated a possible eye-centered frame of reference (SC, FEF, and LIP). Thus, how the inputs to this transformation are actually encoded is an open question.

The output of both of the models is an eye-centered auditory map resembling the primate SC. These units have receptive fields in an eye-centered frame of reference.

C. Neural Architecture

The circuitry of the vector subtraction model is illustrated in Fig. 6B. Neurons in the head-centered auditory map project to a set of intermediate units with a synaptic weighting that is proportional to their receptive field location. Thus, the level of excitation received by these intermediate units will be a linear rate code of the location of the sound with respect to the head. These units also receive an inhibitory synapse from a unit encoding eye position using a linear rate code. This inhibitory connection is the mechanism by which the eye position signal is subtracted from the signal of head-centered target location. The resultant signal encodes the location of the auditory target with respect to the eyes, again as a linear rate code.

The remainder of the circuitry in the model is devoted to converting this rate code into a place code in which neurons have receptive fields in an eye-centered frame of reference, like the map of oculocentric auditory space in the SC. This is accomplished using graded thresholds and inhibitory interneurons.

The simplicity of this mechanism is not without cost: Because of the rate coding at the intermediate stage of processing, the model functions properly for single auditory targets only. In order for an animal to make accurate saccades to the location of one auditory target among distracters, a target selection mechanism must choose the target before the coordinate transformation is executed. If two targets were present, and both were selected as saccade targets, the model as originally proposed would yield the vector sum of two auditory targets as an output. A small modification to this model would allow it to compute the vector average instead of the sum. Vector averaging is a more realistic response to multiple inputs and has been found in a number of other behavioral and physiological experiments. We therefore consider vector averaging of two auditory targets to be more likely than vector summation. Nevertheless, the vector subtraction model is simply not capable of preserving both target locations independently.

The alternative model, the dendrite model (Fig. 6C), can handle multiple targets because there is no rate-coding bottleneck at an intermediate stage of processing. Instead, every unit in the input map projects

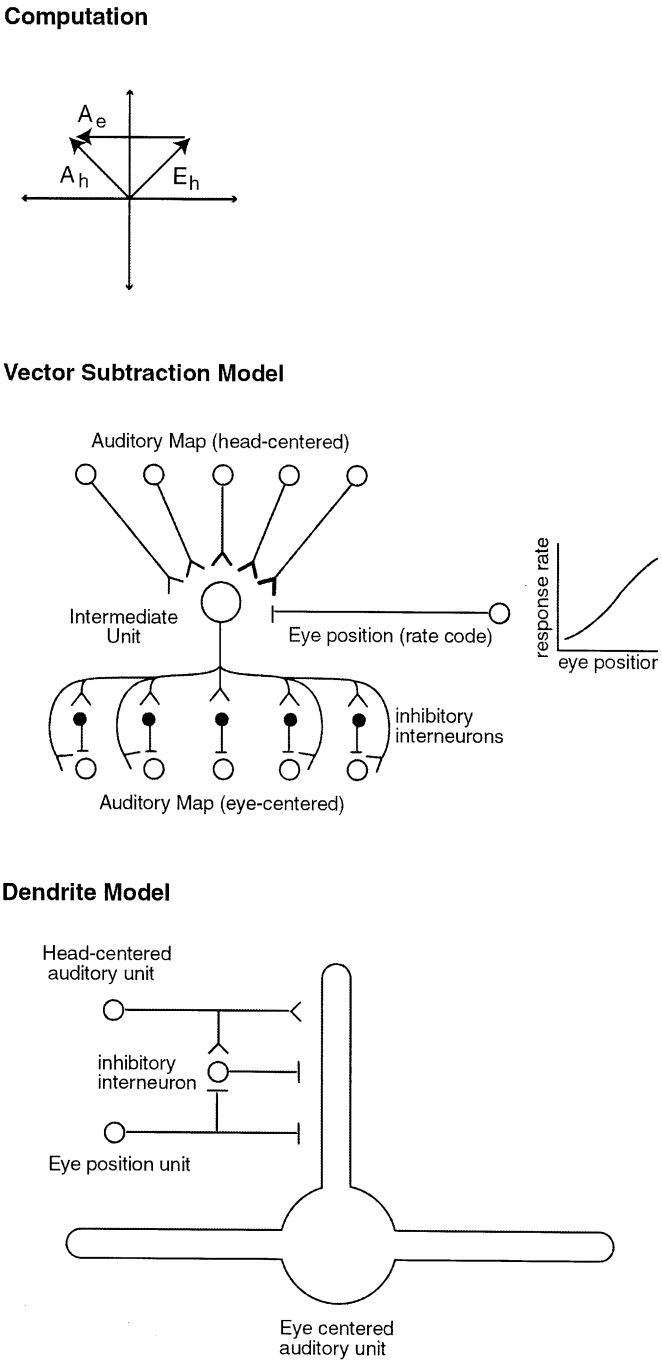


Figure 6 Computation of the eye-centered location of a sound. (A) If the auditory stimulus is located above and to the left of the head ($\boldsymbol{A}_h$), and the eyes are directed upward and rightward in the orbit ($\boldsymbol{E}_h$) as shown here, then the auditory stimulus is located directly to the left of the eyes ($\boldsymbol{A}_e$). (B) The vector subtraction model. The head-centered auditory map projects to the intermediate unit with graded synaptic weights. The eye position signal is subtracted at an inhibitory synapse. The resulting rate code for the target's location with respect to the eyes is converted into a place code for auditory space in an eye-centered frame of reference through graded thresholds and inhibitory interneurons. (C) The dendrite model. Each dendrite receives input from eye position units and one head-centered auditory unit. Thresholds and synaptic weights are balanced so that the cell body receives net excitation if an auditory stimulus is present and the eyes are within a certain range of positions. Each dendrite performs this analysis independently for a particular location in head-centered space. *In toto* the dendrites of an individual neuron create an eye-centered receptive field [for more details, see Groh, J. M., and Sparks, D. L. (1992). Two models for transforming auditory signals from head-centered to eye-centered coordinates. *Biol. Cybernetics* **67,** 291–302].

directly onto separate individual dendrites of every unit in the output map. The eye position units also project directly onto each dendrite. There are also interneurons associated with each dendrite. The local circuitry at each dendrite is set up to activate the unit if (i) there is a sound in the receptive field of the head-centered unit that projects to that dendrite and (ii) the position of the eyes places that sound within the eye-centered receptive field of that unit. Other sounds that are not in the eye-centered receptive field will not activate that particular unit, but they will activate other units in the eye-centered map via the projections to individual dendrites on those units. Because these connections are all independent, the model can accurately transform the locations of an unlimited number of targets.

These are two possible mechanisms whereby auditory signals can be translated from head- to eye-centered coordinates. Other algorithms are undoubtedly also possible. Furthermore, these algorithms are not limited to auditory coordinate transformations nor to translating information from head coordinates into eye coordinates. Translation of visual information from retinal to updated eye-centered coordinates or head-centered coordinates could be accomplished using similar mechanisms.

V. PARALLELS WITH OTHER MULTISENSORY MOTOR SYSTEMS

Thus far, we have concentrated primarily on brain areas implicated in multisensory processing and the control of saccades to sensory stimuli. However, recent work in premotor cortex has revealed that many of the same principles apply for the neural control of skeletal movements guided by sensory stimuli of different modalities. Neurons in premotor cortex can respond to visual or somatosensory stimuli or both. Auditory responses also exist but have been less extensively explored.

Michael Graziano, Charles Gross, and colleagues explored the issues of spatial registry/correspondence and the frames of reference of these signals. They found that bimodal visual/somatosensory neurons show an interesting spatial correspondence, with visual receptive fields tending to occupy a circumscribed three-dimensional volume of space in the immediate vicinity of the region of the body surface that contains the tactile receptive field. Furthermore, the frame of reference appears to be body-part centered. Neurons with tactile receptive fields on the limb and visual receptive fields in nearby space employ a limb-centered frame of reference. The visual receptive fields move in space when the limb moves. If the eyes move but the limb does not, the visual receptive field remains in its original location in space, despite the fact that a different region of the retina would not be activated by stimuli at that location. Thus, the retinal location of a stimulus is irrelevant to these neurons; only the location of the stimulus with respect to the limb is important. Similarly, when the visual receptive fields are located near the face or head, they move when the head moves, even if gaze position in space is constant. In short, like the areas of the oculomotor pathway, the skeletal motor pathway appears to employ frames of reference that are tailored for guiding the movements of individual body parts, regardless of the modality of the stimulus that evokes the movement.

How these visual signals might be translated into a body-part-centered frame of reference is not known. Mechanisms similar to those outlined in the previous section for the translation of auditory signals from head- to eye-centered coordinates could apply in this case as well. For example, the translation of visual signals from eye- to head-centered coordinates is simply the inverse of the auditory transformation and could be accomplished using a retinotopic map as the input. The vector subtraction model would become the vector addition model, with the eye-in-head signal being added to the retinal vector to produce a signal of visual stimulus location with respect to the head. Similar modifications to the dendrite model could also be made. Translation of visual signals from a retinal- to a limb-centered frame of reference is obviously more complicated, but even this computation could be accomplished in an analogous fashion if the brain were to first compute the position of the limb with respect to the eyes. Replacing the eye-in-head signal with an eye-with-respect-to-limb signal in the models might allow the computation to proceed efficiently from that point, without requiring a series of intervening coordinate transformations. Of course, computing such a signal might not be easy, given how many joint angles would be involved.

VI. CONCLUSION

Our brains receive a constant barrage of sensory information from our different sense organs. Making sense of this onslaught requires determining which visual, auditory, somatosensory, olfactory, and

gustatory cues match up with one another and derive from a common event. The multisensory neurons in the brain are likely to play a critical role in this process. Our review of multisensory activity within the saccade pathway reveals evidence for common frames of reference and spatial registry and/or correspondence across sensory modalities. However, we still have much to learn about these cells. Detailed receptive field mapping and more extensive quantitative analysis of the responses to combined modality stimuli in awake animals are needed. Such studies will serve as the foundation on which a bridge between multisensory neural activity and behavior can ultimately be built.

See Also the Following Articles

AUDITORY PERCEPTION • HAND MOVEMENTS • MOTOR SKILL • MULTISENSORY INTEGRATION • OBJECT PERCEPTION • OLFACTION • RECEPTIVE FIELD • TACTILE PERCEPTION • TASTE • VISION: BRAIN MECHANISMS

Suggested Reading

Andersen, R. A., and Mountcastle, V. B. (1983). The influence of the angle of gaze upon the excitability of the light-sensitive neurons of the posterior parietal cortex. *J. Neurosci.* **3,** 532–548.

Bisiach, E., and Luzzatti, C. (1978). Unilateral neglect of representational space. *Cortex* **14,** 129–133.

Brotchie, P. R., Andersen, R. A., Snyder, L. H., and Goodman, S. J. (1995). Head position signals used by parietal neurons to encode locations of visual stimuli. *Nature* **375,** 232–235.

Graziano, M. S. A., and Gross, C. G. (1998). Spatial maps for the control of movement. *Current Opin. Neurobiol.* **8,** 195–201.

Groh, J. M., and Sparks, D. L. (1992). Two models for transforming auditory signals from head-centered to eye-centered coordinates. *Biol. Cybern.* **67,** 291–302.

Groh, J. M., and Sparks, D. L. (1996). Saccades to somatosensory targets. III. Eye-position-dependent somatosensory activity in primate superior colliculus. *J. Neurophysiol.* **75,** 439–453.

Jay, M. F., and Sparks, D. L. (1984). Auditory receptive fields in primate superior colliculus shift with changes in eye position. *Nature* **309,** 345–347.

Mays, L. E., and Sparks, D. L. (1980). Dissociation of visual and saccade-related responses in superior colliculus neurons. *J. Neurophysiol.* **43,** 207–232.

Mazzoni, P., Bracewell, R. M., Barash, S., and Andersen, R. A. (1996). Spatially tuned auditory responses in area LIP of macaques performing delayed memory saccades to acoustic targets. *J. Neurophysiol.* **75,** 1233–1241.

Meredith, M. A., and Stein, B. E. (1996). Spatial determinants of multisensory integration in cat superior colliculus neurons. *J. Neurophysiol.* **75,** 1843–1857.

Russo, G. S., and Bruce, C. J. (1994). Frontal eye field activity preceding aurally guided saccades. *J. Neurophysiol.* **71**(3), 1250–1253.

Snyder, L. H., Grieve, K. L., Brotchie, P., and Andersen, R. A. (1998). Separate body- and world-referenced representations of visual space in parietal cortex. *Nature* **394,** 887–891.

Stein, B. E., Wallace, M. T., and Stanford, T. R. (1999). Development of multisensory integration: Transforming sensory input into motor output. *Mental Retardation Dev. Disabilities Res. Rev.* **5,** 72–85.

Stricanne, B., Andersen, R. A., and Mazzoni, P. (1996). Eye-centered, head-centered, and intermediate coding of remembered sound locations in area LIP. *J. Neurophysiol.* **76,** 2071–2076.

Wilkinson, L. K., Meredith, M. A., and Stein, B. E. (1996). The role of anterior ectosylvian cortex in cross-modality orientation and approach behavior. *Exp. Brain Res.* **112,** 1–10.

Visual Cortex

MARCELLO G. P. ROSA
Monash University

GLOSSARY

blindsight Refers to the visual abilities of patients with lesions of area V1. Despite having no visual sensation, they are capable of correctly guessing the location and some characteristics of visual stimuli presented in the blind field.

blobs Patches of cortical tissue in area V1, where the neuropil has a high content of the enzyme cytochrome oxidase. There are several thousand blobs in the striate cortex of monkeys and humans, each forming a column of cells 200–300 μm in diameter. Although blobs are most clearly defined in layer 3, they also include corresponding patches of layers 2, 5, and 6.

contextual modulation The changes in the responses of visual cortical cells induced by stimuli presented outside the receptive fields. Typically, the cellular responses are suppressed if a similar visual pattern is presented within the receptive field and in the background.

dorsal stream Group of visual areas located in dorsal occipital and parietal cortices, that are involved with the analysis of motion, spatial relationships, and the visual control of movement.

extrastriate cortex Designation that includes most visual areas in the occipital, parietal, temporal, and frontal lobes. Histologically, they lack many of the laminar subdivisions that characterize striate cortex. Extrastriate areas receive their main thalamic inputs from the pulvinar nucleus.

hypercolumn A group of adjacent cortical columns that includes cells selective to each possible variation of a physical parameter of visual stimulus.

striate cortex A histologically distinct area of the occipital lobe, that receives the bulk of the retinal projections to the cortex via the lateral geniculate nucleus. Also known as area V1.

ventral stream Group of visual areas located in ventral occipital and inferior temporal cortices, which are involved with the analysis of shape, color, and texture of objects.

visuotopic map Imaginary projection of the visual field into a cortical area. Visuotopic maps are created by the spatially ordered connections between the retinae, thalamus, and cortex.

visuotopic reorganization Change in the visuotopic map of a cortical area following a change in the pattern of afferent projections to the cortex; occurs after retinal detachments or retinal lesions.

The visual cortex is responsible for the rich ambit of sensations that form our visual experience. Damage to the visual cortex can be sufficient to cause blindness, even though the eyes are unaffected. Moreover, damage to different areas of the visual cortex can result in the disruption of specific aspects of vision, such as the recognition of shapes, the perception of colors, the sensation of motion, or the determination of the positions of different objects in relation to a person's body. Each of these aspects of vision represents an enormous computational problem, which contemporary artificial systems can only handle to a very limited extent. However, every moment the visual cortex solves these problems in a near effortless manner, and integrates the different solutions into a unified percept of the visual world. Understanding how this is accomplished is a major challenge for contemporary neuroscientists.

The rationale for partitioning the cortex into areas is the belief that each of these subdivisions performs a certain set of neuronal computations, which translate into a unique functional contribution for perception, motor control, or cognition. In reality, a good understanding of the likely role of a visual area in perception usually comes long after its existence is proposed on the basis of nonfunctional criteria, such as histological distinctions or anatomical connections. In addition, there is still much debate regarding the exact number

and limits of visual areas, even in intensively studied animal models like the monkey or the cat. Nonetheless, it is clear that one of the main trends in the evolution of the primate brain has been the growth and diversification of the visual cortical areas. Indeed, humans are highly visual animals in comparison with most other mammals, and it can be argued that our sophisticated visual cortex has been a prerequisite for the evolution of the higher cognitive functions that characterize "humanity" as we know it.

I. CHARACTERIZATION OF VISUAL AREAS

A. Striate and Extrastriate Cortices

The number, shape, and location of visual areas vary enormously among species. However, a useful distinction, which applies to the visual cortex of all mammals, is that between striate and extrastriate areas. The striate cortex (also known as the primary visual area or V1) is the largest area of visual cortex and somewhat different from the other visual areas in terms of its histological structure. In addition, there are multiple extrastriate areas, which in humans and other primates include the remainder of the occipital lobe plus subdivisions of the temporal, parietal and frontal lobes (Fig. 1). The histological and anatomical distinction between extrastriate areas can be very subtle. With few exceptions, these areas are difficult to distinguish from each other without physiological mapping techniques, such as single cell recordings or functional magnetic resonance imaging. Striate cortex is also unique in being the principal cortical target of axonal projections from the lateral geniculate nucleus and thus the first cortical area to receive the information coming from the eyes. The thalamic innervation of extrastriate areas, on the other hand, originates mainly from the nuclei of the pulvinar complex, which receives minor retinal projections.

B. Evolution of Visual Cortex

The contemporary representatives of early branches of the mammalian radiation, such as marsupials, have very few visual areas (Fig. 1). Besides V1, only one other visual area (named the second visual area or V2) has been identified with certainty in these animals. In addition, other regions of cortex that receive anatomical projections from V1 and V2 have been identified as putative visual areas, but their exact characteristics remain unclear. They may represent sites of polysensory convergence rather than purely visual areas. Some marsupial species, such as the Australian quoll (*Dasyurus*) and the South American wooly opossum (*Caluromis*), have enormously expanded visual cortices that occupy more than half of their entire neocortex. However, this is accomplished by the enlargement of V1 and V2 rather than by the addition of new areas. This means that there are relatively few stages of visual processing in the cortex, which probably limits the range of visual sensations and visually guided behaviors in these species. From such a simple organization, expansion and addition of extrastriate areas have occurred in many branches of the mammalian radiation (Fig. 1). Among placental mammals, even species with poor vision such as the rat have additional visual areas beyond V1 and V2. Placental mammals with developed vision, such as cats and primates, may have more than 20 extrastriate areas in addition to V2, although the exact number is not known for any species.

In the course of primate evolution, there has been an enormous expansion of the anterior portions of extrastriate cortex (the *inferior temporal* and *posterior parietal* cortices). Some of the areas in these regions appear to exist only in primates. Although the study of the subdivisions of human visual cortex is still in its infancy, the evidence so far indicates that the visual areas demonstrated in monkeys also exist in humans. In addition, recent studies have suggested the existence of evolutionarily new areas, which are probably specific to higher primates. For example, there is a "kinetic occipital" area that is involved in the integration of motion and shape signals; this area seems to have no counterpart in the monkey. Moreover, in the monkey brain cortical areas devoted to different sensory modalities may be located in close apposition, whereas in the human brain the visual, somatosensory, and auditory cortices have become separated by larger expanses of "association" or polysensory areas, some of which receive afferents from visual cortex.

C. Areas and Modules

In addition to the growth and multiplication of extrastriate cortex, another trend in the evolution of complex visual cortices has been the subdivision of some areas, including V1 and V2, into modules in which neurons differ in their patterns of connections with other brain areas, physiological characteristics, and neurochemical composition. These modules

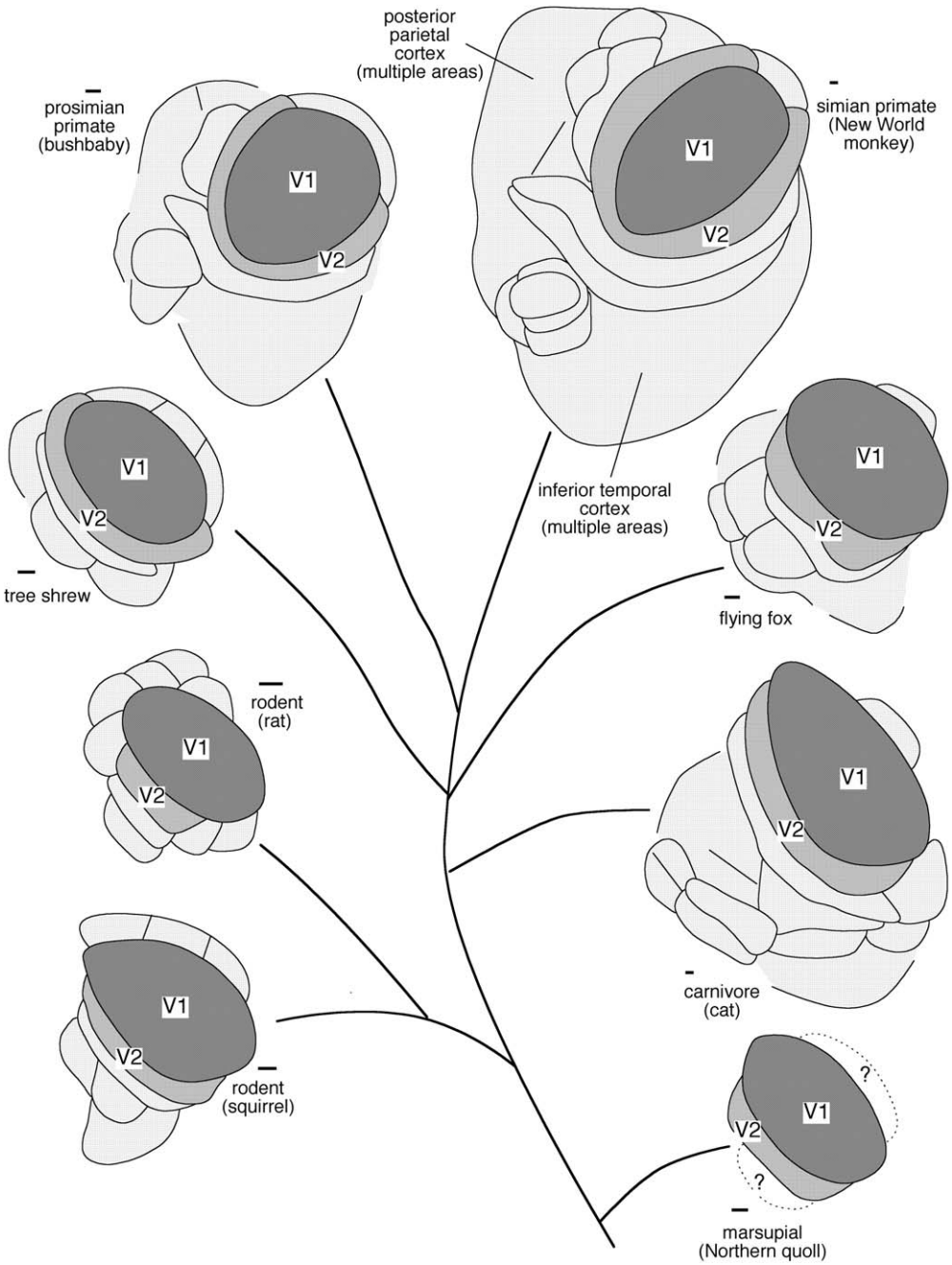


Figure 1 Expansion of extrastriate cortex in mammalian evolution. Each diagram is a schematic view of "unfolded" visual cortex, with striate cortex (V1) shown in dark gray, the second visual area (V2) in medium gray, and other visual areas in light gray. The scale bars, adjacent to the names of the species, correspond to 1 mm.

usually conform to the columnar organization of the cortex, spanning all cortical layers. Although modules are present in mammals from different orders, the most clear-cut examples are found in primates. For example, in V1 of monkeys (Fig. 2) some columns of cells are characterized by a high concentration of neuropil that stains densely for the mitochondrial enzyme cytochrome oxidase. These columns aggregate into modules a few hundred micrometers wide (usually referred to as blobs) that are separated by regions that stain

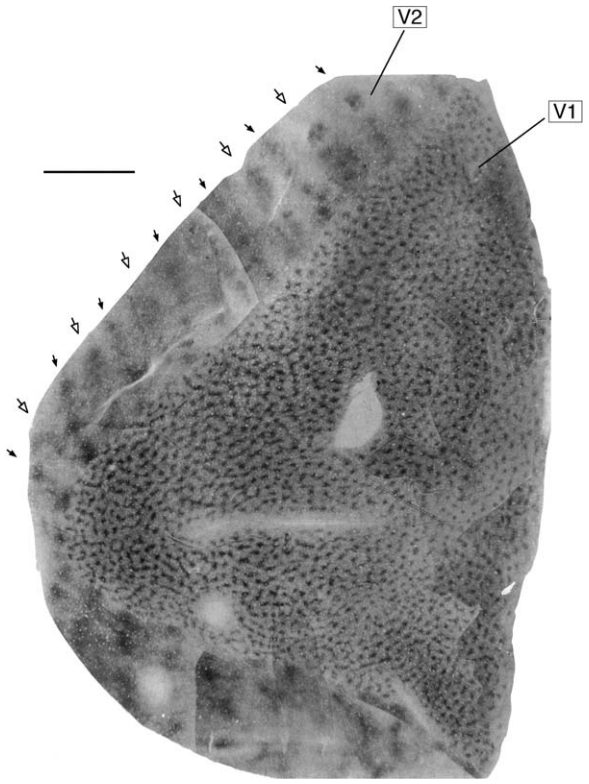


Figure 2 Photographic montage of sections through layer 3 of areas V1 and V2 in a New World monkey (*Cebus apella*). This reconstruction is based on sections tangent to the cortical layers, stained to reveal the mitochondrial enzyme cytochrome oxidase. The V1 blobs and V2 stripes are clearly visible. Thin and thick stripes are indicated by black and white arrows, respectively. Scale bar = 4 mm.

more lightly for cytochrome oxidase (interblobs). Cells in the blobs and interblobs differ even in terms of basic characteristics, such as their dendritic tree morphology and afferent connections from the thalamus. Thus, rather than a homogeneous collection of cell columns, some visual areas are best described as mosaic-like aggregates of two or more types of columns.

It is known that during development, afferents from different sources tend to segregate into adjacent groups of columns, based on the degree of their correlated activity. Thus, one possibility is that modules are formed whenever an evolutionarily new afferent pathway to an already existing visual area emerges. However, why are modules sometimes formed rather than having cells with different characteristics grouped into different areas? One suggestion is that the existence of a modular organization allows more efficient interactions between circuits of neurons that analyze different aspects of the image. For example, a shape can be delineated not only by differences in luminance between an object and the background (e.g., a shadow against a wall) but also by features such as texture, color or the correlated motion of image elements. Shape analysis based on interactions between cells that analyze these different parameters may be facilitated if they are located in modules a few hundred micrometers apart, rather than in separate areas several millimeters apart.

D. What Makes Visual Areas Different from Each Other?

1. Cortical Architecture

The term ''cortical architecture'' refers to the manner in which the different cellular components are distributed and interconnected in the three-dimensional structure of the cortex. In histological sections stained to reveal the neuronal cell bodies, some areas can be seen to differ in characteristics such as the number and thickness of the cortical layers, cell body sizes, and cell densities. These characteristics are referred to as *cytoarchitecture*. In reality, with the exception of the border of V1, cytoarchitecture is not particularly useful in distinguishing visual areas. The study of *myeloarchitecture* (the distribution and density of myelinated axons in the different layers of cortex) has proved more fruitful because several extrastriate areas can be distinguished in this way (Fig. 3). Nonetheless, the myeloarchitectural boundaries are often subtle, and it is normally the case that this type of analysis is used to confirm rather than propose the existence of an area. Finally, particularly in the past 20 years, histochemical and immunocytochemical techniques have revealed that visual areas differ in terms of the distribution of certain chemicals, such as enzymes and neurotransmitters. A good example of this is the detection of modular systems such as the cytochrome oxidase blobs of V1. Combining several architectural techniques in one preparation yields the best precision in determining the boundaries of visual areas because borders that may not be obvious with one stain may be clear with another.

2. Patterns of Connections

In many cases, the first indication that a distinct area exists in a region of cortex is the fact that it receives axonal projections that are different from those of adjacent cortex. For example, only a small region of the superior temporal sulcus of the macaque monkey

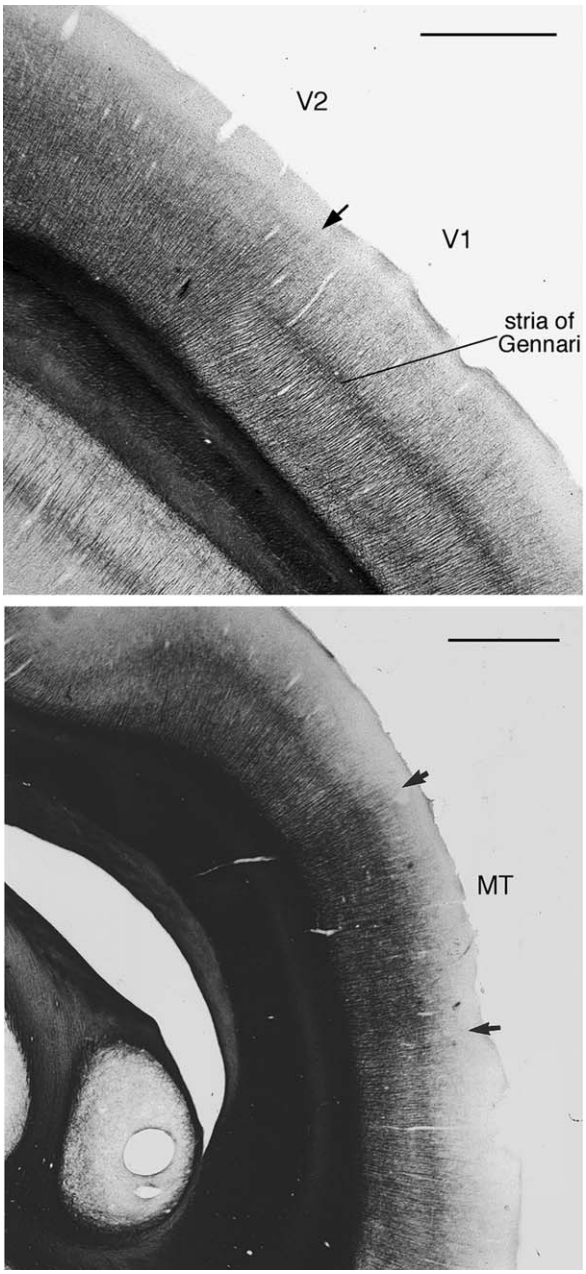


Figure 3 Myelin-stained sections (Schmued's gold chloride stain) illustrating the different myeloarchitectures of visual areas in marmoset monkeys. (Top) Parasagittal section through the V1/V2 border. (Bottom) Coronal section through area MT. The arrows indicate the borders of areas, and the scale bars 1 mm.

receives direct connections from V1. This was one of the first indications of the existence of the middle temporal area (MT), later found to be distinct from adjacent cortex also in terms of architecture and electrophysiological characteristics. Likewise, areas can be identified on the basis of their efferent connections. Thus, in New World monkeys there is a relatively circumscribed region of the inferior temporal cortex (the temporal ventral posterior area) that projects to V1.

3. Receptive Field Characteristics

The *visual receptive field* is the part of the visual field that when stimulated (by the presentation of objects or patterns in front of the eyes) will cause a neuron to respond by changing its firing rate of action potentials (Fig. 4). The receptive fields of neurons in different visual areas are distinct in terms of size, selectivity for stimulus parameters, and other physiological characteristics such as modulation by behavioral context.

Differences in sizes of excitatory receptive fields (i.e., those parts of the visual field in which stimulation causes a cell to *increase* its firing rate) are often a good physiological indicator of the boundary between two areas. In the part of V1 that deals with foveal vision, the receptive field of a single cell may extend only for a fraction of a degree of visual angle (a "window" of visual field about the size of one letter on this page, or even smaller). In the corresponding part of V2, receptive fields are two to five times larger (i.e., enough to include two or three letters), and those in areas anterior to V2 are several times larger again. In general, the average size of receptive fields increases gradually, from the most posterior area (V1) toward anterior areas. Some cells in visual areas of the parietal and temporal lobes may respond to stimuli presented anywhere in the visual field.

The selectivity of a neuron to the characteristics of visual stimuli presented within its receptive field constitutes its *response properties*. This concept is clearly illustrated in the responses of cells in VI (Fig. 4B, C). Some cells are *orientation selective*, i.e., they will only respond if the visual pattern falling within the receptive field contains lines of a given orientation (e.g., vertical). Others will only respond if the visual stimulus is moving in a particular direction (*direction-selective* cells) or if it has a certain wavelength composition (*color-selective* cells). The response properties of cells are thought to be a good indication of the physiological role of an area. For example, nearly all cells in area MT are direction selective, but their responses are nearly the same for stimuli of different colors; this is interpreted as meaning that cells in MT are part of a circuit that analyzes the movement of objects but does not contribute much to the analysis of color.

Finally, the responses of cells in some areas are more strongly modulated by the animal's behavior than are

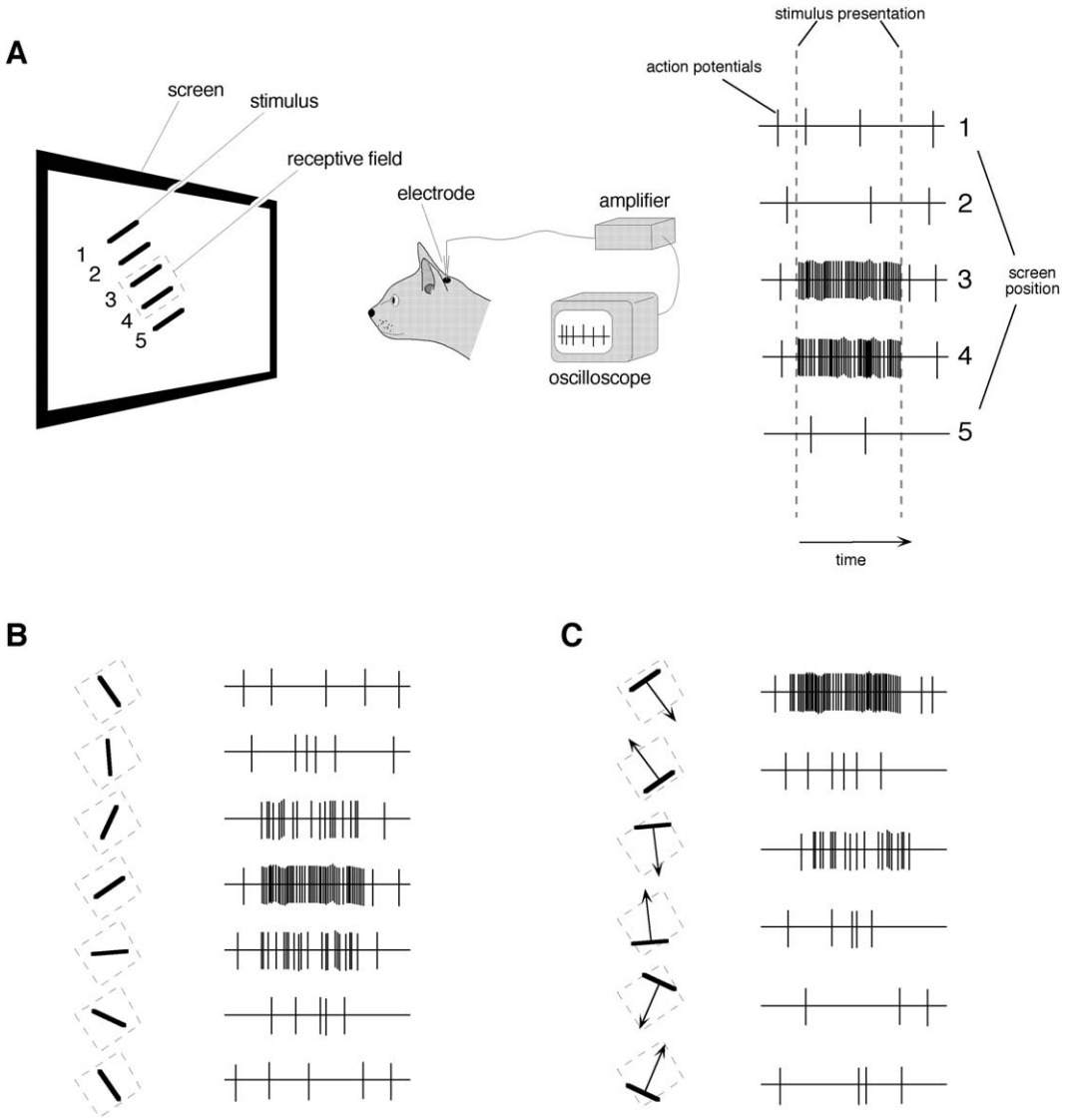


Figure 4 Receptive fields of cortical neurons. (A) Mapping the receptive field extent. The electrical activity of a single cortical cell is monitored while an animal observes a screen. A stimulus (line) is flashed in five different positions (1–5). As indicated on the right, only lines presented in positions 3 and 4 (which are within the receptive field) elicit a change in the cell's firing rate. (B) Orientation selectivity. Lines of different orientations are flashed within the cell's receptive field. The cell's response is greatest to a line of a specific orientation (2–8 o'clock in this case). (C) Direction selectivity. A stimulus moved across the cell's receptive field only elicits a response if it moves downwards, but not upwards.

the responses of cells in other areas. Responses can be elicited from cells in V1 even if the animal is deeply anesthetized, whereas the responses of cells in some extrastriate areas can be suppressed (depending on the type and dosage of anesthetic). In addition, the responses of neurons in some areas can be enhanced or suppressed depending on whether or not the animal is paying attention to that particular part of the visual field, or whether or not it intends to execute an eye movement toward the stimulus. This behavioral modulation is more marked in anterior areas, in comparison with posterior areas such as V1 and V2.

4. Visuotopic Organization

The axonal connections along the visual pathway are organized in such a way that adjacent retinal cells project to adjacent cells in the lateral geniculate nucleus of the thalamus, which in turn project to adjacent cells in V1. As a result of this arrangement, an

ordered map of the visual field (*visuotopic map*) is formed in the cortex. In this map, nearby V1 cells have receptive fields that represent nearby portions of the visual field; moreover, V1 cells on one side of the brain collectively represent the entire contralateral half of the visual field (Fig. 5). The connections between V1 and extrastriate areas, and among extrastriate areas, are also topographically organized; thus, neurons in each extrastriate area form a separate map of the visual field. By counting the number of visuotopic maps in one region of cortex, one can deduce how many areas exist. This has been the basis of many of the proposed subdivisions of primate visual cortex. Recently, non-invasive functional mapping techniques have also been used to map human visual areas based on their visuotopic organization. The visuotopic maps are most precise and therefore more useful as a criterion for subdividing the cortex in V1 and posterior extrastriate areas. Most areas in the inferior temporal and posterior parietal areas have no clear visuotopy.

5. Effects of Lesions

Lesions to different visual areas cause specific deficits in visual perception, and this is one of the best types of evidence to clarify an area's function. It is usually the case that controlled lesion studies are conducted only after there is a good indication of an area's extent. However, this is not always the case. In particular, the study of the effects of lesions in different locations has been instrumental in charting the subdivisions of the inferior temporal cortex, a region without clear visuotopic organization and where architectonic subdivisions are subtle. Long before the existence of multiple extrastriate areas was recognized, clinical evidence revealed that a lesion of a region of the human ventral occipitotemporal cortex causes specific disturbances of color vision (achromatopsia). This syndrome has since been replicated in studies in monkeys, in which the posterior part of the inferior temporal cortex was lesioned.

II. VISUAL AREAS OF PRIMATES

A. Nomenclature

For most of the 20th century, study of the subdivision of visual cortex was based primarily on cyto- and myeloarchitecture. Based on these analyses, it was thought that there were only three visual areas, commonly referred to as areas 17–19 (according to Brodmann's nomenclature). In the 1970s, however, this view was radically changed by the use of intracortical electrophysiological recordings, high-resolution neuroanatomical tract-tracing techniques, and new histological stains, which demonstrated that cytoarchitectural areas 18 and 19 are formed by multiple,

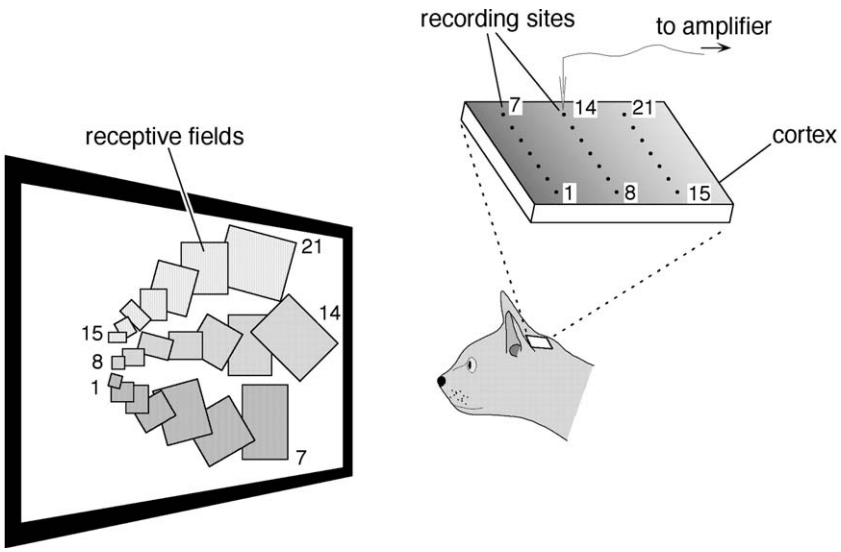


Figure 5 Visuotopic organization of visual cortex. An electrode is moved to different points (1–21) of a cortical area. At each point, the receptive field of a neuron is mapped. The positions of receptive fields in the visual field change systematically as a function of the recording site location.

functionally distinct subdivisions. In addition, the use of single neuron recording techniques demonstrated that visually responsive cells are also common in the inferior part of the temporal lobe (Brodmann's areas 20, 21, and 37), the posterior part of the parietal lobe (area 7), and some parts of the dorsolateral frontal lobe. In each case, these cytoarchitectural subdivisions proved heterogeneous in terms of connections and functions. Thus, although striate cortex is still referred to as area 17 (given that the functional and cytoarchitectural definitions coincide exactly), the numerical nomenclature cannot accurately be used in reference to extrastriate areas.

However, there is no agreement on the best nomenclature of visual areas, a situation that stems from the fact that the borders of these areas are still the subject of debate. Commonly, primate visual areas receive two- or three-letter designations that reflect their location in the brain. For example, area MT was first described in the middle temporal gyrus of the owl monkey, and the lateral intraparietal area (LIP) was first demonstrated in the lateral bank of the intraparietal sulcus of the macaque monkey. This is not an ideal situation, given that the locations of areas in relation to the cortical sulci and gyri vary among species. In other cases, visual areas have received alphanumeric designations that largely reflect the chronology of their discovery (V1 was the first area to be discovered, then V2, V3, V4, etc.). To complicate matters further, some areas have received two or three different designations, and the same name may be used to refer to two different areas. For example, area MT is also known as V5, whereas the name V4 can refer either to a ventrolateral area of the occipital lobe or to the "color center" located in the ventral posterior part of the temporal cortex (to avoid confusion, the latter is best referred to as V8). The emergence of a sensible, unified nomenclature of primate visual areas will probably have to wait until the exact boundaries of extrastriate areas are known and their homologies among species are clarified.

B. The Dorsal and Ventral "Streams"

Figure 6 compares maps of visual areas in a small South American monkey, the marmoset, and man. The likely homologous areas in monkey and man are indicated by similar colors. In monkeys, the boundaries of occipital visual areas have been studied in detail, and most of them can now be identified using a set of reliable criteria. However, the limits of the visual areas of the temporal, parietal, and frontal lobes are still not known with the same level of precision. It should also be kept in mind that relatively few of the human visual areas have had their borders fully determined; thus, most of what we believe about the functions of human visual areas is based on comparative evidence.

As mentioned previously, there are at least a few dozen visual areas in the primate cortex. However, for

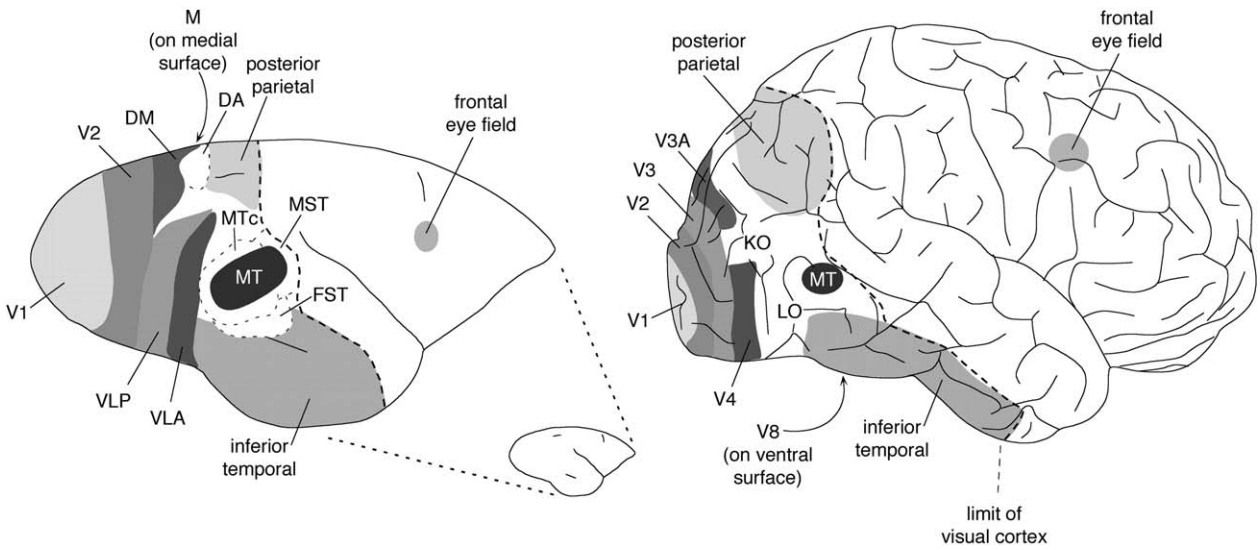


Figure 6 Extent of visual cortex and location of the principal visual areas in a New World monkey (marmoset; left) and man (right). Likely homologous areas are indicated by similar gray shading. The actual size of the marmoset brain in relation to the human brain is indicated in the insert.

didactic purposes, they can be subdivided into functionally related groups. Areas V1 and V2 represent the first cortical stages of visual processing and are characterized by modules that segregate cells that analyze different parameters of vision (contour orientation, luminance, motion, color, texture, etc.). These "*source areas*" project axons to three groups of areas located in the anterior part of the occipital lobe, near the transition with the parietal and temporal lobes: the *lateral*, *dorsomedial*, and *ventrolateral* groups.

Both the lateral group, centered on area MT, and the dorsomedial group, centered on the dorsomedial area (DM), are involved with spatial aspects of vision, including the analysis of motion and the determination of the spatial relationships between objects. These two groups, in turn, project to the *posterior parietal cortex*, in which a unified representation of extrapersonal space, including the location of objects in relation to the observer's body, is created. The ventrolateral group of areas, which in humans includes area V4, performs the early stages of the analyses of shape, color, and texture. These areas project to the inferior temporal cortex, in which these different aspects of object vision are synthesized.

The pathways connecting V1 and V2 to the lateral and dorsomedial groups, and these areas to the posterior parietal cortex, are referred to as the occipitoparietal pathway or "*dorsal stream*" of visual areas; they are involved in telling us *where* things are located. On the other hand, the pathways connecting V1 and V2 to the ventrolateral areas, and these in turn to the inferior temporal cortex, are known as the occipitotemporal pathway or "ventral stream"; they are responsible for establishing *what* things are.

C. Striate Cortex (V1)

Area V1 of primates is located at the most posterior part of the occipital lobe. In humans, it extends from the occipital pole to the rostral tip of the calcarine fissure, near the splenium of the corpus callosum. Flat mounts (histological preparations in which the cortical sulci are physically unfolded and cut parallel to the cortical layers) reveal that V1 forms an elongated oval with fairly sharp boundaries (Figs. 3 and 7). The cytoarchitecture of V1 (Fig. 8) is unlike that of any other cortical area, with each of the six cortical layers being subdivided into sublayers that have distinct afferent and efferent connections. For example, layer 4 is subdivided into two sub-layers (α and β) that are innervated by the magnocellular and parvocellular

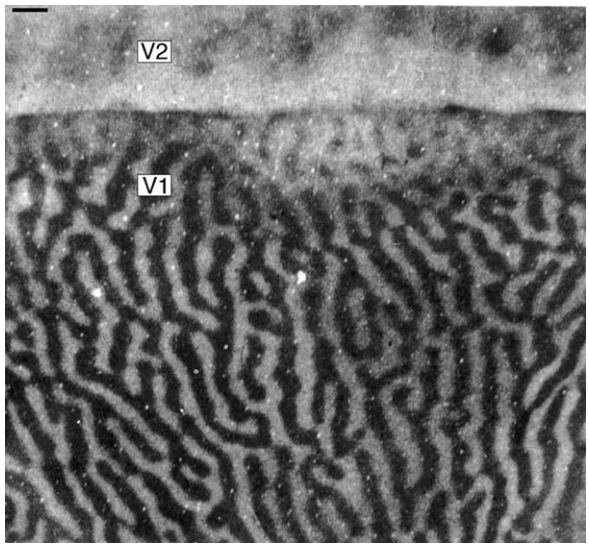


Figure 7 Section tangent to layer 4 of monkey V1, showing ocular dominance stripes revealed by cytochrome oxidase histochemistry. In order to reveal the configuration of stripes, one of the eyes was rendered inactive several weeks prior to the experiment. The light stripes correspond to regions innervated by the inactive eye, and the dark stripes correspond to regions innervated by the active eye. In normal conditions, layer 4 stains uniformly dark for cytochrome oxidase. Scale bar = 800 μm.

subdivisions of the lateral geniculate nucleus, respectively. Indeed, the name "striate cortex" reflects the histological distinctiveness of V1, being derived from the presence of a conspicuous myelinated band (the stria of Gennari; Fig. 3) that lies just above layer 4. The stria of Gennari (which is a subdivision of layer 3, despite being traditionally referred to as layer 4b) allows the easy identification of V1 even in nonstained histological sections. The complex modular and laminar architectures of V1 reflect multiple functional compartments, which process different attributes of vision. For example, the response properties of cells in layer 3 differ depending on whether they are in the blob and interblob compartments. Cells that are strongly selective to stimulus orientation are common in the interblobs, whereas those with selective responses to color contrast tend to reside in the blobs. Moreover, direction-selective cells in V1 concentrate in two layers, the stria of Gennari and layer 6.

Lesions of V1 have devastating effects on visual perception, even when all other visual areas are spared. Partial lesions of V1 cause scotomas, which are blind regions of the visual field whose size and shape reflect the portion of the visuotopic map that was affected by the lesion, whereas larger lesions cause cortical blindness, which is the complete loss of conscious visual

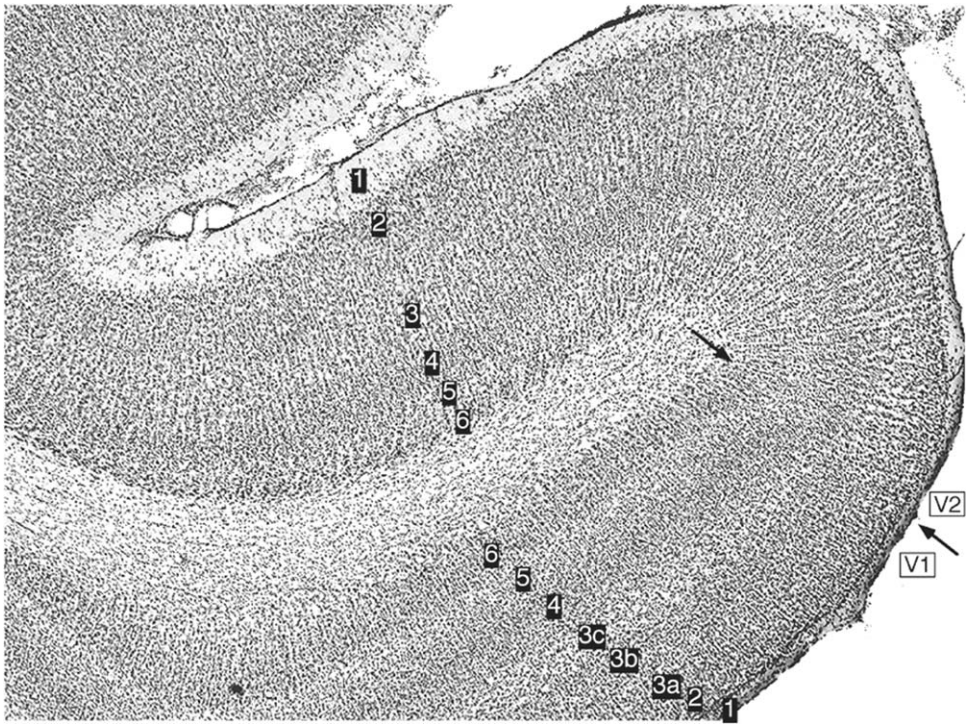


Figure 8 Section through the macaque monkey occipital lobe stained to reveal neuronal cell bodies (cresyl violet stain). The cortical layers are indicated by numbers. The V1/V2 border is indicated by arrows.

perception. This highlights the fact that in primates, V1 receives the bulk of the retinal projections to the cortex via the lateral geniculate nucleus. These projections terminate not only in layer 4 but also in subdivisions of layer 3 (mainly the blob compartments). The corticocortical efferent projections of V1, on the other hand, concentrate in three extrastriate areas: V2, MT, and DM.

D. Second Visual Area (V2)

Area V2 is by far the largest extrastriate area, being on average only slightly smaller than V1. It forms a continuous belt of variable width, which surrounds area V1 except for a small gap near the rostral tip of the calcarine sulcus. The cytoarchitecture of V2 is similar to that of adjacent areas, except V1. However, its limits can be easily determined in flat-mounted preparations stained to reveal the presence of the mitochondrial enzyme cytochrome oxidase. Such preparations reveal a unique architecture of ''stripes,'' alternately rich and poor in cytochrome oxidase content (Fig. 2). Like the blobs and interblobs in V1, these stripes represent modular compartments, containing cells that process different types of visual information. The cytochrome oxidase-poor stripes (or ''interstripes'') concentrate cells whose responses are selective to stimulus orientation and length and thus are likely to be involved in the processing of shape. Cytochrome oxidase-rich stripes, on the other hand, are of two alternate types: thick stripes that contain direction-selective cells and thin stripes that contain cells that respond strongly to color contrast but show little selectivity to either orientation or direction of motion. V2 is the main cortical target of V1 axonal projections. Each type of V2 stripe receives afferents from a different module or layer of V1 (Fig. 9), reflecting its functional specificity; for example, the shape-selective cells in the interstripes of V2 receive projections mainly from the interblobs, whereas the color-selective cells in the cytochrome-rich thin stripes are innervated by the blobs. The efferent connections of V2 are widespread, targeting nearly a dozen other visual areas in the occipital, parietal, temporal, and frontal lobes. Typically, these projections are also specific with respect to the stripe architecture: A given extrastriate area only receives projections that originate in one or two types of stripe.

The effects of V2 lesions have been studied in monkeys, but a clear picture of the exact contribution

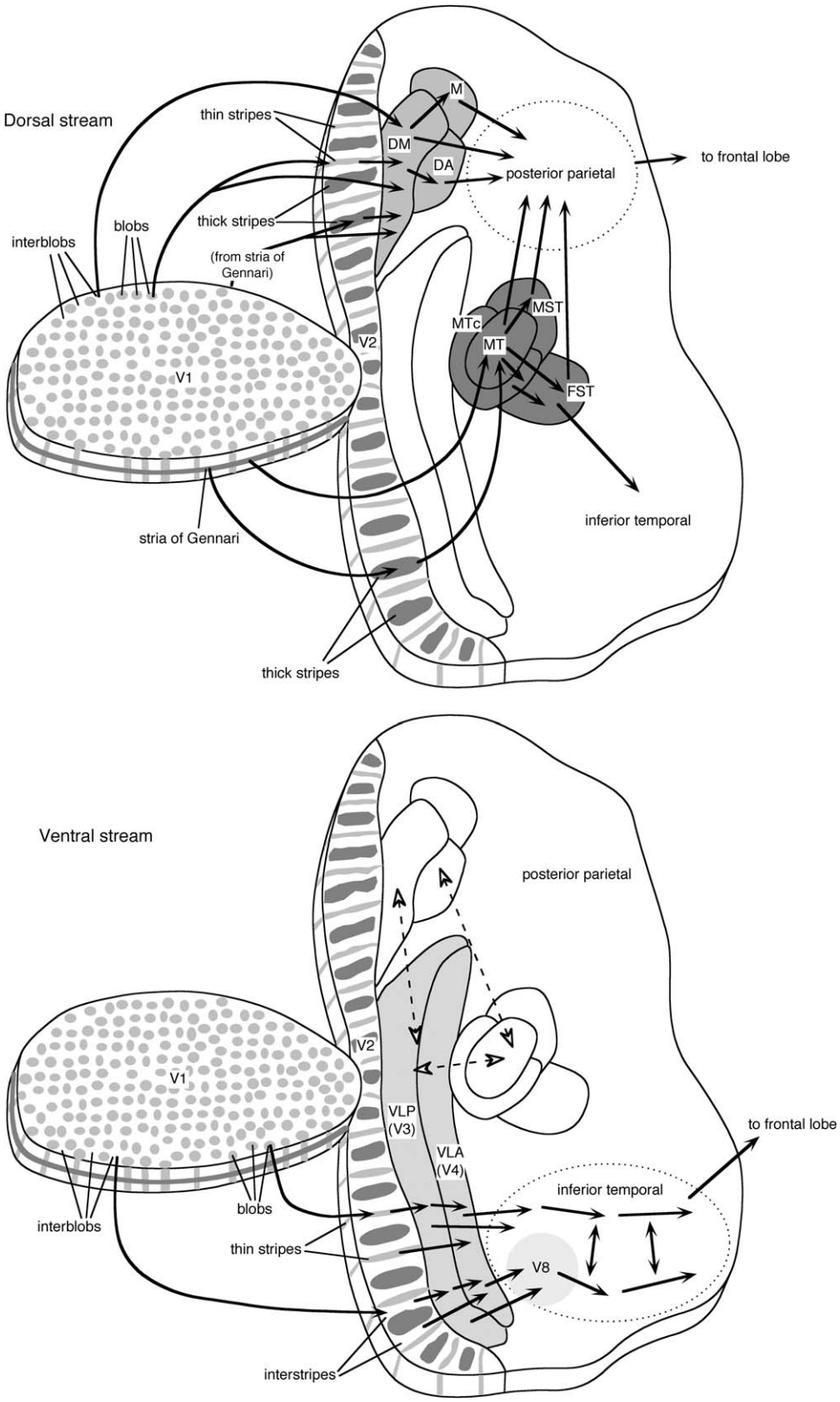


Figure 9 Schematics of the main axonal interconnections between visual areas in the monkey. In these diagrams, V1 has been separated from extrastriate cortex, and the sulci have been "unfolded." (Top) Dorsal stream areas. (Bottom) Ventral stream areas. The white arrows in the bottom diagram indicate interconnections between "streams."

of this area for vision has not yet emerged. Unlike the devastating effects of a V1 lesion, monkeys with V2 lesions are not affected in terms of visual acuity or contrast sensitivity. However, they can be severely impaired in more complex tasks involving spatial discriminations. For example, although these monkeys have no problem detecting the orientation defined by a series of aligned line segments, they perform poorly if the same pattern is presented against a background of line segments of various orientations.

E. Lateral Group

The second main target of V1 projections is MT, an oval-shaped area located near the limit between the occipital and temporal lobes. Area MT is much smaller than either V1 or V2. Unlike most extrastriate areas, it can be easily distinguished from adjacent cortex in histological sections due to an unusually high myelin content (Fig. 3B). Functionally, MT appears to be specialized for the analysis of motion. Nearly 90% of MT neurons are strongly direction selective, and they are far more strongly responsive to moving patterns than to stationary ones. These neurons are organized into two types of modules, which analyze either motion over large parts of the visual field (wide-field motion modules) or the contrast between the direction of motion of one object and that of the background (local-motion modules). Given the response properties of MT cells, it is not surprising that this area's afferent connections from V1 and V2 originate in layer 3 compartments that are themselves involved in motion analysis: the stria of Gennari and the thick cytochrome-rich stripes, respectively (Fig. 9).

The cortex surrounding MT includes a set of small visual areas that are also part of the motion analysis pathway, each being strongly interconnected with MT. The middle temporal crescent (MTc, also known as V4t) is a belt-shaped area that surrounds MT along most of its posterior border, whereas the medial superior temporal area (MST) adjoins MT anteriorly. A third area, FST (first described in the fundus of the superior temporal sulcus of the macaque), is located ventral to MTc (Fig. 9). Although the histological characteristics of MT have allowed the determination of its extent in the human brain, the exact boundaries of MTc, MST, and FST remain unknown. The physiological responses of cells in these areas and their anatomical connections suggest that they are involved in different types of motion analyses. Cells in MST have very large receptive fields, and they respond best to complex moving patterns, such as radial patterns of motion that approach or recede from the fovea. These cells appear suited for the analysis of optic flow, the perceived motion of the visual field that one experiences when moving through the environment. Optic flow is important for determining the direction of an observer's self-motion as well as providing clues about distances (e.g., objects that are closer to the observer move faster than distant ones). In contrast, the neuronal receptive fields in MTc and FST tend to be smaller, and only about half of the cells are strongly direction selective. Whereas the connections of MST focus, to a large extent, on areas of the posterior parietal cortex, those of FST target both parietal and temporal areas. This suggests that FST is involved in the integration of information analyzed by the dorsal and ventral streams of processing.

Due to its small size, lesions of area MT typically also affect its satellite areas to some extent. These lesions cause marked deficits in the perception of motion without affecting the visual acuity for stationary patterns. The deficits can be temporary for simple tests, such as deciding whether an object is moving to the left or to the right. However, more complex discriminations can be permanently affected —for example, in situations of transparent motion (when a group of dots moves coherently in one direction, against a "noisy" background of dots moving in different directions). The specific loss of motion perception caused by bilateral MT lesions (akinetopsia) is extremely rare in humans.

After lesions of V1, some neurons in area MT remain responsive to visual stimuli presented inside the resulting scotoma. These cells are activated by an indirect visual pathway that bypasses V1, involving sequentially the retina, the superior colliculus, the pulvinar, and MT. It is believed that these responsive cells are the physiological basis of *blindsight*, the ability demonstrated by patients and monkeys with V1 lesions to guess correctly the characteristics of visual patterns presented in their blind field, despite their lack of conscious visual perception. Blindsight demonstrates a dissociation between having access to visual information, on the one hand, and being able to consciously experience vision, on the other hand.

F. Dorsomedial Group

The third main cortical target of V1 axonal projections is the DM area, which corresponds to a densely myelinated, wedge-shaped region located anterior to

V2, near the dorsal midline. In humans, the likely homolog of DM has been named V3A (Fig. 6). The exact contribution of DM to visual processing remains unclear. The projections of V1 and V2 to DM originate from multiple modular compartments, suggesting that DM cells can integrate information about different visual parameters. In agreement with this hypothesis, the responses of most DM neurons are strongly dependent not only on the orientation of contours but also on the direction of motion. One possibility is that DM is involved in the separation of objects from background on the basis of motion cues. Another clue to the function of DM neurons is the fact that although the visuotopic maps of most visual areas strongly emphasize the representation of the central part of the visual field, this trend is less marked in DM. This argues against a role in the fine analysis of shape. However, the effects of DM lesions have not been the subject of controlled studies.

In monkeys, the cortex immediately anterior to DM includes other areas that are anatomically and functionally related to DM. These include the dorsoanterior (DA) and medial (M) areas, which are located at the occipitoparietal transition, near the midline. The emphasis on representation of central vision is even less marked in M than in DM; in fact, M remains unique in being the only visual area that represents all parts of the visual field more or less equally. These occipitoparietal areas form an important anatomical link between occipital areas, including DM, and the posterior parietal cortex. This suggests that they represent stages of spatial vision processing and may contribute to visuomotor integration, including the control of eye movements.

G. Posterior Parietal Cortex and Frontal Visual Areas

The designation ''posterior parietal cortex'' encompasses a group of areas located within cytoarchitectural areas 5 and 7, which are important for visuomotor integration and direction of attention to specific parts of extrapersonal space. Most, if not all, areas in this part of visual cortex appear to be specific to primates. One of the best characterized posterior parietal areas is LIP, which is involved in the control of saccadic eye movements and in updating the brain's neural representation of extrapersonal space with respect to eye movements. Other subdivisions of the posterior parietal cortex are involved in the visual coordination of hand movements and the analysis of self-motion. In many cases, this requires combining visual inputs with information processed by the somatosensory and motor systems (e.g., information about the positions of joints and the motor commands sent to the eyes). It is unclear exactly how many areas exist within the posterior parietal cortex. At least six subdivisions have been proposed for the monkey, and more may exist in humans. Lesions involving posterior parietal areas can cause a number of symptoms, including difficulties in initiating or maintaining eye movements, as well as problems with visual orientation toward the side contralateral to the lesion. In some cases, a patient may start to ignore completely the extrapersonal space contralateral to the lesion, a condition referred to as *neglect*.

The posterior parietal cortex is strongly interconnected with the cortex of the frontal lobe, which is involved in complex cognitive functions including motor planning and working memory. Among the frontal areas innervated by the posterior parietal cortex are the frontal eye field and the supplementary eye field, which, as suggested by their names, are closely involved in the control of eye movements including saccades. Cells in both areas respond to the presentation of visual stimuli within their receptive fields but are also active immediately prior to eye movements.

H. Ventrolateral Group

Two relatively large, highly elongated visual areas occupy the regions of cortex located anterior to V2, in the lateral aspect of the occipital lobe and the ventral cortex that overlies the cerebellum. In marmoset monkeys, in which the most complete maps of their extent and organization have been obtained, these are named the ventrolateral posterior and ventrolateral anterior areas, whereas in the most commonly used nomenclature of human areas these have been dubbed visual area 3 (V3) and visual area 4 (V4). An additional area, named LO (lateral occipital area), appears to exist in the human cortex that is functionally related to the ventrolateral group (Fig. 6). Physiological studies in monkeys demonstrate that the ventrolateral areas represent stages in the processing of shape, color, and texture—features that are important for identifying what an object is. For example, the responses of cells in V4 are sensitive to characteristics such as the length/width ratio of a rectangle as well as its size and orientation. Moreover, these cells will respond best if the color of the stimulus is different from that present

in the background. The visuotopic maps of ventrolateral areas are strongly dominated by the representation of the fovea, where visual acuity is maximal. This is in agreement with a role of these areas in the fine analysis of patterns.

Lesions of ventrolateral areas, including V4 and probably parts of V3, have been studied extensively in monkeys. In general, the observed deficits reflect the role of these areas in object recognition. Thus, monkeys with V4 lesions have problems discriminating between objects that are similar in shape but not between very dissimilar objects. They are also profoundly impaired in the recognition of an object presented from different viewpoints, or that has been partially occluded from view by other objects. Finally, V4 lesions cause problems in discriminating the characteristics of an object presented among others, especially if these other objects are more "salient" due to a higher contrast in relation to the background. In humans, lesions involving the ventrolateral areas may be the cause of *object agnosia*, the inability to recognize real objects, and *agnosia for drawings*, the inability to identify a drawn object.

There are relatively few interconnections between V1 and the ventrolateral areas. In contrast, they receive massive projections from V2 that originate from the cytochrome-poor interstripes and the thin cytochrome-rich stripes (Fig. 9). The ventrolateral areas form the bulk of the cortical projections to the inferior temporal areas.

I. Inferior Temporal Cortex

The histological distinctions between inferior temporal areas are subtle, and the best evidence for multiple subdivisions comes from functional studies. We are far from knowing exactly how many areas exist, or even whether the functional subdivisions of the inferior temporal cortex can be adequately described as spatially segregated areas. Nonetheless, evidence indicates that the inferior temporal areas form a posterior-to-anterior gradient of processing, in which increasingly more complex "syntheses" of object features are generated at each stage. Thus, in this simplified model, neurons in the posterior areas of the inferior temporal cortex analyze relatively simple features, such as the number of corners and intersections in a shape or its color composition. These areas project to more anterior areas, which combine these features into increasingly more "explicit" representations of objects. For example, some neurons in anterior inferior temporal cortex are strongly activated upon an animal viewing a picture of a face, but they may respond less strongly to a picture of a face in which the positions of the eyes, nose and mouth have been rearranged (Fig. 10). This suggests that neurons in these areas are not involved in detecting simple physical characteristics but rather the exact combinations of features that characterize an object. It is therefore not surprising that lesions involving the inferior temporal cortex cause deficits not only in object identification but also in the consolidation of visual memory.

Besides the posterior–anterior gradient, there are also distinctions between dorsal and ventral regions of the inferior temporal cortex. Thus, in humans, a ventral area of posterior inferior temporal cortex (V8) appears to be specifically involved in the processing of color. A lesion of this region, which is located in the fusiform gyrus, causes loss or severe distortions of color vision (*achromatopsia*). More anterior lesions of

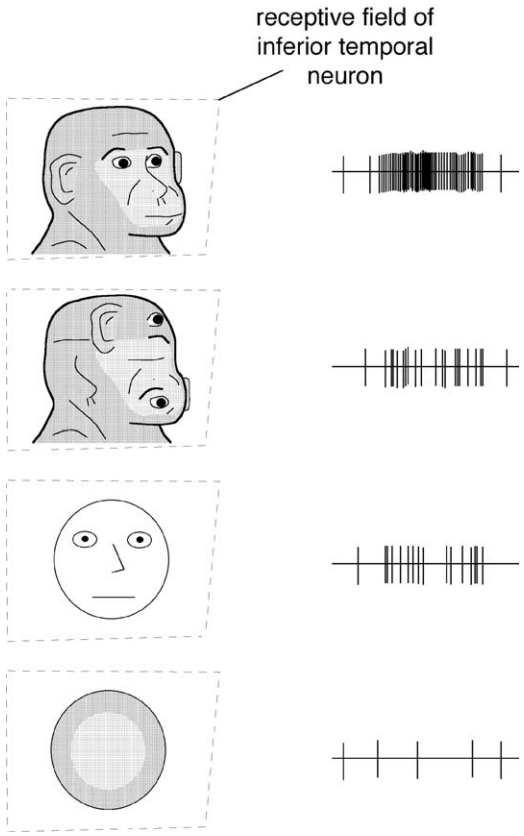


Figure 10 Responses of a face-selective inferior temporal neuron. A photograph of a monkey face presented on the cell's receptive field elicits a strong response (top). Pictures of "scrambled" faces, cartoon faces, or geometric shapes yield weaker responses.

the ventral part of the inferior temporal cortex cause deficits in the discrimination between similar objects that belong to the same category (the most intensively studied of these syndromes is *prosopagnosia*, the inability to distinguish faces). Finally, dorsal areas of the anterior part of the inferior temporal cortex integrate inputs from the ventral and dorsal streams together. In monkeys, neurons in this region, located in the superior temporal sulcus, may respond best to a view of a person walking forward but not backward. This response property requires convergence of shape and motion information onto single cells.

III. FUNCTIONAL ARCHITECTURE OF VISUAL CORTEX

With few exceptions (such as the distinct lamination of V1), the realization that visual areas are often rather distinct in terms of architecture is relatively recent. Indeed, it has been a long-standing view in the fields of cortical anatomy and physiology that the cortical circuits in different areas are essentially similar, and that many of the functional differences between cortical areas are due to the different sources of axonal inputs. Research in the past two decades has demonstrated that this is not an accurate view, and that functional differences are often accompanied by structural differences. This realization is the result of the development of new methods for revealing the chemical architecture of the brain, the detailed morphology of cells, and the way they interconnect within one area.

A. Neurons of the Visual Cortex

Like all neocortical areas, the areas of visual cortex are formed by six layers of cells, numbered sequentially from the surface of the brain to the white matter. Even V1, which has acquired a unique lamination during the course of primate evolution, can be seen as an elaboration of the basic six-layered scheme (Fig. 8). There are two types of excitatory neurons: pyramidal cells and spiny stellate cells. The differences between pyramidal and spiny stellate cells include the shape of the cell body (from which their names derive) and the fact that pyramidal cells have apical dendrites, which project across layers toward the pia mater. Spiny stellate cells tend to concentrate around layer 4, whereas pyramidal cells are more numerous in layers 3, 5, and 6. Whereas spiny stellate cells are usually regarded as local circuit interneurons, there are some exceptions to this rule, including many of the cells that form the pathway from V1 to MT. An elaborate dendritic tree covered in spines (which are membrane specializations for synaptic contact) characterizes both of these cell types. The morphology of pyramidal cells varies systematically between visual areas. For example, the dendritic trees of V1 pyramidal cells tend to be smaller and to have fewer spines per unit length, in comparison with V2 or MT cells. In general, the size of the dendritic tree, the complexity of its branching pattern, and the spine density all tend to increase as one considers cells in successively more anterior visual areas. As a result of these trends, pyramidal neurons in layer 3 of inferior temporal areas, for example, may have 10 times more spines than a V1 cell in the same layer, allowing for the integration of more synaptic inputs.

Smooth stellate are the main cortical inhibitory neurons, which distribute across all cortical layers. They can be further divided into morphological subgroups depending on their axonal branching pattern, the targets of their synapses, and the presence of different calcium-binding proteins. The distribution and connectivity of these inhibitory cell types are distinct in different areas. For example, the inhibitory terminals of parvalbumin-immunoreactive chandelier cells are relatively sparse in V1, more common in V2, and most frequent in inferior temporal cortex.

B. Cortical Layers

Studies of the connections between cells in the same visual area suggest that the six layers of cortex segregate cells that perform distinct sequential stages of visual processing. The basic picture that emerges from these studies (Fig. 11) is as follows:

- Layer 4 is the principal cortical input layer because it receives the majority of the afferent connections that determine an area's function. Thus, the projections from the lateral geniculate nucleus of the thalamus to V1 terminate in layer 4, as do those from V1 to V2. Layer 4 is also known as the *granular layer* in view of the small and closely packed stellate cell bodies contained therein.
- By means of vertical ascending axons, cells in layer 4 project neurons in the *supragranular layers* (layers 2 and 3), either directly or through intervening interneurons. Cells in layers 2 and 3 give origin to *extrinsic connections* (i.e., those that connect an area to

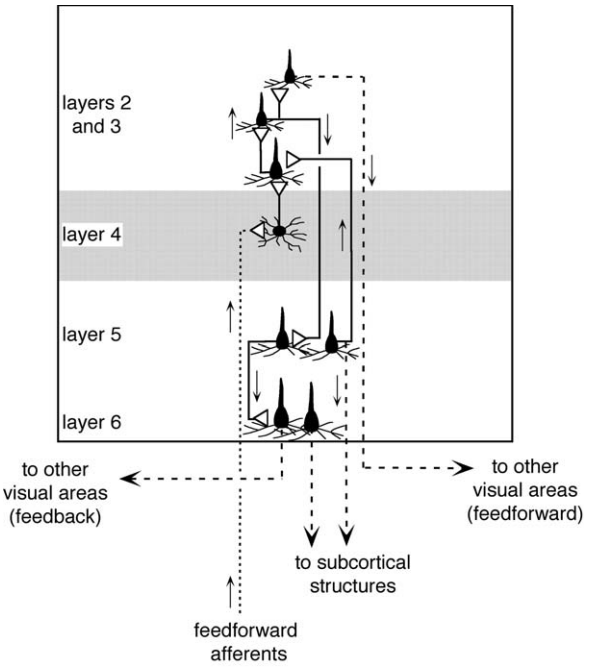


Figure 11 Main laminar excitatory interconnections between cells in a cortical column.

other areas, either in the same or the opposite hemisphere). For example, most V1 cells that project to V2 are located in layer 3.

- Axons of layer 3 cells also project to cells in layer 5, which in turn project to layer 6. Neurons in both layers 5 and 6 (the *infragranular layers*) are the main sources of extrinsic connections to subcortical nuclei (including the lateral geniculate nucleus, pulvinar complex, superior colliculus, pons, and claustrum). In addition, as detailed later, they form "feedback" pathways to other cortical areas (e.g., from MT back to V1).
- In addition to the previously mentioned connections, cells in the different layers interact vertically (i.e., across layers) via other systems of intrinsic axons. For example, cells in layer 5 send axons to layers 2 and 3, and those in layer 6 contact inhibitory cells in layer 4.

C. Columnar Organization

It is clear from the previous discussion that many of the interactions between cells in the same visual area occur via vertical axons, which run across layers. By virtue of this arrangement, the functional organization of visual cortex is dominated by *columns*, groups of tightly interconnected cells that occupy cylindrical compartments of cortex spanning all cortical layers. Functionally, a cortical column is defined by a particular response property, which is shared by all cells that form that column. For example, in many primate species the axons that bring information from the left and right eyes to V1 terminate in interdigitating but largely nonoverlapping territories of layer 4. In these territories (*ocular dominance stripes*; Fig. 7), cells respond exclusively to stimulation of one of the eyes. The inputs from the two eyes then converge onto single cells in the supra- and infragranular layers of V1; nonetheless, binocular cells in these layers continue to respond more strongly to one of the eyes, according to whether they overlie a left- or right-eye layer 4 stripe. Thus, the predominantly vertical interactions between layer 4 and the other layers of V1 result in *ocular dominance columns* spanning all cortical layers. The responses in all cells in a V1 column are also selective for a similar stimulus orientation, characterizing *orientation columns* (Fig. 12).

The columnar organization of different areas reflects the type of analysis performed. For example, area MT is organized into *direction columns* in which all cells respond maximally to stimuli moving in a particular direction, whereas the inferior temporal area is organized into *shape selectivity columns* in which all cells respond maximally to objects sharing certain features, such as curvature, number of intersections, or color.

Normally, the selectivity of cells for a certain stimulus parameter changes gradually between adjacent columns (e.g., the orientation columns in Fig. 12). A group of columns that represent all possible variations of a given stimulus parameter is termed a *hypercolumn*. For example, an ocular dominance hypercolumn in V1 is formed by adjacent left- and right-eye columns, whereas a direction hypercolumn in MT is formed by many columns, which collectively encompass cells sensitive to the full spectrum (360°) of possible directions of motion. Because cells in adjacent columns have receptive fields covering similar regions of the visual field, the hypercolumnar arrangement allows each small region of the visual field to be fully analyzed with respect to the same parameter (i.e., by cells with different sensitivities) within relatively compact "blocks" of cortex. The best studied hypercolumnar systems are approximately 0.5–1.0 mm across (i.e., parallel to the cortical layers), being formed by dozens of individual cell columns. Finally, in some areas many hypercolumns, representing slightly different parts of the visual field, aggregate into modules. The best example of these are in V2, in

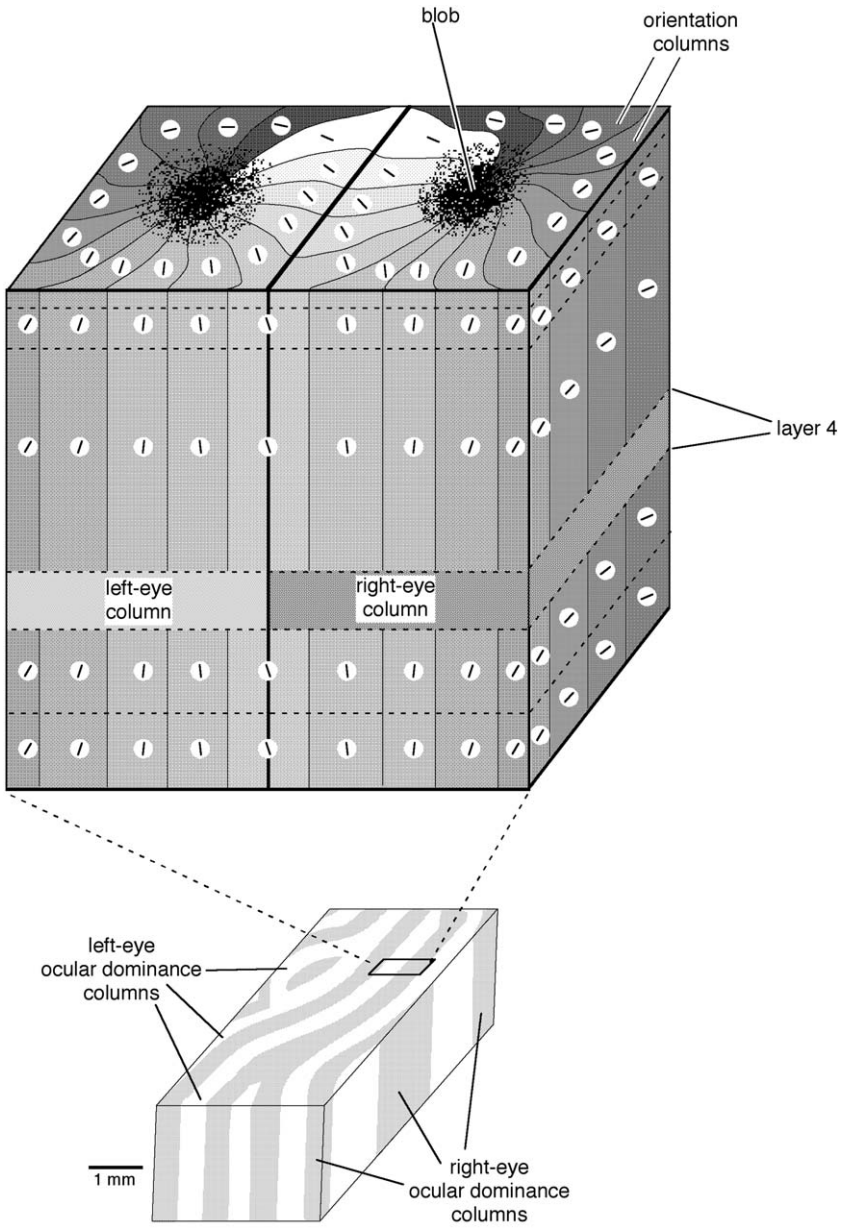


Figure 12 Columnar organization of striate cortex. (Bottom) A small "block" of V1, with ocular dominance columns indicated. A region encompassing two ocular dominance columns is magnified in the top diagram. (Top) Organization of orientation columns. The orientation selectivity of cells in each column is indicated by the line segments. The preferred orientation of neurons changes gradually between adjacent columns. Columns of cells selective to different orientations converge on orientation "pinwheels" centered along the ocular dominance columns. In many cases, cytochrome oxidase blobs (where cells are not strongly selective for orientation) overlie the pinwheels.

which hypercolumns for the same image parameter aggregate to form stripes (Fig. 2).

D. Horizontal Intrinsic Axons

In addition to the vertical axons that connect cells within the same column, visual areas have a system of *horizontal* intrinsic axons that integrate cells in different columns (Fig. 13). Short intrinsic axons connect cells in nearby columns, for example, to mediate inhibitory interactions among cells within the same hypercolumn. Furthermore, there are long horizontal axons whose connections are highly specific: They bypass many columns of cells to form tight clusters of

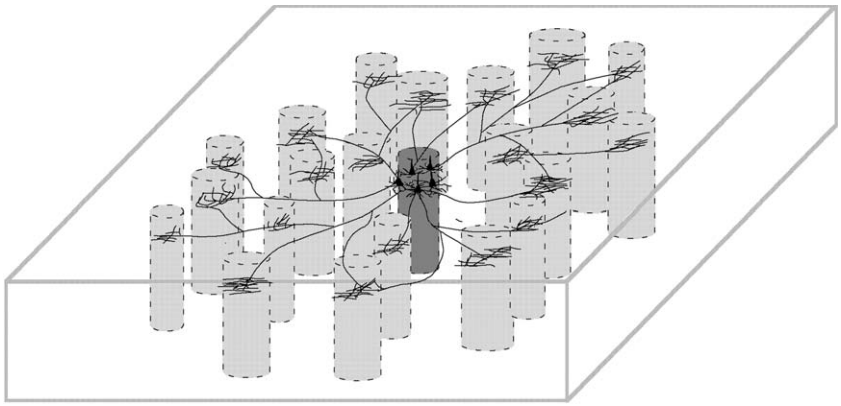


Figure 13 Long-range horizontal connections in layer 3. Cells in one column (dark gray) send horizontal axons to other columns several millimeters away (light gray). Although the horizontal connections are only illustrated for one layer, they in fact extend in register throughout the thickness of the cortex.

terminals in a particular region, sometimes millimiters apart (Fig. 13). The anatomical characteristics of intrinsic axons vary systematically between areas. They are relatively short ranged and have tightly clustered termination patterns in V1. As one moves toward more anterior areas, increasingly longer horizontal axons are observed, and their termination patterns become less tightly clustered. Thus, cells in V1 interact directly only with cells that analyze neighboring parts of the visual field, the interactions of cells in extrastriate areas become progressively more extensive. Horizontal axons are important for a number of physiological processes, most notably contextual modulation, visuotopic reorganization, and sharpening of neuronal tuning to certain stimulus parameters (Fig. 14).

Contextual modulation is the process whereby a neuron's response to a visual pattern is modified by visual stimuli located outside its excitatory receptive field (Fig. 14A). Thus, cells in the local motion modules in area MT will not respond if the background is moving in the same direction as the stimulus presented in the receptive field, whereas cells in V4 may respond less strongly if the object is presented on a background of similar color. These types of comparison, which are mediated by long horizontal axons, require connections between cells with receptive fields located in *different* parts of the visual field but that are selective for the *same* image parameter.

Visuotopic reorganization refers to the changes in the position of neuronal receptive fields that follow changes in the inputs to a visual area. For example, if a region of the retina is inactivated (after a laser lesion or retinal detachment), cells in V1 that had receptive fields in that part of the visual field do not become permanently inactivated (Fig. 14B). Instead, they acquire receptive fields that cover unaffected parts of the visual field surrounding the blind area. This process occurs within a few hours of the lesion and is mediated by long horizontal axons. If the lesion is irreversible, the neuronal responses to stimulation of the new receptive field become gradually stronger over a period of months, a process that appears to be mediated by the growth of new horizontal axons within V1. Visuotopic reorganization has also been observed in area MT after lesions of V1.

The response selectivity of visual cortical cells is sharpened by short-range intrinsic axons. For example, the orientation selectivity of cells in V1 is normally sharp: A cell that is most responsive to vertical contours will only show very weak responses to a similar pattern tilted by 45° and will not respond to a horizontal contour. However, if the strength of inhibitory interactions between adjacent columns is reduced by pharmacological manipulations (Fig. 14C), the same cell starts to show responses to stimuli with those "nonoptimal" orientations. This illustrates the fact that in normal circumstances neighboring columns, which represent different orientations in the same part of the visual field, interact extensively.

E. Feedforward and Feedback Extrinsic Connections

In Section II, the main corticocortical connections between areas were described according to a serial scheme; for example V1 projects to V2, which in turns

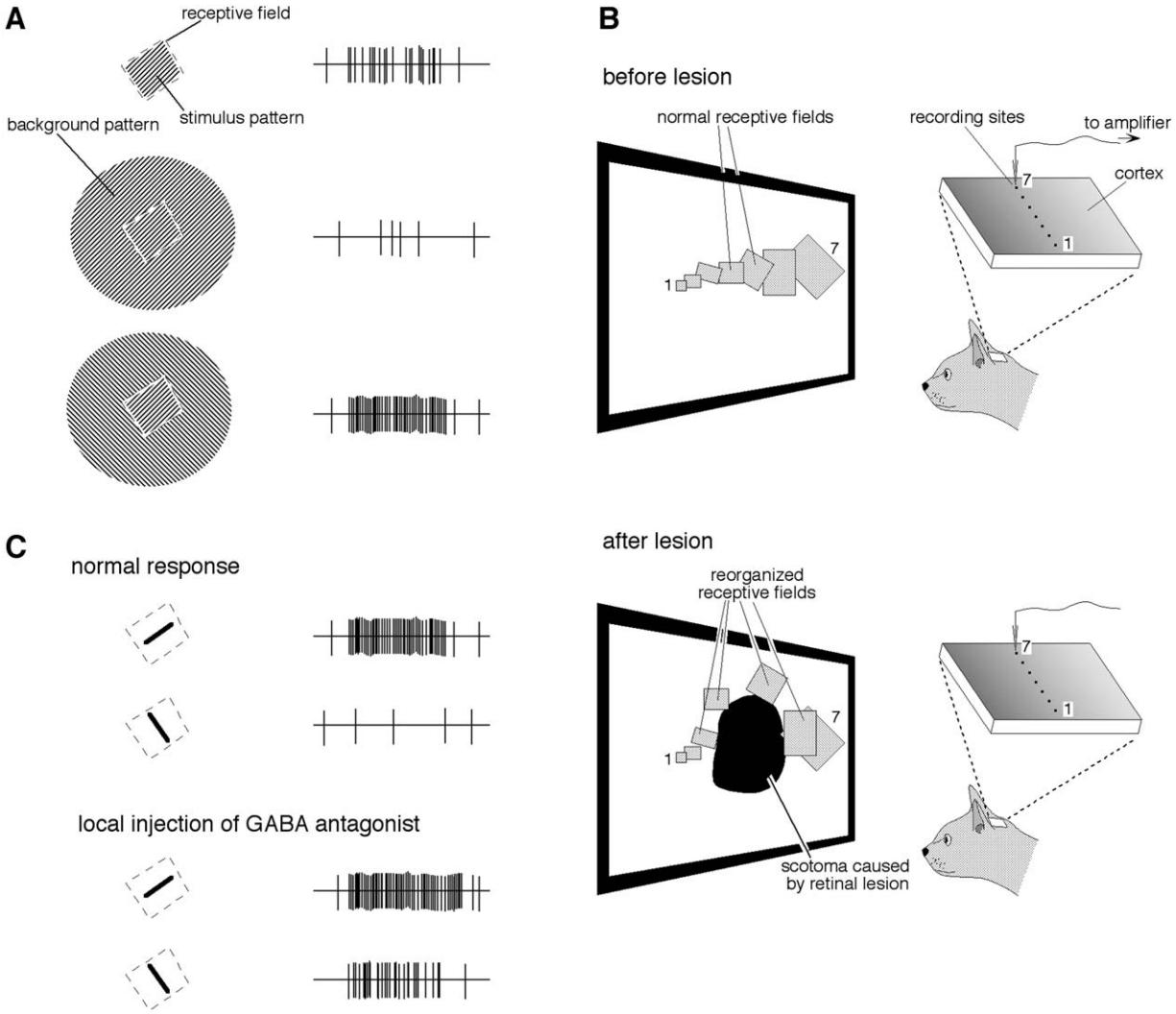


Figure 14 Functions of horizontal connections. (A) Contextual modulation. The neuron's response to a striped pattern presented on the receptive field (top) may be inhibited if the same pattern is presented against a background of similar orientation (middle) or enhanced if the pattern is presented against a background of different orientation (bottom). (B) Visuotopic reorganization. (Top) Normal receptive fields corresponding to a sequence of recording sites in the cortex. (Bottom) After a retinal lesion, the cells that originally represented the affected part of the visual field acquire new receptive fields around the edge of the scotoma. (C) Sharpening of orientation tuning. Upon pharmacological blockage of inhibitory neurotransmission, a neuron starts responding to stimuli of an orientation orthogonal to its preferred orientation.

project to V4, which projects to posterior inferior temporal cortex. This is an oversimplification, because connections between areas are reciprocal: At the same time that V1 projects to V2, some V2 cells project to V1. As a result of this arrangement, connections between different visual areas can be subdivided into two main groups. *Feedforward* connections are those described in Section II. They originate mainly from neurons in layer 3 of one area and contact cells in layer 4 of the target area. In addition, there are *feedback* connections, which originate primarily from infragranular cells and terminate outside layer 4 (Fig. 15). Although cells participating in these two systems appear to be equally numerous, there are good reasons to believe that feedforward connections carry the main flow of information between areas. For example, interruption of the feedforward pathway by means of a lesion of V1 results in a general loss of visual responsiveness in V2 cells, whereas a lesion of V2 only results in subtle changes in the response properties of V1 cells. Moreover, the *response latency* (i.e., the amount of time between the presentation of a visual

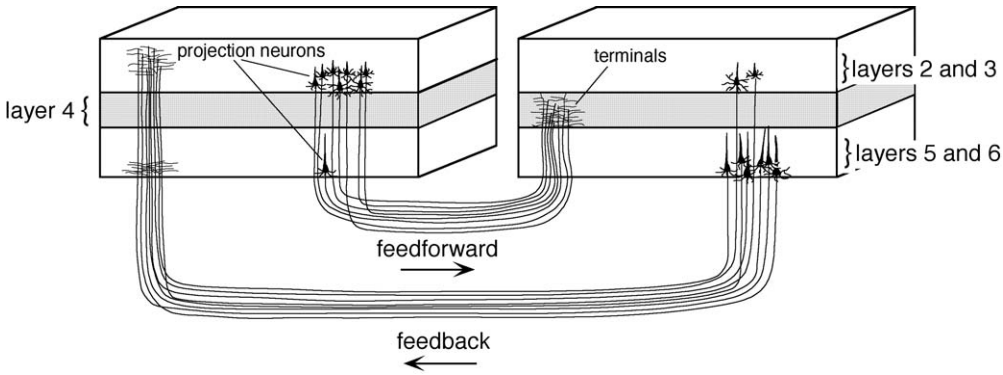


Figure 15 Laminar origin and termination of feedforward and feedback connections between two visual areas.

stimulus and the neuron's response) in different areas reflects, to a large extent, the order suggested by feedforward connections. The patterns of feedforward and feedback connections can be used to rank visual areas according to an *anatomical hierarchy*: V1 occupies the first level of the hierarchy, V2 the second level, etc.

Feedback connections modulate some of the contextual interactions between stimuli presented in the cell's receptive field and the background. Specifically, they enhance these interactions in "low-visibility" situations in which there is little difference between object and background. In addition, it has been suggested that feedback connections are important for visual imagery. According to this view, feedback projections from inferior temporal areas to occipital visual areas (including V1) would be active when one imagines an object while keeping the eyes closed. This would activate the appropriate sets of cells in these areas to form a neural representation of the object that resembles the one elicited by the actual viewing of the same object.

IV. CONCLUSION

The visual cortex can be described at several levels of functional organization. Single cells interconnect according to precise interlaminar patterns to form columns, whereas columns interact not only locally, within the same hypercolumn, but also with distant columns representing other parts of the visual field. Neurons in each cortical area form a separate representation of the visual field, each engineered by natural selection in order to allow the analysis of a different set of features of the visual image. Cells in different areas also interconnect extensively, according to highly precise modular systems that keep cells analyzing different visual parameters more or less segregated in different compartments. However, multiple reentrant pathways, both within and between areas, allow interaction between these different modular systems. In summary, the functioning of visual cortex as a whole is unlikely to resemble that of a linear machine, which generates a representation of the "world outside" on the basis of serial steps of analysis. Instead, the visual cortex could be best described as networks of interconnected cells within progressively larger networks of interconnecting cells. Each of these levels of organization performs certain neuronal operations under the influence of many other neural networks.

The evidence also makes clear that the visual cortex is not a mosaic of specialized areas that are responsible for the analysis of a certain aspect of image. Rather, each visual modality is analyzed by groups of interconnected areas, with the feedforward and feedback interactions between areas being as important as the computations performed within each area. Moreover, each area can contribute to different types of analysis; for example, the information about direction of motion analyzed in MT can be used to detect the movement of an object, a person's self-motion, or the shape of a camouflaged vehicle moving against a background. Thus, the same computation performed in one area contributes to different functions, depending on interactions with other areas. The multilayered, massively parallel organization of visual cortex is probably what allows it to solve difficult computational problems, such as recognizing colors in different conditions of illumination or determining one's heading despite eye and head movements, in real time.

See Also the Following Articles

AREA V2 • COLOR PROCESSING AND COLOR PROCESSING DISORDERS • OBJECT PERCEPTION • SALIENCE • SPATIAL COGNITION • VISION: BRAIN MECHANISMS • VISUAL DISORDERS • VISUAL SYSTEM DEVELOPMENT AND NEURAL ACTIVITY

Acknowledgments

The author acknowledges Rowan Tweedale, James Bourne, and Ramesh Rajan, who reviewed previous versions of the manuscript.

Suggested Reading

Andersen, R. A., Snyder, L. H., Bradley, D. C., and Xing, J. (1997). Multiple representations of space in the posterior parietal cortex and its use in planning movement. *Annu. Rev. Neurosci.* **20,** 303–330.

Gilbert, C. D. (1992). Horizontal integration and cortical dynamics. *Neuron* **9,** 1–13.

Grüsser, O. J., and Landis, T. (1992). The visual world shrinks: Visual field defects, hemianopias and cortical blindness. In *Vision and Visual Dysfunction, Vol. 12: Visual Agnosias and Other Disturbances of Visual Perception and Cognition* (O. J. Grüsser and T. Landis, Eds.), pp. 136–157. Macmillan, London.

Lund, J. S., Wu, Q., Hadingham, P. T., and Levitt, J. B. (1995). Cells and circuits contributing to functional properties in area V1 of macaque monkey cerebral cortex: Bases for neuroanatomically realistic models. *J. Anat.* **187,** 563–581.

Peters, A., and Rockland, K. S. (Eds.) (1994). *Cerebral Cortex, Vol. 10: Primary Visual Cortex in Primates*. Plenum, New York.

Rockland, K. S., Kaas, J. H., and Peters, A. (Eds.) (1997). *Cerebral Cortex, Vol. 12: Extrastriate Cortex in Primates*. Plenum, New York.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.* **19,** 109–139.

Tootell, R. B. H., Mendola, J. D., Hadjikhani, N. K., Ledden, P. J., Liu, A. K., Reppas, J. B., Sereno, M. I., and Dale, A. M. (1997). Functional analysis of V3A and related areas in human visual cortex. *J. Neurosci.* **17,** 7060–7078.

Wurtz, R. H., and Kandel, E. R. (2000). Perception of motion, depth and form. In *Principles of Neural Science*, (E. R. Kandel, J. H. Schwartz and T. M. Jessell, Eds.), 4th ed., pp. 548–571. McGraw-Hill, New York.

Visual Disorders

R. D. JONES and D. TRANEL
University of Iowa College of Medicine

GLOSSARY

achromatopsia Defect in color vision.

alexia Defect in reading.

dyslexia Developmental disability in reading.

Disorders of vision and higher visual functions are common in many patients with neurological disease and are also of considerable scientific and theoretical interest in cognitive neuroscience. The visual system is the most widely researched and best understood higher order sensory system in the brain, owing to decades of research at the levels of cells, physiology, anatomy, systems, and theory. This article is a necessarily selective review of some of the more salient, common, or scientifically interesting disorders of higher order visual functions. Our aim is to review essential neuroanatomy and relate this core knowledge to visual syndromes of clinical and scientific interest. In this context, several case examples of complex visual syndromes and illustrative case material are provided.

I. ANATOMY OF THE VISUAL SYSTEM

Our examination of the visual system begins at the level of gross anatomy. Essential gross anatomic features of the visual system are depicted in Fig. 1. Fibers from the retina form the optic nerve, which is approximately 50 mm in length, and flow posteriorly to the optic chiasm. The chiasm is located just below the third ventricle, close to the hypothalamus, and immediately anterior to the pituitary stalk. At the chiasm, fibers from the nasal side of each retina cross, whereas fibers from the temporal side of each retina remain ipsilateral to the eye from which they came. Posterolateral to the optic chiasm, the optic tract is formed. Fibers in the optic nerve follow from whence they arose in the retina: Superior fibers from the retina course in the superior part of the optic nerve, and inferior fibers from the retina follow an inferior tract in the optic nerve.

The optic tract courses posteriorly through the geniculate bodies, in the posterior ventral thalami. The tracts then form the optic radiations, which connect to primary visual cortex in occipital lobe. Three tracts can be discerned in these optic radiations. First, the superior tract emerges from the lateral geniculate, representing the superior fibers from the retina, and courses posteriorly through deep white matter in the parietal lobe and terminates in the superior aspect of the calcarine fissure. The central tract originates from the central portion of the retina (macula), courses through deep temporal and occipital white matter, and terminates immediately above and below the calcarine fissure. Finally, the lower tract originates in the lower part of the retina, courses through the lateral aspect of the geniculate, follows a wide arc over the temporal horn of the lateral ventricle (Meyer's loop), runs to a point approximately 50 mm posterior to the temporal pole, and terminates in the lower aspect of the calcarine fissure.

With this system, the orientation of objects to brain in the visual system is reversed. That is, objects that are shown to the left visual field only (and thus to the right

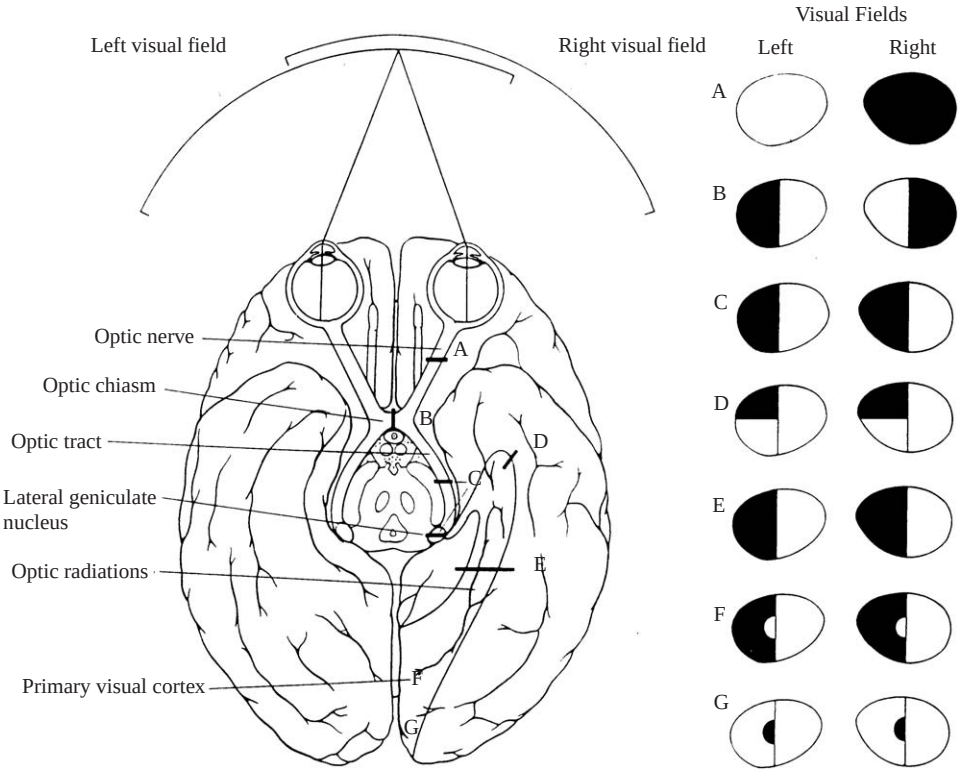


Figure 1 Depiction of the primary components of the visual system and field defects associated with lesions at various levels of the system.

side of the person's retina) are projected only to the right occipital lobe, and objects that appear in the person's right visual field (and thus to fibers in the left side of the retina only) project only to the left occipital lobe. Similarly, objects presented to the upper half of visual space (and thus to the lower half of the retina) are projected only to infracalcarine occipital lobe, whereas objects presented in the lower half of visual space (and thus to the superior aspect of the retinas) project only to supracalcarine occipital lobe.

The site of many lesions of interest in visual function defects can be specified with relative precision. For example, a defect confined to the right eye only would raise suspicion of lesion of the right optic nerve. In the case of *bitemporal hemianopia* (the inability to see on the temporal side of visual space in each eye), a lesion in the midline optic chiasm would be implicated. More commonly, a *homonymous hemianopia* (the inability to see in one-half of visual space in each eye) typically reflects a lesion in the contralateral optic tract, or, with sparing of macular vision, may reflect a contralateral lesion of the entire optic radiation. Finally, a *homonymous quadrantanopia* (the inability to see in one visual quadrant with each eye) reflects a lesion affecting the contralateral optic radiation, with lower (ventral) lesions associated with superior homonymous quadrantanopias and higher (dorsal) lesions associated with inferior homonymous quadrantanopias.

II. DISORDERS OF VISION

A. Pattern Recognition Defects

A broad dichotomy has been observed in the visual system. Specifically, the observation has been made that the visual system can be thought to be composed of two major segments: the superior (dorsal and occipitoparietal) system subserving visual constructive and spatial relationships and the inferior (ventral and occipitotemporal) system subserving object recognition. Lesions in these areas correspond to different identified syndromes, to which we now turn our attention.

1. Visual Agnosia

H. L. Teuber was credited in 1968 with the classic definition of agnosia—that is, a normal percept

stripped of its meaning. In the setting of visual agnosia, such a definition implies normal or near normal sensory perception in the context of impaired recognition of the object being perceived. In fact, given careful neuroophthalmologic and neuropsychological assessment, most patients with visual agnosia can be shown to have some abnormality of perception. With this in mind, it is useful to consider the distinction between "associative" and "apperceptive" agnosia made in 1890 by H. Lissauer. The former corresponds to a "pure" visual agnosia, as reflected in Teuber's definition. The latter, or *apperceptive visual agnosia*, refers to an abnormality of visual recognition in which perceptual features are so prominent as to dominate the clinical picture.

In the course of assessing patients with possible agnosia, it is important to distinguish between defects in naming and defects in recognition. This is a distinction that is commonly either blurred or completely unappreciated in the scientific and clinical literature on agnosia. The key is to understand that an object can be recognized without being named but cannot be named without being recognized. For example, a patient may not be able to produce the name "elephant" but will be able to provide other information that reflects normal recognition (e.g., it has a trunk and tusks, it is found at the zoo/at the circus/in Africa, and there is a story about one named Dumbo). Thus, if a patient is asked to identify an object and fails to produce its name, it is important to probe the patient regarding features of the object to establish that a true recognition defect exists.

2. Prosopagnosia

Prosopagnosia refers to an impairment in familiar face recognition. The term is attributed to Bodamer circa 1947 and derives from the Greek *prosop* (face) and *gnosis* (knowledge). This is a rare condition, but it is of substantial scientific interest.

The patient with prosopagnosia will typically complain of an inability to recognize familiar individuals around them and, in some cases, an inability even to recognize the patient's own face. Individual identities may be recognized by voice, gait, context, or other nonface cues, but the defect may be demonstrated by asking the patient to identify individuals in photographs in which contextual information is deleted. In patients with prosopagnosia caused by acquired brain damage (e.g., by stroke, tumor, or other lesion), there may be some visual field defect, usually a superior quadrantanopia of one or both visual fields. Such patients have otherwise normal neurologic and neuropsychological exams and do not evidence aphasia, amnesia, dementia, or disorders of executive functions. Visual acuity and visual perception may be normal or near normal in associative prosopagnosia, although many cases with clear visual perceptual defects have been described (and probably correspond to an apperceptive type).

In the past 15 years, a series of studies have shed light on a number of the underlying processes involved in prosopagnosia. For example, it has been shown that patients with prosopagnosia demonstrate nonconscious recognition of familiar faces. In one study, patients were shown to have selective electrodermal (skin conductance) responses in relation to familiar but not to unfamiliar faces. These skin conductance responses were elicited to familiar faces despite the fact that the patients were unable to recognize or identify familiar faces. Furthermore, it was demonstrated that patients with prosopagnosia have normal recognition of facial expression, gender, and age based on the perception of faces.

Closer examination of patients with acquired prosopagnosia shows that the recognition defect often extends to objects other than faces. Specifically, patients with prosopagnosia may have difficulty recognizing other objects, provided (i) there are numerous members of the category of the object to be recognized, (ii) there is visual ambiguity (shape similarity) among the objects in the category, and (iii) successful recognition depends on identification at the unique level. Thus, for example, a patient with prosopagnosia may have difficulty recognizing a specific automobile since there is significant visual similarity among exemplars of the category "automobiles," there are numerous members of the category, and the patient is required to identify a specific exemplar of the category.

The lesion associated with acquired prosopagnosia typically involves the occipitotemporal cortex. For the associative type, the lesion is typically bilateral, and the condition is often acquired in stages. That is, particularly with vascular lesions, the patient's history will involve a unilateral occipitotemporal stroke followed some time later by a posterior circulation stroke in a homologous part of the brain in the other hemisphere. For prosopagnosia of the apperceptive type, the lesion may be either bilateral or, in some cases, unilateral. In the case of unilateral lesions associated with apperceptive prosopagnosia, the lesion is typically on the right, is typically large, and includes occipitotemporal cortex, underlying white matter, and may spread

anteriorly and superiorly to include other regions of visual association cortex (Brodmann's areas 18 and 19).

A number of associated clinical conditions can often be observed in prosopagnosic patients. For example, in the associative type of prosopagnosia it is common to see achromatopsia and alexia, whereas in cases of apperceptive prosopagnosia it is common to observe defects in constructional ability and visual attention. These associated conditions are discussed separately later. To merit the term prosopagnosia, however, the central feature of the patient's clinical presentation must be the face recognition defect.

In recent years, a third form of prosopagnosia has been described, termed "developmental prosopagnosia." *Developmental prosopagnosia* refers to a selective face recognition defect that begins in childhood and is not associated with any identifiable structural lesion. A handful of cases have been reported between 1976 and 2001. In general, such cases have shown significant apperceptive features, although one reported case was believed to be an essentially associative prosopagnosia. Similar to cases of acquired prosopagnosia, such patients recognize individuals by gait, voice, or other nonface features. In at least one well-studied case there were features of a broader visual object agnosia. There have been no reported defects in color vision, and neuroophthalmologic exams may be entirely normal. In those cases in which neuroimaging has been completed, no structural brain abnormalities have been identified. Functional imaging has not been done in such cases, but it would be a highly interesting addition to this body of literature. These patients appear to manifest the face recognition disability over many years: Although first identified in childhood, in at least one case the disorder persisted into adulthood. Similar to cases of acquired prosopagnosia, in at least one case nonconscious electrodermal recognition of familiar faces was demonstrated.

3. Defects in Judgment of Emotion and Social Knowledge Based on Face Cues

As noted previously, patients with prosopagnosia, particularly those of the associative type, often have normal judgments of facial emotional expressions. That is, despite their inability to identify the face, judgments related to facial emotional expressions (e.g., happy, sad, angry, and surprised) are similar to those of controls. Recently, the opposite dissociation has been described. In these studies, it was demonstrated that patients with bilateral damage to amygdala were unable to judge emotional facial expressions accurately, despite normal recognition of identity. Similarly, in another study from our laboratory, patients with bilateral amygdala damage were significantly impaired in judging faces that looked "unapproachable" or "untrustworthy." These findings were interpreted to support the proposition that the amygdala was involved in social judgments as well as judgment of emotion based on facial features. The findings indicate the complex, complimentary, and reciprocal relationship between multiple areas of brain—in this case, anatomically quite distinct—in the processing of different aspects of face information (e.g., identity, emotion, and social valence).

4. Visual Object Agnosia

Whereas prosopagnosia is associated with a defect in the identification of individual members of a class, the term *visual object agnosia* is reserved for patients who have difficulty with identification of entities at basic object level. Thus, visual object agnosia refers to a visual recognition defect at the level of nonunique objects rather than at the level of specific members of a category. For example, a patient with visual agnosia may not know that a violin is a violin, that a dog is a dog, or that a car is a car. Patients with visual object agnosia can identify objects by touch, sound, or characteristic movement, thereby ensuring that they know the objects. In fact, in some cases patients are able to accurately draw objects on command but are unable to identify or recognize such objects in the visual modality. This is a rare condition. Memory is normal, and there is no defect in judgment, basic verbal intellectual skills, or language.

Associated conditions often include defects in color recognition and naming and defects in reading. There may be associated visual field defects, although a primary field cut is not always present. Lesions associated with visual object agnosia are typically more extensive than those found in cases of prosopagnosia, including occipitotemporal cortex bilaterally, extending dorsally into visual association cortex, and almost always involving underlying white matter. There is evidence that the left hemisphere plays a more critical role than the right in the development of visual object agnosia.

5. Category-Specific Recognition Defects

In recent years, there has been a considerable amount of research aimed at the issue of differences in

recognition and naming related to different conceptual categories. A common profile, for example, is impaired recognition of living entities, with relative preservation of recognition of nonliving entities. A number of factors may contribute to this observed differentiation, and they are currently being explored. For example, there may be differences in shape, manipulability, or the means by which the entity is learned that operate on subsequent visual recognition of such entities. As an example of this type of research, it was reported in 1997 that defective recognition of persons was associated with lesions in the right temporal polar region, defective recognition of animals was associated with lesions in right mesial occipital/ventral temporal areas and also in the left mesial occipital region, whereas defective recognition of tools was associated with lesions in the left occipital–temporal–parietal junction.

B. Disorders of Reading

Pure alexia refers to the inability to read with preserved ability to write. The term pure alexia refers to an acquired inability rather than a developmental disability in reading (the latter is known as *developmental* dyslexia). This appears to be an apparently bizarre dissociation, one that may evoke the misdiagnosis of a psychiatric disease. These patients are not aphasic, nor is there a defect in memory, orientation, basic intellectual skills, or verbal arithmetic skills. The neurologic exam is normal apart from a typical finding of a right homonymous hemianopia.

In evaluating patients with pure alexia, it is useful to ask them to copy printed material, to write a sentence from dictation, and to write a sentence spontaneously. The latter two tasks are typically completed with relative ease. However, such patients will have difficulty in copying sentences, given their inability to read. However, we have seen patients who are, in fact, able to copy lexical material (written words) accurately, but on questioning these patients it is apparent that they do not appreciate the fact that these are letters or words; rather, they are viewed as a series of geometric shapes. Although such patients are unable to read, if the examiner asks the patient to trace letters with the patient's finger, the patient may be able to discern the meaning of the written word. Furthermore, when visual presentation of the word is enlarged to subsume a large part of the visual field, the patient may be able to read. The mechanism that enables such patients to decode letters and words in a sensorimotor modality is outlined later and may be of particular relevance to rehabilitation efforts.

The lesion associated with pure alexia is in the left occipital lobe and underlying white matter. Such a lesion produces the right homonymous visual field defect that is often found in such cases. The lesion typically abuts the splenium. Thus, visual information is registered only by the right occipital lobe, and such information is blocked from passing to the left hemisphere by virtue of the retrosplenial lesion. This information thus cannot be decoded into lexically meaningful words by association areas in the left parietal cortex, specifically surrounding the left angular gyrus. However, in the instances described previously, when a patient is asked to trace letters with his or her hand, there is no such disconnection between sensorimotor cortex on the left (receiving input from and sending input to the right hand) and the left angular gyrus. Thus, lexically meaningful information can be derived and decoded by connection between sensorimotor cortex on the left (responding to hand movement) and parietal association cortex (decoding motor movements into lexically meaningful units/letters). The key in this case is that the visual system is bypassed. Similarly, in the situation in which words are enlarged to subsume much of the visual field, the motor system of the ocular muscles is invoked; thus, motor movements can be used to decode visual stimuli into lexically meaningful parts.

An interesting but undeveloped area of research in pure alexia has to do with recognition of different levels of lexical material. For example, it has been shown that patients with pure alexia are able to identify words versus nonwords in a lexical decision task. This finding raises questions regarding other aspects of lexical decoding that may be preserved in patients with pure alexia. For example, it is possible that patients with pure alexia will show nonconscious electrodermal evidence of recognition of words. Such patients may also show emotional reactions to words with high emotional valence, as evidenced by psychophysiological measures, in the face of impaired overt recognition. Such questions have not been explored in-depth but are of substantial scientific interest.

1. Case Report: Alexia and Visual Object Agnosia

Case RS is a left-handed, 59-year-old male author and attorney with 20 years of education. He was seen in our laboratory approximately 10 months following the

first of two strokes. Specifically, he experienced sudden visual distortion and rapidly building headache. He noted a subsequent inability to read, although his ability to write, speak, and understand spoken language was normal. The remainder of the neurological exam was normal, and an extensive assessment confirmed a left occipitotemporal stroke. Five months later he experienced sudden onset of left hand, arm, and face numbness which resolved over the subsequent several hours. His ability to read, which had begun to recover, was again lost. At the time of this second event he also experienced difficulty in recognizing other visual objects and had distorted color perception, stating that colors seemed to be "metallic" and "unreal." His impairment in color perception and a mild face recognition defect resolved approximately 3 weeks later. Repeat neuroimaging confirmed a second occipitotemporal stroke, this time in the right hemisphere.

As can be seen in Table I, the patient was intellectually gifted. Basic visual abilities (acuity, contrast sensitivity, and color plates), and visual attention were normal, and visual fields were full. Visual recognition of objects was severely impaired. However, auditory recognition of objects was normal, as was recognition of facial expressions of emotion.

Magnetic resonance imaging (MRI) (Fig. 2) revealed an approximately 3.0×1.5 cm infarct in the left inferior posterior temporal lobe, extending from the surface to the inferior aspect of the lateral ventricle. A second, smaller infarct, measuring approximately 1.0×1.0 cm, involving primarily the cortex, was seen in the right inferior posterior temporal lobe. The first lesion, on the left, accounted for the patient's initial descriptions of dyschromatopsia and pure alexia. The second lesion resulted in visual object agnosia.

A substantial portion of the patient's occupation involved reading; thus, remediation of this ability was the focus of rehabilitation. Two visual dimensions of lexical material were systematically varied in an effort to optimize his speed and accuracy of reading. Specifically, the size of font of letters was systematically tested until both speed and accuracy were optimized, and then it was made smaller in gradual steps until a size was found that was believed to provide the best balance between speed, accuracy, and practicality (very large fonts were impractical for everyday reading). Second, various color contrasts between background and letters were used, again in an effort to maximize speed and accuracy of reading. With these aids, the patient eventually returned to work.

C. Disorders of Color Processing

1. Achromatopsia

Achromatopsia refers to a loss of color vision due to an acquired brain disorder. Patients commonly complain that colors appear to be black and white, washed out, gray, or dirty. Such patients may also demonstrate some color desaturation and dyschromatopsia rather than complete lack of color. Lesions associated with achromatopsia are characteristically found in the infracalcarine medial occipital lobe, typically involving the lingual and fusiform gyri. When the lesion is unilateral, the color defect is in the contralateral visual space. When the lesion is bilateral, the color defect subsumes the entire visual field. There may well be visual field cuts, typically involving superior quadrants.

2. Color Anomia

The term *color anomia* refers to an inability to name colors in the context of normal performances on tests of matching color or of color generation. This disorder is usually associated with pure alexia and right homonymous hemianopia. Such patients are able to match same-color chips and associate colors with objects (e.g., "A banana is… yellow"). However, when shown a specific color, such patients are unable to name it with an appropriate verbal tag. This defect seems to be exacerbated when less contextual information is available. For example, such patients may be able to see a yellow banana and say that the color is yellow but may be unable to name a spot of yellow.

Given the site of the lesion in color anomia and the preponderance of right homonymous hemianopia and pure alexia in such cases, there is some suggestion that this may represent a disconnection syndrome. That is, in the left visual field (ipsilateral to the lesion) color perception is normal, but perception is limited to primary visual cortex in the right hemisphere only. This elementary perceptual information is blocked by a retrosplenial lesion in the mesial left occipital lobe, thereby rendering such information unavailable to the left hemisphere.

3. Color Agnosia

Color agnosia is a rare condition defined as an inability to retrieve color information in the context of normal perception and language. For example, a patient with color agnosia will be unable to remember the colors of

Table I
Standard Neuropsychological Assessments for RS2206[a]

	Score	Interpretation
Intellect		
WAIS verbal IQ (mean = 100 ± 15)	148 (WAIS-III)	Superior
WAIS performance IQ	111	High average
WAIS full-scale IQ	135	Superior
Anterograde memory		
AVLT		
Trial 5 (#/15)	8	Impaired
Delayed recall (#/15)	2	Impaired
CFT, delayed recall (#/36)	16.5	Borderline
BVRT, correct/errors	7/6	Borderline
RMT		
Words (#/50)	48	High average
Faces (#/50)	39	Low average
Visual perception		
FRT (percentile)	15th	Low average
JLO (percentile)	> 74th	High average
Hooper VOT (*t* score; mean = 50 ± 10)	93	Impaired
Visual construction		
WAIS block design (ACSS mean = 10 ± 3)	14	Superior
WAIS object assembly (ACSS mean = 10 ± 3)	6	Low average
CFT (#/36)	35	Normal
Language		
MAE–COWA (percentile)	80th	High average
BDAE reading (#/10)	10	Normal
BNT (#/60)	56	Normal
Visuomotor coordination		
WAIS digit symbol (ACSS mean = 10 ± 3)	12	High average
TMT		
Part A (*t* scores; mean = 50 ± 10)	34	Borderline
Part B	36	Borderline
Experimental visual naming[b]		
Animals		
Recognition (% correct)	71	Impaired
Naming (% correct)	94	Normal
Fruits		
Recognition (% correct)	33	Impaired
Naming (% correct)	86	Normal
Tools		
Recognition (% correct)	73	Impaired
Naming (% correct)	92	Normal
Famous faces		
Recognition (% correct)	86	Normal
Naming (% correct)	81	Normal

[a]Abbreviations used: WAIS, Wechsler Adult Intelligence Scale; AVLT, Auditory Verbal Learning Test; CFT, Complex Figure Test; BVRT, Benton Visual Retention Test; RMT, Recognition Memory Test; FRT, Facial Recognition Test; JLO, Judgment of Line Orientation; Hooper VOT, Visual Orientation Test; CFT, Complex Figure Test; MAE, Multilingual Aphasia Examination; COWA, Controlled Oral Word Association; BDAE, Boston Diagnostic Aphasia Examination; BNT, Boston Naming Test; TMT, Trail Making Test.

[b]Experimental naming scores are listed as percentage named of those items correctly recognized.

common entities (e.g., "What color is a banana?") and will be unable to provide a list of objects that come in various colors (e.g., "Name things that are red"). Based on a handful of well-studied patients, the neuroanatomical correlates of color agnosia include the occipital–temporal region either unilaterally on the left or bilaterally. To date, this lesion localization provides no clear separation from what might be expected with achromatopsia; however, functional imaging studies suggest that areas associated with color knowledge are probably anterior and lateral to areas associated with color perception. Color agnosia is often associated with visual object agnosia and may be associated with category-related recognition defects, particularly for living entities as opposed to artifactual entities.

4. Case Report: Prosopagnosia, Pure Alexia, and Achromatopsia

A 72-year-old right-handed retired female schoolteacher with 18 years of education suffered a myocardial infarction 5 years prior to her clinical presentation. She had a four-vessel coronary artery bypass graft at the time of her heart attack and suffered a right occipitotemporal cardioembolic stroke with resultant left homonymous visual field defect. Approximately 1 month prior to her presentation in our clinic she suffered a sudden change in color vision, noting that colors in traffic lights were all gray. She also experi-

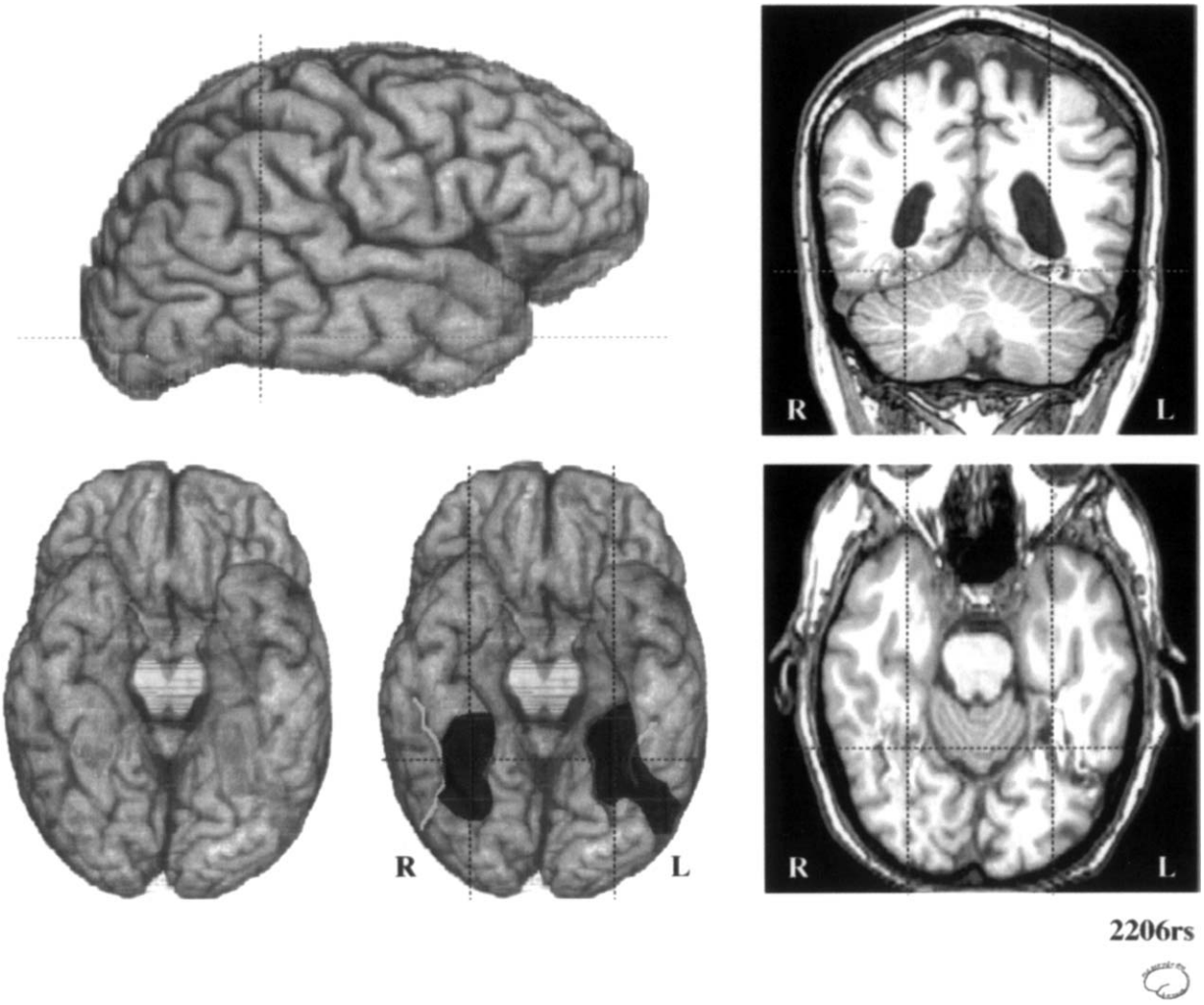


Figure 2 Three-dimensional MRI reconstruction of the brain of a patient with visual object agnosia.

enced vertigo, ataxia, and facial recognition defects. Motor and sensory signs were normal, but she had a wide-based and ataxic gait.

Results of neuropsychological assessment are reflected in Table II. The patient had prosopagnosia (of the associative type), pure alexia, and achromatopsia. Visual acuity was normal with correction, and visual perception of faces was normal (e.g., performance was normal on Benton's Facial Recognition Test). She was able to describe accurately the gender, facial expression, and age of faces in photographs but was unable to indicate the identity of faces. The MRI of the brain is shown in Fig. 3. As can be seen, the patient had bilateral occipitotemporal lesions that accounted for her neuropsychological defects. Basic intellectual skills and language were preserved, as were orientation and attention. Visual spatial judgment was normal.

Her principal concerns were her inability to read (pure alexia) and the face recognition defect. She compensated for her prosopagnosia by identifying individuals by their gait, voice, or, in some cases, a salient visual feature (e.g., baldness). In an effort to identify her favorite family photographs, she wrote the names of the subject of each individual photograph on the back of the picture in large type, and she could read the names by tracing the letters with her finger. Although her reading never returned to baseline levels, she made significant progress by using a magnifying glass and then tracing the letters she wanted to read and thereby invoking the sensorimotor system. In another example of the measures she took to compensate for her pure alexia, she memorized motor patterns associated with emergency telephone numbers (911) and frequently called friends and relatives. She lived

Table II
Standard Neuropsychological Assessments for RM1500[a]

	Score	Interpretation
Intellect		
WAIS-R verbal IQ (mean = 100 ± 15)	91	Average
WAIS-R performance IQ	83	Low average
WAIS-R full scale IQ	86	Low average
Anterograde memory		
AVLT		
Trial 5 (#/15)	9	Borderline
Delayed recall (#/15)	2	Impaired
CFT, delayed recall (#/36)	7	Impaired
BVRT, correct/errors	6/7	Normal
RMT		
Words (#/50)	36	Borderline
Faces (#/50)	31	Impaired
Visual perception		
FRT (percentile)	22nd	Low normal
JLO (percentile)	> 74th	High normal
Hooper VOT (*t* score; mean = 50 ± 10)	70	Impaired
Visual construction		
WAIS block design (ACSS mean = 10 ± 3)	11	Average
WAIS object assembly (ACSS mean = 10 ± 3)	8	Average
CFT (#/36)	25	Normal
Language		
MAE–COWA (percentile)	72nd	Average
BDAE reading (#/10)	8	Normal
Visuomotor coordination		
WAIS-R digit symbol (ACSS mean = 10 ± 3)	7	Low average
Experimental visual naming[b]		
Animals		
Recognition (% correct)	57	Impaired
Naming (% correct)	95	Normal
Fruits		
Recognition (% correct)	59	Impaired
Naming (% correct)	95	Normal
Tools		
Recognition (% correct)	54	Impaired
Naming (% correct)	91	Normal
Famous faces		
Recognition (% correct)	0	Impaired
Naming (% correct)	0	Impaired

[a]Abbreviations used: WAIS, Wechsler Adult Intelligence Scale; AVLT, Auditory Verbal Learning Test; CFT, Complex Figure Test; BVRT, Benton Visual Retention Test; RMT, Recognition Memory Test; FRT, Facial Recognition Test; JLO, Judgment of Line Orientation; Hooper VOT, Visual Orientation Test; CFT, Complex Figure Test; MAE, Multilingual Aphasia Examination; COWA, Controlled Oral Word Association; BDAE, Boston Diagnostic Aphasia Examination.

[b]Experimental Naming scores are listed as percent named of those items correctly recognized.

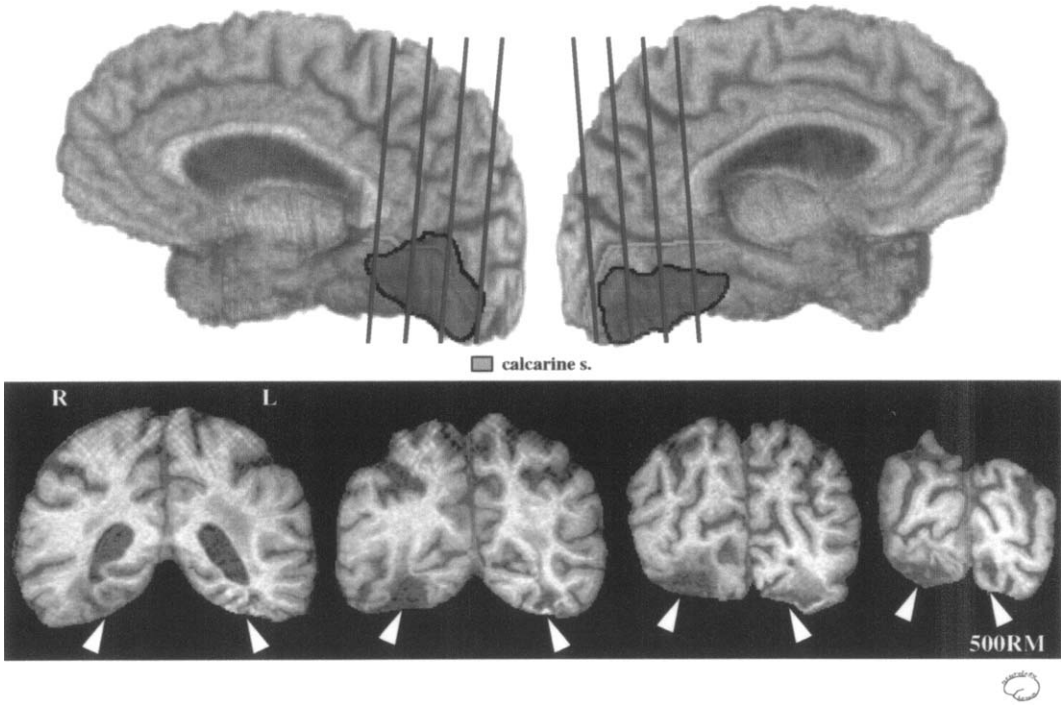


Figure 3 Three-dimensional MRI reconstruction of the brain of a patient with prosopagnosia, pure alexia, and achromatopsia.

alone successfully and walked several blocks each day to visit her husband in a nursing home nearby. She would count the blocks as she walked so as not to become lost.

D. Disorders of Spatial Analysis

1. Balint's Syndrome

The symptom complex necessary for the diagnosis of Balint's syndrome includes simultanagnosia (also known as visual disorientation), optic ataxia, and ocular apraxia.

Simultanagnosia refers to the inability to synthesize separate features of the visual field into a meaningful whole. This is the core feature of Balint's syndrome. Establishing this aspect of the syndrome necessarily requires the self-report of the patient. Typically, patients will describe features of objects as "coming in and out of view" or "sometimes there and sometimes not." Perception of an object in the visual field will often be limited to small parts of any given object. For example, a patient recently studied in our laboratory was shown a line drawing of a Christmas wreath and reported that the only thing she saw was a bow. In fact, there was a bow attached to the Christmas wreath, but she was unable to perceive the whole and unable to provide a synthesized meaningful description of the disparate parts available to her. Typical misperceptions are based on parts of objects, and the astute clinician can determine the features of objects that are used by the patient to generate responses.

The second essential feature for the diagnosis of Balint's syndrome is optic ataxia. *Optic ataxia* refers to an inability to point to targets under visual guidance. Examination of this sign consists of asking the subject to point to or touch the finger of the examiner in a rapid ballistic movement. Patients with optic ataxia will often miss the target or will slowly approach the target with their hand. Optic ataxia can be observed further when patients are reaching for any object in space (e.g., a pencil or food). It is common for such patients to approximate the target and then "crawl" with their fingers to reach the desired object. It is necessary that such patients have difficulty under *visual* guidance. For example, with eyes closed, a patient with optic ataxia can point as accurately as controls to the source of a sound. Optic ataxia is also known as visual motor ataxia and can be found in some patients in isolation from simultanagnosia and ocular apraxia.

Ocular apraxia can be defined as an inability to shift gaze toward a new visual stimulus. At the bedside, the

patient can be requested to fixate on various objects in the immediate environment, and the examiner should take note of the ease with which this is done. Patients with ocular apraxia will show ocular movements that reflect an inability to shift their gaze to a new object and will approximate their gaze and slowly come to appropriate ocular fixation. The saccadic movements of such patients in focusing on new objects may be completely inaccurate or may be much longer and less accurate than those of normal subjects. This third sign of Balint's syndrome is also known as *psychic gaze paralysis*.

The lesion associated with Balint's syndrome has reliably been reported to be in the occipitoparietal visual association cortex (the upper part of Brodmann's areas 18 and 19). A bilateral lesion is necessary to observe the syndrome fully. This is a watershed zone between middle and posterior cerebral arteries; therefore, the lesion is often associated with the medical phenomenon of sudden hypotension, such as in cardiac arrest. Thus, this is one of the defects associated with the dorsal visual stream. Occassionally, area 17 or area 39 may be involved. A field defect is not necessary, but an inferior field cut may be present in some cases.

2. Case Report: Balint's Syndrome

LA is a right-handed 52-year-old female factory worker with a high school education. Approximately 2 years prior to her clinical presentation she began to notice changes in her vision and had sought consultation for new glasses three times in the 6 months preceding our first examination. Neuroophthalmalogic exam was normal, visual acuity was 20/20 (corrected), and visual fields were full. Her chief complaint was that she was having difficulty at work. Her job involved placing objects in a particular orientation in a box, and she had been having increasing difficulty in completing this task. Prior to this job she had been a grocery store manager but had experienced progressive difficulty with using cash registers, using adding machines, keeping invoice records, and other jobs that demanded execution of serial visually guided tasks. She had been fired after several successful years at this job. She described objects as looking ''shattered'' and ''coming and going'' from her sight.

Neuropsychological assessment (Table III) revealed severe defects in constructional praxis. She was oriented, conversational speech was normal, and verbal memory was normal. The presence of visual disorientation (simultanagnosia) was established by extensive testing. For example, in evaluating her responses to specific naming items it became apparent that she was processing only parts of the larger stimulus. For example, she said, ''ribbon'' when shown a line drawing of a wreath with a ribbon on it. Similarly, she said ''roof'' when shown a pyramid; on follow-up questions she explained that she saw one corner of the pyramid, which she then presumed to be a roof.

The presence of optic ataxia was established both by examination and by clinical observation. As an example of the latter, she had difficulty reaching for pencils on the desk in front of her. Also, on formal testing, she was markedly inaccurate and slow in trying to reach for the examiner's hand under visual guidance. Finally, it is noteworthy that her history of job difficulties suggests the presence of optic ataxia. For example, she had difficulty with ballistic hand movements under visual guidance in many of her job responsibilities.

Ocular apraxia was established by clinical test and observation. She was unable to make effective saccadic movements on command when asked to look at various objects in the room.

Functional and structural neuroimaging were completed. MRI showed bilateral occipitoparietal atrophy, although there was evidence of mild atrophy in other areas. Position emission tomography studies showed bilateral occipitoparietal hypometabolism. The patient was ultimately diagnosed with an atypical progressive dementia, though possibly of the visual variant Alzheimer's type. From a neuropsychological perspective, the principal diagnosis was Balint's syndrome.

3. Visual Neglect

Neglect can be defined as ''the failure to report, respond, or orient to novel or meaningful stimuli presented to the side opposite a brain lesion, when this failure cannot be attributed to either sensory or motor defects.'' In the visual modality, the most common form of neglect is hemispatial neglect. The essential feature of spatial neglect is that the patient ignores objects in the visual hemifield contralateral to the lesion, in the absence of a visual field cut. Such neglect is commonly associated with right hemisphere lesions, resulting in left hemispatial visual neglect. Right hemispatial visual neglect, though rare, can be seen during the acute period in some cases of left hemisphere lesions. The typical lesion is in the occipital–temporal–parietal border zone. The finding of neglect

Table III
Standard Neuropsychological Assessments for LA2354[a]

	Score	Interpretation
Intellect		
WAIS-III verbal IQ (mean = 100 ± 15)	76 (WAIS-III)	Borderline
WAIS-III performance IQ	—	
WAIS-III full scale IQ	—	
Anterograde memory		
AVLT		
Trial 5 (#/15)	10	Normal
Delayed recall (#/15)	7	Normal
CFT, delayed recall (#/36)	0	Impaired
BVRT, correct/errors	0/11	Impaired
Visual perception		
FRT (percentile)	2nd	Impaired
JLO (percentile)	< 1st	Impaired
Hooper VOT (*t* score; mean = 50 ± 10)	93	Impaired
Visual construction		
WAIS-III block design (ACSS mean = 10 ± 3)	3	Impaired
WAIS-III object assembly (ACSS mean = 10 ± 3)	3	Impaired
CFT (#/36)	5	Impaired
Language		
MAE–COWA (percentile)	3rd	Impaired
BNT (#/60)	36	Impaired
Visuomotor coordination		
TMT		
Part A (*t* scores; mean = 50 ± 10)	< 20	Impaired
Part B	< 20	Impaired
Experimental visual naming[b]		
Animals		
Recognition (% correct)	50	Impaired
Naming (% correct)	89	Normal
Fruits		
Recognition (% correct)	81	Borderline
Naming (% correct)	85	Normal
Tools		
Recognition (% correct)	77	Impaired
Naming (% correct)	89	Normal
Famous faces		
Recognition (% correct)	67	Impaired
Naming (% correct)	50	Impaired

[a]Abbreviations used: WAIS, Wechsler Adult Intelligence Scale; AVLT, Auditory Verbal Learning Test; CFT, Complex Figure Test; BVRT, Benton Visual Retention Test; RMT, Recognition Memory Test; FRT, Facial Recognition Test; JLO, Judgment of Line Orientation; Hooper VOT, Visual Orientation Test; CFT, Complex Figure Test; MAE, Multilingual Aphasia Examination; COWA, Controlled Oral Word Association; BNT, Boston Naming Test; TMT, Trail Making Test.

[b]Experimental naming scores are listed as percent named of those items correctly recognized.

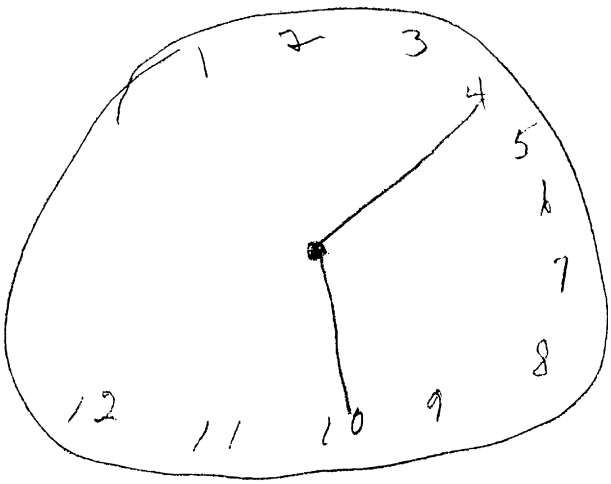


Figure 4 Drawing of a clock showing left hemispatial visual neglect by a 70-year-old right-handed female with a right parietal tumor.

is often not restricted to the visual modality and may be demonstrated in both sensorimotor and auditory systems concurrently with visual neglect.

There are a number of ways to evaluate hemispatial visual neglect in affected patients. One of the more common bedside assessments is to perform a double simultaneous stimulation task, in which the examiner introduces a visual stimulus to the left visual field, to the right visual field, and then to both fields simultaneously. The most clear evidence of neglect from this procedure is that the patient will identify each stimulus individually, but when presented simultaneously the patient will identify only the stimulus presented ipsilateral to the lesion. Furthermore, methods of detecting hemispatial neglect include drawing tasks, line bisection tasks, or line cancelation tasks (Fig. 4). In addition to demonstrating neglect in other modalities (sensorimotor and auditory), associated signs observed in neuropsychological testing of patients with neglect include constructional dyspraxia, anosodiaphoria, and anosognosia.

4. Cortical Blindness and Anton's Syndrome

Bilateral lesions of primary visual cortex result in *cortical blindness*, a condition commonly caused by bilateral infarction of the posterior cerebral arteries. A commonly associated phenomenon with cortical blindness is *Anton's syndrome*, in which the patient with cortical blindness denies all visual difficulties. Anton's syndrome is considered a special form of anosognosia and is frequently seen only in the acute epoch (within 2 or 3 months following the event). Cortical blindness and Anton's syndrome may also result from subcortical lesions to optic radiations. Pupillary responses are preserved, but visual evoked potentials cannot be demonstrated.

5. Topographical Disorientation

The term topographical disorientation refers to an acquired inability to navigate the environment in daily life. This can be conceptualized as a memory defect in the visual realm, but a number of mechanisms of the agnosias can be seen in this disorder. This disorder is highly relevant in patient care and should be considered by clinicians in placement and management of affected patients. Two broad types of topographical disorientation can be discerned. The first may be termed anterograde topographical disorientation, which involves a defect in new learning. Patients with this disorder will complain that they become lost in new places, cannot learn new routes, or cannot retrace their paths in real life. Associated clinical findings may include left hemispatial neglect, left visual field cuts, and constructional dyspraxia. Such patients are not aphasic, nor do they show defects in executive functions or attention. Visual memory may be impaired on formal neuropsychological testing, but verbal memory is typically normal. The site of lesion associated with the disorder is the medial occipitotemporal area, particularly on the right and including the posterior parahippocampal gyrus and lingual gyrus. Anterograde topographical agnosia has been described in cases of right parietal dysfunction and of left mesial occipitotemporal dysfunction, but in these cases the clinical phenomenon appears to be more transient and less severe.

Retrograde topographical disorientation refers to the inability to recognize previously well-known spatial/topographical routes. The patient with retrograde topographical disorientation will be unable to recognize either previously known visual scenes (e.g., rooms in their home and the street on which they lived) or landmarks (e.g., their house and the local grocery). To distinguish this disorder from severe apperceptive defects it is important to note that patients with retrograde topographical disorientation are unable to describe in verbal terms from memory familiar scenes or routes. The lesion most commonly associated with this disorder is in the mesial occipitotemporal cortices of the right hemisphere, thus disconnecting primary visual cortex from higher order association cortices.

6. Defects in Constructional Skills

Constructional ability can be defined as the ability to manipulate materials to form an intended final construction. This ability is known well to most practitioners; examples of tests of constructional ability are the Draw a Clock Test, the block design subtest from the Wechsler Adult Intelligence Test, and the intersecting pentagons from the Mini Mental State Examination. Tests of constructional ability presume that the patient has adequate visual acuity to see the test and that motor skills are adequate to complete the test.

The mechanism whereby a patient manifests defects in constructional abilities may be determined by examination of the process of construction or by examination of the final product. For example, visual neglect may be determined by a drawing task in which the item to be drawn (or copied) is placed far to the side of the page (usually the right side, indicating left hemispatial visual neglect). Hemispatial neglect may also be shown in a case in which one side of the object is distorted or missing. Finally, in the absence of neglect, there may be a lack of appreciation of visual spatial relationships apparent in the production of a drawing by the patient, in which elements of the object to be constructed are present but are distorted in their spatial relationship to one another.

The site of lesion associated with defects in constructional skills is typically the right occipital cortices and adjacent visual association cortex. There may be a greater perceptual component to the defect with more ventral lesions and a greater defect of spatial relationships with more dorsal lesions. In some cases, constructional disabilities can be observed in left-sided lesions in these same sites, but such patients often have a fluent aphasia, a lesser or more transient constructional defect, and the defect is more likely to show a lack of appreciation of spatial relationships and preserved visuoperception.

7. Deficits of Mental Imagery

Visual mental imagery refers to the evocation of visual images in the "mind's eye" in the absence of sensation through the retina. In recent years, research on this topic has resulted in substantial knowledge about the neuroanatomical underpinnings of visual imagery; however, much has yet to be learned.

Early studies demonstrated that patients with various types of deficits in higher order visual functions (e.g., achromatopsia and spatial neglect) tended to show similar deficits in mental imagery. For example, when asked to imagine walking a down a familiar street, a patient with left hemispatial visual neglect would describe only the buildings on the right side of the street. When asked to imagine turning around and walking the other direction up the street, the same patient would describe buildings on the other side of the street that had previously been neglected. Similarly, patients with deficits in color perception due to cerebral lesions have been shown to have deficits in imagining colors of objects. Although not all findings have supported such observations, there appears to be broad general consensus for the idea that visual perceptual deficits in patients can often be demonstrated in some way by similar deficits in visual imagery in these patients.

The neural underpinnings of deficits in visual imagery have been located primarily in visual association cortex (areas 18 and 19) and more anterior association cortices. However, recent data from functional MRI studies suggest that primary visual cortex (area 17) may be more actively involved in mental imagery than had previously been suspected.

III. CONCLUSION

Although disorders of higher visual functions are varied, the visual system as a whole is one of the best understood systems in the brain, due in part to decades of observation of and research on these functions. The visual agnosias, in particular, have been shown to be an area of rich findings in basic brain–behavior relationships. History has demonstrated that such basic findings often serve as the basis of rehabilitation and management of such disorders, an area that is now in its infancy.

See Also the Following Articles

AGNOSIA • ALEXIA • ANOMIA • COLOR PROCESSING AND COLOR PROCESSING DISORDERS • DYSLEXIA • PATTERN RECOGNITION • PROSOPAGNOSIA • READING DISORDERS, DEVELOPMENTAL • UNILATERAL NEGLECT • VISION: BRAIN MECHANISMS • VISUAL CORTEX • VISUAL SYSTEM DEVELOPMENT AND NEURAL ACTIVITY

Acknowledgments

This work was supported by NINDS Program Project Grant NS 19632. The authors are grateful to Thomas J. Grabowski and the Human Neuroanatomy and Neuroimaging Laboratory at the

University of Iowa for providing Figs. 2 and 3, to John Martin for Fig. 1, and to Ken Manzel for collection of data for the tables.

Suggested Reading

Adolphs, R., Tranel, D., Damasio, H., and Damasio, A. R. (1994). Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. *Nature* **372,** 669–672.

Adolphs, R., Tranel, D., Damasio, H., and Damasio, A. R. (1995). Fear and the human amygdala. *J. Neurosci.* **15,** 5879–5891.

Adolphs, R., Tranel, D., and Damasio, A. R. (1998). The human amygdala in social judgment. *Nature* **393,** 470–474.

Ariel, R., and Sadeh, M. (1996). Congenital visual agnosia and prosopagnosia in a child: A case study. *Cortex* **32,** 221–240.

Barrash, J. (1998). A historical review of topographical disorientation and its neuroanatomical correlates. *J. Clin. Exp. Neuropsychol.* **20,** 807–827.

Damasio, A. R., Tranel, D., and Rizzo, M. (2000). Disorders of complex visual processing. In *Principles of Behavioral and Cognitive Neurology* (M. M. Mesulam, Ed.), pp. 332–372. Oxford Univ. Press, New York.

De Haan, E. H. F. (1999). A familial factor in the development of face recognition deficits. *J. Clin. Exp. Neuropsychol.* **21,** 312–315.

De Haan, E. H. F., and Campbell, R. (1991). A fifteen year follow-up of a case of developmental prosopagnosia. *Cortex* **27,** 489–509.

Farah, M. J. (2001). The neuropsychology of mental imagery. In *Handbook of Neuropsychology* (F. Boller and J. Grafman, Eds.), 2nd ed., Vol. 4**,** pp. 239–248. Elsevier, Amsterdam.

Heilman, K. M., Watson, R. T., and Valenstein, E. (1993). Neglect and related disorders. In *Clinical neuropsychology* (K. M. Heilman and E. Valenstein, Eds.), 3rd ed., pp. 279–336. Oxford Univ. Press, New York.

Jones, R., and Tranel, D. (2001). Severe "associative" developmental prosopagnosia in a child with superior intellect. *J. Clin. Exp. Neuropsychol.* **23,** 265–273.

Kracke, I. (1994). Developmental prosopagnosia in Asperger syndrome: Presentation and discussion of an individual case. *Dev. Med. Child Neurol.* **36,** 873–886.

Temple, C. M. (1992). Developmental memory impairment: Faces and patterns. In *Mental Lives: Case Studies in Cognition* (R. Campbell, Ed.), pp. 199–215, Blackwell, Oxford.

Tranel, D. (2001). Central color processing and its disorders. In *Handbook of Neuropsychology* (F. Boller and J. Grafman, Eds.), 2nd ed., Vol. 4**,** pp. 1–14. Elsevier, Amsterdam.

Tranel, D., Damasio, H., and Damasio, A. R. (1997). A neural basis for the retrieval of conceptual knowledge. *Neuropsychologia* **35,** 1319–1324.

Visual System Development and Neural Activity

A. E. WIENCKEN-BARGER and V. A. CASAGRANDE
Vanderbilt University School of Medicine

GLOSSARY

amblyopia The impairment of vision without detectable organic lesion of the eye.

ectopic Positioned abnormally within the body.

neurotrophic factor A molecule, usually a protein, that will facilitate the growth or repair of nerve cells.

ocular dominance column An area in the visual cortex that receives input predominantly from one eye.

transcription factor A protein required for recognition by RNA polymerases of specific stimulatory sequences in eukaryotic genes.

visuotopy The arrangement of cells and the connections between neural structures such that they maintain a topographic representation of the visual field.

Neural structures are specified very early when the future nervous system is still a sheet of cells referred to as the neural plate and before this sheet folds to form the neural tube. These early steps involve evolutionarily conserved inductive signaling pathways that initially establish regional identity. Subsequently, at the time when cells undergo their final cell division, the cells within different regions of the visual system become committed to their specific fates or individual identities. Evidence from studies in a variety of animals suggests that molecular gradients, timing of axon arrival, and correlated spontaneous activity all help to shape early targeting decisions and to establish precise connections. Refinements of the system involve active growth and branching of axons and dendrites, formation of synapses, and elimination of cells, axon collaterals, and some synapses. Further refinements of visual system development depend on activity and competition for limited supplies of neurotrophic factors but do not necessarily require visual experience. However, visual experience, especially during critical periods of active growth, can dramatically modify the final outcome.

I. INTRODUCTION

Mammalian development is an elegant process by which a single cell becomes a complete organism containing multiple organ systems that are highly interconnected. This is never more evident than during neural development. A central issue in the study of neurobiology concerns not only how nerve cells become connected in the first place but also how those connections become organized in such a way that the sensory world is represented. The topographical specificity of different parts of the mammalian visual system is dependent on highly ordered connections. Consider that the mammalian brain contains at least 100 billion neurons and that each of these

neurons can make more than 1000 specific connections with other neurons. Specificity is particularly evident in the visual system, in which a topographic map of visual space is maintained throughout each level of visual processing. For example, in the macaque monkey, a modest estimate of the number of visual areas requires that at least 30 visuotopic maps connect correctly in the developing brain; in humans there may be even more visual areas. Each cell within one of these maps processes information from a specific zone of the visual world, and that cell's neighbor processes information from an adjacent zone and so forth. Neighboring cells connect with neighboring cells in other visual areas. This characteristic is called visuotopy.

One approach that may be useful in elucidating the complex wiring of the mammalian visual system is to assume that axons make connections through a sequence of simpler evolutionarily conserved mechanisms. In mammals, research suggests that the earliest stages in the development of the nervous system are similar across a range of diverse species and involve similar if not identical molecular pathways. The similarities in these early steps of neural development imply the existence of powerful constraints on the regulatory relationships between genes that control early phenotypes (e.g., the characteristics and appearance of different parts of the nervous system). Good examples of these relationships come from genetic studies of eye and head development in flies (*Drosophila*), mice, and humans. In mice and humans the paired box *pax*-6/*aniridia* gene and its homolog in flies, the *eyeless* gene, play major roles in eye and craniofacial development. These genes control transcription factors that regulate a cascade of other genes important for eye and head formation. When loss-of-function mutations are produced in the *eyeless* gene, flies develop with no eyes, very reduced eyes, or defective eyes. Similar phenotypes are seen with genetic mutations in the *pax*-6/*aniridia* gene in mice and humans. Astoundingly, ectopic expression of either the fly *eyeless* gene or the *pax 6* gene results in the development of fully formed insect eyes on parts of the mutant fly's body that do not normally have eyes, such as the leg or wing. These remarkable results not only argue for a common evolutionary origin of eye development, but also reinforce the view that the earliest developmental programs are governed by highly conserved rules.

Although it is not the purpose of this article to focus on these early steps in development, the previous example has relevance. Later steps in neural development, in general, and visual system development, in particular, also involve evolutionarily conserved mechanisms. Since flies, mice, and humans differ in organization, size, and complexity of the brain, it is obvious that there must be developmental differences. Nevertheless, dramatic differences in adult brain organization can involve small changes in basic developmental mechanisms, such as changes in the number of cell divisions that occur before founder populations stop dividing, or changes in the number of cells that are eliminated during the periods of cell death that occur as a part of normal development in all nervous systems. The role of other mechanisms such as neural activity may be more important to the development of connections in large, complex nervous systems, although the basic mechanisms that translate that activity into cell growth or the formation of synaptic connections between nerve cells are also conserved across species.

In this article, we limit our consideration of development to two interconnected and well-studied visual centers in the brain, the lateral geniculate nucleus (LGN) in the thalamus and the primary visual cortex or V1 (Fig. 1). First, we provide an overview of the development of these visual areas in the larger context of brain development and discuss how these structures take on their adult shape. Then, we show how specialized laminar patterns and topographic connections between these structures can develop according to molecular cues. Finally, we explore possible mechanisms for the establishment of specificity, including the role of neural activity in forming and maintaining connections.

II. OVERVIEW OF VISUAL SYSTEM DEVELOPMENT

A. Early Neural Development

The brain begins as a simple plate of progenitor cells that eventually forms a tube that bends and balloons out into three fluid-filled vesicles during the process of development. Even before the neural plate forms a tube, however, communication takes place between cells that determines which progenitor or founder cells will give rise to specific broad regions of the visual system, including regions that contain the retina, the LGN, and the visual cortex (Fig. 2). An enormous array of transcription factors and extracellular molecular signals have been identified that are involved in

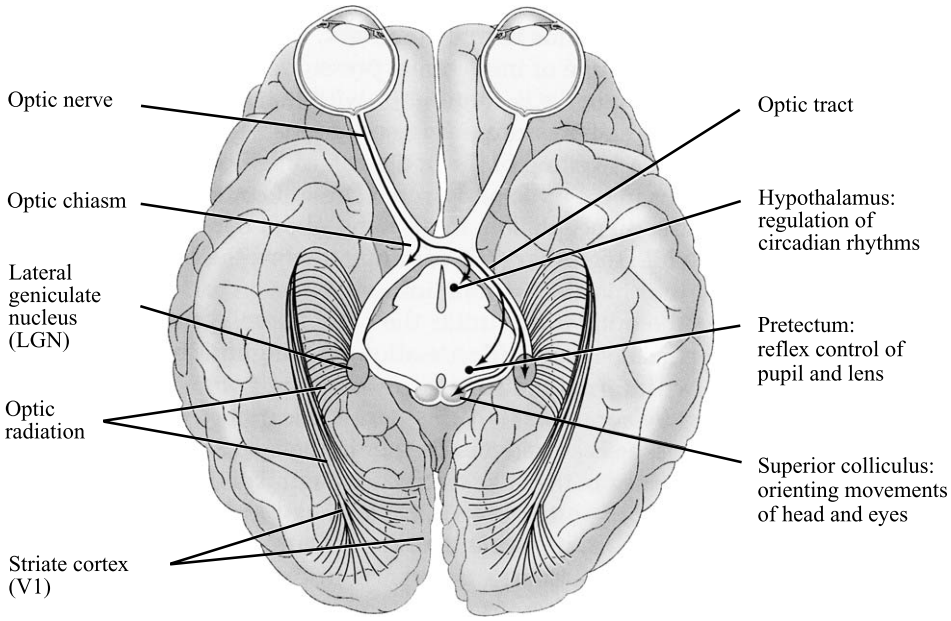


Figure 1 The visual pathways shown as a schematic viewed from the ventral surface of the human brain. The axons from ganglion cells located in the retina leave the eyes as the optic nerve. The ganglion cells located in the temporal retina (shown for the right eye) send axons to the same side (ipsilateral) of the brain (arrow), whereas the axons in the nasal retina cross at the optic chiasm to the other hemisphere or contralateral side of the brain. Once the axons have left the optic chiasm they bundle together as the optic tract. Groups of axons leave the optic tract at various points to make synapses with groups of cells (nuclei) conserved with different visual functions, such as the hypothalamus (circadian rhythms), the lateral geniculate nucleus (LGN); (conscious vision), pretectum (pupillary and lens reflexes), and superior colliculus (orienting head and eyes). The superior colliculus is homologous with the optic tectum of nonmammalian vertebrates. Cells in the LGN send axons to the primary visual cortex or striate cortex, also called V1, located at the back of the brain in the occipital lobe [reproduced with permission from Purves, D., *et al.* (Eds) (1997). *Neuroscience*. Sinauer, Sunderland, MA].

these early stages of regional specification. It has been hypothesized that regional specification of the forebrain, like the hindbrain, involves segmentation, in this case into a series of longitudinal and transverse segments defined by the expression of a number of genes; however, specific molecules that define the exact boundaries of the developing LGN or primary visual cortex have not been identified.

As development progresses, cells divide close to the inner (ventricular) surface of the neural tube. Initially, all cells extend their processes across the full width of the neural tube, but later only a subset of cells, the radial glial cells, span the full width of the developing forebrain. Radial glia are used by other cells as migratory railway guides to move from the ventricular zone to their final destinations, whether in the developing LGN or visual cortex. Each neural progenitor cell, or stem cell, undergoes a certain number of cell divisions in the young animal, until it divides no more in the adult. This final mitotic division is termed the birth date of the neuron. After the final cell division a neuron differentiates or matures. Neurons in different areas of the brain, such as the LGN and visual cortex, exhibit differences, reflecting their variable differentiation programs. Cells within different neural structures mature at different times in the developmental process and at different rates. Additionally, each neural structure may have its own gradient of development, such that some cells in the nucleus mature before others. Finally, as neural structures mature many developing neurons are eliminated through cell death. This natural cell death process not only eliminates cells with incorrect connections but also allows the production of transient cell populations that help to shape the structures and connections of the nervous system by providing temporary scaffolding for migrating neurons and guideposts for developing axons. The development of the mammalian brain is a highly dynamic process involving multiple overlapping gradients of progression. Cells in the brain experience multiple combinatorial genetic and activity-dependent cues at different times in their development that ultimately produce the complex structure of the adult mammalian brain.

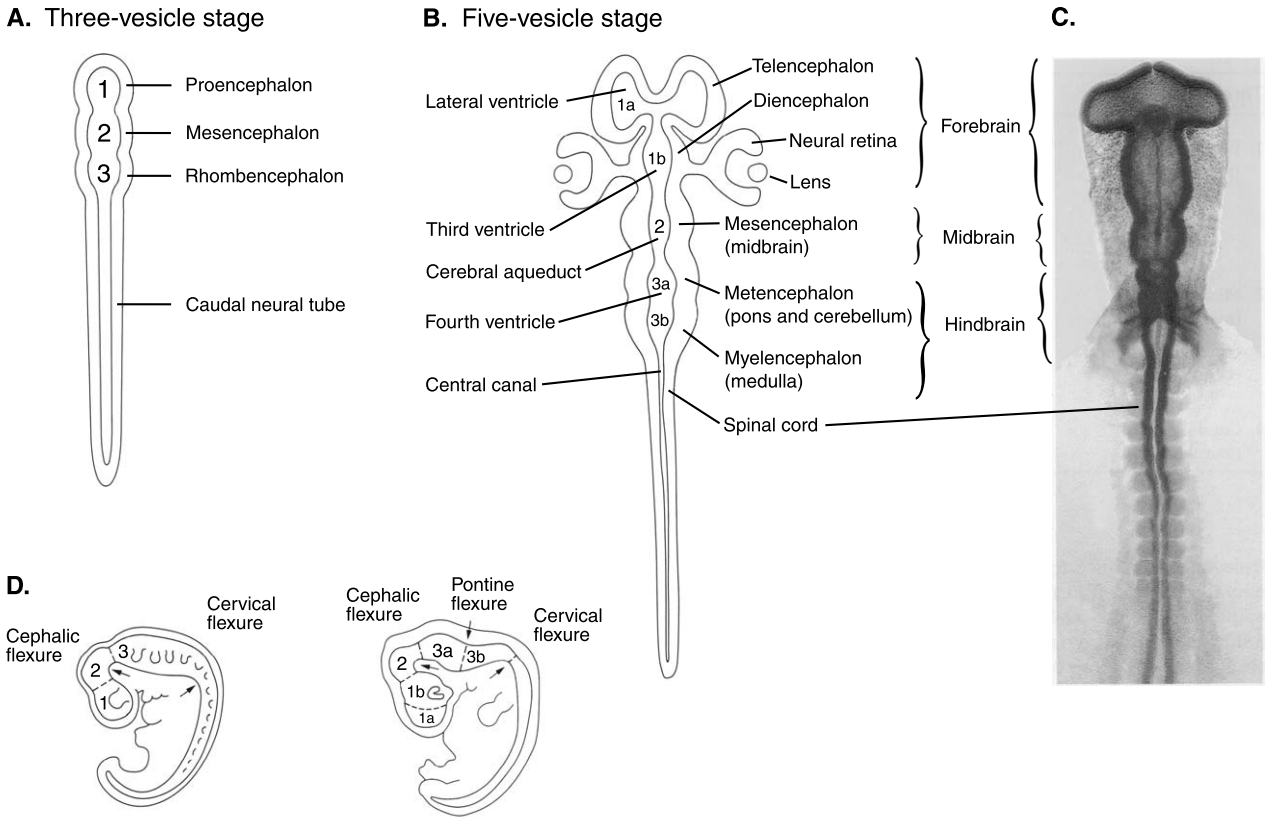


Figure 2 Successive stages of development of the neural tube. The nervous system begins as a plate of tissue (not shown) that rolls up to form a tube. (A) Three-vesicle stage. At early stages of development only three brain vesicles are present. (B) Five-vesicle stage. At later stages, two additional vesicles form, one in the area of the forebrain (1a and 1b) and the other in the hindbrain (3a and 3b). Before the neural tube forms, cells are prespecified such that the retina and lateral geniculate nucleus will form from part of the diencephalon (3b), the visual cortex from the telencephalon (1a), and the superior colliculus from the midbrain (2). (C) Micrograph showing a dorsal view of the neural tube at an early stage of development. The expansion of the future telencephalic vesicle is apparent (micrograph of the chick neural tube provided by G. Schoenwolf). (D) The positions of the cephalic, pontine, and cervical flexures [from Kandel, E. R., Schwartz, J. H., and Jessell, T. M. (Eds.) (2000). *Principles of Neural Science*, 4th ed. Reproduced McGraw-Hill Companies, with permission of the New York].

B. Development of LGN Structure

The adult LGN of many mammals, including humans, is laminated with compartmentalized input. Each LGN layer receives input from only one eye. Each layer also contains a topographic map of one-half of the visual world. The LGN layers are stacked on top of each other like a set of pancakes (Fig. 3). In primates there is layer specialization such that small, medium, and large cells are segregated into different layers and carry different types of visual messages to primary visual cortex. During development, the future large LGN cells are born and begin migrating from the ventricular surface toward their final destination before the future smaller LGN cells. The differing extracellular environment encountered by LGN cells born at different times during development may change the genetic program of those cells, such that the different classes of cells adopt different morphological and neurochemical identities during differentiation. In fact, in visual cortex there is evidence supporting the view that important decisions about cell fate are made at the time progenitor cells divide in the ventricular zone. It is likely that the same rules apply to cells throughout the developing neural tube. During development, LGN progenitor neurons destined to represent central vision versus peripheral vision divide and mature in a gradient such that cells that will carry information about the central visual world develop earlier than cells that will carry information from the visual periphery. Gradients in the timing of cell birth and maturation of LGN

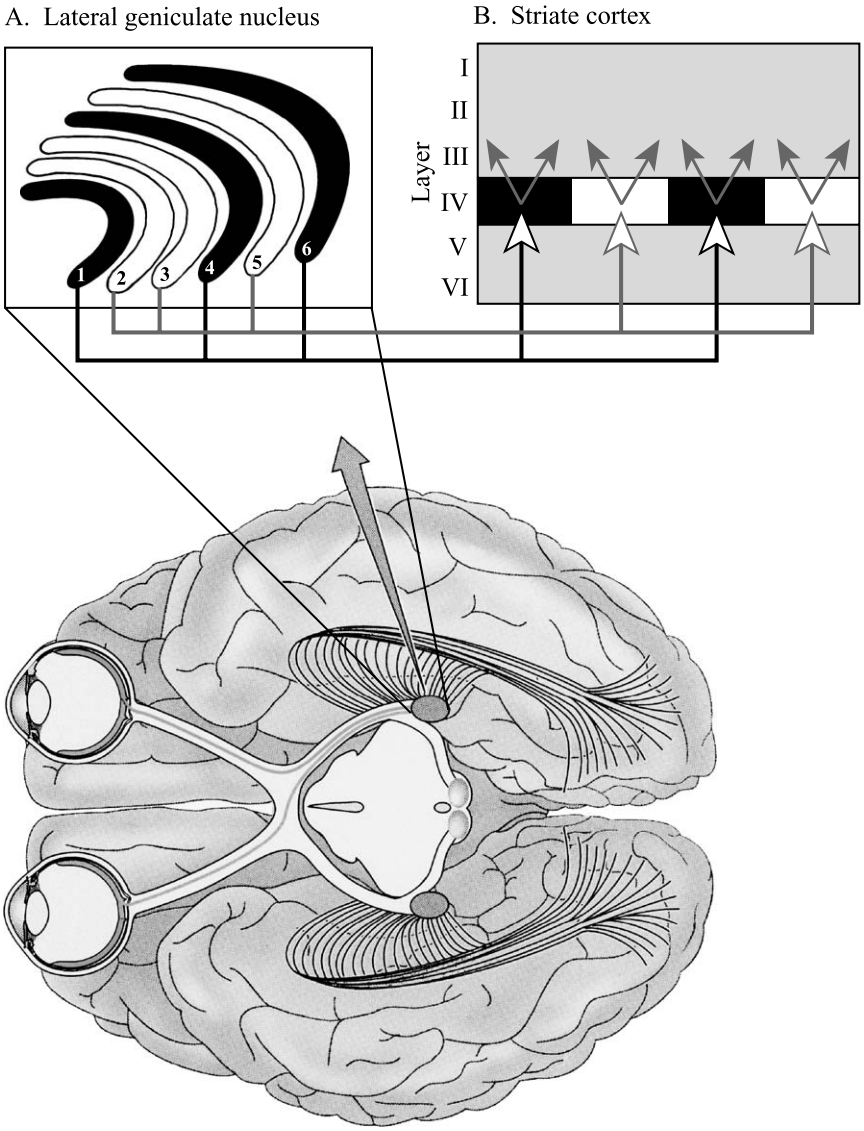


Figure 3 The connections between the lateral geniculate nucleus (LGN) and striate cortex (primary visual cortex). The visual pathway is shown as a schematic viewed from the ventral surface of the human brain. This same view is also shown in Fig. 1. Although the LGN receives inputs from both eyes, these are segregated in separate layers as shown in (A) The first two layers contain large cells (magnocellular layers), and the remaining four layers (parvocellular cells) contain medium-sized cells. Not shown are the small cell (koniocellar) layers that lie between each of the main layers shown. (B) In many species, and most primates, the inputs from the two eyes remain segregated in ocular dominance columns of layer IV, the primary target of LGN axons. Layer IV neurons send their axons to other cortical layers; it is at this stage that information from the two eyes converges onto individual neurons [reproduced with permission from Purves, D., *et al*., (Eds) (1997). *Neuroscience*. Sinauer, Sunderland, MA].

progenitors and within any structures connected to the LGN (such as the eye and the primary visual cortex) may be one of several simple mechanisms that ensure that topographically correspondent regions connect properly.

The retina and the visual cortex, the main inputs to the LGN, both innervate the LGN visuotopically. There is evidence that axons leaving visual cortex and retina grow out very early but may not reach their targets at the same time. In monkeys, retinal ganglion cell axons of different classes grow to the LGN very early and terminate immediately within the appropriate portions of the LGN, demonstrating that some retinal axons are molecularly specified to connect to

specific LGN target regions. However, retinal ganglion cell axons of the same class, arriving from the two eyes initially invade overlapping regions of the LGN. The segregation of these axons into eye-specific layers seems to depend on correlated waves of spontaneous activity in the retina since blocking such activity blocks the normal segregation of retinal axons, at least in carnivores. Cortical axons, like retinal axons, approach the LGN very early in development but, unlike retinal axons, do not invade the nucleus immediately. A special early-born class of cortical cells, the cortical subplate cells, may be the first to send axons to the thalamus. These cells may pioneer the later-born layer VI cell axons that will form permanent connections with the LGN; however, the pioneer axons must still find their way to the LGN. Evidence suggests that the earliest axons use a variety of membrane-bound cues to locate correct targets. When cortical axons do invade the LGN, they immediately grow into visuotopically correct areas of the nucleus and do not appear to require a sorting period, although some smaller refinements may take place as these axons mature.

C. Development of Visual Cortical Structure

Multiple cortical areas, defined by morphology, physiology, and connections with other cortical and subcortical nuclei, make up the neocortex, like a patchwork quilt. One patch of the quilt is the primary visual cortex, located in posterior or occipital cortex. The adult visual cortex contains a precise topographic map and, like all of mammalian cortex, is laminated. The neocortex has six layers. Axons arriving from the LGN synapse primarily in layer IV of visual cortex. In primates, axons from LGN cells of different functional classes (e.g., large versus small LGN cells) synapse in different divisions of cortical layer IV or in distinct compartments located above layer IV. Additionally, LGN cells representing the left and right eye send axons to separate columns of cortex such that cells in a particular column of cortex respond to a stimulus to one eye or the other preferentially (Fig. 3). Cells in layer IV make connections with layers above and below layer IV, and this is where the fusion of information from the two eyes occurs. Cortical cells in layer VI send axons back to the LGN. Also, each ocular dominance column is divided into smaller columns in which cells respond to one orientation of a bar of light. Orientation columns are arranged mainly in pinwheels, such that all orientations are roughly represented within each ocular dominance domain and visuotopy is maintained across columns (Fig. 4). The aforementioned scenario is more simplistic than the biological situation but serves to emphasize some of the complexities of organization that must be achieved during development for the visual cortex to function properly.

As does the developing LGN, the developing cortex begins as a simple epithelium—a thin columnar sheet of cells. As development advances and cells divide, this sheet of cells becomes thicker. The radial glial cells maintain contact with both the ventricular and the pial sides of this sheet, whereas cortical neuroblasts begin dividing at the ventricular zone. Soon, three precortical subdivisions can be recognized. The ventricular zone is the most ventral, where excitatory cortical cells are born from a resident stem cell pool. Interestingly, inhibitory cortical cells, like olfactory bulb cells, may originate from a noncortical stem cell population located in the ganglionic eminence (e.g., the developing basal ganglia). Above the ventricular zone is the intermediate zone, a highway for axons, where corticofugal and corticopetal axons are located. Above the intermediate zone is the preplate, a layer of early born cells that have migrated along radial glial fibers to lie

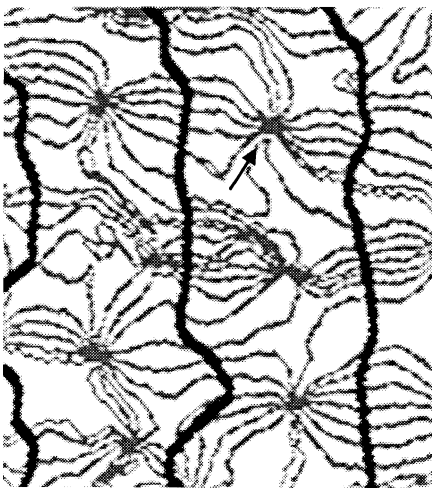


Figure 4 Orientation columns in the visual cortex of the monkey. Comparison of optical image maps of orientation preference and ocular dominance in monkey visual cortex. The thick black lines represent the borders between ocular dominance columns. The thin gray lines represent isoorientation contours, which converge at orientation pinwheel centers (arrow). Each left and right eye ocular dominance column represents a common region of visual space. An adjacent set of columns represents an adjacent region of visual space and so forth across the visual cortex [adapted with permission from Obermeyer, K., and Blasdel, G. G. (1993). Geometry of orientation and ocular dominance columns in monkey striate cortex. *J. Neurosci.* **13,** 114–129].

adjacent to the pial surface. As cell division proceeds, the preplate is split by migrating cells to form the marginal zone (the future layer I of cortex) dorsally and the subplate ventrally. Cells that will form the remaining adult cortical layers will migrate to lie between the marginal zone and subplate. Most of the cells within the preplate and subplate subsequently die. Unlike the LGN, in which cells that are born first lie closest to the pial surface of the neural tube, cortical layers II–VI are formed in an inside–out gradient such that cells born early become the lower layers, whereas cells born later migrate through these layers to form the upper layers (Fig. 5). The transitory layers formed from the original preplate are very important for the proper formation of the other layers of cortex. If the marginal zone is disrupted or not formed, then the cortical layers will not develop in the proper order. For instance, in the reeler mouse, a mutant mouse lacking the gene *reelin* (normally expressed by cells in the marginal zone), the cortical layers develop upside-down or in an outside–in gradient. If the subplate is experimentally removed, the cortical layers develop normally, but axons from the developing LGN are unable to find their proper cortical target.

How the many functionally distinct areas of the neocortex (e.g., the patches of the quilt) differentiate within the forebrain is less well understood. Evidence indicates that parcelation of the forebrain into subdivisions such as the visual cortex involves the expression of a cascade of regulatory genes that first subdivide broad regions of the forebrain and subsequently interact to regulate functionally specific areas. As mentioned previously, in the hindbrain, sequential bands have been identified that differentially express certain genes. These segmental regions are termed rhombomeres and are thought to represent early specification of nuclei in different segments of the brain stem. Some candidates for mediating interactions across brain stem boundaries are members of the Ephrin (Eph) receptor tyrosine kinase family. It is noteworthy that mutant mice lacking connections between the thalamus and cortex still show differential regional expression of genes across the neocortex as well as the normal development of cortical layers independent of this extrinsic input. Nevertheless, there is considerable evidence that extrinsic inputs from the thalamus, particularly those that bring sensory information about the outside world such as visual signals from the retina, can radically alter the pattern of development of cortical areas.

III. AXON PATHFINDING: WIRING THE LGN AND VISUAL CORTEX

A. Molecular Cues

A persistent mystery in developmental neuroscience is how axons travel great distances to find their way to highly specific targets. Correct target selection and the refinements involved in the formation of visuotopic connections between the retina, LGN, and cortex involve both genetic and epigenetic cues. Examples of

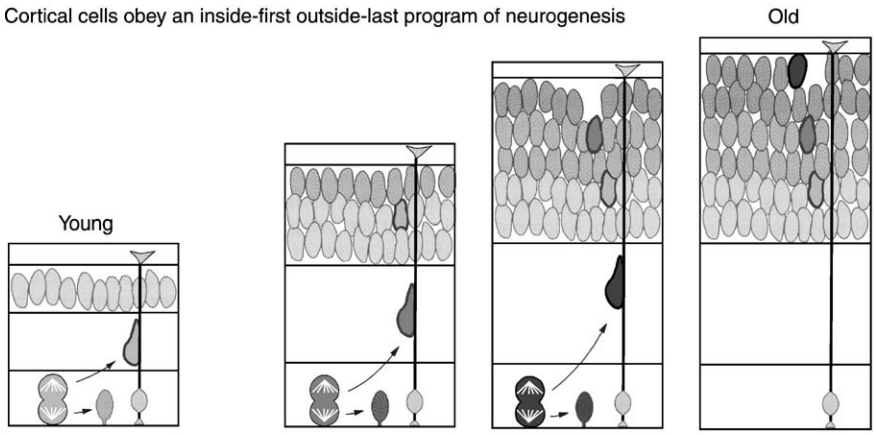


Figure 5 Generation and migration of neurons in the mammalian cerebral cortex. Cortical neurons are generated in an inside-first, outside-last order. Neurons born within the ventricular zone at the early stage migrate to the deepest layers of the cortical plate. Neurons generated at later stages migrate past the earlier generated neurons to form the more superficial layers of the cortex [adapted with permission from McConnel, S. K. (1992). The determinants of neuronal identity in the mammalian cerebral cortex. In *Determinants of Neuronal Identity* (M. Shantiland, and E. Macegno, Eds.), pp. 391–432. Academic Press, New York].

genetic mechanisms include the presence of membrane-bound and secreted molecules, which are permissive, instructive, or obstructive cues for axons with the proper receptors. Axons can grow on various substrates in the brain, including the extracellular matrix, other neural cell membranes, and glial cell membranes (such as the radial glia on which LGN and cortical cells migrate). Developing axons have a specialized structure, a growth cone, at their tip to help them navigate through the extracellular milieu (Fig. 6). This highly motile structure contains receptors capable of transmuting signals about the local environment. The growth cone guides the axon along pathways toward the target region defined by a combination of positive or negative cues.

Once an axon reaches its target, it often will send out a simple T-shaped branch as shown for the LGN axon growing into visual cortex in Fig. 7. Within the visual system, axons often reach their target region while the target cells are still migrating into position or just after migration is complete. Growth cones on the tip of axon branches can detect gradients of membrane-bound molecules to guide axons to their general addresses within a target structure. Axons respond differentially to these molecular gradients within the target based on their initial specification that likely occurred at the time of final cell division and prior to axon outgrowth.

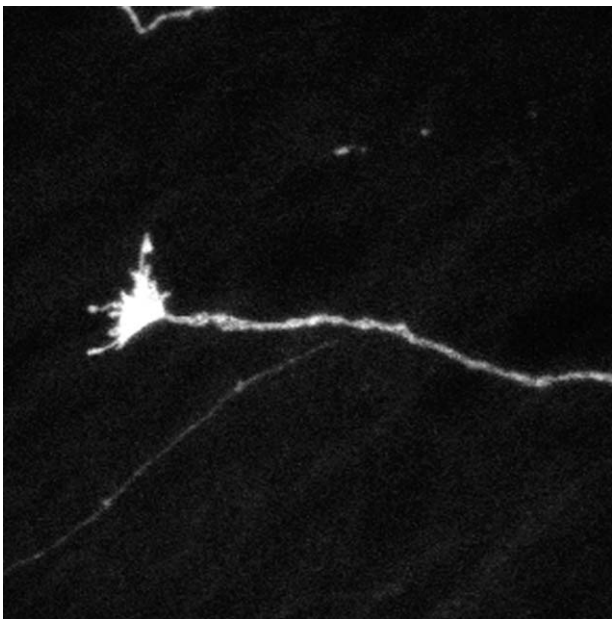


Figure 6 An example of a thalamocortical axon with a growth cone in neonatal mouse brain. The growth cone is a complex, dynamic structure at the tip of the growing axon that samples the local environment, guiding the axon to its target.

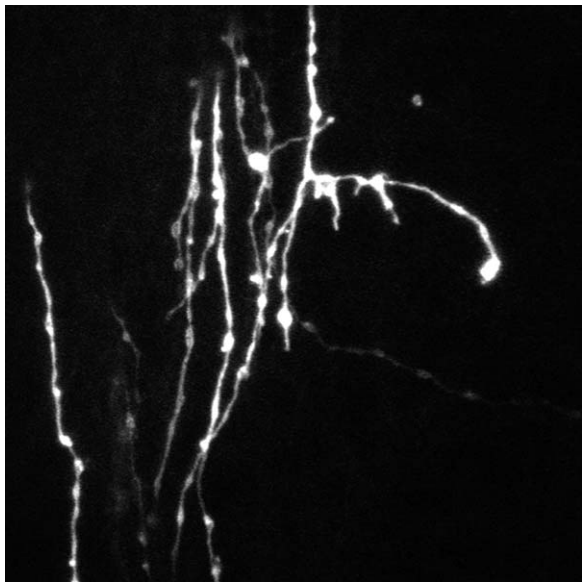


Figure 7 Thalamocortical axons initially exhibit collateral T-shaped side branches when they begin to innervate the cortical plate as shown here from the neocortex of a mouse on the day of birth.

Retinal ganglion cell axons from different retinal locales will therefore respond differently to cues at the target based on their unique growth cone receptors. As mentioned previously, a good example of this process involves the specificity with which different retinal ganglion cell classes target regions of the LGN. Although many molecules have been identified that may guide retinal ganglion cell axons to the right targets, evidence has converged on two membrane-bound ligands of the receptor tyrosine kinases, ephrin A2 and ephrin A5. Both of these ligands are expressed in gradients in the optic tectum, another brain area innervated by retinal ganglion cells. Eph kinases are distributed in gradients in the retina. These two molecular gradients may regulate retinotectal and retinogeniculate topography because ephrins bind to Eph kinases and activation of Eph kinases inhibits axon outgrowth. In particular, the levels of ephrin A2 and A5 are higher in the posterior tectum than in the anterior tectum, and this could inhibit growth of temporal retinal axons, which are rich in the appropriate Eph kinases (Fig. 8). Moreover, retinal axons in mice with mutations in ephrin A2 and A5 exhibit a loss of visuotopy and innervate areas of the optic tectum that they normally would avoid. Interestingly, members of this same family of ephrin molecules are implicated as cues for establishing the proper laminar connections in visual cortex. Therefore, members of the ephrin family may be involved in early regional

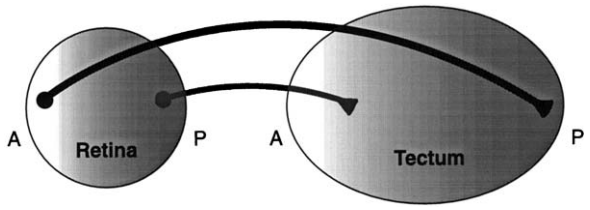


Figure 8 Eph kinases are distributed in gradients in the retina, and ephrins are distributed in gradients in the optic tectum. These two molecular gradients may regulate retinotectal visuotopy because ephrins bind to eph kinases and activation of eph kinases inhibits axon outgrowth.

differentiation in the brain, as well as the establishment of axonal connections later. This example illustrates how a single family of similar molecules can direct several very different developmental programs, sometimes simultaneously, in different neural regions and with overlapping expression patterns, and it underscores the high level of conservation of mechanisms in neural development.

B. The Role of Activity

As neurons mature, they begin to display electrical activity. Neural activity has been shown to play an important role in the correct targeting of axonal connections as well as the formation and maintenance of synapses. The role of neural activity in shaping connections is not restricted to the postnatal period when animals begin to use their senses to explore their environment. Neurons are active well before birth, especially in animals such as primates that are born at a relatively mature developmental stage. Experiments in cats and ferrets have demonstrated that prior to visually driven activity (at a time when the eyelids are still closed), correlated waves of spontaneous activity already exist in the retina and brain. Neurons at this stage are coupled by developmentally transient gap junctions (at least in the retina and the visual cortex, but probably also within other areas of the visual system). The waves of spontaneous activity appear to be important to the segregation of left and right eye inputs within the developing LGN layers and to the segregation of eye-specific inputs into ocular dominance columns in cortex. As mentioned earlier, blocking this activity in prenatal cats and postnatal ferrets disrupts the binocular segregation process. In the examples cited previously, activity likely affects the construction of new axonal branches and synapses and not the retraction of inappropriate connections since the manipulations are initiated when retinal and geniculocortical axons still have a simple stick-like morphology.

During the period of activity-dependent modification of connections, many manipulations are made to the visual system in an effort to define the mechanisms that relate activity to the growth of neural processes and to synaptogenesis. The most famous of these experiments were done in the 1960s by David Hubel and Torsten Wiesel, who showed that preventing one eye of a kitten from seeing during early life resulted in a number of dramatic changes in the connections and function of the visual system. In their experiments, they sutured the lid of one eye closed, but it is now known that the same effects are obtained as long as useful pattern vision is prevented; the amount of light reaching the retina is not as important. In these experiments, they found that kittens could no longer see well out of the deprived eye (i.e., they developed amblyopia) if the deprivation continued past 3 months of age. These investigators also found that geniculocortical axons from the deprived eye occupied less territory in layer IV of visual cortex (the deprived ocular dominance column was smaller) and cells in the cortex now responded mainly to the normal eye. If the kittens were allowed to develop normally for the first 3 months of age, monocular lid suture would no longer affect ocular dominance columns, suggesting that there is a window of time or critical period during which visual experience can have an impact on ocular dominance column formation. If the eye was sutured and shortly afterwards opened again during the critical period, then normal ocular dominance formation would occur. The effects of monocular deprivation are not the consequences of disuse, as Hubel and Wiesel clearly demonstrated, since depriving both eyes of useful pattern vision has much milder effects on the visual system (Fig. 9).

These experiments, along with numerous experiments that followed, suggest that developing geniculocortical axons that receive input from the left and right eyes can compete with one another for territory in cortex. The working hypothesis is that appropriate levels of activity and the temporal correlation of that activity in potential pre- and postsynaptic partners provide axons access to appropriate levels of growth factors (i.e., neurotrophins). These neurotrophins, in turn, activate the cellular machinery that allows axons to grow and develop synapses (Fig. 10). Axon growth is clearly constrained by many other factors in the target tissue discussed earlier, including factors that directly repel or attract axons or substrates that

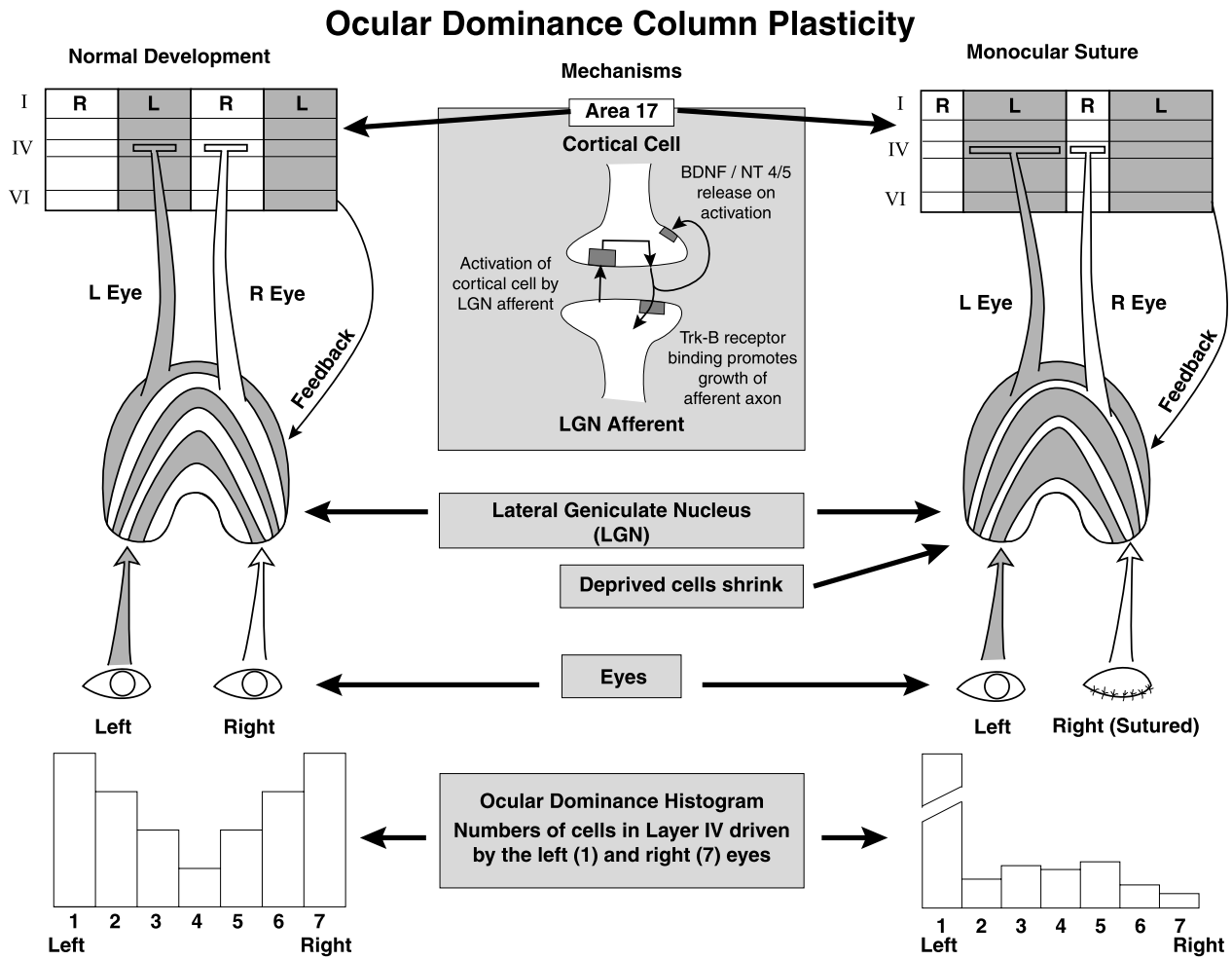


Figure 9 Changes within the LGN and visual cortex (area 17) seen following monocular lid suture in a macaque monkey. Axons from the left and right eyes segregate into six layers within the LGN early in development. LGN axons segregate into ocular dominance columns within layer IV of area 17. Axon segregation within the LGN and cortex takes place before birth, without visual experience. However, correlated spontaneous activity initially within each eye and subsequently via connections between segregated eye inputs in the LGN and cortex including corticogeniculate feedback may help axons segregate into ocular dominance territories. At birth, LGN axons are very immature. Neonatal lid closure eliminates useful patterned activity within the sutured eye. As a result, deprived LGN axons grow less and their LGN cell bodies shrink, and nondeprived LGN axons expand more than normal. Since deprived LGN cell bodies within the monocular segments of the LGN do not show these changes, it is likely that LGN axons innervated by the left (L) and right (R) eyes compete within cortex for limited quantities of neurotrophic factors. Evidence suggests that the neurotrophins, brain-derived neurotrophic factor (BDNF) or neurotrophin (NT) 4/5, are involved, as shown in the center. These neurotrophins bind to tyrosine kinase receptor B (Trk-B), which can be located both pre- and postsynaptically. In this model activation of the cortical cell by axons would cause release of BDNF or NT4/5, which would act to promote growth and survival of the LGN axon. Neurotrophin release could also influence the dendritic growth of cells from which it was released via autocrine mechanisms [adapted with permission from Barker, R. A., and Barasi, S., (with Neal, M. J.) (1999). *Neuroscience at a Glance* Blackwell Science, Oxford].

support growth. In addition, it is evident that pattern vision per se is not essential for setting up ocular dominance columns in the first place since monkeys are born with well-developed ocular dominance columns. Nevertheless, basic cellular mechanisms that normally help axons driven by the left and right eyes to segregate into layers in the LGN and into columns in the cortex are likely the same as those that allow for plastic changes following manipulations of visually driven activity.

How does activity help some developing axons to grow processes or form or strengthen some synapses at the expense of others? A working model for this process was originally derived from the learning model

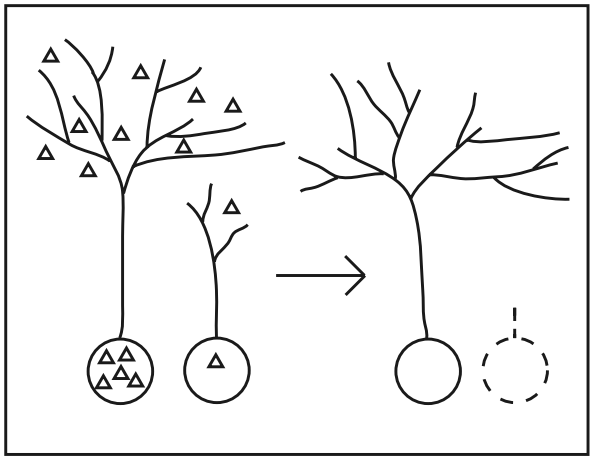


Figure 10 Selective access to growth factors can influence the survival of neurons. Neurons (circles) compete for a limited supply of growth factors (triangles). Neurons that do not get enough of the growth factor may die or have smaller axonal trees.

of Donald Hebb (1949). Hebb proposed that pathways that are active together form or strengthen connections at the expense of those that are inactive or not co-active (Fig. 11). In Hebb's model each instance of coactivation in an assembly of connected cells strengthens these connections. The favored synapses eventually become so strong that memories form. Hebb's idea for the strengthening of synapses in learning and memory provides a good model to explain how activity can influence the development of initial connections.

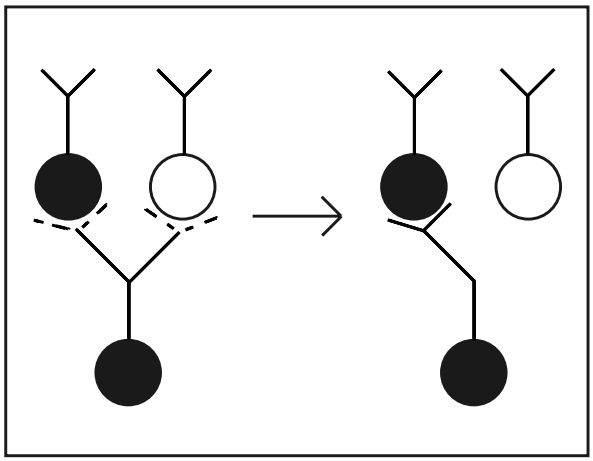


Figure 11 A simple schematic depicting Hebbian mechanisms of axonal plasticity. The two black cells on the left have synchronous action potentials. The white cell does not fire coincidently with the black cells. Initially, the lower black cell sends a synapse to both upper cells (dashed lines). After a certain amount of time, the synapse between the black cells is strengthened, whereas the synapse between the black and white cells is weakened or eliminated.

Hebb's hypothesis can be used to explain the development of ocular dominance columns if sets of correlated inputs tend to exclude uncorrelated inputs and are additionally constrained by molecular gradients related to retinotopy. Support for this idea was best demonstrated in cases in which an extra eye was transplanted to the head of a developing frog (Fig. 12). In frogs, each eye normally sends a crossed projection to each optic tectum. When a third eye is introduced, two eyes divide one tectal territory by developing ocular dominance bands in the dually innervated tectum as would be predicted by the model.

At the cellular level in mammals, the working model that has been proposed to explain how correlated activity can influence the formation and strengthening of synapses and the growth of neural processes involves a special class of receptors for the transmitter glutamate, the *N*-methyl-D-aspartate (NMDA) class of glutamate receptors. The NMDA receptors bind the transmitter glutamate, but such binding has no effect unless another condition is satisfied. At the "resting" membrane potential, the NMDA receptor's channel is blocked by magnesium. Strong depolarization of the postsynaptic membrane removes the magnesium block by electrostatic repulsion. If glutamate now binds to the NMDA receptor while the cell is depolarized, calcium is able to enter the cell via the NMDA channel. Calcium ions that enter through the NMDA channel activate kinases in the cell. Through either kinase activation or some other calcium-dependent series of steps, the postsynaptic cell is modified and potentially more sensitive to transmitter release. Additionally, a retrograde molecular signal or neurotropin (e.g., neurotrophin 4/5 or brain-derived neurotrophic factor) may be released from the postsynaptic cell that can influence presynaptic axon growth or the level of transmitter released by that axon (Fig. 13). Strong evidence in support of the special role for NMDA receptors also comes from work in the three-eyed frog, where infusion of the NMDA antagonist 2-amino-5-phosphonovaleric acid into the optic tectum blocks retinal axon segregation into ocular dominance bands. Subsequently, there have been many other studies that have supported aspects of this model, which has been useful not only in explaining how connections are refined during development but also how plastic changes can occur following different forms of visual experience. Among these studies, those showing that neurotrophin release can be influenced directly by activity and that different classes of neurotrophins and neurotrophin receptors can differentially affect the growth of axons and dendrites depending on the cell

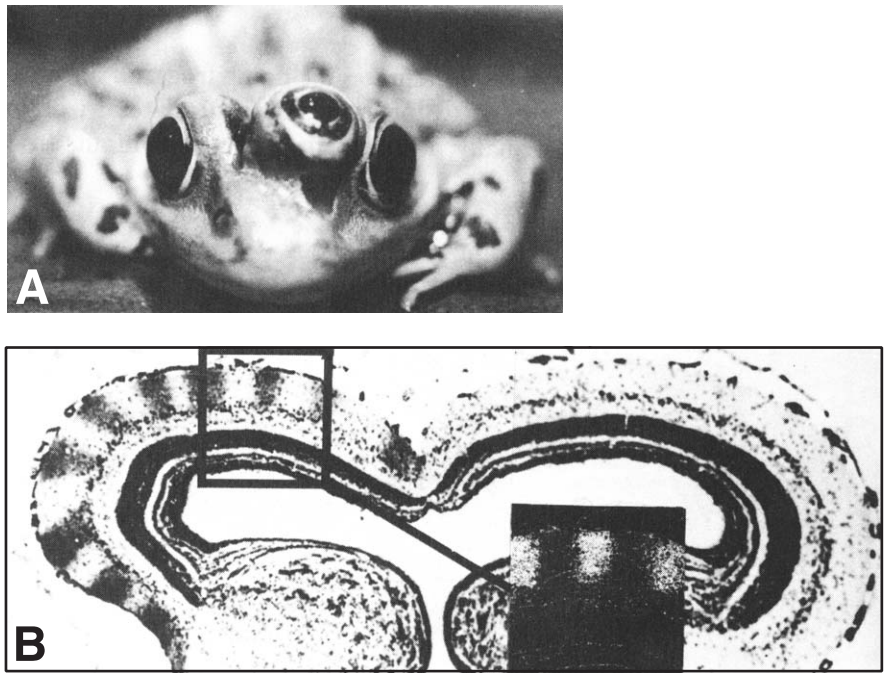


Figure 12 The three-eyed frog. (A) Three-eyed frogs have been studied by Martha Constantine-Paton and colleagues. (B) An autoradiograph of the optic tectum showing the formation of stripes of inputs (black and white) from the normal and implanted eye. The inset shows an enlargement under dark-field illumination [adapted with permission from Constantine-Paton, M., and Law, M.I. (1978). Eye-specific termination bands in tecta of three-eyed frogs *Science* **202,** 639–641. Copyright © 1998 American Association for the Advancement of Science].

type are especially relevant because they link activity with other basic cellular mechanisms involved in cell survival and selective outgrowth and targeting of neural processes. In addition to neurotrophic factors and activity-dependent mechanisms, a variety of other factors can impact the development of visual system connections, including, but not limited to, levels of various hormones, a variety of neurotransmitters, and neuromodulators.

For the examples given previously, it is important to keep in mind that although ocular dominance columns do not require visual experience to form in the first place (at least in primates), there are other aspects of vision that do require appropriate visual experience and are not present before birth even in primates. Stereoscopic depth perception, for example, appears to require adequate coordinated activity from both eyes since it does not exist at birth and becomes disrupted if input from the two eyes is not appropriately correlated. Such a situation can arise in humans or other animals that are born and grow up wall-eyed or cross-eyed. In the latter case, stereoscopic depth perception is lost and binocular connections do not form. Only if the eyes are surgically aligned early in a child's life will he or she develop normal stereoscopic vision. Nevertheless, the same activity-driven mechanisms described previously likely operate to allow appropriate binocular connections to form in visual cortex that, in turn, help us to appreciate our three-dimensional world.

IV. CONCLUSIONS

During neural development, cells proliferate and influence each others' fates through an elaborate interplay of extrinsic and intrinsic signals. Cells become committed to specific fates, such as laminar location in visual cortex, early in development before they begin to migrate. Molecular axon pathways are established through a variety of secreted factors and membrane-bound molecules. Correct axon targeting involves gradients of chemoattractant and chemorepellent molecules. Major waves of cell death take place at this stage that help to sculpt connections. Unlike spontaneous neural activity, activity driven by visual experience is not required for basic aspects of axon targeting, map formation, and segregation of axons into ocular territories within the LGN or visual cortex in precocial species such as primates. In all species,

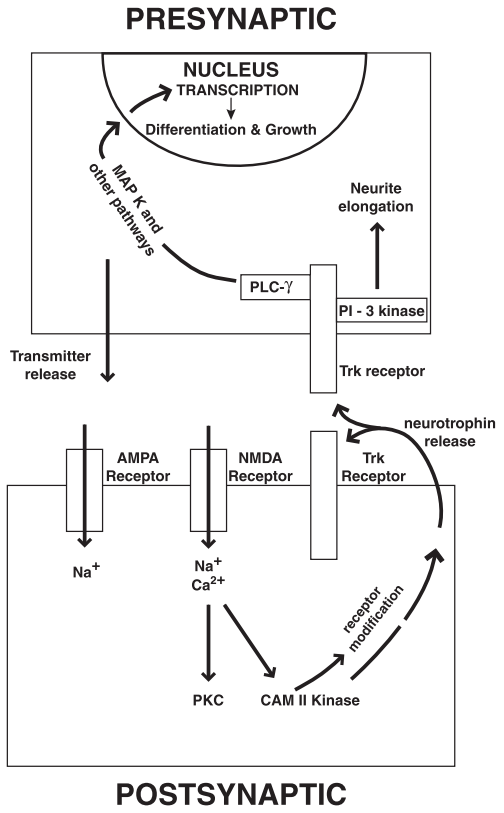


Figure 13 The NMDA receptor channel can open only during depolarization of the postsynaptic neuron from its normal resting level. Depolarization expels Mg^{2+} (not shown) from the NMDA channel, allowing current to flow into the postsynaptic cell. Since the NMDA channel is permeable to Ca^{2+}, there is a significant Ca^{2+} entry into the cell that can trigger other events involving Ca^{2+}. Through either kinase activation (protein kinase C) or a separate Ca^{2+}-dependent mechanism (calmodulin kinase II) or other pathways, neurotrophins can be released from the postsynaptic cell that can act in a paracrine fashion to influence growth of the presynaptic cell/processes or in an autocrine fashion to influence its own growth via pathways shown in the diagram for the presynaptic terminal/cell. Binding to neurotrophin Trk receptors permits Trk molecules to phosphorylate tyrosine residues. Phosphorylation of specific tyrosine residues creates binding sites for PI-3 and phospholipase C (PLC)-γ and recruitment of these proteins into a complex, thus initiating a signaling cascade that can lead to neurite elongation or, via the MAP kinase pathway, to transcription and ultimately differentiation and growth [from Casagrande, V. A., and Wiencken, A. E. (2000). Developmental plasticity in the mammalian visual system. In *The Mutable Brain. Dynamic and Plastic Features*. (J. H. Kaas, Ed.), Copyright © 2000 by Overseas Publishers Association Adapted with permission from Gordon and Breach Publishers].

however, visual system wiring can be profoundly affected by abnormal visual experience, especially during the phase of rapid axonal and dendritic growth.

An elaborate interplay between genetic and epigenetic factors defines specific connections in the highly complex system that is the mammalian nervous system. Molecular cues are important for guiding axons to their general target region and neural activity refines these connections. Even if genetic factors initially define axonal boundaries, it is not necessarily the case that these boundaries are static. The boundaries are partially plastic into adulthood and can be changed even in adults as the result of insult or injury. Neural activity can affect the growth of neural processes and synapses or strengthen existing synapses through Hebbian activity-dependent mechanisms and/or the limited availability of neurotrophins. These same Hebbian mechanisms allow for more limited plasticity in the wiring of the adult brain. Major changes in wiring only occur when processes are growing, but less dramatic changes are possible in the adult and likely involve the same cellular machinery that is used to establish and refine connections normally. The mature visual system at all levels can respond via expression of immediate early genes and regulation of a variety of transmitter and neuropeptide-related molecules to conditions of visual deprivation and damage. Despite the striking similarities that are found between developmental and adult neural plasticity, there are fundamental differences in the developing and adult brain. Younger organisms characteristically seek a variety of forms of visual stimulation and repeat visuomotor activities in ways more sedentary adult animals do not. Adult animals that are forced to use their visual systems via training show greater compensation for early damage than adult animals that are not. Understanding the neural mechanisms that drive early visual self-stimulation and how these factors interact with the activity-dependent mechanisms discussed to allow for compensation after visual system damage is one of the challenges for future investigations.

See Also the Following Articles

BRAIN DEVELOPMENT • NEUROPLASTICITY, DEVELOPMENTAL • SYNAPTOGENESIS • VISION: BRAIN MECHANISMS • VISUAL CORTEX • VISUAL DISORDERS

Suggested Reading

Casagrande, V. A., and Wiencken, A. E. (1996). Prenatal development of axon outgrowth and connectivity. *Prog. Brain Res.* **108,** 83–93.

Casagrande, V. A., and Wiencken, A. E. (2000). Developmental Plasticity in the Mammalian Visual System. In *The Mutable Brain. Dynamic and Plastic Features* (J. H. Kaas, Ed.), Harwood Academic, Reading, UK.

Castellani, V., Yue, Y., Gao, P. P., Zhou, R., and Bolz, J. (1998). Dual action of a ligand for Eph receptor tyrosine kinases on specific populations of axons during the development of cortical circuits. *J. Neurosci.* **18,** 4663–4672.

Chenn, A., Braisted, J. E., McConnell, S. K., and O'Leary, D. D. M. (1997). Development of the cerebral cortex mechanisms controlling cell fate, laminar and areal patterning, and axonal connectivity. In *Molecular and Cellular Approaches to Neural Development* (W. M. Cowan, T. M. Jessell, and S. L. Zipursky, Eds.), pp. 440–473. Oxford Univ. Press, New York.

Cook, P. M., Prusky, G., and Ramoa, A. S. (1999). The role of spontaneous retinal activity before eye opening in the maturation of form and function in the retinogeniculate pathway of the ferret. *Vis. Neurosci.* **16,** 491–501.

Florence, S. L., and Casagrande, V. A. (1990). Development of geniculocortical axon arbors in a primate. *Visual Neurosci.* **5,** 291–309.

Halder, G., Callaerts, P., and Gehring, W. J. (1995). New perspectives on eye evolution. *Curr. Opin. Genet. Dev.* **5,** 602–609.

Hubel, D. H. (1988). *Eye, Brain, and Vision*. Sci. Am., New York.

Irving, C., Flenniken, A., Allduc, G., and Wilkinson, D. G. (1996). Cell-cell interactions and segmentation in the developing vertebrate. *Biochem. Soc. Symp.* **62,** 85–95.

Kandler, K., and Katz, L. C. (1998). Coordination of neuronal activity in developing visual cortex by gap junction-mediated biochemical communication. *J. Neurosci.* **18,** 1419–1427.

Katz, L. C., and Shatz, C. J. (1996). Synaptic activity and the construction of cortical circuits. *Science* **274,** 1133–1138.

Rubenstein, J. L., Shimamura, K., Martinez, S., and Puelles, L. (1998). Regionalization of the prosencephalic neural plate. *Annu. Rev. Neurosci.* **21,** 445–477.

Shatz, C. J. (1990). Impulse activity and the patterning of connections during CNS development. *Neuron* **5,** 745–756.

Tessier-Lavigne, M., and Goodman, C. S. (1996). The molecular biology of axon guidance. *Science* **274,** 1123–1133.

Weliky, M., and Katz, L. C. (1999). Correlational structure of spontaneous neuronal activity in the developing lateral geniculate nucleus in vivo. *Science* **285,** 599–604.

Wernicke's Area

ELEANOR M. SAFFRAN
Temple University School of Medicine

GLOSSARY

neologism Production of a nonword when a word is intended.

phonemic paraphasia A speech error in which some of the sounds of a word are replaced by other sounds.

jargon aphasia A speech pattern in which most of the nongrammatical elements are neologisms.

paragrammatism A pattern denoting the occurrence of grammatical errors in the speech of a fluent aphasic patient.

priming A task in which the presence of one word facilitates processing of another related word.

Wernicke's area, located in the left temporal lobe, was identified by Carl Wernicke in 1874 as the locus of damage of an aphasic syndrome characterized by impairment in language comprehension and production. These patients are impaired in understanding words and sentences, and their speech is riddled with errors, principally affecting the phonological content of words. This article examines the nature of the deficits in detail, along with related disorders such as word deafness. Evidence pertaining to the localization of Wernicke's area is summarized, as is recent functional imaging data that aim at identifying brain regions concerned with the comprehension and production of language.

I. HISTORICAL BACKGROUND

Wernicke's area is named for the physician who first identified the functions associated with this region of the brain. Carl Wernicke, a 26-year-old physician-in-training in the early 1870s, saw several patients with very impaired language comprehension. Their production was disrupted as well. Although they spoke fluently, their speech made little sense; they had difficulty finding words and produced many nonwords (neologisms) along with semantically related word substitutions. When one of the patients died, Wernicke was able to examine her brain. He found evidence of a cerebral infarct (a stroke) involving the middle third of the posterior superior temporal gyrus of the left hemisphere, extending into the middle temporal gyrus below (Fig. 1). (Note, however, that this was not a pure case because the patient also suffered from dementia.) Since the affected area lay close to primary auditory cortex, which receives auditory input from the thalamus, Wernicke reasoned that this region must be part of the auditory association cortex, where auditory input would undergo further analysis. He concluded that this area of the brain must serve as a second "center" for language in the left hemisphere, in addition to the articulatory center in the frontal lobe identified a decade earlier by Paul Broca.

Wernicke's interpretation of his findings was guided by the reflex physiology that dominated neurological thinking at that time. A central assumption was that motor acts were triggered by sensory input. Wernicke took the temporal lobe area to be the repository of "auditory word images," which were the product of language learning. Because it seemed obvious that children must learn the sounds of the words of their

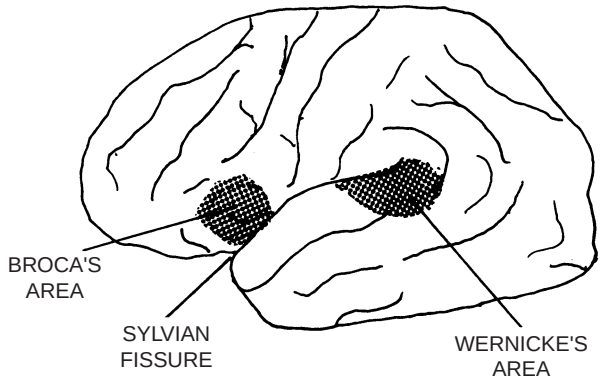


Figure 1 Location of Wernicke's and Broca's areas.

native language before they acquire the ability to produce them, he suggested that these images were not only essential for comprehension but also guided word production. Hence, the sensory and motor centers for language must be connected. Wernicke predicted that disruption of the connections between the two would give rise to a syndrome in which production was impaired but comprehension preserved because the superior temporal area would not be disrupted.

In the mid-1880s, Wernicke's theorizing was extended by Ludwig Lichtheim, who added a "concept center" to Wernicke's model. He termed it a center, although he viewed this information as diffusely represented in the brain. The concept center would serve as the end point for the comprehension of language as well as the starting point for the generation of spoken utterances. Lichtheim developed the model depicted in Fig. 2. The concept center is labeled B in his diagram; Wernicke's area is labeled A and Broca's area M. The only regions that Lichtheim localized in the brain were A and M because he accepted Broca's and Wernicke's placements for these centers.

Lichtheim's model generated predictions for additional syndromes that had not yet been identified. These included the transcortical aphasias—sensory and motor—in which the concept center was isolated either from Wernicke's area (transcortical sensory aphasia; disruption at area 6 in Fig. 2) or Broca's area (transcortical motor aphasia; disruption at 4). In both cases, repetition of the speech of others would be relatively preserved because the connection between Broca's and Wernicke's areas remained intact; however, comprehension and production would be disrupted in transcortical sensory aphasia, and production would be disrupted in the case of transcortical motor aphasia. In contrast, repetition of the speech of others would be impaired by a lesion at 3, which interrupted the connection between A and M (conduction aphasia); as Wernicke predicted, the patient would generate error-ridden speech. Lichtheim also postulated a disorder in which speech perception would be disrupted by removing the sources of auditory input to Wernicke's area (lesion at 7; pure word deafness) and a disorder of production that resulted from disrupting the connections between Broca's area and lower brain centers that control the articulatory musculature (lesion at 5; aphemia). All these disorders were subsequently identified. This approach to the explanation of neurologically based cognitive disorders came to be known as connectionism. This term should not be confused with its current usage, which applies to computer-implemented models that simulate cognitive functions. In these models, connections support the flow of activation from one cognitive subsystem to another.

Although the connectionist approach seemed to be productive, it soon became the target of criticism. One of the early critics was Sigmund Freud, who argued in his 1891 book on aphasia that the approach was overly simplistic. Freud maintained that connectionism could not account for the complexities of language function; nor, he contended, could it explain the language disorders that resulted from damage to the brain. He illustrated this point by attempting to use the connectionist model to generate the symptom patterns of aphasia cases that had been reported in the literature. Today, most would agree that a major limitation of the early work is that it focused on single words. The

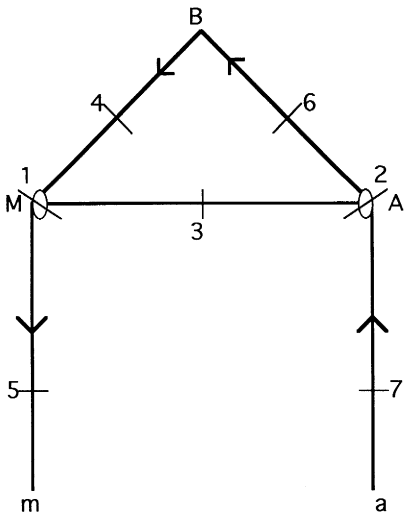


Figure 2 Lichtheim's connectionist model. Numbers indicate sites of lesions. See text for explanation.

connectionist model did not deal with sentence production and comprehension and the syntactic issues raised by these more complex language behaviors.

For much of the first half of the 20th century, researchers rejected the connectionist model and adopted a more diffuse and wholistic view of the representation of language in the brain. In the 1960s, the connectionist approach was revived, largely due to the work of neurologist Norman Geschwind, who offered accounts of a range of cognitive disorders in terms of the interruption of pathways between centers in the brain. Geschwind's theorizing was largely based on case studies of patients with highly selective impairments. At approximately the same time, Hildred Schuell (a speech pathologist) and colleagues at the University of Minnesota promoted a more diffuse view of language organization, arguing that the language area of the brain operated as a whole. Their conclusions were based on factor analyses of data obtained from an unselected series of aphasic individuals, which may have included patients with extensive brain lesions. Similar differences persist in the study of aphasia today: Some researchers advocate intensive single-case studies of patients with specific disorders, whereas others base their conclusions on studies of groups of patients defined by one of the classical aphasia syndromes.

II. FUNCTIONAL DEFICITS ASSOCIATED WITH DAMAGE TO WERNICKE'S AREA

A. General Characteristics

On the widely used Boston Diagnostic Examination for Aphasia (BDAE), constructed by Harold Goodglass and Edith Kaplan in the 1970s, Wernicke's aphasia is defined by impaired comprehension, difficulty with repetition, word-finding deficits, and fluent but paraphasic speech (i.e., the frequent production of speech errors). The BDAE profile that defines the syndrome is illustrated in Fig. 3. Additional information on the performance of a Wernicke's aphasic, NC, studied in our laboratory, is provided in Table I. These data reflect a more psycholinguistic approach to the analysis of aphasic impairments than that of the BDAE. It is important to note two characteristics of this patient: First, his ability to discriminate phonemes is not impaired, at least when they are presented in close proximity; second, his performance with written input is better than his performance with auditory input. It appears, however, that he loses phonological information rapidly. This is demonstrated not only by his difficulty in discriminating two syllables when they are separated by a 5-sec delay but also by his tendency to generate semantic errors in repeating single words (e.g., "windmill" → "lighthouse, something to do with light"). It should be noted, however, that some Wernicke's aphasics perform more poorly than NC on tests of phoneme discrimination, and that some are more impaired with printed words. Other characteristics of these patients include rapidity of spoken output, a pattern referred to as "logorrhea" or "press of speech." Another significant finding is that most of them do not seem to be aware of their speech errors, at least in the initial phase of their illness.

B. The Comprehension Problem

The Russian neuropsychologist Alexander Luria attributed the comprehension deficit in Wernicke's aphasia to a phoneme discrimination problem. He found such impairments in 95% of World War II missile wound casualties with lesions in the superior temporal gyrus. Subsequent investigations did not support this finding. In general, there appears to be little correlation between phoneme discrimination deficits and auditory comprehension impairments in aphasics. One possible explanation of Luria's results is that they reflect the nature of the population he tested. Missile wounds are likely to penetrate into the white matter that delivers input to the superior temporal lobe. As discussed later, depriving the left temporal lobe of auditory input does indeed result in severe phoneme discrimination deficits.

Sheila Blumstein and colleagues found that Wernicke's aphasics were less impaired on phoneme discrimination tasks than other language-impaired patients, and that their performance was dependent on the lexical status of the stimuli. The Wernicke patients demonstrated better discrimination between words than nonwords, a pattern similar to that of other aphasic groups that were tested. Another significant finding from these investigators was that Wernicke's patients were more likely to err by selecting semantically than phonologically related foils on a word–picture matching task, although they did show a greater effect of phonological similarity than Broca's aphasics. Blumstein and colleagues also found that some aphasic patients who did well on phoneme discrimination performed poorly on tasks that required them to identify the syllable they heard

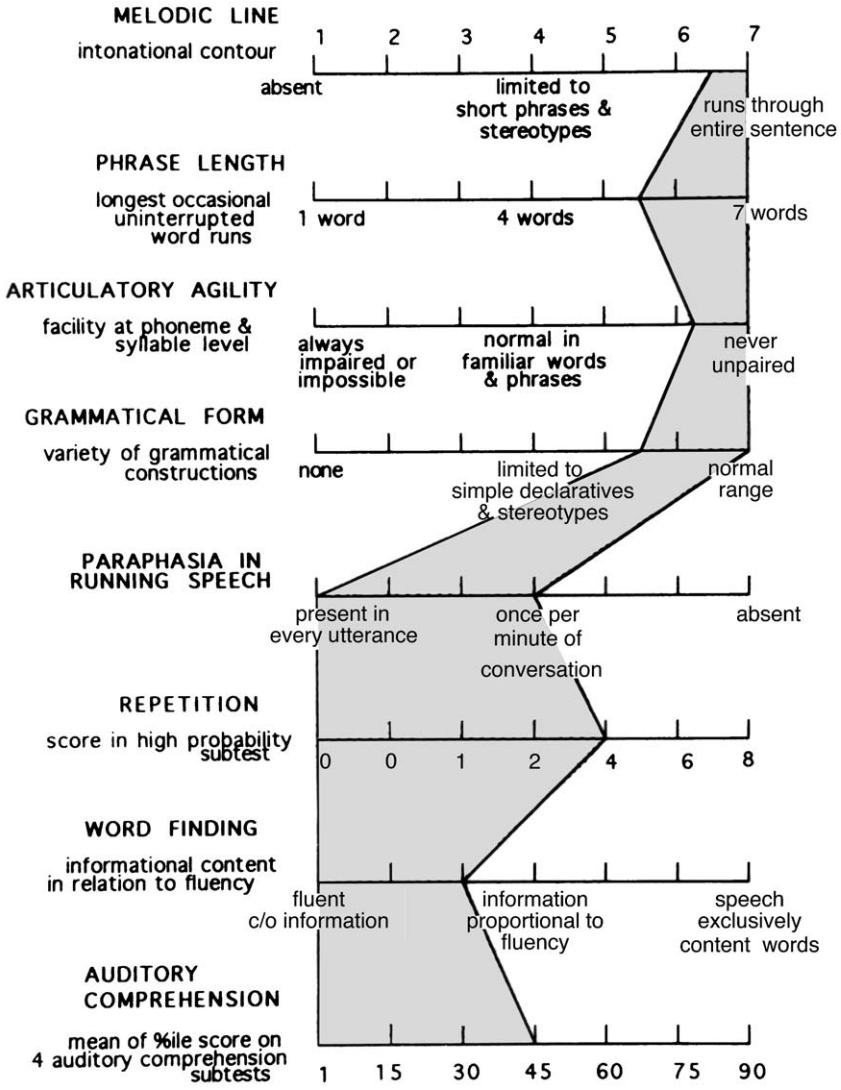


Figure 3 Range of Wernicke's aphasic scores on the Boston Diagnostic Examination for Aphasia (redrawn with permission from Goodglass and Kaplan, 1983).

(e.g., by pointing to the matching syllable in a written array). Many, but not all, researchers assume that phoneme identification is a critical step in language processing. As Blumstein has also noted, speech perception deficits may underlie language comprehension deficits, but these deficits are revealed only in the context of larger streams of speech. Therefore, focusing on the perception of segments in isolated words or syllables may not be a sufficiently sensitive index. This conclusion is supported by the data from NC (Table I), which indicate that phonemes may be perceived accurately but that this information is rapidly lost.

Other findings suggest some imprecision in the responses of Wernicke's aphasics to auditory speech input. For example, Blumstein and colleagues found that Wernicke patients responded to stimuli that were phonologically similar to semantically related words (e.g., "wat" or "gat", where the target was "dog"), whereas Broca's aphasics did not. These words were presented in a priming task, in which the nonword preceded a lexical decision response ("Is the stimulus a word or not?") to the word target. In the Wernicke patients, the phonologically similar nonword speeded the response to the word target in the lexical decision

Table I
Data from Wernicke Patient NC

Peabody Picture Vocabulary Test (matching spoken word to one of four pictures)	< First percentile for normal subjects	
Word-to-picture matching (four choices)		
Within-category choices ($n = 15$)	69%	
Across-category choices ($n = 28$)	100%	
Word similarity judgment (two most related of three)		
Nouns ($n = 15$)	73%	
Verbs ($n = 15$)	87%	
Phoneme discrimination (e.g., pa-ta; $n = 80$)		
no delay	100%	
5-sec. delay	81%	
Lexical decision		
Auditory		
High imageability words ($n = 40$)	56%	
Low imageability words ($n = 40$)	25%	
Nonwords ($n = 80$)	99%	
Written		
High imageability words ($n = 40$)	93%	
Low imageability words ($n = 40$)	95%	
Nonwords ($n = 80$)	94%	
Word naming	Repetition	Oral reading
High imageability ($n = 40$)	57%	85%
Low imageability ($n = 40$)	25%	63%
Auditory sentence comprehension ($n = 60$)		
Lexical choices (1 of 2; $n = 60$)	90%	
Reversible choices (1 of 2; $n = 60$)	68%	

task. In another study, Blumstein and colleagues investigated lexical effects on phoneme classification in aphasics. These studies were based on findings in normal subjects indicating that the location of the boundary between similar phonemes (e.g., /d/ and /t/) depends on the context in which the phoneme is embedded. In normal subjects, the boundary shifted toward /d/ responses when the /d/ stimulus was a word (''dash-tash'') and in the direction of /t/ when the /t/ stimulus was a word (''dask-task''). These investigators found that Broca's aphasics showed the lexical effect—in fact, one that was exaggerated relative to control subjects—whereas the Wernicke patients did not. This finding suggests that the Wernicke's aphasics were less sensitive to the phonological differences among the stimuli: That ''dask'' was as effective as ''task'' in contacting the lexical representation of task. A tendency toward less discriminating responses may reflect some degradation of the lexical network.

One finding that might seem surprising, given the relatively poor performance of Wernicke patients on word comprehension tasks, is that they demonstrate semantic priming effects. Priming refers to the effect of presenting a word that facilitates processing of a second word (e.g., presenting ''horse'' and then the semantically related word ''sheep''). These effects are usually assessed in lexical decision tasks (deciding whether a stimulus is a word or not), in which presentation of a related word speeds the decision to the target word. Sheila Blumstein, William Milberg, and colleagues have shown that Wernicke patients demonstrate semantic priming effects in lexical decision tasks. These effects have been demonstrated with written as well as spoken words. Data cited previously indicate that Wernicke patients frequently make semantic errors in tests of word comprehension. For example, note the relevant data for NC in Table I: He performed well in a word–picture matching task when the foils came from different semantic categories (the

across category condition) but not when they came from the same semantic category (the within-category condition). The priming data suggest that information with respect to semantic category is likely to be retained, although the specific features that differentiate one category member from another may not be accessible from word stimuli.

Wernicke's aphasics are impaired on sentence processing tasks as well. This is evident in the data from NC presented in Table I. Data from 10 Wernicke's aphasics tested in our laboratory on the Philadelphia Comprehension Battery are summarized in Fig. 4. Note that their average scores consistently lie below the aphasic mean (based on data from 64 aphasics). In particular, they perform quite poorly on the grammaticality judgment subtest, in which they are asked to determine whether a spoken sentence adheres to the rules of English syntax, and on the sentence comprehension subtest with lexical distractors, in which one of the pictures depicts an entity or relationship that is not present in the sentence. In contrast, Broca's aphasics ($n = 9$) perform well on these tasks (92 and 97% correct, respectively), although they also break down on the reversible subtest of sentence comprehension (76% correct). The sentence comprehension task included a number of different syntactic structures, from simple actives (e.g., "The girl washed the boy") to passives (e.g., "The boy was washed by the girl") and object relatives (e.g., "The boy that the girl washed was tall"). Although the Broca's aphasics showed a clear effect of syntactic complexity (e.g., worse performance on passives and object relatives compared with active voice sentences), statistical analyses revealed no effect of structural complexity in the Wernicke's aphasics. In Serbo-Croatian, a language that relies heavily on affixes to mark grammatical distinctions, Katarina Lukatela and colleagues found that Wernicke's aphasics also performed more poorly than Broca's aphasics on a grammaticality judgment task. Additional data come from a sentence comprehension study conducted with German-speaking aphasics. Jacqueline Stark and Rudolf Wytek found that Wernicke patients were more likely than Broca's aphasics to choose distractors that incorporated an agent (doer of the action) that was not mentioned in the sentence or one that depicted an action involving the wrong verb. The Wernicke patients made almost three times as many lexical as syntactic errors; in contrast, the Broca patients made almost twice as many syntactic as lexical errors. Note a similar finding in our own study, in which Broca patients rejected picture choices that contained an entity or relationship that was not present in the sentence (the Lexical Sentence Comprehension subtest), whereas the Wernicke patients frequently selected such alternatives instead of the correct target.

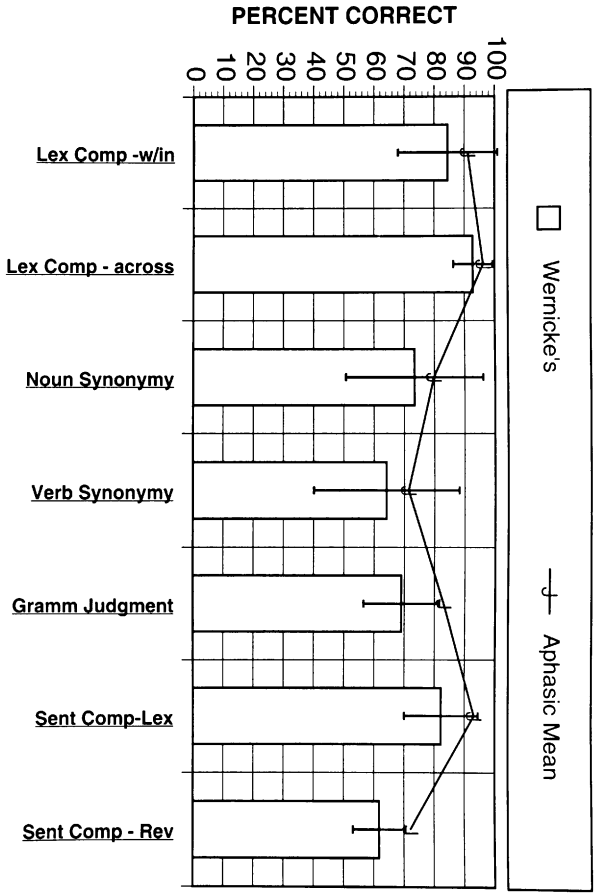


Figure 4 Data for Wernicke's aphasics ($n = 10$) from the Philadelphia Comprehension Battery. Bars indicate standard deviations. The aphasic mean represents average scores for 64 aphasics. Subtests are as follows: lexical comprehension-within—word-to-picture matching with four choices, with the three foils from the same semantic category; lexical comprehension-across—word-to-picture matching with four choices, with the three foils from different semantic categories and with phonologically related foils in some cases; noun synonymy—choose the least related of three nouns, presented in spoken and written form; verb synonymy—choose the least related of three verbs, presented in spoken and written form; grammaticality judgment—sentences representing a range of syntactic forms, half of which violate English syntax; sentence comprehension–lexical—match a sentence to one of two pictures, with the foil containing an item that is not in the sentence; sentence comprehension–reversible—match a sentence (range of syntactic structures) to one of two pictures, with the foil reversing the roles of the two participants.

Although the Wernicke's aphasics resemble Broca's aphasics in demonstrating effects of semantic reversi-

bility, it is difficult to argue that the two sets of phenomena reflect the same underlying impairment. In the case of the Wernicke patients, some of the difficulty with semantic reversibility may reflect confusion with respect to word meaning. Furthermore, it appears from the grammaticality judgment data that the Wernicke's aphasics may not be computing the syntactic structure of the sentence, which the Broca's appear capable of doing, although they are clearly deficient in utilizing the structural information to derive an interpretation for the sentence.

There is an additional comparison worth noting. Semantic dementia is a disorder in which patients have difficulty retrieving words and are impaired in word comprehension as well. This degenerative condition reflects damage to structures outside the classical perisylvian language area—the region on both sides of the Sylvian fissure (Fig. 1) which separates the temporal from the frontal and parietal lobes. Semantic dementia patients show loss of tissue in the inferior, anterior portion of the left (or both) temporal lobe(s). In contrast to Wernicke's aphasics, these individuals perform relatively well on sentence comprehension tasks. To the extent tested, the semantic dementia patients also perform successfully on grammaticality judgment tests, whereas the Wernicke patients do not. It is necessary to conclude, then, that the Wernicke patients are impaired in other aspects of sentence processing in addition to their semantic deficits in word comprehension. However, the nature of these limitations is not clear: Do they involve lexical access, syntactic computation, short-term memory limitations, or perhaps all these mechanisms? Because lexical access is necessary to retrieve syntactic information for words, it is likely that this process is impaired in Wernicke's patients. Also, in view of the Broca patients' greater sensitivity to grammatical violations compared to Wernicke's aphasics, it might be conjectured that the recovery of syntactic structure is supported by left temporal (and perhaps parietal) structures that are damaged in Wernicke's aphasia. Left parietal areas have been implicated in short-term verbal memory, particularly with respect to the retention of phonological information.

C. Word Deafness

There is another impairment that appears to reflect the loss of auditory input to phonological processing areas in the left temporal lobe. This disorder, predicted in Lichtheim's model, is termed pure word deafness. Patients demonstrating this impairment have great difficulty understanding and repeating spoken input, although they speak, read, and write relatively normally. The disorder is rare; it is thought to result from small lesions affecting the pathways that supply auditory input to the left temporal lobe (from the medial geniculate in the thalamus) as well as callosal input from homologous areas of the temporal lobe in the right hemisphere. This account is supported by evidence from a word deaf patient whose lesion involved white matter underlying the left temporal and parietal lobes, leaving Wernicke's area intact. As a consequence of its rarity, all the data in the literature are reported in the form of case studies. In some cases, the lesion is restricted to the left hemisphere; in others, the lesions are bilateral. For example, one patient originally diagnosed as a Wernicke's aphasic following a left temporal lesion became word deaf following a subsequent lesion to the right temporal lobe.

As might be expected from differences in the lesion sites, the behavioral findings are not altogether consistent across patients. Although vowel perception may remain relatively intact, word deaf patients perform very poorly on phoneme discrimination tasks that involve consonants. This difference reflects the nature of the two types of speech sounds: Vowels are longer in duration and have a consistent auditory pattern, whereas consonants involve rapid changes in pitch as well as small differences in the timing of elements of the pattern. There is evidence that the left temporal lobe is specialized to detect these transient changes and timing differences. With respect to consonant perception, some patients have proved more sensitive to place distinctions (the locus of production of the sound, whether it involves the lips or particular placements of the tongue), whereas others appear more sensitive to differences in voicing (whether vocal fold vibration begins with the onset of the sound, as in /b, d, g/ or is slightly delayed, as in /p, t, k/). These individuals are generally able to distinguish speech from other forms of auditory input, but some of them report that the speaker appears to be conversing in a foreign language. In a few cases, dichotic listening tests, which involve presenting two different inputs simultaneously to the two ears, have shown that patients respond better to stimuli delivered to the left ear, which projects more extensively to the right temporal lobe than the left. This finding suggests that these patients are utilizing right temporal lobe areas to process speech input. These data are consistent with the view that the left temporal lobe is specially equipped to deal with the perception of speech.

D. Production in Wernicke's Aphasia

With respect to production, Wernicke patients speak fluently, although it may be difficult to understand what they are trying to say. The sample in Table II illustrates this pattern. Errors include semantic substitutions as well as the production of nonwords (neologisms). The nonwords are sometimes divided into two classes: phonemic paraphasias, which bear some phonological similarity (often defined as 50% or more overlap) to the target word, and nonwords that bear little or no similarity to the target, sometimes referred to as abstruse neologisms. In severe cases, the majority of substantive words (nongrammatical words) are neologisms, a production pattern termed jargon aphasia. It should be noted that these nonwords are nevertheless consistent with the phonotactic constraints of the speaker's language—that is, their phoneme sequences do not violate the language's combinatory rules.

In addition, running speech usually contains grammatical violations, a pattern known as paragrammatism. This production pattern is distinguished from agrammatism, which is found in nonfluent Broca's aphasics with frontal lobe damage, where sentence structure may be absent or vastly simplified. In many cases, the paragrammatic errors involve omissions of grammatical elements or substitutions of one for another. There are also instances in which the errors appear to reflect combinations of two syntactic structures (e.g., "I feel fine"/"I'm feeling fine" → "I'm feel fine"; Table II). Wernicke's aphasics generally appear insensitive to the errors they are producing, suggesting that they are not monitoring their output or not doing so successfully. This pattern contrasts with that of conduction aphasics, who also produce phonological errors but generally attempt (often without success) to correct them. In a study of word production in Wernicke's aphasics carried out in 1982 by Sarah Williams and Gerald Canter, it was found that the patients performed better on picture description tasks, in which the word was generated in the context of a sentence, compared with a task that simply required them to name single items in pictures. This finding suggests that the sentence context facilitates production of the word in these patients, possibly as a result of activation from the syntactic frame for the sentence that stipulates selection of a word from a particular syntactic category (e.g., noun, verb, and adjective). It is interesting that the opposite effect—better performance on picture naming than on picture description—was obtained in Broca's aphasics, who generally have difficulty with sentence production. Another recent report, by Helen Bird and Sue Franklin, examined the relationship between word and sentence production in a Wernicke's aphasic. As word retrieval improved, so did the complexity of the patient's utterances. This finding does not prove that there is a causal relationship between the two language functions, although it does seem reasonable that an increase in the ability to retrieve words would facilitate the construction of more complex sentences.

Clearly, one aspect of the problem with word production in Wernicke's aphasia reflects difficulty in accessing their phonological forms; this difficulty could account for errors that bear a phonological relationship to the target word. However, it is likely that semantic input to the lexicon has also been degraded. It has sometimes been suggested that errors that bear no phonological resemblance to the target word (so-called abstruse neologisms) reflect a combination of the two types of errors: First a semantic error

Table II
Speech Sample from a Wernicke's Aphasic (FL)

Examiner: OK, so tell me what the problem's about.

FL: From the doctors, wh'where, from, from what **eh'spect**?

Examiner: Tell me what you think your problem is.

FL: Ah, where do I start the **tesseinemen** from? They tell me that my brain, physically, my brain is perfect, the attitudes and everything is fine, but the silence now, that I have to **reeh-learn** through **edgit** again, physically nothing wrong with it. It's perfect the doctors tell me. They have attitude. Physically I have **loozing** absolute nothing on my head, but now I have to go through these new attitudes to **looalize** how, some, how can I say to ? Some what that I can **reeh-learn** again so I can **estep** my knowledges, so th'you **kyou** again, what how can I say that …

Examiner: That's OK, I got you, it's clear … so how are you feeling?

FL: I'm feel fine. My wife, I love wife. I'm uh sorry that, I still feel bad, that **he** has been **she** upset by this thing, from 'till three weeks ago nothing was goin,' you know. **Sutting** happened to her, so I feel for her. She was caught, has a, she's been caught in a **herrible** situation by **meem**, that, that, uh there's just no way, you know, I'm just beginning from the doctors that I've learned now, they had my **doctors** open for eight day, **kuntinueh** days.

is generated, and this is followed by failure to access the correct phonological form of the semantically related word. Another proposal, from Brian Butterworth, is that abstruse neologisms are generated by a neologism-generating device that is employed when access to the phonological form of a word is blocked. Butterworth found that the production of neologisms that bore little relationship to the presumed target word followed longer hesitations in running speech compared with other kinds of errors (such as phonemic paraphasias), suggesting that they were generated by a different mechanism. Yet another possibility is that the errors result from the utilization of phonemes that have high levels of activation. Evidence from studies of normal subjects indicates that words that are semantically related to the target are activated along with the target item; if activation of the target word or its phonological form is relatively weak, this would allow phonemes from other words to intrude into the response. Currently, the source (or sources) of these errors, which include phonemic paraphasias as well as errors that are more remote from the target word, has yet to be determined.

Many words represent combinations of morphemes—either inflectional, as in "walk" + "ed", or derivational, as in "distribute" + "tion." Butterworth noted that inflectional affixes were used appropriately by the jargon aphasic he studied. A study of an Italian jargon aphasic, carried out by Marta Panzeri and colleagues, noted the generation of neologisms formed by adding legal derivational affixes to nonword roots. Although these were clearly nonwords, the word category reflecting the affix (noun, verb, adjective, or adverb) was appropriate to the syntactic context in which the erroneous item appeared. For example, the affix "—mente" in Italian turns an adjective into an adverb; the neologism "atamente" occurred in the appropriate context for an adverbial modifier. Thus, it appears that although the patients cannot retrieve the correct root morphemes, they are nevertheless capable of generating nonwords with affixes that are appropriate to the syntactic context.

There is evidence that semantic information is disrupted by lesions affecting the anterior, inferior portions of the left temporal lobe, the area damaged in semantic dementia and in herpes simplex encephalitis, which also results in semantic impairment. The lesion sites in Wernicke's aphasia are likely to sever connections between this region and word forms presumptively stored in the superior temporal gyrus, which may also be lost if this region is damaged as well. Note that in contrast to semantic dementia patients, who have difficulty with semantic tasks that involve pictures, Wernicke's aphasics may perform well on such tests, indicating that conceptual information per se has not been affected. For example, patient NC scored at a normal level (98% correct) on the pictorial version of the Pyramids and Palm Trees test, in which a target picture must be matched to one of two alternatives (e.g., a pyramid to a palm or fir tree). In contrast, he was only 69% correct in the within-category word to picture matching task, in which a spoken word must be related to one of four pictures (e.g., "strawberry" to pictures of a strawberry, watermelon, pear, and pineapple).

In some cases the difficulty with word retrieval may decrease over time, accompanied by some improvement in word comprehension. Hence the frequent change in diagnostic category from Wernicke's aphasia to conduction or anomic aphasia. For example, the profile of patient NC (Table I) changed to that of a conduction aphasic over a period of 18 months. Also, in a number of published cases of Wernicke's aphasia, the production of written words was found to be much better preserved than oral naming. One exceptional case, a Wernicke's aphasic studied in our laboratory, had a lesion involving most of the left temporal lobe. Although his responses on an oral picture-naming test consisted mostly of neologisms, his spontaneous speech, sampled 4 years after the lesion occurred, was free of such errors. To compensate for his deficit, this patient substituted orally spelled words (not always spelled entirely correctly, but recognizable), which were inserted at the proper locations in his utterances. This patient, as in other cases in which written word production is superior to oral production, made semantic errors in writing words to dictation. As has also been reported for other such patients, his sentence writing was agrammatic. We have suggested that this performance pattern may reflect right hemisphere support for written word generation in these patients.

With respect to the production of connected speech, some researchers have noted the persistence or overuse of certain phonemes in Wernicke's aphasics. For example, Butterworth found that the samples from a jargon aphasic contained many examples of infrequent phonemes, such as /z/. One published analysis, by Myrna Schwartz and colleagues, examined the production of the patient FL (whose speech sample is provided in Table II) for the occurrence of anticipation and perseveration errors. Anticipatory errors reflect the production of phonemes in advance of their correct position in the utterance; perseveratory errors reflect

the recurrence of phonemes after they have been produced. Schwartz and colleagues' analysis of published normal speech corpora revealed that anticipatory errors prevailed. However, a bias toward perseverations was noted in the early responses of normal subjects who were learning to produce tongue-twisters (e.g., "flea-free fruit flies"), although there was a shift in the direction of anticipatory errors toward the end of the learning trials. FL's analysis demonstrated a preponderance of perseveratory errors (e.g., "me" → "meem"; Table II). This tendency may reflect the utilization of phonemes that remain activated after they are produced; these phonemes may have higher activation levels than phonemes that are appropriate to the planned utterance, which may be weakly activated in the impaired lexicons of such individuals. In contrast, anticipatory errors might be considered "good" errors in that they reflect the retrieval of upcoming units in the phrase.

Although the speech of Wernicke's aphasics is subject to paragrammatic errors, their production is more syntactically complex than that of Broca's aphasics, whose speech is often characterized as agrammatic. This observation suggests that the retrieval of syntactic form for the purpose of generating sentences may not depend on Wernicke's area. The finding that Wernicke's aphasics are better at retrieving words in sentence contexts than in single picture naming tasks provides additional support for this possibility.

One major question concerns the relationship between the comprehension and production impairments in Wernicke's aphasia, which remains unclear. Do the two deficits reflect reliance on a common set of word forms, or are there different phonological representations of words for comprehension and production? Those who are concerned with the architecture of the language system view this as a central question, which has yet to be resolved.

E. Repetition

As is evident from the profile in Fig. 3, repetition is impaired in Wernicke's aphasia. This is to be expected, given the problems in the perception and production of words, along with what appears to be the rapid loss of phonological information. Examples from NC's single word repetition performance include the following: "drum" → "rum," "train" → "tree," "dragon" → "wagon," "camel" → "captain." Not surprisingly, NC's sentence repetition was seriously impaired, as is evident from the following examples: "Each night the guard checks all the locks" → "Security checks all the doors every night," "While walking under an elm, the man was hit on the head" → "A man was walking down the street and fell down."

One additional pattern deserves mention. In some cases of Wernicke's aphasia, the patients produce semantic errors in single word repetition tasks (e.g., "tobacco" → "cigarettes"). This pattern is termed deep dysphasia, an analogy to the reading disorder termed deep dyslexia, which is characterized by the production of semantic errors in reading single words aloud. The semantically related response indicates that the input signal was processed to a deep, or semantic, level. Other features of this disorder include great difficulty repeating nonwords and better performance on concrete rather than abstract words. One possible account of this pattern is that it reflects abnormally rapid loss of phonological information. The patients hear the word, process it semantically, and regenerate the word from the semantic representation without reference to the phonological form of the input signal. The difference between abstract and concrete words could reflect the richer and less contextually dependent representations attributed to the latter. Deep dysphasics typically perform very poorly on verbal short-term memory tasks, which require them to retain and report strings of digits or words. Patient NC, who demonstrated very rapid loss of phonological input (Table I), was a deep dysphasic. In a functional imaging study of a deep dysphasic patient, Dominique Cardebat and colleagues found high levels of activation in the right temporal lobe, which suggests that this area may support the processing of auditory verbal input in such patients. If this is the case, the generation of semantic errors in word repetition could reflect the nature of right hemisphere semantic relationships, which have been characterized as "coarsely coded" relative to those of the left hemisphere. Furthermore, the phonological capacities of the right hemisphere are thought to be limited.

F. Computational Models

During the past 15 years, a number of investigators have turned to computational models—language processing models implemented as computer simulations—in an attempt to further understand the nature of the deficits in aphasia. One approach has been to utilize, or develop, a model that accounts for normal language behavior and then to "lesion" the model in

order to generate impaired performance. For example, Gary Dell and collaborators took a model that captured features of normal performance in a picture-naming task and created lesions to simulate the error patterns (quantitative distributions of correct responses as well as several error types) generated on the same task for each of 21 aphasics. The model stipulates a set of semantic features for words, which connect to word entries in a lexicon; the word entries then connect to their component phonemes. In this model, the connections are bidirectional; that is, activation flows downward from semantics to the word level to phonemes and back up from phonemes to the word representations and then to the semantic level. The word with the highest activation is selected for production. One way to inflict damage on the model is to reduce the strength of the connections in the network, which decreases the flow of activation between levels. Dell and colleagues showed that reducing the strength of the connections in the model resulted in the production of nonword errors. Intuitively, this is obvious because this lesion essentially disconnects phonemes from the semantic features that are initially activated. Another way to damage the model is to increase the rate of decay of activation in the model. Extreme changes in decay rate will also generate nonword errors, although at intermediate levels of connection weight and decay lesions, the error patterns differ, with the decay rate lesions yielding larger numbers of word substitutions that are semantically or phonologically related to the target word. Overall, the simulations of the performance of the 21 aphasics were quite successful in generating the pattern of each patient's responses. The sample in this study included 5 Wernicke's aphasics. In one case, the lesion was modeled by a reduction in connection strength and in two others by an increase in decay rate; the others were simulated by changes in both parameters. One possible account of the errors generated by Wernicke's aphasics entails reduction in the connection strength among levels of representation in the internal lexicon and/or increases in the rate of decay of activation. However, although the simulations captured the rates of production of various kinds of errors, the model is only capable of generating monosyllabic words. For this reason, there was no attempt to simulate the features of the errors in phonological form—the phonological paraphasias and neologisms that are so common in Wernicke's aphasia. In many cases, these involve the intrusion of phonemes that are not present in the utterance, a rarity in normal speech errors. Attempts to simulate these error patterns await the development of models that generate multisyllabic words, which can then be "lesioned" to generate phonological errors.

III. ANATOMICAL ASYMMETRY AND EVOLUTION OF WERNICKE'S AREA

For many years, neuroanatomists were struck by the similar appearance of the two cerebral hemispheres in the human brain, despite clear evidence of differences in their function. However, measurements performed by Norman Geschwind and Walter Levitsky in the late 1960s uncovered significant differences between them. In particular, the human planum temporale, the area at the junction of the temporal and parietal lobes, was larger in the left than in the right hemisphere in 65% of the brains they examined. This asymmetry is associated with a longer Sylvian fissure on the left side of the brain (Fig. 1). The planum temporale is the location of much of Wernicke's area. Although originally interpreted as the consequence of left hemisphere specialization for language, the same asymmetry has recently been described in chimpanzees. It is possible that the enlargement of this area in other primates led to its eventual support for language function.

In a recent functional magnetic resonance imaging (fMRI) study carried out by Amanda Bischoff-Grethe and colleagues, it was found that the activation of Wernicke's area, as well as its right hemisphere homolog, was negatively correlated with the temporal predictability of a color sequence pattern. The subjects in this study were instructed to perform a different task and were not consciously aware of the predictability of the pattern. The authors suggest that the predictability of sequential information is useful in processing both the phonological and the syntactic aspects of language, and that this capability could conceivably have given rise to the utilization of Wernicke's area for language function.

IV. LOCALIZING WERNICKE'S AREA

Wernicke located his second center for language in the superior temporal gyrus, with some extension into the middle temporal gyrus below. Bogen and Bogen, in a 1976 article titled "Where Is Wernicke's Area?," noted that this localization was not consistently supported by subsequent investigators. Some placed it largely in the parietal lobe, encompassing the angular and supramarginal gyri with some extension to the superior and

middle temporal gyri; some excluded parietal cortex altogether; others extended it into the middle temporal gyrus, whereas some did not. The Russian neuropsychologist Luria, who studied military casualties in World War II, located the center of the area in the superior temporal gyrus, where 95% of the cases exhibited a phonemic perception deficit; of those with lesions in the parietal area posterior to it, 53% had such an impairment. Luria took a probabilistic view of the extent of Wernicke's area, which is probably the correct approach given that brains differ across individuals and that lesions are rarely the same from one case to another. Moreover, language is the product of learning, and there may well be variability across individuals with respect to the brain region(s) that supports this knowledge.

We are in a stronger position today because we can draw not only on computed tomography (CT) and MRI scans to localize the lesions of neurological patients but also on functional imaging studies that track brain activation (indirectly via blood flow to active brain areas) as normal individuals perform language tasks. In the case of patient studies, two factors are critical. The first is temporal. CT scans will not reveal evidence of ischemic infarctions (those that reflect blockage of an artery) soon after the event because cell death and clearing of the infarcted area are required to differentiate the affected area from normal brain. However, if one waits too long, the behavior pattern may change; in many cases, those initially diagnosed as Wernicke's aphasics transmute into conduction or anomic aphasics over time, reflecting an improvement in comprehension and, in some cases, production as well. The neurological basis for the amelioration of symptoms is not always clear, although in some cases there is evidence for right temporal lobe involvement, as indicated by a decline in performance following a right hemisphere lesion or an increase in right hemisphere activation over time in functional imaging studies. Second, it is necessary to adopt clear and replicable parameters for the diagnosis of Wernicke's aphasia. This can be accomplished by using one of the aphasia batteries that specify criteria for such a diagnosis, such as the Boston Diagnostic Examination for Aphasia or the Western Aphasia Battery.

All of the studies that have gathered CT, MRI, or the earlier generation of radionucleotide scans from Wernicke patients have noted variability in the lesions. This is to be expected because the disorder generally results from vascular injuries that vary in the extent of the damage to the circulatory system and, hence, the brain region that is affected. Andrew Kertesz, who superimposed radionucleotide scans from 13 Wernicke's aphasics, found the area of maximum overlap of the lesions in the superior temporal gyrus of the left hemisphere, with some extension into the parietal lobe, both dorsally and posteriorly. Nina Dronkers and colleagues analyzed the lesions in 6 chronic cases of Wernicke's aphasia identified by the Western Aphasia Battery. They found that the region of maximum overlap lay in the middle rather than the superior temporal gyrus; the underlying white matter was affected as well. A sixth patient's lesion was restricted to subcortical white matter, possibly isolating Wernicke's area from input sources. A recent study that examined MRI scans in 10 Wernicke's aphasics noted damage to the posterior portion of the temporal gyri in 7, damage to the anterior portion of the temporal gyri in 6, and damage to the inferior parietal area in 4; lesions in the underlying white matter were present in 8 of the patients. Currently there is disagreement with respect to the localization of Wernicke's area. What is critically needed is a lesion overlap study that includes a larger set of patients, along with the stipulation of quantitative criteria for defining the syndrome.

Regarding functional imaging [positron emission tomography (PET) or MRI] studies, which record blood flow to active areas in the brain, there is evidence that auditory stimuli activate superior temporal cortex bilaterally. The region activated by words as well as nonwords generally includes Brodmann's areas 22, 41, and 42 in both cerebral hemispheres, identified on the basis of histological differences early in the 20th century. All these regions are located in the superior temporal gyrus. When activation patterns for nonwords are subtracted from those for words, much of the left superior temporal activation disappears. However, there is evidence that auditory word stimuli activate areas located more ventrally and posteriorly in the left superior temporal lobe, along with the middle temporal gyrus.

A study conducted with French subjects demonstrated that the left superior temporal lobe was activated by spoken stimuli in an unfamiliar language (Tamil), a finding that implicates this area in prelexical processing of speech input. In a recent magnetoencephalographic (MEG) study conducted in Finland, it was found that the left temporal lobe showed a greater response to a Finnish vowel sound compared to a vowel sound taken from another language (Estonian). No human language draws on all the sounds that the articulatory musculature can produce, and there is evidence from work by Peter Jusczyk and colleagues

that infants become familiar with the sounds of their native language in the first year of life. The Finnish finding suggests that this process is supported by the left temporal lobe.

It should be noted that some phonological processing tasks activate other structures along with those in the temporal lobes. For example, Robert Zatorre and colleagues have shown that tasks that require decomposition of the phonological structure of words, such as determining that both cat and pet contain the /t/ sound, activate frontal lobe areas either within or adjacent to Broca's area, along with areas in the superior temporal gyrus. It may be that subvocal articulatory mechanisms participate in such tasks. Parietal regions that have been implicated in phonological short-term memory processes are activated as well.

Additional evidence that implicates left temporal lobe structures in speech perception and comprehension comes from the use of electrical current to disrupt the operations carried out by the stimulated brain area. Dana Boatman and colleagues examined these effects in three patients with indwelling electrode arrays implanted prior to surgery for intractable epilepsy. These arrays are often used to localize language areas prior to surgery so that these regions can be avoided when tissue is removed. Three types of tasks were administered, with and without electrical stimulation: phoneme discrimination (e.g., pa-ta or pa-pa), phoneme identification (matching a CV syllable to an array of four written choices), and comprehension (matching a spoken word to one of four pictures; Token Test, a measure of comprehension in which the subject manipulates a set of tokens in response to spoken instructions). Stimulation sites in the left superior temporal gyrus elicited several different patterns: comprehension impaired but discrimination and identification spared; comprehension and identification impaired but discrimination spared; and discrimination, identification, and comprehension all impaired. In all three patients, the sites where stimulation disrupted all three functions were located more anteriorly than the sites where comprehension alone was impaired. The behavioral data are in accord with evidence from patients, in that phoneme identification has been found to be impaired independently of discrimination; however, patients impaired on discrimination tasks are also impaired on phoneme identification. As noted previously, there is other evidence that comprehension can be impaired independently of phoneme discrimination and identification because phoneme discrimination scores do not correlate well with comprehension measures in aphasics.

Sentence comprehension has also been examined using functional imaging techniques. Several studies have shown greater activation of Broca's area as a function of increased syntactic complexity. In other studies, activity in both Wernicke's and Broca's areas increased with the structural demands of the sentences. A recent fMRI study of sentence processing showed that both areas were activated, along with a region in the anterior temporal lobe that was active in an earlier PET study involving sentence materials. Finally, Japanese investigators, using a new imaging technique that tracks blood oxygenation through probes attached to the skull (optical topography), found that left superior temporal cortex was activated as subjects listened to a story.

With respect to production, Peter Indefrey and Willem Levelt recently reviewed the literature, which includes functional imaging studies as well as experiments that employ electrophysiological recording and stimulation. Word repetition activates the middle and posterior superior temporal gyrus on the left as well as the middle temporal gyrus. As might be expected, repetition of nonwords involves less activation of the middle and posterior superior temporal gyrus because no other information is associated with these stimuli. Other tasks, such as picture naming, activate a wider range of structures in addition to those implicated in repetition; these include the anterior superior, middle, and inferior temporal gyri as well as the rest of the middle and superior temporal gyri. Additional structures in the frontal lobes and cerebellum are involved in the motor programming of speech.

V. CONCLUSION

The syndrome described by Wernicke more than 125 years ago is still observed today. These individuals demonstrate impaired comprehension at the word and sentence level as well as disordered production. The impact of this disorder on communication can be devastating, although the patient (at least initially) may not be aware of the extent of his or her deficit. The implementation of recent technological advances, such as functional imaging and brain stimulation studies, demonstrates that the superior temporal lobe is involved in the perception of spoken language and also in its generation. Thus, it appears that Wernicke was correct in attributing auditory language processing and aspects of word retrieval to the left temporal lobe. What remains to be delineated is the precise localization of the mechanisms that support these

processes. Answers to these questions are likely to come from functional imaging studies, particularly as their capability for fine-grained resolution of brain structures increases over time. Another important question concerns the relationship between word comprehension and production, which has yet to be defined: Is there a single set of word forms that serve both functions, or are receptive and productive aspects of language served by different sets of word representations? The answer to this question will require additional behavioral investigations in normal individuals as well as further studies of the relationships between comprehension and production in aphasics.

See Also the Following Articles

AGRAPHIA • AUDITORY CORTEX • BROCA'S AREA • LANGUAGE AND LEXICAL PROCESSING • LANGUAGE DISORDERS • LEFT-HANDEDNESS • SPEECH • TEMPORAL LOBES

Suggested Reading

Bischoff-Grethe, A., Proper, S. M., Mao, H., Daniels, K. A., and Berns, G. S. (2000). Conscious and unconscious processing of nonverbal predictability in Wernicke's area. *J. Neurosci.* **20,** 1975–1981.

Blumstein, S. E. (1991). Phonological aspects of aphasia. In *Acquired Aphasia* (M. T. Sarno, Ed.), 2nd ed., pp. 151–180. Academic Press, San Diego.

Boatman, D., Lesser, R.P., and Gordon, B. (1995). Auditory speech processing in the left temporal lobe: An electrical interference study. *Brain Language* **51,** 269–290.

Breedin, S. D., and Saffran, E. M. (1999). Sentence processing in the face of semantic loss. *J. Exp. Psychol. Gen.* **128,** 547–562.

Cardebat, D., Demonet, J.-F., Celsis, P., Puel, M., Viallard, G., and Marc-Vergnes, J.-P. (1994). Right temporal compensatory mechanisms in a deep dysphasic patient: A case report with activation study by PET. *Neuropsychologia* **32,** 97–103.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., and Gagnon, D. A. (1997). Lexical access in aphasic and non-aphasic speakers. *Psychol. Rev.* **104,** 811–838.

Dronkers, N. F., Redfern, B. B., and Knight, R. T. (2000). The neural architecture of language disorders. In *The New Cognitive Neurosciences* (M. S. Gazzaniga, Ed.), pp. 949–958. MIT Press, Cambridge, MA.

Gannon, P. J., Holloway, R. L., Broadfield, D. C., and Braun, A. R. (1998). Asymmetry of chimpanzee planum temporale: Human-like pattern of Wernicke's brain language area homolog. *Science* **279,** 220–222.

Indefrey, P., and Levelt, W. J. M. (2000). The neural correlates of language production. In *The New Cognitive Neurosciences* (M. S., Gazzaniga, Ed.). MIT Press, Cambridge, MA.

Just, M. A., Carpenter, P. A., Keller, T. A., Eddy, W. F., and Thulborn, K. R. (1996). Brain activation modulated by sentence comprehension. *Science* **274,** 114–116.

Martin, N. (1996). Models of deep dysphasia. *Neurocase* **2,** 73–80.

Saffran, E. M., Coslett, H. B., and DeSalme, E. J. F. (2000). How is your B-A-B-Y? Dissociated oral and written production. *Neurocase* **6,** 193–204.

Saffran, E. M., Coslett, H. B., and Fitzpatrick-DeSalme, E. J. (2001). "How is your B-A-B-Y?" Dissociated oral and written production. *Neurocase*, in press.

Schwartz, M. F., Saffran, E. M., Bloch, D. E., and Dell, G. S. (1994). Disordered speech production in aphasic and normal speakers. *Brain Language* **47,** 52–88.

Simons, J. S., and Lambon Ralph, M. A. (1999). The auditory agnosias. *Neurocase* **5,** 379–405.

Zatorre, R. J., Meyer, E., Gjedde, A., and Evans, A. E. (1997). PET studies of phonetic processing of speech: Review, replication and reanalysis. *Cerebral Cortex* **6,** 21–30.

Working Memory

SERGIO DELLA SALA and ROBERT H. LOGIE
University of Aberdeen, Scotland

GLOSSARY

central executive The theoretical component of working memory involved in the control and regulation of the working memory system. It provides a range of executive functions, such as coordinating activity of the other components, focusing, switching and maintaining attention, activating representations from long-term memory, and manipulating information in cognitive tasks.

dual task An experimental paradigm in which volunteer human participants undertake the concurrent performance of two distinct tasks, such as recalling a sequence of digits while tracking a moving target or, in everyday life, walking while talking.

phonological loop The theoretical component of working memory intended to account for the human ability to retain verbal sequences in a range of cognitive tasks through the use of phonological codes and mental, phonologically based rehearsal.

visuo-spatial sketch pad The theoretical component of working memory intended to account for the human ability to retain over periods of a few seconds visual and geometric properties of objects or scenes from the immediate environment or drawn from past experience. A subcomponent is thought to account for retention of movement sequences.

working memory A theoretical model, derived from empirical research, comprising multiple specialized components of cognition that is intended to account for the human capacity to comprehend and mentally represent the immediate environment; to retain information about immediate past experience; to support the acquisition of new knowledge; to solve problems, and to formulate, relate, and act on current goals.

Working memory (WM) can be thought of as the desktop of the brain, a set of cognitive functions that allows humans to keep track of what they are thinking, what they are doing, or where they are moment to moment. It holds information long enough to make a decision, to acquire vocabulary, to mentally image the layout of one's home, to support creative thinking, or to remember what to do next.

I. WHAT IS WORKING MEMORY FOR?

The notion of WM has figured prominently in theories and explanations for human on-line cognition as well as offering a basis for understanding how humans accomplish a wide range of everyday tasks. Moreover, it constitutes a mental tool that aids survival; if an organism is to interact successfully with its environment, then clearly there must be some means to record, on a temporary basis, its own most recent actions and to have some means to plan actions for the immediate future. This faculty gives a significant advantage over organisms entirely driven by automatic responses to external stimuli that signal basic needs. Examples of how this mental tool is used in everyday interaction with the environment highlight its utility. Humans use WM every time they have to solve a problem such as which career path to follow or to perform a mental calculation, to recall where an object has just been laid down, or to repeat a strange foreign word that has just been heard.

From these examples, it is clear that the concept is much broader than the more traditional idea of short-term memory. Indeed, the concept of WM arose from the need to provide a theoretical account of a wide range of everyday tasks that require both temporary memory and on-line cognition. Until the early 1970s, the focus of research on this topic was on retention of random sequences of words, letters, or digits, with the everyday example of recalling telephone numbers while dialing. However, a mental faculty would not have evolved to deal with these relatively artificial tasks; therefore, WM is intended to encapsulate on-line cognition and temporary verbal and visuospatial memory in both laboratory tasks and everyday life.

II. THEORIES OF WORKING MEMORY

Most researchers on memory agree that WM is a useful concept that can account for a substantial body of experimental evidence with relatively few assumptions: It is a system with limited capacity, which loses information over periods of seconds, whose contents are subject to displacement by new input, and which interacts and is supported by other parts of the cognitive system such as stored knowledge or the products of perception. Most theories of WM would incorporate these features, however, it remains a developing concept whose detailed characteristics are the subject of an ongoing debate. One class of theories views WM as the sum total of what is currently activated from our long-term memory, and another class views WM as separate from long-term memory. However, even among these kinds of theories there is discussion about whether the cognitive functions are supported by a single flexible system that provides both temporary storage and on-line processing or whether they comprise multiple components, each specialized for a particular function. The first class of theories can account for a wide range of data from experimental psychology with healthy adults. However, they run into difficulty when we consider dissociations that are apparent from studies of neuropsychological patients. Such patients suffer from impairments of cognitive functions, but the impairments are often selective with, for example, very poor performance on immediate memory tasks coupled with normal performance on long-term retention. The converse dissociation also occurs, namely normal immediate memory contrasting with very poor retrieval from long-term memory or difficulties in long-term learning. Theories that assume temporary activation of long-term memory predict that if long-term memory access is impaired then temporary activation should also result in poor performance. Therefore, theories that can account for the neuropsychological dissociations as well as the data patterns from healthy adults would appear to be more useful.

One theory of WM that has been particularly successful during the past three decades was developed largely in Britain and inspired by the seminal work of Alan Baddeley and Graham Hitch. The theory offers an attractive level of simplicity coupled with success in accounting for many laboratory and everyday demands on moment to moment cognition. It has also helped the interpretation of a range of cognitive deficits following brain injury or brain disease.

III. EVIDENCE FOR A TWO-STORE MODEL OF WORKING MEMORY

The prevailing view in North America during the first half of the 20th century was of human memory as primarily the result of observable learning of associations, and theories of learning dominated. During the same period in Europe, memory as a cognitive concept was more salient, for example, in the work of Bartlett, Broadbent, and Zangwill. By the mid-1950s, this became the more generally accepted view, and a cognitive view of memory gave rise to major new avenues of research. Very quickly it became clear that memory might best be thought of as comprising more than one element, with one focusing on temporary or short-term retention and the other on long-term retention. This revived a dichotomy suggested in the 17th century by John Locke, who suggested that the "faculty of retention" comprised "contemplation" or "keeping an idea in mind" and "memory" or the "storehouse of these ideas." This distinction appeared again in the late 19th century when William James referred to primary memory, which he contrasted with retention. During the 1960s, the contrast between primary memory and what then became secondary memory was demonstrated experimentally. For example, when healthy adult volunteers are asked to retain sequences of words for immediate recall in any order, omissions tend to occur among words from the middle part of the list and much better recall is observed with items near the beginning or near the end of each list. This gives rise to a U-shaped recall pattern known as the serial position curve (Fig. 1). If, instead, the volunteers are asked to recall the items following a delay, performance on the last few items is no better

than it is for items in the middle of the list. However, initial items in the list retain their advantage despite the delay. The pattern for delayed recall is also shown in Fig. 1. The contrasting patterns shown in Fig. 1 have been thought to support the idea that two memory systems are reflected in different parts of the curve: The early list items are thought to be rehearsed and transferred into a longer term memory system, and this part of the curve is typically referred to as the "Primacy effect." As the list progresses, rehearsal becomes more difficult and therefore few items are transferred to a long-term store. The final items are thought to be held in a temporary memory system whose contents are lost unless recall is immediate. The advantage for the final items with immediate recall is known as the recency effect.

Further support for the dichotomy comes from amnesic patients—that is, people who, following brain damage, suffer from severe memory impairment that is characterized by an inability to learn new material. When such patients are tested for immediate free recall of word lists, their serial position curve lacks the advantage for the initial items in the list but retains a normal recency effect. Typical patterns for healthy adults and for amnesic patients are shown in Fig. 2.

The importance of rehearsal is emphasized in a different experimental paradigm whereby participants are given short sequences of letters for immediate recall or for recall after delays from 3 to 18 sec. These delays were either unfilled, and the participants were free to rehearse, or they were filled with verbal tasks such as counting backwards. Counting backwards was thought to prevent verbal rehearsal of the letters, and the number of letters recalled after each period allowed a plot of forgetting over time. After filled delays of

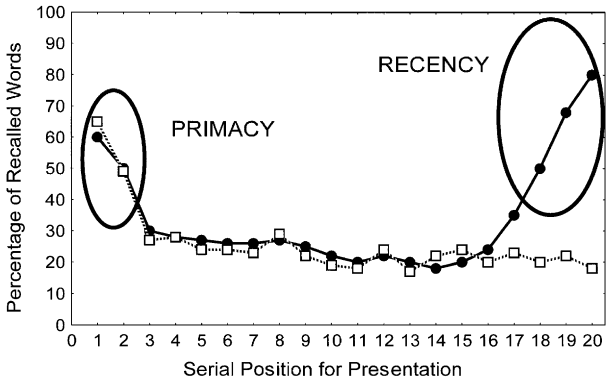


Figure 1 Idealized serial position curve for immediate (●) and delayed (□) free recall of lists of words in healthy participants.

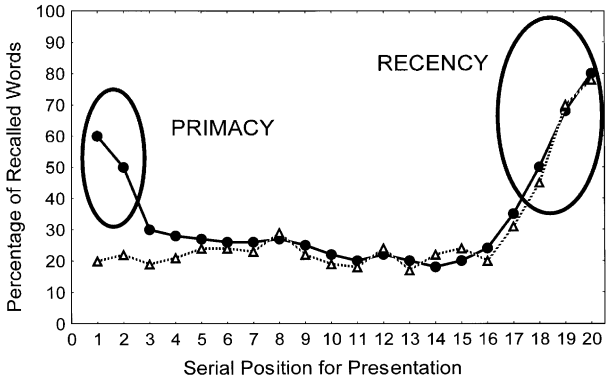


Figure 2 Idealized serial position curve for immediate free recall of lists of words in healthy participants (●) and amnesics (△).

18 sec, recall was extremely poor, indicating that the letters had been lost from the memory system in a relatively short period. This suggested that the passage of information from perceptual input to long-term storage was not automatic, and that information will be lost unless rehearsed.

A third paradigm demonstrated that different memory codes are used in immediate and delayed recall. Here, volunteers were given lists of words that were similar in meaning (e.g., "wide," "broad," "tall," "fat," "huge," "big," "great") or were similar in sound (e.g., "man," "mad," "mat," "map," "cap," "can," "cat") for immediate recall or for recall after a delay. Recall of both lists was compared with memory for lists of words that were neither semantically nor acoustically similar (e.g., "car," "lock," "spoon," "book," "chair," "pen," "dog"). With immediate recall, acoustically similar words were recalled more poorly than were words from the other lists, a phenomenon known as the acoustic similarity or the phonological similarity effect. With delayed recall, the semantically similar words led to poorest recall. Because acoustic similarity affected short-term but not long-term retention, whereas semantic similarity showed the converse, this led to the conclusion that acoustic codes were a signature of a temporary memory system and semantic codes were characteristic of the use of a longer term memory system.

IV. DEVELOPMENT OF THE WORKING MEMORY CONCEPT

The concept of primary and secondary memory provided a useful framework to account for the

accumulating body of experimental data in support of a dichotomous view of the memory system. However, it was limited by the fact that virtually all the research relied on experiments with presentation and recall of words, letters, or digits. It is clear that we can also retain information about the appearance of objects and scenes as well as their names and descriptions. Therefore, in the late 1960s, primary memory began to be seen as a system that could support temporary retention of different kinds of material as well as offering a means to process and manipulate that material. The dominant or ''modal'' concept then became that of a short-term or WM processing and storage system. This modal model retained the notion that information was transferred from the senses via WM to secondary or long-term memory. It also incorporated the assumption that the system has a maximum capacity that can be employed by storage, by processing, or by elements of both.

The idea that information from vision or hearing could reach long-term memory only through WM ran into difficulties when faced with contrasting memory disorders observed in patients with neurological damage. For example, one famous case is that of patient H.M., who, following surgery for the relief of epilepsy, suffered bilateral damage to the temporal lobes and the hippocampus. He showed a very severe deficit in learning anything new and had considerable difficulty in recalling details of events in his life. However, he had normal intelligence and had normal ability to repeat sequences of words that relied on temporary memory. His neuropsychological profile is in stark contrast with that of other patients, such as P.V., who, following a stroke in the Sylvian region of the left hemisphere, presented with a very severe deficit in retaining sequences of words or digits for immediate recall but had no difficulty in retrieving details of autobiographical events and proved well able to learn new material. Several other patients with contrasting patterns similar to those of H.M. and P.V. have been described in the neuropsychological literature. The pattern for patients such as P.V. in particular suggests that access to long-term memory does not depend on normal function in short-term verbal retention, either for new learning or for recall of past events. This undermines one of the key assumptions of the modal model.

A second assumption of the modal model was that of WM as a single, flexible, limited-capacity system. This theme remains in one contemporary strand of WM research that has focused on measures of individual differences in cognitive capacity. In a typical task, healthy adult volunteers are presented with a series of sentences. The requirement is to recall the last word of each sentence after all the sentences have been presented. The number of sentences is gradually increased until the participants can no longer accurately recall all of the final words, and the maximum number of words that can be recalled is taken as a measure of the individual processing and storage capacity. In various forms, the measure has been shown to correlate well with other measures of cognitive ability, such as reading comprehension or mental arithmetic performance. These correlations are then used to argue that WM is a mental capacity for processing and storage that can be applied to a wide range of both laboratory and everyday cognitive tasks. In this sense, the sentence ''span'' (or WM capacity) measure has been quite useful and provides the basis for a standardized test of cognitive ability. Set against its pragmatic value, the underlying theoretical assumptions are that there is a single flexible cognitive system supporting performance. One possible alternative view is that the task taps several aspects of cognition, and that this same range of cognitive functions are employed for other complex cognitive tasks such as reading comprehension. A further limitation is that the test provides only a measure of memory in the context of processing (retention of the final words while reading). In recent tests of the previously mentioned assumptions, a separate measure was taken of processing capacity by asking volunteers to verify each of the sentences for their semantic content without a memory requirement. Memory ability was measured independently by asking volunteers to recall lists of individual words. Participants were then asked to combine verification with recall of all the final words, with the task demands set at the maximum that had been determined for each task component separately. If WM were a single flexible, limited capacity system, we might expect performance to be very poor when these two demanding task elements were combined. Results showed that despite the high processing demand and high memory demand, performance on both elements of the task remained very similar to that obtained when the subjects performed the tasks independently.

These results point to the notion that temporary memory might be served by a different aspect of the cognitive system from that used to support on-line processing. This argument in turn suggests that WM might be best thought of as the orchestrated operation of multiple cognitive instruments, each with a specific role for temporary memory or for processing.

V. COGNITIVE ARCHITECTURE OF A MULTIPLE-COMPONENT WORKING MEMORY

The most influential model of multicomponent WM was proposed in 1974 by British scientists Alan Baddeley and Graham Hitch. The model included some elements of the modal model by incorporating processing and memory; however, these were viewed as distinct. Three components were originally proposed to provide respectively temporary verbal storage, temporary visuospatial storage, and a coordinating or "executive" function. A schematic diagram is shown in Fig. 3. The verbal storage component was originally named the articulatory loop, although subsequently it has been referred to as the phonological loop. The visuospatial component has been known as the visuospatial scratch pad, whereas the coordinating mechanism is generally referred to as the central executive.

The phonological loop comprises a passive phonological store and an articulatory rehearsal process. Auditorily presented verbal information gains direct access to the passive store, which retains information in a phonological form. Information in the store is subject both to decay over time and to interference from new verbal material. Loss of information from the store can be prevented by means of subvocal rehearsal, and with continuing rehearsal the contents of the store could in principle be retained indefinitely. The operation of the store is impaired when it is subject to irrelevant speech. This disruptive effect is even greater when the irrelevant speech is phonologically similar to the material already in the passive store. For example, if the store contains a sequence of digits, its content would be disrupted by listening to digit-sounding words such as "ton," "woo," "tee," "soar," and "thrive."

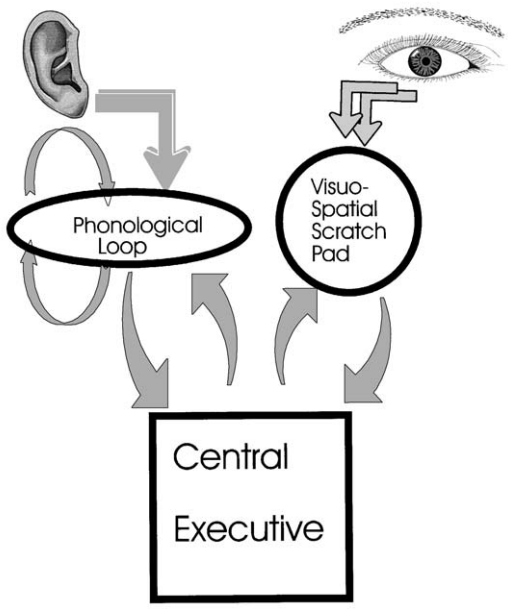


Figure 3 Schematic diagram of an early version of the multiple-component model of working memory.

The store also operates less efficiently when required to retain materials that are phonologically similar to one another compared with when the materials are phonologically distinct. These phonological confusions support the claim that the store is phonologically based. According to the model, subvocal rehearsal is also responsible for translating visually presented verbal material into a phonological code for temporary retention in the phonological store. The evidence for this derives from studies showing that the phonological similarity effect is removed for visually presented words and letters when subjects are required to suppress subvocal rehearsal by repeating aloud an irrelevant word such as "the-the-the." In other words, when subjects are unable to rehearse the visually presented items, they are not translated into a phonological code for retention in the store and may be retained using visual or semantic codes. Finally, the operation of the loop is limited by the length of time taken to pronounce the words for recall. That is, words such as "cricket," "wicket," "bishop," and "parrot" take less time to say than do words such as "Friday," "typhoon," "vendor," "practice." Likewise, words with five syllables, such as "refrigerator," "university," "hippopotamus," and "aluminium" take longer to say than do words of one syllable, such as "salt," "pen," "cat," and "chair." Based on these findings it has been argued that the capacity of the loop was limited by how much speech a subject could produce in approximately of sec. This word-length effect was thought to indicate the rate of decay of information from the phonological store.

The visuospatial scratch pad was thought to serve an analogous function for visual and spatial material and to serve a role in visual imagery tasks. The central executive component of the Baddeley and Hitch model provides the coordinating function for dual-task performance as detailed previously. It also appears to play a role in reasoning and problem solving and to be involved in the allocation of attention.

Although the link between WM and long-term memory has not been specified in detail, there is the general assumption that information can enter one or

more components of WM either from sensory input or from long-term memory. In its original conceptualization, WM was assumed to play an important role in transferring information to long-term memory from sensory input and in particular to be closely involved in learning novel information. For example, as noted previously, auditory input of verbal material is thought to have obligatory access to the phonological store, whereas visual material is thought to have obligatory access to the visuospatial scratch pad. The model accounts for normal long-term learning in patients with verbal short-term memory deficits by arguing that only the phonological loop system is impaired, allowing the central executive and the visuospatial scratch pad to effect learning. This preserves the idea that WM functions as a gateway between perception and long-term memory by means of additional assumptions that are supported by experimental evidence about the nature of the various WM components rather than about the demonstration of their existence. In the next section we address the issue as to whether WM acts as a gateway or more as a mental work space. The following section will then return to the precise characteristics of the WM components and the evidence from which those characteristics have been derived.

VI. WORKING MEMORY AS A WORK SPACE, NOT A GATEWAY

Short-term memory or WM is often thought to function as a form of temporary holding device following perception and to be closely involved in transfer of information from perception to long-term memory. This was a feature of the modal model and also an implicit assumption in the original version of WM as multiple components. It is also an assertion in many contemporary introductory textbooks that discuss the psychology of memory. However, during the past few years it has been increasingly clear that this view is unable to account for new empirical evidence and patterns of impairment and sparing in brain-damaged patients. For example, in normal reading, it is possible for healthy adults to generate the sound characteristics or the phonology of words without having to speak the words aloud. It is possible to have an auditory mental experience of the word from the print on the page. The sounds of the words can be mentally rehearsed and recalled orally or in writing. There is ample evidence that immediate memory for visually presented words relies heavily on the use of mentally rehearsing the phonological properties of those words. This could not be possible if the words being read did not access phonological knowledge in long-term memory before they were rehearsed in WM. The point is also clearly illustrated in studies of patients with acquired anarthria, a syndrome that is characterized by an inability to produce overt speech following damage to the brain stem. However, such patients can clearly access stored knowledge about the phonological properties of words presented visually. As with healthy adults, if the words are being read, the information about their sound has to be extracted from long-term stored knowledge.

Problems with the gateway assumption are not restricted to verbal immediate memory tasks; they also appear with visuospatial WM tasks, particularly in the neuropsychological disorder known as neglect. These patients ignore half of their visual field, of objects and of their body space, most commonly on the left. Thus, when asked to copy a picture they will tend to draw only its right-hand side. When they are asked to describe the scene in front of them they will omit objects on their left. However, although such patients appear to ignore information on their left side, they are nonetheless affected by that information. In one classic experiment, neglect patients were shown pictures of two houses. In one of the pictures, the left half of the house was shown on fire. The patients denied that there was any difference between this picture and a picture of an identical house that was not on fire. That is, they ignored the left half of the picture. However, when asked which house they would prefer to live in, they chose the one that was not on fire.

One well-argued interpretation for these data is that although the sensory information did not appear to be available in WM, it was nevertheless activating semantic knowledge about fires and their unpleasant consequences. Therefore, as with normal subjects, it appears that information can gain direct access to information in long-term memory without going through WM first. Moreover, the results of that access to long-term semantic knowledge, in neglect patients at least, may feed into decisions without the individuals being aware of why they are making a particular decision. One possibility is that the activated information about unpleasant consequences becomes available to WM, even though the visual information from sensory input does not.

Neglect not only impair perception but also can affect mental representations. For example, when some left-sided neglect patients are asked to describe

a familiar scene from memory, such as a town square, they tend to leave out details of the square that would be on their imagined right in the mental image but to report several details on their imagined left. This difficulty does not result from simple forgetting of these details, because if they are asked to imagine standing at the opposite end of the square they can readily report details, now on their imagined right, that they had previously omitted. They also leave out details from what is now their imagined left but which had previously been reported.

Neglect is not a unitary deficit and it appears in several forms. Many patients appear to have difficulties with perception and some of these patients are also reported to have representational problems. However, there are a small number of cases that appear to have a pure form of representational neglect. Such patients have no difficulty describing the immediate scene. However, they omit details from the one side of scenes that are mentally imaged from memory. Moreover, representational neglect occurs even when perception of a scene has been normal. In one study, neglect patients were shown an array of objects and could describe the objects and their locations perfectly. However, if the objects were removed and the patients had to describe the objects and positions from memory they failed to report items that had been shown on the left but were just as good as healthy volunteers in recalling items that had been shown on the right. WM as a gateway between perception and long-term memory cannot account for this pattern of observations since perception was perfectly normal and the objects were correctly identified, showing normal access to information in long-term memory about the objects. However, the information was clearly not available in WM. The pattern is more readily explained if WM is thought of as a work space whose contents are derived from activated knowledge in long-term memory.

The patient data yield insight into functioning of visuospatial WM in the normal brain. Perceiving and interacting with the environment demands the generation of mental representations of that environment. Such representations would incorporate the location of various objects in relation to one another and the location of objects in relation to the observer. Thus, when retrieving information about objects or scenes held in long-term memory, the mental representation contains information about the scenes, such as when they were previously experienced and the names of objects depicted. In this sense, the contents of WM necessarily comprise some of the products of perception and subsequent mental processing, including the semantic properties. WM can then be considered as a multicomponent work space that stores and manipulates information that has been activated from the long-term store rather than as a primary route into long-term memory from the environment. Therefore, WM is better viewed as a system accessed by information stored in long-term memory and by sensory input only *after* that input has been processed by long-term memory.

The model as a work space has the advantage that it specifies something of the way in which visually presented words can be made available to the phonological loop; a feature that was underspecified in the original WM model. For example, if the word ''book'' is heard, the phonological representation of book and the articulatory codes can be used to store a phonological trace in the phonological loop, to rehearse the word, and to articulate it. On hearing the word, the visual and spatial properties of the word book become available to the visuospatial scratch pad, allowing generation of a temporary representation of the visual form of the word that could then be written. Semantic information about books, such as what they are used for, could also be activated. This information would then be available to the central executive, allowing generation of an image of what a book looks like. The central executive could also draw on information stored in the phonological loop and in the visuospatial system to carry out further processing on the properties of the visual or phonological form of the word to allow spelling, segmentation, rotation of an image of the word, and so on.

This conception of WM fits neatly with the notion that people are strategic in how they use their cognitive functions. For example, in recalling a list of words, this can be accomplished by rehearsing the sounds of the words, thinking about their meaning, organizing them into groups, or imagining the objects or concepts to which they refer. As a result, phenomena such as the word length effect only appear if the participants rely on mentally rehearsing the words prior to recall. There is evidence that some participants in experiments spontaneously use other strategies that might rely on visual or semantic codes, and for them the word length effect does not appear. The same argument holds for other well-established phenomena such as the phonological similarity effect. Such strategic use of different aspects of WM would not arise if the material did not activate knowledge stored in long-term memory prior to it being held in WM for the period until recall is required.

VII. FRACTIONATING THE MODEL

The different components of the WM model have been further fractionated and we now describe some of the evidence and rationale for this development. A diagram of the revised version of the model is depicted in Fig. 4.

A. Verbal Storage and Rehearsal

The three-component model provided a useful framework in that it could account for a wide range of laboratory findings with relatively few assumptions. However, it has proved to be rather simplistic when employed to account for details of experimental findings from healthy adults and from patients or as the basis for deriving formal computational models that could simulate human data patterns. For example, the phonological similarity effect and the word-length effect cannot wholly be explained by a single system, and the phonological loop evolved into a two-component system comprising a phonological store, which was subject to decay and that was responsible for the phonological similarity effect, and articulatory rehearsal, which was used to help account for the word-length effect. One source of evidence for this dissociation comes from using a manipulation known as articulatory suppression. This involves asking the volunteer to repeat a word over and over such as "go go go go" while attempting to retain a sequence of words, letters, or digits. The technique is thought to prevent the operation of subvocal articulatory rehearsal, and it typically results in very poor immediate recall performance. In addition, it tends to remove the effect of word length when the words are heard. However, phonological similarity is unaffected by auditory presentation of the items for recall. The model neatly accounts for this contrast by suggesting that because articulatory suppression prevents subvocal rehearsal, this prevents the use of the system that is responsible for the word-length effect. However, this form of suppression has much less impact on phonological similarity because this effect arises from the use of a different part of the system, namely the phonological store.

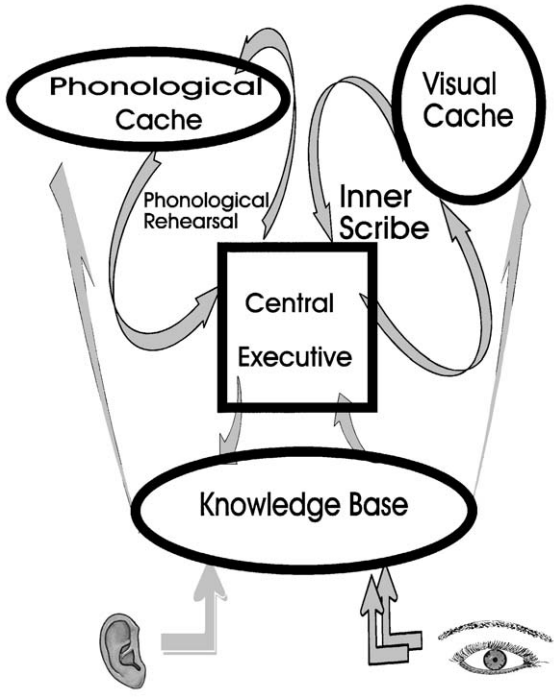


Figure 4 Schematic diagram of a recent version of the multiple-component model of working memory.

A similar argument can be used to account for a pattern of data obtained from anarthric patients whose deficits result from cortical lesions. Such patients show normal phonological similarity effects with auditory presentation but fail to show a word-length effect that they hear. The model explains this by suggesting that the subvocal rehearsal system is damaged, whereas the phonological store is intact. The converse pattern has been observed in patients with severe short-term verbal memory impairments. These patients fail to show phonological similarity effects but do show effects of word length when they are instructed to rehearse the items. They appear to have damage to the phonological store but retain the ability to use articulatory rehearsal. These contrasting patterns point to a dissociation that supports the notion of the two components of the phonological loop.

A final source of evidence is from studies using brain imaging techniques, such as positron emission tomography (PET) and functional magnetic resonance imaging (fMRI). These techniques identify the areas of the brain that are active during the performance of particular tasks. No task will involve only a single area of the brain because aspects of brain function operate in concert. However, current evidence indicates that the storage and rehearsal aspects of the phonological loop could be mapped onto relatively clearly segregated neuroanatomical networks, both making use of functions associated with activation of the left hemisphere. In particular, while storage and rehearsal result in activation of areas of the frontal lobes, rehearsal appears to also involve activation of Broca's area, normally linked to the production of speech, whereas phonological storage involves activation of the supramarginal gyrus. These areas correspond on the whole

with those areas found to be damaged in patients with vocal rehearsal problems or deficits in temporary phonological storage, respectively.

B. Visual and Spatial Storage and Processing

Recent evidence regarding storage and processing of visuospatial material also appears to require further fractionation of the original model. From our own and other laboratories, it is clear that retention of sequences of movements appears to rely on different cognitive functions from those employed in temporary retention of static patterns and visual appearance. One task commonly used to access memory for movement sequences involves an array of randomly arranged wooden blocks on a board. The experimenter taps a sequence of blocks, after which the participant is required to reproduce the sequence. The length of the sequences gradually increases from trial to trial until the participant can no longer recall the sequence accurately. This then allows a measure of individual differences in capacity for immediate recall of movement sequences.

Tests for immediate recall of visual patterns typically involve presenting checker board arrays of black-and-white squares arranged randomly. Participants are then given a blank array of squares and are required to indicate which squares were previously shown in black. As before, the number of squares in the array increases from trial to trial until participants reliably fail to recall all the relevant squares, thereby giving a capacity or span measure for the task.

The previously mentioned tasks have been contrasted in children, in adults, and in adults following brain damage. When the same individuals perform both tasks, the correlations between the span scores are low both for children and for healthy adults. In children, performance on the visual pattern task develops much more rapidly with age than does the movement sequence or block tapping task, indicating that the cognitive functions on which each task relies follow different developmental trajectories. The distinct nature of the two tasks is further emphasized by the observation of double dissociations among brain-damaged patients (strokes or dementia); specifically, some patients show deficits on one task but do well on the other, whereas other patients show the converse. This indicates that the cognitive functions required for each task do not completely overlap.

Further evidence for the separation of visual and spatial cognition comes from dual-task interference studies in healthy adults. In particular, performance on the block tapping task is selectively disrupted if, during retention interval, the participant is required to tap his or her hand on the table top in a different regular pattern (e.g., tapping four buttons arranged in a square). The visual matrix pattern task is not disrupted by concurrent tapping in a square. In contrast, the visual task is selectively disrupted by presentation during a retention interval of irrelevant pictures or patterns, whereas the block-tapping memory task is not. The separation between visual and spatial components is strengthened by evidence from neuroimaging studies showing neuroanatomical segregation in the brain areas involved in either task. Indeed, evidence of independence between visual and spatial neuroanatomical substrates has been reported from studies of nonhuman primates.

The component thought to be responsible for temporary retention of visual information has been referred to as the "visual cache," whereas the component linked with retention of movements has been referred to as the "inner scribe."

C. Executive Functions

The original view of the central executive was of a conceptual Pandora's box, charged with a wide range of complex cognitive functions including reasoning, problem solving, decision making, allocating attention, and coordinating the functioning of the storage and processing systems described previously. One danger with this approach is that it appears to require the notion of a conceptual "fat controller" or homunculus responsible for implementing these executive functions, and we are left with the question as to what is inside the head of the fat controller. One way out of this apparent impasse is to consider how a cognitive system might deliver each of the executive functions, one possibility being that the executive can be fractionated into several specialized components and another being that the executive is a flexible processor that can implement a range of acquired skills and expertise. Before discussing these options, we first consider evidence that there is a need for the concept of WM to include an executive system.

One putative-specific function for the executive is that of coordinating the operation of the phonological loop and the visuospatial system. A number of studies have shown that normal healthy adults can hold verbal sequences such as digits while performing visuospatial tasks, such as following a moving target around a screen or holding visual matrix patterns. This dual-task condition results in little or no decrease in the

performance of either task, even when both tasks are very demanding or when the participants are drawn from the healthy elderly. These data further support the notion that visuospatial WM and verbal WM are served by separate cognitive functions that can operate in parallel. However, the pattern was different in groups of patients suffering from the early stages of Alzheimer's disease. The patients were first tested on each of the tasks individually and the demands of each task were set at the capacity limits for each patient, as done for the healthy subjects in the experiments described previously. In this way, performance on each task was equated between the healthy subjects and the patients. However, when the patients were asked to perform the two tasks together, performance on both was severely disrupted, quite unlike the case for their healthy counterparts. In other studies, patients with early Alzheimer's disease also show difficulty in dealing with information from more than one source, such as in a conversation among several people, although they can adequately cope when conversing with a single individual. These patterns of data have been interpreted to suggest that the patients have a deficit in a coordination function that allows normal healthy subjects to cope with more than one task or source of information almost seamlessly, providing different components of their cognition system are required for each of the tasks involved.

The dual-task impairment found with Alzheimer's patients does not arise simply from the general difficulty of performing two tasks rather than one against the background of a damaged brain. When the individual tasks are gradually increased in difficulty, the relationship between task demand and performance is very similar to that found for healthy elderly volunteers. Therefore, it appears to be a specific deficit in performing two different tasks simultaneously.

Complementary evidence comes from brain imaging studies of healthy adults performing single tasks that recruit either the visuospatial or the verbal component of WM and then require the two tasks to be performed simultaneously. When a single task was increased in difficulty, more of the same brain areas were activated. However, when the two tasks were performed together, both brain networks were involved; in addition, there was activity in the dorsolateral regions of the prefrontal cortex. Consistent with the experimental data, Alzheimer's patients show a significant neurochemical deficit in this same cortical area, and dual-task coordination deficits have also been shown in some stroke patients who have suffered focal damage to the frontal lobes.

The evidence from dual-task performance indicates that there may indeed be a coordination function as suggested for the concept of a central executive. Among the other functions attributed to executive operation, the maintenance and switching of attention appears to also be specifically impaired in early Alzheimer's disease, independently of temporary visuospatial or verbal memory. This again points to the need for a theoretical concept to account for allocation of attention, although any possible links between a coordination function and the control of attention have yet to be explored in detail.

It would be possible to take a similar empirical approach to executive functions such as reasoning, problem solving, decision making or planning, and activation of representations in long-term memory. There is a growing literature on studies of patients with frontal lobe damage, indicating impairments in some of these areas of cognition. However, the link between the range of executive functions and the operation of the frontal lobes is far from clear. For example, the intercorrelations between tests of these functions are far from dramatic, and few of the tests have been subject to adequate task analysis. Such analysis would be essential to establish that each test measures the intended cognitive function. Therefore, it is important to separate conceptually the cognitive frame of reference from its neuroanatomical substrate.

We can avoid some of the problems associated with the homunculus by considering the executive as a flexible mechanism that can serve as the medium for implementing learned strategies, acquiring skills, and accessing stored knowledge relevant to the task at hand. In this view, the executive functions as the central processor of a computer that can run different kinds of software. Without pressing this analogy too much, we then can use the executive to implement the process of language comprehension during reading or listening to conversations, the acquired procedures for mental arithmetic when undertaking calculation, the strategies for mental imagery when holding or manipulating mental representations of objects or scenes, strategies for planning, or the heuristics for switching of attention between tasks. This idea of the executive is compatible with several of the alternative views of WM. However, the notion is very much of an executive that is a distinct feature of the cognitive architecture and not simply the sum total of what is currently activated from long-term memory or the long-term knowledge base. The concept of a flexible executive in this sense would also incorporate the use of the specialist temporary memory systems, such as the

phonological loop or the visuospatial resources, to hold material recently processed or to retain partial solutions while the executive acts on other aspects of a problem.

VIII. APPLICATIONS OF THE WORKING MEMORY CONCEPT

The idea of a multiple component has proved useful in understanding some of the cognitive deficits suffered by individuals with brain damage as well as in accounting for important aspects of healthy cognition, such as temporary memory, acquisition of vocabulary, mental arithmetic, the use of imagery in creative thinking, or remembering to do things. We previously discussed some of the insights into cognitive deficits that have arisen from the use of the model and how studies of patients with such deficits also give possible insight into the organization of normal cognition. The concept of the phonological loop also appears to offer an account of how young children acquire vocabulary. One feature of the loop is that it offers a means to retain a recently heard word—to mentally rehearse the word and then generate the word as articulate speech. A number of studies have examined the ability of young children to repeat aloud new or nonsense words that they have just heard. Their ability to do so correlates well with their general competence in language and predicts their language competence several years later. Patients with severe verbal short-term memory deficits also appear to have difficulty acquiring new foreign vocabulary. These findings led to the suggestion that the phonological loop is a cognitive function that has evolved to support the acquisition of language, and that an emergent property of such as system is a temporary verbal memory system that can be used in a wide variety of tasks.

Among the acquired skills that appear to depend on the phonological loop are counting and mental arithmetic. Keeping track of a running total by counting relies on subvocal rehearsal of the counting sequence, and keeping track of partial solutions in mental arithmetic again appears to rely on subvocalization or the subtotals involved, unless the solution is already well learned. For example, for most people the answer to $6 + 5 + ?$ would be available without the need for mental calculation because it would be recalled as an arithmetic fact. However, the answer to $13 + 18 + 25 + 17 = ?$ would require a series of operations with temporary memory for the subtotals (held in the phonological loop) as each part of the sum is completed. Evidence for these conclusions arises from studies in which volunteers perform counting or mental arithmetic while suppressing articulation through repeating an irrelevant word. Accuracy in counting and in arithmetic is disrupted by this procedure but is not disrupted by performing other irrelevant tasks, such as tapping the table in a regular pattern, or even when presented with irrelevant speech sounds. The phonological loop seems to support accuracy in counting and arithmetic, but when asked simply to make estimates of the final totals while suppressing articulation, the errors tend to be quite close to the correct response. This indicates that generating an estimate may rely on some other aspect of cognition. Additional insight into how this is achieved comes from an additional manipulation in which volunteers are required to generate random sequences of letters while simultaneoulsy adding numbers. People find it almost impossible even to generate estimates under these conditions. This result indicates that generating random letter sequences not only uses up the phonological loop but also requires the operation of the central executive to inhibit the production of well-learned sequences of letters such as "a-b-c-d-e…" and to keep track of which letters of letter sequences they have already generated. This points to the use of the executive for estimating totals in arithmetic. It also supports the idea that the executive might be a flexible system that can implement different cognitive operations according to tasks demands but cannot implement two different operations, such as random generation and arithmetic estimation, at the same time.

Among the other characteristics of WM, mental manipulation and the development of new knowledge are the focus of a developing research area that is sometimes described as mental discovery. The tasks that are employed in this research are ideally suited to illustrate the utility of WM as a bailiwick of specialized functions rather than a general-purpose memory and processing device.

This is best illustrated by attempting a simple mental discovery task. For example, by imagining the capital letter "D" and the capital letter "J," it is possible without the use of any external aids, such as pencil and paper, to mentally construct a recognizable object by rotating the D leftward 90° and putting it on top of the J. The result would be an umbrella (Fig. 5). Similarly, it is possible to construct the shape of an ice cream cone by combining the shapes "V" and "O," and more complex figures can be constructed from combinations of larger numbers of shapes. It is not immediately clear that the outcome of these combinations would be known or predicted from the letters or shapes, and therefore this constitutes a form of mental discovery

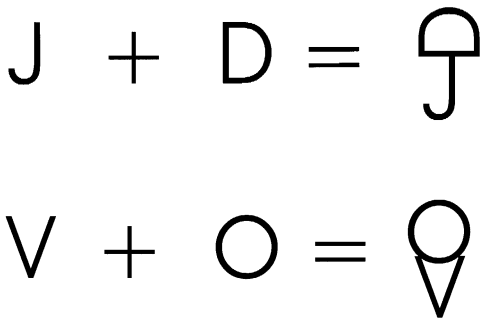


Figure 5 Examples of mental discovery tasks using mental imagery.

using visuospatial WM. One account of the task is that the partial combinations of shapes are held in a temporary visual store, such as the visual cache mentioned previously; the names of the shapes may be rehearsed within the phonological loop; and the manipulation, transformation, and combination would be achieved by the strategies and heuristics of mental imagery run by the central executive.

The visuospatial resources of the WM can support a wide range of other everyday cognitive tasks, such as remembering where we are on the page while reading, maintaining and updating a representation of objects in our immediate environment, mental scanning to plan a route around a familiar town, or imagining the configuration of furniture in an empty room.

Remembering to carry out tasks that we had planned or intended is generally referred to as prospective memory, and it also appears to rely to some extent on WM. To remember to send a letter on the way to work, to switch off the oven when the meal is ready, or to pack the slides and overheads when leaving for a conference requires that we have formed an intention and that we act on that intention at the appropriate time. For some activities, the time of day triggers the intention; for others, it may be a physical cue such as the mailbox. Other activities, such as remembering to call someone later in the day, are triggered by internal cues and may require that the intention is held in WM or written down in order to ensure that it is carried out. Recent evidence from our laboratory suggests that if WM is occupied by a demanding arithmetic task, then prospective memory failures are more frequent.

IX. CONCLUSION

Working memory is an evolving concept that may be subject to change in light of new evidence or questions as to its detailed characteristics. Nevertheless, it allows the interpretation of a great deal of evidence and phenomena from both laboratory tasks and everyday experience.

See Also the Following Articles

COGNITIVE AGING • MEMORY DISORDERS, ORGANIC • MEMORY, EXPLICIT AND IMPLICIT • MEMORY NEUROBIOLOGY • MEMORY, NEUROIMAGING • MENTAL WORKLOAD • NERVE CELLS AND MEMORY • PRIMING • SEMANTIC MEMORY • SHORT-TERM MEMORY • TIME PASSAGE, NEURAL SUBSTRATES

Suggested Reading

Baddeley, A. D. (1997). *Human Memory: Theory and Practice*. Rev. ed., Psychology Press, Hove, UK.

Baddeley, A. D., and Della Sala, S. (1996). Working memory and executive control. *Philos. Trans. R. Soc. London B* **351,** 1397–1404.

Becker, J. T., MacAndrew, D. K., and Fiez, J. A. (1999). A comment on the functional localization of the phonological storage subsystem of working memory. *Brain Cognition* **41,** 27–38.

Beschin, N., Cocchini, G., Della Sala, S., and Logie, R. H. (1997). What the eyes perceive, the brain ignores: A case of pure unilateral representational neglect. *Cortex* **33,** 3–26.

Della Sala, S., and Logie, R. H. (1993). When working memory does not work. The role of working memory in neuropsychology. In *Handbook of Neuropsychology Vol. 8* (F. Boller, and H. Spinnler, Eds.), pp. 1–63. Elsevier, Amsterdam.

Della Sala, S., Gray, C., Baddeley, A., Allamano, N., and Wilson, L. (1999). Pattern span: A tool for unwelding visuo-spatial memory. *Neuropsychologia* **37,** 1189–1199.

Goldman-Rakic, P. S. (1996). The prefrontal landscape: Implications of functional architecture for understanding human mentation and the central executive. *Philos. Trans. R. Soc. London B* **351,** 1445–1453.

Helstrup, T., and Logie, R. H. (Eds.) (1999). *Working Memory and Mental Discovery.* (Special issue of *Eur. J. Cognitive Psychol.*) Psychology Press, Hove, UK.

Hitch, G. J., and Logie, R. H. (Eds.) (1996). *Working Memory* (Special issue of *Q. J. Exp. Psychol.*) Psychology Press, Hove, UK.

Just, M. A., Carpenter, P. A., and Keller, T. A. (1996). The capacity theory of comprehension: New frontiers of evidence and arguments. *Psychol. Rev.* **103,** 773–780.

Logie, R. H. (1995). *Visuo-Spatial Working Memory*. Erlbaum, Hove, UK.

Miyake, A., and Shah, P. (Eds.) (1999). *Models of Working Memory*. Cambridge Univ. Press, New York.

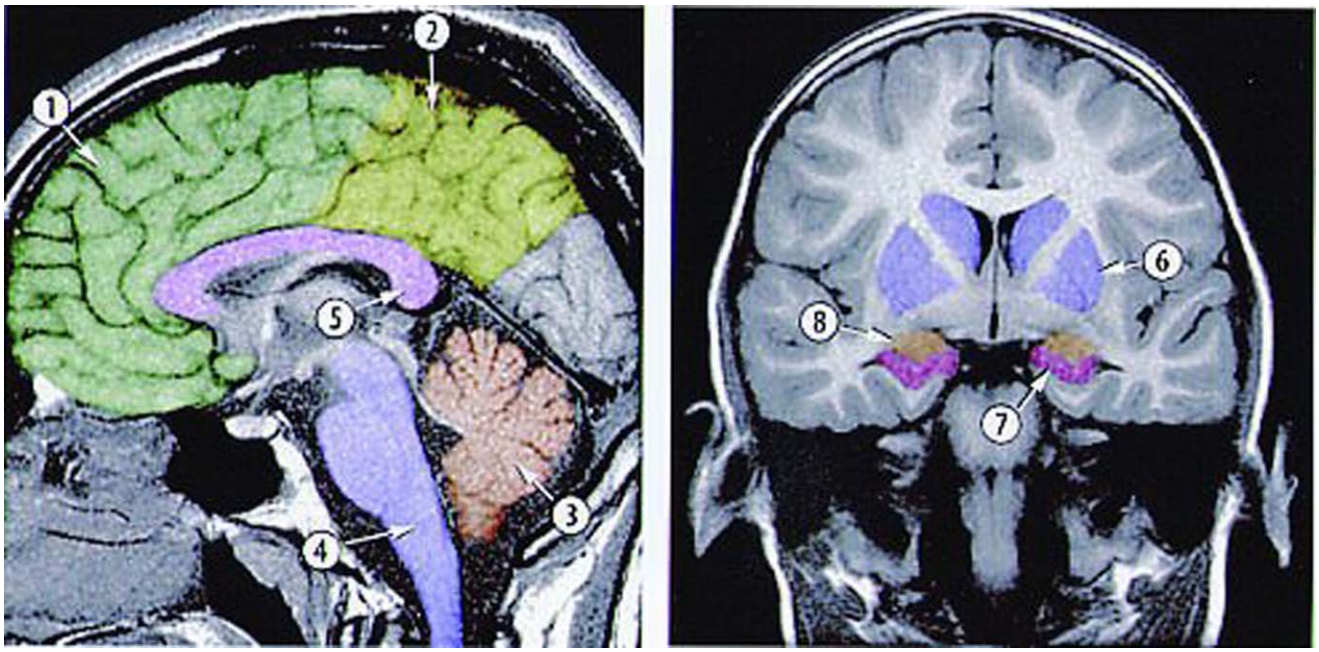
Richardson, J. T. E., Engle, R. W., Hasher, L., Logie, R. H., Stoltzfus, E. R., and Zacks, R. T. (1996). *Working Memory and Human Cognition*. Oxford Univ. Press, New York.

Smith, E. E., and Jonides, J. (1997). Working memory: A view from neuroimaging. *Cognitive Psychol.* **33,** 5–42.

Vallar, G., and Shallice, T. (Eds.) (1990). *Neuropsychological Impairments of Short-Term Memory*. Cambridge Univ. Press, Cambridge, UK.

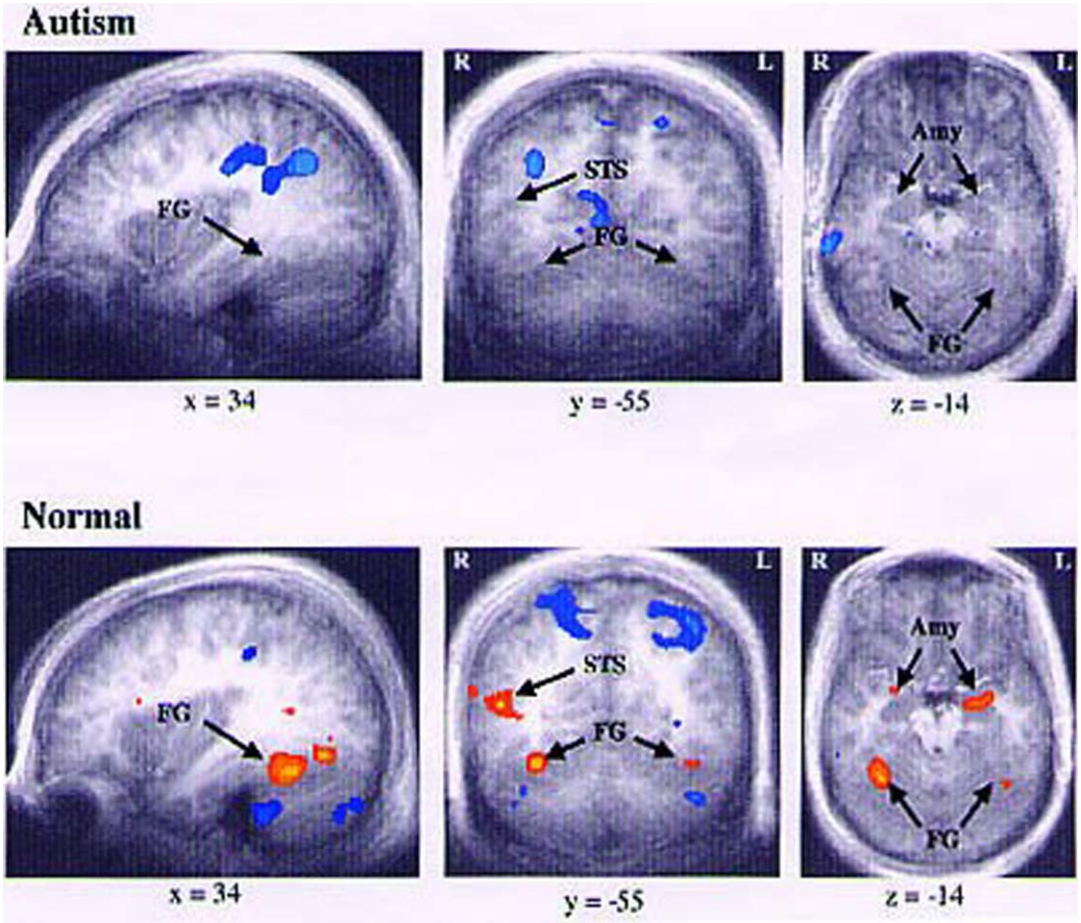
]>

Plate 1



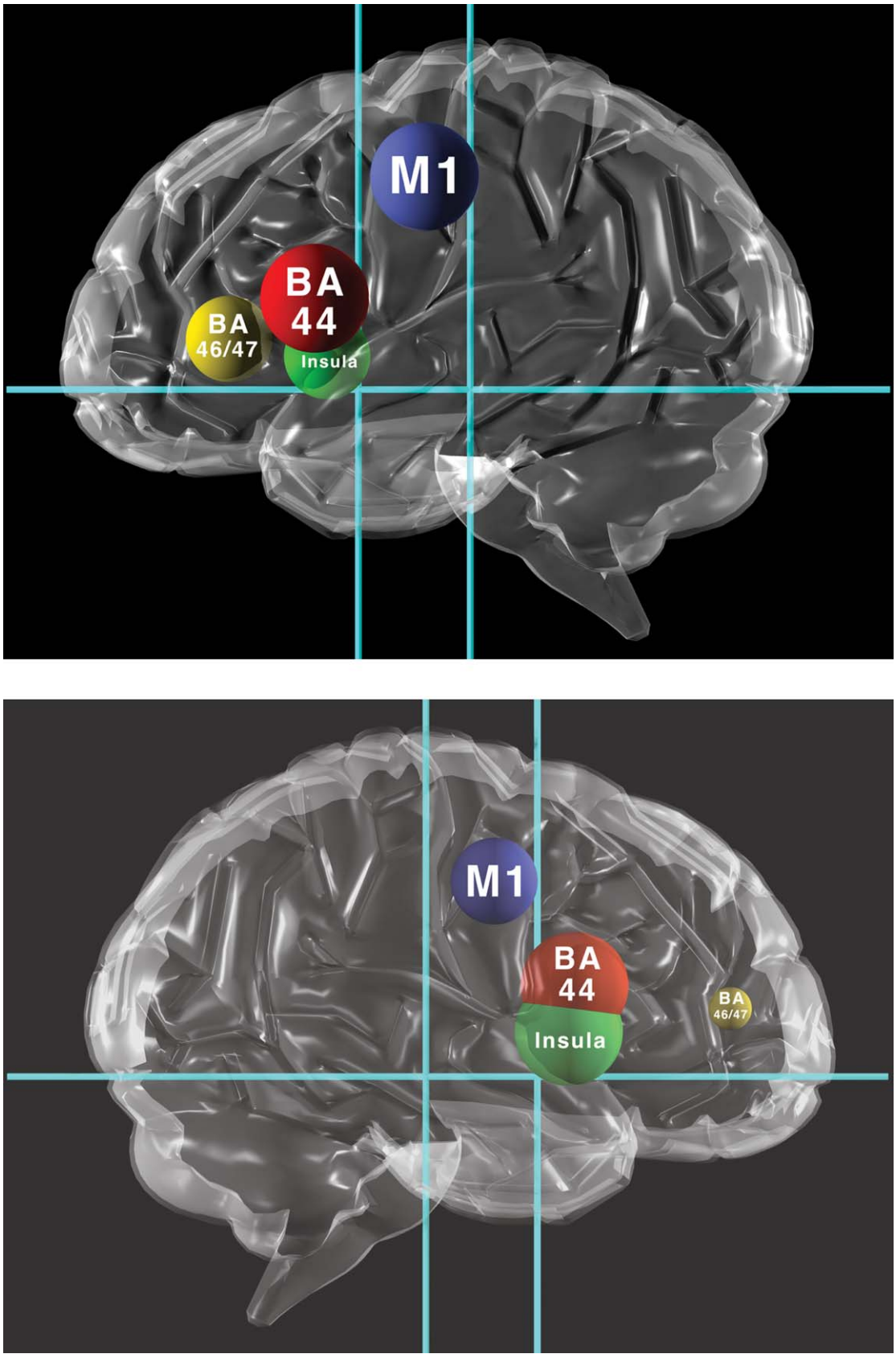
Depiction of cortical and subcortical structures reported as abnormal in autism via autopsy or MRI data from laboratories throughout the United States and Europe. 1, frontal lobes; 2, parietal lobes; 3, cerebellum; 4, brain stem; 5, corpus callosum; 6, basal ganglia; 7, amygdala; 8, hippocampus. See article AUTISM.

Plate 2



Within group t-maps for both autism and normal groups showing significant regions of activation (statistically significant positive activation noted by yellows and orange, deactivation noted by blues). Note fusiform, superior temporal sulcus and amygdala activation in normals, in comparison to a lack of positive activation in the autism group. Decreased functional activity in the autism group is likely due to the inconsistent patterns of activation noted across individual autism subjects, which would fail to beseen when results are averaged (from Pierce et al., 2001). See article AUTISM.

Plate 3

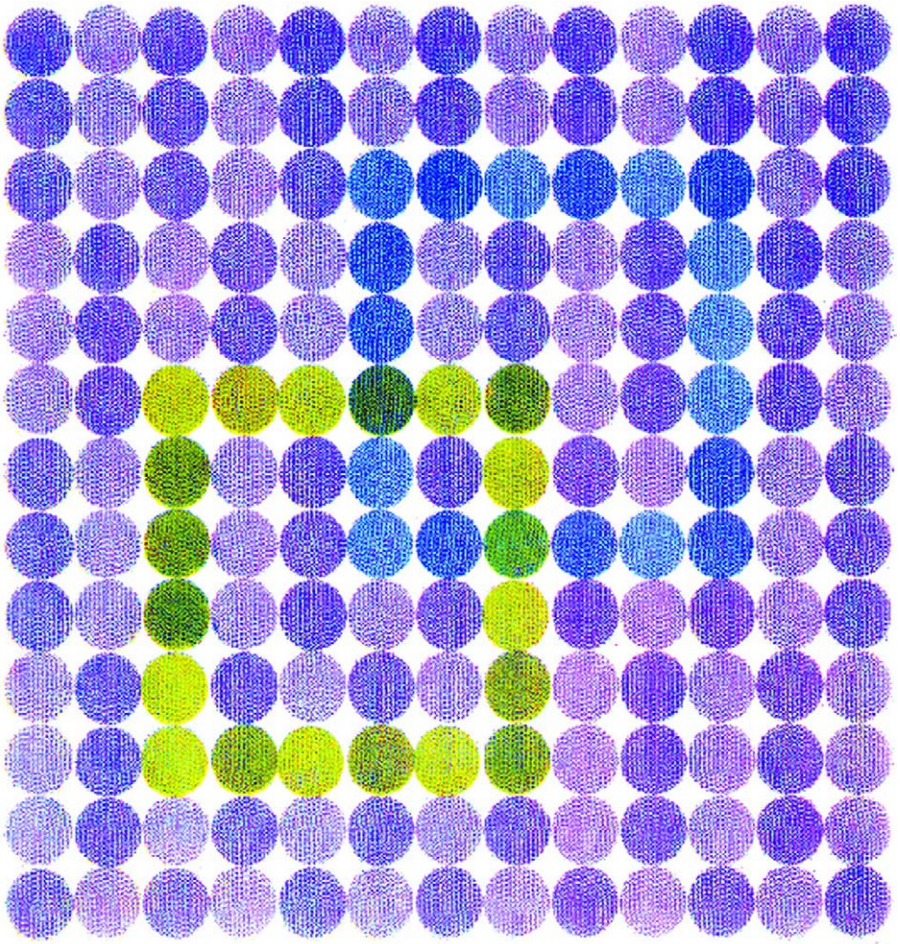


Three-dimensional representations of the three subdivisions of Broca's area and the motor area for the mouth are shown from the left (top) and right (bottom). Broca's area proper (BA 44/6) is shown in red. Sub-Broca's area (anterior insula) is shown in green. Pre-Broca's area (BA 46/47) is shown in yellow. The motor area for the mouth, shown in blue, is laterlized to the left hemisphere and shows activation on that side comparable to that of Broca's area proper. See article BROCA'S AREA.

Plate 4

Segment of a painting by Jan van Eyck (1434). Left, the luminace information has been removed, rendering an image based soley on the chromatic contrasts. Right, the chromatic contrast has been removed, rendering an achormatic black-and-white image of the same scene. Center, the fusion of the two different forms of contrast produces color vision.

Plate 5

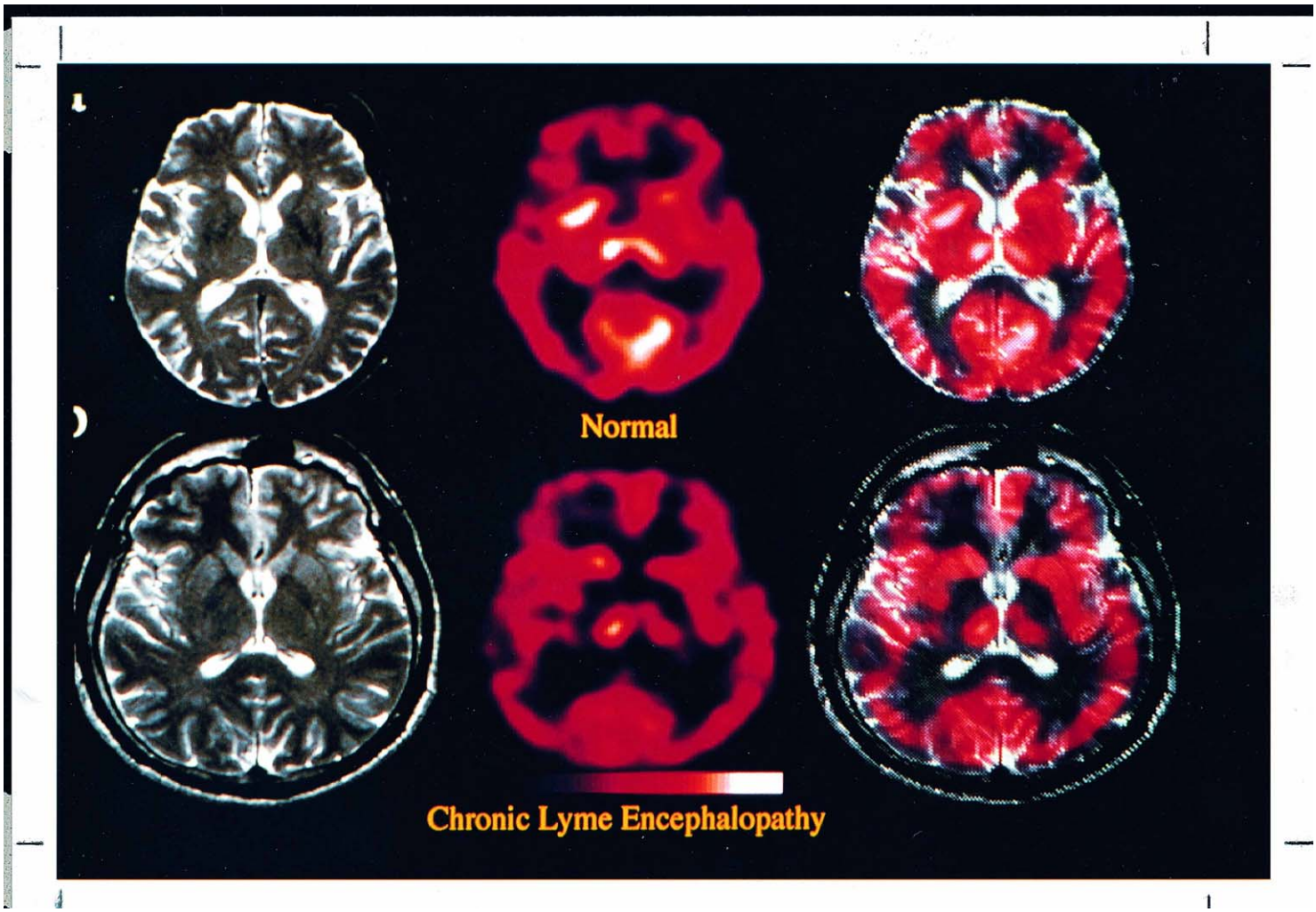


A pseudoisochromatic plate that detects protanopes, deuteranopes, and tritanopes, who fail to see the blue square at the upper right side. See article COLOR PROCESSING.

Plate 6

Some of the human brain areas important for emotion. (Left) medial view of left hemisphere showing orbitofrontal cortex (red) and, deep within the temporal lobe, amygdala (red), as well as hypothalamus and periaqueductal gray (both orange). Right, lateral view of the right hemisphere showing orbitofrontal cortex (red) and, buried in temporal cortex, amygdala (red), and also somatosensory cortex (blue) and, buried under overlying cortex, insula (blue). The amygdala and orbitofrontal cortex receive perceptual information and project to effector structures (principally, the hypothalamus and peraqueductal gray), which effect automatic changes in body state, including blood pressure, heart rate, and so on. See article EMOTION.

Plate 7

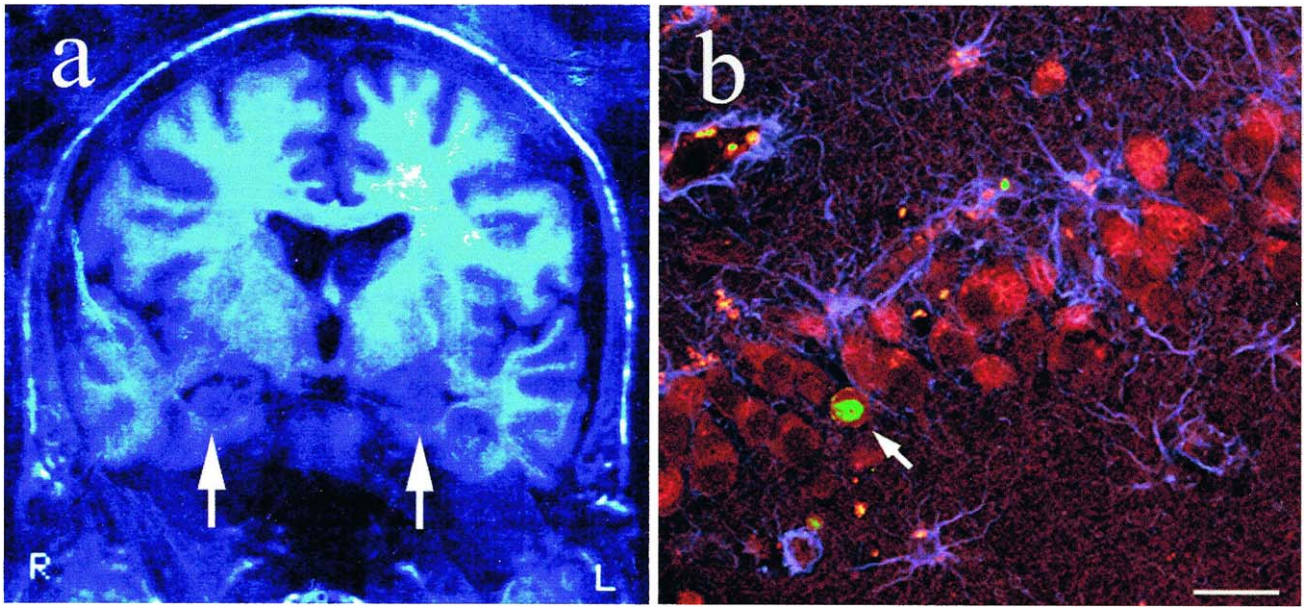


Representative MRI scan (left), SPECT image (center), and superimposed MRI/SPECT image (right) in a normal subject and in a patient with objective evidence of lyme encephalopathy. Blue corresponds to lower perfusion, red to higher perfusion. See article LYME ENCEPHALOPATHY.

Plate 8

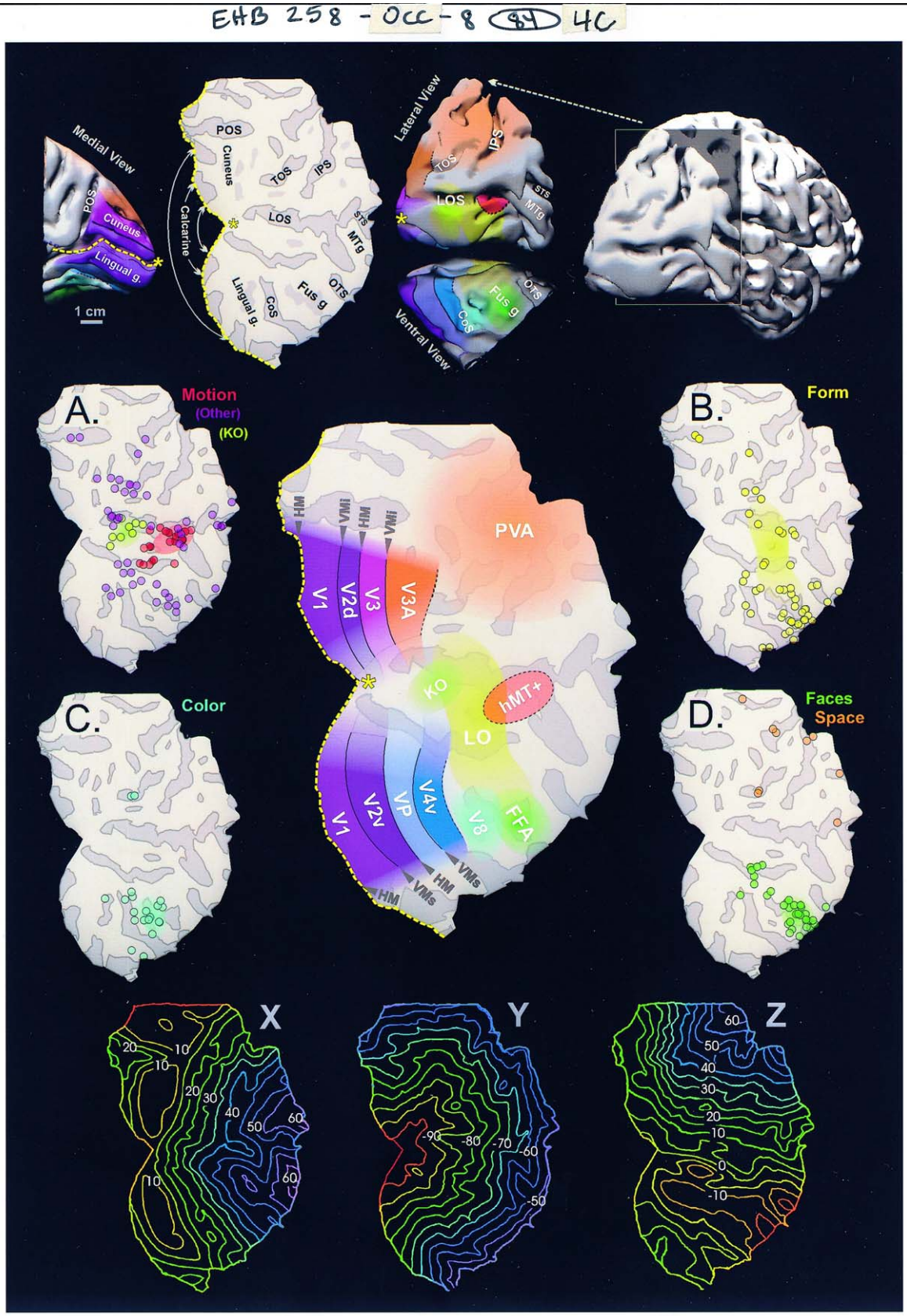
major limbic centers and connections, based on a computer reconstruction of limbic structures traced and digitized from whole brain serial sections. From Cheryl A. Cotman, Jay B. Angevine, Jr., and Kevin Head (illustration by Cheryl A. Cotman; see article NERVOUS SYSTEM, ORGANIZATION OF.

Plate 9



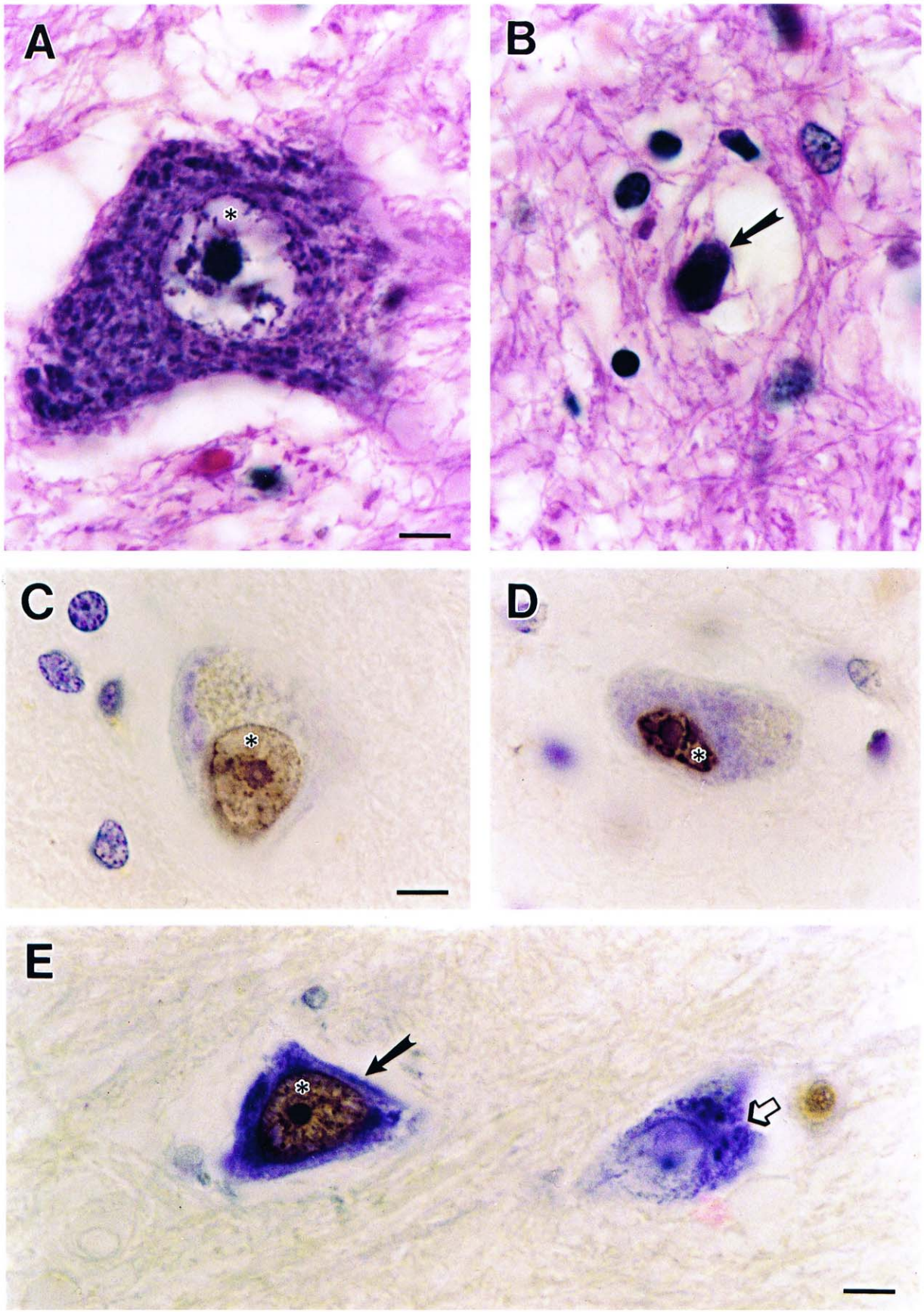
MRI image of the human brain showing the location of the hippocampi (arrows) within the medial portions of the temporal lobe. Right, confocal image of granule cells within the human dentate gyrus (hippocampus). The arrows show the nucleus (green) of a newborn, NeuN immunoreactive, neuron (red) within the adult human brain. Astroglial cells are immunostained for glial fibrillary protein (blue). From Eriksson et al. (1998). Nature Med. 4, 1313–1317. See article NERVE CELLS AND MEMORY.

Plate 10



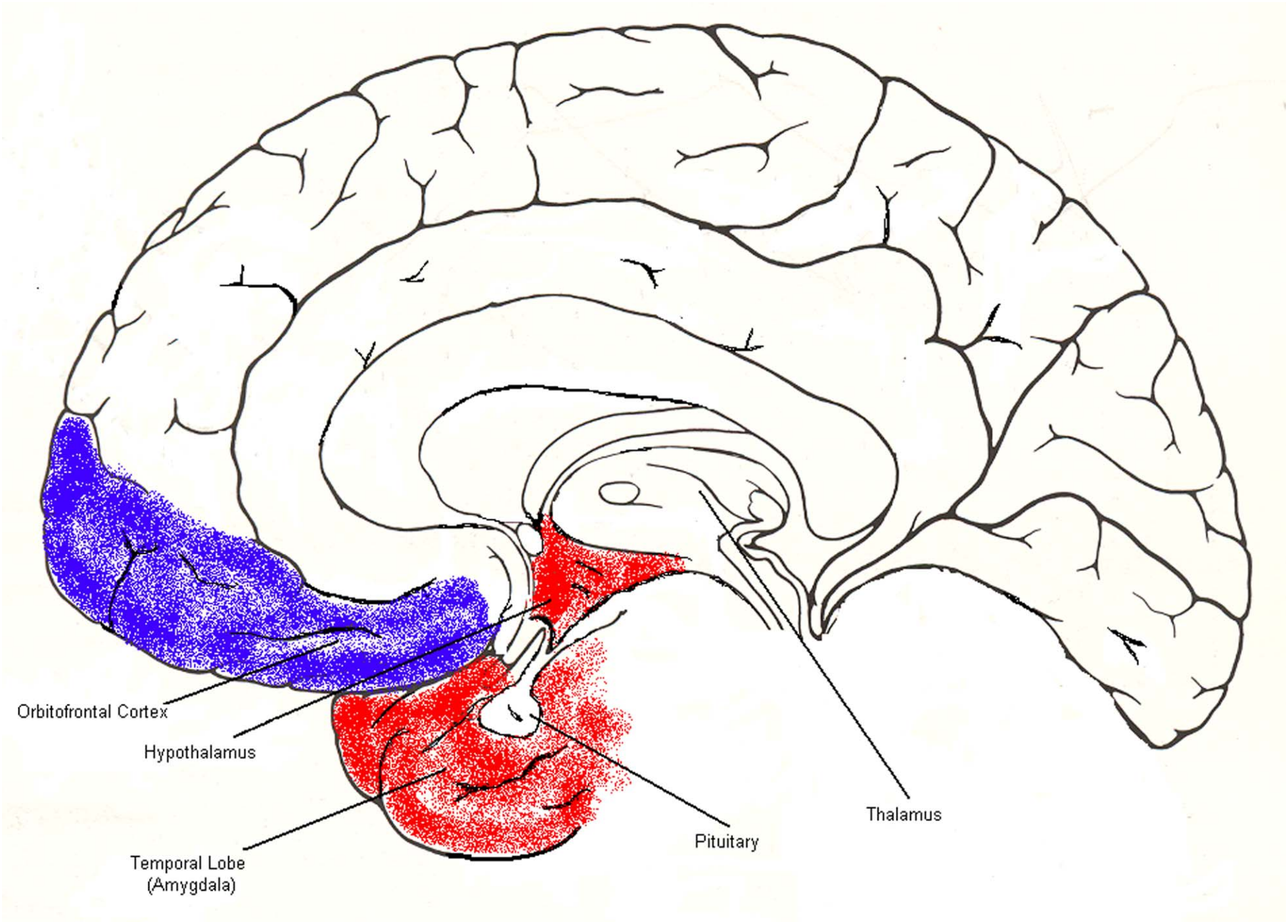
Topography and function of visual areas in human occipital lobe and neighboring cortex. (Top) computer graphics model of the Talairach brain shows anatomical features and topography of visual areas. (Center) Enlarged flat map of occipital cortex shows topograph of visual areas. (Bottom) Flat maps show Talairach coordinate isocontours (in mm). CoS, collateral sulcus; Fus g, fusiform gyrus; IPS, intraparietal sulcus; KO, Kinetic occipital visual area; LO, lateral occipital visual complex; LOS lateral occipital sulcus; MTg, middle temporal gyrus; OTS, occipitotemporal sulcus; POS, parietal occipital sulcus; STS superior temporal sulcus, TOS, transverse occipital sulcus. See article OCCIPITAL LOBE.

Plate 11



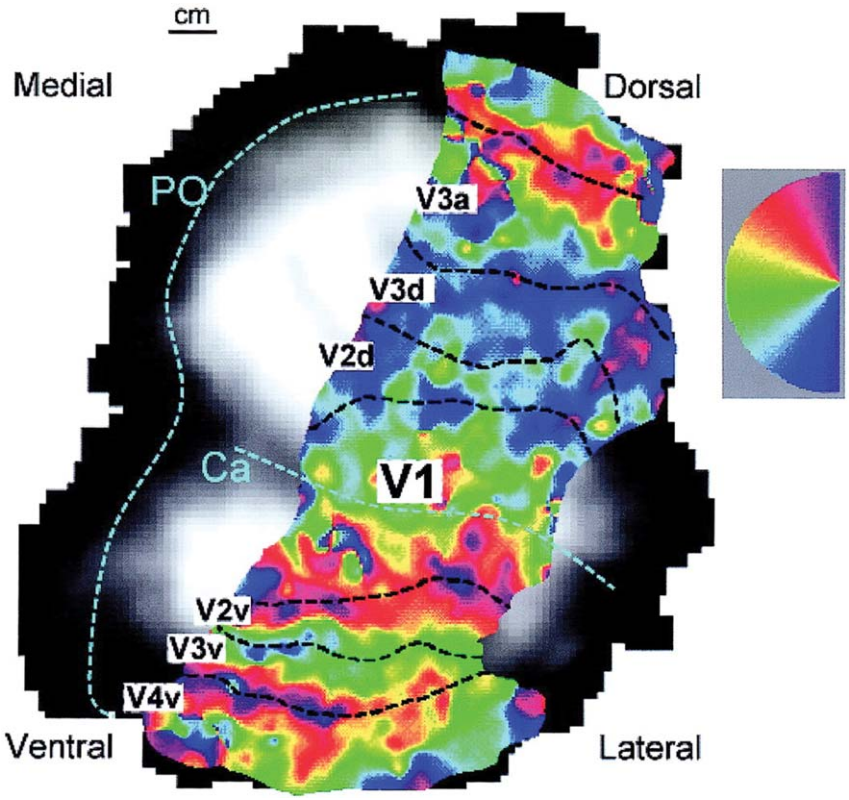
Motor neuron degeneration in ALS (amyotropic lateral sclerosis) is a form of apoptosis that may be mediated by a p53-dependent mechanism triggered by DNA damage. See article NEURODEGENERATIVE DISORDERS.

Plate 12



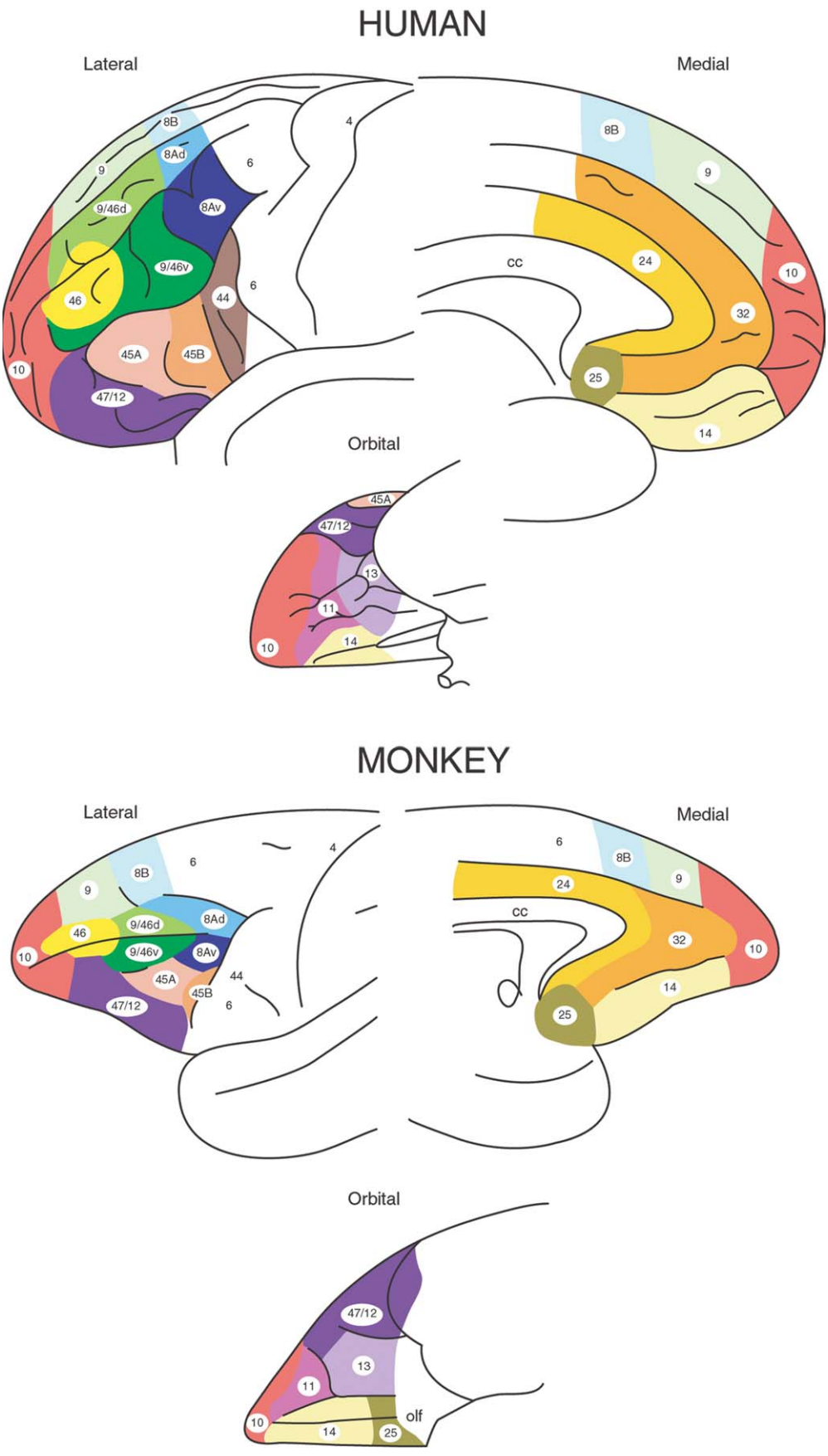
Regulation of aggression by brain regions. Orbitofrontal cortex, shown in blue, modulates decreases in aggression. Limbic regions and hypothalamus shown in red, modulate increases in aggression. See article VIOLENCE.

Plate 13



Retinotopic mapping in the human cortex measured with fMRI. Only the azimuth is mapped; inset gives the color key to angle. The horizontal meridian is green and the vertical meridian Blue. The map, onto a mathematically flattened cortex, shows the transitions between retinotopic areas along the meridians. Reproduced with permission from wandell, 1999. See article VISION.

Plate 14



Schematic diagram illustrating the cytoarchitectonic organization of the lateral, medial, and orbital surfaces of the frontal region in the human (top) and monkey (bottom) cerebral cortices. From Petrides and Pandya, 1994. See article PREFRONTAL CORTEX.